Supporting Information

Sebé-Pedrós et al. 10.1073/pnas.1309748110

SI Methods

Phylogenetic analyses. Alignments were constructed using the online server of the multiple sequence alignment program MAFTT v.6 (1) and then edited in Geneious. Only those species and positions that were unambiguously aligned were included in the final analyses. The best-fit model for our set of proteins was chosen using ProtTest server (2). Maximum likelihood (ML) phylogenetic trees were estimated by RaxML (3) using the PROTGAMMALG + Γ + I model. Statistical support was estimated by performing 100-bootstrap replicates using RaxML. Bayesian analyses were performed with MrBayes3.2 (4), using the Le & Gascuel (LG)+Γ+I model of evolution, with four chains, a subsampling frequency of 100, and two parallel runs. Runs were stopped when the average SD of split frequencies of the two parallel runs was <0.01, at around 18,000,000 generations. Bayesian posterior probabilities (BPP) were used to assess the confidence values of each bipartition.

Protein Binding Microarray. Here, we used two different universal PBM array designs, designated ME and HK, after the initials of their designers (5, 6). The T-box DNA-binding domain of all analyzed T-box genes (Dataset S1), along with 50 amino acid "pads" flanking either side, were cloned as SacI–BamHI fragment into the vector pTH5325, a modified T7-driven GST expression vector. We used 150 ng of plasmid DNA in a 15-μL in vitro transcription/translation reaction using a PURExpress In Vitro Protein Synthesis Kit (New England BioLabs) supplemented with RNase inhibitor (Invitrogen) and 50 μM zinc acetate. After a 2-h incubation at 37 °C, 12.5 mL of the mix was added to 137.5 mL of protein-binding solution for a final mix of PBS/2% skim milk/0.2 mg per mL BSA/50 μM zinc acetate/0.1% Tween-20. This mixture was added to an array previously blocked with PBS/2% skim milk and washed once with PBS/0.1% Tween-20 and once with PBS/0.01% Triton-X 100. After a 1-h incubation at room temperature, the array was washed once with PBS/0.5% Tween-20/50 mM zinc acetate and once with PBS/0.01% Triton-X 100/50 mM zinc acetate. Cy5-labeled anti-GST antibody was added, diluted in PBS/2% skim milk/50 mM zinc acetate. After a 1-h incubation at room temperature, the array was washed three times with PBS/0.05% Tween-20/50 mM zinc acetate and once with PBS/50 mM zinc acetate. The array was then imaged using an Agilent microarray scanner at 2-mM resolution. Image spot intensities were quantified using ImaGene software (BioDiscovery). A position frequency matrix (PFM) motif was created from the PBM data by aligning all 8mers with E-scores > 0.45 (32, 33) using ClustalW (7), trimming the alignment by restricting to positions present in at least half of the sequences in the alignment, and converting each remaining position to frequencies.

1. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30(14):3059–3066.
2. Abascal F, Zardoya R, Posada D (2005) ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics* 21(9):2104–2105.
3. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688–2690.
4. Huelsenbeck JPP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17(8):754–755.
5. Mintseris J, Eisen MB (2006) Design of a combinatorial DNA microarray for protein-DNA interaction studies. *BMC Bioinformatics* 7:429.
6. Philippakis AA, Qureshi AM, Berger MF, Bulyk ML (2008) Design of compact, universal DNA microarrays for protein binding microarray experiments. *J Comput Biol* 15(7):655–665.
7. Chenna R, et al. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* 31(13):3497–3500.

