Nucleic Acids Research, 2003, Vol. 31, No. 2 **653–660**
DOI: 10.1093/nar/gkg156

The phylogenetic diversity of eukaryotic transcription

Richard M. R. Coulson* and Christos A. Ouzounis

Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK

Received September 16, 2002; Revised October 30, 2002; Accepted November 18, 2002

ABSTRACT

Eukaryotic transcription is a highly regulated process involving interactions between large numbers of proteins. To analyse the phylogenetic distribution of the components of this process, six crown eukaryote group genomes were queried with a reference set of transcription-associated (TA) proteins. On average, one in 10 proteins encoded by these genomes were found to be homologous to sequences in the reference set. Analysis of families identified using an accurate sequence clustering algorithm and containing both TA proteins and eukaryotic sequences showed that in two-thirds of the families the homologues originate from a single kingdom. Furthermore, in only 15% of the fungal-specific clusters are the homologues present in both budding and fission yeast, as compared with the metazoan-specific clusters where 53% of the homologues originate from two or more species. Families whose members comprise general transcription factor or RNA polymerase subunits exhibit a low degree of taxon specificity, suggesting that the transcription initiation complex is highly conserved. This contrasts with transcriptional regulator families, that are primarily taxon-specific, indicating proteins controlling gene activation exhibit considerable sequence diversity across the eukaryotic domain.

INTRODUCTION

Mechanisms controlling transcriptional activation are fundamentally different between prokaryotes and eukaryotes (1). In eukaryotes, protein coding genes are usually transcribed by RNA polymerase II (RNAP II) and typically contain common core-promoter elements recognised by general transcription initiation factors (GTFs) and gene-specific DNA elements recognised by regulatory factors (2). Initiation of transcription requires assembly of a pre-initiation complex (PIC); this assembly is nucleated by binding of the TATA-box binding polypeptide subunit of TFIID (TBP). After TBP binding, there follows concerted recruitment of TFIIB, a complex of RNAP II with TFIIF, TFIIE and TFIIH (3). The exact timing of PIC formation varies at different promoters, as do the requirements for chromatin-remodelling events and specific histone modifications (4). Although the general mechanistic principles of gene regulation in eukaryotes are well established (5), the phylogenetic diversity of the transcriptional machinery has not yet been comprehensively explored in the light of entire eukaryotic genome sequences.

Previously, four entire archaeal genomes were profiled using a set of known transcription-associated (TA) proteins, extracted from the protein sequence databases via keyword searches (6). This showed that transcription in Archaea is highly divergent, as represented by the presence of substantially different sets of TA protein homologues in the four species examined. Subsequently, the full clustering of all known TA proteins showed TA protein families to be primarily taxon-specific up to the domain level, with very little sharing between Archaea, Bacteria and Eukarya (7). Furthermore, ~45% of *Arabidopsis thaliana* transcription factors were found to belong to families that are specific to plants (8). To assess the conservation of proteins participating in the eukaryotic transcriptional process, a TA protein sequence reference set was used to identify homologues in the entire genome sequences of six species of the crown eukaryote group: *Schizosaccharomyces pombe* (9), *Saccharomyces cerevisiae* (10), *A.thaliana* (11), *Caenorhabditis elegans* (12), *Drosophila melanogaster* (13) and *Homo sapiens* (14,15). TA proteins from both eukaryotes and prokaryotes and their homologues in the above species were further considered. The degree of divergence and species distribution of proteins involved in the eukaryotic transcriptional process was then assessed by full clustering of both the TA proteins and their identified homologues in the six species examined.

MATERIALS AND METHODS

Extraction of the reference set

TA proteins were obtained using the DESCRIPTION and KEYWORD records of Swiss-Prot and the DESCRIPTION record of SP-TrEMBL (16) containing the word 'transcription', as previously described (7). Given the high degree of manual curation of the Swiss-Prot database, this keyword-based extraction step recovers a wide range of proteins involved in all aspects of transcription. Data extraction, linking and indexing was performed by the Sequence Retrieval System (SRS), version 6.0 and the Icarus language (17). All data were stored in a MySQL relational database system.

*To whom correspondence should be addressed. Tel: +44 1223 494417; Fax: +44 1223 494471; Email: coulson@ebi.ac.uk

Table 1. BLAST search using the TA protein reference set against six eukaryotic genomes

Species	Size[a]	TA homologues	% Genome	TA proteins[b]	% TA data set
H.sapiens	29 304	3471	11.8	4519	45.8
D.melanogaster	13 710	1370	10.0	4277	43.3
C.elegans	19 099	1504	7.9	3711	37.6
A.thaliana	27 406	3273	11.9	2870	29.1
S.cerevisiae	6292	605	9.6	2257	22.9
S.pombe	4882	525	10.8	2154	21.8

Columns: Size, number of protein sequence entries; TA homologues, number of TA protein homologues identified using BLAST; % Genome, percentage of the genome entries identified as TA homologues; TA proteins, sequences in the TA protein reference set having a homologue in the corresponding species; % TA data set, percentage of the TA reference set with a homologue in the corresponding species. Table 1 is sorted using the values of the last two columns.
[a]No. of database targets: 100 693.
[b]No. of query sequences: 9874 (869 678 hits were obtained with an *E*-value cut-off $\leqslant 10^{-6}$).

Sequence comparison

The TA protein sequence reference set was filtered for composition bias using CAST (18) before being used to search against the six complete eukaryotic genomes. The search was performed using BLAST version 2.0 (19) and sequence similarities with an *E*-value threshold $\leqslant 10^{-6}$ were considered as significant. Only TA proteins with eukaryotic homologues were then further considered.

