The role of transposon activity in shaping cis-regulatory element evolution after whole genome duplication

Øystein Monsen[1], Lars Grønvold[1*], Alex Datsomor[1], Thomas Harvey[1], James Kijas[2], Alexander Suh[3,4], Torgeir R. Hvidsten[5 †], Simen Rød Sandve[1†]*

* contributed equally

[†]corresponding authors

[1]Department of Animal and Aquacultural Sciences, Faculty of Bioscience, Norwegian University of Life Sciences

[2]Aquaculture Programme, Commonwealth Scientific and Industrial Research Organisation

[3]School of Biological Sciences – Organisms and the Environment, University of East Anglia, Norwich Research Park, NR4 7TU, Norwich, UK

[4]Department of Organismal Biology – Systematic Biology (EBC), Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden

[5]Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås, Norway

Present address of Alexander Suh: Present address: Centre for Molecular Biodiversity Research, Leibniz Institute for the Analysis of Biodiversity Change, Adenauerallee 160, D-53113 Bonn, Germany

Abstract

Background: Two of the most potent drivers of genome evolution in eukaryotes are whole genome duplications (WGD) and transposable element (TE) activity. These two mutational forces can also play synergistic roles; WGDs result in both cellular stress and functional redundancy, which would allow TEs to escape host-silencing mechanisms and effectively spread with reduced impact on fitness. As TEs can function as, or evolve into, TE-derived cis-regulatory elements (TE-CREs), bursts of TE-activity following WGD are likely to impact evolution of gene regulation. However, the role of TEs in genome regulatory remodelling after WGDs is unclear. Here we used the genome of Atlantic salmon, which is known to have experienced massive expansion of TEs after a WGD ~100 Mya, as a model system to explore the synergistic roles of TEs and WGDs on genome regulatory evolution.

Results: We identified 61,309 putative TE-CREs in Atlantic salmon using chromatin accessibility data from brain and liver. Of these, 82% were tissue specific to liver (43%) or brain (39%) and TE-CREs originating from retroelements were twice as common as those originating from DNA elements. Signatures of selection shaping TE-CRE evolution were evident from depletion of TEs in open chromatin, a bias in tissue-shared TE-CREs towards older TE-insertions, as well as tissue-specific processes shaping the TE-CRE repertoire. The DTT elements (Tc1-Mariners), which exploded in numbers at the time of the WGD, were significantly less prone to evolve into TE-CREs and significantly less potent in driving or repressing transcription compared to other TE-derived sequences. A minority of TEs (16% of consensus TEs) accounted for the origin of 46% of all TE-CREs, but these 'CRE-superspreaders' were not temporally associated with the WGD. Rather, the majority of TE-CREs, including those found to be significantly associated with gene regulatory evolution and thus found to drive or repress transcription, evolved from TE activity occurring across tens of millions of years following the WGD event.

Conclusion: Our results do not support a WGD-associated TE-CRE rewiring of gene regulation. Instead we find that TEs from diverse superfamilies have been particularly effective in spreading TE-CREs and shaping gene regulatory networks under tissue-specific selection pressures, across millions of years following the salmonid WGD.

Introduction

The two most influential mutational mechanisms that have shaped eukaryotic genome evolution are whole genome duplications (WGD) and transposable element (TE) activity. Both WGDs and TEs drive genome size evolution. However, as mobile genetic elements with capacity to replicate (Feschotte and Pritham, 2007), TEs also impact genome evolution in numerous other ways, by generating novel genes (Cosby et al., 2021; Diehl et al., 2020; Elisaphenko et al., 2008; Qin et al., 2015), modulating chromatin looping (Diehl et al., 2020), rearranging genome structure (Bourque et al., 2018) as well as supplying "raw material" for gene regulatory evolution in the form of cis-regulatory elements (CREs) (Bourque et al., 2008; Chuong et al., 2017; Cosby et al., 2019; Diehl et al., 2020; Feschotte, 2008; Sundaram and Wysocka, 2020; Sundaram et al., 2014).

Studies of mammalian genomes have provided deep insights into the role of TEs in CRE-evolution and the potency of TE-derived CREs (TE-CREs) to regulate gene expression (reviewed in Fueyo et al. (2022)). For example, as much as 40% of the mouse and human transcription factor (TF) binding sites have been shown to be within TEs (Sundaram et al., 2014), and as many as 19% of pluripotency factor TFs are located within TEs (Kunarso et al., 2010; Sundaram et al., 2017). Curiously, in mammals TEs associated with gene regulation during development have been shown to be younger than those associated with regulation in adult somatic tissues (reviewed in (Fueyo et al., 2022)), suggesting different evolutionary pressures on TEs with distinct regulatory roles.

Genome evolution through TE activity is also likely influenced by WGDs. Because WGDs result in cellular stress, TEs can escape host-silencing mechanisms following WGDs. This is supported by both experimental (Kashkush et al., 2003, 2002; Kraitshtein et al., 2010) and comparative genomics (Lien et al., 2016; Marburger et al., 2018) studies. Additionally, WGDs result in increased functional redundancy. This will reduce the average negative fitness effects of novel TE insertions and thereby allow for fixation of TE insertions following WGD (Baduel et al., 2019), including insertions that influence gene regulation. In line with this, Gillard et al. (Gillard et al., 2021) recently reported that TE insertions in promoters were associated with regulatory divergence of gene

duplicates following WGD in salmonid fish. However, systematic investigations into the role of TEs in CRE evolution and genome regulatory remodelling after WGDs are still lacking.

Here we address this knowledge gap regarding the role of WGD in TE-associated genome regulatory evolution using salmonids as a model system. Salmonids underwent a WGD 80-100 Mya (Lien et al., 2016) which coincided with the onset of a burst of TE, particularly featuring elements belonging to the DTT/Tc1-mariner superfamily. This observation has led to the hypothesis that increased TE activity in the immediate aftermath of the WGD was a major driver of genome regulatory evolution. To explore this idea we leverage ATAC-seq data from two tissues (brain and liver) to identify putative CREs that have evolved from TE-derived sequences. We then combine these TE-CRE annotations with analyses of the temporal dynamics of TE activity, analyses of gene-coexpression, and massive parallel reporter assays. Our results support a weak link between WGD and TE-CRE evolution, but cast doubts about the power of synergistic interactions between WGDs and TE activity to drive rapid rewiring of genome regulation.

Results

The TE-CRE landscape of Atlantic salmon

To investigate the contributions of different TEs to CRE evolution, we first characterised the TE landscape of the salmon genome using an updated version of the existing TE annotation from (Lien et al., 2016). The total transposable element annotation covered 51.92% of the genome. Consistent with previous findings (Goodier and Davidson, 1994; Lien et al., 2016), the dominating TE group was DNA transposons from the Tc1-Mariner superfamily with >655,000 copies, covering 327 million base pairs, just shy of 10% of the genome (Figure 1A-C). In general, the genomic context of TE insertions was quite similar to the genomic baseline (Figure 1D), but with slightly more TEs in intronic regions and slightly less TEs in exons and intergenic regions. Of the well-represented TE superfamilies (>10k insertions) only the Nimb retrotransposon superfamily was an exception to this pattern, for which 18% of the copies were found in promoter regions

(Figure 1D).

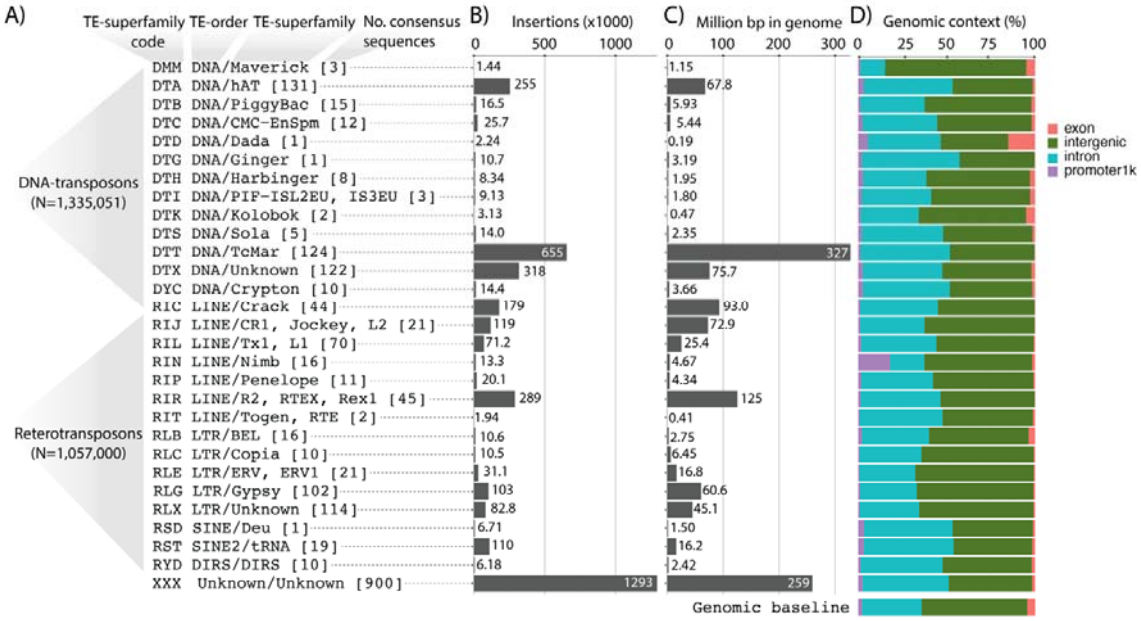

Figure 1. Overview of the genomic TE landscape. A) Superfamily level overview of TE annotations in the Atlantic salmon genome. Number of TE subfamilies per superfamily in square brackets. **B)** TE insertions per superfamily. **C)** Annotated base pairs at the TE superfamily level. **D)** TE annotations (bp proportions) overlapping different genomic contexts. Genomic baseline is the proportion of the entire genomic sequence that is assigned to the four genomic contexts.

Active CREs in tissues and cells are associated with increased chromatin accessibility (Buenrostro et al., 2013; Keene et al., 1981; McGhee et al., 1981). Thus, to study the contribution of TEs to the salmon CRE landscape, we integrated our TE annotation with annotations of accessible chromatin regions identified using ATAC-seq data from liver and brain. Analysis of the overlap between TEs and accessible chromatin revealed a large depletion of TEs in accessible chromatin. While TEs represent ~52% of the genome sequence, only <20% of the regions of accessible chromatin overlapped with TE insertions (Figure 2A), with liver having a higher proportion of annotated TEs in accessible chromatin than brain.

