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 neo-functionalization. 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 stat4 genes in mammals likely arose from a duplication event in early Eutherian evolution, a period from about 310 Mya at the avian-mamalian 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 500 Mya, which include anterior-posterior and dorsal-ventral patterning, head differentiation and nervous systems. 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. 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. 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 gene regulatory networks (GRNs), which can be used repeatedly for many diverse functions throughout animal development processes to achieve necessary organism-specific phenotypic complexity. 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.

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. 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, organized in three tightly linked clusters on different chromosomes in mouse and human genome: stat1 and 4, stat2 and 6, stat3 and Stat5β. 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. Additional support for the gene duplication theory of stat gene evolution comes from the discovery of two stat5 genes in zebras. The existence of a more divergent STAT pathway in Drosophila melanogaster suggests that stat genes arose early in metazoan evolution, 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. 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
gene family from a wide range of organisms can reconstruct their evolutionary history. The availability of high-quality whole genome sequences from a variety of organisms, including Dicyostelium, 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 resulting in over 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. Analysis at the sta-1 locus revealed several partial duplications of the STAT gene, which yielded four annotated genes in Wormbase.org (release WS153), namely y51h4a.18, y51h4a.19, y51h4a.20 and y51h4a.21. 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.3 is a copy of y51h4a.5, 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. 2A), 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 y5354e12. 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). DNA sequencing revealed that the EST clone y5354e12 contains the identical f58e6.1 sequence, in addition to 3’ sequences originally annotated as f58e6.2. Based on the EST sequences of y5354e12, f58e6.2 was merged into f58e6.1 to form f58e6.1b, whereas the original f58e6.1 was named as isoform f58e6.1a (Wormbase). However, our sequence analysis showed that y5354e12 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.](image) Corresponding exons are linked with dashed arrows. Scale: 1,000 bp.
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 revealed that among all the known and predicted peptide sequences, the best matches for F58E6.1 are STATs. Conversely, human STAT5A contains four domains, SH2 (E-value 8.8e-38), which is consistent with our previous studies. For comparison, human STAT5A contains four domains, SH2 (E-value 8.8e-38), which is consistent with our previous studies.12,13 For comparison, human STAT5A contains four domains, SH2 (E-value 8.8e-38), which is consistent with our previous studies.12,13 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, extracting coordinates for the unphosphorylated mouse STAT5A crystal structure. 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. 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. However, among all the known and predicted peptide sequences, the best matches for F58E6.1 are STATs. Conversely, among all the known and predicted peptide sequences, the best matches for F58E6.1 are STATs. Conversely, among all the known and predicted peptide sequences, the best matches for F58E6.1 are STATs. Conversely, among all the known and predicted peptide sequences, the best matches for F58E6.1 are STATs. Conversely, among all the known and predicted peptide sequences, the best matches for F58E6.1 are STATs. Conversely, among all the known and predicted peptide sequences, the best matches for F58E6.1 are STATs.

Table 1. Original and duplicated intron-exon boundaries at the C. elegans sta-7 locus. The length and percentages of identity of duplicated fragments were calculated for exons and flanking sequences.

| Note | intron— | EXON | intron— |
|------|---------|------|---------|
| Exon 1 | original | ...attttcagATACATG... | 203 bp/96% | 310 bp/99% |
| | duplicate | ...attttcagATACATG... | 172 bp/76% |
| Exon 2 | original | ...tttttcagCCGTAACA... | 40 bp/100% | 128 bp/98% |
| | duplicate | ...tttttcagCCGTAACA... | 117 bp/89%* |
| Exon 3 | original | ...tttttcagCCGTAACA... | 8 bp/100% | 238 bp/100% |
| | duplicate | ...tttttcagCCGTAACA... | 238 bp/79% |
| Exon 4 | original | ...tttttcagCCGTAACA... | 8 bp/100% | 238 bp/100% |
| | duplicate | ...tttttcagCCGTAACA... | 238 bp/79% |
| Exon 5 | original | ...tttttcagCCGTAACA... | 157 bp/90% | 135 bp/99% |
| | duplicate | ...tttttcagCCGTAACA... | 42 bp/98%* |
| Exon 6 | original | ...tttttcagCCGTAACA... | 42 bp/98%* | 123 bp/100% |
| | inverse duplicate | ...tttttcagCCGTAACA... | 223 bp/96% |
| Exon 7 | original | ...tttttcagCCGTAACA... | 42 bp/98%* | 123 bp/100% |
| | inverse duplicate | ...tttttcagCCGTAACA... | 223 bp/96% |
| Exon 8* | original | ...tttttcagCCGTAACA... | NA/NA | 84 bp/100% |
| | inverse duplicate | ...tttttcagCCGTAACA... | NA/NA |

