PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

rstb.royalsocietypublishing.org

Review

Cite this article: Sakabe NJ, Nobrega MA. 2013 Beyond the ENCODE project: using genomics and epigenomics strategies to study enhancer evolution. Phil Trans R Soc B 368: 20130022.
http://dx.doi.org/10.1098/rstb.2013.0022

One contribution of 12 to a Theme Issue 'Molecular and functional evolution of transcriptional enhancers in animals'.

Subject Areas:
evolution, genomics

Keywords:
transcription, enhancer, genomics, evolution

Author for correspondence:
Noboru Jo Sakabe
e-mail: nsakabe@uchicago.edu

Beyond the ENCODE project: using genomics and epigenomics strategies to study enhancer evolution

Noboru Jo Sakabe and Marcelo A. Nobrega

Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA

The complex expression patterns observed for many genes are often regulated by distal transcription enhancers. Changes in the nucleotide sequences of enhancers may therefore lead to changes in gene expression, representing a central mechanism by which organisms evolve. With the development of the experimental technique of chromatin immunoprecipitation (ChIP), in which discrete regions of the genome bound by specific proteins can be identified, it is now possible to identify transcription factor binding events (putative *cis*-regulatory elements) in entire genomes. Comparing protein–DNA binding maps allows us, for the first time, to attempt to identify regulatory differences and infer global patterns of change in gene expression across species. Here, we review studies that used genome-wide ChIP to study the evolution of enhancers. The trend is one of high divergence of *cis*-regulatory elements between species, possibly compensated by extensive creation and loss of regulatory elements and rewiring of their target genes. We speculate on the meaning of the differences observed and discuss that although ChIP experiments identify the biochemical event of protein–DNA interaction, it cannot determine whether the event results in a biological function, and therefore more studies are required to establish the effect of divergence of binding events on species-specific gene expression.

1. Introduction

Transcriptional enhancers regulate gene expression in metazoans and control a variety of genetic programmes, particularly during embryonic development. In many cases, enhancers are the key elements that individually regulate the specific spatial and temporal expression of genes with pleiotropic functions. For this reason, changes in enhancers can potentially lead to evolutionary differences in genetic programmes that result in the diversity of form and function in nature.

The multiplicity of enhancers scattered across the human genome, estimated at hundreds of thousands [1], their small sizes (hundreds to a few thousands of base pairs), and the fact that, unlike promoters, they are often located far from their target genes, up to hundreds of kilobases away [2], reviewed by Kleinjan *et al*. [3], makes their identification a formidable problem. Studies of individual enhancers that carefully dissected the mechanisms of gene regulation of a number of genes [4], for example through progressive deletion of select DNA regions, have been critical to our understanding of development at the molecular level and to how mutations in these elements could be a mechanism of evolution. However, only the comparison of large collections of enhancers can be informative about genome-wide changes in gene expression across different species and reveal global patterns of evolution of gene regulation.

In the past few years, techniques previously used in the study of single loci have been adapted and used at the genome scale. Among these, DNAse hypersensitivity and chromatin immunoprecipitation (ChIP) have become the standard way of finding regions that are likely to be regulatory elements (e.g. [5–12]).

The ChIP method (reviewed by Kim & Ren [13]) is based on the fact that the highly dynamic protein–DNA interactions can be 'frozen' and captured by artificially inducing the establishment of covalent bonds between chromatin

Royal Society Publishing
Informing the science of the future

and protein using chemicals such as formaldehyde. Immunoprecipitation of fragmented DNA bound to proteins of interest allows the isolation of DNA where the protein of interest is bound. While, initially, the identification of the DNA sequence bound to proteins was performed in small scale by qPCR, with the advent of microarrays and direct high-throughput sequencing, identification of thousands of DNA fragments spanning the entire genome became common practice. Here, we review studies that used this technique to compare binding maps of orthologous transcription factors and histone modifications in different species and the insights provided by these analyses.

Overall, the trend emerging from multiple studies is that a large fraction of binding events is not conserved across different species, whether they are closely related yeast species [14,15] or more distantly related mammals such as human and mouse [16–22]. The divergence of binding events is also significant between individuals of the same species [23,24] or yeast strains [25]. An association between genetic variation among individuals [23,25] and species-specific repetitive elements [18,21] has been identified, suggesting that variation is largely caused by direct changes to the DNA sequence, although epigenetic differences were also proposed as a cause [26].

The interpretation of the impact of such extensive variation on the regulation of genetic programmes has yet to be explored. The works reviewed here are just the first ones of an exciting era of large-scale studies that has generated an unprecedented amount of data. Analysis of these datasets will deepen our comprehension of how enhancers and other regulatory elements evolved and how and to what extent they changed gene expression programmes and yielded the phenotypic differences that we observe in nature.

2. The pre-ENCODE era: enhancers harbour specific epigenetic signatures

Using ChIP, large collections of enhancers (hundreds to a few thousands) can be identified by targeting transcription factors known to activate transcription. One caveat of this approach is the need for an antibody for each transcription factor, which can be difficult to obtain, and the need for knowledge of the transcription factor networks in a given tissue. Another limitation is the fact that only a specific set of enhancers is identified for each transcription factor. Although this approach is useful for dissecting specific regulatory networks, it prevents analyses of more general enhancer collections.

In addition to transcription factors, many other proteins bind DNA. High levels of chromatin compaction in the nucleus are achieved by wrapping DNA around nucleosomes, protein octamers composed of histones. Histone tails can bear a number of chemical modifications such as acetylation, methylation and phosphorylation [27]. Although the exact role of these chemical modifications in enhancer function is not well established—one possibility being that combinations of histone modifications form a histone code that is recognized by the cell [28]—it is clear that the presence of one modification over another is not random. A large number of studies in the last few years have shown that enhancers can be identified by the presence/absence of specific modifications (reviewed in [29]). Although trimethylation of lysine 4 of histone 3 (H3K4me3) has been found to be present mainly in active promoters, monomethylation (H3K4me1) has been found to be often associated with enhancers [9,10] and has been used to identify hundreds of thousands of enhancers in various tissues and cells [7,8,30].

Later, acetylation of lysine 27 of histone 3 (H3K27ac) was shown to be a better indicator of active enhancers [5,6] and H3K4me1 is now believed to mark both inactive, or poised, and active elements [5–8]. The existence of these molecular signatures makes it possible to identify large collections of enhancers and has been used in several works, including evolutionary studies [21,22]).

3. The ENCODE project

With the development of ChIP at the genome scale, many studies sought to map enhancers and other regulatory elements. A large public consortium named *Encyclopedia of DNA Elements* (*ENCODE*) was created to systematically map functional elements of several cell lines in 1% of the human genome that was selected for its importance to genome research [10]. A functional element was defined as a DNA region that generates a defined product or presents a reproducible biochemical signature and included protein coding genes, non-coding RNAs, promoters and transcriptional regulatory elements. Once the pilot project for the 1% of the human genome was completed [10], the ENCODE project targeted the entire human genome and expanded the number of cells lines being studied [1].

