Genomics 114 (2022) 110434

Contents lists available at ScienceDirect

Genomics

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

ExoPLOT: Representation of alternative splicing in human tissues and developmental stages with transposed element (TE) involvement

Fengjun Zhang [a,*], Carsten Alexander Raabe [a], Margarida Cardoso-Moreira [b,1], Jürgen Brosius [a,c], Henrik Kaessmann [b], Jürgen Schmitz [a,d,*]

[a] Institute of Experimental Pathology, ZMBE, University of Münster, 48149 Münster, Germany
[b] Center for Molecular Biology of Heidelberg University, ZMBH, 69120 Heidelberg, Germany
[c] Institutes for Systems Genetics, Frontiers Science Center for Disease-related Molecular Network, West China Hospital, Sichuan University, 610000 Chengdu, Sichuan, China
[d] EvoPAD-RTG, University of Münster, 48149 Münster, Germany

ARTICLE INFO

Keywords:
ExoPLOT shiny app
Alternative splicing
Human organ transcriptomes
Human developmental stage transcriptomes
CCDS
Exonization
Transposed elements

ABSTRACT

Advances in RNA high-throughput sequencing and large-scale functional assays yield new insights into the multifaceted activities of transposed elements (TE) and many other previously undiscovered sequence elements. Currently, no tool for easy access, analysis, quantification, and visualization of alternatively spliced exons across multiple tissues or developmental stages is available. Also, analysis pipelines demand computational skills or hardware requirements, which often are hard to meet by wet-lab scientists. We developed ExoPLOT to enable simplified access to massive RNA high throughput sequencing datasets to facilitate the analysis of alternative splicing across many biological samples. To demonstrate the functonality of ExoPLOT, we analyzed the contributon of exonized TEs to human coding sequences (CDS). mRNA splice variants containing the TE-derived exon were quantified and compared to expression levels of TE-free splice variants. For analysis, we utilized 313 human cerebrum, cerebellum, heart, kidney, liver, ovary, and testis transcriptomes, representing various pre- and postnatal developmental stages. ExoPLOT visualizes the relative expression levels of alternative transcripts, e.g., caused by the insertion of new TE-derived exons, across different developmental stages of and among multiple tissues. This tool also provides a unique link between evolution and function during exonization (gain of a new exon) and exaptation (recruitment/co-optation) of a new exon. As input for analysis, we derived a database of 1151 repeat-masked, exonized TEs, representing all prominent families of transposons in the human genome and the collection of human consensus coding sequences (CCDS). ExoPLOT screened preprocessed RNA high-throughput sequencing datasets from seven human tissues to quantify and visualize the dynamics in RNA splicing for these 1151 TE-derived exons during the entire human organ development. In addition, we successfully mapped and analyzed 993 recently described exonized sequences from the human frontal cortex onto these 313 transcriptome libraries. ExoPLOT's approach to preprocessing RNA deep sequencing datasets facilitates alternative splicing analysis and significantly reduces processing times. In addition, ExoPLOT's design allows studying alternative RNA isoforms other than TE-derived in a customized – coordinate-based manner and is available at http://retrogenomics3.uni-muenster.de:3838/exz-plot-d/.

1. Introduction

In 1978 Walter Gilbert came up with the most sensible explanation based on evolutionary thoughts, namely that the arrangement of genes in introns and exons favors the evolution of novel gene variants "without destroying the old" and "one product's intron becomes another's exon" concluding that "introns are both frozen remnants of history and the sites of future evolution" [1]. The recruitment of parts of transposed elements support Gilbert's predictions and has been termed exonization [2,3]. TEs are jumping genetic elements that sporadically spread in waves through genomes. After coincidentally inserting into intronic, UTR (untranslated region), or intergenic regions (whereby 60% of TEs in

* Corresponding authors at: Institute of Experimental Pathology, ZMBE, University of Münster, 48149 Münster, Germany.
E-mail addresses: fzhang@uni-muenster.de (F. Zhang), jueschm@uni-muenster.de (J. Schmitz).
[1] Present Address: [Margarida Cardoso-Moreira], Evolutionary Developmental Biology Laboratory, The Francis Crick Institute, London NW1 1AT, UK.

https://doi.org/10.1016/j.ygeno.2022.110434
Received 16 May 2022; Received in revised form 4 July 2022; Accepted 14 July 2022
Available online 18 July 2022
0888-7543/© 2022 Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

genomes of humans and mice are located in introns [4]), the resulting TE sequences are generally not subject to natural selection [2]. However, over time, accumulating mutations can generate favorable splice sites that ultimately lead to novel exons [5,6]. This process of exonization provides an alternative route to structural and functional gene diversification, initially via the alternative inclusion of novel exons, while the original function is maintained by the ancestral variants.

In the beginning, most novel alternative exons contribute only little to gene expression and frequently do not persist [7]. Achieving functional relevance is often accompanied by higher expression levels mediated via the acquisition of selectively favorable splicing-competent mutations. Equal alternative expression or even dominant or constitutive inclusion of TE cassettes are the hallmarks of a successful path towards a novel, exapted protein-coding function. One of the first analyses of tissue-specific expression of exonized *Alu* SINEs was presented by Lin et al. [8], whereby the authors investigated genome-wide exon arrays and conducted RT-PCR experiments. Sorek and colleagues [6] estimated that in humans, *Alu*s already contribute ~5% of alternatively spliced mRNAs (messenger RNAs). Furthermore, thousands of TE-containing exons within 5' UTRs were linked to increased tissue-specific gene expression [9]. On the tail end of the gene, Tajnik et al. [10] showed that human intergenic exonized *Alu* elements can form new exons and alternative 3' UTRs, some of which successfully competed with the original splice form of the host gene and now contribute to tissue-specific regulation. The influence of exonization on gene integrity was demonstrated for modulated splicing efficiency, methylome plasticity, and DNA damage regulation. Furthermore, exonization seems to be actively suppressed in hematological cancer [11], potentially indicating the important regulatory function of exonized isoforms in healthy cells. Recently, Florea et al. [12] identified 45 tissue-specific exonizations from the human frontal cortex, pointing to a function in tissue-dependent variation. It should be mentioned that the exonization of TEs is only one easily traceable prominent pathway leading to alternative splicing. In addition, the inclusion of anonymous random sequences in protein-coding CDS provides another source of variation, which, however, is certainly not as easily detectable and verifiable.