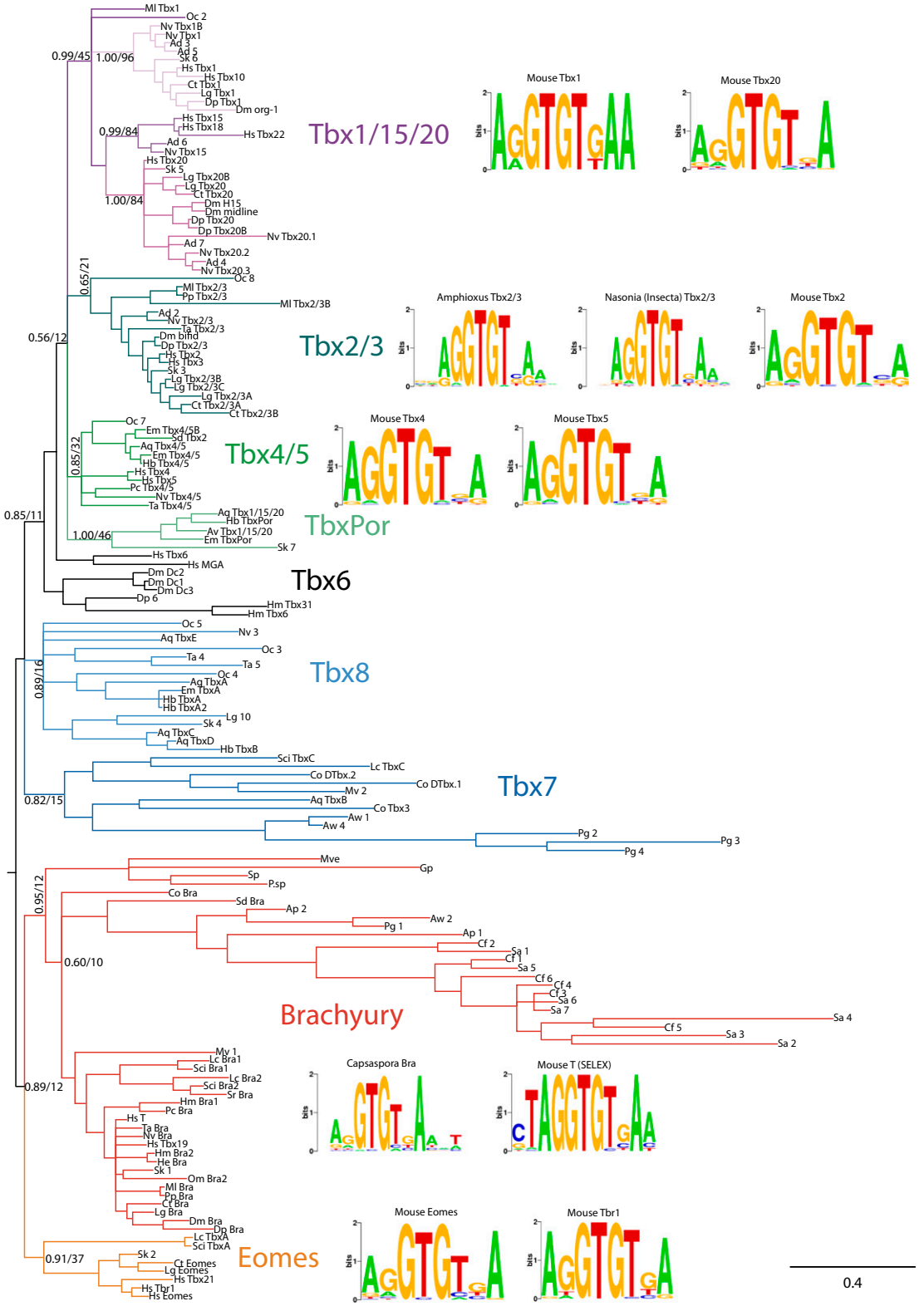

Fig. S1. Bayesian inference tree of T-box domains showing the different T-box families. The tree is rooted using the midpoint-rooted tree option. Statistical support values indicate Bayesian posterior probabilities (BPP) and 1,000-ML bootstrap replicates (BV). Colors correspond to different T-box families (the same as in Fig. 1). Taxa include *Acropora digitifera* (Ad), *Amphimedon queenslandica* (Aq), *Axinella verrucosa* (Av), *Capitella teleta* (Ct), *Capsaspora owczarzaki* (Co), *Drosophila melanogaster* (Dm), *Daphnia pulex* (Dp), *Ephydatia muelleri* (Em), *Gonapodya prolifera* (Gp), *Halichondria bowerbanki* (Hb), *Hydractinia echinata* (He), *Hydra magnipapillata* (Hm), *Homo sapiens* (Hs), *Leucosolenia complicata* (Lc), *Lottia gigantea* (Lg), *Mnemiopsis leydi* (Ml), *Ministeria vibrans* (Mv), *Mortierella verticillata* (Mve), *Nematostella vectensis* (Nv), *Oscarella carmela* (Oc), *Oopsaca minuta* (Om), *Podocoryne carnea* (Pc), *Pleurobrachia pileus* (Pp), *Pyromices* sp. (P.sp), *Sycon ciliatum* (Sci), *Suberites domuncula* (Sd), *Saccoglossus kowalevskii* (Sk), *Spizellomyces punctatus* (Sp), *Sycon raphanus* (Sr), and *Trichoplax adhaerens* (Ta). Co DTbx.1 and Co DTbx.2 are the two T-box domains of the same T-box *C. owczarzaki* gene (for further details, see main text). Protein-binding microarray (PBM) based (except mouse T, based on a SELEX (Systematic Evolution of Ligands by Exponential Enrichment) experiment) DNA-binding motifs for several members of different classes are shown (*Methods*). See Dataset S2 for newly annotated sequences.

