Chapter 6
Methods to Detect and Associate Divergence in Cis-Regulatory Elements to Phenotypic Divergence

Juliana G. Roscito and Michael Hiller

Abstract Understanding which genomic changes are responsible for morphological differences between species is a long-standing question in biology. While evolutionary theory predicts that morphology largely evolves by changing expression of important developmental genes, finding the underlying regulatory mutations is inherently difficult. Here, we discuss how the integration of comparative and functional genomics has provided valuable insights into the regulatory changes involved in morphological changes. By comparing genomes of species exhibiting differences in a morphological trait, comparative genomic methods enable the systematic detection of regulatory elements with divergence in sequence or transcription factor binding sites. To narrow this set of diverged elements down to those that likely contribute to differences in the trait of interest, one can leverage knowledge about gene function to assess which elements are associated with genes known to control the development of this trait. In addition, functional genomics can further prioritize diverged genomic regions based on overlap with experimentally determined regulatory elements that are active in tissues relevant for the trait. Further experiments can then evaluate whether sequence or binding site divergence translates into regulatory differences and affects the development of the trait. Thus, combining comparative and functional genomic approaches provide a widely applicable strategy to reveal regulatory changes contributing to morphological differences, which will enhance our understanding of how nature's spectacular phenotypic diversity evolved.

J. G. Roscito · M. Hiller (✉)
Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
e-mail: hiller@mpi-cbg.de

J. G. Roscito
e-mail: roscito@mpi-cbg.de

Max Planck Institute for the Physics of Complex Systems, Dresden, Germany

Center for Systems Biology, Dresden, Germany

P. Pontarotti (ed.), *Evolution, Origin of Life, Concepts and Methods*,
https://doi.org/10.1007/978-3-030-30363-1_6

6.1 Introduction

The diversity of life forms is one of the most fascinating aspects of biology. The different shapes of the vertebrate limb, the breathtaking diversity of marine invertebrates, and many modifications of body plans in fish are only a few examples of the extent to which the evolutionary process shaped morphology. The question of how this diversity arose is a long-standing question in biology. With recent technological advances in genetics and genomics, we are now better equipped to investigate which changes in the genomes are responsible for morphological differences between species.

Morphology is established during development and involves genes controlling the growth and patterning of different parts of the body. Most genes involved in development are highly pleiotropic, that is, they regulate multiple developmental processes. These genes often code for transcription factors or other signaling proteins that play central roles in various gene regulatory networks that govern the formation of different body structures. Because these high-level regulators are involved in making several structures, they have to be expressed at many tissues and timepoints in the developing embryo.

The specificity in expression is given by cis-regulatory elements (CREs), which directly control the transcriptional output of the gene in a spatial and temporal manner (Davidson et al. 2003; Howard and Davidson 2004). CREs can be promoters, enhancers, silencers, and insulators. While promoters are located in close proximity to the gene's transcription start site, enhancers, silencers, and insulators can be located far from the gene they regulate and can be located upstream or downstream of the transcription start site. CREs are characterized by the presence of shorter sequences to which transacting DNA binding proteins, or transcription factors (TFs), bind to. The binding of specific combinations of transcription factors results in the activation or repression of transcription. Distal CREs such as enhancers and silencers are often modular and independently regulate gene expression in specific tissues or developmental timepoints (Davidson et al. 2003; Howard and Davidson 2004).

Given that morphology is established during development, evolutionary changes in morphology typically require changes affecting developmental genes. Conceptually, such changes can be due to mutations in the protein-coding sequence, leading to changes in protein function, or mutations in the associated regulatory elements that alter the expression pattern of these genes. Coding mutations in pleiotropic developmental genes would likely have widespread effects on development, which is often deleterious for the organism. In contrast, regulatory mutations in modular CREs can have smaller-scale effects, i.e., expression of its target gene would only be altered in a specific tissue or timepoint. The relative contribution of mutations in coding regions and CREs to the evolution of morphology has been the focus of a long-standing debate (Carroll 2008; Wittkopp et al. 2004; Wray 2007), and there are clear examples showing that both coding and regulatory mutations contribute to morphological differences (Stern and Orgogozo 2008; Burga et al. 2017; Sharma et al. 2018). However, the general difference in the degree of pleiotropy of developmental

genes and the modular CREs that control their expression predicts that morphology largely evolves by changes in the CREs, affecting gene expression patterns.

While there is strong support that many morphological changes are associated with spatial or temporal shifts in gene expression patterns, finding the underlying genomic changes in CREs is challenging. Many examples where morphological differences have been mechanistically linked to changes in CREs come from comparative studies between *Drosophila* species. Among those are a variety of abdominal and wing pigmentation patterns that are due to mutations in CREs regulating the *yellow* gene (Gompel et al. 2005; Prud'homme et al. 2006; Wittkopp et al. 2002), the loss of dorsal cuticular hairs in larvae of *Drosophila sechellia* that is a result of changes in CREs of the *shavenbaby* gene (Frankel et al. 2011; McGregor et al. 2007), and the evolution of additional sex combs in the legs and loss of part of the sensory bristles in the genitalia of *Drosophila santomea* that is due to a single nucleotide substitution in an enhancer of the *scute* gene (Nagy et al. 2018). Changes in CREs are also associated with vertebrate morphological evolution. One example is the diversification of artiodactyl digit patterns, which arose by changes in a CRE regulating the *sonic hedgehog* receptor gene *Ptch1* (Lopez-Rios et al. 2014). A second example is the loss of limbs in snakes, which is associated with deletions in a crucial enhancer that regulates the *sonic hedgehog* morphogen in the developing limbs. Snake-specific deletions of TF binding sites (TFBS) in this enhancer result in reduced enhancer activity, which provides an explanation for the reduced expression of this key limb patterning factor in snakes (Kvon et al. 2016; Leal and Cohn 2016). Complete CRE deletions can also change morphology. Loss of pelvic spines in freshwater stickleback fish species is due to recurrent deletions of a pelvic enhancer regulating *Pitx1*, which abolishes expression of this developmental transcription factor in the pelvic region and leads to the loss of pelvic spines (Chan et al. 2010). Finally, as detailed below, the evolution of several human-specific traits is linked to the loss of enhancers (McLean et al. 2011).

While these examples highlight the importance of CRE changes to the evolution of morphology, there are many other morphological changes for which the underlying genetic cause has not been identified yet, despite observations that expression differences in developmental genes are involved. Prominent examples include shifts in the expression domains of *Hox* genes, which pattern the anterior-posterior (AP) axis. In crustaceans, AP shifts in the expression domain of *ultrabithorax* result in the modification of locomotor into feeding appendages (Averof and Patel 1997). In snakes, an anterior expansion of the expression domain of *HoxB* and *HoxC* genes is associated with the homeotic transformation of the cervical region toward a thoracic identity, resulting in the absence of forelimbs (Cohn and Tickle 1999). There are several other examples of divergent limb morphologies within vertebrates that are associated with expression changes of key developmental factors. For example, spatial and temporal changes in expression of many limb patterning genes in the developing hindlimb buds of cetaceans are associated with an arrested development and regression of hindlimbs (Thewissen et al. 2006), *Shh* expression changes are associated with digit reductions in lizards (Shapiro et al. 2003), and changes in *Fgf8*, *Shh,* and *Ptch1* expression in bat wings are associated with digit lengthening and

the extension of interdigital webbing (Hockman et al. 2008). Finally, the variety of beak sizes and shapes in Darwin finches is linked to different expression patterns of the *Bmp4* gene (Abzhanov et al. 2004). While these morphological differences are associated with spatial and/or temporal expression shifts of relevant developmental genes, the genomic changes that underlie the expression changes of these genes remain largely unknown.

We are now witnessing a genomics revolution that can help to identify the genomic changes responsible for morphological changes. Technological and methodological advances are facilitating the acquisition of high-throughput data for an increasing number of species, allowing for large-scale comparative genomic analysis. Furthermore, experimental approaches using high-throughput sequencing enable an in-depth profiling of cell/tissue-specific regulatory landscapes. In this chapter, we discuss how integrating comparative and functional genomics can be a powerful approach to identify the genomic basis of morphological evolution. In the following, we first present a brief overview of computational and functional approaches for genome-wide identification of CREs. Then, we describe how comparative methods can use this data to quantify CRE divergence and to associate divergence patterns to morphological differences.

