Chapter 1

How Do You Find Transcription Factors? Computational Approaches to Compile and Annotate Repertoires of Regulators for Any Genome

Juan M. Vaquerizas, Sarah A. Teichmann, and Nicholas M. Luscombe

Abstract

Transcription factors (TFs) play an important role in regulating gene expression. The availability of complete genome sequences and associated functional genomic data offer excellent opportunities to understand the transcriptional regulatory system of an entire organism. To do so, however, it is essential to compile a reliable dataset of regulatory components. Here, we review computational methods and publicly accessible resources that help identify TF-coding genes in prokaryotic and eukaryotic genomes. Since the regulatory functions of most TFs remain unknown, we also discuss approaches for combining diverse genomic datasets that will help elucidate their chromosomal organisation, expression, and evolutionary conservation. These analysis methods provide a solid foundation for further investigations of the transcriptional regulatory system.

Key words: Transcription factor, Genomics, TF, Transcriptional regulation, Evolution, Gene expression, Genome organisation

1. Introduction

Transcriptional regulation is one of the most fundamental mechanisms for controlling the amount of protein produced by cells under different environmental conditions and developmental stages (1, 2). In eukaryotes, transcription can be modulated at many different levels, from the assembly of the core transcriptional machinery to the architecture and intranuclear localisation of chromosomes (3). A vast array of proteins, including RNA polymerases, histones, histone modifiers, transcription factors, and co-factors, are involved in maintaining the accuracy and specificity of the regulatory process.

Bart Deplancke and Nele Gheldof (eds.), *Gene Regulatory Networks: Methods and Protocols*,
Methods in Molecular Biology, vol. 786, DOI 10.1007/978-1-61779-292-2_1, © Springer Science+Business Media, LLC 2012

Among these, DNA-binding transcription factors (TFs) play a central role, as they are responsible for directing transcription initiation to specific gene promoters based on their sequence-recognition abilities (4). The importance of TFs is highlighted by the amount of research devoted to understanding how they function, ranging from the basis for DNA-sequence recognition to their regulatory function in particular cell types (5, 6).

The availability of fully sequenced genomes and the development of high-throughput experimental techniques over the past decade have expanded our capacity to explore regulatory systems on a whole-organism scale. Using these data, computational studies have characterised repertoires of TFs for organisms across all phylogenetic groups including bacteria, fungi, and plants (see Table 1). Further, integration of these repertoires with functional and evolutionary data has led to insights about basic organisational properties of these regulatory systems. For example, our recent analysis of the human TF repertoire revealed a two-tier organisation of tissue-specific and ubiquitous expression patterns, and a step-wise pattern of evolutionary conservation (7). Since the regulatory functions of many TFs remain unknown, these types of general analyses provide a good starting point for further detailed molecular and computational studies.

In this review, we introduce several online resources describing TF repertoires in different eukaryotic genomes. We also describe possible computational strategies for identifying such a repertoire and discuss approaches to annotate newly identified TF-coding genes with additional information, using our recent publication of the human TF repertoire as an example (7). It is worth noting here that we restrict our definition of TFs as proteins that recognise DNA in a sequence-specific manner, but neither display enzymatic activity nor form part of the core transcriptional machinery. The tutorial refers to our recently published study of human TFs for demonstration purposes (Fig. 1), but the analysis could be extended to any other organism with a sequenced genome.

2. Materials

2.1. Identification of the TF Repertoire

1. Genome sequences and associated annotations. The sequence, gene assembly, and protein-sequence annotation for the human genome, as well as many others, can be obtained from Ensembl (http://www.ensembl.org) (8). Programmatic access to the database is also available through an Application Programming Interface (API).

Table 1
Summary of available online resources for TF annotation

Resource	Organism	References	Link
Prokaryotic			
RegulonDB	*E. coli*	(66)	http://regulondb.ccg.unam.mx/
DBTBS	*B. subtilis*	(53)	http://dbtbs.hgc.jp/
Moreno-Campuzano et al.	*B. subtilis*	(56)	http://www.biomedcentral.com/1471-2164/7/147/additional/
CoryneRegNet	Corynebacteria	(51)	http://www.coryneregnet.de/
cTFbase	Cyanobacteria	(52)	http://cegwz.com/
PRODORIC	Bacteria	(63)	http://prodoric.tu-bs.de
RegTransBase	Bacteria	(65)	http://regtransbase.lbl.gov
ArchaeaTF	Archaea	(49)	http://bioinformatics.zj.cn/archaeatf/
BacTregulators	Prokaryotes	(50)	http://www.bactregulators.org/
GTOP_TF	Prokaryotes	(70)	http://spock.genes.nig.ac.jp/~gtop_tf/index2.html
Eukaryotic			
Vaquerizas et al.	*H. sapiens*	(7)	http://www.valleyofpigs.org/humantfs
Messina et al.	*H. sapiens*	(48)	http://genome.cshlp.org/content/14/10b/2041/suppl/DC1
TFcat	*H. sapiens/M. musculus*	(68)	http://www.tfcat.ca/
TFCONES	*H. sapiens/M. musculus/T. rubripes*	(69)	http://tfcones.fugu-sg.org
Gray et al.	*M. musculus*	(59)	http://www.sciencemag.org/cgi/content/full/306/5705/2255/DC1
TFdb	*M. musculus*	(46)	http://genome.gsc.riken.jp/TFdb/
FlyTF	*D. melanogaster*	(54)	http://flytf.org/
EDGEdb	*C. elegans*	(45)	http://edgedb.umassmed.edu/
RARTF	*A. thaliana*	(64)	http://rarge.gsc.riken.jp/rartf/
AtTFDB	*A. thaliana*	(47)	http://arabidopsis.med.ohio-state.edu/AtTFDB/
SoyDB	*G. max*	(67)	http://casp.rnet.missouri.edu/soydb/
TOBFAC	*N. tabacum*	(71)	http://compsysbio.achs.virginia.edu/tobfac/
wDBTF	*T. aestivum*	(74)	http://wwwappli.nantes.inra.fr:8180/wDBFT/
ITFP	Mammals	(57)	http://itfp.biosino.org/itfp/
FTFD	Fungi	(55)	http://ftfd.snu.ac.kr/
PlanTAPDB	Plants	(60)	http://www.cosmoss.org/bm/plantapdb/
PlantTFDB	Plants	(61)	http://planttfdb.cbi.pku.edu.cn/
PlnTFDB	Plants	(62)	http://plntfdb.bio.uni-potsdam.de/v2.0/
JASPAR	Eukaryotes	(58)	http://jaspar.cgb.ki.se/
TRANSFAC	Eukaryotes	(72)	http://www.gene-regulation.com/pub/databases.html/
TrSDB	Eukaryotes	(73)	http://bioinf.uab.es/cgi-bin/trsdb/trsdb.pl/
Cross-kingdom			
DBD	–	(13)	http://www.transcriptionfactor.org/

