The unicellular ancestry of Groucho mediated repression and the origins of metazoan transcription factors

Richard R Copley

Genome Biology and Evolution Advance Access published May 18, 2016 doi:10.1093/gbe/evw118

Genome Biology and Evolution

SMBE

The unicellular ancestry of Groucho mediated repression and the origins of metazoan transcription factors.

Richard R. Copley

copley@obs-vlfr.fr

Sorbonne Universités, UPMC Univ Paris 06, CNRS, Laboratoire de Biologie du Développement de Villefranche-sur-mer (LBDV), 181 Chemin du Lazaret, 06230 Villefranche-sur-mer, France

Genome Biology and Evolution

Abstract

Groucho is a co-repressor that interacts with many transcription factors playing a crucial role in animal development. The evolutionary origins of Groucho are not clear. It is generally regarded as being a distinct animal specific protein, although with similarities to the yeast Tup-like proteins. Here I show that Groucho has true orthologs in unicellular relatives of animals. Based on their phylogenetic distribution, and an analysis of ligand binding residues, these genes are unlikely to be orthologs of the fungal Tup-like genes. By identifying conserved candidate Groucho interaction motifs in non-metazoan transcription factors, I demonstrate that the details of molecular interactions between Groucho and transcription factors are likely to have been established prior to the origin of animals, but that the association of Groucho interaction motifs with many transcription factor types can be regarded as a metazoan innovation.

Introduction

Many of the genes controlling animal development encode a circumscribed set of transcription factor domains, and many of these domains have a pre-metazoan ancestry, with homologs found in non-metazoan eukaryotes (Sebé-Pedrós et al. 2011; de Mendoza et al. 2013). Although clearly the developmental contexts of these proteins (such as patterning the nervous system) cannot exist in unicellular organisms, what is less clear is how directly analogous their molecular functions are. DNA binding specificities of several transcription factor domains have been shown to be broadly similar between single celled eukaryotes and animals (Kwon et al. 2012; Nakagawa et al. 2013; Sebé-Pedrós et al. 2013), and this is, to a large extent, expected from the conservation of DNA binding amino acids. If there are functional differences between proteins containing these domains (i.e. they are not, within the limits of their expression patterns, completely interchangeable), they are likely to have involved the protein-protein interactions (Wagner 2007; Copley 2008; Sebé-Pedrós et al. 2013; Hudry et al. 2014).

The proteins known as Groucho in *Drosophila*, and the Transducin Like Enhancers of Split (TLEs) in vertebrates are common interaction partners of animal transcription factors. The Groucho-like proteins act as transcriptional co-repressors and orthologs are found in all animal genomes (Copley 2005). No orthologs have been reported outside the Metazoa, but

in fungi the Tup-like (TUP1 in *S. cerevisiae,* tup11 & tup12 in *S. pombe*) proteins, that also act as transcriptional repressors, are sometimes regarded as the equivalent of Groucho (Chen and Courey 2000).

The relationship between TUP1-like and Groucho-like genes (for convenience I will refer to them as Tup and Groucho) has not been well defined. While noting functional similarities, Fisher and Caudy (1998) suggested that "it may be more accurate to consider TUP1 and Groucho proteins as analogous rather than truly homologous". Flores-Saaib and Courey (2000) mentioned that the overall similarity between Groucho and Tup WD40 regions was not significantly greater than between Groucho and other WD40 repeat containing proteins without functional similarities. Based on a more detailed analysis of the sequences of corresponding repeats, they went on to suggest that the proteins were "structurally and therefore perhaps functionally, related", and proceeding on this basis, demonstrated similarities of Groucho and TUP1 histone interactions. Pickles et al. (2002) stated that the two were "increasingly considered as functional equivalents". Other recent authors have more or less explicitly considered them orthologs – that is, encoded in genes related by speciation events (Matsumura et al. 2012; Asada et al. 2015).

There are, however, marked differences in the biology of Tups and Grouchos. Yeast TUP1 proteins form a functional complex with the TPR repeat containing CYC8 (ssn6 in *S. pombe*) protein (Tzamarias and Struhl 1994), but there is no similar co-factor requirement for Groucho. Groucho interacts with EH1 & WRPW protein motifs from a variety of animal transcription factors (Jennings and Ish-Horowicz 2008). Pearl, Ish-Horowicz and co-workers stated that there were no obvious WRPW motif proteins in yeast, suggesting that yeast transcription factors interact with Tup via amphipathic helices similar to the EH1 motif (Jennings et al. 2006). There are not, however, any reported yeast transcription factors with motifs matching the metazoan EH1 consensus. The metazoan EH1 motif as currently described typically begins with a phenylalanine, or less often tyrosine, with a consensus of FS[VI]xx[IL][LM] (see (Copley 2005)). Without the F or Y, the motif is poorly specified and large numbers of amphipathic helices would be expected to match.

The importance of Groucho mediated repression in animal development, its inferred presence in the most recent common ancestor of the animals and absence in other groups, raises the question of its evolutionary origin. The presence of an analogous Tup system in yeasts could, however, potentially shed light this, if its relationship with Groucho were better

understood. To enquire farther into the origins of Groucho, I have examined the evolutionary history of Tup and Groucho and their likely molecular interactions, with a particular focus on recently available genomic and transcriptomic data from close unicellular relatives of the animals.

Taxanomic distribution of Tup and Groucho-like proteins

The Groucho and Tup proteins are composed of N-terminal coiled-coil domains and a C-terminal 7 bladed β-propeller composed of WD40 repeats. WD40 repeats are widespread in animal and eukaryotic proteins, and their repeating nature makes them particularly prone to mis-alignment, making similarity scores hard to interpret. In contrast, structural superimpositions of the 3D structures of the N-terminal domains of TUP1 and TLE suggest that they are distinct from each other and unique to these proteins (**Figure 1**) - the superficial resemblance of the coiled coils is not reflected in any statistically significant sequence similarity. I conjecture that proteins containing a TLE_N Pfam region are orthologs of Groucho/TLE and those containing a Tup_N region, orthologs of TUP1, and that these regions can be used as proxies to determine the phylogenetic distribution of their respective genes. Later phylogenetic analysis of the recovered sequences will show this conjecture to be correct.

