BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 19 2010, pages 2452–2454
doi:10.1093/bioinformatics/btq454

Genome analysis

Advance Access publication August 9, 2010

MASiVE: Mapping and Analysis of SireVirus Elements in plant genome sequences

N. Darzentas[1,*,†], A. Bousios[1,†], V. Apostolidou[1] and A. S. Tsaftaris[1,2]

[1]Institute of Agrobiotechnology, Centre for Research and Technology Hellas, Thessaloniki 57001 and [2]Department of Genetics and Plant Breeding, Aristotle University of Thessaloniki, Thessaloniki 54006, Greece

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: We present MASiVE, an expertly built tool for the large-scale, yet sensitive and highly accurate, discovery, preliminary analysis and insertion age estimation of intact Sirevirus LTR-retrotransposons in plant genomic sequences. Validation was based on the recently available and annotated large maize chromosome one. Results show a considerable improvement in the annotation of Sireviruses, and support our approach as an important addition to the bioinformatics toolbox of plant biologists.

Availability: PERL source code and essential files are available online at http://bat.ina.certh.gr/tools/masive/. The freely available Vmatch, LTRharvest, Wise2, and MAFFT algorithms are required.

Contact: ndarz@certh.gr

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

Received on June 9, 2010; revised on July 20, 2010; accepted on August 4, 2010

1 INTRODUCTION

The availability of large and complex plant genome sequences has necessitated the development of algorithms that can facilitate high-throughput discovery and annotation of transposable elements, especially long terminal repeat retrotransposons (LTR-RTN) that comprise the majority of these genomes (Schnable *et al.*, 2009). Tools like LTR_STRUC (McCarthy and McDonald, 2003), LTR_FINDER (Xu and Wang, 2007) and LTRharvest (Ellinghaus *et al.*, 2008) are available; however, they are all based on general structural characteristics of LTR-RTNs, which may lead to misannotations or incomplete identification of the LTR-RTN set. The recent discovery that the abundant and plant-specific Sireviruses contain highly conserved motifs in key domains of their genome (Bousios *et al.*, 2010) provided the opportunity to develop MASiVE (Mapping and Analysis of Sirevirus Elements), presented herein, a tool able to identify, with high precision and sensitivity, intact Sirevirus elements in plant genomic sequences.

All LTR-RTNs contain polypurine tract (PPT) and primer binding site (PBS) domains. However, their high sequence and length plasticities (Bousios *et al.*, 2010) render them inappropriate for the identification of a specific retrotransposon class. In contrast, Sireviruses have multiple identical PPTs at the upstream border of the 3′LTR and a highly conserved 5′LTR/internal domain junction that includes part of the PBS. These, along with several filters applied sequentially, are used by MASiVE to target full-length Sirevirus elements, while excluding false positives like other retrotransposons or non-retroelement-derived sequences resembling the structure of LTR-RTNs. MASiVE has been calibrated and validated in the largest chromosome of the now available, and expertly annotated, *Zea mays* (maize) B73 genome (Schnable *et al.*, 2009), chromosome one.

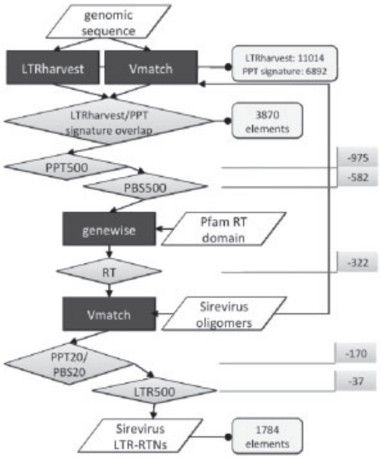

Fig. 1. The MASiVE algorithm pipeline and the effect of filtering on the numbers of Sireviruses finally reported as intact. White parallelograms represent data, dark rectangles tools and grey diamonds filters.

2 METHODS

2.1 LTRharvest and the multiple PPT signature filter

MASiVE (Fig. 1) is written in Perl, and makes use of external algorithms: LTRharvest, Vmatch (www.vmatch.de), Wise2 (Birney *et al.*, 2004) and MAFFT (Katoh and Toh, 2008).