6.2 Methods to Identify CREs

Approaches to identify CRE candidates genome-wide can be mainly divided into computational and experimental methods. It is important to note that each approach has their strengths and intrinsic biases; therefore, combining different methods provide a more complete picture of the regulatory landscape. In-depth information on each of the methods can be found in a number of reviews (Wasserman and Sandelin 2004; Hardison and Taylor 2012; Kleftogiannis et al. 2016; Bell et al. 2011; Pennacchio and Rubin 2001; Li et al. 2015; Noonan and McCallion 2010; Shlyueva et al. 2014; Yip et al. 2013; Aerts 2012; Schmitt et al. 2016).

6.2.1 Computational Identification of CREs

1. Evolutionary conservation

Predicting CRE candidates based on evolutionary conservation relies on the fact that functional genomic sequences often evolve under purifying selection, that is, deleterious mutations affecting the function of such sequences negatively affect fitness and are less likely to be fixed in the population. As a result, functional genomic regions accumulate fewer mutations compared to other non-functional regions of the genome. This principle is exploited in *phylogenetic footprinting* approaches (Tagle et al. 1988) that compare genomes of many species to identify functional genomic

regions (Lindblad-Toh et al. 2011). While conserved genomic regions overlapping coding exons are likely constrained to preserve the encoded protein sequence, conserved sequences in non-coding parts of the genome, referred here as conserved non-coding elements (CNEs), can reliably reveal CREs (Hardison 2000; Pennacchio and Rubin 2001; Woolfe et al. 2005), in particular, when non-coding regions are conserved over large evolutionary timescales (e.g., from human to fish). Indeed, many experimental studies showed that a substantial portion of CNEs have regulatory activity and direct tissue-specific expression patterns (Frazer et al. 2004; Grice et al. 2005; Pennacchio et al. 2006; Visel et al. 2008). While this approach will capture neither recently evolved nor species-specific CREs, sequence conservation is a powerful indicator of CREs.

2. In silico identification of transcription factor binding site clusters

CREs are activated by the binding of TFs to sequence motifs that are generally short (on average 8 bp long) and often degenerate. The observation that CREs are typically bound by multiple TFs made it possible to develop methods that scan the genome for clusters of putative TFBS (Hughes et al. 2000; Kim et al. 2010; Rajewsky et al. 2002; Sinha et al. 2003). Motif-based approaches for predicting CREs are influenced by the set of TFs and a sizeable portion of the TFBS-based predicted CREs are false positives.

6.2.2 Experimental Identification of CREs

Genomic regions comprising active CREs exhibit characteristic physical and chemical signatures that can be detected experimentally. Such functional genomic methods provide a reliable way to identify CREs that are active in the selected tissue or cell type.

1. Open chromatin

Activation of CREs depends on the binding of transcription factors to the CRE sequence. TF binding competes with histone binding, which results in an increased accessibility of the local DNA region. Thus, active CREs can be identified by nucleosome-depleted open chromatin regions. Open chromatin can be targeted with nucleases such as DNaseI (Cockerill 2011; Gross and Garrard 1988; Wu et al. 1979). Another widely used method to profile open chromatin is the ATAC method (assay for transposase-accessible chromatin) (Buenrostro et al. 2013) that uses the Tn5 transposase to digest DNA in areas of reduced nucleosome occupancy. Nucleosome-depleted regions can also be identified as by-product of the detection of nucleosome-bound regions using the micrococcal nuclease MNase (Yuan et al. 2005). Fragmentation-based detection of open chromatin areas is the principle underlying FAIRE (formaldehyde-assisted isolation of regulatory elements) (Simon et al. 2013), which involves the crosslinking and mechanical shearing of DNA and subsequent depletion of those fragments with bound nucleosomes.

2. Chromatin immunoprecipitation (ChIP)

ChIP is a technique that relies on crosslinking chromatin in its native state and the subsequent antibody-based selection of proteins (TFs) bound to DNA, or of specific histone modifications that are hallmarks of active CREs. Given an antibody for the protein of interest, this method enables the direct detection of binding sites of TFs, transcriptional co-activators such as MED or p300, and proteins associated with transcription initiation such as Pol II. It also allows the detection of active and repressed promoters and distal CREs based on the characteristic post-translational histone modifications such as acetylation and methylation of the N-terminal histone tail (Heintzman et al. 2007).

3. Chromosome conformation capture

Chromosome conformation capture assays target the physical interaction between CREs and gene promoters. The detection of interacting regions in a chromosome is a powerful method to map active distal CREs, which are thought to control transcription of their target gene by physically interacting with target promoter regions (de Laat and Duboule 2013; Montavon and Duboule 2012; Noordermeer and Duboule 2013). There are currently a wide variety of chromosome capture techniques (Schmitt et al. 2016), which differ mainly with respect to resolution and the scale of detected interactions. The most commonly used technique for genome-wide profiling of chromosomal interactions is Hi-C.

4. Enhancer RNA (eRNA)

Due to advancements in high-throughput sequencing, it became clear that the pool of RNA in a cell also comprises RNA molecules derived from enhancers (eRNAs) (Andersson et al. 2014; Azofeifa et al. 2018; Lam et al. 2014). While it is not clear whether these short and unstable eRNAs have a function themselves, they are a valuable marker of active CREs. Indeed, eRNA levels are a quantitative readout of CRE activity and correlate with the mRNA levels of the gene they regulate. Thus, selecting nascent RNAs followed by sequencing is another powerful method to determine active CREs in the tissue of interest.

5. Experimental profiling of regulatory activity

The ability to drive expression of a reporter gene is another way to experimentally test whether a CRE candidate has regulatory activity. While such reporter gene assays were previously limited to few elements, high-throughput screening techniques enable now testing the regulatory potential of a large set of DNA fragments (Inoue and Ahituv 2015). High-throughput analysis of regulatory activity involves massive parallel cloning of tagged DNA fragments into reporter vectors and quantifying enhancer activity by deep sequencing. In the Starr-seq method (Arnold et al. 2013), DNA fragments are inserted into a vector downstream of a minimal promoter and transfected into cells; if the DNA fragment acts as an enhancer it will transcribe itself. Sequencing the pool of cells is then used to quantify the transcriptional products, resulting in a parallel readout of enhancer activity for millions of fragments.

6.2.3 *Integrating Computational and Functional Genomics Data for Comprehensive CRE Annotation*

Despite the significant advancements of all above-mentioned methods, allowing for accurate identification of enhancers, each method produces false-positive candidates, and no single method is sufficient to detect all active CREs. By definition, conservation-based or TFBS-based methods will only identify deeply conserved CREs or CREs that exhibit clusters of binding sites for the specified TFs. Likewise, functional genomics methods will only identify CREs active in the tissue or cell type and timepoint where the assay was performed. For these reasons, integrating evidences from multiple approaches can significantly reduce biases and false-positives coming from individual methods alone. A straightforward approach is to combine both computational and experimental predictions by selecting computationally predicted CREs that overlap one or more sets of experimentally discovered CREs. In case many experimental datasets are available, machine learning approaches can be used to model, annotate, and classify CREs genome-wide, which has been shown to increase sensitivity and accuracy in annotating CREs (Erwin et al. 2014; Kleftogiannis et al. 2015; Li et al. 2015; Monti et al. 2017; Narlikar et al. 2010; van Duijvenboden et al. 2016).

6.3 Detecting and Quantifying CRE Divergence

Identifying CREs is the first step in identifying the regulatory changes that are associated with morphological changes. However, narrowing the list down to those CRE candidates actually involved in phenotypic changes is challenging, since each of the above-discussed approaches typically yield hundreds, or even thousands, of CRE candidates. The search for the relevant CREs, i.e., those associated with the morphological change, requires a comparative approach that involves either finding newly evolved CREs specific to species that show the trait of interest, or finding ancestral CREs that changed or lost function in these species, which can be detected as sequence or TF binding site divergence.