The impact of transposed elements on mammalian development has been reviewed by Garcia-Perez et al. [13]. The authors describe how active and inactive TEs influence developing organisms, in particular via their modifying influence on host gene regulation. Less is known about the immediate functional contribution of TEs to protein-coding variants during different phases of organogenesis.

Cardoso-Moreira et al. [14] reported 313 human transcriptomes from the brain (i.e., cerebrum, cerebellum), heart, kidney, liver, ovary, and testis representing the major stages of pre- and postnatal organ development. The samples were collected during different, physiologically relevant phases of organogenesis in embryos and fetuses (4 to 20 weeks post-conception), and during postnatal organ growth in infant, child, adolescent, and adult stages (1 to 63 years). This unique data set provides the basis to comparatively analyze the variations of gene expression during organ development. We developed ExoPLOT to visualize these variations in alternative transcript processing (e.g., triggered by newly evolved TE exons) over different stages of development and among multiple organs and to establish the unique connection between evolution and function during exonization (gain of a new exon) and exaptation (recruitment or co-option) of a new exon.

2. Results and discussion

The ExoPLOT tool combines two fundamental datasets: the genomic coordinates of 313 human RNA-seq libraries [14], and the coordinates of 1151 human genome-wide exonized TEs sorted by TE families and gene names derived from human genome and CCDS databank coordinates [15]. Both coordinate systems are based on the human GRCh37 (hg19) genome assembly and LiftOver GRCh38 (hg38) coordinates (customized approach), which can be automatically interconverted into each other via the integrated UCSC LiftOver toolkit. Additional 993 ExoPLOT-embedded frontal cortex tissue exonizations [12] offer a detailed analysis of brain-specific TE exons and are also integrated in the ExoPLOT exonization database. Further data will be embedded as soon they appear. The ExoPLOT "Customized" approach also permits the flexible and comparable analysis of any alternatively spliced exon of interest. The result of each search request also includes links to other transcriptome and genome databases such as VastDB (https://vastdb.crg.eu/wiki/Main_Page) and UCSC (https://genome.ucsc.edu/). The following are selected examples to demonstrate the usage of ExoPLOT for well-known TE-exonizations.

2.1. *ADARB1*

To demonstrate ExoPLOT's functions, we feature the well-characterized *Alu*Jb SINE exonization of the RNA editing gene *ADARB1* (ENSG00000197381 [8], Fig. 1A). RNA auto-editing of the primary transcript of the RNA editing enzyme led to alternative splicing and consequently *Alu*Jb SINE exonization. The exonized variant of *ADARB1* includes 39 additional TE-encoded amino acids in the catalytical deaminase domain (Fig. 1A). The exonization event results in the suppression of RNA editing via negative autoregulation [16]. Screening for TE expression by querying *ADARB1* chr21:46,604,388-46,604,508 (gene name followed by chromosomal location and chromosomal coordinates at hg19), the ExoPLOT graph reveals increased expressions of the exonized splice form (darker lines in Fig. 1A bottom) compared to the ancestral one in all organs. Fig. 1A displays increased expression and potential translation of both splice variants in heart postnatal developmental stages and the inclusion level of the exonized variant for 2–4-year-old children, which is approximately ten times higher than that of other developmental stages.

2.2. *SUGT1*

An additional well-defined example of an *Alu*Sx SINE exonization is the suppressor of the G2 allele of *SKP1* (*SUGT1*), a cell cycle regulator that evolved recently in great apes [8] and whose alternatively spliced, exonized form includes 33 additional amino acids (Fig. 1B). In contrast to *ADARB1*, in human organs, the alternative splice form of *SUGT1* is expressed at lower levels during organ development compared to the ancestral splice variant(s) devoid of the novel *Alu*-derived exon.

