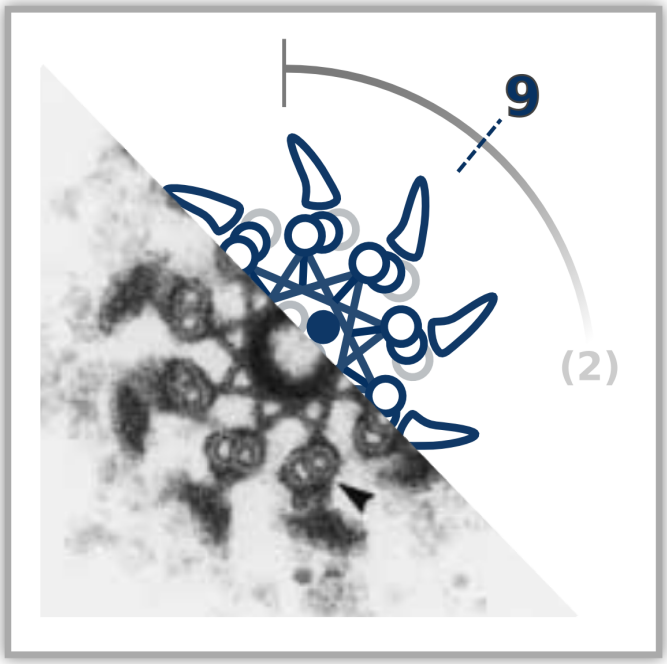
# New tools for old organelles

## Bioinformatics for comparative cell biology

Marc Gouw




Dissertation presented to obtain the Ph.D degree in Bioinformatics

Instituto de Tecnologia Química e Biológica António Xavier | Universidade Nova de Lisboa

Oeiras,
July, 2015

INSTITUTO
DE TECNOLOGIA
QUÍMICA E BIOLÓGICA
ANTÓNIO XAVIER /UNL

Knowledge Creation

# New tools for old organelles

## Bioinformatics for comparative cell biology

Marc Gouw

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

Instituto de Tecnologia Química e Biológica António Xavier | Universidade Nova de Lisboa

Research work coordinated by:

INSTITUTO
GULBENKIAN
DE CIÊNCIA

Oeiras
July, 2015

INSTITUTO
DE TECNOLOGIA
QUÍMICA E BIOLÓGICA
ANTÓNIO XAVIER /UNL

Knowledge Creation

## Acknowledgments

This thesis and the years of work and inspiration put into it would not have been possible without the help of many people.

First of all, I just generally want to thank everyone for being so great. Especially those who in any way contributed to making the past few years in Portugal and at the IGC such a wonderful experience: without all of you, my life and this thesis would have turned out very different indeed.

A special thanks to all of the past and present members of the Computational Genomics Laboratory, who have always been kind enough to be unforgivingly critical of my work, and equally enthusiastic to go enjoy a beer on a Friday afternoon. And especially Beatriz Ferreira Gomes and Yoan Diekmann, who were essential to my PhD project: You have been truly inspiring collogues and close friends. A lot of this work would not have been possible without you. The epic ***mtoc-explorer.org***, and indeed this thesis would also not have been possible without Renato Alves. I blame you, your programming genius and incredible patience for teaching me so much.

I also really want to thank my ever so important supervisors José Pereira-Leal and Mónica Bettencourt-Dias. Your guidance, supervision and support over the past few years has been invaluable. Also Zita Carvalho-Santos, who's initial work laid the foundation of what turned into the major part of my PhD.

The IGC is a wonderful place to study. I wanted to thank Prof. Howard and Prof. Coutinho for giving me the opportunity to be part of the IGC. I wanted to thank Thiago Carvalho, and Élio Sucena for providing a lot of practical and scientific support. Jorge Carneiro also provided incredibly useful feedback and advice throughout various parts of this project. Also my thesis committee Patrícia Beldade and Nuno Moreno, you provided excellent advice (especially towards the end of my PhD), for which I will always be grateful.

There are many at the IGC who I value incredibly as colleagues and friends, to whom I cannot express enough gratitude. Barbara Vreede, your never-ending excitement with science and life in general is truly inspiring (and highly contagious). I can't thank you enough for making the IGC a special place (and for being my gateway to coming to Portugal in the first place). And Dani Bodor, your vast knowledge, curiosity and excitement at the world of cell biology and beyond are incredible, just as your ability to catch a hammer.

There are also a few people I would like to thank for helping me get this thesis finished. Jarek Surkont provided great feedback on chapter 3. Renato Alves & Ricardo Leite for their comments and suggestions on chapter 2.

Lastly I wanted to thank those closest to me for always being there to support me, my work, thesis, or any other important (or meaningless) part in my life. Mom & Dad, thanks for you neverending support, for never doubting

me, and always giving me an extra push to keep on going when I need it. Richard and Bernard, you are incredible brothers, and although you are far away, I know you guys always have my back. And the love of my life Iria, for her endless support while I was writing this thesis, but more generally for being such a wonderful person and part of my life.

*Thank you all !!!*

## Financial Support

Esta dissertação teve o apoio financeiro da FCT e do FSE no âmbito do Quadro Comunitário de Apoio, bolsa de doutoramento #SFRH/BD/51628/2011 e da Fundação Calouste Gulbenkian.

This dissertation had the financial support from FCT and FSE through the Quadro Comunitrio de Apoio, doctoral fellowship #SFRH/BD/51628/2011 and Fundaçõo Calouste Gulbenkian.

## Software

This thesis was written entirely using the free and open source software programs LaTeX, Inkscape & Python.

## Cover image

Cover image is an adaptation of figure 2.1, which includes a modified version of a *Chlamydomonas reinhardtii* Transition Zone from Geimer and Melkonian (2004).

## Summary

For hundreds of years biologists have studied the naturally occurring diversity in plant and animal species. The invention of the electron microscope in the first half of the 1900's reveled that cells also can be incredible complex (and often stunningly beautiful). However, despite the fact that the field of cell biology has existed for over 100 years we still lack a formal understanding of how cells evolve: It is unclear what the extents are in cell and organelle morphology, if and how diversity might be constrained, and how organelles change morphologically over time.

The emergence of the eukaryotic cell over 1 billion years ago marks one of evolutions major transitions. In this branch of life the cellular architecture evolved from a relatively simple plan to a highly complex and compartmentalized system of organelles. One of the most powerful ways to study evolution is to study diversity across a broad range of different species: The "comparative" approach to biology. In the context of eukaryotic evolution we call this "comparative cell biology", which we explore in this thesis.

In this thesis we study two model systems for "comparative cell biology": Microtubule Organizing Centers (MTOCs) in chapters 2 and 3 and RabGTPases in chapter 4. Each of these chapters explores a different angle of cellular evolution, and each chapter proposes new bioinformatics tools to enable a "comparative cell biology" approach.

The first chapter addresses evolution of MTOCs from a purely morphological perspective. In order to achieve this we created *mtoc-explorer.org*, a community driven web-resource in which we collected ultrastructural data on MTOCs from over 100 species. Using this data we were able to determine some of the fundamental principles of the evolution of shape in organelles. We show that although diversity is a prominent theme in MTOC evolution, the total set of possible morphologies is constrained by functional requirements. In doing so we uncover a "spandrel" in cell biology: The requirement for microtubule based motility constraints the overall architecture of a cell's mitotic apparatus. Lastly we develop a model to measure ancestrality of organelles, and show convergent evolution of complex organelles in cells.

One of the major goals in biology is determine the link between a species'

genome and its morphological and functional properties. In chapter 3 we address this issue in a 32 species analysis using the eukaryotic cilium as a model organelle. Using a bioinformatics technique called "phylogenetic profiling" we ask how well we can use the presence and absence of a gene across multiple species to predict if a gene is functionally involved in the biogenesis or maintenance of the cilium. We found that the major improvements in "comparative cell biology" predictions are obtained by maximizing the taxonomic distribution of the species analyzed (representing as many eukaryotic lineages as possible).

Lastly in chapter 4 we explore the comparative approach using only sequence data. Rabs are a family of GTPases that are master regulators of intracellular trafficking, and are present in all major eukaryotic species. Each different family of Rabs is known to participate in different cellular processes. Therefore being able to identify which family a Rab belongs to allows one to make functional predictions about which processes can occur in a cell. In order to make these predictions possible, a bioinformatics pipeline the *Rabifier* and accompanying database *RabDB.org* were developed.

This thesis marks the first application of "comparative cell biology" as a framework to study the evolution of the eukaryotic cell. From an evolutionary perspective, the most important finding of this work is that many of the principles we know from "organism" apply equally to the model systems studied in this thesis. Whether these principals hold for other organelles remains to be explored. Most importantly, in each of these chapters, this thesis provides bioinformatics tools for "comparative cell biology".

# Summário

Durante centenas de anos os biólogos têm estudado a diversidade natural que ocorre em todas as espécies vegetais e animais. A invenção do microscópio electrónico na primeira metade do século passado, ajudou a revelar que as células também podem ser incrivelmente complexas (e muitas vezes de uma beleza apaixonante). No entanto e apesar do facto de que o ramo da biologia celular existe há mais de 100 anos, ainda não temos um conhecimento formal de como as células evoluem: não é claro a vastidão das células e da morfologia dos organelos, e se e de que maneira, a diversidade pode ser limitante, e de que modo os organelos a mudam sua morfologia ao longo do tempo.

O aparecimento da célula eucariótica há mais de 1000 milhões de anos traduz-se numa das mais importantes transições evolutivas. Neste ramo a arquitetura celular evoluiu a partir de um esboço relativamente simples para um sistema altamente complexo e compartimentalizado de organelos.

Uma das formas mais poderosas para estudar a evolução é estudar a diversidade através de uma ampla gama de diferentes espécies: a "abordagem comparativa" para a biologia. No contexto da evolução eucariótica designámos como "Biologia celular comparativa", a abordagem utilizada nesta tese.

Dois sistemas modelo serão estudados usando "biologia celular comparativa": o Centro organizador de microtúbulos (MTOCs) nos captulos 2 e 3 e RabGTPases no capítulo 4. Em cada um destes capítulos exploramos um ângulo diferente da evolução celular, e em cada um deles propomos novas abordagens e ambientes de trabalho bioinformáticos no âmbito da "biologia celular comparativa" . O primeiro capítulo aborda a evolução de MTOCs a partir de uma perspectiva puramente morfológica. De modo a alcançar este objectivo, criamos a ***mtoc-explorer.org***, uma ferramenta web impulsionada pela comunidade, onde foram recolhidos os dados ultra-estruturais de MTOCs de mais de 100 espécies.

Usando estes dados, fomos capazes de determinar alguns dos princpios fundamentais da evolução da forma dos organelos. Mostramos que, embora a diversidade é um tema de proeminente na evolução MTOC, o conjunto total de possíveis morfologias é limitada por requisitos funcionais. Ao fazê-lo descobrimos um "spandrel" na biologia celular: A necessidade de microtúbulos

com base móbil restringe a arquitetura geral do aparelho mitótico da célula. Por fim desenvolvemos um modelo que permite aferir a ancestralidade de organelos, e demostrar a evolução convergente de organelos complexos nas células.

Um dos principais objetivos da biologia é determinar a ligação entre uma espécie e o seu genoma e retirar deste, propriedades morfológicas e funcionais. No capítulo 3 abordamos esta questão analisando 32 espécies e utilizando o cílio eucariótico como modelo de organelo. Usando uma técnica bioinformática chamada de "perfil filogenético", perguntamos o quão bem podemos utilizar a informação da presença ou ausência de genes em várias espécies, de modo a prever se um gene est funcionalmente envolvido na biogénese ou na manutenção do cílio. Descobrimos que a maximização da distribuição taxonómica das espécies analisadas (representando o maior número de linhagens eucarióticas possível) permite grandes melhorias nas previses derivadas da "biologia celular comparativa". Por último, no capítulo 4, exploramos uma abordagem comparativa utilizando apenas dados extraídos de sequências. As Rabs são uma família de GTPases que são as reguladores fundamentais do tráfico intracelular e estão presentes em todas as principais espécies eucarióticas. Sabe-se que cada família diferente de Rabs é capaz de participar em diferentes processos celulares, portanto, a capacidade de identificar a qual a família pertence uma sequência de Rab, permite por si só fazer previsões funcionais sobre os processos que podem ocorrer numa célula. De modo a tornar estas previses possíveis foi desenvolvido um algoritmo bioinformático, o "Rabifier" e respectiva base de dados ***rabdb.org***

Esta tese é a primeira aplicação e abordagem no contexto da "biologia celular comparativa" para estudar a evolução da célula eucariótica. De um ponto de vista evolutivo, a descoberta mais importante deste trabalho é que muitos dos princpios que conhecemos de "organismos" aplicam-se de igual forma aos sistemas modelo estudados nesta tese. Contudo resta explorar se podemos extrapolar esta afirmação para outros organelos. Importante referir que nesta tese, cada um destes capítulos, fornece um estrutura de ambiente de trabalho bioinformático para o estudo da designada "biologia celular comparativa".

# Contents

# Chapter 1

# General Introduction

## 1.1 Diversity, Cells and Bioinformatics

It would be fair to say that the diversity we see in the world around us has been the inspiration for many of us to study evolution; the silent but steady driving force behind this wonderful variation. And it turns out that one of the oldest and most powerful ways to study how evolution works is by studying biological diversity.

For most of its history, biology was the discipline of studying diversity and variation. Aristotle, the father of biology, was the first to devise an organized system of classification of animals (the *scala naturae*). Similarly Linneaus, who in the 18$^{\text{th}}$ century gave rise to the taxonomic system we still use today, classified organisms based on their morphology – forms, shapes and structures – that defined that species and set it apart from others. George Cuvier, during the same era, was the first great comparative morphologist, and invented comparative anatomy and paleontology. However it was Darwin, naturally, who succeeded in using diversity and variation to explain the origin of species. It became clear that studying biological diversity was studying evolution in its most basic form.

For hundreds of years comparative morphologists have been classifying and cataloguing animals and plants from across the globe. In recent years, the focus has shifted to molecular biology and understanding how individual genes and proteins interact and function. However, since the invention of the first

microscopes, an entire new world of diversity has come into view without ever having been properly studied: the cell.

