Published in final edited form as:
Nat Rev Genet. 2014 April ; 15(4): 221–233. doi:10.1038/nrg3481.

Dynamics, mechanisms, and functional implications of transcription factor binding evolution in metazoans

Diego Villar[1], **Paul Flicek**[2], and **Duncan T Odom**[1]

[1]University of Cambridge, Cancer Research UK – Cambridge Institute

[2]European Molecular Biology Laboratory, European Bioinformatics Institute

Synopsis

Transcription factor binding differences can contribute to organismal evolution by altering downstream gene expression programmes. Recent genome-wide studies in *Drosophila* and mammals have revealed common quantitative and combinatorial properties of *in vivo* DNA-binding, as well as significant differences in the rate and mechanisms of metazoan transcription factor binding evolution. Here, we review the recently-discovered, rapid re-wiring of *in vivo* transcription factor binding between related metazoan species and summarize general principles underlying the observed patterns of evolution. We then consider what might explain genome evolution differences between metazoan phyla, and outline the conceptual and technological challenges facing the field.

Complex, multicellular organisms require a means to create hundreds of distinct tissue types from a single genome. Most (if not all) of these tissues are shared among all known vertebrates[1, 2]; for instance, two tissues with distinctive morphologies and evolutionarily conserved functions are the heart, controlling blood flow, and the liver, controlling blood detoxification and circulating lipids. Vertebrate tissues have broadly conserved transcriptional programmes[3, 4], and are often known to be controlled by a highly conserved set of tissue-specific DNA binding transcription factors[5]. Such tissue-specific master regulators include the transcription factors MYOD1 in muscle, HNF4A in liver or NKX2-5 in heart, which have functional roles both in development to establish tissue identity and in adulthood to maintain tissue-specific functions.

It would be reasonable to suppose that the protein-DNA contacts that connect conserved transcription factors and (downstream) conserved tissue-specific gene expression programs are under strong constraint--a paradigm which has prompted the use of many diverse model organisms to model human regulatory and developmental processes. On the other hand, differences in gene regulation have long been recognized as major contributors to phenotypic diversity[6-8], especially between closely related species[9]. Only recently, and mainly due to the advent of functional genomics approaches such as ChIP-Seq, have we been able to experimentally test how widely TF binding differs between species and how

Correspondence to: Paul Flicek; Duncan T Odom.

Competing interests statement None

rapidly these differences accumulate, thus fundamentally reshaping our understanding of how transcription factor binding evolves and the potential consequences for how complex eukaryotic tissues are created.

Here, we will review recent advances in evolutionary analysis of transcription factor binding, with a focus on genome-wide studies in metazoans. We start by briefly discussing current views on how gene expression is controlled by transcription factors. We discuss the early studies that revealed how regulatory sequences controlling specific genes evolved, and the insights gained by sequence-based comparisons of substantial collections of genomes from diverse metazoans. Next, we summarize key findings from experimental ChIP-Seq studies on transcription factor binding evolution and highlight their novel conceptual contributions to models of regulatory evolution and gene expression control. Finally, we discuss how the differences in extent and rate of regulatory evolution among different eukaryotes likely reflect how population genetics acts as a driving force in genome evolution.

(1) Transcription factors and tissue identity

In all multicellular eukaryotes, cellular phenotype is largely dictated by the activity of tissue-specific transcription factors (TFs). Classical gain- and loss-of-function studies demonstrate that tissue-specific TFs often orchestrate the identity of the tissues in which they are selectively expressed. For example, genetic knock-down of MYOD1 and MYF5 (both muscle-specific TFs) in mice completely stalls skeletal muscle development; and forced MYOD1 expression in other cell types is sufficient to induce a muscle-specific gene expression profile[10]. Accordingly, mutations affecting either the sequence of tissue-specific TFs or sequences directly bound by TFs can cause disease[11, 12]: well-studied examples are mutations in HNF1A, HNF1B and HNF4A that cause maturity onset diabetes of the young (MODY)[13].

Determining the genomic regions bound by tissue-specific transcription factors and how they direct gene expression in a specific tissue and developmental time remains a daunting challenge. Classical transgenic studies in fruit flies and mammals have established a central paradigm of tissue-specific regulatory elements, namely, how specific regulatory elements in metazoans can drive transcription in a tissue-specific manner (reviewed in[14]). These studies have often been combined with bioinformatic predictions of TF binding locations in analyzed cis-regulatory modules (CRMs), often revealing clusters of likely recognition sequences. The recent approach to combine genome-wide computational and experimental analyses has complemented and extended site-directed studies---and thus led to refined models of gene regulation (Box 1).

(2) Sequence-based approaches

Site-directed transcription factor binding evolution

Because of the complexity in metazoan transcriptional regulation (Box 1), evolutionary analysis of regulatory sequences and their functional conservation (or lack thereof) has emerged as a powerful approach to infer gene control mechanisms. Several seminal studies analyzed the evolution of transcription factor recognition motifs by sequence comparison of known cis-regulatory modules between species (primarily at strongly conserved developmental enhancers), often in combination with *in vivo* analysis of the resulting gene expression patterns. Well-studied examples are the *endo16* promoter of sea urchins[15, 16] and *even-skipped* enhancers in fruit flies[17-19]. Although selective constraint was often inferred for some transcription factor binding sites, comparative sequence analyses suggested significant turnover of TF binding positions, even between closely related species. Despite this lack of sequence conservation in orthologous enhancers[17, 18], transgenic studies of *even-skipped* revealed conserved expression patterns, arguing for the occurrence of functionally compensatory mutations. Nevertheless, more detailed manipulation of these enhancers by functional complementation also suggested functional divergence. Similar studies in fish and mammals[20, 21] reported a poor correlation between sequence conservation and regulatory function. For example, RET human and zebrafish enhancers drove very similar expression patterns in zebrafish embryos, despite no detectable sequence conservation. These studies have collectively shown that regulatory function can be maintained in the complete absence of sequence conservation, raising the question of how transcription factor binding divergence can be reconciled with functional conservation.

Using collections of well-characterized CRMs, in which transcription factor binding sites have been inferred using alignment of orthologous noncoding sequences, signatures of both (i) neutral evolution and (ii) positive and purifying selection have been found in *Drosophila*[22, 23]. This result suggests that accumulation of regulatory sequence differences reflects a complex mixture of mechanisms. In mammals, the alignment of validated human CRMs to the mouse genome suggested large-scale, functional turnover of TF binding: where experimental data was available for both species, 30-42% of the human regions were not functional in rodents[24].

Complementary whole-genome approaches have recently been used to address two key limitations of the above studies: firstly, the bias towards previously-known CRMs, and secondly the absence of direct, experimental mapping of transcription factor binding in different species.

Whole-genome comparisons and regulatory constraint

Tremendous technological progress over the last decade has resulted in the sequencing of hundreds of metazoan genomes[25, 26]. Comparative analysis of whole genomes can identify specific sequences that have undergone evolutionary selection, such as protein-coding sequences and, to a lesser extent, putative regulatory control sequences[25, 27, 28]. As a tool to identify regulatory control sequences, this analytical strategy relies on the assumption that conserved non-coding sequences have been evolutionarily maintained to control specific gene expression patterns; in several cases, this assumption has been confirmed experimentally[29-31].

These studies have greatly improved our understanding of aspects of the sequence and functional constraints in metazoan genomes[25, 32-34]. Inferences from a recent comparison of 29 mammalian genomes estimate that 3-8% of the human genome is under negative (purifying) selection[25], most of which is presumed to correspond to non-coding regions with regulatory function. Constraint can be inferred in genetic sequences as short as 36 bp, which is comparable to the resolution of experimental ChIP-based TF binding maps. However, sequence-based annotation of constraint cannot resolve spatiotemporal patterns of transcription factor binding, and has limited power to detect novel sequence changes such as lineage-specific regulatory regions[35].

In contrast to mammals, similar analyses of collections of the more compact genomes of *Drosophila*, *C. elegans* and *S. Cerevisiae* species have predicted considerably larger fractions of the genomes to be under evolutionary constraint (37-53% for fruit flies, Table 1)[36-38]. When both the large variations in accessible genome size[39, 40] and the presence of a similar repertoire of genes[36] are considered, it is clear that different metazoans such as mammals and fruit flies have very different genome architectures (Figure 1).

There are two major reasons for the difference in the density of constrained DNA (Figure 1). First, almost a fifth of the fly genome codes for proteins[38], versus approximately 2% for mammals[25]. Second, mammalian genomes typically contain twice as many genes in approximately twenty times as much DNA, much of which is packaged into heterochromatin[40].

(3) Direct global mapping of TF binding

Transcription factor binding patterns can be compared at the whole-genome level by obtaining data from ChIP experiments in matched tissues or cells of different species (Table 1). This approach complements site-directed and computational studies by addressing specific transcriptional contexts such as developmental processes and tissue specificity. Chromatin immunoprecipitation methods also have the specific advantage of providing a quantitative estimate of TF binding, since a linear relationship exists between *in vivo* crosslinking efficiency and occupancy levels on DNA[41-43]. Moreover, in their genome-wide adaptations, ChIP approaches are unbiased as all regulatory regions are interrogated and thus can be included in downstream analyses. However, it is important to note that ChIP-Seq is extremely sensitive and can detect TF binding across a wide range of occupancy levels[43]. Regions bound at low occupancy likely include background binding, thought to be driven by relatively high concentrations per cell of many transcription factors[44]. Nearly all peak calling methods thus employ a statistical cutoff to differentiate biologically meaningful signal from experimental/technical noise, which limits precise cross-study comparisons. Furthermore, both statistical and biological evidence[43] suggests that chromatin immunoprecipitation captures a continuum of functional and nonfunctional TF binding events. It remains challenging to establish the functionality of a specific *in vivo* TF binding event, and we are currently unable to clearly differentiate functional from putatively non-functional/background binding, especially for weakly occupied sites[45].

