bioRxiv preprint doi: https://doi.org/10.1101/2022.11.24.517799; this version posted November 24, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fake IDs? Widespread misannotation of DNA Transposons as a General Transcription Factor

Nozhat T. Hassan[1] and David L. Adelson[1]*

[1]School of Biological Sciences, University of Adelaide, Australia

*Corresponding author: E-mail: david.adelson@adelaide.edu.au

Abstract

Transposable elements (TEs) account for a large proportion of genomic content and are major determinants of genome structure and function in eukaryotes. The annotation of TEs is a critical part of our understanding of genomes, however, the accuracy of gene/TE annotation pipelines remains an issue as TEs are frequently underestimated or misannotated. In this paper, we report that the mammalian General Transcription Factor II-I Repeat Domain-Containing Protein 2 (GTF2IRD2) was used to erroneously annotate DNA transposons in a variety of non-mammalian species because the transcription factor contains a 3' fused hAT transposase domain. The consequences of the misannotation of TEs as *trans*-regulatory elements such as transcription factors can lead to errors in phylogenetic trees based on orthologs and to significant wasted time for investigators interested in gene regulation. This study reiterates the importance of structural analysis and effective annotation methods that minimise transitive error propagation when characterising non-mammalian genomes.

Keywords: Transposable Element, Genome, Annotation, Transcription Factor, GTF2, DNA Transposon

Background

Annotation of TEs is essential for understanding genome structure and function; however, misannotation may result in erroneous classification of TEs as *trans*-regulatory elements such as transcription factors. TEs can introduce genetic novelty to the host genome and an example of this is the exaptation of TEs into the human General Transcription Factor II-I Repeat Domain-Containing Protein 2 (GTF2IRD2). GTF2IRD2 contains a Charlie8-like element positioned at the 3' end (C terminus) of the gene model/full protein that has retained transposable element features such as DDE amino acids required for transposition [1]. The Charlie transposon (DNA transposon; hAT superfamily) is an old autonomous group of transposons

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.24.517799; this version posted November 24, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

abundant in mammalian genomes, including humans [2]. Charlie transposons are defined by their target site duplication (TSDs) and terminally inverted repeats (TIRs). At the same time, the protein-coding sequence of the transposase may vary between Charlie elements from different species [3]. GTF2IRD2 is found in mammals and is predicted to be in many reptiles, amphibians and bony fishes [4]. However, upon closer inspection, some non-mammalian GTF2IRD2 sequences appear to be hAT transposons, not transcription factors. Here we have used structural and phylogenetic analysis to resolve the widespread misannotation of non-mammalian DNA transposons as GTF2IRD2 transcription factors. We believe that this paper addresses an important issue; it has implications not only for the study of TEs but is also relevant as similar misannotations could also cause the misinterpretation of other results that depend on reliable gene annotation.

Results and Discussion

While annotating hAT-6 transposons in Testudines genomes, we noticed that GTF2IRD2/2A, a human general transcription factor was the top BLASTN result when using hAT-6 transposons as a query to search non-mammalian genomes. This was unexpected as hAT-6 has hallmarks of a functional transposase, such as TIRs and TSDs, and has the functional motifs required for transposition, such as the DDE/RW residues in the translated open reading frame (ORF) [4,5]. NCBI's eukaryotic gene annotation pipeline prefers to use experimental evidence when annotating genes but uses an *ab initio* model to predict optimal CDS alignments when there is no experimental data (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/). Specifically, Gnomon is meant to exclude gene predictions with high homology to transposable or retro-transposable elements from the final gene models, however, the eukaryotic annotation pipeline appears to lack a final TE filtering step after integrating RefSeq annotations. This may explain how genes such as GTF2IRD2/2A, which contains an integrated Charlie8-like element (a hAT-like transposase), can lead to hAT-6 being predicted to be a transcription factor in non-mammalian genomes [3]. When searching Interpro for GTF2IRD2/2A sequences, we saw that 485 proteins were annotated as GTF2IRD2/2A proteins, but only 4 of these have been reviewed in human, cow and mouse genomes (https://www.ebi.ac.uk/interpro/entry/InterPro/IPR042224/protein/reviewed/#table, Accessed 11th November 2022). To determine if additional hAT transposons were incorrectly annotated as GTF2IRD2/2A, we examined the phylogeny of a set of protein sequences annotated as GTF2IRD2/2A or GTF2IRD2/2A-like from mammals, reptiles, bony fishes, and amphibians (fig. 1). Birds were excluded from