Similar projects were developed for the mouse [31], *Drosophila melanogaster* [32] and *Caenorhabditis elegans* [33]. Future studies will compare enhancers and other DNA elements identified by the four projects and attempt to probe the extent of conservation of functional elements among organisms and eventually shed some light on the evolution of regulatory elements.

One interesting finding of the pilot ENCODE project was that approximately 50% of the DNA elements identified were not conserved across mammals [10], raising the possibility that species-specific gene regulation is pervasive. While this proposition has since been reinforced by other studies, it is important to note that the ENCODE definition of biological function is rather loose and may result in a large number of false-positive ‘functional regions’ of the genome, as discussed elsewhere [34–38]. The central point of critiques regarding ChIP data is that observation of protein–DNA binding events—a biochemical phenomenon that does not imply a selected biological effect or function—does not necessarily correspond to regulatory function. The function of a DNA element might require many other factors invisible to a single ChIP experiment, and therefore not all binding events might be functional. For this reason, when we discuss ChIP data in this review, we offer alternative interpretations to the differences observed under the hypothesis that many of the binding events might be spurious or have no selected regulatory effect. One of the important conclusions we draw is that a better understanding of how divergence of binding event is related in a causal way to gene expression divergence is still required.

4. Using chromatin immunoprecipitation to study genome-wide enhancer evolution

The possibility of identifying thousands of putative enhancers in different organisms and tissues has led to studies that addressed the question of the prevalence of *cis*-regulatory variation across species. Table 1 lists studies that used genome-wide ChIP of select transcription factors and histone modifications associated with enhancers and promoters to compare the fraction of conserved events in orthologous genomic regions of two or more species. Conserved events do not necessarily imply conservation of the nucleotide sequence, only the presence of a ChIP signal in genomic sequences of two or more species that can be identified as orthologous.

The prevailing scenario is one of extensive variation of binding events. The fractions of conserved events for different transcription factors and histone modifications across species compared vary considerably, as shown in table 1, which could be due to the underlying biological specificities of different transcription factors, although technical aspects cannot be ruled out. If we consider that only regions present in all organisms compared were accounted, then the landscape of binding events and histone modifications across species is certainly more diverse.

Variation seems to be pervasive even in closely related organisms. When comparing three species of *Saccharomyces*, only approximately 20% of the binding events for transcription factors Ste12 and Tec1 were conserved among all three species [14]. Individuals of the same species also display considerable variation albeit at lower levels, as expected. Comparison of 10 human individuals revealed a higher fraction of shared events, but as low as 75% in the case of RNA polymerase II (Pol II) [23]. Similarly, when comparing two *Saccharomyces cerevisiae* strains, around 70% of the genes putatively bound by Ste12 were found to be conserved [25].

Several technical considerations inherent to the ChIP technology could generate spurious variance between experiments of the magnitude reported by the above-mentioned studies. Nevertheless, some of these studies have directly addressed these technical caveats and their data still support the notion of extensive variation. For example, some of the studies shown in table 1 generated biological replicates of each species or related cell lines and showed that biological or experimental variation (sampling differences, different experimental or physiological conditions) in these controls was lower than between species or individuals (column 5 in table 1). Other studies performed ChIP with different antibodies to address the issue of different affinity for orthologous proteins, obtaining similar results [26,47]. In other cases, ChIP of a tagged protein where a known epitope is fused to the protein of interest was performed, eliminating variability between orthologous proteins [14,25]. Finally, other studies controlled for environmental and physiological variation, comparing species under the same condition [14] or using sophisticated approaches in which ChIP of two species was simultaneously performed in the same organism [16].

Two exceptions to the high divergence of binding events reported are comparisons of binding profiles of transcription factors in whole embryos of *Drosophila* flies [26,41]. Around 86–100% concordance of binding event locations for six transcription factors was found between *D. melanogaster* and *D. yakuba* [26] (details in table 1). Analysis of twist showed that more than 80% of binding events were shared between *D. melanogaster* and *D. simulans* and *D. yakuba* whole embryos undergoing mesoderm formation, with 34% shared among six *Drosophila* species.

Possible explanations for the discrepancy of results from other studies were discussed, including the use of whole embryos instead of more homogeneous samples such as *Saccharomyces* or cell cultures [39], different evolutionary distances and methodological differences in accounting for presence/absence of binding events [48]. The use of embryonic versus adult tissues raises the possibility that developmental enhancers are highly conserved owing to the stricter and conserved roles of their target genes. It will be interesting to see comparisons of binding events between vertebrate tissues in early developmental stages. Alternatively, because finding similarity between two different samples is more unlikely to occur by chance than finding differences, these works could indicate that variation is not as high or widespread as believed or that *Drosophila* is a notable exception.

Although binding event locations were highly similar, the *Drosophila* binding maps presented differences in binding intensity [26,41]. Bradley *et al*. [26] noted correlated differences such as increased binding intensity of a repressor (giant) associated with decreased binding intensity of an enhancer (bicoid). Differences in binding affinity were also observed in at least another study [14], in which differences of more than 1.5-fold were seen across 23% of identical binding events of yeast showing that variation can occur in many forms.

(a) Causes of transcription factor binding variation

The differences observed in the studies cited above were assumed to be due to differences in *cis*, although in principle they could also be due to differences between the orthologous transcription factors. One elegant experimental design that provided strong evidence that the cause of divergence is largely due to variation in *cis* used a mouse model of Down syndrome [16]. This mouse carries an entire human chromosome 21 in addition to its normal mouse genome. The authors performed ChIP of transcription factors HNF1A, HNF4A and HNF6 and H3K4me3 in liver, a tissue chosen owing to its homogeneity and conserved function. The data simultaneously provided binding maps of the selected transcription factors in human chromosome 21 and in mouse chromosome 16 (that contains three-quarters of the chromosome 21 syntenic regions), eliminating differences in *trans* and in antibody affinity, environmental, developmental and metabolic factors and species-specific conditions. The results were striking, with only 14–18% of the binding events being conserved in both organisms, showing that the DNA sequence drives binding event locations *per se*.

To gain more insights on the molecular basis of the variation of the location of binding events, two studies listed in table 1 assessed the role of genetic variation such as single nucleotide polymorphisms and indels (figure 1 illustrates possible causes of binding event divergence). Comparing 10 human individuals, 35% of the NFKB and 26% of Pol II diverging binding events were found to coincide with genetic variation [23].

