The Evolution of Function in the Rab Family of small GTPases

Yoan Diekmann

Dissertation presented to obtain the Ph.D. degree in Computational Biology

Instituto de Tecnologia Química e Biológica | Universidade Nova de Lisboa

Research work coordinated by:

FUNDAÇÃO CALOUSTE GULBENKIAN
Instituto Gulbenkian de Ciência

Oeiras,
April, 2014

INSTITUTO
DE TECNOLOGIA
QUÍMICA E BIOLÓGICA
/UNL

Knowledge Creation

Cover: "The Rab Universe". A circular version of Figure 2.5.

The Evolution of Function in the Rab Family of small GTPases

Yoan Diekmann, Computational Genomics Laboratory, Instituto Gulbenkian de Ciência

Declaração: Esta dissertação é o resultado do meu próprio desenvolvido entre Outubro de 2009 e Outubro de 2013 no laboratório do José B. Pereira-Leal, PhD, Instituto Gulbenkian de Ciêcia em Oeiras, Portugal, no âmbito do Programa de Doutoramento em Biologia Computacional (edição 2008-2009).

Apoio financeiro: Apoio financeiro da FCT e do FSE no âmbito do Quadro Comunitário de Apoio, bolsa de doutoramento SFRH/BD/33860/2009 e projectos HMSP-CT/SAU-ICT/0075/2009 e PCDC/EBB-BIO/119006/2010.

Acknowledgements

"Alors que vous achevez la lecture de mon exposé, il est nécessaire de souligner le fait que ce que j'en ai retiré est bien plus conséquent." Mit diesen Worten, die ich meinem Freund Axel schulde, begann was in der vorliegenden Arbeit seinen vorläufigen Höhepunkt findet. So wahr wie damals ist der Satz auch heute, und das verdanke ich einer Reihe von Menschen welche unerwähnt zu lassen unmöglich ist. Obwohl das Schreiben die folgenden Namen notwendigerweise in eine Reihenfolge zwängt, so gebührt doch allen ihr gänzlich eigener Dank. Ich lasse der Müdigkeit freien Lauf und schreibe einfach drauf los.

Als erstes möchte ich meinen supervisor Zé nennen. Danke für all die Möglichkeiten und Lektionen. Ohne eigene, kann man keine Meinung vertreten. An all meine Arbeitskollegen von CGL und BIU, danke für die Hilfe und die Arbeitsumgebung die wir zusammen geschaffen haben. An das IGC für das unvergleichliche 'Science Sommer Camp'-Gefühl, für die brillanten Vorträge und die Chancen ganz nah ranzukommen.

An Daniel, der den vielleicht größten aber sicherlich wichtigsten Teil meiner Ausbildung übernommen hat. Ohne die endlosen Gespräche, den ansteckenden Enthusiasmus und dein Vorbild wüsste ich vielleicht bis heute nicht, dass ich mit science richtig liege. An Jorge und das PDBC, mir die Bewerbungsphotowahl nicht zu sehr angekreidet und die Gelegen-

heit gegeben zu haben all das zu erleben. Ich kann mir keine bessere Art vorstellen zum Biologen zu mutieren.

An die Menschen denen ich zu danken bisher versäumt habe, ohne welche ich aber niemals hier gelandet wäre: Jens Stoye, Eric Tannier und Julien Allali. Danke für das Vertrauen und die Nachsicht.

Schlussendlich, danke an meine Freunde, in Portugal und Deutschland, ohne die ich nie die Kraft gehabt hätte auch nur das kleinste der Opfer zu stemmen die mein “Beruf” fordert. Und an meine Eltern, für die bedingungslose, andauernde Unterstützung, und mir das Einzige beigebracht zu haben, was wirklich zählt, nachzudenken.

Contents

Author contribution: I reviewed the literature and wrote the chapter.

Chapter 1

Evolution of Protein Function

"[...] a general physiology which can describe the essential characteristics of matter in the living state is an ideal [...] we can strive toward [...] by a study of the vital functions in all their aspects throughout the myriads of organisms." [1, p. 4]

—August Krogh, 1929

Abstract Chapter 1

The question how protein function evolves is a fundamental problem with profound implications for both functional end evolutionary studies on proteins. Here, we review some of the work that has addressed or contributed to this question. We identify and comment on three different levels relevant for the evolution of protein function. First, biochemistry. This is the focus of our discussion, as protein function itself commonly receives least attention in studies on protein evolution. We distinguish three basic ways in which protein function evolves: by altering interactions, by changing product outcome or reaction biochemistry in enzymes, or by modification of protein biophysical properties. Second, genetics. This is the level responsible to generate variation, which acts as the substrate of evolution. Third, evolution. Evolutionary forces constantly sieve the pool of random mutations. Of particular interest here are evolutionary models at gene family level, as these provide a framework to integrate functional aspects. We conclude with two major observations. First, functional evolution is extremely dynamic. Not only do we find frequent transitions between distinct functional classes, but most mutational paths realising them are also short. Second and related, functional evolution is more likely than expected. We find the map between sequence and function to be degenerate, *i.e.* various sequences are able to perform the same function, and the same sequence able to carry out different functions. Finally, this review points us to protein switches as a promising model system, and we introduce the Rab family of small GTPases as the object of study in the remainder of this thesis.

1.1 Introduction

PROTEINS are essential macromolecules involved in virtually all cellular and organismic processes of life. Various biological disciplines including biochemistry, cell biology and physiology study the functions and activities of proteins and how these relate to cellular and organismal traits.

Shortly after the first protein sequences became available in the early 1950s, comparing sequences (and later structures) became one of the means to interrogate protein function [2]. Given sequences from the same proteins that evolved in different species (*i.e.* orthologs), identical regions may be important for a function shared amongst these proteins, for instance an enzyme active site. Conversely, if the function differs between species and one seeks to explain these differences, the responsible regions may be expected in varying parts of the protein. Hence, evolution informs about protein function.

With the advent of molecular evolution, the comparative study of similarities and differences amongst proteins was given a formal evolutionary basis. Like species, related proteins descend from common ancestors. They retain similarities and accumulate differences as a consequence of shared history and in response to various evolutionary forces. The observable result of this process is sequence divergence, ultimately caused by mutations whose evolutionary fate is intimately connected to their functional consequences. Hence, function informs about protein evolution.

In conclusion, protein function and evolution are inextricably linked[1]. Arising from this, a fundamental question with profound implications for

[1]The stringent definition of biological function, the modern history view [3] or selected effect definition, is actually an evolutionary concept: "the functions of a trait or feature are all and only those effects of its presence for which it was under positive natural selection in the (recent) past and for which it is under [...] purifying selection now" [4].

comparative approaches in both functional and evolutionary studies is: how does protein function evolve?

Studying protein evolution is essential to understand life in its full diversity at the molecular level. In recent years, the massive accumulation of data has made a growing fraction of this diversity accessible to molecular investigation. At least in principle the full protein repertoires of organisms all over the tree of life are readily available. As a consequence, there are two major reasons why asking how protein function evolves is a timely question.

On one hand, this development generates an immediate necessity to study functional evolution: due to the ever-growing throughput of sequencing technologies, the bottleneck shifted from data gathering to its functional annotation [5]. The predominant strategy for functional annotation is transfer of existing annotations applying the "guilt by association" principle [6, 7]. However, this is becoming insufficient because by definition 'transfer' cannot predict functional novelty [8, p. 164f]. As a result the fraction of uncharacterised genes grows [9]. A deeper understanding of how function evolves has the potential to assist and guide functional genomic efforts.

On the other hand, the amount of freely accessible data presents an opportunity as it greatly facilitates comparative analysis at previously unimaginable scale. Together with other methodological innovations (see Subsection 1.1.3), this has lead to a renewed interest in integrative functional and evolutionary research on proteins [10, 11], with potential to unveil general principles where previously only case-studies had been possible.

Here, we review some of the work that asked or is relevant to the question how protein function evolves. As is true for any biological phe-

nomenon, a satisfactory answer has to incorporate many different levels. What is the genetic and genomic basis of mutations generating protein sequence variation? What are the functional consequences, *i.e.* biochemical, cellular and organismal, of these mutations? And finally, what are the evolutionary mechanisms sieving or favouring a functional variant? The next sections are devoted to these three aspects of the evolution of protein function, *i.e.* genetics, function, and evolution, although a strict separation of the different levels is not always possible. We begin with the biochemistry of evolving protein function, however, before we briefly define what we refer to as the function of a protein, when we consider it to have evolved, and summarise the approaches and tools to study this process. Finally, prior to the conclusion of this chapter, we introduce the Rab family of small GTPases analysed in the remainder of this thesis, and argue why Rabs are a promising model system to study the evolution of protein function.

1.1.1 What is the function of a protein?

Asking when and how protein function has changed throughout evolution first and foremost requires a clear definition of what is considered the function of a protein. Here, we simply consider every effect a protein has as its overall function[2], leaving many philosophical problems of defining biological function aside. However, what these functional effects are depends on the particular biological level that is considered [14, p. 50]. For example, the biochemical function of insulin is to bind the insulin receptor, at the cellular level one of its many effects is the translocation of glucose transporters to the plasma membrane, whereas an organismic function of insulin is the removal of excess glucose from the blood to prevent toxic effects. In the following, we predominantly focus on the lowest, biochem-

[2]More precisely, every effect that does not occur when the protein is experimentally inactivated, the so-called causal role definition of function [12, 13].

ical level of protein function which we simply refer to as 'the function' of a protein. Although any classification remains subjective to a certain degree, we follow reference [14] and distinguish four major functions of a protein: binding, catalysis, switching and as structural elements [15, p. 2]. These functions are not exclusive: the most fundamental one is binding, and is required for all the others [14, p. 50]. Moreover, molecular switches such as small GTPases depend on both binding and catalysis.

1.1.2 When does protein function evolve?

The definition and distinction of basic protein functions provides the means to delineate when protein function is different between two homologs and has therefore evolved. At the biochemical level, we wish to distinguish qualitative (which we further consider) from quantitative changes (which we do not further consider). For interactions, any change in binding specificity, *i.e.* the ability of a protein to bind to a ligand, is considered functional evolution, but not so adjustments in binding affinities defined as the strength of binding. Similar for catalysis, modulation of enzyme kinetics is considered a quantitative change and in principle of no matter here, unlike changes in product outcome or in the biochemistry of a reaction (Box 1.1 introduces EC numbers as a system to systematise biochemical reactions).

This relatively straight-forward distinction between qualitative and quantitative gets blurred as soon as functions at higher levels are considered. For example, a transcription factor (TF) may bind to a newly formed transcription factor binding site affecting the regulation of a cellular process. Although the TF itself has not altered its biochemical function, *i.e.* the binding to a specific DNA sequence motif, it may now nonetheless have a different function at the cellular or organismal level. The same reasoning holds for protein-protein and any other type of protein interaction. Another complication arises from evolutionary change for example

in protein biophysical properties: parameters like length, flexibility and stability are continuous, and can therefore not change qualitatively. Yet, alteration of any of these properties can have functional implications at the cellular and organismic level even without affecting the biochemical functions of proteins like binding and catalysis.

The question when protein function evolves can also be raised from an evolutionary standpoint: we exclude cases of adaptation of proteins without functional innovation, *i.e.* to maintain an existing function. For instance, the authors of reference [16] describe selection for protein stability in Myoglobins that does not affect protein function. Rather, stability is likely to minimise the higher burden inflicted by misfolded proteins resulting from increased expression levels. Other examples are proteins adapting to preserve their function in the face of extreme temperatures or pressures (reviewed in reference [17]). In contrast to adaptation, the related concept of exaptation by definition implies evolution of function [18]: exaptation in proteins designates the co-option of a protein to perform a function different from the one it originated for. It is most readily inferred when phylogenetic analysis suggests that the protein is older than the process it is involved in today, which has for example been found for prominent families such as p53 [19] or various families essential for multicellularity [20–22]. However, exaptation does not necessarily imply this new function to be at the biochemical level (see *e.g.* reference [23]).

Box 1.1: EC numbers

The Enzyme Commission (EC) number is a hierarchical classification scheme assigning numbers serving as identifiers for chemical reactions catalysed by enzymes. Hence, enzymes with the same EC number catalyse the same reaction, but not necessarily by the same structural mechanism, nor are the enzymes necessarily orthologous.

An EC number has four levels. The first level broadly assigns a reaction into one of the six categories shown in the table. The second and third level specify the (sub-)subclass, usually containing information about the type of compound or group involved and detailing the type of reaction. The fourth digits represent the substrate specificity or a simple serial number identifying individual enzymes in a sub-subclass [24].

EC	reaction	examples
EC1	Oxireductases	Dehydrogenase, Oxidase
EC2	Transferases	Transferase, Kinase
EC3	Hydrolases	Lipase, Amylase, Peptidase
EC4	Lysases	Decarboxylase
EC5	Isomerases	Isomerase, Mutase
EC6	Ligases	Synthetase

For example, small GTPases have the EC number 3.6.5.2, classifying them as Hydrolases (EC 3), acting on acid anhydrides (EC 3.6), specifically on GTP (EC 3.6.5), and lastly identifying them as small monomeric GTPases (EC 3.6.5.2).

1.1.3 Studying evolution of protein function

In the following, we briefly describe the recurrent conceptual, computational and experimental tools and techniques used to study the evolution of protein function.

Conceptually

Two related conceptual tools are particularly useful to understand the evolutionary trajectory between homologous proteins. The challenge is twofold: first, represent the steps of this trajectory at the three relevant levels, *i.e.* genotype, phenotype (here restricted to some readout of function) and fitness, and secondly, relate or map these steps across the three levels.

Representing the mutational steps at sequence level is most easily done in protein sequence space [25]. In its most common form, amino acids at positions differing between two proteins are treated as binary alternatives, and the hypercube representing all combinatorially possible intermediates becomes the protein sequence space (see *e.g.* Box 3 in reference [11]). Due to its high-dimensional nature, a meaningful graphical representation is only possible for up to four or five amino acid positions. Adding an additional dimension representing a phenotypic variable generates an explicit map between genotype and phenotype. How proteins 'move', *i.e.* evolve, through this space can for example reveal accessible mutational paths. These indicate constraints on the process of protein evolution for instance resulting from epistasis (see Subsection 1.2.1).

The second related concept is that of a fitness landscape, which we employ here to designate the map between phenotype and organismal fitness[3]. Much has been written as to whether the metaphor of a landscape is more harm- than useful [27], however, it remains in frequent use and allows to connect phenotypic variants and the evolutionary forces acting on an organism carrying them. Fitness differences is one of the crucial ingredients for evolutionary change by natural selection [28].

[3]This is not the original definition [26], neither the frequently encountered broader definition as genotype-fitness map. However, it is complementary to the genotype-phenotype map captured in the protein sequence space concept introduced above, and therefore a useful interpretation in the context of proteins.

Computationally

The computational analysis of protein evolution is based on what has been coined the typical 'phylogenetic pipeline' [29].

Starting from a protein sequence of interest, the fist step is to find homologs in other species. This is usually done based on pairwise structure or sequence comparisons, where similarity above a given threshold is interpreted as evidence for homology. The most fundamental step is then to multiply align the sequences, which is critical because all subsequent analyses directly or indirectly depends on it. Unlike pairwise sequence alignment, which is merely a measure of similarity, a multiple sequence alignment (MSA) represents an evolutionary hypothesis of homology between sites of a protein. MSAs can already be exploited for comparative analysis (see *e.g.* reference [30]), although failure to account for phylogenetic relationships amongst proteins is bad practice and may bias the results of statistical significance tests [31].

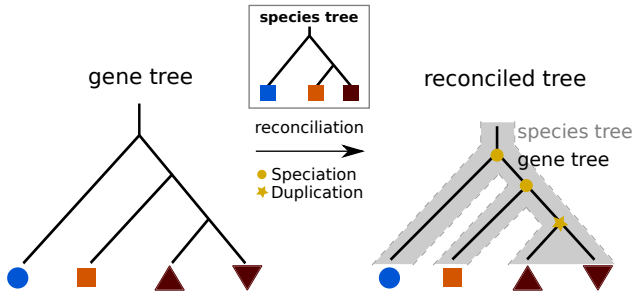
The next step is to derive a phylogenic tree often interpreted as a historical hypothesis of the relationship between the sequences. A proper phylogeny can for example serve to detect sites that convergently evolved and may therefore be functionally important (see *e.g.* reference [32]). If independent information on the species phylogeny is available, it can be used in an additional reconciliation step. Usually, reconciliation looks for the most probable tree that additionally minimises the number of gene duplication and loss events needed to inscribe the gene tree into the species tree (briefly explained in Box 1.2). Reconciliation is required for two reasons. First, reconciliation labels the branching points in a tree as either speciations or duplications, which is the only way to establish the nature of homology relationships between sequences, *i.e.* orthology or paralogy. If independent estimates of species divergence times are available, these can serve for instance to relate duplications to known ecological events [33]. Second, a common use of gene phylogenies is the reconstruction of

ancestral protein sequences. These do obviously only make sense for sequences at branching points that have a historical equivalent. Ancestral sequence reconstructions are an important tool to study the evolution of function, especially when followed by actual experimental resurrection discussed below.

Box 1.2: Gene tree/species tree reconciliation

A gene phylogeny relates a set of sequences by a tree. The tree is commonly inferred to maximise the likelihood of observing the sequences given a model of how proteins are thought to evolve. The resulting tree is then assumed to reflect the historical relationships between the genes.

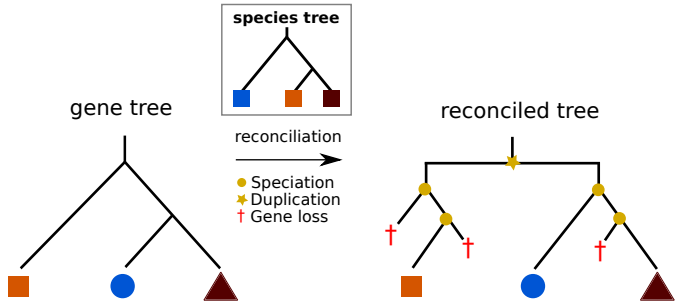
In order to get the homology relationships between the sequences, a reconciliation with the species tree is necessary, that labels the internal nodes in tree as either speciation or duplication (figures adapted from reference [34]). Orthologs and paralogs are all pairs of sequences whose least common ancestor is labeled as a speciation or duplication respectively. Reconciliation can be thought of as inscribing the gene tree into the species tree [35] (see figure below).



One of the oldest applications of gene trees is the inference of species phylogenies, assuming that the branching patterns of the genes and species are the same. This is not always the case: different events like gene duplications, gene loss, recombination and other more complicated phenomena like incomplete lineage

sorting can result in incongruence between gene and species tree. If independent information on the species tree is available, the criterion for reconciliation is usually to minimise the number of duplication and loss events that have to be assumed to reconcile the trees [36].

A different interpretation of reconciliation is not to accept the gene tree as given, but to consider the reconciliation distance (*i.e.* the number of duplication and loss events that have to be assumed) as part of the likelihood of the gene tree (see *e.g.* reference [37]). In the last figure for example, a better gene tree is obtained by inverting the two left leaves resulting in a tree matching the species tree. A possible scenario leading to this incongruence may be accelerated evolution in the orange sequence, provoking the well known long-branch attraction bias.



Another analysis requiring a gene tree is the mapping of sites that show statistical signs of positive selection. Sites inferred to have experienced positive selection are often indicative of functional evolution, and may therefore inspire a functional hypothesis.

Experimentally

In conjunction with computational methods, a mixture of old, repurposed and new experimental techniques forms the methodological basis for the evolutionary analysis of protein function. Two major classes of exper-

imental strategies exist, by analogy comparable to forward and reverse genetics.

Similar to reverse genetics, the first class starts with a protein of interest and manipulates its sequence. The difference is that the mutations are specific and most often introduced by site-directed mutagenesis [38], although comprehensive library-based approaches were recently developed [39, 40]. Therefore, this strategy allows to examine the functional importance and consequences of particular sites. A more sophisticated version of this approach is specifically designed to study mutational trajectories between natural protein homologs. In ancestral protein reconstruction, computationally inferred ancestral proteins are synthesised and functionally characterised [41]. The great advantage of this technique is that the effect of mutations is considered in the correct historical background, for example allowing to account for and study epistatic interactions.

The second type of experimental strategy takes a 'forward' approach: starting with a function of interest, directed evolution of proteins proceeds through repeated rounds of random mutations followed by filtering via selection for the function of interest [42]. The characterisation of intermediates then allows to dissect the process of functional evolution.

These conceptual, computational and experimental techniques and approaches described above have been used to interrogate how protein function evolves, reviewed in the next sections.

1.2 The biochemistry of evolving protein function

This section summarises some of the work that has elucidated the biochemical basis of how protein function evolves. The overwhelming amount of research on many diverse aspects of the topic precludes an exhaustive re-

view. Instead, we focus on some aspects important to start understanding the phenotypic or functional space of proteins, *i.e.* the conceptual space in which protein function evolves. In particular, we highlight some themes that are relevant for the rest of this thesis. In this section, we distinguish three alternative possibilities or modes to evolve protein function: changes in binding, in catalysis and in biophysical properties. Following the distinction of basic protein functions introduced in Subsection 1.1.1, these modes are relevant for all proteins, for all enzymes, and for structural and some non-enzymatic proteins respectively. The discussion is focussed around the biochemical function of proteins, however, higher-level functional evolution is briefly discussed in the context of biophysical properties. Lastly, aspects specific to protein switches are considered later in Section 1.5 when the Rab protein family of small GTPases is introduced as a promising model system to study the evolution of function.

1.2.1 Binding

Binding is the most fundamental function of proteins, all proteins bind other molecules [43, p. 52]. Binding partners or ligands can be other proteins, DNA, or molecules like carbohydrates, lipids, ions, or even nascent ice crystals. About 55% of gene products have no annotated enzymatic activity [44], and may therefore be expected to functionally evolve commonly by changing their set of ligands (with exception of changes in biophysical properties and switching). For enzymes, an alternative possibility is to alter the biochemistry of the reaction. However, a recent large scale analysis of over 2 million enzymes classified into 276 structurally defined superfamilies found ∼1500 changes in substrate specificity as measured by different EC numbers at the fourth level (see Box 1.1), in contrast to ∼1000 at higher EC classification levels [45]. Therefore, altering the set of ligands is most likely the predominant form of functional evolution in proteins.

How many mutations to alter interactions?

The question how many mutations are needed in a protein, *i.e.* few or many, to alter its specificity and in particular form an interaction has been addressed and reviewed mostly as part of the broader question about evolutionary step size (see *e.g.* [10]). Clearly, if mutations only have small effect, one expects that shifts in specificities require many mutations.

Two systems allowing to address this issue have been particularly well studied. One example is the engineered change of coenzyme specificity in decarboxylating dehydrogenases. Two members of this family, isopropylmalate dehydrogenase (IMDH), which exclusively uses nicotinamide adenine dinucleotide NAD^+ as coenzyme, and isocitrate dehydrogenase (IDH), which uses both NAD^+ and nicotinamide adenine dinucleotide phosphate ($NADP^+$), were altered by site-directed mutagenesis. In the first case, seven mutations in *E. coli* IDH caused a total switch from $NADP^+$ to NAD^+ and generated an enzyme kinetically comparable to natural IDHs using NAD^+ [46], although two replacements outside the active site are not related with specificity [33]. In the opposite direction, six replacements switched the *E. coli* IMDH coenzyme specificity to $NADP^+$ [47]. Hence, although IMDH and IDH are ancient paralogs and share only 25% sequence identity [46], as few as five or six targeted mutations are enough to switch binding specificity in both backgrounds.

A second gene family extensively studied in a series of papers are nuclear receptors for steroid hormones, which function as ligand-activated TFs. Vertebrates have two major clades of steroid receptors, one activated by estrogens and another clade containing several paralogous groups bound for example by glucocorticoids and mineralocorticoids. The two clades are derived by duplication from an ancestor that functioned as an estrogen receptor [48]. Using phylogenetic and structural analysis, the authors could identify and later confirm by ancestral protein reconstruction that two amino acid replacements cause the shift in specificity towards

nonaromatised steroids that include glucocorticoids and mineralocorticoids [49]. Notably, unlike the mutations in the decarboxylating dehydrogenases mentioned above, these mutations were not located in the ligand cavity, but in side chains albeit in contact with the ligand.

Hence, these two examples and others [50–56] suggest that few mutations, both engineered and naturally occurring, can be enough to alter binding specificities, both in proteins without catalytic activity and in enzymes. Moreover, these mutations are not necessarily restricted to the binding interface, although these are probably more common [57].

How to lose interactions

An alternative way protein function can evolve is by selective loss of interactions. It is clear that for instance in the engineered cases referenced above, the number of mutations needed to selectively lose an interaction is at most the number that was needed to gain it. Interactions may be disrupted with fewer mutations in case epistatic interactions require the co-occurrence of several residues. Most often interactions are critical aspects of protein function, and as such the interface residues are protected from accumulating mutations by purifying selection. This has for example been demonstrated in ubiquitin, where comprehensively quantifying the fitness effects of mutating each position into every possible residue has revealed that binding is the dominant cause behind purifying selection in ubiquitin [40]. Yet, there are different scenarios in which the loss of an interaction is the crucial step to evolve protein function.

On one hand, there is the loss of interactions previously shaped by selection, which may be lost for several reasons. For instance, after the duplication giving rise to the two clades of steroid receptors already introduced, the estrogen receptor in the *Branchiostoma floridae* lineage lost its ability to interact with ligands by virtue of two replacements, although each of them separately is enough to disrupt the interaction [58]. However,

the duplicate kept the ability to bind DNA. As a result, it competes for the binding sites with the original copy and in this way functions as a repressor. The new repressor function is thus a type of exaptation, *i.e.* the co-option for a new use of a protein previously shaped by selection [18]. An example not for exaptation but for adaptation involving loss of interactions in the evolution of two-component signalling pathways after duplication in α- and β-proteobacteria. These pathways typically consist of sensor histidine kinases which often very specifically phosphorylate a specific response regulator that in turn modulates gene expression. In α- and β-proteobacteria, the duplication of kinase-regulator pair led to crosstalk between the new response regulator NtrX and the conserved kinase PhoR, with negative effects on fitness in phosphate-limited conditions. The two lineages followed a different path to eliminate the cross-talk and insulate the signalling pathways: in α-proteobacteria two mutations in the kinase PhoR restored specificity to the co-evolving response regulator PhoB, whereas β-proteobacteria achieved specificity by four mutations in NtrX [59].

On the other hand, there is the loss of promiscuous, *i.e.* selectively neutral interactions that exist because of a lack of purifying selection. Specificity may often become important for function, large-scale studies for example found that residues surrounding the interfaces involved in transient interactions often contribute to prevent non-native interactions and maximise specificity [60]. A well studied example is the evolution of mineralocorticoid and the cortisol-specific glucocorticoid receptor from a mineralocorticoid-like, promiscuous ancestor. In the glucocorticoid lineage, two mutations largely recapitulate the loss of sensitivity to ligands other than the native ligand cortisol [61], although six additional mutations are needed to evolve the fully cortisol-specific enzyme [62]. It has been argued more generally that a principle of minimal specificity well describes the evolution steroid receptors: as long as promiscuity has no cost

and the receptor works sufficiently well, for instance because the cell will not encounter the promiscuously binding ligand, structural mechanisms leading to higher specificity do not evolve [63]. Note that promiscuity may not always be due to lack of purifying selection, but can reflect a biochemical constraint inherent to the structure and biochemistry of the protein. In this case, the interaction could not be lost without also losing the native function coupled to it, as for example observed in Rubisco that sometimes confuses its substrate CO_2 with its product O_2 [64].

In conclusion, the selective loss of interactions can be important for the evolution of protein function, and can much like gains of interaction be achieved with few mutations. The number of mutations needed is most likely independent of the type of interaction, *i.e.* of promiscuous origin or selected for.

Mutational paths to alter interactions

The question how many mutations are needed to evolve protein function can be elaborated. Commonly, the order of the mutations matters as a mutational path becomes mostly inaccessible if it does not monotonically increase fitness [25]. Therefore, rather than how many mutations, the question becomes how many mutational paths exist. Besides theoretical studies on abstract fitness landscapes models [65], the problem has begun to be addressed experimentally [66]. The pertinent questions are how many of the paths between two functional variants are accessible, what the mechanisms to block the others are, and whether there are processes biasing the choice of which path a protein is actually going to take. Before addressing these questions looking at three scenarios of how mutational paths can behave, we briefly comment on how mutational paths can be studied. It is clear that in naturally diverging proteins, the denser the phylogenetic sampling, *i.e.* the more species are analysed and the closer they are related, the better the order of mutations can be established.

However, in many cases proteins may substantially diverge in time periods between speciations, for instance after duplications. In these cases, hypotheses about the order of mutations can only be generated using experimental techniques like for example site-directed mutagenesis.

All mutational paths are accessible—When considering two functional proteins and the mutational path between them, the simplest way mutations can behave is having additive functional effects, *i.e.* effectively being independent. This has been shown for instance for the engineered switch in coenzyme use from NAD^+ to $NADP^+$ in *E. coli* IMDH [47]. Every mutation creates a functional, intermediate protein on the path between NAD^+ and $NADP^+$ specificity and consequently every mutational path is accessible, at least in principle. The reservation stems from the fact that absence of epistasis at the phenotypic level does not imply that there is none at the level of fitness. In the example above, the system follows a logic of diminishing returns [67], that is the same absolute increase in coenzyme specificity in an already specific enzyme contributes less to organismal fitness than it would in an inefficient enzyme [47]. Hence, the resulting composite genotype-fitness map is a concave function, in this case therefore not changing the accessibility of mutational paths.

Some mutational paths are blocked—The above situation becomes more complicated in the presence of epistasis, *i.e.* the dependence of mutational effects on other mutations. Two different structural mechanisms for epistasis have been described which are particularly important for functional evolution.

First, the structural repositioning of a residue by another mutation alters the mutational effect of the former, a situation which has been coined conformational epistasis [62]. An example is the loss of specificity for hormones other than cortisol in the evolution of the glucocorticoid re-

ceptor from a promiscuous ancestor already mentioned above. While one mutation in isolation has a deleterious effect by destabilising the binding interface, it repositions another residue into contact distance with the ligand. A mutation in this latter residue then allows it to form a new bond specifically with cortisol, which would have had no effect in its old localisation. Hence, in this case a deleterious and a neutral mutation together have a beneficial fitness effect [62], but the mutational path in which the deleterious mutation comes first is blocked or at least its accessibility reduced.

A second epistatic mechanism with equal consequences for the accessibility of mutational paths is the increase in protein stability. Stability often results from functionally neutral mutations, that later become essential to buffer the destabilising effects of mutations altering specificity. The glucocorticoid receptor again serves as an example: a neutral mutation that occurred millions of years before the actual switch in ligand specificity stabilises crucial parts of the protein and becomes necessary to tolerate any of the following mutations that actually modulate binding specificity [62]. This type of "new-function-stability tradeoff" [68] is a common phenomenon: a large analysis of 548 mutations observed in directed evolution experiments of 22 enzymes showed that mutations that affect enzyme specificity destabilise the protein more than other surface mutations on average [69]. At the same time, many functionally neutral mutations in these experiments had stabilising effects. One possible structural explanation for the destabilising effect of mutations altering specificity has been found when analysing the natural and *in vitro* evolution of TEM-1 β-lactamases that gained activity against cephalosporin antibiotics [70]. To accommodate a larger substrate, the binding site has to be enlarged at the expense of internal interactions that stabilise protein structure. As most proteins—including the β-lactamases—are only marginally stable, small effects on stability may have drastic effects on function and there-

with fitness. Thus, the consequence of epistasis is again that mutational paths to altered specificities are blocked unless compensatory mutations have occurred. This form of epistasis provides a good example for the connection between function and basic biophysical properties of a protein.

Finally, an interesting consequence of epistasis is that it can block the reverse mutational path that would undo functional changes in a protein (as for instance reported in reference [53]). Such an "epistatic ratchet" [71] has been described in the glucocorticoid receptor. Five mutations that mildly optimise cortisol specificity in the glucocorticoid receptor destabilise structures needed for the function of the ancestral protein, and as a result reversing the two mutations that are mainly responsible for the switch in specificity yields a non-functional protein. Hence, unless the five mutations are reversed, which have no effect on the ancestral function and would therefore have to be accumulated against purifying selection for cortisol specificity, the reverse path is inaccessible [71].

In conclusion, epistasis provides a mechanism to constrain accessible mutational paths (or depending on the perspective, opening them). How many paths are blocked depends on the strengths of purifying selection that determines the tolerance against mildly deleterious intermediates, and may vary from protein to protein: for some there is no restriction, as for example discussed for *E. coli* IMDH [47], in other cases most paths may be blocked, as has been shown experimentally for instance in TEM β-lactamases [50]. In general, epistasis is a common phenomenon [72, 73] and can therefore be expected to often influence the evolutionary path of proteins [74]. Note that regardless of epistasis, there are mechanisms that enable proteins to evolve via less fit or even non-functional intermediates: redundancy resulting from gene duplication for example has been introduced by Ohno as a solution to precisely this problem [75].

Some mutational paths are more likely than others—An entirely dif-

ferent perspective on mutational paths is not to focus on how many are accessible or blocked, but to ask if mechanisms exist that may bias the actual path taken in the functional evolution of a protein. Of major importance in this context are mostly neutral mutations that contribute to promiscuous enzyme functions different from the native function under purifying selection, but that do not affect the latter [76]. The ability of neutral mutations to affect promiscuous interactions has been demonstrated by experimental evolution for instance in cytochrome P450 [77]. These promiscuous binding activities may poise proteins for future functional evolution, as new specificities can emerge simply by amplifying already existing affinities to an alternative substrate via shorter and therefore more likely mutational paths in sequence space [78]. For example, protein resurrection showed that ancestral enzymes in the SABATH lineage of plant methyltransferases in Solanaceae have promiscuous activity on nicotinic acid (NA), a secondary activity still present in most extant members the family. After a gene duplication within *Nicotiana*, amplification of this promiscuous activity gave rise to a nicotinic acid carboxyl methyltransferase whose preferred substrate is NA [53]. In conclusion, promiscuity is a widespread phenomenon [79] particularly important for the evolution of function because it provides a set of functions that can immediately be tinkered with circumventing the need to evolve them from scratch.

Different mechanisms to alter interactions

So far, we exclusively discussed the effect of point mutations on binding specificities. However, diverse other mechanisms exist to alter protein interactions that can therefore play an important role in functional evolution.

Binding partners can be altered by loss and gain of longer strips of DNA, for example linear motifs, disordered regions or interaction domains. Short linear motifs, or SLiMs, are regions between three and ten amino

acids frequently found outside protein domains, which function as modules mediating weak and transient protein interactions [80]. Since they are short and unlike domains not constrained by the necessity to properly fold, they can be readily gained and lost and are therefore very dynamic throughout evolution. This is best illustrated by the frequent mimicry of SLiMs that convergently evolved for instance in viruses to hijack the host cellular machinery [81]. An important example for a protein that functionally evolved by altering its set of binding partners through gain and loss of SLiMs is the *Drosophila ftz* TF. *ftz* has homeotic functions conserved in most arthropods, mediated by a homeodomain binding DNA and a [FY]PWM motif for co-factor binding. The ancestor of beetles and Drosophila gained a LxxLL motif, allowing it to interact with nuclear hormone receptors, which resulted in *ftz* assuming an additional function in segmentation. However, in *Drosophila* the [FY]PWM motif was lost, therefore the ability to interact with the certain co-factors and as a result the ancestral homeotic function [82]. Interestingly, the importance of functional evolution of the TF itself through altered protein-protein interactions rather than evolution of *cis*-regulatory elements has recently been emphasised as a mode of evolution of gene regulation and developmental evolution in general [83, 84]. Another important mediator of interactions with very similar properties and consequences for functional evolution are intrinsically disordered (ID) regions, which are commonly defined as regions that do not adopt a regular three-dimensional structure. The large-scale analysis of protein-protein interaction networks has shown that interactions mediated by ID regions are more abundant than their ordered counterparts [85], which may be explained by their usually high binding promiscuity [86]. Hence, acquiring and altering ID regions may is an important evolutionary route for master regulators or hub proteins in general. Yet another possibility is the loss or gain of entire domains that mediate interactions. Interaction domains are independently folding modules usually between 35-150 residues

in length that can easily be inserted into loops or terminal regions [87, p. 88f]. Domains with varying specificities are known, some of the most frequent for example being WD40, the Armadillo repeat or SH3 that specifically binds proline-rich sequences [88]. An important protein family that has evolved by combination of domains bringing together distinct activities is Hedgehog. The hog domain, which in itself is already composed of an autoproteolytic and a cholesterol binding domain [89], fused with the receptor-binding Hedge domain at the base of Eumetazoa [90]. The composite is secreted into extracellular space and functions as a diffusible ligand in one of the fundamental signal transduction pathways in animal development.

A fundamentally different mechanism from the intrinsic determinants of binding specificities discussed so far, *i.e.* those found in the protein itself, are extrinsic factors, most importantly the cellular context. Many proteins are not localised in a diffuse manner all over the cell, but specifically targeted and localised to specific compartments or structures within the cell. A good illustration for the fact that localisation is a mechanism providing specificity are kinases. There are many more proteins targeted by kinases than kinases themselves, necessarily requiring them to phosphorylate several substrates. To nonetheless ensure specificity, kinases are targeted to the right localisation effectively preventing interaction with undesired targets [87, p. 90]. An example suggesting a role for relocalisation in the evolution of protein function is the hominoid-specific *CDC14Breto* gene. After duplication from a microtubule-localised phosphatase, mutations in the termini relocalised the protein to the Endoplasmic Reticulum. While sequence analysis suggest it maintained its phosphatase activity, it is very likely that the protein changed its substrates at its new localisation [91]. However, the cause-effect relationship between localisation and protein interactions goes both ways: the *Drosophila* gene *Umbrea* altered its localisation as a result of a mutations in a binding motif entirely changing

its interaction partners [55].

1.2.2 Catalysis

Virtually all chemical reactions in the cell are catalysed [15, p. 2], which speeds them up sometimes more than a billion fold and allows for spatiotemporal control [14, p. 63]. Both speed and control are essential for life to exist. Although some RNAs also have catalytic functions, most catalysts are proteins with 45% of all proteins having an annotated enzyme function [45]. As already discussed in Subsection 1.2.1, changing binding specificities is the most common mode of functional evolution in proteins, including enzymes. Yet, within enzyme superfamilies the evolution of product outcome and biochemistry are frequent and account for roughly two fifth of the recorded events of functional evolution in a survey of over 2 million enzymes [45]. Importantly, functional evolution is not constrained to transformations within enzymes or non-enzymes: loss of enzymatic activity is frequent [92] and discussed later, but also the reverse case, *i.e.* the evolution of an enzyme from a non-enzymatic ancestor, has been found to occur [93, 94].

How many mutations to alter biochemistry?

We begin by asking the same question as above for binding. Clearly, the fact that evolution of enzyme biochemistry is still relatively frequent compared to alterations of binding specificities suggests that the number of mutations needed may be in the same order of magnitude.

One possibility for enzyme function to evolve is to alter the enzyme product. An example of a single amino acid changing the product has been described in the kaurene synthase-like (KSL) gene family of rice. Whereas a particular member of the KSL family in the subspecies *indica* catalyses *ent*-copalyl diphosphate specifically to *ent*-isokaur-15-ene, the

ortholog in subspecies *japonica* produces *ent*-pimara-8(14),15-diene. Using site-directed mutagenesis, the authors could show that only one of three differences in the active site of the enzymes is sufficient to convert the product outcome between the two proteins [95]. That this is a common phenomenon at least in terpene cyclases has been demonstrated by generating and analysing all mutational intermediates between tobacco 5-epi-aristolochene and henbane premnaspirodieme synthase. These enzymes act on the same substrate but have distinct product outputs and can be reciprocally interconverted by mutations in nine amino acids [96]. While most mutations have moderate, additive effects on outcome profiles, punctuated changes as a result of a single amino acid were not rare [97]. Interestingly, much like substrate promiscuity discussed in Subsection 1.2.1 mutational paths proceed through catalitically promiscuous intermediates with equivalent conclusions for the evolution of new enzyme functions [98].

Another, more drastic possibility for the evolution of enzyme function is altering the catalytic activity. A remarkable example of two closely related but functionally distinct enzymes differing by only nine amino acids can be found in bacteria. Melamine deaminase from *Pseudomonas* sp. strain NRRl B-12227 and atrazine chlorohydrolase from *Pseudomonas* sp. strain ADP catalyse different reactions on different substrates, with no detectable activity on the respective other substrate [99]. A greedy strategy illuminating the most likely evolutionary paths between these two enzymes found evidence for many of the phenomena described above in the context of binding: epistatic interactions blocking certain paths, amongst others the direct reversal of the most likely mutational paths from one to tother emzyme, promiscuous intermediates, and a trade-off between enzymatic activity and stability [100].

In conclusion, these few examples suggest no qualitative differences between the evolution of binding specificities and enzymatic product outcome or reaction biochemistry, at least in terms of number of mutations

required. This agrees with previous findings that evolution of new chemistry can also be achieved by small local changes, that is not necessarily requires evolution of entirely new protein structures [45].

How to lose catalytic activity

Perhaps more surprising than in the case of binding, enzymes have also been found to functionally evolve by losing their enzymatic activity. There are several possibilities how this can happen. In a rather straightforward manner comparable to the case of the steroid receptor in the *Branchiostoma floridae* lineage described above, a multidomain protein may lose its enzymatic domain but keep other domains, for example binding domains. An example for such a case has been found in the family of mammalian polo-like kinases (Plks), master regulators of cell division that also have functions outside the cell cycle. A recently described member, Plk5, has lost its kinase activity by inactivating mutations, and the human sequence even lost the domain altogether due to the creation of a premature stop codon [101]. Yet, these proteins kept their substrate-binding domains and have assumed a new function in the brain, likely as anti-proliferative signals [102].

An interesting example, different in the sense that loss of enzymatic activity and gain of a new function happen within a single domain, has been found in the guanylate kinase (GK) enzyme which catalyses phosphotransfer from ATP to GMP. A single mutation is capable of disrupting the enzymatic activity, and at the same time confers the ability to bind an interaction partner that ultimately allows the new protein domain to function in spindle-orientation. Remarkably, the structural mechanism by which the functional transition is achieved does not depend on a specific mutation, rather other mutations preventing the GMP-induced conformational change in the GK enzyme also achieve the same effect and lead to a protein functioning in spindle-orientation. Hence, in this particular

example the evolution from an enzyme to a binding protein as a result of a single mutation is not an oddity but can be achieved via multiple mutational paths [103].

These two examples demonstrate that loss of enzymatic activity is not a death-sentence for genes inevitably leading to pseudogenisation, although this is the most likely outcome as it is generally the case for gene duplicates. How widespread this phenomenon really is has been assessed by several large-scale studies scanning the public protein sequence and structure databases for enzyme homologs with substitutions in their active sites [92, 93]. The surprising result is that inactive enzyme-homologs are the rule, not the exception: 10% of the enzyme domains in humans and even higher percentages in flies and *C. elegans* are predicted to have lost their enzymatic activity, with some superfamilies like RAS small GTPases reaching almost 50% likely inactive members [92]. As has been pointed out [104], these fractions are likely to be overestimates because of atypical enzyme mechanisms (see for instance [105]), nevertheless, the numbers are likely to remain substantial. Many of these inactive enzymes are under purifying selection and have therefore acquired a new beneficial function, which has been hypothesised to mainly be regulatory [92]. This is indeed an attractive and experimentally supported possibility, as the inactive enzymes can exploit their ability to bind substrates and co-factors of their enzymatic counterparts and often retain their tissue-specificity and subcellular localisation [104]. In summary, loss of enzymatic activity emerges as a powerful mechanism for regulatory evolution as well as functional evolution in general.

The degeneracy of the protein sequence-function map

The evolution of enzymes inspires to ask general questions about the relation between sequence and function in proteins. Of particular interest for the evolution of protein function is the degeneracy of this map, or in other

words: can different sequences perform the same specific function, and vice versa, can different functions be performed by the same sequence? In the case of binding the answer is trivial. A short glance at a protein interaction network confirms the degeneracy of the map as proteins can be bound by many other evolutionarily unrelated ones, and a protein may bind many proteins. In the case of enzymes the answer is less immediate.

The question if different, *i.e.* non-homologous, enzymes can perform the same function has recently been addressed by at least two studies [106, 107]. The authors differ by the stringency of their criteria: in one case, equal EC numbers indicate the same enzyme function and different structural folds are used to ensure non-homology [107], in the other case the authors additionally consider enzymes that only share the first three digits of the EC hierarchy but constrain non-homologs only to those cases that convergently evolved the same active sites [106]. Either way, in both cases the conclusion is that for more than 4% of the EC numbers non-homologous enzymes evolved. Hence, there is degeneracy in the map from enzyme sequence to function, meaning that different sequences evolutionary converged on the same enzymatic function.

The other direction is arguably more spectacular, that is a single sequence with different functions. Again, it turns out to be more common than one may expect: a series of multifunctional proteins from all over tree of life have been described coined moonlighting proteins. By definition, they perform autonomous functions not partitioned between different protein domains [108]. Strikingly, only in one of these cases a protein has two different enzymatic functions: an albaflavenone monooxygenase of the soil bacterium *Streptomyces coelicolor* has additional terpene synthase activity [109]. Most other cases consist of enzymes that have additional structural or regulatory functions.

In conclusion, the relationship between sequence and function is degenerate. The consequences for the understanding of the evolution of

function are abstract but fundamental: the fact that a single sequence provides a solution to several functional challenges and that a particular functional challenge can be met by different sequences makes functional evolution overall more probable. Using a popular analogy, not only are several needles in the haystack (several sequences for one function), but different people are looking for them simultaneously (several functions for one sequence).

1.2.3 Biophysical properties

A third possibility how the function of a protein can evolve is via change of biophysical properties. Biophysics has already been mentioned several times in the preceding subsections, for example in the context of new-function-stability tradeoffs [68] frequent in the evolution of new binding specificities or the loss of conformational flexibility as a mechanism to lose catalytic function in the guanylate kinase [103]. However, in these cases the change in a biophysical property like stability or flexibility only indirectly leads to functional change, *i.e.* merely represents a mechanism that affects binding or catalysis which in turn underlie the evolution of function. Yet, properties like length, stability and flexibility can also directly cause the evolution of protein function and examples from each of these categories are discussed below. In comparison to protein binding and catalysis, this mode of functional evolution is probably rare and relatively little is known about it. Rather than from dedicated studies the following examples are inferred from functional and evolutionary analyses conducted mostly for different purposes.

As already argued in Subsection 1.1.2, the continuous nature of biophysical parameters make them incompatible with our definition of functional evolution that requires qualitative differences (see Subsection 1.1.2). In the following, we therefore depart from the strictly biochemical definition of protein function applied so far and consider higher level, *i.e.*

cellular and organismal functional consequences. Higher order functional consequences stemming from quantitative changes are not restricted to biophysical properties, changes in binding affinity or catalytic efficiency can also have important qualitative phenotypic effects. For example, reduction of binding affinity of the garter snake voltage-gated sodium channel to tetrodotoxin, a neurotoxin toxin produced by some of these snakes' preys, confers actual physiological resistance to the toxin [110]. Similarly, the venoms of some shrews and lizards achieve toxicity via a serine protease paralog with enhanced catalytic efficiency. This enzyme releases a catalytic byproduct into the circulatory system which effectively becomes toxic due to its sheer amount [32].

Length

A clear case for the importance of protein length comes from molecular rulers. These molecules function in length control of specific cellular structures, which are synthesised to match the length of the molecular ruler [111].

An example can be found in certain pathogenic bacteria with a type-III secretion system, which consists of secreted effectors and proteins that form the injectisome, a stiff needle-like structure that is thought to function as a conduit for the secreted proteins. The needle length of the injectisome is controlled by a molecular ruler [112], and has been shown by manipulations to be essential for proper translocation of the effectors through the host cell plasma membrane [113]. In *Yersinia enterocolitica*, the protein acting as a ruler has internal repeats that contain stretches forming α-helices which have been shown to be the critical structural elements for length control [114]. Hence, internal repeats or mutations affecting the formation of helices are able to modulate protein length, which likely evolved to match specific distance requirements between the bacterium and its eukaryotic host cell [113]. As a consequence, changing the

length of this protein can evolve its cellular function in the control of an organismic phenotype, the pathogenicity of the bacterium.

Stability

A fascinating functional co-option of protein folding stability is found in bacterial protein thermosensors. Bacteria have evolved different mechanisms to sense temperature that are built around the central principle to exploit physico-chemical change of macromolecules that occur in response to temperature, for example in DNA, RNA, and most relevantly here proteins [115].

Pathogens of the genus *Yersinia* for example need to rapidly adapt their physiology after entering their warm-blooded hosts, often coinciding with the expression of virulence-associated genes. On the other hand, genes important for the initial stages of the infection like surface proteins that mediate binding to host cells are of little use in later stages or may even render the bacterium more susceptible to host immune responses [116]. Different *Yersinia* species do indeed downregulate expression of these genes upon entering the host using a direct temperature-sensing mechanism: the TF RovA reacts to the body-temperature of the host by reversible partial unfolding, which reduces its DNA binding affinity resulting in release from operator-sites and downregulation of transcription [117]. Most interestingly, substitution at three positions in RovA to match the amino acids found in a close homolog of RovA in *Salmonella* that remains stable and active at body temperature result in complete loss of thermosensitivity [118]. Hence, minor sequence changes causing loss of protein stability and resulting in a marginally stable protein domain provide the mean to evolve a new protein function.

Flexibility

Protein flexibility is another example of a biophysical property that can have direct functional consequences and is therefore relevant for the evolution of protein function. Probably the most intuitive class of proteins for which flexibility (or complementarily rigidity) is expected to be important are structural proteins, as for example cytoskeletal components. An intriguing possibility is therefore to hypothesise that different paralogs of cytoskeleton genes, for instance the highly similar actin isoforms, have different flexibilities which could underlie functional differences and ultimately their maintenance by purifying selection. Yet, while measurements of viscoelastic properties of distinct actin isoforms indeed found differences [119], the physiological importance of these differences remains unclear. Rather, the mechanical properties of actin networks are a consequence of varying dynamics [120] and structures [121] of actin-binding proteins (ABP) cross-linking the filaments. The slightly divergent regions of actin isoforms likely contribute to different binding affinities to ABP [122], and actin may therefore represent an interesting case where evolution of binding leads to functional evolution via altered biophysical properties.

