BMC
Evolutionary Biology

RESEARCH ARTICLE   **Open Access**

The evolutionary diversification of LSF and Grainyhead transcription factors preceded the radiation of basal animal lineages

Nikki Traylor-Knowles[1], Ulla Hansen[1], Timothy Q Dubuc[2], Mark Q Martindale[2], Les Kaufman[1] and John R Finnerty*[1]

Abstract

Background: The transcription factors of the LSF/Grainyhead (GRH) family are characterized by the possession of a distinctive DNA-binding domain that bears no clear relationship to other known DNA-binding domains, with the possible exception of the p53 core domain. In triploblastic animals, the LSF and GRH subfamilies have diverged extensively with respect to their biological roles, general expression patterns, and mechanism of DNA binding. For example, *Grainyhead* (GRH) homologs are expressed primarily in the epidermis, and they appear to play an ancient role in maintaining the epidermal barrier. By contrast, LSF homologs are more widely expressed, and they regulate general cellular functions such as cell cycle progression and survival in addition to cell-lineage specific gene expression.

Results: To illuminate the early evolution of this family and reconstruct the functional divergence of LSF and GRH, we compared homologs from 18 phylogenetically diverse taxa, including four basal animals (*Nematostella vectensis*, *Vallicula multiformis*, *Trichoplax adhaerens*, and *Amphimedon queenslandica*), a choanoflagellate (*Monosiga brevicollis*) and several fungi. Phylogenetic and bioinformatic analyses of these sequences indicate that (1) the LSF/GRH gene family originated prior to the animal-fungal divergence, and (2) the functional diversification of the LSF and GRH subfamilies occurred prior to the divergence between sponges and eumetazoans. Aspects of the domain architecture of LSF/GRH proteins are well conserved between fungi, choanoflagellates, and metazoans, though within the Metazoa, the LSF and GRH families are clearly distinct. We failed to identify a convincing LSF/GRH homolog in the sequenced genomes of the algae *Volvox carteri* and *Chlamydomonas reinhardtii* or the amoebozoan *Dictyostelium purpureum*. Interestingly, the ancestral GRH locus has become split into two separate loci in the sea anemone *Nematostella*, with one locus encoding a DNA binding domain and the other locus encoding the dimerization domain.

Conclusions: In metazoans, LSF and GRH proteins play a number of roles that are essential to achieving and maintaining multicellularity. It is now clear that this protein family already existed in the unicellular ancestor of animals, choanoflagellates, and fungi. However, the diversification of distinct LSF and GRH subfamilies appears to be a metazoan invention. Given the conserved role of GRH in maintaining epithelial integrity in vertebrates, insects, and nematodes, it is noteworthy that the evolutionary origin of Grh appears roughly coincident with the evolutionary origin of the epithelium.

Background

In triploblastic animals, LSF/Grainyhead (GRH) transcription factors perform a number of functions essential to both development and homeostasis. They are involved in regulation of the cell cycle, cell division, and cellular differentiation in a range of developmental and non-developmental contexts [1-14].

The LSF/Grainyhead family is split into the LSF/CP2 subfamily and the Grainyhead (GRH) subfamily, which can be distinguished by their distinctive oligomerization domains and differences in their oligomerization behavior [10,15,16]. GRH binds to DNA as a dimer, whereas LSF binds as a tetramer [17,18]. The DNA binding regions in both protein subfamilies show a large amount of conservation [17,18], but each has distinct transcrip-

* Correspondence: jrf3@bu.edu
[1] Department of Biology, Boston University, 5 Cummington Street, Boston, Massachusetts 02215, USA
Full list of author information is available at the end of the article

tional targets. GRH binds to the DNA sequence: (A/T)C(A/C/T)(G/T)GTT(C/G/T), whereas LSF binds to a direct repeat with the consensus sequence of N(C/G/T)N(C/G/T)(C/G)N(C/T)N(C/G/T)NN(C/G/T)(C/G/T)N(A/C/G)N [15,16,18,19]. LSF proteins can also be distinguished from GRH by the possession of a sterile alpha motif (SAM) [20]. Members of both LSF and GRH subfamilies were previously identified in vertebrates, arthropods, and nematodes, so the origin of the family and the diversification into subfamilies is known to predate the evolutionary split between protostomes and deuterostomes [15]. Recently, a common origin for the LSF/GRH family and the p53 family has been proposed based on similarities in the folding of their DNA-binding domains [20].

The differences in the molecular functions of LSF and GRH are accompanied by important differences in their biological roles. In both vertebrates and protostome invertebrates, GRH proteins are involved in the development and maintenance of epithelial integrity [21]. For example, in mice, *grh* is required during embryogenesis where it is expressed exclusively in the developing ectodermal epithelium [22]. Furthermore, embryonic mice lacking *grhl-3* exhibit insufficient wound repair and abnormal skin barrier formation leading to excessive postnatal water loss. The water loss is associated with reduced expression of the gene encoding *TGase1*, an enzyme that promotes cross-linking of parts of the stratum corneum, thus preventing the movement of water and solutes [22]. Likewise, in *Xenopus*, a *Grh*-like gene (*Xgrh1*) has been implicated in the development of the epidermis [13]. One of its primary targets is epidermal keratin. In morpholino studies, knockdown of *Xgrhl* led to loss of surface structures and pigmentation as well as neck and eye defects associated with epidermal instability [13]. In *Drosophila*, GRH plays a critical role in epithelial integrity that is analogous to and perhaps homologous with the role played in vertebrates--GRH maintains the tension of the *Drosophila* cuticle, and it induces cuticle development and cuticle repair following injury [23,24]. Similarly, the CeGrh1 protein of *C. elegans* appears to be required for proper cuticle formation during development, as its knockdown leads to soft, malformed cuticles and embryonic lethality [15].

In addition to its widely conserved role in maintaining epidermal integrity, *Grh* is also involved in the specification and development of the CNS in both *Drosophila* and mice [9,25]. Additionally, in mice, *Grh* mutants exhibit defects of the salivary and kidney ducts and eyelid closure [26-28], and in humans, a single nucleotide polymorphism found in GRHL2 is associated with age-related hearing impairment [29].

The biological roles of LSF are diverse and they have clearly diverged from those of GRH, at least in mammals, where the function of LSF has been well characterized. LSF is ubiquitously expressed [30]. It appears to play a role in liver function, eye development, erythropoesis, neural and immune function, regulation of the cell cycle progression, and cell survival [8,16,31-42].

When the ancestors of the LSF and GRH subfamilies first originated via a gene duplication event from their common ancestor, they would presumably have had identical or largely overlapping functions. However, at least in extant mammals, LSF and GRH have diverged extensively with respect to their biological roles. The basis for this functional diversification is not clear. The common ancestral functional repertoire of LSF and GRH may have become "subfunctionalized" in the two descendants [43]. Alternatively or in concert, LSF and GRH may have independently acquired novel functions since their split from a common ancestral gene ("neofunctionalization") [43,44].

