Gene 562 (2015) 95–106

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Comparative analysis of active retrotransposons in the transcriptomes of three species of heteromyid rodents

Yanzhu Ji [a,*], Nicholas J. Marra [a,1], J. Andrew DeWoody [a,b]

[a] Department of Forestry and Natural Resources, Purdue University, IN 47907, USA
[b] Department of Biological Sciences, Purdue University, IN 47907, USA

ARTICLE INFO

Article history:
Received 17 October 2014
Received in revised form 16 February 2015
Accepted 17 February 2015
Available online 20 February 2015

Keywords:
Transposable element
Exaptation
Transcriptome
Heteromyidae
Differential expression
Multi-read

ABSTRACT

Virtually every eukaryotic genome is replete with transposable elements (TEs). TE activity, or lack thereof, is of considerable evolutionary interest as TE insertions actively drive genome evolution by altering gene expression and/or function through different mechanisms. Herein, we take a comparative approach to better understand the variation in TE transcriptional activity. Our goals were to identify transcriptionally active TEs and to evaluate their relative expression levels in an effort to identify key determinants of TE activity. We do so in three related non-model rodent species (*Dipodomys spectabilis*, *Chaetodipus baileyi*, and *Heteromys desmarestianus*). We used these species (a) because rodents have long been productive models for the study of TEs; (b) because the known and dated relationships among these species permit strong phylogenetic inference (e.g., ancestral character states); and (c) because we have previously characterized key genes that underlie evolutionary adaptations in these species (e.g., osmoregulation). We used RNA-seq to characterize the transcriptomes of two different tissues – kidney and spleen – in each of these three species. Our data revealed a diversity of retrotransposons that were actively transcribed in these rodents (including LINEs, retroviruses/ERVs, and SINEs). We also identified tissue-specific differences in retrotransposon-activity in about half of the retrotransposons detected. We interpret these differences in TE activity in light of overall gene expression, with the transcribed SINEs enriched in differentially expressed sequences in all three species. These transcribed TEs may simply represent transpositionally active TEs. Alternatively, they may be TEs that have been co-opted to represent functionally important components of the host transcriptome (i.e., exaptations). Overall, our results contribute to a growing appreciation for the evolutionary interplay between TEs and their hosts.

1. Introduction

Transposable elements (TEs) are DNA segments that often have the ability to move, and sometimes proliferate, across the genome (McClintock, 1950). Two broad categories of TEs have been established that differ in their mode of transposition (Craig, 2002; Finnegan, 1989; Kazazian, 2011). Class I TEs transpose via an RNA intermediate with the aid of the often cis-encoded reverse transcriptase. In other words, these retrotransposons utilize a "copy and paste" mechanism of transposition whereby elements are first transcribed into RNA, then reverse-transcribed into DNA, and subsequently incorporated into new genomic locations. Class II TEs, or DNA transposons, employ a "cut and paste" mechanism whereby genomic DNA elements transpose without an RNA intermediate. TEs within each of these two classes are further categorized depending on their autonomy, their structural features, and their sequence similarities (Wicker et al., 2007).

Ongoing expression of TEs is of evolutionary interest for proximate and ultimate reasons. TEs may be passively expressed because of proximity to a host gene and thus contribute to the cellular transcriptome with no obvious benefits, but there are also adaptive reasons for the expression of TEs. Proximately, transcription into RNA suggests one of two functional possibilities that include TE transposition and/or exaptation, whereby TEs are transcribed if they have been co-opted by the host as genes or regulatory elements. Ultimately, transcription provides a snapshot of the complex coevolutionary interplay between host and TEs. The study of transcribed TEs can provide insight into whether TEs proliferate and/or whether exapted TEs contribute to host gene expression.