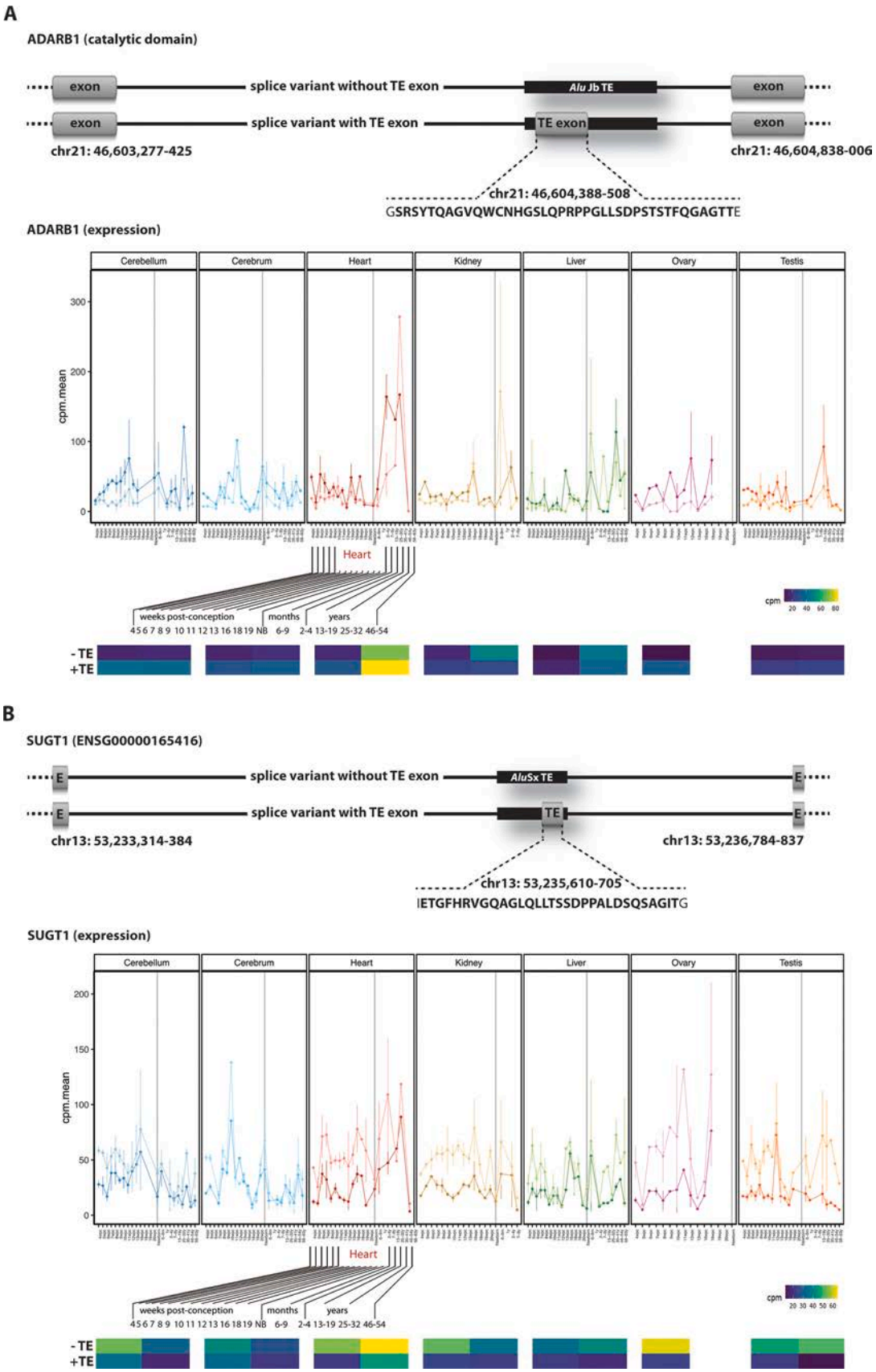
2.3. *Human genome-wide exonization*

The averaged expression profiles in various organs over all developmental stages of the 1151 intron-derived TE-exonized isoforms detected in the CCDS databank are presented in Fig. 2. While different TE families contribute to a different extent to exonization events (see individual TE families below), expression levels of the various exonized isoforms usually were similar among different organs. However, some were also explicitly elevated in specific organs and/or during certain developmental stages. Examples of organ/tissue-specific differentially expressed variant profiles are presented in Supplemental Material 1. For tissue-specificity of TE-expression, we calculated the *Tau* index [17].

The different TE families and their contribution to exonization are presented in the following.

2.3.1. *Alu SINEs*

Owed to their high frequency in the primate genome (>1 million copies), frequent intronic localization [18], and sequence motifs that already resemble internal splice sites [19], *Alu* short interspersed elements (*Alu* SINEs) were the most frequent among all exonized TE classes (329 cases). However, *Alu* SINE exonized isoforms (turquoise in Fig. 2) were generally less expressed than the corresponding ancestral isoforms. *Alu*s represent the youngest TEs in primates and we previously



(caption on next page)

Fig. 1. ExoPLOT functions: *ADARB1* and *SUGT1* genes. A. The exonized sequence of *ADARB1* originated from an intronic *Alu*Jb TE, which integrated in antisense orientation with respect to *ADARB1* gene transcription (chromosome 21). B. In the case of the *SUGT1* gene, the exonized sequence originated from an intronic antisense-oriented *Alu*Sx TE on chromosome 13. Amino acids encoded by exonized sequences are magnified. Flanking amino acids encoded by the ancestral splice variant are displayed in grey. ExoPLOT expression profiles in CPM (counts per million) over different developmental stages and organs are displayed. Dark-colored curves represent the isoform including the exonized sequence; light-colored curves the ancestral isoform(s) devoid of the TE cassette. Often, comparisons are not computed between the TE-exonized mRNA versus a single ancestral state but against two or even many other splice variants devoid of the analyzed TE exon. Vertical dotted lines in each plot divide pre- and postnatal samples. The error bars represent the range of detectable expression levels (CPM) per each data point. Heatmaps below each graph represent the mean number of counts for TE-less (upper line), and TE-containing (lower line) isoforms averaged for pre- and postnatal developmental stages. Wpc = 4–20 weeks post-conception, newborn, 6 m = months to 65 y = years.

