Transposable elements drive the evolution of metazoan zinc fingers

Jonathan N. Wells[1], Ni-Chen (Sylvia) Chang[1], John McCormick[1], Nathalie Ramos[2], Caitlyn Coleman[3], Bozhou Jin[1], Cédric Feschotte[1]

1. Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14850, USA.
2. Mt Sinai?
3. Florida?

Abstract

The Cys2His2 zinc finger genes comprise the largest family of transcription factors in animals, with a multitude of unexplored subfamilies, many of which have arisen via the acquisition of novel accessory domains. Despite their clear importance, our understanding of the function and evolution of most of these families is limited, primarily being restricted to the well-studied KRAB-ZnFs in tetrapods, which repress transposable elements. In this work, we explore the relationship between ZnFs and TEs in a survey of ~3000 metazoan genomes, and characterize a novel subfamily of ZnFs found in cyprinid fish – the most species-rich family of vertebrates. This gene family expansion coincides with the acquisition of a short protein sequence, termed the **Fi**sh **N**-terminal **Z**NF-associated domain (FiNZ).

Introduction

Cys2-His2 zinc finger (ZnF) genes are a family of transcription factors found in all major eukaryotic supergroups. In metazoans they have undergone dramatic, lineage-specific expansions in gene copy number, as well as an increase in the diversity of DNA sequences that they recognize[1,2]. As a result, ZnFs are presently the largest and most dynamically evolving transcription factor family in most metazoan organisms, including ourselves. In extant species, their copy number spans three orders of magnitude, from close to zero in most nematodes, to hundreds in many vertebrate, arthropod and mollusk species. Their lineage-specificity can be seen through comparisons of closely related organisms. For example, while we share ~99% of our genes with mice, less than 30% of human KRAB-ZnFs are clear one to one orthologs with those of mice. Genomic conflict likely plays an important role in explaining the diversity and rapid evolution of ZnFs, and is certainly true of the best-studied ZnF subfamily, namely the Kruppel-associated box (KRAB)-ZnFs, which are typically involved in repressing TE activity.

KRAB-ZnFs were first described in 1991 and were recognized as being broadly conserved across tetrapods[3]. However, their function was unknown until it was observed that the number of ZnF domains correlated strongly with the retroelement copy number across several vertebrate genomes[4]. Since then, numerous KRAB-ZnF genes have been shown to be engaged in arms races with transposable elements (TEs)[5,6], and in a landmark paper documenting the binding specificity of hundreds of human KRAB-ZnFs, it was found that the majority target TE-derived sequences. This strongly suggests that TE repression is a driving force behind the diversification of KRAB-ZnFs in tetrapods. On the other hand, members of another ZnF family, the ZAD-ZnFs, have diverse roles in heterochromatin organization that are seemingly unrelated to TE repression. Despite apparent differences in the primary function of most ZAD- and KRAB-ZnF genes, both groups share key features that are broadly characteristic of all metazoan ZnF genes thus far studied, namely: rapid turnover and sequence evolution, involvement in establishing or maintaining heterochromatin, and expression during early embryogenesis.

Transposable elements (TEs) are selfish genetic elements that replicate within genomes, and in most eukaryotes comprise between 5% and 85% of the host genome. Their existence poses myriad challenges for their hosts, ranging from genomic instability caused by ectopic recombination between TE copies, to dysregulation of gene expression caused by the fact that many TEs carry their own promoters and regulatory machinery. As a result of these threats, metazoans possess a variety of defense systems which control the spread of TEs, including the well-studied piRNA system, tRFs, the HUSH complex, and the previously mentioned KRAB-ZnFs; others undoubtedly remain to be discovered. However, while individual TE insertions are most likely to be neutral at best, in aggregate, TEs are an enormously important source of genetic variation. In primates, most novel regulatory elements are TE-derived, and across eukaryotes there are countless examples of TEs being coopted for the benefit of the host. Finally, while provocative, it has also been suggested that many hallmarks of eukaryotic gene expression machinery have their origins in host defenses against TEs[7,8].

In this work, we investigate the hypothesis that repression of TEs is a driving force underlying the evolution and functional elaboration of ZnF families across all metazoans. Through a survey of ~3000 metazoan genome assemblies, we test whether the number of ZnF open reading frames is positively correlated with the number of reverse transcriptase domains – a protein which is a defining component of all autonomously replicating retroelements. Using in silico predictions of ZnF binding specificity, we test whether ZnFs recognize repetitive sequences. In the latter half of this work, we focus on a specific subfamily of ZnFs found in cyprinid fish – the largest and most species rich group of vertebrates.

Results

Annotation of ZnFs and TEs in metazoan genomes

ZnFs are challenging to annotate due to their highly repetitive nature and restricted patterns of mRNA expression and in even in well annotated genome assemblies, the true number of functional ZnF genes is systematically underestimated[9,10]. To avoid biases caused by incomplete gene annotation, we used HMMER searches to identify ZnF open reading frames (ORFs), here defined as any stretch of protein coding sequence, regardless of whether bounded by start- or stop-codons. To avoid capturing genes with standalone ZnF domains, we restricted our search to ORFs with at least four tandemly repeated domains. Using the same process, we also carried out searches for reverse transcriptase, using the counts for these as a proxy for retroelement abundance in each genome. These searches revealed extensive variation across species, from thousands of ZnF ORFs in some vertebrate and mollusk species, to fewer than ten in many nematode worms (Fig 1A).

