A census of human transcription factors: function, expression and evolution

Juan M. Vaquerizas, Sarah K. Kummerfeld‡§, Sarah A. Teichmann‡ and Nicholas M. Luscombe*||*

Abstract | Transcription factors are key cellular components that control gene expression: their activities determine how cells function and respond to the environment. Currently, there is great interest in research into human transcriptional regulation. However, surprisingly little is known about these regulators themselves. For example, how many transcription factors does the human genome contain? How are they expressed in different tissues? Are they evolutionarily conserved? Here, we present an analysis of 1,391 manually curated sequence-specific DNA-binding transcription factors, their functions, genomic organization and evolutionary conservation. Much remains to be explored, but this study provides a solid foundation for future investigations to elucidate regulatory mechanisms underlying diverse mammalian biological processes.

General transcription factor
One of a group of proteins that are essential for transcription from a eukaryotic promoter. They are involved in the formation of the pre-initiation complex and the recruitment of RNA polymerase.

**EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK.*
‡MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK.
§Present address: Department of Bioinformatics, Genentech Inc., South San Francisco, California 94080, USA.
||EMBL-Heidelberg Gene Expression Unit, Meyerhofstrasse 1, Heidelberg D-69117, Germany.
Correspondence to J.M.V or N.M.L.
e-mails: jvaquerizas@ebi.ac.uk; luscombe@ebi.ac.uk
doi:10.1038/nrg2538
Published online
10 March 2009

Cellular life must recognize and respond appropriately to diverse internal and external stimuli. By ensuring the correct expression of specific genes, the transcriptional regulatory system plays a central part in controlling many biological processes, ranging from cell cycle progression[1] and maintenance of intracellular metabolic and physiological balance, to cellular differentiation and developmental time courses[2–4]. Numerous diseases arise from a breakdown in the regulatory system: transcription factors (TFs) are overrepresented among oncogenes[5], and a third of human developmental disorders have been attributed to dysfunctional TFs[6]. Furthermore, alterations in the activity and regulatory specificity of TFs are likely to be a major source for phenotypic diversity and evolutionary adaptation[7–9]. Indeed, increased sophistication of the transcriptional regulatory system seems to have been a principal requirement for the emergence of metazoan life[10–13].

Much of our basic knowledge of transcriptional regulation derives from molecular biological and genetic investigations. Diverse arrays of proteins are crucial for successful transcription by RNA polymerase in eukaryotic cells. These proteins include general transcription factors, co-factors, histones and chromatin remodelling proteins. In addition, a host of sequence-specific DNA-binding TFs direct transcription initiation to specific promoters[14].

The availability of complete genome sequences and the development of high-throughput experimental techniques in the past decade have and continue to provide complementary information describing the function and organization of these regulatory systems on an unprecedented scale. Computational studies have reported TF repertoires by searching for genes containing DNA-binding domains either across all completely sequenced genomes[15], or for individual organisms and phylogenetic groups, including bacteria (such as *Escherichia coli*[16] and *Bacillus subtilis*[17]), fungi[18] (including *Saccharomyces cerevisiae*[19]), animals (including *Caenorhabditis elegans*[20], *Drosophila melanogaster*[21] and *Mus musculus*[22]) and plants[23] (such as *Arabidopsis thaliana*[24]).

For humans, the initial analyses of the complete genome sequence estimated the presence of 200 to 300 component genes for the basic transcriptional machinery, and 2,000 to 3,000 sequence-specific DNA-binding TFs[25,26]. The automated annotation in the Gene Ontology (GO) database[27] (available at Gene Ontology Home), which is based on mapping InterPro[28] DNA-binding domains, currently predicts 1,052 TF genes; of these, only 62 have been experimentally verified for both DNA-binding and regulatory functions (Supplementary information S1 (PDF)). The DBD database predicts 1,508 human loci as TFs[15]. It automatically annotates sequence-specific DNA-binding TFs for all publicly available

completely sequenced genomes based on a set of hidden Markov models of DNA-binding domain families from the Pfam and SUPERFAMILY databases. Several computational studies have examined individual mammalian TF families in detail, but only a few have attempted to identify the full complement of human TFs[29,30].

Some previous studies of TF repertoires — particularly those in large genomes — may contain misleading predictions for several reasons. Most of these studies depended on identifying genes that are homologous to previously characterized regulators; however, there are technical limitations to sequence search methods, and algorithms can sometimes output false positive hits. Moreover, even among true positives, some DNA-binding domains also exist in non-TF proteins, making these domains unreliable markers of sequence-specific DNA-binding functionality. As a result of these difficulties, we still lack a comprehensive characterization of the human TF repertoire.

Here, we overcome some of these difficulties by focusing on a precise definition of sequence-specific DNA-binding regulators, which are among the best-defined protein domains. We also minimize prediction errors by manually examining each locus that encodes a potential DNA-binding function. In doing so, we present a comprehensive and high-quality census of TFs in the human genome. As most of these TFs have not been experimentally characterized for regulatory function, we evaluate their tissue-specific expression, genomic distribution and evolutionary conservation. Together, these results provide a solid foundation for further systematic characterization of human TFs in their biological context, through traditional molecular approaches and also using genomics techniques, such as chromatin immunoprecipitation, protein-binding microarrays and high-throughput SELEX.

Identifying the TF repertoire

To identify the repertoire of TFs in the human genome we define a class of proteins that binds DNA in a sequence-specific manner, but are not enzymatic or do not form part of the core initiation complex. First we assembled a list of DNA-binding domains and families from the InterPro database (release 17). For each entry we examined the description and associated literature to assess their sequence-specific DNA-binding capabilities, which resulted in an accurate list of 347 domains and families (Supplementary information S1 (PDF), S2 (.txt file)). We then extracted 4,610 proteins from the International Protein Index (IPI) database[31] that show a significant match with these selected DNA-binding domains. This group of proteins mapped to 1,960 human genomic loci in the Ensembl Genome Browser database (release 51)[32].

Next, we manually inspected each locus and grouped them according to our confidence in their TF functionality (Supplementary information S1 (PDF); the full data set is found in Supplementary information S3 (.txt file)): at the highest level, probable TFs have experimental evidence for regulatory function in any mammalian organism or have an equivalent protein domain arrangement; possible TFs contain non-promiscuous InterPro DNA-binding domains that are never found in non-TFs, but for which we do not have further functional evidence; and unlikely TFs comprise predicted genes, genes containing promiscuous InterPro DNA-binding domains or genes with an established molecular function other than transcription (such as nucleoporins, threonine phosphatases or splicing factors). Finally, we also included 27 curated probable TFs from other sources, such as GO or TRANSFAC[33]; these TFs contain undefined DNA-binding domains, and were therefore missed using the above procedure.

This resulted in a high-confidence data set of 1,391 genomic loci (~6% of the total number of protein-coding genes) that encode TFs, which we will focus on for the remainder of this Analysis article, and a further 216 loci representing possible TFs (see Supplementary information S3 (.txt file) for the data set). Estimates of the coverage of our approach range from 85% to 94% (Supplementary information S1 (PDF)) suggesting an upper bound of ~1,700–1,900 TF-coding genes in the human genome.

Despite the care that we have taken in compiling this data set, there are a few possible sources of inaccuracies. Our method depends heavily on the content of the InterPro database, and the ability of the search algorithms to detect these domains in protein sequences. The repertoire should be updated when new InterPro entries for newly discovered DNA-binding domains, or refinements of existing ones, and more sensitive search methods become available. In addition, the annotation of the human genome is still in a state of flux — especially in the annotation of genes — so part of the repertoire will be affected by new releases of the genome. Finally, our manual curation depends on the existing literature about each gene, and our own annotations will need to be updated as new findings are reported. The repertoire will be improved as these underlying data sources are updated. Overall, we expect these limitations to be small compared with the improvements that our data provide over previous resources.

Limited knowledge of TF functions

We gauged the extent of our current knowledge about the regulatory function of these TFs by assessing PubMed abstracts and annotations in the GO database. The literature analysis (FIG. 1a), based on the number of times a TF is cited in an abstract, shows an uneven distribution of information that is biased towards those TFs involved in diseases. Three TFs, including the tumour suppressor p53, have accumulated more citations than all other TFs.

Further analysis using the GO database (FIG. 1b) showed that most human TFs are unannotated, indicating that they remain uncharacterized. In fact, when we inspect the source of these annotations, it is evident that most observations are inferred from studies in other organisms and may not apply directly to human orthologues. Of the assigned regulatory functions, control of developmental processes (such as tissue and organ development), cellular processes (for example,

Co-factor
A protein or small molecule that modulates the activity of an enzyme or of another protein complex.

Histone
A small highly conserved basic protein, found in the chromatin of all eukaryotic cells. Histones associate with DNA to form nucleosomes.

Chromatin remodelling protein
A protein that mediates transient changes in chromatin accessibility by modifying the methylation or acetylation status of histones or the methylation status of cytosine residues in DNA.

Gene Ontology
(GO). A widely used classification system of gene functions and other gene attributes that uses a controlled vocabulary.

InterPro
A database of conserved protein families, domains and motifs that can be used to annotate amino acid sequences. The presence of a protein domain is often indicative of a particular molecular function.

SELEX
A procedure to identify protein ligands. For DNA-binding proteins, the protein is mixed with a pool of double-stranded oligonucleotides that contain a random core of nucleotides flanked by specific sequences. The protein–DNA complex is recovered, the oligonucleotides amplified by PCR and sequenced to reveal the binding specificity of the protein.

Orthologues
Loci in two species that are derived from a common ancestral locus by a speciation event. This is different from paralogous members of a gene family that are derived from duplication events.

