Computational Characterization of Protein Kinase Adaptations in Eukaryotic Pathogens

by

Eric W. Talevich

(Under the Direction of Natarajan Kannan)

Abstract

Parasitic protozoa cause a number of devastating human and veterinary diseases, including malaria, toxoplasmosis, babesiosis and cryptosporidiosis. Protein kinases, a broad class of cellular signaling enzymes which have proven effective drug targets in human cancers, are promising targets in these parasitic diseases as well.

In this work I develop and apply comparative computational techniques in analyses of the protein kinases in specific evolutionary groups of eukaryotic pathogens. First, I comprehensively examine conserved protein kinase families in the Apicomplexa, the phylum that includes the malaria parasites *Plasmodium* spp. and the opportunistic pathogen *Toxoplasma gondii*, to identify conserved genomic and structural features that distinguish parasite kinases from those in their hosts and other eukaryotes. I then explore the structural and evolutionary divergence of the virulence-associated, coccidian-specific rhoptry kinase family. The novel findings presented here shed light on parasite phosphoryl signaling mechanisms as well as provide guidance on potential drug targets for parasitic diseases.

Index words: Apicomplexa, protein kinase, cell signaling, host-pathogen interactions, kinome, phosphorylation

Computational Characterization of Protein Kinase Adaptations in Eukaryotic Pathogens

by

Eric W. Talevich

B.S., University of California, Davis, 2005

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

2013

Computational Characterization of Protein Kinase Adaptations in Eukaryotic Pathogens

by

Eric W. Talevich

Approved:

Major Professor:	Natarajan Kannan
Committee:	Jessica C. Kissinger
	James H. Leebens-Mack
	Zachary A. Wood

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2013

Computational Characterization of Protein Kinase Adaptations in Eukaryotic Pathogens

Eric W. Talevich

August 2013

Dedication

To my loving wife, Brittany Gentile, who made this possible.

Acknowledgments

I am grateful to my committee members Natarajan Kannan, Jessie Kissinger, Jim Leebens-Mack, and Zac Wood; ESBG lab members Amar Mirza, Anish Narayanan, Daniel McSkimming, Gurinder Gosal, Krishnadev Oruganty, Samiksha Katiyar, Shima Dastgheib, Smita Mohanty, Surabhi Maheshwari and Tuan Nguyen; and colleagues at the University of Georgia and elsewhere for their support and guidance.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and literature review

1.1 Motivation

Malaria and related infectious diseases are responsible for thousands of deaths each day, as well as a substantial societal burden due to illness both in humans and in agricultural animals. Current drug treatments for these diseases are unsatisfactory, and parasite strains have been observed to quickly evolve resistance. There is a worldwide need for a consistent drug development pipeline to supply the antiparasitics to treat and cure these diseases.

Protein kinases are attractive pharmaceutical targets because of their important role in the regulation of many cell processes [8]. Since effective protein kinase inhibitors have already been developed to treat human diseases such as cancer, there is significant interest in reusing these drugs and similar compounds to treat infectious diseases caused by eukaryotic pathogens [9]. However, the safety and effectiveness of such treatments relies on the successful targeting of lineage-specific protein kinase features which appear in the parasites but not in their host cells [10].

Many of the pathogens that cause these important global diseases belong to the taxonomic group Apicomplexa, a protozoan phylum consisting mainly of parasitic species. These include *Plasmodium* spp. (malaria), *Toxoplasma gondii* (toxoplasmosis) and *Cryptosporidium* spp. (cryptosporidiosis), among others [36]. However, relatively little is known about the basic biology of these species, compared to that of humans and model organisms. This clade is one of the least-understood branches of Eukaryota, and its medical relevance warrants

deeper investigation into these species' basic biology [25].

We therefore seek to identify distinctive, conserved functional features of apicomplexan protein kinases, some of which could serve as specific targets for therapeutic inhibition. Comparative approaches make it possible to characterize features of species that have not yet been directly investigated in depth. By studying the protein kinases in these parasites we shed light on fundamental features of their cellular biology, both those shared across Eukaryota and those unique to certain lineages, as well as the ancient evolutionary history of the eukaryotic protein kinase superfamily.

1.2 Background

1.2.1 Biology of the Apicomplexa

Impact on health and economic development

Malaria is a devastating parasitic disease that infects hundreds of millions of people and kills more than half a million each year [44]. The single-celled parasite that causes malaria, *Plasmodium* spp., belongs to the phylum Apicomplexa, which includes many other related parasites responsible for human and veterinary diseases.

While malaria is considered one of the "big three" infectious diseases causing worldwide morbidity and mortality (http://www.burnet.edu.au/home/general/focus/big3), a number of "neglected" diseases, including those caused by apicomplexans and other eukaryotic pathogens, receive relatively little research funding despite their substantial global impact. Other apicomplexan diseases affecting humans include toxoplasmosis and cryptosporidiosis, which are generally not fatal in otherwise healthy individuals but can become serious in immunocompromised individuals, such as those with comorbid AIDS. Cryptosporidiosis, caused by *Cryptosporidium* species, afflicts humans worldwide; in otherwise healthy (immunocompetent) adult individuals it causes acute gastroenteritis and diarrhea, generally lasting about a month, and is then resolved by the body's own immune system. *Toxoplasma gondii*, the causative agent of toxoplasmosis, infects an estimated 30% of all humans worldwide; in the vast majority of cases it exists as dormant cysts in the brain and does not

cause physiological symptoms. However, both these diseases are a serious threat to young children, pregnant women and immunocompromised individuals. The AIDS pandemic has created large immunocompromised populations in many of the same tropical areas where apicomplexan diseases are endemic, and such infections may be fatal. *T. gondii* infection has also been linked to schizophrenia and paranoia.

In addition to the significant mortality caused by these diseases, there is a substantial economic and social burden imposed by human illness, as well as veterinary diseases that affect cattle and other animals important to agriculture. These veterinary and agricultural diseases include *Babesia bovis* (haemolytic anemia or babesiosis in cattle), *Theileria annulata* and *T. parva* (tropical theileriosis and East Coast fever, respectively, in cattle), *Eimeria tenella* (coccidiosis in chickens), *Sarcocystis neurona* (myeloencephalitis in horses), and *Neospora caninum* (neosporosis in cattle and neurological problems in dogs).

Most of these diseases have non-existent or poor treatments. The currently available antiparasitic drugs are inconsistently effective, and often have damaging side-effects. Resistance to the available treatments has been identified and is rising. By improving our understanding of the molecular biology of apicomplexan parasites, however, we can improve the treatment, diagnosis and prevention of these diseases they cause.

Evolutionary relationships of the Apicomplexa

As a phylum, Apicomplexa is remarkable large and geographically widespread, consisting of several million species, yet also one of the least characterized, with only about 0.1 percent of species assigned scientific names [2]. Along with ciliates and dinoflagellates, they comprise the kingdom/taxonomic group Alveolata (Figure 1.1).

Phylogenetic studies based on a molecular clock suggest that the first apicomplexans evolved nearly 1 billion years ago, before the Cambrian era and the emergence of land-dwelling animals [14, 39]. For comparison, fossil records date the the divergence of Metazoan phyla to a similar time period [31]. Thus, although we consider *Plasmodium falciparum* and *Toxoplasma gondii* as relatives with shared features in the following studies, evolutionarily speaking, they are about as divergent as humans and mosquitoes, having diverged about 800 million years ago [31].

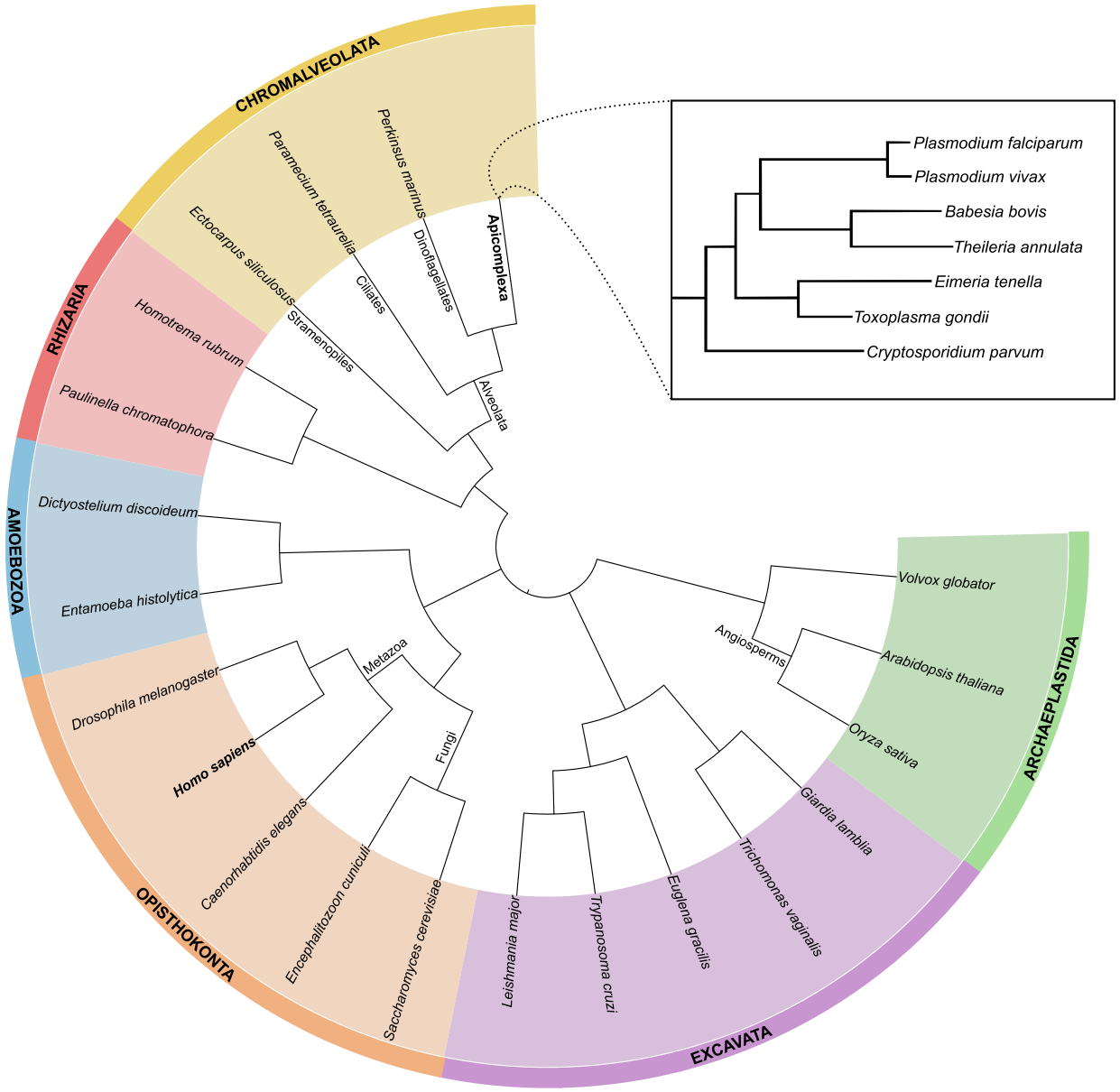


Figure 1.1: Phylogenetic classification of eukaryotes, based on [1], with inset phylogeny of apicomplexan species from [19].

On the other hand, gene families that are conserved across this evolutionary span are likely to be essential, or difficult for the parasite to replace in response to changes in selective pressures. This is useful information because parasites have been observed to quickly develop resistance to existing treatments; there is some correlation between the conservation of a gene across multiple species and its essentiality for survival or reproduction, as evidenced by kinase gene knockout studies [40, 42].

Within Apicomplexa, four sub-clades have been established: haemosporidians (represented by the *Plasmodium* genus here), piroplasmids (including *Theileria* spp. and *Babesia bovis*), coccidians (*Toxoplasma gondii*, *Neospora caninum*, *Sarcocystis neurona*, *Eimeria tenella*, and others), and gregarines. The taxonomic classification of *Cryptosporidium* spp. was inconsistent in early work, but more recent phylogenetic evidence places this lineage as basal to the other apicomplexans [19].

A few outgroup species are notable for their use in understanding apicomplexan genomics through comparison. The ciliates *Tetrahymena thermophila* and *Paramecium tetraulia* are free-living, non-photosynthetic alveolates; the genomes of both species have been sequenced and annotated, including the protein kinases [6, 11]. The genome of the dinoflagellate *Perkinsus marinus*, an oyster parasite, has been fully sequenced an annotated as well, and has been used as an outgroup to root apicomplexan species trees [19]. A photosynthetic alga closely related to apicomplexans, *Chromera velia*, can also fill this role as an outgroup taxon [18, 24].

Cellular organelles and ultrastructure

Apicomplexans are named for a cellular structure at the apex of the cell, known as the apical complex, which the parasite uses to recognize and invade host cells. This structure is a complex of several unique organelles, namely the conoid, rhoptries, micronemes and polar or apical rings. Another type of apicomplexan-specific organelle, the dense granules, are dispersed throughout the cytoplasm. Most apicomplexan also contain a unique non-photosynthetic plastid called the apicoplast which was ancestrally obtained through an ancient secondary endosymbiosis of a red alga [18], leaving apicomplexans with plant-like characteristics including plant-specific gene families and a vulnerability to some herbicides.

Life cycles

Like many parasites, apicomplexans exhibit complex life cycles involving one or two host species, and may pass through multiple stages in each host [36, 41]. For example, the *P. falciparum* life cycle involves an initial sporozoite stage in the *Anopheles* mosquito vector, transmission by the mosquito to a human host followed by a maturation stage in the liver and a blood stage in which parasite cells invade host erythrocytes, multiply and differentiate, and then burst from the cells and return to the bloodstream to continue a period pattern of erythrocyte invasion. Within the erythrocyte, some of the parasite cells parasites differentiate into non-proliferating male or female gametocytes, which may be taken up again by a mosquito to undergo gametogenesis and fertilization, and finally produce oocysts which in turn produce sporozoites to complete the cycle. Other apicomplexan species vary in each of these aspects: *Cryptosporidium parvum* has only a single host species, humans; *Theileria* spp. and *Babesia bovis* escape the parasitophorous vacuole shortly after entering the host lymphocyte cell; *T. gondii* is capable of infecting a wide variety of mammalian hosts and host cell types.

1.2.2 Eukaryotic protein kinases as a model system

The eukaryotic protein kinases (ePK) are a superfamily of proteins which phosphorylate protein substrates, transferring a phosphate group from adenosine triphosphate (ATP) to a serine, threonine or tyrosine residue on the substrate protein. The human genome contains 518 protein kinase genes (known as the "kinome"), about 2% of the entire genome [21]. However, these proteins play a disproportionately large role in cell regulation: An estimated 30% of the proteins in a human cell are phosphorylated, and 40% of cellular pathways involve protein phosphorylation [21].

The ePK superfamily, while sharing a common fold (Figure 1.2), is classified hierarchically into seven major groups according to phylogeny, domain architecture and broad functional roles [17, 21]. Within each group, the kinases are further classified into families and subfamilies, some of which are lineage-specific. There are also a number of "Other" ePK families which do not fit cleanly into any of the major groups, and several "Atypical" families which show protein kinase activity but lack sequence homology to ePKs. The identification

and classification of the ePKs in the genome of a given species thus provides clues as to its regulatory capabilities.

Protozoan pathogens, like all eukaryotes, contain ePKs. The genomic complement of kinases ("kinome") of the apicomplexans *Plasmodium falciparum* [3, 43], *Toxoplasma gondii* [33] and *Cryptosporidium parvum* [4], and the trypanosomatids *Trypanosoma brucei*, *T. cruzi* and *Leishmania major* [32], have been examined. These studies identified expansions, reductions and losses of conserved ePK families, as well as the emergence of novel lineage-specific families, and suggested biological processes that may be affected by these genomic changes. The phyletic patterns of expansion and conservation of protein kinase families can indicate which signaling pathways are biologically important. For instance, multiple copies of the calcium-dependent protein kinase (CDPK) family in apicomplexans suggests calcium signaling may be biologically important. Indeed, the apparent amplification of calcium-dependent protein kinases in *Toxoplasma gondii* and its close relatives correlates with the expanded use of calcium signaling for motility in those species [26, 27]. Other parasite-specific kinases are exported to the host cell and contain distinctive sequence features that can be used to identify them [33, 38]. Through systematic analysis of parasite kinomes, we can leverage our knowledge of signaling pathways in eukaryotic cells to gain insight into the parasite's molecular function, host-parasite interactions, and potential targets for treatment of parasitic disease.

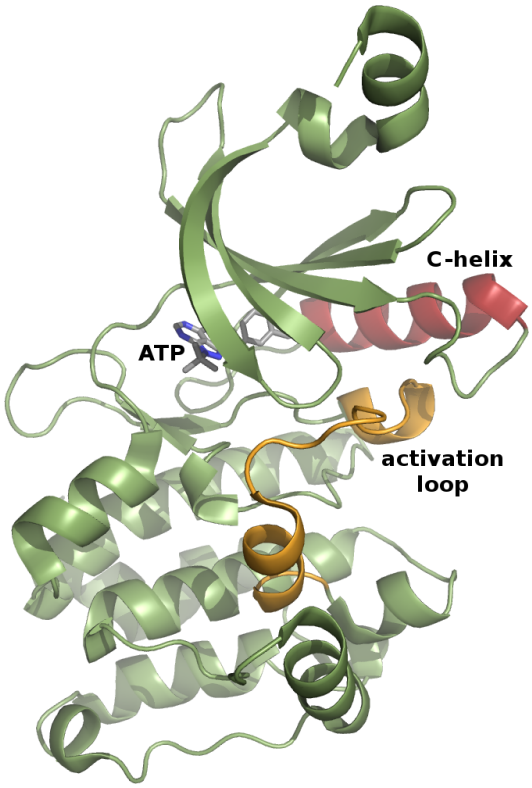


Figure 1.2: Structure of *Cryptosporidium parvum* calcium-dependent protein kinase 1 (CDPK1) with bound ATP (PDB: 2WEI; [4]).

Protein kinases as drug targets in the treatment of apicomplexan diseases

Since effective protein kinase inhibitors have already been developed to treat human diseases such as cancer, there is significant interest in reusing these drugs and similar compounds to treat infectious diseases caused by eukaryotic pathogens [9, 10, 28, 35]. In 2010, GlaxoSmithKline released a library of thousands of pharmacological compounds, known as the Tres Cantos antimalarial compound set, screened for effectiveness against the most virulent malaria parasite species, *P. falciparum*; kinase inhibitors feature prominently among this set of drug candidates. Gene knockout studies have also been conducted to identify the essential kinases in parasite genomes, including 36 essential protein kinases in *P. falciparum* [40]. The work that has been done so far, however, is only a start, and much more will be needed to translate our understanding of kinases in parasites to improving clinical outcomes. Designing parasite-specific inhibitors requires a detailed understanding of the biochemistry and distinguishing characteristics of the protein kinases in parasite genomes.

Characterization of the kinome, the full genomic complement of protein kinases, is a useful useful step in characterizing a species' signaling pathways and identifying possible therapeutic approaches. For example, identification of cell cycle control kinases such as cyclin-dependent kinases (CDKs) can point to some of the most promising targets for inhibition; potentially, inhibiting these kinases could halt parasite growth or replication [16]. Conserved, lineage-specific kinase families are also good targets because drugs that target these proteins are more likely to affect the parasites specifically, with relatively little effect on the host cells.

1.3 Key challenges and unresolved questions

Given the high potential of protein kinases as therapeutic targets, we have a keen interest in the mechanisms and pathways of protein phosphorylation in protozoan pathogens, and how they differ from those in metazoan hosts. An important factor in the effectiveness and safety of kinase inhibitors is that they are specific to the targeted cells, and do not disrupt healthy cells in the host [10]. In the case of parasitism, this requires a detailed understanding of

the biological and biochemical differences between the cells of the targeted parasite and the host. To reduce the likelihood of the targeted cells developing resistance to a treatment, it is effective to target conserved, slowly evolving features, rather than those which are able to mutate quickly.

While metazoan ePKs have been the focus of most research to date, the kinases in non-metazoan eukaryotes, particularly protozoans, are still poorly understood. Previous efforts to perform detailed comparative analysis of protozoan kinomes have largely focused on individual species, typically soon after the draft genome sequence of a species is completed. Consequently, there is no global overview of the sequence and structural features that distinguish apicomplexan and trypanosomatid kinases collectively from their metazoan counterparts.

My working hypothesis is that some protein kinases in parasites have evolved differently than those in their hosts. In several protozoan parasites for which the kinome has been analyzed, unique parasite-specific adaptations in protein kinases have been observed [23, 32, 43]. The systematic comparative analysis of ePKs in protozoan parasites, with respect to their evolutionary relatives and to metazoans, can potentially reveal novel ePK features which are conserved in parasites but not in their metazoan hosts. These lineage-specific features may therefore serve as effective pharmaceutical targets, as well as providing insight into the basic biology of these protozoan species. For example, many kinases are known to be activated by phosphorylation of the activation loop (highlighted in Figure 1.2). If an ePK in apicomplexans shows conserved differences in the activation loop relative to orthologous kinases in other eukaryotes, this may indicate differences in the activation mechanism specific to apicomplexans.

Genome-wide analyses of protein kinases have been performed on many species already [20, 22], which has allowed us to construct sequence profiles for the ePK families and subfamilies that have been characterized. In addition, crystallographic structures of protein kinases from pathogenic protozoan species have been solved, most notably through the efforts of the Structural Genomics Consortium [15]. However, these systematic structural and phylogenetic approaches have not yet been combined to study the mechanistic consequences of lineage-specific adaptations.

1.4 Major research questions addressed

The following chapters address the stated knowledge gaps and overall goal by identifying and examining the distinguishing features of protein kinases in parasitic protozoa. The two research studies investigate the conserved differences between protozoan parasites and hosts at the level of the whole kinome and, within the kinase domain, at the level of specific residues. In each study, I use innovative methodologies that integrate broad genomic analysis with residue-level structural analysis to identify novel features which are specific to certain lineages.

1.4.1 Lineage-specific adaptations in apicomplexan kinomes

Rationale

The kinomes of three medically important apicomplexan species, *Plasmodium falciparum*, *Toxoplasma gondii* and *Cryptosporidium parvum*, were each examined in previous studies by others [3, 4, 33, 43]. However, there was no broad comparison across all of the currently available whole genomes of apicomplexans that synthesized this species-specific information; there was therefore no convincing overview of the phyletic distribution of the novel features identified in individual species.

In this study, I expand on these previous efforts and identify the prevalence and phyletic distribution of lineage-specific protein kinase families and novel sequence features. I undertake a systematic examination and comparision of the kinomes of 15 apicomplexan species for which whole genome sequences are available. This broader analysis effectively uses the available sequence data to identify distinctive conservation patterns at the residue level and identify apicomplexan-specific protein kinase features. In addition, several crystal structures of protein kinases from apicomplexans have been made available in the Protein Data Bank (PDB). I therefore place the results of the comparative analysis in structural context and hypothesize functional and mechanistic consequences.

Research goals

Kinome identification, classification and comparison across apicomplexan species: I systematically catalogue the conserved kinase families in apicomplexans and selected outgroup species, identifying and hierarchically classifying the complement of eukaryotic protein kinases in each genome to reveal lineage-specific expansions, reductions and losses in apicomplexan kinomes.

Specific instances of structural and functional divergence in known ePK families: I use phylogenetic methods to identify divergent apicomplexan- or alveolate-specific ortholog groups within deeply conserved ePK families, specifically those with homologs in both apicomplexan parasites and other branches of Eukaryota. I then perform CHAIN analysis [29] to compare these divergent ortholog groups to the broader families and identify contrasting patterns of residue conservation. Bayesian analysis of selective constraints imposed on these familes indicates distinctive sequence and structural features, distinguishing apicomplexan kinases from their orthologs in model organisms. This residue-level analysis highlights novel protein family adaptations, at the sequence and structural levels, and points to possible functional roles with which these features may be associated. I then discuss the structural and functional implications of these apicomplexan-specific variations.

1.4.2 Subfamily-level diversification of the rhoptry kinase family

Rationale

In this project I investigate a unique family of kinases found in the toxoplasmosis parasite *Toxoplasma gondii* and its close relatives, the Coccidia. This family of protein kinases, called rhoptry kinases (ROPK), have been identified as key determinants of parasite virulence [37]. During the invasion process of *T. gondii*, these kinases are secreted into the parasitophorous vacuole and interact with the host cell's signaling machinery [7]. Much remains to be understood about the molecular mechanisms these parasites use to invade the host cell and co-opt its internal machinery.

A systematic analysis of ROPK sequences has been performed by others, providing names and descriptions for a number of ROPK genes [12, 33]. Four crystal structures of rhoptry

kinases have also been solved [34]. In addition, the whole genomes of several strains of *T. gondii* and other coccidians have been sequenced; there is evidence that differences in the size and composition of ROPK subfamilies between *T. gondii* strains contribute to differences in virulence [5]. However, the results of systematic, multi-species analysis have not yet been integrated with the structural context that can be provided by the solved structures.

Research goals

Subfamily-level phylogenetic structure: My preliminary analysis of apicomplexan kinomes revealed several apparent subfamilies within the ROPK family. Here, I apply phylogenetic methods to identify putative subfamilies in the ROPK family. This analysis can reveal subfamilies shared across species, lineage-specific expansions within the ROPK family, and whether the currently annotated ROPKs are indeed a monophyletic group.

ROPK-shared and subfamily-specific structural features and mechanisms: I determine the sequence and structural features that distinguish these subfamilies from each other, as well as what features distinguish the ROPK family as a whole from typical ePKs. In particular, quantitative methods are used to identify sequence motifs which are conserved in the ROPK family, but not in the broader protein kinase superfamily. The same techniques are also used to identify ROPK subfamily-specific motifs and, where possible, map the motifs onto solved protein structures in order to develop functional hypotheses.

Prediction and comparison of active kinases versus pseudokinases: Most of the ROPK members are believed to be catalytically inactive, but kinase activity has been demonstrated in some [13, 30, 34]. Our analysis applies general knowledge of protein kinase mechanisms to categorize each rhoptry kinase as a likely active, likely pseudokinase, or potentially active but with an atypical catalytic mechanism.

Bibliography

[1] Adl, S. M., Simpson, A. G. B., Farmer, M. A., Andersen, R. A., Anderson, O. R., Barta, J. R., Bowser, S. S., Brugerolle, G., Fensome, R. A., Fredericq, S., James, T. Y., Karpov, S., Kugrens, P., Krug, J., Lane, C. E., Lewis, L. A., Lodge, J., Lynn, D. H., Mann, D. G.,

McCourt, R. M., Mendoza, L., Moestrup, O., Mozley-Standridge, S. E., Nerad, T. A., Shearer, C. A., Smirnov, A. V., Spiegel, F. W., and Taylor, M. F. J. R. (2005). The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *The Journal of Eukaryotic Microbiology*, **52**(5), 399–451.

[2] Adl, S. M., Leander, B. S., Simpson, A. G. B., Archibald, J. M., Anderson, O. R., Bass, D., Bowser, S. S., Brugerolle, G., Farmer, M. A., Karpov, S., Kolisko, M., Lane, C. E., Lodge, D. J., Mann, D. G., Meisterfeld, R., Mendoza, L., Moestrup, O. j., Mozley-Standridge, S. E., Smirnov, A. V., and Spiegel, F. (2007). Diversity, nomenclature, and taxonomy of protists. *Systematic Biology*, **56**(4), 684–689.

[3] Anamika, Srinivasan, N., and Krupa, A. (2005). A genomic perspective of protein kinases in *Plasmodium falciparum*. *Proteins*, **58**(1), 180–189.

[4] Artz, J. D., Wernimont, A. K., Allali-Hassani, A., Zhao, Y., Amani, M., Lin, Y.-H., Senisterra, G., Wasney, G. A., Fedorov, O., King, O., Roos, A., Lunin, V. V., Qiu, W., Finerty, P., Hutchinson, A., Chau, I., von Delft, F., Mackenzie, F., Lew, J., Kozieradzki, I., Vedadi, M., Schapira, M., Zhang, C., Shokat, K., Heightman, T., and Hui, R. (2011). The *Cryptosporidium parvum* Kinome. *BMC Genomics*, **12**(1), 478.

[5] Behnke, M. S., Khan, A., Wootton, J. C., Dubey, J. P., Tang, K., and Sibley, L. D. (2011). Virulence differences in *Toxoplasma* mediated by amplification of a family of polymorphic pseudokinases. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(23), 9631–9636.

[6] Bemm, F., Schwarz, R., Förster, F., and Schultz, J. (2009). A kinome of 2600 in the ciliate Paramecium tetraurelia. *FEBS Letters*, **583**(22), 3589–3592.

[7] Boothroyd, J. C. and Dubremetz, J.-F. (2008). Kiss and spit: the dual roles of *Toxoplasma* rhoptries. *Nature Reviews. Microbiology*, **6**(1), 79–88.

[8] Cohen, P. (2002). Protein kinases–the major drug targets of the twenty-first century? *Nature Reviews. Drug Discovery*, **1**(4), 309–315.

[9] Doerig, C. (2004). Protein kinases as targets for anti-parasitic chemotherapy. *Biochimica et Biophysica Acta*, **1697**(1-2), 155–168.

[10] Doerig, C., Abdi, A., Bland, N., Eschenlauer, S., Dorin-Semblat, D., Fennell, C., Halbert, J., Holland, Z., Nivez, M.-P., Semblat, J.-P., Sicard, A., and Reininger, L. (2010). Malaria: targeting parasite and host cell kinomes. *Biochimica et Biophysica Acta*, **1804**(3), 604–612.

[11] Eisen, J. A., Coyne, R. S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J. R., Badger, J. H., Ren, Q., Amedeo, P., Jones, K. M., Tallon, L. J., Delcher, A. L., Salzberg, S. L., Silva, J. C., Haas, B. J., Majoros, W. H., Farzad, M., Carlton, J. M., Smith, R. K., Garg, J., Pearlman, R. E., Karrer, K. M., Sun, L., Manning, G., Elde, N. C., Turkewitz, A. P., Asai, D. J., Wilkes, D. E., Wang, Y., Cai, H., Collins, K., Stewart, B. A., Lee, S. R., Wilamowska, K., Weinberg, Z., Ruzzo, W. L., Wloga, D., Gaertig, J., Frankel, J., Tsao, C.-C., Gorovsky, M. A., Keeling, P. J., Waller, R. F., Patron, N. J., Cherry, J. M., Stover, N. A., Krieger, C. J., del Toro, C., Ryder, H. F., Williamson, S. C., Barbeau, R. A., Hamilton, E. P., and Orias, E. (2006). Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. *PLoS Biology*, **4**(9), e286.

[12] El Hajj, H., Demey, E., Poncet, J., Lebrun, M., Wu, B., Galéotti, N., Fourmaux, M. N., Mercereau-Puijalon, O., Vial, H., Labesse, G., and Dubremetz, J. F. (2006). The ROP2 family of *Toxoplasma gondii* rhoptry proteins: proteomic and genomic characterization and molecular modeling. *Proteomics*, **6**(21), 5773–5784.

[13] El Hajj, H., Lebrun, M., Arold, S. T., Vial, H., Labesse, G., and Dubremetz, J. F. (2007). ROP18 is a rhoptry kinase controlling the intracellular proliferation of *Toxoplasma gondii*. *PLoS Pathogens*, **3**(2), e14.

[14] Escalante, A. A. and Ayala, F. J. (1995). Evolutionary origin of *Plasmodium* and other Apicomplexa based on rRNA genes. *Proceedings of the National Academy of Sciences of the United States of America*, **92**(13), 5793–5797.

[15] Gileadi, O., Knapp, S., Lee, W. H., Marsden, B. D., Müller, S., Niesen, F. H., Kavanagh, K. L., Ball, L. J., von Delft, F., Doyle, D. A., Oppermann, U. C. T., and Sundström,

M. (2007). The scientific impact of the Structural Genomics Consortium: a protein family and ligand-centered approach to medically-relevant human proteins. *Journal of Structural and Functional Genomics*, **8**(2-3), 107–119.

[16] Hammarton, T. C., Mottram, J. C., and Doerig, C. (2003). The cell cycle of parasitic protozoa: potential for chemotherapeutic exploitation. *Progress in Cell Cycle Research*, **5**, 91–101.

[17] Hanks, S. K. and Hunter, T. (1995). Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB Journal*, **9**(8), 576–596.

[18] Janouskovec, J., Horák, A., Oborník, M., Lukes, J., and Keeling, P. J. (2010). A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(24), 10949–10954.

[19] Kuo, C.-H., Wares, J. P., and Kissinger, J. C. (2008). The Apicomplexan whole-genome phylogeny: an analysis of incongruence among gene trees. *Molecular Biology and Evolution*, **25**(12), 2689–2698.

[20] Manning, G., Plowman, G. D., Hunter, T., and Sudarsanam, S. (2002a). Evolution of protein kinase signaling from yeast to man. *Trends in Biochemical Sciences*, **27**(10), 514–520.

[21] Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002b). The protein kinase complement of the human genome. *Science*, **298**(5600), 1912–1934.

[22] Martin, D. M. A., Miranda-Saavedra, D., and Barton, G. J. (2009). Kinomer v. 1.0: a database of systematically classified eukaryotic protein kinases. *Nucleic Acids Research*, **37**(Database issue), D244–D250.

[23] Miranda-Saavedra, D., Stark, M. J. R., Packer, J. C., Vivares, C. P., Doerig, C., and Barton, G. J. (2007). The complement of protein kinases of the microsporidium *Encephal-*

itozoon cuniculi in relation to those of *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. *BMC Genomics*, **8**, 309.

[24] Moore, R. B., Oborník, M., Janouskovec, J., Chrudimský, T., Vancová, M., Green, D. H., Wright, S. W., Davies, N. W., Bolch, C. J. S., Heimann, K., Slapeta, J., Hoegh-Guldberg, O., Logsdon, J. M., and Carter, D. A. (2008). A photosynthetic alveolate closely related to apicomplexan parasites. *Nature*, **451**(7181), 959–963.

[25] Morrison, D. A. (2009). Evolution of the Apicomplexa: where are we now? *Trends in Parasitology*, **25**(8), 375–382.

[26] Nagamune, K. and Sibley, L. D. (2006). Comparative genomic and phylogenetic analyses of calcium ATPases and calcium-regulated proteins in the apicomplexa. *Molecular Biology and Evolution*, **23**(8), 1613–1627.

[27] Nagamune, K., Moreno, S. N., Chini, E. N., and Sibley, L. D. (2008). Calcium regulation and signaling in apicomplexan parasites. In B. A. Burleigh and D. Soldati-Favre, editors, *Molecular Mechanisms of Parasite Invasion*, volume 47, chapter 5, pages 70–81. Landes Bioscience and Springer Science+Business Media.

[28] Naula, C., Parsons, M., and Mottram, J. C. (2005). Protein kinases as drug targets in trypanosomes and *Leishmania*. *Biochimica et Biophysica Acta*, **1754**(1-2), 151–159.

[29] Neuwald, A. F. (2007). The CHAIN program: forging evolutionary links to underlying mechanisms. *Trends in Biochemical Sciences*, **32**(11), 487–493.

[30] Ong, Y.-C., Reese, M. L., and Boothroyd, J. C. (2010). *Toxoplasma* rhoptry protein 16 (ROP16) subverts host function by direct tyrosine phosphorylation of STAT6. *The Journal of Biological Chemistry*, **285**(37), 28731–28740.

[31] Parfrey, L. W., Lahr, D. J. G., Knoll, A. H., and Katz, L. A. (2011). Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(33), 13624–13629.

[32] Parsons, M., Worthey, E. A., Ward, P. N., and Mottram, J. C. (2005). Comparative analysis of the kinomes of three pathogenic trypanosomatids: *Leishmania major*, *Trypanosoma brucei* and *Trypanosoma cruzi*. *BMC Genomics*, **6**, 127.

[33] Peixoto, L., Chen, F., Harb, O. S., Davis, P. H., Beiting, D. P., Brownback, C. S., Ouloguem, D., and Roos, D. S. (2010). Integrative genomic approaches highlight a family of parasite-specific kinases that regulate host responses. *Cell Host & Microbe*, **8**(2), 208–218.

[34] Qiu, W., Wernimont, A. K., Tang, K., Taylor, S., Lunin, V., Schapira, M., Fentress, S., Hui, R., and Sibley, L. D. (2009). Novel structural and regulatory features of rhoptry secretory kinases in *Toxoplasma gondii*. *The EMBO Journal*, **28**(7), 969–979.

[35] Renslo, A. R. and McKerrow, J. H. (2006). Drug discovery and development for neglected parasitic diseases. *Nature Chemical Biology*, **2**(12), 701–710.

[36] Roos, D. S. (2005). Genetics. Themes and variations in apicomplexan parasite biology. *Science*, **309**(5731), 72–73.

[37] Saeij, J. P. J., Boyle, J. P., Coller, S., Taylor, S., Sibley, L. D., Brooke-Powell, E. T., Ajioka, J. W., and Boothroyd, J. C. (2006). Polymorphic secreted kinases are key virulence factors in toxoplasmosis. *Science*, **314**(5806), 1780–1783.

[38] Schneider, A. G. and Mercereau-Puijalon, O. (2005). A new Apicomplexa-specific protein kinase family: multiple members in *Plasmodium falciparum*, all with an export signature. *BMC Genomics*, **6**(1), 30.

[39] Sogin, M. L. and Silberman, J. D. (1998). Evolution of the protists and protistan parasites from the perspective of molecular systematics. *International Journal for Parasitology*, **28**(1), 11–20.

[40] Solyakov, L., Halbert, J., Alam, M. M., Semblat, J.-P., Dorin-Semblat, D., Reininger, L., Bottrill, A. R., Mistry, S., Abdi, A., Fennell, C., Holland, Z., Demarta, C., Bouza, Y., Sicard, A., Nivez, M.-P., Eschenlauer, S., Lama, T., Thomas, D. C., Sharma, P., Agarwal, S., Kern, S., Pradel, G., Graciotti, M., Tobin, A. B., and Doerig, C. (2011). Global kinomic

and phospho-proteomic analyses of the human malaria parasite *Plasmodium falciparum*. *Nature Communications*, **2**, 565.

[41] Striepen, B., Jordan, C. N., Reiff, S., and van Dooren, G. G. (2007). Building the perfect parasite: cell division in apicomplexa. *PLoS Pathogens*, **3**(6), e78.

[42] Tewari, R., Straschil, U., Bateman, A., Böhme, U., Cherevach, I., Gong, P., Pain, A., and Billker, O. (2010). The systematic functional analysis of *Plasmodium* protein kinases identifies essential regulators of mosquito transmission. *Cell Host & Microbe*, **8**(4), 377–387.

[43] Ward, P., Equinet, L., Packer, J., and Doerig, C. (2004). Protein kinases of the human malaria parasite *Plasmodium falciparum*: the kinome of a divergent eukaryote. *BMC Genomics*, **5**(1), 79.

[44] World Health Organization (2012). *World malaria report : 2012*. Geneva, Switzerland.

Chapter 2

Methods for evolutionary analysis of protein families

In this chapter I review the methods used to characterize eukaryotic kinomes and lineage-specific kinase families and to place the results in the context of molecular and organismal function. Since apicomplexans are divergent from model organisms and are often difficult to study in the lab, computational biology has been important in obtaining preliminary insights into the functions of these species. A common pattern is to use computational methods initially to develop hypotheses, and then test the hypotheses experimentally. More recently, high-throughput experiments on *P. falciparum* and *T. gondii* have provided data that can be mined and integrated; again, the insights obtained for these two species can also be applied to apicomplexan relatives with the appropriate *in silico* methods.

