# Supporting Information

## Pincus *et al.* 10.1073/pnas.0803161105

### SI Methods

**Genomic Datasets and Domain Prediction Algorithm.** To explore the evolution of phospho-tyrosine signaling, we investigated the domain usage, pairwise domain combinations, and architectures of proteins containing tyrosine kinase (TyrK), tyrosine phosphatase (PTP), and SH2 domains across different eukaryotic lineages. For all of these analyses, we employed the Simple Modular Architecture Tool (1) to identify proteins that include TyrK, PTP or SH2 domains as predicted from genomic and transcriptomic databases. With the exception of *Nematostella vectensis* and *Monosiga brevicollis*, all of the genomes we used are available through genomic SMART and are the current filtered, nonredundant working drafts. We obtained the *N. vectensis* predicted protein set from StellaBase (2), and the *M. brevicollis* "best proteins" filtered set from the JGI database (genome.jgi-psf.org/Monbr1/Monbr1.download.ftp.html).

**Quantification of Phospho-Tyrosine Signaling Domains and Proteins.** With the SMART resource, we determined the total number of of TyrK, PTP, and SH2 domains and the number of proteins contiaining each of these domains for all organisms in the SMART genomic database, *N. vectensis* and *M. brevicollis*. Fig. S1 shows all of the values this analysis yielded, and figure 1b shows the number of proteins with each domain for selected genomes as bar graphs next to the species name. For the sake of consistency, we used the publicly available filtered genomic datasets for this quantitative analysis, which, in the case of *M. brevicollis*, does not contain hand-curated, ab-initio models. Thus, the values we obtain represent conservative estimates; the true protein and domain numbers for these genomes are likely to be slightly higher, and may differ slightly from earlier analyses on individual species. The arguments resented here, however, do not depend on an absolutely precise domain count. We constructed the tree in Fig. 1*b* with the user function of the Interactive Tree of Life (3).

**Determination of Pairwise Domain Combinations.** Further investigating the evolution of functional usage of the phsopho-tyrosine (P-Tyr) signaling machinery, we analyzed pairwise combinations of TyrK, PTP, and SH2 domains (query domains). In this analysis, we looked for other protein domains that occur in the same protein (ORF) as the query domains. From our SMART-predicted sets of proteins with TyrK, PTP, and SH2 domains across eukaryotes we searched the sequences for all other domains predicted by either SMART or Pfam (4) algorithms. We removed redundant domains predicted by both algorithms and the N- and C-terminal specifications designated by Pfam. We also merged the representation of the domains listed in Table S1 to simplify Fig. 2*b*. The raw data is available upon request. For each query domain we clustered domain combinations as shown in Fig. 2*b* based on the sets of species in which the combinations cooccur. A schematic for how the analysis is conceptually performed is shown in Fig. 2*a*. Those domain combinations that exist in both *M. brevicollis* and metazoans we have designated "shared core" to indicate that they exist in both unicellular and multicellular organisms.

**Analysis of Expansion and Divergence of Domain Architectures.** We put our domain combination analysis into a more biological context by examining domain architecture of complete proteins containing each of the P-Tyr query domains. We used SMART to predict domain architectures of selected proteins with domain combinations that fit into each region of the Venn diagram in Fig. 3*a*. The proteins are scaled relative to each other. Table S2 lists the protein names (or model identification numbers) and organism of origin. From the results of domain combination and architecture analysis, we observed both divergence and expansion from the shared core in the usage of P-Tyr signaling domains as we depict in Fig. 3*b*.

1. Letunic I, *et al.* (2006) SMART 5: Domains in the context of genomes and networks. *Nucleic Acids Res* 34:D257–D260.
2. Sullivan JC, *et al.* (2006) StellaBase: The Nematostella vectensis genomics database. *Nucleic Acids Res* 34:D495–D499.
3. Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics* 23:127–128.
4. Finn RD, *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res* 36:D281–D288.

