10

11

8

6

5

4

3

7

12

9

Start

RBK GUDMAP BDBAG (ALL)

Study RID

Filter fastq.gz only

Split into Replicate BDBAGs

BDBAG fetch

Replicate RID

file.csv

Experiment RID

File Name

BDBAG .zip

fetch.txt

Make Manifest

Manifest .csv

\*.fastq.gz

fastqc

\*.fastqc.zip

Trim Adapters

Align

Deduplicate

Dedup \*.bam

Count Genes/Transcripts

Make bigWig

\*.count.csv \*.tpm.csv  
\*.fpkm.csv

Average by Experiment RID

QC .json

MultiQC

deriva-upload-cli

End

\*.bigWig

Trimmed \*.fastq.gz

\*.bam

QC

Experiment Settings.csv

Paired\_End

Has\_Strand\_Specific\_Information

Used\_Spike\_Ins

Spike ?

Experiment.csv

Species

Hu/Mu?

References

PCA etc

2

1

Replicate\_\*  
(BDBAG folder split by replicates)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Process** | **Tools** | **Docker Container** | **Input (version if appropriate)** | **Output** | **Parallel Strategy** |
| splitData  1 | Python (3.6.4)  Pandas (0.25.1?) | bicf/bdbag | BDBAG .zip | file.csv  Experiment Settings.csv  Experiment.csv  Replicate\_\* (replicate separated BDBAG folders | Single Process |
| getData  1 | Python (3.6.4)  Pandas (0.25.1?)  BDBAG (1.5.5) | bicf/bdbag | deriva-cookies.txt  Replicate\_\* (replicate separated BDBAG folders | \*.fastq.gz | By replicate (fastq pair) |
| createManifest  2 | Python (3.6.4)  Pandas (0.25.1?) |  | file.csv | Manifest.csv | Single Process |
| trim  3 | Trim Galore (0.6.4?) |  | \*.fastq.gz | Trimmed \*.fastq.gz | By replicate (fastq pair) OR collated |
| align  4 | HISAT2 (v2.1.0?) |  | Experiment.csv  Species  Experiment Settings.csv  Paired\_End  Has\_Strand\_Specific\_Information  Used\_Spike\_Ins  Trimmed \*.fastq.gz  Reference:  Human = GRCh38p12  Mouse = GRCm38p6  Spike-in = ERCC92? | \*.bam  \*.alignment.summary.txt | By sample (fastq pair) |
| dedup  5 | Picard (2.20.7?) |  | \*.bam | Dedup \*.bam | By replicate (fastq pair) |
| count  6 | Subread (2.0.0?)  Python (3.6.4)  Pandas (0.25.1?)  OR  R (3.6.1?) |  | Dedup \*.bam | \*.count.csv  \*.tpm.csv  \*.fpkm.csv | By replicate (fastq pair) |
| aggregateCount  7 | Python (3.6.4)  Pandas (0.25.1?)  OR  R (3.6.1?) |  | \*.count.csv  \*.tpm.csv  \*.fpkm.csv  Manifest .csv | \*.count.average.csv  \*.tpm.averaged.csv  \*.fpkm.averaged.csv  \*.fpkm.cv.csv  PCA.tiff | Collated |
| makeBigWig  8 | samtools  deeptools |  | \*.bam | \*.bw | By replicate (fastq pair) |
| fastqc  9 | Fastqc (0.11.8?) |  | \*.fastq.gz | \*.fastqc.zip | By fastq |
| qc  10 | MultiQC (1.7?)  Python (3.6.4)  Pandas (0.25.1?) |  | \*.fastqc.zip  \*.alignment.summary.txt  \*.fpkm.averaged.csv  \*.fpkm.cv.csv | multiqc\_report.html  \*.qc.json | Collated |
| derivaUpload  11 | Python (3.6.4)  deriva (0.8.8) |  | \*.bam  \*.dedup.bam  \*.bw  \*.count.csv  \*.tpm.csv  \*.fpkm.csv  \*.count.average.csv  \*.tpm.average.csv  \*.fpkm.average.csv  \*.fastqc.zip  \*.qc.json |  | Collated |

Outputs for upload to data hub

Phase 2 pipeline… automatically upload files to the data hub… initially manually go through outputs and QCs and manually upload

Manual process

Nextflow process

DO WE SUPPORT?