Mechanisms of transcription factor evolution in Metazoa

Jonathan F. Schmitz[1], **Fabian Zimmer**[1,2] **and Erich Bornberg-Bauer**[1,*]

[1]Evolutionary Bioinformatics Group, Institute for Evolution and Biodiversity, Hüfferstrasse 1, D-48149 Münster, Germany and [2]Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK

Received January 28, 2016; Revised May 18, 2016; Accepted May 22, 2016

ABSTRACT

Transcriptions factors (TFs) are pivotal for the regulation of virtually all cellular processes, including growth and development. Expansions of TF families are causally linked to increases in organismal complexity. Here we study the evolutionary dynamics, genetic causes and functional implications of the five largest metazoan TF families. We find that family expansions dominate across the whole metazoan tree; however, some branches experience exceptional family-specific accelerated expansions. Additionally, we find that such expansions are often predated by modular domain rearrangements, which spur the expansion of a new sub-family by separating it from the rest of the TF family in terms of protein–protein interactions. This separation allows for radical shifts in the functional spectrum of a duplicated TF. We also find functional differentiation inside TF sub-families as changes in expression specificity. Furthermore, accelerated family expansions are facilitated by repeats of sequence motifs such as C2H2 zinc fingers. We quantify whole genome duplications and single gene duplications as sources of TF family expansions, implying that some, but not all, TF duplicates are preferentially retained. We conclude that trans-regulatory changes (domain rearrangements) are instrumental for fundamental functional innovations, that cis-regulatory changes (affecting expression) accomplish wide-spread fine tuning and both jointly contribute to the functional diversification of TFs.

INTRODUCTION

Transcriptional regulation is crucial for all known processes in life, in particular for growth and development. Consequently, the evolution of gene expression regulation is tightly linked to the apparent evolution of biological complexity, for example as measured in the number of cell types (1–7). The underlying genomic changes are, as yet, only poorly understood but, among others, changes in cis-regulatory elements (8,9), transcription associated proteins (10) and small regulatory RNAs (11,12) have been identified as major contributors to genomic adaptation. Transcription factors (TFs) are proteins which regulate the transcription of DNA to mRNA in all known organisms by binding to specific DNA target sequences. In eukaryotes, TFs play an important role in development, cellular organization and signal response (13) and dis- or non-functional TF genes have been linked to a number of diseases such as cancer (14,15). TFs have also been implicated in evolutionary innovation of novel phenotypes and developmental frameworks (16).

Generally, expansions of gene families involved in signaling and regulation can be observed at a much higher frequency than the expansions of, e.g. metabolic pathways (17). Also, the number of TFs per genome was found to correlate over-proportionally with the number of genes in genomes, resulting in a higher proportion of TF genes in larger genomes (18). This high proportion of TFs in large genomes suggests that higher complexity requires an over-proportional increase in regulatory elements.

The increases in the number of regulatory proteins in general (4,19) and of TFs in particular (5) have repeatedly been connected to phenotypic innovations and the evolution of more complex organisms. For example, TF family expansions (and size reductions) have been implicated in emergence (and loss) of complex features in Stramenopiles (20) and Viridiplantae (21). Another recent example is the expansion of the C2H2 zinc finger (ZF) and the protocadherin families, which has been linked to increased morphological and developmental complexity of the octopus (22). An expansion of the C2H2 ZF TF family has also been linked to the the secondarily evolved multicellularity in the red algae Chondrus (21,23).

Furthermore, the emergence of new TFs has also been shown to play a role in phenotypic changes, especially in animals (24). Taken together, these and other findings suggest that emergence of TFs and growth of TF families are both

[*]To whom correspondence should be addressed. Tel: +49 251 83 21630; Fax: +49 251 83 24668; Email: ebb@uni-muenster.de

related to increases in morphological complexity and the number of cell types (2,3,20,21,25,26). Therefore, a detailed cross-species comparison of TF repertoires is important to delineate which genetic events underlie the expansion of TF families and which ones were instrumental in creating fundamentally new phenotypes which have led to new and possibly more complex body plans. Indeed, with the availability of many genomes such large scale comparisons allow a detailed analysis of origin and nature of important molecular changes in TFs.

On larger evolutionary time scales, the emergence of new TFs or TF sub-families has been linked to many major transitions in morphology and development, e.g. to the emergence of multicellularity (2,25) or the emergence of flowering plants (27). Indeed, most of the largest metazoan TF families originated already before the emergence of Metazoa and thus multicellularity (25). The further expansion of these families then allowed for the evolution of increasingly complex organisms in Metazoa (2).

In some TF families the expansion results clearly from a number of single gene duplications (SGDs) (28). On the other hand, some expansions were suspected to have been triggered mainly by whole genome-duplications (WGDs), coupled with a high retention rate of TFs (3,5,25,29,30). However, it is still unclear if and how WGDs are instrumental in supporting higher regulatory and organismal complexity. First, no consensus has been reached regarding the causes of the high retention rate of TF genes after SGD as well as WGD events (31). Second, WGDs could not be linked to increased complexity as it is documented in the metazoan fossil record (32). Nevertheless, both processes (SGD and WGD) seem to play a role in the expansion of gene families, specifically TF families (31).

It has been proposed that the number of TF family members is limited by the number of possible target sequences (33). These findings imply that dimerization of TFs would allow for TF family expansion by doubling the DNA target sequence length (one target sequence for both proteins in the dimer). Indeed, many TF families form protein dimers or larger protein complexes in order to bind DNA. Within these TF families, some members are only able to form complexes with themselves (homodimerize), while others can dimerize with other members of the family (heterodimerize) (28,34,35). The interactions between TF members form large interaction networks and the structure of these networks depends on the TF family (35). However, most of the interactions in complex formation are context dependent, i.e. preference may change depending on e.g. pH, localization, concentration or salt strength (36,37). This volatility induces a highly entangled combinatorial interaction pattern which helps to increase the capacity for regulatory fine-tuning, way beyond the associated increase in the number of TFs.

Because several hundred millions of years have elapsed since the emergence of most TF families, it is only rarely possible (see e.g. (38)) to track down the precise molecular and genetic origin of new TF families. Nonetheless, in many cases comparative genomics can reveal major rearrangements which shifted functions of TFs and triggered the emergence of new sub-families. For example, the loss and gain of additional domains has been reported in several families of TFs (35,39). Such changes often entail a strongly altered functional spectrum by changing binding specificites to DNA and upstream regulatory proteins, e.g. signaling proteins or other transcriptional regulators (21,28,40). Domain rearrangements (DRs) may thus explain 'functional shifts', i.e. sudden, radical changes in the regulatory potential of TFs.

In this study we ask how strong the effects of WGDs, SGDs and DRs are on the growth of TF families. Additionally, we analyze if any of these genomic events, or a combination of them, have led to functional shifts which my have spurred fundamental developmental innovations. Accordingly, we study the evolution of the five largest TF families (26) and the p53 family in 36 metazoan species to elucidate the evolutionary history of these families during the evolution of more complex, multicellular organisms. The selected genomes and the size of the chosen families provide a relatively dense and even distribution across the metazoan tree along which many complex phenotypes evolved. We determine extant and ancestral TF family sizes to identify branches with accelerated expansions and relate expansions to underlying molecular changes and genomic rearrangements. Finally, we relate these changes to functional properties which can be inferred from annotations and expression profiles of TFs.

