Bioinformatics, 32(16), 2016, 2502–2504
doi: 10.1093/bioinformatics/btw157
Advance Access Publication Date: 7 April 2016
Applications Note

OXFORD

Genome analysis

MGEScan: a Galaxy-based system for identifying retrotransposons in genomes

Hyungro Lee[1,†], Minsu Lee[2,†], Wazim Mohammed Ismail[1], Mina Rho[3], Geoffrey C. Fox[1], Sangyoon Oh[4,*] and Haixu Tang[1,*]

[1]School of Informatics and Computing, Indiana University, Bloomington, IN, USA, [2]Department of Computer Science and Engineering, Ewha Womans University, Seoul, Korea, [3]Department of Computer Science and Engineering, Hanyang University, Seoul, Korea and [4]Department of Software Convergence Technology, Ajou University, Suwon, Korea

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
Associate Editor: John Hancock

Received on December 9, 2015; revised on March 4, 2016; accepted on March 17, 2016

Abstract

Summary: MGEScan-long terminal repeat (LTR) and MGEScan-non-LTR are successfully used programs for identifying LTRs and non-LTR retrotransposons in eukaryotic genome sequences. However, these programs are not supported by easy-to-use interfaces nor well suited for data visualization in general data formats. Here, we present MGEScan, a user-friendly system that combines these two programs with a Galaxy workflow system accelerated with MPI and Python threading on compute clusters. MGEScan and Galaxy empower researchers to identify transposable elements in a graphical user interface with ready-to-use workflows. MGEScan also visualizes the custom annotation tracks for mobile genetic elements in public genome browsers. A maximum speed-up of 3.26× is attained for execution time using concurrent processing and MPI on four virtual cores. MGEScan provides four operational modes: as a command line tool, as a Galaxy Toolshed, on a Galaxy-based web server, and on a virtual cluster on the Amazon cloud.
Availability and implementation: MGEScan tutorials and source code are available at http://mgescan.readthedocs.org/
Contact: hatang@indiana.edu or syoh@ajou.ac.kr
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Transposable elements (TEs) move within or between most eukaryotic genomes and constitute a significant portion of these genomes. It is known that nearly half of the human genome is derived from TEs (Lander *et al.*, 2001). The transposition of TEs generates new mutations that can have different effect, ranging from changes in genes expression to genomic instabilities. Therefore, the identification and analysis of TEs is important for understanding genome evolution.

TEs can be classified into retrotransposons and DNA transposons according to their mechanism of transposition (Kidwell and Lisch, 1997). Whereas DNA transposons cut and insert themselves into new genomic sites, retrotransposons are copied and inserted at new genomic locations using RNA intermediates. Retrotransposons can be classified into two subgroups according to the existence of long terminal repeats (LTRs) at their 5′ and 3′ ends. In addition, both LTR retrotransposons and non-LTR retrotransposons can be further classified into various orders, super-families, and families based on their structure, encoded genes, and phylogeny (Wicker *et al.*, 2007).

We previously developed two separate computational methods for the genome-wide identification of LTR and non-LTR retrotransposons, called MGEScan-LTR and MGEScan-non-LTR, respectively. MGEScan-LTR identifies all types of LTR retrotransposons,

i.e. young intact, old intact and solo LTR retrotransposons, using approximate string matching, protein domain analysis, and profile Hidden Markov Models (Rho *et al.*, 2007, 2010). MGEScan-nonLTR identifies non-LTR retrotransposons based on Gaussian Bayes classifiers and generalized hidden Markov models consisting of 12 super states that correspond to different clades or closely related clades (Rho and Tang, 2009).

MGEScan-LTR and MGEScan-non-LTR have been successfully used for identifying retrotransposons (Colbourne *et al.*, 2011). However, these programs were not designed to support either a web graphical user interface or parallel execution on clusters. Their computational time for eukaryotic genome analysis ranges from a few hours to several days. In addition, because they were developed as command-line tools that produce output files regardless of post-processing, it is not easy to visualize the results or perform further analysis.

In this article, we introduce MGEScan, a Galaxy-based system for identifying retrotransposons in genomes. MGEScan provides an integrated interface to run both computational LTR and non-LTR retrotransposon identification programs using the Galaxy workflow. This design facilitates input data preparation and the further analysis of results. Execution time for retrotransposon identification in genomes is reduced by 68% when MGEScan runs with concurrent and MPI executions. Moreover, because we provide MGEScan in various forms, users can use MGEScan in their preferred way.

2 System descriptions

In this section, we describe the design objectives of MGEScan: to reduce execution time, add versatility, and improve the usability of the MGEScan-non-LTR and MGEScan-LTR programs as well as to map these design objectives with system implementations.

First, we integrated the MGEScan-LTR and MGEScan-non-LTR into MGEScan with a user-friendly interface to facilitate concurrent execution and provide three options for running either program or both. Therefore, a user can execute them concurrently or individually on a local machine or clusters.

Second, we adapted MGEScan to the Galaxy framework to support powerful and flexible data preparation and further analysis of the results. A single input data is shared between the programs. The results of the programs are merged and exported to General Feature Format v.3 (GFF3) for further analyses or visualization by genome browsers (Fig. 1).

