Favorable genomic environments for cis-regulatory evolution: A novel theoretical framework.

Ignacio Maeso*

Corresponding author

nacho.maeso@gmail.com

Centro Andaluz de Biología del Desarrollo (CSIC/UPO/JA), Universidad Pablo de Olavide, 41013 Sevilla, Spain.

Juan J. Tena*

Corresponding author

nacho.maeso@gmail.com

jjtenagu@upo.es

Centro Andaluz de Biología del Desarrollo (CSIC/UPO/JA), Universidad Pablo de Olavide, 41013 Sevilla, Spain.

*These authors contributed equally to this work.

Correspondence should be addressed to nacho.maeso@gmail.com, jjtenagu@upo.es.

Abstract

Cis-regulatory changes are arguably the primary evolutionary source of animal morphological diversity. With the recent explosion of genome-wide comparisons of the cis-regulatory content in different animal species is now possible to infer general principles underlying enhancer evolution. However, these studies have also revealed numerous discrepancies and paradoxes, suggesting that the mechanistic causes and modes of cis-regulatory evolution are still not well understood and are probably much more complex than generally appreciated. Here, we argue that the mutational mechanisms and genomic regions generating new regulatory activities must comply with the constraints imposed by the molecular properties of cis-regulatory elements (CREs) and the organizational features of long-range chromatin interactions. Accordingly, we propose a new integrative evolutionary framework for cis-regulatory evolution based on two major premises for the origin of novel enhancer activity: i) an accessible chromatin environment and ii) compatibility with the 3D structure and

interactions of pre-existing CREs. Mechanisms and DNA sequences not fulfilling these premises, will be less likely to have a measurable impact on gene expression and as such, will have a minor contribution to the evolution of gene regulation. Finally, we discuss current comparative cis-regulatory data under the light of this new evolutionary model, and propose that the two most prominent mechanisms for the evolution of cis-regulatory changes are the overprinting of ancestral CREs and the exaptation of transposable elements.

Keywords: cis-regulation; transcription factor binding site (TFBS); enhancer evolution; chromatin environment; transposable element; enhancer overprinting.

Introduction

Animal form evolves largely by altering the expression of functionally conserved developmental genes and the vast genetic networks they control [1]. The transcriptional regulation of these conserved developmental genes critically depends on the inputs of complex sets of enhancer regions; cis-regulatory elements (CREs) that result from the binding of different transcription factors (TFs) to their specific DNA target sequences. How these mixed assemblages of proteins and DNA emerge as discrete regulatory islands from a vast ocean of chromatin is still not well understood. In contrast to protein coding regions, the biological processing of the genetic information contained in CREs does not rest in a fixed, universal and straightforward informational system such as the genetic code. Thus, defining molecular signatures and functional properties to annotate and characterize CREs genome-wide has been (and still is) a significantly greater challenge, and as such, these type of non-coding genomic elements have traditionally been much more elusive. This makes the investigation of how enhancers appear and are modified during evolution a particularly difficult endeavor. Fortunately, the advent of new experimental techniques and analysis tools, especially those based on Next Generation Sequencing (NGS), is completely changing this picture, and in recent years, numerous genome-wide studies in a variety of animal species have greatly contributed to our understanding of enhancers. These works have revealed two important aspects of regulatory elements that have not been fully integrated yet and are somehow, paradoxical. The first one relates to the molecular makeup of enhancers. Although far more numerous and widespread than protein coding regions, CREs are not just random pieces of chromatin containing sequences matching TF binding motifs. To exert their functions, enhancers need to be biochemically singled out from the surrounding chromatin,

configuring discrete molecular entities that, just from the perspective of their constituent TF binding sites (TFBSs), require the coordinated assembly of multiple factors (reviewed in [2, 3]). The second aspect is evolutionary conservation of CREs. The emerging picture from a wide range of comparative studies profiling several features associated to CRE function is that regulatory DNA is subjected to high rates of turnover. This pattern has been observed using different approaches, such as TF binding events identified by ChIP-seq (reviewed in [4]), chromatin accessibility using deoxyribonuclease I-hypersensitive sites (DHSs) [5, 6], and different histone modifications [7-10]. However, although in all cases evolutionary remodeling appears as a pervasive phenomenon, the rates of CRE conservation vary wildly across studies. Depending on the tissues/cell types, developmental stages, species or biochemical features and techniques used as a proxy to identify regulatory DNA, the proportion of evolutionarily conserved elements ranges from 2 to 75% in the case of mammals and 30 to 99% in *Drosophilid* flies (for a comparative see [4]). The reasons underlying this extreme heterogeneity have not been systematically addressed so far, but whatever the case, these discrepancies reveal that evolutionary processes associated to cis-regulatory changes are probably much more complicated than generally appreciated. As stated above, the structural configuration of enhancers and other CREs is complex enough to allow a precise control of their activity: in the intricate space of the nucleomes of living cells, regulatory DNA elements are exclusively accessed in the appropriate contexts, preventing spurious transcriptional regulation. Furthermore, the distribution of functional elements in the genome, and of CREs in particular, is far from being uniform. Highly distinctive features demarcate the genes and genomic regions around developmental genes, with much larger intergenic distances, longer and more numerous introns and higher evolutionary constraint, both in terms of sequence and synteny conservation [11-17].

Therefore, we argue that to make sense of the current wealth of genome-wide regulatory data, a new conceptual framework for cis-regulatory evolution in animals is urgently needed. The numerous discrepancies and high heterogeneity observed among recent comparative regulatory studies evidence that existing models are not fully capturing the biological complexity of CREs and genomic architecture. Here, to contribute to this debate, we propose a new integrative view. First, we review current knowledge on the molecular properties of CREs and of enhancers in particular, to emphasize how these features may define and constrain which genomic environments are more permissive or prone to enhancer evolution. Second, we discuss which possible evolutionary mechanisms

are more likely to contribute to regulatory evolution. Finally, we discuss these hypotheses in light of the particular signatures associated to the regulatory states, transcriptomic profiles and genomic organization of different tissues, developmental stages and animal species, to better reconcile current data on CRE evolution.

1. Biochemical properties, structure and signatures of cis-regulatory regions: are enhancers molecular ingredients available and abundant in the genome?