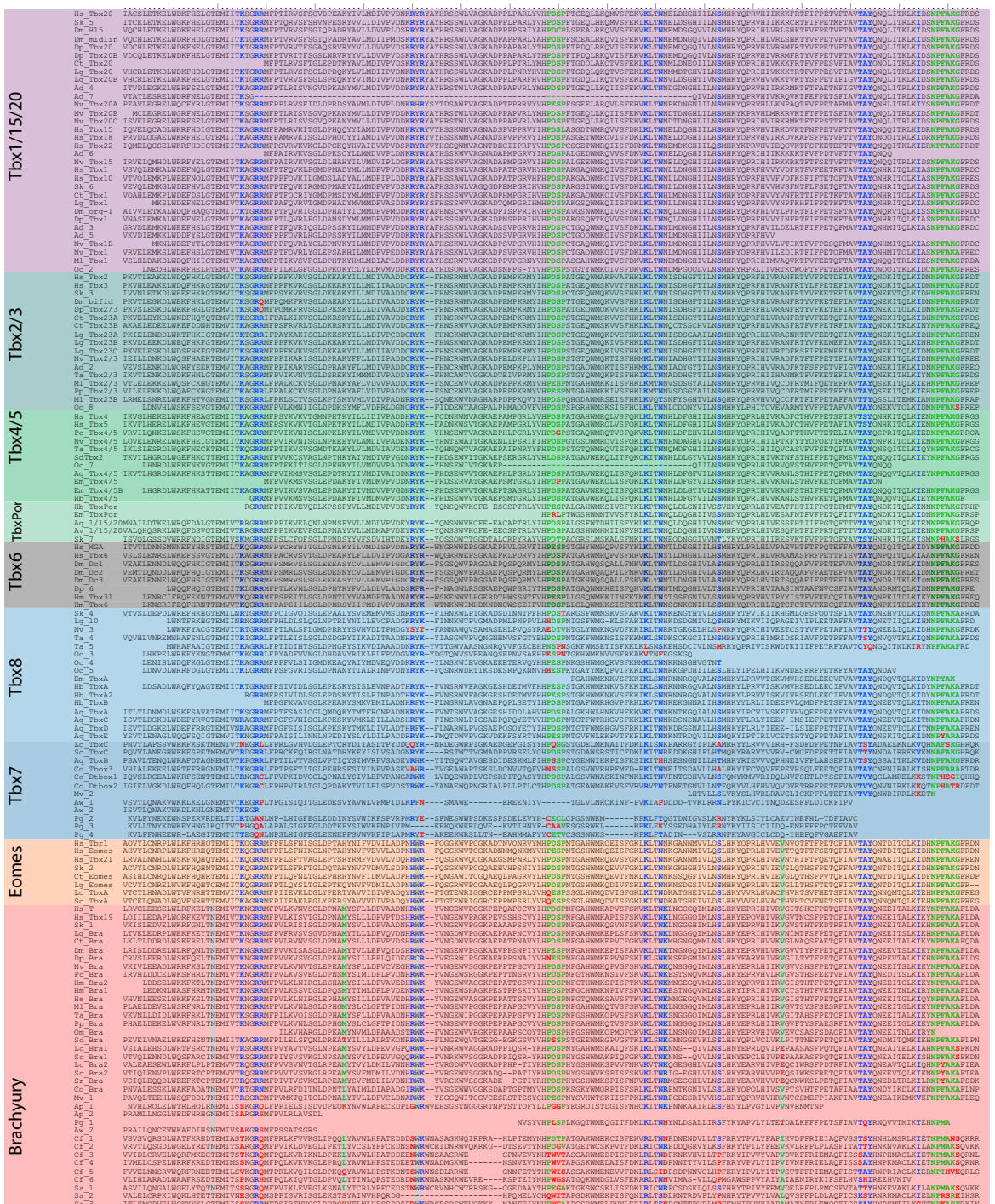

Fig. S2. Alignment of the T-box domain with the different families shown in distinct colors (same as in Fig. 1 and Fig. S1). Key DNA-binding amino acids are highlighted in blue, and dimerization amino acids are highlighted in green. Nonconservative amino acidic changes are depicted in red. Taxa included are the same as in Fig. S1. An arrowhead indicates Lysine149 after ref. 28.