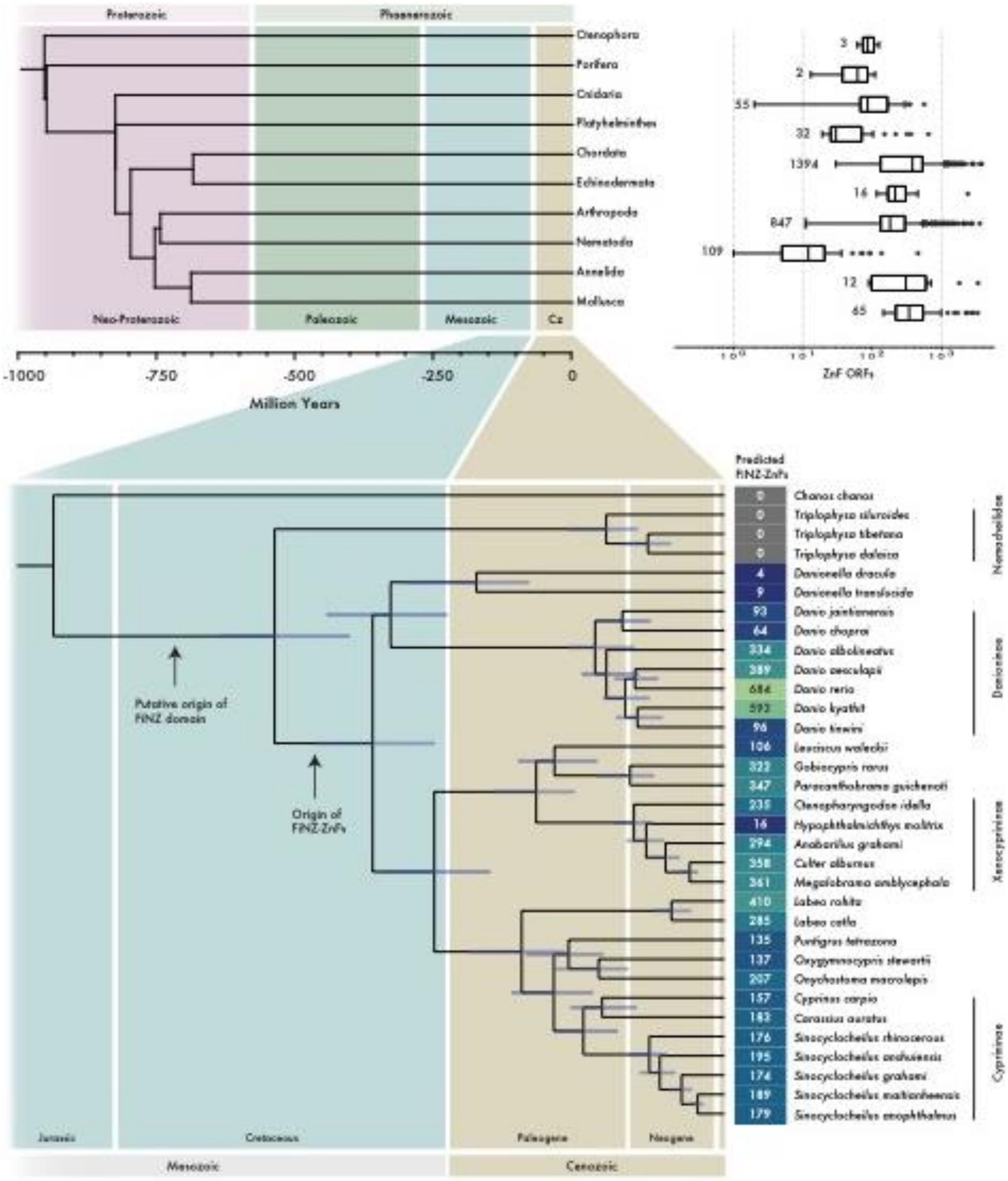


Figure 1. Annotation of ZnF ORFs and full genes in metazoan genome assemblies.
A)

Since estimates of ZnF and retroelement counts could be biased by genome assembly quality, we calculated the correlation between these estimates and scaffold N50, a measure of genome assembly quality (Supp Fig. 1). In both cases, this revealed only weak correlations, reassuring us that using a minimum scaffold N50 of 50kbp (CHECK) was sufficient to avoid serious assembly quality biases in our counts of ZnFs and retroelements.

In addition to our automated counts of ZnF open reading frames, we selected cyprinid fish for more detailed gene annotations. Cyprinids, which include the widely used model organism *Danio rerio* and many economically and culturally important carp species, are the most species-rich family of vertebrates, with approximately 1500 extant members REF. They possess a large subfamily of ZnF genes which are defined by the presence of a characteristic amino-terminal protein sequence, termed the **Fi**sh **N**-terminal **Z**NF-associated domain (FiNZ). FiNZ-ZnFs are present in several hundred copies in zebrafish, almost all of which are located on the long arm of chromosome four, an unusual genomic region that is enriched in TEs, ZnFs and immunity genes, as well as a wide variety of small RNAs[11,12].

Using an iterative approach, we annotated putative FiNZ-ZNF genes using a combination of blast, profile-HMM searches with HMMER[13] and Augustus-PPX[14,15]. The FiNZ domain itself (Supp. Fig. X) is a short, 28 amino acid sequence enriched in negatively charged glutamic acid residues (p = 3e-06, relative to sequences in the SwissProt51 database[16]). The typical structure of FiNZ-ZnF genes resembles that of KRAB-ZnFs, with the FiNZ domain being contained on a single exon upstream of a second exon consisting of an array of ZnF domains. However, the number of copies of the FiNZ domain contained in the first coding exon varies between species – in zebrafish, for example, the FiNZ exon almost always contains a single, stand-alone copy, whereas in goldfish it contains up to four tandemly repeated copies.

Finally, to verify that our ZnF ORF counts are good proxies for the true number of ZnF genes in a species, we compared them to both our FiNZ-ZnF gene annotations, and refined estimates of KRAB-ZnF copy number in tetrapods. In both cases, we found good agreement between the ORF counts and gene annotations (respectively, Spearman's rho = 0.75, p-value = 5.1e-07; rho = 0.91, p-value = 2.6e-59; Supp. Fig. 2). In both cases, the relationship is linear, but ORF counts grow faster than gene counts due to factors such as multiple exons per gene and the presence of pseudogenes or gene fragments.

Correlation between TEs and retroelements

Early evidence for a relationship between TEs and ZnF genes was observed by Thomas and Schneider, who reported a strong correlation between ZnF and retroelement copy number in a set of X vertebrate species[4]. We sought to reproduce this experiment, making use of the more than 3,000 metazoan genome assemblies that are available today. However, a major challenge with this approach is to address the issue of autocorrelation produced by the non-independence of phylogenetically related species. Since there is currently no well-established phylogenetic tree that includes all sequenced metazoan genomes, we opted to select a single species per taxonomic family, leveraging the fact that both ZnF and TE turnover is typically very rapid, so that even closely related species often have largely independent sets of both[6,17].

With this reduced set of 828 genomes, we calculated the correlation between the estimated copy number of ZnF and retroelement ORFs. This revealed positive correlations across all the major animal phyla with at least 15 representative species, with the exception of nematode worms (Fig. 2A). Within smaller taxonomic groups, this relationship holds in the large majority of cases (Supp Table X), with some notable exceptions. For example, birds have secondarily lost most KRAB-ZnFs[6], such that the median number of ZnF ORFs is a quarter of that of other chordates (121 vs. 489, respectively) and there is no correlation between ZnF ORF and retroelement copy number (Supp. Fig. XA). Similarly, there is no correlation in Dipteran flies, which possess ZAD-ZnFs – these have been studied in some depth in

Drosophila melanogaster and have diverse roles in the organization of heterochromatin, but do not appear to specifically target TEs for silencing[18–20].