2.1 Data resources

2.1.1 Pathogen-specific databases

To date, the genomes of 15 apicomplexan species have been fully sequenced and at least partially annotated: *Babesia bovis*, *Cryptosporidium hominis*, *C. muris*, *C. parvum*, *Eimeria tenella*, *Neospora caninum*, *Plasmodium falciparum*, *P. berghei*, *P. chabaudi*, *P. knowlesi*, *P. vivax*, *P. yoelii*, *Theileria annulata*, *T. parva*, and *Toxoplasma gondii*.

GeneDB [25], under the umbrella of Wellcome Trust Sanger Institute (WTSI) Pathogen Genomics, provides a community resource for collecting and accessing the data produced by these projects, including whole-genome sequences, EST data, and community annotation projects. Several other genomes were produced by the J. Craig Venter Institute (JCVI). In addition, sequences have been deposited directly in the universal sequence repositories (NCBI GenBank, EBI ENA, DDBJ).

A family of websites provides an efficient and user-friendly entry point to these and many other large-scale data for specific apicomplexan species: PlasmoDB (*Plasmodium* spp.), ToxoDB (*Toxoplasma gondii* strains, *Neospora caninum*, *Eimeria tenella*, and a planned *Sarcocystis neurona*), CryptoDB (*Cryptosporidium* spp.), and the recently added PiroplasmDB (*Theileria* spp. and *Babesia bovis*). All of these are united behind a portal for eukaryotic pathogens, EuPathDB, formerly ApiDB [4, 5]. EuPathDB also aggregates functional genomics data such as mRNA expression from microarray and RNA-Seq experiments, and proteomics and phospho-proteomics data from mass spectrometry. Datasets are typically taken from published articles and uploaded by the authors with the assistance of EuPathDB staff. The website provides a useful "strategies" interface in which different queries can be combined to filter data sets for specific properties; results can be browsed online, saved or downloaded in batch.

Protein structures

The Protein Data Bank (PDB) is well established as the canonical repository for protein structural data. The Structural Genomics Consortium (SGC) has taken on the challenge of solving neglected parts of the protein structural space on a per-family basis with specific attention to protein kinases [15]. The University of Toronto branch of the SGC has in particular focused on solving the structures of kinases in apicomplexans and other pathogenic protozoa, and since 2004 have deposited many novel structures of apicomplexan kinases, including both eukaryote-conserved and lineage-specific subfamilies, in PDB for public use. Specific findings from this work on kinases in *P. falciparum*, *T. gondii* and *C. parvum* have been described, with particular focus on the calcium-dependent protein kinase family [3, 42, 43]; however, the SGC has also released a number of structures ahead of any

manuscript publication, and these can also be accessed from PDB.

2.1.2 Protein kinase classifications and profiles

The eukaryotic protein kinase (ePK) superfamily is hierarchically classified into eight major groups: (AGC), calcium- and calmodulin-dependent kinases (CAMK), casein kinase 1 and its relatives (CK1), relatives of the cell-cycle control kinases CDK, MAPK, GSK3 and CLK (CMGC), receptor guanylate cyclase (RGC), a group containing the yeast protein Sterile11 (STE), tyrosine kinases (TK), and tyrosine-kinase-like kinases (TKL) [17]. Several families and subfamilies have been defined within each group, as well as a number of families that do not share the characteristics of any of the major groups (known as the "Other" group).

Families that do not share identifiable sequence homology to the "typical" ePKs are designated atypical protein kinases (aPK); some of these, such as Alpha, PIKK and RIO, nonetheless share the overall bilobate structural fold of ePKs, and are known as protein kinase-like kinases (PKL), while others such as pyruvate dehydrogenase kinase (PDHK) and nucleoside diphosphate kinase (NDK) adopt entirely different folds and appear to have independently evolved the functionality of protein phosphorylation [27].

The online database KinBase provides the classification and protein sequences of the kinomes of many model organisms, along with the characteristic domain architectures, descriptions, phyletic profiles, and sequence alignments of each of the recognized families. These family designations correspond to the first comparative analysis of human kinases and the model organisms yeast, fruit fly and nematode [26], and this terminology has been broadly adopted by other annotators.

Other databases have emerged to associate additional types of information with kinases. KinG [21], a database of "kinases in genomes," provides classifications based on sequence analysis methods, with specific attention to the domain architectures characteristic of each kinase family. Kinomer [28] provides automated kinome annotations for a broader range of model organisms, including *Plasmodium falciparum*. In addition, Kinomer provides a service to classify user-supplied sequences using a group-specific HMM profile set, following the same hierarchical classification scheme, though only to the level of major groups and selected atypical families.

The Protein Kinase Resource [33] is an integrated resource for studying protein kinase sequence and structure, providing an interactive visualization of protein kinase structures from PDB alongside with relevant information from UniProt. A focused ontology for protein kinases, ProKinO, has also been developed to provide a controlled vocabulary of terms and relationships unifying kinase sequence, structure and functional information [16].

2.2 Bioinformatic methods for protein subfamily classification

Given the abundance of information available to be mined for new insights, data-driven approaches to the characterization of apicomplexan kinomes are appealing.

2.2.1 Sequence similarity

Many of the kinases in the annotated apicomplexan genomes were assigned functional descriptions based on close matches to homologs, typically based on BLAST search and functional domains identified by Pfam's HMM profile search. The annotation of parasite genomes is often organized through GeneDB, an online community resource provided by the Wellcome Trust Sanger Institute Pathogen Genomics group that combines these automated results with in-progress annotations from curators [25].

The detail of kinase annotations can be improved with group- or family-specific HMM profiles. Kinomer uses such profiles based on the previously annotated kinomes of many species to provide accurate group-level kinase classifications, as well as improve the overall sensitivity of searches over a generic protein kinase profile [28]. Kinannote is another automated kinase classifier based on family-specific sequence profiles; an early version of it was applied to the kinome of the mushroom *Coprinopsis cinerea* [38]. An alternative to HMM profiles is position-specific scoring matrices, as implemented in PSI-BLAST [2] and the related tool MAPGAPS [30].

Domain architectures, as determined by Pfam or similar services, can provide clues to a kinase's classification. For example, members of the cGMP-dependent protein kinase (PKG)

family characteristically contain a series of cyclic-nucleotide-binding domains on the same protein sequence as the kinase domain, and similarly, calcium-dependent protein kinase (CDPK) is associated with four calcium-binding "EF-hand" domains.

2.2.2 Evolution-based methods

Orthology with a characterized kinase in another species serves as a strong signal for transferring functional annotations. The most literal approach to determine orthologous groups of genes is to infer a gene tree using an appropriate phylogenetic model, and compare the topology of the resulting gene tree to that of the accepted species tree. Genes which have been replicated through speciation rather than gene duplication are considered orthologs. Since the species relationships between all of the fully sequenced apicomplexans have been established [22], this approach is feasible with individual apicomplexan kinomes or with specific kinase groups families shared across multiple apicomplexan species. At the cost of some accuracy, orthologous groups can also be inferred in large data sets through a combination of reciprocal best BLAST hits across species and distance-based clustering, as implemented together in the program OrthoMCL [23]. The orthology relationships determined across a large number of complete genomes, including most of the apicomplexan species discussed here, are available through OrthoMCL-DB [7].

Within a kinase group, family, subfamily or ortholog group, it is then useful to examine patterns of conservation and selection in aligned sequences in order to identify possible sites of adaptation, subfunctionalization and neofunctionalization. Peixoto et al. [35] and Reese et al. [37] used the ratio of nonsynonomous and synonymous SNPs in aligned codons to identify regions and sites of positive selection in rhoptry kinases. Talevich et al. [39] used binomial tests of amino acid frequencies and a Bayesian pattern partitioning procedure, as implemented in CHAIN [29], to identify instances of change/gain of function as well as find taxa that share the same selective constraints.

There are several ways misclassification of kinases can occur. When classification is based on similarity to a single sequence, as with BLAST, it is possible to find highly significant matches to paralogous proteins; this risk is greater in a highly expanded protein superfamily such as the protein kinases, and can be compounded by the bias in sequence

representation in databases toward model organisms and the specific gene subfamilies their genomes contain. Curated sequence profiles constructed from diverse sequence sets such as the UniRef50 database, can reduce the bias in taxon sampling across all eukaryotes and thus assists classification of kinase sequences which have diverged from those of model organisms. However, the problem remains that an orphan sequence will be assigned to the best-matching query profile even if it represents a paralogous subfamily which is not represented in the profile set. In the following studies I conducted, I addressed this problem by constructing the ePK sequence profile database using a hierarchical scheme which allows an atypical sequence to be classified according to a broader group (such as CMGC) if it does not show substantial similarity to a more specific family. This reduces the likelihood of incorrect assignment of novel kinase sequences to an established subfamily, and facilitates the identification of divergent ePKs that may warrant deeper investigation, such as the unique MAPK subfamily found in all apicomplexans. In addition, I have continually updated the sequence profiles in accordance with published literature to include newly characterized sequences and ePK families.

2.3 Methods for finding divergent clades

After a gene duplication event produces two copies of the same gene in a single genome, one copy may evolve to gain or lose functions relative to the other copy which is presumed to retain the gene's original function [1, 44]. In this section I review methods to find distinct protein clades that may have emerged through a process similar to this, and to study the fate of this divergent copy, specifically to identify and characterize novel functions or mechanisms it may have evolved.

2.3.1 Phylogenetic analysis

Orthologous genes between species can be predicted by comparing the gene tree and species tree to infer duplication events (Figure 2.1). Note that orthology among multi-gene families is determined with respect to some common ancestral point which is considered the origin of the family of interest. For example, CDPK subfamilies are paralogous with respect to

the root or origin of the CDPK family (assuming a single origin), but can be considered orthologs (or in-paralogs) with respect to a broader grouping such as the CAMK kinase group or the ePK superfamily. Since multiple levels of comparision possible, the point of reference must be decided to address the research problem in question.

Multiple sequence alignment

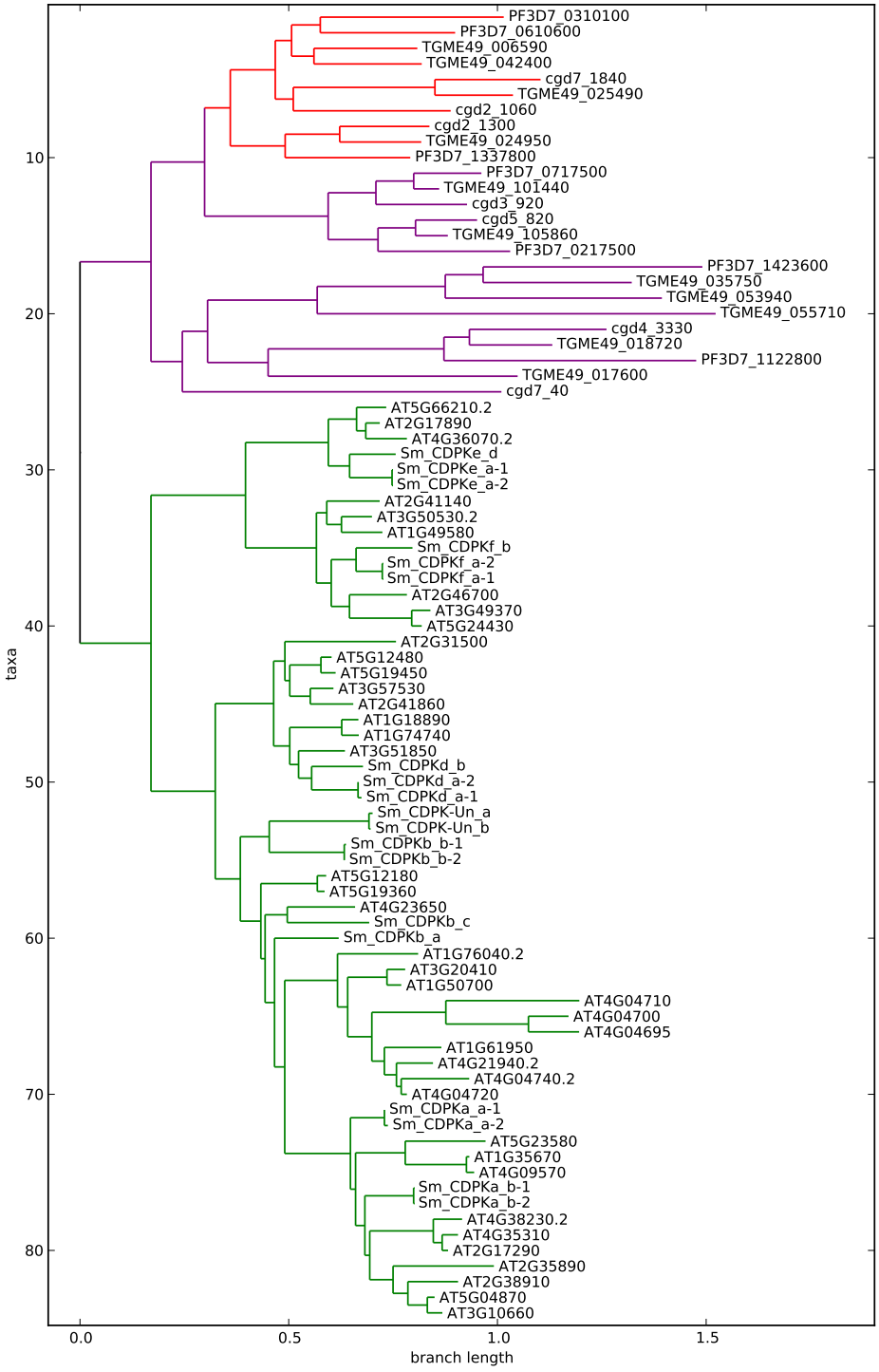
The accuracy of an inferred phylogenetic tree depends critically on the accuracy of the input character alignment. Because the optimal algorithmic solution quickly becomes intractable using dynamic programming methods [24, 41], a variety of methods have been devised to obtain approximate solutions to the problem [12, 19, 20, 31, 34, 40]. These *de novo* multiple sequence alignment methods are more likely to produce incorrect alignments if the input sequences are too few, too numerous, or too divergent. In those cases it is preferable to use structure-based or profile-based methods for sequence alignment.

HMMer can align many sequence to a single profile using the "hmmalign" command [11]. MAPGAPs [30] uses a set of several profiles representing subfamilies of a structurally related protein superfamily, each of which is aligned to the others. Thus, if the profile alignment has been constructed accurately (using 3D alignments of solved protein structures, for example), sequences belonging to divergent subfamilies can be aligned to each other with similar accuracy, potentially improving upon the alignments generated by the single-profile approach used in HMMer.

2.3.2 Orthology prediction methods and databases

Public databases that group genes into ortholog groups have been available for many years. The problem of inferring ortholog groups at the genome scale across many taxa has been

Figure 2.1 ***(following page)*****:** Maximum likelihood gene tree of the calcium-dependent protein kinase family (CDPK), inferred from conserved amino acid sites of the protein kinase domain using FastTree 2 [36]. Colors indicate an apicomplexan-specific subfamily of interest (red), other apicomplexan CDPKs (purple), and plant CDPKs (green).


PF3D7_0310100
PF3D7_0610600
TGME49_006590
TGME49_042400
cgd7_1840
TGME49_025490
cgd2_1060
cgd2_1300
TGME49_024950
PF3D7_1337800
PF3D7_0717500
TGME49_101440
cgd3_920
cgd5_820
TGME49_105860
PF3D7_0217500
PF3D7_1423600
TGME49_035750
TGME49_053940
TGME49_055710
cgd4_3330
TGME49_018720
PF3D7_1122800
TGME49_017600
cgd7_40
AT5G66210.2
AT2G17890
AT4G36070.2
Sm_CDPKe_d
Sm_CDPKe_a-1
Sm_CDPKe_a-2
AT2G41140
AT3G50530.2
AT1G49580
Sm_CDPKf_b
Sm_CDPKf_a-2
Sm_CDPKf_a-1
AT2G46700
AT3G49370
AT5G24430
AT2G31500
AT5G12480
AT5G19450
AT3G57530
AT2G41860
AT1G18890
AT1G74740
AT3G51850
Sm_CDPKd_b
Sm_CDPKd_a-2
Sm_CDPKd_a-1
Sm_CDPK-Un_a
Sm_CDPK-Un_b
Sm_CDPKb_b-1
Sm_CDPKb_b-2
AT5G12180
AT5G19360
AT4G23650
Sm_CDPKb_c
Sm_CDPKb_a
AT1G76040.2
AT3G20410
AT1G50700
AT4G04710
AT4G04700
AT4G04695
AT1G61950
AT4G21940.2
AT4G04740.2
AT4G04720
Sm_CDPKa_a-1
Sm_CDPKa_a-2
AT5G23580
AT1G35670
AT4G09570
Sm_CDPKa_b-1
Sm_CDPKa_b-2
AT4G38230.2
AT4G35310
AT2G17290
AT2G35890
AT2G38910
AT5G04870
AT3G10660
taxa
10
20
30
40
50
60
70
80
0.0
0.5
1.0
1.5
branch length

approached with automated methods. One approach which has been implemented by several teams is to group sequences by reciprocal best BLAST hits: Perform an all-versus-all BLAST search to identify each protein's closest hit in each other genome; those hits for which the query is likewise the best hit performing the search the other direction are possible orthologs. This is implemented in the programs TribeMCL [13] and OrthoMCL [23], and a database of OrthoMCL results run on many whole genomes is available as OrthoMCL-DB [7]. Other teams have used profile-based approaches, including PHOG [8], based on the predictions of the programs FlowerPower and SCI-PHY [6]. KinBase [26] also maps orthologous genes between model organisms and several other species, though it is limited to genes that code for protein kinases.

2.3.3 Sequence profile construction and curation for protein families

The program Fammer (http://github.com/etal/fammer) partially automates the process of sequence profile construction and curation for protein families and subfamilies. Designed for use with MAPGAPS and HMMer 3.0, Fammer creates a tree of profiles, with higher-level alignments representing broader groups and the top-level alignment representing the entire protein superfamily of interest. Several sub-commands are available to: build HMMer 3.0 and MAPGAPS profiles from a directory tree; scan protein sequences to assign hits to best-matching HMM profiles; add new sequences to the profile tree by searching a target database; refine sequence profiles using leave-one-out validation; and cluster a large family-level sequence alignment into phylogenetically supported subfamilies. After profile construction, the program can be applied to scan a proteome to automatically identify and classify the kinome with high accuracy.

2.4 Statistical comparison of clades (sub-alignments)

The selective pressures or constraints on a site are partly revealed through the mutation rates at the site among orthologous proteins. Negative selection is observed as conservation of a site relative to the average mutation rate, while faster mutation rates at a site suggest neutral or positive selection. The effects of selection on individual sites in a protein can be

quantified using an alignment of orthologous gene sequences. One method to distinguish between the two involves comparing the rates of nonsynonymous mutations (assumed to produce potentially functional changes which are subject to selection) to synonomous mutations (assumed to be under approximately neutral selective pressure); if this ratio (dN/dS) is greater than 1, this indicates that the site or gene in question is under positive selection.

At the amino acid level, conservation of a site relative to the other sites in the alignment indicates selective pressure. By comparing two or more sets of sequences, as is done by the programs CHAIN [29] and CladeCompare (http://github.com/etal/cladecompare), statistical tests can identify the sites that show the greatest difference in selective pressures, or different in equilibrium states of the sites, between the compared sets.

2.4.1 Statistical comparison of alignments & models of site divergence

Ball-in-urn model

The urn model poses the question: If we sample N sequences and get k residues of consensus type, what are the odds given the background frequency p? This follows the binomial distribution:

$$P(x|k, N, p) = \sum_{i=k}^{N} \binom{N}{i} p^i (1-p)^{N-i}$$

This test is implemented in CHAIN and CladeCompare. In CHAIN, this test statistic for a sampling of "pattern" sites is used as the optimization criterion for a Markov Chain Monte Carlo algorithm which attempts to assign sequences to "foreground" and "background" sets in order to maximize the contrast between them at the selected pattern sites [32].

G-test

The G-test is a goodness-of-fit test between residue frequencies observed in foreground versus those expected based on the background composition [10]. The test is conceptually

similar to the chi-squared test, but is more robust on smaller samples, including sets with counts of 0.

$$G = 2 \sum_{i \in a.a.} O_i \ln \frac{O_i}{E_i}$$

The test statistic follows the chi-squared distribution with 19 degrees of freedom, for the 20 amino acids minus one:

$$G \sim \chi^2_{19}$$

This model is implemented in CladeCompare. In addition to gain-of-function sites, this test can also reveal loss-of-function sites when comparing a subfamily to its broader family, unlike the urn model.

The program CladeCompare implements several fast statistical tests for finding diagnostic sites between two given clades, using a modular "strategies" framework for modeling different types of site divergence. Unlike CHAIN, there is no Bayesian resampling of the two given sets. Instead, the user must first identify the clades of interest, typically through phylogenetic methods.

2.4.2 Sequence weighting

Both statistical tests described above assume observations are independent. However, because the sequences in each clade are phylogenetically related, they are by definition not independent. Thus, a weighting scheme must be applied to the sequences to correct for the phylogenetic relatedness between sequences [14].

In both CHAIN [32] and CladeCompare, the aligned sequences in each set are weighted according to the Henikoff heuristic [18], following the same approach as and PSI-BLAST [2].

In brief, for each column of the alignment, the algorithm counts the number of rows and the number of occurrences of each amino acid type among all rows within the column. A per-column weight of 1 is first divided by the number of distinct residue types, then for each type, divided again by the number of sequences sharing that residue type and assigned to

each corresponding sequence. (For example, in a column composed of the residues [‘A’, ‘V’, ‘V’, ‘V’], there are two amino acid types ‘A’ and ‘V’; since ‘A’ occurs once, the first sequence receives a weight of $\frac{1}{2}$, and since ‘V’ occurs three times, the remaining weight is divided evenly between then, so each receives a weight of $\frac{1}{6}$.) These site weights are summed across all columns to compute weights for each sequence, and typically normalized for further calculations.

In CladeCompare, the effective overall number of sequences is estimated using another heuristic which, to our knowledge, has not been applied previously. The purpose of this step is to scale the weights obtained above according to the number of distinct residues that would be observed in an equally sized set of random, independent sequences. This value is estimated as the expected number of distinct residues to be observed in a column of given height (following a multinomial distribution, or for speed, pre-computed for a given number of rows by simulation), multiplied by the number of aligned columns, skipping columns which are mostly gaps. This value, the total number of “independent” residues, is then divided by the number of sequences in the alignment to obtain an estimate of the average sequence length. Finally, for each sequence, the sum of per-site weights is divided by the average sequence length. The sum of these final weights is the effective number of independent sequences in the alignment. This method is implemented in the supporting library BioFrills (http://github.com/etal/biofrills).

2.5 Structural mapping

The placement of identified sites of interest into a structural context on a solved protein crystal structure can allow mechanistic interpretation of the features that are unique to the protein subfamily being investigated.

We use CladeCompare and other in-house scripts to map sites of significant constrast onto PDB structures according to a multiple sequence alignment that includes the sequences of the foreground and background sets, as well as the primary sequence of the protein crystal structure itself. The alignment itself is generated using MAPGAPS 1.0 [30] or HMMer 3.0 [11]; in either case, the alignment is made in reference to a previously constructed

profile representing a fixed set of "consensus" or "match" columns, as well as sequence-specific insertions and deletions. Thus, if the foreground, background and PDB sequences are all aligned with the same MAPGAPS or HMMer profile, the equivalent (i.e. homologous) "consensus" columns can be easily identified in all three sets. This property allows us to automatically map significant alignment sites to the structure using the aligned amino acid sequence of the protein structure. CladeCompare uses this information to generate a script in the syntax of the molecular viewer program PyMOL [9] to visualize selected sites on a given PDB structure.

Bibliography

[1] Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M., and Dessimoz, C. (2012). Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Computational Biology*, **8**(5), e1002514.

[2] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402.

[3] Artz, J. D., Wernimont, A. K., Allali-Hassani, A., Zhao, Y., Amani, M., Lin, Y.-H., Senisterra, G., Wasney, G. A., Fedorov, O., King, O., Roos, A., Lunin, V. V., Qiu, W., Finerty, P., Hutchinson, A., Chau, I., von Delft, F., Mackenzie, F., Lew, J., Kozieradzki, I., Vedadi, M., Schapira, M., Zhang, C., Shokat, K., Heightman, T., and Hui, R. (2011). The *Cryptosporidium parvum* Kinome. *BMC Genomics*, **12**(1), 478.

[4] Aurrecoechea, C., Brestelli, J., Brunk, B. P., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O. S., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J. C., Kraemer, E. T., Li, W., Miller, J. A., Nayak, V., Pennington, C., Pinney, D. F., Roos, D. S., Ross, C., Srinivasamoorthy, G., Stoeckert, C. J., Thibodeau, R., Treatman, C., and Wang, H. (2010). EuPathDB: a portal to eukaryotic pathogen databases. *Nucleic Acids Research*, **38**(Database issue), D415–D419.

[5] Aurrecoechea, C., Barreto, A., Brestelli, J., Brunk, B. P., Cade, S., Doherty, R., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O. S., Heiges, M., Hu, S., Iodice, J., Kissinger, J. C., Kraemer, E. T., Li, W., Pinney, D. F., Pitts, B., Roos, D. S., Srinivasamoorthy, G., Stoeckert, C. J., Wang, H., and Warrenfeltz, S. (2012). EuPathDB: The Eukaryotic Pathogen database. *Nucleic Acids Research*, pages 1–8.

[6] Brown, D. P., Krishnamurthy, N., and Sjölander, K. (2007). Automated protein subfamily identification and classification. *PLoS Computational Biology*, **3**(8), e160.

[7] Chen, F., Mackey, A. J., Stoeckert, C. J., and Roos, D. S. (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research*, **34**(Database issue), D363–D368.

[8] Datta, R. S., Meacham, C., Samad, B., Neyer, C., and Sjölander, K. (2009). Berkeley PHOG: PhyloFacts orthology group prediction web server. *Nucleic Acids Research*, **37**(Web Server issue), W84–W89.

[9] Delano, W. (2011). The PyMOL Molecular Graphics System.

[10] Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

[11] Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, **7**(10), e1002195.

[12] Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**(5), 1792–1797.

[13] Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, **30**(7), 1575–1584.

[14] Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist*, **125**(1), 1–15.

[15] Gileadi, O., Knapp, S., Lee, W. H., Marsden, B. D., Müller, S., Niesen, F. H., Kavanagh, K. L., Ball, L. J., von Delft, F., Doyle, D. A., Oppermann, U. C. T., and Sundström,

M. (2007). The scientific impact of the Structural Genomics Consortium: a protein family and ligand-centered approach to medically-relevant human proteins. *Journal of Structural and Functional Genomics*, **8**(2-3), 107–119.

[16] Gosal, G., Kochut, K. J., and Kannan, N. (2011). ProKinO: An Ontology for Integrative Analysis of Protein Kinases in Cancer. *PLoS ONE*, **6**(12), e28782.

[17] Hanks, S. K. and Hunter, T. (1995). Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB Journal*, **9**(8), 576–596.

[18] Henikoff, S. and Henikoff, J. G. (1994). Position-based sequence weights. *Journal of Molecular Biology*, **243**(4), 574–578.

[19] Katoh, K., Kuma, K.-i., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, **33**(2), 511–518.

[20] Kemena, C. and Notredame, C. (2009). Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, **25**(19), 2455–2465.

[21] Krupa, A., Abhinandan, K. R., and Srinivasan, N. (2004). KinG: a database of protein kinases in genomes. *Nucleic Acids Research*, **32**(Database issue), D153–D155.

[22] Kuo, C.-H., Wares, J. P., and Kissinger, J. C. (2008). The Apicomplexan whole-genome phylogeny: an analysis of incongruence among gene trees. *Molecular Biology and Evolution*, **25**(12), 2689–2698.

[23] Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, **13**(9), 2178–2189.

[24] Lipman, D. J., Altschul, S. F., and Kececioglu, J. D. (1989). A tool for multiple sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, **86**(12), 4412–4415.

[25] Logan-Klumpler, F. J., De Silva, N., Boehme, U., Rogers, M. B., Velarde, G., McQuillan, J. A., Carver, T., Aslett, M., Olsen, C., Subramanian, S., Phan, I., Farris, C., Mitra, S.,

Ramasamy, G., Wang, H., Tivey, A., Jackson, A., Houston, R., Parkhill, J., Holden, M., Harb, O. S., Brunk, B. P., Myler, P. J., Roos, D., Carrington, M., Smith, D. F., Hertz-Fowler, C., and Berriman, M. (2012). GeneDB–an annotation database for pathogens. *Nucleic Acids Research*, **40**(Database issue), D98–108.

[26] Manning, G., Plowman, G. D., Hunter, T., and Sudarsanam, S. (2002a). Evolution of protein kinase signaling from yeast to man. *Trends in Biochemical Sciences*, **27**(10), 514–520.

[27] Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002b). The protein kinase complement of the human genome. *Science*, **298**(5600), 1912–1934.

[28] Martin, D. M. A., Miranda-Saavedra, D., and Barton, G. J. (2009). Kinomer v. 1.0: a database of systematically classified eukaryotic protein kinases. *Nucleic Acids Research*, **37**(Database issue), D244–D250.

[29] Neuwald, A. F. (2007). The CHAIN program: forging evolutionary links to underlying mechanisms. *Trends in Biochemical Sciences*, **32**(11), 487–493.

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

[31] Neuwald, A. F. and Liu, J. S. (2004). Gapped alignment of protein sequence motifs through Monte Carlo optimization of a hidden Markov model. *BMC Bioinformatics*, **5**(1), 157.

[32] Neuwald, A. F., Kannan, N., Poleksic, A., Hata, N., and Liu, J. S. (2003). Ran's C-terminal, basic patch, and nucleotide exchange mechanisms in light of a canonical structure for Rab, Rho, Ras, and Ran GTPases. *Genome Research*, **13**(4), 673–692.

[33] Niedner, R. H., Buzko, O. V., Haste, N. M., Taylor, A., Gribskov, M., and Taylor, S. S. (2006). Protein kinase resource: an integrated environment for phosphorylation research. *Proteins*, **63**(1), 78–86.

[34] Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, **302**(1), 205–217.

[35] Peixoto, L., Chen, F., Harb, O. S., Davis, P. H., Beiting, D. P., Brownback, C. S., Ouloguem, D., and Roos, D. S. (2010). Integrative genomic approaches highlight a family of parasite-specific kinases that regulate host responses. *Cell Host & Microbe*, **8**(2), 208–218.

[36] Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE*, **5**(3), e9490.

[37] Reese, M. L., Zeiner, G. M., Saeij, J. P. J., Boothroyd, J. C., and Boyle, J. P. (2011). Polymorphic family of injected pseudokinases is paramount in *Toxoplasma* virulence. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(23), 9625–9630.

[38] Stajich, J. E., Wilke, S. K., Ahrén, D., Au, C. H., Birren, B. W., Borodovsky, M., Burns, C., Canbäck, B., Casselton, L. A., Cheng, C. K., Deng, J., Dietrich, F. S., Fargo, D. C., Farman, M. L., Gathman, A. C., Goldberg, J., Guigó, R., Hoegger, P. J., Hooker, J. B., Huggins, A., James, T. Y., Kamada, T., Kilaru, S., Kodira, C., Kües, U., Kupfer, D., Kwan, H. S., Lomsadze, A., Li, W., Lilly, W. W., Ma, L.-J., Mackey, A. J., Manning, G., Martin, F., Muraguchi, H., Natvig, D. O., Palmerini, H., Ramesh, M. A., Rehmeyer, C. J., Roe, B. A., Shenoy, N., Stanke, M., Ter-Hovhannisyan, V., Tunlid, A., Velagapudi, R., Vision, T. J., Zeng, Q., Zolan, M. E., and Pukkila, P. J. (2010). Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom Coprinopsis cinerea (Coprinus cinereus). *Proceedings of the National Academy of Sciences of the United States of America*, **107**(26), 11889–11894.

[39] Talevich, E., Mirza, A., and Kannan, N. (2011). Structural and evolutionary divergence of eukaryotic protein kinases in Apicomplexa. *BMC Evolutionary Biology*, **11**(1), 321.

[40] Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting,

position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**(22), 4673–4680.

[41] Wang, L. and Jiang, T. (1994). On the complexity of multiple sequence alignment. *Journal of Computational Biology*, **1**(4), 337–348.

[42] Wernimont, A. K., Artz, J. D., Finerty, P., Lin, Y.-H., Amani, M., Allali-Hassani, A., Senisterra, G., Vedadi, M., Tempel, W., Mackenzie, F., Chau, I., Lourido, S., Sibley, L. D., and Hui, R. (2010). Structures of apicomplexan calcium-dependent protein kinases reveal mechanism of activation by calcium. *Nature Structural & Molecular Biology*, **17**(5), 596–601.

[43] Wernimont, A. K., Amani, M., Qiu, W., Pizarro, J. C., Artz, J. D., Lin, Y.-H., Lew, J., Hutchinson, A., and Hui, R. (2011). Structures of parasitic CDPK domains point to a common mechanism of activation. *Proteins*, **79**(3), 803–820.

[44] Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, **18**(6), 292–298.

Chapter 3

Structural and evolutionary divergence of eukaryotic protein kinases in Apicomplexa

Eric Talevich, Amar Mirza and Natarajan Kannan (2011) *BMC Evolutionary Biology* 11:321.
Reprinted here with permission from the publisher.

Abstract

BACKGROUND: The Apicomplexa constitute an evolutionarily divergent phylum of protozoan pathogens responsible for widespread parasitic diseases such as malaria and toxoplasmosis. Many cellular functions in these medically important organisms are controlled by protein kinases, which have emerged as promising drug targets for parasitic diseases. However, an incomplete understanding of how apicomplexan kinases structurally and mechanistically differ from their host counterparts has hindered drug development efforts to target parasite kinases.

RESULTS: We used the wealth of sequence data recently made available for 15 apicomplexan species to identify the kinome of each species and quantify the evolutionary constraints imposed on each family of apicomplexan kinases. Our analysis revealed lineage-specific adaptations in selected families, namely cyclin-dependent kinase (CDK), calcium-dependent protein kinase (CDPK) and CLK/LAMMER, which have been identified as important in the pathogenesis of these organisms. Bayesian analysis of selective constraints imposed on these families identified the sequence and structural features that most distinguish apicomplexan protein kinases from their homologs in model organisms and other eukaryotes. In particular, in a subfamily of CDKs orthologous to *Plasmodium falciparum* crk-5, the activation loop contains a novel PTxC motif which is absent from all CDKs outside Apicomplexa. Our analysis also suggests a convergent mode of regulation in a subset of apicomplexan CDPKs and mammalian MAPKs involving a commonly conserved arginine in the αC helix. In all recognized apicomplexan CLKs, we find a set of co-conserved residues involved in substrate recognition and docking that are distinct from metazoan CLKs.

CONCLUSIONS: We pinpoint key conserved residues that can be predicted to mediate functional differences from eukaryotic homologs in three identified kinase families. We discuss the structural, functional and evolutionary implications of these lineage-specific variations and propose specific hypotheses for experimental investigation. The apicomplexan-specific kinase features reported in this study can be used in the design of selective kinase inhibitors.

3.1 Background

The parasitic protists which comprise the phylum Apicomplexa are responsible for human diseases of global importance, such as malaria (caused by *Plasmodium falciparum* and other members of the *Plasmodium* genus), cryptosporidiosis (*Cryptosporidium* species) and toxoplasmosis (*Toxoplasma gondii*), as well as the agricultural diseases babesiosis (*Babesia bovis* in cattle) and coccidiosis (*Eimeria tenella* in chickens) [82]. In recent years, understanding of the molecular biology and evolution of this phylum has improved dramatically; yet effective treatments for these diseases are still elusive, and there remains an urgent need for deeper research into the basic biology of apicomplexans [80].

Several traits make these pathogens difficult to target therapeutically. As eukaryotes, they share a number of pathways with their mammalian and avian hosts; as intracellular parasites, they have been observed to quickly develop resistance to pharmaceutical treatments [86]. The identification of distinctive protein features which appear conserved across apicomplexan species, but not in their hosts, however, will aid the search for potential new targets for selective inhibition that are more likely to be safe and effective [47]. As protein kinases have been successfully targeted for inhibition in cancer, this diverse protein superfamily warrants consideration as a target for parasitic diseases as well [32, 80].

Recent whole-genome sequencing efforts have targeted a number of apicomplexan species [1, 14, 18, 21, 22, 42, 43, 46, 50, 75, 76, 97]. Several analyses of protein kinases in these organisms, in particular, have pointed out key signaling pathways [34, 61, 91], instances of expansion and loss of kinase gene families [83, 93], and emergence of novel protein kinase families [77, 84, 93], thus providing important insights into biological functions. These comparative studies have furthermore proposed hypotheses which have subsequently been validated by functional and structural studies [35, 61, 91].

The eukaryotic protein kinase (ePK) superfamily is classified into several major groups, corresponding to broad functional categories with distinguishing sequence and structural features [48, 62]. The presence of specific ePK groups and families in a genome is a key indicator of biological functions critical for an organism; likewise, missing groups or families indicate functions less critical for an organism's survival and reproduction. These proteins, and the fundamental cell processes in which they participate, are well characterized in

humans and several model organisms [62].

Previous efforts to perform detailed comparative analysis of apicomplexan kinases have largely focused on the kinomes of individual species within the genera *Plasmodium*, *Toxoplasma* and *Cryptosporidium* [7, 13, 18, 21, 66, 91, 94]. Thus, there is no global overview of the sequence and structural features that distinguish apicomplexan kinases collectively from their metazoan counterparts.

Sequence data from 15 apicomplexan species and several crystallographic structures of a variety of apicomplexan protein kinases are now available. We can use these data to perform a systematic comparison of protein kinases in apicomplexans and model eukaryotes to identify broadly conserved orthologous groups and distinctive residue-level differences.

In this study we use a bioinformatics approach to comprehensively analyze genomic and structural data sets. We perform an exhaustive comparison of apicomplexan kinomes, providing broad coverage of the phylum. We also perform a quantitative, residue-level analysis of the differences between kinases within the Apicomplexa and those in model eukaryotes, in particular humans. We use a Bayesian method [67] to rigorously quantify sequence differences between homologous protein kinases in apicomplexans and other eukaryotes, and reveal contrastingly conserved features that were not apparent previously. Where possible, we then place these sequence features in structural context to postulate specific hypotheses for experimental testing.

Our specific findings include: (i) a detailed accounting of the lineages in which the apicomplexan-specific kinase families FIKK and ROPK appear; (ii) a unique apicomplexan-specific subfamily of cyclin-dependent kinases (CDK), orthologous to *P. falciparum* crk-5, and the motifs that distinguish it; (iii) a hypothesized mechanism of activation by phosphorylation, resembling that of MAP kinases, in a chromalveolate-specific subfamily of calcium-dependent protein kinases (CDPK); and (iv) a description of the adaptation of the substrate-recognition and docking sites in the CLK kinase family in a clade including apicomplexans and other chromalveolates, revealed by the co-evolution of a small set of key residues.

3.2 Results and Discussion

We identified and classified the eukaryotic protein kinases in a total of 17 genomes from 15 species, as well as the solved apicomplexan ePK structures in the Protein Data Bank [12]. We used our classification to broadly describe the conserved ePK families in the Apicomplexa and then performed a residue-level analysis of the lineage-specific differences within several conserved families: CDK, CDPK and CLK. We place our findings in the context of the known evolutionary history of apicomplexans and their relatives.