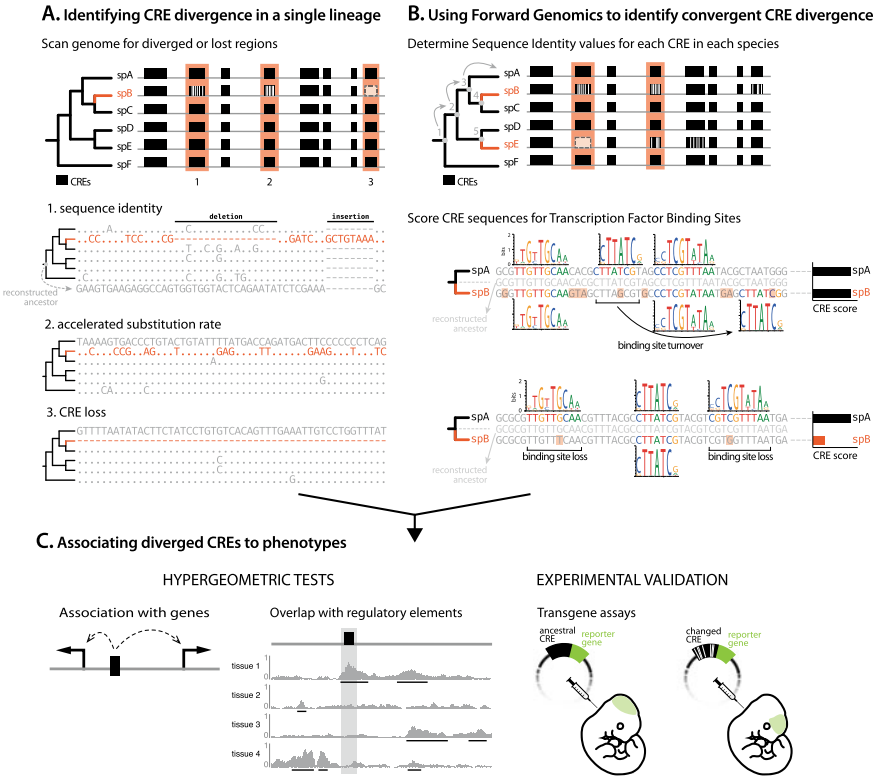
Changes in gene regulation can be due to the evolution of novel CREs in a particular species. Detection of such lineage-specific CREs with computational methods is challenging. Thus, obtaining experimental data for tissues relevant for the trait of interest is often necessary. Newly evolved CREs that may contribute to a newly evolved or changed trait can then be identified by comparing the regulatory landscape of species exhibiting the ancestral phenotype with that of species exhibiting the derived phenotype. However, evolution often also tinkers with existing functional elements. Thus, in addition to newly evolved CREs, changes or losses in conserved CREs can contribute to changes in gene expression and consequently changes in morphological structures. Hence, detecting and quantifying changes in ancestral CREs can highlight regulatory changes that are associated with morphological changes.

In the following, we will describe comparative genomic approaches to detect and quantify sequence and TFBS changes in ancestral CRE candidates whose sequences align across species, and methods to associate divergence patterns to differences in morphological traits. Figure 6.1 illustrates different types of sequence divergence described below, statistical tests to find characteristic enrichments of the identified CREs, and experimental tests to assess whether sequence divergence translates into regulatory changes.

6.3.1 Loss of CREs

The most radical change affecting an ancestral CRE is its complete loss, which can be computationally detected by lineage-specific absences of CNEs (Fig. 6.1a). Following the principle that evolutionary conservation should reflect function, deletion or absence of any aligning sequence in a particular lineage for an otherwise highly conserved CNE can occur when there is a relaxation of the selective pressures to maintain its function. This can happen as a result of different scenarios. First, a CRE can be lost when a new CRE arises at a different locus and functionally replaces the ancestral one, releasing the latter from selective constraint. Second, CREs can be lost following the loss of their target gene. Third, CRE loss can be associated with loss of a particular body structure, reflecting the loss of expression of the CRE's target gene that was once needed for the development of that structure. It should be noted that the absence of conserved sequence in a multiple genome alignment is not always due to CNE loss, as assembly incompleteness of alignment issues can mimic CNE loss. CNE losses should therefore be carefully validated (Hiller et al. 2012a).

CNE loss is not a rare event. A genome-wide screen showed that about 5% of mammalian CNEs are lost in at least one mammal species (Hiller et al. 2012a). More than 600 of those CNE losses are shared by more than one lineage and experiments showed that one of these independently lost CNEs functions as a spinal cord enhancer of the developmental *Gdf11* gene (Hiller et al. 2012a). CNE losses have been linked to lineage-specific morphological changes. For example, a genomic analysis showed that the seahorse has lost a substantially higher number of CNEs compared to other percomorph fish, and several of these CNE losses likely played a role in the evolution of the spectacular seahorse morphology (Lin et al. 2016). McLean et al. performed a screen to detect CNEs that are completely deleted in humans but present in chimpanzees, other primates and other mammals (McLean et al. 2011). Experimental tests of the selected CNEs with transgenic reporter assays in mouse (Fig. 6.1c) allowed them to associate the loss of an enhancer of the androgen receptor gene with the loss of sensory vibrissae and penile spines in humans. Furthermore, they found that humans have lost an enhancer regulating the tumor suppressor gene *Gadd45g*, which could be associated with the expansion of the neocortex in humans. Thus, complete loss of CNEs may affect gene expression and, consequently, morphological traits; however, experimental characterization of the consequences of the loss of a particular CNE are fundamental to reveal whether and which morphological traits are affected.


A. Identifying CRE divergence in a single lineage
Scan genome for diverged or lost regions
spA
spB
spC
spD
spE
spF
CREs
1
2
3
1. sequence identity
deletion
insertion
reconstructed ancestor
2. accelerated substitution rate
3. CRE loss
B. Using Forward Genomics to identify convergent CRE divergence
Determine Sequence Identity values for each CRE in each species
Score CRE sequences for Transcription Factor Binding Sites
reconstructed ancestor
binding site turnover
binding site loss
binding site loss
CRE score
CRE score
C. Associating diverged CREs to phenotypes
HYPERGEOMETRIC TESTS
Association with genes
Overlap with regulatory elements
tissue 1
tissue 2
tissue 3
tissue 4
EXPERIMENTAL VALIDATION
Transgene assays
ancestral CRE
reporter gene
changed CRE
reporter gene

◀**Fig. 6.1** Comparative genomic approaches to identify CRE divergence and methods to associate diverged CREs to functions and phenotypes. **a** To identify lineage-specific CRE divergence, illustrated here for species B (spB, orange), one can screen for CREs that are significantly diverged in this lineage but well conserved in the other species. The most general approach to identify sequence divergence takes the different types of mutations (substitutions, insertions, and deletions) into account and measures sequence divergence between extant species and their reconstructed ancestor (1). Alternatively, one can restrict the screen to identify CREs with accelerated substitution rates (2) or CREs that are completely lost or deleted (3). DNA bases that are identical to the reference are shown as dots in the alignments. **b** The Forward Genomics framework can identify associations between a convergently evolved phenotype and CRE divergence in the respective independent lineages, illustrated here for species B and E (orange). The two orange bars highlight CREs for which a genome-wide screen found a match between the CRE divergence pattern and the phenotypic pattern. These CREs exhibit a significantly higher divergence in species with the derived or lost phenotype (B, E) compared to the other species with the ancestral phenotype (A, C, D, F). The general Forward Genomics framework can consider either sequence or TFBS divergence. Both divergence measurements rely on reconstructing ancestral sequence states for each node in the phylogeny and using both extant and ancestral sequences to compute a branch-specific divergence measure by comparing the start and end node of each branch (illustrated by the gray arrows on the phylogeny). While sequence divergence reflects the amount of nucleotide changes (similar to A1), TFBS divergence reflects changes in binding sites independently of the number of mutations. The top alignment example illustrates similar TFBS scores of species A and B despite a high number of substitutions in species B that still retains all three binding sites because of motif degeneracy (substitutions that do not affect TF binding) and binding site turnover. In contrast, the bottom alignment illustrates that very few mutations can destroy binding sites, leading to a decreased TFBS score in species B. **c** Enrichment tests can be used to test whether diverged CREs are statistically associated with specific functions. Using non-diverged CREs as a control, hypergeometric tests can assess whether diverged CREs are significantly associated with genes belonging to certain functional groups (based on gene ontology terms, pathways, or knockout phenotypes) or whether diverged CRE candidates significantly overlap regulatory elements active in specific tissues. Finally, transgene assays can test whether CRE candidates have regulatory activity and whether CRE divergence translates into expression differences

6.3.2 CREs with Accelerated Substitution Rates

While complete loss of a CRE is the most radical type of change, changes in CRE function can also be detected by divergence in the CRE sequence. For example, a screen for CNEs that exhibit significantly elevated substitution rates in a particular lineage (Fig. 6.1a) can reveal elements evolving under positive selection that are associated with traits specific to this lineage.