Accordingly, I searched the nr protein database from the NCBI with the Pfam hidden Markov models Tup_N and TLE_N, using their associated 'gathering' threshold bit scores as cutoffs (TLE_N, 24.0 bits; Tup_N 22.4 bits) (Finn et al. 2016). Significant (E<0.001) TLE_N hits were to animals (with the exception of *Sphaeroforma arctica* (discussed below). Among the non-bilaterians, significant TLE_N hits were found to sequences in the cnidarians, *Nematostella vectensis* and *Hydra vulgaris*; the placozoan, *Trichoplax adhaerens*; and the sponge *Amphimedon queenslandica*. Further searches of the draft contigs and scaffolds of ctenophore genomes using representative protein sequences revealed likely Groucho candidates in *Mnemiopsis leidyi* and *Pleurobrachia bachei.* These results show that, irrespective of controversies of the branching order of non-bilaterian animals (Pisani et al. 2015; Whelan et al. 2015), Groucho was present in the common ancestor of all extant animals. This is consistent with the same inference drawn from the phylogenetic distribution of EH1 motifs (Copley 2005). A single TLE_N hit was also identified to a non-metazoan eukaryote - *Trimastix pyriformis*, represented in the NCBI TSA archive. Hits to Tup_N were primarily to Fungi. Non-fungal eukaryotes included Oomycetes, Amoebozoa, *Naegleria*

gruberi (Heterolobosea), *Galdieria sulphuraria* (Rhodophyta), *Ectocarpus siliculosus* (stramenopiles) and *Guillardia theta* (Cryptophyta).

To more closely examine the separation between fungal and animal sequences, I searched both Tup_N and TLE_N against the proteins of the "Origin of Multicellularity Project" (Ruiz-Trillo et al. 2007). For TLE_N (i.e. Groucho/TLE) this resulted in significant matches to *Ameobidium parasiticum* and *Sphaeroforma arctica*, but no other species within this project. Both of these taxa belong to the clade of Ichthyosporea. For Tup_N (i.e. Tup), I found significant matches to proteins from *Spizellomyces punctatus*, *Mortierella verticillata* and *Allomyces macrogynus*, the three fungal taxa represented in the project. Notably, I was unable to find matches of either Tup_N or TLE_N to choanoflagellate (*Monosiga, Salpingoeca*) or filasterean (*Capsaspora owczazarki*) protein sets, both of which are more closely related to the Metazoa than the ichthyosporeans. I also searched Tup_N and TLE_N against protein sets generated from the data in Torruella et al. (2015) (see methods)**.** This resulted in further matches of TLE_N (Groucho) to Ichthyosporea and Corallochytrium taxa, and Tup_N matches to Nutomonas and Nuclearia species, essentially confirming the phylogenetic distribution of the 'Multicellularity Project' set.

The use of Groucho and Tup N-terminal domains conveniently avoids cross matching between different WD40-repeat containing proteins, but it is possible that some *bona fide* orthologs of these proteins diverged before the N-terminus became associated with the WD40 repeats, or subsequently lost the N-terminus. To identify possible orthologs of Tup and Groucho that may be lacking these domains in some species, I also performed searches using alignments of the complete β-propeller domain, implemented as a global-local model using the hmmer 2 software package. Two models were used, one built using representative metazoan Groucho sequences, and the other fungal Tup sequences. These models were searched against the nr database of the NCBI, and the eukaryotic sequence databases described above. As the models represent homologous sequences, there is considerable overlap between their hit lists. I conservatively defined the Groucho hit list to be those sequences scoring higher than the first Tup_N domain containing hit, and the Tup hit list to be those sequences scoring higher than the last non-fungal Tup_N domain containing hit with a positive score. There were no Tup_N hits in the Groucho hit list or TLE_N in the Tup hit list.

By combining the sequences matching the HMMs, I produced an alignment of representatives of the two groups of sequences (**Supplementary material**) and hence a single phylogenetic tree using the LG + Gamma model of sequence evolution as implemented in the phyml package (see methods). The tree shows clear separation of the Tup and Groucho groups, essentially mirroring the division between Holozoa (i.e. animals and their closest single celled relatives, but excluding fungi) and non-Holozoa (**Figure 2**). The Tup group includes all fungal sequences and the non-fungal eukaryotic sequences mentioned above, concordant with the analysis based on the presence of the Tup_N domain. It also includes additional non-metazoan non-Tup_N containing proteins, including that from *Fonticula alba,* a member of the Fonticula that together with the Nucleariа forms the sister group to fungi. The only non-holozoans in the Groucho branch of the tree are the TLE_N containing *Trimastix* sequence mentioned above and, in addition, a further *Naegleria* sequence lacking both TLE_N and Tup_N regions in the N-terminus. Importantly, the *Trimastix* and *Naegleria* sequences do not cluster within the bulk of the holozoan sequences, ruling out simple cross-species contamination, but to the base of the holozoan clade, as would be expected if they were *bona fide* Groucho-like sequences from non-holozoan eukaryotes. Analysis using a Bayesian tree reconstruction approach (see methods and **Figure S1**) produced similar results, with a strongly supported Groucho clade and *Naegleria* the first diverging lineage within it. In this analysis, however, the *Trimastix* sequence branched within the Teretosporea, although with weak support.

Again, no hits were identified to choanoflagellates or filastereans. This β-propeller based search also identified orthologs of the human TLE6 gene which, although a Groucho family member, lacks the TLE_N region, and another vertebrate family lacking the TLE_N, exemplified by the human locus 102723796 and mouse Gm21964 gene, for which there does not appear to be evidence of expression.

Structural and sequence features discriminating between Tup and Groucho

Aligning the WD40 repeat containing regions of proteins containing either a Tup_N or TLE_N N-terminal regions enabled the analysis of key residue differences between Grouchos and Tup. In particular, as 3D structures of Groucho (specifically, the human ortholog TLE1) in complex with EH1 and WRPW peptides are available (Jennings et al. 2006), the identity and conservation of residues mediating these interactions could be compared between Groucho and Tup.

I superimposed the structures of the C-terminal domains of TUP1, apo-Groucho, Groucho with EH1 bound, and Groucho with WRPW bound, using the STAMP package (Russell and Barton 1992). Inspection of residues within 5Å of the EH1 peptide shows that they are found in comparable positions (**Figure 3**). The major visual difference lies in the orientation of the side chains of TUP1-Y580 and Groucho-F661. This appears to be a consequence of a further substitution, TUP1-L596 vs Groucho-E677. Whereas the charged Groucho side chain is extended away from the bulk of the protein fold, the non-polar TUP1 residue is half buried within the fold, with the sidechain contacting Y580 and re-orienting it towards the 'pore' of the β-propeller domain, relative to the orientation of Phe found in all the Groucho crystal structures. If the TUP1 residue configuration were observed in Groucho, there would be a steric clash between the Y580 equivalent and the Phe of the EH1 motif (or W of the WRPW motif) (**Figure 3**). From this, it appears as though the Tup fold as observed is incompatible with EH1 or WRPW binding in the configurations seen in current crystal structures. This is in accord with the result that no true EH1 or WRPW like motifs have been reported in yeast transcription factors (although see below).

This residue dichotomy (Y,L in Tup-like and F,E in Groucho-like) is conserved (**Figure 4**). Proteins that have a TUP_N domain have the Y,L pair, whereas those having a TLE_N have the F,E pair, with the exceptions of the most divergent single celled eukaryotic taxa and the vertebrate TLE6 orthologs. As the remainder of the protein binding pocket, interacting with the other ligand residues, appears conserved, it is possible that Tup-like proteins may be able to bind EH1 like motifs that lack the initial Phe residue.

