Evolution of Functional Diversity in the Holozoan Tyrosine Kinome

Wayland, Yeung,[†,1] Annie Kwon,[†,1] Rahil Taujale,[1] Claire Bunn,[2] Aarya Venkat,[3] and Natarajan Kannan*[,1,3]

[1]Institute of Bioinformatics, University of Georgia, Athens, GA, USA

[2]Department of Genetics, University of Georgia, Athens, GA, USA

[3]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA, USA

[†]These authors contributed equally to this work.

****Corresponding author:** E-mail: nkannan@uga.edu.

Associate editor: Naruya Saitou

Abstract

The emergence of multicellularity is strongly correlated with the expansion of tyrosine kinases, a conserved family of signaling enzymes that regulates pathways essential for cell-to-cell communication. Although tyrosine kinases have been classified from several model organisms, a molecular-level understanding of tyrosine kinase evolution across all holozoans is currently lacking. Using a hierarchical sequence constraint-based classification of diverse holozoan tyrosine kinases, we construct a new phylogenetic tree that identifies two ancient clades of cytoplasmic and receptor tyrosine kinases separated by the presence of an extended insert segment in the kinase domain connecting the D and E-helices. Present in nearly all receptor tyrosine kinases, this fast-evolving insertion imparts diverse functionalities, such as post-translational modification sites and regulatory interactions. Eph and EGFR receptor tyrosine kinases are two exceptions which lack this insert, each forming an independent lineage characterized by unique functional features. We also identify common constraints shared across multiple tyrosine kinase families which warrant the designation of three new subgroups: Src module (SrcM), insulin receptor kinase-like (IRKL), and fibroblast, platelet-derived, vascular, and growth factor receptors (FPVR). Subgroup-specific constraints reflect shared autoinhibitory interactions involved in kinase conformational regulation. Conservation analyses describe how diverse tyrosine kinase signaling functions arose through the addition of family-specific motifs upon subgroup-specific features and coevolving protein domains. We propose the oldest tyrosine kinases, IRKL, SrcM, and Csk, originated from unicellular premetazoans and were coopted for complex multicellular functions. The increased frequency of oncogenic variants in more recent tyrosine kinases suggests that lineage-specific functionalities are selectively altered in human cancers.

Key words: **cancer mutations, Bayesian classification, kinase insert domain, protein domain and taxonomic conservation analysis, bidirectional signaling, proteomics.**

Introduction

Tyrosine kinases propagate cellular signals through the phosphorylation of tyrosine residues on protein substrates. Forming a monophyletic group within the larger protein kinase superfamily, tyrosine kinases diverged from serine–threonine kinases prior to the emergence of opisthokonts (animals and fungi) (Suga et al. 2012; Hunter 2014), which are estimated to be over a billion years old (Parfrey et al. 2011). Although their detection in unicellular premetazoans, such as choanoflagellates and filastereans has indicated the fundamental roles of tyrosine kinases in the evolution of multicellularity (King et al. 2008; Miller 2012; Tong et al. 2017), their subsequent expansion throughout metazoan evolution is associated with the evolution of diverse metazoan body plans and complex biological systems, such as the nervous, vascular, and immune systems (Liu et al. 2011; Miller 2012). Given the vast diversity of tyrosine kinases, the diversification events that gave rise to the functional repertoire of tyrosine kinases and the evolutionary timeline of such events has not been fully explored.

A classification of the protein kinome into evolutionarily and functionally related families (here on referred to as the KinBase classification) was achieved two decades ago following the sequencing and comparative genomic analyses of model organism genomes including human (Manning, Whyte, et al. 2002), mouse (Caenepeel et al. 2004), sea urchin (Bradham et al. 2006), fly (Manning, Plowman, et al. 2002), nematode (Plowman et al. 1999), sponge (Srivastava et al. 2010), choanoflagellate (King et al. 2008), and yeast (Manning, Plowman, et al. 2002). In addition, tyrosine kinases can be broadly classified as cytoplasmic or receptor tyrosine kinases based on the presence of transmembrane and extracellular ligand-binding domains; however, unlike kinase groups and families defined in the KinBase classification, cytoplasmic and receptor tyrosine kinases do not form monophyletic clades in the kinome tree because receptor tyrosine kinases are believed to have independently emerged multiple times throughout tyrosine kinase evolution (Robinson et al. 2000; Suga et al. 2012). The KinBase classification has subsequently become a foundation for comparative studies to study the conservation and divergence of kinase sequence, structure, and function. For example, previous studies of the patterns of sequence conservation and variation across kinase families and groups have provided important insights into the unique regulatory spine of tyrosine kinases relative to serine/threonine kinases (Oruganty et al. 2013; Mohanty et al. 2016), as well as into regulatory mechanisms that evolved uniquely in the EGFR (Mirza et al. 2010), Eph (Kwon et al. 2018), and Tec (Amatya et al. 2019) families of tyrosine kinases.

In addition to the uniquely evolved features across different tyrosine kinase families, similarities across some tyrosine kinase families have also been noted. For example, the recently termed "Src module," which consists of a tyrosine kinase domain and N-terminal SH3 and SH2 domains, is found across the Src, Abl, Tec, and Csk families, and structural and solution studies have determined that a similar autoinhibitory configuration of the Src module is shared across members of the Src, Abl, and Tec families (Shah et al. 2018). Because previous classifications of protein kinases were determined by analyzing branching points in phylogenetic trees of diverse kinase domain sequences, along with analysis of common domain structures and known biological functions, the existence of evolutionarily-related and functionally-relevant higher-order groupings of families within the kinase classification has not yet been systematically explored.

Here, we determine a novel hierarchical, constraint-based classification of the tyrosine kinome that newly identifies three evolutionary subgroupings of tyrosine kinase families based on the selective conservation of sequence motifs in the kinase domain, which encode common autoinhibitory conformations. In addition, we illustrate an evolutionary timeline of how unique kinase functions have expanded on shared subgroup-specific features through duplication events, selection of family-specific motifs, and domain shuffling to give rise to the vast repertoire of tyrosine kinase signaling observed throughout metazoans. A closer examination of tyrosine kinase phylogeny in light of constraint-based tyrosine kinase subgroups reveals new insights into the evolutionary conservation or divergence of subgroups, as well as the unique signaling features that may have emerged from three separate monophyletic clades of receptor tyrosine kinases. In particular, we note the early emergence of two major clades of holozoan tyrosine kinases distinguished by the presence (or absence) of an insert between the αD and αE helices of the kinase domain, where tyrosine kinases containing the insert comprise the majority of metazoan receptor tyrosine kinases. Our classification of the tyrosine kinome and the approach used in this study set a new precedent for the classification and evolutionary study of protein kinases and other large protein families.

Results

A Hierarchical, Constraint-Based Classification of the Holozoan Tyrosine Kinome Reveals New Tyrosine Kinase Subgroups

To generate a comprehensive classification of the holozoan tyrosine kinome, we generated a multiple sequence alignment of 44,639 tyrosine kinase sequences spanning 586 species (see Materials and Methods for details). We then used the Bayesian Partitioning with Pattern Selection (BPPS) algorithm to classify aligned sequences into hierarchical clusters based on the patterns of conservation and variation in aligned tyrosine kinase domain sequences (fig. 1A) (Neuwald 2011, 2014). Each cluster is distinguished by cooccurring sequence motifs which are highly conserved within the cluster, but strikingly different outside of the cluster. By sampling different clustering hierarchies and highly distinguishing sequence motifs, we defined an optimal hierarchy for holozoan tyrosine kinases based on the log-probability ratio (LPR) scores, which quantifies the contribution of conserved sequence patterns to the classification/clustering measured in natural units of information (nats) (Neuwald 2011). The total LPR score for the optimized holozoan tyrosine kinome classification was 459825.35 nats, which is higher than the LPR score for KinBase classification of tyrosine kinases (443759.95 nats) (supplementary fig. S1, Supplementary Material online).

Next, we reclassified 34,954 tyrosine kinase sequences from the UniProt reference proteomes database into the optimized hierarchy by quantifying the extent to which individual sequences match cluster-specific motifs (see Materials and Methods for details). We define this reclassification as a constraint-based classification because this post-processing step eliminates spurious or divergent sequences from clusters that do not score over an optimal cut-off score due to their lack of cluster-specific patterns. The spurious sequences eliminated from each cluster are categorized within the unclassified family (fig. 2).

The new constraint-based hierarchical classification of tyrosine kinases, which is broadly similar to the KinBase classification of the tyrosine kinome, reveals several novel subgroupings. In particular, the constraint-based classification defines three new subgroupings of tyrosine kinase families which account for nearly half of the tyrosine kinome: the Src module (SrcM) subgroup, the insulin receptor kinase-like (IRKL) subgroup, and the fibroblast, platelet-derived, and vascular growth factor receptors (FPVR) subgroup (fig. 2). The SrcM subgroup differs significantly from the KinBase classification in that it clusters the SrcA, SrcB, and Frk subfamilies of the KinBase-defined Src family within the same subgroup as the Tec and Abl families. Furthermore, SrcM does not include the SRM and sponge-specific Src (Src-Aque1) families. The FPVR subgroup includes seven distinct receptor tyrosine kinase families, where a Platelet-derived, and Vascular growth factor Receptor (PVR) subgroup subclassifies the VEGFR, PDGFR, Kit, CSF1R, and Flt3 families as