To define a set of TEs that contribute to putative CREs, we narrowed in on those TE annotations overlapping chromatin accessibility peaks (Figure 2B-D). These were defined as putative TE-CREs. Although the majority (55%) of TE annotations (excluding 'unknown' repeats without classification) were DNA elements (1,335,051 insertions), TE-CREs from DNA elements were a minority (27%). Both the proportion (Figure 2C) and number (Figure 2D) of putative TE-CREs were higher in the liver compared to the

brain. Of a total of 61,309 TE-CREs, 18% were shared between tissues, 39% were brain-specific, and 43% were liver-specific (Figure 2D). Tissue-shared TE-CREs were overrepresented about 4-fold in promoters compared to the tissue-specific TE-CREs (Figure 2E). We also found that tissue specific TE-CREs were associated with tissue-bias in gene expression (Figure 2F), supporting a regulatory effect of TE-CREs.

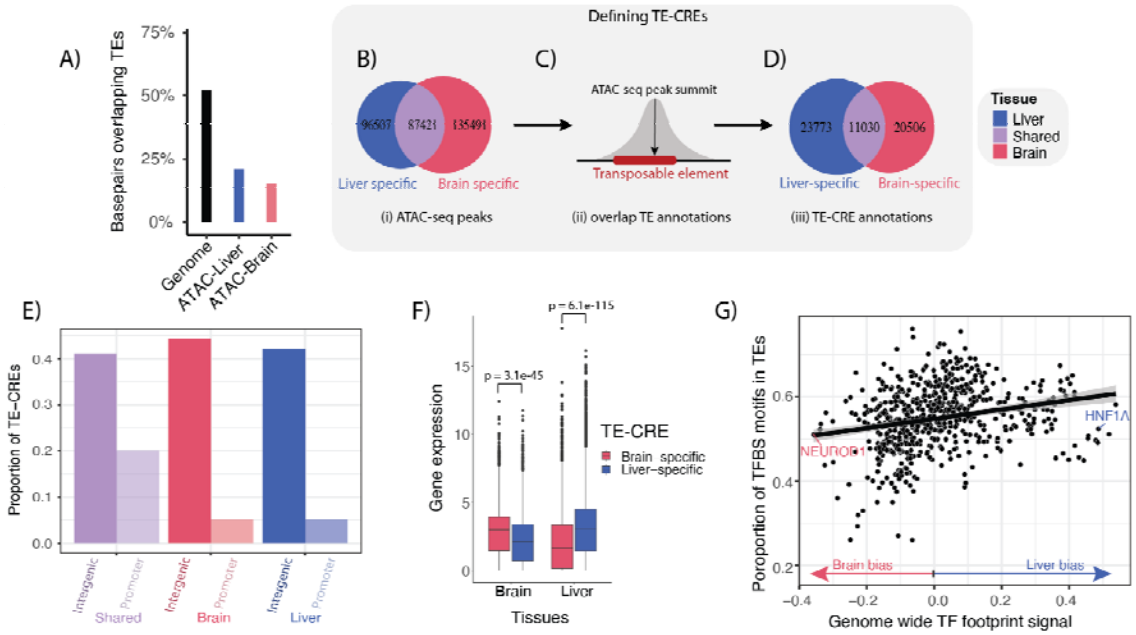

Figure 2. TE-CRE landscape. A) The proportion of base pairs overlapping TEs, either out of all genome-wide bp or those within an ATAC-seq peak. B-D) Pipeline to define putative TE-CREs. B) Venn diagram of tissue specific and shared ATAC-peaks from liver and brain. C) Cartoon showing how TE-CREs are defined as ATAC-seq peak summits when overlapping with a TE. D) Venn diagram of tissue specific and shared TE-CREs from liver and brain. E) Proportion of shared and tissue-specific TE-CREs in promoter vs. intergenic regions. F) Gene expression levels of genes associated with tissue specific TE-CREs in brain and liver. P-values from Wilcoxon-test indicated above tissues. G) Correlation between the proportion of TFBS motifs found in TEs using FIMO and the genome wide TF footprint signal for the TFs predicted to bind these TFBSs.

One reason for TE-CREs tending to be specific to liver rather than brain (Figure 2D) could be due to higher selective constraints on gene regulatory networks important for brain function compared to liver function. One expectation from this hypothesis would be that TFBSs with strong brain bias in TF binding would be depleted in TE sequences. To test this, we first inferred tissue bias in TF binding using genome-wide TFBS occupancy signals through TF-footprinting. We then correlated these signals with the proportion of TFBS motifs found in TE sequences. In line with our expectations, motifs for brain biassed TFs were less frequently found in TEs (regression line in Figure 2G). The most highly liver-biassed TFs, such as HNF1A, were an exception to this general

trend, although these liver-biassed TFs were fewer and much less depleted in TEs compared to the most highly brain-biassed TFs (Figure 2G). Taken together, our results support a role of tissue-specific differences in the selective constraints shaping TE-CRE evolution.

A minority of TEs have CRE superspreader abilities

Next we wanted to understand the contribution of specific TE superfamilies to the TE-CRE landscape. Overall, there was a positive linear relationship between the genomic copy number and the number of TE-CRE for TE superfamilies (Figure 3A). However, some superfamilies (Figure 3A, see data points outside 95% CI), contributed significantly less (RIC and DTT) or more (RIN, RLG, DTA, RIJ) to the TE-CRE landscape than expected based on the genomic copy numbers (Figure 3A). In particular, DTT superfamily elements, which are dominating in terms of numbers of insertions (~27% of all TE copies with an assigned taxonomy), represented only ~4% of the TE-CREs.

To further characterise the TE-CRE landscape in more detail, we identified TEs enriched in open chromatin at the level of TE consensus sequences (Figure 3B). These TEs are hereafter referred to as 'CRE-superspreaders'. Among the 1119 TE consensus sequences with >500 genomic copies, only 178 (16%) were defined as CRE-superspreaders (Figure 3B). Forty nine percent of the superspreaders were enriched in open chromatin in both tissues (88), while 39% (69) and 12% (21) were tissue-specific and enriched in accessible chromatin only in the liver or brain, respectively. The proportion of taxonomically unclassified repeats (three-letter code "XXX") was high among the identified CRE-superspreaders (101 subfamilies). We therefore performed manual curation, resulting in four TEs being discarded from further analyses, and a reduction of taxonomically unclassified TEs to 34 (Supplementary Table 1).

We find that CRE-superspreaders were taxonomically diverse, belonging to 18 different TE superfamilies, but that very few DTT elements evolved into CRE-superspreaders (Figure 3C). Note that superfamilies consisting only of CRE-superspreader TEs is a technical artefact stemming from the manual curation of the taxonomically unknown (three-letter code XXX) superspreaders.

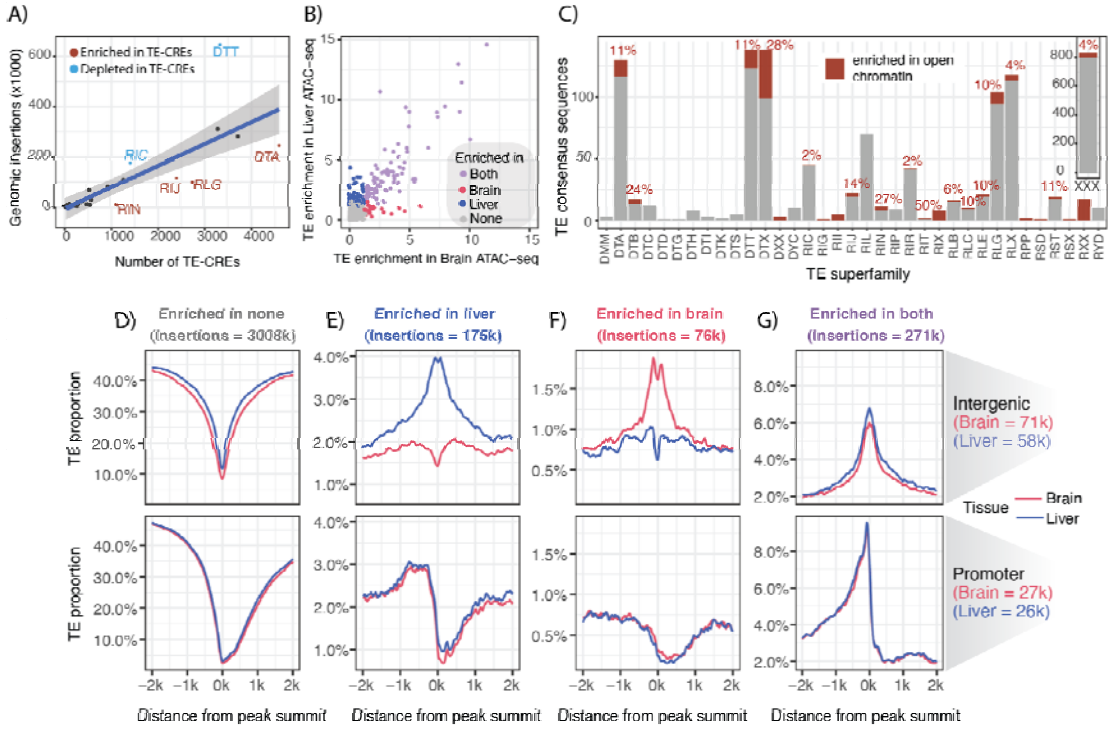

Figure 3. TE subfamilies enriched in open chromatin. A) The number of insertions per superfamily plotted against the number of CREs in each superfamily. The shaded area is a 95% confidence level interval. B) TE consensus sequences plotted according to fold-enrichment within ATAC-seq peaks in brain and liver. TE subfamilies are assigned into categories based on enrichment in liver, brain or both. C) TE consensus sequences enriched in open chromatin after manual curation of significant TEs in enrichment tests. Only TE consensus sequences with > 500 insertions have been included. Percent enriched TE consensus sequences are indicated above bars. D-G) Proportion of bp overlapping TEs from each enrichment category around peak summits in intergenic or promoter regions (summit within 500 bases of TSS). Peaks in promoter regions are oriented according to the corresponding TSS with gene bodies to the right in figures.

