Supplementary Information for "The emergence of Sox and POU transcription factors predates the origins of animal stem cells"

Ya Gao[1,2#], Daisylyn Senna Tan[1,2#], Mathias Girbig[3#], Haoqing Hu[1,2], Xiaomin Zhou[4], Qianwen Xie[4,5], Shi Wing Yeung[1,2], Kin Shing Lee[6], Sik Yin Ho[1], Vlad Cojocaru[7,8,9], Jian Yan[4,5], Georg K.A. Hochberg[3,10], Alex de Mendoza[11,*], Ralf Jauch[1,2*]

[1]School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China.

[2]Center for Translational Stem Cell Biology, Hong Kong SAR, China.

[3]Evolutionary Biochemistry Group, Max Planck Institute for Terrestrial Microbiology, Marburg, Germany.

[4]Department of Biomedical Sciences, City University of Hong Kong, Hong Kong SAR, China.

[5]School of Medicine, Northwest University, Xi'an, China.

[6]Transgenic Core Facility of the Centre for Comparative Medicine Research, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China.

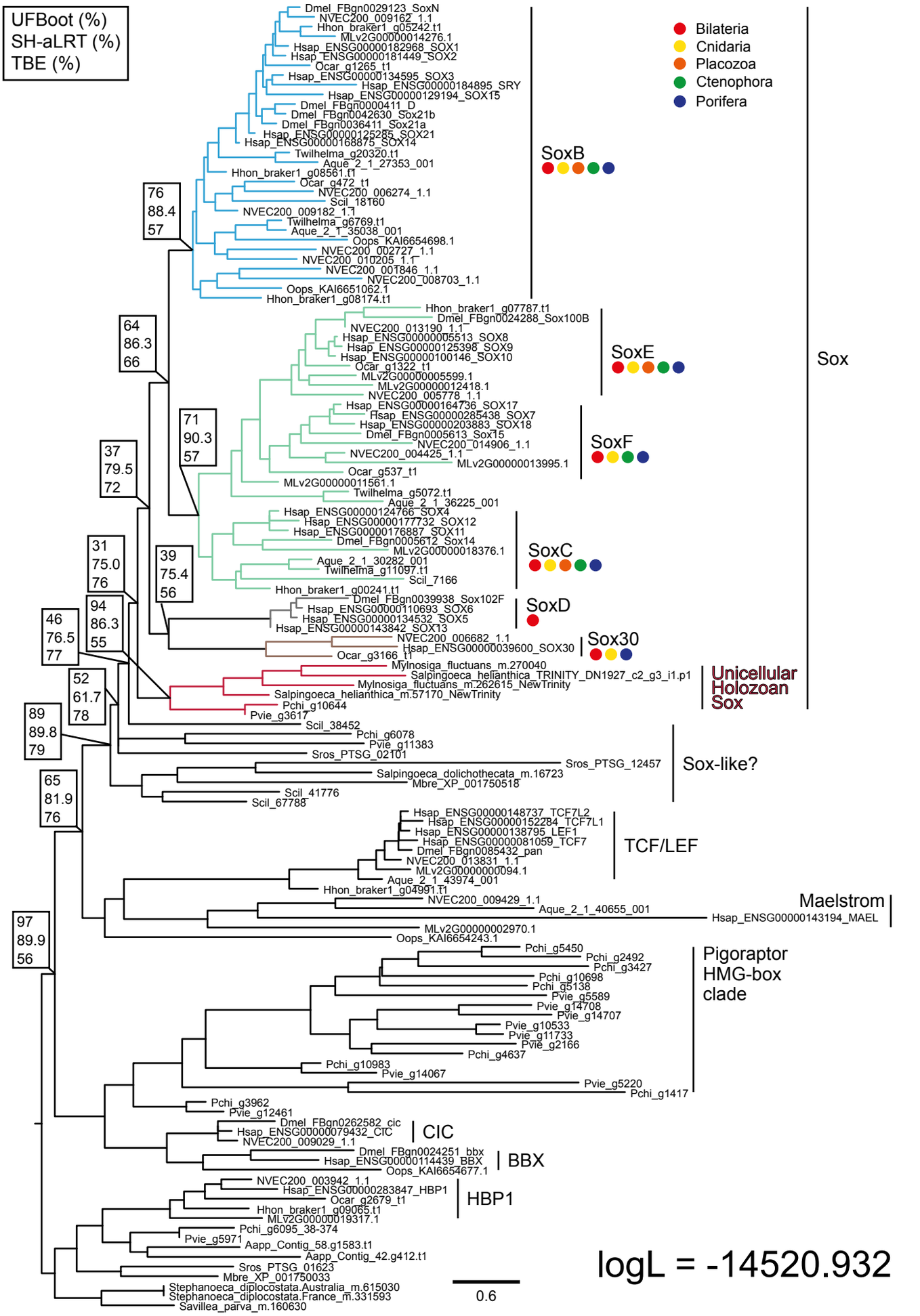
[7]STAR-UBB Institute, Babeş-Bolyai University, Cluj-Napoca, Romania.

[8]Computational Structural Biology Group, Utrecht University, The Netherlands.

[9]Max Planck Institute for Molecular Biomedicine, Münster, Germany.

[10]Department of Chemistry and Center for Synthetic Microbiology, Philipps University, Marburg, Germany.

[11]School of Biological and Behavioural Sciences, Queen Mary University of London, London, United Kingdom.



Bilateria
Hsap - Homo sapiens
Dmel - Drosophila melanogaster

Cnidaria
NVEC - Nematostella vectensis

Placozoa
Hhon - Hoilungia hongkongensis

Porifera
Oops - Oopsacas minuta
Aque - Amphimedon queenslandica
Twhilema - Tethya wilhelma
Ocar - Oscarella carmela
Scil - Sycon ciliatum

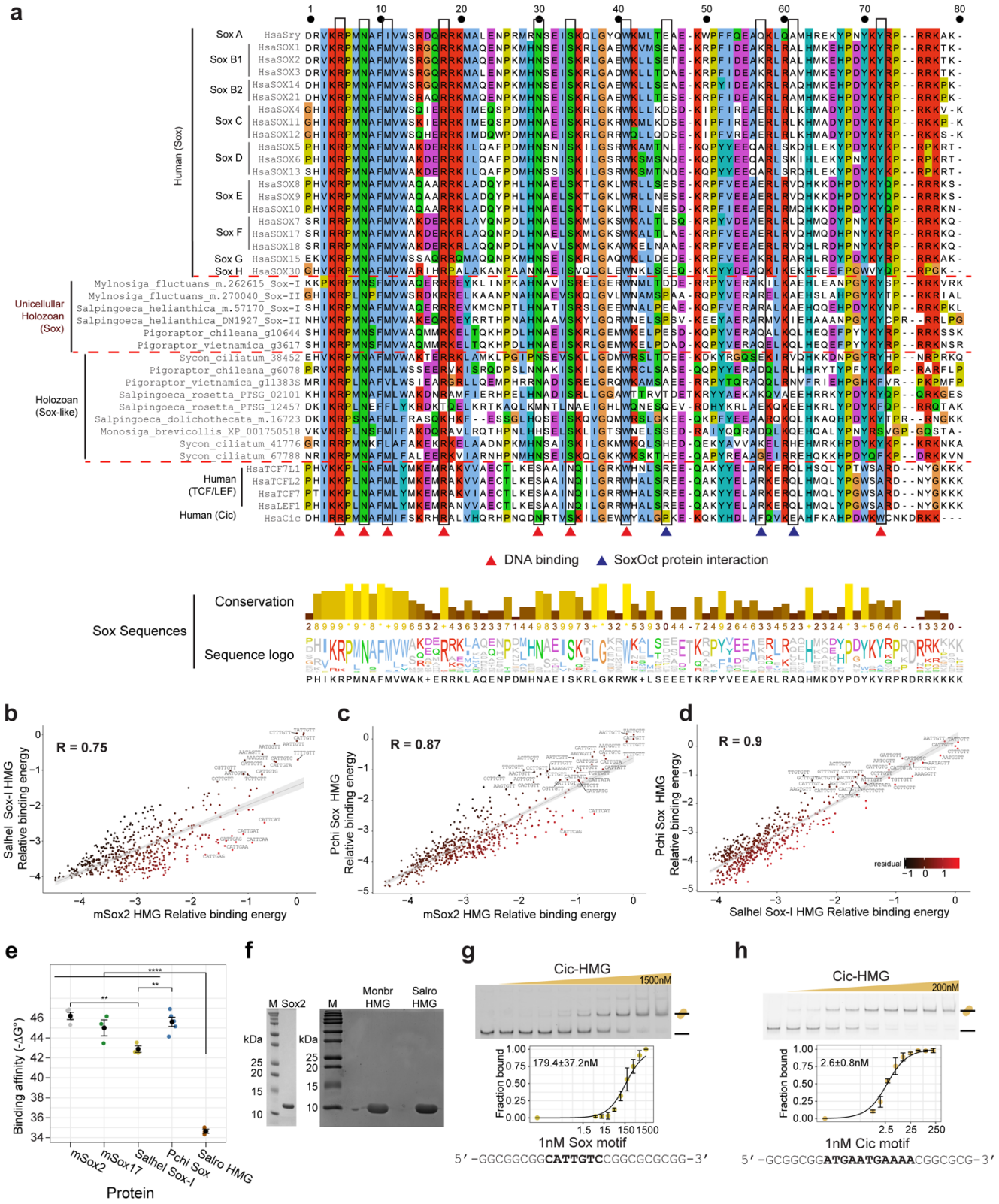
Ctenophora
ML - Mnemiopsis leidyi

Choanoflagellata
Sros - Salpingoeca rosetta
Mbre - Monosiga brevicollis
Mylnosiga fluctuans
Salpingoeca helianthica
Stephanoeca diplocostata

