JAK-STAT 1:1, 23–33; January/February/March 2012; © 2012 Landes Bioscience

Comparative evolutionary genomics of the STAT family of transcription factors

Yaming Wang and David E. Levy*

Departments of Pathology and Microbiology and NYU Cancer Institute; New York University School of Medicine; New York, NY USA

Keywords: evolution, gene duplication, STAT, domain accretion, genome sequencing

The STAT signaling pathway is one of the seven common pathways that govern cell fate decisions during animal development. Comparative genomics revealed multiple incidences of *stat* gene duplications throughout metazoan evolutionary history. While pseudogenization is a frequent fate of duplicated genes, many of these STAT duplications evolved into novel genes through rapid sequence diversification and neofunctionalization. Additionally, the core of STAT gene regulatory networks, comprising *stat1* through *4*, *stat5* and *stat6*, arose early in vertebrate evolution, probably through the two whole genome duplication events that occurred after the split of Cephalochordates but before the rise of Chondrichthyes. While another complete genome duplication event took place during the evolution of bony fish after their separation from the tetrapods about 450 million years ago (Mya), modern fish have only one set of these core *stats*, suggesting the rapid loss of most duplicated *stat* genes. The two *stat5* genes in mammals likely arose from a duplication event in early Eutherian evolution, a period from about 310 Mya at the avian-mammal divergence to the separation of marsupials from other mammals about 130 Mya. These analyses indicate that whole genome duplications and gene duplications by unequal chromosomal crossing over were likely the major mechanisms underlying the evolution of STATs.

Introduction

Organismic complexity ranges widely among the Bilateria, from simple animals such as *C. elegans* with only approximately 1,000 somatic cells, to more complex organisms such as insects and sea urchins, and to the most sophisticated species, mammals. Despite the tremendous diversities among the animal kingdom, there have been few changes in basic body plans since the Early Cambrian period over 600 Mya, which include anterior-posterior and dorsal-ventral patterning, head differentiation and nervous systems.[1] Additionally, while genome sizes may range from 100 million nucleotides in *C. elegans* to about three billion nucleotides in humans, exhibiting some loose correlation to phenotypic complexities, the numbers of genes contained in various animal genomes has been remarkably constant at around 22,000.[2,3]

Consistent with the essentially unchanged body plans, there are only seven major cell-cell signaling pathways that control most developmental decisions across the Bilateria, including *Wnt*, TGFβ, hedgehog, receptor tyrosine kinase, nuclear receptor, STAT and *Notch*.[4,5] These signal transduction pathways consist of a small set of genes; however, they are modular in nature and function as kernels of so-called large **g**ene **r**egulatory **n**etworks (GRNs), which can be used repeatedly for many diverse functions throughout animal development processes to achieve necessary organism-specific phenotypic complexity.[1] The rich evolutionary history of these GRNs, revealed by comparative evolutionary genomic studies of whole-genome data sets, can provide valuable insights into their respective detailed functional mechanisms in mammals and into the evolution of animals in general.[5,6]

STAT proteins are latent cytoplasmic transcription factors activated by tyrosine phosphorylation in response to extracellular signals and are involved in many different regulatory events, including hematopoiesis, immunomodulation and development.[7] In mammals, the STAT family consists of STAT1, 2, 3, 4, 5A, 5B and 6, and share a common set of structural domains: N-terminal, coiled-coil, DNA binding, SH2, linker, and transactivation domains. Genetic mapping of the mammalian STATs indicates an evolutionary pattern that might be related to their functions,[8] organized in three tightly linked clusters on different chromosomes in mouse and human genome: *stat1* and *4*, *stat2* and *6*, *stat3* and *5a/5b*. It has been proposed that a series of tandem gene duplications of an ancestral *stat* locus gave rise to the current seven mammalian family members, followed by dispersion of linked loci to different chromosomes, and that the two *stat5* genes arose most recently.[8,9] Supporting this theory, a single *stat* gene has been identified in Drosophila[10,11] and *C. elegans*.[12,13] Additional support for the gene duplication theory of *stat* gene evolution comes from the discovery of two *stat5* genes in zebrafish.[14] The existence of a more divergent STAT pathway in *Dictyostelium discoideum* suggests that *stat* genes arose early in metazoan evolution,[15] consistent with the fundamental and diverse physiologic roles they serve.

It has long been proposed that gene duplication is a major driving force for genomic and organismal complexity during evolution.[16-19] However, the mechanism and evolutionary details of gene duplication remain largely unknown, and direct insight into this dynamic process will likely come from fine-scale, individualized comparative genomic analyses, particularly those focusing on families of paralogous genes. Focus on particular a

*Correspondence to: David E. Levy; Email: del243@med.nyu.edu
Submitted: 12/28/11; Revised: 01/16/12; Accepted: 01/19/12
http://dx.doi.org/10.4161/jkst.19418

gene family from a wide range of organisms can reconstruct their evolutionary history.[19]

The availability of high-quality whole genome sequences from a variety of organisms, including Dictyostelium, insects, nematodes, sea squirt and various vertebrate animals, allowed us to systematically investigate the evolution of the STAT gene regulatory networks. We identified STAT family member in over 20 eukaryotic genomes and performed phylogenetic analysis. Our results indicated that STAT families rose rapidly from one member to six members during early chordate evolution, likely through the proposed two rounds of whole genome duplications as well by tandem gene duplications. This expansion occurred in parallel to the rapid morphological changes of early vertebrates in the context of two rounds of whole genome duplications.[20] STAT families have undergone few changes since the teleost-tetrapod divergence about 450 Mya. Only the rise of eutherians about 130 Mya saw the duplication of *stat5*, which led to the modern seven-member STAT families in mammals. However, isolated evidence of what is often lineage-specific gene duplications by various mechanisms was found in individual species, suggesting a dynamic mode of evolution for STAT proteins and their functions.

Results

Gene duplications at the *C. elegans sta-1* locus. Previously we characterized the *C. elegans* STAT ortholog, *sta-1*.[12,13] Analysis at the *sta-1* locus revealed several partial duplications of the STAT gene, which yielded four annotated genes in WormBase (release WS155), namely *y51h4a.18, y51h4a.19, y51h4a.20* and *y51h4a.t3*. As illustrated in **Figure 1**, the duplicated exon 1 of *sta-1* formed the basis of *y51h4a.18,* duplicated exons 3 and 4 became part of *y51h4a.19,* and inversely duplicated exons 6 and 7 were annotated as *y51h4a.20.* Along with these duplicated exons, varying lengths of flanking intronic sequences, including complete 6th and 7th introns, were also duplicated (**Table 1**). Thus, *y51h4a.t3* is a copy of *y51h4a.t5,* a tRNA-Gly gene located within the 7th intron of *sta-1.* The 5' end 84 bp portion of exon 8 was also duplicated on the opposite DNA strand (**Fig. 1**), likely together with exons 6 and 7 but interrupted by a subsequent transposon insertion. The rest of the regions do not share any significant sequence homology with the *sta-1* gene. Southern blot analysis of *C. elegans* genomic DNA probed with the *sta-1* cDNA confirmed this complex genomic structure (data not shown).

Since STA-1 domain boundaries do not correspond to exon boundaries, these duplicated genes do not encode individual functional domains. Furthermore, *y51h4a.18* and *y51h4a.19* are likely to form a single transcript, as suggested by northern blot analysis of total RNA from mixed stage worms (data not shown). In addition to a 2.3 kb RNA that corresponded to *sta-1* mRNA, another faster migrating RNA species of about 1.4 kb was detected, which is significantly larger than any one of the three duplicated genes. While this RNA could be a result of alternative splicing of *sta-1* gene, it could also be a transcript that combines *y51h4a.18* (384 bp), *y51h4a.19* (675 bp) and some extra, yet unidentified exonic fragments.

