Bioinformatics, 33(5), 2017, 767–769
doi: 10.1093/bioinformatics/btw732
Advance Access Publication Date: 13 December 2016
Applications Note

OXFORD

Gene expression

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 W. K. Ho[1,2,*]

[1]Victor Chang Cardiac Research Institute, Sydney, NSW, Australia and [2]St. Vincent's Clinical School, University of New South Wales, Sydney, NSW, Australia

*To whom correspondence should be addressed.

Associate Editor: Ziv Bar-Joseph

Received on July 15, 2016; revised on October 27, 2016; editorial decision November 11, 2016; accepted on November 16, 2016

Abstract

Summary: Single-cell RNA-seq (scRNA-seq) is increasingly used in a range of biomedical studies. Nonetheless, current RNA-seq analysis tools are not specifically designed to efficiently process scRNA-seq data due to their limited scalability. Here we introduce Falco, a cloud-based framework to enable paralellization of existing RNA-seq processing pipelines using big data technologies of Apache Hadoop and Apache Spark for performing massively parallel analysis of large scale transcriptomic data. Using two public scRNA-seq datasets and two popular RNA-seq alignment/feature quantification pipelines, we show that the same processing pipeline runs 2.6–145.4 times faster using Falco than running on a highly optimized standalone computer. Falco also allows users to utilize low-cost spot instances of Amazon Web Services, providing a ∼65% reduction in cost of analysis.

Availability and Implementation: Falco is available via a GNU General Public License at https://github.com/VCCRI/Falco/

Contact: j.ho@victorchang.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Major advancements in single-cell technology have resulted in an increasing interest in single-cell level studies, particularly in the field of transcriptomics (Kolodziejczyk *et al.*, 2015b). Single-cell RNA sequencing (scRNA-seq) offers the promise of understanding transcriptional heterogeneity of individual cells, allowing for a clearer understanding of biological process (Darmanis *et al.*, 2015; Grün *et al.*, 2015; Kolodziejczyk *et al.*, 2015a; Patel *et al.*, 2014).

Each scRNA-seq experiment typically generates profiles of hundreds of cells, which is a magnitude larger than the typical amount of data generated by standard bulk RNA-seq experiments. Current RNA-seq processing pipelines are not specifically designed to handle such a large number of profiles. To fully realize the potential of scRNA-seq, we need a scalable and efficient computational solution. The premise of our solution is that state-of-the-art cloud computing technology, which is known for its scalability, elasticity and pay-as-you-go payment model, can allow for a highly efficient and cost-effective scRNA-seq analysis.

There are a number of existing cloud-based next-generation sequencing bioinformatics tools based on the Hadoop framework, an open source implementation of MapReduce (Dean and Ghemawat, 2004), or the Spark framework (Zaharia *et al.*, 2010). Halvade, written in Hadoop MapReduce, is designed to perform variant calling of genomic data from FASTQ files, though it also offer support for transcriptomic analysis (Decap *et al.*, 2015). SparkSeq (Wiewiorka *et al.*, 2014) and SparkBWA (Abuín *et al.*, 2016), both written in Spark, offer interactive sequencing analysis of BAM files and alignment of FASTQ files respectively. These tools have limitations in the context of scRNA-seq analysis. Of the three tools, only SparkSeq allows for multi-sample analysis, although SparkSeq itself is also limited as it does not perform alignment, which is the main bottleneck in sequencing data analysis.

Table 1. scRNA-seq processing time with or without Falco

System	Nodes	Mouse – embryonic stem cell (hours)		Human – brain (hours)	
		S+F*	H+H*	S+F*	H+H*
Standalone	1 (1 process)	93.7	233.6	154.7	198.8
	1 (12 processes)	21.1	32.6	16.4	19.6
	1 (16 processes)	18.5	28.4	13.6	16.2
Falco	10	7	11.2	2.7	4.1
	20	4.1	6.4	1.6	2.3
	30	3.3	4.8	1.4	2.0
	40	2.8	4.0	1.1	1.5

*S + F = STAR for aligner and featureCounts for quantification; H + H = HISAT2 for aligner and HTSeq for quantification. Standalone number of processes indicates the number of FASTQ file pairs that are processed in parallel. Timing for Falco includes initialization and configuration time which are approximately 16 minutes.

Table 2. Falco cost analysis: on-demand versus spot instances for STAR + featureCounts

Dataset	Cluster size	Time (hours)	On-demand cost (USD$)	Spot cost (USD$)	% Savings
Mouse -	10 nodes	8	247.20	85.67	65.34
embryonic	20 nodes	5	301.00	99.09	67.08
stem cell	30 nodes	4	258.00	115.71	55.15
	40 nodes	3	356.40	114.11	67.98
Human -	10 nodes	3	92.70	32.13	65.34
brain	20 nodes	2	120.40	39.64	67.08
	30 nodes	2	179.00	57.86	67.68
	40 nodes	2	237.60	76.08	67.98

Time rounded up to whole hour including cluster startup. Price used for r3.8xlarge instance is USD$2.660/hr (on-demand price) and USD$0.64/hr (average spot price for June 2016).

