DNA-binding specificity changes in the evolution of forkhead transcription factors

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

| As Published | http://dx.doi.org/10.1073/pnas.1310430110 |
| Publisher | National Academy of Sciences (U.S.) |
| Version | Final published version |
| Accessed | Tue Dec 18 04:48:06 EST 2018 |
| Citable Link | http://hdl.handle.net/1721.1/88252 |
| Terms of Use | Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use. |
| Detailed Terms | |
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 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 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 previously undescribed source of modularity and flexibility in gene regulation and may play an important role in the evolution of transcriptional regulatory networks.
similar sequence, GACGC, as the binding site of the Fox3 factor Fhl1 (19); we therefore refer to the GACGC site as the FHL motif (Fig. 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 amino acid 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 subdomain swap experiments, a 20-aa 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 (Dataset S1). We chose these species based on their evolutionary importance and annotation level (27) (Fig. S1). For example, we included Spizellomyces punctatus and F punticola because 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 that a protein can recognize (28). We analyzed the binding specificities of 30 forkhead proteins, combining published data for 9 proteins with data for 21 proteins that we characterized for this study (Dataset S2). We focused on proteins from clades in which we had previously observed alternate binding specificities and clades of unknown specificity. By using two orthogonal means of evaluating the same proteins, we obtained a much richer picture of the evolutionary trajectory of changes in TF DNA-binding specificity than either analysis alone could provide.

**Results**

The published observation of approximately 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 (Materials and Methods), spanning 623 genes from 65 species (Dataset S1 and Fig. S1). We included two distantly related forkhead domains from the opisthokont sister group Amoebazoa as an outgroup. After removing partial domain sequences and those identical throughout the Fox domain, we used 529 Fox domain sequences (340 nonredundant; Dataset S1). We constructed a complete maximum likelihood (ML) tree of all nonredundant Fox domain sequences (Fig. S2). For each branch, the approximate likelihood-ratio test (aLRT) and 100 bootstrap replicates were used to evaluate support for inferred relationships (Materials and Methods). For presentation purposes, we constructed a ML tree of 262 (133 nonredundant) Fox domains from selected informative species (Fig. 2 and Dataset S1).

Various portions of the phylogeny could be determined with high confidence. Our analysis recovered the previously identified subfamily relationships between Fox proteins and also identified a previously unobserved fungal group (Fox4) 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 cross-over 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%; Fig. 2). Additionally, there is a splice site between amino acid positions 46 and 47 in the Pfam fork head domain hidden Markov model (HMM) (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 (Fig. S3A) is significantly less likely than the observed tree (*P* < 10^-10^, likelihood ratio test), and likelihood maximization using this tree as a starting tree separates the FoxN and Fox3 clades (Fig. S3 B and 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 (Fig. 2). This result suggests that FHL binding capacity evolved twice independently within the family and led us to examine these two subclades in more detail.

A phylogenetic tree constructed from only fungal Fox3 domains (Fig. S4) 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 (Fig. 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 (Figs. S4 and 14), 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,
by PBM enrichment (E) score, which indicates the preference of a protein for every possible 8-mer (28)] bound by the Dro sophila melanogaster FoxN2/3 protein checkpoint suppressor homologue (CHES-1-like) are ATAAAACAA and GTAAACAA, perfectly matching the FkhP consensus, and the next two are the FHL matches GACGCTAA and GACGCTAT. FoxR1 also shows bispecificity, despite presumably arising from an FHL-specific N1/4 ancestor. To our knowledge, 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 preferred A immediately 3’ to the core GACG, which we refer to as the FHL-3 motif, whereas FHL motifs from FoxN/R proteins all strongly disfavored A in that position, a variant we refer to as the FHL-N motif (Fig. 1). Similarly, Homo sapiens FoxR1 (which appears to have regained FkhP/S binding from an FHL-only ancestor) strongly preferred a C at position 2 of the FkhP motif, whereas other FkhP-binding FoxN domains strongly preferred T at that position (Fig. 4 and Fig. 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

**Fig. 2.** ML phylogenetic tree of forkhead domains. This compact tree was constructed for presentation purposes from a representative subset of phylogenetically informative species: metazoan oocytes, fly, and sponge; choanoflagellates *Salpingoeca rosetta* and *Monosiga brevicollis*; *Capsaspora owczarzaki* and *Sphaeroforma arctica* from *Ichthyosporea*; *S cerevisiae* from *Dikarya*; *Allomyces macrogynus* from *Blastocladiomycota*; *S punctatus* from *Chytridiomycota*; *Mortierella verticillata* from *Mortierellomycota*. F. alba from *Nuculinaa*; and *Acanthamoeba castellanii* from *Amoebobozoa*. Nodes supported with strong likelihood ratios are indicated with red circles (aRT ≥ 99%) or blue circles (aRT ≥ 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 (Fig. S2).