Supposedly, such variation would lead to creation or deletion of binding sites for transcription factors. Zheng *et al.* [25] observed that 35% of the motifs in non-conserved binding events were affected by variation, whereas only 1% of the conserved events were affected. Borneman *et al.* [14] observed that 14% of Tec1 and 10% of Ste12 binding regions in *Saccharomyces* had a missing motif for the corresponding transcription factor in the unbound orthologous region, offering a possible explanation for the differences observed. Although intuitive, observations such as these should be interpreted with care, because many functional binding events identified using

Table 1. ChIP studies comparing binding events between species, strains or individuals. This table is an extension of the data compiled by Dowell [39]. We added studies that compared histone modifications, the tissue or cell type/line being compared and several notes. Evolutionary distances between organisms were obtained from Hedges *et al*. [40] except where noted. Distance is given in relation to the first species. *D. mel*, *D. melanogaster*; *D. sim*, *D. simulans*.

organisms compared	tissue/cell line or type	transcription factor/histone modification	conservation level of binding/histone modification region or putatively bound genes	reference conservation level of binding/histone modification region or putatively bound genes[a]	notes	reference
D. melanogaster, *D. yakuba* (6.5 Myr)	whole embryo (collected within 1 h preceding gastrulation)	bcd (*bicoid*)	98–99%		— binding intensities presented differences for all six TFs — stronger binding events are less conserved — suggested an epigenetic cause for correlated binding intensity differences	[26]
		hb (*hunchback*)	85–97%			
		Kr (*Kruppel*)	97–99%			
		gt (*giant*)	98–99%			
		kni (*knirps*)	97–100%			
		cad (*caudal*)	98–99%			
D. melanogaster, *D. annanassae* (44.2 Myr), *D. erecta* (12.8 Myr), *D. pseudoobscura* (33.6 Myr), *D. yakuba* (6.5 Myr)	whole embryo (approx. 2–4 h after egg laying)	twi (*twist*)	approximately 80% (*D. mel* versus *D. sim* and *yakuba*) 57–65% (*D. mel* versus all other) 34% shared across all 6 species	98% *D. mel* twi biological replicate 88% *D. mel* sna biological replicate	— clustering using the differences among binding maps recapitulated the known phylogenetic tree — observed differences in binding intensities — conservation of binding is higher for genes with functions related to the known role of twi — approximately 50% of binding events assigned to genes downregulated in twi knockouts are conserved, whereas conservation is lower than average for binding events near genes unresponsive to twi knockout.	[41]
		sna (*snail*)	approximately 91% (*D. mel* versus *D. sim*)			

(*Continued*.)

Table 1. (*Continued.*)

organisms compared	tissue/cell line or type	transcription factor/histone modification	conservation level of binding/histone modification region or putatively bound genes	reference conservation level of binding/histone modification region or putatively bound genes[a]	notes	reference
S. cerevisiae, *S. mikatae* (5–20 Myr [42]), *S. bayanus* (14.3 Myr)	all species kept under low-nitrogen condition (pseudo-hyphal growth)	Ste12 Tec1	21% 20% (genes putatively bound in three species)	97% 95% (genes shared between biological replicates)	— observed differences in binding intensity — different functions for conserved and species-specific putatively bound genes	[14]
two yeast strains: *S. cerevisiae* strains S96 and HS959	grown in presence and absence of mating pheromone α-factor	Ste12	approximately 70% of the genes between S96 and HS959	$>$0.96 mean Pearson correlation between biological replicates, $<$0.89 between strains	— approximately 78% of the variable binding regions exhibit Mendelian segregation — several cases of transgression were noted — 35% of the motifs in *cis*-variable binding traits were affected by variation versus 1% of the non-variable	[25]
S. cerevisiae, *K. lactis* ($>$100 Myr [43]), *C. albicans* (489.8 Myr)	—	Mcm1	7–42% (genes), 13–18% (genes in three species)	duplicates were pooled	— genes bound in all three species are enriched for cell cycle and mating type. Species-specific targets have new functions	[15]
six human individuals (one trio from CEU and one from YRI)	lymphoblastoid cell lines	CTCF	82% (all six individuals), approximately 99% within populations		— 11% of the sites are allele-specific	[24]

(*Continued.*)

Table 1. (*Continued.*)

organisms compared	tissue/cell line or type	transcription factor/histone modification	conservation level of binding/histone modification region or putatively bound genes	reference conservation level of binding/histone modification region or putatively bound genes[a]	notes	reference
10 human individuals	lymphoblastoid cell lines	NFKB Pol II	92.5% 75% NFKB: 94% TSS[b], 92% non-TSS Pol II: 75% TSS[b], 72% non-TSS	Pol II 68% between humans and chimpanzee	— Pol II binding events: 79% Mendelian, 5% transgression. NKFB binding events: 68% Mendelian, 4% transgression — 35% of the NFKB and 26% Pol II divergent binding events coincided with genetic variations — Spearman correlation between gene expression and NFKB: 0.475 and Pol II: 0.461 — stronger binding events are more often conserved — distinct functions associated with common versus species-specific events	[23]
human, mouse (mutant mouse contains a human chromosome 21) (92.3 Myr)	liver	HNF1A HNF4A HNF6 H3K4me3	18% 34% 14% 91% (TSS), 32% (non-TSS)	— mouse chromosome 16 in mutant versus wild-type mouse: 95–100% — human chromosome 21 in mouse versus in human: 86% HNF1A 82% HNF4A 74% HNF6 86% H3K4me3 (TSS) 78% H4K4me3 (non-TSS)	— used a Down syndrome mouse model to control for experimental and species-specific differences	[16]

(*Continued.*)

Table 1. (*Continued.*)

organisms compared	tissue/cell line or type	transcription factor/histone modification	conservation level of binding/ histone modification region or putatively bound genes	reference conservation level of binding/histone modification region or putatively bound genes[a]	notes	reference
human, mouse (92.3 Myr)	liver	FOXA2 HNF1A HNF4A HNF6	11–45% (genes) 18–20% (genes) 31–59% (genes) 13–27% (genes)	HNF6: 66% of the genes bound in human liver were bound in HepG2		[17]
human, mouse (92.3 Myr)	ESCs	OCT4 NANOG CTCF	2%, 3.8%[c] 1.9%, 5.3%[c] 16.7%, 49.6%[c]		— 11/137 genes downregulated in OCT4 knockouts in mouse and human had a nearby OCT4–NANOG binding event — 62/137 genes were cases of turnover — 55/137 of these genes had no binding events — 82-fold enrichment for overlap of OCT4 sites with LTR9B repeats — binding events overlapping repeats: 20.9% for OCT4, 14.6% for NANOG and 11.1% for CTCF in human and 7.2%, 17.1% and 28.3% in mouse — two LTR9B (repeat) binding regions were validated as enhancers	[18]
human, mouse (92.3 Myr)	ESCs	OCT4 NANOG	9.1% (genes) 13% (genes) — 14% of OCT4 peaks <10 kb of the TSS — 21% of NANOG peaks <10 kb of the TSS		— data compared were generated by different studies and platforms (human data from [44])	[19]