Validation

330 sequences in the *Saccharomyces* Genome Database (SGD) (20) contain 'transcription' in their Gene Ontology (GO) annotations (21) and 74% (244) of these sequences have significant matches to the TA protein reference set, corresponding to the majority of known transcription factors. The remaining 86 sequences belonging to the GO class 'transcription' have not been identified as homologues. Of the 605 budding yeast TA protein homologues identified (Table 1), 329 are assigned to other GO classes (mostly related with gene expression) and 32 are unassigned. This suggests that the BLAST searches of complete genomes performed (Table 1) identify the majority of the members of TA protein families within the six eukaryotic genomes. Furthermore, annotations from the TRANSFAC database of transcription factors (22) were used to validate the clustering (see below).

Sequence clustering

Eukaryotic homologues and TA proteins from the reference set were selected for further clustering, using TRIBE-MCL (23), at inflation values 2.0 and 5.0. The TRIBE-MCL clustering algorithm has been used to detect families of related protein sequences on the basis of sequence similarity. This algorithm uses the Markov Clustering aLgorithm (MCL) for graph flow simulation, operating on all-against-all matrices containing similarity values obtained by fast database search algorithms, such as BLAST. Thus, TRIBE-MCL avoids expensive sequence comparison operations, used to delineate the consistency of sequence similarity searches as previously shown (24). TRIBE-MCL is ideally suited to cluster very large sequence data sets, such as the one described herein. Only clusters containing both TA proteins and eukaryotic homologues were further considered (Table 2).

Table 2. Distribution of protein families containing TA proteins and their eukaryotic homologues

Distribution	2.0		5.0	
	#	%	#	%
Viridiplantae	90	12.1	120	13.9
Fungi	178	23.9	171	19.8
Metazoa	249	33.4	378	43.7
Specific to a kingdom	(517)	69.3	(669)	77.3
Fungi::Metazoa	34	4.6	39	4.5
Fungi::Viridiplantae	13	1.7	10	1.2
Metazoa::Viridiplantae	42	5.6	46	5.3
Fungi::Metazoa::Viridiplantae	140	18.8	101	11.7
Shared between kingdoms	(229)	30.7	(196)	22.7
TA proteins and eukaryotic families[a]	746	100.0	865	100.0
TA-protein-only families[b]	17		40	
Eukaryotic-only families	188		418	
Total families[c]	951		1323	

Columns: Distribution, kingdom distribution of the identified families; 2.0 and 5.0, inflation values controlling cluster granularity in the two clustering operations; # and %, the absolute and relative percentage values for the families with members in TA proteins and the six eukaryotic genomes.
[a]Analysed in Figure 1.
[b]These families were not further considered because they did not meet the clustering criteria.
[c]No. of sequences clustered: 16 568 (10 748 eukaryotic and 5820 TA proteins).

All data are available at: http://www.ebi.ac.uk/research/cgg/transcription/crown_eukaryotes/.

RESULTS

To estimate the proportion of each eukaryotic genome that encodes TA proteins, 9874 known TA protein sequences were extracted from Swiss-Prot and SP-TrEMBL (16) and subsequently used as queries to search the entire genomes (100 693 sequences in total) of the six aforementioned species. In total, 5820 TA proteins identify 10 748 eukaryotic homologues. The TA proteins without eukaryotic homologues correspond to archaeal and bacterial sequences. The percentage of the genome that encodes for TA proteins is on average 10%, with some slight variations across species (Table 1). Interestingly,