Tup interaction motifs in fungi

The *Saccharomyces cerevisiae* gene MATα2, a homeobox containing transcription factor and regulator of mating type genes, interacts with the WD40 region of TUP1. Mutation of N-terminal residues Ile4, Leu9 or Leu10, (or also Gly71) disrupts the interaction with TUP1 (Komachi et al. 1994). The characteristic spacing of these residues (i.e. **I**xIxx**LL**) is obviously reminiscent of the EH1 motif (**F**xIxx**IL**). This N-terminal motif is well conserved in other yeast MATα2 orthologs, and has been recently interpreted as a modified version of the EH1 motif by Bürglin and Affolter (2015).

In order to screen for similar motifs in a well characterized fungal genome, I searched protein alignments of orthologous genes from *Schizosaccharomyces* group genomes (*Schizosaccharomyces pombe, Schizosaccharomyces japonicus, Schizosaccharomyces octosporus and Schizosaccharomyces cryophilus*) (see methods). Instances of the motif **[VI]xx[IL][LM]** (essentially the highest scoring residue types of the EH1^{hox} motif without the first two residues) that matched the *S. pombe* sequence in non-domain regions of proteins containing transcription factors domains, and where the motif was conserved in the *Schizosaccharomyces* group alignment, were considered further.

The pombase database (McDowall et al. 2015) annotates 90 genes with a molecular function of 'RNA polymerase II core promoter proximal region sequence-specific DNA binding' (GO:0000978), of which 60 are represented in the set of aligned orthologs. Of these, 10 contain **[VI]xx[IL][LM]** motifs conserved in all pombe group species, and 8 are recorded as interacting with tup11 or tup12 (the pombe TUP1 orthologs) in the biogrid database (Oughtred et al. 2016). Of the proteins containing the motif, three, fep1, res1 and scr1 interact with tup11/12 according to biogrid, with an additional one (sak1) predicted as interacting via homology to an *S. cereveisiae* interacting pair in the STRING database (Szklarczyk et al. 2015). These numbers suggest an enrichment of the conserved motif in transcription factors that interact with tup11/tup12 relative to those that do not (P=0.0343, Fisher's exact test), but it must be noted that as the total number of interactors and genes are small, the result is not especially robust. Furthermore, no studies have specifically focussed on the protein-interaction partners of tup11/tup12 in pombe (or TUP1 in *S. cerevisiae*), leading to the possibility that there are considerable numbers of interacting partners yet to be discovered.

Groucho interaction motifs in Ichthyosporean transcription factors

The analyses of the N-terminal domains and WD40 repeat regions show that orthologs of Groucho are present in the Icthyosporeans *Amoebidium parasiticum* and *Sphaeroforma arctica*. If unicellular Groucho orthologs have the same molecular function, as indicated by the conservation of the key Y,E residue pair in the WD40 domain, we would expect to be able to identify proteins containing EH1 or WRPW motifs. By analogy with metazoan Grouchos, we might further expect these motifs to be preferentially present in transcription factor proteins.

To better understand the role of these proteins, I searched a database of proteins from taxa in the evolution of multicellularity project, with the EH1^{hox} hidden Markov model described in Copley (2005). As these motifs typically occur in non-protein domain contexts, I first masked known Pfam domains. Retrieving hits containing known transcription factor domains (Wilson et al. 2008), 8 out the top 10 were found in *Amoebidium* and *Sphaeroforma*, the only genomes in the set that encode groucho like proteins. As these sequences are all uncharacterized experimentally, I sought evidence of function via evolutionary constraint on sequence evolution by searching for orthologs and paralogs that shared these putative EH1 motifs. In addition to the 'evolution of multicellularity project' proteins, I searched proteins generated from the assembled transcriptomes of the 'Close Relatives of Animals and Fungi' project (Torruella et al. 2015).

Two paralogs within *Amoebidium* mutually supported each other, showing conservation of EH1 motifs in the absence of conservation of surrounding sequence (**Supplementary Figure S2a**). A protein from *Sphaeroforma*, including an N-terminal MYND ZnF and C2H2 Zn fingers contained an EH1 motif that was conserved in an orthologous *Creolimax fragrantissima* sequence (**Supplementary Figure S2b**). An *Amoebidium* sequence including Ankyrin repeats and GATA ZnF had an ortholog in *Ichthyophonus hoferi*, showing conservation of the EH1 motif (**Supplementary Figure S2c**). An additional GATA ZnF protein from *Sphaeroforma arctica* with an N-terminal EH1 motif had readily identifiable orthologs in *Creolimax*, *Amoebidium*, *Pirum gemmata* and *Abeoforma whisleri*. An EH1 motif was identifiable in the *Creolimax* ortholog. The *Amoebidium*, *Pirum* and *Abeoforma* sequences did not contain EH1-like motifs, but instead, and remarkably, conserved WRPW motifs at equivalent positions, suggestive of convergent evolution of distinct binding motifs within orthologous proteins (**Supplementary Figure S2d**).

The Groucho ortholog in *Naegleria gruberi* does not contain the Tyrosine of the Y,E pair, but instead Leucine. Despite the availability of predicted proteins for the complete genome sequence, I was not able to detect significant matches of the EH1 or WRPW motifs associated with *Naegleria* transcription factor domains. Similarly, no EH1 or WRPW matches were detected in the available transcripts from *Trimastix pyriformis,* where the F and Y amino acids correspond to the Y,E pair.

Groucho interaction motifs in non-bilaterian metazoan transcription factors

The major metazoan associations of EH1 motifs are with the homeobox, forkhead and T-box transcription factor domains (Copley 2005). Of these associations, all are found in sponges and ctenophores, with only the EH1 forkhead association being absent from *Trichoplax* (**Figure 5**). The [WF]RP[WY] motif is associated with HLH and Runt transcription factors. The HLH association is present in sponges, ctenophores and *Trichoplax*. The [WF]RP[WY] Runt association is found in *Mnemiopsis* (ML03045a), but not in available sponge or *Trichoplax* sequences. Interestingly, the *Mnemiopsis* T-box protein ML45844a encodes a T-box, EH1 motif and a C-terminal WRPW motif.

Notably, neither the EH1 or WRPW-type motif is found associated with GATA transcription factors, as found in *Ichthyosporea*. Although there are associations between EH1 motifs and C2H2 Zinc fingers in bilateria, none are found in the non-bilaterian Metazoa investigated here. There is thus a discontinuity between single-celled eukaryotes and the Metazoa.