Computational simulations have predicted that in a neutral evolutionary scenario, new TFBSs appear and become fixed at very short timescales, with an estimate of 24 years time for a 6-bp motif to evolve in a 2000 bp region in *Drosophila* [18]. But, is that all that it takes to generate a CRE that actually affects gene expression? For any given TF, the number of occurrences of its specific DNA recognition site in the genome greatly outnumbers the amount of active enhancers that depend on this TF: a 6-mer binding site would be represented over 700.000 times in a human-size genome [19] and given that most TFs recognize degenerate DNA sequences, this figure can be even higher. In contrast, TF binding sites (TFBSs) biochemically identified by ChIP-based techniques typically fall by one or two orders of magnitude [20, 21]. Furthermore, it is still unclear how many of these ChIP-seq binding events are actually part of biologically functional enhancers [22, 23]. Therefore, the amount of TFBSs that have a measurable impact on target gene expression is probably even lower. The presence of a single TF binding event is generally not enough to actively regulate the transcription of its surrounding genes; TFBSs have to be integrated into complex CREs containing additional TF motifs. In fact, the extent to which a TFBS contributes to gene expression critically depends on its spatial association to additional TFBSs, with singleton binding events being less likely to have a significant impact on transcription [20, 24]. According to this, more sophisticated and realistic modeling approaches have been developed for the evolution of more complex CREs containing multiple binding sites for one or more TFs, as typically found in functionally characterized enhancers [25, 26]. In comparison to single TFBSs, this raises time estimates by several orders of magnitude: a CRE of considerable complexity, such as those required to drive A/P patterns in *Drosophila* embryogenesis, may evolve in approximately 0.5 my. Still, this estimate is well in the range of divergence times of very closely related species and microevolutionary timescales, but depending on the complexity of the combinatorial logic encoded by a CRE, this time-to-evolve can vary extensively, with some expression patterns requiring up to 30 times longer [26].

Although these improved simulations could explain the evolution of the complex sequence composition of some enhancers, if TFs are functionally integrated at very specific sites only, what makes an enhancer region special? Leaving aside the combinatorial nature of enhancers and considering just TF recognition and binding preferences, the question is already complex, spanning multiple biological levels (reviewed in [2, 3]). These range from the most basic biophysical and biochemical properties, such as the 3D shape of the DNA molecule [27, 28], to factors affecting nucleosome positioning and chromatin state and accessibility (such as chromatin remodelers/modifiers, [29, 30]) and to complex interactions and cooperativity with other molecules, including additional TFs and their cofactors [31, 32], and noncoding RNAs (reviewed in [33]). The combinations of these factors create appropriate microenvironments, allowing TFs to recognize and stably bind their target sites within a complex nucleoplasm that is crowded with macromolecules (at a really high concentration of 10-40% [34, 35]). This does not necessarily imply that TFs never bind other potential target motifs located in disfavored environments, but those bindings would probably be less stable or efficient. This means that TFBSs contained within enhancer regions will be associated to a precise chromatin environment and certain molecular signatures, features that have been widely exploited by NGS methods for genome-wide enhancer identification. Enhancers are associated with enrichment of lysine 4 monomethylation of histone H3 (H3K4me1) and lysine 27 acetylation of histone H3 (H3K27ac) modifications, are located within accessible chromatin (i.e. DHSs and ATAC-seq peaks [36, 37]) and are able to recruit transcriptional activators, such as p300 and CBP [38-46]. Moreover, enhancers can be subdivided into distinct functional classes based on their epigenetic makeup: whereas H3K4me1 alone marks poised enhancers, active ones are enriched in both H3K4me1 and H3K27ac modifications [38-42]. Importantly, since these combinations of histone marks have been repeatedly found in different animals, ranging from flies to mammals, and more recently even in non bilaterian species such as sea anemones [47] they probably represent evolutionarily conserved signatures of enhancers in animals. However, despite these seemingly distinctive features, a growing number of studies suggest that there appears to be very little difference between promoters and enhancers, challenging the established view separating them as distinct entities (reviewed in [48, 49]). In fact, it now seems clear that functionally active enhancers have the ability to bind Pol II and initiate transcription, with several examples of intragenic enhancers acting as alternative promoters (reviewed in [49]). Collectively, this argues against a strict functional categorization of CREs, leading to the proposition of a unifying model of

transcriptional regulatory elements, in which activity, transcriptional output and regulatory function is context specific [49]. It is thus conceivable that enhancers and promoters can switch functions in evolution, and accordingly, we will use the terms enhancer and CRE in a relaxed and rather interchangeable sense.

In conclusion, it seems very unlikely that all the multiple factors and molecular features associated to enhancer function can evolve de novo at any random genomic location. Even the appearance of an isolated TF binding event would require the mutation(s) generating it to happen in a chromatin region that is accessible to the corresponding TF proteins (Fig. 1). Therefore, at the very least, a pre-existing favorable chromatin environment must precede the appearance of novel regulatory activity.

Figure 1. Generation of new TFBSs due to random mutations in the DNA. Single base random mutations (red asterisks) may generate new TFBS (dotted red boxes) or modify previously active motifs (solid colored boxes). Nevertheless, only TF target motifs that appear in open chromatin regions could be established as new active TFBSs (solid red box).

2. Defining favorable genomic “niches” for cis-regulatory evolution

2.1. New enhancers preferentially arise by modification of pre-existing CREs

Different studies have successfully reconstructed the evolutionary history of several recently evolved enhancers. Setting aside the numerous cases derived from endogenous retrovirus

and other transposable elements (TEs, see the corresponding section below), most other examples of gained enhancer activity actually involved the co-option or modification of DNA regions with pre-existing regulatory functions [50-55]. And this phenomenon is probably widespread. Among tissue-specific DHSs shared between human and mouse, a very substantial proportion has been repurposed or co-opted to a different tissue in one of the two species (from 23 to 69 %, depending on the tissues involved) [6]. Similarly, about 80% of the regions that have gained enhancer activity in human limb bud development after the divergence from macaques have originated from sequences with ancestral regulatory potential in the same or different tissues, and equivalent results have been obtained with elements that have increased their enhancer activity in human developing cortex [7, 8].

In contrast, to our knowledge, functionally validated examples of enhancers originating autonomously from “inert” DNA have not been reported so far. However, a recent study has characterized in detail four cases in which enhancer activity has evolved from pseudogenized coding exons with no previous regulatory activity [56], consistent with an earlier report identifying 38 cis-regulatory coding exons using sequence conservation [57]. In fact, a subsequent work documented two further examples of coding exons with enhancer activity (in these cases, functioning dually as CREs and as coding exons), and inferred many other putative cases of coding CREs regions based on the presence of typical enhancer signatures (deposition of H3K4me1, H3K27ac and p300), suggesting that this phenomenon may be relatively common [57, 58]. These results indicate that new TFBSs can appear outside the context of preexisting CREs. But it is noteworthy that before gaining cis-regulatory activity, these co-opted DNA regions were functional and expressed as mRNAs. In this regard, it is tempting to speculate that the chromatin environment associated to transcriptionally active coding exons could have facilitated the access of TFs to their newly evolved target motifs. Nevertheless, this hypothesis should be treated with caution. First, because in at least three of these coding exon-derived CREs, the immediately flanking introns have conserved enhancer functions, suggesting that the exonic regulatory activity could have arisen by simple extension of a previously existing intronic enhancer [56, 58]. Second, because of technical and experimental design limitations. To precisely document the evolutionary events leading to the appearance of a novel enhancer, it is essential to be able to identify its source DNA sequence in an outgroup species. Therefore, studied examples would be biased towards sequences under purifying selection rather than neutrally evolving DNA regions, thus enriching for conserved enhancers and exons (but see below the section on TEs). Even in