The first studies taking this approach used oligonucleotide microarrays designed against orthologous regions of different species to evaluate TF binding conservation (reviewed in[46]). In an experiment specifically designed to measure conserved tissue-specific TF binding between mouse and human, profiling of four tissue-specific transcription factors revealed large-scale turnover of *in vivo* TF binding in livers of both species with 41 to 89%

of binding regions found to be species-specific[47]. Sequencing-based experiments have rapidly superseded DNA microarrays in interspecies comparisons of TF binding, and numerous recent studies have greatly increased our understanding of the rate and underlying mechanisms of transcription factor binding evolution in metazoans. Although similar analyses have begun to explore the evolutionary stability of histone marks (Box 2), this review focuses on recent discoveries in understanding the rate and mechanisms of TF binding evolution.

Developmental TF binding evolution in Drosophila embryos

A number of recent studies have examined transcription factor binding in embryos of related fruit fly species[48-50], mainly focusing on TFs involved in mesoderm development, such as Twist, Hunchback, Bicoid, and Zelda. Complete gains and losses of TF binding were relatively rare among *Drosophila* species, though pervasive quantitative differences in strength of binding at orthologous loci occurred frequently. In the closely related *D. melanogaster* and *D. yakuba*, 85-98% of binding positions were conserved in the two species[48] for a collection of six developmental regulators. Moreover, binding intensities of six TFs were strongly correlated, suggesting that indirect effects such as chromatin state or cooperativity significantly influence binding patterns within and between species (elements of this are reviewed in[44]).

In an independent study, 60% of the binding peaks for the mesodermal transcription factor Twist were found to be conserved between *D. melanogaster* and *D. pseudoobscura*, at an evolutionary distance estimated to be as divergent as human and chicken by dN/dS ratios[49]. This remarkably high conservation of Twist binding also included lower-occupancy peaks. Fully a third (34%) of Twist binding events were conserved at the exact same syntenic location in six highly divergent *Drosophila* species, and were preferentially located near known functional target genes. A recent report examining the binding of developmental TFs Bicoid, Giant, Hunchback and Krueppel across four *Drosophila* species found similar proportions (15-38%) of binding locations conserved across all species[50], and these were correlated with peak height, location proximal to genes and clustered binding of the other profiled factors. Turnover of binding locations between species was also apparent in this study, and contrasted with higher conservation of gene expression levels[50].

In these studies, a linear relationship was found between quantitative changes in binding and evolutionary distance, with a large proportion of altered binding being associated with turnover in transcription factor recognition sequences. For example, 19% of Twist binding losses were explained by genetic changes to specific motifs directly bound by Twist, and up to 50% of lost Twist events could potentially be explained when mutations in the motifs for partner transcription factors were considered[49].

On the whole, these studies found that developmental transcription factor binding must be under strong constraint in divergent *Drosophila* species (Figure 2)—and contrasted with TF binding evolution results obtained in mammals[47, 51].

Tissue-specific TF binding evolution in mammals

In mammals, studies of transcription factor binding evolution focusing on tissue- or cell-type specific transcription factors have revealed both similarities with the mechanisms driving regulatory evolution in insects, as well as surprising differences in the rate and extent of TF binding divergence—and the forces shaping these differences.

To address how OCT4 and NANOG binding varies between human and mouse embryonic stem cells[51], ChIP-Seq occupancy data was compared with gene expression profiles obtained after OCT4 depletion: although the binding of OCT4 and NANOG was enriched in the vicinity of genes downregulated upon OCT4 depletion in both human and mouse, the precise location of these binding events was often not conserved. In agreement with data from *Drosophila*[17], this study indicates that compensatory changes in TF binding must occur through evolution to maintain similar transcriptional outputs, and further suggests that TF binding may co-evolve combinatorially. Moreover, a similar relationship has been observed in mammals and fruit flies between TF binding variation and changes in the directly bound sequences. Comparison of ChIP-Seq data for the liver-specific TFs CEBPA and HNF4A in human, mouse and dog found 60-85% of binding losses associated to sequence changes in the underlying sequence, and one third of these events had nearby binding events that could be compensatory[52].

Closer evolutionary distances across five mouse species (1-6 MY) were analyzed in a recent report for the genomic binding of CEBPA, HNF4A and FOXA1[53]. The higher resolution and quantitative nature of this data revealed that, as in *Drosophila*, combinatorial TF binding in mammals co-evolves in clusters, and there exists a clear correlation between binding intensity and evolutionary conservation. Moreover, genetic deletion of CEBPA or HNF4A led to loss of co-bound partner TFs in one third of co-bound clusters. Clusters that were more sensitive to genetic deletion also showed sensitivity to evolutionary changes in TF binding motifs across mouse species; for instance, clusters lost after HNF4A deletion were often lost via sequence variant in the HNF4A binding motif in one of the examined species. Furthermore, when compared to *Mus musculus*, a quarter of TF binding peaks that were absent in *Mus caroli* could be associated with genetic variation in the directly bound sequences. On the whole, the features of TF binding evolution—such as strong association with genetic changes, putatively compensatory turnover, combinatorial co-evolution of binding intensity—shared between *Drosophila* and mammals likely reflect the underlying biochemistry and biophysics of protein-DNA interactions.

Contrast in regulatory evolution between mammals and insects

Cross-species studies in *Drosophila* and mammals have also highlighted two perhaps surprising differences that strongly differentiate the activity of mammalian TF evolution from the high-conservation found in *Drosophila*. First, studies on mammalian evolution of tissue-specific TFs have consistently reported much more rapid turnover of binding positions compared to *Drosophila* developmental TFs (Table 1). In liver tissue from five vertebrates (human, macaque, mouse, opossum and chicken), less than fifty CEBPA binding events were ultraconserved in orthologous locations in all five species out of the tens of thousands identified in each species[52]. Even over closer evolutionary distances, mammalian

TF binding variation accumulates rapidly: an exponential relationship was found between evolutionary distance and conservation of TF binding locations for the liver-specific TFs CEBPA, HNF4A and FOXA1 in five closely-related mouse species[53]. Second, the association between conservation of TF binding and regulatory function reported in *Drosophila*[44, 49] seems considerably weaker in mammalian tissues. Across five vertebrates, shared binding events occurring in at least two species were found enriched near functional targets of these factors (as determined by loss-of-function studies), but the bound genomic regions did not show a corresponding increase in sequence constraint[52]. Over closer evolutionary distances, no clear association was found between binding intensity or conservation and functionality for three liver-specific TFs[53]: conserved intensity binding events showed no enrichment at known target genes nor obvious association with liver-related functions.

In summary, genome-wide studies of tissue-specific TF binding evolution in mammals has found concordant biophysical principles with those described in *Drosophila*, but have simultaneously revealed significant differences in the evolutionary stability of TF binding locations (Figure 2) and their association with functionality (see also discussion in Box 3).

CTCF binding evolution in metazoans

Certain transcriptional regulators, such as CTCF, thought to be involved in genome insulation and chromatin loop formation across all tissues, are shared between mammals and fruit flies. Recent studies in each phylum have been published comparing the genome-wide binding in multiple species, providing a useful (and direct) comparison of TF binding evolution between mammals and insects[54, 55].

In contrast to the restricted expression of developmental and tissue-specific transcription factors like Bicoid and HNF4A from above, CTCF is ubiquitously expressed across tissues and developmental states (reviewed in[56]). Notably, in fruit flies, CTCF is one of several known insulator proteins[57, 58], whereas in mammals it is the sole known factor known to regulate genome insulation[56]. Together with cohesins, CTCF[59, 60] (reviewed in[61]) is a central component of chromatin organization that has been the subject of extensive investigation using integrative approaches[62-64].

High-throughput interrogation of CTCF binding locations in different cell types[65] and tissues[66] found that most binding events are tissue-invariant, a property that contrasts with tissue-specific transcription factors (however, see also[67]). Studies focusing on inter-mammalian comparisons[51, 52, 54, 68] revealed that CTCF genomic locations are also more conserved across species than those of most site-specific TFs investigated to date. These findings likely reflect the essential, conserved functions of CTCF, whose binding can often demarcate regulatory domains[68, 69].

In mammals, the most evolutionary diverse inter-species comparison to date profiled CTCF binding in six mammalian species[54]. In agreement with previous reports, this study found highly conserved binding. For example, in human, dog and mouse, CTCF binding events were shared five times more often than binding locations for the tissue-specific TF CEBPA[52], while 60-70% of CTCF binding sites in each of six primates were observed in human[70]. The general mechanism to create new CTCF binding events appears to be via its carriage by specific repeat families (see below), as previously suggested in mouse[51, 71].

Analysis of CTCF binding in four *Drosophila* species[55] found signatures of both purifying and positive selection in the evolution of CTCF binding, and new-born CTCF binding events were correlated with changes in gene expression. In contrast to mammalian data, somewhat higher binding divergence was found for CTCF than for previously-studied *Drosophila* developmental TFs[48, 49]. The differing patterns of CTCF evolution in these two metazoan phyla could be due, at least partially, to different mechanisms of evolution: no clear association was found between CTCF binding evolution in *Drosophila* and the expansion of transposable elements, while compelling evidence points to this mechanisms in mammalian systems[51, 54, 72, 73]. Moreover, the additional presence of multiple insulators in *Drosophila* other than CTCF (such as BEAF or CP190)[57, 58] might relax evolutionary constraint of CTCF binding evolution compared to mammals.