Discussion

The data presented here show that the origins of the metazoan transcriptional co-repressor Groucho predate the Metazoa. I have identified likely Groucho orthologs in the single celled eukaryotes of the Ichthyosporean clade, and further, identified Icthyosporean transcription factors that contain conserved Groucho Interaction Motifs (GIMs). The Ichthyosporean transcription factors with GIMs have no obvious relationships to the typical metazoan proteins containing GIMs, suggesting that the quantitative expansion in transcription factor numbers in the animal stem lineage (de Mendoza et al. 2013) co-occurred with a re-wiring of protein-protein interactions, to make use of Groucho mediated repression.

The Tup and Groucho proteins have long been regarded as functional equivalents in fungi and animals respectively. Their dichotomous phylogenetic distribution (Tup in fungi, Groucho in animals) and shared role in transcriptional repression has been suggestive of an orthologous relationship. Increased sequence sampling of eukaryotic species has extended the range of Tup-like genes beyond the fungi, including non-opisthokont species. In the phylogenetic analysis presented here, the Groucho group clearly does not arise from within this Tup clade, but rather has a sister group relationship with it, suggesting an equally ancient history. Furthermore, the phylogenetic distribution of Groucho and Tup orthologs revealed two excavate species, *Naegleria gruberi* and *Trimastix pyriformis* that appear to encode Groucho-like proteins, with *Naegleria* also encoding a Tup protein. Taken together

with the broad eukaryotic distribution of Tup, this presents a *prima facie* case that both Groucho-like and Tup-like proteins were present in the eukaryotic common ancestor, although clearly distinguishing between this and alternative scenarios of horizontal gene transfer or contamination (or mis-identifed species) and phylogenetic reconstruction artefacts would be made easier by the availability of more non-parasitic eukaryotic genome sequences.

Detailed comparison of the three dimensional protein structures of Groucho and Tup, at the level of the conservation in Tup of the binding site residues of Groucho is further suggestive of a non-orthologous relationship between the two. In particular, two amino acid substitutions play a role in restructuring the binding site of Groucho. The need for multiple substitutions and the presumed biological requirement of functional continuity is more likely to have occurred in a duplicated gene copy. Analysis of yeast transcription factors and their conservation suggests some likely genes encoding candidate interaction motifs, but not, apparently, to the extent seen in Metazoa. Two possibilities suggest themselves: firstly, the ability to discriminate 'F' as the first motif residue enables a greater utility, in the sense that LxxLL is more likely to occur by chance in protein sequences, making it harder to discriminate between 'functional' and non-functional motifs; secondly, that the TPR repeat containing CYC8/ssn6 co-factor plays a crucial role in transcription factor recognition in yeasts, and that specificity is not encoded solely in the WD40 β-propeller domain. Interestingly, the TPR repeats of CYC8/ssn6 appear to be orthologous to the TPR repeats encoded in the human histone demethylase KDM6A/UTY genes (they are reciprocal blast best hits, data not shown), and these latter proteins have been shown to interact with TLE1 (Grbavec et al. 1999).

Among the eukaryotes, plants encode no WD40 β-propeller domains that are obviously orthologs of Groucho or Tup. The *Arabidopsis* protein TOPLESS is frequently described as being a plant equivalent of Groucho/Tup (Liu and Karmarkar 2008), but at the level of primary sequence, contains two WD40 β-propeller domains and distinct N-terminal domains. TOPLESS binds **L**x**L**x**L** motifs present in many plant transcription factor proteins. The recently solved 3D structure of the N-terminal domain, however, demonstrates that the interaction of the peptide motif is with this, rather than the WD40 domain as found in Groucho/Tup, suggesting it has arisen via an independent evolutionary path (Jennings and Ish-Horowicz 2008; Ke et al. 2015).

Groucho proteins also interact with the transcription factor TCF/LEF, the effector of WNT signalling, via an interaction of their N-terminal domains (Chodaparambil et al. 2014). The fact that unicellular eukaryotes encode orthologs of Groucho, but not TCF like transcription factors, suggests that interactions with groucho via EH1 and WPRW type motifs arose before those with TCF/LEF. This inference is consistent with the fact that WNT ligands are found only within the Metazoa.

Methods

Data sources

The NR protein database was downloaded from the NCBI (20th September 2015) ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz.

Proteins from the 'Origins of Multicellularity' project were downloaded from https://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiHome.html Sequence reads from species referred to in Torruella et al. (2015) were downloaded from the EBI ENA database and assembled using Trinity with open reading frames being identified using Transdecoder (Grabherr et al. 2011).

Predicted proteins from Schizosaccharomyces were downloaded from https://www.broadinstitute.org/scientific-community/science/projects/fungal-genome-initiative/schizosaccharomyces-genomes-project

Sponge proteins were downloaded from http://compagen.org/datasets.html
Mnemiopsis leidyi protein models were taken from:
http://research.nhgri.nih.gov/mnemiopsis/download/download.cgi?dl=proteome
Trichoplax adhaerens protein models were taken from:
ftp://ftp.ncbi.nih.gov/genomes/refseq/invertebrate/Trichoplax_adhaerens/

Phylogenetic analysis

Representative WD40 containing regions form Tup and Gro proteins were aligned using the MAFFT program (using the L-INS-i options) (Katoh and Standley 2013). WD40 sequences from WDR5 proteins from *Capsaspora owczarzaki*, *Amphimedon quenslandica*, *Trichoplax*

adhaerens, *Nematostella vectensis* and human were added to serve as an outgroup. Ragged N and C-termini were trimmed, but the alignment was otherwise unedited. Analysis using the protest3 software gave LG+G as the best fitting model (Darriba et al. 2011). Accordingly, phylogenetic analysis was performed using Phyml with the LG+G model (Le and Gascuel 2008), with other parameters left as defaults (Guindon et al. 2010). 100 bootstrap replicates were performed. The data were also analysed with phylobayes, which uses a Bayesian rather than Maximum Likelihood approach (Lartillot et al. 2009), again using the LG+G model and using 2 chains. Chains were run for 35000 generations. A consensus tree was produced using bpcomp from the phylobayes package, discarding the first 20000 generations, giving a maxdiff of 0.1 and a meandiff of 0.003.

Ortholog identification in the Schizosaccharomyces group.

In order to screen for similar motifs in a well characterized fungal genome, I inferred orthologous groups in the *Schizosaccharomyces* group genomes (*Schizosaccharomyces pombe, Schizosaccharomyces japonicus, Schizosaccharomyces octosporus and Schizosaccharomyces cryophilus*), via mutually consistent groups of 4 reciprocal best hits in all against all searches performed with the phmmer program from the hmmer package (http://hmmer.org/). Instances of PFAM domains within *S. pombe* sequences were recorded using hmmsearch from the hmmer package, and only motif matches occurring to sequence regions outside these coordinates were assessed for conservation in the remaining three species.

Acknowledgements

Thanks to Stefania Castagnetti, Lucas Leclère (LBDV, Villefranche-sur-mer) and Max Telford (UCL, London) for helpful discussions.

