frontiers
in Genetics

Evolinc: a tool for the identification and evolutionary comparison of long intergenic non-coding RNAs

Andrew D. Nelson[1*], Upendra K. Devisetty[2], Kyle Palos[1], Asher K. Haug-Baltzell[3], Eric Lyons[4], Mark A. Beilstein[1*]

[1]School of Plant Sciences, University of Arizona School of Plant Sciences, USA, [2]CyVerse, USA, [3]Genetics Graduate Interdisciplinary Group, University of Arizona, USA, [4]School of Plant Sciences, University of Arizona, USA

Submitted to Journal:
Frontiers in Genetics

Specialty Section:
Bioinformatics and Computational Biology

ISSN:
1664-8021

Article type:
Methods Article

Received on:
22 Feb 2017

Accepted on:
12 Apr 2017

Provisional PDF published on:
12 Apr 2017

Frontiers website link:
www.frontiersin.org

Citation:
Nelson AD, Devisetty UK, Palos K, Haug-baltzell AK, Lyons E and Beilstein MA(2017) Evolinc: a tool for the identification and evolutionary comparison of long intergenic non-coding RNAs. *Front. Genet.* 8:52. doi:10.3389/fgene.2017.00052

Evolinc: a tool for the identification and evolutionary comparison of long intergenic non-coding RNAs

Andrew D. L. Nelson[*,1,†], Upendra K. Devisetty[*,2], Kyle Palos[1], Asher K. Haug-Baltzell[3], Eric Lyons[2,3], and Mark A. Beilstein[1,†]

Authors: [1]School of Plant Sciences, University of Arizona, Tucson, Arizona, 85721, [2] CyVerse, Bio5, University of Arizona, Tucson, Arizona, 85721, [3]Genetics Graduate Interdisciplinary Group, University of Arizona, Tucson, Arizona, 85721

* These authors contributed equally to this manuscript.

† Corresponding Authors

Corresponding Authors:

Mark Beilstein, 1140 E. South Campus Drive, 303 Forbes Building, Tucson, Arizona, 85721-0036, 520-626-1562, mbeilstein@email.arizona.edu

Andrew Nelson, 1140 E. South Campus Drive, 303 Forbes Building, Tucson, Arizona, 85721-0036, 520-626-1563, andrewnelson@email.arizona.edu

Running Title: Identification and comparative evolution of lincRNAs with Evolinc

Abstract

Long intergenic non-coding RNAs (lincRNAs) are an abundant and functionally diverse class of eukaryotic transcripts. Reported lincRNA repertoires in mammals vary, but are commonly in the thousands to tens of thousands of transcripts, covering ~90% of the genome. In addition to elucidating function, there is particular interest in understanding the origin and evolution of lincRNAs. Aside from mammals, lincRNA populations have been sparsely sampled, precluding evolutionary analyses focused on their emergence and persistence. Here we present Evolinc, a two-module pipeline designed to facilitate lincRNA discovery and characterize aspects of lincRNA evolution. The first module (Evolinc-I) is a lincRNA identification workflow that also facilitates downstream differential expression analysis and genome browser visualization of identified lincRNAs. The second module (Evolinc-II) is a genomic and transcriptomic comparative analysis workflow that determines the phylogenetic depth to which a lincRNA locus is conserved within a user-defined group of related species. Here we validate lincRNA catalogs generated with Evolinc-I against previously annotated Arabidopsis and human lincRNA data. Evolinc-I recapitulated earlier findings and uncovered an additional 70 Arabidopsis and 43 human lincRNAs. We demonstrate the usefulness of Evolinc-II by examining the evolutionary histories of a public dataset of 5,361 Arabidopsis lincRNAs. We used Evolinc-II to winnow this dataset to 40 lincRNAs conserved across species in Brassicaceae. Finally, we show how Evolinc-II can be used to recover the evolutionary history of a known lincRNA, the human telomerase RNA (TERC). These latter analyses revealed unexpected duplication events as well as the loss and subsequent acquisition of a novel TERC locus in the lineage leading to mice and rats. The Evolinc pipeline is currently integrated in CyVerse's Discovery Environment and is free for use by researchers.

Introduction

A large, and in some cases predominant, proportion of eukaryotic transcriptomes are composed of long non-coding RNAs (lncRNAs) (Hangauer et al., 2013; Guttman et al., 2009; Cabili et al., 2011; H., Wang et al., 2015; Liu et al., 2012). LncRNAs are longer than 200 nucleotides (nt) and exhibit low protein-coding potential (non-coding). While some transcripts identified from RNA-seq are likely the result of aberrant transcription or miss-assembly, others are bona fide lincRNAs with various roles (see [Wang and Chang, 2011; Ulitsky and Bartel, 2013] for a review of lncRNA functions). To help factor out transcriptional “noise”, additional characteristics are used to delineate lncRNA. These additional characteristics focus on factors such as reproducible identification between experiments, degree of expression, and number of exons (Derrien et al., 2012). In general, lncRNAs display poor sequence conservation among even closely related species, are expressed at lower levels than protein-coding genes, and lack functional data.

The function of any particular lncRNA is likely to influence its evolution. One means of inferring that a transcript is a functional lncRNA and not an artefact is the degree of conservation we observe at that locus between two or more species. This conservation can be observed at the sequence, positional, and transcriptional level (Ulitsky, 2016). Comparative approaches to identify conserved and potentially functional lncRNAs typically focus on long *intergenic* non-coding RNAs (lincRNAs), since their evolution is not constrained by overlap with protein-coding genes. In vertebrates, lincRNA homologs have been identified in species that diverged some 400 million years ago (MYA), whereas in plants lincRNA homologs are primarily restricted to species that diverged < 100 MYA (Ulitsky et al., 2011; Necsulea et al., 2014; Nelson et al., 2016; Liu et al., 2012; Li et al., 2014; Zhang et al., 2014; Mohammadin et al., 2015). Importantly, the conserved function of a handful of these lincRNAs have been experimentally verified *in vivo* (Hawkes et al., 2016; Migeon et al., 1999; Quinn et al., 2016).

One major factor inhibiting informative comparative genomics analyses of lincRNAs is the lack of robust sampling and user-friendly analytical tools. Here we present Evolinc, a lincRNA identification and comparative analysis pipeline. The goal of Evolinc is to rapidly and reproducibly identify candidate lincRNA loci, and examine their genomic and transcriptomic conservation. Evolinc relies on RNA-seq data to annotate putative lincRNA loci across the target genome. It is designed to utilize cyberinfrastructure such as the CyVerse Discovery Environment (DE), thereby alleviating the computing demands associated with transcriptome assembly (Merchant et al., 2016). The pipeline is divided into two modules. The first module, Evolinc-I, identifies putative lincRNA loci, and provides output files that can be used for analyses of differential expression, as well as visualization of genomic location using the EPIC-CoGe genome browser (Lyons et al., 2014). The second module, Evolinc-II, is a suite of tools that allows users to identify regions of conservation within a candidate lincRNA, assess the extent to which a lincRNA is conserved in the genomes and transcriptomes of related species, and explore patterns of lincRNA evolution. We demonstrate the versatility of Evolinc on both large and small datasets, and explore the evolution of lincRNAs from both plant and animal lineages.

Materials and Methods

In this section we describe how the two modules of Evolinc (I and II) work, and explain the data generated by each.

Evolinc-I: LincRNA identification

Evolinc-I minimally requires the following input data: a set of assembled and merged transcripts from Cuffmerge or Cuffcompare (Trapnell et al., 2010) in gene transfer format (GTF), a reference genome (FASTA), and a reference genome annotation (GTF/GFF/GFF3). From the transcripts provided in the GTF file, only those longer than 200 nt are kept for further analysis. Transcripts with high protein-coding potential are removed using two metrics: 1) open reading frames (ORF) encoding a protein > 100 amino acids, and 2) similarity to the UniProt protein database (based on a 1E-5 threshold). Filtering by these two metrics is carried out by Transdecoder (https://transdecoder.github.io/) with the BLASTp step included. These analyses yield a set of transcripts that fulfill the most basic requirements of lncRNAs. Due to anticipated lack of sequence homology or simple lack of genome data that users may deal with, we did not include ORF conservation as a filtering step within Evolinc-I, but instead suggest users to perform a PhyloCSF or RNAcode (Washietl et al., 2011) step after homology exploration by Evolinc-II.

The role of transposable elements (TEs) in the emergence and function of lncRNAs is an active topic of inquiry (Wang et al., 2017; Kapusta et al., 2013). To facilitate these studies, Evolinc allows the user to separate lncRNAs bearing similarity to TEs into a separate FASTA file. This is performed by BLASTn (Camacho et al., 2009; Altschul et al., 1990), with the above lncRNAs as query against a user provided TE database (in FASTA format). Many different TE datasets can be acquired from Repbase (http://www.girinst.org/), PGSB-REdat [http://pgsb.helmholtz-muenchen.de/plant/recat/; (Spannagl et al., 2016)], or DPTEdb: Dioecious Plant Transposable Elements Database (http://genedenovoweb.ticp.net:81/DPTEdb/index.php). We considered lncRNAs that exceeded a bit score value of 200 and an E-value threshold of 1E-20 to be TE-derived. These stringent thresholds remove TE-derived lncRNAs with high similarity to TEs, but allow for retention of lncRNAs with only weak similarity to TEs, perhaps reflecting older TE integration events or TE exaptation events (Johnson and Guigó, 2014). To thoroughly identify TE-derived lncRNAs, we suggest building the TE database from as many closely related and relevant species as possible. The output from these analyses includes a sequence file (FASTA) for each TE-derived lncRNAs, and BED files to permit their visualization via a genome browser. These transcripts are excluded from the file of putative lncRNAs used in downstream analyses by Evolinc-I.

