Review

# Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same

Matthew T. Weirauch[1] and Timothy R. Hughes[1,2]

[1] Banting and Best Department of Medical Research and Donnelly Centre for Cellular and Biomolecular Research
[2] Department of Molecular Genetics, University of Toronto, Toronto, ON, M5S 3E1, Canada

**Regulatory regions with similar transcriptional output often have little overt sequence similarity, both within and between genomes. Although *cis*- and *trans*-regulatory changes can contribute to sequence divergence without dramatically altering gene expression outputs, heterologous DNA often functions similarly in organisms that share little regulatory sequence similarities (e.g. human DNA in fish), indicating that *trans*-regulatory mechanisms tend to diverge more slowly and can accommodate a variety of *cis*-regulatory configurations. This capacity to 'tinker' with regulatory DNA probably relates to the complexity, robustness and evolvability of regulatory systems, but cause-and-effect relationships among evolutionary processes and properties of regulatory systems remain a topic of debate. The challenge of understanding the concrete mechanisms underlying *cis*-regulatory evolution – including the conservation of function without the conservation of sequence – relates to the challenge of understanding the function of regulatory systems in general. Currently, we are largely unable to recognize functionally similar regulatory DNA.**

## Evolution of gene transcriptional regulation

In the past decade, it has become clear that most sequence under selection in mammals is outside exons, and much of it seems to function as transcriptional regulatory elements [1–6]. Meanwhile, there are accumulating examples of evolutionary 'tinkering' with gene regulation, in which the gain and loss of transcription factor (TF) binding sites is responsible for the divergence of gene regulatory patterns and alteration of individual phenotypes, consistent with the idea that changes in gene regulation represent a predominant mechanism of evolution [7–10]. A naïve interpretation of these observations is that sequence conservation corresponds to the conservation of function, whereas sequence divergence represents a mechanism for the divergence of function. However, an ever-increasing number of observations emphasize that it is also common for *cis*-regulatory regions to diverge in primary sequences – often bearing little similarity by any current sequence analysis methods – whereas their function (i.e. expression output) remains largely the same. These phenomena are not inconsistent with evolutionary theory or what has been found previously by laboratory investigation, but their existence on a genomic scale exemplifies the challenges in establishing a global mechanistic understanding of genome function and evolution. In this review, we consider examples, molecular mechanisms and potential evolutionary benefits of regulatory alterations that have little effect on gene expression output.

## Examples of conserved expression without overt similarities in *cis*-regulatory sequence

The conservation of expression can be defined in more than one way. Perhaps the most straightforward examples involve expression in specific animal tissues and cell types, developmental stages and/or spatial regions. In such cases, regulatory regions from orthologous genes can be tested on an equal footing in reporter assays, and/or the expression of orthologous genes can be examined in their native context. Results can then be scored by the human eye and sequence identity scored using alignments or other established approaches. Table 1 lists a variety of documented instances in which orthologous genes display conserved expression patterns without overtly sharing *cis*-regulatory sequences. One of the earliest studies to report this phenomenon involved enhancers for the alcohol dehydrogenase (*Adh*) genes in *Drosophila* species [11] (enhancers are distinct subregions – typically <1 000 bp – that control the expression of a nearby gene, often in a specific tissue, organ or cell type). Following this study, several groups reported similar instances for other key *Drosophila* developmental genes, including *even-skipped (eve)* and *runt*, with others reporting analogous results in organisms ranging from nematodes to vertebrates (Table 1). In most cases, although the comparison of orthologous regulatory regions reveals an overall lack of similarity, different combinations and orientations of binding sites for a few shared TFs can be found. One study systematically examined the expression in zebrafish driven by multiple regions of the human RET (receptor tyrosine kinase) gene, and found that expression mimicking zebrafish RET can be driven using human enhancers, despite a striking lack of conserved sequences between the two species [12]. In total,

*Corresponding author:* Hughes, T.R. (t.hughes@utoronto.ca).


0168-9525/$ – see front matter  doi:10.1016/j.tig.2009.12.002

**Table 1. Genes with conserved expression patterns despite divergent *cis*-regulatory regions[a]**

| Gene | Summary of results | Taxa | Refs |
|---|---|---|---|
| *Adh* | One CRM increases expression in the midgut when transferred from *D.af* to *D.ha*. Transferring this CRM to *D.gi* does not increase expression, but transferring another does. | Insects | [11] |
| *unc-119* | Similar expression patterns in *C.br* and *C.el*, despite divergent promoter regions. | Nematodes | [97] |
| *Est-6 / Est-5b* | Expressed in *D.me* and *D.ps* in the third segment of the antenna and the maxillary palps. Expression is controlled by different CREs in the two species. | Insects | [98] |
| *even-skipped (eve)* | Expression driven by stripe 2 enhancers (S2Es) of three *Drosophila* species is similar when transferred to *D.me*, despite shuffling and differing number of CREs for five TFs. | Insects | [99] |
| *runt* | Expressed in similar patterns in *D.me* and *D.vi* embryos. Promoter sequences are largely dissimilar (only five major conserved blocks). | Insects | [100] |
| yolk protein genes | *D.me* can control the proper expression of three introduced *D.gi* yolk protein genes despite shuffled regulatory regions. | Insects | [56] |
| Brachyury genes | Expressed in notochord precursor cells in *H.ro* and *C.in*. The minimal *H.ro* promoter has CREs for one TF, but the *C.in* promoter contains CREs for three different TFs. | Sea squirts | [101] |
| *eve* | Despite the results of [99], chimeric S2Es of *D.me* and *D.ps* no longer drive expression of a reporter gene in the wild type pattern, implying masking by coevolved differences. | Insects | [102] |
| *Hoxb2a* | *M.sa* and *D.re* have similar hindbrain expression patterns. Both species, along with puffer fish, human and mouse, have shuffled CREs for four common TFs. | Fish | [57] |
| *Endo16* | Although expression patterns in *L.va* and *S.pu* are similar during embryonic and larval development, six of seven CRMs cannot be aligned, and lack common CREs. | Sea urchins | [103] |
| *single-minded (sim)* | Transferring the *A.ga* enhancer to *D.me* produces the same embryonic expression pattern despite different CRE combinations and extensive shuffling. | Insects | [104] |
| *hairy* | An 8.8 Kb upstream region can produce a similar striped pattern to that of *D.me* in *T.ca* despite divergent promoters. | Insects | [105] |
| *eve* | *D.me* embryonic lethal phenotype is rescued by *D.ps* enhancer when *D.me* S2E is deleted. Despite divergent enhancer sequences, the expression pattern is identical. | Insects | [106] |
| *Otx* | Same complex expression pattern in the anterior embryo regions of *H.ro* and *C.in*, but no conserved CREs. Instead, CREs for five common TFs are shuffled. | Sea squirts | [24] |
| *tailless (tll)* | The introduced *M.do* enhancer drives the same embryonic expression pattern as the *D.me* enhancer. Regulatory regions are shuffled, with CREs for three shared TFs. | Insects | [107] |
| *RET* | CRMs for *D.re* and human drive reporter gene expression in *D.re* embryos consistent with the endogenous gene, despite highly dissimilar sequences. | Vertebrates | [12] |
| *eve* | Despite dissimilar enhancers in *D.me* and six species of scavenger flies (sepsids), sepsid and *D.me* enhancers drive similar expression patterns in transgenic *D.me* embryos. | Insects | [108] |