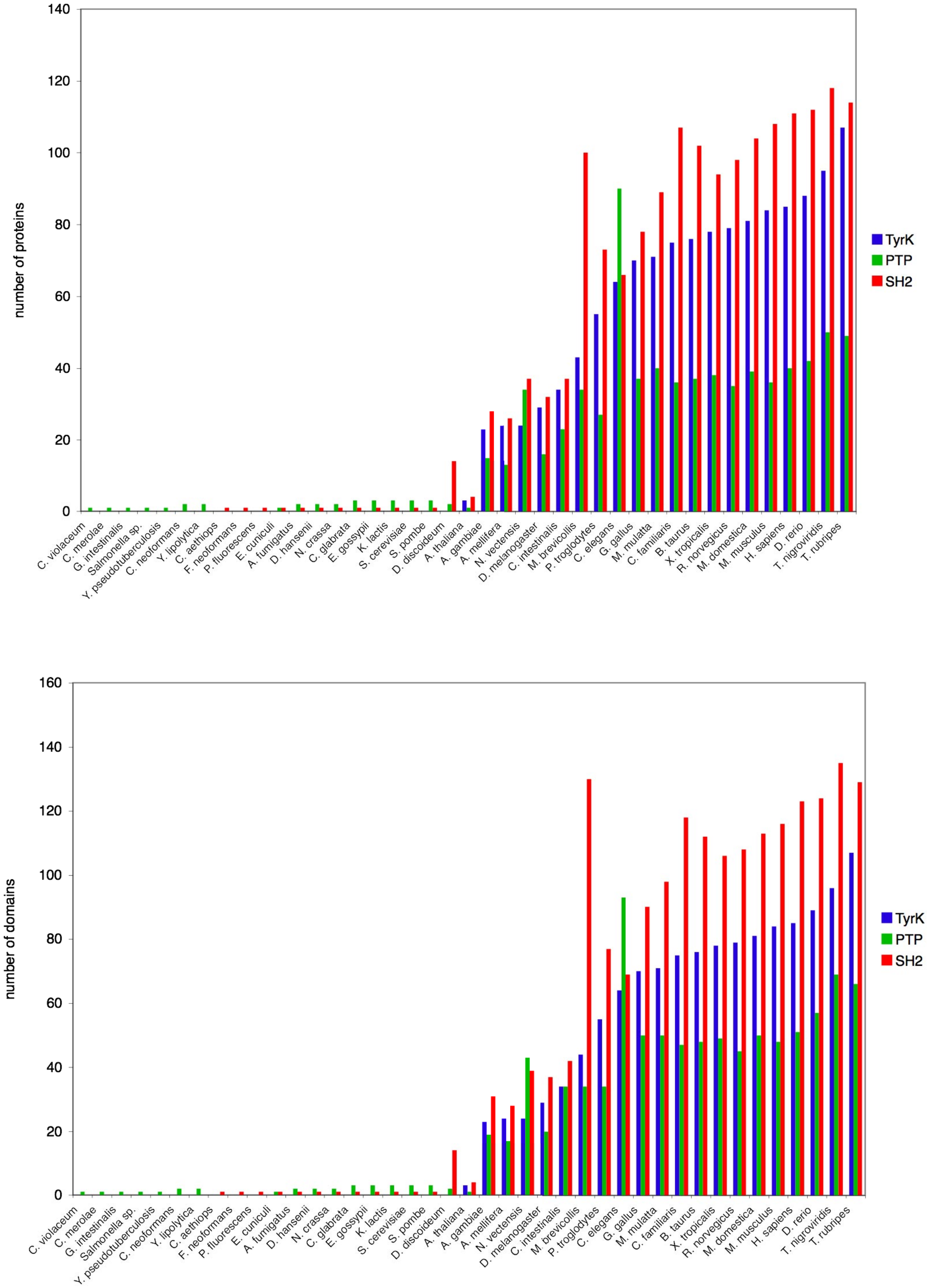


**Fig. S1.** Complete protein number and domain counts for TyrK, PTP, and SH2 domains across all organisms in dataset. The organisms are in ascending order, first by number of TyrK proteins/domains, followed by SH2 and last PTPs.