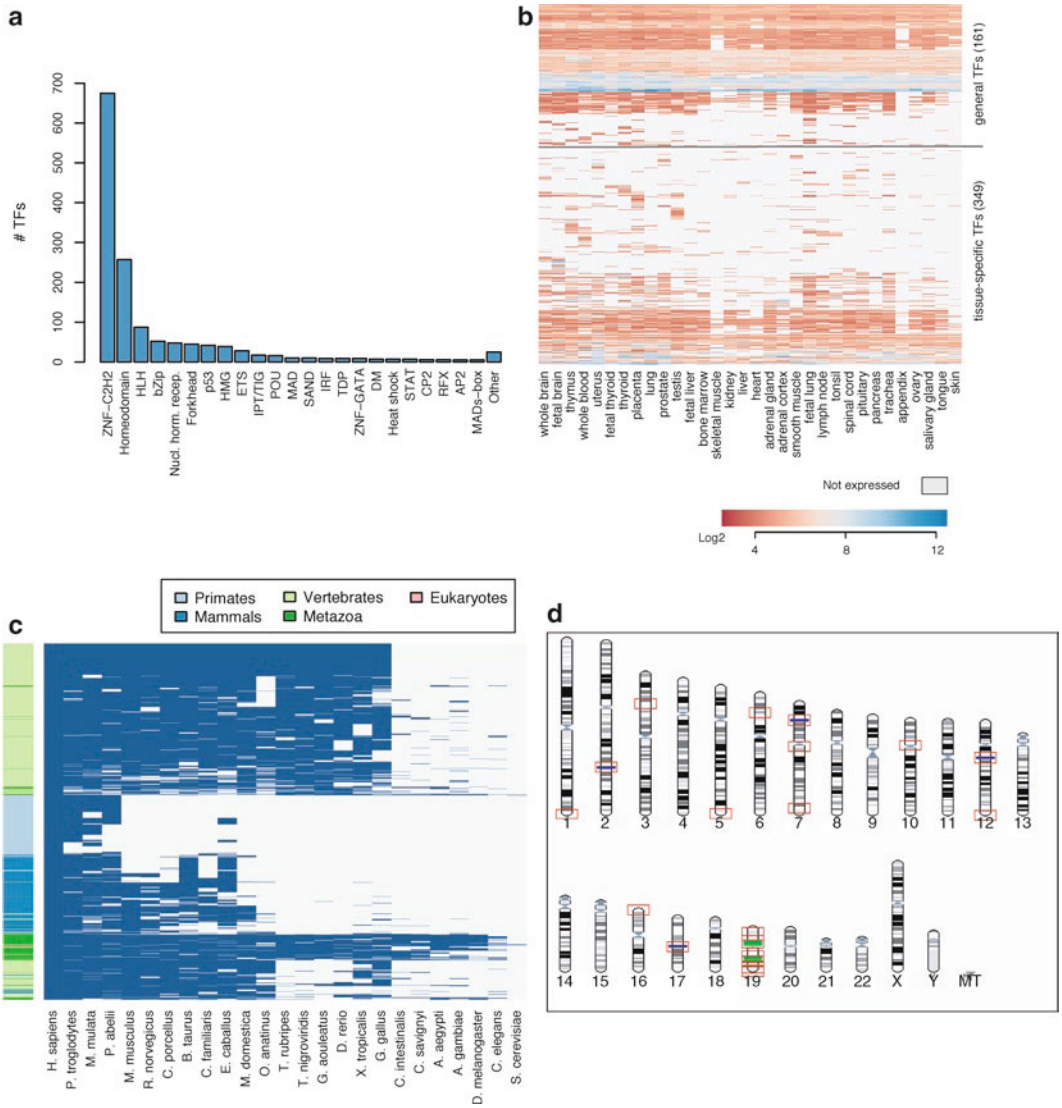

Fig. 1. Computational annotation of sequence-specific DNA binding transcriptional factors. (**a**) Human TFs classified according to their DNA-binding domain composition. Families with less than five members are classified as "other". (**b**) Heat-map representation of transcription factor expression (*rows*) across 32 human tissues and organs (*columns*). Cells are coloured according to the expression level (*dark blue* for high expression, *dark red* for low expression). General (ubiquitous) and tissue-specific TFs are grouped according to their expression values using hierarchical clustering. Expression levels below the threshold of detection are shown as white cells. (**c**) Evolutionary conservation of human transcription factors across 24 eukaryotic genomes. Transcription factors (*rows*) and species (*columns*) are clustered based on the presence (*blue cell*) or absence (*white cell*) of orthologous genes. The coloured bar on the left indicates the level of conservation for each TF depending whether these are primate-specific (*light blue*), mammalian-specific (*dark blue*), vertebrate-specific (*light green*), metazoa-specific (*dark green*) or present in all examined species (*pink*). (**d**) Chromosomal location of transcription factor clusters in the human genome. The positions of 23 chromosomal regions with high density of TFs are *highlighted with red boxes*. The Hox clusters in chromosomes 2, 7, 12, and 17 are depicted in *blue*. Zn-finger clusters in chromosome 19 are depicted in *light green*. First published in [Nature Reviews Genetics, 10(4), 2009, doi:10.1038/nrg2538. © Nature Publishing Group, a division of Macmillan Publishers Limited].