Fig. 2. ML phylogenetic tree of forkhead domains. This compact tree was constructed for presentation purposes from a representative subset of phylogenetically informative species: metazoan oocytes, fly, and sponge; choanoflagellates *Salpingoeca rosetta* and *Monosiga brevicollis*; *Capsaspora owczarzaki* and *Sphaeroforma arctica* from *Ichthyosporea*; *S cerevisiae* from *Dikarya*; *Allomyces macrogynus* from *Blastocladiomycota*; *S punctatus* from *Chytridiomycota*; *Mortierella verticillata* from *Mortierellomycota*. F. alba from *Nuculinaa*; and *Acanthamoeba castellanii* from *Amoebobozoa*. Nodes supported with strong likelihood ratios are indicated with red circles (aRT ≥ 99%) or blue circles (aRT ≥ 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 (Fig. S2).

binding the FkhP/S motifs not better than non-forkhead proteins (percent signs in Fig. 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 Fig. 3A; see also Fig. 1), with no specific binding to either the FkhP/S or FHL motifs.

Similarly, the phylogeny of the holozoan FoxN subfamily is relatively stable (Fig. 3B). Our analysis supports the existence of a fundamental split into FoxN1/A 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 (Fig. 4). For example, the top two 8-mers [ranked

**Fig. 3.** Detailed analysis of Fox3 and FoxN subfamilies. ML phylogenetic trees for Fox domains from a broader range of species for fungal Fox3 (A) and holozoan FoxN/R (B) clades. Red and blue circles indicate node support as in Fig. 2. Bold symbols represent binding capacity for different motif classes as defined in Fig. 1.
II groups, such as Fox4 and FoxM, and assemble them with published PBM data (Fig. S4 and Datasets S3 and S4). 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, as well as 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; Fig. S3D), in that each subfamily is more closely related to proteins that bind only FkhP and FkhS than to each other, suggesting that this finding 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 (Fig. 1). As in the case of FoxN and Fox3, ML inference with a starting tree containing differences in DNA sequence preference (Fig. 1 and Fig. S4). These binding differences are associated with widespread binding specificity.

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 vs. 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 Fox1 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 nonpithosthon Acanthamoeba castellanii did not fall into any of these three clusters, because it binds another distinct motif (Fig. 1 and Fig. S4).
differences in the recognition helix (Fig. S4). Indeed, altered recognition positions (Fig. 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), whereas 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 that we propose for Fox3. Moreover, the sort 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 metazoan 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 amino acid 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 that bind strongly to the FkhP site can achieve an alternate conformation that supports recognition of the FHL motif. It is intriguing, in the context of this observation, that both *M. musculus* FoxJ1 and *A. macrognous* 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 (Fig. 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 previously undescribed phenomenon of DNA-binding bispecificity suggests an alternate 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 Dataset S1. For each annotated protein sequence, we performed a HMM search using HMMER3 (33) with the Forkhead domain (PF00250) in the Pfam database (E value < 10<sup>−10</sup>) (29). Using the hit sequences as queries, we conducted iterative homology search using PSI-BLAST (E value < 10<sup>−10</sup>) (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 Dataset S1. All sequences used for the phylogenetic analysis contain five α-helices and three β-sheets as in human FoxP2 (11).

For phylogenetic analyses, each amino acid sequence of Fox domains was aligned by using five multiple sequence alignment programs: L-INS-i program in MAFFT (35), T-Coffee (36), MUSCLE (37), Clustal Omega (38), and 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 amino acid sites.

**Phylogenetic Inference.** The amino acid replacement models of LG (41) with gamma-distributed rate variation (α = 0.0881) were selected for whole forkhead domains, using the Akaike information criterion implemented in PROTEST 3 (42). Phylogenetic trees were constructed by using the ML method in PhyML 3.0 (43) with robustness evaluated by bootstrapping (100 times) (44) and by aLRT (45, 46). The starting tree for branch swapping was obtained by using a ML tree constructed by RAxML (47). For likelihood ratio tests, two ML trees were constructed from the ML tree in Fig. 2, changing the branching pattern of Fox3 and FoxM (Fig. S3A and B, respectively). RAxML was applied to optimize the lengths of branches and calculate ML scores (−13,422.7 for Fig. S3A and −13,414.9 for Fig. S3B). Comparing the ML score obtained from the tree in Fig. 2 (−13,406.2), P values were calculated based on the χ² 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). Constructs were transferred to the pDEST15 vector, which provides an N-terminal GST tag, using the Gateway re combinational cloning system (Invitrogen). All cloned forkhead domain sequences are provided in Dataset S2. Proteins were expressed by in vitro transcription and translation using the PURExpress in vitro Protein Synthesis kit (New England Biolabs). Concentrations of the expressed GST-fusion proteins were determined by Western blots in comparison with a dilution series of recombinant GST (Sigma).

PBM Experiments and Analysis. Double-stranding of oligonucleotide arrays and PBM experiments were performed essentially as described, except where noted in Dataset S2, using custom-designed “all 10-mer” arrays in the 4 x 44K (Agilent Technologies; AMADID #015681) or 8 x 60K (Agilent Technologies; AMADID #030236) array format (28, 48). Microarray data quantification, normalization, and motif derivation were performed as described (28, 48); some published PBM data (21) were reanalyzed for this study. DNA binding-site motif sequence logos were generated by using logoLOGOS (49). The 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 by using the heatmap.2 function in the gplots R package with the Manhattan distance metric.

ACKNOWLEDGMENTS. We thank Matthew W. Brown and Ifaki Ruiz-Trillo for sharing prepublication 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 Japan Society for the Promotion of Science Research Fellowship for Young Scientists (to S.N.), and by National Institutes of Health Grant R01 HG003985 (to M.L.B.). J.M.R. was supported in part by National Institutes of Health Molecular Biophysics Training Grant T32 GM008313.