BPSM Assignment 1

An RNA-seq analysis pipeline on the effect of life cycle phase on *Trypanosoma Brucei* gene expression.

GitHub Address: https://github.com/IshIshIsh/Learning.git

Pipeline Introduction:

The script for the assignment is named BPSM_Assignment1.sh and contains the basic code for an RNA-Seq analysis pipeline, using paired-end reads with the software fastqc, bowtie2, samtools and bedtools.

The pipeline makes several assumptions:

- Raw sequence files are in the fq or fq.gz format and stored in one folder (set by variable fastqc_data_dir) which also includes a condition mapping file.
- A gzipped fasta format genome to create a bowtie2-format genome index is provided in a separate folder (specified by the genome variable).
- prior instalation of fastqc, bowtie2, samtools and bedtools on unix
- The pipeline will be run from a home directory and will create the relevant folders and within the variable path working directory.
- There are no introns in the genes

Variables should be set by the user either before or during the pipeline. These allow the user to change the input files and folders for use on a new dataset but have default values for the current assignment.

Pipeline Difficulties:

Time has been the main issue in the generation of the pipeline as unfortunately I've had several project and grant deadlines which had to take priority.

Drop-outs/Time-outs in my connection through the VPN have also been a significant issue in creating the pipeline, leading me to investigate running the processes in the background.

I note that my bedtools multicov output has only one count per gene per life stage, while the assignment requests the statistical average of counts therefore further investigation into previous steps is merited. I've put the counts per gene per life stage in the output file rather than the mean counts per gene per life stage.

Pipeline Features:

- Creates folder structure for neat maintenance of files, with a new folder created in the working directory for each analysis stage.
- Creates a summary of the QC results to screen and saves a file with an array of failed reads for fastqc tests output to a file for ease of exclusion in subsequent analysis.
- Additional fastqc output summary format as follows, to be more helpful on selecting individual reads to ignore:

  read test1 test2 test3

  216_1 PASS PASS WARN

  216_2 FAIL WARN FAIL

- Threading can be set with the thread's variable. Note: It is recommended to use the cat /proc/cpuinfo command to select an appropriate level for the active computer. In bowtie memory mapping is also used so that processes can share the same memory image of the index.
- pipes the bowtie output to SAMtools rather than storing the uncompressed format to decrease storage size.

Pipeline Features still to add:

- Investigating optimization of the Bowtie2 command using:
  - --met-file<path> command gives metric options for debugging performance issues
  - --align-paired-reads option
- Calculating the statistical average of the counts per gene per life cycle.
- Run pipeline in a background job so that if ssh connection to server is dropped it will continue processing.

Use of the pipeline:

As long as the assumptions are met and the variables are set correctly the script should run without user input. If a user wishes to exclude or trim reads from the pipeline, additional code would be required. Exclusion of reads based on the fastqc testing is facilitated by the array of reads failing specific tests or based on the table output which gives pass/fail/warn output per read.

Variable	**Default Value**	**Description**
working_directory	BPSM_assignment1	The folder path to the working directory containing all required files and where the output will be stored.
fastqc_data_dir	"$working_directory/fastq"	The folder path to the fq or fq.gz files.
threads	16	threading is used at an appropriate level for the active computer - please check your system before running

genome	"$working_directory/Tbb_genome"	The folder path to the reference genome
zipfilename	Tb927_genome.fasta.gz	The name of the reference genome zip file
fasta	Tb927_genome.fasta	The name of the fasta file used in the creation of a Bowtie2 index genome
genomename	Tb927_genome	The name of the genome to be used for the output of the Bowtie2, Samtools and Bedtools output labels.
mapfile	fqfiles	The name of the file which contains the mapping of which pairs of reads belong to each life cycle
bedfile	"$working_directory/Tbbgenes.bed"	The bedfile for mapping the reads to specified areas of the genome (genes/psudogenes)