Next we explored the local TE landscape around the open chromatin peaks in various genomic contexts (intergenic and promoters) and tissues (Figure 3 D-G). We find that in promoters the proportion of TEs in open chromatin decreases towards TSS and the gene body, reflecting increased purifying selection pressure (less tolerance for TE-insertions). Furthermore we find that close to genes (i.e. in promoters), TE proportions were higher for tissue-shared (Figure 3G) compared to tissue-specific (Figure 3E-F) TE-CREs. In intergenic regions (i.e. enhancers) we find very strong tissue-specific TE enrichment signals not present in promoter TE-CREs (Figure 3 E-F). In sum, we find that CRE-superspreader TEs are biassed towards certain taxonomic groups of TEs and that these TEs are enriched in accessible chromatin with distinct patterns and effect sizes across tissues and genomic contexts.

190 ## The temporal dynamics of TE-CRE evolution

191 The main hypothesis we set out to test in this study was whether the increase in TE-
192 activity associated with salmonid WGD was instrumental in driving TE-CRE evolution.
193 To explore this hypothesis we calculated sequence divergence between TE insertions
194 and their consensus sequence, used this as a proxy for time, and compared it to the
195 expected ~87% sequence similarity between genomic regions arising from the
196 salmonid WGD event (Lien et al., 2016). One challenge with such sequence similarity
197 based comparisons is the intrinsic connection between sequence similarity and
198 purifying selection pressure can bias our results. Here we used the entire TE insertion
199 (not only the part that is in open chromatin) to estimate divergence to consensus, hence
200 we expect this bias to be negligible. Nevertheless, we first analysed the sequence
201 similarity distributions of different classes of TE-CREs as well as TE sequences not in
202 accessible chromatin (Figure 4A). As expected, we do not find that TE-CREs are more
203 similar to their consensus than other TE insertions, supporting that putative purifying
204 selection on TE-CRE function does not transfer to our sequence similarity based age-
205 proxy. If anything, the TE sequences giving rise to tissue-shared CREs are older than TEs
206 not giving rise to TE-CREs (see 'both' in Figure 4A).

207 Next we stratified the TE-CREs on their transposition age-proxy relative to the WGD and
208 their taxonomic order (Figures 4B and 4C). Twelve percent (136) of TE consensus
209 sequences had a mean sequence similarity to TE insertions reflecting activity at the time
210 of, or shortly after the WGD (87-88%). Among the TE consensus sequences with CRE-
211 superspreader ability, a similar proportion of DNA (17.1%) and retroelements (15.2%)
212 were active around the time of the WGD (Figure 4B). However, in absolute numbers
213 retroelements were dominating as CRE-superspreaders (twice as many as DNA-
214 elements) both in terms of the number of consensus TEs (Figure 4B), and the number of
215 TE-CREs originating from these CRE-superspreaders (Figure 4C).

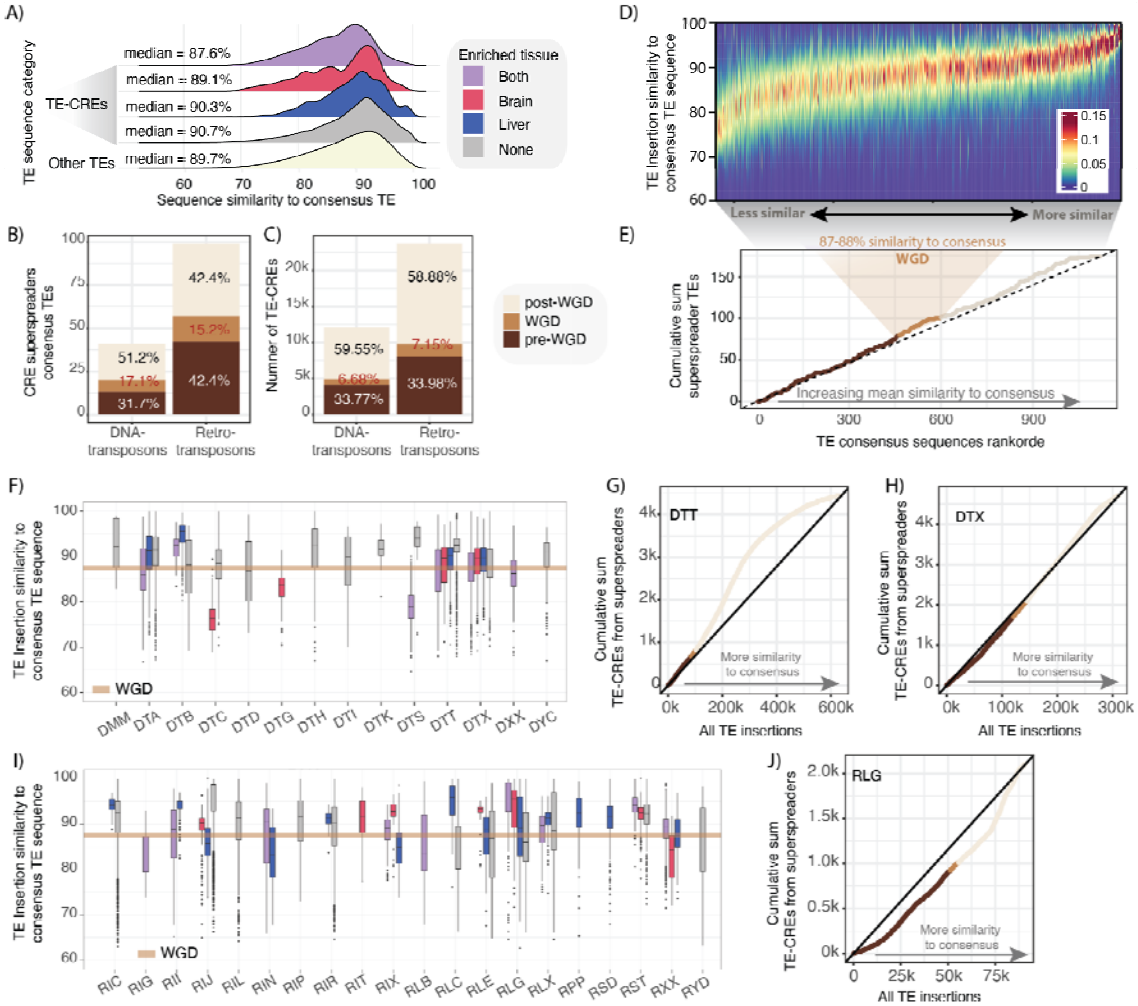

Figure 4. Temporal dynamics of TE-CRE insertion activity by TE taxonomy. A) Distribution of sequence similarity of TE-CREs to their TE consensus sequence. Colours represent if TE-CRE are from TE consensus sequences with superspreader ability (liver, brain, or both) or not (grey). B) Number of TE consensus sequences with superspreader ability subdivided into DNA- and retroelements. Colours represent the age proxy calculated as mean similarity between genomic insertions and their consensus TE sequence. Post-WGD = >88% similarity, WGD = 87-88% similarity, pre-WGD = <87% similarity. C) Number of TE-CREs from TEs with a taxonomic classification (excluding unknown) subdivided into DNA- and retroelements. Colours represent the age proxy as defined by similarity to TE consensus sequence. D) Heatmap of the similarity distributions of TE-CRE insertions to their consensus sequence. TE consensus sequences are ordered based on mean similarity to their consensus. E) Cumulative distribution of CRE-superspreader consensus TEs ordered by mean similarity between genomic copies and TE consensus sequence. Colours represent age proxy as defined by mean similarity to TE consensus sequence F, I) Distribution of sequence similarity between TE-CRE insertions and their consensus TE sequence aggregated at the level of superfamily taxonomy. Colours represent if TE-CRE are from TE consensus sequences with superspreader ability (liver, brain, or both) or not (grey). G, H, J) Cumulative distribution of TE-CREs from superspreader families within single superfamilies (DTT, DTA, and RLG).

To understand temporal dynamics of TE-CRE evolution in more detail we then analysed the temporal dynamics of all TE consensus sequences (>500 genomic copies, Figure 4D). We then plotted the cumulative sum of TE-CRE superspreaders against all TEs ordered

236 by mean similarity to consensus (Figure 4E). If WGD were associated with a general
237 burst of CRE-superspreader activity we expect to see a steeper slope in the cumulative
238 sum distribution around the 87-88% consensus sequence similarity interval. Although
239 we find a slight change in CRE-superspreader accumulation rates around this similarity
240 range (Figure 4E), most of the data points lie on or close to the dotted line (null model)
241 and the age distribution of CRE-superspreaders TE was not significantly different from
242 other TEs (two sided Kolmogorov-Smornov test, p-value = 0.15). Nor did we find a
243 significant increase in the ratio of TE-CRE superspreader consensus sequences to
244 normal TEs in the 87-88% consensus similarity range (Fisher test, p-value = 0.45).
245 Taken together, these results do not support a model whereby the WGD caused a
246 dramatic shift in the transposition activity of TE-CRE superspreaders.

247 When these results are broken down to TE-superfamily resolution (Figure 4 F and I), we
248 see a clearer picture of the temporal heterogeneity emerge. Only a few DNA transposons
249 (DTT and DTA in Figure 4F) and retroelements (RLG in Figure 4I) appear to have a
250 similarity profile that reflect origins at, or just after, the WGD. The TE-CREs from DTT
251 superspreaders seemed to have originated from TEs having a consensus similarity close
252 to the 87-88% range, around the time of the WGD (Figure 4F). In depth analysis
253 confirmed that TE insertions from the DTT CRE-superspreaders giving rise to TE-CREs
254 accumulated at an uneven rate, increasing shortly after the WGD (Figure 4 G). TE-CREs
255 from retroelement superfamilies temporally associated with the WGD (Figure 4I), such
256 as DTX and RLG superspreaders did however not show this tendency (Figure 4 G
257 compared to H and J). Hence, our results do not reflect that TE-CREs in general have a
258 propensity to originate from TE activity around the time of the salmonid WGD (Figure
259 4D), but those TE insertions from DTT CRE-superspreaders giving rise to TE-CREs were
260 biassed towards transposition events happening after the WGD (Figure 4G).

261

262 ## Co-expression analysis support TE-CRE driven regulatory network evolution

263 If TEs are spreading CREs with sequences that either have a potent TF binding motif or are
264 prone to mutate into a TF motif, we expect different genes with similar TE-CREs (TEs insertions
265 belonging to the same consensus sequence) to be more similarly regulated than random gene
266 pairs. To identify such putative cases of TE-CRE driven evolution of gene regulation, we

assigned each TE-CRE to the closest gene and tested if genes with similar TE-CREs were more co-expressed than expected by chance.