If we wish to reconstruct the initial functional diversification of LSF and GRH, it is necessary to identify the ancestor in which the original gene duplication occurred. This may permit us to infer the functional repertoire of the LSF/GRH ancestor, and to compare this ancestral condition with the function of LSF and GRH in a phylogenetic progression of extant taxa. By comparing vertebrates, arthropods, and nematodes, Venkatesan and coworkers previously showed that the origin of distinct LSF and GRH subfamilies predated the diversification of triploblasts into distinct protostome and deuterostome lineages [15]. With the recent availability of sequenced genomes from several basal metazoans, a choanoflagellate, and more distantly related fungal outgroups, we can track the evolution of the LSF/GRH family into the much more distant past. In this study, we report the identification of LSF/GRH family members in 24 previously unreported species. Through a combination of genome prospecting and phylogenetic analysis, we show that the original gene duplication that produced the LSF and GRH subfamilies occurred prior to the evolutionary radiation of basal animal lineages (e.g., Bilateria, Cnidaria, Ctenophora, Porifera, and Placozoa). Interestingly, the GRH protein of the sea anemone *Nematostella vectensis*, a representative cnidarian, appears to have split into two distinct loci. We also identify six protein motifs that are widely shared between the LSF and GRH subfamilies of metazoans, all of which can be traced to the common ancestor of metazoans and fungi. In addition, there is a single motif that appears unique to the LSF subfamily.

Results
Identification of putative LSF/GRH homologs in animals, choanoflagellates, and fungi
BLAST searches identified putative LSF and GRH orthologs in eleven non-mammalian animals (Table 1)

Table 1: LSF/GRH sequences identified by BLAST searches.

Species		Protein	3E-value	Accession
Choanoflagellata				
[1,2]*Monosiga brevicollis*	Mob	LSF-like	3e-10	jgi| Monbr1| 29664| fgenesh2_pg.scaffold_37000034
Fungi/Ascomycota				
[1,2]*Aspergillus niger*	Asn	LSF-like1	8e-18	jgi| Aspni1| 174592| e_gw1.2.1262.1
[1,2]*Aspergillus niger*	Asn	LSF-like2	1e-6	jgi| Aspni1| 206941| estExt_GeneWisePlus.C_21219
[2]*Mycosphaerella fijiensis*	Myf	LSF-like	7e-39	jgi| Mycfi1| 87221| estExt_fgenesh1_pg.C_60133
[2]*Mycosphaerella graminicola*	Myg	LSF-like	9e-8	jgi| Mycgr3| 76636| estExt_Genewise1Plus.C_chr_110062
[2]*Trichoderma virens*	Trv	LSF-like	2e-5	jgi| Trive1| 37466| e_gw1.6.78.1
Fungi/Basidiomycota				
Phanerochaete chrysosporium	Phc		.058	jgi| Phchr1| 428| fgenesh1_pg.C_scaffold_1000428
Sporobolomyces roseus	Spr		.014	jgi| Sporo1| 4513| gw1.5.455.1
Fungi/Zygomycota				
[1,2]*Phycomyces blakesleeanus*	Phb	LSF-like1	2e-32	jgi| Phybl1| 75515| estExt_fgeneshPB_pg.C_10328
[1,2]*Phycomyces blakesleeanus*	Phb	LSF-like2	8e-14	jgi| Phybl1| 63864| fgeneshPB_pg.8__95
Metazoa/Annelida				
[1,2]*Capitella* species	Cap	GRH	7e-41	jgi| Capca1| 198092| fgenesh1_pg.C_scaffold_3000030
[1,2]*Capitella* species	Cap	LSF	e-119	>jgi| Capca1| 222821| estExt_fgenesh1_pg.C_600043
Metazoa/Arthropoda				
Anopheles gambiae	Ang	GRH		ref| XP_308698.4|
Anopheles gambiae	Ang	LSF		ref| XP_315571.4|
[1,2]*Daphnia pulex*	Dap	GRH	8e-43	jgi| Dappu1| 64332| e_gw1.171.21.1

Table 1: LSF/GRH sequences identified by BLAST searches. (Continued)

[1,2]*Daphnia pulex*	Dap	LSF	e-127	jgi| Dappu1| 192185| estExt_Genewise1Plus.C_50175
[1,2]*Drosophilia melanogaster*	Drm	GRH	**	ref| NP_476842.2|
[1,2]*Drosophilia melanogaster*	Drm	Gemini	**	ref| NP_610556.1|
Metazoa/Chordata				
[1,2]*Branchiostoma floridae*	Brf	GRH	e-129	jgi| Brafl1| 106909| fgenesh2_pg.scaffold_554000005
[1,2]*Branchiostoma floridae*	Brf	LSF	e-146	jgi| Brafl1| 129980| estExt_fgenesh2_pg.C_4020027
[1,2]*Ciona intestinalis*	Cii	GRH	e-107	jgi| Cioin2| 262638| gw1.01q.820.1
[1,2]*Ciona intestinalis*	Cii	LSF	e-141	jgi| Cioin2| 262956| gw1.03q.546.1
Fugu rubripes	Fur	GRH	0	jgi| Takru4| 570042| e_gw2.138.33.1
Fugu rubripes	Fur	LSF	0	jgi| Takru4| 710919| fgh5_pm.C_scaffold_60000044
Homo sapiens	Hos	CP2		ref| NP_005644.2|
Homo sapiens	Hos	CP2like3		ref| NP_079191.2|
Homo sapiens	Hos	GRHL1		gb| AAH67519.1|
[1,2]*Homo sapiens*	Hos	GRHL2		gb| AAH69633.1|
[1,2]*Homo sapiens*	Hos	LBP1a		gb| AAH47235.1|
Homo sapiens	Hos	LBP9		ref| NP_055368.1|
Homo sapiens	Hos	LBP32		gb| AAF32276.1| AF198489_1
Homo sapiens	Hos	SOM1		ref| NP_067003.2|
Mus musculus	Mum	BOM		AAM22619
Mus musculus	Mum	CP2		ref| NP_258437.1|
Mus musculus	Mum	CRTR-1		NP_076244
Mus musculus	Mum	NF2d9		AAC52244

Table 1: LSF/GRH sequences identified by BLAST searches. (Continued)

Metazoa/Cnidaria				
[1,2]*Nematostella vectensis*	Nev	GRH1	2e-49	jgi| Nemve1| 95157| e_gw.38.94.1
[1,2]*Nematostella vectensis*	Nev	LSF	3e-88	jgi| Nemve1| 189242| estExt_GenewiseH_1.C_1380090
Metazoa/Ctenophora				
[1,2]* *Vallicula multiformis*	Mnl	GRH		EST
Metazoa/Mollusca				
[1,2]*Lottia gigantea*	Log	GRH	3e-52	jgi| Lotgi1| 157385| fgenesh2_pg.C_sca_14000008
[1,2]*Lottia gigantea*	Log	LSF	e-122	jgi| Lotgi1| 166840| fgenesh2_pg.C_sca_61000088
Metazoa/Placozoa				
[2]*Trichoplax adhaerens*	Tra	GRH-like	4e-20	jgi| Triad1| 25702| e_gw1.5.1214.1
Metazoa/Porifera				
[1,2]*Amphimedon queenslandica*	Amq	GRH	7e-55	Not Available
[1,2]*Amphimedon queenslandica*	Amq	LSF	4e-49	Not Available
Plantae				
Selaginella moellendorffii	Sem		.019	jgi| Selmo1| 411196| fgenesh2_pg.C_scaffold_14000249
Chlamydomonas	Chl		.042	jgi| Chlre3| 194996| estExt_fgenesh2_pg.C_840034

[1]Sequence included in MEME analysis; [2]Sequence included in phylogenetic analysis; [3]Expectation value of match to human query (either GRHL2[4] or LBP1a[5]) observed during BLAST searches. Asterisks identify newly acquired sequences.

including three chordates (*Branchiostoma floridae, Ciona intestinalis, Fugu rubripes*), three arthropods (*Anopheles gambiae, Daphnia pulex*, and *Drosophila melanogaster*), an annelid (*Capitella* spp.), a mollusc (*Lottia gigantea*), a cnidarian (*Nematostella vectensis*), and a sponge (*Amphimedon queenslandica*). We also identified a strong match to human GRHL2 in the ctenophore (*Vallicula multiformis)* and the placozoan *(Trichoplax adhaerens)*. We were not able to identify putative LSF orthologs in either the ctenophore or the placozoan.

