BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 11 2013, pages 1458–1460
doi:10.1093/bioinformatics/btt150

Genome analysis

Advance Access publication March 28, 2013

MCScanX-transposed: detecting transposed gene duplications based on multiple colinearity scans

Yupeng Wang[1,2,*], Jingping Li[1] and Andrew H. Paterson[1,*]

[1]Plant Genome Mapping Laboratory, University of Georgia, Athens, GA 30602 and [2]Computational Biology Service Unit, Cornell University, Ithaca, NY 14853, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Gene duplication occurs via different modes such as segmental and single-gene duplications. Transposed gene duplication, a specific form of single-gene duplication, 'copies' a gene from an ancestral chromosomal location to a novel location. *MCScanX* is a toolkit for detection and evolutionary analysis of gene colinearity. We have developed *MCScanX-transposed*, a software package to detect transposed gene duplications that occurred within different epochs, based on execution of *MCScanX* within and between related genomes. *MCScanX-transposed* can be also used for integrative analysis of gene duplication modes for a genome and to annotate a gene family of interest with gene duplication modes.

Availability: *MCScanX-transposed* is freely available at http://chibba.pgml.uga.edu/mcscan2/transposed/.

Contact: wyp1125@gmail.com or paterson@plantbio.uga.edu

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

Received on December 19, 2012; revised on March 3, 2013; accepted on March 25, 2013

1 INTRODUCTION

Gene duplication is a primary mechanism for increasing genetic complexity and diversity (Ohno, 1970). It has long been known that gene duplication occurs via different modes (Freeling, 2009). Large-scale gene duplications including whole-genome duplication (WGD) and segmental duplication have occurred in different eukaryotic lineages such as fungi, plants and vertebrates (Coghlan *et al.*, 2005; Paterson *et al.*, 2010). Single-gene duplications, also widespread in eukaryotic genomes, by mechanisms include local (tandem or proximal) duplication, which creates two nearby duplicated genes (Wang *et al.*, 2012b), and transposed duplication, which 'copies' a gene from an ancestral location to a novel (transposed) location via either DNA or RNA-based mechanisms (Cusack and Wolfe, 2007). Different modes of gene duplication may contribute differentially to the evolution of specific lineages (Coghlan *et al.*, 2005).

The retention of duplicated genes following different modes of gene duplication is biased (Freeling, 2009). For example, in *Arabidopsis thaliana*, following WGDs, genes related to transcription factors, protein kinases and ribosomal proteins were preferentially retained; local duplicates are enriched for genes that encode membrane proteins and function in abiotic and biotic stress; gene families such as F-box, MADS-box, NBS-LRR and defensins are more likely to have transposed than others (Freeling, 2009). Moreover, different modes of gene duplication often show distinct evolutionary consequences (De Smet and Van de Peer, 2012; Wang *et al.*, 2012b). In rodents, transposed duplications show higher degrees of asymmetric sequence divergence than local duplications (Cusack and Wolfe, 2007). In *A.thaliana*, duplicated genes retained from single-gene duplications, especially transposed duplication, are more likely to show rewiring of gene networks than those retained from WGDs (*Arabidopsis* Interactome Mapping Consortium, 2011; Wang *et al.*, 2012b).

To study the evolutionary dynamics of duplicated genes, it is helpful to investigate gene duplication modes. Duplicated genes originating from WGD, and segmental duplication can be obtained by using software for detection of gene colinearity (homologous genes are in corresponding chromosomal regions and conserved orders). However, to our knowledge, automated tools for detecting transposed gene duplications are not available.