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.24.517799; this version posted November 24, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

analysis as the sequences annotated as GTF2IRD2/2A had no significant similarity to either mammalian GTF2IRD2/2A or to any DNA TE structures. Any sequence homology in birds was limited to the N terminus of the mammalian GTF2IRD2/2A, an indication of potential similarity to the ancestral GTF2IRD2/2A protein prior to the exaptation of Charlie8 (Supplementary Materials).

There was a clear distinction between mammalian and non-mammalian GTF2IRD2/2A gene models. Mammalian GTF2IRD2/2A are correctly annotated as transcription factors as they have an N terminus ~400 aa long containing a GTF2I-Like repeat domain, a zinc finger binding domain, and an integrated Charlie8-like element at the C terminus [1]. However, this is not the case with non-mammalian sequences. Multiple alignment of GTF2IRD2/2A from mammals and non-mammals with hAT-6 transposons exclusively shows high similarity alignment at the 3' end (fig 2). This is consistent with the position of the Charlie-like element in GTF2IRD2/2A, demonstrating that non-mammalian GTF2IRD2/2A are not likely to be TFs, but TEs.

Non-mammalian GTF2IRD2/2A from fish, reptiles, and frogs were manually curated to identify signature motifs of DNA transposons as they appeared closely related to hAT-6 transposons than mammalian GTF2IRD2/2A. 28 sequences were classified as autonomous hAT transposons from several different species, but particularly from the genome of the Atlantic salmon (*Salmo salar*) (Table 1). TSDs and TIRs are characteristic of hAT transposons and were found in most sequences, while sequences without them were classified as partial, non-autonomous transposons. Furthermore, DDE/RW residues required for transposition were identified for most of the newly curated hAT transposons (Supplementary Materials). Finally, the 5' and 3' TIRs were mostly conserved across the misannotated hAT transposons, which further demonstrates the degree of misannotation (fig. 3).

Conclusions

In this study we present a case of widespread incorrect annotation of hAT DNA transposons as GTF2IRD2/2A. This has led to at least 28 or more new instances of hATs from reptiles, bony fishes, and amphibians that had been overlooked and could have been incorrectly used as transcription factors in other analyses. Correct annotation is a vital step in furthering our understanding of genome evolution, and

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.24.517799; this version posted November 24, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

misannotation of TEs as *trans*-regulatory genes such as TFs affects downstream research and can confound phylogenetic analysis.

Methods

Manual curation of hAT transposons from GTF2IRD2/2A sequences

A set of GTF2IRD2/2A and GTF2IRD2/2A-like protein sequences were downloaded from NCBI. The search was limited to species belonging to the *Actinopterygii, Reptilia* and *Amphibia* classes. To determine whether these GTF2IRD2/2A proteins were actually DNA transposons, extensive manual curation was performed to locate characteristic sequence features such as TIRs, TSDs and ORFs. To identify hAT transposons, hAT-6_TSE (in prep) was used as a query in a BLASTP search against a set of mammal, reptile, amphibian, bony fish and bird genomes containing GTF2IRD2/2A gene annotations [6]. The corresponding nucleotide sequence of each top hit was extended 1000 bp in flanking regions where possible and used for manual annotation of TIRs and TSDs characteristic of hAT transposons. ORFs were searched using GENSCAN and searched for DD/E and RW residues. Sequences that contained 5' and 3' TSDs, TIRs, and an intact ORF were classified as autonomous hAT transposons [6,7] (Supplementary Materials). The best match for each new autonomous hAT was found using Repbase and both 5' and 3' TIRs were aligned using MAFFT v7.310 to view conserved nucleotides [8]. hATs misannotated as GTF2IRD2-like were aligned using MAFFT to mammalian GTF2IRD2/2A to confirm they had homology to GTF2IRD2/2A's Charlie8-like domain (Supplementary Materials).