2. Protein sequences representing DNA-binding domains (DBDs). A high-confidence dataset of 347 InterPro hidden Markov models (InterPro release 18) (9) of DBDs can be obtained from (http://www.valleyofpigs.org/humantfs). Complete datasets of protein domains can also be obtained from Pfam (http://pfam.sanger.ac.uk) (10), SUPERFAMILY (http://supfam.cs.bris.ac.uk/SUPERFAMILY_1.73) (11), or PROSITE (http://www.expasy.ch/prosite) (12). The DBD database (http://www.transcriptionfactor.org) also provides hidden Markov models for DBDs from SUPERFAMILY and Pfam (13, 14).
3. Sequence search software. The InterProScan software can be downloaded from (http://www.ebi.ac.uk/Tools/InterProScan) (15).

2.2. Assessing the Coverage of the TF Repertoire

1. Gene-function annotations. Gene Ontology annotations for genes can be downloaded from (http://www.geneontology.org) (16).

2.3. Assigning Regulatory Functions to the TF Repertoire

1. GO functional enrichments. Enrichment of biological functions can be calculated using g:Profiler (http://biit.cs.ut.ee/gprofiler) (17).
2. Literature citations. Journal abstracts citing the genes above can be obtained from the PubMed database (http://www.ncbi.nlm.nih.gov/pubmed) (18). It is possible to query PubMed automatically, but please read the database documentation as uncontrolled querying may lead to restricted access.

2.4. Measuring the Expression of the TF Repertoire

1. Gene-expression data. Datasets can be downloaded from ArrayExpress or GEO. Here we will use the GNF SymAtlas dataset measuring expression across 79 human tissues and cell lines (http://biogps.gnf.org) (19). A GCRMA-normalised version of the dataset is available from ArrayExpress (20) with accession number E-TABM-145.
2. Bioinformatics analysis software. The R statistical software package and the BioConductor bioinformatics suite can be obtained from (http://www.r-project.org) and (http://www.bioconductor.org) (21).

2.5. Examining the Evolutionary Conservation of the TF Repertoire

1. Orthologue predictions. These can be obtained from the Ensembl Compara database (http://www.ensembl.org) for most major eukaryotic genomes (22).

2.6. Assessing the Genomic Organisation of the TF Repertoire

1. Genomic coordinates. Coordinates describing the genomic location of genes can be accessed in Ensembl (see Subheading 2.1).

3. Methods

3.1. Identification of the TF Repertoire

Potential TF-coding genes can be determined by several computational approaches. A common method is to use pair-wise sequence-alignment algorithms such as BLAST (23) to identify homologues of known TFs. A more sensitive approach is to search for genes containing known DBDs using profile-based methods such as InterProScan, HMMER, and PSI-BLAST (24). Resources like InterPro, Pfam, and SUPERFAMILY provide curated hidden Markov models describing the amino-acid sequences for groups of conserved protein-sequence regions and domains. Among these are models representing well-known DBDs for all kingdoms, which we have identified and listed in the dataset from Subheading 2.1.

For most well-annotated genomes, searches can be performed directly against a reference set of protein sequences that are provided by resources such as UniProt (http://www.uniprot.org) (25). For genomes that are still unannotated, searches may have to be performed on the underlying nucleotide sequence itself.

The HMM search will return a set of genes coding for potential TFs containing a DBD. Some DBDs and their sequence models, however, can be promiscuous and produce false-positive hits to non-TF proteins that nonetheless bind DNA. The HMM search will also miss TFs that lack a conventional DBD (26), so generating false negatives. Therefore, we recommend that users refine the dataset to exclude non-TF hits and include known non-standard TFs; this refinement will usually consist of a combination of literature-based curation and inspection of the domain organisation of the protein.

In our own analysis of the human genome, we obtained 1,960 initial hits from a search of 347 DBD models against all human protein sequences. We manually curated this dataset by filtering out 596 probable false positives and included 27 known TFs that were missed. The final high-confidence dataset contained 1,391 TF-coding genes (Fig. 1a) (7).

Below is a step-by-step protocol to identify the TF repertoire for a given genome, using human as an example:

1. Download a list of manually curated DBDs (http://www.valleyofpigs.org/humantfs). Universal prokaryote/eukaryote DBD lists can also be downloaded from DBD (http://dbd.mrc-lmb.cam.ac.uk/DBD/index.cgi?Download).
2. Download and install InterProScan (http://www.ebi.ac.uk/Tools/InterProScan).
3. Download a unique reference set of human proteins from UniProt (http://www.uniprot.org).

4. Run InterProScan on a set of human proteins with default parameters.
5. Examine InterProScan results and filter proteins with hits to a DBD.
6. (Optional) Manually curate the list of proteins obtained in step 5 using the Ensembl gene identifier and online resources such as NCBI's Entrez (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene), GeneCards (http://www.genecards.org) (27), and Uniprot. Also examine the DBD and partner domain composition (i.e. non-DBDs) to remove genes that are unlikely to be TFs (e.g. enzymes, transmembrane proteins, etc.).
7. Finally, augment this list with any known TFs that were missed.

3.2. Assessing the Coverage of the TF Repertoire

Once TFs have been identified, we recommend automated benchmarks against other datasets to gauge the coverage of the repertoire. A common approach is to compare against a reference dataset such as genes annotated as TFs in the Gene Ontology database (GO, molecular function annotation). Note that the reference dataset is not necessarily a true gold standard – indeed the aim is to create a much more comprehensive list of TFs than any other – and these benchmarks will necessarily be estimates. The coverage is calculated as the proportion of entries in the reference dataset that are included in the repertoire.

In our own analysis, we used a list of 62 human and 207 mouse experimentally validated TFs as our reference dataset. The estimated coverage of the repertoire was 85–95% (7).