(4) Sources of metazoan TF binding divergence

Cross-species comparisons of TF binding in metazoans have afforded a number of general insights into the evolutionary origins of TF binding differences between species and individuals and the rules governing TF binding divergence (Figure 3).

Cross-species sequence differences and TF binding

How often are genetic differences in the known motifs that are directly bound by transcription factors responsible for differences in TF binding between species? The comparisons done to date in closely related insect and mammalian species (Table 1) suggest that, at best, a substantial minority of TF binding differences can be attributed to alterations in directly-bound genetic sequences. Other studies in yeast[74] and human cell lines[75, 76] have indicated similar results: namely, that many differences in TF binding can occur in the absence of proximal sequence changes.

However, evidence does exist that the complete ensemble of regulatory sequences may well be ultimately responsible for TF binding differences. Comparison of human chromosome 21 in human liver and in liver tissue from an aneuploid mouse model of trisomy 21 allowed dissection of the relative contributions of genetic sequence versus cellular environment to tissue-specific transcription[77]. The binding locations of three transcription factors (HNF1A, HNF4A and HNF6/ONECUT1) in livers from these mice, carrying a segregating copy of human chromosome 21, were compared with matched experiments in human liver. Almost all transcription factor binding on human chromosome 21 in normal human hepatocytes is recapitulated in the mouse environment by the orthologous transcription factors encoded in the mouse genome. Thus, sufficient information must be encoded in the genetic sequence of human chromosome 21 to recreate transcription factor binding in the corresponding mouse tissues, indicating that differences in the cellular environment between human and mouse tissues contribute significantly less than the DNA sequence itself to transcription factor binding. Other mechanistic studies in yeast and humans have also suggested that the majority of TF binding variation stems from genetic sequence differences, rather than environmental or *trans* effects (reviewed in[46, 78]). Moreover, complementary work on chromatin accessibility changes within[79] and between species[80] implicate variations in chromatin status, such as allele-specific changes in TF binding, in mediating that at least part of the observed TF binding differences.

Mutation of bound sequences and TF binding differences

Most transcription factors bind short and degenerate sequences, and theoretical models based on neutral evolution show that binding sites can arise on relative short timescales upon accumulation of base pair substitutions in a similar sequence [81]. The studies discussed above have shown that a substantial fraction, but probably not most, binding divergence in metazoans can be associated with differences in the underlying sequence, including base-pair substitutions, indels and gaps in the alignment (Figure 3A). For example, the tissue-specific TFs CEBPA and HNF4A bind 10 nucleotide recognition sequences (the average length for binding sequences of eukaryotic TFs), and similar proportions (40-50%) of their *in vivo* binding regions presented underlying point mutations in a second species that could explain the observed absence of binding[52]. Studies looking at the effect of human genetic variation on TF binding[75, 76] also suggest that TF binding divergence partially stems from sequence changes in the bound genetic sequence, as evidenced by an enrichment of TF motif-disrupting mutations in differentially bound sites (whether across species[48-50, 53]) or individuals[75, 76].

However, these studies also indicate that sequence changes in the canonical TF binding motif only explain a minority (12-40%) of TF binding variation. Direct interrogation of several transcription factors (often known to bind combinatorially) in the same study[48, 49, 53, 82] indicates that a substantial fraction of TF binding variation can be explained by disruption in proximal, but not directly bound, TF binding motifs (Figure 3B). For example, a recent study focusing on strain-specific PU.1 and CEBPA binding in macrophages from two mouse inbred strains[82] showed that, while 41% of strain-specific PU.1 binding associated with strain-specific mutations in the PU.1 motif, an additional 15% of strain-specific PU.1 binding could be explained by proximal mutations in CEBPA or AP-1 motifs. Furthermore, and as discussed in the previous section, ChIP-Seq experiments in CEBPA and HNF4A knock-out mice[53] provided direct genetic evidence that TF binding diverge is often a result of altered binding in proximally bound genetic sequences. The effect of genetically knocking-out one factor (i.e. CEBPA) had a strong effect on associated combinatorial binding of the other assayed factors (HNF4A and FOXA1), and the sensitivity to genetic knock-down of a particular TF binding cluster correlated with its evolutionary stability across mouse species.

Repeat-driven expansion of TF binding sites

Whereas point mutations are expected to rapidly create and disrupt shorter TF binding motifs, longer binding sequences could be disrupted, but rarely born, in this manner[81]. Stronger protein-DNA contacts occurring at longer motifs are predicted to be more resilient to genetic drift[83, 84].

A second mechanism to introduce TF binding motifs into large and complex metazoan genomes is the expansion of transposable elements (TEs) (reviewed in[85]), and TE-derived genome content is particularly high in mammals[71, 86]**.** For instance, two studies have highlighted the role that ERV1 repeats have played in the evolution of transcriptional regulation. The detailed analysis of the repeat content of *in vivo* TP53 binding sites in human cells showed that 30% of occupied regions contained primate-specific ERV1 repeats[87]. In addition, OCT4 and NANOG bound regions in human and mouse ES cells also showed significant repeat-element association, which appeared to account for 7-28% of the total TF binding sites[51]. For OCT4 and NANOG, these repeat-associated binding events were mostly species-specific, and ERV1 repeats were the largest contributor of TF binding sequences.

SINE elements have also been implicated in large-scale genome and transcriptional regulatory evolution. A recent study on CTCF binding evolution in six mammalian species[54] found specific sets of motif words bound by CTCF *in vivo* to be embedded in lineage-specific SINE transposons in rodents (mouse and rat), carnivores (dog), and *Didelphimorphia* marsupials (opossum), representing 180 million years of divergence. This observation, combined with the identification of fossilized repeats around some ultraconserved CTCF binding events, suggested that repeat-driven birth for novel CTCF binding events is a shared and ancient mechanism among mammals, although this mechanism has been largely quiescent in primates[70]. Important support for this idea is the observation that newborn motifs appeared to demarcate chromatin and transcriptional domains with a similar frequency as ancient, deeply conserved binding events. The recurrent expansions of retrotransposons has sculpted the CTCF binding landscape over hundreds of millions of years of mammalian (and, most likely, vertebrate) evolution.

An emerging common feature in these studies is a long binding motif for the associated TFs (Figure 3C), likely because longer recognition sequences cannot readily arise by simple point mutations[81]. Such a repeat-carried expansion mechanism, however, may well be active for TFs that bind short motifs as well[72, 88]. Repeat expansions can potentially create highly complex TF binding sequences when a near-perfect match of a TF recognition sequence exists within a repeat family. As has been documented for NSRF[89], transposable elements can carry a low-affinity consensus sequence that can be refined into a high-affinity site with a few key mutations (Figure 3C). The exaptation of selfishly expanding nucleic acids into regulatory sequences that are thus integrated into the functional mammalian genome is a remarkable example of how a host can productively repurpose the selfish DNA of repetitive sequences[90] (for related discussion see[91]).

In comparison to mammals, the contribution of TE expansions to TF binding divergence in *Drosophila* has not been analyzed in detail. Although a few studies have tested the association between experimentally bound regions and specific repeat classes[55], no clear correlations have been conclusively reported. This is likely a reflection of the lower TE genome content in *Drosophila* and other invertebrates versus vertebrate genomes[92], which has been proposed to be a consequence of more efficient selection against transposons in *Drosophila* compared to vertebrates[92-95] (see also discussion in Box 3).

Evolutionary forces and TF binding divergence

What evolutionary forces contribute to these sequence differences? As discussed in the previous section, signatures of both purifying and positive selection have been found through site-directed or whole-genome comparisons of non-coding regions[23], as well as in genome-wide ChIP-Seq studies across species[48, 49]. However, whole-genome interrogation of TF binding evolution has also suggested that many genetic differences in these directly bound sequences are likely a result of nonadaptive forces of evolution such as genetic drift, mutation and recombination—in agreement with the neutral theory of evolution originally proposed by Kimura[96].

(5) Conclusions and future perspectives

The application of high-throughput technologies to comparatively map binding positions of regulatory proteins across related species in different phyla (Table 1) has provided unbiased novel insights into the genomic and molecular complexity of tissue-specific transcriptional programmes and the evolutionary mechanisms that drive regulatory divergence (Figure 3). Comparative genomics studies in diverse metazoans have revealed how the interplay of the continuous genetic drift, mutation, recombination, and retroelement expansion, shaped by natural selection, results in a rapidly evolving regulatory landscape.

From a population genetics perspective, the smaller effective population sizes in mammals should increase their susceptibility to accumulation of neutral--and potentially deleterious--DNA, while selection may overcome drift in insects with considerably larger breeding populations. Ultimately, the lower constraint on mammalian genomes likely explains the different rates of TF binding evolution that have been observed in these two phyla (Figure 2 and Table 1). The rapid evolution of tissue-specific TF binding sites in mammals has also important implications for identifying and understanding human disease-associated non-coding variants. Extensive turnover of TF binding sites suggests that many functional sites will have migrated into lineage-specific sequences that are largely invisible to phylogenetic footprinting[35], potentially undermining attempts to prioritize GWAS hits by underlying sequence constraint. Direct experimental data, ultimately in the correct cell types relevant to a specific disease, will be needed to interpret the molecular disease mechanisms of human genomic variants[97-99].

Despite significant advances in our understanding of metazoan regulatory evolution and its mechanistic basis, many questions remain unresolved. A daunting challenge in the field arguably comes (especially for mammals) from the vastness of metazoan regulatory genomes[100], as well as the combinatorial complexity of tissue-specific transcriptional programs (Box 1).

First, further comparative studies will be needed to address outstanding questions, including: How extensive is regulatory divergence across different classes of regulatory proteins? What are the molecular mechanisms driving these evolutionary differences across different lineages and phyla? How do these mechanisms vary between tissues? These questions remain poorly explored in most species, but fortunately, new technological developments make more detailed studies feasible[101].

