Bioinformatics, 31(11), 2015, 1827–1829
doi: 10.1093/bioinformatics/btv059
Advance Access Publication Date: 1 February 2015
Applications Note

OXFORD

Genome analysis

Transposome: a toolkit for annotation of transposable element families from unassembled sequence reads

S. Evan Staton[1,†,*] and John M. Burke[2]

[1]Department of Genetics and [2]Department of Plant Biology, University of Georgia, Athens, GA 30602, USA

*To whom correspondence should be addressed.
[†]Present address: The Biodiversity Research Centre and Department of Botany, 3529-6270 University Blvd, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada

Associate Editor: Alfonso Valencia

Received on September 11, 2014; revised on January 5, 2015; accepted on June 26, 2015

Abstract

Motivation: Transposable elements (TEs) can be found in virtually all eukaryotic genomes and have the potential to produce evolutionary novelty. Despite the broad taxonomic distribution of TEs, the evolutionary history of these sequences is largely unknown for many taxa due to a lack of genomic resources and identification methods. Given that most TE annotation methods are designed to work on genome assemblies, we sought to develop a method to provide a fine-grained classification of TEs from DNA sequence reads. Here, we present a toolkit for the efficient annotation of TE families from low-coverage whole-genome shotgun (WGS) data, enabling the rapid identification of TEs in a large number of taxa. We compared our software, Transposome, with other approaches for annotating repeats from WGS data, and we show that it offers significant improvements in run time and produces more precise estimates of genomic repeat abundance. Transposome may also be used as a general toolkit for working with Next Generation Sequencing (NGS) data, and for constructing custom genome analysis pipelines.
Availability and implementation: The source code for Transposome is freely available (http://sestaton.github.io/Transposome), implemented in Perl and is supported on Linux.
Contact: statonse@biodiversity.ubc.ca
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genomic repeat annotation is a challenging task, in part because there are dozens of tools available and not all have not been analyzed in terms of performance or accuracy (Leret, 2010). Current approaches for identifying transposable elements (TEs) involve using structural and similarity-based approaches with a genome assembly (Ellinghaus *et al.*, 2008; Steinbiss *et al.*, 2009; Xu and Wang, 2007), mathematical or k-mer based methods from genome assemblies or random sequence reads (Bao and Eddy, 2002; Kurtz et al., 2008), signature-based methods with annotated TEs (Wheeler *et al.*, 2013), and cluster-based approaches from unassembled sequence reads (Novak *et al.*, 2010, 2013). Without question, the most accurate method for identifying TEs would be through a combination of the above methods (Bergman and Quesneville 2007; Leret, 2010; Saha *et al.*, 2008). One caveat with the aforementioned approaches is that most require a reference genome (i.e. assembled genome) as input. It is not, however, practical to generate a reference genome for every species of interest given that genome sequencing costs remain prohibitively high for non-model species, and genome assembly algorithms are not able to easily resolve large and complex genomes (Alkan *et al.*, 2011), such as those of many plant species. Therefore, the ideal solution to improving repeat annotation from understudied

species would be to leverage high-throughput DNA sequencing technologies, enabling phylogenetic descriptions of TE properties in many species simultaneously.

Here we present Transposome, a toolkit that is able to accurately estimate the genomic abundance of TE families from very low coverage whole-genome shotgun (WGS) data, enabling the rapid detection of changes in genome composition across many species. Transposome borrows much of its design from RepeatExplorer (Novak *et al.*, 2013), a recently published tool for detecting repeats from WGS data. RepeatExplorer provides an easy-to-use web interface for users; however, we found this tool to be computationally inefficient and the design is not modular, which makes analyses of multiple species and designing custom pipelines impractical. Transposome builds on the basic approach of RepeatExplorer, which is finding similarity in the genome through a graph-based analysis of similarity between WGS reads. We extend this approach in Transposome to provide a more fine-scale level of annotation of genomic repeats, and we provide direct estimates of genomic repeat abundance in biological terms.

We demonstrate the utility of Transposome by analyzing a set of WGS reads from the well-studied maize (*Zea mays* L.) genome, and we compare our results to published estimates of genome composition for this species. Our findings indicate that Tranposome is both more accurate at quantifying genomic repeats, and more efficient than other approaches (Supplementary Table S1).

2 Methods

The approach we implemented consists of two steps. First, we developed a highly parallel all versus all sequence comparison procedure to find shared similarity within the genome. This step makes use of mgblast (Pertea *et al.*, 2003), a modified version of megablast that is memory efficient. Second, we use a graph-based clustering method that is able to handle very large datasets (Blondel *et al.*, 2008) efficiently and makes use of edge weight information (which corresponds to similar sequence pairs in this application). This algorithm is very fast, graphs grow like $O(n)$, but one caveat is that it over-refines clusters leading to a separation of defined groups (Fortunato, 2010). We modified this algorithm by using paired-end sequence information to find union in the graph constructed during the clustering processes, and this allows us to circumvent the issue of over-refinement of clusters by using biological information.