Below is a step-by-step protocol to estimate the coverage of the TF repertoire, using human as an example:

1. Extract a reference dataset of genes from GO (http://www.geneontology.org) annotated with the terms "transcription factor activity" and "DNA-binding", restricting to experimentally derived evidence (Inferred from Direct Assay, IDA; Inferred from Mutant Phenotype, IMP).
2. Estimate the coverage by calculating the fraction of the reference dataset that is included in your TF repertoire.
3. (Optional) In case the GO annotation is sparse in the organism of interest, it is possible to repeat the analysis using orthologues from a better-annotated genome such as the mouse. Multiple estimates obtained across species will ensure a robust measure of coverage. Alternatively, it is also possible to compare against carefully annotated TF repertoires in other genomes, such as worm or fly.

3.3. Assigning Regulatory Functions to the TF Repertoire

It is possible to survey the cellular and biological processes that are known to be regulated by the TF repertoire by using resources such as GO, Entrez, or PubMed.

In our analysis, we found that the most common GO regulatory functions were for developmental (263 TFs) and cellular processes (221 TFs) (7).

Below is a step-by-step protocol to annotate regulatory functions to TFs:

1. Extract the GO annotations (biological process category) for your repertoire. Depending on stringency, it is possible to restrict annotations to experimentally derived evidence as before (IDA, IMP; very strict) or relax them to include annotations inferred from automatic computational assignments (Inferred from Electronic Annoatation; IEA; very lenient).
2. Upload the TF repertoire to g:Profiler (http://biit.cs.ut.ee/gprofiler) to assess the biological processes that are statistically enriched.
3. (Optional) Query PubMed to access research articles that cite members of the TF repertoire. Note that gene names often resemble common scientific terms, and so careful filtering of hits is required.

3.4. Structural Classification of the TF Repertoire

TFs are most commonly classified by the identity of the DBD. This classification has proved to be very useful in tracing the evolutionary origins of TF families, and understanding how they recognise and bind specific DNA sequences (28). Furthermore, the identity of the DBD itself may indicate the regulatory function of the TF; for example, homeodomains are most commonly found in developmental regulators.

Most importantly, investigating the relative frequency of different TF families across species can bring interesting observations regarding their evolutionary history. For instance, 80% of the human repertoire comprises C_2H_2-Zn fingers, homeodomains, and helix-loop-helix proteins (7). However, this distribution varies substantially between species, as there are lineage-specific expansions such as the nuclear hormone receptor family in worms. Similarly, a recent survey of TFs across the three kingdoms of life described a number of kingdom-specific DBDs (29, 30).

Below is a step-by-step protocol to classify TFs by their DBD:

1. Download a list of TF families and associated InterPro IDs (http://www.valleyofpigs.org/humantfs). It is possible to reconstruct similar groups of TF families using the parent–child relationships between InterPro domains (http://www.ebi.ac.uk/interpro).
2. Calculate the proportions of distinct TF families in the repertoire and compare it against those of other species (see Table 1).

3.5. Measuring Expression of the TF Repertoire

Large-scale gene expression datasets allow us to assess patterns of TF usage across different cell types and conditions (7, 31–33). Two major types of information can be inferred from such data: (1) expression of a TF in specific circumstances is indicative of its function as a regulator in the condition and, therefore, is a good candidate for further investigation; (2) global expression pattern of the TF repertoire might reveal important organisational principles of the regulatory system, such as the hierarchy of control between TFs.

In performing this analysis, it is important to note that many TFs are post-translationally regulated through covalent modifications, subcellular localisation and oligomerisation; therefore, their expression alone may not be indicative of regulatory activity. It is also important to remember that most eukaryotic TFs function in combination with other regulators; therefore, appreciation of combinatorial regulation is important to understand TF functionality (34).

There are large numbers of array-based, and more recently increasing numbers of sequencing-based (RNA-Seq), transcriptomic datasets that can be used to examine tissue-, condition-, and disease-specific TF-expression. The GNF SymAtlas measuring expression across 79 human tissues and cell lines is one of the most widely used datasets (19). More recently, a compendium of over 6,000 array hybridisations has been published, which covers most publicly available datasets generated using the Affymetrix U133a GeneChip (35). One of the limitations of microarrays is that many genes, including TFs, are not represented in the array design. The application of high-throughput sequencing circumvents this problem as does not require prior knowledge of the transcripts that will be present. Additionally RNA-Seq is considered to provide more sensitive and accurate quantification of expression levels (RNA-Seq) (36). Although there are still few datasets available and analysis methods are still being developed, the use of RNA-Seq should increase rapidly in the coming years.

In our analysis, we used a subset of the GNF SymAtlas to examine TF expression across 32 healthy tissue samples. 873 TF-coding genes were present on the Affymetrix HG-U133a GeneChip. We detected expression of 510 TFs in at least one of the tissue types. In general, we found that TFs are expressed at lower levels than non-TFs. In addition, we observed general expression of 161 TFs across most tissues in the dataset, whereas 349 displayed tissue-specific expression (Fig. 1b) (7).

Below is a step-by-step protocol to assess TF expression using the GNF SymAtlas dataset:

1. Download the raw .CEL microarray files for the dataset from the SymAtlas website (http://biogps.gnf.org/downloads). Alternatively, the processed dataset is available from ArrayExpress (accession number E-TABM-145).

2. Download and install the R statistical software (http://www.r-project.org) and the BioConductor bioinformatics analysis suite (http://www.bioconductor.org).
3. Load the .CEL files into R and perform a quality check of the raw data using arrayQualityMetrics package (http://www.bioconductor.org/packages/2.6/bioc/html/arrayQualityMetrics.html) (37).
4. Pre-process and normalise the data using GCRMA (GeneChip Robust Multi-array Analysis) as implemented in Bioconductor. This will output values representing $\log_2$ expression levels for each probe set across all arrays. Alternative normalisation methods can also be used.
5. Download the mapping between probe sets and Ensembl gene IDs using the biomaRt package within BioConductor (http://www.bioconductor.org/packages/2.6/bioc/html/biomaRt.html) (38) or from the Ensembl website (http://www.ensembl.org).
6. Compare $\log_2$ expression values between TF and non-TF-coding genes. Statistical significance for the comparison can be obtained using a Wilcoxon or *t*-test available within R.
7. Determine the "presence" or "absence" (i.e. expressed or not expressed status) for each probe set using the PANP package (http://www.bioconductor.org/packages/2.6/bioc/html/panp.html) or MAS5.0 (http://www.bioconductor.org/packages/2.6/bioc/html/affy.html).
8. Evaluate the number of TF and non-TF-coding genes expressed in each tissue using the presence and absence calls.
9. Calculate tissue-specific expression of genes using for each gene the SpeCond package (http://www.bioconductor.org/packages/2.6/bioc/html/SpeCond.html).
10. Apply hierarchical clustering (available through the *hclust* function in R) to the expression levels and tissue-specific expression values, and represent the output graphically in a heat map.

