Supplemental Text:

We initiated our analyses by performing a PSI-BLAST search (Altschul et al., 1997) of the UniRef50 database (Wu et al., 2006) with the N-terminal $Scm3^{Sc}$ conserved region (corresponding to amino acids 90 to 142). This revealed a Scm3 homologue in the marine choanoflagellate *Monosiga brevicollis* (in the second round with an *E*-value inclusion threshold of 0.05) and, after three rounds, marginal, yet non-significant, sequence similarities between the Scm3 family and bovine HJURP (*E* = 0.82).

Next, we used HMMer (Eddy, 1996) to search UniRef50 for more divergent Scm3 homologues using a hidden Markov model prepared from the N-terminal conserved region of the Scm3 family (which corresponds to $Scm3^{Sc}$ amino acids 90 to 128).

Finally, hidden Markov models of the fungal Scm3 protein alignment and that for the metazoan HJURP alignment were compared using using HHpred (Soding et al., 2005), and found to be highly significant ($E < 10^{-5}$).

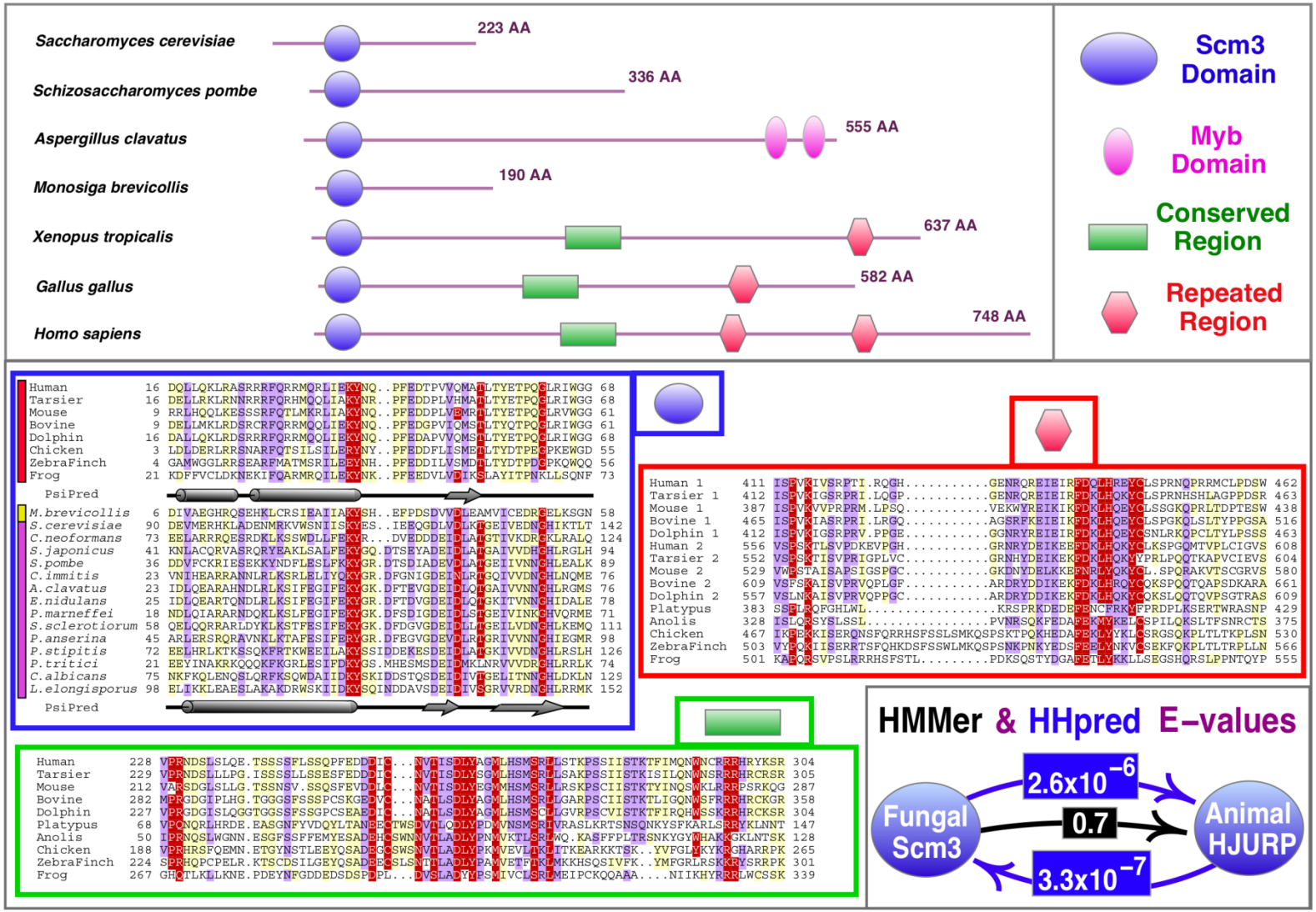
References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research *25*, 3389-3402.

Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Mazumder, R., O'Donovan, C., Redaschi, N., and Suzek, B. (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Research *34*, D187-191.

Eddy, S.R. (1996). Hidden Markov models. Curr. Opin. Struc. Biol. *6*, 361-365.

Soding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Research *33*, W244-248.


Saccharomyces cerevisiae
223 AA
Schizosaccharomyces pombe
336 AA
Aspergillus clavatus
555 AA
Monosiga brevicollis
190 AA
Xenopus tropicalis
637 AA
Gallus gallus
582 AA
Homo sapiens
748 AA
Scm3 Domain
Myb Domain
Conserved Region
Repeated Region
HMMer & HHpred E-values
Fungal Scm3
2.6x10−6
0.7
3.3x10−7
Animal HJURP

Supplemental Figure

Sequence analysis of the Scm3/HJURP protein family.

Top Panel: Schematic representation of evolutionary conserved regions among Scm3 domain-containing proteins.

Bottom Left Panel: Representative multiple sequence alignments of conserved regions (blue, green and red) in Scm3 proteins from tetrapods.

Bottom Right Panel: Numbers correspond to global profile-to-sequence (HMMer) and profile-to-profile (HHpred) comparison *E*-values between the animal HJURP and fungal Scm3 domain alignments (Eddy, 1996; Soding et al., 2005). Arrows indicate the profile search direction. These significant *E*-values and the consistency of secondary structure predictions, provide confidence that the Scm3 domain is present in tetrapod HJURP proteins. Alignments were produced with T-Coffee and HMMer (Eddy, 1996; Notredame et al., 2000) using default parameters, slightly refined manually and viewed with the Belvu program (Sonnhammer and Hollich, 2005). The main groups of Scm3 domain-containing proteins are indicated by coloured bars to the left of the Scm3 domain alignment (blue box): red (tetrapods), yellow (choanoflagellate) and violet (fungi). The colouring scheme indicates average BLOSUM62 scores (correlated with amino acid conservation) for each alignment column: red (greater than 2.5), violet (between 2.5 and 1) and light yellow (between 1 and 0.2). Tetrapod sequences were obtained from UniProt, ENSEMBL, GenBank and GSC-WUSTL databases (Wu et al., 2006, Hubbard et al., 2009), but were supplemented by manually assembled ESTs and FGENESH+-predicted gene models (Solovyev et al., 2002). Tetrapod sequences are named according to their genus or common name. Accession numbers, database of origin and species names are: Human, Q8NCD3, *Homo sapiens*; Tarsier, ENSTSYP00000007653, *Tarsius syrichta*; Mouse, ENSMUSP00000054263; *Mus musculus*; Bovine, UPI0000F33924, *Bos taurus*; Dolphin, ENSTTRP00000004146, *Tursiops truncatus*; Platypus, ENSEMBL, *Ornithorhynchus anatinus*; Anolis, ENSEMBL, *Anolis carolinensis*; Chicken, Q5ZLF3; *Gallus gallus*; ZebraFinch, GSC-WUSTL, *Taeniopygia guttata*; Frog, ENSEMBL, *Xenopus tropicalis*. *Monosiga brevicollis* and fungal sequences, obtained from UniProt database [Wu et al., 2006], are named with their species name abbreviations. Their corresponding accession numbers are: A9V3K2, *Monosiga brevicollis* (choanoflagellate); Q12334, *Saccharomyces cerevisiae*; Q55S59, *Cryptococcus neoformans*; B6K7K7, *Schizosaccharomyces japonicus*; Q9HDY7, *Schizosaccharomyces pombe*; Q1DZJ0, *Coccidioides immitis*; A1C460, *Aspergillus clavatus*; Q5BD66, *Emericella nidulans*; B6QLW7, *Penicillium marneffei*; A7F2V5, *Sclerotinia sclerotiorum*; B2AYH7, *Podospora anserina*; A3LNZ2, *Pichia stipitis*; B2WCW5, *Pyrenophora tritici-repentis*; Q5AJC3, *Candida albicans*; and, A5DY01, *Lodderomyces elongisporus*. Secondary structure predictions were performed independently for the animal and fungal Scm3 domains, using PsiPred (Jones, 1999). Both results predicted the presence of a long alpha-helix located at the N-terminus of the domain (indicated by grey cylinders).

References

Eddy, S.R. (1996). Hidden Markov models. Curr. Opin. Struc. Biol. *6*, 361-365.
Soding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Research *33*, W244-248.
Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. *302*, 205-217.
Sonnhammer, E.L., and Hollich, V. (2005). Scoredist: a simple and robust protein sequence distance estimator. BMC Bioinformatics *6*, 108.
Solovyev, V., Kosarev, P., Seledsov, I., Vorobyev, D. (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol. 7 Suppl 1:S10.1-12.
Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. *292*, 195-202.