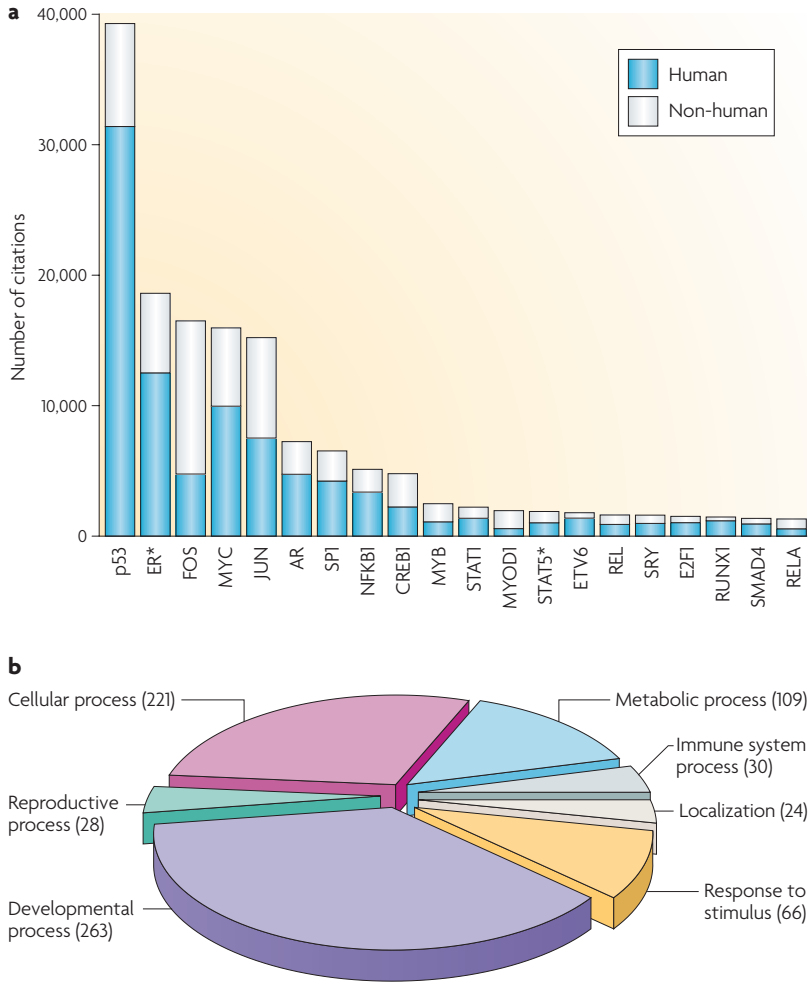

Figure 1 | **Current state of knowledge about transcription factors in the human genome. a** | For the top 20 most cited transcription factors (TFs) in PubMed the number of studies performed in humans (blue bars) and in all other organisms (grey bars) is shown. ER* combines the citations for ERS1 and ERS2, which were indistinguishable in the literature search; similarly, STAT5* includes citations for both STAT5A and STAT5B. **b** | Summary of biological processes regulated by TFs. Annotations were obtained from the Gene Ontology database, excluding those based only in electronic annotation. Numbers of annotated TFs are given in parentheses; each gene can be annotated with more than one function.

signal transduction) and stimulatory response (including immune response and sensory perception) are the most highly represented. Of course, these annotations are often general, and one must return to the original publications in order gain detailed understanding of the functions of the gene.

These observations are not surprising in themselves, but they emphasize how little we know about the biological processes that most of these TFs mediate. Directing research efforts into these uncharacterized TFs — for example, using high-throughput genomic surveys to describe key features combined with detailed examination using traditional molecular approaches — could accelerate our understanding of these regulators. The TF forkhead box P3 (FOXP3) is an excellent example of how research interest can suddenly arise following a key finding, and how follow-up studies can rapidly improve our understanding of regulatory function. FOXP3 was first described in 2001 as the cause of X-linked mouse scurfy and human neonatal diabetes mellitus, enteropathy and endocrinopathy syndrome[34,35]; but only ten further papers were published on this gene during 2001–2002. However, since the discovery of the role of FOXP3 in T-cell development in 2003 (REFS 36,37) there have been 2,382 publications, including several ChIP–chip (chromatin immunoprecipitation coupled with microarray) studies exploring its genome-wide binding sites[38,39]. Greater understanding of how this TF operates has been translated into its clinical use as a marker for transplant rejection[40].

Structural features

The most common classification of TFs is based on the structure of their DNA-binding domains[41]. Grouping TFs by structural domain has been extremely useful in uncovering how they recognize and bind specific DNA sequences, as well as providing insights into their evolutionary histories. Moreover, in some instances the DNA-binding domain provides clues to their function; for example, homeodomain-containing TFs are often associated with developmental processes, and those in the interferon regulatory factor family are generally associated with triggering immune responses against viral infections[41].

We will not describe these families in detail here, as this has been done elsewhere[41]. It is worth noting, however, that three types of TF dominate in the human genome and account for over 80% of the repertoire (FIG. 2; Supplementary information S4 (.txt)): the C_2H_2 zinc-finger (675 TFs), homeodomain (257 TFs) and helix–loop–helix (87 TFs). These results agree with previous studies in the mouse, in which the same three families account for the majority of TFs[22]. We note that C_2H_2 zinc fingers, and some other domains, interact with both DNA and RNA; however, we have included them here as we are currently unable to distinguish proteins that interact with one or the other — further experimental work will be necessary before this is possible.

Tissue-dependent TF expression

We examined the tissue-dependent expression of TFs using the Genome Novartis Foundation SymAtlas data set[42], which contains measurements of transcript levels for 79 human tissues, tumour samples and cell lines, obtained using the Affymetrix GeneChip HG-U133A. We reprocessed the raw data and set robust and objective expression thresholds to define the presence or absence of genes in the biological sample (Supplementary information S1 (PDF)). We excluded tumour samples and cell lines from further analysis, and focused on 32 healthy major tissues and organs. Of course, TF activity often depends on post-translational events, and gene expression does not necessarily indicate regulatory activity. However, it is still useful to assess the extent of TF expression as it provides the first line of evidence for the locations in which they may function.

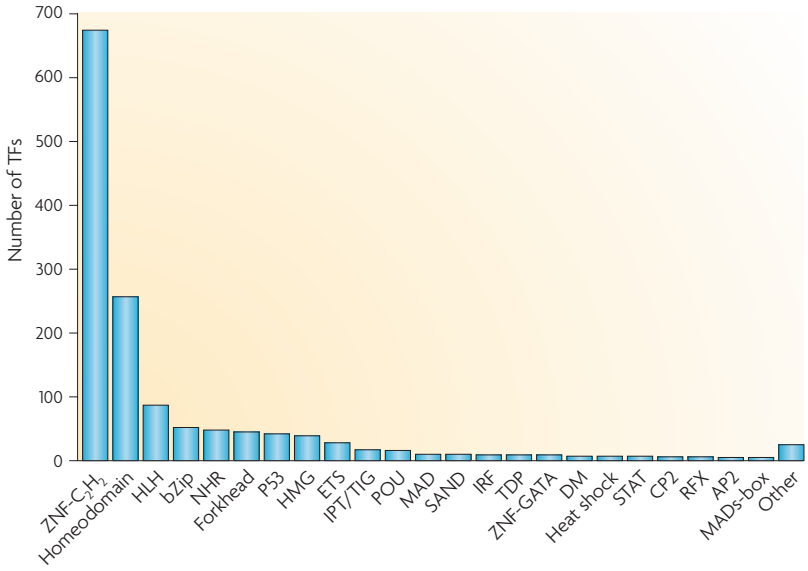

Figure 2 | **Transcription factors classified by DNA-binding domain.** Transcription factors (TFs) were classified into families according to their DNA-binding domain composition. InterPro parent–child relationships between DNA-binding domains were used as the basis for TF family definition (Supplementary information S1 (PDF)). TFs with multiple DNA-binding domains were classified in each of their respective families. Families with less than five members were classified as 'other'.

Levels of TF expression. FIGURE 3a shows the average expression of probe sets mapping to the 873 TF and 10,922 non-TF genes represented on Affymetrix GeneChips across the 32 human samples examined. The plot confirms observations from previous molecular studies, that is, TFs tend to be expressed at lower levels than non-TF genes ($p < 10^{-16}$; t-test)[43]. Mechanistically this makes sense: the effect of a single TF molecule is amplified by transcribing many copies of mRNA from a target gene. Moreover, it is easy to trigger a regulatory event by altering TF concentrations or activity if their expression levels are kept low. Finally, cells need to ensure that TFs recognize the correct target sites in the genome. Maintaining lower expression levels would allow TFs to bind the highest affinity sites, and keep lower affinity sites free for activation under special conditions, or non-functional sites free from undesired binding[44].

TF expression patterns. Of the 873 TFs that are represented on the array, 510 are expressed in at least one tissue. The rest do not rise above the threshold for detection, which means that either they are not present or the arrays are not sensitive enough to detect them.

