BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 18 2014, pages 2656–2658
doi:10.1093/bioinformatics/btu365

Genome analysis

Advance Access publication June 3, 2014

Tedna: a transposable element *de novo* assembler

Matthias Zytnicki[1,*], Eduard Akhunov[2] and Hadi Quesneville[1]

[1]INRA, URGI, Plant Breeding and Biology, Versailles 78026, France and [2]Department of Plant Pathology, Kansas State University, Manhattan, KS 66506, USA

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Recent technological advances are allowing many laboratories to sequence their research organisms. Available *de novo* assemblers leave repetitive portions of the genome poorly assembled. Some genomes contain high proportions of transposable elements, and transposable elements appear to be a major force behind diversity and adaptation. Few *de novo* assemblers for transposable elements exist, and most have either been designed for small genomes or 454 reads.

Results: In this article, we present a new transposable element *de novo* assembler, Tedna, which assembles a set of transposable elements directly from the reads. Tedna uses Illumina paired-end reads, the most widely used sequencing technology for *de novo* assembly, and forms full-length transposable elements.

Availability and implementation: Tedna is available at http://urgi.versailles.inra.fr/Tools/Tedna, under the GPLv3 license. It is written in C++11 and only requires the Sparsehash Package, freely available under the New BSD License. Tedna can be used on standard computers with limited RAM resources, although it may also use large memory for better results. Most of the code is parallelized and thus ready for large infrastructures.

Contact: matthias.zytnicki@toulouse.inra.fr

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

Received on December 9, 2013; revised on April 28, 2014; accepted on May 23, 2014

1 INTRODUCTION

Most laboratories can now afford sequencing a genome, and several *de novo* whole genome assemblers, such as Velvet (Zerbino and Birney, 2008), are available to assemble even larger genomes. However, these assemblers usually do not assemble the transposable elements: current algorithms cannot correctly assemble highly repeated sequences. On the other hand, transposable elements can comprise >90% of a genome, and their role in the evolution of the host genome has been often underlined (Fedoroff, 2012).

Although assembling the *copies* (i.e. the traces of transposable elements, scattered along the genome, often strongly mutated) is usually impossible, assembling the *transposable elements* (a model of the first element that invaded the genome, reconstructed from the observable copies; Flutre *et al.*, 2011b) is a simpler task. Several *de novo* transposable elements assemblers have been presented so far, but most have been designed for small genomes (Pevzner *et al.*, 2004), or assume that the 454 technology have been used (Li *et al.*, 2005; Novák *et al.*, 2010). Only RepeatExplorer (Novák *et al.*, 2013) currently exploits Illumina reads. Here, we present a new tool that reads Illumina paired-end reads, arguably the most widely used sequencing technology for *de novo* assembly, and provides a list of repeated elements.

A transposable element *de novo* assembler is somewhat different from a genome assembler: it assembles sequences from multiple copies, which have evolved through time. A transposable elements assembler should thus correctly handle polymorphism, including long insertions and deletions. Basically, every copy gives a hint about what the transposable element should be, but only the comparison with other copies—and the construction of a consensus—may help building the transposable element. As such, de Bruijn graphs, used by many assemblers, seem well-fitted for this purpose: a k-mer of a copy may be part of the consensus transposable element, whereas other parts of the copy may not. Transposable elements may thus be assembled as a set of highly repeated k-mers. Tedna is the first tool that uses a de Bruijn graph for transposable element assembly. Implementation details are given in Supplementary Material.

2 RESULTS AND CONCLUSION

We compared Tedna with the transposable element assembler RepeatExplorer, and the two other widely used assemblers: Velvet, for genome assembly, and Oases, for transcriptome assembly (Schulz *et al.*, 2012). Detailed information on the benchmark is available in Supplementary Material.

We first benchmarked the tools on the 5143-bp-long *copia* element, present in *Drosophila melanogaster*. We extracted all the copies annotated as *copia* by the REPET pipeline (Flutre *et al.*, 2011a), cut them into reads and mixed them with random reads. We thus produced 100 000 paired-ends reads, of size 2×100. *Copia* is a long terminal repeat (LTR) retrotransposon and thus has long LTRs. With the best parameters of Velvet, we had a 4709-bp-long element, where the low complexity region, in the center of the element, is poorly assembled. Moreover, the predicted element is the concatenation of the 3′ end of the element, one LTR and the 5′ end. The longest element of Oases is 5857-bp long, longer than the actual element because Oases duplicates both LTRs. With other parameters, Oases predicts a 4912-bp-long element, similar to the element predicted by Velvet. Tedna correctly predicts an element with 99% identity when compared with the known element. The predicted element is somewhat shorter (5049 versus 5143 bp) because the predicted

*To whom correspondence should be addressed.