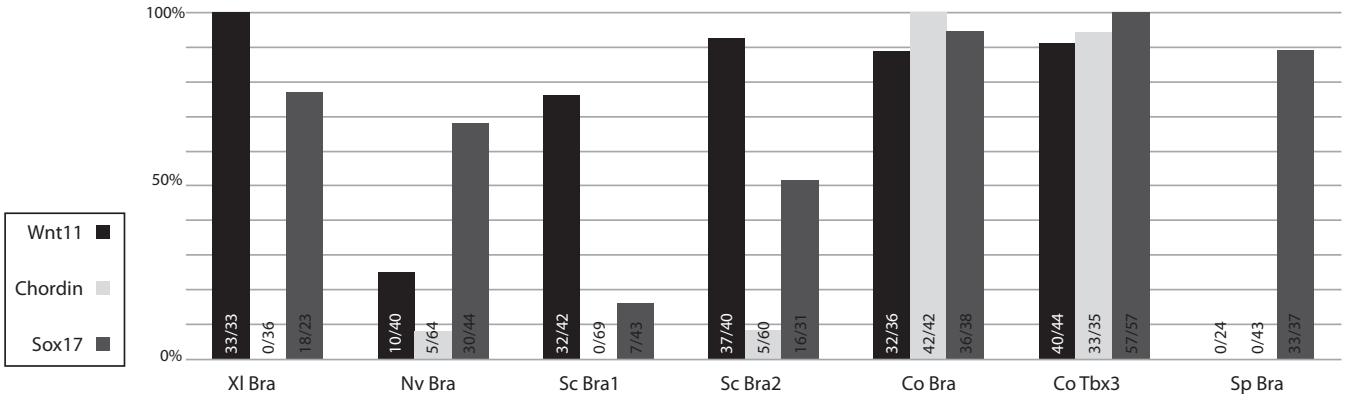

Fig. S3. Molecular phenotype counts of injected *Xenopus* whole embryos based on in situ hybridization for the markers *Wnt11*, *Chrordin*, and *Sox17* (Fig. 3). The bar plot indicates the percentage of injected embryos showing activation for each of the three markers. Total numbers (number of embryos with activation/total injected embryos) are shown at the base.