those examples where the outgroup species is close enough to identify the orthologous DNA region in the absence of sequence constraint, formally, it would be virtually impossible to demonstrate that in this outgroup species, this DNA region does not have any kind of regulatory activity in any possible tissue or developmental stage. In stark contrast with this view, a broad comparative ChIP-seq study of H3K4me3 and H3K27ac distribution in liver cells of 20 mammalian species has reported widespread occurrence of recently evolved enhancers by exaptation of ancestral DNA sequences. The authors identified about 10000 putative enhancers in each species showing H3K27ac enrichment in the liver of that species but not in the rest of studied lineages, a figure 10-20 times higher than the enhancers with conserved liver activity across placental mammals [10]. Interestingly, in up to 77% of the recently gained liver enhancers, the underlying DNA sequence is over 100 my old, predating the radiation of placental mammals (and most of the rest are enriched in TEs and repetitive elements). The conservation of these sequences for such a long period of time is highly suggestive of evolutionary constraint, arguing against the exaptation scenario. Thus, as in the aforementioned examples in human limb buds and cortex [7, 8], these recently gained enhancers in liver (a tissue particularly prone to DHSs repurposing [6]) have most likely evolved from ancestral regulatory sequences rather than from exaptation of inert DNA.

In summary, although there may be important exceptions, the bulk of existing evidence on enhancer evolution indicates that new regulatory activity would preferentially arise within chromatin regions with ancestral regulatory potential.

2.2. Permissive genomic environments for enhancer evolution at the micro-scale level: open chromatin regions.

The cases discussed above exemplified how frequently ancestral enhancers are modified during evolution, suggesting that they constitute particularly favorable environments for regulatory innovations. As discussed before, the complexity of the combinatorial logic encoded by some enhancers could render their de novo appearance particularly unlikely, providing a possible cause for the multiple instances of evolutionary modification of CRE functions. Thus, novel enhancer activities could evolve more easily by reusing part of the regulatory information contained within previously existing CREs, either by mutations in their constituent TFBSs or by extending the size of CREs incorporating new TF motifs at their flanks. However, we argue that, in addition to this, there is another important biological reason that could better explain the rampant reuse and repurposing of ancestral cis-regulatory DNA.

From all possible mutations resulting in the generation of a new TF target motif, only those happening within open chromatin regions could be readily accessible by the corresponding TF protein (Fig. 1). And as previously stated, CREs are the most prominent accessible regions in the genome. In addition to providing a simple mechanistic explanation for the widespread functional changes experienced by CREs [6], this scenario has other implications. First, it could explain the nested structure of many developmental enhancers that are configured as conserved non-coding regions (CNRs): numerous CNR examples exhibit an anciently and deeply conserved core sequence flanked by more divergent regions, conserved only within specific lineages, either because they evolved more recently or because they are less constrained [59, 60]. TFBSs appearing at the more flexible periphery of the enhancer-CNR would be less likely to be deleterious or to disrupt the intricate and conserved structure of the core region, while still benefitting from a more accessible chromatin environment. Second, it would be consistent with the high prevalence of enhancers constituted by homotypic clustering of TFBSs [61, 62]. Since the chromatin of most enhancers is only open at certain regulatory states (i.e. those expressing the TFs recognizing the TFBSs contained in the enhancer), any newly generated target site that is homotypic with those already contained in the enhancer could be immediately recognized and accessed by its cognate TF, thus increasing the chances of being incorporated to the ancestral enhancer sequence. Thus, except in those cases in which this process could have deleterious consequences, homotypic TFBS clusters would tend to appear and accumulate with evolutionary time. This is consistent with a previous hypothesis based on alternative causes arguing that to explain the high genomic frequency of homotypic clusters, there is no need to invoke adaptationist evolutionary explanations [63]. Importantly, depending on the complexity of the regulatory state initially associated to a CRE, this phenomenon could allow the evolutionary switching between homotypic and heterotypic clustering. Third, as commented before in regard of reported cases of exonic enhancers, chromatin states other than those associated to enhancers and promoters could have different propensities for the evolution of new CREs. This is an interesting possibility for future investigation, since it could provide an evolutionary explanation for the frequent genomic intermingling of enhancers and other transcriptional active elements such as lncRNAs and bystander genes or for the recently reported crosstalk between intragenic enhancers, alternative splicing and Ago1 binding events [64, 65]. Finally, under this open chromatin environment scenario, the effective genome size with actual evolutionary potential to generate novel regulatory activities would be drastically reduced.

Therefore, future theoretical models and computational simulations of enhancer evolution could probably be further refined by taking this into account. Importantly, this 'effective genome size' would represent a much higher fraction of the total genome size in animal species with highly compacted genomes versus those with highly expanded genome sizes. As it has been previously proposed, this could potentially have a significant impact on the overall evolutionary trends for regulatory elements present in animal lineages with different degrees of genome compaction [4, 66].

2.3. Permissive genomic environments for enhancer evolution at the macro-scale level: regulatory landscapes and topological associated domains (TADs)

To exert their functions and regulate the transcription of their target genes, individual enhancers do not act in isolation, but in concert with other CREs. The conception of how this coordinated action actually takes place in the nucleus has changed completely over the last years. With the use of different Chromosome Conformation Capture techniques [67-69], the classic linear view of animal genomes, with genes being regulated by neighboring CREs following a sequential arrangement, has been replaced by a three-dimensionally compartmentalized chromatin organized in TADs, genomic regions within which sequences preferentially contact each other. This precise 3D organization of the chromatin enables long-range contacts between sets of CREs and their target genes, ensuring a proper control of transcription. Thanks to the formation of chromatin loops, mediated by the binding of the protein CTCF [70], chromatin elements located at very long linear distances are brought together in the nuclear space; by contrast, the interactions of regions that would otherwise be very close to each other, are separated [71-73]. With this in mind, the evolution of enhancers would not only be constrained by their immediately adjacent chromatin environments, but by the entire 3D topologies and interactions of large genomic regions [74, 75]. At the most simple level, to be able to interact with a given target gene, new enhancers would have to be located within the same 3D chromatin domain. But perhaps more importantly, the enhancer target gene would have to be compatible. This of course includes the promoter selectivity of the gene [76], but also the rest of the gene cis-regulatory apparatus. The genomic region encompassing the full set of CREs controlling the activity of a gene is usually termed the regulatory landscape of that gene [77], or in other contexts, a genomic regulatory block (GRB) [11, 13, 14, 65, 78]. Thus, the regulatory information encoded by a new enhancer, with all the different TFs, cofactors and other molecules recruited by it, would need to productively