3.6. Examining the Evolutionary Conservation of the TF Repertoire

The evolutionary history of TF-coding genes can reveal useful information about their regulatory functions and genomic organisation. For example, in our recent study of the human TF repertoire, we showed how regulators could be separated into five distinct patterns of evolutionary conservation: those present only in primates, predominantly in mammals, vertebrates, metazoans, and finally all eukaryotes (7).

There are many ways to determine the evolutionary conservation of a gene, ranging from orthologue-finding using reciprocal best-matching BLAST hits to detailed phylogenetic methods. There are also online resources that provide automatically calculated phylogenetic relationships such as Ensembl GeneTrees (22), TreeFam (39), and Inparanoid (40). We recommend manual inspection of automated results, if it is important to obtain accurate information about particular genes. A major challenge in examining the evolution of TFs is that they undergo a great deal of lineage-specific expansion; as a result it is often difficult to discriminate between orthologous and paralogous relationships.

In our analysis, we observed dramatic, step-wise increases in the size of the TF repertoire at key points during the evolution of the human lineage (Fig. 1c) (7). Interestingly, different classes of TF families expanded at these stages: for example, C_2H_2-Zn fingers expanded rapidly in vertebrates, whereas helix-loop-helix proteins originated in metazoan organisms and have not expanded significantly since.

Below is a step-by-step description of a possible procedure for identifying orthologues using Ensembl Compara GeneTrees. Similar analyses are possible using orthologue descriptions obtained from other sources. Note that this method highlights the evolutionary history of TFs in the genome of interest only and does not include TF-coding genes that have arisen in other lineages.

1. Use the Ensembl API to obtain Ensembl Compara GeneTrees for the TF repertoire.
2. Create matrices representing the evolutionary conservation of the TF repertoire: rows represent TFs, columns represent organisms, and intersecting cells describe the presence or absence of an orthologue.
3. Visualise the conservation of the TF repertoire by applying a clustering algorithm to the matrix and displaying the results in a heat map.
4. Classify TFs by their pattern of conservation. Human TFs were grouped according to their conservation in specific groups of genomes: for example, to define the five phylogenetic groups (primate-specific, mammalian-specific, vertebrate-specific, metazoan-specific, and eukaryotic), we examined the presence of orthologues in a step-wise manner. For each TF, (1) flag it as eukaryotic if they are present in any single-celled organism such as yeast; (2) for those not present in (1), classify as metazoan if present in two or more metazoan genomes; (3) continue for vertebrates, mammals, and non-human primates.
5. Assess when different TF families appeared in the human lineage by calculating the proportion of DBD types represented among the primate-specific, mammalian-specific, vertebrate-specific, metazoan-specific, and all eukaryotic groups of TFs.

3.7. Assessing the Genomic Organisation of the TF Repertoire

The genomic location of a gene is closely linked to its evolutionary history. As a result, genes with similar functions are often located in clusters at particular loci. Among human TFs, this is exemplified by the HOX clusters, which are found on chromosomes 2, 7, 12, and 17, and the organisation of these genes is extremely well conserved through much of the human lineage. The clustering of HOX genes is important functionally also, since it impacts on their expression; displacement of genes from their original locations results in misexpression and eventually developmental defects (41).

The coordinated expression of gene clusters probably arises from a combination of influencing factors, including shared enhancer elements and the effects of local chromatin structure. Further with the recent availability of high-resolution datasets such as Hi-C and ChIA-PET (42, 43), there is current renewed interest in the spatial arrangement of chromosomes within the nucleus, and the impact this has on transcriptional regulation.

In our study of the human TF repertoire, we found that chromosome 19 is particularly enriched for TF-coding genes compared with others; moreover, we identified 15 new clusters of potentially evolutionarily related TF-coding genes (Fig. 1d) (7). Many of these clusters reside in sub-telomeric and centromeric regions of chromosomes, suggesting that they may experience rapid turnover during evolution compared with other gene types.

Below is a step-by-step description to examine the organisation of TF-coding genes in the human genome:

1. Download the genomic coordinate of TF-coding genes (including chromosome, strand, start and end coordinates) from Ensembl (http://www.ensembl.org).
2. For each chromosome, count the number of TF- and non-TF-coding genes in a 500 kb sliding window using a step size of 100 kb.
3. For each window, construct a 2×2 contingency table of the numbers of TF and non-TF-coding genes present inside and outside the window.
4. Use a Fisher's exact test as implemented in the R statistical software (http://www.r-project.org) to calculate the significance of the TF enrichment for each window. As we are only interested in over-representations of TFs a one-side test can be used.
5. Correct the resulting *p*-values for multiple-testing using FDR as implemented in R (44) and identify windows that are significantly enriched for TFs using a cutoff of $p < 0.05$.
6. To define the TF clusters, merge overlapping windows with significant *p*-values.

4. Conclusions and Further Directions

Transcriptional regulation in eukaryotes is an area of research that attracts a great deal of interest owing to its importance to understanding how cells function. In addition to defining the TF repertoire or a genome, it is possible to integrate large-scale datasets to describe their potential regulatory functions and trace their evolutionary histories. By identifying further datasets and methods of data integration, it should be possible to perform further investigations such as examining the combinatorial usage of TFs, and the regulatory interaction with microRNAs. Such analyses will provide a valuable starting point for more detailed characterisation of individual TFs.