(*Continued.*)

Table 1. (*Continued.*)

organisms compared	tissue/cell line or type	transcription factor/histone modification	conservation level of binding/histone modification region or putatively bound genes	reference conservation level of binding/histone modification region or putatively bound genes[a]	notes	reference
human, mouse (92.3 Myr)	human liver, HepG2, islets, acinar; mouse liver, spleen, kidney, brain, testes, pancreatic islets, pancreatic acinar and Min6	E2F4	20% (genes)	genes overlapping among mouse tissues >65–85% of the genes putatively bound; among human tissues: 70–84%	— did not find a relationship between genes downregulated upon E2F4 knockout — approximately 50 genes with nearby E2F4 binding conserved, GO enrichment for cell cycle, proliferation and DNA repair functions. DNA packaging enriched in mouse-only binding events	[20]
human, mouse (92.3 Myr)	human adipose stromal cells versus 3T3-L1 mouse adipocytes	CTCF PPARG H3K4me3, H3K4me1, H3K4me2, H3K27ac	approximately 53% (L1) 21% (L1) approximately 15–30% (>2kb from TSS)		— species-specific histone marks were associated with species-specific expression — histone conservation was higher near the TSS and increased with enrichment of the peaks — conservation near TSS decreases with enrichment for PPARG and CTCF — cases of turnover observed	[21]
human, mouse (92.3 Myr), chicken (296 Myr)	human CD4+, HeLa and Jurkat cells; mouse ESCs and embryonic fibroblasts; chicken 5- and 10-day-old red blood cells	CTCF	7% of the chicken sites (or 2% of the mouse sites) found in both red blood cell are conserved in all three species	conservation within species (% of the largest set): mouse tissues: 27% human cells: 36% chicken: 16%	— only sites that were present in all cells of each species and across species were considered	[45]

(*Continued.*)

Table 1. (*Continued.*)

organisms compared	tissue/cell line or type	transcription factor/histone modification	conservation level of binding/histone modification region or putatively bound genes	reference conservation level of binding/histone modification region or putatively bound genes[a]	notes	reference
human, chimpanzee (6.3 Myr)	lymphoblastoid cell lines	H3K4me3	69.5 ± 0.3% (human–chimpanzee) 63.9 ± 0.4% (human–rhesus) 63.2 ± 0.3% (chimpanzee–rhesus)	overlap within the same species: 77.7 ± 0.4%	— estimated that at most 7% of gene expression differences between human, chimpanzee and 3% between chimpanzee and rhesus correlated with differences in H3K4me3 — conservation was higher for peaks <1kb of the TSS	[46]
human, mouse (92.3 Myr)	various	H3K4me1 H3K4me3 H3K27ac	approximately 20–40% approximately 80% TSS[b], >approximately 10–40% non-TSS approximately 80% TSS[b], >approximately 20–40% non-TSS		— data compared were generated by different studies and platforms — conserved fraction was higher near TSS — conservation increased when regions were present in multiple cell types — fraction of conserved regions increased with number of transcription factors bound	[22]

(*Continued.*)

Table 1. (*Continued.*)

organisms compared	tissue/cell line or type	transcription factor/histone modification	conservation level of binding/ histone modification region or putatively bound genes	reference conservation level of binding/histone modification region or putatively bound genes[a]	notes	reference
human, mouse (92.3 Myr), dog (94.2 Myr), opossum (162.6 Myr), chicken (296 Myr)	liver	CEBPA	2% human–chicken 7% human–opossum 13% human–mouse 0.3% (all five species)	71% human–human 66% mouse–mouse 61% dog–dog 57% opossum–opossum 56% chicken–chicken	— binding events conserved in at least two species are more likely to occur near genes that are differentially expressed upon knockout of CEPBA/HNF4A — no correlation between binding intensity and conservation — observed many cases of turnover events — 20–40% of the motifs located in species-specific binding events were intact. A 'larger fraction' of the motifs were found to be disrupted	[47]
human, mouse (92.3 Myr), dog (94.2 Myr)	liver	HNF4A	12% human–mouse 9% human–dog	72% human–human 77% mouse–mouse 63% dog–dog		[47]
human, chimpanzee (6.3 Myr)	lymphoblastoid cell lines	Pol II	68%		— species-specific events were associated with different functions than conserved events — higher conservation for stronger Pol II binding events	[23]

[a] Data in this column can be used to assess the significance of the divergence shown in the previous column. When available, the data reports the conservation level in a 'reference' group, such as between biological replicates or between different tissues of the same species, or different species when the previous column refers to a comparison within the same species. For example, a 20% conservation level between species is not surprising if replicates shown 20% conservation only.

[b] <1kb of the TSS.

[c] Top 10% binding events.

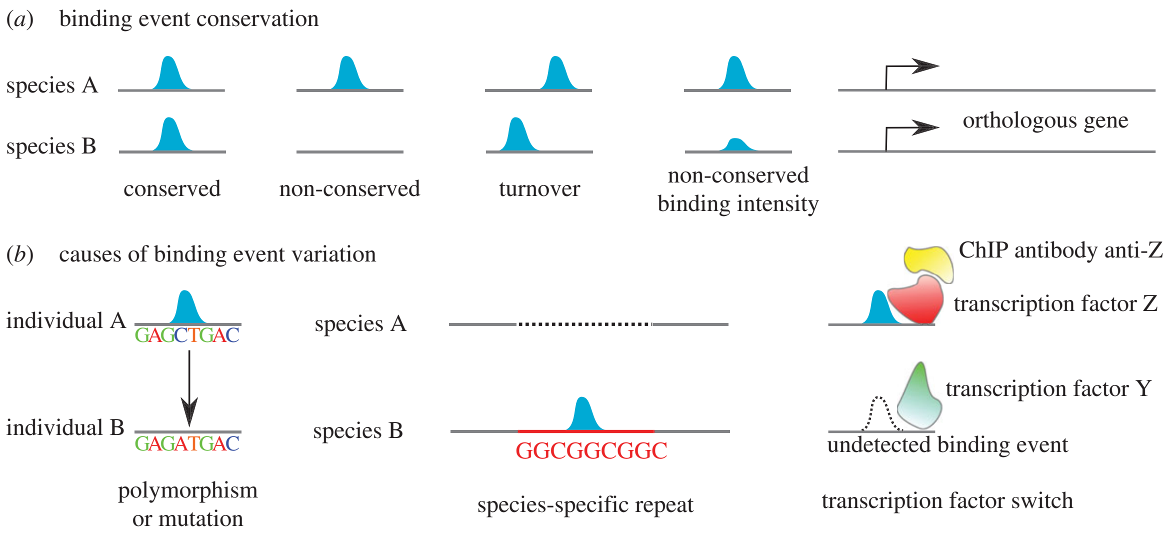

Figure 1. (*a*) Cartoons of ChIP peak signals representing binding events near a target gene. (*b*) Variation in *cis* can potentially alter a DNA motif recognized by a transcription factor and render it unrecognizable and lead to a loss of a binding event. Between species, the appearance of a repeat element or other lineage-specific sequences can create new binding events. Changes of the transcription factor that regulates a given gene can occur during evolution. As ChIP targets specific transcription factors, such changes might be undetected, leading to a false loss of binding event. (Online version in colour.)