3.2.1 Kinome classification and composition: Variations within the Apicomplexa

Recent published evolutionary relationships of eukaryotes provide the basis for our genomic comparison [55]. In this study we have chosen model organisms representing major evolutionary splits — the emergence of Chromalveolata (a proposed super-kingdom of plastid-containing eukaryotes [2]), Alveolata (the kingdom comprising ciliates, dinoflagellates and apicomplexans [52]), and Apicomplexa — to illuminate the origin and divergence of the major ePK groups. For genomic comparison we use the parasitic dinoflagellate *Perkinsus marinus* as an outgroup to the Apicomplexa, the photosynthetic diatom *Thalassiosira pseudonana* as an outgroup to the Alveolata, and the yeast *Saccharomyces cerevisiae* as an outgroup to the Chromalveolata.

Apicomplexan kinome sizes are comparable to those of other unicellular protists

The number of ePKs identified in each of the surveyed apicomplexan genomes varies, with the coccidians (*Toxoplasma gondii*, *Neospora caninum* and *Eimeria tenella*) containing more ePKs than the haemosporidians (*Plasmodium* spp.), and the piroplasms (*Babesia bovis* and *Theileria* spp.) containing fewer (Table 3.1). *Cryptosporidium* spp., the most basal group of apicomplexans considered here, contain a similar number of ePKs to *Plasmodium* spp.

Taken as a percentage of total genome size, the proportions of kinases in apicomplexans are generally either comparable to the 2% observed in yeast and humans [62], as seen in the coccidians and *Cryptosporidium*, or reduced, as in the piroplasms and *Plasmodium*

Table 3.1: Total proteome and protein kinome sizes in each genome. Columns indicate species name, the number of ePKs found using our method, the number of protein-coding genes in each genome, and the calculated proportion of ePKs in each genome for comparison. Atypical protein kinases are excluded from all ePK counts.

Species	**ePKs**	**Genes**	**Ratio**
Plasmodium berghei	69	4904	1.41%
Plasmodium chabaudi	70	5131	1.36%
Plasmodium yoelii	62	5878	1.05%
Plasmodium knowlesi	65	5197	1.25%
Plasmodium vivax	65	5435	1.20%
Plasmodium falciparum	93	5491	1.69%
Theileria annulata	42	3793	1.11%
Theileria parva	43	4035	1.07%
Babesia bovis	43	3671	1.17%
Toxoplasma gondii GT1	137	8102	1.69%
Toxoplasma gondii ME49	146	7993	1.83%
Toxoplasma gondii VEG	133	7846	1.70%
Neospora caninum	141	7082	1.99%
Eimeria tenella	90	8786	1.02%
Cryptosporidium hominis	65	3886	1.67%
Cryptosporidium parvum	75	3805	1.97%
Cryptosporidium muris	77	3934	1.96%
Perkinsus marinus	251	23654	1.06%
Thalassiosira pseudonana	140	11673	1.20%
Saccharomyces cerevisiae	116	5797	2.02%

(Table 3.1). (Note that the quality of genome assemblies and gene model annotations varies, and these differences can affect the number of genes and kinases identified in each genome; the low kinase-to-gene ratios given for *P. yoelii* and *E. tenella* should therefore be interpreted with caution.) There is no evidence of the striking overall expansion of kinases seen in free-living ciliates such as *Paramecium tetraulia* (ePKs 6.6% of the genome [11]), which form a sister clade to Apicomplexa within the kingdom Alveolata. Rather, the number of kinases appears to scale with the total number of protein-coding genes in each genome, with small deviations.

Except for the coccidians and *P. falciparum* (which each contain dramatic expansions of novel kinase families, discussed below), the absolute number of kinases in each apicomplexan genome is markedly reduced relative to free-living eukaryotes (Table 3.1). The piroplasm kinome sizes, for instance, are less than twice the minimal kinome of 29 ePKs exhibited by another obligate intracellular parasite, *Encephalitozoon cuniculi* [64]. The pattern of genome compaction, occasionally offset by lineage-specific expansions of specific gene families, has been noted as a common mode of genomic evolution in unicellular pathogens [59] and apicomplexans specifically [57, 90]. Evidently, the ePKs have evolved according to some of the same adaptive strategies as the overall genomes of these parasites.

3.2.2 Survey of ePK major groups

We classified the kinases in each of the surveyed apicomplexans and model organisms according to a hierachical scheme based on seven major ePK groups, enabling a direct comparison of the group composition between kinomes (Figure 3.1). The CMGC and CAMK groups are especially well conserved across eukaryotes, indicating that the cell functions performed by these proteins are fundamental and essential for eukaryotic life. The casein kinase 1 group (CK1) is conserved in at least one copy among all eukaryotes as well. The tyrosine kinase (TK) and receptor guanylate cyclase (RGC) groups are entirely missing from the Apicomplexa, which has previously been noted [63, 77], as well as the three outgroup genomes. There is an apparent reduction, relative to the outgroup *P. marinus* and *T. pseudonana*, of the cyclic-nucleotide- and calcium/phospholipid-dependent kinases (AGC group) in most of the Apicomplexa (Figure 3.1). The coccidians have between 9

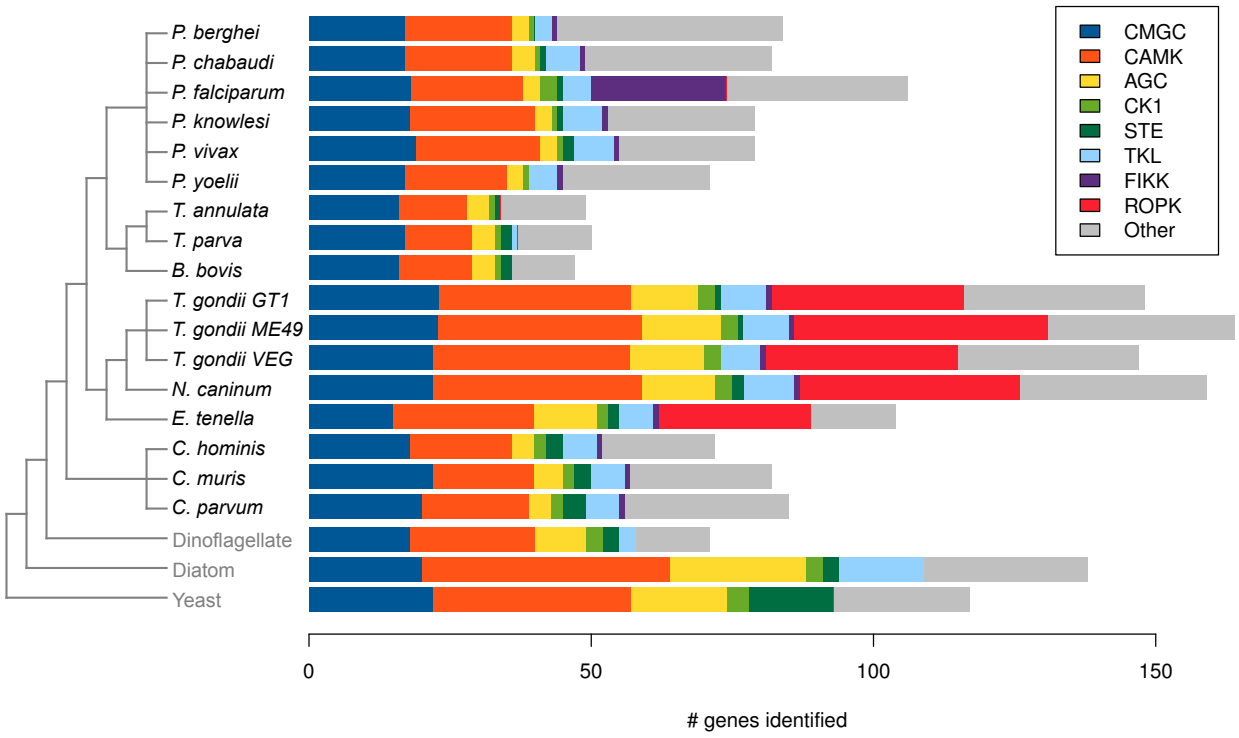


Figure 3.1: Composition of protein kinase major groups and selected apicomplexan-specific families (FIKK and ROPK) in each of the surveyed genomes. The schematic species tree along the left edge is constructed from published sources [2, 58, 78, 96], and includes three outgroup kinomes for comparison: the dinoflagellate *Perkinsus marinus*, the diatom *Thalassiosira pseudonana*, and the yeast *Saccharomyces cerevisiae*. In the stacked bar chart associated with each genome, block width indicates number of genes found belonging to each major group of eukaryotic protein kinases; total bar width indicates total kinome size.

and 13 members of the AGC group, while other apicomplexans have 3 to 5 AGC kinases; PKA is the only AGC family that is found in every genome. The additional AGC members in coccidians appear as 1–3 copies of several known families, suggesting that AGCs were mostly lost in the other lineages and conserved or slightly amplified in coccidians, rather than a significant expansion in coccidians relative to the common ancestor. An even more dramatic loss of kinase families along all lineages is apparent in the STE group, which we discuss below. The tyrosine-kinase-like group (TKL) shows greater variation, appearing in some abundance in coccidians and *Plasmodium* spp. but absent from piroplasms, except for a single instance in *T. annulata* (Figure 3.1). The "Other" group designation collects all the ePK families that share the ePK fold and sub-domain architecture (unlike atypical protein kinases), but do not fall cleanly into any of the recognized major ePK groups found in the human kinome [62]. Many apicomplexan kinases fall in the Other group (Figure 3.1), reflecting their deep evolutionary divergence from humans, the reference genome for the commonly accepted kinase classification scheme [62]. Atypical protein kinases, such as the ABC and RIO families, were excluded from this analysis.

Conservation of cell-cycle-associated kinases (CMGC) in chromalveolates

The CMGC group is named after four protein kinase families it contains: cyclin-dependent kinase (CDK), mitogen-activated protein kinase (MAPK), glycogen synthase kinase (GSK), and cdc-like kinase (CLK, also called LAMMER) [62]. These kinases are involved in various aspects of cell cycle control, and are highly conserved throughout Eukaryota. Though apicomplexans, as obligate parasites, are able to depend on their host for survival, these signaling mechanisms for various aspects of cell cycle control are retained. Their life cycles are generally complex, often involving both a primary and a secondary host, encysted phases, and sudden trigger of reproduction and proliferation in response to some chronological or external stimulus [89]. This seems to suggest elaborate signaling and regulatory mechanisms, and points toward specialization of CMGC kinases in the Apicomplexa [93].

The most abundant family within the CMGC group is CDK; it is found in 3–6 copies in each apicomplexan genome, and 7–11 copies in the outgroup genomes. The CDC2 subfamily

of CDK is found in at least one copy in every genome, while some species contain single instances of additional CDK families. There are also 1–4 CDKs in each genome which could not be classified into known subfamilies, leaving open the possibility of lineage-specific adaptations in these unclassified copies. GSK occurs in 1–3 copies in each apicomplexan genome, and 1–5 in the outgroup genomes, reflecting an essential and conserved role in cellular function. Likewise, MAPK and casein kinase IIα (CK2) are present in a small number of copies in each of the apicomplexan and other eukaryotic genomes surveyed. The MAPK subfamily ERK7 is found in a single copy in every apicomplexan genome, while ERK1 is missing from *Plasmodium* spp. and the piroplasms. The RCK family, comprising the MAK and MOK subfamilies, is present in the three outgroup species but missing from the Apicomplexa.

The CLK and SRPK families, and some subfamilies of DYRK, are involved in phosphorylation of splicing factors such as SR proteins [25, 56]. We found 2–4 DYRKs in each apicomplexan genome. The most conserved subfamily of these, PRP4, was found in 1 copy in each genome except *E. tenella*. A plant-specific subfamily of DYRK, called DYRKP, was found only in coccidians and the outgroups *P. marinus* and *T. pseudonana*. We found 1 copy of CLK in every surveyed genome, and SRPK in 1 copy in all except *P. marinus*, which has 3 copies.

The close relationship between CLK, SRPK and DYRK can confound homology-based classification attempts. However, the families can be distinguished by the presence of family-specific inserts [20] and by the replacement of the arginine in the kinase-conserved catalytic "HRD" motif with threonine ("HTD") in CLK and SRPK, and cysteine ("HCD") or alanine ("HAD") in various DYRK subfamilies [53]. The first comprehensive study of an apicomplexan kinome [93] identified 4 putative CLKs in *P. falciparum*, assigning the names PfCLK-1 through PfCLK-4. Our classification confirmed PfCLK-1 [EupathDB:PF14_0431] as a CLK (discussed in detail below). PfCLK-4 [EupathDB:PFC0105w] has recently been characterized as an SRPK [28]. We assigned PfCLK-3 [EupathDB:PF11_0156] to the PRP4 subfamily of DYRK, supported by the presence of the "HAD" motif in the catalytic loop and homology with putative PRP4 kinases in each of the other *Plasmodium* species. Our classifier placed PfCLK-2 [EupathDB:PF14_0408] in the CMGC group but did not find support for a

more specific family. The portion of the sequence in kinase subdomain X, which is broadly conserved as the "EHLAMMERILG" in CLKs [98], is "RFIYSIVSYIG" in PfCLK-2 — there is no sequence identity except for the C-terminal glycine. PfCLK-2 has the catalytic loop motif "HCD", characteristic of most DYRK subfamilies. The protein sequence also contains long inserts in the catalytic domain in the same locations as those of SRPK. A recent study of PfCLK-1 and PfCLK-2 [3] confirmed SR protein phosphorylation activity and found that PfCLK-1 is localized primarily to the nucleus of the cell, like most CLKs, but PfCLK-2 is found in both the nucleus and the cytoplasm, as has been observed in SRPKs in other eukaryotes [40]. We suggest that this protein is unique, with characteristics of both the SRPK and DYRK families, and that the regulatory functions suggested by typical CLK family members do not fully describe the roles of PfCLK-2 in the cell. The corresponding ortholog group in OrthoMCL-DB [23] [OrthoMCL:OG5_165485] is specific to the *Plasmodium* genus, further evidence that PfCLK-2 and its orthologs are paralagous to apicomplexan CLKs and have diverged significantly.

Distribution of calcium signaling kinases (CAMK) in Eukaryota

Calcium signaling plays an important role in eukaryotic cell biology. Calcium ions serve as important second messengers in signaling pathways, regulated by the calcium- and calmodulin-dependent kinase (CAMK) group [66]. In apicomplexans, calcium signaling regulates motility and other processes associated with host invasion [13].

There are multiple conserved CAMK members in each surveyed genome, though we observed more variation in gene family sizes here than in the CMGC group. We found 19–31 putative CAMK genes in each coccidian genome, 13–16 in *Cryptosporidium* spp., 11–13 in *Plasmodium* spp. and 7 in each piroplasm (Figure 3.1). The closely related dinoflagellate *P. marinus* has 69 putative CAMK genes, and the more distantly related diatom *T. pseudonana* has 42. This points to a slight overall reduction of CAMK and CAMK-like protein kinases in coccidians, and more dramatic reductions in the other apicomplexan lineages, relative to the dinoflagellate and diatom (Figure 3.1). This follows with the overall conservation or reduction of total kinome sizes in each of the genomes (Table 3.1).

The calcium-dependent protein kinase (CDPK) family within CAMK is of particular

interest, as its role in parasite invasion has been investigated recently by several teams [61, 74, 95]. Like plants and some other protists, apicomplexan genomes contain multiple members of the CDPK family [13]. We found 6 CDPKs in *P. falciparum*, 5 in each of the other *Plasmodium* species, 4–5 in the piroplasms, 11–14 in the coccidians and 7–9 in *Cryptosporidium* spp. In *T. gondii* and *N. caninum* there were also 7–10 members of the CAMK group that could not be classified into a known family. The greater number of CDPK copies and unclassified CAMKs in coccidians accounts for most of the apparent expansion of the CAMK group in that lineage relative to other apicomplexans.

Loss and divergence of STE kinase families in apicomplexan lineages

The STE group includes a variety of kinases which participate in MAPK signaling cascades upstream from the MAPK protein [62]. The key families in the group are STE20 (MAP4K), STE11(MAPKKK/MAP3K) and STE7 (MAPKK/MEK), which form a phosphoryl signaling cascade terminating with the phosphorylation of a MAPK on its activation loop at a conserved TxY motif [71]. This MAPK cascade is highly conserved in most eukaryotes, so it is surprising that the STE group has been largely lost from the Apicomplexa, as has been noted previously [5, 93].

According to our analysis, the STE group is entirely missing from the piroplasms, while in the *Plasmodium* genus only *P. knowlesi* and *P. vivax* each retain a single STE gene which could not be further classified into a known STE family (Figure 3.1). There were also unclassified STEs in *T. gondii* strains GT1 and ME49, *E. tenella* and *Cryptosporidium* spp. We did not find any STEs in *T. gondii* strain VEG.

The STE11 family was not found in any of the surveyed apicomplexan genomes. One STE20, showing closest resemblence to the FRAY subfamily (homologs of human OSR1), was found in *N. caninum*; the other apicomplexans had none. STE7 instances appear in *N. caninum*, *C. hominis* and *C. parvum*. For comparison, *Perkinsus marinus* contains 1 instance of STE11 and two instances of STE20, in the MST and PAKA subfamilies (homologs of human MST2 and PAK2, respectively). The ciliate *Tetrahymena thermophila* has multiple representives of STE11, STE20, STE7, and other STE families [37].

Features of the two MAPKs of *P. falciparum* illustrate how apicomplexans can compensate

for the lack of a complete MAP signaling cascade. Pfmap-1 [EupathDB:PF14_0294] was identified as a member of the ERK7 family of MAPK [93], and retains the conserved TxY activation loop motif of most MAPKs. Pfmap-2 [EupathDB:PF11_0147], however, could not be assigned to a known MAPK subfamily in earlier analyses [5, 93] or in ours. In Pfmap-2, the activation loop motif TxY is replaced by TSH [31], and we also note a long insert of about 26 amino acids in the activation loop N-terminal to the TSH motif. Orthologs of Pfmap-2 identified in OrthoMCL-DB [OrthoMCL:OG5_138034] appear in each of the apicomplexan genomes surveyed here, and also retain the long insert in the activation loop and a TSH or TGH motif in place of TxY. Pfmap-2 has been shown to be phosphorylated and activated by the kinase Pfnek-1 [EupathDB:PFL1370w] [33], which is not a member of the STE kinase group but in this case appears to be nonetheless serving as a MAP kinase kinase. As with Pfmap-2, orthologs of Pfnek-1 appear in each of the surveyed apicomplexans [OrthoMCL:OG5_129446]. The conservation patterns of these kinases suggest that the observations made of *P. falciparum*'s unique MAPK signaling mechanisms can be applied usefully to other apicomplexans.

FIKK, an apicomplexan-specific protein kinase family

FIKK is a divergent protein kinase family initially identified in *P. falciparum*, named for a conserved four-residue motif in the kinase subdomain II [93]. Previous studies have found 21 copies in *P. falciparum* and 6 in *P. reichenowi*, but single instances in other *Plasmodium* genomes, indicating rapid expansion along one branch within the genus [84]. In *P. falciparum*, FIKK proteins are generally exported to the host cell and often localized to the host cell membrane [72]. Recent work has found that some *P. falciparum* FIKKs are targeted to the Maurer's clefts, which are formed from or in connection with the parasitophorous vacuole membrane (PVM) as a transport mechanism and eventually reach the host cell surface [73]. A variety of functional domains have also been discovered in the N-terminal tail of the FIKK kinase domain, suggesting that the kinase domain and export signal allow trafficking of parasite proteins or other molecules to the host cell membrane [84].

In addition to the 21 recognized FIKKs in *P. falciparum* [84, 93], we found a single

copy of FIKK in every one of the surveyed apicompexan genomes except *Theileria* spp. and *Babesia bovis* (Figure 3.1). No homologs were found outside the Apicomplexa. The apparent absence of FIKK from the three piroplasm genomes is particularly intriguing. To rule out the possibility that this absence is simply the result of the FIKK gene model having not been included in the available proteomic sequences, we performed an additional search on the full set of translated ORFs from the genomic DNA sequence sets for these three species; again, no FIKK genes were found. The parsimonious conclusion is that the gene was lost along the piroplasmid evolutionary branch. This loss suggests there may be some difference in the physiology of piroplasmids that eliminates the need for the FIKK protein in those species.

We note with some interest that, in the process of entering a host cell, apicomplexans generally envelop themselves in a parasitophorous vacuoule constructed from the host cell membrane. (This is true of all of the species surveyed here.) Unlike *Plasmodium* spp. and most other apicomplexans, however, *Babesia* and *Theileria* species escape from their parasitophorous vacuole shortly after entering the host erythrocyte [18, 85]. Thereafter, the piroplasm interacts directly with the host cell cytoplasm, rather than through the membrane of a vacuole, potentially simplifying the signaling machinery needed by the parasite. Piroplasms are also nonmotile and show other reduced functions compared to other apicomplexans [82]. However, more study of the role of FIKKs and the interaction between the PVM and host cell in apicomplexan species outside *Plasmodium* is needed in order to refine this hypothesis.

ROPK family is specific to the coccidians

The rhoptries are a collection of vesicular organelles within the apical complex, a distinguishing feature of the Apicomplexa. They appear in all of the apicomplexans surveyed here [82]. During the invasion process, a number of proteins contained in the rhoptries are secreted through the apical complex into the parasitophorous vacuole, and in some cases the host cell cytosol [17]. The rhoptry kinase family (ROPK) comprises the protein kinases targeted to the rhoptry. ROPKs play a major role in the infection mechanism of *T. gondii* [15]; they have been characterized in *T. gondii* and to a lesser extent in *N. caninum* [77].

The sequences of ROPKs are divergent from other ePKs, but most can still be recognized by generic protein kinase search profiles [77]. Most rhoptry kinases appear to be catalytically inactive, lacking at least one residue of the catalytic "KDD" triad (the lysine and asparates normally conserved in ePK subdomains II, VI and VII [48]), but kinase activity has been demonstrated in ROP16 and ROP18 [38, 77]. Recent structural studies of ROP2 and ROP8 revealed a unique modification of the N-lobe of the kinase domain, in particular, and suggested important functional roles for these proteins, despite the absence of catalytic activity in these ROPKs [79].

We found the ROPK family only in the coccidian clade (Figure 3.1). Proteins associated with the rhoptries in other lineages appear to be unrelated to coccidian ROPKs or any other ePK families.

Our analysis included three strains of *T. gondii*, corresponding to the three classes of virulence: GT1 (Type I, high virulence), ME49 (Type II, intermediate virulence), and VEG (Type III, non-virulent) [87]. The most dramatic difference in kinase counts between the three strains of *T. gondii* appears in the ROPK family (Figure 3.1). We identified 40 ROPKs in *T. gondii* strain ME49, but 29 in GT1 and VEG. A simple clustering of the sequences (data not shown) did not reveal a clear separation of ME49 ROPK genes that would indicate an expansion in ME49, so the discrepency may instead be due to losses in the other two strains, or simply differences in the quality of genome assembly and annotation.

3.2.3 Sequence and structural features contributing to functional divergence

Our approach revealed several novel and distinct subfamilies within recognized ePK families. Within each family, we then performed a phylogenetic analysis of the protein sequences of kinase domains from apicomplexans and several diverse model organisms to identify putative ortholog groups that include several apicomplexan species, but no metazoan species (see Methods).

Statistical analysis of the sequences using the CHAIN program revealed distinctive sequence and structural features which distinguish apicomplexan kinases from their homologs in other eukaryotes. Specifically, we used each identified apicomplexan-specific ortholog

set as a query against a larger "main" set of sequences representing the corresponding kinase family (CDK, CDPK and CLK) taken from diverse eukaryotic species. CHAIN uses a Bayesian MCMC procedure to concurrently (a) partition the "main" set into a "foreground" of sequences that share distinct residue motifs found in the query, a "background" set of sequences that do not share those motifs, and an "intermediate" set that shares only some of the motifs; and (b) identify the alignment columns defining the motifs that distinguish the foreground and background sets [67]. We then used PyMOL [27] and a set of custom scripts leveraging Biopython [24] to map the most significant residue patterns onto aligned protein structures for comparative structural analysis.

Here we describe three proposed instances of lineage-specific divergence of apicomplexan kinases, within the CMGC and CAMK major groups, with an analysis of the sequence motifs and evolutionary histories that define them. Where crystallographic structures have previously been solved, we map sequence motifs onto the 3D structures to gain insight into possible regulatory mechanisms.

3.2.4 Orthologs of Pfcrk-5 form a novel subfamily of cyclin-dependent kinases

While each apicomplexan kinome contains multiple genes belonging to the cyclin-dependent kinase (CDK) family, we find a novel CDK subfamily which appears in a single copy in 14 of the 17 apicomplexan genomes surveyed, absent only from *Cryptosporidium* spp., and is not found outside Apicomplexa. This subfamily comprises the orthologs of *P. falciparum* Pfcrk-5 [EupathDB:PFF0750w]. This ortholog group is equivalent to a group in OrthoMCL-DB [OrthoMCL:OG5_150603], but with the addition of an ortholog we identified in *Theileria parva* [Genbank:TP04_0791].

The subfamily is distinguished by a unique PTxC motif in the activation loop (Pfcrk-5 positions 255–258), which is strikingly conserved relative to other CDK members in diverse eukaryotes, and absent from diverse eukaryotic homologs, as determined by CHAIN analysis (Figure 3.2). In eukaryotic homologs, the residues at the location of the PTxC motif are most often histidine, glutamate and valine. The threonine in position 254 is also found as either threonine (usually) or serine (more rarely) in homologs; this site is equivalent to

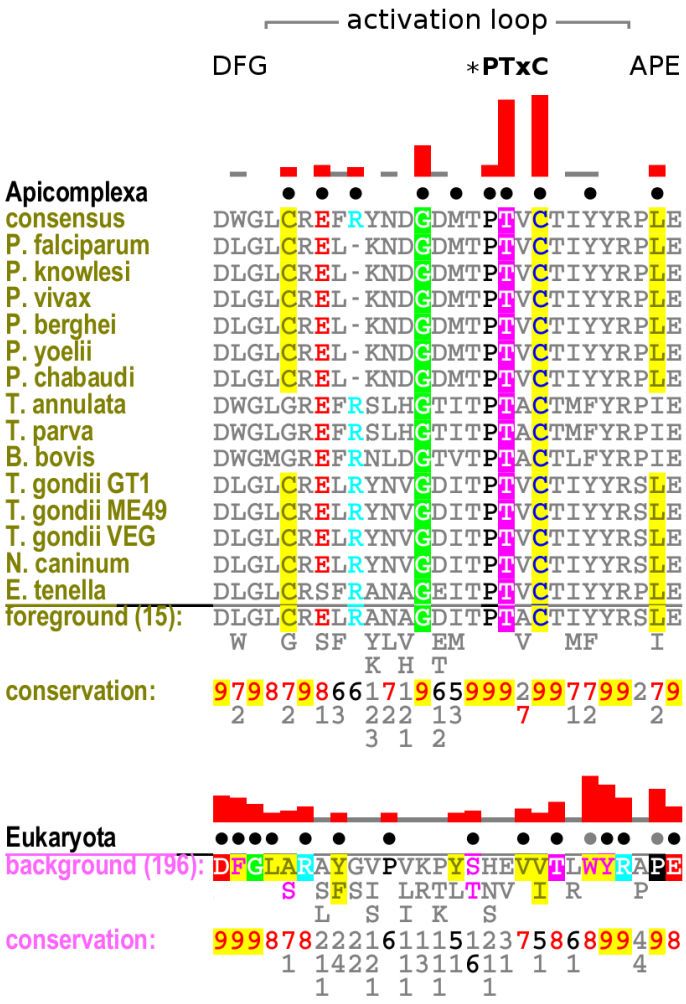


Figure 3.2: CHAIN alignment of the activation loop in the Pfcrk-5-like CDK subfamily ("Foreground") compared to the corresponding region in a large set of diverse eukaryotic CDK sequences ("Background"). The kinase-conserved DFG and APE motifs bordering the activation loop are indicated at the top, along with the subfamily-conserved PTxC motif. An asterisk indicates the position of the threonine observed to be phosphorylated in other CDKs, conserved in both the foreground and background. The histogram above each sequence set represents the differential levels of conservation between the two sets at each position, using logarithmic scaling. Dots above each alignment column indicate the contrasting conservation pattern determined by CHAIN. Note that the Apicomplexa (foreground, top) and Eukaryota (background, bottom) sets have different conservation patterns. In the sequence alignment itself, columns of the conserved pattern are colored according to the consensus residue type. The consensus residue types are listed below the alignment. Weighted residue frequencies are shown in the following rows, in units of integer tenths (e.g. "9" indicates conservation of 90–100%). The number of sequences in each set are shown in parentheses.

T160 in human Cdk2, where phosphorylation of this residue dramatically increases CDK catalytic activity, apparently stabilizing the substrate-binding site by forming a network of hydrogen bonds with surrounding side chains [65].

While T254 is conserved in most CDKs across Eukaryota, the apicomplexan-conserved residues P255, T256 and C258 are strikingly different from those in CDKs of other eukaryotes (Figure 3.2). In particular, T256 in this subfamily appears most often as a glutamate in other CDKs, including the closest-matching known CDK subfamily, CDC2, though it is not strongly conserved overall in eukaryotic CDKs. Given the similarity in chemical properties between glutamate and phosphothreonine, it is tempting to speculate that T256 is a phosphorylation site in this subset of apicomplexan CDKs. An alternative hypothesis is that the residues in the PTxC motif may provide contact points for the substrate, as has been observed for the equivalent residues in the human homolog Cdk2 [19]. Human Cdk2 belongs to the CDK subfamily CDK2, not CDC2, but contains the motif HEVV in place of Pfcrk-5's PTVC, as most CDC2s do. In a solved structure of human Cdk2 [PDB:1QMZ], the residue V164, equivalent to C258 in Pfcrk-5, is located spatially between the bound substrate and the APE motif. It is possible that C258 in Pfcrk-5 and its orthologs packs hydrophobically against the equivalent region in this subfamily. This could also explain the co-conserved change of the APE motif to PLE (Figure 3.2). However, the absence of a solved 3D structure for any member of this subfamily prevents further analysis of the functional role of these residues. Although four structures of apicomplexan CDKs have been published [PDB:1V0O, PDB:1V0B, PDB:1OB3, PDB:2QKR], none of them correspond to genes from the Pfcrk-5 subfamily.

To assess whether the members of this putative subfamily should instead be assigned to the known CDK subfamily CDC2, we used CHAIN again to compare this subfamily to sequences representing the CDC2 subfamily. The same distinguishing pattern of PTxC in the activation loop appears in this comparison as well. In *P. falciparum*, the CDKs Pfcrk-1–4 have all previously been annotated as "cdc2-related" kinases, and have been characterized in previous studies [29, 30]. The canonical CDC2 in *P. falciparum*, as identified by our analysis, is protein kinase 5 [EupathDB:MAL13P1.279], which has the more typical "HEVV" motif in place of Pfcrk-5's "PTVC". Thus, the genes in this apicomplexan-specific subfamily

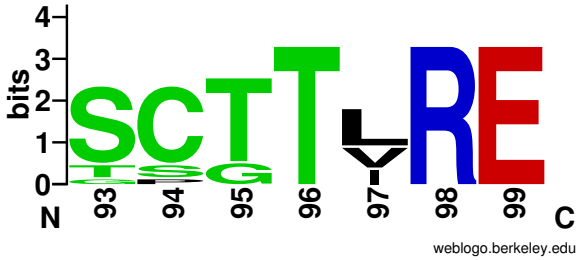


Figure 3.3: Logo of the aligned activation loop sequences in members of the Pfcrk-5-like CDK subfamily, generated by WebLogo [26]. Letter height represents information content; large letters indicated residues conserved within the subfamily.

appear to be paralogous to the known CDC2 subfamily, and may therefore have unique functional roles.

Distinct subfamilies of CDK are sometimes named after the conserved residue sequence in the cyclin-binding helix in the N-lobe of the kinase domain, known as the PSTAIRE helix in CDKs or more generally as the αC helix in protein kinases [30, 65]. In the proposed alveolate-specific subfamily the consensus sequence of the αC motif is SCTTLRE, at Pfcrk-5 sequence positions 93-99 (Figure 3.3). It is not yet known whether Pfcrk-5 is dependent on cyclin binding for activity, like PfPK5, Pfmrk and Pfcrk-3, or independent, like PfPK6 [29, 30]. None of these residues appear in the CHAIN pattern, however, indicating that the individual residues at these positions may occur in some non-apicomplexan CDKs as well, and that this motif did not necessarily co-evolve with the activation loop motif that characterizes this apicomplexan-specific subfamily.

We also identify 5 large inserts in the kinase domain which are conserved to varying degrees across all 14 apicomplexan species, but not found in any other known subfamily of CDK. These inserts occur between subdomains I and II, III and IV, IV and V (in the coccidians), VII and VIII (after the conserved PLE, corresponding to APE in most ePKs, and extending over 100 amino acids in *Plasmodium* spp.), and X and XI (an extension of the CMGC insert, normally involved in substrate binding [53]). The inserts appear to be hydrophilic, and are generally conserved at the sequence level within each genus, but less clearly between different genera, indicating rapid evolution relative to the structurally

conserved portions of the kinase domain.

3.2.5 Features of a chromalveolate-specific CDPK subfamily point to a MAPK-like mode of regulation

The CDPK family is characterized in green plants, and instances of it are also recognized in some protists (specifically, chromalveolates), but there are none in metazoans [49, 66] — this observation by itself encourages study of the CDPK family as a parasite-specific therapeutic target in human diseases. Each apicomplexan genome contains multiple CDPKs; we find and discuss a novel subfamily of these here. The subfamily is found in all of the surveyed apicomplexans as well as the dinoflagellate *Perkinsus marinus*, the ciliates *Tetrahymena thermophila* and *Paramecium tetraulia*, and the diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*, indicating that the subfamily is shared by a clade within the Chromalveolata. It includes the *P. falciparum* protein PfCDPK5, which has been shown to play a key regulatory role during the parasite's blood stage [36]. The subfamily does not correspond cleanly to OrthoMCL-DB groups, but contains some members of the main CDPK group [OrthoMCL:OG5_126600] as well as some small lineage-specific groups (e.g. [OrthoMCL:OG5_170347]).

CHAIN analysis highlighted several key residues that distinguish this subfamily from the larger set of chromalveolate CDPKs, of which two are most striking: an arginine in the αC helix, and a threonine or serine in the activation loop. The conservation of these two residues within the subfamily, but not in the broader CDPK family, suggests they have evolved under a shared functional constraint. Notably, the structure of a member of the subfamily in *C. parvum*, CpCDPK2 [EupathDB:cgd7_1840], has been solved in complex with an inhibitor [PDB:3F3Z] and in *apo* form [PDB:2QG5] [7]. The distinguishing residues numbered according to the crystal structures of CpCDPK2 are R69 and T184. Guided by CHAIN analysis, we compared these structures with that of another *C. parvum* CDPK outside the subfamily, CpCDPK1 [EupathDB:cgd3_920, PDB:3DFA], to understand the sequence and structural basis for possible *C. parvum* CDPK functional divergence.

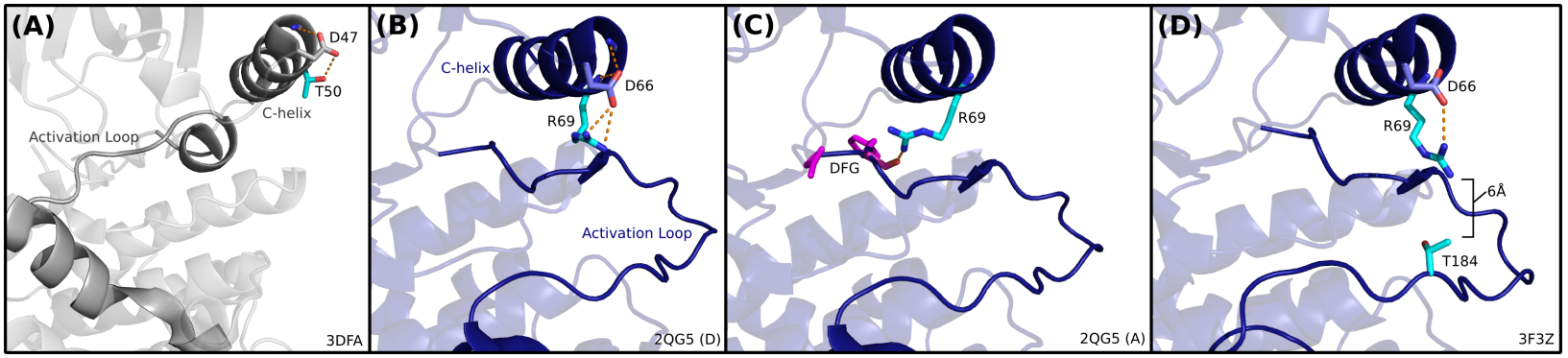


Figure 3.4: Structures of several different CDPKs in *C. parvum*, demonstrating several proposed interactions for the αC helix arginine distinctive of an alveolate-specific CDPK subfamily.
(A) A member of the background set of CDPKs [PDB:3DFA] has a threonine (T50), shown in cyan, in position to form a hydrogen bond with an aspartate (D47), gray, which caps the αC helix. This threonine corresponds to the subfamily-conserved arginine; however, the threonine here is not conserved in the background set of CDPKs.
(B) In a structure of a member of the CDPK subfamily [PDB:2QG5], the subfamily-conserved arginine (R69, cyan) appears similarly positioned to interact with the aspartate (D66, blue) at the end of the αC helix, potentially stabilizing the cap.
(C) Chain A of the same structure shows the distinctive arginine oriented inward, capable of hydrogen-bonding with the kinase-conserved DFG motif (side chains colored magenta).
(D) In another structure of the same CDPK-subfamily protein [PDB:2QG5], the arginine is positioned toward a subfamily-conserved threonine in the activation loop (T184), shown in cyan. The distance between the R69 and T184 side chains is 6Å, which could accomodate a phosphate group attached to the threonine and a hydrogen bond between the phosphothreonine and the arginine.

We analyzed the structural interactions associated with R69 and T184 in the two available crystal structures of CpCDPK2 [PDB:2QG5, PDB:3F3Z] (Figure 3.4). In one of the CpCDPK2 structures [PDB:2QG5], R69 adopts two distinct conformations (Figure 3.4B–D). In chain A, R69 is positioned to form a hydrogen bond to the backbone of a residue (D66) at the αC helix N-terminus, while in chain D, R69 appears to form a 3.1Å hydrogen bond to the backbone of the DFG motif glycine, located at the N-terminus of the activation segment. In chain B, R69 is oriented outward, in a solvent-exposed position. (While the CpCDPK2 structure is presented as three chains, the biological unit has not been described.) B-factors and the different orientations of this residue in each chain indicate that the R69 side chain is flexible in this structure.

In the other CpCDPK2 structure [PDB:3F3Z], R69 is oriented toward the side-chain of T184, separated by a distance of 6.0Å. Previous reports show that threonine autophosphorylation in the activation loop is prevalent in apicomplexan CDPKs [94, 95]. We therefore hypothesize that this threonine (T184^{2QG5,3F3Z}) could also serve as a phosphorylation site in the alveolate-specific CDPK subfamily.

Shared features of MAP kinases suggest a common regulatory mechanism

To obtain additional insights into the role of R69 and T184 in CpCDPK2 functions, we identified and analyzed crystal structures of kinases that contain both an αC arginine and an activation-loop threonine at positions equivalent to CpCDPK2 R69 and T184, respectively. To allow for the flexibility and variable length of the activation loop, we also examined positions adjacent to T184. This revealed a large number of MAPK structures, including human and mouse p38, where a αC-helix arginine (R67) and activation-loop threonine (T180) appear to perform roles analogous to those proposed for R69 and T184 in CpCDPK2. In a crystal structure of p38α [PDB:3NNX], R67 (equivalent to R69 in CpCDPK2) hydrogen bonds with the glycine backbone of the DFG motif at a distance of 2.8Å, in a manner analogous to CpCDPK2. Another structure of p38α complexed with a different inhibitor [PDB:3NNV] shows a similar interaction occurring at 3.2Å. In a structure of mouse p38α [PDB:3PY3], phosphorylated on both a threonine (T180) and a tyrosine (T182) in the activation loop, the αC arginine (R67) coordinates with the phospho-threonine. Thus

the conserved arginine functions as a switch: upon phosphorylation, the activation-loop phospho-threonine interacts with the αC arginine, promoting inter-domain closure and stabilizing the αC helix in an active conformation [4]. An equivalent mechanism has been described for p38γ [PDB:1CM8] as well [10].