Every sequence clustered by Transposome represents a repetitive element within the genome, and these sequences are compared to a reference library of repeat sequences using blastn from the BLAST+ package (Camacho *et al.*, 2009), along with singleton sequences that are not clustered. Perhaps the most useful feature of Transposome is that results are translated directly into estimates of genomic composition of each repeat type, and the full taxonomic lineage of each repeat type is listed to the family level.

The novelty of Transposome is that it offers a programmatic interface, via a Perl API, for manipulating FASTA or FASTQ data, finding similarity in a read set and annotating genomic repeats. Thus, users can construct custom analysis pipelines, or repeat parts of the analysis with varied parameters without unnecessarily repeating an entire pipeline of operations. In addition, Transposome can be used for the general task of manipulating NGS data and BLAST reports. Examples of usage and scripts are provided on the Transposome API Wiki page (https://github.com/sestaton/Transposome/wiki/API-Tutorial) and the separate transposome-scripts repository (https://github.com/sestaton/transposome-scripts).

3 Results

How effectively are we sampling the genomic diversity of repetitive elements? To address this question, we sampled sequence data from maize at varying levels of genome coverage and performed an analysis with Transposome using default parameters on each read set (see Supplementary Information). As we know what fraction of the maize genome is comprised of TEs, we were also able to assess the total fraction of diversity being sampled at varying levels of genome coverage (i.e. the total percent of the genome). The maize genome is >85% TEs (Schnable *et al.*, 2009), with approximately 70% of the genome being composed of just 20 TE families (Baucom *et al.*, 2009). With just 100,000 randomly sampled paired-end sequences (see Supplementary Information), Transposome was able to correctly identify 17 of the top 20 families, with the top 11 families being correctly identified as in Baucom *et al.* (2009), albeit with minor reordering (Supplementary Fig. S1). We note that Transposome also generates estimates of genomic repeat abundance faster than other annotation procedures (Supplementary Table S1).

How many sequence reads are required to capture the majority of the TEs by genome coverage? It is clear that Transposome is able to accurately predict the abundance of TEs and the number of TE families by rank that account for that genome abundance (Supplementary Fig. S2). This result indicates that it is not necessary to have high genome coverage sequence data (i.e. >1%) in order to predict the genomic abundance of TE families, though increasing coverage will yield more accurate estimates of TE diversity and family-level abundance (Supplementary Figs 2 and 3).

4 Conclusions

Transposome is a toolkit for rapidly determining the genomic abundance of repeats from low coverage WGS data, requiring less than 1% genome coverage to obtain reliable estimates of TE family abundance in just a few minutes. We have further demonstrated that a small increase in genome coverage (e.g. 3–5%) allows Transposome to produce even more accurate estimates of family-level genomic abundance and TE diversity in the genome.

Acknowledgement

We thank Petr Novak for discussions about RepeatExplorer and advice on running this program.

Funding

National Science Foundation (DBI-0820451 to J.M.B.); USDA National Institute of Food and Agriculture (2008-35300-19263 to J.M.B.).

Conflict of Interest: none declared.

References

Alkan,C. *et al.* (2011) Limitations of next-generation genome sequence assembly. *Nat. Methods*, **8**, 61–65.

Bao,Z., and Eddy,S.R. (2002) Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res.*, **12**, 1269–1276.

Baucom,R.S. *et al.* (2009) Exceptional diversity, non-random distribution, and rapid evolution or retroelements in the B73 maize genome. *PLoS Genet.*, **5**, 1–13.

Bergman,C.M. and Quesneville,H. (2007) Discovering and detecting transposable elements in genome sequences. *Brief Bioinform.*, **8**, 382–392.

Blondel,V.D. *et al.* (2008) Fast unfolding of communities in large networks. *J. Stat. Mech.*, **10**, P1000.

Camacho,C. *et al*. (2009) BLAST+: architecture and applications. *BMC Bioinform.,* **10**, 421.

Ellinghaus,D. *et al*. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinform.*, **9**, 18.

Fortunato,S. (2010) Community detection in graphs. *Physics Rep*., **486**, 75–174.

Kurtz,S. *et al*. (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics,* **9**, 517.

Leret,E. (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity*, **104**, 520–533.

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

Novak,P. *et al*. (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**, 792–793.

Pertea,G. *et al.* (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.

Saha,S. *et al*. (2008) Empirical comparison of *ab initio* repeat finding programs. *Nucleic Acids Res.*, **36**, 2284–2294.

Schnable,P.S. *et al*. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

Steinbiss,S. *et al*. (2009) Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res*., **37**, 7002–7013.

Wheeler,T.J. *et al*. (2013) Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.*, **41**, D70–D82.

Xu,Z. and Wang,H. (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res*., **35**, W265–W268.