Nematode genomes encode an additional STAT-like protein. In addition to STA-1, the *C. elegans* genome encodes a STAT-like protein F58E6.1, whose expression was confirmed by the matching EST clone *yk354e12*. To study this potential second STAT protein experimentally, a mixed-stage *C. elegans* cDNA library was screened for *f58e6.1*. Four clones with inserts of about 1.8 kb were isolated. Full DNA sequencing revealed a partial 5' trans-splicing leader SL1 sequence followed by a translation initiation codon, suggesting that this 1796 bp clone is a full-length *f58e6.1*. Genomic analysis revealed a very different intron-exon structure than the annotated *f58e6.1b* isoform (**Fig. 2A**). DNA sequencing revealed that the EST clone *yk354e12* contains the identical *f58e6.1* sequence, in addition to 3' sequences originally annotated as *f58e6.2*. Based on the EST sequences of *yk354e12, f58e6.2* was merged into *f58e6.1* to isoform *f58e6.1a,* whereas the original *f58e6.1* was named as isoform *f58e6.1b* (WormBase). However, our sequence analysis showed that *yk354e12* consisted of two ORFs (**Fig. 2A**), and was likely an operon transcript. Therefore, *f58e6.2* should be considered an independent gene.

F58e6.1 is predicted to encode a protein of 567 amino acid residues, with a molecular weight of about 65 kDa (**Fig. 2B**). It shares less than 20% sequence identity with STA-1, with the

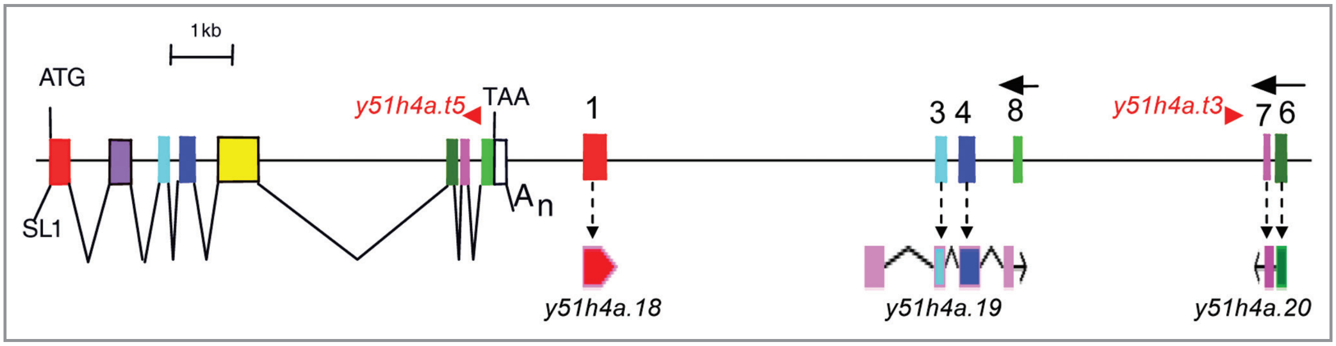


Figure 1. Genomic structures of *sta-1* locus. Exons are shown in boxes, with the same color shade for identical exons, based on comparison of cDNA and genomic sequences,. The left arrows indicate inverse duplications. Y51H4A.18, Y51H4A.19 and Y51H4A.20 are gene models annotated by Wormbase.org (release WS153). Corresponding exons are linked with dashed arrows. Scale: 1,000 bp.

Table 1. Original and duplicated intron-exon boundaries at the *C. elegans sta-1* locus. The length and percentages of identity of duplicated fragments were calculated for exons and flanking sequences

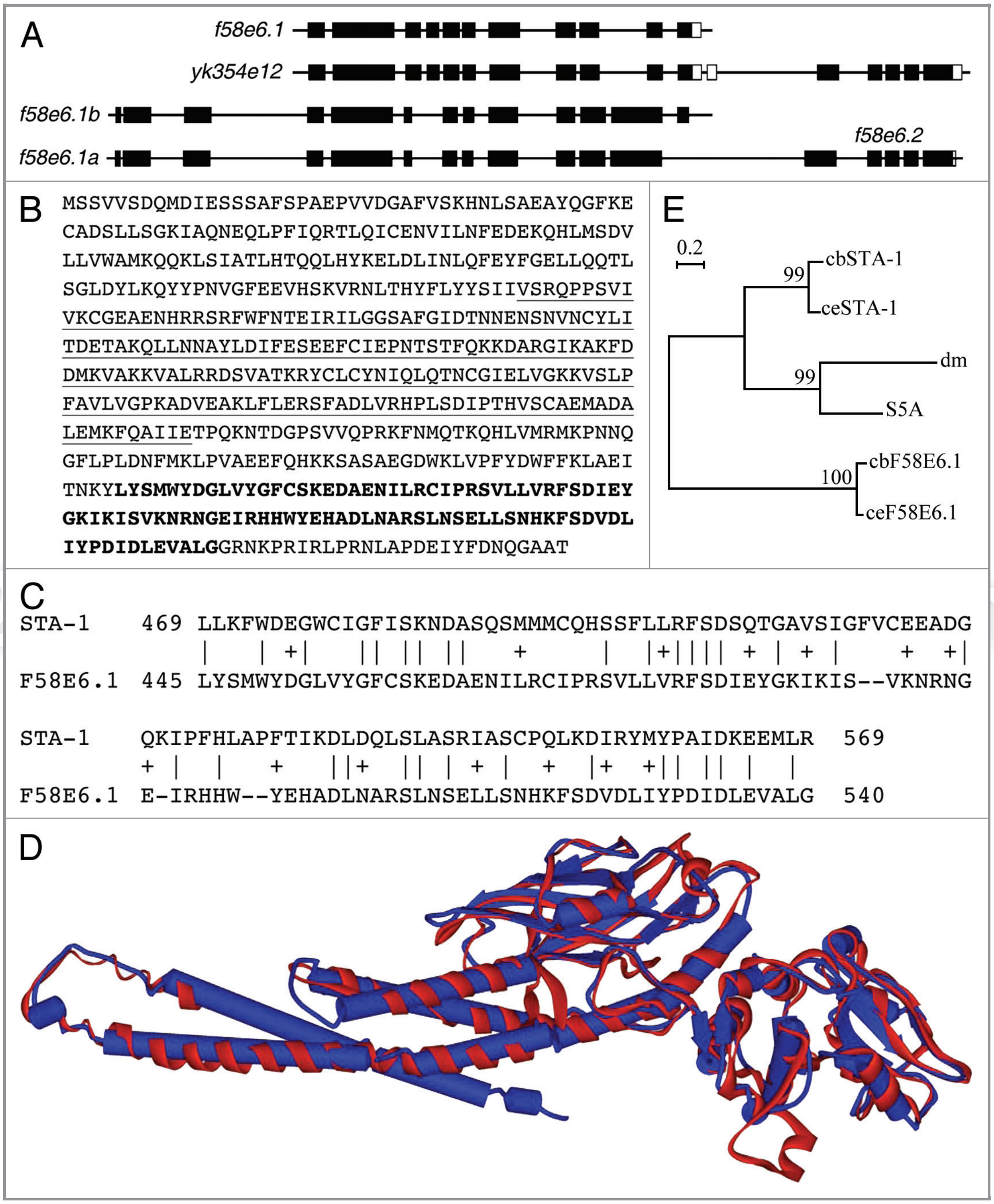
	Note	—intron—	—EXON—	—intron—
Exon 1	original		…atttccagACATGATG…TGCTACAGgttggttc…	
	duplicate		…atttccagACATGATG…TGCTACAGgttggtac…	
	length/identity	203 bp/96%	310 bp/99%	172 bp/76%
Exon 2	original		…atttccagACCCAACT…TCACTGAGgtttgttt…	
Exon 3	original		…atttccagGTCCGTCT…AAGGAAATgtgagttt…	
	duplicate		…atttccagGCCCGTCT…AAGGAAATgtgcgttt	
	length/identity	40 bp/100%	128 bp/98%	117 bp/89%*
Exon 4	original		…tttttcagCCGTAACA…TGAGCAAAgtaagttg…	
	duplicate		…tttttcagCCGTAACA…TGAGCAAAgtaagttg…	
	length/identity	8 bp/100%	238 bp/100%	238 bp/79%
Exon 5	original		…tttaacagGAAGAACA…GATATTCGgttagttt…	
Exon 6	original		…aatttcagATACATGT…ACAAATGGgtaggtta…	
	inverse duplicate		…aatttcagATACATGT…ACAAATGGgtaggtta…	
	length/identity	157 bp/90%	135 bp/99%	42 bp/98%†
Exon 7	original		…tattttagGAATTTCA…ATGGAGATgtgagtga…	
	inverse duplicate		…tattttagGAATTTCA…ATGGAGATgtgagtga…	
	length/identity	42 bp/98%†	123 bp/100%	223 bp/96%‡
Exon 8§	original		…tttccagGTCCGACG…ACTGTAAAtcgaatgt…	
	inverse duplicate		…agtgcagACTTGGTC…ATCACAAAtgtgctat…	
	length/identity	NA/NA	84 bp/100%	NA/NA

