Genomic repertoires of DNA-binding transcription factors across the tree of life
Varodom Charoensawan*, Derek Wilson and Sarah A. Teichmann*

MRC Laboratory of Molecular Biology, Cambridge, CB2 0QH, UK

ABSTRACT
Sequence-specific transcription factors (TFs) are important to genetic regulation in all organisms because they recognize and directly bind to regulatory regions on DNA. Here, we survey and summarize the TF resources available. We outline the organisms for which TF annotation is provided, and discuss the criteria and methods used to annotate TFs by different databases. By using genomic TF repertoires from ~700 genomes across the tree of life, covering Bacteria, Archaea and Eukaryota, we review TF abundance with respect to the number of genes, as well as their structural complexity in diverse lineages. While typical eukaryotic TFs are longer than the average eukaryotic proteins, the inverse is true for prokaryotes. Only in eukaryotes does the same family of DNA-binding domain (DBD) occur multiple times within one polypeptide chain. This potentially increases the length and diversity of DNA-recognition sequence by reusing DBDs from the same family. We examined the increase in TF abundance with the number of genes in genomes, using the largest set of prokaryotic and eukaryotic genomes to date. As pointed out before, prokaryotic TFs increase faster than linearly. We further observe a similar relationship in eukaryotic genomes with a slower increase in TFs.

INTRODUCTION
Regulation of gene expression has always been one of the most prominent areas in the field of genetics. The mechanism of genetic regulation was unveiled for the first time, when Jacob and Monod (1) uncovered the gene regulation apparatus of the lac operon in Escherichia coli. Since then, numerous studies [e.g. (2–4)] have shown that regulation of gene expression is essential to determining organismal complexity and morphological diversity in different species across the tree of life. Transcriptional regulation is a crucial step in gene expression regulation because the genetic information is directly read from DNA by sequence-specific transcription factors (TFs). The unique role of TFs is highlighted by several studies demonstrating their abilities to reprogramme fibroblasts into embryonic stem cells (5,6).

Numerous studies have provided a great deal of insight into the conserved and specific DNA-binding TFs in different lineages, though they tended to concentrate on particular phylogenetic groups. The DNA-binding domains (DBDs), evolutionary components of sequence-specific TFs that mediate the specificity of the TF–DNA interaction, are often used to represent TF families, which is appropriate from functional as well as evolutionary points of view. Despite their importance, the global DBD repertoire was only once reviewed from a structural perspective a decade ago (7).

Being aware of the importance of TFs on genetic regulation, the community has put a great amount of effort into the development of resources for the systematic collection and classification of annotated TFs in genomes from diverse lineages. Here, we summarize key publications of genome-wide studies of TFs and survey TF databases currently available, as well as discuss the criteria and the methods used to obtain TF catalogues. A better understanding of global TF repertoires in species from diverse and related lineages will not only serve as a starting point for experimental design of high-throughput studies for determining the binding sites of TFs in different model organisms (8–12), but will also offer an insight into the evolution of TFs in conjunction with the remainder of the proteins in genomes that they regulate. To summarize our current knowledge on the genomic repertoires of TFs across the tree of life (from Bacteria, Archaea and Eukaryota superkingdoms), we used TFs annotated by the DBD database (13) in ~700 organisms as
representatives to review their abundance with respect to
the number of genes in different genomes, as well as their
complexity in terms of DBD and other protein domain
components in diverse phylogenetic branches.

GENOME-WIDE STUDIES OF TFs IN DIFFERENT
PHYLOGENETIC GROUPS

Since the first whole-genome sequencing was completed in
1995 on a pathogenic bacterium *Haemophilus influenzae*
(14), the number of completely sequenced prokaryotic
genomes has been increasing rapidly, with a doubling
time of ~20 months for bacteria and ~34 months for
archaea (15). Due to the abundance of publicly available
prokaryotic genomes, a large number of whole genome TF
studies have focused on these organisms. Aravind and
Koonin (16) published one of the earlier analyses on the
repertoire of TF families in four complete archaeal
genomes. Using sequence profile methods in conjunction
with protein structure information, they presented the
intriguing finding that the majority of archaeal DBDs
had helix-turn-helix (HTH) structures similar to bacterial
HTH domains. This contrasts with the core components
of the archaeal transcriptional machinery, such as basal
TFs and RNA polymerases, which are more closely
related to eukaryotic systems. A more recent study by
Coulson and coworkers confirmed this finding (17).
Since then, similar types of analysis were conducted by
different groups with larger sets of prokaryotic species.

Perez-Rueda et al. (18) addressed the distribution of
75 TF families across 90 prokaryotes based on the
well-characterized set of TFs in *E. coli* K12. Because the
reference TFs were taken from one bacterial species,
the predicted TFs were restricted to close homologues of
TFs found in *E. coli*. Similarly, Minezaki et al. (19) clas-
sified TFs from 154 complete prokaryotic genomes into
52 TF families. Their TF families were collected from TFs
found in eight different archaea and bacteria, with addi-
tional DBDs documented in Pfam (20). Thus, this refer-
ence TF set was likely to detect additional varieties of TF
homologues across prokaryotic proteins. Different criteria
for constructing the reference TFs notwithstanding, both
studies consolidated the predominance of HTH DBDs in
prokaryotes, especially the winged-HTHs. They also
demonstrated a significant depletion of TF families in
intra-cellular pathogenic and endosymbiotic bacteria
including *Mycoplasma* and *Chlamydia pneumoniae*. These patho-
genic life forms normally inhabit hosts whose environment
lacks selective pressure to maintain the specific genes to
respond to environmental stress. Other groups considered
more restricted lineages of bacteria including Moreno-
Campuzano et al. (21) and Brune et al. (22). Their studies provided comprehensive lists of TF repertoires in
firmicutes and corynebacteria, respectively.

Baker’s yeast *Saccharomyces cerevisiae* was the first eu-
karyotic species to have its genome completely sequenced.
The paper describing the whole-genome sequencing of
baker’s yeast (23) was published in 1996, only slightly
after the first prokaryotic genome *H. influenzae*. The
number of completely sequenced eukaryotic genomes,
however, increases significantly more slowly than that of
prokaryotic genomes. This is likely due to the combina-
tion of larger average size of eukaryotic genomes, and
the difficulty in assembling and annotating the genomes
that contain a great amount of repetitive and non-coding
elements (24). Nonetheless, an increasing number of
studies on the genomic TF repertoires are being conducted
using complete eukaryotic genomes.