Second, despite the major insights that comparative ChIP-Seq analysis of TF binding has provided on regulatory evolution, complementary approaches are needed. For instance, only a few studies exist that disrupt or (more interestingly) genetically re-engineer metazoan regulatory elements (reviewed in[14, 102]. While functional screening of CRMs is comparatively well developed in *Drosophila* and other invertebrates, newly reported genome engineering methodologies in mammals could revolutionize our ability to understand and test the genetic features that differentiate functional from non-functional regulatory elements[103-106]. Perhaps the greatest challenge will be to integrate new experimental methods, such as high-throughput functional perturbations and synthetic biology[107-109] together with an understanding of the regulatory networks active in homologous tissues in multiple species. Such an integrated synthesis would be a powerful approach to mechanistically dissect, quantitatively understand, and successfully manipulate the connections between genetic regulatory sequences and metazoan phenotypes.

Acknowledgments

We thank J.C. Marioni (European Bioinformatics Institute, Cambridge), Anders Eriksson (Evolutionary Ecology Group, Department of Zoology, University of Cambridge), members of the Odom and Flicek laboratories, in particular C. Kutter, K. Stefflova and E. Wong, and three anonymous reviewers for their comments on the manuscript. D.V. thanks Cancer Research UK for funding. This work has been funded by the European Research Council, EMBO Young Investigators Programme, and Cancer Research UK (DTO); EMBL, the Wellcome Trust, and the European Union (PF).

Highlighted references

Ref #15 (Romano and Wray 2003 Development)

Focusing on a well-characterized promoter in sea urchins, the authors showed a largely conserved transcription pattern despite extensive divergence in the promoter sequences of the two species analyzed.

Ref #25 (Lindblad-Toh 2011 Nature)

Sequenced and aligned the genomes of 29 carefully selected mammals, implementing earlier theoretical models to infer, at high resolution and confidence, the constraint of sequence elements in the human genome.

Ref #29 (Pennachio 2006 Nature)

Exploiting human-pufferfish and human-mouse-rat sequence conservation, this study experimentally evaluated the regulatory potential of conserved non-coding sequences in a transgenic mouse enhancer assay.

Ref #36 (Siepel 2005 Genome Res)

A uniform method for estimation of evolutionary conserved elements across groups of related metazoan species, highlighting varying degrees of genome compaction and constraint in metazoans ranging from mammals to yeast.

Ref #45 (Fisher 2012 PNAS)

Convincingly argued that Drosophila genomic regions bound at low occupancy by a set of developmental transcription factors show low functional activity and may not be involved in cis-regulation of transcription.

Ref #48 (Bradley 2010 PLoS Biology)

Documented the high conservation of TF binding locations for five developmental TFs in two Drosophila species, as well as the striking co-evolution of their binding intensities.

Ref #49 (He 2011 Nat genetics)

A demonstration of very high conservation of Twist binding across five fruit fly species, with evolutionary distances estimated to be as divergent as those between human and chicken.

Ref #51 (Kunarso 2010 Nature Genetics)

Combined comparative ChIP-Seq analysis of TF binding in human and mouse embryonic stem cells with gene expression and perturbation studies to show the rapid evolution of TF binding locations and their potentially compensatory turnover.

Ref #52 (Schmidt 2010 Science)

Compared TF binding across divergent vertebrates, revealing extensive turnover of regulatory elements and few deeply shared TF binding sites in vivo.

Ref #54 (Schmidt 2012 Cell)

An extensive analysis of mechanisms of CTCF binding evolution in mammals, showing the large contribution of transposable elements to changes in CTCF binding.

Ref #55 (Ni 2012 PLoS Biology)

Analyzed CTCF binding evolution in fruit fly species and showed rapid evolution of its binding locations, compared to cross-species studies of the same protein in mammals.

Ref #77 (Wilson 2008 Science)

Demonstrated that regulatory sequences are largely sufficient to direct transcriptional programs, even when the cellular environment changes using an unusual mouse model.

Author biographies

Duncan T Odom

Dr Odom obtained his PhD in bioinorganic chemistry from the California Institute of Technology in 2001, and then served five years as a postdoctoral fellow at the Whitehead Institute / MIT. Since 2006, his laboratory at the University of Cambridge has been exploring the functional evolution of mammalian genomes.

Paul Flicek

Dr Flicek earned the Doctor of Science degree from Washington University in 2004 and joined the European Bioinformatics Institute in 2005 initially as a postdoctoral researcher and, since 2007, as head of Vertebrate Genomics and the Ensembl Team. He is a Senior Scientist of the European Molecular Biology Laboratory.

Diego Villar

Dr Villar obtained his PhD in molecular biology and biomedicine from Universidad Autónoma de Madrid (Spain) in 2010, and has been a postdoctoral fellow at the University of Cambridge since 2011. His research interests have focused on the specificity of protein and protein-DNA interactions in mammalian gene regulation.

Glossary terms

Accessible genome Segments of DNA sequence that lie in an open chromatin environment, based on the biophysics of protein-DNA interactions that can occur in these regions. Open or accessible chromatin can be readily bound by transcription factors and other effectors of the transcriptional machinery. Accessible regions are both ubiquitous and tissue-specific and can be inferred from experimental approaches such as DNAse I hypersensitivity or ChIP-Seq.

Average genomic diversity Average synonymous nucleotide heterozygosity, a measure of the number of heterozygotes in a population and, hence, of genomic diversity. It is predicted to decrease in populations with smaller effective population size (e.g. is higher in Drosophila compared to mammals).

ChIP-Seq Chromatin immunoprecipitation coupled to high-throughput sequencing. This technique identifies potential regulatory sequences that are bound by a protein of interest, and is based on immunoprecipitation of covalently-crosslinked chromatin complexes using antibodies against a specific DNA-binding protein.

Cis-regulatory modules (CRMs) discrete arrangements of transcription factor binding sites in the DNA sequence, often containing motifs for several transcription factor proteins. These can be defined using computational predictions and also be investigated through experimental approaches such as ChIP-Seq. Definition of CRMs is often very useful to pinpoint functional regulatory elements.

Effective population size effective number of gametes sampled per generation. The effective population size determines the rate of change in the composition of a population caused by genetic drift.

Exaptation evolutionary co-option of a functionally unrelated DNA sequence for a novel function. This process has been specifically studied for transposable elements, which (in spite of their exogenous origin) are often functionally adopted by the host genome, e.g. as regulatory sequences.

Fossilized repeat ancient repeat events that are (at least partially) visible based on their consensus sequence. Exapted repeat instances (e.g. regulatory elements) derived from transposable elements often become fossilized and have been identified among evolutionarily conserved sequences.

Genetic drift evolutionary change involving random sampling of genetic variants in a finite population, causing the composition of the offspring and parental generations to differ. This process constitutes a ubiquitous source of evolutionary stochasticity.

Neutral evolution a pattern of evolutionary change consistent with random drift of mutant alleles that are neutral or nearly neutral. The neutral theory of evolution states that the dynamics of the majority of changes observed at the molecular level are governed by nonadaptive evolutionary forces, rather than Darwinian (i.e. Positive) natural selection.

Non-adaptive evolutionary forces features of the population-genetic environment that operate in a stochastic manner. These include random genetic drift, recombination and mutation, and the relative power of these forces conditions the types of evolutionary changes that are possible in various contexts.

Non-synonymous to synonymous polymorphisms ratio Lower in larger populations. Reflects the lower probability of segregation of slightly deleterious mutations versus adaptive ones (in other words, the increased efficiency of selection versus drift).

Positive selection also termed directional selection, it is a mode of natural selection that pushes the phenotype towards an extreme, causing the allele frequency to shift over time towards that phenotype. Comparative genomics approaches can often infer positive selection by detecting directional patterns of nucleotide substitutions across species.

Purifying (negative) selection Natural selection against individuals that deviate from an intermediate optimum; this process tends to stabilize the phenotype. Genomics segments that have been subject to purifying selection can be inferred from nucleotide substitution patterns in aligned genomes of multiple species.

Transposable element/retrotransposon A DNA sequence of exogenous origin that inserts itself and can change its position in the genome, thereby altering genome structure and ultimately genome size. A large fraction of mammalian genomes is thought to be derived from transposable elements.

REFERENCES

1. Arendt D. The evolution of cell types in animals: emerging principles from molecular studies. Nat Rev Genet. 2008; 9:868–82. [PubMed: 18927580]
2. Shubin N, Tabin C, Carroll S. Deep homology and the origins of evolutionary novelty. Nature. 2009; 457:818–23. [PubMed: 19212399]
3. Chan ET, et al. Conservation of core gene expression in vertebrate tissues. J Biol. 2009; 8:33. [PubMed: 19371447]
4. Brawand D, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011; 478:343–8. [PubMed: 22012392]
5. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009; 10:252–63. [PubMed: 19274049]
6. Britten RJ, Davidson EH. Gene regulation for higher cells: a theory. Science. 1969; 165:349–57. [PubMed: 5789433]
7. Britten RJ, Davidson EH. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. Q Rev Biol. 1971; 46:111–38. [PubMed: 5160087]
8. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science. 1975; 188:107–16. [PubMed: 1090005]
9. Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. Nat Rev Genet. 2012; 13:505–16. [PubMed: 22705669]
10. Weintraub H. The MyoD family and myogenesis: redundancy, networks, and thresholds. Cell. 1993; 75:1241–4. [PubMed: 8269506]
11. Engelkamp D, van Heyningen V. Transcription factors in disease. Curr Opin Genet Dev. 1996; 6:334–42. [PubMed: 8791518]
12. Wray GA. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 2007; 8:206–16. [PubMed: 17304246]
13. Ryffel GU. Mutations in the human genes encoding the transcription factors of the hepatocyte nuclear factor (HNF)1 and HNF4 families: functional and pathological consequences. J Mol Endocrinol. 2001; 27:11–29. [PubMed: 11463573]
14. Haeussler M, Joly JS. When needles look like hay: how to find tissue-specific enhancers in model organism genomes. Dev Biol. 2011; 350:239–54. [PubMed: 21130761]
15. Romano LA, Wray GA. Conservation of Endo16 expression in sea urchins despite evolutionary divergence in both cis and trans-acting components of transcriptional regulation. Development. 2003; 130:4187–99. [PubMed: 12874137]
16. Balhoff JP, Wray GA. Evolutionary analysis of the well characterized endo16 promoter reveals substantial variation within functional sites. Proc Natl Acad Sci U S A. 2005; 102:8591–6. [PubMed: 15937122]
17. Ludwig MZ, Bergman C, Patel NH, Kreitman M. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature. 2000; 403:564–7. [PubMed: 10676967]
18. Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genet. 2008; 4:e1000106. [PubMed: 18584029]
19. Ludwig MZ, et al. Functional evolution of a cis-regulatory module. PLoS Biol. 2005; 3:e93. [PubMed: 15757364]

20. Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science. 2006; 312:276–9. [PubMed: 16556802]
21. McGaughey DM, et al. Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b. Genome Res. 2008; 18:252–60. [PubMed: 18071029]
22. Kim J, He X, Sinha S. Evolution of regulatory sequences in 12 Drosophila species. PLoS Genet. 2009; 5:e1000330. [PubMed: 19132088]
23. He BZ, Holloway AK, Maerkl SJ, Kreitman M. Does positive selection drive transcription factor binding site turnover? A test with Drosophila cis-regulatory modules. PLoS Genet. 2011; 7:e1002053. [PubMed: 21572512]
24. Dermitzakis ET, Clark AG. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol. 2002; 19:1114–21. [PubMed: 12082130]
25. Lindblad-Toh K, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011; 478:476–82. [PubMed: 21993624]
26. Genome 10K community, o.s. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J Hered. 2009; 100:659–74. [PubMed: 19892720]
27. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE. Human-mouse genome comparisons to locate regulatory sites. Nat Genet. 2000; 26:225–8. [PubMed: 11017083]
28. Pollard KS, et al. An RNA gene expressed during cortical development evolved rapidly in humans. Nature. 2006; 443:167–72. [PubMed: 16915236]
29. Pennacchio LA, et al. In vivo enhancer analysis of human conserved non-coding sequences. Nature. 2006; 444:499–502. [PubMed: 17086198]
30. Woolfe A, et al. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 2005; 3:e7. [PubMed: 15630479]
31. Prabhakar S, et al. Close sequence comparisons are sufficient to identify human cis-regulatory elements. Genome Res. 2006; 16:855–63. [PubMed: 16769978]
32. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010; 20:110–21. [PubMed: 19858363]
33. Ponting CP, Hardison RC. What fraction of the human genome is functional? Genome Res. 2011; 21:1769–76. [PubMed: 21875934]
34. Ward LD, Kellis M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science. 2012; 337:1675–8. [PubMed: 22956687]
35. Alfoldi J, Lindblad-Toh K. Comparative genomics as a tool to understand evolution and disease. Genome Res. 2013; 23:1063–8. [PubMed: 23817047]
36. Siepel A, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005; 15:1034–50. [PubMed: 16024819]
37. Andolfatto P. Adaptive evolution of non-coding DNA in Drosophila. Nature. 2005; 437:1149–52. [PubMed: 16237443]
38. Clark AG, et al. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 2007; 450:203–18. [PubMed: 17994087]
39. Li XY, et al. The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. Genome Biol. 2011; 12:R34. [PubMed: 21473766]
40. Boyle AP, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008; 132:311–22. [PubMed: 18243105]
41. Cao Y, et al. Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. Dev Cell. 2010; 18:662–74. [PubMed: 20412780]
42. Carr A, Biggin MD. A comparison of in vivo and in vitro DNA-binding specificities suggests a new model for homeoprotein DNA binding in Drosophila embryos. EMBO J. 1999; 18:1598–608. [PubMed: 10075930]
43. MacArthur S, et al. Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. Genome Biol. 2009; 10:R80. [PubMed: 19627575]

44. Biggin MD. Animal transcription networks as highly connected, quantitative continua. Dev Cell. 2011; 21:611–26. [PubMed: 22014521]

45. Fisher WW, 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; 109:21330–5. [PubMed: 23236164]

46. Dowell RD. Transcription factor binding variation in the evolution of gene regulation. Trends Genet. 2010; 26:468–75. [PubMed: 20864205]

47. Odom DT, et al. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat Genet. 2007; 39:730–2. [PubMed: 17529977]

48. Bradley RK, et al. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. PLoS Biol. 2010; 8:e1000343. [PubMed: 20351773]

49. He Q, et al. High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species. Nat Genet. 2011; 43:414–20. [PubMed: 21478888]

50. Paris M, et al. Extensive divergence of transcription factor binding in Drosophila embryos with highly conserved gene expression. PLoS Genet. 2013; 9:e1003748. [PubMed: 24068946]

51. Kunarso G, et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet. 2010; 42:631–4. [PubMed: 20526341]

52. Schmidt D, et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science. 2010; 328:1036–40. [PubMed: 20378774]

53. Stefflova K, et al. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. Cell. 2013; 154:530–40. [PubMed: 23911320]

54. Schmidt D, et al. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell. 2012; 148:335–48. [PubMed: 22244452]

55. Ni X, et al. Adaptive Evolution and the Birth of CTCF Binding Sites in the Drosophila Genome. PLoS Biol. 2012; 10:e1001420. [PubMed: 23139640]

56. Phillips JE, Corces VG. CTCF: master weaver of the genome. Cell. 2009; 137:1194–211. [PubMed: 19563753]

57. Hou C, Li L, Qin ZS, Corces VG. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. Mol Cell. 2012; 48:471–84. [PubMed: 23041285]

58. Schwartz YB, et al. Nature and function of insulator protein binding sites in the Drosophila genome. Genome Res. 2012; 22:2188–98. [PubMed: 22767387]

59. Hadjur S, et al. Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. Nature. 2009; 460:410–3. [PubMed: 19458616]

60. Bowers SR, et al. A conserved insulator that recruits CTCF and cohesin exists between the closely related but divergently regulated interleukin-3 and granulocyte-macrophage colony-stimulating factor genes. Mol Cell Biol. 2009; 29:1682–93. [PubMed: 19158269]

61. Merkenschlager M, Odom DT. CTCF and cohesin: linking gene regulatory elements with their targets. Cell. 2013; 152:1285–97. [PubMed: 23498937]

62. Schmidt D, et al. A CTCF-independent role for cohesin in tissue-specific transcription. Genome Res. 2010; 20:578–88. [PubMed: 20219941]

63. Faure AJ, et al. Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. Genome Res. 2012; 22:2163–75. [PubMed: 22780989]

64. Kagey MH, et al. Mediator and cohesin connect gene expression and chromatin architecture. Nature. 2010; 467:430–5. [PubMed: 20720539]

65. Kim TH, et al. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell. 2007; 128:1231–45. [PubMed: 17382889]

66. Shen Y, et al. A map of the cis-regulatory sequences in the mouse genome. Nature. 2012; 488:116–20. [PubMed: 22763441]

67. Wang H, et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res. 2012; 22:1680–8. [PubMed: 22955980]

68. Martin D, et al. Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes. Nat Struct Mol Biol. 2011; 18:708–14. [PubMed: 21602820]

69. Heintzman ND, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009; 459:108–12. [PubMed: 19295514]

70. Schwalie PC, et al. Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes. Genome Biol. 2013; 14:R148. [PubMed: 24380390]

71. Bourque G, et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res. 2008; 18:1752–62. [PubMed: 18682548]

72. Ward MC, et al. Latent regulatory potential of human-specific repetitive elements. Mol Cell. 2013; 49:262–72. [PubMed: 23246434]

73. Jacques PE, Jeyakani J, Bourque G. The majority of primate-specific regulatory sequences are derived from transposable elements. PLoS Genet. 2013; 9:e1003504. [PubMed: 23675311]

74. Zheng W, Zhao H, Mancera E, Steinmetz LM, Snyder M. Genetic analysis of variation in transcription factor binding in yeast. Nature. 2010; 464:1187–91. [PubMed: 20237471]

75. Reddy TE, et al. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. Genome Res. 2012; 22:860–9. [PubMed: 22300769]

76. Kasowski M, et al. Variation in transcription factor binding among humans. Science. 2010; 328:232–5. [PubMed: 20299548]

77. Wilson MD, et al. Species-specific transcription in mice carrying human chromosome 21. Science. 2008; 322:434–8. [PubMed: 18787134]

78. Zheng W, Gianoulis TA, Karczewski KJ, Zhao H, Snyder M. Regulatory variation within and between species. Annu Rev Genomics Hum Genet. 2011; 12:327–46. [PubMed: 21721942]

79. Degner JF, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature. 2012; 482:390–4. [PubMed: 22307276]

80. Shibata Y, et al. Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. PLoS Genet. 2012; 8:e1002789. [PubMed: 22761590]

81. Stone JR, Wray GA. Rapid evolution of cis-regulatory sequences via local point mutations. Mol Biol Evol. 2001; 18:1764–70. [PubMed: 11504856]

82. Heinz S, et al. Effect of natural genetic variation on enhancer selection and function. Nature. 2013

83. Stewart AJ, Plotkin JB. Why transcription factor binding sites are ten nucleotides long. Genetics. 2012; 192:973–85. [PubMed: 22887818]