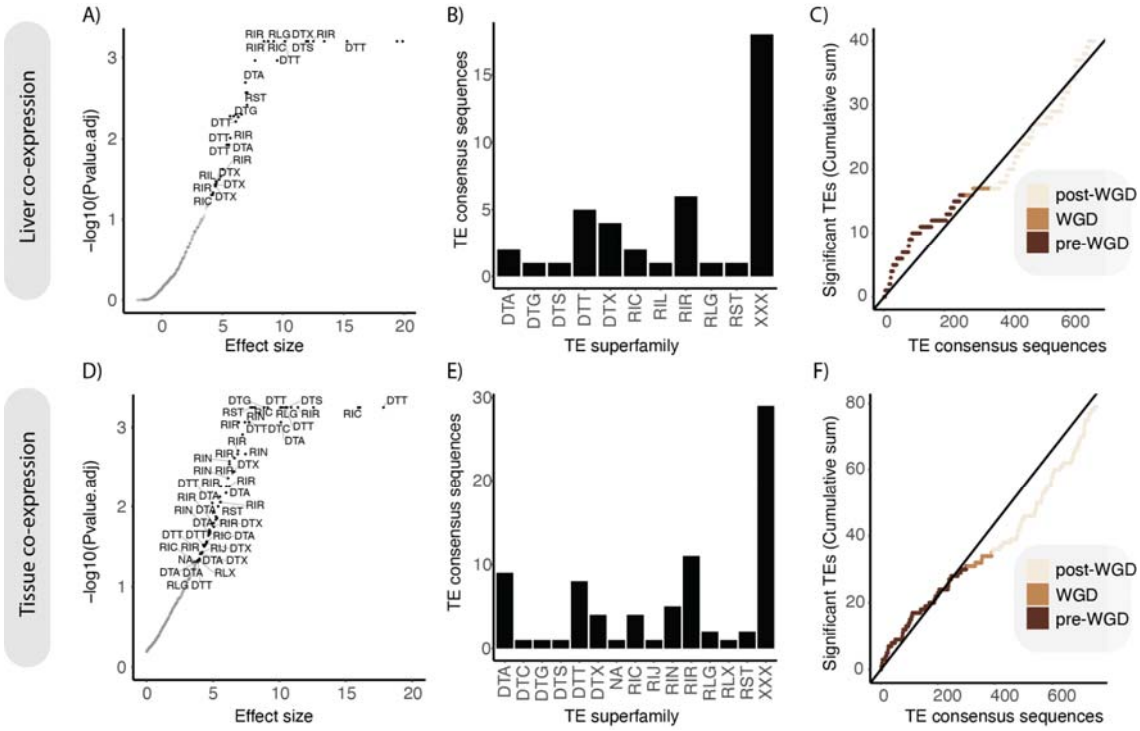

Figure 5. TE-CREs driving co-expression. Top row A-C shows results from liver co-expression. Bottom rows D-F shows results from tissue atlas co-expression. A and D) Significance (FDR-adjusted p-values) plotted against effect size (standard deviations) for each TE consensus sequence, indicating the strength of co-expression of their associated genes in the liver (B) and tissue atlas (D) co-expression networks, respectively. Points with fdr-adjusted p-value < 0.05 are labelled with superfamily names (unknown/XXX superfamilies are not labelled). B,E) Distribution of significant TE consensus sequences on superfamilies in liver (B) and tissue atlas (E) data sets. C and F) Cumulative distribution of TE consensus sequences with significant effect on gene co-expression in liver (C) and tissue atlas (F) data sets. Temporal classification was based on the median similarity of all TE insertions to their TE consensus sequence where post-WGD was defined as >88% similarity, WGD 87-88% similarity and pre-WGD <87% similarity.

We first used RNA-seq data from the liver of 112 individuals spanning different ages, sex and different diets in fresh water. In the context of this liver co-expression network, significant co-expression (low p-values) indicate that TE-CREs from one particular TE consensus are candidates for modulating the gene regulation in the liver depending on developmental and physiological states. Using only TE-CREs from liver, 42 TE consensus sequences (42 of 1395 = 3%) were associated with genes that were significantly co-expressed (FDR-corrected p-value < 0.05) (Figure 5A). Of the significant TE consensus sequences, 23 (55%) were CRE superspreaders. The significant TE consensus sequences came from 11 TE superfamilies, with TEs of unknown origin (XXX) accounting for 43% (Figure 5B). The cumulative distribution of

290 TE-CREs associated with gene co-expression did not suggest a temporal co-occurence of WGD
291 and the TE-CREs with putative gene regulatory effects (Figure 5C).

292 TE-CREs are also known to induce tissue-specific regulatory effects (Karttunen et al., 2023). We
293 therefore conducted the same analyses using RNA-seq data from 13 different tissues. Using TE-
294 CREs from both the liver and brain, 80 TE consensus sequences (80 of 1470 = 5.5%) were
295 associated with significant co-expression (Figure 5D), of which 38 (48%) were superspreaders.
296 The significant TE consensus sequences came from 15 TE superfamilies (Figure 5E). Each
297 significant TE consensus sequence was associated with a tissue TE-CRE-profile (fraction of TE-
298 CREs found in liver, brain or both), and these profiles generally agreed with the tissue
299 expression profiles of the associated genes (RNA-seq expression values across 13 tissues), thus
300 corroborating that our approach indeed identified regulatory-active TE-CREs. Similar to the
301 liver co-expression analyses, the cumulative distribution of TE-CREs impacting tissue-regulation
302 did not suggest any link between the WGD and the TE-CRE evolution shaping gene co-
303 expression (Figure 5F). Taken together, we find evidence for a small proportion (3-5%) of TE
304 consensus sequences spreading CREs that regulate nearby genes by either by modulating their
305 expression in liver or driving tissue-specific expression.

306

307 Functional validation of TE-CREs using massively parallel reporter assay

308 To be able to directly assess regulatory potential of TE-CREs in Atlantic salmon we performed
309 an ATAC-STARR-seq experiment in salmon primary liver cells (Figure 6A). This method
310 assesses the ability of random DNA fragments from accessible chromatin to modulate
311 transcription levels (Wang et al., 2018). In total, 4,267,201 million unique DNA fragments from
312 open chromatin in liver were assayed. Thirty four percent of these fragments (1,456,914) could
313 be assigned to one specific TE insertion site (>50% overlap with a TE annotation) (Figure 6B).
314 Of the TE-derived sequence fragments assayed, 1.2% had transcriptional regulatory activity, a
315 slightly lower proportion than non-TE fragments (1.6%) and, TE-derived regulatory active
316 fragments were more likely to induce transcription compared to non-TE sequences (see "Up" in
317 Figure 6C).

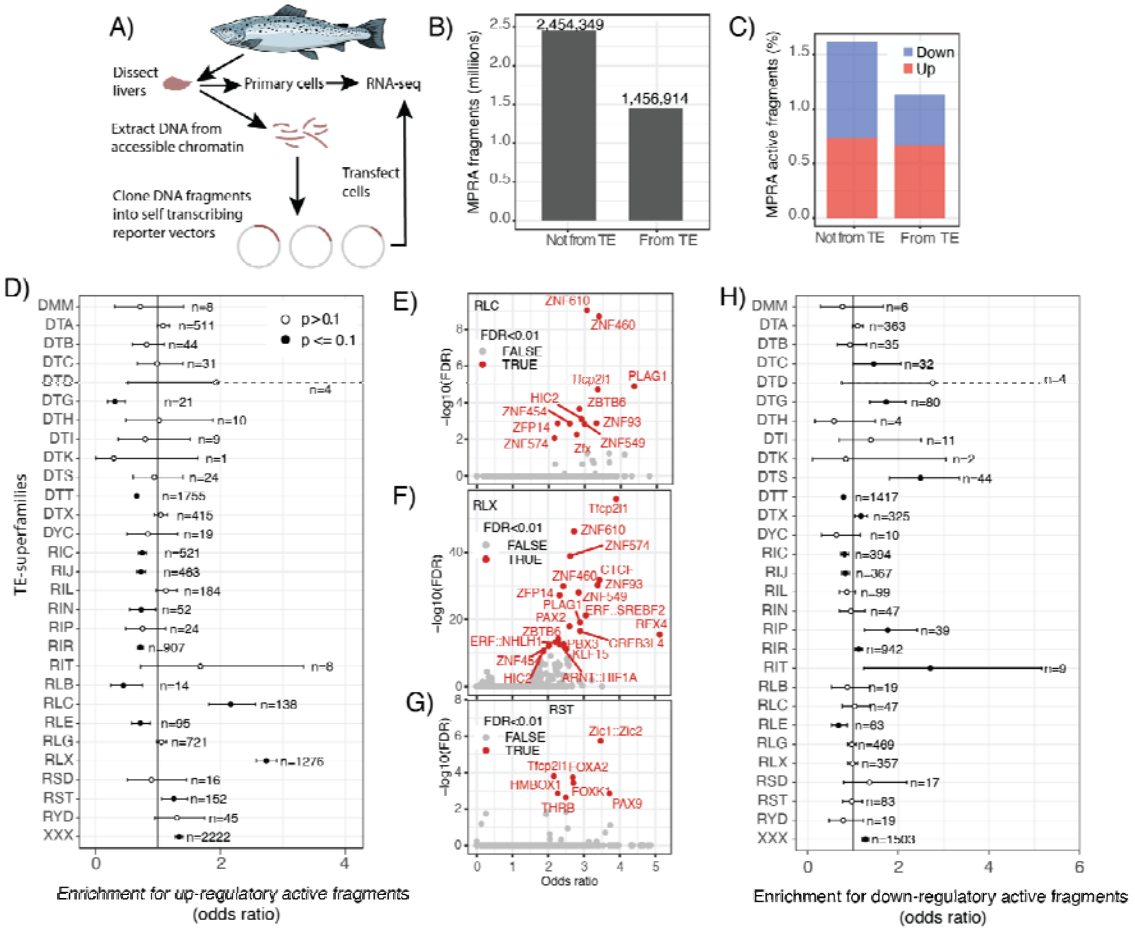

Figure 6. Massive parallel reporter assay screening of regulatory activity. A) schematic overview of the ATAC-STAR-seq MPRA experiment. B) Barplot of the origin of sequence fragments included in the analyses. C) Regulatory activity (inducer or repressor) of MPRA sequence fragments from TE and non-TE sequences. D) Fisher test results for enrichment of transcriptional inducing MPRA fragments within a TE-superfamily compared to all other TEs. Unknown taxonomy and DNA/retrotransposons of unknown origin (DTX/RLX) are considered separate groups. E-G) TFBS motif enrichment in transcriptionally inducing MPRA fragments from TE superfamilies enriched in regulatory active fragments. Number of regulatory active fragments are given for each category (n). H) Fisher test results for enrichment of transcriptional repressing MPRA fragments within a TE-superfamily compared to all other TEs. Unknown taxonomy and DNA/retrotransposons of unknown origin (DTX/RLX) are considered separate groups. Number of regulatory active fragments are given for each category (n).