ChIP do not contain a recognizable motif for the immunoprecipitated transcription factor [11,12], and therefore the lack of a motif may not directly imply lack of function. In addition, Schmidt *et al.* [47] observed motifs disrupted by changes in the sequence, but 20–40% of the motifs located in binding events lost in a given species were unchanged in a comparison of five vertebrates. One possibility is that disruption of motifs for cofactors might cause loss of function. Analysing *twist* binding events in six *Drosophila* species led the authors to propose that instead of loss of *twist* motifs, loss of cofactor motifs (e.g. *bicoid*, *snail* and *Kruppel*) could partially explain species-specific binding events in *D. melanogaster* [41]. The prevalence of such mechanism still needs to be further explored.

The comparison of 10 human individuals allowed the authors to verify that 79% of the Pol II and 68% of the NKFB binding events in the progeny followed Mendelian segregation, and 5% of the cases were transgression events (parents do not have the event) [23]. Analysis of two yeast strains and their progeny led to an estimation of 78% of the variable Ste12 binding events exhibiting Mendelian segregation with several cases of transgression [25]. Although genetic variation co-occurs with a fraction of the divergent binding events, an even larger fraction remains largely unexplained.

In the case of species-specific variation, repeat elements were proposed to have a role in the origin of novelty of *cis*-regulatory elements. Kunarso *et al.* [18] found that binding regions that overlapped repeats accounted for a considerable fraction of the datasets (20.9% for OCT4, 14.6% for NANOG and 11.1% for CTCF in human and 7.2%, 17.1% and 28.3% in mouse) and that most OCT4 binding regions overlapping a repeat (99.1%) were human-specific. The authors found evidence that repeats are more frequent (22.5% versus 12.4%) in human-specific genes that are putatively directly enhanced by OCT4–NANOG than in targets responsive in both human and mouse and showed *in vivo* enhancer activity for two ERV1 binding regions. Similarly, another study reported that 34% of PPARG binding events in mouse 3T3-L1 adipocytes that could not be mapped to the human genome occurred in rodent-specific transposable element insertions, explaining part of the species-specificity of the dataset [21].

Another putative mechanism controlling differential binding could be epigenetic. The *Drosophila* study that found little variation in binding events for six transcription factors suggested that epigenetic changes might be the mechanism behind differences in binding intensity, a more parsimonious explanation than coordinated changes in all six transcription factor expression levels or binding affinities, but experimental evidence is still lacking [26].

(b) Patterns of protein–DNA binding conservation

(i) Clustered binding events tend to be more conserved

Conservation of the location of binding events is not homogeneous across all events. Analysis of *twist* binding events showed that in *Drosophila* whole embryos undergoing mesoderm formation, conservation decreased with the distance between events and that it was less frequent among isolated events (34%) than those that occurred near the same gene (54%) [41]. The existence of enhancers whose role is to provide robustness to a 'primary enhancer', known as 'shadow enhancers', has been demonstrated in *Drosophila* developmental genes [49,50]. It is possible that the observed higher conservation of nearby enhancers reflects such architectural organization of developmental enhancers, but more studies are required.

(ii) Conservation of protein–DNA binding locations is higher among events near the transcription start site and increases with higher binding affinity

Another source of variability of the level of conservation of binding events is their proximity to transcription start sites (TSSs). Most studies observed that a higher fraction of binding events near the TSS are conserved across species than those occurring in intergenic regions [16,19,22,23,46].

At least three scenarios can be envisioned explaining this observation: (i) many binding events are distal enhancers that fine tune gene expression and their variation is therefore central for tissue, temporal and species-specificity; (ii) because enhancers are believed to contact promoters, and ChIP data are static, it is possible that TSS events are a signal of transcription factor binding to a distal enhancer that contacts the promoter—in this scenario, TSS events are a sum of all enhancer contacts, hence the excess conservation; and (iii) promoter–proximal binding events are critical for gene expression, and therefore tend to be maintained, whereas a higher fraction of distal binding events is free to vary owing to lack of function.

The two latter explanations could be related to the fact that the fraction of conserved binding events also increases with binding intensity of the ChIP signal [23], although another study did not observe this trend [47]. Because binding is a probabilistic event, it is possible that a number of the weaker binding events are a result of random and unproductive binding, as has been shown for 15 genomic regions selected by intensity of binding of the *Drosophila* transcription factor *Kruppel* (Kr) [51]. Although five of six highly bound regions were validated as enhancers, only one of nine of the lowly bound regions behaved as an enhancer at the tested stage. The stronger binding events (more frequent or of higher affinity) could be the ones that are actually involved in gene regulation and tend to be conserved. Current ChIP data do not allow us to establish whether a stronger signal is due to higher affinity or to more accessible chromatin that leads to a higher frequency of transcription factor binding. Alternatively, if TSS events tend to be of higher intensity, then the correlation between binding intensity and binding event conservation might simply reflect the correlation between events in the TSS and binding event conservation.

Another possibility is that weaker binding events are functional, but because they are statistically more difficult to identify they tend to be missed and considered non-conserved.

(iii) Protein–DNA binding conservation is higher when multiple transcription factors are bound and when binding events occur in multiple tissues

The fraction of conserved binding events and histone modifications also varies with their presence in multiple cell types and with the number of co-occurring binding events for other transcription factors. Woo *et al.* [22] observed that conservation of the histone modification H3K4me1, an enhancer marker, between human and mouse increased from 20% when present in two cell types to 50% when present in six cell types.

The same study by Woo *et al.*, using transcription factor binding data generated for HepG2 for the ENCODE project [1,10], showed that the fraction of conserved H3K4me1, H3K4me3 and H3K27ac increased with the number of overlapping HepG2 transcription factor binding events, from approximately 20–35% to approximately 50–80%. The same trend was observed for transcription factors (CEBPA, HNF4A, HNF3B from mouse and human liver; OCT4, NANOG from mouse and human embryonic stem cells (ESCs); six transcription factors from *D. melanogaster* and *D. yakuba*). Again, the fraction of conserved CEBPA binding events increased from approximately 18–30% when HNF3B and HNF4A were bound proximally, and similar increases were observed for *D. melanogaster* transcription factors.