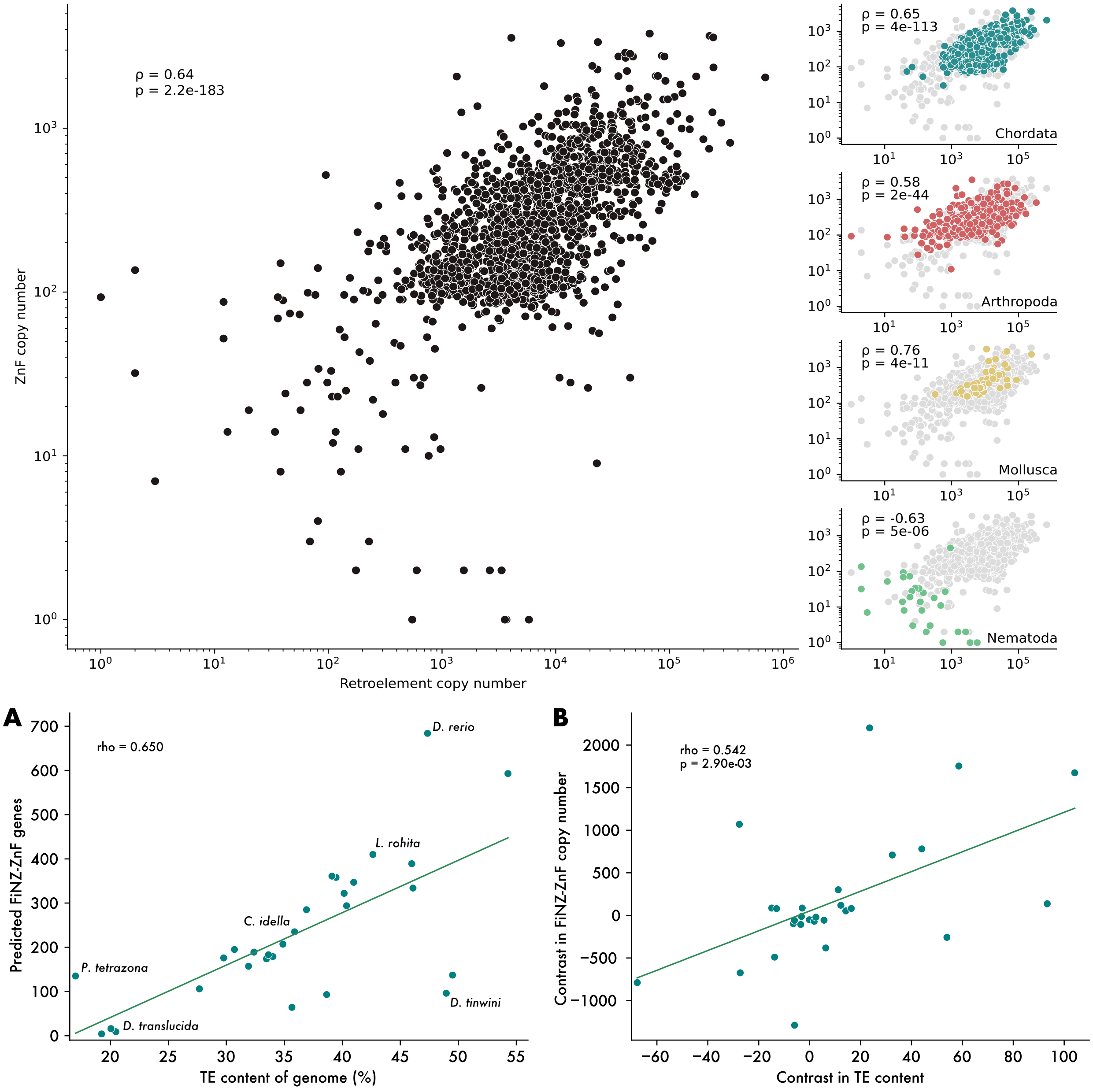


Figure 2. ZnF copy number correlates with genomic TE content

To better control for phylogenetic inter-dependence, we focused on the FiNZ-ZnF family in cyprinid fish, making use of our refined gene annotations and improved estimates of genomic TE content (see methods). Using these more accurate annotations, we find that the correlation between ZnFs and TEs persists (Fig. 2B). Importantly, this relationship holds when explicitly controlling for species relatedness by comparing phylogenetically independent contrasts in FiNZ-ZnF copy number and TE content (Fig. 2C).

Two simple explanations for the correlation between TEs and ZnFs are: one, that ZnFs specifically interact with TEs; or two, that TE content of the genome has a general effect on rates of gene duplication. This latter point is possible as TEs are thought to influence rates of ectopic recombination in a length and copy number-dependent fashion. To explore this possibility, we estimated copy number for mammalian olfactory receptors, using the same procedure as for ZnF ORFs. Olfactory receptors are membrane proteins that do not specifically interact with nucleic acid, and would therefore not expected to coevolve with TEs. In a sample of representative species from 103 mammalian families, we observed a weak, non-significant correlation between the number of retroelement and olfactory receptor ORFs (Spearman's rho = 0.12). While not ruling out the possibility of a small effect of genomic TE content on gene duplication rate, this observation indicates that the effect is insufficient to explain the correlation observed between ZnFs and TEs.

ZnFs are predicted to bind TE sequences.

If ZnFs are directly coevolving with TEs, then it is most likely though sequence-specific recognition of DNA sequence (although several ZnFs are known to bind RNA). Since experimentally confirming the binding specificity of individual ZnF genes is challenging and labor-intensive, computational tools have been developed to predict ZnF binding motifs directly from protein sequence[21–24]. The most recent of these (Najafabadi et al.) predicts binding motifs for individual ZnF domains using a random forest classifier. Using this model, we predicted binding sites for all ZnF ORFs from seven representative metazoan species (Supp Data X).

First, we used previously published ChIP-exo data from human KRAB-ZnFs[6] to confirm that predicted binding sites are similar experimentally observed ones, as has been demonstrated by others working with mice[25] (Supp table X). As a second positive control, we compared the predicted binding motifs for all human and mouse ZnF ORFs to libraries of human- and mouse-specific TEs, and observed a significant, albeit weak, enrichment in both cases, as compared to shuffled motifs (Supp table X). This result is as expected given that KRAB-ZnFs are known to target TEs, and confirms that predicted binding sites recapitulate experimentally obtained motifs.

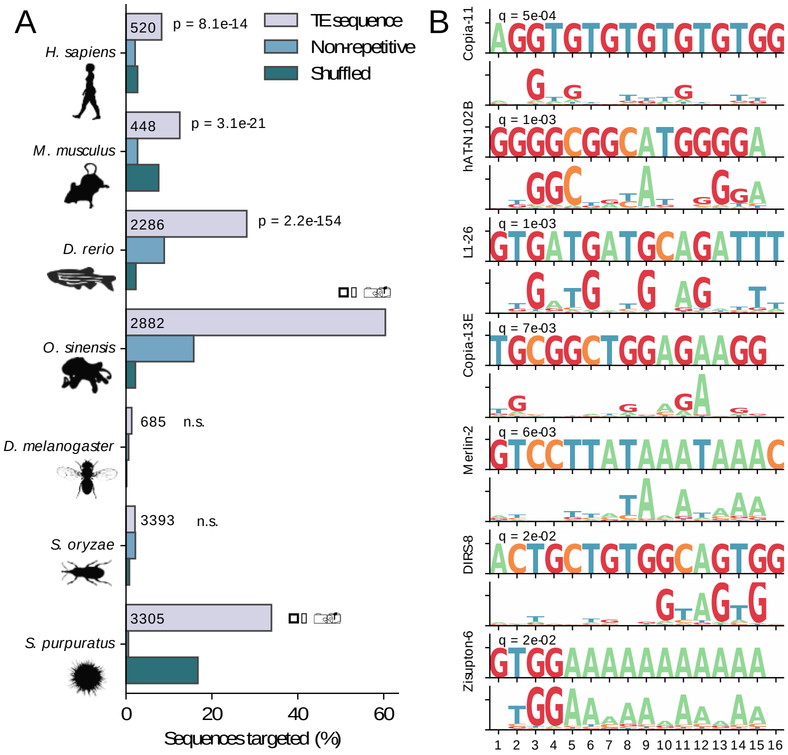


Figure 3. ZnFs are predicted to bind TE sequences.

We then turned our attention to species whose ZnFs are currently understudied, predicting binding motifs for all ZnF ORFs and comparing these to libraries of TE sequences and genomic repeats (Fig. 3B). In two of these, namely zebrafish and octopus, we observed a highly significant enrichment of predicted binding motifs in TE consensus sequence libraries, compared to scrambled motifs. Manual inspection of matches revealed that many were from motifs targeting low-complexity sequences. While this may reflect recognition of genomic satellite DNA by ZnFs, it could also be a result of overly simplistic motif predictions, as the random forest classifier does not consider epistatic interactions between neighboring ZnF domains, which are known to be important[26]. To ensure that our observations were not driven solely by these low-complexity motifs, we repeated the analysis without them, but observed only a minor decrease in enrichment (Supp. Fig. X).

TEs evolve rapidly and are under strong selection to evade repression by their hosts; many KRAB-ZnF proteins are in direct competition with TE families, leading to signatures of arms races between the two [5,27,28]. One such signature is positive selection acting on the DNA contacting residues of ZnFs, i.e. those involved in TE recognition[29]. We therefore tested for evidence of positive selection in *Danio rerio* FiNZ-ZnFs, comparing rates of synonymous and non-synonymous substitutions (dN/dS). Values of dN/dS > 1 suggest that natural selection accepts novel amino-acid changes at a rate faster than expected by chance, and vice versa.

In practice, it is challenging to confidently align ZnFs due to their repetitive nature and tendency to undergo internal rearrangements and gene conversion, and errors in this process can lead to false positive in assessment of dN/dS ratios. To account for this, we restricted our analyses to seven clades of recently duplicated in-paralogs unique to *Danio rerio* (see methods for details). Using PAML[30], we performed likelihood ratio tests to compare two models of evolution for each clade – one in which the genes were evolving under purifying selection, and one in which some sites were assumed to be under positive selection (M2 vs M2a). In four of these seven clades, the model featuring positive selection (M2a) was significantly favored (p-value < 0.01, likelihood ratio tests). Furthermore, in two clades we found that base-contacting residues of ZnF domains were significantly enriched for values of dN/dS > 1 (odds ratio = 2.45, p-value = 0.01; odds ratio = 2.84, p-value = 0.02; Fisher's exact test).