To test if CRE-sequences from particular TEs were more likely to increase gene expression (i.e. act as enhancers) we compared the ratio of regulatorily active vs inactive fragments at the TE-superfamily level (including groups with partially assigned- and unknown taxonomy). These analyses revealed clear differences between superfamilies (Figure 6D). Three retrotransposon superfamilies were significantly enriched for regulatorily active fragments (Fisher test, fdr-corrected p-value < 0.05). Two of these were LTRs (RLC and RLX) which had >2-fold higher ratio of fragments acting as enhancers, while another SINE superfamily (RST) was significantly enriched but with a much lower effect size estimate (Figure 6D). The transcriptionally inducing fragments from these three superfamilies were enriched for a total of 38 unique TFBS (RLC=12,

339 RST=8, and RLX = 19) (Figure 6E-G). Many of these top-enriched TFBS motifs are known to be
340 bound by liver active TFs (i.e. SREBF2, KLF15, FOXA2, THRB) (Chaves et al., 2021; Lau et al.,
341 2018; Tao et al., 2013; Yerra and Drosatos, 2023), including the tfcp2l1 motif (Wei et al., 2019)
342 which were enriched in all three superfamilies (Figure 6E-G). We also tested for enrichment of
343 transcriptionally repressing activity and found 6 superfamilies (in addition to the XXX and DTX
344 groups) that were enriched for transcription repressing fragments (Figure 6H).

345 Interestingly, fragments from the WGD-associated DTTs were significantly less likely to be
346 regulatorily active (both to induce and repress transcription) compared to other TE-derived
347 sequences in the MPRA experiment (i.e. odds ratio < 1 in Figures 6 D and H). This is consistent
348 with our findings that DTTs are depleted in TE-CREs compared to random expectations (Figure
349 3A).

350

351 # Discussion

352 ## The Atlantic salmon TE-CRE landscape

353 Most in-depth characterizations of TE-associated CREs have so far been carried out in
354 mammalian cells and tissues. Our investigations into the Atlantic salmon genome revealed
355 similarities with mammals, but also highlighted some unique features of the salmonid TE-CRE
356 landscape. About ~15-20% of CREs were derived from TE-sequences (Figure 2A, 2C), which is
357 in the lower bound of what has been found in mammals using similar methods to identify TE-
358 CREs (Bourque et al., 2008; Kunarso et al., 2010; Sundaram et al., 2014). Consistent with studies
359 of mammalian genomes (Nishihara, 2019; Simonti et al., 2017), the majority of putative TE-CREs
360 in Atlantic salmon were associated with enhancer function rather than promoters (Figure 2E) .

361 Mammalian TE repertoire (Feschotte and Pritham, 2007) and TE-CRE landscapes (Nishihara,
362 2019; Pehrsson et al., 2019; Roller et al., 2021) are dominated by retroelements. In most fish
363 (Shao et al., 2019), including Atlantic salmon (Figure 1, DNA transposons = 55% of the TEs),
364 DNA transposons are the dominating TEs. However, similar to mammals the majority of Atlantic
365 salmon TE-CREs (73%/45,419) were derived from retroelements (Figure 4C). Our MPRA data
366 (Figure 6) also pointed to retroelements being more likely to induce transcription compared to
367 DNA transposons (Figure 6 D) and that transcription-inducing fragments from these TEs were
368 enriched for TF binding motifs known to be bound by liver-active TFs (Chaves et al., 2021; Lau
369 et al., 2018; Tao et al., 2013; Yerra and Drosatos, 2023). Only one TFBS, the tfcp2l1, was
370 enriched across all three superfamilies enriched for transcription-inducing fragments (Figure 6

371 E-G). Tfcp2l1 has previously been found to bind LTRs in human stem cells (Wang et al., 2014)
372 and is proposed to be a top regulator of human hepatocyte differentiation (Wei et al., 2019).
373 Hence the tfcp2l1 stands out as a key player in shaping evolution of retroelement-associated TE-
374 CRE landscapes in Atlantic salmon.

375 Although retroelements dominate the salmon TE-CRE landscape, the role of DNA elements in
376 TE-CRE evolution cannot be neglected. The TE superfamily contributing to the highest numbers
377 of TE-CREs was in fact the DTA (hATs) DNA elements (Figure 3A). DTAs have also been found
378 important for TE-CRE evolution in several other species. Enrichment of DTA element-insertions
379 in accessible chromatin has also been found in maize (Noshay et al., 2021), and DTA elements
380 make up a significant proportion (15%) of the TE-derived CTCF sites associated with TAD loop
381 anchoring in certain human cell types (Choudhary et al., 2023). Here we find DTA elements to
382 be the second most important superfamily in driving rewiring of tissue gene regulatory
383 networks (Figure 6D). Furthermore, even though DTA sequences were not significantly more
384 likely to drive transcription compared any other TE superfamily (fdr-corrected p-value = 0.18,
385 Figure 6D), DTA sequences were more likely to induce transcription (0.72% of fragments were
386 up-regulatory active) compared to sequence fragments derived from DNA transposons in
387 general (0.51% up-regulatory active). Hence, the DTA group of TEs is a considerable source of
388 CRE sequences that have likely played an important role in the evolution of genome regulation
389 in Atlantic salmon.

390

391 Selection on TE-CRE repertoire

392 Studies of the how evolutionary forces shape the TE landscape highlight strong purifying
393 selection on TE accumulation within protein-coding gene sequences (Bartolomé et al., 2002;
394 Rizzon et al., 2003), but also in non-coding regions (Bergthorsson et al., 2020; Hollister and
395 Gaut, 2009; Langmüller et al., 2023). These selection signatures on TE insertions in non-coding
396 regions indicate selective forces on TE-CRE evolution, which is also evident from several
397 analyses in our study.

398 We find clear underrepresentation of TE sequences in accessible chromatin (Figure 2A), and in
399 particular near the peaks in accessible chromatin in promoters and intergenic regions (Figure
400 3D), consistent with purifying selection against TE accumulation in regulatory active regions
401 (Bergthorsson et al., 2020; Langmüller et al., 2023).

402 In mammals, TEs-CRE are typically from older TE insertions (Pehrsson et al., 2019; Simonti et
403 al., 2017) suggesting that selection pressure on TEs depend on TE insertion age, which is likely
404 related to deterioration of transposition ability as TEs age and accumulate mutations. In Atlantic
405 salmon however, we do not find a general trend of older TE-sequences giving rise to TE-CREs
406 (Figure 4A). This could be linked to a general relaxation of purifying selection pressure after
407 WGD (Baduel et al., 2019; Ronfort, 1999), see section below for in depth discussion. However,
408 we do find that tissue-shared TE-CREs clearly have an older origin compared to tissue specific
409 TE-CREs (Figure 4A), which is difficult to attribute to the WGD. One way to interpret this age
410 bias is that tissue-specific TE-CREs have on average more neutral fitness effects. Conversely,
411 older and tissue-shared TE-CREs are more likely to be advantageous, fixed by selection, and
412 maintained for longer under purifying selection. Under this model we expect higher TE-CRE
413 turnaround rates (loss and gain) for tissue-specific compared to tissue-shared TE-CREs, which
414 has been described in mammals (Roller et al., 2021). Higher evolutionary turnaround rates of
415 tissue-specific TE-CREs is also expected if tissue- or cell-type specific CREs is 'easier' to evolve
416 than tissue-shared CREs, which has recently been suggested to be the case (Luthra et al., 2022).

417 Since gene regulation is under tissue-specific selection pressure (Berthelot et al., 2018;
418 Brawand et al., 2011), we expect CRE-evolution to be under different selection pressures in
419 different tissues. From mammalian studies we know that purifying selection on gene regulation
420 is stronger in the brain than liver (Wang et al., 2020), hence we expect TE-CRE evolution to
421 reflect this asymmetry in selection pressure. COnsistent with this expectation we find clear
422 tissue differences in TE-CRE numbers (Figure 2 D) and that TE sequences were consistently
423 depleted in highly brain biassed TFBS (Figure 2G). We propose that these results may be related
424 to the evolutionary arms race between genomic 'parasites' and the host, and reflect selection
425 pressure to "avoid" having sequences that function as, or can evolve into CREs that can impact
426 brain-specific gene regulatory networks under strong purifying selection pressure.

427

428 ## TE-CRE evolution in aftermath of the WGD

429 The whole genome duplication in the ancestor of salmonids resulted in large scale gene
430 regulatory rewiring (Lien et al., 2016; Varadharajan et al., 2018). These novel gene regulatory
431 phenotypes have been partly linked to divergent TE-insertions in promoters of gene duplicates
432 (Gillard et al., 2021; Sahlström et al., 2023), but the link between WGD and TE-CRE evolution
433 has remained elusive. One hypothesis is that WGD induce a genomic shock which results in
434 bursts of TE activity (the 'genomic shock' model (McClintock, 1984)), and that these novel TE

435 insertions allow for rapid TE-CRE evolution and rewiring of gene regulatory networks in the
436 initial aftermath of a WGD. Another hypothesis is that relaxed purifying selection in polyploids
437 allows for higher rates of TE accumulation (Baduel et al., 2019), which in turn will lead to higher
438 rates of neutral and nearly-neutral TE-CRE evolution. In this scenario, however, there is no
439 expectation of a temporal link between bursts of TE-activity and bursts of TE-CRE evolution.