Filasterea
Pchi - Pigoraptor chileana
Pvie - Pigoraptor vietnamica

Ichthyosporea
Aapp - Amoebidium appalachense

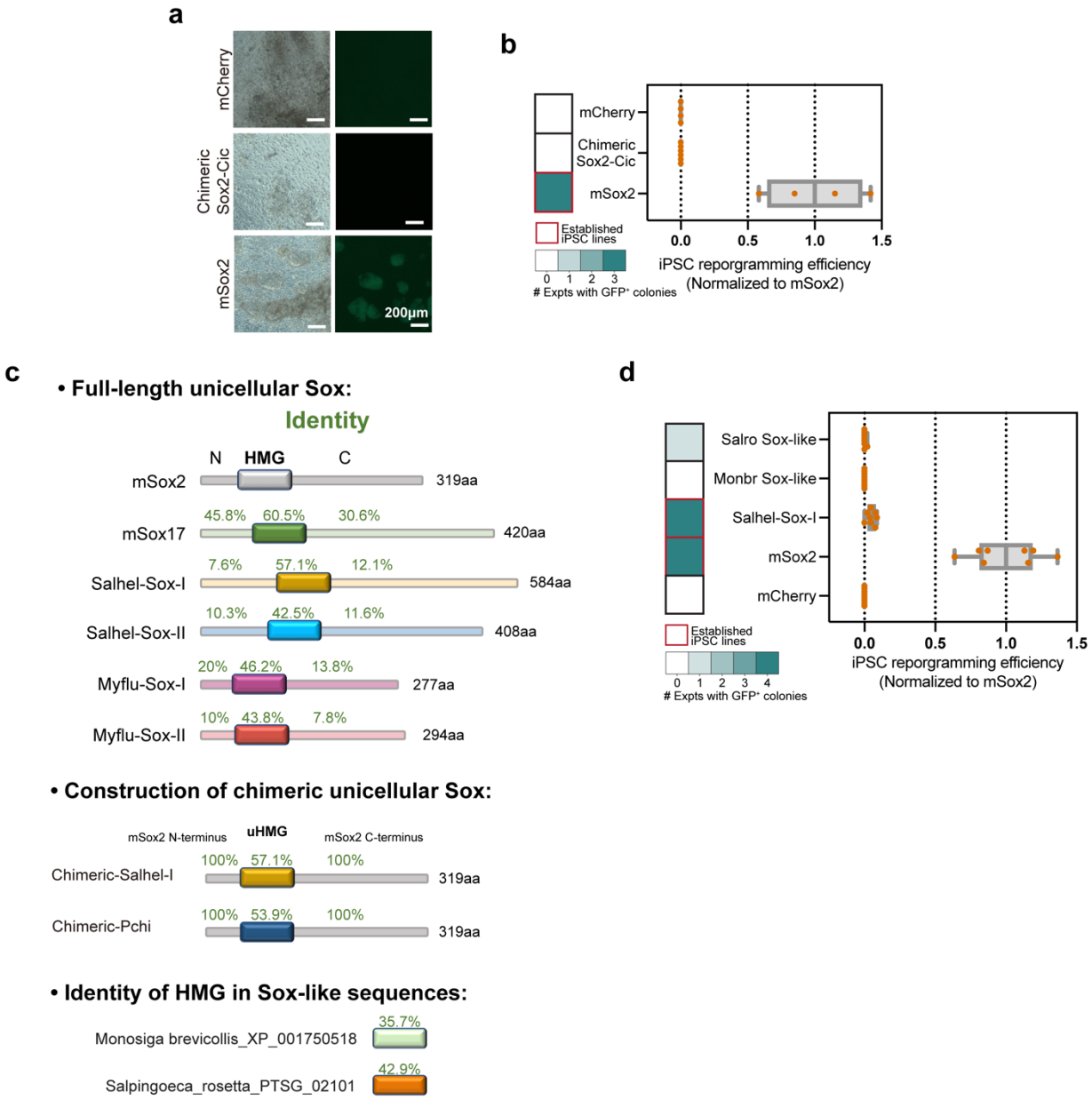
Supplementary Fig. 1 for Fig.1. Detailed phylogenetic tree of Sox transcription factors
Maximum likelihood phylogeny of the Sox HMG domain and homologous Sox-like or non-Sox HMG domains used as outgroups to root the phylogenetic tree. Nodes of interest are labeled with support values for ultrafast bootstrap approximation (UFBoot), SH-like approximate likelihood ratio (SH-aLRT) testing, and transfer bootstrapping (TBE). Colored spheres below Sox groups depict their coverage across the major metazoan clades/phyla (see legend on the top right). Scale bar: substitutions per site. HMG - high-mobility group; TCF/LEF - T cell factor/lymphoid enhancer factor family; CIC - Capicua; BBX - bobby sox homolog; HBP1 - HMG box-containing protein 1. Species identifiers for the sequences: Bilateria - Hsap (*Homo sapiens*), Dmel (*Drosophila melanogaster*), Cnidaria - NVEC (*Nematostella vectensis*), Placozoa - Hhon (*Hoilungia_hongkongensis*), Ctenophora - ML (*Mnemiopsis leidyi*), Porifera - Demosponges - Aque (*Amphimedon queenslandica*), Twhilhelma (*Tethya wilhelma*), Hexactinelid - Oops (*Oopsacas minuta*), Homoscleromorpha - Ocar (Oscarella carmela), Calcarea - Scil (*Sycon ciliatum*), Choanoflagellates – Mbre (*Monosiga brevicollis*), Sros (Salpingoeca rosetta), Filasterea – Pchi (*Pigoraptor chileana*), Pvie (*Pigoraptor vietnamica*), Ichthyosporea – Aapp (*Amoebidium appalachense*).



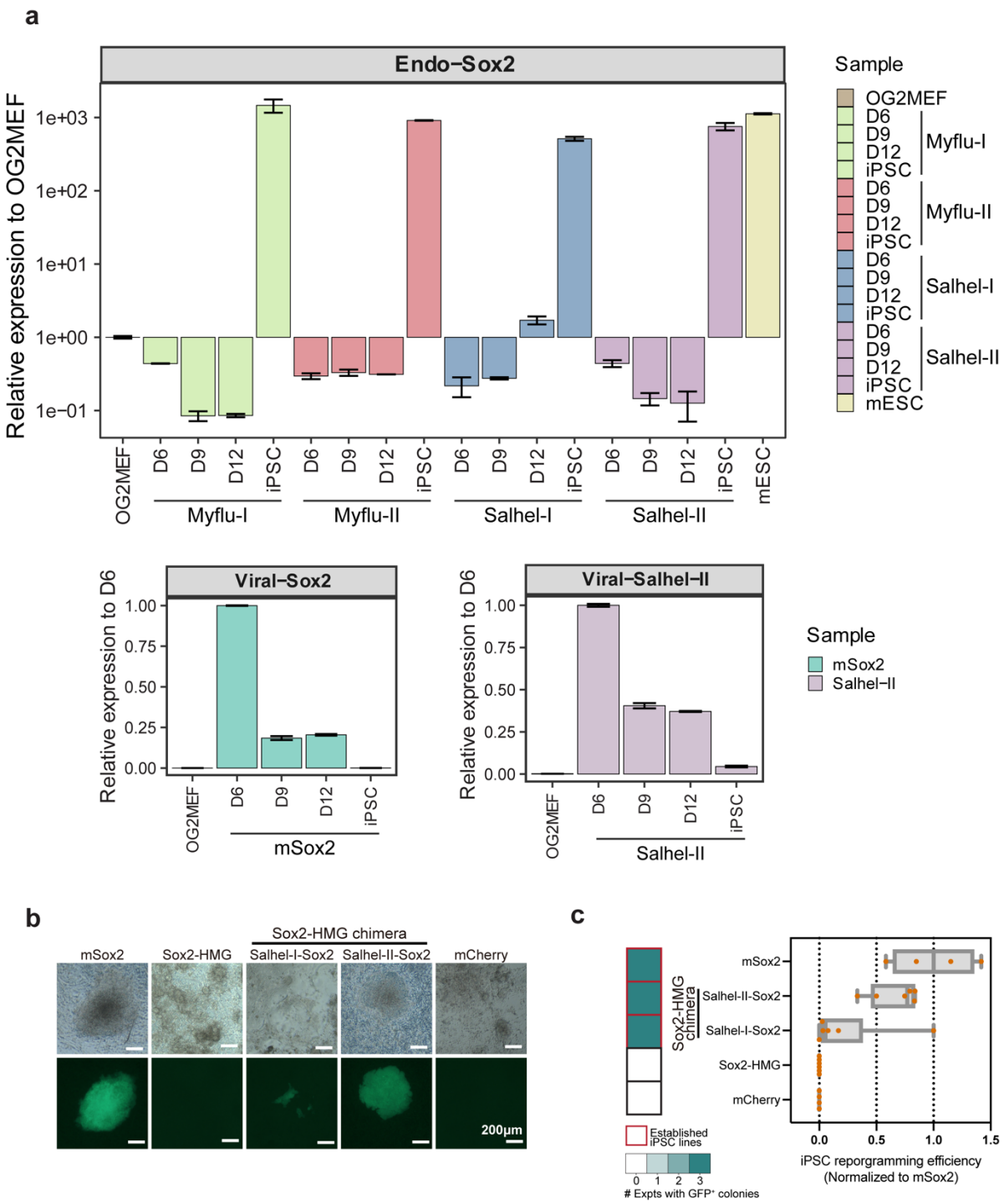
Supplementary Fig.2 for Fig.1. DNA binding of Sox and related HMG box factors.