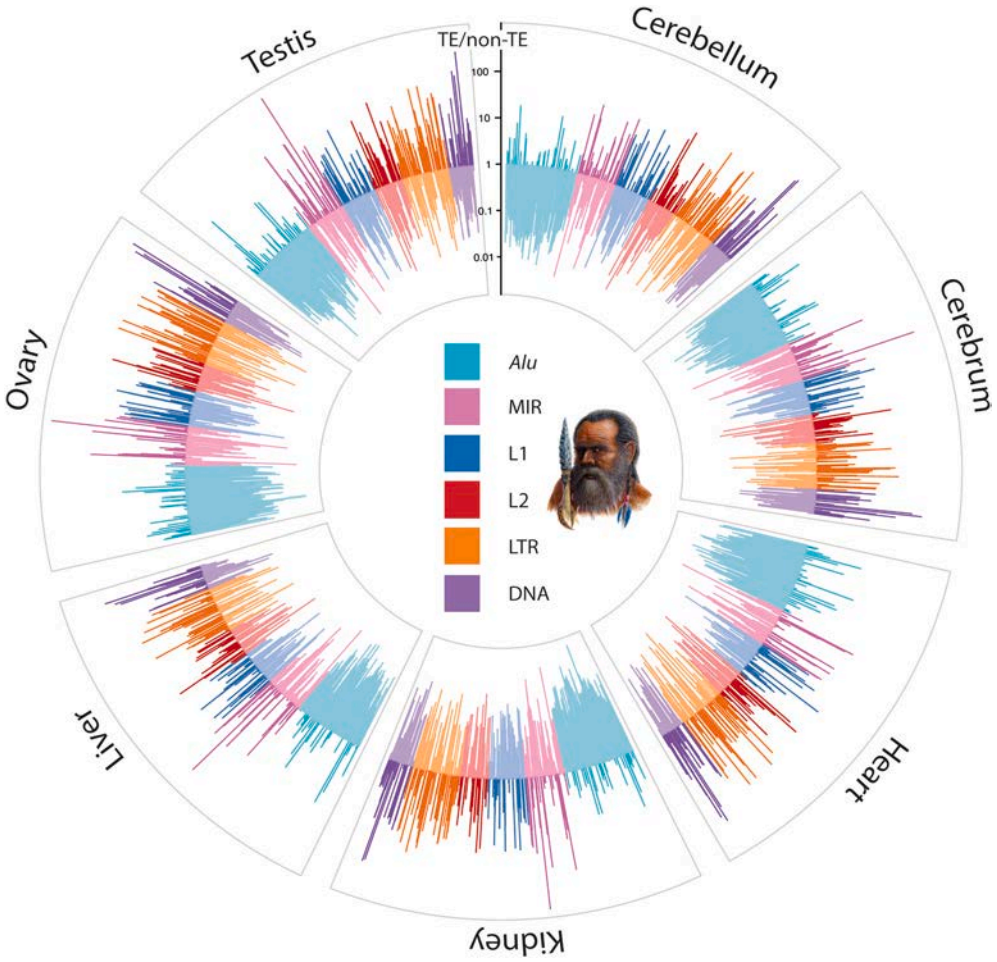


Fig. 2. Relative expression of TE splice variants in cerebellum, cerebrum, heart, kidney, liver, ovary, and testis averaged over all developmental stages. Expression of TE-containing variants are displayed as a logarithm (base = 2) of their fold change plus pseudo count (set to 1) against the variants without TEs (TE/non-TE) from the same locus. Turquoise = *Alu*, pink = MIR, blue = LINE1, red = LINE2, orange = LTR, purple = DNA transposons. The central circle at 1 represents the level of isoforms without TE cassettes. Outward expression (>1) indicates isoforms with TE cassettes that are expressed at higher levels compared to their non-TE variants. Inwardly oriented expression (<1) of isoforms with TE cassettes means a lower expression than the respective non-TE variants. Values are expressed as counts per million (CPM, TE/non-TE; y-labels). All individual data points are presented in Supplementary Material 2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

suggested that the approximately 65 million years of primate evolution has often not been long enough to yield functionally significant or even dominant mRNA isoforms [20].

2.3.2. MIR SINEs

We observed 185 cases of exonized mammalian-wide interspersed repeats (MIR SINEs) among the exonized TEs (pink in Fig. 2). MIR SINEs are not as frequent in the human genome as *Alu* SINEs (<400 thousand MIRs vs. >1 million *Alus* [21]), but their long, mostly pre-mammalian history provided sufficient evolutionary time (>160 MY) to acquire and maintain function, which may explain their higher expression levels in organs compared to the MIR-free isoforms [20].

2.3.3. LINE1 and LINE2 TEs

The still active long interspersed element 1 (LINE1, blue in Fig. 2) and the ancient long interspersed element 2 (LINE2, red in Fig. 2) were underrepresented as exonized domains; we observed only 166 LINE1 (from ~520 thousand genomic insertions [21]) and 135 LINE2 elements (from ~320 thousand genomic insertions [21]) that were exonized. This might be due to their preferential presence in intergenic sequence regions. It is thought that selection acts against the fixation of such autonomous elements, especially in the sense orientation within genes due to the high probability of, e.g., transcriptional interference with the gene regulatory system [22].

2.3.4. LTRs

Approximately 8% of the human genome resemble long terminal repeats (LTR, orange in Fig. 2; ~440 thousand genomic copies [21]), mainly representing fragmented variants. In relative terms, they were frequently found to be exonized (223 cases). Probably due to their long mammalian history, many cases show higher expression levels compared to the splice forms devoid of LTRs. Forty of the exonizations exhibit higher expression of TE-derived alternatives in at least one organ/tissue compared to the ancestral variant(s). Twenty-five of the non-TE variants show higher expression across all organs.

2.3.5. DNA elements

DNA transposons (purple in Fig. 2) with their deep mammalian history were inactivated in the common ancestor of anthropoids [23]. However, despite their low abundance in humans, in relative terms we found many cases of exonized DNA transposons, which we attributed to their long mammalian existence providing sufficient evolutionary time to manifest novel DNA element exons. Nineteen out of 113 exonizations exhibit higher expression levels across all organs, and ten others show higher expression in at least one organ compared to the non-TE variant (s).

Mazin et al. [24] used the Cardoso-Moreira et al. [14] dataset to investigate the developmental dynamics of alternative splicing in cross-species comparisons based on the same RNA-seq dataset included in ExoPLOT. Their analysis of newly emerged cassette exons in organ development revealed that ~40–50% of new species-specific exons overlap with TEs. However, this dataset does not coincide with our 1151 cases of TE exons that generally have a far more ancient origin and are consistent with their cross-species and cross-mammalian lineage exonization history.