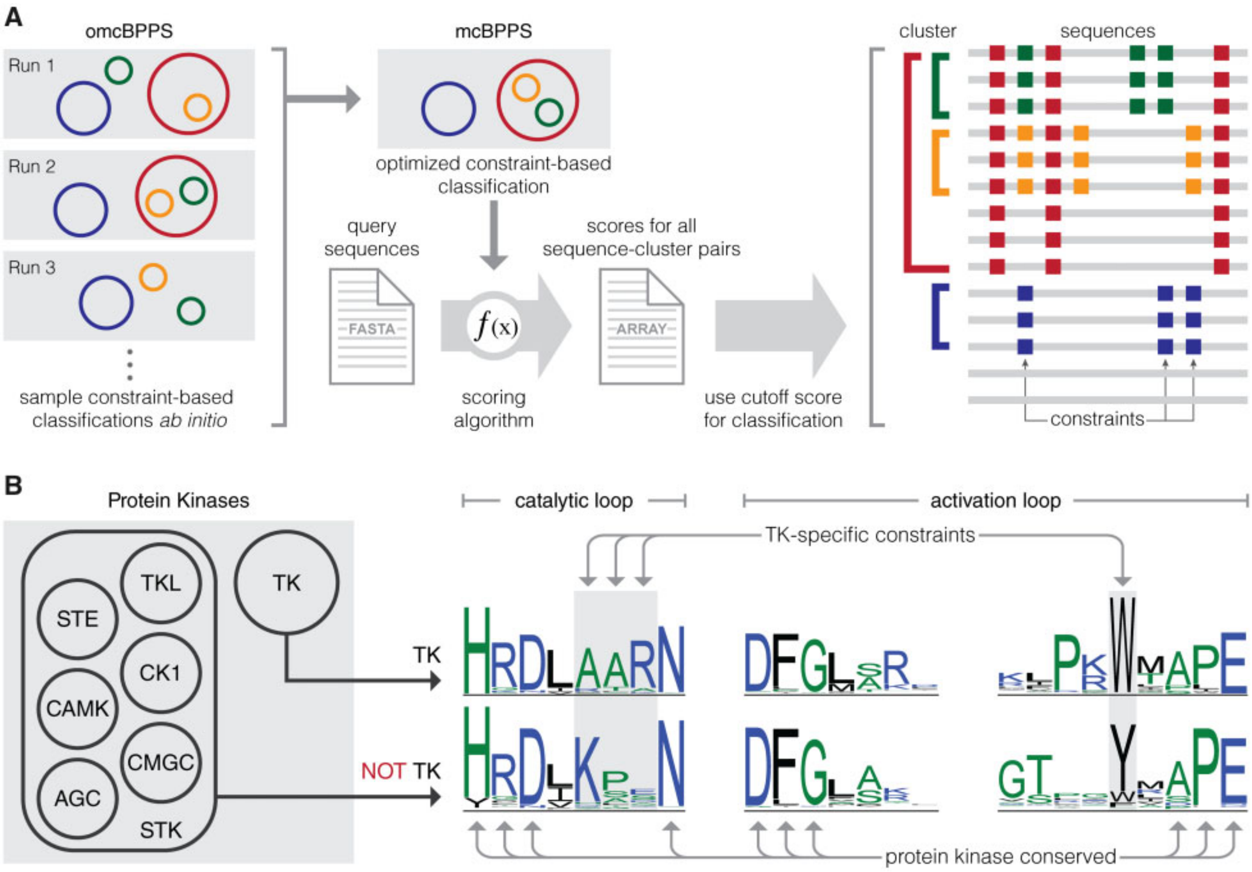

Fig. 1. Workflow for constraint-based hierarchical clustering of tyrosine kinase sequences. (*A*) Multiple omcBPPS runs were used to sample various constraint-based hierarchical classifications for tyrosine kinases. An optimal constraint-based classification was determined based on LPR scores calculated using mcBPPS. Next, a scoring algorithm was used to score sequence-cluster pairs and to include or exclude tyrosine kinase sequences from each cluster defined in the optimal classification. An example constraint-based hierarchical classification is shown on the right. Clusters are represented as colored brackets, sequences are represented as gray lines, and constraints specific to each cluster are represented by squares colored according to the cluster to which they belong. For example, sequences in green cluster share sequence constraints denoted by green squares, which are not found in sequences outside of the green cluster, as well as sequence constraints denoted by red squares, which are not found in sequences outside of the red cluster. The last two sequences are not included in any cluster as they lack any of the cluster-specific constraints defined in the constraint-based classification. (*B*) A visual representation of cluster-specific constraints is shown using previously published data on tyrosine kinase-specific constraints (Mohanty et al. 2016). Seven clusters of protein kinases are shown on the left, where tyrosine kinases are clustered separately from other protein kinases. A sequence logo of tyrosine kinase sequences is shown alongside a sequence logo of all other protein kinases. Sequence motifs such as HRD, DFG, and APE are conserved throughout all protein kinases, whereas the catalytic loop AAR motif and the activation loop tryptophan are defined as tyrosine kinase-specific constraints because their conservation is specific to the tyrosine kinase cluster.

a distinct cluster separate from the FGFR and Ret families. Notably, the new classification separates the KinBase PDGFR family into four families (PDGFR, Kit, CSF1R, and Flt3) due to statistically significant sequence constraints that define each of these families, warranting their designation as distinct families. The IRKL subgroup is the largest subgroup, comprising roughly 16% of tyrosine kinase sequences, and encompasses nine receptor tyrosine kinase families, including the insulin receptor kinase family as well as other poorly studied tyrosine kinases such as the CCK4 family of pseudokinases (Jung et al. 2004; Murphy et al. 2014) and the Lmr family which exhibits serine/threonine kinase activity despite its placement into the tyrosine kinase clade (Wang and Brautigan 2002; Ditsiou et al. 2020). We note that some organism-specific tyrosine kinase families defined in the KinBase classification, such as the unique receptor tyrosine kinase families in choanoflagellates (e.g., RTKA, RTKB, etc.) (King et al. 2008; Manning et al. 2008), were not detected due to the limited number of detectable homologs in current sequence databases (see Materials and Methods).

Subgroup-Specific Motifs Localize to Known Autoregulatory Sites in the Kinase Domain

By examining the sequence constraints that define each of the three novel subgroups in light of existing crystal structures, we observe that subgroup-specific motifs are located in known regulatory regions of the kinase domain. For example, the SrcM subgroup conserves a highly distinguishing GxM motif in the β3-αC loop and a GxKF motif in the activation loop that both form important interactions associated with a common Src-like inactive conformation in the activation loop (fig. 3A) (Xu et al. 1999; Shah et al. 2018). This Src-like inactive conformation has been observed across diverse SrcM families, such as SrcA (Xu et al. 1999), SrcB (Schindler et al.

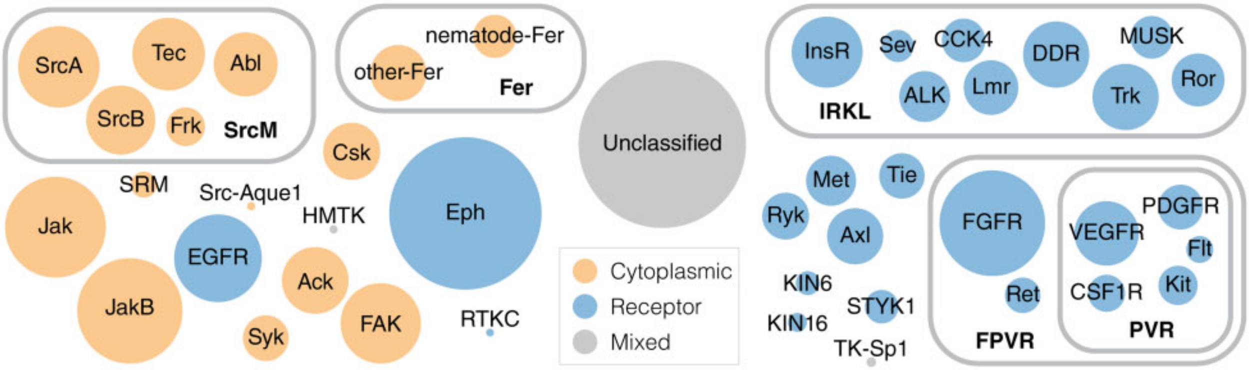

Fig. 2. An evolutionary constraint-based hierarchical classification of the tyrosine kinome. The constraint-based hierarchical classification of tyrosine kinases is depicted as an Euler diagram. Each circle represents a distinct cluster of tyrosine kinases and is scaled to the number of sequences in each cluster. Clusters containing cytoplasmic tyrosine kinases are indicated with orange circles, whereas clusters containing receptor tyrosine kinases are indicated with blue circles. Clusters containing both cytoplasmic and receptor tyrosine kinases are colored gray. For more information, see supplementary file S1, Supplementary Material online.

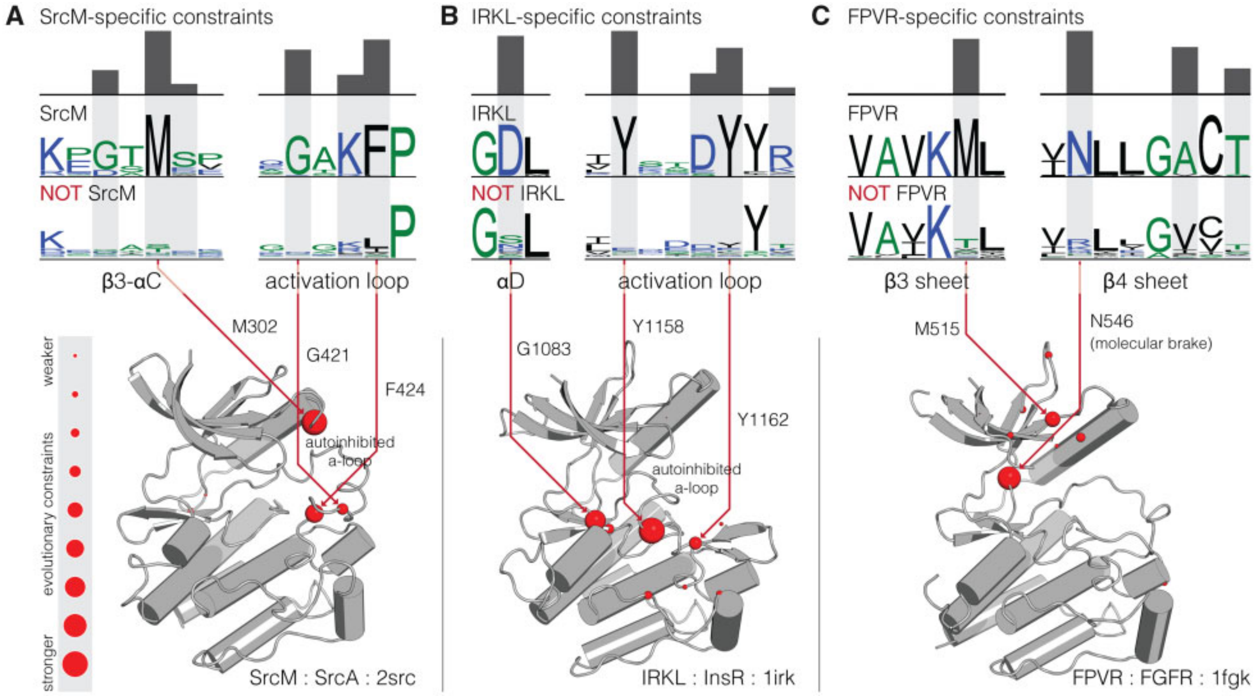

Fig. 3. Structural locations of sequence motifs defining the SrcM, IRKL, and FPVR subgroups. Comparative sequence logos and structural mappings of subgroup-specific motifs are shown for the (*A*) SrcM, (*B*) IRKL, and (*C*) FPVR subgroups of tyrosine kinases. Sequence logos for the strongest evolutionary constraints corresponding to each subgroup are shown on top, with comparative sequence logos for sequences outside each subgroup provided below. Evolutionary constraints are highlighted in gray, with the height of the histogram reflecting the degree of divergence at that position between the subgroup and sequences outside the subgroup. Evolutionary constraints are shown as red balls on representative inactive structures of SrcM (Src) (Xu et al. 1999), IRKL (IRK) (Hubbard et al. 1994), and FPVR (FGFR1) (Mohammadi et al. 1996). The size of the red balls represents the strength of the evolutionary constraint. The strongest constraints are labeled with residue numbers corresponding to the position in the representative structures.