a Multiple sequence alignment of all human Sox family members, TCF7/LEF1, Cic and Sox/Sox-like sequences found in Choanoflagellates and Filastereans. Triangles mark residues mediating sequence-specific DNA binding (red) or location at the Oct4 heterodimer interface (blue). Conservation analysis by Jalview with scores and sequence logo of Sox sequences (human and holozoan) are annotated below. The asterisk denotes highly conserved columns (score of 11), while a plus sign marks columns (score of 10) with mutations but conserved properties. In the sequence logo, consensus is the modal residue while a plus sign represents the equal top residues. **b-d** Correlation scatter plots of the relative binding affinities of **(b)** mSox2 versus Salhel Sox-I HMG, **(c)** mSox2 versus Pchi Sox HMG, and **(d)** Pchi Sox HMG versus Salhel Sox-I HMG, for all 508 sequences tested in Spec-seq. R = Pearson's correlation coefficient. The color indicates the residual score from the correlation line. **e** SDS-PAGE gel of representative HMG proteins purified mSox2 and Sox-like proteins from Monbr and Salro. **f** Calculated standard Gibbs free energy change, ΔG° from apparent Kd's in (**Fig.1 f-h**). Central black dot indicates

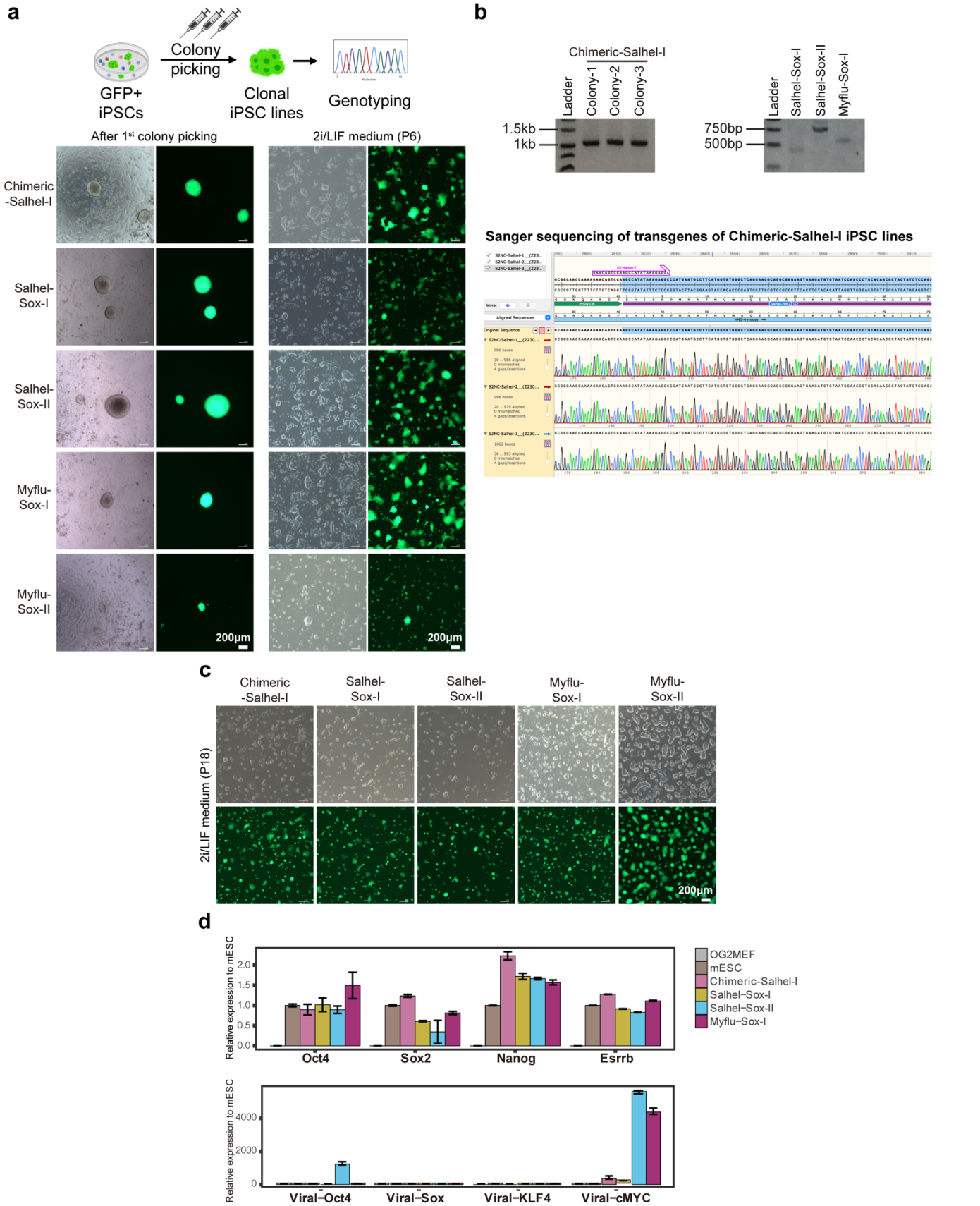
the mean and error bars show SEM. Adjusted p-values were calculated from unpaired t-tests adjusted for multiple comparisons with the Holm–Bonferroni method (**p < 0.01, **** p < 0.0001). **g-h** Binding of the Cic-HMG box DBD to **(g)** consensus Sox DNA and **(h)** Cic DNA element from [1] (n = 3 experiments, apparent Kd shown as mean ± SD).



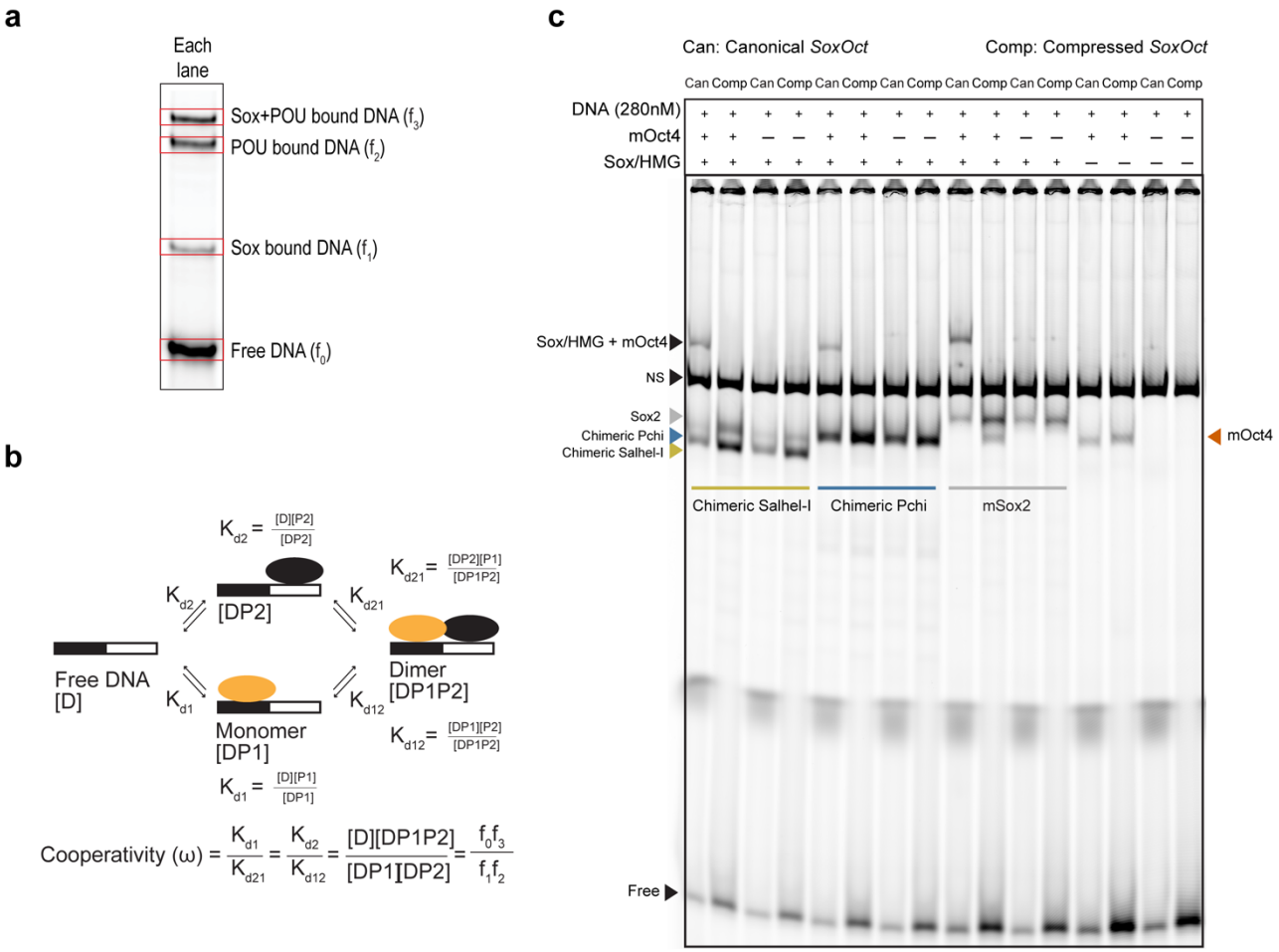
Supplementary Fig 3 for Fig.2. Pluripotency reprogramming of Sox and related HMG box factors.
a Representative microscopy images of iPSC reprogramming by chimeric Sox2-Cic on day 14. Scale bar, 200μm. The chimeric Sox2-Cic was constructed by flanking Cic with N and C-termini of mSox2. **b** Quantification of iPSC reprogramming efficiency of chimeric Sox2-Cic normalized to the colony numbers of mSox2 on day 14 (n=4, 2 biological replicates each with 2 technical replicates). **c** The schematic structure of unicellular Sox tested in this study. The percentage indicates the similarity to the N-terminal domain, HMG box and C-terminal domain of mSox2. mSox2, mouse Sox2; Salhel, *Salpingoeca helianthica*; Myflu, *Mylnosiga fluctuans*; mSox17, mouse Sox17; Pchi, *Pigoraptor chileana*. The middle panel illustrates the design of chimeric unicellular Sox sequences, in which the HMG domain of Salhel-Sox-I or Pchi is flanked by mSox2 N-terminal domain (mSox2-NTD) and C-Terminal domain (mSox2-CTD). Elsewhere in the article, these two chimeric sequences are abbreviated as Chimeric Salhel-I and Chimeric-Pchi, respectively. The sequence identity was calculated using the online tool available at https://www.ebi.ac.uk/Tools/msa/clustalo/. **d** Quantification of iPSC reprogramming efficiency of Sox-like factors normalized to the colony numbers of mSox2 on day 14 (n=6, 4 biological replicates each with 2 technical replicates). Salro Sox-like represents *Salpingoeca rosetta*_PTSG_02101 Sox-like, while Monbr Sox-like represents *Monosiga brevicollis*_XP_001750518 Sox-like.