*Intron 3 is 164 bp long and is not duplicated in its entirety. †Intron 6 is 42 bp long and is not duplicated in its entirety together with exons 6 and 7. ‡Intron 7 is 223 bp, and is duplicated in its entirety together with exon 7 and the first 22 bp of exon 8. §Exon 8 is 407 bp long, only 5′ end 84 bp is duplicated and the flanking sequences do not share any significant identities. Thus identity analysis is not applicable (NA).

most similarity in the SH2 domain, which shares 33% sequence identity (**Fig. 2C**). Searching a library of Hidden Markov Models that represent all proteins of known structure[21] revealed that F58E6.1 might contain two protein domains, SH2 (E-value, 1.3e-14) and DNA binding (E-value 4.1e-05). A similar search with STA-1 produced three domains, SH2 (E-value 7.2e-31), DNA binding (E-value 1.9e-60) and STAT coiled-coil domain (E-value 8.8e-38), which is consistent with our previous studies.[12,13] For comparison, human STAT5A contains four domains, SH2 (E-value 1.4e-34), DNA binding (E-value 8.9e-92), STAT coiled-coil domain (E-value 4.7e-49) and STAT N-domain (E-value 1.6e-39). Thus, F58E6.1 has significantly higher predicted E-values for its two domains than established STAT family members do for corresponding domains, suggesting that F58E6.1 sequences fit poorly with the Hidden Markov Models of SH2 and DNA binding domains. Although F58E6.1 has a single tyrosine residue at the carboxyl-terminus, it has a very short C-terminal fragment, indicating lack of a transactivation domain (**Fig. 2B**). Therefore, F58E6.1 is very different from STA-1 and is unlikely to be a bona fide member of the STAT family. Consistent with this notion, co-expression of F58E6.1 with a tyrosine kinase in mammalian tissue culture system failed to show any DNA binding activities using a STAT consensus DNA sequence motif in electrophoretic mobility shift assay (data not shown), although many other STAT proteins, including STA-1, score positive in this assay.[12,13]

However, among all the known and predicted peptide sequences, the best matches for F58E6.1 are STATs. Conversely, among all the *C. elegans* polypeptides, F58E6.1 is the second best match for STATs after STA-1. To investigate whether F58E6.1 protein sequence was compatible with a STAT-like tertiary structure, a model was generated by using the homology modeling program 3D-JIGSAW,[22] extracting coordinates for the unphosphorylated mouse STAT5A crystal structure.[23] Overall, the resulting model fit well with the STAT5A structure (**Fig. 2D**), suggesting that F58E6.1 may fold into a STAT like structure. It is also noteworthy that a *C. briggsae* homolog shares 96% protein sequence identity with F58E6.1, despite the approximately 100 million years of intervening species divergence.[24] For comparison, the *C. briggsae* STAT shares 80% sequence identity with STA-1 (**Fig. 2E**), suggesting F58E6.1 is under more stringent selection pressure than the *sta-1* locus. Therefore, F58E6.1 probably shared a common ancestor with STA-1, arose through a gene duplication event specific to the nematode lineage, and survived through rapid sequence diversification and neofunctionalization, although its current function remains to be determined.

Ancient STATs in mycetozoa. The mycetozoa represent one of the earliest branches diverged from the last common ancestor



```
MSSVVSDQMDIESSSAFSPAEPVVDGAFVSKHNLSAEAYQGFKE
CADSLLSGKIAQNEQLPFIQRTLQICENVILNFEDEKQHLMSDV
LLVWAMKQQKLSIATLHTQQLHYKELDLINLQFEYFGELLQQTL
SGLDYLKQYYPNVGFEEVHSKVRNLTHYFLYYSIIVSRQPPSVI
VKCGEAENHRRSRFWFNTEIRILGGSAFGIDTNNENSNVNCYLI
TDETAKQLLNNAYLDIFESEEFCIEPNTSTFQKKDARGIKAKFD
DMKVAKKVALRRDSVATKRYCLCYNIQLQTNCGIELVGKKVSLP
FAVLVGPKADVEAKLFLERSFADLVRHPLSDIPTHVSCAEMADA
LEMKFQAIIETPQKNTDGPSVVQPRKFNMQTKQHLVMRMKPNNQ
GFLPLDNFMKLPVAEEFQHKKSASAEGDWKLVPFYDWFFKLAEI
TNKYLYSMWYDGLVYGFCSKEDAENILRCIPRSVLLVRFSDIEY
GKIKISVKNRNGEIRHHWYEHADLNARSLNSELLSNHKFSDVDL
IYPDIDLEVALGGRNKPRIRLPRNLAPDEIYFDNQGAAT
```

```
STA-1    469 LLKFWDEGWCIGFISKNDASQSMMMCQHSSFLLRFSDSQTGAVSIGFVCEEADG
             |   | +|   || || ||   +     |  |+|||| + | + |    +  +|
F58E6.1 445 LYSMWYDGLVYGFCSKEDAENILRCIPRSVLLVRFSDIEYGKIKIS--VKNRNG

STA-1   QKIPFHLAPFTIKDLDQLSLASRIASCPQLKDIRYMYPAIDKEEMLR  569
        + |  |   +   ||+  || | + |  +  |+  +|| || |  |
F58E6.1 E-IRHHW--YEHADLNARSLNSELLSNHKFSDVDLIYPDIDLEVALG  540
```

Figure 2. *C. elegans* genome encodes a STAT-like protein, F58E6.1. (A) Genomic intron-exon structure of f58e6.1. (B) Predicted protein sequence of f58e6.1. (C) Limited sequence homology between STA-1 and F58E6.1. (D) Predicted structural similarity between F58E6.1 and mouse STAT5A. (E) Phylogenetic analysis of *C. elegans* and *C. briggsae* STA-1 and F58E6.1.

of animals, fungi and plants. Therefore, the discovery of STAT family members in the slime mold *Dictyostelium discoideum*[15] placed this phosphotyrosine signaling pathway at the beginning of multicellular evolution. Interestingly, while lower animals generally have only one or two STAT family members, the Dictyostelium genome encodes four STATs, dstA, dstB, dstC and dstD (dictyBase, www.dictybase.org),[25] suggesting extensive usage of SH2 domain-mediated phosphotyrosine signaling in this simple organism.

The N-terminal half of the four slime mold STATs all contain stretches of Asn and Gln amino acid residues, similar to many developmentally regulated genes. Besides this feature that is unlikely to be specific to STAT function, dstA through D proteins contain three attributes that are characteristic of the STAT family, namely DNA-binding and SH2 domains as well as a tyrosine phosphorylation site. However, they lack the N-terminal and transactivation domains characteristic of STAT proteins from higher organisms, consistent with the hypothesis that STATs evolved through domain accretion.[13]