interact with the macromolecular complexes assembled by the rest of CREs located within a regulatory landscape; adopting an appropriate 3D topology (Fig. 2). For instance, an enhancer responding to a regulatory state to which any of the other CREs in the landscape is responsive would be less likely to interact and have any transcriptional impact. Conversely, novel CREs whose regulatory activity is at least partially redundant with that of pre-existing elements would have more chances to become functional. In a similar way to the previous considerations on the evolution of homotypic TFBS clusters, this could provide an explanation for the prevalence of seemingly redundant enhancers [59, 79, 80]. Following this line, if multiple enhancers within a genomic region are recognized by a common set of TFs, they may create a 3D chromatin environment that favor the formation of TF-specific transcription factories (areas of the nucleus where a set of transcription factors and their TFBS concentrate [81, 82]).

Figure 2. Compatibility of new enhancers with the current gene transcription machinery. To settle in a gene regulatory landscape, new enhancers that appear by to random mutations or genomic rearrangements need to be compatible with the promoter complex and other regulatory regions that activate this gene. Blue rounded form represents the promoter complex (all the proteins binding to this region necessary for transcription activation) together with other protein complexes that mediate its interactions with distal enhancer regions. Green forms represent previously active enhancer complexes (both DNA sequence and factors bound to it), whereas orange and purple forms represent new enhancers and their cognate factors generated in the landscape.

3. Main evolutionary mechanisms for the origin and modification of enhancers

Cumulative evidence suggests that certain mutational mechanisms (CRE overprinting and insertion of TEs, see below) have had a particularly prominent role in reshaping the cis-regulatory content of animal genomes. The evolutionary impact of some additional mechanisms (such as reorganization of regulatory landscapes) is still unknown, but they pose interesting biological questions that could lead to new research avenues. Importantly, most of them are not mutually exclusive, so they could potentially act together in the same enhancer regions.

3.1. Overprinting of pre-existing CREs

As we have already mentioned in 2.1 and 2.2, numerous works have demonstrated that modification of ancestral enhancers is a rampant phenomenon, a prevalence that probably stems from the higher accessibility of these genomic regions. To what extent mutations rewriting the sequence of a given CRE also result in changes in CRE activity probably depends on many factors, such as the structure of the CRE, the selective pressures of the target genes and other causes. For instance, Vierstra *et al.* have shown that 41% of DHSs shared between human and mouse lack any conserved TF target site between the two species. Accordingly, the majority of these completely overwritten DHSs were active in different tissues in mouse and humans. On the other hand, although another 19% of shared DHSs have also undergone overwriting of their constituent binding sites, they are operationally conserved (i.e. they contain the same collection of TF recognition motifs, although located at different positions within the DHS) and in many cases, they also maintain the same tissue specific activities [6]. Similar to this latter case, two studies comparing orthologous regulatory landscapes in different insect species have shown that functionally equivalent enhancers are maintained at equivalent genomic positions, even though their underlying sequences have diverged beyond recognition [83, 84], as it has been discussed elsewhere [15, 85].

3.2. De novo autonomous generation by cumulative mutations in 'inert' DNA

In principle, most TFs need a favorable chromatin environment to be able to access the DNA. However, 'pioneer TFs' capable of binding to DNA target sites even in inaccessible regions, subsequently promoting DNA accessibility (and thus creating the appropriate environment for the advent of other types of TFs) [86]. Therefore, new enhancers could appear at random

genomic locations by a stepwise mutational process starting with the creation of a target motif for a pioneer TF. The prevalence of such a scenario has not yet been assessed (for instance it could be checked if TFBSs of pioneer TFs have higher rates of turnover than those of other TFs), but under the aforementioned theoretical simulations [26], this process could lead to the appearance of *de novo* enhancers in a relatively short period of time.

3.3. Enhancer duplication

Tandem and genome duplications have a major impact on the evolution of animal genomes. As any other functional duplicated region, in the case of CREs, after duplication, the resulting copies could undergo non-functionalization (i.e. loss), sub-functionalization or neo-functionalization processes. Interestingly, in comparison with protein coding genes, cases of duplicated enhancers (identified as CNRs) are comparatively rare. Almost half a billion years after the two rounds of whole genome duplications (WGDs) that happened before the diversification of jawed vertebrates, the great majority of developmental genes have retained at least two of their four original WGD copies [87]. In contrast, only 124 of their associated CNRs are duplicated, with only 12 of them present in more than 2 copies [88]. This suggests that either vertebrate genomes were structured very differently before the occurrence of WGDs than in extant species, containing only a very small number of CREs structured as CNRs, or that most of the paralogous CNRs copies have been lost or have diverged beyond recognition. The paucity of tandemly duplicated CNRs adds support to the latter scenario: only 2 CNR paralogous families, one containing *lrx* ultraconserved enhancers and the other one associated to *Sall* genes, have duplicated copies arrayed in tandem [88], in contrast to the high frequency of tandem gene duplication. Consistently, examples of duplicated CNRs in other animal lineages in which the regulatory landscapes of developmental genes have duplicated more recently [16, 89, 90] show numerous duplicated non-coding sequences. Thus, this suggests that in duplicated enhancers, ancestral sequence is rapidly erased or lost. Given the prevalence of the phenomenon and the relaxed selective pressure of a redundant context, we believe that CRE overprinting (see 3.1) is particularly pervasive among duplicated enhancers. This suggests that both genome and local duplications contribute very significantly to regulatory innovation by, for instance, increasing the 'effective genome size' for enhancer evolution.

Figure 3. Gaining of regulatory information by insertion of TEs. TEs carry their own cis-regulatory information within their sequence, including a promoter and diverse TFBSs (colored boxes). In the upper panel, gene A is not being transcribed and its surrounding genomic region is condensed, impeding the TE insertion. However, in the lower panel, gene B shares some regulatory information with the transposable element (represented by the red box), so both will be transcribed in the same cells under their shared regulatory state. Thus, the insertion of the TE in gene B genomic region will be much more probable. If this situation happens, gene B could gain extra regulatory activity by recruiting some of the TE TFBSs and/or the promoter (dashed arrow and shaded boxes).

