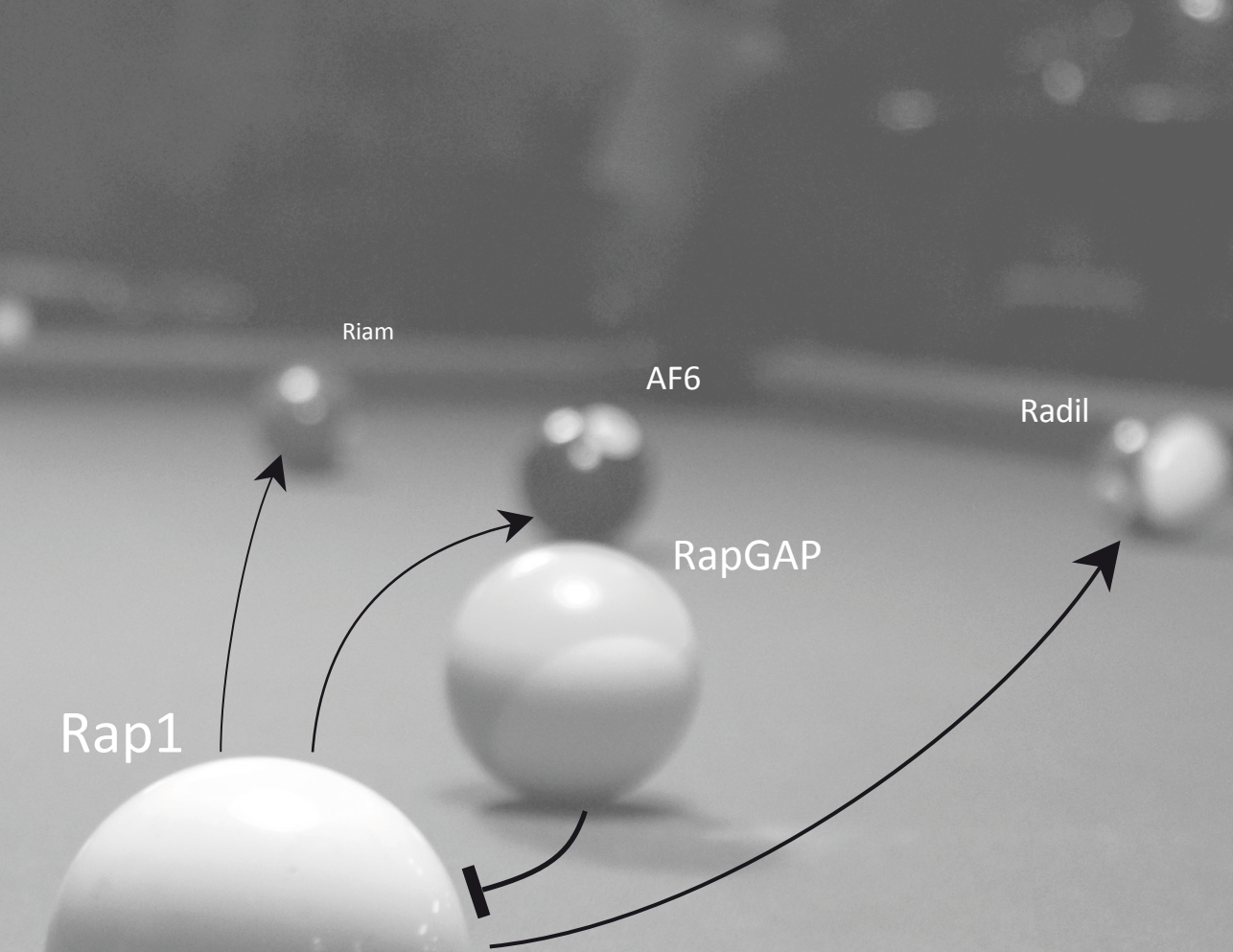


Evolution of Ras-like signaling pathways

Teunis Johannes Pieter van Dam

Reading committee:
Prof. dr. Boudewijn Burgering
Prof. dr. Mark Field
Prof. dr. Sander van den Heuvel
Prof. dr. Martijn Huynen
Dr. Jose Pereira Leal

Paranimfen:
Like Fokkens
Michael Seidl

ISBN/EAN 978-90-8891-244-3

Printed by: Proefschriftmaken.nl || Printyourthesis.com
Published by: Uitgeverij BOXPress, Oisterwijk

The printing of this thesis was financially supported by
The Netherlands Bioinformatics Centre (NBIC)

Evolution of Ras-like GTPase signaling pathways

Evolutie van de Ras-like GTPase signaalnetwerken

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus prof. dr. G.J. van der Zwaan, involge het besluit van het college voor promoties in het openbaar te verdedigen op woensdag 30 maart 2011 des middags te 4.15 uur

door

Teunis Johannes Pieter van Dam

geboren op 11 juni 1981, te Gouda

Promotor:

Prof. dr. J. L. Bos

Co-promotor:

Dr. B. Snel

Dit werk maakt deel uit van het Biorange programma van het Netherlands Bioinformatics Centre (NBIC), dat wordt ondersteund door een BSIK subsidie van het Netherlands Genomics Initiative (NGI).

Contents

CHAPTER 1

Introduction

1.1 General introduction

Signal transduction pathways are crucial for cells. Signaling pathways are networks of interacting proteins that measure and integrate internal and external stimuli and regulate cellular processes accordingly. In these pathways intricate feedback loops are often observed and, as a result, signaling pathways are very complex. Their roles in many diseases and syndromes as well as organismal development, make that signaling pathways are subject of much research. By fully elucidating signaling pathways the research community hopes to better understand the biological processes and the diseases in which disfunctional pathways participate.

Biochemistry and cellular- and molecular biology have produced a huge amount of knowledge on how these pathways operate and thereby the molecular basis of many related diseases. However, the question how a pathway is wired is often followed directly by the question of why the pathway is wired as it is. While the answer to the former question is mostly descriptive in nature, the latter question, once answered, will provide a full understanding of the operation and the mechanisms entailing the workings of these pathways.

Pathways did not appear in their entirety in a single moment in evolution. Instead it is thought that a simpler version of the pathway evolved in to a more complex pathway by subsequent addition and removal of components and interactions. In the course of evolution the pathway is fine tuned and adapted to changing circumstances (e.g. new inputs, new outputs). However, because the pathway needs to be functional in the process of evolution not all changes are as easy to make.

Therefore the question of why a pathway is wired as it is can be partly answered by investigating how it was wired before. By studying the evolution of the individual components of the pathway and their interactions we should in principle be able to

derive the evolution of the pathway. Not only will this give us some insights into how the pathway came to be, but it will also provide a framework onto which experimental data between model organisms can be compared and assessed.

In this thesis we describe bioinformatic and phylogenetic approaches to study the evolution of interaction networks and the complex Ras signaling pathways. We will first study the link between gene conservation and the conservation of interactions. In the subsequent chapters we study distinct protein families that function within the Ras-like signaling pathways as well as make an evolutionary reconstruction of an entire signaling pathway. The Ras-like GTPases belong to the rather large family of small GTPases, all of which share common ancestry. An evolutionary study of the many Ras-like GTPases will therefore greatly benefit our understanding of these pathways.

1.2 Evolution of interaction networks

The link between protein sequence and the function of the protein is strong [1-3]. The identification of similar sequences in other organisms and subsequent transfer of function to these identified sequences has been proven very powerful in molecular biology. However, many proteins carry out their function with the assistance of other proteins. The most intuitive examples for this are protein complexes. The multiple components of a protein complex are needed together to perform its tasks. It is therefore not surprising that we find a connection between protein-protein interactions (PPIs) and conservation on the genomic level. Proteins that are highly connected in an interaction network have been shown to be more strongly conserved on average on the gene level than less connected proteins [3]. Also highly connected proteins tend to evolve slower [2], i.e. their sequence changes less rapidly than on average.

By transferring protein function via sequence similarity we therefore also implicitly transfer knowledge on protein-protein interactions. This is, however, where some issues still remain. Even though the relationship between PPIs and gene conservation is evident, determining the conservation of the interactions themselves has proven difficult. The low coverage of PPI datasets between species has indicated that PPIs might not be as much conserved compared to the conservation of the interaction partners themselves [4]. A comparison of PPIs between yeast and *Caenorhabditis elegans* has resulted in an estimate of PPI conservation of 31% [5]. This may not be enough to faithfully transfer knowledge on PPIs between species based on sequence similarity. Similarly, a comparative genomics study into the composition of protein complexes revealed that complex composition is flexible in evolution as subunits are allowed to be lost or replaced while maintaining a functional protein complex [6]. A great example can be found in the TOR complex, described in Chapter 3. Summarizing, to study conservation and evolution of PPIs is crucial not only for our understanding of how pathways and protein complexes came to be, but also for the transferring of protein function.

1.3 Introduction into the Ras signaling pathways

1.3.1 The Ras-like GTPases

In 1964 the first Ras-like small GTPase was discovered as an oncogene encoded in the viral genomes of Harvey murine sarcoma virus, which caused rapid formation of

sarcomas in mice and rats [7]. Years later it was found that the viral Ras resembles a gene in the host [8] and this discovery jump-started a huge effort to understand the role of the Ras GTPase and its oncogenic properties in cancer (for a complete overview of the history of Ras research see Cox et al. [9]). We now know that in over 15% of all incidences of cancer Ras has been mutated [10].

In subsequent later years many other classes of small GTPases were discovered [9]. Arf, Ran, Rho, Rab, SR-beta, but also many other GTPases that are closely related to Ras, i.e. the Ras-like GTPases. The Ras-like GTPases are classified as such because of sequence similarity [11]. The canonical Ras GTPases (i.e. H-, N-, K-, M-, R-ras and TC21) regulate a wide variety of cellular process such as cell proliferation, cell division and cell differentiation. The Rap GTPases are mostly known for their regulatory function on cell adhesion, cell-cell contacts and junction formation [12]. Ral regulates vesicle transport and exocytosis via the exocyst complex [13,14]. Rheb is a critical regulator in the mTOR pathway regulating cellular and organismal growth [15]. RERG is involved in growth signaling [16]. Other Ras-like GTPases such as GEM, REM and RRAD are negative regulators of calcium currents through voltage gates Ca^{2+} channels [17,18]. The pathways in which Ras-like GTPases take part are diverse but mostly deal with external signals, such as growth factors, insulin, oxidative stress, mechanical stress and more.

Within the Bos lab, research is focused on the elucidation of the spatial-temporal control and molecular mechanisms of Rap1 on cell-cell junction formation, cell migration and adhesion. The goal of the Rap1 research is to understand the inhibition by Rap1 of oncogenic Ras mediated transformation of somatic cells to cancer cells.

1.3.2 Small GTPases are molecular switches

Small GTPases have two states depending on whether they have bound a GTP or GDP, and can therefore act as a switch [19]. In the GTP-bound state the GTPase is considered active as the additional phosphate induces a conformational change in the protein that allows the GTPase to bind other proteins that then propagate the signaling cascade. Subsequent deactivation occurs by hydrolysis of GTP into GDP thereby returning the protein to an inactive conformation [19].

The small GTPase superfamily is a large family of highly similar proteins that has expanded significantly during eukaryotic evolution [20]. In the human genome alone we can find well over 150 small GTPases. Most of these small GTPases take part in a large variety of signaling and regulatory pathways. The elucidation of their phylogenetic relationships could enable us to gain oversight and understanding of the many subfamilies, their functions and their regulation in the cell.

A possible reason that small GTPases are so found so frequently as regulatory switches in signaling pathways may be because of the versatility and modularity in regulation of small GTPases. This modularity is due to two intrinsic biochemical reasons. The first reason is that small GTPases are rather poor at their biochemical function (i.e. hydrolyzing GTP) [19]. Small misalignments or lack of critical catalytic residues prevents the small GTPases from being efficient GTP converting enzymes [21]. Therefore, the GTP conversion is exceptionally slow. Secondly, the ratio of GTP-bound vs. GDP-bound GTPases is mostly limited by the thermodynamic exchange of GDP by the more plentiful GTP in the cytosol. The affinity of the small GTPases for both GDP and GTP is high, therefore limiting the exchange of GDP for GTP [22,23]. The equilibrium of these two processes determines

the default state of the small GTPase.

The GDP for GTP exchange can be boosted by Guanine Exchange Factors (GEFs) [22,23] and the intrinsic GTPase activity can be boosted by GTPase Activating Proteins (GAPs) [24,25]. The GEFs can induce a conformational change in the small GTPase, temporarily lowering the binding affinity for GTP or GDP [22]. This allows for the bound nucleotide to be exchanged by the abundant GTP from the cytosol thereby activating the small GTPase. GAPs provide missing catalytic residues and/or induce a conformational change in the small GTPase aligning the catalytic residues, dramatically increasing the rate of conversion of GTP into GDP [24,25]. Note that the term GTPase Activating Protein indicates that these proteins activate the biochemical function of the GTPase but thereby inactivate its signaling function, i.e. GDP bound Ras does not signal. By regulating the GEFs and GAPs it is possible to tightly regulate the state of the small GTPases.

The small GTPase superfamily is a large family of structurally highly similar proteins [20]. The superfamily is subdivided based on sequence similarity, structural features, and cellular function [26,27]. To the superfamily of small GTPases belong the Arf, Ran, Rab, Rho, Ras, SRβ and SAR GTPases [9,11]. The Arf subfamily GTPases are known to regulate trafficking of proteins and vesicles [28,29]. Rab GTPases regulate many processes involved in phagocytosis and vesicle transport [29,30]. Rho is involved in cytoskeleton rearrangements [31], Ran is a regulator of nuclear import and export via the nucleopore complex [32] and Ras-like GTPases regulate cell division, movement, cell adhesion, apoptosis and many other diverse cellular processes [33]. Each of the small GTPase subfamilies has GEFs and GAPs that are non-homologous between the subfamilies [21,34].

While the function of Arf, Ran, Rab and Rho subfamilies can easily be summarized as regulating internal processes (e.g. actin and microtubule remodeling, vesicle transport), the Ras-like GTPases regulate a myriad of cellular processes that could be described as responses to outside stimuli. For instance, Rheb integrates a number of signals from the cell surroundings, such as insulin and nutrient availability, and promotes or inhibits growth accordingly [35-37]. RERG is upregulated upon estrogen stimulation [16]. Rap integrates outside signals and mechanical stress and promotes cell adhesion, cell-cell junction formation and migration [12,38]. The above examples clearly illustrate that Ras-like GTPases regulate very different cellular processes.

1.3.3 Flicking the switch

Above I discussed the need for Guanine Exchange Factors and GTPase Activating Proteins in small GTPase regulation. For the Ras-like GTPases the prominent GEFs and GAPs harbor one of three regulatory domains: the RasGEF domain (also known as the CDC25 homology domain), the RasGAP domain or the RapGAP domain. GEFs harboring the RasGEF domain regulate Ras, Rap and Ral GTPases. GAPs harboring the RasGAP domain (RasGAPs) regulate the canonical Ras GTPases, but one class of RasGAPs also regulates Rap GTPases (for more information see Chapter 5). Lastly, the GAPs harboring the RapGAP domain regulate Rap (Sipa, Rap1GAP), Rheb (TSC2) or Ral (RalGAPA/B complex) [39,40,36,41].

A key feature of the RasGEF, RasGAP and RapGAP domain containing proteins is the large variety of protein domain architectures (see Figure 1.1). The additional domains regulate the activity and localization of the GEF or GAP. For instance, the cAMP binding

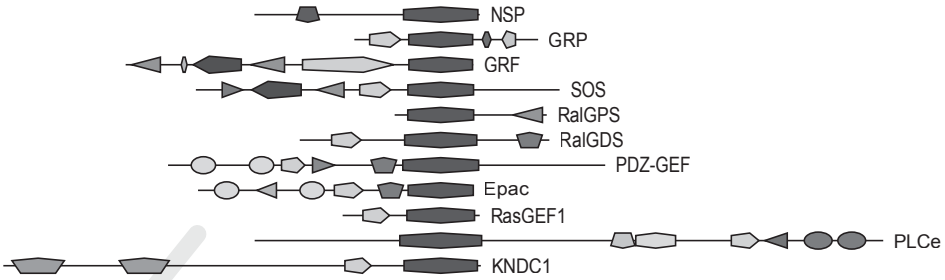


Figure 1.1. Examples of domain architectures for the RasGEF protein family. Depicted are human RasGEF domain containing proteins. Note the diverse set of additional domains and reoccurrence of specific domains at different positions such as the RA and PH domains.

domain in Epac1 induces a conformational change in the whole protein upon binding of cAMP, activating Epac by freeing the catalytic site of the CDC25HD [42]. The DEP domain is crucial for localization of Epac1 to the plasma membrane [43]. It is postulated that temporal-spatial regulation of GEF and GAP activity, and therefore GTPase activity, is critical for the signaling function of the Ras-like GTPases [21]. Interestingly, the known GEF and GAP domains only regulate part of the Ras-like GTPases, namely the Ras, Rap, Ral and Rheb GTPases. To our knowledge, no GEF or GAP has been reported for other members of the Ras-like GTPase subfamily, such as Rit, RERG, GEM and DiRas.

Upon activation of the Ras-like GTPases the switch regions I and II change conformation [44]. This conformational change allows for downstream effectors to bind and propagate the signal or effectuate the response.

1.4 Evolution of the small GTPase superfamily

The small GTPase superfamily is large and their phylogenetic relations are complicated due to the many duplications that occurred during eukaryotic evolution. Additionally the changes in domain composition of the regulatory proteins we observe in the RasGAPs and RasGEFs increases the complexity of Ras signaling greatly. Therefore, by studying the evolution of the Ras-like GTPases and their regulators and pathways we hope to gain answers to how, when and why this complexity arose.

With the publication of more and more genomes of eukaryotic organisms from all major phyla, it becomes increasingly more interesting to compare these genomes. The available genomes now allow us to time duplications to specific points in evolution and show the relationships in time between the GTPases and their regulators and pathways. By studying the similarities and differences between Ras pathway components in several species we hope to identify common themes or find how conserved certain interactions within Ras pathway are. Additionally, these studies can help in selecting the correct model organism for experiments and prevent wrong assumptions being made for the transfer of functional data between species. They can also help by providing a framework on which a complete overview of the elaborate Ras pathways or large protein families can be projected.

Many of the small GTPase subfamilies are involved in the regulation of hallmark traits of eukaryotes. Jekeley [45] and Yutin et al. [46] described antagonizing scenarios for the evolution of the ancestral eukaryote (the ancestral unicellular organism that bears

hallmarks typical for eukaryotes and from which all extant eukaryotes stem) based amongst others on the phylogeny of the complete small GTPase superfamily.

In the evolution of the eukaryote (eukaryogenesis, the emergence of eukaryotic features in the eukaryote ancestor), the process of phagocytosis is deemed a crucial invention. According to the archezoan hypothesis [47], phagocytosis allowed for the envelopment of the mitochondrion and the chloroplast by the proto-eukaryote. In the "fateful encounter" theory, phagocytosis is considered the process by which the intracellular membranes were obtained by the eukaryotic ancestor.

Jekely describes a detailed scenario of the acquisition of the eukaryotic hallmark traits by the eukaryote ancestor, based on an evolutionary reconstruction of the order of appearance of the small GTPase subfamilies. He argues that phagocytosis is not the first major novel trait that separated eukaryotes from prokaryotes in terms of cellular structure, but is most likely one of the last gained. By studying the order of appearance of the small GTPases and thereby ordering the emergence the cellular processes they regulate, Jekeley comes to the conclusion that intracellular membranes, exocytosis, the nuclear membrane, the cytoskeleton and vesicle trafficking predates phagocytosis.

Yutin et al. argue for the "fateful encounter" scenario based on an evolutionary reconstruction of the phagoproteome. Two families of proteins involved in phagocytosis that are conserved in bacteria are the actin and the small GTPase superfamilies. Yutin et al. identify additional bacterial homologs of the small GTPase superfamily. These additional homologs divide the superfamily of small GTPases into two super groups, one containing the SAR1, SRβ and ARF GTPases and one containing the Ran, Rab, Rho and Ras-like GTPases. This division within the small GTPase superfamily is confirmed by Dong et al. [48]. Yutin et al. argue that the ancestral Ran-Rab-Ras-Rho GTPase has been acquired via horizontal gene transfer suggesting the "fateful encounter" model for eukaryogenesis.

The studies into the evolution of eukaryotes by Jekely and Yutin et al. stress the evolutionary importance of the small GTPases. Brighouse et al. reconstruct the evolution of intracellular membranes and vesicle transport by analyzing the phylogeny of the Rab subfamily of GTPases [49]. Pereira-Leal and Seabra show that Rab GTPase orthologs maintained their cellular function in evolution similar to the distinct segregation of function between the Ras superfamily members [50]. Pereira-Leal describes in detail the evolutionary dynamics of the Rab GTPases fungi [51]. He shows that the Rab family of small GTPases in fungi are kept small compared to animals and plants while maintaining the spectrum of Rab dependend processes, suggesting a different route to multi-cellularity [51]. Phylogenetic studies have been performed for other small GTPase subfamilies as well, such as Rho [52] and also Ras [11,53]. However, these studies focus mostly on making classifications and/or distribution of GTPases in species.

By studying the evolution of the Ras signaling pathways, and more particularly the evolution of the Ras-like GTPases themselves, we hope to gain more insight into how these networks are wired by investigating how they were wired. The time of invention and duplications of the GTPases, their effectors and their regulatory proteins should give us insight into when the observed complexity arose and which entities have had a significant role in this.

1.5 Phylogenomics and bioinformatics

This thesis covers topics that are on the cusp of comparative genomics and molecular biology. In this section some methods and principles will be discussed in the hope that it will allow a reader without extensive background knowledge in comparative genomics to better appreciate the work herein.

Comparative genomics is a field in which we try to use information provided by signatures in genomic sequences left by evolution. By exploiting similarities and differences on the sequence level between species or strains we are able to discern functional elements and how they change in evolution. Common descent with modification, as described in the theory of evolution, is reflected in genomic sequences. By comparing genomic sequences between species, modifications become traceable. Key in these comparisons is the determination of common ancestry of the elements studied. Elements sharing common ancestry are called homologous. Homology is reflected in sequence similarity between genes and is an important concept in comparative genomics.

1.5.1 Sequence similarity and homology

The determination and quantification of sequence similarity is a critical step in any comparative genomics study. Many methods have been developed to compare sequences, but one of the best known and used methods is the Basic Local Alignment Search Tool (BLAST). Using heuristics and a generic model describing preferred substitutions in sequences, BLAST is able to rapidly search large databases of sequences for similar sequences. However BLAST often fails to identify highly divergent sequences. Methods that use sequence specific models such as Position Specific Iterative BLAST (PSI-BLAST) or Hidden Markov Model methods (such as HMMER) are better suited as they use information from a set of predetermined homologous sequences to detect divergent sequences instead of relying on a generic model of sequence evolution.

Although these methods are very powerful and provide useful metrics for sequence similarity like e-values, they do not automatically infer homology (e.g. that these sequences share common ancestry). Homology is determined best by manual sequence comparisons rather than using predetermined cut-off values for sequence similarity searches.

Sequences are not always found to share homology over the full length of its sequence. Instead, conserved regions can often be identified which can be found in many sequences in different compositions (e.g. protein domains). This implies that genes do not necessarily hold a linear line of decent as domains each represent distinct evolutionary units. In this thesis we will encounter these types of sequences in Chapters 3 and 5 and describe how to deal with them in phylogenetic analyses.

1.5.2 Multiple Sequence Alignments

Identified similar sequences that are homologous according to criteria[1] can be compared to each other by making multiple sequence alignments (MSA). The previously discussed programs and methods only provide pair-wise comparisons. The goal of MSA is to

1 Note that homology is a qualitative term and not a quantitative term. Two sequences are either homologous (i.e. share common ancestry) or they are not. As sequence similarity often implies homology, homology is often used in the context of sequence similarity. Sequence similarity is a quantitative term to describe to what extent two sequences look alike. However sequence similarity does not always imply common ancestry.

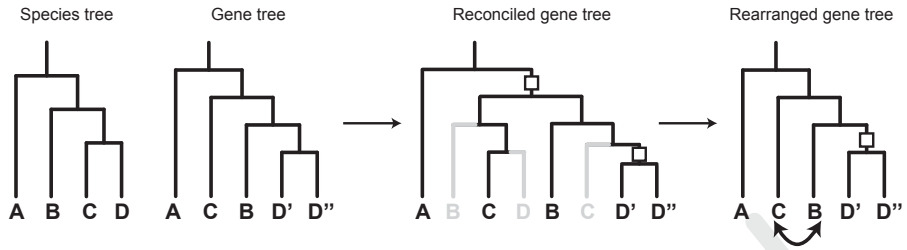


Figure 1.2. Reconciling a gene tree with a species tree. The letters represent species or "homolog of gene in species". Squares represent duplications and gray lines indicate gene loss. The placement of genes from species C and B in the original gene tree results in the reconstruction of multiple gene losses and additional duplication. By flipping B and C the topology of the resulting reconciled tree is greatly simplified.

organize the sequences in such a way that homologous characters are identified and can be compared (i.e. bases or amino acid residues that represent the same residues or positions in other sequences, such as a specific conserved residue in the catalytic pocket of a protein crucial for catalysis). Because we can mostly consider sequences of extant species and because of the necessity for heuristics for calculating MSAs[2], errors are introduced into the alignment where phylogenetic characters are misaligned. Misalignment results in loss of phylogenetic signal and the introduction of noise. The quality of the MSA is crucial for the final phylogenetic inference and often needs manual editing. New programs, such as MAFFT, are constantly being developed to cope with the increasing amount of sequences while improving or maintaining the level of accuracy.

1.5.3 Phylogenetic inference

The MSA is used as the information for phylogenetic inference, i.e. determining the evolutionary relationships between all sequences considered. There are several methods available that differ mainly in the underlying model of sequence evolution used to calculate the final phylogenetic gene tree [54]. Because evolution is not clock-like, trees only infer relationships and no root is implied. Phylogenetic trees need to be rooted in order to correctly represent the phylogenetic relationships. Trees can be rooted in a number of ways: by midpoint rooting, by a priori including an outgroup (e.g. a sequence for which it is known that it is the most diverged sequence in the set) or by choosing the root which will reduce the number of loss and duplication events. Midpoint rooting does not guarantee any evolutionary or biologically relevant rooting. The other two methods rely on a reconciliation of the gene tree with the tree of life, a predetermined model of evolution which describes the order of divergence of all species represented in the gene tree. A figure depicting the tree of life as used in this thesis can be found in appendix A.1. Under the assumption that the evolution of a single gene will show a similar pattern as the evolution of the species we can determine the outgroup as the gene of the most distant species and root accordingly. In case a clear outgroup is not available (for instance when multiple homologous genes are present in each species) we need to make a full reconciliation of the gene tree with the species tree (see Figure 1.2).

2 Calculating MSAs using exhaustive methods is an NP-complete problem (ref) which basically means that there is no algorithm that calculates a solution in a reasonable amount of time. Computation time increases exponentially with sequence length or the number of sequences. Exhaustive methods therefore become computationally prohibitive very quickly and heuristics must be applied.

1.5.4 Phylogenetic tree reconciliation and evolutionary reconstruction

Once rooted, we can compare the gene tree to the species tree and identify and time duplications and gene loss (see Figure 1.2 – reconciled tree). By reconciling the gene tree with the species tree we come to a full evolutionary reconstruction based on our data.

The gene tree rarely exactly fits the species tree since MSAs can be noisy and true phylogenetic relationships can be obscured. We therefore need to correct the reconciled tree. The criteria for rearranging a gene tree are based on parsimony, i.e. we try to optimize the gene tree in such a way that the number of duplications and gene losses are minimized (Figure 1.2, compare reconciled tree to rearranged tree). Often we will encounter tree topologies that have no unique parsimonious solution and the final rearrangement (and thereby the evolutionary reconstruction) are subjective or biased but most notably, can be incorrect. However there is no way to determine incorrect from incorrect rearrangements other than correlating the final rearrangement to additional data or adding more sequences. The last option does not necessarily solve the problem of multiple possible arrangements or eliminate bias. The final rearranged and rooted tree represents the evolutionary reconstruction. From the evolutionary reconstruction orthologous (genes separated by speciation) and paralogous (genes separated by duplication) relationships can be derived.

This could be the endpoint but the power in phylogenetic and comparative analysis lies in correlating the evolutionary reconstruction to functional data. Mapping functional aspects such as substrate specificity, or cellular localization can form the basis for an analysis of the molecular mechanism of a functional aspect that was previously not apparent in the original gene of interest. By comparing evolutionary reconstructions of multiple genes of which the proteins participate in a protein complex or pathway, we obtain an evolutionary profile of the entire pathway or complex. An example of how powerful such an approach can be is shown by Li et al. [55]. They successfully applied phylogenetic profiles for the cilium proteome to search for genes with similar phylogenetic profiles. This method successfully identified many new genes involved in the cilium.

1.6 Outline of this thesis

In this thesis we use a comparative genomics approach to investigate the evolution of interaction networks and signaling pathways. In particular we focus on the Ras-like GTPases and their signaling networks in eukaryotes.

Understanding the evolution of interaction networks, and even more importantly, the conservation of individual protein-protein interactions is crucial for the elucidation of pathway evolution by comparative genomics. Previously, low estimates were given for the conservation of interactions between proteins, but no real number was given, complicating reconstruction of pathways in other organisms. To safely transfer functional knowledge on protein interaction networks and (signaling) pathways to other eukaryotic species based on gene orthology, we first needed to know how well interactions between proteins are conserved. Since there is no experimental evidence for most organisms we cannot assume that components are still involved in the same pathway. The publication of two large TAP-MS PPI datasets for yeast allowed us to compare the protein-protein interaction networks between human and yeast in **Chapter 2** and determine how well

protein-protein interactions are conserved.

As a first step into researching the evolution of Ras-like GTPase pathways we investigate the evolution of the RasGEF domain (also known as the CDC25 homology domain or CDC25HD) in **Chapter 3**. RasGEFs are ideally suited to investigate evolutionary dynamics of a GTPase regulatory protein family and the role of protein domain acquisitions therein, as RasGEFs contain a myriad of protein domains but still share common ancestry via CDC25HD.

In **Chapter 4** we describe the evolution of an entire signaling pathway. We investigate the eukaryotic TOR pathway which contains the Rheb GTPase. Components of the TOR pathway are a highly conserved and the pathway is well researched and therefore provides a perfect opportunity to study the evolution of a Ras-like GTPase containing pathway in detail.

Finally in **Chapter 5** we investigate the complex phylogeny of the Ras-like subfamily of small GTPases. We combine the Ras-like GTPase phylogeny and the detailed evolutionary reconstructions of the regulatory RasGEF (Chapter 3), RapGAP and RasGAP domains (Chapter 5), to explore the early divergence between the different Ras-like GTPase subtypes and give an answer as to how and when they diverged.

1.7 References

1. Pereira-Leal JB, Levy ED, Teichmann SA (2006) The origins and evolution of functional modules: lessons from protein complexes. 361: 507-517.
2. Fraser HB, Wall DP, Hirsh AE (2003) A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC evolutionary biology 3: 11. doi:10.1186/1471-2148-3-11
3. Wuchty S, Oltvai ZN, Barabási A-L (2003) Evolutionary conservation of motif constituents in the yeast protein interaction network. Nature genetics 35: 176-9. doi:10.1038/ng1242
4. Suthram S, Sittler T, Ideker T (2005) The Plasmodium protein network diverges from those of other eukaryotes. Nature 438: 108-12. doi:10.1038/nature04135
5. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, et al. (2001) Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or “interologs”. Genome research 11: 2120-6. doi:10.1101/gr.205301
6. Fokkens L, Snel B (2009) Cohesive versus flexible evolution of functional modules in eukaryotes. PLoS computational biology 5: e1000276. doi:10.1371/journal.pcbi.1000276
7. Malumbres M, Barbacid M (2003) RAS oncogenes: the first 30 years. Nature reviews. Cancer 3: 459-65. doi:10.1038/nrc1097
8. Stehelin D, Varmus HE, Bishop JM, Vogt PK (1976) DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. Nature 260: 170-3.
9. Cox AD, Der CJ (2010) Ras history. Small GTPases 1: 2-27. doi:10.4161/sgtp.1.1.12178
10. Bos JL (1989) ras oncogenes in human cancer: a review. Cancer Res 49: 4682-4689.
11. Wennerberg K, Rossman KL, Der CJ (2005) The Ras superfamily at a glance. Journal of cell science 118: 843-6. doi:10.1242/jcs.01660
12. Bos JL (2005) Linking Rap to cell adhesion. Current opinion in cell biology 17: 123-8. doi:10.1016/j.ceb.2005.02.009
13. Moskalenko S, Henry DO, Rosse C, Mirey G, Camonis JH, et al. (2002) The exocyst is a Ral effector complex. Nature cell biology 4: 66-72. doi:10.1038/ncb728
14. Sugihara K, Asano S, Tanaka K, Iwamatsu A, Okawa K, et al. (2002) The exocyst complex binds the small GTPase RalA to mediate filopodia formation. Nature cell biology 4: 73-8. doi:10.1038/ncb720
15. Aspuria P-J, Tamanoi F (2004) The Rheb family of GTP-binding proteins. Cellular signalling 16: 1105-12. doi:10.1016/j.cellsig.2004.03.019
16. Finlin BS, Gau CL, Murphy GA, Shao H, Kimel T, et al. (2001) RERG is a novel ras-related, estrogen-regulated and growth-inhibitory gene in breast cancer. The Journal of biological chemistry 276: 42259-67. doi:10.1074/jbc.M105888200
17. Béguin P, Nagashima K, Gonoi T, Shibasaki T, Takahashi K, et al. (2001) Regulation of Ca2+ channel expression at the cell surface by the small G-protein kir/Gem. Nature 411: 701-6. doi:10.1038/35079621
18. Finlin BS, Crump SM, Satin J, Andres DA (2003) Regulation of voltage-gated calcium channel activity by the Rem and Rad

GTPases. Proceedings of the National Academy of Sciences of the United States of America 100: 14469-74. doi:10.1073/pnas.2437756100

19. Gibbs JB (1984) Intrinsic GTPase Activity Distinguishes Normal and Oncogenic ras p21 Molecules. Proceedings of the National Academy of Sciences 81: 5704-5708. doi:10.1073/pnas.81.18.5704
20. Bourne HR, Sanders DA, McCormick F (1990) The GTPase superfamily: a conserved switch for diverse cell functions. Nature 348: 125-32. doi:10.1038/348125a0
21. Bos JL, Rehmann H, Wittinghofer A (2007) GEFs and GAPs: critical elements in the control of small G proteins. Cell 129: 865-77. doi:10.1016/j.cell.2007.05.018
22. Klebe C, Prinz H, Wittinghofer A, Goody RS (1995) The kinetic mechanism of Ran--nucleotide exchange catalyzed by RCC1. Biochemistry 34: 12543-52.
23. Lenzen C, Cool RH, Prinz H, Kuhlmann J, Wittinghofer A (1998) Kinetic analysis by fluorescence of the interaction between Ras and the catalytic domain of the guanine nucleotide exchange factor Cdc25Mm. Biochemistry 37: 7420-30. doi:10.1021/bi972621j
24. Scheffzek K (1997) The Ras-RasGAP Complex: Structural Basis for GTPase Activation and Its Loss in Oncogenic Ras Mutants. Science 277: 333-338. doi:10.1126/science.277.5324.333
25. Rittinger K, Walker PA, Eccleston JF, Smerdon SJ, Gamblin SJ (1997) Structure at 1.65 A of RhoA and its GTPase-activating protein in complex with a transition-state analogue. Nature 389: 758-62. doi:10.1038/39651
26. Drivas GT, Palmieri S, D'Eustachio P, Rush MG (1991) Evolutionary grouping of the RAS-protein family. Biochemical and biophysical research communications 176: 1130-5. doi:10.1016/0006-291X(91)90402-S
27. Valencia A, Chardin P, Wittinghofer A, Sander C (1991) The ras protein family: evolutionary tree and role of conserved amino acids. Biochemistry 30: 4637-48.
28. Boman A, Kahn R (1995) Arf proteins: the membrane traffic police? Trends in Biochemical Sciences 20: 147-150. doi:10.1016/S0968-0004(00)88991-4
29. Chavrier P, Goud B (1999) The role of ARF and Rab GTPases in membrane transport. Current opinion in cell biology 11: 466-75. doi:10.1016/S0955-0674(99)80067-2
30. Novick P (1993) Friends and family: The role of the rab GTPases in vesicular traffic. Cell 75: 597-601. doi:10.1016/0092-8674(93)90478-9
31. Spiering D, Hodgson L (2011) Dynamics of the Rho-family small GTPases in actin regulation and motility. Cell adhesion & migration 5: 22-21.
32. Pemberton LF, Paschal BM (2005) Mechanisms of receptor-mediated nuclear import and nuclear export. Traffic (Copenhagen, Denmark) 6: 187-98. doi:10.1111/j.1600-0854.2005.00270.x
33. Karnoub AE, Weinberg RA (2008) Ras oncogenes: split personalities. Nature reviews. Molecular cell biology 9: 517-31. doi:10.1038/nrm2438
34. Jiang S-Y, Ramachandran S (2006) Comparative and evolutionary analysis of genes encoding small GTPases and their activating proteins in eukaryotic genomes. Physiological genomics 24: 235-51. doi:10.1152/physiolgenomics.00210.2005
35. Garami A, Zwartkruis FJT, Nobukuni T, Joaquin M, Roccio M, et al. (2003) Insulin Activation of Rheb, a Mediator of mTOR/S6K/4E-BP Signaling, Is Inhibited by TSC1 and 2. Molecular Cell 11: 1457-1466. doi:10.1016/S1097-2765(03)00220-X
36. Inoki K, Li Y, Xu T, Guan K-L (2003) Rheb GTPase is a direct target of TSC2 GAP activity and regulates mTOR signaling. Genes & development 17: 1829-34. doi:10.1101/gad.1110003
37. Avruch J, Long X, Ortiz-Vega S, Rapley J, Papageorgiou A, et al. (2009) Amino acid regulation of TOR complex 1. American journal of physiology. Endocrinology and metabolism 296: E592-602. doi:10.1152/ajpendo.90645.2008
38. Pannekoek W-J, Kooistra MRH, Zwartkruis FJT, Bos JL (2009) Cell-cell junction formation: the role of Rap1 and Rap1 guanine nucleotide exchange factors. Biochimica et biophysica acta 1788: 790-6. doi:10.1016/j.bbamem.2008.12.010
39. Maric D, Epting CL, Engman DM (2010) Composition and sensory function of the trypanosome flagellar membrane. Current opinion in microbiology. doi:10.1016/j.mib.2010.06.001
40. Gridley S, Chavez JA, Lane WS, Lienhard GE (2006) Adipocytes contain a novel complex similar to the tuberous sclerosis complex. Cellular signalling 18: 1626-32. doi:10.1016/j.cellsig.2006.01.002
41. Zhang Y, Gao X, Saucedo LJ, Ru B, Edgar BA, et al. (2003) Rheb is a direct target of the tuberous sclerosis tumour suppressor proteins. Nature cell biology 5: 578-81. doi:10.1038/ncb999
42. Rehmann H, Arias-Palomo E, Hadders MA, Schwede F, Llorca O, et al. (2008) Structure of Epac2 in complex with a cyclic AMP analogue and RAP1B. Nature 455: 124-7. doi:10.1038/nature07187
43. Ponsioen B, Gloerich M, Ritsma L, Rehmann H, Bos JL, et al. (2009) Direct spatial control of Epac1 by cyclic AMP. Molecular and cellular biology 29: 2521-31. doi:10.1128/MCB.01630-08
44. Stouten PF, Sander C, Wittinghofer A, Valencia A (1993) How does the switch II region of G-domains work? FEBS letters 320: 1-6.
45. Jékely G (2003) Small GTPases and the evolution of the eukaryotic cell. BioEssays : news and reviews in molecular, cellular and developmental biology 25: 1129-38. doi:10.1002/bies.10353
46. Yutin N, Wolf MY, Wolf YI, Koonin EV (2009) The origins of phagocytosis and eukaryogenesis. Biology direct 4: 9.

doi:10.1186/1745-6150-4-9

47. Poole A, Penny D (2007) Eukaryote evolution: engulfed by speculation. Nature 447: 913. doi:10.1038/447913a

48. Dong J-H, Wen J-F, Tian H-F (2007) Homologs of eukaryotic Ras superfamily proteins in prokaryotes and their novel phylogenetic correlation with their eukaryotic analogs. Gene 396: 116-24. doi:10.1016/j.gene.2007.03.001

49. Brighouse A, Dacks JB, Field MC (2010) Rab protein evolution and the history of the eukaryotic endomembrane system. Cellular and molecular life sciences : CMLS: 1-17-17. doi:10.1007/s00018-010-0436-1

50. Pereira-Leal JB, Seabra MC (2001) Evolution of the Rab family of small GTP-binding proteins. Journal of molecular biology 313: 889-901. doi:10.1006/jmbi.2001.5072

51. Pereira-Leal JB (2008) The Ypt/Rab family and the evolution of trafficking in fungi. Traffic (Copenhagen, Denmark) 9: 27-38. doi:10.1111/j.1600-0854.2007.00667.x

52. Boureux A, Vignal E, Faure S, Fort P (2007) Evolution of the Rho family of ras-like GTPases in eukaryotes. Molecular biology and evolution 24: 203-16. doi:10.1093/molbev/msl145

53. Colicelli J (2004) Human RAS superfamily proteins and related GTPases. Science's STKE : signal transduction knowledge environment 2004: RE13. doi:10.1126/stke.2502004re13

54. Lemey P (2009) The phylogenetic handbook. 2nd ed. Cambridge University Press. p.

55. Li JB, Gerdes JM, Haycraft CJ, Fan Y, Teslovich TM, et al. (2004) Comparative Genomics Identifies a Flagellar and Basal Body Proteome that Includes the BBS5 Human Disease Gene. Cell 117: 541-552. doi:10.1016/S0092-8674(04)00450-7

CHAPTER 2

Protein complex evolution does not involve extensive network rewiring

Teunis J.P. van Dam and Berend Snel

PLoS Computational Biology 2008 4(7) : e1000132
doi:10.1371/journal.pcbi.1000132

2.1 Abstract

The formation of proteins into stable protein complexes plays a fundamental role in the operation of the cell. The study of the degree of evolutionary conservation of protein complexes between species and the evolution of protein-protein interactions has been hampered by lack of comprehensive coverage of the high-throughput (HTP) technologies that measure the interactome.