ExoPLOT was designed to utilize a range of structured transcriptomic data as a backbone to analyze expression patterns of novel exons throughout the entire development of human organs. The unique ExoPLOT data visualization enabled the analysis of the general distribution of novel exon cassettes and their expression among multiple organs and different developmental stages compared with the expression of the ancestral TE-free splice forms of the same genes. Such extensive transcriptome-level expression data is important for any comparative analysis of TE exonization events, where functional and evolutionary considerations overlap. For evolutionary studies, ExoPLOT visualizes the differential organ- or tissue-specific regulation of alternative splicing as well as the variation of expression during consecutive developmental stages in human life. To broaden the applicability of this approach, we added a strategy ("customized input") to compare not only alternatively spliced TE exonized genes but also any alternatively expressed locus, by allowing the input of user-defined coordinates to study the general impact of alternative splicing of genes.

ExoPLOT uses a reverse approach compared to the conventional analysis of RNA splicing. While most available tools build on user-provided BAM files or spliced alignment/gene models for analysis and data visualization (e.g., [25,26]), ExoPLOT uses preprocessed RNA deep sequencing datasets that can be scanned for alternative splicing. The design of our tool reduces time requirements and workload for the end-user. Standard desktop or laptop computers are more than sufficient to analyze big data for splicing events of interest. Also, because simple lists of genomic coordinates represent the actual input for ExoPLOT, even scientists with little or no experience in bioinformatics can conduct even complex analyses with our tool, which we imagine to constitute a "swiss army knife" for wet-lab scientists. Unlike many of the existing genome browsers, ExoPLOT not only provides the display of alternatively spliced exons, but also offers accurate quantification of the exon inclusion levels (see also [26,27]). The hyperlinking option to the UCSC genome browser enables the user to comprehensively analyze the splicing event in the context of huge genomic datasets. ExoPLOT fills in an existing gap in available tools to investigate alternative splicing. In the future, ExoPLOT datasets continue to provide access to a constantly growing collection of RNA high-throughput data for further analysis.

3. Conclusion

ExoPLOT enables the specific comparison of expression levels of different splice variants. We used the tool to detect 1151 exonized TE sequences detected in 313 transcriptome libraries to differentiate organ- and developmental-specific expression patterns in humans. In addition to these 1151 exonizations, ExoPLOT can also be applied to any coordinate-based analysis of interest to compare, for example, any other kind of alternative splicing or the expression of otherwise alternatively regulated gene variants. Notably, the currently employed 313 transcriptome libraries can be easily expanded by incorporating new transcriptome data and additional species. ExoPLOT can therefore be adapted to the specific needs of the user. We are confident that in the future, ExoPLOT will complement other existing tools for the analysis of alternative expression in mammalian organ development, such as those recently presented in Mazin et al. [24].

4. Methods

ExoPLOT, designed as an online tool for the science community, was inspired by the Shiny application of Cardoso-Moreira et al. [14] (https://apps.kaessmannlab.org/evodevoapp/) for exploring alternative expression profiles of genes with exonized protein-coding portions throughout different organs and developmental stages in humans. However, a comparative view and analysis of different splice forms is not possible with the original Shiny application. Because web applications require a fast response to users' requests, we designed an ultrafast system based on GRCh37/hg19-mapped coordinates of 2 billion exon-exon junction reads. ExoPLOT also includes Rtracklayer, an R wrapper for UCSC LiftOver, to retrieve loci from GRCh38/hg38 coordinates. Prescreening the 313 transcriptomes for billions of splice junctions enabled quick and reliable access to alternative TE expression. That enables comparative measurements of the expression level of the currently largest dataset of exonized sequence regions and allows real-time comparisons, for example, of the expression levels of alternatively spliced TE exons and their corresponding TE-free splice variants.

4.1. TE exon detection from the human CCDS databank and UCSC genome annotation

RepeatMasker can recognize (exonized) TEs and parts thereof (> 30 nts) in protein-coding sequence regions [28]. We first screened for TE inclusion in the NCBI Consensus CDS databank (CCDS; https://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi; a reliable set of >35,000 high-quality human protein-coding sequences that were independently identified and verified by different research groups). To detect even smaller fragments of TEs in CCDS data, we also compared genomic positions of TEs using the annotation report of the GRCh37/hg19 human reference genome (https://genome.ucsc.edu/) with the CCDS coordinates. This strategy could detect additional TE exonizations (< 30 nts) in CDS sequences that continue into the neighboring introns. For our test cases, we removed all exonizations shorter than 12 nts to ensure a clear assignment of exonized regions to TEs.

We detected and verified 1151 TE inclusions in this set and sorted them according to their gene names and corresponding genomic coordinates (e.g., *ADARB1* chr21:46604388–46,604,508). All these are now embedded in ExoPLOT and can be used for comparative expression analyses.

4.2. Organs and developmental stages

To work with the Cardoso-Moreira [14] database, we retrieved the 313 human RNA-seq transcriptomes, i.e., the fastq sequence files for each experiment, the mapping results against the human reference genome, and the corresponding metadata, from ArrayExpress, the Archive of Functional Genomics Data (https://www.ebi.ac.uk/arrayexpress/), accession code E-MTAB-6814 (Human RNA-seq time-series of the development of seven major organs).

4.3. Exon-exon junction coordinate comparison

For rapid responses to users' requests for screening of exonized cassettes, we first extracted >2 billion coordinates of cDNA reads that cover all the exon-exon junctions from the 313 transcriptome libraries to produce the pre-established backbone dataset. To reduce the

computational processing time, non-junction reads (i.e., exon internal reads), which were less informative for the differentiation between individual splicing events, were filtered out in this step (but were later included for the estimations of gene counts, see below). Junction reads overlapping with genomic regions of interest were organized into (1) TE exclusion junction-included exons (e.g., Human.Testis.28ypb.Male *ADARB1* | HISEQ:213 | chr21:46603392-46603425 chr21:46604388-46604508 chr21:46604838-46604904, interval cleavages indicating exon boundaries of this spliced form), and (2) TE inclusion junction-included exons (e.g., Human.Testis.28ypb.Male | *ADARB1* | HISEQ:213 | chr21:46604444-46604508 chr21:46604838–46604873, for human testis of a 28-year-old male).

