TRACING THE EVOLUTION OF THE TYROSINE KINOME

FROM SEQUENCE TO FUNCTION

by

ANNIE HYUNJIN KWON

(Under the Direction of Natarajan Kannan)

ABSTRACT

Tyrosine kinases are regulated through diverse mechanisms that control enzymatic outputs based on the presence of specific molecular cues. The biomedical relevance of tyrosine kinases has driven extensive research on a number of tyrosine kinases such as Src and the oncogene epidermal growth factor receptor, however, regulatory mechanisms in many other tyrosine kinases still remain unknown. In this work I apply rigorous, large scale sequence comparisons to study the evolution of tyrosine kinases. First, I determine a new hierarchical classification of the tyrosine kinome, which reflects the modular evolution of distinct and anciently conserved regulatory cores. Next, I use computational and experimental approaches to demonstrate how sequence features unique to the Ephrin family of tyrosine kinases contribute to an allosteric coupling between the juxtamembrane and sterile α motif linker to uniquely regulate the activation of Ephrin kinases. Last, through a comprehensive analysis of pseudokinase sequences, I propose a new classification of pseudokinases families that reflects their evolution from diverse canonical kinases and through the selection of sequence motifs that stabilize unique inactive kinase domain conformations and contribute to other non-enzymatic functions. The findings presented here shed light on how diverse functions and regulatory mechanisms have

evolved in the tyrosine kinome through sequence evolution, and we also provide novel hypotheses of undiscovered kinase functions.

INDEX WORDS: tyrosine kinase, kinase evolution, kinase regulation, allostery

TRACING THE EVOLUTION OF THE TYROSINE KINOME FROM SEQUENCE TO FUNCTION

by

ANNIE HYUNJIN KWON

B.S., Emory University, 2012

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2019

TRACING THE EVOLUTION OF THE TYROSINE KINOME

FROM SEQUENCE TO FUNCTION

by

ANNIE HYUNJIN KWON

Major Professor:	Natarajan Kannan
Committee:	Eileen J. Kennedy
	Zachary A. Wood
	Jonathan Arnold

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2019

To my loving family, who made this possible

♦ ♦ ♦

ACKNOWLEDGEMENTS

I am grateful to my dissertation advisor, Natarajan Kannan, for his continued guidance and support during my Ph.D. study, as well as to my committee members, Eileen J. Kennedy, Zachary A. Wood, and Jonathan Arnold, for their invaluable advice. I acknowledge members of my lab, past and present, who I have had the pleasure to work with over the years: Samiksha Katiyar, Wayland Yeung, Safal Shrestha, Aaryka Venkat, Rahil Taujale, Liang-Chin Huang, Zheng Ruan, Mihir John, Steven Scott, Smita Mohanty, Krishnadev Oruganty, Daniel McSkimming, Eric Talevich, and Shima Dastgheib. I also express my gratitude to Patrick A. Eyers, the Franklin College, and the University of Liverpool for the unique opportunity to conduct research abroad under the UGA-Liverpool travel fellowship.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter 1

Introduction and literature review

1.1 Motivation

Specific mutation patterns in tyrosine kinases can be linked to a variety of diseases including immune disorders, diabetes, and numerous cancer types. Their biomedical relevance has warranted substantial efforts to understand their biological functions, how they become dysregulated to cause disease, and how they can be targeted pharmaceutically. While the molecular machinery that controls tyrosine kinase function is well understood for a handful of cases, the atomic-level details of how most tyrosine kinases are regulated are still unknown or difficult to piece together from disparate literature and data sources. As a consequence, it is difficult to predict which of the many disease-associated mutations in tyrosine kinases alter enzyme function and contribute to disease ('driver' mutations), and which of these mutations arise as a consequence of random mutation and have no effect on disease progression ('passenger' mutations) (*1, 2*). Their implication in many diseases also make tyrosine kinases some of the most pursued targets for the development of new drugs (*1-6*). However, despite the development of over 40 FDA approved tyrosine kinase inhibitors, limited specificity and off-target effects remain major hurdles to the efficacy and safety of inhibitors and occur due to the

highly similar drug-binding pocket shared across all protein kinases. Recently, scientists have developed some allosteric inhibitors that increase selectivity by binding outside of the primary drug-binding pocket and into other functional regions of the enzyme (*5, 7*). Still, identifying which allosteric sites can be targeted remains a challenge due to the insufficient understanding of the ninety human tyrosine kinases and their unique functional mechanisms.

We therefore seek to computationally and experimentally characterize the unique features of distinct tyrosine kinases. Through a comparative approach, we identify how the tyrosine kinases diversified through the evolution of protein sequence, making it possible to discover previously unknown molecular mechanisms in tyrosine kinases. By studying the unique features of various tyrosine kinases, we shed light on the fundamental aspects of their functions and on the evolutionary constraints that shape them.

1.2 Background

1.2.1 Protein Kinases

Protein kinases comprise one of the largest enzyme families and make up nearly 2% of the human genome (*8*). These enzymes catalyze protein phosphorylation, where the γ phosphate of an adenosine tri-phosphate molecule is transferred to a protein substrate. This modification of substrate proteins, which is reversible through separate enzymes called phosphatases, transduces biochemical signals by affecting the structure and biochemical properties of substrate proteins, and often also their protein interactions and subcellular locations (*9, 10*). The universality of phosphorylation in biology is mirrored in the fact that proteins across all living organisms from archaea, to bacteria, to eukaryotes and viruses are phosphorylated (*11, 12 , 13, 14*), and it is estimated that approximately 75% of the human proteome undergoes phosphorylation (*15*). As

such, protein kinases play fundamental roles in many key biological functions including cell division and differentiation, metabolism, and immune processes (*13*) and reflects why mutations in protein kinases can lead to a variety of diseases including many forms of cancer, developmental disorders, diabetes, cardiovascular disease, and immune disorders (*1, 7, 16*).

1.2.2 Conserved features of protein kinase sequence, structure, and function

Protein kinases are evolutionarily related, therefore, they share common features in protein sequence, structure, and function. Multiple sequence alignments of diverse protein kinases have shown that some amino acids are nearly invariant across all protein kinase sequences. For example, some of the most evolutionarily conserved amino acids, or sequence motifs, include a 'HRDxxxxN' motif and a 'DFG' motif (*17, Hanks, 1995 #17, 18*). Protein kinases also share a common protein fold as observed from crystal structures of diverse protein kinases that have been solved using X-ray crystallography (*19-21*). The protein kinase fold is characterized by an N-terminal lobe containing five beta strands (termed the β1, β2, β3, β4, and β5 strands) and a helical segment called the αC helix and a C-terminal lobe consisting primarily of alpha helices (termed the αD, αE, αF, αG, αH, and αI helices). Structurally, the N-terminal lobe sits above the C-terminal lobe, and two loop regions containing the 'HRDxxxxN' and 'DFG' motifs termed the catalytic loop and the activation loop lie between the two lobes.

Such conserved sequence and structural motifs give way to a common catalytic mechanism that is shared by all protein kinases. For example, ATP binds within the N-terminal lobe of the kinase domain, where the α and β phosphates are stabilized by a salt bridge between a conserved lysine from the β3 strand and a conserved glutamate from the C-helix. In addition, a glycine rich GxGxxG motif in the loop region between the β1 and β2 strands caps the phosphate

groups of ATP to further stabilize the molecule for phosphotransfer (*20-22*). Protein kinases also require binding of metal ions such as magnesium to chelate the β and γ phosphates of ATP, which is achieved through a conserved asparagine of the 'HRDxxxxN' motif in the catalytic loop and the conserved aspartate of the 'DFG' motif of the activation loop, which also helps bind ATP (*20, 21*). Whereas ATP binds in the N-terminal lobe, protein substrates bind mostly to the C-terminal lobe of the kinase domain, particularly to the αG helix and the activation loop (*19*). Importantly, the aspartate of the catalytic loop 'HRDxxxxN' motif serves as the catalytic base for the phosphotransfer reaction and is invariant across protein kinase sequences like other amino acids important for ATP and substrate binding (*22*).

Because phosphorylation by protein kinases propagates cellular signals, the enzymatic activities of protein kinases must be tightly controlled such that phosphorylation reactions only occur given the appropriate cellular cues. As such, protein kinases toggle between 'on' and 'off' states, often through modifying the structure and dynamics of the kinase domain (*18, 23*). To achieve this, several regulatory elements are also conserved across protein kinases. For example, the flexibility of the activation loop allows it to adopt numerous different conformations and is therefore utilized as a regulatory element across all protein kinases (*18*). Specifically, in the active state, the activation loop adopts an outward and extended conformation, which permits the binding of protein substrates, and this active conformation is further stabilized by phosphorylation of the activation loop in most protein kinases (*23, 24*). In addition, the αC helix can undergo dynamic conformational transitions to regulate activity, where an inward conformation of the αC helix allows the conserved salt bridge between the β3 strand lysine and the αC helix glutamate to form and thus allows effective ATP binding, whereas an outward movement of the αC helix prevents formation of the salt bridge and prohibits effective ATP

binding (*23*). Another regulatory element that is conserved across protein kinases is the hydrophobic regulatory spine, which is assembled in the active state such that it spans the N- and C-terminal lobes, but is disassembled in the inactive state (*25*). As such, conserved sequence and structural elements in protein kinases not only preserve a common catalytic mechanism, but also a common dynamic framework to regulate catalysis.

1.2.3 Kinase evolution and the diversification of kinase function and regulation

Protein kinases occupy diverse functional niches, and despite their conservation of common sequence, structural, and functional elements, they have also diversified such that distinct protein kinases respond to their own unique cellular, spatial, and temporal cues. For example, the intracellular tyrosine kinase domain of the Epidermal Growth Factor Receptor (EGFR) only becomes activated when its extracellular ligand binding domain senses the presence of specific ligands such as epidermal growth factor (EGF) and transforming growth factor α (TGFα) (*26*). Specifically, extracellular ligand binding prompts the intracellular EGFR kinase domains to dimerize, which allosterically activates one EGFR molecule to phosphorylate tyrosine residues on the other. Specific phosphorylation patterns on these sites can then initiate a variety of downstream pathways such as the ERK MAPK or the AKT-PI3K pathways that control different aspects of cell cycle progression (*26*).

Such precise regulation of signaling pathways requires that the cellular cues that prompt kinase activation or inactivation are translated to the kinase domain in some way to modulate its catalytic output. This occurs through a variety of mechanisms that include dimerization, as in the case with EGFR, as well as other protein-protein interactions, the binding or unbinding of protein domains or flexible linkers flanking the kinase domain, post-translational modifications, or localization to subcellular compartments (*16, 23, 27*). Often, these mechanisms impart control

on the catalytic output of the kinase domain by structurally manipulating conserved regulatory features of protein kinases. For example, dimerization allosterically activates EGFR because one molecule in the dimer sits on the αC helix of the other, stabilizing the αC helix in an inward conformation that permits ATP binding (*28*).

Such regulatory mechanisms are often shared among closely related protein kinases and encoded within shared sequence motifs (*27, 29-32*). The rapidly increasing number of available protein kinase sequences through the last three decades has allowed the protein kinases to be classified into closely related groups, families, and subfamilies based on sequence similarity, protein domain organization, and conservation across species (*8, 13, 19*). The resulting classification identified seven major kinase groups, as well as an "Other" group that includes protein kinases that do not distinctly classify into one of the major groups and an "Atypical" group that includes protein kinases that exhibit protein kinase activity but lack significant sequence homology to canonical protein kinases (*13*). The tyrosine kinases comprise one of the major seven groups and have uniquely evolved from other protein kinases to specifically phosphorylate tyrosine residues on substrate proteins rather than serine or threonine. Interestingly, sequence motifs unique to tyrosine kinases not only contribute to the specific recognition of tyrosine on substrate proteins, but also contribute to a tyrosine kinase-specific regulatory spine that operates in a biochemically dissimilar manner than that in other protein kinases (*31*). In this way, tyrosine kinases have evolutionarily diverged from other protein kinases through its specific phosphorylation of tyrosine residues, as well as through a fundamental difference in how they toggle between 'on' and 'off' states.

While distinct tyrosine kinase families have further diversified to fulfill different functional niches through the evolution of unique regulatory mechanisms, one peculiar way in

which the tyrosine kinases have diversified is through the evolution of pseudokinases, which are protein kinases that lack catalytic function due to the loss of key catalytic residues such as the 'HRDxxxxN' and 'DFG' aspartates . Nine of the ninety human tyrosine kinases can be classified as pseudokinases due to their lack of at least one key catalytic residue, and they each play important biological roles despite their lack of catalytic output (*8, 33, 34*). For example, the Jak family pseudokinases and HER3 pseudokinase can all form complexes with active tyrosine kinase domains to regulate their activity (*33*). However, how pseudokinases evolved from active protein kinases and how their noncatalytic functions emerged still remain unknown. Furthermore, the elusive functions of lesser studied tyrosine pseudokinases such as PTK7 and STYK1 are difficult to discover due to the lack of measurable catalytic output and the intrinsic difficulty in discovering protein-protein interactions (*34*).

1.3 Key challenges and unresolved questions

Understanding the molecular mechanisms controlling different tyrosine kinases is critical for understanding and treating diseases such as cancer. However, discovering new molecular mechanisms in proteins is difficult due to the restricted scope of mutagenesis experiments, limited information in crystal structure data, if it is available, and the inherent difficulty in discovering protein-protein interactions and linking such interactions to specific cell signaling pathways and biological phenotypes. Investigating the patterns of conservation and variation in sequence across a protein superfamily can offer a complementary approach to study protein function by spotlighting potential regions of interest. My working hypothesis is that tyrosine kinases have evolved new functions through sequence evolution, and therefore their molecular mechanisms can be inferred by examining how tyrosine kinase sequences have changed over time. Examining the sequence features unique to specific tyrosine kinases and tyrosine kinase

families can provide novel testable hypotheses about the distinct mechanisms of tyrosine kinases, which can be further explored using traditional biochemical methods. In addition, adapting this information into an ontological format can provide a minable resource to inform future studies on tyrosine kinases and tyrosine kinase evolution.

1.4 Major research questions addressed

The following chapters address the stated knowledge gaps by investigating the diversification of tyrosine kinase function through the evolution of tyrosine kinase sequence. The research studies outlined below use sequence comparisons to identify sequence motifs selected throughout different branches of tyrosine kinase evolution, which we link to various aspects of tyrosine kinase function and regulation using further data mining and computational and experimental investigations. A broader investigation into pseudokinase evolution, both in the tyrosine kinases as well as in serine/threonine kinases sheds new insight into pseudokinase functions and the evolutionary constraints that mold them. In each study, I identify and annotate residue-level features unique to distinct tyrosine kinase lineages to provide in an ontological format.

1.4.1 A novel classification of the tyrosine kinome reveals anciently conserved regulatory modules

The tyrosine kinases can be classified into thirty major evolutionarily distinct families (*13, 19*). However, certain similarities in structure, domain architecture, and regulatory features have been noted to be shared across multiple tyrosine kinase families (*35-37*), suggesting that some tyrosine kinase families might be classified into subgroups. In this study, I derive an alternative classification of the tyrosine kinases that arranges them hierarchically into subgroups,

families, and subfamilies based on unique patterns of sequence conservation and divergence. From this new classification, I infer how sequence motifs unique to distinct subgroups of tyrosine kinases encode anciently conserved regulatory cores that distinguish each subgroup. Examining sequence motifs associated with the divergence of tyrosine kinase subgroups into families sheds additional insight into how these regulatory cores have been fine-tuned to provide additional regulation. The study provides an emerging scheme of how tyrosine kinases evolve new regulatory functions through the modular evolution of a common regulatory core.

1.4.2 Coupled regulation by the juxtamembrane and sterile α motif (SAM) linker is a hallmark of ephrin tyrosine kinase evolution

The ephrin (Eph) family of receptor tyrosine kinases comprises the largest family of tyrosine kinases with fourteen distinct members in humans and plays important roles in developmental processes and adult tissue homeostasis (*38*). The role of the juxtamembrane in the autoinhibition of Eph kinases is well understood (*39, 40*), as is the general role of activation loop phosphorylation in the activation of protein kinases (*24*). However, other aspects of Eph regulation is not well understood. For example, how Eph members oligomerize and how oligomerization affects catalytic activity is still unknown.

In this study, sequence motifs unique to the Eph family were identified to tether flanking protein segments to the kinase domain. Mutating residues that tether either of these flanking segments, the N-terminal juxtamembrane and the C-terminal sterile α motif (SAM) linker caused an increase in the rate of autophosphorylation of EphA3, thus these residues were determined to be important for the autoinhibition of Eph activity. Examining the behavior of double and triple mutants harboring mutations in one or both distinct networks, as well as additional molecular dynamics studies, demonstrates how the SAM linker is dynamically, and thus functionally,

coupled to the juxtamembrane. I propose a model in which the SAM linker provides an additional source of entropy that can be modulated through the tethering/untethering onto a conserved site in the Eph kinase domain, which in turn influences Eph activation by controlling the rate of juxtamembrane autophosphorylation. The identification of the SAM linker tethering interface highlights a new potential allosteric site that might be utilized in novel Eph-targeting small molecule drugs. The study also identifies other highly distinguishing motifs unique to the Eph family that may also be involved in other aspects of Eph regulation and that warrant further investigation.

1.4.3 Tracing the Origin and Evolution of Pseudokinases Across the Tree of Life

Pseudokinases are protein kinases that lack catalytic activity but nonetheless perform critical signaling roles through non-enzymatic functions such as molecular recruitment and allosteric modulation of active enzymes (*33*). Several pseudokinases have evolved within the tyrosine kinase group, including the Janus tyrosine kinase family (JAKs), the epidermal growth factor receptor (EGFR) family member HER3, the Eph members EphA10 and EphB6, and the lesser studied PTK7 and STYK1 pseudokinases. Despite the estimate that some 10% of protein kinases are predicted to be pseudokinases in vertebrates, the prevalence of pseudokinases outside of vertebrates, for example, in bacteria and fungi, as well as how pseudokinases have emerged through the evolution of the ancient protein kinase fold has not yet been systematically examined.

In this study, we perform a comprehensive analysis of 10,092 archaeal, bacterial, and eukaryotic proteomes to determine the prevalence of pseudokinase sequences across the tree of life. The study establishes that pseudokinases are present across all three domains of life, and these diverse pseudokinase sequences can be classified into 86 pseudokinase families.

Expansions of pseudokinase families in certain organisms and taxonomic groups, such as the plant specific LysM family and the *R. irregularis*-specific Rig1 family of pseudokinases, suggest that pseudokinases may have evolved to play roles unique to the distinct biologies of various organisms. Further examining the evolutionary constraints associated with the evolution of pseudokinases shows how different pseudokinase families have evolved novel ways to stabilize distinct inactive kinase domain conformations. A curated resource of annotated pseudokinase sequences provides a minable dataset that can be used to further characterize the unique functions of various pseudokinases and how these emerged through protein kinase evolution.

Bibliography

1. L. J. Wilson *et al.*, New Perspectives, Opportunities, and Challenges in Exploring the Human Protein Kinome. *Cancer Res* **78**, 15-29 (2018).

2. C. Greenman *et al.*, Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-158 (2007).

3. A. L. Hopkins, C. R. Groom, The druggable genome. *Nat Rev Drug Discov* **1**, 727-730 (2002).

4. A. Gschwind, O. M. Fischer, A. Ullrich, The discovery of receptor tyrosine kinases: targets for cancer therapy. *Nat Rev Cancer* **4**, 361-370 (2004).

5. W. R. Montor, A. Salas, F. H. M. Melo, Receptor tyrosine kinases and downstream pathways as druggable targets for cancer treatment: the current arsenal of inhibitors. *Mol Cancer* **17**, 55 (2018).

6. L. K. Shawver, D. Slamon, A. Ullrich, Smart drugs: tyrosine kinase inhibitors in cancer therapy. *Cancer Cell* **1**, 117-123 (2002).

7. K. S. Bhullar *et al.*, Kinase-targeted cancer therapies: progress, challenges and future directions. *Molecular Cancer* **17**, (2018).

8. G. Manning, D. B. Whyte, R. Martinez, T. Hunter, S. Sudarsanam, The protein kinase complement of the human genome. *Science* **298**, 1912-1934 (2002).

9. A. Ullrich, J. Schlessinger, Signal transduction by receptors with tyrosine kinase activity. *Cell* **61**, 203-212 (1990).

10. T. Hunter, Why nature chose phosphate to modify proteins. *Philos Trans R Soc Lond B Biol Sci* **367**, 2513-2516 (2012).

11. D. Esser *et al.*, Protein phosphorylation and its role in archaeal signal transduction. *FEMS Microbiol Rev* **40**, 625-647 (2016).

12. N. Kannan, S. S. Taylor, Y. Zhai, J. C. Venter, G. Manning, Structural and functional diversity of the microbial kinome. *PLoS Biol* **5**, e17 (2007).

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

14. D. P. Leader, M. Katan, Viral aspects of protein phosphorylation. *J Gen Virol* **69 (Pt 7)**, 1441-1464 (1988).

15. K. Sharma *et al.*, Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep* **8**, 1583-1594 (2014).

16. M. A. Lemmon, J. Schlessinger, Cell signaling by receptor tyrosine kinases. *Cell* **141**, 1117-1134 (2010).

17. S. K. Hanks, A. M. Quinn, T. Hunter, The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science* **241**, 42-52 (1988).

18. N. Kannan, A. F. Neuwald, Did protein kinase regulatory mechanisms evolve through elaboration of a simple structural component? *J Mol Biol* **351**, 956-972 (2005).

19. S. K. Hanks, T. Hunter, Protein Kinases .6. The Eukaryotic Protein-Kinase Superfamily - Kinase (Catalytic) Domain-Structure and Classification. *Faseb Journal* **9**, 576-596 (1995).

20. S. R. Hubbard, L. Wei, L. Elis, W. A. Hendrickson, Crystal-Structure of the Tyrosine Kinase Domain of the Human Insulin-Receptor. *Nature* **372**, 746-754 (1994).

21. D. R. Knighton *et al.*, Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* **253**, 407-414 (1991).

22. J. A. Adams, Kinetic and catalytic mechanisms of protein kinases. *Chem Rev* **101**, 2271-2290 (2001).

23. M. Huse, J. Kuriyan, The conformational plasticity of protein kinases. *Cell* **109**, 275-282 (2002).

24. B. Nolen, S. Taylor, G. Ghosh, Regulation of protein kinases; controlling activity through activation segment conformation. *Mol Cell* **15**, 661-675 (2004).

25. C. L. McClendon, A. P. Kornev, M. K. Gilson, S. S. Taylor, Dynamic architecture of a protein kinase. *Proc Natl Acad Sci U S A* **111**, E4623-4631 (2014).

26. P. Wee, Z. Wang, Epidermal Growth Factor Receptor Cell Proliferation Signaling Pathways. *Cancers (Basel)* **9**, (2017).

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

28. X. Zhang, J. Gureasko, K. Shen, P. A. Cole, J. Kuriyan, An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell* **125**, 1137-1149 (2006).

29. N. Kannan, N. Haste, S. S. Taylor, A. F. Neuwald, The hallmark of AGC kinase functional divergence is its C-terminal tail, a cis-acting regulatory module. *Proc Natl Acad Sci U S A* **104**, 1272-1277 (2007).

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

31. S. Mohanty *et al.*, Hydrophobic Core Variations Provide a Structural Framework for Tyrosine Kinase Evolution and Functional Specialization. *PLoS Genet* **12**, e1005885 (2016).

32. T. Nguyen, Z. Ruan, K. Oruganty, N. Kannan, Co-conserved MAPK features couple D-domain docking groove to distal allosteric sites via the C-terminal flanking tail. *PLoS One* **10**, e0119636 (2015).

33. J. M. Murphy, P. D. Mace, P. A. Eyers, Live and let die: insights into pseudoenzyme mechanisms from structure. *Curr Opin Struct Biol* **47**, 95-104 (2017).

34. J. Boudeau, D. Miranda-Saavedra, G. J. Barton, D. R. Alessi, Emerging roles of pseudokinases. *Trends Cell Biol* **16**, 443-452 (2006).

35. S. C. Artim, J. M. Mendrola, M. A. Lemmon, Assessing the range of kinase autoinhibition mechanisms in the insulin receptor family. *Biochem J* **448**, 213-220 (2012).

36. H. B. Chen *et al.*, A molecular brake in the kinase hinge region regulates the activity of receptor tyrosine kinases. *Molecular Cell* **27**, 717-730 (2007).

37. N. H. Shah, J. F. Amacher, L. M. Nocka, J. Kuriyan, The Src module: an ancient scaffold in the evolution of cytoplasmic tyrosine kinases. *Crit Rev Biochem Mol Biol* **53**, 535-563 (2018).

38. E. M. Lisabeth, G. Falivelli, E. B. Pasquale, Eph receptor signaling and ephrins. *Cold Spring Harb Perspect Biol* **5**, (2013).

39. T. L. Davis *et al.*, Autoregulation by the juxtamembrane region of the human ephrin receptor tyrosine kinase A3 (EphA3). *Structure* **16**, 873-884 (2008).

40. L. E. Wybenga-Groot *et al.*, Structural basis for autoinhibition of the Ephb2 receptor tyrosine kinase by the unphosphorylated juxtamembrane region. *Cell* **106**, 745-757 (2001).

Chapter 2

A novel classification of the tyrosine kinome reveals anciently conserved regulatory modules

Kwon, Annie. To be submitted to *Genome biology*

Abstract

Tyrosine kinases have undergone millions of years of evolution, acquiring diverse mechanisms to control their enzymatic activity. Understanding how evolution has shaped such diverse regulatory mechanisms across the tyrosine kinases can facilitate the discovery of currently unknown mechanisms, especially in lesser studied tyrosine kinases, as well as to provide insight into how disease mutations deregulate kinases. However, the current, widely accepted classification of protein kinases into distinct evolutionarily and functionally related families fails to reflect that common kinase functions and regulatory features are often shared across multiple protein kinase families. We present a comprehensive study of tyrosine kinases to determine a new hierarchical classification that reclassifies the tyrosine kinases into larger subgroups and families that are distinguished by common functional features. We identify three major subgroupings within the tyrosine kinase group that are each defined by a common regulatory module shared across multiple tyrosine kinase families. Examining the family-level diversification of conserved regulatory modules demonstrates the modular evolution of tyrosine kinases and illustrates how a hierarchical classification scheme sets a foundation for meaningful comparative studies of protein function and evolution.

2.1 Introduction

Tyrosine kinases propagate cellular signals by phosphorylating tyrosine residues on substrate proteins. Their integral role in signaling pathways controlling diverse functions from cell growth and differentiation to angiogenesis, cell migration, and apoptosis is mirrored in their propensity to be mutated in cancer (*1-3*). The catalytic mechanism of tyrosine kinases is ubiquitously conserved and shared among a larger group of evolutionarily related proteins that include serine/threonine kinases (*4-9*). Despite the conservation of a common enzymatic

mechanism, catalytic output is regulated in incredibly diverse and intricate ways across the kinome, allowing this single superfamily of enzymes to fill many different biological niches. However, how evolution has shaped the regulatory landscape of different protein kinases is still poorly understood.

The regulatory mechanisms that control kinase activity are numerous and diverse and include post-translational modifications (*1, 10*), oligomerization (*1, 10, 11*), interactions with other proteins or domains (*10*), tethering/untethering of flexible protein linkers (*11-14*), and various conformational transitions within the kinase domain (*1, 15-17*). These mechanisms can facilitate kinase activation; for example, extracellular ligand stimulation induces the formation of an intracellular asymmetric dimer by EGFR family members, which stabilizes the "receiver" kinase in an active state (*11*). In contrast, autoinhibitory mechanisms can inhibit catalytic activity, as seen in many receptor tyrosine kinases that are autoinhibited by the juxtamembrane until extracellular signals are received (*14*). In addition, mechanisms exist to regulate the localization, degradation, and downstream signaling of kinases (*10*).

Regulatory mechanisms are often shared among closely related protein kinases (*15, 18*). For example, Src family kinases are autoinhibited by their N-terminal SH3 and SH2 domains, which assemble onto the backside of the kinase domain and are locked in place by a phosphorylated tyrosine conserved within the C-terminal tails of Src kinases. The Src kinases were one of the first recognized protein kinase families in early catalogues and classifications of kinases, and they comprise one of the largest tyrosine kinases families (*4, 5, 7*). Further population and refinement of the protein kinase classification by Manning *et al.* was based on phylogenetic sequence comparisons, domain organization, and biological function and yielded 30 major tyrosine kinase families, 20 of which are receptor tyrosine kinase families (*7, 19*).

Large expansions of tyrosine kinase families in some organisms and taxonomic groups, such as in the choanoflagellate *Monosiga brevicollis* or in Nematoda, have led to species-specific divergence and often constitute their own distinct families (*20, 21*). As such, Manning's classification designates over 20 organism-specific tyrosine kinase families.

Manning's classification of protein kinases 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 (R-spine) of tyrosine kinases relative to serine/threonine kinases (*22, 23*), as well as into the regulatory mechanisms that evolved uniquely in the EGFR (*24*) and Eph (*13*) families of tyrosine kinases. In addition to the unique features of 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 (*18*). In addition, recent work by Andreotti et al demonstrated the Tec-specific evolution of the Src module, where sequence motifs selected along the Tec family branch of evolution build upon the Src module to accommodate further regulation by the PHTH domain that is unique to Tec kinases (in submission).

The emerging picture of protein kinase diversification reflects a modular evolution where common core regulatory features are further evolved to accommodate fine-tuned control by additional regulatory elements. Nevertheless, a systematic analysis to define common regulatory features shared among multiple tyrosine kinase families is still lacking. While previous

classifications have identified some higher order groupings of tyrosine kinase families, such as the SrcA, SrcB, Frk, and SRM subfamilies contained in the Src family (*5, 7, 19*), these were primarily based on clustering of sequences in phylogenetic trees, and other possible hierarchical organizations have not yet been systematically explored. Here, we determine a novel, pattern-based hierarchical classification of the tyrosine kinases based on the modular selection of sequence motifs in the kinase domain. We hypothesize that ancient regulatory modules are conserved in sequence across multiple tyrosine kinase families, which have further diverged to accomodate unique regulatory functions adapted to the specific biological niches of distinct tyrosine kinases. Our resulting classification identifies three major subgroups consisting of numerous tyrosine kinase families and that share distinct regulatory cores. We demonstrate how the new classification provides a foundation for more meaningful comparative analyses between tyrosine kinase families to determine how common regulatory modules have diverged. Moreover, the identification of sequence motifs associated with tyrosine kinase evolution provides novel hypotheses that can be further examined to gain novel insights into unknown or poorly characterized tyrosine kinases and their functions. Further systematic characterization of tyrosine kinase regulation will have important implications in understanding the effects of the many disease-associated mutations in tyrosine kinases and in examining the efficacy of tyrosine kinase inhibitors.

2.2 Results and Discussion

2.2.1 A novel hierarchical, pattern-based classification of tyrosine kinases

To determine an optimal hierarchical classification of the tyrosine kinases, we first explored all possible classification structures of diverse tyrosine kinase sequences *a priori* using

an optimal multiple-category Bayesian Partitioning with Pattern Selection (omcBPPS) algorithm, which classifies sequences based on patterns of amino acid conservation and variation in a large multiple sequence alignment. We used multiple sequence sets from the NCBI nr and UniProt (*25*) databases and different clustering parameters to sample all possible classification structures of the sequence sets. The 30 major tyrosine kinase families were consistently identified in all runs, however, multiple statistically significant classification structures varied with respect to the subgroupings of various tyrosine kinase families. In order to compare the various omcBPPS subgroupings amongst each other and with the standard Manning classification, we then used a larger sequence set (33,769 sequences) and multiple-category Bayesian Partitioning with Pattern Selection (mcBPPS), which classifies sequences based on a pre-specified classification structure with seed sequences. By comparing the log-probability ratio (LPR) scores for each subgroup structure, we then iteratively optimized the classification structure by accepting the subgroup structures that maximized the LPR scores for the subgroup (see methods).

The final optimized classification improved the total LPR score substantially compared to the Manning classification of tyrosine kinases (16,065.4 nats) and designates three subgroups of tyrosine kinase families: the Src module subgroup (SrcM), the InsR-like subgroup (InsRL), and the FGFR/VEGFR/PDGFR receptor (FPVR) subgroup (Figure 2.1). The SrcM subgroup differs significantly from the Manning classification in that it clusters the SrcA, SrcB, and Frk subfamilies of the Src family within the same subgroup as the Tec and Abl families, notably without the SRM and sponge-specific Src (Src-Aque1) families and without a separate subgrouping of the Src kinases from Tec and Abl. The FPVR subgroup includes seven distinct receptor tyrosine kinase families, where a PDGFR/VEGFR receptor (PVR) subgroup subclassifies the VEGFR, PDGFR, Kit, CSF1R, and Flt3 families as a distinct subgroup from the

FGFR and Ret families. Notably, the new classification separates the Manning PDGFR family into four separate families (PDGFR, Kit, CSF1R, and Flt3) suggesting that statistically significant sequence patterns define each of these families as to warrant their designation as distinct families. The InsRL subgroup is the largest subgroup, comprising roughly 20% of tyrosine kinases, and encompasses nine receptor tyrosine kinase families, including the archetypal Insulin receptor kinase family as well as other poorly studied tyrosine kinases such as the the CCK4 family of pseudokinases and the Lmr family. In total, nearly half (45.3%) of all tyrosine kinase sequences could be classified into one of the three subgroups.

In addition, we also detect a novel family of tyrosine kinases only detected in nematode species (Figure 2.1). These kinases share sequence homology to the Fer family of tyrosine kinases, however, they form a distinct family unique from canonical Fer kinases found in other metazoan species. While vertebrate tyrosine kinomes typically have two kinase members within the Fer family (Fer and Fes) and invertebrate tyrosine kinomes typically have one member, several nematode species have more than ten members within the nematode-Fer family, indicating that the Fer family has undergone a large expansion in nematodes. For example, *Caenorhabditis remanei* has 39 unique sequences in the nematode-Fer family, and *Pristionchus pacificus* has 11. We note that other organism-specific tyrosine kinase families, such as the unique receptor tyrosine kinase families in choanoflagellates (*21, 26*), were not detected due to the limited number of sequences available.

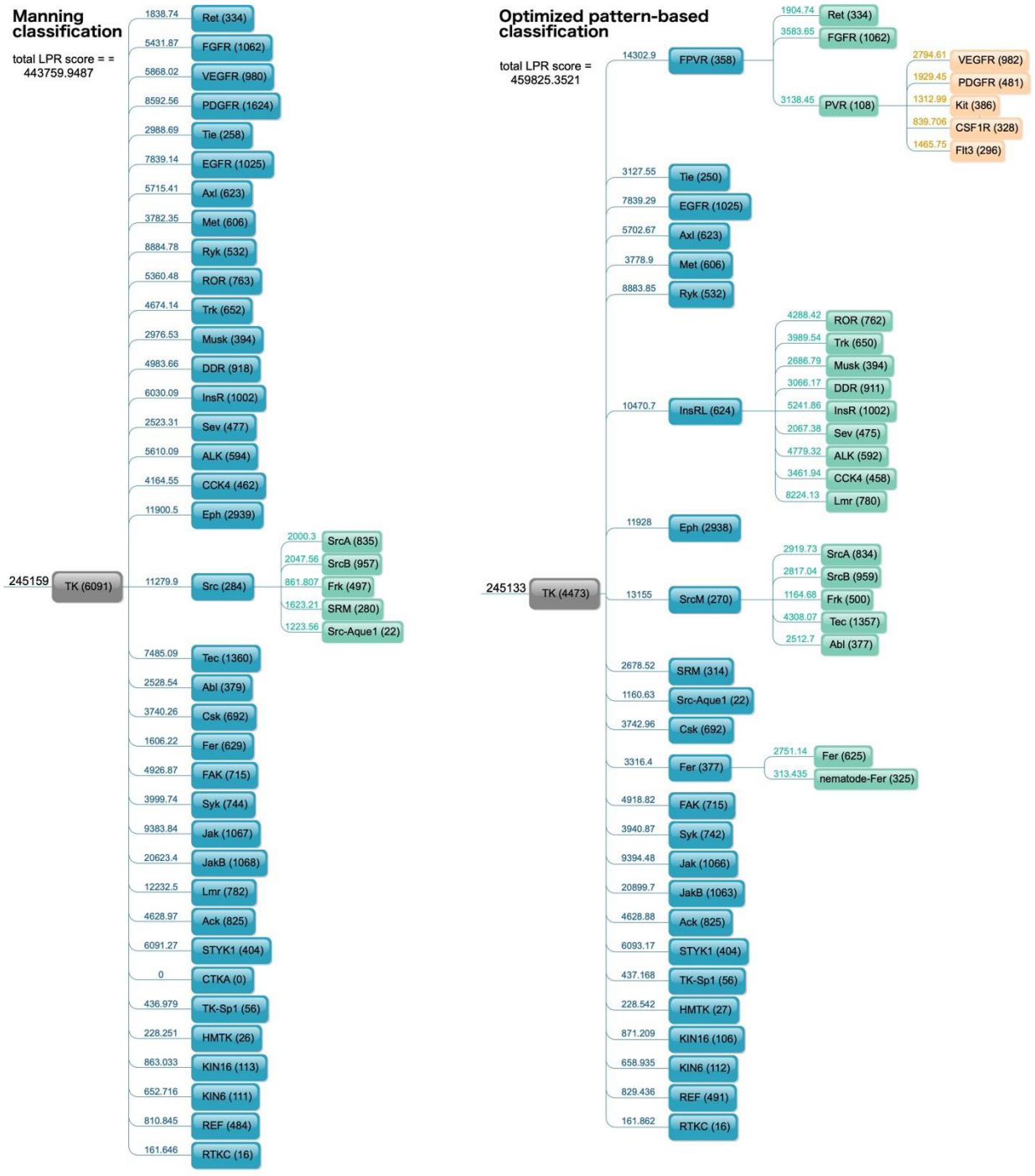

Figure 2.1: A novel hierarchical, pattern-based classification of tyrosine kinase sequences. The previously determined Manning classification (left) and the new hierarchical, pattern-based classification (right) are depicted as cladeograms. Each node represents a tyrosine kinase subgroup or family, with the number in parentheses indicating the number of sequences classified into each subgroup/family. Labels above each branch indicate log probability ratio (LPR) scores for each subtree in the classification (calculated by the mcBPPS algorithm). The total LPR scores are indicated above each cladeogram.

2.2.2 Common regulatory cores define tyrosine kinase subgroups

From our hierarchical pattern-based classification of the tyrosine kinome, we next investigated the defining sequence and structural characteristics that distinguish each subgroup from other tyrosine kinases. By examining representative crystal structures from each subgroup, we observe that subgroup-specific motifs locate to important regulatory regions on the kinase domain and encode unique regulatory modules. Below, we define the regulatory modules associated with the SrcM, InsRL, and FPVR subgroups, and we also examine how these regulatory modules have diverged in unique ways across various tyrosine kinase families.

2.2.2.1 Regulation of Src Module kinases through a common SH3-SH2-KD framework

Sequence motifs defining the SrcM kinases locate to two primary regions of the kinase domain. In autoinhibited structures of SrcM kinases, one set of SrcM-specific residues mediate key interactions with the linker region between the kinase and SH2 domains (SH2 linker) and with the SH2 domain, and another set of SrcM-specific residues mediate key interactions associated with unique conformations of the ATP binding site and the substrate binding site (Figure 2.2).

The regulatory roles of interactions by the SH2 domain and SH2 linker with the kinase domain been extensively examined in various members of the SrcA, SrcB, Tec, and Abl families (*18, 27-31*). The tethering of a "WEI" motif in the SH2 linker by SrcM-specific residues in the N-lobe forms a network on top of the R-spine termed the "hydrophobic stack" (Figure 2.2, Table 2.1), allosterically linking the SH2 linker to the active site. While the interactions between the SH2 linker and the SrcM-specific residues, W286 and Y326, are well characterized (*18, 27-29, 32*), we note for the first time the tethering of an arginine in the SH2 linker (R264) via unique cation-pi interactions with a SrcM-specific tyrosine in the β5 strand. Important interactions also

occur between the SH2 domain and the C-lobe in across autoinhibited structures of SrcM kinases (*18, 30, 31*), including a salt bridge (Figure 2.2, Table 2.1). Like the SH2 linker interactions with the N-lobe, SH2 domain interactions with the C-lobe also appear to be linked to the R-spine, in this case, via a SrcM-specific residue leucine in the αC-β4 loop (L322). Though structural studies on SrcM constructs with SH2/SH3 domains and/or the SH2 linker, as well as subsequent mutagenesis studies, have been able to identify these networks as important regulatory regions, our sequence studies identify the SrcM-specific β5 strand tyrosine (Y335) and αC-β4 loop leucine (L322) and the SH2 linker arginine (R264) as integral components of these networks for the first time. Therefore, the contribution of these residues to SH2/SH3-mediated regulatory functions in SrcM kinases warrants further investigation.

Another set of SrcM-specific sequence motifs locate to the front of the kinase domain and contribute to unique conformations of the ATP- and substrate-binding sites in inactive structures of SrcM kinases (Figure 2.2). The unique conformation of these sites is largely due to a specific, inactive activation loop conformation, which was identified as one of the defining characteristics of Src module kinases (*18*) and is characterized by a 3^{10} helix in the N-terminal portion of the activation loop, which abuts the αC helix and pushes it into an outward conformation typically unfavorable for nucleotide binding. The most distinguishing residue of the SrcM subgroup is a methionine in the β3-αC loop (M302), which bridges the coiled activation loop and the αC helix to form a tightly packed hydrophobic network adjacent to the ATP-binding site. The recent finding that SRC and the Tec family kinase, ITK, readily and preferentially bind ADP in their inactive states, while preferring ATP in their active states, raises the possibility that ADP binding in the inactive state plays an important regulatory role in SrcM kinases (*33*). As such, the conservation of SrcM-specific residues in this region (G300, M302, S303, I334, I336) (Figure

2.2, Table 2.1) may facilitate ADP binding in addition to stabilizing the inactive conformation of the activation loop and αC helix. Furthermore, mutational studies on the SH2 linker/R-spine hydrophobic stack determined that the switching behavior of SRC and ITK between ADP- and ATP- binding preferences is mediated by the assembly or disassembly of the hydrophobic stack, respectively (*33*), illustrating how the hydrophobic stack network allosterically modulates the ATP-binding site. The substrate-binding site is also uniquely formed in inactive SrcM structures. SrcM-specific residues in the activation loop contribute to a unique conformation of the activation loop that allows the orientation of the activation loop phosphorylation site as a pseudosubstrate in the active site. Most notably, a highly distinguishing phenylalanine (F424) in the activation loop forms unique pi stacking interactions with the activation loop tyrosine (Y416). Other SrcM-specific residues in the activation loop such as G421 appear to be important for maintaining the conformation of the activation loop such that this pi-pi interaction can occur.