Table 1. Comparison of the tools

Dataset	Tool	Number of sequences	Sensitivity (%)	Specificity (%)	Average maximum size (%)
Wheat	RepeatExplorer	982	35	78	6
	Velvet	836	2	6	1
	Oases	6505	33	37	13
	Tedna	1365	38	66	11
Arabidopsis thaliana	RepeatExplorer	160	3	14	2
	Velvet	67 615	73	2	42
	Oases	1963	41	38	30
	Tedna	1263	24	26	17

LTRs are slightly too short, and the low complexity region is not accurately assembled. The three tools needed < 1 min to produce the assembly. RepeatExplorer produced an internal error, probably because it expects more reads. This shows that a transposable element can be reconstructed from its copies with a de Bruijn graph on the most frequent *k*-mers.

We then tested Tedna by assembling sequence data generated for the wheat genome, 90% of which is composed of repetitive elements. We used the unassembled reads (size 2×100 bp) produced by the International Wheat Genome Sequencing Consortium for wheat chromosome arm 3AL (unpublished data) and a manually curated library of 335 wheat transposable elements (Josquin Daron, submitted for publication). We finally compared Tedna on an *Arabidopsis thaliana* resequencing project, available from the Sequence Read Archive under code SRR616966, which contains 2×100 paired-end reads and an insert size of 500 bp. We used the *A.thaliana* RepBase data (Jurka *et al.*, 2005) as reference, which contains 390 transposable elements.

For the two latter datasets, we gave the number of putative transposable elements given by each tool, as well as their sensitivity and specificity in Table 1. For each reference transposable element, we computed the size of the longest predicted fragment and expressed it as a ratio of the reference transposable element size (100% would be a predicted full-length element). The average ratio size is given in the last column. The wheat dataset shows that Tedna has the best sensitivity and a good specificity (although not as good as RepeatExplorer). When ranked by size of fragments, Tedna is the second best. Oases performs well because it also contains dedicated algorithms for merging contigs into full-length transcripts, and it handles read coverage better than Velvet (which expects uniform coverage). The *A.thaliana* dataset is clearly favorable to Oases, which gives almost everywhere the best results. The only exception is the fragment size, where Velvet performs better. This is a usual trade-off between specificity, which is low for Velvet, and accuracy. These results suggest that Tedna performs better when used on genomes with high transposable element density, which is observed in most of higher eukaryotes.

We then detailed the results given by Tedna for each major transposable element class of *A thaliana* (see Supplementary Table). Results vary greatly, and there is no clear reason as to why the DNA transposons are better assembled than other elements. Up to now, Tedna only assembles repeated sequences and cannot discriminate transposable elements. In the *A.thaliana* dataset, we located the sequences assembled by Tedna that did not match the RepBase sequences. Among them, we had 1618 matches in genes. Most map to protein of unknown function; some of them could be misannotated transposable elements. The other fragments map to known duplicated genes (Agamous-like proteins, cellulose synthase, expansins, etc.). A total of 305 fragments map to inserted copies and have been thus missed in our classification protocol, most likely because they significantly diverged from the consensus. A total of 26 fragments mapped the *gypsy* element, 19 MuDR and 16 *copia*.

In a future version, we would like to provide an annotation of the output of Tedna, as RepeatExplorer does, that would classify transposable elements and possibly discriminate genes or other repeated elements.

ACKNOWLEDGEMENTS

The authors wish to thank the URGI laboratory, and especially F. Maumus for his helpful comments, the International Wheat Genome Sequencing Consortium for providing unpublished sequencing data, J. Daron for sharing his manually curated transposable element library and R. Chikhi for sharing his knowledge on assembly.

Conflict of Interest: none declared.

REFERENCES

Fedoroff,N. (2012) Transposable elements, epigenetics, and genome evolution. *Science*, **338**, 758–767.

Flutre,T. *et al.* (2011a) Considering transposable element diversification in *de novo* annotation approaches. *PLoS One*, **6**, e16526.

Flutre,T. *et al.* (2011b) In search of lost trajectories: recovering the diversification of transposable elements. *Mob. Genet. Elements*, **1**, 151–154.

Jurka,J. *et al.* (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, **110**, 462–467.

Li,R. *et al.* (2005) ReAS: recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Comput. Biol.*, **1**, e43.

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

Novák,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.

Pevzner,P.A. *et al.* (2004) De novo repeat classification and fragment assembly. *Genome Res.*, **14**, 1786–1796.

Schulz,M.H. *et al.* (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, **28**, 1086–1092.

Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.