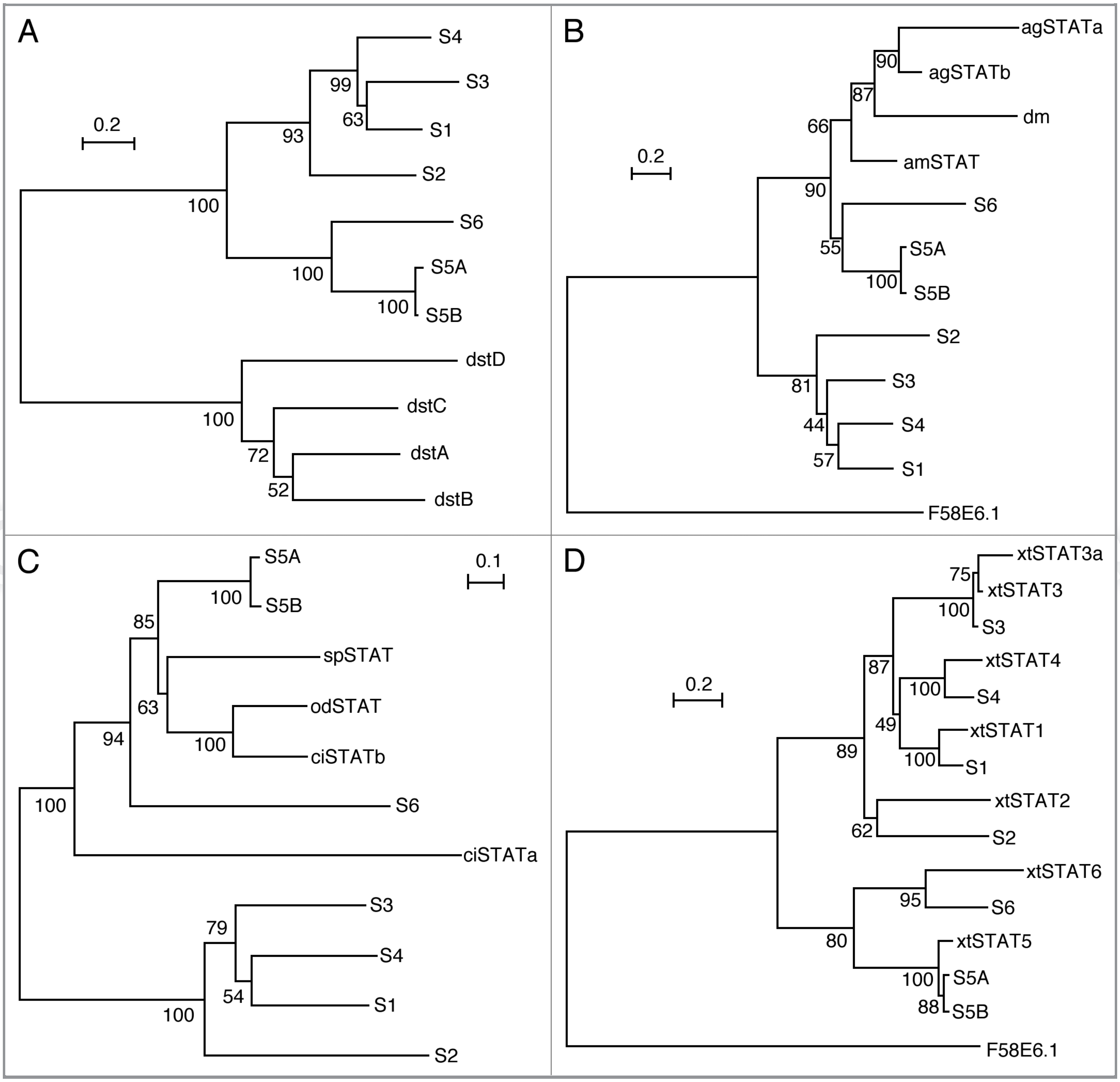


Figure 3. Phylogenetic relationships of STAT proteins. Phylogenetic trees are shown for (A) Dictyostelium, (B) arthropods, (C) deuterostomes and (D) Xenopus.

Protein sequence analysis revealed that the four slime mold STATs share less than 10% overall sequence identities with the seven mammalian STATs, while they share about 22–38% identity with each other. Therefore and not surprisingly, molecular phylogenetic analysis indicated that the slime mold STATs form a distinct clade (**Fig. 3A**), which raises the possibility of mixed concerted and birth-and-death evolution of STAT family of transcription factors.

STATs in the arthropods. With over one million species, the arthropods represent the most diverse group of animals and likely shared the last common ancestor with vertebrates at least one billion years ago.[26] A canonical STAT signaling pathway, analogous to that in mammals, exists in the fruit fly *Drosophila melanogaster*.[27] Genome sequencing revealed a single *stat* gene in Drosophila and the honey bee *Apis mellifera*, but two *stat* genes in the mosquito *Anopheles gambiae*. All insect STATs are predicted to be identical to mammalian STATs in domain structure, in contrast to the partial identities in the Dictyostelium and nematode STATs, suggesting that STAT evolution by domain accretion stopped before the rise of Deuterostomes over a billion years ago. The insect STATs also form a single clade in phylogenetic analysis, and constitute an ancient class of STATs with the clade consisting of STAT5s and 6 (**Fig. 3B**).

The two mosquito STATs are almost identical in protein length, but share only 47% overall sequence identity. Ag-STATa (Ensembl gene ID: ensangg00000021440), previously found to be involved in immune responses to bacterial infections,[28] is located at one end of chromosome 3L, encoded by a single exon, the only documented instance of a single-exon STAT. The ag-STATb gene (Ensembl gene ID ensangg00000006157) is on the X chromosome, encoded by 8 exons. Intron loss suggests that ag-STATa was derived from ag-STATb through gene duplication by retrotransposition, an event that likely happened after the split of the Drosophila and Anopheles about 400~500 Mya. As most duplicated genes are rapidly degenerated into pseudogenes and disappear,[19] ag-STATa has likely survived through rapid sequence diversification and neofunctionalization. Consistent with this prediction, comparative proteomics revealed an expansion of immunity-related genes, including the *stat* genes, in Anopheles spp relative to the Drosophila spp, likely driven by exposure to an expanded set of pathogens.[29]

STATs in the invertebrate subgroup of deuterostomes. The extant deuterostomes are the echinoderms, hemichordates and chordates, including the urochordates (ascidians, thaliaceans and larvaceans), cephalochordates and vertebrates. The California purple sea urchin, *Strongylocentrotus purpuratus,* is an echinoderm, and its 800 Mb genome contains a single STAT protein that shares 48% overall sequence identity with human STAT5 and less than 30% with any other human STAT. Similarly, the larvacean *Oikopleura dioica* contains a single STAT with 47% overall sequence identity to human STAT5 in its miniature 70 Mb genome. However, genome sequences of two ascidians, *Ciona intestinalis* and *C. savignyi*, revealed two STATs that share only 27% overall sequence identity with each other. Phylogenetic analysis suggests that STATs in invertebrate deuterostomes belong to the ancient class of STATs (**Fig. 3C**). Sequences of the sea urchin and pelagic tunicate STATs as well as the Ciona STATb are most similar to STAT5. This similarity is in contrast with STATs in protostomes, which formed a sister clade to both STAT5 and 6 (**Fig. 3C**). Assuming the absence of a second STAT in *S. purpuratus* and *O. dioica*, early deuterostomes likely evolved a single STAT protein and the two STATs in Ciona would have resulted from a gene duplication event in acidians.

STATs in the teleost fish. After the divergence of vertebrates from urochordates about 770 Mya, ancestral vertebrates likely underwent two rounds of whole genome duplication, followed by the divergence of ray-finned fish from tetrapods about 450 Mya.[20,26] Analysis of three fish genomes, the zebrafish *Danio rerio* and the two pufferfish *Fugu rubripes* and *Tetraodon nigroviridis,* identified multiple STAT proteins that are clearly orthologous to mammalian STATs, specifically STAT1–4, STAT6 and one STAT5 (**Fig. 4**), suggesting the presence of all these STATs in their common ancestor with tetrapods, which existed about 450 Mya. The presence of these multiple STATs further suggests that the expansion of the STAT family as largely due to the whole genome duplications early in vertebrate evolution.[30]

In addition to these 6 STATs, the zebrafish has acquired two extra orthologous STATs, one similar to STAT1 and the other similar to STAT5 (**Fig. 4**). These extra STATs were likely gained through the whole genome duplication that occurred early in teleost evolution after divergence from the tetrapods,[31,32] and thus unlikely to be present in their common ancestor with the mammals. These duplicated STATs may have survived due to expression pattern diversification, similar to the duplicated zebrafish JAK2.[33] Furthermore, they were likely lost in the pufferfish lineage, presumably in favor of a much more compact genome.

In addition to those orthologous STATs, the two pufferfish genomes also contain a non-orthologous STAT that appears to form an outgroup to the clade consisting of STAT1–4 (**Fig. 4**). Interestingly, this extra *stat* gene is located immediately adjacent to *stat4* in a tail-to-tail configuration in both pufferfish genomes, suggesting an inverted gene duplication event followed by rapid sequence diversification.