The sequence motifs defining SrcM kinases point to shared SrcM regulatory module that can be defined by a unique allosteric network linking the N-terminal SH domain components, the R-spine, and the ATP- and substrate-binding sites. Through such a common foundational module, different molecular events can signal specific enzymatic outputs. For example, binding of specific peptides to the SH3 domain and subsequent disassociation of the kinase domain interactions with the SH machinery can signal fine-tuned adjustments to nucleotide binding and the arrangement of catalytic residues, and removal of the pseudosubstrate tyrosine from the active site and its autophosphorylation not only primes the C-lobe for substrate binding but also allows the αC helix to rotate inward to fully assemble the R-spine and facilitate ATP binding. Nevertheless, more specificity in these regulatory mechanisms are likely to be established through family-specific variations in the SrcM regulatory module. Indeed, despite the

conservation of SrcM-specific residues across most SrcM members, some important differences across different SrcM families exist. For example, while the Abl family conserves the most highly distinguishing SrcM-specific residues such as the β3-αC loop methionine and the activation loop phenylalanine, Abl sequences lack many other SrcM-specific residues, and the activation loop conformations observed in inactive Abl structures are notably divergent from other inactive SrcM structures (Figure 2.2, Table 2.1). To further explore family-specific divergence within the SrcM subgroup, we next identified and examined sequence motifs defining each family compared to other members of the subgroup. Family-specific motifs across the SrcM subgroup illustrate how the Src module has been adapted in various ways to further fine-tune regulation.

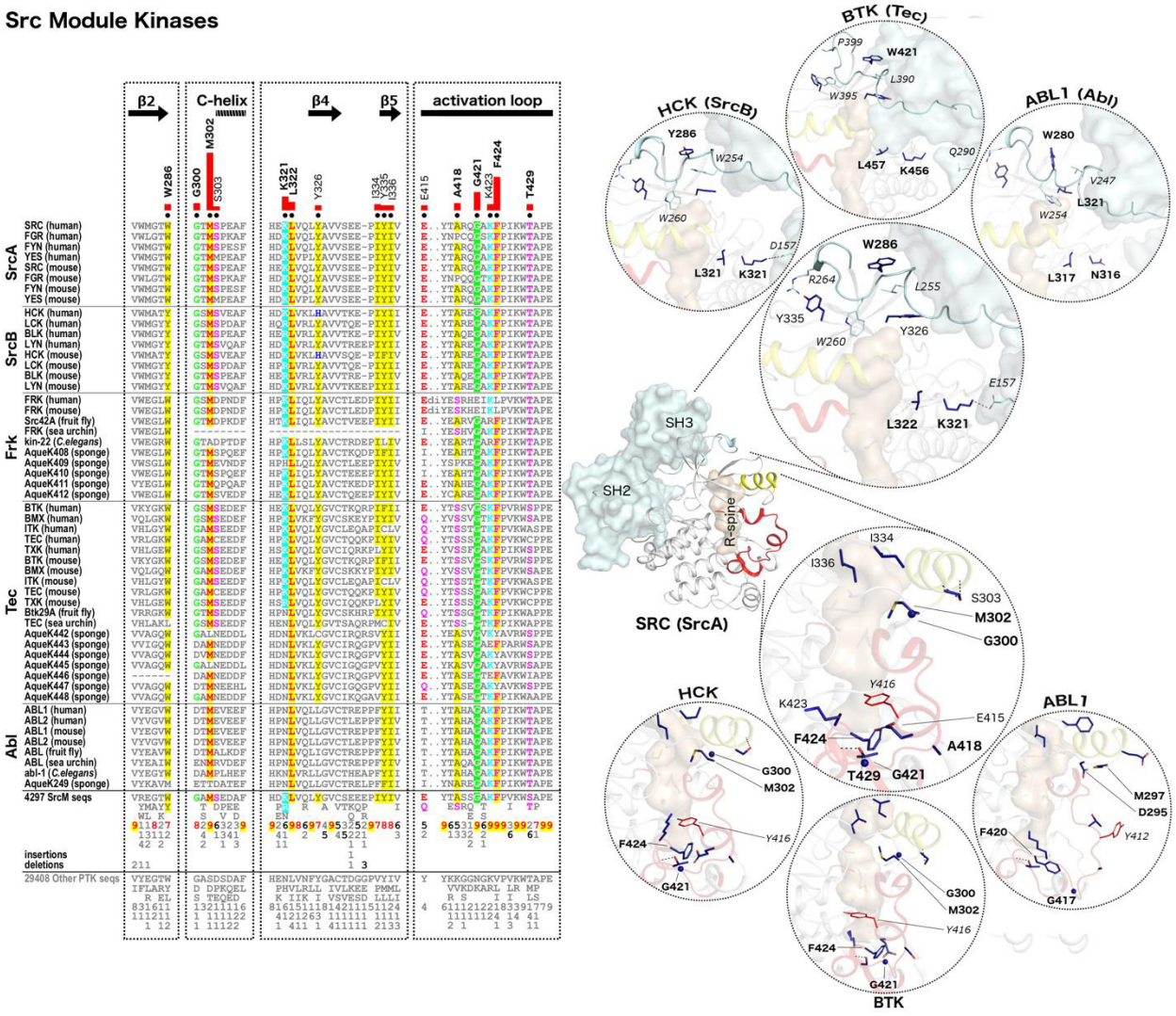

Figure 2.2: Sequence features defining Src module tyrosine kinases. In the alignment, columns are highlighted where amino acids are highly conserved in SrcM sequences but non-conserved or biochemically dissimilar in other tyrosine kinase sequences. Histograms quantify the degree of divergence between SrcM and other tyrosine kinase sequences. Column-wise amino acid and insertion/deletion frequencies are indicated in integer tenths where a "5" indicates an occurrence of 50 – 60% in the given (weighted) sequence set. Kinase secondary structures are annotated above the alignment. Sequence numbering above the alignment corresponds to the chicken SRC sequence. (b) In the structure, the SH3 and SH2 domains are shown in cyan, the regulatory spine in tan, the αC helix in yellow, and the activation loop in red. SrcM-specific sequence motifs are shown as dark blue sticks. The activation loop tyrosine phosphorylation site is shown in red lines. Glycines are depicted as spheres. Hydrogen bonding interactions are depicted by black dashes. The most highly distinguishing sequence motifs are labeled in bold font.

	SrcM-specific residue	Functional annotation
SH2 linker tether	β2 strand Trp/Tyr (W286)	tethers a conserved hydrophobic residue in the SH2 linker (L255)
	β4 strand Tyr (Y326)	tethers a conserved hydrophobic residue in the SH2 linker (L255); lost in Abl kinases
	β5 strand Tyr (Y335)	forms cation-pi interactions a conserved arginine in the SH2 linker (R264)
SH2 interface	αC-β4 loop Lys/Arg (K321)	forms a salt bridge to a conserved aspartate/glutamate in the SH2 domain; lost In Abl and sponge Tec kinases
	αC-β4 loop Leu (L322)	interacts with and undergoes coordinated side-chain flips with the DFG-phenylalanine in the R-spine
nucleotide binding pocket	β3-αC loop Gly (G300)	imparts flexibility to the β3-αC loop; lost In Abl kinases
	β3-αC loop Met (M302)	forms a hydrophobic network between the 3^{10} helix in the activation loop, β3-αC loop, and αC helix
	β3-αC loop Ser (G303)	caps the N-terminus of the αC helix; lost In Abl, Frk, and sponge Tec kinases
	β4-β5 loop Ile (I334)	forms a hydrophobic network between the activation loop 3^{10} helix, β3-C loop, and αC helix; lost In Abl and sponge Tec kinases
	β5 strand Ile (I336)	forms a hydrophobic network between the 3^{10} helix in the activation loop, β3-αC loop, andαC helix
substrate binding pocket	activation loop Glu/Gln (E415)	forms hydrogen bonds within the activation loop; lost In Abl and some Frk and Tec kinases
	activation loop Ala/Ser (A418)	imparts flexibility to the activation loop
	activation loop Gly (G421)	imparts flexibility to the activation loop; lost in most vertebrate Frk kinases (retained in fish)
	activation loop Lys (K423)	forms hydrogen bonds within the activation loop
	activation loop Phe (F424)	forms edge-to-face pi stacking interactions with the activation loop tyrosine (Y416); lost in most vertebrate Frk kinases (retained in fish)
	activation loop Thr/Ser (T429)	forms hydrogen bonds within the activation loop; lost in some Tec kinases

Table 2.1: Functional annotation of SrcM-specific sequence motifs

Motifs unique to the SrcA family facilitate a novel configuration of the SH domain components and of the C-terminal tail, as observed in one active structure of SRC (*34*). In particular, a distinct SH2 linker conformation in this active structure reconfigures the critical hydrophobic stack residue, W260, and also reconfigures the SH3 domain onto the αC helix, rather than on the back of the kinase domain (Figure 2.3, panel A). These rearrangements of both the SH2 linker and the SH3 domain in this manner is believed to further stabilize the R-spine and αC helix in an active conformation (*34*) and suggests a unique SrcA-specific activating mechanism of the Src module. However, these hypotheses have not yet been tested experimentally. In addition, the active structure shows a number of highly distinguishing SrcA motifs that tether the (unphosphorylated) C-terminal tail between the αE, αF, αH, and αI helices (Figure 2.3, panel A). Interestingly, the same site is used in Abl to dock a myristate moiety in an autoinhibitory manner (*31*), however, whether the site in SrcA can also dock myristate, as well as the functional consequence of this C-tail tethering interaction is still unknown. One possibility is that tethering the C-tail in the active state prevents phosphorylation of the C-tail by Csk (*35*), thereby averting inactivation. Our identification of striking SrcA-specific motifs mediating unique SH domain and C-tail interactions further support the functional significance of these regions and pinpoint specific residues that are likely to be informative in future mutagenesis studies.

While SrcA family motifs build upon the Src module by modulating specific conformations in the active state, the unique sequence motifs of the SrcB family make rather subtle adjustments to the inactive state of the SrcM module. In particular, the SH2 linker leucine (L255 in SRC) is selectively substituted to a bulkier tryptophan (W254, human HCK numbering) in SrcB kinases, and this amino acid change is accompanied by a concomitant change in the β2-

strand tethering residue from tryptophan to a smaller SrcB-specific tyrosine (Y286)(Figure 2.3, panel B). The positioning of the SH2 linker and the orientation of the SH2 linker and hydrophobic stack residues are nearly identical between SRC and HCK (Figure 2.2), however, the SrcB-specific substitution of the SH2 linker leucine to tryptophan introduces unique pi-pi interaction forces within the network. Another unique feature in the SH2 linker network is a SrcB-specific asparagine in the αC helix (N312) that forms hydrogen bonds with the SH2 linker tryptophan, W260, which may enable a unique coupling between W260 and the R-spine in a manner that is divergent from other SrcM kinases. In addition, a SrcB-specific glutamine (Q318) and lysine (K324) in the αC-β4 loop do not directly contact the SH2 linker but may indirectly contribute to SH2 linker tethering through other interactions, such as the C-terminal capping of the αC helix by the glutamine residue Q318. These unique autoinhibitory SH2 linker interactions suggest a SrcB-specific mode of stabilizing the inactivating the Src module.

Sequence motifs unique to the Tec family have instead built upon the activation loop portion of the inactive Src module, where Tec-specific residues form unique surface exposed patches in regions spanning the activation loop, β3-αC loop, and αC helix (Figure 2.3, panel C). As recently determined through solution studies on BTK, these residues comprise an interface for PHTH domain binding shared across Tec kinases (*35*). The utilization of the SrcM-specific inactive conformation around the activation loop as a binding interface suggests that PHTH domain binding may further stabilize other canonical inactive Src module conformations, such as the SH3/SH2 domain assembly on the backside of the kinase domain. Indeed, other SrcM families have evolved their own ways to stabilize this assembly through C-tail phosphorylation or binding of myristoyl groups in the case of SrcA/SrcB and Abl, respectively (*18, 31*).

The Abl family is the most divergent family of the SrcM subgroup and lacks several key SrcM-specific motifs including the β3-αC loop glycine and the αC-β4 loop lysine that forms part of the SH3 interface, (Figure 2.2, Table 2.1). Comparing available Abl structures with those of other SrcM kinases demonstrates why some SrcM-specific residues may be selectively missing in the Abl family and substituted for by Abl-specific residues. For example, the most striking difference between the Abl family from other SrcM members comes from the range of observed activation loop conformations in Abl structures, which have only been observed to adopt the SRC-like activation loop conformation when bound to an ATP-peptide conjugate (*36*). In a structure of ABL1 bound to an ATP-peptide conjugate, the N-terminal 3^{10} helix in the activation loop packs against the outwardly displaced αC helix via hydrophobic interactions with the β3-αC methionine (M297, human ABL1 numbering), consistent with other inactive SrcM structures (Figure 2.2). One study of ABL1 kinase showed that mutation of the SrcM-specific β3-αC loop methionine to glycine had an activating effect whereas mutation to leucine decreased activity (*37*), suggesting that this residue and its specific biochemical properties are in fact important for the inactive-to-active transition, as predicted for all SrcM kinases (Figure 2.2). However, the exact mechanistic role of the methionine in the inactive-to-active transition still remains unknown for all SrcM members, including Abl. Importantly, despite the structural similarities in the N-terminal portion of the activation loop between Abl and other SrcM members, the activation loop tyrosine (Y412) is not accommodated as a pseudosubstrate in the SRC-like Abl structure, which is a canonical feature of inactive SrcM structures. The loss of the pseudosubstrate interaction in this structure may be associated with the Abl-specific substitution of the SrcM-specific β3-αC glycine to aspartate (D295), which hydrogen bonds the activation loop backbone such that the second helical turn in the middle of the activation loop typical of

inactive SrcM structures cannot form (Figure 2.3, panel D). Similarly, the Abl-specific threonine (T411) that replaces the SrcM-specific glutamate in the activation loop also forms hydrogen bonds to the backbone of the activation loop. Despite the lack of several significant SrcM-specific sequence motifs in Abl kinases and the lack of Abl crystal structures with the canonical SrcM inactive activation loop conformation, we cannot rule out the possibility that Abl may still adopt this canonical conformation. Indeed, we expect that binding of the ATP-peptide conjugate, though it promotes a SRC-like inactive conformation in the N-terminal portion of the activation loop, it may also prevent adoption of the SRC-like conformation in the C-terminal portion because the ligand blocks the substrate binding site, and thus prevents the arrangement of the activation loop tyrosine (Y412) into a pseudosubstrate position. Therefore further investigation, for example through molecular dynamics simulations or mutagenesis experiments, is needed to determine if Abl kinases can adopt the exact canonical conformation of the inactive SrcM activation loop and to characterize the exact functional roles of Abl-specific features of this region.

Aside from distinct features in the activation loop, the Abl family also conserves other distinguishing sequence motifs that contribute unique functions to the Src module. For example, like SrcB kinases, SH2 linker interactions with the kinase domain are slightly modified in Abl. The SH2 linker leucine (L255 in SRC) is selectively changed in the Abl family to valine (V247), which is accompanied by a substitution in the β4 strand tethering residue from a tyrosine to an Abl-specific leucine (L321) and the presence of an additional tethering residue from the β3 strand, Y283 (Figure 2.3, panel D). The subtle changes in this network may contribute to the different orientation of the SH2 linker tryptophan (W254) observed in Abl compared to that seen in SRC and HCK (Figure 2.2). In addition, Abl kinases lack the canonical lysine residue that

forms a salt bridge to the SH3 domain. Instead, Abl kinases have evolved a unique pair of tyrosines between the αE helix and the SH3 domain that form face-to-face pi-stacking interactions unique to the Abl family (Figure 2.3, panel D). Despite differences in SH2 linker- and SH3 domain-tethering residues, the conservation of SrcM-specific features that connect the SH domain components to the R-spine such as the αC-β4 leucine (L321) and the SH2 linker hydrophobic stack demonstrates that a similar, albeit distinct, allosteric communication is maintained in Abl kinases as in other SrcM families. Furthermore, as mentioned previously, the ways in which the autoinhibitory SH3/SH2 domain assembly is stabilized between various SrcM families are distinct. As such, the conservation of Abl-specific residues (A452, T364) that dock a myristoyl group in the C-lobe (*31*) demonstrates how family-specific features contribute to distinct mechanisms to stabilize the inactive Src module (Figure 2.3, panel D).

In addition to Abl-specific residues that contribute to the unique autoinhibitory interactions, the Abl family also conserves unique features to regulate the active state of the Src module. In particular, the Abl family has evolved a separate conserved interface on the top of the N-lobe to dock the SH2 domain in an activating manner (*28*) (Figure 2.3, panel D). While the activating SH2 domain interaction is fairly well understood, adjacent to this interface is another cluster of highly distinguishing residues that undergo large structural rearrangements between inactive and active structures and whose relevance to function has not yet been scrutinized. This network includes a tyrosine phosphorylation site in the P-loop (Y272), which inhibits kinase activity and transforming activity when phosphorylated (*38*), however, the mechanism underlying phospho-Y272-dependent inhibition is unknown. Interestingly, the striking structural transitions of the glycine rich loop allows Y272 to interact with various other important residues such as the catalytic aspartate (D382) and DFG-phenylalanine (F401) or Abl-specific residues

G269, N341, and M407 in different states (Figure 2.3, panel D), suggesting that autophosphorylation of this site may be regulated through dynamic structural transitions involving both canonical SrcM features such as the SH domains and activation loop and also the glycine rich loop.

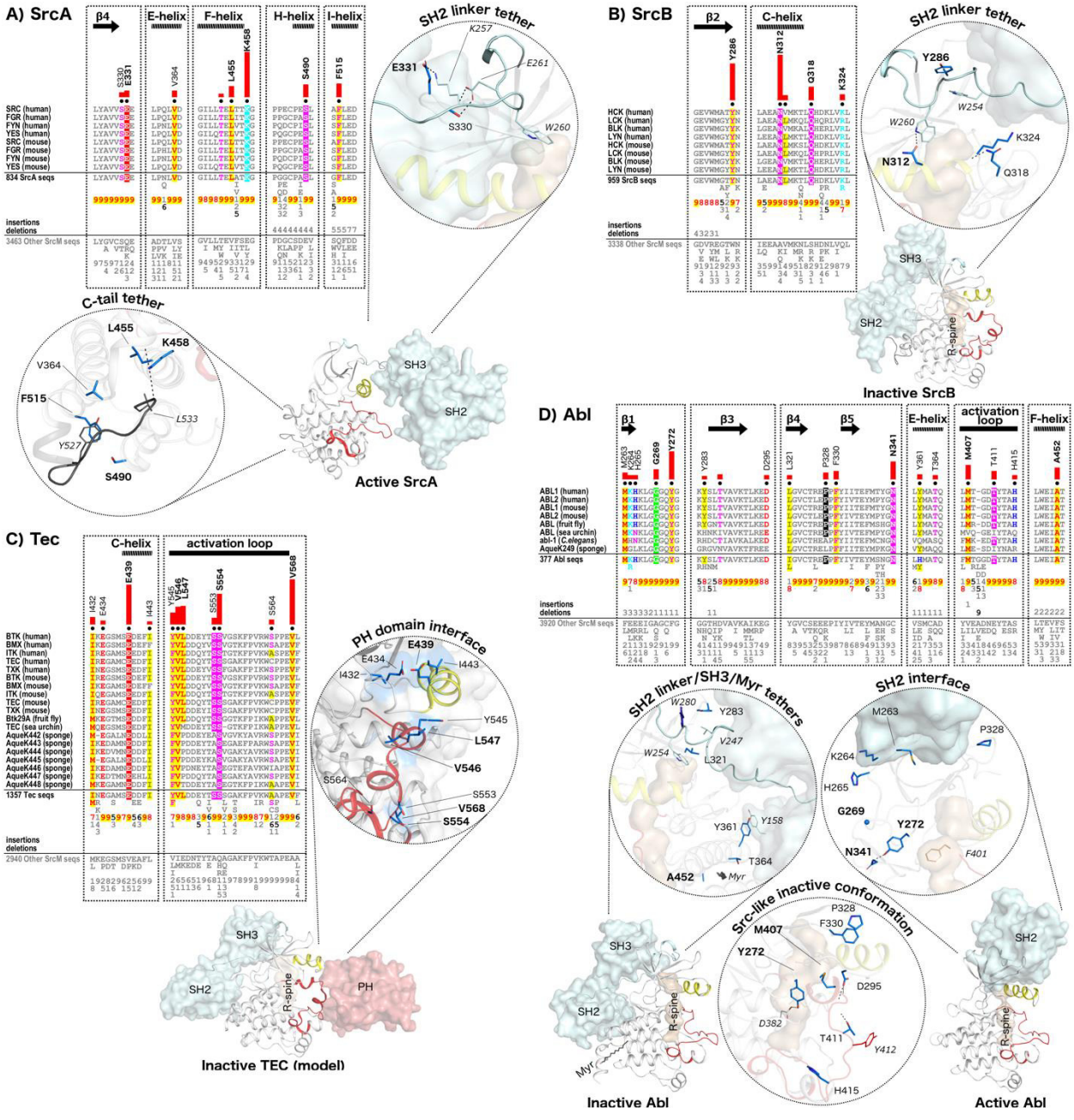

Figure 2.3: Family-specific divergence of the Src module. Sequence features defining the A) SrcA, B) SrcB, C) Tec, and D) Abl families are shown. In the structure, the SH2 and SH3

domains and SH2 linker are shown in cyan, the regulatory spine in tan, the αC helix in yellow, and the activation loop in red. Family-specific sequence motifs are shown as blue sticks. The activation loop tyrosine phosphorylation site is shown in red lines. Glycines are depicted as spheres. Hydrogen bonding interactions are depicted by black dashes. The most highly distinguishing sequence motifs are labeled in bold font.

2.2.2.2 A common inactive conformation of InsR-like kinases

One of the most distinguishing features of the InsRL kinases is the conservation of three tyrosine phosphorylation sites within the activation loops of InsRL kinases, two of which are unique to InsRL kinases. The phosphorylation of each of these residues cause specific effects on the catalytic outputs of the kinase domain (i.e. shifts in Km and/or Vmax), therefore these multiple phosphorylation events and the order in which they occur are precisely regulated (*1*). As such, sequence motifs that distinguish the InsRL subgroup from other tyrosine kinases are involved in stabilizing a unique autoinhibitory conformation of the kinase domain, which has been well characterized for the insulin receptor (INSR)(*39*), insulin-like growth factor 1 receptor (IGF1R, InsR family)(*40, 41*), tropompysin receptor kinase A (TRKA, Trk family)(*42*), ROR2 (Ror family)(*42*), and MUSK (MuSK family)(*43*). This autoinhibitory conformation, like the inactive SrcM conformation, is characterized by a coiled activation loop and one phosphorylation site in the activation loop that serves as a pseudosubstrate (Figure 2.4). The manner in which the autoinhibitory activation loop is achieved in InsRL kinases, however, are distinct from SrcM kinases. For example, unlike the activation loops of SrcM kinases, the inactive activation loops of InsRL kinases are folded inward into the active site such that the ATP-binding pocket is completely occluded, and the folded activation loop is stabilized polar interactions mediated by two pairs of InsRL-specific residues. Specifically, the N-terminal tyrosine phosphorylation site (Y1161, human IGF1R numbering) forms a hydrogen bonding network with the InsRL-specific aspartate in the αD helix (D1086), as well as with the backbone

of the αD helix. The second pair of hydrogen bonded residues, an aspartate in the activation loop (D1164) and a glutamine in the FG-loop (Q1181), additionally stabilizes the folded activation loop. The folded activation loop conformation in turn allows the second tyrosine phosphorylation site to occupy the active site as a pseudosubstrate, binding to active site residues D1135 and R1139 from the catalytic loop HRDXXXXN motif. The selection of negatively charged residues in this network may additionally be important in that these residues impart negative charges that can repel the activation loop from its inhibitory conformation once phosphorylated (*44*). We also note a phenylalanine/tyrosine (F1131) consistently found C-terminal to the αE helix across InsRL kinases, which interacts with the R-spine and produces important interactions with the DFG-phenylalanine. In this way, the InsRL kinases appear to conserve a unique mechanism to modulate the R-spine analogously to the way in which the SrcM kinases use the SH domains to modulate unique conformations of the R-spine. For example, one way in which the InsRL-specific phenylalanine may communicate signaling inputs to the R-spine is through interactions with the juxtamembrane, as seen in the autoinhibited structure of TRKA (Figure 2.4)(*45*).

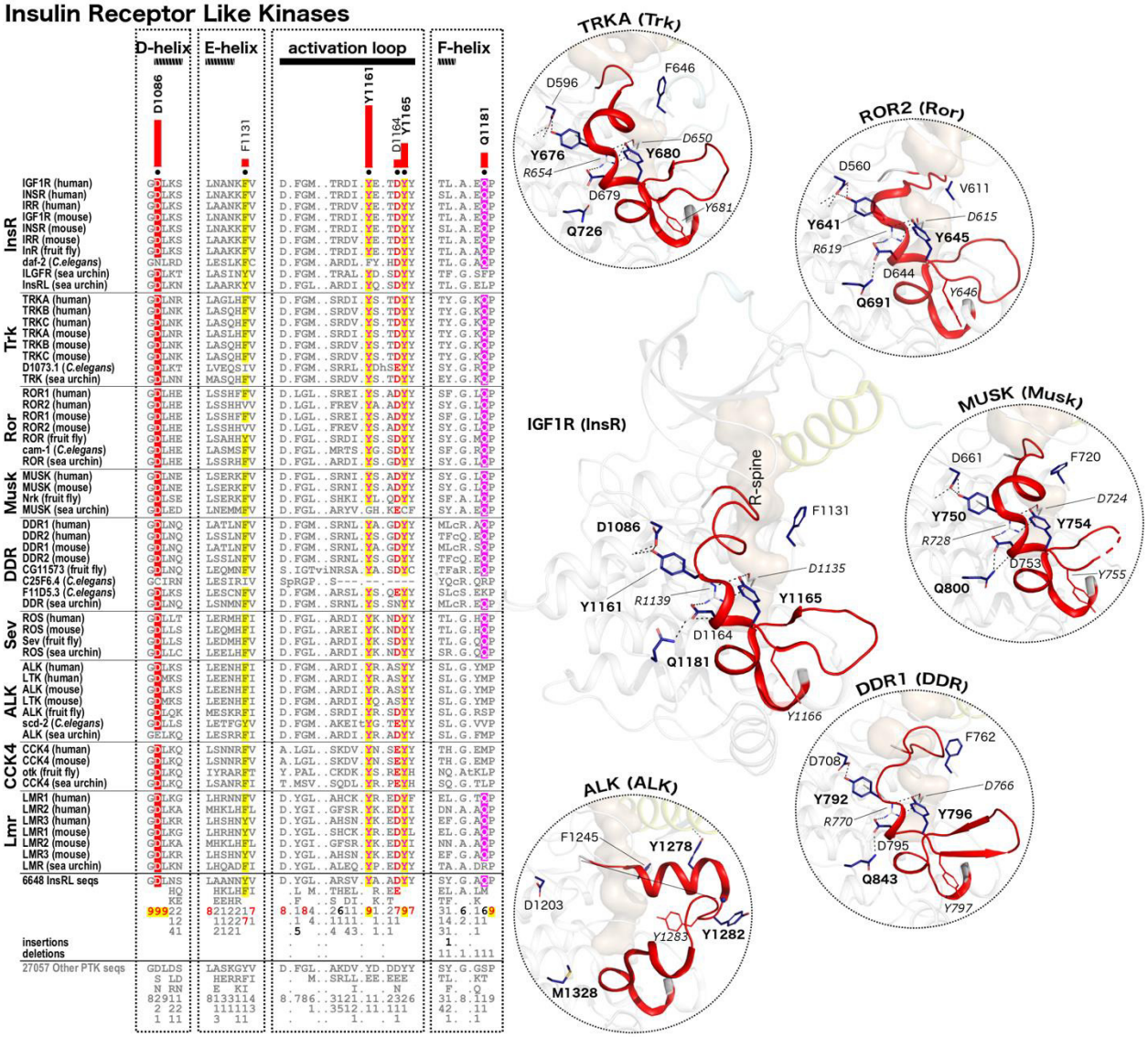

Figure 2.4: Sequence features defining Insulin receptor-like tyrosine kinases. A) In the alignment, columns are highlighted where amino acids are highly conserved in InsRL sequences but non-conserved or biochemically dissimilar in other tyrosine kinase sequences. In the structure, the juxtamembrane is shown in cyan, the regulatory spine in tan, the αC helix in yellow, and the activation loop in red. InsRL-specific sequence motifs are shown as dark blue sticks. The activation loop tyrosine phosphorylation site is shown in red lines. Glycines are depicted as spheres. Hydrogen bonding interactions are depicted by black dashes. The most highly distinguishing sequence motifs are labeled in bold font.

While sequence motifs associated with InsRL autoinhibition are well conserved across this large group of tyrosine kinases, we note some variations of InsRL motifs in certain families and organisms. For example, the FG-loop glutamine is lost in the ALK and PTK7 families and

predominantly replaced with methionine (Figure 2.4). While crystal structures of PTK7 have note yet been obtained, interestingly, inactive structures of ALK exhibit an activation loop that is highly distinct from other InsRL members, where the activation loop is extended outward and tethered to the αC helix side rather than folded inward toward the αD helix. Notably, the activation loop aspartate, which typically hydrogen bonds the FG loop glutamine, was also lost between the evolution of invertebrate and vertebrate ALK's (Figure 2.4). The correlated loss of both residues in one of hydrogen bonded pairs (Q1181, D1164 in IGF1R; M1328, S1281 in ALK) suggests a co-evolution, and thus a functional association, of residues at these positions. An examination of ALK-specific sequence motifs demonstrate how InsRL-specific features are reconfigured by ALK-specific motifs to accommodate an alternative conformation of the activation loop in ALK (Figure 2.5). In particular, the N-terminal phosphorylation site in ALK (Y1278) forms a network of pi-stacking interactions with the InsRL-specific phenylalanine near the αE helix (Y1245) and a conserved phosphorylation site in the juxtamembrane (Y1096). This network between the activation loop and juxtamembrane is further stabilized by hydrogen bonding between Y1278 and the juxtamembrane backbone, as well as by the tethering of this network to the αC helix via hydrophobic interactions with the αC helix through ALK-specific residues L1169 and M1166. In this way, ALK-specific residues not only accommodate a distinct activation loop conformation divergent from canonical InsRL kinases, but also bridge the activation loop to the juxtamembrane in a unique way, perhaps to regulate both activation loop and juxtamembrane phosphorylation concurrently. However, whether the canonical InsRL activation loop conformation is also functionally relevant in ALK kinases is still unknown, and future studies of ALK conformations using molecular dynamics and other structural studies are

needed to shed light on how the conformational landscape of ALK kinases may differ from other InsRL members.

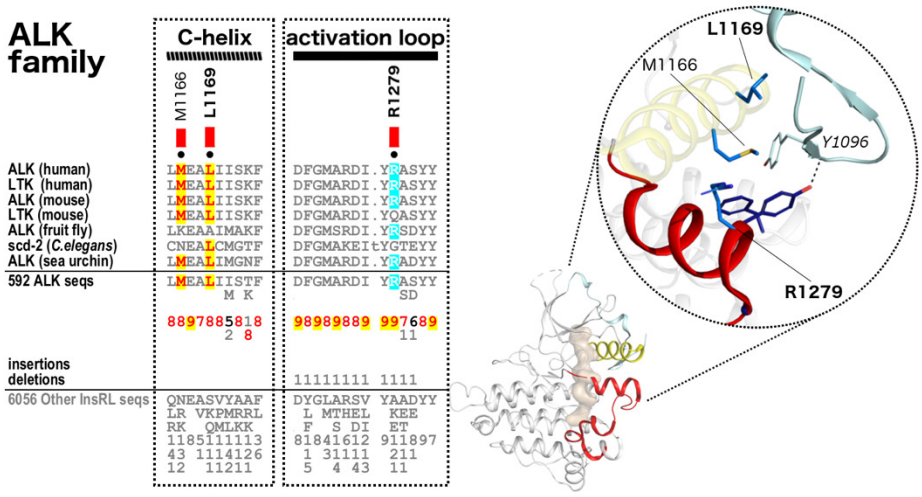

Figure 2.5: Sequence features defining the ALK family. In the alignment, columns are highlighted where amino acids are highly conserved in ALK sequences but non-conserved or biochemically dissimilar in other InsRL tyrosine kinase sequences. In the structure, the juxtamembrane is shown in cyan, the regulatory spine in tan, the αC helix in yellow, and the activation loop in red. InsRL-specific sequence motifs are shown as dark blue sticks. The juxtamembrane tyrosine phosphorylation site is shown in cyan sticks. Glycines are depicted as spheres. Hydrogen bonding interactions are depicted by black dashes. The most highly distinguishing sequence motifs are labeled in bold font.

2.2.2.3 Regulation of FPVR kinases through a common N-lobe framework

The FPVR subgroup includes seven distinct receptor tyrosine kinase families, and therefore all FPVR kinases contain a juxtamembrane segment linking the intracellular kinase domain to the transmembrane domain. The FPVR kinases are defined by highly distinguishing sequence motifs in the N-lobe of the kinase domain, which encode a common regulatory module that spanning the juxtamembrane and the hinge region. Interestingly, sequence motifs unique to the FPVR subgroup facilitate the tethering of the juxtamembrane to the N-lobe of the kinase domain that is reminiscent of SH2 linker tethering by SrcM kinases. The juxtamembrane sequence of FPVR kinases contain a highly conserved "WEF" motif, which aligns to an

analogous “WEI” motif in the SH2 linker of SrcM kinases. Just as the tryptophan of the “WEI” SH2 linker motif is allosterically linked to the active site through interactions with R1 residue of the R-spine, the tryptophan of the “WEF” motif in FPVR kinases analogously docks onto the top of the R-spine (Figure 2.6). Despite similarities, the R-spine interaction with the “WEX” motif is accompanied by FPVR-specific interactions between the kinase domain and N-terminal linker that distinguishes the FPVR module from the Src module. In particular, the phenylalanine of the “WEF” motif is docked into a hydrophobic pocket formed by FPVR-specific residues in the β2-β3 loop (A608, G610, L611, human KIT numbering). The hydrophobic tethering of the tryptophan and phenylalanine of the “WEF” motif is also accompanied by hydrogen bonding interactions between the glutamate of the “WEF” motif and an FPVR-specific threonine (T661) and the β4 strand backbone. Furthermore, a conserved hydrophobic network between the αC helix and β4 strand (L637, L641, A649) mediate further interactions between the N-lobe and more N-terminal portions of the juxtamembrane. Interestingly the Ret and FGFR families have variations at one of these positions and also exhibit juxtamembrane conformations that are notably divergent compared to the PVR families.

The most highly distinguishing residue of FPVR kinases is a well characterized asparagine in the αC-β4 loop in the hinge region of the kinase domain. The asparagine (N655) plays an essential autoinhibitory role by engaging a “molecular brake” that stabilizes an inactive state via a hydrogen bonding network between the αC-β4 loop asparagine, a glutamate in the β5-αD loop, a lysine in the β8 strand, and the backbone of the αC-β4 loop, straining the movement of the N-terminal lobe towards the C-terminal lobe (*15*). The co-conservation of the molecular brake network with the juxtamembrane tethering network and their structural proximity to each other raises the possibility that these two networks function cooperatively to allosterically

modulate the active site. Indeed, both the R-spine, which can be modulated via the tethering/untethering of the juxtamembrane, as well as the molecular brake, are known to adjust the integrity of the ATP binding pocket to regulate activity. However the Ret family is the only family of the subgroup that does not co-conserve both the juxtamembrane and molecular brake networks since it lacks the critical αC-β4 loop asparagine. Also interestingly, another tyrosine kinase family outside of the FPVR subroup, the Tie family, conserves the αC-β4 loop asparagine and the molecular brake mechanism (*15*) but lacks residues in the justamembrane tethering network such as the β2 strand alanine (A608) and the "WEF" motif. Therefore the co-conservation of the juxtamembrane and molecular brake network are hallmarks of the FPVR subgroup, however, further evolutionary divergence has allowed parts of this shared network to be degraded in some families..

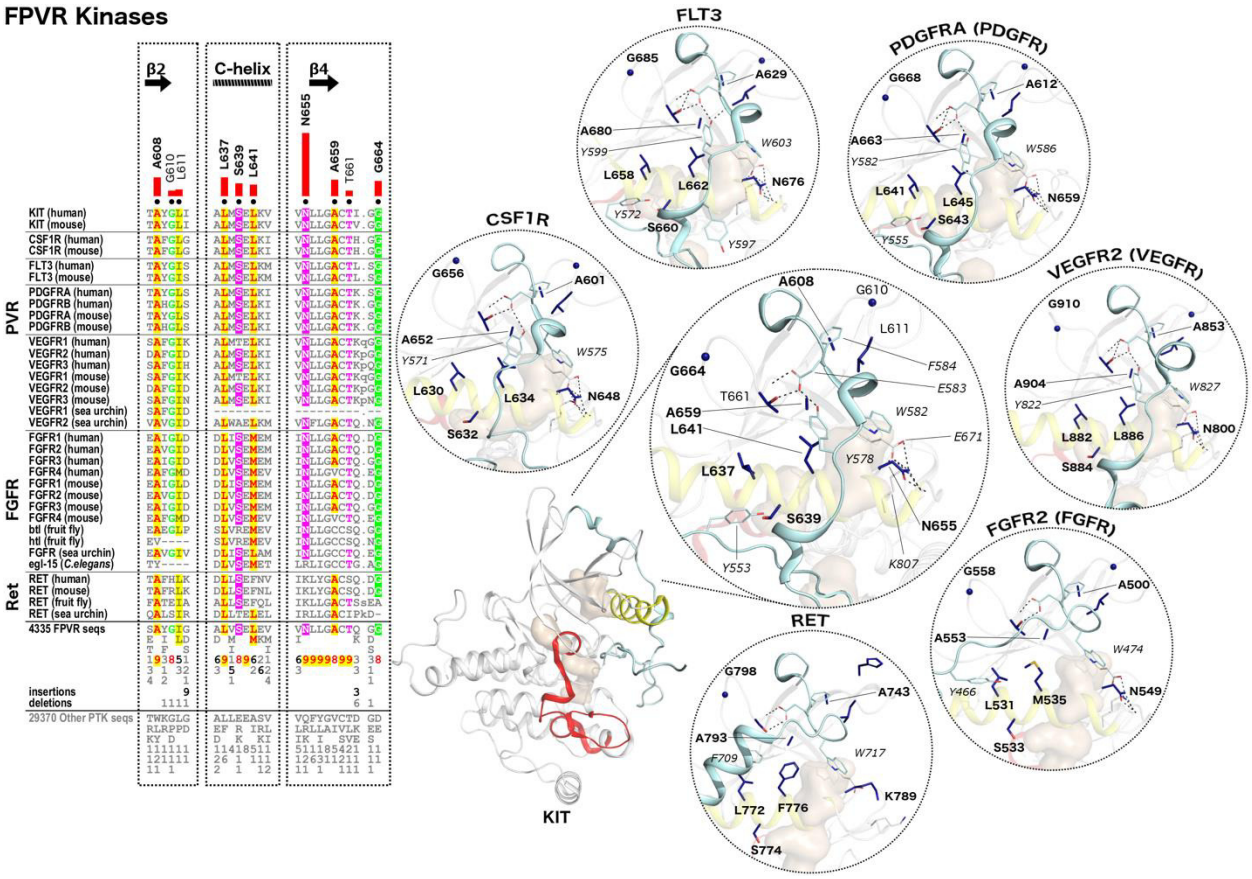

Figure 2.6: Sequence features defining FPVR tyrosine kinases. In the alignment, columns are highlighted where amino acids are highly conserved in FPVR sequences but non-conserved or biochemically dissimilar in other tyrosine kinase sequences. In the structure, the juxtamembrane is shown in cyan, the regulatory spine in tan, the αC helix in yellow, and the activation loop in red. FPVR-specific sequence motifs are shown as dark blue sticks. Glycines are depicted as spheres. Hydrogen bonding interactions are depicted by black dashes. The most highly distinguishing sequence motifs are labeled in bold font.

Aside from variations of FPVR-specific residues, PVR-, FGFR-, and Ret-specific sequence motifs also contribute to notable differences in the shared FPVR module. While FPVR-specific hydrophobic residues between the αC helix and β4 strand form hydrophobic interactions with the juxtamembrane across all FPVR families, the specific interactions between the PVR subgroup and FGFR and Ret families are further refined to accommodate variations in the juxtamembrane sequence and structure. For example, in the PVR families, the FPVR-specific juxtamembrane docking residues (L641, A659, L637, Kit numbering) in the αC helix and β4 strand are absolutely invariant, presumably because they dock a universally conserved tyrosine autophosphorylation site in the juxtamembrane that causes kinase activation when phosphorylated (Figure 2.7). In contrast, the both the Ret and FGFR families lack the C-terminal αC helix leucine, which is unique to PVR kinases. Instead, the FGFR family conserves several highly distinguishing motifs in the αC helix that facilitate a unique tethering of the juxtamembrane across the αC helix, where a hydrophobic pocket formed by FGFR-specific residues (V532 and M535, FGFR2 numbering) dock a juxtamembrane phosphorylation site conserved across FGFR kinases (Figure 2.8. Similarly, sequence motifs unique to the Ret family form unique hydrophobic interfaces on the αC helix, where a conserved phenylalanine in the juxtamembrane forms pi-pi interactions with the highly distinguishing αC helix phenylalanine of Ret kinases (F776, Ret numbering) (Figure 2.8. The PVR kinases conserve an additional PVR-specific methionine in the αC helix which distinguishes PVR kinases from Ret and FGFR

kinases and further facilitate tethering of the juxtamembrane tyrosine (Figure 2.7). In addition, PVR-specific residues A636 and E633 in the αC helix form interactions with the N-terminal portion of the juxtamembrane such that the juxtamembrane is wedged between the αC helix and inactive activation loop.