Tree-building

GTF2IRD2/2A and GTF2IRD2/2A-like protein sequences from mammals, reptiles, amphibians, and bony fishes were aligned to the Hermes and hAT-6 transposons using MAFFT v7.319 [9]. The alignment was trimmed using CLipKit and IQTree was used for tree reconstruction with JTT+F+I+G4 as the best-fit model with 20 maximum likelihood trees and 1000 bootstraps [10–13].

Abbreviations

GTF2IRD2: General Transcription Factor II-I Repeat Domain-Containing Protein 2

GTF2IRD2A: General Transcription Factor II-I Repeat Domain-Containing Protein 2A

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.24.517799; this version posted November 24, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgements

We would like to thank Terry Bertozzi and Christine Elsik for their expertise and helpful comments. We would also like to thank our lab members for their ongoing support throughout this study and beyond.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.24.517799; this version posted November 24, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

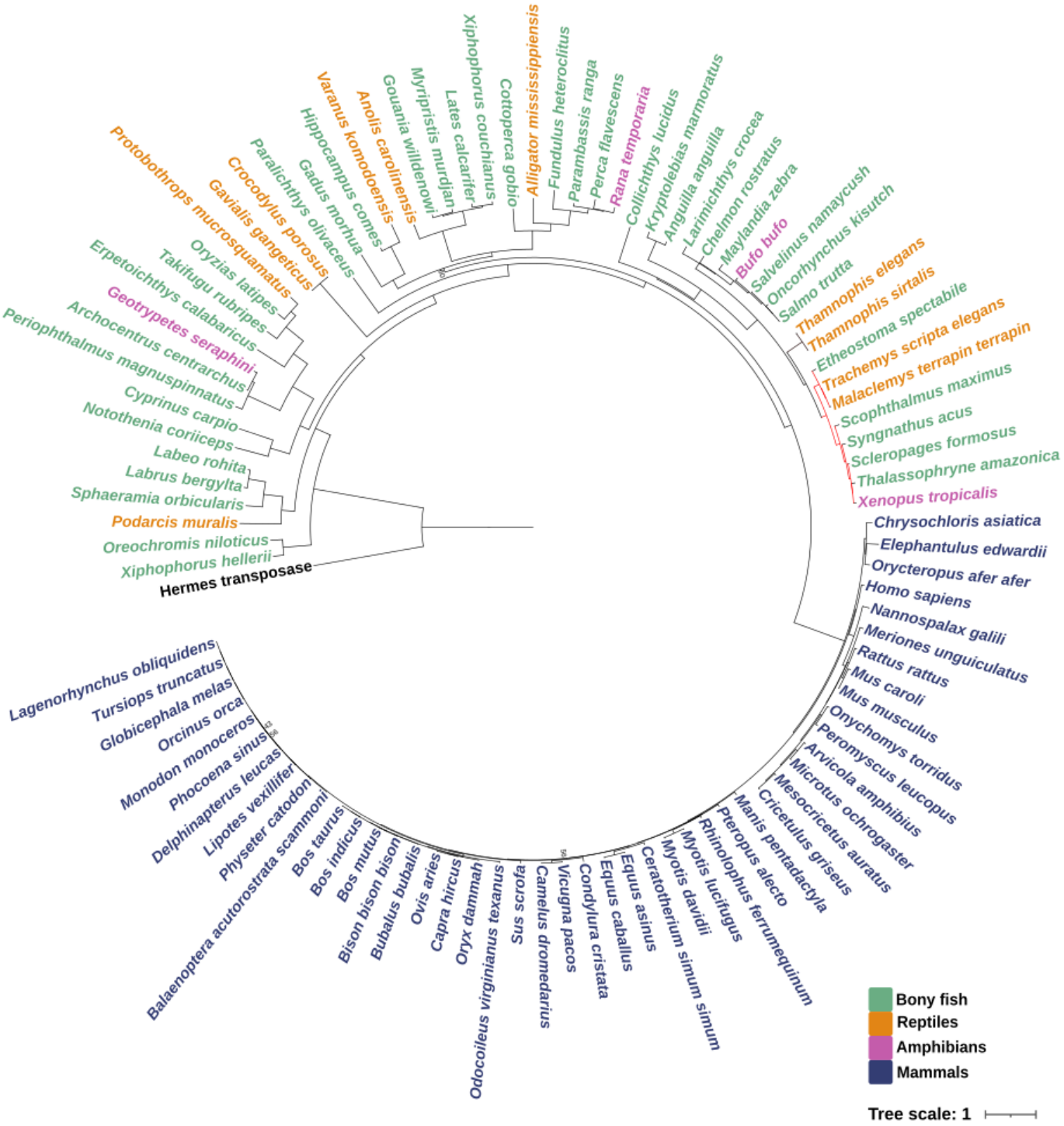

Figure 1: The phylogenetic relationship of sequences annotated as GTF2IRD2/2A from bony fishes, reptilian, amphibian, and mammalian genomes from NCBI. Bony fish are coloured green, reptiles are orange, amphibians are pink and mammals are blue. Branches labelled in red are hAT-6 transposons. The Hermes transposase was used as an outgroup. Support values under 60 are shown at nodes. For the full tree and support values, see Supplementary materials.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.24.517799; this version posted November 24, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