Candidate lncRNAs are next compared against reference annotation files using the BEDTools package (Quinlan and Hall, 2010) to determine any overlap with known genes. Some reference annotations distinguish between protein-coding and other genes (lncRNAs, pseudogenes, etc). If this style of reference annotation is available, we suggest running Evolinc-I twice, once with an annotation file containing only protein-coding genes (generated with a simple grep command) and once with all known genes. This is a simple way to distinguish between the identification of novel putative lncRNAs and known (annotated) lncRNAs. We also recommend using an annotation file that contains 5' and 3' UTRs where possible. If this is unknown, the genome coordinates within the reference annotation file should be manually adjusted to include additional sequence on either end of known genes (i.e., 500bp). This number can be adjusted to adhere to community-specific length parameters for intergenic space. We provide two simple ways to update genome annotation files, either for the command line: (https://github.com/Evolinc/Accessory-scripts) or an app within the DE (Modify_GFF_Coordinates) Evolinc-I identifies lncRNAs whose coordinates overlap with those of a known gene. These gene-associated lncRNAs are then sorted into groups based on direction of overlap to known genes: sense or antisense-overlapping lncRNA transcripts (SOT or AOT, respectively). Keep in mind that in order for these inferences to be made, either strand-specific RNA-sequencing must be performed or the lncRNA must be multi-exonic. Sequence FASTA and BED files for each group of overlapping lncRNAs are generated by Evolinc-I for the user to inspect. Demographic data are also generated for each of these lncRNA types (explained further below).

LncRNAs that do not overlap with known genes and have passed all other filters are considered (putative) lincRNAs. Evolinc-I also deals with optional input data that may increase the confidence in the validity of particular candidate lincRNAs. For example, when users provide transcription start site coordinates (in BED format), Evolinc-I identifies lincRNAs in which the 5' end of the first exon is within 100bp of any transcriptional start site (TSS). LincRNAs with TSS are annotated as “CAGE_PLUS” in the FASTA sequence file (lincRNAs.FASTA), and the identity of such lincRNAs is recorded in the final summary table (Final_summary_table.tsv). Optionally, Evolinc-I identified lincRNAs (termed Evolinc-lincRNAs) can also be tested against a set of user-defined lincRNAs that are not found in the reference annotation (i.e., an in-house set of lincRNAs not included in the genome annotation files). When the coordinates for a set of such lincRNA loci are provided in general feature format (GFF), Evolinc-I will use these data to determine if any putative Evolinc-lincRNAs are overlapping. These loci are appended with “_overlapping_known_lncRNA” in the lincRNA.FASTA file. The identity of the overlapping (known lncRNA) is listed for each Evolinc-lincRNA in the final summary table (Final_summary_table.tsv).

Output from Evolinc-I

Evolinc-I generates a sequence file and BED file for TE-derived lncRNAs, AOT or SOT lncRNAs, and intergenic lncRNAs (lincRNAs). We highly recommend scanning the FASTA files for the presence of ribosomal and other RNAs against the Rfam database (http://rfam.xfam.org/search#tabview=tab1) and removing these before further analysis. The BED file is useful for direct visualization in a genome browser (Buels et al., 2016) or intersecting with other BED files generated from different Evolinc-I analyses (Quinlan and Hall, 2010). An updated genome annotation file is created, appending only the lincRNA loci to the user-supplied reference annotation file. This file can then be used with differential expression analysis programs such as DESeq2 or edgeR (Anders and Huber, 2010; Robinson and Oshlack, 2010). In addition, two types of demographic outputs are generated. For SOT, AOT, and lincRNAs, a report is created that describes the total number of transcripts identified for each class (isoforms and unique loci), GC content, minimum, maximum, and average length. For lincRNAs only, a final summary table is generated with the length and number of exons for each lincRNA, as well as TSS support and the ID of any overlapping, previously curated lincRNA. The Evolinc-I workflow is shown in Figure 1A.

Additional Evolinc-I resources

We have also included in the DE and in the GitHub repository (https://github.com/Evolinc/Accessory-scripts/) an assortment of scripts and workflows that will prevent known errors from occurring in transcript assembly and lncRNA identification. For instance, genome FASTA files often have chromosome headers prefaced with lcl| or gi|, whereas the corresponding genome annotation (GFF) file does not. Some tools such as Cuffmerge and Cuffcompare cannot parse genome associated files with non-matching chromosome IDs, resulting in an output file that will not work with Evolinc-I. To address this issue, we have included a short script called “clean_fasta_header.sh” to the GitHub repository and an app with the same name in the DE.

We also created an additional workflow to streamline the read mapping and transcript assembly process to generate input for Evolinc-I. This workflow is available as an app in the DE called Hisat2-Cuffcompare v1.0 and as a script in our GitHub repository under Accessory_scripts. Hisat2-Cuffcompare requires one or more SRA IDs, a genome sequence file (FASTA), and a genome reference annotation file (GFF) as input. Hisat2-Cuffcompare uses HISAT2 (Pertea et al., 2016) to map reads, either Cufflinks or StringTie (Trapnell et al., 2010; Pertea et al., 2016) to assemble transcripts, and then Cuffmerge or Cuffcompare to generate the input file for Evolinc-I.

Identifying lincRNA conservation with Evolinc-II

Evolinc-II minimally requires the following input data: a FASTA file of lincRNA sequences, FASTA file(s) of all genomes to be interrogated, and a single column text file with all species listed in order of phylogenetic relatedness to the query species (example and further elaboration on the species list in File S1). Many of these genomes can be acquired from CoGe (www.genomevolution.org) or the genome_data folder for Evolinc within the DE (/iplant/home/shared/iplantcollaborative/example_data/Evolinc.sample.data), and lincRNA sequences can be obtained from either the output of Evolinc-I or from another source. Genome FASTA files should be cleaned of pipe (|) characters (see above) and lincRNA FASTA files should not include underscores. The number, relationship, and divergence times of the genomes chosen will depend on the hypotheses the user intends to test. We recommend using many closely related species (intra-family), where possible, and then picking species outside of the family of interest depending on quality of genome annotation and number of lincRNAs identified. To determine the transcriptional status of lincRNA homologs across a group of species, Evolinc-II can optionally incorporate genome annotation files (GFF) and known lincRNA datasets from target species in FASTA format. In addition, Evolinc-II can incorporate motif and structure data, in BED format, to highlight any potential overlap between conserved regions and user-supplied locus information.

Evolinc-II starts by performing a series of reciprocal BLASTn (Camacho et al., 2009) searches against provided target genomes, using a user-defined set of lincRNAs as query and user chosen E-value cutoff. We suggest starting at an E-value cutoff of 1E-20 because we found that across 10 Brassicaceae genomes, and independently among human, orangutan and mouse, this value was optimal for recovering reciprocal and syntenic sequence homologs (Nelson et al., 2016). While 1E-20 represents a starting point for these analyses, lincRNA homolog recovery relies on a variety of factors (i.e., background mutation rate, genome stability, evolutionary distance of species / taxa being analyzed, and genome size) that could affect the E-value cutoff most likely to return homologous loci among related genomes. Thus, we recommend “calibrating” Evolinc-II using varying E-values with at least three genomes (two genomes aside from the query) of varying evolutionary distances from the query species before including a larger (> 3) set of genomes. If few sequence homologs are recovered for distantly related species, the user should try lowering the E-value. For command-line users examining transposable element derived lncRNAs identified by Evolinc-I, it might be useful to replace all instances of “blastn” within “Building_Families.sh” with “rmblastn”. RmBLASTn is a version of BLASTn with Repeat Masker extensions, which will provide more sensitivity when examining conservation of this set of lncRNAs (www.repeatmasker.org). After BLASTn (or RmBLASTn), the top blast hit (TBH) to the query lincRNA is identified for each additional genome included. Multiple, non-redundant hits falling within the same genomic region, which is likely to occur when the query lincRNA is multi-exonic, are merged as a single TBH. Sequence for all TBHs are then used as query in reciprocal BLAST searches (see below). For researchers interested in inferring orthology versus paralogy of a sequence homolog in a particular subject species, the coordinates of all BLAST hits that passed the E-value cutoff are retained in the file: Homology_search/Subject_species.out.merged.gff. However, to reduce computing time, subsequent analyses are confined to TBHs. Query lincRNAs for which a TBH is not identified in the first iteration (i.e., did not pass the E-value cutoff), are subdivided into non-overlapping segments of 200 nt and each segment is used as query in a second set of BLAST searches using similar parameters as the initial search. This reiterative step can be useful in finding short regions of sequence similarity in long query lincRNAs.

TBHs from each species included in the analysis are then used as query sequences in a reciprocal BLAST against the genome of the species whose lincRNA library was used in the original query. For a locus to be considered homologous to the original query lincRNA locus, both loci must be identified as the TBH to each other. This is especially useful when performing searches using a low E-value cutoff, as it reduces the chance of random sequence being returned as a sequence homolog. TBHs that pass the reciprocity test are appended with “Homolog” in the final FASTA sequence alignment file (“query_lincRNA_1”_alignment.FASTA).

As TBHs from each target genome are identified, they are scanned for overlap against optional genome reference annotation datasets (GFF) and known lincRNA files (FASTA). The identifier number (ID) of all TBHs with overlap against these two datasets is appended with either “Known_gene” or “Known_lincRNA”. The identity of the overlapping gene is retained in the final summary table (final_summary_table.tsv) as well as in each FASTA sequence alignment file (see below). Many genes and almost all lincRNAs are annotated based on transcriptional evidence. Thus, this is a simple way of determining if a query lincRNA corresponds to a locus with evidence of transcription in another species. In addition, when working with a poorly annotated genome, comparing against well-annotated species can provide additional levels of information about the putative function of query lincRNAs. For example, if the homologous locus of a query lincRNA overlaps a protein-coding gene in that species, it could indicate that the query lincRNA is a protein-coding gene, or a pseudogene.