84. Johnson R, et al. Evolution of the vertebrate gene regulatory network controlled by the transcriptional repressor REST. Mol Biol Evol. 2009; 26:1491–507. [PubMed: 19318521]

85. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 2008; 9:397–405. [PubMed: 18368054]

86. de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 2011; 7:e1002384. [PubMed: 22144907]

87. Wang T, et al. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. Proc Natl Acad Sci U S A. 2007; 104:18613–8. [PubMed: 18003932]

88. Bolotin E, et al. Nuclear receptor HNF4alpha binding sequences are widespread in Alu repeats. BMC Genomics. 2011; 12:560. [PubMed: 22085832]

89. Johnson R, et al. Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. Nucleic Acids Res. 2006; 34:3862–77. [PubMed: 16899447]

90. Lowe CB, Bejerano G, Haussler D. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc Natl Acad Sci U S A. 2007; 104:8005–10. [PubMed: 17463089]

91. Eddy SR. The C-value paradox, junk DNA and ENCODE. Curr Biol. 2012; 22:R898–9. [PubMed: 23137679]

92. Lynch M, Bobay LM, Catania F, Gout JF, Rho M. The repatterning of eukaryotic genomes by random genetic drift. Annu Rev Genomics Hum Genet. 2011; 12:347–66. [PubMed: 21756106]

93. Gonzalez J, Petrov DA. Evolution of genome content: population dynamics of transposable elements in flies and humans. Methods Mol Biol. 2012; 855:361–83. [PubMed: 22407716]

94. Bartolome C, Maside X, Charlesworth B. On the abundance and distribution of transposable elements in the genome of Drosophila melanogaster. Mol Biol Evol. 2002; 19:926–37. [PubMed: 12032249]

95. Eickbush TH, Furano AV. Fruit flies and humans respond differently to retrotransposons. Curr Opin Genet Dev. 2002; 12:669–74. [PubMed: 12433580]

96. Kimura M. Evolutionary rate at the molecular level. Nature. 1968; 217:624–6. [PubMed: 5637732]

97. Maia AT, et al. Effects of BRCA2 cis-regulation in normal breast and cancer risk amongst BRCA2 mutation carriers. Breast Cancer Res. 2012; 14:R63. [PubMed: 22513257]

98. Zhang X, Cowper-Sal lari R, Bailey SD, Moore JH, Lupien M. Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus. Genome Res. 2012; 22:1437–46. [PubMed: 22665440]

99. Schodel J, et al. Common genetic variants at the 11q13.3 renal cancer susceptibility locus influence binding of HIF to an enhancer of cyclin D1 expression. Nat Genet. 2012; 44:420–5. S1–2. [PubMed: 22406644]

100. Bernstein BE, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]

101. Garber M, et al. A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. Mol Cell. 2012; 47:810–22. [PubMed: 22940246]

102. Dickel DE, Visel A, Pennacchio LA. Functional anatomy of distant-acting mammalian enhancers. Philos Trans R Soc Lond B Biol Sci. 2013; 368:20120359. [PubMed: 23650633]

103. Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD. Genome editing with engineered zinc finger nucleases. Nat Rev Genet. 2010; 11:636–46. [PubMed: 20717154]

104. Joung JK, Sander JD. TALENs: a widely applicable technology for targeted genome editing. Nat Rev Mol Cell Biol. 2013; 14:49–55. [PubMed: 23169466]

105. Cong L, et al. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013; 339:819–23. [PubMed: 23287718]

106. Mali P, et al. RNA-guided human genome engineering via Cas9. Science. 2013; 339:823–6. [PubMed: 23287722]

107. Chevrier N, et al. Systematic discovery of TLR signaling components delineates viral-sensing circuits. Cell. 2011; 147:853–67. [PubMed: 22078882]

108. Sharon E, et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat Biotechnol. 2012; 30:521–30. [PubMed: 22609971]

109. Raveh-Sadka T, et al. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. Nat Genet. 2012; 44:743–50. [PubMed: 22634752]

110. Spitz F, Furlong EE. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012; 13:613–26. [PubMed: 22868264]

111. Ravasi T, et al. An atlas of combinatorial transcriptional regulation in mouse and man. Cell. 2010; 140:744–52. [PubMed: 20211142]

112. Voss TC, et al. Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism. Cell. 2011; 146:544–54. [PubMed: 21835447]

113. Zhu J, et al. Genome-wide Chromatin State Transitions Associated with Developmental and Environmental Cues. Cell. 2013; 152:642–54. [PubMed: 23333102]

114. Bernstein BE, et al. Genomic maps and comparative analysis of histone modifications in human and mouse. Cell. 2005; 120:169–81. [PubMed: 15680324]

115. Cain CE, Blekhman R, Marioni JC, Gilad Y. Gene expression differences among primates are associated with changes in a histone epigenetic modification. Genetics. 2011; 187:1225–34. [PubMed: 21321133]

116. Mikkelsen TS, et al. Comparative epigenomic analysis of murine and human adipogenesis. Cell. 2010; 143:156–69. [PubMed: 20887899]

117. Xiao S, et al. Comparative epigenomic annotation of regulatory DNA. Cell. 2012; 149:1381–92. [PubMed: 22682255]

118. Cotney J, et al. The evolution of lineage-specific regulatory activities in the human embryonic limb. Cell. 2013; 154:185–96. [PubMed: 23827682]

119. Charlesworth B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 2009; 10:195–205. [PubMed: 19204717]

120. Lynch M. The origins of eukaryotic gene structure. Mol Biol Evol. 2006; 23:450–68. [PubMed: 16280547]

121. Shapiro JA, et al. Adaptive genic evolution in the Drosophila genomes. Proc Natl Acad Sci U S A. 2007; 104:2271–6. [PubMed: 17284599]

122. Yu N, Jensen-Seaman MI, Chemnick L, Ryder O, Li WH. Nucleotide diversity in gorillas. Genetics. 2004; 166:1375–83. [PubMed: 15082556]

123. Lusk RW, Eisen MB. Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in Drosophila enhancers. PLoS Genet. 2010; 6:e1000829. [PubMed: 20107516]

124. Gayral P, et al. Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. PLoS Genet. 2013; 9:e1003457. [PubMed: 23593039]

125. Loh YH, et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat Genet. 2006; 38:431–40. [PubMed: 16518401]

126. Conboy CM, et al. Cell cycle genes are the evolutionarily conserved targets of the E2F4 transcription factor. PLoS One. 2007; 2:e1061. [PubMed: 17957245]

127. Woo YH, Li WH. Evolutionary conservation of histone modifications in mammals. Mol Biol Evol. 2012; 29:1757–67. [PubMed: 22319170]

128. Kutter C, et al. Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. Nat Genet. 2011; 43:948–55. [PubMed: 21873999]

Box 1

Current models of gene regulation by tissue-specific transcription factors

Two recent reviews have summarized our current models for eukaryotic gene regulation[44, 110]. Eukaryotic transcription factors recognize short, partially degenerate sequences (6-12 nucleotides long), and many appear to be expressed at high concentrations (typically 1,000-100,000 molecules per cell)[44]. They bind the genome over a continuum of occupancy levels that includes many lowly-occupied regions, often interpreted as background binding. TF binding in metazoans is also highly combinatorial[110]. For example, the human genome codes for an estimated 2000-3000 TFs, hundreds of which are expressed in a typical somatic tissue[5]. Combinatorial binding can be mediated by direct protein-protein interactions, often in a tissue-dependent manner[111], or via indirect cooperativity facilitated by co-binding of the same DNA[112]. This combinatorial complexity occurs over vast regulatory regions, ranging from hundreds to thousands of megabases of accessible genome sequence. These observations have led to an interconnected, continuous model of transcriptional networks, where biological significance of TF binding is proposed to correlate with combinatorial complexity[110] and occupancy levels[44]: indeed, strongly bound regions have been reported to be biologically relevant more often than those bound at low occupancy[45]. Mainly based on *Drosophila* studies, properties of low- and high-occupancy regions, such as evolutionary conservation or distance to functional target genes, are additionally correlated with regulatory function of TF binding locations. Interestingly, high occupancy TF binding in mammals may not be as tightly correlated with TF function or TF binding conservation[53] and this difference is an area of active investigation.

Box 2

Comparative ChIP-Seq studies of histone modifications

Genome-wide mapping of other aspects of chromatin structure (in addition to TF binding) can elucidate regulatory regions. Particular histone modifications preferentially mark promoter regions (e.g. H3K4me3), distal enhancers (e.g. H3K4me1 or H3K27Ac) and actively transcribed regions (H3K36me3). Comparative studies in human cells lines[69] and mouse tissues[66] have shown that these epigenetic modifications often show tissue-specific patterns, and can be used as a proxy to functionally annotate a species' genome without prior knowledge of what TFs are active in a particular tissue. Recent analysis of promoter (H3K4me3) and enhancer marks (H3K4me1 and H3K27Ac) in a panel of adult and embryonic mouse tissues found a large fraction of marked regions to be tissue-specific[66]. H3K4 monomethylated regions were the most tissue-specific, probably due to the high tissue-specificity of enhancers; whereas most regions occupied by H3K4me3 were so across many tissues. Similar conclusions regarding the tissue-specificity of histone modifications were reported in a recent study across human tissues[113]. Many tissue-specific regulatory regions are enhancers, leading to the question of how chromatin modifications evolve in different species (Table 1).

In primary human and mouse lung fibroblasts, typically 55-68% of syntentic regions in human and mouse are similarly enriched for H3K4 di- and trimethylation[110, 114]. High conservation of H3K4me3 locations was also found in lymphoblastoid cell lines from closely related primate species (human, chimpanzees and rhesus macaques), where 65% of orthologous regions were occupied in all three species[115]. At locations proximal to transcriptional start sites, 90% overlap was found between human and macaque, which is similar to the epigenetic conservation found in orthologous mouse and human proximal promoters occupied by H3K4me3[77]. These chromatin differences between species were (partly) predictive of changes in nearby gene expression[2, 115].