A group of proteins for which flexibility itself has been suggested to be the key to functional evolution are pathogen effectors. The realisation that intrinsically disordered regions are overrepresented in secreted effectors of plant pathogens has led to the hypothesis that the conformational flexibility of disordered regions is a key structural feature of effectors. It could allow the effectors to be translocated through the type-III secretion system, to avoid recognition by the plant innate immune system, and mimic eukaryotic proteins [123]. Whereas in the latter two cases function is affected by flexibility only indirectly by altering the binding properties of the effectors, the requirement of flexibility for translocation can modulate if the protein reaches the eukaryotic cell altogether. Thus, in this case changing protein flexibility is an evolutionary degree of freedom with

drastic consequences for protein function.

1.3 The genetic basis of evolving protein function

In the following, we change the focus from biochemical aspects of functional evolution to its genetic basis. One of the fundamental prerequisites for evolution is the existence of phenotypic variation [28], serving as the raw material that evolutionary forces act upon. This variation has its root in genotypic variation which exists as the result of mutations in DNA, and is translated into phenotypes by physiology and development. In the following, we briefly summarise three different types of mutations leading to variation in protein function: modifications, duplications and rearrangements. In particular, we specify their rate whenever available and highlight consequences, biases and constraints of different types of mutations for the evolution of protein function.

1.3.1 Modification

Although point mutations affect only the minimum of one 'quantum' of genetic information, the functional consequences of a single point mutation can already be striking as discussed in Section 1.2. In particular, drastic effects are expected when mutations affect gene structure, for instance by causing gain (for example [101]) or loss (for example [124]) of stop codons. Note that even synonymous mutations may have functional consequences [125], yet effects are usually regulatory and therefore most likely affect higher functional levels. Synonymous mutations can reveal mutational biases deriving from genomic fluctuations in GC-content, as for instance reported in reference [126], implying a potential for the genomic location to influence the functional evolution of a gene.

While point mutations conserve protein length (with exception of those affecting stop codons), indels have the potential to alter it. Most indels are short (smaller than 5 residues [127, 128]), and indels are less common than substitutions: analyses found around 1 indel per 40 substitutions in coding sequence for both primates and bacteria [129]. Although less common, their accumulation has clear effects, as exemplified by the analysis of over 350 structural domain superfamilies, 60% of which showed at least 5% length variation from their typical size [127].

1.3.2 Duplication

Duplication of DNA is fundamentally important for the evolution of function as it introduces redundancy which can free the duplicated segment from purifying selective pressure [75]. Thus, duplication is a constant source of new and potentially functioning DNA material that can be tinkered with. The easiest way to discuss duplications is by order of how much DNA is affected. In general, there is a negative relationship between the length of the duplication and its frequency, *i.e.* small duplications are observed much more often than long ones [130].

Repeats are the smallest duplications, Microsatellites for example consist of repeating units of less than 10 nucleotides. Duplications altering the number of repeats may cause variation in repeat number in between 12% and 22% of all yeast genes [131] and can have important functional consequences. For instance, strength of cell adherence has been shown to scale with the number of repeats in the yeast cell surface adhesin FLO1 [132], likely mediated by an increased hydrophobic surface [133].

Larger internal duplications of gene segments are also frequent and have been found in between 8% and 16% of genes from six diverse eukaryotes [134]. They frequently lead to creation of novel introns and splice sites [134], therewith contributing to the functional diversification of proteins.

The duplication of segments containing an entire gene is of particular

interest, as unlike repeats and most internal segments the new stretch of DNA codes for an independent functional unit. Gene duplications are surprisingly frequent and have been found to occur at the same rate as point mutations [135]. They are probably the best studied duplication scenario, and evolutionary models of sequence divergence and functional dynamics are discussed in the next Section 1.4.2. Although all gene duplications by definition result in two copies of the original gene, the actual genetic mechanism of duplication influences the further evolution of the copies. A basic difference arises from genomic context: tandem duplications creating adjacent gene copies may be expected to preserve the regulatory context of a gene, whereas relocating one copy for example onto a new chromosome may not. A special case is retrotransposition: not only can the gene be inserted into a new regulatory landscape, but it is also cleared from all introns which often contain regulatory elements. The altered genomic context may therefore bias the evolutionary trajectory and predispose the gene for regulatory and functional evolution [136].

The most extreme cases of gene duplication are changes in ploidy, for instance whole genome duplication events. Whole genome duplications are interesting from a functional perspective because unlike duplication of a single gene, all interaction partners of a gene are also duplicated and therefore their relative dosage conserved. This may prevent immediate deleterious effects arising from changed post-duplication stoichiometry and give duplicated genes time to functionally evolve.

1.3.3 Rearrangements

A third class of mutations relevant for the evolution of protein function is genomic rearrangements. Basically, changing the mere localisation and context of a gene on the chromosome can influence its evolution, *e.g.* by fluctuations in GC-content already mentioned in Subsection 1.3.1. Another example are subtelomeric regions, which have been shown to be particu-

larly dynamic with elevated rates of duplication and recombination [137].

Recombination, *i.e.* the exchange of genetic information between two DNA molecules, is an important evolutionary force, as it breaks the associations imposed by the linear structure of the DNA strand. For example, intergenic recombination in the context of pralogous gene families has been observed to increase the diversity of alleles and functional variants by generating new genes carrying combinations of the mutations that separately arose in the different copies [138]. Because each individual mutation has been subject to purifying selection before, the recombined variants may be expected to be less often deleterious, as shown for instance using engineered β-lactamase chimera [139]. Recombination thus provides a mechanism to increase the probability of finding and fixing positively epistatic mutations. In directed evolution experiments, recombination has been found to be able to generate both new ligand specificities [140] and more stable proteins [141, 142].

Gene fusions are another outcome of rearrangements. Fused genes are frequent, an early study for example found around 5% of *E. coli* genes to have been involved in fusion events [143]. An interesting example in eukaryotes are three independent fusions of the *adh* gene in *Drosophila*, which were hypothesised to subsequently have followed convergent path of functional evolution [144]. The impressive potential of fusion to couple functionally unrelated genes has been demonstrated using a directed evolution strategy. The fusion of an *E. coli* maltose binding protein and TEM1 β-lactamase generated a β-lactamase with switch like catalytic activity dependent on the presence of maltose [145]. Thus, gene fusions emerge as a powerful mutational mechanism to generate new complex functions.

1.4 Evolution and protein function

So far, this chapter discussed the biochemical basis of how protein function evolves, and what mechanisms generate the necessary genetic variation serving as raw material for this process. Here, we consider the forces that sieve the pool of random mutations by guiding the path of functional change. In other words, we discuss evolution.

How protein function evolves has been studied in molecular evolution for a long time, however, usually implicitly by following a 'statistical paradigm' focussed around the origin and maintenance of genetic variation [10]. The strength of this approach is the generality of its results which avoid the reference to any function in particular. This is achieved by relating sequence variation to fitness directly and leaving the phenotypic level out altogether, or by considering abstract phenotypic spaces for the study of adaptation as for example in Fisher's geometric model [146]. The price of this generality is that mechanistic explanations for fitness differences are impossible, as those can only be derived from an explicit representation of a concrete phenotypic space. In particular, protein function evolves in phenotypic space, and can therefore only be studied indirectly when this level is omitted. In order to complement the classical statistical approach and study phenomena as for instance the evolution of protein function, alternative paradigms have been out forth which stress the benefits of integrating the three levels genotype, phenotype, and fitness [11, 147, 148]. They can be seen as part of a larger 'functional synthesis' subsuming most of what has been discussed in this chapter [10]. Yet, also as part of this synthesis the comparative analysis of sequences in search of statistical patterns left by evolution remains fruitful for functional analysis, in particular as more and more genomes become available. In the following, we detail aspects of gene sequence and -family evolution that are relevant for the study of protein function.

1.4.1 Gene sequence evolution

The evolutionary fate of a mutation is tightly linked to its effect on organismal fitness. Although fitness effects are continuous, one can qualitatively distinguish three categories: beneficial, neutral and deleterious mutations. The proportions of mutations falling into each of these categories and in particular the functional form of mutational effects are old, important yet mostly unsolved problems in evolution [149]. At least in experimental evolution studies, typically $30-50\%$ of point mutations are strongly deleterious, $50-70\%$ can be considered neutral and roughly $0.5-0.01\%$ are beneficial [51]. The type of selective pressure a mutation will be exposed to, if any, critically depends on which category it belongs to: beneficial mutations are swept to fixation by positive selection, neutral mutations are invisible to selection but may be fixed as a result of drift, and sufficiently deleterious mutations are purged from the population due to purifying selection.

The value of positive selection for studying the evolution of protein function mostly lies in the following reasoning: if a statistical footprint left by positive selection can be detected, one can infer that the site in question is functionally important and its alteration contributed to the beneficial consequences. Therefore, inferring selection can serve to guide further functional investigation (see *e.g.* [150] and references therein). However, the concrete functional consequences of a positively selected site are by no means obvious, in particular they do not need to be at the biochemical level. For example, selection for protein stability already mentioned in Subsection 1.1.2 does not necessarily alter the biochemical function of a protein itself, yet, at the cellular level the function may be overall physiological robustness.

In the case of neutral mutations, the importance for the evolution of function may be less obvious: as neutrality by definition implies no effect on fitness, neutral mutations may be expected to be irrelevant for

functional evolution. However, this turns out not to be the case. Two examples were already mentioned in Subsection 1.2.1 [51]: first, epistatic interactions frequently alter the fitness effect of mutation that are neutral in isolation. These can for instance stabilise protein structures and therewith open evolutionary paths to new substrate specificities otherwise blocked due to biophysical constraints resulting from marginal protein stability. Second, they may contribute to binding promiscuity, which in turn can facilitate for example the evolution of new specificities to previously low-affinity binding partners.

Lastly, there is purifying selection, which is a conservational force preventing the accumulation of mutational changes and therefore the loss of protein function. The hallmark of purifying selection, conservation of sequence beyond neutral expectation, is an important statistical footprint for function and at the base of most genomic screens for functional elements [151]. Conceptually, these methods follow the basic comparative reasoning mentioned in the introduction and applied since the early 1950s, however, in addition they properly account for the phylogenetic relationships which underlie the evolutionary process [31].

1.4.2 Gene family evolution

Besides protein sequence itself, another level of evolution is crucial for innovation in protein function and functional change in general: the evolution of gene families. Without the occurrence of gene duplications, a gene family consists of at most one ortholog per species as the only family-level event is gene loss. Thus, in this case the only dynamics important for functional evolution is sequence divergence at the level of genes, governed by the processes described above. In contrast, gene duplications lead to more intricate pattens, and various models for gene duplication have been developed. Note however that despite its importance duplication is not a prerequisite for functional evolution, as orthologs may also diverge in

function [59, 152–154].

The most basic models merely describe the patterns of gene family evolution. They distinguish two basic scenarios for the divergence within and between orthologs and paralogs: divergent- and concerted evolution. Divergent evolution was the first pattern observed in the 1960s for example in hemoglobins, and is characterised by higher sequence divergence after duplication than after speciation, leading to multiple clades of orthologs in the gene phylogeny. Concerted evolution is the opposite case, *i.e.* lower sequence divergence after duplication, resulting in clades of paralogs in the phylogeny. Concerted evolution can be caused by frequent gene conversion [155], and is the mode of evolution for instance of rRNA genes. Another scenario stresses the effect of different rates of duplication and loss, modelling gene family evolution as a birth death process [156]. These basic evolutionary patterns can already be correlated with functional patterns: the vertebrate cytochrome P450 family for example can be partitioned into a stable part evolving by divergent evolution and having core functions, and an unstable one better modelled by a birth-death process and having accessory functions [134].

Gene duplication has also been studied specifically in the context of functional evolution. The importance of gene duplication as a source of new genetic material for functional evolution has been hypothesised since the 1930s [157]. However, the link with innovation is mostly mentioned in connection with Susumu Ohno, who boldly advocated the importance of duplication: "natural selection merely modified, while redundancy created" [75]. In the following, we summarise three scenarios for gene duplication. Ohno initially formulated two models connecting evolutionary patterns and functional consequences, and a third possibility only requiring neutral processes has been proposed later. While one of these scenarios often well describes actual events, they are not exclusive and may overlap or co-occur in the evolution of a gene family (see *e.g.* [54]).

Neofunctionalisation

Ohno presented redundancy as a solution to the problem of how a mutational path could proceed through non-functional intermediates that would otherwise be purged. In other words, redundancy functions a mechanism to defy purifying selection. The neofunctionalisation model applies this idea to the evolution of duplicate genes: after gene duplication, the copies are functionally redundant, therefore effectively freeing one gene copy from the conservative force of purifying selection and allowing it to explore the sequence neighbourhood independently of deleterious effects of mutations on the original function. If this path encounters a protein sequence able to perform a new beneficial function, it is picked up by positive selection and fixed in the population, forming a new gene with a new function that is in turn protected from accumulating deleterious mutations by purifying selection [75]. Hence, neofunctionalisation is a general model for the innovation of function, satisfying the requirements of the selected effects definition of function [3] (*i.e.* past positive selection, current purifying selection [4]). It is general in the sense that it follows the statistical paradigm of molecular evolution ignoring the phenotypic level. Most relevantly, it does not specify how mutations can bring about a new function (discussed in Section 1.2).

Dosage

A second possibility presented by Ohno is retention of a gene duplicate due to beneficial effects of higher dosage. In this case, both copies are conserved by purifying selection, leading to pattern of concerted gene family evolution. This model does not lead to functional innovation at the biochemical level, and is therefore less interesting in the context of this chapter.

However, dosage can nonetheless play an important role in functional

evolution as part of more complex gene duplication models. For instance, poxviruses transiently increase the copy number of genes that counteract host defence in order to make up for the initially low efficacy of these genes in a species they are maladapted for [158]. The greater number of genes and their protection from deleterious mutations by purifying selection increases the probability of mutations optimising at least one of the copies [159]. This mechanism does not necessarily imply evolution of protein function beyond quantitative adjustments of efficacy or affinity, however, the same principle also works in conjunction with neofunctionalisation [160]. In what has been called the Innovation-Amplification-Divergence (IAD) model [161], promiscuous side-functions which may not be efficient are the cause for purifying selection initially maintaing the multiple copies. One copy subsequently neofunctionalises by the action of positive selection turning one of these promiscuous side-activities for example into the main substrate of an enzyme [53].

Subfunctionalisation

Subfunctionalisation offers an alternative explanation for gene duplicate retention. In the tradition of the critique of adaptionism [162], it demonstrates the important possibility that nonadaptive processes alone, *i.e.* without requiring the action of positive selection, can lead to increased genomic complexity [163]. After duplication, the initially redundant gene copies accumulate deleterious or neutral mutations that partially disrupt complementary parts of the function. This renders both copies necessary to perform the original function, and both are conserved by purifying selection. Complementarity can for instance be achieved by splitting the set of tissues a gene is active in [164], a reduction in individual performance requiring two copies to attain the original activity level, or by partitioning any type of discrete function like interactions [163]. Unlike neofunctionalisation, subfunctionalisation is therefore a model in which function evolves

by loss or reduction.

An interesting variation of subfunctionalisation is the EAC-model (Escape from Adaptive Conflict), in which a constraint of any sort prevents the independent optimisation of any of (at least) two functions of an unduplicated protein, creating an adaptive conflict. After duplication, partitioning the functions between the copies alleviates the conflict and positive selection is free to optimise the functions separately (see *e.g.* [165, 166]).

Finally, more complicated scenarios can result by introducing functional coupling, for example when distinct protein domains interact to perform the protein's function or if the protein homodimerises. A scenario of this kind has been described for duplications of steroid hormone receptors, ligand-activated TFs that consist of a DNA and ligand binding domain. Degenerative mutations in either of these two domains can result in a partially inactive protein that competes for either DNA binding sites or the ligand and therefore functions as a repressor [58]. Thus, a new protein function emerges, yet solely by virtue of degenerative mutations which increases the probability of this scenario and therefore of functional innovation. However, the evolution of a new function is not a necessary outcome. This scenario may simply lead to an increase in the number of molecular components and therefore complexity of molecular assemblies without change in function[167, 168].

1.5 Rab GTPases: the evolution of function in protein switches

In the present section, we argue that protein switches represent a relevant and interesting class of proteins to explore the evolution of function. Furthermore, we introduce the Rab family of small GTPases analysed in the remainder of this thesis as a promising model gene family for protein switches.

Switching, *i.e.* the conformational change of a protein in response to a signal or event, can be considered one of the fundamental biochemical function of proteins (see Subsection 1.1.1). The nature of the of the trigger leading to switching can be very diverse: conformational changes result from hydrolysis of GTP or ATP, binding of ions like for example calcium, posttranslational modifications, change in PH or even from light.

The biochemical function of protein switches can evolve in unique ways. For example, the transcription factor CEBPB is essential in placental mammals and initiates transcription after being activated via conformational switching in response to phosphorylation. Using ancestral sequence reconstruction, it could be shown that three mutations introduced and removed phosphorylation sites that lead to an inversion of CEBPB behaviour: rather than being repressed, phosphorylation activates the transcription factor [169]. Hence, in this case very few mutations are capable of altering the switching mechanism representing a mode of functional evolution exclusive to protein switches. While this is a striking example of a direct effect of switching on function, it is likely that evolution of switching leads to different functions mostly indirectly via alteration of binding specificities. An illustration for this mechanism is the calcium-binding C_2-domain. It has been shown that skipping an exon of only nine amino acids by alternative splicing abrogates the conformational change needed to activate the ability to bind calcium, generating an isoform able to permanently bind calcium [170]. Hence, in this case small changes are enough to alter switching and indirectly affect protein function via binding.

As already mentioned in Subsection 1.2.1, this latter mode of functional evolution by tinkering with binding specificities is the most common form of evolution of function, and therefore of upmost importance. Protein switches are most relevant models to study this phenomenon as they are a critical class of proteins, involved in the regulation of key biological processes and intimately linked to central organisational features

of the cell like compartmentalisation. They are compelling models as they are subject to a unique set of constraints that is interesting to consider in the context of functional evolution: maintenance of enough flexibility to switch conformations, yet without resulting in unstable proteins; multi-specificity to often to a large number of interaction partners; and enzyme function in case of switching triggered by nucleotide hydrolysis. With this argument in mind, we introduce a particular family of protein switches in more detail—the Rab family of small GTPases—that serves as a model to inquire into the evolution of function in the remainder of this thesis.

1.5.1 The Rab family of small GTPases

Rabs are GTPase enzymes belonging to the Ras superfamily, which is also referred to as small GTPases. Structurally, the Ras superfamily is a representative of the P-loop NTPases, one of the chain folds found in proteins that bind and hydrolyse nucleoside triphosphates including ATP and GTP. P-loop NTPases are the most populous protein fold in many cellular organisms, comprising $10 - 18\%$ of all gene products [171]. GTPases are a monophyletic superclass within P-loop NTPases, that can be—based on sequence and structure—further divided into two large classes and have been coined SIMIBI (after its three biggest subgroups, the Signal recognition GTPases, the MinD superfamily and the BioD superfamily) and TRAFAC (for translation factor-related). The latter includes the extended Ras-like superfamily, and other important GTPases like heterotrimeric G proteins that are activated by G-protein coupled receptors [171].

Ras-like proteins are found in all three superkingdoms of life, however, in prokaryotes the only member is the MglA family. In the bacterium *Myxococcus xanthus*, MglA functions in motility, sporulation, and morphogenesis. Interestingly, a MglA gene deletion leads to sporulation defects which can be rescued by a yeast small GTPase [172]. This emphasises that not only the overall structure but also fundamental aspects of the func-

tional mechanism are conserved across the superkingdoms. In eukaryotes, small GTPases have massively expanded as early as in the last eukaryotic common ancestor, laying the foundation for the extraordinary diversity of Ras-like families and proteins observed in extant eukaryotes [171]. For example, the human Ras superfamily comprises over 150 genes [173], and excavate genomes with over 320 small GTPases have been sequenced [174]. The Ras superfamily owes its name to its founding members, genes that were identified in rats as the targets of transformation by rat-derived Harvey and Kirsten murine sarcoma retroviruses and therefore named Ras (from 'rat sarcoma') [175]. Ras have signalling functions and remain the most extensively studied family within the superfamily because of their critical role in human oncogenesis [176].

Among the members of the Ras superfamily, Rabs are the largest family, for instance accounting for over 60 of 150 Ras superfamily genes present in the human genome. Rabs were originally described in yeast due to their similarity with Ras [177], and suggested to function in microtubule organisation [178]. Shortly after, another member of the the yeast Rab family was implicated in secretion [179], and its functional mechanism began to be worked out [180]. In the meantime, Rabs were shown to be widespread in eukaryotes and conserved in sequence [181, 182] and function [183]. The name is derived from the origin of the library they were identified in, 'Ras genes from rat brain' [182].

Sequence and structure

Rabs are relatively short proteins of length around 220 amino acids. They can be partitioned roughly into a central conserved region that contains various GTPase and Rab-specific motifs and that is flanked by two more divergent so-called hypervariable termini (see Figure 1.1).

The conserved region harbours six different types of motifs hierarchically defining the sequence as a Rab. First, the catalytic triad, *i.e.* three

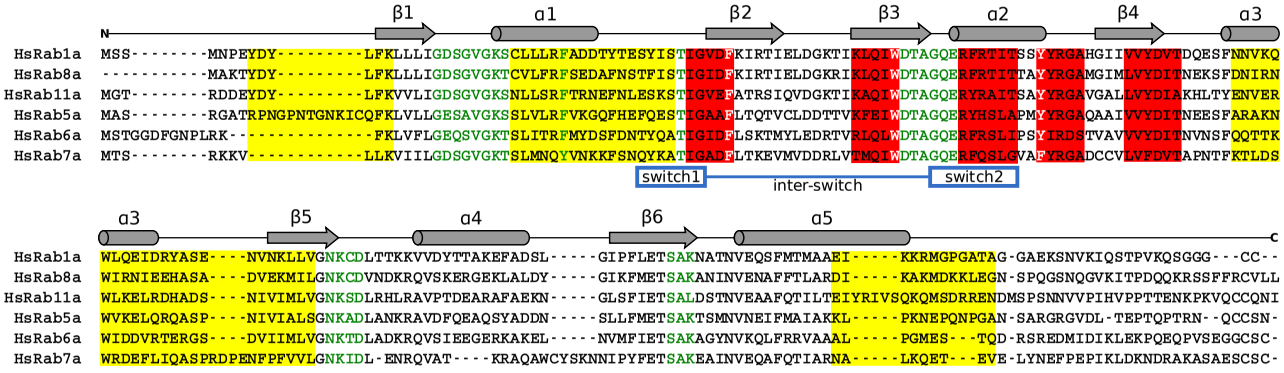


FIGURE 1.1: *Sequence motifs defining Rabs*—Multiple sequence alignment (generated with Prank [184]) of some human Rabs (Uniprot accessions P62820, P61006, P62491, P20339, P20340, P51149), representing ancestral subfamilies found in most eukaryotes. The coloured residues and regions correspond to motifs required for different aspects of Rab function that define the identity of the sequence as a Rab (see main text). Shown are the catalytic triad (white residues); motifs required for the enzymatic function which are referred to as G1−G3 (regions interacting with guanine) and PM1−PM3 (regions interacting with phospate or Mg) (green residues) and occuring in the order PM1 (corresponding to the P-loop), G1, PM2 (the Walker B motif), PM3, G2 (the GTP-specificity motif), G3; the RabF1 − 5 (Rab family) motifs [185] (red areas); and the RabSF1 − 4 (Rab subfamily) motifs [185, 186] (yellow areas). Furthermore, the secondary structure elements are indicated above the sequence, and the switch region is shown which is the site of conformational change upon nucleotide-binding (blue) (see main text).

amino acids shared by certain hydrolase and transferase enzymes. Second, two motifs that define P-loop NTPases, the N-terminal Walker A (GxxxGK[ST]) motif or P-loop that binds phosphate and the distal Walker B motif (DxxG) that binds a Mg ion. Third, besides having a specific form of the Walker B motif, GTPases have an additional distal [NT]KxD motif conferring specificity to GTP and thus not found in other P-loop NTPases [171]. Fourth, small GTPases share four further motifs, two single conserved [FY] and T residues, the latter located in the second loop also known as the effector loop, and two conserved strips with consensus DTAGQ and SAK respectively. These residues participate in the interactions with guanine, phosphate and Mg. Fifth, bioinformatic analysis determined five Rab family motifs distinguishing them from other families of the Ras superfamily [185]. These regions are important for the interaction with Rab-specific regulators discussed later. Finally, Rab subfamily motifs recapitulate the partitioning of the family into subfamilies based on overall sequence similarity, with the first and last motif extending into the hypervariable termini [185, 186]. The regions corresponding to the subfamily motifs have been proposed to confer specificity to the distinct sets of effectors of Rab subfamilies.

Rabs are protein switches belonging to the structural class of α/β proteins, with an antiparallel sheet opposite to the strand with the Walker B motif distinctive of GTPases of the TRAFAC class [171]. A diagram of the relative order of the six β-sheets and five α-helices common to small GTPases is shown in Figure 1.2. The region before the antiparallel β-sheet $\beta 2$ and the one after the adjacent β-sheet $\beta 3$ constitute the switch 1 and 2 respectively. The switch region as a whole including the inter-switch region is the primary structural determinant of nucleotide-state dependent function, undergoing major conformational change between the GTP- and GDP-bound state. In the inactive GDP-bound conformation the

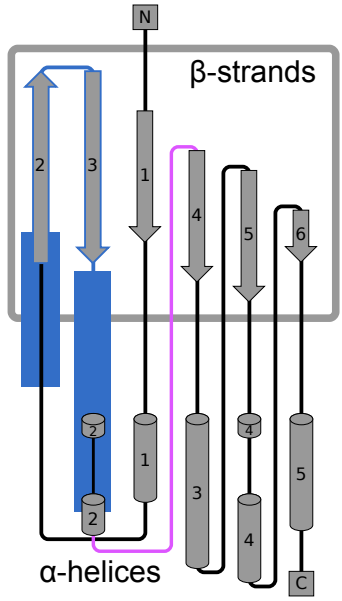


FIGURE 1.2: *Rab structural topology*—The organisation of Rab secondary structure elements represented as a 2D topology diagram (generated with Pro-origami [189] and manually recoloured). The sheets and helices are numbered $\beta 1-6$ and $\alpha 1-5$ respectively corresponding to the sequences shown in Figure 1.1. The switch region changing conformation upon nucleotide-binding is highlighted (blue) (see main text). For clarity, overlapping connections between secondary structure elements are drawn in a different colour (violet) [189].

switch region tends to be unstructured, whereas a structured conformation defines the active state and is involved in conferring specificity to the binding of effectors [187, 188]. The full spatial organisation of Rab proteins and in particular the switch region is exemplified by Figure 1.3.

Function

Rabs function as molecular switches cycling between a GTP-bound active and GDP-bound inactive state, primarily defined by the conforma-

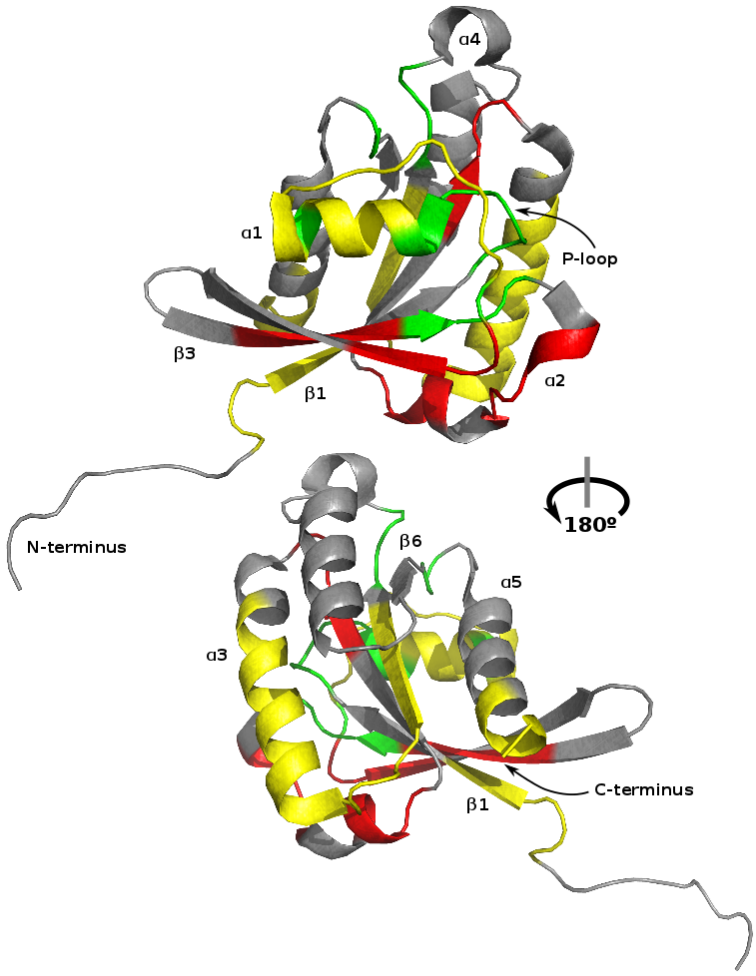


FIGURE 1.3: *Ypt1 (yeast Rab1 ortholog) crystal structure*—Crystal structure of Ypt1, the yeast Rab1 ortholog, in active conformation bound to a GTP analog (not shown) [190]. Regions corresponding to the different Rab motifs shown in Figure 1.1 are coloured equivalently (green: Rab G1−3 and PM1−3 motifs, red: RabF motifs [185], yellow: RabSF motifs [186, 191]). The switch region (see main text) covers β-sheets $\beta 2$ (labelled) and the immediately adjacent $\beta 3$ and is located in front in the upper view, on the back in the rotated view.

tion of the switch region that changes depending on the nucleotide. The conformation of the active state is recognised by effector proteins that preferentially bind the GTP-bound form, although proteins specific to the GDP-bound form are also known [192]. The recruited effectors perform the functions regulated by Rabs in their respective pathways, which are discussed below.

Besides the switching between active and inactive states, Rabs spatially cycle between a cytosolic and membrane-bound form. The full Rab cycle is summarised in Figure 1.4. After translation, a GDP-bound Rab associates with a Rab escort protein (REP) that presents it to a Rab geranylgeranyl transferase (RabGGT) which posttranslationally modifies a C-terminal prenylation motif. The modified Rab is delivered to the target membrane, where its now hydrophilic tail allows membrane insertion that may be assisted by a GDP dissociation inhibitor (GDI) dissociation factor (GDF). In the membrane-inserted state ('In'), the Rab is activated ('On') by the aid of a guanine exchange factor (GEF) that catalyses exchange of GDP for GTP. It is now that effector-binding occurs. The low intrinsic GTP-hydrolysis is accelerated by a GTPase-activating protein (GAP) that promotes the formation of the inactive GDP-bound state ('Off'). Finally, the spatial cycle between the membrane-bound and cytoplasmic pool ('Out') of Rabs is closed by GDIs, that extract Rabs from the membrane and deliver them back to GDFs for reinsertion (see [173, 188] for reviews).

In the cellular context, Rabs function as master regulators of membrane traffic. The active forms are localised in the membrane of a variety of compartments where the recruited effectors mediate different steps of vesicular trafficking. Figure 1.5 summarises the known intracellular localisation of most vertebrate Rab subfamilies (adapted from reference [188]). The mechanisms targeting Rabs to their respective compartment are not fully understood, but the hypervariable C-terminus [193, 194] and cer-

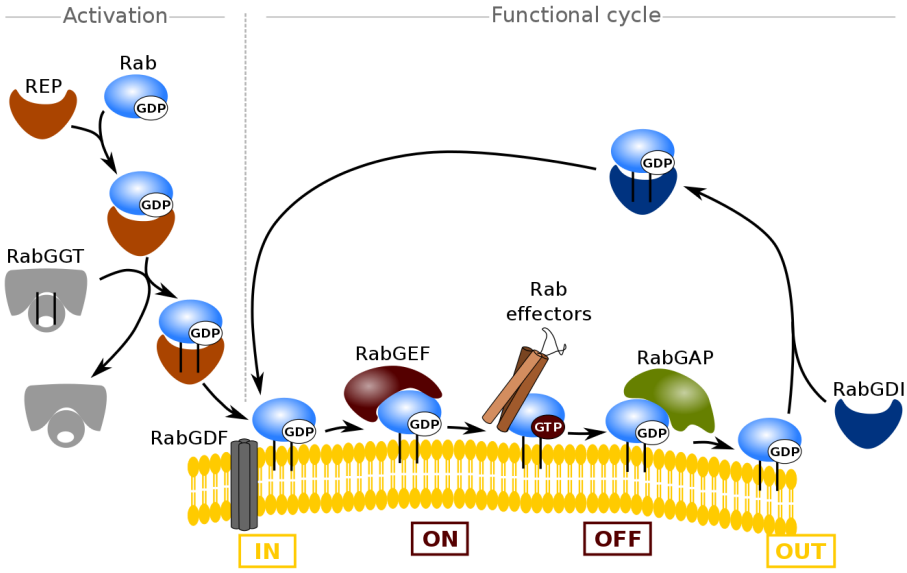


FIGURE 1.4: *Rab cycle*—The full Rab cycle consisting of Rab activation and two cycles switching between the membrane-inserted 'In' and cytosolic 'Out' state, and the GTP-bound 'On' and GDP-bound 'Off' state (see main text). *Abbreviations:* Rab escort protein (REP), Rab geranylgeranyl transferase (RabGGT), GDP dissociation inhibitor (GDI), GDI dissociation factor (GDF), guanine exchange factor (GEF), GTPase-activating protein (GAP),

tain RabF and RabSF regions [195] have been experimentally implicated. The functions performed by effectors that Rabs regulate can be associated to distinct steps of vesicular trafficking. First, cargo is selected for coated vesicle trafficking and enclosed in a forming bud. Afterwards, the matured vesicle is transported by molecular motors and cytoskeletal filaments to the acceptor membrane. After uncoating, tethering factors ensure that the vesicle is then fused with the right membrane. Effectors involved in each of these steps have been found to be regulated by specific Rabs (see [188, 196] for reviews).

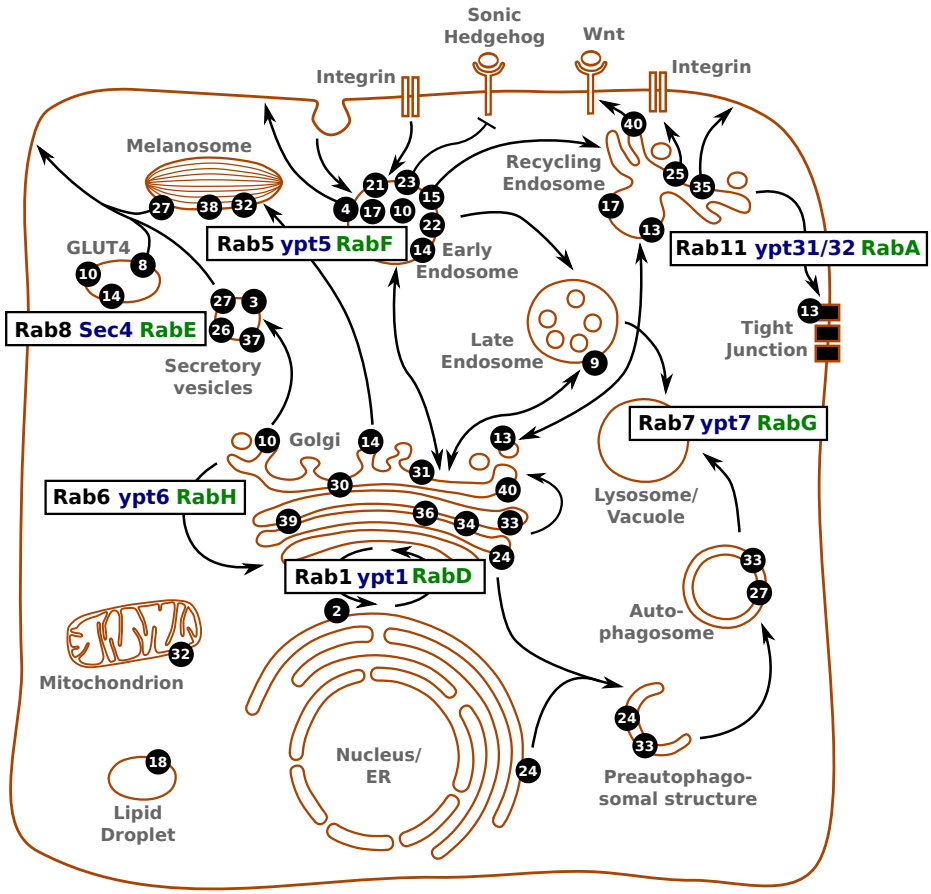


FIGURE 1.5: *Intracellular localisation of vertebrate Rab subfamilies*—Rab subfamilies found in vertebrates are shown at their intracellular localisation together with the associated trafficking pathways (adapted from reference [188]). The boxed subfamilies are found in most eukaryotes, and the names of the orthologs in vertebrates (black), fungi (blue) and plants (green) are given.

1.6 Conclusion

Biological phenomena occur on many levels. Here, we consider the fascinating question how protein function evolves, and identify three levels

essential to a complete answer and relevant for the rest of this thesis: biochemistry, genetics and evolution. This list could be further extended: an obvious perspective missing for a complete understanding is ecology, which is concerned for example with the agents of selection, *i.e.* the causes for natural selection [197].

The main conclusions from reviewing work on the structure of the protein genotype-phenotype map and the protein phenotype or function space itself are twofold. First, the evolution of protein function is incredibly dynamic. Probably the most powerful examples come from proteins that evolve across functional categories: enzymes evolve from non-enzymes [93, 94] and vice-versa [92]. We find that regardless of the type of protein the mutational paths to new functions can be short, frequently even single amino acid changes have dramatic effects. This implies that mutations of large effect not only exist, but also occur in the evolution of protein function. Yet, this does not contradict the fact that functions can be conserved even in very divergent proteins, also a consequence of the unintuitive nature of the high-dimensional sequence space. Even if mutational paths are short, they are not straight. There is ample evidence for epistasis in the evolution of protein function, which leads to rugged landscapes and complicates the prediction of mutational effects. Second, the evolution of protein function is not unlikely. This is a common criticism of evolutionary theory in general, where apparent unlikeliness is perverted to justify the existence of higher powers. In proteins, the argument of unlikeliness has been most prominently addressed by Maynard-Smith [25], but research since then adds a few nuances. Besides the dynamic nature of functional evolution discussed above, this is most relevantly the degeneracy of the protein genotype-phenotype map. A protein can serve many functions, and a function can be carried out by different proteins, which increases the likelihood that particular pairs of protein and function evolve.

1.6.1 Rabs to study the evolution of function

In Section 1.5, we argue that protein switches are interesting yet understudied models for the evolution of protein function, and introduce Rabs as the model family for the rest of this thesis. What makes Rabs well-suited for the study of functional evolution?

First of all, Rabs can be seen (and are even used in experimental practice) as markers for certain compartments and trafficking pathways, due to their essential functions in vesicle transport and organelle identity [198]. Their evolution is therefore relevant for the function and evolution of the entire endomembrane system, that fundamentally shapes intracellular organisation and defines eukaryotes. As protein switches, they are subject to various constraints already mentioned above, *e.g.* flexibility, stability, multispecificity and enzyme function, providing the opportunity to study the influence of these factors on the evolution of function. With publication of the first complete genomes, it became clear that the Rab family greatly expanded in animal evolution, which suggests that despite multiple constraints Rab proteins easily evolve new functions. This counterintuitive observation stimulates interesting questions about the mechanisms of functional evolution. Not only in animals, but throughout the eukaryotic tree of life the Rab family displays the interesting pattern of a conserved core of subfamilies complemented by lineage-specific evolutionary events, resulting in many independent cases of functional evolution that can be studied. The annotation of eukaryotic Rabs and the resulting patterns of Rab subfamily-level evolution are subject of Chapter 2 within this thesis.

In more practical terms, Rabs are a good model proteins because a rich body of work established publicly accessible data on sequence, structure and function, including expression, localisation and interactions. Some isolated cases of functional evolution have already been studied. For instance, the duplication of a Rab6 by retrotransposition in the ancestor of apes and humans lead to paralogous gene copy that lost its ability to

bind GTP and therefore catalytic activity. In lieu of localisation to the Golgi like other Rab6, the new isoform localises to the centrosome where it functions in cell cycle progression [124]. Hence, cases do exist in which Rab function evolved in non-trivial ways underscoring the interest in Rabs as a model. Chapter 4 asks how Rab function in general evolves after duplication, and in particular analyses the Rab11 duplication that gave rise to Rab25.

On the other hand, testing hypotheses about the evolution of function in Rabs also poses technical challenges. Most importantly, Rabs carry little phylogenetic information due to their size and mixture of highly conserved and highly divergent sequence regions. This complicates obtaining phylogenetic trees, which form the basis of for all further computational analysis of Rab sequences. This particular challenge is addressed later in Chapter 3.

References

[1] August Krogh. “The Progress of Physiology”. In: *8th International Physiological Congress*. Baltimore, Aug. 1929, pp. 1–9.

[2] Bruno J Strasser and Soraya de Chadarevian. “The Comparative and the Exemplary: Revisiting the Early History of Molecular Biology”. In: *Science History* 49.3 (Sept. 2011), p. 317.

[3] Peter Godfrey-Smith. “A Modern History Theory of Functions”. In: *NOÛS* 28.3 (Apr. 1994), pp. 344–362.

[4] W Ford Doolittle. “Is junk DNA bunk? A critique of ENCODE”. In: *Proceedings of the National Academy of Sciences of the United States of America* 110.14 (Apr. 2013), pp. 5294–5300.

[5] Crysten E Blaby-Haas and Valérie de Crécy-Lagard. “Mining high-throughput experimental data to link gene and function”. In: *Trends in biotechnology* 29.4 (Apr. 2011), pp. 174–182.

[6] L Aravind. “Guilt by association: contextual information in genome analysis”. In: *Genome Research* 10.8 (July 2000), pp. 1074–1077.

[7] William C Earnshaw. “Deducing protein function by forensic integrative cell biology”. In: *PLoS Biology* 11.12 (Dec. 2013), e1001742.

[8] Gregory A Petsko and Dagmar Ringe. “From Sequence to Function: Case Studies in Structural and Functional Genomics”. In: *Protein Structure and Function*. New Science Press, Jan. 2004.

[9] Andrew D Hanson et al. “’Unknown’ proteins and ’orphan’ enzymes: the missing half of the engineering parts list—and how to find it”. In: *The Biochemical Journal* 425.1 (2010), pp. 1–11.

[10] Antony M Dean and Joseph W Thornton. “Mechanistic approaches to the study of evolution: the functional synthesis”. In: *Nature Reviews Genetics* 8.9 (Sept. 2007), pp. 675–688.

[11] Michael J Harms and Joseph W Thornton. “Evolutionary biochemistry: revealing the historical and physical causes of protein properties”. In: *Nature Reviews Genetics* 14.8 (Aug. 2013), pp. 559–571.

[12] Robert Cummins. “Functional Analysis”. In: *The Journal of Philosophy* 72.20 (1975), pp. 741–765.

[13] Ron Amundson and George V Lander. “Function Without Purpose: the Uses of Causal Role Function in Evolutionary Biology”. In: *Biol Philos* 9 (Oct. 1994), pp. 443–469.

[14] Gregory A Petsko and Dagmar Ringe. *Protein Structure and Function*. Primers. New Science Press, Jan. 2004.

[15] Gregory A Petsko and Dagmar Ringe. “From Sequence to Structure”. In: *Protein Structure and Function*. New Science Press, Jan. 2004.

[16] Pouria Dasmeh et al. “Positively selected sites in cetacean myoglobins contribute to protein stability”. In: *PLoS Computational Biology* 9.3 (Mar. 2013), e1002929.

[17] George N Somero. “Protein adaptations to temperature and pressure: complementary roles of adaptive changes in amino acid sequence and internal milieu”. In: *Comparative biochemistry and physiology. Part B, Biochemistry & molecular biology* 136.4 (Dec. 2003), pp. 577–591.

[18] Stephen Jay Gould and Elisabeth S Vrba. “Exaptation—A missing term in the Science of Form”. In: *Paleobiology* 8.1 (1982), pp. 4–15.

[19] Wan-Jin Lu, James F Amatruda, and John M Abrams. “p53 ancestry: gazing through an evolutionary lens”. In: *Nature Reviews Cancer* 9.10 (Oct. 2009), pp. 758–762.

[20] Arnau Sebé-Pedrós et al. “Ancient origin of the integrin-mediated adhesion and signaling machinery”. In: 107.22 (June 2010), pp. 10142–10147.

[21] Arnau Sebé-Pedrós et al. “Unexpected repertoire of metazoan transcription factors in the unicellular holozoan Capsaspora owczarzaki”. In: 28.3 (Mar. 2011), pp. 1241–1254.

[22] Arnau Sebé-Pedrós et al. “Premetazoan origin of the hippo signaling pathway”. In: *Cell reports* 1.1 (Jan. 2012), pp. 13–20.

[23] Zachary D Blount et al. “Genomic analysis of a key innovation in an experimental Escherichia coli population”. In: *Nature* 489.7417 (Sept. 2012), pp. 513–518.

[24] K Tipton and S Boyce. “History of the enzyme nomenclature system”. In: *Bioinformatics* 16.1 (Jan. 2000), pp. 34–40.

[25] John Maynard Smith. “Natural Selection and the Concept of a Protein Space”. In: *Nature* 225 (1970), pp. 563–564.

[26] Sewall Wright. “The roles of Mutation, Inbreeding, Crossbreeding and Selection in Evolution”. In: *Proceedings of the 6th International Congress of Genetics* 1 (Aug. 1932), pp. 356–366.

[27] Massimo Pigliucci. “Sewall Wright’s adaptive landscapes: 1932 vs. 1988”. In: *Biology and Philosophy* 23.5 (Nov. 2008), pp. 591–603.

[28] Richard C Lewontin. “The units of selection”. In: *Annual Review of Ecology and Systematics* 1 (1970), pp. 1–18.

[29] Maria Anisimova et al. “State-of the art methodologies dictate new standards for phylogenetic analysis”. In: *BMC Evolutionary Biology* 13 (2013), p. 161.

[30] Antonio Rausell et al. “Protein interactions and ligand binding: from protein subfamilies to functional specificity”. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.5 (Feb. 2010), pp. 1995–2000.

[31] Joseph Felsenstein. “Phylogenies and the comparative method”. In: *The American Naturalist* 125.1 (1985), pp. 1–15.

[32] Yael T Aminetzach et al. “Convergent evolution of novel protein function in shrew and lizard venom”. In: *Current Biology* 19.22 (Dec. 2009), pp. 1925–1931.

[33] Guoping Zhu, G Brian Golding, and Antony M Dean. “The selective cause of an ancient adaptation”. In: *Science* 307.5713 (Feb. 2005), pp. 1279–1282.

[34] Adrian M Altenhoff and Christophe Dessimoz. “Inferring Orthology and Paralogy”. In: *Evolutionary Genomics: Statistical and Computational Methods, Volume 1*. Ed. by Maria Anisimova. Springer Science+Business Media, Feb. 2012, pp. 259–279.

[35] Wayne P Maddison. “Gene trees in species trees”. In: *Systematic Biology* 46.3 (1997), pp. 523–536.

[36] Jean-Philippe Doyon et al. “Models, algorithms and programs for phylogeny reconciliation”. In: *Briefings in Bioinformatics* 12.5 (Sept. 2011), pp. 392–400.

[37] Yi-Chieh Wu et al. “TreeFix: Statistically Informed Gene Tree Error Correction Using Species Trees”. In: *Systematic Biology* 62 (Nov. 2012), pp. 110–120.

[38] Oded Edelheit, Aaron Hanukoglu, and Israel Hanukoglu. “Simple and efficient site-directed mutagenesis using two single-primer reactions in parallel to generate mutants for protein structure-function studies”. In: *BMC biotechnology* 9 (2009), p. 61.

[39] Steven F Field and Mikhail V Matz. “Retracing evolution of red fluorescence in GFP-like proteins from Faviina corals”. In: *Molecular Biology and Evolution* 27.2 (Feb. 2010), pp. 225–233.

[40] Benjamin P Roscoe et al. “Analyses of the effects of all ubiquitin point mutants on yeast growth rate”. In: *Journal of Molecular Biology* 425.8 (Apr. 2013), pp. 1363–1377.

[41] Michael J Harms and Joseph W Thornton. “Analyzing protein structure and function using ancestral gene reconstruction”. In: *Current Opinion in Structural Biology* 20.3 (June 2010), pp. 360–366.

[42] Philip A Romero and Frances H Arnold. “Exploring protein fitness landscapes by directed evolution”. In: *Nature Reviews Molecular Cell Biology* 10.12 (Dec. 2009), pp. 866–876.

[43] Gregory A Petsko and Dagmar Ringe. “From Structure to Function”. In: *Protein Structure and Function*. New Science Press, Jan. 2004.

[44] Nicola Jane Mulder et al. “In silico characterization of proteins: UniProt, InterPro and Integr8”. In: *Molecular biotechnology* 38.2 (Feb. 2008), pp. 165–177.

[45] Nicholas Furnham et al. “Exploring the evolution of novel enzyme functions within structurally defined protein superfamilies”. In: *PLoS Computational Biology* 8.3 (2012), e1002403.

[46] R Chen, A Greer, and Antony M Dean. “A highly active decarboxylating dehydrogenase with rationally inverted coenzyme specificity”. In: *Proceedings of the National Academy of Sciences of the United States of America* 92.25 (Dec. 1995), pp. 11666–11670.

[47] Mark Lunzer et al. “The biochemical architecture of an ancient adaptive landscape”. In: *Science* 310.5747 (Oct. 2005), pp. 499–501.

[48] Joseph W Thornton, Eleanor Need, and David Crews. “Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling”. In: *Science* 301.5640 (Sept. 2003), pp. 1714–1717.

[49] Michael J Harms et al. “Biophysical mechanisms for large-effect mutations in the evolution of steroid hormone receptors”. In: *Proceedings of the National Academy of Sciences of the United States of America* 110.28 (July 2013), pp. 11475–11480.

[50] Daniel M Weinreich et al. “Darwinian evolution can follow only very few mutational paths to fitter proteins”. In: *Science* 312.5770 (Apr. 2006), pp. 111–114.