All TBH sequences for a given query lincRNA are clustered into a family. For example, an Evolinc-II analysis that queries ten lincRNAs across a set of target genomes will result in ten lincRNA families, populated with the TBH from each target genome. Genomes that do not return a TBH at the specified E-value cutoff (from either full-length or segmented searches), or whose TBH does not pass the reciprocity test, will not be represented in the family. These lincRNA families are then batch aligned using MAFFT under default settings with 1000 iterations (Katoh and Standley, 2013). Command-line users wishing to modify the MAFFT parameters can do so on line 27 of the Batch_MAFFT script available in our GitHub repository (below). The alignment file for each lincRNA family can be downloaded into a sequence viewer. Evolinc-II will also infer phylogeny from the sequence alignment using RAxML v8.2.9 (Stamatakis, 2014) under the GTRGAMMA model, with rapid bootstrap analysis of 1000 bootstrap datasets. Parameters for RAxML are viewable and modifiable in the Batch_RAxML file. Gene trees are reconciled with a user-provided species tree, in Newick format, using Notung (Durand et al., 2006). This latter analysis pinpoints duplication and loss events that may have occurred during the evolution of the lincRNA locus. Bootstrap support of 70 is required for Notung to choose the gene tree model over the species tree. The Notung reconciled tree is available to view in PNG format within the CyVerse DE. Duplication and loss events are denoted by a red D or L, respectively (Example in Figure S4). The Evolinc-II workflow is shown in Figure 2A.

Output from Evolinc-II

Evolinc-II generates sequence files containing lincRNA families with all identified sequence homologs from the user-defined target genomes. In addition, a summary statistics table of identified lincRNA loci based on depth of conservation and overlapping features (e.g., genes, lincRNAs, or other user defined annotations) is generated. The identity of overlapping features (e.g., gene, known lincRNAs) in each genome for which a sequence homolog was identified is listed (Shown for the Liu-lincRNAs in File S3). To visualize conserved regions of all query lincRNAs, a query-centric BED file is generated that is ready for import into any genome browser. An example using the genome browser embedded within CoGe (Tang and Lyons, 2012) is shown below (Figure 2C). Following phylogenetic analysis, a reconciled gene tree is produced with predicted duplication and loss events indicated. Lastly, to provide the user with a broad picture of lincRNA conservation within their sample set, a bar graph is produced that indicates the number and percent of recovered sequence homologs in each species (Figure S2A).

Data and software availability

All genomes used in this work, including version and source, are listed in File S1. The accession number of all short read archive files (SRA) used in this work, including project ID, TopHat (Kim et al., 2013) read mapping rate, and total reads mapped for each SRA are shown in File S1. Genomic coordinates for lincRNAs identified by Evolinc-I are listed by species in BED/GFF format in File S2. LincRNAs were scanned for the presence of ribosomal and other known RNAs by batch searching against the Rfam database (http://rfam.xfam.org/search#tabview=tab1). Novel lincRNAs have also been deposited within the CoGe environment as tracks for genome browsing (Links found in File S2). Evolinc is available as two apps (Evolinc-I and Evolinc-II) in CyVerse's DE (https://de.cyverse.org/de/), for which a tutorial and sample data are available (https://wiki.cyverse.org/wiki/display/TUT/Evolinc+in+the+Discovery+Environment). Evolinc is also available as self-contained Docker images (https://hub.docker.com/r/evolinc/evolinc-i/ and https://hub.docker.com/r/evolinc/evolinc-ii/) for use in a Linux or Mac OSX command-line environment. The code for Evolinc is available to download/edit as a GitHub repository (https://github.com/Evolinc). Information for installation of the Docker image in a command-line environment, as well as FAQs associated with this process are available in the Evolinc GitHub repository readme file. Both Evolinc tools make use of several open source tools, such as BLAST for sequence comparisons (Altschul et al., 1990; Camacho et al., 2009), Cufflinks (Trapnell et al., 2010) for GFF to FASTA conversion, Bedtools (Quinlan and Hall, 2010) for sequence intersect comparisons, MAFFT (Katoh and Standley, 2013) for sequence alignment, RAxML (Stamatakis, 2014) for inferring phylogeny, Notung (Durand et al., 2006) for reconciling gene and species trees, and python, perl, and R for file manipulation and data reporting.

RNA-seq read mapping and transcript assembly

SRA files were uploaded directly into CyVerse DE from (http://www.ncbi.nlm.nih.gov/sra) by using the "Import from URL" option. All further read processing was performed using applications within DE. Briefly, uncompressed paired end reads were trimmed (5 nt from 5' end and 10 nt from 3' end) using FASTX trimmer, whereas single end read files were filtered with the FASTX quality filter so that only reads where ≥ 70% of bases with a minimum quality score of 25 were retained (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Reads were mapped to their corresponding genomes using TopHat2 version 2.0.9 (Kim et al., 2013). TopHat2 settings varied based on organism and SRA, and are listed in File S1. Transcripts were assembled using the Cufflinks2 app version 2.1.1 under settings listed in File S1 (Trapnell et al., 2010). TopHat2 and Cufflinks2 were executed on reads from each SRA file independently.

Validation of lincRNA expression in vivo

RNA was extracted from 2-week old seedlings and flower buds from 4-week old Arabidopsis Col-0 using Trizol (ThermoFisher Life Sciences catalog # 15596018). These tissues and age at extraction most closely matched the experiments from which the RNA-seq data was obtained (Liu et al., 2012). cDNA was synthesized using SuperScript III (ThermoFisher Life Sciences catalog # 18080051) and 2µg of RNA as input. Primers were first validated by performing PCR with genomic DNA as template using GoTaq Green polymerase master mix (Promega catalog #M712) with 95°C for 3' to denature, followed by 35 cycles of 95°C for 15", 55°C for 30" and 72°C for 30" and a final extension step of 5' at 72°C. Primers used are listed in File S2.

Results

An overview of lincRNA identification with Evolinc-I

Evolinc-I validation

After establishing a workflow using the most commonly accepted parameters for defining a lincRNA (detailed in Materials and Methods), we wanted to evaluate its efficiency at distinguishing between unknown or novel protein-coding genes and non-coding loci. For this, we used a random set of 5,000 protein-coding transcripts selected from the TAIR10 annotation to determine Evolinc-I's false discovery rate (FDR) (i.e., protein-coding transcripts erroneously classified as lincRNAs). ORFs for this test dataset ranged in length from 303 to 4182 nts, with an average ORF of 1131 nts (File S3). Because Evolinc is designed to automatically remove transcripts that map back to known genes, we removed these 5,000 genes from the reference genome annotation file, and then generated a transcript assembly file from RNA-seq data where these 5,000 genes were known to be expressed. We fed the transcript assembly file to Evolinc-I. Out of 5,000 protein-coding genes, only 11 were categorized as non-coding by Evolinc-I (0.22% FDR; File S3). Further investigation of the 11 loci revealed that they were predominantly low coverage transcripts with ORFs capable of producing polypeptides greater than 90, but less than 100 amino acids (aa). Moreover, low read coverage for these transcripts led to incomplete transcript assembly. Together these factors were responsible for the miss-annotation of these loci as non-coding. Importantly, our results indicate that read depth and transcript assembly settings impact lincRNA identification, a finding also noted by Cabilli et al. (2011). Therefore, exploring transcript assembly parameters may be necessary prior to running Evolinc-I. In sum, Evolinc-I has a low FDR that can be further reduced by increasing read per base coverage thresholds during transcript assembly as performed in Cabilli et al. (2011).

We determined the overlap of Evolinc predicted lincRNAs with previously published datasets from humans and Arabidopsis, following as closely as possible the methods published for each dataset. We first used Evolinc-I to identify lincRNAs from an RNA-seq dataset generated by Liu et al. (2012) in Arabidopsis (File S1). From nearly one billion reads generated from four different tissues (siliques, flowers, leaves, and roots), Liu et al. (2012) identified 278 lincRNAs (based on the TAIR9 reference genome annotation). Using the Liu et al. (2012) SRA data, we mapped RNA-seq reads and assembled transcripts with Tophat2 and Cufflinks2 in the DE. From these transcripts, Evolinc-I, identified 571 lincRNAs. We then reconciled the lincRNAs identified in Liu et al. (Liu-lincRNAs) with those from Evolinc-I (Evolinc-lincRNAs), by identifying overlapping genomic coordinates for lincRNAs from the two datasets using the Bedtools suite (Quinlan and Hall, 2010). Of the 278 Liu-lincRNAs, 261 were also recovered by Evolinc-I (Table S1). Cufflinks failed to assemble the 17 unrecovered Liu-lincRNAs, due to low coverage, and thus differences in recovery for these loci reflect differences in the Cufflinks parameters employed.

The Arabidopsis genome reference has been updated since Liu et al. (2012), from TAIR9 to TAIR10 (Lamesch et al., 2012). We also ran Evolinc-I with the TAIR10 annotation and found that only 198 of the 261 Liu-lincRNAs were still considered intergenic (Figure 1B). The remaining 63 were reclassified as overlapping a known gene (either sense overlapping transcript, SOT, or antisense overlapping transcript, AOT). This highlights an important aspect of Evolinc-I. While Evolinc-I is able to identify long non-coding RNAs without a genome annotation, genome annotation quality can impact whether an lncRNA is considered intergenic versus AOT or SOT. In sum, 198 of the 571 lincRNAs identified by Evolinc-I correspond to a previously identified Liu-lincRNA (Figure 1B).

Of the 571 lincRNAs identified by Evolinc-I, 373 were not classified as lincRNAs by Liu et al. (2012). Evolinc-I removes transcripts that overlap with the 5' and 3' UTRs of a known gene, whereas Liu et al. (2012) removed transcripts that were within 500 bp of a known gene (Liu et al., 2012). This difference in the operational definition of intergenic space accounts for the omission of 197 Evolinc-lincRNAs from the Liu et al. (2012) lincRNA catalog. In addition, Evolinc-I removes transcripts with high similarity to transposable elements, but not tandem di- or trinucleotide repeats. We could see no biological reason for excluding these simple repeat containing transcripts, and in fact, transcripts with simple tandem repeats have been attributed to disease phenotypes and therefore might be of particular interest (Usdin, 2008). The inclusion of these transcripts accounts for 106 of the unique Evolinc-lincRNAs.

Finally, 70 of the 571 Evolinc-lincRNAs were entirely novel, and did not correspond to any known Liu-lincRNA or gene within the TAIR10 genome annotation. To determine whether these represented *bona fide* transcripts, we tested expression of a subset (*n* = 20) of single and multi-exon putative lincRNAs by RT-PCR using RNA extracted from two different tissues (seedlings and flowers, Figure S1A). We considered expression to be positive if we recovered a band in two different tissues or in the same tissue but from different biological replicates. We recovered evidence of expression for 18 of these putative lincRNAs out of 20 tested. Based on these data we conclude that a majority of the 70 novel lincRNAs identified by Evolinc-I for Arabidopsis are likely to reflect *bona fide* transcripts, and thus valid lincRNA candidates.