We show that new high-throughput datasets on protein co-purification in yeast have a substantially lower false negative rate than previous datasets when compared to known complexes. These datasets are therefore more suitable to estimate the conservation of protein complex membership than hitherto possible.

We perform comparative genomics between curated protein complexes from human and the HTP data in *Saccharomyces cerevisiae* to study the evolution of co-complex memberships. This analysis revealed that out of the 5960 protein pairs that are part of the same complex in human, 2216 are absent because both proteins lack an ortholog in *S. cerevisiae*, while for 1828 the co-complex membership is disrupted because one of the two proteins lacks an ortholog. For the remaining 1916 protein pairs, only 10% were never co-purified in the large scale experiments. This implies a conservation level

of co-complex membership of 90% when the genes coding for the protein pairs that participate in the same protein complex are also conserved.

We conclude that the evolutionary dynamics of protein complexes are, by and large, not the result of network rewiring (i.e. acquisition or loss of co-complex memberships), but mainly due to genomic acquisition or loss of genes coding for subunits. We thus reveal evidence for the tight interrelation of genomic and network evolution.

2.2 Introduction

Many proteins perform their functions together with other proteins to form distinct complexes which are responsible for specific processes in a cell. Understanding how, why and when proteins associate into stable protein complexes is a pivotal part of understanding cellular life. The evolution of protein complexes is intrinsically of interest, as protein complexes are important functional units. In addition, evolutionary information can help us to clean noisy high-throughput data on protein complexes and interactions [1,2]. In general, measuring the evolutionary dynamics of protein complexes should improve the framework for function prediction and comparative analysis of interactome networks. For example, knowledge on interactome evolution can help us to establish how reliably we can transfer measured interactions of a protein in *S. cerevisiae* to its ortholog in human for function prediction.

Various aspects of the evolution of protein complexes and interactomes have been studied [3]. Work on interaction networks so far has revealed that highly connected proteins tend to be more conserved than less connected proteins when looking for the presence or absence in other species [4]. Also, higher connected proteins tend to evolve slower than less connected proteins [5]. Moreover, it has been shown that the subunits of protein complexes seem to evolve uncohesively: the genomes of many species contain only a subset of the genes that make up a protein complex of a particular species [6,7]. However, all these studies did not analyze the evolution of interactions or co-complex membership, but only the evolution of the genes.

The actual conservation of protein interactions themselves is still debated, in part because information and direct measurements of interactions in multiple species is sparse. Suthram and co-workers [8] for instance, have found remarkably low overlap in interaction networks between *P. falciparum* and other eukaryotic interaction networks, like those of yeast and human. They also concluded that even between closer and well studied eukaryotes like *S. cerevisiae* and *D. melanogaster*, many interactions and complexes have been lost. This study, and others like it, have been careful to equate small overlap with a low degree of conservation and has pointed out that the analysis of complex evolution has been hampered by the quality of the available high throughput data. In contrast, anecdotal evidence based on specific cases studied from the literature suggest high conservation of co-complex membership such as observed in the ribosome [9]. Therefore it remains unresolved to what extent protein interactions and protein complexes are conserved.

When analyzing interaction conservation we need to acknowledge that proteins can keep, lose or gain interactions. To properly measure interaction conservation we need data which not only contains protein-protein interactions but also contains data on proteins which do not seem to interact [6]. The measurements as done in interaction

experiments initially provided data on the former. Yet when the coverage of the data is such that it approximates 'complete', the probability that a protein pair without measurable interaction does indeed not interact should increase rapidly.

With the publication of two new datasets of high throughput tandem affinity purification-mass spectrometry (TAP-MS) experiments in *S. cerevisiae* [10,11], data has become available which is seemingly of high enough quality [11,12] to warrant a new look at interaction conservation. We revisit therefore the question of how complexes evolve and how well protein-protein interactions are conserved.

Measuring evolution of protein complexes obviously depends on a reasonable definition of what constitutes a complex: proteins can associate strongly to other proteins and form a stable protein complex (e.g. proteasomes) or proteins can associate transiently to often many other proteins (e.g. a kinase and its substrate) and not be truly part of one stable complex. We chose to study the evolution of the first (stable) type. In addition new insights propose a world view where complexes are not static entities but fluctuate in time and space [10]. Unlike the manner in which it is by necessity stored in reference databases such as MIPS or SGD, the composition of protein complexes is condition and sub cellular localization dependent. This also makes it difficult to study the evolution of protein complexes; i.e. if only a subset of the subunits is involved in a complex in another species, is the complex then conserved? We here adapt to the latter problem by choosing as the unit of which we want to measure conservation "a pair of proteins that are part of the same protein complex". For brevity we will refer to this as "co-complex membership" or sometimes the even shorter and arguably inappropriate term "interaction".

In this study we extend interaction data by defining non-interactions in order to examine co-complex membership conservation between *S. cerevisiae* and human. Estimating the absence of interactions allows us to look at the conservation and not just the overlap between two interaction networks. The analysis reveals that the main processes of evolution for complexes are the acquisition of new or the loss of old subunits as the co-complex interaction network is highly conserved between orthologous proteins in *S. cerevisiae* and human.

2.3 Results

2.3.1 Dataset quality and false negative rate assessed by yeast complexes.

The new TAP-MS datasets seem to be very complete and accurate [10,11,12]. We explicitly test the completeness of the datasets by specifically analyzing to what extent different HTP datasets are able to predict all interactions and absence of interactions, i.e. the false negative rate (type 2 error). A false negative will result in the observation that an association is absent while in reality the interaction is present but the experiment failed to detect it. We use the false negative rate because it is a measure of how complete the actual connectivity of a given protein is represented in the datasets. Such false negative pairs are crucial for the study of evolution, because these false negatives will erroneously lower the degree of conservation.

A reference set of known complexes is needed to assess which co-complex memberships are erroneously reported as absent in the various HTP datasets (false negatives). In the light of the ongoing discussion on what constitutes a complex [10,11], we used

different independent sources such as MIPS and SGD and their intersection (see table 2.3 in Methods section). We use the latter as the main reference, because it provides a reference set in which both MIPS and SGD agree and therefore more is reliable in terms of co-complex memberships and complex definition.

Naturally, there is a trade-off between the false negative rate and false positive rate when choosing an appropriate cut-off value for the TAP-MS datasets. The optimal cut-off value for the socio-affinity scores was determined by plotting a Receiver-Operator Curve (see supplementary material). We found that a relatively low cut-off value of 0 provides an optimal balance between specificity and sensitivity for measuring complex interactions.

We observe that the new datasets achieve very low false negative rates. The Gavin dataset has a false negative rate of 0.23 whereas the Krogan dataset has a false negative rate of 0.32 (table 2.1). Combining the TAP-MS datasets (both union and intersection) does not only increase the number of true positives but also reduces the number of false negatives and consequently the false negative rate (table 2.1), e.g. the intersection of the Gavin and Krogan datasets has a false negative rate of 0.11 (see Figure 2.1 and Methods on dataset construction). These low false negative rates reveal that when the TAP-MS datasets report an absence of interaction only a small percentage is a "failure" of the experimental assay. The new datasets are therefore a substantial improvement for the study of co-complex membership conservation relative to what was available previously.

In addition to the TAP-MS datasets we also analyzed other high-throughput Yeast-2-Hybrid datasets (Y2H) by Uetz et al. [13] and Ito et al. [14] in order to compare them to the new datasets (for an overview on all datasets see table 2.4 in the Methods section). We see that the false negative rate in these Y2H assays is much higher, when we define absence of an interaction from Y2H conventionally, that is to say, an absence is a prey and bait pair that failed to report an interaction. The higher false negative rate of the Y2H datasets is of course to be expected because Y2H measures direct protein-protein interactions rather than co-complex memberships. Mass-spec co-purifications are expected to retrieve co-memberships more easily [15]. At the same time it might also be that Y2H does have a slightly higher natural level of false negatives as implied previously [2]. To test this, we redefined our Y2H negatives for the Uetz dataset as follows: both the bait-prey as the prey-bait has been tested and both failed to report

Table 2.1. False Negative Rates for Different Datasets Compared to Complex Definitions

Datasets	Intersection of MIPS and GO			MIPS			SGD GO		
	FNR*	#FN¶	#TP§	FNR*	#FN¶	#TP§	FNR*	#FN¶	#TP§
Gavin et al.	0.23	1226	4083	0.33	2284	4687	0.37	3769	6328
Krogan et al.	0.32	2209	4644	0.44	4208	5372	0.52	7927	7406
Intersection	0.11	517	4396	0.21	1356	5203	0.25	2378	7233
Inclusive	0.21	1517	5732	0.34	3370	6622	0.42	6572	9247
Uetz et al.	0.66	91	46	0.75	194	63	0.76	270	87
Uetz et al. strict	0.1	5	46	0.11	8	63	0.15	15	87
Ito et al.	0.92	822	76	0.93	1427	114	0.95	2358	114

*False Negative Rate, ¶False Negatives, §True Positives

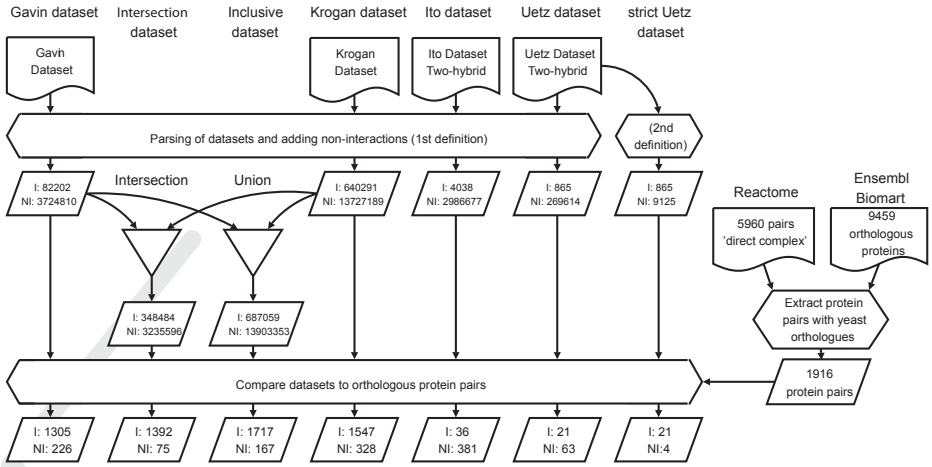


Figure 2.1: Data flow diagram. NI = non-interaction, I = interaction. The non-interactions are calculated for each dataset before they are combined in a union or intersection dataset. The complex definition of Reactome and ortholog definitions from Ensembl are combined to find the conserved protein pairs. The interaction data of the conserved protein pairs are extracted from the datasets and the interaction conservation is calculated.

an interaction. We see a very dramatic decrease in the false negative rate for the Uetz 'strict' dataset (table 2.1). In fact Uetz strict has a false negative rate comparable to the intersection of the two mass-spec datasets (0.10 for Uetz strict as compared to 0.11 for the Intersection of the Gavin and Krogan datasets, see table 2.1). This shows it is possible to obtain reliable indications of the absence of an interaction from apparently less complete datasets. However, this requires specific attention to the method by which an absence of interaction is inferred from the primary data. Due to coverage of this Uetz strict dataset we cannot use it as the main source for the study of the conservation of interaction, but we can use it to test how general our findings from the mass-spec source are, and whether or not they depend on the precise experimental method for detecting interactions.

2.3.2 Interaction network evolution in complexes

The Gavin and Krogan datasets and in particular the combination of these datasets (union and intersection) show a very low false negative rate: i.e. only a small fraction of the true co-complex memberships are not reported by these datasets. Given that these datasets are available with substantially improved false negative rates we have an excellent starting point for comparative genomics to see to what extent co-complex membership is conserved between species. Reactome for human [16] was used as a highly reliable reference set for calculating interaction conservation. Reactome is a high quality manually curated database based on expert opinion. Recently a Co-IP interaction dataset has been published for the human interactome by Ewing et al. [17]. We use this dataset as complementary source to confirm our qualitative trends, rather than our main reference set, because this dataset is only slightly larger than Reactome (6463 interactions vs. 5960), but has less protein pairs with orthologs in yeast (650 vs. 1916) and contains experimental noise (See the supplementary material for analysis performed with the Ewing dataset).

We extracted protein pairs that were part of the same core protein complex according

Table 2.2. Conservation of Protein-Protein Interactions Defined by Reactome in Yeast.

Datasets	Interactions	Non-interactions	Conservation[1]	Coverage[2]	Complex coverage by dataset[3]
Gavin et al.	1305	226	85.20%	68.10%	135
Krogan et al.	1547	328	82.50%	80.70%	150
Intersection	1392	75	94.90%	72.70%	133
Inclusive	1717	167	91.10%	89.60%	152
Uetz et al.	21	63	24.10%	1.10%	26
Uetz et al. strict	21	4	84.00%	1.10%	17
Ito et al.	36	381	8.60%	1.90%	65

Human Datasets	Interactions	Non-interactions	Overlap[1]	Coverage[2]	Complex coverage by dataset[3]
Ewing et al.	16	434	3.60%	0.80%	56
Rual et al.	3	5	37.50%	0.20%	5
Stelzl et al.	4	79	4.80%	0.20%	15

to Reactome. Orthology data was extracted from Ensembl (see methods) in order to transfer the yeast interaction data onto Reactome (Figure 2.1). This analysis revealed that out of the 5960 human co-complex memberships 4044 are absent in yeast due to the absence of either one (1828) or both (2216) of the interaction partners, leaving 1916 pairs with orthologs in yeast. In terms of complexes we found that 66% of human complexes contain less than 50% subunits with orthologs in yeast with an average of 35% over all complexes, which is similar to the percentage of protein pairs. These results are confirmed by orthology calculated with inparanoid [18] (see supplementary material). Thus a large number of co-complex membership pairs are not conserved because either one or both of the genes was lost in fungi or acquired in animals. This is consistent with previous findings on the evolutionary cohesiveness of protein complexes [6]. Therefore a tremendous amount of flexibility in the evolution of protein complexes is not due to the evolution of the co-complex membership (the interactions) itself, but rather due to the acquisition and loss of subunits from the genome.

We subsequently asked how many of the 1916 gene pairs are also part of the same protein complex in yeast and, more importantly, we also counted how many pairs are not interacting according to our inferred non-interacting pairs. In case of inparalogs conservation of interaction was inferred when one of the inparalogs returned a positive interaction from the datasets (see methods). We observe a high rate of co-complex membership conservation: 82.5% to 85.2% for the Gavin and Krogan datasets respectively and 91.1% to 94.9% for the Inclusive and Intersection datasets respectively (table 2.2). Although this seems in contrast to the Y2H datasets (Uetz dataset reaches 24.1%, Ito dataset 8.6%), the Uetz strict dataset returns 84% conservation. The Y2H thus in fact confirms the observation on conservation from the TAP-MS datasets.

The rate of conservation that we obtain from the protein purification experiment datasets are not based on a small subset of protein pairs but on a very large proportion of all associated protein pairs. The TAP-MS datasets have coverage of up to 90% when combined as the union of both datasets. The coverage of Reactome by the Krogan and Gavin datasets is substantial (81% and 68% resp.), whereas the Y2H datasets cover at most 2% (Ito dataset) of the 1916 orthologous protein pairs in Reactome. Moreover the conservation rates are based for e.g. the intersection on 133 distinct complexes

(table 2.2). From the high conservation rates as well as the percentage of coverage as determined from our analysis based on the TAP-MS datasets, we conclude that the evolution of protein complexes is mainly due to the acquisition or loss of subunits and not due to network rewiring.

Analogous to the yeast datasets and the yeast complex definitions, we analyzed the overlap of the human Co-IP [17] dataset and Y2H datasets [19,20] with Reactome. To prevent bias we only took those Reactome gene pairs that have orthologs in yeast. The overlap between the human datasets and Reactome is surprisingly so small that they perform worse than the Y2H datasets from yeast. The small coverage of the human datasets is perhaps caused by the fact that the human HTP interaction studies targeted proteins that are presumably of more interest to mammalian systems.

2.3.3 Loss and acquisition of co-complex associations in human

From the high conservation rates as determined from our analysis we conclude that the evolution of protein complexes is mainly due to the acquisition or loss of subunits and not due to network rewiring. The non-conserved interactions are those associations between protein pairs that are present in yeast and human as orthologs but whose interaction seems to have been either lost in yeast or acquired in human.

These associations are potentially interesting because they tell us about the evolution of new interactions. Out of the 1884 associations covered by the inclusive dataset only 167 seem to be not conserved (see table 2.4). We scanned this list manually searching for possible errors in annotation, false negatives and true negatives (actual non-conserved protein-protein interactions). Of the 167 protein pairs 139 pairs are present in the same complex in yeast according to GO and/or MIPS or based on literature. In other words, a large portion of these pairs seem to be a member of the same protein complex in yeast and human according to the literature, but were never co-purified in either Krogan or Gavin. I.e. these 139 are possible false negatives of the experimental assays rather than non-conserved interaction pairs. The remaining 28 non-conserved interactions (see supplementary material) consist of errors in orthology of one gene (5 interactions), incorrect assignment of two proteins to a complex in Reactome (10 interactions) and possible neo-functionalisation after duplication in human (3 proteins, 13 interactions).

Based on the analysis of the proteins pairs which did not have an interaction according to the HTP datasets, it seems that the actual conservation of co-complex membership might be higher than follows from our analysis, because we mostly ran into potential errors in orthology assignment, conceptual issues in the curated database of Reactome, or false negatives in the HTP assay. Interestingly, in this analysis the three proteins which represent potentially new complex memberships, are all proteins which have retained the same or similar function as their orthologs in yeast but have acquired additional functions and interactions in human.

2.4 Discussion

We have shown that with the publication of the TAP-MS datasets by Gavin et al. [10] and Krogan et al. [11] we now have datasets which are sufficiently large to reliably estimate the level of co-complex membership conservation. Specifically, we have shown that the false negative rate of these datasets can be reduced to 7%. This means that we are now

able to do comparative network studies with substantially less coverage problems for the yeast interactome than previous studies. This is important as estimates of the level of co-complex membership conservation do not only depend on reliable measures for the presence of a link but also on reliable measures for the absence of a link.

Unfortunately similar interaction data is not available for other species. We have therefore chosen to use a curated interaction database called Reactome and extracted complex definitions. Combining the human Reactome complex definition and the interaction data for yeast reveals that the complex protein pairs which have been conserved in both species do not lose their interaction in contrast to what has been previously suggested [8,21]. We conclude therefore that evolution of protein complexes does not involve extensive network rewiring, but is mostly due to loss of subunits and the acquisition of novel proteins.

This type of behavior is clearly illustrated by the eIF3 protein complex from human and its comparison to the complex in yeast (see Figure 2.2). The eIF3 complex in yeast (left) and human (right) are depicted in a network with similar topology relevant to the orthologs (connected by dotted lines). Although the eIF3 complex in human has expanded compared to yeast, all yeast proteins are also part of the same complex in human (light gray). Modifications of the complex during evolution have been through the acquisition of new proteins (dark gray).

The high degree of co-complex membership conservation could potentially arise from some degree of circularity: the protein complexes in human have been originally identified in yeast. However, our knowledge of human complexes is not limited by what we know about complexes in yeast, as can be deduced by many human subunits which do not have orthologs in yeast such as EF3C or IF36 in the example of the eIF3 protein complex (Figure 2.2). In general many human interactions are disrupted in yeast due to the absence of either one (1828) or both (2216) of the interaction partners. All these subunits are part of a complex in human but are absent in yeast. The knowledge about these subunits is the result of direct intensive biochemical analysis in human or other animal systems. Therefore, we have a substantial degree of trust in our estimate of interaction conservation, because the knowledge on the protein complexes deposited in Reactome is the result of direct extensive experimentation in animal systems and is not only based on experimentation in yeast.

An important aspect of protein-protein interaction evolution is that the physical interaction surface is often provided by distinct protein domains. In evolution of protein-protein interactions they play an important role as acquisition or loss of a particular domain can result in the combination of new interactions with new functions. Itzhaki et al. [22] report that 9% of protein-protein interactions in yeast and 20% in human can be ascribed

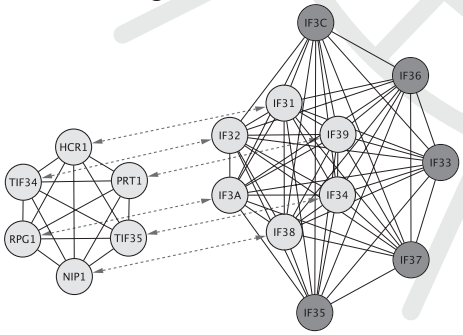


Figure 2.2: The eIF3 protein complex. The eIF3 complexes in yeast (left) and human (right) are depicted in a network with similar topography relevant to the orthologs (connected by dotted lines). Although the eIF3 complex in human has expanded compared to yeast, all yeast proteins still have orthologs in the human eIF3 complex. Modification of the complex seems to have been mainly through the acquisition of new proteins.

to domain-domain interactions. It therefore bears to mind that a small part of co-complex membership conservation might not be due to the conservation of whole proteins but due to specific domains which have maintained the interaction. This would leave a conserved interaction network the freedom to add or change function without having to compromise interaction integrity. Another possible theoretical framework for our observations is given by Kirschner and Gerhart [23], who argue that conserved mechanisms or processes are conserved because they "deconstrain" phenotypic variations in other processes. Our observations neatly fit their theory: the conserved proteins and their conserved interactions represent a "backbone" to which variable subunits are observed to be added or removed.

The possible new interactions that we have found, XAB2, PCBP1 and PABP2, still have the same or similar function as their yeast orthologs, but have acquired new functions and new interactions in human. Additions to the functionality were made only through minor instead of radical adjustments leaving the interaction network intact and added upon. In the light of co-complex membership this might imply that it is easier to add function and interactions than it is to remove the interaction while retaining the gene. The high conservation of co-complex memberships is also support for bioinformatic function prediction by transfer of information on complex-membership between orthologs: this aspect of gene function can be reliably transferred between evolutionary divergent species such as yeast and human when the partner gene is also present.

We have shown that the gain of interactions by existing proteins in complexes seems quantitatively not important in evolution. Rather the evolution of protein complexes is dominated by co-complex memberships that are acquired or lost concomitantly with acquiring or losing the gene. However, the precise order of events in the latter case is difficult to determine. If we for example suppose that the absence of an ortholog in yeast of a human protein complex member is the result of a gene loss (deletion) in the fungal lineage (rather than being acquired in animals), then there are two scenarios than can explain this loss. On the one hand the loss of membership to a protein complex could have preceded the evolutionary loss of the gene. On the other hand a co-complex membership is by definition disrupted by the deletion of the gene coding for the subunit from the genome.

For both examples divergence in transcriptional regulation could mediate less dramatic scenarios of interaction loss. Transcriptional regulation diverges significantly between relatively close species [24] and is therefore a faster process than for example gene loss or acquisition. Loss of membership could have preceded a fast transcriptional down regulation to avoid expression of potentially rogue proteins before the actual loss of the gene. If a subunit is no longer needed deletion of this subunit could have been preceded by down regulation, which could have given the organism some time to adapt (stabilize the complex) to the missing of the subunit before its deletion from the genome.

Although gene loss preceded by interaction loss seems somewhat more likely, the high level of co-complex membership conservation that we observe in those cases were the protein pairs are present in both species, suggest a low frequency of such evolutionary intermediate stages. Because we find such low frequency of intermediate stages and a high conservation rate of interactions between conserved proteins we reveal evidence of the tight interrelation of genomic and network evolution.

2.5 Methods

2.5.1 Interaction Datasets

Mass spectrometry datasets

The Gavin dataset with socio-affinity scores was obtained from the embl website (http://yeast-complexes.embl.de/) as referred to in the original article [10]. The Krogan dataset has been obtained from Vera van Noort who kindly provided us with a processed tab delimited file in which the raw Krogan data had been converted into socio-affinity scores as defined by Gavin et al. [10]. For an overview of all interaction datasets used in this publication see table 2.4.

Yeast-2-Hybrid datasets *S. cerevisiae*

Yeast-2-Hybrid interaction data for *S. cerevisiae* was downloaded from BIOGRID (http://www.thebiogrid.org/ 01/03/2007). The Y2H datasets from Uetz et al. [13] and Ito et al. [14] were extracted by pubmed id.

Yeast-2-Hybrid datasets for Human

Stelz [20] and Rual [19] datasets were obtained from the IntAct database (http://www.ebi.ac.uk/intact/) on 01/12/2007. Files were downloaded in PSI MI 2.5 XML format and data was extracted by using XMLMakerFlattener.

Co-IP dataset for Human

The Ewing [17] dataset was downloaded from the IntAct database on 01/14/2008. The four PSI MI XML files where parsed for primary UNIPROT identifiers. Id's of proteins which did not have a primary identifier where retrieved from the UNIPROT database by blast (100% identity, lowest E-value).

Defining Non-Interactions

We have defined absence as interactions between proteins that have been successfully purified as either bait or prey, but have not occurred together. For the 'Uetz strict' dataset we have defined absence of interactions between proteins that have been successfully purified as both bait and prey, but have not occurred together. 'Absence of interaction' was included into the datasets by assigning the protein pair the socio-affinity score 0 which did not occur in each of the original datasets.

Combining the mass spectrometry datasets

The Gavin and Krogan datasets were combined in two ways. Firstly the Intersection dataset represents the intersection of protein pairs of both Gavin and Krogan datasets after the addition of non-interactions. The socio-affinity scores were averaged. Secondly the Inclusive dataset represents the union of protein pairs of both Gavin and Krogan datasets after the addition of non-interactions. The socio-affinity scores where averaged where appropriate. It may be noted that the total number of positive interactions in the intersection dataset is larger than the Gavin dataset. This is because the dataset was combined by identical protein pairs which allowed for many interactions in Krogan to be included which are non-interactions in Gavin.

Table 2.3. Overview of Complex Definitions

Definition	Reactome	MIPS	SGD GO
Source	Reactome Database	MIPS Database	SGD Database
Processing	"direct complex" interactions	Subunits pooled by complex ID	By GO category
Date	9/19/2006	5/18/2006	5/9/2007
Nr of complexes	3915	217	225
Min complex size	2	2	2
Max complex size	140	81	94
Avg. complex size	7.72	6.33	7.55
Median	2	4	4
Co-complex memberships	5960	15613	19073
Proteins	973	1194	1467

Table 2.4. Overview of PPI Datasets

Datasets	Interactions	Non-interactions	Species	Source	Method	Advantages	Disadvantages
Gavin et al.	82202	3724810	Yeast	Gavin et al.	TAP-MS	Large datasets. Repeated purifications.	Does not detect low affinity interactions. Does not detect 1-to-1 interactions but clusters of proteins.
Krogan et al.	640291	13727189	Yeast	Krogan et al.	TAP-MS	"	"
Intersection	348484	3235596	Yeast	This publication	TAP-MS	"	"
Inclusive	687059	13903353	Yeast	This publication	TAP-MS	"	"
Uetz et al.	865	269614	Yeast	BioGRID	Y2H	Can also detect low affinity interactions. Measures 1-to-1 interactions.	Low coverage.
Uetz et al. strict	865	9125	Yeast	This publication	Y2H	"	"
Ito et al.	4038	2986677	Yeast	BioGRID	Y2H	"	"
Rual et al.	1911	614341	Human	IntAct	Y2H	"	"
Stelzl et al.	1967	249857	Human	IntAct	Y2H	"	"
Ewing et al.	5761	1804013	Human	IntAct	PI-HTMS	Larger than human Y2H datasets.	Purifications done only once.

2.5.2 Complex Definitions

Yeast complex definitions

For an overview of all complex definitions in this publication, see table 2.3.

The MIPS complex definition was downloaded from ftp://ftpmips.gsf.de/yeast/catalogues/complexcat, last updated 05/18/2006. Proteins were pooled per complex ID and interactions were defined between proteins which are present in the same complex.

The SGD GO complex definition was provided by Patrick van Kemmeren. SGD GO (as of 05/09/2007) was parsed, keeping only those components which contain the following

words in their GO description: complex, subunit, ribosome, proteasome, nucleosome, repairosome, degradosome, apoptosome, replisome, holoenzyme or snRNP. Only the lowest annotation level was maintained. Associations that were obtained from high-throughput data have been removed to avoid pollution with false positive interactions. Specifically the following publications were excluded: Ito et al, (PMID: 10655498), Ito et al, (PMID: 11283351), Uetz et al, (PMID: 10688190), Ho et al, (PMID: 11805837), Gavin et al, (PMID: 11805826), Tong et al, (PMID: 14764870), Davierwala et al, (PMID: 16155567), Gavin et al, (PMID: 16429126), Schuldiner et al, (PMID: 16269340), Krogan et al, (PMID: 16554755), Pan et al, (PMID: 16487579) and Miller et al, (PMID: 16093310).

2.5.3 Reactome and Orthology

Human protein-protein interaction pairs as defined by Reactome were downloaded from http://www.reactome.org/download/current/homo_sapiens.interactions.txt.gz on August 19 2006. According to the Reactome annotation standard, protein pairs in direct complex are not necessarily directly interacting but are part of the same core complex, while indirect complex means that two proteins are in the same meta-complex, i.e. two direct complexes that under certain cellular conditions associate (for example the TFII transcription factors and RNA polymerase II forming the pre-initiation complex). We extracted protein pairs which were designated 'direct complex' as interaction type and excluded protein pairs designated 'indirect complex'. We only kept protein pairs assigned 'direct interaction', because we want only core complex proteins to keep our definition strict.

Orthology data was retrieved from the Ensembl database [25] version 41 using the BioMart mining tool (http://www.ensembl.org/biomart/martview/ accessed on October 26 2006). For deriving orthology Ensembl uses a pipeline for ortholog/paralog prediction based on best reciprocal similarity relationship as of June 2006. The method includes determining gene families by best reciprocal match, tree construction by PHYML and MUSCLE and tree reconciliation by the RAL algorithm. We have provided results based on orthology defined by the inparanoid program [18] in de supplementary materials. Although inparanoid is a less-advanced orthology inference method than Ensembl it shows slightly higher conservation of co-complex memberships.

In case of inparalogs in yeast interaction between the human protein pair was inferred from yeast when at least one combination of the yeast orthologs has an interaction according to the interaction dataset. We assumed that if one of the combinations has an interaction, the interaction is conserved in evolution and the other orthologs have lost the interaction after function divergence. This means that the conserved interaction does not have to be between orthologs which are closest in sequence which is consistent with Notebaart et al. [26] who state that an ortholog which has identical function, does not necessarily have to be closest in sequence.

2.5.4 Data handling

All data was handled by Perl scripts (Perl 5.8.8) on a 64 bit Linux machine.

2.6 Acknowledgments

This work is funded by the BioRange program SP 2.1.3.4. Vera van Noort has kindly provided us with the converted Krogan dataset and Patrick van Kemmeren has provided

us with the SGD GO based complex definition. For their input in discussions we also like to thank Jos Boekhorst and Like Fokkens.

2.7 Supplementary materials

2.7.1 Determining optimal cut-off value for TAP-MS datasets

There is a tradeoff between the false negative rate and the false positive rate when determining the optimal cut-off value for the TAP-MS datasets. Because false positives are hard to determine due to the lack of larger reference sets, we have calculated a Receiver-Operator curve based on the agreement of two complex definitions, GO and MIPS (See table S2.1 and Figure S2.1). We have defined as positive interacting protein pairs two proteins in the same complex according to both MIPS and GO. Negative interacting protein pairs are defined as two proteins of which each is in another complex according to both MIPS and GO. As our benchmark dataset we have used the intersection dataset. We have also plotted Y2H datasets by Uetz et al. and Ito et al. for comparison to the Intersection dataset based on Krogan et al. and Gavin et al.. From this curve we find that a cutoff value of 0 is optimal for our question.

Table S2.1. Effect of cut-off on FN and FP

Cutoff	TP	FP	FN	TN
0	7130	33960	2481	185684
1	6542	14431	3069	205213
2	5223	3252	4388	216392
3	3510	696	6101	218948
4	2834	199	6777	219445
5	2397	97	7214	219547
6	1933	66	7678	219578
7	1614	48	7997	219596
8	1301	33	8310	219611
9	989	23	8622	219621
10	734	15	8877	219629
11	514	11	9097	219633
12	310	4	9301	219640
13	193	2	9418	219642
14	111	0	9500	219644
15	53	0	9558	219644

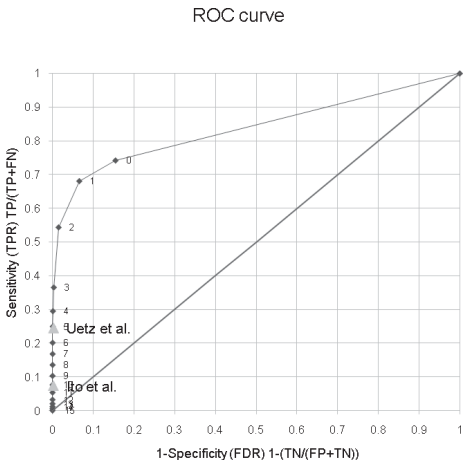


Figure S2.1. Reciever-Operator Curve for the TAP-MS datasets

2.7.2 Ewing et al. HTP IP-HTMS dataset used to calculate conservation between human and yeast interactome

We have chosen Reactome as our reference set in human for calculating the conservation of co-complex membership because it is manually curated and based on expert opinion

Table S2.2.

Cut-off: none	**Total interactions:**	5761	**Total conserved:**	650
Dataset	**PPI**	**NO-PPI**	**Conservation(%)**	**Coverage(%)**
Uetz	6	40	13.04	0.92
Ito	10	256	3.76	1.54
Uetz Int	6	3	66.67	0.92
Gavin	75	292	20.44	11.54
Krogan	154	450	25.50	23.69
Intersection	117	245	32.32	18.00
Inclusive	171	433	28.31	26.31

Cut-off: 0.3	**Total interactions:**	2039	**Total conserved:**	219
Dataset	**PPI**	**NO-PPI**	**Conservation(%)**	**Coverage(%)**
Uetz	5	10	33.33	2.28
Ito	9	63	12.50	4.11
Uetz Int	5	0	100.00	2.28
Gavin	58	81	41.73	26.48
Krogan	99	108	47.83	45.21
Intersection	78	59	56.93	35.62
Inclusive	105	102	50.72	47.95

Cut-off: 0.4	**Total interactions:**	695	**Total conserved:**	91
Dataset	**PPI**	**NO-PPI**	**Conservation(%)**	**Coverage(%)**
Uetz	2	3	40.00	2.20
Ito	5	22	18.52	5.49
Uetz Int	2	0	100.00	2.20
Gavin	33	18	64.71	36.26
Krogan	52	31	62.65	57.14
Intersection	37	14	72.55	40.66
Inclusive	53	30	63.86	58.24

Cut-off: 0.5	**Total interactions:**	245	**Total conserved:**	34
Dataset	**PPI**	**NO-PPI**	**Conservation(%)**	**Coverage(%)**
Uetz	0	1	0.00	0.00
Ito	2	5	28.57	5.88
Uetz Int	0	0	NA	0.00
Gavin	18	5	78.26	52.94
Krogan	26	5	83.87	76.47
Intersection	20	3	86.96	58.82
Inclusive	26	5	83.87	76.47

Table S2.3.
Conservation based on inparanoid (2596 conserved protein pairs)

Datasets	Interactions	Non-interactions	Conservation[1]	Coverage[2]
Gavin et al.	1646	274	85.73%	63.41%
Krogan et al.	2084	462	81.85%	80.28%
Intersection	2317	239	90.65%	89.25%
Inclusive	1761	84	95.45%	67.84%
Uetz et al.	23	83	21.70%	0.89%
Uetz et al. strict	23	4	85.19%	0.89%
Ito et al.	37	634	5.51%	1.43%

Human Datasets	Interactions	Non-interactions	Overlap[1]	Coverage[2]
Ewing et al.	2	9	18.18%	0.08%
Rual et al.	5	111	4.31%	0.19%
Stelzl et al.	46	546	7.77%	1.77%

Conservation based on Ensembl (1916 conserved protein pairs) Identical to table 2.3 in the main text

Datasets	Interactions	Non-interactions	Conservation[1]	Coverage[2]
Gavin et al.	1305	226	85.20%	68.10%
Krogan et al.	1547	328	82.50%	80.70%
Intersection	1392	75	94.90%	72.70%
Inclusive	1717	167	91.10%	89.60%
Uetz et al.	21	63	24.10%	1.10%
Uetz et al. strict	21	4	84.00%	1.10%
Ito et al.	36	381	8.60%	1.90%

Human Datasets	Interactions	Non-interactions	Overlap[1]	Coverage[2]
Ewing et al.	16	434	3.60%	0.80%
Rual et al.	3	5	37.50%	0.20%
Stelzl et al.	4	79	4.80%	0.20%

and therefore is likely to contain fewer errors. A new CoIP dataset for human by Ewing et al. has become available and we show here the same calculations when Reactome is substituted by this dataset below (Table S2.2).

The authors state that interactions with a confidence score higher or equal to 0.3 should be regarded as high confidence. When using a higher cut-off value we see a steady rise in conservation (87% for >= 0.5 against the intersection dataset) but also see the total number of conserved protein pairs plummet towards small numbers. The number of conserved protein pairs in Ewing when no cut-off value was used is significantly less than for Reactome and the conservation calculated is therefore less representative.

Ewing shows a much lower preservation of orthologs of protein pairs than Reactome (11% and 32% resp.). It is reported by Ewing et al. explicitly that they have based their

bait selection on human disease association. Ewing therefore does not represent the basal conserved eukaryotic machinery as well as Reactome, which would account for the low conservation of protein pairs.

2.7.3 Orthology: results are not sensitive to orthology definition.

We also performed our analysis with another orthology definition. We have used inparanoid [18] to calculate orthology between human sequences from the UniProt database and yeast sequences from SGD. Inparanoid is a script which uses BLAST to obtain homology and calculated orthologs taking into account the existence of paralogs and in-paralogs. We have used the standard settings for inparanoid. Table S2.3, is similar to table 2.4 in the publication but based on the inparanoid orthology. We see that the orthology based on inparanoid results in slightly higher conservation and more conserved protein pairs. We feel that the orthology based on Ensembl is more advanced as it is based on reciprocal match, phylogenetic tree construction and tree reconciliation. We therefore used the Ensembl definition in our main analysis as opposed to InParanoid.

2.7.4 Errors in orthology, complex definition and neo-functionalisation

Of the 167 non-interactions as found using Reactome and the Inclusive dataset, 139 appear to be potential false negatives. The remaining 28 non-conserved interactions consist of errors in orthology of one gene (5 interactions), incorrect assignment of two proteins to a complex in Reactome (10 interactions) and possible neo-functionalisation after duplication in human (3 proteins, 13 interactions).

Five protein pairs do not show an interaction due to incorrect orthology assignment in Ensembl. The human protein TF2H4 [Swiss-Prot:Q92759] is annotated as orthologous to VAS1 [SGD:YGR094W] and is present in five conserved protein pairs in Reactome. We could not confirm any homology between these proteins (let alone orthology) and it seems unlikely as well from the annotation: TF2H4 is a subunit of Transcription Factor IIH complex whereas VAS1 is a valyl-tRNA synthetase.

Ten protein pairs are probably erroneously assigned to the spliceosome complex in Reactome, based on our re-analysis of the available literature. Of these ten pairs, five protein pairs contain NFX1 [Swiss-Prot:Q12986] and five protein pairs contain SMC1 alpha [Swiss-Prot:Q14683]. Human NFX1 [Swiss-Prot:Q12986] is associated with the spliceosome complex and “Export Receptor bound mature mRNA Complex” according to Reactome (internal id’s 72022, 72074, 72057, 159329, 159259, 113815). NFX1 however is a transcription factor for MHCII genes and is not implicated in pre-mRNA modifications or nuclear export. Confusingly the NXF1 protein (mind the spelling of NXF1) is a known nuclear export factor. NXF1 is not listed as part of the “Export Receptor bound mature mRNA Complex”. A misspelling of NXF1 might have caused a mix-up in Reactome. (for example NXF1 [Swiss-Prot:Q9UBU9] is misspelled in Cohen and Panning [27] as NFX1.)