Latent TEs can become transpositionally active if the TE copy is transcribed and expressed, and the reproductive efforts of these elements can actively alter the structure of the host genome as well as directly influence fitness (Doolittle and Sapienza, 1980; Baillie et al.,

Abbreviations: APOBEC3, apolipoprotein B mRNA editing enzyme catalytic polypeptide 3; ERV, endogenous retrovirus; FPKM, fragments per kilobase transcript per million mapped reads; LINE, long interspersed element; L1, long interspersed element-1; MS, mappability score; MSA, multiple sequence alignment; ORF, open reading frame; POT1, protection of telomeres; RT, reverse transcriptase; SINE, short interspersed element; SNP, single nucleotide polymorphism; TE, transposable element; Tie-1, tyrosine-protein kinase receptor-1; TPM, transcripts per million.

* Corresponding author.
E-mail addresses: ji20@purdue.edu (Y. Ji), njm77@cornell.edu (N.J. Marra), dewoody@purdue.edu (J.A. DeWoody).
[1] Present address: Department of Population Medicine and Diagnostic Sciences, Cornell University, Ithaca, NY 14853, USA.

http://dx.doi.org/10.1016/j.gene.2015.02.058
0378-1119/

2011; Eickbush and Eickbush, 2011). Alternatively, TEs can senesce or die when mutation and/or recombination render them unable to transpose (i.e., reproduce), yet their sequences can still impact the host (Kines and Belancio, 2012; Schumann et al., 2010). Dead TEs can be co-opted by the host genome into novel functional genes (e.g., Fugmann et al., 2000) or as regulatory elements such as enhancers, promoters, or long non-coding RNAs (Brosius and Gould, 1992; Emera et al., 2012; Faulkner et al., 2009; Zhang et al., 2013). These exapted TEs can become essential to hosts (e.g., Levis et al., 1993; Fugmann et al., 2000). As with host genes, exapted TEs are expected to be expressed under proper cellular environment (Emera et al., 2012; Schmidt et al., 2012; Xie et al., 2013). As the cellular environment differs among tissues, patterns of tissue-specific expression could provide clues as to the function (if any) of transcriptionally active TEs.

The transcriptional activity of TEs has long been studied in model organisms such as human and mouse (Flockerzi et al., 2008; Medstrand and Blomberg, 1993; Zhang et al., 2013). However, model organisms are not always representatives of TE activities (e.g., Casavant et al., 2000; Ray et al., 2007). For example, the invasion of new environments can shape the TE repertoire of a species (Kim et al., 2014; Wijayawardena et al., in press). Modern sequencing technologies provide new opportunities to expand such studies to non-model species by screening for transcribed TEs directly at the transcriptome level in species with or without a reference genome (Hobbs et al., 2014; Jiang et al., 2012; Lopes et al., 2013; Yan et al., 2014).

Herein we report our transcriptomic work on TEs in three heteromyid species. These three species (*Dipodomys spectabilis*, *Chaetodipus baileyi*, and *Heteromys desmarestianus*) are new world rodents that diverged from Muroid rodents (e.g., *Mus musculus* and *Rattus rattus*) roughly 60–90 mya (Adkins et al., 2003; Honeycutt, 2009). Although quite divergent from Murid rodents, our suite of species is well-suited for comparative analyses because all three lineages diverged from one another roughly 22 mya. Phylogenetic analyses have revealed that *C. baileyi* and *H. desmarestianus* form a sister clade relative to *D. spectabilis* (Hafner et al., 2007). Thus, this known phylogeny [i.e., (*Chaetodipus* + *Heteromys*) + *Dipodomys*] allows us to polarize differences in TE abundance and diversity. Furthermore, transcriptomes have been characterized for osmoregulation and immunity via RNA-seq (Marra et al., 2012, 2014; Marra and DeWoody, 2014). Herein, we compared transcriptional activities of TEs between tissues and across species. We focused on retrotransposons because they are more prevalent than DNA transposons in rodent genomes (Gibbs et al., 2004; Waterston et al., 2002) and because some retrotransposons are transpositionally active whereas DNA transposons are inactive in most mammals (reviewed in Huang et al., 2012; but see Ray et al., 2007, 2008). Our data reveal a diversity of retrotranposons actively expressed in these rodents as well as differences in retrotransposon activity between tissues as indicated by differential expression.

