Supplementary Information

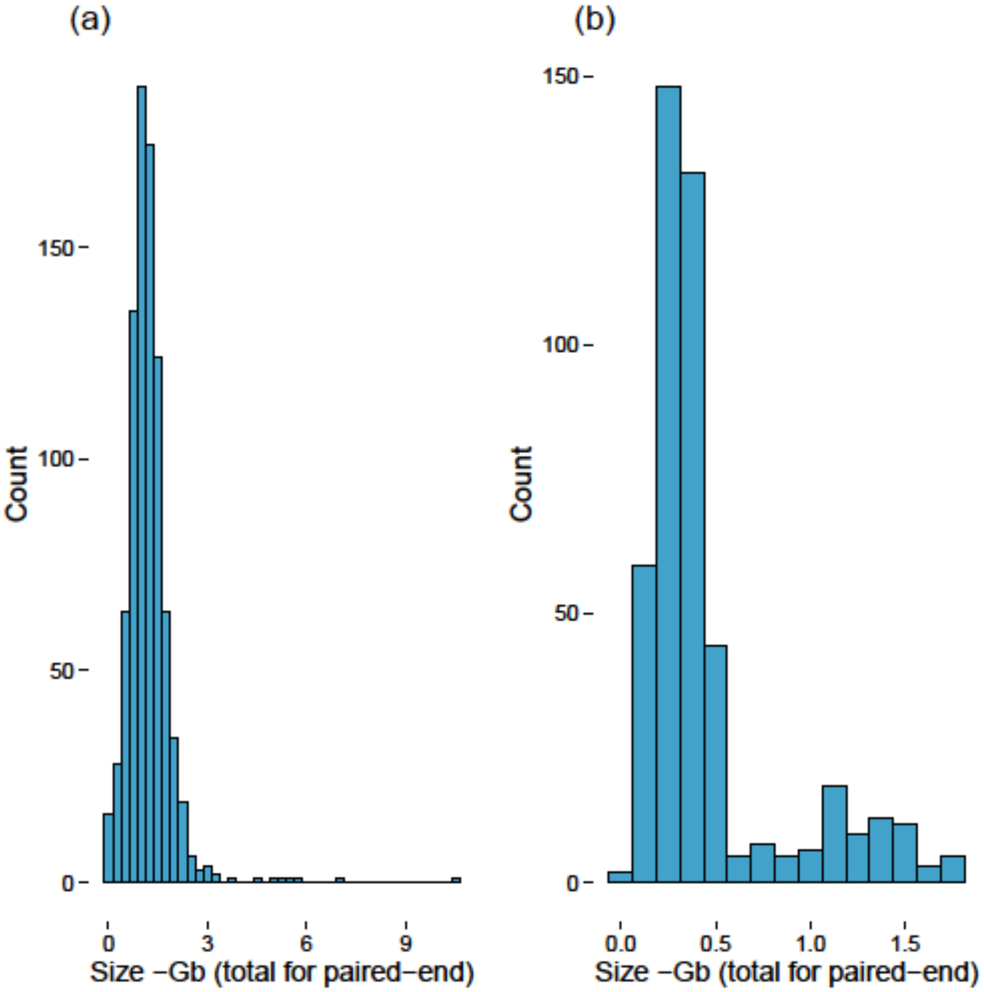
Falco: A quick and flexible single-cell RNA-seq processing framework on the cloud

Andrian Yang[1,2], Michael Troup[1], Peijie Lin[1,2] and Joshua WK Ho[1,2*]

[1]Victor Chang Cardiac Research Institute, Sydney, NSW, Australia
[2]St. Vincent's Clinical School, University of New South Wales, Sydney, NSW, Australia
[*]To whom correspondence should be addressed: j.ho@victorchang.edu.au