Riechmann et al. (25) surveyed specific TF families
occurring in four eukaryotic genomes: *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*
and *S. cerevisiae*. They demonstrated that a number of
DBD families are shared across all three major eukaryotic
kingdoms, i.e. Metazoa (animal), Fungi and Viridiplantae
(plant), but the domain combinations of DBDs and other
domains in TFs are highly kingdom-specific. According
to Coulson and Ouzounis (26), each eukaryotic kingdom
possesses not only the families common to all eukaryotes,
but also a number of kingdom-specific transcriptional
regulators, which possibly participate in kingdom-specific
processes. Other studies focused on particular eukaryotic
kingdoms. In plant, Shiu et al. (27) pointed out that not
only were the TF families more diverse compared with
fungi and animals, but the expansion and duplication
rates in plants were also considerably greater. This
suggests a more frequent adaptive response to selection
pressure among plants since they do not have mobility
to avoid stress stimuli in the same way as other eukary-
otes. More recent work by Shelest (28) concentrated on
TFs in fungi, reporting 37 TF families in 62 fungal species,
of which only six families were fungal specific. Being
phylogenetically distant from animals, fungi and plants,
the genomes of parasitic protists such as apicomplexans
and ciliates are known to be substantially divergent from
the current model eukaryotic genomes and thus less
well-understood. Iyer and coworkers (29) were the first
to investigate the repertoires of TFs and chromatin
proteins in these parasitic unicellular eukaryotes.

In the Metazoa (animal) kingdom, TFs are particularly
essential to the morphological development of animals’
organ systems. Messina et al. (30) compiled one of the
first lists of metazoan TFs by focusing on human. They
aimed to produce a starting point for array experiments
across species. By taking known TFs from TRANSFAC
(31), InterPro (32) and FlyBase (33) as seed sequences, addition-
al human TFs were discovered using hidden Markov
model (HMM) searches, followed by manual curation. As
part of the initiative to characterize the transcription regu-
larly network in mammalian cells, the International
Regulome Consortium (IRC) have put together a compre-
hsive list of mouse TFs by mapping cDNA sequences
from several libraries to the NCBI mouse genome. More
recently, Vaquerizas et al. (34) have manually compiled a
human TF repertoire and analyzed their expression
patterns and evolutionary conservation. These studies on
mammalian TFs will contribute to a better understanding
of gene expression control in higher organisms.

In summary, several key publications mentioned here
highlight the importance of TFs in the development and
maintenance of cellular phenotypes in different kinds of
organisms. These genome-wide studies provide a starting
point for a systematic comparative analysis of genomic TF repertoires in both closely and distantly related genomes.

TF RESOURCES FOR ORGANISMS FROM DIVERSE LINEAGES

In this section, we survey the TF resources available to date and summarize them in Table 1. The resources are categorized according to the organisms for which the TF annotations are provided. We also illustrate the list of these resources on a timeline (Figure 1), which indicates the year in which different resources were first developed. The figure shows the trends in methods used to annotate TFs, with respect to the number of complete genomes available over the past 16 years according to the Genome OnLine Database (35).