4.4. Library counts and normalization

Cardoso-Moreira et al. [14] used the RPKM (reads per kilobase per million mapped reads) units for their *Evo-devo Mammalian Organs* Shiny app (https://apps.kaessmannlab.org/evodevoapp/), which was not specifically designed for estimating alternatively spliced variants. ExoPLOT, on the other hand, focuses on the analysis of alternatively spliced variants and was built based on exon-exon junction reads instead of fully annotated genes. We, therefore, roughly scaled the counts for exonized vs. non-exonized alternatives in counts per million (CPM) based on normalized library sizes. The minimum number of cDNA reads in one library must be above 10 to be considered, and the corresponding genomic coordinates must overlap with the hg19 Ensembl annotation (a LiftOver for hg38 is implemented, see above). We calculated the efficiency of a specific exon being spliced into the transcript population using the normalized percent spliced in index (PSI) [29], which can be obtained from the junction read counts and combined this with the overall gene counts obtained from traditional annotation counting approaches. The counts of both TE variants and non-TE variants in every sample were computed using normalized PSI and normalized gene counts. Subsequently, ExoPLOT took the average counts of all samples from one organ as the representative for organ/tissue-specific expression. The calculation of raw PSI between exonized TE reads and other alternative splicing events is shown below:

$$PSI_{raw} = \frac{Spliced\ in\ reads\ (reads\ with\ TEs)}{Spliced\ in\ reads + Spliced\ out\ reads\ (reads\ without\ TEs)}$$

The PSI_{raw} values are normalized and used to calculate the counts of the TE exons multiplied by the total number of counts of the gene.

ExoPLOT users are free to select coordinates from our 1151 pre-established TE exonization cases or to directly submit their own customized coordinate set (e.g., for detecting the expression level of interesting alternatively spliced exons; for the *ADARB1* example, the coordinates of a spliced exon is chr21 46,604,388 46,604,508). However, it should be mentioned that the customized coordinates should represent clear boundaries of unique alternative exons. Including multiple exons will lead to inconclusive results. Fast screening speed for diagnostic coordinates was realized by automated coordinate-relevant subscreenings, thus drastically reducing the number of comparisons.

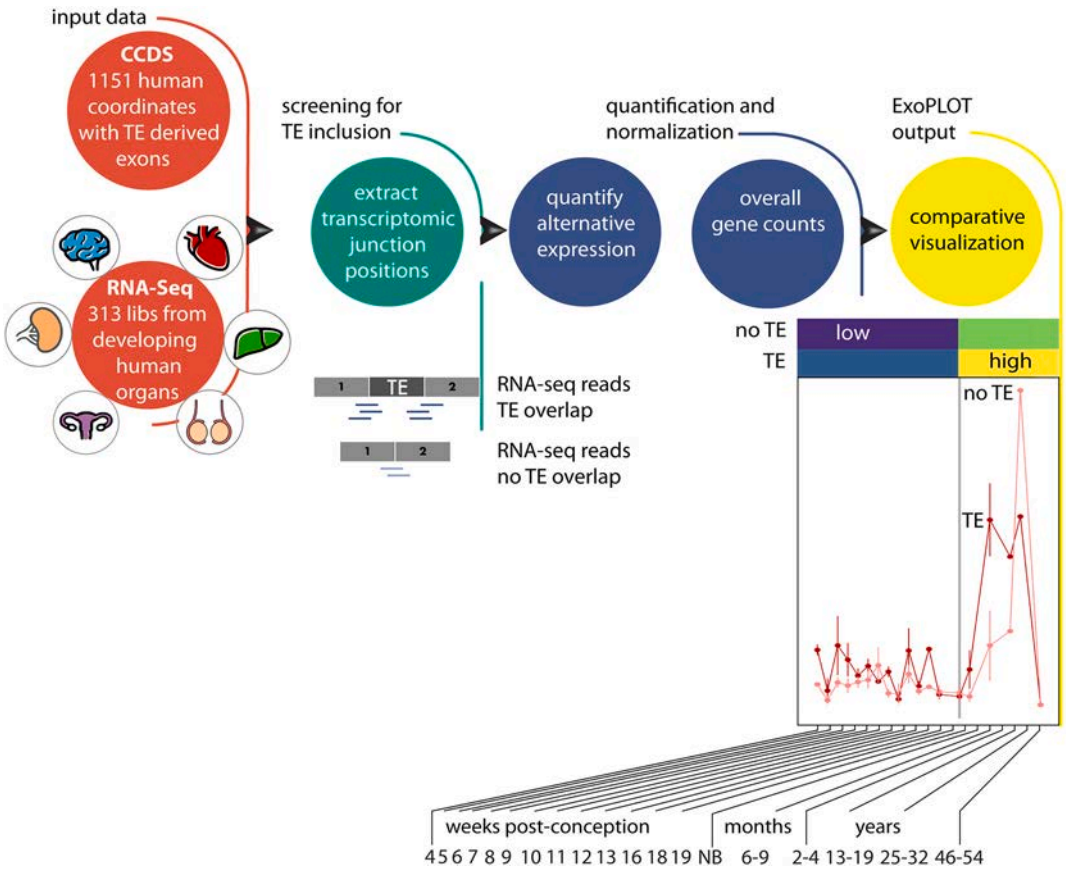


Fig. 3. ExoPLOT: an ultrafast tool to detect differentially expressed exonizations and alternatively spliced variants. ExoPLOT utilizes 313 transcriptomic libraries from various developmental stages of seven different organs: cerebrum, cerebellum, heart, kidney, liver, ovary, and testis (Cardoso-Moreira et al. [14]). Based on CCDS and genomic annotations, we built a pipeline to capture all reads with or without TE exon cassettes, detect alternative expressions, normalize read numbers, and finally visualize the results in a line graph. A heatmap indicates low and high expression levels averaged over pre- and postnatal stages. The genomic coordinates derived from a human genome-wide screening resulting in 1151 exonized TE cassettes are accessible via a drop-down menu. Alternatively, user-defined coordinates can be inserted to examine expression patterns of various other alternatively spliced exon regions.