2. Materials and methods

2.1. Sample collection, library construction, and RNA-seq

Specimens and molecular data were initially collected for studies related to osmoregulation and immunity (Marra and DeWoody, 2014; Marra et al., 2012, 2014). Only a brief overview is presented here; detailed descriptions are available in the original papers. Four wild *D. spectabilis* were collected from a desert population near Portal, Arizona and four wild *C. baileyi* were collected at the same site. In an environmental contrast, four wild *H. desmarestianus* were collected from the tropical lowlands (i.e., rainforest) of La Selva biological station in Costa Rica. Kidney and spleen tissues were collected, minced, and frozen immediately in Trizol reagent. A total of 24 cDNA libraries (3 species × 4 individuals × 2 tissue types) were constructed and barcoded using standard techniques. All twenty-four libraries, after barcoding and pooling, were sequenced using 100 bp paired-end reads on two lanes of the Illumina HiSeq 2000 platform in the Purdue Genomics facility.

2.2. Quality control, transcriptome assembly, and validation

Illumina reads that contained sequencing primers or had low quality scores were trimmed using Trimmomatic (Lohse et al., 2012). Bowtie2 (Langmead and Salzberg, 2012) was used to remove rRNA and phiX sequences from paired reads. Following these quality control measures, the remaining Illumina reads were pooled by species and assembled. With all Illumina reads from eight libraries per species pooled, the Trinity package with default parameters (Grabherr et al., 2011) was used to generate transcriptome assemblies for each species.

To validate the three transcriptomes, we first removed all sequences that had extremely low expression across libraries using RSEM (Li and Dewey, 2011). Briefly, we estimated expression levels for each of the eight libraries of each species by mapping back reads per library to the corresponding species-specific assembly as an integrated step in RSEM. Next, we utilized fragments per kilobase transcript per million mapped reads (FPKM) measured by RSEM to filter out sequences in each species that have FPKM < 1 across all eight libraries (Mortazavi et al., 2008). Using the filtered assemblies of each species, we employed the program CEGMA to determine that the dataset contains a complete collection of core eukaryote genes (Parra et al., 2007).

Reads that map to multiple positions have the potential to introduce unknown biases into measures of gene expression, and some mapping ambiguity is expected for isoforms and for highly similar TE copies. We aimed to reduce such biases by checking the Trinity assemblies for redundant sequences that could cause mapping problems. Trinity assemblies have three hierarchical levels whereby "components" are comprised of "subcomponents", and "subcomponents" are comprised of "isoforms". Isoform-level sequences are bound to share identical regions and be redundant because they often originate from the same gene (and are represented by the same de Bruijn graph in Trinity). Thus we grouped isoforms into higher categories (i.e., components or subcomponents) to reduce read mapping ambiguity. We further checked whether sequences from subcomponents or components represent the smallest group of redundant sequences by mappability analysis using GEM package and custom perl scripts (Derrien et al., 2012; see Supplementary material).

2.3. Autonomous retrotransposon identification, verification, and classification

TEs in genomes can be detected using homology-based, structure-based, and de novo approaches (Bergman and Quesneville, 2007; Feschotte et al., 2009; Lerat, 2010). However, only homology-based approaches are appropriate for the detection of TEs in transcriptomes (Jiang et al., 2012). We expected protein-level homologies to be more robust than nucleotide-level homologies given (1) the ancient divergence between our study species and model (Murid) rodents (Hafner et al., 2007); (2) the complexity of transcriptomes, which contain a mixture of immature mRNA and mature mRNA (Pelechano et al., 2013); and (3) the lack of fully annotated reference genomes of our study species. Therefore, we relied on protein-level homology for autonomous TEs (which encode enzymes to transpose) but considered DNA-level homology for non-autonomous TEs that do not encode proteins.

For autonomous retrotransposons that all encode reverse transcriptases (RTs), we utilized RT to aid in both identification and classification (Eickbush and Jamburuthugoda, 2008; Tristem, 2000; Xiong and Eickbush, 1990). We used the pfam entry of RT as the query to search against translated Trinity assemblies with HMMER (Eddy, 1998). To include RTs with frameshift mutations, all hits with matched length ≥ 80 amino acids of queries were retained whereas shorter matches were discarded.