STATs in amphibians and birds. Within the tetrapods, amphibians diverged from amniotes, which include birds and mammals, about 370 Mya, and birds and mammals diverged around 310 Mya.[34] Analysis of the Xenopus genome indicated that amphibians likely have a single set of STATs, similar to early tetrapods (**Fig. 3D**). However, Xenopus has an extra, nearly identical STAT3, with over 95% protein sequence identity. The two *stat3s* are located on the same chromosome in a tail-to-head configuration, separated by only 14 kilobases. Despite the highly identical exonic sequence, intronic sequences have completely diverged. This structure is indicative of a recent gene duplication event, with the divergence of intronic sequences while maintenance of ORFs indicates that the survival of both genes was driven by neofunctionalization.[35]

Genome analysis of birds suggests five members, STAT1, 3, 4 a single STAT5 and STAT6 (data not shown). Both the clawed frog and the red jungle fowl genomes have a single STAT5 that is equally diverged from mammalian STAT5A and STAT5B, suggesting that the gene duplication event that led to two mammalian STAT5s occurred after speciation about 310 Mya. Additionally, the chicken *stat5* gene is flanked on the same DNA strand by *stat3* upstream and *lgp1* downstream, a configuration identical to the *stat5b* locus minus the inverted *stat5a* in mammals, consistent with a mammalian specific *stat5* duplication.

Mammalian *stat* genes are tightly linked in three chromosomal clusters. The seven mammalian *stat* genes exist in three linked clusters on different chromosomes, *stat1* and *4* on chromosome 1, *stat2* and *6* on chromosome 10, *stat3* and the two *stat5s* on chromosome 11 in mice.[8] Similar linkages occur in humans and are likely maintained throughout the mammals (**Table 2**), despite large variations of chromosome number, genome sizes, and extensive chromosomal rearrangements that occurred during the 130 million years of the therian mammals. While the small cluster sizes certainly contribute significantly to this pattern of preservation, inherent local chromosomal stability

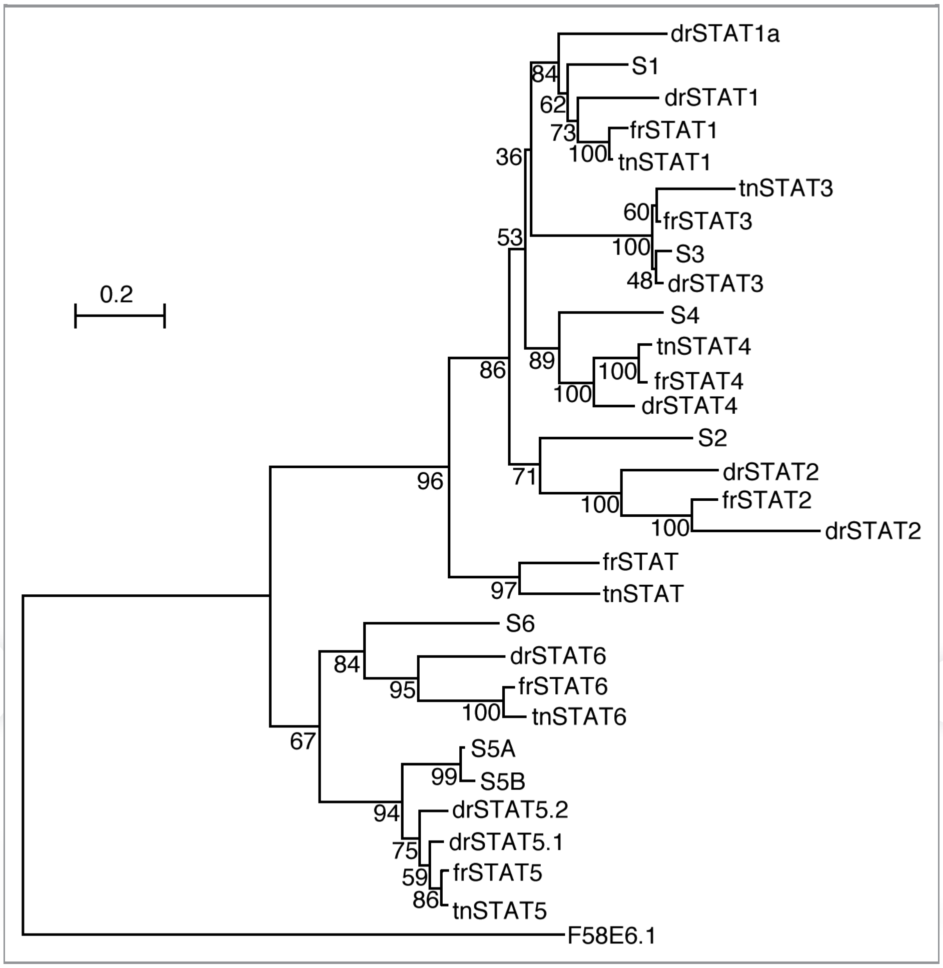


Figure 4. Phylogenetic relationships of fish STAT proteins. Phylogenetic trees are shown for *Danio rerio, Fugu rubripes* and *Tetraodon nigroviridis* STAT proteins.

or even a selective constraint on tight-linkage may also have contributed to this conserved arrangement.

Discussion

The partial *sta-1* duplications in *C. elegans*. The duplication at the *sta-1* locus was not found in the genome of a related nematode, *C. briggsae* (WormBase WS155), suggesting that it occurred after the split of the two worm species about 100 Mya.[24] Since generally intron sequences drift rapidly, the fact that duplicated intronic fragments showed significant sequence identities with their corresponding regions (**Table 1**) suggests the duplication occurred quite recently. Additionally, the nearly identical exons also strongly support their recent birth. Alternatively, while these duplicated exons do not encode any intact functional protein domains, they could still be transcribed and

Table 2. Chromosomal clustering patterns of STAT genes in mammals

	STAT1	STAT4	STAT5B	STAT5A	STAT3	STAT2	STAT6
	— < ——— < —		— < ———— > —— < —			— < ——— < —	
Monodelphis domestica	Scaffold 2		Scaffold 18			N/A	Sca. 303
Bos taurus	ChrUn. 103		Chromo. 19			Chromo. 5	
Canis familiaris	Chromo. 37		Chromo. 9			Chromo. 10	
Homo sapiens	Chromo. 2		Chromo. 17			Chromo. 12	
Mus musculus	Chromo. 1		Chromo. 11			Chromo. 10	
Pan troglodytes	Chromo. 2B		Chromo. 17			Chromo. 12	
*Rattus norvegicus**	Chr. 9		Chromo. 10			Chromo. Seven	

serve some unknown critical function, which could provide strong pressure against nucleotide mutations.

Gene duplication is thought to be generated by three types of mechanisms, chromosomal unequal crossing over, retrotransposition, and chromosomal (or genome) duplication, the outcomes of which are quite different.[19] The tandem nature of these *sta-1* duplications and the presence of highly identical, yet partial intronic sequences exclude the possibility of generation by the latter two mechanisms. Therefore, it is possible that a recent, unequal crossing-over led to a *sta-1* complete tandem duplication, which rapidly degraded into a non-functional, possibly transcriptionally active pseudogene through a series of complex genomic rearrangements, including the loss of exons 2, 5 and partial 8, the inversion of exons 6 through 8, and an insertion-disruption of the remain exon 8.

However, a more likely scenario, which would require fewer discrete evolutionary steps, would be duplication by exon shuffling.[36] Supporting evidence for exon shuffling, which is also referred to as domain shuffling, comes from comparative genomics which revealed that protein domains correlate strongly with exons and that exon-bordering domains tend to be bounded by same phase introns.[37,38] However, direct evidence for the exon shuffling theory should come from detailed, genome-wide analysis of newly arisen partial gene duplications. A recent analysis of *C. elegans* genome identified 290 pairs of gene duplicates with less than 10% sequence divergence at synonymous sites, of which approximately 60% are partial or chimeric in nature.[39,40] About 36% of these duplicate pairs are located on different chromosomes, and even among the rest that do reside on the same chromosome, majorities are separated by other non-duplicated annotated genes. A re-examination of these partial or chimeric duplicates revealed a common feature. In all but three cases, a set of exon(s), together with intronic sequences, was duplicated as a unit to various locations in the genome, suggesting that these represent actual exon shuffling events.