SMC1 alpha (human [Swiss-Prot:Q14683], yeast [SGD:YFL008W]), responsible for another five protein pairs, is part of the cohesin complex but also takes part in the spliceosome formation according to Reactome (internal id’s 72159, 72022, 72074, 77505, 72057). However, we could not find any literature which linked SMC1 alpha directly to the spliceosome. The link between the cohesin complex and the spliceosome is one of its alleged co-complex members CD2B2 [Swiss-Prot:O95400] of which LIN1 [SGD:YHR156C is its ortholog in yeast. LIN1 is implicated to link chromatine modification

and the cohesin complex to the spliceosome complex [28]. But the similarity between CD2B2 and LIN1 is weak, and both have very different functions. CD2B2 is involved in immunity and binds to antibodies, whereas LIN1 is a non-essential component of U5 snRNP. CD2B2 and the spliceosome are mentioned together in an article by Monos et al. [29] because an antibody raised against CD2B2 also reacted with the spliceosomal Sm B/B' proteins. The experimental link between SMC1 alpha and the spliceosome is weak and it can therefore be argued that SMC1 is not part of the spliceosome complex.

We identified 13 protein pairs which could be possible new interactions. Each of these pairs contain one of three proteins: PCBP1 [Swiss-Prot:Q15365], PABP2 [Swiss-Prot:Q86U42] and XAB2 [Swiss-Prot:Q9HCS7]. The human PCBP1 is involved in regulating the spliceosome [30]. Its yeast ortholog PBP2/HEK1 [SGD:YBR233W] is involved in the regulation of telomere position effect and telomere length [31]. However PCBP1 is not the only ortholog of PBP2. 14 human proteins are orthologs to PBP2. These are active in different processes, some of them still perform the ancestral function [32]. So whereas PBP2 solely has a function in the regulation of telomere position effect and telomere length in yeast, the human PCBP family of inparalogs has gained many other functions and interaction partners after several rounds of duplications in the course of evolution (neofunctionalization of inparalogs).

The human PABP2 is a poly(A)-binding protein and is part of the "3' end cleaved, ligated exon containing complex" in the nucleus according to Reactome. Its ortholog in yeast, SGN1 [SGD:YIR001C], is a poorly characterized poly(A)-binding protein that localizes to the cytoplasm and not to the nucleus [33]. Hence some degree of functional differentiation took place in either human or yeast.

The human protein XAB2 is involved in transcription coupled-nucleotide excision repair (TC-NER) [34] and also in mRNA splicing (spliceosome) [35] albeit indirectly. The ortholog of XAB2 in yeast, SYF1 [SGD:YDR416W], is a component of the spliceosome [36,37]. SYF1 however has not been implied with nucleotide excision repair. XAB2 apparently has gained a new function, and new interaction partners, in human TC-NER, but also seemingly retained its ancestral function (or some of it), like its yeast ortholog SYF1, in the spliceosome.

2.8 References

1. Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, et al. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc Natl Acad Sci U S A 100: 11394-11399.
2. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, et al. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417: 399-403.
3. Pereira-Leal JB, Levy ED, Teichmann SA (2006) The origins and evolution of functional modules: lessons from protein complexes. Philos Trans R Soc Lond B Biol Sci 361: 507-517.
4. Wuchty S, Oltvai ZN, Barabasi AL (2003) Evolutionary conservation of motif constituents in the yeast protein interaction network. Nat Genet 35: 176-179.
5. Fraser HB, Wall DP, Hirsh AE (2003) A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol Biol 3: 11.
6. Snel B, Huynen MA (2004) Quantifying modularity in the evolution of biomolecular systems. Genome Res 14: 391-397.
7. Glazko GV, Mushegian AR (2004) Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. Genome Biol 5: R32.
8. Suthram S, Sittler T, Ideker T (2005) The *Plasmodium* protein network diverges from those of other eukaryotes. Nature 438: 108-112.
9. Smits P, Smeitink JA, van den Heuvel LP, Huynen MA, Ettema TJ (2007) Reconstructing the evolution of the mitochondrial ribosomal proteome. Nucleic Acids Res.

10. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440: 631-636.
11. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, et al. (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature 440: 637-643.
12. Hart GT, Lee I, Marcotte EM (2007) A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC Bioinformatics 8: 236.
13. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 403: 623-627.
14. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 98: 4569-4574.
15. Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, et al. (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends Genet 18: 529-536.
16. Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, et al. (2007) Reactome: a knowledge base of biologic pathways and processes. Genome Biol 8: R39.
17. Ewing RM, Chu P, Elisma F, Li H, Taylor P, et al. (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. Molecular Systems Biology 3: 89.
18. Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. Journal of Molecular Biology 314: 1041-1052.
19. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, et al. (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437: 1173-1178.
20. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122: 957-968.
21. Beltrao P, Serrano L (2007) Specificity and evolvability in eukaryotic protein interaction networks. PLoS Comput Biol 3: e25.
22. Itzhaki Z, Akiva E, Altuvia Y, Margalit H (2006) Evolutionary conservation of domain-domain interactions. Genome Biology 7: R125.
23. Kirschner M, Gerhart J (1998) Evolvability. Proceedings of the National Academy of Sciences of the United States of America 95: 8420-8427.
24. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, et al. (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat Genet 39: 730-732.
25. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, et al. (2007) Ensembl 2007. Nucleic Acids Res 35: D610-617.
26. Notebaart RA, Huynen MA, Teusink B, Siezen RJ, Snel B (2005) Correlation between sequence conservation and the genomic context after gene duplication.[erratum appears in Nucleic Acids Res. 2005;33(22):7176]. Nucleic Acids Research 33: 6164-6171.
27. Cohen HR, Panning B (2007) XIST RNA exhibits nuclear retention and exhibits reduced association with the export factor TAP/NXF1. Chromosoma 116: 373-383.
28. Bialkowska A, Kurlandzka A (2002) Proteins interacting with Lin 1p, a putative link between chromosome segregation, mRNA splicing and DNA replication in *Saccharomyces cerevisiae*. Yeast 19: 1323-1333.
29. Monos D, Heliopoulos J, Argyris E, Cordopatis P, Zompra A, et al. (2006) Analysis of the CD2 and spliceosomal Sm B/B' polyproline-arginine motifs defined by a monoclonal antibody using a phage-displayed random peptide library. J Mol Recognit 19: 535-541.
30. Meng Q, Rayala SK, Gururaj AE, Talukder AH, O'Malley BW, et al. (2007) Signaling-dependent and coordinated regulation of transcription, splicing, and translation resides in a single coregulator, PCBP1. Proc Natl Acad Sci U S A 104: 5866-5871.
31. Denisenko O, Bomsztyk K (2002) Yeast hnRNP K-like genes are involved in regulation of the telomeric position effect and telomere length. Mol Cell Biol 22: 286-297.
32. Makeyev AV, Liebhaber SA (2002) The poly(C)-binding proteins: a multiplicity of functions and a search for mechanisms. Rna 8: 265-278.
33. Winstall E, Sadowski M, Kuhn U, Wahle E, Sachs AB (2000) The *Saccharomyces cerevisiae* RNA-binding protein Rbp29 functions in cytoplasmic mRNA metabolism. J Biol Chem 275: 21817-21826.
34. Nakatsu Y, Asahina H, Citterio E, Rademakers S, Vermeulen W, et al. (2000) XAB2, a novel tetratricopeptide repeat protein involved in transcription-coupled DNA repair and transcription. J Biol Chem 275: 34931-34937.
35. Yonemasu R, Minami M, Nakatsu Y, Takeuchi M, Kuraoka I, et al. (2005) Disruption of mouse XAB2 gene involved in pre-mRNA splicing, transcription and transcription-coupled DNA repair results in preimplantation lethality. DNA Repair (Amst) 4: 479-491.
36. Ben-Yehuda S, Dix I, Russell CS, McGarvey M, Beggs JD, et al. (2000) Genetic and physical interactions between factors involved in both cell cycle progression and pre-mRNA splicing in *Saccharomyces cerevisiae*. Genetics 156: 1503-1517.
37. Russell CS, Ben-Yehuda S, Dix I, Kupiec M, Beggs JD (2000) Functional analyses of interacting factors involved in both pre-mRNA splicing and cell cycle progression in *Saccharomyces cerevisiae*. Rna 6: 1565-1572.

CHAPTER 3

Phylogeny of the CDC25 homology domain reveals rapid differentiation of Ras pathways between early animals and fungi

Teunis J.P. van Dam, Holger Rehmann, Johannes L. Bos and Berend Snel

Cellular Signaling 2009 21(11) Pages 1579-85
doi: 10.1016/j.cellsig.2009.06.004

3.1 Abstract

The members of the Ras-like superfamily of small GTP-binding proteins are molecular switches that are in general regulated in time and space by guanine nucleotide exchange factors and GTPase activating proteins. The Ras-like G-proteins Ras, Rap and Ral are regulated by a variety of guanine nucleotide exchange factors that are characterized by a CDC25 homology domain. Here we study the evolution of the Ras pathway by determining the evolutionary history of CDC25 homology domain coding sequences. We identified CDC25 homology domain coding sequences in animals, fungi and a wide range of protists, but not in plants. This suggests that the CDC25 homology domain originated in or before the Last Eukaryotic Ancestor but was subsequently lost in plant. We provide evidence that at least seven different ancestral Ras guanine nucleotide exchange factors were present in the ancestor of fungi and animals. Differences between present day fungi and animals are the result of loss of ancestral Ras guanine nucleotide exchange factors early in fungal and animal evolution combined with lineage specific duplications and domain acquisitions. In addition, we identify Ral guanine exchange factors and Ral in early diverged fungi, dating the origin of Ral signaling back to before the divergence of animals and fungi. We conclude that the Ras signaling pathway evolved by gradual

change as well as through differential sampling of the ancestral CDC25 homology domain repertoire by both fungi and animals. Finally, comparison of the domain composition of the Ras guanine nucleotide exchange factors shows that domain addition and diversification occurred both prior to and after the fungal-animal split.

3.2 Introduction

Ras signaling pathways are involved in the regulation of a wide variety of cellular processes such as cell proliferation, cell adhesion, cell movement, division, secretion and cell differentiation. It is therefore not surprising that Ras G-proteins and their up- or downstream effectors are mutated in many types of cancer [1], making Ras signaling pathways a prime target for cancer research.

The Ras G-proteins are members of the Ras-like superfamily of small G-proteins, which includes Ras, Rho, Rab, Arf and Ran proteins [2]. These proteins cycle between an inactive GDP-bound and an active GTP-bound conformation. This cycling is regulated by Guanine nucleotide Exchange Factors (GEFs) and GTPase Activating Proteins (GAPs). GEFs exchange the G-protein bound GDP for the cellular more abundant GTP, while GAPs increases the intrinsic GTP hydrolyzing activity of G-proteins by several orders of magnitude. Each family of small G-proteins has its own set of GEFs and GAPs that are not homologous [3], e.g. the catalytic region of RhoGEFs is a tandem PH-DH domain, whereas RasGEFs have a CDC25 Homology Domain (CDC25 HD) with an unrelated protein fold.

The CDC25 HD was first identified in the budding yeast *Saccharomyces cerevisiae* protein CDC25 and GEFs containing this domain regulate exclusively the Ras, Rap and Ral proteins (commonly called here the Ras* proteins), a subset of the Ras-like superfamily of small G-proteins. In the yeast *S. cerevisiae* three different proteins with CDC25 HD are present for five Ras* proteins, whereas in humans cells thirty CDC25 HD proteins are presents for sixteen Ras* proteins. *S. cerevisiae* does not contain Ral proteins, suggesting that Ral is an animal innovation. Fungi and animals have very different sets of RasGEFs, but we do not know how these differences arose.

An implicit consensus within the field was that Ras* G-proteins were a fungal-animal (Opisthokont) invention, but recent genomic evidence seems to suggest that Ras* is much older and its origins can be traced back to the Last Eukaryotic Common Ancestor (LECA) [4]. However no study has been done to investigate the origins of the CDC25 HD. Given the role of CDC25 HD-containing proteins as the principal activator of the Ras* G-proteins, knowledge of its origin and evolution may give new insight into the evolution of the Ras* signaling pathways. We therefore studied the origin and evolution of this domain using the increasing amount of sequenced genomes of eukaryote species from all major phyla.

We found that, similar to Ras* G-proteins, the CDC25 HD is an ancient protein domain for which the origin dates back to the LECA. Furthermore, the data suggests that at least seven and possibly twelve ancestral domains were already present at that stage in evolution. Interestingly, we observe an unusually strong functional and evolutionary relationship of Ras* and the CDC25 HD. Finally, we found that the Ral protein as well as RalGEFs are present in early diverged fungi (e.g. primitive fungi) suggesting an earlier origin of these proteins than originally thought.

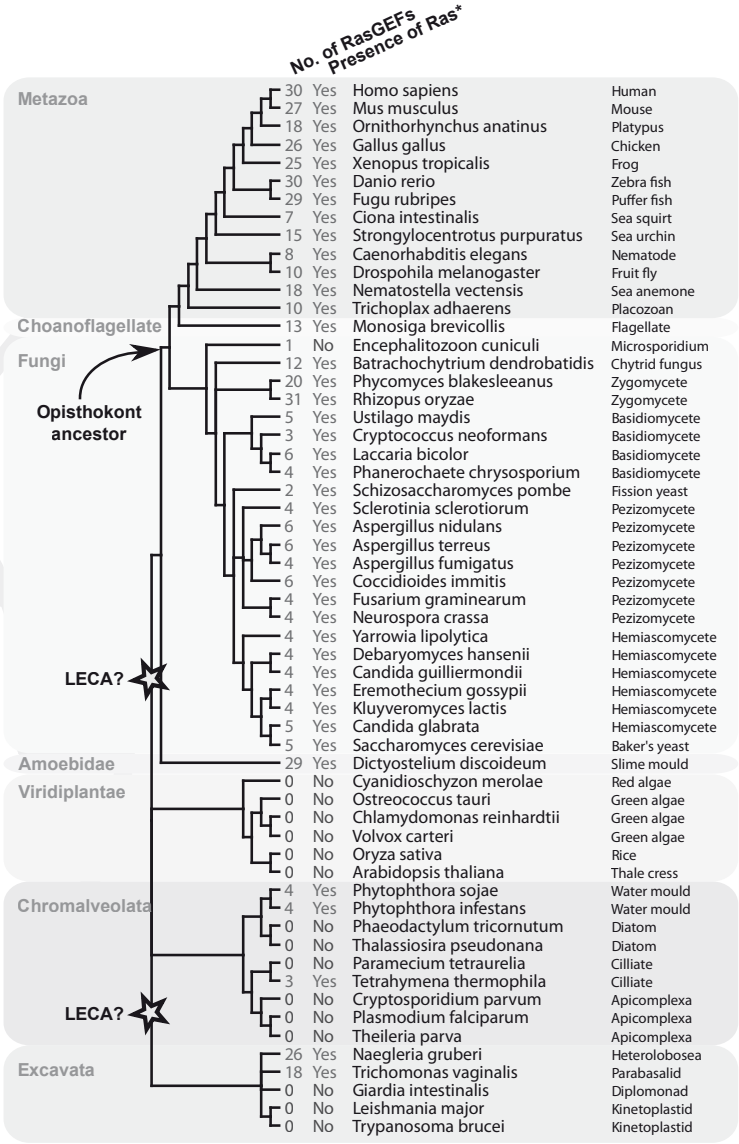


Figure 3.1 Species tree according to Simpson and Roger[10] and NCBI taxonomy database of the selected genomes with number of identified CDC25 HD containing proteins and occurrence of Ras* G-proteins. Possible alternative rooting of the eukaryotic tree of life are indicated with a star.

3.3 Results

3.3.1 Presence of CDC25 homology domain containing proteins in eukaryotic genomes

We have analyzed the genomes of a large variety of different eukaryotic species representing all major phyla for the presence of the CDC25 HD using a custom HMM model and PSI-BLAST. We identified 509 CDC25 HD coding sequences in 42 out of 58 genomes (Figure 3 1). In addition to genomes in which CDC25 HD was already identified

(i.e. animals and fungi), the CDC25 HD were found in oomycetes (*Phytophthora sojae* [5], *Phytophthora infestans* [6]), a ciliate (*Tetrahymena thermophila* [7]) and excavates (*Naegleria gruberi* [8], **Trichomonas vaginalis** [9]). As these divergent eukaryotes represent phyla probably diverged from the unikonts (animals, fungi and amoebozoa) at the eukaryotic root [10], the CDC25 HD was most likely present in the LECA (see Figure 3.1). This implies that the absence of Ras-like signaling in plants is the result of loss within this phylum.

The number of CDC25 HD coding sequences per genome is highly variable throughout species (7 to 30 in animals, see Figure 3.1). Complex animals, such as mammals, are characterized by a diverse and expanded RasGEF repertoire – they contain up to 30 CDC25 HD coding sequences. Interestingly, we observed similarly high numbers of CDC25 HD coding sequences in the early diverging fungus *Rhyzopus oryzae* [11] , the amoeba *Dictyostelium discoideum* [12] and the excavate *N. gruberi* [8]. These species have complex life cycles with divergent cell morphologies. The most notable are the amoeba *D. discoideum*, whose lifecycle includes both single cell and multi-cellular stages and *N. gruberi*, whose lifecycle includes amoeba and flagellate morphologies. It is therefore tempting to speculate that the CDC25 HD paralogy number might correlate with cellular lifestyle complexity rather than organismal complexity.

3.3.2 Co-occurrence of Ras* G-proteins and the CDC25 homology domain

Previously the origin of Ras* G-proteins has been firmly put in the LECA [4,13,14]. Since our results also suggest an early origin of the CDC25 HD we searched the genomes for co-occurrence of Ras* coding sequences and CDC25 HD coding sequences. We observed that species with a CDC25 HD coding sequence also have Ras* G-proteins and vice versa, with the sole exception of *Encephalitozoon cuniculi* (Figure 3.1). This microsporidium has a CDC25 HD coding sequence and thus presumably a RasGEF, but none of the small G-protein in *E. cuniculi* could be classified as a Ras* G-protein.

We observe that Ras* G-proteins and CDC25 HD containing proteins have a high degree of phylogenetic co-occurrence. Such high co-occurrence is not generally observed in eukaryotes [15] and is thus indicative of an unusually strong functional and evolutionary link between the CDC25 HD and Ras* proteins.

3.3.3 Phylogenetic reconstruction of CDC25 HD evolution

Due to high variability in the domain composition of RasGEFs in general, large parts of these proteins are not homologous and it is therefore impossible to make full-length alignments. Therefore, to determine how all CDC25 HD containing proteins are related and to time duplication and loss events we constructed multiple phylogenetic trees based on the CDC25 HD. Additionally, the use of only the CDC25 HD allows us to analyze the evolution of domain compositions independently and to infer when domains have been acquired or lost relative to the duplication and speciation events of the CDC25 HD.

We constructed phylogenetic trees from a multiple sequence alignment using various phylogenetic methods such as Maximum Likelihood and Neighbor Joining (see methods). We analyzed a total of three gene trees based on different methods by comparing them to the species phylogeny (tree of life) and annotated the nodes in the trees in terms of duplication and speciation events. This allows for a comparative genomics interpretation of the resulting large phylogeny. Species that can be used to

define orthologous groups for fungi and animals (the amoeba *D. discoideum*, the Phytophthora species *P. sojae*, *P. infestans*, the ciliate *T. thermophila*, and the excavata *N. gruberi* and *T. vaginalis*) where used as outgroups.

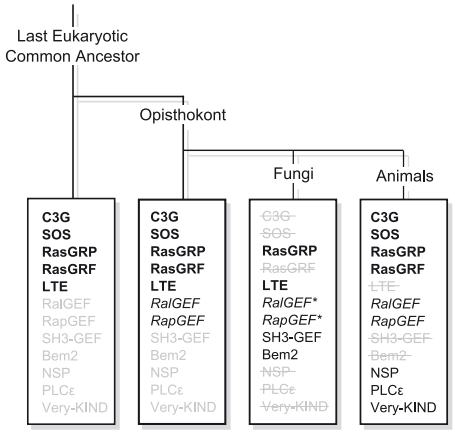


Figure 3.2 Evolutionary reconstruction of the CDC25 Homology Domain. The twelve classes of CDC25-HD are shown. The earliest origin of a class based on outgroup species is indicated by the font style: bold, originated in the Last Eukaryotic Common Ancestor; italics, originated in the Opisthokont and plain, originated either in fungi or animals. The earlier origins of several classes as suggested by phylogenetic tree topology and annotation are shown in gray. Classes that have been lost are stricken. Classes indicated with * have been lost in derived fungi (basidiomycetes and ascomycetes, but not in early derived fungi e.g. chytrids and zygomycetes).

We derived twelve fungal/animal orthologous groups of CDC25 HD coding sequences (Figure 3.2) by annotating the phylogenetic trees in terms of duplication and speciation events. We will refer to these orthologous groups as classes henceforth. Of the twelve classes, five (C3G, SOS, RasGRP, RasGRF and LTE) consistently contained outgroup species, suggesting with high certainty that the origin of these classes lies within LECA (Figure 3.2). Two classes (RalGEF and RapGEF) consist of fungal and animal sequences, strongly suggesting that these classes represent ancestral genes in the opisthokont ancestor (ancestor of animals and fungi). The remaining five classes contain either exclusively fungal sequences (SH3-GEF, Bem2) or animal sequences (NSP, PLCε and Very-KIND). However, since no close relationship with each of these five classes with any other class is observed and their connection to other classes suggests an origin by duplication that predates the LECA, we postulate here that these five classes are potentially older than either the fungal ancestor or animal ancestor. There is no evidence, such as domain composition or consistent proximity of clusters within the trees that indicate whether separate clusters should be merged.

By analyzing the phylogenetic trees we derived the gene family dynamics for each of the twelve classes (see the supplementary material for a phylogenetic description for each class). If the LECA indeed contained all 12 classes, it implies that not four, but nine classes have been lost in either the animal or fungal ancestor (Figure 3.2). In the case that the five classes which contain only fungal or animal sequences should have been merged into one or two classes, our results are affected only quantitively. In addition to loss in the animal and fungal ancestors we also find two classes (RalGEF and RapGEF) which contain animals and early diverged fungi, thereby also displaying a similar pattern of losses early in fungal evolution, although not immediately after the animal fungal split.

The selective loss of the CDC25 HD classes during animal and fungal evolution offers an explanation on the substantial difference in the RasGEF repertoire between fungi and animals. Our results complement the prevailing notion that in animals Ras* signaling has gained complexity largely by duplications of CDC25 HD coding sequences as our results

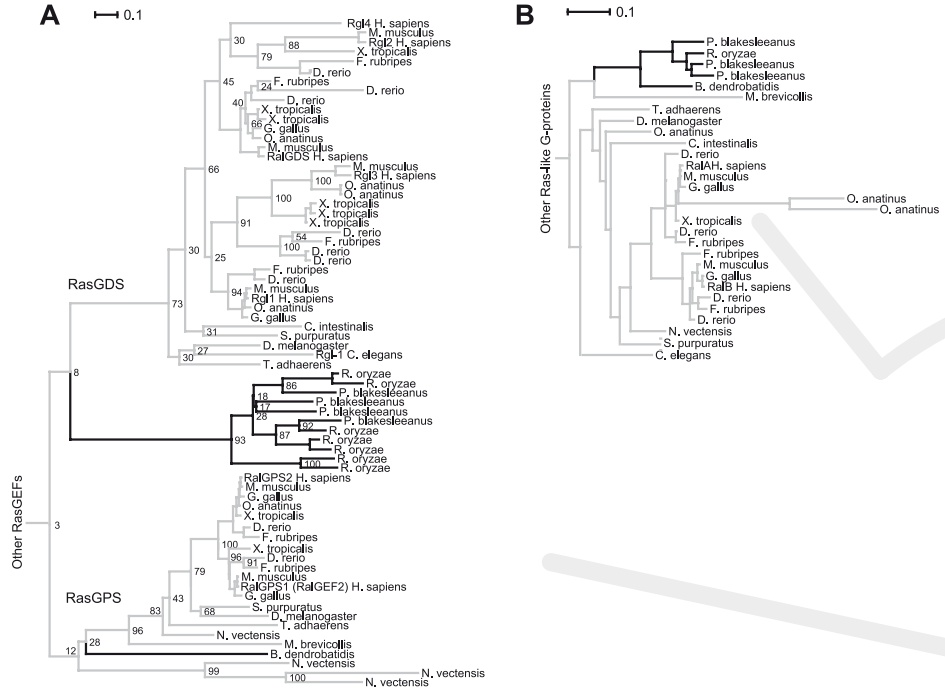


Figure 3.3 A) Sub tree of RasGEFs containing both types of RalGEFs: RalGPS and RalGDS. Zygomycota proteins (black lines) cluster consistently with RalGDS proteins while a *B. dendrobatides* gene (black lines) clusters with RalGPS. B) Subtree of Ras-like G-proteins showing Ral-like G-proteins. Fungal Ral-like proteins are indicated by black lines.

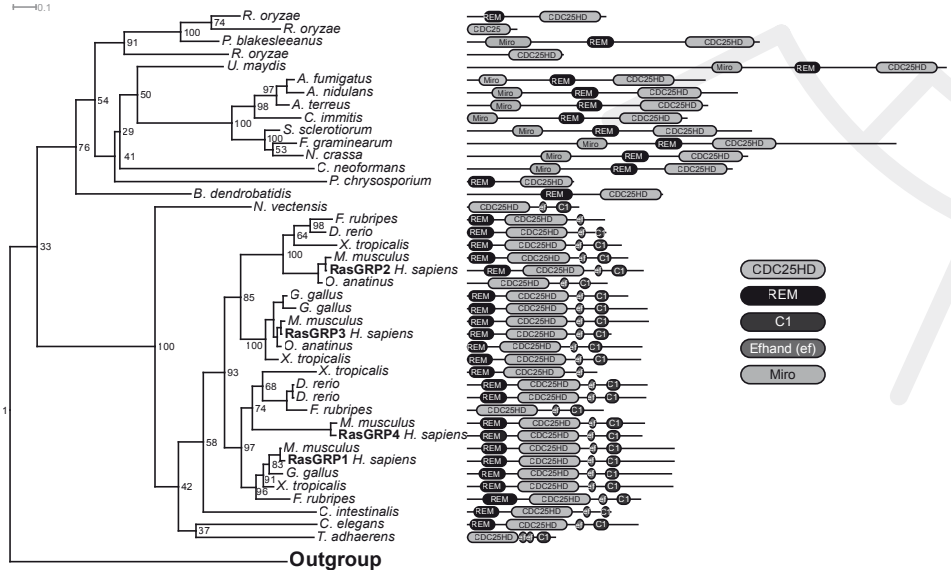


Figure 3.4 Subtree of the RasGRP class from the PhyML CDC25 HD phylogenetic tree and the domain compositions of each protein.

imply that fungi have greatly reduced their Ras* signaling complexity.

3.3.4 Ral signaling in fungi

The RalGEF class contains CDC25 HD coding sequences of RasGEF proteins that have unique specificity for the Ral proteins but not for the other Ras* proteins. The presence of orthologs of RalGEF in primitive fungi is fascinating because Ral proteins and the RalGEFs (RalGDS and RalGPS) were previously only identified in animals [16]. The phylogeny of the CDC25 HD displays a total of twelve fungal genes clustering consistently with RalGDS and RalGPS genes (Figure 3.3A). Eleven of these fungal genes belong to the Zygomycota fungal species *R. oryzae* and *Phycomyces blakesleeanus* and consistently cluster at the base of animal RalGDS-like genes but share no other characteristics, such as for example the Ras Association (RA) domains in animal RalGDS. A single gene of *B. dendrobatidis* (chytrid fungus) is positioned at the base of animal RalGPS-like genes and very interestingly shares the Pleckstrin Homology (PH) domain and the SH3 binding motif with the mammalian RalGPS. The phylogeny of Ras-like proteins shows that these fungal species also contain small G-proteins orthologous to animal Ral (Figure 3.3B). This strongly suggests that the fungal RalGEFs that cluster with animal RalGEFs are very likely to be functional RalGEFs. Moreover, it also shows that the origin of Ral signaling does not lie within the animal ancestor but in the opisthokont ancestor and that Ral signaling has been subsequently lost in derived fungi (i.e. the basidiomycetes and ascomycetes).

3.3.5 RasGEF domain compositions and domain shuffling

Based on our CDC25 HD phylogeny we next investigated the conservation of the domain composition of the various classes. Figure 3.4 shows the tree and the domain compositions of the RasGRP class, one of three orthologous groups that contains both animal and fungal genes (RasGRP, RalGEF and RapGEF). The animal RasGRP has a CDC25 HD, the Ras Exchange Motif (REM, also known as RasGEFN), a Protein kinase C conserved region 1 domain (C1) and EF-hand motifs, while the fungal orthologs have a Miro domain followed by the REM and CDC25 HD. This implies that the various domains were added to the CDC25 HD domain after the animal-fungal split. Similarly, in the RalGEF and RapGEF classes we find genes belonging to early diverged fungi displaying only the basic RasGEF structure (REM followed by the CDC25 HD) while the animal orthologs contain additional domains (e.g. cAMP binding domain, RA domain, PH domain) (see supplementary Figure S3.1-3). As indicated, the *B. dendrobatidis* RalGPS gene is unique in that it is the only fungal CDC25 HD containing ortholog whose domain composition is identical to its animal orthologs. Since additional domains play a critical role in the spatial and temporal control of GEFs [17] the difference in acquired domains between animal and fungal CDC25 HD containing proteins likely represent the evolving complexity of regulation.

The projection of the domain composition also reveals three potential cases of convergent evolution of domain composition in the RasGEF family (RA-, PH- and RhoGEF domains). For instance the RA domain is present in RalGDS-like proteins but also in Epac and PDZ-GEF proteins. Yet these genes do not cluster together in the phylogenetic trees and instead are separated by ancient duplications. Convergent evolution is also supported by the position of the RA domain in these classes: the RalGDS genes have the RA domain at the C-terminus while Epac and PDZ-GEFs have their RA domain N-terminally of the CDC25 HD. Indeed phylogenetic analysis of the RA domain reveals that the Epac and PDZ-GEF RA domains are more similar to human unconventional Myosin-9b and

to fungal Cyr1 (*S. cerevisiae*), an adenylate cyclase, than to RalGDS-like proteins (data not shown). This indicates that identical domains have been acquired independently in different RasGEF classes early in animal evolution.

The Bem2 gene in *S. cerevisiae* and its orthologs in other yeasts have a peculiar domain composition: the REM domain is located C-terminal of the CDC25 HD according to Pfam [18] and SMART [19] (Figure 3.5A), while it is normally located N-terminal to the CDC25 HD. However, by comparing the full length sequences to a database of single domains previously identified using HMMs (this is similar to the snipsel approach as used in the SMART database, see methods) we detected a second CDC25 HD C-terminal to the REM domain (Figure 3.5B). Similarly, the Bem2 ortholog in *R. oryzae* (an early diverged fungus) contains a second REM domain N-terminal to the first CDC25 HD (Figure 3.5C). The identification of a second CDC25 HD domain in yeast and early diverged fungi strongly implies that the Bem2-like genes in fungi were originally comprised out of two REM-CDC25 HD cassettes. Such a domain organization is not observed in any other known RasGEF and may suggest internal domain cassette duplication. However the phylogeny of the CDC25 HD that includes both CDC25 HDs of Bem2 is inconclusive: the two domains do not cluster together using RaxML or Quicktree and are intersected by few other sequences using PhyML.

The addition of regulatory domains to CDC25 HD containing proteins early in the evolution of animals and fungi represents the evolving complexity in the regulation of Ras* proteins. Additionally, with the identification of a tandem REM-CDC25 HD cassette in Bem2 we find that the CDC25 HD itself can be added to existing CDC25 HD containing proteins, potentially increasing the complexity of Ras* regulation by its GEFs even further.

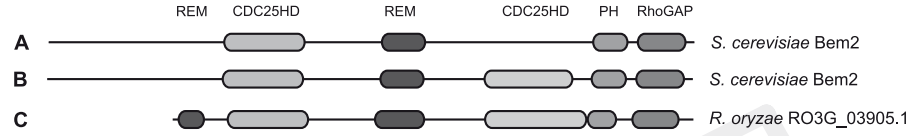


Figure 3.5 A) Domains as detected by the Pfam database in yeast Bem2. B) Our method for detecting domains resulted in the detection of an additional CDC25 domain in Bem2. C) An orthologous gene in *R. oryzae* reveals another REM domain N-terminal of the first CDC25 domain.

3.4 Discussion

Recently, many (draft) genomes of diverge eukaryotic species have become available allowing detailed analysis of the evolution of protein families. We have investigated the evolution of the regulation of Ras* signaling by reconstruction of the evolutionary history of the CDC25 HD, a domain that functions as GEF for Ras* proteins (Ras, Rap, Ral).

We made the following conclusions. First, coding sequences for the CDC25 HD can be found in animals, fungi and species belonging to the Chromalveolate and Excavate phyla, but not in plants. This indicates that the CDC25 HD most likely evolved in a common ancestor of all extant eukaryotic species but was subsequently lost in plants. Previously, Xu et al. [4] have described a Ras* G-protein in Trichmonas vaginalis and came to the conclusion that Ras* G-proteins originate in LECA. We found that nearly all species in our genome set that contain a CDC25 HD coding sequence also have a Ras* protein indicating an unusually strong functional and evolutionary link between the CDC25 HD and Ras* proteins. Our results thus extend on the observations by Xu et al. that in

addition to Ras* various CDC25 HD containing proteins were already present in LECA. We therefore conclude that Ras* signaling was already diverse and complex even before the radiation of all extant eukaryotic species. The observed Ras* signaling complexity in the LECA as well as the observed high numbers of CDC25 HD coding sequences in unicellular organisms exhibiting complex life styles might imply that Ras* signaling complexity evolved to serve complex processes in single cells rather than in complex organisms.

Secondly we divided the CDC25 HD coding sequences into twelve classes based on orthology relative to the common ancestor of fungi and animals. At least seven of these classes were found in animals, fungi or protist species, suggesting presence in LECA (C3G, SOS, RasGRP, RasGRF, LTE1) or at least prior to the animal-fungal split (RalGEF, RapGEF). Five classes are found only in animals (Very-KIND, PLCε, NSP) or in fungi (SH3-GEF, BEM2) suggesting a later origin (an animal or fungal ancestral gene). However, since no close relationship of each of these five classes is observed with any other class, we postulate that these five classes are older and were present in or before the Opisthokont ancestor.

Our third conclusion is that, based on the annotation of the phylogenetic trees in terms of duplication nodes and speciation nodes, the fungal and animal RasGEF repertoire has differentiated in large part by the loss of six ancestral CDC25 HD containing proteins in fungi and three in animals. The differential loss suggest that the substantial difference between fungal and animal Ras* signaling networks is not only due to gradual changes in the signaling networks but also to a great extend due to differential sampling of existent Ras* signaling pathways in the ancestor by fungi and animals. The large difference between fungal and animal Ras* signaling is further increased by the different domain compositions of orthologous fungal and animal RasGRP proteins.

Fourthly, our reconstruction uncovered Ral proteins and RalGEFs in Chytrids and Zygomycota (primitive or early diverged fungi). Previously, Ral signaling has been identified in animals but not for instance in *S. cerevisiae* or *Schizosaccharomyces pombe*. Our results show that this absence of Ral in yeast is due to loss of Ral signaling in higher fungi. This is surprising since one of the functions of Ral in mammals is the control of the exocyst complex [20] that is also present in yeast [16]. It would be interesting to know whether in the early fungi Ral also regulates the exocyst complex and why this regulation has been subsequently lost.

A fifth conclusion is that the introduction of addition domains to the REM-CDC25-HD cassette may have occurred in large part later during evolution and included convergent evolution for instance for the addition of the RA domain. These additional domains are largely responsible for the spatial and temporal control of the GEFs [3] which apparently needed adjustment later during evolution.

A final layer to the dynamics within the evolution of Ras* signaling pathways is added by the observed changeability in Ras* G-protein specificity. Although RalGEFs are exclusive activators for Ral, for the classes RasGRP and SH3-GEFs it is known that their members individually can activate Ras or Rap G-proteins or both [21,22]. The positions of these genes within the CDC25 HD phylogeny suggest that G-protein specificity is not fixed in evolution but can change by small modifications of the sequence and post translational modifications to the protein.

The evolutionary dynamics of the CDC25 HD provides us with a more complete picture of the evolution of Ras* signaling pathways. We observe not only duplications and

losses but we also gain insight into how the pathway was wired. We can see differential sampling from the ancestor by fungi and animals, which substantially determined the shape of fungal and animal Ras signaling pathways respectively. Together with the radical change in domain composition in the RasGRP class this might be indicative for a need to change the Ras signaling pathways to accommodate the different environments or life styles of fungi and animals.

3.5 Methods

3.5.1 Genomes

We have acquired a number of genomes from the string database, version 7 [23]. In addition, we have acquired best model protein sequences of the remaining genomes from their respective project download sites. For full overview of genomes and their source, see supplementary material table 1.

3.5.2 Genome search and CDC25 domain identification

The SMART RasGEF Hidden Markov Model (HMM) was used to initially search 30 genomes. All CDC25 HD sequences with an E-value lower than 1e-5 were gathered. All gathered sequences were aligned using Muscle 3.6 [24]. Bad aligning sequences or sequences which introduce large insertions in the alignment were removed and the remaining sequences were re-aligned. A custom HMM profile was created from this alignment using hmmbuild and hmmcalibrate of the HMMER package version 2.3.2 [25]. The HMMer profile is provided as supplementary data. The genome protein dataset was then searched using the custom CDC25 HD HMM profile. All proteins with E-value of 1e-5 or lower were collected and the sequences with E-values above 1e-05 were manually checked for significant sequence similarity using local PSI-BLAST.

3.5.3 Multiple sequence alignment and phylogenetic tree construction

All gathered protein sequences were aligned using MAFFT [26] with the --localpair option. As many gaps were introduced by only a few sequences, positions were discarded where less than 10% of the sequences had a residue. Phylogenetic trees were constructed using Maximum Likelihood (PhyML [27] and RaxML [28]) and Neighbor joining (Quicktree [29]). For PhyML the WAG model for amino acid substitution was used with a discrete gamma model using 6 categories with estimated gamma parameter and estimated proportions of invariable sites. The bootstrap analysis is based on 100 iterations. RaxML was run using the PROTGAMMAIWAG model and 100 iterations for bootstrap analysis was performed with 1688 as seed number. The phylogenetic trees were analyzed by manual annotation in terms of duplication, loss and speciation using the species tree (see Figure 3.1) bootstrap values and domain compositions. The alignment and the phylogenetic trees (in newick format) are provided as supplementary data.

3.5.4 Domain evolution analysis

The domain compositions of the RasGEF sequences were analyzed by projecting domain structures of the RasGEF sequences onto the phylogenetic trees. Firstly, domains were identified in the full length RasGEF sequences by using the hmmpfam program of the HMMer package and the Pfam_ls collection of HMM profiles from the Pfam database

using the provided gathering cutoff values. Secondly, insignificant Pfam hits were called 'true' when the hsp in a blast all vs. all RasGEFs returned a hit (E-value < 1) within a protein with a significant Pfam hit of the same domain model. Thirdly, all RasGEF sequences where blasted against a sequence database containing sequences of all previously called domains in step 1 and 2 using a cut-off value of 1e-5. Hits which overlap more than 20% with Pfam hits were excluded. A custom Perl script was used to draw domain structures in a phylogenetic tree in Scalable Vector Graphics format. Sequences were analyzed manually where needed, to detect domains which are not detected using the described method but for which contextual evidence exists that they should be present.

3.5.5 Phylogenetic analysis of Ras-like G-proteins

The genome protein dataset was searched using the Pfam HMM profile for Ras. All sequences with a bit score above zero where selected. Due to high similarity of Ras to other small G-protein sequences the HMM profile will not distinguish between the various small G-protein super families and this step is therefore very inclusive. In order to select only the Ras-like G-protein subfamily members an alignment of all selected protein sequences was made using MAFFT with the "--localpair" option. A neighbor joining tree was constructed with the Quicktree program and analyzed. All known Ras-like sub family members were contained in one clade which did not contain proteins of the other sub families. All sequences in this clade were selected and human RhoA/B/C sequences were added as out-group. These selected sequences were aligned using MAFFT with the "--localpair" option. To construct the final phylogenetic tree PhyML was used with a discrete gamma model using four categories with estimated gamma parameter and estimated proportions of invariable sites.

3.6 Acknowledgements

We like to thank Jos Boekhorst, Gabino Sanchez-Perez, Like Fokkens and Michael Seidl for their valuable input and support. This work is part of the BioRange program of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI). The sequence data of selected genomes were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/, in collaboration with the user community and the Fungal Genome Initiative of the Broad Institute. For a full overview of the genomes see Supplementary Table 1.