We next compared Evolinc-I against a well-annotated set of human lincRNAs characterized by Cabili et al. (2011). Cabili et al. (2011) used RNA-seq data from 24 different tissues and cell types, along with multiple selection criteria to identify a "gold standard" reference set of 4,662 lincRNAs. We assembled transcripts from RNA-seq data for seven of these tissues (File S1) using Cufflinks under the assembly parameters and read-per-base coverage cut-offs of Cabili et al. (2011) (see Materials and Methods). We then fed these transcripts to Evolinc-I. To directly compare Evolinc-I identified lincRNAs with the Cabili et al. (2011) reference dataset (Cabili-lincRNAs), we used the BED files generated by Evolinc-I to identify a subset of 360 multi-exon putative lincRNAs that were observed in at least two tissues (consistent with criteria employed in Cabili et al. [2011] when using a single transcript assembler). We then asked whether these 360 Evolinc-I lincRNAs were found in either the Cabili-lincRNAs, or the hg19 human reference annotation (UCSC). A total of 317 (88%) of the Evolinc-I lincRNAs matched known lincRNAs from the two annotation sources (Figure 1C). The remaining 43 transcripts (12% of the 360 tested) passed all other criteria laid out by Cabili et. al. (2011) and therefore may be *bona fide* lincRNAs, but will require further testing.

Evolution of lincRNA loci with Evolinc-II

Evolinc-II validation

Evolinc-II is an automated and improved version of a workflow we previously used to determine the depth to which Liu-lincRNAs (Liu et al., 2012) were conserved in other species of the Brassicaceae (A., D., L., Nelson et al., 2016). The Evolinc-II workflow is outlined in Figure 2A. While most Liu-lincRNAs were restricted to Arabidopsis, or shared only by Arabidopsis and *A. lyrata*, 3% were conserved across the family, indicating that the lincRNA-encoding locus was present in the common ancestor of all Brassicaceae ~54 MYA (Beilstein et al., 2010). We used Evolinc-II to recapitulate our previous analysis in three ways. First, to provide replicates for statistical analysis, we randomly divided the 5,361 Liu-lincRNAs into 200-sequence groups prior to Evolinc-II analysis (*n* = 27; Figure 2B and Figure S2B). Second, we performed a separate comparison by dividing the Liu-lincRNAs based upon chromosomal location (*n* = 5). Lastly, we used Evolinc-II to search for sequence homologs using the complete Liu-lincRNA dataset but querying with varying E-value cutoffs (E-20, E-15, E-10, E-05, and E-01). This analysis allowed us to test the impact of the requirement for reciprocity on the recovery of putative homologs under different E-value criteria (Figure 2B and Figure S2D). The number of sequence homologs increased for each decrement in BLAST stringency (Figure S2D), indicating that a significant number of putative homologs fulfill the reciprocity requirement even as sequence similarity decreases. The percentage of sequence homologs retrieved by Evolinc-II was statistically indistinguishable for lincRNAs assigned to groups, chromosomes, or the average from all E-value cutoffs (Figure 2B and Figure S2C). Thus, Evolinc-II is a robust method to identify sets of lincRNAs that are conserved across a user-defined set of species, such as the Brassicaceae.

In addition to identifying sets of conserved lincRNAs, Evolinc-II also highlights conserved regions within each query lincRNA. To demonstrate these features, we scanned through the Liu-lincRNA Evolinc-II summary statistics file (at 1E-10; File S4) to identify a conserved lincRNA. At1NC023160 is conserved as a single copy locus in eight of the ten species we examined. It was identified by Liu et al. (2012) based on both RNA-seq and tiling array data, as well as validated by Evolinc-I. During the comparative analyses, Evolinc-II generates a query-centric coordinate file that allows the user to visualize within a genome browser (e.g., JBrowse; [Buels et al., 2016]) what regions of the query lincRNA are most conserved. Using this query-centric coordinate file, we examined the 332 nt At1NC023160 locus in the CoGe genome browser and determined that the 3' end was most highly conserved (Figure 2C). We used the MAFFT multiple sequence alignment generated by Evolinc-II for At1NC023160 to perform structure prediction with RNAalifold (Figure S3A; (Lorenz et al., 2011)). The structural prediction based on the multiple sequence alignment had a greater base pair probability score and lower minimum free energy than the structure inferred from the Arabidopsis lincRNA alone (Figure S3B and S3C). Conserved regions of a lincRNA serve as potential targets for disruption via genome editing techniques, thereby facilitating its functional dissection.

Using Evolinc-II to infer the evolution of the human telomerase RNA locus TERC

In addition to exploring the evolutionary history of a lincRNA catalog, Evolinc-II is an effective tool to infer the evolution of individual lincRNA loci. To showcase the insights Evolinc-II can provide for datasets comprised of a small number of lincRNAs, we focused on the well-characterized human lincRNA, TERC. TERC is the RNA subunit of the ribonucleoprotein complex telomerase that is essential for chromosome end maintenance in stem cells, germ-line cells, and single-cell eukaryotes (Theimer and Feigon, 2006; Zhang et al., 2011; Blackburn and Collins, 2011). TERC is functionally conserved across almost all eukarya, but is highly sequence divergent. Building on work performed by Chen et al. (2000) we used Evolinc-II to examine the evolutionary history of the human TERC locus in 26 mammalian species that last shared a common ancestor between 100-130 MYA (Figure 3) (Glazko, 2003; Arnason et al., 2008).

Evolinc-II identified a human TERC sequence homolog in 23 of the 26 species examined (Figure 3; raw output shown in Figure S4). We were unable to identify a human TERC homolog in *Ornithoryhnchus anatinus* (platypus), representing the earliest diverging lineage within class Mammalia, using our search criteria. In addition, *Mus musculus* (mouse) and *Rattus norvegicus* (rat) were also lacking a human TERC homolog. However, close relatives of mouse and rat, such as *Ictidomys tridecemlineatus* (squirrel) and *Oryctolagus cuniculus* (rabbit) retained clear human TERC sequence homologs, suggesting that loss of the human TERC-like locus is restricted to the Muridae (mouse/rat family). This is in agreement with the previous identification of the mouse TERC, which exhibits much lower sequence similarity with the human TERC than do other mammals (Chen et al, 2000). All identified human TERC homologs also share synteny, suggesting similar evolutionary origins for this locus throughout mammals (Figure 3). Evolinc-II also identified lineage-specific duplication events for the human TERC-like locus in the orangutan, lemur, and galago genomes (Figure 3), similar to previous observations in pig and cow (Chen et al., 2000). In sum, Evolinc-II can be applied to both large and small datasets to uncover patterns of duplication, loss, and conservation across large phylogenetic distances.

Discussion

Rapid identification of lincRNAs using Evolinc-I

With Evolinc-I our goal was to develop an automated and simple pipeline for rapid lincRNA discovery from RNA-seq data. In addition to identification, Evolinc-I generates output files that put downstream analyses and data visualization into the hands of biologists, making it simpler for researchers to discover and explore lincRNAs. Evolinc-I makes use of standard lincRNA discovery criteria, and packages each step into easy-to-use applications within the CyVerse DE or for command-line use via a Docker image with all dependencies pre-installed. We recommend the DE-version of Evolinc-I for novice users, whereas the command-line version of Evolinc-I is useful for knowledgeable users wishing to tweak parameters to fit their system or question. By using Evolinc-I within the DE, the user can take advantage of the cyberinfrastructure support of CyVerse (Merchant et al., 2016). One of the key advantages of combining Evolinc-I with cyberinfrastructure such as the CyVerse's DE is the ability to combine various applications together in one streamlined workflow, and making the workflow easier to implement by interested researchers. For instance, a user can download an RNA-seq SRA file into their DE account, quickly process and map reads, assemble transcripts, and execute Evolinc-I. All of this occurs within the DE without downloading a single file or installing a program on a desktop computer.

We demonstrated the ability of Evolinc-I to identify lincRNAs from previously curated catalogs for plants and mammals. Note that we were able to account for all differences between results from Evolinc-I and the published studies, indicating that our pipeline is operating under definitions and filters currently used by the community. Moreover, because we have formalized the process by which annotations of genome data can be incorporated into the search strategy, Evolinc-I gives researchers the ability to easily explore the contributions of TEs, repetitive elements, or other user defined features to the prediction of lincRNA loci. Finally, we stress that this tool permits experiments to be repeated by researchers to compare the contribution of recently released annotations, or to repeat experiments from other groups. This latter point cannot be overemphasized as interest in lincRNAs grows.

Examining evolutionary history and patterns of conservation of lincRNA loci using Evolinc-II

Evolinc-II is designed to perform a series of comparative genomic and transcriptomic analyses across an evolutionary timescale of the user's choosing and on any number (1-1000s) of query lincRNAs. Similar to the lncRNA discovery and evolutionary analysis tool Slncky (J., Chen et al., 2016), the analyses performed by Evolinc-II highlight conserved lincRNA loci, conserved regions within those loci, and overlap with transcripts in other species. To develop an informative evolutionary profile, we recommend users incorporate as many genomes as possible for closely related species and then choose more distantly related species based on the level of genome annotation, genome quality, and quantity of lncRNAs identified for those species. The computationally intensive nature of these analyses is ameliorated by taking advantage of a high-performance computing cluster such as CyVerse. While sequence conservation is certainly not the only filtering mechanism to identify functional lncRNAs, we believe that is a critical first step. In the future, as more becomes known about structural conservation within lncRNAs, this aspect of lncRNA evolution will be added as an additional filter. We envision Evolinc-II being useful for both scientists attempting to identify functional regions of a lincRNA as well as those wanting to understand the pressures impacting lincRNA evolution.