Table 1. TF resources currently available

<table>
<thead>
<tr>
<th>Database</th>
<th>Annotation</th>
<th>Organism</th>
<th>Link</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>I GTOP_TF</td>
<td>A/M</td>
<td>Prokaryotes</td>
<td><a href="http://spock.genes.nig.ac.jp/~gtop_tf/index2.html">http://spock.genes.nig.ac.jp/~gtop_tf/index2.html</a></td>
<td>Covers over 150 prokaryotic genomes</td>
</tr>
<tr>
<td>BacTregulators*</td>
<td>A/M</td>
<td>Prokaryotes</td>
<td><a href="http://www.bactregulators.org/">http://www.bactregulators.org/</a></td>
<td>AraC-Xyls and TetR transcription regulator families. Last updated 2004</td>
</tr>
<tr>
<td>PRODORIC</td>
<td>M</td>
<td>Bacteria</td>
<td><a href="http://prodoric.tu-bs.de">http://prodoric.tu-bs.de</a></td>
<td>Contains protein-DNA interaction information</td>
</tr>
<tr>
<td>RegTransBase</td>
<td>M</td>
<td>Bacteria</td>
<td><a href="http://regtransbase.lbl.gov">http://regtransbase.lbl.gov</a></td>
<td>Contains protein-DNA interaction information</td>
</tr>
<tr>
<td>ArchaeaTF</td>
<td>A/M</td>
<td>Archaea</td>
<td><a href="http://bioinformatics.zj.cn/archaeatf/">http://bioinformatics.zj.cn/archaeatf/</a></td>
<td>Covers 37 archaeal genomes</td>
</tr>
<tr>
<td>CoryneRegNet</td>
<td>A/M</td>
<td>Corynebacteria</td>
<td><a href="http://www.coryneregnet.de/">http://www.coryneregnet.de/</a></td>
<td>Contains protein-DNA interaction information</td>
</tr>
<tr>
<td>cTFbase</td>
<td>A/M</td>
<td>Cyanobacteria</td>
<td><a href="http://cegwz.com/">http://cegwz.com/</a></td>
<td>Covers 21 cyanobacterial genomes</td>
</tr>
<tr>
<td>DTBBS</td>
<td>M</td>
<td>Bacillus subtilis</td>
<td><a href="http://dbtbs.hgc.jp/">http://dbtbs.hgc.jp/</a></td>
<td>Contains other literature-curated information for B. subtilis</td>
</tr>
<tr>
<td>RegulonDB</td>
<td>M</td>
<td>E. coli K-12</td>
<td><a href="http://regulondb.ccg.unam.mx/">http://regulondb.ccg.unam.mx/</a></td>
<td>Contains other literature-curated information for E. coli K-12</td>
</tr>
<tr>
<td>II TRANSFAC</td>
<td>M</td>
<td>Eukaryotes</td>
<td><a href="http://www.gene-regulation.com/pub/databases.html">http://www.gene-regulation.com/pub/databases.html</a></td>
<td>Partially commercial. Licence required to access some restricted areas</td>
</tr>
<tr>
<td>JASPAR</td>
<td>A/M</td>
<td>Eukaryotes</td>
<td><a href="http://jaspar.cgb.ki.se">http://jaspar.cgb.ki.se</a></td>
<td>Contains collections of experimentally defined TF binding sites</td>
</tr>
<tr>
<td>TrsSDB*</td>
<td>A/M</td>
<td>Eukaryotes</td>
<td><a href="http://bioinf.uab.es/cgi-bin/trsdb/trsdb.pl/">http://bioinf.uab.es/cgi-bin/trsdb/trsdb.pl/</a></td>
<td>Covers nine eukaryotic proteomes. Last updated 2004</td>
</tr>
<tr>
<td>ITFP</td>
<td>A</td>
<td>Mammals</td>
<td><a href="http://itfp.biosino.org/itfp/">http://itfp.biosino.org/itfp/</a></td>
<td>Contains TFs and target genes from human, mouse and rat</td>
</tr>
<tr>
<td>TFCat</td>
<td>M</td>
<td>Mammals</td>
<td><a href="http://www.tfcat.ca/">http://www.tfcat.ca/</a></td>
<td>Contains manually curated TFs from human, mouse</td>
</tr>
<tr>
<td>TFdb*</td>
<td>A/M</td>
<td>Mouse</td>
<td><a href="http://genome.gsc.riken.jp/TFdb/">http://genome.gsc.riken.jp/TFdb/</a></td>
<td>Based on LocusLink and GO annotations. Last update: 2004</td>
</tr>
<tr>
<td>FTFD</td>
<td>A/M</td>
<td>Fungi</td>
<td><a href="http://fffd.snu.ac.kr/">http://fffd.snu.ac.kr/</a></td>
<td>Covers 69 fungal and three oomycete genomes</td>
</tr>
<tr>
<td>PlanTAPDB</td>
<td>A/M</td>
<td>Plants</td>
<td><a href="http://www.cosmoss.org/bm/plantapdb/">http://www.cosmoss.org/bm/plantapdb/</a></td>
<td>Contains taxonomic information of transcription associated protein families</td>
</tr>
<tr>
<td>PlantTFDB</td>
<td>A/M</td>
<td>Plants</td>
<td><a href="http://planttfdb.cbi.pku.edu.cn/">http://planttfdb.cbi.pku.edu.cn/</a></td>
<td>Integrates other plant databases: DPTF (poplar), DRTF (rice), DATF (Arabidopsis)</td>
</tr>
<tr>
<td>PhnTFDB</td>
<td>A/M</td>
<td>Plants</td>
<td><a href="http://plantfdb.bio.uni-potsdam.de/v2.0">http://plantfdb.bio.uni-potsdam.de/v2.0</a></td>
<td>Covers five model plant genomes</td>
</tr>
<tr>
<td>RARTF</td>
<td>A/M</td>
<td>A. thaliana (thale cress)</td>
<td><a href="http://rarge.gsc.riken.jp/rartf/">http://rarge.gsc.riken.jp/rartf/</a></td>
<td>TF database devoted to A. thaliana</td>
</tr>
<tr>
<td>AtTFDB*</td>
<td>A/M</td>
<td>A. thaliana</td>
<td><a href="http://arabidopsis.med.ohio-state.edu/AtTFDB/">http://arabidopsis.med.ohio-state.edu/AtTFDB/</a></td>
<td>Sister database, AtsicDB, contains cis-regulatory data. Last updated 2004</td>
</tr>
<tr>
<td>SoyDB</td>
<td>A/M</td>
<td>Glycine Max (soybean)</td>
<td><a href="http://casp.rnet.missouri.edu/soydb/">http://casp.rnet.missouri.edu/soydb/</a></td>
<td>Predicts TFs using InterProScan</td>
</tr>
<tr>
<td>wDBTF</td>
<td>A/M</td>
<td>Triticum aestivum (wheat)</td>
<td><a href="http://www.appli.nantes.inra.fr:8180/wDBTF/">http://www.appli.nantes.inra.fr:8180/wDBTF/</a></td>
<td>Predicts TFs from wheat Expressed Sequence Tags (ESTs) and mRNA.</td>
</tr>
<tr>
<td>TOBFAC</td>
<td>A/M</td>
<td>Nicotiana tabacum (Tobacco)</td>
<td><a href="http://compysbio.achs.virginia.edu/tobfac/">http://compysbio.achs.virginia.edu/tobfac/</a></td>
<td>Predicts TFs from tobacco gene-space sequence reads (GSRs)</td>
</tr>
<tr>
<td>FlyTF</td>
<td>M</td>
<td>D. melanogaster (fruit fly)</td>
<td><a href="http://flytf.org/">http://flytf.org/</a></td>
<td>TF database devoted to D. melanogaster</td>
</tr>
<tr>
<td>EDGEdb</td>
<td>A/M</td>
<td>C. elegans (worm)</td>
<td><a href="http://edgedb.umassmed.edu/">http://edgedb.umassmed.edu/</a></td>
<td>Contains protein-DNA interaction information</td>
</tr>
<tr>
<td>III DBD</td>
<td>A</td>
<td>Cellular organisms</td>
<td><a href="http://www.transcriptionfactor.org/">http://www.transcriptionfactor.org/</a></td>
<td>Contains TF predictions of more than 1000 cellular organisms</td>
</tr>
</tbody>
</table>

The databases can be divided into three categories: (I) prokaryotic TF databases; (II) Eukaryotic TF databases; (III) databases that provide TF annotations in genomes from different superkingdoms. The databases which have ceased to be developed or not been updated since 2004 are marked with asterisks. The years of the latest update are included in the comment field. Annotation methods are indicated as A (Automated) and M (Manually curated).
Prokaryotic TF resources

It has been thought that the level of transcription in prokaryotes is largely governed by the binding strengths of RNA polymerase and TFs to DNA (36). In addition, prokaryotic genomes are typically smaller, with fewer repeats and lower fractions of non-coding DNA, compared to eukaryotes. Consequently, the current knowledge on prokaryotic transcriptional mechanisms is believed to be more complete. Early established databases for prokaryotes serve as integrated resources of transcriptional machineries for specific model organisms, instead of focusing on TF annotation per se. RegulonDB (37) was one of the first TF databases to be established. The database provides high-quality information relating to the transcriptional regulatory network in the Gram-negative bacterium *E. coli* K12. Apart from the literature curated TFs, RegulonDB contains other molecular categories such as small RNAs and operon structures. Similarly, DBTBS (38) provides TFs as well as other transcriptional regulation-related knowledge such as recognition sequences and target genes for the Gram-positive bacterium *Bacillus subtilis*.

Due to the continuous development of sequencing technology, large number of completely sequenced genomes can be generated within shorter periods of time. This undoubtedly facilitates TF annotation and thus has resulted...
in a number of TF databases with larger and wider ranging prokaryotic taxonomic groups. In addition to RegulonDB and D8TBS, there are other resources, such as PRODORIC (39) and RegTransBase (40), that provide different aspects of data related to genetic regulation including lists of manually curated TFs and regulatory sites for multiple bacterial species. Other databases, such as the GTPF_TF database (19) provide a comparative insight into TF repertoires in >150 species from taxonomically diverse prokaryotic groups. Similarly, BacTregulators (41) is another resource providing TF collections for 123 genomes from archaea and bacteria, although the annotations are restricted to only two TF families: AraC-XylS and TetR. CoryneRegNet (42) integrates data on TFs and gene regulatory networks of eight corynebacteria, two mycobacteria and E. coli K12. eTFbase (43) is a database devoted to TF prediction in cyanobacteria, an ancient group of Gram-negative bacteria which reside in diverse environments. They obtain energy through photosynthesis and are believed to be the origin of eukaryotic chloroplasts according to the endosymbiotic theory (44). ArchaeaTF (45) is a unique database which is dedicated to archaea. Among the three main superkingdoms, Archaea are the least studied lineage. By providing TF prediction for 37 archaeal species, Archaea TF serves as an important platform for understanding the genetic regulation mechanisms in these life forms.