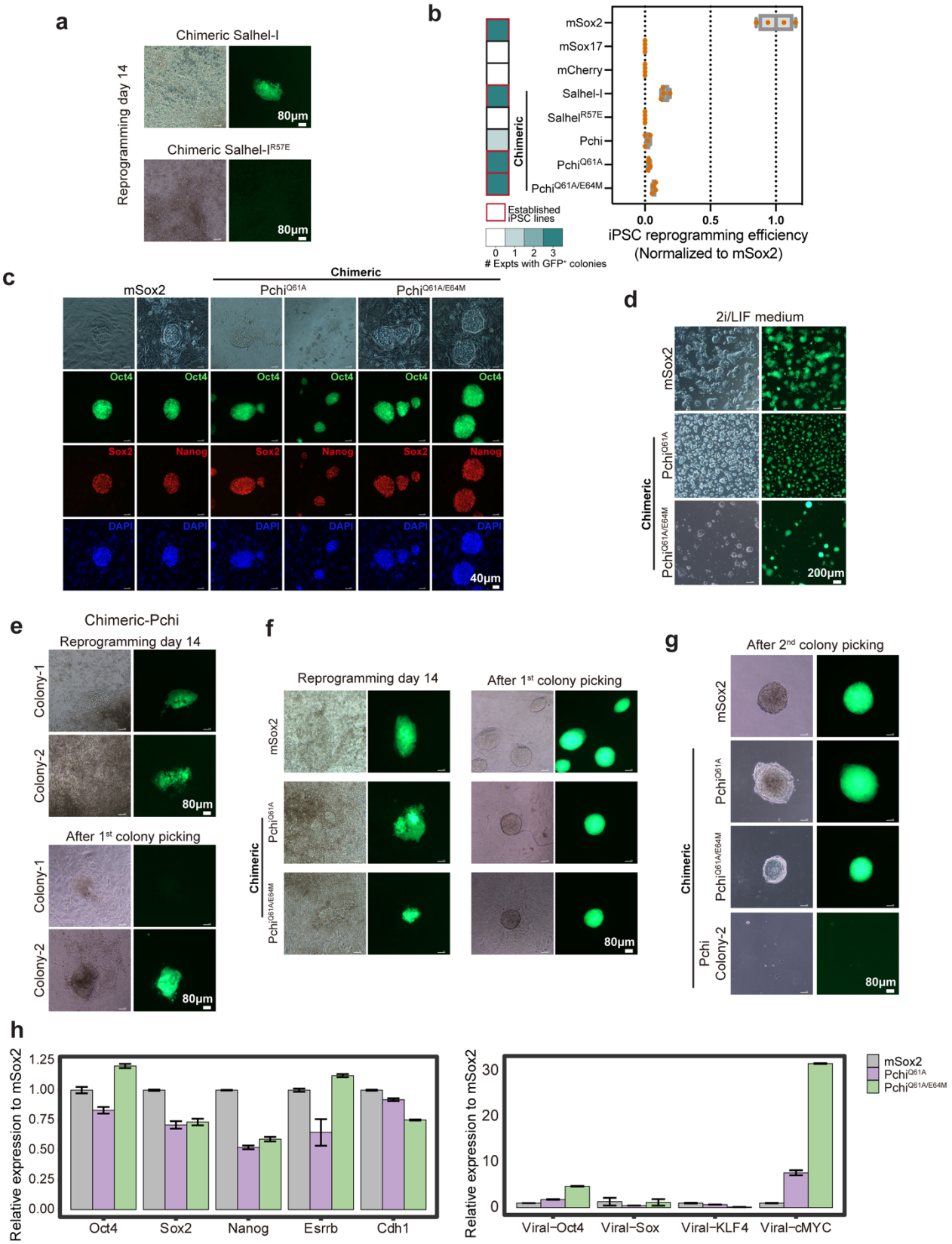
Supplementary Fig.4 for Fig.2. Pluripotency reprogramming with Choanoflagellate Sox.
a RT-qPCR analysis of the expression level of endogenous Sox2 (upper panel) and viral Sox2 or viral Salhel-Sox-II (lower panel) during days 6, 9 and 12 (D6-12) of iPSC reprogramming, in starting OG2MEFs and established clonal iPSC colonies. **b** Representative microscopy images of colonies observed during iPSC reprogramming using Sox2-HMG only or chimeric Sox2-HMG variants. Two chimeric Sox-HMG variants were created by flanking the Sox2-HMG with the N- and C-termini of Salhel-Sox-I or Salhel-Sox-II. Scale bar, 200μm. **c** Quantification of iPSC reprogramming efficiency of Sox2-HMG variants normalized to the colony numbers of mSox2 on day 14 (n=3, 2 biological replicates each with 1 or 2 technical replicates).



Supplementary Fig.5 for Fig.2. Validation of iPSCs generated with choanoflagellate Sox.
a The flow chart of clonal iPSC lines established by three rounds of single-cell colony picking. Representative images of uSox iPSCs after the first round of colony picking and cell morphology in 2i/LIF medium at passage 6. Scale bar, 200μm. **b** The validation of uSox iPSC lines by PCR genotyping and Sanger sequencing. **c** Representative images of clonal iPSC lines generated with indicated Sox factors at passage 18 with prolonged culture in 2i/LIF medium**.** Scale bar, 200μm. **d** Expression of pluripotency genes and viral transgenes in uSox iPSC lines examined by RT-qPCR (2 replicates), normalized to the expression of mouse embryonic stem cells (mESC). The illustration for a was created using BioRender.com with license for publication.