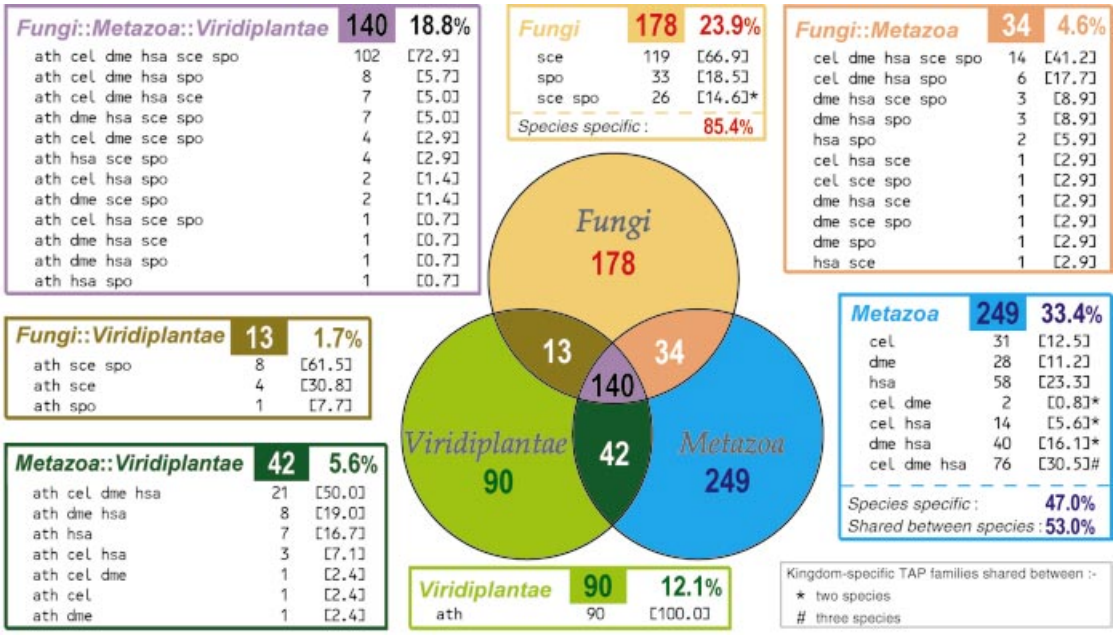


Figure 1. Taxonomic distribution of TRIBE-MCL families. The sectors of the Venn diagram represent the seven possible eukaryotic kingdom groupings from which TA protein homologues within a cluster could derive. The number shown in each sector is the number of low granularity clusters in which the constituent homologues originate from the kingdom/s that comprise the grouping. The panels show (except for the one bounded in grey) the number of times a particular species distribution of homologues is observed within a kingdom grouping—the figure in square brackets is the percentage of clusters within the grouping that show the pattern. The colour of the box (displaying the number of clusters in the group) in these panels is specific for a particular group of kingdoms and is identical in colour to the corresponding sector of the Venn diagram. The percentage next to the box is the proportion of clusters out of the total number of clusters (746) that fall into the grouping.

there is more variation within the animal kingdom, ranging from 8% for worms to 12% for humans, compared with the fungi.

The degree of conservation between the eukaryotic TA homologues was assessed by clustering a data set consisting of the TA proteins with homologues in the six target genomes (5820 sequences) and their corresponding homologues (10 748 sequences). Clustering was performed using TRIBE-MCL, an efficient algorithm for the large-scale detection of protein families, previously used to identify protein families within the draft human genome (23). Two clustering operations were performed with inflation values of 2.0 and 5.0 to produce clusters of different granularity levels. Inflation value determines cluster granularity, with lower values producing larger clusters, containing more distantly-related sequences. The clustering operation performed at the lower inflation value generates 951 families of which 78% contain both eukaryotic sequences and TA proteins (Table 2), with the remaining proportion representing either clusters of very remote sequences or false positive BLAST hits (data not shown). At the higher inflation value, 65% of the 1323 families similarly contain sequences from both sets (Table 2). The precision of the clustering algorithm TRIBE-MCL has been estimated to lie between 87 and 98% for InterPro families (using two different criteria), and over 80% for inflation value 2 and up to 87% for inflation value 5 with SCOP families (23).

Complete genome sequences used in this analysis provide a representative view of the TA protein content of a genome, as opposed to just the inclusion of all available sequences from the database where biases may occur. At both granularity levels, the majority of protein families contain eukaryotic sequences that originate from a single kingdom, while only 30% of these families are shared across any of the three eukaryotic kingdoms (Table 2). This observation suggests that the sharing of TA proteins in the eukaryotic domain is limited, consistent with previous observations (7,8). It is also striking that for the families shared by at least two different kingdoms, more than half of them are present in all three kingdoms, suggesting that once a TA protein family is shared, it is more likely to be widely spread within the eukaryotic domain (Table 2).

To investigate the species specificity of eukaryotic transcription, the clusters of low granularity (encompassing the higher number of protein family members) were further analysed (Fig. 1). Only clusters containing at least one TA protein and at least one eukaryotic sequence were selected. This criterion was necessary in order to obtain reliable descriptions of the clusters, supported by functional annotations from the Swiss-Prot database records. For families shared between kingdoms, the most common species distributions are the ones in which all species are represented. For example, out of 140 families shared between all three kingdoms, 102 (73%) contain sequences from all six species (Fig. 1, top left-hand panel). This pattern is similar to the one observed at the kingdom level, suggesting that if a TA protein family is shared across kingdoms, then it is likely to have members present in all the species within the kingdom.

In the kingdom-specific clusters, the pattern of TA protein family sharing is less pronounced with 30.5% of the families present in all three metazoan species (Fig. 1, bottom right-