In addition to the juxtamembrane conformation, the activation loop conformation of inactive PVR kinases is also commonly shared across the PVR subgroup and mediated by PVR-specific residues. The inactive activation loop conformation of PVR kinases is highly similar to that of InsRL kinases in that it is folded completely within the active site and forms a beta hairpin at the C-terminal end of the activation loop. In addition, like the InsRL kinases, the activation loop phosphorylation site is sequestered into the active site by the catalytic aspartate, fulfilling a pseudosubstrate role. Nevertheless, a distinct set of sequence motifs unique to the PVR subgroup facilitates the adoption of this activation loop conformation in PVR kinases that are unrelated to the residues facilitating the similar activation loop conformation in InsRL kinases. In particular, the activation loop conformation in the PVR subgroup is mediated by polar interactions between PVR-specific residues D820 and N822 in the activation loops, and may also be facilitated by PVR-specific residues G827 and V824. Importantly, the conformation of the activation loop in PVR kinases appears to be facilitated also by the juxtamembrane segment, which inserts not the kinase cleft in inactive structures across the PVR subgroup. Therefore the cooperation of the juxtamembrane with the activation loop to stabilize the inactive conformation appears to be unique to the PVR module, whereas the activation loop in InsRL kinases is solely stabilized by InsRL-specific residues in the kinase domain.

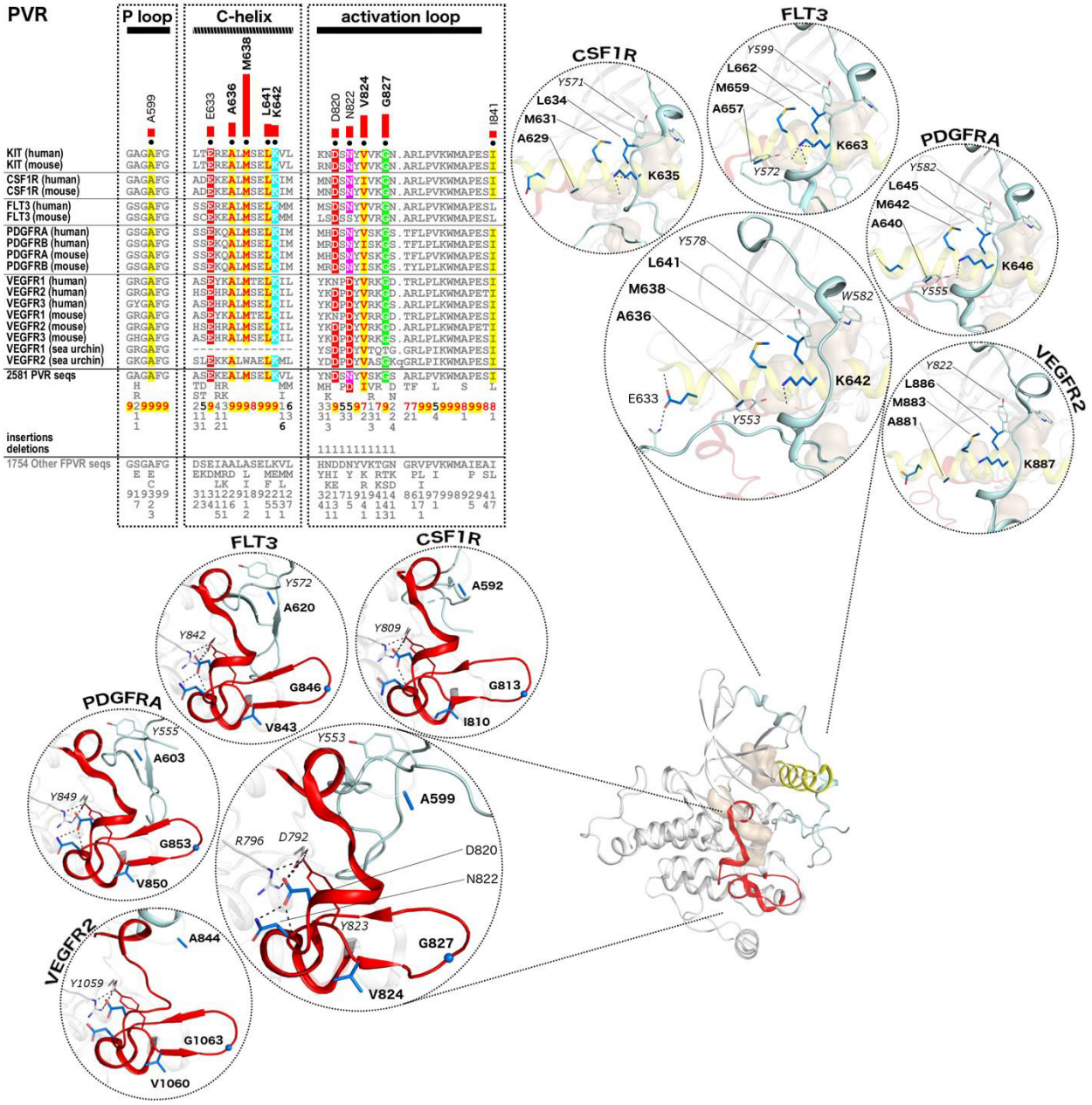

Figure 2.7: PVR-specific divergence of the FPVR module. Sequence features defining the PVR subgroup are shown. The juxtamembrane is shown in cyan, the regulatory spine in tan, the αC helix in yellow, and the activation loop in red. Subgroup- or family-specific sequence motifs are shown as blue sticks. The activation loop tyrosine phosphorylation site is shown in red lines. Glycines are depicted as spheres. Hydrogen bonding interactions are depicted by black dashes. The most highly distinguishing sequence motifs are labeled in bold font.

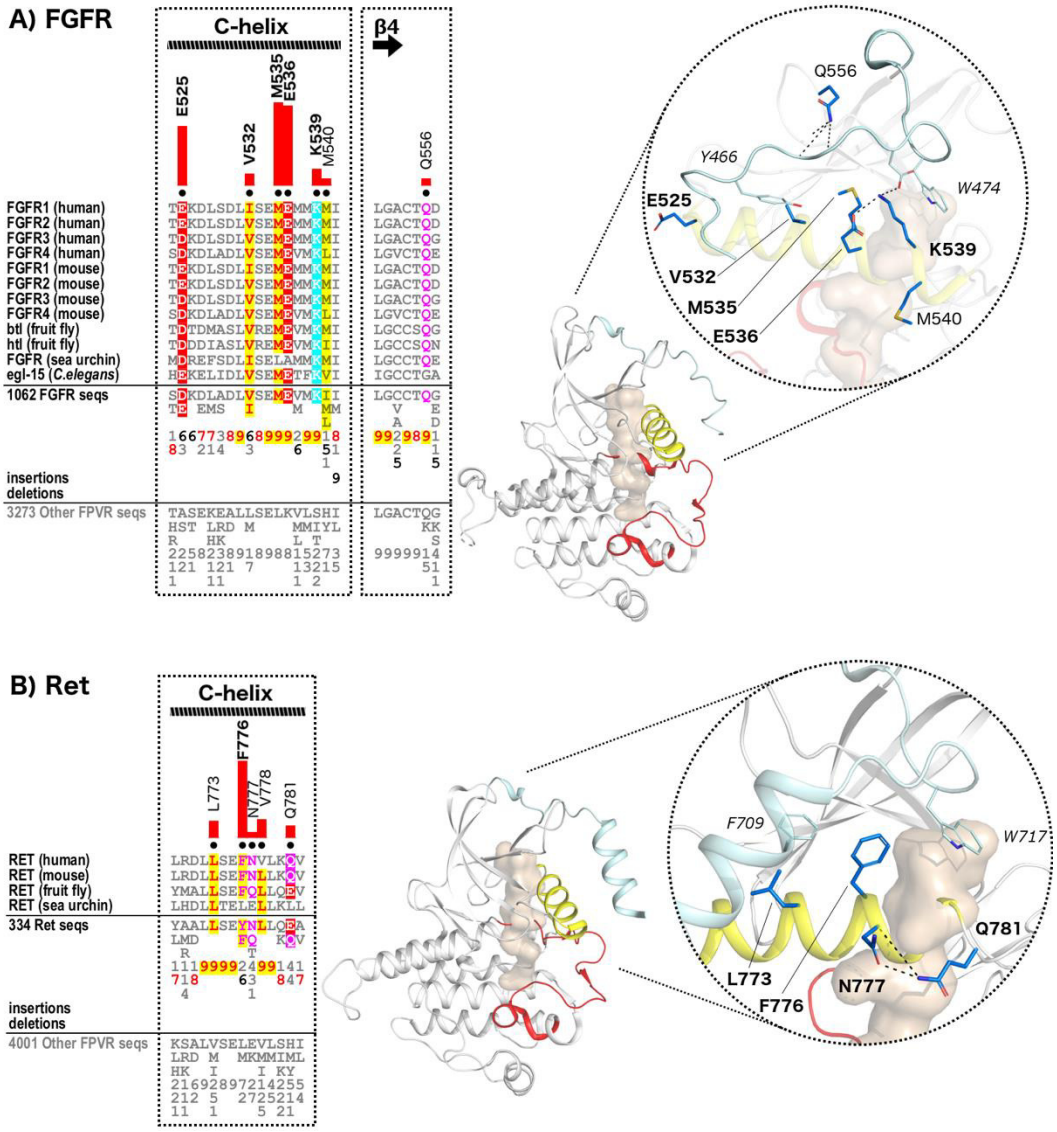

Figure 2.8: Family-specific divergence of the FPVR module in FGFR and Ret kinases. Sequence features defining the A) FGFR family and the B) Ret family are shown. The juxtamembrane is shown in cyan, the regulatory spine in tan, the αC helix in yellow, and the activation loop in red. Subgroup- or family-specific sequence motifs are shown as blue sticks. The activation loop tyrosine phosphorylation site is shown in red lines. Glycines are depicted as spheres. Hydrogen bonding interactions are depicted by black dashes. The most highly distinguishing sequence motifs are labeled in bold font.

2.2.3 Structural and functional prediction of understudied tyrosine kinases

Based on our knowledge of the conserved regulatory modules of various tyrosine kinase subgroups, we modelled the structures of four understudied tyrosine kinases and predicted specific functions encoded in their sequence (Figure 2.9).

A model of the Frk kinase based on the archetypal inactive Src structure shows that the Frk sequence can adopt the canonical Src module configuration. By mapping SrcM-specific residues onto the structure, as well as Frk-specific residues, we can determine that the residues that mediate kinase domain interactions with the SH2 linker and SH2/SH3 in SrcM kinases are preserved in Frk. In addition, no Frk-specific features locate to this region, suggesting that there is no significant Frk-specific divergence with regard to SH2 linker and SH2 domain docking, and therefore Frk is likely to share this common regulatory module with other SrcM families. Interestingly, however, a 'TxxWK' motif in the αI helix is uniquely found in Frk kinases that structurally locates proximally to the C-terminal tail of Frk. The C-tail is known to play essential roles in locking the autoinhibitory conformation of the Src module through interactions with the SH2 domain in SrcA and SrcB kinases. In addition, the observation of a SrcA-specific C-tail tether in an active Src structure compels us to reason that the 'TxxWK' motif may also regulate the Src module by sequestering the C-tail in specific kinase states.

Despite the conservation of SrcM-specific residues mediating SH2 linker and SH2 domain interactions, SrcM-specific residues in the activation loop are not well conserved in Frk and may suggest an alternative conformation of the activation loop in Frk that is divergent from other SrcM families. Notably, the SrcM-specific GxxF motif in the C-terminal portion of the activation loop that typically forms pi-interactions with the activation loop tyrosine is substituted by a ExxL sequence, which would both dramatically affect the flexibility of this region of the

activation loop as well as the stabilization of the activation loop tyrosine as a pseudosubstrate. In addition, the SrcM-specific glutamate in the activation loop is selectively substituted by an isoleucine in Frk. The change of this residue from a charged amino acid to a bulky hydrophobic residue at a surface-exposed region of the activation loop further suggests that the canonical SrcM activation loop conformation would not be favorable in Frk. Frk-specific residues near the activation loop may also contribute to a unique activation loop conformation; for example, a salt bridge between the Frk-specific K380 from the activation loop may hydrogen bond Frk-specific D270 in the αC helix to accommodate an alternative conformation of the activation loop.

Interestingly, many Frk-specific residues cluster on the backside of the kinase domain adjacent to the SH2 and SH3 interface, including a number of hydrophobic residues that we predict would be surface-exposed. Despite the proximity of the surface cluster to the SH2/SH3 interface, the high conservation of SrcM-specific SH2 and SH3 tethering residues in Frk suggest that the surface cluster is not the SH2/SH3 interface in Frk. It could perhaps represent an alternative binding interface for these domains, similarly to that seen in the secondary SH2 interface specific to the Abl family of the SrcM subgroup.

The Lmr family of kinases in the InsRL subgroup represents one of the most understudied tyrosine kinase families, and no crystal structures of these kinase domains have been solved to date. By modeling a structure of the LMR1 kinase using a prototypic inactive IGF1R structure and mapping InsRL- and Lmr family-specific residues, we highlight key Lmr-specific regulatory features, including one putative regulatory C-tail tethering site. First, the conservation of canonical InsRL-specific sequence motifs in the activation loop suggests that Lmr kinases can adopt the canonical InsRL inactive conformation. However, many Lmr-specific residues form a unique network in and around the R-spine. Most notably, the R-spine base

residue, the αF helix aspartate, which is one of the most anciently conserved amino acids across the eukaryotic and eukaryotic-like protein kinases, is selectively substituted to an asparagine. The aspartate residue typically stabilizes the active kinase conformation by forming two hydrogen bonds to the backbone atoms of the HRD motif backbone. Because the asparagine would only be able to form one of these hydrogen bonds, we hypothesize that the active conformation in Lmr kinases would likely have to be stabilized in an alternative manner. One way in which this may be achieved is through the stabilization of the αF helix asparagine and the catalytic loop backbone through Lmr-specific residues in this region, such as the αE helix histidine (H245, human LMR1 numbering) and the αF helix glutamine and lysine (Q314, K316). Another set of Lmr-specific residues occur in the C-lobe and are surface-exposed near the C-tail. The Lmr kinases are known to have extremely long C-terminal tails, however, their functions are not yet known. We propose that these residues may function to tether the C-tail.

The CCK4 (PTK7) family of receptor tyrosine kinases in the InsRL subgroup represents a pseudokinase family, which lacks catalytic activity due to the loss of key residues in the catalytic site. Despite the loss of enzymatic activity, pseudokinases often play functional roles, such as scaffolding, molecular recruitment, and the allosteric activation of other (active) kinases. However the function of CCK4 has not yet been determined, and no crystal structures of its kinase domain have been solved. In addition to the loss of canonical catalytic residues, we find that CCK4-specific residues further facilitate the inactivation of the kinase domain. For example, the presence of bulky hydrophobic residues in the active site (L828, F850) can prevent nucleotide binding. Other CCK4-residues form a network of charged residues around the αC helix and may act as a juxtamembrane tether, similarly to many other receptor tyrosine kinases. Notably, a highly distinguishing arginine pair (R864, R864) could potentially served as a

tethering site for a phosphorylated residue. Indeed, there is evidence that a serine in the juxtamembrane (S784) becomes phosphorylated and is located structurally in a manner that would allow tethering to the arginine pair.

The Sev family is also part of the InsRL subgroup, and while one active crystal structure of the ROS kinase domain exists, currently, there are not crystal structures of Sev kinases in the inactive state. Modeling ROS into a canonical inactive InsRL conformation and mapping InsRL- and Sev family-specifc residues points to an interesting cluster of hydrophobic residues in the InsRL inactive activation loop conformation of Sev. The InsRL activation loop conformation is associated with a network of two pairs of hydrogen bonded residues. While the Sev family conserves all four residues in these hydrogen bonded pairs, hydrophobic Sev-specific residues map directly adjacent to these hydrogen bonded networks. For example, a highly distinguishing leucine in the αD helix (L2035) and an alanine in the activation loop (A2106) seem to thether the InsRL-specific activation loop tyrosine which hydrogen bonds the InsRL-specific αD helix aspartate. However, whether these hydrophobic residues facilitate or impede the canonical inactive InsRL-specific conformation from being formed is still unknown and warrants further investigation. In addition, we note a lysine pair in the αC helix that may function as a juxtamembrane tether.

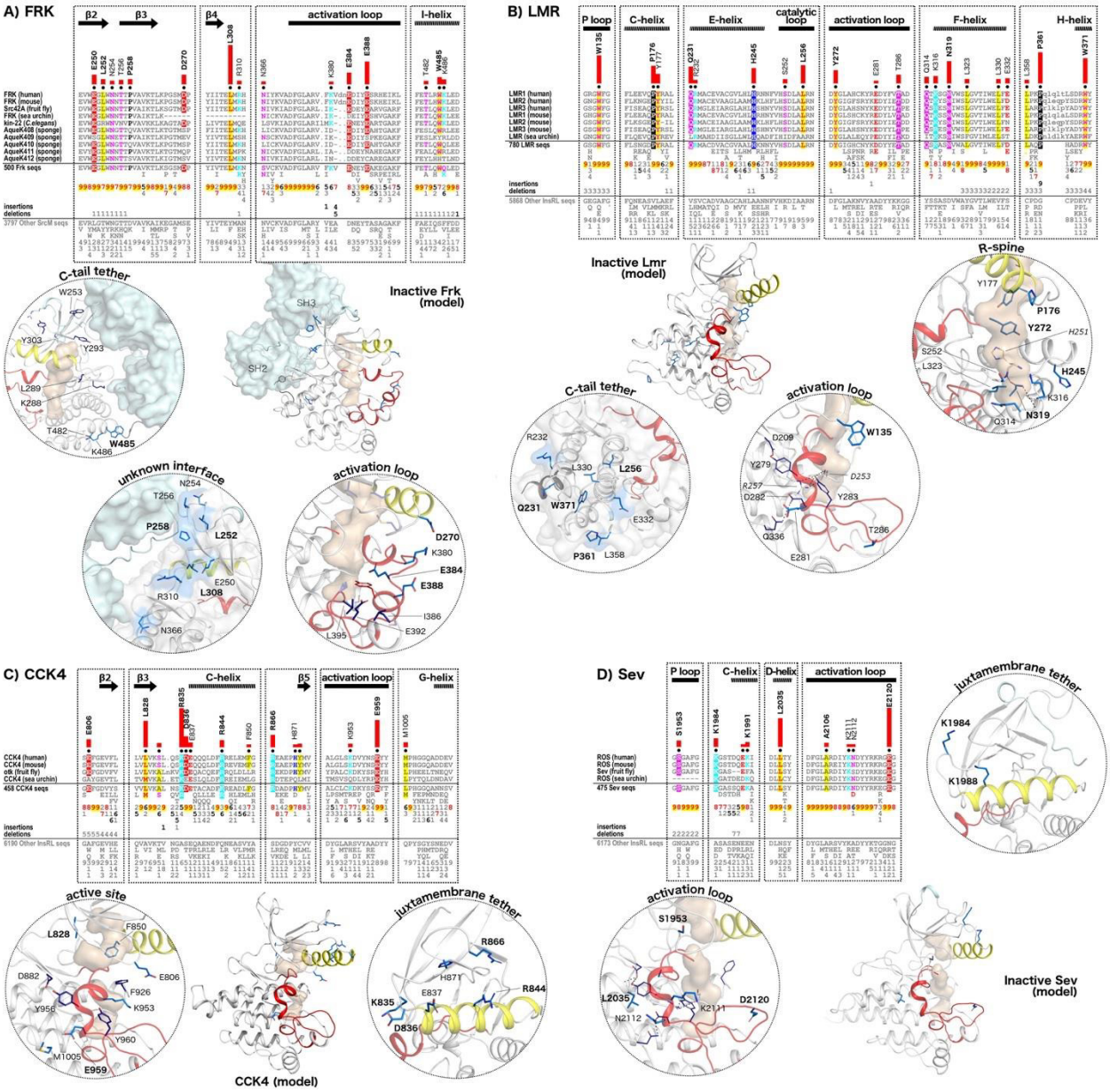

Figure 2.9: Regulatory models of understudied tyrosine kinases. Sequence features defining the A) Frk, B) Lmr, C) CCK4, and D) Sev families are shown. The N-terminal regions are shown in cyan, the regulatory spine in tan, the αC helix in yellow, and the activation loop in red. Family-specific sequence motifs are shown as blue sticks. The activation loop tyrosine phosphorylation site is shown in red lines. Glycines are depicted as spheres. Hydrogen bonding interactions are depicted by black dashes. The most highly distinguishing sequence motifs are labeled in bold font.

2.3 Conclusions

The classification of protein kinases since its inception has been used to compare and contrast regulatory functions of kinases and to infer unknown functions of lesser studied kinases, such as the SrcM kinase Frk. We demonstrate for the first time that tyrosine kinase families can be reliably classified into higher order subgroups based on the statistically significant conservation of distinguishing sequence motifs within these subgroups. The sequence motifs defining the SrcM, InsRL, and FPVR subgroups each contribute to unique regulatory modules that facilitate autoinhibitory interactions and/or their release during kinase activation. For example, the SrcM, InsRL, and PVR subgroups each have a defining autoinhibitory activation loop conformation that is facilitated by subgroup-specific motifs. In turn, despite the conservation of common regulatory cores, tyrosine kinases achieve further specialized regulation via the addition of family-specific features onto these regulatory cores. For example, several activating mechanisms of regulatory cores are conserved at the family level, such as the SrcA-specific activating modes of C-tail and SH2 linker/SH3 tethering.

Defining common regulatory modules and conformations is important not only for determining the functions of kinases, but also to study or infer the drug binding properties of kinases. Type II kinase inhibitors, which bind inactive conformations of kinases, often exhibit more selectivity toward their kinase targets than Type I inhibitors, which bind the active conformation, mostly due to the fact that inactive conformations of kinases are much more diverse than the active conformation. Despite this, like we have been able to show, many tyrosine kinases also share similar inactive conformations that can lead to diminished selectivity of kinase inhibitors. For example, the type II inhibitor ponatinib can target multiple members of the FPVR subgroup such as PDGFRA/B, Kit, Flt3, Ret and VEGFR and FGFR kinases(*46*).

Interestingly, some inhibitors can bind to and inhibit tyrosine kinases across broad subgroups, as demonstrated by ponatinib's ability to target SrcM kinases such as SRC and ABL as well. This is due to small similarities in inactive structures (i.e. DFG out) despite differences elsewhere. Therefore defining the common characteristics of inactive kinase conformations, as well as defining how they are unique in individual kinases or kinase families is an important piece of information to infer what kinases a drug will bind to, as well as to guide the development of more selective inhibitors that utilize unique characteristics of inactive conformations.

The conservation of particular sequence motifs across large subgroups of tyrosine kinases reflects an ancient conservation that, despite millions of years of evolution that allowed tyrosine kinases to diversify functions such as ligand specificity, suggests some indispensable function that has been retained. Interestingly the SrcM, FPVR, and InsRL subgroups represent some of the most anciently conserved tyrosine kinase lineages. For example, orthologs of each subgroup can be detected in choanoflagellates, which represent the closest relative to metazoan, suggesting that the diversification of these subgroups occurred as early as the emergence of tyrosine kinases and multicellularity associated with the emergence of metazoans in evolutionary history. This ancient evolutionary history may also be reflected in some shared sequence and functional characteristics as well. For example, both the SrcM and FPVR subgroups share a "WEX" motif in the linkers N-terminal to their kinase domains. In both subgroups, the tryptophan of the "WEX" motif docks onto the R-spine to modulate the active site. While the tethering interactions that facilitate docking of the "WEX" tryptophan onto the R-spine differs between subgroups and families, the conservation of the motif across such a large number of tyrosine kinase families suggests that it played a significant functional role across both early and later periods of tyrosine kinase evolution.

2.4 Materials and Methods

2.4.1 Classification of tyrosine kinase sequences

Tyrosine kinase sequences were identified using previously curated profiles of diverse tyrosine kinases (*7*) and the rapid and accurate alignment procedure MAPGAPS (*47*). Sequences that did not span from at least the β3-lysine to the DFG-aspartate were deemed fragmentary and removed. Two sequence sets were used as input into the optimal multiple-category Bayesian Partitioning with Pattern Selection algorithm (omcBPPS)(*48, 49*) to examine various classification hierarchies. One set included 17,071 representative diverse tyrosine kinase sequences from the NCBI nr database (downloaded 2/13/19) purged at 98% sequence identity, and the second sequence set comprised 12,137 tyrosine kinase sequences extracted from the UniProt reference proteome database (release 2019_02)(*25*). A cluster size cut off of 50 sequences was used to identify the major sequence families, and omcBPPS runs were performed with each sequence set using two different "minnats" parameters of 1 and 5, which determines the minimum information score (in nats) to classify a sequence into a family/subgroup node. All runs were replicated twice.

In order to compare the various omcBPPS classifications amongst each other and with the previously established Manning classification, we used a larger sequence set comprising all tyrosine kinase sequences detectable from the nr database (33,769 sequences) and multiple-category Bayesian Partitioning with Pattern Selection (mcBPPS) to classify the large sequence set based on various classification structures and compare statistical significance scores across the different classification structures. Seed sequences for each tyrosine kinase family were obtained from Kinbase (*7*), and the seed sequences for each family and the pre-determined classification of the families (based from various omcBPPS runs, explained above) were

supplied as inputs to mcBPPS. The log-probability ratios calculated by mcBPPS to score each classification structure and each subtree within each structure were used to compare and optimize the final classification structure.

2.4.2 Structural modeling and molecular dynamics simulations

Structural models were generated using Modeller (version 9.19)(*50*). All-atom unbiased MD simulations were produced with GROMACS 2016.4 (*51*). Hydrogens atoms were represented as virtual sites to remove the fastest vibrational freedom. Structures were parameterized using the AMBER- 99SB-ILDN force field, solvated with TIP3P water, and neutralized using sodium and chloride ions. The system was contained in a dodecahedral box at least 1 nm larger than the protein from all sides with periodic boundary conditions. Long range interactions were calculated using particle mesh Ewald. Neighbor lists were maintained by the Verlet cutoff scheme. The system underwent energy minimization using steepest descent minimization until the maximum force is < 100 kJ/mol. Canonical ensemble was used to warm the system from 0 to 310 K in 100 ps. Isothermal-isobaric ensemble (1 bar, 310 K) as applied for 100 ps. Positional restraints were applied during equilibration. The MD simulation used 5 fs timesteps.

Bibliography

1. M. A. Lemmon, J. Schlessinger, Cell signaling by receptor tyrosine kinases. *Cell* **141**, 1117-1134 (2010).

2. M. K. Paul, A. K. Mukhopadhyay, Tyrosine kinase - Role and significance in Cancer. *Int J Med Sci* **1**, 101-115 (2004).

3. G. Vlahovic, J. Crawford, Activation of tyrosine kinases in cancer. *Oncologist* **8**, 531-538 (2003).

4. S. K. Hanks, A. M. Quinn, T. Hunter, The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science* **241**, 42-52 (1988).

5. S. K. Hanks, T. Hunter, Protein Kinases .6. The Eukaryotic Protein-Kinase Superfamily - Kinase (Catalytic) Domain-Structure and Classification. *Faseb J* **9**, 576-596 (1995).

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

7. G. Manning, D. B. Whyte, R. Martinez, T. Hunter, S. Sudarsanam, The protein kinase complement of the human genome. *Science* **298**, 1912-1934 (2002).

8. Z. Wang, P. A. Cole, Catalytic mechanisms and regulation of protein kinases. *Methods Enzymol* **548**, 1-21 (2014).

9. J. A. Adams, Kinetic and catalytic mechanisms of protein kinases. *Chem Rev* **101**, 2271-2290 (2001).

10. S. R. Hubbard, J. H. Till, Protein tyrosine kinase structure and function. *Annu Rev Biochem* **69**, 373-398 (2000).

11. Y. Huang *et al.*, Molecular basis for multimerization in the activation of the epidermal growth factor receptor. *Elife* **5**, (2016).

12. A. C. Register, S. E. Leonard, D. J. Maly, SH2-catalytic domain linker heterogeneity influences allosteric coupling across the SFK family. *Biochemistry* **53**, 6910-6923 (2014).

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

14. S. R. Hubbard, Juxtamembrane autoinhibition in receptor tyrosine kinases. *Nat Rev Mol Cell Biol* **5**, 464-471 (2004).

15. H. B. Chen *et al.*, A molecular brake in the kinase hinge region regulates the activity of receptor tyrosine kinases. *Molecular Cell* **27**, 717-730 (2007).

16. M. Huse, J. Kuriyan, The conformational plasticity of protein kinases. *Cell* **109**, 275-282 (2002).

17. B. Nolen, S. Taylor, G. Ghosh, Regulation of protein kinases; controlling activity through activation segment conformation. *Mol Cell* **15**, 661-675 (2004).

18. N. H. Shah, J. F. Amacher, L. M. Nocka, J. Kuriyan, The Src module: an ancient scaffold in the evolution of cytoplasmic tyrosine kinases. *Crit Rev Biochem Mol Biol* **53**, 535-563 (2018).

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

20. T. L. Davis *et al.*, Autoregulation by the juxtamembrane region of the human ephrin receptor tyrosine kinase A3 (EphA3). *Structure* **16**, 873-884 (2008).

21. G. Manning, S. L. Young, W. T. Miller, Y. Zhai, 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**, 9674-9679 (2008).

22. S. Mohanty *et al.*, Hydrophobic Core Variations Provide a Structural Framework for Tyrosine Kinase Evolution and Functional Specialization. *PLoS Genet* **12**, e1005885 (2016).

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

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

25. C. The UniProt, UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**, D158-D169 (2017).

26. N. King *et al.*, The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature* **451**, 783-788 (2008).

27. G. Gosal, K. J. Kochut, N. Kannan, ProKinO: an ontology for integrative analysis of protein kinases in cancer. *PLoS One* **6**, e28782 (2011).

28. A. J. Lamontanara, S. Georgeon, G. Tria, D. I. Svergun, O. Hantschel, The SH2 domain of Abl kinases regulates kinase autophosphorylation by controlling activation loop accessibility. *Nat Commun* **5**, 5470 (2014).

29. R. E. M. Joseph, Lie; Andreotti, Amy, H., The Linker between SH2 and Kinase Domains Positively Regulates Catalysis of the Tec Family Kinases. *Biochemistry* **46**, 5455-5462 (2007).

30. Q. Wang *et al.*, Autoinhibition of Bruton's tyrosine kinase (Btk) and activation by soluble inositol hexakisphosphate. *Elife* **4**, (2015).

31. B. Nagar *et al.*, Structural basis for the autoinhibition of c-Abl tyrosine kinase. *Cell* **112**, 859-871 (2003).

32. T. Schindler *et al.*, Crystal structure of Hck in complex with a Src family-selective tyrosine kinase inhibitor. *Mol Cell* **3**, 639-648 (1999).

33. F. von Raussendorf, A. de Ruiter, T. A. Leonard, A switch in nucleotide affinity governs activation of the Src and Tec family kinases. *Sci Rep* **7**, 17405 (2017).

34. S. W. Cowan-Jacob *et al.*, The crystal structure of a c-Src complex in an active conformation suggests possible steps in c-Src activation. *Structure* **13**, 861-871 (2005).

35. N. Amatya *et al.*, Autoinhibition beyond the SRC module: the multifacted pleckstrin homology domain of the TEC family *eLife*, (2019).

36. N. M. Levinson *et al.*, A Src-like inactive conformation in the abl tyrosine kinase domain. *PLoS Biol* **4**, e144 (2006).

37. N. Dolker *et al.*, The SH2 domain regulates c-Abl kinase activation by a cyclin-like mechanism and remodulation of the hinge motion. *PLoS Comput Biol* **10**, e1003863 (2014).

38. J. Colicelli, ABL tyrosine kinases: evolution of function, regulation, and specificity. *Sci Signal* **3**, re6 (2010).

39. S. R. Hubbard, L. Wei, L. Ellis, W. A. Hendrickson, Crystal structure of the tyrosine kinase domain of the human insulin receptor. *Nature* **372**, 746-754 (1994).

40. S. Munshi, D. L. Hall, M. Kornienko, P. L. Darke, L. C. Kuo, Structure of apo, unactivated insulin□like growth factor□1 receptor kinase at 1.5 Å resolution. *Acta Crystallographica* **59**, 1725 (2003).

41. S. Munshi *et al.*, Crystal structure of the Apo, unactivated insulin-like growth factor-1 receptor kinase. Implication for inhibitor specificity. *J Biol Chem* **277**, 38797-38802 (2002).

42. S. C. Artim, J. M. Mendrola, M. A. Lemmon, Assessing the range of kinase autoinhibition mechanisms in the insulin receptor family. *Biochem J* **448**, 213-220 (2012).

43. J. H. Till *et al.*, Crystal structure of the MuSK tyrosine kinase: insights into receptor autoregulation. *Structure* **10**, 1187-1196 (2002).

44. P. L. Liu, L. Du, Y. Huang, S. M. Gao, M. Yu, Origin and diversification of leucine-rich repeat receptor-like protein kinase (LRR-RLK) genes in plants. *BMC Evol Biol* **17**, 47 (2017).

45. N. Furuya *et al.*, The juxtamembrane region of TrkA kinase is critical for inhibitor selectivity. *Bioorg Med Chem Lett* **27**, 1233-1236 (2017).

46. P. Canning *et al.*, Structural mechanisms determining inhibition of the collagen receptor DDR1 by selective and multi-targeted type II kinase inhibitors. *J Mol Biol* **426**, 2457-2470 (2014).

47. A. F. Neuwald, Rapid detection, classification and accurate alignment of up to a million or more related protein sequences. *Bioinformatics* **25**, 1869-1875 (2009).

48. A. F. Neuwald, Evaluating, comparing, and interpreting protein domain hierarchies. *J Comput Biol* **21**, 287-302 (2014).

49. A. F. Neuwald, A Bayesian sampler for optimization of protein domain hierarchies. *J Comput Biol* **21**, 269-286 (2014).

50. A. Fiser, R. K. Do, A. Sali, Modeling of loops in protein structures. *Protein Sci* **9**, 1753-1773 (2000).

51. M. J. Abraham *et al.*, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19-25 (2015).

Chapter 3

Coupled regulation by the juxtamembrane and sterile α motif (SAM) linker is a hallmark of Ephrin tyrosine kinase evolution

Accepted by *J Biol Chem*. Annie Kwon, et al. (2018) *J Biol Chem* 293(14): 5102-16
Reprinted here with permission from the publisher.

Abstract

Ephrin (Eph) receptor tyrosine kinases have evolutionarily diverged from other tyrosine kinases to respond to specific activation and regulatory signals that require close coupling of kinase catalytic and regulatory functions. However, the evolutionary basis for such functional coupling is not fully understood. We employed an evolutionary systems approach involving statistical mining of large sequence and structural data sets to define the hallmarks of Eph kinase evolution and functional specialization. We found that some of the most distinguishing Eph-specific residues structurally tether the flanking juxtamembrane and sterile α motif (SAM) linker regions to the kinase domain, and substitutions of these residues in EphA3 resulted in faster kinase activation. We report for the first time that the SAM domain linker is functionally coupled to the juxtamembrane through co-conserved residues in the kinase domain and that together these residues provide a structural framework for coupling catalytic and regulatory functions. The unique organization of Eph-specific tethering networks and the identification of other Eph-specific sequence features of unknown functions provide new hypotheses for future functional studies and new clues to disease mutations altering Eph kinase–specific functions.

3.1 Introduction

The ephrin (Eph) family of receptor tyrosine kinases comprises the largest family of tyrosine kinases and controls critical developmental processes by modulating cell adhesion, cell migration, and cytoskeletal organization in a variety of cell types (1, 2). The human Eph family includes 14 members, which are further divided into two subclasses, EphA and EphB, based on sequence similarity and ligand affinity. Two members, EphA10 and EphB6, are classified as pseudokinases based on their divergence in sequence motifs essential for catalytic activity (3-5). However, all Eph members share a common domain organization, characterized by an

extracellular ligand-binding region and an intracellular region that includes the juxtamembrane segment, tyrosine kinase domain, a sterile α motif (SAM) domain, and a PDZ-binding motif (1, 2). The implication of Eph family kinases in a broad array of diseases from Alzheimer's disease, viral pathogenesis, and multiple types of cancer (2, 6) underscores the biomedical significance of these enzymes and rationalizes the interest and investment in investigating their molecular mechanisms.

Like other receptor tyrosine kinases, the enzymatic activity of Eph kinases is tightly regulated by molecular mechanisms that allow specific responses to ligand binding, post-translational modifications, and other molecular events. Activation of the receptor is achieved when the extracellular domain binds membrane-anchored ephrin ligands, which triggers the autophosphorylation of tyrosines in the cytoplasmic kinase domain (1, 7). Autophosphorylation on two conserved tyrosines in the juxtamembrane (Y596 and Y602 in EphA3) is a prerequisite for catalytic activity (8-13) and precedes the autophosphorylation of the activation loop tyrosine (Y779 in EphA3) (8). Activation loop phosphorylation increases catalytic efficiency as seen in many protein kinases (14), however, unlike the juxtamembrane tyrosines, mutation of the activation loop tyrosine to phenylalanine decreases but does not abolish catalytic or autophosphorylation activity (10, 12). Additionally, phosphorylation of these three major sites as well as other minor sites serve to bind SH2 domain-containing proteins, such as Src, Crk, Nck, and RasGAP (15-17). Thus, once the juxtamembrane becomes autophosphorylated, further autophosphorylation on additional sites as well as direct substrate phosphorylation propagates signaling through downstream pathways.

While the roles of juxtamembrane and activation loop autophosphorylation in Eph kinase functions are well recognized, little is known about how other molecular events are involved in

Eph regulation and autophosphorylation. For example, regulation via the formation of higher order complexes is well understood for some receptor tyrosine kinases, for example, in the EGFR family of receptor tyrosine kinases (18, 19); however, the mechanisms of Eph kinase regulation by oligomerization are not well established. In particular, the difficulty of determining the structure of oligomeric complexes and the lack of full-length crystal structures for the large Eph family, whose members are thought to hetero-oligomerize (20), have contributed to the difficulty in understanding the intricate details of Eph oligomerization. Despite the determination of several dimeric structures of isolated SAM domains from Eph kinases (20, 21), the contribution of the SAM domain in oligomerization or other forms of regulation is yet to be fully characterized. Conflicting reports on the effects of SAM domain deletion mutations on dimerization and activity in EphA3 (22), EphA2 (23), and EphB2 (10) further confound our understanding of the Eph SAM domain. Furthermore, though the SAM domain linker is resolved in multiple crystal structures (9), whether it plays any functional role has never been explored.

Inferring mechanisms from protein sequences provides an alternative and complementary approach to experimental and biochemical characterization methods. In particular, currently unanswered questions such as the evolutionary bases for Eph kinase functional divergence through oligomerization or interactions with regulatory domains and protein segments can be inferred through statistical mining of protein sequences. Indeed, previous statistical analyses of evolutionary constraints acting on protein kinase sequences have provided important insights into protein kinase evolution and allosteric regulation (24-29). The Eph family serves as an ideal system for evolutionary sequence studies because it is a monophyletic family with detectable orthologs throughout extremely diverse metazoan phyla, from chordates to nematodes,

poriferans, and choanoflagellates (30-32). The same domain structure is conserved across most metazoans, indicating that the overall structure and function of Eph receptors is likely well conserved throughout metazoan evolution (33). The Eph kinase domain conserves prototypical features of the protein kinase domain, which is comprised of the N-terminal ATP binding lobe (N-lobe) and the C-terminal substrate binding lobe (C-lobe)(9, 11, 34).

In this study, we use a Bayesian statistical framework (35) to identify sequence features most characteristic of the Eph family. We show for the first time that nearly all residues that distinguish the Eph family from other tyrosine kinases occur on the surface of the protein, and that some of the most distinguishing residues tether flanking protein segments to the kinase domain. The selective conservation of these residues within the Eph family suggests that they play important roles in Eph kinase functions. Single mutations of juxtamembrane and SAM linker tethering residues both resulted in more rapid activation of EphA3, for the first time highlighting the SAM domain linker as a critical functional component in the activation of Eph kinases. Simultaneous mutation of both networks demonstrated that the SAM linker plays a regulatory role via allosteric coupling to the juxtamembrane. The emerging model of Eph evolution spotlights the co-evolution of the juxtamembrane and the SAM linker with the kinase domain to precisely coordinate kinase catalytic and regulatory functions. We also identify highly distinct Eph-specific motifs of unknown functions that warrant further investigation.

3.2 Results

3.2.1 Unique sequence features distinguish the Eph family from other tyrosine kinases

To identify sequence features that most distinguish Eph kinases from other tyrosine kinases, we systematically compared large and diverse sequence sets of Eph kinases and non-Eph tyrosine kinases using computational methods previously described (35). We note that the regions flanking the Eph kinase domain, namely the N-terminal juxtamembrane and the C-terminal SAM domain linker, are unique to the Eph family in that they share detectable sequence similarity within the Eph family, but share no similarity with kinase sequences outside of the Eph family. The juxtamembrane and SAM domain linker are both well conserved, sharing 47.9% and 45.9% sequence similarity, respectively, within the Eph family. Eph kinases are the only family of tyrosine kinases with a conserved SAM domain, therefore the SAM domain, as well as the SAM domain linker, are unique to Eph kinase sequences.

To identify sequence constraints that distinguish the Eph kinase domain from other tyrosine kinase domains, we used a Bayesian pattern partitioning procedure, which classifies multiply aligned sequences based on patterns of amino acid conservation and variation (35). Residues that are highly conserved in Eph kinase sequences but non-conserved and/or biochemically dissimilar in non-Eph tyrosine kinase sequences are highlighted in Figure 3.1, where the histograms represent the relative strengths of the constraints imposed on these residues. Some Eph-specific residues occur in the N-lobe, specifically in the αC-helix, β4 strand, and β5 strand. Additionally, highly distinguishing Eph-specific residues are scattered throughout the C-lobe and are located in the αD/αE/αF/αG-helices, β7/β8 strands, αF-αG loop, and the αH-αI loop.

In the following sections, we first provide a brief overview of the structural locations of Eph-specific residues. Next, we report our detailed structural and experimental analysis of Eph-specific residues that take part in well-conserved networks involving the unique N- and C-terminal segments flanking the kinase domain. Eph-specific residues of unknown functions were also investigated and are reported below.