The cnidarian *Nematostella* is unusual in that its GRH homolog appears to be split between two loci. Nev-GRH1, which had been reported previously [45], emerged as a strong match to the entire human GRHL2 protein. However, as Nev-GRH1 appears to be truncated relative to the human protein, we conducted a separate BLAST search using only the carboxy terminal region of the human protein as a query sequence. Nev-GRH2, which was identified in this second BLAST search, is a strong match to the carboxy terminal portion of the human GRHL2 protein.

Among choanoflagellates and fungi, we were also able to identify members of the LSF/GRH family, but clear evidence for distinct LSF and GRH family members was less compelling. The sequenced genome of the choanoflagellate *Monosiga brevicollis* appears to encode only a single LSF/GRH related gene. Likewise, we could identify only a single LSF/GRH homolog in the genomes of four

fungi (*Mycosphaerella fijiensis*, *Mycosphaerella gramini-cola*, *Phanerochaete chrysosporium* and *Trichoderma virens*). We did identify two LSF/GRH-related sequences in *Aspergillus niger* (phylum Ascomycota) and *Phycomyces brevicollis* (phylum Zygomycota), but in both cases, the two sequences appeared most similar to each other, suggesting that they might have resulted from lineage-specific gene duplications.

In members of the kingdom Plantae, evidence of LSF/GRH family members was far more tenuous. Using a less stringent E-value cut off (e-1), we identified two proteins with limited resemblance to LSF/GRH in *Selaginella moellendorffii*, a lycophyte. In addition, we also identified a protein with similarity to LSF in the green algae *Chlamydomonas reinhardtii*. Using this *Chlamydomonas* sequence to query the genome of *Volvox carteri*, we identified the corresponding gene in this alga.

Protein motif identification
MEME analysis (Additional file 1) reveals extensive conservation in motif architecture within and between the LSF and GRH proteins of animals; it also reveals extensive conservation between these animal proteins and the LSF/GRH-related proteins of the choanoflagellate and the fungi (Figure 1). Overall, the MEME analysis identified 19 motifs that exhibit significant conservation between two or more sequences (Figure 2). Six motifs (4, 5, 6, 9, 10 and 11) are almost universally conserved among animal, choanoflagellate, and fungal sequences. Several of these motifs either correspond to previously identified functional domains, or they reside within such domains. Motif 1 corresponds to the activation domain [3,16,46]. Motifs 4, 5, 6, 9, 10 and 11 reside within the DNA binding domain [18,20,47]. Motif 15 corresponds to the SAM domain, and motifs 18 and 19 correspond to the dimerization domain. Two adjacent motifs (13 and 15) are well conserved among LSF proteins. While motif 15 was also identified in the choanoflagellate protein, the co-occurrence of motifs 13 and 15 appears characteristic of the LSF subfamily, with the exception of the sponge LSF sequence that did not exhibit a significant match to motif 13.

The motif analysis reveals strong similarities between the pairs of sequences identified in each of the two fungal species. The two proteins from the ascomycote fungus, *Aspergillus*, are nearly identical to each with respect to motif architecture, and they can be distinguished from other sequences by the possession of motifs 14, 16, and 17. Likewise, the two sequences from the zygomycote fungus, *Phycomyces*, are most similar to each other with respect to the arrangement of conserved motifs.

The motif analysis also supports the conclusion that the GRH locus of the cnidarian *Nematostella* has experienced a split. Nev-GRH1 encompasses six conserved motifs (4, 5, 6, 9, 10, 11), and these motifs occupy the same relative positions as in the GRH proteins of fruit fly and sponge. Nev-GRH2 encompasses conserved motifs 18 and 19, which occupy the same relative position in most other metazoan GRH sequences.

Phylogenetic analysis
All phylogenetic analyses that we performed can be rooted so that the fungal sequences and the metazoan sequences form mutually exclusive monophyletic groups (Figure 3; Additional file 2). On the neighbor-joining tree (Figure 3), the metazoan clade can be further subdivided into putative LSF and GRH clades. Within the LSF clade, the triploblastic animals form a monophyletic group to the exclusion of two diploblastic animals (*Nematostella* and *Amphimedon*). Similarly, within the GRH clade, the triploblastic animals form a monophyletic group to the exclusion of four diploblastic animals (*Nematostella*, *Amphimedon*, *Trichoplax*, and *Vallicula*), implying that both LSF and GRH subfamilies had originated prior to the evolutionary split between diploblasts and triploblasts. The single *Trichoplax* sequence groups within the GRH clade. Though the bootstrap support for this grouping is low, along with the motif analysis, this suggests that the *Trichoplax* sequence may be a true GRH ortholog (implying that the LSF ortholog of *Trichoplax* has either been lost or we failed to find it). The single *Monosiga* sequence appears at the base of the LSF clade, suggesting that it might be a true LSF ortholog (which would imply that LSF and GRH diverged before the split between animals and choanoflagellates). The single *Vallicula* sequence groups with GRH sequences of other diploblastic animals.

The maximum-likelihood analysis (Additional file 2) supports most of the major divisions that appear on the neighbor-joining tree. The animal sequences and fungal sequences comprise discrete subtrees. The LSF sequences form a putative clade, and within this clade, the LSF sequences of triploblasts cluster together to the exclusion of LSF sequences from diploblasts. Likewise, the GRH sequences of triploblasts also group together. However, the putative GRH sequences of diploblastic animals do not form a monophyletic group with the GRH sequences of triploblasts as they do on the neighbor-joining tree. Instead, the sponge and ctenophore sequences appear more closely related to the LSF clade, while the precise position of the anemone and placozoan GRH sequences is not resolved.

On both the neighbor-joining tree and the maximum-likelihood tree, bootstrap support for individual nodes is generally low because the analyses are based on a small number of highly conserved residues. However, both phylogenies are consistent with divisions between animal and fungal sequences and between LSF and GRH sequences, the same divisions that are implied by the motif analysis.