The phosphorylated threonine in p38 corresponds to the TxY motif which is conserved across MAPKs [71], including JNK and ERK1. A sequence alignment of CpCDPK2 and PfCDPK5 along with human p38, JNK1 and ERK1 shows that the CDPK subfamily-conserved threonine is centered on the MAPK TxY motif. Another threonine, located 4 residues C-terminal to this site, is broadly conserved in both MAPK and CDPK.

We draw parallels between the observed conformations of CpCDPK2 and p38. An analogous role for R69 and T184 in CpCDPK2 would suggest a regulatory mechanism wherein phosphorylation of T184 leads to kinase activation by repositioning R69 from a DFG-stabilizing or solvent-exposed orientation toward the activation loop, consequently moving the regulatory αC helix in an active conformation.

In a paralogous *C. parvum* CDPK that does not belong to the CpCDPK2 subfamily, CDPK1 [PDB:3DFA, EupathDB:cgd3_920], the αC arginine is replaced by T50, and the activation loop threonine by D165 (Figure 3.4A). Rather, the interactions described here are distinctive of the alveolate-specific subfamily of CDPKs including CpCDPK2. The minor expansion of the CDPK family in chromalveolates has created an evolutionary opportunity for certain copies of CDPK genes to subfunctionalize, adapting the additional regulatory role for promoting phosphorylation-dependent inter-domain closure.

3.2.6 Lineage-specific mechanisms of substrate recognition and binding in CLK

Within the CLK family, we again find a residue pattern that distinguishes chromalveolate CLKs from those in all other eukaryotic lineages. This pattern appears in all apicomplexans surveyed, as well as several dinoflagellates, ciliates, diatoms, and the brown alga *Ectocarpus siliculosus*. The phyletic distribution of this set of co-conserved motifs points to an origin near the base of Chromalveolata, prior to the emergence of alveolates, and a deep evolutionary divergence between chromalveolates and metazoans.

These chromalveolate CLKs are distinguished most prominently by residues in the substrate-recognition and docking sites. Numbered according to the representative *P. falciparum* protein serine/threonine kinase 1 [EupathDB:PF14_0431], also called PfLAMMER [60], the distinguishing residues include Q739, L772 and R775 in the primary docking site, N736 and S755 in the secondary substrate-recognition site, and the acidic residue D653 in the αE helix (Figures 3.5, 3.6 and 3.7; discussed below). Taken together, this set of amino acid differences represents a statistically significant partition between chromalveolate and other eukaryotic CLK sequences.

A crystallographic structure of PfLAMMER is available [PDB:3LLT], but has not been previously discussed in detail. We compared this structure to two human CLK homologs, Clk1 [PDB:1Z57, PDB:2VAG] and Clk2 [PDB:3NR9], as well as human SRPK1 [PDB:1WAK], to predict structural and functional roles of the lineage-specific residues.

Mechanisms of substrate recognition, binding and processive phosphorylation

The typical substrate of CLK is an SR protein, characterized by an N-terminal RNA-binding domain and an unstructured C-terminal tail of varying length, called the RS domain, which is enriched in arginine and serine, often occurring as "RS" dipeptide repeats [25]. The SR proteins in a cell play multiple roles in spliceosome formation and mRNA splicing activity, including regulation of alternative splicing [45, 51]. CLKs are closely related to SRPKs, which also phosphorylate the RS domain of SR proteins. Both kinases are constitutively active, and perform processive phosphorylation on the RS domain of an SR protein substrate, proceeding in the carbonyl-to-amino direction along the substrate peptide [92]. However, differences in substrate binding and the extent of RS domain phosphorylation between SRPK and CLK allow interplay between these proteins to affect the activity and subcellular localization of the SR protein in a complementary fashion [69]. Thus, the complementary regulation of SR proteins by CLK and SRPK has an important functional impact on mRNA splicing in the cell [28].

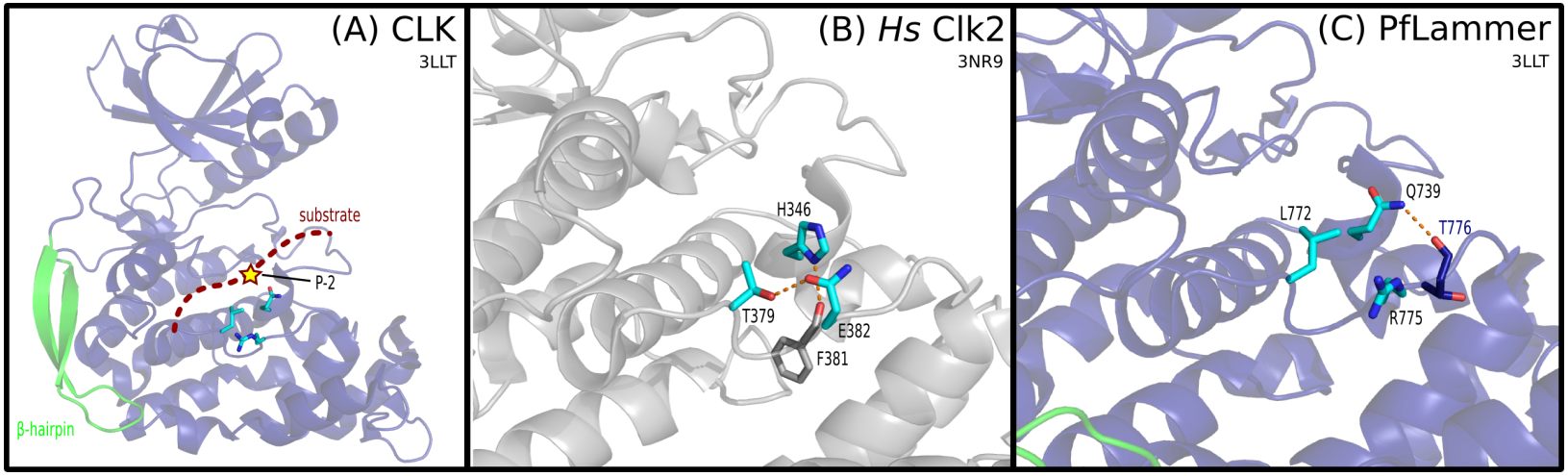


Figure 3.5: Three contrastingly conserved residues involved in substrate recognition and docking in human Clk2 [PDB:3NR9] and the *P. falciparum* CLK, PfLAMMER [PDB:3LLT].
(A) Global view of the docking site, illustrating the position of the substrate RS domain and phosphorylation site. The contrastingly conserved resides are shown in cyan.
(B) Human Clk2. A trio of constrastingly conserved residues (cyan), along with a nearby phenylalanine (gray), form a network of hydrogen bonds. The conserved histidine (H346) is positioned to interact with the substrate P−2 position.
(C) In PfLAMMER, the three residues (cyan) are conserved as different types. A glutamine (Q739) replaces the histidine in human Clk2 seen to interact with the substrate P−2 position. The hydrogen bonding network is different: A leucine (L772) replaces the threonine seen in Clk2; an arginine (R775), corresponding to a glutamate in Clk2, is directed away from the other two conserved residues; and the glutamine (Q739) instead forms a hydrogen bond with a nearby threonine.

Substrate-recognition site: Three residues responsible for initial recognition of the substrate, Q739, L772 and R775, are contrastingly conserved within the chromalveolate clade (Figure 3.5). In human Clk2, the equivalent residues H346, T379 and E382 form the substrate-recognition site, with the histidine interacting with the substrate P$-$2 residue (P indicates the phosphorylatable residue), preferentially selecting for glutamate [20]. In PfLAMMER the histidine is replaced by a glutamine; the change in chemical properties suggests a different substrate preference for the protein. Additionally, in human Clk2 the three conserved residues form hydrogen bonds with each other and with a nearby F381 (Figure 3.5B); in PfLAMMER, Q739 only potentially forms a hydrogen bond with nearby residue T776, while L772 appears in place of human T379, losing the bond (Figure 3.5C). The E382 in Clk2 is replaced in PfLAMMER by R775, which does not form hydrogen bonds with the nearby trio of substrate-recognition residues but is instead oriented outward, free to interact with other atoms, such as the substrate (Figure 3.5C). The location of the residues L772 and R775 in the loop connecting the αF and αG helices, in particular, is also significant because the αF-αG loop is also involved in substrate binding; it is therefore likely that the chromalveolate-specific variations observed in this loop also contribute to a difference in substrate recognition.

P+1 binding pocket: As mentioned above, apicomplexan CLKs have conserved lineage-specific residues located at the substrate-binding pocket. One such residue is the chromalveolate-specific asparagine (N736) in the P+1 pocket. N736 is conserved as a glutamine in SRPKs, as a serine in human Clk1 and Clk2, as a cysteine in GSK, and as a valine in CDK [53]. These variations may contribute to the substrate specificity by subtly altering the geometry of the P+1 pocket. Alternatively, the variation observed at the P+1 pocket may reflect the unique mode of allosteric coupling between the substrate-binding site and active site in CMGC kinases. Notably, both the backbone and side-chain of N736 in PfLAMMER are involved in hydrogen bonding to the backbone of the catalytically important HTD motif (Figure 3.6B), while in other CMGC kinases, the coupling between the P+1 pocket and catalytic site is largely mediated through backbone hydrogen bonds (Figure 3.6C,D).

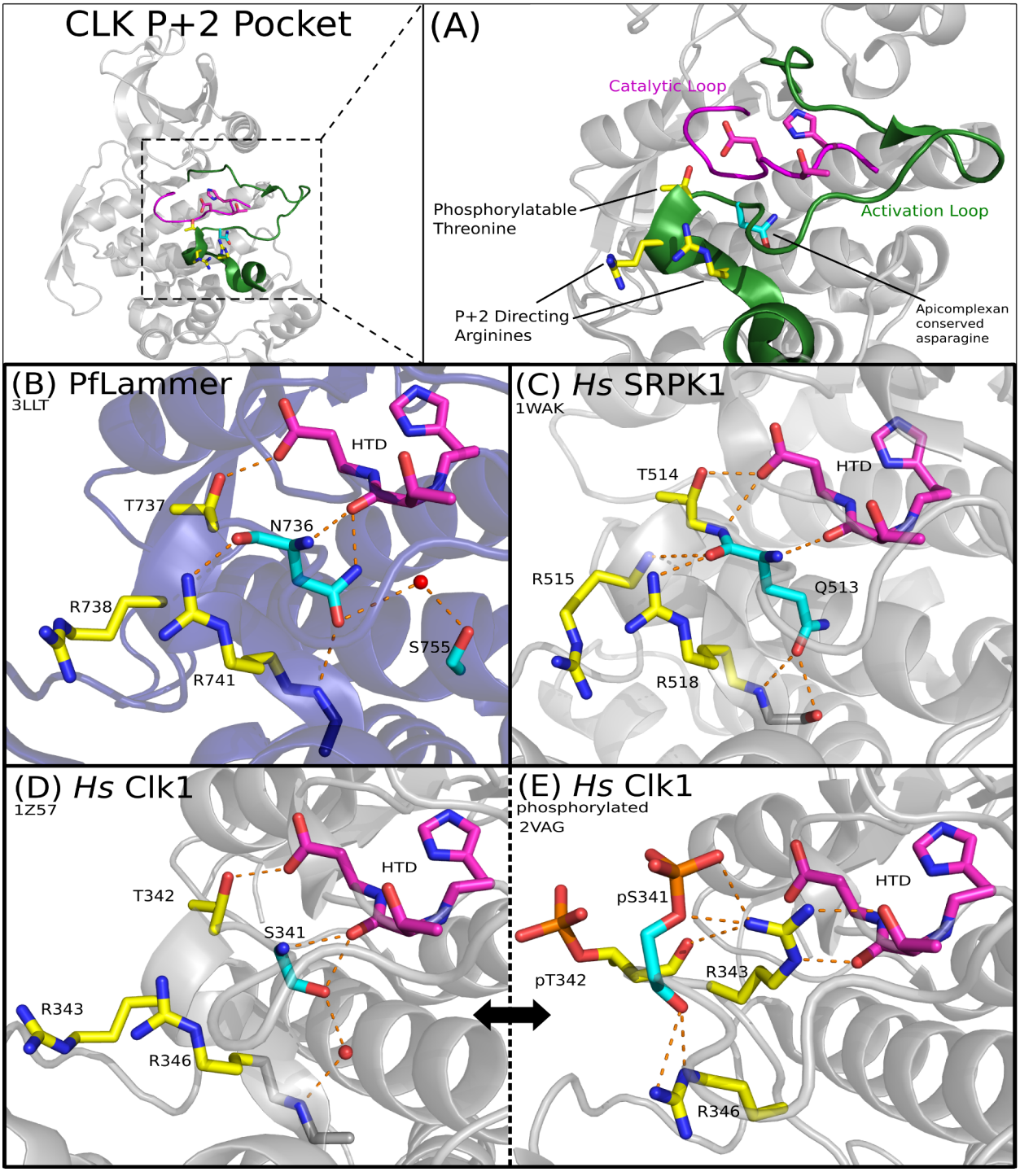
We used the program Coot [39] to examine N736 in the structure of PfLAMMER and found that its backbone conformation lies in a disallowed region of the Ramachandran plot,

indicating that torsion-angle strain occurs here. This position has been reported to be in a strained position in SRPK1 and other CMGCs prior to substrate binding; substrate binding relieves this strain, highlighting the importance of this residue in the substrate binding mechanism [53]. It is also significant that in one of the human Clk1 structures [PDB:2VAG] (Figure 3.6E), S341 (equivalent to N736 in PfLAMMER) and T342 are phosphorylated, which dramatically alters the geometry of the P+1 pocket and inactivates the kinase [81]. This indicates that the P+1 pocket is conformationally malleable and can contribute to the unique modes of allosteric regulation.

Proline-directed and processive phosphorylation: The CLK family, and related members of the CMGC group, conserve several distinctive residues in the substrate-binding site that contribute to the substrate specificity of CMGC kinases. One such residue is the

---

Figure 3.6 ***(following page)*****:** Interactions between key residues in the substrate-binding region and the catalytic HTD motif are mediated by conserved residues in the activation loop.
(A) Structural context of features in PfLAMMER [PDB:3LLT], showing the activation loop in green and the catalytic loop in magenta. Conserved residues are displayed in "sticks" representation. A contrastingly conserved asparagine, distinctive of chromalveolate CLKs, is indicated in cyan, and three other residues conserved throughout the CLK family are shown in yellow.
(B) In PfLAMMER, the distinctive asparagintex e (N736) forms hydrogen bonds with the CMGC-conserved arginine (R741), the backbone of the alanine in the APE motif, the backbone of the threonine in the catalytic HTD motif, and, mediated by a water molecule, a subfamily-conserved serine in the αF helix.
(C) In human SRPK1, several of the hydrogen bonds formed by the glutamine Q513 are analogous to those formed by the N736 in apicomplexans.
(D) and **(E)** Two structures of human Clk1. In the unphosphorylated structure [PDB:1Z57], left, the serine corresponding to PfLAMMER N736 (S341) and the adjacent CLK-conserved threonine (T342) are oriented in an "in" conformation, interacting with the catalytic motif (HTD) but not with the conserved arginines (R343, R346). In the phosphorylated structure [PDB:2VAG], right, the serine (pS341) and threonine (pT342) are flipped to an "out" conformation, breaking the interaction with the catalytic motif. One arginine (R343) moves to occupy the area vacated by the phosphorylated serine S341, while the other (R346) now interacts with the backbone of the phosphorylated serine. Phosphates are shown in orange. Images of PDB structures were rendered using PyMOL [27].


CLK P+2 Pocket
(A)
Catalytic Loop
Phosphorylatable Threonine
Activation Loop
P+2 Directing Arginines
Apicomplexan conserved asparagine
(B) PfLammer
3LLT
HTD
T737
N736
R738
S755
R741
(C) Hs SRPK1
1WAK
T514
HTD
R515
Q513
R518
(D) Hs Clk1
1Z57
HTD
T342
S341
R343
R346
(E) Hs Clk1
phosphorylated
2VAG
pS341
HTD
pT342
R343
R346

distinctive CMGC-arginine [53] (R741 in Figure 3.6B) located at the C-terminal end of the activation loop. The CMGC-arginine contributes to substrate specificity by creating a favorable hydrophobic environment for a proline at the P+1 position of the substrate. Specifically, the CMGC-arginine caps the backbone carbonyl oxygen of a residue (N736 in PfLAMMER) in the P+1 pocket that typically hydrogen bonds to the backbone amide of a residue at the P+1 position. Because proline lacks a backbone amide, the capping of carbonyl oxygen by the CMGC-arginine allows selective binding of substrates with proline at the P+1 position [19]. The presence of the CMGC-arginine and the hydrogen bonds in the P+1 pocket of PfLAMMER (Figure 3.6C) suggest that chromalveolate CLKs, like other CMGC kinases [25, 70], are likely to be proline-directed.

PfLAMMER also conserves the P−2 arginine (R738 in Figure 3.6C), which in human CLKs and SRPKs contributes to the processive phosphorylation of substrates by stabilizing a phosphorylated serine or threonine at the P−2 position in the substrate [20, 53]. This feature suggests that chromalveolate CLKs, like human and plant CLKs and SRPKs, may processively phosphorylate substrates with phosphorylatable serine or threonine at the P−2 position. Indeed, a search for protein sequences with an RNA-binding domain [Pfam:RRM_1] and "RS" repeat regions identified at least three possible SR proteins in *P. falciparum* [EupathDB:PF10_0217, PFE0865c, PFE0160c], each with orthologs in other apicomplexan species [OrthoMCL:OG5_127971, OG5_128933, OG5_127418].

Chromalveolate-specific features in the distal substrate-recognition site

The CLK family, as it appears in all eukaryotes including apicomplexans, has a characteristic β-hairpin insert in the C-lobe between the β7 and β8 strands, which blocks its SR protein substrate from docking in what is a distal substrate-recognition groove in other CMGCs (such as the MAP kinase p38) [20]. Blocking this docking interaction is critical for CLK substrate specificity, the primary means by which CLKs are regulated [20].

CHAIN analysis revealed a strikingly conserved acidic residue (aspartate or glutamate) in the αE helix of chromalveolate CLKs which in other eukaryotic CLKs is generally a histidine or a glutamine. This difference is reflected in the anchoring of the β-hairpin insert to the C-lobe of the kinase domain (Figure 7). In PfLAMMER, the conserved acidic

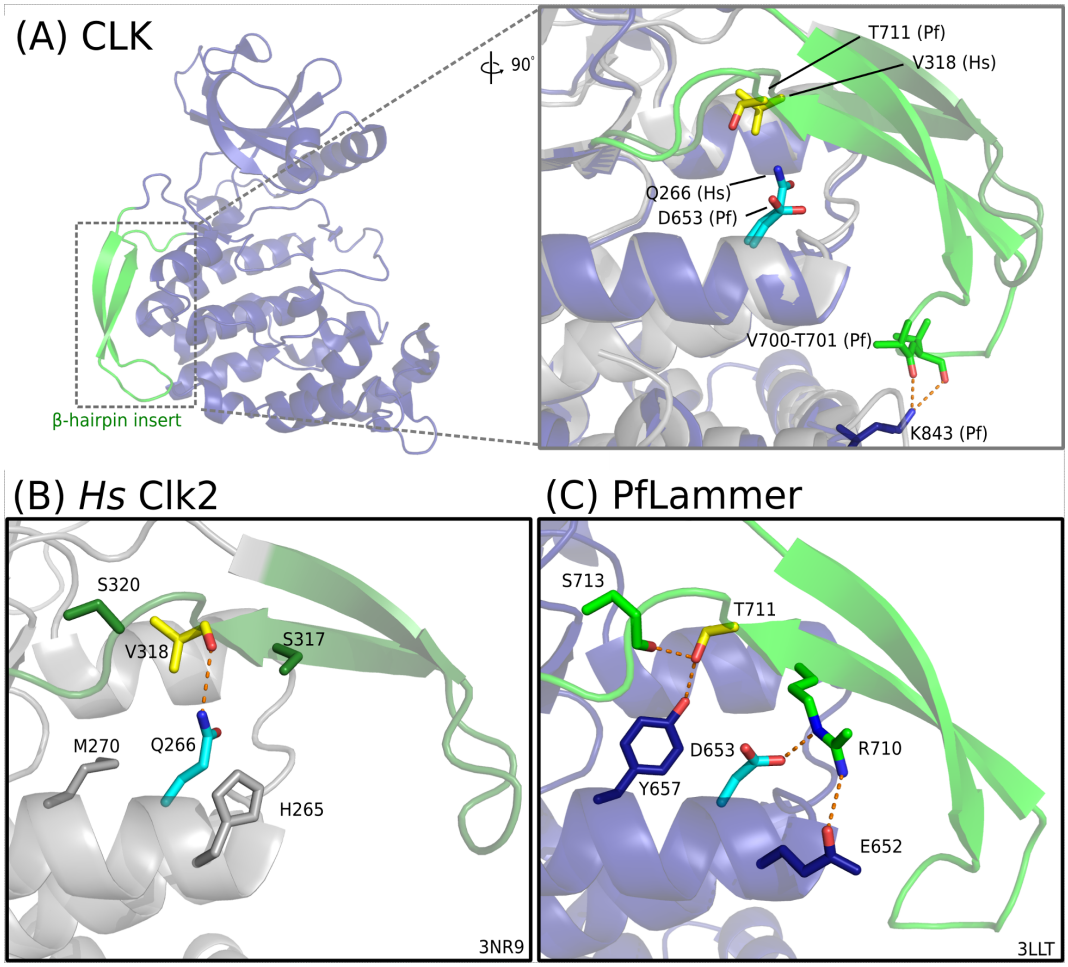


Figure 3.7: Comparison of the residue interactions anchoring the β-hairpin insert to the kinase C-lobe in solved structures of PfLAMMER [PDB:3LLT] and human Clk2 [PDB:3NR9]. **(A)** Both structures superimposed, with corresponding key residues shown in "sticks" representation. The contrastingly conserved residue (from CHAIN analysis) is highlighted cyan: D653 in PfLAMMER, Q266 in human Clk2. A residue of interest near the base of the hairpin insert, discussed in the text, is shown in yellow; its type is not strongly conserved within apicomplexan CLKs. Two residues in the loop of the hairpin, colored green, are inserts in PfLAMMER relative to Clk2; they appear anchored to the kinase C-lobe by interactions with a lysine, dark blue.
(B) Human Clk2, showing side chains near the residues of interest. A hydrogen bond appears between the αE-helix glutamine (cyan) and the backbone of a valine (yellow) near the base of the hairpin insert.
(C) In PfLAMMER, the two residues of interest, D653 (cyan) and T711 (yellow), do not interact directly; each instead forms several novel hydrogen bonds with other nearby residues, shown in green and blue, corresponding to those shown in green and gray in the human Clk2 structure.

residue is D653; the equivalent residue in human Clk2 [PDB:3NR9] is Q266. In Clk2, the MAPK substrate-recognition groove is occupied by a hydrophobic V318; Q266 stabilizes the backbone of V318 in human Clk2 (Figure 3.7B). In contrast, the distinctive D653 in PfLAMMER participates in a network of hydrogen bonds involving an arginine in the β-hairpin insert; the V318 in Clk2 is replaced by T711, which itself forms hydrogen bonds with two other residues in the αE helix and at the base of the insert, rather than with D653 (Figure 3.7C). Together these changes appear to further stabilize the beta-hairpin insert in *P. falciparum* by forming additional interactions. The changes also make the pocket more hydrophilic relative to Clk2.

The β-hairpin insert is several residues longer in chromalveolate CLKs than in human Clk2. In the PfLAMMER structure [PDB:3LLT], the hairpin loop is also anchored to the kinase C-lobe by a hydrogen bond between a lysine (K843) in the C-lobe and the backbone of the hairpin loop — this lysine, and consequently the hydrogen bond, is not seen in human Clk2 (Figure 3.7A). However, it is also possible that the interaction occurs in the solved structure as a consequence of crystal packing, in which case there may be no functional significance *in vivo*.

These variations, along with the variations in the primary substrate-binding site, indicate that apicomplexan and other chromalveolate CLKs have diverged from their human counterparts and specifically recognize and phosphorylate selected protein substrates.

3.3 Conclusions

We have used an approach based on evolutionary analysis to identify statistically distinct subfamilies of CDK and CDPK in the Apicomplexa and Chromalveolata, and explore the structural adaptations of CLK for substrate binding among chromalveolates. We discussed the functional implications of these distinguishing variations, confirmed and clarified previously published results regarding protein kinases in apicomplexan species, and proposed a set of new testable functional hypotheses, which we hope will focus future experimental efforts.

This methodology has provided a means for identifying clade-specific sequence and

structural features which may be associated with functional specialization. We presented three well-supported lineage-specific groups of kinases that emerged from our analysis, supported by existing structural and functional data about related proteins, and inferred additional functional hypotheses and the mechanisms that might enable these functions. Two of these sub-groups are members of the CMGC kinase group, which is highly conserved across Eukaryota, allowing strong homologies to be drawn between extant species to reveal ancient divergences along evolutionary branches. The third family, CDPK, is largely specific to plastid-containing eukaryotes in the Chromalveolata and Viridiplantae (but also found in other protozoans), but is also relatively more highly duplicated in each genome; the additional gene copies enhanced the statistical support for a proposed subfamily. The public availability of whole-genome sequences from diverse apicomplexan species likewise enabled the detection of deeply conserved sequence patterns. The work of the Structural Genomics Consortium [44] has also been invaluable in providing structural evidence for this neglected branch of protozoa.

Not every eukaryotic protein kinase family in apicomplexans yielded a distinctive feature set, however. Many of the "Other" kinase families are difficult to classify precisely; some are lineage-specific, and some have a mix of sequence features shared by multiple kinase families — the PfPK7 family, in fact, presents both problems [34]. The previously identified apicomplexan-specific families, FIKK and ROPK, are not strong candidates for CHAIN analysis, either: Since all of the species containing these families belong to the same phylum, shared sequence features within a sub-clade are likely to be the result of recent common ancestry rather than functional constraints on their molecular evolution. Despite these limitations, the approach we have presented will be useful for further analysis of apicomplexans as additional whole-genome sequences and protein kinase structures become available.

In the search for potential therapeutic targets for parasitic diseases, identification of these features and the molecular mechanisms they represent could lead to potential candidates for selective targeting. The taxonomic distribution of these novel protein features also provides insight into the evolution of apicomplexans and chromalveolates, lending support to the current understanding of these species' history.

Table 3.2: Genome data sources.

Genomes	Source
Plasmodium berghei ANKA, *P. chabaudi* AS, *P. falciparum* 3D7, *P. knowlesi* H, *P. vivax* Salvador I, *P. yoelii* 17XNL	PlasmoDB v.8.0 [8, 21, 22, 42, 46, 76]
Babesia bovis T2Bo; *Theileria annulata* Ankara, *T. parva* Mugaga	PiroplasmaDB v.1.1 [18, 43, 75]
Neospora caninum; *Toxoplasma gondii* GT1, ME49, VEG; *Eimeria tenella* Houghton	ToxoDB v.7.0 [41]
Cryptosporidium hominis, *C. muris*, *C. parvum* Iowa II	CryptoDB v.4.5 [1, 50, 97]
Perkinsus marinus ATCC 50983	NCBI genome project 12737
Thalassiosira pseudonana CCMP1335	NCBI genome project 34119 [6, 16]
Saccharomyces cerevisiae	Kinbase (http://kinase.com/kinbase/), Saccharomyces Genome Database (http://yeastgenome.org/)

3.4 Methods

3.4.1 Genome data sources

The protein complements of 17 complete genomes, from 15 distinct apicomplexan species, were retrieved from EupathDB [9]. The genomes of three non-apicomplexan species were also obtained for comparison (Table 3.2).

To obtain a sequence set of all solved apicomplexan ePK structures, the August 2011 release of PDBAA, the protein sequence database derived from PDB, was downloaded from NCBI. Phylum labels were added to the sequence headers according to GI number using the NCBI taxonomy data set, and sequences from the phylum Apicomplexa were selected.

3.4.2 Identification, classification and alignment of eukaryotic protein kinases (ePKs) in selected genomes

We constructed a curated set of ePK family profiles using previously annotated sequences from diverse model organisms. The classification scheme is based on the kinase groups and

families described in previous kinomic analyses [48, 54, 62]. Additional profiles for the FIKK, ROPK and PfPK7 families were built from apicomplexan sequences with annotations supported by experimental studies in published literature [77, 93].

We used the MAPGAPS program with the curated profile sets to identify, classify and align the protein kinases in the genomic sequences, as well as the apicomplexan ePK structures in PDB. MAPGAPS selects all sequences with a kinase domain containing key motifs, assigns each sequence with a significant hit to the best-matching family in the query profile, and accurately aligns each hit to the kinase consensus sequence, capturing conserved motifs [68]. Fragmentary sequences were then deleted.

Identification and classification of the ePKs in each genome revealed certain families present in multiple copies, providing enough data for further comparative analysis. The sequence counts in this scan generally agree with previously published kinome analyses, though because these and most previous annotations are produced by different computational methods there is occasional disagreement over the classification of more divergent sequences lacking clear orthologs in model organisms.

3.4.3 Gene tree inference to find divergent apicomplexan ortholog groups

Within each assigned ePK family, we concatenated the three sequence sets (apicomplexan genomic; a profile of sequences from model organisms including human; apicomplexan PDB sequences) and realigned the kinase domains using MAPGAPS to prepare a sequence alignment for phylogenetic analysis. To infer a gene tree from each of these alignments, we used RAxML with the fast bootstrap and maximum likelihood tree estimation procedure [88], PROTGAMMAWAG model (WAG amino acid substitution model with the rate heterogeneity), and 500 bootstrap replicates. We then used a custom script based on Biopython [24] to collapse branches with less than 50% bootstrap support in the resulting gene trees.

A resolved clade in the gene tree containing sequences from a monophyletic group of species, in agreement with the established species tree, indicates that the genes are orthologous. We selected clades that contained sequences from several apicomplexan species, but did not include any metazoan sequences, and with particular interest in clades

containing PDB structures, for further analysis.

3.4.4 Patterns of functional divergence

We queried related families of diverse sequences with selected clusters using the CHAIN program [67]. For each apicomplexan-specific cluster, we used the sequences from each gene clade of interest (described above) as the query set, and the sequences of diverse eukaryotic species in the corresponding kinase family as the main set, constructed from all kinase family members found in NCBI-nr. Both the query and main sequence sets were aligned with MAPGAPS for comparison.

The Bayesian Pattern Partitioning Search (BPPS) procedure in CHAIN simultaneously identifies selective constraints imposed on the foreground sequences, and pulls any sequences from the background that share the identified patterns in the query into the foreground, precisely defining a statistically supported family or subfamily if one exists [67].

Authors contributions

NK designed and conceived the project. ET performed the bioinformatics analyses. ET, AM and NK examined sequences and structural features and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Jessica C. Kissinger for valuable discussions related to this project and a critical review of a draft of this manuscript. We also thank the two anonymous reviewers for their detailed comments and advice on this manuscript.

We acknowledge the important contributions of the Structural Genomics Consortium in providing relevant crystal structures, many of which have been made public prior to the publication of an accompanying journal article. The Wellcome Trust Sanger Institute Pathogen Sequence Unit generously provided pre-publication access to genomic sequences.

Funding for NK from the University of Georgia is acknowledged.

Bibliography

[1] Abrahamsen, M. S., Templeton, T. J., Enomoto, S., Abrahante, J. E., Zhu, G., Lancto, C. A., Deng, M., Liu, C., Widmer, G., Tzipori, S., Buck, G. A., Xu, P., Bankier, A. T., Dear, P. H., Konfortov, B. A., Spriggs, H. F., Iyer, L., Anantharaman, V., Aravind, L., and Kapur, V. (2004). Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science*, **304**(5669), 441–445.

[2] Adl, S. M., Simpson, A. G. B., Farmer, M. A., Andersen, R. A., Anderson, O. R., Barta, J. R., Bowser, S. S., Brugerolle, G., Fensome, R. A., Fredericq, S., James, T. Y., Karpov, S., Kugrens, P., Krug, J., Lane, C. E., Lewis, L. A., Lodge, J., Lynn, D. H., Mann, D. G., McCourt, R. M., Mendoza, L., Moestrup, O., Mozley-Standridge, S. E., Nerad, T. A., Shearer, C. A., Smirnov, A. V., Spiegel, F. W., and Taylor, M. F. J. R. (2005). The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *The Journal of Eukaryotic Microbiology*, **52**(5), 399–451.

[3] Agarwal, S., Kern, S., Halbert, J., Przyborski, J. M., Baumeister, S., Dandekar, T., Doerig, C., and Pradel, G. (2011). Two nucleus-localized CDK-like kinases with crucial roles for malaria parasite erythrocytic replication are involved in phosphorylation of splicing factor. *Journal of Cellular Biochemistry*, **112**(5), 1295–1310.

[4] Ahn, Y. M., Clare, M., Ensinger, C. L., Hood, M. M., Lord, J. W., Lu, W.-P., Miller, D. F., Patt, W. C., Smith, B. D., Vogeti, L., Kaufman, M. D., Petillo, P. A., Wise, S. C., Abendroth, J., Chun, L., Clark, R., Feese, M., Kim, H., Stewart, L., and Flynn, D. L. (2010). Switch control pocket inhibitors of p38-MAP kinase. Durable type II inhibitors that do not require binding into the canonical ATP hinge region. *Bioorganic & Medicinal Chemistry Letters*, **20**(19), 5793–5798.

[5] Anamika, Srinivasan, N., and Krupa, A. (2005). A genomic perspective of protein kinases in *Plasmodium falciparum*. *Proteins*, **58**(1), 180–189.

[6] Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., Zhou, S., Allen, A. E., Apt, K. E., Bechner, M., Brzezinski, M. A., Chaal, B. K., Chiovitti, A., Davis, A. K., Demarest, M. S., Detter, J. C., Glavina, T., Goodstein, D., Hadi, M. Z., Hellsten, U., Hildebrand, M., Jenkins, B. D., Jurka, J., Kapitonov, V. V., Kröger, N., Lau, W. W. Y., Lane, T. W., Larimer, F. W., Lippmeier, J. C., Lucas, S., Medina, M., Montsant, A., Obornik, M., Parker, M. S., Palenik, B., Pazour, G. J., Richardson, P. M., Rynearson, T. A., Saito, M. A., Schwartz, D. C., Thamatrakoln, K., Valentin, K., Vardi, A., Wilkerson, F. P., and Rokhsar, D. S. (2004). The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science*, **306**(5693), 79–86.

[7] Artz, J. D., Wernimont, A. K., Allali-Hassani, A., Zhao, Y., Amani, M., Lin, Y.-H., Senisterra, G., Wasney, G. A., Fedorov, O., King, O., Roos, A., Lunin, V. V., Qiu, W., Finerty, P., Hutchinson, A., Chau, I., von Delft, F., Mackenzie, F., Lew, J., Kozieradzki, I., Vedadi, M., Schapira, M., Zhang, C., Shokat, K., Heightman, T., and Hui, R. (2011). The *Cryptosporidium parvum* Kinome. *BMC Genomics*, **12**(1), 478.

[8] Aurrecoechea, C., Brestelli, J., Brunk, B. P., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O. S., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J. C., Kraemer, E., Li, W., Miller, J. A., Nayak, V., Pennington, C., Pinney, D. F., Roos, D. S., Ross, C., Stoeckert, C. J., Treatman, C., and Wang, H. (2009). PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Research*, **37**(Database issue), D539–D543.

[9] Aurrecoechea, C., Brestelli, J., Brunk, B. P., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O. S., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J. C., Kraemer, E. T., Li, W., Miller, J. A., Nayak, V., Pennington, C., Pinney, D. F., Roos, D. S., Ross, C., Srinivasamoorthy, G., Stoeckert, C. J., Thibodeau, R., Treatman, C., and Wang, H. (2010). EuPathDB: a portal to eukaryotic pathogen databases. *Nucleic Acids Research*, **38**(Database issue), D415–D419.

[10] Bellon, S., Fitzgibbon, M., Fox, T., and Hsiao, H. (1999). The structure of phosphorylated P38 is monomeric and reveals a conserved activation-loop conformation. *Structure*, pages 1057–1065.

[11] Bemm, F., Schwarz, R., Förster, F., and Schultz, J. (2009). A kinome of 2600 in the ciliate Paramecium tetraurelia. *FEBS Letters*, **583**(22), 3589–3592.

[12] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, **28**(1), 235–242.

[13] Billker, O., Lourido, S., and Sibley, L. D. (2009). Calcium-dependent signaling and kinases in apicomplexan parasites. *Cell Host & Microbe*, **5**(6), 612–622.

[14] Bontell, I. L., Hall, N., Ashelford, K. E., Dubey, J. P., Boyle, J. P., Lindh, J., and Smith, J. E. (2009). Whole genome sequencing of a natural recombinant *Toxoplasma gondii* strain reveals chromosome sorting and local allelic variants. *Genome Biology*, **10**(5), R53.

[15] Boothroyd, J. C. and Dubremetz, J.-F. (2008). Kiss and spit: the dual roles of *Toxoplasma* rhoptries. *Nature Reviews. Microbiology*, **6**(1), 79–88.

[16] Bowler, C., Allen, A. E., Badger, J. H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R. P., Rayko, E., Salamov, A., Vandepoele, K., Beszteri, B., Gruber, A., Heijde, M., Katinka, M., Mock, T., Valentin, K., Verret, F., Berges, J. A., Brownlee, C., Cadoret, J.-P., Chiovitti, A., Choi, C. J., Coesel, S., De Martino, A., Detter, J. C., Durkin, C., Falciatore, A., Fournet, J., Haruta, M., Huysman, M. J. J., Jenkins, B. D., Jiroutova, K., Jorgensen, R. E., Joubert, Y., Kaplan, A., Kröger, N., Kroth, P. G., La Roche, J., Lindquist, E., Lommer, M., Martin-Jézéquel, V., Lopez, P. J., Lucas, S., Mangogna, M., McGinnis, K., Medlin, L. K., Montsant, A., Oudot-Le Secq, M.-P., Napoli, C., Obornik, M., Parker, M. S., Petit, J.-L., Porcel, B. M., Poulsen, N., Robison, M., Rychlewski, L., Rynearson, T. A., Schmutz, J., Shapiro, H., Siaut, M., Stanley, M., Sussman, M. R., Taylor, A. R., Vardi, A., von Dassow, P., Vyverman, W., Willis, A., Wyrwicz, L. S., Rokhsar, D. S., Weissenbach, J., Armbrust, E. V., Green, B. R., Van de Peer, Y., and Grigoriev, I. V. (2008). The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, **456**(7219), 239–244.

[17] Bradley, P. J., Ward, C., Cheng, S. J., Alexander, D. L., Coller, S., Coombs, G. H., Dunn, J. D., Ferguson, D. J., Sanderson, S. J., Wastling, J. M., and Boothroyd, J. C. (2005). Proteomic analysis of rhoptry organelles reveals many novel constituents for host-parasite interactions in *Toxoplasma gondii*. *The Journal of Biological Chemistry*, **280**(40), 34245–34258.