1999), Abl (Levinson et al. 2006), and Tec (Wang et al. 2015), and similarities between their inactive structures have been previously noted (Shah et al. 2018). SrcM-specific sequence motifs are located in key regions in the Src-like inactive conformation which suggests that these residues play key roles in the conformational control of SrcM kinase activity. Likewise, the strongest sequence constraints on IRKL tyrosine kinases are associated with a common autoinhibitory conformation of the activation loop (fig. 3*B*), which has been observed across crystal structures of diverse IRKL members (Hubbard et al. 1994; Artim et al. 2012; Canning et al. 2014; Ditsiou et al. 2020), and is distinct from the autoinhibitory activation loop conformation of SrcM tyrosine kinases. Lastly, the FPVR subgroup of tyrosine kinases is defined by a highly conserved asparagine in the hinge region of the kinase domain (fig. 3*C*), which engages an autoinhibitory "molecular brake"

(Chen et al. 2007) shared across these kinases. Other FPVR-specific sequence motifs are structurally located near the juxtamembrane, which is an important regulatory segment for many receptor tyrosine kinases (Griffith et al. 2004), and may play common structural and functional roles in juxtamembrane-mediated regulation across FPVR tyrosine kinases.

Tyrosine Kinase Subgroups Are Anciently Conserved across Diverse Holozoan Taxa

In order to infer when each of these tyrosine kinase subgroups and families emerged in evolution, we organized tyrosine kinase subgroups and families based on taxonomic conservation (fig. 4). Tyrosine kinases from the SrcM subgroup, IRKL subgroup, and Csk family are detected across the most diverse holozoan taxa, including in unicellular relatives of metazoans (premetazoans), such as filastereans and choanoflagellates. The conservation of these early emerging tyrosine kinases suggests that they likely played important roles in the evolution of metazoan multicellularity. The FPVR subgroup appears to have emerged later in eumetazoans, following the divergence from early metazoans such as poriferans, which lack the organized tissues observed across eumetazoans. Interestingly, the PVR subgroup within the FPVR subgroup evolved much later in metazoan evolution and can only be detected in deuterostomes. The fact that tyrosine kinases from the SrcM, IRKL, and FPVR subgroups emerged in early stages of metazoan evolution suggests that these tyrosine kinases, their defining sequence motifs, and the regulatory functions associated with the motifs (fig. 3) were important in the evolution of metazoan morphologies, such as multicellularity and organized tissues. We also note that, as found in previous kinome studies, our constraint-based classification defines several organism-specific tyrosine kinase families, such as the sponge-specific Src-Aque1 family (Srivastava et al. 2010), the nematode-specific KIN6 and KIN16 families (Plowman et al. 1999), and the choanoflagellate-specific HMTK and RTKC families (King et al. 2008) (fig. 4).

Domain Shuffling Contributed to Diverse Functions of the SrcM, IRKL, and FPVR Kinase Domains

In order to further explore the functional diversity of SrcM, IRKL, and FPVR tyrosine kinases, we surveyed the diversity of protein domains present across these subgroups and analyzed their conservation across holozoan taxa (fig. 5). As previously noted, the SrcM tyrosine kinases, as well as the SRM, Csk, and Src-Aque1 families of tyrosine kinases share a core SH3-SH2-kinase domain organization, an anciently conserved (Shah et al. 2018), coevolving unit (Nars and Vihinen 2001), which can be detected in SrcM orthologs in unicellular premetazoans, such as choanoflagellates and filastereans (fig. 5D). The Tec family domain architecture, which includes an N-terminal lipid-targeting PH-domain (except in the Tec family member Txk), can be detected in choanoflagellates and filastereans, therefore the SH3-SH2-kinase and PH-SH3-SH2-kinase domain architectures represent the most anciently conserved tyrosine kinase domain structures. Because sequence motifs defining the SrcM subgroup and the Tec family are also anciently conserved (fig. 4), we suggest that these motifs have coevolved with the SH3-SH2 and PH domains, respectively, and play key functional roles that link the kinase domain with their associated domains. Interestingly, the Abl family domain architecture, which includes a C-terminal F-actin binding domain appended to the SH3-SH2-kinase domain structure, emerged later in metazoan evolution after the

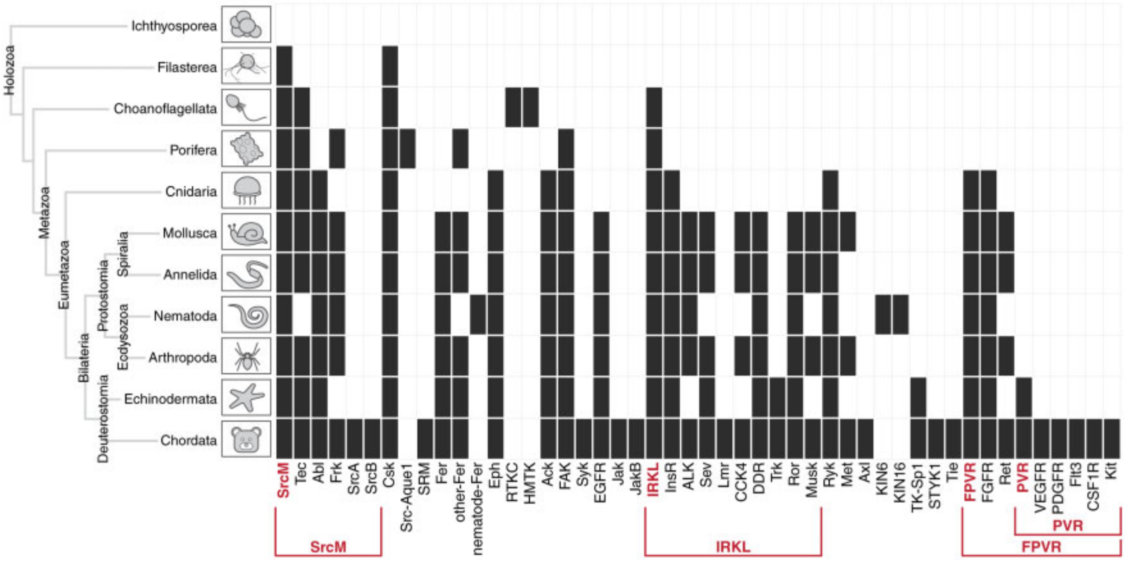

FIG. 4. Emergence of tyrosine kinase subgroups and families throughout diverse holozoan taxa. The detection of each tyrosine kinase subgroup and family defined in the constraint-based classification is shown across diverse holozoan taxa, including single-celled relatives of metazoa. Constraint-based tyrosine kinase subgroups and families are shown across the x-axis, and diverse holozoan taxa, and their evolutionary relationships are shown on the y-axis. Cells are marked black if one or more members of a subgroup or family could be detected within a given taxa. For more information, see supplementary file S2 and supplementary figure S7, Supplementary Material online.

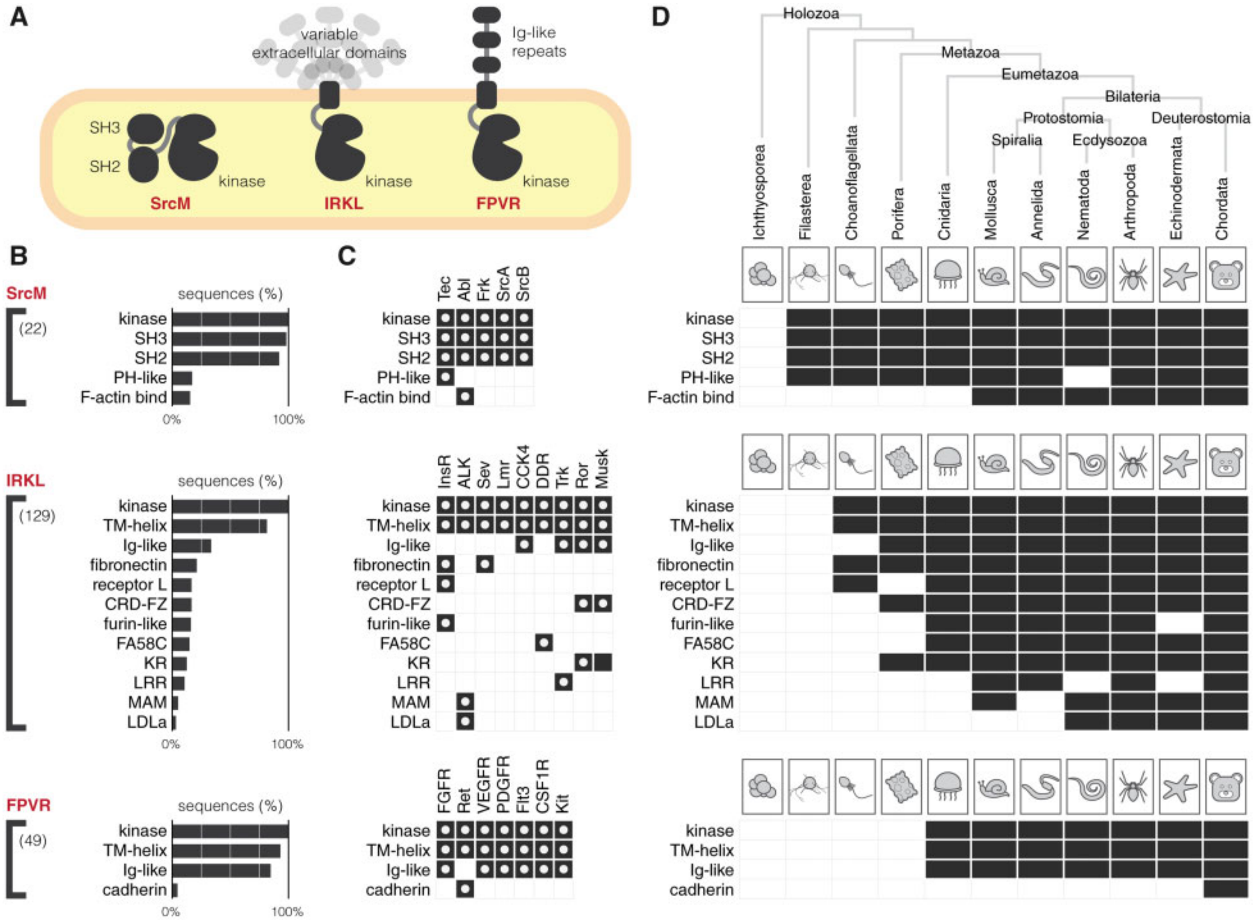

Fig. 5. Protein domains associated with SrcM, IRKL, and FPVR tyrosine kinases. (*A*) A graphical depiction of common protein domain architectures observed across the SrcM, IRKL, and FPVR subgroups. (*B*) Bar graphs show the frequencies of protein domains detected in SrcM, IRKL, and FPVR sequences. Protein domains that occur in at least 3% of sequences are shown. The value in parentheses denotes the total number of unique protein domains found for sequences of a given subgroup. Consecutive repeat domains have been compressed into a single domain for simplicity. (*C*) Protein domains found across individual tyrosine kinase families within the SrcM, IRKL, and FPVR subgroups. White dots indicate domains found in human tyrosine kinases in each family. (*D*) The conservation of protein domains associated with SrcM, IRKL, and FPVR tyrosine kinases across diverse holozoan taxa.

divergence of bilaterians from other eumetazoans. However, evolutionary sequence constraints that distinguish the Abl family kinase domain from other SrcM members emerged earlier in metazoan evolution after the emergence of eumetazoans (fig. 4). This suggests that Abl-specific functions of the kinase domain, perhaps substrate-specificity or Abl-specific regulation of catalysis, predated the additional functions imparted by the F-actin binding domain.