References

1. Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2001) Serial regulation of transcriptional regulators in the yeast cell cycle, *Cell* ***106***, 697–708.
2. Bain, G., Maandag, E. C., Izon, D. J., Amsen, D., Kruisbeek, A. M., Weintraub, B. C., Krop, I., Schlissel, M. S., Feeney, A. J., and van Roon, M. (1994) E2A proteins are required for proper B cell development and initiation of immunoglobulin gene rearrangements, *Cell 79*, 885–892.
3. Lemon, B., and Tjian, R. (2000) Orchestrated response: a symphony of transcription factors for gene control, *Genes Dev* ***14***, 2551–2569.
4. Mitchell, P. J., and Tjian, R. (1989) Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins, *Science* ***245***, 371–378.
5. Levine, M., and Tjian, R. (2003) Transcription regulation and animal diversity, *Nature* ***424***, 147–151.
6. Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C., Coburn, D., Newburger, D. E., Morris, Q., Hughes, T. R., and Bulyk, M. L. (2009) Diversity and complexity in DNA recognition by transcription factors, *Science* ***324***, 1720–1723.
7. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009) A census of human transcription factors: function, expression and evolution, *Nat. Rev. Genet* ***10***, 252–263.
8. Flicek, P., Aken, B. L., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Gräf, S., Haider, S., Hammond, M., Howe, K., Jenkinson, A., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Koscielny, G., Kulesha, E., Lawson, D., Longden, I., Massingham, T., McLaren, W., Megy, K., Overduin, B., Pritchard, B., Rios, D., Ruffier, M., Schuster, M., Slater, G., Smedley, D., Spudich, G., Tang, Y. A., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S. P., Zadissa, A., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X. M., Herrero, J., Hubbard, T. J. P., Parker, A., Proctor, G., Smith, J., and Searle, S. M. J. (2010) Ensembl's 10th year, *Nucleic Acids Res 38*, D557–562.
9. Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist, C. J. A., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., and Yeats, C. (2009) InterPro: the integrative protein signature database, *Nucleic Acids Res 37*, D211–215.
10. Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., and Bateman, A. (2010) The Pfam protein families database, *Nucleic Acids Res 38*, D211–222.
11. Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C., and Gough, J. (2009) SUPERFAMILY – sophisticated comparative genomics, data mining, visualization and phylogeny, *Nucleic Acids Res 37*, D380–386.
12. Sigrist, C. J. A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., and Hulo, N. (2010) PROSITE, a protein domain database for functional characterization and annotation, *Nucleic Acids Res 38*, D161–166.
13. Kummerfeld, S. K., and Teichmann, S. A. (2006) DBD: a transcription factor prediction database, *Nucleic Acids Res* 34, D74–81.
14. Wilson, D., Charoensawan, V., Kummerfeld, S. K., and Teichmann, S. A. (2008) DBD – taxonomically

broad transcription factor predictions: new content and functionality, *Nucleic Acids Res* ***36***, D88–92.

15. Mulder, N., and Apweiler, R. (2007) InterPro and InterProScan: tools for protein sequence classification and comparison, *Methods Mol. Biol* ***396***, 59–70.
16. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat. Genet* ***25***, 25–29.
17. Reimand, J., Kull, M., Peterson, H., Hansen, J., and Vilo, J. (2007) g:Profiler – a web-based toolset for functional profiling of gene lists from large-scale experiments, *Nucleic Acids Res* ***35***, W193–200.
18. Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., Madej, T., Maglott, D. R., Marchler-Bauer, A., Miller, V., Mizrachi, I., Ostell, J., Panchenko, A., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Wang, Y., John Wilbur, W., Yaschenko, E., and Ye, J. (2010) Database resources of the National Center for Biotechnology Information, *Nucleic Acids Res* ***38***, D5–16.
19. Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes, *Proc. Natl. Acad. Sci. USA* ***101***, 6062–6067.
20. Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M., Abeygunawardena, N., Berube, H., Dylag, M., Emam, I., Farne, A., Holloway, E., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Rayner, T. F., Rezwan, F., Sharma, A., Williams, E., Bradley, X. Z., Adamusiak, T., Brandizi, M., Burdett, T., Coulson, R., Krestyaninova, M., Kurnosov, P., Maguire, E., Neogi, S. G., Rocca-Serra, P., Sansone, S., Sklyar, N., Zhao, M., Sarkans, U., and Brazma, A. (2009) ArrayExpress update – from an archive of functional genomics experiments to the atlas of gene expression, *Nucleic Acids Res* ***37***, D868–872.
21. Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004) Bioconductor: open software development for computational biology and bioinformatics, *Genome Biol* 5, R80.
22. Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates, *Genome Res* ***19***, 327–335.
23. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool, *J. Mol. Biol* ***215***, 403–410.
24. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res* ***25***, 3389–3402.
25. (2010) The Universal Protein Resource (UniProt) in 2010, *Nucleic Acids Res* ***38***, D142–148.
26. Hu, S., Xie, Z., Onishi, A., Yu, X., Jiang, L., Lin, J., Rho, H., Woodard, C., Wang, H., Jeong, J., Long, S., He, X., Wade, H., Blackshaw, S., Qian, J., and Zhu, H. (2009) Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling, *Cell* ***139***, 610–622.
27. Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., Sirota-Madi, A., Olender, T., Golan, Y., Stelzer, G., Harel, A., and Lancet, D. (2010) GeneCards Version 3: the human gene integrator, *Database (Oxford)* 2010, baq020.
28. Luscombe, N. M., Austin, S. E., Berman, H. M., and Thornton, J. M. (2000) An overview of the structures of protein-DNA complexes, *Genome Biol* ***1***, REVIEWS001.
29. Charoensawan, V., Wilson, D., and Teichmann, S. A. (2010) Genomic repertoires of DNA-binding transcription factors across the tree of life, *Nucleic Acids Res.*
30. Charoensawan, V., Wilson, D., and Teichmann, S. A. (2010) Lineage-specific expansion of DNA-binding transcription factor families, *Trends Genet* ***26***, 388–393.
31. Zaslaver, A., Mayo, A. E., Rosenberg, R., Bashkin, P., Sberro, H., Tsalyuk, M., Surette, M. G., and Alon, U. (2004) Just-in-time transcription program in metabolic pathways, *Nat. Genet* ***36***, 486–491.