[51] Jesse D Bloom and Frances H Arnold. “In the light of directed evolution: pathways of adaptive protein evolution”. In: *Proceedings of the National Academy of Sciences of the United States of America* 106 Suppl 1 (June 2009), pp. 9995–10000.

[52] Patrick A Alexander et al. “A minimal sequence code for switching protein structure and function”. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.50 (Dec. 2009), pp. 21149–21154.

[53] Ruiqi Huang et al. “Enzyme functional evolution through improved catalysis of ancestrally nonpreferred substrates”. In: *Proceedings of the National Academy of Sciences of the United States of America* 109.8 (Feb. 2012), pp. 2966–2971.

[54] Karin Voordeckers et al. “Reconstruction of Ancestral Metabolic Enzymes Reveals Molecular Mechanisms Underlying Evolutionary Innovation through Gene Duplication”. In: *PLoS Biology* 10.12 (Dec. 2012), e1001446.

[55] Benjamin D Ross et al. “Stepwise evolution of essential centromere function in a Drosophila neogene”. In: *Science* 340.6137 (June 2013), pp. 1211–1214.

[56] C Tony Liu et al. “Functional significance of evolving protein sequence in dihydrofolate reductase from bacteria to humans”. In: *Proceedings of the National Academy of Sciences of the United States of America* 110.25 (June 2013), pp. 10159–10164.

[57] Jamie T Bridgham et al. “Protein evolution by molecular tinkering: diversification of the nuclear receptor superfamily from a ligand-dependent ancestor”. In: *PLoS Biology* 8.10 (2010).

[58] Jamie T Bridgham et al. “Evolution of a new function by degenerative mutation in cephalochordate steroid receptors”. In: *PLoS Genetics* 4.9 (2008), e1000191.

[59] Emily J Capra et al. “Adaptive Mutations that Prevent Crosstalk Enable the Expansion of Paralogous Signaling Protein Families”. In: *Cell* 150.1 (July 2012), pp. 222–232.

[60] A Stein and Patrick Aloy. “Contextual specificity in peptide-mediated protein interactions.” In: *PLoS ONE* 3.7 (2008), e2524.

[61] Jamie T Bridgham, Sean Michael Carroll, and Joseph W Thornton. “Evolution of hormone-receptor complexity by molecular exploitation”. In: *Science* 312.5770 (Apr. 2006), pp. 97–101.

[62] Eric A Ortlund et al. “Crystal structure of an ancient protein: evolution by conformational epistasis”. In: *Science* 317.5844 (Sept. 2007), pp. 1544–1548.

[63] Geeta N Eick et al. “Evolution of minimal specificity and promiscuity in steroid hormone receptors”. In: *PLoS Genetics* 8.11 (Nov. 2012), e1003072.

[64] Guillaume G B Tcherkez, Graham D Farquhar, and T John Andrews. “Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized”. In: *Proceedings of the National Academy of Sciences of the United States of America* 103.19 (May 2006), pp. 7246–7251.

[65] Jasper Franke et al. “Evolutionary accessibility of mutational pathways”. In: *PLoS Computational Biology* 7.8 (Aug. 2011), e1002134.

[66] Frank J Poelwijk et al. “Empirical fitness landscapes reveal accessible evolutionary paths”. In: *Nature* 445.7126 (Jan. 2007), pp. 383–386.

[67] Daniel L Hartl, D E Dykhuizen, and Antony M Dean. “Limits of adaptation: the evolution of selective neutrality”. In: *Genetics* 111.3 (Nov. 1985), pp. 655–674.

[68] Misha Soskine and Dan S Tawfik. “Mutational effects and the evolution of new protein functions”. In: *Nature Reviews Genetics* 11.8 (Aug. 2010), pp. 572–582.

[69] Nobuhiko Tokuriki et al. “How protein stability and new functions trade off”. In: *PLoS Computational Biology* 4.2 (Feb. 2008), e1000002.

[70] Xiaojun Wang, George Minasov, and Brian K Shoichet. “Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs”. In: *Journal of Molecular Biology* 320.1 (June 2002), pp. 85–95.

[71] Jamie T Bridgham, Eric A Ortlund, and Joseph W Thornton. “An epistatic ratchet constrains the direction of glucocorticoid receptor evolution”. In: *Nature* 461.7263 (Sept. 2009), pp. 515–519.

[72] Mark Lunzer, G Brian Golding, and Antony M Dean. “Pervasive cryptic epistasis in molecular evolution”. In: *PLoS Genetics* 6.10 (Oct. 2010), e1001162.

[73] Michael S Breen et al. “Epistasis as the primary factor in molecular evolution”. In: *Nature* (Oct. 2012).

[74] Merijn L M Salverda et al. “Initial mutations direct alternative pathways of protein evolution”. In: *PLoS Genetics* 7.3 (Mar. 2011), e1001321.

[75] Susumu Ohno. *Evolution by Gene Duplication*. Springer-Verlag, 1970.

[76] Amir Aharoni et al. “The ’evolvability’ of promiscuous protein functions”. In: *Nature Genetics* 37.1 (Jan. 2005), pp. 73–76.

[77] Jesse D Bloom et al. “Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution”. In: *Biology Direct* 2.1 (2007), p. 17.

[78] Eric J Hayden, Evandro Ferrada, and Andreas Wagner. “Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme”. In: *Nature* 474.7349 (June 2011), pp. 92–95.

[79] Olga Khersonsky and Dan S Tawfik. “Enzyme promiscuity: a mechanistic and evolutionary perspective”. In: *Annual Review of Biochemistry* 79 (2010), pp. 471–505.

[80] Victor Neduva and Robert B Russell. “Linear motifs: evolutionary interaction switches”. In: *FEBS Letters* 579.15 (June 2005), pp. 3342–3345.

[81] Norman E Davey, Gilles Travé, and Toby J Gibson. “How viruses hijack cell regulation”. In: *Trends in Biochemical Sciences* 36.3 (Mar. 2011), pp. 159–169.

[82] U Löhr, M Yussa, and L Pick. “Drosophila fushi tarazu: a gene on the border of homeotic function”. In: *Current Biology* 11.18 (Sept. 2001), pp. 1403–1412.

[83] Cheryl C Hsia and William McGinnis. “Evolution of transcription factor function”. In: *Current Opinion in Genetics & Development* 13.2 (Apr. 2003), pp. 199–206.

[84] Vincent J Lynch and Günter P Wagner. “Resurrecting the role of transcription factor change in developmental evolution”. In: *Evolution* 62.9 (Sept. 2008), pp. 2131–2154.

[85] Roberto Mosca, Roland A Pache, and Patrick Aloy. “The role of structural disorder in the rewiring of protein interactions through evolution”. In: *Molecular & Cellular Proteomics* 11.7 (July 2012), p. M111.014969.

[86] Tanya Vavouri et al. “Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity”. In: *Cell* 138.1 (July 2009), pp. 198–208.

[87] Gregory A Petsko and Dagmar Ringe. “Control of Protein Function”. In: *Protein Structure and Function*. New Science Press, Jan. 2004.

[88] J R Bock and D A Gough. “Predicting protein-protein interactions from primary structure”. In: *Bioinformatics* 17.5 (May 2001), pp. 455–460.

[89] Thomas R Bürglin. “The Hedgehog protein family”. In: *Genome Biology* 9.11 (2008), p. 241.

[90] Maja Adamska et al. “The evolutionary origin of hedgehog proteins”. In: *Current Biology* 17.19 (Oct. 2007), R836–7.

[91] Lia Rosso et al. “Birth and rapid subcellular adaptation of a hominoid-specific CDC14 protein”. In: *PLoS Biology* 6.6 (June 2008), e140.

[92] Birgit Pils and Jörg Schultz. “Inactive enzyme-homologues find new function in regulatory processes”. In: *Journal of Molecular Biology* 340.3 (July 2004), pp. 399–404.

[93] Annabel E Todd, Christine A Orengo, and Janet M Thornton. “Sequence and structural differences between enzyme and nonenzyme homologs”. In: *Structure* 10.10 (Oct. 2002), pp. 1435–1451.

[94] Micheline N Ngaki et al. “Evolution of the chalcone-isomerase fold from fatty-acid binding to stereospecific catalysis”. In: *Nature* 485.7399 (May 2012), pp. 530–533.

[95] Meimei Xu, P Ross Wilderman, and Reuben J Peters. “Following evolution’s lead to a single residue switch for diterpene synthase product outcome”. In: *Proceedings of the National Academy of Sciences of the United States of America* 104.18 (May 2007), pp. 7397–7401.

[96] Bryan T Greenhagen et al. “Identifying and manipulating structural determinates linking catalytic specificities in terpene synthases”. In: *Proceedings of the National Academy of Sciences of the United States of America* 103.26 (June 2006), pp. 9826–9831.

[97] Paul E O’Maille et al. “Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases”. In: *Nature chemical biology* 4.10 (Oct. 2008), pp. 617–623.

[98] P J O’Brien and D Herschlag. “Catalytic promiscuity and the evolution of new enzymatic activities”. In: *Chemistry & biology* 6.4 (Apr. 1999), R91–R105.

[99] J L Seffernick et al. “Melamine deaminase and atrazine chlorohydrolase: 98 percent identical but functionally different”. In: *Journal of Bacteriology* 183.8 (Apr. 2001), pp. 2405–2410.

[100] Sajid Noor et al. “Intramolecular epistasis and the evolution of a new enzymatic function”. In: *PLoS ONE* 7.6 (2012), e39822.

[101] Guillermo de Cárcer, Gerard Manning, and Marcos Malumbres. “From Plk1 to Plk5—Functional evolution of polo-like kinases”. In: *Cell Cycle* 10.14 (July 2011), pp. 2255–2262.

[102] Guillermo de Cárcer et al. “Plk5, a polo box domain-only protein with specific roles in neuron differentiation and glioblastoma suppression”. In: *Molecular and Cellular Biology* 31.6 (Mar. 2011), pp. 1225–1239.

[103] Christopher A Johnston et al. “Conversion of the enzyme guanylate kinase into a mitotic-spindle orienting protein by a single mutation that inhibits GMP-induced closing”. In: *Proceedings of the National Academy of Sciences of the United States of America* 108.44 (Nov. 2011), E973–8.

[104] Colin Adrain and Matthew Freeman. “New lives for old: evolution of pseudoenzyme function illustrated by iRhoms”. In: *Nature Reviews Molecular Cell Biology* 13.8 (Aug. 2012), pp. 489–498.

[105] Konark Mukherjee et al. “CASK Functions as a Mg2+-independent neurexin kinase”. In: *Cell* 133.2 (Apr. 2008), pp. 328–339.

[106] Pier Federico Gherardini et al. “Convergent Evolution of Enzyme Active Sites Is not a Rare Phenomenon”. In: *Journal of Molecular Biology* 372.3 (Sept. 2007), pp. 817–845.

[107] Marina V Omelchenko et al. “Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution”. In: *Biology Direct* 5 (2010), p. 31.

[108] Daphne H E W Huberts and Ida J van der Klei. “Moonlighting proteins: an intriguing mode of multitasking”. In: *Biochimica Et Biophysica Acta* 1803.4 (Apr. 2010), pp. 520–525.

[109] Bin Zhao et al. “Crystal structure of albaflavenone monooxygenase containing a moonlighting terpene synthase active site”. In: *The Journal of biological chemistry* 284.52 (Dec. 2009), pp. 36711–36719.

[110] Shana L Geffeney et al. “Evolutionary diversification of TTX-resistant sodium channels in a predator-prey interaction”. In: *Nature* 434.7034 (Apr. 2005), pp. 759–763.

[111] Wallace F Marshall. “Cellular length control systems”. In: *Annual Review of Cell and Developmental Biology* 20 (2004), pp. 677–693.

[112] Laure Journet et al. “The needle length of bacterial injectisomes is determined by a molecular ruler”. In: *Science* 302.5651 (Dec. 2003), pp. 1757–1760.

[113] Luís Jaime Mota et al. “Bacterial injectisomes: needle length does matter”. In: *Science* 307.5713 (Feb. 2005), p. 1278.

[114] Stefanie Wagner et al. “The helical content of the YscP molecular ruler determines the length of the Yersinia injectisome”. In: *Molecular Microbiology* 71.3 (Feb. 2009), pp. 692–701.

[115] S Eriksson, R Hurme, and M Rhen. “Low-temperature sensors in bacteria”. In: *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences* 357.1423 (July 2002), pp. 887–893.

[116] Ann Kathrin Heroven et al. “Regulatory elements implicated in the environmental control of invasin expression in enteropathogenic Yersinia”. In: *Advances in experimental medicine and biology* 603 (2007), pp. 156–166.

[117] Katharina Herbst et al. “Intrinsic thermal sensing controls proteolysis of Yersinia virulence regulator RovA”. In: *PLoS Pathogens* 5.5 (May 2009), e1000435.

[118] Nick Quade et al. “Structural basis for intrinsic thermosensing by the master virulence regulator RovA of Yersinia”. In: *The Journal of biological chemistry* 287.43 (Oct. 2012), pp. 35796–35803.

[119] P G Allen et al. “Phalloidin binding and rheological differences among actin isoforms”. In: *Biochemistry* 35.45 (Nov. 1996), pp. 14062–14069.

[120] D H Wachsstock, W H Schwarz, and T D Pollard. “Cross-linker dynamics determine the mechanical properties of actin gels”. In: *Biophysical Journal* 66.3 (Mar. 1994), pp. 801–809.

[121] B Wagner et al. “Cytoskeletal polymer networks: the molecular structure of cross-linkers determines macroscopic properties”. In: *Proceedings of the National Academy of Sciences of the United States of America* 103.38 (Sept. 2006), pp. 13974–13978.

[122] Vera Dugina et al. “Actin”. In: *Cytoskeleton and Human Disease*. Ed. by M Kavallaris. Springer Science+Business Media, Feb. 2012, pp. 1–29.

[123] Macarena Marín, Vladimir N Uversky, and Thomas Ott. “Intrinsic disorder in pathogen effectors: protein flexibility as an evolutionary hallmark in a molecular arms race”. In: *The Plant Cell* 25.9 (Sept. 2013), pp. 3153–3157.

[124] Joanne Young, Julie Ménétrey, and Bruno Goud. “RAB6C is a retrogene that encodes a centrosomal protein involved in cell cycle progression”. In: *Journal of Molecular Biology* 397.1 (Mar. 2010), pp. 69–88.

[125] J V Chamary, Joanna L Parmley, and Laurence D Hurst. “Hearing silence: non-neutral evolution at synonymous sites in mammals”. In: *Nature Reviews Genetics* 7.2 (Feb. 2006), pp. 98–108.

[126] Federico Abascal et al. “Subfunctionalization via Adaptive Evolution Influenced by Genomic Context: The Case of Histone Chaperones ASF1a and ASF1b”. In: *Molecular Biology and Evolution* 30.8 (Aug. 2013), pp. 1853–1866.

[127] Sankaran Sandhya et al. “CUSP: an algorithm to distinguish structurally conserved and unconserved regions in protein domain alignments and its application in the study of large length variations”. In: *BMC Structural Biology* 8 (2008), p. 28.

[128] Reed A Cartwright. “Problems and solutions for estimating indel rates and length distributions”. In: *Molecular Biology and Evolution* 26.2 (Feb. 2009), pp. 473–480.

[129] Jian-Qun Chen et al. “Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria”. In: *Molecular Biology and Evolution* 26.7 (July 2009), pp. 1523–1531.

[130] Vaishali Katju and Michael Lynch. “The structure and early evolution of recently arisen gene duplicates in the Caenorhabditis elegans genome”. In: *Genetics* 165.4 (Dec. 2003), pp. 1793–1803.

[131] Rita Gemayel et al. “Variable tandem repeats accelerate evolution of coding and regulatory sequences”. In: *Annual Review of Genetics* 44 (2010), pp. 445–477.

[132] Kevin J Verstrepen et al. “Intragenic tandem repeats generate functional variability”. In: *Nature Genetics* 37.9 (Sept. 2005), pp. 986–990.

[133] Saul M Honigberg. “Cell signals, cell contacts, and the organization of yeast communities”. In: *Eukaryotic Cell* 10.4 (Apr. 2011), pp. 466–473.

[134] Xiang Gao and Michael Lynch. “Ubiquitous internal gene duplication and intron creation in eukaryotes”. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.49 (Dec. 2009), pp. 20818–20823.

[135] Michael Lynch and John S Conery. “The evolutionary fate and consequences of duplicate genes”. In: *Science* 290.5494 (Nov. 2000), pp. 1151–1155.

[136] Brian P Cusack and Kenneth H Wolfe. “Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates”. In: *Molecular Biology and Evolution* 24.3 (Mar. 2007), pp. 679–686.

[137] Chris A Brown, Andrew W Murray, and Kevin J Verstrepen. “Rapid expansion and functional divergence of subtelomeric gene families in yeasts”. In: *Current Biology* 20.10 (May 2010), pp. 895–903.

[138] Joaquin F Christiaens et al. “Functional divergence of gene duplicates through ectopic recombination”. In: *EMBO Reports* 13.12 (Nov. 2012), pp. 1145–1151.

[139] D Allan Drummond et al. “On the conservative nature of intragenic recombination”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.15 (Apr. 2005), pp. 5380–5385.

[140] Marco Landwehr et al. “Diversification of catalytic function in a synthetic family of chimeric cytochrome p450s”. In: *Chemistry & biology* 14.3 (Mar. 2007), pp. 269–278.

[141] Yougen Li et al. “A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments”. In: *Nature Biotechnology* 25.9 (Sept. 2007), pp. 1051–1056.

[142] Frances H Arnold. “How proteins adapt: lessons from directed evolution”. In: *Cold Spring Harbor Symposia on Quantitative Biology* 74 (2009), pp. 41–46.

[143] Anton J Enright et al. “Protein interaction maps for complete genomes based on gene fusion events”. In: *Nature* 402.6757 (Nov. 1999), pp. 86–90.

[144] Corbin D Jones and David J Begun. “Parallel evolution of chimeric fusion genes”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.32 (Aug. 2005), pp. 11373–11378.

[145] Gurkan Guntas et al. “Directed evolution of protein switches and their application to the creation of ligand-binding proteins”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.32 (Aug. 2005), pp. 11224–11229.

[146] Ronald A Fisher. “The Genetical Theory of Natural Selection”. In: *Oxford University Press* (Nov. 1930), pp. 1–308.

[147] Theodore Garland and P A Carter. “Evolutionary Physiology”. In: *Annual review of physiology* 56 (1994), pp. 579–621.

[148] Martin E Feder, Albert F Bennett, and Raymond B Huey. “Evolutionary Physiology”. In: *Annual Review of Ecology and Systematics* 31 (Apr. 2000), pp. 315–341.

[149] Adam Eyre-Walker and Peter D Keightley. “The distribution of fitness effects of new mutations”. In: *Nature Reviews Genetics* 8.8 (Aug. 2007), pp. 610–618.

[150] Ziheng Yang. “The power of phylogenetic comparison in revealing protein function”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.9 (Feb. 2005), pp. 3179–3180.

[151] Adam Siepel et al. “Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes”. In: *Genome Research* 15.8 (Aug. 2005), pp. 1034–1050.

[152] Romain A Studer and Marc Robinson-Rechavi. “How confident can we be that orthologs are similar, but paralogs differ?” In: *Trends in Genetics : TIG* 25.5 (May 2009), pp. 210–216.

[153] Walid H Gharib and Marc Robinson-Rechavi. “When orthologs diverge between human and mouse”. In: *Briefings in Bioinformatics* 12.5 (Sept. 2011), pp. 436–441.

[154] Emily S W Wong et al. “A limited role for gene duplications in the evolution of platypus venom”. In: *Molecular Biology and Evolution* 29.1 (2012), pp. 167–177.

[155] Jian-Min Chen et al. “Gene conversion: mechanisms, evolution and human disease”. In: *Nature Reviews Genetics* 8.10 (Oct. 2007), pp. 762–775.

[156] Masatoshi Nei and Alejandro P Rooney. “Concerted and birth-and-death evolution of multigene families”. In: *Annual Review of Genetics* 39 (2005), pp. 121–152.

[157] John S Taylor and Jeroen Raes. “Duplication and Divergence: The Evolution of New Genes and Old Ideas”. In: *Annual Review of Genetics* 38.1 (Dec. 2004), pp. 615–643.

[158] Nels C Elde et al. “Poxviruses Deploy Genomic Accordions to Adapt Rapidly against Host Antiviral Defenses”. In: *Cell* 150.4 (Aug. 2012), pp. 831–841.

[159] M Pilar Francino. “An adaptive radiation model for the origin of new gene functions”. In: *Nature Genetics* 37.6 (June 2005), pp. 573–577.

[160] Gavin C Conant and Kenneth H Wolfe. “Turning a hobby into a job: how duplicated genes find new functions”. In: *Nature Reviews Genetics* 9.12 (Dec. 2008), pp. 938–950.

[161] Ulfar Bergthorsson, Dan I Andersson, and John R Roth. “Ohno’s dilemma: evolution of new genes under continuous selection”. In: *Proceedings of the National Academy of Sciences of the United States of America* 104.43 (Oct. 2007), pp. 17004–17009.

[162] Stephen Jay Gould and Richard C Lewontin. “The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme”. In: *Proceedings of the Royal Society B: Biological Sciences* 205.1161 (Sept. 1979), pp. 581–598.

[163] Arlin Stoltzfus. “On the possibility of constructive neutral evolution”. In: *Journal of Molecular Evolution* 49.2 (Aug. 1999), pp. 169–181.

[164] A Force et al. “Preservation of duplicate genes by complementary, degenerative mutations”. In: *Genetica* 151.4 (Apr. 1999), pp. 1531–1545.

[165] Chris Todd Hittinger and Sean B Carroll. “Gene duplication and the adaptive evolution of a classic genetic switch”. In: *Nature* 449.7163 (Oct. 2007), pp. 677–681.

[166] Cheng Deng et al. “Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict”. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.50 (Dec. 2010), pp. 21593–21598.

[167] Gregory C Finnigan et al. “Evolution of increased complexity in a molecular machine”. In: *Nature* 481.7381 (Jan. 2012), pp. 360–364.

[168] Christopher R Baker, Victor Hanson-Smith, and Alexander D Johnson. “Following gene duplication, paralog interference constrains transcriptional circuit evolution”. In: *Science* 342.6154 (Oct. 2013), pp. 104–108.

[169] Vincent J Lynch, Gemma May, and Günter P Wagner. “Regulatory evolution through divergence of a phosphoswitch in the transcription factor CEBPB”. In: *Nature* 480.7377 (Dec. 2011), pp. 383–386.

[170] Jesus Garcia et al. “A conformational switch in the Piccolo C2A domain regulated by alternative splicing”. In: *Nature Structural & Molecular Biology* 11.1 (Jan. 2004), pp. 45–53.

[171] Detlef D Leipe et al. “Classification and evolution of P-loop GTPases and related ATPases”. In: *Journal of Molecular Biology* 317.1 (Mar. 2002), pp. 41–72.

[172] P L Hartzell. “Complementation of sporulation and motility defects in a prokaryote by a eukaryotic GTPase”. In: *Proceedings of the National Academy of Sciences of the United States of America* 94.18 (Sept. 1997), pp. 9881–9886.

[173] Krister Wennerberg, Kent L Rossman, and Channing J Der. “The Ras Superfamily at a Glance”. In: *Journal of Cell Science* 118.5 (Mar. 2005), pp. 843–846.

[174] Jane M Carlton et al. “Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis”. In: *Science* 315.5809 (Jan. 2007), pp. 207–212.

[175] D DeFeo et al. “Analysis of two divergent rat genomic clones homologous to the transforming gene of Harvey murine sarcoma virus”. In: *Proceedings of the National Academy of Sciences of the United States of America* 78.6 (June 1981), pp. 3328–3332.

[176] Antoine E Karnoub and Robert A Weinberg. “Ras oncogenes: split personalities”. In: *Nature Reviews Molecular Cell Biology* 9.7 (July 2008), pp. 517–531.

[177] D Gallwitz, C Donath, and C Sander. “A yeast gene encoding a protein homologous to the human c-has/bas proto-oncogene product”. In: *Nature* 306.5944 (Dec. 1983), pp. 704–707.

[178] H D Schmitt et al. “The ras-related YPT1 gene product in yeast: a GTP-binding protein that might be involved in microtubule organization”. In: *Cell* 47.3 (Nov. 1986), pp. 401–412.

[179] A Salminen and P J Novick. “A ras-like protein is required for a post-Golgi event in yeast secretion”. In: *Cell* 49.4 (May 1987), pp. 527–538.

[180] Bruno Goud et al. “A GTP-binding protein required for secretion rapidly associates with secretory vesicles and the plasma membrane in yeast”. In: *Cell* 53.5 (June 1988), pp. 753–768.

[181] H Haubruck et al. “The ras-related ypt protein is an ubiquitous eukaryotic protein: isolation and sequence analysis of mouse cDNA clones highly homologous to the yeast YPT1 gene”. In: *The EMBO Journal* 6.13 (Dec. 1987), pp. 4049–4053.

[182] N Touchot, P Chardin, and A Tavitian. “Four additional members of the ras gene superfamily isolated by an oligonucleotide strategy: molecular cloning of YPT-related cDNAs from a rat brain library”. In: *Proceedings of the National Academy of Sciences of the United States of America* 84.23 (Dec. 1987), pp. 8210–8214.

[183] H Haubruck et al. “The ras-related mouse ypt1 protein can functionally replace the YPT1 gene product in yeast”. In: *The EMBO Journal* 8.5 (May 1989), pp. 1427–1432.

[184] Ari Löytynoja and Nick Goldman. “An algorithm for progressive multiple alignment of sequences with insertions”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.30 (July 2005), pp. 10557–10562.

[185] José B Pereira-Leal and Miguel C Seabra. “The mammalian Rab family of small GTPases: definition of family and subfamily sequence motifs suggests a mechanism for functional specificity in the Ras superfamily”. In: *Journal of Molecular Biology* 301.4 (Aug. 2000), pp. 1077–1087.

[186] Ian Moore, J Schell, and K Palme. “Subclass-specific sequence motifs identified in Rab GTPases”. In: *Trends in Biochemical Sciences* 20.1 (Jan. 1995), pp. 10–12.

[187] Meng-Tse Gabe Lee, Ashwini Mishra, and David G Lambright. “Structural mechanisms for regulation of membrane traffic by rab GTPases”. In: *Traffic* 10.10 (Oct. 2009), pp. 1377–1389.

[188] Alex H Hutagalung and Peter J Novick. “Role of Rab GTPases in membrane traffic and cell physiology”. In: *Physiological Reviews* 91.1 (2011), pp. 119–149.

[189] Alex Stivala et al. “Automatic generation of protein structure cartoons with Pro-origami”. In: *Bioinformatics* 27.23 (Dec. 2011), pp. 3315–3316.

[190] Sudharshan Eathiraj et al. “Structural basis of family-wide Rab GTPase recognition by rabenosyn-5”. In: *Nature* 436.7049 (July 2005), pp. 415–419.

[191] J B Pereira-Leal and M C Seabra. “Evolution of the Rab family of small GTP-binding proteins”. In: *Journal of Molecular Biology* 313.4 (Nov. 2001), pp. 889–901.

[192] Michiko Shirane and Keiichi I Nakayama. “Protrudin induces neurite formation by directional membrane trafficking”. In: *Science* 314.5800 (Nov. 2006), pp. 818–821.

[193] Dikran Aivazian, Ramon L Serrano, and Suzanne R Pfeffer. “TIP47 is a key effector for Rab9 localization”. In: *The Journal of Cell Biology* 173.6 (June 2006), pp. 917–926.

[194] P Chavrier et al. “Hypervariable C-terminal domain of rab proteins acts as a targeting signal”. In: *Nature* 353.6346 (Oct. 1991), pp. 769–772.

[195] Bassam R Ali et al. “Multiple regions contribute to membrane targeting of Rab GTPases”. In: *Journal of Cell Science* 117.Pt 26 (Dec. 2004), pp. 6401–6412.

[196] Harald Stenmark. “Rab GTPases as coordinators of vesicle traffic”. In: *Nature Reviews Molecular Cell Biology* 10.8 (Aug. 2009), pp. 513–525.

[197] Andrew D C Maccoll. “The ecological causes of evolution”. In: *Trends in Ecology and Evolution* 26.10 (Oct. 2011), pp. 514–522.

[198] Rudy Behnia and Sean Munro. “Organelle identity and the signposts for membrane traffic”. In: *Nature* 438.7068 (Dec. 2005), pp. 597–604.

This chapter has been published as: Yoan Diekmann, Elsa Seixas, Marc Gouw, Filipe Tavares-Cadete, Miguel C Seabra, and José B Pereira-Leal. "Thousands of Rab GTPases for the Cell Biologist". In: *PLoS Computational Biology* 7.10 (Oct. 2011), e1002217.

Author contribution: In accordance with the author contributions stated in the reference above, I participated in conceiving and designing the experiments together with the senior authors Miguel C. Seabra (Instituto Gulbenkian de Ciência and Centro de Estudos de Doenças Chrónicas), and José B. Pereira-Leal (Instituto Gulbenkian de Ciência), performed most of the experiments with exception of the gene expression experiments reported in Figure 2.8 done by Elsa Seixas (Instituto Gulbenkian de Ciência and Centro de Estudos de Doenças Chrónicas) and the website shown in Figure 2.3 implemented by Marc Gouw (Instituto Gulbenkian de Ciência), I analysed the data, and wrote the paper together with José B. Pereira-Leal.

Chapter 2

Evolutionary patterns of the Rab family of small GTPases

"[...] the study of pattern must be divorced as much as possible from the study of process, to provide an unbiased baseline for the evaluation of alternative hypotheses about process." [1, p. 4f]

—Eldredge, Cracraft, 1980

Abstract Chapter 2

Rab proteins are small GTPases that act as essential regulators of vesicular trafficking. 44 subfamilies are known in humans, performing specific sets of functions at distinct subcellular localisations and tissues. Rab function is conserved even amongst distant orthologs. Hence, the annotation of Rabs yields functional predictions about the cell biology of trafficking. So far, annotating Rabs has been a laborious manual task not feasible for the genomic output of deep sequencing technologies. We developed, validated and benchmarked the Rabifier, an automated bioinformatic pipeline for the identification and classification of Rabs, which achieves up to 90% accuracy. We cataloged ∼8000 Rabs from 247 genomes covering the entire eukaryotic tree. The full Rab database and a web tool implementing the pipeline are publicly available at `www.RabDB.org`. For the first time, we describe and analyse the evolution of Rabs over the whole eukaryotic phylogeny. We found a highly dynamic family undergoing frequent taxon-specific expansions and losses. We dated the origin of human subfamilies using phylogenetic profiling, which enlarged the Rab repertoire of the eukaryotic ancestor with Rab14, 32 and L4. A detailed analysis of the Choanoflagellate *M. brevicollis* Rab family pinpointed the changes that accompanied animal multicellularity, mainly an expansion and specialisation of the secretory pathway. Lastly, we experimentally establish tissue specificity of mouse Rabs and suggest that neo-functionalisation best explains the emergence of new Rab subfamilies. The Rabifier and RabDB allow non-bioinformaticians to integrate thousands of Rabs in their analyses. They are designed for the cell biology community to keep pace with the increasing number of genomes and change the scale at which we perform comparative analysis in cell biology.

2.1 Introduction

Intracellular compartmentalisation is found in all cellular lifeforms, yet eukaryotes have evolved extensive membranous compartments unique to this domain of life. Protein trafficking pathways accomplish the movement of cellular components like proteins and lipids between the cellular compartments. These essential pathways play house-keeping roles, such as transport of proteins destined for secretion to the plasma membrane via the secretory pathway, or recycling of membrane receptors via the endocytic pathway. In addition, they play a variety of specialised roles, such as bone resorption in osteoclasts, pigmentation in melanocytes and antigen presentation in immune cells. Malfunction of protein trafficking components leads to a large number of human diseases, ranging from hemorrhagic disorders and immunodeficiencies to mental retardation and blindness [2–5], as well as cancer [6–9]. Furthermore, protein trafficking pathways are frequently exploited by human pathogens to gain entry and survive within host cells [10–13].

The endomembrane system accounts for a large fraction of the protein coding sequences in eukaryotic genomes [14], and a plethora of data on molecules and interactions in different model organisms is available. However, it is unclear how these data map across organisms, and how general the mechanisms characterised in single species are. To answer these question we need to understand the evolution of the protein trafficking pathways and organelles. An evolutionary framework for protein trafficking is particularly important given the overwhelming accumulation of genomes, many from pathogenic organisms. Their comparative analysis can distinguish conserved from taxon-specific machineries, with clear practical applications. For example, conservation of genes led to the discovery of novel components and mechanisms in ciliogenesis [15], whereas the presence of taxon-specific pathways allowed the identifica-

tion of Fosmidomycin as a potential antimalarial drug [16]. Studying the evolution of protein trafficking is essential to understand the origins of eukaryotes. Comparative genomics and phylogenetics have established that the Last Eukaryotic Common Ancestor (LECA) already had a complex membrane trafficking system [17] including most types of extant molecular components [18]. These are believed to have expanded by duplication and specialisation giving rise to the full diversity of organelles and trafficking pathways observed today (see [17] for a detailed description of this evolutionary scenario).

Rabs are central regulators of protein trafficking. They are small GTPases that work as molecular switches to regulate vesicle budding, motility, tethering and fusion steps in vesicular transport [19]. Most recently the authors of [20] also linked Rabs to membrane fission. They recruit molecular motors to organelles and transport-vesicles, coordinate intracellular signalling with membrane trafficking, organise distinct subdomains within membranous organelles and play a critical role in the definition of organelle identity (recently reviewed in reference [21]). Rab subfamilies localise to distinct cellular locations, and regulate trafficking in a pathway-, organelle- and tissue-specific manner. This makes them ideal markers for the majority of trafficking-processes and compartments. Among trafficking-associated proteins, the Rab family expanded most in evolution [17, 22], suggesting that it provided the primary diversification element in the evolution of trafficking [22]. An important feature of the Rab family is that Rab orthologs tend to perform similar functions even in divergent taxa. For example, the mouse Rab1 has been shown to be able to functionally replace its ortholog YPT1 in yeast [23]. Hence assigning a Rab to a known and functionally described subfamily, *e.g.* Rab1, is a strong functional prediction, *i.e.* functioning in the early secretory pathway in the case of Rab1. Together with the ability to classify them into subfamilies based on sequence alone, this allows to establish the presence

or loss of pathways and organelles solely based on the annotation of the Rab repertoire—a procedure we subsequently refer to as Rab profiling.

Previously, we defined criteria to identify and classify Rab proteins [24], which have been used as a basis for detailed manual analysis of the Rab families in a variety of organisms [24–32]. However, manual identification of Rab repertoires is tedious and time-consuming and not compatible with the deluge of fully sequenced eukaryotic genomes that new sequencing technologies are generating. We thus need to develop methods that enable the automated annotation of Rab proteins. Several characteristics of the Rab family make this a challenging bioinformatics problem. First, there is a strong non-specific signal from GTPase motifs spread throughout the protein sequence [33], which makes it hard to distinguish Rabs from other small GTPases. Second, the Rab family is large due to extensive duplication in several branches of the eukaryotic tree (*e.g.* [27, 28]). Together with high sequence similarity amongst Rabs this causes difficulties to correctly classify Rabs into subfamilies and to further discern yet unseen subfamilies. Lastly, any automated scheme has to respect and perpetuate as much as possible the current naming conventions, despite any inconsistencies stemming from the decentralised nature of scientific discovery and the huge bias of existing annotations towards Opisthokonts. This requires a flexible, learning scheme both able to cope with the contingency of the field and to easily incorporate new naming consensuses.

Here, we overcame these problems and developed an automated bioinformatic pipeline for the identification and classification of Rabs. We termed our pipeline the 'Rabifier', which we describe, validate and benchmark. Using our tool, we cataloged nearly 8.000 Rabs from 247 genomes covering the major taxa of the eukaryotic tree, which we make available along with our pipeline at `www.RabDB.org`.

Based on this comprehensive dataset of Rab proteins, we describe and analyse the evolution of Rabs. We found a highly dynamic family under-

going frequent taxon-specific expansions and losses. We extend the Rab repertoire previously reported to have been present in the LECA, identify the changes in the Rab family that accompanied the emergence of multicellularity and show that neofunctionalisation best explains the emergence of new human Rab subfamilies.

2.2 Results / Discussion

2.2.1 The Rabifier

We implemented a bioinformatics pipeline to identify and classify Rab GTPases in any set of protein sequences independently of taxonomical information, which we term 'Rabifier'. The Rabifier proceeds in two major phases, which are schematised in Figure 2.1. First, it decides whether a protein sequence belongs to the Rab family, *i.e.* that it is not a Ras, a Rho, etc., and in the second phase it classifies the predicted Rab sequence into a Rab subfamily (*e.g.* Rab1). We describe the rationale for this procedure below—technical details are given in Sections 2.4 and 2.A.

Phase 1 (Figure 2.1A), which classifies protein sequences to the Rab family, proceeds in three stages. First, we check that the protein has a G-protein family domain. As the presence of such a domain can be decided with near certainty, this step drastically reduces the number of candidate Rabs while not excluding any real Rab. In order to do so, we align the sequence against a profile Hidden Markov Model (HMMs) [38] describing the known GTPase structures, as provided by the Superfamily database [39]. Secondly, we search for local sequence similarity by performing a BLASTp [35] query against an internal reference set of manually curated GTPases and discard the protein if it is most similar to a GTPase other than a Rab. At this stage of the workflow, the majority of non-Rab sequences has already been rejected (see Figure 2.1C, where the number of sequences

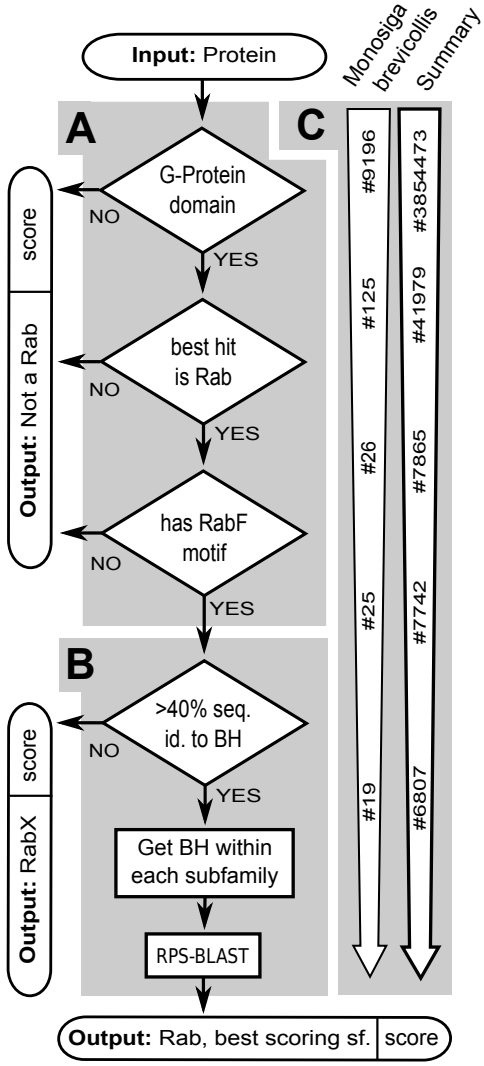


FIGURE 2.1: *Flowchart of the Rabifier*—(A) Identification- and (B) classification-procedure implemented by the Rabifier, see Section 2.2 for details on the two phases. Panel (C) shows descriptive statistics from the application of the Rabifier to 247 genomes in the Superfamily database [34], and details about *M. brevicollis*. *Abbreviations:* best BLAST hit (BH) [35], Rab family motif (RabF) [36], reverse Ψ-BLAST (RPS-BLAST) [37], subfamily (sf.), Rab not classified to any subfamily within our internal reference set (RabX)

that transition between these phases is shown for *M. brevicollis* and for a database of 247 genomes described below). However, small GTPases are so similar to each other that a residual amount of false positives still remains undetected. We remove them in the third stage, where we scan the sequence for the presence of at least one of five characteristic RabF motifs defined in reference [36]. If no motif is found, it is concluded that the protein cannot be a Rab and rejected. Remaining sequences are all assigned to the Rab family at an individual confidence level computed for each Rab. The confidence score is derived from the combination of the individual statistics generated by the three stages according to a procedure described in Text S1.

The second phase (Figure 2.1B) proposes a classification into one of the Rab subfamilies present in our internal reference set, or suggests no similarity to any of those. It proceeds in two stages. First, we test whether the Rab respects a 40% identity cut-off to its BH that prevents assignment of too disparate sequences to any of the pre-defined subfamilies. If the cut-off is met, a classification is proposed, if not, the Rab is classified as belonging to the undetermined subfamily RabX. The use of a 40% threshold is supported in Figure 2.13, and has previously been employed for example in reference [29]. The actual subfamily classification is based on the computation of a likelihood score for each of the subfamilies in our reference set. Intuitively, the protein is classified as belonging to the highest scoring subfamily, however, all scores are kept and thus provide an estimate of the relative uncertainty associated with each call. Like the Rab family score generated in the first phase of the Rabifier, the computation integrates output statistics from different tools, namely from local alignments via BLAST and from alignments using reverse Ψ-BLAST (RPS-BLAST [37]). Similar to HMMs, RPS-BLAST compares a sequence against a summary of a set of sequences, in our case summaries of all sequences in our reference set belonging to a single Rab subfamily, and measures how likely the in-

put belongs to any the subfamilies. This way we take information from all sequences in the internal reference set into account. For details on the procedure check Section 2.4 and Supplementary Methods Text S1.

2.2.2 Validation of the Rabifier classifications and design

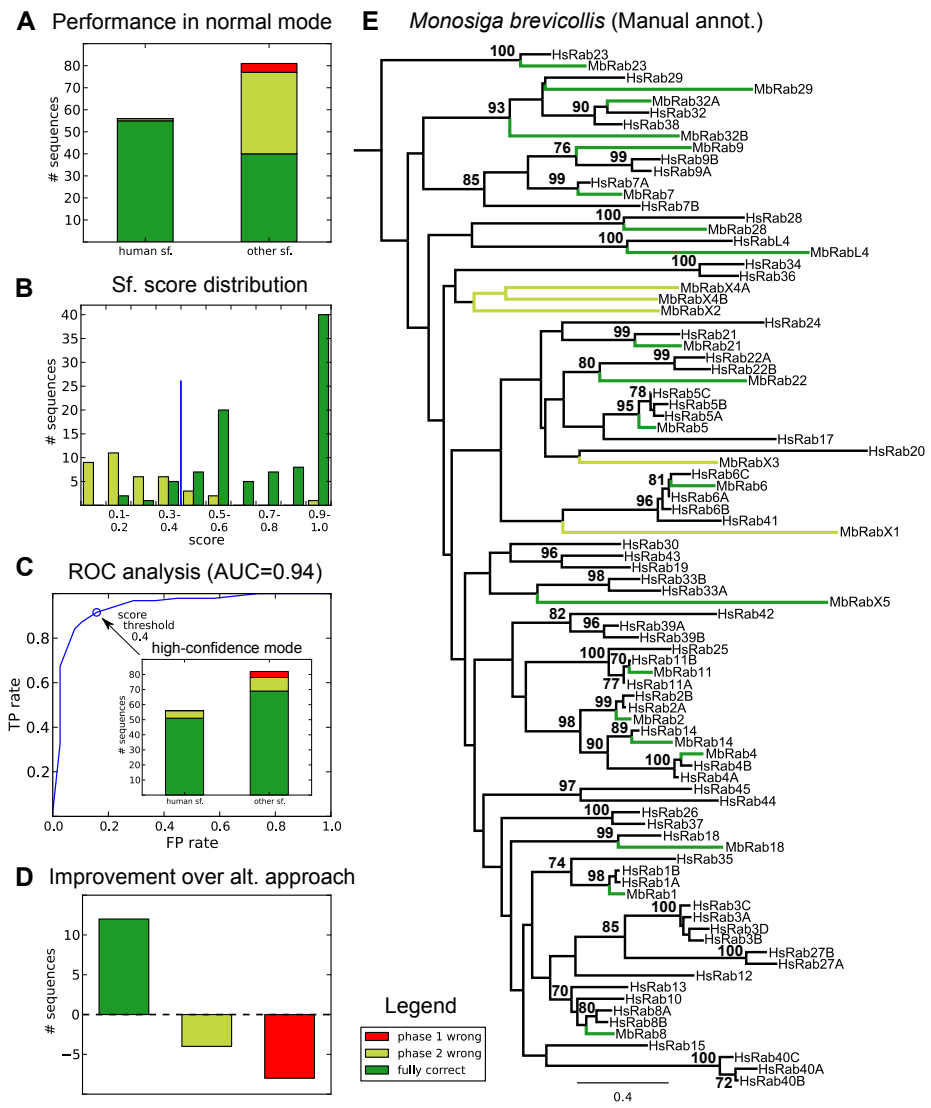
Any new methodology has to be validated. Ideally this is based on a test data set fulfilling three requirements: the test data is correctly and comprehensively annotated with those features the tool automatically detects, it is large enough to provide robust statistics, and it covers the entire range of possible inputs the tool might encounter in its real-world application, at best even respecting the expected proportions of worst- to best-case inputs. In our case, no dataset is available which fulfils the three requirements simultaneously: Rab repertoires are only available for a limited number of organisms which are not evenly distributed across eukaryotic phylogeny, and whose annotation was manually performed by different groups, hence may be inconsistent or even incorrect (in some cases a 'correct', *i.e.* consensual, classification might not even exist).

In the absence of a suitable validation dataset, we opted to validate the Rabifier against the manually curated Rab families of three organisms representing distinct worst case scenarios for the Rabifier (Figure 2.2A-C, see Table S1 for a list of all sequences used). This ensures that the validation is meaningful, as it provides a strict lower bound on the expected performance in every day use. First, we chose the Excavate *Trypanosoma brucei* [31], which is one of the most distantly related organism to our reference sequences, which are dominated by Opisthokonts (an unranked scientific classification sometimes also called 'Fungi/Metazoa group'). The second is *Entamoeba histolytica* [29], a Unikont from the phylum of Amoebozoa that is thus marginally closer to the sequences that dominate our reference database, but has a heavily expanded and diverse Rab repertoire which makes it challenging to assign Rab subfamilies. The third organism,

Monosiga brevicollis from the class of Choanoflagellates, was chosen as a representative of a phylum (Choanozoa) for which no information on the Rab family is available yet. In this third case, we compare the automated predictions against a manual analysis we performed in this study (Figure 2.2E), and which we will discuss below.

The first aspect we assessed is the ability of the Rabifier to distinguish Rabs from other GTPases (summarised in Figure 2.2A). We present the Rabifier with the set of GTPases from the above organisms and count how often we miss a Rab (false negative—FN), and how often we incorrectly classify a non-Rab as a Rab (false positive—FP). For *T. brucei*, we correctly classified 101 out of 102 GTPases as being a Rab or not, 292 out of 295 in *E. histolytica* and finally all 125 GTPases in *M. brevicollis*. Altogether, we have no FP and 4 FN, which means that for this particular set of genomes we make correct decisions about whether a protein is a Rab in 99.2% of the cases with no differences amongst the organisms. In order to understand the sources of the misannotations at family level, we inspected the false negatives individually. The Rabifier disagrees with the manual curation of [31] in *T. brucei* for TbRabX3, a RabL2-like protein, that is counted as a false negative. We explicitly added RabL2 sequences to our negative data set as we do not consider these proteins as members of the Rab family (see section 2.4). The remaining disagreements between the Rabifier and the manual annotations are three false negative proteins in *E. histolytica* in which we cannot find any detectable RabF motif, and one protein which has no similarity to any member of our reference dataset of small GTPases. We conclude that these proteins are likely misclassified in reference [29], and hence that the above failures of the Rabifier to identify Rabs are artificially introduced by our validation procedure.

Secondly, we established the accuracy by which a given Rab sequence is assigned to the right subfamily (summarised in Figure 2.2A). Concretely, for those sequences which were correctly identified as Rabs, we checked


A Performance in normal mode
sequences
human sf.
other sf.
B Sf. score distribution
sequences
score
C ROC analysis (AUC=0.94)
score threshold 0.4
high-confidence mode
TP rate
FP rate
D Improvement over alt. approach
sequences
E Monosiga brevicollis (Manual annot.)
HsRab23
MbRab23
HsRab29
MbRab29
MbRab32A
HsRab32
HsRab38
MbRab32B
MbRab9
HsRab9B
HsRab9A
HsRab7A
MbRab7
HsRab7B
HsRab28
MbRab28
HsRabL4
MbRabL4
HsRab34
HsRab36
MbRabX4A
MbRabX4B
MbRabX2
HsRab24
HsRab21
MbRab21
HsRab22A
HsRab22B
MbRab22
HsRab5C
HsRab5B
HsRab5A
MbRab5
HsRab17
HsRab20
MbRabX3
HsRab6C
MbRab6
HsRab6A
HsRab6B
HsRab41
MbRabX1
HsRab30
HsRab43
HsRab19
HsRab33B
HsRab33A
MbRabX5
HsRab42
HsRab39A
HsRab39B
HsRab25
HsRab11B
MbRab11
HsRab11A
HsRab2B
HsRab2A
MbRab2
HsRab14
MbRab14
MbRab4
HsRab4B
HsRab4A
HsRab45
HsRab44
HsRab26
HsRab37
HsRab18
MbRab18
HsRab35
HsRab1B
HsRab1A
MbRab1
HsRab3C
HsRab3A
HsRab3D
HsRab3B
HsRab27B
HsRab27A
HsRab12
HsRab13
HsRab10
HsRab8A
HsRab8B
MbRab8
HsRab15
HsRab40C
HsRab40A
HsRab40B
0.4
Legend
phase 1 wrong
phase 2 wrong
fully correct

whether the proposed subfamily agreed either with the public annotation or our own one for *M. brevicollis*. We distinguished between two operating modes of the Rabifier: a normal one which does not consider the confidence levels the Rabifier attributes to its classifications, and a high-confidence mode which accepts only the high-confidence annotations above a certain confidence threshold, whereas those below are classified as belonging to the undetermined subfamily RabX. Ignoring the information provided by the classification confidence, we correctly called 16 out of 17 Rabs for *T. brucei*, 59 out of 91 in *E. histolytica* and 20 out of 25 for *M. brevi-*

FIGURE 2.2 *(preceding page)*: *Validation and benchmarking of the Rabifier*—(A) summarises the validation in normal mode, *i.e.* without taking the subfamily score produced by Rabifier into account, against the Rab families of *Trypanosoma brucei* [31], *Entamoeba histolytica* [29] and *Monosiga brevicollis*, which we annotated in (E). Three quantities needed to judge the performance of the Rabifier are shown for Rabs belonging to human and other subfamilies separately: sequences erroneously classified as not being a Rab by the Rabifier (red), sequences correctly identified as Rabs, however, wrongly classified at subfamily level (light green), and those which were entirely correct (dark green). (B) displays the distribution of confidence scores associated to each subfamily call, respecting the same colour code as above. The blue line indicates the threshold which we propose on default, and below which subfamily classification may be rejected and treated as a undefined RabX. That choice is based on the ROC-curve [40] analysis shown in (C), which plots the true positive rate against the false positive rate for each possible confidence threshold [40] and provides a combined measure of the accuracy of a classifier (Area under the curve, AUC [41]). The effect of choosing an 0.4 confidence threshold (blue circle) on the classification accuracy, *i.e.* running the Rabifier in high confidence mode, is shown in the inlay. (D) plots the improvement in terms of the three quantities discussed above the Rabifier achieves compared to an alternative strategy (see Results and Discussion for details on its implementation). (E) Phylogenetic tree of the human and *M. brevicollis* Rab family on which the manual classification of the latter Rab family was based (bootstrap support above 70% shown). Colours indicate the results of the corresponding automated annotation for that specific sequence. *Abbreviations:* subfamily (sf.), annotation (annot.)

collis, leading to an overall fraction of 71.4% correct decisions (79.7% on average per organism). However, if one defines a threshold below which a classification is systematically considered as belonging to the undefined subfamily RabX, the accuracy can be substantially improved. To illustrate this, Figure 2.2B displays the distribution of scores associated to correct and wrong calls, which shows that wrong calls clearly have lower confidence scores on average. In order to test for all possible thresholds exploiting this difference, we performed a ROC curve analysis presented in Figure 2.2C. This machine learning technique allows to summarise and quantify the classification performance for all thresholds (Area Under the Curve (AUC) [41], here 0.94), and enables to objectively choose a threshold providing an optimal TP/FP-tradeoff. Here, we opted for 0.4, which we propose as a default choice for the interpretation of the Rabifier's results. Yet, the use of this threshold is not fixed as it may vary depending on the dataset, and can be freely modified by users of the Rabifier. The consequences of applying a cutoff on the classification accuracy are quantified by the inlay in Figure 2.2C: only trusting calls with confidence higher or equal to 0.4 greatly reduces the amount of misclassified Rabs from non-human subfamilies and improves the overall accuracy to 90% (92.01% on average per organism).

