BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 6 2013, pages 792–793
doi:10.1093/bioinformatics/btt054

Genome analysis

Advance Access publication February 1, 2013

RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads

Petr Novák, Pavel Neumann, Jiří Pech, Jaroslav Steinhaisl and Jiří Macas*

Institute of Plant Molecular Biology, Biology Centre ASCR, Branišovská 31, České Budějovice, CZ-37005, Czech Republic

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Repetitive DNA makes up large portions of plant and animal nuclear genomes, yet it remains the least-characterized genome component in most species studied so far. Although the recent availability of high-throughput sequencing data provides necessary resources for in-depth investigation of genomic repeats, its utility is hampered by the lack of specialized bioinformatics tools and appropriate computational resources that would enable large-scale repeat analysis to be run by biologically oriented researchers.

Results: Here we present RepeatExplorer, a collection of software tools for characterization of repetitive elements, which is accessible via web interface. A key component of the server is the computational pipeline using a graph-based sequence clustering algorithm to facilitate *de novo* repeat identification without the need for reference databases of known elements. Because the algorithm uses short sequences randomly sampled from the genome as input, it is ideal for analyzing next-generation sequence reads. Additional tools are provided to aid in classification of identified repeats, investigate phylogenetic relationships of retroelements and perform comparative analysis of repeat composition between multiple species. The server allows to analyze several million sequence reads, which typically results in identification of most high and medium copy repeats in higher plant genomes.

Implementation and availability: RepeatExplorer was implemented within the Galaxy environment and set up on a public server at http://repeatexplorer.umbr.cas.cz/. Source code and instructions for local installation are available at http://w3lamc.umbr.cas.cz/lamc/resources.php.

Contact: macas@umbr.cas.cz

Received on November 23, 2012; revised on January 7, 2013; accepted on January 28, 2013

1 INTRODUCTION

In spite of the recent progress in genome sequencing technologies, accurate quantification and sequence characterization of repetitive DNA in complex plant and animal genomes remains a challenging task. Most of the existing bioinformatics tools either require completed genome assemblies for repeat identification or rely on similarity searches to databases of known repetitive elements. However, the number of fully assembled genomes is still small, and many exhibit poor assembly quality across repetitive regions where highly repetitive elements are mostly omitted. On the other hand, database scanning enables repeat identification in any type of sequence data, but it in principle does not allow detection of novel repetitive elements that lack similarity to the database. Consequently, both of these methods work well for extensively investigated groups of model species but their application to a broader range of diverse taxa is limited.

A principally different approach for global repeat analysis that allows for *de novo* repeat identification and is suitable for utilizing unassembled reads was developed by Novak *et al.* (2010). It is based on finding and quantifying similarities between individual sequence reads typically obtained by next-generation sequencing of randomly sheared genomic DNA. These similarities are used to construct a graph in which the vertices correspond to sequence reads, overlapping reads are connected with edges and their similarity scores are expressed as edge weights. Graph topology is then analyzed to identify and separate clusters of frequently connected reads that represent individual families of repetitive elements. This strategy proved to be efficient in analyzing repeat composition of both plant (Macas *et al.*, 2011; Novak *et al.*, 2010) and animal (Pagan *et al.*, 2012) genomes.

Here we report on the development of a computational pipeline implementing an improved version of graph-based clustering and a number of additional tools for downstream analysis of the identified repeats. All tools were adapted to run under the online data analysis platform Galaxy (Goecks *et al.*, 2010), which provides a user-friendly web interface and facilitates easy execution, documentation and sharing of analysis protocols and results.

2 DESCRIPTION OF TOOLS

A schematic representation of the RepeatExplorer components is depicted in Figure 1. Sequence reads are uploaded and pre-processed using tools included in the Galaxy platform. Optionally, RepeatExplorer tools can be used to manipulate read names (e.g. adding specific codes for different datasets that are later merged) and to perform random sampling of single or paired-end reads to reduce the size of analyzed data.

The main component of RepeatExplorer is the clustering pipeline, which performs all-to-all similarity comparisons of sequence reads followed by their graph-based clustering to identify groups of reads derived from repetitive elements. These groups are

*To whom correspondence should be addressed.