32. Freilich, S., Massingham, T., Bhattacharyya, S., Ponsting, H., Lyons, P. A., Freeman, T. C., and Thornton, J. M. (2005) Relationship between the tissue-specificity of mouse gene expression and the evolutionary origin and function of the proteins, *Genome Biol 6*, R56.

33. Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes, *Nature 431*, 308–312.

34. Ravasi, T., Suzuki, H., Cannistraci, C. V., Katayama, S., Bajic, V. B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., Carninci, P., Daub, C. O., Forrest, A. R. R., Gough, J., Grimmond, S., Han, J., Hashimoto, T., Hide, W., Hofmann, O., Kamburov, A., Kaur, M., Kawaji, H., Kubosaki, A., Lassmann, T., van Nimwegen, E., MacPherson, C. R., Ogawa, C., Radovanovic, A., Schwartz, A., Teasdale, R. D., Tegnér, J., Lenhard, B., Teichmann, S. A., Arakawa, T., Ninomiya, N., Murakami, K., Tagami, M., Fukuda, S., Imamura, K., Kai, C., Ishihara, R., Kitazume, Y., Kawai, J., Hume, D. A., Ideker, T., and Hayashizaki, Y. (2010) An atlas of combinatorial transcriptional regulation in mouse and man, *Cell 140*, 744–752.

35. Lukk, M., Kapushesky, M., Nikkilä, J., Parkinson, H., Goncalves, A., Huber, W., Ukkonen, E., and Brazma, A. (2010) A global map of human gene expression, *Nat. Biotechnol 28*, 322–324.

36. Wang, Z., Gerstein, M., and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics, *Nat. Rev. Genet 10*, 57–63.

37. Kauffmann, A., Gentleman, R., and Huber, W. (2009) arrayQualityMetrics – a bioconductor package for quality assessment of microarray data, *Bioinformatics 25*, 415–416.

38. Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis, *Bioinformatics 21*, 3439–3440.

39. Ruan, J., Li, H., Chen, Z., Coghlan, A., Coin, L. J. M., Guo, Y., Hériché, J., Hu, Y., Kristiansen, K., Li, R., Liu, T., Moses, A., Qin, J., Vang, S., Vilella, A. J., Ureta-Vidal, A., Bolund, L., Wang, J., and Durbin, R. (2008) TreeFam: 2008 Update, *Nucleic Acids Res 36*, D735–740.

40. Ostlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D. N., Roopra, S., Frings, O., and Sonnhammer, E. L. L. (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis, *Nucleic Acids Res 38*, D196–203.

41. Garcia-Fernàndez, J. (2005) The genesis and evolution of homeobox gene clusters, *Nat. Rev. Genet 6*, 881–892.

42. Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome, *Science 326*, 289–293.

43. Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., Chew, E. G. Y., Huang, P. Y. H., Welboren, W., Han, Y., Ooi, H. S., Ariyaratne, P. N., Vega, V. B., Luo, Y., Tan, P. Y., Choy, P. Y., Wansa, K. D. S. A., Zhao, B., Lim, K. S., Leow, S. C., Yow, J. S., Joseph, R., Li, H., Desai, K. V., Thomsen, J. S., Lee, Y. K., Karuturi, R. K. M., Herve, T., Bourque, G., Stunnenberg, H. G., Ruan, X., Cacheux-Rataboul, V., Sung, W., Liu, E. T., Wei, C., Cheung, E., and Ruan, Y. (2009) An oestrogen-receptor-alpha-bound human chromatin interactome, *Nature 462*, 58–64.

44. Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society Series B-Methodological 57*, 289–300.

45. Reece-Hoyes, J. S., Deplancke, B., Shingles, J., Grove, C. A., Hope, I. A., and Walhout, A. J. M. (2005) A compendium of *Caenorhabditis elegans* regulatory transcription factors: a resource for mapping transcription regulatory networks, *Genome Biol 6*, R110.

46. Kanamori, M., Konno, H., Osato, N., Kawai, J., Hayashizaki, Y., and Suzuki, H. (2004) A genome-wide and nonredundant mouse transcription factor database, *Biochem. Biophys. Res. Commun 322*, 787–793.

47. Palaniswamy, S. K., James, S., Sun, H., Lamb, R. S., Davuluri, R. V., and Grotewold, E. (2006) AGRIS and AtRegNet. A platform to link cis-regulatory elements and transcription factors into regulatory networks, *Plant Physiol 140*, 818–829.

48. Messina, D. N., Glasscock, J., Gish, W., and Lovett, M. (2004) An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression, *Genome Res 14*, 2041–2047.

49. Wu, J., Wang, S., Bai, J., Shi, L., Li, D., Xu, Z., Niu, Y., Lu, J., and Bao, Q. (2008) ArchaeaTF: an integrated database of putative transcription factors in Archaea, *Genomics 91*, 102–107.