3.4. Insertion and exaptation of TEs

An increasingly growing number of genome-wide studies is showing the paramount importance of TEs as global drivers of cis-regulatory evolution, with widespread occurrence of promoters and enhancers derived from exapted TEs [6, 91-98]. For instance, 25% of mouse lineage specific OCT4 binding events are derived from a particular LTR/ERV family [97]. Similar trends have been observed in lineage specific expansion of CTCF binding sites by independent exaptation of SINEB elements in different mammalian species [94]. And the same pattern is also observed in more global approaches not focused on particular TFs: up to 40% of mouse lineage-specific DHSs (i.e. with no sequence counterpart in humans) overlap

mobile elements, and between 23 to 48% of H3K27ac regions in liver located within species-specific sequences were enriched for mobile elements [6, 10]. These frequent exaptation phenomena are most probably due to the intrinsic regulatory potential contained within the DNA sequences of many TE families. Consistent with their behavior as autonomous molecular entities, several types of TEs contain an endogenous promoter element and other regulatory motifs at their UTRs that are able to recruit different TF proteins and either the Pol II or the Pol III complexes [99]. Thus, in contrast to other mutational mechanisms, TEs would carry with them their own regulatory activities and in this regard, to give rise to new CREs, TEs would be less influenced by the pre-existing chromatin environment of the insertion place. However, the TE insertion process itself could be highly dependent on the chromatin accessibility of the potential target regions (Fig. 3), as indicated by the high affinity for open chromatin regions exhibited by certain transposases (a phenomenon that has been recently exploited by the ATAC-seq method to efficiently and systematically profile CRE-associated open chromatin regions [36]). The preferential insertion of TEs within accessible chromatin regions could have important evolutionary consequences. To prevent an excess of deleterious insertions, the transcription of most retroelements (to be mobilized, this type of TEs first need to be expressed and retrotranscribed) is either always repressed or active at very specific regulatory states only. Thus, retroelement insertions would preferentially happen in chromatin regions that are accessible at the same regulatory state recognized by the TE itself to activate its transcription. Therefore, similar to the scenario proposed in 2.2 for the preferential TF binding to novel target motifs that are recognized by the same regulatory state that a nearby open chromatin region (see 2.2 on the evolution of homotypic clusters), newly inserted TEs and their accessible target locations would share at least part of their encoded CRE activities (Fig. 3). The implications of this are manifold. First, since most of the genes undergoing regulatory rewiring by the insertions of a particular class of retroelements would be transcriptionally active under a regulatory state shared by all of them (and therefore corresponding to the same tissues and stages), their biological functions would probably be already connected and in many cases, they would even be part of the same genetic network. This means that potentially, entire gene networks (or a large fraction of their components) could be rewired concertedly; therefore, expression changes associated to the TE insertion would impact the network in a coherent manner, being less likely to cause dosage compensation problems. Second, apart from those TFBSs associated to the regulatory state activating the TE, the inserted mobile elements can harbor target motifs for additional TFs

that, acting alone or in combination with those present in the target sites, could result in co-option of rewired networks, modification of network architectures or even the creation of new gene networks by establishing new regulatory links. Thus, by recruiting hundreds of genes at the same time, TE rewiring processes could be a major mechanism for the evolution of complex traits. Attesting to the great potential of TE rewiring, a recent study demonstrated that the gene regulatory network of endometrial stromal cells originated by rewiring in placental mammals during the evolution of pregnancy, and that this network reorganization and rewiring was partly mediated by the transposable element MER20 [100]. Similarly, another study recently reported that multiple MER130 elements were recruited as neocortical enhancers in mammals [101]. Third, to cause heritable insertions, regulatory states driving both the TE activation and the accessibility of the target sites should be active in the germline and/or early totipotent stages. This is in fact consistent with the bulk of reported cases of widespread exaptation of TEs as CREs, since many of them correspond to regulatory elements and gene networks operating in pluripotent cells and early developmental stages [93, 95, 96]. Importantly, this does not necessarily mean that TEs would not have the potential to create enhancers and reorganize gene networks in somatic contexts, since as mentioned above, TEs contain additional regulatory information to that directly involved in TE mobilization and insertion processes. Consistent with this, multiple examples of TE-derived enhancers active in different cellular contexts have been identified, and the number of reported cases is still growing [100, 102-106].

3.5. Changes in the architecture of regulatory landscapes

The alteration of the 3D topologies responsible for the compartmentalization of long-range chromatin interactions is likely to promote changes in CREs and gene regulation [74]. A possible mutational mechanism that could render this kind of changes would be the gain or loss of CTCF binding sites (as in the case of those exapted from retroelements [94]). However, it has been shown that recently evolved CTCF binding sites do not have a major impact in TAD architecture and that main changes in interaction topologies are instead driven by chromosome rearrangements (i.e. insertions, deletions and translocations). Furthermore, in those cases, the observed rearrangement occurs at positions that do not disrupt the organization of individual TADs and their borders, at least at the timescale of the radiation of placental mammals [107]. In fact, another study has shown that some TADs are strikingly

stable in evolutionary time, having maintained their architecture for 600 my since the origin of deuterostomes during the Cambrian explosion [72]. This would be consistent with a recent evolutionary model for the evolution of genome architecture that predicts that genomic rearrangements would preferentially occur if they do not disrupt the organization of TADs (Fig. 4A) [108]. However, a recent study characterizing the 3D architecture of the mouse chromosomal region containing *Wnt6*, *Ihh*, *EphA4* and *Pax3*, attests for the evolutionary potential of mutations altering the compartmentalization of long-range chromatin interactions. Inversions and deletions affecting CTCF-associated TAD boundaries within this region result in spurious regulatory contacts between the *Ihh* promoter and an *EphA4* limb bud enhancer. This causes *Ihh* to be ectopically expressed in the limb bud, leading to limb developmental and morphological aberrations such as polydactyly. Interestingly, it has been suggested that another paralog of the hedgehog family, *Shh*, could have gained its limb bud activity by a similar genomic rearrangement that would have made possible the interaction between the *Shh* promoter and its long-range limb bud enhancer [109, 110]. Although examples like this are very suggestive, unfortunately, similar cases demonstrating a causal link between changes in the 3D architecture of regulatory landscapes and morphological and cis-regulatory evolution, have not been identified yet. Nevertheless, it seems reasonable to hypothesize that this type of rearrangements may have played an important role in the evolution of gene expression (Fig. 4B), and that future systematic approaches could elucidate their potential contribution to animal evolution.