[a]Abbreviations: *A.ga, Anopheles gambiae; B.mo, Bombyx mori; C.br, Caenorhabditis briggsae; C.el, Caenorhabditis elegans; C.in, Ciona intestinalis; D.af, Drosophila affinidisjuncta; D.gi, Drosophila grimshawi; D.ha, Drosophila hawaiiensis; D.me, Drosophila melanogaster; D.ps, Drosophila pseudoobscura; D.re, Danio rerio; D.vi, Drosophila virilis; H.ro, Halocynthia roretzi; L.va, Lytechinus variegatus; M.do, Musca domestica; M.sa, Morone saxatilis; S.pu, Strongylocentrotus purpuratus; T.ca, Tribolium castaneum.* CRE, *cis*-regulatory element; CRM, *cis*-regulatory module; TF, transcription factor.

11 out of 13 human enhancers drove expression consistently in zebrafish, even in cell types absent in mammals.

Genomic studies suggest that these examples are only the 'tip of the iceberg'. All vertebrates have a related body plan, use similar tissues, organs and cell types, and their development occurs by comparable mechanisms; moreover, their gene complements are similar, and it is reasonable to assume that orthologs will function similarly and will display related expression patterns. Indeed, a recent study showed that, in a microarray expression analysis of 10 large organs, at least one-third of all single-copy orthologs shared some degree of conservation in tissue-specific expression patterns across all vertebrate lineages [13]. However, there is little conserved nonexonic sequence at evolutionary distances farther than those among mammals [14].

The conservation of gene expression can also be defined by the conserved coregulation of orthologs; that is, if two genes have correlated expression across conditions in different species then their coregulation can be inferred to be conserved. This metric has been used to examine the conservation of expression characteristics over longer evolutionary distances [15]. The coexpression of components of pathways and complexes can occur even if expression responses in specific states and conditions are not conserved, so here we will focus our discussion on species with closely related physiology, for which it is likely that genes have largely conserved function and expression, and are likely to be induced or repressed in similar conditions (e.g. in response to nutrient availability). Computational studies have identified coexpressed *Saccharomyces cerevisiae* gene groups using expression data taken from multiple studies, and demonstrated that in many cases *cis* elements enriched upstream of *S. cerevisiae* genes are completely different from enriched sequences upstream of orthologous genes in other yeast species [16–18] (reviewed in [19]). In particular, differences were found upstream of genes encoding subunits of the ribosome (ribosomal proteins or RPs), which are renowned for their tight coexpression pattern across many conditions [20], typically in response to perturbations that influence growth rate. This example represents a case in which at least one mechanism for sequence divergence is known: recent work has discovered that the major controller of ribosomal subunit expression has 'switched' from Tbf1 to Rap1 in *S. cerevisiae* [21] (Figure 1). Other instances of yeast TF substitution between *Candida albicans* and *S. cerevisiae* have been documented in the control of genes involved in galactose metabolism [22] and mating [23], the latter showing that the conservation of cell type-specific gene expression without the conservation of the *cis*-regulatory sequence is