The IRKL and FPVR subgroups encompass the majority of receptor tyrosine kinase families and, despite sharing common kinase domain mechanisms within these subgroups (fig. 3), have diversified functions through the incorporation of various extracellular domains. In fact, the diverse assortment of extracellular domains architectures found throughout the IRKL subgroup may be consistent with extracellular domain shuffling throughout holozoan evolution. Interestingly, the emergence of family-specific extracellular domains often precedes the emergence of family-specific motifs that define receptor tyrosine kinase families. For example, the extracellular fibronectin domain, which is associated with InsR and Sev family receptor tyrosine kinases, can be detected in diverse holozoan taxa, from chordates to choanoflagellates (fig. 5D), however, InsR and Sev-specific sequence motifs in the kinase domain emerged later in metazoan evolution during the emergence of eumetazoans and bilaterians, respectively (fig. 4). Similarly, Frizzled cysteine-rich (CRD-FZ), Coagulation factor 5/8 C-terminal (FA58C), and leucine-rich repeat (LRR) domains, which are associated with the Ror/Musk, DDR, and Trk families, respectively, emerged before their respective family-specific motifs. The evolutionary emergence of extracellular domains before family-specific motifs in the kinase domain suggests that evolution first diversified extracellular ligand binding before the fine tuning of family-specific kinase domain functions such as downstream substrate specificity, regulation of catalysis, or intracellular protein–protein interactions. In contrast, in the ALK and Ret families of receptor tyrosine kinases, which uniquely contain LDLa and cadherin extracellular domains, respectively, the emergence of their family-specific sequence motifs predated the addition of their distinctive extracellular domains. These cases suggest that unique family-specific functions in the intracellular portion of receptor tyrosine kinases can also be expanded upon by subsequent shuffling of extracellular domains such that intracellular signaling functions are newly adapted to alternative extracellular ligands. Generally, despite the high degree of conservation of family-specific core domain structures (fig. 5), the extreme diversification of SrcM, IRKL, and FPVR kinase domains across holozoans is evident in the huge number of unique protein domains that can be detected across sequences.

A Representative Phylogeny of the Holozoan Tyrosine Kinome Reveals New Insights into Tyrosine Kinase Evolution

To better understand the evolutionary relationships between tyrosine kinase subgroups and families, we constructed a phylogenetic tree using maximum likelihood, which models the natural process of sequence variation and finds a tree that best describes the evolutionary history of diverse protein sequences (see Materials and Methods for details). By integrating our constraint-based classification of tyrosine kinases with our phylogenetic tree (fig. 6*A*), we can now infer the evolutionary history of sequence constraints imposed on tyrosine kinase subgroups and families defined in our constraint-based classification. For example, the IRKL, FPVR, and PVR subgroups, which were defined in our constraint-based classification due to their conservation of a set of subgroup-specific motifs, each form monophyletic clades in the phylogenetic tree, demonstrating that all tyrosine kinases within these subgroups have both maintained their respective subgroup-specific motifs and have descended from a common evolutionary ancestor. In contrast, the SrcM subgroup does not form a monophyletic clade in the phylogenetic tree. Instead, families within the SrcM subgroup share a monophyletic clade with the SRM, Src-Aque1, and Csk families, all of which likely descended from a common ancestor which conserved SrcM-specific motifs. However, that the SRM, Src-Aque1, and Csk families were not included in our constraint-based definition of the SrcM subgroup signifies that these families independently diverged from the rest of the SrcM subgroup through variations in the canonical SrcM-specific motifs, as well as accumulating additional variations contributing to functional divergence. Furthermore, examining SRM-specific, Src-Aque1-specific, and Csk-specific motifs unique to each of these families alongside SrcM subgroup-specific motifs will reveal how a common SH3-SH2-kinase domain organization (fig. 5) (Shah et al. 2018) has diverged along these various lineages to innovate divergent regulatory functions on a shared domain architecture. Our phylogenetic tree also confirms, along with our constraint-based classification, that the Lmr family belongs within the IRKL subgroup/clade despite detectable serine/threonine activity. Further sequence analysis shows that the Lmr kinases have regained a serine/threonine kinase-specific histidine (LMTK3$^{\mathrm{His260}}$) in the αE helix which tyrosine kinases selectively lost upon diverging from the serine/threonine kinases (supplementary fig. S5, Supplementary Material online) (Mohanty et al. 2016). The reemergence of this histidine may explain why Lmr kinases have regained serine/threonine activity. In addition, the evolutionary relationships described by our phylogeny of the holozoan tyrosine kinome are independently supported by conserved intron and phase positions in the kinase domain (supplementary fig. S6, Supplementary Material online) (Brunet et al. 2016, 2017).

A further examination of our phylogenetic tree of tyrosine kinases reveals several new insights about tyrosine kinase evolution. As noted in previous phylogenetic studies of tyrosine kinase sequences, tyrosine kinases form a monophyletic clade separate from the closely related tyrosine kinase-like (TKL) group of serine/threonine kinases. Holozoan tyrosine kinases (Group A) also form a monophyletic clade that is distinct from a paraphyletic group of divergent tyrosine kinase sequences found in pre-opisthokonts (Group B), such as those found in the amoebozoans *Dictyostelium discoideum* or in the green algae *Chlamydomonas reinhardtii* (fig. 6) (Suga et al. 2012). We note for the first time another major branching point in the evolution of tyrosine kinases which is associated with the presence or absence of an insert between the αD and αE helices of the kinase domain that we refer to as the DE insert, historically referred to as the kinase insert domain (Locascio and Donoghue 2013). Although the DE insert segment was not explicitly used in building the phylogenetic tree due to lack of detectable sequence similarity in this region across families, the phylogenetic tree exhibits a clear division of two clades, one containing kinases with a short αD-αE loop, which we refer to as the shortDE clade, and one predominantly containing kinases with a long αD-αE loop, which we refer to as the longDE clade (fig. 6*A*). The evolutionary separation of the longDE clade is also independently supported by a common phase-2 intron at the αH helix (supplementary fig. S6, Supplementary Material online). Although the variation in the length of the DE insert has been previously noted (fig. 6*B*), and previous studies have shown the functional significance of the insert on downstream signaling and kinase activation (Locascio and Donoghue 2013; Manni et al. 2014), the evolutionary history of the DE insert has not been examined.

In light of the evolutionary divergence between longDE and shortDE clades, we also note that the longDE clade primarily contains receptor tyrosine kinase families, whereas the shortDE clade contains predominantly cytoplasmic receptor tyrosine kinases (with the exception of the EGFR, Eph, and choanoflagellate-specific RTKC families). This correlation between the presence of the longDE insert with the presence of transmembrane and extracellular domains, alongside evidence that the insert plays important roles in kinase activation and protein recruitment for downstream signaling, suggests that the longDE insert evolved as a means to facilitate downstream intracellular signaling upon the activation of receptor tyrosine kinases by extracellular signals. In addition, though the DE insert is difficult to align across families due to the lack of sequence conservation, the DE insert is alignable within families and often conserves sequence motifs including phosphorylatable tyrosine, serine, or threonine residues, suggesting that individual receptor tyrosine kinase families along the longDE clade have rapidly and frequently

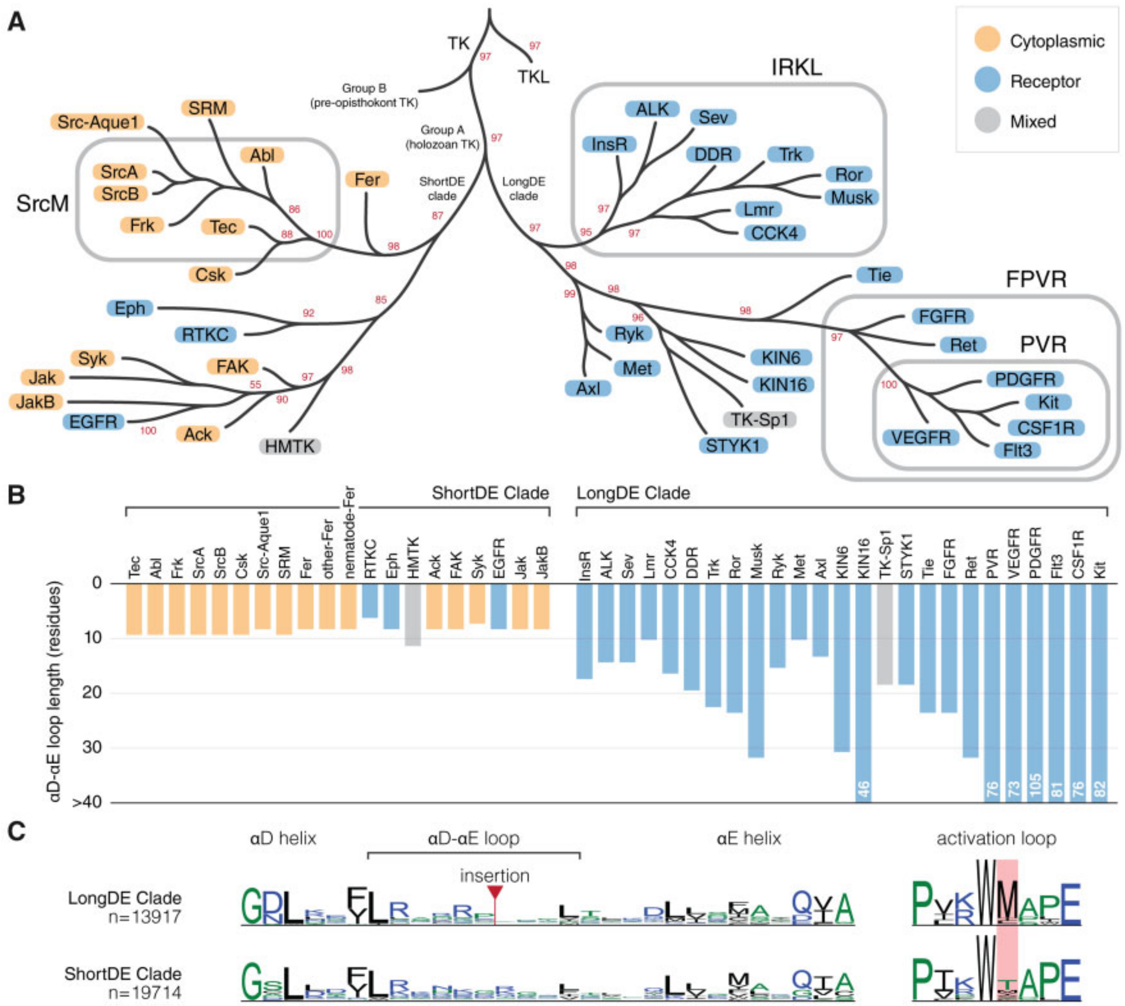

Fig. 6. Evolution of sequence constraint-defined subgroups/families in the holozoan tyrosine kinome. (*A*) An abridged depiction of our representative phylogeny for the holozoan tyrosine kinome. Branch tips represent clades of sequence constraint-defined families. Cytoplasmic tyrosine kinase families are indicated with orange circles, whereas receptor tyrosine kinase families are indicated with blue circles. Subgroups are indicated by rounded rectangles which encompass their respective constituents. Bootstrap support values for select clades are shown in red. Branch lengths are not drawn to scale. The full tree is provided in supplementary file S4, Supplementary Material online. (*B*) A bar chart showing the median αD-αE loop length of each tyrosine kinase family, which is shown on the x-axis separated by LongDE or ShortDE clade. On the y-axis, loop length is truncated at 30 residues. Any αD-αE loop lengths surpassing this limit are designated as >30. (*C*) Comparative sequence logos show differences between ShortDE and LongDE clade kinases.