MATERIALS AND METHODS

Taxon sampling and sequence data

The 36 species analyzed here were selected to represent a large sample of sequenced Metazoa with a high quality genome available. *Saccharomyces cerevisiae* was chosen as a non-metazoan outgroup with a high genome quality. To enable phylogenetic analyses, a dated tree was reconstructed based on the study by Erwin *et al.* (41). Dating for species not included in the Erwin *et al.* study were added manually according to various sources (see Supplementary Data). The sequence data for most species were obtained from Ensembl release 74 (42) or from Ensembl Genomes release 21 (43). Species not available on Ensembl were downloaded from various sources, see Supplementary Table S3. Only the longest splicing variant of each gene was considered in our analyses.

Domain annotation

Domains were annotated using the hidden Markov models (HMMs) of Pfam-A version 27.0 (44). The PfamScan script provided by Pfam was used to perform the annotation. A list of HMMs representing the TF families' DNA-binding domains (DBDs) was used to identify TF proteins. For the list defining the relationship DBD–HMMs see Supplementary Table S2. A protein's domain arrangement was defined as the sequence of domains, domain repetitions were not collapsed. All proteins sharing a domain arrangement were grouped into a domain arrangement cluster (DAC).

Ancestral family size reconstruction

Ancestral TF family sizes for all nodes in the species tree were reconstructed using Count (45) in symmetrical Wagner parsimony mode by setting the ratio of gain- to loss-penalties to 1. In Count, the DACs were used as subfamilies. The number of annotated proteins per DAC was used as input for Count.

Comparison of gene/DAC gain/loss rates. For each of the branches of the species tree, the gene/DAC gain and loss rates were calculated by dividing the number of events per category by the branch length in million years. This analysis was performed using a custom R script (46). Figures comparing the distribution of rates were produced using the ggplot2 R library (47). To test the rate distribution of the four categories for differences, the Wilcoxon signed rank test was used. The wilcox.test function of the R base package was used for this purpose (46).

Plotting of TF family evolution per lineage. The TF family evolution of a lineage was represented by plotting the TF family size and DAC composition for each ancestral node. The plotting was performed using a custom Python script utilizing the matplotlib plotting library (48).

Gene Ontology enrichment testing

Gene Ontology (GO) annotation data were downloaded from Ensembl for the model organisms *Homo sapiens*, *Danio rerio*, *Drosophila melanogaster* and *Caenorhabditis elegans* (42). Using the topGO R library (49), the proteins of each DAC were tested for GO enrichment using all proteins of the respective TF family as background. topGO's weighted Fisher test method was used. The minimum number of annotations per GO term was set to 3 (Node Size = 3) to ensure a certain stability of the GO annotations. Consequently, only DACs with at least three protein members were taken into account for this analysis. A *P*-value cutoff of 0.05 was chosen to select only significant hits. A multiple testing correction was performed by multiplying *P*-values with the number of DACs for which GO enrichment tests were executed. In this analysis, only the biological process class of GO was considered.

Gene expression pattern comparison

Expression data for eight human organs (50) were used to compare the expression of the TFs. Pre-computed FPKM values for this experiment were obtained from the Expression Atlas Website (51). To compare the expression patterns among the TF genes, the genes were clustered according to their expression profile similarity using the cosine function as a similarity measurement. Clusters of genes with similar expression profiles were then manually inspected for the proteins' domain arrangements. The vector of expression strengths per organ, given as the FPKM value, was used as expression profile for each gene. This approach was chosen since FPKM values can not be used to reliably compare expression strength across experiments (52). The analysis was conducted in R (46) utilizing the lsa packages' cosine function (53) and hclust in complete mode from the R base package. Custom python scripts were used to analyze expression breadth using a cutoff of 1 FPKM for presence of expression. The first node with DAC presence generated by Count (see above) was used to determine domain arrangement age. GOATOOLS (https://github.com/tanghaibao/Goatools) in Fisher's exact test mode was used to determine GO enrichment in clusters of genes with similar expression patterns. Clusters of genes with similar expression were extracted using the hierarchical clustering function of SciPy (54).

RESULTS AND DISCUSSION

We annotated TFs in 36 metazoan species using HMMs to find the TF family-specific DBDs. We leave aside other transcriptional regulators (see also (10)), because most of these have multiple, more general, roles such that their evolutionary functional impacts are even more difficult to characterize than those of TFs. We do, however, include family members of TFs that have lost their DNA binding abilities in a secondary event. Such a loss of DNA binding affinity can provide valuable information on the molecular triggers of functional shifts and family expansions and can be clearly delineated by comparative genomics.

To analyze TF family sizes, we first determined the TF families in our set of 36 metazoan species and baker's yeast (see Figure 1). The first family we annotate is the bHLH TF family, which is characterized by the basic helix-loop-helix domain, in which the basic region binds DNA and the helix-loop-helix motif facilitates dimerization and DNA binding. Next to the bHLH domain, other protein domains can be found in bHLH proteins (55), such as the Orange, PAS or Leucine zipper (LZ) domains (28). These domains can have various functions, such as environmental sensing, signal transduction and dimerization facilitation (56,57).

The second family is the bZIP TF family, whose proteins contain a basic region that binds DNA, just as bHLH proteins do. However, the bZIP basic region does not show any detectable homology to the bHLH basic region and is likely an example of convergent evolution. In bZIP proteins the basic region directly extends into an LZ which, convergently to bHLH proteins again, facilitates dimerization (58). The bZIP family comprises many well known TFs such as JUN and FOS, which are involved in cell proliferation, apoptosis, and survival, as well as cancer development in case of loss-of-function mutations (59,60).

The third family is the Homeobox TF family, which is defined by the Homeobox domain that consists of 60 amino acids forming three α-helices (61). Proteins carrying Homeobox domains can be found in all eukaryotes (25) and play an important role in regulating development, especially in Metazoa (61). The Hox genes are the best-known metazoan homeobox genes and are crucial in Bilateria for determining the body axis during development among other functions (62).

Fourth, the Nuclear Receptor (NR) family was analyzed. NR proteins contain a DBD and a ligand-binding domain (LBD). The LBD binds a number of cofactors such as steroid hormones or lipids (63,64) and can also facilitate dimerization (65). The NR family is Metazoa-specific (25) and important for the regulating of development, metabolism and reproduction.

Next, the C2H2 ZF family is defined by a sequence motif in which two Cystein (C) and two Histidine (H) amino acid residues coordinate a zinc ion. The C2H2 ZF domain fa-