Nev-GRH1 and Nev-GRH2

The sea anemone, *Nematostella vectensis*, is unique in that the GRH locus has been split in two, with Nev-Grh1 encoding primarily the DNA-binding domain and Nev-Grh2 encoding primarily the dimerization domain. In the current draft assembly of the genome, Nev-Grh1 maps to scaffold 2, and Nev-Grh1 maps to scaffold 38 (Joint Genome Institute, *Nematostella vectensis* v1.0; Figure 4). Nev-Grh1 is flanked by a QRSL1 like gene and a B9D1-like gene. Nev-Grh2 is flanked by an arylsulfatase-like gene and an opsin-like gene. Even if these two scaffolds reside on the same chromosome, based on the location of each gene within its respective scaffold, the two loci must be separated by no less than 580 kilobases of intervening

sequence. Both of the Grh loci are represented by multiple ESTs (NevGRH1, EST cluster: 2655293_3; NevGRH2, EST cluster: 2664076_1), and none of the individual ESTs overlap (thus, there is no evidence for trans-splicing).

Potential homologs in plants?

Given that the origin of the LSF/GRH family predates the divergence of animals and fungi, we searched for LSF and GRH homologs in amoebozoans and plants to see if this gene family might predate the origin of opisthokonts. Plant genomes and amoebozoan genomes do not appear to encode any proteins with extensive similarity to the LSF/GRH proteins of animals and fungi. In tblastn searches of assembled genomes at the JGI Genome Portal
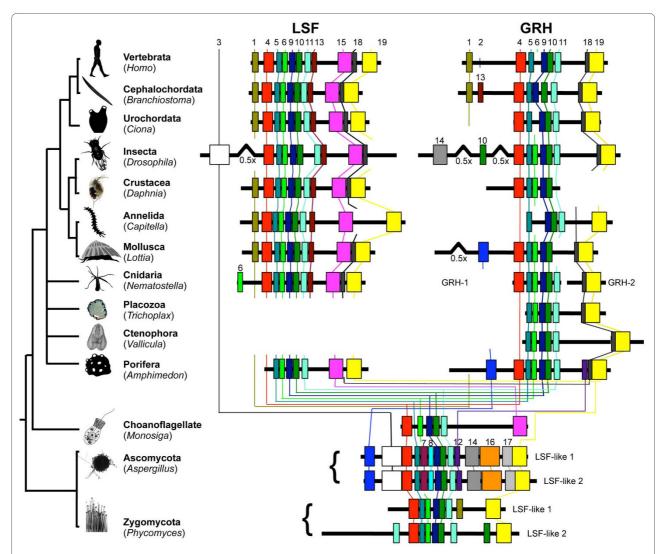

Figure 1 Motif architecture of LSF and GRH proteins from 10 metazoan taxa, a choanoflagellate, and two fungi. Conserved motifs were identified using MEME, as described in the methods. Motifs (colored boxes) and inter-motif regions (thick black lines) were drawn to scale except for certain lengthy inter-motif regions, which were truncated by 50% (0.5×). Thin colored lines highlight motif conservation between proteins. The relative relationships among taxa depicted here reflect a general consensus among molecular phylogenetic analyses [62-68], although there continues to be controversy surrounding key elements of the phylogeny including the placement of ctenophores [69] and the monophyly of the triploblasts [70-72].
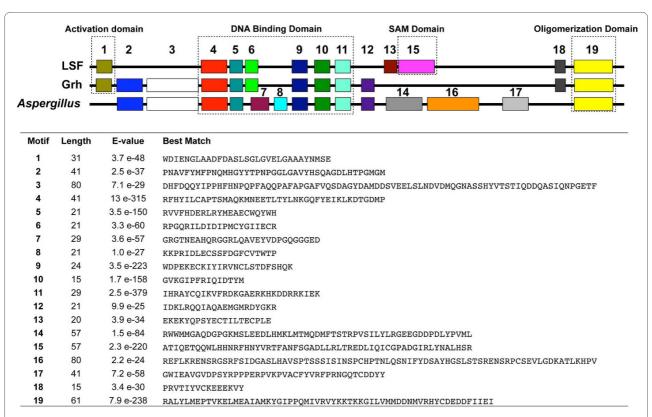
Motif	Length	E-value	Best Match
1	31	3.7 e-48	WDIENGLAADFDASLSGLGVELGAAAYNMSE
2	41	2.5 e-37	PNAVFYMFPNQMHGYYTPNPGGLGAVYHSQAGDLHTPGMGM
3	80	7.1 e-29	DHFDQQYIPPHFHNPQPFAQQPAFAPGAFVQSDAGYDAMDDSVEELSLNDVDMQGNASSHYVTSTIQDDQASIQNPGETF
4	41	13 e-315	RFHYILCAPTSMAQKMNEETLTYLNKGQFYEIKLKDTGDMP
5	21	3.5 e-150	RVVFHDERLRYMEAECWQYWH
6	21	3.3 e-60	RPGQRILDIDIPMCYGIIECR
7	29	3.6 e-57	GRGTNEAHQRGGRLQAVEYVDPGQGGGED
8	21	1.0 e-27	KKPRIDLECSSFDGFCVTWTP
9	24	3.5 e-223	WDPEKECKIYIRVNCLSTDFSHQK
10	15	1.7 e-158	GVKGIPFRIQIDTYM
11	29	2.5 e-379	IHRAYCQIKVFRDKGAERKHKDDRRKIEK
12	21	9.9 e-25	IDKLRQQIAQAEMGMRDYGKR
13	20	3.9 e-34	EKEKYQPSYECTILTECPLE
14	57	1.5 e-84	RWWMMGAQDGPGKMSLEEDLHMKLMTMQDMFTSTRPVSILYLRGEEGDDPDLYPVML
15	57	2.3 e-220	ATIQETQQWLHHNRFHNYVRTFANFSGADLLRLTREDLIQICGPADGIRLYNALHSR
16	80	2.2 e-24	REFLKRENSRGSRFSIDGASLHAVSPTSSSISINSPCHPTNLQSNIFYDSAYHGSLSTSRENSRPCSEVLGDKATLKHPV
17	41	7.2 e-58	GWIEAVGVDPSYRPPPERPVKPVACFYVRFPRNGQTCDDYY
18	15	3.4 e-30	PRVTIYVCKEEEKVY
19	61	7.9 e-238	RALYLMEPTVKELMEAIAMKYGIPPQMIVRVYKKTKKGILVMMDDNMVRHYCDEDDFIIEI

Figure 2 Top-scoring motif sequences and consensus motif architecture. Metazoan LSF proteins, metazoan GRH proteins, and fungal proteins can be distinguished by their consensus motif architectures (top). The consensus diagrams include all motifs that were found in at least one member of the respective group (Fig. 1). The best matches for each sequence motif identified by MEME are shown below the diagrams. The correspondence between these conserved motifs and known functional domains are indicated by boxes.

[48] using a permissive E value cut-off (e-1), the lycophyte, *Selaginella moellendorffii*, yielded a hit for GRH (E value 0.07), the alga *Chlamydomonas* yielded a hit for LSF (E value 0.07), and the amoebozoan *Dictyostelium purpureum* yielded a hit for LSF (E value 0.04; Additional file 3). When the top hit from *Selaginella* and *Dictyostelium* were BLASTed back against the human genome, the search yielded no significant hits.