The number of expressed TFs varies greatly between tissue types, ranging from approximately 150 in the appendix, skeletal muscle and skin, to over 300 in the whole brain, thyroid and placenta (FIG. 3b). The proportion of TFs relative to all expressed genes, however, is remarkably stable at ~6% across all samples. Two related factors could account for the variation between tissues in the number of expressed TFs. First, tissues contain multiple cell types and the number of TFs will rise with increasing varieties of cells in the sample; for example, the thyroid expresses a greater number of TFs than the liver, as the liver has a more homogeneous composition consisting mainly of hepatocytes. Second, some cells need more genes to function normally, and the number of expressed TFs might vary in line with the corresponding regulatory requirements. Intuitively, it would be attractive to propose that complex or metabolically active tissues, such as the brain, utilize more TFs than simple tissues, such as the appendix; however, it is difficult to provide a firm conclusion for this observation as we do not know the precise origins of the tissue samples or the way in which they were obtained. Further work using higher-resolution data — generated through new techniques such as transcriptomic sequencing (RNA-Seq)[45,46] — should shed more light on this matter.

The heat map in FIG. 4 displays the pattern of TF expression across the 32 major tissues examined. We calculated propensity values as a measure of tissue-specific expression for each TF (Supplementary information S1 (PDF)). This groups TFs into two categories: 161 TFs that are present in all or most tissues with similar expression levels (ubiquitous TFs); and 349 TFs that are selectively expressed in a few tissues (specific TFs). The ubiquitous category includes familiar TFs, such as the circadian regulator CLOCK, the oncogenic and growth-factor-activated Kruppel family member GLI2, and T-box 1 (TBX1). Though many of these entries are annotated with very specific and localized regulatory functions, their broad expression profiles suggest participation in a much wider range of processes. For example, TBX1, although primarily known as a developmental factor[47], also continues to be expressed in the adult organism.

The 349 specifically expressed TFs are interesting as they are involved in defining the precise nature of individual tissues. 123 of these factors display distinct expression levels in one tissue compared with all other samples and can be considered as potential markers; these include TFs that are expressed only in one tissue as well as those that are expressed widely, but with significantly elevated expression in a single tissue. Examples include the testis zinc-finger protein ZBTB32 (REF. 48) or the heart-specific NKX2-1 transcription factor[49]. Furthermore, 226 tissue-specific TFs display shared specificity among groups of related tissues. In general, there is substantial overlap in TF expression between the developing and adult stages of the same tissues. For example, fetal and adult lungs both express 14 lung-specific TFs — including the epithelial PAS domain protein 1 (EPAS1), which is thought to upregulate hypoxia-induced genes[50]. Adult and embryonic thyroids share seven thyroid-specific regulators; these include NK-homeobox 1 (NKX2-1), which activates genes that are essential for maintaining the differentiated cellular phenotype[51].

There is also shared specificity among adult tissues with similar physiological function and cellular composition. We observe that components of the central nervous system (whole brain, spinal cord and fetal brain) have seven specific TFs in common, including the thyroid hormone receptor alpha (THRA) and aryl-hydrocarbon receptor nuclear translocator 2 (ARNT2)[52]. Each of these tissues also utilizes unique TFs: for example, the fetal

RNA-Seq
The use of high-throughput sequencing techniques for transcriptomic profiling.

Propensity values
A measure of tissue specificity that normalizes the expression value of a TF across all samples, and the expression of all TFs in a single sample. It is commonly used to measure the distribution of amino acids types in different features of protein structures.

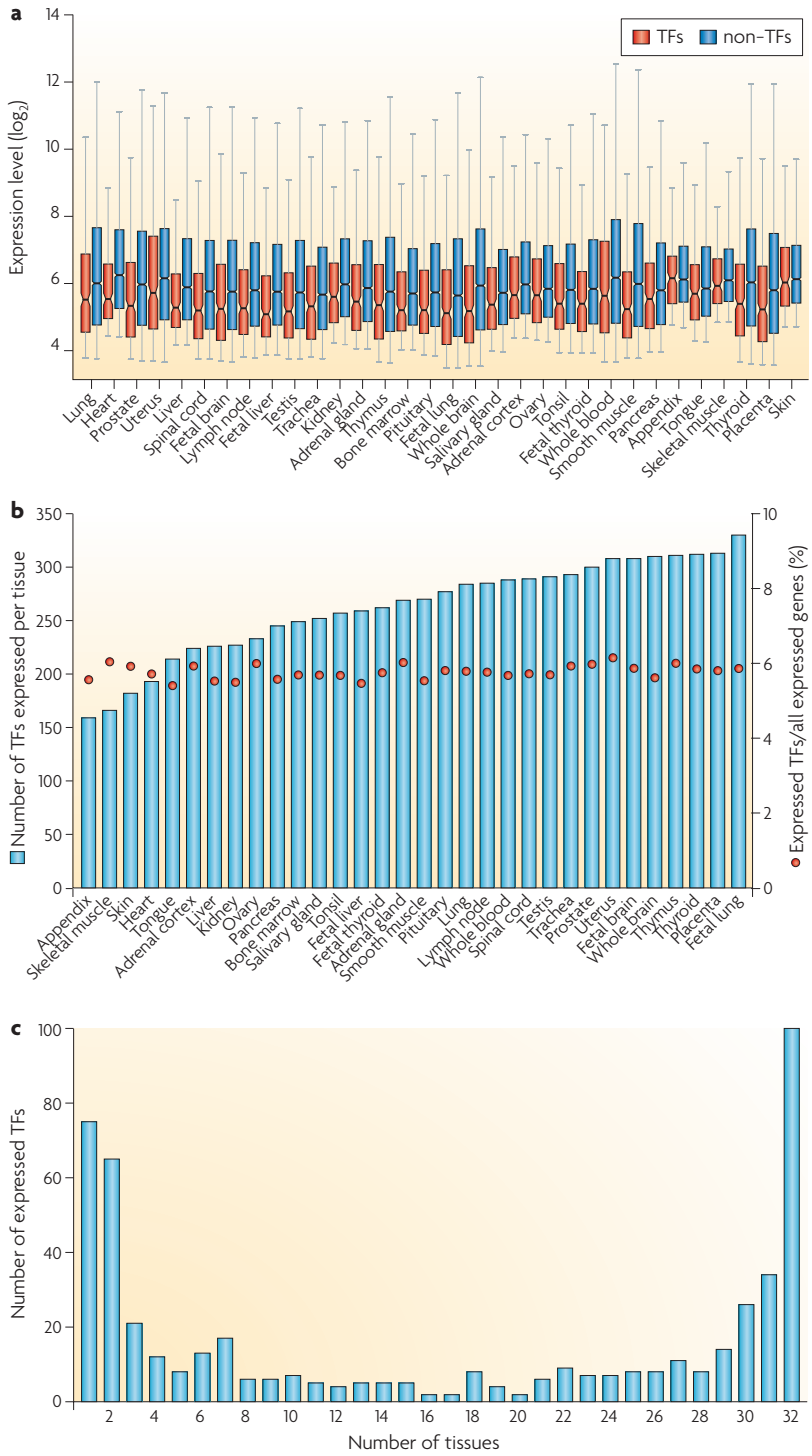

Figure 3 | **Expression level of transcription factors in 32 human organs and tissues.** **a** | Distribution of gene expression levels for transcription factors (TFs) (red) and non-TFs (blue) in different tissue types, shown as a box plot. In all samples, non-TF genes have higher average expression than TFs. **b** | Numbers of TFs expressed in each sample (blue bars) and the proportion of expressed TFs versus all expressed genes, given as a percentage (red points). The numbers of expressed regulators vary widely, ranging from about 150 in the appendix to over 300 in the fetal lung. However, in all tissues, TFs constitute ~6% of expressed genes. **c** | Number of tissues in which transcription factors are expressed (see also Supplementary information S1 (PDF)). Regulators are either expressed generally (30–32 samples) or specifically (1–3 samples). The U-shaped distribution in which factors are expressed in most tissues or in few tissues is robust against outliers (data not shown).

brain contains the N-myc proto-oncogene (MYCN), the inactivation of which impairs control of cell proliferation, differentiation and nuclear size in neuronal progenitors[53].

Finally, apparent multiple-tissue specificity could result from cross-contamination during sample preparation. This is most evident in the case of whole blood, with TFs found in the blood also being detected in seemingly unrelated organs containing many blood vessels (such as lungs), and related tissues with a high concentration of white blood cells (such as tonsil, thymus, lymph node and bone marrow tissue).

Unannotated TFs. The expression data for the 172 completely unannotated TFs is of particular interest, as their expression patterns provide preliminary insights into their potential regulatory functions. There are 69 unannotated TFs that fall in the category of general expression. For example, ZNF444 or FOXJ2 are highly and ubiquitously expressed, and this pattern could indicate an important function. 103 of the unannotated TFs have a tissue-specific expression, suggesting that they control processes that are characteristic of individual tissues. For example, the expression of ZNF337 in the fetal brain suggests that it might be involved in brain development, and therefore is an excellent candidate for further investigation.

Combinatorial usage. Another aspect of TFs that requires further research is their combinatorial usage, which allows great precision and flexibility in dictating the transcriptional programme of different tissues[13]. The two-tier system of general and specific TFs (FIGURE 3c), suggests different potential regulatory scenarios. Ubiquitous TFs alone — in isolation or in combination with each other — might control the general cellular machinery, and combinations of specific TFs might regulate tissue-specific genes. Alternatively, and we expect this to occur most commonly, ubiquitous TFs might serve as a platform to regulate a broad set of genes, which are then fine-tuned by specific regulators.

An interesting example is the serum response factor (SRF): it is a ubiquitously expressed TF that is involved in controlling multiple processes, including cell proliferation and differentiation[54–57]. SRF activity is modulated at several levels, including by the Rho family GTPase signalling pathway[58] as well as by interactions with other proteins[59,60]. A recent ChIP–chip study has shown that SRF binds to a large number of locations[61], acting as a global regulator, and thus particular target genes are likely to be activated when SRF combines with other specifically expressed TFs[62]. An illustration of this is SRF function in the determination of the smooth muscle phenotype, which is dependent on the interaction of SRF with the TFs NFAT and HERP1 (REFS 63,64), or its role in prostate differentiation and function, which is dependent on its interaction with NKX3-1 (REF. 65).