[18] Brayton, K. A., Lau, A. O. T., Herndon, D. R., Hannick, L., Kappmeyer, L. S., Berens, S. J., Bidwell, S. L., Brown, W. C., Crabtree, J., Fadrosh, D., Feldblum, T., Forberger, H. A., Haas, B. J., Howell, J. M., Khouri, H., Koo, H., Mann, D. J., Norimine, J., Paulsen, I. T., Radune, D., Ren, Q., Smith, R. K., Suarez, C. E., White, O., Wortman, J. R., Knowles, D. P., McElwain, T. F., and Nene, V. M. (2007). Genome sequence of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa. *PLoS Pathogens*, **3**(10), 1401–1413.

[19] Brown, N. R., Noble, M. E., Endicott, J. A., and Johnson, L. N. (1999). The structural basis for specificity of substrate and recruitment peptides for cyclin-dependent kinases. *Nature Cell Biology*, **1**(7), 438–443.

[20] Bullock, A. N., Das, S., Debreczeni, J. E., Rellos, P., Fedorov, O., Niesen, F. H., Guo, K., Papagrigoriou, E., Amos, A. L., Cho, S., Turk, B. E., Ghosh, G., and Knapp, S. (2009). Kinase domain insertions define distinct roles of CLK kinases in SR protein phosphorylation. *Structure*, **17**(3), 352–362.

[21] Carlton, J. M., Angiuoli, S. V., Suh, B. B., Kooij, T. W., Pertea, M., Silva, J. C., Ermolaeva, M. D., Allen, J. E., Selengut, J. D., Koo, H. L., Peterson, J. D., Pop, M., Kosack, D. S., Shumway, M. F., Bidwell, S. L., Shallom, S. J., van Aken, S. E., Riedmuller, S. B., Feldblyum, T. V., Cho, J. K., Quackenbush, J., Sedegah, M., Shoaibi, A., Cummings, L. M., Florens, L., Yates, J. R., Raine, J. D., Sinden, R. E., Harris, M. A., Cunningham, D. A., Preiser, P. R., Bergman, L. W., Vaidya, A. B., van Lin, L. H., Janse, C. J., Waters, A. P., Smith, H. O., White, O. R., Salzberg, S. L., Venter, J. C., Fraser, C. M., Hoffman, S. L., Gardner, M. J., and Carucci, D. J. (2002). Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature*, **419**(6906), 512–519.

[22] Carlton, J. M., Adams, J. H., Silva, J. C., Bidwell, S. L., Lorenzi, H., Caler, E., Crabtree, J., Angiuoli, S. V., Merino, E. F., Amedeo, P., Cheng, Q., Coulson, R. M. R., Crabb, B. S., Del Portillo, H. A., Essien, K., Feldblyum, T. V., Fernandez-Becerra, C., Gilson, P. R., Gueye, A. H., Guo, X., Kang'a, S., Kooij, T. W. A., Korsinczky, M., Meyer, E. V.-S., Nene, V., Paulsen, I., White, O., Ralph, S. A., Ren, Q., Sargeant, T. J., Salzberg, S. L., Stoeckert, C. J., Sullivan, S. A., Yamamoto, M. M., Hoffman, S. L., Wortman, J. R., Gardner, M. J., Galinski, M. R., Barnwell, J. W., and Fraser-Liggett, C. M. (2008). Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature*, **455**(7214), 757–763.

[23] Chen, F., Mackey, A. J., Stoeckert, C. J., and Roos, D. S. (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research*, **34**(Database issue), D363–D368.

[24] Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**(11), 1422–1423.

[25] Colwill, K., Feng, L. L., Yeakley, J. M., Gish, G. D., Cáceres, J. F., Pawson, T., and Fu, X. D. (1996). SRPK1 and Clk/Sty protein kinases show distinct substrate specificities for serine/arginine-rich splicing factors. *The Journal of Biological Chemistry*, **271**(40), 24569–24575.

[26] Crooks, G. E., Hon, G., Chandonia, J.-m., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Research*, **14**(6), 1188–1190.

[27] Delano, W. (2011). The PyMOL Molecular Graphics System.

[28] Dixit, A., Singh, P. K., Sharma, G. P., Malhotra, P., and Sharma, P. (2010). PfSRPK1, a novel splicing-related kinase from *Plasmodium falciparum*. *The Journal of Biological Chemistry*, **285**(49), 38315–38323.

[29] Doerig, C. (2005). Protein kinases regulating *Plasmodium* proliferation and development. In I. W. Sherman, editor, *Molecular Approaches to Malaria*, chapter 15, pages 290–310. ASM Press, Washington, D.C.

[30] Doerig, C., Endicott, J., and Chakrabarti, D. (2002). Cyclin-dependent kinase homologues of *Plasmodium falciparum*. *International Journal for Parasitology*, **32**(13), 1575–1585.

[31] Doerig, C., Billker, O., Pratt, D., and Endicott, J. (2005). Protein kinases as targets for antimalarial intervention: Kinomics, structure-based design, transmission-blockade, and targeting host cell enzymes. *Biochimica et Biophysica Acta*, **1754**(1-2), 132–150.

[32] Doerig, C., Abdi, A., Bland, N., Eschenlauer, S., Dorin-Semblat, D., Fennell, C., Halbert, J., Holland, Z., Nivez, M.-P., Semblat, J.-P., Sicard, A., and Reininger, L. (2010). Malaria: targeting parasite and host cell kinomes. *Biochimica et Biophysica Acta*, **1804**(3), 604–612.

[33] Dorin, D., Le Roch, K., Sallicandro, P., Alano, P., Parzy, D., Poullet, P., Meijer, L., and Doerig, C. (2001). Pfnek-1, a NIMA-related kinase from the human malaria parasite *Plasmodium falciparum*. *European Journal of Biochemistry*, **268**(9), 2600–2608.

[34] Dorin, D., Semblat, J.-P., Poullet, P., Alano, P., Goldring, J. P. D., Whittle, C., Patterson, S., Chakrabarti, D., and Doerig, C. (2005). PfPK7, an atypical MEK-related protein kinase, reflects the absence of classical three-component MAPK pathways in the human malaria parasite *Plasmodium falciparum*. *Molecular Microbiology*, **55**(1), 184–196.

[35] Dorin-Semblat, D., Sicard, A., Doerig, C., Ranford-Cartwright, L., and Doerig, C. (2008). Disruption of the PfPK7 gene impairs schizogony and sporogony in the human malaria parasite *Plasmodium falciparum*. *Eukaryotic Cell*, **7**(2), 279–285.

[36] Dvorin, J. D., Martyn, D. C., Patel, S. D., Grimley, J. S., Collins, C. R., Hopp, C. S., Bright, a. T., Westenberger, S., Winzeler, E., Blackman, M. J., Baker, D. A., Wandless, T. J., and Duraisingh, M. T. (2010). A plant-like kinase in *Plasmodium falciparum* regulates parasite egress from erythrocytes. *Science*, **328**(5980), 910–912.

[37] Eisen, J. A., Coyne, R. S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J. R., Badger, J. H., Ren, Q., Amedeo, P., Jones, K. M., Tallon, L. J., Delcher, A. L., Salzberg, S. L., Silva, J. C., Haas, B. J., Majoros, W. H., Farzad, M., Carlton, J. M., Smith, R. K., Garg, J., Pearlman, R. E., Karrer, K. M., Sun, L., Manning, G., Elde, N. C., Turkewitz, A. P., Asai, D. J., Wilkes, D. E., Wang, Y., Cai, H., Collins, K., Stewart, B. A., Lee, S. R., Wilamowska, K., Weinberg, Z., Ruzzo, W. L., Wloga, D., Gaertig, J., Frankel, J., Tsao, C.-C., Gorovsky, M. A., Keeling, P. J., Waller, R. F., Patron, N. J., Cherry, J. M., Stover, N. A., Krieger, C. J., del Toro, C., Ryder, H. F., Williamson, S. C., Barbeau, R. A., Hamilton, E. P., and Orias, E. (2006). Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. *PLoS Biology*, **4**(9), e286.

[38] El Hajj, H., Lebrun, M., Fourmaux, M. N., Vial, H., and Dubremetz, J. F. (2007). Inverted topology of the *Toxoplasma gondii* ROP5 rhoptry protein provides new insights into the association of the ROP2 protein family with the parasitophorous vacuole membrane. *Cellular Microbiology*, **9**(1), 54–64.

[39] Emsley, P. and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta rcystallographica. Section D, Biological Crystallography*, **60**, 2126–2132.

[40] Fluhr, R. (2008). Regulation of Splicing by Protein Phosphorylation. In A. S. N. Reddy and M. Golovkin, editors, *Nuclear pre-mRNA Processing in Plants*, volume 326 of *Current Topics in Microbiology and Immunology*, chapter 7, pages 119–138. Springer Berlin Heidelberg, Berlin, Heidelberg.

[41] Gajria, B., Bahl, A., Brestelli, J., Dommer, J., Fischer, S., Gao, X., Heiges, M., Iodice, J., Kissinger, J. C., Mackey, A. J., Pinney, D. F., Roos, D. S., Stoeckert, C. J., Wang, H., and Brunk, B. P. (2008). ToxoDB: an integrated *Toxoplasma gondii* database resource. *Nucleic Acids Research*, **36**(Database issue), D553–D556.

[42] Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A., Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M.-S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather,

M. W., Vaidya, A. B., Martin, D. M. A., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S. A., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Newbold, C., Davis, R. W., Fraser, C. M., and Barrell, B. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, **419**(6906), 498–511.

[43] Gardner, M. J., Bishop, R., Shah, T., de Villiers, E. P., Carlton, J. M., Hall, N., Ren, Q., Paulsen, I. T., Pain, A., Berriman, M., Wilson, R. J. M., Sato, S., Ralph, S. A., Mann, D. J., Xiong, Z., Shallom, S. J., Weidman, J., Jiang, L., Lynn, J., Weaver, B., Shoaibi, A., Domingo, A. R., Wasawo, D., Crabtree, J., Wortman, J. R., Haas, B., Angiuoli, S. V., Creasy, T. H., Lu, C., Suh, B., Silva, J. C., Utterback, T. R., Feldblyum, T. V., Pertea, M., Allen, J., Nierman, W. C., Taracha, E. L. N., Salzberg, S. L., White, O. R., Fitzhugh, H. A., Morzaria, S., Venter, J. C., Fraser, C. M., and Nene, V. (2005). Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. *Science*, **309**(5731), 134–137.

[44] Gileadi, O., Knapp, S., Lee, W. H., Marsden, B. D., Müller, S., Niesen, F. H., Kavanagh, K. L., Ball, L. J., von Delft, F., Doyle, D. A., Oppermann, U. C. T., and Sundström, M. (2007). The scientific impact of the Structural Genomics Consortium: a protein family and ligand-centered approach to medically-relevant human proteins. *Journal of Structural and Functional Genomics*, **8**(2-3), 107–119.

[45] Golovkin, M. and Reddy, a. S. (1999). An SC35-like protein and a novel serine/arginine-rich protein interact with Arabidopsis U1-70K protein. *The Journal of Biological Chemistry*, **274**(51), 36428–36438.

[46] Hall, N., Karras, M., Raine, J. D., Carlton, J. M., Kooij, T. W. A., Berriman, M., Florens, L., Janssen, C. S., Pain, A., Christophides, G. K., James, K., Rutherford, K., Harris, B., Harris, D., Churcher, C., Quail, M. A., Ormond, D., Doggett, J., Trueman, H. E., Mendoza, J., Bidwell, S. L., Rajandream, M.-A., Carucci, D. J., Yates, J. R., Kafatos, F. C., Janse, C. J., Barrell, B., Turner, C. M. R., Waters, A. P., and Sinden, R. E. (2005). A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. *Science*, **307**(5706), 82–86.

[47] Hammarton, T. C., Mottram, J. C., and Doerig, C. (2003). The cell cycle of parasitic protozoa: potential for chemotherapeutic exploitation. *Progress in Cell Cycle Research*, **5**, 91–101.

[48] Hanks, S. K. and Hunter, T. (1995). Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB Journal*, **9**(8), 576–596.

[49] Harper, J. F. and Harmon, A. (2005). Plants, symbiosis and parasites: a calcium signalling connection. *Nature Reviews. Molecular Cell Biology*, **6**(7), 555–566.

[50] Heiges, M., Wang, H., Robinson, E., Aurrecoechea, C., Gao, X., Kaluskar, N., Rhodes, P., Wang, S., He, C.-Z., Su, Y., Miller, J., Kraemer, E., and Kissinger, J. C. (2006). CryptoDB: a *Cryptosporidium* bioinformatics resource update. *Nucleic Acids Research*, **34**(Database issue), D419–D422.

[51] Iriko, H., Jin, L., Kaneko, O., Takeo, S., Han, E.-T., Tachibana, M., Otsuki, H., Torii, M., and Tsuboi, T. (2009). A small-scale systematic analysis of alternative splicing in *Plasmodium falciparum*. *Parasitology International*, **58**(2), 196–199.

[52] Joseph, S. J., Fernández-Robledo, J. A., Gardner, M. J., El-Sayed, N. M., Kuo, C.-H., Schott, E. J., Wang, H., Kissinger, J. C., and Vasta, G. R. (2010). The Alveolate Perkinsus marinus: biological insights from EST gene discovery. *BMC Genomics*, **11**, 228.

[53] Kannan, N. and Neuwald, A. F. (2004). Evolutionary constraints associated with functional specificity of the CMGC protein kinases MAPK, CDK, GSK, SRPK, DYRK, and CK2alpha. *Protein Science*, **13**(8), 2059–2077.

[54] Kannan, N., Taylor, S. S., Zhai, Y., Venter, J. C., and Manning, G. (2007). Structural and functional diversity of the microbial kinome. *PLoS Biology*, **5**(3), e17.

[55] Keeling, P. J., Burger, G., Durnford, D. G., Lang, B. F., Lee, R. W., Pearlman, R. E., Roger, A. J., and Gray, M. W. (2005). The tree of eukaryotes. *Trends in Ecology & Evolution*, **20**(12), 670–676.

[56] Kojima, T., Zama, T., Wada, K., Onogi, H., and Hagiwara, M. (2001). Cloning of human PRP4 reveals interaction with Clk1. *The Journal of Biological Chemistry*, **276**(34), 32247–32256.

[57] Kuo, C.-H. and Kissinger, J. C. (2008). Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites *Plasmodium* and *Theileria*. *BMC Evolutionary Biology*, **8**(1), 108.

[58] Kuo, C.-H., Wares, J. P., and Kissinger, J. C. (2008). The Apicomplexan whole-genome phylogeny: an analysis of incongruence among gene trees. *Molecular Biology and Evolution*, **25**(12), 2689–2698.

[59] Lawrence, J. G. (2005). Common themes in the genome strategies of pathogens. *Current Opinion in Genetics & Development*, **15**(6), 584–588.

[60] Li, J. L., Targett, G. A., and Baker, D. A. (2001). Primary structure and sexual stage-specific expression of a LAMMER protein kinase of *Plasmodium falciparum*. *International Journal for Parasitology*, **31**(4), 387–392.

[61] Lourido, S., Shuman, J., Zhang, C., Shokat, K. M., Hui, R., and Sibley, L. D. (2010). Calcium-dependent protein kinase 1 is an essential regulator of exocytosis in *Toxoplasma*. *Nature*, **465**(7296), 359–362.

[62] Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. *Science*, **298**(5600), 1912–1934.

[63] Martin, D. M. A., Miranda-Saavedra, D., and Barton, G. J. (2009). Kinomer v. 1.0: a database of systematically classified eukaryotic protein kinases. *Nucleic Acids Research*, **37**(Database issue), D244–D250.

[64] Miranda-Saavedra, D., Stark, M. J. R., Packer, J. C., Vivares, C. P., Doerig, C., and Barton, G. J. (2007). The complement of protein kinases of the microsporidium *Encephalitozoon cuniculi* in relation to those of *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. *BMC Genomics*, **8**, 309.

[65] Morgan, D. O. (1997). Cyclin-dependent kinases: engines, clocks, and microprocessors. *Annual Review of Cell and Developmental Biology*, **13**, 261–291.

[66] Nagamune, K. and Sibley, L. D. (2006). Comparative genomic and phylogenetic analyses of calcium ATPases and calcium-regulated proteins in the apicomplexa. *Molecular Biology and Evolution*, **23**(8), 1613–1627.

[67] Neuwald, A. F. (2007). The CHAIN program: forging evolutionary links to underlying mechanisms. *Trends in Biochemical Sciences*, **32**(11), 487–493.

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

[69] Ngo, J. C. K., Chakrabarti, S., Ding, J.-H., Velazquez-Dones, A., Nolen, B., Aubol, B. E., Adams, J. A., Fu, X.-D., and Ghosh, G. (2005). Interplay between SRPK and Clk/Sty kinases in phosphorylation of the splicing factor ASF/SF2 is regulated by a docking motif in ASF/SF2. *Molecular cell*, **20**(1), 77–89.

[70] Nikolakaki, E., Du, C., Lai, J., Giannakouros, T., Cantley, L., and Rabinow, L. (2002). Phosphorylation by LAMMER protein kinases: determination of a consensus site, identification of in vitro substrates, and implications for substrate preferences. *Biochemistry*, **41**(6), 2055–2066.

[71] Nishida, E. and Gotoh, Y. (1993). The MAP kinase cascade is essential for diverse signal transduction pathways. *Trends in Biochemical Sciences*, **18**(4), 128–131.

[72] Nunes, M. C., Goldring, J. P. D., Doerig, C., and Scherf, A. (2007). A novel protein kinase family in *Plasmodium falciparum* is differentially transcribed and secreted to various cellular compartments of the host cell. *Molecular Microbiology*, **63**(2), 391–403.

[73] Nunes, M. C., Okada, M., Scheidig-Benatar, C., Cooke, B. M., and Scherf, A. (2010). *Plasmodium falciparum* FIKK kinase members target distinct components of the erythrocyte membrane. *PloS One*, **5**(7), e11747.

[74] Ojo, K. K., Larson, E. T., Keyloun, K. R., Castaneda, L. J., Derocher, A. E., Inampudi, K. K., Kim, J. E., Arakaki, T. L., Murphy, R. C., Zhang, L., Napuli, A. J., Maly, D. J.,

Verlinde, C. L. M. J., Buckner, F. S., Parsons, M., Hol, W. G. J., Merritt, E. A., and Van Voorhis, W. C. (2010). *Toxoplasma gondii* calcium-dependent protein kinase 1 is a target for selective kinase inhibitors. *Nature Structural & Molecular Biology*, **17**(5), 602–607.

[75] Pain, A., Renauld, H., Berriman, M., Murphy, L., Yeats, C. A., Weir, W., Kerhornou, A., Aslett, M., Bishop, R., Bouchier, C., Cochet, M., Coulson, R. M. R., Cronin, A., de Villiers, E. P., Fraser, A., Fosker, N., Gardner, M., Goble, A., Griffiths-Jones, S., Harris, D. E., Katzer, F., Larke, N., Lord, A., Maser, P., McKellar, S., Mooney, P., Morton, F., Nene, V., O'Neil, S., Price, C., Quail, M. A., Rabbinowitsch, E., Rawlings, N. D., Rutter, S., Saunders, D., Seeger, K., Shah, T., Squares, R., Squares, S., Tivey, A., Walker, A. R., Woodward, J., Dobbelaere, D. a. E., Langsley, G., Rajandream, M.-A., McKeever, D., Shiels, B., Tait, A., Barrell, B., and Hall, N. (2005). Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*. *Science*, **309**(5731), 131–133.

[76] Pain, A., Böhme, U., Berry, a. E., Mungall, K., Finn, R. D., Jackson, a. P., Mourier, T., Mistry, J., Pasini, E. M., Aslett, M. A., Balasubrammaniam, S., Borgwardt, K., Brooks, K., Carret, C., Carver, T. J., Cherevach, I., Chillingworth, T., Clark, T. G., Galinski, M. R., Hall, N., Harper, D., Harris, D., Hauser, H., Ivens, A., Janssen, C. S., Keane, T., Larke, N., Lapp, S., Marti, M., Moule, S., Meyer, I. M., Ormond, D., Peters, N., Sanders, M., Sanders, S., Sargeant, T. J., Simmonds, M., Smith, F., Squares, R., Thurston, S., Tivey, a. R., Walker, D., White, B., Zuiderwijk, E., Churcher, C., Quail, M. A., Cowman, a. F., Turner, C. M. R., Rajandream, M. A., Kocken, C. H. M., Thomas, a. W., Newbold, C. I., Barrell, B. G., and Berriman, M. (2008). The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature*, **455**(7214), 799–803.

[77] Peixoto, L., Chen, F., Harb, O. S., Davis, P. H., Beiting, D. P., Brownback, C. S., Ouloguem, D., and Roos, D. S. (2010). Integrative genomic approaches highlight a family of parasite-specific kinases that regulate host responses. *Cell Host & Microbe*, **8**(2), 208–218.

[78] Pick, C., Ebersberger, I., Spielmann, T., Bruchhaus, I., and Burmester, T. (2011). Phylogenomic analyses of malaria parasites and evolution of their exported proteins. *BMC Evolutionary Biology*, **11**(1), 167.

[79] Qiu, W., Wernimont, A. K., Tang, K., Taylor, S., Lunin, V., Schapira, M., Fentress, S., Hui, R., and Sibley, L. D. (2009). Novel structural and regulatory features of rhoptry secretory kinases in *Toxoplasma gondii*. *The EMBO Journal*, **28**(7), 969–979.

[80] Renslo, A. R. and McKerrow, J. H. (2006). Drug discovery and development for neglected parasitic diseases. *Nature Chemical Biology*, **2**(12), 701–710.

[81] Rodgers, J. T., Haas, W., Gygi, S. P., and Puigserver, P. (2010). Cdc2-like kinase 2 is an insulin-regulated suppressor of hepatic gluconeogenesis. *Cell Metabolism*, **11**(1), 23–34.

[82] Roos, D. S. (2005). Genetics. Themes and variations in apicomplexan parasite biology. *Science*, **309**(5731), 72–73.

[83] Sargeant, T. J., Marti, M., Caler, E., Carlton, J. M., Simpson, K., Speed, T. P., and Cowman, A. F. (2006). Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. *Genome Biology*, **7**(2), R12.

[84] Schneider, A. G. and Mercereau-Puijalon, O. (2005). A new Apicomplexa-specific protein kinase family: multiple members in *Plasmodium falciparum*, all with an export signature. *BMC Genomics*, **6**(1), 30.

[85] Shaw, M. K. (2003). Cell invasion by *Theileria sporozoites*. *Trends in Parasitology*, **19**(1), 2–6.

[86] Sibley, L. D. (2004). Intracellular parasite invasion strategies. *Science*, **304**(5668), 248–253.

[87] Sibley, L. D. and Ajioka, J. W. (2008). Population structure of *Toxoplasma gondii*: clonal expansion driven by infrequent recombination and selective sweeps. *Annual Review of Microbiology*, **62**, 329–351.

[88] Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**(21), 2688–2690.

[89] Striepen, B., Jordan, C. N., Reiff, S., and van Dooren, G. G. (2007). Building the perfect parasite: cell division in apicomplexa. *PLoS Pathogens*, **3**(6), e78.

[90] Templeton, T. J., Iyer, L. M., Anantharaman, V., Enomoto, S., Abrahante, J. E., Subramanian, G. M., Hoffman, S. L., Abrahamsen, M. S., and Aravind, L. (2004). Comparative analysis of apicomplexa and genomic diversity in eukaryotes. *Genome Research*, **14**(9), 1686–1695.

[91] Tewari, R., Straschil, U., Bateman, A., Böhme, U., Cherevach, I., Gong, P., Pain, A., and Billker, O. (2010). The systematic functional analysis of *Plasmodium* protein kinases identifies essential regulators of mosquito transmission. *Cell Host & Microbe*, **8**(4), 377–387.

[92] Velazquez-Dones, A., Hagopian, J. C., Ma, C.-T., Zhong, X.-Y., Zhou, H., Ghosh, G., Fu, X.-D., and Adams, J. A. (2005). Mass spectrometric and kinetic analysis of ASF/SF2 phosphorylation by SRPK1 and Clk/Sty. *The Journal of Biological Chemistry*, **280**(50), 41761–41768.

[93] Ward, P., Equinet, L., Packer, J., and Doerig, C. (2004). Protein kinases of the human malaria parasite *Plasmodium falciparum*: the kinome of a divergent eukaryote. *BMC Genomics*, **5**(1), 79.

[94] Wernimont, A. K., Artz, J. D., Finerty, P., Lin, Y.-H., Amani, M., Allali-Hassani, A., Senisterra, G., Vedadi, M., Tempel, W., Mackenzie, F., Chau, I., Lourido, S., Sibley, L. D., and Hui, R. (2010). Structures of apicomplexan calcium-dependent protein kinases reveal mechanism of activation by calcium. *Nature Structural & Molecular Biology*, **17**(5), 596–601.

[95] Wernimont, A. K., Amani, M., Qiu, W., Pizarro, J. C., Artz, J. D., Lin, Y.-H., Lew, J., Hutchinson, A., and Hui, R. (2011). Structures of parasitic CDPK domains point to a common mechanism of activation. *Proteins*, **79**(3), 803–820.

[96] Xiao, L., Sulaiman, I. M., Ryan, U. M., Zhou, L., Atwill, E. R., Tischler, M. L., Zhang, X., Fayer, R., and Lal, A. A. (2002). Host adaptation and host-parasite co-evolution in

Cryptosporidium: implications for taxonomy and public health. *International Journal for Parasitology*, **32**(14), 1773–1785.

[97] Xu, P., Widmer, G., Wang, Y., Ozaki, L. S., Alves, J. M., Serrano, M. G., Puiu, D., Manque, P., Akiyoshi, D., Mackey, A. J., Pearson, W. R., Dear, P. H., Bankier, A. T., Peterson, D. L., Abrahamsen, M. S., Kapur, V., Tzipori, S., and Buck, G. A. (2004). The genome of *Cryptosporidium hominis*. *Nature*, **431**(7012), 1107–1112.

[98] Yun, B., Farkas, R., Lee, K., and Rabinow, L. (1994). The Doa locus encodes a member of a new protein kinase family and is essential for eye and embryonic development in Drosophila melanogaster. *Genes & Development*, **8**(10), 1160–1173.

Chapter 4

Structural and evolutionary adaptation of rhoptry kinases and pseudokinases, a family of coccidian virulence factors

Eric Talevich and Natarajan Kannan (2013) *BMC Evolutionary Biology* 13:117.
Reprinted here with permission from the publisher.

Abstract

BACKGROUND: The widespread protozoan parasite *Toxoplasma gondii* interferes with host cell functions by exporting the contents of a unique apical organelle, the rhoptry. Among the mix of secreted proteins are an expanded, lineage-specific family of protein kinases termed rhoptry kinases (ROPKs), several of which have been shown to be key virulence factors, including the pseudokinase ROP5. The extent and details of the diversification of this protein family are poorly understood.

RESULTS: In this study, we comprehensively catalogued the ROPK family in the genomes of *Toxoplasma gondii*, *Neospora caninum* and *Eimeria tenella*, as well as portions of the unfinished genome of *Sarcocystis neurona*, and classified the identified genes into 42 distinct subfamilies. We systematically compared the rhoptry kinase protein sequences and structures to each other and to the broader superfamily of eukaryotic protein kinases to study the patterns of diversification and neofunctionalization in the ROPK family and its subfamilies. We identified three ROPK sub-clades of particular interest: those bearing a structurally conserved N-terminal extension to the kinase domain (NTE), an *E. tenella*-specific expansion, and a basal cluster including ROP35 and BPK1 that we term ROPKL. Structural analysis in light of the solved structures ROP2, ROP5, ROP8 and in comparison to typical eukaryotic protein kinases revealed ROPK-specific conservation patterns in two key regions of the kinase domain, surrounding a ROPK-conserved insert in the kinase hinge region and a disulfide bridge in the kinase substrate-binding lobe. We also examined conservation patterns specific to the NTE-bearing clade. We discuss the possible functional consequences of each.

CONCLUSIONS: Our work sheds light on several important but previously unrecognized features shared among rhoptry kinases, as well as the essential differences between active and degenerate protein kinases. We identify the most distinctive ROPK-specific features conserved across both active kinases and pseudokinases, and discuss these in terms of sequence motifs, evolutionary context, structural impact and potential functional relevance. By characterizing the proteins that enable these parasites to invade the host cell and co-opt its signaling mechanisms, we provide guidance on potential therapeutic targets for the diseases caused by coccidian parasites.

4.1 Introduction

Toxoplasma gondii is an intracellular parasite that infects a wide range of hosts, including an estimated one-third of the world's human population [44]. The resulting disease toxoplasmosis can be serious in pregnant women and immunocompromised individuals, and as an opportunistic infection associated with AIDS and cancer patients [33]. *T. gondii* and its evolutionary relatives, the Coccidia, form a clade of parasitic protozoa involved in many human and veterinary diseases such as toxoplasmosis and coccidiosis. Coccidians are a lineage within the protozoan phylum Apicomplexa, which also includes the deadly malaria pathogen *Plasmodium falciparum*. Thus, *T. gondii* also serves as an experimentally tractable model organism for studying the shared and contrasting biological properties of the Apicomplexa and other intracellular parasites [32, 72].

Apicomplexans contain a unique system of apical organelles called the apical complex, consisting of rhoptries, micronemes and dense granules [46]. At the initiation of host cell invasion, the contents of the rhoptries are injected into the host cell and the forming parasitophorous vacuole which protects the intracellular parasite [7]. Once there, the parasite proteins can disrupt host cell signaling and defense mechanisms and assist in recruiting host organelles [28].

Proteomic profiling of *T. gondii* rhoptries [10] and analyis of apicomplexan genomic sequences [21, 43, 56, 76] revealed that many of the proteins secreted by coccidians are protein kinases, a class of enzymes that regulate cell signal transduction through phosphorylation. This expanded, rapidly evolving family of kinases and pseudokinases has been termed the rhoptry kinase (ROPK) family [56], or ROP2 family, in reference to a representative member of the family [21]. While rhoptry kinases appear to be unique to the Coccidia, the involvement of lineage-specific protein kinase families in host-parasite interactions is observed across the Apicomplexa [40]. Several rhoptry kinases have been shown to be involved in virulence and alteration of host cell transcription [9, 28]. These include ROP18, a key modulator of parasite growth and virulence which is localized to the parasitophorous vacuole membrane (PVM) [66, 78], and ROP5, another PVM-associated protein which assists ROP18 in blocking the host immune response [3, 4, 23, 63, 75]. ROP16 localizes to the host cell nucleus and interacts with the STAT3 and STAT6 immune-

response signaling pathways [12, 53, 54, 67, 83], and ROP38 has been implicated in the modulation of host MAPK signaling [56].

Protein kinases are a diverse family of enzymes which have been successfully targeted for inhibition in human cancers, and show promise for treating infections by protozoan pathogens as well [17]. ATP-competitive small-molecule inhibitors have been developed to specifically target catalytically active protein kinases in parasitic protozoa [65]. Since many of the ROPKs appear to also be catalytically active, there may be an opportunity to target these kinases for infectious diseases. However, the "catalytic triad" of residues considered essential for kinase enzymatic activity [34] is altered in about half of the identified ROPKs [56]. Pseudokinases have been observed to perform important functions in other systems, typically through inducing allosteric changes in other interacting partners (e.g. [71, 85]; reviewed in [8, 35, 84]). The overall expansion of pseudokinases in the ROPK family underscores observations that some catalytically inactive ROPKs nonetheless play important, functional roles through interaction with other proteins [4, 23, 62]. Structural studies showed that the pseudokinase virulence factors ROP2, ROP8 and ROP5 do indeed form a protein kinase fold; ROP2 and ROP8 were indicated to be unable to bind ATP [37], while ROP5 was shown to bind ATP in an atypical, noncatalytic conformation [61]. An interplay between ROP5, the active kinase ROP18 and a host immunity-related GTPase has been identified [4, 23], demonstrating the potential for complex interplay between rhoptry kinases and the host cell signaling pathways. However, the full extent of the diversity in ROPK family, in terms of function, potential interacting partners, protein structure and structural mechanisms, is poorly understood. With the availability of molecular sequence and structural data from multiple strains of *T. gondii* and related apicomplexans, we can use comparative methods to examine the molecular evolution of ROPKs and identify functional shifts that may point to distinct regulatory roles and mechanisms.

We catalogued the rhoptry kinases in several fully sequenced coccidian genomes, including *Toxoplasma gondii*, *Neospora caninum*, *Sarcocystis neurona* and *Eimeria tenella*, and compared them to the broader eukaryotic protein kinase (ePK) superfamily and to each other to study the patterns of diversification and neofunctionalization in the ROPK family and its subfamilies. We propose previously unidentified rhoptry kinases in each of

these genomes, including several putative new ROPK subfamilies. We studied the variation in these subfamilies in light of the solved structures of ROP2, ROP8 and ROP5 proteins, and relative to "typical" eukaryotic protein kinases. Both pseudokinases and catalytically active kinases appear to be prevalent throughout the ROPK family. We found a striking co-evolution of structural inserts within the canonical protein kinase domain and the residues that interact with them. Most noteworthy among these is a pattern of residues surrounding the ROPK-specific αC' helix in the kinase "hinge" region. We also recovered another pattern of co-conserved cysteines that form a disulfide bond in the substrate-binding C-lobe. We then discuss some possible functional consequences of these distinguishing features of the ROPK family.

4.2 Results

To examine the molecular evolution and functional shifts in ROPKs, we used the genomic, mRNA and proteomic sequences of multiple *T. gondii* strains, *Neospora caninum*, *Sarcocystis neurona* and *Eimeria tenella* to develop profiles for 42 subfamilies of ROPK, reflecting orthology as well as chromosomal patterns of tandem repeats (see Methods).

We used these sequence profiles to perform an analysis of evolutionary constraints, applying statistical tests of contrasting conservation between gene clades to identify potential sites of subfunctionalization and neofunctionalization in the ROPK family and each ROPK subfamily. We then mapped the sites and regions of interest onto solved structures of ROP2, ROP8 and ROP5 to examine the structural and possible functional roles these features may play within the parasite proteins.

4.2.1 Global trends in the ROPK family

We used a set of HMM profiles derived from our subfamily sequence alignments to scan the translated gene model sequences available for *T. gondii* strains GT1, ME49 and VEG, *N. caninum* and *E. tenella* and classify putative ROPK genes into the identified subfamilies. We found 37, 55 and 38 ROPK genes in *T. gondii* strains GT1, ME49 and VEG, respectively, 44 in *N. caninum* and 27 in *E. tenella*. The elevated ROPK counts in *T. gondii* ME49 relative

to the other strains is probably due to differences in sequencing depth and the quality of assembly and gene model annotation; we also found genomic evidence of unannotated orthologs in the other strains. As suggested by Reese and Boyle [62], ROPK genes are often present in expanded loci (sites of gene duplication, usually in tandem array) and are probably undercounted in annotated genomes.

By incorporating sequences from multiple coccidian species into HMM profiles, we were able to identify several putative ROPKs that were not identified in previous computational surveys [21, 56]. These include the proposed subfamilies ROP47, ROP48, ROP49 and ROP50, present in *T. gondii* and *N. caninum*, and the *E. tenella*-specific subfamilies ROPK-Eten1, ROPK-Eten2a, ROPK-Eten2b, ROPK-Eten3, ROPK-Eten4, ROPK-Eten5 and ROPK-Eten6. We suggest these to be likely rhoptry kinases on the basis of sequence homology, phylogenetic placement, signal peptide presence, and existing experimental evidence. Protein or mRNA expression has been previously observed for at least one member of each of these proposed subfamilies, indicating that they are not pseudogenes. ROP47, ROP49 and ROP50 are predicted to contain a signal peptide. The gene coding for ROP48 has only been annotated in *T. gondii* strain ME49 (TGME49_234950, numbered TGME49_034950 in ToxoDB prior to version 8.0), but we identified genomic regions with 95% sequence identity to this protein sequence on chromosome X of strains VEG and GT1 as well. Recently, a proteomics study observed two *E. tenella* proteins expressed during the sporozoite stage and localized in the rhoptries: ETH_00027700, which we assigned to the ROPK-Eten1 subfamily, and ETH_00005190, which we assigned to the ROPK-Unique category [52]. A search of the available *S. neurona* expressed sequence tags (ESTs) and genomic scaffolds indicates that ROPKs are prevalent in this species as well, though we cannot assign a specific number until the assembly is complete. The subfamilies that have clear representatives in all four of the surveyed species are ROP21/27 and ROP35.

In *S. neurona*, rhoptries are present in the sporozoite [41] and bradyzoite [18] stages but absent from schizonts and merozoites [74]. Surprisingly, we found *S. neurona* genomic regions and EST sequences from the schizont and merozoite stages that appear to code for rhoptry kinases. Of the ESTs currently available in the NCBI GenBank EST database, we identified seven putative rhoptry kinases [GenBank:BM303139.1, BM303688.1, BQ749596.1,

BQ750005.1, BU085181.1, CO748650.1, CV193082.1], all obtained from the *S. neurona* merozoite stage, evidence that these genes are indeed expressed despite the absence of rhoptry organelles during this life stage. We also examined genomic open reading frames (ORFs) for signal peptides using the program SignalP [58] and identified likely signal peptide regions and cleavage sites in several of the ORFs that we predicted to encode rhoptry kinases, suggesting that at least some of these are likely to be exported.

Both pseudokinases and catalytically active kinases appear to be prevalent throughout the ROPK family, in roughly equal numbers of subfamilies. The pseudokinase subfamilies are distributed throughout the phylogenetic tree, rather than forming any distinct clade, suggesting that the evolutionary pressures that lead to the degeneration of paralogs into pseudokinases have applied throughout the ROPK family.

Phylogenetic clustering reveals distinct sub-clades

We inferred a phylogenetic tree from the consensus sequences of each of the ROPK subfamilies to illustrate evolutionary patterns within the ROPK family (Figure 4.1). Several distinct clades emerged, which we examined more specifically: rhoptry kinases with homology to the N-terminal extension (NTE) observed in ROP2, ROP8 and ROP5 structures (discussed below); an expanded clade of seven subfamilies specific to *E. tenella*; and a basal clade of divergent, ROPK-like protein kinases, including ROP35 and BPK1, which we refer to as ROPKL here.

Within the *E. tenella*-specific clade, the putative ROPK proteins ETH_00028855, ETH_00020620 and ETH_00000075, which we placed in the subfamilies Eten2b, Eten3 and Eten4, respectively, were recently observed to be expressed solely in merozoites [52]. The emergence of this gene clade reflects the significant phylogenetic and phenotypic divergence of the oocyst-forming *E. tenella* from the other tissue-cyst-forming coccidian species we have examined here [45]. *E. tenella* also contains several putative ROPKs outside this clade, more closely related to the ROPKs found in *T. gondii* and *N. caninum*, which we placed in the ROPK-Unique category.