evolved the longDE insert in family-specific contexts, presumably to carry out family-specific downstream signaling functions. We also note that the longDE tyrosine kinases highly conserve a unique activation loop methionine, which is not observed in shortDE tyrosine kinases (fig. 6C); however, the role of this methionine, or whether it is significant for DE insert function or for receptor tyrosine kinase function is unknown.

Interestingly, the shortDE clade in the tyrosine kinase phylogeny, which predominantly consists of cytoplasmic tyrosine kinases, includes two monophyletic clades of receptor tyrosine kinases: the EGFR family of receptor tyrosine kinases and a separate monophyletic clade that includes the Eph and choanoflagellate-specific RTKC families of receptor tyrosine kinases. Thus, our phylogeny suggests at least three independent origins of highly expanded receptor tyrosine clades, with the majority of receptor tyrosine kinases emerging from the longDE clade. That these disparate branches along the tyrosine kinase phylogeny have convergently evolved to include transmembrane and extracellular domains highlights the importance of relaying extracellular signals into intracellular responses across various signaling niches. Although the longDE receptor tyrosine kinases are distinguished by extra functionalities imparted by the longDE insert, the Eph and EGFR families also exhibit unusual signaling functions so far unobserved in other longDE receptor tyrosine kinases. Eph receptor tyrosine kinases have a unique capacity for bidirectional signaling, where the binding of ephrin ligands, which are also membrane bound, can activate signaling both in the receptor-bearing cell, as well as in the ligand-bearing cell (Pasquale 2010). Furthermore, our tree suggests that the RTKC family may also share this unique capacity for bidirectional signaling. The EGFR family is also a unique family of receptor tyrosine kinases in that ligand binding induces dramatic conformational changes in the dimerization arm extracellularly, also inducing a unique allosterically activating dimer in the intracellular portion (Zhang et al. 2006; Jura et al. 2009). These mechanisms of EGFR family kinases, as well as their activating (rather than autoinhibitory) juxtamembrane and their lack of activation via activation loop phosphorylation distinguishes the EGFR family from receptor tyrosine kinases in the longDE clade (Lemmon and Schlessinger 2010; Lemmon et al. 2014).

Discussion

The classification of protein kinases into evolutionarily related families has provided the foundation for decades of comparative sequence-structure-function studies on protein kinases (Hanks and Hunter 1995; Manning, Plowman, et al. 2002; Manning, Whyte, et al. 2002; Kwon et al. 2019). Here, we propose a new constraint-based classification of tyrosine kinases that newly defines the SrcM, IRKL, and FPVR subgroups (fig. 2) each of which maintains core subgroup-specific sequence motifs associated with subgroup-specific autoinhibited conformations (fig. 3). Subsequent taxonomic conservation analysis suggests that expansion of tyrosine kinase subgroups and evolution of family-specific motifs within these subgroups, along with domain shuffling, elaborated on subgroup-specific functions to diversify cell signaling functions (figs. 4 and 5). However, these core regulatory motifs are likely conserved because they ensure that these kinases are activated at the right place and time. For instance, the strongest SrcM-specific constraint is a methionine in the αC-β4 loop which packs against the autoinhibited activation loop conformation, suggesting an anciently conserved role in stabilizing the Src-like inactive conformation (fig. 3A) that may be relieved in various manners, such as the binding of substrates to the coevolved SH3-SH2 domains. Furthermore, the ancient conservation of Csk and SrcM-specific constraints supports the premetazoan origins of SrcM inhibition via C-terminal tail phosphorylation by Csk (Taskinen et al. 2017). Further study of family-specific motifs across various lineages of tyrosine kinases is expected to reveal unique regulatory mechanisms across distinct tyrosine kinase families.

We constructed a new representative phylogenetic tree of the holozoan tyrosine kinome which revealed larger evolutionarily related clades of tyrosine kinases associated with additional defining features (fig. 6A). Dividing the holozoan tyrosine kinome into two roughly equal halves, the basal longDE and shortDE clades are distinguished by the presence or absence (respectively) of a fast-evolving kinase domain insertion in the αD-αE loop (fig. 6B). Functionally important insertion regions shared by large groups of evolutionarily-related protein kinases have also been described in the CMGC group, which conserves a kinase domain insertion between the αH and αI helices (Kannan and Neuwald 2004). The diverse functions of the longDE insertion remains understudied; however, the region is documented to possess many functionally important phosphorylation sites in multiple tyrosine kinase families (Locascio and Donoghue 2013). Our tree also reveals three distinct evolutionary lines of receptor tyrosine kinases: longDE, RTKC-Eph, and EGFR, each of which are distinguished by unique lineage-specific variations on receptor tyrosine kinase signaling and regulation (Pasquale 2010; Locascio and Donoghue 2013; Lemmon et al. 2014). Overall, the definition of these evolutionarily related tyrosine kinases will enable the inference of sequence-structure-function relationships in the understudied kinases based on the known functions of well-studied kinase families.

The expansion and diversification of the tyrosine kinome across the animal kingdom highlights its central role in metazoan biology. Although many previous studies have speculated on the role of tyrosine kinases in the evolution of multicellularity (Miller 2012; Suga et al. 2014; Hunter and Manning 2015; Brunet et al. 2017), our findings suggest key evolutionary innovations which likely contributed to the adoption of tyrosine kinase signaling for multicellular functions (fig. 7A). Although elaborate tyrosine kinase signaling networks have been discovered in unicellular premetazoans, they generally display low orthology to tyrosine signaling networks in metazoans (Manning et al. 2008). Our analyses identify sparse similarities between premetazoan and metazoan tyrosine kinase signaling in that SrcM, Tec, Csk, and IRKL tyrosine kinases originated in premetazoans and have remained conserved throughout diverse metazoan taxa (fig. 4). These components may represent a core phosphotyrosine signaling machinery that have been expanded through the addition of taxa-specific tyrosine kinases to suit unique organismal needs. We further speculate that additional capacity for bidirectional signaling emerged in the RTKC-Eph lineage in an extinct ancestral premetazoan organism, as previous studies have identified distant Eph orthologs in various premetazoan organisms (Richter and King 2013; Tong et al. 2017) which are not shown in our conservation table because they lack Eph-specific sequence constraints. These core signaling functions likely enabled the evolution of primordial Metazoa, a mass of cells capable of moving and growing in a coordinated fashion, lacking much of the complexity found in modern day metazoans. Throughout the course of metazoan evolution, the emergence of new complex biological systems such as the nervous system, circulatory system, and adaptive immunity appears to coincide with the emergence of functionally associated tyrosine kinase families (table 1). Additional complexity was gained through whole-genome duplication and small-scale duplication events which played a major role in tyrosine kinase evolution in vertebrates, especially the PVR kinases whose evolution has been heavily influenced by both types of duplication events (Grassot et al. 2006; D'Aniello et al. 2008; Brunet et al. 2016, 2017).

Given the important roles of tyrosine kinases in multicellular metazoan biology, it comes as no surprise that sequencing efforts have identified many disease-related variants across the tyrosine kinome (Hunter 2009). Multiple studies have proposed reversal of cancer cells to a more unicellular-like state through the release of "multicellular constraints" (Chen et al. 2015). Although many oncogenes predate the origin of multicellularity, evolutionary studies have identified a second burst of oncogene emergence which cooccurred