Eukaryotic TF resources

Transcription mechanisms in eukaryotes are known to be sophisticated but less well-understood, compared with the prokaryotes. Among the early eukaryotic TF resources is TRANSFAC (31). The database contains literature-curated information on sequence-specific TFs together with their binding sites, nucleotide distribution matrices as well as composite elements. While TRANSFAC is semi-commercial and some parts of the database require registration to access, JASPAR (46) is an open-access database which also mainly focuses on eukaryotic TF-binding sites. It also contains useful information on structural classes of TFs that share binding properties.

Owing to the continuous improvement of eukaryotic genome assembly and annotation, the number of databases containing TF annotation for eukaryotic species has increased rapidly in the past 5 years. A great deal of attention has been focused on plants in particular, possibly due to their importance as model organisms (e.g. Arabidopsis), food (e.g. rice, potato, wheat) and alternative energy sources (e.g. corn, sugarcane). Consequently, databases have been created to concentrate solely on TFs from groups of plant species: PlanTAPDB (47), PlantTFDB (48) and PmTFDB (49). In addition, there are many databases devoted to single plant species including RARTF (50) and AThTFDB (51) for A. thaliana (thale cress), SoyDB (52) for Glycine max (soybean), wDBTF (53) for Triticum aestivum (wheat) and TOBFAC (54) for Nicotiana tabacum (tobacco).

The only resource focusing on TF annotation in fungi is FTFD (55). The database has been developed to provide TF predictions for >70 fungal species. For animals, the currently available databases contain TF collections for specific model animal species but not for larger taxonomic groups. These databases include FlyTF (56) which provides a manually curated list of D. melanogaster TFs. EDGEdb (57) contains protein–DNA interaction data in addition to a comprehensive collection of TFs in C. elegans. A number of databases which provide TF catalogues specifically for vertebrates include TFlb (58) (mouse), ITFP (59) (human, mouse and rat) and TFcat (60) (human and mouse).

Among the currently available TF resources, the DBD TF database (13,61) is one of the most comprehensive and unique TF collections in terms of number and variety of species available. The current version of the DBD database contains TF predictions for >1000 cellular organisms from the three major superkingdoms (Bacterial, Archaea, Eukaryota).

Although our survey is by no means exhaustive, it is the most comprehensive compilation of DNA-binding TF resources to date. Apart from the databases in Table 1, there are relevant databases that concentrate on TF-binding sites but also provide some TF predictions, e.g. MATCH (62) and MAPPER (63). In addition, we also noted that there are general resources for specific genomes that contain literature-curated gene annotations, as well as listings of known TFs such as FlyBase (33) and FlyMine (64), both for Drosophila. Because these two groups of databases focus more on the TF-DNA interactions and particular genomes, respectively, and not specifically TF annotations, we did not include them in our table.

In our survey, we collected >25 TF databases that are available to date. Almost all databases concentrate on model organisms or specific taxonomic groups. Evidently, we still lack a universal platform that systematically integrates and validates the TFs annotated using criteria from a variety of resources.

VARIETY OF TF ANNOTATION METHODS FOR COMPILING TF CATALOGUES

A variety of TF identification methods have been used by different authors. The databases launched before the year 2000 rely on manual literature curation (Figure 1). These databases include RegulonDB (E. coli), DBTBS (B. subtilis) and TRANSFAC (eukaryotes). The other resources that also exclusively contain a compilation of manually curated TFs but were published more recently are PRODORIC and RegTransBase (both for bacteria), FlyTF (fruit fly) and TFcat (human and mouse).

Computational approaches became more popular after the year 2000 when a large number of fully sequenced genomes became available. This is because automated methods allow scientists to identify putative TF sets from completely sequenced genomes within a short
period of time. Sequence-based pair-wise alignment applications such as BLAST (65) were widely used to detect homologues of known DBDs from numerous protein sequences. TFs can then be annotated based on the presence of DBDs in the protein chains. Due to its low sensitivity in remote homologue detection, the technique has become less popular and most of the second-generation databases relying on pair-wise comparisons have now ceased to be developed further (indicated by asterisks in Table 1, last update in 2004 or earlier).

In contrast, most regularly updated databases use profile-based searches such as HMM and PSI-BLAST (66) as opposed to conventional pair-wise alignments such as BLAST. The profile-based approaches have a number of advantages, including the superior sensitivity and specificity in recognizing remote homologues of sequence-specific DBDs (61). These third-generation TF databases include GTOP_TF, RATTF, DBD, PlanTAPDB, PlantTFDB, PfTFDB, cTFbase and ArchaeaTF. The sources of reference query sets and refinement processes after the profile searches vary across the databases. Reference libraries can either be taken directly from existing projects including Pfam (20), SUPERFAMILY (67), PROSITE (68) or newly compiled from the literature and text mining. More recently, a support vector machine (SVM) was used in the ITPF database (59) instead of the profile-based searches for detecting DBDs. Apart from the methods involving sequence comparisons, structural alignment models were designed to identify members of superfamily, which do not contain splice variants. These third-generation TF databases include GTOP_TF, RATTF, DBD, PlanTAPDB, PlantTFDB, PfTFDB, cTFbase and ArchaeaTF. The TF predictions for 683 non-redundant genomes (449 bacteria; 49 archaea; 185 eukaryotes) from the DBD database. To obtain the non-redundant genome set, we minimized the bias due to well-characterized species by manually excluding multiple strains of pathogenic bacteria and fungi, according to species name. These genomes are important from medical and agricultural points of view and thus have been intensively studied. Only the well-studied strains are included. For instance, E. coli K12 and Candida albicans SC5314 were used to represent E. coli and C. albicans, respectively. For eukaryotic genomes, only the longest transcript of each gene is included in this study. We noted that a survey on splice variants across multiple genomes is confounded by the heterogeneity of the data available for different organisms. For instance, mouse is extremely well-characterized, while chimpanzee is not. As a result, alternative splicing was excluded entirely from this study. This also allows the numbers of eukaryotic TFs to be compared with the bacterial TFs, which do not contain splice variants.