Figure Legends

Fig. 1. The N-terminal domains of Groucho/TLE (PDB: 4om2) and TUP1 (PDB: 3vp8) do not share statistically significant sequence similarity: **a)** they adopt different quaternary structures - chains coloured from blue at N-terminus to red at C-terminus; **b)** both contain coiled-coil regions but with differing degrees of curvature and Groucho/TLE contains additional C-terminal helices.

Fig. 2a) Phylogenetic tree of aligned WD40 sequences from Groucho and Tup, with representative WDR5 proteins as an outgroup. The Tup and Groucho clades are boxed and labelled. Sequences within the Tup group typically include a Tup_N N-terminal motif, and those within the Groucho group a TLE_N motif. The *Naegleria* sequences are indicated with arrows. 'Teretosporea' is a clade of Icthyosporea and *Corallochytrium,* defined in Torruella et al. (2015). Black circles on nodes represent complete boostrap support, with numbers giving values for other nodes central to the discrimination of Tup and Groucho. Sequences that uniquely define the leaves are available in the supplementary information.

Fig. 2b) The distribution of Groucho and Tup orthologs identified in this study with respect to the three major eukaryotic groups (eukaryotic tree adapted from (He et al. 2014))

Fig. 3. Side view of the ligand binding pocket residues of TLE1 (3 structures, with bound EH1 (PDB: 2ce8), **W**RPW (PDB: 2ce9) and non-bound forms (PDB: 1gxr)), with equivalent TUP1 (PDB: 1erj) residues superimposed. The F and W of the TLE1 bound ligands are shown (Sprague et al. 2000; Jennings et al. 2006).

Fig. 4. Sequence conservation of the ligand binding pocket of the Groucho and Tup proteins (outgroup WDR5 members are also shown). The region is extracted from the full multiple sequence alignment, columns that are more than 80% identical within a class are coloured by amino acid type (Taylor 1997). The dichotomous F,E (in Groucho) and Y,L (in Tup) residues, likely to contribute to ligand recognition, are marked with arrows.

Fig. 5. Example associations from non-bilaterian metazoans of EH1 motifs with a) Homeobox, b) Forkhead and c) T-box domains. Sc = *Sycon ciliatum*, a sponge; ML = *Mnemiopsis leidyi*, a ctenophore; Ta = *Trichoplax adhaerens*. Sequence accessions

correspond to the databases described in the methods. Domain diagrams represent the *Sycon* sequences.

Supplementary files

1) FASTA format alignment of Groucho and Tup WD40 repeat containing regions

2) Newick format tree, including bootstrap values, for the Phyml analysis

3) Newick format tree, including posterior probabilities, for the phylobayes analysis

2) Supplementary figures 1 & 2.

References:

Asada R, Takemata N, Hoffman CS, Ohta K, Hirota K. 2015. Antagonistic controls of chromatin and mRNA start site selection by Tup family corepressors and the CCAAT-binding factor. Mol. Cell. Biol. 35:847–855.

Bürglin TR, Affolter M. 2015. Homeodomain proteins: an update. Chromosoma.

Chen G, Courey AJ. 2000. Groucho/TLE family proteins and transcriptional repression. Gene 249:1–16.

Chodaparambil JV, Pate KT, Hepler MRD, Tsai BP, Muthurajan UM, Luger K, Waterman ML, Weis WI. 2014. Molecular functions of the TLE tetramerization domain in Wnt target gene repression. EMBO J. 33:719–731.

Copley RR. 2005. The EH1 motif in metazoan transcription factors. BMC Genomics 6:169.

Copley RR. 2008. The animal in the genome: comparative genomics and evolution. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 363:1453–1461.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinforma. Oxf. Engl. 27:1164–1165.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44:D279–D285.

Fisher AL, Caudy M. 1998. Groucho proteins: transcriptional corepressors for specific subsets of DNA-binding transcription factors in vertebrates and invertebrates. Genes Dev. 12:1931–1940.

Flores-Saaib RD, Courey AJ. 2000. Analysis of Groucho-histone interactions suggests mechanistic similarities between Groucho- and Tup1-mediated repression. Nucleic Acids Res. 28:4189–4196.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644–652.

Grbavec D, Lo R, Liu Y, Greenfield A, Stifani S. 1999. Groucho/transducin-like enhancer of split (TLE) family members interact with the yeast transcriptional co-repressor SSN6 and mammalian SSN6-related proteins: implications for evolutionary conservation of transcription repression mechanisms. Biochem. J. 337 (Pt 1):13–17.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New

algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.

He D, Fiz-Palacios O, Fu C-J, Fehling J, Tsai C-C, Baldauf SL. 2014. An alternative root for the eukaryote tree of life. Curr. Biol. CB 24:465–470.

Hudry B, Thomas-Chollier M, Volovik Y, Duffraisse M, Dard A, Frank D, Technau U, Merabet S. 2014. Molecular insights into the origin of the Hox-TALE patterning system. eLife 3:e01939.

Jennings BH, Ish-Horowicz D. 2008. The Groucho/TLE/Grg family of transcriptional co-repressors. Genome Biol. 9:205.

Jennings BH, Pickles LM, Wainwright SM, Roe SM, Pearl LH, Ish-Horowicz D. 2006. Molecular recognition of transcriptional repressor motifs by the WD domain of the Groucho/TLE corepressor. Mol. Cell 22:645–655.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

Ke J, Ma H, Gu X, Thelen A, Brunzelle JS, Li J, Xu HE, Melcher K. 2015. Structural basis for recognition of diverse transcriptional repressors by the TOPLESS family of corepressors. Sci. Adv. 1:e1500107.

Komachi K, Redd MJ, Johnson AD. 1994. The WD repeats of Tup1 interact with the homeo domain protein alpha 2. Genes Dev. 8:2857–2867.

Kwon E-JG, Laderoute A, Chatfield-Reed K, Vachon L, Karagiannis J, Chua G. 2012. Deciphering the transcriptional-regulatory network of flocculation in Schizosaccharomyces pombe. PLoS Genet. 8:e1003104.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinforma. Oxf. Engl. 25:2286–2288.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25:1307–1320.

Liu Z, Karmarkar V. 2008. Groucho/Tup1 family co-repressors in plant development. Trends Plant Sci. 13:137–144.

Matsumura H, Kusaka N, Nakamura T, Tanaka N, Sagegami K, Uegaki K, Inoue T, Mukai Y. 2012. Crystal structure of the N-terminal domain of the yeast general corepressor Tup1p and its functional implications. J. Biol. Chem. 287:26528–26538.