In addition to highlighting large-scale lincRNA patterns of conservation, we also demonstrated how Evolinc-II can be used to examine the detailed evolutionary history of a single lincRNA, using the human TERC as a test-case. We performed an Evolinc-II analysis with human TERC on 26 genomes in the class Mammalia, 14 of which had not been included in previous studies (Chen et al., 2000). As expected, we recovered a human TERC-like locus in most mammals, as well as three previously unrecorded lineage-specific duplication events. Whether these duplicate TERC loci are expressed and interact with telomerase is unknown; if so they may represent potential regulatory molecules, similar to TER2 in Arabidopsis (Xu et al., 2015; A., D., L., Nelson and Shippen, 2015). We also determined that the human TERC-like locus was lost (or experienced an accelerated mutation rate relative to other mammals) in the common ancestor of mouse and rat. The conservation of the TERC locus across mammals, characterized by rare evolutionary transitions such as that in mouse and rat, stands in stark contrast to the evolution of the telomerase RNA in Brassicaceae (Beilstein et al., 2012), despite the fact that other telomere components are highly conserved (Nelson et al., 2014). Interestingly, mammalian TERCs appear to evolve more slowly than their plant counterparts, similar to the protein components of telomerase (Wyatt et al., 2010). These discoveries highlight the novel insights that can be uncovered using Evolinc-II on even well studied lincRNAs.

In summary, Evolinc streamlines lincRNA identification and evolutionary analysis. Given the wealth of RNA-seq data being uploaded on a daily basis to NCBI's SRA, and the increased availability of high performance computing resources, we believe that Evolinc will prove to be tremendously useful. Combining these resources, Evolinc can uncover broad and fine-scale patterns in the way that lincRNAs evolve and ultimately help in linking lincRNAs to their function.

Acknowledgements

We thank Evan Forsythe (University of Arizona) and Dr. Molly Megraw (Oregon State University) for thoughtful comments pertaining to Evolinc parameters. We thank the PaBeBaMo discussion group at the University of Arizona, and in particular Drs. David Baltrus, Rebecca Mosher, and Ravi Palanivelu, for critical comments on this work. Particular thanks to Dr. Yazhou Chen at the John Innes Center and Ali Reda Ali at Middle Tennessee State University for beta testing of the Evolinc apps. We would also like to thank the CoGe group at the University of Arizona. We are very grateful for the cloud computing resources at CyVerse, and in particular the CyVerse DE group for their support in establishing Evolinc as a set of apps in the discovery environment. This work was supported by National Science Foundation Plant Genome Research Program Grant IOS-1444490 to M.A.B. and E.L. and NSF-MCB #1409251 to MAB.

Figure Legends

Figure 1. Schematic representation of the Evolinc-I workflow and validation. (A) Evolinc-I takes assembled transcripts as input and then filters over several steps (1-4). Evolinc generates output files detailed in the materials and methods. **(B)** Evolinc validation on RNA-seq data from Liu et al. (2012). Four tissues were sequenced by Liu et al., as indicated by the red circles, including (from top to bottom) flowers, siliques, leaves, and roots. Assembled transcripts were fed through Evolinc-I to identify lincRNAs, Antisense Overlapping Transcripts (AOTs), and Sense Overlapping Transcripts (SOTs). A reconciliation was performed between the Evolinc-I identified lincRNAs and the Liu et al. dataset. Gene associated transcriptional unit (GATU) and repeat containing transcriptional unit (RCTU) terminology comes from Liu et al. (2012). **(C)** Evolinc validation of Cabili et al. (2011) RNA-seq data. RNA-seq data was assembled and then filtered through additional Cabili-specific parameters (shown in box). The pie chart shows Evolinc-identified lincRNAs that correspond to Cabili et al. or are novel.

Figure 2. Schematic representation of the Evolinc-II workflow and validation of Liu-lincRNA and Evolinc-identified lincRNAs. (A) Evolinc-II uses lincRNAs as a query in reciprocal BLAST analyses against any number of genomes. Sequences that match the filters (see Materials and Methods) are grouped into families of sequences based on the query lincRNA. Each sequence homolog is classified using user-defined

data or annotations, such as expression or overlap with known gene or lincRNA. Sequences are aligned to highlight conserved regions and to infer phylogeny. These steps can be performed on thousands to tens of thousands of query lincRNAs. Gene trees are inferred for each sequence family using RAxML. The resulting trees are reconciled with the known species tree using Notung 2.0. Notung delineates gene loss and duplication events by marking the output tree with a D (duplication) and blue branch, or L (loss) and red branch. Phylogenetic inference is computationally intensive, and thus we suggest limiting the number of sequence families for which the analysis is performed. Data files generated by Evolinc-II are described in the Materials and Methods. **(B)** Validation of Evolinc-II by repeating the Liu-lincRNA dataset in three different ways. The ~5400 Liu-lincRNAs were randomly divided into 200 sequence bins (blue bar), each bin was run through Evolinc-II (total number of runs = 27), and then the results were averaged, with standard deviation denoted. In the second analysis, the Liu-lincRNAs were divided based on chromosome, and then each set of Liu-lincRNAs (five groups) were run through Evolinc-II separately. Lastly, all Liu-lincRNAs were run through Evolinc using different BLAST E-value cutoffs (E-1, -5, -10, -15, -20), and the results averaged. Bars represent the percent of Liu-lincRNAs for which sequence homologs were identified. *A. tha = Arabidopsis thaliana*, *A. lyr = Arabidopsis lyrata*, *C. rub = Capsella rubella*, *L. ala = Leavenworthia alabamica*, *B. rap = Brassica rapa*, *B. ole = Brassica oleracea*, *S. par = Schrenkiella parvula*, *E. sal = Eutrema salsugineum*, *A. ara = Aethionema arabicum*, and *T. has = Tarenaya hassleriana*. **(C)** Genome browser visualization of the At1NC023160 locus and its conservation in other Brassicaceae. Regions of the Arabidopsis locus that Evolinc-II identified to be conserved are shown in green, with species of origin listed to the right. The blue bar indicates the length of the locus in Arabidopsis, with the arrow indicating direction of transcription. The region of the locus selected for structural prediction is shown in the red dashed box.

Figure 3. Evolinc-II analysis of the human TERC locus in mammals. Species tree of 26 species within class Mammalia with duplication (D) or loss (L) events hung on the tree (left). A micro-synteny profile is shown to the right for each species, showing the TERC locus in red, and adjacent protein-coding genes in black. Direction of each gene is indicated with arrows. The mouse and rat TERC loci are indicated by blue arrows to represent the poor sequence similarity between these two loci and human TERC. Divergence times are approximate and extracted from Arnason et al. (2008). A key is shown below, with gene names indicated. All pertinent links are shown below to regenerate micro-synteny analyses with CoGe (genomeevolution.org) for all species on the tree.

File S1 List of publically available genome and sequence files used, as well as conditions and results from TopHat and Cufflinks for each assembly.

File S2 Evolinc-I output for all species from which lincRNAs were identified, as well as bed files for genome browser viewing, and primers used in RT-PCR verification of transcription of novel Arabidopsis lincRNAs. Also contains CoGe genome browser links to the novel lincRNAs identified.

File S3 False-positive testing of Evolinc-I with Arabidopsis protein-coding genes.

Figure S1 RT-PCR validation of lincRNAs identified in Arabidopsis by Evolinc-I. LincRNA IDs match those found in File S2. G = genomic DNA positive control. F = flower cDNA, S = seedling cDNA.

Figure S2 Examining conservation of Liu-lincRNAs in multiple ways with Evolinc-II. **(A)** Example of the type of bar graph produced by Evolinc-II, in this case for the Liu-lincRNAs at 1E-20. **(B)** Bar graph of level of lincRNA conservation observed when dividing the Liu-lincRNAs into 27 random bins of 200 lincRNAs each. Standard deviation is based on the difference seen between the 27 bins. **(C)** Bar graph depicting the level of lincRNA conservation seen when dividing the Liu-lincRNAs by Arabidopsis chromosome (E-cutoff value of 1E-20). **(D)** Bar graph demonstrating the level of conservation of the Liu-lincRNAs throughout Brassicaceae at different E-cutoff values.

Figure S3 Using At1NC023160 to highlight the structural information that can be gleaned from Evolinc-II. (A) Multiple sequence alignment, generated by MAFFT and visualized within Geneious v7.1 (Kearse et al., 2012). Similar sequences are highlighted, with the consensus sequence shown on top. Nucleotide identity is shown below the consensus sequence, with green representing 100% identity across all sequences. **(B)** RNAalifold (Lorenz et al., 2011) consensus secondary structure prediction based on multiple sequence alignment in **(A)**. Base-pair probabilities are shown, with red being more probable and blue least probable. **(C)** RNAfold structure prediction based on the same region as in **(B)**, but limited to just the Arabidopsis sequence. Base-pair probabilities are shown as in **(B)**.

Figure S4 Raw phylogenetic output from Evolinc-II for TERC. **(A)** A gene tree for the TERC sequence homologs identified in each of the species shown. Sequences without

"TBH" indicate paralogs. **(B)** Notung (Durand et al., 2006) reconciliation of the gene tree shown in (A) to the known species tree. Duplication (red "D") and loss events (grey "LOST") are shown. Support for duplication or loss events are indicated by the green numbers at the nodes that represent the predicted origin of those events.

Table S1 Percent similarity between transcripts identified following transcript assembly and lincRNA identification.

References:

Altschul, Stephen F., Warren Gish., Webb Miller., Eugene W. Myers., and David J. Lipman. 1990. Basic Local Alignment Search Tool. *Journal of Molecular Biology* 215 (3). Academic Press: 403–10. doi:10.1016/S0022-2836(05)80360-2.

Anders, Simon., and Wolfgang Huber. 2010. Differential Expression Analysis for Sequence Count Data. *Genome Biology* 11 (10). BioMed Central: R106. doi:10.1186/gb-2010-11-10-r106.

Arnason, Ulfur., Joseph A Adegoke., Anette Gullberg., Eric H Harley., Axel Janke., and Morgan Kullberg. 2008. Mitogenomic Relationships of Placental Mammals and Molecular Estimates of Their Divergences. *Gene* 421 (1–2): 37–51. doi:10.1016/j.gene.2008.05.024.

Beilstein, Mark a., Amy E. Brinegar., and Dorothy E. Shippen. 2012. Evolution of the Arabidopsis Telomerase RNA. *Frontiers in Genetics* 3 (SEP). Frontiers: 1–8.

doi:10.3389/fgene.2012.00188.