Given that the observation of overlapping binding events and reproducibility in independent cell types is unexpected by random chance, one interpretation of these patterns is that conservation tends to occur among truly functional events and non-conserved events may be enriched for unproductive ones. In molecular terms, because transcription factors act with cofactors, many of the singleton transcription factor binding events observed might not be biologically productive. Therefore, it will be important to generate more comprehensive maps to better understand how binding event variation is related to species-specific gene expression changes.

Another reason why analysing multiple transcription factors at once is important is the knowledge that genetic programmes controlled by a transcription factor in one species might be regulated by a different one in another species. Specific examples of these switches include ribosomal subunit expression regulated by Tbf1 in *Candida albicans* and by Rap1 in *S. cerevisae* [52] or in the case of mating type in these fungi where a complex rewiring took place [53] or galactose metabolism regulated by Cph1p in *C. albicans* and by Gal4p in *S. cerevisiae* [54]. In other words, it is possible that a transcription factor takes the place of another one and, therefore, analysing one single binding map might give the erroneous impression of extensive genetic programme rewiring (illustrated in figure 1).

In conclusion, the correlation between higher fractions of conserved events and their proximity, binding intensity, and enrichment near the TSS and across cell types/tissues might be indicative of their nature; more studies will be required to clarify these patterns.

(c) Binding event turnover and differences in biochemical pathways

Many studies report conservation of putative gene targets instead of location of binding events as shown in table 1 [14,15,17,19,25]. Surprisingly, the fraction of conserved genes is also generally low. This means that considerable differences in the biochemical pathways regulated by the transcription factors analysed are to be expected. Indeed, when analysing the Gene Ontology [55] categories that are enriched in different gene sets, species-specific gene targets correspond to novel functions regulated by the transcription factor under study [14,15,20,23].

Tuch *et al.* [15] found that the transcription factor Mcm1 bound a conserved set of only 12 genes in *S. cerevisiae*, *Kluyveromyces lactis* and *C. albicans*, and there was enrichment for functional terms related to cell cycle and mating type. Another set of 378 genes putatively bound by Mcm1 only in *K. lactis* was found to be enriched for ribosomal genes, a case of a transcription factor regulating an entirely new functional category. Observations such as this provide a mechanistic basis for how changes in *cis*-regulation are an important means as to how species and tissues become different from each other.

On the other hand, there are many cases in which gene targets are conserved, i.e. binding events in two different species are located near the same orthologous genes, but the specific locations of these events are not the same [17,18,21,47] (figure 1). Mikkelsen *et al.* [21] noted that orthologous genes with similar expression patterns often had nearby H3K27ac regions, but the particular locations of these modifications were generally species-specific. In another study [18], only 11 of 137 human and mouse orthologous genes found to respond to OCT4 knockdown shared binding events for OCT4–NANOG, although 72 had at least one binding event for these transcription factors. The other 61 were cases of binding event turnover. Interestingly, Schmidt *et al.* [47] noted that half of the species-specific losses had another binding event within 10 kb. In such cases, the biochemical pathways are maintained, but expression differences might still be driven by the different enhancers, although one idea is that turnover of binding events might be a compensatory mechanism that maintains the local concentration of transcription factors constant near gene targets. Our understanding of how enhancers behave in such cases and how they evolved while maintaining the expression pattern is still anecdotal, and more studies are required to characterize the effect of turnover on gene expression regulation genome-wide.

(d) Binding event divergence and species-specific regulation of gene expression

The correlation between the reported variation of binding event location and phenotypic variability across species is an intuitive and tempting explanation as an evolutionary mechanism of diversity. Linking binding event variation and species-specific gene regulation is therefore critical to support the notion of causality of binding event variation.

The difficulty of such task hinges on the fact that the correlation of ChIP data and gene expression *per se* is complex. Because binding intensity might indicate the frequency an enhancer is active, some studies analysed the correlation between ChIP and gene expression signals. In two studies, the correlation was on the low side (less than 0.5) [23,25]. At least three factors might explain the low correlation: (i) transcription factors act in conjunction with other factors or other mechanisms (e.g. long non-coding RNAs) and each contributes only a fraction of the gene expression pattern; (ii) a considerable fraction of the protein–DNA binding events is non-functional—in fact, a fraction of ChIP binding events often lie near unexpressed genes or genes unresponsive to a transcription factor knockout [11,56,57]; and (iii) many transcription factors act as both activators and repressors, and therefore the correlation of binding intensity with expression levels is a mixture of both positive and negative effects [11,24,58].

One way of separating activating and repressive effects is by analysing a large number of biological replicates or individuals. Analysis of 43 yeast segregants still showed that the vast majority of the negatively (repressor) and positively (enhancer) correlated binding events had Pearson's $R < |0.5|$ [25].

Another way to separate activating and repressive effects is to categorize binding events based on the response of their assigned genes (putative targets) to the absence of the transcription factor in *in vivo* gene knockout studies. Under-expressed genes are assumed to be gene targets normally enhanced by the transcription factor, and the opposite for over-expressing genes. Kunarso *et al.* [18] found that 55 of 137 downregulated genes in OCT4 knockouts in human and mouse did not have a nearby OCT4–NANOG binding event, hinting at indirect regulation, i.e. the knocked out transcription factor binds genes that bind the targets. Such cases raise the question of the role of binding events that are not near responsive genes and highlight the possibility that they are non-functional or not related to gene expression.

Another example of the difficulty of correlating binding events and gene expression comes from a comparison between ChIP-seq data of E2F4 between human and mouse, in which only 20% conservation was observed [20]. Using an E2F4 knockout mouse, the authors were unable to identify a link between this transcription factor and its targets. Transcription factor redundancy might explain the results, but the link between low E2F4 binding event conservation and species-specific gene expression cannot be established without further analyses.

Some of the data available, however, make a case for variation of binding events playing a role in species-specific gene expression. In a comparison between human and chimpanzee, the authors were able to estimate that between 3% and 7% of variation of gene expression could be attributed to H3K4me3, concluding the effect was modest, despite 27–30% differences in H3K4me3 locations [46]. The identification of 378 putative genes regulated by Mcm1 only in *K. lactis* that were highly enriched for ribosomal genes provides an example of a transcription factor acquiring a new regulatory function [15].

The other important aspect of the correlation between regulatory divergence and gene expression is that the latter has been shown to be largely conserved across vertebrates [59–61]. Although it is possible that the small fraction of conserved binding events is responsible for conserved expression patterns, with the majority of non-conserved events providing species-specific regulation, the observation of small numbers of conserved targets [17,19] challenges this hypothesis. In any case, even if extensive changes in regulatory circuits might have occurred, the expression output has been largely maintained. Indeed, radically different circuits may lead to the same genetic programme, as shown by Tsong *et al.* [53] and reviewed by Weirauch & Hughes [62]. Therefore, the actual impact of regulatory variation might not be as profound as the small fraction of conserved events might suggest.