3.7 Supplementary material

Description of the 12 RasGEF Classes from the phylogenetic analysis.

3.7.1 SOS

The SOS (Son of Sevenless) class is represented by animal and outgroup sequences but not fungi, indicating loss in early fungal evolution. SOS has duplicated in the vertebrate ancestor.

3.7.2 RasGRF

The RasGRF (Ras Guanine nucleotide-Releasing Factor) class, similar to SOS is also represented by animal and outgroup sequences but does not contain fungal sequences.

The RasGRF genes have a similar domain composition as SOS genes (RhoGEF, Ph, REM, and CDC25HD, though N-terminus differs). This might indicate that SOS and RasGRF proteins are related but based on the phylogeny of the CDC25 domain this appears not to be the case. RasGRF has also duplicated in the vertebrate ancestor.

3.7.3 C3G

A third class contains animal C3G (also known as RapGEF1) and outgroup sequences and based on the position of the outgroup sequences also represents an ancestral gene in the ancestor of animals and fungi. Like SOS and RasGRF, the group of C3G genes also lacks fungal genes.

3.7.4 RasGRP

The Ras Guanyl Releasing Protein (RasGRP) class contains both fungal, animal and outgroup sequences, but the fungal sequences do not share the same domain composition as the animal orthologs (see main text section 3.5). Apparently fungal RasGRP has been lost in hemiascomycota (yeasts). RasGRP has undergone at least two rounds of duplication in the vertebrate ancestor, resulting in four vertebrate RasGRPs.

3.7.5 RalGEFs

Ral Guanine nucleotide Dissociation Stimulator (RalGDS) and Ral Guanine nucleotide exchange factor with Ph domain and Sh3 binding motif (RalGPS) each represent an ancestral gene in the opisthokont as they both contain animals and fungal sequences. Because the clustering of RalGDS and RalGPS is never intersected with a gene from the outgroup species and these are both RalGEFs we suggest that the RalGDS and RalGPS ancestral genes originate from a duplication in an ancestor of animals and fungi, but before the last common ancestor of animals and fungi. Because the observed relation between RalGPS and RalGDS we have combined them into a single class named RalGEFs.

The RalGEFs cluster with outgroup sequences in the PhyML tree. In the RaxML tree the RalGEFs clusters with the first domain of Bem2 and outgroup sequences. In the Quicktree NJ tree RalGEFs cluster with RasGRF including outgroup sequences. Because of the inconsistent clustering with other classes and the presence of outgroup sequences in these clusters we postulate that the RalGEFs are older than the ancestor of animals and fungi and might have been present in LECA.

3.7.6 RapGEFs

Similar to RalGEFs the Epac and PDZ-GEF genes share a single ancestral gene. Epac and PDZ-GEF genes share a Ras Association (RA) domain and Cyclic AMP binding domains. Genes belonging to zygomycota and chytrid fungi cluster together with Epac and PDZ-GEF indicating that Epac and PDZ-GEF arose from a duplication of an animal ancestral gene after the split with fungi. The Epac/PDZ-GEF genes cluster consistently with another animal group of RapGEFs represented by human RasGEF1A, B and C. This group of genes is never intersected by genes from any of the outgroup species, though it is intersected by the NSP cluster in the RaxML tree. Also, RasGEF1, Epac and PDZ-GEF genes are reported to be Rap specific GEFs. Both RasGEF1 and Epac/PDZ-GEF represent ancestral genes in the last common ancestor of animals and fungi and both originate from a single ancestral gene in the common ancestor before the last common ancestor

of animals and fungi. Similar to the RalGEFs we have combined Epac, PDZ-GEF and RasGEF1 genes into a single class named RapGEFs. The RapGEFs are clustered twice with LTE including outgroup sequences (PhyML, RaxML) but cluster together with a group of outgroup sequences in the Quicktree NJ tree which include *N. gruberi*, *D. discoideum* and *T. vaginalis*. This suggests that although the RapGEF class does not have consistent clustering with outgroup sequences itself, it might be older than the Opisthokont ancestor and might have been present in LECA.

3.7.7 PLCε and very-KIND

PLCε and very-KIND are two classes which are not well defined in the phylogenetic trees or even within animals, but because they do not share domain composition and are not clearly related to each other or any of the other RasGEF classes we postulate that each originated from a single ancestral gene in the common ancestor of animals and fungi.

3.7.8 SH3-GEFs

CDC25, SDC25 and Bud5 (*S. cerevisiae*) are part of a large group of fungal genes characterized by a SH3 domain at the N-terminus of the protein (SH3-GEF class). CDC25 and SDC25 are recent duplicates in yeasts and coincide with the yeast Whole Genome Duplication. CDC25 and SDC25 are distinct from Bud5 in that CDC25 and SDC25 are RasGEFs while Bud5 is a GEF for Bud1 which is a Rap-like G-protein. Nearly all fungi have orthologs for both Bud5 and for CDC25/SDC25 suggesting a duplication event early within fungal evolution. It appears that the duplication early in fungal evolution might coincide with a change in specificity between the early paralogs. The SH3 class represents a single ancestral gene in the ancestor of animals and fungi but has subsequently been lost in animals.

3.7.9 LTE

A cluster of fungal and outgroup sequences which contain LTE1 in *S. cerevisiae* also represents an ancestral gene in the ancestor of animals and fungi and appears to be lost in animals, similar to the SH3-GEF cluster.

3.7.10 NSP (BCAR and SHEP1)

A gene cluster of animal genes represented by human NSP1, NSP2 (BCAR) and NSP3 (SHEP1) clusters inconsistently within the trees of all three phylogenetic methods. It is associated with fungal Bem2 and a collection of dissimilar (i.e. they display different domain architectures) fragmented fungal genes (PhyML), is placed within the Epac/PDZ-GEF class (RaxML) or as a separate clade with genes from outgroup species. It is therefore unclear what the evolutionary history of the NSP genes is.

3.7.11 Bem2

Similarly to animal NSP genes, Bem2 is not well defined by the three phylogenetic methods. The fungal group itself is well defined and seems to be lost in pezizomycota only. Clear phylogenetic relations to animal RasGEFs is missing but these genes cluster together with largely fragmented and degenerate RasGEF sequences including animal NSP genes. To complicate matters, we have detected two CDC25 domains within each of these fungal sequences (see the main text). Although we suspect these domains to have

arisen from internal domain duplication it is not unambiguously supported by each of the phylogenetic methods. Like the NSP class it is unclear what the evolutionary history of the Bem2 genes is.

Table S3.1. Genomes used in this study

Species	Genome version	Genome source	References
Homo sapiens	Version 7	STRING DB	[21]
Mus musculus	Version 7	STRING DB	[21]
Ornithorhynchus anatinus	Version 46	ENSEMBL	[28]
Gallus gallus	Version 46	ENSEMBL	[28]
Xenopus tropicalis	Version 4.1	JGI	[8]
Danio rerio	Version 46	ENSEMBL	[28]
Fugu rubripes	Version 4	JGI	[29]
Ciona intestinalis	Version 2	JGI	[30]
Strongylocentrotus purpuratus	Spur 2.1, Sep 2006	SpBase	[31]
Drosophila melanogaster	Version 7	STRING DB	[21]
Caenorhabditis elegans	Version 7	STRING DB	[21]
Nematostella vectensis	Version 1	JGI	[32]
Trichoplax adhaerens	Version 1	JGI	[33]
Monosiga brevicollis	Version 1	JGI	[34]
Encephalitozoon cuniculi	11/24/2001	Genoscope	[35]
Batrachochytrium dendrobatidis	3/1/2007	BROAD	[36]
Rhizopus oryzae	5-Apr	BROAD	[37]
Phycomyces blakesleeanus	Version 1	JGI	[8]
Cryptococcus neoformans	1/7/2005	SGTC and TIGR	[38]
Laccaria bicolor	Version 1	JGI	[39]
Phanerochaete chrysosporium	Version 1	JGI	[40]
Ustilago maydis	3-May	BROAD	[41]
Schizosaccharomyces pombe	10/25/1996	Sanger	[42]
Yarrowia lipolytica	7/2/2004	Génolevures	[43]
Debaryomyces hansenii	7/2/2004	Génolevures	[43]
Candida guilliermondii	4-Sep	BROAD	[44]
Eremothecium gossypii	3/6/2004	Zoologisches Institut der Univ. Basel, Switzerland	[45]
Kluyveromyces lactis	7/2/2004	Génolevures	[43]
Candida glabrata	7/2/2004	Génolevures	[43]
Saccharomyces cerevisiae	10/25/1996	SGD	[46]
Aspergillus terreus	5-Aug	BROAD	[47]
Aspergillus nidulans	2005	BROAD	[48]
Aspergillus fumigatus	7-Jan	TIGR	[49]
Coccidioides immitis	Release 2 Aug 2005	BROAD	[50]
Fusarium graminearum	Release 2 Nov 2003	BROAD	[51]
Neurospora crassa	25/04/05	BROAD	[52]
Sclerotinia sclerotiorum	5-Apr	BROAD	[53]
Dictyostelium discoideum	Version 2.5	"DictyBase	[54,55]"
Phytophthora sojae	Version 1	JGI	[5]

Species	Genome version	Genome source	References
Phytophthora infestans	Version 1	BROAD	[6]
Phaeodactylum tricornutum	Version 2	JGI	[56]
Thalassiosira pseudonana	Version 3	JGI	[57]
Paramecium tetraurelia	Version 1.2	Genoscope	[58]
Tetrahymena thermophila	4-Aug	TIGR	[7]
Cryptosporidium parvum	7/1/2005	CryptoDB	[59]
Plasmodium falciparum	2002	PlasmoDB	[60]
Theileria parva	2005	TIGR	[61]
Cyanidioschyzon merolae	8/4/2004	National Institute of Genetics, Japan	[62]
Ostreococcus tauri	6-Jan	Laboratoire Arago, France	
Chlamydomonas reinhardtii	Version 3	JGI	[63]
Volvox certeri	Version 1	JGI	[8]
Oryza sativa	12/18/2002	International Rice Genome Sequencing Project	[64]
Arabidopsis thaliana	12/14/2000	Arabidopsis Information Resource (TAIR)	[65]
Giardia intestinalis		GiardiaDB	[66]
Leishmania major	Version 5.2	Sanger	[67]
Trypanosoma brucei	Version 4, 05/2008	Sanger	[68]
Trichomonas vaginalis	1	TrichDB	[66]
Naegleria gruberi	Version 1	JGI	[8]

3.8 References

1. Bos JL (1989) ras oncogenes in human cancer: a review. Cancer Res 49: 4682-4689.
2. Wennerberg K, Rossman KL, Der CJ (2005) The Ras superfamily at a glance. J Cell Sci 118: 843-846.
3. Bos JL, Rehmann H, Wittinghofer A (2007) GEFs and GAPs: critical elements in the control of small G proteins. Cell 129: 865-877.
4. Xu MY, Liu JL, Zhang RL, Fu YC (2007) Isolation of a novel ras gene from Trichomonas vaginalis: a possible evolutionary ancestor of the Ras and Rap genes of higher eukaryotes. Biochem Cell Biol 85: 239-245.
5. Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RHY, et al. (2006) Phytophthora Genome Sequences Uncover Evolutionary Origins and Mechanisms of Pathogenesis. Science 313: 1261-1266.
6. Phytophthora infestans Sequencing Project. Broad Institute of Harvard and MIT (http://www.broad.mit.edu).
7. Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, et al. (2006) Macronuclear Genome Sequence of the Ciliate Tetrahymena thermophila, a Model Eukaryote. PLoS Biology 4: e286.
8. These sequence data were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ in collaboration with the user community.
9. Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, et al. (2007) Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis. Science 315: 207-212.
10. Simpson AG, Roger AJ (2004) The real 'kingdoms' of eukaryotes. Curr Biol 14: R693-696.
11. Rhizopus oryzae Sequencing Project. Broad Institute of Harvard and MIT (http://www.broad.mit.edu).
12. Eichinger L, Pachebat JA, Glockner G, Rajandream MA, Sucgang R, et al. (2005) The genome of the social amoeba Dictyostelium discoideum. Nature 435: 43-57.
13. Jekely G (2003) Small GTPases and the evolution of the eukaryotic cell. Bioessays 25: 1129-1138.
14. Dong JH, Wen JF, Tian HF (2007) Homologs of eukaryotic Ras superfamily proteins in prokaryotes and their novel phylogenetic correlation with their eukaryotic analogs. Gene 396: 116-124.
15. Fokkens L, Snel B (2009) Cohesive versus flexible evolution of functional modules in eukaryotes. PLoS Comput Biol 5: e1000276.
16. Moskalenko S, Henry DO, Rosse C, Mirey G, Camonis JH, et al. (2002) The exocyst is a Ral effector complex. Nat Cell Biol 4: 66-72.
17. Ponsioen B, Gloerich M, Ritsma L, Rehmann H, Bos JL, et al. (2009) Direct spatial control of Epac1 by cAMP. Mol Cell Biol.
18. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, et al. (2008) The Pfam protein families database. Nucleic Acids Res 36: D281-288.

19. Schultz J, Milpetz F, Bork P, Ponting CP (1998) SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A 95: 5857-5864.

20. Lipschutz JH, Mostov KE (2002) Exocytosis: the many masters of the exocyst. Curr Biol 12: R212-214.

21. Quilliam LA, Rebhun JF, Castro AF (2002) A growing family of guanine nucleotide exchange factors is responsible for activation of Ras-family GTPases. Prog Nucleic Acid Res Mol Biol 71: 391-444.

22. Camus C, Boy-Marcotte E, Jacquet M (1994) Two subclasses of guanine exchange factor (GEF) domains revealed by comparison of activities of chimeric genes constructed from CDC25, SDC25 and BUD5 in Saccharomyces cerevisiae. Mol Gen Genet 245: 167-176.

23. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, et al. (2007) STRING 7--recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 35: D358-362.

24. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.

25. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755-763.

26. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30: 3059-3066.

27. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696-704.

28. Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21: 456-463.

29. Howe K, Bateman A, Durbin R (2002) QuickTree: building huge Neighbour-Joining trees of protein sequences. Bioinformatics 18: 1546-1547.

30. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, et al. (2009) Ensembl 2009. Nucleic Acids Res 37: D690-697.

31. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, et al. (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297: 1301-1310.

32. Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, et al. (2002) The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. Science 298: 2157-2167.

33. Cameron RA, Samanta M, Yuan A, He D, Davidson E (2009) SpBase: the sea urchin genome database and web site. Nucleic Acids Res 37: D750-754.

34. Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, et al. (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317: 86-94.

35. Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, et al. (2008) The Trichoplax genome and the nature of placozoans. Nature 454: 955-960.

36. King N, Westbrook MJ, Young SL, Kuo A, Abedin M, et al. (2008) The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. Nature 451: 783-788.

37. Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, et al. (2001) Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. Nature 414: 450-453.

38. Batrachochytrium dendrobatidis Sequencing Project. Broad Institute of Harvard and MIT (http://www.broad.mit.edu).

39. C. neoformans Genome Project, Stanford Genome Technology Center, funded by the NIAID/NIH under cooperative agreement AI47087, and The Institute for Genomic Research, funded by the NIAID/NIH under cooperative agreement U01 AI48594.

40. Martin F, Aerts A, Ahren D, Brun A, Danchin EG, et al. (2008) The genome of Laccaria bicolor provides insights into mycorrhizal symbiosis. Nature 452: 88-92.

41. Martinez D, Larrondo LF, Putnam N, Gelpke MD, Huang K, et al. (2004) Genome sequence of the lignocellulose degrading fungus Phanerochaete chrysosporium strain RP78. Nat Biotechnol 22: 695-700.

42. Ustilago maydis Sequencing Project. Broad Institute of MIT and Harvard (http://www.broad.mit.edu).

43. Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, et al. (2002) The genome sequence of Schizosaccharomyces pombe. Nature 415: 871-880.

44. Sherman DJ, Martin T, Nikolski M, Cayla C, Souciet JL, et al. (2009) Genolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes. Nucleic Acids Res 37: D550-554.

45. Candida guilliermondii Sequencing Project. Broad Institute of Harvard and MIT (http://www.broad.mit.edu).

46. Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, et al. (2004) The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. Science 304: 304-307.

47. Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, et al. (1997) Genetic and physical maps of Saccharomyces cerevisiae. Nature 387: 67-73.

48. Aspergillus Comparative Genome Project. Broad Institute of Harvard and MIT (http://www.broad.mit.edu).

49. Galagan JE, Calvo SE, Cuomo C, Ma LJ, Wortman JR, et al. (2005) Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. Nature 438: 1105-1115.

50. Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, et al. (2005) Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus. Nature 438: 1151-1156.

51. Coccidioides immitis Sequencing Project. Broad Institute of Harvard and MIT (http://www.broad.mit.edu).

52. Cuomo CA, Guldener U, Xu JR, Trail F, Turgeon BG, et al. (2007) The Fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization. Science 317: 1400-1402.

53. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, et al. (2003) The genome sequence of the filamentous fungus Neurospora crassa. Nature 422: 859-868.

54. Sclerotinia sclerotiorum Sequencing Project. Broad Institute of Harvard and MIT (http://www.broad.mit.edu).

55. Fey P, Gaudet P, Curk T, Zupan B, Just EM, et al. (2009) dictyBase--a Dictyostelium bioinformatics resource update. Nucleic Acids Res 37: D515-519.

56. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, et al. (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature 456: 239-244.

57. Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, et al. (2004) The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. Science 306: 79-86.

58. Aury JM, Jaillon O, Duret L, Noel B, Jubin C, et al. (2006) Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature 444: 171-178.

59. Heiges M, Wang H, Robinson E, Aurrecoechea C, Gao X, et al. (2006) CryptoDB: a Cryptosporidium bioinformatics resource update. Nucleic Acids Res 34: D419-422.

60. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, et al. (2009) PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res 37: D539-543.

61. Gardner MJ, Bishop R, Shah T, de Villiers EP, Carlton JM, et al. (2005) Genome sequence of Theileria parva, a bovine pathogen that transforms lymphocytes. Science 309: 134-137.

62. Nozaki H, Takano H, Misumi O, Terasawa K, Matsuzaki M, et al. (2007) A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga Cyanidioschyzon merolae. BMC Biol 5: 28.

63. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, et al. (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science 318: 245-250.

64. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, et al. (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res 35: D883-887.

65. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, et al. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res 36: D1009-1014.

66. Aurrecoechea C, Brestelli J, Brunk BP, Carlton JM, Dommer J, et al. (2009) GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens Giardia lamblia and Trichomonas vaginalis. Nucleic Acids Res 37: D526-530.

67. Sequence data for L. major chromosome X was obtained from The Sanger Institute website at http://www.sanger.ac.uk/Projects/L_major/. Sequencing of L. major chromosome X was accomplished as part of the Leishmania Genome Network with support by The Wellcome Trust.

68. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, et al. (2005) The genome of the African trypanosome Trypanosoma brucei. Science 309: 416-422.

CHAPTER 4

Evolution of the TOR pathway

Teunis J.P. van Dam, Fried J.T. Zwartkruis, Johannes L. Bos and Berend Snel

Manuscript in preparation

4.1 Abstract

The TOR kinase is a major regulator of growth in eukaryotes. Many components of the TOR pathway are implicated in cancer and metabolic diseases in humans. Analysis of the evolution of TOR and its pathway may provide fundamental insight into the evolution of growth regulation in eukaryotes and provide a practical framework on which experimental evidence can be compared between species. Here we performed phylogenetic analyses on the components of the TOR pathway and determined their point of invention. We find that the two TOR complexes and a large part of the TOR pathway originate from before the Last Eukaryotic Common Ancestor and form a core to which new inputs have been added during animal evolution. In addition we provide an insight in how duplications and subfunctionalization of the S6K, RSK, SGK and PKB kinases shaped the complexity of the TOR pathway. In yeast we identify novel AGC kinases that are orthologous to the S6 kinase. These results demonstrate how a vital signaling pathway can be both highly conserved and flexible in eukaryotes.

4.2 Introduction

Target of rapamycin (TOR) is a major regulator of growth in eukaryotes [1]. It integrates both intracellular signals that depend on nutrient availability and extracellular signals such as growth factors [2-4]. Therefore, dysfunction of TOR or other proteins in the

TOR pathway is involved in organismal and cancer development in mammals [5]. The TOR protein is a kinase that participates in two distinct protein complexes. The TOR Complex 1 (TORC1) promotes translation by phosphorylating the S6 subunit of the ribosomal complex via S6K and by phosphorylating 4E-BP, which then dissociates from eIF4E allowing translation [6,7]. TORC1 is activated by the Rheb G-protein, which in its turn is regulated by the TSC1/2 complex [8,9]. The best characterized activation route of the TORC1 in animals is the insulin signaling pathway, which includes well characterized oncogenes and tumor suppressors such as PI3K, PDK1, PTEN and PKB (also known as AKT) [5,10-12]. It is not fully understood how TORC1 is regulated by nutrient levels [13], although it is known that rag GTPases play an important role in amino acid regulated TORC1 activity [14,15,13]. The TOR Complex 2 (mTORC2) regulates cytoskeleton rearrangement in response to growth. The regulation of TORC2 is unknown but TORC2 positively regulates TORC1 via PKB [16]. In addition, negative feedback from S6K lowers PKB activity [17].

TOR is a conserved kinase and has been functionally characterized in animals [18], fungi (yeast) [19] and plants [20]. However, not all of the TOR complex subunits or TOR pathway components are equally conserved. For instance, the yeast *Saccharomyces cerevisiae* has retained the Rheb G-protein [21], but lacks the Rheb regulatory genes TSC1 and TSC2, whereas *Schizosaccharomyces pombe* has retained all three genes [22]. Plants appear to lack Rheb and the TSC1/2 complex altogether [23,15] despite the presence of TORC1 [24,20].

To obtain novel insights into the evolution of the TOR pathway we performed phylogenetic analyses on the components of the TOR pathway, starting from the mammalian TOR pathway. We are specifically interested in the distribution of each subunit over all eukaryotes and the point of invention of single genes or even modules in evolution. To that end we searched in 64 diverse eukaryotic genomes for homologous and orthologous genes for components of the mTOR pathway. The recent publication of genomes of highly divergent eukaryotic organisms such as *Naegleria gruberi* [25] now permits us to make detailed phylogenetic analyses of genes that play a key role in the mTOR pathway and determine the earliest possible point of invention in evolution for each subunit. We analyzed each evolutionary or phylogenetic profile in light of distinct components of the pathway (e.g. the mTOR complexes, Rheb regulation) and the pathway as a whole.

Importantly, we found that TOR, all subunits of TORC1 and 2 and a large part of the TOR pathway components form an evolutionary core. New regulatory inputs, such as insulin and TNFα signaling, have been added to the core TOR pathway during animal evolution. We show that TORC1 and TORC2 appear to behave as independent evolutionary modules, even though the majority of the subunits are shared between the two complexes. We infer the presence of a large common evolutionary core, including Rheb and TSC2, in the Last Eukaryotic Common Ancestor (LECA), the last ancestral eukaryote that gave rise to all current eukaryotic species. We provide an in-depth phylogenetic analyses of TOR, Rheb, TSC2 and TCTP as well as other components of the TOR signaling pathway. In addition, we reveal the remarkable role of the duplications of an ancestral AGC kinase that gave rise to the S6K, RSK, PKB and SGK kinases, in the increasing complexity of the TOR pathway. We conclude that a vital signaling pathway can be both highly conserved and flexible in eukaryotes.

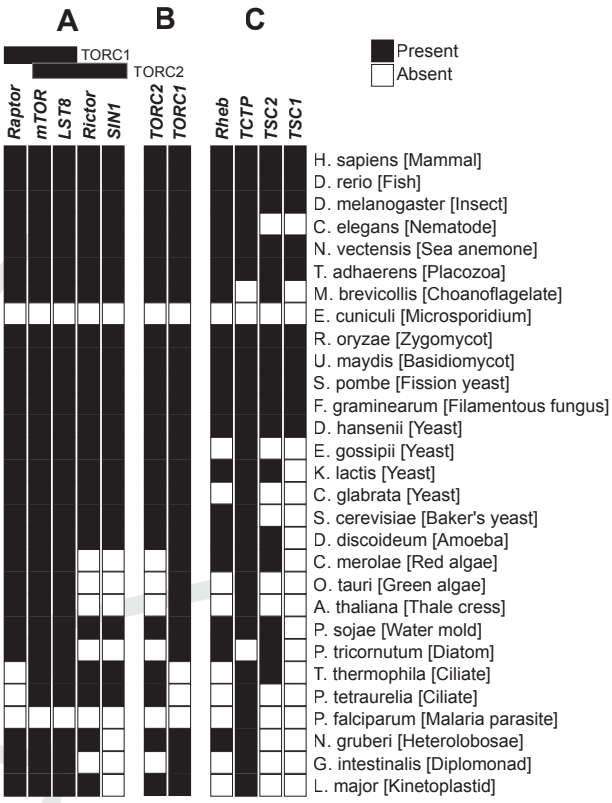


Figure 4.1 Absence/presence plots in a subset of 65 eukaryotic genomes. A) The absence and presence plot of the individual mTOR complex subunits. B) Translation of (A) into an absence and presence plot of the TOR complexes. To the left of the mTORC subunits is shown to which complex each subunit belongs. Animals and fungi have both TORC1 and TORC2 while plants have only TORC1 and ciliates have only TORC2. Apparently it is possible to lose either one of the complexes while maintaining the other. C) The absence and presence plot of Rheb, TSC1 & 2 and TCTP. The GAP domain of TSC2 is well conserved and is found (with few exceptions) in species that also contain Rheb throughout the eukaryotic lineages. TSC1 is an animal/fungal invention and therefore newer than TSC2 and Rheb. The occurrence of TCTP in species lacking Rheb and vice versa, raises additional doubt on the debated Guanine Exchange Factor function of TCTP.

4.3 Results and discussion

4.3.1 The evolution of TOR complex 1 and 2 are decoupled

TOR functions as part of two distinct protein complexes: the TORC1 and TORC2 complexes [26]. In mammalian cells TORC1 contains mTOR, LST8 and Raptor, while TORC2 contains mTOR, LST8, Rictor and SIN1 (Figure 4.1A). TOR and LST8 are both present in genomes in all major eukaryotic lineages and therefore form the evolutionary core of the TOR complexes (Figure 4.1A). In addition we also observe that TOR and LST8 co-occur with either Raptor (TORC1) or Rictor (TORC2) or both, indicating that both TOR complexes are old and were likely present in LECA, the common ancestor of all current eukaryotic species.

We find TORC1 together with TORC2 in all major lineages, except plants, which possess only TORC1 (Figure 4.1B). Interestingly we detect TORC2, but not TORC1 in the ciliates *Tetrahymena thermophila* and *Paramecium tetraurelia*. It therefore appears that the

two distinct TOR complexes are decoupled in evolution as either one can be lost while the other is maintained.

We do not detect any of the TOR Complex subunits in the microsporidium E*ncephalitozoon cuniculi* and the apicomplexa *Plasmodium falciparum*, *Cryptosporidium parvum* and *Theileria parva,* indicating at least two independent loss events for TOR signaling in eukaryotes. All four species are intracellular parasites and have reduced genomes and cellular structure. Host-parasite interactions might have replaced the function of TOR in these organisms as growth of the parasite is directly linked to conditions in the host cell.

Distinct protein complexes that share subunits (i.e. hyperlinks) provide a selective reason for maintaining duplicate copies of these shared subunits [27]. Therefore, we could expect to find duplications of TOR in some species, resulting in a dedicated TOR for each of the two TOR complexes. Indeed we find duplications of TOR in *Saccharomyces cerevisiae*, *Candida glabrata*, *Schizosaccharomyces pombe*, *Batrachochytrium dendrobatidis*, *Populus trichocarpa*, *Emiliania huxleyi*, *Trypanosoma brucei*, *Leishmania major*, *Phytophthora infestans* and *Phytophthora sojae* (see Figure S4.1). The duplication of TOR in *S. cerevisiae* and *C. glabrata* likely originates from the whole genome duplication event. In *S. cerevisiae* both TOR1 and TOR2 can be part of TORC1, while TOR2 is specific for the TORC2 [26], which indicates that the two TOR duplicates have not yet reached complete functional divergence. The TOR duplication in *S. pombe* and *B. dendrobatidis* are lineage-specific duplications and occurred independently from each other and the duplications in *S. cerevisiae* and *C. glabrata*. While we have no functional descriptions for *B. dendrobatidis*, it has been shown for *S. pombe* that TOR1 and TOR2 function as part of TORC1 [28], while TOR1 is specific for TORC2 [29]. (Note that the naming of TOR1 and TOR2 in *S. cerevisiae* and *S. pombe* can cause confusion as the genes resulted from independent duplication events, the naming of TOR1 and TOR2 in both yeasts does not reflect one-to-one orthologous relationships but is based on order of discovery [30].) Surprisingly, LST8 has not been duplicated in any of the species examined. This raises the question as to why the two hyperlinks TOR and LST8 behave differently in evolution. We hypothesize that duplication and subsequent functional divergence of LST8 may have implications for the structural integrity of the two TOR complexes, while minor modifications to the TOR duplicate genes increased functional divergence without compromising complex stability.

4.3.2 TSC2-Rheb signaling, a highly conserved signaling route to TORC1

The Rheb G-protein, conserved throughout the eukaryotic lineage

The Rheb G-protein is one of the major regulators of TOR activity in animals and directly regulates the activity of TORC1 but not TORC2 [16]. Rheb is a Ras-like small GTPase and the sequences of small GTPases are highly conserved [31,32]. We therefore reconstructed the phylogeny of the Ras-like small GTPases. From this phylogeny we identified Rheb orthologs (see methods section and Figure S4.2) and derived a phylogenetic profile of orthologs (Figure 4.1C). We identified Rheb orthologs in all animals and fungi (except in *C. glabrata*, *Eremothecium gossypii* and *E. cuniculi*). Additionally we identify orthologs in distantly related organisms such as diatoms, oomycetes, the amoeba *Dictyostelium discoideum*, the heterolobosida *Naegleria gruberi* and the red alga *Cyanidioschyzon merolae*. To our knowledge, the presence of a Rheb ortholog in red

algae is the first time a G-protein belonging to the Ras-like sub-family of small GTPases has been identified in the Archaeplastida (i.e. plants, red and green algae).

The identification of Rheb orthologs in distantly related species strongly suggests that Rheb originated in or before LECA. We observe Rheb orthologs in species that also contain TORC1, which indicates that the regulation of TOR by Rheb is strongly conserved. However, we do not observe the opposite, e.g. species that have TORC1 do not necessarily have a Rheb ortholog. The most notable of these species are the green algae and plants, but also the yeasts *C. glabrata* and *E. gossypii*. The presence of a Rheb ortholog in *C. merolae* indicates a loss of Rheb in the ancestor of the green algae and plants.

TSC1 is an animal-fungal innovation in TSC2 regulation of Rheb

Next we investigated the phylogeny of the only known regulator of Rheb, the GTPase activating protein (GAP) TSC2. TSC2 integrates many inputs such as MAPK/Ras signaling via RSK1 [33], Wnt signaling via GSK3β [34] and insulin signaling via PI3K and PKB [35] in animals. The TSC2 GAP domain occurs in all major eukaryotic phyla, except the excavates, indicating it is much older than previously suggested [36] and likely originated in or before the LECA. We find the TSC2 GAP domain orthologs in species that also have a Rheb ortholog, including the red algae *C. merolae* mentioned above. The only exception is the ciliate *T. thermophila*. Therefore we predict that the TSC2 GAP orthologs will regulate the Rheb orthologs in *D. discoideum*, *C. merolae* and the *Phytophthora* species.

Furthermore, we find that while TSC1 orthologs are always observed together with TSC2 orthologs in the same genomes, TSC2 can be found on its own in additional eukaryotic species (Figure 4.1C). Interestingly in some of these species (*D. discoideum*, *C. merolae*, *P. infestans*, *P. sojae*, *Phaeodactylum tricornutum*) we were able to identify the GAP domain but not the Tuberin domain that is necessary to dimerize with TSC1. Strikingly, we find TSC1 orthologs in animals and fungi, which is the same phylogenetic distribution as the tuberin domain of TSC2. Therefore it is likely that TSC1 itself and the ability of TSC2 to dimerize with TSC1 via the tuberin domain are inventions in the Opisthokont ancestor (i.e. in the animal and fungal ancestor).

The absence of the TSC1/2 complex in *C. elegans* and *S. cerevisiae* (Figure 4.1C) suggests caution should be taken when comparing regulatory mechanisms of TOR between these two and other species. For instance, regulatory mechanisms for Rheb and TOR discovered in animals, such as in *D. melanogaster* do not necessarily hold for *C. elegans* and vice versa.

Translationally Controlled Tumor Protein 1, no phylogenetic linkage with Rheb

TCTP (also known as TPT1) has been reported to be the Guanine Exchange Factor (GEF) for Rheb [37] but this function has been debated by us and others [38,39]. Hence it is interesting to study the phylogenetic profiles of TCTP and TSC2 and compared them to Rheb (Figure 4.1C). We constructed the phylogenetic profile of TCTP and found that it is well conserved in many eukaryotic species, including plants, chromalveolata and excavata. Interestingly we observe eukaryotic species which have a TCTP ortholog but do not have the Rheb G-protein (i.e. green plants and algae, apicomplexa, ciliates, *C. glabrata*, *E. gossypii*, see Figure 4.1C) and vice versa (i.e. the choanoflagellate *M. brevicollis* and the diatom *P. tricornutum*). In addition we observe in the apicomplexa (i.e. a group of unicellular intracellular parasites including *Plasmodium falciparum*), that

even though a TCTP ortholog is present, Rheb and TOR orthologs are abscent.

In a previous study on Ras Guanine Exchange Factors (RasGEFs) we observed a strong evolutionary link between the RasGEFs and their respective Ras-like GTPases [40]. The RasGEF functional domain, the CDC25 homology domain, and the Ras, Rap and Ral GTPases were found to be present or absent together in a diverse set of eukaryotic genomes. In contrast we observe no such strong presence/absence pattern for TCTP and Rheb. It therefore seems that there is no evolutionary linkage between TCTP and Rheb.

Further doubts about TCTP's GEF activity for Rheb arise from experimental evidence of TCTP function in *Arabidopsis thaliana* when put in an evolutionary framework. Berkowitz et al. [41] studied the function of the ortholog of TCTP in *A. thaliana*. They found that TCTP acts as an important regulator of growth, implying a function for TCTP in TOR activity, which resembles the situation in animals [37]. However *Arabidopsis* does not have a Rheb ortholog (Vernoud et al. [15] and this study). Berkowitz et al. postulate that *Arabidopsis* TCTP regulates another GTPase (either a Rhop- or Rab-like G-protein) that might function equivalent to Rheb in plants. However, the function of TCTP in *A. thaliana* in absence of a Rheb ortholog and our observation of an evolutionary 'mismatch' between Rheb and TCTP do not support TCTP as a RhebGEF, but instead suggest that TCTP regulates TOR via an alternative route.

4.3.3 Evolution of the mammalian TOR pathway; gaining inputs

We extended our phylogenetic study to include upstream and downstream components of the mammalian TOR pathway and thereby put Rheb and TORC in a wider biological context. We focus on the mammalian TOR pathway because TOR signaling in animals and particularly mammals is the most extensively studied intact TOR pathways (e.g. not lacking key components such as the TSC1/2 complex in *S. cerevisiae* or TORC2 in *A. thaliana*). The mTOR pathway was assembled from literature to reflect current consensus. For each protein we constructed the phylogenetic profile and determined the point of invention (e.g. the age of a protein; see Table S4.2 for all phylogenetic profiles). We depict the age of each protein along the metazoan natural history towards LECA in the representation of the mTOR pathway (Figure 4.2).

Very recently, Serfontein et al. [36] published an evolutionary survey on the components of the TOR pathway in a representative selection of eukaryotic genomes. Our results concerning the evolution of TSC1/TSC2-TOR pathway underlines select observations made by Serfontein and coworkers but differ considerably in others. While Serfontein and co-workers find that the evolutionary "core" of the pathway that was present in LECA consisted out of the TORC1 complex (TOR, LST8, Raptor), AMPK, PI3K and PTEN and S6K, we find that TORC2 (TOR, LST8, Rictor, SIN1), Rheb, TSC2, PDK1 and the remaining AGC kinases PKB, RSK and SGK are also part of this evolutionary core. We show that Rheb and TSC2 have not been "bolted on" in evolution but are infect part of the evolutionary core that originated in or before LECA. We do however observe other regulatory elements of the TOR pathway that have been added at a later evolutionary stage. We made three observations concerning the evolution of new TOR signaling regulation.

The first observation is that the regulation of TOR activity by insulin is an animal-specific addition to the pathway and may be linked to the necessity in animals for growth regulation on a multi cellular scale. The second observation is that the more recently invented TSC1 introduces novel regulatory input onto the Rheb-TOR cascade (i.e. FIP200

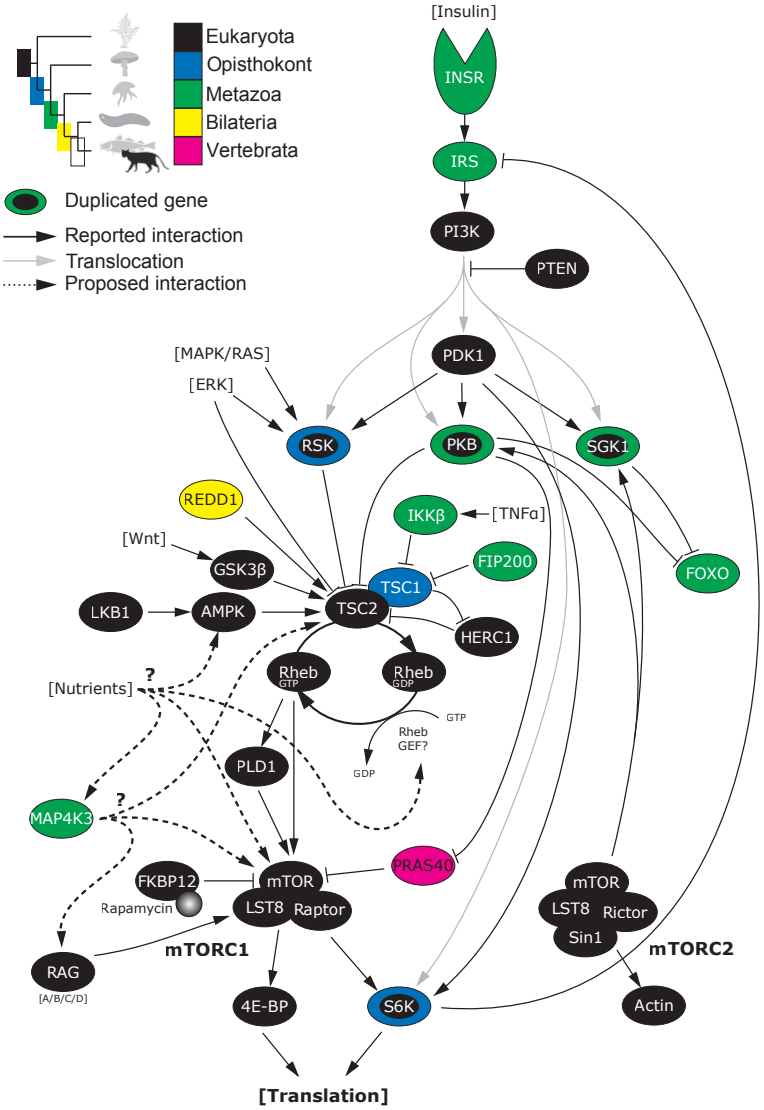


Figure 4.2 Evolution of the mTOR pathway. We reconstructed the mTOR pathway based on literature. The coloring indicates the age for each gene based on a reconstruction from phylogenetic profiles. For duplicate genes the color of the outer circle denotes the time of duplication and the inner circle denotes the origin of the complete orthologous group. The regulation of TOR via insulin and TNFα are animal specific additions onto an ancient TOR pathway. The invention of TSC1 in the animal and fungal ancestor allowed for new regulatory inputs onto TSC2. Duplications of an ancestral AGC kinase that gave rise to S6K, RSK, PKB and SGK played a significant role in the evolution of the TOR pathway.

and IKKβ/TNFα). The third observation is that PRAS40 is a vertebrate invention and therefore an uncommonly late addition to the TOR pathway. This suggests that although Rheb-TSC1/2 route is an ancient functioning route, new routes like PRAS40 can be added to PKB regulation of mTORC1. These three observations suggest that although TOR signaling is highly conserved within eukaryotes, it is also flexible enough to accept new inputs and can be adapted to suit new environments (e.g. multi cellular tissues).