Transposed duplications in a genome, and the epochs during which they occurred, can be inferred based on colinearity conservation within and between related genomes (Woodhouse *et al.*, 2011). MCScan is an algorithm to scan genomes to identify colinear homologs and duplicated chromosomal regions (Tang *et al.*, 2008). *MCScanX* is a toolkit for colinearity detection based on an adjusted MCScan (i.e. *MCScanX*) algorithm and evolutionary analyses of identified colinearity (Wang *et al.*, 2012a). Here, we present *MCScanX-transposed*, a software package based on execution of *MCScanX* within and between related genomes, for detecting transposed gene duplications, which occurred within different epochs, as well as integrative analysis of gene duplication modes.

2 DESCRIPTION

2.1 Overview of the *MCScanX-transposed* package

The *MCScanX-transposed* package is comprised of one core program named *MCScanX-transposed.pl*, for detecting transposed gene duplications that occurred within different epochs and other modes of gene duplication including segmental, tandem and proximal duplications, and five downstream analysis programs, for further analysis of identified gene duplications.

The *MCScanX-transposed* package needs to be executed using command lines, on Mac OS or Linux computers, where Xcode

*To whom correspondence should be addressed.

or the Java SE Development Kit and 'libpng' should be pre-installed, respectively. The user is advised to read the 'readme.txt' file that provides detailed instructions on input file formats and command line options.

2.2 Input of *MCScanX-transposed.pl*

Before the execution of *MCScanX-transposed.pl*, '.gff' and '.blast' files, containing the chromosomal positions and homologous relationships (generated by BLASTP), respectively, for all genes in a target genome, and at least one outgroup (i.e. related genome that shows gene colinearity with the target genome), should be available. Two types of BLASTP files are needed: (i) an intra-species BLASTP file for the target genome and (ii) inter-species BLASTP files between the target genome and each of the outgroup genomes. We suggest that BLASTP retain no more than five homologs for each query gene (Tang *et al.*, 2008; Wang *et al.*, 2011) to reduce redundant duplication relationships while allowing the identification of fast-evolving duplicated genes.

2.3 Algorithm for detecting transposed gene duplications

Transposed duplications 'copy' genes from ancestral loci (parental duplicates) to novel loci (transposed duplicates). Technically, with ancestral and novel loci in the genome identified, a transposed duplication can be discerned if a pair of dispersed (i.e. distant and non-colinear) duplicated genes consists of an ancestral and a novel locus. The *MCScanX-transposed* algorithm adopts a procedure for identifying transposed duplications based on processing BLASTP output, which has been previously shown to work well in *A. thaliana* and rice genomes (Wang *et al.*, 2011., 2013).

The intra-species BLASTP file of the target genome is deemed the whole set of potential gene duplications. The *MCScanX-transposed* algorithm first executes the *MCScanX* algorithm based on this intra-species BLASTP file, to generate colinear gene pairs. These colinear gene pairs are regarded as segmental duplications. The *MCScanX-transposed* algorithm then detects tandem and proximal duplications, according to the following criteria: tandem duplications are paralogs consecutive to each other, and proximal duplications are paralogs near each other but separated by a few non-paralogous genes (configurable) on the chromosomes. The remaining dispersed gene duplications, i.e. the BLASTP hits after segmental, tandem and proximal duplications are removed, are searched for transposed duplications. Ancestral loci in the target genome are the inter-species colinear genes generated by executing *MCScanX* on the inter-species BLASTP files between the target and outgroup genomes. If a dispersed gene duplication consists of an ancestral locus and a novel locus, this duplication is deemed a transposed duplication. For a transposed duplicate, there may be multiple ancestral loci within an epoch. In this case, the transposed duplicate and its ancestral locus with the highest sequence identity are deemed a transposed duplication.

Based on sequential exclusion of the closest outgroup, the *MCScanX-transposed* algorithm generates transposed duplications that occurred after different speciation events. The transposed duplications that occurred within different epochs (i.e. between different speciation events) are derived by set subtraction of the transposed duplications that occurred after different speciation events.