The comparison of human chromosome 21 and the orthologous mouse chromosome 16 in the same genetically engineered mouse provided strong evidence that binding event variation was not a product of experimental variation. However, using genome-wide RNA-seq data from Brawand *et al.* [60] for 5321 orthologues, we calculated that the Spearman correlation between human and mouse liver samples was approximately 0.8 and between replicates approximately 0.95. This result supports the idea that the liver is similar in structure and function across mammals and that it is possible that the high variability of transcription factor binding observed does not necessarily result in profound differences in the transcriptome.

4. Concluding remarks

The advent of ChIP experiments at the genome scale has caused a revolution in the way we identify and think of enhancers. The studies reviewed here have made use of the technique to compare thousands of binding events for specific transcription factors, a proxy for putative enhancers and promoters, at once. Some of the studies analysed multiple transcription factors and we can expect that the collection of binding maps available will only increase with time and allow more comprehensive analyses. It is possible that our view of binding event variation might change in the light of more data as we gain access to more complete collections of enhancers in different conditions and species.

Despite the richness of the data generated by ChIP experiments, current studies have limited power to demonstrate a causative role for binding event divergence and gene expression. Collectively, these works highlight an important limitation of ChIP, namely that it is able to identify the biochemical event of protein–DNA interaction, but not able to directly infer whether this event results in a biological function subjected to the scrutiny and constraint of evolution. The extent to which the differences in protein–DNA binding across species reflect underlying biological differences between these species remains an unresolved question, and more studies directly addressing this problem will be required. Challenges involve better identification of functional enhancers, and an understanding of how enhancers in the same loci compensate for each other and how the differences translate into quantitative gene expression.

The fact that different regulatory circuits can produce the same gene expression output adds another layer of complexity to the interpretation of genome-wide data, because multiple regulatory networks are likely to be contained in the same dataset. The challenge of dissecting these large collections of enhancers will require innovative approaches of data analysis but have the potential of revealing new aspects of gene regulation unapproachable by the study of individual circuits. Although small-scale studies will still be absolutely critical for our understanding of the evolution of *cis*-regulation, genome-wide studies will bring new ways of addressing problems and create new avenues of research on enhancer evolution.

Lastly, if the divergence of biochemical pathways regulated by orthologous transcription factors is indeed as extensive as suggested by the studies discussed here, the implication is that much of the knowledge of specific regulatory elements gathered from model organisms such as the mouse are of little practical application in humans.

Acknowledgements. N.J.S. is a recipient of an American Heart Association post-doctoral fellowship (12POST12100008). The authors declare no competing interests.

Funding statement. This work is partially supported by the National Institutes of Health grants nos. R01DK093972 and R01HL114010 (M.A.N.)

References

1. Dunham I *et al*. 2012 An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74. (doi:10.1038/nature11247)
2. Lettice LA *et al*. 2003 A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735. (doi:10.1093/hmg/ddg180)
3. Kleinjan DA, van Heyningen V. 2005 Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* **76**, 8–32. (doi:10.1086/426833)
4. Davidson EH. 2006 *The regulatory genome: gene regulatory networks in development and evolution*. Pasadena, CA: Academic Press.
5. Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2011 A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283. (doi:10.1038/nature09692)
6. Creyghton MP *et al*. 2010 Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA* **107**, 21 931–21 936. (doi:10.1073/pnas.1016071107)
7. Cui K, Zang C, Roh T-Y, Schones DE, Childs RW, Peng W, Zhao K. 2009 Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell* **4**, 80–93. (doi:10.1016/j.stem.2008.11.011)
8. Ghisletti S *et al*. 2010 Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity* **32**, 317–328. (doi:10.1016/j.immuni.2010.02.008)
9. Heintzman ND *et al*. 2007 Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318. (doi:10.1038/ng1966)
10. Birney E *et al*. 2007 Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816. (doi:10.1038/nature05874)
11. Sakabe NJ, Aneas I, Shen T, Shokri L, Park SY, Bulyk ML, Evans SM, Nobrega MA. 2012 Dual transcriptional activator and repressor roles of TBX20 regulate adult cardiac structure and function. *Hum. Mol. Genet.* **21**, 2194–2204. (doi:10.1093/hmg/dds034)
12. Vokes SA, Ji H, Wong WH, McMahon AP. 2008 A genome-scale analysis of the *cis*-regulatory circuitry underlying sonic hedgehog-mediated patterning of the mammalian limb. *Genes Dev.* **22**, 2651–2663. (doi:10.1101/gad.1693008)
13. Kim TH, Ren B. 2006 Genome-wide analysis of protein–DNA interactions. *Annu. Rev. Genomics Hum. Genet.* **7**, 81–102. (doi:10.1146/annurev.genom.7.080505.115634)
14. Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M. 2007 Divergence of transcription factor binding sites across related yeast species. *Science* **317**, 815–819. (doi:10.1126/science.1140748)
15. Tuch BB, Galgoczy DJ, Hernday AD, Li H, Johnson AD. 2008 The evolution of combinatorial gene regulation in fungi. *PLoS Biol.* **6**, e38. (doi:10.1371/journal.pbio.0060038)
16. Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VL, Fisher EMC, Tavare S, Odom DT. 2008 Species-specific transcription in mice carrying human chromosome 21. *Science* **322**, 434–438. (doi:10.1126/science.1160930)
17. Odom DT *et al*. 2007 Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet.* **39**, 730–732. (doi:10.1038/ng2047)
18. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng H-H, Bourque G. 2010 Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42**, 631–634. (doi:10.1038/ng.600)
19. Loh YH *et al*. 2006 The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* **38**, 431–440. (doi:10.1038/ng1760)
20. Conboy CM *et al*. 2007 Cell cycle genes are the evolutionarily conserved targets of the E2F4 transcription factor. *PLoS ONE* **2**, e1061. (doi:10.1371/journal.pone.0001061)
21. Mikkelsen TS, Xu Z, Zhang X, Wang L, Gimble JM, Lander ES, Rosen ED. 2010 Comparative epigenomic analysis of murine and human adipogenesis. *Cell* **143**, 156–169. (doi:10.1016/j.cell.2010.09.006)
22. Woo YH, Li WH. 2012 Evolutionary conservation of histone modifications in mammals. *Mol. Biol. Evol.* **29**, 1757–1767. (doi:10.1093/molbev/mss022)
23. Kasowski M *et al*. 2010 Variation in transcription factor binding among humans. *Science* **328**, 232–235. (doi:10.1126/science.1183621)
24. McDaniell R *et al*. 2010 Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328**, 235–239. (doi:10.1126/science.1184655)
25. Zheng W, Zhao H, Mancera E, Steinmetz LM, Snyder M. 2010 Genetic analysis of variation in transcription factor binding in yeast. *Nature* **464**, 1187–1191. (doi:10.1038/nature08934)
26. Bradley RK *et al*. 2010 Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related *Drosophila* species. *PLoS Biol.* **8**, e1000343. (doi:10.1371/journal.pbio.1000343)
27. Ruthenburg AJ, Li H, Patel DJ, Allis CD. 2007 Multivalent engagement of chromatin modifications by linked binding modules. *Nat. Rev. Mol. Cell Biol.* **8**, 983–994. (doi:10.1038/nrm2298)
28. Jenuwein T, Allis CD. 2001 Translating the histone code. *Science* **293**, 1074–1080. (doi:10.1126/science.1063127)
29. Sakabe NJ, Nobrega MA. 2010 Genome-wide maps of transcription regulatory elements. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2**, 422–437. (doi:10.1002/wsbm.70)
30. Heintzman ND *et al*. 2009 Histone modifications at human enhancers reflect global cell-type-specific