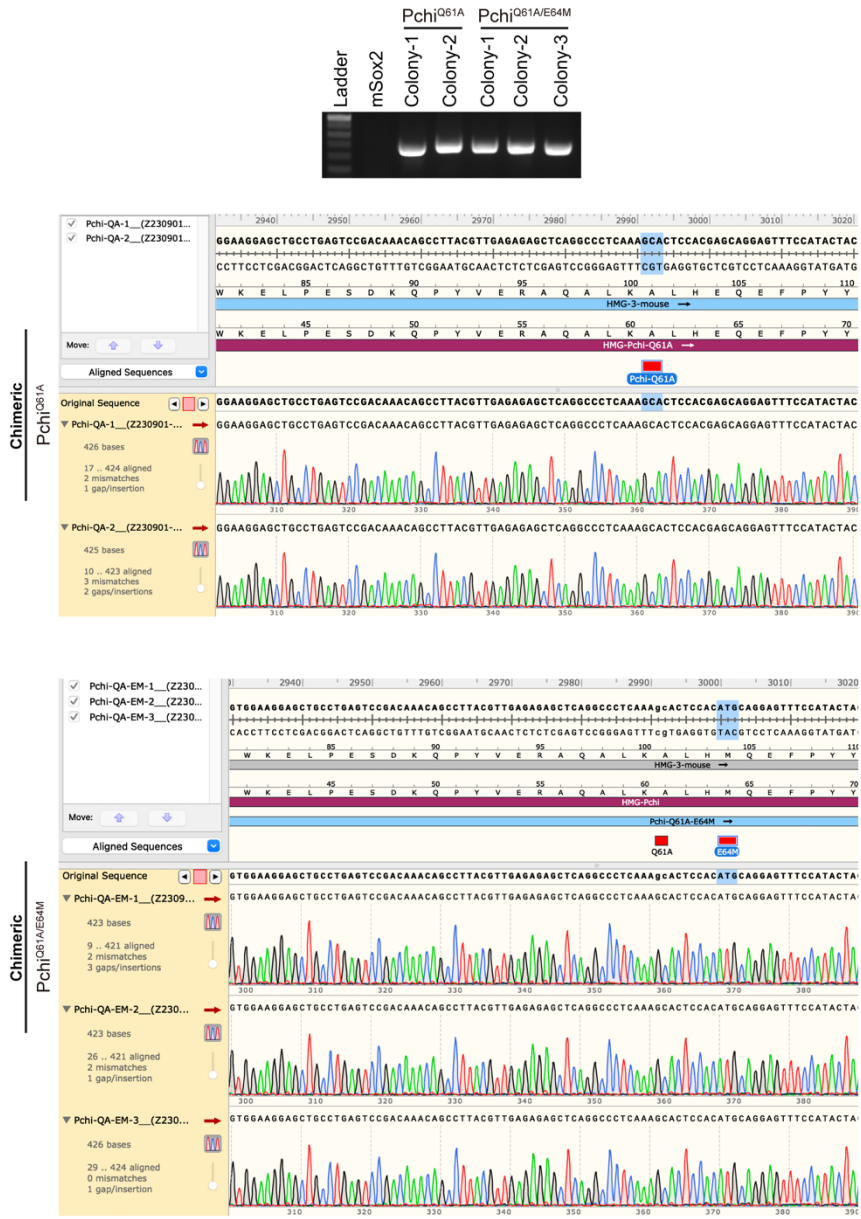


Supplementary Fig.6 for Fig.3. Dimerization of choanoflagellate Sox with mouse Oct4.
a An illustration of a lane of an EMSA gel with the four microstates of cy5-labelled DNA complexed with Sox and POU proteins. The fractional contribution of the microstates of the free DNA (f0), Sox-DNA (f1), POU-DNA (f2) and ternary complex (f3) can be quantified using densitometry. **b** Schematic diagram highlighting the approach to calculating the cooperativity of heterodimer formation on DNA as originally detailed in [2]. Monomeric binding of either Sox or POU on each of their respective half sites followed by heterodimer complex formation by binding to the remaining site. Kd_1/Kd_2 and Kd_{12}/Kd_{21} are the equilibrium dissociation constants for monomer or dimer formation. [D], [P], [DP1], [DP2] and [DP1P2] are the fractional concentrations of the free DNA, protein, monomer–DNA complexes and dimer-DNA complex, respectively. These concentrations can be estimated through the fractions (f0-f3 detailed in A). The cooperativity factor (ω) can then be calculated directly from these fractional contributions. **c** Representative EMSA titrations from 2 replicate experiments using full-length proteins from HEK293T whole cell extracts of Soxes (Sox2, Chimeric Salhel-I and Chimeric Pchi) dimerizing with mOct4 on the canonical or compressed SoxOct DNA.

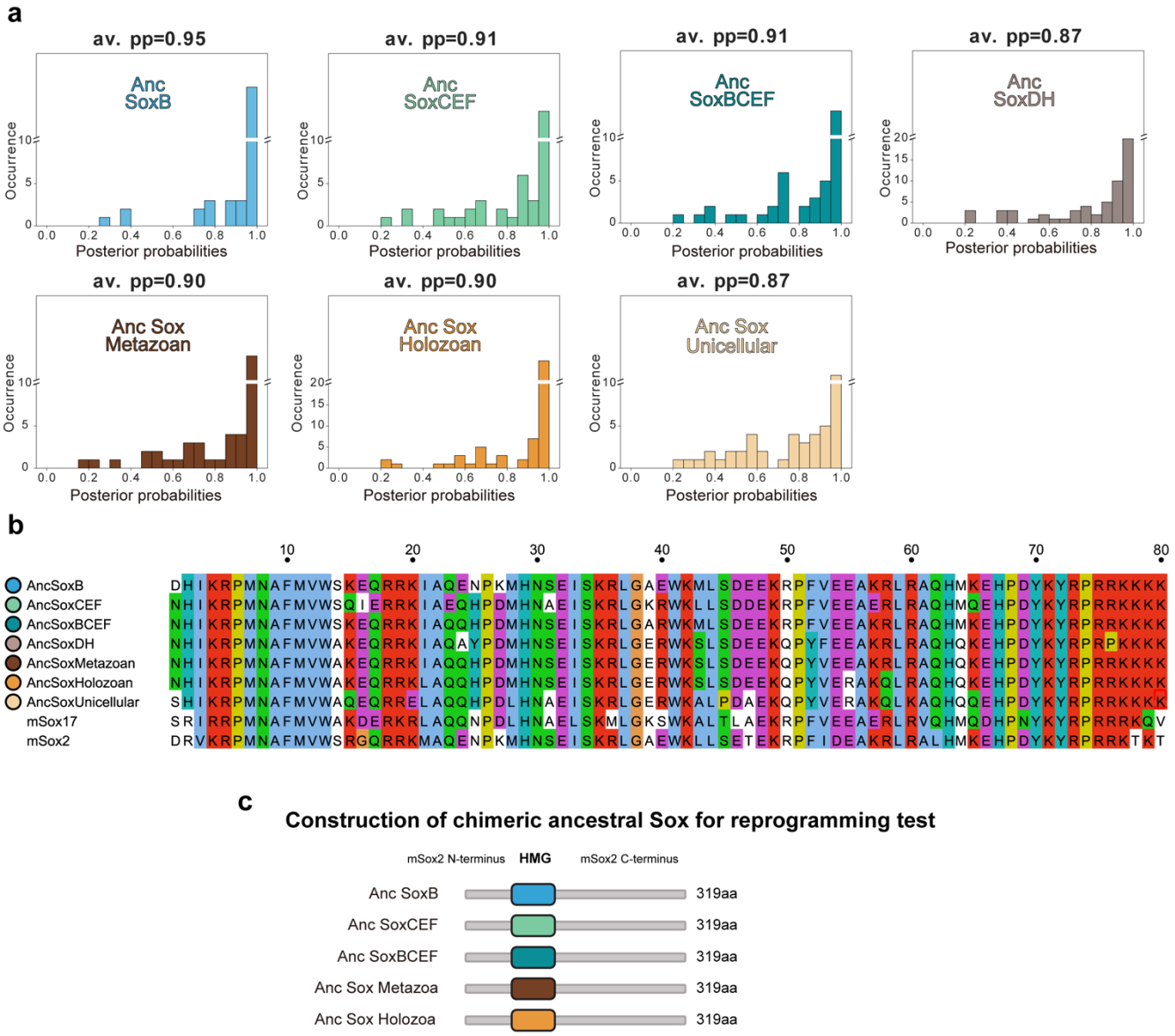


Supplementary Fig.7 for Fig.3. Re-engineering of filasterean Sox into a reprogramming factor.
a Representative microscope images of colonies generated by Chimeric Salhel-I and its mutants Chimeric Salhel-I^{57E}. Scale bar, 80μm. **b** Quantification of Oct4-GFP+ iPSC colonies counted on day 14 of reprogramming. The heatmap on the left panel illustrates the reprogramming efficiency of unicellular Sox (uSox), normalized by the number of iPSC colonies generated by mSox2. The right panel shows the statistic quantification of GFP+ iPSC colonies count on day 14 of reprogramming (n=4, mean ± SD, 2 biological replicates each with 2 technical replicates). **c** Characterization of iPSC lines derived by Chimeric Pchi mutants by immunocytochemistry staining. Scale bar, 40μm. **d** Representative images of Pchi mutants derived iPSCs cultured in 2i/LIF medium. Scale bar, 200μm. **e** Representative microscope images of iPSC colonies reprogrammed by Chimeric-Pchi on reprogramming day 14 and after the first round of colony picking. Scale bar, 80μm. **f** Representative microscope images show iPSC colonies generated by mSox2 and mutants of Chimeric-Pchi on reprogramming day 14 and

after the first round of colony picking. Scale bar, 80μm. **g** Representative images of uSox iPSCs after the second round of colony picking and cell morphology in 2i/LIF medium at passage 6. The iPSC Colony-2 generated by Chimeric Pchi differentiated after the second colony picking. Each set of images is accompanied by its respective scale bar. **h** Expression level of pluripotency markers in iPSC lines derived by Chimeric Pchi mutants by qRT-PCR (2 replicates).



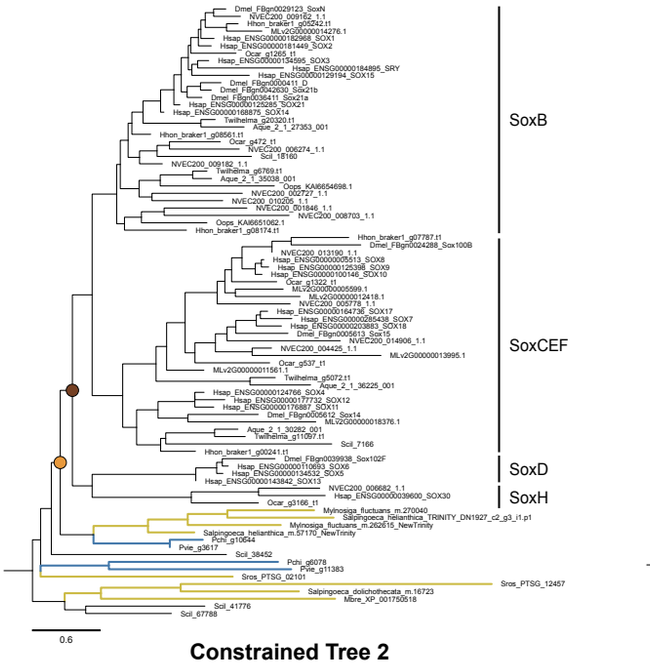
Supplementary Fig.8 for Fig.3. Validation that iPSCs were generated with re-engineered filasterean Sox. PCR genotyping of iPSC lines reprogrammed with rationally engineered Pchi mutants to verify the identity of the integrated transgenes.