Cooperativity between TFs is known to involve extensive protein–protein interactions, both within families of homomeric and heteromeric TFs[66] and between structurally unrelated TFs[67]. Such interactions are not included in this Analysis article, but their incorporation

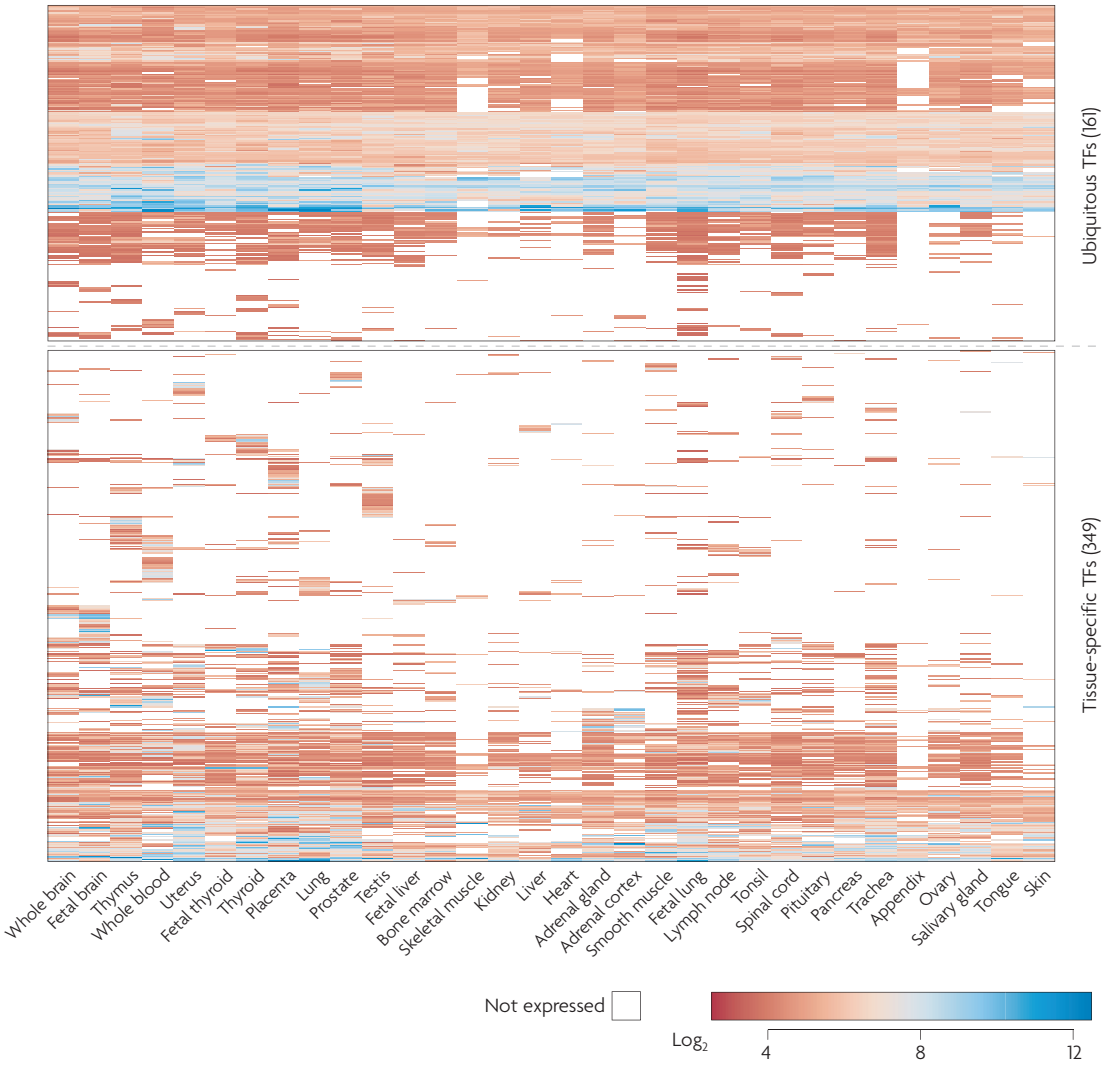

Figure 4 | **Heat map representation of transcription factor expression in 32 human organs and tissues.** Heat map of transcription factor (TF) expression (rows) in 32 organs and tissues (columns). Intersecting cells are shaded according to expression level (dark red for low expression and blue for high expression). Ubiquitous and specific TFs are grouped according to their expression profiles using hierarchical clustering (before setting an expression level threshold). Ubiquitous regulators are expressed at similar levels across most tissues, whereas specific regulators are expressed at significantly different levels in certain tissues (Supplementary information S1 (PDF)). Expression levels below the threshold of detection are depicted as white cells.

in future studies will help to elucidate patterns of combinatorial regulation and ultimately the regulatory functions of these TFs.

Evolutionary history of human TFs

The function and genomic organization of genes are intimately connected with how they have evolved[12]. Therefore, in order to gain an insight into the highly structured and coordinated regulatory functions of TFs, we studied several aspects of their evolution.

Evolution of the TF repertoire. First we studied the history of the TF repertoire, using phylogenetic relationships provided by the Ensembl Compara database (version 51). FIGURE 5a shows the evolutionary history of the 1,391 TF genes and their orthologues for which data are available, across 24 eukaryotic genomes ranging from yeast to chimpanzee.

There are five groups of TFs with distinct patterns of conservation: those that are present only in primates; predominantly in mammals, vertebrates or metazoa; and finally in most eukaryotes including yeast (Supplementary information S1 (PDF)). These groups appeared through periodic expansions in the TF repertoire along the human lineage — the proliferation of new regulatory genes coincided with the emergence of increasing organismal complexity, and they enabled organisms to develop new functionalities. For instance, the homeodomain family of TFs appeared during the emergence of a body plan in animals[68], and the Hox proteins — a sub-group of regulators in this family — have a central role in controlling segmental patterning during development. In another

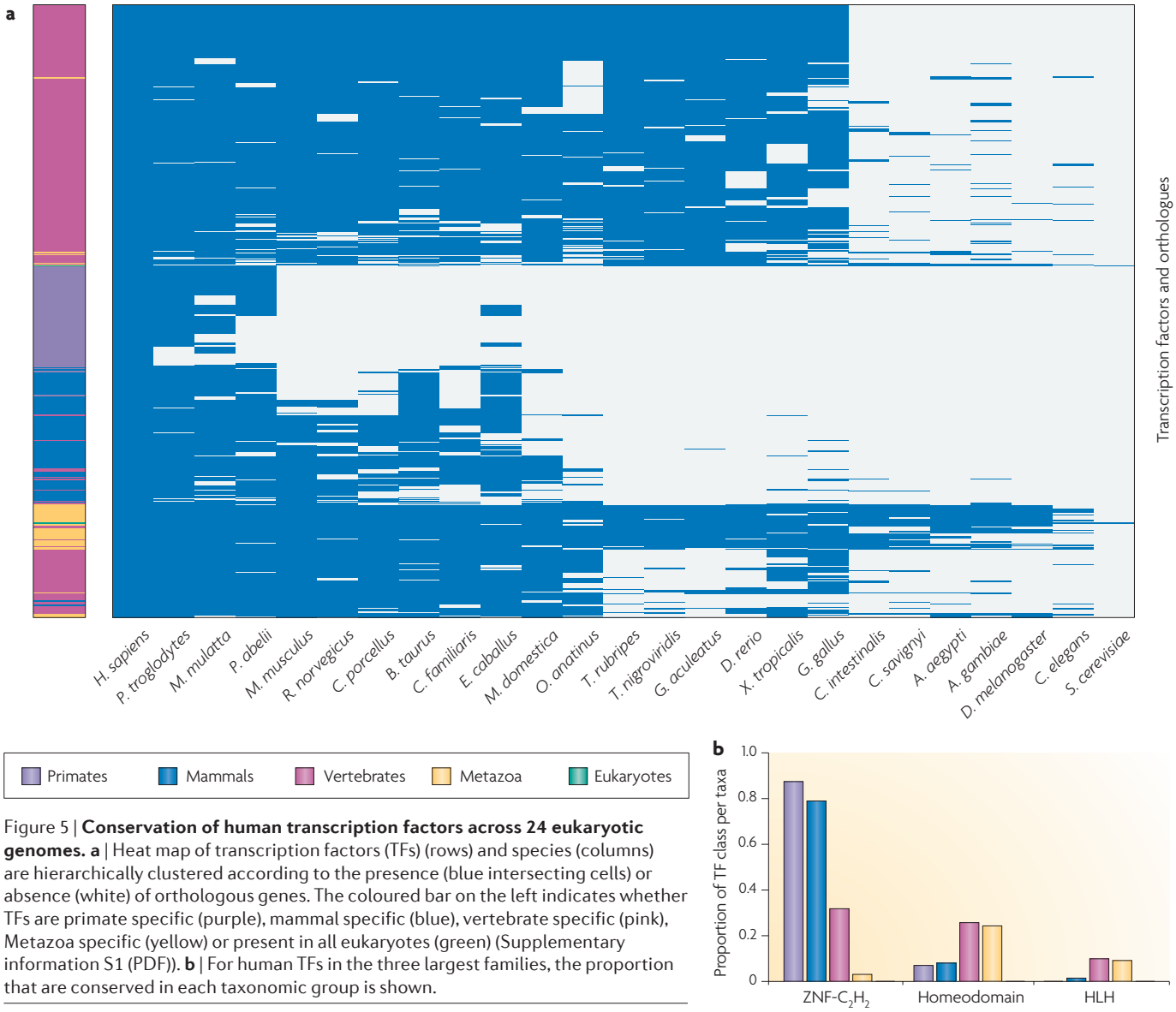

Figure 5 | **Conservation of human transcription factors across 24 eukaryotic genomes. a** | Heat map of transcription factors (TFs) (rows) and species (columns) are hierarchically clustered according to the presence (blue intersecting cells) or absence (white) of orthologous genes. The coloured bar on the left indicates whether TFs are primate specific (purple), mammal specific (blue), vertebrate specific (pink), Metazoa specific (yellow) or present in all eukaryotes (green) (Supplementary information S1 (PDF)). **b** | For human TFs in the three largest families, the proportion that are conserved in each taxonomic group is shown.

example, the large group of primate-specific regulators suggests that TF expansions continued until recently in human evolution. This point is stressed by the fact that 13% of the human TF repertoire appeared in primates, whereas only 2% of metabolic enzymes originate from this period (data not shown).