FiNZ-ZnF and retroelement expression coincide during zebrafish embryogenesis

Multiple studies have reported widespread ZnF expression during early zebrafish development. Using our de-novo FiNZ-ZnF annotations, we remapped previously published sequencing data from White et al. covering the early stages of embryogenesis[10]. Setting a lower limit of 0.5 transcripts per million (TPM) to call genes as expressed, we find that approximately half of all FiNZ-ZnFs (306 out of 684) are active during development. Recently published work on zygotic genome activation (ZGA) revealed two distinct waves of ZnF expression in zebrafish, named "sharp peak" and "broad peak"[31,32]. Sharp peak FiNZ-ZnFs share a distinct promoter architecture and are expressed in the early wave of ZGA, followed shortly after by the broad peak ZnFs in the major wave.

With our updated FiNZ-ZnF annotations, we recapitulated these findings, finding 80 sharp peak and 204 broad peak genes expressed during development (Fig. 4A). In addition, we found a small subset of approximately 22 maternally deposited genes. To our knowledge, these have not previously been described, likely due to their relative scarcity and incomplete annotation in Ensembl and RefSeq gene sets. Furthermore, maternally deposited mRNAs in zebrafish and other animals are initially not fully poly-adenylated[33–35], and are therefore difficult to detect when using polyA-selected RNA-seq platforms, as is the case for the White et al. data. To account for this, we remapped rRNA-depleted reads from a second dataset covering early embryogenesis[33], and found that a subset of FiNZ-ZnF mRNAs are indeed present in the egg and 1-cell zygote (Fig 4A, inset).

Embryogenesis represents a critical opportunity for TEs to be inherited by future host generations, as the germline cells are relatively unprotected in the earliest stages of development, compared to adult gonads. Consequently, many TE families are transpositionally active – retroelements particularly so (CITE) – during embryogenesis, presenting a selective pressure on the host to control their activity. In *Danio rerio*, the onset of TE expression broadly coincides with that of FiNZ-ZnFs (Fig 4A). Separating by TE class and FiNZ-ZnF type, we find that sharp and broad peak FiNZ-ZnFs tend to have strong positive correlations with LINEs and LTR retroelements, whereas DNA transposon expression is significantly less likely to correlate with FiNZ expression (Fig. 4B). In contrast to most FiNZ-ZnFs, the typical expression pattern of the 22 maternally deposited genes is anticorrelated with that of all TE classes.

In addition to the timing of expression, maternally deposited FiNZ-ZnFs differ from their zygotically expressed counterparts in other respects: while the majority of *Danio rerio* FiNZ-ZnFs are very young, particularly sharp-peak genes, those that are maternally deposited conserved across species and are significantly older (Fig. 4C). They are also physically co-localized at the telomeric end of Chr 4q, in a region that lacks the repeat density that characterizes much of Chr 4q (Fig. 4D, Supp. Fig. XA). This fact is surprising since Chr 4q in zebrafish is subject to rapid turnover and rearrangement, such that there is effectively no correlation between physical proximity and evolutionary relatedness (Supp. Fig. XB). This sharply contrasts with KRAB-ZnFs, which form characteristic clusters of tandem duplications, suggesting that selection may be acting to compartmentalize maternally deposited FiNZ-ZnFs.

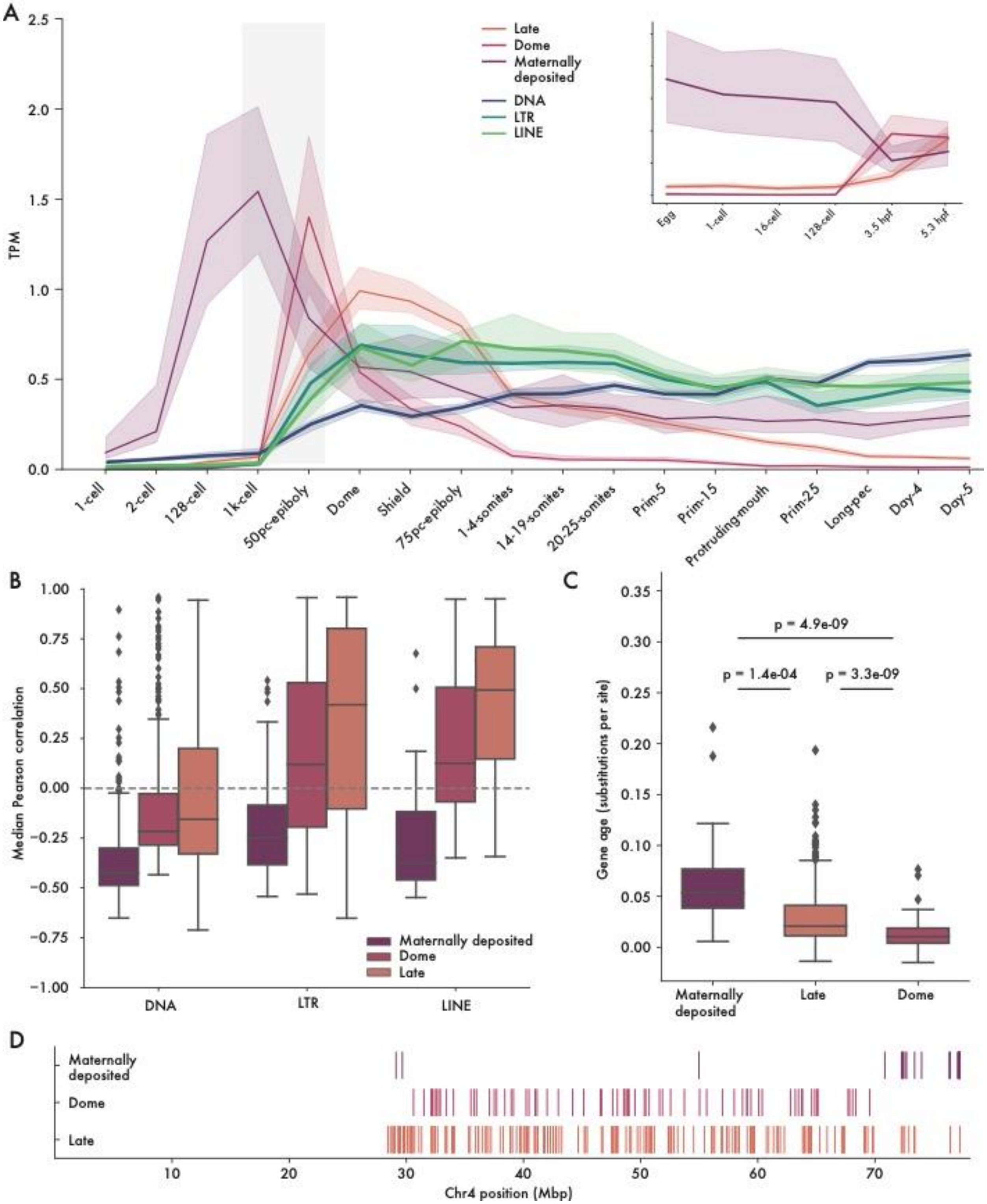


Figure 4. Embryonic expression of FiNZ-ZnF genes in zebrafish

Check/remove zga line. Fix stage order. Recolor. Add significance. Fix < 0 values for dome

A) Median expression trajectory of 306 genes with expression of at least 0.5 TPM in at least 1 stage, separated by cluster. Shaded borders represent 90% confidence intervals. Light grey bar shows *approximate timing of ZGA. In-set panel shows FiNZ-ZnF expression from rRNA-depleted reads. B) Maternally deposited genes are significantly older, whilst those expressed specifically at Dome stage are young, and generally unique to D. rerio. Ages were calculated C) The majority of maternally deposited FiNZ genes are located in a specific, sub-telomeric cluster. Each vertical line represents the midpoint of an expressed FiNZ-ZnF gene.*

FiNZ-ZnFs repress LTR retroelement expression during early development

Finally, we turned to the function of the FiNZ-ZnF genes. Based on their strong correlation with TE abundance, embryonic gene expression and predicted binding specificity, we predicted that they were likely to silence TE expression. Using a translation-blocking morpholino predicted to target approximately 400 FiNZ-ZnF genes, we compared gene expression between shield stage zebrafish embryos injected with this morpholino, to those injected with a scrambled control (Fig. 5A). Principal component analyses of batch-corrected samples revealed good separation between the FiNZ MO and control groups (Supp. Fig. XA).