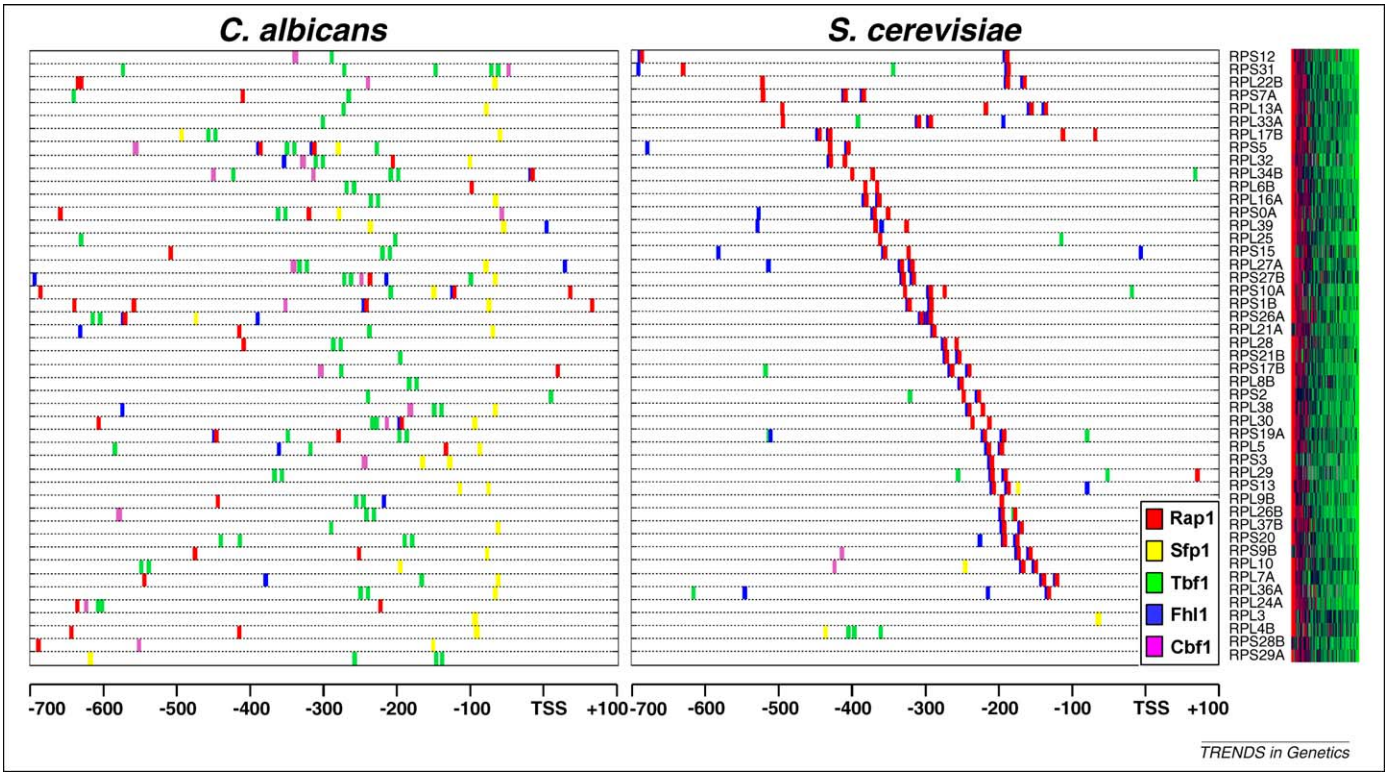


**Figure 1**. An overview of yeast Ribosomal Protein (RP) promoters. Promoter schematics and expression profiles of RPs in two yeast species. Each line displays information for one RP gene (left to right: promoter regions in *C. albicans* and *S. cerevisiae* (from −700 to +100 bp, relative to transcription start site, TSS), gene name and expression pattern). RPs were chosen based on annotations in *S. cerevisiae,* and restricted to those that have a 1:1 ortholog mapping to *C. albicans,* using InParanoid [109]. Colored boxes indicate locations of predicted TF binding sites (see key in bottom right corner). TFs were selected that have at least threefold binding site enrichment in promoter regions of RP genes in either species (relative to randomly selected promoter sets). Genes are sorted (top to bottom) in order of most distal appearance of Rap1 binding sites in *S. cerevisiae* (relative to TSS). Expression values range from twofold downregulated (green) to twofold upregulated (red). 111 experiments from [110] are shown that meet the criterion that at least 10 of the 47 genes have an absolute $\log_{10}$ ratio of at least 0.1, sorted from highest to lowest average ratio among the 47 genes.

also found in single-celled eukaryotes. These studies clearly indicate that the control of crucial processes can change across species for a group of functionally related genes despite an apparent maintenance of overall expression patterns.

An additional relevant observation is that, within a single genome, the precise arrangement of TF binding sites in promoters and enhancers often differs substantially among genes that have very similar expression patterns (Figure 1). Functionally similar enhancers in individual genomes can also have very different arrangements of a related set of binding sites [24,25]. Such cases presumably reflect convergent rather than divergent evolution (i.e. the regulatory regions evolved to have similar output to each other, rather than starting out as identical). Nonetheless, this phenomenon clearly indicates that there are many ways to achieve a similar expression output. On a genomic scale, among the hundreds of thousands of conserved elements in mammals – many of which are believed to be enhancers [26]- most bear little or no sequence similarity to each other [27], indicating that regulatory regions are virtually all unique.

Collectively, these observations show that expression output can be conserved despite divergence and dissimilarity in regulatory regions. Next, we consider two major questions: what mechanisms facilitate the conservation of *cis*-regulatory function in the face of sequence divergence? And what is the benefit, if any, of a system that constantly changes over evolutionary time?

## Mechanisms underlying conserved *cis*-regulation without primary sequence conservation

There are two general types of mechanisms by which a regulatory region can change sequence without affecting the transcriptional regulation of the gene under its control. First, the *trans*-acting factors can change with the regulatory DNA sequence compensating for this change. Second, the *cis*-regulatory sequences can rearrange into a different but functionally equivalent configuration. The first mechanism is easier to understand and is exemplified by several recently described concrete cases in yeast. The second mechanism seems more prevalent, but specific details are harder to pinpoint. Almost certainly, present-day *cis*-regulatory sequences resulted from the combined effect of multiple mechanisms.

### *Changes in trans-acting factors: TF substitution*

Regulatory proteins are often members of families with similar binding affinities. One mechanism for TF substitution might therefore involve mutating a binding site to favor an alternative factor, a phenomenon that has been identified as polymorphisms within human populations [28]. An alternative mechanism for TF substitution, which lacks the requirement for binding affinity similarity, has recently been documented by three different cases in yeast [21–23]. All three require an intermediate stage between the ancestral and current stages (Figure 2), as suggested by True and colleagues [29,30].

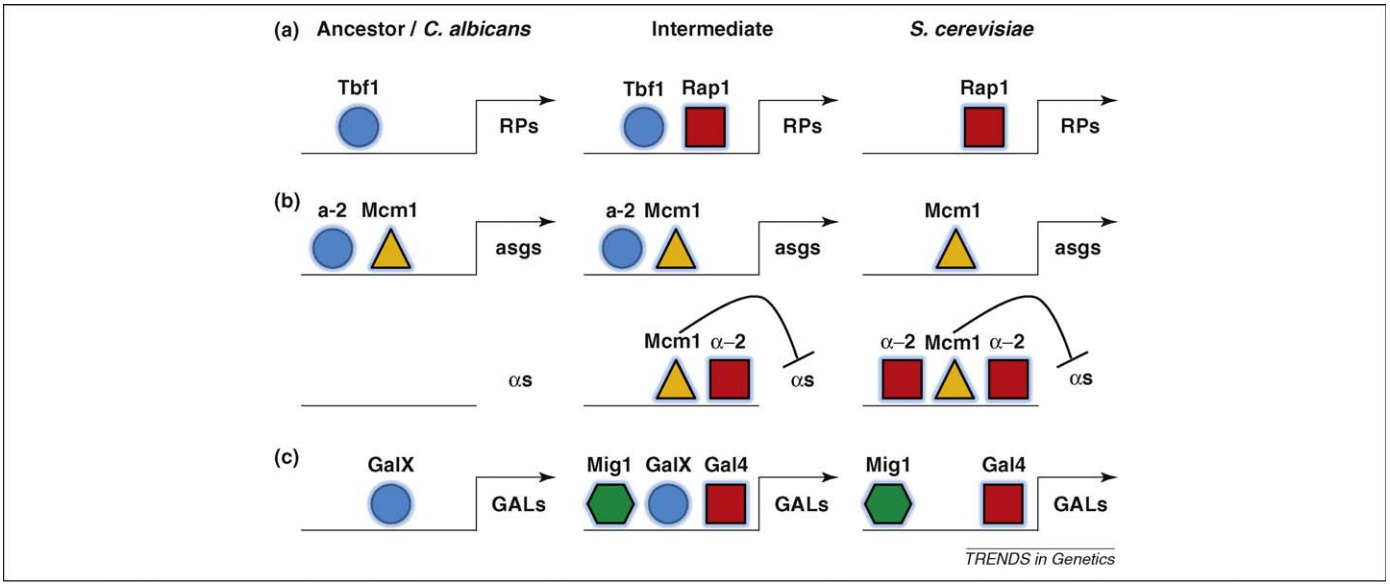


**Figure 2**. Mechanisms of TF switching in yeast. Three documented cases of TF switching between *C. albicans* and *S. cerevisiae*. TFs are depicted as colored shapes, with names depicted above. Gene regulatory regions are depicted as straight lines, with regulated genes indicated to the right. Arrows indicate the activation of the corresponding genes; thick lines ending in a bar indicate repression. **(a)** Mechanism for switching of RP gene control from Tbf1 to Rap1. **(b)** Mechanism for altering the transcriptional control of the mating type (MAT) locus while maintaining the same regulatory output of only expressing asgs in **a**-type cells. Promoter schematics for the MAT**a** and MAT**α** loci are depicted on the top and bottom, respectively. **(c)** Mechanism for switching the control of galactose metabolism (*GAL*) genes from an unknown ancestral regulator (GalX) to Gal4.