gene expression. *Nature* **459**, 108–112. (doi:10.1038/nature07829)

31. Stamatoyannopoulos JA *et al*. 2007 An encyclopedia of mouse DNA elements (mouse ENCODE). *Genome Biol.* **13**, 418. (doi:10.1186/gb-2012-13-8-418)
32. Roy S *et al*. 2010 Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**, 1787–1797. (doi:10.1126/science.1198374)
33. Gerstein MB *et al*. 2010 Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330**, 1775–1787. (doi:10.1126/science.1196914)
34. Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. 2013 On the immortality of television sets: 'function' in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol.* **5**, 578–590. (doi:10.1093/gbe/evt028)
35. Doolittle WF. 2013 Is junk DNA bunk? A critique of ENCODE. *Proc. Natl Acad. Sci. USA* **110**, 5294–5300. (doi:10.1073/pnas.1221376110)
36. Eddy SR. 2012 The C-value paradox, junk DNA and ENCODE. *Curr Biol.* **22**, R898–R899. (doi:10.1016/j.cub.2012.10.002)
37. Eddy SR. 2013 The ENCODE project: missteps overshadowing a success. *Curr. Biol.* **23**, R259–R261. (doi:10.1016/j.cub.2013.03.023)
38. Niu DK, Jiang L. 2013 Can ENCODE tell us how much junk DNA we carry in our genome? *Biochem. Biophys. Res. Commun.* **430**, 1340–1343. (doi:10.1016/j.bbrc.2012.12.074)
39. Dowell RD. 2010 Transcription factor binding variation in the evolution of gene regulation. *Trends Genet.* **26**, 468–475. (doi:10.1016/j.tig.2010.08.005)
40. Hedges SB, Dudley J, Kumar S. 2006 TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971–2972. (doi:10.1093/bioinformatics/btl505)
41. He Q, Bardet AF, Patton B, Purvis J, Johnston J, Paulson A, Gogol M, Stark A, Zeitlinger J. 2011 High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species. *Nat. Genet.* **43**, 414–420. (doi:10.1038/ng.808)
42. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003 Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254. (doi:10.1038/nature01644)
43. Rusche LN, Rine J. 2010 Switching the mechanism of mating type switching: a domesticated transposase supplants a domesticated homing endonuclease. *Genes Dev.* **24**, 10–14. (doi:10.1101/gad.1886310)
44. Boyer LA *et al*. 2005 Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–956. (doi:10.1016/j.cell.2005.08.020)
45. Martin D *et al*. 2011 Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes. *Nat. Struct. Mol. Biol.* **18**, 708–714. (doi:10.1038/nsmb.2059)
46. Cain CE, Blekhman R, Marioni JC, Gilad Y. 2011 Gene expression differences among primates are associated with changes in a histone epigenetic modification. *Genetics* **187**, 1225–1234. (doi:10.1534/genetics.110.126177)
47. Schmidt D *et al*. 2010 Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036–1040. (doi:10.1126/science.1186176)
48. Weirauch MT, Hughes TR. 2010 Dramatic changes in transcription factor binding over evolutionary time. *Genome Biol.* **11**, 122. (doi:10.1186/gb-2010-11-6-122)
49. Frankel NS, Davis GK, Vargas D, Wang S, Payre FO, Stern DL. 2010 Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466**, 490–493. (doi:10.1038/nature09158)
50. Hong JW, Hendrix DA, Levine MS. 2008 Shadow enhancers as a source of evolutionary novelty. *Science* **321**, 1314. (doi:10.1126/science.1160631)
51. Fisher WW *et al*. 2012 DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in *Drosophila*. *Proc. Natl Acad. Sci. USA* **109**, 21 330–21 335. (doi:10.1073/pnas.1209589110)
52. Hogues H, Lavoie H, Sellam A, Mangos M, Roemer T, Purisima E, Nantel A, Whiteway M. 2008 Transcription factor substitution during the evolution of fungal ribosome regulation. *Mol. Cell* **29**, 552–562. (doi:10.1016/j.molcel.2008.02.006)
53. Tsong AE, Tuch BB, Li H, Johnson AD. 2006 Evolution of alternative transcriptional circuits with identical logic. *Nature* **443**, 415–420. (doi:10.1038/nature05099)
54. Martchenko M, Levitin A, Hogues H, Nantel A, Whiteway M. 2007 Transcriptional rewiring of fungal galactose-metabolism circuitry. *Curr. Biol.* **17**, 1007–1013. (doi:10.1016/j.cub.2007.05.017)
55. Ashburner M *et al*. 2000 Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29. (doi:10.1038/75556)
56. Marson A *et al*. 2007 Foxp3 occupancy and regulation of key target genes during T-cell stimulation. *Nature* **445**, 931–935. (doi:10.1038/nature05478)
57. Gitter A, Siegfried Z, Klutstein M, Fornes O, Oliva B, Simon I, Bar-Joseph Z. 2009 Backup in gene regulatory networks explains differences between binding and knockout results. *Mol. Syst. Biol.* **5**, 276. (doi:10.1038/msb.2009.33)
58. Yu M *et al*. 2009 Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Mol. Cell* **36**, 682–695. (doi:10.1016/j.molcel.2009.11.002)
59. Chan ET *et al*. 2009 Conservation of core gene expression in vertebrate tissues. *J. Biol.* **8**, 33. (doi:10.1186/jbiol130)
60. Brawand D *et al*. 2011 The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348. (doi:10.1038/nature10532)
61. Barbosa-Morais NL *et al*. 2012 The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587–1593. (doi:10.1126/science.1230612)
62. Weirauch MT, Hughes TR. 2010 Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet.* **26**, 66–74. (doi:10.1016/j.tig.2009.12.002)