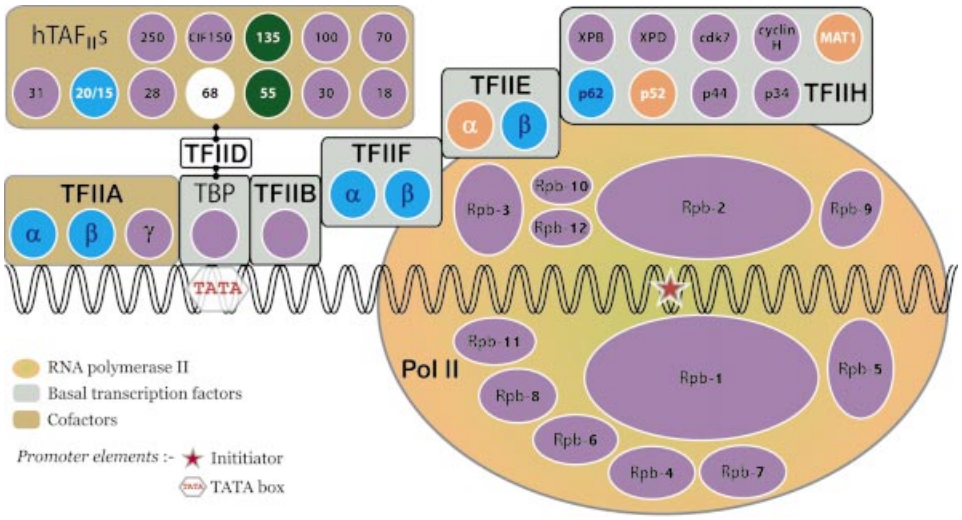


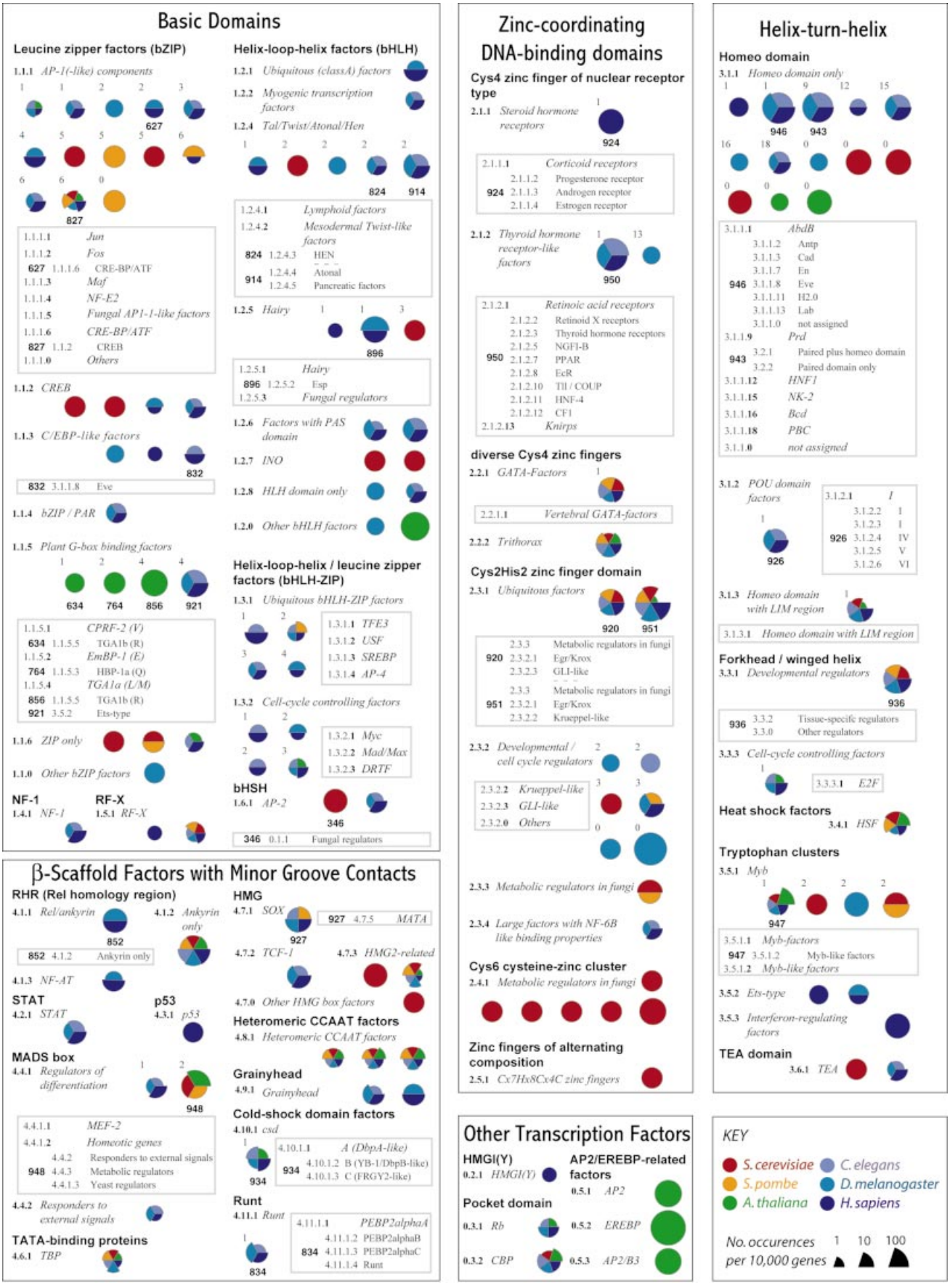
Figure 2. Phylogenetic diversity of human GTF subunit families. The 12 subunits of RNAP II (Rpb-1 to -12) are indicated as white-bounded ovals. The subunit components of the basal transcription factors (TBP and TFII-B, -E, -F, -H) and cofactors (TFIIA and hTAF$_{II}$s) are shown as white-bounded circles. The colour used to fill the circles and ovals indicates the kingdom origins of the homologues within these human subunit-containing clusters (see Fig. 1 for the colour codes). hTAF$_{II}$68 is displayed in white as the sequence is absent from the analysis because its corresponding Swiss-Prot entry does not meet the TA protein extraction criteria. There are only two clusters encoding the three subunits of TFIIA, as TFIIAα and TFIIAβ are produced post-translationally from the same precursor polypeptide. The total number of clusters encompassing the proteins shown is 40.

hand panel) and only 14.6% present in both fungal species (Fig. 1, top middle panel). Additionally, the degree of species specificity is markedly different for yeasts and animals: 85.4% of the fungal-specific TA protein families are present either in *S.pombe* or *S.cerevisiae*, whereas only 47% of the metazoan families are restricted to a single species. Hence, it appears that the degree of sharing of kingdom-specific TA protein families is related to the evolutionary distance between the species within a kingdom and not their level of cellular complexity. In summary, 33.4, 23.9 and 12.1% of the TA protein families, contain homologues derived solely from metazoa, fungi and viridiplantae, respectively (Fig. 1). The comparatively low percentage of viridiplantae-specific clusters may result from the relative lack of experimental data on plant transcription and fewer plant-specific proteins in the TA protein reference set.