McDowall MD, Harris MA, Lock A, Rutherford K, Staines DM, Bähler J, Kersey PJ, Oliver SG, Wood V. 2015. PomBase 2015: updates to the fission yeast database. Nucleic Acids Res. 43:D656–D661.

de Mendoza A, Sebé-Pedrós A, Šestak MS, Matejcic M, Torruella G, Domazet-Loso T, Ruiz-Trillo I. 2013. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. Proc. Natl. Acad. Sci. U. S. A. 110:E4858–E4866.

Nakagawa S, Gisselbrecht SS, Rogers JM, Hartl DL, Bulyk ML. 2013. DNA-binding specificity changes in the evolution of forkhead transcription factors. Proc. Natl. Acad. Sci. U. S. A. 110:12349–12354.

Oughtred R, Chatr-Aryamontri A, Breitkreutz B-J, Chang CS, Rust JM, Theesfeld CL, Heinicke S, Breitkreutz A, Chen D, Hirschman J, et al. 2016. BioGRID: A Resource for Studying Biological Interactions in Yeast. Cold Spring Harb. Protoc. 2016:pdb.top080754.

Pickles LM, Roe SM, Hemingway EJ, Stifani S, Pearl LH. 2002. Crystal structure of the C-terminal WD40 repeat domain of the human Groucho/TLE1 transcriptional corepressor. Struct. Lond. Engl. 1993 10:751–761.

Pisani D, Pett W, Dohrmann M, Feuda R, Rota-Stabelli O, Philippe H, Lartillot N, Wörheide G. 2015. Genomic data do not support comb jellies as the sister group to all other animals. Proc. Natl. Acad. Sci. U. S. A. 112:15402–15407.

Ruiz-Trillo I, Burger G, Holland PWH, King N, Lang BF, Roger AJ, Gray MW. 2007. The origins of multicellularity: a multi-taxon genome initiative. Trends Genet. TIG 23:113–118.

Russell RB, Barton GJ. 1992. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins 14:309–323.

Sebé-Pedrós A, Ariza-Cosano A, Weirauch MT, Leininger S, Yang A, Torruella G, Adamski M, Adamska M, Hughes TR, Gómez-Skarmeta JL, et al. 2013. Early evolution of the T-box transcription factor family. Proc. Natl. Acad. Sci. U. S. A. 110:16050–16055.

Sebé-Pedrós 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. 28:1241–1254.

Sprague ER, Redd MJ, Johnson AD, Wolberger C. 2000. Structure of the C-terminal domain of Tup1, a corepressor of transcription in yeast. EMBO J. 19:3016–3027.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, et al. 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 43:D447–D452.

Taylor WR. 1997. Residual colours: a proposal for aminochromography. Protein Eng. 10:743–746.

Torruella G, de Mendoza A, Grau-Bové X, Antó M, Chaplin MA, del Campo J, Eme L, Pérez-Cordón G, Whipps CM, Nichols KM, et al. 2015. Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. Curr. Biol. CB 25:2404–

2410.

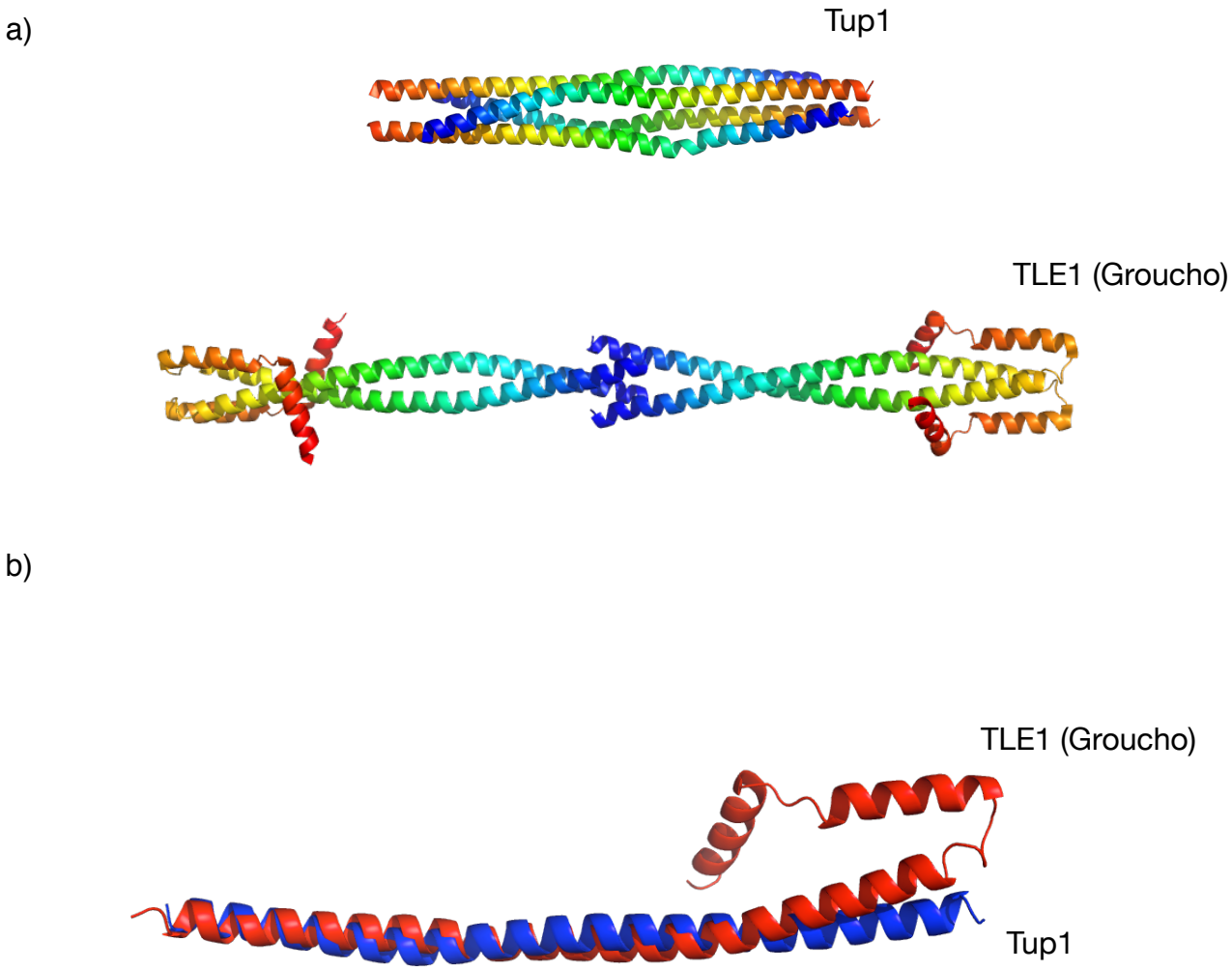
Tzamarias D, Struhl K. 1994. Functional dissection of the yeast Cyc8-Tup1 transcriptional co-repressor complex. Nature 369:758–761.