Therefore, the duplications at the *sta-1* locus are likely to result from three exon shuffling events, of which exon 1, exons 3–4 and exons 6–partial 8 each represent a shuffling unit. The duplicated exons 6 through 8 were likely further disrupted by a transposon insertion followed by excision at the beginning of exon 8, since a blast search yielded 19 significant matches including one next to the transposon gene *k10f12.5*.

STAT origin and early evolution. The discovery of STAT signaling in Dictyostelium extended this intercellular phosphotyrosine pathway beyond the Metazoa and raised the possibility of a single origin of STATs during the single cell-metazoan evolutionary process.[41] However the controversial phylogenetic status of Dictyostelids undermines a single STAT origin theory. At issue is whether the slime mold diverged before or after fungi from the line that later evolved into metazoans. While many Dictyostelium proteins are more similar to human orthologs than those of yeasts, phylogenetic analysis suggested the divergence occurred before that of fungi,[25] which don't employ SH2-mediated phosphotyrosine signaling. If there existed a single STAT ancestor after the plant-animal split, then it was lost along with the phosphotyrosine signaling pathway in the single-cellular fungal lineage, while it expanded in the multicellular mycetozoan and metazoan lineages. Alternatively, STAT signaling may have arisen during the transition to multicellularity early in the metazoan evolution, and acquired by primordial mycetozoas through horizontal gene transfer, consistent with the observation that many Dictyostelium proteins are more similar to human orthologs, in contrast to the divergence of yeast and human orthologs.[25]

It is likely that the common ancestors of the so-called "crown eukaryotes" are single-cell organisms with very diverse genomes[25] extant when the plant-animal split took place about 1.6 billion years ago. Since then, plants, animals and slime molds may have independently evolved multicellularity while fungi retained ancestral single cellularity. As phosphotyrosine-based signaling pathways are considered to be a tool specific for intercellular communications within a multicellular organism,[41] the STAT signaling pathway might have arisen through convergent evolution early in the mycetozoan lineage independently of its origin in the metazoan lineage during their respective transitions from single cellularity to multicellarity. Further evidence for an independent STAT origin in mycetozoa, a distant ortholog of the STAT linker-SH2 domain was found in Saccharomyces and Arabidopsis, suggesting its more ancient origin before the plant-animal split.[42] Thus, the linker-SH2 domain may be the original evolutionary foundation upon which STATs later evolved through domain accretion in the mycetozoa and metazoan lineages.

STAT evolution in the early metazoa. Since the mycetozoa is no longer considered a direct ancestor of metazoans,[25] another ancestral STAT likely arose early in metazoan evolution after the fungi-animal divergence 1.5 billion years ago. As nematodes diverged from other metazoans about 1.2 billion years ago,[26] characterizations of the STAT in *C. elegans* provides insight into this STAT ancestor.[12,13] Prior to nematode divergence, the ancestral STAT likely had the same domain structure as the nematode STAT, lacking the N-terminal domain, which is conserved among all other known animal STATs.

The Cnidaria, including very simple-bodied animals such as corals, sea anemones, hydras and jellyfishes, are likely diverged from the metazoan lineage shortly before the rise of bilaterians. A study of *wnt* signaling in the sea anemone *Nematostella vectensis* revealed an unexpectedly diverse gene family, providing significant clues to early animal body-plan evolution as well as insights into the signaling pathway evolution and functions in protostomes and deuterostomes.[43] Interestingly, a search of the freshwater polyp *Hydra magnipapillata* EST database yielded two clones that encode a single STAT. This STAT has a stretch of charged residues after the putative tyrosine phosphorylation site, similar to STAT5s in mammals, suggesting a potential transcription activation domain in this early STAT.

STAT evolution in deuterostomes. The ancestors of deuterostomes were likely to have a single *stat* in their genome. Both the Pseudocoelomates like *C. elegans* and most protosomes appear to have a single *stat* gene, which is unlikely to represent a loss of STATs in these two lineages since the genome sequencing of the California purple sea urchin from the most basal deuterostome lineage Echinodermata also revealed a single STAT.

How many *stat* genes were there in the early vertebrates? The answer will likely come from the genome sequencing of species from its two sister groups, the cephalochordates and the urochordates. The two STATs in the sea squirt Ciona were likely a result of gene duplications specific in the acidian lineage. Though genome sequencing of the larvacean *O. dioica* revealed a single STAT, whether this status is representative of the tunicates in general is not clear, since *O. dioica* clearly underwent a drastic reduction in genome size. A search of the amphioxus *Branchiostoma floridae* EST database revealed two clones that encode a single STAT (data not shown).

Clearly six of the seven vertebrate STATs arose before the divergence of ray-finned fish from the tetrapods 450 Mya (**Fig. 4**). Since two rounds of whole genome duplication likely occurred before that divergence, the details of the STAT family expansion from one or two to six members could come from comparative analysis of genomes from two basal lineages, the lampreys and the cartilaginous fish. Interestingly, a search of the dogfish shark *Squalus acanthias* EST database revealed two clones, each encoding a different STAT. One clone (GenBank accession number DV500695) showed 77% protein sequence identity to human STAT1 while the other (accession number DV497815) showed 71% to human STAT5. As their respective sequence identities to other human STATs are significantly lower, they likely represent *stat1* and *5* genes in the cartilaginous fish. As the whole genome duplication events probably occurred close to the origin of the teleost fish,[32] it is tempting to suggest that these two *stat* genes are the two founding members of vertebrate STATs. Specifically, through whole genome duplications and subsequent survivals, ancestral *stat1* likely produced the present-day *stat1–4* genes whereas ancestral *stat5* yielded *stat6* and two *stat5s.*

Dynamic STAT evolution by duplications. Gene duplications are proposed to be a major driving force for genomic and organismal complexities during evolution.[16,17] Our comparative genomics of the STAT family of transcription factors has provided strong and detailed evidence for the gene duplication theory. Despite it rareness, whole genome duplications provided the most genomic raw material for evolutionary selections. A clear example is the whole genome duplication event in the ray-finned fish lineage after its divergence from the tetrapods.[31,32] While all the duplicated STATs were likely lost in the pufferfishes during the drastic reduction of their genome size, at least two duplicated STATs survived in the zebrafish genome (**Fig. 4**), probably by rapid changes to their expression profiles as is the case for JAK2 duplicates.[33] Additionally, whole genome duplications were likely responsible for the major expansion in STAT family members in the early vertebrate evolution. Similar evidence for genome duplication and divergence in the evolution of STAT proteins has recently been provided by analysis of STAT genes in teleostean fishes.[30]

The tightly linked chromosomal clusters of *stat* genes in many of the vertebrate genomes (**Table 2**) suggest that gene duplication by unequal chromosomal crossing over also contribute significantly to the STAT family expansion. In addition to the two *stat5* genes in mammals, the two *stat3* genes in the Xenopus genome clearly resulted from such a recent duplication event, as well as the extra *stat* that was likely duplicated from *stat4* in the pufferfish. In contrast, gene duplication by retrotransposition likely did not play important roles in STAT evolution. The malaria mosquito *A. gambiae* provides the only example where a retrotransposition event resulted in two functional *stat* genes. The only other retrotransposition example uncovered by this study is *stat2* gene in the domesticated dog *Canis familiaris*. A reverse transcription and insertion event led to the duplicated *stat2* gene on the X chromosome. While the duplicated copy retains over 90% DNA sequence identity, multiple tiny indels (insertion/deletion) completely disrupted the ORF, thus rendering it a pseudogene.

While gene duplication obviously is the major mechanism underlying STAT evolution, it likely also provides diversity to the genome beyond the signaling pathway. Due to the strong selection pressure of maintaining single-copy status on major developmental signaling pathways, the majority of the full duplications of *stat* genes were lost during evolution while the rest were fixed in the genome by rapid sequence or expression diversifications and sub- or neo-functionalizations. Rarely, the sequence diversification would be such that the duplicated gene became a novel gene with little resemblance to the ancestral copy. One such example is the *C. elegans f58e6.1* gene. Protein sequence analysis suggested that it is likely a duplicated copy of the worm *stat* gene; however, it lost the major characteristics of the STAT family and can no longer be considered a bona fide member (**Fig. 2**).