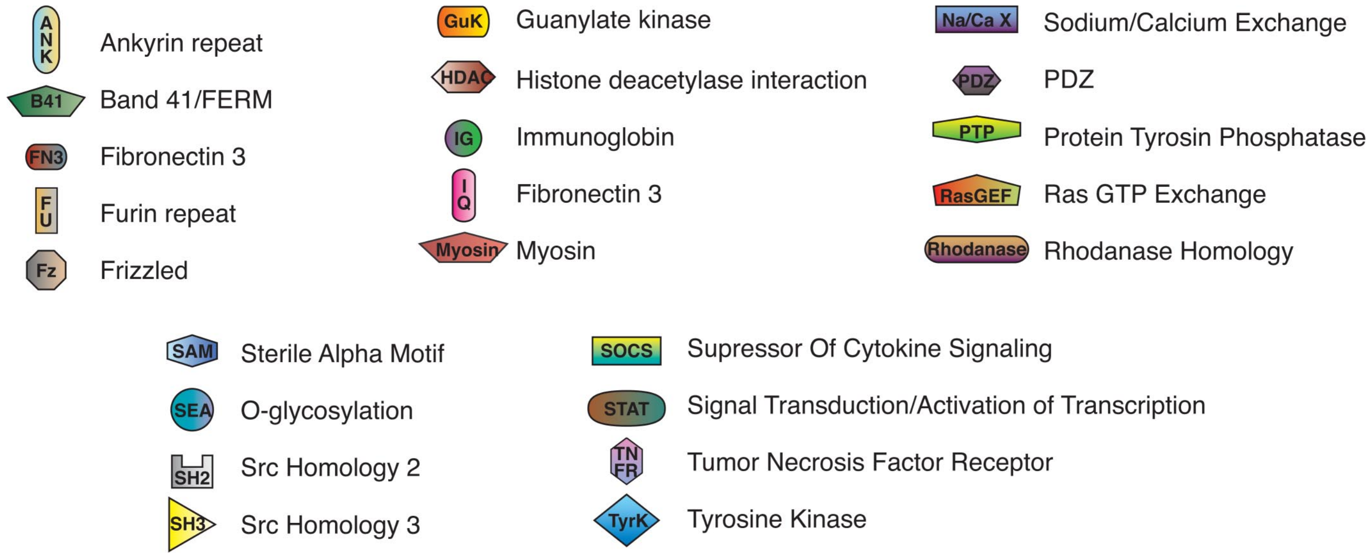
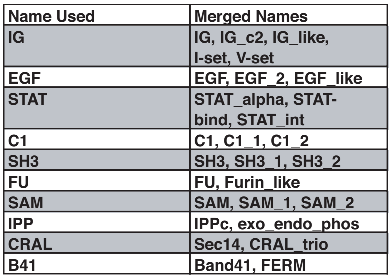


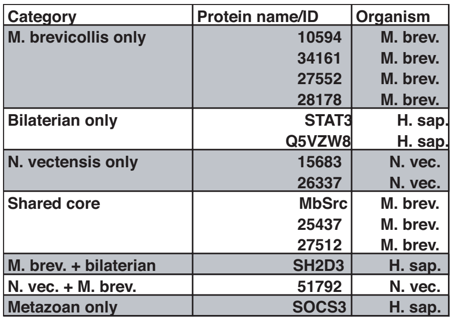
**Fig. S2.** Definitions of domains depicted in Fig. 3.

### Table S1. Domains merged into single classifier domain



| Name Used | Merged Names |
|---|---|
| IG | IG, IG_c2, IG_like, I-set, V-set |
| EGF | EGF, EGF_2, EGF_like |
| STAT | STAT_alpha, STAT-bind, STAT_int |
| C1 | C1, C1_1, C1_2 |
| SH3 | SH3, SH3_1, SH3_2 |
| FU | FU, Furin_like |
| SAM | SAM, SAM_1, SAM_2 |
| IPP | IPPc, exo_endo_phos |
| CRAL | Sec14, CRAL_trio |
| B41 | Band41, FERM |

**Table S2. Names and identification number of protein architectures depicted in Fig. 3**



| Category | Protein name/ID | Organism |
|---|---|---|
| M. brevicollis only | 10594 | M. brev. |
| | 34161 | M. brev. |
| | 27552 | M. brev. |
| | 28178 | M. brev. |
| Bilaterian only | STAT3 | H. sap. |
| | Q5VZW8 | H. sap. |
| N. vectensis only | 15683 | N. vec. |
| | 26337 | N. vec. |
| Shared core | MbSrc | M. brev. |
| | 25437 | M. brev. |
| | 27512 | M. brev. |
| M. brev. + bilaterian | SH2D3 | H. sap. |
| N. vec. + M. brev. | 51792 | N. vec. |
| Metazoan only | SOCS3 | H. sap. |

Proteins are listed from the top down.