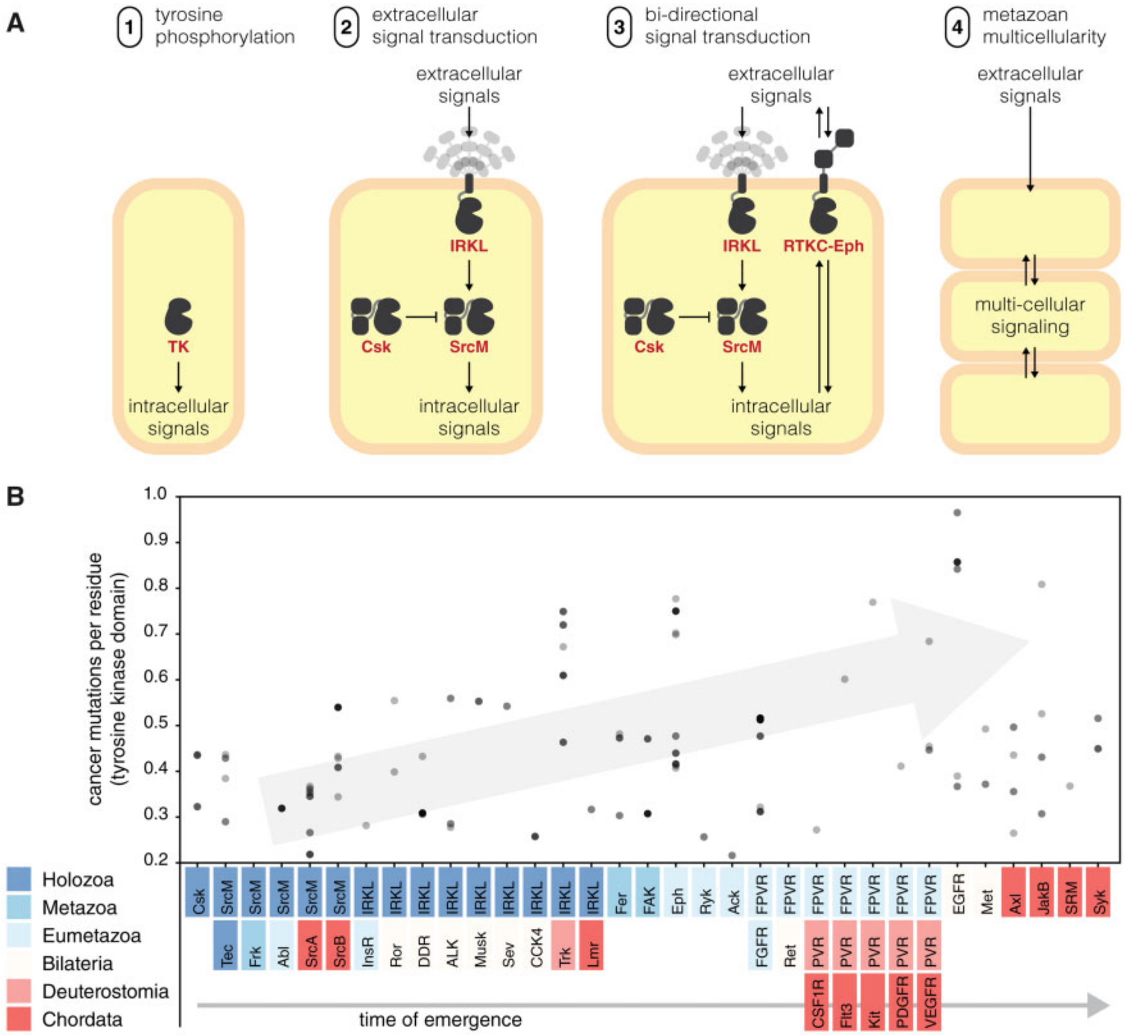

Fig. 7. Tyrosine kinases in the evolution of multicellularity and cancer. (*A*) We propose a series of evolutionary innovations of the tyrosine kinome signaling which potentially contributed to the emergence of metazoan multicellularity. (*B*) Disease-related mutations tend to occur in more recent tyrosine kinase families. A scatter plot shows how frequently human tyrosine kinases are found to be mutated in genome-wide cancer sequencing studies from the Catalogue Of Somatic Mutations In Cancer (COSMIC) database (Tate et al. 2019). On the y-axis, mutation frequency is measured by the average number of mutations per residue in the tyrosine kinase domain. On the x-axis, each kinase is sorted into a cluster defined by our hierarchy and sorted by their time of emergence. All points were drawn with transparency; as a result, points which appear darker indicate multiple points falling in the same location.

with the appearance of multicellular metazoans, and is comprised of genes (including tyrosine kinases) involved in cellular signaling and growth processes (Domazet-Lošo and Tautz 2010). In support of this finding, our data offer a closer look at this event, suggesting that cancer variants occur more frequently in recently evolved tyrosine kinases (fig. 7*B*). Overall, the deep connection between phosphotyrosine signaling and cancer encourages further studies on the tyrosine kinome and the unique sequence constraints that guide their evolution and function. Focused analyses are expected to reveal new insights on metazoan multicellular signaling and its malfunctions in disease states.

Materials and Methods

Evolutionary Constraint-Based Clustering

We sampled alternative classification hierarchies ab initio, using the omcBPPS algorithm (Neuwald 2014) which employs a Markov chain Monte Carlo sampling strategy to classify aligned sequences into hierarchical categories/clusters based on shared constraints, that is, slow evolving sites. Category/cluster-specific constraint refers to alignment positions that are highly conserved within a given cluster, but divergent in sequences outside of the cluster. The omcBPPS algorithm iteratively optimizes for two interdependent criteria: 1) which alignment positions should be defined as cluster-specific constraints and 2) what is the optimal hierarchical clustering scheme based on cluster-specific constraints.

We ran omcBPPS on two different sequence sets from UniProt reference proteomes (retrieved February 13, 2019) and NCBI nonredundant (nr) proteins database (retrieved February 13, 2019) (Pruitt et al. 2007). Within these sequence sets, we identified and aligned tyrosine kinase sequences using the Multiply-Aligned Profiles for Global Alignment of Protein Sequences (MAPGAPS) algorithm (Neuwald 2009). This

Table 1. The Emergence of Tyrosine Kinase Families Is Associated with the Metazoan Phenotypes.

Family/Subgroup	Kinase Functions	Time of Emergence	Predicted Association to Metazoan Evolution	References
FAK	Dynamically regulation of cell adhesion and motility	Porifera (Metazoa)	Organized cell layer (found in all metazoans)	Mitra et al. (2005); Mitchell and Nichols (2019)
Eph	Regulation of cell migration during development	Cnidaria (Eumetazoa)	Organized germ layers (emerged in cnidarians)	Poliakov et al. (2004)
FGFR	Regulation of early metazoan development, such as gastrulation and neural induction	Cnidaria (Eumetazoa)	Simple nervous system (emerged in eumetazoans)	Green et al. (1996); Matus et al. (2007); Bertrand et al. (2014)
FPVR	Many FPVR kinases play roles in vasculature-related functions	Cnidaria (Eumetazoa)	Vasculature (emerged in bilaterians)	Manning, Plowman, et al. (2002)
EGFR	Regulated cell growth via the EGFR pathway	Bilateria	EGFR pathway (emerged in bilaterians)	Barberán et al. (2016)
Ror, Musk, DDR, Sev, ALK, CCK4	Many IRKL kinase families play roles in neural development	Bilateria	True brain characterized by neuropile (emerged in bilaterians)	Manning, Plowman, et al. (2002); Chiu and Cline (2010)
VEGFR, PDGFR, Kit, CSF1R, Flt3	Many PVR kinase families play roles in blood vasculature-related functions and are implicated in blood cancers	Chordata	Closed circulatory system lined with endothelium (emerged in vertebrates)	Monahan-Earley et al. (2013); Berenstein (2015)
JAK, JAK-b, Syk	Immune signaling via the JAK-STAT pathway	Chordata	Adaptive immunity (emerged in vertebrates)	Mócsai et al. (2010); Liongue et al. (2016)

Note.—The taxa in which many tyrosine kinase families first emerged (fig. 4) is correlated with the emergence of various taxa-specific phenotypes which are associated with functions performed by the corresponding tyrosine kinase family. Families whose functions do not seem to exhibit a clear connection to the emergence of a metazoan phenotype are not shown.

alignment was restricted to the protein kinase domain starting from the β1 strand (PKA^{Phe43}) and ending at the αI helix (PKA^{Lys292}). Kinase sequences which did not span from at least the β3-Lys to the DFG-Asp were deemed fragmentary and removed from the alignment. The UniProt sequence set contained 12,137 tyrosine kinase sequences. The nr sequence set was further purged at 98% sequence identity and contained 17,071 tyrosine kinase sequences. We performed hierarchical clustering on both sequence sets using omcBPPS. For both sequence sets, we optimized the “minnats” parameters (minnat = 1 and minnat = 5) which changed the minimum log-likelihood required to form a cluster. All runs were performed twice. To create a consensus of the hierarchical classification schemes found by multiple runs of omcBPPS, we used the mcBPPS algorithm. Clusters which were consistently identified throughout multiple runs were refined using the mcBPPS algorithm (Neuwald 2011). We ran mcBPPS on a maximally diverse set of 33,769 sequences containing all tyrosine kinases detectable from nr to generate an optimal model that is consistent with the existing data.

Fitting New Sequences to a Constraint-Based Clustering Model

Using our consensus model, we developed quantitative means of evaluating how well any given sequence fit into each of the clusters defined by our model (supplementary algorithm S1, Supplementary Material online). Within our model, each cluster was associated with a list of constraints that dictated which amino acid(s) were likely to be found at a given alignment position. Furthermore, each constraint was associated with a log-likelihood score which described how specific a constraint was to its respective cluster. To score a sequence against a cluster, we added the log-likelihoods of all the constraints which were true for the query sequence. In order to make this score comparable across clusters, we divided this number by the total log-likelihood of all the cluster’s constraints. This resulted in a range from 0 to 1. For example, a sequence which followed all of a given cluster’s constraints would have received a maximum score of 1 for that cluster, whereas a sequence which did not follow any of a given cluster’s constraints would have received a minimum score of 0 for that cluster. For the purposes of discrete classification, we defined a cut-off score for classifying a sequence into a cluster. Based on the distribution of scores from all possible sequence-cluster pairs across multiple test data sets, we defined the optimal cut-off at the global minima of 0.7.

Using supplementary algorithm S1, Supplementary Material online, we scored a representative set of sequences from UniProt proteomes (retrieved April 2, 2020) and estimated the size of each cluster. A representative set of 44,639 tyrosine kinase sequences spanning 586 species were identified and aligned to a common alignment profile using the MAPGAPS algorithm (Neuwald 2009). We determined the size of each cluster based on how many sequences scored above the cut-off. We also quantified the similarity between clusters by evaluating how well sequences in a given cluster scored against all other clusters using an all-versus-all comparison (supplementary fig. S2, Supplementary Material online). We defined a nonsymmetric similarity metric for any two given clusters, A and B, as the average score of cluster A sequences when classified against cluster B constraints. As a consequence of our cut-off, the average score when A and B were the same cluster was always greater than 0.7 which we observe across the diagonal. Indicative of our hierarchical classification scheme, we also observed a distinct signature for subclusters (such as Tec) which were classified under a larger supercluster (such as SrcM): subclusters would score highly (>0.7) for the constraints of its respective supercluster but not vice versa. Our comparison matrix also indicated that evolutionarily related sequences within clusters share more constraints in common than sequences outside of the cluster. Finally, we observed that the majority of values outside of the diagonal were quite low which indicated that our model defined well-separated sequence clusters, each defined by their own unique sequence constraints.

Comparative Sequence Analysis

All comparative sequence analyses were performed in Python 3 using the HelperBunny library (provided with our computational notebooks) which implements NumPy array-based sequence alignment manipulation. All sequence features (including features pertaining to motifs, evolutionary constraints, domain composition, taxonomic descriptors, insertions, and deletions) were represented as Boolean arrays as a function of the full sequence alignment. More complex queries pertaining to multiple sequence features (such as the presence of a given domain in a given taxon) were constructed by Boolean algebra. These Boolean arrays were applied as filters to our sequence alignment using NumPy indexing routines (Harris et al. 2020). Sequence logos were generated using the WebLogo 3 API (Crooks et al. 2004).

Taxonomic Conservation Analysis

After we classified our representative set of tyrosine kinase into discrete clusters, we determined the taxonomic conservation of each cluster. We determined the source of each tyrosine kinase sequence using the organism identifier number (OX) provided in the FASTA header of UniProtKB sequences. OX numbers were traced back to their parent node identifiers using the nodes dump file provided in the NCBI taxdump database. All node identifiers were translated to their respective scientific names using the taxonomy names dump file. We determined the distribution of taxa across each cluster of tyrosine kinases and selected an optimal mix to depict diverse taxa ranging from unicellular premetazoans to more complex metazoans such as chordates (supplementary file S2, Supplementary Material online). Because our definition of protein family is based on sequence constraints, our methods may sometimes yield differential results compared with more traditional methods which utilize sequence homology (Fairclough et al. 2013; Junqueira Alves et al. 2019). We report taxonomic conservation using our optimized cut-off score of 0.7 (fig. 4) as well as a relaxed cut-off score of 0.6 (supplementary file S2 and Supplementary fig. S7, Supplementary Material online).