The previously identified proteins in the ROPKL clade are ROP33, ROP34, ROP35 and ROP46. The clade also contains the brazyzoite-expressed pseudokinase BPK1 [11]. The

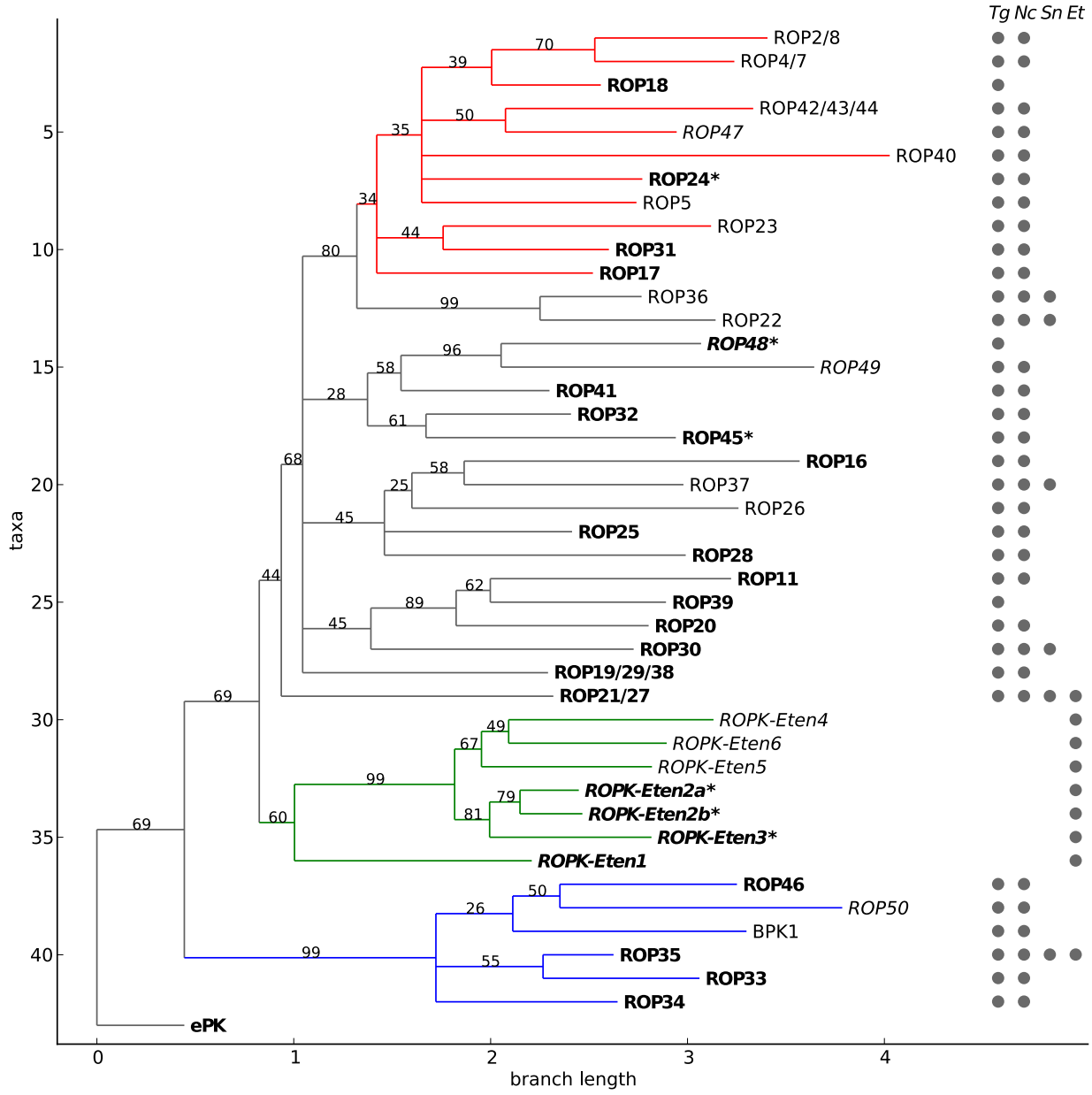


Figure 4.1: Phylogeny of rhoptry kinase subfamilies. Predicted or known active kinases are labeled in bold text, and kinases that may have a noncanonical catalytic mechanism are marked with an asterisk. Newly proposed ROPK subfamilies are labeled in italic text. The clade indicated in red contains the ROPK subfamilies with a homologous N-terminal extension to the kinase domain (NTE). The clade in green is specific to *E. tenella*. The divergent "ROPKL" clade is shown in blue. Branch labels indicate bootstrap support. The grid along the right side indicates the species in which each subfamily appears: *T. gondii* (Tg), *N. caninum* (Nc), *S. neurona* (Sn) and *E. tenella* (Et).

gene models of the ROPKL proteins in *T. gondii* ME49, the best-annotated strain, all contain at least one intron, in contrast to most other ROPK genes, which are typically encoded by a single exon.

Known or likely catalytic kinases

In our analysis, we consider the catalytically essential residues to be the aspartate in the catalytic loop ("HRD" motif, $D166^{PKA}$) and the aspartate in the Mg-binding loop at the start of the activation segment ("DFG" motif, $D184^{PKA}$); we categorize the ROPK subfamilies missing either of these residues as pseudokinases. Additionally important residues involved in ATP positioning or conformational changes necessary for catalytic activity include a glycine in subdomain I ($G52^{PKA}$), lysine in subdomain II ("VAIK" motif, $K72^{PKA}$), glutamate in subdomain III ($E91^{PKA}$) and asparagine in the catalytic loop ($N171^{PKA}$) [34, 68, 79], as well as the F-helix aspartate which positions the catalytic loop ("DxW" motif, $D220^{PKA}$) [55]. While catalysis has been observed in kinases that lack one or more of these residues, their absence usually indicates a noncanonical mechanism or impairment of activity [47, 71, 82].

The subfamilies ROP11, ROP16, ROP17, ROP18, ROP19/29/38, ROP20, ROP21/27, ROP25, ROP28, ROP30, ROP31, ROP32, ROP35, ROP39 and ROP41 were previously suggested to be active kinases based on the conserved catalytic triad [56]. Phosphoryl transfer has been demonstrated experimentally for ROP18 [60] and ROP16 [53], and molecular modelling simulations have shown that ATP could dock in a typical conformation to ROP11, ROP16, ROP17 and ROP18 [37]. Our analysis additionally found the catalytically essential residues conserved in ROP33, ROP34 and ROP46, suggesting these may also be active kinases. Of the *E. tenella*-specific subfamilies we identified, ROPK-Eten1 also retains all of the key residues needed for catalysis (Figure 4.2).

Known or likely pseudokinases

Kinases that lack one or more of the residues necessary for catalysis are likely to be non-catalytic pseudokinases. The apparent pseudokinase ROPK subfamilies are ROP2/8, ROP4/7, ROP5, ROP22, ROP23, ROP26, ROP36, ROP37, ROP40 and ROP42/43/44, as identified previously [56]. We include BPK1, previously noted as a *T. gondii* brazyzoite-

expressed pseudokinase [11], in the ROPK family based on sequence similarity. Additionally, our proposed subfamilies ROP47, ROP49, ROP50, and the *E. tenella*-specific ROPK-Eten4, ROPK-Eten5 and ROPK-Eten6, are also missing key aspartates involved in the kinase catalytic mechanism and are likely to be pseudokinases (Figure 4.3). ROP50 does have an aspartate at the HRD+3 position (Figure 4.3), so in absence of a structure we cannot rule out that this nearby residue may play a compensatory role in catalysis.

Several of these pseudokinase subfamilies share the unusual characteristic of replacing the catalytic aspartate (in the kinase-conserved "HRD" motif) with a basic residue: ROP4/7 (HGK), ROP5 (HG[R/K/H]), ROP22 (HTH), ROP36 (HGH), ROP40 (LRR) and ROP42-43-44 (HGK), as previously noted [61].

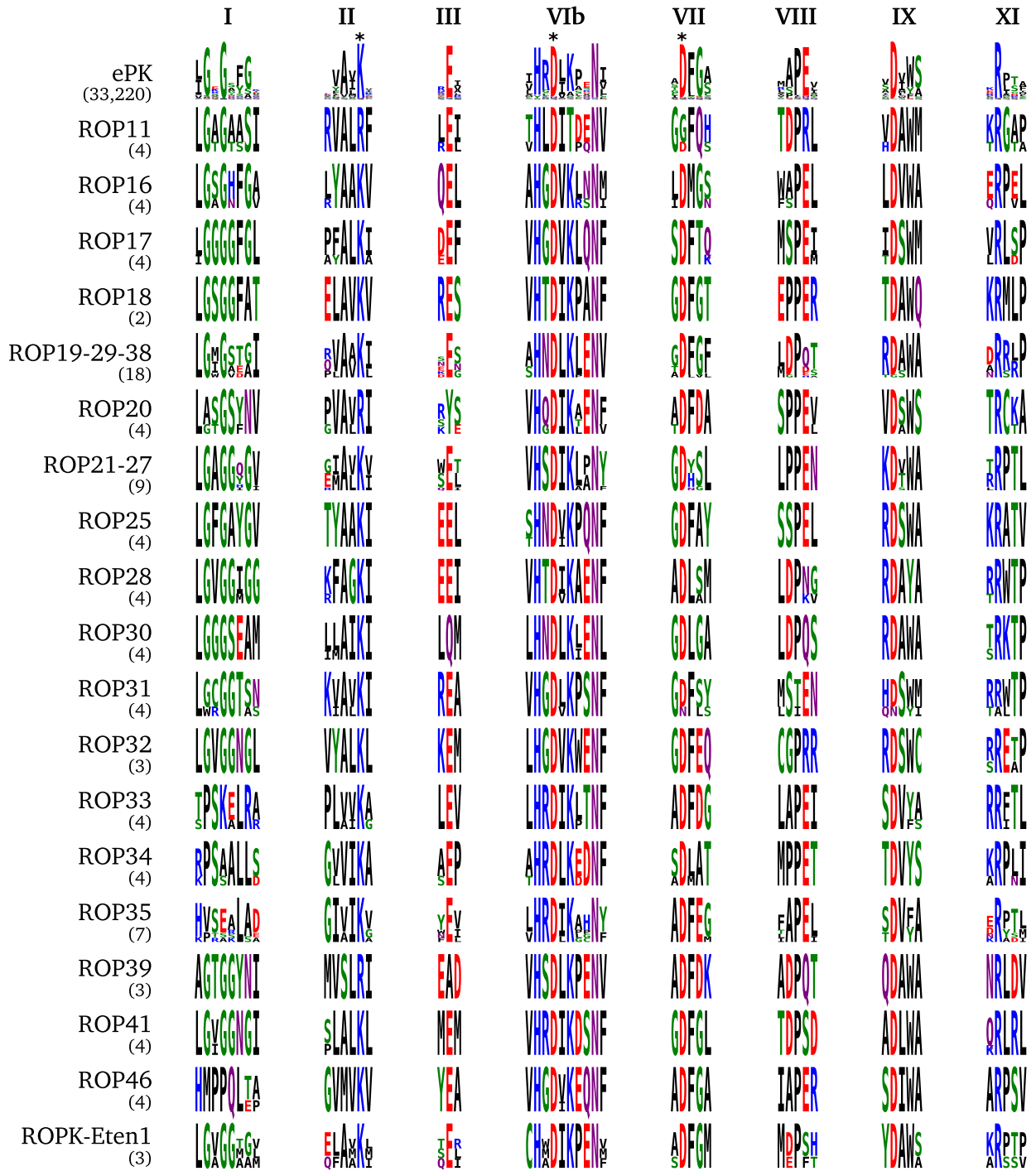
Noncanonical kinases

The subfamilies ROP24, ROP45 and the proposed ROP48, ROPK-Eten2a and ROPK-Eten2b have most of the residues necessary for catalysis, but with some differences in other typically conserved residues that suggest the mechanisms may be noncanonical (Figure 4.4).

In most active ePKs, an asparagine in the catalytic loop ($N171^{PKA}$) coordinates a magnesium ion to position ATP in the active site [34]. This residue varies among some ROPKs: In ROP24, ROP45 and ROP48, the asparagine is replaced by a basic residue

---

Figure 4.2 ***(following page)*:** Conserved motifs of catalytically active rhoptry kinase subfamilies. Sequence logos of key regions in the kinase domain of the broader ePK superfamily and of predicted active ROPK subfamilies as they occur in the coccidian species examined. Letter height at each sequence position indicates greater conservation of that character in a multiple sequence alignment of a large set of ePK sequences (first row) and the annoted genomic sequences of each ROPK subfamily. Row labels indicate subfamily names, with the number of sequences in each alignment shown in parentheses. The ePK-conserved motifs shown are the glycine-rich loop in subdomain I, catalytic lysine in subdomain II, αC glutamate in subdomain III, catalytic loop in subdomain VIb, "DFG" in subdomain VII, "APE" in subdomain VIII, αF "DxW" in subdomain IX, and arginine in subdomain XI. The adjacent sequence sites surrounding each motif are included for context. Asterisks indicate above ePK motifs indicate the catalytic triad. Generated using the WebLogo [15] and ReportLab [64] libraries.


I
II
III
VIb
VII
VIII
IX
XI
ePK (33,220)
ROP11 (4)
ROP16 (4)
ROP17 (4)
ROP18 (2)
ROP19-29-38 (18)
ROP20 (4)
ROP21-27 (9)
ROP25 (4)
ROP28 (4)
ROP30 (4)
ROP31 (4)
ROP32 (3)
ROP33 (4)
ROP34 (4)
ROP35 (7)
ROP39 (3)
ROP41 (4)
ROP46 (4)
ROPK-Eten1 (3)

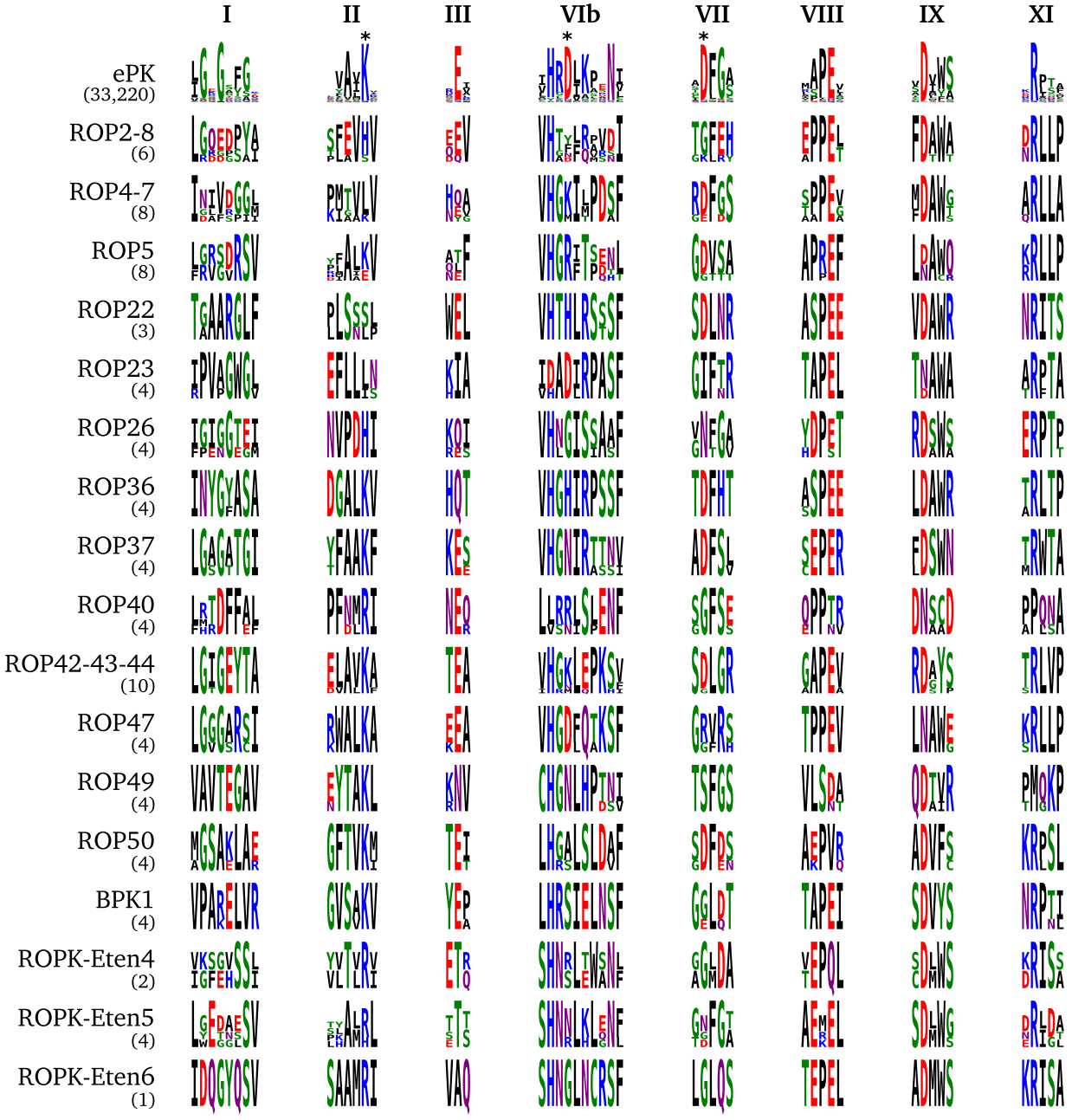


Figure 4.3: Conserved motifs of likely inactive rhoptry kinase subfamilies. Sequence logos of conserved motif regions in the kinase domain of the broader ePK superfamily and of predicted pseudokinase ROPK subfamilies as they occur in the coccidian species examined.

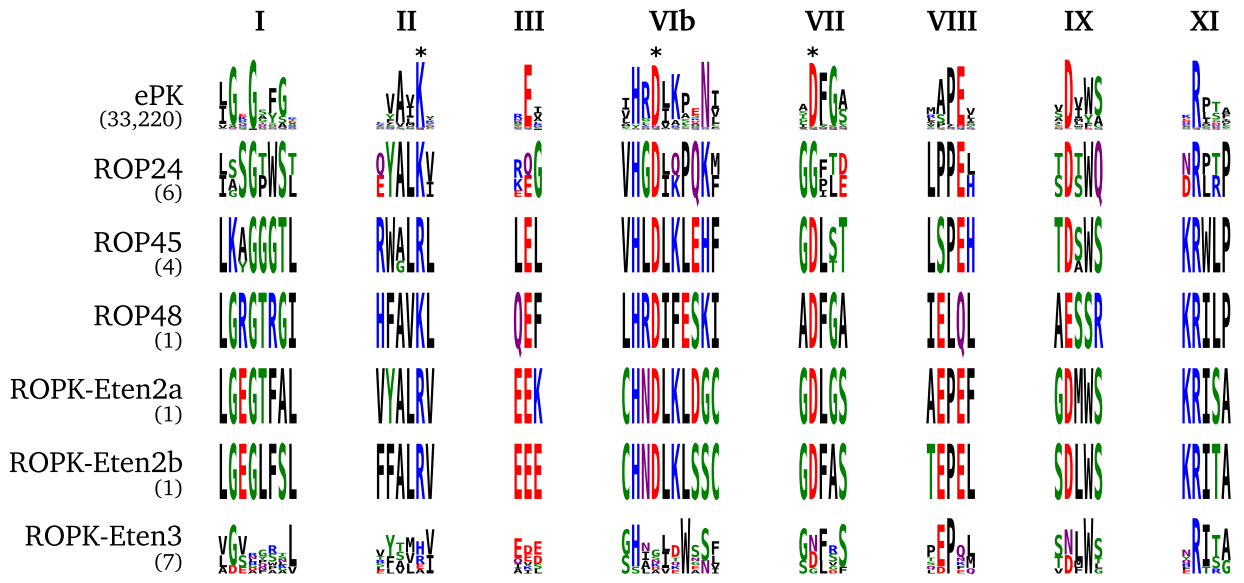


Figure 4.4: Conserved motifs of ROPK subfamilies with potentially noncanonical catalytic mechanisms. Sequence logos of conserved motif regions in the kinase domain of the broader ePK superfamily and of ROPK subfamilies with predicted noncanonical catalytic mechanisms as they occur in the coccidian species examined.

(lysine, histidine and lysine, respectively). The closely related *E. tenella*-specific subfamilies ROPK-Eten2a and ROPK-Eten2b have the catalytic loop motifs HNDLKLDG and HNDLKLSS, respectively, each replacing the ePK-conserved asparagine with a different residue type. Such replacements are rare in catalytically active kinases; in an alignment of ePK sequences (not shown), we observed only two cases in which the "HRD" motif is conserved without the accomanying asparagine, both of which have been shown to have noncanonical catalytic mechanisms: CASK [47], which replaces the asparagine with a cysteine, and Type II PAK [22], which has a serine.

The ePK-conserved lysine in subdomain II (β3) is replaced with arginine in ROP45, ROPK-Eten2a and ROPK-Eten2b, though the conserved C-helix glutamate is retained, suggesting the necessary salt bridge could still form in the active state of these kinase as in other ePKs. In ROP24, however, the lysine is retained but the corresponding C-helix glutamate is instead alanine, precluding a salt bridge. The DFG motif is replaced with the sequence GFT, though a potentially compensatory acidic residue appears at the DFG+1

position. These observations suggest that the activation mechanism [29, 36] in ROP24 could be different from that of other ePKs. ROP48 retains the β3 lysine, αC glutamate and DFG motif; however, the substrate-binding lobe is quite divergent, with a dramatically shortened activation loop and F-helix, and the F-helix DxW motif is replaced with ESS, which suggests that the positioning of the catalytic loop occurs differently from other ePKs.

The *E. tenella*-specific subfamily ROPK-Eten3, in contrast to all the other identified ROPK subfamilies, appears to comprise both active and inactive kinases. The locus appears as a tandem repeat of 5 similar genes, with pairwise identity ranging from 32% to 52% (mean 41%), only one of which (ETH_00020585) retains the key residues indicating catalytic function (Figure 4.4).

4.2.2 ROPK-conserved inserts within the protein kinase domain

ROPK- and subfamily-specific inserts within the kinase domain are widespread, suggesting unique functional adaptations [37, 60, 61]. We found six conserved inserts in the ROPK domain relative to the PK domain (Figure 4.5). They are:

(i) An extension of the β3–αC loop, residues 289–293$^{\mathrm{ROP2}}$, of varying length across ROPK subfamilies; it is fairly short (4–5 amino acids) in the NTE-bearing clade, missing altogether in ROPKL, but extends up to 13 amino acids other ROPKs including the *E. tenella*-specific clade.

(ii) C-terminal to the αC helix, residues 309–318$^{\mathrm{ROP2}}$, present in all subfamilies except the ROPKL clade in roughly equal size. In the ROP2/8 structures [PDB:2W1Z,3DZO,3BYV] it was observed to form an additional helix, termed αC' [37], in the kinase inter-lobe hinge area (discussed below), while in the ROP5 structures [PDB:3Q5F,3Q60] it is disordered.

(iii) In β4–β5 loop, residues 335–351$^{\mathrm{ROP2}}$, present in most subfamilies, including ROP33 but not the other ROPKLs, in similar size. In a ROP2 structure [PDB:2W1Z] this appears as two β strands, termed β' and β", that extend the loop to form a β-hairpin in the kinase N-lobe [37], spatially near the α" helix of the NTE. In the other structure of ROP2, ROP8 and ROP5 [PDB:3DZO,3BYV,3Q5F,3Q60] this region is mostly disordered, though the protein sequences indicate the insert is present in this subfamily as well.

(iv) Between the kinase APE motif (end of the activation segment) and the αF helix, residues 453–462$^{\mathrm{ROP2}}$, present in varying lengths across the ROPK subfamilies including each of the major clades (NTE, Eten, ROPKL). This is near the substrate-binding site in typical protein kinases. The insert appears as a short 4aa loop in ROP5 [PDB:3Q60], but in ROP2 [PDB:3DZO] and ROP8 [PDB:3BYV] it forms an additional single-turn helix in crystal structures [PDB:3DZO, PDB:3BYV] [60], though this feature may have been stabilized in the crystals because of crystal packing.

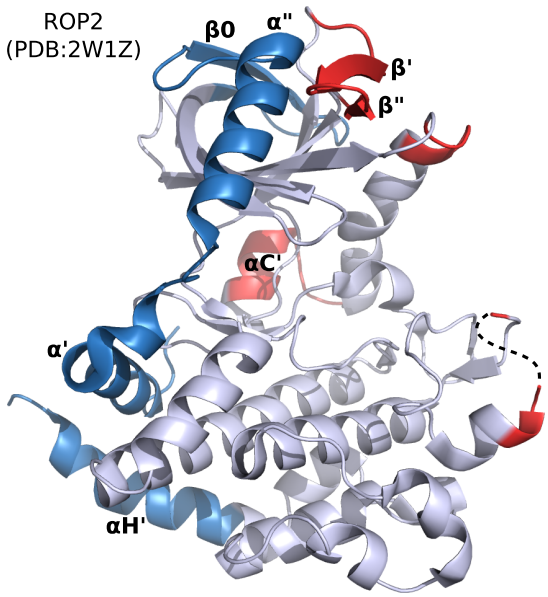


Figure 4.5: Structural location of ROPK-conserved inserts. Inserts relative to the conserved ePK fold are highlighted in red. The N-terminal extension (NTE) and C-terminal extended helix (αH') are shown in blue. Novel secondary structures are labeled according to Labesse et al. [37].

(v) An extension of the αF–αG loop, absent from ROP2/8, ROP40 and ROP49 and the ROPKL clade, but present in ROP5 and the other ROPK subfamilies in the region of residues 467–478$^{\mathrm{ROP5}}$. In the ROP5 structures [PDB:3Q5F,3Q60], B-factors indicate this elongation of the αF–αG loop is relatively flexible compared to the adjacent regions; the G-helix itself appears unfolded. Sequences of other ROPKs, including ROP24, suggest it is even longer in those subfamilies.

(vi) In the αG–αH loop, near the C-terminus of the αG helix, a 5aa insert absent from ROP2/8, ROP5, ROP18, ROP23, ROP25, ROP26, ROP30 and ROP40 and the ROPKLs but present in the other ROPK subfamilies including the *E. tenella*-specific clade. The ROPKLs appear to have large deletions in this region, and may be missing the αG helix structure altogether. We note that the αG–αH loop is extended in many other protein kinases, most notably CMGC kinases [30].

4.2.3 Distinguishing ROPK-specific conserved sites in the protein kinase domain, and corresponding structural features

We evaluated shifts in site-specific residue conservation between the ROPK family and overall PK superfamily by performing a goodness-of-fit test of residue frequences in the two sequence sets at each aligned column of the PK domain (see Methods). The same comparisons were also performed with each subfamily versus the other ROPKs.

Hinge region

The most statistically significant sites distinguishing ROPKs from PKs overall are in the kinase hinge region. Numbered according to ROP2 [PDB:2W1Z], these are: sites [E/R/Y]320, L321, [R/G]322, [V/L/A]325 and P326 in the αC'–β4 loop; P358 in the β5–αD loop, and [L/F/Y]424 in the β8 strand (Figure 4.6). Two residues in the αE helix, [L/A/S]396 and [H/N/S]399, are oriented toward the hinge region and under the αC' helix.

The residue P358$^{\mathrm{ROP2}}$ is typically a glutamate in most eukaryotic protein kinases (e.g. E121$^{\mathrm{PKA}}$, E565$^{\mathrm{FGFR}}$), where it contributes to the opening/closing motion of the kinase during activation by forming a lobe-bridging salt bridge interaction [38]. In fibroblast growth factor receptor kinase (FGFR), for example, the equivalent residue E565 hydrogen-bonds with K641 in the β8 strand conditionally upon phosphorylation of the FGFR activation loop [13] (Figure 4.6D,E). In ROP2, the residues equivalent to E565 and K641 are P358 and F424, respectively (Figure 4.6A,B). Since proline and phenylalanine are not charged residues, the ROP2 structure is incapable of forming the same interaction. The residue P358$^{\mathrm{ROP2}}$ is conserved as a proline throughout most of the ROPK family, with the exception of subfamilies ROP18 (methionine), ROP21/27 (aspartate, though a Phe appears in the β8 strand), ROP26 (serine), ROP32 (histidine), ROP41 (lysine), and the *E. tenella*-specific subfamilies (retained as glutamate, though only ROPK-Eten1 also retains a basic residue in the β8 strand).

The residues at sites P358$^{\mathrm{ROP2}}$ and P326$^{\mathrm{ROP2}}$ appear to have instead taken on another structural role. In ROPKs, the residue immediately N-terminal to P358$^{\mathrm{ROP2}}$, a site known as the kinase "gatekeeper" residue, is a large, usually hydrophobic residue oriented toward the αC'–β4 strand and, in the ROP2 structure, packing against the ROPK-conserved P326;

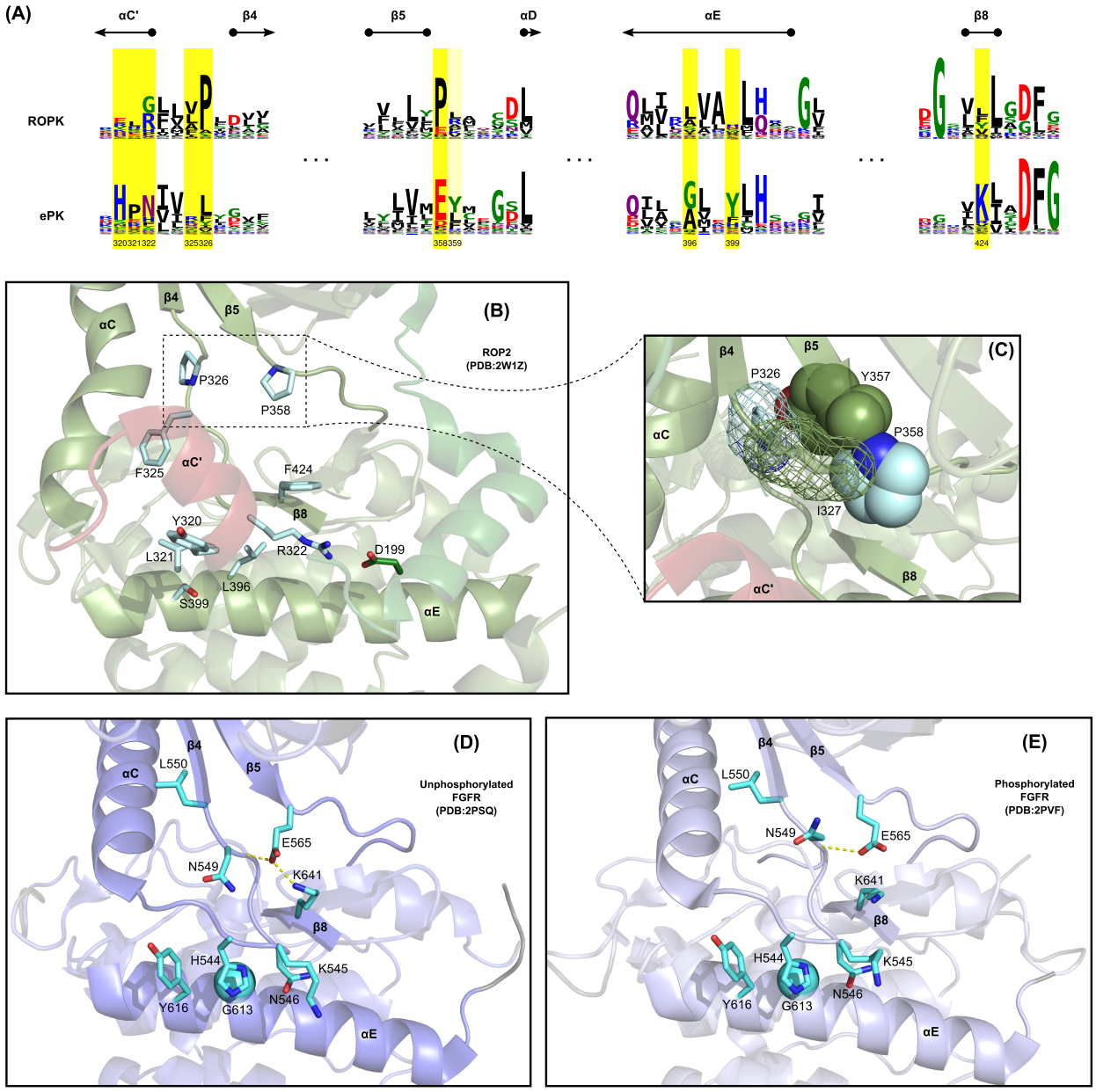


Figure 4.6: Contrasting sites between ROPK and PK in the kinase hinge region. **(A)** Sequence logos of regions surrounding the αC' helix insert and kinase hinge in ROPK (top) and PK (bottom), with selected contrasting sites highlighted. **(B)** ROP2 structure with selected contrasting sites shown in "sticks" representation. **(C)** Inset of the ROP2 structure around the contrasting prolines in the αC'–β4 loop and linker. Two other adjacent residues, I327 and Y357, pack against the prolines on opposite sides. (D) and (E) Two structures of the protein kinase FGFR show the conditional salt bridge between the linker glutamate (E565) and β8 lysine (K641), commonly observed in typical protein kinases, dependent upon kinase activation.

the hydrophobic residue immediately N-terminal to P326 (most commonly valine but also varyingly leucine, alanine, phenylalanine, isoleucine and methionine in ROPKs) is likewise oriented toward the linker in the ROP2 structure, packing against P358 (Figure 4.6C). These four residues thus form a stable packing "box" bridging the αC'–β4 and β5–αD loops.

F-helix "WC" motif and disulfide bridge

A distinctive "WC" motif appears at the end of the αF helix (Figure 4.7) in most ROPKs. The cysteine (C485$^{\mathrm{ROP2}}$), together with another ROPK-conserved cysteine (C506$^{\mathrm{ROP2}}$) [21] in the αG–αH insert described above, forms a disulfide bond which has been proposed to stabilize the two helices [60]. The tryptophan (W484$^{\mathrm{ROP2}}$) appears to pack against the extended αD and αE helices, pushing the αE helix futher outward. Thus the "WC" motif couples two ROPK-specific inserts to the substrate-binding lobe of the kinase core. There are no other known protein kinase families or subfamilies in which cysteines at the end of the F-helix and in the αG–αH loop co-occur in positions that could potentially interact. Additionally, both the WC motif and the αG–αH cysteine are absent from the *E. tenella* and ROPKL clades.

Another site in the αF helix (W482$^{\mathrm{ROP2}}$) is conserved as a glutamate in most ePKs (E230$^{\mathrm{PKA}}$), but unconserved in ROPKs, suggesting that a selective constraint that conserves glutamate at this site in most ePKs has been lost in the ROPK family. In at least some other ePKs, it appears that this glutamate can interact with a basic residue on the polar/charged surface of the amphipathic αD helix (R133$^{\mathrm{PKA}}$), as well as a conserved tyrosine in the P+1 pocket (Y204$^{\mathrm{PKA}}$) at the end of the activation segment (Figure 4.7D,E). Notably, the mutation of E230 to glutamine in PKA not only disrupted substrate recognition and phosphoryl transfer, but also resulted in higher temperature factors in the αD helix, particularly in R133 [81]. However, in ROPKs the interaction between the F and D helices occurs somewhat differently: in ROP5, R455 interacts with E345 and Y427, and in ROP2, W482 packs with H365, while the P+1-pocket Tyr replaced by F446, a side chain not capable of hydrogen bonding (Figure 4.7A,B).

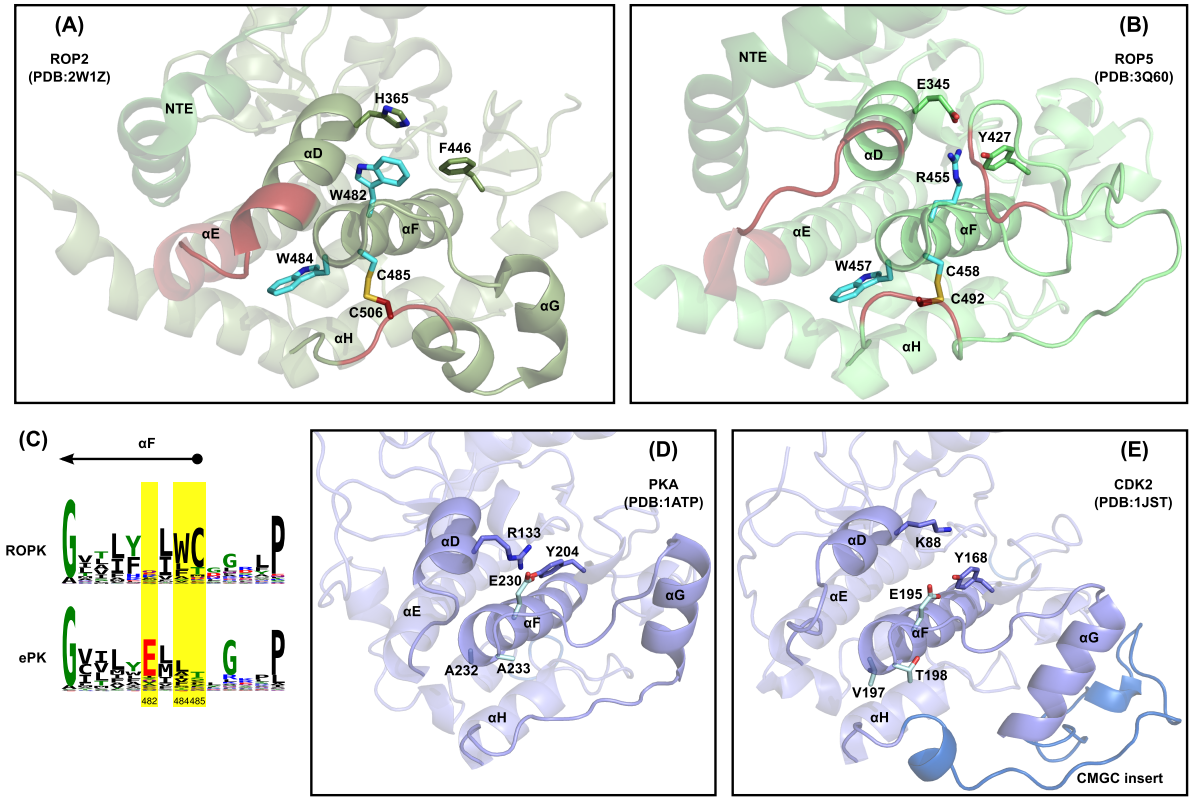


Figure 4.7: Contrasting sites between ROPK and PK: C-lobe WC motif and loss of Glu constraint.
(A) and **(B)** Structures of ROP2 and ROP5 with WC motif and ROPK-conserved disulfide bridge residues shown in "sticks" representation.
(C) Sequence logos of F helix region in ROPK (top) and ePK (bottom), with contrasting sites highlighted.
(D) PKA, a representative typical protein kinase, with equivalent residues shown as sticks.
(E) CDK2, another typical protein kinase. The "CMGC insert" occurs in the αG–αH loop but does not perform the same structural role as the ROPK-specific insert in the same region.

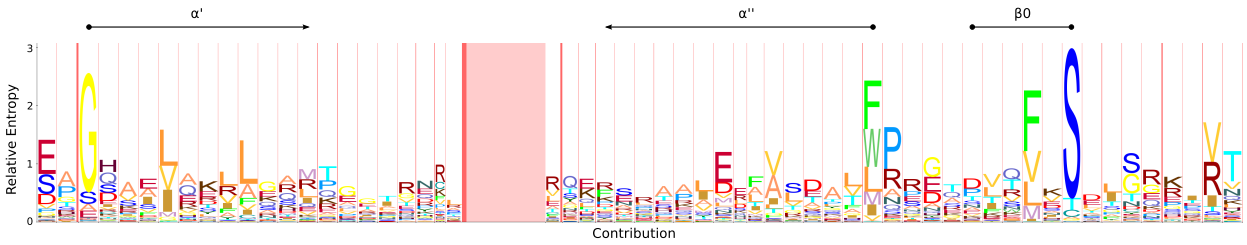


Figure 4.8: HMM sequence logo of the NTE region. Conserved secondary structures are indicated above the corresponding sequence positions. Generated with the HMM-Logos server LogoMat-M [70].