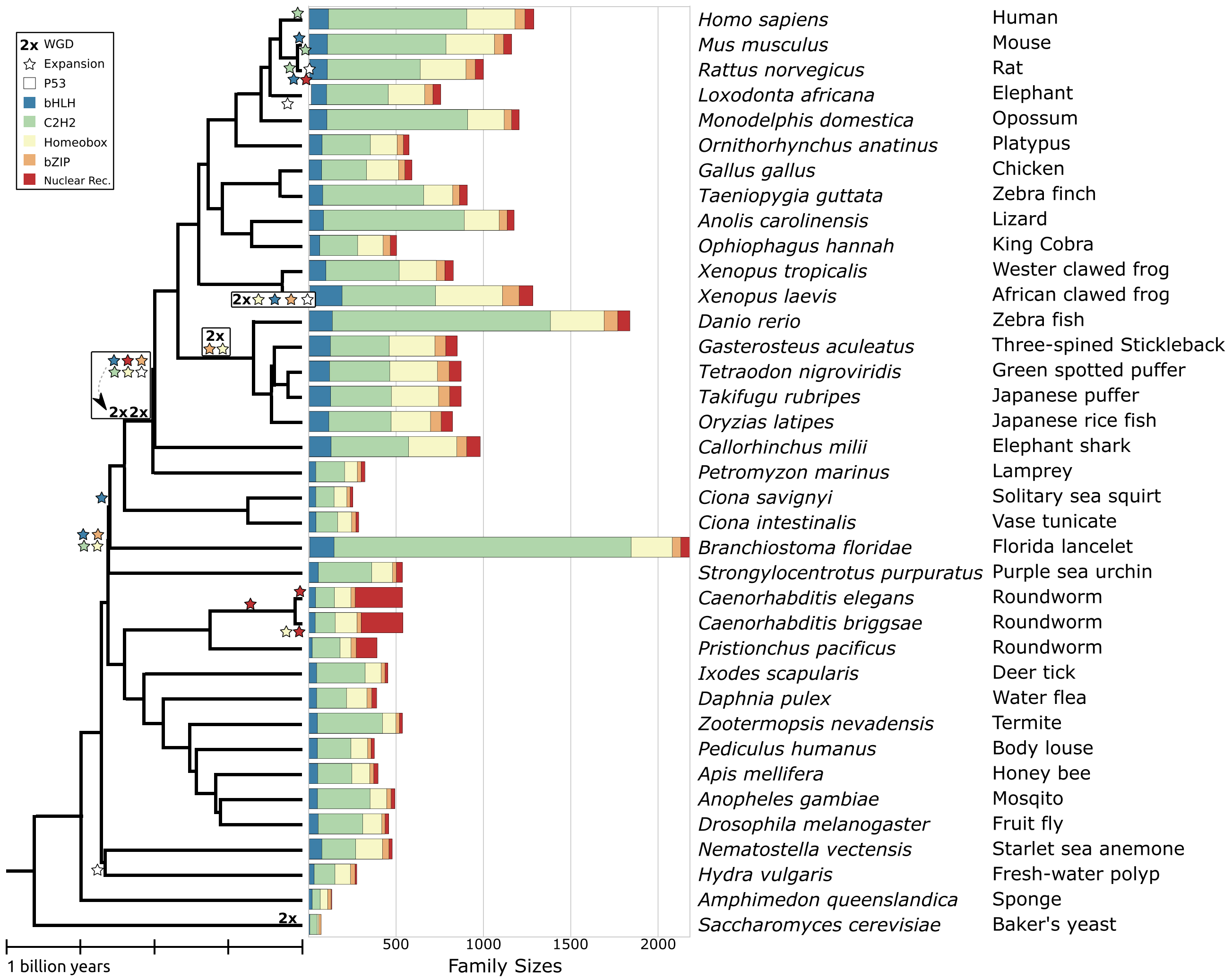

Figure 1. Stacked bar plots depicting TF family sizes in analyzed species. The left-hand side of the graph shows a phylogenetic tree of the analyzed species. WGD events are denoted with a '2x'-symbol. Branches with accelerated gene gain rates are highlighted with stars. The stars are colored according to the TF family with accelerated gain rate. Time scale is approximate and largely based on (41). See 'Materials and Methods' section for more details.

cilitates DNA binding as well as dimerization and is made up of two β-sheets and one α-helix. C2H2 ZF genes can be found in all eukaryotes (25) and have various functions such as regulation of stress response (66). C2H2 ZFs have been proposed to play a role in a number of important evolutionary processes such as speciation in the primate lineage (67,68).

Finally, p53 proteins consist of the p53 DBD, a 200 amino acids long domain consisting mainly of β-sheets, the p53 tetramerization domain that facilitates oligomerization of p53 proteins and in some cases additional domains (69). p53 genes can be found in Holozoa and are not Metazoa-specific (25). In Metazoa, p53 proteins are important, mainly in controlling the cell cycle. Loss of function of a p53 gene can entail a cancer risk (70). The p53 family was included because of this high relevance for medical issues. The other families were analyzed because they are the largest TF families in human and as such represent the bulk of the TFs.

TF family sizes in Metazoa show a pattern of repeated expansions

We analyzed the evolution of TF families in Metazoa using the TF family sizes determined in the previous step. TF family sizes vary drastically in Metazoa for different species and TF families. More specifically, some lineages, such as the ray-finned fishes, have experienced expansion of all TF families. Also, all TF families are expanded compared to most non-vertebrate species. Interestingly, the lancelet lineage has much larger TF families than any of the closely related lineages. On the other hand, in some lineages only one specific TF family is expanded, like the nematode clade in which the largest NR families of all Metazoa can be found (already noted in (71)). Many species' genomes contain a markedly larger C2H2 ZF family compared to closely related species. Examples for species with expanded C2H2 ZF family are *Anolis carolinensis*, *D. rerio*, *Anopheles gambiae*, *Zootermopsis nevadensis* and *Ixodes scapularis*. Additionally, most mammals except for the elephant possess larger

C2H2 ZF families than most other vertebrates. Generally, repeating patterns of clade or lineage-specific expansions of one or multiple TF families can be observed.

The differences in TF family sizes between different animals raises the question of which proportion of species' proteomes the TF families take up. A comparison between the TF family sizes and proteome sizes shows that differences in the proportion of TFs between clades and lineages can be observed (Supplementary Figure S2). In many clades with TF family expansions, TF families make up a larger portion of the proteome. One example are vertebrates in which the bHLH and bZIP families make up a larger portion of the proteome than in non-vertebrate species. The C2H2 ZF family forms a different pattern characterized by lineage-specific expansions. Consequently, the C2H2 TF family makes up a noticeably larger portion of some species' proteome compared to closely related species. C2H2 ZFs are noticeably expanded in the proteome of *D. rerio*, *Branchiostoma floridanus* and *A. carolinensis* for example. The C2H2 family is exceptional in showing such taxonomically restricted bursts, resulting in a larger fraction of the species' proteome being made up of the C2H2 family. The p53 family is expanded in the elephant (*Loxodonta africana*) without an expansion of the elephant proteome. This finding confirms previous ones about a p53 family expansion in elephants (72). However, in general, lineage-specific TF expansions should be interpreted cautiously as they can be an artifact of incorrect genome annotations. In general, TF family expansions often lead to a higher proportion of TFs in the proteome. These expansions can be stable in clades, like for the bHLH and bZIP families in vertebrates.

Given the high variability in TF family sizes it can be concluded that TF family expansion/reduction has occurred along many branches of the metazoan tree. However, findings of burst-like TF family expansions the evolution of Metazoa have only been reported for the proto-metazoan stem (2). In cases where a large clade has significantly larger TF families for all TF families (Vertebrata, ray-finned fishes), a clear connection between WGD events on the branches leading to these clades and the TF family expansions can be made. Additionally, the WGD event on the branch leading to *Xenopus laevis* seems to have doubled the size of most TF families except for the C2H2 ZFs. However, in other cases larger TF families can not be linked to WGD events. The lancet *B. floridae*, for example, has a high number of genes for all TF families, but no WGD event has been proposed to have occurred in that lineage. Also, the many cases of significantly larger C2H2 families do not seem to be connected to WGD events, just as the large NR TF family in nematodes. The hypothesis that C2H2 family expansion is more often connected to SGD than to WGD is further supported by smaller median pairwise gene distances in human (Supplementary Figure S4) and the small amount of C2H2 expansion after the *X. laevis* WGD (see Figure 1). To clarify the relationship between TF family expansions and WGD events, we analyzed ancestral TF family sizes in a next step.

Reconstructed ancestral TF family sizes reveal branches with accelerated gene gain

To locate points in the evolution of Metazoa with accelerated TF family expansion, we reconstructed the TF family sizes of the ancestral nodes of our phylogenetic tree. Using the ancestral TF family sizes we compared the gain/loss rates of genes as well as DACs (genes sharing a domain arrangement) along the branches of the phylogenetic tree. The gain or loss of a DAC describes the gain of at least one gene with a certain domain arrangement or respectively the loss of all genes with a certain domain arrangement in a tree node compared to the parental node. Box plots of gain and loss rates for the six TF families (Supplementary Figure S1) show that the analyzed TF families mainly evolve via gene gain. For all families the gene gain rate distribution has a higher median than the other event types (Wilcoxon signed rank test; $P < 0.01$ for all families). The DAC gain rates are also relatively high compared to the loss rates, which complies with DAC gain being linked to gene gain. The loss rates, for DACs as well as genes, are lower than either of the gain rates, showing that gain of genes seems to be the more important process in TF family evolution. This finding indicates a largely constant growth of the TF families. The magnitude of gene gain rates differs between the six TF families. In p53, for example, the maximum observed gene gain rate is below 0.2 genes per million years, while for C2H2 ZF more than 25 gene gains per million years can be observed on the branch leading to *Mus musculus* since the split from the *Rattus norvegicus* branch.

The gene gain rate distributions (Supplementary Figure S1) feature a number of prominent outliers. These outliers indicate branches with strongly accelerated TF family evolution, indicative of events that we call 'bursts'. For outlier branches with such bursts see Table 1 and Figure 1. Many branches show up for more than one TF family burst, for example the branch leading to *X. laevis* or the Gnathostomata branch. In some cases the bursts in gene gain rate can be linked to WGD events. For the branches leading to *X. laevis* and Percomorphia (ray-finned fishes), WGD events have been proposed (73,74). These two branches show accelerated gene gain rates for four and two TF families, respectively. For the Gnathostomata branch no WGD has been proposed directly, but for its parent branch, the branch leading to Vertebrata, the 2R WGD events have been proposed (75,76). The only non-gnathostome vertebrate in our species set is the lamprey. The *Petromyzon marinus* genome likely caused an artifact in the ancestral reconstruction of TF family sizes because of its vertebrate-atypical small proteome size, 30% smaller than the next smallest analyzed vertebrate (*P. marinus*: 10 415 proteins, *Gallus gallus*: 15 508 proteins, no splice variants counted, from ensembl annotation). Consequently the accelerated gene gain rate on the Gnathostomata branch is likely connected to the 2R WGD events.

However, in other cases accelerated gene gain rates can not be linked to WGD events. The branches leading to *R. norvegicus* and Deuterostomia, for example, show accelerated gene gain rates for four TF families while no WGD has occurred on these branches. Other branches without WGD event show accelerated gene gain rates only for one

Table 1. Tree branches with an exceptionally high gene gain rate for one or more of the TF families and the evolutionary events that can be linked with the accelerated gene gain rate

Branch	Event	TF families
Caenorhabditis	SGD	Nuclear Receptor
Caenorhabditis elegans	SGD	Nuclear Receptor
Caenorhabditis briggsae	SGD	Homeobox, Nuclear Receptor
Chordata	SGD	bZIP
Cnidaria	SGD	p53
Deuterostomia	SGD	bHLH, bZIP, C2H2, Homeobox
Gnathostomata	WGD	bHLH, bZIP, C2H2, Homeobox, Nuclear Receptor, p53
Homo sapiens	SGD	C2H2
Loxodonta africana	SGD	p53
Mus musculus	SGD	bHLH, C2H2
Percomorpharia	WGD	Homeobox, bZIP
Rattus norvegicus	SGD	bHLH, C2H2, Nuclear Receptor, p53
Xenopus laevis	WGD	bHLH, bZIP, Homeobox, p53

For each branch the name of the node at the younger end of the branch was used as name.

or two TF families. *A priori*, WGDs would be expected to be linked to an accelerated gene gain rate in most TF families since all genes get duplicated and only families where many genes are lost afterward would show no acceleration in gene gain rates. It has previously been suggested that TF families show high retention rates after WGD events (31,77,78). Family expansion largely caused by SGD events, however, could be a sign of evolutionary pressure for innovation on the affected TF family. In such a case, not all TF families would be expected to be under this evolutionary pressure. Consequently, only few TF families would be expected to show accelerated gene gain rates on branches without WGD event. Many, but not all, branches seem to follow these patterns in our case. A low retention rate of some TF families after a WGD event can be explained by less evolutionary pressure for innovation on this TF family. For example, gene losses in some parts of the teleost fish lineage could explain the small number of TF families with accelerated gene gain rate on the Percomorphia branch in our reconstruction. On the other hand, evolutionary pressure for regulatory innovation could explain the accumulation of TF families with accelerated gene gain rate in the branches leading to, e.g. Deuterostomia, where no WGD event occurred. Nevertheless, we find that WGD events lead to accelerated TF family expansion rates for all analyzed branches with WGD in Metazoa, at least in some TF families. Additionally, we find a number of branches with increased TF family expansion rates caused by SGDs. These findings show that WGD as well as SGD both contribute to TF family expansions. To further understand the mechanism of TF family expansion, we analyzed the domain arrangements found in the TF families.

TF family size is correlated with number of DACs and unique domains

We analyzed the relationship between DRs and TF family expansion to elucidate the role of DRs for the expansion of TF families. All TF families show a positive correlation between TF family size and the number of DACs (Figure 2A). However, the strength of the correlation varies between the TF families (Table 2). The strongest correlation (0.93) can be found for the Homeobox and C2H2 ZF TF families, which are also the two largest TF families in most of the analyzed species. The increase in the number of DACs per TF family with TF family expansion could either be a by-product of the TF family evolution or a required step during TF family expansion. Given that protein domains are seen as the functional subunits of proteins it seems logical that DRs strongly influence TF function in various ways. Additional domains can also restrict the dimerization partners of dimerizing proteins and thereby modify the TF family's dimerization network (35). Creating dimerization sub-networks could facilitate functional diversification of TFs by minimizing cross-talk between different functions. An additional domain could also facilitate interaction with other molecules in the cell, i.e. signaling. The PAS domain is an example for a protein domain that can facilitate signaling in a protein (57) and can be found in the bHLH TF family (35).

Apart from additional domains, rearrangement of existing domains can also influence TF function (79). Such changes have been reported for many families, e.g. a number of plant gene families (80), many genes involved in signal transduction (81) and globins (82). In the C2H2 ZF TF family the C2H2 domain can be repeated as often as 30 times. The repetition of the DBD could in this case augment the number of possible target sequences in the DNA and thereby facilitate functional diversification. Additionally, this higher number of target sequences could allow family expansion, since previously the number of target sequences was suggested to be limiting to family size (33).

There is also a correlation between the number of unique domains and the number of genes per TF family (Figure 2 and Table 2). The implications of this correlation are quite similar to the implications of the correlation between number of DACs and number of genes. The main difference between the two analyses is that, when counting DACs, all possible arrangements of domains, i.e. repititions or changed order, are counted separately. When counting unique domains, each domain is only counted once, regardless of the number of separate arrangements it occurs in. Counting all DRs has the advantage of also considering events such as domain duplications that are common, especially in C2H2 ZFs (83,84). In practice, both measures

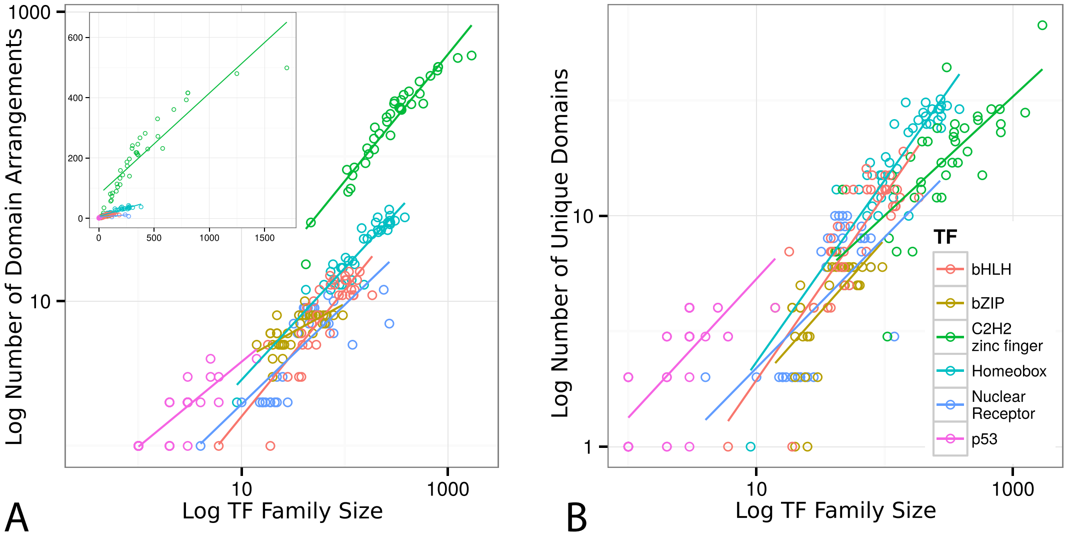

Figure 2. Relationship between domains and number of genes per TF family for all analyzed species. Linear regression lines are shown for each TF family. (**A**) Number of DACs (different domain arrangements) per TF family plotted against number of genes in the respective TF family. Full scale graph shows a log–log plot, inset shows linear axis. Each of the points represents one species. (**B**) Number of unique domains per TF family plotted against number of genes in the respective TF family. Each point represents one species.

Table 2. Correlation between TF family gene number and number of DACs in the respective TF family

TF family	Correlation to DAC number	Correlation to number of unique domains
bHLH	0.76	0.72
bZIP	0.66	0.67
C2H2 zinc finger	0.92	0.78
Homeobox	0.91	0.84
Nuclear Receptor	0.52	0.35
p53	0.79	0.56

The correlation of gene number and number of unique domains per TF family is also shown. The values given are product-moment correlation coefficients.

are meaningful, as the number of unique domains can show gain of novel functions and the number of domain arrangements can show events of major restructuring of TF proteins.

DACs are functional subunits of TF families

To determine the influence of DRs on TF function we tested the DACs of each TF family for GO term enrichment. In human, most DACs showed significant enrichment for certain GO terms, except in the C2H2 ZF family where only less than half of the DACs showed functional enrichment (Supplementary Table S1). For other species fewer DACs showed enrichment of GO terms. This result is likely caused by an incomplete annotation of TFs in species other than human. The enrichment of GO terms in the DACs shows that functions differ between the DACs of a TF family and at least some genes in each DAC share a function. The enriched GO terms of a DAC can cover a range of completely different functions (Figure 3). For example, DACs can show enrichment for GO terms as different as muscle cell differentiation and nephron tubule development. The enrichment for different GO terms shows that the genes belonging to each DAC can facilitate a wide range of functions.

The enrichment of certain GO terms in the DACs' genes could be caused by an influence of the domain arrangement on the function of proteins. An influence of domain arrangement on function would explain differences in function between the DACs. As mentioned previously, there are various ways in which changes in domain arrangements can influence TF function, e.g. by adding signaling or dimeriza-

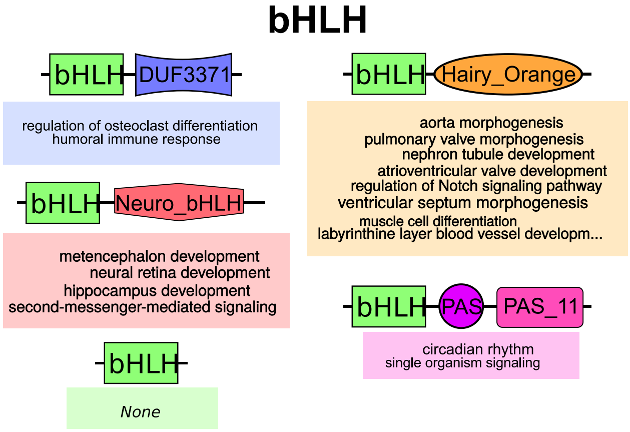
bHLH
bHLH
DUF3371
regulation of osteoclast differentiation
humoral immune response
bHLH
Hairy_Orange
aorta morphogenesis
pulmonary valve morphogenesis
nephron tubule development
atrioventricular valve development
regulation of Notch signaling pathway
ventricular septum morphogenesis
muscle cell differentiation
labyrinthine layer blood vessel developm...
bHLH
Neuro_bHLH
metencephalon development
neural retina development
hippocampus development
second-messenger-mediated signaling
bHLH
PAS
PAS_11
circadian rhythm
single organism signaling
bHLH
None

Figure 3. Wordclouds of the GO terms found to be enriched in the DACs of the bHLH TF family in human. For each DAC a pictogram of the domain arrangement is shown. Each GO term is scaled according to the *P*-value found in the enrichment test (smaller *P*-values mean bigger font size).

tion functionality to certain genes through the gain of certain domains.

Expression patterns differ between DACs

Since genes with similar expression patterns are expected to have similar functions (85–87), we analyzed the expression patterns of the TF family members. We determined if DAC members share the same expression pattern as an alternative explanation for the enrichment of GO terms in DACs found in the previous section. However, we find that TFs of a DAC rarely share the same expression pattern, i.e. many genes that have the same domain arrangement do not share the same expression pattern (see Supplementary Figure S3). Expression clusters consist of genes that all show high expression in some tissues, but low expression in the rest of the tissues. The domain arrangements of the genes found in the expression clusters differ, with several different arrangements present among them. Also, TFs with the same domain arrangement can be found in various clusters of TFs with similar expression patterns. Still, enrichment of GO terms could be found in clusters of TFs with similar expression patterns. But the GO terms enriched in clusters of TFs with similar expression patterns are different from the terms found enriched in DACs (compare Supplementary Table 4 and Figure S3). This finding suggests that DACs and expression clusters both represent functional subunits of TF families. However, these subunits are not congruent, meaning that genes with the same domain arrangement show different expression patterns that are necessary to carry out the specific regulation in multiple tissues. Additionally, members of different DACs are present in the same expression cluster. Joint expression could lead to interference between TF family members. Likely, DRs represent a mechanism via which interference can be inhibited due to changed dimerization preferences. In this way, DRs could also facilitate TF family growth.

In an additional step, we analyzed the breadth of TF expression, i.e. the number of organs a TF was found to be expressed in human (FPKM >= 1; Supplementary Figure S5). Across all TFs, most TFs were found to be expressed either in most organs or few/none of the analyzed organs. Only few of the TFs being expressed in an intermediate number of organs. Globally, this pattern has already been found in previous studies which did not differentiate TF families and DACs (26,88). However, when analyzing expression breadth of the TF families separately, our results reveal a different pattern. For the Homeobox family, for example, most genes are expressed in few tissues and only few are expressed in more than four organs. For the bZIP family, on the other hand, most genes are expressed in more than four organs. These differences in expression breadth most likely stand in relation to the TF function. Homeobox genes are often associated with developmental functions and would as such not be expected to be expressed in many adult organs. When analyzing the expression breadth of the genes in the various DACs according to the DAC's evolutionary age, the pattern visible for the whole TF family is also represented in most of the DACs (Supplementary Figure S6). There does not seem to be a relationship between DAC age and expression breadth, all patterns of expression breadth appear in all age groups. In this, our results are somewhat in contrast to previous results that proposed a more specialized expression of recently duplicated genes (89). According to our results, the C2H2 family with many recent duplications is broadly expressed. However, this might also be related to specific functions of the C2H2 family in silencing mobile elements in the genome (90,91).

CONCLUSIONS

Our results show that the expansion of TF families is often accompanied by a functional diversification that follows modular DRs. According to our findings, gene duplications offer the potential for sequence changes in one of the copies, in agreement with the established theories about gene duplications (75,92). Among the possible mutations, DRs offer the largest shift in function. By gaining a dimerization and sensing domain such as the PAS domain in bHLH, a gene copy can establish new functions such as binding signaling molecules in the cell and also act independently from the rest of the family through a new dimerization specificity. Through further gene duplications (especially WGD events), a new sub-family can be established. According to our model, WGD events per se do not add much complexity; however, functional diversification of expanded gene families after a certain time can do so.

Our study offers a solution to the riddle of how WGDs and seemingly gradual molecular changes can both help increase the complexity although WGDs alone seem to have little effect (see above). True innovation in function often requires a predating molecular change as trigger. Such a trigger can be a rearrangement of domains or the exaptation of a duplicate for a new function and both may lead to a radical shift in function. DRs and emergence as a trigger for functional shifts across a wide range of regulatory proteins have also been reported in recent studies concentrating on genomic comparisons of closely related insect species (81,93). An additional mechanism of functional diversification found in this study is change of expression patterns. These two mechanisms can help explaining the expansion of TF families by laying out how novel functions can be obtained.

Once established, such true novelties are receptive to further expansions and fine tuning which may allow for a rapid expansion of TF families and diversification of functions of family members. A possible WGD leads to a large amount of raw material which is, according to our data, in many cases rapidly utilized. However, these subsequent changes in TF protein sequence are mostly subtle, at least initially, leaving the overall architecture of regulation in order. This relationship is obvious, for example in the maintenance of interaction patterns in bZIP proteins (see above and (34)) and helps to explain why WGDs can not easily be linked to sudden organismic innovations (21,32,94). WGDs may of course still be instrumental, for example for adaptation under rapidly changing environmental conditions (29,74,95), but their adaptive value is likely not primarily related to the innovative potential of novel TFs but rather to the changes in gene expression brought by the WGDs (95). SGDs, on the other hand, can also contribute to network growth, since their duplicates also inherit their interaction preferences.

A remarkable case in point is the MADS TF family which has only five copies in human (26) and no known major expansions in any metazoan linage, but up to a hundred copies in plants (27). The MADS TF family has probably evolved by exaptation from a DNA topoisomerase (38). In plants, an array of several domains, which are mostly involved in the dimerization (or multimerization) of MADS proteins, has been acquired in a group of paralogs which became known as MIKC-type MADS proteins. These, but not any of the MIKC-free MADS proteins, then duplicated to form a dense interaction network (39). This interaction network mainly evolved from a starting point of nine to eleven interacting MIKC proteins via WGDs that left the core-interaction patterns intact (78). MIKC-type MADS proteins are key determinants of plant flower development (ABC model) and are thus instrumental for the intricacies of petal development (96). Therefore, in striking resemblance to the recruitment of domains by metazoan bHLH proteins (28), the acquisition of the IKC domains in the MADS TF family seems to have triggered a functional shift which allowed for subsequent expansion via WGDs, as was also the case in metazoan bZIP proteins (34).

In all scenarios, continuous changes in function, such as gradual shifts of sub-optimal functions as they can be observed in some enzymes (97) have not been reported for TF evolution. A possible reason may be that TF functions are more specific such that minor changes may render them non-functional and prone to rapid loss as has been hypothesized from mutational experiments on bHLH proteins (40). Modular rearrangements of domains offer a solution to this problem because readily approved subunits are recombined.

By delineating the relationships between TF family expansions, TF expression patterns and domain arrangements we make another step toward understanding the evolutionary history of Metazoa. We help explain how the TF families could expand during the evolution of Metazoa, an event that likely facilitated the evolution of more biological complexity (1–7). Our findings further our understanding of how the functional diversification of expanding TF families works in detail, namely by DRs and changes in expression pattern. This functional diversification seems necessary for family growth as it would help explain why only some genes are retained after duplication events. In detail, we find DRs and changes in expression to both contribute to functional diversification independently which we demonstrated by showing distinct GO enrichment in DACs and expression clusters. Overall, these findings shed a new light on how the evolution of more complex organisms with differing body plans and rising numbers of cell types occurred in a number of metazoan lineages.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors would like to thank Amelie S. Veron and Tristan Bitard Feildel for stimulating discussions, Andreas Schüler for pilot studies and Carsten Kemena for computational support.

FUNDING

DFG [BO-2544/2-2 to EBB]; EBB ORCID [0000-0001-2345-6789]; EBB RESEARCHID [A-1563-2013]. Funding for open access charge: DFG [BO-2544/2-2].

Conflict of interest statement. None declared.

REFERENCES

1. Valentine,J.W., Collins,A.G. and Meyer,C.P. (1994) Morphological complexity increase in metazoans. *Paleobiology*, **20**, 131–142.
2. Degnan,B.M., Vervoort,M., Larroux,C. and Richards,G.S. (2009) Early evolution of metazoan transcription factors. *Curr. Opin. Genet. Dev.*, **19**, 591–599.
3. Charoensawan,V., Wilson,D. and Teichmann,S.A. (2010) Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Res.*, **38**, 7364–7377.
4. Miyata,T. and Suga,H. (2001) Divergence pattern of animal gene families and relationship with the Cambrian explosion. *Bioessays*, **23**, 1018–1027.
5. Levine,M. and Tjian,R. (2003) Transcription regulation and animal diversity. *Nature*, **424**, 147–151.
6. McCarthy,M.C. and Enquist,B.J. (2005) Organismal size, metabolism and the evolution of complexity in metazoans. *Evol. Ecol. Res.*, **7**, 681–696.
7. Vogel,C. and Chothia,C. (2006) Protein family expansions and biological complexity. *PLoS Comput. Biol.*, **2**, e48.
8. Wray,G.A. (2007) The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.*, **8**, 206–216.
9. Gaunt,S.J. and Paul,Y.-L. (2012) Changes in cis-regulatory elements during morphological evolution. *Biology*, **1**, 557–574.
10. Lang,D., Weiche,B., Timmerhaus,G., Richardt,S., Riaño-Pachón,D.M., Corrêa,L.G.G., Reski,R., Mueller-Roeber,B. and Rensing,S.A. (2010) Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. *Genome Biol. Evol.*, **2**, 488–503.
11. Grimson,A., Srivastava,M., Fahey,B., Woodcroft,B.J., Chiang,H.R., King,N., Degnan,B.M., Rokhsar,D.S. and Bartel,D.P. (2008) Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*, **455**, 1193–1197.
12. Technau,U. (2008) Evolutionary biology: small regulatory RNAs pitch in. *Nature*, **455**, 1184–1185.
13. Weirauch,M.T. and Hughes,T. (2011) A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. In: Hughes,TR (ed). *A Handbook of Transcription Factors*. Springer, Dordrecht, Vol. **52**, pp. 25–73.
14. Tomlins,S.A., Rhodes,D.R., Perner,S., Dhanasekaran,S.M., Mehra,R., Sun,X.-W., Varambally,S., Cao,X., Tchinda,J., Kuefer,R. *et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–648.
15. Gordon,S., Akopyan,G., Garban,H. and Bonavida,B. (2005) Transcription factor YY1: structure, function, and therapeutic implications in cancer biology. *Oncogene*, **25**, 1125–1142.
16. Lynch,V.J., Leclerc,R.D., May,G. and Wagner,G.P. (2011) Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat. Genet.*, **43**, 1154–1159.
17. Lespinet,O., Wolf,Y.I., Koonin,E.V. and Aravind,L. (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.*, **12**, 1048–1059.
18. Babu,M.M., Luscombe,N.M., Aravind,L., Gerstein,M. and Teichmann,S.A. (2004) Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.*, **14**, 283–291.
19. Kusserow,A., Pang,K., Sturm,C., Hrouda,M., Lentfer,J., Schmidt,H.A., Technau,U., von Haeseler,A., Hobmayer,B., Martindale,M.Q. *et al.* (2005) Unexpected complexity of the Wnt gene family in a sea anemone. *Nature*, **433**, 156–160.
20. Buitrago-Flórez,F.J., Restrepo,S. and Riaño-Pachón,D.M. (2014) Identification of transcription factor genes and their correlation with the high diversity of stramenopiles. *PLoS One*, **9**, e111841.
21. Lang,D. and Rensing,S.A. (2015) The evolution of transcriptional regulation in the viridiplantae and its correlation with morphological

complexity. In: Ruiz-Trillo,I and Nedelcu,AM (ed). *Evolutionary Transitions to Multicellular Life*. Springer, The Netherlands, pp. 301–333.
22. Albertin,C.B., Simakov,O., Mitros,T., Wang,Z.Y., Pungor,J.R., Edsinger-Gonzales,E., Brenner,S., Ragsdale,C.W. and Rokhsar,D.S. (2015) The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature*, **524**, 220–224.
23. Collén,J., Porcel,B., Carré,W., Ball,S.G., Chaparro,C., Tonon,T., Barbeyron,T., Michel,G., Noel,B., Valentin,K. *et al.* (2013) Genome structure and metabolic features in the red seaweed Chondrus crispus shed light on evolution of the Archaeplastida. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 5247–5252.
24. Jovelin,R. (2009) Rapid sequence evolution of transcription factors controlling neuron differentiation in caenorhabditis. *Mol. Biol. Evol.*, **26**, 2373–2386.
25. de Mendoza,A., Sebé-Pedrós,A., Šestak,M.S., Matejčić,M., Torruella,G., Domazet-Lošo,T. and Ruiz-Trillo,I. (2013) Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, E4858–E4866.
26. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
27. Becker,A. and Theißen,G. (2003) The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Mol. Phylogenet. Evol.*, **29**, 464–489.
28. Amoutzias,G.D., Robertson,D.L., Oliver,S.G. and Bornberg-Bauer,E. (2004) Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. *EMBO Rep.*, **5**, 274–279.
29. Van de Peer,Y., Maere,S. and Meyer,A. (2009) The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.*, **10**, 725–732.
30. Smet,R.D., Adams,K.L., Vandepoele,K., Montagu,M.C.E.V., Maere,S. and de Peer,Y.V. (2013) Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 2898–2903.
31. Rensing,S.A. (2014) Gene duplication as a driver of plant morphogenetic evolution. *Curr. Opin. Plant Biol.*, **17**, 43–48.
32. Crow,K.D. and Wagner,G.P. (2006) What is the role of genome duplication in the evolution of complexity and diversity? *Mol. Biol. Evol.*, **23**, 887–892.
33. Itzkovitz,S., Tlusty,T. and Alon,U. (2006) Coding limits on the number of transcription factors. *BMC Genomics*, **7**, 239.
34. Amoutzias,G.D., Veron,A.S., Weiner,J., Robinson-Rechavi,M., Bornberg-Bauer,E., Oliver,S.G. and Robertson,D.L. (2007) One Billion Years of bZIP Transcription Factor Evolution: Conservation and Change in Dimerization and DNA-Binding Site Specificity. *Molecular Biology and Evolution*, **24**, 827–835.
35. Amoutzias,G.D., Robertson,D.L., Van de Peer,Y. and Oliver,S.G. (2008) Choose your partners: dimerization in eukaryotic transcription factors. *Trends Biochem. Sci.*, **33**, 220–229.
36. Roberts,D., Keeling,R., Tracka,M., van der Walle,C.F., Uddin,S., Warwicker,J. and Curtis,R. (2014) The role of electrostatics in protein–protein interactions of a monoclonal antibody. *Mol. Pharm.*, **11**, 2475–2489.
37. Roberts,D., Keeling,R., Tracka,M., van der Walle,C.F., Uddin,S., Warwicker,J. and Curtis,R. (2015) Specific ion and buffer effects on protein–protein interactions of a monoclonal antibody. *Mol. Pharm.*, **12**, 179–193.
38. Gramzow,L., Ritz,M.S. and Theißen,G. (2010) On the origin of MADS-domain transcription factors. *Trends Genet.*, **26**, 149–153.
39. Kaufmann,K., Melzer,R. and Theißen,G. (2005) MIKC-type MADS-domain proteins: structural modularity, protein interactions and network evolution in land plants. *Gene*, **347**, 183–198.
40. Maerkl,S.J. and Quake,S.R. (2009) Experimental determination of the evolvability of a transcription factor. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 18650–18655.
41. Erwin,D.H., Laflamme,M., Tweedt,S.M., Sperling,E.A., Pisani,D. and Peterson,K.J. (2011) The cambrian conundrum: early divergence and later ecological success in the early history of animals. *Science*, **334**, 1091–1097.
42. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2014) Ensembl 2014. *Nucleic Acids Res.*, **42**, D749–D755.
43. Kersey,P.J., Allen,J.E., Christensen,M., Davis,P., Falin,L.J., Grabmueller,C., Hughes,D. S.T., Humphrey,J., Kerhornou,A., Khobova,J. *et al.* (2014) Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res.*, **42**, D546–D552.
44. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2011) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
45. Csűös,M. (2010) Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, **26**, 1910–1912.
46. R Core Team and others. (2012) R: a language and environment for statistical computing. http://cran.case.edu/web/packages/dplR/vignettes/timeseries-dplR.pdf.
47. Wickham,H. (2009) *ggplot2: Elegant Graphics for Data Analysis*, Springer, NY.
48. Hunter,J.D. (2007) Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95.
49. Alexa,A. and Rahnenführer,J. (2010) topGO: enrichment analysis for gene ontology. *R package version 2.12*. Available from: http://www.bioconductor.org/packages/2.12/bioc/html/topGO.html.
50. Brawand,D., Soumillon,M., Necsulea,A., Julien,P., Csárdi,G., Harrigan,P., Weier,M., Liechti,A., Aximu-Petri,A., Kircher,M. *et al.* (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343–348.
51. Petryszak,R., Burdett,T., Fiorelli,B., Fonseca,N.A., Gonzalez-Porta,M., Hastings,E., Huber,W., Jupp,S., Keays,M., Kryvych,N. *et al.* (2014) Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **42**, D926–D932.
52. Wagner,G.P., Kin,K. and Lynch,V.J. (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.*, **131**, 281–285.
53. Wild,F. (2014) lsa: latent semantic analysis. http://CRAN.R-project.org/package=lsa.
54. Jones,E., Oliphant,T., Peterson,P. *et al.* (2001) *SciPy: Open source scientific tools for Python*. http://www.scipy.org/ (28 January 2016, date last accessed).
55. Morgenstern,B. and Atchley,W.R. (1999) Evolution of bHLH transcription factors: modular evolution by domain shuffling? *Mol. Biol. Evol.*, **16**, 1654–1663.
56. McIntosh,B.E., Hogenesch,J.B. and Bradfield,C.A. (2010) Mammalian Per-Arnt-Sim proteins in environmental adaptation. *Annu. Rev. Physiol.*, **72**, 625–645.
57. Mei,Q. and Dvornyk,V. (2014) Evolution of PAS domains and PAS-containing genes in eukaryotes. *Chromosoma*, **123**, 385–405.
58. Bornberg-Bauer,E., Rivals,E. and Vingron,M. (1998) Computational approaches to identify leucine zippers. *Nucleic Acids Res.*, **26**, 2740–2746.
59. Karin,M., Liu,Z.-g. and Zandi,E. (1997) AP-1 function and regulation. *Curr. Opin. Cell Biol.*, **9**, 240–246.
60. Shaulian,E. and Karin,M. (2001) AP-1 in cell proliferation and survival. *Oncogene*, **20**, 2390–2400.
61. Bürglin,T.R. (2011) Homeodomain subtypes and functional diversity. In: Hughes,TR (ed). *A Handbook of Transcription Factors*. Springer, The Netherlands, pp. 95–122.
62. Gehring,W.J., Affolter,M. and Burglin,T. (1994) Homeodomain proteins. *Annu. Rev. Biochem.*, **63**, 487–526.
63. Robinson-Rechavi,M., Garcia,H.E. and Laudet,V. (2003) The nuclear receptor superfamily. *J. Cell Sci.*, **116**, 585–586.
64. Pardee,K., Necakov,A.S. and Krause,H. (2011) Nuclear receptors: small molecule sensors that coordinate growth, metabolism and reproduction. In: Hughes,TR (ed). *A Handbook of Transcription Factors*, Springer, Dordrecht, pp. 123–153.
65. Glass,C.K. (1994) Differential recognition of target genes by nuclear receptor monomers, dimers, and heterodimers. *Endocr. Rev.*, **15**, 391–407.
66. Görner,W., Durchschlag,E., Martinez-Pastor,M.T., Estruch,F., Ammerer,G., Hamilton,B., Ruis,H. and Schüller,C. (1998) Nuclear localization of the C2H2 zinc finger protein Msn2p is regulated by stress and protein kinase A activity. *Genes Dev.*, **12**, 586–597.