*Intron 3 is 164 bp long and is not duplicated in its entirety. †Intron 6 is 124 bp long and is not duplicated in its entirety with exon 6 and 7. ‡Intron 7 is 233 bp, and is duplicated in its entirety 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).

www.landesbioscience.com JAK-STAT 25

© 2012 Landes Bioscience.
of animals, fungi and plants. Therefore, the discovery of STAT family members in the slime mold *Dictyostelium discoideum* 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), 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 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.
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. A canonical STAT signaling pathway, analogous to that in mammals, exists in the fruit fly Drosophila melanogaster. 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.
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: ensangg00000006157) 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: ensangg00000001540), previously found to be involved in immune responses to bacterial infections, is located at the other end of chromosome 3L, encoded by a single exon. 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 pseudo-genes and disappear, 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.26

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 STATs are most similar to STAT5. This similarity is in contrast with STATs in protostomes, which formed a sister clade to both STATs 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 ascidians.

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.27 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.

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 tetrapod evolution after divergence from the tetrapods, 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 JAK2.29 Furthermore, they were likely lost in the pufferfish lineage, presumably in favor of a much more compact genome.

In addition to these 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 invented gene duplication event followed by rapid sequence diversification.

STATs in amphibians and birds. Within the tetrapods, amphibians diverged from anamniotes, which include birds and mammals, about 370 Mya, and birds and mammals diverged around 310 Mya. 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 STAT5, with over 95% protein sequence identity. The two stat5s 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.

Genome analysis of birds suggests five members, STAT1, 3, 4 and two STAT5s, and these data are 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 stat5a gene is flanked on the same DNA strand by stat3 upstream and igf1 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, and stat3 and the two stat5s on chromosome 11 in mice.30 Similar linkages also 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
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

| Genus                  | Linear position | Linear position | Linear position | Linear position |
|------------------------|-----------------|-----------------|-----------------|-----------------|
| Monodelphis domestica  | Scaffold 2      | Scaffold 18     | N/A             | Sca. 303        |
| B. taurus              | ChrUn. 103      | Chromo. 19      | Chromo. 10      |                 |
| 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. 28      | Chromo. 17      | Chromo. 12      |                 |
| Rattus norvegicus*     | Chr. 9          | Chromo. 10      | Chromo. Seven   |                 |
serve some unknown critical function, which could provide strong pressure for STAT 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. The tandem nature of these sta-1 duplicates and the presence of highly identical, yet partial intrinsic sequences exclude the possibility of generation by the latter two mechanisms. Therefore, it is possible that a recent, unequal crossing over led to the 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. 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. 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. 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 intrinsic 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 k10q2.5.