The amount of taxon sharing was then investigated in TA protein clusters containing either GTF components or regulatory factors. Human sequences present in the extracted TA protein sequence set that participate in transcription initiation and mRNA synthesis were identified by their gene description entries. Analysis of the taxonomic distribution of the TA protein homologues within the low granularity (larger) clusters containing these human sequences, showed that the 12 families containing the subunits of human RNAP II also contain homologues from all three kingdoms (Fig. 2). However, in the 15 families containing the human proteins that comprise the 'basal' transcription factors (TBP, TFIIB, TFIIE, TFIIF and TFIIH), only in 53% of the clusters do the homologues originate from all three eukaryotic kingdoms (Fig. 2, silver panel). In three of these families, the homologues are observed only in metazoa and fungi, and in four of them they are metazoan-specific. The homologues present in 69% of the 13 families containing the TFIIA subunits and the coactivator subunits of TFIID, TBP-associated factors (hTAF$_{II}$s), are detected in all of the three kingdoms (Fig. 2, gold panel). Of the above 40 families, none are species-specific: for any *H.sapiens* sequence belonging to one of the clusters, there exist homologues either in the other metazoan species or even in other kingdoms. This indicates the PIC is highly conserved in eukaryotes.

The degree of taxon restrictedness of regulatory proteins was determined by analysing the phylogenetic distribution of the eukaryotic sequences present in low granularity families whose TA protein members are also present in the TRANSFAC database of transcription factors (22). This procedure identified 136 clusters of which 125 (92%) are linked to only one TRANSFAC family (Fig. 3). In only 21%

Figure 3. (Opposite) Association of TRANSFAC transcription factor classifications with TA protein homologues. The five TRANSFAC superclasses (labelled top-centre in the black framed boxes) are divided into classes and the classes are divided into families. The decimal classification number and description of a family is listed if any of its members are also present in the low granularity TRIBE-MCL clusters. If these TRANSFAC families are further divided into subfamilies, the last digits of the subfamily decimal classification number is displayed top-left of the pie chart. The description and full classification number (last digits are in bold) of the subfamily is shown in the grey rectangle. The colour of each segment of the pie chart represents the species from which the TA protein homologues in the cluster originate (see KEY box for colour codes). The radius of each segment is proportional to the number of occurrences of these homologues normalised by genome size. The bottom of the KEY box shows the relationship between the normalised number of occurrences and segment radius. Pie charts with TRIBE-MCL cluster identifiers beneath them indicate that TA proteins in the cluster are present in other TRANSFAC families and/or subfamilies. The classifications of these families and subfamilies, along with the cluster identifier, are noted in the grey rectangles.