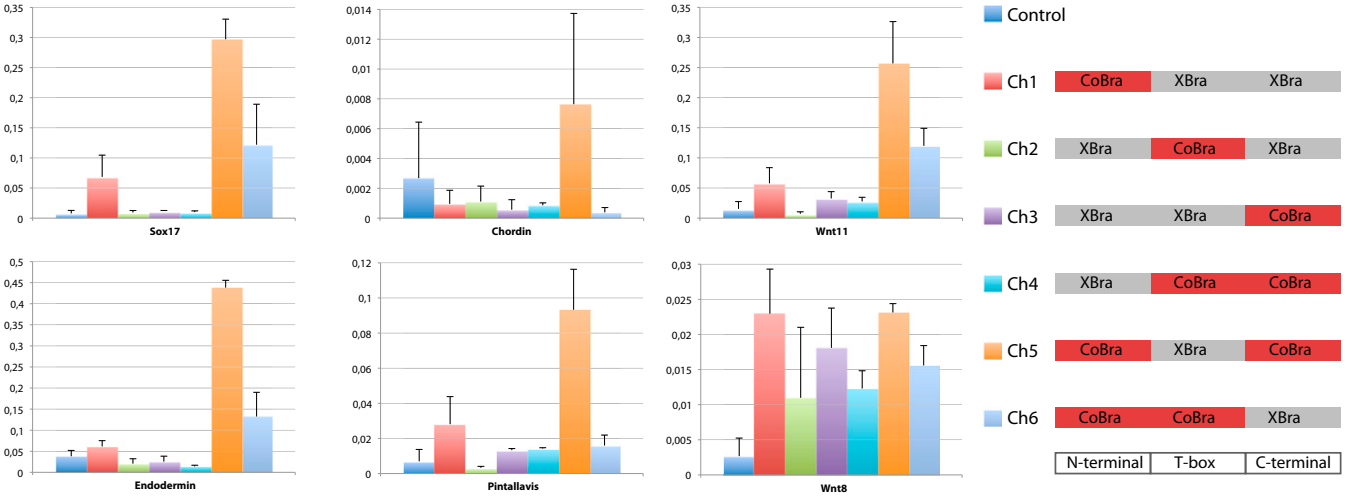

Fig. S4. Molecular phenotype of *Xenopus* animal caps injected with mRNAs of different *Capsaspora* Brachyury-*Xenopus* Brachyury (*CoBra-XBra)* gene chimeras. Quantitative RT-PCR experiments show several mesendoderm markers in animal caps injected with the mRNAs of the different chimeras. Controls are noninjected embryos. Bar plots show relative expression, normalized with endogenous *Histone 4*. Error bars represent SD from at least two different biological replicates.

Table S1. Primers used

Gene	Forward	Reverse
Sox17α	5′ -GCAAGATGCTTGGCAAGTCG	5′ -GCTGAAGTTCTCTAGACACA
pintallavis	5′ -GCACCCAACAAGATGATGACAC	5′ -TCCAATAGGAGCCTTTACCTGG
H4	5′ -GGGATAACATTCAGGGTATC	5′ -CATGGCGGTAACTGTCTTC
chordin	5′ -AACTGCCAGGACTGGATGGT	5′ -GGCAGGATTTAGAGTTGCTTC
Wnt11	5′ -TGGGCTCTGCTTTTGTTGACG	5′ -GGATGACTTTAGTGGCACCAAGG
endodermin	5′ -TATTCTGACTCCTGAAGGTG	5′ -GAGAACTGCCCATGTGCCTC
wnt8	5′ -CTGATGCCTTCAGTTCTGTGG	5′ -CTACCTGTTTGCATTGCTCGC
MyoD	5′ -AGGTCCAACTGCTCCGACGGCATGAA	5′ -AGGAGAGAATCCAGTTGATGGAAACA
muscle actin	5′ -GCTGACAGAATGCAGAAG	5′ -TTGCTTGGAGGAGTGTGT
shh	5′ -AACACACCTGGGCACACCTC	5′ -TCCAAAAGCCAAGTCCCTAT

Dataset S1. Newly annotated genes in Fig. S1, including source, sequence, and classification

Dataset S1

Dataset S2. Genes included in Fig. 4 and Fig. S1, including source and motif type

Dataset S2