Beilstein, Mark A., Nathalie S Nagalingum., Mark D Clements., Steven R Manchester., and Sarah Mathews. 2010. Dated Molecular Phylogenies Indicate a Miocene Origin for Arabidopsis Thaliana. *Proceedings of the National Academy of Sciences of the United States of America* 107 (43): 18724–28. doi:10.1073/pnas.0909766107.

Blackburn, Elizabeth H., and Kathleen Collins. 2011. Telomerase: An RNP Enzyme Synthesizes DNA. *Cold Spring Harbor Perspectives in Biology* 3 (5): a003558-. doi:10.1101/cshperspect.a003558.

Buels, Robert., Eric Yao., Colin M. Diesh., Richard D. Hayes., Monica Munoz-Torres., Gregg Helt., et al. 2016. JBrowse: A Dynamic Web Platform for Genome Visualization and Analysis. *Genome Biology* 17 (1). BioMed Central: 66. doi:10.1186/s13059-016-0924-1.

Cabili, Moran N., Cole Trapnell., Loyal Goff., Magdalena Koziol., Barbara Tazon-Vega., Aviv Regev., et al. 2011. Integrative Annotation of Human Large Intergenic Noncoding RNAs Reveals Global Properties and Specific Subclasses. *Genes & Development* 25 (18). Cold Spring Harbor Laboratory Press: 1915–27. doi:10.1101/gad.17446611.

Camacho, C., G Coulouris., V Avagyan., N Ma., J Papadopoulos., K Bealer., et al. 2009. BLAST+: Architecture and Applications. *BMC Bioinformatics* 10: 421. doi:10.1186/1471-2105-10-421.

Chen, Jenny., Alexander A. Shishkin., Xiaopeng Zhu., Sabah Kadri., Itay Maza., Mitchell Guttman., et al. 2016. Evolutionary Analysis across Mammals Reveals

Distinct Classes of Long Non-Coding RNAs. *Genome Biology* 17 (1). BioMed Central: 19. doi:10.1186/s13059-016-0880-9.

Chen, Jiunn-Liang., Maria A Blasco., and Carol W Greider. 2000. Secondary Structure of Vertebrate Telomerase RNA. *Cell* 100 (5): 503–14. doi:10.1016/S0092-8674(00)80687-X.

Derrien, Thomas., Rory Johnson., Giovanni Bussotti., Andrea Tanzer., Sarah Djebali., Hagen Tilgner., et al. 2012. The GENCODE v7 Catalog of Human Long Noncoding RNAs: Analysis of Their Gene Structure, Evolution, and Expression. *Genome Research* 22 (9): 1775–89. doi:10.1101/gr.132159.111.

Durand, Dannie., Bjarni V Halldórsson., and Benjamin Vernot. 2006. A Hybrid Micro-Macroevolutionary Approach to Gene Tree Reconstruction. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology* 13 (2). Mary Ann Liebert, Inc. 2 Madison Avenue Larchmont, NY 10538 USA: 320–35. doi:10.1089/cmb.2006.13.320.

Glazko, G. V. 2003. Estimation of Divergence Times for Major Lineages of Primate Species. *Molecular Biology and Evolution* 20 (3): 424–34. doi:10.1093/molbev/msg050.

Guttman, Mitchell., Ido Amit., Manuel Garber., Courtney French., Michael F. Lin., David Feldser., et al. 2009. Chromatin Signature Reveals over a Thousand Highly Conserved Large Non-Coding RNAs in Mammals. *Nature* 458 (7235). Nature Publishing Group: 223–27. doi:10.1038/nature07672.

Hangauer, Matthew J., Ian W. Vaughn., and Michael T. McManus. 2013. Pervasive

Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *PLoS Genetics* 9 (6). doi:10.1371/journal.pgen.1003569.

Hawkes, Emily J., Scott P. Hennelly., Irina V. Novikova., Judith A. Irwin., Caroline Dean., Karissa Y. Sanbonmatsu., et al. 2016. COOLAIR Antisense RNAs Form Evolutionarily Conserved Elaborate Secondary Structures. *Cell Reports* 16 (12). Elsevier: 3087–96. doi:10.1016/j.celrep.2016.08.045.

Johnson, Rory., and Roderic Guigó. 2014. The RIDL Hypothesis: Transposable Elements as Functional Domains of Long Noncoding RNAs. *RNA (New York, N.Y.)* 20 (7): 959–76. doi:10.1261/rna.044560.114.

Kapusta, Aurélie., Zev Kronenberg., Vincent J Lynch., Xiaoyu Zhuo., LeeAnn Ramsay., Guillaume Bourque., et al. 2013. Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. *PLoS Genetics* 9 (4). Public Library of Science: e1003470. doi:10.1371/journal.pgen.1003470.

Katoh, Kazutaka., and Daron M Standley. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* 30 (4): 772–80. doi:10.1093/molbev/mst010.

Kearse, Matthew., Richard Moir., Amy Wilson., Steven Stones-Havas., Matthew Cheung., Shane Sturrock., et al. 2012. Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data. *Bioinformatics (Oxford, England)* 28 (12): 1647–49.

doi:10.1093/bioinformatics/bts199.

Kim, Daehwan., Geo Pertea., Cole Trapnell., Harold Pimentel., Ryan Kelley., Steven L Salzberg., et al. 2013. TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions. *Genome Biology* 14 (4). BioMed Central: R36. doi:10.1186/gb-2013-14-4-r36.

Lamesch, Philippe., Tanya Z Berardini., Donghui Li., David Swarbreck., Christopher Wilks., Rajkumar Sasidharan., et al. 2012. The Arabidopsis Information Resource (TAIR): Improved Gene Annotation and New Tools. *Nucleic Acids Research* 40 (Database issue). Oxford University Press: D1202-10. doi:10.1093/nar/gkr1090.

Li, Lin., Steven R Eichten., Rena Shimizu., Katherine Petsch., Cheng-Ting Yeh., Wei Wu., et al. 2014. Genome-Wide Discovery and Characterization of Maize Long Non-Coding RNAs. *Genome Biology* 15 (2): R40. doi:10.1186/gb-2014-15-2-r40.

Liu, Jun., Choonkyun Jung., Jun Xu., Huan Wang., Shulin Deng., Lucia Bernad., et al. 2012. Genome-Wide Analysis Uncovers Regulation of Long Intergenic Noncoding RNAs in Arabidopsis. *The Plant Cell* 24 (11): 4333–45. doi:10.1105/tpc.112.102855.

Lorenz, Ronny., Stephan H Bernhart., Christian Höner Zu Siederdissen., Hakim Tafer., Christoph Flamm., Peter F Stadler., et al. 2011. ViennaRNA Package 2.0. *Algorithms for Molecular Biology : AMB* 6 (1): 26. doi:10.1186/1748-7188-6-26.

Lyons, Eric., Matthew Bomhoff., Fan Li., and Brian D. Gregory. 2014. EPIC-CoGe: Functional and Diversity Comparative Genomics. https://pag.confex.com/pag/xxii/webprogram/Paper10615.html.

Merchant, Nirav., Eric Lyons., Stephen Goff., Matthew Vaughn., Doreen Ware., David Micklos., et al. 2016. The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLOS Biology* 14 (1). Public Library of Science: e1002342. doi:10.1371/journal.pbio.1002342.

Migeon, Barbara R., Ethan Kazi., Camille Haisley-Royster., Jie Hu., Roger Reeves., Linda Call., et al. 1999. Human X Inactivation Center Induces Random X Chromosome Inactivation in Male Transgenic Mice. *Genomics* 59 (2): 113–21. doi:10.1006/geno.1999.5861.

Mohammadin, Setareh., Patrick P Edger., J Chris Pires., and Michael Eric Schranz. 2015. Positionally-Conserved but Sequence-Diverged: Identification of Long Non-Coding RNAs in the Brassicaceae and Cleomaceae. *BMC Plant Biology* 15 (1): 217. doi:10.1186/s12870-015-0603-5.

Necsulea, Anamaria., Magali Soumillon., Maria Warnefors., Angélica Liechti., Tasman Daish., Ulrich Zeller., et al. 2014. The Evolution of lncRNA Repertoires and Expression Patterns in Tetrapods. *Nature* 505 (7485). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 635–40. doi:10.1038/nature12943.

Nelson, Andrew D. L., and Dorothy E. Shippen. 2015. Evolution of TERT-Interacting lncRNAs: Expanding the Regulatory Landscape of Telomerase. *Frontiers in Genetics* 6 (September): 1–6. doi:10.3389/fgene.2015.00277.

Nelson, Andrew D L A.D.L., Evan S. E.S. Forsythe., Xiangchao Gan., Miltos Tsiantis., and M.A. Mark a. Beilstein. 2014. Extending the Model of Arabidopsis Telomere

Length and Composition across Brassicaceae. *Chromosome Research* 22 (2): 153–66. doi:10.1007/s10577-014-9423-y.

Nelson, Andrew D L., Evan S Forsythe., Upendra K Devisetty., David S Clausen., Asher K Haug-Batzell., Ari M R Meldrum., et al. 2016. A Genomic Analysis of Factors Driving lincRNA Diversification: Lessons from Plants. *G3&amp;#58; Genes|Genomes|Genetics* 6 (September). Genetics Society of America: 2881–91. doi:10.1534/g3.116.030338.

Pertea, Mihaela., Daehwan Kim., Geo M Pertea., Jeffrey T Leek., and Steven L Salzberg. 2016. Transcript-Level Expression Analysis of RNA-Seq Experiments with HISAT, StringTie and Ballgown. *Nat Protocols* 11 (9): 1650–67. doi:10.1038/nprot.2016.095.

Quinlan, Aaron R., and Ira M Hall. 2010. BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. *Bioinformatics (Oxford, England)* 26 (6). Oxford University Press: 841–42. doi:10.1093/bioinformatics/btq033.