Basic Domains
Leucine zipper factors (bZIP)
1.1.1 AP-1(-like) components
1 1 2 2 3
627
4 5 5 5 6
6 6 0
827
1.1.1.1 Jun
1.1.1.2 Fos
627 1.1.1.6 CRE-BP/ATF
1.1.1.3 Maf
1.1.1.4 NF-E2
1.1.1.5 Fungal AP1-1-like factors
1.1.1.6 CRE-BP/ATF
827 1.1.2 CREB
1.1.1.0 Others
1.1.2 CREB
1.1.3 C/EBP-like factors
832
832 3.1.1.8 Eve
1.1.4 bZIP / PAR
1.1.5 Plant G-box binding factors
1 2 4 4
634 764 856 921
1.1.5.1 CPRF-2 (V)
634 1.1.5.5 TGA1b (R)
1.1.5.2 EmBP-1 (E)
764 1.1.5.3 HBP-1a (Q)
1.1.5.4 TGA1a (L/M)
856 1.1.5.5 TGA1b (R)
921 3.5.2 Ets-type
1.1.6 ZIP only
1.1.0 Other bZIP factors
NF-1
1.4.1 NF-1
RF-X
1.5.1 RF-X
Helix-loop-helix factors (bHLH)
1.2.1 Ubiquitous (classA) factors
1.2.2 Myogenic transcription factors
1.2.4 Tal/Twist/Atonal/Hen
1 2 2 2 2
824 914
1.2.4.1 Lymphoid factors
1.2.4.2 Mesodermal Twist-like factors
824 1.2.4.3 HEN
1.2.4.4 Atonal
914 1.2.4.5 Pancreatic factors
1.2.5 Hairy
1 1 3
896
1.2.5.1 Hairy
896 1.2.5.2 Esp
1.2.5.3 Fungal regulators
1.2.6 Factors with PAS domain
1.2.7 INO
1.2.8 HLH domain only
1.2.0 Other bHLH factors
Helix-loop-helix / leucine zipper factors (bHLH-ZIP)
1.3.1 Ubiquitous bHLH-ZIP factors
1 2 3 4
1.3.1.1 TFE3
1.3.1.2 USF
1.3.1.3 SREBP
1.3.1.4 AP-4
1.3.2 Cell-cycle controlling factors
1 2 2 3
1.3.2.1 Myc
1.3.2.2 Mad/Mxx
1.3.2.3 DRTF
bHSH
1.6.1 AP-2
346
346 0.1.1 Fungal regulators
β-Scaffold Factors with Minor Groove Contacts
RHR (Rel homology region)
4.1.1 Rel/ankyrin
852
4.1.2 Ankyrin only
852 4.1.2 Ankyrin only
4.1.3 NF-AT
STAT
4.2.1 STAT
p53
4.3.1 p53
MADS box
4.4.1 Regulators of differentiation
1 2
948
4.4.1.1 MEF-2
4.4.1.2 Homeotic genes
4.4.2 Responders to external signals
948 4.4.3 Metabolic regulators
4.4.1.3 Yeast regulators
4.4.2 Responders to external signals
TATA-binding proteins
4.6.1 TBP
HMG
4.7.1 SOX
927
927 4.7.5 MATA
4.7.2 TCF-1
4.7.3 HMG2-related
4.7.0 Other HMG box factors
Heteromeric CCAAT factors
4.8.1 Heteromeric CCAAT factors
Grainyhead
4.9.1 Grainyhead
Cold-shock domain factors
4.10.1 csd
1
934
4.10.1.1 A (DbpA-like)
934 4.10.1.2 B (YB-1/DbpB-like)
4.10.1.3 C (FRGY2-like)
Runt
4.11.1 Runt
1
834
4.11.1.1 PEBP2alphaA
4.11.1.2 PEBP2alphaB
834 4.11.1.3 PEBP2alphaC
4.11.1.4 Runt
Zinc-coordinating DNA-binding domains
Cys4 zinc finger of nuclear receptor type
2.1.1 Steroid hormone receptors
1
924
2.1.1.1 Corticoid receptors
2.1.1.2 Progesterone receptor
924 2.1.1.3 Androgen receptor
2.1.1.4 Estrogen receptor
2.1.2 Thyroid hormone receptor-like factors
1 13
950
2.1.2.1 Retinoic acid receptors
2.1.2.2 Retinoid X receptors
2.1.2.3 Thyroid hormone receptors
2.1.2.5 NGFI-B
950 2.1.2.7 PPAR
2.1.2.8 EcR
2.1.2.10 Tll / COUP
2.1.2.11 HNF-4
2.1.2.12 CF1
2.1.2.13 Knirps
diverse Cys4 zinc fingers
2.2.1 GATA-Factors
1
2.2.1.1 Vertebral GATA-factors
2.2.2 Trithorax
Cys2His2 zinc finger domain
2.3.1 Ubiquitous factors
920 951
2.3.3 Metabolic regulators in fungi
920 2.3.2.1 Egr/Krox
2.3.2.3 GLI-like
2.3.3 Metabolic regulators in fungi
951 2.3.2.1 Egr/Krox
2.3.2.2 Krueppel-like
2.3.2 Developmental / cell cycle regulators
2 2
3 3
0 0
2.3.2.2 Krueppel-like
2.3.2.3 GLI-like
2.3.2.0 Others
2.3.3 Metabolic regulators in fungi
2.3.4 Large factors with NF-6B like binding properties
Cys6 cysteine-zinc cluster
2.4.1 Metabolic regulators in fungi
Zinc fingers of alternating composition
2.5.1 Cx7Hx8Cx4C zinc fingers
Helix-turn-helix
Homeo domain
3.1.1 Homeo domain only
1 1 9 12 15
946 943
16 18 0 0 0
0 0 0
3.1.1.1 AbdB
3.1.1.2 Antp
3.1.1.3 Cad
3.1.1.7 En
946 3.1.1.8 Eve
3.1.1.11 H2.0
3.1.1.13 Lab
3.1.1.0 not assigned
3.1.1.9 Prd
943 3.2.1 Paired plus homeo domain
3.2.2 Paired domain only
3.1.1.12 HNF1
3.1.1.15 NK-2
3.1.1.16 Bcd
3.1.1.18 PBC
3.1.1.0 not assigned
3.1.2 POU domain factors
1
926
3.1.2.1 I
3.1.2.2 I
3.1.2.3 I
926 3.1.2.4 IV
3.1.2.5 V
3.1.2.6 VI
3.1.3 Homeo domain with LIM region
1
3.1.3.1 Homeo domain with LIM region
Forkhead / winged helix
3.3.1 Developmental regulators
936
936 3.3.2 Tissue-specific regulators
3.3.0 Other regulators
3.3.3 Cell-cycle controlling factors
1
3.3.3.1 E2F
Heat shock factors
3.4.1 HSF
Tryptophan clusters
3.5.1 Myb
1 2 2 2
947
3.5.1.1 Myb-factors
947 3.5.1.2 Myb-like factors
3.5.1.2 Myb-like factors
3.5.2 Ets-type
3.5.3 Interferon-regulating factors
TEA domain
3.6.1 TEA
Other Transcription Factors
HMGI(Y)
0.2.1 HMGI(Y)
Pocket domain
0.3.1 Rb
0.3.2 CBP
AP2/EREBP-related factors
0.5.1 AP2
0.5.2 EREBP
0.5.3 AP2/B3
KEY
S. cerevisiae
C. elegans
S. pombe
D. melanogaster
A. thaliana
H. sapiens
No. occurences per 10,000 genes 1 10 100

(28) of the families do the TA protein homologues originate from multiple kingdoms and of these families, almost half (13) contain homologues present in all three kingdoms (with eight of these families containing sequences present in all six species). The remaining 108 (79%) of the TRANSFAC-linked TA protein families are identified as kingdom-specific: nine (6%) have homologues found uniquely in plants, 30 (22%) in yeasts and 69 (51%) in animals, possibly reflecting some bias in the experimental analyses of transcription in these organisms. The degree of sharing of the kingdom-specific homologues across species is markedly different for fungi and metazoa. Two-thirds (46) of the metazoan-specific families are shared between at least two species, whereas only one tenth (3) of the fungal-specific families occur in both yeasts. Hence, the levels of taxon specificity observed for families containing well-characterised eukaryotic transcription factors are similar to those observed for the entire data set (Fig. 1).

The levels of abundance of the TA protein homologues (present in the TRANSFAC-linked clusters; Fig. 3) within the eukaryotic genomes were assessed by comparing the number of occurrences of the homologues, normalised by genome size, per 10 000 genes. The kingdom-specific TA protein homologues are present in each of the genomes on average 3.4 ($\pm$6.2) times in every 10 000 genes, whereas the homologues present in multiple kingdoms are nearly twice as abundant (6.7 $\pm$ 19.3). Additionally, the variation in abundance of the kingdom-shared families is 3.1 times higher. For example, in the three clusters that contain members of the heteromeric CCAAT factor family (Fig. 3, Family 4.8.1), the levels of occurrence of the homologues range from 0.5 to 4.7 per 10 000 genes. However, for a C2H2-zinc finger family (Fig. 3, Family 2.3.1) the range is from 0.4 to 185.6 and for a MYB family 1.0 to 55.5 (Fig. 3, Family 3.5.1), per 10 000 genes. This increased range in variation primarily results from the relative under- and over-representation, respectively, of the plant homologues in these clusters (Fig. 3).