Amongst all significant differentially expressed genes and TEs, TEs were enriched in the upregulated group (odds ratio = 5.94, p-value = 0.0049, Fisher's exact test) and LTR elements particularly so (odds ratio = 10.04, p-value = 0.0041) (Fig. 5A). This is consistent with previous observations showing an enrichment of repressive H3K9me3 histone marks over LTR elements following ZGA[36], and the fact that LTRs are transcriptionally active during zebrafish embryogenesis[37] (Fig. 4D). However, we observed mild developmental delays in the treatment group, and therefore reasoned that increases in LTR expression could be caused by differences in developmental stage, rather than by the FiNZ ZnF knock-down itself. To rule this out, we compared our samples with samples from White et al., covering shield stage, and 50% epiboly and 75% epiboly – the two stages preceding and following shield stage. Treating these samples as additional control groups, and after batch correction, we found that our samples clearly clustered with shield stage samples from White et al. (Supp. Fig. XB,C). Finally, we compared the magnitude of the changes in LTR expression between treatment and control with those between 50% and 75% epiboly and found that the effects of the FiNZ morpholino treatment were greater than those caused by stage differences (Supp. Fig. XD).

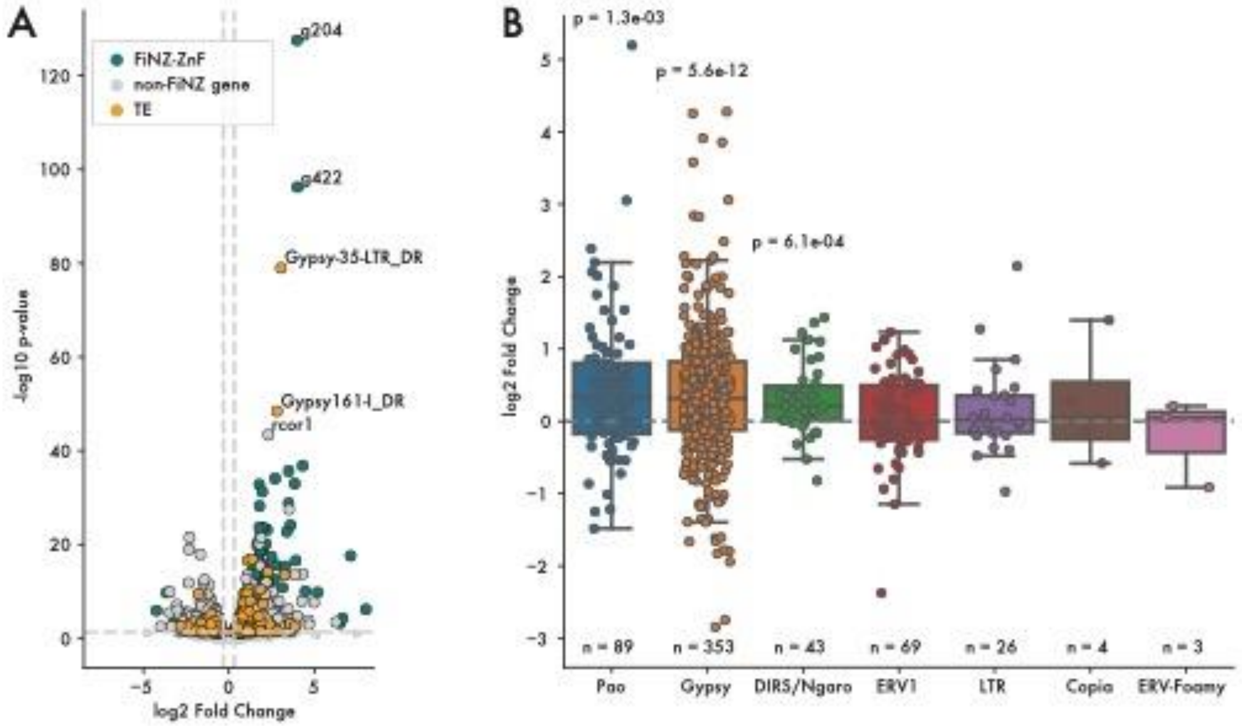


Figure 5. Up-regulation of LTR retroelements in response to FiNZ-ZnF knock down
A) Log2 fold change vs p. value, comparing FiNZ translation blocking morpholinos to scrambled morpholino control. B) Log2 Fold change of different LTR superfamilies.

An unexpected effect of blocking FiNZ-ZnF translation was that the most up-regulated group of genes were FiNZ-ZnFs themselves. While we cannot currently rule out the possibility that the morpholinos themselves inhibit the degradation of their target sequences, such effects have not previously been reported in the literature. We also note that levels of *tp53*, which is known to account for spurious morpholino phenotypes, is unchanged between treatment and control. Similarly, we observed no significant GO-term enrichment for genes involved in stress responses. A possible explanation for the up-regulation of FiNZ-ZnFs is that they target themselves as a form of negative feedback. In this way, their expression would be robustly switched off once sufficient FiNZ-ZnF protein had been generated, without the need for additional regulatory control.

We then returned to LTR retroelement expression: while the overall magnitude of upregulation is low, there is considerable variation between TE families (Fig. 5B), and we therefore sought to identify factors influencing the degree of differential expression. Overall, retroelement expression peaks at shield stage, but specific TE families have expression trajectories that peak earlier or later. We reasoned that those TEs whose expression peaks later in development ought to be less likely to be affected by FiNZ repression at shield stage, and we therefore separated all LTR families into two groups – those which are expressed prior to shield stage, and those expressed later. Comparing these two groups, we found early-expressed LTR families were significantly up-regulated relative to those with later expression profiles.

Discussion

In this work, we have demonstrated that there is a deep coevolutionary relationship between TEs and ZnFs that spans the breadth of metazoans. This is most clearly seen in the strong correlation between TE and ZnF copy number, a relationship that is independent of the accessory domains that define different ZnF subfamilies and persists across at least 750 million years of animal evolution. We also found that ZnFs from distantly related species are predicted to preferentially bind TE consensus sequences, and in the case of FiNZ-ZnFs in zebrafish, are under positive selection at these binding sites. Finally, through simultaneous knockdown of approximately 400 FiNZ-ZnFs during zebrafish embryogenesis, we demonstrated that FiNZ-ZnFs repress transcription of LTR retroelements. Previously, the only ZnF family with a known role in TE repression were the KRAB-ZnFs in tetrapods. Since the FiNZ domain is unique to cyprinid fish, which lack both KRAB and its cofactor KAP1 (a.k.a. TRIM28), this implies that FiNZ-ZnFs and KRAB-ZnFs have independently evolved to repress TEs, suggesting that TE repression is an intrinsic feature of ZnFs themselves, rather than their specific accessory domains.

An open question is whether repression of TE transcription is the primary function of metazoan ZnFs. While both KRAB- and FiNZ-ZnFs have now been shown to have a repressive effect on TE expression, there are many paradoxical observations that suggest repression itself may be secondary to other functions. Most obvious of these is the fact that gene duplication and diversification is very slow, relative to the speed at which novel TE families can become established and propagate within a species. For example, P-element has famously spread throughout all wild populations of *Drosophila melanogaster* in the space of approximately 30 years, and the retrovirus KoRV has become widespread in koalas through both exogenous and endogenous transmission in a similar time frame[38]; on the other hand, observations in primates imply that the emergence of novel KRAB-ZnF repressors takes place over millions of years[6,39]. In contrast, the piRNA system is able to adapt much faster, as TE insertions themselves form the raw materials for novel piRNAs – in the case of both KoRV and P-element, piRNAs have already emerged to mitigate their deleterious effects[40,41].