### 1.1.1 Diversity in Cell Biology

Over the past few centuries researchers have discovered that biology is not simply limited to species and life forms visible to the naked eye: most life is unicellular. The "Eukaryotic" kingdom, one of the three (or possibly two[1]) major branches of life has existed for approximately 1.5 billion years (Yoon et al., 2004). In contrast to bacteria and archea, eukaryotes have a highly complex, organized and compartmentalized cellular structure (Diekmann et al., 2011). Many of the organelles considered hallmarks of eukaryotes (such as the nucleus, Golgi apparatus, peroxisomes and also cilia) are thought to date back to the Last Common Eukaryotic Ancestor (LECA). However, despite a common origin, eukaryotes show a tremendous amount of morphological diversity in cellular structure (Figure 1.1).

This diversity exists at multiple different levels of cellular organization. Although a typical eukaryotic cell is $10 - 100\mu m$ in diameter, the single celled ciliate *Stentor coeruleus* can measure up to to $2.8mm$ in length (Marshall et al., 2012; Morgan, 1901). Cell shape can also vary greatly: diatoms alone display an incredibly vast amount of (often stunningly beautiful) variation in shape (see the illustration titled "Diatomea" in (Haeckel, 1904) for examples). Cell morphology also can differ greatly between different cells of a single species, exemplified by the structurally intricate and complex shapes of neurons. Other than shape and size, there is also a large amount of variation in intracellular composition of cells. The first and most obvious diversity is in the presence and absence of certain organelles (for chloroplasts, which exist in plants, and in a derived state in diatoms). Organelles themselves also show a large amount of morphological diversity. One example is the Golgi apparatus, which can take on a variety of different shapes including stacked and single cistern, and may even be invisible[2](Mowbrey and Dacks, 2009). Another example is the microtubule organizing centers (MTOCs) of eukaryotes – cilia and centrosomes – which

[1]Recent work has provided evidence that favours the 2 domain tree of life, in which Eukaryotes belong to archea (Spang et al., 2015).

[2]At least, not visible using standard electron microscopy techniques.

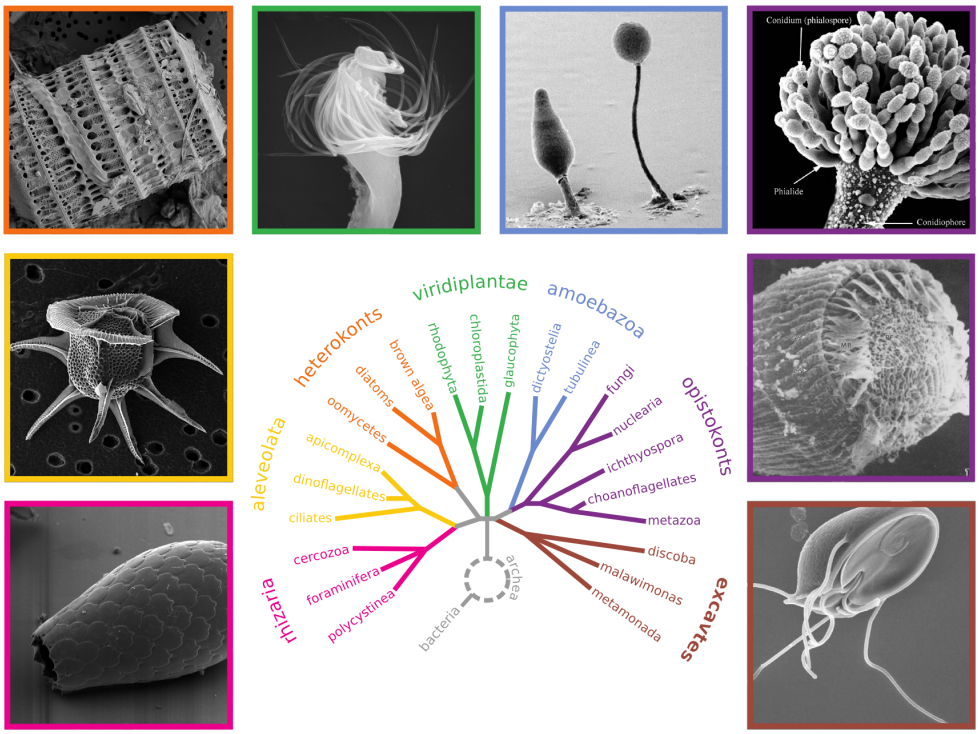



Figure 1.1: **Diversity in the eukaryotic kingdom (a few examples)** The Eukaryotic kingdom, despite consisting largely of unicellular life, is filled with morphological diversity. From bottom left to bottom right (clockwise): *Euglypah sp.*, *Ceratocorys horrida*, *Paralia sulcata*, *Equisetum hyemale*, *Dictyostelium discoideum*, *Aspergillus flavus*, *Stentor coeruleus*, and *Giardia lamblia*. Evolutionary tree and color scheme adapted from Adl et al. (2012) and Baldauf (2003a).

show structural diversity including different radial symmetries, stacked configurations, and the presence and absence of specific subcomponents. MTOCs are morphologically so diverse (and interesting) that we will be using these as the "model organelle" throughout most of this thesis, and will be further discussed in section 1.2.

Previous work shows that the existence of an organelle in a certain species can be related to a small number of genes. For example, the presence or absence of peroxisomes in a species can be predicted based on the presence or absence of only 4 genes in a species' genome (Schlüter et al., 2006). Similarly there is a core set of at least 3 genes required for centriole formation whose presence in a species' genome predicts the presence of the structure in a species (Carvalho-Santos et al., 2010). These studies both further suggest that the structural diversity observed in organelle structure and context may be linked to the presence and absence of other genes biologically related to the organelle.

Diversity exists in cell biology, and for a small handful of organelles biologists have identified genes who's presence directly correlates with the presence of that organelle. However these particular case-studies are limited in scope and require a large amount of manual curation of species' genotypes and phenotypes, nor do they address morphological diversity beyond the presence or absence of an organelle. They do not provide a framework to understand how morphological diversity evolves in cell biology. What they do show is that if we wish to understand the diversity of cell & organelle morphology, we first will need to obtain a detailed characterization of extents and types of diversity that exist. Subsequently we can look if and how a species' genome contributes to the evolutionary origins of diversity in cells. I propose that we turn to bioinformatics to solve both of these issues.

### 1.1.2 Bioinformatics for Comparative Cell Biology

Bioinformatics has been a part of biology since the early days of computing and the internet. One of the major reasons bioinformatics emerged as a discipline was to find ways to store, share and analyse the rapidly growing collection of biological data (Moore, 2007; Neerincx and Leunissen, 2005). In the past 3 decades bioinformatics has become central to many different fields of biology, and most molecular and cell biologists have become familiar with some of the

most basic bioinformatics techniques, for example BLAST (Altschul et al., 1990). At the moment, there are two major ways in which bioinformatics contributes to our understanding of evolution: genome and morphology databases. If we wish to study the evolution of cells, we can use techniques and concepts currently being used in both of these fields of evolutionary biology.

The vast majority of bioinformatics resources are dedicated to storing and analysing of genomic sequence data. One of the major goals of these projects is to understand how a species' genome is responsible for the shapes, function and behaviour of that species. An important step in this process is determining the function of all the genes in a species' genome. This is typically done by identifying genes in other species, sequence motifs or protein domains who's function is known. The set of techniques used to make these inferences between different genes and different species is called "comparative genomics".

Another group of scientists using bioinformatics to study evolution are those working in systematics and taxonomy, who study the shapes of limbs, skeletons, roots, trunks and organs between different species to determine their evolutionary relationships. Computers and the internet are helping researchers working in "comparative morphology" around the globe to work together and share their knowledge in ways never before possible.

Once again, we see that current tools and techniques in bioinformatics exist on exactly two different levels; that of large organisms (animals & plants) and that of molecules (DNA & protein sequences). However, we have no bioinformatics resources dedicated to studying diversity in cells. In this thesis we propose to combine tools and techniques from "comparative genomics" and "comparative morphology" to study the evolution of diversity at the level of the cell. In the same way that the microscope provided the hardware to see how wonderfully diverse cells are, we propose to use bioinformatics as the lens through which to see the evolutionary processes behind this diversity.

### 1.1.3 MTOCs, ontologies and databases

Studying the evolution of the eukaryotes is a daunting task: They have been evolving for over 1.5 billion years, and show a large amount of morphological diversity, much of which we probably have not yet discovered. Instead of trying to solve this entire complex puzzle, we will be using the aforementioned

"microtubule organizing centers" (MTOCs) as "model organelles" for this task. In order to start understanding how morphological diversity evolves in MTOCs we will first need to create a quantified database of the extends of this diversity. In order to achieve this we will borrow two "comparative morphology" techniques: 'ontologies' and "morphological databases". The remainder of the introduction is dedicated to these three topics.

## 1.2 The Microtubule Organizing Centers of Eukaryotes

The microtubule cytoskeleton was one of the major innovations during the early evolution of eukaryotes. Alongside the nucleus, a complex endomembrane system, the Golgi apparatus and mitochondria, the microtubule cytoskeleton is considered a hallmark of eukaryotes (Jékely, 2007, Chapter 1). The microtubule cytoskeleton is both unique to and ubiquitous in eukaryotes. The fact that no species have been identified with intermediate stages of the microtubule cytoskeleton suggests that this innovation gave an immense selective advantage to its ancestor that it gave rise to all currently existing eukaryotes (see Chapter 11 in Jékely (2007) or Mitchell (2007)). They are the main contributors to cellular architecture, and also play major roles in cell division, motility, signalling, trafficking and establishing cell polarity. The overall architecture and dynamics of the microtubule cytoskeleton are coordinated by microtubule organizing centers (MTOCs). There are two organelles that are considered the main MTOCs: the "cilium" (also known as the "flagellum") and the "centrosome".

The term "microtubule organizing center", as well as the terms "cilium", "flagellum" and "centrosome" have been interpreted and defined in many different ways, an issue addressed in more detail in section 1.3. However, for the sake of clarity, I will define these terms as they are used throughout the remainder of this section and thesis. Although many organelles have microtubule organizing capabilities (including the Golgi apparatus and condensed DNA) I will be using the term "MTOC" to refer to the two main MTOCs: the "cilium/flagellum" and "centrosome". Unfortunately, these terms have historically also been used in various ways. I will use the term "cilium" to refer to both "cilia" and "flagella" as there is no structural or functional distinction between the two, and they in

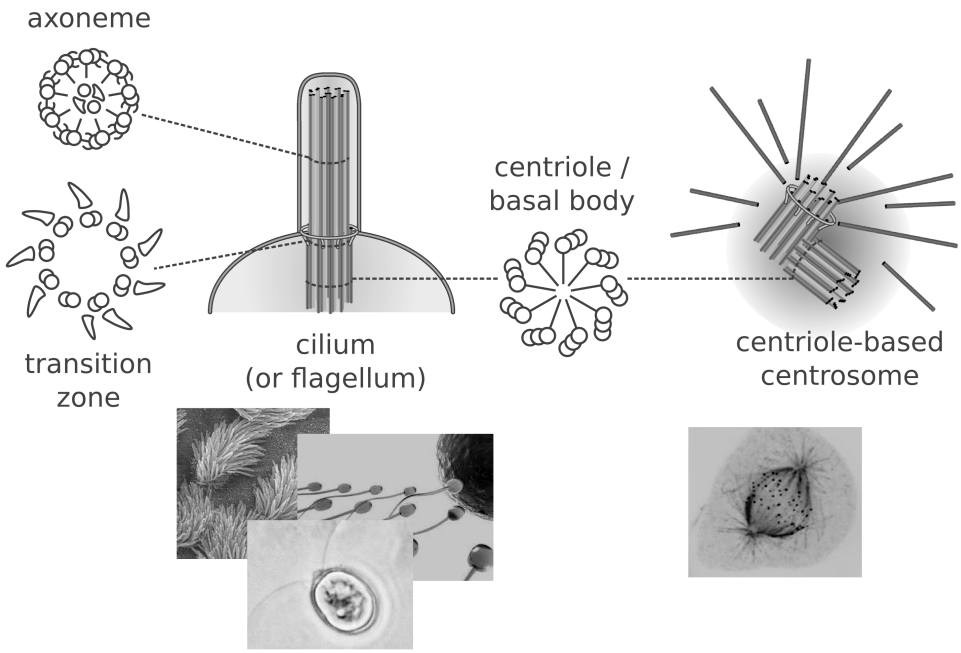



Figure 1.2: **The eukaryotic cilium and the centriole-based centrosome.** The "cilium" and the "centriole-based centrosome" (the centrosome in almost all animals) share a common component during the lifetime of a cell. This (typically) cylindrical organelle composed of 9-fold symmetrical microtubule triplets is referred to as the "basal body" when anchoring the cilium, the "centriole" when participating as part of the mitotic apparatus during cell division, and jointly as the "CBB". (Images are reproduced (with modifications) from (Carvalho-Santos et al., 2010)).

fact refer to a homologous organelle. The term "centrosome" has classically been used to describe the microtubule based organelles observed at the spindle poles during mitosis in animal cells. More recently the term has been adopted to include microtubule based organelles in fungi and amoebozoa (Azimzadeh, 2014) that localize to the spindle poles during mitosis. Although it is currently not known whether these are homologous structures in different eukaryotic lineages, I will be using the term "centrosome" to refer to any microtubule based structure which is functionally and behaviourally equivalent to the classical animal "centrosome". The base of the "cilium" – the "basal body" – is now known to be the same organelle as the "centriole" (Figure 1.2), and collectively these are referred to as the "CBB". The term "basal body" will be used exclusively when the "CBB" is anchoring a "cilium", and "centriole" to refer to the "CBB" when it is part of the mitotic apparatus. Finally, to distinguish the canonical animal "centrosome" from others, we will be using the term "centriole-based centrosome" if the mitotic apparatus contains "centrioles".

As the two major components of the microtubule cytoskeleton, the evolutionary histories of these organelles is both complicated and fascinating. These organelles play different roles in which their capacity to coordinate and modify the microtubule cytoskeleton plays a major role. Aside from functional differences, both of these structures are a source of structural diversity. Due to their presence in (almost) all eukaryotes, the ease of viewing them under a microscope, and the beautiful morphologies they display, it is no wonder that these organelles have a rich history as model organelles. This section starts with a historical introduction of MTOCs as model organelles to study diversity and evolution. Subsequently I will proceed to describe what is known about cilia and centrosomes in the present day including their functional roles, as well as the large amount of structural diversity we now know to exist.