We chose TF predictions from the DBD database to illustrate the genome-wide TF repertoires in related and diverse species mainly because all the TF annotation in all the species was performed using a single platform. Although single-species databases such as RegulonDB (E. coli K12) and FlyTF (fruit fly) are more comprehensively and thoroughly curated, inconsistent criteria and methods used by different resources hinder an interspecies comparison. The DBD database is thus most suitable for a cross-lineage TF repertoire comparison. The initial benchmark of the DBD database against the proteins from PDB (70) and UniProt (71) classified by Gene Ontology (72) as TFs showed 95–99% accuracy and 66–67% coverage (61). More recent databases that benchmark their TF predictions against the DBD database include FlyTF, cTFbase and TFcat. Note that, in many cases, the annotation of TFs in one database relies on the annotated TFs from other databases as a starting point. Furthermore, many are based on similar sets of HMMs from Pfam or SUPERFAMILY. That is, TF annotations in different resources are not entirely independent. Consequently, the benchmarks and comparisons of TFs annotated by different databases need to be treated with care.

The TF prediction in the DBD database was performed based on the presence of DBDs (DBDs), scored by two HMM libraries: SUPERFAMILY (67) and Pfam (20). The DBD lists were manually curated and undergo occasional refinement. For SUPERFAMILY, the HMM models were designed to identify members of superfamilies, based on the domain definitions from the Structural Classification of Proteins (SCOP) (73). Since protein domain members in SCOP superfamilies tend to
be functionally diverse, manual curation in the DBD database was done at the SCOP family level instead (61). Moreover, it has been shown that many SCOP families have homologous connections to Pfam families (74). For these reasons, we performed our analysis based on the Pfam and SCOP family DBDs. Although here we only discuss the number of domains and DBDs, assigned to proteins and TFs based on the presence of the Pfam HMMs, it is worth noting that the same observations also hold for SCOP families (see Supplementary Data for SCOP family results).

Table 2 describes the median values of various structural features of proteins and TFs in the three superkingdoms, predicted based on Pfam domain assignment to all proteins in each genome. We observed large variations among the eukaryotic species so we further divided them into three major kingdoms: Metazoa (animal), Fungi and Viridiplantae (plant) in Table 3. These results will be discussed in the next section. The means with standard deviations, SCOP family results and a table containing the numbers of TFs annotated and DBDs in the 683 genomes are available in Supplementary Data.

**TF and DBD repertoires in the major superkingdoms of life**

According to Table 2, eukaryotic species have much larger protein repertoires that contain longer average peptide sequences than bacteria, the superkingdom that dominates the prokaryotic group (medians of 465 and 322 residues, respectively). Note that the *P*-values (*P*), calculated using a non-parametric test (Mann–Whitney), are <10<sup>-15</sup> for all comparisons discussed here unless specified otherwise. The longer eukaryotic proteins also contain more domains per protein chain than bacteria (1.53 versus 1.41 domains per protein). This might allow protein sequences to possess more functionality such as enzymatic properties, DNA binding, as well as binding to other proteins.

The average length of TFs is greater than the average length of all proteins in most eukaryotic species (Table 2). Within the Eukaryota superkingdom, the average length

---

**Table 2. TF repertoires in the three main superkingdoms of life:** Bacteria, Archaea, and Eukaryota

<table>
<thead>
<tr>
<th>Proteins</th>
<th>Bacteria</th>
<th>Archaea</th>
<th>Eukaryota</th>
<th>Cellular organisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteins per species</td>
<td>3140</td>
<td>1966</td>
<td>14141</td>
<td>3885</td>
</tr>
<tr>
<td>Length of all proteins (residues)</td>
<td>322</td>
<td>289</td>
<td>465</td>
<td>328</td>
</tr>
<tr>
<td>Domains assigned per protein</td>
<td>1.41</td>
<td>1.30</td>
<td>1.53</td>
<td>1.42</td>
</tr>
<tr>
<td>Distinct domain families per protein</td>
<td>1.33</td>
<td>1.25</td>
<td>1.29</td>
<td>1.32</td>
</tr>
<tr>
<td>Length of protein domains (residues)</td>
<td>180</td>
<td>171</td>
<td>161</td>
<td>177</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TFs</th>
<th>Bacteria</th>
<th>Archaea</th>
<th>Eukaryota</th>
</tr>
</thead>
<tbody>
<tr>
<td>TFs per species</td>
<td>131</td>
<td>60</td>
<td>325</td>
</tr>
<tr>
<td>Distinct architectures per species</td>
<td>39</td>
<td>19</td>
<td>45</td>
</tr>
<tr>
<td>Length of TFs (residues)</td>
<td>242</td>
<td>196</td>
<td>560</td>
</tr>
<tr>
<td>DBDs per TF</td>
<td>1.04</td>
<td>1.00</td>
<td>1.41</td>
</tr>
<tr>
<td>Distinct DBD families per TF</td>
<td>1.00</td>
<td>1.00</td>
<td>1.01</td>
</tr>
<tr>
<td>Length of DBDs (residues)</td>
<td>62</td>
<td>60</td>
<td>64</td>
</tr>
<tr>
<td>Partner domains per TF</td>
<td>0.58</td>
<td>0.25</td>
<td>0.24</td>
</tr>
<tr>
<td>Distinct partner domain families per TF</td>
<td>0.57</td>
<td>0.24</td>
<td>0.21</td>
</tr>
<tr>
<td>Length of partner domains (residues)</td>
<td>153</td>
<td>97</td>
<td>85</td>
</tr>
<tr>
<td>TF content in genome (%)</td>
<td>4.39</td>
<td>2.94</td>
<td>2.91</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DBDs</th>
<th>Bacteria</th>
<th>Archaea</th>
<th>Eukaryota</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBD families</td>
<td>61</td>
<td>15</td>
<td>77</td>
</tr>
<tr>
<td>Superkingdom-specific DBD families</td>
<td>43</td>
<td>0</td>
<td>69</td>
</tr>
<tr>
<td>Partner domain families</td>
<td>228</td>
<td>55</td>
<td>795</td>
</tr>
<tr>
<td>Superkingdom-specific partner domain families</td>
<td>116</td>
<td>12</td>
<td>693</td>
</tr>
<tr>
<td>Distinct domain architectures</td>
<td>605</td>
<td>118</td>
<td>2209</td>
</tr>
<tr>
<td>DBDs per species</td>
<td>109</td>
<td>42</td>
<td>206</td>
</tr>
<tr>
<td>Distinct DBD families per species</td>
<td>23</td>
<td>12</td>
<td>27</td>
</tr>
</tbody>
</table>

Domain assignments are from Pfam. Median values of all species in each lineage are displayed. Mean values and their SDs for each property are described in ‘Supplementary Data’.