4.2.4 N-terminal extension to the protein kinase domain

Structural studies of ROP2, ROP8 and ROP5 revealed another feature common to each of these proteins, an N-terminal extension (NTE) to the canonical protein kinase domain consisting of at least two additional helices and a beta sheet, with the region between the two helices varying between ROP2/8 and ROP5 [37, 60, 61]. The NTE has also been suggested to be present in ROP18, ROP4/7 and ROP17 based on sequence homology, though its presence does not appear to be universal among rhoptry kinases [60, 61]. We investigated the distinguishing features of NTE-containing rhoptry kinases to determine whether other ROPKs may also contain the NTE, and to look for additional conserved features that characterize this gene clade (see Methods).

In addition to ROP2/8 and ROP5, we found significant matches in ROP4/7, ROP17 and ROP18, as expected, and also a number of additional subfamilies which appear to form a clade (Figure 4.1): ROP23, ROP24 (originally known as ROP2L8 [7]), ROP31, ROP40, ROP42/43/44, and the proposed ROP47. Four proteins in the ROPK-Unique (species-specific) category also showed evidence for NTE homology: TGME49_296000 (TGME49_096000 in ToxoDB prior to version 8.0), also known as ROP2L12 and previously identified as a pseudogene [7]; its orthologs TGVEG_050080 and TGGT1_054010; and the *E. tenella* protein ETH_00005190. A small number of sites in the NTE sequence region show strong conservation (Figure 4.8).

Having identified the NTE-bearing clade, we then compared this clade to all other

identifid ROPKs to identify clade-specific residue conservation patterns. In the solved structures of ROP2, ROP8 and ROP5, several of these distinctive sites in the NTE clade are spatially located around the NTE itself, primarily near the conserved β0 and α' secondary structure elements. In ROP2, V330 and P333 in the β4 sheet β4–β4' loop are positioned on either side of the β0 sheet of the NTE, close to the conserved S244; in ROP5, the equivalent residues are V310 and Q313. In each of the solved crystal structures of ROP2 [PDB:2W1Z], ROP8 [PDB:3BYV] and ROP5 [PDB:3Q60], the β0 sheet passes directly between these two side chains, suggesting a structural selective constraint in NTE-bearing ROPKs.

Three significantly contrasting sites in the E-helix may also have some bearing on the NTE conformation or placement: H378 near the αE N-terminus, oriented toward the NTE in the ROP2 structure [PDB:2W1Z]; V382, a small, nonpolar residue oriented toward the extended αD; and Q388 in the middle of the αE helix, where in the ROP2 structure it interacts with the backbone of the conserved G198 at the N-terminus of the NTE α' — though in the ROP5 structure the equivalent residue is I368 which despite having the same orientation cannot form an identical interaction.

Also in the αE helix, a hydrophobic residue (L391$^{\mathrm{ROP2}}$, A371$^{\mathrm{ROP5}}$), in place of a usually basic residue outside the NTE clade, is oriented toward a helix which extends beyond the kinase C-terminus in the ROP2, ROP8 and ROP5 structures, previously described as the αH' helix [37]. Though this short, weakly conserved region is difficult to detect by sequence analysis, the conservation of the hydrophobic residue in the αE helix and the presence of this helix in the available structures does suggest a correlation between the presence of the NTE and C-terminal αH' helix.

4.3 Discussion

We classified the ROPKs into likely active kinases, likely pseudokinases, and predicted kinases that may be active, but with a noncanonical catalytic mechanism, based on differences in ePK-conserved residues surrounding the ATP binding pocket. Our alignment shows that conserved residues in or near the key ePK-conserved motifs, including the histidine of the canonical "HRD" motifs, are well aligned for each of these categories, so it is unlikely

that the absence of the key aspartates in predicted pseudokinases is due to misalignment. Structural investigation of the unusual motifs in noncanonical subfamilies ROP24 and ROP45 in *T. gondii* could reveal novel kinase mechanisms of activation, ATP positioning and catalysis. Relatedly, analysis of the equivalent motifs in the ROPK pseudokinases could improve our understanding of pseudokinases in general.

Our phylogenetic tree of ROPK subfamilies revealed three specific clades of interest: the NTE-bearing ROPKs, the only clade for which crystal structures have been solved or even homology models reliably constructed; an *E. tenella*-specific expansion of ROPKs; and the divergent, intron-bearing ROPKLs. Notably, each of these clades contains both predicted active kinases and pseudokinases, indicating a pattern of evolution in which, in a parsimonious interpretation, pseudokinases repeatedly emerge from an ancestral state shared with active kinases, rather than a single or small number of expansions of pseudokinases.

We were unable to find conclusive published evidence that the ROPKL proteins are indeed localized to the rhoptry during the tachyzoite stage of coccidians and expelled during invasion at the same time and through the same mechanism as other ROPKs. ROP35 protein expression has been detected during the *T. gondii* tachyzoite stage [80] and the *E. tenella* merozoite stage (ETH_00005905) [52]. Signal peptides were predicted for ROP33, ROP50 and BPK1, but not ROP35, while the gene models of ROP34 and ROP46 contain a short or nonexistent N-tail to the kinase domain which could indicate a trunctated gene model. However, transcription levels across the cell cycle do not match the distinctive two-peaked pattern of *T. gondii* rhoptry proteins in any of the *T. gondii* ROPKLs [2]; the secretory organelle of BPK1 was not identified in the study that described the protein [11]. Our HMM profile search and gene trees indicated that the ROPKL proteins show stronger sequence similarity to typical ROPKs than to any other characterized protein kinase family, leaving open the question of how deep their functional similarity goes.

A common theme we observe in structural features unique to the ROPK family is the interaction between ROPK-specific inserts or structural motifs, including the N-terminal extension (NTE), and conserved sites within the kinase domain that show contrasting selection in ROPKs. Two regions in particular, the kinase hinge region surrounding the αC'

helix and and the dusulphide bridge at the end of the αF helix, suggest several possible functional or mechanistic consequences.

Our observations in the ROPK hinge region raise several hypotheses. The αC' insert in the αC–β4 loop has possible structural analogues in other kinases. The vaccinia-related kinase (VRK) family has a similar insert which packs hydrophobically against the αE helix and was proposed to promote a closed conformation of the kinase domain in the pseudokinase VRK3 [69]; the authors of that study suggested that related active kinases that retain the same feature would be constitutively active. Comparison of the structure of VRK3 [PDB:2JII] to that of ROP2 [PDB:2WIZ] indicates that the ROPK-conserved site L396$^{\mathrm{ROP2}}$ (Figure 4.6A,B) may perform a similar role to the VRK3-conserved F296$^{\mathrm{VRK3}}$ in hydrophobically coupling the two lobes of the kinase domain. Interestingly, the ATP-bound and *apo* structures of the pseudokinase ROP5 show very little overall conformational change [61]. As another example, crystal structures of the yeast SRPK protein Sky1 conserve a short αC' helix insert, and the flexibility of this region is indicated to be critical for interlobe closure [50]. Together with the ROPK-specific conservation of prolines in the αC–β4 loop and linker, this could indicate the possibility that these differences modulate interlobe closure (the kinase hinging mechanism) in ROPKs.

Another hypothesis regarding the function of the αC' helix, not necessarily conflicting with the above hypothesis, is that it could serve as a binding interface or protein-protein interaction site. We observed that the αC' helix does not pack hydrophobically against the N-lobe of the kinase domain in the available ROP2 structures; instead, there appear to be water molecules in between [PDB: 3Q60] [61]. The B-factors are somewhat higher than in the immediately surrounding areas, and the symmetry of the ROP2 structure suggests that the insert may have been stabilized in this structure by crystal packing. Given that the same region is disordered in the available ROP5 structures, it appears possible that αC' may be relatively flexible, capable of unfolding from the helical secondary structure into a mobile loop. For comparison, in VRK3, a surface patch centered on the αC–αC' region has been proposed as a binding site [69].

In the kinase C-lobe, a pair of ROPK-conserved cysteines form a disulfide bridge between the end of the αF helix and the αG–αH loop, which is extended in most ROPKs. A conserved

tryptophan adjacent to the αF cysteine packs hydrophobically against the αD and αE helices, which are also extended in ROPKs; thus the "WC" motif appears to couple both ROPK inserts to the kinase C-lobe. Notably, this stabilization occurs in the surface region of the protein that was identified as polymorphic between ROP5 alleles in *T. gondii* [61], and was recently shown to be the interface of an interaction with the host (mouse) immunity-related GTPase (IRG) protein [23]. Reese et al. proposed an allosteric network involving the NTE and αF helix to link the polymorphic surfaces in the C-lobe and kinase active site in ROP5 [61]. The variability of this site in ROPKs may therefore be justified by its involvement in that network, which itself appears to be variable in ROPKs. We can hypothesize that, at least in ROP5, the increased structural stability provided by the WC motif in this region permits these subfamily-specific mutations to proliferate at this surface without compromising the folding or stability of the kinase C-lobe [6]. This hypothesis assumes that the disulfide bridge is indeed maintained throughout the lifespan of the protein; while it appears as such in the available solved structures, we note that once the protein is inside the host cell, the cytosolic environment is not conducive to disulfide bond formation. The two cysteines involved are co-conserved in not only the PVM-associated ROP2, ROP8, ROP5 and ROP18, but also ROP16, which has been shown to be localized to the host nucleus [67], among other ROPKs.

We also searched for sites that showed conservation specific to the NTE-bearing ROPK clade, rather than ROPKs as a whole. Interestingly, only a small number of strongly contrasting sites emerged as specific to this clade. This could indicate that the mechanistic roles of the NTE vary across even the NTE-bearing clade of ROPKs.

More structural information will be essential to further understand the ROPK family. Currently, only ROPKs from the ROP2/8 and ROP5 subfamilies within the NTE clade have been solved [37, 60, 61]. While these structures have been invaluable in understanding ROPK mechanisms and possible functions, the low sequence identity and presence of indels across subfamilies makes it difficult to produce reliable homology models for ROPK subfamilies outside this clade. We can suggest several important ROPKs outside the NTE clade which appear to be active kinases, are highly expressed [56], and from which we could gain important insights from the solved crystal structure. ROP16 was indirectly

implicated in virulence differences between *T. gondii* strains in mice [66], and also shown to to modulate the host STAT3 and STAT6 pathway response [12, 53, 54, 67, 83], but the precise mechanisms of this action remain to be discovered. Peixoto *et al.* [56] found evidence that ROP38 is involved in modulating the MAPK cascade; the ROP19/29/38 subfamily was also found to be independently duplicated in *T. gondii* and *N. caninum*, thus the other subfamily members could easily be modeled if a ROP38 structure were available. Finally, ROP35 is a representative member of the divergent, poorly understood ROPKL clade; the presence of several indels relative to other ROPKs at structurally important locations in the sequence suggest that a crystal structure would almost certainly reveal surprising variations on the ePK fold and catalytic mechanisms.

4.4 Conclusion

In this study, we developed novel bioinformatic methods to study patterns of diversification and neofunctionalization in the rhoptry kinase family, and integrated the results of a systematic, multi-species analysis with the structural context provided by the solved structures. Our phylogenetic analysis revealed a subfamily-level structure shared across species, as well as lineage-specific expansions within the ROPK family and three distinct sub-clades of ROPK. We applied general knowledge of protein kinase mechanisms to categorize each rhoptry kinase as a likely active, likely pseudokinase, or potentially active but with an atypical catalytic mechanism. We determined the sequence and structural features that distinguish these subfamilies from each other, as well as those that distinguish the ROPK family as a whole from typical ePKs. Where possible, ROPK-specific motifs were placed into structural context to develop functional hypotheses.

This work sheds light on several important but previously unrecognized features shared among rhoptry kinases, as well as the essential differences between active and degenerate protein kinases or pseudokinases. Our studies provide specific hypothesis for further characterizing ROPK structure and function and also inform ongoing efforts to design protein kinase inhibitors for global diseases caused by coccidian parasites.

4.5 Methods

4.5.1 Data collection

The sequences of translated gene models, unannotated genomes and ESTs from the species *Toxoplasma gondii*, *Neospora caninum* and *Eimeria tenella* were retrieved from ToxoDB version 8.1 [24]. Pre-release genomic sequences and ESTs of *Sarcocystis neurona* were provided by the laboratories of Dan Howe, Christopher Schardl and Jessica Kissinger.

After constructing the initial ROPK subfamily profiles (below), additional ROPK sequences were identified in the NCBI databases `est_others` and `nr` and added to the profiles. To obtain putative ROPK sequences from the unannotated *T. gondii* and *S. neurona* genomes, we used the program `exonerate` (https://www.ebi.ac.uk/~guy/exonerate/; also see [73]) to align the ROPK subfamily consensus sequences to each genome scaffold sequence, omitting introns according to likely splice sites. A script using Biopython [14] was then used to extract the highest-scoring putative protein sequences from the `exonerate` output and combine identical sequences and sequence fragments.

4.5.2 Subfamily classification

We previously constructed a database of HMM profiles for every known protein kinase family and subfamily defined in KinBase [42], as well as several apicomplexan-specific kinase families [76]. The ROPK profile in this set was initially constructed from annotated ROPK sequences in ToxoDB, similar to the techinique described by Peixoto *et al.* [56]. Sequences were aligned using MAFFT version 6.940 [31] with a "seed" alignment of the protein kinase domain constructed using published PDB structures [PDB: 2W1Z, 3BYV, 3DZO, 3Q5Z, 3Q60] [37, 60, 61] and the structure alignment program TM-align (May 2012 release) [86]. Finally, HMM profiles were constructed from each sequence alignment and compiled into an HMM profile database. We used this HMM profile database to search the protein and translated EST sequences described in the previous section; those which scored as stronger matches to the ROPK-specific HMM profile than to our ePK profiles were taken as an initial set of putative rhoptry kinases.

We developed a program called Fammer to partially automate the construction and curation of hierarchical protein subfamily sequence profiles for use with HMMer 3.0 [20] and MAPGAPS 1.0 [49], and to use these HMM and MAPGAPS profiles for sequence search, classification and alignment. The Fammer software package, including source code, documentation and the ROPK sequence profiles used in this study, is available at http://github.com/etal/fammer.

The full-length ROPK sequences identified in each annotated coccidian genome and translated EST set were clustered using OrthoMCL version 2.0.3 [39]. We manually trimmed the sequences in each OrthoMCL cluster to the canonical protein kinase domain and aligned the sequence sets with Fammer version 0.1.0 to create an initial set of ROPK subfamily profiles, as well as a set of "unique" or orphan ROPKs which matched the ROPK HMM profile but were not placed into a larger cluster by OrthoMCL.

Iteratively, we performed the following steps to refine the ROPK subfamily classification. We constructed a phylogenetic tree of the consensus sequences of each putative ROPK subfamily, using FastTree version 2.1.5 [59], and merged ortholog groups which were separated by short branches in the tree and, for subfamilies that appeared in multiple copies within a single genome (e.g. ROP2/8, ROPK-Eten3), showed co-localization in the chromosome. Existing descriptions of the annotated *T. gondii* proteins were used to assign names to subfamilies. Unannotated subfamilies that were phylogenetically placed basally to the known ROPKs, indicating closer relationship to other ePKs, were removed. We visually inspected each subfamily sequence set for potential outlier sequences, on the basis of conserved motifs in key regions of the kinase domain, and moved any of these to the "unique" sequence set. We used the Fammer *build* command to realign all sequences and to construct an HMM profile database of all subfamily profiles, then used this database with the Fammer *scan* command to reclassify the "unique" or outlier ROPK sequences. We included a profile of non-ROPK protein kinase sequences in this HMM database in order to identify and remove false positives in the "unique" set as well as subsequent searches of the coccidian proteome, genome and EST sequences. Finally, we used the Fammer *refine* command to perform leave-one-out validation of each subfamily profile versus the "unique" sequence set, following the approach described by Hedlund *et al.* [26]. This process yielded

42 stable subfamilies of ROPK, along with a "ROPK-Unique" profile set of unclassified orphan sequences. We then identified the ROPK complement in each annotated proteome by running the Fammer *scan* command with the final ROPK HMM profile database, each coccidian species' proteome sequences, and an expectation-value cutoff of 10^{-10}.

4.5.3 Subfamily tree inference

We used the curated alignment of consensus sequences from each ROPK subfamily profile and the non-ROPK protein kinase profile as input to infer phylogenic trees. To quickly examine the structure of the ROPK family during profile refinement, we used FastTree [59] with the WAG scoring matrix, gamma model of rate variation and pseudocount correction for gaps. To infer the final tree shown in Figure 4.1, we first used the GUIDANCE server [57] with 100 replicates of PRANK and removed columns with less than 5% support, in order to remove alignment columns that were likely to have been misaligned while retaining most of the potentially phylogenetically informative columns. We then used a script to remove columns that were more than 30% gap characters. This filtering yielded an alignment of 279 columns, slightly less than the length of the top-level ROPK HMM profile (288 columns). We inferred the tree from this alignment using PhyML (December 2011 release) [25], with the LG scoring matrix, gamma model of rate variation, empirically estimated amino acid frequencies and 100 bootstrap runs, taking the output of FastTree as the user-supplied starting tree. Finally, we used script based on the Bio.Phylo module of Biopython [77] to reroot the tree with ePK as the outgroup, collapse all splits with less than 25% bootstrap support, colorize the specific clades of interest and visualize the tree. The alignment of subfamily consensus sequences and the inferred tree have been deposited in TreeBase (http://www.treebase.org/; Study ID: 14212).

4.5.4 Analysis of evolutionary constraints

To identify sites of contrasting conservation between ROPK subfamilies, and between all ROPKs and the broader protein kinase superfamily, we compared aligned sites between two given sequence sets by applying a multinomial log-likelihood test (G-test) [19] of the

residue compositions of each column in the two sets. The test statistic G is derived from the frequencies of each amino acid type as observed in the "foreground" set, O_i, and as expected based on the "background" set, E_i, including pseudocounts taken from the amino acid frequencies of the full alignment.

$$G = 2 \sum_{i \in a.a.} O_i \ln \frac{O_i}{E_i}$$

To adjust for the non-independence of sequences in each set due to phylogenetic relatedness, the aligned sequences in each set are weighted according to the Henikoff heuristic [27], and the amino acid counts in each column are adjusted according to these sequence weights, an approach also used in PSI-BLAST [1]. The test statistic G follows the chi-squared distribution with 19 degrees of freedom (for the 20 amino acid types).

We implemented this test in a program called CladeCompare, available at http://github.com/etal/cladecompare. The output of the program includes (i) a table of the probabilities (p-values) of each site in the combined alignment, (ii) a list of the significantly contrasting sites after adjusting for multiple testing using the Benjamini-Hochberg false discovery rate method [5], and (iii) images of paired "background" and "foreground" sequence logos to illustrate the contrast at significant sites, generated using the WebLogo [15] and ReportLab [64] libraries.

4.5.5 Detection of the N-terminal extension in additional subfamilies

To identify which ROPK subfamilies share sequence homology to the NTE region observed in the ROP2, ROP8 and ROP5 structures, and suggested to be present in ROP18, ROP4/7 and ROP17, we used the CHAIN program [48] with the previously identified NTE-bearing sequences as the query set and the complete set of full-length ROPK sequences as the main set. CHAIN identified a "foreground" partition corresponding to the clade highlighted in Figure 4.1.

We then constructed an alignment of the sequence regions N-terminal to the kinase domain in the identified using the "accurate" mode of T-Coffee [51], built an HMM profile from this alignment, and used HMMer 3.0 [20] to search the full-length ROPK sequences.

This recovered the same ROPK subfamilies identified by CHAIN, confirming the presence of homologous NTE regions in those subfamilies.

4.5.6 Structural analysis

Sites of interest were mapped onto PDB protein structures with a script and visualized in PyMOL [16] for manual inspection.

Author's contributions

ET performed the bioinformatics analyses. ET and NK conceived and designed the study, examined sequences and structural features and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Krishnadev Oruganty and Smita Mohanty for critical feedback and helpful discussions. Pre-release data of the genome and transcriptome of *Sarcocystis neurona* were generated by the laboratories of Dan Howe (Gluck Equine Research Center, University of Kentucky), Christopher Schardl (Advanced Genetic Technologies Center, University of Kentucky) and Jessica Kissinger (Department of Genetics, Institute of Bioinformatics, Center for Tropical & Emerging Global Diseases, University of Georgia), and funded by a USDA National Institute of Food and Agriculture award. We also thank Joshua Bridgers and Sivaranjani Namasivayam for assistance with obtaining the *S. neurona* sequencing data.

This work was supported in part by the National Science Foundation (grant number MCB-1149106).

Bibliography

[1] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402.

[2] Behnke, M. S., Wootton, J. C., Lehmann, M. M., Radke, J. B., Lucas, O., Nawas, J., Sibley, L. D., and White, M. W. (2010). Coordinated progression through two subtranscriptomes underlies the tachyzoite cycle of *Toxoplasma gondii*. *PLoS ONE*, **5**(8), e12354.

[3] Behnke, M. S., Khan, A., Wootton, J. C., Dubey, J. P., Tang, K., and Sibley, L. D. (2011). Virulence differences in *Toxoplasma* mediated by amplification of a family of polymorphic pseudokinases. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(23), 9631–9636.

[4] Behnke, M. S., Fentress, S. J., Mashayekhi, M., Li, L. X., Taylor, G. A., and Sibley, L. D. (2012). The polymorphic pseudokinase ROP5 controls virulence in *Toxoplasma* gondii by regulating the active kinase ROP18. *PLoS Pathogens*, **8**(11), e1002992.

[5] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.

[6] Bloom, J. D., Labthavikul, S. T., Otey, C. R., and Arnold, F. H. (2006). Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(15), 5869–5874.

[7] Boothroyd, J. C. and Dubremetz, J.-F. (2008). Kiss and spit: the dual roles of *Toxoplasma* rhoptries. *Nature Reviews. Microbiology*, **6**(1), 79–88.

[8] Boudeau, J., Miranda-Saavedra, D., Barton, G. J., and Alessi, D. R. (2006). Emerging roles of pseudokinases. *Trends in Cell Biology*, **16**(9), 443–452.

[9] Bradley, P. J. and Sibley, L. D. (2007). Rhoptries: an arsenal of secreted virulence factors. *Current Opinion in Microbiology*, **10**(6), 582–587.

[10] Bradley, P. J., Ward, C., Cheng, S. J., Alexander, D. L., Coller, S., Coombs, G. H., Dunn, J. D., Ferguson, D. J., Sanderson, S. J., Wastling, J. M., and Boothroyd, J. C. (2005). Proteomic analysis of rhoptry organelles reveals many novel constituents for host-parasite interactions in *Toxoplasma gondii*. *The Journal of Biological Chemistry*, **280**(40), 34245–34258.

[11] Buchholz, K. R., Fritz, H. M., Chen, X., Durbin-Johnson, B., Rocke, D. M., Ferguson, D. J., Conrad, P. A., and Boothroyd, J. C. (2011). Identification of tissue cyst wall components by transcriptome analysis of in vivo and in vitro *Toxoplasma gondii* bradyzoites. *Eukaryotic Cell*, **10**(12), 1637–1647.

[12] Butcher, B. A., Fox, B. A., Rommereim, L. M., Kim, S. G., Maurer, K. J., Yarovinsky, F., Herbert, D. R., Bzik, D. J., and Denkers, E. Y. (2011). *Toxoplasma gondii* rhoptry kinase ROP16 activates STAT3 and STAT6 resulting in cytokine inhibition and arginase-1-dependent growth control. *PLoS Pathogens*, **7**(9), e1002236.

[13] Chen, H., Ma, J., Li, W., Eliseenkova, A. V., Xu, C., Neubert, T. A., Miller, W. T., and Mohammadi, M. (2007). A molecular brake in the kinase hinge region regulates the activity of receptor tyrosine kinases. *Molecular Cell*, **27**(5), 717–730.

[14] Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**(11), 1422–1423.

[15] Crooks, G. E., Hon, G., Chandonia, J.-m., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Research*, **14**(6), 1188–1190.

[16] Delano, W. (2011). The PyMOL Molecular Graphics System.

[17] Doerig, C. (2004). Protein kinases as targets for anti-parasitic chemotherapy. *Biochimica et Biophysica Acta*, **1697**(1-2), 155–168.

[18] Dubey, J. P., Lindsay, D. S., Fritz, D., and Speer, C. A. (2001). Structure of Sarcocystis neurona sarcocysts. *The Journal of Parasitology*, **87**(6), 1323–1327.

[19] Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

[20] Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, **7**(10), e1002195.

[21] El Hajj, H., Demey, E., Poncet, J., Lebrun, M., Wu, B., Galéotti, N., Fourmaux, M. N., Mercereau-Puijalon, O., Vial, H., Labesse, G., and Dubremetz, J. F. (2006). The ROP2 family of *Toxoplasma gondii* rhoptry proteins: proteomic and genomic characterization and molecular modeling. *Proteomics*, **6**(21), 5773–5784.

[22] Eswaran, J., Lee, W. H., Debreczeni, J. E., Filippakopoulos, P., Turnbull, A., Fedorov, O., Deacon, S. W., Peterson, J. R., and Knapp, S. (2007). Crystal Structures of the p21-activated kinases PAK4, PAK5, and PAK6 reveal catalytic domain plasticity of active group II PAKs. *Structure*, **15**(2), 201–213.

[23] Fleckenstein, M. C., Reese, M. L., Könen-Waisman, S., Boothroyd, J. C., Howard, J. C., and Steinfeldt, T. (2012). A *Toxoplasma gondii* pseudokinase inhibits host IRG resistance proteins. *PLoS Biology*, **10**(7), e1001358.

[24] Gajria, B., Bahl, A., Brestelli, J., Dommer, J., Fischer, S., Gao, X., Heiges, M., Iodice, J., Kissinger, J. C., Mackey, A. J., Pinney, D. F., Roos, D. S., Stoeckert, C. J., Wang, H., and Brunk, B. P. (2008). ToxoDB: an integrated *Toxoplasma gondii* database resource. *Nucleic Acids Research*, **36**(Database issue), D553–D556.

[25] Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, **59**(3), 307–321.

[26] Hedlund, J., Jörnvall, H., and Persson, B. (2010). Subdivision of the MDR superfamily of medium-chain dehydrogenases/reductases through iterative hidden Markov model refinement. *BMC Bioinformatics*, **11**(1), 534.

[27] Henikoff, S. and Henikoff, J. G. (1994). Position-based sequence weights. *Journal of Molecular Biology*, **243**(4), 574–578.

[28] Hunter, C. A. and Sibley, L. D. (2012). Modulation of innate immunity by *Toxoplasma gondii* virulence effectors. *Nature Reviews. Microbiology*, **10**(11), 766–778.

[29] Huse, M. and Kuriyan, J. (2002). The conformational plasticity of protein kinases. *Cell*, **109**(3), 275–282.

[30] Kannan, N. and Neuwald, A. F. (2004). Evolutionary constraints associated with functional specificity of the CMGC protein kinases MAPK, CDK, GSK, SRPK, DYRK, and CK2alpha. *Protein Science*, **13**(8), 2059–2077.

[31] Katoh, K., Kuma, K.-i., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, **33**(2), 511–518.

[32] Kim, K. and Weiss, L. M. (2004). *Toxoplasma gondii*: the model apicomplexan. *International Journal for Parasitology*, **34**(3), 423–432.

[33] Kim, K. and Weiss, L. M. (2008). *Toxoplasma*: the next 100 years. *Microbes and Infection*, **10**(9), 978–984.

[34] Knighton, D., Zheng, J., Ten Eyck, L., Ashford, V., Xuong, N., Taylor, S., and Sowadski, J. (1991). Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science*, **253**(5018), 407–414.

[35] Kornev, A. P. and Taylor, S. S. (2009). Pseudokinases: functional insights gleaned from structure. *Structure*, **17**(1), 5–7.

[36] Kornev, A. P., Haste, N. M., Taylor, S. S., and Eyck, L. F. T. (2006). Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(47), 17783–17788.

[37] Labesse, G., Gelin, M., Bessin, Y., Lebrun, M., Papoin, J., Cerdan, R., Arold, S. T., and Dubremetz, J.-F. (2009). ROP2 from *Toxoplasma gondii*: a virulence factor with a protein-kinase fold and no enzymatic activity. *Structure*, **17**(1), 139–146.

[38] Lamers, M. B., Antson, A. A., Hubbard, R. E., Scott, R. K., and Williams, D. H. (1999). Structure of the protein tyrosine kinase domain of C-terminal Src kinase (CSK) in complex with staurosporine. *Journal of Molecular Biology*, **285**(2), 713–725.

[39] Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, **13**(9), 2178–2189.

[40] Lim, D. C., Cooke, B. M., Doerig, C., and Saeij, J. P. J. (2012). *Toxoplasma* and *Plasmodium* protein kinases: roles in invasion and host cell remodelling. *International Journal for Parasitology*, **42**(1), 21–32.

[41] Lindsay, D. S., Mitchell, S. M., Vianna, M. C., and Dubey, J. P. (2004). *Sarcocystis neurona* (Protozoa: Apicomplexa): description of oocysts, sporocysts, sporozoites, excystation, and early development. *The Journal of Parasitology*, **90**(3), 461–465.

[42] Manning, G., Plowman, G. D., Hunter, T., and Sudarsanam, S. (2002). Evolution of protein kinase signaling from yeast to man. *Trends in Biochemical Sciences*, **27**(10), 514–520.

[43] Miranda-Saavedra, D., Gabaldón, T., Barton, G. J., Langsley, G., and Doerig, C. (2012). The kinomes of apicomplexan parasites. *Microbes and Infection*, **14**(10), 796–810.

[44] Montoya, J. G. and Liesenfeld, O. (2004). Toxoplasmosis. *Lancet*, **363**(9425), 1965–1976.

[45] Morrison, D. A., Bornstein, S., Thebo, P., Wernery, U., Kinne, J., and Mattsson, J. G. (2004). The current status of the small subunit rRNA phylogeny of the coccidia (Sporozoa). *International Journal for Parasitology*, **34**(4), 501–514.

[46] Morrissette, N. S. and Sibley, L. D. (2002). Cytoskeleton of Apicomplexan Parasites. *Microbiology and Molecular Biology Reviews*, **66**(1), 21–38.

[47] Mukherjee, K., Sharma, M., Urlaub, H., Bourenkov, G. P., Jahn, R., Südhof, T. C., and Wahl, M. C. (2008). CASK Functions as a Mg2+-independent neurexin kinase. *Cell*, **133**(2), 328–339.

[48] Neuwald, A. F. (2007). The CHAIN program: forging evolutionary links to underlying mechanisms. *Trends in Biochemical Sciences*, **32**(11), 487–493.

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

[50] Nolen, B., Ngo, J., Chakrabarti, S., Vu, D., Adams, J. A., and Ghosh, G. (2003). Nucleotide-induced conformational changes in the *Saccharomyces cerevisiae* SR protein kinase, Sky1p, revealed by X-ray crystallography. *Biochemistry*, **42**(32), 9575–9585.

[51] Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, **302**(1), 205–217.

[52] Oakes, R. D., Kurian, D., Bromley, E., Ward, C., Lal, K., Blake, D. P., Reid, A. J., Pain, A., Sinden, R. E., Wastling, J. M., and Tomley, F. M. (2013). The rhoptry proteome of *Eimeria tenella* sporozoites. *International Journal for Parasitology*, **43**(2), 181–188.

[53] Ong, Y.-C., Reese, M. L., and Boothroyd, J. C. (2010). *Toxoplasma* rhoptry protein 16 (ROP16) subverts host function by direct tyrosine phosphorylation of STAT6. *The Journal of Biological Chemistry*, **285**(37), 28731–28740.

[54] Ong, Y.-C., Boyle, J. P., and Boothroyd, J. C. (2011). Strain-dependent host transcriptional responses to *Toxoplasma* infection are largely conserved in mammalian and avian hosts. *PLoS ONE*, **6**(10), e26369.

[55] Oruganty, K., Talathi, N. S., Wood, Z. A., and Kannan, N. (2013). Identification of a hidden strain switch provides clues to an ancient structural mechanism in protein kinases. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(3), 924–929.

[56] Peixoto, L., Chen, F., Harb, O. S., Davis, P. H., Beiting, D. P., Brownback, C. S., Ouloguem, D., and Roos, D. S. (2010). Integrative genomic approaches highlight a family of parasite-specific kinases that regulate host responses. *Cell Host & Microbe*, **8**(2), 208–218.

[57] Penn, O., Privman, E., Ashkenazy, H., Landan, G., Graur, D., and Pupko, T. (2010). GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Research*, **38**(Web Server issue), W23–W28.

[58] Petersen, T. N., Brunak, S. r., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, **8**(10), 785–786.

[59] Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE*, **5**(3), e9490.

[60] Qiu, W., Wernimont, A. K., Tang, K., Taylor, S., Lunin, V., Schapira, M., Fentress, S., Hui, R., and Sibley, L. D. (2009). Novel structural and regulatory features of rhoptry secretory kinases in *Toxoplasma gondii*. *The EMBO Journal*, **28**(7), 969–979.

[61] Reese, M. L. and Boothroyd, J. C. (2011). A conserved noncanonical motif in the pseudoactive site of the ROP5 pseudokinase domain mediates its effect on *Toxoplasma* virulence. *The Journal of Biological Chemistry*, **286**(33), 29366–29375.

[62] Reese, M. L. and Boyle, J. P. (2012). Virulence without catalysis: how can a pseudokinase affect host cell signaling? *Trends in Parasitology*, **28**(2), 53–57.

[63] Reese, M. L., Zeiner, G. M., Saeij, J. P. J., Boothroyd, J. C., and Boyle, J. P. (2011). Polymorphic family of injected pseudokinases is paramount in *Toxoplasma* virulence. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(23), 9625–9630.

[64] Reportlab Inc. (2010). The ReportLab PDF generation library.

[65] Rotella, D. P. (2012). Recent results in protein kinase inhibition for tropical diseases. *Bioorganic & Medicinal Chemistry Letters*, **22**(22), 6788–6793.

[66] Saeij, J. P. J., Boyle, J. P., Coller, S., Taylor, S., Sibley, L. D., Brooke-Powell, E. T., Ajioka, J. W., and Boothroyd, J. C. (2006). Polymorphic secreted kinases are key virulence factors in toxoplasmosis. *Science*, **314**(5806), 1780–1783.

[67] Saeij, J. P. J., Coller, S., Boyle, J. P., Jerome, M. E., White, M. W., and Boothroyd, J. C. (2007). *Toxoplasma* co-opts host gene expression by injection of a polymorphic kinase homologue. *Nature*, **445**(7125), 324–327.

[68] Scheeff, E. D. and Bourne, P. E. (2005). Structural evolution of the protein kinase-like superfamily. *PLoS Computational Biology*, **1**(5), e49.

[69] Scheeff, E. D., Eswaran, J., Bunkoczi, G., Knapp, S., and Manning, G. (2009). Structure of the pseudokinase VRK3 reveals a degraded catalytic site, a highly conserved kinase fold, and a putative regulatory binding site. *Structure*, **17**(1), 128–138.

[70] Schuster-Böckler, B., Schultz, J., and Rahmann, S. (2004). HMM Logos for visualization of protein families. *BMC Bioinformatics*, **5**, 7.

[71] Shi, F., Telesco, S. E., Liu, Y., Radhakrishnan, R., and Lemmon, M. A. (2010). ErbB3/HER3 intracellular domain is competent to bind ATP and catalyze autophosphorylation. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(17), 7692–7697.

[72] Sibley, L. D. (2011). Invasion and intracellular survival by protozoan parasites. *Immunological Reviews*, **240**(1), 72–91.

[73] Slater, G. S. C. and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

[74] Speer, C. A. and Dubey, J. P. (2001). Ultrastructure of schizonts and merozoites of *Sarcocystis neurona*. *Veterinary Parasitology*, **95**(2-4), 263–271.

[75] Steinfeldt, T., Könen-Waisman, S., Tong, L., Pawlowski, N., Lamkemeyer, T., Sibley, L. D., Hunn, J. P., and Howard, J. C. (2010). Phosphorylation of mouse immunity-related GTPase (IRG) resistance proteins is an evasion strategy for virulent *Toxoplasma gondii*. *PLoS Biology*, **8**(12), e1000576.

[76] Talevich, E., Mirza, A., and Kannan, N. (2011). Structural and evolutionary divergence of eukaryotic protein kinases in Apicomplexa. *BMC Evolutionary Biology*, **11**(1), 321.

[77] Talevich, E., Invergo, B. M., Cock, P. J. A., and Chapman, B. A. (2012). Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics*, **13**(1), 209.

[78] Taylor, S., Barragan, A., Su, C., Fux, B., Fentress, S. J., Tang, K., Beatty, W. L., Hajj, H. E., Jerome, M., Behnke, M. S., White, M., Wootton, J. C., and Sibley, L. D. (2006). A secreted serine-threonine kinase determines virulence in the eukaryotic pathogen *Toxoplasma gondii*. *Science*, **314**(5806), 1776–1780.

[79] Taylor, S. S., Knighton, D. R., Zheng, J., Sowadski, J. M., Gibbs, C. S., and Zoller, M. J. (1993). A template for the protein kinase family. *Trends in Biochemical Sciences*, **18**(3), 84–89.

[80] Treeck, M., Sanders, J. L., Elias, J. E., and Boothroyd, J. C. (2011). The phosphoproteomes of *Plasmodium falciparum* and *Toxoplasma gondii* reveal unusual adaptations within and beyond the parasites' boundaries. *Cell Host & Microbe*, **10**(4), 410–419.

[81] Wu, J., Yang, J., Kannan, N., Madhusudan, Xuong, N.-h., Ten Eyck, L. F., and Taylor, S. S. (2005). Crystal structure of the E230Q mutant of cAMP-dependent protein kinase reveals an unexpected apoenzyme conformation and an extended N-terminal A helix. *Protein Science*, **14**(11), 2871–2879.

[82] Xu, B., English, J. M., Wilsbacher, J. L., Stippec, S., Goldsmith, E. J., and Cobb, M. H. (2000). WNK1, a novel mammalian serine/threonine protein kinase lacking the catalytic lysine in subdomain II. *The Journal of Biological Chemistry*, **275**(22), 16795–16801.

[83] Yamamoto, M., Standley, D. M., Takashima, S., Saiga, H., Okuyama, M., Kayama, H., Kubo, E., Ito, H., Takaura, M., Matsuda, T., Soldati-Favre, D., and Takeda, K. (2009). A single polymorphic amino acid on *Toxoplasma gondii* kinase ROP16 determines the direct and strain-specific activation of Stat3. *The Journal of Experimental Medicine*, **206**(12), 2747–2760.

[84] Zeqiraj, E. and van Aalten, D. M. F. (2010). Pseudokinases-remnants of evolution or key allosteric regulators? *Current Opinion in Structural Biology*, **20**(6), 772–781.

[85] Zeqiraj, E., Filippi, B. M., Deak, M., Alessi, D. R., and van Aalten, D. M. F. (2009). Structure of the LKB1-STRAD-MO25 complex reveals an allosteric mechanism of kinase activation. *Science*, **326**(5960), 1707–1711.

[86] Zhang, Y. and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, **33**(7), 2302–2309.

Chapter 5

Discussion and concluding remarks

5.1 Achievement of goals

I have developed and applied novel approaches to address each of the research questions stated in the begining of this thesis. My original contributions in the analysis of protein families, in particular divergent, lineage-specific subfamilies of the protein kinase superfamily, enabled the identification and characterization of functionally important cases of structural and evolutionary adaptation in the protein kinases of eukaryotic pathogens. In each case I discussed the lineage-specific features that distinguish each subfamily in terms of sequence motifs, evolutionary context, structural impact and potential functional relevance.

5.1.1 Lineage-specific adaptations in apicomplexan kinomes

With colleagues, I perfomed an *in silico* characterization of the eukaryotic protein kinase superfamily in 15 apicomplexan species in order to identify the unique protein subfamilies and novel structural features that distinguish the kinases of apicomplexan parasites from those of their hosts [20]. This was the most comprehensive survey to date of the kinases in apicomplexan genomes. I used the available sequences to identify and classifify the kinome of each species and identify strongly conserved protein kinase families, as well as putative parasite-specific subfamilies within the characterized kinase families. By combining structural and evolutionary information, I identified several parasite-specific features that

could potentially serve as targets for inhibition, and identified other specific kinase families that warrant follow-up work using experimental and/or other bioinformatics methods.