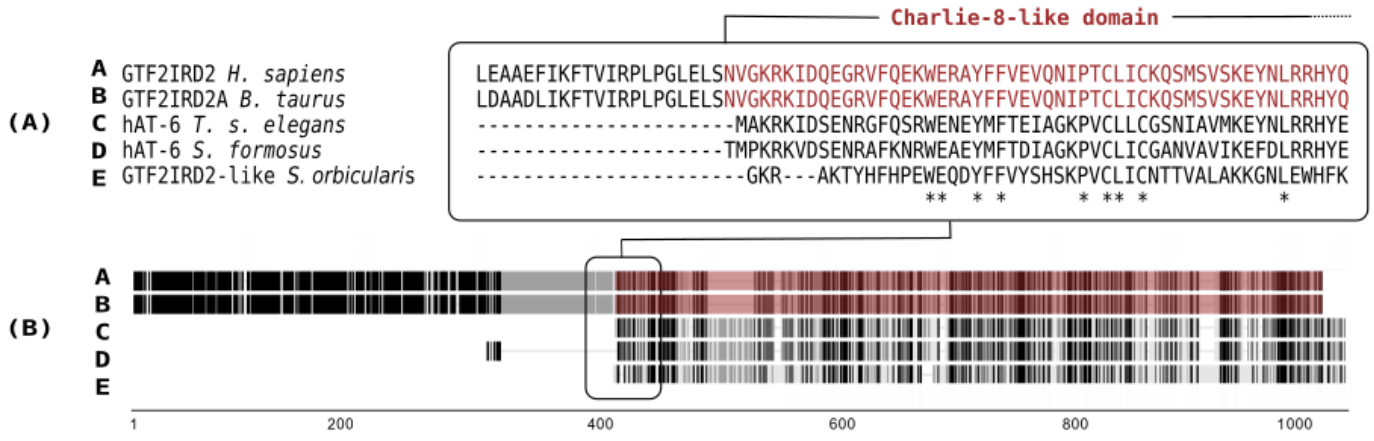

Figure 2: (A) Boundary of non-transposon and Charlie8-like domain of a multiple sequence alignment (MSA) of mammalian GTF2IRD2/2A to hAT-6 transposons and a TE sequence misannotated as GTF2IRD2-like. (B) Full schematic MSA of the selected sequences. The length of sequences in amino acids is shown on the x-axis from the 5' to 3' direction. The Charlie8-like domain of mammalian GTF2IRD2/2A is highlighted in red text and red shading.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.24.517799; this version posted November 24, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Class	Species Name	Length (bp)	Repbase Best Match	Pairwise Identity (%)
Actinopterygii	*Sphaeramia orbicularis*	2921	hAT-1_SSa	81
	Xiphophorus hellerii	3231	hAT-1_XT	79
	Archocentrus centrarchus	2522	hAT-2_PM	83
	Takifugu rubripes	2612	hAT-7_PM	82
	Erpetoichthys calabaricus	2591	hAT-4_SSa	88
	Oryzias latipes	2614	hAT-1_SSa	67
	Xiphophorus couchianus	2756	hAT-1_SSa	75
	Gouania willdenowi	2789	CHAPLIN4_FR	91
	Cottoperca gobio	2656	hAT-1_SSa	74
	Kryptolebias marmoratus	2673	hAT-1_SSa	74
	Larimichthys crocea	2787	hAT-6_XT	87
	Chelmon rostratus	2530	hAT-1_SSa	80
	Maylandia zebra	2551	hAT-1_SSa	87
	Salvelinus namaycush	2488	hAT-1_SSa	99
	Oncorhynchus kisutch	2508	hAT-1_SSa	98
	Salmo trutta	2893	hAT-1_SSa	99
	Labrus bergylta	2191	hAT-2B_PM	96
	Labeo rohita	2528	hAT-8_PM	72
	Paralichthys olivaceus	2255	hAT-17_SSa	79
	Gadus morhua	2803	hAT-22_DR	70
	Lates calcarifer	2788	hAT-9_DR	83
	Collichthys lucidus	3127	CHAPLIN8_FR	91
	Anguilla anguilla	3562	hAT-1_SSa	95
	Perca flavescens	2442	hAT-1_SSa	80
	Parambassis ranga	2630	hAT-4_SSa	70
Amphibia	*Bufo bufo*	2392	hAT-1_SSa	93
Reptilia	*Anolis carolinensis*	2250	hAT-2_AC	100
	Alligator mississippiensis	4120	hAT-7_AMi	87

Table 1: Species from the *Actinopterygii, Reptilia* and *Amphibia* class containing hAT transposons that were incorrectly annotated as GTF2IRD2/2A. The best match of each hAT transposon to the Repbase database and the corresponding pairwise identity is shown.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.24.517799; this version posted November 24, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Species name	5' TIR	[hAT-6]	3'TIR
Sphaeramia orbicularis	-**CA**CATA**T**GTC**AAA**C-T	----------	**AGTTT**GACA**CCCCT**G