In summary, we conclude that our workflow is able to correctly discern Rabs from other GTPases. Furthermore, calls both at family and subfamily level have an associated confidence score which correctly captures uncertainty in the decision. Relying on the information provided by the confidence level, the Rabifier suggests correct subfamilies around 90% of the time even in difficult and phylogenetically isolated cases.

2.2.3 Benchmarking the Rabifier

After having established the correctness of our procedure, we wished to assess the improvement it represents over possible alternative large-scale

approaches in an objective manner. This excludes benchmarking against methods for example based on phylogenetic trees, as reasoning over them is difficult to automate and not feasible for thousands of sequences.

We chose to compare the Rabifier to the Conserved Domain Database at the NCBI [42], the only resource we are aware of that specifically scores for RabF motifs. To this end, we implemented an alternative decision scheme which given a protein retrieves the protein name and CDD domain annotation of its BH in the NCBI protein database. Note that if the protein is in the NCBI database, the BH retrieves the protein itself. As for the choice of genome, the Rabifier has to be benchmarked against an organism whose Rab family has not been manually curated, as our alternative procedure would simply retrieve that annotation. Moreover, an organism from a taxon which is both close to Metazoa and for which no information on the Rab family exists best ensures an unbiased measurement. These requirements are met by the Choanoflagellate *M. brevicollis*, which we analysed ourselves and is thus an ideal candidate for a direct comparison.

The results of this experiment are detailed in Figure 2.2D (see also Table S1). As above, we distinguished between the ability to discern Rabs from other GTPases and to actually propose the correct subfamily for a given Rab. First, while the Rabifier achieved 100% accuracy in separating Rabs from other GTPases in *M. brevicollis*, the alternative strategy—although not introducing false positives—misses 8 of 25 Rabs leading to an overall drop in sensitivity. On top of these eight sequences, the Rabifier correctly suggests subfamilies for four further proteins wrongly classified by the alternative strategy, leading to an overall difference of 12 sequences correctly classified only by the Rabifier.

Thus, our annotation pipeline represents a significant improvement over currently available large scale approaches, both in terms of sensitive identification of Rabs and especially with regards to the difficult automatic classification of Rabs into subfamilies.

2.2.4 Availability of the Rabifier and its predictions

In order to make our pipeline useful to the cell biology community interested in Rabs, we provide access to the Rabifier in form of a web tool (Figure 2.3A). Via the graphical interface users can submit up to five protein sequences at a time, and the classifications generated by our workflow are returned together with their associated degree of confidence. We envisage users who want to quickly generate hypotheses about one or a few candidate proteins. Users wishing to classify more sequences are encouraged to contact us. We emphasise that the Rabifier works without need for phylogenetic information about the input, hence any set of protein sequences can be submitted. In addition, we generated a database of nearly 8,000 classified Rab sequences in 247 eukaryotic genomes, which we make publicly available at `www.RabDB.org` (Figure 2.3A) together with basic browsing and visualisation tools. Our database is built on top of the Superfamily database [34] (September 2009 release), which allows us to follow its release cycle and include predictions for all newly sequenced genomes contained therein. Figure 2.3B details the phylogenetic distribution of genomes in RabDB and the number of Rabs we predict in each of those eukaryotic branches. The correctness of the content in `www.RabDB.org` is not manually confirmed systematically. However, we constantly inspect and manually curate the generated predictions and update our internal reference database accordingly. Furthermore, we provide users the possibility to notify us of a potential mis-annotation found in the database such that we can correct the classification of the Rab in question. These measures further enhance the expected quality of future releases of `www.RabDB.org`.

2.2.5 New hypothetical subfamilies

As can be noticed from Figure 2.3B, the Rabifier detected a large number of Rabs not belonging to any subfamily represented in our reference set,

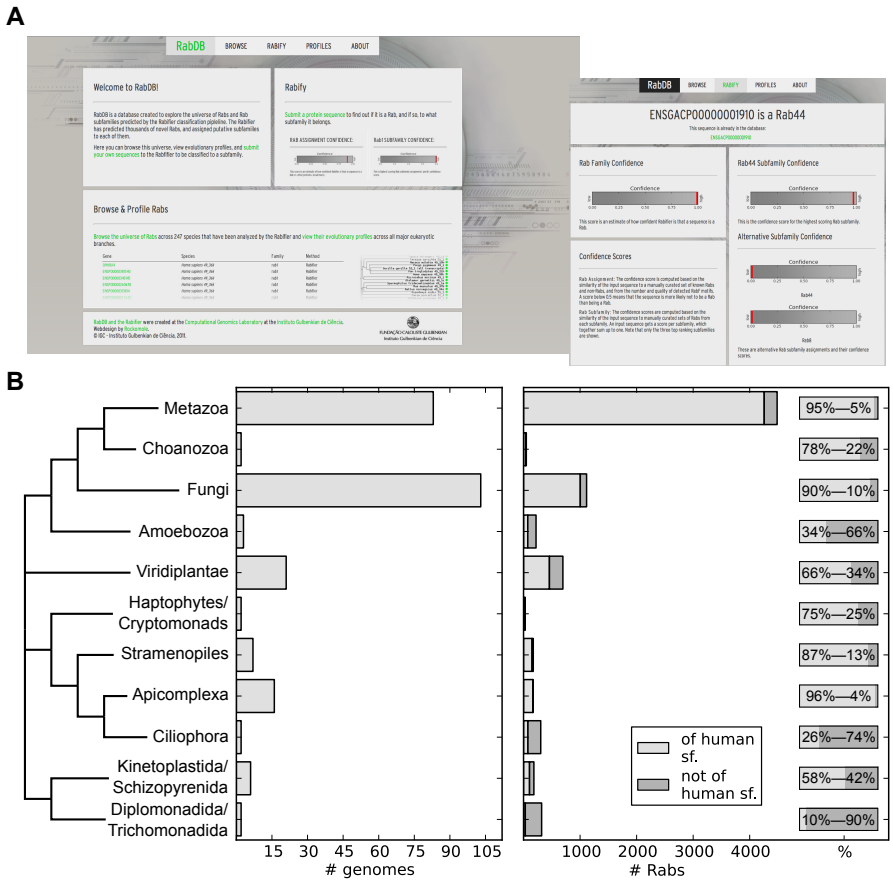


FIGURE 2.3: *Resources we make available*—(A) Snapshots of the database `www.RabDB.org` which provides public access to the results of the Rabifier applied to the Superfamily database [34] and the online version of the Rabifier. (B) Statistics of the current content of `www.RabDB.org` in terms of number of genomes (left), absolute number of Rabs either belonging to a subfamily also present in humans or not (middle), and the relative fraction of the two types of Rabs for a given branch (right). The cladogram (*i.e.* the branch length are arbitrary, see [43]) of the eukaryotic taxa is derived from [44].

i.e. most subfamilies which have been described before. By definition these sequences show no similarity to any functionally characterised Rab, hence

a bioinformatic annotation is not possible. However, in order to structure the space of new sequences and provide a starting point to study this yet unexplored diversity, we clustered these Rabs with respect to their sequence identity and propose several hypothetical Rab subfamilies (see Section 2.4 for details). The result of this procedure is shown in Figure 2.4, which details the amount of hypothetical subfamilies according to the breadth of their occurrence (see Figure 2.16 for an overview of the amount of Rabs falling into each of these classes). We integrated these new subfamilies both in our database, where they can be browsed with help of the visualisation tools we provide, and in the online version of the Rabifier. Note that in addition to these new hypothetical subfamilies we still find hundreds of Rabs that we cannot group with others. Those may result from erroneous gene models in less well curated genomes, represent cases where our simple clustering procedure failed, or indeed be bona fide singletons. A detailed phylogenetic analysis may be required to resolve these cases which is out of the scope of this study.

2.2.6 Global Dynamics of the Rab sequence space

A dataset of 8,000 Rabs allows us to take a global view of the Rab sequence space, and to address previously inaccessible questions. Here, we investigate the patterns of Rab repertoire expansion in the eukaryotic tree (Figure 2.5). Expansion of certain protein families has been found to correlate with organismal complexity [45]. The anecdotal evidence of Rab profiles in different organisms suggests at least three possible scenarios: a conserved core of Rabs present in all organisms; tinkering with a core of subfamilies by taxon- or species-specific expansions of existing subfamilies; a major variation of the Rab machinery with taxon- or species-specific Rab repertoires. We asked whether any such scenario is apparent for the Rab family across the eukaryotic tree, or if different ones predominate in different branches.

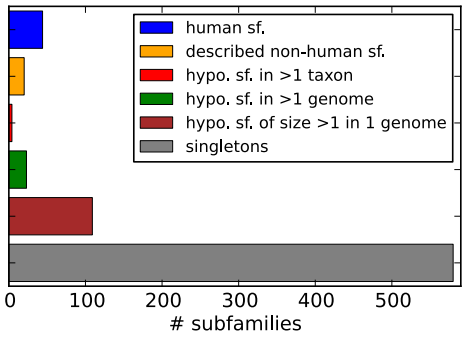
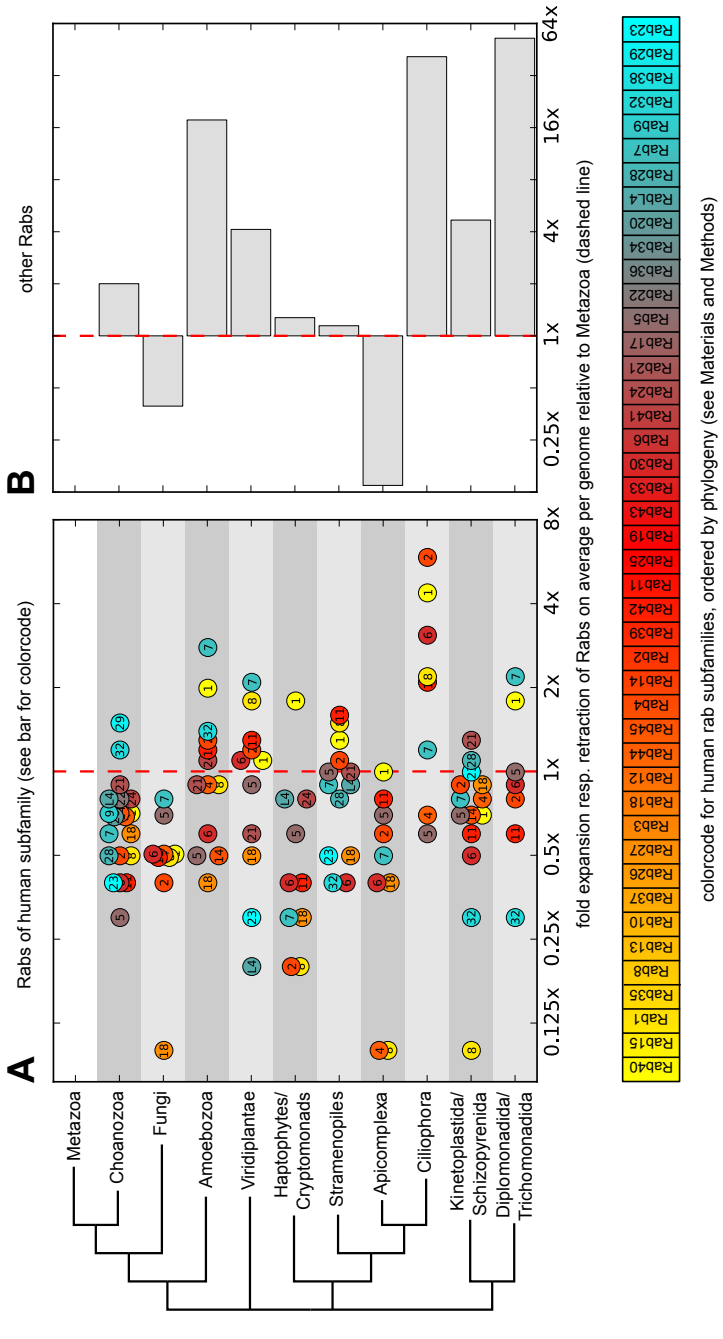


FIGURE 2.4: *Rab subfamilies in our dataset*—Number of different Rab subfamilies found in our dataset. Human sf. are shown in blue, and other known sf. in orange. The last four categories are hypothetical subfamilies we propose in the context of this paper (see Section 2.4 for details on the procedure): subfamilies whose members span more than one taxon (red), those spanning more than on genome (green), subfamilies with several members yet only present in one organism (brown) and finally singletons (grey) which are not similar to any other known Rab. All members and subfamilies can be browsed in our website at `www.RabDB.org`. *Abbreviations:* hypothetical (hypo.), subfamily (sf.)

We observe a tremendous heterogeneity in the sizes of Rab repertoires, ranging from five to several hundreds of Rabs in *Encephalitozoon cuniculi* and *Trichomonas vaginalis* respectively. Genomic analyses have shown a general trend for more and larger families in bigger genomes [46, 47]. In the case of Rabs, linear regression over all taxa reveals that genome size explains roughly 60% of the observed variance in numbers of Rabs in an organism (Figure 2.14). However, due to the current bias in fully sequenced genomes towards Opisthokonts (compare Figure 2.3B), it is unclear whether these numbers will remain as such in the future. We find that closely related organisms tend to have similar Rab repertoires in size, but at the level of phyla we encounter marked differences indicating taxon-specific adaptations. For example, although Ciliophora and Apicomplexa belong to the same superphylum (Alveolata), these sister


A
Rabs of human subfamily (see bar for colorcode)
Metazoa
Choanozoa
Fungi
Amoebozoa
Viridiplantae
Haptophytes/
Cryptomonads
Stramenopiles
Apicomplexa
Ciliophora
Kinetoplastida/
Schizopyrenida
Diplomonadida/
Trichomonadida
0.125x
0.25x
0.5x
1x
2x
4x
8x
B
other Rabs
0.25x
1x
4x
16x
64x
fold expansion resp. retraction of Rabs on average per genome relative to Metazoa (dashed line)
Rab40
Rab15
Rab1
Rab35
Rab8
Rab13
Rab10
Rab37
Rab26
Rab27
Rab3
Rab18
Rab12
Rab44
Rab45
Rab4
Rab14
Rab2
Rab39
Rab42
Rab11
Rab25
Rab19
Rab43
Rab33
Rab30
Rab6
Rab41
Rab24
Rab21
Rab17
Rab5
Rab22
Rab36
Rab34
Rab20
Rab4
Rab28
Rab7
Rab9
Rab32
Rab38
Rab29
Rab23
colorcode for human rab subfamilies, ordered by phylogeny (see Materials and Methods)

phyla show very different repertoires, highly expanded in the first case, and streamlined in the second. The smaller Rab repertoires in Apicomplexan genomes, mostly dominated by intracellular parasites, may be due to secondary gene loss, similar to that reported in bacterial intracellular parasites and endosymbionts [48] and in the obligate intracellular parasitic Microsporidia [48]. Another example of reduction of Rab repertoires is observed in the fungal branch, as we reported previously [24] and now confirm based on an extended set of 103 genomes. It is noteworthy that Fungi are Unikonts, a taxon which comprises Metazoa and Amoebozoa, *i.e.* branches that appeared to have suffered independent expansions of their Rab repertoires [29, 36]. We observe large expansions in DiplomonadidaTrichomonadida, Ciliophora and Amoebozoa. Much of these expansions are accounted for by species-specific subfamilies (see Figure 2.4). This demonstrates that there is frequent invention of new Rabs, perhaps in a taxon-specific manner—a hypothesis that will have to await broader sampling of the genomes space to be tested in most taxa. On the other hand, inspection of Figure 2.5 reveals that for those Rabs that can be clas-

FIGURE 2.5 *(preceding page)*: *Rab subfamily expansions relative to Metazoa in a dataset of 247 genomes*—For each of the eukaryotic taxa (as derived from [44]), (A) displays the relative size compared to Metazoa of each human Rab subfamily on average per genome. The dashed line represents the average in Metazoan genomes, *i.e.* any circle lying on that line represents a human subfamily that has the same amount of members on average per genome than on average in Metazoa. Similarly, any circle to the left represents a subfamily that is smaller compared to Metazoa, finally, all on the right are expanded compared to the Metazoan average. Note that the axis are in logarithmic scale. In addition to the numbers indicating the human Rab subfamily, a colour code to distinguish subfamilies is shown below, where similar colours indicate proximity in the phylogenetic tree of human Rabs. The same plot for all other Rabs is shown in (B), again on a logarithmic scale. All sequences used are accessible at `www.RabDB.org`. *Abbreviations:* subfamily (sf.)

sified, different subfamilies expanded in each branch of the tree. For example, Rab7 forms the largest subfamily in Diplomonadida/Trichomonadida and Amoebozoa, whereas Ciliophora's most expanded subfamily is Rab2. This suggests that these are independent expansions, which has already been observed for example within the Rab5 subfamily [24, 49]. Note that we repeated these analyses for different confidence cutoffs and observed no significant consequences on the broad picture.

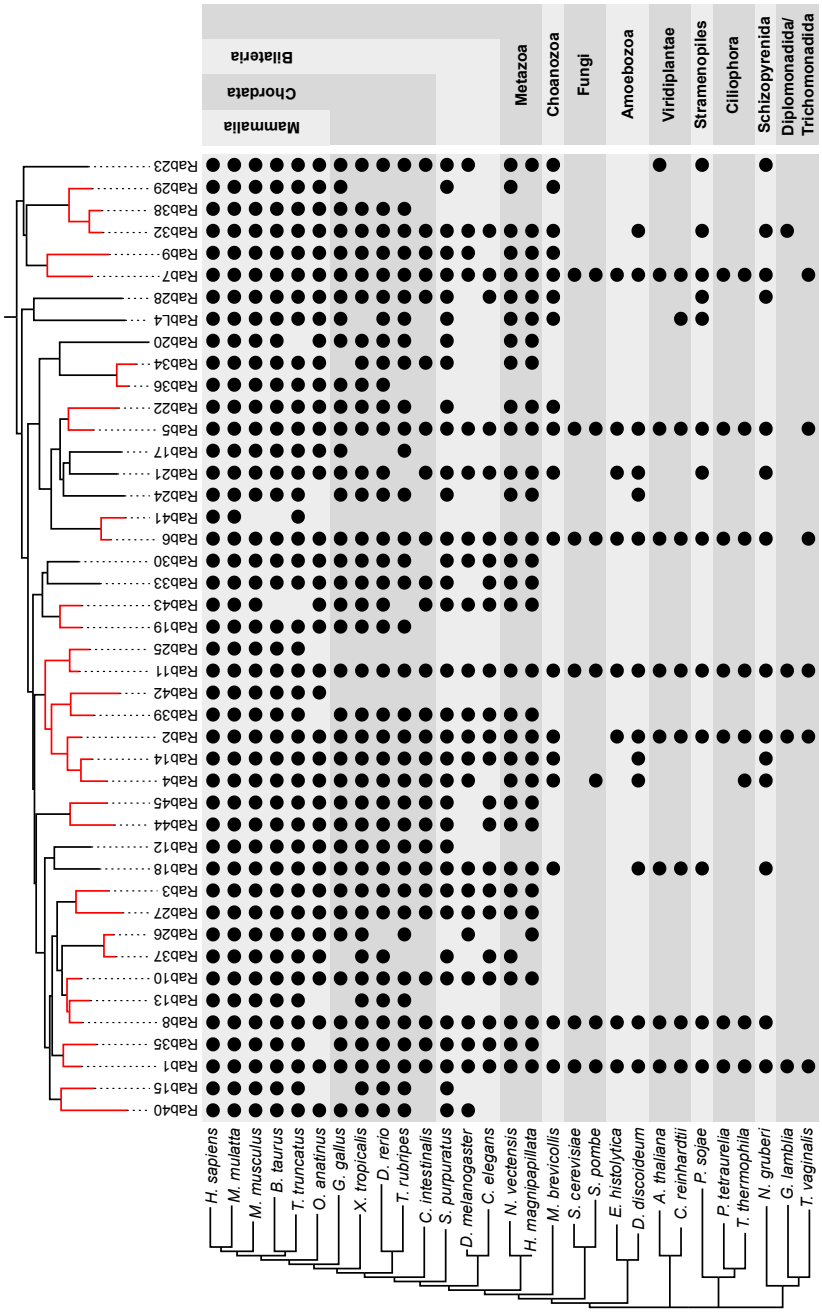
In summary, the global evolution of Rab repertoires is highly dynamic with frequent taxon-specific subfamily expansions, gain of new Rabs and losses. Hence, we observe a scenario where a core set of Rabs tends to be universally conserved, and can coexist in different taxa with subfamily expansions and/or taxon- or species-specific Rabs. It is clear that no unique path to cellular complexity and specialisation exists, implying that any conclusion about the evolution of Rabs in a given taxon is not necessarily true for other eukaryotic taxa.

2.2.7 Dating the origin of Rabs and expanding the LECA

The systematic identification and classification of Rab repertoires in multiple branches of the eukaryotic tree of life allows the establishment of a phylogenetic profile for each Rab subfamily. As Metazoa and Fungi are the most extensively sampled and best annotated groups, we profiled human subfamilies (Figure 2.6) and determined their likely time of origin (Figure 2.7). For a detailed analysis of fungal Rabs see [24]. We further established the direction of duplication, *i.e.* from which Rab subfamily another emerged by duplication and subsequent divergence, by crossing their likely time of origin with a phylogenetic tree of the human Rab family. We reasoned that for two closely related Rabs, the one that is present in more taxa is likely the ancestral one. Since all Rabs are by definition paralogs and especially the deeper evolutionary relationships are unclear, we restricted the inference of direction of duplication to well supported

branches. Here, we define well supported branches as those with bootstrap support higher than 58% in a tree of human Rabs, which is chosen to include the branch between Rab5 and Rab22 as their association is commonly accepted [50–54]. As further support, we note that all branches selected according to this criterion are also present in the tree of mouse Rabs we present below, however, in general 58% is not a strong branch support and should not be used indiscriminately on trees of other Rabs. Based on a 58% cutoff, one obtains directed duplication scenarios for a number of subfamilies as summarised in Figure 2.7. We term subfamilies with a clear origin as 'derived'.

This analysis suggests new candidates for ancestral Rabs. Previously Rab1, 2, 4, 5, 6, 7, 8 and Rab11 [17], Rab18 [30, 59], Rab21 [29, 60] as well as Rab23 and 28 [31] could be mapped to more than one major branch of the eukaryotic tree, making them likely candidates to be present in the LECA. Our results support these assignments and reveal a new set of proteins that can be found in two or more basal eukaryotic taxa, namely Rab14, 32 and RabL4. Applying the same parsimony argument as previous studies suggests that these Rabs were part of the ancestral set of Rab in the LECA. Are these putative ancestral Rabs an artefact due to incorrect assignments or convergent evolution? We validated the automated subfamily classification by phylogenetic trees, and could not disprove their annotation (Figures S4 A-C from reference [61]). The possibility of convergent evolution is however harder to rule out. Regardless, an organism with 15 Rabs is not surprising and comparable with some unicellular eukaryotes [31, 32], and free living fungi frequently have less [24]. It is remarkable that with every new analysis the LECA appears to become increasingly more complex [62]. On functional grounds, mapping these Rabs to the LECA is plausible. RabL4, also known as IFT27, plays a role in ciliogenesis as part of the Intra Flagella Transport (IFT) machinery [63]. Flagella are believed to be ancestral characters, present in


Mammalia
Chordata
Bilateria
Metazoa
Choanozoa
Fungi
Amoebozoa
Viridiplantae
Stramenopiles
Ciliophora
Schizopyrenida
Diplomonadida/
Trichomonadida
RAB23
RAB29
RAB38
RAB32
RAB9
RAB7
RAB28
RABL4
RAB20
RAB34
RAB36
RAB22
RAB5
RAB17
RAB21
RAB24
RAB41
RAB6
RAB30
RAB33
RAB43
RAB19
RAB25
RAB11
RAB42
RAB39
RAB2
RAB14
RAB4
RAB45
RAB44
RAB12
RAB18
RAB3
RAB27
RAB26
RAB37
RAB10
RAB13
RAB8
RAB35
RAB1
RAB15
RAB40
H. sapiens
M. mulatta
M. musculus
B. taurus
T. truncatus
O. anatinus
G. gallus
X. tropicalis
D. rerio
T. rubripes
C. intestinalis
S. purpuratus
D. melanogaster
C. elegans
N. vectensis
H. magnipapillata
M. brevicollis
S. cerevisiae
S. pombe
E. histolytica
D. discoideum
A. thaliana
C. reinhardtii
P. sojae
P. tetraurelia
T. thermophila
N. gruberi
G. lamblia
T. vaginalis

the LECA [64, 65]. Rab32 regulates transport to the pigmentedsecretory granules [66], an animal-specific function, but it has also been claimed to have a mitochondria-related function [67, 68]. The known function of Rab14 in phagosome maturation and a recycling step at the TGN [69, 70] is less clearly ancestral, but it may lend support for a phagotrophic LECA as previously proposed [71].

In summary, our results support the claim that the LECA had a highly complex endomembrane system, and that secondary Rab losses have been dominant in the evolution of the major eukaryotic taxa [17].

2.2.8 The Rab family in *Monosiga brevicollis* and the origin of animals

The emergence of multicellularity is one of the major transitions in evolution [72], which happened independently multiple times (see [73] for a recent review). There are several critical features necessary for the evolution of multicellular organisms, for example mechanisms for cell adhesion, cell polarity and inter-cellular communication. Little is known about how protein trafficking has evolved during this transition. We take advantage of our extensive annotation of the Rab family to derive the Rab complement prior to and after the emergence of multicellularity in Metazoa.

Monosiga brevicollis belongs to the Choanozoa, the closest unicellular relatives of Metazoa. The genome of this organism was only recently

FIGURE 2.6 *(preceding page)*: *Phylogenetic profiles of human Rab subfamilies in selected organisms*—A black dot reads as presence of the corresponding subfamily in the respective species. Rab subfamilies are ordered according to the top phylogenetic tree generated as explained in Materials and Methods. Branches with bootstrap support above 58 are coloured in red. The tree on the left represents the species' branching order and is derived from [44, 55–57] together with the naming of the partially nested monophyletic groups on the right.

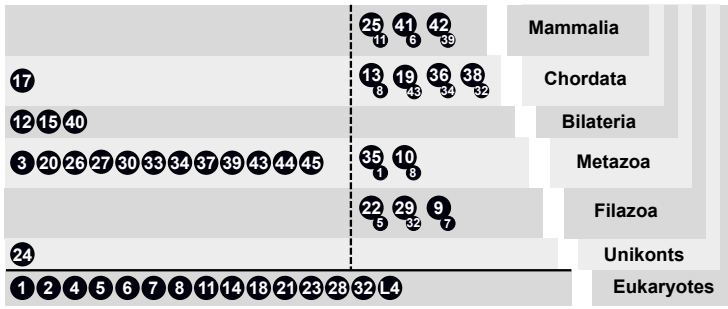


FIGURE 2.7: *Summary of evolutionary age and duplication origin of human subfamilies*—Each level represents a nested evolutionary stage from the LECA to humans (derived from [44, 58]) with one circle per human subfamily. Those subfamilies for which we could establish a clear origin, that is which subfamily it was derived from by duplication, are right from the dotted line with the subfamily it was derived from attached at the bottom right.

sequenced [74], and in the context of the validation of the Rabifier we conducted a detailed analysis of its Rab family. The phylogenetic tree in Figure 2.2E reveals a relatively large Rab family with nearly no sub-family expansions (see also Figure 2.5), *i.e.* mostly with a single member per subfamily (only Rab32 has two members). This is also observed in simpler animals like *D. melanogaster* and *C. elegans* [59], suggesting that larger subfamilies observed in mammals represent taxon-specific duplications. Secondly, we observe several organism-specific Rabs, which we labeled MbRabX. Consistent with results from the last section, the "invention" of new Rabs is a recurrent feature in multiple branches of the tree of life (*e.g.* [27, 29, 31, 59]). We observed the emergence of three novel sub-families, Rab9, 22, 29, none playing 'animal-specific' roles. The function of Rab29 is unknown, but Rab9 and Rab22 both appear to be involved in late endocytic traffic [52, 53, 75, 76]. Surprisingly, the genome of *M. brevicollis* codes for proteins previously believed to be specific to multicellular organisms, for example Cadherins [74, 77]. In animals, trafficking of the cell adhesion molecules Integrins and Cadherins is regulated

by Rab4, 5, 11, 21 and 25 [78–81], and Rab5 and 7 [82, 83], respectively. Interestingly, these Rabs are also found in *M. brevicollis*, and—with the exception of Rab25—are all likely ancestral proteins. That highlights that complex new functions, as are for example the regulation of Cadherin and Integrin and ultimately cell adhesion, can be gained without inventing new subfamilies.

Our analysis revealed 14 Rab subfamilies that emerged at the base of Metazoa (Figure 2.7). Surveying the currently known functions of these animal-specific subfamilies suggests roles mainly in regulated secretion (Rab3 [84–87], Rab26 [88], Rab27 [87, 89–91], Rab33 [87], Rab37 [87, 92], Rab39 [93]), trafficking from (Rab10 [94]) and to the Golgi (Rab43 [95]) and more generally localisation at the Golgi (Rab30 [96–98], Rab33 [99], Rab34 [100], Rab43 [101]). Hence, our analysis suggests that the appearance of animals cooccurred with an important expansion and specialisation of the secretory pathway.

2.2.9 A model for Rab subfamily innovation

Gene duplication is a frequent mode of gene gain in eukaryotes. This is well illustrated by the expansion of the Rab family in emergence and evolution of Metazoa. Following gene duplication, the most common fate for one of the duplicates is accumulation of mutations up to the point of pseudogenisation. In the alternative case, the retention of both duplicates has been explained by different theoretical scenarios, recently surveyed in reference [102]. Most prominently, either divergence results in gain of a beneficial new function (neo-functionalisation) by one of the duplicates, or disruption of complementary parts of the function in each of the genes leaves both paralogs indispensable to perform the original function (sub-functionalisation). As discussed in reference [102], those models predict distinct types and strengths of selective forces acting on the two duplicates allowing to test and distinguish amongst putative scenarios. Namely,

while in both neo- and subfunctionalisation the new copy indistinguishably evolves neutrally, detecting purifying selection acting on the original copy is an indication of neofunctionalisation, whereas relaxed purifying selection or neutral evolution is suggestive for subfunctionalisation. In the case of Rabs, Figure 2.6 shows that the original copy is conserved and keeps its identity as the original subfamily, whereas the new copy initiates a distinct subfamily defined by a discernible level of sequence divergence. We interpret this pattern as evidence that the mode by which the Metazoan Rab family expands is most probably neofunctionalisation rather than subfunctionalisation.

To gain further insights into the nature of the gain of function, we asked whether the derived Rab subfamilies show differences in tissue-specificity that could hint at the type of newly evolved functions. To this end, we investigated tissue-specificity in expression of Rabs in mouse tissues and cell lines (Figure 2.8) by means of PCR (see Section 2.4). We also analysed publicly available microarrays (Figures 2.15 and S5 from reference [61]) which overall corroborate the trends described in the following.

First, we observed that all ancestral Rabs are widely expressed (*i.e.* in all tested tissues), most probably performing general functions required in all tissues. Similarly, Rabs that predate the advent of multicellularity are also broadly expressed, a general phenomenon that has been described for genes which emerged prior to multicellularity [103]. Second, for the derived subfamilies in which a clear directionality of duplication could be established (see Figure 2.7), we detected a trend for an increase in tissue specificity, *i.e.* a reduction in number of tissues in which the Rab is expressed relative to its progenitor subfamily. For example, Rab34 is expressed in all tissues investigated but the liver, whereas the derived Rab36 is only expressed in lung and brain. Thirdly, at no time we observe complementary expression, *i.e.* a pair of subfamilies which have opposite tissue specificities.

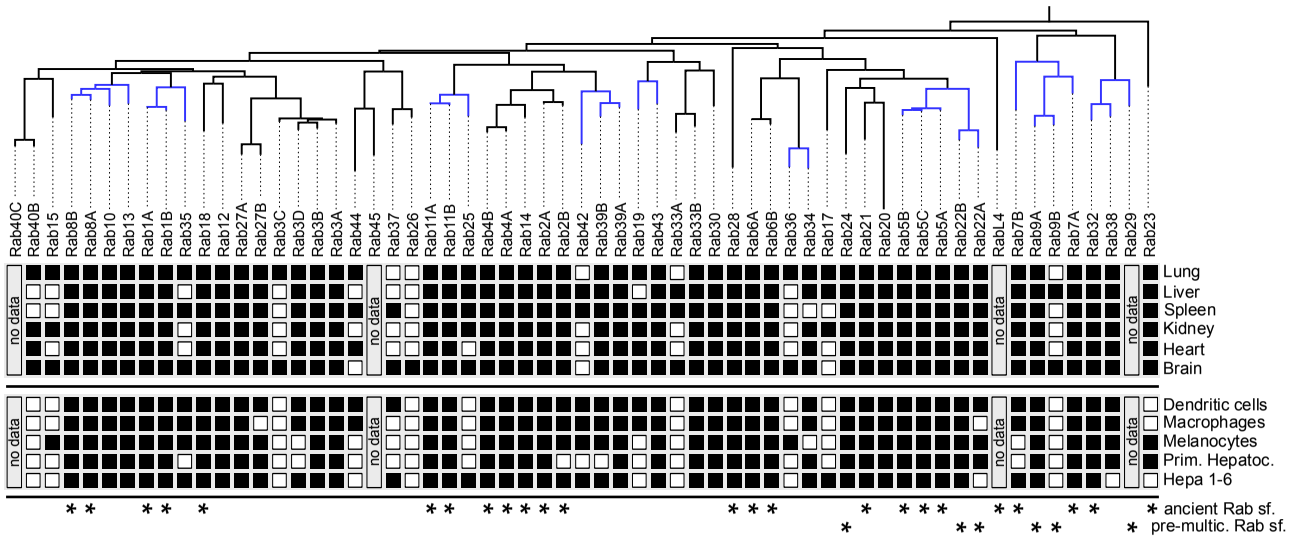


FIGURE 2.8: *Increasing tissue specificity in expression of derived Rabs in mice*—Summary of PCR experiments establishing expression (black squares) or lack thereof (white squares) of mouse Rabs in six tissues and five mouse cell lines. Stars on the bottom indicate subfamilies which we found already present in LECA, and that predate the evolution of multicellularity (see Figure 2.7). Branches coloured in blue in the phylogenetic tree of mouse Rabs on the left are those for which we test the hypothesis that derived subfamilies are expressed in the same or in a subset of tissues of the Rab they were derived from (see Figure 2.7 for a summary of which Rabs have a clear origin). *Abbreviations:* subfamily (sf.), primary Hepatocytes (Prim. Hepatoc.), multicellularity (multic.), last eukaryotic common ancestor (LECA)

Overall, these observations are strong indications that derived subfamilies are retained for a new tissue-specific functions, different from or at least complementing the progenitor ones. Thus, our results support a neo-functionalisation model explaining the retention of novel Rab subfamilies in Metazoa. This model makes several predictions about expression patterns of Metazoan Rabs for which we could not derive expression data. Concretely, Rab41 which we only find in primates and dolphin is expected to show a restricted tissue expression, as its origin from Rab6 is statistically well supported. Rab29 is expected to be ubiquitously expressed despite its clear origin from Rab32 as it predates the evolution of multicellularity, a prediction at least supported by our microarray-based analysis (Figure S5 in reference [61]). One notable observation is that the tested mouse tissues express an unexpectedly high number of distinct Rabs. This is also observed in individual cell lines, which indicates that it is not an artefact from multiple cell types mixed in the tissue. While it is clear that Rabs are expressed at different levels [104] (see also Figure 2.15), our results from a more sensitive method than microarrays reveal that the tissue-specific Rabs may be more widely expressed than previously anticipated. It remains to be investigated whether the low levels of expression we can detect by PCR are functionally significant.

2.3 Conclusions

We developed the 'Rabifier', a bioinformatics tool to identify and classify Rabs from any set of protein sequences with no need for additional phylogenetic information, which we make available as a web tool for the community. We deployed the Rabifier on 247 proteomes predicted from complete genome sequences, generating the first comprehensive view of the Rab sequence space, which we also make available in form of a browsable database of Rab proteins. We envisage that cell biologists interested

in specific organisms may use RabDB and the Rabifier as a first description of the family, at accuracy levels we showed to be very high. In fact, our predictions are well suited to be the first step towards high quality manual annotations. Furthermore, we introduced unified and objective criteria for the annotation of Rabs which is especially important for large-scale comparative studies, which can now be grounded on a coherent body of data.

The classification of Rab repertoires in hundreds of genomes gives us the first global view of the Rab family in evolution, revealing that this family followed different routes in each branch of the tree. Massive expansions co-exist with extensive losses. These expansions can vary from taxon to taxon, suggesting that care must be taken when transferring information amongst different branches of the tree of life. In this respect, future work may focus on understanding the detailed evolutionary patterns in eukaryotic taxa other than Metazoa, which we analysed here. It appears that plants are ideal candidates for such a study as multiple genomes have been sequenced covering both unicellular and multicellular organisms.

One of the perhaps most surprising observations we made was the extension of RabXs, *i.e.* Rabs that cannot be assigned to any previously characterised subfamily. Hence, a major bioinformatic and cell biological challenge now is to identify how many Rab subfamilies exist overall, and to establish their conservation or taxon-specificity. Here, we started this classification by proposing new Rab subfamilies derived from clustering of RabXs with respect to their sequence similarity. We hope to stimulate further research which may allow the refinement of our criteria and ultimately the definition of a Rab subfamily. The notion of Rab subfamily is supposed to reflect both evolutionary history and functional information, but has historically been mixed with less clear criteria. In the absence of functional information for all Rabs, phylogenetic analysis becomes particularly important, especially for functional prediction. In

this context, it is all the more serious that we found a notorious frailty of Rab trees. Factors such as choice of sequences, outgroups, alignment program, probabilistic model and program implementing it contribute to very different trees (compare for example [59, 105, 106] and Figures S4A-C in reference [61]). We thus need to derive objective criteria that define a Rab subfamily which go beyond the clearly outdated yet still useful sequence identity cutoff [36]. Possibilities are for example to introduce soft thresholds depending on background divergence levels within a given taxon, or to restrain the area considered to measure sequence divergence to the functionally relevant regions.

We focused on the evolutionary path from the LECA to mammals in order to gain insight into the mechanism of functional innovation within the Rab family. Based on objective and re-usable criteria we were able to map directionality to duplications clarifying the origin of some human subfamilies. Crossing these relations with data on tissue-expression patterns of Rab genes, we proposed that neo-functionalisation best explains the emergence of new subfamilies. More recent subfamilies are most likely retained for newly evolved tissue-specific functions and coexist with older ones in a subset of tissues. It remains to be determined whether the same happens within a subfamily, *i.e.* whether a RabXa and a RabXb represent cases of neo- or sub-functionalisation [107]. This is particularly relevant to conceptually tell apart isoforms and distinct subfamilies. As we restricted our analysis to subfamilies present in humans, it is important now to test whether the same neo-functionalisation scenario is observed in other branches of the tree of life. As mentioned before, plants appear to be ideal candidates to extend this analysis. Finally, while we studied the fate of new subfamilies in the context of tissue-specific expression, it will be important to understand the contribution of subcellular re-localisation to neo-functionalisation [108, 109].

New generations of sequencing methods promise to change that scale

at which we perform comparative analysis in cell biology. But for this change to reach the cell biology community, we need the appropriate tools that allow the non-bioinformatician to take advantage of all the emerging data. The Rabifier is one such tool, tailored to enable the cell biologist to analyse protein repertoires in hundreds of genomes.

2.4 Materials and Methods

2.4.1 Ethics Statement

C57BL/6 mice were bred and housed in the pathogen-free facilities of the Instituto de Gulbenkian de Ciência (IGC). Mouse experimental protocols were approved by the Institutional Ethical Committee and the Portuguese Veterinary General Division.

2.4.2 The set of human Rabs

Before we devised a workflow able to identify and classify Rabs, we decided which protein subfamilies we considered being human Rab subfamilies. Since the early genomic analyses of the human Rab repertoire reporting subfamilies 1 to 40 (with exception of 16) [36], five subfamilies have been newly discovered (41 to 45/RasEF) [110]. Besides those clear cases, the distinction remained less obvious for those which are termed 'Ran' and 'Rab-like', each of which we briefly discuss in the following.

Rans control nucleocytoplasmic shuttling [111], and are frequently considered to be members of the Rab family [105, 110]. This view is supported by our own phylogenetic analysis (see tree in Figure S3 in reference [61]), although without strong bootstrap support. Due to the distinct function and localisation [111] partly within the nucleus we do not further consider Rans in our dataset. However, Rans have recently been linked to ciliary entry of certain kinesins [112], and they may be included in the future.

RabL2 proteins were already mentioned in reference [36] where it is concluded that they are not Rabs, amongst others due to non-conforming RabF motifs. In reference [105], RabL2s are said to cluster together with Rans, which we do not include in our analysis. The tree of human GTPases shown in reference [106] suggests that RabL2 proteins branch of Rhos at an early stage. Finally, our own tree of human GTPases (Figure S3 in reference [61]) positions RabL2s at the periphery of the Rab branch, yet with little bootstrap support. Altogether, we do not see enough evidence for RabL2 proteins to be considered Rabs. The situation is similar for RabL3 and RabL5. Colicelli clusters them together with Rans [105], whereas in reference [106] both reside on a branch with Arfs though classified as belonging to none of the classes Rab, Ras, Arf, Rho or Ran. Our tree of human GTPases suggests that RabL5 and Arfs have a common ancestor, equally so RabL3 and RabL2, hence we ignored both in our further analysis. Rab7L1 is nearly identical to Rab29 and represents a simple case of naming ambiguity, as has already been pointed out in reference [36].

The last case is RabL4, which all [105, 106, 110] consider being a Rab. We confirmed that interpretation by detecting and validating four RabF motifs, as well as by our phylogenetic tree, which places RabL4 within Rabs. However, we only group RabL4 together with Rab28 as suggested in reference [105, 110] when no GTPase other than the human Rab subfamilies 1 to 45 are included (see trees in Figure S3 and Figures S4 A-B both in reference [61]). In mouse, RabL4 is not classified as being monophyletic with Rab28 (see Figure S4 C in reference [61]).

2.4.3 The Rabifier

We give some technical details about the implementation of the Rabifier which for the sake of brevity have been omitted above. For information on the computation of the confidence scores see Text S1.

In the first phase (Figure 2.1A), the profile HMMs representing the

G-protein family domain are either run manually using Perl scripts (as of June 2010) provided by Superfamily [39] and HMMER 2.3.2 [38], or in the case the sequences have been retrieved from the Superfamily database [34] the domain structure is taken directly from Superfamily. Note that Superfamily is a pure protein resource that contains proteomes predicted from genome sequences. It does not provide information about the underlying genes systematically, hence counts of how many Rab genes are present in a specific genome can generally not be derived from Superfamily. BLASTp [35] queries are performed with soft masking (parameters -F m S) and considered up to an e-value threshold of 10^{-10}. Our reference set of sequences not being Rabs is provided as Dataset S1, whereas the reference database of Rabs are the sequences accessible at `www.RabDB.org` with redundancy removed using CDHit (at a 90% sequence identity threshold) [113]. Our reference data set of Rabs covers more than just the human subfamilies, namely previously published and functionally described subfamilies from *Arabidopsis thaliana* (AtRabA1, AtRabA3-AtRabA6, AtRabC2, AtRabD1, AtRabF1, AtRabG1) [30], yeast (yptA, ypt10, ypt11), *Drosophila melanogaster* (DmRabX1-X6, DmRab9D, DmRab9F) and *C. elegans* (CeRabY6) [59]. Furthermore, as detailed in the main text we proposed a set of hypothetical subfamilies which we integrated into our reference set. The members and phylogenetic distribution of these hypothetical subfamilies can browsed directly on our web site `www.RabDB.org`. The last stage of the first phase is performed using the Motif Alignment & Search Tool (MAST) (motif finding threshold 0.0005) [114] from the MEME-suite [115], with probabilistic representations of the motifs 'IGVDF', 'KLQIW', 'RFxxxT', 'YYRGA', 'LVYDIT' [36] as input generated on our reference database of Rabs beforehand using MEME.

In the second phase (Figure 2.1B), RPS-BLAST queries [37] are performed with standard parameters and an e-value threshold of 10^{-5}, with position-specific scoring matrices (PSSM) previously generated by Ψ-BLAST

on all members of each of the Rab subfamilies present in our reference database.

2.4.4 Hypothetical subfamilies

The hypothetical subfamilies result from two distinct clustering steps. First, we clustered sequences classified as RabX by the Rabifier and belonging to the same genome at a sequence identity threshold of 70% [36]. In order to resolve the potential conflicts caused by sequences that belong to several clusters at the same time, we applied MCL [116] (inflation parameter 2.0), which resulted in a clean partition, *i.e.* non-overlapping clustering, of the sequences. In a second step, we merged the resulting clusters across genomes if at least one pair of sequences across clusters shared a sequence identity over 70%. We chose this threshold as it is the lowest which ensures meaningful clusters, that is clusters which in their majority respect taxa boundaries.

2.4.5 Phylogenetic trees

All phylogenetic trees of Rabs and GTPases presented in this article have been generated with PhyML [117], which implements a Maximum Likelihood probabilistic model, using standard parameters and 100 bootstraps. Alignments were performed with MAFFT [118], and manually edited to remove sites with deletions using Jalview [119]. The human trees have been generated using human kRas as an outgroup, the mouse trees using mouse kRas as outgroup, and the mixed tree of human and *Monosiga brevicollis* Rabs uses both human and *M. brevicollis* kRas as outgroups. Sequence accessions of all sequences can be taken from Table S2. Tree visualisations have been generated with Figtree[1]. The tree of human Rabs not displaying isoforms (see Figure 2.5, Figure 2.6) has been generated by

[1] `http://tree.bio.ed.ac.uk/software/figtree/`

removing isoforms and keeping the longest branch as representative of the corresponding subfamily.

2.4.6 Rab PCR of mouse organs and cells

Cell lines and primary cells

We decided to use both cell lines and primary cells. Cell lines are populations of cells that grow and replicate continuously, *i.e.* that have undergone genetic transformations which result in indefinite growth potential. They are prone to genotypic and phenotypic drifting, and can both lose tissue-specific functions and acquire a molecular phenotype quite different from primary cells. In contrast to that, primary cells have a finite lifespan but reflect the in vivo situation, despite their added complexity. In the following, we list the protocols we followed to obtain our cell material.

Mouse hepatoma Hepa 1-6 cells were cultured in DMEM supplemented with 10% FCS, 100 U/ml penicillin and 100 μg/ml streptomycin, maintained at 37°C in 10% CO_2 until the cells were 80% confluent and then used to extract RNA. The melanocyte cell line melan-ink was cultured in RPMI 1640 with glutamax and hepes, supplemented with 10% FCS, 0.1 mM 2-mercaptoethanol, 200 nM phorbol 12-myristate 13-acetate, 100 Uml penicillin and 100 μg/ml streptomycin at 37°C with 5% CO_2. We extracted RNA when the cells were 80% confluent. Primary dendritic cells (DC) were isolated from the bone marrow of C57BL6 mice. Femurs and tibia were removed, both ends of the bones cut and the bone marrow flushed using a syringe. Cells were cultured in plates (2-4x106 cells per plate) with 10 ml of Iscove's medium with glutamax and hepes, supplemented with 10% FCS, 100 Uml of penicillin, 100 μgml streptomycin, 5x10-5 M 2-mercaptoethanol, 0.5 mM sodium pyruvate, containing 2% of culture supernatant from X630 myeloma cells transfected with mouse GM-CSF cDNA. After 3 days of culture, new medium with GM-CSF was

added to each plate. After 7 days of culture, the non-adherent cells were collected and processed for purification with magnetic beads on MACS columns (Miltenyi Biotec). Cells were incubated with CD11c+ magnetic beads and passed through the column. The positively selected cells were pelleted by centrifugation for RNA extraction. Typically more than 90% of the positive cell population expressed the dendritic cell marker CD11c+ as determined by flow cytometry. Primary macrophages were isolated from the bone marrow of C57BL6 mice using the same procedure as for the DC and matured in M-CSF-containing media. Cells were cultured in plates (4x106 cells per plate) with 10 ml of Iscove's medium containing 30% of L929 cell-conditioned media as a source of M-CSF. After 4 days of culture, additional media with M-CSF was added. Macrophages were used after 8 days in culture for RNA extraction after removing non-adherent cells. Typically more than 90% of the cell population expressed the macrophage marker CD11b (Mac-1) as determined by flow cytometry. Primary hepatocytes were obtained from C57BL6 mice as previously described in reference [120] and used to extract RNA.