We found that the expansions occurred unevenly for TFs containing different types of DNA-binding domains (FIG. 5b). As mentioned above, homeodomain TFs first appeared in metazoan organisms and expanded rapidly in vertebrates, whereas helix–loop–helix TFs originated in metazoan organisms and have not expanded significantly since. The C_2H_2 zinc-finger family grew at several evolutionary stages, including with the appearance of vertebrates, and most substantially during the emergence of mammals and primates. Other domains, such as the CCAAT factor domain, are present in all eukaryotes, and have not increased substantially since their appearance. These TF expansions are interesting as they might have provided evolution with a way of modifying or creating different expression patterns for TFs, such as tissue-specific ones, by duplication followed by promoter divergence. This would explain, for example, why TFs such as ETS1 and ETS2 share a functional redundancy, but are expressed in different parts of the body[69,70]. In this Analysis article we do not consider the expansions that occurred in non-human lineages, but different TF families are known to have proliferated in other organisms (such as the nuclear hormone receptors in worms[20]).

Previous work suggested that the size of TF families is influenced in part by the number of different DNA sequences that they are able to recognize[71,72]. In other words, DNA-binding domains that can diversify their collection of target sequences should occur in greater numbers in a genome. This could explain why C_2H_2 zinc-finger proteins — with their ability to mutate amino acid positions that directly interact with DNA bases, and their capacity to extend the length of

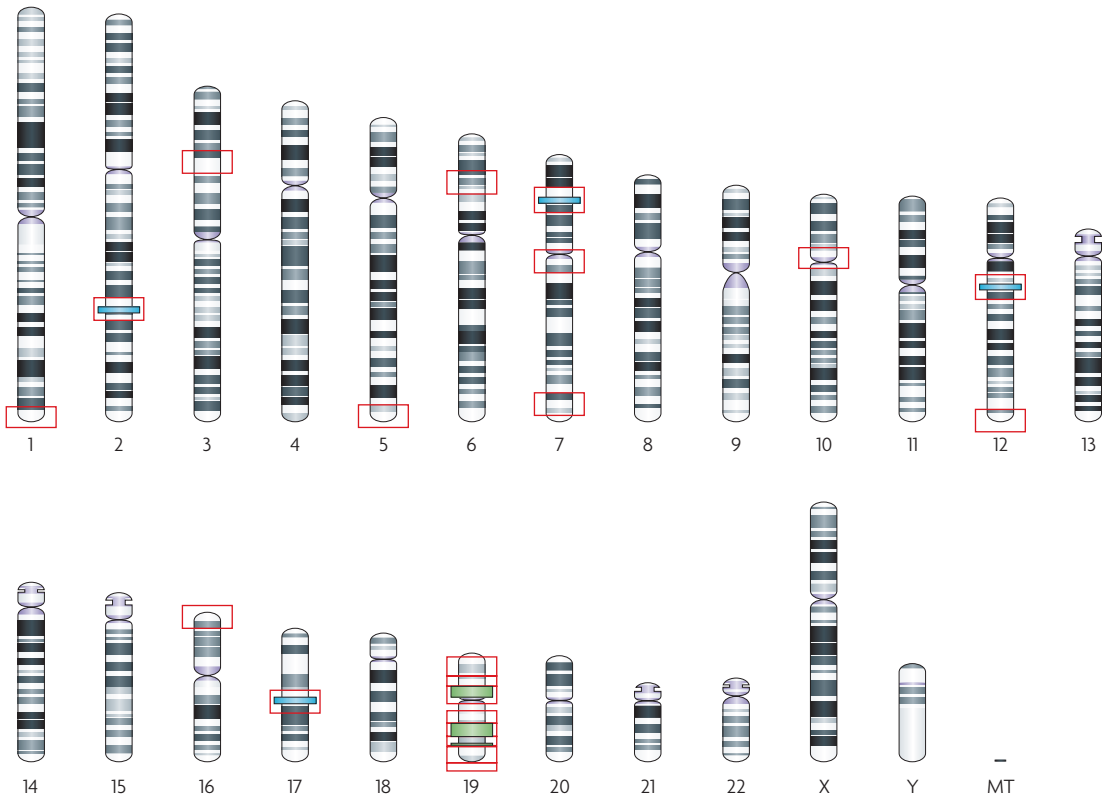

Figure 6 | **Locations of transcription factor clusters in the human genome.** There are 23 chromosomal loci that contain a high density of transcription factor (TF) genes (red boxes). The Hox clusters are present on chromosomes 2, 7, 12 and 17 (blue bars). Green bars represent previously described zinc-finger clusters on chromosome 19. MT, mitochondrial DNA.

the target-binding site by linking multiple domains in a sequential manner — constitute the largest group of TFs[73,74]. We also emphasize that the success of a particular protein family does not necessarily indicate the importance of individual TFs, and there are many TFs containing unusual DNA-binding domains that are central to human transcriptional regulation (for example, the interferon regulatory family[75]).

Chromosomal distribution of TFs. The local distribution of TF genes relative to each other is also of interest. Previous studies have described tandem clusters of paralogous genes, such as olfactory and immune receptors, that arose from large-scale intra-chromosomal duplications[76,77]. Similarly, for TFs, there are well-documented cases of biologically important clusters of Hox genes on chromosomes 2, 7, 12 and 17 (REF. 78) (FIG. 6).

To identify similar clusters, we searched for genomic regions that contain an unusually high proportion of TFs. We scanned each chromosome using a 500 kb sliding window and counted the number of TF and non-TF genes that occur in this region (Supplementary information S1 (PDF)). In total, there are 23 high-density clusters containing 284 TF genes (20% of all TFs), including the Hox regions and previously reported C_2H_2 zinc-finger-containing genes on chromosome 19 (REFS 79,80). Except for the Hox genes, all the clusters consist entirely of C_2H_2 zinc-fingers.

We can distinguish between two types of clusters based on the evolutionary history of the TF-coding genes that they contain. The first type, comprising 15 clusters, consists of a series of paralogues, suggesting that they arose through repeated tandem duplications from a founding locus. The Hox clusters belong to this category and their origins are documented elsewhere[68,81]. The C_2H_2 zinc-finger clusters on chromosome 19 arose from a similar process. These consist entirely of a particular class of TFs called KRAB–ZNF, which combine C-terminal DNA-binding zinc-fingers with an N-terminal Kruppel-associated box domain. Having undergone recent and rapid expansion, these TFs constitute the single largest family of regulators in the human genome (~400 genes), but their biological functions are largely unknown[82]. One of the clusters is primate specific, and seems to be undergoing further expansion: interspersed within the clusters and other genomic locations are several human specific paralogues of the gene *ZNF705A*[83]. A recent study showed that the KRAB–ZNF genes on chromosome 19 are extensively bound by proteins that promote heterochromatin formation[84]. The resulting packed chromosomal structure probably prevented recombination-mediated deletion of recently duplicated TFs, thereby contributing to the expansion of this family.

By contrast, the second type of cluster — comprising eight regions — does not consist of paralogues. These clusters reside in centromeres and telomeres, which

participate in intense genomic shuffling through recombination. We anticipate that the TF-coding genes in these clusters arose independently of each other at diverse locations of the genome, and relocated over time to form these clusters. In support of this view, there is a cluster on chromosome 16 containing nine genes; in the mouse genome, orthologues of these genes exist as two separate clusters on human chromosomes 16 and 17.

It is not obvious why so many TFs should co-localize in this way; but the fact that some clusters have been retained for hundreds of millions of years indicates a strong selective pressure to stabilize the spatial organization of these genes[68]. Hox genes can be used to explain one possible reason for cluster formation: in order to function properly, TFs within each Hox region must be expressed in a precisely coordinated manner, and the clustered arrangement allows them to be controlled by a common set of distal enhancers[85,86].

Using the SymAtlas data set we assessed whether any other gene clusters display similar patterns of co-expression. Although we observed the coordinated repression of members of some clusters, we did not find any clusters that were coordinately upregulated in specific tissues. This type of localized gene expression control has been reported previously[87], and one possible explanation for coordinated gene expression control is chromatin condensation, which limits TF and polymerase access to entire genomic regions. However, these types of observations are extremely sensitive to the quality of the underlying microarray data, and further analysis with additional data sets will be required before we can draw firm conclusions.