**Table 3. TF repertoires in three major eukaryotic kingdoms:** Viridiplantae (plants), Fungi, and Metazoa (animals), plus all eukaryotes combined

<table>
<thead>
<tr>
<th>Proteins</th>
<th>Viridiplantae</th>
<th>Fungi</th>
<th>Metazoa</th>
<th>Eukaryota</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteins per species</td>
<td>27235</td>
<td>9997</td>
<td>16371</td>
<td>14141</td>
</tr>
<tr>
<td>Length of all proteins (residues)</td>
<td>387</td>
<td>466</td>
<td>479</td>
<td>465</td>
</tr>
<tr>
<td>Domains assigned per protein</td>
<td>1.48</td>
<td>1.47</td>
<td>2.00</td>
<td>1.53</td>
</tr>
<tr>
<td>Distinct domain families per protein</td>
<td>1.24</td>
<td>1.29</td>
<td>1.37</td>
<td>1.29</td>
</tr>
<tr>
<td>Length of protein domains (residues)</td>
<td>158</td>
<td>185</td>
<td>150</td>
<td>161</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TFs</th>
<th>Viridiplantae</th>
<th>Fungi</th>
<th>Metazoa</th>
<th>Eukaryota</th>
</tr>
</thead>
<tbody>
<tr>
<td>TFs per species</td>
<td>591</td>
<td>203</td>
<td>806</td>
<td>325</td>
</tr>
<tr>
<td>Distinct architectures per species</td>
<td>77</td>
<td>38</td>
<td>160.5</td>
<td>45</td>
</tr>
<tr>
<td>Length of TFs (residues)</td>
<td>375</td>
<td>604</td>
<td>545</td>
<td>560</td>
</tr>
<tr>
<td>DBDs per TF</td>
<td>1.13</td>
<td>1.36</td>
<td>2.75</td>
<td>1.41</td>
</tr>
<tr>
<td>Distinct DBD families per TF</td>
<td>1.01</td>
<td>1.00</td>
<td>1.03</td>
<td>1.01</td>
</tr>
<tr>
<td>Length of DBDs (residues)</td>
<td>73</td>
<td>65</td>
<td>56</td>
<td>64</td>
</tr>
<tr>
<td>Partner domains per TF</td>
<td>0.23</td>
<td>0.14</td>
<td>0.40</td>
<td>0.24</td>
</tr>
<tr>
<td>Distinct partner domain families per TF</td>
<td>0.20</td>
<td>0.13</td>
<td>0.35</td>
<td>0.21</td>
</tr>
<tr>
<td>Length of partner domains (residues)</td>
<td>90</td>
<td>83</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>TF content in genome (%)</td>
<td>2.12</td>
<td>2.53</td>
<td>4.65</td>
<td>2.91</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DBDs</th>
<th>Viridiplantae</th>
<th>Fungi</th>
<th>Metazoa</th>
<th>Eukaryota</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBD families</td>
<td>38</td>
<td>34</td>
<td>58</td>
<td>78</td>
</tr>
<tr>
<td>Kingdom specific DBD families</td>
<td>12</td>
<td>6</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>DBDs per species</td>
<td>602</td>
<td>152</td>
<td>2151</td>
<td>206</td>
</tr>
<tr>
<td>Distinct DBD families per species</td>
<td>37</td>
<td>25</td>
<td>53</td>
<td>27</td>
</tr>
</tbody>
</table>

The domain assignments are Pfam families. Median values of all species in each lineage are displayed.
of TFs versus all proteins is 545 versus 479 residues in animals \((P = 10^{-6})\), and 604 versus 466 in fungi. Interestingly, the length of TFs in plants is not significantly different from other proteins on average, 375 versus 387 \((P = 0.4)\) (Table 3). In contrast, TFs in bacteria are significantly shorter than their average proteins. One possible explanation for longer TFs compared to all proteins in animals and fungi is a high fraction of intrinsically disordered (ID) regions, which are absent in bacterial TFs (75,76). These ID segments in proteins are naturally unfolded and unstructured but may serve as flexible linkers that aid protein interactivity (77). Through promoting protein–protein interaction, these long-ID regions aid formation of composite regulatory protein elements in eukaryotes.

A single eukaryotic TF typically contains 1.41 DBDs on average but these DBDs only belong to \(\sim 1\) distinct DBD family per TF. This suggests that many eukaryotic TFs include \(\geq 1\) repeated DBD from the same family. Among the eukaryotic species, the number of DBD repeats is greatest in the animals where DBDs from the same family can occur almost three times on average \((2.75 \text{ DBDs, median})\) in a single TF chain; while fungi \((1.36)\) and plants \((1.13)\) possess significantly fewer DBD repeats per TF. Zinc fingers are among the DBD families that combine with 228 partner domains and give rise to 605 okotypes (29,80).

In both eukaryotic and prokaryotic groups, the number of distinct DBD families per TF is close to 1.00. Indeed the combination of DBDs from more than one family in the same TF chain is extremely rare and restricted to certain phylogenetic groups. For instance, we observed two bacterial DBD families, HTH_AraC and AraC_N, that occur in the same TF but this combination is restricted to proteobacteria. In animals, the HLH DBD mainly appears in single domain TFs but can also combine with two other DBD families, Myc_N and Basic, to form a TF. These are examples of only eight combinations in total (five in eukaryotes and three in bacteria) of multiple DBD families occurring in the same TF chain (79). One possible advantage of having few TFs with multiple DBD families is the minimization of cross-talk between two or more distinct DBDs on the same TF, and a large number of possible binding sites of different DBD families on DNA.

Apart from DBDs, TFs may contain non-DBD domains of different functions, which we hereby call ‘partner domains’. In both prokaryotes and eukaryotes, repeats of the same partner domains within one TF chain are rare. The average number of total domains \(\text{(DBDs plus partner domains)}\) in TFs is greater than in other proteins on average: 1.65 versus 1.53 for eukaryotes \((P = 0.002)\), and 1.62 versus 1.41 for bacteria. However, only in eukaryotes is the number of distinct families per protein fewer in TFs than in other proteins, i.e. 1.22 versus 1.29 \((P = 10^{-6})\), suggesting a higher rate of domain repeats in TFs than in eukaryotic proteins in general.

One might think that the greater number of domains in eukaryotic TFs could be a probable reason that explains why they are longer in sequence than non-TF proteins. Nonetheless, we have shown that this is not the case because bacterial TFs also contain more domains than other proteins on average, but are shorter in length. An alternative explanation would be that the protein domains present in TFs (DBDs and partner domains) are longer than other domains found in non-TF proteins. Interestingly, we demonstrate here, for the first time, that the average length of DBDs does not vary by much across different superkingdoms \(\text{(60–64 amino acid residues, medians)}\). Furthermore, they are significantly shorter than other protein domain families. Thus, the average length of domains, as well as their number of occurrence, cannot explain the longer eukaryotic TFs compared to other proteins. Instead, this is more likely to be due to long stretches of intrinsically disordered regions detected in eukaryotic TFs but not in bacteria as we mentioned earlier.