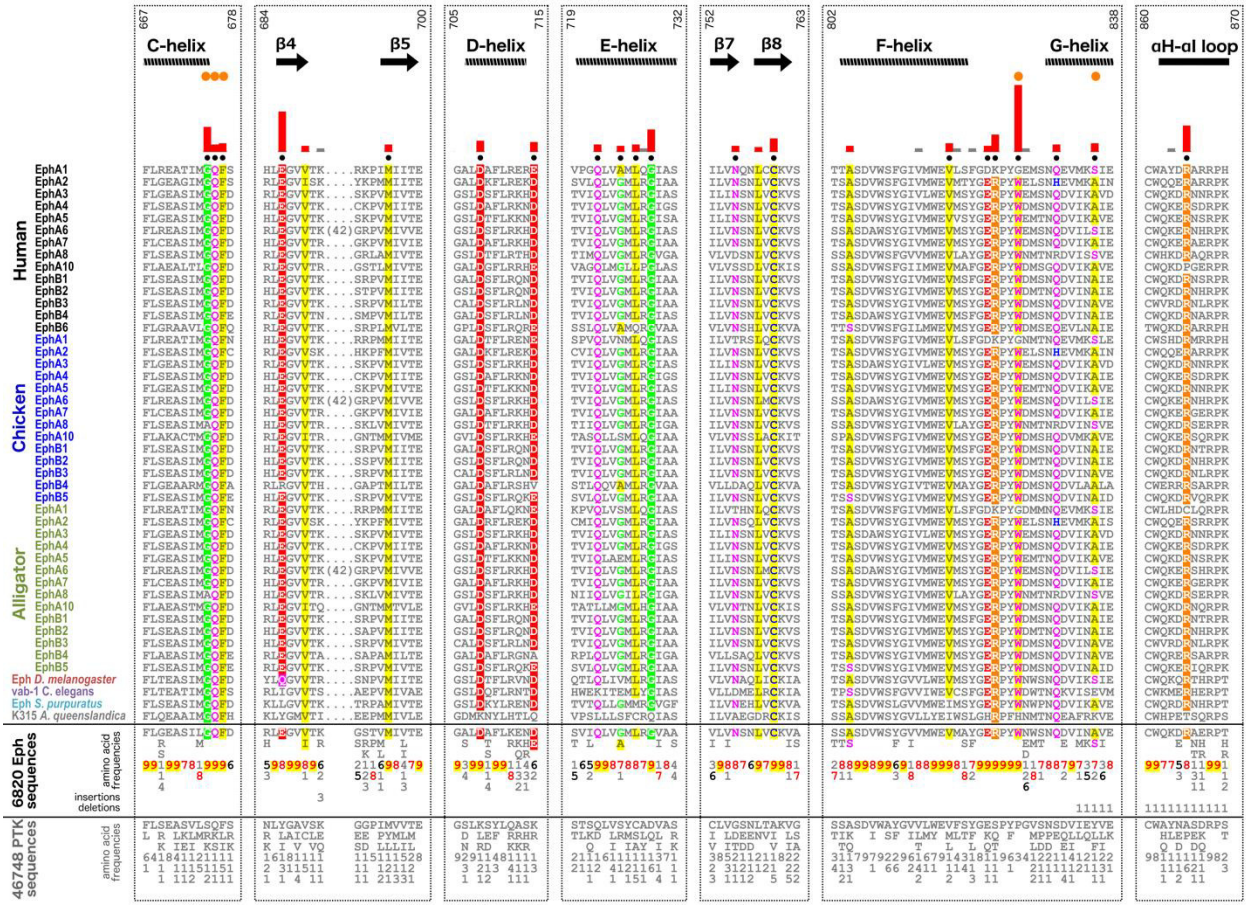

Figure 3.1: Unique sequence features defining the Eph family kinase domain. Columns are highlighted where amino acids are highly conserved in Eph family sequences and non-conserved and/or biochemically dissimilar in other tyrosine kinase (PTK) sequences. Histograms quantify the degree of divergence between Eph and other tyrosine kinase sequences. Column-wise amino acid and insertion/deletion frequencies are indicated in integer tenths, where a '5' indicates an occurrence of 50-60% in the given (weighted) sequence set. Columns used by the Bayesian partitioning procedure to sort Eph sequences from other tyrosine kinase sequences are marked with black dots. Kinase secondary structures are annotated above the alignment. Eph-specific residues that participate in juxtamembrane and SAM linker tethering networks are marked by

orange dots. Sequence numbering above the alignment corresponds to the human EphA3 sequence.

3.2.2 Structural location of Eph-specific residues

Sequence features distinguishing the Eph kinase domain are dispersed in primary sequence, however, a structural mapping of these residues onto crystal structures of EphA3 shows that Eph-specific residues cluster into networks that are largely located on the surface of the protein (Figure 3.2). One of these networks has previously been noted to be significant for Eph function and tethers the unique juxtamembrane segment to the kinase domain (Figure 3.2, panel A). However, many other Eph-specific residues form networks dispersed across the N-lobe, hinge region, and C-lobe of the kinase domain whose functions are yet to be determined. In order to shed light on the roles of these conserved residues, we have performed detailed structural analyses and mutational studies using human EphA3 as a model Eph family kinase. In our experimental analysis, we mutated Eph-specific residues to those observed in other tyrosine kinases and determined the impact of mutations on kinase activation via autophosphorylation. We also determined the effects on substrate phosphorylation using enolase as a generic protein substrate.

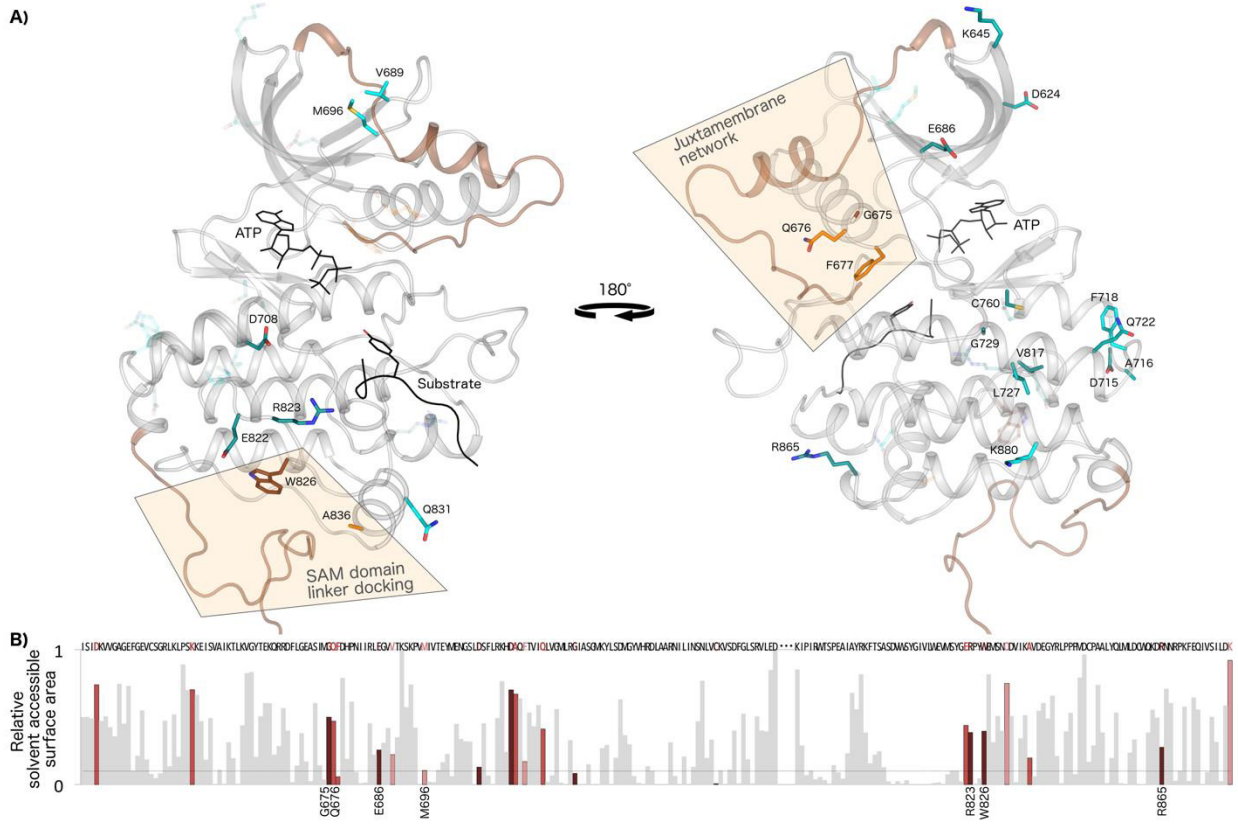

Figure 3.2: Networks of Eph-specific residues in distinct functional regions of the kinase tertiary structure. A) Key distinguishing residues of the Eph family mapped onto a model structure of human EphA3. ATP and peptide substrate (PDB: 3FY2) are shown in black. The juxtamembrane (modeled from PDB: 2QO2) and the SAM domain linker (modeled from PDB: 2QOC) is shown in brown. The activation loop was modeled using active EphA2 (PDB: 4TRL) and active human Lck as a template (PDB: 3LCK). The 25 most distinguishing Eph-specific residues (columns with black dot in Figure 3.1) are shown as sticks, which are colored darker to lighter based on the greater or lesser contrast between Eph's and other tyrosine kinases at that position. B) Relative solvent accessible surface area of residues across the kinase domain calculated from an active EphA3 structure (PDB: 2QOC). Eph-specific residues are highlighted in dark to light red relative to the strength of the sequence constraint, with darker red indicating more highly constrained residues. Residues that are further investigated in the study are labeled. Data for the disordered activation loop is omitted and marked by ellipses.

3.2.3 Eph-specific networks tether unique flanking segments onto the kinase domain

3.2.3.1 Eph-specific interactions between the kinase domain and juxtamembrane

The "GQF" motif in the αC-β4 loop (Figure 3.1) forms a network of critical interactions with the juxtamembrane, and these interactions have been highlighted in previous biochemical and structural studies on various Eph members (8-13). In crystal structures of autoinhibited Eph kinase domains, the N-terminal juxtamembrane phosphorylation site, Tyr596 (EphA3 numbering), docks into a pocket formed by Eph-specific residues Phe677 and Gln676. The side chain of Gln676 also forms two hydrogen bonds to the backbone of the juxtamembrane (Figure 3.3) (9). This autoinhibitory configuration prohibits the activation loop from adopting an active conformation due to steric hindrance by another important tyrosine residue in the network, Tyr742 (9). In comparison, the active structure of EphA3 exhibits shifts in the side chain orientations of residues in this network. For example, the hydrogen bonds by Gln676 to the juxtamembrane backbone are instead satisfied by the side chain of Tyr742, which is rotated away from the active cleft in a manner that allows the open conformation of the activation loop (Figure 3.3). Gly675 is the most highly distinguishing residue in the GQF motif (Figure 3.1) and has not yet been recognized for its role in juxtamembrane function in previous studies. This glycine residue caps the C-terminal end of the αC-helix, causing the αC-helix to be a half to full turn shorter than observed in other tyrosine kinases (Figure 3.3). Interestingly, a variation of the GQF motif is observed in chicken and alligator EphA8 sequences, which harbor a biochemically similar "AQF" variation, though they conserve all other canonical residues of the juxtamembrane network (Figure 3.1). Through further taxonomic analyses of EphA8 sequences, we have determined that the alanine variant of the GQF motif is selectively conserved in non-mammalian

vertebrate EphA8 sequences.

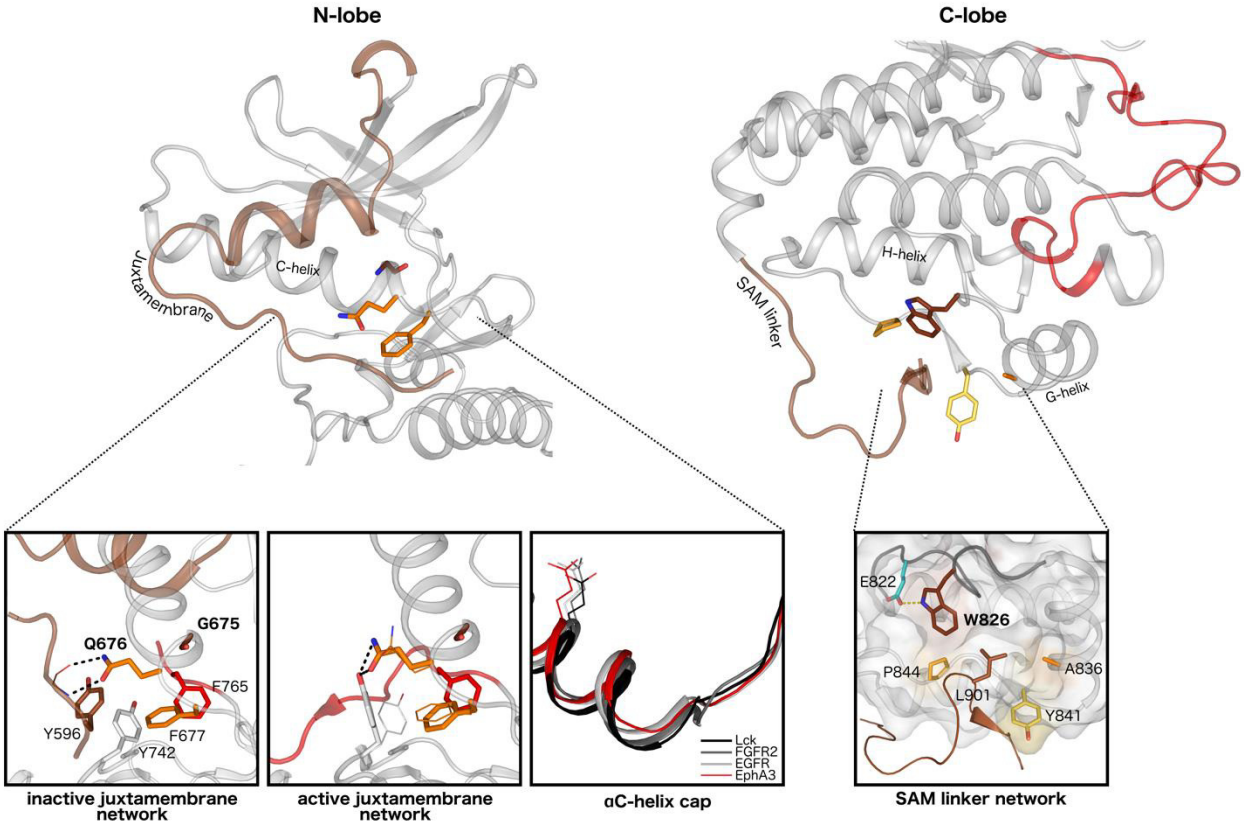

Figure 3.3: Structural interactions in the juxtamembrane network and SAM domain linker network. A) Key interactions between the juxtamembrane, the GQF motif, and kinase domain residues are highlighted in the autoinhibited (PDB: 2QO2) and active (PDB: 2QOC) structures of EphA3. The juxtamembrane tyrosine is shown in brown, the tethering GQF motif in orange, and the DFG-phenylalanine in red. Different side chain occupancies of residues observed in the active structure are shown as lines. An alignment of αC-helixes of various tyrosine kinases reveals the earlier helix capping observed in Eph structures. The conserved ATP-binding αC-helix glutamate is shown as a reference for the structural alignment. B) Interactions between the SAM domain linker and the C-lobe of the kinase domain are highlighted in an active structure of EphA3 (PDB: 2QOC). The hydrophobic pocket in the C-lobe is depicted in surface representation, and the surface of Eph-specific residues that make up the pocket is colored. The αF-αG loop backbone is highlighted in dark gray.

3.2.3.2 Eph-specific interactions between the kinase domain and SAM domain linker

Another group of Eph-specific residues map to the basal face of the C-lobe and form a network of interactions tethering the SAM domain linker to the kinase domain (Figure 3.3). Though these tethering interactions have been observed in multiple inactive and active EphA3

structures (9), the functional significance of these interactions has not been explored. In crystal structures where the SAM domain linker is ordered, a leucine in the linker (Leu901 in EphA3) is tethered onto the C-lobe of the kinase domain between the αF-αG loop, αG-helix, and αG-αH loop (9). The side chain of Trp826, which is the most distinguishing Eph-specific residue in this network, forms a hydrophobic tethering pocket along the αF-αG loop. The αF-αG loop is observed in a unique conformation that is divergent from other tyrosine kinase structures (Figure 3.6), and this conformation is associated with a hydrogen bond between the indole nitrogen of Trp826 to the side chain oxygen of another Eph-specific residue in the αF-αG loop, Glu822. The hydrophobic tethering pocket is additionally formed by other Eph-specific residues (Ala836 in αG-helix, and Pro844 and Tyr841 in the αG-αH loop).

3.2.3.3 Mutation of N- and C-lobe tethering networks results in more rapid activation of EphA3

To shed light on the roles of these N-lobe and C-lobe tethering networks, we mutated select Eph-specific residues in these networks to residues observed at the equivalent position in other tyrosine kinases and determined the impact of mutations on autophosphorylation and enolase phosphorylation. As seen in Figure 3.4, WT EphA3 begins to exhibit activation loop autophosphorylation (pY779) after 5 minutes of incubation with MgATP and approaches maximum phosphorylation levels at 1 hour. To validate the role of the GQF motif from the juxtamembrane network, we mutated the GQF-glutamine (Q676) to glutamate, which is found at this position in other tyrosine kinase sequences such as Abl and Musk. We also examined the functional impact of the AQF variant by mutating the GQF-glycine (G675) to alanine, which is observed in select EphA8 sequences. The glycine mutant, G675A, exhibited similar activation time to WT EphA3, whereas the glutamine mutant, Q676E, achieved much faster activation, with autophosphorylation detected as early as 30 seconds (Figure 3.4, panel A). In contrast,

neither the Q676E or G675A mutants exhibited any differences compared to WT with respect to rates of enolase phosphorylation (Figure 3.4, panel B). In the SAM domain linker network, we mutated the tryptophan (W826) to proline, which is the residue observed in over 30% of other tyrosine kinase sequences. We also perturbed this network via a truncation mutant that removes the SAM domain linker at position T892 (ΔSAMlinker). The W826P mutant exhibited significantly increased rates of autophosphorylation with high levels of autophosphorylation detected at 30 seconds. In contrast, the ΔSAMlinker mutation reduced rates of autophosphorylation, with activation loop phosphorylation detected only after 10 minutes, and decreased total autophosphorylation levels to about 60% compared to WT (Figure 3.4, panel A). Neither of the SAM linker network mutations caused any observable effect on enolase phosphorylation (Figure 3.4, panel B). We additionally characterized the rates of activation and peptide phosphorylation for WT and W826P proteins using an NADH-coupled assay, which also showed that the W826P mutant achieves faster rates of activation compared to WT (Table 3.1).

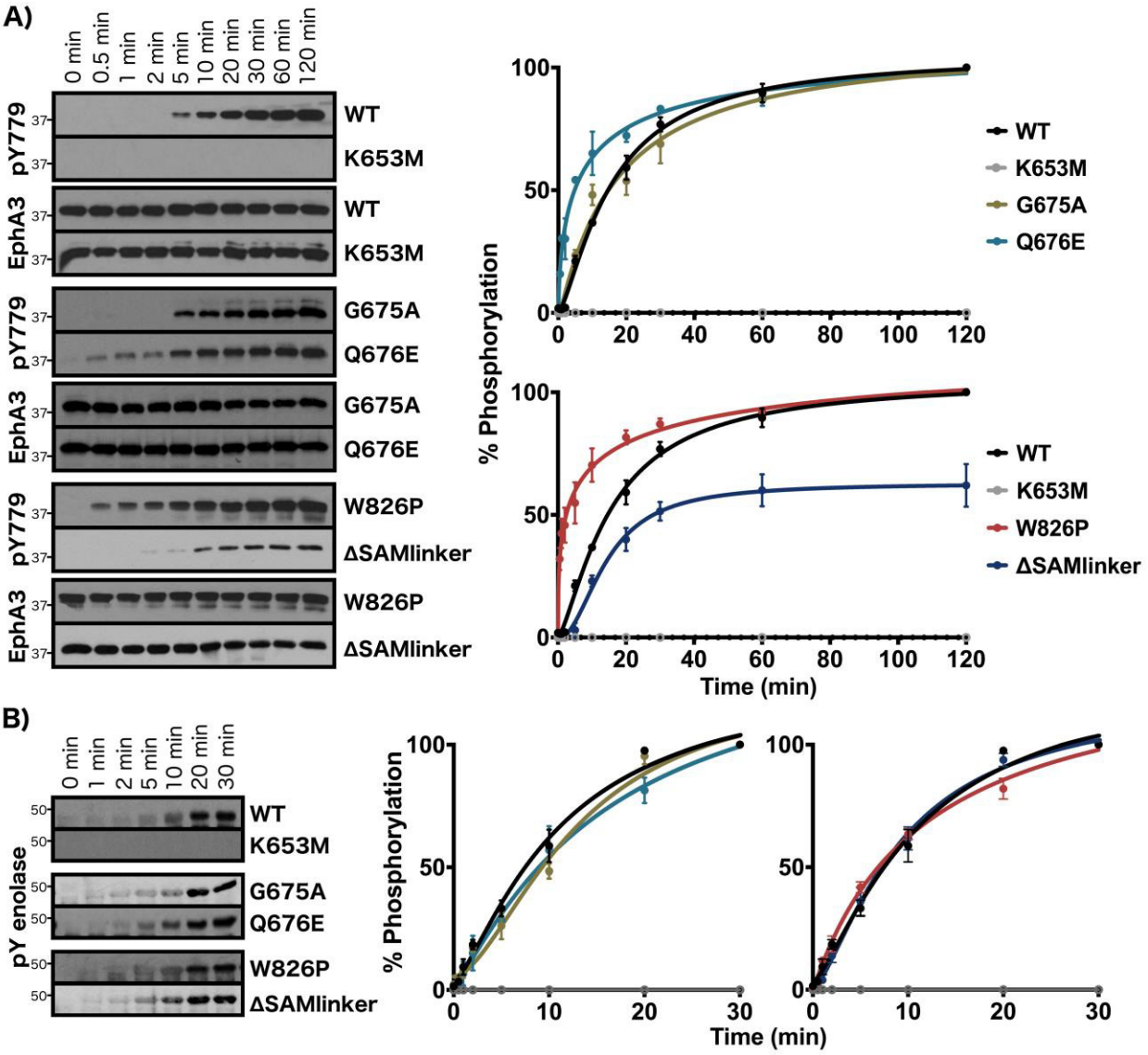

Figure 3.4: Experimental analysis of juxtamembrane and SAM linker network residues. A) Autophosphorylation time series of WT EphA3, kinase dead EphA3 (K653M), and juxtamembrane and SAM linker network mutants of EphA3. Phosphorylation levels of the activation loop site were detected (pY779). B) Enolase phosphorylation time series. Phosphorylation levels in (B) and (C) were quantified by densitometry and normalized against total protein levels. Data shown represent experiments performed in triplicates or more.

	Specific activity (nmol/min×μmol)	
	active	inactive
WT	3.70 ± 0.19	0.607 ± 0.08
W826P	2.93 ± 0.16	0.954 ± 0.18

Table 3.1: Specific activities of phosphorylated (active) and unphosphorylated (inactive) WT and W826P EphA3.

3.2.4 SAM linker tethering residue is coupled to juxtamembrane autoregulatory function

Mutational analysis of the highly distinct SAM linker tethering residue (W826P) resulted in significantly more rapid activation similar to that seen when directly mutating the autoinhibitory juxtamembrane network (Q676E). Because the most critical step in Eph activation is autophosphorylation of the juxtamembrane, which is believed to precede any other autophosphorylation event, we sought to determine whether the mutation directly affected juxtamembrane autophosphorylation. We used MALDI peptide mass fingerprinting and LC-MS/MS to identify phosphorylated juxtamembrane residues (pY596, pY602) for WT and W826P samples after 30 seconds and after 10 minutes of incubation with MgATP (Table 3.2). For the W826P mutant, we detected the singly-phosphorylated juxtamembrane peptide for the 30s sample, and we detected the doubly-phosphorylated juxtamembrane peptide for the 10 min sample. In contrast, only the singly-phosphorylated juxtamembrane peptide was readily detected for WT for the 10 min sample. Further LC-MS/MS and CID analysis of the W826P 30s sample was used to distinguish between phosphorylated N-terminal and C-terminal tyrosine residues in the juxtamembrane. Our data showed that the juxtamembrane peptide phosphorylated on the C-terminal tyrosine is readily detected in contrast to the N-terminally phosphorylated or doubly-

phosphorylated peptide, corroborating that the C-terminal tyrosine is phosphorylated first in the sequence of autophosphorylation events as has been determined in a previous study (8).

			WT 30s			WT 10m		
Sequence	**PTMs**	**Theoretical m/z**	**Observed m/z**	**Intensity**	**Mass error (ppm)**	**Observed m/z**	**Intensity**	**Mass error (ppm)**
TYVDPHTYEDPTQTVHEFAK		2379.56	2379.3	140280.0	-88.9	2379.5	23529.0	-17.0
TYVDPHTYEDPTQTVHEFAK	phos. (1)	2459.54				2459.1	19835.0	-167.2
TYVDPHTYEDPTQTVHEFAK	phos. (2)	2539.52						
			W826P 30s			**W826P 10m**		
Sequence	**PTMs**	**Theoretical m/z**	**Observed m/z**	**Intensity**	**Mass error (ppm)**	**Observed m/z**	**Intensity**	**Mass error (ppm)**
TYVDPHTYEDPTQTVHEFAK		2379.56	2379.6	108491.0	14.5	2379.3	25899.0	-110.3
TYVDPHTYEDPTQTVHEFAK	phos. (1)	2459.54	2459.8	54618.0	106.4	2459.3	41238.0	-84.3
TYVDPHTYEDPTQTVHEFAK	phos. (2)	2539.52				2539.3	50184.0	-90.6

Table 3.2: Tryptic juxtamembrane peptides identified by MALD peptide mass fingerprinting.

In addition, we produced double- and triple- mutations at sites in the juxtamembrane and SAM linker networks to investigate the cooperativity of these networks. We produced a YFYF mutant with double tyrosine-to-phenylalanine mutations in the juxtamembrane, as well as a triple mutant containing the W826P mutation in addition to the YFYF mutations. Substitution of both juxtamembrane tyrosines to phenylalanine was previously found to completely abolish catalytic activity in various Eph family kinases (8-13). We observed that the juxtamembrane double mutant (YFYF) showed almost no autophosphorylation activity, with minor levels of autophosphorylation detected from 20 min to 2 hours (Figure 3.5, panel A). Interestingly, the W826P mutation, which significantly increased the rate of EphA3 activation when examined as a single mutant, was able to rescue activity when added in the YFYF double mutant background (Figure 3.5, panel A). The triple YFYF+W826P mutant exhibited significant levels of

autophosphorylation by 10 to 20 minutes, similar to WT EphA3, but lesser than levels exhibited by the single mutant alone (W826P).

In order to determine whether the mutations may be disrupting autoinhibition by destabilizing the autoinhibited state, we examined the global protein stability of the inactive, dephosphorylated forms of WT EphA3 and its mutants through a dye-based thermal shift assay. As shown in Figure 3.5, panel B, the W826P mutation decreased global stability and resulted in a T_m 1.5°C lower than the T_m of WT. The addition of the YFYF double mutation had a stabilizing effect on both the WT protein and the W826P mutant. The YFYF mutant exhibited a T_m 2°C higher than the T_m of WT, and the YFYF+W826P triple mutant resulted in a 2.5°C increase in T_m relative to the W826P single mutant. We also examined the thermal stability of the SAM domain linker deletion mutant, ΔSAMlinker, which was slightly more stable than WT, with a ΔT_m of +0.83°C. The ΔSAMlinker mutant was also much more stable than the W826P mutant by 2.1°C. In addition, we used molecular dynamics simulations to analyze the dynamics and stability of autoinhibited models of WT and W826P EphA3. As shown in Figure 3.5, panel C, the W826P mutation increases the positive correlative motions between the juxtamembrane (residues 595-620) and SAM linker (residues 892-906). The regions that show the most difference in cross-correlated motions between WT and W826P are also highlighted in Figure 3.5, panel C and include the activation loop (residues 764-794), F-G loop (residues 820-829), SAM linker, and other regions in the C-lobe. Notably, the positive correlations between the F-G loop with other regions of the C-lobe (G-H loop, H- and I-helices, SAM linker) are markedly decreased in the W826P mutant. Similarly, positive correlations between the activation loop and the substrate-binding regions (F-G loop/G-helix) and between the activation loop and SAM linker are decreased in the W826P mutant as well.

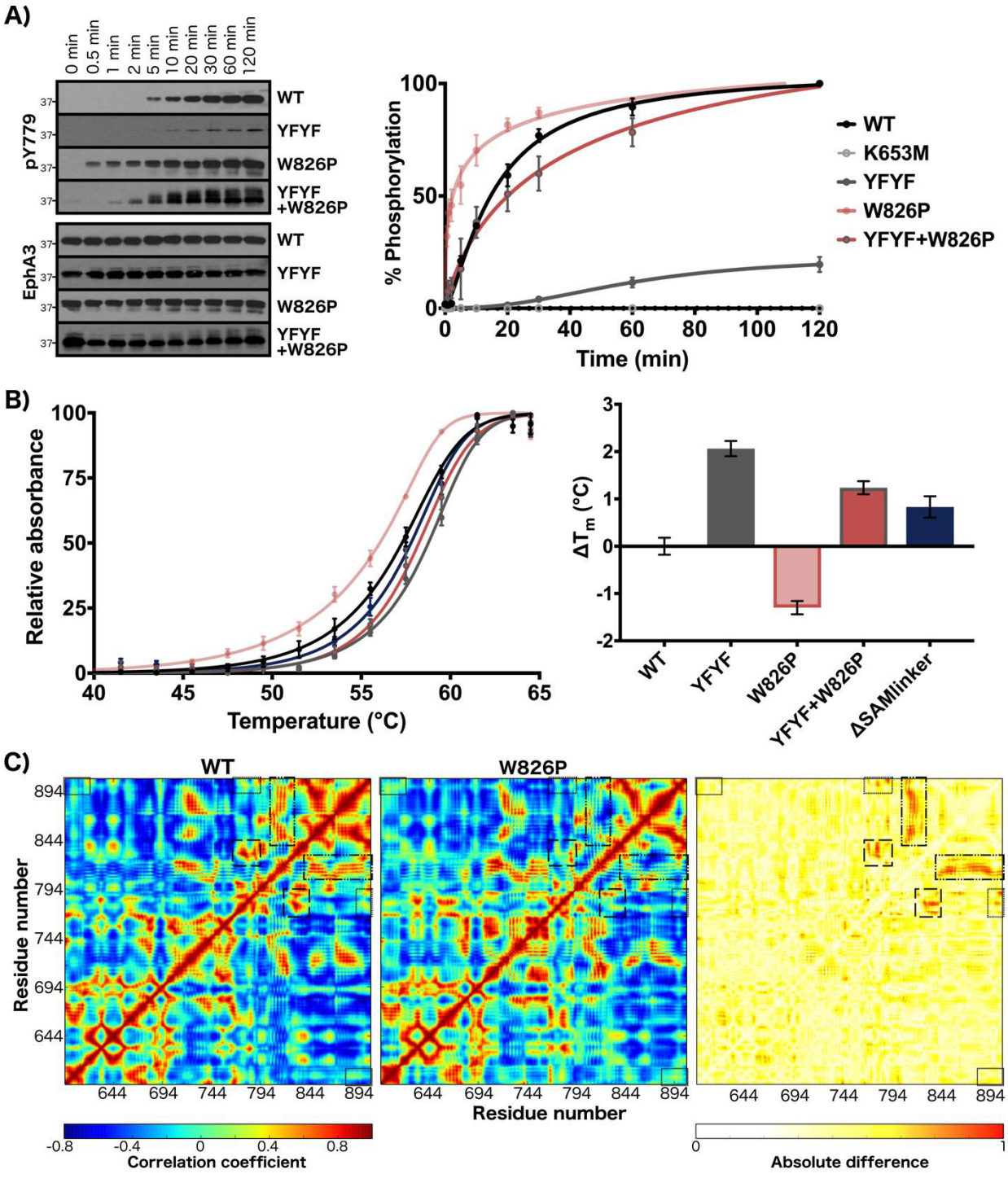

Figure 3.5: Characterization of coupled behavior between the juxtamembrane and SAM domain linker. A) Autophosphorylation time series of YFYF, W826P, and YFYF+W826P EphA3 mutants. Phosphorylation levels of the activation loop site were detected (pY779). Data for WT and W826P EphA3 autophosphorylation from Figure 3.4 are shown for comparison. B) Thermal stability analysis. The change in T_m for the mutants were calculated with respect to WT. Data shown represent experiments performed in triplicates or more. C) Cross-correlation analysis for

500 ns WT and W826P molecular dynamics trajectories. Correlation coefficient values are colored from red (positive) to blue (negative). Cross-correlations between the juxtamembrane (residues 595-620) and SAM linker (residues 892-906) are highlighted in a black box. Also highlighted are cross-correlations of the F-G loop (820-829) with the G-H loop (840-849), H-helix (850-860), H-I loop (861-869), I-helix (870-891), and SAM linker in the large dash-dot box, cross-correlations between the activation loop (764-794) and F-G loop/G-helix (820-839) in the long-dash box, and cross-correlations between the activation loop and SAM linker in the short-dash box. A difference map of cross-correlation coefficients between WT and W826P simulations (right) shows the regions exhibiting the most difference.

3.2.5 Eph-specific networks of unknown function

Eph-specific residues of unknown functions locate to the β1, β3, β4, and β5 strands of the N-lobe, and to the αD-αE loop, αE-helix, and β7/β8 strands in the hinge region of the kinase domain. Several Eph-specific residues additionally map to the C-lobe, spanning the αF-helix, αF/αG-loop, αG-helix, αH/αI-loop, and αI-helix (Figure 3.6). Here, we describe the structural interactions observed for these residues in light of their uniqueness relative to other tyrosine kinase structures, however, the precise roles of these conserved residues are not readily apparent from available crystal structures alone.

Eph-specific residues in the C-lobe contribute to several unique structural interactions in the substrate binding region, which includes the αD and αG helices, αF-αG loop, and the activation loop (Figure 3.6, panel A). Two residues that appear to be involved in substrate binding, Arg823 and Glu827, may be connected to the SAM linker tethering network via the αF-αG loop, which contains the SAM linker docking residue, Trp826, and is observed to be in a highly unique conformation relative to the αF-αG loop conformations seen in other tyrosine kinase structures (Figure 3.6, panel A). Many distinguishing Eph-specific residues also broadly map to the N-lobe and hinge region of the kinase domain (Figure 3.6, panel B). The most distinctive residue networks include a charged three residue network in the N-lobe comprising Asp624, Lys645, and Glu686, a hydrophobic network in the N-lobe comprising Met696 and

Val689, and a large network spanning the hinge region to the αD-, αE-, and αF-helices in the C-lobe, including the highly distinctive residues Cys760 and Gly729 (Figure 3.6, panel B).

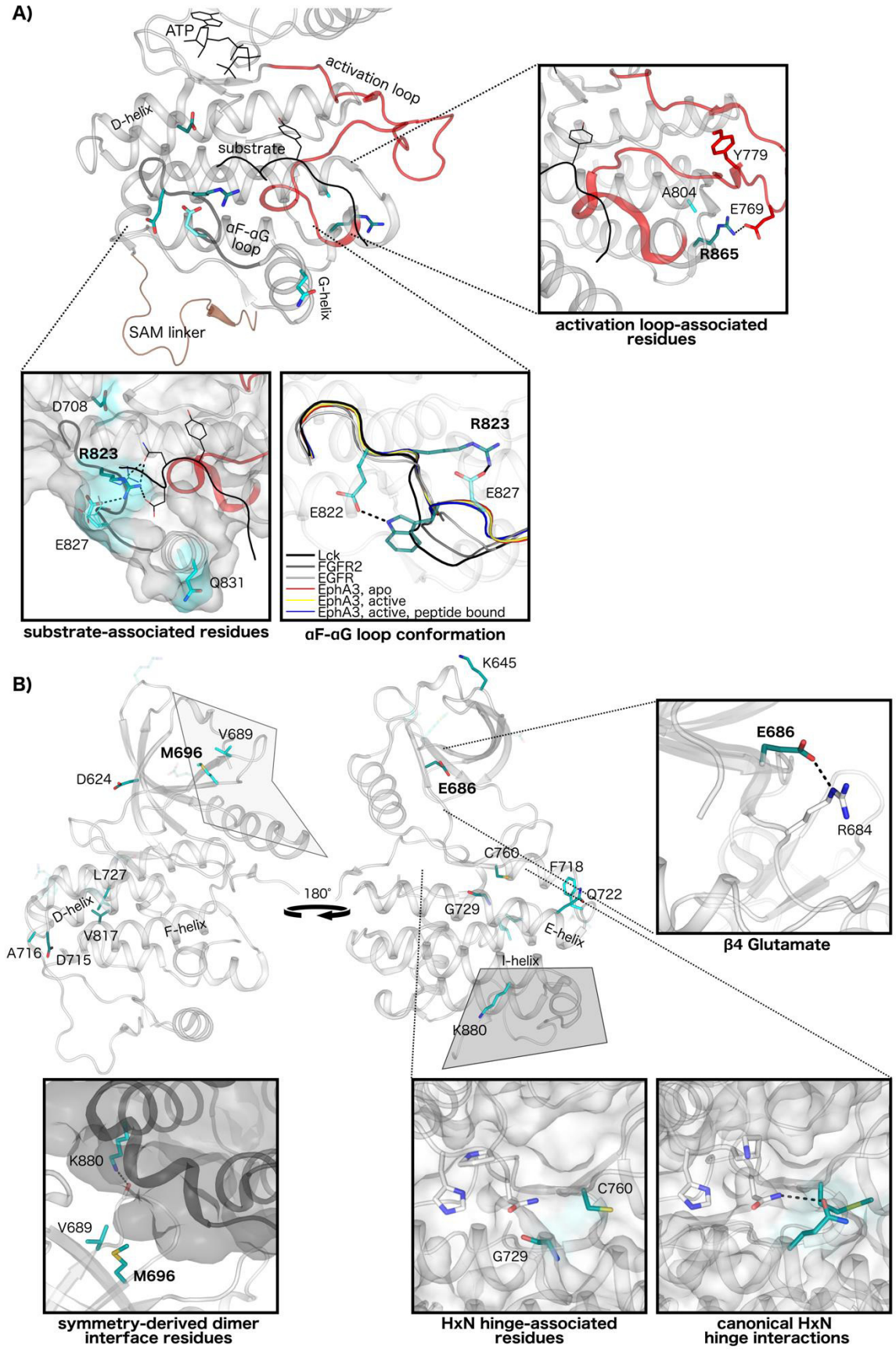
A)
ATP
activation loop
D-helix
substrate
αF-αG loop
G-helix
SAM linker
Y779
A804
E769
R865
activation loop-associated residues
D708
R823
E827
Q831
substrate-associated residues
R823
E827
E822
Lck
FGFR2
EGFR
EphA3, apo
EphA3, active
EphA3, active, peptide bound
αF-αG loop conformation
B)
K645
V689
M696
D624
E686
L727
D-helix
V817
F-helix
A716
D715
180°
C760
F718
Q722
G729
E-helix
I-helix
K880
E686
R684
β4 Glutamate
K880
V689
M696
symmetry-derived dimer interface residues
C760
G729
HxN hinge-associated residues
canonical HxN hinge interactions

Figure 3.6: Structural interactions of Eph-specific residues of unknown function. A) Eph-specific residues contribute to unique interactions in the substrate binding region and activation loop. Key conserved residues in this region are mapped to structures of human EphA3, with darker shading indicating greater sequence constraints at that position. ATP and peptide substrate are shown in black, the activation loop in red, and the SAM domain linker in brown. Interactions of Eph-specific residues and peptide substrate are shown, with different rotamer conformations between peptide-bound (shown as sticks)(PDB: 3FXX) and peptide-unbound structures (shown as lines) (PDB: 2QOC). The unique conformation of the αF-αG loop in Eph structures is shown relative to other tyrosine kinase structures: EphA3, apo (PDB: 2QO2), EphA3, active (PDB:2QOC), EphA3, active, peptide-bound (PDB: 3FXX), Lck (PDB: 3LCK), FGFR2 (PDB: 2PVF), and EGFR (PDB: 2GS2). Interactions between Eph-specific residues R865 and A804 with the activation loop are shown (EphA2, PDB: 4TRL). B) Eph-specific residues in the N-lobe and hinge region. Interactions of the β4-glutamate are shown (PDB:2QOC). The N-lobe-to-C-lobe interface in a symmetry-derived dimer observed in several EphA3 structures is shown with key Eph-specific residues in the interface (PDB: 2QOC) highlighted. K880 of this network forms a hydrogen bond to a backbone residue in the β4-β5 loop in the dimeric partner. Comparisons between canonical HxN hinge network interactions, which are conserved across diverse protein kinases, and the Eph family HxN hinge network are shown from representative structures (FGFR2, PDB:2PVF and EphA3, PDB: 2QOC).

3.2.5.1 Experimental analysis of Eph-specific residues of unknown function

We performed an initial characterization of Eph-specific residues of unknown function by mutating conserved residues in distinct regions of the C-lobe and N-lobe in EphA3 and determining the effects on autophosphorylation and enolase phosphorylation. Mutation of the substrate-associated residue Arg823 to a glutamine (R823Q), which is conserved at the equivalent position in other RTKs (i.e. Ror and DDR families), resulted in faster autophosphorylation relative to WT (Figure 3.7, panel A). Similar to the juxtamembrane and SAM linker tethering mutants (Q676E and W826P), significant levels of autophosphorylation was detected for the R823Q mutant after just 30 seconds of incubation with ATP. Similarly, mutation of the activation loop-associated Arg865 to proline, which is observed in over 60% of other tyrosine kinase sequences, caused an increase in the rate of autophosphorylation relative to WT, though this mutant activated slightly less quickly than the R823Q mutant (significant autophosphorylation levels detected after 1 minute). In the N-lobe, both the β4-strand glutamate and the β5-strand methionine were mutated to leucine (E686L, M696L), which is the most

common residue found at the equivalent positions in other tyrosine kinases. Similar to mutations in the C-lobe, both mutations (E686L, M696L) caused increases in the rate of autophosphorylation with respect to WT (Figure 3.7, panel A). Interestingly, while R823Q, R865P, E686L, and M696L mutants all exhibited increased rates of autophosphorylation, no effects were observed on enolase phosphorylation compared to WT (Figure 3.7, panel B).