Discussion

Evolutionary origins of the LSF/GRH family and subfamilies

Prior to the present study, members of the LSF/GRH family had been reported from a number of triploblastic animals but not from diploblastic animals, choanoflagellates, or fungi. We recovered clear LSF and GRH orthologs from two diploblastic animals (sea anemone and sponge) revealing that the evolutionary divergence between these two subfamilies must have predated the diploblast-triploblast split. Furthermore, fungi possess clear LSF/GRH homologs, although the fungal sequences cannot be assigned to either the LSF or GRH subfamilies. Therefore, while the family clearly originated prior to the metazoan-fungal divergence, the diversification of sub-

families occurred more recently, perhaps in an ancient animal lineage.

Nev-GRH1 and Nev-GRH2

The sea anemone, *Nematostella vectensis*, is the only species where the sequences encoding the ancestral GRH protein are known to be split between two loci. As sponges, ctenophores, and triploblasts exhibit full-length GRH proteins, this condition must be derived in the sea anemone. The splitting of the ancestral Grh locus in *Nematostella* must have profound consequences for the regulation and function of GRH. In other animals, GRH binds DNA targets as a dimer. However, in *Nematostella*, the DNA-binding domain and the oligomerization reside on different proteins. Perhaps Nev-GRH1 is able to interact with the DNA singly, or perhaps a partnership with Nev-GRH2 allows it to form the equivalent of a GRH-dimer on DNA, reminiscent of other GRH proteins. This latter possibility implies that Nev-GRH1 and Nev-GRH2 will be co-expressed in the same cells. This will need to be confirmed experimentally. Interestingly, a comparable split seems to have occurred in the NF-κB gene of this species, with distinct loci encoding different functional domains of the ancestral protein [49].
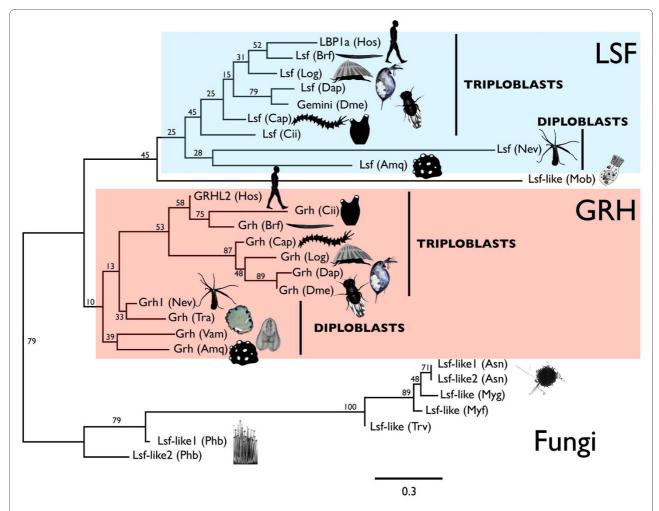
Figure 3 Phylogeny of LSF and GRH proteins. The tree shown is based on a neighbor-joining analysis of the amino acids in the gap free alignment. Numbers at nodes represent bootstrap support. The tree is drawn as though rooted between the metazoan sequences and the fungal sequences. Branch length is shown in terms of expected number of substitutions per residue (bar at lower right).

Identification of LSF/GRH Homologs in Fungi

Convincing matches to human LSF and/or GRH query sequences were found in the genomes of representative ascomycote, basiomycote and zygomycote fungi (Table 1). The phylum Basiomycota is the sister group to the phylum Ascomycota, with the Zygomycota being more distantly related [50], and the phylogenetic analysis we performed grouped the LSF-like proteins of the ascomycotes *Aspergillus*, *Mycosphaerella*, and *Trichoderma* to the exclusion of the LSF-like proteins from the zygomycote *Phycomyces*. In the MEME analysis, the two LSF/GRH proteins identified in the zygomycote *Phycomyces* were found to possess all of the conserved motifs that were identified within the DNA-binding domain of animals (motifs 4, 5, 6, 9, 10, and 11). The two LSF/GRH proteins of the ascomycote, *Aspergillus*, also possess motifs 4, 5, 9, 10, and 11, but in place of motif 6, these proteins share motifs 7 and 8, which are unique to this

fungus. All four fungal sequences subjected to the MEME analysis were found to contain motif 19, which corresponds to the dimerization domain. Given the strong conservation of motifs between fungi and animals in the DNA-binding and dimerization domains, we hypothesize that the molecular function of these fungal proteins will be very similar to their animal homologs, *i.e.*, they are transcription factors that will bind DNA targets, most likely as dimers (like GRH). However, if the novel fungal-specific motifs functionally replace the SAM domain, which is likely to represent the second protein-protein interaction domain in LSF subfamily members, they might instead bind DNA as tetramers (like LSF).

Insights into the hypothesized ancestral role of GRH from basal animals

Because GRH plays a comparable role in the maintenance and repair of the surface epithelium in mouse [22], clawed frog [13], fruit fly [23], and soil nematode [15], it
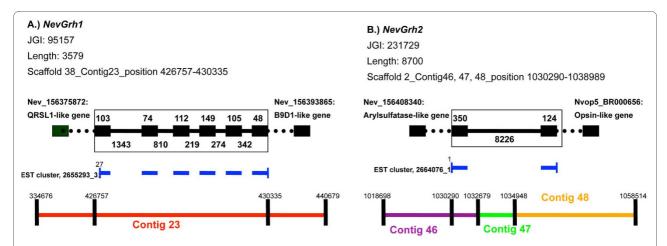
Figure 4 Mapping of *Nev-Grh1* and *Nev-Grh2* ESTs to separate loci. The *NevGrh1* and *NevGrh2* loci (enclosed in boxes) are flanked by distinct genes and are found on distinct, non-overlapping genomic scaffolds. Exons are indicated by black boxes, and introns are represented by solid black lines. Dotted lines represent the intergenic sequence leading to the nearest flanking genes. Flanking genes are named by species and NCBI number. The EST contigs for each locus are represented as thick blue lines beneath the exons that encode them. Figure is not to scale.

has been hypothesized that this role is homologous among triploblastic bilaterians [21,24]. Given that the shared possession of an epithelium is thought to be homologous across the Metazoa, it is possible that the functional evolution of GRH is connected to the origin and early evolution of the epithelium. The presence of an epithelial boundary is a plesiomorphic character of triploblastic animals, and therefore, we cannot explore the early evolution of animal epithelia using only triploblastic model systems. The identification of clear GRH homologs in cnidarians, ctenophores, and sponges, and the apparent absence of a true *Grh* gene in the choanoflagellate *Monosiga* suggests the origin of *Grh* may be coincident with the origin of the metazoan epithelium. Historically, sponges have been said to lack an epithelium, but more recently, the identification of a genuine basement membrane in homoscleromorph sponges removes this distinction between poriferans and other metazoans [51]. If the role of *Grh* in maintaining epithelial integrity dates to the origin of the epithelium, then *Grh* should be expressed in the epidermal epithelium of cnidarians, ctenophores and sponges. Furthermore, *Grh* should regulate proteins involved in epithelial differentiation and maintenance, although the exact targets of *Grh* transcriptional regulation may vary among basal animals as they vary among triploblasts. Additionally, we may expect that *Grh* will be upregulated in response to injury, while knockdown of *Grh* expression may undermine epithelial integrity and inhibit wound healing. All of these questions are amenable to testing in one or more basal model systems.