Each request returns a collection of comparative alternative splices (e.g., exonized TEs sorted by organs and developmental stages). The procedure by which ExoPLOT starts with data from a transcriptome library and eventually compares the expression of alternatively spliced exons is illustrated in Fig. 3.

ExoPLOT is located on a publicly accessible server embedded in a complex data environment. Pre-established exon lists are available via the "Exonization" tab for comparative analysis of exonized TEs from our survey and other sources within different tissues. Software and data are continuously curated and updated. For visualization of our 1151 uncovered exonized TE cassettes, custom track files containing the corresponding genomic regions are embedded, which can be uploaded at the UCSC server to identify the area of interest in a browser window (for details and further information http://www.genome.ucsc.edu/cgi-bin/hgCustom).

Data availability

The tool is available at: http://retrogenomics3.uni-muenster.de:3838/exz-plot-d/. All source codes are uploaded to https://github.com/RetroWWU/ExoPLOT. Supplementary Data are available online.

Author contributions

FZ and JS designed the study. MCM and HK provided early access to their transcriptome data and advised on designing ExoPLOT. FZ developed the ExoPLOT Shiny application, CAR added the customized track for exonized sequences in UCSC. JS, CAR, and FZ wrote the manuscript, and JB contributed helpful comments.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - RTG2220 - project number 281125614 to JS.

Author statement

The authors have no conflicts of interest to declare. The authors declared that the work described has not been published previously, that it was not under consideration for publication elsewhere, that its publication is approved by all authors.

Acknowledgements

We thank Norbert Grundmann for advice in developing the ExoPLOT Shiny App and improving the ultrafast algorithm for screening coordinate-based alternative splice forms. Many thanks go to Marsha Bundman for editorial help.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2022.110434.