### 1.2.1 MTOCs as classical model organelles for cell biology

Since the invention of the earliest microscopes, MTOCs have been a focal point for studying cells.[3] In 1676 Antonie van Leeuwenhoek described a "second set of animalcules" with "little feet, or little legs" (see Haimo and Rosenbaum (1981)). History would have to wait another 200 years until 1887 for the first insights on the structure of the cilium to emerge when Jensen proposed that the cilium

contained multiple "fibrils" (i.e. microtubules). Incidentally 1887 also marked the discovery of the animal "centrosome", when both Boveri (Boveri, 1887) and van Beneden (Beneden and Neyt, 1887) simultaneously discovered a dense and conserved structure at the heart of spindle poles in *Ascaris megalocephala* (Scheer, 2014). Once again, history would have us wait almost a full century until the commercialization of the Electron Microscope (EM) in the 1950's before any more significant insights were obtained in the underlying ultrastructure of these enigmatic organelles.

The mid and late 1900's would prove to be a very interesting time for cilium & centrosome biologists. Crude yet incredibly elegant scanning EM experiments (looking at shadows of microtubule bundles created with a low angle emission source) suggested that the cilium was composed of bundles of 9+2 "fibrils" (Manton and Clarke, 1952; Fawcett and Porter, 1954). Sorokin in (1962) described the difference between motile and non-motile cilia, and associate it to the presence of a central pair of microtubules in motile cilia. Towards the end of the same decade Dingemans (1969) and Wheatley (2005) followed by confirmation by Fulton and Dingle (1971) showed that the basal body and the centriole were one and the same organelle. In that same year Archer and Wheatley (1971) also noted that many plants do not have any distinguishable MTOC. During the same period it was revealed that other species have all together different microtubule based MTOC's: spindle pole bodies (SPB) in yeast (Robinow, 1966) and nucleus associated bodies (NAB) in amoebas (Roos, 1975).

### 1.2.2 Eukaryotic Cilia

The eukaryotic cilium is a membrane bound protrusion extending from the cell, involved in multiple cellular processes including motility, chemo-, photo- & mechanosensation, and signalling. Internally, the cilium is build on a scaffold of microtubule arrays that cover the entire length of the cilium, and are anchored to the cell via the plasma membrane. Typically the structure is a 9-fold

[3]For a more comprehensive history of cilia and centrosomes the reader is referred to the excellent review on cilia by Haimo and Rosenbaum (1981) and an insightful review on the discovery of the centrosome by Scheer (2014). Much of the text in this section is based on these works.

symmetrical cylinder of microtubule doublets, although the exact structure can vary greatly between different types of cilia and different species.

The cilium is currently thought to have evolved to combine motility, sensation and trafficking into a single organelle (Carvalho-Santos et al., 2011; Jékely and Arendt, 2006). These three major functions are also observed in extant species in all major branches of eukaryotes. In multicellular organisms the requirement for cell motility is greatly diminished, and only a handful of cell types have motile cilia. However the immotile cilium is present in almost all cells in animals and acts as the central hub of cell-to-cell signalling (Singla and Reiter, 2006; Goetz and Anderson, 2010). In animals, motile cilia can still be found in sperm cells and in multiciliated epithelial cells (for example the trachea and oviduct).

The cilium consists of three major components: the "axoneme", "transition zone" and "basal body" (Figure 1.2). It is typically described as a scaffold of microtubule doublets with 9-fold radial symmetry. The "basal body" is a short barrel shaped organelle which forms the base of the structure, and generally consists of microtubule triplets and may or may not contain a cartwheel. The upper part of the cilium, the "axoneme" is an extension of the two inner microtubules of the "basal body". Typically the axoneme is also 9-fold symmetrical. In between the "basal body" and the "axoneme" is the aptly named "transition zone", in which the array of microtubules transitions from its "basal body" structure to its "axoneme" structure, and the membrane anchoring machinery of the cilium can usually be found. The transition zone is the gateway that filters which components enter and leave the ciliary compartment. There are many structures which may or may not be present in these cells, which is often reflected by whether the cilium is motile or not. Motile cilia (as shown in Figure 1.2) typically have many additional components including a central pair of microtubules, 2 sets of dynein motor proteins, and radial spokes. Non-motile cilia typically have none of these.

Although the canonical 9-fold symmetrical cilium is a highly conserved structure, the cilium also shows a tremendous amount of structural diversity. These differences go well beyond motile vs. immotile cilia: Especially when we look beyond well characterized model systems we find a whole new world of structural diversity (Figure 1.3). Insects show an incredibly rich diversity in

fold symmetry ranging between 3 and 20 (and possibly even more), as well as non-symmetrical microtubule sheets and spirals (see Mencarelli et al. (2008) for some examples). Other structures are thought be taxon specific, such as the "plates" in *P. tetraurelia* (Dippell, 1968) or the "stellate fibers" in *C. reinhardtii* (Geimer and Melkonian, 2004). This diversity can be observed between cilia of different species, but also between different cells in the same species, and even in different life cycle stages of a single cell.

### 1.2.3 Eukaryotic Centrosomes

The "centrosome" is the generic name given to any organelle, or organelle-like structure, which is at the spindle poles during mitosis (Bornens, 2012). Unlike cilia, many cells exist which do not have a centrosome (at least, not readily visible by electron microscopy). In many cells (typically animal cells) the centrosome is formed by a pair of centrioles (Azimzadeh and Bornens, 2007). In other cells, microtubule based structures can clearly be seen organizing the spindles, however they are structurally (and sometimes molecularly) different from the animal centrosome. Thus it appears that centrosomes are not essential for cell division, although they exist in many different species, and when they exist can take on a number of different forms.

In animal cells the "centrosome" is typically formed by a pair of 9-fold symmetrical centrioles aligned orthogonally, and surrounded by a peri-centriolar matrix. They were long thought to be required for cell division, although we now know that this is not always true. Multiple experiments show that the centrosome is not required for mitosis in somatic cells in *D. melanogaster* (Debec et al., 2010). The fact that "centriole-based centrosomes" are not the major coordinators of mitosis is supported by the fact that many (in fact, almost all) eukaryotes do not have "centriole-based centrosomes" in any part of their cell cycle. Also, recently Azimzadeh et al. (2012) have identified an animal (the planarian flatworm *Schmidtea mediterranea*) that has evolutionarily lost its centrosomes completely. In many fungi, the spindle poles display stacks of disks which have (unimaginatively) been called "Spindle Pole Bodies" (SPB) (Kilmartin, 2014). Amoebozoa, the sister group of fungi & metazoa, also have a layered structure which appears to function as a mitotic MTOC called the "Nucleus Associated Body" (NAB) (Daunderer et al., 1999).

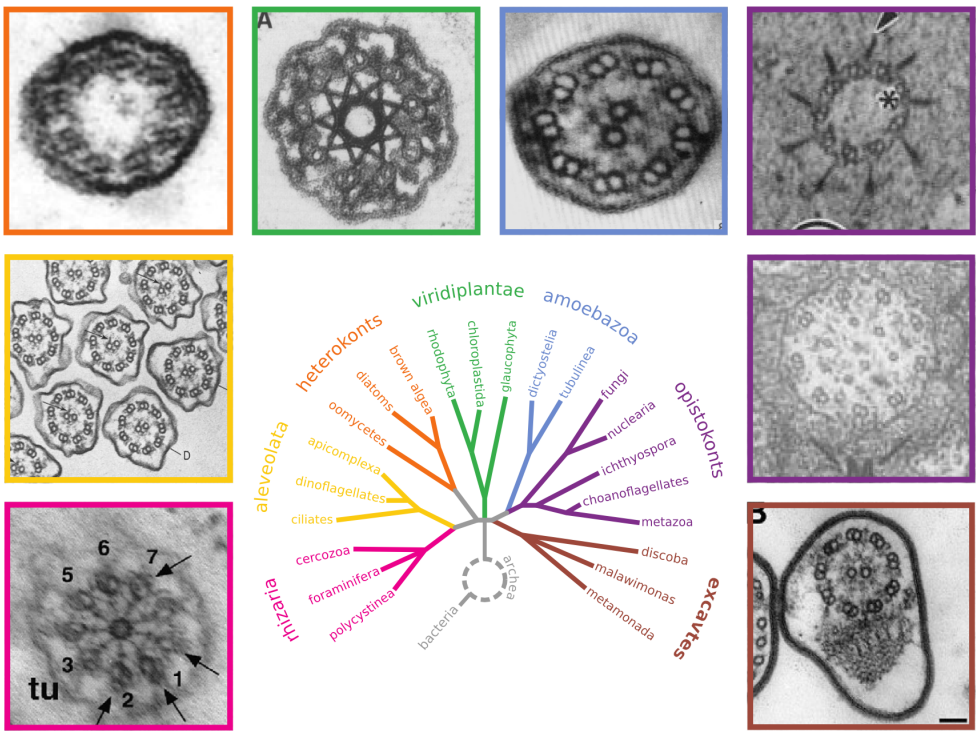



Figure 1.3: **The Eukaryotic Cilium (a few examples).** The eukaryotic cilium dates back to the Last Eukaryotic Common Ancestor, and shows both a remarkable amount of morphological conservation as well as diversity. These images show just a few examples of the conservation and diversity. References & ***mtoc-explorer.org*** image ID, bottom left to bottom right (clockwise): *Sainouron acronematica* (Cavalier-Smith et al., 2008) (670), *Tetrahymena pyriformis* (Allen, 1968) (401), *Lithodesmium undulatum* (Manton et al., 1970) (253), *Chlamydomonas reinhardtii* (Sanders, 1989) (62), *Physarum flavicomum* (Aldrich, 1968) (737), *Batrachochytrium dendrobatidis* (Longcore et al., 1999) (102), *Caenorhabditis elegans* (Perkins et al., 1986) (333), *Trypanosoma brucei* (Gadelha et al., 2006) (81).

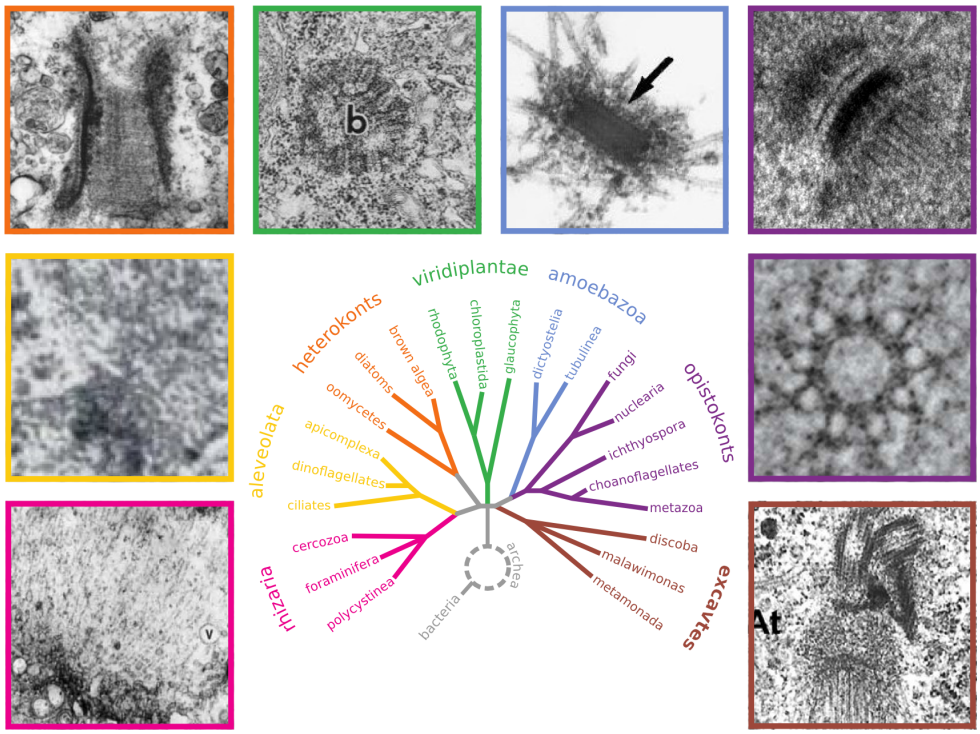



Figure 1.4: **The Eukaryotic Centrosome (a few examples)** Eukaryotes have a variety of different microtubule based organelles as part of the mitotic apparatus, and this figure shows a few examples. Apart from diversity in centrosome structure, many species exist for which no organelle is visible (by EM) at the spindle poles. References & ***mtoc-explorer.org*** image ID, bottom left to bottom right (clockwise): *Leptophrys vorax* (Ropstorf et al., 1994) (not on ***mtoc-explorer.org***), *Plasmodium fallax* (Aikawa, 1966) (288), *Lithodesmium undulatum* (Manton et al., 1969) (230), *Ceratopteris richardii* (Hoffman and Vaughn, 1995) (738), *Dictyostelium discoideum* (Ueda et al., 1999) (224), *Ashbya gossypi* (original microscopy contributed by Sue Jaspersen) (702), *Caenorhabditis elegans* (Pelletier et al., 2006) (326), *Trichomonas vaginalis* (Bricheux et al., 2007) (33).

### 1.2.4 MTOCs as model organelles for comparative cell biology

The eukaryotic MTOC is the ideal "model organelle" to study the evolution of morphological diversity in cell biology. The primary reason is that MTOCs have a rich and complex evolutionary history: They were present in the LECA, have been lost multiple times in different lineages, and have diversified structurally as well as functionally. Also, MTOCs have been extensively studied for over one hundred years, resulting in a large collection of published work across hundreds of different species. In the following two sections we explore two different techniques which we will use use to catalogue the morphological diversity in these enigmatic organelles.

## 1.3 Ontologies for Cell Biology

Cell biology, for the first few hundred years, existed almost entirely as a descriptive discipline. During this time thousands of articles were published containing ultrastructural descriptions of novel species and cells. These studies would usually consist of EM images of one or more organelles, accompanied by highly detailed text descriptions of the structures visible in each image.