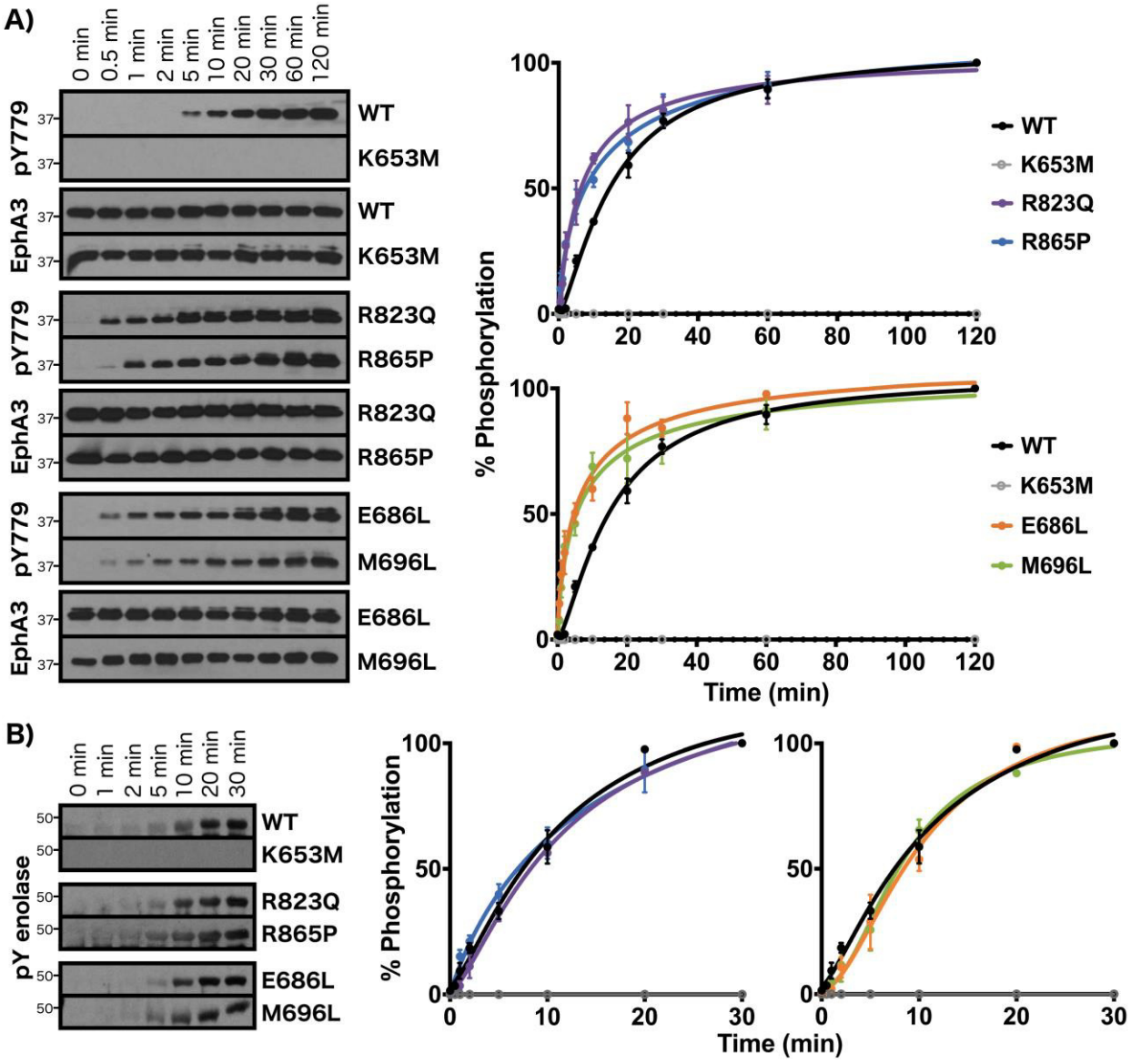

Figure 3.7: Experimental analysis of Eph-specific residues of unknown functions. A) Autophosphorylation time series of R823Q, R865P, E686L, and M696L EphA3 mutants. Phosphorylation levels of the activation loop site were detected (pY779). B) Enolase

phosphorylation time series. Phosphorylation levels in (B) and (C) were quantified by densitometry and normalized against total protein levels. Data for WT and K653M EphA3 autophosphorylation and enolase phosphorylation from Figure 3.4 are shown for comparison. Data shown represent experiments performed in triplicates or more.

3.3 Discussion

Eph kinases have evolutionarily diverged from other tyrosine kinases to respond to specific temporal and cellular signals in signaling pathways. Here, we have identified strikingly conserved sequence features shared among Eph kinases and across extremely diverse evolutionary phyla that tether the flexible juxtamembrane and SAM domain linker to the kinase domain, and we have also discovered novel Eph-specific residues of unknown function. Mutation of the most highly distinguishing Eph-specific residues from various regions of the kinase domain resulted in significantly reduced activation times with negligible effects on enolase phosphorylation. We also show for the first time that the SAM linker and its tethering interactions are essential for the regulation of activation through allosteric coupling of the SAM linker to the critical auto-regulatory juxtamembrane.

Linker regions in multi-domain proteins have been commonly observed in protein kinases and other diverse proteins to play autoregulatory roles (24, 27, 36-38), promote protein interactions (27, 28, 36, 39, 40), and serve as evolutionary hotspots for neo-functionalization (41, 42). The identification of highly distinct Eph-specific residues that tether the N- and C-terminal flanking segments indicates that these flexible segments have evolved to play important and conserved functional roles unique to this family of tyrosine kinases. Mutation of the juxtamembrane tethering GQF motif corroborated the autoinhibitory role of the juxtamembrane, which has previously been characterized in several biochemical and structural studies (8-13). The replacement of the GQF-glutamine with glutamate caused faster activation, indicating that the glutamine residue and its electrostatic interactions with the juxtamembrane backbone are

indispensable for proper autoinhibition. Though the role of the glutamine has been noted in previous studies (11), we identify for the first time the GQF-glycine as a critical part of the juxtamembrane network. Despite it being the most uniquely conserved residue of the network, we identified an alanine variant of this residue in non-mammalian, vertebrate EphA8 sequences (Figure 3.1), We observed that producing this variation in EphA3 (G675A) resulted in a similar activation time to WT EphA3 (Figure 3.4). These results reflect that the AQF variant is well tolerated with respect to auto-regulatory function, however, this variation could reflect functional specialization of mechanisms other than autophosphorylation. We note that the GQF-glycine is also a recurrent hotspot for cancer-associated variants. The glycine has been observed to be mutated in 7 different cancer samples across three different Eph members (EphA7, EphA8, and EphB3) to arginine, glutamate, valine, and cysteine (SI Dataset 1).

In addition, we investigated the unique network of Eph-specific residues that tether the C-terminal SAM domain linker to the C-lobe of the kinase domain. Although the functional relevance of the SAM linker has remained elusive, the unique conservation of residues that tether the linker reflects an evolutionary pressure to maintain the SAM linker tethering pocket and indicates that the linker and its interaction with the kinase domain play an important role. Through mutational analysis of the SAM linker network, we discovered that the SAM domain linker is important for activation via autophosphorylation but dispensable for enolase phosphorylation. Mutation of the SAM linker tethering tryptophan residue, Trp826, which is the most distinguishing residue of the Eph family, significantly increased the speed of activation. Structurally, we expect that the W826P mutation would cause ineffective tethering of the SAM linker by reorienting the αF-αG loop portion of the interface and by widening the tethering pocket on the basal face of the C-lobe. The weaker tethering of the SAM linker via the W826P

mutation is corroborated by our thermal stability studies, which showed significantly decreased global stability of the W826P mutant with respect to WT (Figure 3.5), and by our molecular dynamics studies, which showed increased SAM linker dynamics in the W826P mutant. Further studies of the SAM linker network via a SAM linker deletion mutant surprisingly showed that the ΔSAMlinker mutant exhibited diminished autophosphorylation activity. Thermal stability analysis of this mutant showed increased global stability compared to WT, which suggests that the SAM linker, when present, contributes to the general flexibility of the WT protein. Interestingly, we note that the SAM linker tethering tryptophan is substituted with glycine in EphA1 sequences (Figure 3.1), which reflects a drastic variation of the SAM linker network. We predict that this variation would have dramatic effects on Eph-specific interactions in the αF-αG loop and would prevent effective tethering of the SAM linker. Given the effects of the W826P mutation on EphA3 activation, this variation may allow EphA1 kinases to autophosphorylate and activate more rapidly than other Eph members. However, the possible functional consequences of this variation and whether there are EphA1-specific variations that compensate for the tryptophan-to-glycine substitution warrant further investigation. Furthermore, the tryptophan in EphA7 is observed to be mutated to leucine in two cancer samples (SI Dataset 1). Whether this disease variant of EphA7 contributes to abnormal signaling and/or cancer progression should also be explored.

The strikingly faster activation by the W826P mutant was unexpected given the distal location of the SAM linker network to the activation loop or to the juxtamembrane, which to this point have been the only identified autoinhibitory regulators of Eph activity. We confirmed through mass spectrometry studies that the effects of the W826P mutation on activation were directly linked to changes in the rate of juxtamembrane autophosphorylation. In addition, by

examining the effects of coupled mutations between the juxtamembrane and SAM linker networks, we find that the SAM linker network adds an additional layer of regulation to the autoinhibitory juxtamembrane to regulate autophosphorylation. Specifically, the W826P mutation was able to rescue activity when added to the nearly inactive YFYF juxtamembrane mutant, which demonstrates that this mutation is able to communicate to the distal juxtamembrane and disrupt its autoinhibitory function despite the lack of structural and biochemical changes induced by phosphorylation. In light of these results, we propose a model where the C-terminal SAM domain linker plays an essential and cooperative role in disengaging the conserved juxtamembrane autoinhibitory mechanism during activation (Figure 3.8). Given the faster activation by the W826P mutant, we believe that the untethered, flexible SAM domain linker is able to allosterically disrupt the juxtamembrane from its autoinhibitory conformation, as supported by our molecular dynamics studies, and cause the autophosphorylation of the juxtamembrane to occur more readily. We also conclude that the SAM domain linker is essential for juxtamembrane autophosphorylation and kinase domain activation to occur because the SAM domain deletion mutant exhibited decreased autophosphorylation efficiency. The co-conservation of the coupled SAM linker and juxtamembrane networks highlights the evolution of an allosteric communication network that employs both flanking segments to regulate the Eph kinase domain.

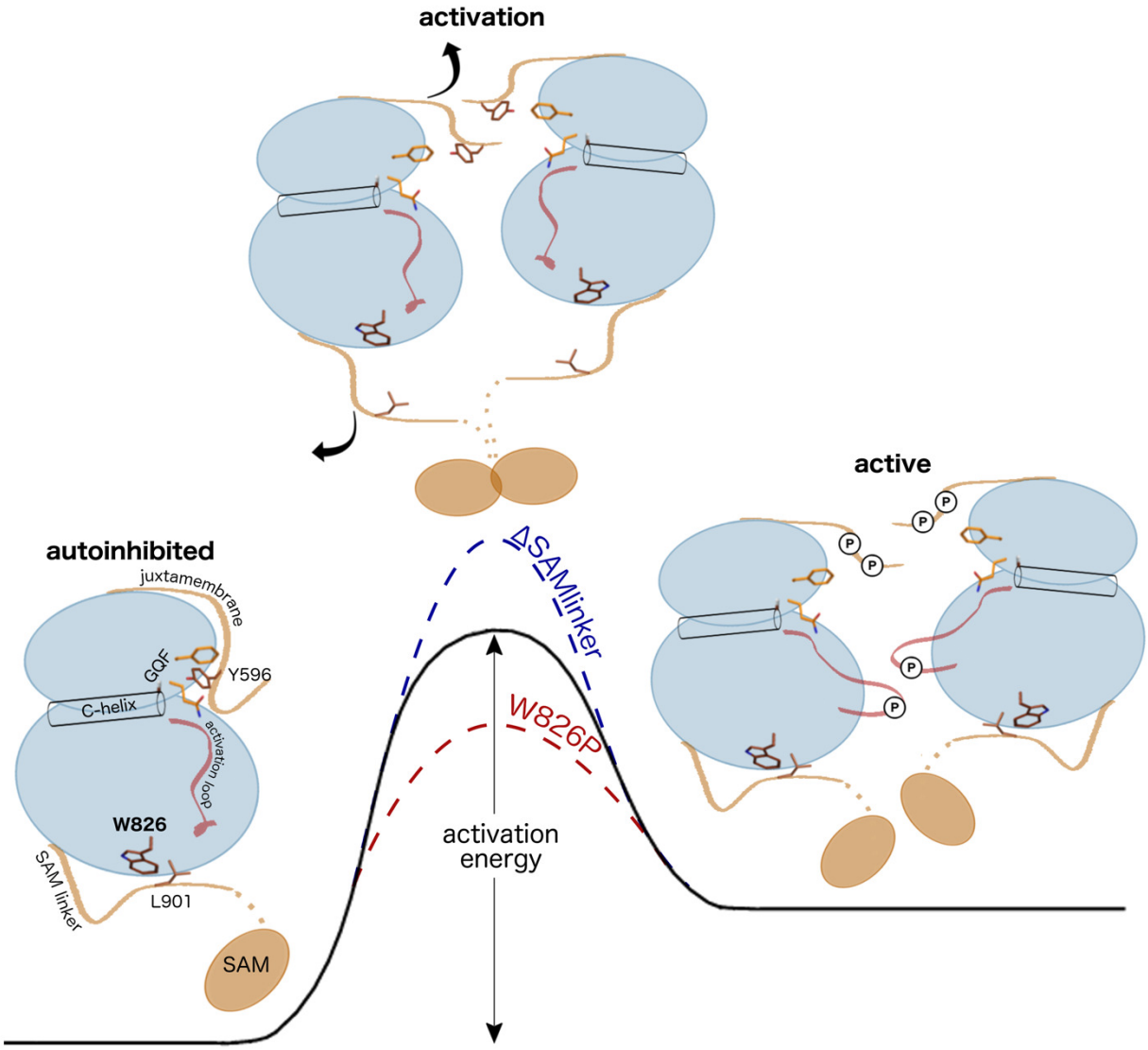

Figure 3.8: Proposed energetic model for Eph activation by the juxtamembrane and SAM domain linker. The initial step of Eph activation occurs with the displacement of the juxtamembrane from its autoinhibitory conformation. Displacement of the autoinhibitory conformation occurs with an increase in entropy and is facilitated by 1) autophosphorylation of the juxtamembrane and/or 2) untethering of the SAM domain linker. This allows for further autophosphorylation to occur (*i.e.* autophosphorylation on the activation loop shown in red) to complete the activation process. SAM domain network mutants affect the dynamics of the autoinhibitory juxtamembrane, altering the activation energy required for activation.

Conserved allosteric communication networks have been observed in many protein kinases that couple distal protein regions and molecular events such as substrate-binding, post-translational modifications, and protein interactions to the active site in ways that impact

catalytic activity (25, 27, 29, 43-46). We believe that the SAM linker comprises part of an allosteric communication network in Eph kinases, which allows the coupling of other important functional regions to the SAM linker and juxtamembrane. For example, oligomerization by the SAM domain may trigger the untethering of the SAM domain linker from the kinase domain and subsequently prime the juxtamembrane for autophosphorylation. Furthermore, as suggested by our molecular dynamics analysis (Figure 3.5, panel C), the untethering of the SAM linker may allosterically affect the conformation of other functionally important regions such as the activation loop or substrate-binding region through other Eph-specific residues, such as R865 and R823 (Figure 3.6). Though it was unexpected that the R823Q mutation did not negatively affect autophosphorylation or substrate phosphorylation despite its location in the substrate binding pocket, whether this result is attributed to changes in substrate recognition or other conformational changes remains unknown. It is also worth noting that Eph-specific residues, most of which are located on the protein surface distal to known juxtamembrane-, SAM linker-, or active site-associated regions, may serve as protein-protein interaction sites that can communicate protein interaction events to regulatory networks within the kinase domain. In light of this view, we have observed Eph-specific residues that comprise the interface in a symmetry-related dimer, though the biological relevance of this dimer remains unknown. We also note that the same position of the β4-glutamate, which is the second-most distinguishing Eph-specific residue, has been observed in other tyrosine kinases such as Abl and Src to be important in regulatory inter- and intramolecular interactions (47, 48); this common docking site may have been refashioned as a unique regulatory interaction site in Eph kinases as well. In this study, mutation of Eph-specific residues resulted in faster activation in all but one case (G675A), therefore, further investigating how these residues may allosterically affect catalysis and

regulation, and how they may be involved in protein interactions and pathways will be important directions for future studies.

3.4 Materials and Methods

3.4.1 Identification of Eph-specific sequence features

We used a dataset of 53,568 tyrosine protein kinase sequences identified from the NCBI-nr protein database using curated hierarchical sequence profiles generated using the MAPGAPS suite (49). The 53,568 sequences were aligned using MAPGAPS, and this alignment was used as an input for mcBPPS (35). N-terminal juxtamembrane and C-terminal SAM linker and SAM domain sequences were identified using Pfam domain (50) and TMHMM transmembrane (51) predictions, then aligned using Clustal Omega (52).

3.4.2 Expression and purification of EphA3 WT and mutant proteins

WT and mutant EphA3 proteins were expressed and purified as described previously (25). The EphA3 construct used in the study corresponds to residues Asp577-Ser947 with an N-terminal 6X His tag. Additional details are provided as Supporting Information.

3.4.3 Autophosphorylation assays

To obtain starting samples for autophosphorylation assays, we produced inactive, dephosphorylated EphA3 by incubating 25-50 mg of purified protein samples with 100 units of CIP (New England Biolabs) overnight at 4°C. The CIP was removed by purification on Ni-NTA agarose columns. The complete dephosphorylation and removal of CIP was verified using western blotting with anti-pY779 (Cell Signaling #8862) and anti-CIP (Fitzgerald 20C-CR2110RP) antibodies. For autophosphorylation reactions, reactions were set up with a final

reaction volume of 20 μL and with final concentrations of the following: 0.375 mg/mL EphA3, 5 mM ATP, and 10 mM MgCl2. To quench the reaction at each time point, 1 μL of the reaction was diluted into 74 μL of SDS-PAGE sample buffer, then the sample was boiled for 5 minutes. Additional details of western blotting methods are provided as Supporting Information.

3.4.4 Enolase phosphorylation assays

To obtain starting samples for enolase phosphorylation assays, we produced fully active EphA3 by incubating 25-50 mg of purified protein samples with ATP and MgCl2 at final concentrations of 10 mM and 20 mM, respectively, for 6 hours at 4°C. Excess MgATP/ADP was dialyzed out overnight. For enolase phosphorylation reactions, rabbit muscle enolase (Sigma) was denatured in 25 mM AcOH for 15 minutes. Reactions were set up with a final reaction volume of 250 μL containing 70 μg of denatured enolase, and final concentrations of the following: 0.1 mg/mL EphA3, 50 μM ATP, and 25 mM MgCl2. To quench the reaction at each time point, 30 μL of the reaction was mixed into 10 μL of SDS-PAGE sample buffer, then the sample was boiled for 5 minutes.

3.4.5 Western blotting

Samples were run on 10% SDS-PAGE and transferred to PVDF membranes. EphA3 phosphorylation levels were detected by western blotting using anti-pY779 antibody (Cell Signaling #8862) and total EphA3 levels were detected using anti-EphA3 antibody (Cell Signaling #8793). Enolase phosphorylation levels were detected by western blotting using anti-pY100 antibody (Cell Signaling #9411)

3.4.6 Thermal melt assays

Thermal melt assays were performed as described previously (25). In brief, thermal melt assays were performed by quantifying SYPRO orange dye (Sigma) fluorescence as a reporter for protein unfolding. Reactions were set up with a final volume of 100 μL, with final EphA3 concentrations of 0.05-0.1 mg/mL. SYPRO orange dye was used at a final dilution of 1:8000. These assays were performed in a Synergy H4 microplate Reader with increasing temperature from 25°C-65°C with a step size of 2°C. Fluorescence was measured with excitation at 470nm and emission at 570nm. For all experiments, Tm values were calculated by fitting data to a Boltzmann sigmoidal curve.

3.4.7 Mass spectrometry analysis of autophosphorylated samples

WT and W826P 30s and 10 min autophosphorylation samples were analyzed by in-gel trypsin digestion followed by peptide mass fingerprinting by MALDI-TOF mass spectrometry. The W826P 30s sample was further analyzed by LC-MS/MS, CID, and HCD MS/MS. Further details are provided as Supporting Information. WT and W826P 30s and 10 min autophosphorylation samples were analyzed by in-gel trypsin digestion followed by peptide mass fingerprinting by MALDI-TOF mass spectrometry. The W826P 30s sample was further analyzed by LC-MS/MS, CID, and HCD MS/MS. Further details are provided as Supporting Information.

3.4.8 Molecular dynamics simulations

All-atom MD simulations were performed using GROMACS (5.0.2)(53). Structural models of autoinhibited WT and W826P EphA3 were generated using Rosetta (3.8) (54) and Modeller (9.19)(55) and used as starting structures. Simulation data was acquired for 500 ns. Analysis of MD trajectories was done using packages from the Gromacs suite, and the cross-correlation

matrices of Cα-Cα motions were generated using Prody (56). Further details are provided as

Supporting Information

Bibliography

1. Kullander K & Klein R (2002) Mechanisms and functions of Eph and ephrin signalling. *Nat Rev Mol Cell Biol* 3(7):475-486.

2. Pasquale EB (2005) Eph receptor signalling casts a wide net on cell behaviour. *Nat Rev Mol Cell Biol* 6(6):462-475.

3. Murphy JM*, et al.* (2014) A robust methodology to subclassify pseudokinases based on their nucleotide-binding properties. *Biochem J* 457(2):323-334.

4. Zeqiraj E & van Aalten DM (2010) Pseudokinases-remnants of evolution or key allosteric regulators? *Curr Opin Struct Biol* 20(6):772-781.

5. Boudeau J, Miranda-Saavedra D, Barton GJ, & Alessi DR (2006) Emerging roles of pseudokinases. *Trends Cell Biol* 16(9):443-452.

6. Barquilla A & Pasquale EB (2015) Eph receptors and ephrins: therapeutic opportunities. *Annu Rev Pharmacol Toxicol* 55:465-487.

7. Himanen JP*, et al.* (2001) Crystal structure of an Eph receptor-ephrin complex. *Nature* 414(6866):933-938.

8. Singla N, Erdjument-Bromage H, Himanen JP, Muir TW, & Nikolov DB (2011) A semisynthetic Eph receptor tyrosine kinase provides insight into ligand-induced kinase activation. *Chem Biol* 18(3):361-371.

9. Davis TL*, et al.* (2008) Autoregulation by the juxtamembrane region of the human ephrin receptor tyrosine kinase A3 (EphA3). *Structure* 16(6):873-884.

10. Binns KL, Taylor PP, Sicheri F, Pawson T, & Holland SJ (2000) Phosphorylation of tyrosine residues in the kinase domain and juxtamembrane region regulates the biological and catalytic activities of Eph receptors. *Mol Cell Biol* 20(13):4791-4805.

11. Wybenga-Groot LE*, et al.* (2001) Structural basis for autoinhibition of the Ephb2 receptor tyrosine kinase by the unphosphorylated juxtamembrane region. *Cell* 106(6):745-757.

12. Shi G, Yue G, & Zhou R (2010) EphA3 functions are regulated by collaborating phosphotyrosine residues. *Cell Res* 20(11):1263-1275.

13. Wiesner S, *et al.* (2006) A change in conformational dynamics underlies the activation of Eph receptor tyrosine kinases. *EMBO J* 25(19):4686-4696.

14. Adams JA (2003) Activation Loop Phosphorylation and Catalysis in Protein Kinases: Is There Functional Evidence for the Autoinhibitor Model? *Biochemistry* 42(3):601-607.

15. Zisch AH, *et al.* (2000) Replacing two conserved tyrosines of the EphB2 receptor with glutamic acid prevents binding of SH2 domains without abrogating kinase activity and biological responses. *Oncogene* 19(2):177-187.

16. Lawrenson ID, *et al.* (2002) Ephrin-A5 induces rounding, blebbing and de-adhesion of EphA3-expressing 293T and melanoma cells by CrkII and Rho-mediated signalling. *J Cell Sci* 115(Pt 5):1059-1072.

17. Holland SJ, *et al.* (1997) Juxtamembrane tyrosine residues couple the Eph family receptor EphB2/Nuk to specific SH2 domain proteins in neuronal cells. *EMBO J* 16(13):3877-3888.

18. 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.

19. Huang Y, *et al.* (2016) Molecular basis for multimerization in the activation of the epidermal growth factor receptor. *Elife* 5.

20. Thanos CD, Goodwill KE, & Bowie JU (1999) Oligomeric structure of the human EphB2 receptor SAM domain. *Science* 283(5403):833-836.

21. Stapleton D, Balan I, Pawson T, & Sicheri F (1999) The crystal structure of an Eph receptor SAM domain reveals a mechanism for modular dimerization. *Nat Struct Biol* 6(1):44-49.

22. Singh DR, *et al.* (2015) Unliganded EphA3 dimerization promoted by the SAM domain. *Biochem J* 471(1):101-109.

23. Singh DR, *et al.* (2017) The SAM domain inhibits EphA2 interactions in the plasma membrane. *Biochim Biophys Acta* 1864(1):31-38.

24. 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.

25. Mohanty S, *et al.* (2016) Hydrophobic Core Variations Provide a Structural Framework for Tyrosine Kinase Evolution and Functional Specialization. *PLoS Genet* 12(2):e1005885.

26. 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.

27. Kannan N, Haste N, Taylor SS, & Neuwald AF (2007) The hallmark of AGC kinase functional divergence is its C-terminal tail, a cis-acting regulatory module. *Proceedings of the National Academy of Sciences of the United States of America* 104(4):1272-1277.

28. Eyers PA, Keeshan K, & Kannan N (2017) Tribbles in the 21st Century: The Evolving Roles of Tribbles Pseudokinases in Biology and Disease. *Trends Cell Biol* 27(4):284-298.

29. Nguyen T, Ruan Z, Oruganty K, & Kannan N (2015) Co-conserved MAPK features couple D-domain docking groove to distal allosteric sites via the C-terminal flanking tail. *PLoS One* 10(3):e0119636.

30. Brunet FG, Volff JN, & Schartl M (2016) Whole Genome Duplications Shaped the Receptor Tyrosine Kinase Repertoire of Jawed Vertebrates. *Genome Biol Evol* 8(5):1600-1613.

31. Fairclough SR*, et al.* (2013) Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate Salpingoeca rosetta. *Genome Biol* 14(2):R15.

32. Srivastava M*, et al.* (2010) The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature* 466(7307):720-726.

33. Drescher U (2002) Eph family functions from an evolutionary perspective. *Curr Opin Genet Dev* 12(4):397-402.

34. Knighton DR*, et al.* (1991) Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* 253(5018):407-414.

35. Neuwald AF (2010) Bayesian classification of residues associated with protein functional divergence: Arf and Arf-like GTPases. *Biol Direct* 5:66.

36. Dyson HJ & Wright PE (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6(3):197-208.

37. Hubbard SR (2004) Juxtamembrane autoinhibition in receptor tyrosine kinases. *Nat Rev Mol Cell Biol* 5(6):464-471.

38. Plaza-Menacho I*, et al.* (2016) RET Functions as a Dual-Specificity Kinase that Requires Allosteric Inputs from Juxtamembrane Elements. *Cell Rep* 17(12):3319-3332.

39. Gajiwala KS (2013) EGFR: tale of the C-terminal tail. *Protein Sci* 22(7):995-999.

40. Ma B, Tsai CJ, Haliloglu T, & Nussinov R (2011) Dynamic allostery: linkers are not merely flexible. *Structure* 19(7):907-917.

41. Brown CJ, Johnson AK, Dunker AK, & Daughdrill GW (2011) Evolution and disorder. *Curr Opin Struct Biol* 21(3):441-446.

42. Harrison SC (2003) Variation on an Src-like theme. *Cell* 112(6):737-740.

43. Kornev AP & Taylor SS (2015) Dynamics-Driven Allostery in Protein Kinases. *Trends Biochem Sci* 40(11):628-647.

44. Laine E, Auclair C, & Tchertanov L (2012) Allosteric communication across the native and mutated KIT receptor tyrosine kinase. *PLoS Comput Biol* 8(8):e1002661.

45. McClendon CL, Kornev AP, Gilson MK, & Taylor SS (2014) Dynamic architecture of a protein kinase. *Proc Natl Acad Sci U S A* 111(43):E4623-4631.

46. Schulze JO*, et al.* (2016) Bidirectional Allosteric Communication between the ATP-Binding Site and the Regulatory PIF Pocket in PDK1 Protein Kinase. *Cell Chem Biol* 23(10):1193-1205.

47. Chen S, Brier S, Smithgall T, & Engen J (2007) The Abl SH2-kinase linker naturally adopts a conformation competent for SH3 domain binding. *Protein Sci* 16(4):572-581.

48. Xi G, Shen X, & Clemmons DR (2010) p66shc inhibits insulin-like growth factor-I signaling via direct binding to Src through its polyproline and Src homology 2 domains, resulting in impairment of Src kinase activation. *J Biol Chem* 285(10):6937-6951.

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

50. Bateman A*, et al.* (2004) The Pfam protein families database. *Nucleic Acids Res* 32(Database issue):D138-141.

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

52. Sievers F*, et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539.

53. Abraham MJ*, et al.* (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1-2:19-25.

54. Rohl CA, Strauss CEM, Misura KMS, & Baker D (2004) Protein structure prediction using rosetta. *Numerical Computer Methods, Pt D* 383:66-+.

55. Fiser A, Do RK, & Sali A (2000) Modeling of loops in protein structures. *Protein Sci* 9(9):1753-1773.

56. Bakan A, Meireles LM, & Bahar I (2011) ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics* 27(11):1575-1577.

Chapter 4

Tracing the origin and evolution of pseudokinases across the tree of life

Accepted by *Science Signaling*. In print. Annie Kwon, et al. (2019)
Reprinted here with permission from the publisher.

Abstract

Protein phosphorylation by eukaryotic protein kinases (ePKs) represents a fundamental mechanism of cell signaling in all organisms. In model vertebrates, ~10% of ePKs are classified as pseudokinases, which possess amino acid changes within the catalytic machinery of the kinase domain that distinguish them from their canonical kinase counterparts. However, pseudokinases still regulate a wide variety of signaling pathways, usually doing so in the absence of their own catalytic output. To investigate the prevalence, evolutionary relationships, and biological diversity of these pseudoenzymes, we performed a comprehensive analysis of putative pseudokinase sequences in available eukaryotic, bacterial, and archaeal proteomes. We found that pseudokinases are present across all domains of life, and we classified nearly 30,000 eukaryotic, 1,500 bacterial, and 20 archaeal pseudokinase sequences into 86 pseudokinase families, including ~30 families that were previously unknown. We uncovered a rich variety of pseudokinases with notable expansions not only in animals, but also in plants, fungi, and bacteria, where pseudokinases have previously received cursory attention. These expansions are accompanied by domain shuffling, which suggests roles for pseudokinases in plant innate immunity, plant-fungal interactions, and bacterial signaling. Mechanistically, the ancestral kinase fold has diverged in many distinct ways through the enrichment of unique sequence motifs to generate new families of pseudokinases in which the kinase domain is repurposed for non-canonical nucleotide binding or to stabilize unique, inactive kinase conformations. We provide a classified collection of annotated pseudokinase sequences in the Protein Kinase Ontology (ProKinO) as a new minable resource for the signaling community.

4.1 Introduction

Protein phosphorylation catalyzed by eukaryotic protein kinases (ePKs) controls multiple aspects of prokaryotic and eukaryotic-based cell signaling (1, 2), and its dysregulation contributes to many major diseases. The conserved architecture of the ePK domain is very well understood from both structural (3-5) and biochemical (6-8) perspectives, and the versatility of the kinase fold has been exploited many times during evolution to impart mechanistic control over diverse cell signaling processes (9, 10). A vast amount of genomic and proteomic datasets can now be mined to map the evolution of kinases and their associated signaling pathways across multiple species (11-17). In this context, some 10% of model vertebrate ePKs contain amino acid changes at specific positions that are predicted to lead to catalytic inactivation, which led to the coining of the term 'pseudokinase' (5, 15, 18-21). A number of well-studied pseudokinases are thought to play central roles in signaling despite impaired catalytic function (22-26), for example through allosteric modulation of other active kinases or the transduction of cellular signals via dynamic scaffolding functions (9, 19, 21, 27-30). However, whether pseudokinases have evolved to control fundamental aspects of signaling across all organisms has never been scrutinized in depth, and much remains to be understood about the origin of pseudokinases and how they became embedded in signaling networks during prokaryotic and eukaryotic evolution.

Protein pseudokinases represent the best understood members of the growing classes of pseudoenzymes, which include pseudophosphatases (31) and pseudoproteases (32), both of which are also predicted to have lost canonical catalytic function, but nonetheless perform critical non-enzymatic roles (9, 20, 33, 34). By definition, pseudokinases lack canonical phosphotransferase activity, and they can be predicted bioinformatically by identifying sequences that lack at least one key amino acid normally required for metal and ATP binding and

for catalysis (3, 7, 8, 18-20). Prominent catalytic motifs include the 'catalytic triad' residues, comprised of the ATP-binding β3-lysine, the catalytic aspartate within the catalytic loop His-Arg-Asp-X-X-X-X-Gln (HRDXXXXN) motif, and the metal binding aspartate of the activation loop Asp-Phe-Gly (DFG) motif. Some examples of human pseudokinases with variations at these catalytic triad residues are summarized in Table 4.1. Notably, loss of these canonical residues does not always abolish nucleotide-binding or phosphoryl transfer, and in some cases residual kinase activity or ATP binding may fulfill a unique functional role. However, we still define these catalytically-competent proteins as pseudokinases, in recognition of their non-canonical amino acid composition. For example, in the human epidermal growth factor receptor (EGFR)-related receptor pseudokinase HER3, where the catalytic triad is conserved except for the substitution of the catalytic HRD-aspartate for asparagine, low amounts of catalytic activity support HER3 trans-autophosphorylation in vitro, although this vestigial activity is insufficient for phosphorylation of exogenous substrates in cells (22, 35, 36). In other cases, degenerated catalytic residues can be compensated by similar amino acids found elsewhere in the active site to rescue catalytic function. This is best illustrated by the With No Lysine kinases (WNKs), which lack the canonical β3-lysine, but maintain ATP-binding and catalytic activity via a conserved compensatory lysine in the glycine rich loop (37-39). Less predictably, pseudokinases contain co-evolving amino acids that are far-removed from the active site and contribute critical non-catalytic signaling roles, as described recently for the Tribbles (TRIB) family of pseudokinases, where a co-evolved C-terminal tail docking site in the pseudokinase domain negatively regulates binding of the E3 ubiquitin-protein ligase COP1 (26 , 40-42). Finally, amino acid shuffling offers new biochemical opportunities as described recently for the

atypical selenoprotein-O (SelO) pseudokinase in which dramatic active site variations facilitate "inverted" ATP binding to support protein AMPylation instead of phosphorylation (43, 44).

Human pseudokinase	Degraded catalytic residue(s)	Observed residue(s)
KSR1, KSR2	K	R
WNK1, WNK2, WNK3, WNK4	K	C
HER3	HRD-D	N
JAK1 (domain2)	HRD-D	N
JAK2 (domain2)	HRD-D	N
ILK	HRD-D	A
TRIB3	DFG-D	N
TRIB2	DFG-D	S
CASK	DFG-D	G
GCN2 (domain2)	K, HRD-D	Y, V
ULK4	K, DFG-D	L, N
VRK3	HRD-D, DFG-D	N, G
MLKL	HRD-D, DFG-D	K, G
STRADB (STLK6)	HRD-D, DFG-D	S, G
EphB6	K, HRD-D, DFG-D	Q, S, R
SCYL1, SCYL2, SCYL3	K, HRD-D, DFG-D	F, N, G
NRBP1, NRBP2	K, HRD-D, DFG-D	N, N, S

Table 4.1. Examples of human pseudokinases. Degraded catalytic triad residues and the amino acids that replace them in each pseudokinase are noted.

Approximately 50 protein pseudokinases are encoded by the human genome, nearly all of which are also found in rodents, suggesting conserved vertebrate-wide signaling roles (5, 15, 18). Half of vertebrate pseudokinases also have clear orthologues in well-characterized genetic model organisms, including flies and worms, supporting the assumption that pseudokinases are part of ancient genetic lineages, rather than extraneous remnants of evolutionary 'experiments' (13). Several pseudokinases have been analyzed in depth in human cells, including HER3 (23, 35 , 45, 46), the RAF/MEK modulators kinase suppressor of Ras 1 and 2 (KSR1/2) (24), the Janus tyrosine kinases (JAKs) (25), which contain a disease-associated pseudokinase domain positioned adjacent to an active kinase domain, and the Tribbles (TRIB) family of pseudokinases

(26, 40, 47). However, over half of the predicted human pseudokinome remains understudied at the molecular level, despite clear evidence for expression in cells. Pseudokinase-based signaling has also been described in simple model organisms (48), such as in the small genome of the intestinal parasite Giardia lamblia (16) and in commercially important plants (49, 50, 51 , 52, 53). However, the origin and evolution of pseudokinases across the tree of life has not been explored in any depth.

In this resource, we present a systematic identification of predicted pseudokinase sequences, ranging from archaea and bacteria through to simple eukaryotes, fungi, plants, and vertebrates. Based on the well understood catalytic machinery in canonical ePKs (6), we find that ePK-like pseudokinase sequences can even be detected in some archaeal and bacterial proteomes, though they are much rarer than in eukaryotic proteomes, where they appear to be ubiquitous. Corroborating previous kinome studies, we also find that the number of pseudokinases remains relatively constant among vertebrates and correlates with the relative size of the kinome. Indeed, our broad analysis permits us to establish that ~10% of ePK members should be classified as pseudokinases in swathes of vertebrate animal species. In several phyla, specific pseudokinase expansions linked to lifestyle are observed within the different kinase families, whose shared sequence signatures and domain structures permit specific functions to be deciphered. In particular, we note the expansions of interleukin-1 receptor-associated kinase (IRAK)-like pseudokinases in plants, increases of tyrosine kinase-like (TKL) pseudokinases in fungi, and a diversified family of PknB pseudokinases in bacteria. Notably, most pseudokinases exhibit lineage-specific sequence variations that might facilitate previously unknown modes of ATP binding, unusual catalytic outputs, and/or allosteric coupling between distal protein binding and regulatory sites. As such, pseudokinases cannot be remnants of evolution, but must instead

operate as fundamental, and function-specific, signaling proteins across organisms covering some 4 billion years of evolution. Our analysis includes a minable, comprehensive classification of pseudokinase sequences from diverse organisms, providing a conceptual starting point for future hypothesis-driven characterization of pseudokinase signaling from bacteria to humans.

4.2 Results

4.2.1 Identification of pseudokinomes across the domains of life

To detect the prevalence of pseudokinases across the domains of life, we used curated multiple sequence alignment profiles of 592 protein kinase families (52 , 54-58) to scan all available eukaryotic, bacterial, and archaeal proteomes in the UniProt reference proteome database (10,092 proteomes in release 2018_9) (59). Aligned sequences that lacked one or more of the canonical catalytic triad residues, namely the β3-lysine residue, the HRD-aspartate, or the DFG-aspartate, were classified as pseudokinase sequences. Such pseudokinase sequences with non-canonical residues at the three catalytic triad positions were detected in 100% of eukaryotic proteomes analyzed, whereas only 5.8% and 2.5% of bacterial and archaeal proteomes, respectively, contained putative pseudokinase sequences (Table 4.2). The prevalence of kinases and pseudokinases across 93 diverse representative species throughout the domains of life is summarized in Figures 4.1 and 4.2. Consistent with previous studies, we identified 55 pseudokinases in the human kinome (Figure 4.1), including the pseudopodium-enriched atypical kinase 3 (PEAK3), which was recently reported as a pseudokinase sharing similarity to the pragmin (PRAG1) and PEAK1/Sgk269 pseudokinases (30, 60). We also identified a previously unidentified putative human pseudokinase (A0A1B0GUL7) that is homologous to the dual-specificity mitogen-activated kinase kinase (MEK2) and possesses a pseudokinase ortholog in

chimpanzee (Pan troglodytes). Other previously studied pseudokinases such as transformation/transcription domain-associated protein (TRRAP) (19, 61), SelO (43 , 44), and family with sequence similarity 20 A (Fam20A) (29) are also considered pseudokinases; however, these atypical kinases and other small-molecule kinases such as aminoglycoside kinases and lipid kinases were not considered in our analysis because they are markedly divergent from ePKs and cannot be reliably placed within the ePK evolutionary framework (further described in the Materials and Methods).

	Archaea	Bacteria	Eukaryota
Total # proteomes	441	8818	833
Proteomes with kinases	149 (33.8%)	3523 (40.0%)	833 (100%)
Proteomes with pseudokinases	11 (2.49%)	508 (5.76%)	819 (98.3%)
Total # pseudokinase sequences	20	1386	30049

Table 4.2. Detection of protein kinase and pseudokinase sequences across archaeal, bacterial, and eukaryotic proteomes available from the UniProt database (*59*). Protein kinase sequences were identified and aligned from each reference proteome using diverse sequence profiles of ePKs, and pseudokinase sequences were identified based on their lack of at least one residue in the catalytic triad (N=1).

Pseudokinase complements of kinomes (hereafter referred to as pseudokinomes) are nearly always proportional to the size of the kinome in vertebrate species (Figure 4.1), which is consistent with previous estimates that pseudokinases generally account for ~10% of kinome content (18). However, both kinome and pseudokinome sizes are much more diverse across other (non-vertebrate) eukaryotic clades. For example, pseudokinase sequences account for between 8-17% of plant kinomes, which are often drastically expanded in size when compared to metazoan kinomes. Moreover, a mycorrhizal fungal species, Rhizophagus irregularis, and two protist species, Paramecium tetraurelia and Tetrahymena thermophila, have substantially expanded kinomes respective to other fungi and protists, and the pseudokinomes of R. irregularis

and T. thermophila are also markedly expanded, comprising 32% and 25% of each kinome respectively. We also note a remarkable expansion of pseudokinases in eukaryotic pathogens, including Plasmodium falciparum and Giardia lamblia, which possess relatively small kinomes (16, 62), but contain highly expanded pseudokinomes that account for more than a half of their kinomes (Figure 4.1).