Furthermore, while the large majority of KRAB-ZnFs are broadly expressed during early embryogenesis and specifically target TEs, approximately a third target non-TE sequences, such as satellite repeats and gene promoters. There are also several examples in both mouse and human of conserved clusters of KRAB-ZnFs that target TEs younger than themselves, implying they must have had some function predating the emergence of their cognate TE families [6,25,42]. These observations have led others to propose that, rather than simply repressing TEs, KRAB-ZnFs play an important role in their domestication, incorporating them into existing gene regulatory networks[27,39,43]. In this way, KRAB-ZnFs act as tolerogenic agents which facilitate the evolution of species-specific gene regulatory networks. While we agree with this interpretation, it is important to clarify that this feature of KRAB-ZnFs cannot be the initial driver of their rapid evolution and turnover

An alternate, non-mutually exclusive hypothesis is that, rather than ZnFs being selected for their repressive effect on TE transcription, it is instead a matter of stabilizing repetitive DNA more generally. There is a substantial body of evidence that points to non-allelic homologous recombination being the major limiting factor for the accumulation of TEs in genomes, and heterochromatin has a strong repressive effect on recombination rates.

The exceptions: Drosophila/Anopheles ZnFs not apparently under selection [29]… In addition to lack of TE binding, lack of reported TE function, lack of correlation. Also important that piRNAs are known to be involved in establishing H3K9me3 during early embryogenesis in Dropsophila[44] and in yeast small RNAs do this[45,46].

Methods

Annotation of FiNZ-ZnF genes

We first identified genomic regions containing candidate FiNZ-ZnF genes using BLAST to search for matches with a consensus FiNZ domain sequence. This consensus sequence was generated from a set of *D. rerio* FiNZ-ZnF genes previously identified as being expressed during development[10]. Using the "PPX" module of Augustus (v3.3)[14,15], we carried out a first pass of the genomic regions identified in our set of cypriniform species, using a protein profile generated from the previously mentioned set of *D. rerio* annotations. To avoid biasing our search towards zebrafish, we used the results of this round of annotation to generate a new FiNZ-ZnF profile generated by sampling genes from all species. We then repeated the above procedure, and performed a final filtering step with HMMER[13] to identify annotated genes containing both the FiNZ domain and ZnF domain (PFAM: PF00096).

For *D. rerio* specifically, we produced a separate, high-quality set of gene predictions by retraining Augustus specifically for FiNZ-ZnFs using a manually curated set of Ensembl predictions. This allowed us to generate predictions for up- and downstream untranslated regions, for use in analyses of gene expression. We explored the effect of including transcripts generated by Trinity[47] as hints for Augustus, but found that these reduced the quality of resulting annotations, likely as a result of the difficulty of assembling accurate transcripts for such repetitive genes. Full parameter details, configuration files and training data are available in the GitHub repository.

Cypriniformes phylogeny

To generate a phylogenetic tree of the Cypriniformes order, we first used the Actinopterygii-specific BUSCO database to extract intact single-copy orthologues from our set of 28 species, including an outgroup, *Chanos chanos*. We then selected those protein sequences found in at least 10 out of 28 species, producing a final set of 3,581 proteins. These were aligned separately using mafft v7.475 with parameters –globalpair and –maxiterate 1000 (i.e. ginsi)[48] and trimmed to remove large insertions or poorly aligned regions using trimAl v1.4.rev15 with the –automated1 parameter set[49]. All resulting files were combined to produce a concatenated super-gene alignment, which was used to generate a time-calibrated phylogeny.

IQ-TREE v2.0.6 was used to generate a maximum likelihood tree from a partitioned analysis of the super-gene alignment[50], such that model selection and parameter estimation was performed independently for each gene. THIS NEEDS UPDATING TO INCLUDE ALL RELEVANT CITES. Ultrafast bootstraps and Shimodaira–Hasegawa approximate likelihood ratio tests were used to determine branch support values, with 5000 and 1000 replicates respectively. A time-calibrated tree was generated using the integrated LSD2 module[51], constraining the date of the split between Gonorynchiformes and

Cypriniformes (i.e. between *Chanos chanos* and all other species) to 162 Mya, based on a recent comprehensive phylogeny of teleost fish[52].

Testing for positive selection

Predicting genomic TE content

We estimated the TE-derived proportion of genomes using dnaPipeTE[53], a tool that offers considerable speed improvements over RepeatMasker/Modeller with comparable accuracy. This program requires short reads and genome size as input: for the former, we simulated reads using ART[54] and for the latter, we used genome assembly size as a proxy for true genome size.