2.4 Execution of *MCScanX-transposed.pl*

To execute *MCScanX-transposed.pl*, the target and outgroup genomes should be specified via the '-t' and '-c' options of the command line. To detect transposed duplications that occurred within a user-specified number ('−x' option) of epochs that are marked by different speciation events, multiple outgroups must be entered in the order of divergence from the target genome (most recent first).

On the successful execution of *MCScanX-transposed.pl*, different modes of gene duplication including segmental, tandem, proximal and transposed duplications are output as separate '.pairs' files, with each line containing one gene duplication. In transposed duplications, the first duplicated gene is the transposed locus. If transposed duplications are classified into different epochs, the transposed duplications belonging to each epoch are also output as a separate file.

2.5 Downstream analysis programs

Five downstream analysis programs are available in the *MCScanX-transposed* package. The program *add_ka_ks.pl* can annotate gene duplication modes with non-synonymous (Ka) and synonymous (Ks) substitution rates based on the Nei and Gojobori method (1986). The program *detect_dup_modes_for_a_gene.pl* provides single-gene query from the *MCScanX-transposed.pl* output. The program *detect_dup_modes_for_a_family.pl* provides query for a gene family from the *MCScanX-transposed.pl* output, which can be further visualized via a phylogenetic tree in which duplicated gene pairs of different modes are connected by curves of different colors, by the program *annotate_tree_with_dup_modes*, or a phylogenetic tree linked to a chromosome ideogram in which each pair of transposed and parental duplicates is demonstrated by a curve of unique color, by the program *annotate_tree_with_tra_dup*.

3 USAGE EXAMPLE

To illustrate the usefulness of *MCScanX-transposed*, we applied it to the genome of *Arabidopsis thaliana*, using five other eudicot genomes as possible outgroups including *Arabidopsis lyrata*, *Brassica rapa*, *Carica papaya*, *Populus trichocarpa* and *Vitis vinifera*. The genome annotations were downloaded from Phytozome (http://www.phytozome.net). For genes with multiple transcripts, the longest transcript was used. The phylogenetic relationships of the six genomes is shown in Figure 1A. BLASTP was executed within *A.thaliana* and between *A.thaliana* and each outgroup. For each gene, top five non-self BLASTP hits with *E*-value $< e^{-10}$ were retained. Detailed commands for the usage example is described on the software website.

First, we executed *MCScanX-transposed.pl* for *A.thaliana*, specifying *A.lyrata*, *B.rapa*, *C.papaya*, *P.trichocarpa* and *V.vinifera* as the outgroups and three epochs to be identified ('−x 3'). *MCScanX-transposed.pl* identified 4299 WGD/segmental, 2130 tandem, 784 proximal and 3766 transposed duplications, respectively. Further, *MCScanX-transposed.pl* identified 904 *A.thaliana*

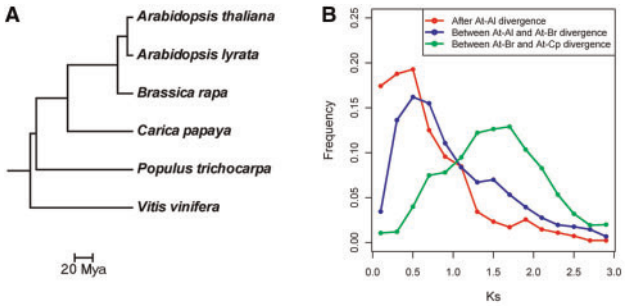

Fig. 1. A usage example of *MCScanX-transposed*. Species abbreviations: At, *Arabidopsis thaliana*; Al, *Arabidopsis lyrata*; Br, *Brassica rapa*; Cp, *Carica papaya*; Pt, *Populus trichocarpa*; and Vv, *Vitis vinifera*. (**A**) Phylogenetic relationships between *A.thaliana* and the outgroups. (**B**) Comparison of Ks distributions among the *A.thaliana* transposed duplications that occurred after At-Al divergence, between At-Al and At-Br divergence and between At-Br and At-Cp divergence

transposed duplications that occurred after *A.thaliana-A.lyrata* divergence, 1114 that occurred between *A.thaliana-A.lyrata* and *Arabidopsis-Brassica* divergence and 1748 between *Arabidopsis-Brassica* and *Arabidopsis-Carica* divergence. Comparison of the Ks (a proxy of time since duplication) distributions of the transposed duplications that occurred within these respective epochs (Fig. 1B) is consistent with the phylogeny of the four genomes.