RNA isolation and cDNA synthesis

Tissue samples (Spleen, Liver, Kidney, Brain, Heart and Lung) were rapidly dissected and immediately homogenised in Trizol reagent. Total RNA was purified from the cells or tissues using a RNeasy Mini Kit (Qiagen) following the manufacturer's instructions. For cDNA synthesis 500ng of total RNA was reverse transcribed using the "First-Strand cDNA synthesis kit" (Roche) following the manufacturer's instructions.

PCR and DNA analysis of Rab GTPase expression profiles

PCR was performed on the cDNA product to assess the expression of Rab GTPases. The primers used for amplification can be taken from Table S3.

The PCR amplification was performed in a reaction mixture containing 1x green Go Taq buffer (Promega), 1 mM MgCl2, 0.2 mM of dNTP mix, 2.5 U of Taq polymerase (Promega) and specific primers at a final concentration of 0.5 μM, followed by a denaturation step of 3 min at 94°C and a 32-cycle program consisting of 94°C for 40 s, 58°C for 40 s and 72°C for 1 min. The final amplification mixture was separated in 1.2% agarose gel containing ethidium bromide and photographed under UV illumination.

2.A Supplementary text / figures

Files, missing supplementary figures and tables can be obtained from the author upon request.

This supplementary text describes the computation of the statistical confidence scores generated with each call in phase one and two of the Rabifier (see Figure 2.1). Generally, the procedure implements a naive Bayesian classifier [121]. This well studied probabilistic machine learning approach is one of the simplest yet most performant classifiers in a supervised setting.

The basis for our score is a feature vector computed for a given input. In our case, the input is a sequence to be classified, and the features are the output of different tools we feed with the input sequence (see workflow in Figure 2.1). In the following, we describe the two distinct steps necessary to perform the classification. As a prerequisite, we first establish distributions from our reference data or training set, and second, we evaluate them and combine the results to produce a single value per input sequence. This result represents the actual confidence score. The procedure is equivalent for both Rab family and Rab subfamily scores, with the difference that the Rab family score is binary and only generates two values for the classes 'Rab' and 'non-Rab', whereas the subfamily score

produces one per subfamily in our reference set. For the sake of simplicity, we describe the procedure for the binary case, however, all descriptions equivalently apply to the subfamily score.

2.A.1 Step 1

The purpose of this phase is to establish how likely certain feature values are under the assumption that the input is a Rab and that it is not. The tools we use to obtain features or measure properties of the input are BLAST [35] and MAST [114] from the MEME-suite [115] to get the sequence identity, similarity and e-value of the alignment to the best hit in our reference set, and the number and alignment e-value of the RabF motifs [36] respectively. We used the same manually compiled reference set of Rabs and sequences which are not Rabs to measure the values described above, however, to ensure we did not bias the distributions by aligning sequences against themselves we excluded this case for each sequence. The result are two times (both for Rabs and non-Rabs) five histograms (sequence identity, similarity and e-value of the alignment to the best hit in our reference set plus number and alignment e-value of the RabF motifs). For illustration purposes Figures 2.9 and 2.10 show two such histograms. Note that we did not fit any distribution to obtain true densities, but used the empirical distributions as they are shown in the Figures.

2.A.2 Step 2

Given these ten histograms, five per possible outcome (Rab or non-Rab), the computation of the confidence score given an input sequence is straightforward. Once the sequence in question has been BLASTed against the internal reference set of the Rabifier and MAST has detected the motifs and their e-value, the obtained values are evaluated under both possible outcomes with help of the density functions defined by the histograms.

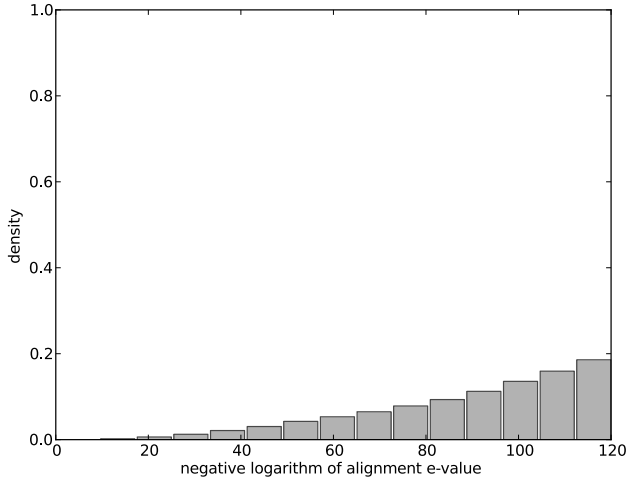


FIGURE 2.9: *Cumulative distribution of the negative logarithm of the BLAST [35] alignment e-value of our reference set of Rabs against itself*—Self hits are excluded.

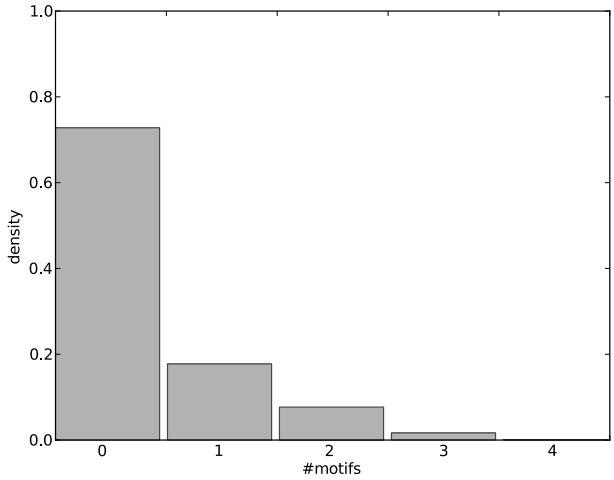


FIGURE 2.10: *Cumulative distribution of number of RabF motifs detected by MAST [114] in the reference set of non-Rabs.*

The final score is then obtained by applying Bayes formula:

$$P(C|\vec{F}) = \frac{P(\vec{F}|C)}{P(\vec{F}|C = \text{Rab}) \times P(\vec{F}|C = \text{non-Rab})}$$

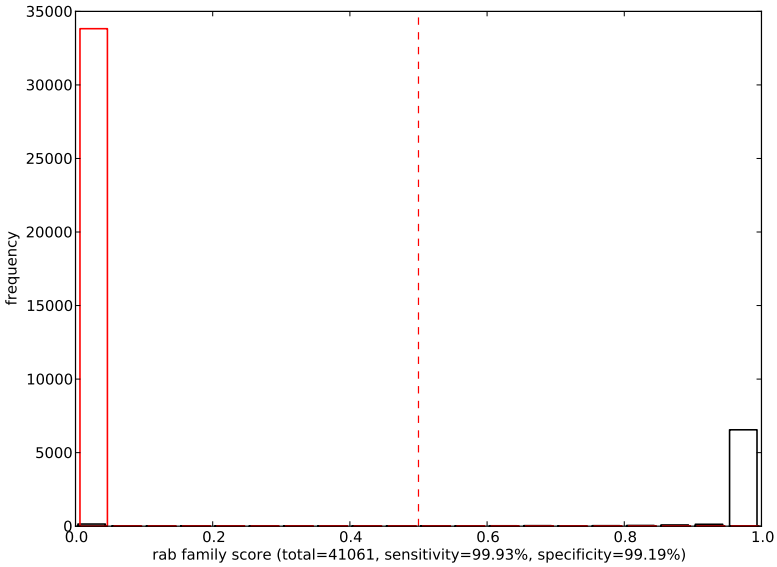
where $\vec{F}$ is the feature vector with five individual components being the sequence identity, similarity and e-value, as well as the motif count and e-value, and C are the possible outcomes or classes, *i.e.* Rab or non-Rab.

Figure 2.11 shows the distribution of scores we obtained from the application of the Rabifier to 247 genomes taken from the Superfamily database as described in the main manuscript. Note that in a true classification setting, any score below 0.5 would lead to consider a sequence as not being a Rab and vice versa. However, as presented in Figure 2.1 and unlike in the second phase, the family score does directly influence the decision of calling a sequence a Rab or not, and is to be understood as a pure confidence level. In fact, the Rabs with scores lower than 0.5 are mostly accounted for by exceptions as for example very short sequences or those with long strips of masked residues, where motif detection and alignments generally tend to fail.

Figure 2.12 summarises the result for phase 2 of the Rabifier.

Acknowledgments

We would like to thank members of the Computational Genomics Lab for helpful discussions, and in particular Renato Alves for help with setting up the database. We also thank Thiago Carvalho for critical reading of the manuscript. We wish to acknowledge one anonymous referee whose comments helped us to greatly improve the Rabifier.


35000
30000
25000
20000
15000
10000
5000
0
frequency
0.0
0.2
0.4
0.6
0.8
1.0
rab family score (total=41061, sensitivity=99.93%, specificity=99.19%)

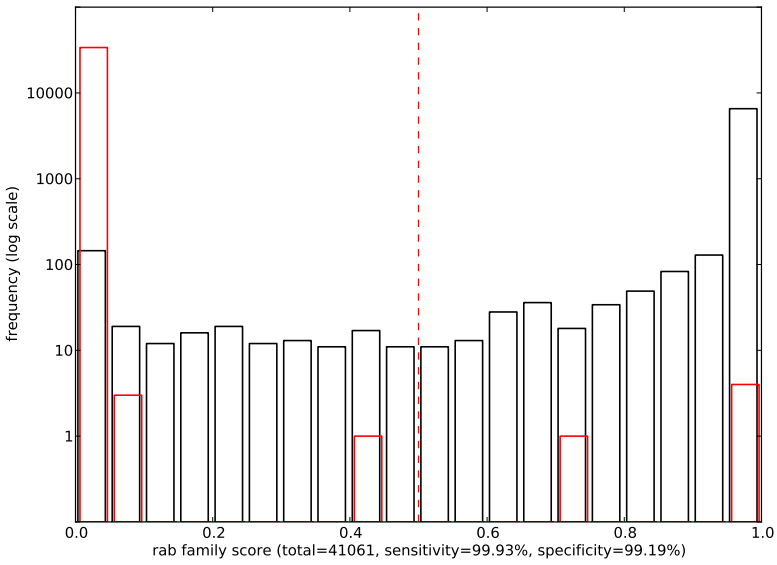

10000
1000
100
10
1
frequency (log scale)
0.0
0.2
0.4
0.6
0.8
1.0
rab family score (total=41061, sensitivity=99.93%, specificity=99.19%)

FIGURE 2.11 *(preceding page)*: *Rab family scores obtained for all G-protein family domain containing proteins from the 247 genomes described in the main text*—Black bars capture sequences the Rabifier classified as Rabs and are browsable at our public website `www.RabDB.org`, in red are those we classified as not being Rabs. The lower histogram shown the same quantities in log-scale. Sensitivity and specificity refer to the threshold at 0.5 marked by the red dashed line.

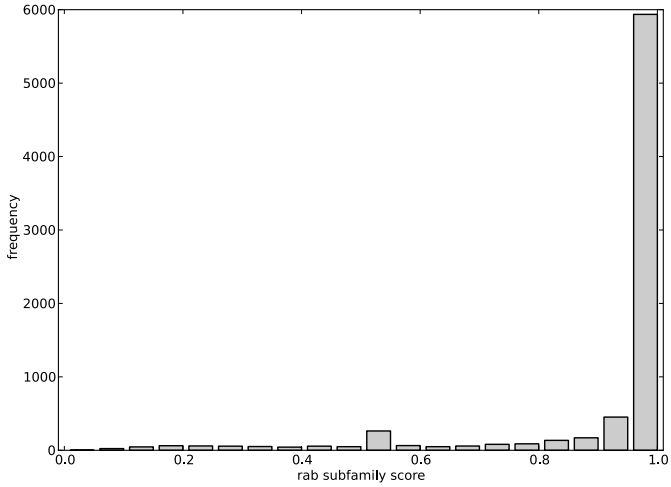


FIGURE 2.12: *Distribution of subfamily scores of the highest scoring subfamily for all Rabs in our database*—Database accessible at `www.RabDB.org`.

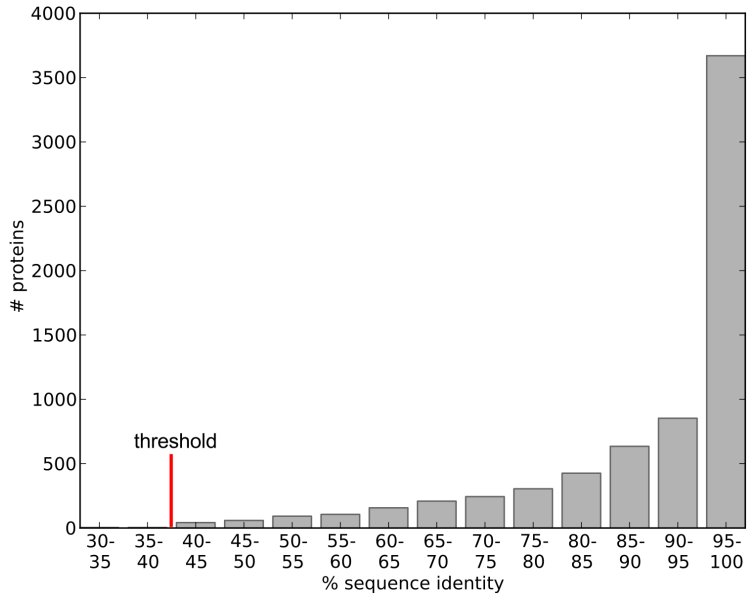


FIGURE 2.13: *Sequence identity to best hit within same subfamily*— Histogram of sequence identity of all sequences in our reference database to their respective best hit within the same subfamily (itself excluded). Subfamilies can contain sequences from organisms anywhere in the eukaryotic tree. The threshold is the minimal required identity for a sequence to be attributed to the subfamily of its best hit (see Figure 2.1). It is chosen to minimise the number of times a sequence is annotated as belonging to the unspecified subfamily RabX although it is a member of the same subfamily as its best hit.

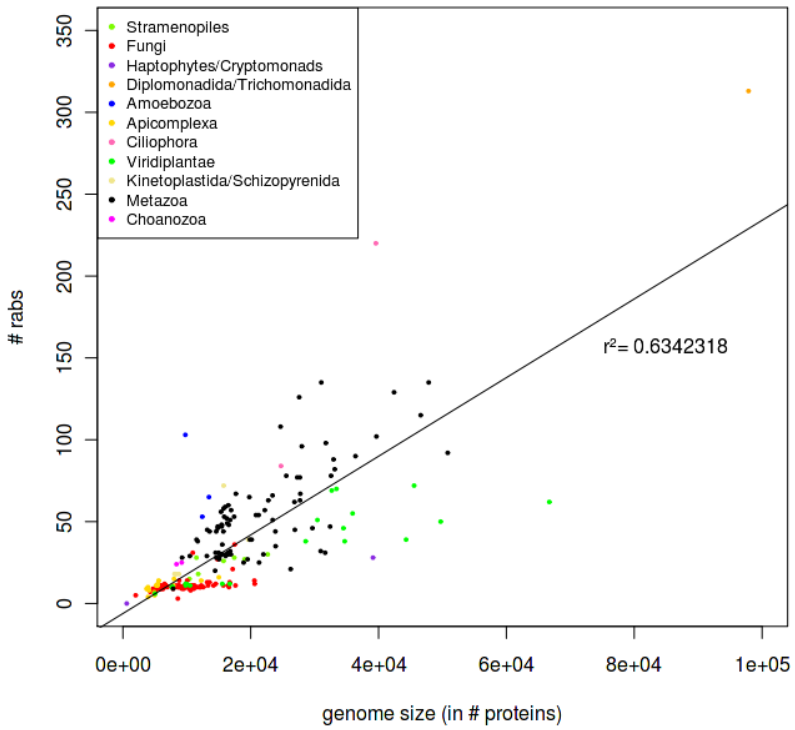


FIGURE 2.14: *Linear regression of number of Rabs against genome size*— Data consists of the 247 genomes profiled by the Rabifier. The taxa are shown in different colours.

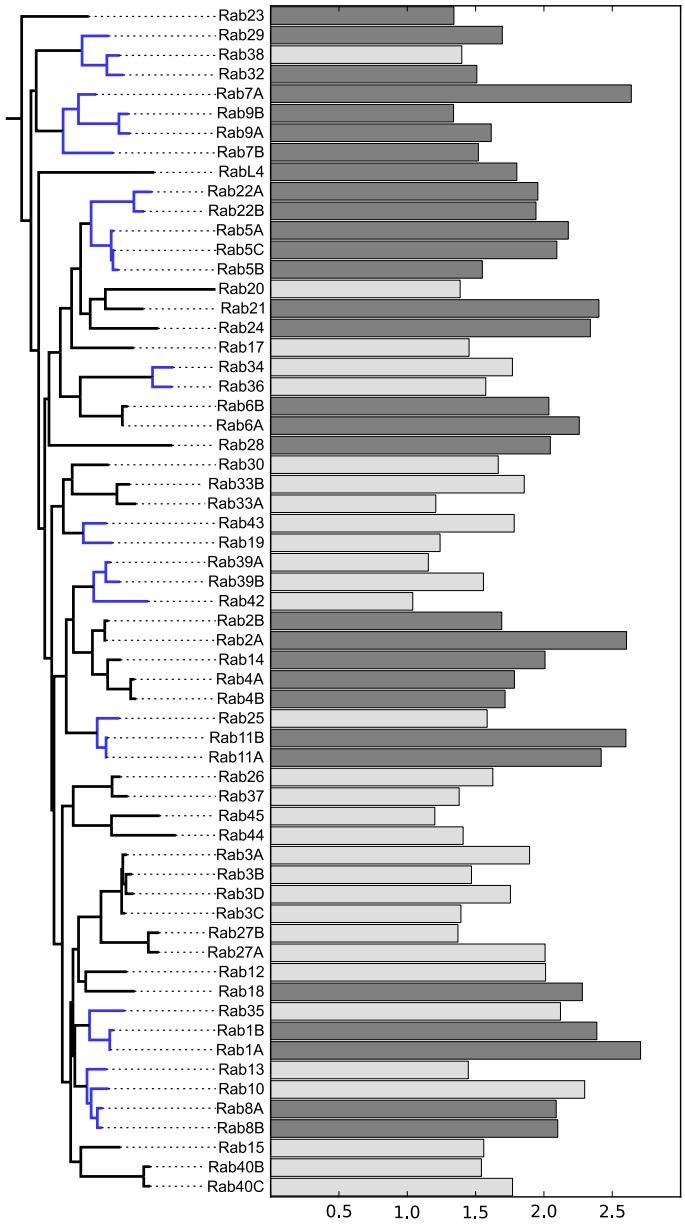

Rab23
Rab29
Rab38
Rab32
Rab7A
Rab9B
Rab9A
Rab7B
RabL4
Rab22A
Rab22B
Rab5A
Rab5C
Rab5B
Rab20
Rab21
Rab24
Rab17
Rab34
Rab36
Rab6B
Rab6A
Rab28
Rab30
Rab33B
Rab33A
Rab43
Rab19
Rab39A
Rab39B
Rab42
Rab2B
Rab2A
Rab14
Rab4A
Rab4B
Rab25
Rab11B
Rab11A
Rab26
Rab37
Rab45
Rab44
Rab3A
Rab3B
Rab3D
Rab3C
Rab27B
Rab27A
Rab12
Rab18
Rab35
Rab1B
Rab1A
Rab13
Rab10
Rab8A
Rab8B
Rab15
Rab40B
Rab40C
0.5
1.0
1.5
2.0
2.5

FIGURE 2.15 *(preceding page)*: *Microarrays of Rabs in mouse tissues*—The average expression across the mouse tissues (cell lines not included) in the same data as shown in Figure S5 in reference [61].

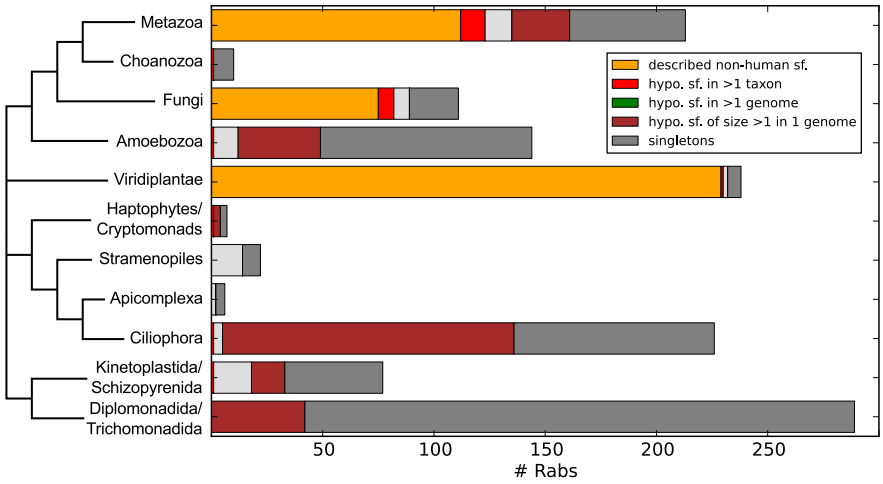


FIGURE 2.16: *Distribution of Rabs belonging to non-human subfamilies*—The histogram details for each taxon how we classified those Rabs not belonging to human subfamilies. Subfamilies falling into the orange category have been previously described in the literature, whereas all other subfamilies result from clustering of the sequences as described in Section 2.4. See Figure 2.4 for an overview of the number of subfamilies in each category.

References

[1] Niles Eldredge and Joel Cracraft. *Phylogenetic Patterns and the Evolutionary Process—Method and Theory in Comparative Biology.* New York: Columbia University Press, Dec. 1980.

[2] Meir Aridor and Lisa A Hannan. "Traffic jam: a compendium of human diseases that affect intracellular transport processes". In: *Traffic* 1.11 (Nov. 2000), pp. 836–851.

[3] Meir Aridor and Lisa A Hannan. "Traffic jams II: an update of diseases of intracellular transport". In: *Traffic* 3.11 (Nov. 2002), pp. 781–790.

[4] Miguel C Seabra, Emilie H Mules, and Alistair N Hume. "Rab GTPases, intracellular traffic and disease". In: *Trends in molecular medicine* 8.1 (2002), pp. 23–30.

[5] Shreya Mitra, Kwai W Cheng, and Gordon B Mills. "Rab GTPases Implicated in Inherited and Acquired Disorders". In: *Seminars in Cell & Developmental Biology* 22 (2011), pp. 57–68.

[6] Roshan Agarwal et al. "The emerging role of the RAB25 small GTPase in cancer". In: *Traffic* 10.11 (Nov. 2009), pp. 1561–1568.

[7] Uri David Akavia et al. "An integrated approach to uncover drivers of cancer". In: *Cell* 143.6 (Dec. 2010), pp. 1005–1017.

[8] Wan Jie Chia and Bor Luen Tang. "Emerging roles for Rab family GTPases in human cancer". In: *Biochimica Et Biophysica Acta* 1795.2 (Apr. 2009), pp. 110–116.

[9] Kwai Wa Cheng et al. "The RAB25 small GTPase determines aggressiveness of ovarian and breast cancers". In: *Nature medicine* 10.11 (Nov. 2004), pp. 1251–1256.

[10] Stefan S Weber, Curdin Ragaz, and Hubert Hilbi. “Pathogen trafficking pathways and host phosphoinositide metabolism”. In: *Molecular Microbiology* 71.6 (Mar. 2009), pp. 1341–1352.

[11] Amit P Bhavsar, Julian A Guttman, and B Brett Finlay. “Manipulation of host-cell pathways by bacterial pathogens”. In: *Nature* 449.7164 (Oct. 2007), pp. 827–834.

[12] Nicolas Frei dit Frey and Silke Robatzek. “Trafficking vesicles: pro or contra pathogens?” In: *Current Opinion in Plant Biology* 12.4 (Aug. 2009), pp. 437–443.

[13] John H Brumell and Marci A Scidmore. “Manipulation of rab GTPase function by intracellular bacterial pathogens”. In: *Microbiology and Molecular Biology Reviews : MMBR* 71.4 (Dec. 2007), pp. 636–652.

[14] Andrew Brighouse, Joel B Dacks, and Mark C Field. “Rab protein evolution and the history of the eukaryotic endomembrane system”. In: *Cellular and Molecular Life Sciences : CMLS* 67.20 (Oct. 2010), pp. 3449–3465.

[15] Tomer Avidor-Reiss et al. “Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis”. In: *Cell* 117.4 (May 2004), pp. 527–539.

[16] H Jomaa et al. “Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs”. In: *Science* 285.5433 (Sept. 1999), pp. 1573–1576.

[17] Joel B Dacks and Mark C Field. “Evolution of the eukaryotic membrane-trafficking system: origin, tempo and mode”. In: *Journal of Cell Science* 120.17 (Sept. 2007), pp. 2977–2985.

[18] Gáspár Jékely. “Small GTPases and the evolution of the eukaryotic cell”. In: *BioEssays* 25.11 (Nov. 2003), pp. 1129–1138.

[19] Harald Stenmark. “Rab GTPases as coordinators of vesicle traffic”. In: *Nature Reviews Molecular Cell Biology* 10.8 (Aug. 2009), pp. 513–525.

[20] Stéphanie Miserey-Lenkei et al. “Rab and actomyosin-dependent fission of transport vesicles at the Golgi complex”. In: *Nature Cell Biology* 12.7 (July 2010), pp. 645–654.

[21] Bianka L Grosshans, Darinel Ortiz, and Peter J Novick. “Rabs and their effectors: achieving specificity in membrane traffic”. In: *Proceedings of the National Academy of Sciences of the United States of America* 103.32 (Aug. 2006), pp. 11821–11827.

[22] Cemal Gurkan, Atanas V Koulov, and William E Balch. “An evolutionary perspective on eukaryotic membrane trafficking”. In: *Advances in experimental medicine and biology* 607 (2007), pp. 73–83.

[23] H Haubruck et al. “The ras-related mouse ypt1 protein can functionally replace the YPT1 gene product in yeast”. In: *The EMBO Journal* 8.5 (May 1989), pp. 1427–1432.

[24] José B Pereira-Leal. “The Ypt/Rab family and the evolution of trafficking in fungi”. In: *Traffic* 9.1 (2008), pp. 27–38.

[25] Philippe Abbal et al. “Molecular characterization and expression analysis of the Rab GTPase family in Vitis vinifera reveal the specific expression of a VvRabA protein”. In: *Journal of Experimental Botany* 59.9 (2008), pp. 2403–2416.

[26] Lydia J Bright et al. “Comprehensive analysis reveals dynamic and evolutionary plasticity of Rab GTPases and membrane traffic in Tetrahymena thermophila”. In: *PLoS Genetics* 6.10 (2010), e1001155.

[27] Kalpana Lal et al. “Identification of a very large Rab GTPase family in the parasitic protozoan Trichomonas vaginalis”. In: *Molecular and Biochemical Parasitology* 143.2 (Oct. 2005), pp. 226–235.

[28] Yumiko Saito-Nakano et al. “Marked amplification and diversification of products of ras genes from rat brain, Rab GTPases, in the ciliates Tetrahymena thermophila and Paramecium tetraurelia”. In: *The Journal of Eukaryotic Microbiology* 57.5 (2010), pp. 389–399.

[29] Yumiko Saito-Nakano et al. “The diversity of Rab GTPases in Entamoeba histolytica”. In: *Experimental parasitology* 110.3 (June 2005), pp. 244–252.

[30] Stephen Rutherford and Ian Moore. “The Arabidopsis Rab GTPase family: another enigma variation”. In: *Current Opinion in Plant Biology* 5.6 (Dec. 2002), pp. 518–528.

[31] John P Ackers, Vivek Dhir, and Mark C Field. “A bioinformatic analysis of the RAB genes of Trypanosoma brucei”. In: *Molecular and Biochemical Parasitology* 141.1 (May 2005), pp. 89–97.

[32] Emmanuel Quevillon et al. “The Plasmodium falciparum family of Rab GTPases”. In: *Gene* 306 (Mar. 2003), pp. 13–25.

[33] Alfonso Valencia et al. “The Ras protein family: evolutionary tree and role of conserved amino acids”. In: *Biochemistry* 30.19 (May 1991), pp. 4637–4648.

[34] Derek Wilson et al. “SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny”. In: *Nucleic Acids Research* 37.Database issue (2009), pp. D380–6.

[35] S F Altschul et al. “Basic local alignment search tool”. In: *Journal of Molecular Biology* 215.3 (Oct. 1990), pp. 403–410.

[36] José B Pereira-Leal and Miguel C Seabra. “The mammalian Rab family of small GTPases: definition of family and subfamily sequence motifs suggests a mechanism for functional specificity in the Ras superfamily”. In: *Journal of Molecular Biology* 301.4 (Aug. 2000), pp. 1077–1087.

[37] Stephen F Altschul et al. “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”. In: *Nucleic Acids Research* 25.17 (Sept. 1997), pp. 3389–3402.

[38] Sean R Eddy. “Hidden Markov models”. In: *Current Opinion in Structural Biology* 6.3 (June 1996), pp. 361–365.

[39] Julian Gough and Cyrus Chothia. “SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments”. In: *Nucleic Acids Research* 30.1 (Jan. 2002), pp. 268–272.

[40] Tom Fawcett. “An introduction to ROC analysis”. In: *Pattern Recognition Letters* 27 (Apr. 2006), pp. 861–874.

[41] James A Hanley and Barabra J McNeil. “The meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve”. In: *Radiology* 143 (1982), pp. 29–36.

[42] Aron Marchler-Bauer et al. “CDD: a Conserved Domain Database for the functional annotation of proteins”. In: *Nucleic Acids Research* 39.Database issue (2011), pp. D225–9.

[43] Sandra L Baldauf. “Phylogeny for the faint of heart: a tutorial”. In: *Trends in Genetics : TIG* 19.6 (June 2003), pp. 345–351.

[44] Fabien Burki, Kamran Shalchian-Tabrizi, and Jan Pawlowski. “Phylogenomics reveals a new ’megagroup’ including most photosynthetic eukaryotes”. In: *Biology letters* 4.4 (Aug. 2008), pp. 366–369.

[45] Christine Vogel and Cyrus Chothia. “Protein family expansions and biological complexity”. In: *PLoS Computational Biology* 2.5 (May 2006), e48.

[46] I King Jordan et al. “Lineage-specific gene expansions in bacterial and archaeal genomes”. In: *Genome Research* 11.4 (Apr. 2001), pp. 555–565.

[47] Ravindra Pushker, Alex Mira, and Francisco Rodríguez-Valera. “Comparative genomics of gene-family size in closely related bacteria”. In: *Genome Biology* 5.4 (2004), R27.

[48] Andrés Moya et al. “Learning how to live together: genomic insights into prokaryote-animal symbioses”. In: *Nature Reviews Genetics* 9.3 (Mar. 2008), pp. 218–229.

[49] H Field et al. “Complexity of trypanosomatid endocytosis pathways revealed by Rab4 and Rab5 isoforms in Trypanosoma brucei”. In: *The Journal of biological chemistry* 273.48 (Nov. 1998), pp. 32102–32110.

[50] Lucas Pelkmans et al. “Caveolin-stabilized membrane domains as multifunctional transport and sorting devices in endocytic membrane traffic”. In: *Cell* 118.6 (Sept. 2004), pp. 767–780.

[51] Dmitry Poteryaev et al. “Identification of the switch in early-to-late endosome transition”. In: *Cell* 141.3 (Apr. 2010), pp. 497–508.

[52] Maria Kauppi et al. “The small GTPase Rab22 interacts with EEA1 and controls endosomal membrane trafficking”. In: *Journal of Cell Science* 115.5 (Mar. 2002), pp. 899–911.

[53] R Mesa et al. “Rab22a affects the morphology and function of the endocytic pathway”. In: *Journal of Cell Science* 114.22 (Nov. 2001), pp. 4041–4049.

[54] M A Barbieri et al. “Epidermal growth factor and membrane trafficking. EGF receptor activation of endocytosis requires Rab5a”. In: *The Journal of Cell Biology* 151.3 (Oct. 2000), pp. 539–550.

[55] Chris P Ponting. “The functional repertoires of metazoan genomes”. In: *Nature Reviews Genetics* 9.9 (Sept. 2008), pp. 689–698.

[56] Mark S Springer and William J Murphy. “Mammalian evolution and biomedicine: new views from phylogeny”. In: *Biological reviews of the Cambridge Philosophical Society* 82.3 (Aug. 2007), pp. 375–392.

[57] Marek Eliáš. “Patterns and processes in the evolution of the eukaryotic endomembrane system”. In: *Molecular Membrane Biology* 27.8 (Nov. 2010), pp. 469–489.

[58] Kamran Shalchian-Tabrizi et al. “Multigene phylogeny of choanozoa and the origin of animals”. In: *PLoS ONE* 3.5 (2008), e2098.

[59] J B Pereira-Leal and M C Seabra. “Evolution of the Rab family of small GTP-binding proteins”. In: *Journal of Molecular Biology* 313.4 (Nov. 2001), pp. 889–901.

[60] Lillian K Fritz-Laylin et al. “The genome of Naegleria gruberi illuminates early eukaryotic versatility”. In: *Cell* 140.5 (Mar. 2010), pp. 631–642.

[61] Yoan Diekmann et al. “Thousands of Rab GTPases for the Cell Biologist”. In: *PLoS Computational Biology* 7.10 (Oct. 2011), e1002217.

[62] Eugene V Koonin. “The incredible expanding ancestor of eukaryotes”. In: *Cell* 140.5 (Mar. 2010), pp. 606–608.

[63] Hongmin Qin et al. “Intraflagellar transport protein 27 is a small G protein involved in cell-cycle control”. In: *Current Biology* 17.3 (Feb. 2007), pp. 193–202.

[64] Zita Carvalho-Santos et al. “Stepwise evolution of the centriole-assembly pathway”. In: *Journal of Cell Science* 123.9 (May 2010), pp. 1414–1426.

[65] Matthew E Hodges et al. “Reconstructing the evolutionary history of the centriole from protein components”. In: *Journal of Cell Science* 123.9 (May 2010), pp. 1407–1413.

[66] Christina Wasmeier et al. “Rab38 and Rab32 control post-Golgi trafficking of melanogenic enzymes.” In: *The Journal of Cell Biology* 175.2 (Oct. 2006), pp. 271–281.

[67] Neal M Alto, Jacquelyn Soderling, and John D Scott. “Rab32 is an A-kinase anchoring protein and participates in mitochondrial dynamics”. In: *The Journal of Cell Biology* 158.4 (Aug. 2002), pp. 659–668.

[68] Michael Bui et al. “Rab32 modulates apoptosis onset and mitochondria-associated membrane (MAM) properties”. In: *The Journal of biological chemistry* 285.41 (Oct. 2010), pp. 31590–31602.

[69] George B Kyei et al. “Rab14 is critical for maintenance of Mycobacterium tuberculosis phagosome maturation arrest”. In: *The EMBO Journal* 25.22 (Nov. 2006), pp. 5250–5259.

[70] Tassula Proikas-Cezanne et al. “Rab14 is part of the early endosomal clathrin-coated TGN microdomain”. In: *FEBS Letters* 580.22 (Oct. 2006), pp. 5241–5246.

[71] Thomas Cavalier-Smith. “The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa”. In: *International journal of systematic and evolutionary microbiology* 52.2 (Mar. 2002), pp. 297–354.

[72] John Maynard Smith and Eörs Szathmáry. *The Major Transitions in Evolution*. Oxford University Press, Oct. 1997.

[73] Antonis Rokas. “The molecular origins of multicellular transitions”. In: *Current Opinion in Genetics & Development* 18.6 (Dec. 2008), pp. 472–478.

[74] Nicole King et al. “The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans”. In: *Nature* 451.7180 (Feb. 2008), pp. 783–788.

[75] Ian G Ganley et al. “Rab9 GTPase regulates late endosome size and requires effector interaction for its stability”. In: *Molecular Biology of the Cell* 15.12 (Dec. 2004), pp. 5420–5430.

[76] A G Rodriguez-Gabin et al. “Role of rRAB22b, an oligodendrocyte protein, in regulation of transport of vesicles from trans Golgi to endocytic compartments”. In: *Journal of neuroscience research* 66.6 (Dec. 2001), pp. 1149–1160.

[77] Monika Abedin and Nicole King. “Diverse evolutionary paths to cell adhesion”. In: *Trends in Cell Biology* 20.12 (Dec. 2010), pp. 734–742.

[78] M Roberts et al. “PDGF-regulated rab4-dependent recycling of alphavbeta3 integrin from early endosomes is necessary for cell adhesion and spreading”. In: *Current Biology* 11.18 (Sept. 2001), pp. 1392–1402.

[79] Aimee M Powelka et al. “Stimulation-dependent recycling of integrin β1 regulated by ARF6 and Rab11”. In: *Traffic* 5.1 (Jan. 2004), pp. 20–36.

[80] Teijo Pellinen et al. “Small GTPase Rab21 regulates cell adhesion and controls endosomal traffic of beta1-integrins”. In: *The Journal of Cell Biology* 173.5 (June 2006), pp. 767–780.

[81] Patrick T Caswell et al. “Rab25 associates with $\alpha 5\beta 1$ integrin to promote invasive migration in 3D microenvironments”. In: *Developmental Cell* 13.4 (Oct. 2007), pp. 496–510.

[82] Toshihiro Kimura et al. “Involvement of the Ras-Ras-activated Rab5 guanine nucleotide exchange factor RIN2-Rab5 pathway in the hepatocyte growth factor-induced endocytosis of E-cadherin”. In: *Journal of Biological Chemistry* 281.15 (Apr. 2006), pp. 10598–10609.

[83] Marieke A M Frasa et al. “Armus is a Rac1 effector that inactivates Rab7 and regulates E-cadherin degradation”. In: *Current Biology* 20.3 (Feb. 2010), pp. 198–208.

[84] Mikhail V Khvotchev et al. “Divergent functions of neuronal Rab11b in Ca2+-regulated versus constitutive exocytosis”. In: *The Journal of neuroscience* 23.33 (Nov. 2003), pp. 10531–10539.

[85] M Rupnik et al. “Distinct role of Rab3A and Rab3B in secretory activity of rat melanotrophs”. In: *American Journal of Physiology, Cell Physiology* 292.1 (2007), pp. C98–105.

[86] Oliver M Schlüter et al. “Localization versus function of Rab3 proteins—Evidence for a common regulatory role in controlling fusion”. In: *The Journal of biological chemistry* 277.43 (Oct. 2002), pp. 40919–40929.

[87] Takashi Tsuboi and Mitsunori Fukuda. “Rab3A and Rab27A cooperatively regulate the docking step of dense-core vesicle exocytosis in PC12 cells”. In: *Journal of Cell Science* 119.11 (May 2006), pp. 2196–2203.

[88] S Yoshie et al. “Expression, characterization, and localization of Rab26, a low molecular weight GTP-binding protein, in the rat parotid gland”. In: *Histochemistry and cell biology* 113.4 (Apr. 2000), pp. 259–263.

[89] Duarte C Barral et al. “Functional redundancy of Rab27 proteins and the pathogenesis of Griscelli syndrome”. In: *The Journal of clinical investigation* 110.2 (July 2002), pp. 247–257.

[90] Clare E Futter. “The molecular regulation of organelle transport in mammalian retinal pigment epithelial cells”. In: *Pigment cell research* 19.2 (Apr. 2006), pp. 104–111.

[91] Tanya Tolmachova et al. “Rab27b regulates number and secretion of platelet dense granules”. In: *Proceedings of the National Academy of Sciences of the United States of America* 104.14 (Apr. 2007), pp. 5872–5877.

[92] E S Masuda et al. “Rab37 is a novel mast cell specific GTPase localized to secretory granules”. In: *FEBS Letters* 470.1 (Mar. 2000), pp. 61–64.

[93] Christine E Becker, Emma M Creagh, and Luke A J O’Neill. “Rab39a binds caspase-1 and is required for caspase-1-dependent interleukin-1beta secretion”. In: *The Journal of biological chemistry* 284.50 (Dec. 2009), pp. 34531–34537.

[94] Sebastian Schuck et al. “Rab10 is involved in basolateral transport in polarized Madin-Darby canine kidney cells”. In: *Traffic* 8.1 (2007), pp. 47–60.

[95] Selma Y Dejgaard et al. “Rab18 and Rab43 have key roles in ER-Golgi trafficking”. In: *Journal of Cell Science* 121.Pt 16 (Aug. 2008), pp. 2768–2781.

[96] H P de Leeuw et al. “Small GTP-binding proteins in human endothelial cells”. In: *British journal of haematology* 103.1 (Oct. 1998), pp. 15–19.

[97] Rita Sinka et al. “Golgi coiled-coil proteins contain multiple binding sites for Rab family G proteins”. In: *The Journal of Cell Biology* 183.4 (Nov. 2008), pp. 607–615.

[98] Chloe Thomas, Raphaël Rousset, and Stéphane Noselli. “JNK signalling influences intracellular trafficking during Drosophila morphogenesis through regulation of the novel target gene Rab30”. In: *Developmental Biology* 331.2 (July 2009), pp. 250–260.

[99] R Valsdottir et al. “Identification of rabaptin-5, rabex-5, and GM130 as putative effectors of rab33b, a regulator of retrograde traffic between the Golgi apparatus and ER”. In: *FEBS Letters* 508.2 (Nov. 2001), pp. 201–209.

[100] Neil M Goldenberg, Sergio Grinstein, and Mel Silverman. “Golgi-bound Rab34 is a novel member of the secretory pathway”. In: *Molecular Biology of the Cell* 18.12 (Dec. 2007), pp. 4762–4771.

[101] Alexander K Haas et al. “Analysis of GTPase-activating proteins: Rab1 and Rab43 are key Rabs required to maintain a functional Golgi complex in human cells”. In: *Journal of Cell Science* 120.17 (Sept. 2007), pp. 2997–3010.

[102] Hideki Innan and Fyodor A Kondrashov. “The evolution of gene duplications: classifying and distinguishing between models”. In: *Nature Reviews Genetics* 11.2 (Feb. 2010), pp. 97–108.

[103] Shiri Freilich et al. “Relating tissue specialization to the differentiation of expression of singleton and duplicate mouse proteins”. In: *Genome Biology* 7.10 (2006), R89.

[104] Cemal Gurkan et al. “Large-scale profiling of Rab GTPase trafficking networks: the membrome”. In: *Molecular Biology of the Cell* 16.8 (Aug. 2005), pp. 3847–3864.

[105] John Colicelli. “Human RAS superfamily proteins and related GTPases”. In: *Science’s STKE* 2004.250 (Sept. 2004), RE13.

[106] Krister Wennerberg, Kent L Rossman, and Channing J Der. “The Ras Superfamily at a Glance”. In: *Journal of Cell Science* 118.5 (Mar. 2005), pp. 843–846.

[107] Joanne Young, Julie Ménétrey, and Bruno Goud. “RAB6C is a retrogene that encodes a centrosomal protein involved in cell cycle progression”. In: *Journal of Molecular Biology* 397.1 (Mar. 2010), pp. 69–88.

[108] Ana Claudia Marques et al. “Functional diversification of duplicate genes through subcellular adaptation of encoded proteins”. In: *Genome Biology* 9.3 (2008), R54.

[109] S Ashley Byun-McKay and R Geeta. “Protein subcellular relocalization: a new perspective on the origin of novel genes”. In: *Trends in Ecology and Evolution* 22.7 (July 2007), pp. 338–344.

[110] Samantha L Schwartz et al. “Rab GTPases at a glance”. In: *Journal of Cell Science* 120.Pt 22 (Nov. 2007), pp. 3905–3910.

[111] Jomon Joseph. “Ran at a glance”. In: *Journal of Cell Science* 119.Pt 17 (Sept. 2006), pp. 3481–3484.

[112] John F Dishinger et al. “Ciliary entry of the kinesin-2 motor KIF17 is regulated by importin-beta2 and RanGTP”. In: *Nature Cell Biology* 12.7 (July 2010), pp. 703–710.

[113] W Li and A Godzik. “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences”. In: *Bioinformatics* 22.13 (June 2006), pp. 1658–1659.

[114] Timothy L Bailey and M Gribskov. “Combining evidence using p-values: application to sequence homology searches”. In: *Bioinformatics* 14.1 (1998), pp. 48–54.

[115] T L Bailey and C Elkan. “Fitting a mixture model by expectation maximization to discover motifs in biopolymers”. In: *Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB)* 2 (1994), pp. 28–36.

[116] Stijn van Dongen. “A cluster algorithm for graphs”. In: *Technical report* INS-R0010 (2000), pp. 1–42.

[117] Stéphane Guindon and Olivier Gascuel. “A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood”. In: *Systematic Biology* 52.5 (Oct. 2003), pp. 696–704.

[118] Kazutaka Katoh and Hiroyuki Toh. “Recent developments in the MAFFT multiple sequence alignment program”. In: *Briefings in Bioinformatics* 9.4 (July 2008), pp. 286–298.

[119] Andrew M Waterhouse et al. “Jalview Version 2—a multiple sequence alignment editor and analysis workbench”. In: *Bioinformatics* 25.9 (May 2009), pp. 1189–1191.

[120] Lígia A Gonçalves, Ana M Vigário, and Carlos Penha-Gonçalves. “Improved isolation of murine hepatocytes for in vitro malaria liver stage studies”. In: *Malaria journal* 6 (2007), p. 169.

[121] Tom M Mitchell. “Generative and discriminative classifiers: naive bayes and logistic regression”. In: *Machine Learning*. McGraw Hill, Jan. 2010.

Author contribution: I conceived, designed and performed the experiments, analysed the data and wrote the chapter.

Chapter 3

Phylogeny and the inference of episodic positive selection

"Detecting selection needs comparative data" [1]

—Nielsen, Hubisz, 2005

"Comparative biological analysis can be carried out only in the context of a phylogeny." [2]

—Thornton, DeSalle, 2000

Abstract Chapter 3

The Branch-Site Test of Positive Selection is a standard approach to detect past episodic positive selection in *a priori* specified branches of a gene phylogeny. Here, we ask if errors in the topology of the gene tree have any influence on its ability to infer positively selected sites. Using simulated sequences, we compare the results obtained for the true and erroneous topologies, and find a strong linear effect on the ability to predict sites if an erroneous tree topology changes how long sequences are inferred to have experienced selection. Moreover, reanalysing a previously published data set we show that the choice of gene tree also alters the results obtained for real-world sequences. This is the first time a clear effect of the gene tree topology on the inference of positive selection is demonstrated. We conclude that the gene tree is an important factor for the branch-site analysis of positive selection so far unrecognised.

3.1 Introduction

The branch-site test of positive selection (BSPS) [3, 4] is a standard approach to detect sites that evolved under episodic positive selection, *i.e.* in a subset of branches in a phylogeny. It is based on a codon model of sequence evolution [5] with an explicit parameter ω representing the nonsynonymous to synonymous substitution rate ratio (dN/dS), which is commonly interpreted as evidence for positive selection when above one. Given a multiple sequence alignment (MSA), a gene tree relating these sequences, and a partition of the branches into so called 'foreground' and 'background', a likelihood ratio test (LRT) determines if a codon model with ω greater one for some sites in the foreground branches fits the data significantly better than a null model with no site above one. If this is the case, the actual sites most likely evolving under positive selection in the foreground branches can be determined in a second step by a Bayes Empirical Bayes (BEB) procedure [6]. For each site, BEB outputs the probability that it evolved under selection in the foreground, with values above 0.95 generally considered as significant.

The performance of the BSPS, usually defined by the type I and type II errors of the LRT, has been assessed mostly using simulations, as in general the true history of selection pressures cannot be known.

The most common scheme is to generate different sets of sequences under a varying simulation parameter and compare the performance of the BSPS on each of those data sets. Examples of variables that have been examined are sequence length, strength of positive selection, proportion of sites under positive selection [7], indels and alignment errors [8], synonymous substitution saturation and variations in GC-content [9]. In contrast, to the best of our knowledge the effect of the gene tree on BSPS performance has not been systematically quantified. At least for

site models, the gene tree does not seem to be of great concern as long as it is "reasonably good" [10], *i.e.* inferred from the data for example by Maximum Likelihood (ML).

Here, we ask if and how the gene tree topology impacts on the performance of the BSPS. Note that we define performance here as the ability to retrieve the actual sites under positive selection by BEB and not as the errors committed by the LRT. Except for a short paragraph in reference [8], BEB performance has so far only been measured in the context of site models [6, 11–13]. However, although statistically "[i]dentifying amino acid residues under positive selection along the lineages of interest is clearly much more difficult than testing for the presence of such sites" [3], the actual sites are often most useful to molecular biologists (see for example [14] and the references therein). Therefore, our performance metric is relevant in practice and novel in the context of the BSPS.

3.2 Results / Discussion

First, we measured the performance of BEB on simulated sequences given the true topology. This establishes the baseline against which the results on erroneous topologies can be compared.