Our study also uncovered another type of gene duplication during the evolutionary history of the *stat* genes. Detailed genomic sequence analysis of the *stat* locus in *C. elegans* revealed immediately adjacent, recent partial duplications, which likely result from three independent exon-based events such as exon-shuffling (**Fig. 1 and Table 1**). Such partial duplications are not rare; a recent survey identified at least 39 such instances in the *C. elegans* genome, including the three at the *stat* locus.[40] Nor are such duplications nematode specific. We found in the human genome, about 400 bp of the *stat2* exon 24 was duplicated from chromosome 12 to chromosome 8, retaining 87% DNA sequence identity. In the rat genome, about 500 bp of the *stat6* genes, consisting exon 13, intron 13, exon 14 and part of intron 14, were duplicated from chromosome 7 to chromosome 13, retaining 99% sequence identity. While partial gene duplications cannot yield a full copy of the ancestral gene, it can fuse into other genes in the genome to form chimeric novel genes, as demonstrated by various examples in *C. elegans*.[40] Perhaps this is the same process as exon-shuffling, which is thought to be a major mechanism underlying domain accretion, a process where novel protein domains confer novel functions to existing genes.[44]

The STAT proteins are an ancient family of signaling molecules that arose early in evolution and have diversified during the radiation of animal species. Selective retention of the basic domains required for phosphotyrosine signaling, an SH2 domain and a site for protein tyrosine phosphorylation, indicate the importance of this module throughout multicellular organisms. It is thus of particular interest that these diverse proteins have also acquired additional, apparently phosphotyrosine-independent functions, without losing their participation in this basic mode of signaling,[45-48] documenting the parsimony of evolution.

Methods

Cloning and characterization of *f58e6.1*. A Blast search using human STAT1 of *C. elegans* protein database yielded another significant hit F58E6.1, in addition to STA-1. The matching EST clone *y354e12* (gift of Y. Kohara, NIG, Japan) was used to screen 5×10^6 colonies from a mixed-stage *C. elegans* λgt11 cDNA library (gift of P. Okkema, IL). Eight positive colonies were isolated and grouped into three categories based on insert sizes, ~1.8 kb (4 colonies), ~1.4 kb (2 colonies) and ~1 kb (2 colonies). DNA sequencing revealed that these three groups shared the same 3' end sequences. Clone #1 which has a ~1.8 kb insert and EST clone *yk354e12* were then fully sequenced. Clone #1 appeared to be a full-length cDNA clone, as it started with a partial trans-splicing leader SL1 sequence and contained a single ORF. Clone *yk354e12* contained identical clone #1 sequences except the 5' SL1 and the beginning ORF sequences, plus ~800 bp extra 3' sequences. Sequence analysis suggested *yk354e12* to be operon transcript as it had two ORFs. The 5' ORF was identical to clone #1, but the 3' ORF matched with F58E6.2, which was originally annotated as a single gene but later merged with annotated F58E6.1 into an isoform F58E6.1a, based on the end sequences of *yk354e12*. Clone #1 was FLAG-tagged and biochemically characterized as described previously.[13]

Domain prediction and structure modeling. Predicted F58E6.1 protein sequences were used to search the Superfamily database,[21] which is a collection of Hidden Markov Models from all proteins of known structure, for potential domain structures at www.supfam.org. For 3D structure modeling, the homology-based structure prediction program 3D-JIGSAW[22] (www.bmm.icnet.uk/servers/3djigsaw) was used to first identify the mouse unphosphorylated STAT5A crystal structure as the one that F58E6.1 was mostly related to, and to subsequently build a potential structure model, which was further fitted to STAT5A crystal structure using SwissPdb Viewer[49] at www.expasy.org/spdbv. The final superimposed structures were then visualized in ViewerLite v5.0 (Accelrys, CA).

Sequence data. The genome sequences and annotations used in this study are from the following sources: WormBase (release WS155, www.wormbase.org) for nematodes *C. elegans* and *C. briggsae*; dictyBase (www.dictybase.org) for the slime mold Dictyostelium; Ensembl (www.ensembl.org) for fruit fly *D. melanogaster,* honeybee *A. mellifera,* malaria mosquito *A. gambiae,* ascidians *C. intestinalis* and *C. savignyi,* zebrafish *D. rerio,* pufferfish *F. rubripes* and *T. nigroviridis,* clawed frog *X. tropicalis,* opossum *M. domestica,* dog *C. familiaris,* cow *B. taurus,* mouse *M. musculus,* rat *R. norvegicus,* chimpanzee *P. troglodytes* and human *H. sapiens.* All genome annotation data were manually checked for errors and for incompleteness before phylogenetic analysis. GenBank for genome sequences of the California purple sea urchin *S. purpuratus,* pelagic tunicate *O. dioica,* for EST sequences of the freshwater polyp *H. magnipapillata,* the amphioxus *B. floridae,* the chicken *G. gallus,* and the dogfish shark *S. acanthias.*

Phylogenetic analysis. A multiple sequence alignment for each STAT group was created using ClustalW 1.81 with default parameters. A phylogenetic tree based on the neighbor joining method was generated with 1,000 bootstrap replicates of the alignment, excluding positions with gaps and correcting for multiple substitutions. The resulting un-rooted tree was then visualized in the NJPLOT program and manually rooted.

Acknowledgments

We thank L. Pan for expert technical assistance. This work was supported by a grants from the NIH (AI28900, U54AI057158) and a gift from the G. Harold and Leila Y. Mathers Charitable Foundation.

References

1. Davidson EH, Erwin DH. Gene regulatory networks and the evolution of animal body plans. Science 2006; 311:796-800; PMID:16469913; http://dx.doi.org/10.1126/science.1113832
2. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, et al. Ensembl 2006. Nucleic Acids Res 2006; 34(Database issue):D556-61; PMID:16381931; http://dx.doi.org/10.1093/nar/gkj133
3. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, et al. Ensembl 2012. Nucleic Acids Res 2012; 40(Database issue):D84-90; PMID:22086963; http://dx.doi.org/10.1093/nar/gkr991
4. Gerhart J. 1998 Warkany lecture: signaling pathways in development. Teratology 1999; 60:226-39; PMID:10508976; http://dx.doi.org/10.1002/(SICI)1096-9926(199910)60:4<226::AID-TERA7>3.0.CO;2-W
5. Pires-daSilva A, Sommer RJ. The evolution of signalling pathways in animal development. Nat Rev Genet 2003; 4:39-49; PMID:12509752; http://dx.doi.org/10.1038/nrg977
6. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, et al. Comparative genomics of the eukaryotes. Science 2000; 287:2204-15; PMID:10731134; http://dx.doi.org/10.1126/science.287.5461.2204
7. Levy DE, Darnell JE, Jr.. Stats: transcriptional control and biological impact. Nat Rev Mol Cell Biol 2002; 3:651-62; PMID:12209125; http://dx.doi.org/10.1038/nrm909
8. Copeland NG, Gilbert DJ, Schindler C, Zhong Z, Wen Z, Darnell JE, Jr., et al. Distribution of the mammalian Stat gene family in mouse chromosomes. Genomics 1995; 29:225-8; PMID:8530075; http://dx.doi.org/10.1006/geno.1995.1235
9. Ihle JN. STATs: signal transducers and activators of transcription. Cell 1996; 84:331-4; PMID:8608586; http://dx.doi.org/10.1016/S0092-8674(00)81277-5
10. Yan R, Small S, Desplan C, Dearolf CR, Darnell JE, Jr. Identification of a Stat gene that functions in Drosophila development. Cell 1996; 84:421-30; PMID:8608596; http://dx.doi.org/10.1016/S0092-8674(00)81287-8
11. Hou XS, Melnick MB, Perrimon N. Marelle acts downstream of the Drosophila HOP/JAK kinase and encodes a protein similar to the mammalian STATs. Cell 1996; 84:411-9; PMID:8608595; http://dx.doi.org/10.1016/S0092-8674(00)81286-6
12. Wang Y, Levy DE. *C. elegans* STAT cooperates with DAF-7/TGF-beta signaling to repress dauer formation. Curr Biol 2006; 16:89-94; PMID:16401427; http://dx.doi.org/10.1016/j.cub.2005.11.061
13. Wang Y, Levy DE. *C. elegans* STAT: evolution of a regulatory switch. FASEB J 2006; 20:1641-52; PMID:16873887; http://dx.doi.org/10.1096/fj.06-6051com
14. Lewis RS, Ward AC. Conservation, duplication and divergence of the zebrafish stat5 genes. Gene 2004; 338:65-74; PMID:15302407; http://dx.doi.org/10.1016/j.gene.2004.05.012
15. Williams JG. The STAT proteins of *Dictyostelium*. In: Sehgal PB, Levy DE, Hirano T, eds. Signal Transducers and Activators of Transcription (STATs) Activation and Biology. Dordrecht: Kluwer Academic Publishers, 2003:105-21.
16. Kimura M. The neutral theory of molecular evolution. Cambridge Cambridgeshire; New York: Cambridge University Press, 1983.
17. Ohno S. Evolution by gene duplication. Berlin, New York: Springer-Verlag, 1970.
18. Hurles M. Gene duplication: the genomic trade in spare parts. PLoS Biol 2004; 2:E206; PMID:15252449; http://dx.doi.org/10.1371/journal.pbio.0020206
19. Zhang J. Evolution by gene duplication: an update. Trends Ecol Evol 2003; 18:292-8; http://dx.doi.org/10.1016/S0169-5347(03)00033-8