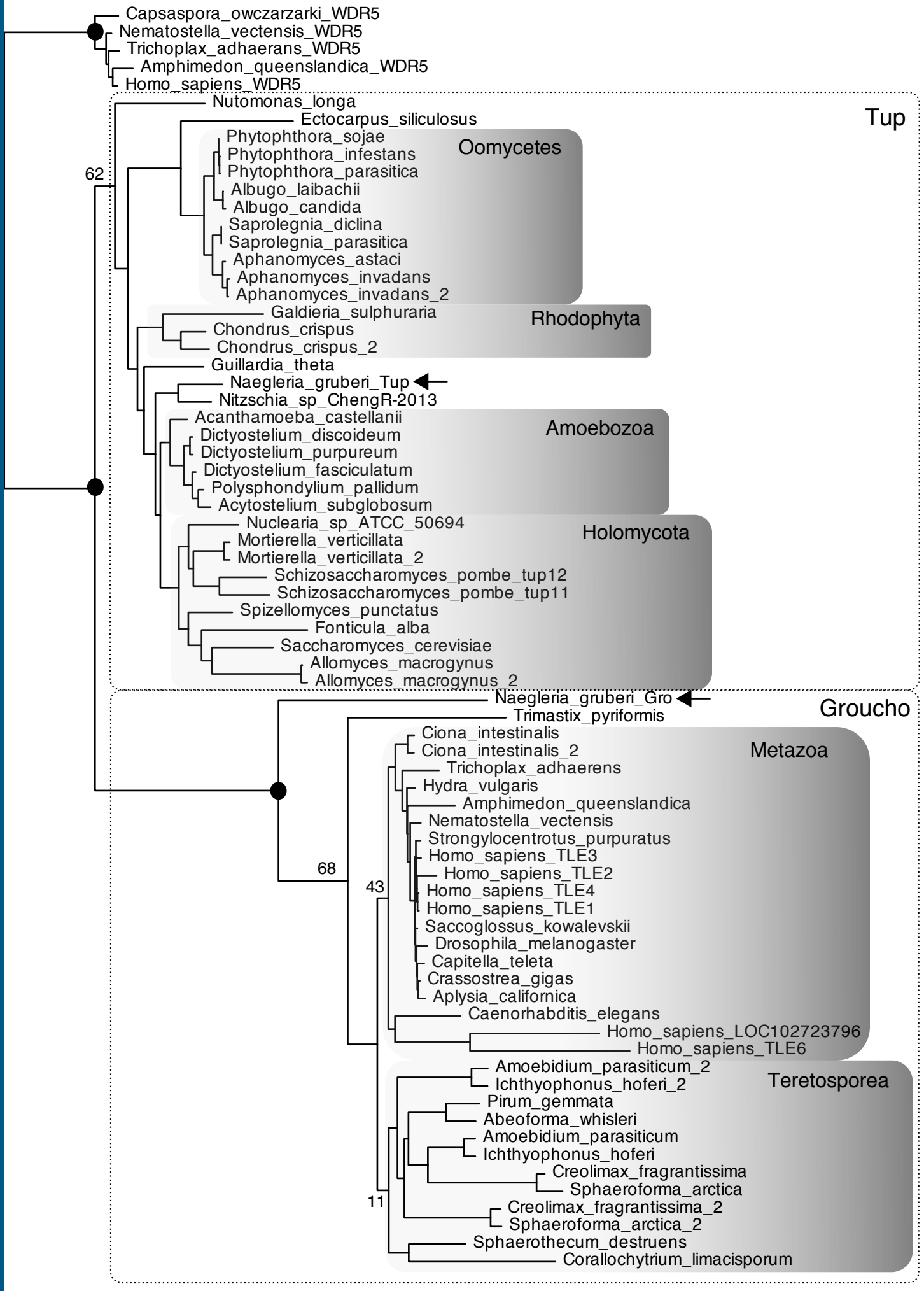
Wagner GP. 2007. The developmental genetics of homology. Nat. Rev. Genet. 8:473–479.

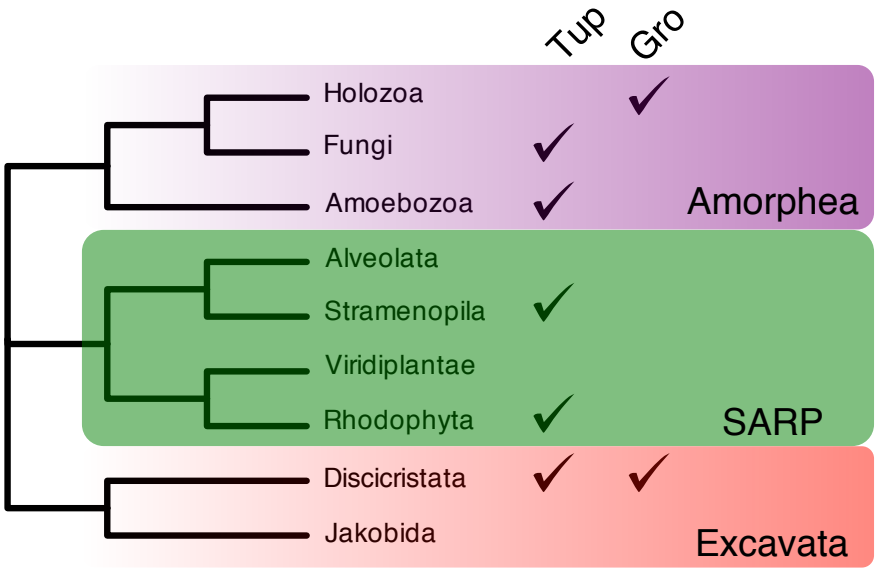
Whelan NV, Kocot KM, Moroz LL, Halanych KM. 2015. Error, signal, and the placement of Ctenophora sister to all other animals. Proc. Natl. Acad. Sci. U. S. A. 112:5773–5778.

Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA. 2008. DBD--taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res. 36:D88–D92.

igure 1







Genome Biology and Evolution
SMBE

igure 3

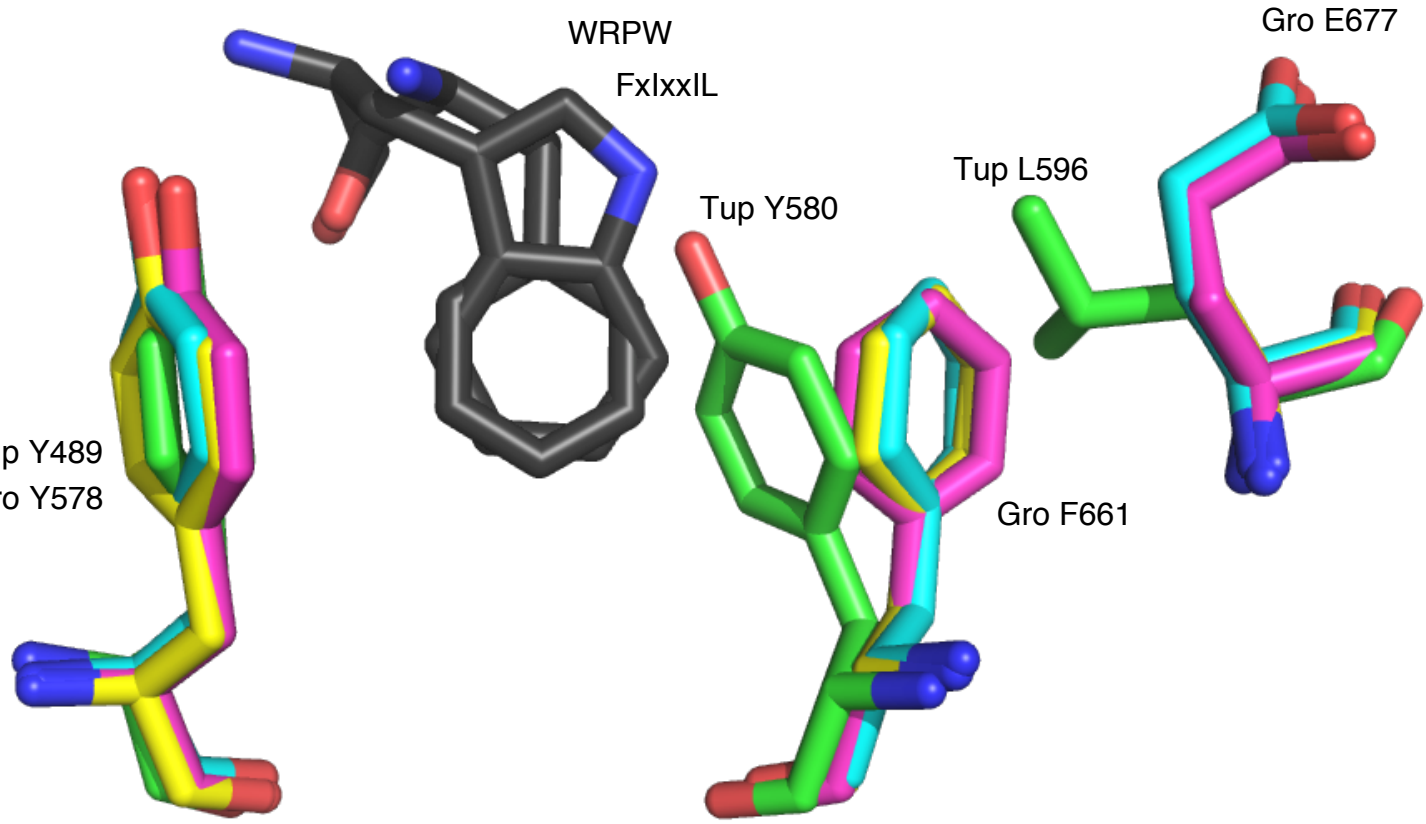


Figure 4

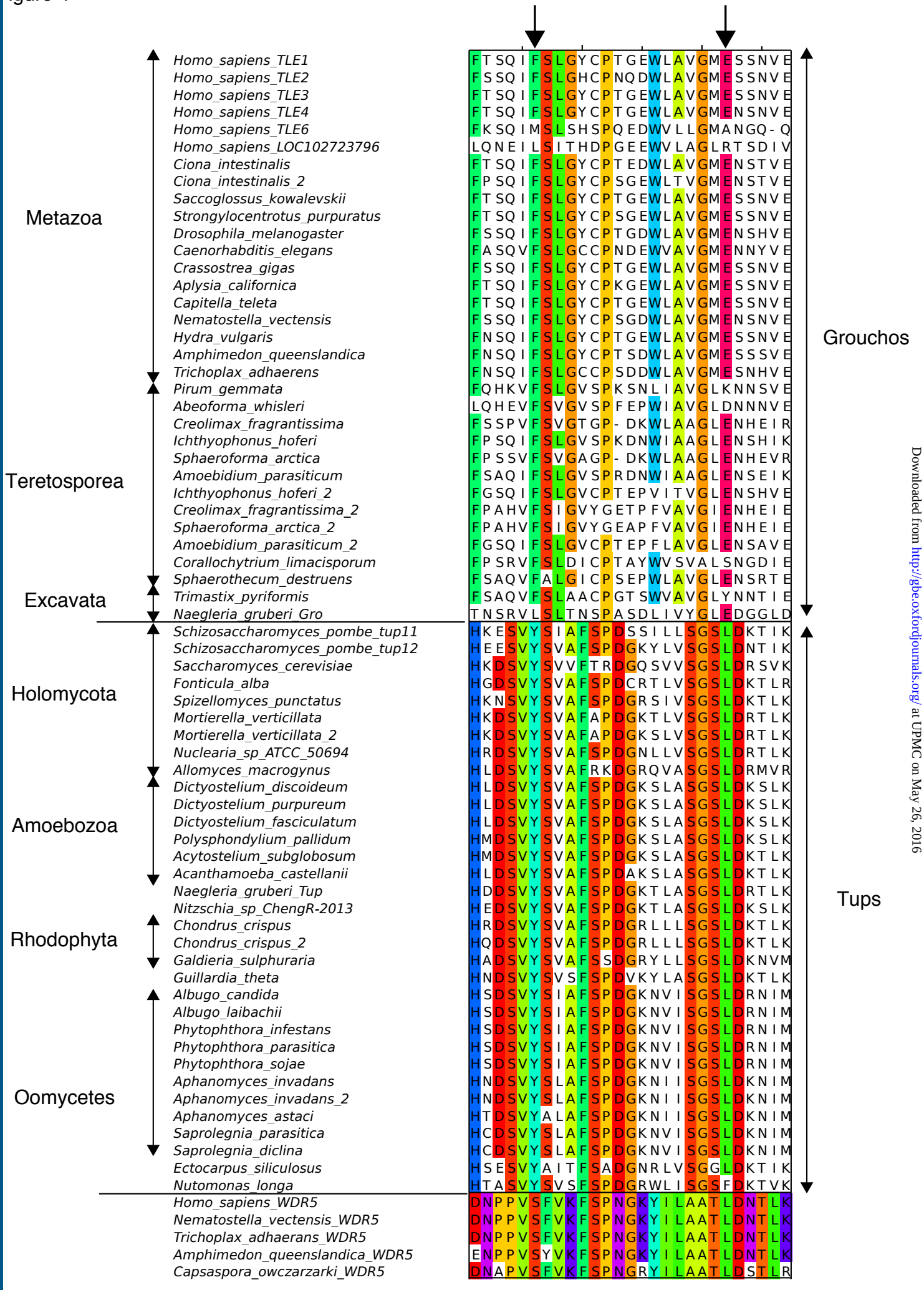


Figure 5