440 Our results are more consistent with the 'relaxed selection' model than the 'genomic shock'
441 model, as there was little evidence for a temporal co-occurrence between TE-CRE evolution and
442 WGD (Figure 4, Figure 5C, F). In fact, the TEs with the highest activity following the WGD, the
443 DTTs (Tc1-Mariner superfamily) (Lien et al., 2016), has contributed significantly less to the TE-
444 CRE landscape than expected (Figure 2G) and is also significantly less likely to impact
445 transcription compared to other TEs (Figure 5D, H). This is in line with other studies showing
446 that the DTT superfamily does not contain many TBFSs (Simonti et al., 2017; Zeng et al., 2018).
447 Beyond the DTTs, many individual TEs, including those with CRE-superspreader capabilities,
448 have been active long after the salmonid-specific whole genome duplication (Figure 4E). These
449 include TEs impacting gene regulatory networks (Figure 5) and those enriched for
450 transcriptional modulatory capabilities (RLX, RLC, RST in Figure Figure 6E-G).

451 In conclusion, our results cast doubts about the role TE-activity bursts at the time of the WGD in
452 TE-CRE driven gene regulatory evolution. However, we find that certain TEs have been
453 particularly effective in spreading TE-CREs, and regulating gene transcription, but that many of
454 these TEs remained active long after the initial 'genome shock' following WGD. To further
455 quantify the importance of selection on TE-CRE evolution, a need for a larger comparative
456 approach (Andrews et al., 2023) is warranted.

457

458 # Methods

459 ## TE annotation

460 The TE library (ssal_repeats_v5.1) used to annotate TEs in this study is described in detail in
461 (Richard Minkley, 2018). To generate a TE annotation of the salmon genome (ICSASG v2
462 assembly) we used RepeatMasker version 4.1.2-p1 (Smit et al., 2015) under default settings
463 with the ssal_repeats_v5.1 library. RepeatMasker takes a library of TE consensus sequences and
464 detects whole and fragmented parts of these consensuses across the genome using a BLAST-like
465 algorithm. The output file contains the genomic coordinates of the annotation, and various
466 quality measures such as completeness, and divergence from consensus. The latter measure

467 was used to estimate relative ages of TE activity. TE superfamilies were assigned a three letter
468 tag based on the classifications from Figure 1 in (Wicker et al., 2007). Where there was no
469 obvious categorisation, a literature review was conducted to determine the taxonomic status of
470 a superfamily, and a new tag name introduced based on available letters (so e.g. Nimb is here
471 called RIN as a superfamily of LINE elements).

472 Manual curation of specific TE subfamilies was done following an adapted version of Goubert et
473 al's process (Goubert et al., 2022), under inspiration from Suh (Suh et al., 2018): Using BLASTn
474 (Altschul et al., 1990), we aligned each transposable element consensus to the genome,
475 extracted the twenty best matches and extended them by 2000bp upstream and downstream.
476 We checked the extended matches against the RepBase (Bao et al., 2015) database using
477 BLASTn and xBLAST with standard settings, before we aligned them using MAFFT's 'einsi'
478 variant (Katoh and Standley, 2013) . Then, we inspected these alignments for structural features
479 in BioEdit (Hall, 1999) and, if conservation across the sequence was deemed interesting, in
480 JalView (Waterhouse et al., 2009) . In addition, we ran the TE-Aid package
481 (https://github.com/clemgoub/TE-Aid) on each consensus to help guide curation efforts and
482 check each consensus according to its annotation profile and self-alignment. This helped screen
483 for technical noise such as microsatellite sequences near sites of local annotation enrichment. If
484 the annotating consensus was deemed to be incomplete (i.e. if parts of the extended sequence
485 aligned well outside of the consensus), we used Advanced Consensus Generator
486 (https://www.hiv.lanl.gov/content/sequence/CONSENSUS/AdvCon.html) to generate a new
487 consensus from the most complete of the extracted alignments for classification.

488

489 ## ATAC-seq peak calling

490 To annotate regions of accessible chromatin we used ATAC-seq data from four brains and livers
491 from Atlantic salmon (ENA project number PRJEB38052). The ATAC-seq reads were mapped to
492 the salmon genome assembly (ICSASG v2, refseq ID: GCF_000233375.1) using BWA-MEM.
493 Genrich v.06 (https://github.com/jsh58/Genrich) was then used to call open chromatin regions
494 (also referred to as 'peaks') with default parameters, apart from '-m 20 -j' (minimum mapping
495 quality 20; ATAC-Seq mode). Genrich uses all four replicates to generate peaks, resulting in one
496 set of peaks for each tissue. The summit of each peak is identified as the midpoint of the peak
497 interval with highest significance.

498

499 ## TE-CRE definition

500 To define TE-CREs we combined the ATAC-seq peak set with our TE annotations and classified
501 an ATAC-seq peak as a TE-CRE if the peak summit is inside a TE-annotation. TE-CREs were
502 defined as shared between tissues if (i) the brain ATAC-seq peak summit was within the liver
503 ATAC-seq peak interval and (ii) both the liver and brain peak summits are inside the same TE
504 annotation.

505

506 ## Defining genomic context

507 Based on the NCBI gene annotation (refseq ID: GCF_000233375.1), each part of the genome was
508 assigned as promoter, exon, intron or intergenic. For Figure 1D the promoter was defined as
509 1000 bp upstream to 200 bp downstream of each transcription start site (TSS). Gene
510 annotations can overlap, e.g. because of multiple transcript isoforms, so overlapping
511 annotations were merged by prioritising promoter > exon > intron > intergenic. For TE-CREs
512 (Figure 2E and 3D-G) each peak was classified as promoter if the summit is less than 500bp
513 upstream or downstream from start of gene (i.e. first TSS per gene) or intergenic if summit is
514 more than 500bp from any gene (exon and intron TE-CREs are not specifically mentioned).
515

516 ## Identification of TE subfamilies enriched in open chromatin

517 To identify TE subfamilies which had contributed more to TE-CREs than expected by chance we
518 counted the number of ATAC-seq peak summits that are inside an annotated TE for each
519 subfamily and compare that with the total number of bases covered by that TE subfamily
520 genome wide. The enrichment value for each subfamily was calculated as the proportion of
521 summits in TEs divided by the proportion of basepairs in the genome that is annotated as TE.
522 Subfamilies with less than 500 insertions were excluded. We defined TE subfamilies enriched in
523 open chromatin as those containing more ATAC-seq peak summits than chance (binomial,
524 $p<0.05$), either in the ATAC-seq peak set from liver, brain, or in both tissues.

525

526 ## Estimating evolutionary timing of TE activity

527 To temporal activity of TEs, we used the sequence divergence between insertions and
528 consensus TE. Divergence from consensus for each insertion was extracted from RepeatMasker
529 software output (Smit et al., 2015).

530

531 Transcription factor binding and footprinting

532 We annotated transcription factor binding sites (TFBS) in two different ways. First, we used
533 FIMO (Grant et al., 2011) on the whole genome with the JASPAR CORE vertebrates non-
534 redundant motif database (https://jaspar.genereg.net). Secondly, we used the TOBIAS software,
535 which uses a FIMO-like TFBS scan but also integrate ATAC-seq data to detect signals of local TF
536 occupancy (i.e. a sudden, local drop in chromatin accessibility) and assigns each TFBS motif a
537 "bound" or "not bound" status. We used the TOBIAS software to estimate a single genome wide
538 TF binding score for each TFBS in liver and brain tissues (Bentsen et al., 2020).

539

540 Testing TE-TFBS enrichment

541 For each TE subfamily we counted the number of overlaps between each jaspar-TFBS motif (i.e.
542 the entire motif within the annotated TE) and calculated these numbers for TFBSs "bound" by
543 TFs and those " not bound" by TFs as according to the TOBIAS software (Bentsen et al., 2020)
544 results. We then did the same counting for all TFBS instances outside the particular TE
545 subfamily in question, and used this 2*2 contingency table (Table 1) in a Fisher exact test in R
546 using the fisher.test() function in R (R Core Team, 2021).

547 **Table 1.** Example of a 2*2 contingency table for fisher exact tests for TFBS-TE associations.

	TE subfamily	All other genomic positions
Not bound TFBS	x	y
Bound TFBS	z	w

548

549 Co-expression analysis

550 We used two RNA-seq expression data sets to analyse the effect of TE-CREs on gene expression:
551 (1) A liver data set comprising 112 samples spanning different diets and life stages in fresh-
552 water (Gillard et al., 2018) and (2) a tissue atlas comprising 13 different tissues (Lien et al.,
553 2016).

554

555 TE-CREs in liver, brain or both (the ATAC-seq peak summits of the liver and brain TE-CREs
556 reciprocally overlapped the peak in the other tissue) were assigned to genes with the closest
557 transcription start site (TSS). For each TE consensus sequence, we computed the network

558 density of the associated genes (mean pairwise Pearson correlation). False Discovery Rate
559 (FDR)-corrected p-values were obtained by comparing these network densities to those of
560 randomly selected genes. We ran 100 000 simulations drawing the same number of genes,
561 containing the same number of WGD-derived duplicates (which are often co-expressed), as
562 found in the original data. Effect sizes were calculated as the number of standard deviations
563 away from the mean of randomised network densities.
564

565 ## Massive parallel reporter assay

566 Transcriptional regulatory potential of TE-CREs in Atlantic salmon was assessed using ATAC-
567 STARR-seq as previously described in Wang et al. (2018). We used the pSTARR-seq reporter
568 plasmid with the core promoter of Atlantic salmon elongation factor 1 alpha, EF1α
569 (NC_027326.1: 7785458-7785702) instead of the super core promoter 1 (SCP1) originally
570 adapted in human cells (Arnold et al., 2013). ATAC DNA fragments were extracted from Atlantic
571 salmon liver cell nuclei following the OmniATAC protocol (Corces et al., 2017). A clean-up step
572 was performed using Qiagen MinElute PCR purification kit and PCR-amplified using NEBNext
573 Ultra Q5 DNA polymerase master mix (New England Biolabs®) with forward primer (5'-
574 TAGAGCATGCACC GGCAAGCAGAAGACGGCATACGAGAT[N10]ATGTCTCGTGGGCTCGGAGATGT-
575 3', where N10 corresponds to a random 10 nucleotide i7 barcode sequence) and reverse primer
576 (Rv:5'-GGCCGAATTCGTCGATCGTCGGCAGCGTCAGATGTG-3'). Thermo cycling conditions were
577 72 ^{0}C for 5 min, 98 ^{0}C for 30 sec, 8 cycles of 98 ^{0}C for 10 sec, 63 ^{0}C for 30 sec and 72 ^{0}C for 1 min.
578 PCR products were purified using Qiagen MinElute PCR purification kit and size-selected (~30-
579 280 bp) using Ampure XP beads (Beckman Coulter). Reporter plasmid libraries were made by
580 cloning amplified ATAC fragments into AgeI-HF- and SalI-HF-linearized pSTARR-seq plasmid
581 using InFusion HD cloning kit (Takara) and then propagated in MegaX DH10B T1R
582 electrocompetent bacteria. Plasmids were isolated using the NucleoBond® PC 2000 Mega kit
583 (MACHEREY-NAGEL). An aliquot of plasmid library was PCR-amplified with i5 and i7 primers
584 and sequenced on Novaseq (150 bp Paired-end) and aligned to salmon genome to ensure
585 sufficient complexity and proportions of cloned fragments within open chromatin region.
586 Plasmid library was electroporated into primary salmon hepatocytes as previously described
587 (Datsomor et al., 2022). Total RNA was isolated 24 hours post-transfection using the Qiagen
588 RNeasy Midi columns. Poly A+ RNA from total RNA was extracted using the mRNA isolation kit
589 (Roche). Remaining genomic DNA in isolated mRNA were digested with Turbo DNase (Thermo
590 Fisher). Complementary DNA (cDNA) from mRNA was generated using the Superscript III
591 Reverse transcriptase (Thermo Fisher) with a gene-specific primer (5'-
592 CAAACTCATCAATGTATCTTATCATG-3). Sequencing-ready libraries from cDNA and the input

593 (reporter plasmid library) were prepared as previously described by Wang et al. (2018) and
594 Tewhey et al. (Tewhey et al., 2016).