Evolution of new regulatory functions. Finally, having observed the conservation of TFs, we evaluated whether orthologues from different organisms retain equivalent regulatory functions. This is of interest as there is current discussion about the extent to which the evolution of regulatory proteins contributes to phenotypic differences between species[88]. In practical terms, this could also help us to identify the roles of human TFs, as it would be possible to infer functions from similar regulators in other organisms.

Previous studies of enzymes have shown that pairs of proteins with sequence similarities above 40% tend to share the same catalytic activity[89,90]. Related work has examined the relationship between the age of a gene and its biological function: older, more conserved genes usually take part in basic cellular functions, whereas newer, less conserved genes are associated with specialized or species-specific tasks. From this it follows that older genes should be broadly expressed as they are required throughout the organism, and tissue specificity should increase with decreasing age of genes[91]. The gene expression data did not demonstrate this trend as we found that TFs of all ages were both ubiquitously and specifically expressed.

A recent paper suggested that TFs that control developmental processes tend to be older than other types of regulators[9]. We observed several examples of proteins that control fundamental cellular processes that follow this pattern. For example, cell division cycle 5-like (*CDC5L*) encodes a ubiquitously expressed cell cycle regulator, with a function that is conserved from yeast to humans[92]. However, conservation of function is difficult to assess across the entire repertoire, as so few TFs have known functions.

The idea that phenotypic differences between organisms — particularly closely related ones such as humans and chimpanzees — may arise from changes in gene regulation, as well as from alterations to protein-coding genes, is not new[93,94]. Support for this opinion has increased recently, as comparative surveys of primate genomes have failed to reveal dramatic differences among genes that mediate the most distinguishing traits, such as cognitive, behavioural and dietary processes[7,95].

In support of this view, several studies have shown that TF-coding genes tend to be under greater positive evolutionary selection compared with other genes[7,8]. A highly publicized example is forkhead box 2 (FOXP2), a TF that is relevant to language development in humans[96]. Although the gene is one of the most conserved among mammals, it contains two amino acid changes that are present in the human gene but not in that of other primates, strongly suggesting that it was targeted for selection during recent human evolution[97]. In parallel, there is evidence for positive selection within the promoter sequences in the human genome — regions that are rich in TF-binding sequences — and that this has occurred primarily upstream of genes that are known for their involvement in neural function and nutrition[98]. In turn, these differences have had a direct impact on the DNA-binding activity of TFs. This was demonstrated in a ChIP–chip study of four highly conserved liver-specific regulators in humans and mice, in which 40% to 90% of binding sites differed between the two organisms[99]. At the level of gene expression, comparisons of primate transcriptomes showed that, whereas most genes have maintained similar profiles, a small subset of genes — particularly TFs — display significantly changed expression levels in the human lineage[100,101].

One of the striking observations from these studies is that apparently minor differences in the underlying nucleotide sequence can have profound effects on regulatory function. This effect might depend on where the nucleotide changes occur; for example, the mutations in *FOXP2* correspond to the DNA-binding and dimerization interfaces of the protein[102]. However, this case is in contrast to the observation that the liver-specific TFs in the ChIP–chip study mentioned above seem to function similarly despite the large differences in their binding locations. In fact, a large-scale rewiring of the transcriptional regulatory network in *E. coli*, which was achieved by adding new binding sites to promoters, showed that the bacteria is robust against most perturbations[103]. Therefore, the picture that emerges is one in which TF-coding genes and their target sites evolve quickly — probably faster for binding sites than for the TFs — but the full impact of these changes on regulatory function is yet to be understood.

TFs and human diseases

Transcriptional misregulation has been associated with a diverse set of diseases, including cancer and developmental syndromes[6,104,105]. We do not provide an extended analysis here, as it is covered in greater depth elsewhere[106,107]. Briefly, to evaluate the overall impact of TFs in human diseases, we examined the proportion of TF genes included within the OMIM database, which contains information of Mendelian-inherited monogenic diseases[108]. We identified 164 TFs (~12% of the total repertoire; $p = 0.018$ for association of TFs with diseases using a Fisher's exact test) that are directly responsible for 277 diseases or syndromes. Among these, a significant proportion is related to developmental defects, highlighting the importance of TFs during the early stages of development[6].

Misregulation of TF genes themselves also has important implications for more complex, or multigenic, diseases such as cancer or Parkinson's disease[109]. We studied TF expression levels across five leukaemia and lymphoma samples included in the SymAtlas data set, and found 25 TFs that were expressed exclusively in the disease samples but not in the healthy tissues that we discussed above (data not shown). However, it is not possible to determine whether the change in TF expression is directly responsible for the disease or is an indirect effect of other defects. This type of information will be revealed through further experiments and advances in the methods used to determine the transcriptional regulatory networks.

Conclusions

This Analysis article has presented a census of sequence-specific DNA-binding TFs in the human genome. Transcriptional regulation in higher organisms is of great current research interest; however, much of the work so far has focused on identifying binding sites through a combination of experimental methods, such as chromatin immunoprecipitation and computational predictions using motif searches. A major limitation to studying the regulators themselves has been the lack of a reliable data set of TFs in the human genome, a difficulty that is compounded by the large number of false predictions that can accompany the automated identification of genes encoding DNA-binding domains. By manually inspecting every entry in the data set, we have minimized such errors. As a result the repertoire here is of high quality, and can be used as a basis to design further computational and experimental studies.

Most of these TFs remain uncharacterized: this is highlighted by the observation that just three regulators (p53, ER and FOS) have more publications than all the other TFs put together. By incorporating publicly available genomic data sets, we provide clues into how these TFs might operate. For example, the SymAtlas microarray data set allowed us to investigate the expression of TFs across 32 major tissues.

Our analysis expanded the existing knowledge of ubiquitous and specific regulators by illustrating this phenomenon on a genome-wide scale. Intriguingly, few TFs are present in an intermediate number of tissues. This pattern of expression will have implications for our understanding of how TFs combine to exert their regulatory effect. By indicating where TFs are present we provide a starting point for future studies into the activity of individual TFs and how groups of TFs mediate biological processes of interest. Analysis of chromosomal clustering of TF genes and examination of their evolutionary histories also yield insights into how these regulators may function. In addition, we build on previous findings of disease-causing regulators: large numbers of TFs that we do not detect in normal tissues become highly expressed in diseased conditions.

Of course, the observations discussed here constitute only a first attempt to describe these regulators, and there are important caveats that should be considered when using the data. The SymAtlas data set is restricted by the fact that not all TFs are represented on the arrays used in this study; this means that we were unable to assess the expression of more than 500 TFs. Nonetheless, as the existing data covers two-thirds of TFs, we expect our findings to be robust. The repertoire itself also has weaknesses: at present, we include all genes encoding C_2H_2 zinc-fingers, as they are a major group of regulators; however, a major future challenge will be to distinguish between proteins that bind DNA from those that primarily complex with RNA. We will overcome many of these limitations as new data types become available. Indeed, transcriptomic data derived from high-throughput sequencing methods, such as RNA-Seq, will yield precise expression levels for every TF, as the method does not depend on array design. The quality of the repertoire itself will also continue to improve, as better annotations of the human genome are released and a greater range of descriptions of TF functions are published.

Finally, the data and observations presented here will be a valuable and reliable resource for a broad range of research into human transcriptional regulation. It would be possible to use the repertoire computationally to assess how the TFs themselves are regulated through transcriptional, post-transcriptional and post-translational mechanisms in order to transform incoming signals into a regulatory output. For example, TFs are thought to integrate numerous signals through the action of microRNAs, and it would be straightforward to carry out a preliminary survey to identify nucleic acids that are predicted to target transcripts encoding for TFs[110,111].

Experimentally, it will be possible to use the repertoire as a foundation for genome-scale assays to establish the precise DNA-binding specificities of these TFs, using technologies such as protein-binding arrays[112] and high-throughput SELEX[113]. *In vivo*, ChIP–chip[114] or ChIP-seq[115,116] will determine genomic binding locations under different cellular conditions. A combination of these approaches, along with detailed follow-up studies using traditional molecular approaches, will dramatically improve our understanding of the control underlying important biological processes. Ultimately, it might be possible to predict how the actions of different regulators lead to particular outcomes, and to apply these predictions in clinical settings, such as directing the progression of stem cell differentiation[117]. In conclusion, we are only just beginning to understand what TFs do in humans.

Fisher's exact test
A statistical test of independence between two categorical variables.

Protein-binding array
A high-throughput technique to determine DNA-binding affinities of proteins using microarrays displaying synthetic oligonucleotides.

ChIP-Seq
The combination of chromatin immunoprecipitation (ChIP) experiments with high-throughput sequencing techniques to quantitate protein targeting or chromatin modifications across the entire genome.