**STAT origin and early evolution.** The discovery of STAT signaling in Dicyostelium extended this intercellular phosphorylase pathway beyond the Metazoa and raised the possibility of a single origin of STATs during the single cell-metazoan evolutionary process. However the controversial phylogenetic status of Dicyostelids 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 Dicyostelium proteins are more similar to human orthologs than those of yeasts, phylogenetic analysis suggested the divergence occurred before that of fungi, which don’t employ SH2-mediated phosphorylase signaling. If there existed a single STAT ancestor after the plant-animal split, then it was lost along with the phosphorylase 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 mycetozoans through horizontal gene transfer, consistent with the observation that many Dicyostelium proteins are more similar to human orthologs, in contrast to the divergence of yeast and human orthologs. It is likely that the common ancestors of the so-called “crown eukaryotes” are single-cell organisms with very diverse genomes.11-13 It is thought an STAT was already present 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 phosphorylase-based signaling pathways are considered to be a tool specific for intercellular communications within a multicellular organism,14 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 multicellularity. Further evidence for an independent STAT origin in mycetozoa, a distant ortholog of the STATs in C. elegans was found in Saccharomyces and Arabidopsis, suggesting its more ancient origin before the plant-animal split. Thus, the linker-SH2 domain may be the original evolutionary foundation upon which STATs later evolved through domain accretion in the mycetozoans and metazoans.

**STAT evolution in the early metazoa.** Since the mycetozoans are no longer considered a direct ancestor of metazoans,15 another ancestral STAT likely arose early in metazoan evolution after the fungi-animal divergence 1.5 billion years ago. As nematodes diverged from other metazoa about 1.2 billion years ago,25,26 characterizations of the STAT in C. elegans provides insight into this STAT ancestor.16-17 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, hydrids and jellyfishes, are likely diverged from the metazoan lineage shortly before the rise of bilaterians. A study of STAT signaling in the sea anemone *Nemastoma helianthum* 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 proto-stomes and deuterostomes. Interestingly, a search of the fresh-water 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 STATs 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 out 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 *Echinodermeta* 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 ascidian lineage. Though genome sequencing of the lancelet 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 to 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 STAT5. As their respective sequence identities to other human STATs are significantly lower, they likely represent two founding members of vertebrate STATs. Specifically, through whole genome duplications and subsequent survivals, ancestral stat1 likely produced the present-day stat1a and stat1b genes whereas ancestral stat5 yielded stat5a and two stat5b genes.

Dynamic STAT evolution by duplications. Gene duplications are proposed to be a major driving force for genomic and organismal complexities during evolution. Our comparative genomics of the STAT family of transcription factors has provided strong and detailed evidence for the gene duplication theory. Despite its rarity, 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. 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. 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.

The tightly linked chromosomal clusters of stat genes in many of the vertebrate genomes (Table 2) suggest that gene duplication by unequal chromosome 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 stat5 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, the copy has 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 selective 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 fshb-1 gene. Protein sequence analysis suggested that it is a likely 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 two novel stat genes, one of which was likely duplicated from stat4 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, the copy has completely disrupted the ORF, thus rendering it a pseudogene.

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, documenting the parsimony of evolution.
Methods

Cloning and characterization of fs86E.1. A Blast search using human STAT1 of C. elegans protein database yielded another significant hit FS86E.1, in addition to STA-1. The matching EST clone y354e12 (gift of Y. Kohara, NIG, Japan) was used to screen 5 × 10⁴ colonies from a mixed-stage C. elegans λgt11 cDNA library (gift of P. O’Keeffe, IL). Eight positive colonies were isolated and grouped into three categories based on insert sizes, ~1 kb (4 colonies), ~1.4 kb (2 colonies) and ~1.8 kb (2 colonies). DNA sequencing revealed that these three groups shared the same 5’ end sequences. Clone #1 which has a ~1.8 kb insert and EST clone y354e12 were then fully sequenced. Clone #1 appeared to be a full-length cDNA clone, as it started with a trans-pairing leader SL1 sequence and contained a single ORF. Clone y354e12 contained identical clone #1 sequences except the 5’ SL1 and the beginning ORF sequences, plus ~800 bp extra 3’ sequences. Sequence analysis suggested y354e12 to be an open transcript as it had two ORFs. The 5’ ORF was identical to clone #1, but the 3’ ORF matched with FS86E.2, which was originally annotated as a single gene but later merged with annotated FS86E.1 into an isoform F58E6.1a, based on the end sequences of F58E6.1. Clone #1 was FLAG-tagged and biochemically characterized as described previously.13

