Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Applications Note

Sequence analysis

SeqBox: RNAseq/ChIPseq reproducible analysis on a consumer game computer

Marco Beccuti[1,*], Francesca Cordero[1], Maddalena Arigoni[2], Riccardo Panero[2], Elvio G. Amparore[1], Susanna Donatelli[1] and Raffaele A Calogero[2]

[1]Department of Computer Sciences, University of Torino, Corso Svizzera 185, Torino, Italy, [2]Department of Molecular Biotechnology and Health Sciences, Via Nizza 52, Torino, Italy.

*beccuti@di.unito.it.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Short reads sequencing technology has been used for more than a decade now. However, the analysis of RNAseq and ChIPseq data is still computational demanding and the simple access to raw data does not guarantee results reproducibility between laboratories. To address these two aspects, we developed **SeqBox**, a cheap, efficient and reproducible RNAseq/ChIPseq hardware/software solution based on NUC6I7KYK mini-PC (an Intel consumer game computer with a fast processor and a high performance SSD disk), and Docker container platform. In SeqBox the analysis of RNAseq and ChIPseq data is supported by a friendly GUI. This allows access to fast and reproducible analysis also to scientists with/without scripting experience.

Availability and Implementation: Docker container images, docker4seq package and the GUI are available at http://www.bioinformatica.unito.it/reproducibile.bioinformatics.html.

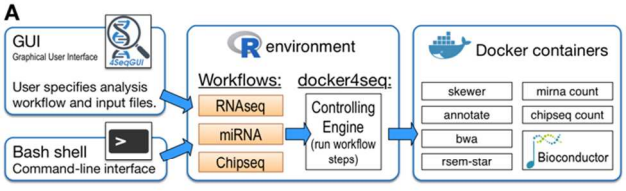
Contact: beccuti@di.unito.it

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Whole transcriptome sequencing (WTS) and ChIPseq made obsolete the corresponding array hybridization based technologies. Short reads sequencing technology has been used for more than a decade now, and experience shows that the main bottleneck in sequencing workflows is the time spent in analyzing and interpreting data (Wong et al. 2017).

The primary analysis of the data, i.e. mapping short read sequences on the reference genome, is still computationally demanding and requires computer performances that are not commonly available in laptops. In particular WTS requires a significant amount of RAM and multicores processors. The needs of high performance computing infrastructure for the analysis of sequencing data has brought to the development of cloud based analysis tools, e.g. Illumina BaseSpace (https://basespace.illumina.com/home/index), Galaxy (https://usegalaxy.org/), etc. However, cloud based solutions suffer of some criticalities, e.g. data uploading speed, limited storage space and significant computing and data transfer costs. Moreover, although all available data analysis platforms guarantee a certain level of reproducibility, typically storing the version of the software being used, tracking changes in the system libraries, which might lead to sneaky reproducibility issues, is not provided.



B

	Nº of CPU cores	CPU type	Clock Speed	Total RAM	Storage type	Hardware type	Software configuration
SeqBox	8	i7-6770HQ	3.5 GHz	32 GB	SSD	Mini-pc (1)	Fedora
SGI UV2000	16	E5-4650	2.4 GHz	1024 GB	SATA	Server (10)	Red Hat

Figure 1: A) SeqBox framework: depicting the structure of SeqBox and its functionalities from a user point of view. The analysis flows from left to right. B) Characteristics of the hardware used to evaluate the SeqBox performances.

To combine reproducible data generation with cost effective but efficient hardware we have developed **SeqBox**, a software/hardware ecosystem providing the most common analyses of RNAseq and ChIPseq data (i.e. genomic mapping, experimental power evaluation, differential expression, transcription factors/histone-marks peaks identification, etc.) on a consumer game computer (Figure 1A).

2 Methods

The SeqBox ecosystem (Figure 1A) is the union of SeqBox software and SeqBox target hardware.

A user can access the system either through a Java-based graphical interface (GUI, see supplementary data), or through R console (see supplementary data). Independently of the access type, the user can exploit three different workflows: RNAseq, miRNA and ChipSeq, which are managed by a Controlling Engine written in R (Figure 1A). The functions that realize the workflows are either standard analysis algorithms or a set of supporting functions that have been developed and included in Bioconductor packages (e.g. DESeq2, ChIPpeakAnno). The algorithms used for sequencing data analysis include STAR (Dobin, et al., 2013) for RNAseq genomic mapping, DESeq2 (Love, et al., 2014) for differential expression analysis, BWA (Li and Durbin, 2009) for ChIPseq genomic mapping, MACS (Zhang, et al., 2008), and SICER (Xu, et al., 2014) for ChIP peaks detection (see supplementary data). All of them are encapsulated into Docker images.