This creates a major challenge for those wishing to obtain a detailed overview of these structures across the eukaryotic kingdom. The first major problem to overcome (as we have seen in section 1.2) is that cell biology (as many other fields of biology) is prone to discrepancies in nomenclature. The second major problem is that images and written descriptions on their own are not "data" in the sense that they are not systematically quantified. This makes meaningful cross-species and cross-organelle comparisons difficult, and completely rules out the possibility of computational analysis. These particular problems have been encountered in multiple other fields of biology. The solution most commonly used is to create a formal language for describing morphology: an "ontology".

In this section, I introduce some of the basic concepts of ontologies, and review a selection of ontologies currently used in biology, focussing on those useful for studying morphological variation.

### 1.3.1 Ontologies, a formal introduction

The concept of an "ontology" dates back to the (pre-Socratic) Greek philosopher Parmenides as the "study of the nature of being". Over the past few decades this concept has been adopted in a more practical sense in computer sciences to structure domains of knowledge. More recently, ontologies have been introduced to biology, to structure and order biological concepts, and the past decade and a half have seen an explosion in the number of "bio-ontologies" (Deans et al., 2012; Howe and Yon, 2008; Blake, 2004). Interestingly, these ontologies have been successfully implemented at the highest level of biology (the whole organism), and at the lowest level (genes and proteins), but only recently have a few attempts been made at the level of the cell.

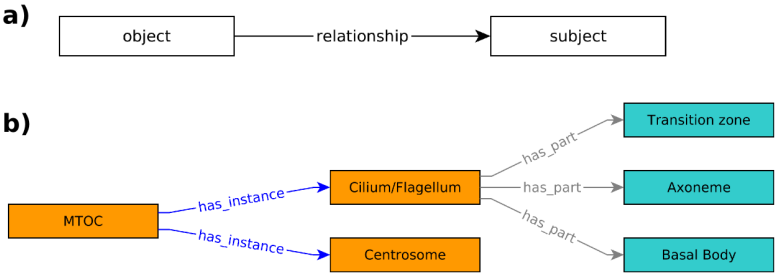



Figure 1.5: **Ontologies for biology.** Ontologies are a formal way to translate real-world entities into a conceptual graph. **a)** There are two parts to an ontology: The "terms" (in this example SUBJECT and OBJECT) which represent physical objects or concepts and the "relationships" which define how these "terms" are related to one another. **b)** An example of a basic ontology for "MTOC"s. This example shows two different types of terms, "classes" and "instances": Although not required to defined an ontology, these types help organize what each "term" represents. There are also two types of "relationships": The ***has_instance*** relationship establishes that CENTROSOMES and cilia are both different instances of class MTOC. The ***has_part*** shows that a CILIUM may have any one of the three components AXONEME, TRANSITION ZONE and BASAL BODY. This example is a subset of the ontology used later in chapter 2 (section 2.2.1).

Formally an ontology can be described as a "hierarchical controlled vocabulary". A "controlled vocabulary" simply means a set of strictly defined terms, thereby removing ambiguities that results from using natural languages (Vogt et al., 2009). These terms are organized hierarchically: terms "descend" from others in a logical fashion (Figure 1.5). The second part of an ontology is

the "relationships", which define how each term is related to its ancestors (or descendants). One of the powerful aspects of ontologies is "transitivity": logical rules can be used to traverse the hierarchy. If a NUCLEUS ***is_a*** ORGANELLE, and an ORGANELLE is ***part_of*** a CELL, it logically follows that a NUCLEUS is ***part_of*** a CELL. As a conceptual framework ontologies allow for the structured expression of almost any object of interest. Temporal aspects can be captured by relating "terms" via (for example) ***precedes***. Quantities can be described by using terms as values, for instance NUMBER OF MITOCHONDRIA ***has_value*** 9.

Ontologies are highly flexible in what they are able to describe: their strictly defined frameworks remove linguistic ambiguities, and they allow for a structured representation of quantified descriptions. Ontologies are highly suited tools for any field comparative biology, and their use in cell biology is long overdue.

### 1.3.2 Bio-ontolgies

There are many bio-ontolgies in existence today, and the number keeps on growing. A comprehensive list can be found at the OBO foundry, the official repository for biological ontologies, which as of June 2015 lists 10 officially recognized & 121 candidate bio-ontologies (Smith et al., 2007). There are two major types of ontologies dedicated to capturing morphological diversity: ontologies for taxonomy & systematics, and model organism ontologies for annotating gene and protein data.

#### Bio-ontologies for taxonomy & systematics

One of the main applications of bio-ontologies is in the field of systematics to aid the classification of species. The scope of these ontologies typically ranges from high resolution "natural diversity" of a closely related group of species, to large all-encompassing ontologies that allow cross-species comparisons.

There are numerous taxon specific ontologies for cataloguing natural diversity (for a review, see (Deans et al., 2012)). Some noteworthy examples include the Hymenoptera Anatomy Ontology (HAO) (Yoder et al., 2010), the Teleost Anatomy Ontology (TAO) (Dahdul et al., 2010) and the Xenopus Anatomy Ontology (XAO) (Segerdell et al., 2008). What these projects have in common

is that they allow for the complete (or partial) morphological annotation of organisms belonging to a closely related group of species: the ontologies allow for the description of the presence or absence of structures, sizes, colors, numbers, etc. However their scope is typically "small", each ontology consisting only of terms relevant to the particular set of species being studied.

### Bio-ontologies for model organisms

Model organisms are the work-horses of molecular and cell biology, and the results from high throughput phenotype screens are all available online. In an effort to study the organisms as a whole, and to integrate studies in different parts of each model organism, many model-organism specific ontologies have been developed. In most cases there are three general types of ontologies for each model organism: anatomical, developmental and (mutant) phenotype ontologies. These are almost exclusively usually used to describe the localization, timing and functional properties of genes (or gene products).

Anatomical ontologies exist for the major (metazoan) model organisms: *D. melanogaster*, *X. laevis*, *C. elegans*, *M. musculus*, *D. rerio* and also *H. sapiens* (see Dress et al. (2008) for an overview). The Zebrafish Anatomy Ontology (ZAO) (Sprague et al., 2006), includes terms for the complete anatomy of all major Zebrafish organs across 44 development stages, and allows mutation phenotype annotation via GO. As part of the Zebrafish Information Network (ZFIN) (Sprague et al., 2006) these ontologies are integrated with many other online resources including genome browsers, orthology predictions, antibodies & experimental protocols (Bradford et al., 2011). From a genetic perspective *D. melanogater* is one of the best studied model organisms, and Flybase (St Pierre et al., 2014) has created the Drosophila Anatomy Ontology (DAO) (Costa et al., 2013) as well as the Drosophila Phenotype Ontology (DPO) (Osumi-Sutherland et al., 2013). The *C. elegans* community has the *C. elegans* Cell and Anatomy Ontology (CECAO) (Lee and Sternberg, 2003) which includes anatomy, development and cell type annotations. In mice, the e-Mouse Atlas Project (EMAP) (Hayamizu et al., 2013; ema, 2015) aims to be a complete 3D atlas of mouse anatomy and development, and includes a phenotype ontology (Gkoutos et al., 2005). Lastly there are ontologies for Humans, mainly the FMA (Hunter et al., 2003), which has an ontology for the complete Human anatomy

as well as a developmental ontology for the first 20 Carnegie stages.

Given the large amount of model-organism databases, it is no surprise that there are efforts to unite these under common frameworks. UBERON, the "uber ontology" is striving to create a single reference ontology to relate all model-organism specific ontologies (Mungall et al., 2009; Mungall et al., 2012). The Common Reference Ontology (CARO) was designed as a species-independent (animal) anatomy framework (see chapter 16 in Dress et al. (2008)), and was the basis for development of the XAO and UBERON, and is also cross-referenced by the DAO. The Phenotype Annotation Ontology (PATO) (pato, 2015) is a similar project for annotating phenotypes (both natural and mutant).

Although the ontology is created for an organism as a whole, they are used for annotating properties of genes (or gene products). Model-organism ontologies are not used for describing naturally occurring diversity, and therefore are not suited to studying morphological evolution.

### Bio-ontologies for cells

There are very few ontologies dedicated to describing morphological diversity in cells. The foremost cell ontology is the Cell Ontology (CO), an ontology for cell types during development (Bard et al., 2005). This ontology spans all major branches of the tree of life, and contains terms for different cell types. It has been incorporated into several model-organism ontologies including the DAO, DPO, FMA and Mouse ontologies. The Subcellular Anatomy Ontology (SAO) (Larson et al., 2007) is the only ontology that contains terms for organelles and parts of cells. Although initially intended to capture the entire morphology of cells and their organelles, the only part actively developed is dedicated to neurons. This ontology contains terms for all major components of nervous system cells and cell types, including terms for describing morphology (anatomical properties). In 2013 this project was successfully integrated with the Gene Ontology (Roncaglia et al., 2013).

Neither of these ontologies serve the general purpose of studying naturally occurring diversity in cells and organelle morphology.

**Ontologies for genes & proteins**

Possibly the most known "bio-ontology" is the Gene Ontology (GO) (Ashburner et al., 2000). This was one of the first successful ontologies in molecular & cell biology, and is used to annotate gene function to a remarkable level of detail. Gene products can be annotated in all three major parts of the ontology: "Molecular Function", "Biological Process" & "Cell Compartment". However GO does not (as was never intended to) be used to describe morphological diversity *per se*. Even though it contains terms highly relevant to this thesis (such as CILIUM, AXONEME, etc.), the ontology cannot be used to study diversity from an morphological diversity perspective.

### 1.3.3 Bio-ontologies for Comparative Cell Biology

There are a large number of ontologies in the field of biology, and many of them have proven successful as a means to create large datasets of biological data. However, there are no ontologies suitable for studying morphological diversity in cells & organelles. The CO is targeted at describing cell types, and much like the model organism ontologies is intended to annotate gene function and not morphological diversity. The SAO makes room for morphological annotations, although is limited in scope to neurons. Likewise, the "cell compartment" of GO has terms for all major organelles in the cell, however lacks the terms required to describe morphological diversity.

It is from the non-model-organism ontologies used by taxonomists and systematics that we stand to learn the most about describing morphological diversity. First of all, an ontology for describing cell morphology should allow for quantitative descriptions of diversity. Secondly, it is important that this ontology be taxon-independent: otherwise cross-species studies become impossible. The concepts introduced in this section will become important in chapter 2 when we develop an ontology dedicated to studying the morphological & functional evolution of MTOCs across the eukaryotic kingdom.

## 1.4 Morphology Databases for Cell Biology

For most of its history the field of evolutionary biology has been an exercise in comparative morphology. Whether studying animals, plants or cells, the typical work flow is similar: Researchers create collections of images & illustrations on a set of closely related species, and only those with access to this entire collection could study their evolution. Cell biology has a similar history: Ultrastructural studies were published containing hand drawn illustrations and microscopy image "plates" accompanied by text descriptions. The age of computers and the internet has changed this work flow in many areas of biology, allowing scientists to collaborate and share data, as well as to create a single centralized repository to store data. As discussed previously (section 1.1.1), the past few decades have revealed that there is a large amount of biological diversity in cells and organelle morphology. In order to study the evolutionary mechanisms behind this diversity, we will first have to create a catalogue of this diversity.

### 1.4.1 Morphology Databases

Computers and the internet now allow for comparative morphologists to collaborate and share information as never before possible (Bisby, 2000; Sugden and Pennisi, 2000). Many different morphology databases exist, each tailored to address different research questions. I will briefly describe two main types of morphology databases; those dedicated to systematics and taxonomy, and those dedicated to studying model organisms, and end with a brief description of cell morphology databases.

#### Morphology databases for taxonomy & systematics

Unequivocally the greatest efforts to create detailed and comprehensive catalogs of biodiversity find their origins in taxonomy & systematics. These comparative morphologists have a strong history in collecting & describing biodiversity, although classically they have only been able to share their work as published (paper) material. The introduction of large online repositories has enabled taxonomists and systematicists world wide to embark on what might be considered their "holy grail": to catalogue and classify all existing biodiversity.

The Encyclopedia of Life (eol, 2015; Parr et al., 2014a) serves as a portal to "gather information and pictures of all species known to science". Although their emphasis is on collecting images, some collections are annotated using Traitbank (Parr et al., 2014b), an ontology created to allow a complete description of a species' behaviours, habitat, and some morphological descriptors.

There are two major projects which aim to facilitate collaboration between groups of systematicists. MorphoBank (O'Leary and Kaufman, 2011; morphobank, 2015) (O'Leary and Kaufman, 2011; morphobank, 2015) hosts several small projects for scientists working on a particular project to share and annotate specimens of species via character matrices. MorphBank (morphbank, 2015) is a similar project, which allows users to upload and annotate images. Although in theory neither of these is limited in scope, most of their collections are small projects centered collaborations on animal and plant diversity.

The most important goal of these projects is to collect images representing biodiversity, and when possible to quantify the morphological observations using ontologies or character matrices. Each of these projects also collects data in a similar fashion: By creating a community driven resource in which members can upload and annotate images. This aspect is incredibly valuable as the database content can be contributed from people around the globe, and moreover is not limited to images and specimens published in academic journals. But most notably, in all of these projects the "image" is the central point of reference: annotations and character matrices are always tied to the "raw data" (Ramírez et al., 2007).

### Morphology databases for model organisms

There are many online resources dedicated to housing information on model organisms. Although most of this information is typically centered around genes and proteins many of these resources are also making room for annotated image collections. Typically these images show gene expression localization, knockout/knockdown phenotypes, and occasionally developmental stages. ZFIN (Sprague et al., 2006) has a large collection of images annotated for expression localization for individual genes (annotated with GO), as well as images of various development stages, accompanied by text descriptions, and annotated using the ZAO. Flybase (St Pierre et al., 2014) contains scanning electron

microscopy images of life cycle stages and developmental stages, annotated with their ontology, but have no quantified morphological data. XenBase has illustrations of development stages, as well as links to gene expression (Karpinka et al., 2014; Bowes et al., 2009), but also lacks quantification of the information in these images. Yeast is a very suitable model organism for high throughput genomics and phenotype screens, and there are currently two major projects focussed on *S. cerevisiae* with a strong morphological component. These projects are The Phenomics of yeast Mutants (PhemoM) (Jin et al., 2012) and the Saccharomyces cerevisiae Morphology Database (SCMD) (Saito et al., 2004). These projects use automated image capturing & analysis to measure the morphological changes for thousands of mutants, including fluorescent labelling of specific cellular components including the nucleus, actin & microtubules.