595 Sequenced reads were mapped to the salmon genome assembly (ICSASG v2, refseq ID:
596 GCF_000233375.1) using BWA-MEM. The number of read-pairs mapped to each unique location
597 was counted. Each unique location, i.e. having a specific start and end, was assumed to come
598 from a unique fragment. These counts were fed into DESeq2 using the DNA (input plasmid
599 library) as control and contrasted with the RNA (cDNA) samples. Fragments with significant
600 RNA to DNA ratio were used to define fragments with significant regulatory activity. Prior to
601 DESeq2 the fragment counts were split into bins by length.

602

603 Data availability and code

604 We produced plots using base R's (R Core Team, 2021) plot function, as well as the packages
605 ggplot2 (Wickham, 2016) and cowplot. Both the Tidyverse (Wickham et al., 2019) and
606 data.table packages were used for analysis, summary statistics and data management. All scripts
607 to reproduce figures and analyses are available at GitLab repo: https://gitlab.com/sandve-
608 lab/TE-CRE. Raw data used in the analyses can be downloaded here:
609 https://arken.nmbu.no/~lagr/share/TE-CRE-DATA.zip.

610 Raw sequencing data from the ATAC-STARR-seq experiment will be deposited to ENA
611 (accession number PRJEB71627) prior to publication. Until then, this data can be acquired from
612 corresponding authors upon request.

613

614 Acknowledgements and funding

615 This research was funded by NMBU and the Norwegian Research Council through the projects
616 Transpose (275310), Rewired (274669), and DigiSal (248792). We thank Sigbjørn Lien for
617 comments on earlier versions of the manuscript.

618

619 Author contributions

620 SRS and TRH conceived the study. SRS and TRH acquired funding. AD performed all lab
621 experiments related to the massive parallel reporter assays. ØM, LG, SRS, and TRH performed
622 analyses. ØM, LG, TRH, and SRS drafted the manuscript. All authors took part in critical

discussions of various aspects of lab-work and/or analytical approaches relevant to their expertise. All authors critically reviewed the manuscript.

Supplementary material

Table S1: Curation notes and classification. Every CRE-superspreader TE-consensus has been inspected manually as per the procedure in Materials and Methods. 'consensus_TE' is the ID of the annotating consensus in question, 'original_annotation' is the automatic classification, 'manual_curation_three_letter_code' is the post-curation three-letter ID.

Bibliography

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Andrews G, Fan K, Pratt HE, Phalke N, Zoonomia Consortium§, Karlsson EK, et al. Mammalian evolution of human cis-regulatory elements and transcription factor binding sites. Science 2023;380:eabn7930. https://doi.org/10.1126/science.abn7930.

Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature 2020;587:246–51. https://doi.org/10.1038/s41586-020-2871-y.

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 2013;339:1074–7. https://doi.org/10.1126/science.1232542.

Baduel P, Quadrana L, Hunter B, Bomblies K, Colot V. Relaxed purifying selection in autopolyploids drives transposable element over-accumulation which provides variants for local adaptation. Nat Commun 2019;10:5818. https://doi.org/10.1038/s41467-019-13730-0.

Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 2015;6:11. https://doi.org/10.1186/s13100-015-0041-9.

Bartolomé C, Maside X, Charlesworth B. On the abundance and distribution of transposable elements in the genome of Drosophila melanogaster. Mol Biol Evol 2002;19:926–37. https://doi.org/10.1093/oxfordjournals.molbev.a004150.

Bentsen M, Goymann P, Schultheis H, Klee K, Petrova A, Wiegandt R, et al. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. Nat Commun 2020;11:4267. https://doi.org/10.1038/s41467-020-18035-1.

Bergthorsson U, Sheeba CJ, Konrad A, Belicard T, Beltran T, Katju V, et al. Long-term

657 experimental evolution reveals purifying selection on piRNA-mediated control of transposable
658 element expression. BMC Biol 2020;18:162. https://doi.org/10.1186/s12915-020-00897-y.

659 Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P. Complexity and conservation of regulatory
660 landscapes underlie evolutionary resilience of mammalian gene expression. Nat Ecol Evol
661 2018;2:152–63. https://doi.org/10.1038/s41559-017-0377-2.

662 Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you
663 should know about transposable elements. Genome Biol 2018;19:199.
664 https://doi.org/10.1186/s13059-018-1577-z.

665 Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, et al. Evolution of the mammalian
666 transcription factor binding repertoire via transposable elements. Genome Res 2008;18:1752–
667 62. https://doi.org/10.1101/gr.080663.108.

668 Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, et al. The evolution of gene
669 expression levels in mammalian organs. Nature 2011;478:343–8.
670 https://doi.org/10.1038/nature10532.

671 Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin
672 for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and
673 nucleosome position. Nat Methods 2013;10:1213–8. https://doi.org/10.1038/nmeth.2688.

674 Chaves C, Bruinstroop E, Refetoff S, Yen PM, Anselmo J. Increased Hepatic Fat Content in
675 Patients with Resistance to Thyroid Hormone Beta. Thyroid 2021;31:1127–34.
676 https://doi.org/10.1089/thy.2020.0651.

677 Choudhary MNK, Quaid K, Xing X, Schmidt H, Wang T. Widespread contribution of transposable
678 elements to the rewiring of mammalian 3D genomes. Nat Commun 2023;14:634.
679 https://doi.org/10.1038/s41467-023-36364-9.

680 Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts
681 to benefits. Nat Rev Genet 2017;18:71–86. https://doi.org/10.1038/nrg.2016.139.

682 Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, et al. An
683 improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues.
684 Nat Methods 2017;14:959–62. https://doi.org/10.1038/nmeth.4396.

685 Cosby RL, Chang N-C, Feschotte C. Host-transposon interactions: conflict, cooperation, and
686 cooption. Genes Dev 2019;33:1098–116. https://doi.org/10.1101/gad.327312.119.

687 Cosby RL, Judd J, Zhang R, Zhong A, Garry N, Pritham EJ, et al. Recurrent evolution of vertebrate
688 transcription factors by transposase capture. Science 2021;371.
689 https://doi.org/10.1126/science.abc6405.

690 Datsomor AK, Wilberg R, Torgersen JS, Sandve SR, Harvey TN. Efficient transfection of Atlantic
691 salmon primary hepatocyte cells for functional assays and gene editing. BioRxiv 2022.

692 Diehl AG, Ouyang N, Boyle AP. Transposable elements contribute to cell and species-specific
693 chromatin looping and gene regulation in mammalian genomes. Nat Commun 2020;11:1796.
694 https://doi.org/10.1038/s41467-020-15520-5.

695 Elisaphenko EA, Kolesnikov NN, Shevchenko AI, Rogozin IB, Nesterova TB, Brockdorff N, et al. A
696 dual origin of the Xist gene from a protein-coding gene and a set of transposable elements. PLoS
697 ONE 2008;3:e2521. https://doi.org/10.1371/journal.pone.0002521.

698 Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. Annu Rev
699 Genet 2007;41:331–68. https://doi.org/10.1146/annurev.genet.40.110405.090448.

700 Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet
701 2008;9:397–405. https://doi.org/10.1038/nrg2337.

702 Fueyo R, Judd J, Feschotte C, Wysocka J. Roles of transposable elements in the regulation of
703 mammalian transcription. Nat Rev Mol Cell Biol 2022;23:481–97.
704 https://doi.org/10.1038/s41580-022-00457-y.

705 Gillard G, Harvey TN, Gjuvsland A, Jin Y, Thomassen M, Lien S, et al. Life-stage-associated
706 remodelling of lipid metabolism regulation in Atlantic salmon. Mol Ecol 2018;27:1200–13.
707 https://doi.org/10.1111/mec.14533.

708 Gillard GB, Grønvold L, Røsæg LL, Holen MM, Monsen Ø, Koop BF, et al. Comparative regulomics
709 supports pervasive selection on gene dosage following whole genome duplication. Genome Biol
710 2021;22:103. https://doi.org/10.1186/s13059-021-02323-0.

711 Goodier JL, Davidson WS. Tc1 transposon-like sequences are widely distributed in salmonids. J
712 Mol Biol 1994;241:26–34. https://doi.org/10.1006/jmbi.1994.1470.

713 Goubert C, Craig RJ, Bilat AF, Peona V, Vogan AA, Protasio AV. A beginner's guide to manual
714 curation of transposable elements. Mob DNA 2022;13:7. https://doi.org/10.1186/s13100-021-
715 00259-7.

716 Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics
717 2011;27:1017–8. https://doi.org/10.1093/bioinformatics/btr064.

718 Hall TA. BioEdit: A User-Friendly Biological Sequence Alignment Editor and Analysis Program
719 for Windows 95/98/NT. Nucleic Acids Symposium Series 1999;41:95–8.

720 Hollister JD, Gaut BS. Epigenetic silencing of transposable elements: a trade-off between
721 reduced transposition and deleterious effects on neighboring gene expression. Genome Res
722 2009;19:1419–28. https://doi.org/10.1101/gr.091678.109.