1. Simon, I. *et al.* Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **106**, 697–708 (2001).
2. Accili, D. & Arden, K. C. FoxOs at the crossroads of cellular metabolism, differentiation, and transformation. *Cell* **117**, 421–426 (2004).
3. Bain, G. *et al.* E2A proteins are required for proper B cell development and initiation of immunoglobulin gene rearrangements. *Cell* **79**, 885–892 (1994).
4. Dynlacht, B. D. Regulation of transcription by proteins that control the cell cycle. *Nature* **389**, 149–152 (1997).
5. Furney, S. J. *et al.* Structural and functional properties of genes involved in human cancer. *BMC Genomics* **7**, 3 (2006).
6. Boyadjiev, S. A. & Jabs, E. W. Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders. *Clin. Genet.* **57**, 253–266 (2000).
7. Bustamante, C. D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–1157 (2005).
Genome-wide study demonstrating that human TFs are under strong positive selection.
8. De, S., Lopez-Bigas, N. & Teichmann, S. A. Patterns of evolutionary constraints on genes in humans. *BMC Evol. Biol.* **8**, 275 (2008).
9. Lopez-Bigas, N., De, S. & Teichmann, S. A. Functional protein divergence in the evolution of *Homo sapiens*. *Genome Biol.* **9**, R33 (2008).
10. van Nimwegen, E. Scaling laws in the functional content of genomes. *Trends Genet.* **19**, 479–484 (2003).
11. Vogel, C. & Chothia, C. Protein family expansions and biological complexity. *PLoS Comput. Biol.* **2**, e48 (2006).
12. Kirschner, M. & Gerhart, J. Evolvability. *Proc. Natl Acad. Sci. USA* **95**, 8420–8427 (1998).
13. Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147–151 (2003).
A discussion of how progressively more elaborate transcriptional regulation has contributed to organismal complexity.
14. Lemon, B. & Tjian, R. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.* **14**, 2551–2569 (2000).
15. Wilson, D. *et al.* DBD — taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res.* **36**, D88–D92 (2008).
16. Perez-Rueda, E. & Collado-Vides, J. The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. *Nucleic Acids Res.* **28**, 1838–1847 (2000).
17. Moreno-Campuzano, S., Janga, S. C. & Perez-Rueda, E. Identification and analysis of DNA-binding transcription factors in *Bacillus subtilis* and other Firmicutes — a genomic approach. *BMC Genomics* **7**, 147 (2006).
18. Park, J. *et al.* FTFD: an informatics pipeline supporting phylogenomic analysis of fungal transcription factors. *Bioinformatics* **24**, 1024–1025 (2008).
19. Cherry, J. M. *et al.* SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.* **26**, 73–79 (1998).
20. Reece-Hoyes, J. S. *et al.* A compendium of *Caenorhabditis elegans* regulatory transcription factors: a resource for mapping transcription regulatory networks. *Genome Biol.* **6**, R110 (2005).
21. Adryan, B. & Teichmann, S. A. FlyTF: a systematic review of site-specific transcription factors in the fruit fly *Drosophila melanogaster*. *Bioinformatics* **22**, 1532–1533 (2006).
22. Gray, P. A. *et al.* Mouse brain organization revealed through direct genome-scale TF expression analysis. *Science* **306**, 2255–2257 (2004).
23. Riano-Pachon, D. M. *et al.* PlnTFDB: an integrative plant transcription factor database. *BMC Bioinformatics* **8**, 42 (2007).
24. Riechmann, J. L. *et al. Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science* **290**, 2105–2110 (2000).
The first genome-wide survey of TF repertoires for eukaryotic organisms.
25. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
26. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
27. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).
28. Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res.* **35**, D224–D228 (2008).
29. Messina, D. N. *et al.* An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res.* **14**, 2041–2047 (2004).
30. Roach, J. C. *et al.* Transcription factor expression in lipopolysaccharide-activated peripheral-blood-derived mononuclear cells. *Proc. Natl Acad. Sci. USA* **104**, 16245–16250 (2007).
31. Kersey, P. J. *et al.* The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4**, 1985–1988 (2004).
32. Hubbard, T. J. *et al.* Ensembl *Nucleic Acids Res.* **35**, D610–D617 (2007).
33. Wingender, E. *et al.* TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* **28**, 316–319 (2000).
34. Brunkow, M. E. *et al.* Disruption of a new forkhead/winged-helix protein, scurfin, results in the fatal lymphoproliferative disorder of the scurfy mouse. *Nature Genet.* **27**, 68–73 (2001).
35. Wildin, R. S. *et al.* X-linked neonatal diabetes mellitus, enteropathy and endocrinopathy syndrome is the human equivalent of mouse scurfy. *Nature Genet.* **27**, 18–20 (2001).
36. Hori, S., Nomura, T. & Sakaguchi, S. Control of regulatory T cell development by the transcription factor *Foxp3*. *Science* **299**, 1057–1061 (2003).
37. Fontenot, J. D., Gavin, M. A. & Rudensky, A. Y. Foxp3 programs the development and function of $CD4^+CD25^+$ regulatory T cells. *Nature Immunol.* **4**, 330–336 (2003).
38. Marson, A. *et al.* Foxp3 occupancy and regulation of key target genes during T-cell stimulation. *Nature* **445**, 931–935 (2007).
39. Zheng, Y. *et al.* Genome-wide analysis of Foxp3 target genes in developing and mature regulatory T cells. *Nature* **445**, 936–940 (2007).
40. Satoda, N. *et al.* Value of FOXP3 expression in peripheral blood as rejection marker after miniature swine lung transplantation. *J. Heart Lung Transplant.* **27**, 1293–1301 (2008).
41. Luscombe, N. M. *et al.* An overview of the structures of protein–DNA complexes. *Genome Biol.* **1**, REVIEWS001 (2000).
A review of DNA-binding domain structures and their interactions with DNA.
42. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA* **101**, 6062–6067 (2004).
This paper presents the microarray experiments underlying the SymAtlas gene expression data for human and mouse organs, tissues and cell lines.
43. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
44. Liu, X. & Clarke, N. D. Rationalization of gene regulation by a eukaryotic transcription factor: calculation of regulatory region occupancy from predicted binding affinities. *J. Mol. Biol.* **323**, 1–8 (2002).
45. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
46. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10**, 57–63 (2009).
A recent overview of the use of ultra-high-throughput sequencing technologies for measuring transcript levels.
47. Merscher, S. *et al. TBX1* is responsible for cardiovascular defects in velo-cardio-facial/DiGeorge syndrome. *Cell* **104**, 619–629 (2001).
48. Tang, C. J. *et al.* The zinc finger domain of Tzfp binds to the *tbs* motif located at the upstream flanking region of the *Aie1* (*aurora-C*) kinase gene. *J. Biol. Chem.* **276**, 19631–19639 (2001).
49. Pashmforoush, M. *et al.* Nkx2–5 pathways and congenital heart disease; loss of ventricular myocyte lineage specification leads to progressive cardiomyopathy and complete heart block. *Cell* **117**, 373–386 (2004).
50. Koizume, S. *et al.* Heterogeneity in binding and gene-expression regulation by HIF-2α. *Biochem. Biophys. Res. Commun.* **371**, 251–255 (2008).
51. Kimura, S. *et al.* The T/ebp null mouse: thyroid-specific enhancer-binding protein is essential for the organogenesis of the thyroid, lung, ventral forebrain, and pituitary. *Genes Dev.* **10**, 60–69 (1996).
52. Morte, B. *et al.* Deletion of the thyroid hormone receptor α1 prevents the structural alterations of the cerebellum induced by hypothyroidism. *Proc. Natl Acad. Sci. USA* **99**, 3985–3989 (2002).
53. Hirose, K. *et al.* cDNA cloning and tissue-specific expression of a novel basic helix-loop-helix/PAS factor (Arnt2) with close sequence similarity to the aryl hydrocarbon receptor nuclear translocator (Arnt). *Mol. Cell Biol.* **16**, 1706–1713 (1996).
54. Alberti, S. *et al.* Neuronal migration in the murine rostral migratory stream requires serum response factor. *Proc. Natl Acad. Sci. USA* **102**, 6148–6153 (2005).
55. Miano, J. M. *et al.* Restricted inactivation of serum response factor to the cardiovascular system. *Proc. Natl Acad. Sci. USA* **101**, 17132–17137 (2004).
56. Gauthier-Rouviere, C. *et al.* p67SRF is a constitutive nuclear protein implicated in the modulation of genes required throughout the G1 period. *Cell Regul.* **2**, 575–588 (1991).
57. Arsenian, S. *et al.* Serum response factor is essential for mesoderm formation during mouse embryogenesis. *EMBO J.* **17**, 6289–6299 (1998).
58. Hill, C. S., Wynne, J. & Treisman, R. The Rho family GTPases RhoA, Rac1, and CDC42Hs regulate transcriptional activation by SRF. *Cell* **81**, 1159–1170 (1995).
59. Treisman, R. Ternary complex factors: growth factor regulated transcriptional activators. *Curr. Opin. Genet. Dev.* **4**, 96–101 (1994).
60. Mo, Y. *et al.* Crystal structure of a ternary SAP-1/SRF/*c-fos* SRE DNA complex. *J. Mol. Biol.* **314**, 495–506 (2001).
61. Cooper, S. J. *et al.* Serum response factor binding sites differ in three human cell types. *Genome Res.* **17**, 136–144 (2007).
62. Gineitis, D. & Treisman, R. Differential usage of signal transduction pathways defines two types of serum response factor target gene. *J. Biol. Chem.* **276**, 24531–24539 (2001).
63. Gonzalez Bosc, L. V. *et al.* Nuclear factor of activated T cells and serum response factor cooperatively regulate the activity of an α-actin intronic enhancer. *J. Biol. Chem.* **280**, 26113–26120 (2005).
64. Doi, H. *et al.* HERP1 inhibits myocardin-induced vascular smooth muscle cell differentiation by interfering with SRF binding to CArG box. *Arterioscler. Thromb. Vasc. Biol.* **25**, 2328–2334 (2005).
65. Zhang, Y., Fillmore, R. A. & Zimmer, W. E. Structural and functional analysis of domains mediating interaction between the *bagpipe* homologue, Nkx3.1 and serum response factor. *Exp. Biol. Med. (Maywood)* **233**, 297–309 (2008).
66. Amoutzias, G. D. *et al.* One billion years of bZIP transcription factor evolution: conservation and change in dimerization and DNA-binding site specificity. *Mol. Biol. Evol.* **24**, 827–835 (2007).
67. Uht, R. M. *et al.* A conserved lysine in the estrogen receptor DNA binding domain regulates ligand activation profiles at AP-1 sites, possibly by controlling interactions with a modulating repressor. *Nucl. Recept.* **2**, 2 (2004).
68. Garcia-Fernandez, J. The genesis and evolution of homeobox gene clusters. *Nature Rev. Genet.* **6**, 881–892 (2005).
69. De la Houssaye, G. *et al.* ETS-1 and ETS-2 are upregulated in a transgenic mouse model of pigmented ocular neoplasm. *Mol. Vis.* **14**, 1912–1928 (2008).
70. Albagli, O. *et al.* A model for gene evolution of the ets-1/ets-2 transcription factors based on structural and functional homologies. *Oncogene* **9**, 3259–3271 (1994).
71. Itzkovitz, S., Tlusty, T. & Alon, U. Coding limits on the number of transcription factors. *BMC Genomics* **7**, 239 (2006).
72. Luscombe, N. M. & Thornton, J. M. Protein–DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol.* **320**, 991–1009 (2002).
73. Pavletich, N. P. & Pabo, C. O. Zinc finger–DNA recognition: crystal structure of a Zif268–DNA complex at 2.1 A. *Science* **252**, 809–817 (1991).
74. Nardelli, J. *et al.* Base sequence discrimination by zinc-finger DNA-binding domains. *Nature* **349**, 175–178 (1991).
75. Honda, K. & Taniguchi, T. IRFs: master regulators of signalling by Toll-like receptors and cytosolic pattern-recognition receptors. *Nature Rev. Immunol.* **6**, 644–658 (2006).
76. Bulger, M. *et al.* Conservation of sequence and structure flanking the mouse and human β-globin loci: the β-globin genes are embedded within an array of odorant receptor genes. *Proc. Natl Acad. Sci. USA* **96**, 5129–5134 (1999).