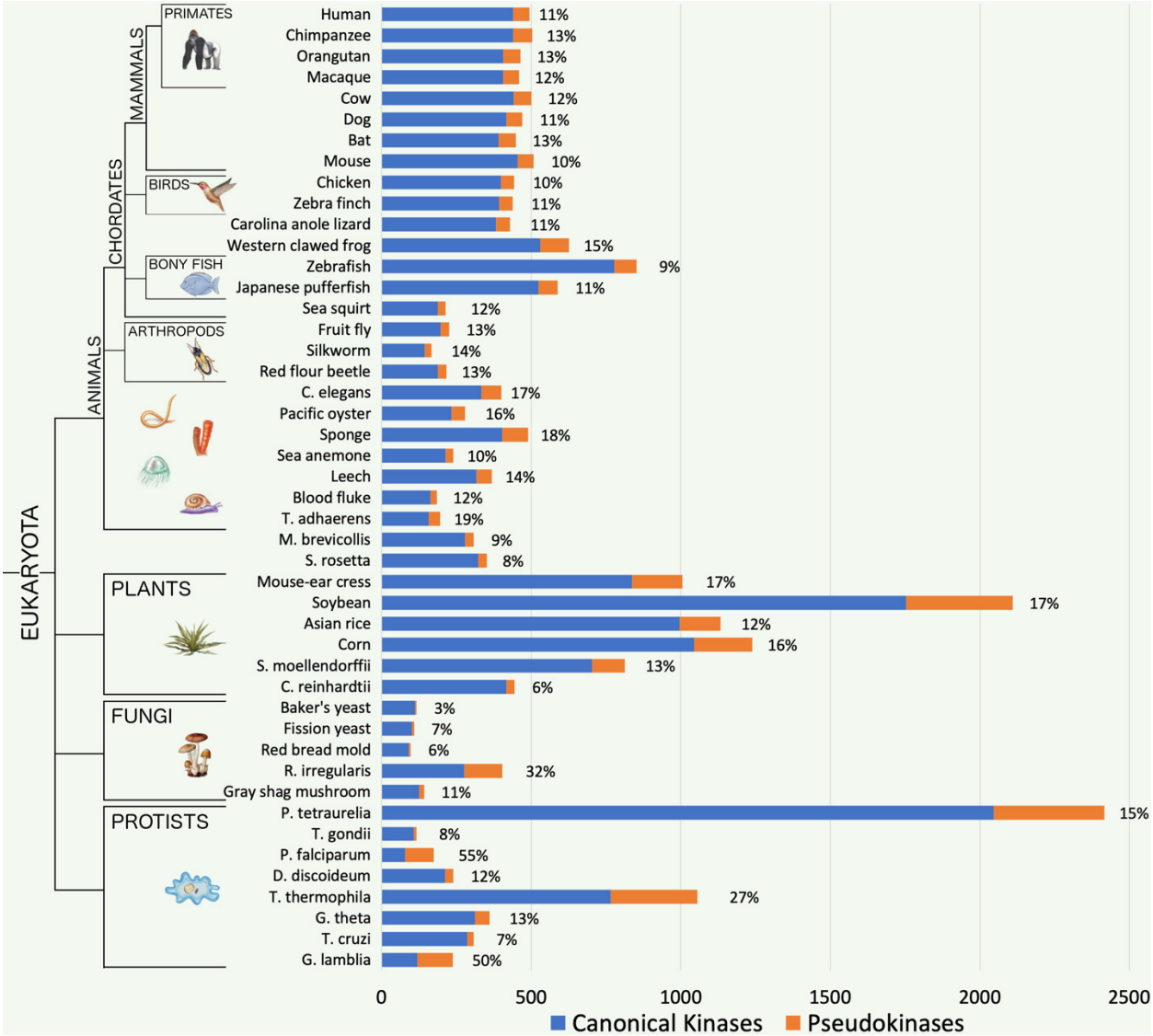

Figure 4.1. Kinome and pseudokinome sizes evaluated in 46 eukaryotic species. The counts of protein kinases and pseudokinases detectable in various eukaryotic proteomes are shown. Blue bars represent the number of canonical kinases detected in the proteome of each eukaryotic species, and orange bars represent the number of predicted pseudokinases. Percentages indicate the fraction of total kinases from each proteome that were determined to be pseudokinases based

on the lack of at least one residue in the catalytic triad. The tree on the left indicates major evolutionary kingdoms and phyla.

Also unprecedented was the varied detection of pseudokinases across diverse bacterial phyla with high sequence similarity to PknB kinases (Figure 4.2). Bacterial kinomes and pseudokinomes exhibit a large amount of diversity in size, particularly when compared to those in eukaryotes. For example, we note the large expansion of pseudokinases in Streptomyces coelicolor, which has 31 protein kinases and 5 pseudokinases, whereas the proteomes of Shigella flexneri and Escherichia coli like most bacterial proteomes lack pseudokinases, containing only 1 protein kinase sequence each. In contrast, the proteomes of some bacterial species, including Treponema denticola, do not contain any detectable protein kinase sequences. Nonetheless, pseudokinases were detected in at least one species for every bacterial phylum that we examined, except in Chlamydiae, suggesting that pseudokinases are not confined to any specific bacterial classifications such as gram-negative or gram-positive bacteria, but rather are found across diverse bacterial phyla.

Remarkably, a small number of ePK sequences were also detected in archaea (Figure 4.2), where 11 archaeal proteomes contained putative pseudokinase sequences. In particular, while most of these archaeal proteomes contained only one or two pseudokinase sequences, 7 out of the 8 ePK sequences detected in Halorientalis regularis were identified as pseudokinases.

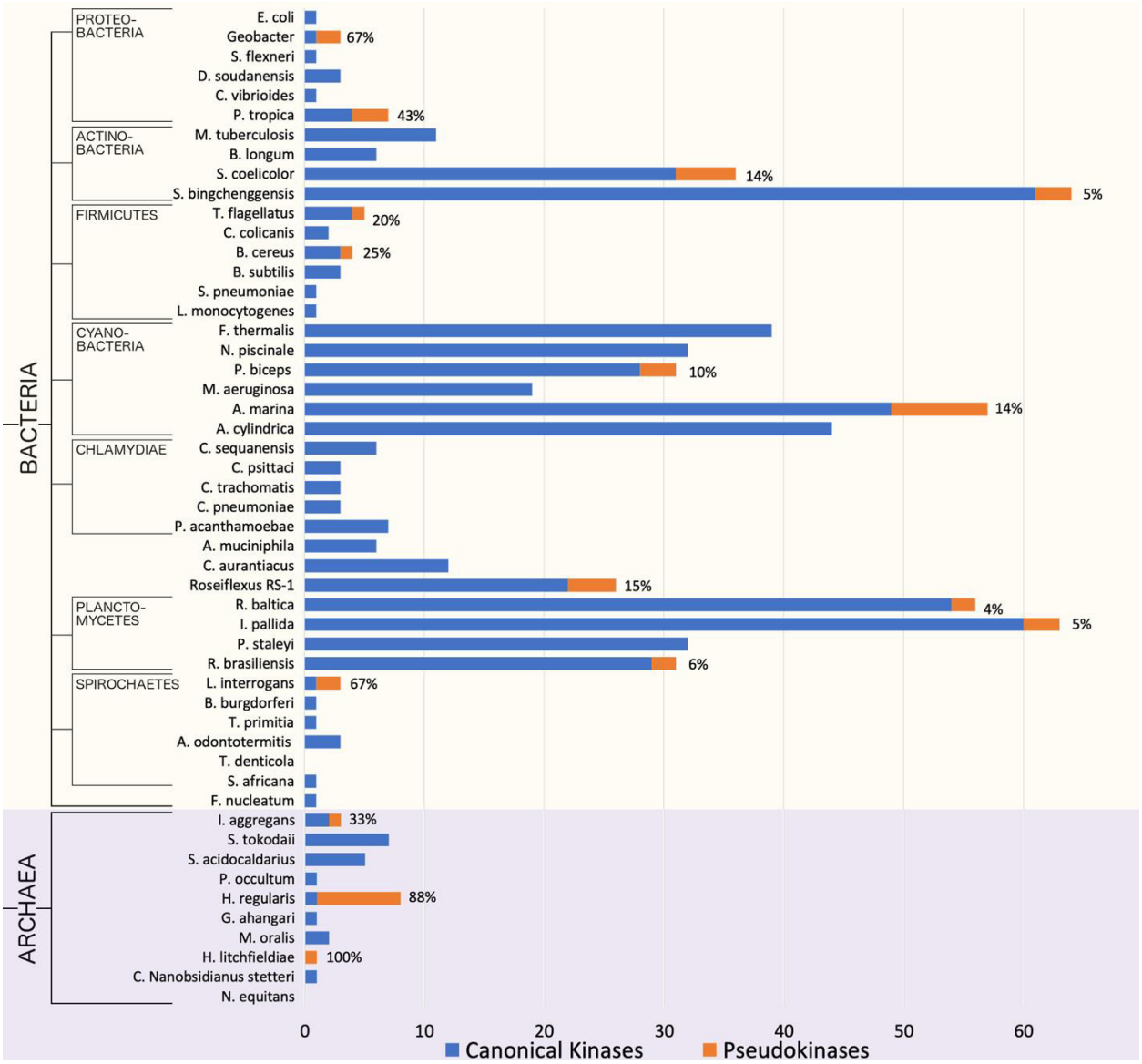

Figure 4.2**.** Kinome and pseudokinome sizes evaluated in 51 bacterial and archaeal species. The counts of protein kinases and pseudokinases detectable in various bacterial and archaeal proteomes are shown. Blue bars represent the number of canonical kinases detected in the proteome of each bacterial and archaeal species, and orange bars represent the number of predicted pseudokinases. Percentages indicate the fraction of total kinases from each proteome that were determined to be pseudokinases based on the lack of at least one residue in the catalytic triad. The tree on the left indicates major evolutionary kingdoms and phyla.

4.2.2 Classification of pseudokinase sequences

We next classified pseudokinase sequences into evolutionarily related clusters using an optimal multiple-category Bayesian Partitioning with Pattern Selection (omcBPPS) algorithm,

which classifies sequences based on patterns of amino acid conservation and variation in a large multiple sequence alignment (see methods) (63, 64). Due to the large number of pseudokinase sequences analyzed, we first classified a diverse representative set of 26,273 pseudokinase sequences (see methods for details). Some of the well-known metazoan pseudokinase families such as the JAK and TRIB families were found to fall into distinct pseudokinase sequence clusters (26). The identified pseudokinase clusters were incorporated within an existing evolutionary hierarchy of kinase groups and families (18) (see methods), and a resulting hierarchy of 592 sequence profiles containing both canonical kinase families and pseudokinase families were used to detect, classify, and align all pseudokinase sequences from the UniProt reference proteome database.

Overall, we detected pseudokinase families across all major kinase groups (Figure 4.3). Whereas some kinase groups defined by Manning and colleagues (18), such as the STE and AGC groups, possess relatively few pseudokinases, the Tyrosine Kinase-Like (TKL) group is highly enriched with pseudokinases. This increase is caused by the diversification of TKL pseudokinase sequences in various plant and fungal species, where the general expansion of canonical TKL kinases has previously been noted (52, 65). We also identify species-specific divergence of pseudokinase sequences, such as the divergence of metazoan-specific poly(A)-nuclease deadenylation complex subunit 3 (PAN3) from other eukaryotic PAN3, and the diversification of plant WNK kinases from the classical chordate WNK kinases. In contrast, some pseudokinase clusters comprise orthologs found across diverse taxonomic groups, such as Haspin, which is found across diverse eukaryotes, and the TRIB family of pseudokinases, which is found across diverse metazoa, where they function as regulators of protein ubiquitylation (26).

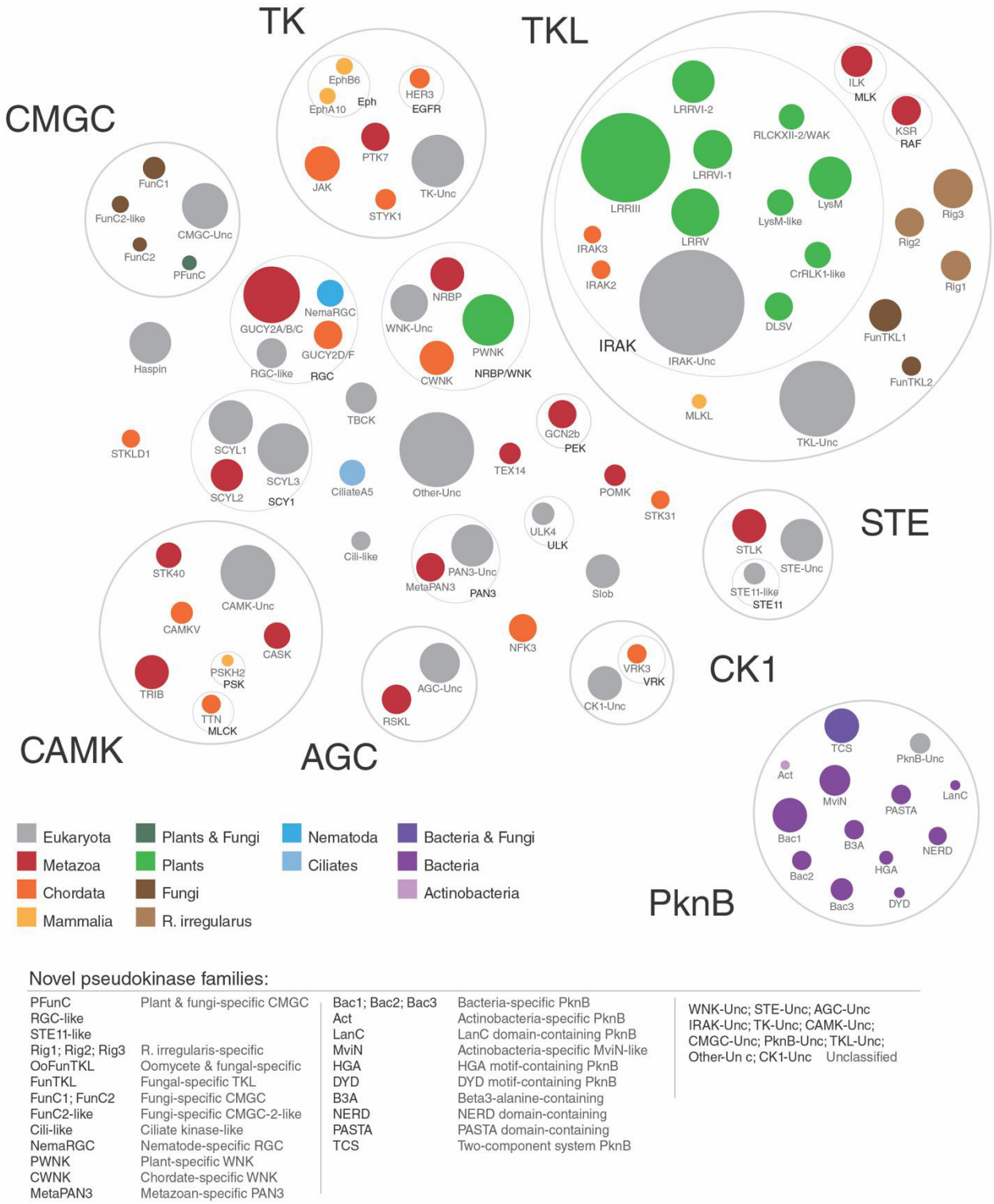

Figure 4.3**.** A new classification of pseudokinase families. Each colored circle represents a distinct pseudokinase family, which is colored according to the taxonomic group(s) in which it is found. The size of each circle represents the relative size of the corresponding pseudokinase family in log scale, and the distances between circles reflect the approximate sequence

divergence between families as determined by HMM-to-HMM distances. Kinase groups and families from the human kinome classification (*18*) are depicted by gray circles and labeled in bold font.

Notably, using our integrated evolutionary hierarchy of pseudokinase clusters with previously established kinase groups and families, we can now identify some canonical kinase sequences conserving the catalytic triad residues that classify into pseudokinase families. For example, four sequences containing the canonical HRD motif classified with the HER3 pseudokinases, which typically contain a conserved HRN motif that severely blunts catalysis. These sequences were identified to be HER3 orthologs in four separate rodent species (Mus musculus, Rattus norvegicus, Cricetulus griseus, and Mesocricetus auratus) (66). Similarly, whereas orthologs of the KSR family can be detected across diverse metazoan species, only chordate KSR's are classified as predicted pseudokinases due to the replacement of the β3-lysine by an arginine within the canonical 'VAIK' motif. In addition, we often identify large expansions of pseudokinase families in plant species that are accompanied by the presence of few canonical sequence members. For example, in black cottonwood (Populus trichocarpa), we identify one canonical PWNK member containing the β3-lysine ('VAWK' motif), along with 15 pseudokinase members containing a 'VAWN' motif characteristic of PWNK pseudokinases. The clustering of canonical sequences within pseudokinase families is summarized in Table 4.3.

Pseudokinase family	# Pseudokinase sequences	# Canonical sequences	%Canonical
JAK	282	0	0.00%
EphA10	41	0	0.00%
EphB6	45	0	0.00%
STYK1	75	0	0.00%
ILK	210	0	0.00%
MLKL	32	0	0.00%
IRAK2	58	0	0.00%

IRAK3	49	0	0.00%
LysM	465	0	0.00%
LysM-like	139	0	0.00%
Rig1	192	0	0.00%
FunTKL2	62	0	0.00%
RSKL	181	0	0.00%
CAMKV	89	0	0.00%
STK40	126	0	0.00%
GUCY2D/GUCY2F	165	0	0.00%
NemaRGC	144	0	0.00%
GCN2b	162	0	0.00%
Slob	259	0	0.00%
STK31	56	0	0.00%
NKF3	159	0	0.00%
TEX14	83	0	0.00%
NRBP	257	0	0.00%
WNK-Unc	334	0	0.00%
CWNK	265	0	0.00%
SCYL1	503	0	0.00%
SCYL2	228	0	0.00%
SCYL3	722	0	0.00%
TBCK	210	0	0.00%
PAN3-Unc	444	0	0.00%
MetaPAN3	168	0	0.00%
Bac2	63	0	0.00%
Act	10	0	0.00%
MviN	203	0	0.00%
HGA	27	0	0.00%
NERD	54	0	0.00%
PASTA	65	0	0.00%
PWNK	738	1	0.14%
STLK	262	1	0.38%
TRIB	261	1	0.38%

GUCY2A/GUCY2B/GUCY2C	935	4	0.43%
LRRV	618	3	0.48%
RGC-like	190	1	0.52%
Haspin	436	4	0.91%
LRRIII	2935	29	0.98%
VRK3	63	1	1.56%
TTN	60	1	1.64%
PTK7	164	3	1.80%
CASK	138	3	2.13%
Bac3	91	2	2.15%
LRRVI-2	455	11	2.36%
POMK	82	2	2.38%
LRRVI-1	364	9	2.41%
B3A	64	2	3.03%
HER3	70	4	5.41%
Bac1	269	22	7.56%
LanC	12	1	7.69%
STKLD1	54	8	12.90%
TCS	273	54	16.51%
RLCKXII-2/WAK	128	28	17.95%
CiliateA5	131	32	19.63%
CrRLK1-like	137	36	20.81%
Rig3	377	127	25.20%
DYD	14	5	26.32%
Rig2	174	96	35.56%
KSR	179	118	39.73%
ULK4	93	73	43.98%
PSKH2	19	18	48.65%
FunTKL1	230	263	53.35%
Cili-like	62	103	62.42%
FunC2	29	54	65.06%
FunC2-like	46	94	67.14%
DLSV	164	408	71.33%

PFunC	32	136	80.95%
FunC1	91	424	82.33%
STE11-like	84	424	83.46%

Table 4.3: Counts of canonical sequences classified into pseudokinase families

Additionally, our in-depth analysis led to the identification of many previously undefined pseudokinase families (Table 4.4). These include the plant lysin motif (LysM)-like pseudokinase family and the bacterial HGA motif-containing PknB pseudokinase family, as well as unclassified pseudokinase families that do not readily fall within the identified pseudokinase clusters, which we term the tyrosine kinase (TK)-unclassified, TKL-unclassified, STE-unclassified, CK1-unclassified, AGC-unclassified, CAMK-unclassified, CMGC-unclassified, PknB-unclassified, and Other-unclassified pseudokinase families (Figure 4.3). As noted previously, pseudokinases in the TKL group appear to have expanded substantially, particularly in non-metazoan species. Consistently, five distinct fungal-specific pseudokinase families have emerged in the TKL group, of which three (termed Rig 1-3) are currently found predominantly in the species R. irregularis, which has a considerably expanded kinome and pseudokinome compared to other fungi (Figure 4.1). In addition, IRAK pseudokinases (part of the TKL group) are massively expanded in plants, which classify into 9 unique families. This mirrors the well-known plant-specific expansions of the canonical IRAK (also known as RLK/Pelle) kinase family, which have previously been classified into 65 subfamilies (52). A surprising finding from our analysis was the detection of over 1,200 putative bacterial PknB pseudokinases, which we have classified into 12 distinct clusters. Some bacterial pseudokinase clusters are specific to selected bacterial phyla, such as the Actinobacteria-specific PknB pseudokinase family (which we term Act), whereas other families such as the NERD domain-containing PknB pseudokinase family are found more broadly across diverse bacterial phyla. Sequences and alignments for

each pseudokinase family identified here are available through the Protein Kinase Ontology (ProKinO, http://vulcan.cs.uga.edu/prokino/about/browser). In the following sections, we build on the first comprehensive pseudokinase catalogue by examining the putative functions of plant IRAK pseudokinases and pseudokinases in R. irregularis and in sequenced bacteria.

Pseudokinase family	Kinase family	Kinase group	Example member(s)
TK-Unclassified		TK	Camponotus floridanus putative tyrosine-protein kinase Wsck (E2ACX7_CAMFO)
LysM	IRAK	TKL	Arabidopsis thaliana LYK5/LysM-containing receptor-like kinase 5/At2g33580 (LYK5_ARATH)
LysM-like	IRAK	TKL	Arabidopsis thaliana Protein kinase superfamily protein/At3g57120 (Q8VYG5_ARATH)
DLSV	IRAK	TKL	Arabidopsis thaliana Cysteine-rich receptor-like protein kinase 45/Cysteine-rich RLK45/At4g11890 (CRK45_ARATH)
RLCKXII-2/WAK	IRAK	TKL	Arabidopsis thaliana Non-functional pseudokinase ZED1/hopZ-ETI-deficient 1/At3g57750 (ZED1_ARATH)
CrRLK1-like	IRAK	TKL	Cajanus cajan Receptor-like protein kinase ANXUR2 /KK1_043523 (A0A151QYL0_CAJCA)
LRRIII	IRAK	TKL	Arabidopsis thaliana PRK1/Pollen receptor-like kinase 1/At5g35390 (PRK1_ARATH)
LRRV	IRAK	TKL	Arabidopsis thaliana SUB/STRUBBELIG/At1g11130 (SUB_ARATH)
LRRVI-1	IRAK	TKL	Arabidopsis thaliana Leucine-rich repeat protein kinase family protein/MUA22.21/At5g14210 (A0A178US29_ARATH)
LRRVI-2	IRAK	TKL	Arabidopsis thaliana MDIS1/Male discoverer 1/At5g45840 (MDIS1_ARATH)
Rig1		TKL	Rhizophagus irregularis Rad53p/RirG_173180 (A0A015K1H5_9GLOM)
Rig2		TKL	Rhizophagus irregularis Skt5p/RirG_180750 (A0A015ISZ5_9GLOM)
Rig3		TKL	Rhizophagus irregularis Ssk22p/RirG_232660 (A0A015IBX1_9GLOM)
FunTKL1		TKL	Rhizoctonia solani Uncharacterized protein/ RSAG8_12895 (A0A066UWY2_9HOMO)
FunTKL2		TKL	Penicillium subrubescens Uncharacterized protein/PENSUB_13048 (A0A1Q5SUF1_9EURO)
IRAK-Unclassified		TKL	Arabidopsis thaliana BIR2/BAK1-interacting receptor-like kinase 2/At3g28450 (BIR2_ARATH)
STE11-like		STE	Arabidopsis thaliana Protein kinase superfamily protein/At2g40580 (O22877_ARATH)
STE-Unclassified		AGC	Homo sapiens Uncharacterized protein (A0A1B0GUL7_HUMAN); Pan troglodytes Uncharacterized protein (A0A2I3RK43_PANTR)
AGC-Unclassified		AGC	Paramecium tetraurelia hypothetical protein/ GSPATT00021006001 (A0DWE7_PARTE)
CAMK-Unclassified		CAMK	Paramecium tetraurelia Uncharacterized protein/GSPATT00003535001 (A0E5V5_PARTE)

PFunC		CMGC	Sorghum bicolor Uncharacterized protein/SORBI_3005G165300 (A0A1Z5RK15_SORBI)
FunC1		CMGC	Penicillium patulum Uncharacterized protein/ PGRI_024730 (A0A135LI09_PENPA)
FunC2		CMGC	Penicillium subrubescens SRSF protein kinase 3/ PENSUB_10992 (A0A1Q5T6W4_9EURO)
FunC2-like		CMGC	Coprinopsis cinerea CMGC/SRPK protein kinase/CC1G_09673 (A8P9H1_COPC7)
NemaRGC	RGC	Other	Caenorhabditis elegans Receptor-type guanylate cyclase gcy-1/AH6.1 (GCY1_CAEEL)
RGC-like	RGC	Other	Amphimedon queenslandica Guanylate cyclase (A0A1X7VH93_AMPQE)
CWNK	WNK	Other	Homo sapiens WNK1/Protein kinase with no lysine 1 (WNK1_HUMAN)
PWNK	WNK	Other	Arabidopsis thaliana AtWNK1/Protein kinase with no lysine 1/At3g04910 (WNK1_ARATH)
WNK-Unclassified	WNK	Other	Caenorhabditis elegans Serine/threonine-protein kinase WNK/Protein kinase with no lysine 1/C46C2.1 (WNK_CAEEL)
Cili-like		Other	Neurospora crassa Mitotic spindle checkpoint component mad3/NCU10043 (V5IRN2_NEUCR)
MetaPAN3		PknB	Homo sapiens PAN3/Poly(A)-nuclease deadenylation complex subunit 3 (PAN3_HUMAN)
Bac1		PknB	Streptomyces aurantiacus JA 4570 Uncharacterized protein/STRAU_1734 (S3ZNQ3_9ACTN)
Bac2		PknB	Frankia sp. EI5c Serine/threonine protein kinase /UG55_100447 (A0A166T1F0_9ACTN)
Bac3		PknB	Trichodesmium erythraeum (strain IMS101) Serine/threonine protein kinase with TPR repeats/Tery_4781 (Q10VJ1_TRIEI)
Act		PknB	Streptomyces sp. NTK 937 Aminoglycoside phosphotransferase/DT87_12690 (A0A069K2D6_9ACTN)
MviN		PknB	Mycobacterium shimoidei Transmembrane serine/threonine-protein kinase D PknD/STPK D/ BHQ16_10175 (A0A1E3TG29_MYCSH)
LanC		PknB	Streptomyces vietnamensis Uncharacterized protein /SVTN_34785 (A0A0B5I836_9ACTN)
DYD		PknB	Frankia sp. (strain EAN1pec) Nucleic acid binding OB-fold tRNA/helicase-type/Franean1_4264 (A8L9E5_FRASN)
HGA		PknB	Actinosynnema mirum (strain ATCC 29888) Uncharacterized protein/Amir_3792 (C6WD53_ACTMD)
B3A		PknB	Lentisphaera araneosa HTCC2155 Serine/threonine-protein kinase/LNTAR_17493 (A6DFI7_9BACT)
PASTA		PknB	Firmicutes bacterium CAG:94 PASTA domain protein/BN815_00934 (R6ZDX7_9FIRM)
NERD		PknB	Singulisphaera acidiphila (strain ATCC BAA-1392) Nuclease-like protein/protein kinase family protein/Sinac_1959 (L0DAB9_SINAD)
TCS		PknB	Rhodopirellula islandica Two-component signal transduction/RISK_006087 (A0A0J1B5P6_RHOIS)

Table 4.4. List of newly identified pseudokinase families. Representative members of previously unidentified pseudokinase families are listed.

4.2.3 A massive, plant-specific IRAK pseudokinase expansion

The plant IRAK kinases have been previously classified into 65 subfamilies (52). We now identify 9 unique IRAK pseudokinase families (Figure 4.3), which are conserved across multiple plant species, including two pseudokinase families that resemble the lysin-motif receptor-like (LysM RLK) kinases, termed LysM and LysM-like, and the leucine-rich repeat receptor-like (LRR RLK) pseudokinase families, termed LRRIII, LRRV, LRRV1-1, and LRRVI-2. In addition, we define three "mixed" pseudokinase families that contain domains homologous to multiple previously defined IRAK subfamilies, such as the receptor-like cytoplasmic kinase XII-2 (RLCKXII-2)/wall-associated kinase (WAK) family and the Catharanthus roseus RLK 1 (CrRLK1)-like family.

To specifically examine the evolutionary expansion of IRAK pseudokinases in plants, we constructed a rooted phylogenetic tree of both canonical and pseudokinase IRAK sequences using diverse non-IRAK TKL sequences and non-TKL sequences as an outgroup. The plant IRAK pseudokinase families form distinct clades in the phylogenetic tree that clearly distinguish them from metazoan IRAK pseudokinase sequences (Figure 4.4, panelA). In addition, metazoan IRAK members are cytoplasmic and typically contain a death domain N-terminal to the kinase domain (Figure 4.4, panel C), which is not observed in any plant IRAK kinases. We note that the three mixed pseudokinase families we identified (RLCKXII-2/WAK, CrRLK1-like, DLSV) form distinct monophyletic clades in the phylogenetic tree, indicating close homology and common descent, despite their homology to multiple IRAK subfamilies. In addition, the IRAK pseudokinases are generally divergent from canonical plant IRAK members, as shown by the long branch lengths separating the pseudokinase family clades (red lines) from the canonical IRAK sequences (gray lines) (Figure 4.4, panel A). In some pseudokinase families, several

canonical sequences cluster within pseudokinase family clades (namely RLCKXII-2/WAK, CrRLK1-like, DLSV, and LRRIII, and LRRV families), indicating that these pseudokinase families also have very close, likely catalytically active, homologs.

We next examined the relative amount of degradation of the canonical catalytic motifs and the domain organizations of the major IRAK pseudokinase families, allowing us to expand on the potential functions of some IRAK pseudokinases. The LysM pseudokinase family shares sequence homology in the kinase domain with known catalytically active LysM RLKs and conserves a similar domain architecture, which is characterized by an intracellular kinase domain, and one or more extracellular LysM domains (Figure 4.4, panel C). In plants, LysM RLKs play diverse sensing functions, recognizing chitin oligosaccharides in plant defense responses towards fungi (67), as well as binding peptidoglycans on bacterial cell walls to aid recognition of symbiotic bacteria (68). Recent studies of two LysM pseudokinases, MtNFP and LjNFR5 from Medicago, demonstrate the importance MtNFP and LjNFR5 during the Rhizobium pre-infection response and to the specificity of Rhizobium-legume symbiosis, respectively. Despite their lack of catalytic activity, these LysM pseudokinases are believed to contribute to their appropriate signaling pathways via interactions with other active LysM RLKs (69). We also detect another distinct family, which we term the LysM-like pseudokinase family, which shares sequence homology in the kinase domain with active LysM RLKs, but lacks LysM and transmembrane domains, suggesting a uniquely cytoplasmic function for this family (Figure 4.4, panel C). Moreover, the LysM and LysM-like families share different patterns of residue conservation, as identified in our pattern-based classification, and they form distinct clades in the phylogenetic tree (Figure 4.4, panel A), suggesting that they likely have divergent functions.

The LRRV pseudokinase family comprises the Strubbelig receptor family, which consists predominately of pseudokinases, with 9 members in Arabidopsis thaliana alone (70). The best characterized member of this family, Strubbelig (SUB), plays roles in organ development and cellular morphogenesis, however, the mechanism by which it contributes to cell signaling despite a lack of catalytic activity, and the functions of other LRRV members are not yet known (70, 71). While the domain organizations for most IRAK pseudokinase families such as LRRV are rather well conserved, pseudokinases in the DLSV family possess diverse domain architectures, indicating that a common pseudokinase domain has co-evolved with a variety of different protein domains in order to diversify biological function. For example, DLSV pseudokinases homologous to the DUF26 IRAK subfamily contain extracellular domains associated with salt-stress and antifungal responses (72, 73) (Figure 4.4, panel C). Although DUF26 kinases have been associated with plant-specific functions in ROS/redox signaling and stress adaptation (74), to our knowledge, pseudokinase complements of DUF26 kinases have not been described previously. Other DLSV pseudokinases exhibit tandem pseudokinase and canonical kinase domains (Figure 4.4, panel C). Previously described IRAK pseudokinases in plants include MDIS2 (MRH1) (75), ZED1 (51), RSK1 (76), BSK8 (77), BIR2 (49), SOBIR1 (78), and CRN (79), and their placement in the expanded IRAK pseudokinase classification is described in Table 4.5.

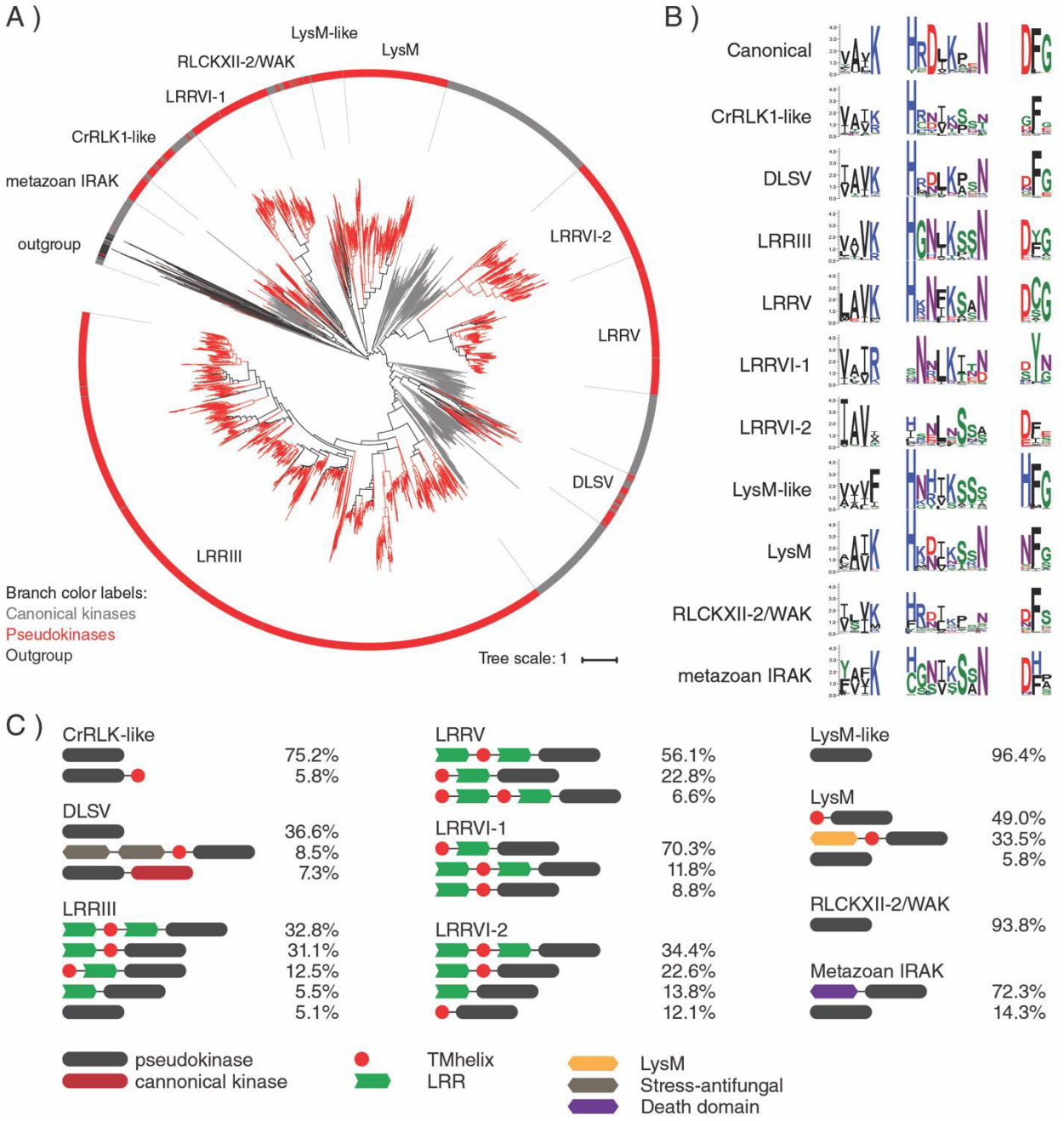

Figure 4.4**.** Plant-specific IRAK pseudokinase families. (A) Phylogenetic tree of catalytically active and pseudokinase members of the IRAK family. The 9 plant IRAK pseudokinase families are labeled, and IRAK pseudokinase sequences are shown in red. Canonical IRAK sequences are shown in gray. Outgroup sequences are shown in black. (B) Sequence logos of catalytic motifs for IRAK pseudokinase families. (C) Unique domain structures observed in plant IRAK pseudokinase families. The most common domain structures observed in each family are shown (occurring >5%), with frequencies of each domain structure indicated.

Pseudokinase	Pseudokinase Classification	IRAK subclassification (Shiu)
MDIS2 (MRH1)	LRRVI-2	LRRVI-2
ZED1	RLCK-XII-2/WAK	RLXK-XII-2
ZAR1	LRRIII	LRRIII
RSK1	RLCK-XII-2/WAK	RLXK-XII-2
BSK8	IRAK-unclassified	RLCK-XII-1
BIR2	IRAK-unclassified	LRR-Xa
SOBIR1	IRAK-unclassified	LRR-XI-2
CRN	IRAK-unclassified	LRR-XI-2
SUB	LRRV	LRRV
SRF1	LRRV	LRRV
SRF2	LRRV	LRRV
SRF3	LRRV	LRRV
SRF4	LRRV	LRRV
SRF5	LRRV	LRRV
SRF6	LRRV	LRRV
SRF7	LRRV	LRRV
SRF8	LRRV	LRRV

Table 4.5: Known plant IRAK pseudokinases and their classifications

4.2.4 Expansion of TKL pseudokinases in Rhizophagus irregularis

We identified and analyzed 3 previously unidentified fungal pseudokinase families (termed Rig1, Rig2, and Rig3), which are currently comprised predominantly of sequences from a single species, the commercially important soil inoculant Rhizophagus irregularis (Figure 4.3). This organism has a highly expanded proteome when compared to other fungal species, including an expanded kinome (65, 80). Notably, the R. irregularis fungal kinome comprises ~2.6% of the entire proteome, a larger proportion than is observed for any other fungal kinome. 32% of these kinases possess pseudokinase domains (Figure 4.1) that most closely resemble the

TKL kinases. Several lines of evidence suggest that the genes coding these pseudokinases are expressed at the protein level (65, 81), suggesting that protein pseudokinases must contribute to an important biological function in R. irregularis.

Sequence comparisons to known TKL kinase families demonstrated that R. irregularis-specific TKL kinases are divergent from canonical TKL families and are most homologous to the leucine rich repeat kinase (LRRK) family of TKL's. LRRK's are intracellular kinases distinct from IRAK LRR RLK's and are conserved in metazoans and expanded in the slime-mold Dictyostelium discoidium but are otherwise absent in fungi. In order to understand the evolutionary events that led to the expansion of these pseudokinase families, we conducted a phylogenetic analysis of the entire R. irregularis kinome, which, like the human kinome (18), consists of 7 major ePK groups. R. irregularis kinase sequences from the AGC, CMGC, STE, CAMK, and CK1 groups form mostly separate monophyletic clades in a phylogenetic tree of the R. irregularis kinome, whereas TK sequences are clustered within various groupings of TKL's (Figure 4.5, panel A). Interestingly, the R. irregularis kinome additionally includes PknB-like kinase sequences, which are typically associated with bacteria. However, as reflected by the phylogenetic tree (Figure 4.5, panel A), TKL sequences comprise the majority of the R. irregularis kinome. Thus, the expanded kinome of R. irregularis can be attributed to a substantial expansion amongst TKL kinases. Furthermore, we found that 166 of the 183 pseudokinases (90%) identified in R. irregularis are TKL-like, indicating that R. irregularis pseudokinases emerged primarily from within the TKL group. Of the three distinct pseudokinases families, Rig1 forms a monophyletic group comprised entirely of pseudokinases, whereas Rig2 and Rig3 cluster into clades including both pseudokinase and canonical kinase members. Rig2 and Rig3

are the most diverse pseudokinase families in R. irregularis, and both cluster with canonical sequences from the TK and "Other" groups, based on the human kinome classification (18).

Further analysis of the domain organization in these pseudokinase sequences revealed that Rig 1 and Rig2 pseudokinases often have an additional putative tetratricopeptide (TPR) domain C-terminal to the pseudokinase domain, whereas Rig3 pseudokinases are single domain pseudokinases (Figure 4.5, panel C). TPR repeats are short structural motifs that are classically involved in mediating protein-protein interactions crucial for cell signaling. Apart from the TPR repeat regions, no additional domains were found in the pseudokinase members of the 3 families.

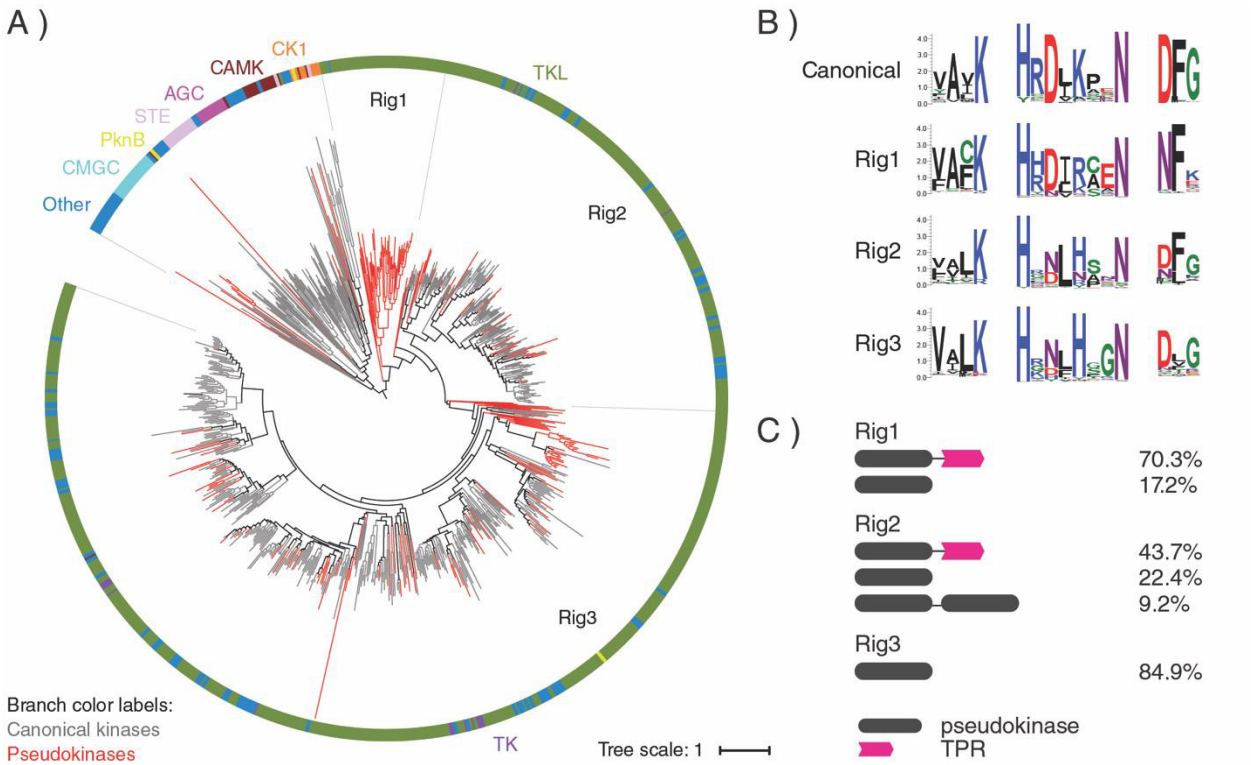

Figure 4.5**.** *Rhizophagus irregularis*-specific TKL pseudokinase families. (A) Phylogenetic tree of the *R. irregularis* kinome. Canonical kinase branches are colored in gray and pseudokinases in red. Major kinase groups are labelled using different colors in the outer circle. The 3 major *R. irregularis* specific pseudokinase families are labelled as Rig1, Rig2 and Rig3. (B) Sequence logos of catalytic motifs for Rig1, Rig2, and Rig3 pseudokinase families. (C) The most common domain structures observed in Rig pseudokinase families are shown (occurring >5%), with frequencies of each domain structure indicated.