Conclusions

Until recently, existing hypotheses on cis-regulatory evolution in animals could only be sustained on empirical results obtained from isolated examples and on purely theoretical models. The unprecedented wealth of genome-wide CRE data from a continuously growing number of animal species has led us to a privileged scenario to identify global trends on enhancer evolution and to a much better understanding of the molecular makeup of CREs. Taking advantage of this, we have proposed a new evolutionary framework for cis-regulatory evolution in animals that emphasizes the importance of chromatin accessibility, the regulatory states operating in different cellular contexts, long-range 3D regulatory interactions and the regulatory rewiring by insertion and exaptation of TEs. This novel unifying view provides simple mechanistic explanations for the evolution of some common features of CREs, such as homotypic TFBS clustering, the abundance of redundant enhancers or the prevalence of

TEs exaptations at early totipotent stages. Furthermore, our model proposes that the specific regulatory architectures of a genomic region, together with the degree of complexity of the regulatory state of a given cellular context, are major determinants of the evolutionary trajectories and turnover propensity of their associated CREs. Finally, since different tissues and developmental stages will be characterized by varying degrees of complexity in the architecture of their active regulatory landscapes and defining regulatory states, we believe that our model will help to reconcile the observed discrepancies among current genome-wide comparative studies of cis-regulation in animals.

Figure 4. Influence of TAD borders position in regulatory information rearrangements. Regulatory information loss or gain within a gene landscape highly depends on the TAD borders distribution. A) An inversion covering a gene and some of its associated regulatory elements does not change their relative position with respect to TAD borders in the area. In this case, genes located at both sides of the TAD border maintain intact their regulatory landscapes. B) If the inversion covers the TAD border and an enhancer flips to the other side of the border, some regulatory information will change from one gene to its neighbor. Light blue squares represent inversions. Red bars represent inverted sections of the DNA. Colored ovals and arrows represent enhancers and genes, respectively. Purple rings represent proteins involved in regulatory isolation (TAD borders).

Acknowledgments

We thank J.L. Gómez-Skarmeta, J.R. Martinez-Morales, M. Irimia, R.R. Acemel and I. Almudi for fruitful discussions and critically reading the manuscript. J.J.T. is funded by a contract associated to Spanish government project (BFU2013-41322-P granted to J.L. Gómez-Skarmeta) and I.M. holds a Juan de la Cierva postdoctoral contract.

References

1. Carroll, S.B., *Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution*. Cell, 2008. **134**(1): p. 25-36.
2. Slattery, M., et al., *Absence of a simple code: how transcription factors read the genome*. Trends Biochem Sci, 2014. **39**(9): p. 381-99.
3. Todeschini, A.L., A. Georges, and R.A. Veitia, *Transcription factors: specific DNA binding and specific gene regulation*. Trends Genet, 2014. **30**(6): p. 211-9.
4. Villar, D., P. Flicek, and D.T. Odom, *Evolution of transcription factor binding in metazoans - mechanisms and functional implications*. Nat Rev Genet, 2014. **15**(4): p. 221-33.
5. Stergachis, A.B., et al., *Conservation of trans-acting circuitry during mammalian regulatory evolution*. Nature, 2014. **515**(7527): p. 365-70.
6. Vierstra, J., et al., *Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution*. Science, 2014. **346**(6212): p. 1007-12.
7. Cotney, J., et al., *The evolution of lineage-specific regulatory activities in the human embryonic limb*. Cell, 2013. **154**(1): p. 185-96.
8. Reilly, S.K., et al., *Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis*. Science, 2015. **347**(6226): p. 1155-9.
9. Tena, J.J., et al., *Comparative epigenomics in distantly related teleost species identifies conserved cis-regulatory nodes active during the vertebrate phylotypic period*. Genome Res, 2014. **24**(7): p. 1075-85.
10. Villar, D., et al., *Enhancer evolution across 20 mammalian species*. Cell, 2015. **160**(3): p. 554-66.
11. Engstrom, P.G., et al., *Genomic regulatory blocks underlie extensive microsynteny conservation in insects*. Genome Res, 2007. **17**(12): p. 1898-908.
12. Irimia, M., et al., *Contrasting 5' and 3' evolutionary histories and frequent evolutionary convergence in Meis/hth gene structures*. Genome Biol Evol, 2011. **3**: p. 551-64.
13. Irimia, M., et al., *Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints*. Genome Res, 2012. **22**(12): p. 2356-67.
14. Kikuta, H., et al., *Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates*. Genome Res, 2007. **17**(5): p. 545-55.
15. Maeso, I., et al., *Deep conservation of cis-regulatory elements in metazoans*. Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1632): p. 20130020.
16. Maeso, I., et al., *An ancient genomic regulatory block conserved across bilaterians and its dismantling in tetrapods by retrogene replacement*. Genome Res, 2012. **22**(4): p. 642-55.
17. Nelson, C.E., B.M. Hersh, and S.B. Carroll, *The regulatory content of intergenic DNA shapes genome architecture*. Genome Biol, 2004. **5**(4): p. R25.

18. Stone, J.R. and G.A. Wray, *Rapid evolution of cis-regulatory sequences via local point mutations*. Mol Biol Evol, 2001. **18**(9): p. 1764-70.
19. Georges, A.B., et al., *Generic binding sites, generic DNA-binding domains: where does specific promoter recognition come from?* FASEB J, 2010. **24**(2): p. 346-56.
20. Ballester, B., et al., *Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways*. Elife, 2014. **3**: p. e02626.
21. Robertson, G., et al., *Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing*. Nat Methods, 2007. **4**(8): p. 651-7.
22. Fisher, W.W., et al., *DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila*. Proc Natl Acad Sci U S A, 2012. **109**(52): p. 21330-5.
23. Sakabe, N.J. and M.A. Nobrega, *Beyond the ENCODE project: using genomics and epigenomics strategies to study enhancer evolution*. Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1632): p. 20130022.
24. Pasquali, L., et al., *Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants*. Nat Genet, 2014. **46**(2): p. 136-43.
25. Duque, T., et al., *Simulations of enhancer evolution provide mechanistic insights into gene regulation*. Mol Biol Evol, 2014. **31**(1): p. 184-200.
26. Duque, T. and S. Sinha, *What does it take to evolve an enhancer? A simulation-based study of factors influencing the emergence of combinatorial regulation*. Genome Biol Evol, 2015. **7**(6): p. 1415-31.
27. Levo, M., et al., *Unraveling determinants of transcription factor binding outside the core binding site*. Genome Res, 2015. **25**(7): p. 1018-29.
28. Zhou, T., et al., *Quantitative modeling of transcription factor binding specificities using DNA shape*. Proc Natl Acad Sci U S A, 2015. **112**(15): p. 4654-9.
29. Becker, P.B. and J.L. Workman, *Nucleosome remodeling and epigenetics*. Cold Spring Harb Perspect Biol, 2013. **5**(9).
30. Chen, T. and S.Y. Dent, *Chromatin modifiers and remodellers: regulators of cellular differentiation*. Nat Rev Genet, 2014. **15**(2): p. 93-106.
31. Panne, D., *The enhanceosome*. Curr Opin Struct Biol, 2008. **18**(2): p. 236-42.
32. Wasson, T. and A.J. Hartemink, *An ensemble model of competitive multi-factor binding of the genome*. Genome Res, 2009. **19**(11): p. 2101-12.
33. Lee, N. and J.A. Steitz, *Noncoding RNA-guided recruitment of transcription factors: A prevalent but undocumented mechanism?* Bioessays, 2015. **37**(9): p. 936-41.
34. Hancock, R., *The crowded nucleus*. Int Rev Cell Mol Biol, 2014. **307**: p. 15-26.
35. Nolin, F., et al., *Changes to cellular water and element content induced by nucleolar stress: investigation by a cryo-correlative nano-imaging approach*. Cell Mol Life Sci, 2013. **70**(13): p. 2383-94.
36. Buenrostro, J.D., et al., *ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide*. Curr Protoc Mol Biol, 2015. **109**: p. 21 29 1-9.
37. Thurman, R.E., et al., *The accessible chromatin landscape of the human genome*. Nature, 2012. **489**(7414): p. 75-82.
38. Bogdanovic, O., et al., *Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis*. Genome Res, 2012. **22**(10): p. 2043-53.
39. Bonn, S., et al., *Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development*. Nat Genet, 2012. **44**(2): p. 148-56.