Quinn, Jeffrey J., Qiangfeng C Zhang., Plamen Georgiev., Ibrahim A Ilik., Asifa Akhtar., and Howard Y Chang. 2016. Rapid Evolutionary Turnover Underlies Conserved lncRNA-Genome Interactions. *Genes & Development* 30 (2). Cold Spring Harbor Laboratory Press: 191–207. doi:10.1101/gad.272187.115.

Robinson, Mark D., and Alicia Oshlack. 2010. A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data. *Genome Biology* 11 (3). BioMed Central: R25. doi:10.1186/gb-2010-11-3-r25.

Spannagl, Manuel., Thomas Nussbaumer., Kai C. Bader., Mihaela M. Martis., Michael

Seidel., Karl G. Kugler., et al. 2016. PGSB plantsDB: Updates to the Database Framework for Comparative Plant Genome Research. *Nucleic Acids Research* 44 (D1): D1141–47. doi:10.1093/nar/gkv1130.

Stamatakis, Alexandros. 2014. RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics (Oxford, England)* 30 (9): 1312–13. doi:10.1093/bioinformatics/btu033.

Tang, Haibao., and Eric Lyons. 2012. Unleashing the Genome of Brassica Rapa. *Frontiers in Plant Science* 3 (January). Frontiers: 172. doi:10.3389/fpls.2012.00172.

Theimer, Carla A., and Juli Feigon. 2006. Structure and Function of Telomerase RNA. *Current Opinion in Structural Biology* 16 (3): 307–18. doi:10.1016/j.sbi.2006.05.005.

Trapnell, Cole., Brian A Williams., Geo Pertea., Ali Mortazavi., Gordon Kwan., Marijke J van Baren., et al. 2010. Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation. *Nature Biotechnology* 28 (5). Nature Publishing Group: 511–15. doi:10.1038/nbt.1621.

Ulitsky, Igor. 2016. Evolution to the Rescue: Using Comparative Genomics to Understand Long Non-Coding RNAs. *Nature Reviews Genetics* 17 (10). Nature Research: 601–14. doi:10.1038/nrg.2016.85.

Ulitsky, Igor., and David P Bartel. 2013. Ulitsky, Igor, and David P Bartel. 2013. “lincRNAs: Genomics, Evolution, and Mechanisms.” Cell 154 (1): 26–46. doi:10.1016/j.cell.2013.06.020.lincRNAs: Genomics, Evolution, and Mechanisms. *Cell* 154 (1): 26–46. doi:10.1016/j.cell.2013.06.020.

Ulitsky, Igor., Alena Shkumatava., Calvin H Jan., Hazel Sive., and David P Bartel. 2011. Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution. *Cell* 147 (7): 1537–50. doi:10.1016/j.cell.2011.11.055.

Usdin, Karen. 2008. The Biological Effects of Simple Tandem Repeats: Lessons from the Repeat Expansion Diseases. *Genome Research* 18 (7). Cold Spring Harbor Laboratory Press: 1011–19. doi:10.1101/gr.070409.107.

Wang, Dong., Zhipeng Qu., Lan Yang., Qingzhu Zhang., Zhi-Hong Liu., Trung Do., et al. 2017. Transposable Elements (TEs) Contribute to Stress-Related Long Intergenic Noncoding RNAs in Plants. *The Plant Journal*, January. doi:10.1111/tpj.13481.

Wang, Huan., Qi-Wen Niu., Hui-Wen Wu., Jun Liu., Jian Ye., Niu Yu., et al. 2015. Analysis of Non-Coding Transcriptome in Rice and Maize Uncovers Roles of Conserved lncRNAs Associated with Agriculture Traits. *The Plant Journal* 84 (2): 404–16. doi:10.1111/tpj.13018.

Wang, Kevin C., and Howard Y Chang. 2011. Molecular Mechanisms of Long Noncoding RNAs. *Molecular Cell* 43 (6): 904–14. doi:10.1016/j.molcel.2011.08.018.

Washietl, Stefan., Sven Findeiss., Stephan A Müller., Stefan Kalkhof., Martin von Bergen., Ivo L Hofacker., et al. 2011. RNAcode: Robust Discrimination of Coding and Noncoding Regions in Comparative Sequence Data. *RNA (New York, N.Y.)* 17 (4): 578–94. doi:10.1261/rna.2536111.

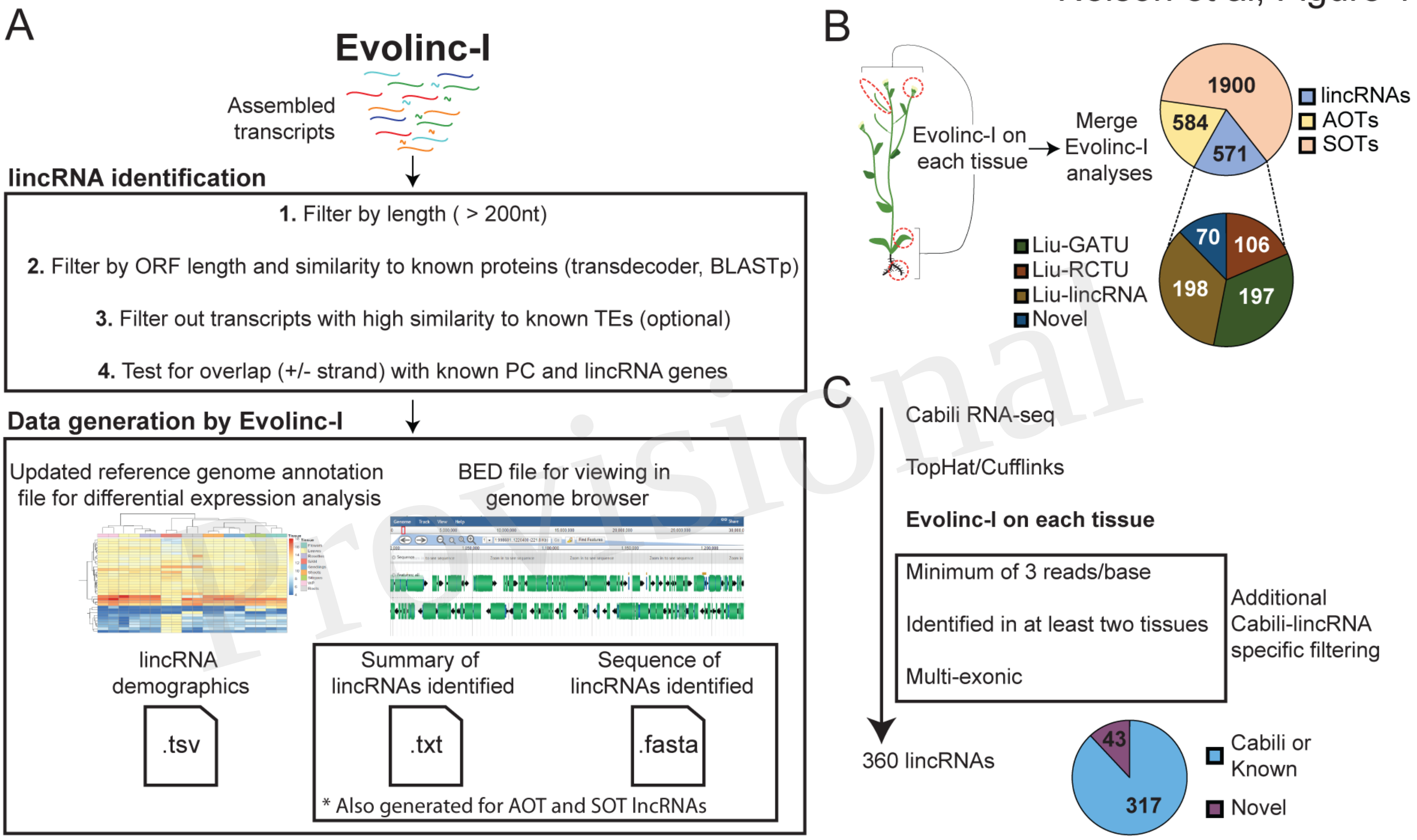
Wyatt, Haley D M., Stephen C West., and Tara L Beattie. 2010. InTERTpreting Telomerase Structure and Function. *Nucleic Acids Research* 38 (17). Oxford University Press: 5609–22. doi:10.1093/nar/gkq370.

Xu, Hengyi;, Andrew D L; A.D.L. Nelson., and Dorothy E; D.E. Shippen. 2015. A Transposable Element within the Non-Canonical Telomerase RNA of Arabidopsis Thaliana Modulates Telomerase Activity in Response to DNA Damage. *PLoS Genetics* In press (6). Public Library of Science: e1005281. doi:10.1371/journal.pgen.1005281.

Zhang, Qi., Nak-Kyoon Kim., and Juli Feigon. 2011. Architecture of Human Telomerase RNA. *Proceedings of the National Academy of Sciences of the United States of America* 108 (51): 20325–32. doi:10.1073/pnas.1100279108.

Zhang, Yu-Chan., Jian-You Liao., Ze-Yuan Li., Yang Yu., Jin-Ping Zhang., Quan-Feng Li., et al. 2014. Genome-Wide Screening and Functional Analysis Identify a Large Number of Long Noncoding RNAs Involved in the Sexual Reproduction of Rice. *Genome Biology* 15 (12). BioMed Central: 512. doi:10.1186/s13059-014-0512-1.