4.2.5 Conservation of PknB-like pseudokinases in bacteria

We identified 12 unique families of bacterial pseudokinases that are most closely related to the PknB group of canonical prokaryotic protein kinases (Figure 4.3) and have been classified and named based on their conserved domains, taxonomic specificity, and/or their similarity to previously identified kinases. Three bacterial pseudokinase families (DYD, HGA, B3A) were named based on the unique conservation of noncanonical catalytic motifs. In order to examine the evolutionary relationships between these families, we built a phylogenetic tree of the 1,145 PknB pseudokinase sequences identified in this study combined with a representative set of canonical PknB kinases (Figure 4.6, panel A). Each bacterial pseudokinase family falls into a distinct clade and mostly segregates away from canonical kinase sequences.

Analysis of the catalytic motifs for each PknB pseudokinase family shows that different PknB pseudokinase families diverge in the canonical catalytic motifs in a variety of ways (Figure 4.6, panel B). For example, the MviN and PASTA families exhibit the most extreme degeneration of catalytic residues and lack all three canonical residues associated with catalysis (β3-lysine, HRD-aspartate, DFG-aspartate), in addition to the conserved magnesium-binding asparagine located at the end of the catalytic loop HRDXXXXN motif. Conversely, many bacterial pseudokinase families appear to conserve the catalytic aspartates of the HRD and DFG motifs, as well as the magnesium-binding asparagine residue; this is the case for the Act, DYD, B3A, and LanC pseudokinase families. Of these, the Act, DYD, and LanC families all possess a chemically comparable arginine residue in place of the ATP-binding β3-lysine, and thus might still retain ATP-binding and catalytic functions, as demonstrated for canonical kinases such as Aurora A (82).

Analysis of the common domain organizations of bacterial PknB pseudokinase families reveals that, like eukaryotic pseudokinases, bacterial pseudokinases also co-occur with protein signaling domains involved in a wide range of biological functions (Figure 4.6, panel C). Two families in particular, termed NERD and TCS, have highly conserved domain architectures of particular interest. For example, the domain architecture of NERD is characterized by an N-terminal NERD domain, which is predicted to function as a nuclease (83), followed by a PknB pseudokinase domain and often a C-terminal canonical (catalytically active) PknB kinase domain. This multi-domain architecture is also observed in PglW, a constituent of the Streptomyces coelicolor A(3)2 phage growth limitation system (84). The active kinase domain of the PglW is catalytically competent (85), whereas the function of the pseudokinase domain remains unknown. Additionally, we identified putative protein domains that resemble the C-terminal domain of the bacterial RNA polymerase α-subunit in ~45% of the NERD family members; this domain classically functions in DNA binding and protein-protein interactions in bacterial RNA polymerases (86). The co-I of these nucleic acid-associated domains with pseudokinase domains suggests that the NERD pseudokinase family may be involved in signaling pathways relevant to transcriptional regulation. We also note the unique domain structure within the TCS family, which has clear orthologs in both bacteria and fungi (Figure 4.3) contains a pseudokinase domain that often co-occurs with other protein domains typically found in two-component signaling systems involving histidine and aspartate phosphorylation. Nearly 40% of TCS family pseudokinases contain an N-terminal pseudokinase domain followed by an ATPase domain, a GAF-domain, and a C-terminal histidine kinase domain (Figure 4.6, panel C). This domain architecture is reminiscent of two-component system proteins that have been previously identified in bacterial species from the Cyanobacteria, Proteobacteria, and

Spirochaete phyla, as well as fungal species such as Candida albicans and Schizosaccharomyces pombe (87-92). These proteins have been associated with a wide range of functions, including nitrogen metabolism and glycolipid synthesis in Anabaena sp. PCC7120, hyphal development and pathogenicity in C. albicans, and cell cycle regulation and oxidative stress response in S. pombe (87, 91, 93-101). To our knowledge, our analysis is the first to reveal pseudokinase domains that are likely to be associated with two-component signaling in bacteria.

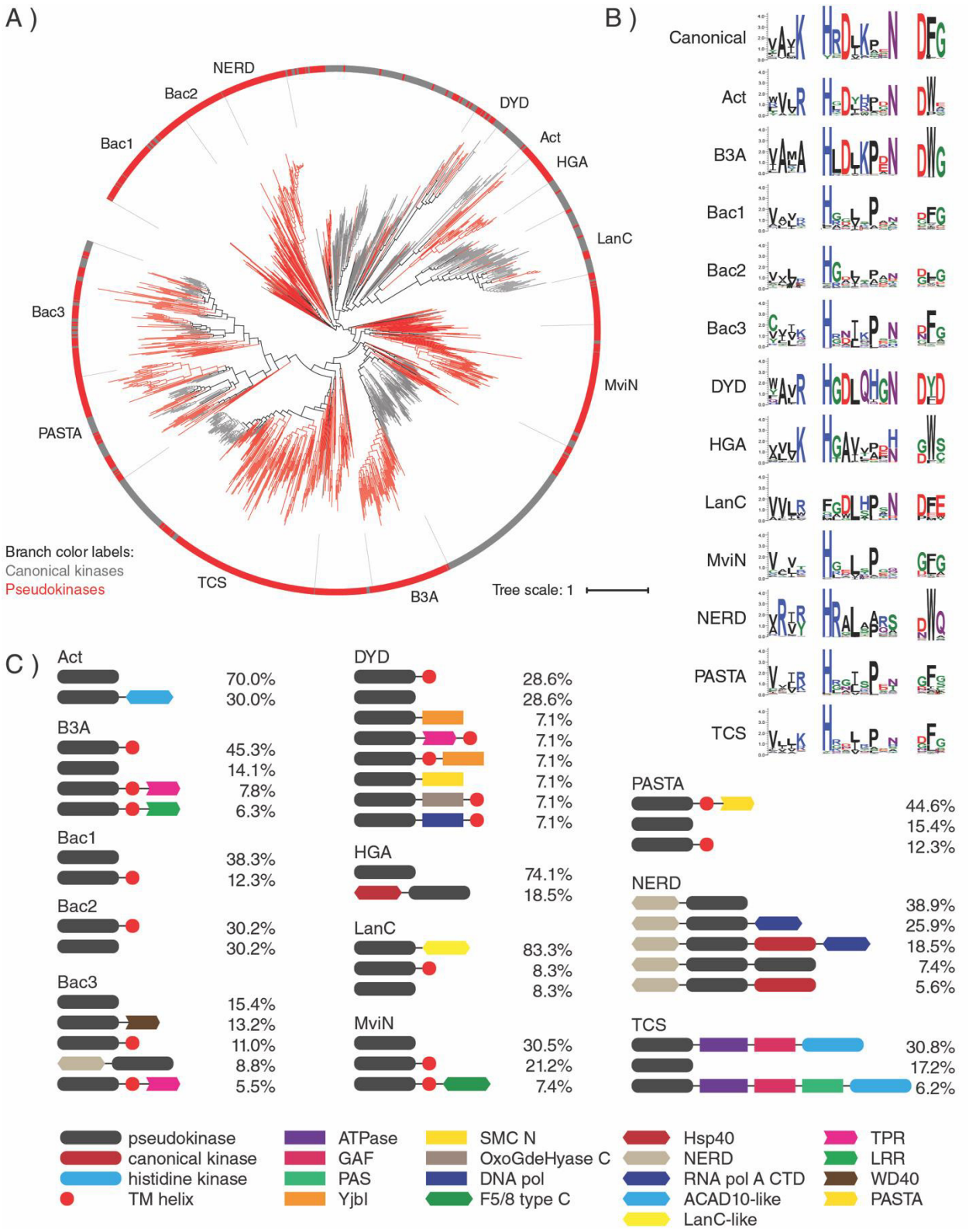

Figure 4.6**.** Bacterial PknB pseudokinase families. (A) Phylogenetic tree of PknB canonical kinases and pseudokinases. The 12 PknB-related pseudokinase families are labeled and shown in

red branches. Representative canonical PknB kinases are shown in gray. (B) Sequence logos of catalytic motifs in PknB pseudokinase families. (C) Unique domain structures observed in bacterial pseudokinase families. The most common domain structures observed in each family are shown (occurring >5%), with percentages indicating the frequency of each domain structure.

4.2.5 Sequence and structural basis for pseudokinase evolutionary divergence: A case study on two plant IRAK pseudokinase families

Conformational flexibility in pseudokinases has permitted the structurally conserved ancestral protein kinase fold to be 're-purposed' for multiple cellular signaling roles, including the evolution of new ways through which to bind and modulate cellular targets. Using the LRRVI-2 and RLCKXII-1 pseudokinase families as examples, we evaluated how evolution has constrained sequences in different pseudokinase families to disrupt canonical ATP-binding and constrain kinase domain conformations in a multitude of ways that serve to abolish catalytic activity. This leads us to propose previously unknown molecular functions that may have evolved in LRRVI-2 pseudokinases through the selection of unique motifs on the surface of the protein.

The LRRVI-2 pseudokinase family includes the previously described pseudokinase, MDIS2, which interacts with the plant potassium channel AKT2 (75) associated with root hair formation (102). However, little is known about the molecular functions of other pseudokinases in this family. Using the crystal structure of a LRRVI-2 pseudokinase from Zea mays (GRMZM2G135359), we examined the structural role of LRRVI-2 specific motifs. ATP binding in the GRMZM2G135359 structure appears to be completely inhibited due to the complete obstruction of the ATP-binding site by the activation loop. This activation loop conformation is stabilized by a LRRVI-2-specific lysine (Lys275) that replaces the ATP-binding C-helix glutamate (Figure 4.7, panel A) and hydrogen bonds to the backbone of the activation loop. Notably, a glutamate in the activation loop "DLE" motif, which replaces the canonical

DFG motif, hydrogen bonds to the glycine rich loop to occlude ATP binding. The inhibitory activation loop conformation is additionally stabilized by hydrophobic interactions between LRRVI-2-specific residues including a phenylalanine in the gatekeeper position on the β5 strand (Phe306), a phenylalanine in the αC-β4 loop (Phe287), and a cysteine in the E-helix (Cys341). These residues form hydrophobic interactions with Ala254 (which replaces the ATP-binding β3-lysine) and with Leu375 (which replaces the DFG-phenylalanine). Together these hydrophobic interactions appear to stabilize the C-helix in an inactive, outward conformation and the activation loop in an autoinhibitory conformation that occludes ATP-binding. In addition, a LRRVI-2 family-specific asparagine replaces the canonical F-helix aspartate, which typically participates in a switch-like mechanism and stabilizes the active conformation of the kinase domain by forming key hydrogen bonds with the catalytic loop backbone (103).

The F-helix aspartate is typically conserved as an asparagine in LRRVI-2 pseudokinases, although the GRMZM2G135359 structure has a hydrophobic isoleucine at this position, and other LRRVI-2 members are noted to have a valine, methionine or aspartate (Figure 4.7, panel A). Mutation of the F-helix-aspartate residue to an asparagine or a leucine in canonical kinases such as Aurora A abrogates activity (103), thus, substitution of the F-helix-aspartate to an asparagine or hydrophobic residue is predicted to inactivate LRRVI-2 pseudokinases. Likewise, variations in the catalytic loop (replacement of the HRD motif by LRN motif) may contribute to the observed inactive structural conformation in LRRVI-2. In addition, we found that several LRRVI-2 family-specific motifs occur well outside of the catalytically important regions, including the surface of the protein near the N- and C-terminal tails. We note one such example in the GRMZM2G135359 structure, where LRRVI-2-specific residues appear to dock the I-helix and C-terminal tail onto the backside of the kinase domain. Specifically, a methionine (Met336)

and a tyrosine (Tyr340) on the E-helix tether the C-tail through hydrophobic and hydrogen bonding interactions, respectively (Figure 4.7, panel A). Another cluster of LRRVI-2-specific, surface-exposed residues (A309, T365, A369) may also participate in tethering the C-tail and extend this interaction to the hinge region of the kinase domain.

To further investigate the evolutionary basis for IRAK pseudokinase functional specialization, we next quantified the evolutionary constraints imposed on the RLCKXII-1 family of pseudokinases, which comprise the brassinosteroid signaling kinases (BSKs). BSKs are involved in regulating plant growth and physiology in response to brassinosteroid hormone signals. The crystal structure of BSK8 has been determined (77) and shown to bind the non-hydrolyzable ATP analog AMP-PNP, despite the atypical DFG motif (CFG in BSK8) conformation. In addition to an unusual glycine-rich loop structure and the presence of a conserved small amino acid at the gatekeeper position (Ala132) (77), our studies also reveal family-specific variations in both the active site (Tyr185, Arg186, Asn205) and in allosteric regions such as the F-helix-aspartate (Val234), which provide clues to the unusual mode of AMP-PNP binding in RLCKXII-1 (Figure 4.7, panel B). Family-specific replacement of the canonical HRD-motif histidine and arginine in the catalytic loop (Tyr179 and His180) facilitates unique hydrogen bonding and hydrophobic interactions that stabilize the activation loop in a unique inactive conformation (Figure 4.7, panel B). Other RLCKXII-1 features contribute to unique inactive conformations of the C-helix via hydrophobic packing interactions with the ATP-binding C-helix glutamate (Glu103, Ala104, Met203, and Trp94) and by promoting a unique secondary structure of the β3-αC loop and C-helix (Pro95 and Asp96). These findings add to the seemingly limitless ways in which ePK superfamily members can evolve new

sequence features to affect kinase conformations and to ultimately modulate signaling outputs from the kinase domain.

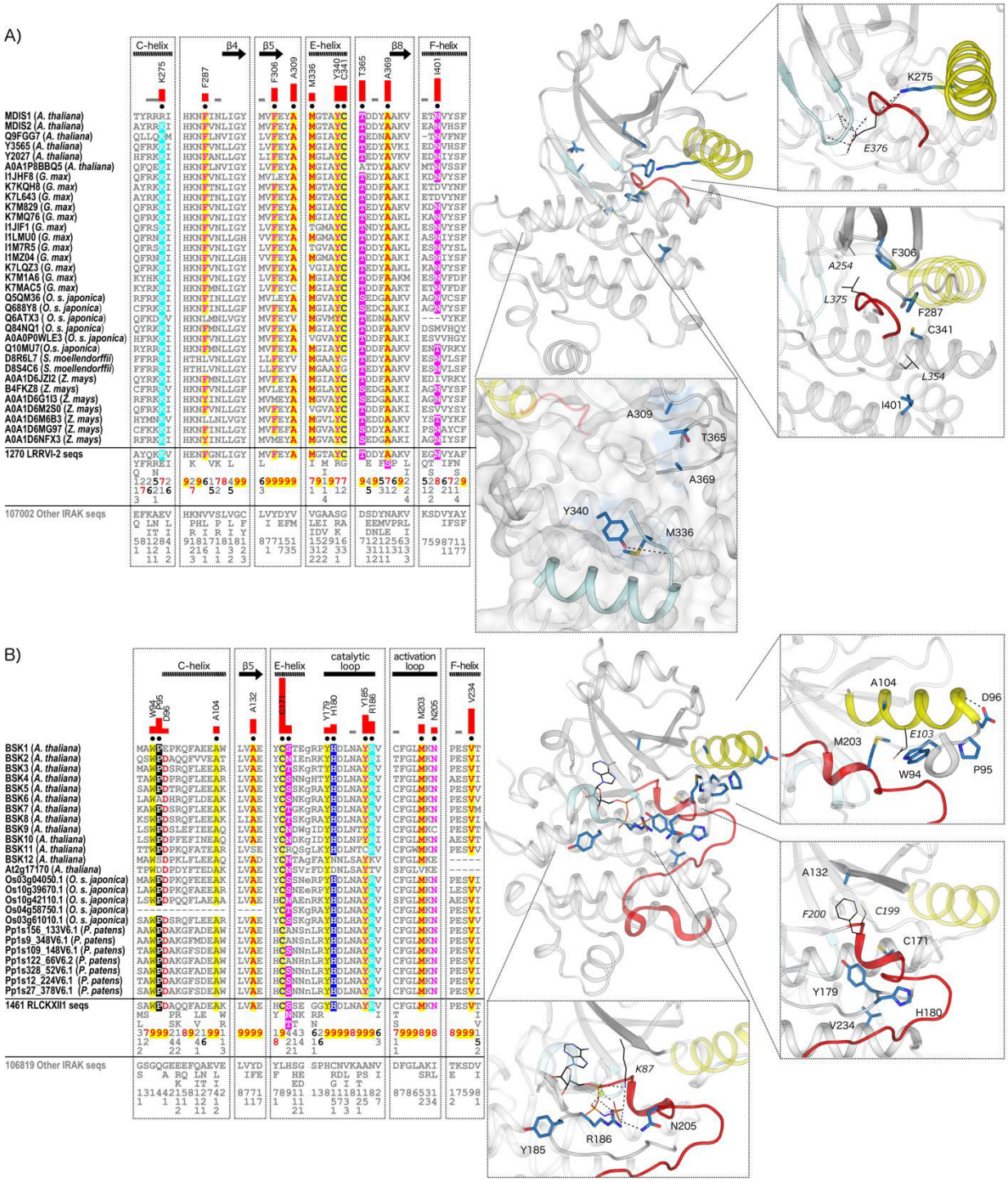

Figure 4.7. IRAK pseudokinase-specific features contribute to unique conformations in key catalytic regions. (A) LRRVI-2 pseudokinase family-specific sequence motifs. In the alignment, columns are highlighted where amino acids are highly conserved in LRRVI-2 pseudokinase family sequences and non-conserved and/or biochemically dissimilar in other IRAK sequences. Red bar lengths quantify the degree of divergence between LRRVI-2 and other IRAK sequences. Column-wise amino acid and insertion/deletion frequencies are indicated in integer tenths where a "5" indicates an occurrence of 50–60% in the given (weighted) sequence set. Columns used by the Bayesian partitioning procedure to sort LRRVI-2 sequences from other IRAK sequences are marked with black dots. Kinase secondary structures are annotated above the alignment. A structure of apo GRMZM2G135359 pseudokinase from *Zea mays* (PDB 6CPY) was used to analyze LRRVI-2-specific sequence motifs. In the structure, the glycine-rich loop is colored in light cyan, the C-helix in yellow, and the activation loop in red. Family-specific residues are shown in blue sticks. Residues occurring in canonical catalytic motifs are shown in black lines. Hydrogen bonds are shown in dashed black lines. (B) RLCKXII-1 pseudokinase family-specific sequence motifs. A structure of BSK8 from *Arabidopsis thaliana* (PDB 4I94) was used to analyze RLCKXII-1-specific sequence motifs.

4.3 Discussion

Using a large-scale bioinformatics analysis, we have considerably expanded the classification of pseudokinases. We identified a total of 86 putative pseudokinase families and demonstrated that pseudokinases are present across the tree of life. Our analysis strongly suggests that pseudokinases are polyphyletic, emerging through numerous events during the course of protein kinase evolution, presumably to fulfill different biological niches and signaling roles through non-catalytic functions, the broad extent of which is revealed by our analysis.

Pseudokinases have evolved in all the major ePK groups, however, the TKL group is particularly enriched with pseudokinases, largely due to expansions of TKL kinases in plants and fungi. These TKL group expansions occur in the IRAK family in plants, whereas TKL expansions in fungi, including R. irregularis, comprise distinct pseudokinase families unrelated to any known TKL families in other organisms, corroborating previous descriptions of 'unclassifiable' kinases in fungi (104, 105). Why have TKL's in particular been selected during evolution for such marked kinome and 129seudokinomes expansions in both plants and fungi? One possible explanation is the lack of TKs in these organisms, which, in metazoa, evolved and

duplicated to play crucial phosphotyrosine-dependent roles in multicellular signal transduction (106). Whereas tyrosine kinases comprise most of the receptor protein kinase repertoire in metazoans, the IRAK family of TKL's comprise a receptor kinase-like repertoire in plants, and thus the expansion of IRAK kinases and pseudokinases in plants may be analogous to the expansion of receptor tyrosine kinases in metazoa. In line with this view, LysM pseudokinases in Medicago are believed to contribute to Rhizobium interactions by interacting with active LysM RLK members, which is reminiscent of metazoan tyrosine pseudokinases, such as HER3, which allosterically modulates closely related, active EGFR family members (36, 69). Thus, the expansion of IRAK pseudokinases in plants mirrors the expansion of canonical IRAK kinases, perhaps due to regulatory interactions between co-evolved kinases and pseudokinases. Mechanistically, plant IRAK kinases are known to play vital roles in plant-fungi interactions, both during symbiotic interactions as well as during pathogen defense, suggesting that the expansion of fungal TKL kinases and pseudokinases may have arisen due to a close symbiotic co-evolution. R. irregularis participates in arbuscular mycorrhizal symbiotic relationships with more than two-thirds of all known plant species (65), and recent studies have evaluated the roles of the expanded R. irregularis proteome (107) as well as its kinome (81, 108) in this symbiotic relationship. Nevertheless, additional investigation of the contribution of R. irregularis pseudokinases in plant-fungal symbiosis is warranted. Furthermore, an understanding of how the symbiotic or infectious nature of host-pathogen interactions operate in the context of bacterial and eukaryotic pseudokinomes is likely to yield important information explaining how such relationships emerged and were propagated during the cellular 'wiring' of both physiological and pathological signaling pathways. As such, the prevalence of pseudokinases in some pathogenic protists such as Plasmodium falciparum and Giardia lamblia and in some bacteria suggest

possible roles in pathogenicity. An understanding of the extent and biological niches for symbiotic and pathogenic pseudokinases will also create further opportunities for pseudokinase-based targeting with small molecules in the future.

Examining the clustering of canonical kinase sequences within pseudokinase families suggests that some pseudokinase families include very closely-related, and potentially catalytically-active, members (Table 4.3). Although it is currently believed that pseudokinases most likely evolved from gene-duplicated canonical kinases (33), the clustering of canonical sequences within pseudokinase families does not rule out the possibility that some catalytically-active kinases might have evolved from 'pseudokinases', or other poorly defined ancestral non-enzymatic proteins, as recently reported for other pseudoenzymes (109-111). Indeed, some pseudokinase families have quite subtle chemical substitutions at the catalytic triad positions and might therefore be poised to 'revert' to a catalytically-active enzyme in response to a random mutagenic event or appropriate evolutionary pressure. An artificially-guided evolutionary pathway for reversion has been demonstrated for the relatively well-understood pseudokinase CASK, with five steps required to regenerate catalytic activity comparable to a canonical CAMK homolog (112). Likewise, pseudokinase families such as Rig3, HER3, STKLD1, and PSKH2 have conserved all canonical catalytic motifs except for the HRD-aspartate, which is normally substituted to asparagine; canonical members are likely to have evolved in these families by a simple substitution 'back' to the canonical HRD-aspartate. Alternatively, kinases can retain low (or very low) amounts of catalytic activity even with aspartate-to-asparagine substitutions in the HRD motif (36, 113, 114), suggesting that pseudokinase families with few relatively benign catalytic motif substitutions may sometimes represent low activity kinases with the capacity for signaling, and may explain why canonical sequences sometimes cluster within pseudokinase

families. A case for latent catalytic activity can also be made for pseudokinase families such as Act, DYD, and LanC, which conserve the catalytic triad residues except for a benign lysine-to-arginine substitution in the β3-strand. Nevertheless, the detection of pseudokinase families that have been evolutionarily retained across diverse species suggest that these predicted pseudokinases cannot be mere 'remnants of evolution', but rather that they play important biological roles, either through non-canonical, enzyme-based signaling or, in the majority of cases, via non-catalytic functions that await discovery.

Pseudokinases are likely to have evolved due to a relaxed constraint on usually invariant catalytic residues. However, in this study, our examination of the conserved sequence motifs associated with pseudokinase evolutionary divergence reveals variations not only in catalytic motifs, but also in conserved 'non-catalytic' regions distal from the pseudo-active site (Figure 4.7). For example, the observation of pseudokinase family-specific variations at the highly conserved F-helix aspartate suggests that this allosteric region is indeed important for kinase domain activation, as proposed in previous work (103, 115). As a result, future evaluation of pseudokinases may benefit from examining other sequence motifs that contribute to catalysis, including the glycine rich loop, C-helix glutamate, and the metal-coordinating asparagine, alongside regulatory motifs such as the catalytic and regulatory spines, which comprise allosteric networks distant from the active site. In addition, by comparing two plant IRAK pseudokinase families, we found that inactive kinase domain conformations are stabilized in divergent ways between different pseudokinase families through distinct sets of sequence motifs that have been selectively constrained during evolution. The conservation of pseudokinase family-specific motifs on the surface of the kinase domain further suggest that pseudokinase families have evolved unique interactions with other proteins domains or with flexible linker regions. To

evidence this, we identified a patch of LRRVI-2 family-specific residues on the surface of the pseudokinase domain that helps tether the C-terminal tail. In terms of regulation, one of the tethering residues is a tyrosine, whose phosphorylation could impart a switch-like function to alter tethering of the C-terminal tail, similar to that observed in the SRC family kinases (116), or for recruiting proteins via the binding of SH2 domains (1, 117). Regulatory tyrosine phosphorylation by dual-specificity LRR RLK's has recently been recognized in plants, where it likely represents a fundamental signaling role despite the lack of conventional tyrosine kinases encoded in plant kinomes (118-120). The tethering and untethering of flexible flanking linkers is also emerging as a common theme in canonical kinase regulation (55, 121-124) and for modulation of kinase-protein interactions (26, 55, 125, 126). Consistently, these flexible segments are often evolutionary hot-spots for neofunctionalization among signaling proteins (116, 127).

This comprehensive curated resource represents a comparative analysis of >30,000 pseudokinase sequences, which includes numerous representative species covering archaea, bacteria, protists, fungi, plants, and animals. It provides a new conceptual framework for characterizing pseudokinase evolution, and for the experimental dissection of pseudokinase-dependent lifestyles ranging across a very wide variety of model organisms. Our sub-classification of pseudokinases, which includes more than 30 previously unrecognized families, points to fundamental roles for pseudokinases in nearly all biological systems, where re-use of the versatile protein kinase fold has permitted a vast array of noncatalytic signaling mechanisms, many of which might be targeted therapeutically. Our data also represent a useful starting-point for the evaluation of other types of pseudoenzymes in diverse biological systems and sets the

stage for future evolutionary analyses of other pseudoenzyme families across the kingdoms of life.

4.4 Materials and Methods

4.4.1 Detection of pseudokinase sequences

Protein kinase sequences were extracted from the NCBI non-redundant (nr) (downloaded 4/4/18) and UniProt reference proteome databases (Release 2018_09) (59). Protein kinase sequences were identified and aligned using previously curated profiles of diverse eukaryotic and eukaryotic-like protein kinases (2, 52, 54, 57, 58) and the rapid and accurate alignment procedure MAPGAPS (56). Some sequences contained many low complexity regions (for example in P. falciparum and T. thermophila), and therefore a filter was used to mask these low complexity regions during eukaryotic kinase domain detection. Sequences that did not span from at least the β3-lysine to the G-helix (like many atypical kinases and small-molecule kinases) were deemed \

4.4.2 Bayesian pattern-based classification of pseudokinase sequences

We used the pseudokinomes extracted from the UniProt proteomes as well as additional diverse pseudokinase sequences extracted from the NCBI nr protein database as input into the optimal multiple-category Bayesian Partitioning with Pattern Selection algorithm (omcBPPS) (63, 64). To remove closely related sequences, we purged the nr pseudokinase sequences using 75% sequence identity cutoff (22,152 total nr pseudokinase sequences) and restrained the number of UniProt proteomes to 83 diverse representative archaeal, bacterial, and eukaryotic species (4,121 total UniProt sequences). A combined total of 26,273 distinct sequences of aligned pseudokinase domains was then used as an input for omcBPPS using a cluster size cut off of 50 sequences to identify the major sequence families(63, 64). Based on this clustering, we

initially identified 68 unique pseudokinase clusters. Some human pseudokinases were not classified by the algorithm due to the limited number of sequences (<50 sequences). In these cases, we ran separate clustering (omcBPPS analyses) within these groups using a minimum size of 15 sequences per cluster, which allowed us to further sub-classify these clusters. From these sub-classifications, we took only the clusters containing human pseudokinases in order to cover the entire human 135seudokinomes in our classification, yielding a total of 77 unique pseudokinase clusters in total.

The 77 pseudokinase clusters were incorporated within an existing hierarchical profile of ePK sequences, yielding 592 total ePK sequence profiles using MAPGAPS (2, 52, 57, 58). Using the resulting hierarchical sequence profile, we then classified all sequences from each of the 10,092 proteomes available in the UniProt proteomes database to all kinase/pseudokinase families in the profile. Pseudokinase sequences that did not classify into the 77 pseudokinase clusters were placed within one of the 9 unclassified pseudokinase families based on sequence similarity to the major kinase groups (i.e. TK-, TKL-, STE-, CK1-, AGC-, CAMK-, CMGC-, Other-, and PknB-unclassified groups), resulting in a total of 86 pseudokinase families. For 38 pseudokinase families, canonical kinase sequences clustered into the pseudokinase family. For these cases, canonical sequences were removed from the pseudokinase sequence set and are noted in Table 4.3. Pseudokinase family alignments were manually evaluated for possible misalignments.

4.4.3 Phylogenetic tree building

To understand the relationships between the identified pseudokinase families, Hidden Markov Models (HMMs) were built using alignments for each family, and HMM-to-HMM distances were computed. From these distances, a distance matrix was created to build a

neighbor joining tree, which was used to approximate the distances of families shown in Figure 4.3. This method was implemented using pHMM-Tree (128).

The IRAK, R.irregularis, and PknB phylogenetic trees were built from aligned kinase domains using FastTree version 2.1.10 using default settings, which implements the JTT ML model and calculates local support values for internal nodes via the Shimodaira-Hasegawa test (129, 130). The rooted IRAK phylogenetic tree was created using a total of 7,647 diverse plant and metazoan IRAK sequences. These sequences included all plant IRAK pseudokinase sequences (5,405 sequences), canonical plant kinase sequences purged at 80% sequence identity (1,894 sequences), metazoan IRAK pseudokinases purged at 90% sequence identity (113 sequences), and canonical metazoan kinases purged at 80% sequence identity (145 sequences). Also included were diverse non-IRAK TKL sequences (41 sequences) and non-TKL kinase sequences (49 sequences), which were used as an outgroup to root the tree. The unrooted R. irregularis phylogenetic tree was created using all 787 members of its kinome. The unrooted PknB phylogenetic tree was created using 769 representative pseudokinase sequences assigned to a PknB-related family (removed 376 divergent sequences with conservation log odds score <-10) in addition to a set of representative canonical PknB kinase sequences (511 sequences). iTOL was used to generate the final trees (131).

4.4.4 Weblogo creation

Weblogos were created using version 3.6 of the WebLogo 3 online server (132). Amino acids were colored according to their biochemical properties: basic in blue, acidic in red, amide groups in purple, nonpolar residues in black, and polar or uncharged residues in green.

4.4.5 Determination of domain organizations

Additional domains for each of the IRAK, R. irregularis, and PknB-related pseudokinase families were initially identified using NCBI's Batch Web CD-Search Tool against the CDD database with an expected value threshold of 0.01 and a maximum of 500 hits per CD-search (133, 134). Transmembrane (TM) helices were identified using TMHMM 2.0 (135) and the Phobius web server (136). TM-helices were annotated as such only if both TMHMM and Phobius predicted the same TM-helix region within a 10-residue margin. We verified kinase domain hits using our manually curated protein kinase profiles and removed any overlapping domain predictions.

4.4.6 Pattern analysis of plant IRAK families

To detect sequence motifs associated with the evolution of IRAK pseudokinase families, we used sequence sets from the omcBPPS classification of IRAK pseudokinase families as seed alignments and used mcBPPS (137) to optimally partition 136,068 diverse IRAK sequences extracted from the NCBI nr database (including active sequences) into either pseudokinase family sets or a "background" set that includes unclassified IRAK sequences. Some pseudokinase families have very close active homologs, thus separate seed alignments for canonical IRAK families were also included to ensure that pseudokinase partitions did not include active members. mcBPPS identifies the amino acid patterns that most distinguishes each pseudokinase partition from other sequences, and the 30 most statistically significant patterns for each family were analyzed using available crystal structures.

4.4.7 Analysis of fungal proteomes and kinomes

To compare the kinome and proteome sizes across different fungal species, we analyzed a total of 448 fungal proteomes obtained from the EnsemblFungi database (138). Kinases and pseudokinases were identified using previously curated multiple protein alignment profiles and examination of the β3-lysine, HRD-aspartate, and DFG-aspartate positions, as detailed above.

Bibliography

1. P. Cohen, The origins of protein phosphorylation. *Nat Cell Biol* **4**, E127-130 (2002).
2. N. Kannan, S. S. Taylor, Y. Zhai, J. C. Venter, G. Manning, Structural and functional diversity of the microbial kinome. *PLoS Biol* **5**, e17 (2007).
3. D. R. Knighton *et al.*, Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* **253**, 407-414 (1991).
4. D. R. Knighton *et al.*, Structure of a peptide inhibitor bound to the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* **253**, 414-420 (1991).
5. E. D. Scheeff, J. Eswaran, G. Bunkoczi, S. Knapp, G. Manning, Structure of the pseudokinase VRK3 reveals a degraded catalytic site, a highly conserved kinase fold, and a putative regulatory binding site. *Structure* **17**, 128-138 (2009).
6. J. A. Adams, Kinetic and catalytic mechanisms of protein kinases. *Chem Rev* **101**, 2271-2290 (2001).
7. S. K. Hanks, A. M. Quinn, T. Hunter, The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science* **241**, 42-52 (1988).
8. S. K. H. Hanks, T.;, The eukaryotic protein kinase superfamily-kinase catalytic domain structure and classification *FASEB journal* **9**, 576-596 (1995).
9. J. M. Murphy, P. D. Mace, P. A. Eyers, Live and let die: insights into pseudoenzyme mechanisms from structure. *Curr Opin Struct Biol* **47**, 95-104 (2017).
10. E. D. Scheeff, P. E. Bourne, Structural evolution of the protein kinase-like superfamily. *PLoS Comput Biol* **1**, e49 (2005).
11. M. Kostich *et al.*, Human members of the eukaryotic protein kinase family. *Genome Biol* **3**, RESEARCH0043 (2002).

12. L. J. Wilson *et al.*, New Perspectives, Opportunities, and Challenges in Exploring the Human Protein Kinome. *Cancer Res* **78**, 15-29 (2018).

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

14. G. Manning, S. L. Young, W. T. Miller, Y. Zhai, 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**, 9674-9679 (2008).

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

16. G. Manning *et al.*, The minimal kinome of Giardia lamblia illuminates early kinase evolution and unique parasite biology. *Genome Biol* **12**, R66 (2011).

17. M. Zulawski, The Arabidopsis Kinome: phylogeny and evolutionary insights into functional diversification. *BMC genomics* **15**, (2014).

18. G. Manning, D. B. Whyte, R. Martinez, T. Hunter, S. Sudarsanam, The protein kinase complement of the human genome. *Science* **298**, 1912-1934 (2002).

19. J. Boudeau, D. Miranda-Saavedra, G. J. Barton, D. R. Alessi, Emerging roles of pseudokinases. *Trends Cell Biol* **16**, 443-452 (2006).

20. Patrick A. Eyers, J. M. Murphy, Dawn of the dead: protein pseudokinases signal new adventures in cell biology. *Biochemical Society Transactions* **41**, 969-974 (2013).

21. V. Reiterer, P. A. Eyers, H. Farhan, Day of the dead: pseudokinases and pseudophosphatases in physiology and disease. *Trends Cell Biol* **24**, 489-505 (2014).

22. J. M. Murphy *et al.*, A robust methodology to subclassify pseudokinases based on their nucleotide-binding properties. *Biochem J* **457**, 323-334 (2014).

23. J. M. Mendrola, F. Shi, J. H. Park, M. A. Lemmon, Receptor tyrosine kinases with intracellular pseudokinase domains. *Biochem Soc Trans* **41**, 1029-1036 (2013).

24. N. S. Dhawan, A. P. Scopton, A. C. Dar, Small molecule stabilization of the KSR inactive state antagonizes oncogenic Ras signalling. *Nature* **537**, 112-116 (2016).

25. J. J. Babon, I. S. Lucet, J. M. Murphy, N. A. Nicola, L. N. Varghese, The molecular regulation of Janus kinase (JAK) activation. *Biochem J* **462**, 1-13 (2014).

26. P. A. Eyers, K. Keeshan, N. Kannan, Tribbles in the 21st Century: The Evolving Roles of Tribbles Pseudokinases in Biology and Disease. *Trends Cell Biol* **27**, 284-298 (2017).

27. E. Zeqiraj, D. M. van Aalten, Pseudokinases-remnants of evolution or key allosteric regulators? *Curr Opin Struct Biol* **20**, 772-781 (2010).

28. M. Sierla *et al.*, The Receptor-like Pseudokinase GHR1 Is Required for Stomatal Closure. *Plant Cell* **30**, 2813-2837 (2018).

29. H. Zhang *et al.*, Structure and evolution of the Fam20 kinases. *Nat Commun* **9**, 1218 (2018).

30. C. Lecointre *et al.*, Dimerization of the Pragmin Pseudo-Kinase Regulates Protein Tyrosine Phosphorylation. *Structure* **26**, 545-554 e544 (2018).

31. M. J. Chen, J. E. Dixon, G. Manning, Genomics and evolution of protein phosphatases. *Sci Signal* **10**, (2017).

32. M. Zettl, C. Adrain, K. Strisovsky, V. Lastun, M. Freeman, Rhomboid family pseudoproteases use the ER quality control machinery to regulate intercellular signaling. *Cell* **145**, 79-91 (2011).

33. P. A. Eyers, J. M. Murphy, The evolving world of pseudoenzymes: proteins, prejudice and zombies. *BMC Biol* **14**, 98 (2016).

34. J. M. Murphy, H. Farhan, P. A. Eyers, Bio-Zombie: the rise of pseudoenzymes in biology. *Biochemical Society Transactions* **45**, 537-544 (2017).

35. C. J. Novotny *et al.*, Overcoming resistance to HER2 inhibitors through state-specific kinase binding. *Nat Chem Biol* **12**, 923-930 (2016).

36. F. Shi, S. E. Telesco, Y. Liu, R. Radhakrishnan, M. A. Lemmon, ErbB3/HER3 intracellular domain is competent to bind ATP and catalyze autophosphorylation. *Proc Natl Acad Sci U S A* **107**, 7692-7697 (2010).

37. B. Xu *et al.*, WNK1, a novel mammalian serine/threonine protein kinase lacking the catalytic lysine in subdomain II. *J Biol Chem* **275**, 16795-16801 (2000).

38. A. Y. Lee *et al.*, Protein kinase WNK3 regulates the neuronal splicing factor Fox-1. *Proc Natl Acad Sci U S A* **109**, 16841-16846 (2012).

39. R. Ahlstrom, A. S. Yu, Characterization of the kinase activity of a WNK4 protein complex. *Am J Physiol Renal Physiol* **297**, F685-692 (2009).

40. P. A. Eyers, TRIBBLES: A Twist in the Pseudokinase Tail. *Structure* **23**, 1974-1976 (2015).

41. S. Uljon *et al.*, Structural Basis for Substrate Selectivity of the E3 Ligase COP1. *Structure* **24**, 687-696 (2016).

42. S. A. Jamieson *et al.*, Substrate binding allosterically relieves autoinhibition of the pseudokinase TRIB1. *Sci Signal* **11**, (2018).

43. J. B. Sheetz, M. A. Lemmon, Flipping ATP to AMPlify Kinase Functions. *Cell* **175**, 641-642 (2018).

44. A. Sreelatha *et al.*, Protein AMPylation by an Evolutionarily Conserved Pseudokinase. *Cell* **175**, 809-821 e819 (2018).

45. N. V. Sergina *et al.*, Escape from HER-family tyrosine kinase inhibitor therapy by the kinase-inactive HER3. *Nature* **445**, 437-441 (2007).

46. J. Claus *et al.*, Inhibitor-induced HER2-HER3 heterodimerisation promotes proliferation through a novel dimer interface. *Elife* **7**, (2018).

47. F. P. Bailey *et al.*, The Tribbles 2 (TRB2) pseudokinase binds to ATP and autophosphorylates in a metal-independent manner. *Biochem J* **467**, 47-62 (2015).

48. G. Labesse *et al.*, ROP2 from Toxoplasma gondii: a virulence factor with a protein-kinase fold and no enzymatic activity. *Structure* **17**, 139-146 (2009).

49. B. S. Blaum *et al.*, Structure of the pseudokinase domain of BIR2, a regulator of BAK1-mediated immune signaling in Arabidopsis. *J Struct Biol* **186**, 112-121 (2014).

50. L. A. Gish, S. E. Clark, The RLK/Pelle family of kinases. *Plant J* **66**, 117-127 (2011).

51. J. D. Lewis *et al.*, The Arabidopsis ZED1 pseudokinase is required for ZAR1-mediated immunity induced by the Pseudomonas syringae type III effector HopZ1a. *Proc Natl Acad Sci U S A* **110**, 18722-18727 (2013).