Initially, LTRharvest is deployed to scan the genomic sequence for the presence of LTR-RTNs (see Supplementary Table 1 for LTRharvest parameters). MASiVE proceeds by matching the Sirevirus-specific PPT octamer (AGGGGGAG) and 12-mer (AGGGGGAGATTG) on the genomic sequence using the Vmatch algorithm. The majority of PPTs cluster within 500 bp of the 3′LTR (Supplementary Fig. 1), while the terminal octamer is always followed by the conserved ATTG attachment site. Therefore, we describe the multiple PPT signature as the requirement to find at least 2 and a maximum of 15 upstream PPTs and the terminal PPT/attachment site within 1000 bp (Supplementary Table 2). Since LTRharvest does not search for PPT-like structures, we can instead use the highly conserved Sirevirus

[*]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.

multiple PPT signature and retain, as putative full-length Sireviruses, only LTRharvest-predicted elements that overlap with a multiple PPT signature, in our case 3870 candidates.

2.2 PPT500 and PBS500 filters

2.2.1 PPT500 Misannotations of LTR-RTNs may either include random sequences, or a fusion of different retroelement fragments in the complex-nested retrotransposon landscape over long chromosomal regions that is characteristic of plants with large genomes (Baucom *et al*., 2009). As the 12-mer of each multiple PPT signature precisely defines the Sirevirus internal domain/3′LTR junction, if the distance between the 12-mer and the overlapping LTRharvest-predicted inner side of the 3′LTR is too long, then it is highly likely that LTRharvest reported an element-artifact. In 3870 putative full-length Sireviruses in maize chromosome one, in 85%, 75% and 53% of the cases the distance between the 12-mer and the 3′LTR was smaller than 1000, 500 and 100 bp, respectively. Thus, we decided to filter out elements with a distance longer than 500 bp (PPT500 filter).

2.2.2 PBS500 The second strongest sequence pattern of the Sirevirus genome is the 5′LTR/internal domain junction. The conserved 16-mer consists of the CAATTGG attachment site and the TATCAGAGC motif that complements the 3′ end of the $^{\mathrm{met}}$tRNA. Likewise the multiple PPT signature, the PBS site can be used to reject false positives. In 2895 putative intact Sireviruses that had passed the PPT500 filter, the distance between the 16-mer (with up to three mismatches) and the 3′ side of the 5′LTR in 95%, 84% and 75% of the elements was shorter than 1000, 500 and 100 bp, respectively. Thus, we set the threshold at 500 bp (PBS500 filter).

2.3 Pfam reverse transcriptase filter

Although it is very likely that elements passing through the PPT500 and PBS500 filters represent intact Sireviruses, we also searched for hits of the Pfam reverse transcriptase domain (PF07727) for Ty1/*copia* retrotransposons in the internal region between the two LTRs of the putative Sireviruses.

2.4 PPT20 and PBS20 filters

These two filters require the full length elements to possess: (i) a 20-mer that includes the terminal two nucleotides of the 12-mer PPT in the 2.5-kbp-long upstream region of the PBS, and (ii) a 20-mer that includes the first two nucleotides of the PBS in the 2.5-kbp-long downstream region of the PPT (with up to three mismatches for either 20-mer). Importantly, this step also allows for the determination of the exact LTR borders of the Sirevirus element.

2.5 LTR filters

Finally, all Sireviruses with (i) LTR lengths of at least 500 bp and (ii) difference between their 3′ and 5′ LTR lengths <500 bp, are placed in the final set of intact Sireviruses.