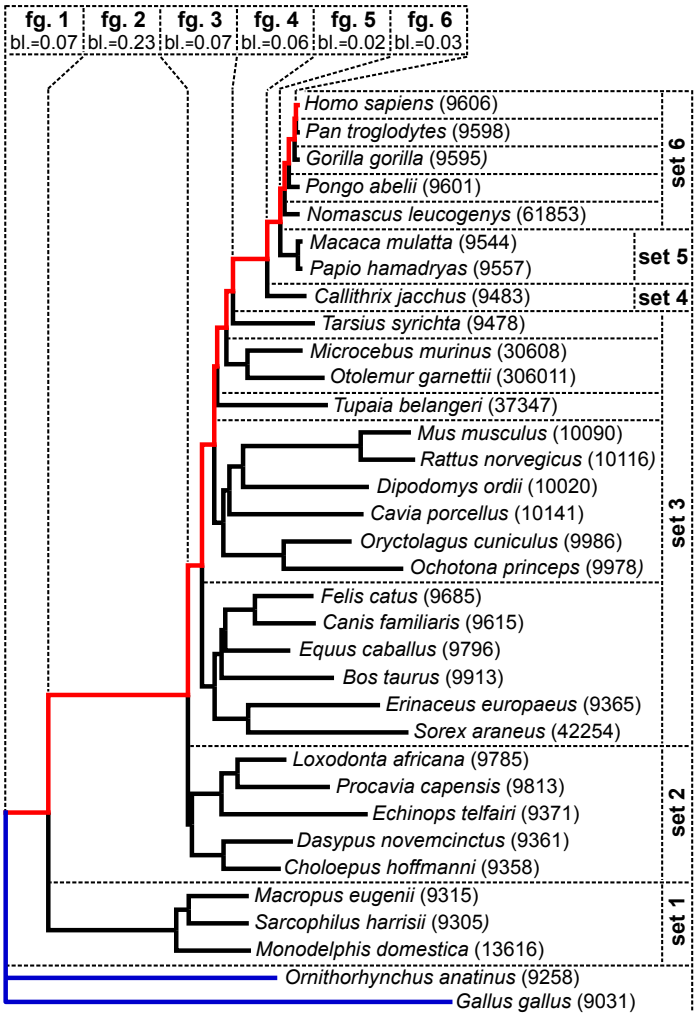
We simulated sequences along the Ensembl Compara [15] species tree for selected mammals and chicken shown in Figure 3.1 in order to ensure a realistic tree topology. We generated eight independent replica for 21 different foreground branches (each contiguous subset of labeled branches in Figure 3.1, *i.e.* {fg1}, ..., {fg6}, {fg1, fg2}, ..., {fg5, fg6}, ..., {fg1, fg2, fg3, fg4, fg5, fg6}), resulting in 168 sets of 32 sequences in total. Each time, the foreground was simulated with 30% of the sites at a dN/dS above one. Although shown to be of critical importance for the BSPS [8],

we did not simulate insertions or deletions (indels) as we did not want to confound the effect of tree topology alone. More details on the simulation procedure are given in the Materials and Methods Section 3.4. For each of the sets of sequences we determined the sites inferred to be under positive selection in the foreground branch by BEB (at site-specific posterior probability > 0.95) [6] using PAML [16]. We compared those to the true simulated sites and summarised the elements of the confusion matrix by computing sensitivity, specificity and Matthew's Correlation Coefficient (MCC) (definitions given in Section 3.4). By default, we average derivations of the confusion matrix over the eight replica to mask the variation across replica.

The overall distribution of sensitivity, specificity and MCC values is shown in Figure 3.2. It is obvious that specificity is generally high, *i.e.* very few sites are erroneously inferred to have evolved under selection in the foreground branches. Hence, in the following we focus on sensitivity as a measure of BEB performance, which has the advantage of being easier to interpret (analogous to the power of a statistical test, or verbally: of all sites under selection, what fraction has been correctly found) than for example MCC. Interestingly, at least for the more lenient simulation scenario without indels followed here, we attain higher power than the previous report of less than 1% [8].

In Figure 3.3, we detail the sensitivity for each of the 21 different foreground branches to visualise potential differences in performance depending on the foreground. Indeed, we find marked differences, with almost no sites inferred for foreground branch six and one (referring to the labels from Figure 3.1), and highest power for foregrounds stretching (nearly) the entire path from root to leaf.

These performance differences can be explained in terms of properties of the foreground branch. For the power of the LRT, two aspects have previously been shown to be important: the foreground branch length,


fg. 1
bl.=0.07
fg. 2
bl.=0.23
fg. 3
bl.=0.07
fg. 4
bl.=0.06
fg. 5
bl.=0.02
fg. 6
bl.=0.03
Homo sapiens (9606)
Pan troglodytes (9598)
Gorilla gorilla (9595)
Pongo abelii (9601)
Nomascus leucogenys (61853)
set 6
Macaca mulatta (9544)
Papio hamadryas (9557)
set 5
Callithrix jacchus (9483)
set 4
Tarsius syrichta (9478)
Microcebus murinus (30608)
Otolemur garnettii (306011)
Tupaia belangeri (37347)
Mus musculus (10090)
Rattus norvegicus (10116)
Dipodomys ordii (10020)
Cavia porcellus (10141)
Oryctolagus cuniculus (9986)
Ochotona princeps (9978)
set 3
Felis catus (9685)
Canis familiaris (9615)
Equus caballus (9796)
Bos taurus (9913)
Erinaceus europaeus (9365)
Sorex araneus (42254)
Loxodonta africana (9785)
Procavia capensis (9813)
Echinops telfairi (9371)
Dasypus novemcinctus (9361)
Choloepus hoffmanni (9358)
set 2
Macropus eugenii (9315)
Sarcophilus harrisii (9305)
Monodelphis domestica (13616)
set 1
Ornithorhynchus anatinus (9258)
Gallus gallus (9031)

and to a lesser extent its age, loosely formalised as the distance to the root [8, 9]. We confirm the major influence of the length of the foreground also on the sensitivity of the BEB procedure (simple linear regression $r^2 = 0.85$, $p = 1.95 \times 10^{-9}$). Moreover, adding the age of the foreground (see Section 3.4 for the definition used here) as a second explanatory variable leads to a better model fit (multiple linear regression $r^2 = 0.90$, $p = 5.26 \times 10^{-10}$, the resulting regression plane is shown in Figure 3.4). Although the gain is modest, it is favoured by model selection (Bayesian Information Criterion (BIC), ΔBIC$=-6.1$).

Hence, we show for the first time that the foreground branch length is a major factor for the performance of the BEB also in the context of the BSPS, as well as the age of the foreground to a lesser extent. In our simulations, these factors together account for roughly 90% of the observed variation in mean sensitivity. In the following, we use the obtained sensitivities as reference values for comparison, allowing to assess the effect of erroneous tree topologies on the performance of the BEB procedure.

Next, we turned to our principal question and asked if errors in the

FIGURE 3.1 *(preceding page)*: *Gene tree underlying sequence simulations*— The gene tree along which the sequences were simulated using INDELible [17]. The tree represents a subset of the Ensembl Compara [15] species tree, however, the species names and the corresponding NCBI taxon IDs are only given for orientation purposes as the simulations start from a random sequence and are therefore not related to any of the species' sequences. Every contiguous subset of branches labeled as foreground and highlighted in red, *i.e.* {fg1}, ..., {fg6}, {fg1, fg2}, ..., {fg5, fg6}, ..., {fg1, fg2, fg3, fg4, fg5, fg6}, is simulated under a foreground selection scheme described in Section 3.4, with the remaining branches simulated using a background selection scheme, leading to 21 different foregrounds for which eight replica each have been generated. Most basal branches serving as outgroups are shown in blue. Dashed lines are solely for clarity and labelling of sets of leaves for reference in the main text. *Abbreviations:* foreground (fg.), branch length (bl.)

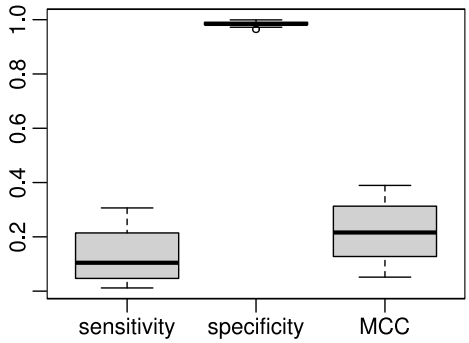


FIGURE 3.2: *Overall distribution of derivations from the confusion matrix*—The boxplots show the distributions of mean sensitivity, specificity and (MCC) across eight replica for all foreground branches. *Abbreviations:* Matthew's Correlation Coefficient (MCC)

topology of the gene tree have any influence on the performance of BEB.

Our approach is to manually introduce topological errors (the trees analysed in the following are listed in Table 3.2), pass these trees as input to PAML, and quantify the effect on the BEB procedure. Hence, the trees are not inferred from the simulated sequences for example by ML, allowing to freely manipulate the topology independently of its likelihood on the MSA of the simulated sequences.

First, we established that erroneous trees had an effect at all. The distribution of effects on sensitivity is shown in Figure 3.5, indicating that sensitivity can drop by over 0.15 in the most extreme cases, which we recall represents over 50% of the maximal observed sensitivity (see Figure 3.3). This is the first time a clear effect of the gene tree topology on the inference of positive selection is demonstrated.

Second, we sought to explain this effect as a function of the tree and foreground. We hypothesised that the most drastic effects on the results occur when the altered topology affects how long sequences are inferred to have experienced selection. To quantify this phenomenon, Figure 3.6 in-

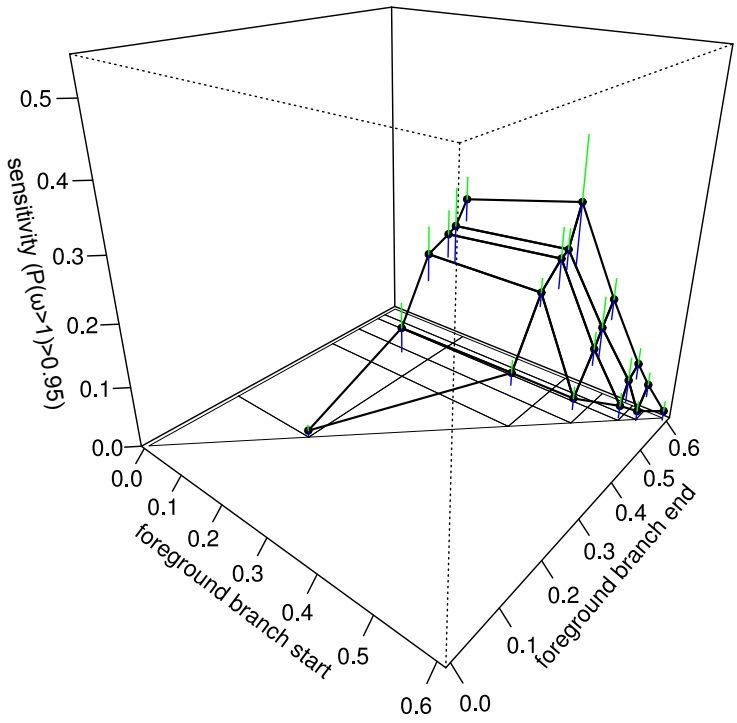


FIGURE 3.3: *Sensiticity of the BEB procedure on the true gene tree*—For every foreground branch, specified here by the pair of distances from the root to the start- and end-point of the branch on the tree, the mean sensitivity across the eight replica is shown. Green and blue lines are standard deviations. *Abbreviations:* Bayes Empirical Bayes (BEB)

troduces the "conflicting branch length" of a sequence, which we define as the difference in branch length that a sequence is inferred to have evolved under positive selection in the true and in a topologically perturbed tree. Note that conflicting branch length depends both on the gene tree and the position and lengths of the foreground branch simulated to be under selection, meaning that the same tree can result in different conflicting branch lengths.

In the major finding of this chapter, we validate our hypothesis by

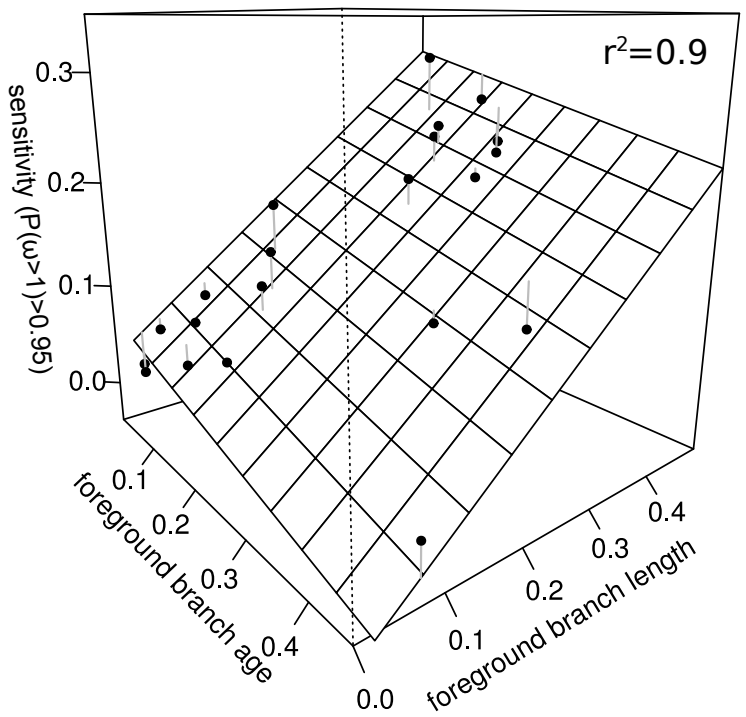


FIGURE 3.4: *Performance of the BEB procedure on the true gene tree*— Multiple linear regression of BEB performance inferring the sites simulated under positive selection in the foreground given the true tree. Performance is measured as MCC and averaged over the eight replica. The explanatory variables are foreground branch length and age (see Section 3.4 for the definition of age employed here). *Abbreviations:* Bayes Empirical Bayes (BEB), Matthew's Correlation Coefficient (MCC)

demonstrating a linear relationship between the loss in sensitivity of the BEB procedure and conflicting branch length (simple linear regression $r^2 = 0.87$, $p < 2.2 \times 10^{-16}$, Figure 3.6). Note that we also observe a small (here below 0.02) yet significant increase of specificity with conflicting branch length (simple linear regression $r^2 = 0.45$, $p < 2.2 \times 10^{-16}$, data not shown), resulting in two opposite effects of conflicting branch length on the overall accuracy.

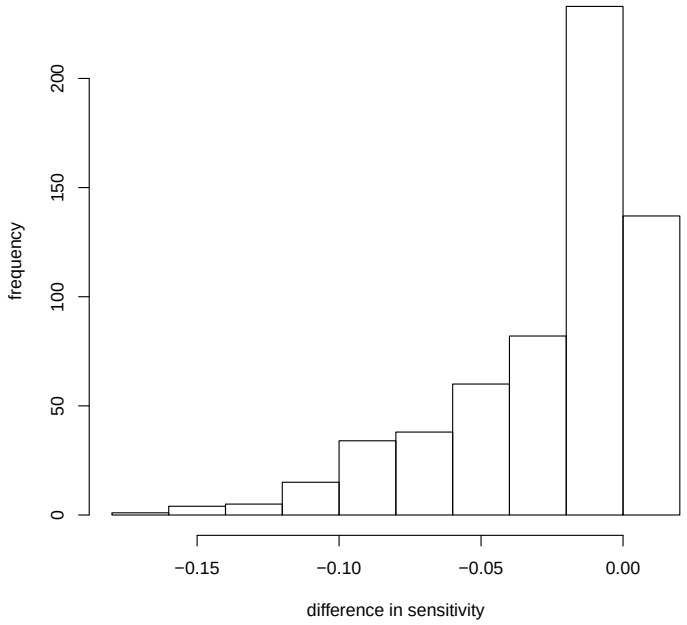
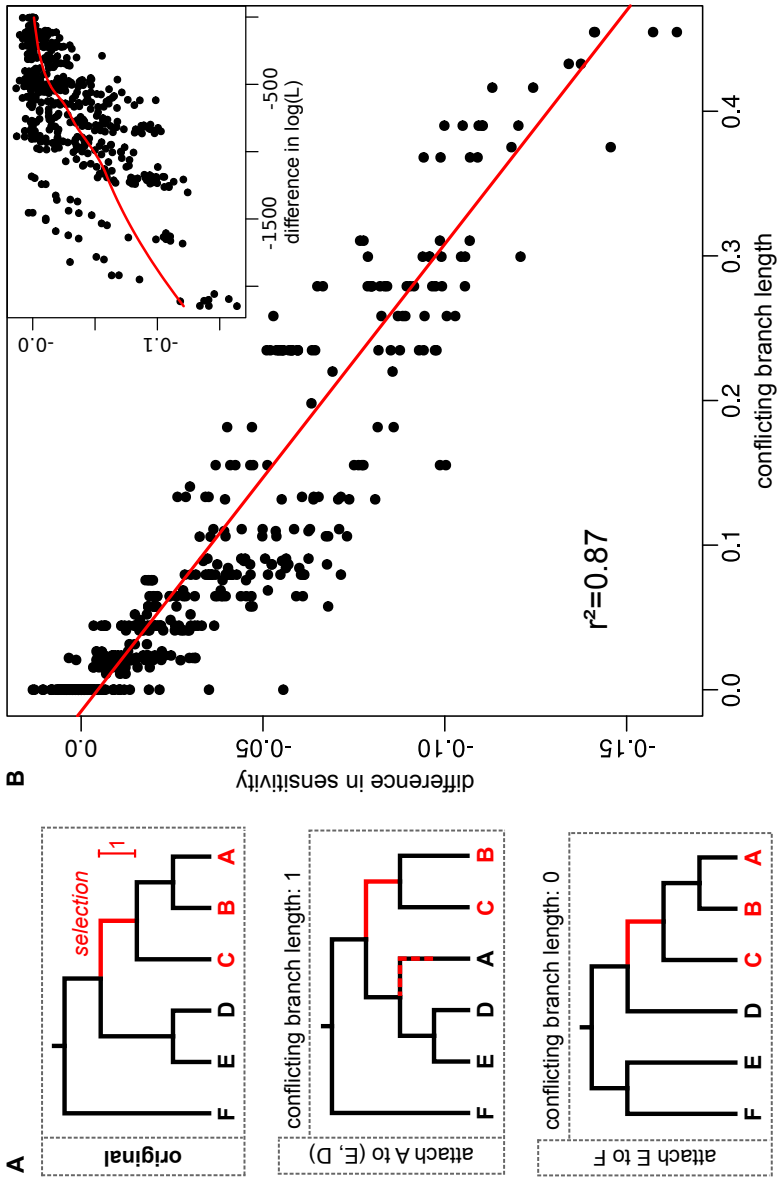


FIGURE 3.5: *Differences in sensitivity observed with erroneous topologies*— Histogram of mean differences in sensitivity across the eight replica observed for each of the erroneous topologies (listed in Table 3.2) and each of the foreground branches.

To further understand this phenomenon, we asked if the likelihood that a site evolves with $\omega > 1$ in the foreground, *i.e.* the numerical result of the BEB procedure per site, is also dependent on conflicting branch length. Surprisingly, Figure 3.7 shows that the likelihood of a site which was inferred to have evolved with $\omega > 1$ given the true topology, but not so given the erroneous topology , is reduced largely independently of conflicting branch length. In other words, higher conflicting branch length causes more sites to be inferred not to have evolved under selection in the


A
original
selection
1
F E D C B A
conflicting branch length: 1
attach A to (E, D)
F E D A C B
conflicting branch length: 0
attach E to F
F E D C B A
B
difference in sensitivity
0.0
-0.05
-0.10
-0.15
0.0
0.1
0.2
0.3
0.4
conflicting branch length
r²=0.87
-0.0
-0.1
-500
-1500
difference in log(L)

foreground, but the strength of the effect on a single site remains the same. Interestingly, this is also true for trees that introduce no conflicting branch length.

In summary, model violations introduced by erroneous tree topologies can have a strong detrimental effects in the context of the BSPS. We define and single out one parameter—conflicting branch length—as an explanatory variable for a strong linear loss in sensitivity. Furthermore, it appears that the overall tree quality, which has previously been suggested to be important for site models, is only indirectly related to the loss of sensitivity observed here (see inlay in Figure 3.6), as it has less explanatory power (simple linear regression $r^2 = 0.50$, $p < 2.2 \times 10^{-16}$). We conclude that the effect we describe is important whenever the branch-site analysis is performed and competing gene tree topologies exist. This may even

FIGURE 3.6 *(preceding page)*: *Conflicting branch length explains the loss in sensitivity of the BEB procedure*—Panel **A** illustrates the definition of "conflicting branch length" as the difference in branch length that a sequence is inferred to have evolved under positive selection in the true and in an altered tree. In the middle panel, moving leaf A to another part of the tree causes it to seemingly never have experienced positive selection, *i.e.* the set of sequences {A, B, C} which were inferred to have experienced positive selection in upper panel is changed to {B, C}. Changing the tree as seen in the lower panel does not create such an error. The definition of conflicting branch length is naturally extended to an entire tree as the sum of the conflicting branch lengths of all of its sequences. Note that the definition does not distinguish the direction of change, meaning that changing the set of sequences which are inferred to have experienced positive selection for example to {A, B, C, D} also leads to a conflicting branch length of one. In Panel **B**, each data point in the simple linear regression gives the change in sensitivity averaged over the eight replica corresponding to a manually altered topology (all listed in Table 3.2) and a specific foreground branch. The inlay depicts the same quantity, however, against the difference in the log-likelihood of the codon models fitted by PAML [16] on the true versus the erroneous topology. The line is a non-parametric local regression (LOESS). *Abbreviations:* Bayes Empirical Bayes (BEB)

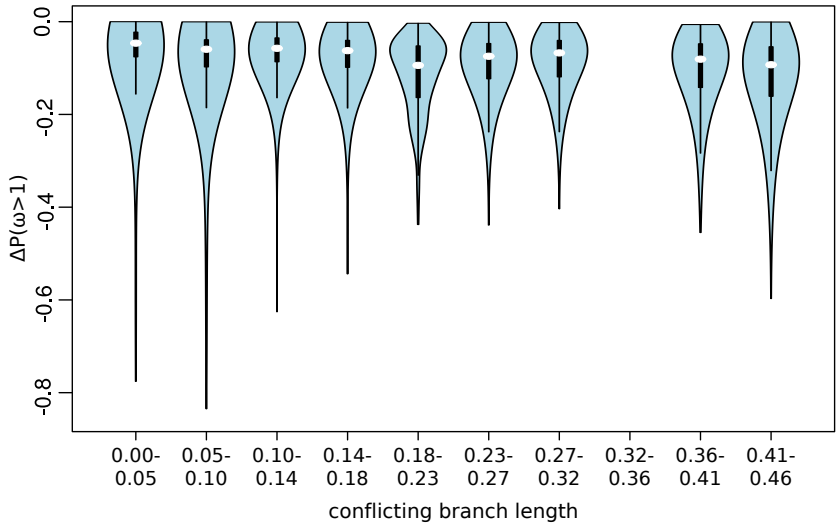


FIGURE 3.7: *Individual sites behave the same qualitatively independent of conflicting branch length*—For ten equally sized partitions of simulations by conflicting branch length (varying interval lengths suggested by the x-axis labels are caused by rounding), the distribution of losses in the likelihood that a site has evolved under selection (*i.e.* with $\omega > 1$) in the foreground is shown. Only sites that were inferred to have evolved under selection given the true topology, but not not so given the erroneous topology are included.

hold if the position of the foreground prevents alternative topologies to introduce conflicting branch lengths, as single sites (even if few) can still substantially lose likelihood to evolve under selection in the foreground. We address this last point in the following paragraph.

Lastly, we ask if the choice of gene tree also affects the results in real sequences. Real sequences usually evolve in more complex manners than it is possible and desirable to simulate, in particular compared to the rudimentary scheme without indels used here. Hence, although we loose the certainty about the right tree topology and selective regimes, only these conditions can show if the effect we described remains detectable, or

TABLE 3.1: *Results of the Branch Site Test of positive Selection on data from [18] obtained with original and reconciled gene tree*—Branch names refer to the labels given in Figure 3.8.

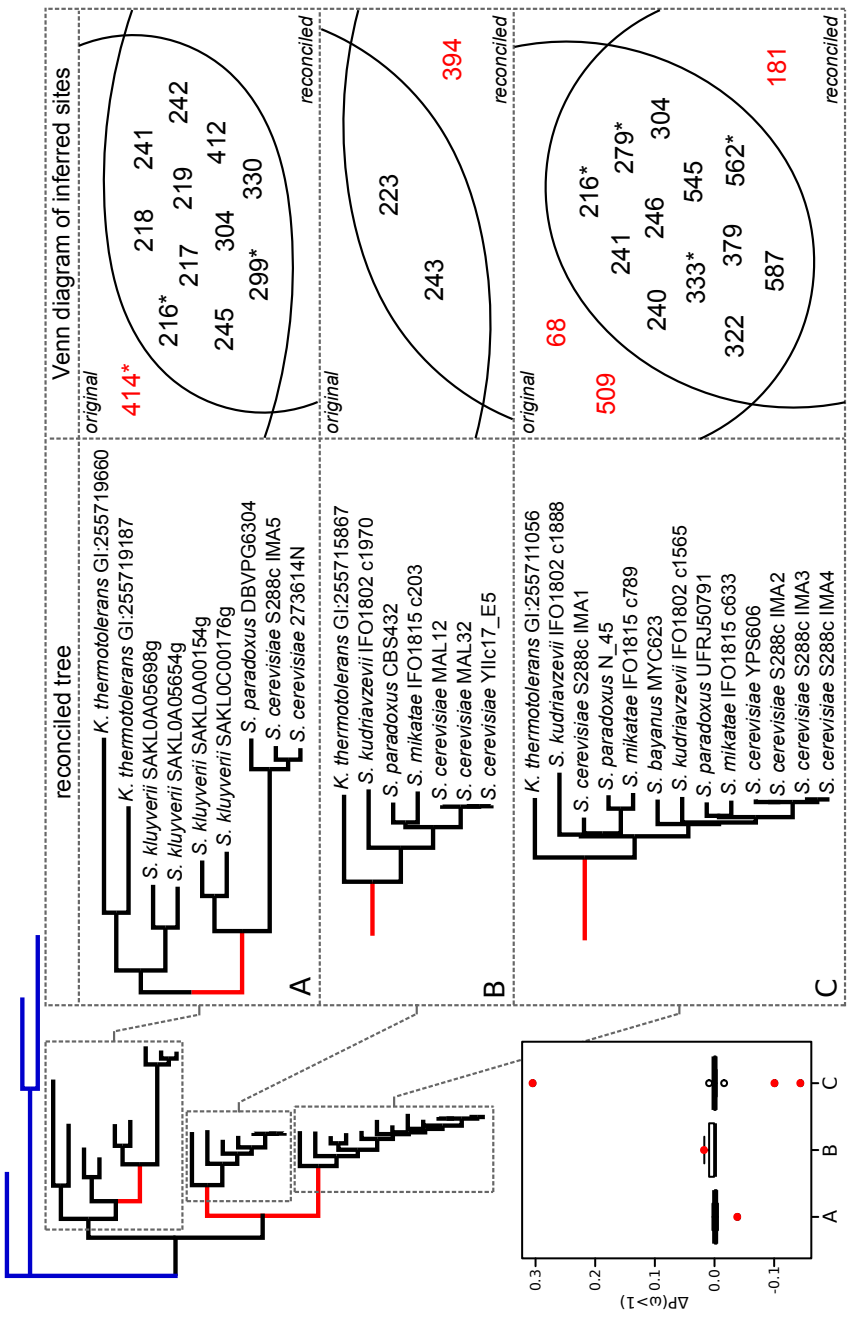
branch	tree	H_0	H_1	LRT	p-value
A	original	-25007.8	-24994.3	27	< 0.0005
	reconciled	-25130.1	-25116.0	28.1	< 0.0005
B	original	-25026.7	-25024.4	4.4	< 0.025
	reconciled	-25147.1	-25144.9	4.3	< 0.025
C	original	-25018.3	-25005.3	26	< 0.0005
	reconciled	-25139.1	-25125.8	26.6	< 0.0005

if it is minor and therefore without consequence for real sequences.

To answer this question, we reanalysed a data set of fungal glucosidase genes that have been studied by the authors with respect to their functional specialisation after gene duplication [18]. We chose this data set as it exemplifies a situation in which alternative tree topologies commonly arise, namely when the gene tree of orthologs and the species tree are incongruent, or alternatively when a gene tree / species tree reconciliation (see Box 1.2) yields a competing gene tree in the presence of paralogs. There seems to be no consensus on which tree to use (see [19] and [20] for examples of a gene- and a reconciled gene tree respectively) which suggests that both generally represent plausible choices.

The authors of [18] analysed the sequences using a gene tree inferred by MrBayes [21] testing for sites under positive selection in three different foreground branches. Based on the author's species tree of yeasts, manual gene tree / species tree reconciliation alters the subtree under each of these foreground branches. When comparing the BEB results obtained with the original tree and our reconciled tree, we observe that all three lists of sites changed as summarised in Figure 3.8 (see Table 3.1 for the results of the LRT).

This demonstrates that the gene tree influences the results also on


reconciled tree
K. thermotolerans GI:255719660
K. thermotolerans GI:255719187
S. kluyveri SAKL0A05698g
S. kluyveri SAKL0A05654g
S. kluyveri SAKL0A00154g
S. kluyveri SAKL0C00176g
S. paradoxus DBVPG6304
S. cerevisiae S288c IMA5
S. cerevisiae 273614N
A
K. thermotolerans GI:255715867
S. kudriavzevii IFO1802 c1970
S. paradoxus CBS432
S. mikatae IFO1815 c203
S. cerevisiae MAL12
S. cerevisiae MAL32
S. cerevisiae YIIc17_E5
B
K. thermotolerans GI:255711056
S. kudriavzevii IFO1802 c1888
S. cerevisiae S288c IMA1
S. paradoxus N_45
S. mikatae IFO1815 c789
S. bayanus MYC623
S. kudriavzevii IFO1802 c1565
S. paradoxus UFRJ50791
S. mikatae IFO1815 c633
S. cerevisiae YPS606
S. cerevisiae S288c IMA2
S. cerevisiae S288c IMA3
S. cerevisiae S288c IMA4
C
Venn diagram of inferred sites
original
reconciled
414*
218
241
216*
217
219
242
245
304
412
299*
330
original
reconciled
223
243
394
original
reconciled
68
509
216*
241
279*
240
246
304
333*
545
322
379
562*
587
181
ΔP(ω > 1)
0.3
0.2
0.1
0.0
-0.1
A
B
C

real sequences, including alignments containing indels. Yet, the experimental setting does not allow to tell if one inference of sites and tree or the other are better, because outside simulations the truth is generally unknown. Interestingly, these incongruences are obtained without introducing conflicting branch length. However, with exception of an outlier, the difference in likelihood at single sites are small and in the range of the values we predominantly observed in our simulations (see Figure 3.7). Hence, on this small scale where the result consists only of a few sites, the comparably small fluctuations in likelihood per site resulting from alternative tree topologies that we also report in simulations are enough to affect the output. This means that also in this case the gene tree matters, and it does so beyond the sole effect of conflicting branch length shown in Figure 3.6.

3.3 Conclusion

For the first time, we assessed the performance of the Bayes Empirical Bayes procedure in the context of branch-site models. We have shown

FIGURE 3.8 *(preceding page)*: *Different gene trees can lead to different BEB results on real sequences*—The upper left panel represents the gene tree published as Figure 4 in reference [18], with the analysed foreground branches highlighted in red and basal branches in blue. The three boxed areas correspond to the subtrees which we manually reconciled in the middle column using the species tree from Figure 1 in reference [18]. The results of inferring sites under positive selection with the two different trees (original and reconciled) are compared by Venn diagrams in the right column. Note that we were not able to reproduce the original results (here indicated by stars after the common sites) despite using the same sequences, alignments, trees and program version. None of the differences we report change the author's conclusions. Additionally, the lower left panel details the difference in likelihood for each of the sites reported by BEB. Red dots correspond to the sites differently classified using the original and reconciled tree. *Abbreviations:* Bayes Empirical Bayes (BEB)

that the length and age of the foreground not only determine the power of the LRT as reported before, but also that of the BEB procedure. Most importantly, we found evidence for an effect of the gene tree on the inference of selection in both simulated and real-word sequences. In simulations, we are able to explain this effect by virtue of a single parameter we coin conflicting branch length.

We conclude that the gene tree is an important factor for the branch-site analysis of positive selection so far unrecognised. Further investigations are needed to understand the precise effects and interplay with other known factors, and most relevantly develop guidelines for the choice of the gene tree in the analysis of real data sets.

3.4 Materials and Methods

Genes consisting of 522 codons were simulated along the tree depicted in Figure 3.1 starting from a random sequence at the root using INDELible (version 1.03) [17]. Simulation parameters were the transition/transversion ratio κ=2.1, chosen to match the average reported for the human genome (see for example [22]), a background selection scheme [1, 1, 0.8, 0.8, 0.5, 0.5, 0.2, 0.2, 0, 0] with every class making up 10% of the sites (the same as background scheme X from reference [8]), and a foreground selection scheme [0.5, 1, 4, 0.8, 4, 0.5, 4, 0.2, 0.8, 0.5]. We did not simulate indels.

The simulated sequences were analysed with PAML (version 4.6) [16], labelling the branches as foreground which were simulated as such. Branch lengths are estimated by PAML. The sites under selection in the foreground branches were obtained by BEB, but only in case the LRT was significant (here > 2.71, *i.e.* $p < 0.05$). Note that we did not correct for multiple testing, as every foreground branch and replicate is based on an independent simulation and we therefore never interrogate the same data twice.

We defined the age a of a foreground branch delimited by nodes n_1 and n_m and—in case of foregrounds spanning several individual branches—with additional internal nodes $n_2, \ldots, n_{m-1}$ as

$$a := \frac{1}{m} \cdot \sum_{i=1}^{m} d(n_i,\ n_{human}),$$

where $d(k, l)$ represents the distance between node k and l on the tree and n_{human} designates the node corresponding to human, *i.e.* the leaf at then end of fg6 in Figure 3.1.

We summarised the elements of the confusion matrix (*i.e.* true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN)) computing sensitivity, specificity and MCC according to their standard definitions $\frac{TP}{TP+FN}$, $\frac{TN}{FP+TN}$, $\frac{TP \cdot TN - FP \cdot FN}{\sqrt{P \cdot N \cdot P' \cdot N'}}$ respectively. Although the numerical values are hard to interpret, MCC provides an attractive measure as it summarises all values of the confusion matrix, while correcting for unequal amounts of positives (here selected sites, 30%) and negatives (other sites, 70%).

Tree manipulations were done in Python using Biopython [23] and the ETE library [24].

The DNA MSA for the reanalysis of the sequences from [18] were generated by using Pal2Nal (version 14) [25] on amino-acid alignment provided in the Supplementary material of reference [18]. The original tree was reproduced and reconciled manually, with branch length of the constrained topologies shown in Figure 3.8 computed by PhyML (version 3.1) [26]. All sequences, alignments and trees used to generate Figure 3.8 can be obtained from the author upon request. In order to exclude potential differences in our results stemming from different versions of PAML, we

performed the computations with PAML version 4.4 used by the authors in reference [18].

3.A Supplementary tables

TABLE 3.2: *List of perturbed topologies tested for their effect on the performance of the Bayes Empirical Bayes procedure.*—The leaf names refer to the NCBI taxon IDs also listed in the species tree in Figure 3.1.

name	operations
within set 1	swap leaves 13616, 9315
within set 2	swap leaves 9371, 9361
within set 3a	swap leaves 42254, 9615
within set 3b	swap leaves 10116, 9986
across set 5/4	attach 9557 to 9483
across set 5/3e	attach 9557 to 9478
across set 5/3d	attach 9557 to 30611
across set 5/3c	attach 9557 to 37347
across set 5/3b	attach 9557 to 9986
across set 5/3a	attach 9557 to 9365
across set 5/2	attach 9557 to 9358
across set 5/1	attach 9557 to 13616
across set 3d/3c	attach 30608 to 37347
across set 3d/3b	attach 30608 to 9986
across set 3d/3a	attach 30608 to 9365
across set 3d/2	attach 30608 to 9358
across set 3d/1	attach 30608 to 13616
across set 3b/3a	attach 10116 to 9365
across set 3b/2	attach 10116 to 9358

name	operations
across set 3b/1	attach 10116 to 13616
across set 3a/2	attach 42254 to 9358
across set 3a/1	attach 42254 to 13616
across set 2/1	attach 9371 to 13616
across set 1/6d	attach 9315 to 9598
across set 2/6d	attach 9371 to 9598
across set 1/3b	attach 9315 to 10116
across set 2/3b	attach 9371 to 10116
across double1	attach 30608 to 9358; attach 9557 to 9483
across double2	attach 30608 to 9358; attach 9557 to 13616

Acknowledgments

We wish to thank Karin Voordeckers for providing help retrieving the DNA sequences reanalysed here.

References

[1] Rasmus Nielsen and Melissa J Hubisz. “Detecting selection needs comparative data”. In: *Nature* 433.7023 (Jan. 2005), E6; discussion E7–8.

[2] J W Thornton and Rob DeSalle. “Gene family evolution and homology: genomics meets phylogenetics”. In: *Annual Review of Genomics and Human Genetics* 1 (2000), pp. 41–73.

[3] Ziheng Yang and Rasmus Nielsen. “Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages”. In: *Molecular Biology and Evolution* 19.6 (June 2002), pp. 908–917.

[4] Jianzhi Zhang, Rasmus Nielsen, and Ziheng Yang. “Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level”. In: *Molecular Biology and Evolution* 22.12 (Dec. 2005), pp. 2472–2479.

[5] Nick Goldman and Ziheng Yang. “A codon-based model of nucleotide substitution for protein-coding DNA sequences”. In: *Molecular Biology and Evolution* 11.5 (Sept. 1994), pp. 725–736.

[6] Ziheng Yang, Wendy S W Wong, and Rasmus Nielsen. “Bayes empirical bayes inference of amino acid sites under positive selection”. In: *Molecular Biology and Evolution* 22.4 (Apr. 2005), pp. 1107–1118.

[7] Ziheng Yang and Mario dos Reis. “Statistical properties of the branch-site test of positive selection”. In: *Molecular Biology and Evolution* 28.3 (Mar. 2011), pp. 1217–1228.

[8] William Fletcher and Ziheng Yang. “The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection”. In: *Molecular Biology and Evolution* 27.10 (Oct. 2010), pp. 2257–2267.

[9] Walid H Gharib and Marc Robinson-Rechavi. “The Branch-Site Test of Positive Selection Is Surprisingly Robust but Lacks Power under Synonymous Substitution Saturation and Variation in GC”. In: *Molecular Biology and Evolution* (May 2013).

[10] Ziheng Yang et al. “Codon-substitution models for heterogeneous selection pressure at amino acid sites”. In: *Genetics* 155.1 (May 2000), pp. 431–449.

[11] Eyal Privman, Osnat Penn, and Tal Pupko. “Improving the performance of positive selection inference by filtering unreliable alignment regions”. In: *Molecular Biology and Evolution* (July 2011).

[12] Benjamin P Blackburne and Simon Whelan. “Class of Multiple Sequence Alignment Algorithm Affects Genomic Analysis”. In: *Molecular Biology and Evolution* (Dec. 2012).

[13] Gregory Jordan and Nick Goldman. “The effects of alignment error and alignment filtering on the sitewise detection of positive selection.” In: *Molecular Biology and Evolution* 29.4 (Apr. 2012), pp. 1125–1139.

[14] Ziheng Yang. “The power of phylogenetic comparison in revealing protein function”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.9 (Feb. 2005), pp. 3179–3180.

[15] Albert J Vilella et al. “EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates”. In: *Genome Research* 19.2 (Feb. 2009), pp. 327–335.

[16] Ziheng Yang. “PAML 4: phylogenetic analysis by maximum likelihood”. In: *Molecular Biology and Evolution* 24.8 (Aug. 2007), pp. 1586–1591.

[17] William Fletcher and Ziheng Yang. “INDELible: a flexible simulator of biological sequence evolution”. In: *Molecular Biology and Evolution* 26.8 (Aug. 2009), pp. 1879–1888.

[18] Karin Voordeckers et al. “Reconstruction of Ancestral Metabolic Enzymes Reveals Molecular Mechanisms Underlying Evolutionary Innovation through Gene Duplication”. In: *PLoS Biology* 10.12 (Dec. 2012), e1001446.

[19] Krishanu Mukherjee, Henry Campos, and Bryan Kolaczkowski. “Evolution of animal and plant dicers: early parallel duplications and recurrent adaptation of antiviral RNA binding in plants”. In: *Molecular Biology and Evolution* 30.3 (Mar. 2013), pp. 627–641.

[20] Pouria Dasmeh et al. “Positively selected sites in cetacean myoglobins contribute to protein stability”. In: *PLoS Computational Biology* 9.3 (Mar. 2013), e1002929.

[21] Fredrik Ronquist et al. “MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space.” In: *Systematic Biology* 61.3 (May 2012), pp. 539–542.

[22] Mark A Depristo et al. “A framework for variation discovery and genotyping using next-generation DNA sequencing data”. In: *Nature Genetics* 43.5 (May 2011), pp. 491–498.

[23] Peter J A Cock et al. “Biopython: freely available Python tools for computational molecular biology and bioinformatics”. In: *Bioinformatics* 25.11 (May 2009), pp. 1422–1423.

[24] Jaime Huerta-Cepas, Joaquín Dopazo, and Toni Gabaldón. “ETE: a python Environment for Tree Exploration”. In: *BMC Bioinformatics* 11 (2010), p. 24.

[25] Mikita Suyama, David Torrents, and Peer Bork. “PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments”. In: *Nucleic Acids Research* 34.Web Server issue (July 2006), W609–12.

[26] Stéphane Guindon and Olivier Gascuel. “A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood”. In: *Systematic Biology* 52.5 (Oct. 2003), pp. 696–704.

Author contribution: I conceived, designed and performed the experiments, analysed the data and wrote the chapter. The data and the corresponding Figure 4.6 has been generated with help of Krzysztof Kus (Protein-Nucleic Acids Interactions Group, Instituto Gulbenkian de Ciência).

Chapter 4

Functional Innovation in the Rab family of small GTPases

"Population genetic models take such a familiar form that it is easy to overlook a respect in which they are odd. These models begin with selection coefficients but say nothing whatever about where these coefficients come from. It is vaguely assumed of course that selection coefficients emerge from the phenotypic effects of mutations [...]. Although this shortcut suffices for many evolutionary questions, it leaves us in an awkward position when thinking about adaptation." [1]

—H. Allen Orr, 2005

Abstract Chapter 4

Neofunctionalisation is a popular model to explain the retention of gene duplicates with new functions, however, it does not explain the evolution of new functions themselves. Here, we use a pair of subfamilies from the Rab family of small GTPases to ask how mutations can lead new protein function in the context of gene duplication. Rabs are protein switches that function as master regulators of vesicular trafficking. In their active GTP-bound state, they recruit so-called effector proteins that then exert their function. We hypothesise that Rabs neofunctionalise by changing these sets of effectors, a model we refer to as effector switching. First, we demonstrate that Rab11 and its duplicate Rab25 indeed evolved by neofunctionalisation. Next, we test our hypothesis about the mechanism of neofunctionalisation by confirming three predictions of the effector switching model: Rab25 has replacements mostly on its surface, Rab11 and 25 have different effectors, and Rab25 has been selected for new binding interfaces. Moreover, we discuss alternative but non-exclusive modes of functional evolution and their support. Finally, we detail the temporal dynamics of the process. We present indications that Rab25 function may be different between zebrafish and mammals and has therefore evolved over an extended period after duplication. In conclusion, we suggest that our model for the evolution of Rab function represents a general mechanism for neofunctionalisation particularly attractive for the important class of protein master regulators.

4.1 Introduction

Evolutionary innovation, *i.e.* the process by which novel characters emerge in evolution [2], persists as a fascinating and fundamental problem in biology [3, 4] transcending the boundaries of biological disciplines. At the morphological level, it is one of the defining topics of an entire field merging evolution and development or EvoDevo [2–4]. At the molecular level, it is subsumed under or even equivalent to asking how protein function evolves (see Chapter 1), a question receiving growing attention as part of what has been coined the "functional synthesis" of evolutionary and molecular biology [5].

A particularly favourable and frequent condition for the evolution of novel protein functions—and therefore of upmost importance—is the presence of redundancy, mostly resulting from duplication of DNA. This connection has been put forth most prominently by Susumo Ohno, who emphasised the role of gene duplication as a mechanism to free genes from functional constraints [6]. Duplicates can thus diverge even through non-functional intermediates without deleterious effects on fitness, until a new beneficial function evolves, is picked up by positive selection, and the gene is ultimately preserved by purifying selection. The resulting model coined neofunctionalisation therefore provides an evolutionary mechanism to explain retention of gene duplicates. However, even though functional innovation is implied by definition, no explicit biochemical mechanism is given specifying how mutations lead to a new function. In other words, while neofunctionalisation explains the evolutionary emergence of genes with new functions, the model is not concerned with the emergence of new functions themselves.

Here, we ask how mutations lead to new protein function in the context of neofunctionalisation. While "[i]t is more important to understand the

general than the particular, [...] the first is achieved only through the second" [7] and we therefore study the question focussing on Rab small GTPases as a model gene family. Rabs are an especially interesting system for various reasons.

Firstly, Rabs frequently neofunctionalise. The Rab family is large and grew from roughly 20 Rab subfamilies in the last common ancestor of animals [8] to 44 subfamilies and over 60 genes in humans. The evidence that those duplications represent neofunctionalisations is threefold. Primarily, the available functional annotation of Rabs supports a conserved core of ancient Rabs with equivalent or equal functions all over the eukaryotic tree of life. The functional characterisation of some Metazoan subfamilies that emerged later by duplication found distinct functions different from those of the subfamilies they are derived from (see *e.g.* Figure 1.5). Next, the overall patterns of Rab family evolution are not consistent with sub- and rather suggest neofunctionalisation, both in terms of sequence divergence and tissue specificity of expression (see Chapter 2) [8]. Lastly, at least one clear case of neofunctionalisation has been experimentally established and studied, the evolution of a catalytically inactive Rab6 duplicate that lost Golgi localisation and functions in cell cycle progression at the centrosome instead [9]. Secondly, Rabs are a promising model system to study the process of neofunctionalisation because they are protein switches. As such, they are subject to a diverse set of constraints, for example flexibility, stability, multispecificity and enzyme function. The impact of these factors on the evolution of function is interesting to consider. Thirdly, Rabs are master regulators of vesicular trafficking and essential for the organisation of the endomembrane system. They are markers for many organelles and transport pathways, and therefore allow to indirectly study one of the hallmarks of the Eukaryotic cell, its extensive membranous compartments. Finally, due to their importance a rich body of work has contributed public data on sequence, structure and function, including expression, localisa-

tion and interactions that may be integrated in an evolutionary analysis.

Here, we address the hypothesis that Rabs neofunctionalise by changing their set of binding proteins. The hypothesis is based on the known biochemistry of Rabs: as depicted for example in Figure 1.4, Rabs exert their function via effector proteins that by definition recognise the GTP-bound form specifically and are recruited to the compartment the active Rab is attached to. Hence, an intuitive hypothesis is that altering this set of effectors represents a way that evolution tinkers with Rab function. Subsequently, we refer to this process as effector switching. Alternatively, other possibilities to evolve Rab function may be for example by affecting the interactions with GAPs and GEFs that activate and inactivate Rabs or by changes in overall tissue-specificity. These options are also briefly considered.

While the above hypothesis can in principal be tested for any pair of Rab duplicates, in the remainder of the chapter we focus on the ancestral Rab11 and Rab25 that emerged from the former by duplication at the base of jawed vertebrates. There are three major reasons for this choice. First and foremost, as far as we can tell from the analysis presented in Chapter 2, the patterns of evolution do not differ between the human Rab subfamilies [8]. Hence, we expect the results presented for the pair of subfamilies Rab11 and 25 to be general and also apply to most other cases of Rab neofunctionalisation, at least in Metazoa. Note that exceptions do exist: as already mentioned, in at least one case a Rab6 gained a new function by losing its catalytic activity, a scenario clearly and fundamentally different from our hypothesis [9]. However, as it is to the best of our knowledge the only documented case in animal Rabs we expect this evolutionary path to be rare and therefore represent an exception rather than the rule. Second, both Rab11 and Rab25 have been extensively studied. They function at recycling endosomes, many effectors are known, and

three crystals of Rab11 bound to an effector have been published so far. Additionally, both Rab11 and Rab25 play a role in various diseases, most importantly Rab25 which has been implicated in cancer progression (see references [10, 11] for reviews). A third and more practical reason is that the duplication that gave rise to Rab25 is not yet too remote to analyse patterns of synonymous and nonsynonymous substitutions, which allow to generate hypothesis about past positive selection. Generally, for events beyond this time frame too many synonymous substitution have occurred and the resulting saturation precludes the reliable estimation of the rates required for the analysis.

In the following, we begin by establishing a gene phylogeny of some animal Rab11 and 25 sequences. This tree allows us to confirm that the pattern of sequence evolution is indeed best explained by a process of neofunctionalisation. As part if this, we map the sites that show signs of positive selection in the lineage leading to the extant human Rab25. Next, we test our hypothesis about the mechanism of neofunctionalisation by verifying and confirming three predictions of the proposed effector switching model: Rab25 has replacements mostly on its surface, Rab11 and 25 have different effectors, and Rab25 has been selected for new binding interfaces. Additionally, we discuss the support for alternative but non-exclusive modes of functional evolution in Rab25. Finally, we present indications that the function of Rab25 may be different between zebrafish and mammals and has therefore not ceased to evolve after duplication. In conclusion, we suggest that our model for the evolution of Rab function represents a general mechanism for neofunctionalisation particularly attractive for the important class of protein master regulators.

4.2 Results / Discussion

4.2.1 Rab25 evolved by neofunctionalisation

To ensure that Rab11 and 25 represent a valid model system to test our hypothesis, we begin by confirming that Rab25 indeed evolved by neofunctionalisation. We consider two different levels: function, including intracellular localisation and tissue specificity of expression, and sequence.

The neofunctionalisation model makes two predictions about the function of involved Rabs. One one hand, one should find that the function of Rab11 outside jawed vertebrates, *i.e.* in those organisms that branched off before the emergence of Rab25, is equivalent or the same as in jawed vertebrates where Rab11 coexists with Rab25. We can test this prediction by reviewing published data on the function of Rab11 in mammals and for example in budding yeast. After its initial discovery [12, 13], mammalian Rab11 has been found to localise both to the *trans*-Golgi network [14] and recycling endosomes [15] where it functions in the secretory pathway. The direct functional comparison between Rab11 and its yeast orthologs ypt31/32 is complicated by the fact that there is no one to one relationship between the compartments of the endocytic and secretory pathways in mammals and budding yeast (see *e.g.* reference [16]). Yet, the localisation of ypt31/32 at the *trans*-Golgi and their involvement in secretion [17] and recycling [18] clearly indicate functional equivalence and homology. On the other hand, neofunctionalisation predicts that Rab25 should have a distinct function, different from Rab11. The first indication that this is indeed the case dates back to the identification and first description of Rab25: while Rab11 is ubiquitously expressed [19], Rab25 shows a restricted epithelial distribution [20]. Subsequently, Rab25 has been found to co-localise with Rab11 at recycling endosomes, however, functional characterisation pointed to related but distinct functions [21].