These identified gene duplication modes can be further used to study the evolution of a gene family of interest. Here, we show an analysis of 105 *A. thaliana* MADS-box genes (Parenicova *et al.*, 2003). The program *detect_dup_modes_for_a_family.pl* identified 14 WGD/segmental, 6 tandem, 10 proximal and 31 transposed duplications. Based on a 'Newick' format tree of the MADS-box family, the program *annotate_tree_with_dup_modes* visualized the phylogenetic tree and annotated it with different duplication modes (Supplementary Fig. S1), and the program *annotate_tree_with_tra_dup* visualized the phylogenetic tree and demonstrated the gene transpositions between genomic locations within different epochs for the MADS-box family (Supplementary Figs S2–S4).

4 CONCLUSIONS

MCScanX-transposed is a software package able to detect transposed gene duplications that occurred within different epochs, also useful for integrative analysis of gene duplication modes. *MCScanX-transposed* can be used to study gene duplication mechanisms on both genome-wide and gene family levels.

ACKNOWLEDGEMENTS

The authors thank Dr Haibao Tang for helpful discussion, and Barry Marler and the Georgia Advanced Computing Resource Center for IT support.

Funding: A.H.P. appreciates funding from the National Science Foundation (NSF: DBI 0849896, MCB 0821096, MCB 1021718).

Conflict of Interest: none declared.

REFERENCES

Arabidopsis Interactome Mapping Consortium. (2011) Evidence for network evolution in an *Arabidopsis* interactome map. *Science*, **333**, 601–607.

Coghlan,A. *et al.* (2005) Chromosome evolution in eukaryotes: a multi-kingdom perspective. *Trends Genet.*, **21**, 673–682.

Cusack,B.P. and Wolfe,K.H. (2007) Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. *Mol. Biol. Evol.*, **24**, 679–686.

De Smet,R. and Van de Peer,Y. (2012) Redundancy and rewiring of genetic networks following genome-wide duplication events. *Curr. Opin. Plant Biol.*, **15**, 168–176.

Freeling,M. (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.*, **60**, 433–453.

Nei,M. and Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.

Ohno,S. (1970) *Evolution by Gene Duplication*. Springer Verlag, New York.

Parenicova,L. *et al.* (2003) Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in *Arabidopsis*: new openings to the MADS world. *Plant Cell*, **15**, 1538–1551.

Paterson,A.H. *et al.* (2010) Insights from the comparison of plant genome sequences. *Annu. Rev. Plant Biol.*, **61**, 349–372.

Tang,H. *et al.* (2008) Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.*, **18**, 1944–1954.

Wang,Y. *et al.* (2011) Modes of gene duplication contribute differently to genetic novelty and redundancy, but show parallels across divergent angiosperms. *PLoS One*, **6**, e28150.

Wang,Y. *et al.* (2012a) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.*, **40**, e49.

Wang,Y. *et al.* (2012b) Genome and gene duplications and gene expression divergence: a view from plants. *Ann. N. Y. Acad. Sci.*, **1256**, 1–14.

Wang,Y. *et al.* (2013) Gene body methylation shows distinct patterns associated with different gene origins and duplication modes and has a heterogeneous relationship with gene expression in Oryza sativa (rice). *New Phytol.*, **198**, 274–283.

Woodhouse,M.R. *et al.* (2011) Different gene families in *Arabidopsis thaliana* transposed in different epochs and at different frequencies throughout the rosids. *Plant Cell*, **23**, 4241–4253.