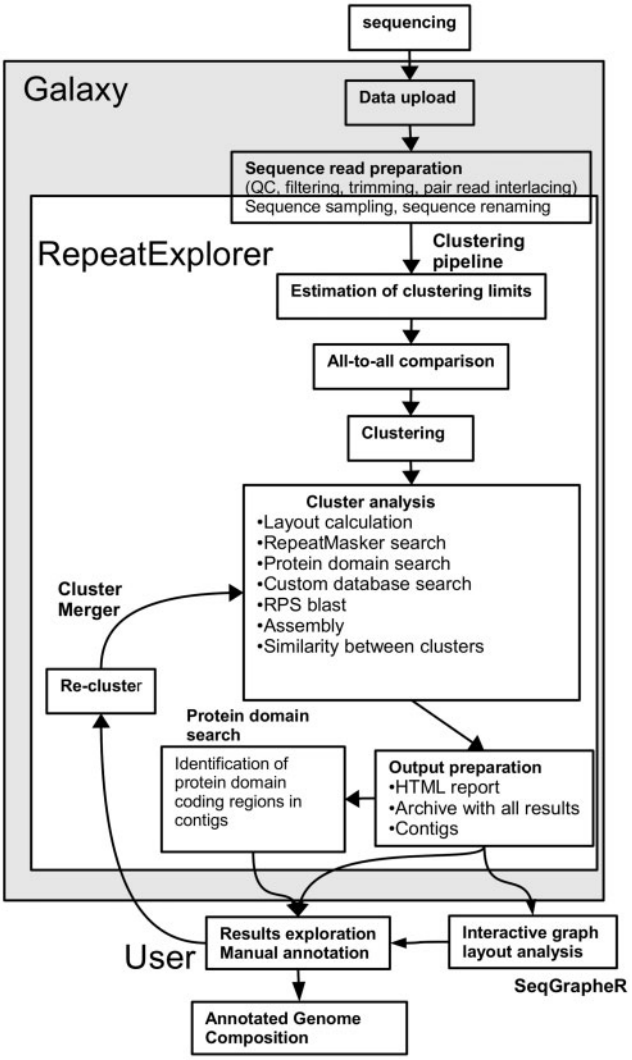

Fig. 1. RepeatExplorer components and analysis workflow

further characterized with respect to their proportion in the genome and similarity to known repetitive elements and conserved protein domains. Cluster graph layouts are calculated, which in combination with the results of similarity searches, can be used for manual classification and annotation of repeats in individual clusters (Novak *et al*., 2010). As the clustering procedure tends to split large repeats into several clusters, there is a separate re-clustering tool for user-aided merging of clusters. This can be performed taking into account read similarities between clusters and information about paired-end reads, if available.

A set of tools is provided for identification, extraction and analysis of sequence regions corresponding to conserved domains of retrotransposon proteins. The domains are identified in contigs assembled from reads in individual clusters based on their similarity to a representative set of reference sequences. The domains can be further submitted to tools that facilitate multiple sequence alignments and calculation of phylogenetic trees.

The tools can be combined within the Galaxy environment to generate analysis workflows aimed at specific tasks. Example workflows are provided and further explained in the RepeatExplorer manual (available online), covering the topics of global repeat characterization in a single species, comparative analysis of repeat composition between multiple species and phylogenetic analysis of Long terminal repeat (LTR)-retrotransposon sequences. They can be used to better understand principles of analysis or as templates that can be modified and re-used for user-provided data.

3 PERFORMANCE

Several million reads can be analyzed in one run on the current RepeatExplorer server configuration, which allocates up to 16GB RAM for individual processes. The actual number of reads that can be processed depends on the number of similarity hits they produce because all read overlaps must be loaded into the computer memory during the graph-based clustering step (Blondel *et al*., 2008). About 3.4×10^8 similarity hits is the current limit of the server, which corresponds, for example, to the number of hits generated by about 2 million Illumina reads (100 nt) from the highly repeated genome of pea (*Pisum sativum*). Proportionally larger numbers of reads can be analyzed for species with lower repeat content like soybean (Novak *et al*., 2010), where the limiting number of hits is produced by about 3 million reads. It should be noted that the proportion of satellite DNA is often the main factor affecting total number of similarity hits because its repeated units have highly conserved sequences and can occur in the genome in millions of copies. Should the current RepeatExplorer server configuration be limiting for users requiring analysis of larger datasets, they are welcome to set up their own instance of the server using the provided source code.

ACKNOWLEDGEMENTS

We thank CESNET for providing data storage facility (project no. CZ.1.05/3.2.00/08.0142) and Jasper E. Manning for his help with manuscript preparation.

Funding: This work was supported by grants P501/12/G090 from the Czech Science Foundation, OC10037 from Ministry of Education, Youth and Sports and RVO:60077344 from the Academy of Sciences of the Czech Republic.

Conflict of Interest: none declared.

REFERENCES

Blondel,V.D. *et al.* (2008) Fast unfolding of communities in large networks. *J. Stat. Mech.*, **10**, P10008.

Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Macas,J. *et al.* (2011) Next generation sequencing-based analysis of repetitive DNA in the model dioecious plant *Silene latifolia*. *PLoS One*, **6**, e27335.

Novák,P. *et al.* (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, **11**, 378.

Pagan,H.J.T. *et al.* (2012) Survey sequencing reveals elevated DNA transposon activity, novel elements, and variation in repetitive landscapes among bats. *Genome Biol. Evol.*, **4**, 575–585.