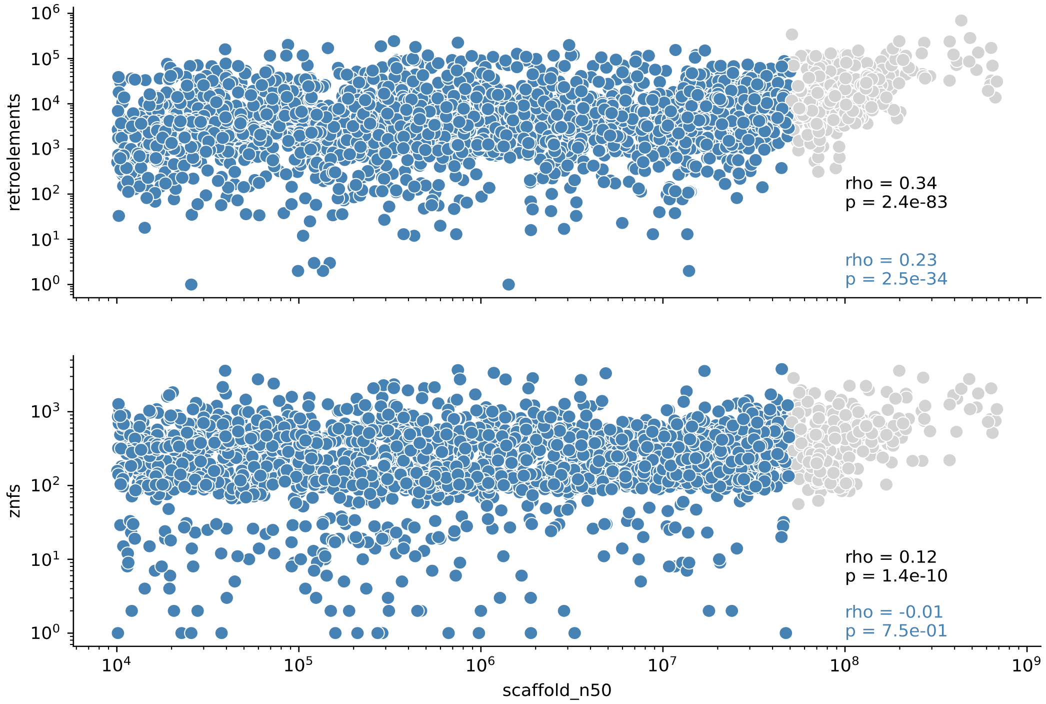
Predicting ZnF binding sites

Calculating expression

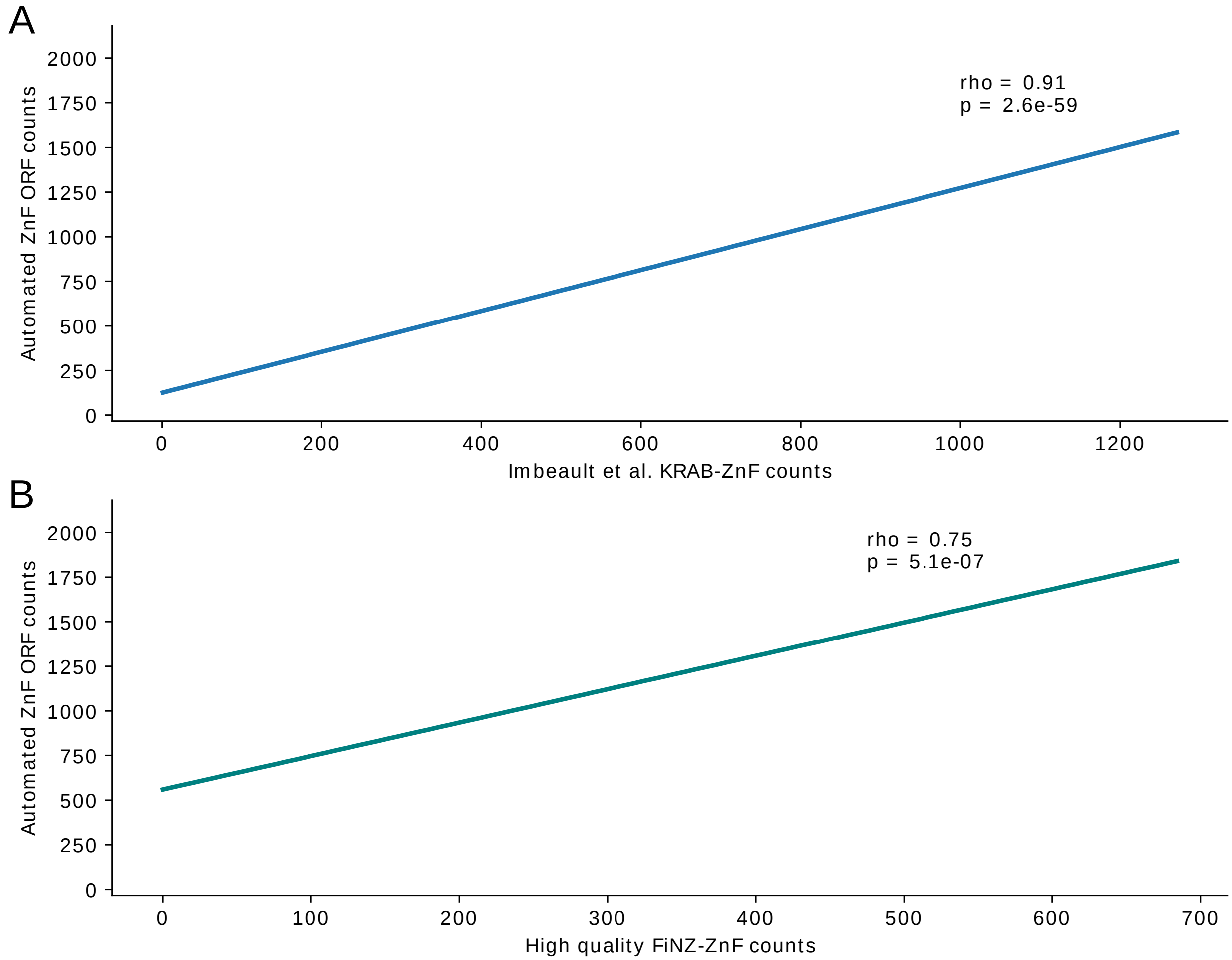
To calculate FiNZ-ZnF and TE expression, we used STAR in multi-mapping mode to align reads to the v11 reference genome. TEcounts to count reads. We used GTFtools (python package) to calculate the median length of each gene. We used these measurements to calculate TPM, allowing comparison of expression across stages. Specifically, for each transcript we calculate the sum of the union of exons. For genes, we then take the median length of transcripts associated with each gene, whereas for TEs we use the sum of all insertions in the genome, thus partially controlling for insertion copy number.

Morpholino knockdown of FiNZ-ZnF translation

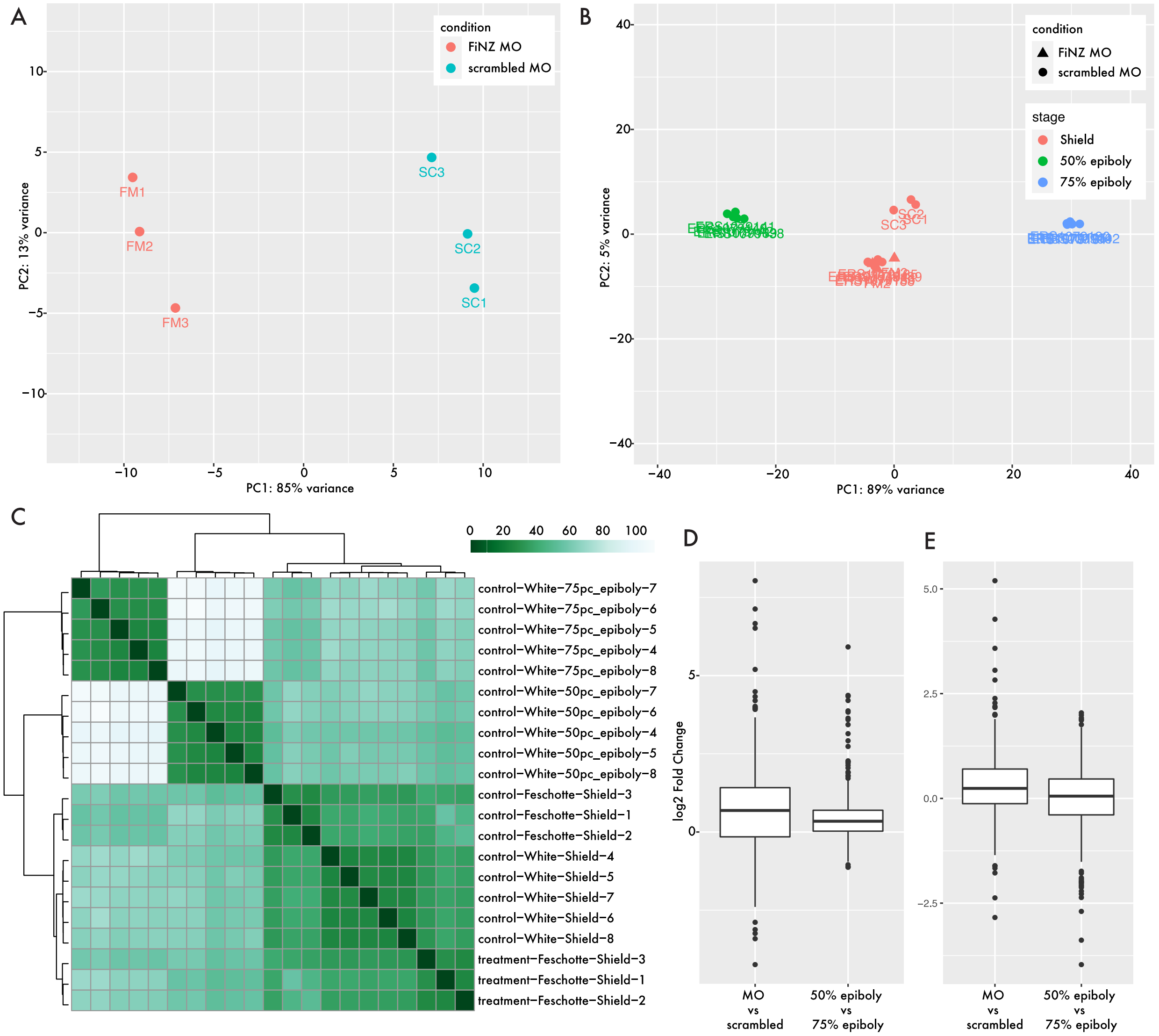
Supplementary Figures



Supplementary Figure 1. Correlation between scaffold N50 and retroelement and ZnF counts.
To assess the degree to which genome assembly quality biases our counts of ORFs, we compared scaffold N50 with A) retroelement and B) ZnF ORF copy number. This revealed low correlations, primarily driven by the fact that very large, high-quality genomes (and thus high scaffold N50 scores) almost always have high numbers of both retroelements and ZnFs.



Supplementary Figure 2. Correlation between automated ZnF ORF counts and annotated gene copy number.



Supplementary Figure 4. Comparison of treatment vs control and between-stage differential gene expression.