40. Hawkins, R.D., et al., *Dynamic chromatin states in human ES cells reveal potential regulatory sequences and genes involved in pluripotency*. Cell Res, 2011. **21**(10): p. 1393-409.
41. Rada-Iglesias, A., et al., *A unique chromatin signature uncovers early developmental enhancers in humans*. Nature, 2011. **470**(7333): p. 279-83.
42. Creyghton, M.P., et al., *Histone H3K27ac separates active from poised enhancers and predicts developmental state*. Proc Natl Acad Sci U S A, 2010. **107**(50): p. 21931-6.
43. Kim, T.K., et al., *Widespread transcription at neuronal activity-regulated enhancers*. Nature, 2010. **465**(7295): p. 182-7.
44. Visel, A., E.M. Rubin, and L.A. Pennacchio, *Genomic views of distant-acting enhancers*. Nature, 2009. **461**(7261): p. 199-205.
45. Heintzman, N.D., et al., *Histone modifications at human enhancers reflect global cell-type-specific gene expression*. Nature, 2009. **459**(7243): p. 108-12.
46. Heintzman, N.D., et al., *Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome*. Nat Genet, 2007. **39**(3): p. 311-8.
47. Schwaiger, M., et al., *Evolutionary conservation of the eumetazoan gene regulatory landscape*. Genome Res, 2014. **24**(4): p. 639-50.
48. Kim, T.K. and R. Shiekhattar, *Architectural and Functional Commonalities between Enhancers and Promoters*. Cell, 2015. **162**(5): p. 948-59.
49. Andersson, R., *Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model*. Bioessays, 2015. **37**(3): p. 314-23.
50. Koshikawa, S., et al., *Gain of cis-regulatory activities underlies novel domains of wingless gene expression in Drosophila*. Proc Natl Acad Sci U S A, 2015. **112**(24): p. 7524-9.
51. Boyd, J.L., et al., *Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex*. Curr Biol, 2015. **25**(6): p. 772-9.
52. Rogers, W.A., et al., *Recurrent modification of a conserved cis-regulatory element underlies fruit fly pigmentation diversity*. PLoS Genet, 2013. **9**(8): p. e1003740.
53. Pearson, J.C., J.D. Watson, and S.T. Crews, *Drosophila melanogaster Zelda and Single-minded collaborate to regulate an evolutionarily dynamic CNS midline cell enhancer*. Dev Biol, 2012. **366**(2): p. 420-32.
54. Rebeiz, M., et al., *Evolutionary origin of a novel gene expression pattern through co-option of the latent activities of existing regulatory sequences*. Proc Natl Acad Sci U S A, 2011. **108**(25): p. 10036-43.
55. Prabhakar, S., et al., *Human-specific gain of function in a developmental enhancer*. Science, 2008. **321**(5894): p. 1346-50.
56. Eichenlaub, M.P. and L. Ettwiller, *De novo genesis of enhancers in vertebrates*. PLoS Biol, 2011. **9**(11): p. e1001188.
57. Dong, X., et al., *Exonic remnants of whole-genome duplication reveal cis-regulatory function of coding exons*. Nucleic Acids Res, 2010. **38**(4): p. 1071-85.
58. Birnbaum, R.Y., et al., *Coding exons function as tissue-specific enhancers of nearby genes*. Genome Res, 2012. **22**(6): p. 1059-68.
59. de la Calle-Mustienes, E., et al., *A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts*. Genome Res, 2005. **15**(8): p. 1061-72.
60. McEwen, G.K., et al., *Early evolution of conserved regulatory sequences associated with development in vertebrates*. PLoS Genet, 2009. **5**(12): p. e1000762.
61. Berman, B.P., et al., *Exploiting transcription factor binding site clustering to identify cis-*