Model organism centered databases are beginning to recognize the value in collecting images. By directly integrating them with other sources of data (gene & protein function, human disease, etc), these images can be used to make predictions and sometimes even novel discoveries. This is of tremendous value to experimental biologists working in these model systems. For our purposes however, the major drawback is that these databases are limited to a single species, making it impossible to do cross-species studies.

### Morphology databases for Cells

There are very few morphology databases focussed on cells.[4] The major cell image database is the Cell Image Library (Orloff et al., 2013), which has recently merged with what was started as the Cell Centered Database (CCDB) (Martone et al., 2002). This project has 2 main goals: To serve as a central repository for collecting annotated images of cell ultrastructure, and to provide a free and open collection of images for education and the public. This project is remarkable in being one of the few cell centered databases focussed on capturing natural variation in cell ultrastructure. Although not compulsory images may be annotated using (up to) 14 ontologies. However these ontologies (including many of those mentioned in section 1.5) are not for quantifying morphological diversity, but rather for annotating the organ or tissue source of the image

[4]We exclude SCMD and PhenoM from "cell" databases as these focus on the model organism aspects of yeast, and not on morphological diversity.

(and consequently are limited to animal specific ontologies). The Cell Image Library is an extremely valuable resource for collecting and sharing images of biodiversity in cells, and has an impressively broad species coverage. However, it does not allow for the annotation of morphological diversity in organelle shape, and therefore cannot be used to study the evolution of organelles.

### 1.4.2 Morphological Databases for Comparative Cell Morphology

Projects such as the Cell Image Library show that there is a growing interest in studying diversity at the level of cells and organelles. Like many biodiversity catalogues this project harnesses the power of community efforts to unite researchers from around the globe to work together on a single centralized project. In order to ensure congruity between different projects, these repositories standardize methods for annotation and quantification. Most of these projects also have a very strong "human" component: Annotations are done by knowledgable domain experts (as opposed to automated image analysis). Lastly, whether studying model organisms, cells, animals or plants, the central unit of data is the image, which remains linked to the data derived from it. From each of these projects we can learn valuable lessons which will guide the creation of a catalogue of diversity in cilia and centrosomes.

## 1.5 Outline of this thesis

The eukaryotic kingdom is brimming with morphological diversity, and despite decades of research, we lack a general understanding of how (and why) cells are the way they are. In this thesis we will take the "comparative" approach to cell biology, and study diversity as it naturally occurs in species throughout the eukaryotic kingdom. The chapters presented in this thesis address different aspects of comparative cell biology ranging from the evolution of shape to the evolution of amino acid motifs. What they have in common is that each of these requires development of novel bioinformatics approaches.

Before we can understand the evolution of diversity in organelle morphology and function we first need to obtain an overview of what this diversity is, what its limits are, and how it is distributed in the tree of life. In chapter 2, I present

***mtoc-explorer.org***: a community based resource to study the evolution of MTOCs. Over a period of two years reseachers from around the globe uploaded and annotated EM images of MTOCs from species covering the entire eukaryotic tree. Using this dataset we can for the first time adress how organelles evolve from a morpholoigcal perspective.

In the chapter 3 we combine "comparative morphology" and "comparative genomics" benchmark how well we can link genotypes with phenotypes in cell biology. "Phylogenetig profiling" can be used to predict protein's function based on whether or not it is present in all species with a particular phenotype (or morphology). Using the dataset generated in chapter 2 we ask how well "phylogenetic profiling" works using the presence of the eukaryotic cilium as a target phenotype, and what the main factors are that affect its performance.

Lastly in chapter 4 we study the evolution of the eukaryotic trafficking system using sequence based function prediction across a species genome. Rab GTPases are a family of proteins that are master regulators of intracellular trafficking. The *Rabifier* is a bioinformatics pipeline developed to predict which cellular processes are being regulated by a given Rab based on its amino acid sequence alone. The Rabifier was run on all eukaryotes who's genome had been sequenced, and the entire dataset and pipeline made available at ***RabDB.org***.

Finally, I conclude with a brief discussion on "comparative cell biology", and on what the age of bioinformatics means for the study of ancient organelles.

# Chapter 2

# The Evolutionary Cell Biology of Cilia and Centrosomes


## Abstract

One of the cornerstones of evolutionary biology is the study of morphological diversity, and how functional constraints shape the landscape through which this diversity is explored. Although this concept has been studied in organisms, its role in shaping cell biology is still poorly understood. By using cilia and centrosomes as model organelles we identify how function dictates morphological diversity in cells. Cilia and centrosomes are microtubule organizing centers (MTOCs) observed in all major eukaryotic branches, and play key roles in cell motility and division. Their stereotypical arrangement of 9-fold symmetrical doublet and triplet microtubules strongly suggests this conformation originated in the first eukaryote over a billion years ago. However these organelles have diversified in both structure and function in different eukaryotic branches.

To catalogue the diversity of MTOCs we created mtoc-exlorer.org: a community resource to collect and share images of microtubule derived organelles. Each image is annotated using an ontology designed to allow a detailed structural description of these organelles. With over 500 images from more than 100 species, this unique resource allows us to study the evolution of organelles.

Using the Morphological Diversity Index – a measure of observed vs. expected diversity – we show that the diversity in MTOC morphology is governed by constraints. Although the motile cilium has many different structures which define it, its overall morphology is greatly limited compared to immotile cilia. More surprisingly we also discover that the requirement for ciliary motility con-

straints the morphology of the mitotic apparatus, an evolutionary phenomenon known as a "spandrel". Lastly we develop Maximum Parsimony Landscapes, a method to test for convergent evolution across long evolutionary time-spans, and show that the centriole-based centrosome has evolved multiple independent times in almost all eukaryotic branches. This research shows that principles known to govern the evolution of plants and animals also operate in cell biology.

This is the first time that the evolution of a set of organelles has been studied quantitatively in great detail across such a broad taxonomic range. By creating a centralized resource to collect images, and a language to communicate and measure morphological diversity, we show that the interplay between structure and function also operates at a cellular level. We feel that the conceptual framework we present will not only offer novel insights into cell biology, but that it also can be used to study morphological diversity at any biological scale.

## Publication

This chapter is currently being prepared as a manuscript for publication.

## Author's contributions

This chapter marks the major part of the work performed during my PhD. As such I was responsible for conducting the experiments for this chapter, as well as writing the text and creating the figures. This work and writing was done in close collabration with José Pereira-Leal and Mónica Bettencourt-Dias.

We recieved extraorinarily valueable input from: Zita Carvalho-Santos, Renato Alves, Yoan Diekmann, Juliette Azimzadeh, Keith Gull and Michel Bornens.

We also thank all of the members of the **MTOC consortium** who contributed data to this project: Marlene Benchimol, James Braselton, Chris Bowler, Giuliano Callaini, Jean Cohen, Janet Chenevert, Alex Dammermann, Lillian Fritz-Laylin, Elisaveta Gonobobleva, Beatriz Ferreira Gomes, Ralph Gr af, Mathew Hoges, Kazuo Inaba, Sue Jaspersen, Grant Jensen, Hua Jin, Sergey Karpov, Katsaros, Alu Konno, Ryoko Kuriyama, Jadranka Loncarek, Pedro Machado, Shinichiro Maruyama, Neuza Matias, Brian Mitchell, Naomi Morrissette, Maxence Nachury, Geral Schatten, Nicole Scheumann, Chad Pearson, Jiri Vavra, Bill Wickstead and Mark Winey.

## 2.1 Introduction

The cell is the fundamental unit of life, as posited by the "cell theory" (see (Mazzarello, 1999) for a historical review). Cells accomplish their function in highly diverse spatial and environmental conditions, as unicellular organisms or part of large consortia of multicellular organisms, and with very diverse and distinct intracellular organisation. The evolution of this diversity is unclear. While major efforts have been put into understanding the molecular and developmental mechanisms behind the evolution of species, little attention has been devoted to evolution of the cell itself (Lynch et al., 2014). One of the major challenges in cell and molecular biology, as well as in evolutionary biology is thus to determine how cells originate, acquire and diversify their internal and external architecture (Biggins and Welch, 2014; Lynch et al., 2014).

The emergence of the eukaryotic cell, dubbed one of the major transitions in evolutionary biology (Maynard Smith and Szathmáry, 1995) is particularly fascinating as it represents the transition from a simple cell plan to a complex, highly compartmentalised one (Diekmann and Pereira-Leal, 2013). Many recent studies have focused on analysing the evolution of gene families that are associated with a specific organelle function and/or structure in order to gain an understanding about the origin of the organelle. The advantage of this approach is that as gene function is frequently conserved, the presence of a gene implies the presence of the function in that organism. One example is the peroxisome, whose presence is perfectly predicted by 4 highly conserved genes of the PEX family (Schlüter et al., 2006). However, studying gene repertoires is limited by the availability of sequenced genomes that are representative of any one specific biological trait. Furthermore, this approach can only be informative for well characterised gene families. In addition, molecules may indicate the presence of a given organelle or structure, but not its structure and regulation/context.

The cellular organization of ancestors of major taxonomic groups of eukaryotes is unclear, the role of physical constraints, historical contingency and adaptation in creating cellular organization and function is unresolved and, finally, the mapping of major cellular innovations on the tree of life is not obvious (Lynch et al., 2014). To answer these questions we need to understand the diversity of cellular life beyond the restricted number of model organisms that

have been the staple of molecular cell biology research programs. We also need to look beyond sequences alone, and form a basic understanding of the structure and function of cells and organelles, and evolutionary forces that may shape them. Recent technological advances in DNA sequencing and genetics make the investigation of non-model organisms more tractable, heralding the birth of an evolutionary cell biology (Lynch et al., 2014; Brodsky et al., 2012). However, a major challenge still remains in describing, quantifying and interpreting the cellular diversity, a challenge that we address in this study, focusing on the eukaryotic microtubule organizing center (MTOC).

The microtubule cytoskeleton is both unique to and ubiquitous in eukaryotes, and was one of the major innovations during the evolution of eukaryotes (Mitchell, 2007). Cilia and centrosomes, the two major microtubule organizing centers (MTOCs) of the cell, have been model organelles for morphological diversity for over 50 years (see section 1.2 of this thesis for a comprehensive overview). Here we focus on MTOC evolution to develop a conceptual framework for evolutionary cell biology.

Cilia and centrosomes are the major MTOCs of the cell, and have existed since the LECA (Mitchell, 2007; Jékely, 2007) (for a more comprehensive review of MTOCs, the reader is referred to section 1.2 in this thesis). They are involved in many cellular processes including sensation, motility, division, and establishing cell polarity. These two structures are linked in many species through the centriole/basal body (CBB): a cylindrical organelle of 9-fold symmetrical microtubule triplets which anchors the cilium (as the basal body), and in pairs forms part of the centrosomal complex for mitosis (as the centriole). The relationship between the basal body and the centriole, as well as the canonical 9 fold symmetrical architechture associated with each of these, is by no means the norm: The microtubule cytoskeleton shows an incredibly large amount of both structural and functional diversity throughout the eukaryotic kingdom, including complete losses and re-inventions of entire organelles (for a review see (Carvalho-Santos et al., 2010)).

The cilium is typically stated to be composed of 3 components: The basal body (BB), transition zone (TZ), and axoneme (Ax). The full length of the cilium is characterized by 9-fold symmetrical microtubule doublets (in the axoneme) and triplets (in the basal body). The motile axoneme is typically

decorated with dynein arms, nexins, radial spokes, and a central pair of microtubules. Immotile cilia are usually devoid of these decorations. However, this classical view of cilia is limited: there are many structural variations ranging from the presence and absence of different substructures to differences in numbers of microtubules and fold symmetry.

Likewise, the centrosome may take on a variety of different forms (for a review, see (Azimzadeh, 2014)). In animal cells the centrosome is almost always a pair of centrioles (9 fold symmetrical microtubule triplets), each surrounding a cartwheel scaffold and surrounded by a pericentriolar matrix (PCM) (Azimzadeh and Bornens, 2007; Bornens, 2012). There are however many variations on this theme, including the stacked centrosomes observed in many Fungi (Spindle Pole Bodies / SPBs, (Kilmartin, 2014)) and Amoebozoa (Nucleus Associated Bodies / NABs, (Roos, 1975; Ueda et al., 1999)) and the plates in Diatoms (Polar plaques, (Tippit et al., 1977)). Many other species have no mitotic MTOC visible using EM at all including many protists, higher plants and even the planarian (metazoan) *Schmidtea mediterranea* (Azimzadeh et al., 2012).

Cilia and centrosomes have been evolving for around 1.5 billion years (Yoon et al., 2004), and have essential roles in many extant species. The combination of structural and functional diversity balanced with a high degree of conservation makes the eukaryotic MTOC the perfect "model organelle" in which to study the morphological evolution in cell biology.

However the existence of morphological diversity alone is not enough to gain an understanding of how cilia and centrosomes evolve. In order to gain an evolutionary cell biological understanding of how organelles evolve, we will first need to create a catalogue of morphological diversity quantified in a manner that lends it amenable to computational analysis. The early decades of cell biology were characterized by a vast number publications showing Electron Microscopy (EM) images with ultrastructural details of species from throughout the tree of life. Although images are a valuable source of information, the real value lies in the expert interpretation of the structures visible in the image (Ramírez et al., 2007).

In the first sections of the results we specifically address how to obtain a comprehensive and highly quantified catalog of morphological diversity. In section 2.2.1 we outline the development of a novel ontology specifically designed

to annotate the morphological diversity of cilia and centrosomes from all eukaryotes. In section 2.2.2 we describe how we setup **mtoc-explorer.org**: an online community driven resource on MTOC diversity. This database allows users (members of the MTOC consortium) to upload and annotate (using the ontology) electron microscopy images of the species they work with. This website was used to create the catalogue of diversity covering over 100 species.