Domain prediction and structure modeling. Predicted FS86E.1 protein sequences were used to search the Superfamily database,21 which is a collection of Hidden Markov Models (HMMs) that describe structural classes at www.supfam.org. For 3D structure modeling, the homology-based structure prediction program 3D-JIGSAW22 (http://www.bmm.icn.rwth-aachen.de/services/3djigasw) was used to first identify the mouse unphosphorylated STAT5A crystal structure as the template for building a potential structure model, which was further fitted to STAT5A crystal structure using SwissPDB Viewer 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, dicytosteliumBase (www.dicytosteliumbase.org) for the slime mold Dicytostelium; Ensembl (www.ensembl.org) for fruit fly D. melanogaster; honeybee A. mellifera; malaria mosquito A. gambiae; ascidians C. intestinalis and S. purpuratus; 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. GrankB for genome sequences of the California purple sea urchin S. purpuratus, pelagic tunicate O. dioica, and EST sequences of the freshwater polyp H. magnipapillata, the amphibian B. florisana, 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 grant from the NIH (AI28990, U54AI057518) 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. Nature 2006; 438:746-53. doi:10.1038/nature04293
2. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Cox T,デルナム D, Durbin R, Elnitski L, Gibbons A, Guigo R, Haakonsen D, Hall N, Himananthena A, Hunt S, Jones M, Kitchen S, Kettani M, Kucukural A, Liu Y, McWilliam H, Myers C, Zabeau M, Zeger S, Devulapala P, Gilbert J, Gingeras T, Grocock R, Haussler D, Hibbs J, Hotz H, Ito T, Jordan S, Kong X, Kulp D, Lee C, Minneci C, Morris W, Ondov B, Orsini A, Pachter L, Parra E, Pertierra K, Pimentel M, Powell N, Quackenbush J, Ratan S, Rosenbloom K, Searls D, Thrash K, Venter J, Vershon K, Watermen S, Wilson RK, Word C, Zody M. Ensembl (release 22) provides enhanced analysis and interpretation of human genome sequence. Genome Res. 2006;16:193-203. doi:10.1101/gr.4636806
3. Gerhart J. 1998 Warkany lecture: signaling pathways and the evolution of animal body plans. Science 2006; 310:796-800. PMID:16408193. doi:10.1126/science.1113869
4. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Cooper G, Duarte-Carvajalino J, Durbin R, Cattoretti G, Elnitski L, Flicek P, Goekdel DG, Green EK, Haakonsen D, Holder A, Hunt S, Jordan S, Kong X, Kulp D, Parra E, Pertierra K, Pimentel M, Powell N, Quackenbush J, Ratan S, Rosenbloom K, Searls D, Thrash K, Venter J, Word C, Zody M. Ensembl (release 23) provides enhanced analysis and interpretation of human genome sequence. Genome Res. 2006;16:204-9. PMID:16408193. doi:10.1101/gr.4279506
5. Davidson EH, Erwin DH. Gene regulatory networks and the evolution of animal body plans. Science 2006; 310:796-800. PMID:16408193. doi:10.1126/science.1113869
6. Gojobori T. 1998 Workshop lecture: signaling pathways in development. Taxon 1999; 48:265-39. PMID:10387879. doi:10.2307/2335170
7. Bhan A, Sambamurti K. 1998 Workshop lecture: signaling pathways in development. Taxon 1999; 48:265-9. PMID:10387879. doi:10.2307/2335170
8. Copeland NG, Gilbert DJ, Schindler C, Zhong Z, Wu S, Darnell JE, Jr., et al. Distribution of the mammalian Stat gene family in mouse chromosomes. Genome Res. 1995; 5:235-42. PMID:8490877. doi:10.1101/gr.8490877
9. Bhan A, Sambamurti K. 1998 Workshop lecture: signaling pathways in development. Taxon 1999; 48:265-39. PMID:10387879. doi:10.2307/2335170
10. Yan R, Small S, Desplan C, Dearolf CR, Darnell JE, Jr. Downstream of the Drosophila HOP/JAK kinase and transcription. Cell 1996; 84:331-4; PMID:8608586; http://dx.doi.org/10.1016/S0092-8674(00)81277-5
11. Hou XS, Melnick MB, Perrimon N. Marelle acts as a Stat-regulator in Drosophila development. Cell 1996; 84:411-9; PMID:8608595; http://dx.doi.org/10.1016/S0092-8674(00)81286-6
12. Wang Y, Levy DE. STAT: evolution of a transcriptional regulatory switch. FASEB J 2006; 20:1641-52; PMID:16873943; http://dx.doi.org/10.1096/fj.06-6051com
13. Wang Y, Levy DE. STAT: evolution of a transcriptional regulatory switch. FASEB J 2006; 20:1641-52; PMID:16873943; 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 JH. The STAT protein of Drosophila. In: Schibler P, Levy DE, Hsuie 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: Cambridge University Press, 1983.
17. Ohta S. Evolution by gene duplication. Berlin, New York: Springer-Verlag, 1979.
18. Hurles M. Gene duplication: the genomic trade in spare parts. Philos Trans R Soc B 2006; 361:281-9. PMID:16523449. http://dx.doi.org/10.1098/rstb.2005.1962
19. Zhang J. Evolution by gene duplication: an update. Trends Ecol Evol 2005; 20:292-8. http://dx.doi.org/10.1016/j.tree.2005.05.008
Drosophila

24. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, et al. 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

23. Bates PA, Kelley LA, MacCallum RM, Sternberg MJ. Enhancement of protein modeling by human intervention in applying the automated program SOBAS and 3D-PSSM. Proteins 2005;Suppl 8:99-46; PMID:16846980; http://dx.doi.org/10.1002/prot.20308

22. Madera M, Vogel C, Kummerfeld SK, Chothia C, et al. Structure of the unphosphorylated, transcriptionally active and functional form of eukaryotic STAT5a dimer. J Biol Chem 2005; 280:40782-7; PMID:16192273; http://dx.doi.org/10.1074/jbc.M507662200

21. Dehal P, Boore JL. Comparative genomics of the innate immune system in the malaria vector Anopheles gambiae. EMBO J 2003; 22:1056-7; PMID:12846193; http://dx.doi.org/10.1093/emboj/cdg117

20. Gorissen M, de Vrieze E, Flik G, Huising MO. STAT5 is not expressed during the embryonic development of the nematode Caenorhabditis elegans. Proc Natl Acad Sci USA 1997; 94:11767-9; PMID:9342310; http://dx.doi.org/10.1073/pnas.94.22.11767

19. Gough J. The SUPERFAMILY database in 2004: structural alignment. Nucleic Acids Res 2004; 32(Database issue):D235-9; PMID:14681402; http://dx.doi.org/10.1093/nar/gkh117

18. Kusserow A, Pang K, Sturm C, Hovanes M, Lortz J, et al. Unexpected complexity of the Wnt gene family in a nearer ancestor. Nature 2005; 431:156-60; PMID:15830582; http://dx.doi.org/10.1038/nature02902

17. Darnell JE, Jr.. Phosphotyrosine signaling and the single-catalytic domain. Proc Natl Acad Sci U S A 1997; 94:11767-9; PMID:9342310; http://dx.doi.org/10.1073/pnas.94.22.11767

16. Lee JE, Yang Y-M, Liang F-X, Gough DJ, Levy DE, et al. Comparative genomics on our understanding of evolution. Cell 2009; 136:573-9; PMID:19606213; http://dx.doi.org/10.1016/j.cell.2009.02.040

15. Patthy L. Protein evolution by exon-shuffling. New York: Oxford University Press, 2003:1-26.

14. M507662200

13. Stein LD, Acevedo-Arozena E, Blasiar D, Bao Z, Lowary PT, et al. The genome sequence of Caenorhabditis elegans. Science 2003; 301:143-54; PMID:12870262; http://dx.doi.org/10.1126/science.1082791