The average fraction of TFs in genomes (TF content) is highest in animal groups whereas \(\sim 4.7\%\) of proteins are TFs. Fungi and plant genomes possess significantly smaller TF contents of between 2–2.5\%, which surprisingly are less than the average TF fraction in bacteria of 4.2\%. This is because the DBD repertoires in different eukaryotic kingdoms are highly lineage-specific (79), while plant and fungal TF repertoires are less well-characterized than animals. The difference within the animal kingdom is most apparent between vertebrates \(\text{(dominated by Chordata)}\) and invertebrates \(\text{(dominated by Arthropoda)}\). The average number of proteins and TFs per species, as well as TF contents are significantly greater in Chordata than in Arthropoda. This could be a result of whole genome duplication events, a greater rate of segmental duplication in vertebrates, or simply due to better characterized TF catalogues in Chordata.

Unicellular obligate parasites such as apicomplexa and euglenozoa, e.g. \textit{Plasmodium falciparum} (malaria apicomplexan) and \textit{Trypanosoma brucei} (sleeping sickness euglenozoa), contain surprisingly small TF fractions of their genomes. Their entire protein repertoires typically contain only 0.5\% TFs. To illustrate the point, as many as 6\% of human proteins are classified as TFs, as opposed to only 0.3\% in \textit{P. falciparum}. These parasitic organisms have different lifestyles from the other eukaryotes considered here, as they only survive or replicate in a relatively stable environment inside their hosts. The low fraction of predicted TFs in these life forms is likely to be due to their reduced number of proteins and regulatory components, as well as their less well-characterized TF repertoire (29,80).

Bacteria make use of 61 Pfam DBDs in total, which combine with 228 partner domains and give rise to 605 distinct domain architectures (Table 2). When considering them individually, each bacterial organism possesses 131 TFs on average but only 39 distinct architectures (all averages are medians). This corresponds to the
previous finding that the majority of bacterial TFs have arisen through gene duplication events (18,81). Although there are fewer complete genomes available, the eukaryotic superkingdom possesses a large DBD repertoire of 77 distinct families. They combine with 795 partner domain families and form 2209 distinct domain architectures. Besides the larger DBD repertoire, a greater number of partner domains utilized by eukaryotes also plays a part in creating more diverse architectures.

Within the eukaryotic genomes, the Metazoa kingdom possess a considerably larger DBD repertoire than the Fungi and Viridiplantae kingdoms (Table 3). This reflects the greater morphological complexity and number of body structures in animals, as well as a potential bias towards the study of animal model organisms. On average, eukaryotic species possess 325 TFs per genome but make use of only 45 distinct arrangements (medians). This suggests that a large fraction of eukaryotic TFs also emerged through gene duplication (82), possibly even at a higher rate than in bacteria. The result is in accordance with previous work suggesting that as many as 90% of eukaryotic genes have arisen by duplication (83).

The wealth of completely sequenced genomes and automatic TF annotation allow us to predict the TF sets from entire genomes and analyze at a global level the genomic TF repertoires in species from diverse phylogenetic groups. We observe distinct features of TFs and DBDs such as TF length, TF content in genome, number of DBD families and number of DBD repeats per TF in different lineages. During the course of evolution, as genomes expand via gene duplication, a greater number of TFs are required to orchestrate the expression of these expanded genes. In the next section, we will investigate whether or not the TF expansion with respect to the total number of genes is also lineage-specific.

**TF ABUNDANCE FOLLOWS A POWER LAW INCREASE WITH NUMBER OF GENES**

As morphological complexity increases, organisms require a greater proportion of TFs for gene expression control. As well as TFs, cell adhesion molecules and proteins involved in extra-cellular processes have been shown to be greatly expanded in animals (84,85). A power law increase in TF numbers with gene numbers has previously been observed in several bacterial genomes and a very limited number of eukaryotic genomes (15,86–90). In accordance with these previous studies, we do not confirm a linear trend of TF abundance with the number of genes on the log–log scale using the TFs obtained from the DBD database for a large set of bacteria (449 genomes), but also extend this analysis to eukaryotes (185 genomes). This implies a power law relationship between the two variables in both prokaryotic and eukaryotic genomes.

**Power law relationship between TFs and number of genes**

In bacteria, as the number of genes becomes larger, the TF expansion strictly follows a power law increase with an exponent close to 2, which infers a quadratic increase (power law exponent of 1.98, coefficient of determination, $R^2$ of 0.87, Figure 2A). It is worth noting that sigma factors and other non-sequence-specific TFs were not included in this dataset. We observe similar exponents when the numbers of TFs from different bacterial phyla are correlated separately with their number of genes (Supplementary Figure S1). A similar exponential TF expansion can also be seen in eukaryotic genomes but with a slower increase (power law exponent of 1.23) and less fitting quality (lower $R^2$ of 0.61). The exponents >1 observed in both cell types mean that the TF repertoire expands faster than linearly for every gene added to the genome.

Two possible implications of this power law relationship were proposed separately in the context of metabolic networks (89) and microeconomics (88). From the metabolic network point of view, when organisms evolve to explore a new environment, a new set of TFs are required to monitor new tasks necessary to adapt to different conditions. On the other hand, some of the metabolic enzymes can be reused and fewer new ones are required to regulate each task. This may explain why the number of new tasks and their regulators increase faster than linearly with the number of genes encoding enzymes (89).

The necessity of a sharper TF increase with number of genes in bacteria might be linked to the absence of a nucleus and other eukaryote-specific transcriptional mechanisms that might hinder the organisms from having a larger genome. This observation corresponds with a previous study (88) where a microeconomic model was used to speculate that bacteria already have a maximal number of genes, given their transcriptional mechanisms. A further increase in number of genes would be ‘economically’ ineffective since the average cost to regulate a gene becomes prohibitively expensive.

Eukaryotes employ more complex mechanisms for gene expression control compared to bacterial systems that may partly explain the slower TF increase with the number of genes observed. For instance, the high degree of combinatorial regulation in eukaryotes (91) means that TFs are involved in many different multi-protein transcription complexes. The greater fraction of non-coding DNA in the eukaryotic genomes has an important role in producing small RNAs that provide an additional layer of gene regulation. This large amount of non-coding DNA also harbours cis-regulatory sequences with more complex-binding site architecture than in prokaryotes (92–94). Eukaryotic DNA is packaged into chromatin repressed in the transcriptional ground state and the promoter is only accessible in the presence of chromatin remodelling proteins (2). This system also acts as an extra switch for expression control. Although some bacterial chromosome packaging has been observed (95), the system is less well-characterized. Tissue-specific regulatory circuits are another way multi-cellular eukaryotes utilize the same transcriptional associated elements to temporally and spatially control gene expression (96,97). The existence of splice variants in eukaryotes is another possible explanation for the slower TF increase (98). This is,
however, not considered here as only the longest transcript per gene is included, due to the heterogeneity in splice variant datasets as discussed earlier. These machineries unique to eukaryotes together enhance genetic regulation beyond the context of TFs and their target genes.