References

1. Heger, P., Zheng, W., Rottmann, A., Panfilio, K. A. & Wiehe, T. The genetic factors of bilaterian evolution. *Elife* **9**, e45530 (2020).
2. Najafabadi, H. S. et al. Non-base-contacting residues enable kaleidoscopic evolution of metazoan C2H2 zinc finger DNA binding. *Genome Biology* **18**, 1-15 (2017).
3. Bellefroid, E. J., Poncelet, D. A., Lecocq, P. J., Revelant, O. & Martial, J. A. The evolutionarily conserved Kruppel-associated box domain defines a subfamily of eukaryotic multifingered proteins. *Proceedings of the National Academy of Sciences* **88**, 3608-3612 (1991).
4. Thomas, J. H. & Schneider, S. Coevolution of retroelements and tandem zinc finger genes. *Genome Research* **21**, 1800-1812 (2011).

5. Jacobs, F. M. J. et al. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* **516**, 242-245 (2014).
6. Imbeault, M., Helleboid, P.-Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550-554 (2017).
7. Fedoroff, N. V. Transposable Elements, Epigenetics, and Genome Evolution. *Science* **338**, 758-767 (2012).
8. Madhani, H. D. The Frustrated Gene: Origins of Eukaryotic Gene Expression. *Cell* **155**, 744-749 (2013).
9. Huntley, S. et al. A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. *Genome Res* **16**, 669-677 (2006).
10. White, R. J. et al. A high-resolution mRNA expression time course of embryonic development in zebrafish. *eLife* **6**, (2017).
11. Howe, K. et al. Structure and evolutionary history of a large family of NLR proteins in the zebrafish. *Open Biol* **6**, 160009 (2016).
12. Howe, K. et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498-503 (2013).
13. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Computational Biology* **7**, e1002195 (2011).
14. Hoff, K. J. & Stanke, M. Predicting Genes in Single Genomes with AUGUSTUS. *Current Protocols in Bioinformatics* e57 (2018).
15. Keller, O., Kollmar, M., Stanke, M. & Waack, S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27**, 757-763 (2011).
16. Vacic, V., Uversky, V. N., Dunker, A. K. & Lonardi, S. Composition Profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics* **8**, 211 (2007).
17. Liu, H., Chang, L.-H., Sun, Y., Lu, X. & Stubbs, L. Deep Vertebrate Roots for Mammalian Zinc Finger Transcription Factor Subfamilies. *Genome Biology and Evolution* **6**, 510-525 (2014).
18. Kasinathan, B. et al. Innovation of heterochromatin functions drives rapid evolution of essential ZAD-ZNF genes in Drosophila. *eLife* **9**, (2020).
19. Chung, H. R., Löhr, U. & Jäckle, H. Lineage-specific expansion of the zinc finger associated domain ZAD. *Mol Biol Evol* **24**, 1934-1943 (2007).
20. Baumgartner, L., Handler, D., Platzer, S., Duchek, P. & Brennecke, J. The Drosophila ZAD zinc finger protein Kipferl guides Rhino to piRNA clusters. *bioRxiv* (2022).
21. Persikov, A. V. & Singh, M. De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. *Nucleic Acids Res* **42**, 97-108 (2014).
22. Najafabadi, H. S. et al. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nature Biotechnology* **33**, 555-562 (2015).
23. Molparia, B., Goyal, K., Sarkar, A., Kumar, S. & Sundar, D. ZiF-Predict: a web tool for predicting DNA-binding specificity in C2H2 zinc finger proteins. *Genomics Proteomics Bioinformatics* **8**, 122-126 (2010).
24. Kaplan, T., Friedman, N. & Margalit, H. Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput Biol* **1**, e1 (2005).
25. Wolf, G. et al. Non-essential function of KRAB zinc finger gene clusters in retrotransposon suppression. (2020).

26. Zuo, Z. et al. Why Do Long Zinc Finger Proteins have Short Motifs? *bioRxiv* (2019).
27. Bruno, M., Mahgoub, M. & Macfarlan, T. S. The Arms Race Between KRAB–Zinc Finger Proteins and Endogenous Retroelements and Its Impact on Mammals. *Annual Review of Genetics* **53**, annurev-genet (2019).
28. Fernandes, J. D. et al. KRAB Zinc Finger Proteins coordinate across evolutionary time scales to battle retroelements. (2018).
29. Emerson, R. O. & Thomas, J. H. Adaptive evolution in zinc finger transcription factors. *PLoS Genetics* **5**, (2009).
30. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24**, 1586-1591 (2007).
31. Hadzhiev, Y. et al. A cell cycle-coordinated Polymerase II transcription compartment encompasses gene expression before global genome activation. *Nat Commun* **10**, 691 (2019).
32. Hadzhiev, Y. et al. The miR-430 locus with extreme promoter density is a transcription body organizer, which facilitates long range regulation in zygotic genome activation. (2021).
33. Winata, C. L. et al. Cytoplasmic polyadenylation-mediated translational control of maternal mRNAs directs maternal-to-zygotic transition. *Development* **145**, (2018).
34. Cui, J., Sartain, C. V., Pleiss, J. A. & Wolfner, M. F. Cytoplasmic polyadenylation is a major mRNA regulator during oogenesis and egg activation in Drosophila. *Developmental Biology* **383**, 121-131 (2013).
35. Potireddy, S., Vassena, R., Patel, B. G. & Latham, K. E. Analysis of polysomal mRNA populations of mouse oocytes and zygotes: Dynamic changes in maternal mRNA utilization and function. *Developmental Biology* **298**, 155-166 (2006).
36. Laue, K., Rajshekar, S., Courtney, A. J., Lewis, Z. A. & Goll, M. G. The maternal to zygotic transition regulates genome-wide heterochromatin establishment in the zebrafish embryo. *Nature Communications* **10**, (2019).
37. Chang, N.-C., Rovira, Q., Wells, J. N., Feschotte, C. & Vaquerizas, J. M. Zebrafish transposable elements show extensive diversification in age, genomic distribution, and developmental expression. *Genome Research* gr-275655 (2022).
38. Tarlinton, R. E., Meers, J. & Young, P. R. Retroviral invasion of the koala genome. *Nature* **442**, 79-81 (2006).
39. Ecco, G., Imbeault, M. & Trono, D. KRAB zinc finger proteins. *Development* **144**, 2719-2729 (2017).
40. Brennecke, J. et al. An Epigenetic Role for Maternally Inherited piRNAs in Transposon Silencing. *Science* **322**, 1387-1392 (2008).
41. Yu, T. et al. The piRNA Response to Retroviral Invasion of the Koala Genome. *Cell* **179**, 632-643.e12 (2019).
42. Wolf, G. et al. The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses. *Genes &amp; Development* **29**, 538-554 (2015).
43. Pontis, J. et al. Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. *Cell Stem Cell* **24**, 724-735.e5 (2019).
44. Wei, K. H.-C., Chan, C. & Bachtrog, D. Establishment of H3K9me3-dependent heterochromatin during embryogenesis in Drosophila miranda. *eLife* **10**, (2021).

45. Hall, I. M. et al. Establishment and Maintenance of a Heterochromatin Domain. *Science* **297**, 2232-2237 (2002).
46. Volpe, T. A. et al. Regulation of Heterochromatic Silencing and Histone H3 Lysine-9 Methylation by RNAi. *Science* **297**, 1833-1837 (2002).
47. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652 (2011).
48. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772-780 (2013).
49. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).
50. Minh, B. Q. et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534 (2020).
51. To, T. H., Jung, M., Lycett, S. & Gascuel, O. Fast Dating Using Least-Squares Criteria and Algorithms. *Syst Biol* **65**, 82-97 (2016).
52. Hughes, L. C. et al. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *Proceedings of the National Academy of Sciences* **115**, 6249-6254 (2018).
53. Goubert, C. et al. De novo assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti). *Genome Biol Evol* **7**, 1192-1205 (2015).
54. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593-594 (2012).