Conclusions
The LSF/GRH family had already originated by the time of the opisthokont ancestor, and the overall domain architecture of LSF/GRH proteins has been largely conserved in extant fungi, animals, and choanoflagellates. The LSF subfamily had diverged from the GRH family prior to the divergence of sponges, cnidarians, and triploblastic animals. Consistent differences in domain architecture distinguish the LSF and GRH proteins of both diploblastic and triploblastic animals, suggesting that the functional divergence between these proteins had been established prior to the evolutionary divergence between diploblasts and triploblasts. The sea anemone *Nematostella* appears unique in that the DNA-binding domain and the dimerization domain of the ancestral GRH protein are now encoded on two separate loci.

Methods
Identification of LSF/Grainyhead family members in outgroup taxa
The human proteins LSF [NP_005644.2] and GRHL2 [AAH69633.1] were used to query online genomic databases (Joint Genome Institute Eukaryotic Genomes and NCBI) for LSF-like and GRH-like proteins respectively using BlastP. The following search settings were employed: gap opening penalty = 11; gap extension penalty = 1. Potential homologs that matched one of the query sequences with an expectation score < e-1 were used to query the human genome (using BLASTp) to determine if their top human match was to the original human query sequence (LSF or GRHL2). Sequences were kept for phylogenetic and protein motif identification only if they met this criterion.

Protein motif identification
To identify conserved protein motifs, LSF/GRH proteins were evaluated using MEME (<u>M</u>ultiple <u>E</u>xpectation <u>M</u>aximization for <u>M</u>otif <u>E</u>licitation; http://meme.nbcr.net;

([52]; Additional file 1). LSF/GRH family members were chosen to represent ten metazoan phyla, the choanoflagellate *Monosiga brevicollis*, an ascomycote fungus (*Aspergillus*) and a zygomycote fungus (*Phycomyces*; Table 1). The following settings were used in the motif search: maximum number of motifs = 20; occurrences of a single motif = any number; minimum length of a motif = 3 amino acids; maximum length of a motif = 300.

Multisequence Alignment

Twenty-eight of the twenty-nine LSF/GRH protein sequences included in the MEME analysis were aligned in preparation for phylogenetic analysis (Additional file 4). The GRH2 protein of *Nematostella vectensis* was excluded from the alignment because it is substantially truncated relative to the full-length LSF and GRH proteins of other animals. Since motif 4 was identified near the amino terminal of all but three of the proteins, and motif 19 was identified near the carboxy terminal of all but one of the proteins (Figure 1), these motifs were used to bracket the alignment. To ensure that the motifs identified by MEME were maintained in register, the motifs themselves were manually aligned. Then, the regions between conserved motifs were multiply aligned using the Clustal alignment tool found in the application MEGA [53]. The following settings were specified: protein weight matrix = Gonnet, gap opening penalty = 10; gap extension penalty = 0.2. The resulting alignment spans 2045 characters. All positions in the alignment containing gaps were deleted to produce a gap-free alignment comprising 44 characters (Additional file 4).

Phylogenetic Analysis

Phylogenetic relationships among taxa were inferred from both the gap-free alignment and the full alignment using neighbor-joining [54] and maximum-likelihood [55]. All 44 residues in the gap-free alignment derive from motifs 9-11, which are part of the DNA-binding domain (Additional file 2). First, eighty alternate models of the amino acid substitution process were compared using the program ProtTest 1.3 [56]. The substitution process was optimized along with the tree topology and branch lengths. For both the full alignment and the gap-free alignment, the empirically determined JTT substitution matrix [57] outperformed other substitution matrices, and incorporating rate variation among sites significantly improved the model (the shape coefficient of the Gamma distribution, α = .837; the coefficient of rate variation among sites = $1/\alpha^{1/2}$ = 1.093). The JTT matrix with gamma-distributed rate variation among sites was specified in subsequent phylogenetic analyses.

For the neighbor joining analysis, pairwise distances between proteins were calculated using the Prodist program, and the tree topology was determined using the

Neighbor program, both in the Phylip package (v. 3.6; [58]). Maximum-likelihood analysis was performed using RAxML (v 7.0.3; [59]) as implemented on the CIPRES Portal (v. 2.0; [60]). In both the neighbor-joining analysis and the maximum-likelihood analysis, support for specific clades was assessed using the bootstrap [61]: 1,000 replicates of the bootstrap were performed for the neighbor-joining analysis, and 100 replicates were performed for the maximum-likelihood analysis.

Additional material

> **Additional file 1 Output from the motif discovery program MEME.**
>
> **Additional file 2 The tree shown is based on a maximum-likelihood analysis of the amino acids in the gap free alignment using the program RAxML**[73]. Numbers at nodes represent bootstrap support. The tree is drawn as though rooted between the metazoan sequences and the fungal sequences. Branch length is shown in terms of expected number of substitutions per residue (bar at lower left).
>
> **Additional file 3 The assembled genomes of** *Chlamydomonas reinhardtii, Dictyostelium purpureum,* **and** *Selaginella moellendorffii* (**housed at the JGI Genome Portal** [48]) **were queried with a human LSF sequence (>gi|21361278|ref|NP_005644.2| transcription factor CP2) and GRH sequence (>gi|46854865|gb|AAH69633.1| GRHL2 protein).** A permissive E value cut-off was specified (e-1). *Dictyostelium* and *Chlamydomonas* each returned a single hit for LSF, and *Selaginella* returned a single hit for GRH.
>
> **Additional file 4 All positions in the full alignment containing gaps were deleted to produce this gap-free alignment comprising 44 characters.**

Author Details

[1]Department of Biology, Boston University, 5 Cummington Street, Boston, Massachusetts 02215, USA and [2]Kewalo Marine Laboratory, Pacific Biomedical Research Center, University of Hawaii, 41 Ahui Street, Honolulu, HI 96813, USA

References

1. Auden A, Caddy J, Wilanowski T, Ting SB, Cunningham JM, Jane SM: Spatial and temporal expression of the Grainyhead-like transcription factor family during murine development. *Gene Expr Patterns* 2006, **6(8)**:964-970.
2. Bray SJ, Kafatos FC: Developmental function of Elf-1: an essential transcription factor during embryogenesis in Drosophila. *Genes Dev* 1991, **5(9)**:1672-1683.
3. Kudryavtseva EI, Sugihara TM, Wang N, Lasso RJ, Gudnason JF, Lipkin SM, Andersen B: Identification and characterization of Grainyhead-like

epithelial transactivator (GET-1), a novel mammalian Grainyhead-like factor. *Dev Dyn* 2003, **226**(4):604-617.

4. Hayashi Y, Yamagishi M, Nishimoto Y, Taguchi O, Matsukage A, Yamaguchi M: A binding site for the transcription factor Grainyhead/Nuclear transcription factor-1 contributes to regulation of the Drosophila proliferating cell nuclear antigen gene promoter. *J Biol Chem* 1999, **274**(49):35080-35088.

5. Lim LC, Swendeman SL, Sheffery M: Molecular cloning of the alpha-globin transcription factor CP2. *Mol Cell Biol* 1992, **12**(2):828-835.

6. Ramamurthy L, Barbour V, Tuckfield A, Clouston DR, Topham D, Cunningham JM, Jane SM: Targeted disruption of the CP2 gene, a member of the NTF family of transcription factors. *J Biol Chem* 2001, **276**(11):7836-7842.