77. Ben-Arie, N. *et al.* Olfactory receptor gene cluster on human chromosome 17: possible duplication of an ancestral receptor repertoire. *Hum. Mol. Genet.* **3**, 229–235 (1994).
78. Scott, M. P. Vertebrate homeobox gene nomenclature. *Cell* **71**, 551–553 (1992).
79. Dehal, P. *et al.* Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* **293**, 104–111 (2001).
80. Grimwood, J. *et al.* The DNA sequence and biology of human chromosome 19. *Nature* **428**, 529–535 (2004).
81. Abbasi, A. A. & Grzeschik, K. H. An insight into the phylogenetic history of HOX linked gene families in vertebrates. *BMC Evol. Biol.* **7**, 239 (2007).
82. Looman, C. *et al.* KRAB zinc finger proteins: an analysis of the molecular mechanisms governing their increase in numbers and complexity during evolution. *Mol. Biol. Evol.* **19**, 2118–2130 (2002).
83. Huntley, S. *et al.* A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. *Genome Res.* **16**, 669–677 (2006).
84. Vogel, M. J. *et al.* Human heterochromatin proteins form large domains containing KRAB-ZNF genes. *Genome Res.* **16**, 1493–1504 (2006).
85. Kikuta, H. *et al.* Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* **17**, 545–555 (2007).
86. Lee, A. P. *et al.* Highly conserved syntenic blocks at the vertebrate Hox loci and conserved regulatory elements within and outside Hox gene clusters. *Proc. Natl Acad. Sci. USA* **103**, 6994–6999 (2006).
87. Kim, S. K. *et al.* A gene expression map for *Caenorhabditis elegans*. *Science* **293**, 2087–2092 (2001).
88. Arendt, D. The evolution of cell types in animals: emerging principles from molecular studies. *Nature Rev. Genet.* **9**, 868–882 (2008).
89. Wilson, C. A., Kreychman, J. & Gerstein, M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297**, 233–249 (2000).
90. Todd, A. E., Orengo, C. A. & Thornton, J. M. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143 (2001).
91. Freilich, S. *et al.* Relationship between the tissue-specificity of mouse gene expression and the evolutionary origin and function of the proteins. *Genome Biol.* **6**, R56 (2005).
92. Hirayama, T. & Shinozaki, K. A cdc5$^+$ homolog of a higher plant, *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **93**, 13371–13376 (1996).
93. Monod, J. & Jacob, F. Teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harb. Symp. Quant. Biol.* **26**, 389–401 (1961).
94. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
95. Nielsen, R. *et al.* A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3**, e170 (2005).
96. Lai, C. S. *et al.* A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* **413**, 519–523 (2001).
97. Enard, W. *et al.* Molecular evolution of *FOXP2*, a gene involved in speech and language. *Nature* **418**, 869–872 (2002).
98. Haygood, R. *et al.* Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nature Genet.* **39**, 1140–1144 (2007).
99. Odom, D. T. *et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature Genet.* **39**, 730–732 (2007).
A ChIP–chip study showing that functionally equivalent TFs bind to different sites in the human and mouse genomes.
100. Khaitovich, P. *et al.* Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **309**, 1850–1854 (2005).
101. Gilad, Y. *et al.* Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* **440**, 242–245 (2006).
A microarray study suggesting that changes in gene expression levels might be important in distinguishing between different primate species.
102. Stroud, J. C. *et al.* Structure of the forkhead domain of FOXP2 bound to DNA. *Structure* **14**, 159–166 (2006).
103. Isalan, M. *et al.* Evolvability and hierarchy in rewired bacterial gene networks. *Nature* **452**, 840–845 (2008).
104. Lopez-Bigas, N., Blencowe, B. J. & Ouzounis, C. A. Highly consistent patterns for inherited human diseases at the molecular level. *Bioinformatics* **22**, 269–277 (2006).
105. Darnell, J. E. Transcription factors as targets for cancer therapy. *Nature Rev. Cancer* **2**, 740–749 (2002).
106. Engelkamp, D. & van Heyningen, V. Transcription factors in disease. *Curr. Opin. Genet. Dev.* **6**, 334–342 (1996).
107. Jimenez-Sanchez, G., Childs, B. & Valle, D. Human disease genes. *Nature* **409**, 853–855 (2001).
108. Wheeler, D. L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **35**, D5–D12 (2007).
109. Scherzer, C. R. *et al.* GATA transcription factors directly regulate the Parkinson's disease-linked gene α-synuclein. *Proc. Natl Acad. Sci. USA* **105**, 10907–10912 (2008).
110. Sinha, S. *et al.* Systematic functional characterization of *cis*-regulatory motifs in human core promoters. *Genome Res.* **18**, 477–488 (2008).
111. Martinez, N. J. *et al.* A *C. elegans* genome-scale microRNA network contains composite feedback motifs with high flux capacity. *Genes Dev.* **22**, 2535–2549 (2008).
112. Berger, M. F. *et al.* Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**, 1266–1276 (2008).
A high-throughput assay of DNA-binding specificities for 168 mouse TFs using protein-binding microarrays.
113. Hallikas, O. *et al.* Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**, 47–59 (2006).
A study combining SELEX and motif-finding methods to identify the DNA-binding specificities and target regions for five mammalian TFs.
114. Horak, C. E. *et al.* GATA-1 binding sites mapped in the β-globin locus by using mammalian chIp–chip analysis. *Proc. Natl Acad. Sci. USA* **99**, 2924–2929 (2002).
115. Johnson, D. S. *et al.* Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* **316**, 1497–1502 (2007).
116. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* **4**, 651–657 (2007).
Refs 115 and 116 are the first uses of chromatin immunoprecipitation and ultra-high-throughput sequencing to determine genome-wide binding sites of mammalian TFs
117. Gaspard, N. *et al.* An intrinsic mechanism of corticogenesis from embryonic stem cells. *Nature* **455**, 351–357 (2008).

Acknowledgements

The authors thank J. Herrero and A. Vilella for assistance using Ensembl Compara. J.M.V. thanks the Spanish Ministry of Education and Science, the European Science Foundation Exchange Grant and the Marie Curie Biostar programmes for funding. N.M.L. acknowledges funding from EMBL.

FURTHER INFORMATION

Human TF repertoire: http://www.valleyofpigs.org/humantfs
Luscombe laboratory: http://www.ebi.ac.uk/luscombe
Teichmann laboratory: http://www2.mrc-lmb.cam.ac.uk/SS/Teichmann_S
DBD: http://www.transcriptionfactor.org
EdgeDB: http://edgedb.umassmed.edu
Ensembl Compara database: http://nov2008.archive.ensembl.org/info/docs/compara/index.html
Ensembl Genome Browser: http://www.ensembl.org
FlyTF: http://www.flytf.org
FTFD: http://ftfd.snu.ac.kr
Gene Ontology Home: http://www.geneontology.org
InterPro: http://www.ebi.ac.uk/interpro
IPI: http://www.ebi.ac.uk/ipi
JASPAR: http://jaspar.genereg.net
OMIM: http://www.ncbi.nlm.nih.gov/omim
Pfam: http://pfam.sanger.ac.uk
PlnTFDB: http://plntfdb.bio.uni-potsdam.de
PubMed: http://www.pubmed.org
RegulonDB: http://regulondb.ccg.unam.mx
SGD: http://www.yeastgenome.org
SUPERFAMILY: http://supfam.cs.bris.ac.uk/SUPERFAMILY
SymAtlas: http://symatlas.gnf.org/SymAtlas
TRANSFAC: http://www.gene-regulation.com

SUPPLEMENTARY INFORMATION

See online article: S1 (PDF) | S2 (.txt file) | S3 (.txt file) | S4 (.txt file)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF