DNA-Binding Specificity Changes in the Evolution of Forkhead Transcription Factors

DNA Binding Specificity Changes in the Evolution of Forkhead Transcription Factors

So Nakagawa[a,b,1,2], Stephen S. Gisselbrecht[c,1], Julia M. Rogers[c,e,1], Daniel L. Hartl[a,3] and Martha L. Bulyk[c,d,e,3]

[a]Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA [b]Center for Information Biology, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan [c]Division of Genetics, Department of Medicine, [d]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA. [e]Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, MA 02138, USA. [1]S.N., S.S.G., and J.M.R. contributed equally to this work. [2]Present address: Department of Molecular Life Science, Tokai University School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1193, Japan [3]To whom correspondence should be addressed.

Submitted to Proceedings of the National Academy of Sciences of the United States of America

The evolution of transcriptional regulatory networks entails the expansion and diversification of transcription factor (TF) families. The forkhead family of TFs, defined by a highly conserved winged helix DNA-binding domain (DBD), has diverged into dozens of subfamilies in animals, fungi, and related protists. We have used a combination of maximum likelihood phylogenetic inference and independent, comprehensive functional assays of DNA binding capacity to explore the evolution of DNA binding specificity within the forkhead family. We present converging evidence that similar alternative sequence preferences have arisen repeatedly and independently in the course of forkhead evolution. The vast majority of DNA binding specificity changes we observed is not explained by alterations in the known DNA-contacting amino acid residues conferring specificity for canonical forkhead binding sites. Intriguingly, we have found forkhead DBDs that retain the ability to bind very specifically to two completely distinct DNA sequence motifs. We propose an alternate specificity-determining mechanism whereby conformational rearrangements of the DBD broaden the spectrum of sequence motifs that a TF can recognize. DNA binding bispecificity suggests a new source of modularity and flexibility in gene regulation and may play an important role in the evolution of transcriptional regulatory networks.

DNA binding specificity | evolution | transcription factor

Introduction

The regulation of gene expression by the interaction of sequence-specific transcription factors (TFs) with target sites (cis-regulatory elements) near their regulated genes is a central mechanism by which organisms interpret regulatory programs encoded in the genome to develop and interact with their environment. The emergence of new species has depended in part on the evolution of the network of interactions by which an organism's TFs control gene expression. Much attention has been paid to changes in cis-regulatory sequences over evolutionary time, as these changes can result in incremental modifications of organismal phenotypes without large-scale "rewiring" of transcriptional regulatory networks that would result from changes in TF DNA binding specificity (1). Nevertheless, TFs and their DNA binding specificities have changed over time (2). Gene duplication, followed by divergence of the resulting redundant TFs, has resulted in the emergence of families of paralogous TFs with diversified DNA binding specificities and functions (3). Thus, identifying mechanisms by which related DNA-binding domains (DBDs) have acquired novel specificities is important for understanding TF evolution.

The forkhead box (Fox) family of TFs spans a wide range of species, and is one of the largest classes of TFs in humans. In metazoans, Fox proteins have vital roles in development of a variety of organ systems, metabolic homeostasis, and regulation of cell cycle progression, while fungal Fox proteins are involved in cell cycle progression and the expression of ribosomal proteins. The Fox family of TFs shares a conserved DBD that is structurally identifiable as a subgroup of the much larger winged helix superfamily, which includes both sequence-specific DNA-binding proteins and linker histones, which appear to bind DNA nonspecifically (4, 5). Proteins with unambiguous sequence homology to the forkhead domain are present throughout opisthokonts—the phylogenetic grouping which includes all descendants of the last common ancestor of animals and fungi—but have diverged so extensively over approximately one billion years of evolution that distantly related Fox proteins are not generally alignable outside the forkhead domain (6, 7). Moreover, distantly related Fox-like domains have been found in Amoebozoa, a sister group to opisthokonts (8). Three distinct subfamilies (Fox1 through Fox3) of fungal Fox proteins have been identified. Metazoan Fox proteins are classified into 19 subfamilies (FoxA through FoxS), some of which have been further subdivided on phylogenetic grounds.

The Fox domain itself is roughly 80–100 amino acids (a.a.) in length and, like other winged helix domains, comprises a bundle of three α-helices connected via a small β-sheet to a pair of loops or "wings". In available structures of forkhead domain-DNA complexes, helix 3 forms a canonical recognition helix positioned in the major groove of the DNA target site by the helical bundle, while the wings, which often contain a poorly alignable region rich in basic residues, lie along the adjacent DNA backbone (9-13).

Several groups have studied the evolutionary history of the family using multiple sequence alignment and phylogenetic inference methods; however, the results of these studies are in many cases inconsistent. Published forkhead phylogenies lack statistical support for deep branches and the relative positions of forkhead subfamilies, especially of the fungal groups (14, 15). Thus, the relationships among Fox genes have remained unclear.

In separate studies, the DNA binding specificities of various forkhead proteins have been examined. In most cases, *in vitro* binding has been observed to variants of the canonical forkhead target sequence RYAAAYA (16-21), which we refer to as the forkhead primary (FkhP) motif (Figure 1). A similar variant, AHAACA, has been observed in *in vitro* selection (SELEX) (17) and protein-binding microarray (PBM) experiments (20); this specificity appears to be common to several Fox proteins, and we refer to it as the forkhead secondary (FkhS) motif (22). However, a SELEX study of the FoxN1 TF mutated in the famous *nude*

Reserved for Publication Footnotes

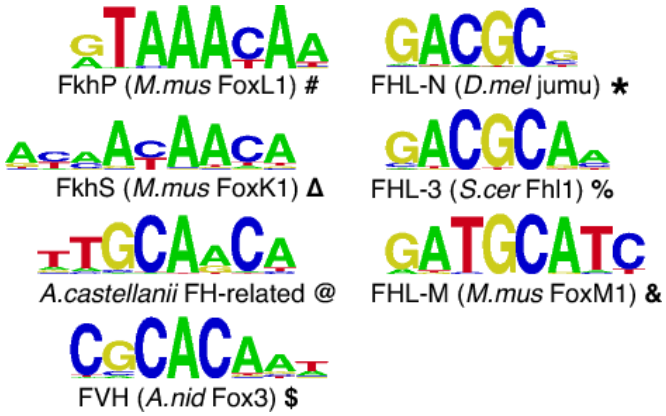


Fig. 1. DNA binding site motifs bound by forkhead domain proteins. A representative member of each class of binding site discussed in the text is shown. Bold symbols are used to represent binding specificities in subsequent figures.

mouse identified an entirely different sequence, ACGC, as its preferred binding site (23). The closely related *Mus musculus* FoxN4 has been shown to bind ACGC *in vivo* (24). A PBM survey of *Saccharomyces cerevisiae* TFs identified a very similar sequence, GACGC, as the binding site of the Fox3 factor Fhl1 (19); we therefore refer to the GACGC site as the FHL motif (Figure 1).

Previous work on differences in forkhead DNA binding specificity has focused on preferential recognition of FkhP and FkhS variants by forkhead proteins (17, 18). Contrary to the common mechanism of varying specificity by changing a.a. residues that make base-specific DNA contacts (25), the positions in the forkhead recognition helix that make base-specific contacts are conserved across proteins with different binding specificities (9, 17). In sub-domain swap experiments, a 20-a.a. region immediately N-terminal to the recognition helix was shown to switch DNA-binding specificities between forkhead proteins (17). Interestingly, this region has been shown by NMR to adopt different secondary structures in forkheads with distinct DNA binding specificities (26). However, a similar analysis of sequence features conferring binding to the FHL motif has not been performed.

The observation of binding to such different sequences –RYAAAYA and GACGC –within widely diverged members of the Fox family raises the question of how the binding specificity of these proteins has evolved. We have addressed this question using a combined phylogenetic and biochemical approach. We conducted a phylogenetic analysis of Fox domains from 10 metazoans, 30 fungi, and 25 protists (Table S1). We chose these species based on their evolutionary importance and annotation level (27) (Figure S1). For example, we included *Spizellomyces punctatus* and *Fonticula alba*, since they are very close to the root of fungi and a closely related outgroup, respectively. We considered conserved splice junctions along with multiple sequence alignment to infer the phylogeny. We assayed DNA binding specificity *in vitro* using universal PBM technology, in which a DNA-binding protein is applied to a double-stranded DNA microarray containing 32 replicates of all possible 8-bp sequences (8-mers) and is fluorescently labeled, permitting the exhaustive cataloguing of the range of sequences a protein can recognize (28). We analyzed the binding specificities of 30 forkhead proteins, combining previously published data for 9 proteins with data for 21 proteins that we newly characterized for this study (Table S4). We focused on proteins from clades where we had previously observed alternate binding specificities and clades of unknown specificity. By using two orthogonal means of evaluating the same proteins, we obtain a much richer picture of the evolutionary trajectory of changes

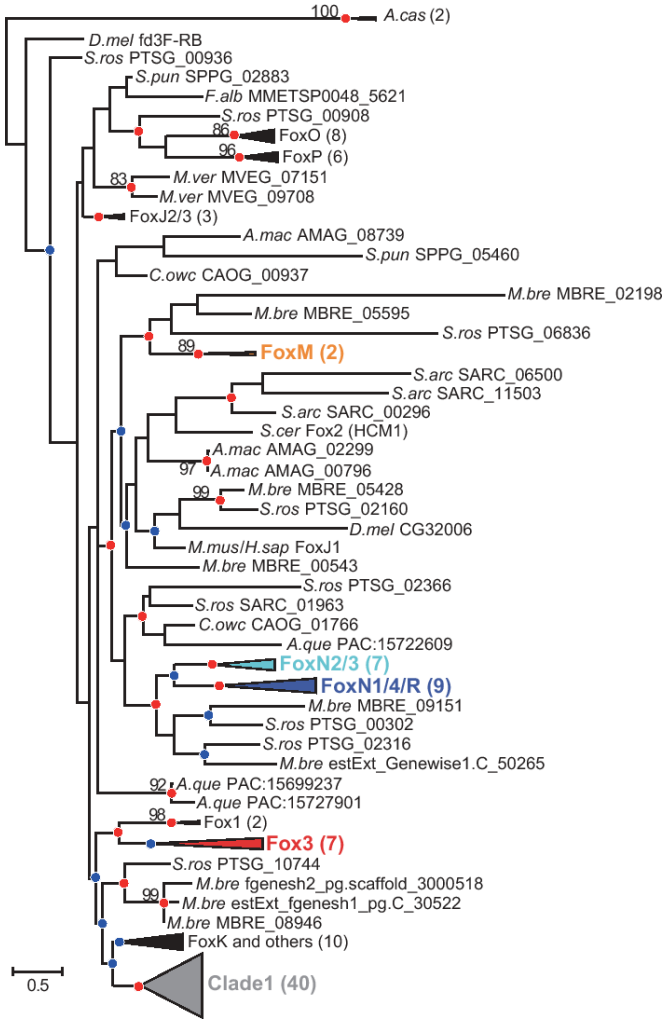


Fig. 2. Maximum likelihood phylogenetic tree of forkhead domains. This compact tree was constructed for presentation purposes from a representative subset of phylogenetically informative species: metazoans mouse, fly and sponge; choanoflagellates *Salpingoeca rosetta* and *Monosiga brevicollis*; *Capsaspora owczarzaki* and *Sphaeroforma arctica* from Ichthyosporea; *Saccharomyces cerevisiae* from Dikarya; *Allomyces macrogynus* from Blastocladiomycota; *Spizellomyces punctatus* from Chytridiomycota; *Mortierella verticillata* from Mortierellomycotina; *Fonticula alba* from Nucleariida; *Acanthamoeba castellanii* from Amoebozoa. Nodes supported with strong likelihood ratios are indicated with red circles (aLRT ≥99%) or blue circles (aLRT ≥95%); bootstrap support values are shown for nodes with ≥80% support. Clades containing alternate binding specificities are highlighted in color (see text). Importantly, the groupings of subfamilies in this tree and the complete tree with all Fox domains are almost identical to each other (see Figure S2).

in TF DNA binding specificity than either analysis alone can provide.

Results

The published observation of roughly the same alternate binding motif (FHL) for metazoan FoxN1/4 and fungal Fox3 suggests the parsimonious hypothesis that they derive from a common FHL-binding ancestral protein in the last common ancestor of opisthokonts. To explore this hypothesis, we performed phylogenetic inference on a broad group of Fox domain sequences (see Materials and Methods), spanning 623 genes from 65 species (Table S1, Figure S1)). We included two distantly related forkhead domains from the opisthokont sister group Amoebozoa as an outgroup. After removing partial domain sequences and those identical throughout the Fox domain, we used 529 Fox domain

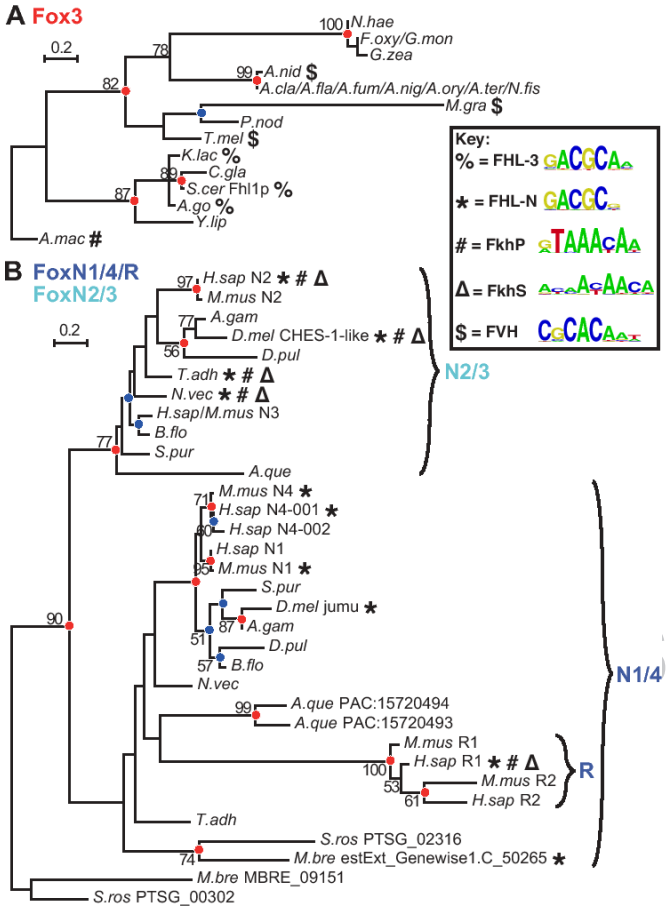


Fig. 3. Detailed analysis of Fox3 and FoxN subfamilies. Maximum likelihood phylogenetic trees for Fox domains from a broader range of species for **(A)** fungal Fox3, and **(B)** holozoan FoxN/R clades. Red and blue circles indicate node support as in Figure 2. Bold symbols represent binding capacity for different motif classes as defined in Figure 1.

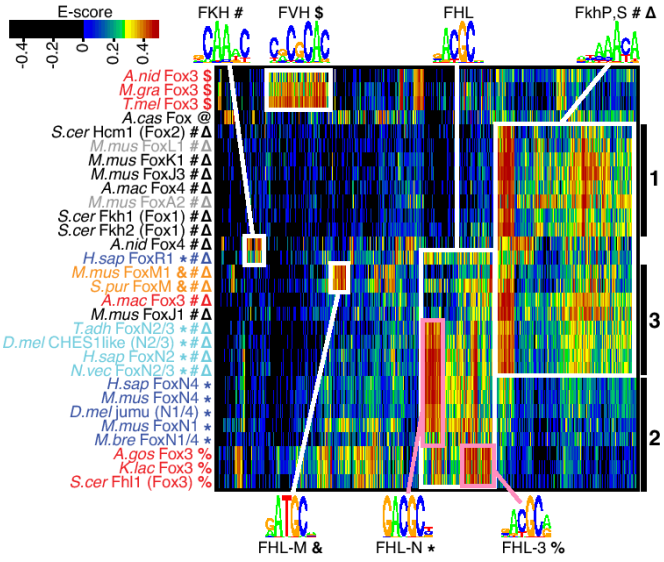


Fig. 4. Biclustering of Fox domain binding data reveals multiple functional classes. E-score binding profiles were clustered both by protein (rows) and by contiguous 8-mer (columns) for any 8-mer bound (E-score ≥ 0.35) by at least one assayed Fox protein. Fox domains fall into functional classes (bold symbols represent binding capacity for different motifs as defined in Figure 1) that do not uniformly correlate with phylogeny (protein names are colored by phylogenetic grouping as in Figure 2). Cluster 1 (black bar) comprises proteins specific only for the FkhP,S motifs, cluster 2 proteins are specific only for FHL variants, and cluster 3 proteins have more complex specificity; see text for details. Sequence motifs shown were generated by alignment of the indicated clusters of 8-mers and are for visualization purposes only.

sequences (340 non-redundant, Table S1). We constructed a complete maximum likelihood (ML) tree of all non-redundant Fox domain sequences (Figure S2). For each branch, the approximate Likelihood-Ratio Test (aLRT) and 100 bootstrap replicates were used to evaluate support for inferred relationships (see Materials and Methods). For presentation purposes, we constructed a ML tree of 262 (133 non-redundant) Fox domains from selected informative species (Figure 2, Table S1).

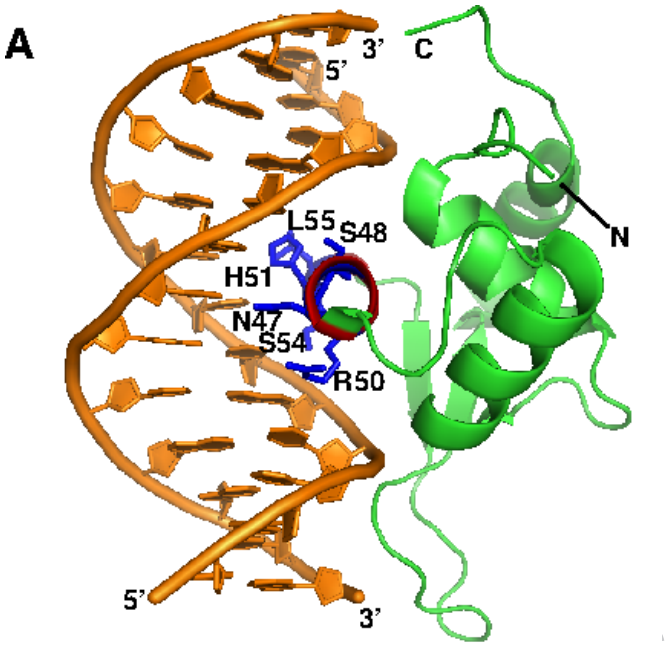
Various portions of the phylogeny could be determined with high confidence. Our analysis recovered the previously identified subfamily relationships between Fox proteins, as well as identifying a new fungal group (Fox4) not previously observed because it is not represented in *S. cerevisiae*. However, the structure of the deep portions of the Fox tree could not be resolved for two major reasons. First, the number of alignable positions within the Fox domain is too small to resolve the phylogenetic history of such a broadly and deeply diverged family, and regions outside the domain are not alignable among distantly related members. Second, some Fox genes appear to have evolved through gene conversion and/or crossover events (15), as evinced by the appearance of species-specific Fox domain signatures.

The ML tree inferred here strongly supports the hypothesis of Larroux *et al.* that a monophyletic group of forkhead domains (which they refer to as clade I) emerged in the common ancestor of metazoans (14) (aLRT value = 0.9999, bootstrap value = 4%) (Figure 2). Additionally, there is a splice site between a.a. positions 46 and 47 in the Pfam Fork_head domain hidden Markov model (29) conserved in various clade II forkhead proteins across kingdoms; no clade I genes share this splice site, further supporting the monophyly of clade I in metazoans.

Surprisingly, there is no support for a tree topology in which metazoan FoxN and fungal Fox3 subfamilies form a monophyletic, FHL-binding clade. A tree containing a FoxN+3 clade (Figure S3A) is significantly less likely than the observed tree ($p < 10^{-8}$, likelihood ratio test), and likelihood maximization using this as a starting tree separates the FoxN and Fox3 clades (Figure S3B,C). Moreover, we see separate, well-supported clades (aLRT values ≥ 0.99) combining each of these groups with others that bind only the FkhP and FkhS motifs (Figure 2). This result suggests that FHL binding capacity evolved twice independently within the family, and led us to examine these two subgroups in more detail.

A phylogenetic tree constructed from only fungal Fox3 domains (Figure 3A) is much more stable than the larger, more complex tree, with acceptable bootstrap support at major branch points; moreover, it follows the species tree closely (see Figure S1), suggesting radiation of a family of orthologs. The most basally diverged member of this group, *Allomyces macrogynus* Fox3, binds only the canonical FkhP and FkhS motifs (Figure 3A and Figure 4), providing experimental support for the hypothesis that FHL binding arose within the Fox3 clade after its divergence from other forkhead domains. The remaining Fox3 proteins considered here fall into two distinct groups. Those most closely related to Fhl1 (*S. cerevisiae* Fox3) show the same FHL-binding specificity, binding the FkhP,S motifs no better than non-forkhead proteins (percent signs in Figure 3A). Members of the other group, including *Aspergillus nidulans* Fox3, bind another motif entirely, which we term the Forkhead Variant Helix (FVH) motif (dollar signs in Figure 3A; see Figure 1), with no specific binding to either the FkhP,S or FHL motifs.

Similarly, the phylogeny of the holozoan FoxN subfamily is relatively stable (Figure 3B). Our analysis supports the existence



B

	45	50	55
M.mus FoxA2 #Δ	Q R W Q N S	I R H S L	S F N D
M.mus FoxK1 #Δ	K G W Q N S	I R H N L	S L N R
S.cer Fkh1 (Fox1) #Δ	M A W Q N S	V R H N L	S L N K
M.mus FoxJ1 #Δ	P T W Q N S	I R H N L	S L N K
S.pur FoxM &#Δ	L G W K N S	I R H N L	S L H K
M.mus FoxM1 &#Δ	P G W K N S	I R H N L	S L H D
T.adh FoxN2/3 *#Δ	T G W K N S	V R H N L	S L N K
H.sap FoxN2 *#Δ	T G W K N S	V R H N L	S L N K
H.sap FoxR1 *#Δ	E G W K N T	V R H N L	C F R D
M.bre FoxN1/4 *	S G W K N S	I R H N L	S L N K
H.sap FoxN4 *	D G W K N S	V R H N L	S L N K
A.mac Fox3 #Δ	H S W Q N S	V R H N L	S L N R
S.cer Fhl1 (Fox3) %	D G W Q S S	V R H N L	S L N K
A.nid Fox3 $	E G W T S S	V R H N L	N G N G
A.cas Fox @	R N W K A A	I R S C L	S K N E

Fig. 5. Canonical Fox base-contacting residues do not explain most alternate specificity. (A) A previous co-crystal structure of mouse FoxK1 bound to the canonical FkhP site GTAAACA (PDB ID 2c6y, (10)). The recognition helix is highlighted; side chains are shown in blue and labeled for those a.a. that make base-specific contacts in at least two existing structures. **(B)** Protein sequence alignment of the recognition helix (red underscore) and adjacent positions for a sample of Fox domains representing various specificity classes (bold symbols represent binding capacity for different motif classes as defined in Figure 1). Numbers above alignment represent positions within the Pfam Fork_head domain HMM. **Supporting Information:**

of a fundamental split into FoxN1/4 and N2/3 clades, with FoxR (initially called N5 (30)) placed within the N1/4 group (14). As expected, FoxN1 and other N1/4 proteins are highly specific for the FHL motif. Surprisingly, all FoxN2/3 proteins assayed by PBMs exhibited high sequence specificity for both the FkhP,S and FHL motifs (see Figure 4). For example, the top two 8-mers (ranked by PBM enrichment (E) score, which indicates the preference of a protein for every possible 8-mer (28)) bound by the *Drosophila melanogaster* FoxN2/3 protein CHES-1-like are ATAAACAA and GTAAACAA, perfectly matching the FkhP consensus, while the next two are the FHL matches GACGC-TAA and GACGCTAT. FoxR1 also shows bispecificity, despite presumably arising from an FHL-specific N1/4 ancestor. Such bispecificity for two seemingly unrelated sequence motifs by a single DBD (*i.e.*, excluding proteins with multiple DNA-binding subdomains) has not been observed previously.

Consistent with the hypothesis that FHL binding arose independently in the fungal Fox3 and holozoan FoxN groups, we observed slight variations between the versions of the FHL motif bound by each of these two groups. Specifically, all tested FHL-binding Fox3 proteins strongly prefer A immediately 3' to the core GACGC, which we refer to as the FHL-3 motif, while FHL motifs from FoxN/R proteins all strongly disfavor A in that position, a variant we refer to as the FHL-N motif (Figure 1). Similarly, *Homo sapiens* FoxR1 (which appears to have regained FkhP,S binding from an FHL-only ancestor) strongly prefers a C at position 2 of the FkhP motif, while other FkhP-binding Fox domains strongly prefer T at that position (Figure 4 and Figure S4).

The unexpected variety in Fox domain binding specificity led us to perform additional PBM experiments on a range of Fox domains, focusing on representative proteins from other clade II groups such as Fox4 and FoxM, and assemble them with previously published PBM data (Figure S4, Table S2, Table S3). In addition to finding more examples of proteins that exhibit the sequence preferences described above, we also discovered a third instance of binding to an FHL-like motif. Two metazoan FoxM proteins exhibit high specificity for the FkhP and FkhS motifs, and for a third FHL variant, GATGC, which we refer to as FHL-M. The most preferentially bound 8-mer matching this motif is an overlapping inverted repeat, GATGCATC; human FoxM1 has previously been shown to bind overlapping multimers of the FkhP motif *in vitro*, which suggests that these two FoxM proteins might bind as dimers to GATGCATC. Phylogenetic analysis strongly supports an independent origin of the FoxM subfamily from FoxN ($p < 10^{-4}$, likelihood ratio test, Figure S3D), in that each subfamily is more closely related to proteins that bind only FkhP and FkhS than to each other, suggesting that this represents yet a third independent emergence of a form of FHL binding (FHL-M), with each one characterized by slight differences in DNA sequence preference (Figure 1). As in the case of FoxN and Fox3, ML inference with a starting tree containing a FoxM+N clade leads to separation of the subfamilies (Figure S3E,F).

Biclustering of the 30 total Fox proteins and bound 8-mers according to PBM enrichment (E) scores reveals three major functional protein classes (Figure 4). The first prominent cluster of proteins is characterized by specificity only for the FkhP and FkhS motifs. Binding to these motifs tracks together across proteins; the motif constructed from these 8-mers is an average over both motifs. This FkhP,S-binding cluster comprises representatives of widely varying subfamilies, including clade I (*M. musculus* FoxA2 and FoxL1), metazoan clade II (*M. musculus* FoxJ3 and FoxK1), and fungal Fox1, Fox2, and Fox4 (*S. cerevisiae* Fkh1, Fkh2, and Hcm1, and *A. macrogynus* Fox4). This broad distribution of FkhP,S binding specificity supports the hypothesis that it is the ancestral binding specificity of the entire forkhead family.

The second large cluster comprises domains that are uniquely specific for the FHL motif: holozoan FoxN1/4 and fungal Fox3 (*S. cerevisiae* subgroup). This cluster is further divided into holozoan and fungal groups, based on preference for the FHL-N versus FHL-3 variants, as described above.

The third major cluster combines several proteins exhibiting broad specificity. The bispecific metazoan FoxN2/3 and FoxM subfamilies are present in this cluster, along with *M. musculus* FoxJ1 and *A. macrogynus* Fox3, both of which show strong preference for the FkhP and FkhS motifs and weaker preference for the FHL motif variants.

One of the forkhead-like domains from the non-opisthokont *Acanthamoeba castellanii* did not fall into any of these three clusters, as it binds another distinct motif (see Figure 1, Figure S4). These binding differences are associated with widespread differences in the recognition helix (Figure 5A). Indeed, altered recognition positions (Figure 5B) can clearly explain the non-FkhP,S specificities of the forkhead-related protein from *A. castellanii* and *A. nidulans* Fox3; furthermore, there are sufficient differences in the recognition helix of *H. sapiens* FoxR1 that it is perhaps surprising that its specificity is so similar to that of other Fox proteins. Surprisingly, however, the majority of specificity changes in the Fox family, including FHL binding and bispecificity, do not correlate with changes in canonical specificity-determining positions. Indeed, although *H. sapiens* FoxN4 is highly specific for only the FHL motif, and *H. sapiens* FoxN2 is bispecific and robustly binds FkhP and FkhS sites as well as the FHL motif, these two FoxNs are identical throughout the entire recognition helix; thus, the inability of FoxN4 to recognize FkhP sites is not strictly a function of the canonical DNA-contacting residues in the recognition helix.

Discussion

The previously unappreciated diversity in DNA binding specificity of Fox domain TFs that we have discovered raises the question of how specificity has evolved in this family. We have presented evidence that major changes in specificity have occurred separately in three different Fox subfamily lineages. In fungal Fox3 proteins, two different alternate specificities (FHL-3 and FVH) have arisen, with alteration of the canonical recognition positions in the FVH-binding but not the FHL-3-binding proteins. In metazoan FoxM proteins, binding to the canonical FkhP and FkhS sites has been supplemented with binding to a very different site, the FHL-M motif, with the same proteins binding well to both motifs. In addition, in the holozoan FoxN subfamily some proteins (FoxN2/3) exhibit this kind of bispecificity for two very different motifs (FkhP,S and FHL-N), while others (FoxN1/4) have completely lost the ability to bind the classic forkhead site (FkhP,S) in favor of the FHL-N motif. Finally, a derived subfamily unique to vertebrates (FoxR) appears to have regained specificity for a variant of the canonical FkhP motif from a more recent, exclusively FHL-specific ancestor. Formally, it is possible that lineages containing only proteins that bind only the FkhP,S sequences are derived from a more promiscuously binding ancestor with loss of FHL binding; however, this model would require a much larger number of specificity changes than the model we put forth here. Moreover, each instance of specificity change inferred from phylogenetic analyses is corroborated by minor but consistent differences in the motifs that have arisen; for example, all FoxN proteins bind to a version of the FHL motif that is distinguishable from the very similar FHL motif of fungal Fox3 proteins by preferences at a flanking position.

Our strategy of combining phylogenetic inference with comprehensive assays of DNA binding specificity permits us to study the evolution of DNA binding specificity in more detail using information from these complementary approaches. The monophyly of clade I, for example, is supported both by a high-confidence node in the inferred phylogeny and by the observed uniformity of binding specificity within this group. In the absence of phylogenetic analyses, the observation of an alternate specificity (GAYGC) appearing three times in different Fox domain subfamilies would lead to a parsimonious hypothesis that one ancestral FHL-binding forkhead domain arose before the last common ancestor of metazoa and fungi and gave rise to fungal Fox3 and metazoan FoxM and N groups. However, this hypothesis is strongly refuted by ML phylogenetic inference, which instead suggests independent origins of all three groups of alternate-specificity proteins. Further support for this surprising model comes from the observation that fine differences in FHL specificity distinguish these three groups, as discussed above.

This model raises the question of how such similar alternate specificities could have arisen independently in three different forkhead lineages. In the group of Fox3 proteins from fungi related to *A. nidulans*, the alteration in specificity to the FVH motif with concomitant loss of binding to FkhP,S sequences might be due to the extensive changes observed in the recognition helix. However, the appearances of the FHL motif variants during forkhead evolution, whether along with FkhP binding in bispecific proteins or as a replacement, do not correlate with any changes at a.a. positions known to specify FkhP binding, and suggest an alternate mechanism for changes in DNA binding specificity.

We propose that the existence of bispecific proteins that bind both FkhP,S and FHL sequences with high specificity points to a possible explanation — that some Fox domain proteins which bind strongly to the FkhP site can achieve an alternate conformation which supports recognition of the FHL motif. It is intriguing, in the context of this observation, that both *M. musculus* FoxJ1 and *A. macrogynus* Fox3 show weak binding to a subset of FHL-containing 8-mers, and exhibit binding similarity to bispecific factors that bind much more strongly and specifically to the FHL motif (see Figure 4). We suggest that the Fox domain can adopt an alternate DNA binding mode, and thus possesses an inherent "evolvability" of DNA sequence specificity that has permitted the emergence of FHL binding multiple independent times.

Allostery is a widespread and fundamental phenomenon in biological regulation, and in principle the use of alternate binding modes to recognize multiple sequence motifs could result in alternate protein interaction surfaces of a TF, thus creating a new regulatory role for the alternate binding motifs as allosteric effectors of interactions with cofactors (31, 32). Exploring the mechanisms of such regulatory consequences will require an approach combining structural studies of distinct TF-DNA complexes, such as those identified here, with *in vivo* analyses of binding site utilization and function. This newly discovered phenomenon of DNA binding bispecificity suggests a novel source of modularity and flexibility in the structure of TFs and transcriptional regulatory networks. Improved understanding of the evolution of TF binding specificity will provide insights into the evolution of transcriptional regulatory networks, which ultimately will shed light on the processes underlying the evolution of new body plans and environmental responses.

Materials and Methods

Forkhead sequences

The genome sequences and annotations used in this study are summarized in Table S1. For each annotated protein sequence, we performed a hidden Markov model (HMM) search using HMMER3 (33) with the Fork_head domain (PF00250) in the Pfam database (E-value < 10^{-10}) (29). Using the hit sequences as queries, we conducted iterative homology search using PSI-BLAST (E-value < 10^{-10}) (34). We then constructed a HMM from each multiple alignment of forkhead sequences, and searched against all protein sequences again. All obtained genes are described with their identification method in Table S1. All sequences used for the phylogenetic analysis contain five alpha-helices and three beta-sheets as in human FoxP2 (11).

For phylogenetic analyses, each a.a. sequence of Fox domains was aligned using five multiple sequence alignment programs: a) L-INS-i program in MAFFT (35), b) T-Coffee (36), c) MUSCLE (37)), d) Clustal Omega (38), and e) Clustal W (39). The accuracies of multiple sequence alignments were evaluated by FastSP (40), and the MAFFT alignment was selected by the number of homologous a.a. sites.

Phylogenetic inference

The a.a. replacement models of LG (41) with gamma-distributed rate variation (α = 0.881) were selected for whole forkhead domains, using the Akaike information criterion implemented in PROTTEST 3 (42). Phylogenetic trees were constructed using the maximum-likelihood method in PhyML 3.0 (43) with robustness evaluated by bootstrapping (100 times) (44) and by approximate likelihood-ratio test (aLRT) (45, 46). The starting tree for branch swapping was obtained using a ML tree constructed by RAxML (47). For likelihood ratio tests, two ML trees were constructed from the ML tree in Figure 2, changing the branching pattern of Fox3 and FoxM (Figure S3A and S3B, respectively). RAxML was applied to optimize the lengths of branches

and calculate ML scores (-13422.7 for Figure S3A and -13414.9 for Figure S3B). Comparing the ML score obtained from the tree in Figure 2 (-13406.2), *p*-values were calculated based on the chi-square distribution with one degree of freedom.

Cloning and protein expression

The DBDs of the forkhead proteins, flanked by *attB* recombination sites, were constructed by gene synthesis and cloned into the pUC57 vector (GenScript USA, Inc.). Constructs were transferred to the pDEST15 vector, which provides an N-terminal glutathione S-transferase (GST) tag, using the Gateway recombinational cloning system (Invitrogen). All cloned forkhead domain sequences are provided in Table S4. Proteins were expressed by *in vitro* transcription and translation (IVT) using the PURExpress *in vitro* Protein Synthesis kit (New England BioLabs, Inc.). Concentrations of the expressed GST-fusion proteins were determined by Western blots in comparison to a dilution series of recombinant GST (Sigma).

PBM experiments and analysis

Double-stranding of oligonucleotide arrays and PBM experiments were performed essentially as described previously, except where noted in Table S4, using custom-designed "all 10-mer" arrays in the 4x44K (Agilent Technologies, Inc.; AMADID #015681) or 8x60K (Agilent Technologies, Inc.; AMADID #030236) array format (28, 48). Microarray data quantification, normalization, and motif derivation were performed as described previously (28, 48); some published PBM data (21) were reanalyzed for this study. DNA binding site motif sequence logos were generated using enoLOGOS (49). 8-mer E-score data were collected for any contiguous 8-mer bound (E-score ≥ 0.35) by at least one assayed Fox protein and clustered using the heatmap.2 function in the gplots R package with the Manhattan distance metric.

Author contributions: M.L.B., D.H., S.N., S.S.G., and J.M.R. designed research; D.H. and M.L.B. supervised the research; J.M.R. performed expriments; S.N., S.S.G., and J.M.R. analyzed data; S.N, S.S.G., and J.M.R. wrote the paper.

The authors declare no conflicts of interest.

Acknowledgements.

We thank Matthew W. Brown and Iñaki Ruiz-Trillo for sharing pre-publication forkhead sequences from *F. alba* and *A. castellanii*, Anastasia Vedenko and Leila Shokri for technical assistance, and Anton Aboukhalil, Shamil Sunyaev and Ivan Adzhubey for helpful discussion. This study was supported by a Research Fellowship for Young Scientists from the Japan Society for the Promotion of Science to S.N., and by National Institutes of Health grant # R01 HG003985 to M.L.B. J.M.R. was supported in part by the Molecular Biophysics Training Grant # T32 GM008313 from the National Institutes of Health. This article contains Supporting Information online.

1. Carroll S, Grenier J, & Weatherbee S (2001) *From DNA to Diversity* (Blackwell Science, Malden, MA).
2. Gasch AP, *et al.* (2004) Conservation and evolution of cis-regulatory systems in ascomycete fungi. *PLoS Biol* 2(12):e398.
3. Teichmann SA & Babu MM (2004) Gene regulatory network growth by duplication. *Nat Genet* 36(5):492-496.
4. Gajiwala KS, *et al.* (2000) Structure of the winged-helix protein hRFX1 reveals a new mode of DNA binding. *Nature* 403(6772):916-921.
5. Ramakrishnan V, Finch JT, Graziano V, Lee PL, & Sweet RM (1993) Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. *Nature* 362(6417):219-223.
6. Shimeld SM, Degnan B, & Luke GN (2010) Evolutionary genomics of the Fox genes: origin of gene families and the ancestry of gene clusters. *Genomics* 95(5):256-260.
7. Kaestner KH, Knochel W, & Martinez DE (2000) Unified nomenclature for the winged helix/forkhead transcription factors. *Genes & development* 14(2):142-146.
8. Sebe-Pedros A, de Mendoza A, Lang BF, Degnan BM, & Ruiz-Trillo I (2011) Unexpected repertoire of metazoan transcription factors in the unicellular holozoan Capsaspora owczarzaki. *Mol Biol Evol* 28(3):1241-1254.
9. Clark KL, Halay ED, Lai E, & Burley SK (1993) Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* 364(6436):412-420.
10. Tsai KL, *et al.* (2006) Crystal structure of the human FOXK1a-DNA complex and its implications on the diverse binding specificity of winged helix/forkhead proteins. *J Biol Chem* 281(25):17400-17409.
11. Stroud JC, *et al.* (2006) Structure of the forkhead domain of FOXP2 bound to DNA. *Structure* 14(1):159-166.
12. Littler DR, *et al.* (2010) Structure of the FoxM1 DNA-recognition domain bound to a promoter sequence. *Nucleic Acids Res* 38(13):4527-4538.
13. Boura E, Rezabkova L, Brynda J, Obsilova V, & Obsil T (2010) Structure of the human FOXO4-DBD-DNA complex at 1.9 A resolution reveals new details of FOXO binding to the DNA. *Acta crystallographica. Section D, Biological crystallography* 66(Pt 12):1351-1357.
14. Larroux C, *et al.* (2008) Genesis and expansion of metazoan transcription factor gene classes. *Mol Biol Evol* 25(5):980-996.
15. Wang M, Wang Q, Zhao H, Zhang X, & Pan Y (2009) Evolutionary selection pressure of forkhead domain and functional divergence. *Gene* 432(1-2):19-25.
16. Kaufmann E, Muller D, & Knochel W (1995) DNA recognition site analysis of Xenopus winged helix proteins. *J Mol Biol* 248(2):239-254.
17. Overdier DG, Porcella A, & Costa RH (1994) The DNA-binding specificity of the hepatocyte nuclear factor 3/forkhead domain is influenced by amino-acid residues adjacent to the recognition helix. *Mol Cell Biol* 14(4):2755-2766.
18. Pierrou S, Hellqvist M, Samuelsson L, Enerback S, & Carlsson P (1994) Cloning and characterization of seven human forkhead proteins: binding site specificity and DNA bending. *EMBO J* 13(20):5002-5012.
19. Zhu C, *et al.* (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res* 19(4):556-566.
20. Badis G, *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science* 324(5935):1720-1723.
21. Badis G, *et al.* (2008) A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol Cell* 32(6):878-887.
22. Zhu X, *et al.* (2012) Differential regulation of mesodermal gene expression by Drosophila cell type-specific Forkhead transcription factors. *Development* 139(8):1457-1466.
23. Schlake T, Schorpp M, Nehls M, & Boehm T (1997) The nude gene encodes a sequence-specific DNA binding protein with homologs in organisms that lack an anticipatory immune system. *Proc Natl Acad Sci U S A* 94(8):3842-3847.
24. Luo H, *et al.* (2012) Forkhead box N4 (Foxn4) activates Dll4-Notch signaling to suppress photoreceptor cell fates of early retinal progenitors. *Proceedings of the National Academy of Sciences* 109(9):E553–E562.
25. Lehming N, Sartorius J, Kisters-Woike B, von Wilcken-Bergmann B, & Muller-Hill B (1990) Mutant lac repressors with new specificities hint at rules for protein--DNA recognition. *EMBO J* 9(3):615-621.
26. Marsden I, Chen Y, Jin C, & Liao X (1997) Evidence That the DNA Binding Specificity of Winged Helix Proteins Is Mediated by a Structural Change in the Amino Acid Sequence Adjacent to the Principal DNA Binding Helix. *Biochemistry* 36(43):13248-13255.
27. Rodríguez-Ezpeleta N, *et al.* (2007) Toward Resolving the Eukaryotic Tree: The Phylogenetic Positions of Jakobids and Cercozoans. *Current Biology* 17(16):1420-1425.
28. Berger MF, *et al.* (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* 24(11):1429-1435.
29. Punta M, *et al.* (2011) The Pfam protein families database. *Nucleic Acids Res* 40(D1):D290-D301.
30. Katoh M & Katoh M (2004) Identification and characterization of human FOXN5 and rat Foxn5 genes in silico. *Int J Oncol* 24(5):1339-1344.
31. Meijsing SH, *et al.* (2009) DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science* 324(5925):407-410.
32. Scully KM, *et al.* (2000) Allosteric Effects of Pit-1 DNA Sites on Long-Term Repression in Cell Type Specification. *Science* 290(5494):1127-1131.
33. Eddy SR (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol* 7(10):e1002195.
34. Schäffer AA, *et al.* (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res* 29(14):2994-3005.
35. Katoh K, Kuma K, Toh H, & Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33(2):511-518.
36. Notredame C, Higgins DG, & Heringa J (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology* 302(1):205-217.
37. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792-1797.
38. Sievers F, *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:1-6.
39. Thompson JD, Higgins DG, & Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22(22):4673-4680.
40. Mirarab S & Warnow T (2011) FastSP: linear time calculation of alignment accuracy. *Bioinformatics* 27(23):3250-3258.
41. Le SQ & Gascuel O (2008) An improved general amino acid replacement matrix. *Mol Biol Evol* 25(7):1307-1320.
42. Darriba D, Taboada GL, Doallo R, & Posada D (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27(8):1164-1165.
43. Guindon S, *et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59(3):307-321.
44. Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*:783-791.
45. Anisimova M & Gascuel O (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol* 55(4):539-552.
46. Anisimova M, Gil M, Dufayard J-F, Dessimoz C, & Gascuel O (2011) Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. *Syst Biol* 60(5):685-699.
47. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688-2690.
48. Berger MF & Bulyk ML (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature protocols* 4(3):393-411.
49. Workman CT, *et al.* (2005) enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res* 33(Web Server issue):W389-392.