67. Nowick,K., Fields,C., Gernat,T., Caetano-Anolles,D., Kholina,N. and Stubbs,L. (2011) Gain, loss and divergence in primate zinc-finger genes: a rich resource for evolution of gene regulatory differences between species. *PLoS One*, **6**, e21553.
68. Nowick,K., Carneiro,M. and Faria,R. (2013) A prominent role of KRAB-ZNF transcription factors in mammalian speciation? *Trends Genet.*, **29**, 130–139.
69. Ho,W.C., Fitzgerald,M.X. and Marmorstein,R. (2006) Structure of the p53 core domain dimer bound to DNA. *J. Biol. Chem.*, **281**, 20494–20502.
70. Levine,A.J. and Oren,M. (2009) The first 30 years of p53: growing ever more complex. *Nat. Rev. Cancer*, **9**, 749–758.
71. Reece-Hoyes,J.S., Deplancke,B., Shingles,J., Grove,C.A., Hope,I.A. and Walhout,A.J. (2005) A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks. *Genome Biol.*, **6**, R110.
72. Abegglen,L.M., Caulin,A.F. and Chan,A. (2015) Potential mechanisms for cancer resistance in elephants and comparative cellular response to dna damage in humans. *JAMA*, **314**, 1850–1860.
73. Uno,Y., Nishida,C., Takagi,C., Ueno,N. and Matsuda,Y. (2013) Homoeologous chromosomes of *Xenopus laevis* are highly conserved after whole-genome duplication. *Heredity*, **111**, 430–436.
74. Meyer,A. and Van de Peer,Y. (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays*, **27**, 937–945.
75. Ohno,S. (1970) *Evolution by gene duplication*. George Alien & Unwin Ltd/Springer-Verlag, London/Berlin, Heidelberg and NY.
76. Sidow,A. (1996) Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.*, **6**, 715–722.
77. Maere,S., Bodt,S.D., Raes,J., Casneuf,T., Montagu,M.V., Kuiper,M. and de Peer,Y.V. (2005) Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 5454–5459.
78. Veron,A.S., Kaufmann,K. and Bornberg-Bauer,E. (2007) Evidence of interaction network evolution by whole-genome duplications: a case study in MADS-Box proteins. *Mol. Biol. Evol.*, **24**, 670–678.
79. Björklund,A.K., Ekman,D., Light,S., Frey-Skött,J. and Elofsson,A. (2005) Domain rearrangements in protein evolution. *J. Mol. Biol.*, **353**, 911–923.
80. Kersting,A.R., Mizrachi,E., Bornberg-Bauer,E. and Myburg,A.A. (2015) Protein domain evolution is associated with reproductive diversification and adaptive radiation in the genus Eucalyptus. *New Phytol.*, **206**, 1328–1336.
81. Moore,A.D., Grath,S., Schüler,A., Huylmans,A.K. and Bornberg-Bauer,E. (2013) Quantification and functional analysis of modular protein evolution in a dense phylogenetic tree. *Biochim. Biophys. Acta*, **1834**, 898–907.
82. Projecto-Garcia,J., Jollivet,D., Mary,J., Lallier,F.H., Schaeffer,S.W. and Hourdez,S. (2015) Selective forces acting during multi-domain protein evolution: the case of multi-domain globins. *Springerplus*, **4**, 354.
83. Iuchi,S. (2001) Three classes of C2H2 zinc finger proteins. *Cell. Mol. Life Sci.*, **58**, 625–635.
84. Stubbs,L., Sun,Y. and Caetano-Anolles,D. (2011) Function and evolution of C2H2 zinc finger arrays. In: Hughes,TR (ed). *A Handbook of Transcription Factors*. Springer, Dordrecht, pp. 75–94.
85. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 14863–14868.
86. 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. *Nat. Genet.*, **25**, 25–29.
87. Stuart,J.M., Segal,E., Koller,D. and Kim,S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.
88. Freilich,S., Massingham,T., Bhattacharyya,S., Ponsting,H., Lyons,P.A., Freeman,T.C. and Thornton,J.M. (2005) Relationship between the tissue-specificity of mouse gene expression and the evolutionary origin and function of the proteins. *Genome Biol.*, **6**, R56.
89. Freilich,S., Massingham,T., Blanc,E., Goldovsky,L. and Thornton,J.M. (2006) Relating tissue specialization to the differentiation of expression of singleton and duplicate mouse proteins. *Genome Biol.*, **7**, R89.
90. Thomas,J.H. and Schneider,S. (2011) Coevolution of retroelements and tandem zinc finger genes. *Genome Res.*, **21**, 1800–1812.
91. Jacobs,F. M.J., Greenberg,D., Nguyen,N., Haeussler,M., Ewing,A.D., Katzman,S., Paten,B., Salama,S.R. and Haussler,D. (2014) An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature*, **516**, 242–245.
92. Zhang,J. (2003) Evolution by gene duplication: an update. *Trends Ecol. Evol.*, **18**, 292–298.
93. Moore,A.D. and Bornberg-Bauer,E. (2012) The dynamics and evolutionary potential of domain loss and emergence. *Mol. Biol. Evol.*, **29**, 787–796.
94. Donoghue,P. C.J. and Purnell,M.A. (2005) Genome duplication, extinction and vertebrate evolution. *Trends Ecol. Evol.*, **20**, 312–319.
95. Fawcett,J.A., Maere,S. and de Peer,Y.V. (2009) Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proc. Natl. Acad. Sci.*, **106**, 5737–5742.
96. Honma,T. and Goto,K. (2001) Complexes of MADS-box proteins are sufficient to convert leaves into floral organs. *Nature*, **409**, 525–529.
97. Voordeckers,K., Brown,C.A., Vanneste,K., van der Zande,E., Voet,A., Maere,S. and Verstrepen,K.J. (2012) Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS Biol.*, **10**, e1001446.