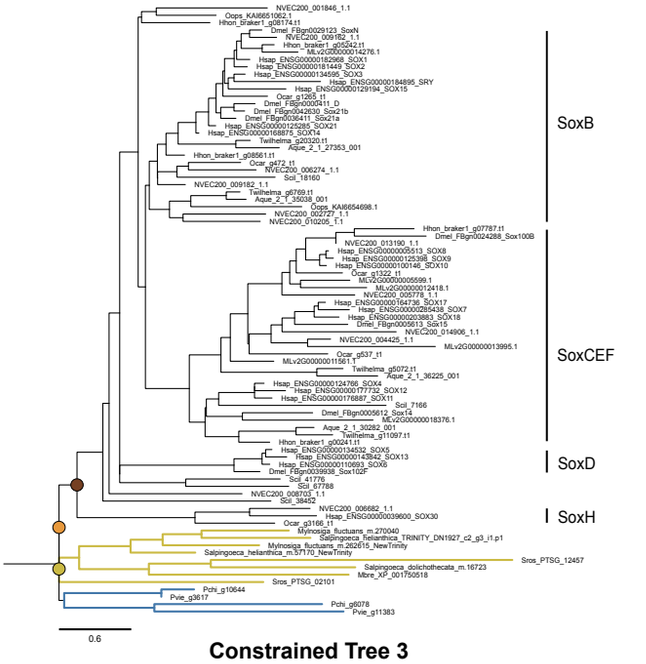
Supplementary Fig.9 for Fig.4. Ancestral sequence reconstruction using the maximum-likelihood phylogeny.

a Posterior probability distributions of ancestral Sox HMG domains reconstructed via ancestral sequence reconstruction (ASR) from the ML phylogeny. See Fig.4 and Supplementary Fig.1 for the corresponding phylogeny and highlighted reconstructed nodes. The posterior probabilities represent the likelihood of specific ancestral states at each site, providing a measure of confidence in the inferred ancestral states. The x-axis shows the range of posterior probabilities (from 0 to 1) and the y-axis represents the occurrence of sites with corresponding posterior probabilities. The average posterior probability (av. pp) value across all sites reflects the overall confidence in the reconstruction. **b** Multiple sequence alignments of the maximum a posteriori (MAP) sequences of selected ancestral Sox proteins, inferred via ASR from the ML tree, and of the mouse (m) Sox17 and Sox2. The MSA is colored according to the Clustal X color scheme. **c** Schematic of the chimeric constructs of ancestral (Anc) Sox HMG domain with the N- and C-terminus of mSox2 used in iPSC reprogramming.

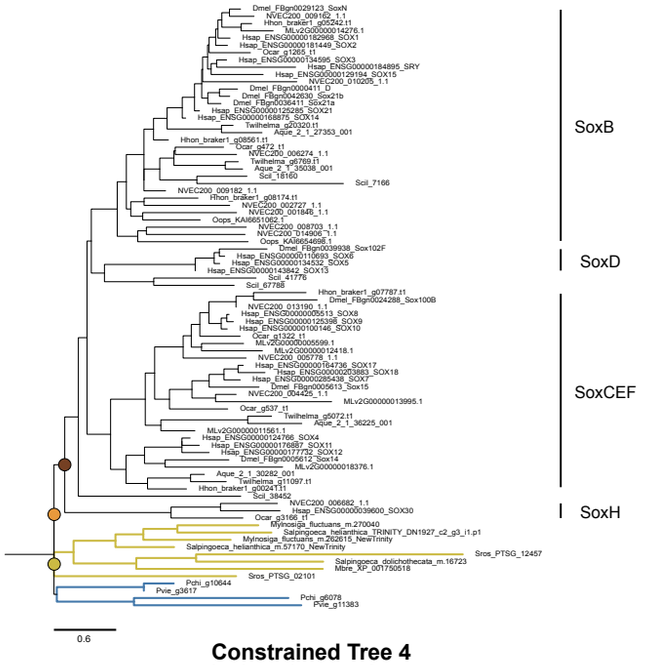
Maximum Likelihood (ML) Tree
logL = -14521.38
p-AU: 0.8



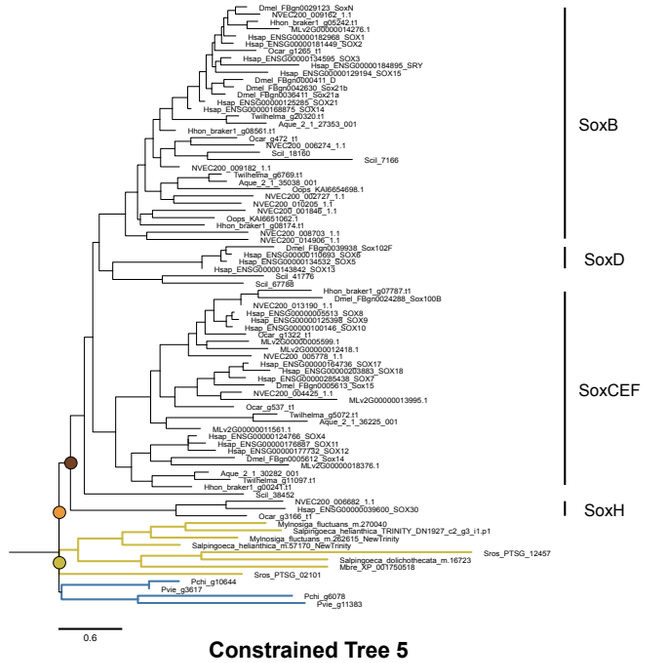
Constrained Tree 1
logL = -14536.18
p-AU: 0.374



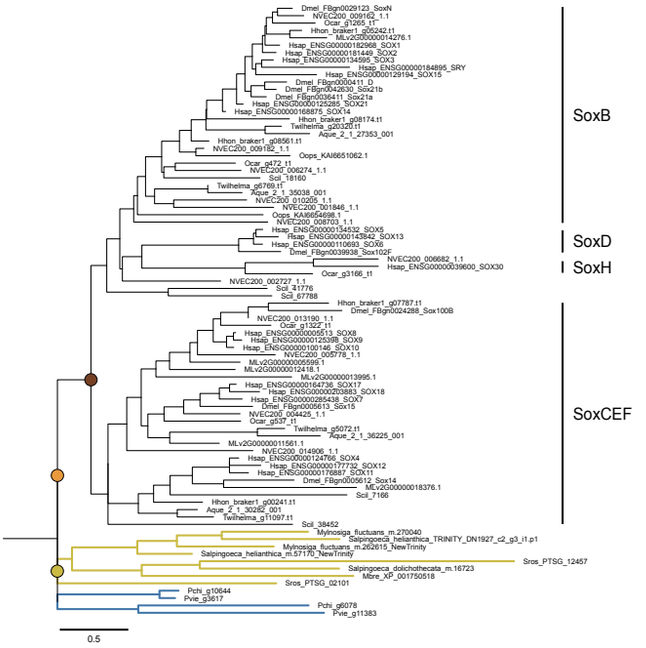
Constrained Tree 2
logL = -14541.35
p-AU: 0.296



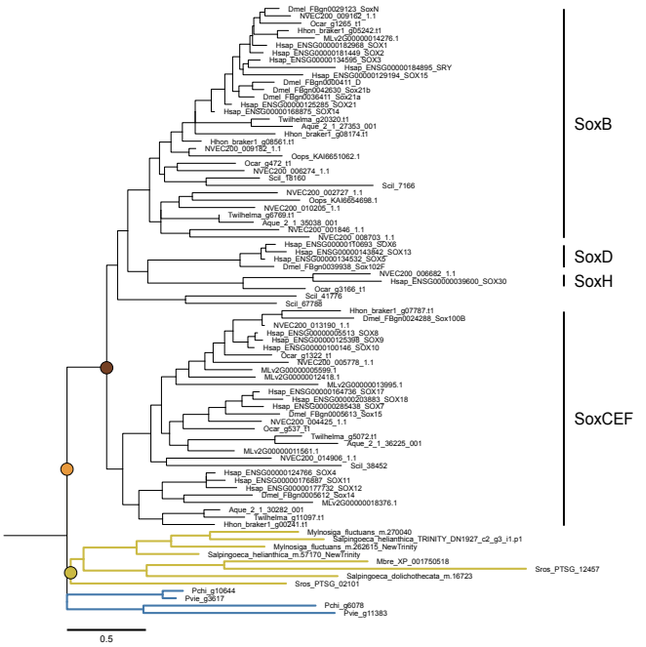
Constrained Tree 3
logL = -14541.49
p-AU: 0.286