This approach has been widely used (Cotney et al. 2013; Lindblad-Toh et al. 2011) and, together with follow-up experimental characterizations (Fig. 6.1c), revealed insights into regulatory changes involved in many traits. For example, a screen for CNEs with accelerated substitution rates in the placental mammal ancestor identified almost 5000 loci (Holloway et al. 2016). Subsequent transgenic reporter assays revealed that some of these accelerated CREs are associated with general mammalian morphological and physiological characteristics, such as hormonal regulation of milk production and uterine contractions, and development of the central nervous system

(in particular the visual system) (Holloway et al. 2016). Screens for CNEs with accelerated substitution rates in the human lineage revealed CREs regulating expression of genes involved in brain and neuronal development, including the RNA gene *Har1f* (Capra et al. 2013; Pollard et al. 2006a, b; Prabhakar et al. 2006). Experiments also revealed that a human-specific accelerated CRE drives stronger expression in the developing limbs than the chimpanzee or rhesus macaque sequence, which likely contributes to human-specific aspects of limb and digit morphology such as shortening of the toes and increase in thumb length (Prabhakar et al. 2008). Screens for accelerated CNEs also contributed to understanding traits in birds and bats. For example, vocal learning in independent bird lineages is linked to convergent acceleration of CNEs that likely regulate genes expressed in the song nuclei of the brain, such as the speech gene *FoxP1* (Zhang et al. 2014). The evolution of the bat wing is associated with bat-specific accelerated CREs regulating *HoxD* and other genes (Booker et al. 2016). Reporter assays showed that the sequence of these CREs often drives a novel expression pattern in the developing limbs compared to the orthologous mouse sequence (Booker et al. 2016).

In addition to detecting a gain or change of CRE function by identifying those CREs evolving under positive selection in specific lineages, accelerated substitution rates can also be used to detect otherwise conserved CREs that evolve under relaxed selection or neutrally in particular lineages. For example, several CNEs near *Pax6* and other transcription factors important for eye development exhibited accelerated substitution rates in subterranean mammals that have highly degenerated eyes (Partha et al. 2017). This acceleration was largely the result of neutral evolution, consistent with an association between the reduced visual system and relaxation of constraint on eye-regulatory elements (Partha et al. 2017).

6.3.3 Sequence Identity

By definition, accelerated substitution rates consider only substitutions and ignore insertions and deletions that can also affect CRE function. An exclusive focus on substitutions is a severe limitation when screening for neutrally evolving CREs as the clearest indication of loss of function is the complete deletion of a CRE. To take substitutions, insertions, and deletions into account when measuring sequence divergence, we previously developed an approach that is based on reconstructing the sequence of common ancestor of the species of interest (Hiller et al. 2012b; Prudent et al. 2016). This ancestral sequence is then aligned to sequences of extant species to determine the percent of identical bases (Fig. 6.1a). If the ancestral sequence is unchanged in an extant species, percent identify will be 100%. Complete divergence or a deletion of the ancestral sequence will result in percent identity of 0%. A framework termed "Forward Genomics" uses this sequence identity measure to perform a genome-wide screen to identify those conserved regions that exhibit elevated sequence divergence in those lineages where a particular trait was independently changed or lost (Fig. 6.1b) (Hiller et al. 2012b). For example, screening for genomic

regions that are preferentially diverged in mammals that are unable to synthesize vitamin C successfully identified genomic loci overlapping the vitamin C synthesizing *Gulo* gene (Hiller et al. 2012b).

To increase the power of the Forward Genomics framework, we developed the "branch method" to control for phylogenetic relatedness between species and differences in their evolutionary rates (Prudent et al. 2016). This method reconstructs the sequence of all ancestral states, represented as internal nodes in the phylogenetic tree (Fig. 6.1b), and computes branch-specific sequence divergence values by comparing the sequence at the start and end of each branch. These values are normalized by the length of the branch to control for differences in evolutionary rates. Since each branch represents independent evolution, branch-specific sequence divergence values circumvent the problem that the sequences of extant species are phylogenetically related (Prudent et al. 2016). As a consequence, this measure can accurately highlight specific branches of the tree along which the given sequence significantly accumulated mutations. Finally, to associate sequence divergence with a given trait, the branch method tests, separately for each CRE, whether the sequence identity values of branches along which the trait was changed or lost are lower than the values of branches along which the ancestral trait was preserved.

We applied this approach to perform the first genome-wide screen for CREs associated with the convergent eye degeneration in subterranean mammals (Roscito et al. 2018). This screen revealed more than 9000 CNEs with significantly increased sequence divergence on the branches leading to subterranean mammals. Using computational enrichment tests (Fig. 6.1c), we found that these CNEs are preferentially located near genes involved in eye development and eye function. To test whether these diverged CNEs correspond to eye-regulatory elements, we performed ATAC-seq in the developing mouse eye tissues and found that diverged CNEs significantly overlap eye CREs (Roscito et al. 2018) (Fig. 6.1c). Corroborating these results, diverged CNEs also significantly overlap functional genomics datasets obtained for adult mouse eye tissues. Together, our genome-wide analysis integrating functional and comparative genomics demonstrated that eye degeneration in subterranean mammals is associated with a widespread sequence divergence in the eye-regulatory landscape.

We used a similar ancestral reconstruction-based approach to investigate regulatory changes associated with the loss of limbs in snakes. Our genome-wide screen detected more than 5000 CNEs that exhibit significantly higher sequence divergence in snakes in comparison to other limbed species. Among those snake-diverged CNEs is the *Shh* limb enhancer that was experimentally shown to have significantly decreased its limb regulatory activity in snakes (Kvon et al. 2016; Leal and Cohn 2016). By combining ATAC-seq profiling and computational enrichment tests, we could show that snake-diverged CNEs are preferentially located near genes involved in limb development and significantly overlap CREs active in the developing limbs. This provides evidence that loss of limbs in the snake lineage is associated with genome-wide divergence of the limb regulatory landscape. Together with our results

on eye degeneration, these genome-wide screens suggest that the widespread divergence of the trait-specific regulatory landscape is a general evolutionary principle following the loss of complex morphological traits.

Another study devised a computational "Reverse Genomics" approach to associate sequence divergence to changes in morphological traits (Marcovitz et al. 2016). Instead of using ancestral sequence reconstruction, sequence divergence was identified based on a pairwise comparison between each mammalian species and human, which served as the reference. The study identified CNEs that were independently lost during placental mammalian evolution and used a large phenotypic character matrix to associate CRE loss patterns to trait change patterns. This screen uncovered numerous cases where independently lost CNEs are located near genes that are known to affect the development of the changed morphological structure. Striking examples include diverged CNEs associated with differences in pelvic or forelimb skeleton or differences in cochlea and brain morphology (Marcovitz et al. 2016).

Taken together, associating CNE sequence divergence in independent lineages with convergently evolved traits represents a powerful approach to reveal regulatory changes involved in morphological changes.