Subsequently, we proceed to address the fundamental biological questions regarding the evolution of cells and organelles. In section 2.2.3 we ask how diverse these structures are, and how this diversity is distributed throughout the eukaryotic kingdom. Afterwards (in section 2.2.4) we ask if this diversity is constrained, for which we develop the Morphological Diversity Index (MoDI), a new metric to quantify constraints in morphology. Next (in section 2.2.5), we examine one possible source of constraints by examining the effect of the requirement of ciliary motility on the morphological diversity of the mitotic MTOC. Lastly, we look at the historic relationship between the ciliary and mitotic machineries in section 2.2.6 , and make quantitative predictions on the presence of cilia and of centriole-based centrosomes in the LECA.

## 2.2 Results

### 2.2.1 The MTOC-ontology: An ontology for MTOC morphology

From the mid 1800's to the late 20$^{\text{th}}$ century, cell biology was almost entirely a descriptive field of science. Improvement after improvement in microscopy allowed early cell biologists to construct an increasingly accurate understanding of the inner workings of cells. The advent of electron microscopy (EM) in the 1940's spawned an era of prolific ultrastructural descriptions of cells from all branches of life, showcasing EM images of cells, organelles and the morphological diversity that characterized them. Microtubule based organelles, due to their ease of observation and diversity, were often a focal point of these studies (Haimo and Rosenbaum, 1981).

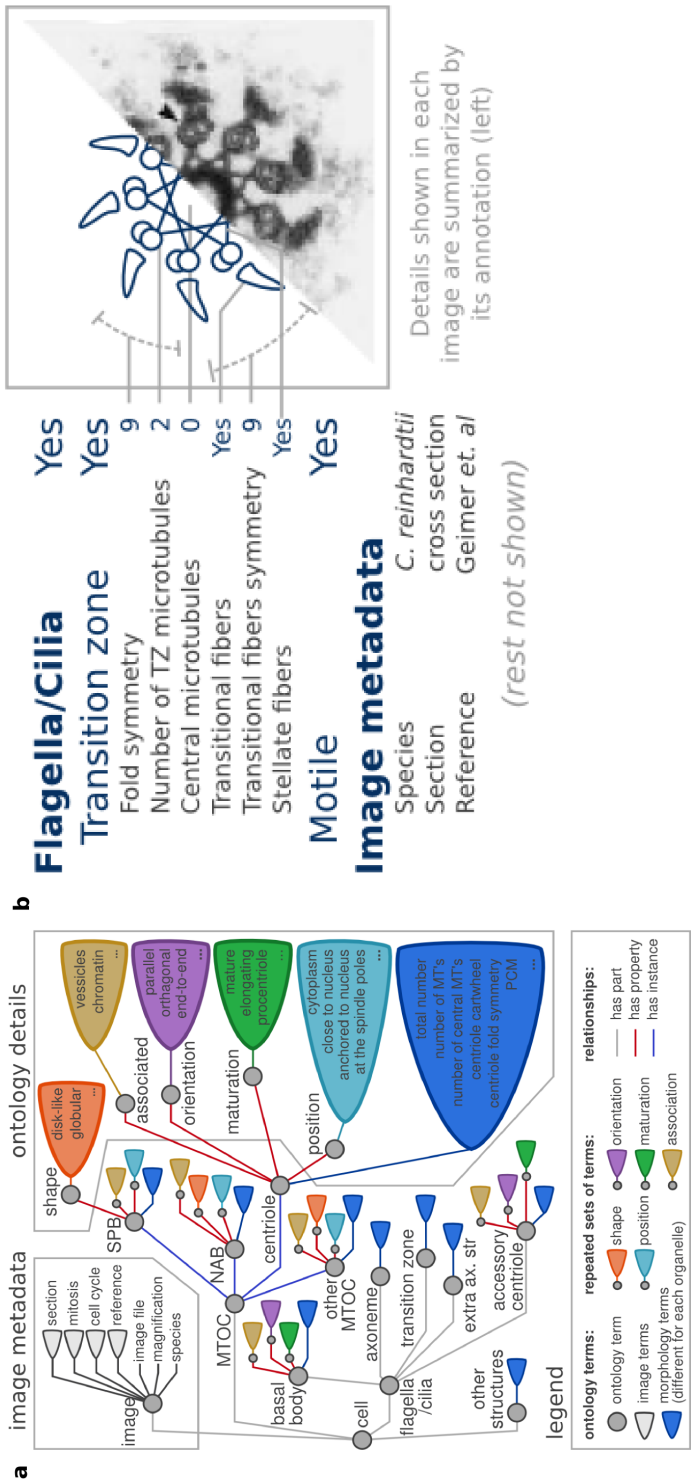
Early cell biologists frequently worked (and published) in separation from others working in related (and sometimes identical) organelles and cells. As a result organelles we now know to be the same were initially published under

different names, and these names continue to exist to this very day: "centrioles" vs. "basal bodies" and "cilia" vs. "flagella" vs. "undulipodia" (Margulis, 1980) are just two (particularly relevant) examples. These discrepancies in nomenclature make it impossible to directly compare results between different studies, and hence to obtain a clear picture of how organelles evolve. However, this problem is not unique to cell biology, and one of the most applied solutions to circumvent the errors introduced by natural languages is to define a structured controlled vocabulary, or "ontology" (Vogt, 2008; Vogt et al., 2009).

The use of ontologies in biology dates back to Linnaeus, who first outlined the taxonomic system of classification we still use today (Vogt, 2008). Recently, advances in computation have resulted in an explosion of ontologies, aimed at different fields of biology (for a review see section 1.3 of this thesis). In summary, there are 2 different types of ontology commonly used in different fields of biology. Firstly, there are ontologies used in evolutionary biology to systematize taxonomic classification based on morphology (for instance UBERON (Mungall et al., 2012), Phenex/Phenoscape (Balhoff et al., 2010; Dahdul et al., 2010), and PATO (pato, 2015; Mungall et al., 2010)). Each of these ontologies (or ontology frameworks) enables a complete description of morphological diversity of a collection of different species. However none of these ontologies are inherently capable of dealing with organelles. Second are ontologies aimed at molecular biology, including the Gene Ontology (Ashburner et al., 2000)[1], as well as model organism specific databases aimed at characterizing mutant phenotypes (ZFIN (Sprague et al., 2006), FlyBase (St Pierre et al., 2014) and SGD (Cherry et al., 2012). But these ontologies are not suited for describing morphological diversity, but rather for describing the localization or process involvement of gene products. None of the existing "bio-ontologies" address the fundamental problem of how to quantify morphological diversity in organelles.

We set out to create an ontology to describe the structural and functional diversity observed in microtubule derived organelles throughout the eukaryotic kingdom. The result is an ontology of over 300 terms specifically designed for the detailed annotation of microtubule derived organelles in a species independent manner (Figure 2.1). The higher levels of the ontology contain terms for the major microtubule based organelles which include CILIA/FLAGELLA, CENTRIOLES,

[1]For a note on why we did not use the Gene Ontology, please see page 60.


a
image metadata
ontology details
section
mitosis
cell cycle
reference
image file
magnification
species
image
shape
disk-like
globular
...
vessicles
chromatin
...
SPB
associated
parallel
orthagonal
end-to-end
...
orientation
MTOC
NAB
centriole
maturation
mature
elongating
procentriole
...
position
cytoplasm
close to nucleus
anchored to nucleus
at the spindle poles
...
total number
number of MTs
number of central MTs
centriole cartwheel
centriole fold symmetry
PCM
...
basal body
other MTOC
cell
axoneme
transition zone
flagella /cilia
extra ax. str
accessory centriole
other structures
legend
ontology terms:
ontology term
image terms
morphology terms
(different for each organelle)
repeated sets of terms:
shape
position
orientation
maturation
association
relationships:
has part
has property
has instance
b
Flagella/Cilia
Yes
Transition zone
Yes
Fold symmetry
9
Number of TZ microtubules
2
Central microtubules
0
Transitional fibers
Yes
Transitional fibers symmetry
9
Stellate fibers
Yes
Motile
Yes
Image metadata
Species
C. reinhardtii
Section
cross section
Reference
Geimer et. al
(rest not shown)
Details shown in each
image are summarized by
its annotation (left)

SPBs, NABs, and OTHER MTOCs. [2] The lower levels of the ontology allow for comprehensive description of the organelle(s), including terms for ORIENTATION, MATURATION and POSITION WITHIN THE CELL (where applicable), as well as organelle specific descriptors for structural components and morphology.

Other than a detailed description of the morphological features shown in each image each annotation is also accompanied by its IMAGE METADATA: information associated with the image, including its SOURCE, SPECIES, TISSUE TYPE, LIFE CYCLE/DEVELOPMENTAL STAGE and CELL CYCLE STAGE. Each annotated image is therefore a detailed ultrastructural description of an organelle including information on the cellular and species context. Thus it becomes possible to compare the MTOCs of (for example) interphase vs. mitotic cells of a single species or spermatozoids between different species. This information is typically not visible in the EM image itself, and needs to be extracted from the image's publication (or annotated by the person responsible for creating the sample).

The ontology was initially developed in collaboration with a team of experts working in various fields of cilia/flagella and MTOC research, and new terms were added to the ontology as novel specimens of diversity were encountered. During the development process decisions were made on precise definitions for each term to remove the ambiguities prevalent in the MTOC literature. For instance we strictly define a BASAL BODY as "a cylindrical shaped microtubule based organelle that anchors a CILIUM/FLAGELLUM" vs. a CENTRIOLE, which is

[2] We use THIS FONT to specify when we refer to terms in the ontology and ***this_one*** do denote relationships.

Figure 2.1 *(previous page)*: **An ontology for MTOC morphology.** The mtoc-explorer ontology is a hierarchical controlled vocabulary designed to allow a detailed annotation of cilia and MTOC's morphology. **a)** A diagram showing the major components of the ontology. A portion of the ontology (top left) contains "metadata" about the image. The remainder of the ontology is dedicated to structural annotation, and is split in two major types of organelles: MTOC's and Cilia/Flagella. Many parts of the ontology (position, shape, maturation stage, orientation & association) are applicable to various organelles, and are repeated throughout the ontology. Terms for describing morphology are different for each organelle. A detailed view of some of the 'leaf' terms for "centriole" and "SPB" is shown on the right. **b)** An example of an EM image annotated using the mtoc-explorer ontology: A Transition Zone from *Chlamydomonas reinhardtii* (Geimer and Melkonian, 2004) showing 9+0 fold symmetry, transitional fibers and stellate fibers.

"a cylindrical shaped microtubule based organelle forming part of the mitotic apparatus". More details about the ontology and how it was developed can be found in the Methods section 2.4.2.

The process of annotating an image consists of describing the observed structures exclusively in terms of the ontology. Figure 2.1B gives an example of an annotated cilium (from Geimer and Melkonian (2004)). The image shows a from *Chlamydomonas reinhardtii* Transition Zone (part of the cilium) with 9 fold symmetry, 2 microtubules per rotation, 0 central microtubules, transitional fibers (which are also 9 fold symmetric) and stellate fibers. The remainder of the information is available from the publication and surrounding text.

By annotating an image using an ontology, a single EM image becomes translated to a series of datapoints that contain all of the important information contained in the image and its expert interpretation. This includes information on both morphological diversity as well as its context, and any information not directly visible in the image. And lastly, by creating a single unified language, we can directly compare annotations created by individuals from different parts of the globe working in different model systems. A large collection of images can thus be translated to a dataset, which is now amenable to computational analysis.

### 2.2.2 mtoc-explorer.org: a database for MTOC diversity

One of the challenges in any comparative morphology project is obtaining a collection of well annotated data from a wide range of species. In many other fields of comparative biology, this gap is being filled by creating online resources and community projects which enable large communities of experts in different species to combine their knowledge into a single centralized repository. For a comprehensive review, the reader is referred to section 1.4 of this thesis.

Most of the projects dedicated to cataloguing morphological diversity are specific to animals & plants, such as MorphoBank (morphobank, 2015), MorphBank (morphbank, 2015), MorphDBase (morphDbase, 2015) and DigiMorph (digimorph, 2015). These resources allow groups of researchers to create a catalog of morphological diversity (often along with detailed trait matrices and other measurements) for the systematic classification of species. However, none of these projects are targeted towards annotating morphological data in cells.

One recent project has started to collect images on morphological diversity in cells: "The Cell: An Image Library" (library, 2015) is a resource in which users can upload images of cells and organelles, and includes the option to annotate these images using the Gene Ontology (Ashburner et al., 2000). However the Gene Ontology does not allow the annotation of morphological diversity (see section 2.4.2). What all of these projects have in common is that they involve community projects in which experts in certain fields upload and annotate images. This aspect of morphology is very important: A raw image alone may not convey all of the important morphological data, but remains essential to be able to return to in case of doubt, ambiguity or to serve as a reference (Ramírez et al., 2007).

Mtoc-explorer.org is a community resource that was created to capture the diversity in cilia and centrosomes from the entire eukaryotic kingdom (Figure 2.2). On the site members can upload and annotate EM images showcasing morphological diversity in cilia or centrosome structure using the mtoc-explorer ontology (see section 2.2.1). Members can annotate EM images from previously published data or contribute original (unpublished) EM. The complete content of the site is publicly available, and can be searched and browsed by species, structure or publication/authorship.

Mtoc-explorer.org is the first effort to characterize morphological evolution of organelles with the same completeness and comprehensiveness as is common in characterizing the evolution of larger species. After receiving over 500 contributions from over 40 members, the database now contains detailed ultrastructural descriptions from EM data of over 100 species from all major eukaryotic lineages (Figure 2.2). The species incorporated in the database were selected to a) represent a broad taxonomic range, b) to encompass the range of structural diversity and c) to include the most model organisms. Although the total number of species annotated in the database represents a fraction of extant species, we made sure to include any known species with clearly interesting morphologies. By attempting to include species beyond the well characterized model systems, the contents in the database are a fair approximation of existing morphological diversity as observed in nature. Although this resource is sure to grow in the future, the current contents are enough to start understanding the evolutionary cell biology of cilia and centrosomes.