Here we propose a different approach to utilizing cloud-based big data technology. Our framework—Falco—is a software bundle that allows users to 'plug-in' their chosen RNA-seq alignment, quality control, preprocessing and feature quantification tools, and enables the resulting pipeline to run multi-sample analysis of large-scale transcriptomic data on the cloud (See Supplementary Note 1 for details of the framework). Falco utilizes Amazon Elastic MapReduce (EMR), a big data processing service for deploying managed Hadoop and Spark clusters on the Amazon Web Services (AWS) cloud.

2 Evaluation

To evaluate the performance of Falco, the runtimes of two popular RNA-seq pipelines, STAR (Dobin *et al.*, 2013) followed by featureCounts (Liao *et al.*, 2014) (S + F), and HISAT2 (Kim *et al.*, 2015) followed by HTSeq (Anders *et al.*, 2014) (H + H), are evaluated using two scRNA-seq datasets (Darmanis *et al.*, 2015; Kolodziejczyk *et al.*, 2015a) with and without using the Falco framework. A number of realistic scenarios for analysis in a single computing node were devised—from the naïve single processing approach to a highly parallelized approach. Furthermore, to demonstrate the scalability of Falco, EMR clusters with increasing numbers of core nodes (from 10 to 40) were used to show the effect of adding more computational resources on the runtime of Falco (See Supplementary Note 2 for experimental details).

Comparing the performance of a single node, with different parallelization approaches, against Falco shows that running the S + F pipeline on Falco results in a speed-up of 2.6× (10 nodes versus 16 processes) to 33.4× (40 nodes versus 1 process) for the mouse dataset and 5.1× (10 nodes versus 16 processed) to 145.4× (40 nodes versus 1 process) for the human dataset. For the H + H pipeline, Falco gives a speedup of 2.5× (10 nodes versus 16 processes) to 58.4× (40 nodes versus 1 process) and 4.0× (10 nodes versus 16 processes) to 132.5× (40 nodes versus 1 process) for the mouse and human datasets respectively (Table 1). The disparity in the speed-up between the two datasets is due to different pre-processing tools being employed, with the human dataset utilizing more pre-processing steps in the original publication (Darmanis *et al.*, 2015). We also note that the gene expression quantification produced by a given pipeline is the same regardless of whether the Falco framework was used.

For the scalability comparison, it can be seen that the runtime of the pipeline decreases with increasing cluster size (Table 1), though the trend is gradual rather than linear. Analysis of the runtime for each step in the framework shows a similar gradual decrease in runtime for pre-processing and analysis steps (Supplementary Fig. S2). For the splitting step, a different trend is seen where there is little to no decrease in runtime for cluster size $\geq$ 20 nodes. The lack of speed up for splitting is due to the number of executors exceeding the number of files to be split and the limitation of time taken to split large files as the distribution of file size in both test datasets is uneven (Supplementary Fig. S1).

To save cost, EMR allows for the usage of reduced price *spot* computing resources. The spot prices fluctuate depending on the availability of the unused computing resource and the spot instance is obtained by supplying a bid for the resource. The use of spot instances for analysis provides a substantial saving of around ~65% compared to using on-demand instances (Table 2 and Supplementary Table S1). The trade-off with using spot instances is that the computing resource could be terminated should the market price for that resource exceed the user's bid price. In summary, Falco is a cloud-based framework that enables massively parallelized sequence alignment, quality control and feature quantification of single-cell transcriptomic data in the AWS cloud-computing environment.

Funding

This work was supported in part by funds from the New South Wales Ministry of Health, a National Health and Medical Research Council/National Heart Foundation Career Development Fellowship (1105271), a Ramaciotti Establishment Grant (ES2014/010), an Australian Postgraduate Award and Amazon Web Services (AWS) Credits for Research.

Conflict of Interest: none declared.

References

Abuín,J.M. *et al*. (2016) SparkBWA: speeding up the alignment of high-throughput DNA sequencing data. *Plos One*, **11**, e0155461.

Anders,S. *et al*. (2014) HTSeq A Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

Darmanis,S. *et al*. (2015) A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci*., **112**, 7285–7290.

Dean,J. and Ghemawat,S. (2004) MapReduce: Simplified Data Processing on Large Clusters. In: *Proceedings of the Sixth Symposium on Operating System Design and Implementation (OSDI), OSDI'04*. USENIX Association.

Decap,D. *et al*. (2015) Halvade: scalable sequence analysis with MapReduce. *Bioinformatics*, **31**, 2482–2488.

Dobin,A. *et al*. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, **29**, 15–21.

Grün,D. *et al*. (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, **525**, 251–255.

Kim,D. *et al*. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.

Kolodziejczyk,A.A. *et al*. (2015a) Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, **17**, 471–485.

Kolodziejczyk,A.A. *et al*. (2015b) The technology and biology of single-cell RNA sequencing. *Mol. Cell*, **58**, 610–620.

Liao,Y. *et al*. (2014) FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.

Patel,A.P. *et al*. (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, **344**, 1396–1401.

Wiewiorka,M.S. *et al*. (2014) SparkSeq: fast, scalable and cloud-ready tool for the interactive genomic data analysis with nucleotide precision. *Bioinformatics*, **30**, 2652–2653.

Zaharia,M. *et al*. (2010). Spark: Cluster Computing with Working Sets. In: *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, p. 10. USENIX Association.