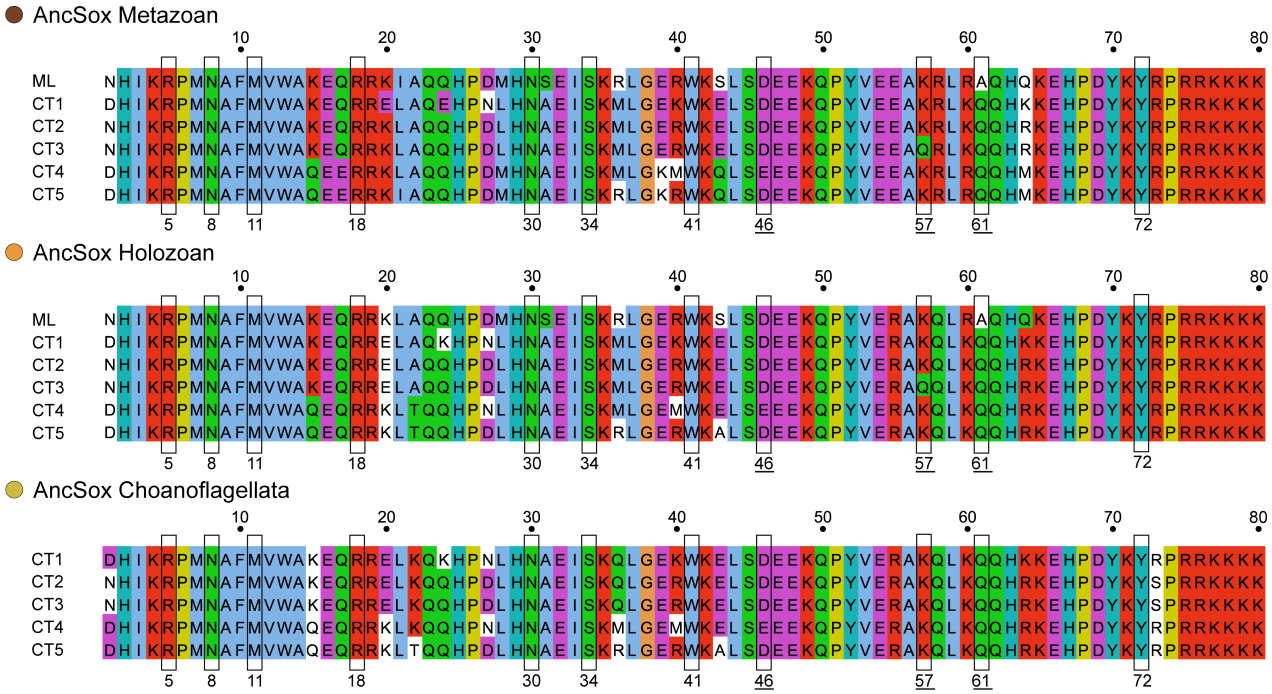
Constrained Tree 4
logL = -14545.56
p-AU: 0.34



Constrained Tree 5
logL = -14554.03
p-AU: 0.197



Supplementary Fig.10 for Fig.4. Comparison of maximum likelihood and constrained trees.
Side-by-side comparison and approximately unbiased (AU) test [3] of the maximum-likelihood (ML) tree and five alternative constrained trees (CT). The ML tree is the same is in Supplementary Figure S1 but only the Sox/Sox-like clade is shown. The constrained trees were all loosely constrained according to the holozoan species tree by generating a guide tree with the Sox/Sox-like genes of metazoans (black branches), choanoflagellates (yellow branches), and filastereans (blue branches) grouped into separate clades that were arranged into the following topology: (Filasteria, (Choanoflagellata, Metazoa)). For generation of the guide tree, polytomies were applied to all internal branches and to those of the outgroup sequences and all branch length were set to 1.0. The constrained tree searches were repeated 10 times using IQ-TREE and the five trees with the highest log-Likelihood (logL) values are shown. The ML and constrained trees were subjected to an AU-test (implemented into IQ-TREE) using the original multiple-sequence alignment, the substitution model LG+G4 and 10,000 replicates. The computed logL values and p-values of the AU-test (p-AU) are shown above the trees. None of the five tested constrained trees were rejected by the AU-test as their computed p-AU values were all above 0.05. The alternative ancestral nodes for metazoan (brown), holozoan (orange) and choanoflagellate (gold) Sox/Sox-like genes are indicated with circles. Species acronyms as in Supplementary Fig.1.



Supplementary Fig.11 for Fig.4. Possible sequences of ancestral Sox variants.
Multiple sequence alignments (MSA) of selected ancestral Sox proteins, inferred via ancestral sequence reconstruction (ASR) from the maximum-likelihood (ML) tree (see Supplementary Figure 1 and 9) and the constrained trees (CT) shown in Supplementary Fig.10. Shown are the inferred ancestral sequences of the following nodes: AncSox Metazoan, AncSox Holozoan, AncSox Choanoflagellata. The corresponding nodes are indicated in Supplementary Fig.10. The MSA is colored according to the Clustal X color scheme. The black boxes highlight positions that are, in metazoan Sox proteins, critical for DNA recognition or selective pairing with Oct (additionally highlighted with underscores).

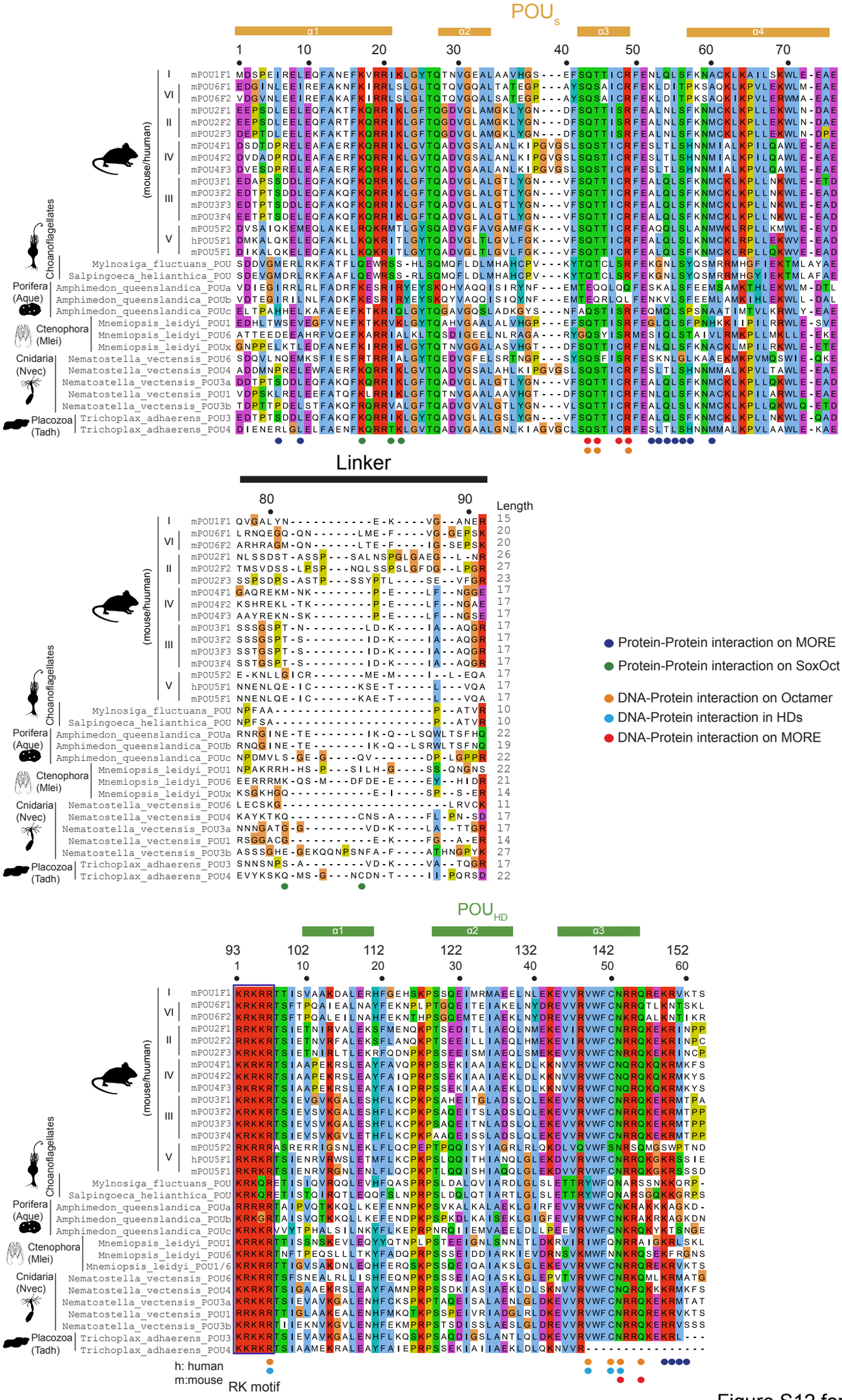
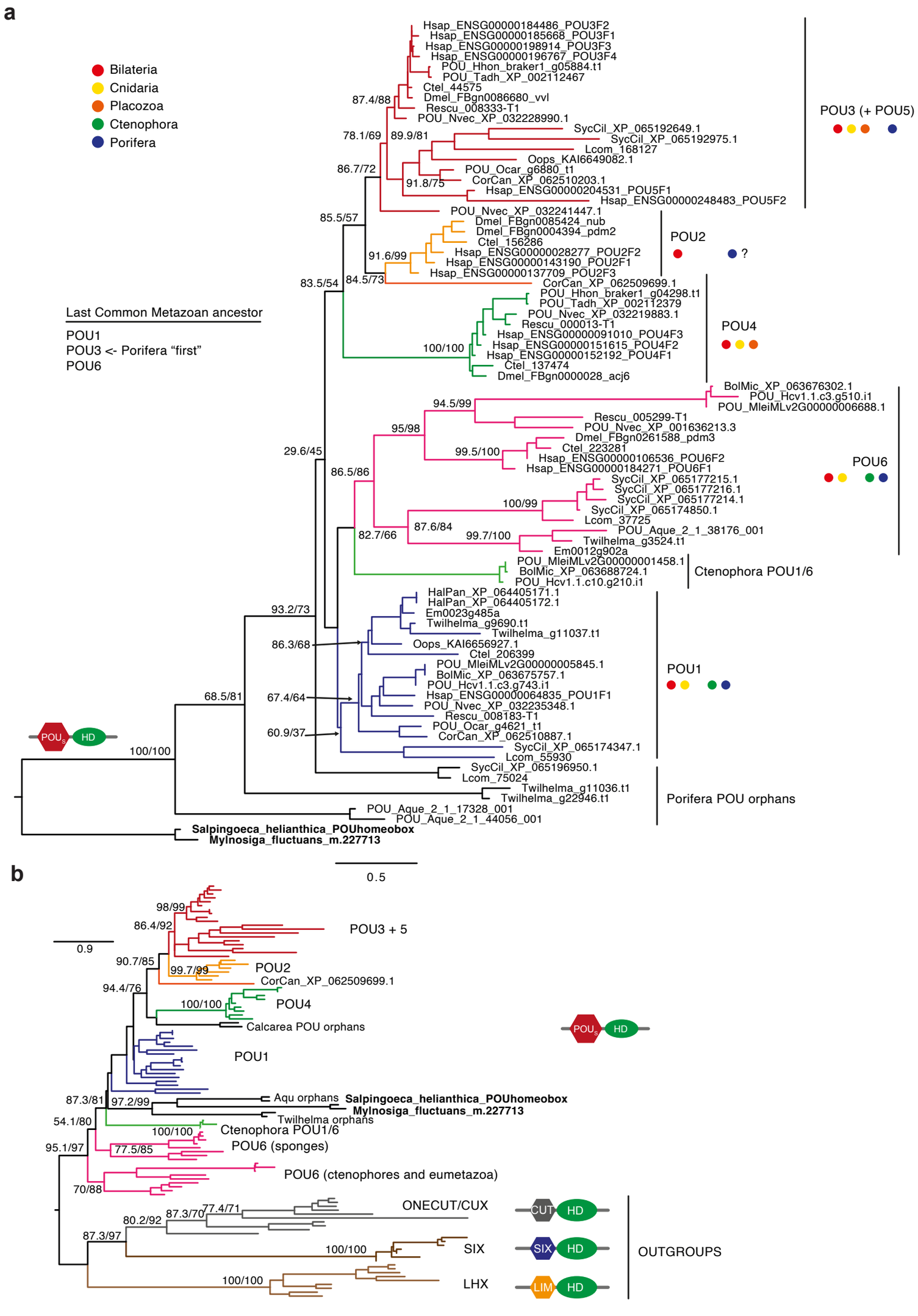