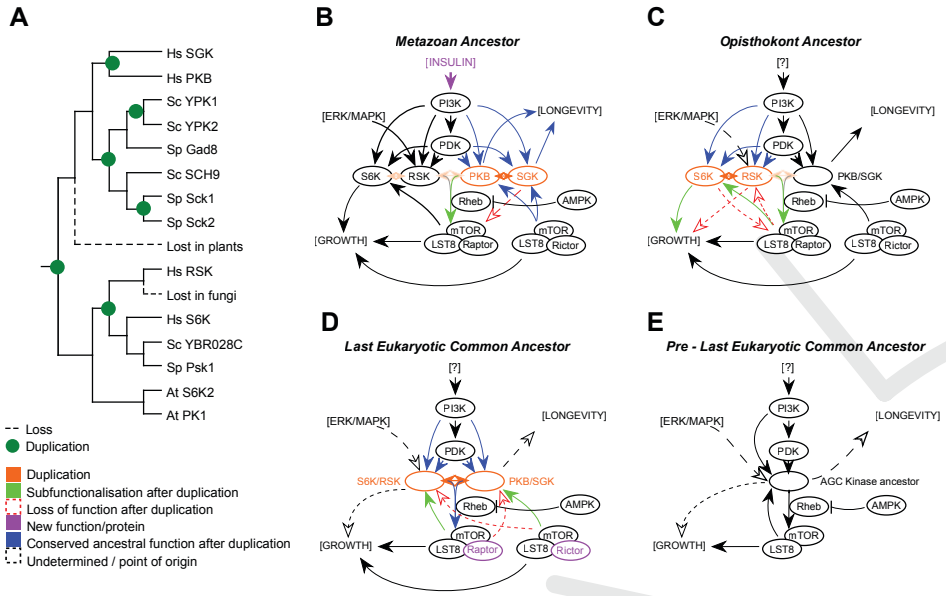


Figure 4.3 Evolutionary reconstruction of the ancestral TOR signaling pathway based on the evolutionary reconstruction of the AGC kinase ancestral genes of S6K, RSK1, PKB and SGK1. A) Simplified representation of the phylogenetic tree of the AGC kinases (see Figure S4.3). Species indication: Hs *Homo sapiens*, Sc *S. cerevisiae*, Sp *S. pombe*, At *A. thaliana.* B) Reconstruction of TOR signaling in the animal ancestor. An AGC kinase duplicated in the ancestor of animals and choanoflagellates to give rise to PKB and SGK. SGK lost the ability to inhibit TSC2. New signaling inputs where invented in the animal ancestor, among which is the insulin signaling. C) TOR signaling in the ancestor of animals and fungi. An AGC kinase duplicated in the ancestor of animals and fungi (Opisthokonta) to give rise to S6K and RSK. The duplication was followed by subsequent subfunctionalization of ancestral functions between the two paralogs D) TOR signaling in LECA. The duplication of the ancestral AGC kinase that give rise to the RSK-S6K and PKB-SGK ancestral precursor genes corresponds to the differential activation by either TORC1 or TORC2 respectively. E) Reconstruction of pre-LECA TOR signaling. The shared subunits TOR and LST8 of both TOR complexes suggest that before LECA there was at one point only one proto-TOR complex. As the AGC kinases also share single ancestry we can reduce the complexity of the TOR pathway in pre-LECA even further.

Blenis et al. [10] suggested that PRAS40 might represent a conserved PKB regulation route to mTORC1 while TSC1/2 represented a newer additional pathway in higher eukaryotes because TSC1 and TSC2 were previously not found in lower eukaryotes like *S. cerevisiae*. However due to the availability of many newly sequenced eukaryotic genomes and with more extensive phylogenetic profiling as done here, it becomes evident that PRAS40 represents a new vertebrate specific additional route of PKB-mTORC1 activation and that the Rheb/TSC1/TSC2 route is in fact much older.

Serfontein and coworkers [36] have detected PKB orthologs only in animals, amoebozoa and excavate species, but not in plants and chromalveolates. However they have included only those sequences that included both the kinase domain and a Pleckstrin Homology (PH) domain that is characteristic of the animal PKB. However, the PH domain is known to be very promiscuous in eukaryotes [42] and is therefore not suitable to use as a restriction criterion for orthology.

We find that the AGC kinases PKB and SGK are paralogs that have arisen from a duplication event in the animal and choanoflagellar ancestor, but we also find PKB/SGK orthologs in chromalveolates and excavates. PKB and SGK are closely related to the AGC kinases S6K

and RSK, all involved in TOR signaling. This suggests that the AGC kinases and duplication events play a striking and complex evolutionary role in the TOR pathway. We therefore focused on the evolution of the AGC-kinases.

4.3.4 Duplication of AGC kinases has increased internal TOR pathway complexity

In the mTOR pathway, the AGC family kinases S6K, RSK1, PKB and SGK1 are located both upstream and downstream of mTOR. S6K and RSK1 arose from a duplication event in the ancestor of animals and fungi (Opisthokont ancestor) while PKB and SGK1 arose from a duplication event in the ancestor of animals and the closely related choanoflagellate *Monosiga brevicollis* [43] (see Figure 4.3A and Figure S4.3). The S6K-RSK and PKB-SGK ancestral genes themselves have arisen from a gene duplication in or before LECA. The evolutionary relation between S6K, RSK1, PKB and SGK1 make it uniquely possible to reconstruct the evolution of their regulatory interactions within the TOR pathway. In Figure 4.3 we have reconstructed the TOR pathway at several points in evolution based on events in the evolution of the AGC kinases and experimental characterization of orthologous genes in *H. sapiens*, *S. cerevisiae*, *S. pombe* and *A. thaliana*.

Duplication of the PKB-SGK1 ancestral gene in the ancestor of animals and choanoflagellates

The PKB and SGK genes duplicated from a single ancestral gene in the filozoan ancestor, i.e. the ancestor of animals and choanoflagellates (see Figure 4.3A and Figure S4.3). However, additional new components have been invented specifically in animals, such as the insulin and TNFα signaling pathways. Therefore we reconstructed the TOR pathway in the metazoan ancestor. PKB and SGK are both activated by PI3K, PDK1 and the TORC2 complex (the blue edges in Figure 3B) [44-48]. It is therefore very likely that the PKB/SGK ancestral protein was also activated by PI3K, PDK1 and TORC2 ancestral proteins.

PKB inhibits the TSC1/2 complex by phosphorylating TSC2 [49,50] and inhibits FOXO transcription factors by directly phosphorylating them [51]. SGK also inhibits FOXO by phosphorylation [52] but has not been reported to phosphorylate TSC2. We can partly derive the ancestral functions by comparing PKB and SGK functions to the gene functions of the co-orthologous genes in *S. pombe* and *S. cerevisiae*.

In *S. pombe* and *S. cerevisiae* there are three genes co-orthologous to both PKB and SGK (Sck1, Sck2 and Gad8 in *S. pombe*, YPK1, YPK2 and SCH9 in *S. cerevisiae*, see Figure S4.3). The *S. pombe* and *S. cerevisiae* PKB/SGK1 orthologs Sck1, Sck2, Gad8 and SCH9 are implied to have function in oxidative stress responses and aging [53,54], similar to PKB and SGK1 in animals. Therefore, the role of PKB and SGK in regulating longevity is conserved and likely an ancestral function in the Opisthokont (Figure 4.3C).

In contrast to stress response and aging, the origin of TSC2 phosphorylation by PKB is not immediately apparent. In *S. cerevisiae* TSC2 has been lost and we have been unable to find any references that implicate the *S. pombe* PKB/SGK orthologs Sck1, Sck2 in growth regulation via TOR (Sck1/2) or that Gad8 has been associated with the TSC1/2 complex. This makes it difficult to determine if TSC2 phosphorylation by PKB is an ancestral function or has been newly acquired. Nevertheless, there are similarities in function of PKB with RSK, and we can therefore reconstruct the ancestral function of PKB and SGK by comparing their functions to their paralogs RSK and S6K.

Similar to PKB, RSK also inhibits the TSC1/2 complex by phosphorylating TSC2 in mammals [55]. Because the GAP domain of TSC2 is conserved throughout the eukaryotic lineage, the most plausible scenario is that PKB and RSK inhibition of TSC2 is an ancestral function of the PKB-SGK-S6K-RSK ancestral gene (henceforth we will refer to this ancestor as the ancestral AGC kinase for brevity). In this scenario the TSC2 regulation by PKB is a subfunctionalization of the ancestral gene, e.g. an ancestral function that is maintained by one paralog and lost in the other (the green edge and red dashed edge in Figure 4.3B).

Interestingly, similar to the filozoan PKB-SGK duplication event, fungi seem to have undergone a similar duplication event of the ancestral PKB-SGK kinase. YPK1 and YPK2 function in *S. cerevisiae* can be rescued by rat SGK, but not mouse PKB or rat S6K [56]. Therefore, although the PKB-SGK and YPK-SCH9 duplication events in animals and fungi occurred independently, the resulting animal and fungal paralogs appear to have evolved in a functionally similar way.

Duplication of the S6K-RSK1 ancestral gene in the animal and fungal ancestor

The S6K and RSK genes duplicated from a single ancestral gene in the fungal and animal ancestor (Figure 4.3C). Like PKB and SGK, both are regulated by PI3K and PDK1 (the blue edges in Figure 4.3C).

However S6K and RSK are not regulated by the TORC2 complex [57]. S6K is regulated by the TORC1 complex instead, while RSK is regulated via MAPK signaling. While it is very likely that the S6K-RSK ancestral protein was regulated by PI3K and PDK1 ancestral proteins, the regulation of S6K by TORC1 and RSK by MAPK initially obscures whether the ancestral kinase was either activated by TORC1 or MAPK or both.

We can infer whether the ancestral protein was activated by TORC1 and or MAPK by inferring ancestral function from experimental evidence for TOR signaling in the plant *A. thaliana*. RSK1 and S6K are co-orthologs to the *A. thaliana* S6K kinase. The S6K kinase of *A. thaliana* is regulated by the *A. thaliana* TOR complex [6], which is the same as TORC1 in other organisms, and we can therefore infer that the S6K-RSK ancestral protein was likely activated by TORC1. Thus S6K regulation by TORC1 is a subfunctionalization from the ancestral gene (the green edge in Figure 4.3C from TORC1 to S6K), i.e. RSK has lost the regulation by TORC1 (the red dashed edge in Figure 4.3C from TORC1 to RSK). Activation by MAPK of the ancestral protein cannot be inferred from the *A. thaliana* S6K as there is no published link between *A. thaliana* MAPK and *A. thaliana* TOR signaling but we cannot exclude loss of this function in plants (the dotted edge from MAPK in Figure 3C).

Previously we deduced that RSK regulation of mTORC1 activity via TSC2 is likely an ancestral function from the symmetry with PKB. Therefore TSC2 regulation by RSK is an ancestral function that has been lost in S6K (the green edge in Figure 4.3C from RSK to TORC1 and the red dashed edge in Figure 4.3C from S6K to TORC1).

Intriguingly, both *S. pombe* and *S. cerevisiae* have a one-to-one ortholog with S6K that has not been fully characterized yet (see Figure 4.3A and Figure S4.3). The *S. cerevisiae* locus YBR028C codes for a kinase, but has not been reported on in literature. The S6K ortholog in *S. pombe*, psk1, has been reported to be involved in phenylarsine oxide resistance and disruption of the psk1 gene did not result in growth defects [58]. Instead, the S6K-like cellular function in *S. cerevisiae* has been ascribed to SCH9 [59], which from our analysis is an ortholog of mammalian PKB and SGK. SCH9 shows that the AGC kinases

are capable of performing cellular functions that have been ascribed to their paralogs, possibly increasing the complexity of the roles the AGC kinases play in TOR signaling. Therefore, given that the poorly characterized YBR028C and psk1 are clearly orthologs of S6K, we suggest that there is a substantial role for these two genes to be uncovered in TOR signaling.

Back to the root: the ancestral AGC kinase and the ancestral TOR pathway

Above we have described the ancestral states of the S6K-RSK and PKB-SGK ancestral genes. We observe symmetric functions between the two ancestral genes and therefore we are able to (partly) reconstruct the ancestral AGC kinase (e.g. the ancestral gene of S6K, RSK, PKB and SGK). All four kinases are regulated by PDK1 and PI3K and we can therefore infer that the ancestral AGC kinase was also regulated by the PDK1 and PI3K ancestral genes (the blue edges in Figure 4.3D from PI3K and PDK1). Above, we also deduced that the S6K-RSK and PKB-SGK ancestral genes possibly regulated TORC1 activity via the TSC2 ancestral gene as it is a shared function of both RSK and PKB (the blue edges in Figure 4.3D from the AGC kinases to TORC1).

Phosphorylation of the S6 ribosomal subunit by the S6K-RSK ancestral kinase cannot be reconstructed beyond the LECA, because the PKB and SGK kinases do not share this function and we are therefore unable to determine if the S6 activation was a function of the ancestral AGC kinase that has been lost by the PKB-SGK ancestral gene or that it is an acquired function of the S6K-RSK ancestral gene (the dotted edge from S6K/RSK in Figure 4.3D).

Our phylogenetic reconstruction suggest the existence of a TORC1 and a TORC2 complex in LECA that functions in conjunction with Rheb/TSC2 to activate at least two distinct AGC kinases (the green dashed edges in Figure 4.3D). These AGC kinases arose from duplication and thus required (partially) independent regulation by TOR for their sub-functionalization (compare Figure 4.3E and the red and green dashed edges in Figure 4.3D). TORC1 and TORC2-specific proteins like Raptor and Rictor most likely contributed situation-specific activation of AGC kinases, i.e. determine context-relevant downstream outputs.

The positive feedback loop that emerges in the reconstruction of the ancestral TOR pathway in and before LECA (Figure 4.3D and E) imposes a network structure that is undesirable as the feedback loop could easily result in a constitutively activated TOR and therefore uncontrolled growth. In animals negative feedback from S6K and mTOR to PKB or more upstream elements in the insulin pathway have been documented [60,17,61]. It is very likely, given the importance of proper regulation of the TOR pathway that negative feedback routes were also in place in LECA. However we are unable to reconstruct these negative feedbacks back to LECA. An explanation for this could be that the negative feedbacks in the TOR pathway have been subject to change in evolution. Subsequent duplications of the ancestral AGC kinase and subfunctionalization however might have provided the opportunity to increase the possibility for more precise or additional regulation of TOR activity.

4.3.5 Flexibility in a conserved signaling pathway

The TOR pathway is an universal regulator of cell growth in eukaryote species. TOR is the master regulator and integrates many inputs such as growth signals and nutrient

availability in a cell. We show that the two TOR complexes are highly conserved and originated in the last common ancestor of all eukaryotes. We analyzed the origin and evolution of each subunit of the TOR complexes separately as well as other components of the TOR signaling pathway. We show that TORC1 and TORC2 behave as separate evolutionary modules that can be individually lost (i.e. loss of either Raptor (TORC1) or Rictor (TORC2) or both complexes as a whole). We find that the TOR pathway has a conserved "core" to which new inputs have been added early in animal evolution, such as insulin and TNFα signaling. We also find evidence that the core itself has been extensively modified in evolution by duplications of ancestral AGC kinases that gave rise to S6K, RSK, SGK and PKB. The evolution of TOR and conversely the whole TOR pathway demonstrates that a vital signaling pathway, can be both highly conserved and flexible in eukaryotes and can be adapted to fulfill changing requirements of growth regulation by eukaryotic organisms.

4.4 Methods

4.4.1 Genome selection

We acquired best model protein sequences of 64 divergent eukaryotic species from EnsEMBL [62], JGI, the Broad institute or their respective genome project sites. We have selected a wide range of animal and fungal genomes as most research on TOR signaling is being done in either animal or fungal model organisms. We also included a wide range of genomes belonging to other major phyla, such as the archeaplastida, chromalveolates and excavates, to be able to accurately time the origin of each TOR pathway component. For a full overview of genomes, source and version information see Table S4.1. The species tree as used in this chapter can be found in Appendix A.1.

4.4.2 Phylogenetic analyses

Orthology determination and phylogenetic profiles

Orthology was determined automatically by applying MCL [63] on InParanoid [64] species-species comparisons. For InParanoid an all vs. all BLAST [65] was performed on the whole set of 64 genomes with the options –p blastp –m 8 –v 1000000 –b 1000000. The BLAST results were then split into separate data files as required for InParanoid. InParanoid was applied to all possible combinations of the 64 genomes using the default settings (score_cutoff = 50, outgroup_cutoff = 50, seq_overlap_cutoff = 0.5, conf_cutoff = 0.05, group_overlap_cutoff = 0.5, grey_zone = 0). A matrix dataset for MCL was constructed from the InParanoid analyses by constructing edges between genes for each InParanoid cluster taking the lowest confidence value of both genes. Edges were only drawn between genes of different species. The MCL analysis was run with parameters –abc –I 1.5 –write-graph.

Phylogenetic profiles were constructed from the InParanoid-MCL clusters by determining which species were (not) represented in each cluster. The phylogenetic profiles of genes of interest were manually verified.

Rheb orthology identification and phylogenetic analysis

The sequences of the selected genomes were searched using the Pfam [66] HMM profile for the Ras family (Pfam accession PF00071.12, Pfam version 23) and hmmsearch of the

HMMER package [67] version 2.3.2. All sequences with a bitscore larger than 0 were selected. Due to the high sequence similarity of Ras to other small GTPases many other small GTPases are included in this set. An alignment of all sequences was made using the MAFFT program [68] with the –globalpair option. A neighbor joining tree was constructed using the Quicktree program [69]. A subtree was selected which contained all Ras-like subfamily members but no other small GTPases. The sequences were gathered from the initial alignment as manual inspection of the alignment produced from the subset showed it was suboptimal to the initial alignment. Subsequently a phylogenetic tree was constructed over all Ras-like subfamily members using RAxML [70] (-T 4 -x 488761235 -f a -N 1000 -m PROTGAMMAIWAG). From the resulting tree a sub tree was selected that contained Rheb and as many genes from the searched species as possible without including other known small GTPases. All phylogenetic trees were visualized using Dendroscope [71].

The AGC kinases SGK1, PKB, RSK1 and S6K

Protein sequences belonging to the cluster of orthologous groups that contains S6K, RSK, PKB and SGK, were aligned using MAFFT [68] --globalpair. The resulting alignment was analyzed and a segment that showed high sequence similarity between all sequences was excised (positions 893 to 2050, corresponding to the kinase domain and PKC terminal domain). A Phylogenetic tree, including bootstrap analysis, was constructed using RAxML [70] (-T 8 -x 78382369 -f a -N 1000 -m PROTGAMMAIWAG).

TSC2 ortholog identification

TSC2 orthologs were initially identified using the automated orthology determination as described above. Closer examination revealed that orthology was mainly based on the GAP domain sequence. Since the GAP domain of TSC2 belongs to a larger family of GAP domains, the Ran/RapGAP domain family, we used phylogenetic methods to faithfully determine true orthology based on the RapGAP domain sequences. We gathered RapGAP domain sequences from the sequence set by using a custom made HMM model and hmmsearch of the HMMER package [67] version 2.3.2. The domain sequences were aligned using MAFFT [68] –globalpair. A phylogenetic tree was build using RAxML [70] (-T 6 -x 23421421 -f a -N 1000 -m PROTGAMMAIWAG).

4.5 Acknowledgements

We would like to thank Jos Boekhorst, Gabino Sanchez-Perez, Like Fokkens, Michael Seidl for their help in performing the analyses and support. We would especially like to thank Prof. Michael Hall and Thomas Sturgill for their suggestions and critical appraisal. The sequence data of selected genomes were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/, in collaboration with the user community, or the Fungal Genome Initiative of the Broad Institute. For a full overview of the genomes and references see Table S4.1. This work was supported by the BioRange program of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

4.6 Supplementary material

Due to the amount of detail Supplemetary figures S4.1, S4.2, S4.3 are best viewed electronically. Supplemetary figures S4.1, S4.2 and S4.3 can be downloaded from http://bioinformatics.bio.uu.nl/john/thesis/.

Table S4.1. Genomes used in this study

Species	Genome source	Genome version	References
Homo sapiens	EnsEMBL	NCBI 36	[72,73]
Mus musculus	EnsEMBL	NCBI m37 Apr 2007	[74]
Ornithorhynchus anatinus	EnsEMBL	v5.0, Dec 2005	[75]
Gallus gallus	EnsEMBL	WASHUC2, May 2006	[76]
Xenopus tropicalis	EnsEMBL	JGI 4.1, Aug 2005	[77]
Danio rerio	EnsEMBL	Zv7, Apr 2007	[62]
Fugu rubripes	EnsEMBL	FUGU 4.0, Jun 2005	[78]
Branchiostoma floridae	JGI	v.1.0 (December 5, 2006)	[79]
Ciona intestinalis	EnsEMBL	JGI 2, Mar 2005	[80]
Strongylocentrotus purpuratus	SUGP	Spur 2.1, Sep 2006	[81]
Caenorhabditis elegans	EnsEMBL	WS190, Apr 2008	[82]
Anopheles gambiae	EnsEMBL	AgamP3, February 2006	[83]
Drospohila melanogaster	EnsEMBL	BDGP 5.4, Nov 2007	[84]
Lottia gigantea	JGI	v1.0 (July 24, 2007)	[85]
Daphnia pulex	JGI	v.1.0 (July 5, 2007) frozen catalog	[85]
Nematostella vectensis	JGI	JGI v1.0	[86]
Trichoplax adhaerens	JGI	JGI v1.0 22 July 2007	[87]
Monosiga brevicollis	JGI	JGI v1.0 July 2006	[88]
Encephalitozoon cuniculi	EMBL	EMBL	[89]
Batrachochytrium dendrobatidis	BROAD	BROAD 3/1/2007	[90]
Phycomyces blakesleeanus	JGI	JGI v1.0 January 8, 2007	[85]
Rhizopus oryzae	BROAD	RO3 12/6/05	[91]
Laccaria bicolor	JGI	JGI v1.0 May 22, 2006	[92]
Phanerochaete chrysosporium	JGI	JGI v2.0 February 2005	[93]
Cryptococcus neoformans	BROAD	BROAD v3.0 2/17/2006	[94]
Ustilago maydis	BROAD	BROAD v2.0 April 1, 2004	[95]
Schizosaccharomyces pombe	Sanger	Sanger v19 07/16/2008	[96]
Yarrowia lipolytica	Genolevures	V 2	[97]
Debaryomyces hansenii	Genolevures	V 1	[97]
Candida guilliermondii	BROAD	V 1	[98]
Eremothecium gossypii	EMBL		[99]
Candida glabrata	Genolevures	V 2	[97]
Kluyveromyces lactis	Genolevures	V 2	[97]
Saccharomyces cerevisiae	SGD	SGD 06/06/2008	[100]
Aspergillus terreus	BROAD	BROAD v1.0	[101]
Coccidioides immitis	BROAD	C. immitis RS v2	[102]
Fusarium graminearum	BROAD	BROAD v3	[103]
Neurospora crassa	BROAD	BROAD v7	[104]

Species	Genome source	Genome version	References
Sclerotinia sclerotiorum	BROAD	BROAD v2.0	[105]
Dictyostelium discoideum	DictyDB	created: 03-01-2009 01:29	[106,107]
Cyanidioschyzon merolae	C. merolae Genome Project	April 13 2004 (ORF)	[108]
Ostreococcus tauri	JGI	JGI v2	[109]
Chlamydomonas reinhardtii	JGI	JGI v3.1	[110]
Volvox carteri	JGI	JGI v1.0 June 1, 2007	[111]
Physcomitrella patens ssp patens	JGI	v.1.1 (March 2007)	[112]
Selaginella moellendorffii	JGI	v1.0 (December 20, 2007)	[85]
Oryza sativa	Rice Genome Annotation (TIGR)	v5 TIGR January 24, 2007	[113]
Arabidopsis thaliana	Arabidopsis genome initiative	TIER v8	[114]
Populus trichocarpa	JGI	v1.0 (June 2004)	[115]
Phaeodactylum tricornutum	JGI	v.2.0 (November 16, 2006)	[116]
Thalassiosira pseudonana	JGI	v3.0 (August 2006)	[117]
Phytophthora sojae	JGI	v1.0 (April 2004)	[118]
Phytophthora infestans	BROAD	03-14-07	[119]
Emiliania huxleyi CCMP1516	JGI	v.1.0 (April 25, 2008)	[85]
Aureococcus anophagefferens	JGI	JGI v1.0 September 27, 2007	[85]
Paramecium tetraurelia	ParameciumDB	V1.21 07/29/2008	[120]
Tetrahymena thermophila	TIGR	aug-04	[121]
Cryptosporidium parvum	CryptoDB	CryptoDB 3.7	[122]
Plasmodium falciparum	PlasmoDB	5,5	[123]
Theileria parva	TIGR	1	[124]
Naegleria gruberi	JGI	v.1.0 (October 23, 2006)	[25]
Trichomonas vaginalis	TIGR	02-15-07	[125,126]
Gardia intestinalis	GiardiaDB	V1.1 may 2008	[126]
Leishmania major	Sanger	11-5-2006	[127]
Trypanosoma brucei	Sanger	May 08 v4	[128]

4.7 References

1. Thomas G, Hall MN (1997) TOR signalling and control of cell growth. Current opinion in cell biology 9: 782-7.
2. Goberdhan DCI, Ogmundsdóttir MH, Kazi S, Reynolds B, Visvalingam SM, et al. (2009) Amino acid sensing and mTOR regulation: inside or out? Biochemical Society transactions 37: 248-52. doi:10.1042/BST0370248
3. Oldham S, Hafen E (2003) Insulin/IGF and target of rapamycin signaling: a TOR de force in growth control. Trends in Cell Biology 13: 79-85. doi:10.1016/S0962-8924(02)00042-9
4. Wullschleger S, Loewith R, Hall MN (2006) TOR signaling in growth and metabolism. Cell 124: 471-84. doi:10.1016/j.cell.2006.01.016
5. Menon S, Manning BD (2008) Common corruption of the mTOR signaling network in human tumors. Oncogene 27 Suppl 2: S43-51. doi:10.1038/onc.2009.352
6. Mahfouz MM, Kim S, Delauney AJ, Verma DPS (2006) Arabidopsis TARGET OF RAPAMYCIN interacts with RAPTOR, which regulates the activity of S6 kinase in response to osmotic stress signals. The Plant cell 18: 477-90. doi:10.1105/tpc.105.035931
7. Tee AR, Blenis J (2005) mTOR, translational control and human disease. Seminars in cell & developmental biology 16: 29-37. doi:10.1016/j.semcdb.2004.11.005
8. Zhang Y, Gao X, Saucedo LJ, Ru B, Edgar BA, et al. (2003) Rheb is a direct target of the tuberous sclerosis tumour suppressor proteins. Nature cell biology 5: 578-81. doi:10.1038/ncb999
9. Inoki K, Li Y, Xu T, Guan K-L (2003) Rheb GTPase is a direct target of TSC2 GAP activity and regulates mTOR signaling. Genes & development 17: 1829-34. doi:10.1101/gad.1110003
10. Ma XM, Blenis J (2009) Molecular mechanisms of mTOR-mediated translational control. Nature reviews. Molecular cell

Table S4.2; Presence/Absence plot of all TOR pathway components in all 64 eukaryotic genomes

biology 10: 307-18. doi:10.1038/nrm2672

11. Shaw RJ, Cantley LC (2006) Ras, PI(3)K and mTOR signalling controls tumour cell growth. Nature 441: 424-30. doi:10.1038/nature04869
12. Engelman JA, Luo J, Cantley LC (2006) The evolution of phosphatidylinositol 3-kinases as regulators of growth and metabolism. Nature reviews. Genetics 7: 606-19. doi:10.1038/nrg1879
13. Avruch J, Long X, Ortiz-Vega S, Rapley J, Papageorgiou A, et al. (2009) Amino acid regulation of TOR complex 1. American journal of physiology. Endocrinology and metabolism 296: E592-602. doi:10.1152/ajpendo.90645.2008
14. Vega-Rubin-de-Celis S, Abdallah Z, Kinch L, Grishin NV, Brugarolas J, et al. (2010) Structural analysis and functional implications of the negative mTORC1 regulator REDD1. Biochemistry. doi:10.1021/bi902135e
15. Vernoud V, Horton AC, Yang Z, Nielsen E (2003) Analysis of the small GTPase gene superfamily of Arabidopsis. Plant physiology 131: 1191-208. doi:10.1104/pp.013052
16. Cybulski N, Hall MN (2009) TOR complex 2: a signaling pathway of its own. Trends in biochemical sciences 34: 620-7. doi:10.1016/j.tibs.2009.09.004
17. Kockel L, Kerr KS, Melnick M, Brückner K, Hebrok M, et al. (2010) Dynamic switch of negative feedback regulation in Drosophila Akt-TOR signaling. PLoS genetics 6: e1000990. doi:10.1371/journal.pgen.1000990
18. Hall MN (2008) mTOR-what does it do? Transplantation proceedings 40: S5-8. doi:10.1016/j.transproceed.2008.10.009
19. Lorberg A, Hall MN (2004) TOR: the first 10 years. Current topics in microbiology and immunology 279: 1-18.
20. Deprost D, Yao L, Sormani R, Moreau M, Leterreux G, et al. (2007) The Arabidopsis TOR kinase links plant growth, yie... [EMBO Rep. 2007] - PubMed result. EMBO reports 8: 864-70. doi:10.1038/sj.embor.7401043
21. Urano J (2000) The Saccharomyces cerevisiae Rheb G-protein Is Involved in Regulating Canavanine Resistance and Arginine Uptake. Journal of Biological Chemistry 275: 11198-11206. doi:10.1074/jbc.275.15.11198
22. Matsumoto S, Bandyopadhyay A, Kwiatkowski DJ, Maitra U, Matsumoto T (2002) Role of the Tsc1-Tsc2 complex in signaling and transport across the cell membrane in the fission yeast Schizosaccharomyces pombe. Genetics 161: 1053-63.
23. Díaz-Troya S, Pérez-Pérez ME, Florencio FJ, Crespo JL (2008) The role of TOR in autophagy regulation from yeast to plants and mammals. Autophagy 4: 851-65.
24. Menand B, Desnos T, Nussaume L, Berger F, Bouchez D, et al. (2002) Expression and disruption of the Arabidopsis TOR (target of rapamycin) gene. Proceedings of the National Academy of Sciences of the United States of America 99: 6422-7. doi:10.1073/pnas.092141899
25. Fritz-Laylin LK, Prochnik SE, Ginger ML, Dacks JB, Carpenter ML, et al. (2010) The Genome of Naegleria gruberi Illuminates Early Eukaryotic Versatility. Cell 140: 631-642. doi:10.1016/j.cell.2010.01.032
26. Loewith R, Jacinto E, Wullschleger S, Lorberg A, Crespo JL, et al. (2002) Two TOR Complexes, Only One of which Is Rapamycin Sensitive, Have Distinct Roles in Cell Growth Control. Molecular Cell 10: 457-468. doi:10.1016/S1097-2765(02)00636-6
27. Shevchenko A, Roguev A, Schaft D, Buchanan L, Habermann B, et al. (2008) Chromatin Central: towards the comparative proteome by accurate mapping of the yeast proteomic environment. Genome biology 9: R167. doi:10.1186/gb-2008-9-11-r167
28. Hartmuth S, Petersen J (2009) Fission yeast Tor1 functions as part of TORC1 to control mitotic entry through the stress MAPK pathway following nutrient stress. Journal of cell science 122: 1737-46. doi:10.1242/jcs.049387
29. Otsubo Y, Yamamato M (n.d.) TOR signaling in fission yeast. Critical reviews in biochemistry and molecular biology 43: 277-83. doi:10.1080/10409230802254911
30. Soulard A, Cohen A, Hall MN (2009) TOR signaling in invertebrates. [Curr Opin Cell Biol. 2009] - PubMed result. Current opinion in cell biology 21: 825-36. doi:10.1016/j.ceb.2009.08.007
31. Colicelli J Human RAS superfamily proteins and related GTPases. Science's STKE : signal transduction knowledge environment 2004: RE13. doi:10.1126/stke.2502004re13
32. Wennerberg K, Rossman KL, Der CJ (2005) The Ras superfamily at a glance. Journal of cell science 118: 843-6. doi:10.1242/jcs.01660
33. Carrière A, Cargnello M, Julien L-A, Gao H, Bonneil E, et al. (2008) Oncogenic MAPK signaling stimulates mTORC1 activity by promoting RSK-mediated raptor phosphorylation. Current biology : CB 18: 1269-77. doi:10.1016/j.cub.2008.07.078
34. Inoki K, Ouyang H, Zhu T, Lindvall C, Wang Y, et al. (2006) TSC2 integrates Wnt and energy signals via a coordinated phosphorylation by AMPK and GSK3 to regulate cell growth. Cell 126: 955-68. doi:10.1016/j.cell.2006.06.055
35. Vivanco I, Sawyers CL (2002) The phosphatidylinositol 3-Kinase AKT pathway in human cancer. Nature reviews. Cancer 2: 489-501. doi:10.1038/nrc839
36. Serfontein J, Nisbet RER, Howe CJ, de Vries PJ (2010) Evolution of the TSC1/TSC2-TOR Signaling Pathway. Science Signaling 3: ra49-ra49. doi:10.1126/scisignal.2000803
37. Hsu Y-C, Chern JJ, Cai Y, Liu M, Choi K-W (2007) Drosophila TCTP is essential for growth and proliferation through regulation of dRheb GTPase. Nature 445: 785-8. doi:10.1038/nature05528
38. Wang X, Fonseca BD, Tang H, Liu R, Elia A, et al. (2008) Re-evaluating the roles of proposed modulators of mammalian

target of rapamycin complex 1 (mTORC1) signaling. The Journal of biological chemistry 283: 30482-92. doi:10.1074/jbc.M803348200

39. Rehmann H, Brüning M, Berghaus C, Schwarten M, Köhler K, et al. (2008) Biochemical characterisation of TCTP questions its function as a guanine nucleotide exchange factor for Rheb. FEBS letters 582: 3005-10. doi:10.1016/j.febslet.2008.07.057

40. van Dam TJP, Rehmann H, Bos JL, Snel B (2009) Phylogeny of the CDC25 homology domain reveals rapid differentiation of Ras pathways between early animals and fungi. Cellular signalling 21: 1579-85. doi:10.1016/j.cellsig.2009.06.004

41. Berkowitz O, Jost R, Pollmann S, Masle J (2008) Characterization of TCTP, the translationally controlled tumor protein, from Arabidopsis thaliana. The Plant cell 20: 3430-47. doi:10.1105/tpc.108.061010

42. Basu MK, Carmel L, Rogozin IB, Koonin EV (2008) Evolution of protein domain promiscuity in eukaryotes. Genome research 18: 449-61. doi:10.1101/gr.6943508

43. Shalchian-Tabrizi K, Minge MA, Espelund M, Orr R, Ruden T, et al. (2008) Multigene phylogeny of choanozoa and the origin of animals. PloS one 3: e2098. doi:10.1371/journal.pone.0002098

44. Alessi DR, James SR, Downes CP, Holmes AB, Gaffney PRJ, et al. (1997) Characterization of a 3-phosphoinositide-dependent protein kinase which phosphorylates and activates protein kinase Bα. Current Biology 7: 261-269. doi:10.1016/S0960-9822(06)00122-9

45. Burgering BM, Coffer PJ (1995) Protein kinase B (c-Akt) in phosphatidylinositol-3-OH kinase signal transduction. Nature 376: 599-602. doi:10.1038/376599a0

46. Kobayashi T, Cohen P (1999) Activation of serum- and glucocorticoid-regulated protein kinase by agonists that activate phosphatidylinositide 3-kinase is mediated by 3-phosphoinositide-dependent protein kinase-1 (PDK1) and PDK2. The Biochemical journal 339 (Pt 2: 319-28.

47. Stokoe D (1997) Dual Role of Phosphatidylinositol-3,4,5-trisphosphate in the Activation of Protein Kinase B. Science 277: 567-570. doi:10.1126/science.277.5325.567

48. Park J, Leong ML, Buse P, Maiyar AC, Firestone GL, et al. (1999) Serum and glucocorticoid-inducible kinase (SGK) is a target of the PI 3-kinase-stimulated signaling pathway. The EMBO journal 18: 3024-33. doi:10.1093/emboj/18.11.3024

49. Potter CJ, Pedraza LG, Xu T (2002) Akt regulates growth by directly phosphorylating Tsc2. Nature cell biology 4: 658-65. doi:10.1038/ncb840

50. Dan HC, Sun M, Yang L, Feldman RI, Sui X-M, et al. (2002) Phosphatidylinositol 3-kinase/Akt pathway regulates tuberous sclerosis tumor suppressor complex by phosphorylation of tuberin. The Journal of biological chemistry 277: 35364-70. doi:10.1074/jbc.M205838200

51. Burgering BMT, Kops GJPL (2002) Cell cycle and death control: long live Forkheads. Trends in Biochemical Sciences 27: 352-360. doi:10.1016/S0968-0004(02)02113-8

52. Brunet A, Park J, Tran H, Hu LS, Hemmings BA, et al. (2001) Protein kinase SGK mediates survival signals by phosphorylating the forkhead transcription factor FKHRL1 (FOXO3a). Molecular and cellular biology 21: 952-65. doi:10.1128/MCB.21.3.952-965.2001

53. Ikeda K, Morigasaki S, Tatebe H, Tamanoi F, Shiozaki K (2008) Fission yeast TOR complex 2 activates the AGC-family Gad8 kinase essential for stress resistance and cell cycle control. Cell cycle (Georgetown, Tex.) 7: 358-64.

54. Chen B-R, Runge KW (2009) A new Schizosaccharomyces pombe chronological lifespan assay reveals that caloric restriction promotes efficient cell cycle exit and extends longevity. Experimental gerontology 44: 493-502. doi:10.1016/j.exger.2009.04.004

55. Roux PP, Ballif BA, Anjum R, Gygi SP, Blenis J (2004) Tumor-promoting phorbol esters and activated Ras inactivate the tuberous sclerosis tumor suppressor complex via p90 ribosomal S6 kinase. Proceedings of the National Academy of Sciences of the United States of America 101: 13489-94. doi:10.1073/pnas.0405659101

56. Casamayor A, Torrance PD, Kobayashi T, Thorner J, Alessi DR (1999) Functional counterparts of mammalian protein kinases PDK1 and SGK in budding yeast. Current Biology 9: 186-S4. doi:10.1016/S0960-9822(99)80088-8

57. Jacinto E, Lorberg A (2008) TOR regulation of AGC kinases in yeast and mammals. The Biochemical journal 410: 19-37. doi:10.1042/BJ20071518

58. Mukai H (1995) Identification of Schizosaccharomyces pombe gene psk1+, encoding a novel putative serine/threonine protein kinase, whose mutation conferred resistance to phenylarsine oxide. Gene 166: 155-159. doi:10.1016/0378-1119(95)00553-1

59. Urban J, Soulard A, Huber A, Lippman S, Mukhopadhyay D, et al. (2007) Sch9 is a major target of TORC1 in Saccharomyces cerevisiae. Molecular cell 26: 663-74. doi:10.1016/j.molcel.2007.04.020

60. Findlay GM, Harrington LS, Lamb RF (2005) TSC1-2 tumour suppressor and regulation of mTOR signalling: linking cell growth and proliferation? Current opinion in genetics & development 15: 69-76. doi:10.1016/j.gde.2004.11.002

61. Manning BD (2004) Balancing Akt with S6K: implications for both metabolic diseases and tumorigenesis. The Journal of cell biology 167: 399-403. doi:10.1083/jcb.200408161

62. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, et al. (2009) Ensembl 2009. Nucleic Acids Res 37: D690-7. doi:10.1093/nar/gkn828

63. Van Dongen S (2008) Graph Clustering Via a Discrete Uncoupling Process. SIAM Journal on Matrix Analysis and

Applications 30: 121-141.

64. Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. Journal of molecular biology 314: 1041-52. doi:10.1006/jmbi.2000.5197

65. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of molecular biology 215: 403-10. doi:10.1006/jmbi.1990.9999

66. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, et al. (2008) The Pfam protein families database. Nucleic Acids Res 36: D281-8. doi:10.1093/nar/gkm960

67. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755-763. doi:10.1093/bioinformatics/14.9.755

68. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30: 3059-3066. doi:10.1093/nar/gkf436

69. Howe K, Bateman A, Durbin R (2002) QuickTree: building huge Neighbour-Joining trees of protein sequences. Bioinformatics 18: 1546-1547. doi:10.1093/bioinformatics/18.11.1546

70. Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21: 456-463. doi:10.1093/bioinformatics/bti191

71. Huson D, Richter D, Rausch C, Dezulian T, Franz M, et al. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. BMC bioinformatics 8: 460. doi:10.1186/1471-2105-8-460

72. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921. doi:10.1038/35057062

73. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science (New York, N.Y.) 291: 1304-51. doi:10.1126/science.1058040

74. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520-62. doi:10.1038/nature01262

75. Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, et al. (2008) Genome analysis of the platypus reveals unique signatures of evolution. Nature 453: 175-83. doi:10.1038/nature06936

76. Wallis JW, Aerts J, Groenen MAM, Crooijmans RPMA, Layman D, et al. (2004) A physical map of the chicken genome. Nature 432: 761-4. doi:10.1038/nature03030

77. Hellsten U, Harland RM, Gilchrist MJ, Hendrix D, Jurka J, et al. (2010) The Genome of the Western Clawed Frog Xenopus tropicalis. Science 328: 633-636. doi:10.1126/science.1183670

78. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, et al. (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297: 1301-1310. doi:10.1126/science.10721041072104 [pii]

79. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453: 1064-71. doi:10.1038/nature06967

80. Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, et al. (2002) The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. Science (New York, N.Y.) 298: 2157-67. doi:10.1126/science.1080049

81. Cameron RA, Samanta M, Yuan A, He D, Davidson E (2009) SpBase: the sea urchin genome database and web site. Nucleic Acids Res 37: D750-4. doi:gkn887 [pii]10.1093/nar/gkn887

82. The C. elegans Sequencing Consortium (1998) Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science 282: 2012-2018. doi:10.1126/science.282.5396.2012

83. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. (2002) The genome sequence of the malaria mosquito Anopheles gambiae. Science (New York, N.Y.) 298: 129-49. doi:10.1126/science.1076181

84. Adams MD (2000) The Genome Sequence of Drosophila melanogaster. Science 287: 2185-2195. doi:10.1126/science.287.5461.2185

85. These sequence data were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ in collaboration with the user community. (n.d.).

86. Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, et al. (2007) Science (New York, N.Y.) 317: 86-94. doi:10.1126/science.1139158

87. Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, et al. (2008) Nature 454: 955-960. doi:nature07191 [pii]10.1038/nature07191

88. King N, Westbrook MJ, Young SL, Kuo A, Abedin M, et al. (2008) Nature 451: 783-788. doi:nature06617 [pii]10.1038/nature06617

89. Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, et al. (2001) Nature 414: 450-453. doi:10.1038/3510657935106579 [pii]

90. Batrachochytrium dendrobatidis Sequencing Project. Broad Institute of Harvard and MIT (http://www.broad.mit.edu) (n.d.).

91. Rhizopus oryzae Sequencing Project. Broad Institute of Harvard and MIT (http://www.broad.mit.edu) (n.d.).

92. Martin F, Aerts A, Ahren D, Brun A, Danchin EG, et al. (2008) Nature 452: 88-92. doi:nature06556 [pii]10.1038/nature06556

93. Martinez D, Larrondo LF, Putnam N, Gelpke MD, Huang K, et al. (2004) Nat Biotechnol 22: 695-700. doi:10.1038/nbt967nbt967 [pii]

94. C. neoformans Genome Project, Stanford Genome Technology Center, funded by the NIAID/NIH under cooperative agreement AI47087, and The Institute for Genomic Research, funded by the NIAID/NIH under cooperative agreement U01 AI48594. (n.d.).

95. Kämper J, Kahmann R, Bölker M, Ma L-J, Brefort T, et al. (2006) Nature 444: 97-101. doi:10.1038/nature05248

96. Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, et al. (2002) Nature 415: 871-880. doi:10.1038/nature724nature724 [pii]

97. Sherman DJ, Martin T, Nikolski M, Cayla C, Souciet JL, et al. (2009) Nucleic Acids Res 37: D550-4. doi:gkn859 [pii]10.1093/nar/gkn859

98. Candida guilliermondii Sequencing Project. Broad Institute of Harvard and MIT (http://www.broad.mit.edu) (n.d.).

99. Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, et al. (2004) Science 304: 304-307. doi:10.1126/science.10957811095781 [pii]

100. Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, et al. (1997) Nature 387: 67-73.

101. Aspergillus Comparative Genome Project. Broad Institute of Harvard and MIT (http://www.broad.mit.edu) (n.d.).

102. Coccidioides immitis Sequencing Project. Broad Institute of Harvard and MIT (http://www.broad.mit.edu) (n.d.).

103. Cuomo CA, Guldener U, Xu JR, Trail F, Turgeon BG, et al. (2007) Science 317: 1400-1402. doi:317/5843/1400 [pii]10.1126/science.1143708

104. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, et al. (2003) Nature 422: 859-868. doi:10.1038/nature01554nature01554 [pii]

105. Sclerotinia sclerotiorum Sequencing Project. Broad Institute of Harvard and MIT (http://www.broad.mit.edu) (n.d.).

106. Fey P, Gaudet P, Curk T, Zupan B, Just EM, et al. (2009) Nucleic Acids Res 37: D515-9. doi:gkn844 [pii]10.1093/nar/gkn844

107. Eichinger L, Pachebat JA, Glockner G, Rajandream MA, Sucgang R, et al. (2005) Nature 435: 43-57. doi:nature03481 [pii]10.1038/nature03481

108. Nozaki H, Takano H, Misumi O, Terasawa K, Matsuzaki M, et al. (2007) BMC Biol 5: 28. doi:1741-7007-5-28 [pii]10.1186/1741-7007-5-28

109. Palenik B, Grimwood J, Aerts A, Rouzé P, Salamov A, et al. (2007) Proceedings of the National Academy of Sciences of the United States of America 104: 7705-10. doi:10.1073/pnas.0611046104

110. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, et al. (2007) Science 318: 245-250. doi:318/5848/245 [pii]10.1126/science.1143609

111. Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, et al. (2010) Science 329: 223-226. doi:10.1126/science.1188800

112. Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, et al. (2008) Science (New York, N.Y.) 319: 64-9. doi:10.1126/science.1150646

113. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, et al. (2007) Nucleic Acids Res 35: D883-7. doi:gkl976 [pii]10.1093/nar/gkl976

114. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, et al. (2008) Nucleic Acids Res 36: D1009-14. doi:gkm965 [pii]10.1093/nar/gkm965

115. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, et al. (2006) Science (New York, N.Y.) 313: 1596-604. doi:10.1126/science.1128691

116. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, et al. (2008) Nature 456: 239-244. doi:nature07410 [pii]10.1038/nature07410

117. Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, et al. (2004) Science 306: 79-86. doi:10.1126/science.1101156306/5693/79 [pii]

118. Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RHY, et al. (2006) Science 313: 1261-1266. doi:10.1126/science.1128796

119. Haas BJ, Kamoun S, Zody MC, Jiang RHY, Handsaker RE, et al. (2009) Nature 461: 393-8. doi:10.1038/nature08358

120. Aury JM, Jaillon O, Duret L, Noel B, Jubin C, et al. (2006) Nature 444: 171-178. doi:nature05230 [pii]10.1038/nature05230

121. Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, et al. (2006) PLoS Biology 4: e286.

122. Heiges M, Wang H, Robinson E, Aurrecoechea C, Gao X, et al. (2006) Nucleic Acids Res 34: D419-22. doi:34/suppl_1/D419 [pii]10.1093/nar/gkj078

123. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, et al. (2009) Nucleic acids research 37: D539-43. doi:10.1093/nar/gkn814

124. Gardner MJ, Bishop R, Shah T, de Villiers EP, Carlton JM, et al. (2005) Science 309: 134-137. doi:309/5731/134 [pii]10.1126/science.1110439

125. Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, et al. (2007) Science 315: 207-212. doi:315/5809/207 [pii]10.1126/

science.1132894

126. Aurrecoechea C, Brestelli J, Brunk BP, Carlton JM, Dommer J, et al. (2009) Nucleic Acids Res 37: D526-30. doi:gkn631 [pii]10.1093/nar/gkn631

127. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, et al. (2005) Science (New York, N.Y.) 309: 436-42. doi:10.1126/science.1112680

128. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, et al. (2005) Science 309: 416-422. doi:309/5733/416 [pii]10.1126/science.1112642

CHAPTER 5

Evolution of the Ras-like small GTPases and their regulators

Teunis J.P. van Dam, Johannes L. Bos and Berend Snel

Accepted for publication
Small GTPases, February 9 2011

5.1 Abstract

Small GTPases are molecular switches at the hub of many signaling pathways and the expansion of this protein family is interwoven with the origin of unique eukaryotic cell features. We have previously reported on the evolution of CDC25 Homology Domain containing proteins, which act as guanine nucleotide exchange factors (GEFs) for Ras-like proteins. We now report on the evolution of both the Ras-like small GTPases as well as the GTPase activating proteins (GAPs) for Ras-like small GTPases. We performed an in depth phylogenetic analysis in 64 genomes of diverse eukaryotic species. These analyses revealed that multiple ancestral Ras-like GTPases and GAPs were already present in the Last Eukaryotic Common Ancestor (LECA), compatible with the presence of RasGEFs in LECA. Furthermore, we endeavor to reconstruct in which order the different Ras-like GTPases diverged from each other. We identified striking differences between the expansion of the various types of Ras-like GTPases and their respective GAPs and GEFs. Altogether, our analysis forms an extensive evolutionary framework for Ras-like signaling pathways and provides specific predictions for molecular biologists and biochemists.

5.2 Introduction

The Ras-like GTPases perform central regulatory functions in a myriad of cellular processes, such as cell division, differentiation, cell-cell adhesion, growth and apoptosis.

The well-studied oncogenes H-, N-, and K-Ras belong to this family. The Ras-like GTPase family belong to the small GTPase superfamily of which all members share high sequence similarity and structural features [1]. Other well-known members of this superfamily are the Rab, Ran, Rho and Arf GTPases.

The expansion and evolution of the small GTPases have been related to the evolution of unique eukaryotic cellular features, such as phagocytosis [2,3] and specific studies on the evolution of the Rab and Rho GTPases have been published [4-6]. However, to our knowledge no large scale and broad phylogenetic analysis has been reported for the Ras-like GTPases.

The small GTPases are inefficient GTPases that require GTPase Activating Proteins (GAPs) to enhance the intrinsic GTPase activity, and Guanine Exchange Factors (GEFs) to efficiently exchange bound GDP for GTP. Regulation of GAP and GEF activity allows for small GTPases to function as an efficient, highly regulated molecular switch that can quickly alternate between its active GTP bound state and its inactive GDP bound state. Although virtually all small GTPases share this common principle of regulation, each subfamily of GTPases has its own set of evolutionary unrelated GAPs and GEFs [7]. Importantly, whereas Ras-like small GTPase are almost uniquely regulated by CDC25 HD domain containing proteins as GEF, the GAP proteins belong to two unrelated families, the RasGAPs and the RapGAPs [7].

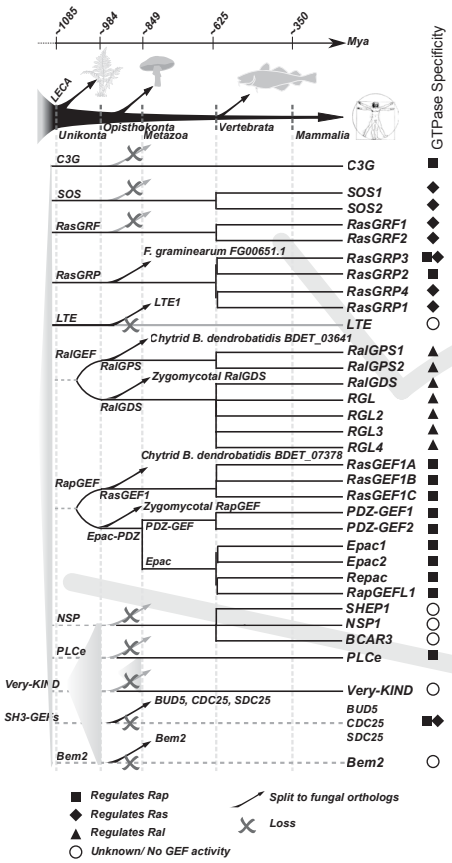


Figure 5.1 Graphical summary of the evolution of the CDC25 Homology Domain as described in van Dam et al. For each class of RasGEFs we have traced back their perceived and reconstructed time of origin. For the mammalian RasGEFs we reconstructed the time of duplications to specific points in eukaryotic evolution. Note the differential loss of ancestral RasGEFs between animals and fungi. All gene names of fungal RasGEFs are from orthologs in *S. cerevisiae* unless specified otherwise. The evolutionary time scale is based on Douzery et al.[34] but note that molecular dating is highly inaccurate[42] and that these dates are therefore approximate at best.

Specific members of the Ras-like GTPase family regulate a multitude of diverse cellular processes. Therefore analyzing the phylogeny of the Ras-like GTPases and their GAPs and GEFs could clarify and explain the differences as well as provide an in depth evolutionary history of one of the most enigmatic families belonging to the small GTPase superfamily.

We have previously shown a surprisingly tight evolutionary relationship between Ras-like GTPases and their Guanine Exchange Factors [8]. The CDC25 Homology Domain (CDC25HD), or RasGEF domain, was found in animal, fungal, chromalveolate and excavate species. The CDC25HD co-occurs perfectly with the Ras, Rap and Ral GTPases

(i.e. either both GTPase and CDC25HD are present or both are absent in all species). We observed that the significant differences in the RasGEF repertoire between fungi and animals has been the result of differential loss of RasGEFs in the animal and fungal ancestor (see Figure 5.1). The tight evolutionary relationship between Ras-like GTPases and their RasGEFs raises the possibility that the evolution of the regulatory domains likely holds information on the evolution of the Ras-like GTPases as well.

We now report on the evolutionary history both of Ras- and RapGAPs as well of Ras-like small GTPases. We found that the origin of the RapGAP and RasGAP domains can reliably be traced back to four ancestral RapGAPs and five ancestral RasGAPs in the Last Eukaryotic Common Ancestor (LECA). We identified a previously unrecognized RapGAP domain in RalGAPB, a subunit of the RalGAPA/B dimer, and we are able to show that this RalGAPA/B dimer was already present in LECA. The domain architecture of the RasGAP protein family is extremely well-conserved throughout the eukaryotic kingdom. Phylogenetic analysis of the Ras-like GTPase protein sequences reveal that multiple Ras-like GTPases were present in LECA. We reconstruct the early evolution of the Ras-like GTPases and its regulatory domains and show in which order the Ras, Rap, Ral and Rheb GTPases diverged from each other. Lastly we combine the analyses of the phylogenies of the Ras-like GTPases, their GAPs and GEFs and examine the evolutionary dynamics of the whole Ras regulatory system (i.e. the GTPase and its GAPs and GEFs). We show that Rap and Ras display strikingly different behavious in evolution with respect to the expansion of their respective GEFs and GAPs.

5.3 Results and discussion

5.3.1 Evolution of the Rap GTPase Activating Protein Domain; conservation of GTPase specificity

We collected RapGAP domain (Pfam: PF02145) sequences from our genome dataset containing 64 different eukaryotics genomes and performed phylogenetic analysis. The resulting phylogenetic tree clearly shows a trifurcation into Rap, Rheb and Ral specific monophyletic groups of RapGAP domain proteins (Figure 5.2). For a detailed phylogenetic tree including domain architecture of each protein see Figure S5.1.

Rheb specific RapGAP domain containing proteins

The cluster of Rheb specific RapGAP domain containing proteins encompasses the TSC2 gene and its orthologs as described in chapter 4. The TSC2 protein contains a DUF3384 domain, a Tuberin domain and a RapGAP domain. This domain architecture is strongly conserved in animal and fungal TSC2-like proteins (see Figure S4.1). Additionally, we find sequences containing the TSC2-like RapGAP domain in *Dictyostelium discoideum* (amoeba), *Phytophthora* species (oomycetes), *Phaeodactylum tricornutum* (diatom), *Tetrahymena thermophila* (ciliate), and the red algae *C. merolae*. The species distribution for the TSC2 GAP domain strongly suggests that the TSC2-like RapGAP domain originated in or before LECA. Except for *T. thermophila* the TSC2-like RapGAP domain can be found exclusively in species that also contain a Rheb ortholog (Chapter 4 and this chapter), indicating that there is an exceptionally strong evolutionary link between Rheb and the TSC2-like GAP domain.

The Ral specific RapGAP domain containing proteins; evidence for an ancient heterodimer complex

The GAP specific for the Ral GTPases was already discovered earlier [9] but has only been recently shown to specifically regulate Ral [10]. RalGAP is a protein complex that consists of RalGAPB with either RalGAPA1 or RalGAPA2. RalGAPA1 and 2 have been shown to harbor the RapGAP domain and perform the GAP function towards Ral. The RalGAP complex has been described as a TSC1/2-like complex as RalGAPB initially seemed to have no protein domain or other obvious structural features, much like TSC1. However, we identified a putative RapGAP domain at the C-terminal part of RalGAPB by sequence analysis (custom HMM model, e-value 7.1e-08 for the human RalGAPB, see also Figure S5.1). Although the RapGAP domain sequence is highly conserved, RalGAPB has an insertion/deletion where normally the catalytic loop can be found (positions 282 to 290 in the Rap1GAP protein sequence, see Scrima et al. [11]). This raises questions about the specific function of RalGAPB in the protein complex as it has no sequence similarity to TSC1 but also apparently lacks the ability to provide GAP function towards Rap, Ral or Rheb.

Analysis of the phylogenetic tree as represented in Figure 5.2 indicates that the RapGAP domain of RalGAPA1/2 and RalGAPB are closely related and distinct from the Rap specific and Rheb specific RapGAP domain containing proteins. The close relation between RalGAPA1/2 and RalGAPB suggests that the RalGAP complex originated from a single gene, of which the gene product likely formed a homodimeric complex that after gene duplication became an heterodimeric complex. Interestingly, based on the species distribution in the gene tree we can reconstruct both RalGAPA1/2 (the result of a duplication in the vertebrate ancestor) and RalGAPB as ancestral genes in LECA. The species distribution of RalGAPA and RalGAPB are identical suggesting that the RalGAP complex is perfectly conserved in evolution and was already present as a heterodimeric complex in LECA.

Intriguingly, the species distribution of the RalGAP complex matches the species distribution of the Ral subtype GTPases (animals + chytrid and zygomycota fungi), but can additionally be found in the amoeba *D. discoideum*, the oomycetes *Phytophthora sojae* and *Phytophthora infestans* and the excavate *N. gruberi*. We have not identified Ral-like GTPases nor putative RalGEFs [8] in these organisms. It could therefore be possible that the RalGAP complexes in these species perform GAP functions towards other small GTPases. Accordingly, this then raises the question if the RalGAP complex might also still perform GAP activity to other GTPases in human since RalGAP complex activity has been tested only on the Ras-like GTPases RalA and B, H-Ras, Rap1 and Rheb [10]. This leaves many Ras-like GTPases still to be tested.

The Rap subtype specific RapGAP domain containing proteins; the odd ones

As the Rap specific RapGAPs form a well defined cluster outside of the TSC2 and RalGAP clusters (100% bootstrap support), they too can be reconstructed into the LECA gene repertoire. The Rap specific RapGAPs are by far the most expanded RapGAP domain containing proteins in mammals. In this group we find SIPA and the SIPA-like RapGAPs as well as the Rap1GAPs. We also find an uncharacterized RapGAP domain containing protein called GARNL3 (bootstrap support of 100%). GARNL3 contains in addition to a RapGAP domain a CNH domain (InterPro IPR001180, Pfam PF00780) which is present in

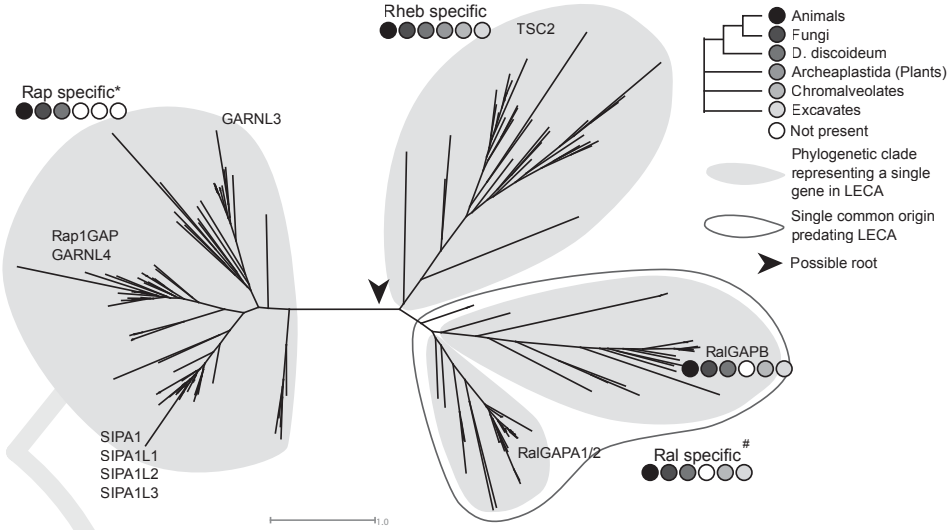


Figure 5.2 Representation of the phylogenetic tree of RapGAP domain containing proteins. We observe four clades, each of which represents a single ancestral gene in LECA. RalGAPB and RalGAPA each represent a single ancestral gene in LECA but cluster together, indicating a common ancestral gene preceding LECA. The species present in each clade is depicted as a colored barcode. The differences in sequence between the Rap specific GAP sequences and the Rheb/Ral specific GAP sequences is well defined (100% bootstrap support) and a possible root may therefore lie between these two groups. *In the Rap specific clade we also observe GARNL3, a putative RapGAP, but it cellular function has not yet been reported. #It is unknown if RalGAPB harbors any GAP activity.

a number of proteins that interact with Rho, Rac and or CDC42 [12]. GARNL3 therefore could possibly link Rho GTPase activity to RapGAP activity as GARNL3 clearly clusters together with known RapGAPs.

The Rap1GAPs and SIPA-like RapGAPs each occur in all animals and likely originated from a duplication in the animal ancestor. The GARNL3 cluster also contains protein sequences from the chytrid fungus *B. dendrobaditis* and the amoeba *D. discoideum*. Sequences belonging to the fungi *B. dendrobaditis*, *R. oryzae* and *P. blakesleeanus* cluster outside of the RapGAPs including GARNL3. These sequences potentially belong to the SIPA-like and Rap1GAP cluster. The species distribution suggests that GARNL3 and the SIPA/Rap1GAP genes originated in the Unikont or Opisthokont ancestor from a gene duplication. Interestingly the Rap specific RapGAP group of proteins lack sequences from a large group of fungal species (i.e. basidiomycetes and ascomycetes) while these species do contain a Rap ortholog (e.g. BUD1 in *S. cerevisiae*). It is likely that the Rap specific RapGAP domain containing proteins have been replaced by the C2 domain containing RasGAPs (the RASA-like RasGAPs, BUD5 in *S. cerevisiae*, see below).

5.3.2 Evolution of the Ras GTPase Activating Protein Domain; conservation of domain architecture

The phylogenetic tree of the Ras GTPase Activating Proteins or RasGAPs shows a similar behavior as the RapGAP tree. We observe five clusters which each represent an ancestral RasGAP gene in LECA (Figure 5.3). For two of the five ancestral groups there is evidence to support a common origin predating LECA (GAP1 and IQGAP-like RasGAPs). However

in contrast to the RapGAP domain phylogeny, these five ancestral groups are not only supported by the gene tree, but also by their domain architectures (i.e. the composition and order of protein domains in a protein). The domain architectures of the RasGAP domain containing proteins are highly conserved and can, in part, be reconstructed to the ancestral RasGAPs in LECA. Below we discuss the phylogenetic reconstruction and ancestral domain architectures of the RasGAP domain containing proteins.

The GAPVD1-like RasGAPs; an ancestral link between Rab and Ras signaling

GAPVD1 (also known as RAP6, Gapex-5 or RME-6) represents a conserved type of RasGAP that can be found in animals, the amoeba *D. discoideum*, the oomycete *P. infestans* and the ciliates *T. tetrahymena* and *P. tetraurelia*. Interestingly we identified the RasGAP domain followed by the VPS9 domain in both the animal sequences and in the ciliate sequences. Even though there are no sequences found in representative species of the other phyla (e.g. plants and excavata), the identification of the GAPVD1 orthologs in the Chromalveolata (i.e. the ciliates and oomycetes) indicates that GAPVD1 is ancient and perhaps originated from LECA.

GAPVD1 has been characterized in *C. elegans*, mouse and human [13-15]. The VSP9 domain exhibits Guanine Exchange Factor activity towards Rab5-like GTPases and the RasGAP domain has been shown to bind to H-Ras and stimulate its GTPase hydrolysis in human[15]. GAPVD1 therefore forms a conserved bridge between Ras and Rab signaling.

The GAP1 and IQGAP-like RasGAPs; diversification between the animal and fungal RasGAP repertoire

The human IQGAPs and the *S. pombe* GAP1 [genedb: SPBC646.12c] each represent an ancestral RasGAP gene in LECA. Both families are characterized by a highly conserved RasGAP-RasGAP_C domain architecture, which does not occur in other GAP proteins. We identified IQGAP-like genes in animals, fungi (e.g. IQG1 in *S. cerevisiae*) including *Enchephalitozoon cuniculi*, the amoeba *D. discoideum* and the excavate *N. gruberi*. The N-terminal CH domain of the IQGAPs is highly conserved in all species, even though the IQ repeats, characteristic for this RasGAP family, is only conserved in animals and fungi. The CH-RasGAP-RasGAP_C architecture therefore represents the ancestral domain architecture of the IQGAP ancestral gene in LECA.

The GAP1-like RasGAP does not occur in higher animals but can be found in the excavate *N. gruberi*, the amoeba *D. discoideum*, the choanoflagellate *Monosiga brevicollis*, the primitive animal *Trichoplax adhaerens* and fungi. In fungi however the GAP1 gene appears to have been lost in the *Saccharomyceae* (*S. cerevisiae*, *Candida glabrata*, *Kluyveromyces lactis*, *Eremothecium gossypii*, *Candida guilliermondii* and *Debaryomyces hansenii*). The presence of a GAP1 ortholog in the placozoan *T. adhaerens* and choanoflagellate *M. brevicollis* indicates that GAP1 has been lost early in the animal lineage. The conserved RasGAP-RasGAP_C domain architecture and position in the gene tree next to the IQGAP-like sequences indicates that both groups originated from a single ancestral gene that predates LECA and already had the RasGAP-RasGAP_C domain architecture. This particular domain architecture is not shared with any of the other RasGAP types.

The NF1-like RasGAPs; an ancient link between Ras and Rheb signaling?

Neurofibromin 1 (NF1) is RasGAP and a known tumor suppressor. NF1 mediates Ras

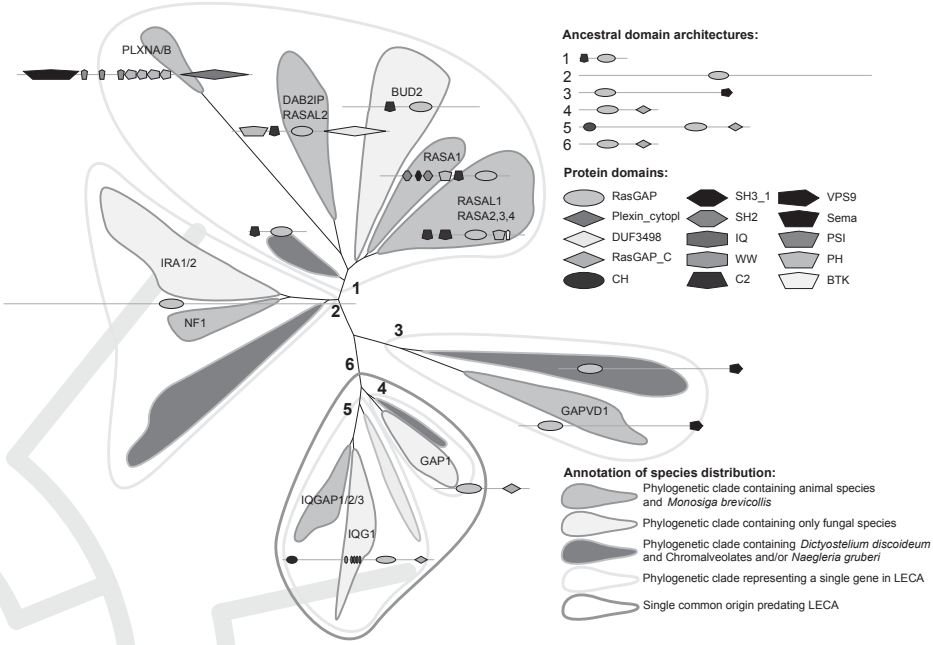


Figure 5.3 Simplified representation of the phylogenetic tree of RasGAP containing protein sequences. Gene names from human and *S. cerevisiae* are depicted as representative genes. Representative domain architectures for each clade are depicted in the tree and the five clades representing a single ancestral gene in or before LECA are numbered. The GAP1 and IQGAP clades from a combined clade in which a domain architecture is shared that predates LECA. The tree depicted has been modified to reflect proper grouping based on domain architecture: the clade containing RASA-like genes in *D. discoideum*, chromalveolates and *N. gruberi* is found between clusters 2 and (3,6) in the ML tree, see Figure S5.2.

dependent mTOR activation via TSC2 [16]. Dysfunction of NF1 in mammals results in a disorder called neurofibromatosis which is characterized by the development of neurofibromas, that are benign tumors found on and around the peripheral nerves. We find NF1 orthologs in all animals and in fungi where the gene has duplicated in the yeasts *S. cerevisiae* and *C. glabrata* (IRA1 and IRA2 in *S. cerevisiae*). We also find multiple NF1-like RasGAP sequences in *N. gruberi*, *D. discoideum*, *P. sojae* and *P. infestans*. The sequences of one gene in *P. infestans* [PITG_09962] and *P. sojae* [131833] and two genes in the excavate *N. gruberi* [80941, 73399] are conserved over the full length of NF1. This strongly indicates that the ancestral gene in LECA likely coded for a NF1-like protein.

The RASA-like RasGAPs; origins of the C2-RasGAP domain architecture, linking Ras and Rap signaling in evolution

The RASA-like RasGAPs contain by far the most human RasGAPs (10 of 17, excluding the Plexin genes, see below) and includes mammalian RASA1-4, SynGAP1 and yeast Bud2. While the other classes of RasGAPs contain only one group of animal genes, the RASA-like RasGAPs have undergone an animal specific expansion. Additionally, this class has undergone extensive protein domain rearrangements and domain acquisitions. However, nearly all sequences contain at least the C2 domain followed by the RasGAP domain. This indicates that although the overall bootstrap values for this class is low, the clustering itself is correct.

Proteins belonging to the RASA-like RasGAPs have been reported to be dual specific towards Ras- and Rap GTPases [17-19]. Pena et al. [20] has shown that the individual RasGAP domain of SynGAP1 specifically deactivates Ras, while the full length protein is a Rap specific GAP. They also showed that the C2 domain is necessary for its Rap specificity. The phylogenetic distribution of the C2 – RasGAP domain architecture matches the phylogenetic distribution of the Rap GTPase subtype suggesting that the C2 – RasGAP architecture may represent conservation of specificity towards the Rap GTPases and therefore represents an evolutionary link between the RASA-like RasGAPs and Rap signaling.

We note above that Rap specific RapGAP domain containing proteins do not occur in fungal species belonging to the basidiomycetes and ascomycetes. A known GAP for the fungal Rap ortholog BUD1 in the yeast *S. cerevisiae* is the BUD2 gene [17]. BUD2 belongs to the RASA-like GTPases and contains the C2-RasGAP domain architecture. Therefore, BUD2 and its fungal orthologs fill the gap left by the loss of the RapGAP domain containing proteins in fungal Rap signaling.

The Plexin type RasGAPs; unclear origin

The Plexin type RasGAPs cluster in the RASA-like RasGAPs but do not share the C2 domain or any other domain except for the RasGAP domain. The Plexin type RasGAPs from an animal specific group of genes but show a particularly long branch indicating a period of fast sequence evolution of the RasGAP domain sequence early in or just before the emergence of animals in evolution. Therefore it is uncertain if the Plexin type RasGAPs indeed belongs to the RASA-like RasGAPs or if its placement is a Long Branch Attraction artifact (LBA) of the Maximum Likelihood inference.

Localized signaling and cross-signaling in Ras signaling network

An interesting observation is that although both the RasGAP and RapGAP domains had multiple copies in LECA we observe conservation of protein domain architecture only in the RasGAP family. Two of three protein domains that are conserved in LECA together with the RasGAP domain are involved in localization (CH for binding to the actin skeleton, C2 for Ca^{2+} dependent membrane targeting). This strengthens the idea that sub-localization is an crucial factor in Ras-like GTPase signaling [21]. The conservation of the VPS9 domain (a GEF domain for Rab5-like GTPases) and the C2 domain with the RasGAP domain across the eukaryotic kingdom is indicative for the interconnectivity of Ras signaling pathways with other signaling pathways. This interconnectivity, or cross-signaling, is one of the hallmarks of Ras signal transduction.

5.3.3 A high resolution phylogeny of the Ras-like subfamily of small GTPases

We identified small GTPase protein sequences in the genomes of 64 divergent eukaryotic genomes using the Pfam [22] Ras HMM model (Pfam PF00071.15). The small GTPase super family members all share a high degree of sequence similarity and the Pfam Ras model therefore also matches significantly to hits belonging to the other small GTPase subfamilies, such as Rho, Rab and Ran. In order to identify Ras-like subfamily members we aligned all small GTPase sequences and constructed a phylogenetic tree using the neighbor joining algorithm. We identified a cluster that contained all currently classified mammalian and yeast Ras-like GTPase subfamily members but did not contain small

GTPase members belonging to the other subfamilies. Therefore we can confidently classify all GTPases in this cluster, mostly belonging to species for which there is no gene function information available, as belonging to the Ras-like subfamily of GTPases.

A phylogeny of the sequences belonging to the Ras-like subfamily cluster was constructed by using RAxML [23] and PhyML [24]. The evolutionary relationships between the Ras-like GTPases were determined as well as the timing of the duplications based on species distribution in each phylogenetic branch (see Figure 5.4 and Supplementary data).

We define Ras-like GTPase subtypes based on their relatedness as observed in clustering in the phylogenetic trees. We can trace back the origin of four subtypes to in or before the LECA (i.e. Ras, Ral, Rheb and Rap), but for other subtypes we are unable to do so. These other subtypes do not affiliate specifically to any other subtypes and encompasses only metazoan species implicating a more recent origin. The many isoforms of each subtype observed in vertebrates, for instance Rap1A and Rap1B, are likely the results of multiple whole genome duplications (WGD) that occurred in the vertebrate ancestor. Below we will discuss the evolution of each Ras-like GTPase subtype in more detail.

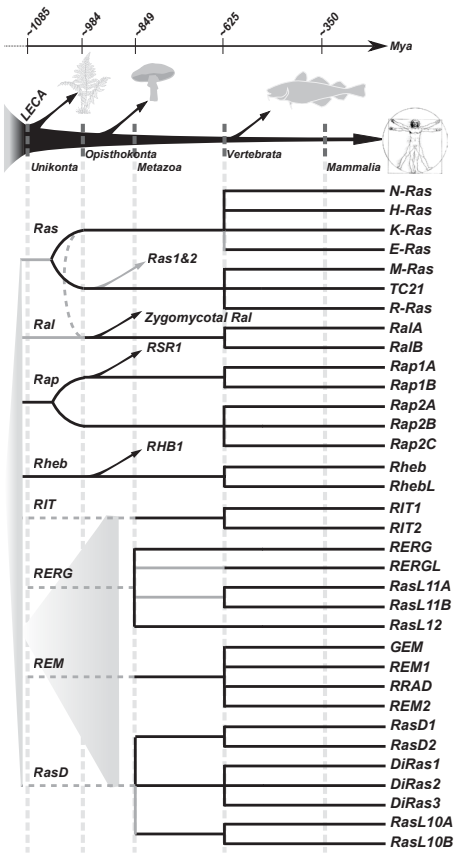


Figure 5.4 Evolutionary timeline of the Ras-like subfamily members. Many duplication events occurred in the common ancestor of the metazoa and vertebrates. The LECA likely contained multiple ancestral Ras-like GTPases although it is impossible to resolve how many there exactly were. Confident branches are depicted in black. Grey lines indicate possible alternative interpretations but are supported by circumstantial evidence. Gray dotted lines are based on a strict interpretation of the position of these GTPases in the phylogenetic tree but are otherwise unsupported and may originate from a more recent ancestral gene in the Unikont or Opisthokont ancestor. Fungal gene names are based on the orthologous genes in the yeast *S. cerevisiae*. The evolutionary time scale is based on Douzery et al. [34] but note that molecular dating is highly inaccurate [42]and that these dates are therefore approximate at best.

The Ras subtype (canonical)

The canonimal Ras-like GTPases N-, H-, K-Ras and the closely related TC21, R- and M-Ras regulate diverse cellular functions such as cell differentiation, apoptosis, cell proliferation, cell division and functions as input into many other signaling pathways. All canonical Ras GTPases require many of the same GEF and GAP proteins for regulation [7]. Indeed we find that all six are closely related, but form a bi-partitioned group: one consisting out of N-, H-, K-Ras and the other consisting out of TC21, R- and M-Ras (Figure 5.4). We find that E-Ras, generally not considered to be closely related to the Ras subtype [25] is directly related to N-, H-, K-Ras (bootstrap support of 82%). We

find that the fungal orthologs of Ras (Ras1 and Ras2 in *Saccharomyces cerevisiae*, Ras1 in *Schizosaccaromyces pombe*) are orthologous (i.e. sharing common ancestry) to TC21, R- and M-Ras but are paralogous to N-, H-, K- and E-Ras, whom appear to have diverged earlier. However, the phylogenetic signal is weak (0% and 1% bootstrap support for the PhyML and RAxML trees respectively) and therefore an alternative evolutionary scenario could be that the N-, H-, K-, E-Ras and TC21, M- and R-Ras have arisen from animal specific duplications and all are orthologous to the fungal Ras genes. Nonetheless, GTPase genes from the slime mould *Dictyostelium discoideum* are found in both the N-, H-, K-, E-Ras partition and the TC21, M- and R-Ras partition, supporting the first scenario. Because of the uncertainty concerning the placement of the fungal Ras orthologs, Ras1 & Ras2 are colored gray in Figure 5.4.

The Ral subtype; Opisthokont origin or much older?

The Ral GTPases (RalA and RalB) function in a number of different cellular processes, including the regulation of exocytosis [26,27]. Ral has its own complement of GEFs (RalGDS- and RalGPS-like proteins) and GAPs (RalGAPA/B complex [10]). Previously we observed that the RalGDS and RalGPS genes form a single evolutionary gene family with fungal orthologs in the Chytrid and Zygomycetae fungi [8]. We also identified putative Ral orthologs in the same fungi indicating a strong evolutionary link between Ral and the RalGEFs.

It is unclear whether Ral clusters into the Ras subtype or forms a separate cluster, since different phylogenetic methods (e.g. maximum likelihood, neighbor joining) and programs (RAxML or PhyML) reached different conclusions (see supplementary material). This raises the question of how old the Ral subtype truly is. The species that contain a clear Ral ortholog all belong to the Opisthokonta (animals and fungi), so clear evidence is lacking that Ral may be older than the Opisthokonta. However, the fact that Ral has a GAP which is distinct from Rap- and RhebGAPs (see section RalGAPs) may indicate that Ral form their own evolutionary subtype and that Ral emerged earlier in eukaryotic evolution and subsequently has been lost in all other phyla.

The Rap subtype;

Rap has a suppressing effect on the Ras oncogenes [28] and have therefore attracted much attention in the GTPase field. Rap1 (Rap1A and Rap1B) is known to regulate many aspects of cell-cell adhesion, development of focal adhesions and junction formation. Although similar functions for Rap2 (Rap2A/B/C) are observed it is currently still unclear what the distinct cellular function of Rap2 is relative to Rap1 [21].

We identify Rap orthologs in animals, fungi, the amoeba *D. discoideum* and the excavate *N. gruberi*. The *N. gruberi* Rap ortholog (55569) indicates that Rap might be older than the Unikonta and might even have originated in or before LECA. This is supported by the presence of RapGEF orthologs and a RASA RasGAP ortholog in *N. gruberi* (Chapter 3 [8] and this chapter). Rap must have subsequently been lost in the Archaeplastida, Chromalveolates and specific lineages within the Excavata. It has been observed that *N. gruberi* contains many genes belonging to orthologous groups of genes previously thought to be Unikont specific [29]. Therefore our observation of the phylogenetic distribution of the Rap-like GTPases is not uncommon.

The Rap1 and Rap2 genes from two separate groups within the Rap subtype and have arisen from a duplication in the Unikont or Opisthokont ancestor, depending on the

placement of orthologous genes from the amoeba *D. discoideum* (Figure 5.4). Fungal Rap orthologs (e.g. BUD1 in *S. cerevisiae*) consistently cluster to Rap1 but not Rap2, indicating that fungi have lost the Rap2 gene. *S. pombe* appears to also have lost the Rap1 ortholog as well.

The Rheb subtype

Rheb is an integral part of the TOR pathway and plays a central role in the regulation of the TOR complex 1 (TORC1). Activation of Rheb results in the activation of TORC1 which leads to increased translation and growth [30]. Via its GAP, the TSC1/2 complex, Rheb integrates many signals including nutrient availability and growth factor signaling [31-33]. The TSC2 protein harbors a RapGAP domain that is needed for its catalytic activity towards Rheb.

Previously, the phylogeny of Rheb has been extensively discussed in light of the TOR pathway by us (Chapter 4). We found that the Rheb subtype is well conserved in all eukaryotic lineages and we have been able to identify clear orthologs in animals, fungi, amoebazoa, stramenophila (chromalveolates), excavates and even archaeplastida (specifically in the red alga *Cyanidioschyzon merolae*). The Rheb-like GTPases are generally found as a single copy in most eukaryotic species, except vertebrates in which Rheb has duplicated giving rise to the Rheb and RhebL1 genes. The Rheb subtype is well established and consistent in the phylogenetic trees. The Rheb subtype represents the least ambiguous orthologous group of Ras-like GTPases spanning all major eukaryotic phyla and very likely represents a single ancestral GTPase in LECA (see Figure 5.4).