7. Rodda S, Sharma S, Scherer M, Chapman G, Rathjen P: CRTR-1, a developmentally regulated transcriptional repressor related to the CP2 family of transcription factors. *J Biol Chem* 2001, **276**(5):3324-3332.

8. Sueyoshi T, Kobayashi R, Nishio K, Aida K, Moore R, Wada T, Handa H, Negishi M: A nuclear factor (NF2d9) that binds to the male-specific P450 (Cyp 2d-9) gene in mouse liver. *Mol Cell Biol* 1995, **15**(8):4158-4166.

9. Uv AE, Harrison EJ, Bray SJ: Tissue-specific splicing and functions of the Drosophila transcription factor Grainyhead. *Mol Cell Biol* 1997, **17**(11):6727-6735.

10. Wilanowski T, Tuckfield A, Cerruti L, O'Connell S, Saint R, Parekh V, Tao J, Cunningham JM, Jane SM: A highly conserved novel family of mammalian developmental transcription factors related to Drosophila grainyhead. *Mech Dev* 2002, **114**(1-2):37-50.

11. Wilanowski T, Caddy J, Ting SB, Hislop NR, Cerruti L, Auden A, Zhao LL, Asquith S, Ellis S, Sinclair R, et al.: Perturbed desmosomal cadherin expression in grainy head-like 1-null mice. *Embo J* 2008, **27**(6):886-897.

12. Ting SB, Wilanowski T, Auden A, Hall M, Voss AK, Thomas T, Parekh V, Cunningham JM, Jane SM: Inositol- and folate-resistant neural tube defects in mice lacking the epithelial-specific factor Grhl-3. *Nat Med* 2003, **9**(12):1513-1519.

13. Tao J, Kuliyev E, Wang X, Li X, Wilanowski T, Jane SM, Mead PE, Cunningham JM: BMP4-dependent expression of Xenopus Grainyhead-like 1 is essential for epidermal differentiation. *Development* 2005, **132**(5):1021-1034.

14. Parekh V, McEwen A, Barbour V, Takahashi Y, Rehg JE, Jane SM, Cunningham JM: Defective extraembryonic angiogenesis in mice lacking LBP-1a, a member of the grainyhead family of transcription factors. *Mol Cell Biol* 2004, **24**(16):7113-7129.

15. Venkatesan K, McManus HR, Mello CC, Smith TF, Hansen U: Functional conservation between members of an ancient duplicated transcription factor family, LSF/Grainyhead. *Nucleic Acids Res* 2003, **31**(15):4304-4316.

16. Veljkovic J, Hansen U: Lineage-specific and ubiquitous biological roles of the mammalian transcription factor LSF. *Gene* 2004, **343**(1):23-40.

17. Attardi LD, Tjian R: Drosophila tissue-specific transcription factor NTF-1 contains a novel isoleucine-rich activation motif. *Genes Dev* 1993, **7**(7B):1341-1353.

18. Shirra MK, Hansen U: LSF and NTF-1 share a conserved DNA recognition motif yet require different oligomerization states to form a stable protein-DNA complex. *J Biol Chem* 1998, **273**(30):19260-19268.

19. Frith MC, Hansen U, Weng Z: Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics* 2001, **17**(10):878-889.

20. Kokoszynska K, Ostrowski J, Rychlewski L, Wyrwicz LS: The fold recognition of CP2 transcription factors gives new insights into the function and evolution of tumor suppressor protein p53. *Cell Cycle* 2008, **7**(18):2907-2915.

21. Harden N: Cell biology. Of grainy heads and broken skins. *Science* 2005, **308**(5720):364-365.

22. Ting SB, Caddy J, Hislop N, Wilanowski T, Auden A, Zhao LL, Ellis S, Kaur P, Uchida Y, Holleran WM, et al.: A homolog of Drosophila grainy head is essential for epidermal integrity in mice. *Science* 2005, **308**(5720):411-413.

23. Mace KA, Pearson JC, McGinnis W: An epidermal barrier wound repair pathway in Drosophila is mediated by grainy head. *Science* 2005, **308**(5720):381-385.

24. Moussian B, Uv AE: An ancient control of epithelial barrier formation and wound healing. *Bioessays* 2005, **27**(10):987-990.

25. Cenci C, Gould AP: Drosophila Grainyhead specifies late programmes of neural proliferation by regulating the mitotic activity and Hox-

26. dependent apoptosis of neuroblasts. *Development* 2005, **132**(17):3835-3845.

26. Yu Z, Bhandari A, Mannik J, Pham T, Xu X, Andersen B: Grainyhead-like factor Get1/Grhl3 regulates formation of the epidermal leading edge during eyelid closure. *Dev Biol* 2008, **319**(1):56-67.

27. Yamaguchi Y, Yonemura S, Takada S: Grainyhead-related transcription factor is required for duct maturation in the salivary gland and the kidney of the mouse. *Development* 2006, **133**(23):4737-4748.

28. Gustavsson P, Greene ND, Lad D, Pauws E, de Castro SC, Stanier P, Copp AJ: Increased expression of Grainyhead-like-3 rescues spina bifida in a folate-resistant mouse model. *Hum Mol Genet* 2007, **16**(21):2640-2646.

29. Van Laer L, Van Eyken E, Fransen E, Huyghe JR, Topsakal V, Hendrickx JJ, Hannula S, Maki-Torkko E, Jensen M, Demeester K, et al.: The grainyhead like 2 gene (GRHL2), alias TFCP2L3, is associated with age-related hearing impairment. *Hum Mol Genet* 2008, **17**(2):159-169.

30. Swendeman SL, Spielholz C, Jenkins NA, Gilbert DJ, Copeland NG, Sheffery M: Characterization of the genomic structure, chromosomal location, promoter, and development expression of the alpha-globin transcription factor CP2. *J Biol Chem* 1994, **269**(15):11663-11671.

31. Bing Z, Huang JH, Liao WS: NFkappa B interacts with serum amyloid A3 enhancer factor to synergistically activate mouse serum amyloid A3 gene transcription. *J Biol Chem* 2000, **275**(41):31616-31623.

32. Zhou W, Clouston DR, Wang X, Cerruti L, Cunningham JM, Jane SM: Induction of human fetal globin gene expression by a novel erythroid factor, NF-E4. *Mol Cell Biol* 2000, **20**(20):7662-7672.

33. Volker JL, Rameh LE, Zhu Q, DeCaprio J, Hansen U: Mitogenic stimulation of resting T cells causes rapid phosphorylation of the transcription factor LSF and increased DNA-binding activity. *Genes Dev* 1997, **11**(11):1435-1446.

34. Kashour T, Burton T, Dibrov A, Amara FM: Late Simian virus 40 transcription factor is a target of the phosphoinositide 3-kinase/Akt pathway in anti-apoptotic Alzheimer's amyloid precursor protein signalling. *Biochem J* 2003, **370**(Pt 3):1063-1075.

35. Jane SM, Nienhuis AW, Cunningham JM: Hemoglobin switching in man and chicken is mediated by a heteromeric complex between the ubiquitous transcription factor CP2 and a developmentally specific protein. *Embo J* 1995, **14**(1):97-105.