The percentage of metazoan genomes encoding TA protein homologues increases as the level of cellular complexity of the organism increases (Table 1). To examine if this trend is also observed with transcriptional regulators, the percentage of the genome that encodes the TA protein homologues present in the 51 TRANSFAC-linked families containing sequences from all three animal species was determined. This analysis revealed that the trend is similar with 1.7% of the worm genome encoding sequences homologous to TRANSFAC entries and 3.1% of the fly and 4.4% of the human genomes encoding such sequences. This absolute increase of 2.7% from worms to humans accounts for 69.2% of the overall increase of 3.9% in the proportion of the human genome encoding TA proteins as compared with the worm genome, possibly also reflecting the extent of experimental work on gene expression in vertebrates (Table 1). These increases are also observed in Figure 3, where the average difference between the number of *H.sapiens* and *C.elegans* sequences (normalised by genome size) in a family is 186.4%. The average difference between *D.melanogaster* and *C.elegans* is 93.3% and between *H.sapiens* and *D.melanogaster*, 76.8%. Thus, the increase in the number of homologues of these DNA-binding proteins within the genomes of these species appears to correlate with their increasing cellular diversity.

DISCUSSION

A salient feature of the protein families identified by the clustering procedure described above, is the degree of taxon specificity displayed by the TA protein homologues within the families. A prior expectation based on previous studies (7,8) of the outcome of this analysis may have been that a ~10% of families would contain homologues present in all three eukaryotic kingdoms. Of the remaining families, a minority would be specific to unicellular organisms with the majority being specific to multicellular organisms, with a large proportion of the animal homologues being species-specific. Figure 1 shows 14% of the low granularity clusters contain homologues present in all six eukaryotes (in agreement with the above expectation), implying that proteins closely related to these sequences are widely dispersed throughout the eukaryotic domain. In 69% of the families identified, the TA protein homologues within a family are unique to one of the three kingdoms. As these data are based on sequences identified by complete genome searches, this suggests that the sharing of TA proteins at the eukaryotic kingdom level is limited and taxon-specific, although the number of complete eukaryotic genomes is still limited. However, the degree of species specificity differs markedly in the fungal and animal kingdoms: 53% of metazoan-specific clusters contain homologues present in multiple species and this contrasts greatly with the fungal-specific clusters in which only 15% of them contain homologues shared between both yeasts. Given the similarities of the biologies of budding and fission yeast (25) in comparison with the very substantial differences in the levels of cellular diversification between nematodes and vertebrates, the observation that fungal TA protein families show greater species specificity than do the metazoan TA protein families is unexpected.

The taxonomic origins of the homologues present in the clusters whose TA protein members play a role in transcription initiation are far more uniform than the origins of the regulatory factor homologues. Homologues from all three eukaryotic kingdoms are represented in each of the 12 human RNAP II subunit families, suggesting that RNAP II is highly conserved across the eukaryotic domain. In addition, 53% of families encoding basal transcription factor components and 69% encoding transcriptional cofactors contain homologues originating from the three eukaryotic kingdoms, implying that proteins encoding these functions show higher rates of evolution than do the subunits of RNAP II. This increased level of divergence may be expected of the cofactors as they mediate gene-specific transcription, but not of the basal components whose role in transcription initiation is generic. In only 9% of the well-characterised transcriptional regulator families (i.e. those containing TA proteins linked to TRANSFAC) do the homologues originate from the three kingdoms. The percentage of metazoan-specific, regulator families in which the homologues originate from two or more species is 67% (compared with 53% for the entire data set), whereas in only 10% (15% overall) of the fungal-specific families are the homologues present in both yeasts. Hence, the taxon distribution patterns observed for the transcriptional regulator families are similar, but slightly more pronounced, than those observed for the entire data set.

The TA protein families present solely in animals display 3.6 times (6.7 times for the transcriptional regulator families) the level of homologue sharing between species than do fungi, despite their considerable differences in cellular complexity. Additionally, the proportion of each metazoan genome encoding TA proteins increases with complexity and 70% of this increase from *C.elegans* to *H.sapiens* can be accounted for by sequences encoding regulators. However, there is strong evidence that evolutionary changes in developmental gene regulation have shaped large-scale differences in animal body plans and parts, and regulatory DNA is the predominant source of genetic diversity underlying this variation (26). The low inflation clustering operation identified 56, 77 and 170 *Hox* genes in *C.elegans*, *D.melanogaster* and *H.sapiens*, respectively. These homeobox genes are involved in the determination of the animal body plan. After normalising for genome size, it would appear that humans have 2.0 times more *Hox* genes than worms (58.0 versus 29.3) and flies 1.9 times the number (56.2 versus 29.3), whereas the vertebrate genome encodes only 3% more than the insect genome (Fig. 3, Family 3.1.1).