12. Madera M, Vogel C, Kummerfeld SK, Chothia C, et al. Structure of the unphosphorylated, transcriptionally active and functional form of eukaryotic STAT5a dimer. J Biol Chem 2005; 280:40782-7; PMID:16192273; http://dx.doi.org/10.1074/jbc.M507662200

11. Madera M, Vogel C, Kummerfeld SK, Chothia C, et al. Structure of the unphosphorylated, transcriptionally active and functional form of eukaryotic STAT5a dimer. J Biol Chem 2005; 280:40782-7; PMID:16192273; http://dx.doi.org/10.1074/jbc.M507662200

10. Madera M, Vogel C, Kummerfeld SK, Chothia C, et al. Structure of the unphosphorylated, transcriptionally active and functional form of eukaryotic STAT5a dimer. J Biol Chem 2005; 280:40782-7; PMID:16192273; http://dx.doi.org/10.1074/jbc.M507662200

9. Madera M, Vogel C, Kummerfeld SK, Chothia C, et al. Structure of the unphosphorylated, transcriptionally active and functional form of eukaryotic STAT5a dimer. J Biol Chem 2005; 280:40782-7; PMID:16192273; http://dx.doi.org/10.1074/jbc.M507662200

8. Madera M, Vogel C, Kummerfeld SK, Chothia C, et al. Structure of the unphosphorylated, transcriptionally active and functional form of eukaryotic STAT5a dimer. J Biol Chem 2005; 280:40782-7; PMID:16192273; http://dx.doi.org/10.1074/jbc.M507662200

7. Madera M, Vogel C, Kummerfeld SK, Chothia C, et al. Structure of the unphosphorylated, transcriptionally active and functional form of eukaryotic STAT5a dimer. J Biol Chem 2005; 280:40782-7; PMID:16192273; http://dx.doi.org/10.1074/jbc.M507662200

6. Madera M, Vogel C, Kummerfeld SK, Chothia C, et al. Structure of the unphosphorylated, transcriptionally active and functional form of eukaryotic STAT5a dimer. J Biol Chem 2005; 280:40782-7; PMID:16192273; http://dx.doi.org/10.1074/jbc.M507662200

5. Madera M, Vogel C, Kummerfeld SK, Chothia C, et al. Structure of the unphosphorylated, transcriptionally active and functional form of eukaryotic STAT5a dimer. J Biol Chem 2005; 280:40782-7; PMID:16192273; http://dx.doi.org/10.1074/jbc.M507662200

4. Madera M, Vogel C, Kummerfeld SK, Chothia C, et al. Structure of the unphosphorylated, transcriptionally active and functional form of eukaryotic STAT5a dimer. J Biol Chem 2005; 280:40782-7; PMID:16192273; http://dx.doi.org/10.1074/jbc.M507662200

3. Madera M, Vogel C, Kummerfeld SK, Chothia C, et al. Structure of the unphosphorylated, transcriptionally active and functional form of eukaryotic STAT5a dimer. J Biol Chem 2005; 280:40782-7; PMID:16192273; http://dx.doi.org/10.1074/jbc.M507662200

2. Madera M, Vogel C, Kummerfeld SK, Chothia C, et al. Structure of the unphosphorylated, transcriptionally active and functional form of eukaryotic STAT5a dimer. J Biol Chem 2005; 280:40782-7; PMID:16192273; http://dx.doi.org/10.1074/jbc.M507662200

1. Madera M, Vogel C, Kummerfeld SK, Chothia C, et al. Structure of the unphosphorylated, transcriptionally active and functional form of eukaryotic STAT5a dimer. J Biol Chem 2005; 280:40782-7; PMID:16192273; http://dx.doi.org/10.1074/jbc.M507662200

www.andebioscience.com JAK-STAT 33