723 Karttunen K, Patel D, Xia J, Fei L, Palin K, Aaltonen L, et al. Transposable elements as tissue-
724 specific enhancers in cancers of endodermal lineage. Nat Commun 2023;14:5313.
725 https://doi.org/10.1038/s41467-023-41081-4.

726 Kashkush K, Feldman M, Levy AA. Transcriptional activation of retrotransposons alters the
727 expression of adjacent genes in wheat. Nat Genet 2003;33:102–6.
728 https://doi.org/10.1038/ng1063.

729 Kashkush K, Feldman M, Levy AA. Gene loss, silencing and activation in a newly synthesized
730 wheat allotetraploid. Genetics 2002;160:1651–9.
731 https://doi.org/10.1093/genetics/160.4.1651.

732 Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements
733 in performance and usability. Mol Biol Evol 2013;30:772–80.
734 https://doi.org/10.1093/molbev/mst010.

735 Keene MA, Corces V, Lowenhaupt K, Elgin SC. DNase I hypersensitive sites in Drosophila
736 chromatin occur at the 5' ends of regions of transcription. Proc Natl Acad Sci USA 1981;78:143–
737 6. https://doi.org/10.1073/pnas.78.1.143.

738 Kraitshtein Z, Yaakov B, Khasdan V, Kashkush K. Genetic and epigenetic dynamics of a
739 retrotransposon after allopolyploidization of wheat. Genetics 2010;186:801–12.
740 https://doi.org/10.1534/genetics.110.120790.

741 Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, Chan Y-S, et al. Transposable elements have
742 rewired the core regulatory network of human embryonic stem cells. Nat Genet 2010;42:631–4.
743 https://doi.org/10.1038/ng.600.

744 Langmüller AM, Nolte V, Dolezal M, Schlötterer C. The genomic distribution of transposable
745 elements is driven by spatially variable purifying selection. Nucleic Acids Res 2023;51:9203–13.
746 https://doi.org/10.1093/nar/gkad635.

747 Lau HH, Ng NHJ, Loo LSW, Jasmen JB, Teo AKK. The molecular functions of hepatocyte nuclear
748 factors - In and beyond the liver. J Hepatol 2018;68:1033–48.
749 https://doi.org/10.1016/j.jhep.2017.11.026.

750 Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, et al. The Atlantic salmon genome
751 provides insights into rediploidization. Nature 2016;533:200–5.
752 https://doi.org/10.1038/nature17164.

753 Luthra I, Chen XE, Jensen C, Rafi AM, Salaudeen AL, de Boer CG. Biochemical activity is the
754 default DNA state in eukaryotes. BioRxiv 2022. https://doi.org/10.1101/2022.12.16.520785.

755 Marburger S, Alexandrou MA, Taggart JB, Creer S, Carvalho G, Oliveira C, et al. Whole genome
756 duplication and transposable element proliferation drive genome expansion in Corydoradinae
757 catfishes. Proc Biol Sci 2018;285. https://doi.org/10.1098/rspb.2017.2732.

758 McClintock B. The significance of responses of the genome to challenge. Science 1984;226:792–
759 801. https://doi.org/10.1126/science.15739260.

760 McGhee JD, Wood WI, Dolan M, Engel JD, Felsenfeld G. A 200 base pair region at the 5' end of the
761 chicken adult beta-globin gene is accessible to nuclease digestion. Cell 1981;27:45–55.
762 https://doi.org/10.1016/0092-8674(81)90359-7.

763 Nishihara H. Retrotransposons spread potential cis-regulatory elements during mammary gland
764 evolution. Nucleic Acids Res 2019;47:11551–62. https://doi.org/10.1093/nar/gkz1003.

765 Noshay JM, Marand AP, Anderson SN, Zhou P, Mejia Guerra MK, Lu Z, et al. Assessing the
766 regulatory potential of transposable elements using chromatin accessibility profiles of maize
767 transposons. Genetics 2021;217:1–13. https://doi.org/10.1093/genetics/iyaa003.

768 Pehrsson EC, Choudhary MNK, Sundaram V, Wang T. The epigenomic landscape of transposable
769 elements across normal human development and anatomy. Nat Commun 2019;10:5640.

770 https://doi.org/10.1038/s41467-019-13555-x.

771 Qin S, Jin P, Zhou X, Chen L, Ma F. The role of transposable elements in the origin and evolution
772 of micrornas in human. PLoS ONE 2015;10:e0131365.
773 https://doi.org/10.1371/journal.pone.0131365.

774 Richard Minkley D. Transposable Elements in the Salmonid Genome. Master thesis. University of
775 Victoria, 2018.

776 R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R
777 Foundation for Statistical Computing; 2021.

778 Rizzon C, Martin E, Marais G, Duret L, Ségalat L, Biémont C. Patterns of selection against
779 transposons inferred from the distribution of Tc1, Tc3 and Tc5 insertions in the mut-7 line of
780 the nematode Caenorhabditis elegans. Genetics 2003;165:1127–35.
781 https://doi.org/10.1093/genetics/165.3.1127.

782 Roller M, Stamper E, Villar D, Izuogu O, Martin F, Redmond AM, et al. LINE retrotransposons
783 characterize mammalian tissue-specific and evolutionarily dynamic regulatory regions. Genome
784 Biol 2021;22:62. https://doi.org/10.1186/s13059-021-02260-y.

785 Ronfort J. The mutation load under tetrasomic inheritance and its consequences for the
786 evolution of the selfing rate in autotetraploid species. Genet Res 1999;74:31–42.
787 https://doi.org/10.1017/S0016672399003845.

788 Sahlström HM, Datsomor AK, Monsen Ø, Hvidsten TR, Sandve SR. Functional validation of
789 transposable element-derived cis-regulatory elements in Atlantic salmon. G3 (Bethesda)
790 2023;13. https://doi.org/10.1093/g3journal/jkad034.

791 Shao F, Han M, Peng Z. Evolution and diversity of transposable elements in fish genomes. Sci Rep
792 2019;9:15399. https://doi.org/10.1038/s41598-019-51888-1.

793 Simonti CN, Pavlicev M, Capra JA. Transposable Element Exaptation into Regulatory Regions Is
794 Rare, Influenced by Evolutionary Age, and Subject to Pleiotropic Constraints. Mol Biol Evol
795 2017;34:2856–69. https://doi.org/10.1093/molbev/msx219.

796 Smit AFA, Hubley R, Green P. RepeatMasker. 2015.

797 Suh A, Smeds L, Ellegren H. Abundant recent activity of retrovirus-like retrotransposons within
798 and among flycatcher species implies a rich source of structural variation in songbird genomes.
799 Mol Ecol 2018;27:99–111. https://doi.org/10.1111/mec.14439.

800 Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of transposable
801 elements to the innovation of gene regulatory networks. Genome Res 2014;24:1963–76.
802 https://doi.org/10.1101/gr.168872.113.

803 Sundaram V, Choudhary MNK, Pehrsson E, Xing X, Fiore C, Pandey M, et al. Functional cis-
804 regulatory modules encoded by mouse-specific endogenous retrovirus. Nat Commun
805 2017;8:14550. https://doi.org/10.1038/ncomms14550.

806 Sundaram V, Wysocka J. Transposable elements as a potent source of diverse cis-regulatory

807 sequences in mammalian genomes. Philos Trans R Soc Lond B Biol Sci 2020;375:20190347.
808 https://doi.org/10.1098/rstb.2019.0347.

809 Tao R, Xiong X, DePinho RA, Deng C-X, Dong XC. Hepatic SREBP-2 and cholesterol biosynthesis
810 are regulated by FoxO3 and Sirt6. J Lipid Res 2013;54:2745–53.
811 https://doi.org/10.1194/jlr.M039339.

812 Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, et al. Direct Identification of Hundreds
813 of Expression-Modulating Variants using a Multiplexed Reporter Assay. Cell 2016;165:1519–29.
814 https://doi.org/10.1016/j.cell.2016.04.027.

815 Varadharajan S, Sandve SR, Gillard GB, Tørresen OK, Mulugeta TD, Hvidsten TR, et al. The
816 Grayling Genome Reveals Selection on Gene Expression Regulation after Whole-Genome
817 Duplication. Genome Biol Evol 2018;10:2785–800. https://doi.org/10.1093/gbe/evy201.

818 Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, et al. Primate-specific endogenous
819 retrovirus-driven transcription defines naive-like stem cells. Nature 2014;516:405–9.
820 https://doi.org/10.1038/nature13804.

821 Wang X, He L, Goggin SM, Saadat A, Wang L, Sinnott-Armstrong N, et al. High-resolution genome-
822 wide functional dissection of transcriptional regulatory regions and nucleotides in human. Nat
823 Commun 2018;9:5380. https://doi.org/10.1038/s41467-018-07746-1.

824 Wang Z-Y, Leushkin E, Liechti A, Ovchinnikova S, Mößinger K, Brüning T, et al. Transcriptome
825 and translatome co-evolution in mammals. Nature 2020;588:642–7.
826 https://doi.org/10.1038/s41586-020-2899-z.

827 Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a multiple
828 sequence alignment editor and analysis workbench. Bioinformatics 2009;25:1189–91.
829 https://doi.org/10.1093/bioinformatics/btp033.

830 Wei J, Ran G, Wang X, Jiang N, Liang J, Lin X, et al. Gene manipulation in liver ductal organoids by
831 optimized recombinant adeno-associated virus vectors. J Biol Chem 2019;294:14096–104.
832 https://doi.org/10.1074/jbc.RA119.008616.

833 Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification
834 system for eukaryotic transposable elements. Nat Rev Genet 2007;8:973–82.
835 https://doi.org/10.1038/nrg2165.

836 Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the
837 tidyverse. JOSS 2019;4:1686. https://doi.org/10.21105/joss.01686.

838 Wickham H. ggplot2 - Elegant Graphics for Data Analysis. 2nd ed. Cham: Springer International
839 Publishing; 2016. https://doi.org/10.1007/978-3-319-24277-4.

840 Yerra VG, Drosatos K. Specificity Proteins (SP) and Krüppel-like Factors (KLF) in Liver
841 Physiology and Pathology. Int J Mol Sci 2023;24. https://doi.org/10.3390/ijms24054682.

842 Zeng L, Pederson SM, Kortschak RD, Adelson DL. Transposable elements and gene expression
843 during the evolution of amniotes. Mob DNA 2018;9:17. https://doi.org/10.1186/s13100-018-
844 0124-5.

845

846