The published article earned "highly accessed" status on the *BMC Evolutionary Biology* journal website, with over 3,500 readers. To make this information more broadly accessible, I uploaded annotations for over 700 apicomplexan protein kinase genes to each corresponding gene page on the eukaryotic pathogen database EuPathDB.

Recent studies by other groups have since obtained experimental support for some of our findings, including a phosphoproteomic survey [23] which observed phosphorylation of members of our proposed CDK and CDPK subfamilies at the sites we predicted, supporting our proposed mechanism. We have also performed detailed characterization and review of the *Plasmodium falciparum* 3D7 kinome [21], based substantially on this analysis.

5.1.2 Subfamily-level diversification of the rhoptry kinase family

My work on the rhoptry kinase family sheds light on several important but previously unrecognized features shared among rhoptry kinases, as well as the essential differences between active and degenerate protein kinases. My structural analysis revealed novel features which will be informative to future studies of the mechanisms of protein kinases and pseudokinases.

5.2 Significance & broader impact

The novel findings presented in these studies shed light on parasite phosphoryl signaling pathways and introduce novel bioinformatic methods to study patterns of diversification and neofunctionalization in protein families. This work also directly informs ongoing efforts to design protein kinase inhibitors for global diseases caused by apicomplexan parasites. The identification of unique features and molecular mechanisms conserved in pathogens but not their mammalian hosts could lead to potential diagnostic markers and candidates for targeted drug therapies for diseases caused by apicomplexans.

5.2.1 Insights into apicomplexan molecular biology

Alongside international collaborators, I am working to close gaps in the scientific knowledge of how these single-celled parasites function at the biochemical level, using integrative approaches in biology, chemistry, statistics and computation. Following up on this work, my advisor and I collaborated with Dr. Andrew Tobin at the (University of Leicester, UK) and Dr. Christian Doerig of (Monash University, Australia) to analyze recent experimental datasets related to the kinase interactions in the proteins of *P. falciparum*, and the effect of inhibition or deletion of specific kinases on these pathways. We have summarized the basis for this work in a focused review of our current knowledge of the kinases in the malaria [21]. My advisor and I were also invited by Dr. Diego Miranda-Saavedra of Osaka University (Japan) to co-author a book chapter focusing on computational approaches for studying the kinases in all of the targeted parasite species a collectively. This work will appear in an upcoming book on drug development approaches for kinase inhibitors in parasitic diseases (Talevich, Kannan and Miranda-Saavedra, in press).

5.2.2 Development of novel methods and computational tools

We have developed a set of sequence profiles which can be used to identify, classify and align the protein kinases in a given set of protein sequences. The development of specific profiles for novel ePK families allows us to identify kinases which would not be found by the more generic "protein kinase" profiles currently available in Pfam or the NCBI Conserved Domains Database (CDD), and have already proven useful for kinase identification and classification in diverse eukaryotic genomes.

As another product of these research efforts, I developed new software tools to assist in these analyses that combine structural and evolutionary analytical techniques. These methods were being developed with an additional interest in generalizability, so that they may be applied to protozoan clades other than Apicomplexa and Kinetoplastida.

During the course of this work I developed two stand-alone programs that can be used in further investigation of the evolution of protein kinase families, called Fammer and CladeCompare, in addition to the supporting libraries BioCMA (http://github.com/etal/biocma) and BioFrills (http://github.com/etal/biofrills), and a general-purpose phylogenetics mod-

ule distributed with the larger framework Biopython [22]. Fammer is a generalized method for large-scale annotation of protein families in sequence databases, including protein kinases in protozoan genomes. I used Fammer to construct a detailed classification scheme and profile database for the protein kinase superfamily, based on KinBase and a previously published survey of ePK-like microbial kinases found in oceanic metagenomics datasets [12]. This profile database consists of over 500 profiles and can be used to accurately identify and classify the kinases in a given proteome or protein sequence database. The program CladeCompare complements Fammer in that once a distinct protein subfamily has been identified and profiles constructed for the subfamily of interest, related subfamilies and the broader family, these profiles can be quickly compared using CladeCompare to identify the features that most uniquely characterize the subfamily of interest.

These methods are not specific to kinases and could be applied to other expanded protein families with a conserved domain or fold, particularly those for which representative structures have been solved. Nor are the research approaches based on these methods specific to parasites; since major protein superfamilies such as protein kinases are still unexplored for much of the tree of life, computational analysis of ePKs or other families in other clades using these methods is likely to be fruitful.

5.3 Future directions

The work discussed in this thesis suggests several avenues for future investigation of parasite signaling mechanisms, as well as the basic biology and evolution of parasitic protists.

5.3.1 Exploration of lineage-specific divergence in protein kinases

The novel methods for annotation and characterization of protein kinase subfamilies I have developed can be applied to other research questions, specifically to investigate the differences between protozoan parasites and hosts at the level of the whole kinome and, within the kinase domain, at the level of specific residues.

The Structural Genomics Consortium [9] has deposited many crystallographic structures in the Protein Data Bank (PDB) that await detailed structural and bioinformatic analysis.

Using the tools Fammer and CladeCompare, computational researchers can use the available data sets to efficiently explore and annotate these new structures after they are deposited. For example, the *Cryptosporidium parvum* protein (cgd4_240), which we classified as an unusual member of the glycogen synthase kinase 3 family (CMGC/GSK3) [20], was solved and deposited [PDBID:3EB0] but has not been described in a publication. The orthologous *P. falciparum* protein PfPK1 was previously discussed in absence of structural information or genomic context [13]. The available datasets are sufficient for this protein subfamily, which appears to be present in many apicomplexan genomes, to be computationally characterized using the tools described here. Below we consider several other research areas where a similar approach can be applied.

Divergence of apicomplexan MAPK cascade kinases and interacting partners.

In our preliminary analysis of apicomplexan kinomes, we noted that the MAPK cascade appears degenerate in apicomplexans. Specifically, no members of the STE group, the typical upstream regulators of MAPK, are conserved across multiple apicomplexan genera. In addition, the ERK1 subfamily of MAPK, which is highly conserved across Eukaryota, is missing from the *Plasmodium* and piroplasmid lineages. However, two other MAPK subfamilies are conserved in all of the surveyed apicomplexans: an ERK7 (Pfmap-1), whose upstream regulator is unknown, and a unique alveolate-specific MAPK (Pfmap-2) which is phosphorylated by a NEK kinase (Pfnek-1) [7].

We have another opportunity to build on our preliminary results by further investigation of the unusual adaptations involving MAPK signaling in apicomplexans. Residue analysis with CHAIN and CladeCompare can identify distinguishing features of apicomplexan ERK7 and the unique MAPK subfamily, as well as the ERK1 instances that we identified in coccidian species and *Cryptosporidium* spp.

The MAPK analysis is anticipated to yield residue-level characterizations of the divergence of the three MAPK subfamilies of interest in the Apicomplexa: ERK1, ERK7 and the unique MAPK subfamily related to Pfmap-2. We can place the identified sequence features in structural context based on solved structures and homology models.

In addition, we can use CHAIN and CladeCompare to provide an analysis of the co-

evolution of MAPK subfamily members and their proposed interacting partners, including Pfnek-1, the divergent apicomplexan STE kinase, and any other proposed members of the upstream MAPK cascade.

By identifying patterns of co-evolution in related sequences it may also be possible to identify potential upstream regulators of these MAPK instances.

Other apicomplexan-specific kinase families

Two other apicomplexan-specific kinase families have garnered recent interest: FIKK and PfPK7. FIKK, named after a shared Phe-Ile-Lys-Lys sequence motif in subdomain II of the protein kinase domain, is found in 1 copy in most apicomplexan genomes but expanded to 20 copies in *P. falciparum* [21, 25]. Members of this family appear to be exported to the host cell membrane, but little is known about their function [18].

Another conserved protein of interest is PfPK7. The initial investigation of the *P. falciparum* kinome found that this kinase is an orphan with an unclear relationship to other kinase families: the N-lobe of the kinase domain showed greatest similarity to AGC kinases, while the C-lobe instead matched STE kinases [8, 25]. My phylogenetic analysis and HMM profile search indicated that, of the characterized kinase families in KinBase, this protein is most closely related to Ca^{2+}/calmodulin-dependent protein kinase kinase (CaMKK). A detailed analysis using CHAIN and CladeCompare could identify the motifs that characterize PfPK7 and look for similar motifs in AGC, STE and CaMKK kinases to resolve this question.

Protein kinases in trypanosomatids and the Kinetoplastida

Trypanosomatids are responsible for a number of neglected tropical diseases. Three are most prominent: African trypanosomiasis ("sleeping sickness", caused by *Trypanosoma brucei*), Chagas disease (*Trypanosoma cruzi*), and leishmaniasis (*Leishmania* spp.). These diseases (African trypanosomiasis, Chagas disease and leishmaniasis), along with malaria, are the top targets for the Drugs for Neglected Diseases initiative (http://www.dndi.org). Trypanosomatids, though parasitic, are evolutionarily, biologically and biochemically distinct from the Apicomplexa and other protist phyla [15], and are classified within the taxonomic group Kinetoplastida, separated from Apicomplexa by an evolutionary distance of over 1

billion years [2]. As with the Apicomplexa, a systematic analysis of the protein kinases in these species and their evolutionary relatives would help address the biological and medical knowledge gaps and assist the search for effective treatments.

While the kinomes of three medically important trypanosomatids have been studied with phylogenetic methods [17], an expansion of this analysis can identify differences in related species and strains that may be functionally significant. In addition, structural analysis of trypanosomatid kinase families of interest can lead to more specific functional insights that could not be obtained from sequence data alone. Conducting this analysis in light of the results of our completed studies on apicomplexan protein kinases may reinforce our understanding of both the Apicomplexa and the Kinetoplastida, and potentially provide support for investigation of other protist clades as well.

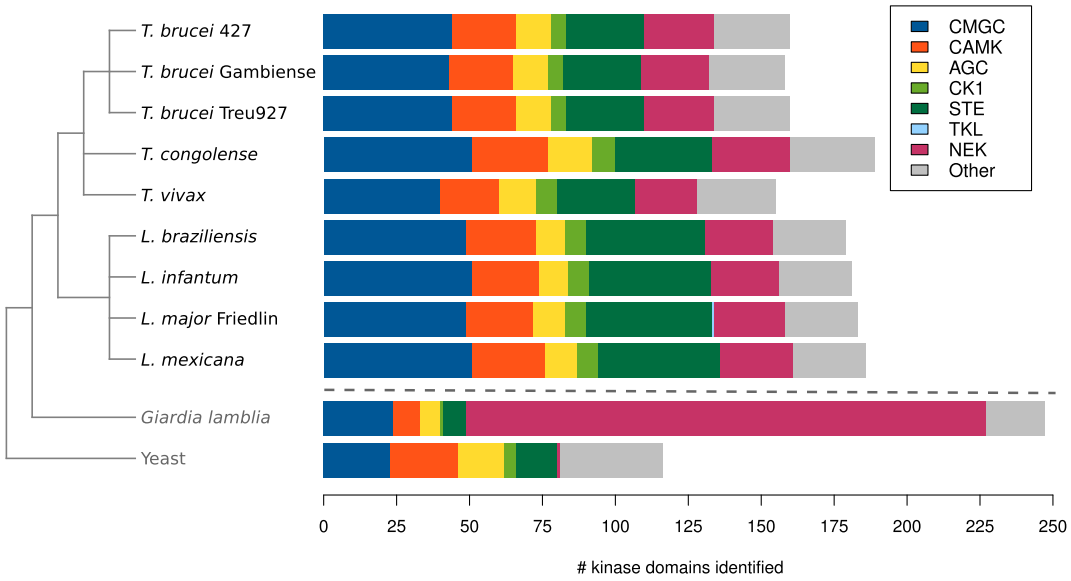


Figure 5.1: Composition of protein kinase major groups and the NEK family in each of the surveyed genomes. The outgroup kinomes of baker's yeast (*Saccharomyces cerevisiae*) and the parasitic euglenid *Giardia lamblia* are included for comparison.

Trypanosomatid kinome identification and classification: The kinomes of *Trypanosoma brucei*, *T. cruzi* and *Leishmania major* were previously examined with phylogenetic methods [17]. In contrast to the Apicomplexa, trypanosomatid kinomes do not show a dramatic overall reduction. Several ePK families were identified as dramatically expanded in these trypanosomatids: CMGC, STE11, and NEK. In the CMGC group, the CLK and MAPK (ERK1)

families are particularly expanded. The co-expansion of ERK1 and STE11 families may be significant because in other eukaryotes, ERK1 and STE11 are two interacting components of the MAPK signaling cascade. The expansion of CLK also correlates with the importance of RNA metabolism as a mode of post-transcriptional regulation in the Kinetoplastida [6]. The two major groups of receptor kinases, TK and TKL, are both missing in trypanosomatids; few putative transmembrane domains in protein kinases have been found, indicating that receptor kinases of any kind are rare in trypanosomatids [17]. There also appear to be smaller expansions of dual-specificity kinases such as DYRK and WEE in the trypanosomatids genomes, which may account for the observed tyrosine phosphorylation in trypanosomatid cells despite the lack of kinases in the TK group [17]. These findings were originally based on a survey of three trypanosomatid genomes, but we have replicated the results in a broader set of taxa (Figure 5.1).

Identify and classify the kinomes of multiple species in Kinetoplastida: During the analysis of apicomplexan kinomes, we developed a set of sequence profiles which can be used to identify, classify and align the protein kinases in a given set of protein sequences. This profile set can be applied to each of the 11 trypanosomatid genomes available, plus several outgroup genomes, to identify known subfamilies of kinase.

Identify lineage-specific ortholog groups and their distinguishing features in known ePK families: Phylogenetic analysis of each protein kinase gene family can identify divergent, lineage-specific ortholog groups in trypanosomatid kinomes. The divergent ortholog groups identified this way can also be compared to those in OrthoMCL-DB [4] to support or refine the initial findings. Comparison of these ortholog groups to the typical members of the kinase family can pinpoint specific sequence motifs that distinguish the divergent ortholog group, specifically by using CHAIN or CladeCompare to identify the sequence motifs that distinguish each divergent gene clade from the larger ePK family. This analysis is expected to provided guidance to characterize the lineage-specific adaptations in certain kinase families.

To address the issue of a shortage of sequence/taxon diversity for quantitative analysis within the Trypanosomatida or Kinetoplastida, this analysis can be expanded to include more outgroup species at different evolutionary distances from the Trypanosomatida, such

as the mosquito parasite *Crithidia fasciculata* within Trypanosomatida, the free-living *Bodo saltans* within Kinetoplastida, and the free-living *Euglena gracilis* within Excavata. In particular, the kinomes of the metamonad species *Giardia lamblia* and *Trichomonas vaginalis* have been annotated recently [16]. These more distant relatives may still be useful for comparison, and the high quality of these outgroup kinome annotations may also provide more support for annotations of the trypanosomatids of interest, by orthology.

Place distinguishing motifs in structural context to develop functional hypotheses: For gene clades where structures have been solved, the motifs discovered in the previous step can be mapped onto representative protein structures. Examination of the structures and relevant published literature can help develop hypotheses as to possible functions and functional differences related to these motifs. In gene clades where no representative crystal structure is available, a model of a protein kinase of interest can be constructed based on homologous structures.

Investigate novel trypanosomatid-specific ePK subfamilies: Our preliminary search identified several expanded ePK families, as well as some protein kinases which were not assigned to known families or subfamilies and may be divergent. These kinases in particular can be the focus of a search for the emergence of novel lineage-specific families, supported by orthology across multiple species, that may be functionally important in trypanosomatids. This study is expected to yield evidence of lineage-specific expansions and reductions of ePK families, a set of distinguishing sequence motifs for divergent kinase subfamilies, and several mechanistic or functional hypotheses related to the identified sequence patterns. Since several trypanosomatid-specific expansions of kinase families have been noted previously, but not carefully characterized, a deeper analysis of the structural and functional features of these expansions could be particularly valuable.

5.3.2 Structural analysis of kinase phosphorylation sites

The phosphorylation of certain residues in a protein kinase, particularly in the activation loop and substrate-binding region, has been observed to have an important functional effects on a kinase or substrate. High-throughput experimental methods such as mass spectrometry allow us to examine phosphorylation patterns across the proteome. The

phosphopeptide data produced by mass spectrometry experiments reveals the precise sequence locations of phosphorylation events on the proteins in a cell, and can also indicate the relative abundance of phosphorylated proteins. Subsequent computational analysis can identify differentially phosphorylated sites under different celluar conditions, and provide hints as to the functional effect of these post-translational modifications.

I have conducted preliminary analyses of the protein phosphorylation patterns in *P. falciparum*, in collaboration with Drs. Andrew Tobin (University of Leicester) and Christian Doerig (Monash University).

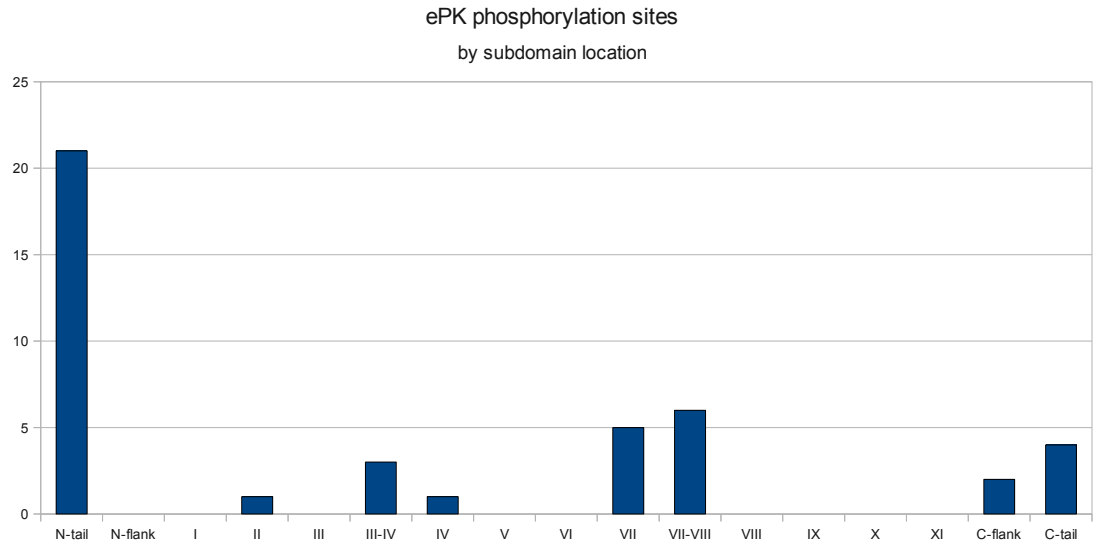


Figure 5.2: Subdomain location of observed *P. falciparum* phosphorylation sites within the kinase domain.

Using the data set originally published in [19], I selected the phosphorylation sites (phospho-sites) that occurred on protein kinases and used a sequence motif model to determine the kinase subdomain location of each phosphosite (Figure 5.2). This preliminary analysis revealed which conserved structural locations of the kinase domain are most frequently phosphorylated. Specifically, within the kinase domain, activation loop phospho-sites are the most common, followed by sites in the N-lobe surrounding (but not in) the αC helix. Outside the kinase domain, distant N-tail phospho-sites are by far the most common, followed by C-tail phospho-sites.

5.3.3 Prediction of kinase-substrate interactions and signaling cascades

A lingering question is whether differences at the sequence and structure level are reflected at the pathway level — that is, whether the signaling pathways of parasitic protozoans are distinct from those of other eukaryotes, including their metazoan hosts and related free-living protists. A phosphoproteomics-based analysis of the malaria parasite *P. falciparum* could yield important insights into parasite signaling machinery and phosphoregulation. Comparison of the predicted *P. falciparum* signaling pathways with those of humans and other eukaryotes could also indicate important differences in pathways.

Preliminary analysis of potential kinase-substrate interactions in *P. falciparum*

In a preliminary analysis of potential kinase-substrate interactions, I integrated the phosphopeptide data from [19] and [23] with yeast two-hybrid interactions in *P. falciparum*, as described in [14], kinase family-specific substrate-recognition consensus motifs, available from the Human Protein Reference Database (HPRD) [1], and the *P. falciparum* protein kinase family classifications I previously curated [21]. I found three putative kinase-substrate interactions supported by the following evidence:

1. The phosphorylation site in question was observed in two independently collected *P. falciparum* phospho-peptide data sets [19, 23].
2. A yeast two-hybrid interaction was observed between the proposed *Plasmodium* kinase and substrate proteins.
3. The sequence region surrounding the phosphorylation site matches the literature-supported substrate-recognition consensus motif for the proposed kinase's family, per HPRD, according to the current classification.
4. The available mRNA expression data indicate the proposed kinase and substrate are co-expressed.
5. An independent tool for predicting kinase-substrate interactions, NetPhosK [3], supports the prediction to at least some extent.

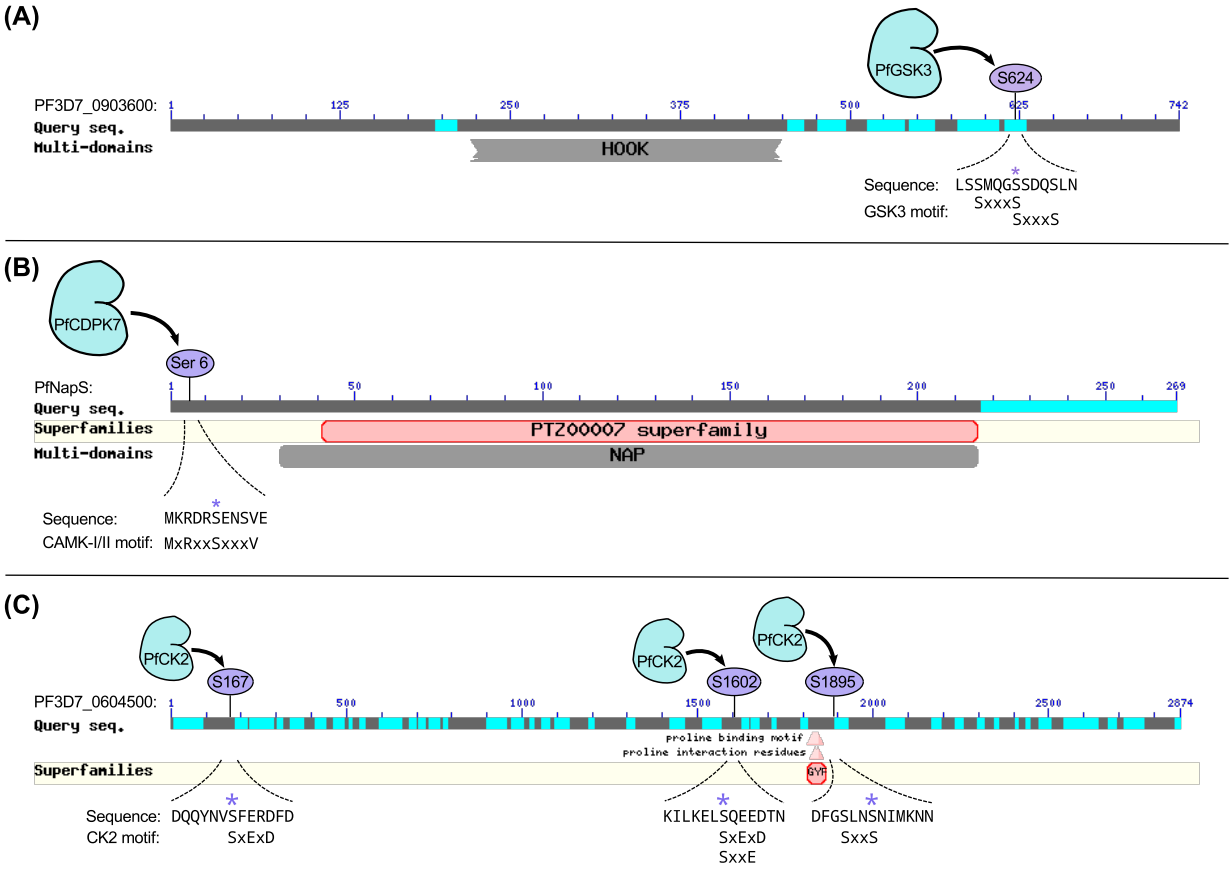


Figure 5.3: Three predicted kinase-substrate interactions in *P. falciparum*.
(A) Predicted phosphorylation of PfHOOK1 in a low-complexity C-terminal region by PfGSK3.
(B) Predicted phosphorylation of PfNapS at the C-terminus by PfCDPK7.
(C) Predicted phosphorylation of PfGYF at three sites by PfCK2.

I also attempted to assess the biological plausibility of these interactions, guided by GeneDB, PlasmoDB, homology, and any relevant publications.

1. *PfGSK3 to uncharacterized protein at Ser624* (Figure 5.3A): GSK3 is typically involved in control of microtubule assembly and stabilization [5]. The yeast 2-hybrid study shows interaction of PfGSK3 (PF3D7_0312400) with only one protein, the proposed substrate (PF3D7_0903600). Orthologs of this substrate are mostly specific to *Plasmodium*, with single hits in *T. gondii* and a sea anenome, according to OrthoMCL-DB.

The substrate gene is annotated as uncharacterized but appears to contain a HOOK domain (microtubule-binding), according to Pfam and CDD. GO terms for the substrate include function "actin binding" and component "myosin complex". These terms would be associated with the proposed microtubule assembly pathway. The matched phospho-site region is a low-complexity serine-rich region, C-terminal to the HOOK domain, matching the GSK3 recognition motif `SxxxS`. NetPhosK also predicts GSK3 as the most likely kinase to phosphorylate this site in the substrate protein.

2. *PfCDPK7 to PfNapS at Ser6* (Figure 5.3A): There are two nucleosome assembly proteins in the *P. falciparum* genome, NapS (PF3D7_0919000) and NapL. These perform non-redundant roles in the cell, distinguished by different localization (NapL to the cytoplasm, NapS primarily in or near the nucleus) and phosphorylation. A proposed mechanism is the chaperoning of histones from the cytoplasm to the nucleus, with a hand-off from NapL to NapS. It is plausible that phosphorylation of PfNapS at this site could influence its localization or binding to the nuclear membrane, as those are common effects of phosphorylation in other proteins. Both NAP proteins are phosphorylated by CKII; the predicted sites for this on NapS are Ser91, Thr190 and Thr191 (NetPhosK). However, CKII is not predicted to phosphorylate another observed site, Ser6. Instead, this region matches the CAMK consensus motif. NetPhosK predicts the three most likely kinases to phosphorylate this substrate as PKA, RSK and CAMK2. There is no CAMK2 homolog in apicomplexans; the most similar sequence to human CAMK1 and CAMK2 in *P. falciparum* is PfCDPK7. Yeast 2-hybrid interactions between PfCDPK7 and PfNapS are strongly supported (26 observations, 5 reproductions). The unusual domain architecture of PfCDPK7 (PF3D7_1123100), including a PH domain (associated with lipid binding and regulation), suggests non-canonical functions for this kinase relative to other CDPKs (such as those in plants). The PH domain in particular suggests a phospholipid membrane interaction. A crystal structure of PfNapS has been solved (PDB: 3KYP), but Ser6 was not included in the the crystalized protein. Still, one can infer that Ser6 would be at the end of the long alpha-helix.

3. *PfCK2 to uncharacterized protein at Ser167, Ser1602 and Ser1895* (Figure 5.3C): Casein kinase II is typically involved in many cell processes and phosphorylates a

variety of substrates. The proposed substrate (PF3D7_0604500) is a long (2874 aa), uncharacterized protein containing many low-complexity regions and a somewhat divergent GYF domain (named for a a Gly-Tyr-Phe motif), which is involved in ligand binding and recognition of proline-rich sequence regions. I'll call this protein by the bespoke name PfGYF here. Both PfCK2 (PF3D7_1108400) and PfGYF are very highly expressed throughout the intraerythrocyte stage. While CK2 is deeply conserved in Eukaryota, OrthoMCL-DB indicates orthologs of this substrate protein are specific to *Plasmodium*. The GO terms associated with PfGYF show it was electronically predicted to have nucleotide-binding function. PfGYF is highly phosphorylated. Three of these sites match the CKII consensus recognition motif: Ser167, Ser1602, Ser1895. (In general, CK2 phosphorylates acidic regions, or sites near other phospho-serines. The first two sites contain several acidic residues C-terminal to the phospho-site, while the third is near another serine.) NetPhosK predicts CK2 as the second-best match for Ser167 (just behind cdc2), the best match for Ser1602, and the sixth-best match for Ser1895 (behind PKC, CDC2, CAMK2, GSK3 and CK1). The latter two phospho-sites appear to be near the GYF domain.

Iterative homology search with jackhmmer found the human homolog GIGYF1, a member of the PERQ family which interacts with another binding partner (GRB10, SH2-domain-containing) to regulate TK receptor signaling, specifically insulin growth factor (IGF-1). No homologs of GRB10 exist in the Pf genome, though. However, different binding properties of PERQ family members have been noted between yeast and human homologs, indicating that this protein family is flexible/adaptable, and different binding partners may be present in *Plasmodium*. PfGYF shows yeast 2-hybrid interactions with 19 proteins, including itself. The GYF domain is noted to bind to the motif PPG[FILMV]. Searching the *P. falciparum* proteome for this motif matches 30 proteins; one of these also appears among the yeast 2-hybrid interactions for the GYF protein: splicing factor 3B subunit 4 (PF3D7_1420000). Alongside the GO annotation, this suggests a role for the complex in mRNA processing.

It is surprising that only one yeast 2-hybrid interaction was observed for PfCK2, given its many known roles and interacting partners. This is probably due to the transient nature

of kinase-substrate interactions, and is a limitation of the yeast 2-hybrid method for use in this analysis.

Kinase substrate-recognition motif identification and prediction

Phospho-sites based on several criteria. First, proteins can be grouped according to existing gene and Gene Ontology annotations. Proteins can also be grouped according to domain architecture, either by searching for specific domains of interest (such as the protein kinase domain), developing a custom system, or by using the mappings provided by Pfam2GO or Interpro2GO (http://www.geneontology.org/external2go/). I extracted a sequence set composed of the 7 residues flanking each phospho-site in the phosphoproteomic data set provided in [19]. At the sequence level, I have performed a BLAST-based clustering of the extracted sequence regions immediately surrounding each phospho-site, and found by examining the existing annotations of the source proteins that sequences with similar functional annotations often clustered together. Alternative sequence-based clusterings could be performed by constructing a gene tree from the phospho-site regions using phylogenetic methods, or by using the Gibbs Motif Sampler to identify recurring similar motifs in the set of phospho-site sequences.

Based on the clustering results obtained in the previous step, a position-specific scoring matrix (PSSM) can be constructed for each cluster. Each PSSM can then be used to search the full proteomic sequence set for similar regions. The similarity scores for these matches can be used to prioritize putative kinase-substrate interactions, in combination with other functional data.

This study would yield a set of predicted kinase-substrate interactions in the *P. falciparum* proteome. These predicted interactions, as well as the the observed phosphorylation sites and the surrounding sequence motifs, can be compared to known phosphorylation sites and interactions in model organisms such as yeast and human to refine the motif models used for predictions.

Prediction of signaling cascades

A network biology approach can be used to infer possible signaling pathways and use these predictions to compare selected pathways between *P. falciparum* and other model eukaryotes, including humans. This study would yield detailed predictions of the interactions constituting one or more phosphoryl signaling pathways, based on the predicted kinase-substrate interactions and other relevant data.

Reconstruct cellular phosphorylation pathways: The set of proposed interactions generated in the previous step could be used to construct proposed networks representing the signaling pathways of *P. falciparum*. Two alternate approaches can be used to construct these networks. In the simpler approach, the predicted kinase-substrate interactions are be linked together transitively to form a network. For example, if kinase a is predicted to phosphorylate kinase b, and b to phosphorylate protein c, then the pathway $a \rightarrow b \rightarrow c$ is predicted as a signaling cascade. A more sophisticated approach could use significance scores assigned to each possible kinase-substrate interaction, and construct a graph in which proteins are nodes and interactions are edges, weighted by the score for each interaction. Algorithms such as Markov Cluster [24] and max-flow can then be applied to delineate well-supported clusters or routes within the graph.

Compare predicted P. falciparum pathways to known pathways of model eukaryotes: We wish to identify signaling pathways in *P. falciparum* that may be substantially different from those in other eukaryotes, most importantly the human host. To accomplish this, specific signaling pathways predicted in the previous step, such as the MAPK cascade, are selected for comparison to the known pathways in model organisms. The phosphoproteome of yeast has been studied in detail [10, 26] and thus is a promising model for comparison. The interactions in the human proteome are also well-studied, and this data can be easily obtained from sources such as Reactome (http://www.reactome.org) [11]. The computational model of the predicted signaling networks can be refined iteratively, incorporating new functional and phosphoproteomic information as it becomes available. The phosphoproteomes of model organisms can also be used as a benchmark to test and validate the general prediction method described here.

Bibliography

[1] Amanchy, R., Periaswamy, B., Mathivanan, S., Reddy, R., Tattikota, S. G., and Pandey, A. (2007). A curated compendium of phosphorylation motifs. *Nature Biotechnology*, **25**(3), 285–286.

[2] Battacharya, D., Yoon, H. S., Hedges, S. B., and Hackett, J. D. (2009). Eukaryotes (Eukaryota). In S. Hedges and S. Kumar, editors, *The Timetree of Life*, chapter Eukaryotes, pages 116–120. Oxford University Press.

[3] Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., and Brunak, S. r. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, **4**(6), 1633–1649.

[4] Chung, Y. and Ané, C. (2011). Comparing two Bayesian methods for gene tree/species tree reconstruction: simulations with incomplete lineage sorting and horizontal gene transfer. *Systematic Biology*, **60**(3), 261–275.

[5] Daire, V. and Poüs, C. (2011). Kinesins and protein kinases: Key players in the regulation of microtubule dynamics and organization. *Archives of Biochemistry and Biophysics*, **510**(2), 83–92.

[6] De Gaudenzi, J., Frasch, A. C., and Clayton, C. (2005). RNA-binding domain proteins in Kinetoplastids: a comparative analysis. *Eukaryotic Cell*, **4**(12), 2106–2114.

[7] Dorin, D., Le Roch, K., Sallicandro, P., Alano, P., Parzy, D., Poullet, P., Meijer, L., and Doerig, C. (2001). Pfnek-1, a NIMA-related kinase from the human malaria parasite *Plasmodium falciparum*. *European Journal of Biochemistry*, **268**(9), 2600–2608.

[8] Dorin, D., Semblat, J.-P., Poullet, P., Alano, P., Goldring, J. P. D., Whittle, C., Patterson, S., Chakrabarti, D., and Doerig, C. (2005). PfPK7, an atypical MEK-related protein kinase, reflects the absence of classical three-component MAPK pathways in the human malaria parasite *Plasmodium falciparum*. *Molecular Microbiology*, **55**(1), 184–196.

[9] Gileadi, O., Knapp, S., Lee, W. H., Marsden, B. D., Müller, S., Niesen, F. H., Kavanagh, K. L., Ball, L. J., von Delft, F., Doyle, D. A., Oppermann, U. C. T., and Sundström, M. (2007). The scientific impact of the Structural Genomics Consortium: a protein family and ligand-centered approach to medically-relevant human proteins. *Journal of Structural and Functional Genomics*, **8**(2-3), 107–119.

[10] Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology*, **19**(3), 1720–1730.

[11] Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G. R., Wu, G. R., Matthews, L., Lewis, S., Birney, E., and Stein, L. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, **33**(Database issue), D428–D432.

[12] Kannan, N., Taylor, S. S., Zhai, Y., Venter, J. C., and Manning, G. (2007). Structural and functional diversity of the microbial kinome. *PLoS Biology*, **5**(3), e17.

[13] Kappes, B., Yang, J., Suetterlin, B. W., Rathgeb-Szabo, K., Lindt, M. J., and Franklin, R. M. (1995). A *Plasmodium falciparum* protein kinase with two unusually large kinase inserts. *Molecular and Biochemical Parasitology*, **72**(1-2), 163–178.

[14] LaCount, D. J., Vignali, M., Chettier, R., Phansalkar, A., Bell, R., Hesselberth, J. R., Schoenfeld, L. W., Ota, I., Sahasrabudhe, S., Kurschner, C., Fields, S., and Hughes, R. E. (2005). A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature*, **438**(7064), 103–107.

[15] Lawrence, J. G. (2005). Common themes in the genome strategies of pathogens. *Current Opinion in Genetics & Development*, **15**(6), 584–588.

[16] Manning, G., Reiner, D. S., Lauwaet, T., Dacre, M., Smith, A., Zhai, Y., Svard, S., and Gillin, F. D. (2011). The minimal kinome of Giardia lamblia illuminates early kinase evolution and unique parasite biology. *Genome Biology*, **12**(7), R66.

[17] Parsons, M., Worthey, E. A., Ward, P. N., and Mottram, J. C. (2005). Comparative analysis of the kinomes of three pathogenic trypanosomatids: *Leishmania major*, *Trypanosoma brucei* and *Trypanosoma cruzi*. *BMC Genomics*, **6**, 127.

[18] Schneider, A. G. and Mercereau-Puijalon, O. (2005). A new Apicomplexa-specific protein kinase family: multiple members in *Plasmodium falciparum*, all with an export signature. *BMC Genomics*, **6**(1), 30.

[19] Solyakov, L., Halbert, J., Alam, M. M., Semblat, J.-P., Dorin-Semblat, D., Reininger, L., Bottrill, A. R., Mistry, S., Abdi, A., Fennell, C., Holland, Z., Demarta, C., Bouza, Y., Sicard, A., Nivez, M.-P., Eschenlauer, S., Lama, T., Thomas, D. C., Sharma, P., Agarwal, S., Kern, S., Pradel, G., Graciotti, M., Tobin, A. B., and Doerig, C. (2011). Global kinomic and phospho-proteomic analyses of the human malaria parasite *Plasmodium falciparum*. *Nature Communications*, **2**, 565.

[20] Talevich, E., Mirza, A., and Kannan, N. (2011). Structural and evolutionary divergence of eukaryotic protein kinases in Apicomplexa. *BMC Evolutionary Biology*, **11**(1), 321.

[21] Talevich, E., Tobin, A. B., Kannan, N., and Doerig, C. (2012a). An evolutionary perspective on the kinome of malaria parasites. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **367**(1602), 2607–2618.

[22] Talevich, E., Invergo, B. M., Cock, P. J. A., and Chapman, B. A. (2012b). Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics*, **13**(1), 209.

[23] Treeck, M., Sanders, J. L., Elias, J. E., and Boothroyd, J. C. (2011). The phosphoproteomes of *Plasmodium falciparum* and *Toxoplasma gondii* reveal unusual adaptations within and beyond the parasites' boundaries. *Cell Host & Microbe*, **10**(4), 410–419.

[24] van Dongen, S. (2000). *A Cluster algorithm for graphs*. Thesis (phd), Amsterdam, Netherlands.

[25] Ward, P., Equinet, L., Packer, J., and Doerig, C. (2004). Protein kinases of the human malaria parasite *Plasmodium falciparum*: the kinome of a divergent eukaryote. *BMC Genomics*, **5**(1), 79.

[26] Wilson-Grady, J. T., Villén, J., and Gygi, S. P. (2008). Phosphoproteome analysis of fission yeast. *Journal of Proteome Research*, **7**(3), 1088–1097.