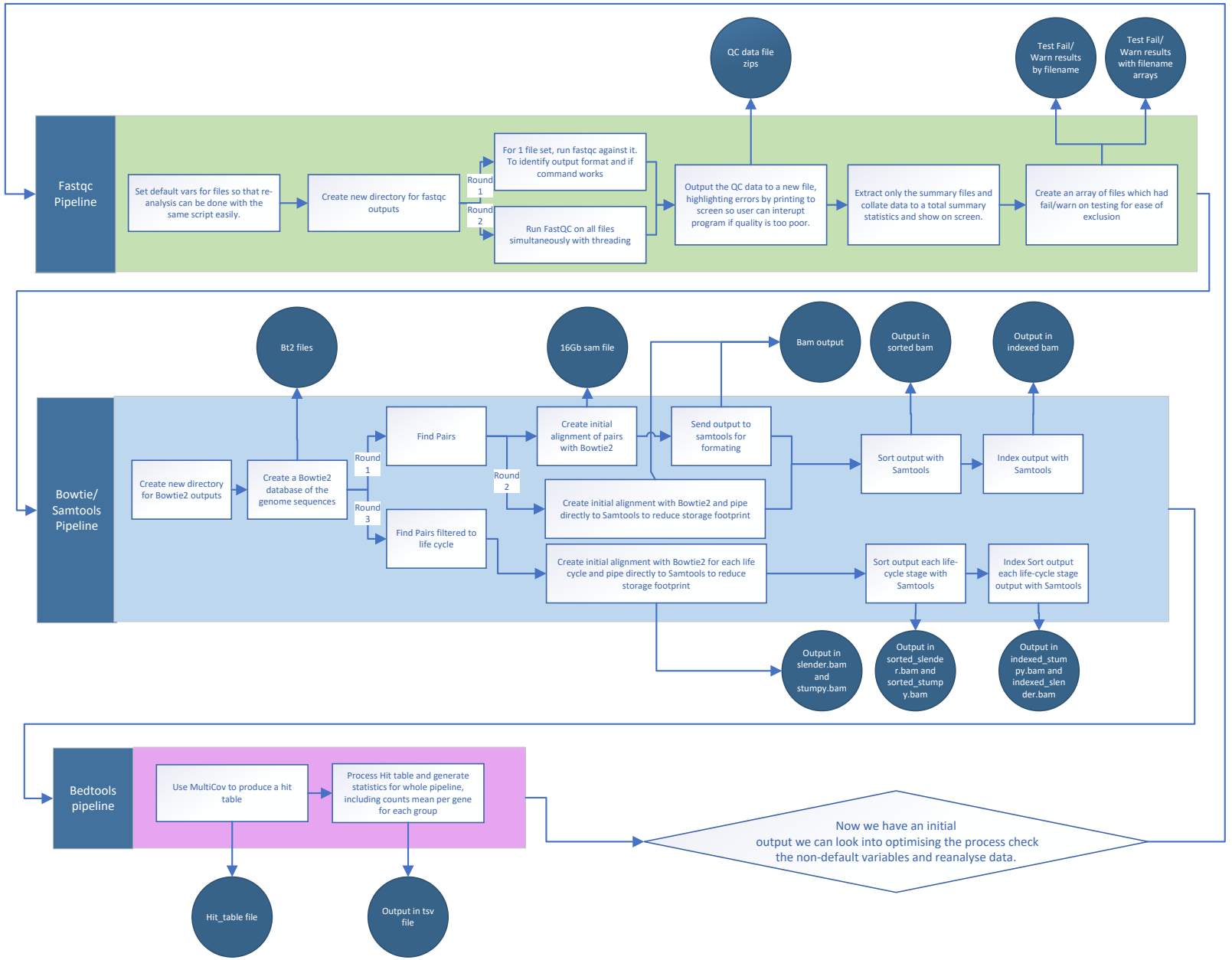
Bowtie2 outputs:

Stumpy Bowtie2 Output

```
6000000 reads; of these:
  6000000 (100.00%) were paired; of these:
    331767 (5.53%) aligned concordantly 0 times
    2903264 (48.39%) aligned concordantly exactly 1 time
    2764969 (46.08%) aligned concordantly >1 times
    ----
    331767 pairs aligned concordantly 0 times; of these:
      41322 (12.46%) aligned discordantly 1 time
    ----
    290445 pairs aligned 0 times concordantly or discordantly; of these:
      580890 mates make up the pairs; of these:
        360258 (62.02%) aligned 0 times
        71511 (12.31%) aligned exactly 1 time
        149121 (25.67%) aligned >1 times
97.00% overall alignment rate
```

Slender Bowtie2 Output

```
6000000 reads; of these:
  6000000 (100.00%) were paired; of these:
    972062 (16.20%) aligned concordantly 0 times
    2515176 (41.92%) aligned concordantly exactly 1 time
    2512762 (41.88%) aligned concordantly >1 times
    ----
    972062 pairs aligned concordantly 0 times; of these:
      31445 (3.23%) aligned discordantly 1 time
    ----
    940617 pairs aligned 0 times concordantly or discordantly; of these:
      1881234 mates make up the pairs; of these:
        1607159 (85.43%) aligned 0 times
        79607 (4.23%) aligned exactly 1 time
        194468 (10.34%) aligned >1 times
86.61% overall alignment rate
```


Fastqc Pipeline
Set default vars for files so that re-analysis can be done with the same script easily.
Create new directory for fastqc outputs
Round 1
Round 2
For 1 file set, run fastqc against it. To identify output format and if command works
Run FastQC on all files simultaneously with threading
Output the QC data to a new file, highlighting errors by printing to screen so user can interupt program if quality is too poor.
QC data file zips
Extract only the summary files and collate data to a total summary statistics and show on screen.
Create an array of files which had fail/warn on testing for ease of exclusion
Test Fail/ Warn results by filename
Test Fail/ Warn results with filename arrays
Bowtie/ Samtools Pipeline
Bt2 files
Create new directory for Bowtie2 outputs
Create a Bowtie2 database of the genome sequences
Round 1
Round 3
Find Pairs
Round 2
Find Pairs filtered to life cycle
16Gb sam file
Create initial alignment of pairs with Bowtie2
Create initial alignment with Bowtie2 and pipe directly to Samtools to reduce storage footprint
Create initial alignment with Bowtie2 for each life cycle and pipe directly to Samtools to reduce storage footprint
Send output to samtools for formating
Bam output
Output in sorted bam
Output in indexed bam
Sort output with Samtools
Index output with Samtools
Sort output each life-cycle stage with Samtools
Index Sort output each life-cycle stage output with Samtools
Output in slender.bam and stumpy.bam
Output in sorted_slender.bam and sorted_stumpy.bam
Output in indexed_stumpy.bam and indexed_slender.bam
Bedtools pipeline
Use MultiCov to produce a hit table
Process Hit table and generate statistics for whole pipeline, including counts mean per gene for each group
Hit_table file
Output in tsv file
Now we have an initial output we can look into optimising the process check the non-default variables and reanalyse data.