During the dynamic remodelling of histone modifications in human and mouse adipogenesis models, the majority of chromatin marks were species-specific and only 15-30% of orthologous genomic locations shared histone marks in human and mouse[116]. Consistent with other studies, though, the divergence was far higher among distal histone modifications, such as regions enriched for the enhancer mark H3K27Ac. A more detailed view of chromatin differences among mammals was recently obtained through extensive comparison of eight histone modifications and DNA methylation in human, mouse and pig pluripotent stem cells[117]. In contrast to previous observations[2, 114], and with the exception of the repressive mark H3K9me3, genomic regions occupied by histone modifications were correlated with conserved genomic sequences. However, reference 117 found no direct correlation between sequence similarity and epigenomic conservation, and most modifications showed conservation in both rapidly and slowly evolving sequences. Finally, a recent study where H3K27Ac profiles in human, rhesus and mouse embryonic limb were used to infer human gains of regulatory activity[118], most of the identified regions did not involve highly conserved elements, further suggesting rapid evolution of H3K27Ac locations. Moreover, comparison of the ChIP signal in orthologous locations across the three species indicated that most H3K27Ac human gains may arise through modification of pre-existing regulatory regions, marked at lower levels in rhesus and mouse[118].

BOX 3

Population genetics and metazoan transcription factor binding evolution

Natural selection operates at the level of a single organism's fitness, which manifests in the population as enhanced reproductive success. Thus selection influences the rate of evolutionary change in a species, the types of paths that are open to evolutionary exploration and, ultimately, expansions or contractions in genome size[92]. In particular, the effective size of a population (N_e) directly influences the rate of evolutionary change due to genetic drift (reviewed in[119]). Furthermore, the effective breeding population of a species is related to the rates of non-adaptive evolutionary processes, with smaller effective population sizes displaying elevated drift, higher mutation rates and lower rates of recombination. Because estimated effective population sizes vary widely across metazoan and eukaryotic phyla, species with smaller population sizes are thought to have reduced intensity of selection simultaneous with an increasing accumulation of mildly deleterious mutations via genetic drift[92]. Because there is a mutational bias towards insertions versus deletions, smaller effective population sizes of multicellular eukaryotes allow accumulation of putatively non-functional DNA and thus the observed expansion in genome size[120]. Therefore, effective population differences between vertebrates and invertebrates could underlie the observed differences in TF binding evolution between *Drosophila* and mammals, whose estimated N_e values differ by two orders of magnitude ($1{,}15*10^6$ [121] and 10^4 [122], respectively). According to population genetics theory, genomes in *Drosophila* species are under stronger selective pressures, which has led to genome compression[93]. Indeed, it is possible that very few (if any) nucleotides in the compact genomes of *Drosophila* species evolve completely free of selection[119]. Furthermore, multiple selective sweeps are likely to occur simultaneously in Drosophila populations, possibly interfering with each other.

In sum, large effective population sizes probably underlie key features of *Drosophila* TF binding evolution: stronger conservation of TF binding found in fruit flies[48, 49], signatures of selection across TF binding regions in multiple *Drosophila* species[23, 123] and lack of consistent evidence on the involvement of TEs in fruit fly TF binding evolution[55]. Conversely, mammals have much lower effective population sizes and much larger genomes, where genetic drift likely dominates over selection. This situation leads naturally to rapid evolution of TF binding in mammals and may mask signatures of natural selection.

A population genetics hypothesis is an attractive way to reconcile the differing evolution rates of TF binding observed in mammals and fruit flies, yet more data is needed--both to confirm these differences across a wider range of DNA binding proteins and to prove the dependence of such differences on effective population size. A recent population genomics study analyzed RNA-Seq data in a collection of (mainly) non-model vertebrate and invertebrate species and reported findings that were partially consistent with a vertebrate/invertebrate divide. Consistent with effective population size estimates, average genomic diversity was higher in invertebrates than in vertebrates, but the expected differences in the non-synonymous to synonymous polymorphisms ratio appeared to be absent[124]. Unfortunately, TF binding evolution has only been explored in a few genomes within the broader vertebrate and invertebrate phyla (Table 1), and comparative studies in more representative species would be invaluable to understand the forces that shape TF binding evolution.

Online summary

- Transcription factors orchestrate tissue-specific gene expression and thus tissue identity. Metazoan gene regulation is highly complex and comparative analyses of TF binding across species have revealed mechanisms underlying both genome evolution and gene regulation.
- Early studies focused on individual loci, and showed both conservation and divergence of putative TF binding sites across metazoan species.
- Direct global mapping of TF binding locations in multiple mammalian and fruit fly species discovered that tissue-specific TF binding evolves rapidly in mammals, whereas developmental TF binding in fruit flies appears under substantially greater constraint.
- Comparative studies in mammals and fruit flies have also highlighted common properties of metazoan TF binding evolution, such as dependence on genetic sequence changes, combinatorial co-evolution of binding, and partially compensatory turnover.
- Observed differences in TF binding evolution and densities of conserved non-coding elements among different metazoan families may be the result of different pressures from extreme differences in effective population sizes.
- In mammals, cross-species ChIP-Seq studies have further revealed how transposable element-derived sequences help generate novel lineage-specific TF binding.

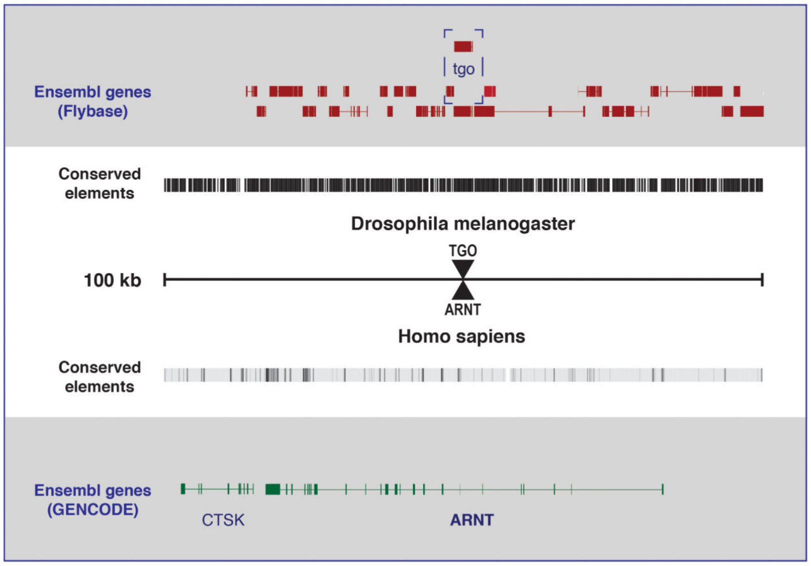

Figure 1. Insects and mammals have dramatically different densities of conserved elements

The central panel (white background) compares sequence constraint (Conserved elements tracks) in a 100 kb window around the tgo gene in *Drosophila melanogaster* (top) and the homologous ARNT gene in *Homo sapiens* (bottom); the difference in constraint density in this window is representative of whole-genome differences[36]. A higher fraction of the fruit fly genome is conserved (across fifteen insect genomes) compared to the human genome (across 33 placental mammals): at the whole-genome level, 37-53% of the D. melanogaster genome lies in conserved elements, compared to 3-8% of the human genome. Grey background panels show gene annotations in these regions (Ensembl tracks). For the gene-dense region in *Drosophila*, forward strand genes are on the top track and reverse strand genes on the bottom. Figure adapted from Ensembl Genome Browser, with conserved elements from phastCons/UCSC[36]. See main text for further discussion.