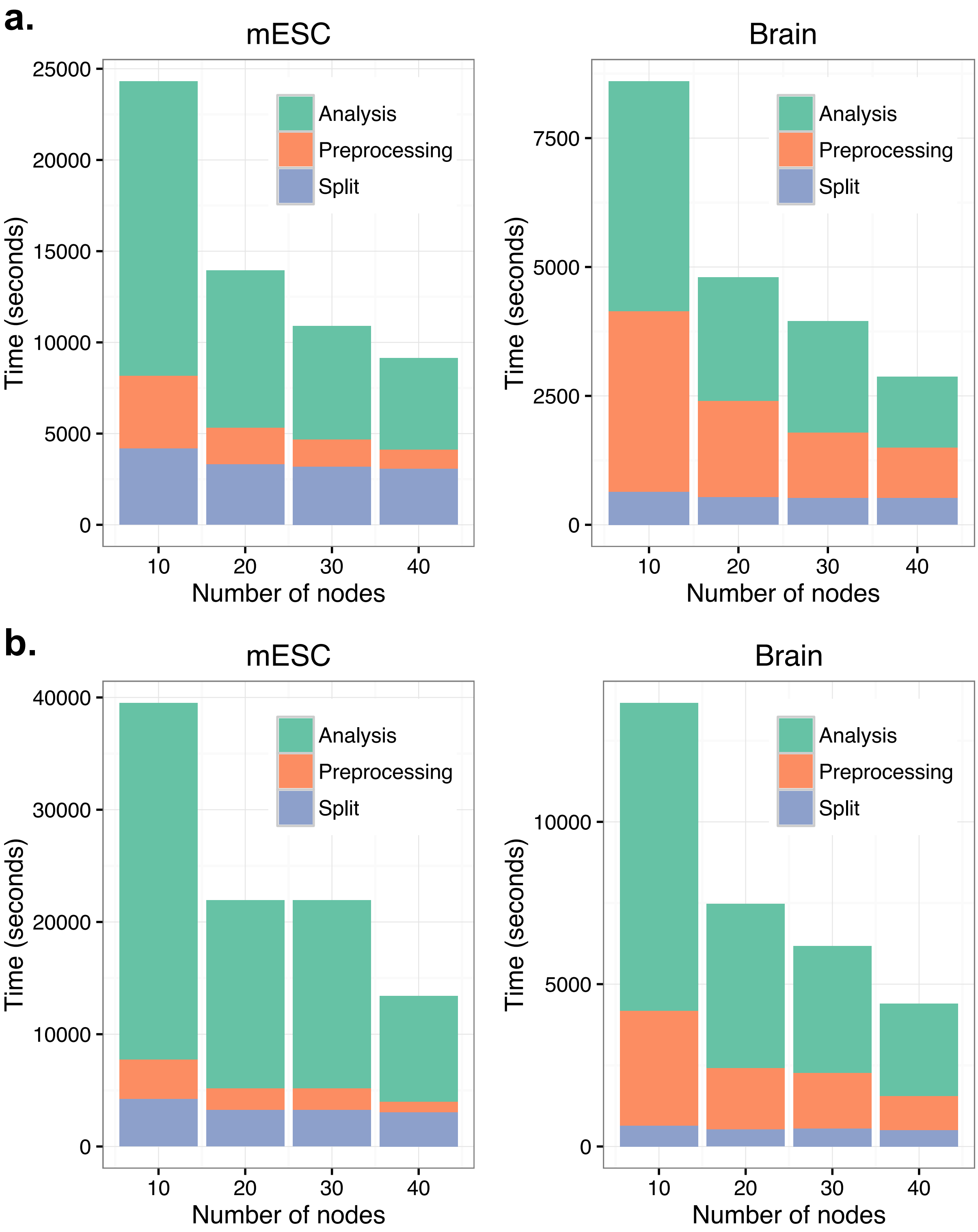
Availability: https://github.com/VCCRI/Falco

Supplementary Figure 1 – Datasets file sizes



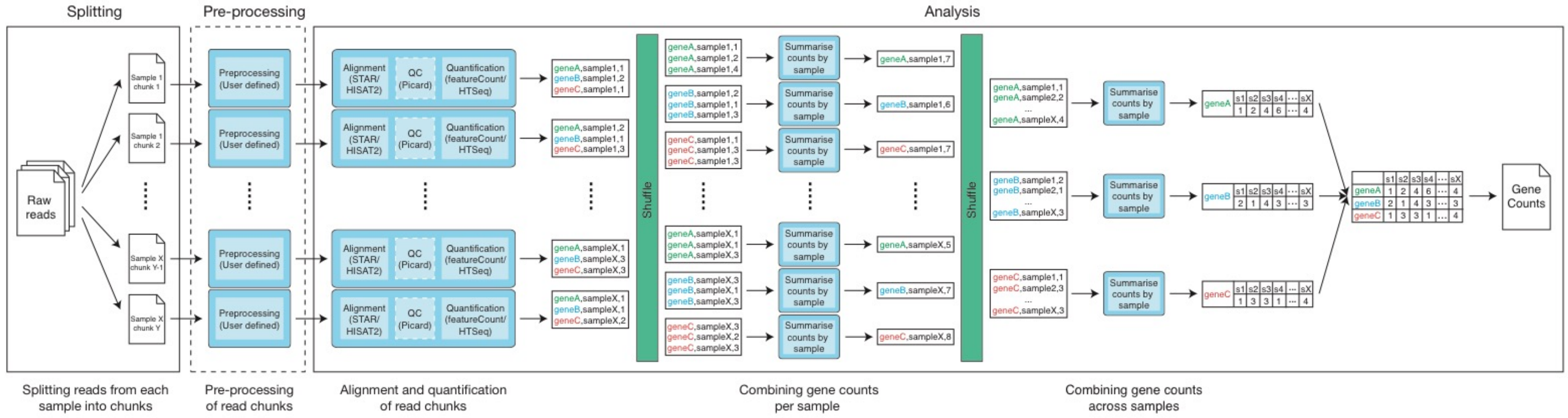
Supplementary Figure 1. Size of the FASTQ file (in terms of gigabytes) of individual cells in each single-cell RNA-seq data set. **(a)** Mouse embryonic stem cells (SRA accession: ERP005988). **(b)** Human brain cells (SRA accession: SRP057196).

Supplementary Figure 2 – Runtime by steps



Supplementary Figure 2. Falco processing time split by steps for STAR+featureCounts (a) and HISAT2+HTSeq (b) pipelines for mouse embryonic stem cell and human brain data analysis.

Supplementary Figure 3 – Process pipeline



Supplementary Figure 3. The Falco framework process pipeline. In the splitting step, reads from one or multiple FASTQ files are split into multiple chunks of size 256 Mb uncompressed. This step makes use of Apache Hadoop MapReduce. A pre-processing step is executed if the data require pre-processing using MapReduce. In the main analysis step, sequence alignment and gene expression quantification are carried out in a highly parallelised fashion using the Apache Spark framework.