References

[1] W. Gilbert, Why genes in pieces? Nature 271 (1978) 501, https://doi.org/10.1038/271501a0.
[2] J. Schmitz, J. Brosius, Exonization of transposed elements: a challenge and opportunity for evolution, Biochimie. 93 (2011) 1928–1934, https://doi.org/10.1016/j.biochi.2011.07.014.
[3] L. Schrader, J. Schmitz, The impact of transposable elements in adaptive evolution, Mol. Ecol. 28 (2019) 1537–1549, https://doi.org/10.1111/mec.14794.
[4] N. Sela, B. Mersch, N. Gal-Mark, G. Lev-Maor, A. Hotz-Wagenblatt, G. Ast, Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome, Genome Biol. 8 (2007) R127, https://doi.org/10.1186/gb-2007-8-6-r127.
[5] R. Sorek, G. Ast, D. Graur, *Alu*-containing exons are alternatively spliced, Genome Res. 12 (2002) 1060–1067, https://doi.org/10.1101/gr.229302.
[6] R. Sorek, G. Lev-Maor, M. Reznik, T. Dagan, F. Belinky, D. Graur, G. Ast, Minimal conditions for exonization of intronic sequences: 5′ splice site formation in *Alu* exons, Mol. Cell 14 (2004) 221–231, https://doi.org/10.1016/S1097-2765(04)00181-9.
[7] M. Krull, J. Brosius, J. Schmitz, Alu-SINE exonization: en route to protein-coding function, Mol. Biol. Evol. 22 (2005) 1702–1711, https://doi.org/10.1093/molbev/msi164.
[8] L. Lin, S. Shen, A. Tye, J.J. Cai, P. Jiang, B.L. Davidson, Y. Xing, Diverse splicing patterns of exonized *Alu* elements in human tissues, PLoS Gen. 4 (2008), e1000225, https://doi.org/10.1371/journal.pgen.1000225.
[9] J.J. Merkin, P. Chen, M.S. Alexis, S.K. Hautaniemi, C.B., Burge, Origins and impacts of new mammalian exons, Cell Rep. 10 (2015) 1992–2005, https://doi.org/10.1016/j.celrep.2015.02.058.
[10] M. Tajnik, A. Vigilante, S. Braun, H. Hänel, N.M. Luscombe, J. Ule, K. Zarnack, J. König, Intergenic *Alu* exonisation facilitates the evolution of tissue-specific transcript ends, Nucleic Acids Res. 43 (2015) 10492–10505, https://doi.org/10.1093/nar/gkv956.
[11] N. Avgan, J.I. Wang, J. Fernandez-Chamorro, R.J. Weatheritt, Multilayered control of exon acquisition permits the emergence of novel forms of regulatory control, Genome Biol. 20 (2019) 141, https://doi.org/10.1186/s13059-019-1757-5.
[12] L. Florea, L. Payer, C. Antonescu, G. Yang, K. Burns, Detection of *Alu* exonization events in human frontal cortex from RNA-seq data, Front. Mol. Biosci. 8 (2021), 727537, https://doi.org/10.3389/fmolb.2021.727537.
[13] J.L. Garcia-Perez, T.J. Widmann, I.R. Adams, The impact of transposable elements on mammalian development, Dev. 143 (2016) 4101–4114, https://doi.org/10.1242/dev.132639.
[14] M. Cardoso-Moreira, J. Halbert, D. Valloton, B. Velten, C. Chen, Y. Shao, A. Liechti, K. Ascenção, C. Rummel, S. Ovchinnikova, P.V. Mazin, I. Xenarios, K. Harshman, M. Mort, D.N. Cooper, C. Sandi, M.J. Soares, P.G. Ferreira, S. Afonso, M. Carneiro, J.M.A. Turner, J.L. VandeBerg, A. Fallahshahroudi, P. Jensen, R. Behr, S. Lisgo, S. Lindsay, P. Khaitovich, W. Huber, J. Baker, S. Anders, E. Yong, Y.E. Zhang, H. Kaessmann, Gene expression across mammalian organ development, Nature. 571 (2019) 505–509, https://doi.org/10.1038/s41586-019-1338-5.
[15] K.D. Pruitt, J. Harrow, R.A. Harte, C. Wallin, M. Diekhans, D.R. Maglott, S. Searle, C.M. Farrell, J.E. Loveland, B.J. Ruef, E. Hart, M.M. Suner, M.J. Landrum, B. Aken, S. Ayling, R. Baertsch, J. Fernandez-Banet, J.L. Cherry, V. Curwen, M. Dicuccio, M. Kellis, J. Lee, M.F. Lin, M. Schuster, A. Shkeda, C. Amid, G. Brown, O. Dukhanina, A. Frankish, J. Hart, B.L. Maidak, J. Mudge, M.R. Murphy, T. Murphy, J. Rajan, B. Rajput, L.D. Riddick, C. Snow, C. Steward, D. Webb, J. A. Weber, L. Wilming, W. Wu, E. Birney, D. Haussler, T. Hubbard, J. Ostell, R. Durbin, D. Lipman, The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes, Genome Res. 19 (2009) 1316–1323, https://doi.org/10.1101/gr.080531.108.
[16] K. Nishikura, A-to-I editing of coding and non-coding RNAs by ADARs, Nat. Rev. Mol. Cell Biol. 17 (2016) 83–96, https://doi.org/10.1038/nrm.2015.4.
[17] I. Yanai, H. Benjamin, M. Shmoish, V. Chalifa-Caspi, M. Shklar, R. Ophir, A. Bar-Even, S. Horn-Saban, M. Safran, E. Domany, D. Lancet, O. Shmueli, Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification, Bioinformatics. 21 (2005) 650–659, https://doi.org/10.1093/bioinformatics/bti042.
[18] P. Deininger, *Alu* elements: know the SINEs, Genome Biol. 12 (2011) 236, https://doi.org/10.1186/gb-2011-12-12-236.
[19] N. Gal-Mark, S. Schwartz, G. Ast, Alternative splicing of Alu exons – two arms are better than one, Nucleic Acids Res. 36 (2008) 2012–2023, https://doi.org/10.1093/nar/gkn024.
[20] M. Krull, M. Petrusma, W. Makalowski, J. Brosius, J. Schmitz, Functional persistence of exonized mammalian-wide interspersed repeat elements (MIRs), Genome Res. 17 (2007) 1139–1145, https://doi.org/10.1101/gr.6320607.
[21] E. Lander, E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, et al., Initial sequencing and analysis of the human genome, Nature. 409 (2001) 860–921, https://doi.org/10.1038/35057062.
[22] P. Medstrand, L.N. van de Lagemaat, D.L. Mager, Retroelement distributions in the human genome: variatons associated with age and proximity to genes, Genome Res. 12 (2002) 1483–1495, https://doi.org/10.1101/gr.388902.
[23] J. Schmitz, A. Noll, C.A. Raabe, G. Churakov, R. Voss, M. Kiefmann, T. Rozhdestvensky, J. Brosius, R. Baertsch, H. Clawson, C. Roos, A. Zimin, P. Minx, M.J. Montague, R.K. Wilson, W.C. Waren, Genome sequence of the basal haplorrhine primate *Tarsius syrichta* reveals unusual insertions, Nat. Commun. 7 (2016) 12997, https://doi.org/10.1038/ncomms12997.
[24] P.V. Mazin, P. Khaitovich, M. Cardoso-Moreira, H. Kaessmann, Alternative splicing during mammalian organ development, Nat. Genet. 53 (2021) 925–934, https://doi.org/10.1038/s41588-021-00851-w.
[25] E.D. Harrington, P. Bork, Sircah: a tool for the detection and visualization of alternative transcripts, Bioinformatics. 24 (2008) 1959–1960, https://doi.org/10.1093/bioinformatics/btn361.
[26] Y. Katz, E.T. Wang, J. Silterra, S. Schwartz, B. Wong, H. Thorvaldsdóttir, J. T. Tobinson, J.P. Mesirov, E.M. Airolgi, C.B. Burge, Quantitative visualization of alternative exon expression from RNA-seq data, Bioinformatics. 31 (2015) 2400–2402, https://doi.org/10.1093/bioinformatics/btv034.

[27] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler, The human genome browser at UCSC, Genome Res. 12 (2002) 996–1006, https://doi.org/10.1101/gr.229102.
[28] S.S. Singer, D.N. Männel, T. Hehlgans, J. Brosius, J. Schmitz, From "Junk" to Gene: *Curriculum vitae* of a primate receptor isoform gene, J. Mol. Biol. 341 (2004) 883–886, https://doi.org/10.1016/j.jmb.2004.06.070.
[29] S. Schafer, K. Miao, C.C. Benson, M. Heinig, S.A. Cook, N. Hubner, Alternative splicing signatures in RNA-seq data: percent spliced in (PSI), Curr. Protoc. Hum. Genet. 87 (2015), https://doi.org/10.1002/0471142905.hg1116s87, 11.161.1-11.16.14.