All eukaryotic obligate parasites have less than 50 TFs (grey dots below the red line in Figure 2A). They are known to be divergent in sequence as well as structure from other eukaryotes (29,80) and less well-characterized. Even when such parasitic species were excluded from the model fitting, the degree of correlation of the linear model in eukaryotes is drastically weaker than in bacteria. The poor correlation could be due to a greater organismal complexity in multi-cellular eukaryotes, which cannot be captured by the total number of proteins alone.

**Slow increase in the number of DBD families reveals TF evolution via gene duplication**

In addition to the total number of TFs per genome, in Figure 2B we illustrate the repertoires of distinct DBD families in each species. Bacterial organisms with larger numbers of genes contain more distinct DBD families. As opposed to the quadratic increase in the total number of TFs with respect to number of genes (power law exponent of 1.98), the increase in number of distinct DBD families recruited by larger prokaryotic organisms is close to linear (power law exponent of 1.00, $R^2 = 0.71$). From this finding, we can infer that the number of TFs per DBD family gradually increases as the total number of genes grows larger. Most likely the TFs belonging to the same DBD family have arisen through multiple gene duplication events followed by a series of protein sequence divergence and domain re-combination events. This has also been shown previously (18,81).

In contrast to bacteria, we did not observe any clear correlation between the number of distinct DBD families and the number of genes per genome in eukaryotes. When we performed a linear regression separately for different eukaryotic kingdoms, we found that the number of unique eukaryotic DBD families per species is relatively conserved in all animals, regardless of the number of genes, and the same is true for fungi (power law exponents are 0.09 and 0.13 for animals and fungi, respectively). This suggests there might be a minimal requirement of DBD diversity in animals, and similarly in fungi. Evidently, there seem to be at least two major bursts of DBD family expansion: the first when eukaryotes branched off from prokaryotes and the second at the common ancestral node of animals and fungi. This gives rise to the unique set of DBD repertoires in the eukaryotic kingdoms, which reflects the organismal complexity and morphological diversity of the two lineages. As discussed previously, some eukaryotic DBD families can occur repeatedly within the same TF chain. This may also allow eukaryotes to boost the length and diversity of DNA-recognition sequence without recruiting additional DBD families.

Apart from the lineage-specific increase in TFs with the total number of genes, as well as the distinct structural features of TFs and DBDs in different lineages we discussed in the previous section, we also noted...
lineage-specific presence and absence of DBD families, and their domain combinations with partner domains. We discuss this analysis in greater detail in a separate article (79). Examples of the most frequently observed lineage-specific Pfam DBD families and their architectures in TFs are shown in Figure 3.

CONCLUSIONS AND PERSPECTIVE

Sequence-specific TFs are a vital class of proteins because they directly bind to DNA and thus regulate differential expression of genes. Importantly, they determine physiological diversity of organisms in different lineages across the tree of life (99). We survey and summarize TF resources currently available, as well as discuss the criteria and methods used to annotate TFs by different authors. Comprehensive and high quality TF catalogues serve as a starting point for the experimental design of high-throughput studies on TF–DNA interactions (8–12), as well as being a platform for protein engineering such as in the emerging field of synthetic biology, e.g. engineered zinc finger nucleases (100,101).

As we are moving into the post-genomic era, computational tools have been employed more often to help TF prediction for a large number of completely sequenced genomes. Exclusive literature curation of entire genomes is still available for a small number of model organisms, e.g. FlyTF for fly, and TFCat for human and mouse. Over 25 databases providing genomic TF catalogues have become available over the past 15 years, however, most of them focus on a small number of model organisms or specific taxonomic groups. A universal platform that systematically integrates and validates TFs annotated using different criteria from different TF databases would be of great benefit to the community, in a similar way to InterPro (32) that integrates protein families, domains and functional sites from other protein databases such as Pfam (20), SUPERFAMILY (67) and PROSITE (68).

We used TF annotations obtained from the DBD database to demonstrate the current knowledge on the global TF repertoire in ~700 genomes across the tree of life. TF catalogues of different species in the DBD database are automatically annotated based on consistent criteria and this eliminates the biases due to different methods of TF annotation. Although the biases due to different levels of knowledge of TF repertoires in diverse lineages remain (model organisms are better studied), the datasets we use can serve as representative examples for summarizing the community’s current understanding of the genomic TF abundance and structural complexity.

We observe several features of TF families and their protein domain architectures unique to specific lineages,
most apparently between prokaryotic and eukaryotic genomes. Firstly, the eukaryotic TFs are significantly longer than eukaryotic proteins of other functions while this relationship is reversed in prokaryotes. This could be due to the presence of long intrinsic disordered segments in eukaryotic TFs that are required to leverage the formation of multi-protein transcription complexes (76). Second, repeats of the same DBD family in one polypeptide chain are common only in eukaryotes. This has been suggested as one mechanism used by eukaryotes to increase the length and diversity of DNA-binding recognition sequence from a limited number of DBD families (78). It also potentially explains why the number of unique DBD families keeps increasing when the total number of prokaryotic genes grows larger, while the abundance of DBD family repertoires seem to be relatively conserved in animals and fungi.

We not only confirmed the quadratic increase in TFs with the number of genes in prokaryotes observed by previous studies (92–94), but also extended the model fitting to a large group of eukaryotic species. We observe a similar exponential TF expansion in eukaryotic genomes but with a lower exponent and fitting quality than bacteria. We speculate that this may be due to the complex mechanisms for gene expression control utilized only by eukaryotes such as a greater fraction of regulatory non-coding DNA, combinatorial regulation of multiple TFs and chromatin repressed transcriptional ground state.

We demonstrate the lineage-specific structural features, distinct rates of increase with respect to the total number of genes of TFs and DBD families in different lineages. In addition, we observe distinct patterns of DBD family expansion and their domain combinations with partner domains in diverse phylogenetic groups. The lineage-specific characteristic of DBD families and TF architectures can be used as signatures for the genetic regulatory circuits, which can improve methods for remote homology detection and thus the discovery of new TFs in genomes. Coin and coworkers (102,103) have shown that techniques along these lines can be used to enhance protein domain discovery.

**SUPPLEMENTARY DATA**

Supplementary Data are available at NAR Online.

**ACKNOWLEDGEMENTS**

We thank Daniel Hebenstreit, Joseph Marsh, Anuphon Laohavisit, as well as anonymous referees for critical commentary on the manuscript.

**FUNDING**

Medical Research Council, and a Royal Thai Government Scholarship (to V.C.). Funding for open access charge: Medical Research Council.

**Conflict of interest statement.** None declared.

**REFERENCES**