Intron and Phase Position Analysis

Intron/exon annotations were mapped using Scipio (Keller et al. 2008). Intron phases were determined by calculating the modulus of three (representing the codon size) of the cumulative sum of lengths of each exon which preceded an intron within an open reading frame. A phase-0 intron is located between two consecutive codons; a phase-1 intron is located between the first and second nucleotides of a codon; and a phase-2 intron is located between the second and third nucleotides of a codon.

Protein Domain Conservation Analysis

We produced protein domain annotations for each full-length tyrosine kinase sequence using the NCBI Conserved Domain Database (CDD v3.18—55,570 PSSMs) database of conserved protein domains (Marchler-Bauer et al. 2011, 2017). Queries to the database were programmatically submitted using the bwrpsb PERL script which was provided in the Batch CD-Search API. Search parameters included an expected value threshold of 0.01 with only the best scoring domain model being returned. Transmembrane (TM) helix annotations were identified and appended to our domain annotation results using TMHMM 2.0 (Krogh et al. 2001). Synonymous domain names were manually identified and merged in post-processing. We combined these domain annotations with previously generated data to evaluate the conservation of protein domains across various constraint-defined clusters and taxonomic clades.

Phylogenetic Analysis

We inferred the evolution of tyrosine kinase families/subgroups using a maximum-likelihood approach. In order to create a representative phylogeny of the holozoan tyrosine kinome, we sampled a taxonomically diverse set of sequences from each constraint-defined cluster of tyrosine kinase sequences. Our representative set of sequences consisted of 1) one randomly selected sequence from each cluster-taxon pair (based on supplementary file S2, Supplementary Material online) excluding Chordata, 2) all human tyrosine kinase sequences which represented the chordate taxon, 3) three early diverging, preopisthokont tyrosine kinase sequences from amoebozoan, *Acanthamoeba castellanii* (Suga et al. 2012), and 4) eight human TKL group kinases which was used as an outgroup. We produced multiple representative sequence sets with different random samples of taxonomically diverse sequences. Using these sequence sets, we generated multiple phylogenies using IQTREE v1.6.11 (Nguyen et al. 2015), with ModelFinder (Kalyaanamoorthy et al. 2017). Branch support values were generated using ultrafast bootstrap with 1000 resamples (Hoang et al. 2018). Results indicated that sequences from the same clusters consistently formed paraphyletic groups or monophyletic clades with high bootstrap support (supplementary files S4 and S5, Supplementary Material online). We compared topologies generated from different random samples and found no major changes in the evolutionary relationships between evolutionary constraint-defined sequence clusters. Furthermore, our topology was robust to the inclusion of unclassified tyrosine kinase sequences.

The consensus tree with the highest bootstrap support values for the three major tyrosine subgroups was selected as the final tree. The optimal substitution model for our final topology was determined to be LG+R8 based on the Bayesian Information Criterion as determined by ModelFinder (Kalyaanamoorthy et al. 2017). We rooted our final tree against the TKL outgroup using ETE Toolkit v3.1.1 (Huerta-Cepas et al. 2016). Consistent with previous studies, the most divergent tyrosine kinases in our tree were the paraphyletic preopisthokont "Group B" tyrosine kinases which diverged prior to the emergence of the monophyletic "Group A" holozoan tyrosine kinases (Suga et al. 2012). We observed high concordance between our evolutionary constraint-based clustering and phylogenetic inference; this allowed us to simplify our tree topology by condensing monophyletic clades and pruning paraphyletic groups (fig. 6A). The simplified representation describes evolutionary relationships between constraint-defined families. Furthermore, we represented each of the three major tyrosine kinase subgroups by drawing an enclosure around each subgroup's constituent families.

We compared our tree to previously published phylogenies which also sampled diverse tyrosine kinase families (Robinson et al. 2000; Manning, Whyte, et al. 2002; Suga et al. 2008; Suga et al. 2012; Brunet et al. 2016; Modi and Dunbrack 2019). However, many of these studies focused on vertebrate, basal metazoan, and premetazoan kinase sequences leaving out many diverse protostome sequences. We observed key differences with the current widely accepted phylogeny of the human tyrosine kinome (supplementary fig. S3, Supplementary Material online) (Manning, Whyte, et al. 2002) and found the most similarities to a tree published by Robinson et al. (2000). Previously placed near the base of the tyrosine kinase clade, our placement of STYK1 was supported by common introns shared with Tie and various IRKL families, whereas our placement of Lmr was supported by common introns shared by various IRKL kinases (Brunet et al. 2016). Furthermore, the majority of human longDE kinases share a common phase-2 intron in the genomic region which codes for the αH helix (supplementary fig. S6, Supplementary Material online). Overall, our tree has high support for all major clades, including historically difficult-to-place families such as JAK (Shiu and Li 2004). Providing additional support, we note that our tree is highly concordant with the evolutionary progression of holozoan taxa in that anciently conserved tyrosine kinase families tend to diverge first, whereas more recently diverged tyrosine kinase families appear closer to the tips (supplementary fig. S4, Supplementary Material online).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Dr James H. Leebens-Mack for valuable feedback. This work was supported by the National Institutes of Health (R01 GM114409 and R35 GM139656 to N.K.).

Data Availability

The data sets and code for our analyses are freely available for download from GitHub at: https://github.com/esbgkannan/holozoan_tk_evolution.

References

Amatya N, Wales TE, Kwon A, Yeung W, Joseph RE, Fulton DB, Kannan N, Engen JR, Andreotti AH. 2019. Lipid-targeting pleckstrin homology domain turns its autoinhibitory face toward the TEC kinases. *Proc Natl Acad Sci U S A*. 116(43):21539–21544.

Artim SC, Mendrola JM, Lemmon MA. 2012. Assessing the range of kinase autoinhibition mechanisms in the insulin receptor family. *Biochem J*. 448(2):213–220.

Barberán S, Martín-Durán JM, Cebrià F. 2016. Evolution of the EGFR pathway in Metazoa and its diversification in the planarian *Schmidtea mediterranea*. *Sci Rep*. 6:28071.

Berenstein R. 2015. Class III receptor tyrosine kinases in acute leukemia – biological functions and modern laboratory analysis. *Biomark Insights*. 10(Suppl 3):1.

Bertrand S, Iwema T, Escriva H. 2014. FGF signaling emerged concomitantly with the origin of eumetazoans. *Mol Biol Evol*. 31(2):310–318.

Bradham CA, Foltz KR, Beane WS, Arnone MI, Rizzo F, Coffman JA, Mushegian A, Goel M, Morales J, Geneviere A-M, et al. 2006. The sea urchin kinome: a first look. *Dev Biol*. 300(1):180–193.

Brunet FG, Lorin T, Bernard L, Haftek-Terreau Z, Galiana D, Schartl M, Volff J-N. 2017. Cham case studies of seven gene families with unusual high retention rate since the vertebrate and teleost whole-genome duplications. In: Pontarotti P, editor. Evolutionary biology: self/nonself evolution, species and complex traits evolution, methods and concepts. New York: Springer International Publishing. p. 369–396. Available from: 10.1007/978-3-319-61569-1_19

Brunet FG, Volff J-N, Schartl M. 2016. Whole genome duplications shaped the receptor tyrosine kinase repertoire of jawed vertebrates. *Genome Biol Evol*. 8(5):1600–1613.

Caenepeel S, Charydczak G, Sudarsanam S, Hunter T, Manning G. 2004. The mouse kinome: discovery and comparative genomics of all mouse protein kinases. *Proc Natl Acad Sci U S A*. 101(32):11707–11712.

Canning P, Tan L, Chu K, Lee SW, Gray NS, Bullock AN. 2014. Structural mechanisms determining inhibition of the collagen receptor DDR1 by selective and multi-targeted type II kinase inhibitors. *J Mol Biol*. 426(13):2457–2470.

Chen H, Lin F, Xing K, He X. 2015. The reverse evolution from multicellularity to unicellularity during carcinogenesis. *Nat Commun*. 6:6367.

Chen H, Ma J, Li W, Eliseenkova AV, Xu C, Neubert TA, Miller WT, Mohammadi M. 2007. A molecular brake in the kinase hinge region regulates the activity of receptor tyrosine kinases. *Mol Cell*. 27(5):717–730.

Chiu S-L, Cline HT. 2010. Insulin receptor signaling in the development of neuronal structure and function. *Neural Dev*. 5:7.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res*. 14(6):1188–1190.

D'Aniello S, Irimia M, Maeso I, Pascual-Anaya J, Jiménez-Delgado S, Bertrand S, Garcia-Fernàndez J. 2008. Gene expansion and retention leads to a diverse tyrosine kinase superfamily in amphioxus. *Mol Biol Evol*. 25(9):1841–1854.

Ditsiou A, Cilibrasi C, Simigdala N, Papakyriakou A, Milton-Harris L, Vella V, Nettleship JE, Lo JH, Soni S, Smbatyan G, et al. 2020. The structure-function relationship of oncogenic LMTK3. *Sci Adv*. 6(46):eabc3099.

Domazet-Lošo T, Tautz D. 2010. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biol.* 8:66.

Fairclough SR, Chen Z, Kramer E, Zeng Q, Young S, Robertson HM, Begovic E, Richter DJ, Russ C, Westbrook MJ, et al. 2013. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. *Genome Biol.* 14(2):R15.

Grassot J, Gouy M, Perrière G, Mouchiroud G. 2006. Origin and molecular evolution of receptor tyrosine kinases with immunoglobulin-like domains. *Mol Biol Evol.* 23(6):1232–1241.

Green PJ, Walsh FS, Doherty P. 1996. Promiscuity of fibroblast growth factor receptors. *Bioessays* 18(8):639–646.

Griffith J, Black J, Faerman C, Swenson L, Wynn M, Lu F, Lippke J, Saxena K. 2004. The structural basis for autoinhibition of FLT3 by the juxtamembrane domain. *Mol Cell.* 13(2):169–178.

Hanks SK, Hunter T. 1995. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification1. *Faseb J.* 9(8):576–596.

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, et al. 2020. Array programming with NumPy. *Nature.* 585(7825):357–362.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol.* 35(2):518–522.

Hubbard SR, Wei L, Ellis L, Hendrickson WA. 1994. Crystal structure of the tyrosine kinase domain of the human insulin receptor. *Nature* 372(6508):746–754.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol.* 33(6):1635–1638.

Hunter T. 2009. Tyrosine phosphorylation: thirty years and counting. *Curr Opin Cell Biol.* 21(2):140–146.

Hunter T. 2014. The genesis of tyrosine phosphorylation. *Cold Spring Harb Perspect Biol.* 6(5):a020644.