Xiphophorus hellerii	-**C**CAGTG**T**ATTGATG-T	----------	**AGTTT**GACA**CCCCT**G
Archocentrus centrarchus	-**CA**GGGG**T**GTCC**AAA**-C	----------	**AGTTT**GGACA**CCCCT**G
Takifugu rubripes	-**CA**GGGG**T**GGGC**AAA**-C	----------	**AGTTT**GCCCATG**CCT**-
Erpetoichthys calabaricus	-**CA**CAGG**T**GTC**AAA**C-T	----------	TATAAGCCTTGCT**T**G
Oryzias latipes	-**CA**GACC**T**GGGC**AAA**-C	----------	AA**TT**GCCCA**CCCCT**G
Xiphophorus couchianus	-**CAC**ATA**T**GTCAGAG-T	----------	AC**TTT**GACA**CCCCT**G
Gouania willdenowi	-**CA**GGGG**T**GTC**AAA**C-G	----------	**AGTTT**GACA**CCCCT**G
Cottoperca gobio	-**CA**CATA**T**GTCAGAG-T	----------	AC**TTT**GACA**CCCCT**G
Kryptolebias marmoratus	-**CA**CATA**T**GTCAGAG-T	----------	AC**TTT**GACA**CCCCT**G
Larimichthys crocea	-**CA**TATG**T**GTC**AAA**C-T	----------	TAAAGAACCCAGC**T**A
Chelmon rostratus	-**CA**TATGAGTC**AAA**C-T	----------	**AGTTT**GACA**CCCCT**G
Maylandia zebra	-**CA**GGGG**T**GTC**AAA**C-T	----------	**AGTTT**GACACCTA**T**G
Salvelinus namaycush	-**CA**TGGG**T**GTC**AAA**C-T	----------	GA**TTT**GACA**CCCCT**G
Oncorhynchus kisutch	-**CA**TGGG**T**GTC**AAA**C-T	----------	**AGTTT**GACA**CCCCT**G
Salmo trutta	-**CA**TGGG**T**GTC**AAA**C-T	----------	**AGTTT**GACA**CCCCT**G
Labrus bergylta	-**CA**GGGG**T**GTCC**AAA**-G	----------	**AGTTT**GGACA**CCCC**T-
Labeo rohita	-**CA**GGGG**T**CTC**AAA**C-T	----------	**AGTTT**GAGA**CCCCT**G
Acipenser ruthenus	-**CA**GGGA**T**AGTCAAT-T	----------	TA**TT**GACTA**CCCCT**G
Paralichthys olivaceus	-**CA**GGGG**T**GGGGAAC-G	----------	AG**GTT**CCCCA**CCCCT**-
Gadus morhua	-**CA**GGGG**T**GCCC**AAA**-T	----------	AATGGGGCTGCCC**T**G
Lates calcarifer	--**A**GTGG**T**CCCC**AAA**CT	----------	AAG**TTT**GGGGA**CCCC**--
Collichthys lucidus	-**CA**GGGG**T**CTC**AAA**C-T	----------	**AGTTT**GAGA**CCCCT**G
Anguilla anguilla	-**CA**GGGG**T**GTC**AAA**C-T	----------	AGA**TT**GTAG**CCCCT**A
Perca flavescens	-**CA**CATG**T**GTC**AAA**C-T	----------	**AGTTT**GACACCCT**T**G
Parambassis ranga	-**CA**GGGG**T**GGGCAAT-T	----------	CAATGCCCAGGTC**T**G
Bufo bufo	-**CA**GGGG**T**GTC**AAA**C-T	----------	**AGTTT**GACACCTA**T**G
Anolis carolinensis	T**CA**GGGG**T**GCTTTGA-T	----------	**AGTTT**GGGGA**CCCCT**G
Alligator mississippiensis	AG**CA**GGG**T**GGGC**AAA**-A	----------	AA**TT**GCCCA**CCCCT**G