3 RESULTS

3.1 Sirevirus specificity—the multiple PPT signature

Analysis of the 12-mers and octamers reported by Vmatch revealed that a large proportion belonged to multiple PPT signatures, which together with their sequence integrity, denotes the conservation of this domain and its capacity to be used by MASiVE (Supplementary Table 3). We also searched for the multiple PPT signatures in non-plant genomes (human, fly and yeast) with negative results (data not shown). Moreover, to verify that the multiple PPT signature is a Sirevirus-specific motif, we isolated 200 bp downstream of each signature, predicted to be the first section of the 3′LTR, and compared them to the respective 200 bp fragments of the maize Sireviruses Opie, Ji and Giepum, which comprise nearly 20% of the maize genome, and of the other annotated maize Ty1/*copia* retrotransposons (Baucom *et al*., 2009). More than 99% of the 6892 sequences clustered with the maize Sireviruses (data not shown), demonstrating that the vast majority of sequences after the signature are indeed the 3′LTR of Sireviruses.

3.2 Effects of filtering and insertion age estimation

The stepwise application of the filters rejected 2086 elements in total that did not satisfy the criteria for being categorized as full-length Sireviruses, resulting in a final set of 1784 high-quality intact Sireviruses (Fig. 1). Analysis along the MASiVE pipeline of the full lengths and the lengths of LTRs of the different Sirevirus populations showed a gradual convergence around the expected averages (Supplementary Fig. 2) (Baucom *et al*., 2009). Finally, following pairwise alignment of LTRs, an optional capability of MASiVE, we estimated that Sireviruses in chromosome one of maize have an average insertion age of 0.92 mya (million years ago). Apparently, an amplification burst occurred ∼0.1–0.4 mya (Supplemenatry Fig. 3).

3.3 Validation

We studied the overlap between our set of full-length Sirevirus elements in chromosome one of maize and that recently reported by the Maize Transposable Element Consortium (MTEC) (Schnable *et al*., 2009). The results (Supplementary Fig. 4) show that MASiVE not only recovered 76% of the MTEC set (951/1244), but also added 50% more intact elements to the pool (625). In fact, almost 90% of the MTEC Sireviruses that MASiVE failed to match (293 false negatives) are actually reported by MASiVE either as multiple PPT signatures or as problematic elements (thus, potentially fragmented and filtered out during the MASiVE pipeline).

A look into *Arabidopsis thaliana* provided no mentionable results due to the degeneracy of the Endovir family (nearly all elements are fragmented), and thus the very low abundance of LTR-RTN retrotransposons in this small-genome plant.

4 CONCLUSIONS AND OUTLOOK

Based on highly detailed genome structure information specific for Sireviruses, we developed MASiVE, a method which we show is able to discover, with high sensitivity and even higher precision, intact Sireviruses in plant genomic sequences. MASiVE can also provide detailed information on Sireviruses that failed the successive filters, therefore expanding the view on the genome distribution of Sirevirus elements. Currently, the method is based on LTRharvest results, but we plan to remove this requirement since the multiple PPT signature appear to be powerful enough to support MASiVE on its own. Finally, the Sirevirus population can be further parsed by available tools that conduct detailed annotation of LTR-RTN genomes (Sperber *et al*., 2007). We believe MASiVE to be a valuable addition to the methodological arsenal for scientists interested in untangling the intriguingly complex genome landscape of plants.

Funding: FP6 Biosapiens Network of Excellence (contract number LSHG-CT-2003-503265).

Conflict of Interest: none declared.

REFERENCES

Baucom,R.S. *et al.* (2009) Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *Plos Genet.*, **5**, e1000732.

Birney,E. *et al.* (2004) Genewise and genomewise. *Genome Res.*, **14**, 988–995.

Bousios,A. *et al.* (2010) Highly conserved motifs in non-coding regions of Sirevirus retrotransposons: the key for their pattern of distribution within and across plants? *BMC Genomics*, **11**, 89.

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

Katoh,K. and Toh,H. (2008) Recent developments in the MAFFT multiple sequence alignment program, *Brief. Bioinform.*, **9**, 286–298.

McCarthy,E.M. and McDonald,J.F. (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics*, **19**, 362–367.

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

Sperber,G.O. *et al.* (2007) Automated recognition of retroviral sequences in genomic data—RetroTector (c). *Nucleic Acids Res.*, **35**, 4964–4976.

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.