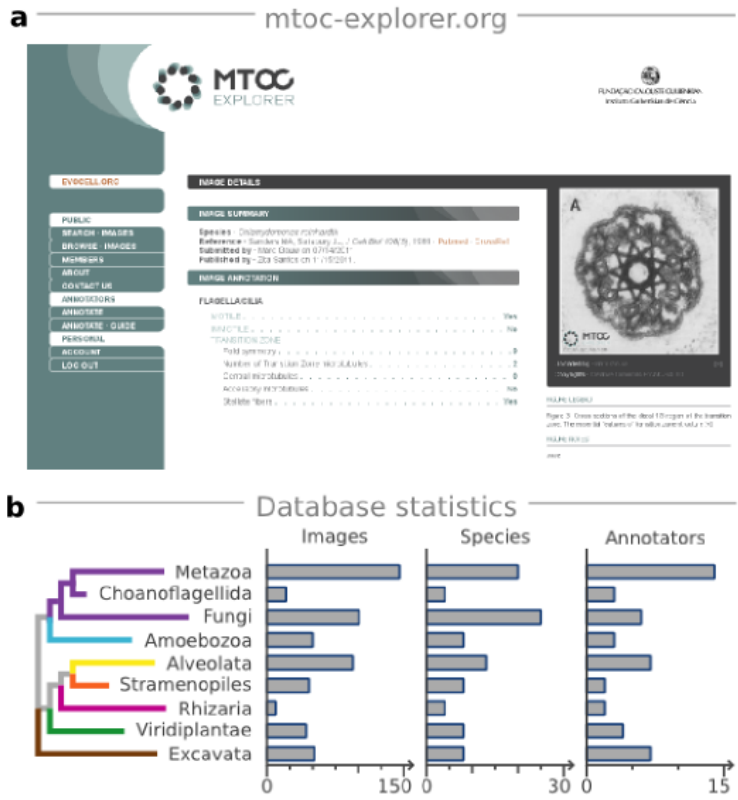



Figure 2.2: **mtoc-explorer.org: a community resource for studying the evolution of MTOCs and cilia.** ***mtoc-explorer.org*** is a community resource where members upload and annotate EM images of centrosomes and cilia. The collection of annotated images can be searched and browsed at ***mtoc-explorer.org***. **a)** A screenshot of an image annotated using the ontology (see Figure 2.1). **b)** The database currently contains over 500 annotated images from over 100 species from all major eukaryotic branches, each annotated by an expert annotator (statistics as of June 2015).

### 2.2.3 Morphological diversity in cilia across the eukarotic kingdom

Cilia and centrosomes are among the most prominent examples of diversity in cell and organelle morphology. To determine the evolutionary processes behind this diversity, we must first determine how this diversity is distributed throughout the eukaryotic kingdom. The first task we undertook with the catalog of annotated images in ***mtoc-explorer.org*** is a purely descriptive study in which we ask: "What is where?".

The cilium exists in all major branches of eukaryotes, however we do not know if they all have "the same" cilium, or whether different morphologies exist as taxon specific innovations. Specifically we wanted to address this question for the three major components of the eukaryotic cilium (the Axoneme, Transition Zone and Basal Body). This challenge involves summarizing the morphological annotations from over 200 images representing 75 different species described by over 150 ontological terms.

Inspired by the use of color gradients in heatmaps, we developed a "morphological heatmap" (Figure 2.3) in which we represent the diversity observed in a specific taxon by plotting the frequency of each annotation across all images belonging to that taxon. For example: STELLATE FIBERS characteristic of the *Chlamydomonas reinhardtii* Tranzition Zone appears to be a structure specific to Viridiplantae.

The structures typically associated to the "canonical cilium" (9-FOLD SYMMETRY with 2 MICROTUBULES in the AXONEME and 3 MICROTUBULES in the BASAL BODY along with a BASAL BODY CARTWHEEL and TRANSITIONAL FIBERS in the TRANSITION ZONE) are present in all major eukaryotic lineages. The structures typically associated with cilium motility (such as the CENTRAL PAIR of microtubules, INNER DYNEIN ARMS and OUTER DYNEIN ARMS, RADIAL SPOKES and NEXIN FIBERS) are likewise present in all major eukaryotic lineages. These results favour the notion that the "canonical motile cilium" was probably present in the Last eukaryotic Common Ancestor, a hypothesis we test later in section 2.2.6.

The fact that the cilium is a morphologically diverse structure is one of the reasons it was selected as the model organelle for this project. However, this

raises the question of whether or not there are any limits or constraints to this diversity.

### 2.2.4 Measuring morphological constraints in the eukaryotic cilium

One of the most striking phenomena in biology is that the observed variation in shape, although abundant, is much smaller that what we could imagine to be possible. This lack of morphological diversity is attributed to constraints in the morphological space (or "morphospace") (Hall, 2008; Raup, 1966; Raup and Michelson, 1965) available to living organisms. While evolutionary constraint has been abundantly explored in areas of biology such as paleontology, quantitative genetics and evo-devo (Arnold, 1992; Smith et al., 1985), little attention has been paid to the role of constraint in the evolution of cellular architecture and function. We set out to quantify morphological constraints in the evolution of cells using the annotated image collection in ***mtoc-explorer.org***.

In evolutionary biology the term "constraint" is often loosely defined (Pigliucci, 2007; Antonovics and Tienderen, 1991). Some argue that a "constraint" must be the result of (bio-) physical restrictions on form or function (Pigliucci, 2007; Pigliucci and Kaplan, 2000). A lack of morphological diversity may simply originate from a lack of genetic variation available for selection, resulting in historical contingency, and is therefore not indicative of a *bona fide* "constraint". Others, however, argue that historical contingency is simply another level of "constraint" in the evolution of an organism (Shanahan, 2008), and we will view "constraints" as any limitations in the outcome of evolution (see the reply to (Pigliucci and Kaplan, 2000) by (Getty, 2000)). This approach becomes more meaningful in large pan-species analyses (Sansom, 2008), in which the lack of any existing morphological variation directly implies some type of constraint on the outcome of evolution, regardless of its source (Arnold, 1992; Mezey and Houle, 2005).

In the previous section (2.2.3) we show that diversity exists in all three components of the cilium across all major eukaryotic lineages. What we did not quantify is the extent to which variation co-occurs both within and between different components of the cilium. As a hypothetical example, imagine all 9-fold symmetrical axonemes possessed both inner and outer dynein arms, as well

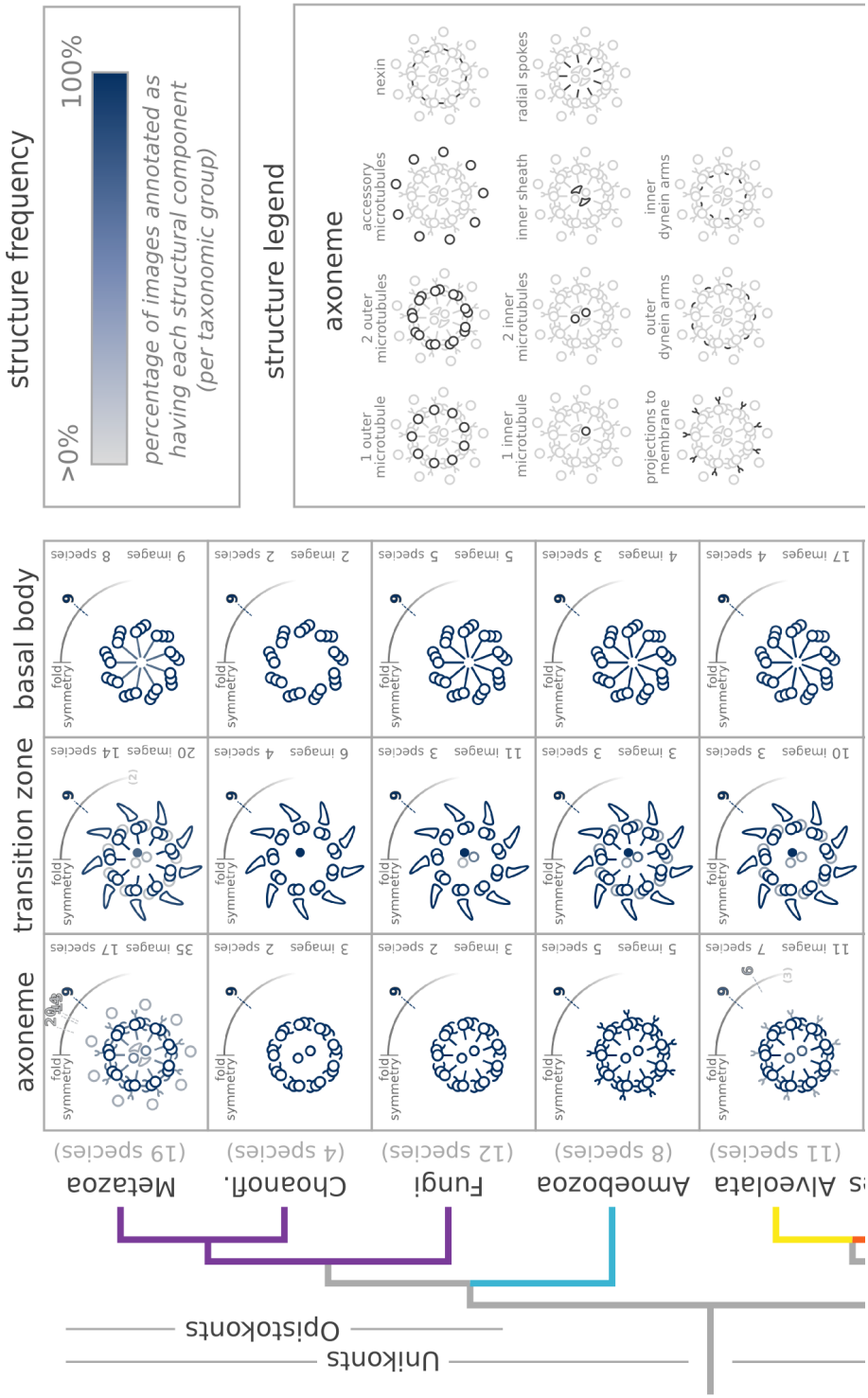
as a central pair of microtubules. This would directly imply an "all or nothing" constraint on the presence/absence of these three structures, compared to the theoretical possibility of their presence/absence being independent of each other. We set out to determine if the cilium is morphological constrained, whether these constraints are differently distributed between different components of the cilium, and whether they are affected by functional requirements.

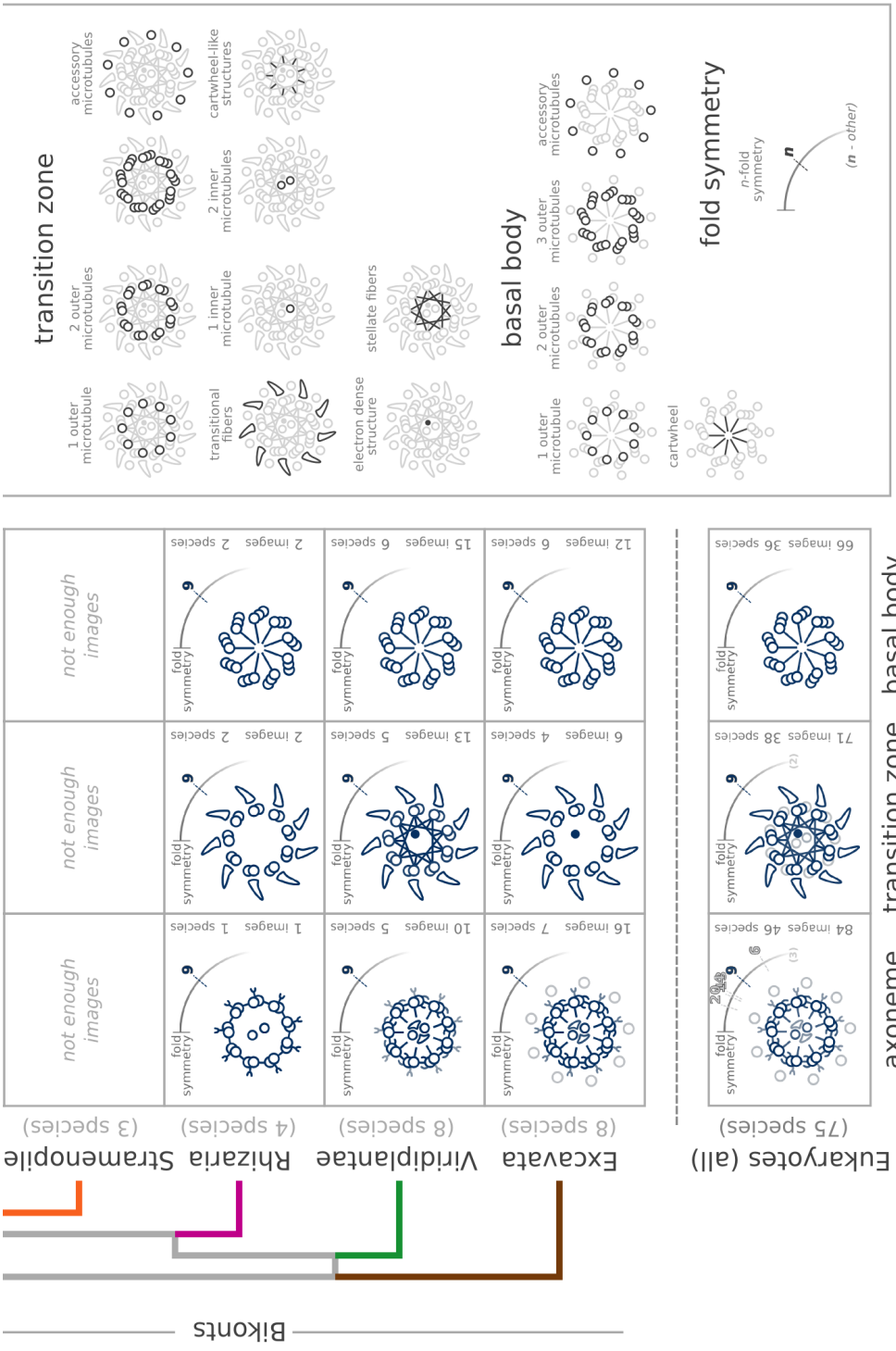
The level of constraint was measured by examining the co-occurrence of annotations across all images in ***mtoc-explorer.org*** compared to what would be expected by unconstrained evolution. This amounts to estimating the size of the theoretically available "morphospace", and measuring the fraction of this space inhabited by the observed data. To achieve this we developed the Morphological Diversity Index (MoDI): a metric that quantifies constraint for a collection of (morphological) data annotated with an ontology.