The second level of predictions from neofunctionalisation relate to sequence divergence and the selective regimes causing it. Gene copies evolving via subfunctionalisation each experience relaxed purifying selection and one therefore expects divergence in both sequences relative to the common ancestor. In contrast, neofunctionalisation predicts only one copy to diverge, and unlike subfunctionalisation under the influence of positive selection. We first consider sequence divergence independently of the evolutionary forces. Figure 2.6 suggests the latter of the above patterns, however, divergence cannot be assessed properly due to the categorical classification into subfamilies. Therefore, we analysed divergence after duplication directly by reconstructing the ancestral sequences before duplication in a phylogentic framework (see Subsection 4.4.2 in Materials and Methods for details). Figure 4.1 summarises the results. We observe that Rab11 barely changes after Rab25 emerges (96% identical to the extant human Rab11a), while the gene copy that gives rise to Rab25 shows important sequence divergence after duplication (40% of Rab25 sites have been replaced in humans compared to the last common ancestor of Rab11a/25). This pattern clearly points towards neofunctionalisation. The last prediction to be confirmed is that the observed divergence in Rab25 is at least partly caused by positive selection. Using the gene tree already presented in Figure 4.1, we detect statistical signatures in extant sequences most likely left by past positive selection acting all over the branch leading to the human Rab25 sequence (shown in Figure 4.2).

Hence, functional data for Rab11 and 25, sequence divergence patterns in both Rabs, and the selective regimes under which Rab25 evolved all conform to the predictions made by the neofunctionalisation model. Therefore, we conclude that the pair Rab11/25 indeed represents a valid model system to test our hypothesis about the mechanisms of functional evolution in the context of neofunctionalisation.

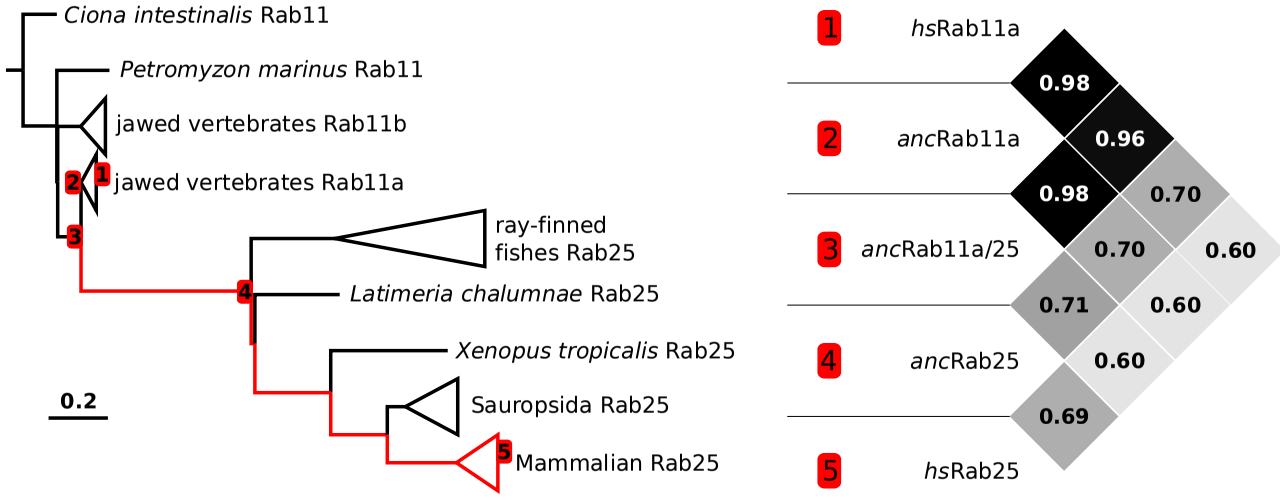
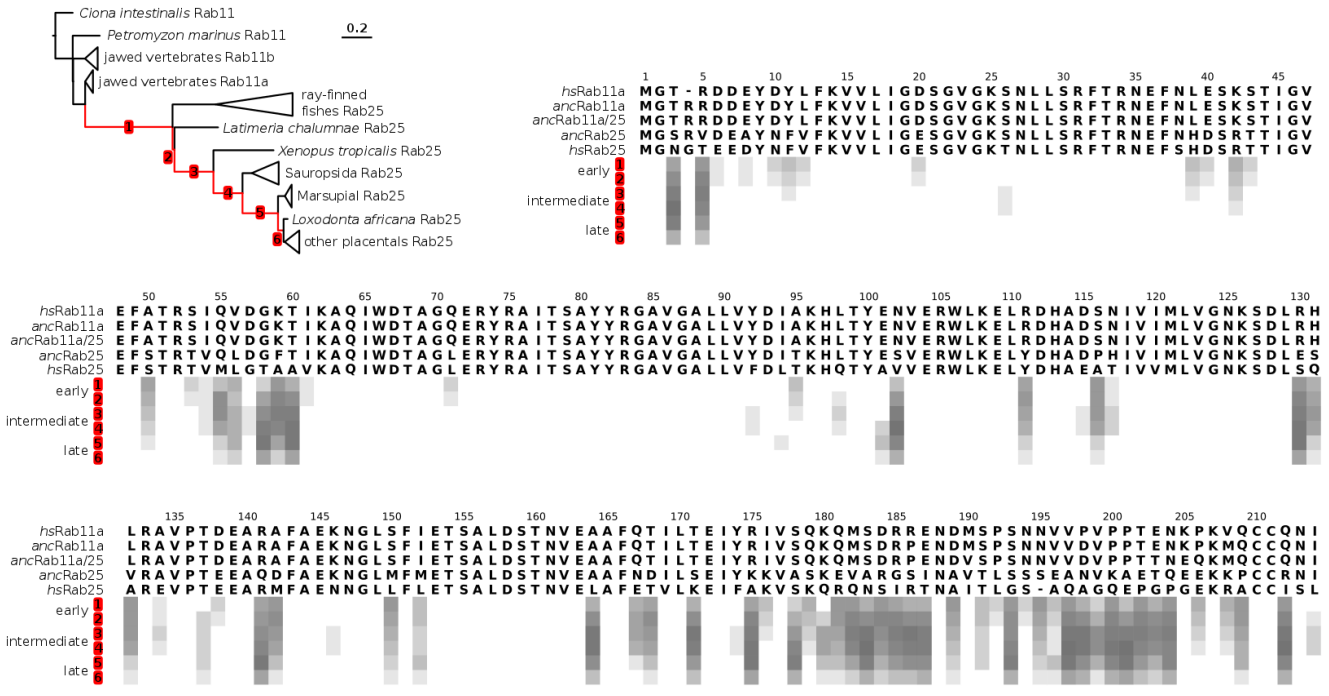


FIGURE 4.1: *The ancestral Rab25 was a Rab11a*—Reconciled gene tree of Rab11 and Rab25 sequences (see Subsection 4.4.1 for details), and pairwise sequence identities between reconstructed ancestral sequences (labeled $2-4$) and extant human Rab11 and 25 sequences (labeled 1, 5). The branches highlighted in red represent the evolutionary path of the extant human Rab25, and were marked as foregound in the model used for the inference of the ancestral sequences (see Subsection 4.4.2 for details).


Ciona intestinalis Rab11
Petromyzon marinus Rab11
0.2
jawed vertebrates Rab11b
jawed vertebrates Rab11a
ray-finned fishes Rab25
Latimeria chalumnae Rab25
Xenopus tropicalis Rab25
Sauropsida Rab25
Marsupial Rab25
Loxodonta africana Rab25
other placentals Rab25
1
2
3
4
5
6
hsRab11a MGT-RDDEYDYLFKVVLIGDSGVGKSNLLSRFTRNEFNLESKSTIGV
ancRab11a MGTRRDDEYDYLFKVVLIGDSGVGKSNLLSRFTRNEFNLESKSTIGV
ancRab11a/25 MGTRRDDEYDYLFKVVLIGDSGVGKSNLLSRFTRNEFNLESKSTIGV
ancRab25 MGSRVDEAYNFVFKVVLIGESGVGKSNLLSRFTRNEFNHDSRTTIGV
hsRab25 MGNGTEEDYNFVFKVVLIGESGVGKTNLLSRFTRNEFSHDSRTTIGV
early
intermediate
late
hsRab11a EFATRSIQVDGKTIKAQIWDTAGQERYRAITSAYYRGAVGALLVYDIAKHLTYENVERWLKELRDHADSNIVIMLVGNKSDLRH
ancRab11a EFATRSIQVDGKTIKAQIWDTAGQERYRAITSAYYRGAVGALLVYDIAKHLTYENVERWLKELRDHADSNIVIMLVGNKSDLRH
ancRab11a/25 EFATRSIQVDGKTIKAQIWDTAGQERYRAITSAYYRGAVGALLVYDIAKHLTYENVERWLKELRDHADSNIVIMLVGNKSDLRH
ancRab25 EFSTRTVQLDGFTIKAQIWDTAGLERYRAITSAYYRGAVGALLVYDITKHLTYESVERWLKELYDHADPHIVIMLVGNKSDLES
hsRab25 EFSTRTVMLGTAAVKAQIWDTAGLERYRAITSAYYRGAVGALLVFDLTKHQTYAVVERWLKELYDHAEATIVVMLVGNKSDLSQ
early
intermediate
late
hsRab11a LRAVPTDEARAFAEKNGLSFIETSALDSTNVEAAFQTILTEIYRIVSQKQMSDRRENDMSPSNNVVPVPPTENKPKVQCCQNI
ancRab11a LRAVPTDEARAFAEKNGLSFIETSALDSTNVEAAFQTILTEIYRIVSQKQMSDRPENDMSPSNNVVDVPPTENKPKMQCCQNI
ancRab11a/25 LRAVPTDEARAFAEKNGLSFIETSALDSTNVEAAFQTILTEIYRIVSQKQMSDRPENDVSPSNNVVDVPPTTNEQKMQCCQNI
ancRab25 VRAVPTEEAQDFAEKNGLMFMETSALDSTNVEAAFNDILSEIYKKVASKEVARGSINAVTLSSSEANVKAETQEEKKPCCRNI
hsRab25 ARVPTEEARMFAENNGLLFLETSALDSTNVELAFETVLKEIFAKVSKQRQNSIRTNAITLGS-AQAGQEPGPGEKRACCISL
early
intermediate
late

4.2.2 Rab25 function evolved by effector switching

Given that Rab25 evolved from Rab11 by neofunctionalisation, we can now test the effector switching model that we propose for this process. Our hypothesis is that Rab25 gained a new function by changing its set of effectors that it initially inherited from Rab11. Effector switching makes at least three predictions that we address in turn.

First, Rab25 should have replacements mostly on its surface, as we expect that these replacements represent a more direct way to alter interaction interfaces and as a result the set of interactors. Accepting this reasoning and because we suggest that gain and loss of interactions provides the beneficial function required as part of the neofunctionalisation model, we expect in particular that sites showing signs of positive selection preferentially localise to the protein surface. Second, altering the set of effectors in evolution should have the clear consequence that extant Rab11 and Rab25 do not interact with the same set of effectors. Third and elaborating on the previous two points, we should see that the interaction interfaces of effectors not shared by Rab11 and Rab25 have been shaped by positive selection.

The next paragraphs present data in favour of each of these predictions. Therefore, we find no evidence to reject our hypothesis from which we conclude that effector switching represents a promising mechanism to explain Rab25 neofunctionalisation. In the last paragraph we briefly comment on alternative but non-exclusive modes of functional evolution and

FIGURE 4.2 *(preceding page)*: *Positive selection on Rab25*—Traces of past positive selection on the six different branches leading from the emergence of Rab25 by duplication from an ancestral Rab11a (see Figure 4.1) to the extant Rab25 sequence in humans. The branch labels $1-6$ refer to the gene tree on the upper left. The darker the area below the sequence corresponding to a particular site and branch, the stronger the evidence that positive selection indeed acted there (see Subsection 4.4.3 for details).

their support, namely via changes in intracellular localisation, regulation by GAPs and GEFs and overall tissue-specificity.

Rab25 has replacements mostly on its surface

Although Rabs are small and the majority of the residues localises to the surface, Figure 4.3 confirms a significant association between surface residues and those that have been replaced in Rab25 during its evolution from the common ancestor of Rab11 and 25. This association is even clearer when differing residues are restricted only to those that show signs of positive selection, as all of them lie on the surface of the Rab25 molecule.

These results support the notion that Rab25 evolved its function by altering interaction interfaces located on the surface, rather than for example aspects of its structure like stability or flexibility that are determined mostly by internal interactions between buried residues. Furthermore, these alterations of the interaction interfaces have been driven by positive selection and therefore had to have beneficial functional effects. This represents a powerful way to test the effector switching model, however, so far no actual effectors are considered and it thus remains to be shown that specific interactions are indeed affected by the observed changes of the surface residues. Next, we therefore review the known Rab11 and 25 effectors.

Rab11 and 25 have different effectors

Probably the most intuitive prediction from the effector switching model is that pairs of Rabs that evolved by duplication and neofunctionalisation should interact with different sets of effectors. This can be tested by reviewing the literature on known Rab11 and Rab25 effectors. Note that this small-scale approach is expected to miss actual effectors that have currently not yet been described. However, for the sake of testing our

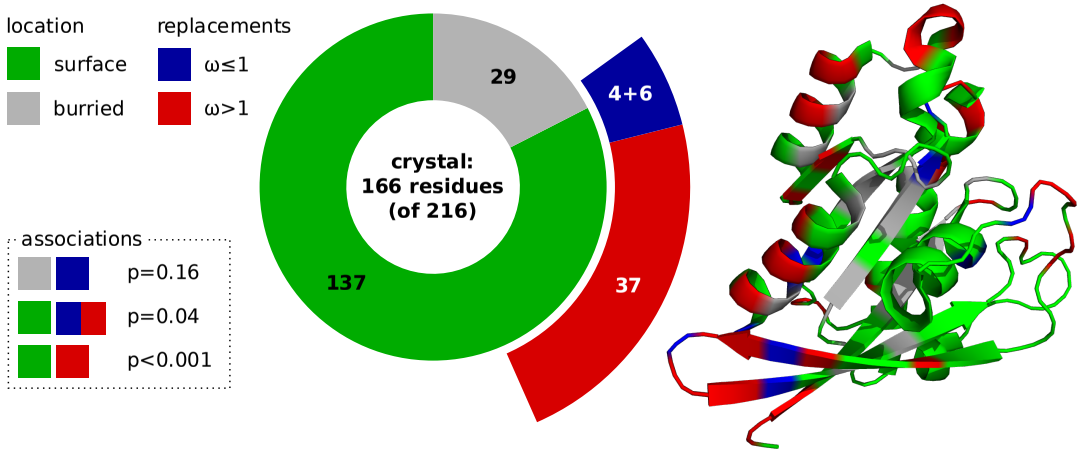


FIGURE 4.3: *Sites with replacements in Rab25 localise to the protein surface*—Sites that differ in Rab25 compared to the ancestral Rab11 (see Figure 4.1) are mapped onto the Rab25 structure (right), distinguishing sites with signs of past episodic positive selection ($\omega > 1$, see Figure 4.2) and those without. Furthermore, their association with the protein surface (obtained from PyMOL) and the buried core is tested using Fisher's exact test (one-tailed). PDB accession of Rab25 structure 1OIW [22]. *Abbreviations:* ratio of nonsynonymous to synonymous substitution rate or dN/dS (ω)

model it is enough if at least one interaction specific to either Rab11 or 25 has been found. Figure 4.4 summarises the current set of known Rab11 and 25 effectors, revealing that both Rab11- and Rab25-specific effectors exist in agreement with the effector switching model.

This represents a necessary condition following from the model proposed here. Finally, we combine both of the above aspects and test if specific interactions can be shown to have been lost or gained in Rab25.

Rab25 has been selected for new binding interfaces

The availability of detailed structural information on some of the interactions reviewed above allows to verify if loss and/or gain of these interactions has specifically been driven by positive selection. Finding evidence for positive selection is a powerful conceptual tool as it allows to conclude that switching effectors had beneficial functional effects. In the context of neofunctionalisation, loss and/or gain of effectors therefore becomes the mechanism by which a new function evolves, which is precisely what we propose to be the dominant form of functional evolution in Rabs. As shown in Figure 4.4, structural details about the interaction are known for four effectors. Unfortunately, these do not include a Rab11-specific effector, which may have allowed to ask if active loss of interactions (*i.e.* driven by positive selection) takes part in the neofunctionalisation process. We briefly comment on each of the effectors in turn, beginning with the most revealing one: the Rab25-specific Integrin β1-subunit (ITGB1).

Using chimera, the interaction interface between Rab25 and ITGB1 has been mapped to the Rab25 hypervariable C-terminal region starting around residue 170. The Rab hypervariable termini are unstructured regions that do not adopt a stable fold and are therefore usually excluded from crystallographic analysis. Figure 4.2 reveals that this region is heavily shaped by positive selection. Hence, despite the lack of resolution at the residue-level, it is reasonable to assume that the formation of the in-

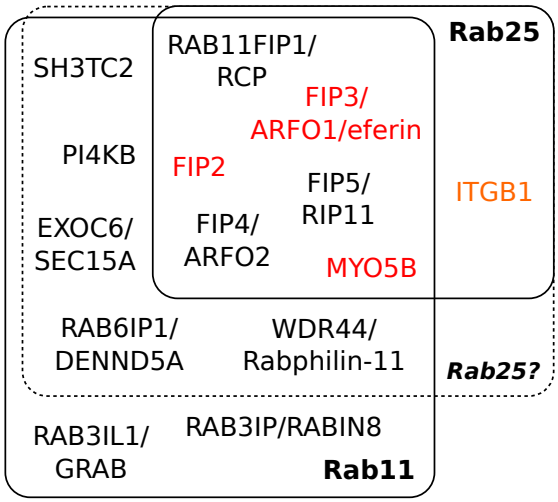


FIGURE 4.4: *Rab25 both lost and gained effectors after duplication from a Rab11*—Currently known Rab11 and 25 effectors. The dashed box comprises proteins for which the inability to interact with Rab25 has not been demonstrated. Effectors in red have been crystallised together with Rab11 [23–26], additionally a Rab25-FIP2 crystal has been published [27]. ITGB1 is highlighted in orange to indicate that a less detailed mapping of the interaction interface with Rab25 based on chimeras exists [28]. Additional effectors are RAB11FIP1/RCP [29], FIP2 [29, 30], FIP3/ARFO1/eferin [29, 31], FIP4/ARFO2 [32, 33], FIP5/RIP11 [29, 30, 34], MYO5B [35, 36], RAB3IP/RABIN8 [30], RAB3IL1/GRAB [37], RAB6IP1/DENND5A [38], SH3TC2 [39], EXOC6/SEC15A [40], PI4KB [41], WDR44/Rabphilin-11 [42]. *Abbreviations:* Integrin β1 subunit (ITGB1), Rab11 family interacting protein 1 (RAB11FIP1), Rab-coupling protein (RCP), Arfophilin-1 (ARFO1), Rab11-interacting protein (RIP11), Unconventional myosin-Vb (MYO5B), Rab3A-interacting-like protein 1 (RAB3IL1), guanine nucleotide exchange factor for Rab3a (GRAB), DENN domain-containing protein 5A (DENND5A), SH3 domain and tetratricopeptide repeat-containing protein 2 (SH3TC2), Exocyst complex component Sec15A (EXOC6), Phosphatidylinositol 4-kinase β (PI4KB), WD repeat-containing protein 44 (WDR44)

teraction between Rab25 and ITGB1 has been driven by positive selection, confirming the prediction of the effector switching model. To add confidence that formation of the binding site was indeed driven by selection on

the Rab sequence rather than by evolution of the complementary interface in the binding partner, we asked if ITGB1 is also found outside jawed vertebrates. Figure 4.5 shows that this is the case, which suggests that Rab25 evolved a binding site to an existing evolutionarily constrained protein and renders the possibility that the observed selection in the Rab25 C-terminus is entirely unrelated to binding ITGB1 less likely.

Interactions conserved amongst Rab11 and 25 are less interesting to validate the effector switching model, however, they nonetheless reveal an interesting pattern. The interaction with FIP2 is mediated amongst others by residues around the switch region differing between Rab11 and 25 [23, 24, 27], some of which we found to that have experienced positive selection. However, although the characterisation of the Rab25-FIP2 interaction found a threefold reduced affinity of Rab25 to FIP2 compared to Rab11, the molecular basis of this drop in affinity remains elusive. As the authors point out, this is because the divergent residues do not alter the number of polar contacts, *i.e.* electrostatics and hydrogen bonds [27]. Here, we analysed the interaction of Rab25 with FIP3, an effector which has solely been crystallised bound to Rab11. We scanned the replaced residues that show signs of positive selection and that are involved in the interaction at least in Rab11 for their putative impact on the interaction. Figure 4.6 presents evidence that Rab25 may also have reduced affinity to FIP3, caused by the loss of a critical van der Waals interaction stabilising the C-terminal helical element of the Rab-binding domain of FIP3 that forms a hook-like structure [25]. An intriguing possibility that remains to be investigated is that this putative structural mechanism also provides the basis for the reduced affinity of Rab25 to FIP2. This finding is interesting in the context of Rab25 neofunctionalisation as it implies that some of the conserved effectors evolved reduced affinities to Rab25 driven by positive selection. Several possibilities exist for the beneficial func-

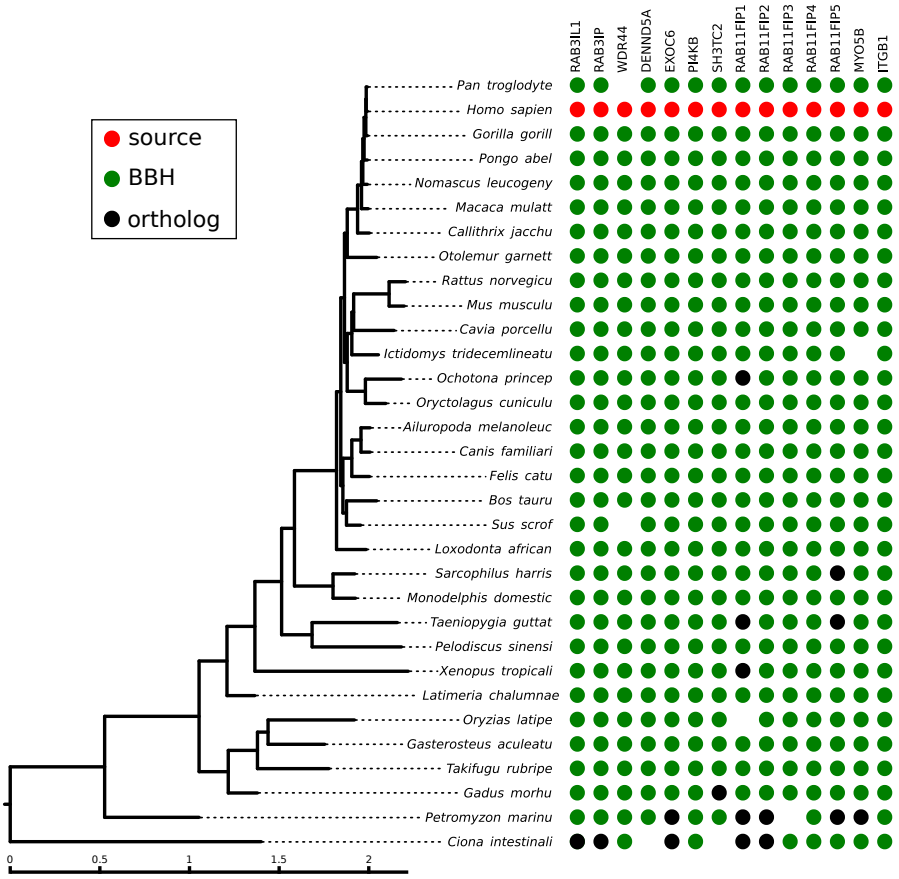


FIGURE 4.5: *Phylogenetic profile of Rab11/25 effectors*—Phylogenetic profile of the currently published Rab11/25 effectors (see Figure 4.4) in all species covered by the sequences analysed in Figures 4.1 and 4.2. Whenever no BBH is found, it is checked whether an ortholog is annotated in Ensembl Compara [43] (see Subsection 4.4.4 for details on BBHs and orthologs). *Abbreviation:* Bidirectional Best Hit (BBH)

tional effects of a reduction, for example reduced cross-talk with Rab11 also suggested in reference [27] (see also the discussion about loss of interactions in Subsection 1.2.1) or the resulting higher specificity of Rab25 to its other effectors. Obviously, these speculations are only warranted if our interpretations of the structural data we present in Figure 4.6 are validated.

In summary, we confirm the central prediction of the effector switching model that the gain of a new Rab25-specific effector had a beneficial functional effect and therefore represents at least part of the mechanism by which Rab25 evolved a new function. Although no structural data on lost effectors is available, tinkering with the strength of conserved interactions may also be part of this mechanism.

Alternative modes of Rab25 functional evolution

The above data supports our hypothesis of effector switching as a mechanism for neofunctionalisation. However, other non-exclusive modes of functional evolution may also contribute the the overall process of neofunctionalisation. At least three possibilities are supported by the available functional information on Rab25.

First, as discussed in Subsection 1.2.1 every protein functions in a cellular context primarily defined by its intracellular localisation, which determines for example which other proteins it can interact with. Hence, altering the localisation of a protein represents an interesting mechanism to evolve function [45, 46]. The Rab11 effector PI4KB has been shown to be sufficient for the localisation of Rab11 to the Golgi [41]. While there is no definite negative evidence, it is likely that Rab25 lost the ability to interact with PI4KB (see Figure 4.4). Hence, this would have resulted also in loss of Rab25 localisation to the Golgi, which unlike Rab11 is indeed exclusively found at recycling endosomes. Effector switching can thus indirectly explain other aspects of functional evolution in Rabs such as

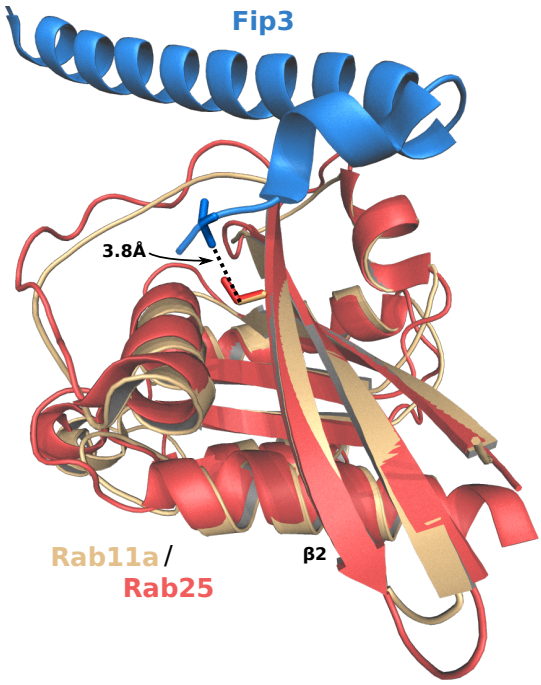


FIGURE 4.6: *Putative structural basis for reduced affinity of Rab25 to FIP3 compared to Rab11*—Rab25 loses a potentially critical van der Waals interaction between an Alanine (residue 50 in Figure 4.2) and a Valine in FIP3 by replacement with a Serine (residues with side-chains shown in stick representation). The interaction stabilises the C-terminal helical element of the Rab-binding domain of FIP3 (region in front after the turn, before the highlighted Valine residue) that forms a hook-like structure. Rab11 interacting with FIP3 [25] and Rab25 (interacting with FIP2, only the Rab shown [27]) have been overlaid without additional structural modelling (RMSD of superposition 0.484). Measurement between Rab11 Alanine and FIP3 Valine reveals ideal distance for van der Waals interaction. The interaction and its loss is additionally corroborated by the destabilisation of the interaction upon A50S replacement in Rab11 predicted to be 0.71 $\Delta\Delta G_{Bind}$ (kcal/mol) which is in agreement with loss of one van der Waals interaction (computation by BeAtMuSiC (version 1.0) [44]). Overlay, measurment and graphical representation have been done in PyMOL. PDB accessions: 3TSO (human Rab25), 2D7C (human Rab11, FIP3). *Abbreviations:* root mean square deviation (RMSD)

relocalisation.

Second, Rabs are part of complex networks that regulate the Rab itself via the action of GAPs and GEFs and couples them to other compartments, pathways and corresponding Rabs by what has been coined Rab cascades (see [47] for review). It appears natural that changing the interactions with these regulatory proteins affects the function of a Rab. Indeed, one of the few known regulators of Rab11, the GAP Ecotropic viral integration site 5 protein homolog (EVI5), does not interact with Rab25 [48, 49]. Hence, loss of the interaction with EVI5 may have provided a mechanism for functional evolution of Rab25 by allowing for a Rab11-independent regulation. Furthermore, several Rab11 effectors are known to be GEFs for other Rabs: RAB3IP/RABIN8 for Rab8 [50], RAB3IL1/GRAB for Rab8 [51] and Rab3a [52], and RAB6IP1/DENND5A for Rab39 [53]. At least RAB3IP/RABIN8 and RAB3IL1/GRAB are known not to interact with Rab25 (see Figure 4.4), and loss of these interactions therefore abolished cross-talk with Rab8 and 3a and their pathways. As a result, Rab25 was freed from regulatory constraints. Hence, a more general version of the effector switching model also covering the interactions with regulatory proteins accounts for an even greater mechanistic diversity involved in the evolution of Rab function.

Third, as already mentioned several times Rab25 has lost expression in all but epithelial cells compared to the ubiquitous expression profile of Rab11. Figure 2.8 shows that increasing tissue-specificity is a general phenomenon observed for all Metazoan-specific Rab subfamilies [8]. Hence, a sharpening expression profile represents another dimension of functional evolution well supported in the case of Rab25.

In conclusion, the above examples emphasise that besides the biochemical level we focus on with our effector switching model, neofunctionalisation also has a cell biological (localisation) and physiological (tissue specificity) dimension. Moreover, these different levels can be linked in non-trivial ways, as exemplified by localisation and regulatory coupling

which are achieved by the interaction with effector proteins.

4.2.3 Rab25 function evolved even long after duplication

A distinct question from the actual mechanism by which function evolved is to ask about the temporal dynamics of this process. In the case of Rabs, this is particularly interesting as the general consensus states that Rab function is stable even over long evolutionary time periods. The question to what extent this is indeed the case becomes relevant for example to interpret the bioinformatic annotation of the Rab family in new species: in Chapter 2, we argue that classifying a Rab sequence is a strong functional statement ultimately even allowing to infer the presence of entire compartments and pathways. The availability of functional information on various model organisms allows to detail the temporal dynamics of the evolution of function in Rab25. Concretely, we ask if the function of Rab25 is the same in zebrafish and mammals. As can be seen for example from the tree in Figure 4.1, the fish clade is the first one to branch off and therefore comprises the Rab25 proteins in our dataset that evolved independently from the mammalian Rab25 for the longest time.

Our approach is to consider two aspects of Rab25 function: the interaction with ITGB1 and the pattern of tissue specificity. In order to assess how likely it is that zebrafish Rab25 interacts with ITGB1, we compared the C-terminal region which we know harbours the binding site for ITGB1 in mammals [28]. As shown in Figure 4.7a, the termini have highly diverged between zebrafish and mammals. They share only around 30% sequence identity and two indels have occurred. Hence, it is possible that the binding site evolved later and zebrafish Rab25 does not interact with ITGB1. We agree that this indirect experiment has limited power, amongst others because binding sites in unstructured regions are generally less constrained and conserved [54]. Secondly, we compared the tissue-specificity of Rab25 in zebrafish and mammals represented by mouse and human.

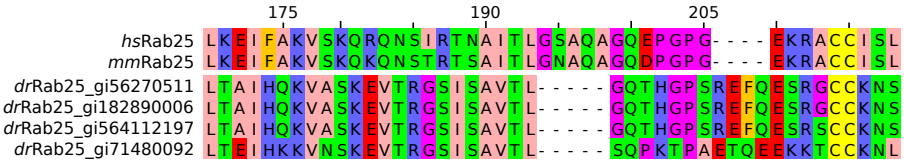
Figure 4.7b summarises data indicating that the loss of Rab25 expression in heart observed in mammals occurred after the branching of the fish clade. Again, there may be alternative explanations for this pattern as epithelial expression is not directly assessed. Yet, both experiments suggest the hypothesis that Rab25 functions differently in zebrafish and mammals.

In conclusion, Rab25 function may have evolved for a period largely extending beyond the time shortly after duplication. This agrees with the relatively long Rab25 branch highlighted in Figure 4.2 which suggests continuous sequence divergence before the appearance of mammals. Hence, Rab function at least in non-ancestral subfamilies is more dynamic than suggested previously for example by successful rescue experiments between mouse and yeast [59], and may differ qualitatively between organisms. Bioinformatic annotations can therefore only be interpreted as rough functional statements.

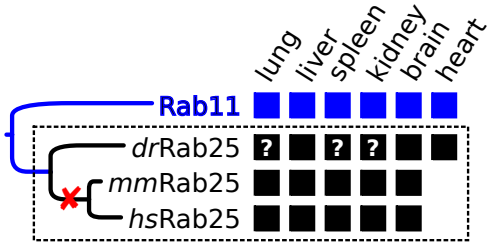
4.3 Conclusion

In this chapter, we have asked about the mechanisms by which new functions evolve in the context of gene duplication. We used Rab11 and 25 as a model system, which we demonstrate evolved by neofunctionalisation. We suggested a model of effector switching (summarised in Figure 4.8) which is able to account for the evolutionary patterns we observe in Rab25. Furthermore, we found that this process took place in a time window extending far beyond the actual gene duplication that gave rise to Rab25.

Rabs show an interesting behaviour with respect to the evolution of their function: on one hand, mainly ancestral Rabs but also for example Rab25 in mammals are highly constrained, *i.e.* subject to strong purifying



(A)



(B)

FIGURE 4.7: *Zebrafish and mammalian Rab25 may have different functions*—(**A**) Multiple sequence alignment of the C-terminal tail of Rab25 sequences in zebrafish, mouse, and human, generated with PRANK [55], colour and graphical representation by Jalview [56]. (**B**) Expression of Rab25 in various tissues in zebrafish, mouse, and human. Rab11 and 25 expression in mouse assessed by PCR, data replicated from Figure 2.8 [8]. Protein expression in humans obtained from the Human Protein Atlas [57] (accession ENSG00000132698). Zebrafish expression from GEO profiles [58], accessions: heart (32317921, 66454322), liver (11990612), brain (35418821). No data found on zebrafish lung, spleen, kidney (indicated by white question marks). The red cross indicates the parsimonious inference of when expression in heart has been lost.

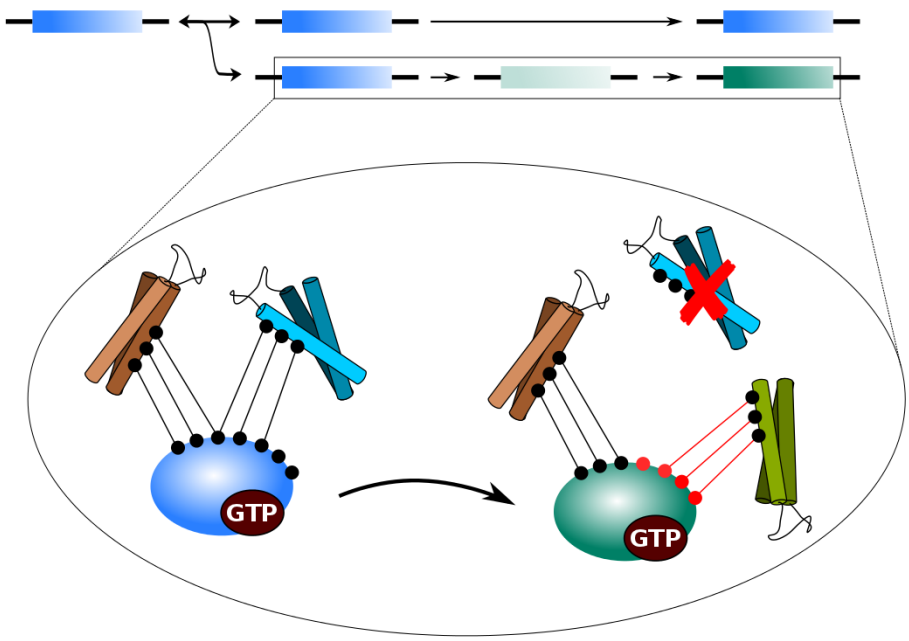


FIGURE 4.8: *The effector switching model for Rab neofunctionalisation*— On top, the classical neofunctionalisation model [6] is shown including an intermediate state between the old and new gene. Below, the proposed biochemical mechanism for the evolution of a new function is detailed. New Rab functions evolve by replacements of residues (small circles) that alter interaction interfaces and ultimately lead to the loss and gain of effectors. In agreement with the neofunctionalisation model, this process is driven by positive selection (red circles). Blue and green stand for distinct functions. The intermediate state included in the above panel represents the possibility that new Rab functions evolve over an extended period of time starting with reduction of affinity to existing effectors without having evolved interaction interfaces to new effectors yet (compare Figures 4.6 and 4.7).

selection, and barely diverge in sequence and function over long periods of time. On the other hand, Rabs duplicate and neofunctionalise frequently, resulting in a gene family that roughly tripled its size at least in Metazoan evolution [8]. This represents a striking contrast: functional stasis on one side and a dynamic family with great innovatory potential on the

other. An interesting argument made in the context of Hox genes may be able to resolve this apparent conflict. Due to the strong constraints, Rabs are not evolvable. Duplications temporarily lift these constraints on one copy, creating a "window of opportunity" in which the dramatically increased innovatory potential can be exploited given the right ecological and developmental conditions [60]. Moreover, as we reviewed in Chapter 1 forming new interactions may not require a lot of mutations which would allow Rabs to functionally evolve rather easily. This argument opens a promising avenue for future work on Rabs, namely to investigate possible agents of selection that link Rab evolution and its impact on intracellular organisation to the historical ecological and developmental context.

In conclusion, we have presented an integrated evolutionary and biochemical model for neofunctionalisation, addressing both the complementary aspects of emergence of new genes and new functions. While we focussed on Rabs, we believe that effector switching provides a promising basis for a more general model especially well suited for the attractive and important class of master regulators. These proteins are expected nearly by definition to exert their function via the interaction with numerous partners, an archetypical functional paradigm also followed by Rabs. Future work will therefore focus on different families of master regulators with aim to confirm and generalise the model corroborated here for Rabs only.

4.4 Materials and Methods

4.4.1 Alignment and gene tree

The protein sequences corresponding to the accessions listed in Table 4.1 were multiply aligned with PRANK [61] and a preliminary gene tree

generated with PhyML [62]. The resulting tree clearly showed three well supported clades corresponding to Rab11a, Rab11b and Rab25. Each of these clades was manually reconciled based on the species tree provided by Ensembl [63] (shown as part of Figure 4.5). The branch lengths of the reconciled gene tree were obtained on the fixed reconciled topology with PhyML. Next, the multiple sequence alignment (MSA) was recomputed with PRANK providing the reconciled gene tree as input. Lastly, the resulting MSA was trimmed using ZORRO [64] with standard parameters and providing the reconciled gene tree as input.

4.4.2 Ancestral sequence reconstruction

First, the protein alignments described above were translated into codon alignments with Pal2Nal [65]. Then, ancestral sequences were reconstructed on these and the reconciled gene tree also described in Subsection 4.4.1 using PAML (version 4.6) [66] under a branch-site model. The entire branch from the initial duplication to the extant human Rab25 sequence was marked as foreground (highlighted in red in Figure 4.1), *i.e.* allowed to evolve with ω (defined as the nonsynonymous to synonymous substitution rate ratio) greater than one.

4.4.3 Past episodic positive selection

The presence of sites that evolved under positive selection in the foreground branches was tested using the standard Branch-Site Test of Positive Selection [67, 68] as implemented by PAML (version 4.6) [66]. The likelihood ratio test (LRT) was considered significant when above 2.71 corresponding to $p < 0.05$. In case of a significant test, the putative sites were obtained by Bayes Empirical Bayes [69]. We used a relatively low threshold, 0.7, where usually 0.95 is common. However, we tested all intervals defined by every pairwise choice of branch labeled by $1 - 6$ in

Figure 4.2 as foreground and therefore expect to exclude potential false positives by only considering sites that are above the likelihood threshold of 0.7 in more than one case. We note that we failed to apply a multiple testing correction for the LRT, however, we do not expect this to influence our results. Nonetheless, future versions will include the correct procedure advised in reference [70].

In each case, the reconciled topology was used as input rather than a common maximum likelihood tree. Branch lengths are estimated by PAML. This choice follows from the experiments presented in Chapter 3 suggesting that tree quality is a critical parameter for the inference of sites that evolved under episodic positive selection.

4.4.4 Bidirectional Best Hits and orthologs

Bidirectional Best Hits (BBHs) are computed with Blast [71] (e-value threshold 10^{-3}) respecting two additional constraints defined by the INPARANOID project [72]: coverage, *i.e.* 50% of the residues must be covered by the local alignment, and overlap, meaning that 25% of the residues must be aligned [73]. Additionally, in case of zero e-values all hits are kept, meaning that one-to-many and even many-to-many relationships are possible although the intuitive definition of BBHs seems to exclude this possibility. Orthologs from Ensembl Compara [43] are obtained by retrieving pairs of leaves whose last common ancestor in the gene trees is marked as a speciation event.

4.A Supplementary tables

TABLE 4.1: *List of Rab11 and Rab25 sequences used in phylogenetic analysis.*—The initial annotations of Rab sequences were generated by the Rabifier (see Chapter 2) [8] on full genomes downloaded from the Ensembl (version 70) [63] and Ensembl Genomes (version 17) [74] databases. Apparently truncated sequences and other outliers visually identified in the multiple sequence alignment and through long branches in the gene tree (see Subsection 4.4.1) were removed.

Ensembl accession	species	Rab subfamily
ENSOCUT00000005636	*Oryctolagus cuniculus*	Rab25
ENSXETT00000054294	*Xenopus tropicalis*	Rab11
ENSSSCT00000007117	*Sus scrofa*	Rab25
ENSSTOT00000015254	*Ictidomys tridecemlineatus*	Rab11
ENSLACT00000019294	*Latimeria chalumnae*	Rab11
ENSMODT00000004807	*Monodelphis domestica*	Rab11
ENSSSCT00000024720	*Sus scrofa*	Rab11
ENSTRUT00000026540	*Takifugu rubripes*	Rab11
ENSPTRT00000045045	*Pan troglodytes*	Rab25
ENSSHAT00000007680	*Sarcophilus harrisii*	Rab11
ENSAMET00000015643	*Ailuropoda melanoleuca*	Rab25
ENSCINT00000023970	*Ciona intestinalis*	Rab11
ENSORLT00000019146	*Oryzias latipes*	Rab11
ENSCAFT00000048184	*Canis familiaris*	Rab11
ENSMMUT00000031379	*Macaca mulatta*	Rab25
ENSORLT00000022954	*Oryzias latipes*	Rab11
ENSXETT00000044022	*Xenopus tropicalis*	Rab25
ENSLACT00000012758	*Latimeria chalumnae*	Rab25
ENSGACT00000017553	*Gasterosteus aculeatus*	Rab25
ENSGACT00000020470	*Gasterosteus aculeatus*	Rab11
ENSPPYT00000007758	*Pongo abelii*	Rab11
ENSPSIT00000011394	*Pelodiscus sinensis*	Rab11

Ensembl accession	species	Rab subfamily
ENSTGUT00000003823	*Taeniopygia guttata*	Rab25
ENSTRUT00000000486	*Takifugu rubripes*	Rab11
ENST00000261890	*Homo sapiens*	Rab11
ENSTRUT00000046937	*Takifugu rubripes*	Rab11
ENSGGOT00000014301	*Gorilla gorilla*	Rab11
ENSPMAT00000010308	*Petromyzon marinus*	Rab11
ENSBTAT00000003206	*Bos taurus*	Rab11
ENSRNOT00000010197	*Rattus norvegicus*	Rab11
ENSOPRT00000004669	*Ochotona princeps*	Rab11
ENSAMET00000005460	*Ailuropoda melanoleuca*	Rab11
ENSTGUT00000000085	*Taeniopygia guttata*	Rab11
ENSMUST00000172298	*Mus musculus*	Rab11
ENSCPOT00000006527	*Cavia porcellus*	Rab11
ENSSHAT00000015768	*Sarcophilus harrisii*	Rab25
ENSLACT00000013937	*Latimeria chalumnae*	Rab11
ENSSTOT00000008074	*Ictidomys tridecemlineatus*	Rab11
ENSBTAT00000025235	*Bos taurus*	Rab11
ENSXETT00000014541	*Xenopus tropicalis*	Rab11
ENSTRUT00000046916	*Takifugu rubripes*	Rab11
ENSGMOT00000008127	*Gadus morhua*	Rab11
ENSORLT00000019171	*Oryzias latipes*	Rab11
ENSLAFT00000011417	*Loxodonta africana*	Rab11
ENSFCAT00000015220	*Felis catus*	Rab11
ENST00000361084	*Homo sapiens*	Rab25
ENSFCAT00000025321	*Felis catus*	Rab11
ENSOCUT00000014754	*Oryctolagus cuniculus*	Rab11
ENSCPOT00000028204	*Cavia porcellus*	Rab11
ENSSTOT00000000636	*Ictidomys tridecemlineatus*	Rab25

Ensembl accession	species	Rab subfamily
ENSPTRT00000019157	*Pan troglodytes*	Rab11
ENSNLET00000002364	*Nomascus leucogenys*	Rab11
ENSCJAT00000015098	*Callithrix jacchus*	Rab25
ENSOPRT00000002901	*Ochotona princeps*	Rab25
ENSGMOT00000010104	*Gadus morhua*	Rab11
ENSPSIT00000014617	*Pelodiscus sinensis*	Rab25
ENSRNOT00000015598	*Rattus norvegicus*	Rab11
ENSPPYT00000011079	*Pongo abelii*	Rab11
ENSMMUT00000011193	*Macaca mulatta*	Rab11
ENSLAFT00000013596	*Loxodonta africana*	Rab11
ENSPTRT00000013265	*Pan troglodytes*	Rab11
ENSFCAT00000011777	*Felis catus*	Rab25
ENSPSIT00000013351	*Pelodiscus sinensis*	Rab11
ENSMUST00000057373	*Mus musculus*	Rab11
ENSAMET00000003419	*Ailuropoda melanoleuca*	Rab11
ENSSHAT00000007917	*Sarcophilus harrisii*	Rab11
ENSNLET00000016408	*Nomascus leucogenys*	Rab11
ENSGACT00000006586	*Gasterosteus aculeatus*	Rab25
ENSMMUT00000014295	*Macaca mulatta*	Rab11
ENSGMOT00000004555	*Gadus morhua*	Rab11
ENSPPYT00000000878	*Pongo abelii*	Rab25
ENSOGAT00000007646	*Otolemur garnettii*	Rab11
ENSCAFT00000026699	*Canis familiaris*	Rab25
ENSORLT00000020191	*Oryzias latipes*	Rab25
ENSOGAT00000016959	*Otolemur garnettii*	Rab25
ENSGGOT00000014564	*Gorilla gorilla*	Rab11
ENSRNOT00000032355	*Rattus norvegicus*	Rab25
ENST00000328024	*Homo sapiens*	Rab11

Ensembl accession	species	Rab subfamily
ENSGGOT00000006848	*Gorilla gorilla*	Rab25
ENSGACT00000017634	*Gasterosteus aculeatus*	Rab11
ENSCJAT00000014411	*Callithrix jacchus*	Rab11
ENSCAFT00000027321	*Canis familiaris*	Rab11
ENSOGAT00000003230	*Otolemur garnettii*	Rab11
ENSMUST00000008745	*Mus musculus*	Rab25
ENSORLT00000015209	*Oryzias latipes*	Rab25
ENSMODT00000012486	*Monodelphis domestica*	Rab11
ENSSSCT00000014854	*Sus scrofa*	Rab11
ENSBTAT00000025170	*Bos taurus*	Rab25
ENSTGUT00000009676	*Taeniopygia guttata*	Rab11
ENSMODT00000021525	*Monodelphis domestica*	Rab25
ENSLAFT00000009222	*Loxodonta africana*	Rab25
ENSNLET00000015706	*Nomascus leucogenys*	Rab25
ENSTRUT00000018665	*Takifugu rubripes*	Rab25
ENSTRUT00000037427	*Takifugu rubripes*	Rab25
ENSCPOT00000002505	*Cavia porcellus*	Rab25

References

[1] H Allen Orr. “Theories of adaptation: what they do and don’t say”. In: *Genetica* 123.1-2 (Feb. 2005), pp. 3–13.

[2] Gerd B Müller and Stuart A Newman. “The innovation triad: an EvoDevo agenda”. In: *Journal of Experimental Zoology Part B, Molecular and Developmental Evolution* 304.6 (Nov. 2005), pp. 487–503.

[3] Massimo Pigliucci. “What, if Anything, Is an Evolutionary Novelty?” In: *Philosophy of Science* 75 (2008), pp. 887–898.

[4] Günter P Wagner and Vincent J Lynch. “Evolutionary novelties”. In: *Current Biology* 20.2 (Jan. 2010), R48–52.

[5] Antony M Dean and Joseph W Thornton. “Mechanistic approaches to the study of evolution: the functional synthesis”. In: *Nature Reviews Genetics* 8.9 (Sept. 2007), pp. 675–688.

[6] Susumu Ohno. *Evolution by Gene Duplication*. Springer-Verlag, 1970.

[7] Emile Zuckerkandl and Linus Pauling. “Evolutionary divergence and convergence in proteins”. In: *Evolving genes and proteins* (1965), pp. 97–166.

[8] Yoan Diekmann et al. “Thousands of Rab GTPases for the Cell Biologist”. In: *PLoS Computational Biology* 7.10 (Oct. 2011), e1002217.

[9] Joanne Young, Julie Ménétrey, and Bruno Goud. “RAB6C is a retrogene that encodes a centrosomal protein involved in cell cycle progression”. In: *Journal of Molecular Biology* 397.1 (Mar. 2010), pp. 69–88.

[10] Eoin E Kelly, Conor P Horgan, and Mary W McCaffrey. “Rab11 proteins in health and disease”. In: *Biochemical Society Transactions* 40.6 (Dec. 2012), pp. 1360–1367.

[11] Roshan Agarwal et al. “The emerging role of the RAB25 small GTPase in cancer”. In: *Traffic* 10.11 (Nov. 2009), pp. 1561–1568.