A Docker image is a lightweight, stand-alone, executable package that includes everything needed to run a specific software. A runtime instance of an image, called container, runs completely isolated from the host environment except for user-specified host files. The advantage of using Docker images is that the whole environment is fixed, the images are available in the Docker repository, and the identity of the images is the only element needed to reproduce the results. The execution of the Docker images, implementing the workflow chosen by the user, is done by docker4seq, a R package which embeds a set of functions providing the running parameters to the mapping and counting engine.

SeqBox provides six Docker images: (1) skewer.2017.01, which uses skewer (Jiang, et al., 2014) for adapter trimming; (2) rsemstar.2017.01, which uses STAR (Dobin, et al., 2013) to map short reads mapping on the reference genome and RSEM (Li and Dewey, 2011) for gene and isoform-level quantification; (3) annotate.2017.01, which is used to associate RSEM output id with gene symbols; (4) mirnaseq.2017.01, which implements the miRNAseq analysis workflow described in (Cordero, et al., 2012); (5) r332.2017.01, which allows differential expression analysis via Bioconductor package DESeq2; (6) chipseq.2017.01, which uses BWA (Li and Durbin, 2009) to map short reads on the reference genome, MACS (Zhang, et al., 2008) to detect transcription factors binding sites, and SICER (Xu, et al., 2014) to define histone-marks.

The GUI provides a graphical access to the docker4seq functions allowing the use of the tools to biologists without scripting experience.

SeqBox hardware: The parameters setting of the algorithms (in terms of memory size vs number of assigned cores) is optimized for an execution on the game computer NUC617KYK (Figure 1A), which is based on an Intel Core I7, featuring 4 cores running up to 8 threads that share a common memory of 32Gb and a SSD disk of 256 GB.

3 Results

We benchmarked SeqBox with respect to a high-end server (SGI UV2000, Figure 1B). The performance comparison was done for the three workflows (see supplementary data Figures 1, 2 and 3). In brief, we compared the workflows using increasing amounts of reads (see supplementary data) on SeqBox, using 8 threads, and on the SGI server increasing the number of threads from 8 to 160 (Supplementary Figures 1, 2 and 3). Parallelization provided by the SGI server did not improve very much the overall performances in the RNAseq workflow (Supplementary Figure 1). SeqBox significantly outperformed the server, because of the presence of a SSD with high I/O performance which can cope with the limited parallelism of SeqBox. In the case of miRNA and ChIPseq workflows the parallelization is only available for the reads mapping procedures. The limited parallelization of these two workflows combined with the higher I/O performances of the SSD with respect to the SATA array makes SeqBox extremely effective even with very high number of reads to be processed (Supplementary Figures 2 and 3).

4 Conclusion

The majority of the algorithms used in the considered bioinformatics workflows is strongly I/O bound and exhibits a limited exploitation of parallelism. Our experiments show that a combination of a consumer computer with a fast storage is able to over-perform a high-end server. The integration of Docker technology within a mini-PC consumer computer such as Intel NUC6I7KYK provides therefore, to small biology laboratories, a solution for Next Generation Sequence (NGS) analysis which is cheap, efficient and reproducible.

Funding

This work has been supported by the EPIGEN FLAG PROJECT

Conflict of Interest: none declared.

References

Cordero, F.*, et al.* Optimizing a massive parallel sequencing workflow for quantitative miRNA expression analysis. *PloS one* 2012;7(2):e31630.

Dobin, A.*, et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;29(1):15-21.

Jiang, H.*, et al.* Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC bioinformatics* 2014;15:182.

Li, B. and Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* 2011;12:323.

Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25(14):1754-1760.

Love, M.I., Huber, W. and Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* 2014;15(12):550.

Wong, S et al. ARC 2017 Proceedings, Wong S, Beck AC, Bertels K, Carro L Eds Springer 2017

Xu, S.*, et al.* Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. *Methods Mol Biol* 2014;1150:97-111.

Zhang, Y.*, et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* 2008;9(9):R137.