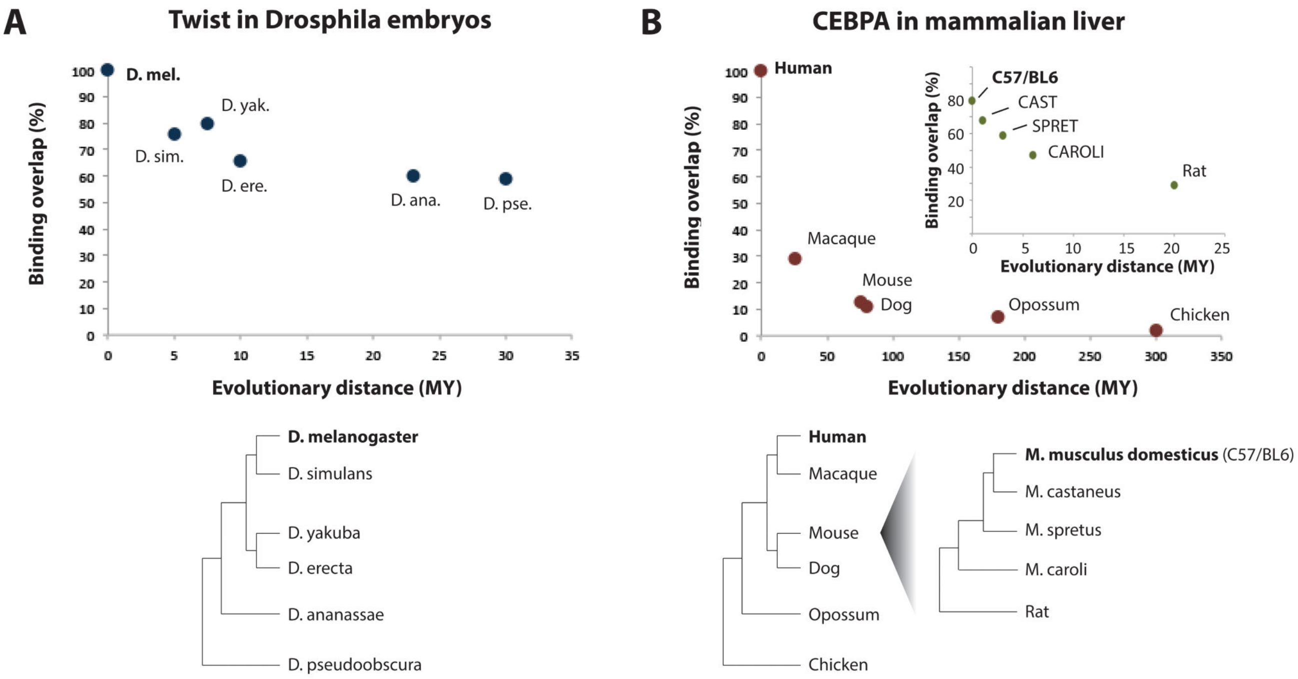

Figure 2. Genome-wide TF binding profiling in *Drosophila* and Mammals

Percentage binding overlaps are shown for the developmental TF Twist in whole embryos from divergent *Drosophila* species[49] (**A**) and for the tissue-specific TF CEBPA in livers from mammalian species (**B**). In **B**, the main graph shows overlaps for 6 vertebrate species[52], while the inset data[53] is for four mouse species and rat. In all cases, species are ordered by their evolutionary relationships as shown in the phylogenetic trees below each graph. Species in bold were used as the reference genome for comparison of the corresponding ChIP-Seq data. Where name abbreviations are used in the main graph, full species names are shown in the phylogenetic trees below.

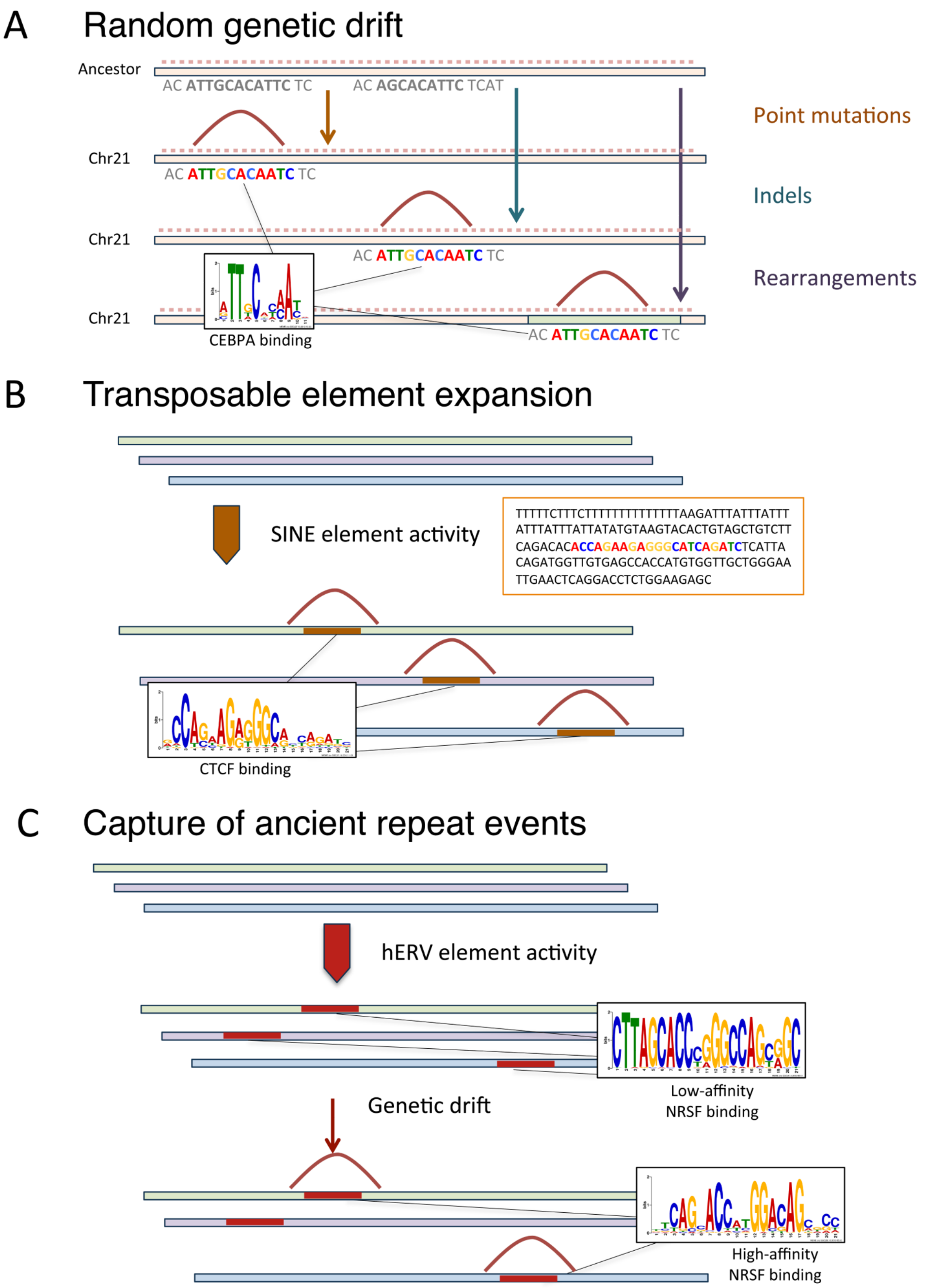

Figure 3. Sources of metazoan TF binding divergence

A. Random genetic drift. Point mutations, indels and genomic rearrangements can lead to binding events from non-bound sequences in the last common ancestor. This mechanism is most efficient for transcription factors with short binding sequences, such as CEBPA (binding motif logo shown on the left). From top to bottom, the examples in the diagram exemplify the birth of CEBPA binding events from the ancestor sequence by a point mutation, an insertion, or a genomic rearrangement with a different chromosome. **B. Repetitive element expansions**. Expansion of repetitive sequences carrying binding motifs by transposable elements can give rise to numerous binding events across mammalian genomes. This mechanism is especially relevant for transcriptional regulators such as CTCF, whose long binding sequence cannot easily arise by genetic drift. The diagram depicts birth of multiple CTCF binding sites through expansion of SINE transposable elements. The central inset contains a partial B2 element sequence harbouring a high-affinity CTCF binding event. **C. Capture of ancient repeat events**. In contrast to B, some repetitive elements contain low affinity binding motifs that differ in a few key mutations from high-affinity binding sequences. Once expanded throughout the genome by transposable elements, these binding sequences can easily mutate to high-affinity binding events by genetic drift. This mechanism is exemplified in the diagram for the transcriptional repressor NRSF. The hERV family of transposons contains low-affinity, non-binding motifs for NRSF[89] that can be exapted as high-affinity binding sites upon a few key mutations.

Table 1
Cross-species ChIP-Seq studies in metazoans

(Sub)phylum	Category	Species	Sample	Evolutionary distance (MY)	Antibody	Reported binding conservation
					Bicoid	99-98%
					Hunchback	94-86%
		D. melanogaster	Whole embryo	6-15	Krüppel	97%
		D. yakuba			Giant	99-97%
					Knirps	99.7-97%
					Caudal	98%
		D. melanogaster				
Insecta	*Transcription factors*	D. simulans				
		D. yakuba	Whole embryo	2.5-30	Twist	80-60%
		D. erecta				
		D. ananassae				
		D. pseudoobscura				
		D. melanogaster				
		D. simulans	Whole embryo	2.5-30	CTCF	85-30%
		D. yakuba				
		D. pseudoobscura				
		Human	ES cells	80	OCT4	9.1%
		Mouse			NANOG	13%
					HNF4A	48-29%
		Human	Primary hepatocytes	80	HNF1A	32-7%
		Mouse			HNF6	32-19%
					FOXA2	32-15%
		Human	Assorted tissues	80	E2F4	20%
		Mouse				
		Human		80-300	CEBPA	14-2% (0.3% utrashared in all five)
		Mouse				
		Dog	Liver tissue			
		Opossum				
		Chicken				
		Human				
		Mouse	Liver tissue	80	HNF4A	22-12%
		Dog				
	Transcription factors	Human	ES cells	80	OCT4	2%
		Mouse			NANOG	1.9%
					CTCF	16.7%
		Human	Assorted cell lines	80-300	CTCF	16.8-6.8%
		Mouse				
		Chicken				
		Human				
		Macaque				
Vertebrata		Mouse	Liver tissue	23-80	CTCF	60-38%
		Rat				
		Dog				

(Sub)phylum	Category	Species	Sample	Evolutionary distance (MY)	Antibody	Reported binding conservation
		M. musculus domesticus			CEBPA	74-29%
		M. castaneus	Liver tissue	0.5-20	FOXA1	77-28%
		M. spretus			HNF4A	74-28%
		M. caroli				
		Rat				
		Human	Primary lung fibroblasts	80	H3K4me3	68-55%
		Mouse				
		Human	Lymphoblastoid cell lines	6-23	H3K4me3	69.5-63.2%
		Chimpanzee				
		Macaque				
	Histone marks	Human	Assortment of cell lines		H3K4me3	75-50%
		Mouse		80	H3K4me1	40-30%
					H3K27Ac	70-50%
		Human				
		Macaque	Limb buds	23-80	H3K27Ac	79-40%
		Mouse				
		Human	Immortalized B cell line	6	pol II	68%
		Chimpanzee				
		Human				
	Polymerases	Macaque				
		Mouse	Liver tissue	23-80	pol III	52-23%
		Rat				
		Dog				