Figure S12 for figure 5

Supplementary Fig.12 for Fig.5. Alignment and annotated sequences of holozoan POU factors.
Multiple sequence Alignment of all mouse POU family members, human OCT4 (hPOU5F1) and the POU found in indicated animal and unicellular species. Roman numbers I-VI indicate the POU subfamilies POU1-POU6. α-helices are marked with colored bars and the flexible linker with a black bar. Blue and green circles indicate MORE and SoxOct protein-protein interaction sites, respectively. Orange and red circles mark residues engaged in base-specific binding in the context of Octamer and MORE DNA, respectively. Light blue circles indicate residues involved in sequence recognition by homeodomain proteins (HD). The numbering scheme refers to the POU DBD numbering scheme. For the POU_{HD} the conventional homeodomain numbering scheme is shown underneath. The silhouettes of the species are sourced from BioRender.com with license for publication or PhyloPic (http://phylopic.org).


a
Bilateria
Cnidaria
Placozoa
Ctenophora
Porifera
Last Common Metazoan ancestor
POU1
POU3 <- Porifera “first”
POU6
Hsap_ENSG00000184486_POU3F2
Hsap_ENSG00000185668_POU3F1
Hsap_ENSG00000198914_POU3F3
Hsap_ENSG00000196767_POU3F4
POU_Hhon_braker1_g05884.t1
POU_Tadh_XP_002112467
Ctel_44575
Dmel_FBgn0086680_vvl
Rescu_008333-T1
POU_Nvec_XP_032228990.1
SycCil_XP_065192649.1
SycCil_XP_065192975.1
Lcom_168127
Oops_KAI6649082.1
POU_Ocar_g6880_t1
CorCan_XP_062510203.1
Hsap_ENSG00000204531_POU5F1
Hsap_ENSG00000248483_POU5F2
POU_Nvec_XP_032241447.1
Dmel_FBgn0085424_nub
Dmel_FBgn0004394_pdm2
Ctel_156286
Hsap_ENSG00000028277_POU2F2
Hsap_ENSG00000143190_POU2F1
Hsap_ENSG00000137709_POU2F3
CorCan_XP_062509699.1
POU_Hhon_braker1_g04298.t1
POU_Tadh_XP_002112379
POU_Nvec_XP_032219883.1
Rescu_000013-T1
Hsap_ENSG00000091010_POU4F3
Hsap_ENSG00000151615_POU4F2
Hsap_ENSG00000152192_POU4F1
Ctel_137474
Dmel_FBgn0000028_acj6
BolMic_XP_063676302.1
POU_Hcv1.1.c3.g510.i1
POU_MleiMLv2G00000006688.1
Rescu_005299-T1
POU_Nvec_XP_001636213.3
Dmel_FBgn0261588_pdm3
Ctel_223281
Hsap_ENSG00000106536_POU6F2
Hsap_ENSG00000184271_POU6F1
SycCil_XP_065177215.1
SycCil_XP_065177216.1
SycCil_XP_065177214.1
SycCil_XP_065174850.1
Lcom_37725
POU_Aque_2_1_38176_001
Twilhelma_g3524.t1
Em0012g902a
POU_MleiMLv2G00000001458.1
BolMic_XP_063688724.1
POU_Hcv1.1.c10.g210.i1
HalPan_XP_064405171.1
HalPan_XP_064405172.1
Em0023g485a
Twilhelma_g9690.t1
Twilhelma_g11037.t1
Oops_KAI6656927.1
Ctel_206399
POU_MleiMLv2G00000005845.1
BolMic_XP_063675757.1
POU_Hcv1.1.c3.g743.i1
Hsap_ENSG00000064835_POU1F1
POU_Nvec_XP_032235348.1
Rescu_008183-T1
POU_Ocar_g4621_t1
CorCan_XP_062510887.1
SycCil_XP_065174347.1
Lcom_55930
SycCil_XP_065196950.1
Lcom_75024
Twilhelma_g11036.t1
Twilhelma_g22946.t1
POU_Aque_2_1_17328_001
POU_Aque_2_1_44056_001
Salpingoeca_helianthica_POUhomeobox
Mylnosiga_fluctuans_m.227713
POU3 (+ POU5)
POU2
?
POU4
POU6
Ctenophora POU1/6
POU1
Porifera POU orphans
POUs
HD
94.5/99
95/98
99.5/100
86.5/86
100/99
87.6/84
99.7/100
82.7/66
29.6/45
93.2/73
86.3/68
68.5/81
67.4/64
60.9/37
100/100
87.4/88
78.1/69
89.9/81
86.7/72
91.8/75
85.5/57
91.6/99
83.5/54
84.5/73
100/100
0.5
b
98/99
86.4/92
0.9
90.7/85
99.7/99
POU3 + 5
POU2
CorCan_XP_062509699.1
94.4/76
100/100
POU4
Calcarea POU orphans
POU1
87.3/81
97.2/99
Aqu orphans
Salpingoeca_helianthica_POUhomeobox
Mylnosiga_fluctuans_m.227713
54.1/80
Twilhelma orphans
100/100
Ctenophora POU1/6
95.1/97
77.5/85
POU6 (sponges)
POU6 (ctenophores and eumetazoa)
70/88
87.3/70
77.4/71
ONECUT/CUX
80.2/92
87.3/97
100/100
SIX
100/100
LHX
CUT
HD
SIX
LIM
OUTGROUPS

Supplementary Fig.13 for Fig.5. Phylogeny of holozoan POU factors.
a Maximum likelihood phylogeny of POU family members in Holozoans, rooted using the choanoflagellate sequences. The tree is constructed based on an alignment of both POU domains: POU specific domain and the homeobox. Major known animal POU families are highlighted with branch colors, and the legend of presence/absence in early diverging animal lineages is depicted with colored circles as shown in legend on the top left. **b** Maximum likelihood phylogeny of the POU family members as in panel a, but including non-POU Homeobox outgroups for rooting, selecting those that have been previously shown to be closely related to the POU lineage and that have structurally analogous N-terminal domains to the Homeobox: ONECUT, LHX and Six. Same clade color coding as in panel a, with PFAM domain structure configurations shown on the right-hand side. For both trees, nodes of interest are labelled with support values for ultrafast bootstrap approximation (UFBoot) and SH-like approximate likelihood ratio (SH-aLRT) testing. Species identifiers for the sequences: Bilateria - Hsap (*Homo sapiens*), Dmel (*Drosophila melanogaster*), Ctel (*Capitella teleta*), Cnidaria - Nvec (*Nematostella vectensis*), Rescu (*Rhopilema esculentum*), Placozoa - Tadh (*Trichoplax adhaerens*), Hhon (*Hoilungia_hongkongensis*), Ctenophora - BolMic (*Bolinopsis microptera*), Mlei (*Mnemiopsis leidyi*), Hc (*Hormiphora californensis*), Porifera - Demosponges - Aque (*Amphimedon queenslandica*), Twhilhelma (*Tethya wilhelma*), Em (*Ephydatia muelleri*), HalPan (*Halichondria panicea*), Hexactinelid - Oops (*Oopsacas minuta*), Homoscleromorpha - Ocar (Oscarella carmela), CorCan (*Corticium candelabrum*), Calcarea - SycCil (*Sycon ciliatum*), Lcom (*Leucosolenia complicata*).

References

1. Lee H, Song JJ. The crystal structure of Capicua HMG-box domain complexed with the ETV5-DNA and its implications for Capicua-mediated cancers. *FEBS J* **286**, 4951-4963 (2019).

2. Ng CK, Li NX, Chee S, Prabhakar S, Kolatkar PR, Jauch R. Deciphering the Sox-Oct partner code by quantitative cooperativity measurements. *Nucleic Acids Res* **40**, 4933-4941 (2012).

3. Shimodaira H. An approximately unbiased test of phylogenetic tree selection. *Syst Biol* **51**, 492-508 (2002).