Hunter T, Manning G. 2015. The eukaryotic protein kinase superfamily and the emergence of receptor tyrosine kinases. In: Wheeler DL, Yarden Y, editors. Receptor tyrosine kinases: structure, functions and role in human disease. New York: Springer. p. 1–15. Available from: 10.1007/978-1-4939-2053-2_1

Jung J-W, Shin W-S, Song J, Lee S-T. 2004. Cloning and characterization of the full-length mouse Ptk7 cDNA encoding a defective receptor protein tyrosine kinase. *Gene* 328:75–84.

Junqueira Alves C, Yotoko K, Zou H, Friedel RH. 2019. Origin and evolution of plexins, semaphorins, and Met receptor tyrosine kinases. *Sci Rep.* 9(1):1970.

Jura N, Endres NF, Engel K, Deindl S, Das R, Lamers MH, Wemmer DE, Zhang X, Kuriyan J. 2009. Mechanism for activation of the EGF receptor catalytic domain by the juxtamembrane segment. *Cell* 137(7):1293–1307.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods.* 14(6):587–589.

Kannan N, Neuwald AF. 2004. Evolutionary constraints associated with functional specificity of the CMGC protein kinases MAPK, CDK, GSK, SRPK, DYRK, and CK2alpha. *Protein Sci.* 13(8):2059–2077.

Keller O, Odronitz F, Stanke M, Kollmar M, Waack S. 2008. Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC Bioinformatics.* 9:278.

King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I, et al. 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* 451(7180):783–788.

Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J Mol Biol.* 305(3):567–580.

Kwon A, John M, Ruan Z, Kannan N. 2018. Coupled regulation by the juxtamembrane and sterile α motif (SAM) linker is a hallmark of ephrin tyrosine kinase evolution. *J Biol Chem.* 293(14):5102–5116.

Kwon A, Scott S, Taujale R, Yeung W, Kochut KJ, Eyers PA, Kannan N. 2019. Tracing the origin and evolution of pseudokinases across the tree of life. *Sci. Signal.* 12(578):eaav3810./

Lemmon MA, Schlessinger J. 2010. Cell signaling by receptor tyrosine kinases. *Cell* 141(7):1117–1134.

Lemmon MA, Schlessinger J, Ferguson KM. 2014. The EGFR family: not so prototypical receptor tyrosine kinases. *Cold Spring Harb Perspect Biol.* 6(4):a020768.

Levinson NM, Kuchment O, Shen K, Young MA, Koldobskiy M, Karplus M, Cole PA, Kuriyan J. 2006. A Src-like inactive conformation in the Abl tyrosine kinase domain. *PLoS Biol.* 4(5):e144.

Liongue C, Sertori R, Ward AC. 2016. Evolution of cytokine receptor signaling. *J Immunol.* 197(1):11–18.

Liu BA, Shah E, Jablonowski K, Stergachis A, Engelmann B, Nash PD. 2011. The SH2 domain–containing proteins in 21 species establish the provenance and scope of phosphotyrosine signaling in eukaryotes. *Sci Signal.* 4(202):ra83.

Locascio LE, Donoghue DJ. 2013. KIDs rule: regulatory phosphorylation of RTKs. *Trends Biochem Sci.* 38(2):75–84.

Manni S, Kisko K, Schleier T, Missimer J, Ballmer-Hofer K. 2014. Functional and structural characterization of the kinase insert and the carboxy terminal domain in VEGF receptor 2 activation. *Faseb J.* 28(11):4914–4923.

Manning G, Plowman GD, Hunter T, Sudarsanam S. 2002. Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci.* 27(10):514–520.

Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. 2002. The protein kinase complement of the human genome. *Science* 298(5600):1912–1934.

Manning G, Young SL, Miller WT, Zhai Y. 2008. The protist, *Monosiga brevicollis*, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan. *Proc Natl Acad Sci U S A.* 105(28):9674–9679.

Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. 2017. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* 45(D1):D200–D203.

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, et al. 2011. CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res.* 39(Database issue):D225–D229.

Matus DQ, Thomsen GH, Martindale MQ. 2007. FGF signaling in gastrulation and neural development in *Nematostella vectensis*, an anthozoan cnidarian. *Dev Genes Evol.* 217(2):137–148.

Miller WT. 2012. Tyrosine kinase signaling and the emergence of multicellularity. *Biochim Biophys Acta.* 1823(6):1053–1057.

Mirza A, Mustafa M, Talevich E, Kannan N. 2010. Co-conserved features associated with cis regulation of erbb tyrosine kinases. *PLoS One.* 5(12):e14310.

Mitchell JM, Nichols SA. 2019. Diverse cell junctions with unique molecular composition in tissues of a sponge (Porifera). *Evodevo* 10:26.

Mitra SK, Hanson DA, Schlaepfer DD. 2005. Focal adhesion kinase: in command and control of cell motility. *Nat Rev Mol Cell Biol.* 6(1):56–68.

Mócsai A, Ruland J, Tybulewicz VLJ. 2010. The SYK tyrosine kinase: a crucial player in diverse biological functions. *Nat Rev Immunol.* 10(6):387–402.

Modi V, Dunbrack RL. 2019. A structurally-validated multiple sequence alignment of 497 human protein kinase domains. *Sci Rep.* 9(1):19790.

Mohammadi M, Schlessinger J, Hubbard SR. 1996. Structure of the FGF receptor tyrosine kinase domain reveals a novel autoinhibitory mechanism. *Cell* 86(4):577–587.

Mohanty S, Oruganty K, Kwon A, Byrne DP, Ferries S, Ruan Z, Hanold LE, Katiyar S, Kennedy EJ, Eyers PA, et al. 2016. Hydrophobic core variations provide a structural framework for tyrosine kinase evolution and functional specialization. *PLoS Genet.* 12(2):e1005885.

Monahan-Earley R, Dvorak AM, Aird WC. 2013. Evolutionary origins of the blood vascular system and endothelium. *J Thromb Haemost*. 11(Suppl 1):46.

Murphy JM, Zhang Q, Young SN, Reese ML, Bailey FP, Eyers PA, Ungureanu D, Hammaren H, Silvennoinen O, Varghese LN, et al. 2014. A robust methodology to subclassify pseudokinases based on their nucleotide-binding properties. *Biochem J*. 457(2):323–334.

Nars M, Vihinen M. 2001. Coevolution of the domains of cytoplasmic tyrosine kinases. *Mol Biol Evol*. 18(3):312–321.

Neuwald AF. 2009. Rapid detection, classification and accurate alignment of up to a million or more related protein sequences. *Bioinformatics* 25(15):1869–1875.

Neuwald AF. 2011. Surveying the manifold divergence of an entire protein class for statistical clues to underlying biochemical mechanisms. *Stat Appl Genet Mol Biol*. 10:Article 36.[Internet] Available from: https://www.degruyter.com/document/doi/10.2202/1544-6115.1666/html

Neuwald AF. 2014. A Bayesian sampler for optimization of protein domain hierarchies. *J Comput Biol*. 21(3):269–286.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32(1):268–274.

Oruganty K, Talathi NS, Wood ZA, Kannan N. 2013. Identification of a hidden strain switch provides clues to an ancient structural mechanism in protein kinases. *Proc Natl Acad Sci U S A*. 110(3):924–929.

Parfrey LW, Lahr DJG, Knoll AH, Katz LA. 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc Natl Acad Sci U S A*. 108(33):13624–13629.

Pasquale EB. 2010. Eph receptors and ephrins in cancer: bidirectional signalling and beyond. *Nat Rev Cancer*. 10(3):165–180.

Plowman GD, Sudarsanam S, Bingham J, Whyte D, Hunter T. 1999. The protein kinases of *Caenorhabditis elegans*: A model for signal transduction in multicellular organisms. *Proc Natl Acad Sci U S A*. 96(24):13603–13610.

Poliakov A, Cotrina M, Wilkinson DG. 2004. Diverse roles of Eph receptors and ephrins in the regulation of cell migration and tissue assembly. *Dev Cell*. 7(4):465–480.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*. 35(Database issue):D61–D65.

Richter DJ, King N. 2013. The genomic and cellular foundations of animal origins. *Annu Rev Genet*. 47:509–537.

Robinson DR, Wu Y-M, Lin S-F. 2000. The protein tyrosine kinase family of the human genome. *Oncogene* 19(49):5548–5557.

Schindler T, Sicheri F, Pico A, Gazit A, Levitzki A, Kuriyan J. 1999. Crystal structure of Hck in complex with a Src family—selective tyrosine kinase inhibitor. *Mol Cell*. 3(5):639–648.

Shah NH, Amacher JF, Nocka LM, Kuriyan J. 2018. The Src module: an ancient scaffold in the evolution of cytoplasmic tyrosine kinases. *Crit Rev Biochem Mol Biol*. 53(5):535–563.

Shiu S-H, Li W-H. 2004. Origins, lineage-specific expansions, and multiple losses of tyrosine kinases in eukaryotes. *Mol Biol Evol*. 21(5):828–840.

Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier MEA, Mitros T, Richards GS, Conaco C, Dacre M, Hellsten U, et al. 2010. The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature* 466(7307):720–726.

Suga H, Dacre M, A de M, Shalchian-Tabrizi K, Manning G, Ruiz-Trillo I. 2012. Genomic survey of premetazoans shows deep conservation of cytoplasmic tyrosine kinases and multiple radiations of receptor tyrosine kinases. *Sci Signal*. 5(222):ra35.

Suga H, Sasaki G, Kuma K, Nishiyori H, Hirose N, Su Z-H, Iwabe N, Miyata T. 2008. Ancient divergence of animal protein tyrosine kinase genes demonstrated by a gene family tree including choanoflagellate genes. *FEBS Lett*. 582(5):815–818.

Suga H, Torruella G, Burger G, Brown MW, Ruiz-Trillo I. 2014. Earliest holozoan expansion of phosphotyrosine signaling. *Mol Biol Evol*. 31(3):517–528.

Taskinen B, Ferrada E, Fowler DM. 2017. Early emergence of negative regulation of the tyrosine kinase Src by the C-terminal Src kinase. *J Biol Chem*. 292(45):18518–18529.

Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, et al. 2019. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res*. 47(D1):D941–D947.

Tong K, Wang Y, Su Z. 2017. Phosphotyrosine signalling and the origin of animal multicellularity. *Proc R Soc B Biol Sci*. 284:20170681.

Wang H, Brautigan DL. 2002. A novel transmembrane Ser/Thr kinase complexes with protein phosphatase-1 and inhibitor-2. *J Biol Chem*. 277(51):49605–49612.

Wang Q, Vogan EM, Nocka LM, Rosen CE, Zorn JA, Harrison SC, Kuriyan J. 2015. Autoinhibition of Bruton's tyrosine kinase (Btk) and activation by soluble inositol hexakisphosphate. *eLife* 4:e06074.

Xu W, Doshi A, Lei M, Eck MJ, Harrison SC. 1999. Crystal structures of c-Src reveal features of its autoinhibitory mechanism. *Mol Cell*. 3(5):629–638.

Zhang X, Gureasko J, Shen K, Cole PA, Kuriyan J. 2006. An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell* 125(6):1137–1149.