6.3.4 Scoring Gains and Losses of Transcription Factor Binding Sites

CREs are comprised of short-sequence motifs bound by TFs that determine regulatory activity. Gains, losses, and changes in TFBS, therefore, have strong potential to modify the regulatory activity of a CRE. Hence, investigating sequence instead of TFBS changes in a CRE may not always be informative about predicting functional CRE changes. For example, very few nucleotide changes (irrespective of whether they are substitutions or insertions/deletions) may affect key TFBS and thus change function, whereas numerous nucleotide changes may leave TFBS unaffected or result in TFBS turnover (creation of new equivalent TFBS and destruction of the original site, Fig. 6.1b) (Dermitzakis and Clark 2002; Huang et al. 2007; Otto et al. 2009; Villar et al. 2014). An analogy to that are nucleotide changes in protein-coding regions where few non-synonymous (amino-acid changing) differences can have a large impact on protein function, whereas numerous synonymous differences result in the same protein. Consequently, methods that take TFBS divergence into account can provide a more reliable way of identifying CRE changes that are more likely to affect regulatory activity and result in morphological changes across species.

A key requirement for identifying TFBS changes across species is predicting TFBS in a given CRE sequence. The binding motif of a TF can often be described by a position weight matrix (PWM), which is a representation of the per-base affinity of a TF to the respective binding site. Given that TFs of related species often have highly similar binding motifs (Boyle et al. 2014; Nitta et al. 2015), the same PWM can be used to predict TFBS in orthologous CRE sequences of multiple species. Applied

genome-wide, this information can highlight CREs with altered TFBS and can reveal statistical associations between TFBS differences and trait changes (Fig. 6.1c). For example, significant TFBS differences were observed in the promoter of genes associated with social behavior in bees and these differences correlated with differential expression of these genes in different bee castes (Kapheim et al. 2015; Sinha et al. 2006). Widespread gains and losses of limb-related TFBS in bat-accelerated CREs corroborate an association between changes in CRE function and the evolution of the bat wing (Booker et al. 2016). Another study demonstrated that genome-wide detection of TFBS losses combined with enrichment tests can reveal associations between TFBS losses and morphological differences (Berger et al. 2018). For example, TFBS preferentially lost in cetaceans and sirenians that lack hindlimbs are significantly associated with genes involved in hindlimb development, and TFBS preferentially lost in subterranean mammals are significantly associated with genes involved in eye development (Berger et al. 2018).

To pinpoint individual CREs where changes in clusters of TFBS are associated with convergent trait changes in independent lineages, we developed the REforge (regulatory element forward genomics) method (Langer et al. 2018). REforge is built on the Forward Genomics framework explained above, but measures TFBS divergence instead of sequence divergence. Given a set of TFs that may be relevant for the trait of interest (e.g., eye-related TFs could be relevant for eye degeneration), REforge performs a genome-wide screen to find CREs for which divergence of binding sites of these TFs is statistically associated with a given convergent trait change. In detail, REforge first reconstructs all ancestral sequence states of each CRE and uses Stubb (Sinha et al. 2003, 2006) to compute per-sequence TF binding scores. For a single (extant or ancestral) sequence, this score reflects the collective binding affinity of the input TF set. Then, for each branch in the phylogeny, REforge computes the difference between the scores for the sequences at the start and end of each branch, which reflects evolutionary gains and losses of TFBS along this branch (Fig. 6.1b). While the TFBS scores of sequences of extant species are not independent as species share evolutionary histories, the branch-specific scores are independent, enabling a direct comparison between branches. REforge then uses the Forward Genomics principle to test whether the given CRE exhibits a TFBS divergence pattern that is associated with the convergent change of the trait of interest. If this CRE is involved in the trait change, one would expect that the branches leading to species with the derived trait exhibit a stronger divergence of the TFBS ensemble than branches leading to lineages in which the ancestral trait is present. This can be directly tested by comparing the branch-specific scores of the two groups of branches.

We applied REforge to identify CREs that preferentially lost TFBS in subterranean mammals exhibiting convergent eye degeneration. Given a set of eye-related TFs, REforge identified thousands of CREs with a significant TFBS loss in subterranean mammals (Langer et al. 2018). These CREs are preferentially located near genes known to be involved in eye development and significantly overlap eye-specific regulatory elements identified in mouse by several experimental methods. A direct comparison to the sequence identity-based Forward Genomics approach showed that REforge has a substantially improved ability to detect functionally relevant CRE

divergence. Overall, this shows that convergent degeneration of the visual system in these species involved the loss of binding sites for many eye-relevant TFs in gene regulatory regions (Langer et al. 2018).

The input set of TFs influences the results of REforge. The choice of TFs relevant for the phenotype of interest can come from the literature, expression data of TFs (reasoning that relevant TFs should be expressed in the respective tissue), or TF knockout studies in model organisms, as we have done for analyzing eye degeneration. For cases where prior knowledge of relevant TFs is limited, we developed TFforge (transcription factor forward genomics) (Langer and Hiller 2019) that allows for the discovery of trait-involved TFs by large-scale divergences of their binding sites in species in which the trait has changed. As for REforge, TFforge relies on ancestral sequence reconstruction and scoring TFBS changes along individual branches. In contrast to REforge, TFforge jointly considers a large set of CNEs and screens all given TF motifs for those that exhibit a preferential and widespread binding site divergence in the CREs of species with the modified phenotype. Applying TFforge to genomic regions bound by the eye TFs CRX and NRL revealed that subterranean mammals have not only lost a significant number of binding sites for these two TFs but also other eye-related TFs that interact or co-bind with CRX and NRL (Langer et al. 2018). The joint application of TFforge and REforge has strong potential to reveal TFBS changes in CREs that could contribute to convergent changes in morphology.

6.4 Concluding Remarks

The identification of genomic changes associated with changes in morphological structures provides a foundation to understand how nature's phenotypic diversity evolved. While there is strong support that mutations in CRE sequences are an important source for morphological evolution, finding the regulatory mutations that contribute to morphological differences is difficult. Indeed, identifying CREs, detecting differences that likely affect CRE function and associating such differences to morphological change is not straightforward. Nevertheless, facilitated by technological and methodological advances, the combination of comparative and functional genomics approaches makes it possible to reveal regulatory differences associated with morphological traits. Specifically, a variety of comparative genomic approaches has been developed to detect differences in the sequence of CRE candidates such as accelerated substitution rates, increased sequence divergence or complete loss. Furthermore, recent methods are able to detect TFBS differences in CREs, which is a more powerful way of identifying changes likely affecting regulatory activity. Results from such genome-wide screens can be intersected with functional genomics data to reveal which of the diverged regulatory element candidates are active in a relevant tissue or timepoint, and thus may contribute to the trait of interest. To explore whether CRE divergence truly translates into regulatory differences, reporter assays

can subsequently test whether the CRE sequence of different species drives different expression patterns. Finally, genome engineering can be used to assess whether introducing the modified CRE in a model organism recapitulates trait changes seen in natural species, and thus establish a causal connection between differences in CREs and the trait. In summary, combining comparative and functional genomics provides a general and widely applicable strategy to reveal insights into the genomic basis of morphological differences.

References

Abzhanov A, Protas M, Grant BR, Grant PR, Tabin CJ (2004) Bmp4 and morphological variation of beaks in Darwin's finches. Science 305(5689):1462–1465. https://doi.org/10.1126/science.1098095

Aerts S (2012) Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. Curr Top Dev Biol 98:121–145. https://doi.org/10.1016/B978-0-12-386499-4.00005-7

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, Ntini E, Arner E, Valen E, Li K, Schwarzfischer L, Glatz D, Raithel J, Lilje B, Rapin N, Bagger FO, Jorgensen M, Andersen PR, Bertin N, Rackham O, Burroughs AM, Baillie JK, Ishizu Y, Shimizu Y, Furuhata E, Maeda S, Negishi Y, Mungall CJ, Meehan TF, Lassmann T, Itoh M, Kawaji H, Kondo N, Kawai J, Lennartsson A, Daub CO, Heutink P, Hume DA, Jensen TH, Suzuki H, Hayashizaki Y, Muller F, Forrest ARR, Carninci P, Rehli M, Sandelin A (2014) An atlas of active enhancers across human cell types and tissues. Nature 507(7493):455–461. https://doi.org/10.1038/nature12787

Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 339(6123):1074–1077. https://doi.org/10.1126/science.1232542

Averof M, Patel NH (1997) Crustacean appendage evolution associated with changes in Hox gene expression. Nature 388(6643):682–686. https://doi.org/10.1038/41786

Azofeifa JG, Allen MA, Hendrix JR, Read T, Rubin JD, Dowell RD (2018) Enhancer RNA profiling predicts transcription factor activity. Genome Res. https://doi.org/10.1101/gr.225755.117

Bell O, Tiwari VK, Thoma NH, Schubeler D (2011) Determinants and dynamics of genome accessibility. Nat Rev Genet 12(8):554–564. https://doi.org/10.1038/nrg3017

Berger MJ, Wenger AM, Guturu H, Bejerano G (2018) Independent erosion of conserved transcription factor binding sites points to shared hindlimb, vision and external testes loss in different mammals. Nucleic Acids Res 46(18):9299–9308. https://doi.org/10.1093/nar/gky741

Booker BM, Friedrich T, Mason MK, VanderMeer JE, Zhao J, Eckalbar WL, Logan M, Illing N, Pollard KS, Ahituv N (2016) Bat accelerated regions identify a bat forelimb specific enhancer in the HoxD locus. PLoS Genet 12(3):e1005738. https://doi.org/10.1371/journal.pgen.1005738

Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, Gardner K, Hillier LW, Janette J, Jiang L, Kasper D, Kawli T, Kheradpour P, Kundaje A, Li JJ, Ma L, Niu W, Rehm EJ, Rozowsky J, Slattery M, Spokony R, Terrell R, Vafeados D, Wang D, Weisdepp P, Wu YC, Xie D, Yan KK, Feingold EA, Good PJ, Pazin MJ, Huang H, Bickel PJ, Brenner SE, Reinke V, Waterston RH, Gerstein M, White KP, Kellis M, Snyder M (2014) Comparative analysis of regulatory information and circuits across distant species. Nature 512(7515):453–456. https://doi.org/10.1038/nature13668

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10(12):1213–1218. https://doi.org/10.1038/nmeth.2688

Burga A, Wang W, Ben-David E, Wolf PC, Ramey AM, Verdugo C, Lyons K, Parker PG, Kruglyak L (2017) A genetic signature of the evolution of loss of flight in the Galapagos cormorant. Science 356(6341). https://doi.org/10.1126/science.aal3345

Capra JA, Erwin GD, McKinsey G, Rubenstein JL, Pollard KS (2013) Many human accelerated regions are developmental enhancers. Philos Trans R Soc Lond B Biol Sci 368(1632):20130025. https://doi.org/10.1098/rstb.2013.0025

Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134(1):25–36. https://doi.org/10.1016/j.cell.2008.06.030

Chan YF, Marks ME, Jones FC, Villarreal G Jr, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, Myers RM, Petrov D, Jonsson B, Schluter D, Bell MA, Kingsley DM (2010) Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science 327(5963):302–305. https://doi.org/10.1126/science.1182213

Cockerill PN (2011) Structure and function of active chromatin and DNase I hypersensitive sites. FEBS J 278(13):2182–2210. https://doi.org/10.1111/j.1742-4658.2011.08128.x

Cohn MJ, Tickle C (1999) Developmental basis of limblessness and axial patterning in snakes. Nature 399(6735):474–479. https://doi.org/10.1038/20944

Cotney J, Leng J, Yin J, Reilly SK, DeMare LE, Emera D, Ayoub AE, Rakic P, Noonan JP (2013) The evolution of lineage-specific regulatory activities in the human embryonic limb. Cell 154(1):185–196. https://doi.org/10.1016/j.cell.2013.05.056

Davidson EH, McClay DR, Hood L (2003) Regulatory gene networks and the properties of the developmental process. Proc Natl Acad Sci USA 100(4):1475–1480. https://doi.org/10.1073/pnas.0437746100

de Laat W, Duboule D (2013) Topology of mammalian developmental enhancers and their regulatory landscapes. Nature 502(7472):499–506. https://doi.org/10.1038/nature12753

Dermitzakis ET, Clark AG (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol 19(7):1114–1121. https://doi.org/10.1093/oxfordjournals.molbev.a004169

Erwin GD, Oksenberg N, Truty RM, Kostka D, Murphy KK, Ahituv N, Pollard KS, Capra JA (2014) Integrating diverse datasets improves developmental enhancer prediction. PLoS Comput Biol 10(6):e1003677. https://doi.org/10.1371/journal.pcbi.1003677

Frankel N, Erezyilmaz DF, McGregor AP, Wang S, Payre F, Stern DL (2011) Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. Nature 474(7353):598–603. https://doi.org/10.1038/nature10200

Frazer KA, Tao H, Osoegawa K, de Jong PJ, Chen X, Doherty MF, Cox DR (2004) Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. Genome Res 14(3):367–372. https://doi.org/10.1101/gr.1961204

Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB (2005) Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature 433(7025):481–487. https://doi.org/10.1038/nature03235

Grice EA, Rochelle ES, Green ED, Chakravarti A, McCallion AS (2005) Evaluation of the RET regulatory landscape reveals the biological relevance of a HSCR-implicated enhancer. Hum Mol Genet 14(24):3837–3845. https://doi.org/10.1093/hmg/ddi408

Gross DS, Garrard WT (1988) Nuclease hypersensitive sites in chromatin. Annu Rev Biochem 57:159–197. https://doi.org/10.1146/annurev.bi.57.070188.001111

Hardison RC (2000) Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet 16(9):369–372

Hardison RC, Taylor J (2012) Genomic approaches towards finding cis-regulatory modules in animals. Nat Rev Genet 13(7):469–483. https://doi.org/10.1038/nrg3242

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39(3):311–318. https://doi.org/10.1038/ng1966

Hiller M, Schaar BT, Bejerano G (2012a) Hundreds of conserved non-coding genomic regions are independently lost in mammals. Nucleic Acids Res 40(22):11463–11476. https://doi.org/10.1093/nar/gks905

Hiller M, Schaar BT, Indjeian VB, Kingsley DM, Hagey LR, Bejerano G (2012b) A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species. Cell Rep 2(4):817–823. https://doi.org/10.1016/j.celrep.2012.08.032

Hockman D, Cretekos CJ, Mason MK, Behringer RR, Jacobs DS, Illing N (2008) A second wave of Sonic hedgehog expression during the development of the bat limb. Proc Natl Acad Sci USA 105(44):16982–16987. https://doi.org/10.1073/pnas.0805308105

Holloway AK, Bruneau BG, Sukonnik T, Rubenstein JL, Pollard KS (2016) Accelerated evolution of enhancer hotspots in the mammal ancestor. Mol Biol Evol 33(4):1008–1018. https://doi.org/10.1093/molbev/msv344

Howard ML, Davidson EH (2004) Cis-Regulatory control circuits in development. Dev Biol 271(1):109–118. https://doi.org/10.1016/j.ydbio.2004.03.031

Huang W, Nevins JR, Ohler U (2007) Phylogenetic simulation of promoter evolution: estimation and modeling of binding site turnover events and assessment of their impact on alignment tools. Genome Biol 8(10):R225. https://doi.org/10.1186/gb-2007-8-10-r225

Hughes JD, Estep PW, Tavazoie S, Church GM (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. J Mol Biol 296(5):1205–1214. https://doi.org/10.1006/jmbi.2000.3519

Inoue F, Ahituv N (2015) Decoding enhancers using massively parallel reporter assays. Genomics 106(3):159–164. https://doi.org/10.1016/j.ygeno.2015.06.005

Kapheim KM, Pan H, Li C, Salzberg SL, Puiu D, Magoc T, Robertson HM, Hudson ME, Venkat A, Fischman BJ, Hernandez A, Yandell M, Ence D, Holt C, Yocum GD, Kemp WP, Bosch J, Waterhouse RM, Zdobnov EM, Stolle E, Kraus FB, Helbing S, Moritz RF, Glastad KM, Hunt BG, Goodisman MA, Hauser F, Grimmelikhuijzen CJ, Pinheiro DG, Nunes FM, Soares MP, Tanaka ED, Simoes ZL, Hartfelder K, Evans JD, Barribeau SM, Johnson RM, Massey JH, Southey BR, Hasselmann M, Hamacher D, Biewer M, Kent CF, Zayed A, Blatti C, 3rd, Sinha S, Johnston JS, Hanrahan SJ, Kocher SD, Wang J, Robinson GE, Zhang G (2015) Social evolution. Genomic signatures of evolutionary transitions from solitary to group living. Science 348(6239):1139–1143. https://doi.org/10.1126/science.aaa4788

Kim J, Cunningham R, James B, Wyder S, Gibson JD, Niehuis O, Zdobnov EM, Robertson HM, Robinson GE, Werren JH, Sinha S (2010) Functional characterization of transcription factor motifs using cross-species comparison across large evolutionary distances. PLoS Comput Biol 6(1):e1000652. https://doi.org/10.1371/journal.pcbi.1000652

Kleftogiannis D, Kalnis P, Bajic VB (2015) DEEP: a general computational framework for predicting enhancers. Nucleic Acids Res 43(1):e6. https://doi.org/10.1093/nar/gku1058

Kleftogiannis D, Kalnis P, Bajic VB (2016) Progress and challenges in bioinformatics approaches for enhancer identification. Brief Bioinform 17(6):967–979. https://doi.org/10.1093/bib/bbv101

Kvon EZ, Kamneva OK, Melo US, Barozzi I, Osterwalder M, Mannion BJ, Tissieres V, Pickle CS, Plajzer-Frick I, Lee EA, Kato M, Garvin TH, Akiyama JA, Afzal V, Lopez-Rios J, Rubin EM, Dickel DE, Pennacchio LA, Visel A (2016) Progressive loss of function in a limb enhancer during snake evolution. Cell 167(3):633–642, e611. https://doi.org/10.1016/j.cell.2016.09.028

Lam MT, Li W, Rosenfeld MG, Glass CK (2014) Enhancer RNAs and regulated transcriptional programs. Trends Biochem Sci 39(4):170–182. https://doi.org/10.1016/j.tibs.2014.02.007

Langer BE, Hiller M (2019) TFforge utilizes large-scale binding site divergence to identify transcriptional regulators involved in phenotypic differences. Nucleic Acids Res 47(4):e19. https://doi.org/10.1093/nar/gky1200

Langer BE, Roscito JG, Hiller M (2018) REforge associates transcription factor binding site divergence in regulatory elements with phenotypic differences between species. Mol Biol Evol 35(12):3027–3040. https://doi.org/10.1093/molbev/msy187

Leal F, Cohn MJ (2016) Loss and re-emergence of legs in snakes by modular evolution of sonic hedgehog and HOXD enhancers. Curr Biol 26(21):2966–2973. https://doi.org/10.1016/j.cub.2016.09.020

Li Y, Chen CY, Kaye AM, Wasserman WW (2015) The identification of cis-regulatory elements: a review from a machine learning perspective. Biosystems 138:6–17. https://doi.org/10.1016/j.biosystems.2015.10.002

Lin Q, Fan S, Zhang Y, Xu M, Zhang H, Yang Y, Lee AP, Woltering JM, Ravi V, Gunter HM, Luo W, Gao Z, Lim ZW, Qin G, Schneider RF, Wang X, Xiong P, Li G, Wang K, Min J, Zhang C, Qiu Y, Bai J, He W, Bian C, Zhang X, Shan D, Qu H, Sun Y, Gao Q, Huang L, Shi Q, Meyer A, Venkatesh B (2016) The seahorse genome and the evolution of its specialized morphology. Nature 540(7633):395–399. https://doi.org/10.1038/nature20595

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, Ward LD, Lowe CB, Holloway AK, Clamp M, Gnerre S, Alfoldi J, Beal K, Chang J, Clawson H, Cuff J, Di Palma F, Fitzgerald S, Flicek P, Guttman M, Hubisz MJ, Jaffe DB, Jungreis I, Kent WJ, Kostka D, Lara M, Martins AL, Massingham T, Moltke I, Raney BJ, Rasmussen MD, Robinson J, Stark A, Vilella AJ, Wen J, Xie X, Zody MC, Broad Institute Sequencing P, Whole Genome Assembly T, Baldwin J, Bloom T, Chin CW, Heiman D, Nicol R, Nusbaum C, Young S, Wilkinson J, Worley KC, Kovar CL, Muzny DM, Gibbs RA, Baylor College of Medicine Human Genome Sequencing Center Sequencing T, Cree A, Dihn HH, Fowler G, Jhangiani S, Joshi V, Lee S, Lewis LR, Nazareth LV, Okwuonu G, Santibanez J, Warren WC, Mardis ER, Weinstock GM, Wilson RK, Genome Institute at Washington U, Delehaunty K, Dooling D, Fronik C, Fulton L, Fulton B, Graves T, Minx P, Sodergren E, Birney E, Margulies EH, Herrero J, Green ED, Haussler D, Siepel A, Goldman N, Pollard KS, Pedersen JS, Lander ES, Kellis M (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478(7370):476–482. https://doi.org/10.1038/nature10530

Lopez-Rios J, Duchesne A, Speziale D, Andrey G, Peterson KA, Germann P, Unal E, Liu J, Floriot S, Barbey S, Gallard Y, Muller-Gerbl M, Courtney AD, Klopp C, Rodriguez S, Ivanek R, Beisel C, Wicking C, Iber D, Robert B, McMahon AP, Duboule D, Zeller R (2014) Attenuated sensing of SHH by Ptch1 underlies evolution of bovine limbs. Nature 511(7507):46–51. https://doi.org/10.1038/nature13289

Marcovitz A, Jia R, Bejerano G (2016) "Reverse genomics" predicts function of human conserved noncoding elements. Mol Biol Evol 33(5):1358–1369. https://doi.org/10.1093/molbev/msw001

McGregor AP, Orgogozo V, Delon I, Zanet J, Srinivasan DG, Payre F, Stern DL (2007) Morphological evolution through multiple cis-regulatory mutations at a single gene. Nature 448(7153):587–590. https://doi.org/10.1038/nature05988

McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke DB, Schaar BT, Wenger AM, Bejerano G, Kingsley DM (2011) Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature 471(7337):216–219. https://doi.org/10.1038/nature09774

Montavon T, Duboule D (2012) Landscapes and archipelagos: spatial organization of gene regulation in vertebrates. Trends Cell Biol 22(7):347–354. https://doi.org/10.1016/j.tcb.2012.04.003

Monti R, Barozzi I, Osterwalder M, Lee E, Kato M, Garvin TH, Plajzer-Frick I, Pickle CS, Akiyama JA, Afzal V, Beerenwinkel N, Dickel DE, Visel A, Pennacchio LA (2017) Limb-enhancer genie: an accessible resource of accurate enhancer predictions in the developing limb. PLoS Comput Biol 13(8):e1005720. https://doi.org/10.1371/journal.pcbi.1005720

Nagy O, Nuez I, Savisaar R, Peluffo AE, Yassin A, Lang M, Stern DL, Matute DR, David JR, Courtier-Orgogozo V (2018) Correlated evolution of two copulatory organs via a single cis-regulatory nucleotide change. Curr Biol 28(21):3450–3457, e3413. https://doi.org/10.1016/j.cub.2018.08.047

Narlikar L, Sakabe NJ, Blanski AA, Arimura FE, Westlund JM, Nobrega MA, Ovcharenko I (2010) Genome-wide discovery of human heart enhancers. Genome Res 20(3):381–392. https://doi.org/10.1101/gr.098657.109

Nitta KR, Jolma A, Yin Y, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EE, Taipale J (2015) Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. Elife 4. https://doi.org/10.7554/elife.04837

Noonan JP, McCallion AS (2010) Genomics of long-range regulatory elements. Annu Rev Genomics Hum Genet 11:1–23. https://doi.org/10.1146/annurev-genom-082509-141651

Noordermeer D, Duboule D (2013) Chromatin looping and organization at developmentally regulated gene loci. Wiley Interdiscip Rev Dev Biol 2(5):615–630. https://doi.org/10.1002/wdev.103