20. Dehal P, Boore JL. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol 2005; 3:e314; PMID:16128622; http://dx.doi.org/10.1371/journal.pbio.0030314
21. Madera M, Vogel C, Kummerfeld SK, Chothia C, Gough J. The SUPERFAMILY database in 2004: additions and improvements. Nucleic Acids Res 2004; 32(Database issue):D235-9; PMID:14681402; http://dx.doi.org/10.1093/nar/gkh117
22. Bates PA, Kelley LA, MacCallum RM, Sternberg MJ. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins 2001(Suppl 5):39-46; PMID:11835480; http://dx.doi.org/10.1002/prot.1168
23. Neculai D, Neculai AM, Verrier S, Straub K, Klumpp K, Pfitzner E, et al. Structure of the unphosphorylated STAT5a dimer. J Biol Chem 2005; 280:40782-7; PMID:16192273; http://dx.doi.org/10.1074/jbc.M507682200
24. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, et al. The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. PLoS Biol 2003; 1:E45; PMID:14624247; http://dx.doi.org/10.1371/journal.pbio.0000045
25. Eichinger L, Pachebat JA, Glöckner G, Rajandream MA, Sucgang R, Berriman M, et al. The genome of the social amoeba Dictyostelium discoideum. Nature 2005; 435:43-57; PMID:15875012; http://dx.doi.org/10.1038/nature03481
26. Blair JE, Shah P, Hedges SB. Evolutionary sequence analysis of complete eukaryote genomes. BMC Bioinformatics 2005; 6:53; PMID:15762985; http://dx.doi.org/10.1186/1471-2105-6-53
27. Bach EA, Perrimon N. Prime time for the *Drosophila* JAK/STAT pathway. In: Sehgal PB, Levy DE, Hirano T, eds. Signal Transducers and Activators of Transcription (STATs) Activation and Biology. Dordrecht: Kluwer Academic Publishers, 2003:87-104.
28. Barillas-Mury C, Han YS, Seeley D, Kafatos FC. Anopheles gambiae Ag-STAT, a new insect member of the STAT family, is activated in response to bacterial infection. EMBO J 1999; 18:959-67; PMID:10022838; http://dx.doi.org/10.1093/emboj/18.4.959
29. Christophides GK, Vlachou D, Kafatos FC. Comparative and functional genomics of the innate immune system in the malaria vector Anopheles gambiae. Immunol Rev 2004; 198:127-48; PMID:15199960; http://dx.doi.org/10.1111/j.0105-2896.2004.0127.x
30. Gorissen M, de Vrieze E, Flik G, Huising MO. STAT genes display differential evolutionary rates that correlate with their roles in the endocrine and immune system. J Endocrinol 2011; 209:175-84; PMID:21330334; http://dx.doi.org/10.1530/JOE-11-0033
31. Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, et al. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 2004; 431:946-57; PMID:15496914; http://dx.doi.org/10.1038/nature03025
32. Mulley J, Holland P. Comparative genomics: small genome, big insights. Nature 2004; 431:916-7; PMID:15496903; http://dx.doi.org/10.1038/431916a
33. Oates AC, Brownlie A, Pratt SJ, Irvine DV, Liao EC, Paw BH, et al. Gene duplication of zebrafish JAK2 homologs is accompanied by divergent embryonic expression patterns: only jak2a is expressed during erythropoiesis. Blood 1999; 94:2622-36; PMID:10515866
34. Blair JE, Hedges SB. Molecular phylogeny and divergence times of deuterostome animals. Mol Biol Evol 2005; 22:2275-84; PMID:16049193; http://dx.doi.org/10.1093/molbev/msi225
35. Huminiecki L, Wolfe KH. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res 2004; 14(10A):1870-9; PMID:15466287; http://dx.doi.org/10.1101/gr.2705204
36. Gilbert W. Why genes in pieces? Nature 1978; 271:501; PMID:622185; http://dx.doi.org/10.1038/271501a0
37. Liu M, Grigoriev A. Protein domains correlate strongly with exons in multiple eukaryotic genomes–evidence of exon shuffling? Trends Genet 2004; 20:399-403; PMID:15313546; http://dx.doi.org/10.1016/j.tig.2004.06.013
38. Patthy L. Protein evolution by exon-shuffling. New York, Austin.: Springer-Verlag; R.G. Landes, 1995.
39. Katju V, Lynch M. The structure and early evolution of recently arisen gene duplicates in the Caenorhabditis elegans genome. Genetics 2003; 165:1793-803; PMID:14704166
40. Katju V, Lynch M. On the formation of novel genes by duplication in the Caenorhabditis elegans genome. Mol Biol Evol 2006; 23:1056-67; PMID:16500928; http://dx.doi.org/10.1093/molbev/msj114
41. Darnell JE, Jr.. Phosphotyrosine signaling and the single cell:metazoan boundary. Proc Natl Acad Sci U S A 1997; 94:11767-9; PMID:9342310; http://dx.doi.org/10.1073/pnas.94.22.11767
42. Gao Q, Hua J, Kimura R, Headd JJ, Fu XY, Chin YE. Identification of the linker-SH2 domain of STAT as the origin of the SH2 domain using two-dimensional structural alignment. Mol Cell Proteomics 2004; 3:704-14; PMID:15073273; http://dx.doi.org/10.1074/mcp.M300131-MCP200
43. Kusserow A, Pang K, Sturm C, Hrouda M, Lentfer J, Schmidt HA, et al. Unexpected complexity of the Wnt gene family in a sea anemone. Nature 2005; 433:156-60; PMID:15650739; http://dx.doi.org/10.1038/nature03158
44. Koonin EV, Aravind L, Kondrashov AS. The impact of comparative genomics on our understanding of evolution. Cell 2000; 101:573-6; PMID:10892642; http://dx.doi.org/10.1016/S0092-8674(00)80867-3
45. Gough DJ, Corlett A, Schlessinger K, Wegrzyn J, Larner AC, Levy DE. Mitochondrial STAT3 supports Ras-dependent oncogenic transformation. Science 2009; 324:1713-6; PMID:19556508; http://dx.doi.org/10.1126/science.1171721
46. Gough DJ, Sehgal PB, Levy DE. Nongenomic functions of STAT3. In: Decker T, Müller M, eds. Jak-Stat Signaling: From Basics to Disease. Wien: Springer-Verlag Gmbh, 2011:in press.
47. Lee JE, Yang Y-M, Liang F-X, Gough DJ, Levy DE, Sehgal PB. STAT5-dependent nongenomic effects on Golgi/ER structure and function and their involvement in idiopathic pulmonary hypertension. Am J Physiol Cell Physiol 2012; 302:C804-20; PMID:22159083; http://dx.doi.org/10.1152/ajpcell.00379.2011
48. Wegrzyn J, Potla R, Chwae YJ, Sepuri NB, Zhang Q, Koeck T, et al. Function of mitochondrial Stat3 in cellular respiration. Science 2009; 323:793-7; PMID:19131594; http://dx.doi.org/10.1126/science.1164551
49. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997; 18:2714-23; PMID:9504803; http://dx.doi.org/10.1002/elps.1150181505