Figure 3: Multiple alignment of the 5' and 3' TIRs of *Actinopterygii*, *Amphibia* and *Reptilia* hAT transposons derived from sequences misannotated as GTF2IRD2/2A. Nucleotides in bold text are conserved in most of the species.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.24.517799; this version posted November 24, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References:

1. Tipney HJ, Hinsley TA, Brass A, Metcalfe K, Donnai D, Tassabehji M. Isolation and characterisation of GTF2IRD2, a novel fusion gene and member of the TFII-I family of transcription factors, deleted in Williams–Beuren syndrome [Internet]. European Journal of Human Genetics. 2004. p. 551–60. Available from: http://dx.doi.org/10.1038/sj.ejhg.5201174

2. Arensburger P, Hice RH, Zhou L, Smith RC, Tom AC, Wright JA, et al. Phylogenetic and functional characterization of the hAT transposon superfamily. Genetics. 2011;188:45–57.

3. Jurka J, Naik A, Kapitonov V: CHARLIE8 RepBase entry RepBase release 7.2.

4. Blum M, Chang H-Y, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, et al. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 2021;49:D344–54.

5. Nesmelova IV, Hackett PB. DDE transposases: Structural similarity and diversity. Adv Drug Deliv Rev. 2010;62:1187–95.

6. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol. 2000;7:203–14.

7. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA [Internet]. Journal of Molecular Biology. 1997. p. 78–94. Available from: http://dx.doi.org/10.1006/jmbi.1997.0951

8. Kuraku S, Zmasek CM, Nishimura O, Katoh K. aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity. Nucleic Acids Res. 2013;41:W22–8.

9. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.

10. Steenwyk JL, Buida TJ 3rd, Li Y, Shen X-X, Rokas A. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. PLoS Biol. 2020;18:e3001007.

11. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.24.517799; this version posted November 24, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2020;37:1530–4.

12. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol. 2018;35:518–22.

13. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587–9.