Other Ras-like GTPase subtypes

The Ras, Ral, Rap and Rheb subtypes represent the most well characterized and frequently studied Ras-like GTPases. However, there are more Ras-like GTPases (18 out of 34 Ras-like GTPases in human) including the RIT, RERG, GEM, and DiRAS GTPases. To the best of our knowledge there has been no (confirmed) report on the characterization of any GAP or GEF proteins specific for these Ras-like GTPases considerably hampering biochemical characterization of the pathways in which these GTPases are involved.

Interestingly for the Ras-like GTPases other than the Ras, Rap, Ral and Rheb subtypes we cannot identify any orthologs in *S. cerevisiae*, *S. pombe* or other fungi. In fact we only find well defined animal specific clusters, but very quickly the resolution in the trees is lost due to weak bootstrap support and/or absence of sequences from non-animal species (see Figure 5.4). The subtypes rarely cluster together (except the RERG and REM subtypes in the NJ tree and RERG and RIT subtypes in the RAxML tree, but both with a bootstrap value of 0). In fact none of the Ras-like GTPase subtypes cluster together consistently, for which the low bootstrap support is indicative. It is therefore impossible to determine with any confidence the origin of groups we have termed the REM, RasD, RERG and the RIT subtypes.

The number of ancestral GTPases that gave rise to the REM, RasD, RERG and the RIT subtypes may vary between one and four. Either these subtypes are the result of multiple rounds of duplications from a single ancestral gene in the animal ancestor and was lost only once in the other major phyla (i.e. a single ancestral gene in LECA) or all four subtypes emerged in or before LECA and all subtypes were individually lost in the major phyla, except the Metazoa. The first possibility requires the least evolutionary events to describe the observed tree and must therefore be considered as the most

likely scenario.

5.3.4 The early emergence of the Ras, Rap, Ral and Rheb GTPases an their regulatory domains

The short sequence length of the small GTPases (average sequence length of 159 aa) is restricting the reliability and reconstruction of the phylogenetic relationships between the GTPases. We may however be able to at least refine the phylogenetic relationships between the Ras, Rap, Ral and Rheb GTPase subtypes. We previously observed a strong evolutionary relation between Ras-like GTPases Ras, Rap and Ral and the CDC25HD [8] and also between Rheb and the TSC2 GAP domain (Chapter 4 and this chapter) and Rap and RASA-like RasGAPs. The RasGEF domain (the CDC25HD) and the GAP domains (RasGAP and RapGAP) provide catalytic activity to two or more of the Ras, Ral, Rap and Rheb subtypes and therefore could possibly be used to infer the order of divergence into these four subtypes. Below we discuss our results on the GAPs and our previous study on the evolution of the CDC25HD (i.e. the RasGEF domain) in light of the Ras-like GTPase subtypes.

In Figure 5.5A we have condensed the specificity of the regulatory domains to their GTPase subtypes into a matrix. For each possible evolutionary scenario for the Ras-like

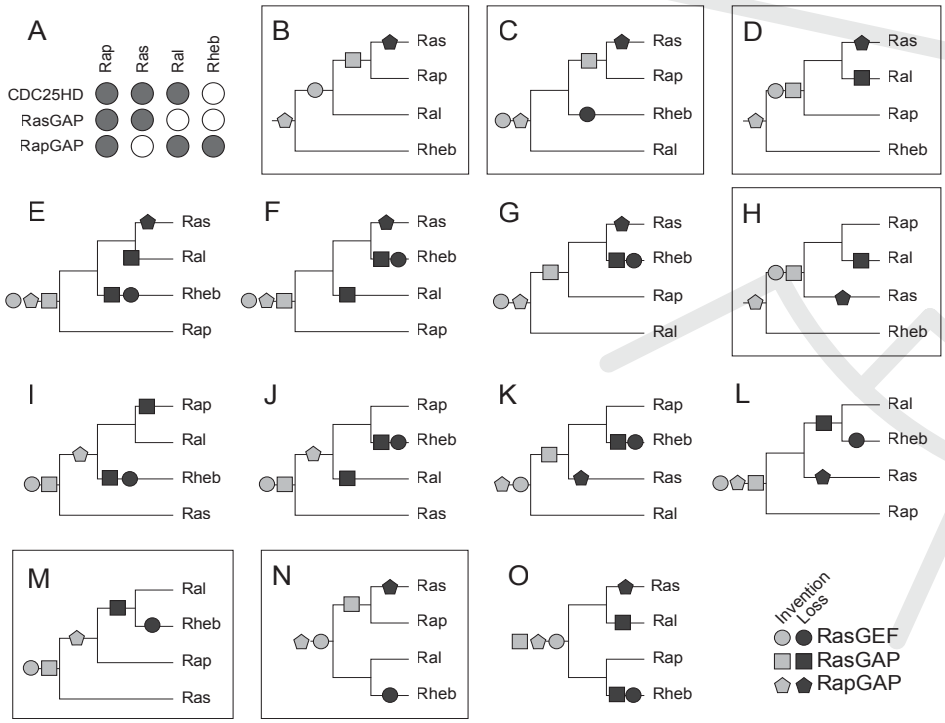


Figure 5.5 Evolutionary reconstruction of the Ras, Ral, Rap and Rheb GTPases based on co-evolution with their regulatory domains. A) Matrix describing for each GTPase subtype which regulatory domain is active. B-O) All scenarios possible for the order of duplications that gave rise to Ras, Rap, Ral and Rheb. For each scenario we reconstructed the order of invention and loss of regulation by the regulatory domains to fit the observed regulation of the GTPases as depicted in panel A. Panels that exhibit the minimal number of events needed to fit the matrix in panel A (panels B (4) and panels C,D,H,M,N (5 events) are marked.

GTPase subtypes we have reconstructed the point of invention and or loss of subtype specific regulatory domains while trying to minimize the number of events (i.e. the number of inventions and losses, see Figure 5.5B-O). Of the 14 possible scenarios there is only one scenario (Figure 5.5B) that requires the least amount of events (four) to describe the matrix in Figure 5.5A. Five scenarios require five events (Figure 5.5C, D, H, M and N) and the remaining eight scenario's require 6 or 7 events. The scenario depicted in Figure 5.5B therefore represents the most parsimonious scenario of how the Ras-like subtypes could have differentiated from each other before the LECA. It also depicts clearly when and in which order the regulatory domains emerged that regulate current day Ras-like GTPases.

The parsimonious scenario in Figure 5.5B describes a chain of successive inventions of regulatory domains and the duplications of the Ras-like GTPases in the eukaryotic ancestors before the radiation in all the current eukaryotic phyla. According to this scenario the invention of the RapGAP domain preceded the emergence of the Ras, Rap, Ral and Rheb GTPases in an early eukaryotic ancestor predating LECA. Subsequently Rheb diverged from the Ras/Rap/Ral ancestral GTPase and the RasGEF domain was invented. The early divergence of Rheb explains the lack of observed Rheb specific RasGAP and RasGEFs. After the addition of the RasGEF domain the Ral subtype diverged from the Ras/Rap ancestral GTPase. At this point all regulatory domains for Ral already existed (i.e. the RasGEF and RapGAP domains). Finally the RasGAP protein domain was invented after which the Rap and Ras subtypes diverged from each other. The Ras subtype then lost regulation by RapGAP domain containing regulatory proteins. Only after the divergence of all four Ras-like GTPase subtypes did the radiative expansion of eukaryotes take place.

5.3.5 Expansion of the GTPases and their regulatory domains in animals

We have shown that the Ras-like GTPases and their regulatory domains (i.e. the RasGEF, RasGAP and RapGAP domain) are highly conserved in the eukaryotic lineages and that there is compelling evidence that all protein families occurred in multiple copies in the LECA. We have described in detail the phylogeny of all domains and protein families involved. Here we will discuss the evolutionary dynamics (i.e. expansion of the protein families over time) that these protein families display. In Figure 5.6 we show the expansion of the GTPases and their specific GEF and GAP proteins for each of the Ras, Ral, Rap and Rheb subtypes.

The number of GTPases and regulatory proteins remained relatively equal for a long time in evolution (LECA leading up to the Metazoa) but have greatly expanded (doubled) in the vertebrate ancestor (a period of roughly 500 million years [34] see Figure 5.1, Figure 5.2, Figure 5.3 and Figure 5.4). The Ras and Rap GTPase subtypes and their regulators contributed the most to this expansion. The Rheb and Ral subtypes and regulators remained mostly constant in numbers with only a significant increase of the Ral specific GEFs in the vertebrate ancestor.

Interestingly the Ras and Rap subtype exhibit different dynamics compared to each other in regards to their regulatory domains. In the early expansion of the eukaryotic lineage the Ras subtype supported more GAP and GEF regulatory genes than there were Ras GTPases (five GEFs and five GAPs to one ancestral Ras GTPase). However, leading up to the vertebrate lineages the number of Ras specific GEFs remained equal as duplications were balanced by multiple loss events (Figure 5.6 and see Figure 5.1). In contrast, the

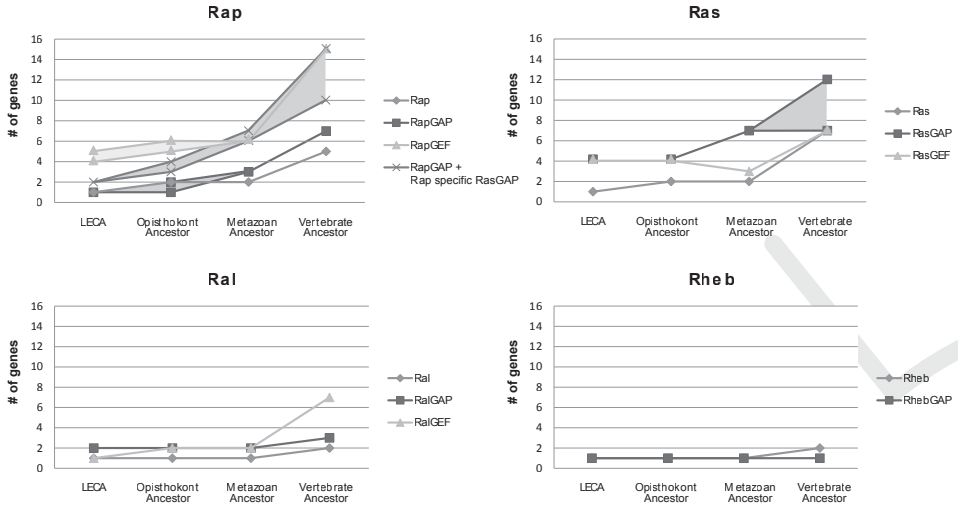


Figure 5.6 Expansion of Ras-like GTPase subtypes and their respective GEFs and GAPs (Ras-like regulatory system) in time leading up to mammalian organisms. We incorporated upper and lower estimates for GEFs and GAPs based on how the phylogenetic trees can be interpreted. The difference between higher and lower estimates for RapGAPs and RasGAPs is mainly caused by the RASA1 RasGAP protein family as GTPase specificity shifts within this family (see main text). The Rap GTPase regulatory system display a many to one GEF regulation network. This scheme is maintained throughout eukaryotic evolution leading up to vertebrates. However with the inclusion of the C2-RasGAPs, the GAP regulation of Rap shows a similar trend. In contrast, the Ras regulatory system displays a many to one GAP regulation, while maintaining a relatively equal amount of Ras GTPase specific GEFs. The difference in the rate of expansion of the regulatory proteins for the Ras and Rap GTPase subtype indicates that there is a fundamental difference between the regulatory networks of Ras and Rap. Where Rap relies on multiple GEF and GAP proteins for its signaling diversification, Ras seems to rely mostly on its GAPs. The Ral and Rheb GTPases display a relatively compact regulatory system although in the vertebrates the number of RalGEFs expand significantly.

RasGAP domain greatly increased in numbers, hence indicating a relevant importance of regulation via GAP proteins over regulation by GEF proteins in the expansion of the Ras subtype.

In contrast to the Ras subtype, the Rap subtype shows a dependency on GEF regulation over GAP regulation in the early eukaryote as GTPase and GAP numbers are similar (one and two genes respectively), but already supported seven Rap specific GEFs. Interestingly, the number of GAPs increased consistently in time and match the number of Ras specific GEFs in the vertebrate ancestor (14 RapGEFs and 14 RapGAPs of which seven contain the RapGAP domain and another seven contain the C2 – RasGAP domain architecture). The Rap subtype therefore displays an ancestral dependency on regulation by its GEFs in strong contrast to Ras. However, regulation by GAP activity at some point in the Opisthokont lineage acquired importance and the Rap specific GAPs started to expand. The slow expansion of the Ral and Rheb subtype regulatory systems is most likely linked to their specialized role in cellular signaling and regulation. The Rheb GTPase has a singular role in the eukaryotic TOR pathway, namely to regulating TOR activity based on nutrient availability and growth signals [35]. The Ral GTPase plays a critical role in exocytosis in animals [26,27]. The Rap and Ras subtypes play important roles in multiple pathways and provide many cross-signaling routes between distinct signaling pathways. These multiple roles and a suspected need to separate the cellular role of activated Ras

and Rap in the multitude of cross-signaling pathways might provide a reason for the expansion of the Ras and Rap regulatory systems compared to Ral and Rheb.

5.3.6 Evolution of Ras-like GTPase regulation

We have discussed in detail the evolution of the Ras-like GTPases and their GAPs and GEFs and focused mainly on the co-evolution of the GTPases and their regulatory domains (i.e. the RasGAP, RapGAP and RasGEF domains). However, protein domains other than the GAP and GEF domains play an important role in the cellular functions of RasGEFs, RasGAPs and RapGAPs (for instants localization and regulation of catalytic activity by the GAP or GEF domains). We find that the domain architectures of these regulatory protein families are generally not conserved beyond the Metazoa (RasGEFs see Chapter 3 [8], RapGAPs see Figure S5.1), suggesting intrinsic adaptability of Ras-like GTPase regulation via acquisition and loss of regulatory domains in RasGEFs and RapGAPs. However RasGAPs clearly show strong conservation of ancestral domain architectures. This indicates that the adaptability of Ras regulation via GAPs is constrained. Additionally, the expansion of the Ras-like GTPases and their regulatory proteins in evolution show that Ras GTPases are preferably regulated by GAPs while Rap GTPases are preferably regulated by GEFs. Together, the conservation of RasGAP domain architecture and an evolutionary preference for regulation via GAPs indicates a necessity for strict down regulation of canonical Ras GTPases.

By analyzing the evolution of not only the Ras-like GTPases but also the evolution of their regulatory domains we are able to paint a complete evolutionary scenario of the Ras-like GTPase protein family and it regulation. We have shown that the RasGAP, RapGAP and RasGEF domains co-evolved coherently with their respective GTPases indicating a robustness in the way Ras-like GTPase activity are regulated throughout evolution. On the other hand, we observe flexibility in how the GAPs and GEFs themselves are regulated as observed in the variability of the domain architectures observed in the specific subtypes of GAPs and GEFs. We believe that this aspect, the regulation of the regulators, makes the Ras-like GTPase family a most versatile molecular switch in eukaryotic evolution. Altogether, our analysis provides a detailed evolutionary framework but also provides specific predictions for molecular biologists and biochemists working on Ras-like GTPases and their signaling pathways.

5.4 Methods

5.4.1 Genome selection

We acquired best model protein sequences of 64 divergent eukaryotic species from EnsEMBL [36], JGI, the Broad institute or their respective genome project sites. We have selected a wide range of animal and fungal genomes as most research on Ras signaling is being done in either animal or fungal model organisms. We also included a wide range of genomes belonging to other major phyla, such as the archeaplastida, chromalveolates and excavates, to be able to accurately time the duplication and loss events of the Ras-like GTPases as well as for the RasGAP and RapGAP domains. For a full overview of genomes, source and version information see Table S5.1.

5.4.2 Phylogenetic analyses

Identification of Ras-like GTPases and phylogenetic analysis

The sequences of the selected genomes were searched using the Pfam [22] HMM profile for the Ras family (Pfam accession PF00071.12, Pfam version 23) and hmmsearch of the HMMER package version 2.3.2 [37]. All sequences with a bitscore larger than 0 were selected. Due to the high sequence similarity of Ras to other small GTPases many other small GTPases are included in this set. An alignment of all sequences was made using the MAFFT program [38] with the ginsi option. A neighbor joining tree was constructed using the Quicktree program [39]. A sub tree was selected which contained all Ras-like subfamily members but no other small GTPases. The sequences were gathered from the initial alignment as manual inspection of the alignment produced from the subset showed it was suboptimal to the initial alignment. Subsequently a phylogenetic tree was constructed over all Ras-like subfamily members using RAxML [23] (-T 4 -x 488761235 -f a -N 1000 -m PROTGAMMAIWAG), PhyML [24] (phyml <file> 1 i 1 0 WAG e 6 e BIONJ y y) and Quicktree. For the Quicktree and RAxML analyses a 1000 bootstrap runs were performed. However the bootstrapping method implemented by the PhyML program is very slow compared to the RAxML bootstrap algorithm. We therefore used the bootstrap data from the RAxML run to calculate bootstrap values for the PhyML tree. All phylogenetic trees were visualized using Dendroscope [40].

Identification of RapGAP domain containing proteins and phylogenetic analysis

We gathered RapGAP domain sequences from the sequence set by using a custom made HMM model and hmmsearch of the HMMER package version 2.3.2. The custom RapGAP HMM model is based on an edited alignment of RapGAP domain sequences from a previous hmmsearch run using the Pfam RapGAP model. We did not use HMMER version 3 as the latest HMMER version does not allow full length domain detection (ls) The domain sequences were aligned using MAFFT ginsi. A phylogenetic tree was constructed using RAxML (-T 6 -x 23421421 -f a -N 1000 -m PROTGAMMAIWAG). The domain architectures of the RapGAP containing proteins were visualized using the iTOL webserver application[41]. The domain architecture datasets as needed by the iTOL server were build using a custom perl script and data from hmmscan (HMMER package version 3.0b3) and Pfam version 24 hmm models.

Identification of RasGAP domain containing proteins and phylogenetic analysis

We gathered RasGAP domain containing proteins sequences from the sequence set by using the Pfam version 24 HMM model with hmmsearch of the HMMER package version 3.0b3. The full length sequences were aligned using MAFFT ginsi to include conserved sequence positions bordering the detected Pfam domain (positions 2870-5010). A phylogenetic tree was constructed using RAxML (-T 8 -x 57231793 -f a -N 1000 -m PROTGAMMAIWAG). The domain architectures of the RasGAP containing proteins were visualized using the iTOL webserver application.

5.5 Acknowledgements

We would like to thank Jos Boekhorst, Gabino Sanchez-Perez, Like Fokkens, Michael Seidl for their help in performing the analyses and support. This work is part of the BioRange program of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK

grant through the Netherlands Genomics Initiative (NGI). The sequence data of selected genomes were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/, in collaboration with the user community, and the Fungal Genome Initiative of the Broad Institute. For a full overview of the genomes see supplementary Table S5.1.

5.6 Supplementary material

Due to the amount of detail Supplemetary figures S5.1 and S5.2 are best viewed electronically. Supplemetary figures S5.1, S5.2 and newick format trees of the Ras-like GTPases can be downloaded from http://bioinformatics.bio.uu.nl/john/thesis/.

Table S5.1. Genomes used in this study

Species	Genome source	Genome version	References
Homo sapiens	EnsEMBL	NCBI 36	[72,73]
Mus musculus	EnsEMBL	NCBI m37 Apr 2007	[74]
Ornithorhynchus anatinus	EnsEMBL	v5.0, Dec 2005	[75]
Gallus gallus	EnsEMBL	WASHUC2, May 2006	[76]
Xenopus tropicalis	EnsEMBL	JGI 4.1, Aug 2005	[77]
Danio rerio	EnsEMBL	Zv7, Apr 2007	[62]
Fugu rubripes	EnsEMBL	FUGU 4.0, Jun 2005	[78]
Branchiostoma floridae	JGI	v.1.0 (December 5, 2006)	[79]
Homo sapiens	EnsEMBL	NCBI 36	[42,43]
Mus musculus	EnsEMBL	NCBI m37 Apr 2007	[44]
Ornithorhynchus anatinus	EnsEMBL	v5.0, Dec 2005	[45]
Gallus gallus	EnsEMBL	WASHUC2, May 2006	[46]
Xenopus tropicalis	EnsEMBL	JGI 4.1, Aug 2005	[47]
Danio rerio	EnsEMBL	Zv7, Apr 2007	[36]
Fugu rubripes	EnsEMBL	FUGU 4.0, Jun 2005	[48]
Branchiostoma floridae	JGI	v.1.0 (December 5, 2006)	[49]
Ciona intestinalis	EnsEMBL	JGI 2, Mar 2005	[50]
Strongylocentrotus purpuratus	Sea Urchin Genome Project	Spur 2.1, Sep 2006	[51]
Caenorhabditis elegans	EnsEMBL	WS190, Apr 2008	[52]
Anopheles gambiae	EnsEMBL	AgamP3, February 2006	[53]
Drospohila melanogaster	EnsEMBL	BDGP 5.4, Nov 2007	[54]
Lottia gigantea	JGI	v1.0 (July 24, 2007)	[55]
Daphnia pulex	JGI	v.1.0 (July 5, 2007) frozen catalog	[55]
Nematostella vectensis	JGI	JGI v1.0	[56]
Trichoplax adhaerens	JGI	JGI v1.0 22 July 2007	[57]
Monosiga brevicollis	JGI	JGI v1.0 July 2006	[58]
Encephalitozoon cuniculi	EMBL	EMBL	[59]
Batrachochytrium dendrobatidis	BROAD	BROAD 3/1/2007	[60]
Phycomyces blakesleeanus	JGI	JGI v1.0 January 8, 2007	[55]
Rhizopus oryzae	BROAD	RO3 12/6/05	[61]
Laccaria bicolor	JGI	JGI v1.0 May 22, 2006	[62]

Species	Genome source	Genome version	References
Phanerochaete chrysosporium	JGI	JGI v2.0 February 2005	[63]
Cryptococcus neoformans	BROAD	BROAD v3.0 2/17/2006	[64]
Ustilago maydis	BROAD	BROAD v2.0 April 1, 2004	[65]
Schizosaccharomyces pombe	Sanger	Sanger v19 07/16/2008	[66]
Yarrowia lipolytica	Genolevures	V 2	[67]
Debaryomyces hansenii	Genolevures	V 1	[67]
Candida guilliermondii	BROAD	V 1	[68]
Eremothecium gossypii	EMBL		[69]
Candida glabrata	Genolevures	V 2	[67]
Kluyveromyces lactis	Genolevures	V 2	[67]
Saccharomyces cerevisiae	SGD	SGD 06/06/2008	[70]
Aspergillus terreus	BROAD	BROAD v1.0	[71]
Coccidioides immitis	BROAD	C. immitis RS v2	[72]
Fusarium graminearum	BROAD	BROAD v3	[73]
Neurospora crassa	BROAD	BROAD v7	[74]
Sclerotinia sclerotiorum	BROAD	BROAD v2.0	[75]
Dictyostelium discoideum	DictyDB	created: 03-01-2009 01:29	[76,77]
Cyanidioschyzon merolae	C. merolae Genome Project	April 13 2004 (ORF)	[78]
Ostreococcus tauri	JGI	JGI v2	[79]
Chlamydomonas reinhardtii	JGI	JGI v3.1	[80]
Volvox carteri	JGI	JGI v1.0 June 1, 2007	[81]
Physcomitrella patens ssp patens	JGI	v.1.1 (March 2007)	[82]
Selaginella moellendorffii	JGI	v1.0 (December 20, 2007)	[55]
Oryza sativa	Rice Genome Annotation (TIGR)	v5 TIGR January 24, 2007	[83]
Arabidopsis thalia	Arabidopsis genome initiative	TIER v8	[84]
Populus trichocarpa	JGI	v1.0 (June 2004)	[85]
Phaeodactylum tricornutum	JGI	v.2.0 (November 16, 2006)	[86]
Thalassiosira pseudonana	JGI	v3.0 (August 2006)	[87]
Phytophthora sojae	JGI	v1.0 (April 2004)	[88]
Phytophthora infestans	BROAD	3/14/2007	[89]
Emiliania huxleyi CCMP1516	JGI	v.1.0 (April 25, 2008)	[55]
Aureococcus anophagefferens	JGI	JGI v1.0 September 27, 2007	[55]
Paramecium tetraurelia	ParameciumDB	V1.21 07/29/2008	[90]
Tetrahymena thermophila	TIGR	Aug-04	[91]
Cryptosporidium parvum	CryptoDB	CryptoDB 3.7	[92]
Plasmodium falciparum	PlasmoDB	5,5	[93]
Theileria parva	TIGR	1	[94]
Naegleria gruberi	JGI	v.1.0 (October 23, 2006)	[29]
Trichomonas vaginalis	TIGR	2/15/2007	[95,96]
Gardia intestinalis	GiardiaDB	V1.1 may 2008	[96]
Leishmania major	Sanger	11/5/2006	[97]
Trypanosoma brucei	Sanger	May 08 v4	[98]

5.7 References

1. Wennerberg K, Rossman KL, Der CJ (2005) The Ras superfamily at a glance. Journal of cell science 118: 843-6. doi:10.1242/jcs.01660
2. Yutin N, Wolf MY, Wolf YI, Koonin EV (2009) The origins of phagocytosis and eukaryogenesis. Biology direct 4: 9. doi:10.1186/1745-6150-4-9
3. Jékely G (2003) Small GTPases and the evolution of the eukaryotic cell. BioEssays : news and reviews in molecular, cellular and developmental biology 25: 1129-38. doi:10.1002/bies.10353
4. Brighouse A, Dacks JB, Field MC (2010) Rab protein evolution and the history of the eukaryotic endomembrane system. Cellular and molecular life sciences : CMLS: 1-17-17. doi:10.1007/s00018-010-0436-1
5. Boureux A, Vignal E, Faure S, Fort P (2007) Evolution of the Rho family of ras-like GTPases in eukaryotes. Molecular biology and evolution 24: 203-16. doi:10.1093/molbev/msl145
6. Elias M, Patron NJ, Keeling PJ (n.d.) The RAB family GTPase Rab1A from Plasmodium falciparum defines a unique paralog shared by chromalveolates and rhizaria. The Journal of eukaryotic microbiology 56: 348-56. doi:10.1111/j.1550-7408.2009.00408.x
7. Bos JL, Rehmann H, Wittinghofer A (2007) GEFs and GAPs: critical elements in the control of small G proteins. Cell 129: 865-77. doi:10.1016/j.cell.2007.05.018
8. Dam TJP van, Rehmann H, Bos JL, Snel B (2009) Phylogeny of the CDC25 homology domain reveals rapid differentiation of Ras pathways between early animals and fungi. Cellular signalling 21: 1579-85. doi:10.1016/j.cellsig.2009.06.004
9. Gridley S, Chavez JA, Lane WS, Lienhard GE (2006) Adipocytes contain a novel complex similar to the tuberous sclerosis complex. Cellular signalling 18: 1626-32. doi:10.1016/j.cellsig.2006.01.002
10. Shirakawa R, Fukai S, Kawato M, Higashi T, Kondo H, et al. (2009) Tuberous sclerosis tumor suppressor complex-like complexes act as GTPase-activating proteins for Ral GTPases. The Journal of biological chemistry 284: 21580-8. doi:10.1074/jbc.M109.012112
11. Scrima A, Thomas C, Deaconescu D, Wittinghofer A (2008) The Rap-RapGAP complex: GTP hydrolysis without catalytic glutamine and arginine residues. The EMBO journal 27: 1145-53. doi:10.1038/emboj.2008.30
12. Chen X-Q (1999) The Myotonic Dystrophy Kinase-related Cdc42-binding Kinase Is Involved in the Regulation of Neurite Outgrowth in PC12 Cells. Journal of Biological Chemistry 274: 19901-19905. doi:10.1074/jbc.274.28.19901
13. Lodhi IJ, Chiang S-H, Chang L, Vollenweider D, Watson RT, et al. (2007) Gapex-5, a Rab31 guanine nucleotide exchange factor that regulates Glut4 trafficking in adipocytes. Cell metabolism 5: 59-72. doi:10.1016/j.cmet.2006.12.006
14. Sato M, Sato K, Fonarev P, Huang C-J, Liou W, et al. (2005) Caenorhabditis elegans RME-6 is a novel regulator of RAB-5 at the clathrin-coated pit. Nature cell biology 7: 559-69. doi:10.1038/ncb1261
15. Hunker CM, Galvis A, Kruk I, Giambini H, Veisaga ML, et al. (2006) Rab5-activating protein 6, a novel endosomal protein with a role in endocytosis. Biochemical and biophysical research communications 340: 967-75. doi:10.1016/j.bbrc.2005.12.099
16. Johannessen CM, Reczek EE, James MF, Brems H, Legius E, et al. (2005) The NF1 tumor suppressor critically regulates TSC2 and mTOR. Proceedings of the National Academy of Sciences of the United States of America 102: 8573-8. doi:10.1073/pnas.0503224102
17. Park HO, Chant J, Herskowitz I (1993) BUD2 encodes a GTPase-activating protein for Bud1/Rsr1 necessary for proper bud-site selection in yeast. Nature 365: 269-74. doi:10.1038/365269a0
18. Krapivinsky G, Medina I, Krapivinsky L, Gapon S, Clapham DE (2004) SynGAP-MUPP1-CaMKII synaptic complexes regulate p38 MAP kinase activity and NMDA receptor-dependent synaptic AMPA receptor potentiation. Neuron 43: 563-74. doi:10.1016/j.neuron.2004.08.003
19. Kupzig S, Deaconescu D, Bouyoucef D, Walker SA, Liu Q, et al. (2006) GAP1 family members constitute bifunctional Ras and Rap GTPase-activating proteins. The Journal of biological chemistry 281: 9891-900. doi:10.1074/jbc.M512802200
20. Pena V, Hothorn M, Eberth A, Kaschau N, Parret A, et al. (2008) The C2 domain of SynGAP is essential for stimulation of the Rap GTPase reaction. EMBO reports 9: 350-5. doi:10.1038/embor.2008.20
21. Bos JL (2005) Linking Rap to cell adhesion. Current opinion in cell biology 17: 123-8. doi:10.1016/j.ceb.2005.02.009
22. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, et al. (2008) The Pfam protein families database. Nucleic Acids Res 36: D281-8. doi:10.1093/nar/gkm960
23. Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21: 456-463. doi:10.1093/bioinformatics/bti191
24. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696-704. doi:10.1080/10635150390235520
25. Colicelli J Human RAS superfamily proteins and related GTPases. Science's STKE : signal transduction knowledge environment 2004: RE13. doi:10.1126/stke.2502004re13
26. Moskalenko S, Henry DO, Rosse C, Mirey G, Camonis JH, et al. (2002) The exocyst is a Ral effector complex. Nat Cell Biol 4: 66-72. doi:10.1038/ncb728
27. Lipschutz JH, Mostov KE (2002) Exocytosis: the many masters of the exocyst. Curr Biol 12: R212-4. doi:10.1016/S0960-

9822(02)00753-4

28. Kitayama H, Sugimoto Y, Matsuzaki T, Ikawa Y, Noda M (1989) A ras-related gene with transformation suppressor activity. Cell 56: 77-84. doi:10.1016/0092-8674(89)90985-9
29. Fritz-Laylin LK, Prochnik SE, Ginger ML, Dacks JB, Carpenter ML, et al. (2010) The Genome of Naegleria gruberi Illuminates Early Eukaryotic Versatility. Cell 140: 631-642. doi:10.1016/j.cell.2010.01.032
30. Wullschleger S, Loewith R, Hall MN (2006) TOR signaling in growth and metabolism. Cell 124: 471-84. doi:10.1016/j.cell.2006.01.016
31. Inoki K, Ouyang H, Zhu T, Lindvall C, Wang Y, et al. (2006) TSC2 integrates Wnt and energy signals via a coordinated phosphorylation by AMPK and GSK3 to regulate cell growth. Cell 126: 955-68. doi:10.1016/j.cell.2006.06.055
32. Inoki K, Zhu T, Guan K-L (2003) TSC2 Mediates Cellular Energy Response to Control Cell Growth and Survival. Cell 115: 577-590. doi:10.1016/S0092-8674(03)00929-2
33. Garami A, Zwartkruis FJT, Nobukuni T, Joaquin M, Roccio M, et al. (2003) Insulin Activation of Rheb, a Mediator of mTOR/S6K/4E-BP Signaling, Is Inhibited by TSC1 and 2. Molecular Cell 11: 1457-1466. doi:10.1016/S1097-2765(03)00220-X
34. Douzery EJP, Snell EA, Bapteste E, Delsuc F, Philippe H (2004) The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? Proceedings of the National Academy of Sciences of the United States of America 101: 15386-91. doi:10.1073/pnas.0403984101
35. Aspuria P-J, Tamanoi F (2004) The Rheb family of GTP-binding proteins. Cellular signalling 16: 1105-12. doi:10.1016/j.cellsig.2004.03.019
36. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, et al. (2009) Ensembl 2009. Nucleic Acids Res 37: D690-7. doi:10.1093/nar/gkn828
37. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755-763. doi:10.1093/bioinformatics/14.9.755
38. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30: 3059-3066. doi:10.1093/nar/gkf436
39. Howe K, Bateman A, Durbin R (2002) QuickTree: building huge Neighbour-Joining trees of protein sequences. Bioinformatics 18: 1546-1547. doi:10.1093/bioinformatics/18.11.1546
40. Huson D, Richter D, Rausch C, Dezulian T, Franz M, et al. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. BMC bioinformatics 8: 460. doi:10.1186/1471-2105-8-460
41. Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics (Oxford, England) 23: 127-8. doi:10.1093/bioinformatics/btl529
42. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921. doi:10.1038/35057062
43. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science (New York, N.Y.) 291: 1304-51. doi:10.1126/science.1058040
44. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520-62. doi:10.1038/nature01262
45. Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, et al. (2008) Genome analysis of the platypus reveals unique signatures of evolution. Nature 453: 175-83. doi:10.1038/nature06936
46. Wallis JW, Aerts J, Groenen MAM, Crooijmans RPMA, Layman D, et al. (2004) A physical map of the chicken genome. Nature 432: 761-4. doi:10.1038/nature03030
47. Hellsten U, Harland RM, Gilchrist MJ, Hendrix D, Jurka J, et al. (2010) The Genome of the Western Clawed Frog Xenopus tropicalis. Science 328: 633-636. doi:10.1126/science.1183670
48. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, et al. (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297: 1301-1310. doi:10.1126/science.1072104 1072104 [pii]
49. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453: 1064-71. doi:10.1038/nature06967
50. Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, et al. (2002) The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. Science (New York, N.Y.) 298: 2157-67. doi:10.1126/science.1080049
51. Cameron RA, Samanta M, Yuan A, He D, Davidson E (2009) SpBase: the sea urchin genome database and web site. Nucleic Acids Res 37: D750-4. doi:gkn887 [pii] 10.1093/nar/gkn887
52. The C. elegans Sequencing Consortium (1998) Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science 282: 2012-2018. doi:10.1126/science.282.5396.2012
53. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. (2002) The genome sequence of the malaria mosquito Anopheles gambiae. Science (New York, N.Y.) 298: 129-49. doi:10.1126/science.1076181
54. Adams MD (2000) The Genome Sequence of Drosophila melanogaster. Science 287: 2185-2195. doi:10.1126/science.287.5461.2185
55. These sequence data were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ in collaboration with the user community. (n.d.).
56. Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, et al. (2007) Science (New York, N.Y.) 317: 86-94. doi:10.1126/

science.1139158

57. Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, et al. (2008) Nature 454: 955-960. doi:nature07191 [pii] 10.1038/nature07191
58. King N, Westbrook MJ, Young SL, Kuo A, Abedin M, et al. (2008) Nature 451: 783-788. doi:nature06617 [pii] 10.1038/nature06617
59. Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, et al. (2001) Nature 414: 450-453. doi:10.1038/35106579 35106579 [pii]
60. Batrachochytrium dendrobatidis Sequencing Project. Broad Institute of Harvard and MIT (http://www.broad.mit.edu) (n.d.).
61. Rhizopus oryzae Sequencing Project. Broad Institute of Harvard and MIT (http://www.broad.mit.edu) (n.d.).
62. Martin F, Aerts A, Ahren D, Brun A, Danchin EG, et al. (2008) Nature 452: 88-92. doi:nature06556 [pii] 10.1038/nature06556
63. Martinez D, Larrondo LF, Putnam N, Gelpke MD, Huang K, et al. (2004) Nat Biotechnol 22: 695-700. doi:10.1038/nbt967 nbt967 [pii]
64. C. neoformans Genome Project, Stanford Genome Technology Center, funded by the NIAID/NIH under cooperative agreement AI47087, and The Institute for Genomic Research, funded by the NIAID/NIH under cooperative agreement U01 AI48594. (n.d.).
65. Kämper J, Kahmann R, Bölker M, Ma L-J, Brefort T, et al. (2006) Nature 444: 97-101. doi:10.1038/nature05248
66. Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, et al. (2002) Nature 415: 871-880. doi:10.1038/nature724 nature724 [pii]
67. Sherman DJ, Martin T, Nikolski M, Cayla C, Souciet JL, et al. (2009) Nucleic Acids Res 37: D550-4. doi:gkn859 [pii] 10.1093/nar/gkn859
68. Candida guilliermondii Sequencing Project. Broad Institute of Harvard and MIT (http://www.broad.mit.edu) (n.d.).
69. Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, et al. (2004) Science 304: 304-307. doi:10.1126/science.1095781 1095781 [pii]
70. Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, et al. (1997) Nature 387: 67-73.
71. Aspergillus Comparative Genome Project. Broad Institute of Harvard and MIT (http://www.broad.mit.edu) (n.d.).
72. Coccidioides immitis Sequencing Project. Broad Institute of Harvard and MIT (http://www.broad.mit.edu) (n.d.).
73. Cuomo CA, Guldener U, Xu JR, Trail F, Turgeon BG, et al. (2007) Science 317: 1400-1402. doi:317/5843/1400 [pii] 10.1126/science.1143708
74. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, et al. (2003) Nature 422: 859-868. doi:10.1038/nature01554 nature01554 [pii]
75. Sclerotinia sclerotiorum Sequencing Project. Broad Institute of Harvard and MIT (http://www.broad.mit.edu) (n.d.).
76. Fey P, Gaudet P, Curk T, Zupan B, Just EM, et al. (2009) Nucleic Acids Res 37: D515-9. doi:gkn844 [pii] 10.1093/nar/gkn844
77. Eichinger L, Pachebat JA, Glockner G, Rajandream MA, Sucgang R, et al. (2005) Nature 435: 43-57. doi:nature03481 [pii] 10.1038/nature03481
78. Nozaki H, Takano H, Misumi O, Terasawa K, Matsuzaki M, et al. (2007) BMC Biol 5: 28. doi:1741-7007-5-28 [pii] 10.1186/1741-7007-5-28
79. Palenik B, Grimwood J, Aerts A, Rouzé P, Salamov A, et al. (2007) Proceedings of the National Academy of Sciences of the United States of America 104: 7705-10. doi:10.1073/pnas.0611046104
80. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, et al. (2007) Science 318: 245-250. doi:318/5848/245 [pii] 10.1126/science.1143609
81. Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, et al. (2010) Science 329: 223-226. doi:10.1126/science.1188800
82. Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, et al. (2008) Science (New York, N.Y.) 319: 64-9. doi:10.1126/science.1150646
83. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, et al. (2007) Nucleic Acids Res 35: D883-7. doi:gkl976 [pii] 10.1093/nar/gkl976
84. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, et al. (2008) Nucleic Acids Res 36: D1009-14. doi:gkm965 [pii] 10.1093/nar/gkm965
85. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, et al. (2006) Science (New York, N.Y.) 313: 1596-604. doi:10.1126/science.1128691
86. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, et al. (2008) Nature 456: 239-244. doi:nature07410 [pii] 10.1038/nature07410
87. Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, et al. (2004) Science 306: 79-86. doi:10.1126/science.1101156 306/5693/79 [pii]

88. Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RHY, et al. (2006) Science 313: 1261-1266. doi:10.1126/science.1128796

89. Haas BJ, Kamoun S, Zody MC, Jiang RHY, Handsaker RE, et al. (2009) Nature 461: 393-8. doi:10.1038/nature08358

90. Aury JM, Jaillon O, Duret L, Noel B, Jubin C, et al. (2006) Nature 444: 171-178. doi:nature05230 [pii] 10.1038/nature05230

91. Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, et al. (2006) PLoS Biology 4: e286.