- regulatory modules involved in pattern formation in the Drosophila genome*. Proc Natl Acad Sci U S A, 2002. **99**(2): p. 757-62.
62. Gotea, V., et al., *Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers*. Genome Res, 2010. **20**(5): p. 565-77.
 63. He, X., T.S. Duque, and S. Sinha, *Evolutionary origins of transcription factor binding site clusters*. Mol Biol Evol, 2012. **29**(3): p. 1059-70.
 64. Allo, M., et al., *Argonaute-1 binds transcriptional enhancers and controls constitutive and alternative splicing in human cells*. Proc Natl Acad Sci U S A, 2014. **111**(44): p. 15622-9.
 65. Becker, T.S. and B. Lenhard, *The random versus fragile breakage models of chromosome evolution: a matter of resolution*. Mol Genet Genomics, 2007. **278**(5): p. 487-91.
 66. Peterson, B.K., et al., *Big genomes facilitate the comparative identification of regulatory elements*. PLoS One, 2009. **4**(3): p. e4688.
 67. de Wit, E. and W. de Laat, *A decade of 3C technologies: insights into nuclear organization*. Genes Dev, 2012. **26**(1): p. 11-24.
 68. Dekker, J., *Two ways to fold the genome during the cell cycle: insights obtained with chromosome conformation capture*. Epigenetics Chromatin, 2014. **7**(1): p. 25.
 69. Gorkin, D.U., D. Leung, and B. Ren, *The 3D genome in transcriptional regulation and pluripotency*. Cell Stem Cell, 2014. **14**(6): p. 762-75.
 70. Ong, C.T. and V.G. Corces, *CTCF: an architectural protein bridging genome topology and function*. Nat Rev Genet, 2014. **15**(4): p. 234-46.
 71. Phillips-Cremins, J.E. and V.G. Corces, *Chromatin insulators: linking genome organization to cellular function*. Mol Cell, 2013. **50**(4): p. 461-74.
 72. Gomez-Marin, C., et al., *Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders*. Proc Natl Acad Sci U S A, 2015. **112**(24): p. 7542-7.
 73. Rao, S.S., et al., *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping*. Cell, 2014. **159**(7): p. 1665-80.
 74. Montavon, T., et al., *A regulatory archipelago controls Hox genes transcription in digits*. Cell, 2011. **147**(5): p. 1132-45.
 75. Gonzalez, F., D. Duboule, and F. Spitz, *Transgenic analysis of Hoxd gene regulation during digit development*. Dev Biol, 2007. **306**(2): p. 847-59.
 76. van Arensbergen, J., B. van Steensel, and H.J. Bussemaker, *In search of the determinants of enhancer-promoter interaction specificity*. Trends Cell Biol, 2014. **24**(11): p. 695-702.
 77. Spitz, F., F. Gonzalez, and D. Duboule, *A global control region defines a chromosomal regulatory landscape containing the HoxD cluster*. Cell, 2003. **113**(3): p. 405-17.
 78. Irimia, M., et al., *Ancient cis-regulatory constraints and the evolution of genome architecture*. Trends Genet, 2013. **29**(9): p. 521-8.
 79. Hong, J.W., D.A. Hendrix, and M.S. Levine, *Shadow enhancers as a source of evolutionary novelty*. Science, 2008. **321**(5894): p. 1314.
 80. Tena, J.J., et al., *An evolutionarily conserved three-dimensional structure in the vertebrate *Ir* clusters facilitates enhancer sharing and coregulation*. Nat Commun, 2011. **2**: p. 310.
 81. Edelman, L.B. and P. Fraser, *Transcription factories: genetic programming in three dimensions*. Curr Opin Genet Dev, 2012. **22**(2): p. 110-4.
 82. Razin, S.V., et al., *Transcription factories in the context of the nuclear and genome organization*. Nucleic Acids Res, 2011. **39**(21): p. 9085-92.

83. Cande, J., Y. Goltsev, and M.S. Levine, *Conservation of enhancer location in divergent insects*. Proc Natl Acad Sci U S A, 2009. **106**(34): p. 14414-9.
84. Hare, E.E., et al., *Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation*. PLoS Genet, 2008. **4**(6): p. e1000106.
85. Martinez-Morales, J.R., *Toward understanding the evolution of vertebrate gene regulatory networks: comparative genomics and epigenomic approaches*. Brief Funct Genomics, 2015.
86. Zaret, K.S. and J.S. Carroll, *Pioneer transcription factors: establishing competence for gene expression*. Genes Dev, 2011. **25**(21): p. 2227-41.
87. Putnam, N.H., et al., *The amphioxus genome and the evolution of the chordate karyotype*. Nature, 2008. **453**(7198): p. 1064-71.
88. McEwen, G.K., et al., *Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis*. Genome Res, 2006. **16**(4): p. 451-65.
89. Irimia, M., I. Maeso, and J. Garcia-Fernandez, *Convergent evolution of clustering of Iroquois homeobox genes across metazoans*. Mol Biol Evol, 2008. **25**(8): p. 1521-5.
90. Jimenez-Delgado, S., et al., *Evolutionary genomics of the recently duplicated amphioxus Hairy genes*. Int J Biol Sci, 2006. **2**(2): p. 66-72.
91. Fort, A., et al., *Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance*. Nat Genet, 2014. **46**(6): p. 558-66.
92. Gifford, W.D., S.L. Pfaff, and T.S. Macfarlan, *Transposable elements as genetic regulatory substrates in early development*. Trends Cell Biol, 2013. **23**(5): p. 218-26.
93. Macfarlan, T.S., et al., *Embryonic stem cell potency fluctuates with endogenous retrovirus activity*. Nature, 2012. **487**(7405): p. 57-63.
94. Schmidt, D., et al., *Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages*. Cell, 2012. **148**(1-2): p. 335-48.
95. Kunarso, G., et al., *Transposable elements have rewired the core regulatory network of human embryonic stem cells*. Nat Genet, 2010. **42**(7): p. 631-4.
96. Bourque, G., *Transposable elements in gene regulation and in the evolution of vertebrate genomes*. Curr Opin Genet Dev, 2009. **19**(6): p. 607-12.
97. Bourque, G., et al., *Evolution of the mammalian transcription factor binding repertoire via transposable elements*. Genome Res, 2008. **18**(11): p. 1752-62.
98. Sundaram, V., et al., *Widespread contribution of transposable elements to the innovation of gene regulatory networks*. Genome Res, 2014. **24**(12): p. 1963-76.
99. Richardson, S.R., et al., *The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes*. Microbiol Spectr, 2015. **3**(2): p. MDNA3-0061-2014.
100. Lynch, V.J., et al., *Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals*. Nat Genet, 2011. **43**(11): p. 1154-9.
101. Notwell, J.H., et al., *A family of transposable elements co-opted into developmental enhancers in the mouse neocortex*. Nat Commun, 2015. **6**: p. 6644.
102. Lynch, V.J., et al., *Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy*. Cell Rep, 2015. **10**(4): p. 551-61.
103. Domene, S., et al., *Enhancer turnover and conserved regulatory function in vertebrate evolution*. Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1632): p. 20130027.
104. Emera, D., et al., *Convergent evolution of endometrial prolactin expression in primates, mice, and elephants through the independent recruitment of transposable elements*.

- Mol Biol Evol, 2012. **29**(1): p. 239-47.
105. Sasaki, T., et al., *Possible involvement of SINEs in mammalian-specific brain formation*. Proc Natl Acad Sci U S A, 2008. **105**(11): p. 4220-5.
 106. Bejerano, G., et al., *A distal enhancer and an ultraconserved exon are derived from a novel retroposon*. Nature, 2006. **441**(7089): p. 87-90.
 107. Vietri Rudan, M., et al., *Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture*. Cell Rep, 2015. **10**(8): p. 1297-309.
 108. Farre, M., T.J. Robinson, and A. Ruiz-Herrera, *An Integrative Breakage Model of genome architecture, reshuffling and evolution: The Integrative Breakage Model of genome evolution, a novel multidisciplinary hypothesis for the study of genome plasticity*. Bioessays, 2015. **37**(5): p. 479-88.
 109. Irimia, M., et al., *Comparative genomics of the Hedgehog loci in chordates and the origins of Shh regulatory novelties*. Sci Rep, 2012. **2**: p. 433.
 110. Lettice, L.A., et al., *A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly*. Hum Mol Genet, 2003. **12**(14): p. 1725-35.