Third, to improve the performance (i.e. throughput), we replaced HMMER version 2.3.2 in MGEScan with version 3.1b1 (Finn *et al.*, 2015) and re-trained the profile Hidden Markov Models that are used to identify retrotransposons in a genome. HMMER 3.1b1 performs better and ensures improved sensitivity because of its new heuristic acceleration algorithm for sequence searches. In addition to the concurrent execution of MGEScan-LTR and MGEScan-non-LTR, if a user runs MGEScan on an MPI-capable machine, the overall process execution is faster than non-MPI execution.

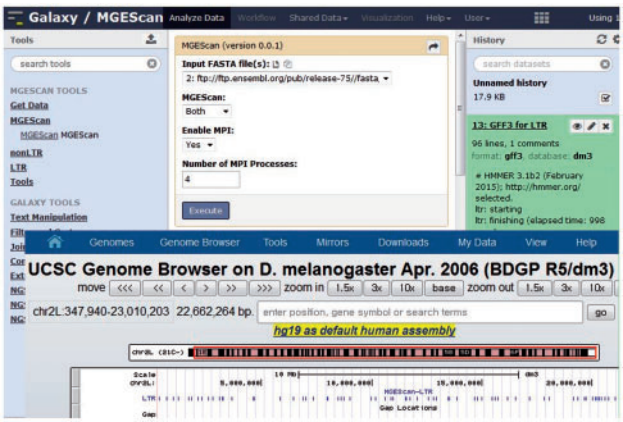

Fig. 1. Screenshots of the Galaxy-based MGEScan

Last, we provide four different distribution forms of the MGEScan program: the command-line program without the Galaxy framework, the Galaxy Toolshed, the Galaxy-based web server, and the MGEScan virtual clusters on the Amazon cloud. If a user is looking for a light-weight and simple program but is not familiar with the Galaxy system, the MGEScan command-line program can be downloaded and installed on a local machine without the Galaxy framework. In contrast, if a user needs to use both the graphical and command-line interfaces with full features but lacks the necessary skills or understanding of the MGEScan programs, the Galaxy-based MGEScan webserver or the MGEScan Galaxy Toolshed can be downloaded and installed. MGEScan programs are included in the webserver with helper tools to reduce the effort of using workflows and tools. We also provide a one-liner command to support the quick and simple installation of the required software and configuration for the command-line program or the Galaxy-based webserver. Finally, a user may want to run MGEScan in a cloud environment or in clusters. In this case, the user can use a virtual machine image of MGEScan with the Galaxy framework. Currently, we provide a virtual image of Galaxy-based MGEScan on Amazon EC2 (Amazon machine image ID: ami-10672b7a on the US East region). Thus, a virtual single or multiple servers for MGEScan can be easily deployed on Amazon Web Services.

3 Results and conclusion

To evaluate the improvement in throughput of MGEScan, we compared the execution time of MGEScan-LTR, MGEScan-non-LTR, MGEScan without MPI, and MGEScan with MPI using test genome datasets. The consistency of prediction results across different versions of MGEScan was also evaluated (see Supplementary Materials). We described our experiments and results in detail in the Supplementary Materials.

When MGEScan ran without MPI, the execution time was, on average, 1.43× faster than MGEScan-non-LTR and MGEScan-LTR because of the concurrent execution support and version upgrade of the HMMER in MGEScan. In addition, MGEScan with MPI (four vCPUs) was 3.26× faster than MGEScan-non-LTR and MGEScan-LTR. Also, the performance comparison of MGEScan over a different number of MPI processes shows that the MGEScan programs for LTR and non-LTR provide better performance with additional MPI processes. These results show that MGEScan improves the throughput of retrotransposons identification in genomes while the prediction performance remains consistent across the different versions of program and different MPI settings.

In this article, we provide the integrated and improved MGEScan for identifying retrotransposons in eukaryotic genomes. MGEScan, with its improved usability, throughput and versatility, can be easily used to identify all types of retrotransposons. Future work will include developing a module for identifying DNA transposons in MGEScan and incorporating newly discovered families of TEs to fully support the identification of all kinds of TEs in eukaryotic genomes.

Funding

This work was supported by National Research Foundation of Korea grants funded by the Ministry of Education (NRF-2015R1D1A1A01059557),

by the Korean government (MSIP) (KW-2014PPD0053 and NRF-2015R1C1A1A01054305), by the National Science Foundation grant (DBI-1262588) and by Marine Biotechnology Program (PJT200620) of Ministry of Oceans and Fisheries.

Conflict of Interest: none declared.

References

Colbourne,J.K. *et al*. (2011) The ecoresponsive genome of Daphnia pulex. *Science*, **331**, 555–561.

Finn,R.D. *et al*. (2015) HMMER web server: 2015 update. *Nucleic Acids Res*., **43**, w30–w38.

Kidwell,M.G., and Lisch,D. (1997) Transposable elements are sources of variation in animals and plants. *Proc. Natl. Acad. Sci. USA*, **94**, 7704–7711.

Lander,E.S. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Rho,M. *et al*. (2007) De novo identification of LTR retrotransposons in eukaryotic genomes. *BMC Genomics*, **8**, 90.

Rho,M. *et al*. (2010) LTR retroelements in the genome of Daphnia pulex. *BMC Genomics*, **11**, 425.

Rho,M. and Tang,H. (2009) MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes. *Nucleic Acids Res*., **37**, e143.

Wicker,T. *et al*. (2007) A unified classification system for eukaryotic transposable elements. *Nature Rev. Genet*., **8**, 973–982.