52. M. D. Lehti-Shiu, S. H. Shiu, Diversity, classification and function of the plant protein kinase superfamily. *Philos Trans R Soc Lond B Biol Sci* **367**, 2619-2639 (2012).

53. P. L. Liu, L. Du, Y. Huang, S. M. Gao, M. Yu, Origin and diversification of leucine-rich repeat receptor-like protein kinase (LRR-RLK) genes in plants. *BMC Evol Biol* **17**, 47 (2017).

54. A. Bateman *et al.*, The Pfam protein families database. *Nucleic Acids Res* **32**, D138-141 (2004).

55. N. Kannan, N. Haste, S. S. Taylor, A. F. Neuwald, The hallmark of AGC kinase functional divergence is its C-terminal tail, a cis-acting regulatory module. *Proc Natl Acad Sci U S A* **104**, 1272-1277 (2007).

56. A. F. Neuwald, Rapid detection, classification and accurate alignment of up to a million or more related protein sequences. *Bioinformatics* **25**, 1869-1875 (2009).

57. D. I. McSkimming *et al.*, KinView: a visual comparative sequence analysis tool for integrated kinome research. *Mol Biosyst* **12**, 3651-3665 (2016).

58. E. Talevich, A. Mirza, N. Kannan, Structural and evolutionary divergence of eukaryotic protein kinases in Apicomplexa. *BMC Evol Biol* **11**, 321 (2011).

59. C. UniProt, UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204-212 (2015).

60. O. Patel *et al.*, Structure of SgK223 pseudokinase reveals novel mechanisms of homotypic and heterotypic association. *Nat Commun* **8**, 1157 (2017).

61. S. B. McMahon, H. A. Van Buskirk, K. A. Dugan, T. D. Copeland, M. D. Cole, The novel ATM-related protein TRRAP is an essential cofactor for the c-Myc and E2F oncoproteins. *Cell* **94**, 363-374 (1998).

62. P. Ward, L. Equinet, J. Packer, C. Doerig, Protein kinases of the human malaria parasite Plasmodium falciparum: the kinome of a divergent eukaryote. *BMC Genomics* **5**, 79 (2004).

63. A. F. Neuwald, A Bayesian sampler for optimization of protein domain hierarchies. *J Comput Biol* **21**, 269-286 (2014).

64. A. F. Neuwald, Evaluating, comparing, and interpreting protein domain hierarchies. *J Comput Biol* **21**, 287-302 (2014).

65. E. Tisserant *et al.*, Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 20117-20122 (2013).

66. S. L. Sierke, K. Cheng, H. H. Kim, J. G. Koland, Biochemical characterization of the protein tyrosine kinase homology domain of the ErbB3 (HER3) receptor protein. *Biochem J* **322 (Pt 3)**, 757-763 (1997).

67. A. Miya *et al.*, CERK1, a LysM receptor kinase, is essential for chitin elicitor signaling in Arabidopsis. *Proc Natl Acad Sci U S A* **104**, 19613-19618 (2007).

68. S. Radutoiu *et al.*, Plant recognition of symbiotic bacteria requires two LysM receptor-like kinases. *Nature* **425**, 585-592 (2003).

69. A. Pietraszewska-Bogiel *et al.*, Interaction of Medicago truncatula lysin motif receptor-like kinases, NFP and LYK3, produced in Nicotiana benthamiana induces defence-like responses. *PLoS One* **8**, e65055 (2013).

70. B. Eyuboglu *et al.*, Molecular characterisation of the STRUBBELIG-RECEPTOR FAMILY of genes encoding putative leucine-rich repeat receptor-like kinases in Arabidopsis thaliana. *BMC Plant Biol* **7**, 16 (2007).

71. D. Chevalier *et al.*, STRUBBELIG defines a receptor kinase-mediated signaling pathway regulating organ development in Arabidopsis. *Proceedings of the National Academy of Sciences* **102**, 9074-9079 (2005).

72. T. Miyakawa, K. Miyazono, Y. Sawano, K. Hatano, M. Tanokura, Crystal structure of ginkbilobin-2 with homology to the extracellular domain of plant cysteine-rich receptor-like kinases. *Proteins* **77**, 247-251 (2009).

73. L. Zhang *et al.*, Identification of an apoplastic protein involved in the initial phase of salt stress response in rice root by two-dimensional electrophoresis. *Plant Physiol* **149**, 916-928 (2009).

74. G. Bourdais *et al.*, Large-Scale Phenomics Identifies Primary and Fine-Tuning Roles for CRKs in Responses Related to Oxidative Stress. *PLoS Genet* **11**, e1005373 (2015).

75. K. Sklodowski *et al.*, The receptor-like pseudokinase MRH1 interacts with the voltage-gated potassium channel AKT2. *Sci Rep* **7**, 44611 (2017).

76. C. Huard-Chauveau *et al.*, An atypical kinase under balancing selection confers broad-spectrum disease resistance in Arabidopsis. *PLoS Genet* **9**, e1003766 (2013).

77. C. Grutter, S. Sreeramulu, G. Sessa, D. Rauh, Structural characterization of the RLCK family member BSK8: a pseudokinase with an unprecedented architecture. *J Mol Biol* **425**, 4455-4467 (2013).

78. M. Gao *et al.*, Regulation of cell death and innate immunity by two receptor-like kinases in Arabidopsis. *Cell Host Microbe* **6**, 34-44 (2009).

79. Z. L. Nimchuk, P. T. Tarr, E. M. Meyerowitz, An evolutionarily conserved pseudokinase mediates stem cell production in plants. *Plant Cell* **23**, 851-854 (2011).

80. A. Kuo, A. Kohler, F. M. Martin, I. V. Grigoriev, Expanding genomics of mycorrhizal symbiosis. *Front Microbiol* **5**, 582 (2014).

81. Y. Handa *et al.*, RNA-seq Transcriptional Profiling of an Arbuscular Mycorrhiza Provides Insights into Regulated and Coordinated Gene Expression in Lotus japonicus and Rhizophagus irregularis. *Plant Cell Physiol* **56**, 1490-1511 (2015).

82. C. E. Haydon *et al.*, Identification of novel phosphorylation sites on Xenopus laevis Aurora A and analysis of phosphopeptide enrichment by immobilized metal-affinity chromatography. *Mol Cell Proteomics* **2**, 1055-1067 (2003).

83. M. Grynberg, A. Godzik, NERD: a DNA processing-related domain present in the anthrax virulence plasmid, pXO1. *Trends Biochem Sci* **29**, 106-110 (2004).

84. P. S. Sumby, M.C. , Genetics of the phage growth limitation (Pgl) system ofStreptomyces coelicolor A3(2) *Molecular microbiology* **44**, 489-500 (2002).

85. P. A. S. Hoskisson, P.; Smith, M., The phage growth limitation system in Streptomyces coelicolor A(3)2 is a toxin/antitoxin system, comprising enzymes with DNA methyltransferase, protein kinase and ATPase activity. *Virology*, (2015).

86. R. H. Ebright, S. Busby, The Escherichia coli RNA polymerase alpha subunit: structure and function. *Curr Opin Genet Dev* **5**, 197-203 (1995).

87. K. Aoyama, H. Aiba, T. Mizuno, Genetic analysis of the His-to-Asp phosphorelay implicated in mitotic cell cycle control: involvement of histidine-kinase genes of Schizosaccharomyces pombe. *Biosci Biotechnol Biochem* **65**, 2347-2352 (2001).

88. M. K. Ashby, J. Houmard, Cyanobacterial two-component proteins: structure, diversity, distribution, and evolution. *Microbiol Mol Biol Rev* **70**, 472-509 (2006).

89. J. A. Calera, G. H. Choi, R. A. Calderone, Identification of a putative histidine kinase two-component phosphorelay gene (CaHK1) in Candida albicans. *Yeast* **14**, 665-674 (1998).

90. A. S. Krupa, N., Diversity in domain architectures of Ser/Thr kinases and their homologues in prokaryotes. *BMC genomics* **6**, 129 (2005).

91. V. L. Phalip, J.; Zhang, C., HistK a cyanobacterial protein with both a serine/threonine kinase domain and a histidine kinase domain: implication for the mechanism of signal transduction. *The biochemical journal* **360**, 639-644 (2001).

92. X. Zhang *et al.*, Genome-wide survey of putative serine/threonine protein kinases in cyanobacteria. *BMC Genomics* **8**, 395 (2007).

93. V. Buck *et al.*, Peroxide sensors for the fission yeast stress-activated mitogen-activated protein kinase pathway. *Mol Biol Cell* **12**, 407-419 (2001).

94. J. A. Calera, X. J. Zhao, F. De Bernardis, M. Sheridan, R. Calderone, Avirulence of Candida albicans CaHK1 mutants in a murine model of hematogenously disseminated candidiasis. *Infect Immun* **67**, 4280-4284 (1999).

95. J. A. C. Calera, R., Flocculation of hyphae is associated with a deletion in the putative CaHKl two-component histidine kinase gene from Candida albicans *Microbiology* **145**, 1431-1442 (1999).

96. Y. Cheng *et al.*, A pair of iron-responsive genes encoding protein kinases with a Ser/Thr kinase domain and a His kinase domain are regulated by NtcA in the Cyanobacterium Anabaena sp. strain PCC 7120. *J Bacteriol* **188**, 4822-4829 (2006).

97. M. Kruppa *et al.*, The role of the Candida albicans histidine kinase [CHK1) gene in the regulation of cell wall mannan and glucan biosynthesis. *FEMS Yeast Res* **3**, 289-299 (2003).

98. J. Quinn *et al.*, Two-component mediated peroxide sensing and signal transduction in fission yeast. *Antioxid Redox Signal* **15**, 153-165 (2011).

99. L. Shi *et al.*, Two genes encoding protein kinases of the HstK family are involved in synthesis of the minor heterocyst-specific glycolipid in the cyanobacterium Anabaena sp. strain PCC 7120. *J Bacteriol* **189**, 5075-5081 (2007).

100. A. Torosantucci, Deletion of the Two-Component Histidine Kinase Gene (CHK1) of Candida albicans Contributes to Enhanced Growth Inhibition and Killing by Human Neutrophils In Vitro. *Infection and Immunity* **70**, 985-987 (2002).

101. T. Yamada-Okabe *et al.*, Roles of three histidine kinase genes in hyphal development and virulence of the pathogenic fungus Candida albicans. *J Bacteriol* **181**, 7243-7247 (1999).

102. M. A. Jones, M. J. Raymond, N. Smirnoff, Analysis of the root-hair morphogenesis transcriptome reveals the molecular identity of six genes with roles in root-hair development in Arabidopsis. *Plant J* **45**, 83-100 (2006).

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

104. T. Hunter, G. D. Plowman, The protein kinases of budding yeast: six score and more. *Trends Biochem Sci* **22**, 18-22 (1997).

105. J. E. Stajich *et al.*, Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom Coprinopsis cinerea (Coprinus cinereus). *Proc Natl Acad Sci U S A* **107**, 11889-11894 (2010).

106. W. T. Miller, Tyrosine kinase signaling and the emergence of multicellularity. *Biochim Biophys Acta* **1823**, 1053-1057 (2012).

107. G. D. A. Werner, Y. Zhou, C. M. J. Pieterse, E. T. Kiers, Tracking plant preference for higher-quality mycorrhizal symbionts under varying CO2 conditions over multiple generations. *Ecol Evol* **8**, 78-87 (2018).

108. K. Nanjareddy, M. K. Arthikala, B. M. Gomez, L. Blanco, M. Lara, Differentially expressed genes in mycorrhized and nodulated roots of common bean are associated with defense, cell wall architecture, N metabolism, and P metabolism. *PLoS One* **12**, e0182328 (2017).

109. A. E. O. Todd, C.A.; Thornton, J.M., Sequence and structural differences between enzyme and nonenzyme homologs. *Structure* **10**, 1435-1451 (2002).

110. M. Kaltenbach *et al.*, Evolution of chalcone isomerase from a noncatalytic ancestor. *Nat Chem Biol* **14**, 548-555 (2018).

111. B. E. Clifton *et al.*, Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein. *Nature Chemical Biology* **14**, 542-547 (2018).

112. K. Mukherjee, M. Sharma, R. Jahn, M. C. Wahl, T. C. Sudhof, Evolution of CASK into a Mg2+-sensitive kinase. *Sci Signal* **3**, ra33 (2010).

113. D. M. Williams, P. A. Cole, Proton demand inversion in a mutant protein tyrosine kinase reaction. *J Am Chem Soc* **124**, 5956-5957 (2002).

114. V. T. Skamnaki *et al.*, Catalytic mechanism of phosphorylase kinase probed by mutational studies. *Biochemistry* **38**, 14718-14730 (1999).

115. S. Mohanty *et al.*, Hydrophobic Core Variations Provide a Structural Framework for Tyrosine Kinase Evolution and Functional Specialization. *PLoS Genet* **12**, e1005885 (2016).

116. S. C. Harrison, Variation on an Src-like theme. *Cell* **112**, 737-740 (2003).

117. J. G. Williams, M. Zvelebil, SH2 domains in plants imply new signalling scenarios. *Trends Plant Sci* **9**, 161-163 (2004).

118. M. H. Oh *et al.*, Tyrosine phosphorylation of the BRI1 receptor kinase emerges as a component of brassinosteroid signaling in Arabidopsis. *Proc Natl Acad Sci U S A* **106**, 658-663 (2009).

119. S. Luan, Tyrosine phosphorylation in plant cell signaling. *Proc Natl Acad Sci U S A* **99**, 11567-11569 (2002).

120. A. Perraki *et al.*, Phosphocode-dependent functional dichotomy of a common co-receptor in plant signalling. *Nature*, (2018).

121. S. R. Hubbard, Juxtamembrane autoinhibition in receptor tyrosine kinases. *Nat Rev Mol Cell Biol* **5**, 464-471 (2004).

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

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

124. I. Plaza-Menacho *et al.*, RET Functions as a Dual-Specificity Kinase that Requires Allosteric Inputs from Juxtamembrane Elements. *Cell Rep* **17**, 3319-3332 (2016).

125. K. S. Gajiwala, EGFR: tale of the C-terminal tail. *Protein Sci* **22**, 995-999 (2013).

126. B. Ma, C. J. Tsai, T. Haliloglu, R. Nussinov, Dynamic allostery: linkers are not merely flexible. *Structure* **19**, 907-917 (2011).

127. C. J. Brown, A. K. Johnson, A. K. Dunker, G. W. Daughdrill, Evolution and disorder. *Curr Opin Struct Biol* **21**, 441-446 (2011).

128. L. Huo *et al.*, pHMM-tree: phylogeny of profile hidden Markov models. *Bioinformatics* **33**, 1093-1095 (2017).

129. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**, 1641-1650 (2009).

130. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

131. I. Letunic, P. Bork, Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **44**, W242-245 (2016).

132. G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190 (2004).

133. A. Marchler-Bauer *et al.*, CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res* **45**, D200-D203 (2017).

134. A. Marchler-Bauer *et al.*, CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* **39**, D225-229 (2011).

135. A. Krogh, B. Larsson, G. von Heijne, E. L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567-580 (2001).

136. L. Kall, A. Krogh, E. L. Sonnhammer, Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res* **35**, W429-432 (2007).

137. A. F. Neuwald, Surveying the manifold divergence of an entire protein class for statistical clues to underlying biochemical mechanisms. *Stat Appl Genet Mol Biol* **10**, Article 36 (2011).

138. P. J. Kersey *et al.*, Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res* **46**, D802-D808 (2018).

Chapter 5

Discussion and Concluding remarks

5.1 Achievement of goals

I have applied computational and experimental approaches to investigate each of the research questions stated in the first chapter. The studies presented here provide novel insights into tyrosine kinase functions by identifying and characterizing key sequence motifs that have been selected across various branches of tyrosine kinase evolution. Through these contributions I have newly defined three tyrosine kinase subgroups, identified a new mechanism for allosteric regulation in Ephrin kinases, and investigated the evolution of pseudokinases. The work provides new insights to deepen our understanding of how the diverse functions of tyrosine kinases have evolved, and it demonstrates how similar approaches may be used to understand the evolution of other protein families.

5.1.1 New hierarchical classification of the tyrosine kinome

The rearrangement of tyrosine kinase families into a new hierarchical, pattern-based classification revealed the ancient conservation of sequence motifs defining three major subgroups of tyrosine kinase families. The study represents the first major attempt to reclassify

the tyrosine kinome since the establishment of Manning's classification in the early 2000s (*1-3*). The resulting classification, as I demonstrate, can guide informative sequence comparison studies, which I used reveal the modular evolution of regulatory cores in unique to different tyrosine kinase families. Recent collaborative work on the Src module kinase and Tec family kinase, BTK, experimentally determined that Tec family-specific features build upon the common Src module regulatory core by contributing to a PHTH domain-binding interface that allows PHTH-mediated autoinhibition unique to the Tec family (*4*).

5.1.2 Allosteric regulation in the Ephrin kinases

The juxtamembrane plays an important regulatory role in Eph kinases, as it does in many receptor tyrosine kinases (*5-7*). My work on the Eph family showed that the C-terminal flanking segment, the SAM domain linker, also plays an important regulatory role in Eph activation through allosteric coupling to the juxtamembrane. Using a combination of molecular dynamics simulations and experimental approaches, I demonstrated that mutations in the network tethering the SAM linker onto the C-terminal lobe allosterically affected kinase activity by modulating the dynamics of the juxtamembrane and its rate of autophosphorylation. Soon after the publication of this work in the *Journal of Biological Chemistry*, one group developed a hetero-bivalent kinase inhibitor that targeted the SAM linker tethering interface of the Eph kinase, EphA3, suggesting that this allosteric site can be used to increase the selectivity of Eph kinase inhibitors to modulate Eph activity in unique ways (*8*).

5.1.3 Evolution of pseudokinases

The comprehensive investigation of pseudokinase sequences across diverse proteomes showed that pseudokinases are prevalent across the tree of life, demonstrating that pseudokinases

are present in archaea for the first time. The detection of pseudokinase families homologous to diverse canonical protein kinase families and across diverse taxonomic groups reflect that pseudokinases have emerged numerous times during protein kinase evolution. This is captured in a new classification of pseudokinase sequences that place them into over eighty pseudokinase families in the context of canonical kinase families. An examination of pseudokinase sequences illustrates that pseudokinases evolve not only through the degeneration of key catalytic motifs, but also through the selection of amino acids distal to the active site to deactivate the kinase domain and/or contribute to non-catalytic functions. Recent work to experimentally characterize pseudokinases in plants and fungi (*9-11*) highlight the important and largely undiscovered roles of pseudokinases outside of humans and other vertebrates, and thus this study represents an important resource that catalogues and organizes pseudokinase data from diverse organisms.

5.2 Future directions

The findings presented in these studies shed novel insights into the sequence, structure, function, and evolution of tyrosine kinases. In light of these contributions, I also highlight several avenues for future investigation.

5.2.1 Annotation of motifs in an ontological framework

Our identification and characterization of sequence motifs distinguishing different branches of tyrosine kinase evolution in the studies presented here demonstrate how analyzing these sequence motifs can lead to new insights on various tyrosine kinase functions. Furthermore, decades of experimental research on protein kinases have provided extensive data on the regulatory functions in various tyrosine kinases and the specific amino acids that contribute to these functions, however, these data are scattered across disparate literature sources

and difficult to piece together or mine in a large-scale, systematic manner. In order to organize amino acid-level information on protein kinases in a manner that would be insightful to other researchers, we have developed an annotation scheme to systematically annotate the functions of sequence motifs in protein kinases, as well as an ontological framework to make sequence motif and annotation data publicly available.

While many existing databases and ontologies provide annotated information of protein functions such as post-translational modifications, biological pathways and reactions, gene expression, and subcellular location, the detailed analyses of individual amino acids in protein kinases called for a standardized vocabulary to describe residue-level functions particularly pertaining to protein kinases. To this end, I have developed a structured, kinase-centric vocabulary scheme that encompasses all known functions of individual residues in protein kinases (Figure 5.1). The vocabulary scheme has a hierarchical format, wherein the highest-level residue-level functions include catalysis, which would describe residues such as the catalytic HRD-aspartate (*12*), PTM-site, molecular interaction, positive regulation, and negative regulation. I include positive and negative regulation as residue-level functions in order to broadly annotate residues that participate in inhibitory or activating mechanisms that regulate catalytic function, which will additionally be useful to translate information from mutagenesis studies that identify a loss or gain of activity as a result of a mutation (*5, 13*). Such annotations can be potentially useful to identify potential oncogenic mutations (*14*). Vocabulary terms under molecular interaction include intermolecular interaction descriptors such as metal binding (i.e. DFG-aspartate), ATP binding (i.e. β3-lysine), and protein-protein interaction (i.e. substrate binding), as well as intramolecular interaction descriptors, which can be used to describe

common kinase-specific regulatory conformations such as those involving the αC-helix, activation loop, or interlobe hinge (*15* , *16*).

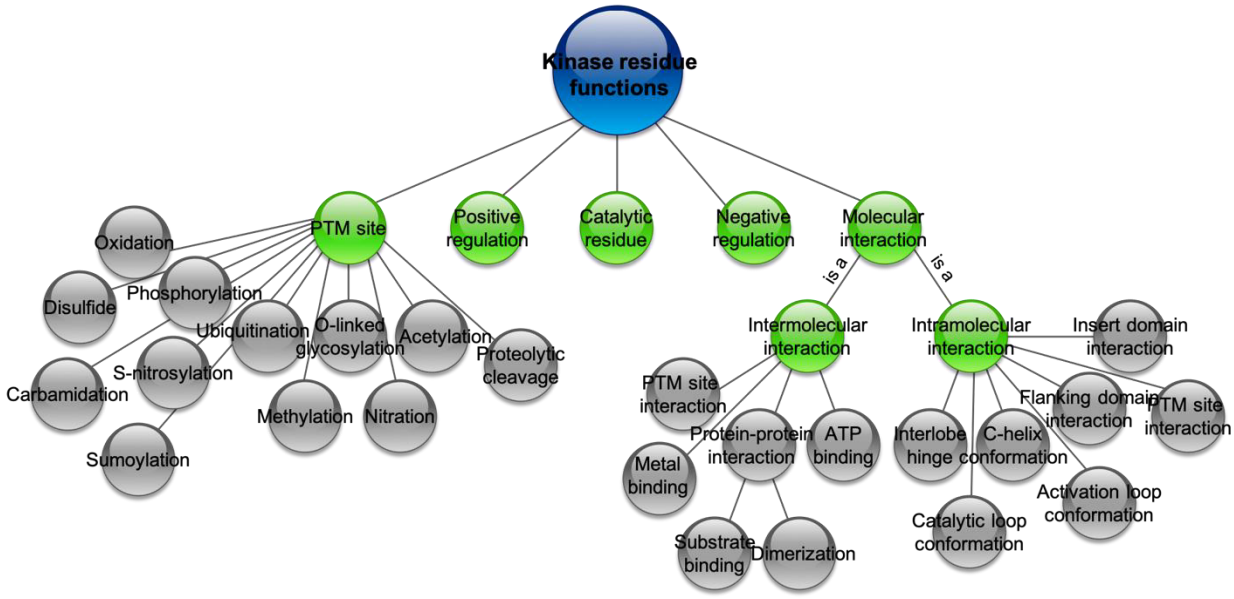

Figure 5.1: A structured, kinase-centric vocabulary to annotate residue functions in protein kinases.

Using this structured vocabulary, I can integrate motif data with available biochemical and structural data in literature and public databases to annotate residue-level motif functions across the major tyrosine kinase families. In addition to annotating the residue-level functions of motifs with respect to kinase function and regulation, I also provide a framework to directly associate individual motif residues to larger molecular and cellular functions. For example, an FGFR-specific aspartate in the αI-helix makes up part of the protein interaction interface with the SH2 domain of phospholipase C-gamma, whose interaction with the C-terminal lobe of the FGFR kinase domain facilitates phospholipase C-gamma phosphorylation, leading to increased cell motility (*17*). In this case, I can annotate this motif residue as having a kinase residue function of protein-protein interaction, a molecular function of phospholipase C-gamma phosphorylation, and a cellular function of cell motility (Figure 5.3).

In order to report the level of support substantiating our annotations, I provide evidence codes from the Evidence Codes Ontology (*18*), including experimental evidence, which includes structure determination evidence as well as other experimental methods, non-traceable author statement evidence, in which an author makes a statement that is not associated with results presented or a cited reference, and curator inference, which is used to describe all annotations where there is no direct experimental evidence available. When applicable, I provide the appropriate references to published articles and/or database access codes (i.e. Protein Data Bank, Reactome, Gene Ontology) (*19, 20*). For each motif residue, we additionally provide sequence information pertaining to the motif sequence, the sequence position of the motif, the family to which the motif belongs, the frequency of the motif within the family, and the frequency of the motif outside the family (Figure 5.3).

To provide our motif and annotation data to the greater research community, I have modified the framework of ProKinO, an ontological resource that integrates disparate data on protein kinases (*21, 22*), to incorporate motif information (Figure 5.2). In particular, I have developed a new "Motif" class, which includes distinct "SequenceMotif," "SecondaryStructureMotif," "Subdomain," "SpatialMotif," and "FunctionalMotif" subclasses, as well as a new "Function" class, in which "KinaseResidueFunction," "MolecularFunction," and "CellularFunction" subclasses are housed (Figure 5.2). In this way, identified sequence motifs are populated as instances of the "SequenceMotif" class, which are linked to instances of the "Function" class. Through the ontological framework provided by ProKinO, motif and annotation data will be linked to other important information relevant to protein kinases such as disease-associated mutations. Using the ontology query language, SPARQL, already implemented in ProKinO, researchers will be able to mine large amounts of data in biologically

and biomedically relevant ways to develop novel and testable hypotheses. For example, querying which disease-associated mutations affect residues involved in negative regulation will identify mutations across all kinases that potentially lead to over-activation and thus have may have oncogenic potential. Moreover, simply querying the functions of a residue that is mutated in disease will facilitate the determination of whether the mutation has any functional impact, and if so, what the molecular and functional consequences of the mutation may be. In this way, the detailed functional annotations of individual residues, rather than annotations of whole proteins, is particularly conducive for studying specific point mutations that often occur in disease.

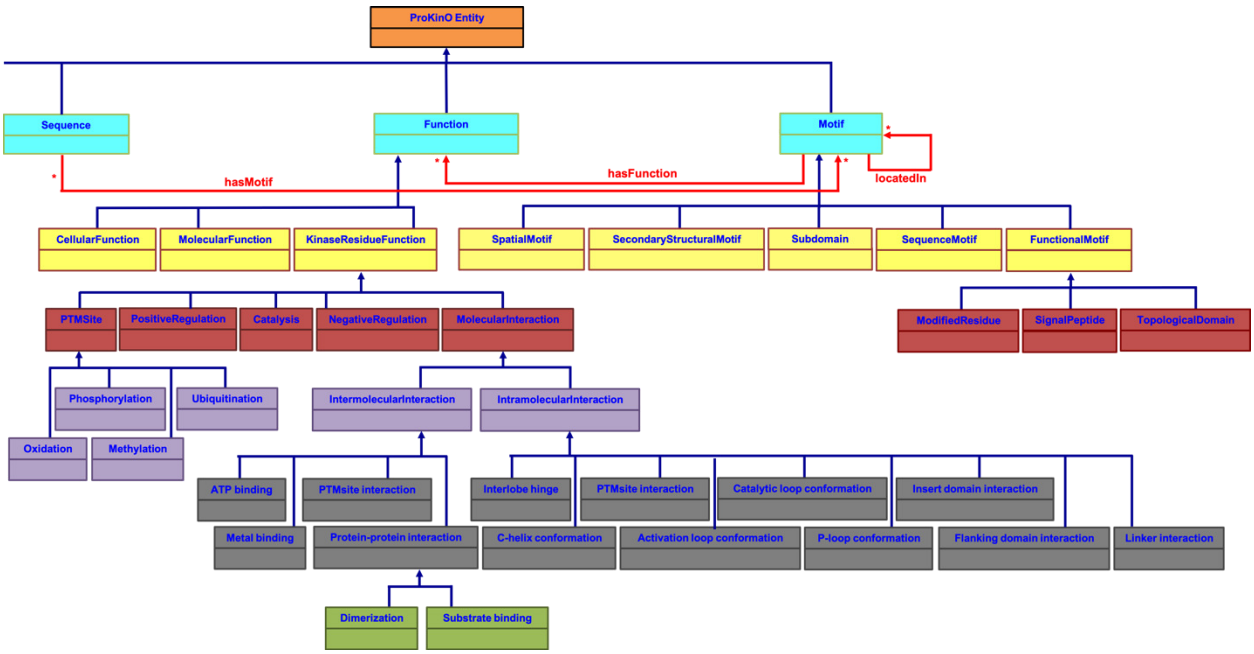

Figure 5.2: Schema of Motif and Function classes in ProKinO

Many tyrosine kinases (as well as most serine/threonine kinases) are still poorly studied, and some do not yet have crystal structures (i.e. PTK7, Lmr family), however, annotating conserved motif residues through the examination of structural models can provide novel testable hypotheses of the functions of these kinases (as demonstrated in Section 5.3.2), which can be provided to the larger research community through ProKinO. While the current study has

specifically focused on sequence motifs occurring across the tyrosine kinome, the same approach can be readily applied to study sequence motifs and their associated functions elsewhere in the protein kinome. Similarly, this comparative sequence approach and annotation scheme may be applied to study the evolution of and functional divergence of other protein families.

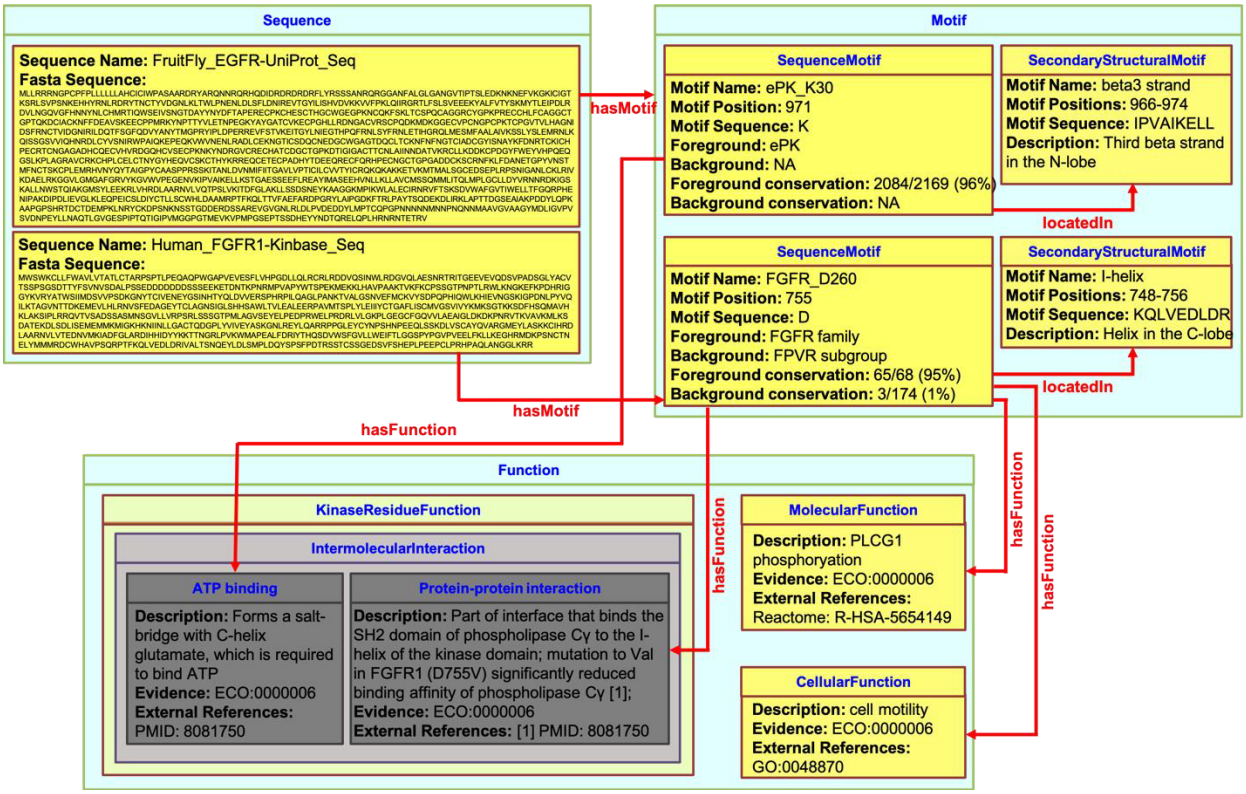

Figure 5.3: Example instances of Sequence, Motif, and Function classes and their relations

5.2.2 Further characterization of sequence motifs across different branches of tyrosine kinase evolution

The identification of sequence motifs associated with different branches of tyrosine kinase evolution can spotlight functional regions of tyrosine kinases, however, the mechanistic details of how these regions operate in regard to protein function are often lacking due to limited information available in crystal structures, biochemical databases, and in the literature. As a result, further examination of these sequence motifs through computational studies, such as

molecular dynamics simulations, and complementary mutagenesis experiments are often needed to piece together the full mechanisms encoded in these sequence motifs, as demonstrated by my studies on the Eph family of tyrosine kinases (*23*). Here, I propose a hypothesis-driven study to characterize regulatory mechanisms unique to the SrcM subgroup of tyrosine kinases.

The most highly distinguishing sequence motif of SrcM kinases is a methionine in the β3-αC loop that mediates unique interactions between the αC helix and activation loop in inactive structures of SrcM kinases. Molecular dynamics simulations on inactive strucures of wild-type and mutant BTK indicate that these interactions are stable during the course of simulation, however, *in silico* mutation of the methionine to a glycine significantly increases the flexibility of the kinase domain, implying that this residue may be important to stabilizing the unique SrcM inactive conformation (Figure 5.4). In contrast, *in silico* mutation of methioinine to leucine causes negligible changes in the backbone flexibility of the protein, however, the side chain of leucine is much less dynamic compared to the side chain of methionine, suggesting that the conserved methionine side chain may also be important for facilitating an inactive-to-active transition in SrcM kinases.

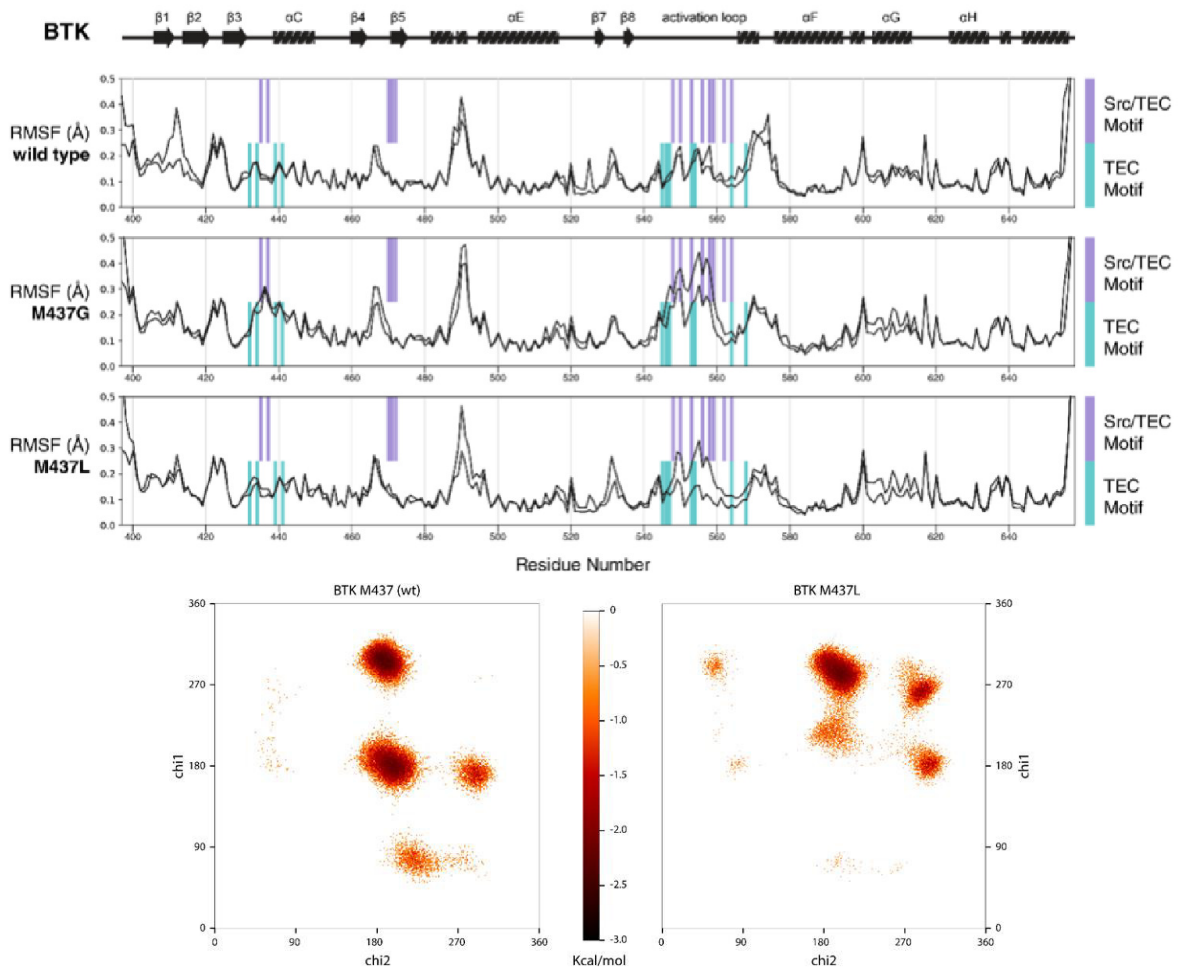

Figure 5.4: Molecular dynamics simulation of inactive WT and mutant BTK (performed in replicate). A) Backbone fluctuations of WT, M437G, and M347L BTK were calculated and plotted versus the time span of the simulation. Secondary structures are annotated above the plots. Distinguishing SrcM-specific motifs and TEC family-specific motifs are highlighted in lavender and cyan, respectively. B) Free energy surfaces for the β3-αC loop residue 437 in WT BTK and BTK M437L. Free energy surfaces show conformational preference throughout the simulation for chi1-chi2 dihedrals at residue 437.

The important structural and dynamic contributions by the SrcM-specific methionine is well established by the presented preliminary data, however, whether the methionine plays consistent structural and dynamic roles across all diverse SrcM members such as Frk, Lyn, ITK, and Abl is still unknown. One study of Abl kinase showed that mutation of the β3-αC loop methionine to glycine had an activating effect on Abl, whereas mutation to leucine decreased activity (*24*), which is consistent with our functional prediction of the methionine from our *in*

silico studies as explained above. Further studies of the methionine across multiple SrcM members, and importantly, examination of potential differences in this network across members, will further refine the evolutionary models of common regulatory cores in tyrosine kinases as proposed in the dissertation.

Bibliography

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

2. G. Manning, D. B. Whyte, R. Martinez, T. Hunter, S. Sudarsanam, The protein kinase complement of the human genome. *Science* **298**, 1912-1934 (2002).

3. D. R. Robinson, Y. M. Wu, S. F. Lin, The protein tyrosine kinase family of the human genome. *Oncogene* **19**, 5548-5557 (2000).

4. N. Amatya *et al.*, Autoinhibition beyond the SRC module: the multifacted pleckstrin homology domain of the TEC family *eLife*, (2019).

5. S. R. Hubbard, Juxtamembrane autoinhibition in receptor tyrosine kinases. *Nat Rev Mol Cell Biol* **5**, 464-471 (2004).

6. L. E. Wybenga-Groot *et al.*, Structural basis for autoinhibition of the Ephb2 receptor tyrosine kinase by the unphosphorylated juxtamembrane region. *Cell* **106**, 745-757 (2001).

7. T. L. Davis *et al.*, Autoregulation by the juxtamembrane region of the human ephrin receptor tyrosine kinase A3 (EphA3). *Structure* **16**, 873-884 (2008).

8. S. R. Kedika, D. G. Udugamasooriya, Converting a weaker ATP-binding site inhibitor into a potent hetero-bivalent ligand by tethering to a unique peptide sequence derived from the same kinase. *Org Biomol Chem* **16**, 6443-6449 (2018).

9. D. P. Bastedo *et al.*, Perturbations of a Kinase-Pseudokinase Module Activate Plant Immunity. *BioRxiv*, (2019).

10. V. Klymiuk *et al.*, Cloning of the wheat Yr15 resistance gene sheds light on the plant tandem kinase-pseudokinase family. *Nat Commun* **9**, 3735 (2018).

11. M. D. Berg, J. Genereaux, J. Karagiannis, C. J. Brandl, The Pseudokinase Domain of Saccharomyces cerevisiae Tra1 Is Required for Nuclear Localization and Incorporation into the SAGA and NuA4 Complexes. *G3: Genes|Genomes|Genetics* **8**, 1943-1957 (2018).

12. J. A. Adams, Kinetic and catalytic mechanisms of protein kinases. *Chem Rev* **101**, 2271-2290 (2001).

13. M. A. Lemmon, J. Schlessinger, Cell signaling by receptor tyrosine kinases. *Cell* **141**, 1117-1134 (2010).

14. A. Torkamani, N. J. Schork, Prediction of cancer driver mutations in protein kinases. *Cancer Res* **68**, 1675-1682 (2008).

15. M. Huse, J. Kuriyan, The conformational plasticity of protein kinases. *Cell* **109**, 275-282 (2002).

16. H. B. Chen *et al.*, A molecular brake in the kinase hinge region regulates the activity of receptor tyrosine kinases. *Molecular Cell* **27**, 717-730 (2007).

17. J. H. Bae *et al.*, The selectivity of receptor tyrosine kinase signaling is controlled by a secondary SH2 domain binding site. *Cell* **138**, 514-524 (2009).

18. M. C. Chibucos *et al.*, Standardized description of scientific evidence using the Evidence Ontology (ECO). *Database (Oxford)* **2014**, (2014).

19. M. Ashburner *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).

20. D. Croft *et al.*, Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* **39**, D691-697 (2011).

21. D. I. McSkimming *et al.*, ProKinO: a unified resource for mining the cancer kinome. *Hum Mutat* **36**, 175-186 (2015).

22. G. Gosal, K. J. Kochut, N. Kannan, ProKinO: an ontology for integrative analysis of protein kinases in cancer. *PLoS One* **6**, e28782 (2011).

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

24. N. Dolker *et al.*, The SH2 domain regulates c-Abl kinase activation by a cyclin-like mechanism and remodulation of the hinge motion. *PLoS Comput Biol* **10**, e1003863 (2014).