50. Martínez-Bueno, M., Molina-Henares, A. J., Pareja, E., Ramos, J. L., and Tobes, R. (2004) BacTregulators: a database of transcriptional regulators in bacteria and archaea, *Bioinformatics* ***20***, 2787–2791.
51. Baumbach, J., Brinkrolf, K., Czaja, L. F., Rahmann, S., and Tauch, A. (2006) CoryneRegNet: an ontology-based data warehouse of corynebacterial transcription factors and regulatory networks, *BMC Genomics* *7*, 24.
52. Wu, J., Zhao, F., Wang, S., Deng, G., Wang, J., Bai, J., Lu, J., Qu, J., and Bao, Q. (2007) cTFbase: a database for comparative genomics of transcription factors in cyanobacteria, *BMC Genomics* ***8***, 104.
53. Sierro, N., Makita, Y., de Hoon, M., and Nakai, K. (2008) DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information, *Nucleic Acids Res* ***36***, D93–96.
54. Pfreundt, U., James, D. P., Tweedie, S., Wilson, D., Teichmann, S. A., and Adryan, B. (2010) FlyTF: improved annotation and enhanced functionality of the Drosophila transcription factor database, *Nucleic Acids Res* ***38***, D443–447.
55. Park, J., Park, J., Jang, S., Kim, S., Kong, S., Choi, J., Ahn, K., Kim, J., Lee, S., Kim, S., Park, B., Jung, K., Kim, S., Kang, S., and Lee, Y. (2008) FTFD: an informatics pipeline supporting phylogenomic analysis of fungal transcription factors, *Bioinformatics* ***24***, 1024–1025.
56. Moreno-Campuzano, S., Janga, S. C., and Pérez-Rueda, E. (2006) Identification and analysis of DNA-binding transcription factors in *Bacillus subtilis* and other Firmicutes – a genomic approach, *BMC Genomics* *7*, 147.
57. Zheng, G., Tu, K., Yang, Q., Xiong, Y., Wei, C., Xie, L., Zhu, Y., and Li, Y. (2008) ITFP: an integrated platform of mammalian transcription factors, *Bioinformatics* ***24***, 2416–2417.
58. Portales-Casamar, E., Thongjuea, S., Kwon, A. T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W. W., and Sandelin, A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles, *Nucleic Acids Res* ***38***, D105–110.
59. Gray, P. A., Fu, H., Luo, P., Zhao, Q., Yu, J., Ferrari, A., Tenzen, T., Yuk, D., Tsung, E. F., Cai, Z., Alberta, J. A., Cheng, L., Liu, Y., Stenman, J. M., Valerius, M. T., Billings, N., Kim, H. A., Greenberg, M. E., McMahon, A. P., Rowitch, D. H., Stiles, C. D., and Ma, Q. (2004) Mouse brain organization revealed through direct genome-scale TF expression analysis, *Science* ***306***, 2255–2257.
60. Richardt, S., Lang, D., Reski, R., Frank, W., and Rensing, S. A. (2007) PlanTAPDB, a phylogeny-based resource of plant transcription-associated proteins, *Plant Physiol* ***143***, 1452–1466.
61. Guo, A., Chen, X., Gao, G., Zhang, H., Zhu, Q., Liu, X., Zhong, Y., Gu, X., He, K., and Luo, J. (2008) PlantTFDB: a comprehensive plant transcription factor database, *Nucleic Acids Res* ***36***, D966–969.
62. Pérez-Rodríguez, P., Riaño-Pachón, D. M., Corrêa, L. G. G., Rensing, S. A., Kersten, B., and Mueller-Roeber, B. (2010) PlnTFDB: updated content and new features of the plant transcription factor database, *Nucleic Acids Res* ***38***, D822–827.
63. Grote, A., Klein, J., Retter, I., Haddad, I., Behling, S., Bunk, B., Biegler, I., Yarmolinetz, S., Jahn, D., and Münch, R. (2009) PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes, *Nucleic Acids Res* *37*, D61–65.
64. Iida, K., Seki, M., Sakurai, T., Satou, M., Akiyama, K., Toyoda, T., Konagaya, A., and Shinozaki, K. (2005) RARTF: database and tools for complete sets of Arabidopsis transcription factors, *DNA Res* ***12***, 247–256.
65. Kazakov, A. E., Cipriano, M. J., Novichkov, P. S., Minovitsky, S., Vinogradov, D. V., Arkin, A., Mironov, A. A., Gelfand, M. S., and Dubchak, I. (2007) RegTransBase – a database of regulatory sequences and interactions in a wide range of prokaryotic genomes, *Nucleic Acids Res* ***35***, D407–412.
66. Gama-Castro, S., Jiménez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Peñaloza-Spinola, M. I., Contreras-Moreira, B., Segura-Salazar, J., Muñiz-Rascado, L., Martínez-Flores, I., Salgado, H., Bonavides-Martínez, C., Abreu-Goodger, C., Rodríguez-Penagos, C., Miranda-Ríos, J., Morett, E., Merino, E., Huerta, A. M., Treviño-Quintanilla, L., and Collado-Vides, J. (2008) RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation, *Nucleic Acids Res* ***36***, D120–124.
67. Wang, Z., Libault, M., Joshi, T., Valliyodan, B., Nguyen, H. T., Xu, D., Stacey, G., and Cheng, J. (2010) SoyDB: a knowledge database of soybean transcription factors, *BMC Plant Biol* ***10***, 14.
68. Fulton, D. L., Sundararajan, S., Badis, G., Hughes, T. R., Wasserman, W. W., Roach, J. C., and Sladek, R. (2009) TFCat: the curated catalog of mouse and human transcription factors, *Genome Biol* ***10***, R29.

69. Lee, A. P., Yang, Y., Brenner, S., and Venkatesh, B. (2007) TFCONES: a database of vertebrate transcription factor-encoding genes and their associated conserved noncoding elements, *BMC Genomics* ***8***, 441.
70. Fukuchi, S., Homma, K., Sakamoto, S., Sugawara, H., Tateno, Y., Gojobori, T., and Nishikawa, K. (2009) The GTOP database in 2009: updated content and novel features to expand and deepen insights into protein structures and functions, *Nucleic Acids Res* ***37***, D333–337.
71. Rushton, P. J., Bokowiec, M. T., Laudeman, T. W., Brannock, J. F., Chen, X., and Timko, M. P. (2008) TOBFAC: the database of tobacco transcription factors, *BMC Bioinformatics* ***9***, 53.
72. Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., and Wingender, E. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, *Nucleic Acids Res* ***34***, D108–110.
73. Hermoso, A., Aguilar, D., Aviles, F. X., and Querol, E. (2004) TrSDB: a proteome database of transcription factors, *Nucleic Acids Res* ***32***, D171–173.
74. Romeuf, I., Tessier, D., Dardevet, M., Branlard, G., Charmet, G., and Ravel, C. (2010) wDBTF: an integrated database resource for studying wheat transcription factor families, *BMC Genomics* ***11***, 185.