One case involves the switching of interacting TFs. The results of computational analyses indicated a strong divergence in the *cis*-regulatory programs of the RPs of *S. cerevisiae* and *C. albicans* [16,17]. A recent study established that RPs are indeed controlled by different TFs in these species [21]. The key players involved in regulatory control in *S. cerevisiae* are the DNA binding proteins Rap1 and Fhl1, which form a complex with Ifh1 [31]. By contrast, the control of the RPs of *C. albicans* is largely achieved by Tbf1, which works in conjunction with Cbf1 [21] (Figure 1).

The reason for this switch remains a topic of speculation, and might relate to the environmental cues to which RP transcription responds [21], but many aspects of the mechanism by which it occurred can be inferred with some confidence. Phylogenetic analysis in 15 sequenced yeast genomes indicates that ancestral RP genes were likely to be regulated by Tbf1 (Figure 2a). Following the split between *S. cerevisiae* and *C. albicans*, RP genes might have been regulated by both Rap1 and Tbf1, allowing an eventual switch to Rap1 regulation in *S. cerevisiae* and relatives [21] (Figure 2a). There are ~150 RP genes in yeast, so numerous *cis* element changes were required for this switch to occur. Figure 1 illustrates that the locations of Tbf1 and Rap1 sites typically do not correspond in orthologous promoters, so more has happened in the interim than the direct conversion of Tbf1 to Rap1 sites. However, Rap1 and Tbf1 do both bind DNA using Myb domains and recognize G/C-rich sequences that share the core sequence ACCC, which might have facilitated the direct mutation of Tbf1 sites to Rap1 sites. Rap1 control of *S. cerevisiae* RP gene expression is thought to function by establishing an open chromatin environment; Tbf1 is also capable of this function [32]. These changes might have also been facilitated by established physical interactions between Rap1 and Tbf1 [21].

The second recently documented example of TF switching involves a common interaction partner. Mating type in *S. cerevisiae* and *C. albicans* is controlled by the MAT locus, which has two forms (MAT**a** and MAT**α**), each of which encodes master regulators for **a**-cells or **α**-cells, respectively [23]. Both yeasts utilize the same basic regulatory output of only expressing **a**-specific genes (asgs) in **a**-type cells. However, in *S. cerevisiae* asgs are on by default, and are repressed by the regulator **α**-2 in **α** cells (Figure 2b). In *C. albicans*, asgs are off by default, and are activated by the regulator **a**-2 in **a**-type cells (Figure 2b). Strikingly, the same end result is achieved in these two species through opposite mechanisms: activation and repression.

This switch from positive to negative regulation was possibly facilitated by a common interaction partner shared by **a**-2 and **α**-2, the ubiquitous activator Mcm1. The ancestral regulatory control of yeast mating genes was likely to be similar to that of *C. albicans* (Figure 2b). Subsequently, a protein interaction evolved between **α**-2 and Mcm1, coincident with the emergence of an **α**-2 site and a strengthening of the Mcm1 binding site in asg regulatory regions (Figure 2b). Through time, the progression towards high A/T content surrounding asg Mcm1 binding sites, which allows Mcm1 to function without a cofactor in modern day *S. cerevisiae* [33], removed the requirement for the **a**-2 activation of asgs. Negative control was strengthened in *S. cerevisiae* by the addition of a second **α**-2 binding site (Figure 2b). This handoff from positive to negative control was made possible by the presence of Mcm1, which allowed the proper regulation of asgs throughout all stages of the transition.

The final recent example involves the switching of non-interacting TFs with unrelated binding affinities. The control of galactose metabolism (*GAL*) genes is one of the best characterized regulatory systems in *S. cerevisiae*

[34]. The activation of *GAL* genes in *S. cerevisiae* is largely achieved by Gal4, which binds the sequence $CGG(N_{11})CCG$ in the promoter regions of GAL1, 2, 3, 7, 10 and 80. Repression is achieved through Mig1. The galactose metabolism pathway is largely conserved (and syntenic) between *C. albicans* and *S. cerevisiae*, and galactose induces the *GAL* genes in both species.

Surprisingly, the *GAL* genes in *C. albicans* are controlled not by Gal4 but by an unknown regulator (GalX), which has a binding sequence (TGTAACGTTACA) unrelated to that of Gal4 [19,22]. The results of a recent study suggest that *C. albicans* Gal4 has a similar binding affinity to *S. cerevisiae* Gal4, but regulates other processes [35]. Phylogenetic analysis of nine other yeast species suggests that the ancestral control of *GAL* genes was achieved by GalX (Figure 2c). Furthermore, *GAL* promoters of intermediate species *Saccharomyces castellii* and *Kluyveromyces lactis* contain both Gal4-like and weak GalX-like and Mig1-like motifs (Figure 2c), suggesting that evolution of the control of *GAL* genes enabled tighter transcriptional control via the concerted efforts of an alternative activator and a new repressor (Figure 2c). The tighter control of *GAL* genes in *S. cerevisiae* might have enabled its specialization for using large quantities of glucose.

### *Changes in trans-acting factors: evolution of TFs*

A handful of studies have documented examples of the evolution of TF binding affinity [16,36–39]. In such cases, the changes in binding affinity presumably require changes in the *cis* sequences that the TF binds. In addition, there are cases in which the repertoire of TFs has expanded in specific lineages by duplication and divergence; for example, the C2H2 zinc fingers in mammals [40], and nuclear receptors in *Caenorhabditis elegans* [41]. However, it is likely that these cases represent exceptions rather than the rule. On the whole, regulatory proteins are among the most slowly evolving of all protein classes [7], and the amino acid sequences of DNA binding domains are usually highly conserved [42]. Likewise, most TF sequence preferences are thought to be largely unchanged over large evolutionary distances [32,43]. Presumably, TF binding affinities tend to be conserved because changes will impact all genes under the TF's control. Indeed, amino acid substitutions in DNA binding domains can have large effects on gene expression [36,37] and can even result in changes at the phenotypic level [44].

### *Cis-regulatory turnover and 'shuffling'*

There is abundant evidence that *cis*-regulatory sequences have a high rate of turnover (i.e. gradual gain and loss) in yeast [45,46], fly [47,48] and vertebrates [49,50]. In animals, known individual functional sites are gained and lost over a timescale of typically $\sim 10^6$–$10^8$ years [49,51,52]. In addition, 'shuffling' (i.e. the local relocation and/or inversion) of sequences seems to be prevalent in the noncoding regions of distantly related vertebrates [50].

Both turnover and shuffling should, in many (if not most) cases, have the capacity to produce regulatory sequences with a similar function to the original. First, enhancers are classically defined by being independent of position and orientation relative to the impacted promoter, so shuffling should generally be tolerated. Second, enhancers typically contain multiple copies of binding sites for one or more TFs [49,53–59], and this can allow them to tolerate the alteration of individual sites. Historically, two general models have been used to explain enhancer identity and function. The *enhanceosome* model invokes specific protein–protein interactions that produce specific spacing and orientation constraints among TF binding sites, whereas the *billboard* model considers enhancers to behave as an ensemble of separately acting units that independently interact with their targets [60]. The enhanceosome model is particularly useful for explaining certain regulatory behaviors [61], and highly conserved noncoding DNA sequences [2] might represent candidate enhanceosome-type enhancers. However, the billboard model is most consistent with existing data [24,58,59] and also consistent with *cis*-regulatory turnover and shuffling. Studies have reported a lack of constraint on binding site orientation in enhancers in a variety of species, including sea urchins, nematodes, insects and mammals (Table 1). Similar principles might apply to proximal promoters: the enrichment of TF binding sequences in groups of target promoters identified by ChIP-chip in mammals is often conserved broadly across vertebrates (in 10 of 16 cases examined extending to chicken, frog or fish), even though the individual binding sites are generally not conserved in alignments [62,63].

Because most eukaryotic TFs are capable of binding multiple sequences with similar affinities [32,43], binding sites can be easily gained (as well as lost). Are newly created TF binding sites utilized? At the very least, it seems as if they are frequently occupied. Recent ChIP-chip and ChIP-seq studies have shown that the number of TF binding sites *in vivo* is large, i.e. many binding sequences are "functional" in binding proteins, even though many of them might not function in gene regulation in any particular situation. For example, CREB (cAMP response element binding protein) binds ~4 000 human promoters *in vivo*, but only a small proportion are induced by cAMP in any cell type [64]. Even in yeast, Gao *et al.* found no correlation between occupancy patterns and gene expression profiles for the majority (67%) of yeast TFs [65]. Moreover, actual binding sites vary dramatically even among species with relatively close evolutionary distances. Odom *et al.* examined binding sites for four TFs in human and mouse hepatocytes and found that 41 to 89% of binding events are species-specific [66], even though the function of the TFs is conserved [67], and liver-specific gene expression programs are highly correlated [13,68]. Furthermore, the species-specific binding events are, for the most part, recapitulated when a human chromosome is placed in mouse, indicating that the *trans*-regulatory apparatus is largely conserved, and that the differing sequences of the chromosome determine the differing arrangement of proteins [69]. Thus, the conservation of expression patterns is tolerant not only of the gain and loss of binding sequences but also the resulting rearrangement of protein binding events.

We stress that multiple factors are involved in the function and evolution of regulatory sequences and that redundancy among these factors presumably plays a role in the malleability of *cis*-regulatory DNA. Moreover, owing to our incomplete understanding of gene regulation pro-

cesses (discussed below), it is possible that factors not normally considered determinants of *cis*-regulatory function might be important and might even be conserved, but not detected in standard multiple sequence alignments. For example, recent papers indicate that DNA structure and topology tend to be conserved and correlate with general or specific aspects of regulation [70,71]. Even simple GC content, which is known to be elevated at promoters and other regulatory regions [72], correlates with nucleosome sequence preferences [73,74].

## Potential benefits of tinkering

These studies demonstrate that it is mechanistically possible to achieve the same expression pattern despite a lack of overt *cis*-regulatory sequence conservation. However, the question still remains as to why *cis*-regulatory regions would be so prone to change in the first place, relative to coding sequences, given that the end output of gene transcription often remains so similar. Are there benefits to continuously altering the mechanics of an already useful and functional regulatory program, especially if the output remains largely the same much of the time? Or at least of having the ability to do so?

It is easy to imagine that providing evolutionary plasticity, while minimizing the risk of failure, might be an advantage. A system that inherently allows minor variation – for example, the addition (or removal) of a sequence that drives expression in a given tissue, without otherwise altering the regulatory properties of the gene – might have long-term advantages. In this sense, *cis*-regulatory turnover and shuffling might be a byproduct of organizational schemes that ensure consistent function while facilitating variation and neofunctionalization (evolutionary 'tinkering' [75]). Tolerance to changes in *trans* might also contribute to regulatory evolution by broadening the array of *cis*-regulatory sequences that produce an appropriate transcriptional output. A mutation that strengthens an interaction between a TF and its cofactor can compensate for a mutation to the cofactor–DNA interaction, and so can promote *cis* sequence turnover and increase the possibility of interaction with a new cofactor. For example, Mcm1, which enabled the regulatory switch of yeast mating gene control (Figure 2b), has at least five different interaction partners, and these partners vary substantially across yeast species [76]. It is also likely that the redundancy offered by such a system plays a role in avoiding the deleterious effects of uncontrollable mutations. Altogether, it is easy to postulate that existing regulatory schemes, and their capacity to tinker, might contribute to redundancy, robustness, modularity, complexity and evolvability – all concepts now broadly associated with regulatory network properties and hypothesized to be a product of evolutionary processes, or at least favorable for both survival and adaptation [77].

What might lead to the creation of such organizational schemes in the first place? One possibility is that regulating gene expression in animals is itself sufficiently demanding to initiate such a scheme. Müller and Stelling used results of simulations of a model yeast RP gene promoter to show that more complex regulatory architectures are better suited to creating precise expression dynamics [78]. Furthermore, using simulated (arbitrary) regulatory networks, Siegal and Bergman showed that robustness is an inherent property of complex and highly connected networks – even without selective pressure to stabilize outputs, complex regulatory network outputs are inherently stable [79]. Simulation-based studies of *Drosophila* patterning networks also concluded that consistent transcriptional outputs are produced across a wide range of binding site perturbations and TF concentrations [80]. Thus, it is possible that the capacity of gene regulatory systems to tinker could be a byproduct of the fact that complex regulatory architectures are needed to successfully create an animal. It is intriguing to speculate that the drive for complex regulatory architectures and/or the capacity to tinker might also provide an explanation for the fact that more complicated organisms tend to have a larger number of TFs, but that these TFs tend to have less sequence specificity (and, often, more promiscuous binding *in vivo*), relative to simpler organisms. Wunderlich and Mirny systematically documented this phenomenon [81], and concluded (citing [7]) that the promiscuity of eukaryotic TFs is likely to constitute one of many eukaryotic evolutionary novelties, which might enable more evolvable gene regulation and thereby be essential for the evolution of a variety of structures.

At least some of these speculations are likely to be correct – possibly most. But it is difficult to obtain definitive proof for the forces and mechanisms of evolution, even for basic observations and enduring concepts. As noted by Lynch, referring to modeling higher order regulatory schemes, 'systems with this level of complexity do not yield simple analytical solutions' [77]. Nonetheless, a system that allows for tinkering in the form of *cis*-regulatory turnover and shuffling seems to be a recipe for success over evolutionary time, apparently being utilized by most (if not all) of the metazoan lineages existing on the earth today.

## Future directions

Regardless of evolutionary origin, understanding the mechanistic basis of *cis*-regulatory turnover and shuffling is tied to understanding gene regulation in general. To better understand how gene regulation evolves, it would be valuable to first understand how it works. The utility of comparative genomics as a tool to identify regulatory mechanisms is likely to increase as more genomes are sequenced and more expression data become available. Xie *et al.* discovered hundreds of motifs enriched among conserved noncoding elements in the human genome, many of which correspond to known TF binding sites [82]. Even when they are not conserved, these motifs display intriguing properties in their occurrence in the genome, often being enriched near transcription start sites. It might yet be possible to derive rules of gene regulation by pure sequence analysis given enough data. At the level of primary sequences, the initial observations by Sanges *et al.* [50] suggest that there is a residual signal even in human vs. fish alignments that allows for shuffling. It will be intriguing to see what more can be learned by having multiple genome sequences at mammal-like distances among different branches of vertebrates, together with ChIP-chip, ChIP-seq and expression data from multiple

species. Although there are inherent signal-to-noise limits associated with direct sequence comparisons [13,50], it might be feasible to align binding sites, or combinations of binding sites, rather than primary sequences [83], and compare these to measured TF binding and gene expression data. Strategies such as 'network-level conservation' [84] might also be useful, particularly in smaller genomes, in identifying species-specific binding sites.

Although all of these approaches are informative, they generally do not attempt to solve what is likely to be the single major problem in understanding both the function and evolution of *cis*-regulatory sequences. As stated by Carroll, 'while we are often able to infer coding sequence function from primary sequences, we are generally unable to decipher functional properties from mere inspection of noncoding sequences' [7]. This problem underlies the difficulty in identifying functionally similar regulatory sequences both within and among genomes, as well as a host of related frustrations, such as the challenge of understanding how TFs select binding sites and targets, how chromatin configuration is established, and how individual genes are regulated [6]. Advances in constructing general models of *cis*-regulatory function will be instrumental in mapping the evolution of gene regulatory mechanisms, because they would allow us to ask if sequences of genes with conserved regulatory characteristics do indeed encode functionally equivalent *cis*-regulatory information given the cellular environment.

Despite some successes [85–88], learning to recognize the 'rules' of *cis*-regulatory function of DNA sequences from examples – the foundation of most supervised learning approaches – has, in our view, been less successful than might have been hoped, even in relatively small genomes. The difficulty of learning logical *cis*-regulatory rules, and the observation that regulatory regions are, in general, unique within a genome [25,27,89], raises the possibility that specific logical rules might not be the best way to encapsulate the functional specificity of most *cis*-regulatory mechanisms. Generative modeling [86,89–91], which incorporates knowledge of the physical mechanisms at work, might provide a way forward. This strategy uses the known properties of DNA binding proteins, including nucleosomes, to predict the configurations that they are likely to adopt on chromosomes as an ensemble. Activities beyond DNA binding, such as mapping these configurations to transcriptional output, could also be included. This strategy is obviously complicated – it involves learning physical models rather than logical rules and requires either concrete knowledge or inferred properties of the activities of individual contributing factors. Nonetheless, it is appealing because it bypasses learning sequence rules from example and inherently incorporates 'context dependence' (the fact that TF binding sites behave differently at different loci and that orthologous loci are often differentially occupied by TFs in different genomes) [62,66,69]. Once constructed, such models could be applied to different genomes, and the key features that determine transcriptional outputs could be compared among related genomes [86].

Building and testing models requires both inputs and training/test cases. In this regard, it will be invaluable to have a more complete knowledge of the inherent activities of TFs and chromatin proteins. Several groups have reported systematic analyses of the binding specificities of TFs [32,43,92–94]. To build models of physical mechanisms and compare them among species with different sets of TFs, it will be important to have a complete index for multiple related species. Nucleosome sequence preferences are also a subject of intense activity [95], as is the determination and modeling of the impact that TF binding has on chromatin configuration and transcription [32,96]. *In vivo* data from species to be compared – for example, ChIP-chip and ChIP-seq and, in particular, responses to experimental perturbations – will provide not only a means to train models and confirmation that models are accurate, but itself can make fundamental findings, as noted above. Such data sets are, however, relatively sparse.

## Concluding remarks

*Cis*-regulatory turnover and shuffling over evolutionary time might represent a rule rather than an exception, and is most frequently described as contributing to evolutionary divergence, presumably via the alteration of gene expression. However, in many cases gene expression patterns are conserved despite extensive shuffling. The robustness of regulatory output to the configuration of individual sequence features might be tied to other aspects and requirements of transcriptional regulatory networks. A more detailed understanding of the molecular mechanisms governing chromatin organization and the regulation of transcription will be invaluable to understanding how *cis*-regulatory sequences tolerate random mutations and respond to selection pressures.

### Acknowledgements

MTW is supported by a scholarship from the Canadian Institute for Advanced Research (CIFAR) Junior Fellows Genetic Networks Program, and by funding from the Ontario Research Fund and Genome Canada through the Ontario Genomics Institute.

### References

1 Waterston, R.H. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562
2 Siepel, A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm and yeast genomes. *Genome Res.* 15, 1034–1050
3 Dermitzakis, E.T. *et al.* (2002) Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* 420, 578–582
4 Cooper, G.M. *et al.* (2004) Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Res.* 14, 539–548
5 Visel, A. *et al.* (2009) Genomic views of distant-acting enhancers. *Nature* 461, 199–205
6 Loots, G.G. (2008) Genomic identification of regulatory elements by evolutionary sequence comparison and functional analysis. *Adv. Genet.* 61, 269–293
7 Carroll, S.B. (2005) Evolution at two levels: on genes and form. *PLoS Biol.* 3, e245
8 Wray, G.A. (2007) The evolutionary significance of *cis*-regulatory mutations. *Nat. Rev. Genet.* 8, 206–216
9 Whitehead, A. and Crawford, D.L. (2006) Variation within and among species in gene expression: raw material for evolution. *Mol. Ecol.* 15, 1197–1211
10 Khaitovich, P. *et al.* (2006) Evolution of primate gene expression. *Nat. Rev. Genet.* 7, 693–702
11 Wu, C.Y. and Brennan, M.D. (1993) Similar tissue-specific expression of the Adh genes from different *Drosophila* species is mediated by distinct arrangements of *cis*-acting sequences. *Mol. Gen. Genet.* 240, 58–64

12 Fisher, S. *et al.* (2006) Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* 312, 276–279
13 Chan, E.T. *et al.* (2009) Conservation of core gene expression in vertebrate tissues. *J. Biol.* 8, 33
14 Thomas, J.W. *et al.* (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424, 788–793
15 Stuart, J.M. *et al.* (2003) A gene–coexpression network for global discovery of conserved genetic modules. *Science* 302, 249–255
16 Gasch, A.P. *et al.* (2004) Conservation and evolution of *cis*-regulatory systems in ascomycete fungi. *PLoS Biol.* 2, e398
17 Tanay, A. *et al.* (2005) Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proc. Natl. Acad. Sci. U. S. A.* 102, 7203–7208
18 Ihmels, J. *et al.* (2005) Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science* 309, 938–940
19 Lavoie, H. *et al.* (2009) Rearrangements of the transcriptional regulatory networks of metabolic pathways in fungi. *Curr. Opin. Microbiol.* 12, 655–663
20 Gasch, A.P. *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* 11, 4241–4257
21 Hogues, H. *et al.* (2008) Transcription factor substitution during the evolution of fungal ribosome regulation. *Mol. Cell* 29, 552–562
22 Martchenko, M. *et al.* (2007) Transcriptional rewiring of fungal galactose-metabolism circuitry. *Curr. Biol.* 17, 1007–1013
23 Tsong, A.E. *et al.* (2006) Evolution of alternative transcriptional circuits with identical logic. *Nature* 443, 415–420
24 Oda-Ishii, I. *et al.* (2005) Making very similar embryos with divergent genomes: conservation of regulatory mechanisms of Otx between the ascidians *Halocynthia roretzi* and *Ciona intestinalis*. *Development* 132, 1663–1674
25 Brown, C.D. *et al.* (2007) Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science* 317, 1557–1560
26 Pennacchio, L.A. *et al.* (2006) *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* 444, 499–502
27 Bejerano, G. *et al.* (2004) Into the heart of darkness: large-scale clustering of human non-coding DNA. *Bioinformatics* 20 (Suppl 1), i40–48
28 Rockman, M.V. and Wray, G.A. (2002) Abundant raw material for *cis*-regulatory evolution in humans. *Mol. Biol. Evol.* 19, 1991–2004
29 True, J.R. and Haag, E.S. (2001) Developmental system drift and flexibility in evolutionary trajectories. *Evol. Dev.* 3, 109–119
30 True, J.R. and Carroll, S.B. (2002) Gene co-option in physiological and morphological evolution. *Annu. Rev. Cell Dev. Biol.* 18, 53–80
31 Rudra, D. *et al.* (2005) Central role of Ifh1p–Fhl1p interaction in the synthesis of yeast ribosomal proteins. *Embo. J.* 24, 533–542
32 Badis, G. *et al.* (2008) A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol. Cell* 32, 878–887
33 Acton, T.B. *et al.* (1997) DNA-binding specificity of Mcm1: operator mutations that alter DNA-bending and transcriptional activities by a MADS box protein. *Mol. Cell Biol.* 17, 1881–1889
34 Lohr, D. *et al.* (1995) Transcriptional regulation in the yeast GAL gene family: a complex genetic network. *Faseb. J.* 9, 777–787
35 Askew, C. *et al.* (2009) Transcriptional regulation of carbohydrate metabolism in the human pathogen *Candida albicans*. *PLoS Pathog.* 5, e1000612
36 Conlon, F.L. *et al.* (2001) Determinants of T box protein specificity. *Development* 128, 3749–3758
37 D'Elia, A.V. *et al.* (2001) Missense mutations of human homeoboxes: a review. *Hum. Mutat.* 18, 361–374
38 Bustamante, C.D. *et al.* (2005) Natural selection on protein-coding genes in the human genome. *Nature* 437, 1153–1157
39 Lopez-Bigas, N. *et al.* (2008) Functional protein divergence in the evolution of *Homo sapiens*. *Genome Biol.* 9, R33
40 Tadepally, H.D. *et al.* (2008) Evolution of C2H2-zinc finger genes and subfamilies in mammals: species-specific duplication and loss of clusters, genes and effector domains. *BMC Evol. Biol.* 8, 176
41 Haerty, W. *et al.* (2008) Comparative analysis of function and interaction of transcription factors in nematodes: extensive conservation of orthology coupled to rapid sequence evolution. *BMC Genomics* 9, 399
42 Luscombe, N.M. and Thornton, J.M. (2002) Protein–DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol.* 320, 991–1009
43 Berger, M.F. *et al.* (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* 133, 1266–1276
44 Brickman, J.M. *et al.* (2001) Molecular effects of novel mutations in Hesx1/HESX1 associated with human pituitary disorders. *Development* 128, 5189–5199
45 Doniger, S.W. and Fay, J.C. (2007) Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput. Biol.* 3, e99
46 Raijman, D. *et al.* (2008) Evolution and selection in yeast promoters: analyzing the combined effect of diverse transcription factor binding sites. *PLoS Comput. Biol.* 4, e7
47 Moses, A.M. *et al.* (2006) Large-scale turnover of functional transcription factor binding sites in *Drosophila*. *PLoS Comput. Biol.* 2, e130
48 Kim, J. *et al.* (2009) Evolution of regulatory sequences in 12 *Drosophila* species. *PLoS Genet.* 5, e1000330
49 Dermitzakis, E.T. and Clark, A.G. (2002) Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* 19, 1114–1121
50 Sanges, R. *et al.* (2006) Shuffling of *cis*-regulatory elements is a pervasive feature of the vertebrate lineage. *Genome Biol.* 7, R56
51 Dermitzakis, E.T. *et al.* (2003) Tracing the evolutionary history of *Drosophila* regulatory regions with models that identify transcription factor binding sites. *Mol. Biol. Evol.* 20, 703–714
52 Costas, J. *et al.* (2003) Turnover of binding sites for transcription factors involved in early *Drosophila* development. *Gene* 310, 215–220
53 Wray, G.A. *et al.* (2003) The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* 20, 1377–1419
54 Ludwig, M.Z. and Kreitman, M. (1995) Evolutionary dynamics of the enhancer region of even-skipped in *Drosophila*. *Mol. Biol. Evol.* 12, 1002–1011
55 Hancock, J.M. *et al.* (1999) High sequence turnover in the regulatory regions of the developmental gene hunchback in insects. *Mol. Biol. Evol.* 16, 253–265
56 Piano, F. *et al.* (1999) Evidence for redundancy but not *trans* factor-*cis* element coevolution in the regulation of *Drosophila* Yp genes. *Genetics* 152, 605–616
57 Scemama, J.L. *et al.* (2002) Evolutionary divergence of vertebrate Hoxb2 expression patterns and transcriptional regulatory loci. *J. Exp. Zool.* 294, 285–299
58 Davidson, E.H. *et al.* (2002) A genomic regulatory network for development. *Science* 295, 1669–1678
59 Kulkarni, M.M. and Arnosti, D.N. (2003) Information display by transcriptional enhancers. *Development* 130, 6569–6575
60 Arnosti, D.N. and Kulkarni, M.M. (2005) Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J. Cell Biochem.* 94, 890–898
61 Struhl, K. (2001) Gene regulation. A paradigm for precision. *Science* 293, 1054–1055
62 Conboy, C.M. *et al.* (2007) Cell cycle genes are the evolutionarily conserved targets of the E2F4 transcription factor. *PLoS One* 2, e1061
63 Ettwiller, L. *et al.* (2008) Analysis of mammalian gene batteries reveals both stable ancestral cores and highly dynamic regulatory sequences. *Genome Biol.* 9, R172
64 Zhang, X. *et al.* (2005) Genome-wide analysis of cAMP-response element binding protein occupancy, phosphorylation and target gene activation in human tissues. *Proc. Natl. Acad. Sci. U. S. A.* 102, 4459–4464
65 Gao, F. *et al.* (2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics* 5, 31
66 Odom, D.T. *et al.* (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet.* 39, 730–732
67 Boj, S.F. *et al.* (2009) Functional targets of the monogenic diabetes transcription factors HNF-1alpha and HNF-4alpha are highly conserved between mice and humans. *Diabetes* 58, 1245–1253