36. Lim LC, Fang L, Swendeman SL, Sheffery M: Characterization of the molecularly cloned murine alpha-globin transcription factor CP2. *J Biol Chem* 1993, **268**(24):18008-18017.

37. Drouin EE, Schrader CE, Stavnezer J, Hansen U: The ubiquitously expressed DNA-binding protein late SV40 factor binds Ig switch regions and represses class switching to IgA. *J Immunol* 2002, **168**(6):2847-2856.

38. Chae JH, Kim CG: CP2 binding to the promoter is essential for the enhanced transcription of globin genes in erythroid cells. *Mol Cells* 2003, **15**(1):40-47.

39. Chae JH, Lee YH, Kim CG: Transcription factor CP2 is crucial in hemoglobin synthesis during erythroid terminal differentiation in vitro. *Biochem Biophys Res Commun* 1999, **263**(2):580-583.

40. Casolaro V, Keane-Myers AM, Swendeman SL, Steindler C, Zhong F, Sheffery M, Georas SN, Ono SJ: Identification and characterization of a critical CP2-binding element in the human interleukin-4 promoter. *J Biol Chem* 2000, **275**(47):36605-36611.

41. Bruni P, Minopoli G, Brancaccio T, Napolitano M, Faraonio R, Zambrano N, Hansen U, Russo T: Fe65, a ligand of the Alzheimer's beta-amyloid precursor protein, blocks cell cycle progression by down-regulating thymidylate synthase expression. *J Biol Chem* 2002, **277**(38):35481-35488.

42. Powell CM, Rudge TL, Zhu Q, Johnson LF, Hansen U: Inhibition of the mammalian transcription factor LSF induces S-phase-dependent apoptosis by downregulating thymidylate synthase expression. *Embo J* 2000, **19**(17):4665-4675.

43. Lynch M, Force A: The probability of duplicate gene preservation by subfunctionalization. *Genetics* 2000, **154**(1):459-473.

44. Rastogi S, Liberles DA: Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol Biol* 2005, **5**(1):28.

45. Reitzel AM, Sullivan JC, Traylor-Knowles N, Finnerty JR: Genomic survey of candidate stress-response genes in the estuarine anemone Nematostella vectensis. *Biol Bull* 2008, **214**(3):233-254.

46. Ting SB, Wilanowski T, Cerruti L, Zhao LL, Cunningham JM, Jane SM: The identification and characterization of human Sister-of-Mammalian

Grainyhead (SOM) expands the grainyhead-like family of developmental transcription factors. *Biochem J* 2003, **370**(Pt 3):953-962.

47. Uv AE, Thompson CR, Bray SJ: **The Drosophila tissue-specific factor Grainyhead contains novel DNA-binding and dimerization domains which are conserved in the human protein CP2.** *Mol Cell Biol* 1994, **14**(6):4020-4031.

48. JGI Genome Portal   [http://genome.jgi-psf.org/]

49. Sullivan JC, Kalaitzidis D, Gilmore TD, Finnerty JR: **Rel homology domain-containing transcription factors in the cnidarian Nematostella vectensis.** *Dev Genes Evol* 2007, **217**(1):63-72.

50. James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, Celio G, Gueidan C, Fraker E, Miadlikowska J, *et al.*: **Reconstructing the early evolution of Fungi using a six-gene phylogeny.** *Nature* 2006, **443**(7113):818-822.

51. Aouacheria A, Geourjon C, Aghajari N, Navratil V, Deleage G, Lethias C, Exposito JY: **Insights into early extracellular matrix evolution: spongin short chain collagen-related proteins are homologous to basement membrane type IV collagens and form a novel family widely distributed in invertebrates.** *Mol Biol Evol* 2006, **23**(12):2288-2302.

52. Bailey TL: **Discovering novel sequence motifs with MEME.** *Curr Protoc Bioinformatics* 2002, **Chapter 2**(Unit 2):4.

53. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.** *Mol Biol Evol* 2007, **24**(8):1596-1599.

54. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4**(4):406-425.

55. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *J Mol Evol* 1981, **17**(6):368-376.

56. Abascal F, Zardoya R, Posada D: **ProtTest: selection of best-fit models of protein evolution.** *Bioinformatics* 2005, **21**(9):2104-2105.

57. Jones D, Taylor W, Thornton J: **The rapid generation of mutation data matrices from protein sequences.** *Computer Applications in the Biosciencesx* 1992, **8**:275-282.

58. Felsenstein J: **PHYLIP (Phylogeny Inference Package).** In *Seattle: Distributed by the author* 3.6th edition. Department of Genome Sciences, University of Washington; 2005.

59. Stamatakis A, Ludwig T, Meier H: **RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees.** *Bioinformatics* 2005, **21**(4):456-463.

60. CIPRES. Cyberinfrastructure for Phylogenetic Research   [http://www.phylo.org/sub_sections/portal/]

61. Felsenstein J: **Confidence limits on phylogenies: an approach using the bootstrap.** *Evolution* 1985, **39**(4):783-791.

62. Halanych KM, Bacheller JD, Aguinaldo AM, Liva SM, Hillis DM, Lake JA: **Evidence from 18S ribosomal DNA that the lophophorates are protostome animals.** *Science* 1995, **267**(5204):1641-1643.

63. Bourlat SJ, Nielsen C, Economou AD, Telford MJ: **Testing the new animal phylogeny: a phylum level molecular analysis of the animal kingdom.** *Mol Phylogenet Evol* 2008, **49**(1):23-31.

64. Telford MJ, Bourlat SJ, Economou A, Papillon D, Rota-Stabelli O: **The evolution of the Ecdysozoa.** *Philos Trans R Soc Lond B Biol Sci* 2008, **363**(1496):1529-1537.

65. Rogozin IB, Wolf YI, Carmel L, Koonin EV: **Analysis of rare amino acid replacements supports the Coelomata clade.** *Mol Biol Evol* 2007, **24**(12):2594-2597.

66. Zheng J, Rogozin IB, Koonin EV, Przytycka TM: **Support for the Coelomata clade of animals from a rigorous analysis of the pattern of intron conservation.** *Mol Biol Evol* 2007, **24**(11):2583-2592.

67. Philippe H, Lartillot N, Brinkmann H: **Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia.** *Mol Biol Evol* 2005, **22**(5):1246-1253.

68. Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA: **Evidence for a clade of nematodes, arthropods and other moulting animals.** *Nature* 1997, **387**(6632):489-493.

69. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, *et al.*: **Broad phylogenomic sampling improves resolution of the animal tree of life.** *Nature* 2008, **452**(7188):745-749.

70. Irimia M, Maeso I, Penny D, Garcia-Fernandez J, Roy SW: **Rare coding sequence changes are consistent with Ecdysozoa, not Coelomata.** *Mol Biol Evol* 2007, **24**(8):1604-1607.

71. Roy SW, Irimia M: **Rare genomic characters do not support Coelomata: RGC_CAMs.** *J Mol Evol* 2008, **66**(3):308-315.

72. Roy SW, Irimia M: **Rare genomic characters do not support Coelomata: intron loss/gain.** *Mol Biol Evol* 2008, **25**(4):620-623.

73. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22**(21):2688-2690.