Nelson et al, Figure 1


A
Evolinc-I
Assembled transcripts
lincRNA identification
1. Filter by length (> 200nt)
2. Filter by ORF length and similarity to known proteins (transdecoder, BLASTp)
3. Filter out transcripts with high similarity to known TEs (optional)
4. Test for overlap (+/- strand) with known PC and lincRNA genes
Data generation by Evolinc-I
Updated reference genome annotation file for differential expression analysis
BED file for viewing in genome browser
lincRNA demographics
.tsv
Summary of lincRNAs identified
.txt
Sequence of lincRNAs identified
.fasta
* Also generated for AOT and SOT lncRNAs
B
Evolinc-I on each tissue
Merge Evolinc-I analyses
1900
584
571
lincRNAs
AOTs
SOTs
Liu-GATU
Liu-RCTU
Liu-lincRNA
Novel
70
106
198
197
C
Cabili RNA-seq
TopHat/Cufflinks
Evolinc-I on each tissue
Minimum of 3 reads/base
Identified in at least two tissues
Multi-exonic
Additional Cabili-lincRNA specific filtering
360 lincRNAs
43
317
Cabili or Known
Novel

Nelson et al, Figure 2

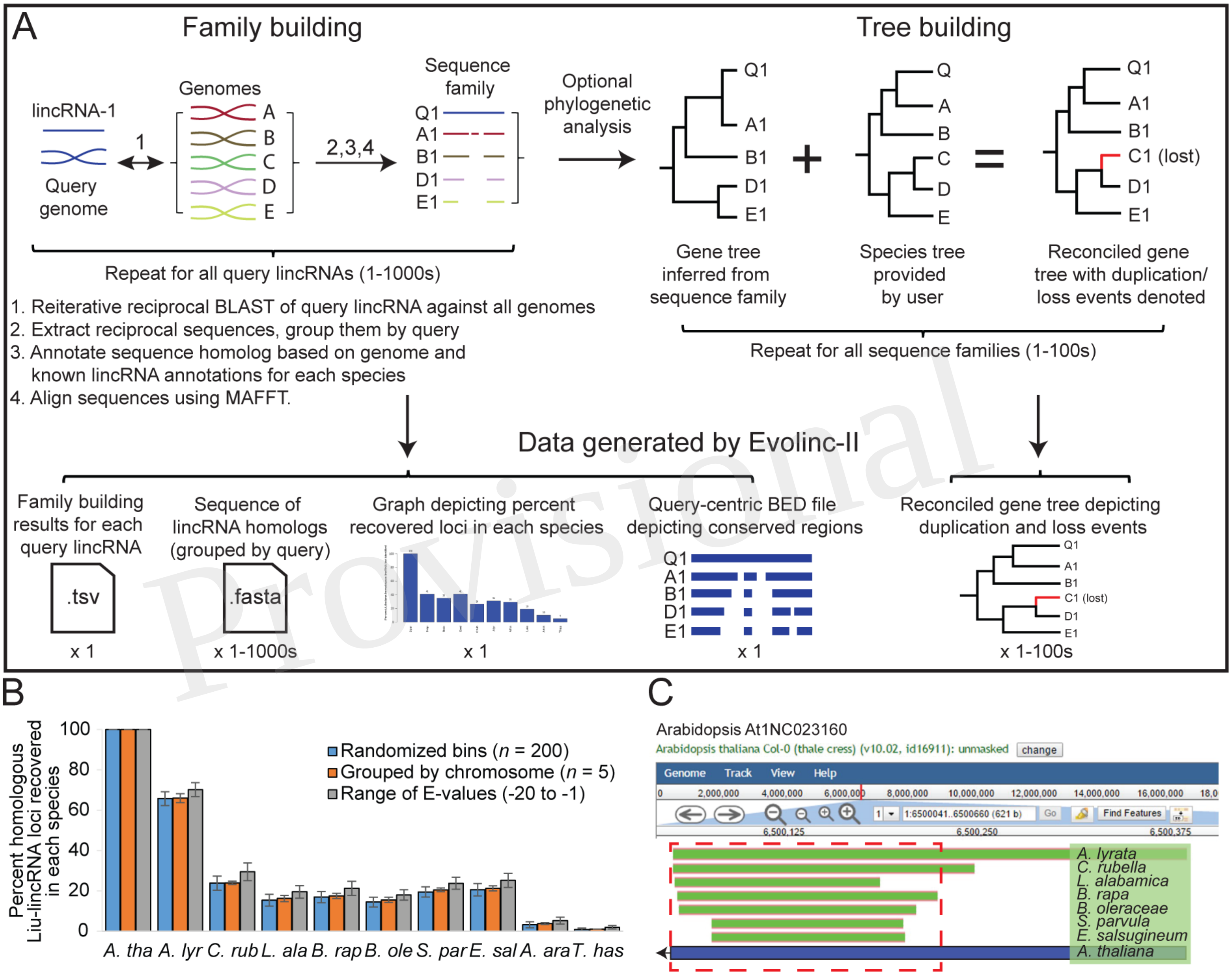
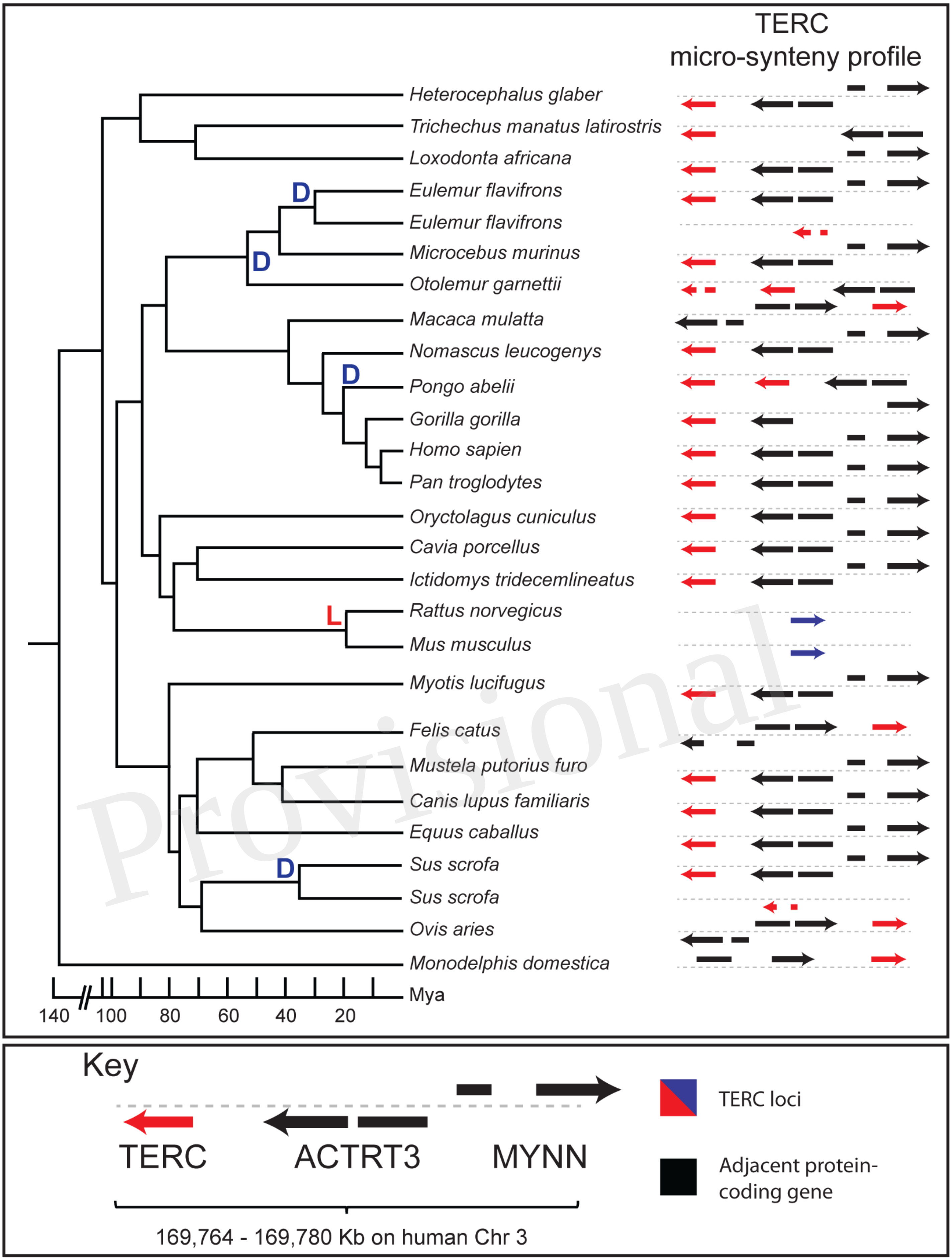

A
Family building
Tree building
lincRNA-1
Query
genome
1
Genomes
A
B
C
D
E
2,3,4
Sequence
family
Q1
A1
B1
D1
E1
Optional
phylogenetic
analysis
Q1
A1
B1
D1
E1
Gene tree
inferred from
sequence family
+
Q
A
B
C
D
E
Species tree
provided
by user
=
Q1
A1
B1
C1 (lost)
D1
E1
Reconciled gene
tree with duplication/
loss events denoted
Repeat for all query lincRNAs (1-1000s)
1. Reiterative reciprocal BLAST of query lincRNA against all genomes
2. Extract reciprocal sequences, group them by query
3. Annotate sequence homolog based on genome and known lincRNA annotations for each species
4. Align sequences using MAFFT.
Repeat for all sequence families (1-100s)
Data generated by Evolinc-II
Family building results for each query lincRNA
.tsv
x 1
Sequence of lincRNA homologs (grouped by query)
.fasta
x 1-1000s
Graph depicting percent recovered loci in each species
x 1
Query-centric BED file depicting conserved regions
Q1
A1
B1
D1
E1
x 1
Reconciled gene tree depicting duplication and loss events
Q1
A1
B1
C1 (lost)
D1
E1
x 1-100s
B
Percent homologous Liu-lincRNA loci recovered in each species
100
80
60
40
20
0
Randomized bins (n = 200)
Grouped by chromosome (n = 5)
Range of E-values (-20 to -1)
A. tha
A. lyr
C. rub
L. ala
B. rap
B. ole
S. par
E. sal
A. ara
T. has
C
Arabidopsis At1NC023160
Arabidopsis thaliana Col-0 (thale cress) (v10.02, id16911): unmasked
change
Genome
Track
View
Help
0
2,000,000
4,000,000
6,000,000
8,000,000
10,000,000
12,000,000
14,000,000
16,000,000
18,0
1
1:6500041..6500660 (621 b)
Go
Find Features
6,500,125
6,500,250
6,500,375
A. lyrata
C. rubella
L. alabamica
B. rapa
B. oleraceae
S. parvula
E. salsugineum
A. thaliana

Figure 03.TIF

Nelson et al, Figure 3



To regenerate the micro-synteny analyses
https://genomevolution.org/r/lxvp
https://genomevolution.org/r/lxvo
https://genomevolution.org/r/lxvn
https://genomevolution.org/r/lxz6