68 Su, A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* 101, 6062–6067
69 Wilson, M.D. *et al.* (2008) Species-specific transcription in mice carrying human chromosome 21. *Science* 322, 434–438
70 Rohs, R. *et al.* (2009) The role of DNA shape in protein-DNA recognition. *Nature* 461, 1248–1253
71 Parker, S.C. *et al.* (2009) Local DNA topography correlates with functional noncoding regions of the human genome. *Science* 324, 389–392
72 Di Filippo, M. and Bernardi, G. (2008) Mapping DNase-I hypersensitive sites on human isochores. *Gene* 419, 62–65
73 Peckham, H.E. *et al.* (2007) Nucleosome positioning signals in genomic DNA. *Genome Res.* 17, 1170–1177
74 Tillo, D. and Hughes, T.R. G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics*, 10, 442.
75 Jacob, F. (1977) Evolution and tinkering. *Science* 196, 1161–1166
76 Tuch, B.B. *et al.* (2008) The evolution of combinatorial gene regulation in fungi. *PLoS Biol.* 6, e38
77 Lynch, M. (2007) The evolution of genetic networks by non-adaptive processes. *Nat. Rev. Genet.* 8, 803–813
78 Muller, D. and Stelling, J. (2009) Precise regulation of gene expression dynamics favors complex promoter architectures. *PLoS Comput. Biol.* 5, e1000279
79 Siegal, M.L. and Bergman, A. (2002) Waddington's canalization revisited: developmental stability and evolution. *Proc. Natl. Acad. Sci. U. S. A.* 99, 10528–10532
80 von Dassow, G. *et al.* (2000) The segment polarity network is a robust developmental module. *Nature* 406, 188–192
81 Wunderlich, Z. and Mirny, L.A. (2009) Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet* 25, 434–440
82 Xie, X. *et al.* (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc. Natl. Acad. Sci. U. S. A.* 104, 7145–7150
83 Palin, K. *et al.* (2006) Locating potential enhancer elements by comparative genomics using the EEL software. *Nat. Protoc.* 1, 368–374
84 Pritsker, M. *et al.* (2004) Whole-genome discovery of transcription factor binding sites by network-level conservation. *Genome Res.* 14, 99–108
85 Beer, M.A. and Tavazoie, S. (2004) Predicting gene expression from sequence. *Cell* 117, 185–198
86 Segal, E. *et al.* (2008) Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature* 451, 535–540
87 Yuan, Y. *et al.* (2007) Predicting gene expression from sequence: a reexamination. *PLoS Comput. Biol.* 3, e243
88 Bussemaker, H.J. *et al.* (2001) Regulatory element detection using correlation with expression. *Nat. Genet.* 27, 167–171
89 Segal, E. and Widom, J. (2009) From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat. Rev. Genet.* 10, 443–456
90 Raveh-Sadka, T. *et al.* (2009) Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res.* 19, 1480–1496
91 Frey, B.J. *et al.* (2005) Genrate: a generative model that finds and scores new genes and exons in genomic microarray data. *Pac. Symp. Biocomput.* 495–506
92 Noyes, M.B. *et al.* (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* 133, 1277–1289
93 Badis, G. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science* 324, 1720–1723
94 Grove, C.A. *et al.* (2009) A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. *Cell* 138, 314–327
95 Kaplan, N. *et al.* (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* 458, 362–366
96 Whitehouse, I. *et al.* (2007) Chromatin remodelling at promoters suppresses antisense transcription. *Nature* 450, 1031–1035
97 Maduro, M. and Pilgrim, D. (1996) Conservation of function and expression of unc-119 from two *Caenorhabditis* species despite divergence of non-coding DNA. *Gene* 183, 77–85
98 Tamarina, N.A. *et al.* (1997) Divergent and conserved features in the spatial expression of the *Drosophila pseudoobscura* esterase-5B gene and the esterase-6 gene of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U. S. A.* 94, 7735–7741
99 Ludwig, M.Z. *et al.* (1998) Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change. *Development* 125, 949–958
100 Wolff, C. *et al.* (1999) Structure and evolution of a pair-rule interaction element: runt regulatory sequences in *D. melanogaster* and *D. virilis*. *Mech. Dev.* 80, 87–99
101 Takahashi, H. *et al.* (1999) Evolutionary alterations of the minimal promoter for notochord-specific Brachyury expression in ascidian embryos. *Development* 126, 3725–3734
102 Ludwig, M.Z. *et al.* (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* 403, 564–567
103 Romano, L.A. and Wray, G.A. (2003) Conservation of Endo16 expression in sea urchins despite evolutionary divergence in both *cis* and *trans*-acting components of transcriptional regulation. *Development* 130, 4187–4199
104 Markstein, M. *et al.* (2004) A regulatory code for neurogenic gene expression in the *Drosophila* embryo. *Development* 131, 2387–2394
105 Eckert, C. *et al.* (2004) Separable stripe enhancer elements for the pair-rule gene hairy in the beetle Tribolium. *EMBO Rep.* 5, 638–642
106 Ludwig, M.Z. *et al.* (2005) Functional evolution of a *cis*-regulatory module. *PLoS Biol.* 3, e93
107 Wratten, N.S. *et al.* (2006) Evolutionary and functional analysis of the tailless enhancer in *Musca domestica* and *Drosophila melanogaster*. *Evol. Dev.* 8, 6–15
108 Hare, E.E. *et al.* (2008) Sepsid even-skipped enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. *PLoS Genet.* 4, e1000106
109 Berglund, A.C. *et al.* (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic. Acids Res.* 36, D263–266
110 Wu, L.F. *et al.* (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.* 31, 255–265