[12] A Kikuchi et al. “Purification and characterization of a novel GTP-binding protein with a molecular weight of 24,000 from bovine brain membranes”. In: *Journal of Biological Chemistry* 263.6 (Feb. 1988), pp. 2897–2904.

[13] P Chavrier et al. “Molecular cloning of YPT1/SEC4-related cDNAs from an epithelial cell line”. In: *Molecular and Cellular Biology* 10.12 (Dec. 1990), pp. 6578–6585.

[14] S Urbé et al. “Rab11, a small GTPase associated with both constitutive and regulated secretory pathways in PC12 cells”. In: *FEBS Letters* 334.2 (Nov. 1993), pp. 175–182.

[15] O Ullrich et al. “Rab11 regulates recycling through the pericentriolar recycling endosome”. In: *The Journal of Cell Biology* 135.4 (Nov. 1996), pp. 913–924.

[16] Hugh R B Pelham. “Insights from yeast endosomes”. In: *Current Opinion in Cell Biology* 14.4 (Aug. 2002), pp. 454–462.

[17] G Jedd, J Mulholland, and N Segev. “Two new Ypt GTPases are required for exit from the yeast trans-Golgi compartment”. In: *The Journal of Cell Biology* 137.3 (May 1997), pp. 563–580.

[18] Shu Hui Chen et al. “Ypt31/32 GTPases and their novel F-box effector protein Rcy1 regulate protein recycling”. In: *Molecular Biology of the Cell* 16.1 (Jan. 2005), pp. 178–192.

[19] K Sakurada et al. “Molecular cloning and characterization of a ras p21-like GTP-binding protein (24KG) from rat liver”. In: *Biochemical and Biophysical Research Communications* 177.3 (June 1991), pp. 1224–1232.

[20] J R Goldenring et al. “Identification of a small GTP-binding protein, Rab25, expressed in the gastrointestinal mucosa, kidney, and lung”. In: *Journal of Biological Chemistry* 268.25 (Sept. 1993), pp. 18419–18422.

[21] J E Casanova et al. “Association of Rab25 and Rab11a with the apical recycling system of polarized Madin-Darby canine kidney cells”. In: *Molecular Biology of the Cell* 10.1 (Jan. 1999), pp. 47–61.

[22] Sebastiano Pasqualato et al. “The structural GDP/GTP cycle of Rab11 reveals a novel interface involved in the dynamics of recycling endosomes”. In: *Journal of Biological Chemistry* 279.12 (Mar. 2004), pp. 11480–11488.

[23] William N Jagoe et al. “Crystal structure of rab11 in complex with rab11 family interacting protein 2”. In: *Structure* 14.8 (Aug. 2006), pp. 1273–1283.

[24] T Shiba et al. “Structural basis for Rab11-dependent membrane recruitment of a family of Rab11-interacting protein 3 (FIP3)/Arfophilin-1”. In: *Proceedings of the National Academy of Sciences of the United States of America* 103.42 (Oct. 2006), pp. 15416–15421.

[25] Sudharshan Eathiraj et al. “Structural basis for Rab11-mediated recruitment of FIP3 to recycling endosomes”. In: *Journal of Molecular Biology* 364.2 (Nov. 2006), pp. 121–135.

[26] Olena Pylypenko et al. “Structural basis of myosin V Rab GTPase-dependent cargo recognition”. In: *Proceedings of the National Academy*

of Sciences of the United States of America 110.51 (Dec. 2013), pp. 20443–20448.

[27] Patrick Lall et al. “Structural and functional analysis of FIP2 binding to the endosome-localised Rab25 GTPase”. In: *Biochimica Et Biophysica Acta* 1834.12 (Dec. 2013), pp. 2679–2690.

[28] Patrick T Caswell et al. “Rab25 associates with $\alpha 5\beta 1$ integrin to promote invasive migration in 3D microenvironments”. In: *Developmental Cell* 13.4 (Oct. 2007), pp. 496–510.

[29] C M Hales et al. “Identification and characterization of a family of Rab11-interacting proteins”. In: *Journal of Biological Chemistry* 276.42 (Oct. 2001), pp. 39067–39075.

[30] Mitsunori Fukuda et al. “Large scale screening for novel rab effectors reveals unexpected broad Rab binding specificity”. In: *Molecular & Cellular Proteomics* 7.6 (June 2008), pp. 1031–1042.

[31] R Prekeris, J M Davies, and R H Scheller. “Identification of a novel Rab11/25 binding domain present in Eferin and Rip proteins”. In: *Journal of Biological Chemistry* 276.42 (Oct. 2001), pp. 38966–38970.

[32] Deborah M E Wallace et al. “Rab11-FIP4 interacts with Rab11 in a GTP-dependent manner and its overexpression condenses the Rab11 positive compartment in HeLa cells”. In: *Biochemical and Biophysical Research Communications* 299.5 (Dec. 2002), pp. 770–779.

[33] Gilles R X Hickson et al. “Arfophilins are dual Arf/Rab 11 binding proteins that regulate recycling endosome distribution and are related to Drosophila nuclear fallout”. In: *Molecular Biology of the Cell* 14.7 (July 2003), pp. 2908–2920.

[34] R Prekeris, Judith Klumperman, and R H Scheller. “A Rab11/Rip11 protein complex regulates apical membrane trafficking via recycling endosomes”. In: *Molecular Cell* 6.6 (Dec. 2000), pp. 1437–1448.

[35] L A Lapierre et al. “Myosin Vb is associated with plasma membrane recycling systems”. In: *Molecular Biology of the Cell* 12.6 (June 2001), pp. 1843–1857.

[36] Joseph T Roland, Lynne A Lapierre, and James R Goldenring. “Alternative splicing in class V myosins determines association with Rab10”. In: *Journal of Biological Chemistry* 284.2 (Jan. 2009), pp. 1213–1223.

[37] Conor P Horgan, Sara R Hanscom, and Mary W McCaffrey. “GRAB is a binding partner for the Rab11a and Rab11b GTPases”. In: *Biochemical and Biophysical Research Communications* 441.1 (Nov. 2013), pp. 214–219.

[38] Stéphanie Miserey-Lenkei et al. “Rab6-interacting protein 1 links Rab6 and Rab11 function”. In: *Traffic* 8.10 (Oct. 2007), pp. 1385–1403.

[39] Claudia Stendel et al. “SH3TC2, a protein mutant in Charcot-Marie-Tooth neuropathy, links peripheral nerve myelination to endosomal recycling”. In: *Brain* 133.Pt 8 (Aug. 2010), pp. 2462–2474.

[40] Xiang-Ming Zhang et al. “Sec15 is an effector for the Rab11 GTPase in mammalian cells”. In: *Journal of Biological Chemistry* 279.41 (Oct. 2004), pp. 43027–43034.

[41] Petra de Graaf et al. “Phosphatidylinositol 4-Kinaseβ is critical for functional association of rab11 with the Golgi complex”. In: *Molecular Biology of the Cell* 15.4 (Apr. 2004), pp. 2038–2047.

[42] J Zeng et al. “Identification of a putative effector protein for rab11 that participates in transferrin recycling”. In: *Proceedings of the National Academy of Sciences of the United States of America* 96.6 (Mar. 1999), pp. 2840–2845.

[43] Albert J Vilella et al. “EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates”. In: *Genome Research* 19.2 (Feb. 2009), pp. 327–335.

[44] Yves Dehouck et al. “BeAtMuSiC: Prediction of changes in protein-protein binding affinity on mutations”. In: *Nucleic Acids Research* 41.Web Server issue (July 2013), W333–9.

[45] Ana Claudia Marques et al. “Functional diversification of duplicate genes through subcellular adaptation of encoded proteins”. In: *Genome Biology* 9.3 (2008), R54.

[46] Xiujuan Wang et al. “Comparative study of human mitochondrial proteome reveals extensive protein subcellular relocalization after gene duplications”. In: *BMC Evolutionary Biology* 9 (2009), p. 275.

[47] Emi Mizuno-Yamasaki, Felix Rivera-Molina, and Peter Novick. “GTPase networks in membrane traffic”. In: *Annual Review of Biochemistry* 81 (2012), pp. 637–659.

[48] Christopher J Westlake et al. “Identification of Rab11 as a small GTPase binding protein for the Evi5 oncogene”. In: *Proceedings of the National Academy of Sciences of the United States of America* 104.4 (Jan. 2007), pp. 1236–1241.

[49] J T S Dabbeekeh et al. “The EVI5 TBC domain provides the GTPase-activating protein motif for RAB11”. In: *Oncogene* 26.19 (Apr. 2007), pp. 2804–2808.

[50] Andreas Knödler et al. “Coordination of Rab8 and Rab11 in primary ciliogenesis”. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.14 (Apr. 2010), pp. 6346–6351.

[51] Zhong Guo et al. “Intermediates in the guanine nucleotide exchange reaction of Rab8 protein catalyzed by guanine nucleotide exchange factors Rabin8 and GRAB”. In: *The Journal of biological chemistry* 288.45 (Nov. 2013), pp. 32466–32474.

[52] H R Luo et al. “GRAB: a physiologic guanine nucleotide exchange factor for Rab3A, which interacts with inositol hexakisphosphate kinase”. In: *Neuron* 31.3 (Aug. 2001), pp. 439–451.

[53] Shin-ichiro Yoshimura et al. “Family-wide characterization of the DENN domain Rab GDP-GTP exchange factors”. In: *The Journal of Cell Biology* 191.2 (Oct. 2010), pp. 367–381.

[54] Bálint Mészáros et al. “Molecular Principles of the Interactions of Disordered Proteins”. In: *Journal of Molecular Biology* 372.2 (Sept. 2007), pp. 549–561.

[55] Ari Löytynoja and Nick Goldman. “webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser”. In: *BMC Bioinformatics* 11 (2010), p. 579.

[56] Andrew M Waterhouse et al. “Jalview Version 2—a multiple sequence alignment editor and analysis workbench”. In: *Bioinformatics* 25.9 (May 2009), pp. 1189–1191.

[57] Mathias Uhlen et al. “Towards a knowledge-based Human Protein Atlas”. In: *Nature Biotechnology* 28.12 (Dec. 2010), pp. 1248–1250.

[58] Tanya Barrett et al. “NCBI GEO: archive for functional genomics data sets—update”. In: *Nucleic Acids Research* 41.Database issue (Jan. 2013), pp. D991–5.

[59] H Haubruck et al. “The ras-related mouse ypt1 protein can functionally replace the YPT1 gene product in yeast”. In: *The EMBO Journal* 8.5 (May 1989), pp. 1427–1432.

[60] Günter P Wagner, Chris T Amemiya, and F Ruddle. “Hox cluster duplications and the opportunity for evolutionary novelties”. In: *Proceedings of the National Academy of Sciences of the United States of America* 100.25 (Dec. 2003), pp. 14603–14606.

[61] Ari Löytynoja and Nick Goldman. “An algorithm for progressive multiple alignment of sequences with insertions”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.30 (July 2005), pp. 10557–10562.

[62] Stéphane Guindon et al. “New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0”. In: *Systematic Biology* 59.3 (May 2010), pp. 307–321.

[63] Paul Flicek et al. “Ensembl 2013”. In: *Nucleic Acids Research* 41.D1 (Dec. 2012), pp. D48–D55.

[64] Martin Wu, Sourav Chatterji, and Jonathan A Eisen. “Accounting for alignment uncertainty in phylogenomics”. In: *PLoS ONE* 7.1 (2012), e30288.

[65] Mikita Suyama, David Torrents, and Peer Bork. “PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments”. In: *Nucleic Acids Research* 34.Web Server issue (July 2006), W609–12.

[66] Ziheng Yang. “PAML 4: phylogenetic analysis by maximum likelihood”. In: *Molecular Biology and Evolution* 24.8 (Aug. 2007), pp. 1586–1591.

[67] Ziheng Yang and Rasmus Nielsen. “Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages”. In: *Molecular Biology and Evolution* 19.6 (June 2002), pp. 908–917.

[68] Jianzhi Zhang, Rasmus Nielsen, and Ziheng Yang. “Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level”. In: *Molecular Biology and Evolution* 22.12 (Dec. 2005), pp. 2472–2479.

[69] Ziheng Yang, Wendy S W Wong, and Rasmus Nielsen. “Bayes empirical bayes inference of amino acid sites under positive selection”. In: *Molecular Biology and Evolution* 22.4 (Apr. 2005), pp. 1107–1118.

[70] Maria Anisimova and Ziheng Yang. “Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites”. In: *Molecular Biology and Evolution* 24.5 (May 2007), pp. 1219–1228.

[71] Christiam Camacho et al. “BLAST+: architecture and applications”. In: *BMC Bioinformatics* 10 (2009), p. 421.

[72] M Remm, C E Storm, and Erik L L Sonnhammer. “Automatic clustering of orthologs and in-paralogs from pairwise species comparisons”. In: *Journal of Molecular Biology* 314.5 (Dec. 2001), pp. 1041–1052.

[73] Gabriel Ostlund et al. “InParanoid 7: new algorithms and tools for eukaryotic orthology analysis”. In: *Nucleic Acids Research* 38.Database issue (Jan. 2010), pp. D196–203.

[74] Paul J Kersey et al. “Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species”. In: *Nucleic Acids Research* 40.Database issue (Jan. 2012), pp. D91–7.

Chapter 5

Conclusion

"Nothing retains the form that seems its own, and Nature, the renewer of all things, continually changes every form into some other shape." [1, book XV, verse 252]

—Ovid, 8AD

How does protein function evolve? This is the fundamental question we have asked in this thesis, using the Rab family of small GTPases as a model system.

We began most generally in Chapter 1 by charting the phenotypic space of protein function and reviewing some previous work on the evolutionary dynamics of proteins through that space. We distinguished three levels, genetics, biochemistry and evolution, all of which are pertinent to the question how protein function evolves. They cover distinct but related aspects of the phenomenon. We focussed our discussion around the biochemical level, as protein function itself is commonly receiving least attention in studies on protein evolution. At the same time, the "function" of a protein at the level of biochemistry is less complex and interconnected than at higher levels of biological organisation like the cell or the multicellular organism. Therefore, biochemistry represents a natural starting point to integrate function and evolution [2, 3]. At the end of Chapter 1, we introduced the Rab family of small GTPases as a promising model system to contribute to the understanding of the evolution of function. However, while the rest of the thesis focusses on Rabs, the scope of our conclusions are broader.

Evolution is dynamic, and understanding dynamics requires comparing. Comparative approaches in Biology have a long tradition, and start with a simple question: "what is there?". We therefore began our analysis of the evolution of function in Rabs by establishing the patterns of Rab evolution. Doing so posed the first challenge, how to identify and classify Rabs in the many—with exception of some model organisms—incomplete, badly annotated and constantly updated genomes. We solved this problem in Chapter 2 by designing a bioinformatic tool dedicated to the detection and annotation of Rabs in whole genomes that we coined the Rabifier. While 'annotation' may seem a rather prosaic preoccupa-

tion to most including myself, all following analyses are dependent on the quality of the data set of Rabs. The Rabifier thus remains an apt solution to an important problem and a major deliverable of this thesis. At the end of Chapter 2 stands a map of the Rab universe, and emerging from it the old model of neofunctionalisation [4] as the new hypothesis about the dominant process of functional evolution in Rabs. To a 'Rabologist', this is no striking news, as the ongoing functional characterisation since the early 1980's continuously expands the functional repertoire of the Rab family and thus leaves little doubt that new functions must have evolved at some point. However, it is the process itself that is interesting in the context of this thesis, and less so its precise functional outcome.

Even more than for the comparative study of pattern, analysing an evolutionary process can only be done in the context of a phylogeny [5]. Rabs however turn out to be a challenging family for phylogenetic methods. Their sequences are short and consist nearly exclusively of very conserved or highly divergent parts. As a result, Rabs carry little phylogenetic information that can be exploited for gene tree inference. Yet, a gene tree is strictly required for example to verify one of the central predictions of the neofunctionalisation model, that divergence in sequence and function is driven by positive selection. This problem motivated the work presented in Chapter 3. While the effects of various factors relating to sequences and alignments on the standard approach to test for past positive selection have been recognised and studied, the gene tree had so far received no attention and it remained unclear if tree topology is important at all. Motivated by the expectation of topologically erroneous trees for Rabs and the need to detect selection, we therefore tested if the gene tree has any impact at all on the quality of the inference of sites under positive selection. Using simulated sequences, we indeed found a negative effect of erroneous gene trees. Because this result is based on simulations, it is not specific to Rabs and holds for any protein family. As a consequence

of the findings in Chapter 3, the following phylogenetic experiments were all performed based on a manually reconciled tree topology. While we did not show that reconciled trees are better, in the case of Rabs we expect them to more adequately reflect the true historical gene divergence patterns than trees inferred for example by Maximum Likelihood.

Finally, Chapter 4 addresses the core issue, the biochemical mechanism behind neofunctionalisation. The latter had been found to describe the evolution of the Rab family in Chapter 2. Our approach is that of classical hypothesis testing. Based on the review of some of what is known about the evolution of function and in particular of the biochemistry and cell biology of Rab function in Chapter 1, we propose and test a model we coin 'effector switching'. The reasoning is straight-forward. The most frequent evolutionary mechanism to evolve function is altering interactions, and Rabs are known to mediate their function by interaction with effectors that are recruited to the membranous compartments. Hence, we suggest that Rabs evolve by altering their set of effectors. We test our hypothesis by successfully verifying three predictions of effector switching for a representative pair of Rab subfamilies, Rab11 and 25. Furthermore, we discuss extensions of the simple model restricted to interactions with effectors: some effectors determine Rab subcellular localisation, other interactions have regulatory roles for the Rab itself and others, and finally restriction of expression profiles to certain tissues provide additional mechanisms for the evolution of function at the level of Rab biochemistry, cell biology and physiology. In conclusion, Chapter 4 begins to answer how protein function evolves in Rabs, putting forth a basic but general mechanistic scaffold that can easily be extended to account for additional more specific phenomena.

Concluding, what have we learned about the evolution of function? By our choice of model system, we have focussed on proteins that func-

tion as master regulators, and our contribution is thus most relevant to proteins following this functional paradigm. First, master regulators are commonly subject to strong evolutionary constraints, and gene duplication is therefore expected to play an important role in the functional evolution of this class of proteins. After duplication, the former strength of these constraints is no more inversely related to innovatory potential or evolvability. Second, almost by definition master regulators have to interact with numerous partners to exert the high-level control of cellular and organismal processes. Loss of some of these interactions and gain of others therefore represents a general mechanism for the evolution of function in these proteins.

5.1 Outlook

The work presented here will continue in three different ways. First and as an immediate goal, several technical aspects can be improved. This most importantly concerns the Rabifier. The initial design of the Rabifier was facing the classical tradeoff in algorithm and tool design: speed versus accuracy. While bioinformatic approaches are generally less accurate, they are much faster. On the other hand, phylogenetic tools are believed to be more accurate but do generally not scale well to data sets containing more than a hundred sequences. The underlying difference is that phylogenetic tools implement an explicit model of sequence evolution, usually taking the form of a continuous-time Markov process branching over a tree. While this means that computationally expensive operations like matrix inversions become necessary, bioinformatic tools are based most of the time on efficient pattern-matching algorithms on sequences that rarely exceed quadratic running time complexity. Several developments in recent years open new avenues and offer a way forward. Most importantly, these include phylogenetic tools that are based to a large extent on precomputed

results. In the context of the Rabifier, 'phylogenetic sequence placement' as implemented by PAGAN [6] is particularly interesting. Given a precomputed multiple sequence alignment and tree, new sequences are placed considering phylogenetic relatedness and added to the precomputed alignment and tree without need to recompute them. Therefore, this hybrid approach provides a potentially accurate way to classify a sequence as part of a subfamily represented by a clade in a phylogenetic tree without sacrificing speed. The implementation of this and other minor improvements is planned as part of a Rabifier version 2.

Second, a mid-term goal is to apply the model for evolution of function presented here in a predictive manner. While we tested several consequences of 'effector switching' on existing data in Chapter 4 and failed to disproof it, the ultimate demonstration of usefulness and a great boost in confidence about its adequacy is to be gained from predicting yet unknown aspects of Rab function and confirm those experimentally. A promising pair of Rabs is Rab6 and 41: Rab41 is the evolutionarily most recent addition to the Rab family in humans [7] and very little is known about this protein. Since its identification [8], we are only aware of one dedicated functional study on Rab41 [9]. Starting from known Rab6 effectors and regulators, predicting residues that show signs of positive selection in Rab41 may generate hypotheses about the loss and gain of interactions. However, new types of experiments potentially including explicit structural modelling may become necessary to grant enough confidence in the predictions and justify the investment into experimental validation. A different way forward is to focus on protein families other than Rabs. As already discussed above, we believe that effector switching is be a general model applicable to other master regulators. The ability to understand and predict the evolution of function in other proteins would underline the value of our contribution.

Third, the most exciting goal is to use Rabs as markers to bridge

levels and inquire into the evolution of the Endomembrane System and cellular organisation *per se*. This is surely most ambitious and therefore a long-term goal, potentially requiring to integrate the evolutionary analysis of other important gene families involved in vesicular trafficking such as SNAREs. However, a small step in that direction has already been made in Chapter 2: analysing the expansion of the Rab family at the base of Metazoa and mapping the new Rabs into their functional categories suggested that regulated secretion greatly diversified and complexified in animal evolution. Another interesting suggestion is the extrapolation of the mode of gene evolution to organelle evolution: the organelle paralogy model proposes that new organelles evolve by a process equivalent to duplication-divergence of genes [10]. The great challenge now is to go beyond these correlations, and design models and experiments that are able to contribute to a mechanistic understanding of cellular evolution. In other words, Rabs may be have the potential to be founding members of a new evolutionary cell biology.

References

[1] Ovid. *Metamorphoses*. Boston: Cornhill Publishing Co., Jan. 1922.

[2] Antony M Dean and Joseph W Thornton. "Mechanistic approaches to the study of evolution: the functional synthesis". In: *Nature Reviews Genetics* 8.9 (Sept. 2007), pp. 675–688.

[3] Michael J Harms and Joseph W Thornton. "Evolutionary biochemistry: revealing the historical and physical causes of protein properties". In: *Nature Reviews Genetics* 14.8 (Aug. 2013), pp. 559–571.

[4] Susumu Ohno. *Evolution by Gene Duplication*. Springer-Verlag, 1970.

[5] J W Thornton and Rob DeSalle. "Gene family evolution and homology: genomics meets phylogenetics". In: *Annual Review of Genomics and Human Genetics* 1 (2000), pp. 41–73.

[6] Ari Löytynoja, Albert J Vilella, and Nick Goldman. "Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm". In: *Bioinformatics* 28.13 (July 2012), pp. 1684–1691.

[7] Yoan Diekmann et al. "Thousands of Rab GTPases for the Cell Biologist". In: *PLoS Computational Biology* 7.10 (Oct. 2011), e1002217.

[8] Jin-Hu Guo et al. "Isolation, expression pattern of a novel human RAB gene RAB41 and characterization of its intronless homolog RAB41P". In: *DNA sequence* 14.6 (Dec. 2003), pp. 431–435.

[9] Shijie Liu, Lauren Hunt, and Brian Storrie. "Rab41 is a novel regulator of Golgi apparatus organization that is needed for ER-to-Golgi trafficking and cell growth". In: *PLoS ONE* 8.8 (2013), e71886.

[10] Joel B Dacks and Mark C Field. “Evolution of the eukaryotic membrane-trafficking system: origin, tempo and mode”. In: *Journal of Cell Science* 120.17 (Sept. 2007), pp. 2977–2985.

Summary

"Es ist schon alles gesagt, nur noch nicht von allen."
('Everything has been said, but not yet by everyone.')

—Karl Valentin

Summary

WHY does a protein function the way it does? The question provokes different answers. For example, one could invoke the biochemistry of the protein, the cellular network it is part of, the evolutionary forces that shaped the sequence or its importance for organismal survival. The type of reply will most likely depend on who is asked: a biochemist or cell biologist may be inclined to answer in terms of the first two possibilities, whereas an evolutionary biologist probably prefers the latter two. It is clear that all interpretations are correct, the difference merely reflects the focus of biological disciplines on distinct causes of a biological phenomenon. Each of the aspects are important and may be studied independently, yet, all are required for a complete understanding [1].

The appeal of the above reasoning has long been recognised, but actual examples of research integrating multiple disciplines and causes of a phenomenon are still underrepresented. This has at least two reasons: first, the undeniable success that biology had so far studying causes of biological phenomena in isolation, and more pragmatically, the broad expertise required to bridge the highly intricate and specialised disciplines. However, research programmes are now settling into mainstream that stress the mutual benefit and unique value of integrative approaches, such as evolutionary biochemistry [2] and evolutionary cell biology [3] which have been identified as part of a larger "functional synthesis" [4]. The times seem promising for such a synthesis: not only have innovative formal and experimental strategies contributed to form its methodological foundation, but the general availability and burgeoning of public molecular data greatly facilitate research bridging the disciplines.

Here, I follow the functional synthesis and ask a question that falls squarely between the disciplines: how does protein function evolve?

The question of how function evolves is a founding theme of an entire discipline, evolutionary physiology, one of the earliest formulations of an integrative research programme [5, 6]. More recently, the question is also gaining fundamental importance for functional genomics. Due to the ever-growing throughput of genome sequencing, the bottleneck shifted from data gathering to its functional annotation. The predominant strategy for functional annotation is transfer of existing annotations applying the "guilt by association" principle [7]. However, this is becoming insufficient mostly due to the fact that by definition 'transfer' cannot predict true functional novelty, and as a result the fraction of uncharacterised genes grows [8]. A deeper understanding of how function evolves has the potential to assist and guide functional genomic efforts.

In this thesis, I focus on the Rab family of small GTPases as a model system. Rabs are an essential family of master regulators of vesicular trafficking in eukaryotes, the largest family within the Ras superfamily of small GTPases. In human, there are more than 60 Rab genes that fall into roughly 45 subfamilies, each of which reside on characteristic organelles and their respective vesicles and regulate the distinct steps of vesicular trafficking, *i.e.* cargo selection, budding, scission, transport, tethering and fusion. Rabs function as molecular switches that cycle between a GTP-bound 'on', and GDP-bound 'off' state. In their GTP-bound state, they recruit a set of effectors that usually exert a function, for example the Retromer complex [9], members of all classes of molecular motors [10] or the Exocyst complex [11]. Rab sequences are usually well conserved but for the two so-called hypervariable termini that significantly diverged between different Rab subfamilies.

Rabs are in many ways well suited to ask questions about the evolution of function. They are an essential family and therefore found in all eukaryotes sequenced so far providing an abundance of sequence data to work

with. Due to their importance, they have been extensively characterised and a rich body of experimental studies and data is publicly available. In general, Rab function shows impressive stability throughout evolution, for example, the mouse Rab1 is still able to rescue a Rab1 knockout in yeast [12]. Even without ability to rescue, orthologous Rabs have usually been found to perform equivalent functions at corresponding compartments in all eukaryotes. In contrast to the functional stasis, the composition and size of the family greatly varies throughout the eukaryotic tree of life, implying the existence of numerous lineage-specific gene losses and duplications and therefore potentially interesting functional changes. For these reasons, the evolution of Rab function mostly becomes a question of how Rabs evolve after gene duplication, for which classical evolutionary models exist [13]. Lastly, Rabs function by transient protein-protein interactions with numerous effectors, which may be archetypical for master-regulators. To the best of my knowledge, this can be considered a different 'functional paradigm' than most of the proteins that have been studied so far with respect to their functional evolution. These include for example historically well-studied proteins like haemoglobins [14] and Cytochrome c [15], proteins that are characterised by exquisite affinity to a specific ligand like metabolic enzymes [16, 17] or hormone receptors [18, 19], but also transcription factors [20, 21] that possibly come closest to Rabs in the way they function.

I divide the question how Rab function evolves in two parts and present the results separately: after reviewing the relevant literature on Rab function in the introductory Chapter 1, I describe the patterns of Rab family evolution in Chapter 2. Based on these observations, I infer the evolutionary processes that may have generated the observed patterns in the second part consisting of Chapters 3 and 4.

Chapter 2 tackles the bioinformatic problem of large scale annotation

of the Rab family across all sequenced eukaryotic genomes. Annotating Rabs has been a laborious manual task because of the high similarity amongst all members of the Ras-superfamily of small GTPases and the lack of a well-annotated reference set of Rabs to transfer annotations. I develope, validate and benchmarke the 'Rabifier', an automated bioinformatic pipeline for the identification and classification of Rabs, which achieves up to 90% classification accuracy. I cataloge roughly 8.000 Rabs from 247 genomes covering the entire eukaryotic tree. The full Rab database and a web tool implementing the pipeline are publicly available at `www.RabDB.org`. For the first time, I describe and analyse the evolution of Rabs in a dataset covering the whole eukaryotic phylogeny. I find a highly dynamic family undergoing frequent taxon-specific expansions and losses. I date the origin of human subfamilies using phylogenetic profiling, which enlarges the Rab repertoire of the Last Eukaryotic Common Ancestor to additionally include Rab14, 32 and RabL4. Furthermore, a functional survey of the Choanoflagellate *Monosiga brevicollis* Rab family pinpoints the changes that accompanied the emergence of Metazoan multicellularity, mainly an important expansion and specialisation of the secretory pathway. Lastly, I establish tissue specificity in expression of mouse Rabs using public microarray data. I conclude by hypothesising that neofunctionalisation (rather than the alternative dosage- and subfunctionalisation models) best explains the emergence of new human Rab subfamilies. Neofunctionalisation predicts that one copy remains functionally unaltered, while the other copy accumulates mutations that are going to bring about a new beneficial function which is going to be driven to fixation by positive selection [22].

Chapter 3 addresses a technical problem resulting from verifying the major prediction of the neofunctionalisation model: the past action of positive selection on Rab, which is detected in a phylogenetic framework based on gene trees. However, Rabs are particularly challenging for gene

tree inference methods, as Rab sequences carry little phylogenetic signal due to their shortness and structure consisting only of highly conserved and highly divergent parts. I therefore begin by asserting if the gene tree has any effect on the inference of positive selection at all. The Branch-Site Test of Positive Selection [23, 24] is a standard approach to detect episodic positive selection in a priori specified branches. In this chapter, I ask if errors in the topology of the gene tree have any influence on its ability to infer positively selected sites. Using simulated sequences, I compare the results obtained for the true and erroneous topologies, and find a strong linear effect on the ability to predict sites if the tree changes how long sequences are inferred to have experienced selection. Moreover, I show by reanalysing a previously published data set that the choice of gene tree alters the results not only for simulated but also for actual sequences. This is the first time a clear effect of the gene tree topology on the inference of positive selection is demonstrated. I conclude that the gene tree is an important factor for the branch-site analysis of positive selection that has so far been overlooked. As a consequence, the following analyses are based on manually curated Rab phylogenies.

Chapter 4 addresses the neofunctionalisation hypothesis put forth at the end of Chapter 2 and finally answers how Rab function evolves. Neofunctionalisation is a classical model to explain the emergence of novel gene functions. Its original emphasis was to propose a mechanism that could account for a mutational path between two proteins that included mutations disrupting the original function. The resulting problem why those intermediates would not immediately be purged by purifying selection is elegantly solved by the redundancy introduced via gene duplication. However, while the model solves the evolutionary problem, the biochemical or cell biological dimension of the problem is neglected, and thecrucial functional challenge remains disguised as a tacit assumption: how can mutations lead to a new function? Here, I address this question for Rab

GTPases. First, I confirm based on public functional and sequence data that the Rab family frequently expands by neofunctionalisation. I identify the interactions with effectors as the central aspect of Rab function, and therefore hypothesise that Rabs gain new functions by changing their set of effectors. I verify the necessary condition that related Rab subfamilies have overlapping sets of effectors by surveying the known effector interactions. Finally, I focus on the vertebrate subfamily Rab25, which evolved from a duplicated Rab11. In accordance with my hypothesis, I discover signatures of selection in the binding interfaces of these two Rab subfamilies. This suggests a scenario where shortly after duplication Rab11-specific interactions are lost and new interfaces to Rab25-specific effectors are formed and fixed by positive selection. Furthermore, I find a trend at the organismal level to reduce the breadth of tissue expression of new Rab subfamilies established concurrently with the alterations of the set of effectors. These results open the functional "black box" of the original neofunctionalisation model and showcase how mutations can lead to new function. Rabs are a particularly relevant model system, as their evolution is tightly linked to that of the eukaryotic cell itself due to their central importance for the endomembrane system and intracellular organisation in general.

The last chapter concludes by summarising the results and discussing the implications of this case study for the evolution of function in general. In particular, at least for Metazoan Rabs I find an interesting contrast between functional stasis maintained by strong purifying selection on one hand, and frequent duplications and neofunctionalisation on the other. It is interesting to speculate that this observation is explained by the vast interaction network Rabs are part of that renders any alteration disastrous for the organism. However, that it is for the same reason that duplicates easily form new interactions and therefore show the great potential for

functional innovation that played its part in shaping the Metazoan cell as we know it.

Resumo

Porque é que uma proteína funciona de uma determinada forma? Esta pergunta suscita diferentes respostas. Por exemplo, poder-se-ia invocar a sua estrutura bioquímica, as vias celulares a que pertence, as forças evolutivas que moldaram a sequência ou a sua importância para a sobrevivência do organismo. A resposta irá provavelmente depender de quem for inquirido: um bioquímico ou um biólogo celular estariam mais inclinados a considerar as duas primeiras possibilidades, enquanto um biólogo evolucionista preferiria as duas últimas. Ambas as interpretações estariam correctas, baseando-se, no entanto, no estudo de diferentes causas para o mesmo fenómeno biológico. Cada um destes aspectos é importante e pode ser estudado independentemente. No entanto, todos eles são necessários para que haja um completo entendimento do problema [1].

A importância do argumento supramencionado foi há muito reconhecida. Contudo, estudos integrando múltiplas facetas e disciplinas estão ainda sub-representados. Isto deve-se a pelo menos dois motivos: primeiro, o sucesso inegável que a biologia tem tido a estudar as causas dos fenómenos biológicos isoladamente; segundo, e mais pragmático, é o vasto conhecimento necessário para conectar disciplinas altamente complexas e especializadas. Não obstante, o crescente interesse em programas de investigação considerando abordagens integrativas realça o benefício mútuo e valor único destas, como por exemplo, bioquímica evolutiva [2] e biologia celular evolutiva [3], que foram apontadas como parte de uma grande "síntese funciona"' [4]. O momento parece promissor para tal síntese: não só foram criadas estratégias inovadoras, teóricas e experimentais, que têm contribuído para a sua base metodológica, mas também a maior disponi-

bilidade e crescimento de dados moleculares publicamente acessíveis têm facilitado a conexão destas disciplinas.

Aqui, no seguimento da síntese funcional, tenciono responder a uma pergunta que se centraliza entre as disciplinas: como é que a função das proteínas evolve? Esta questão é um tema fundador de toda uma disciplina, Fisiologia Evolutiva, pioneira na investigação integrativa e interdisciplinar [5, 6]. Recentemente, esta questão tornou-se fundamental também para a disciplina de Genómica Funcional. Devido aos avanços tecnológicos na sequenciação genómica, o passo limitante passou da obtenção para a anotação funcional de dados. A estratégia predominante para a anotação funcional é a transferência da anotação já existente, aplicando o princípio de "culpa por associação" [7]. Mas este tem-se revelado insuficiente principalmente porque, por definição, uma "transferência" não pode predizer novas funções génicas, e como consequência, a fracção de genes não caracterizados aumenta [8]. A compreensão aprofundada da evolução da função proteica pode fornecer importantes pistas para a problemática da Genómica Funcional.

Na presente tese, foquei-me nas proteínas Rab, uma família proteica de pequenas GTPases, como um sistema-modelo. As proteínas Rab são uma das principais famílias reguladoras do tráfego vesicular em eucariotas, assim como a maior família dentro da superfamília de pequenas GTPases Ras. Existem, em humanos, mais de 60 genes Rab, categorizados em aproximadamente 45 subfamílias, cada uma localizada num determinado organelo e respectivas vesículas. Desta forma, as proteínas Rab são responsáveis pela regulação das diferentes etapas do tráfego vesicular, *i.e.*, selecção da carga, budding, cisão, transporte, ancoramento e fusão vesicular. As proteínas Rab funcionam como interruptores moleculares; ligam-se a GDP ou GTP, adquirindo uma conformação inactiva ou activa,

respectivamente. Quando se encontram ligadas a GTP, *i.e.* sob a forma activa, recrutam conjuntos de moléculas efectoras, tais como o complexo *Retromer* [9], vários membros de todas as classes de proteínas motoras [10] ou o complexo Exocisto [11]. As sequências das proteínas Rab são geralmente bem conservadas, à excepção de dois terminais hipervariáveis que divergem significativamente entre as diferentes subfamílias de Rabs.

As proteínas Rab são, por diversos motivos, uma ferramenta adequada para investigar a evolução da função proteica. Estas são uma família essencial ao funcionamento celular e, desta forma, estão presentes em todos os organismos eucariotas até agora sequenciados, fornecendo assim numerosas sequências para análise. Devido à sua importância, as proteínas Rab encontram-se extensamente caracterizadas, existindo um grande volume de dados e estudos experimentais acessíveis publicamente. Em geral, estas proteínas apresentam uma estabilidade evolutiva extraordinária. Por exemplo, a Rab1 de ratinho é capaz de resgatar o fenótipo provocado pelo knockout de Rab1 em levedura [12]. Mesmo não tendo prova experimental desta capacidade de resgate para todas as Rabs, os ortólogos das Rabs, na maioria das vezes, exercem funções equivalentes nos compartimentos correspondentes em todos os organismos eucariotas. Opostamente a esta imutabilidade funcional, a composição e tamanho da família varia grandemente ao longo da árvore dos eucariotas, sugerindo a existência de numerosas perdas e duplicações de genes e por isso potenciais alterações interessantes a nível funcional. Por estas razões, a evolução da função nas proteínas Rab torna-se maioritariamente uma questão de como estas evolvem após a sua duplicação, para a qual existem modelos evolutivos clássicos [13]. Por último, as proteínas Rab estabelecem interacções transientes proteína-proteína com várias proteínas efectoras, que podem ser o arquétipo do regulador mestre. Tanto quanto sei, este facto representa um paradigma funcional diferente, quando comparado com a evolução funcional da maior parte das proteínas já estudadas. Estas incluem, por ex-

emplo: as famílias das hemoglobinas [14] e Cytochrome C [15], proteínas que são caracterizadas pela sua afinidade a um ligando específico como as enzimas metabólicas [16, 17]; mas também factores de transcrição [20, 21] que possivelmente se mais se assemelham a Rabs no modo como funcionam.

Este trabalho será composto por duas partes. Na primeira parte, farei uma revisão da literatura relevante acerca da função das proteínas Rab, no Capítulo 1, e descrevo os padrões evolutivos desta família, no Capítulo 2. A segunda parte consistirá na inferência dos processos evolutivos que poderão ter originado os padrões observados (Capítulos 3 e 4).

O capítulo 2 aborda o problema bioinformático de anotação em larga escala da família das proteínas Rab entre todos os genomas de organismos eucariotas sequenciados. A anotação das proteínas Rab foi uma tarefa manual trabalhosa devido à elevada semelhança entre todos os membros da superfamília Ras de pequenas GTPases assim como devido à ausência de um grupo-referência das proteínas Rab bem anotado para proceder com a transferência de anotações. Eu desenvolvi, validei e fiz o *benchmark* da *pipeline* bioinformática Rabifier, automatizada para a identificação e classificação de proteínas Rab, que atinge até 90% de precisão. Adicionalmente, cataloguei cerca de 8.000 Rabs de 247 genomas, abrangendo todas as famílias de organismos eucariotas. A base de dados das proteínas Rab e a *web tool* implementada com a *pipeline* estão publicamente acessíveis em `www.RabDB.org`. Pela primeira vez, descrevi e analisei a evolução das proteínas Rab utilizando um conjunto de dados que engloba toda a filogenia eucariota. Assim, deparei-me com uma família proteica extremamente dinâmica, que sofreu frequentes expansões e perdas em táxon específicos. Estabeleci a datação da origem das subfamílias das proteínas Rab em humano baseando-me em perfis filogenéticos que permitiram aumentar o repertório do último ancestral comum eucari-

ota (*Last Eukaryotic Common Ancestor*) das proteínas Rab, adicionando Rab14, Rab32 e RabL4. Além disso, uma análise funcional da família proteica Rab presente no coanoflagelado *Monosiga brevicollis* exemplifica precisamente as alterações que acompanharam o surgimento da multicelularidade nos Metazoa, evidenciando uma forte expansão e especialização da via secretora. Por último, estabeleci a especificidade da expressão das proteínas Rab em diferentes tecidos de ratinho, utilizando dados públicos de microarrays. Eu concluo com a hipótese de que a neofuncionalização (em vez dos modelos alternativos de dosagem e subfuncionalização) explica melhor o surgimento de novas subfamílias de Rabs em humanos. O modelo de neofuncionalização prediz que, após um evento de duplicação génica, uma cópia permanece funcionalmente inalterada, enquanto que a outra acumula mutações que irão resultar numa nova função. Caso esta seja benéfica será fixada por selecção positiva [22].

No capítulo 3 abordei um problema técnico que surgiu da verificação das previsões do modelo de neofuncionalização: a acção da anterior selecção positiva nas proteínas Rab que é detectada num enquadramento filogenético baseado em árvores de genes. As proteínas Rab são um desafio para os métodos de inferência de árvores de genes dado que as suas sequências apresentam um fraco sinal filogenético, uma vez que são proteínas pequenas e possuem uma estrutura que consiste em regiões muito conservadas e regiões muito divergentes. Desta forma, comecei por verificar se a árvore de genes teria alguma influência na inferência de selecção positiva, utilizando o teste *Branch-Site Test of Positive Selection* [23, 24]. Este teste é a abordagem padrão para a detecção de selecção positiva episódica em ramos especificados *a priori*. Posteriormente, pretendi esclarecer se erros na topologia da árvore de genes teriam alguma influência na sua capacidade de inferir locais positivamente seleccionados. Utilizando simulação de sequências, comparei os resultados obtidos em topologias verdadeiras e erróneas e encontrei uma forte correlação linear

na capacidade de predizer locais consoante a árvore é modificada para inferir como sequências longas foram sujeitas a pressão selectiva. Além disso, através da re-análise de dados previamente publicados, mostrei que a escolha da árvore de genes altera os resultados não só para sequências simuladas mas também para sequências verdadeiras. Esta é a primeira vez que um efeito evidente da topologia da árvore de genes na inferência de selecção positiva é demonstrado. Com base nos resultados obtidos, concluí que a árvore de genes é um factor determinante para a previsão de locais sujeitos a selecção positiva, aspecto que tem sido negligenciado. Consequentemente, as análises seguintes basearam-se em filogenias das proteínas Rab manualmente curadas.

O Capítulo 4 aborda a hipótese da neofuncionalização mencionada no final do capítulo 2 e finalmente responde a como as proteínas Rab evolvem. Neofuncionalização é um modelo clássico que explica o surgimento de novas funções génicas. O seu objectivo original foi propor um mecanismo que considerava uma via mutacional entre duas proteínas incluindo proteínas em estados intermediários com mutações que alteravam a sua função original. O problema resultante destas proteínas intermediárias não serem imediatamente removidas através de uma selecção purificadora é elegantemente resolvido pela redundância introduzida aquando da duplicação de genes. No entanto, enquanto que o modelo de neofuncionalização resolve o problema evolucionista, a problemática a nível bioquímico ou celular continua por explicar, e o desafio funcional crucial permanece, mascarado por uma suposição tácita: como podem as mutações levar a uma nova função? Aqui, eu abordo esta questão para as proteínas Rab. Primeiro, confirmei, baseado em dados públicos, de sequências e funcionais, que a família proteica Rab expande frequentemente por neofuncionalização. Para além disso, identifiquei as suas interacções com moléculas efectoras como o aspecto central da função das Rabs. Assim, proponho a hipótese de que as proteínas Rab adquirem novas funções através da alteração do con-

junto de moléculas efectoras com que interagem. Através da comparação da interacção das proteínas Rab com efectores já estudados, verifiquei a condição fundamental de que subfamílias de proteínas Rab relacionadas entre si possuem conjuntos de moléculas efectoras comuns. Finalmente, foquei-me na subfamília Rab25, em vertebrados, que evoluiu de uma duplicação do gene codificante para a proteína Rab11. Em concordância com a hipótese proposta, identifiquei traços de selecção nas interfaces de ligação destas duas subfamílias de Rabs. Isto sugere um cenário em que, pouco tempo após a duplicação, as interacções específicas à Rab11 foram perdidas e novas interacções com efectores específicos da Rab25 foram formadas e fixadas por selecção positiva. Adicionalmente, encontrei uma tendência ao nível do organismo para reduzir a amplitude da expressão tecidual de novas subfamílias de proteínas Rab estabelecidas simultaneamente com as alterações do conjunto de efectores. Estes resultados põem a descoberto o modelo de neofuncionalização original e demonstram como as mutações podem levar a uma nova função. As proteínas Rab são um sistema modelo particularmente relevante, uma vez que a sua evolução está fortemente interligada com a da própria célula eucariota devido ao seu papel fulcral no sistema endomembranar e organização intracelular.

No último capítulo sumarizei os resultados obtidos e discuti as implicações do estudo das proteínas Rab para a evolução generalizada da função proteica. Em particular, pelo menos para as proteínas Rab em Metazoa, há um contraste intrigante entre a manutenção de uma homeostase funcional por evolução fortemente purificadora, e frequentes duplicações seguidas de processos de neofuncionalização. É interessante especular que esta observação poderá ser explicada pela vasta rede de interacções em que as proteínas Rab participam. Sendo a sua função fundamental para a célula, a sua alteração revelar-se ia desastrosa para o organismo. No entanto, é pela mesma razão que os eventos de duplicação

favorecem a criação de novas interacções proteína-proteína, mostrando assim um grande potencial para a inovação funcional que desempenhou o seu papel na modelação da célula metazoária como a conhecemos.

References

[1] Ernst Mayr. “Cause and Effect in Biology”. In: *Science* 134.3489 (1961), pp. 1501–1506.

[2] Michael J Harms and Joseph W Thornton. “Evolutionary biochemistry: revealing the historical and physical causes of protein properties”. In: *Nature Reviews Genetics* 14.8 (Aug. 2013), pp. 559–571.

[3] Frances M Brodsky, Mukund Thattai, and Satyajit Mayor. “Evolutionary cell biology: Lessons from diversity”. In: *Nature Cell Biology* 14.7 (July 2012), p. 651.

[4] Antony M Dean and Joseph W Thornton. “Mechanistic approaches to the study of evolution: the functional synthesis”. In: *Nature Reviews Genetics* 8.9 (Sept. 2007), pp. 675–688.

[5] Theodore Garland and P A Carter. “Evolutionary Physiology”. In: *Annual review of physiology* 56 (1994), pp. 579–621.

[6] Martin E Feder, Albert F Bennett, and Raymond B Huey. “Evolutionary Physiology”. In: *Annual Review of Ecology and Systematics* 31 (Apr. 2000), pp. 315–341.

[7] L Aravind. “Guilt by association: contextual information in genome analysis”. In: *Genome Research* 10.8 (July 2000), pp. 1074–1077.

[8] Andrew D Hanson et al. “’Unknown’ proteins and ’orphan’ enzymes: the missing half of the engineering parts list—and how to find it”. In: *The Biochemical Journal* 425.1 (2010), pp. 1–11.

[9] Juan S Bonifacino and James H Hurley. “Retromer”. In: *Current Opinion in Cell Biology* 20.4 (Aug. 2008), pp. 427–436.

[10] Anna Akhmanova and John A Hammer. “Linking molecular motors to membrane cargo”. In: *Current Opinion in Cell Biology* 22.4 (Aug. 2010), pp. 479–487.

[11] Margaret R Heider and Mary Munson. “Exorcising the exocyst complex”. In: *Traffic* 13.7 (July 2012), pp. 898–907.

[12] H Haubruck et al. “The ras-related mouse ypt1 protein can functionally replace the YPT1 gene product in yeast”. In: *The EMBO Journal* 8.5 (May 1989), pp. 1427–1432.

[13] Hideki Innan and Fyodor A Kondrashov. “The evolution of gene duplications: classifying and distinguishing between models”. In: *Nature Reviews Genetics* 11.2 (Feb. 2010), pp. 97–108.

[14] M F Perutz. “Species adaptation in a protein molecule”. In: *Molecular Biology and Evolution* 1.1 (Dec. 1983), pp. 1–28.

[15] Richard E Dickerson. “The structure of cytochrome c and the rates of molecular evolution”. In: *Journal of Molecular Evolution* 1 (1971), pp. 26–45.

[16] D A Powers et al. “Genetic mechanisms for adapting to a changing environment”. In: *Annual Review of Genetics* 25 (1991), pp. 629–659.

[17] Karin Voordeckers et al. “Reconstruction of Ancestral Metabolic Enzymes Reveals Molecular Mechanisms Underlying Evolutionary Innovation through Gene Duplication”. In: *PLoS Biology* 10.12 (Dec. 2012), e1001446.

[18] Jamie T Bridgham et al. “Evolution of a new function by degenerative mutation in cephalochordate steroid receptors”. In: *PLoS Genetics* 4.9 (2008), e1000191.

[19] Sean Michael Carroll, Eric A Ortlund, and Joseph W Thornton. “Mechanisms for the evolution of a derived function in the ancestral glucocorticoid receptor”. In: *PLoS Genetics* 7.6 (June 2011), e1002117.

[20] Cheryl C Hsia and William McGinnis. “Evolution of transcription factor function”. In: *Current Opinion in Genetics & Development* 13.2 (Apr. 2003), pp. 199–206.

[21] Günter P Wagner and Vincent J Lynch. “Molecular evolution of evolutionary novelties: the vagina and uterus of therian mammals”. In: *Journal of Experimental Zoology Part B, Molecular and Developmental Evolution* 304.6 (Nov. 2005), pp. 580–592.

[22] Susumu Ohno. *Evolution by Gene Duplication*. Springer-Verlag, 1970.

[23] Ziheng Yang and Rasmus Nielsen. “Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages”. In: *Molecular Biology and Evolution* 19.6 (June 2002), pp. 908–917.

[24] Jianzhi Zhang, Rasmus Nielsen, and Ziheng Yang. “Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level”. In: *Molecular Biology and Evolution* 22.12 (Dec. 2005), pp. 2472–2479.