The normalised number of *MYB* genes in *C.elegans*, *D.melanogaster* and *H.sapiens* is, respectively, 1.1, 5.1 and 2.4 and in *A.thaliana* this is substantially elevated to 55.5 (Fig. 3, Family 3.5.1). Hence, TRIBE-MCL identifies 152 MYB-containing sequences from the predicted plant proteome, compared with 190 members of the MYB superfamily using the entire set of genomic sequences (8). However, both data sets show that the *Arabidopsis* MYB family is highly amplified in comparison with the animal families. MYB proteins are widely spread throughout the eukaryotic domain and control cell proliferation and differentiation, with the additional *Arabidopsis* genes controlling many aspects of plant secondary metabolism as well (27). The small variation in normalised *MYB* gene number across the three metazoan species is similar to that observed with the *Hox* genes and supports the idea that it is primarily the evolution of gene expression that underlies morphological diversity. An important caveat however is the possibility of a higher degree of differential splicing and post-translational modification, as well as the presence of unidentified TA-protein-coding genes with complex gene structure, in higher animals concealing TA protein diversity.

An explanation for the very low amount of homologue sharing between budding and fission yeast in the fungal-specific TA protein families could lie in their evolutionary distances: *C.elegans* is closer to *H.sapiens* than *S.pombe* is to *S.cerevisiae* (28). This analysis suggests that the eukaryotic transcriptional process, despite (or because of) it having a fundamental role is highly diversified in the six species of the crown eukaryote group examined, as exemplified by the presence or absence of TA protein homologues in their entire genomes. Possible reasons for this high degree of specificity may include the different needs to respond to environmental and genetic stimuli as well as some form of 'genetic immunity': it may be beneficial for a species to retain a specific repertoire of TA proteins, possibly avoiding catastrophic changes in genetic control by laterally transferred genes from other species. Speciation itself may be largely due to subtle changes in gene regulation via the selective acquisition or loss of TA proteins, as well as changes in their specificity.

Metabolic enzymes are in general highly conserved in sequence and phylogenetic distribution (29), yet a direct comparison of the phylogenetic distribution of TA proteins in eukaryotes with other biological processes, including metabolism, is not currently feasible using the same approach. It would be interesting to examine the extent to which other fundamental biological processes, such as translation or DNA repair, exhibit similar patterns of taxon specificity. To achieve this, a more elaborate classification of the roles of gene products for eukaryotes is required, such as the one provided by the GO classification scheme (21). Once sufficient coverage of eukaryotic genomes has been achieved, it would be interesting to perform comparative studies across species and other biological processes.

ACKNOWLEDGEMENTS

R.M.R.C. and C.A.O. would like to thank the Medical Research Council for supporting this work through a Special Training Fellowship in Bioinformatics to R.M.R.C.

REFERENCES

1. Struhl,K. (1999) Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, **98**, 1–4.
2. Roeder,R.G. (1996) The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.*, **21**, 327–335.
3. Woychik,N.A. and Hampsey,M. (2002) The RNA polymerase II machinery: structure illuminates function. *Cell*, **108**, 453–463.
4. Emerson,B.M. (2002) Specificity of gene regulation. *Cell*, **109**, 267–270.
5. Ptashne,M. and Gann,A. (2002) *Genes and Signals*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
6. Kyrpides,N.C. and Ouzounis,C.A. (1999) Transcription in archaea. *Proc. Natl Acad. Sci. USA*, **96**, 8545–8550.
7. Coulson,R.M., Enright,A.J. and Ouzounis,C.A. (2001) Transcription-associated protein families are primarily taxon-specific. *Bioinformatics*, **17**, 95–97.
8. Riechmann,J.L., Heard,J., Martin,G., Reuber,L., Jiang,C., Keddie,J., Adam,L., Pineda,O., Ratcliffe,O.J., Samaha,R.R. *et al.* (2000) *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.
9. Wood,V., Gwilliam,R., Rajandream,M.A., Lyne,M., Lyne,R., Stewart,A., Sgouros,J., Peat,N., Hayles,J., Baker,S. *et al.* (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature*, **415**, 871–880.
10. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 546, 563–547.
11. The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
12. The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.
13. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
14. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
15. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

16. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
17. Etzold,T. and Argos,P. (1993) SRS—an indexing and retrieval tool for flat file data libraries. *Comput. Appl. Biosci.*, **9**, 49–57.
18. Promponas,V.J., Enright,A.J., Tsoka,S., Kreil,D.P., Leroy,C., Hamodrakas,S., Sander,C. and Ouzounis,C.A. (2000) CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. *Bioinformatics*, **16**, 915–922.
19. Altschul,S.F., Madden,T.L., Schaffer,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.
20. Dwight,S.S., Harris,M.A., Dolinski,K., Ball,C.A., Binkley,G., Christie,K.R., Fisk,D.G., Issel-Tarver,L., Schroeder,M., Sherlock,G. *et al.* (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.*, **30**, 69–72.
21. 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. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
22. Wingender,E., Dietze,P., Karas,H. and Knuppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.
23. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
24. Enright,A.J. and Ouzounis,C.A. (2000) GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics*, **16**, 451–457.
25. Sipiczki,M. (2000) Where does fission yeast sit on the tree of life? *Genome Biol.*, **1**, r1011.1011–r1011.1014.
26. Carroll,S.B. (2000) Endless forms: the evolution of gene regulation and morphological diversity. *Cell*, **101**, 577–580.
27. Stracke,R., Werber,M. and Weisshaar,B. (2001) The R2R3-MYB gene family in *Arabidopsis thaliana*. *Curr. Opin. Plant Biol.*, **4**, 447–456.
28. Feng,D.F., Cho,G. and Doolittle,R.F. (1997) Determining divergence times with a protein clock: update and reevaluation. *Proc. Natl Acad. Sci. USA*, **94**, 13028–13033.
29. Doolittle,R.F., Feng,D.F., Tsang,S., Cho,G. and Little,E. (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science*, **271**, 470–477.