Otto W, Stadler PF, Lopez-Giraldez F, Townsend JP, Lynch VJ, Wagner GP (2009) Measuring transcription factor-binding site turnover: a maximum likelihood approach using phylogenies. Genome Biol Evol 1:85–98. https://doi.org/10.1093/gbe/evp010

Partha R, Chauhan BK, Ferreira Z, Robinson JD, Lathrop K, Nischal KK, Chikina M, Clark NL (2017) Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. Elife 6. https://doi.org/10.7554/elife.25884

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM (2006) In vivo enhancer analysis of human conserved non-coding sequences. Nature 444(7118):499–502. https://doi.org/10.1038/nature05295

Pennacchio LA, Rubin EM (2001) Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet 2(2):100–109. https://doi.org/10.1038/35052548

Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R, Rosenbloom KR, Kent J, Haussler D (2006a) Forces shaping the fastest evolving regions in the human genome. PLoS Genet 2(10):e168. https://doi.org/10.1371/journal.pgen.0020168

Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, Kern AD, Dehay C, Igel H, Ares M Jr, Vanderhaeghen P, Haussler D (2006b) An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443(7108):167–172. https://doi.org/10.1038/nature05113

Prabhakar S, Noonan JP, Paabo S, Rubin EM (2006) Accelerated evolution of conserved noncoding sequences in humans. Science 314(5800):786. https://doi.org/10.1126/science.1130738

Prabhakar S, Visel A, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Morrison H, Fitzpatrick DR, Afzal V, Pennacchio LA, Rubin EM, Noonan JP (2008) Human-specific gain of function in a developmental enhancer. Science 321(5894):1346–1350. https://doi.org/10.1126/science.1159974

Prud'homme B, Gompel N, Rokas A, Kassner VA, Williams TM, Yeh SD, True JR, Carroll SB (2006) Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. Nature 440(7087):1050–1053. https://doi.org/10.1038/nature04597

Prudent X, Parra G, Schwede P, Roscito JG, Hiller M (2016) Controlling for phylogenetic relatedness and evolutionary rates improves the discovery of associations between species' phenotypic and genomic differences. Mol Biol Evol 33(8):2135–2150. https://doi.org/10.1093/molbev/msw098

Rajewsky N, Vergassola M, Gaul U, Siggia ED (2002) Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. BMC Bioinf. 3:30

Roscito JG, Sameith K, Parra G, Langer BE, Petzold A, Moebius C, Bickle M, Rodrigues MT, Hiller M (2018) Phenotype loss is associated with widespread divergence of the gene regulatory landscape in evolution. Nat Commun 9(1):4737. https://doi.org/10.1038/s41467-018-07122-z

Schmitt AD, Hu M, Ren B (2016) Genome-wide mapping and analysis of chromosome architecture. Nat Rev Mol Cell Biol 17(12):743–755. https://doi.org/10.1038/nrm.2016.104

Shapiro MD, Hanken J, Rosenthal N (2003) Developmental basis of evolutionary digit loss in the Australian lizard Hemiergis. J Exp Zool B Mol Dev Evol 297(1):48–56

Sharma V, Lehmann T, Stuckas H, Funke L, Hiller M (2018) Loss of RXFP2 and INSL3 genes in Afrotheria shows that testicular descent is the ancestral condition in placental mammals. PLoS Biol 16(6):e2005293. https://doi.org/10.1371/journal.pbio.2005293

Shlyueva D, Stampfel G, Stark A (2014) Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet 15(4):272–286. https://doi.org/10.1038/nrg3682

Simon JM, Giresi PG, Davis IJ, Lieb JD (2013) A detailed protocol for formaldehyde-assisted isolation of regulatory elements (FAIRE). Curr Protoc Mol Biol Chap 21:Unit 21 26. https://doi.org/10.1002/0471142727.mb2126s102

Sinha S, Ling X, Whitfield CW, Zhai C, Robinson GE (2006) Genome scan for cis-regulatory DNA motifs associated with social behavior in honey bees. Proc Natl Acad Sci USA 103(44):16352–16357. https://doi.org/10.1073/pnas.0607448103

Sinha S, van Nimwegen E, Siggia ED (2003) A probabilistic method to detect regulatory modules. Bioinformatics 19(Suppl 1):i292–i301

Stern DL, Orgogozo V (2008) The loci of evolution: how predictable is genetic evolution? Evolution 62(9):2155–2177. https://doi.org/10.1111/j.1558-5646.2008.00450.x

Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol 203 (2):439–455

Thewissen JG, Cohn MJ, Stevens LS, Bajpai S, Heyning J, Horton WE Jr (2006) Developmental basis for hind-limb loss in dolphins and origin of the cetacean bodyplan. Proc Natl Acad Sci USA 103(22):8414–8418. https://doi.org/10.1073/pnas.0602920103

van Duijvenboden K, de Boer BA, Capon N, Ruijter JM, Christoffels VM (2016) EMERGE: a flexible modelling framework to predict genomic regulatory elements from genomic signatures. Nucleic Acids Res 44(5):e42. https://doi.org/10.1093/nar/gkv1144

Villar D, Flicek P, Odom DT (2014) Evolution of transcription factor binding in metazoans—mechanisms and functional implications. Nat Rev Genet 15(4):221–233. https://doi.org/10.1038/nrg3481

Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, Pennacchio LA (2008) Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat Genet 40(2):158–160. https://doi.org/10.1038/ng.2007.55

Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5(4):276–287. https://doi.org/10.1038/nrg1315

Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in cis and trans gene regulation. Nature 430(6995):85–88. https://doi.org/10.1038/nature02698

Wittkopp PJ, Vaccaro K, Carroll SB (2002) Evolution of yellow gene regulation and pigmentation in Drosophila. Curr Biol 12(18):1547–1556

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards YJ, Cooke JE, Elgar G (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3(1):e7. https://doi.org/10.1371/journal.pbio.0030007

Wray GA (2007) The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8(3):206–216. https://doi.org/10.1038/nrg2063

Wu C, Bingham PM, Livak KJ, Holmgren R, Elgin SC (1979) The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. Cell 16(4):797–806

Yip KY, Cheng C, Gerstein M (2013) Machine learning and genome annotation: a match meant to be? Genome Biol 14(5):205. https://doi.org/10.1186/gb-2013-14-5-205

Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ (2005) Genome-scale identification of nucleosome positions in *S. cerevisiae*. Science 309(5734):626–630. https://doi.org/10.1126/science.1112178

Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, Odeen A, Cui J, Zhou Q, Xu L, Pan H, Wang Z, Jin L, Zhang P, Hu H, Yang W, Hu J, Xiao J, Yang Z, Liu Y, Xie Q, Yu H, Lian J, Wen P, Zhang F, Li H, Zeng Y, Xiong Z, Liu S, Zhou L, Huang Z, An N, Wang J, Zheng Q, Xiong Y, Wang G, Wang B, Wang J, Fan Y, da Fonseca RR, Alfaro-Nunez A, Schubert M, Orlando L, Mourier T, Howard JT, Ganapathy G, Pfenning A, Whitney O, Rivas MV, Hara E, Smith J, Farre M, Narayan J, Slavov G, Romanov MN, Borges R, Machado JP, Khan I, Springer MS, Gatesy J, Hoffmann FG, Opazo JC, Hastad O, Sawyer RH, Kim H, Kim KW, Kim HJ, Cho S, Li N, Huang Y, Bruford MW, Zhan X, Dixon A, Bertelsen MF, Derryberry E, Warren W, Wilson RK, Li S, Ray DA, Green RE, O'Brien SJ, Griffin D, Johnson WE, Haussler D, Ryder OA, Willerslev E, Graves GR, Alstrom P, Fjeldsa J, Mindell DP, Edwards SV, Braun EL, Rahbek C, Burt DW, Houde P, Zhang Y, Yang H, Wang J, Avian Genome C, Jarvis ED, Gilbert MT, Wang J (2014) Comparative genomics reveals insights into avian genome evolution and adaptation. Science 346(6215):1311–1320. https://doi.org/10.1126/science.1251385