92. Heiges M, Wang H, Robinson E, Aurrecoechea C, Gao X, et al. (2006) Nucleic Acids Res 34: D419-22. doi:34/suppl_1/D419 [pii] 10.1093/nar/gkj078

93. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, et al. (2009) Nucleic acids research 37: D539-43. doi:10.1093/nar/gkn814

94. Gardner MJ, Bishop R, Shah T, Villiers EP de, Carlton JM, et al. (2005) Science 309: 134-137. doi:309/5731/134 [pii] 10.1126/science.1110439

95. Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, et al. (2007) Science 315: 207-212. doi:315/5809/207 [pii] 10.1126/science.1132894

96. Aurrecoechea C, Brestelli J, Brunk BP, Carlton JM, Dommer J, et al. (2009) Nucleic Acids Res 37: D526-30. doi:gkn631 [pii] 10.1093/nar/gkn631

97. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, et al. (2005) Science (New York, N.Y.) 309: 436-42. doi:10.1126/science.1112680

98. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, et al. (2005) Science 309: 416-422. doi:309/5733/416 [pii] 10.1126/science.1112642

99. Roger AJ, Hug LA (2006) Philosophical transactions of the Royal Society of London. Series B, Biological sciences 361: 1039-54. doi:10.1098/rstb.2006.1845

CHAPTER 6

General Discussion

6.1 Evolution of interaction networks

In this thesis we contributed to the knowledge on the evolution of protein networks and pathways in eukaryotes with a particular interest in Ras-like signaling pathways. In Chapter 2 we have been able to show that co-complex memberships between conserved interaction partners are highly conserved between human and yeast (>90%). This is in strong contrast to previous more pessimistic estimates made in literature [1,2]. The establishment of high conservation of PPIs allows for reliable transfer of interaction data between species based on orthology.

Reliable transfer of functional data on signaling pathways and networks is critical for comparative genomics analyses. Simply, because if this were not the case, many conclusions on the evolution of pathways and protein networks would be unreliable as there would be no certainty that orthologous proteins participate in the same pathway in other species without experimental validation beforehand.

The protein-protein interactions (PPIs) as studied in Chapter 2 only entail the interaction between protein complex components and are often stable interactions. However, many PPIs are of a more transient type, that is to say these interactions are often short lived because of weak interactions or there is a bind-and-release mechanism. These transient interactions are difficult to measure using the Tandem Affinity Purification – Mass Spectrometry (TAP-MS) method and are therefore underrepresented in the datasets we used to determine PPI conservation. It is therefore entirely possible that transient interactions are less conserved in evolution.

Other methods such as Yeast-2-Hybrid (Y2H) are more suitable to determine transient interactions as the interaction partners are not put under considerable stress and force during the experimental procedures, like in TAP-MS. However, there are no Y2H data sets available that support nearly as large an overlap of the entire yeast proteome as the

TAP-MS data sets by Krogan et al. [3] and Gavin et al. [4]. Nevertheless, when we apply rigorous filtering onto the Uetz et al. Y2H data set [5] to remove noise, comparable high conservation percentages (84%) are obtained as found for the TAP-MS data sets. The results described in Chapter 2 therefore indicate that gene loss is the most predominant process by which interaction networks change. This implies that orthology can be used to confidently describe differences and similarities between pathways and interaction networks between species pending experimental verification.

The interactions between Ras-like GTPases and their effectors are of the transient type, i.e. short lived interactions which are regulated in space and time [6]. This type is qualitatively different to the PPIs we studied in Chapter2. That doesn't mean that interactions with Ras GTPases are weak. The interaction between Ras GTPases and their Guanine Exchange Factors (GEF) are quite strong [7], but only for specific steps in the Guanine Exchange reaction. This reaction is fast and therefore the interaction is short lived, making the interaction transient.

If we look to the protein-protein interactions between the Ras GTPases and their regulatory proteins we observe that the specificity of the interactions can be very flexible, i.e. individual GAPs and GEFs can change specificity over time (e.g. the RasGRPs in Chapter 3). Specific interactions therefore seem relatively short lived in evolution for the Ras GTPases. Interestingly, when we zoom out from individual proteins to the protein domains, we actually observe that the interactions between Ras-like GTPases and their regulatory protein domains are extremely well conserved.

6.2 Evolution of domain-architectures and protein families; how typically animal are the animal type Ras-like signaling pathways?

We analyzed protein families that share common protein-domains, but have radically different domain-architectures (e.g. the RasGEF, RapGAP and RasGAP protein families) in Chapters 3 and 5. Therefore the terms orthology and homology do not need to apply for the full length genes. Instead many separate homology and orthology relationships can exist for these genes, i.e. different sets of homologs and orthologs can be defined for each individual protein domain. For example, SOS and RasGRF have similar domain architectures (X-RhoGEF-PH-REM-CDC25HD). However individual phylogenetic analysis of the RasGEF, PH and RhoGEF domains (not described in this thesis) indicated that each domain has been acquired separately by the SOS and RasGRF genes.

Because of so called domain promiscuity observed within the RasGEF, RasGAP and RapGAP protein families [8] evolutionary reconstruction of the Ras-like signaling transduction pathways becomes complicated. If we want to transfer cellular function, we cannot go further back than the last ancestor for which we can reconstruct a specific domain architecture. We have been able to push the point of invention of a specific domain architecture for most RasGEFs, RapGAPs and RasGAPs marginally further than the animal ancestor as *Monosiga brevicollis*, a unicellular choanoflagellate, contains many animal-specific domain architectures for RasGEFs, RasGAPs and GAPs. In the case of RalGPS we have been able to identify an ortholog with previously animal specific domain architectures in the chytrid fungus *Batrachichodon dendrobatidis*, pushing back the point of origin for RalGPS even further.

The identification of bona fide orthologs of typical animal RasGEFs, RapGAP, and RasGAPs in chytrid and zygomycotal fungi (although not all have the animal domain architecture) indicates that at least parts of the typical animal Ras-like GTPase signaling pathway components are not specific for animals, but instead were invented before the animal-fungal split and have subsequently been lost in (higher) fungi. The phylogenetic analysis suggests that these genes did not arise from horizontal gene transfer. This indicates that fungi, and not animals, completely rewired their Ras-like signaling pathways. Our observations raise many interesting questions: What is the function of these animal type Ras pathways in the basal fungi? Can we differentiate between multi-cellular and cellular function of Ras-like pathways by studying Ras signaling components in these organisms? To which extend does the animal type Ras-pathway predate animal-fungal split? Why did fungi refurbished their Ras-like signaling pathways? Therefore, it would be very interesting to search for additional animal-specific Ras/Rap pathway components in the chytrid and zygomycotal fungi and map in detail which parts predate animal-fungal split.

We are aware that our selection of genomes shows a bias towards animal and fungal genomes, this due to a couple of reasons. The first reason is that Ras signaling has mainly or exclusively been described in animals and fungi. Focusing on these species increases resolution in these phyla and ultimately increases the usefulness of our analyses to other researchers. Secondly, the main focus for sequencing project lies with animal and fungal species, therefore there are more fungal and animal genomes available. Many interesting protists and early divergent species of eukaryotes have not been sequenced yet. The genomes of early divergent species that will become available in the future should be very interesting as they can provide many new insights, like the excavate protist *Naegleria gruberi* [9]. The *N. gruberi* genome has resulted in the identification of many pan-eukaryotic protein families that were thought to be specific for only a few phyla. It is also in *N. gruberi* that we identify extended Ras signaling pathways which were previously observed only in vertebrates.

6.3 Evolution of an entire pathway; easy integration of new inputs via GTPases

We investigated the evolution of Rheb GTPase and the TOR signaling pathway in Chapter 4. The reason we picked the Rheb-TOR pathway and not for instance the Rap signaling pathway is three fold. One, we attempted to find a Rheb specific Guanine Exchange Factor by phylogenetic co-occurrence and we therefore wanted to profile the entire pathway in order to get insight into which phylogenetic patterns we should search for. Two, only a single GAP protein has been described for Rheb and no GEF. This suggested a simpler pathway topology around the Rheb GTPase compared to the multi-GEF, multi-GAP regulation of the Ras and Rap GTPases (see Rap1 pathway of Bos et al. [10]). Three, the TOR pathway is well described in multiple organisms and critical components of the pathway (e.g. the TOR complexes) have been reported to be highly conserved in eukaryotic evolution. This allowed us to investigate the GTPase in a distinct environment that would be more easily tractable in evolution compared to the Ras and Rap pathways for which much is still unknown.

We found that the Rheb GTPase and its GAP, TSC2, are strongly linked in evolution. We showed that the many signaling inputs to Rheb are integrated via TSC2. This is in strong

contrast to Ras and Rap signaling wherein many inputs are relayed to the GTPase via multiple GEFs. Therefore a qualitative difference between Rheb signaling and Ras/Rap signaling is the level on which the input is integrated. The input to Rheb is integrated by the TSC2 while the input for Ras and Rap GTPases is integrated by the GTPases themselves. Our analysis showed that many new signaling inputs can be acquired and integrated with only minor changes (for instance by the addition of TSC1) to pathway topology. I feel that this is the intrinsic strength of the small GTPase signaling in eukaryotic evolution.

6.4 Evolution of Rap1 effectors

We see strong co-evolution of GTPases and GEFs/GAPs and we could use this info to learn more than from GTPase phylogeny alone. Similarly looking at whole TOR pathway gave us insights into Rheb GTPase signaling. We suspect looking at effectors of Ras-like GTPases will also yield more knowledge on how Ras pathways are wired. We would like to investigate specifically the Rap1 pathway. As previously discussed, basal fungi contain orthologs of previously thought animal specific regulators. It would be very interesting to search for additional animal-specific Ras pathway components in the chytrid and zygomycotal fungi as well.

6.5 Vertebrate Ras signaling

We find that many genes have duplicated in the vertebrate ancestor (e.g. Rap1A/B, RalA/B, RasGRP1/2/3, etc). It is well established that multiple events of whole genome duplications (WGD) occurred early in vertebrate evolution [11] which most likely accounts for the duplications of the Ras-like GTPases and their regulators. However it does not necessarily account for the fact that most duplicates have been maintained. Also it is unclear why some genes have retained all four duplicates versus only two. One reason could be that tissue specific needs for Ras signaling components in vertebrates could not be solved by adding more regulatory elements to the same locus. Interestingly, in plants signaling pathway components were retained more than average after a WGD [12-14]. This suggests that there might be a general evolutionary advantage to completely duplicate a signaling pathway via WGD. The emergence of vertebrates coincides with one of the largest expansions of Ras-signaling components in eukaryote evolution. Understanding why this expansion occurred can greatly increase our knowledge on how signal transduction pathways evolve in general.

6.6 Evolution of small GTPases; GAPs and GEFs tell their stories.

In order to proceed to a full evolutionary reconstruction of Ras-like signaling we proceeded with reconstructing the evolution of the entire Ras-like GTPase family. We find that the Ras-like GTPase phylogeny is difficult to interpret. Inconsistent clustering and low bootstrap support, especially in the deep branches, make it hard to draw definite conclusions on a single phylogenetic tree alone. To make sense of the Ras-like GTPase phylogeny we needed additional information.

In this thesis we made use of our observation of multiple cases of tight co-evolution of the Ras-like GTPases and their Guanine Exchange Factors and GTPase Activating proteins. The evolutionary patterns on the level of protein families seem less complex. We find tight

co-evolution of the CDC25HD with Rap, Ras and Ral GTPases (Chapter 3), the RapGAP domain with Rheb, Ral and Rap (Chapter 4 and 5) and also the RasGAP domain with Ras and Rap (Chapter 5). We used phylogenetic data from the Ras-like specific GEF and GAP protein domains to refine the Ras-like GTPase phylogeny itself as the evolutionary interpretation should fit the results of the phylogenetic analysis of all co-evolving protein families. By studying the expansion of the individual GAP and GEF protein families of the Ras-like GTPases we find a subtle preference for a type of regulation by Rap and Ras GTPases (i.e. negative regulation via GAPs or positive regulation via GEFs). Our results suggests that more weight lies on fluxes via GAP proteins in the regulation of Ras signaling compared to fluxes via GEF proteins in the regulation of Rap signaling.

On the level of protein families we find that Ras-like GTPases show a similar pattern compared to other small GTPase subfamilies like Arf (Laura Hamerslag University College Utrecht SCI 301 Research Thesis, " Evolutionary development of the family of ArfGAP proteins") and Rho (Fokkens & van Dam unpublished). Each of these subfamilies occurred in LECA together with their specific GEF or GAP domains. At first glance domain acquisition seem to play a more important role in signaling diversity for the Ras-like GTPase subfamily compared to other small GTPase subfamilies. For instance, we found in our lab that for the ARF GTPase family the domain architectures of the ARF-GAP protein family are highly conserved and less variable in evolution (Laura Hamerslag).

The addition of new domains to GTPase regulatory proteins increases signaling diversity and allows for a 'plug and play' approach in acquiring new regulatory inputs. One could therefore wonder why each small GTPase subfamily supports its own set of GEF and GAP protein domains instead of maintaining the primordial GEF and GAP domains with distinct domain architectures that allow them to specifically regulate the distinct GTPase subfamilies. One reason for the observed subfamily specific GEF/GAP domain sets could be the necessity to thoroughly separate regulation of the distinct small GTPase pathways and abolish any possibility of leaking cross-activation. This separation of small GTPase signaling by non-homologous replacement of GEF and GAP domains between small GTPase subfamilies mostly occurred before LECA. However we observed that this process of non-homologous replacements of GAP and GEF domains also occurred later in eukaryotic evolution, albeit at lower frequency. For instance, for Rho GTPases, two major GEF domains have been described: the canonical RhoGEF domain and the plant specific PRONE domain. In preliminary analysis into Rho GTPase evolution we observed a transition in non-homologous RhoGEF protein domains in plants from the canonical RhoGEF domain towards the plant specific PRONE RhoGEF domain (Fokkens & van Dam unpublished). In the red algae Cyanidioschizon merolae we find both the canonical RhoGEF and the plant specific PRONE. This indicates that the transition is smooth where the old domain temporarily co-exists with a new non-homologous domain but is slowly phased out and replaced.

6.7 References

1. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, et al. (2001) Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". Genome research 11: 2120-6. doi:10.1101/gr.205301

2. Suthram S, Sittler T, Ideker T (2005) The Plasmodium protein network diverges from those of other eukaryotes. Nature 438: 108-12. doi:10.1038/nature04135

3. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440: 637-43. doi:10.1038/nature04670

4. Gavin A-C, Aloy P, Grandi P, Krause R, Boesche M, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440: 631-6. doi:10.1038/nature04532

5. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. 403: 623-627.

6. Ponsioen B, Gloerich M, Ritsma L, Rehmann H, Bos JL, et al. (2009) Direct spatial control of Epac1 by cyclic AMP. Molecular and cellular biology 29: 2521-31. doi:10.1128/MCB.01630-08

7. Lenzen C, Cool RH, Prinz H, Kuhlmann J, Wittinghofer A (1998) Kinetic analysis by fluorescence of the interaction between Ras and the catalytic domain of the guanine nucleotide exchange factor Cdc25Mm. Biochemistry 37: 7420-30. doi:10.1021/bi972621j

8. Basu MK, Carmel L, Rogozin IB, Koonin EV (2008) Evolution of protein domain promiscuity in eukaryotes. Genome research 18: 449-61. doi:10.1101/gr.6943508

9. Fritz-Laylin LK, Prochnik SE, Ginger ML, Dacks JB, Carpenter ML, et al. (2010) The Genome of Naegleria gruberi Illuminates Early Eukaryotic Versatility. Cell 140: 631-642.

10. Kooistra MRH, Dubé N, Bos JL (2007) Rap1: a key regulator in cell-cell junction formation. Journal of cell science 120: 17-22. doi:10.1242/jcs.03306

11. Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS biology 3: e314. doi:10.1371/journal.pbio.0030314

12. Seoighe C, Gehring C (2004) Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome. Trends in genetics : TIG 20: 461-4. doi:10.1016/j.tig.2004.07.008

13. Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. The Plant cell 16: 1679-91. doi:10.1105/tpc.021410

14. Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, et al. (2005) Modeling gene and genome duplications in eukaryotes. Proceedings of the National Academy of Sciences of the United States of America 102: 5454-9. doi:10.1073/pnas.0501102102

A

Appendices

A.1 List of online supplementary material

Because of their nature not all supplementary material could be enclosed within this thesis. Additional online supplementary material can be found on http://bioinformatics.bio.uu.nl/john/thesis/. Additionally, the supplementary material for Chapters 2 and 3 can be found on the publishers' websites.

Chapter 3:

Figure_S3.1.pdf	Unrooted phylogenetic tree (PhyML) of the CDC25HD including domain compositions of the full length sequences.
Figure_S3.2.pdf	Unrooted phylogenetic tree (RAxML) of the CDC25HD including domain compositions of the full length sequences.
Figure_S3.3.pdf	Unrooted phylogenetic tree (Quicktree) of the CDC25HD including domain compositions of the full length sequences.
CDC25HD.hmm	HMM model (HMMER2) for the CDC25HD
CDC25_alignment.fa	Sequence alignment of the collected CDC25HDs
CDC25_phyml.phb	PhyML tree in newick format
CDC25_raxml.phb	RAxML tree in newick format
CDC25_quicktree.phb	Quicktree tree in newick format

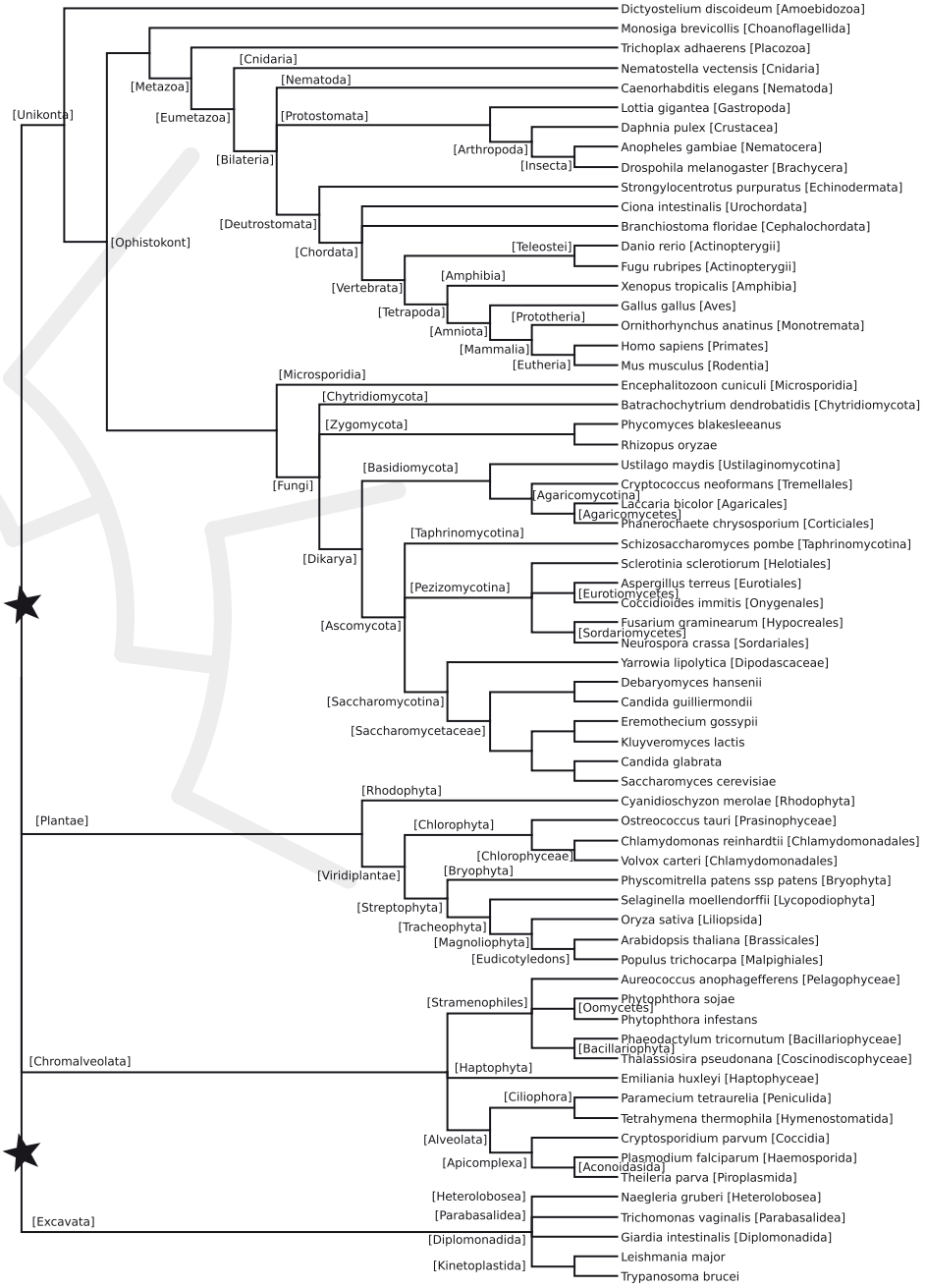
Chapter 4:

Figure_S4.1.pdf	mTOR kinase phylogeny
Figure_S4.2.pdf	Rheb phylogeny
Figure_S4.3.pdf	AGC kinase phylogeny

Chapter 5:

Figure_S5.1.pdf	Phylogenetic tree of the RapGAP domain and domain architecture of full length sequences
Figure_S5.2.pdf	Phylogenetic tree of the RasGAP domain and domain architecture of full length sequences
Supp_File_Ras_Phylogeny_PhyML_rooted.phb	PhyML tree in newick format
Supp_File_Ras_Phylogeny_RAxML_rooted.phb	RAxML tree in newick format

A.2 Eukaryotic tree of life



Eukaryotic tree of life as used throughout this thesis. The tree is based on NCBI Taxonomy and Simpson AG, Roger AJ (2004) The real 'kingdoms' of eukaryotes. Curr Biol 14: R693-696. The stars indicate possible roots for the tree.

Samenvatting in het Nederlands

Signaaltransductienetwerken zijn netwerken van inter-acterende eiwitten die cellulaire processen reguleren door interne en externe signalen te meten en integreren. Signaaltransductienetwerken zijn cruciaal voor de cel om zich aan te kunnen passen aan zijn omgeving. Hun rol in vele ziekten en syndromen en de ontwikkeling van een organisme maakt dat veel onderzoek zich centreert op signaalnetwerken. De wetenschap hoopt dat de biologische processen en onderliggende problemen van gerelateerde ziekten beter begrepen worden door het volledig uitpluizen van deze signaalnetwerken.

De biochemische en moleculaire biologie hebben veel kennis vergaard over hoe signaal netwerken en hebben daarmee de moleculaire basis van veel ziekten blootgelegd. Echter, de vraag hoe een signaal netwerk er uit ziet wordt vaak gevolgd door de vraag waarom het netwerk er zo uitziet. Waar de eerste vraag hoofdzakelijk beschrijvend is zal de tweede, wanneer die beantwoord wordt, een diepgaand inzicht geven in de werking en functie van het netwerk.

De complexe signaalnetwerken zijn niet in één keer ontstaan, maar zijn in de loop van de evolutie steeds complexer geworden door voortdurende aanpassingen en het toevoegen en verwijderen van onderdelen. In de loop van de evolutie wordt het netwerk verfijnd en aangepast naar veranderende omstandigheden (nieuwe signalen en nieuwe regulaties). Echter, een signaal netwerk moet vaak te allen tijde functioneel te blijven zodat de cel kan blijven leven. Daarom zullen niet alle aanpassingen even gemakkelijk te maken zijn.

De vraag waarom een netwerk op een bepaalde manier ìs gelegd kan dus voor een deel beantwoordt worden door uit te zoeken hoe het wàs gelegd. Het bestuderen van de evolutie van de individuele componenten van het netwerk en hun interacties zou ons in staat moeten stellen om de evolutie van het gehele netwerk in kaart te brengen. Dit geeft ons niet alleen inzicht in hoe een signaaltransductienetwerk ontstaan is, maar geeft ons ook een raamwerk waarop experimentele data afkomstig van verschillende modelorganismen onderling kunnen worden vergeleken en beoordeeld.

De Ras-signaalnetwerken zijn zeer complexe signaalnetwerken in eukaryote organismen. In de mens zijn defecten in deze netwerken vaak te vinden in kankercellen omdat ze kritieke processen reguleren zoals celdeling, celdifferentiatie, celgroei, cel-cel adhesie en geprogrammeerde celdood. De centrale eiwitten in deze netwerken, de Ras GTPases, gedragen zich als schakelaars. Ze kunnen aan, of uit naar gelang ze Guanosinetrifosfaat (GTP) of Guanosinedifosfaat (GDP) gebonden hebben. Wanneer de Ras GTPases ‘aan’ staan worden signalen doorgegeven naar het volgende niveau in de signaal cascade (een signaal wordt vaak op verschillende momenten ge-evalueerd en doorgegeven in een signaalnetwerk waarbij Ras GTPases functioneren als hoofdschakelaars). Het aanzetten gebeurt door het gebonden GDP te vervangen met GTP. Guanine Exchange Factoren (GEFs) zijn hiervoor verantwoordelijk en hun activiteit wordt in sterke mate gereguleerd. Voor het uitzetten zijn de GTPase Activerende Eiwitten (GAPs) verantwoordelijk, ze helpen de GTPases zelf de GTP om te zetten in GDP.

De bedrieglijke eenvoud van dit systeem maakt Ras-signaalnetwerken zo complex omdat er meerdere GEFs en GAPs zijn die elk op een andere manier gereguleerd worden. Ook

is het zo dat de effecten van Ras activatie anders zijn, naar mate de omstandigheden van activatie anders zijn (bijvoorbeeld activatie door de ene GEF lijd tot de doorgifte van een ander signaal dan wanneer Ras door een ander GEF wordt geactiveerd). De regulatie in tijd en ruimte lijkt van grote invloed te zijn op het doorgegeven signaal. Door de evolutie van deze complexe schakelaars te bestuderen hopen we meer inzicht te krijgen op het functioneren van de Ras-signaalnetwerken.

In dit proefschrift beschrijven we een bioinformatische en fylogenetische aanpak om de evolutie te bestuderen van interactienetwerken en de complexe Ras-signaalnetwerken. Als eerste bestuderen we de connectie tussen de conservering van genen en de conservering van interacties in evolutie. In de daarop volgende hoofdstukken bekijken we specifieke eiwitfamilies die een belangrijke rol spelen in de Ras-signaalnetwerken en maken een reconstructie van een volledig signaaltransductienetwerk.

Het begrijpen van de evolutie van interactienetwerken en de interacties is belangrijk voor het ontrafelen van de evolutie van signaalnetwerken door comparative genomics. Lage schattingen voor de conservatie van eiwit-eiwit interacties bemoeilijkten de reconstructie van netwerken in andere organismen, maar het was nog niet met zekerheid bepaald. Er zijn veel organismen waarvan geen of weinig experimentele data, maar wel het genoom beschikbaar is. We kunnen in deze soorten wel dezelfde componenten identificeren die bijvoorbeeld in de menselijke signaalnetwerken functioneren, maar we wisten niet of ze ook daadwerkelijk nog in hetzelfde signaalnetwerk functioneerden.

Om kennis over eiwit-eiwit interacties en netwerken toe te kunnen passen in andere organismen moeten we eerst bepalen hoe sterk of slecht de interacties tussen eiwitten zijn geconserveerd. In 2006 zijn twee grote interactiedatasets gepubliceerd voor bakkersgist. Nu hadden we de gegevens in handen om de interactienetwerken tussen mens en bakkersgist te kunnen vergelijken (Hoofdstuk 2) en te bepalen hoe goed eiwit-eiwit interacties zijn geconserveerd in de evolutie.

We vergeleken eiwitcomplexen van mens met de experimentele data uit bakkersgist met als doel de evolutie te bestuderen van interacties binnen eiwitcomplexen. Een eiwitcomplex is een structuur van twee of meer eiwitten die vaak gezamenlijk één of meerdere functies uitvoeren. We beperkten ons tot eiwitcomplexen omdat de experimentele methoden die gebruikt zijn voor de data uit bakkersgist hoofdzakelijk stabiele vormen van interacties meten, zoals in eiwitcomplexen. We vonden dat van de 5960 eiwitparen binnen hetzelfde complex in mens, 2216 afwezig waren in gist omdat beide eiwitten niet aanwezig waren in gist. Van nog eens 1828 eiwitparen was een mogelijke interactie verstoord in gist doordat één van het eiwitpaar afwezig was. Van de overgebleven 1916 geconserveerde eiwitparen bleek slechts 10% van hun onderlinge interacties niet gemeten in gist. Dit suggereert dat 90% van de interacties tussen eiwitparen uit hetzelfde eiwit complex geconserveerd is gebleven tussen gist en mens. We kunnen concluderen dat in eiwitcomplexen de onderlinge interacties hoofdzakelijk veranderen door dat een of beide eiwitten die een interactie hebben in mens niet aanwezig zijn in gist en dat onderlinge interacties tussen geconserveerde eiwitparen sterk geconserveerd zijn. We hebben kunnen aantonen dat er een sterke relatie is tussen evolutie op genoom en op netwerk niveau. Dit geeft ons meer vertrouwen dat als we componenten van een menselijk interactie- of signaalnetwerk identificeren in het genoom van een organisme dat deze ook tot hetzelfde signaalnetwerk behoren.

Als een eerste verkenning naar de evolutie van de Ras-signaalnetwerken onderzochten

we de evolutie van het RasGEF domein in hoofdstuk 3. Zoals eerder genoemd zijn de RasGEFs verantwoordelijk voor het activeren van de Ras GTPases. De RasGEF familie van eiwitten zijn ideaal voor het bestuderen van de evolutionaire eigenschappen van een Ras regulerende groep van eiwitten. RasGEFs combineren namelijk een grote verscheidenheid aan verschillende regulerende eiwitdomeinen (dit zijn duidelijk onderscheidbare functionele onderdelen van een eiwit) met een gemeenschappelijke en traceerbare evolutionaire afstamming via het RasGEF domein.

We hebben sequenties van RasGEF domeinen gevonden in dieren, schimmels en een grote verscheidenheid aan ééncelligen, maar niet in planten. Het feit dat we het RasGEF domein in zoveel soorten vinden suggereert dat het RasGEF domein waarschijnlijk al ontstaan is in of voor de voorouder van alle huidige eukaryote soorten (1 tot 2 miljard jaar geleden) en dat planten het dus later weer zijn verloren. We laten zien dat al ten minste zeven voorouderlijke RasGEFs aanwezig waren in de voorouder van de schimmels en de dieren. Dieren en schimmels hebben waarschijnlijk los van elkaar voorouderlijke RasGEFs verloren en nieuwe eiwitdomeinen opgepikt. Hierdoor zijn er verschillen in het RasGEF repertoire tussen de schimmels en dieren van nu. We vinden echter ook bewijs dat de Ras-signaalnetwerken in sommige schimmels meer overeenkomen met dieren dan dat eerder werd aangenomen. In primitieve schimmels vinden we zowel Ral als Ral specifieke RasGEFs (Ral is een type Ras GTPase) die normaal alleen in dieren gevonden worden. Dit betekend dat het Ral-signaalnetwerk veel ouder is dan eerst werd aangenomen. Het lijkt er dus op dat Ras-signaalnetwerken al vroeg (voor het ontstaan van de dieren) vrij complex waren, maar dat bepaalde groepen van organismen het hun unieke manier eigen hebben gemaakt.

In hoofdstuk 4 beschrijven we de evolutie van het TOR-signaalnetwerk, die het Rheb GTPase bevat. Het TOR kinase, die het signaalnetwerk zijn naam geeft, is een belangrijke regulator van groei in cellen van eukaryoten. Veel eiwitten in dit netwerk hebben een rol in het ontstaan van kanker en metabole ziekten in de mens. De verschillende eiwitten van het TOR-netwerk zijn sterk geconserveerd in de evolutie en er is al veel experimenteel onderzoek naar gedaan. Het TOR-signaalnetwerk is hierdoor uitermate geschikt om de evolutie van een Ras-signaalnetwerk in detail te bestuderen. We hebben van elk eiwit in het TOR-netwerk een fylogenetische analyse gemaakt en hebben bepaald wanneer ze zijn ontstaan in de evolutie. We zien dat de twee TOR complexen (deze twee complexen bevatten het TOR kinase) en een groot deel van de rest van het signaalnetwerk al bestonden voor de voorouder van alle huidige eukaryote soorten. Deze onderdelen vormen de kern van het netwerk waaraan nieuwe onderdelen zijn toegevoegd tijdens de evolutie van de dieren. We laten ook zien hoe genduplicaties die hebben geleid tot de S6K, RSK, SGK en PKB kinasen hebben bijgedragen voor de complexheid van het TOR-signaalnetwerk. De evolutie van het TOR-signaalnetwerk laat zien hoe een belangrijk signaalnetwerk zowel sterk geconserveerd kan zijn als flexibel.

Als laatste, in hoofdstuk 5, onderzoeken we de complexe evolutie van de Ras GTPases zelf. We combineren de gedetailleerde evolutionaire reconstructies van de RasGEFs (hoofdstuk 3), RapGAPs en RasGAPs (hoofdstuk 5), met de onduidelijke stamboom van de Ras-achtige GTPases om deze vervolgens te kunnen begrijpen en verduidelijken. De RasGAP en RapGAP eiwitten hebben een belangrijke functie door Ras GTPases uit te kunnen schakelen. Net als de RasGEFs waren er al enkele RasGAPs en RapGAPs in de voorouder van alle huidige eukaryote soorten. Alle Ras-achtige GTPases zijn ooit

ontstaan uit een enkel gen dat in de evolutie meerdere keren is gedupliceerd wat de Ras-signaalnetwerken zeer complex maken. We reconstrueren de volgorde van differentiatie van de verschillende Ras-achtige GTPases en zien in de patronen van de evolutie dat de Ras type en de Rap type Ras-achtige GTPases verschillend worden gereguleerd, n.l. Rap via GEFs (dus positieve regulatie) en Ras via de GAPs (negatieve regulatie).

We hebben nu veel meer inzicht in de evolutie van de complexe Ras-signaalnetwerken. Echter, er blijven veel interessante vragen over om te onderzoeken. Zo weten we nog niet hoe goed andere typen interacties, anders dan de stabiele, geconserveerd zijn. Voor de Ras-signaalnetwerken hebben we interessante overeenkomsten tussen primitieve schimmels en dieren gevonden. Het zal interessant zijn om ook andere onderdelen van de dierlijke Ras-signaalnetwerken te zoeken in schimmels. Dit zal helpen om meer algemene cellulaire functies van deze Ras-signaalnetwerken te onderscheiden van dierspecifieke cellulaire functies. In de gewervelde dieren zijn veel onderdelen van de Ras-signaalnetwerken één of meerdere keren gedupliceerd. Het uitzoeken naar de reden hiervan zal ons naast meer inzicht in de menselijke Ras-signaalnetwerken, ook meer leren over de evolutie van signaalnetwerken in het algemeen.

Dit proefschrift beschrijft in detail de evolutie van één van de meest intrigerende en complexe signaaltransductienetwerken in de eukaryote organismen, maar geeft ook specifieke handvaten aan moleculaire biologen voor het uitvoeren van nieuwe experimenten en het projecteren en vergelijken van resultaten die vergaard zijn in meerdere organismen.

Curriculum Vitae

Teunis Johannes Pieter (John) van Dam was born on June 11th, 1981, in Gouda, The Netherlands. From 1993 to 1999 he attended St. Antonius College, Gouda, The Netherlands, where he obtained his VWO diploma (A-levels). In 1999 John started his study in biology at Leiden University with a main focus on molecular biology with internships in the Fungal Genetics group and the Molecular Biochemistry department. During the last two years of his study in Leiden, John became increasingly more interested in bioinformatics and computational modeling. While finishing his Masters in biology in Leiden, John started a second Masters at Theoretical Biology and Bioinformatics at Utrecht University in 2004. In 2005 John received his Masters degree in biology from Leiden University and in 2007 he received his Masters degree in Theoretical Biology and Bioinformatics from Utrecht University.

In the final stages of his Theoretical Biology Masters he started his PhD research in Utrecht in 2006 which was a cooperation between Prof.dr. Johannes L. Bos of the Molecular Cancer Research Center, University Medical Center Utrecht and dr. Berend Snel of the Theoretical Biology group, Science Faculty, Utrecht University. As of the 1st of October 2010 John is doing post-doctoral work at the Centre for Molecular and Biomolecular Informatics, St. Radboud University Medical Center, Nijmegen, The Netherlands.

List of publications

van Dam TJP, Snel B (2008) **Protein complex evolution does not involve extensive network rewiring.** PLoS computational biology 4: e1000132.

van Dam TJP, Rehmann H, Bos JL, Snel B (2009) **Phylogeny of the CDC25 homology domain reveals rapid differentiation of Ras pathways between early animals and fungi.** Cellular signalling 21: 1579-85.

van Dam TJP, Bos JL, Snel B (2011) **Evolution of the Ras-like small GTPases and their regulators.** Accepted for publication at Small GTPases, February 9, 2011.

Dankwoord

Berend, je ondersteuning, enthousiasme, kennis, begrip, voorbeeld en je twijfel heeft me tot de onderzoeker gemaakt die ik nu ben en heeft me laten groeien. Heel erg bedankt voor al die jaren dat ik met je hebt mogen werken en dat ik gebruik heb mogen maken van je kennis en input. Ik denk dat ik niemand beter had kunnen treffen als begeleider.

Hans, bedankt voor de vrijheid en ruimte die je me gegeven hebt de afgelopen vier jaar. Het perspectief vanuit de moleculaire en celbiologische hoek is mij veel waard en je enthousiasme voor mijn onderzoek heeft me erg geholpen.

Holger and **Fried**, thank you for your biological inputs for chapters 3 and 4. You forced me to think more like a biologist, which was very useful indeed.

Like, bedankt voor je enthousiasme, kritische blik, je groene vingers die onze kamer leefbaar maakte, je inzet voor TBB en je positive blik op de mensheid afgewisseld met licht kwaadaardige gedachten richting falende infrastructuur. **Michael**, I will miss your humor, our early morning talks and the hourly walk to the coffee machine! You are personally responsible for a 50% increase in my tea consumption. **Gabino**, I've learned a lot from you in the last years. You have a very balanced view on the world which I greatly admire. **Lidija**, I admire your directness that has corrected my line of thought more than once. Your text editing skills are impressive! Thanks! **Erik**, ik heb maar een maand gehad om je zoveel mogelijk kennis en kunde mee te geven voordat ik in Nijmegen begon. Ik hoop dat je aan het einde van je stage door de fylogenetische bomen nog steeds het bos kunt zien. Ik heb daar in ieder geval alle vertrouwen in en ik ben zeer benieuwd naar je resultaten! **Jos**, ik zie je nu weer in Nijmegen, maar ik heb je de laatste maanden in Utrecht wel gemist. Je kritische blik, humor en de meesterlijke telefoongesprekken met helpdesks zijn ongeëvenaard.

Paulien en **Rob**, samen zijn jullie grotendeels verantwoordelijk voor mijn theoretisch biologische opleiding waar ik erg van genoten heb. Door jullie inzichten bekijk ik veel biologische vraagstukken door de ogen van een modelleur en waardeer ik het dynamische aan biologische systemen.

Thanks to the whole **TBB Group** for the warm environment that you created in which anyone can thrive and grow on an academic level but as a person as well. Thank you all for your kindness, interest, enthusiasm and fun times. Ik wil graag **Henk-Jan**, **Hanneke** en **Marian** extra bedanken voor de thee op de vroege ochtend en De Grote Dalmuti tijdens de lunch.

I would also like to thank the whole **Bos-Burgeringen Group** at the UMCU. Thank you for making me feel welcome! I know that some of you have gained some interest in the evolution of molecular systems, for which I am glad. Thank you too for all the effort to understand my presentations, which sometimes must have sounded like gibberish. You proved a strong and necessary foothold for me in experimental biology, for which I am truly grateful.

Geert en **Saskia**, bedankt voor jullie interesse in comparative genomics. De directe koppeling van phylogenetische analyses naar experimenten aan BubR1 en jullie enthousiasme is voor mij heel speciaal. Bedankt voor het geduld in de afgelopen periode. Ik ben vast besloten om onze samenwerking tot een succes te maken.

NBIC, bedankt voor het timmeren aan een infrastructuur en het samensmeden van een community van bioinformatici in Nederland.

Joost, **Erik**, **Wiebe** en **Folco**, bedankt voor de gezellige avonden D&D, overgoten met een grote hoeveelheid nerderigheid en flauwe grappen. Ze zijn een belangrijke uitlaatklep geweest in de afgelopen jaren.

Zelfs met de beste collega's is familie een van de belangrijkste factoren om een promotie te volbrengen, of zelfs aan te beginnen. Daarom, **Oma** heel erg bedankt voor de elleboog in mijn zij (letterlijk) die ik nodig had vijf jaar geleden. **Pap en mam**, **i**k ben erg blij dat jullie mijn keuze voor de wetenschap al sinds mijn vierde hebben gevoed en gesteund. Bedankt ook **Marco** en **Anouk**, **Kik** en **Ben** en **Sandór**. **Ingrid**, je promoveert een week eerder en we hebben daarom samen kunnen genieten van alle voorpret/stress. Het is fijn om zo nu en dan te kunnen praten met een lotgenoot.

Mijn liefste **Krista**, zonder jou zou ik nu niet op 30 maart mijn proefschrift verdedigen, had ik zeker niet gestudeerd in Utrecht, en had ik ook zeker jaren langer over mijn studie in Leiden gedaan. Bedankt voor je liefde, je enthousiasme, je bedachtzaamheid, je spontaniteit, je raad en daad. Bedankt voor al je steun en je hulp bij moeilijke keuzes en de nodige schoppen onder mijn kont.