The Morphological Diversity Index (MoDI) is based on finding the dimensionality of the existing "morphospace" compared to a theoretically possible "morphospace", by measuring co-occurrence of annotations across images (Figure 2.4). To calculate the MoDI we first need to quantify the amount of co-variance in morphology (annotations). The rank of the covariance matrix of annotations represents the number of independent "morphospace" dimensions of the phenotype (Pavlicev et al., 2009). Essentially this amounts to performing Principal Component Analysis (PCA) on all annotations for a set of images, thereby removing all internal correlations between correlated morphologies. This number of components equals the "morphospace" populated by species annotated in the database.

The size of the theoretically possible "morphospace" estimated by creating 1000 random permutations of the existing annotations, simulating evolution in the absence of constraints, and performing PCA on each of these simulated datasets. The final MoDI is the number of principal components of observed data (i.e. observed diversity) divided by the average number of principal components of theoretically possible data (i.e. possible diversity). For more details on how the MoDI is calculated see the methods section 2.4.4. A MoDI close to 1 indicates that most of the possible morphologies exist, and hence low level of constraint, whereas a low MoDI suggests a high level of constraint.

The cilium is an organelle with extensive morphological diversity. Using


structure frequency
>0%
100%
percentage of images annotated as having each structural component (per taxonomic group)
structure legend
axoneme
1 outer microtubule
2 outer microtubules
accessory microtubules
nexin
1 inner microtubule
2 inner microtubules
inner sheath
radial spokes
projections to membrane
outer dynein arms
inner dynein arms
axoneme
transition zone
basal body
Unikonts
Opistokonts
Metazoa (19 species)
Choanofl. (4 species)
Fungi (12 species)
Amoebozoa (8 species)
Alveolata (11 species)
35 images 17 species
3 images 2 species
3 images 2 species
5 images 5 species
11 images 7 species
20 images 14 species
6 images 4 species
11 images 3 species
3 images 3 species
10 images 3 species
9 images 8 species
2 images 2 species
5 images 5 species
4 images 3 species
17 images 4 species
fold symmetry


transition zone
1 outer microtubule
2 outer microtubules
1 inner microtubule
2 inner microtubules
accessory microtubules
transitional fibers
stellate fibers
cartwheel-like structures
electron dense structure
basal body
1 outer microtubule
2 outer microtubules
3 outer microtubules
accessory microtubules
cartwheel
fold symmetry
n-fold symmetry
n
(n - other)
Bikonts
Excavata
Viridiplantae
Rhizaria
Stramenopile
(8 species)
(8 species)
(4 species)
(3 species)
Eukaryotes (all)
(75 species)
axoneme
transition zone
basal body
not enough images
16 images 7 species
10 images 5 species
1 images 1 species
6 images 4 species
13 images 5 species
2 images 2 species
12 images 6 species
15 images 6 species
2 images 2 species
84 images 46 species
71 images 38 species
66 images 36 species
fold symmetry

the data of the 101 species in ***mtoc-explorer.org***, we examined whether the observed morphological diversity showed any signs of constraint using the MoDI. Figure 2.5 shows that cilium is morphologically constrained, and only displays approximately 0.50 of theoretically available morphologies. As the MoDI is generic, it is not limited to measuring the level of constraint in an organelle, but can also be used to measure and compare constraints across different organelles, and different parts of an organelle. Figure 2.5 shows that the axoneme and transition zone are both structurally constrained to 0.55 and 0.58 respectively. The basal body, however, appears to be much more constrained than the axoneme and transition zone. These results show that both organelles as well as their components are morphologically constrained, and that there are different levels of constraint in different parts of cilium.

There are many different types of cilia which all fall into one of two functional categories: motile and immotile. We split the collection of cilia annotations into two groups – motile and immotile – and asked if motility had an effect on their morphological diversity. The MoDI of motile cilia (0.47) is significantly lower than that of immotile cilia (0.63), suggesting that the functional requirement of motility constrains the number of available morphologies (Figure 2.6).

Figure 2.3 *(previous page)*: **Diversity in cilia across the eukaryotic kingdom** Morphological diversity exists throughout the eukaryotic kingdom in the three major components of the Cilium. The morphological diversity of Basal bodies, Transition Zones and Axonemes across all major eukaryotic lineages is shown as a 'morphological heatmap', where the color intensity corresponds to the proportion of times each annotation occurs. For example, the Metazoa "Axoneme" morphological heatmap shows the annotations from 20 annotated images from 9 different species, and the "structure frequency" is measured as the number of images with YES vs. NO across all the images. The bottom panel shows the frequency of annotations for all species annotated in the database, and represents the "archetypal eukaryotic cilium". We only used images of cross-sections, since it is not possible to see many of the structures (for example the number of microtubules) in non-cross section images. Also, we only used a subset of of the terms deemed to be relevant for structural properties (for example, we removed all terms related to POSITION IN THE CELL).

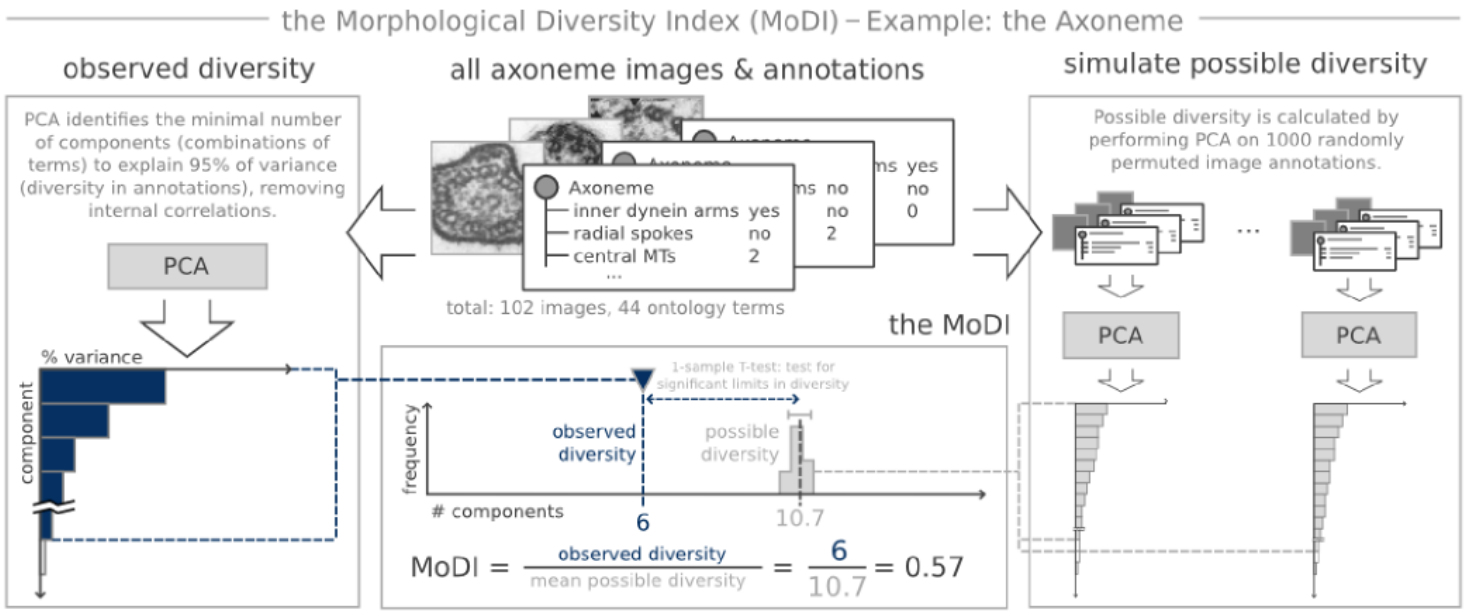



Figure 2.4: **The Morphological Diversity Index.** The Morphological Diversity Index (MoDI) is a measure of observed diversity as a fraction of the possible diversity. In this example, the "observed diversity" is measured by selecting all 102 annotations of images containing "Axonemes" and performing Principal Component Analysis (PCA) to quantify the degree of similarity within this set of images. In order to determine the number of components expected by unconstrained evolution we perform the same calculation using random permutations of the original annotations. The MoDI is defined as the fraction of "observed diversity" vs. the mean "possible diversity". Values below 1 indicate that the "observed diversity" is limited, i.e. less than the "possible diversity".

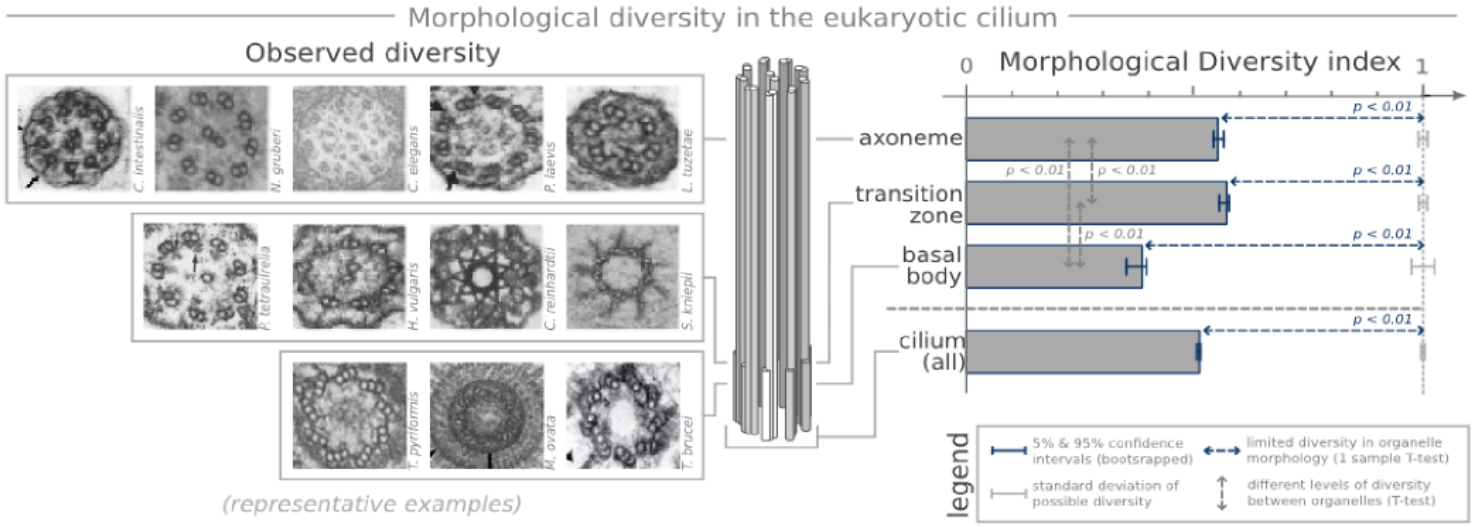



Figure 2.5: **The MoDI of cilia and cilium compartments.** The "Cilium" as well as its three major components, the "Axoneme", "Transition Zone" & "Basal body", are significantly limited in diversity. Moreover, different parts of the cilium are differentially limited in diversity: The "Basal body" appears to be most constrained, and the "Axoneme" the least. Figure references: Axoneme (left to right): (Konno et al., 2010), (Koonce et al., 1992), (Dingle and Fulton, 1966), (Perkins et al., 1986), (Heath and Darley, 1972), (Schrevel and Besse, 1975). Transition Zone (left to right): (Schrevel and Besse, 1975), (Wood, 1979), (Sanders, 1989), (Olson and Fuller, 1968). Basal Body (left to right): (Olson and Fuller, 1968), (Heath and Darley, 1972).

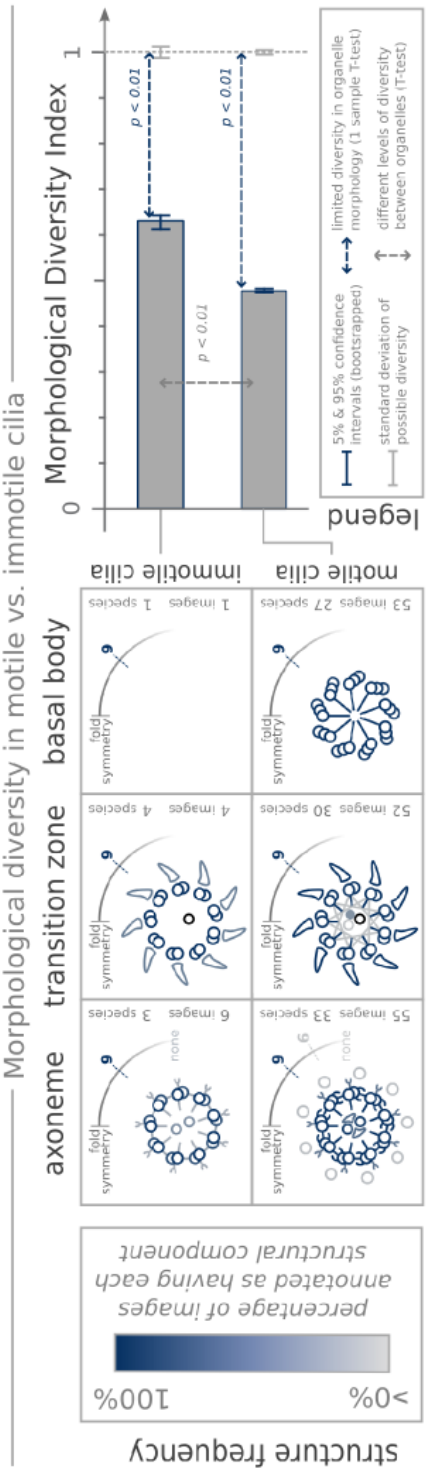



Figure 2.6: **The MoDI of motile and immotile cilia.** Motile cilia are structurally less diverse than immotile cilia. The MoDI of all annotations of motile cilia is significantly less than that of immotile cilia, suggesting that the functional requirement of motility constraints the morphology of motile cilia.

### 2.2.5 Cilia & centrosomes: a spandrel in cell biology

Most recent studies on mitosis from the past few decades have been in animal model organisms, in which the mitotic MTOC is a centriole-based centrosome: a pair of 9-fold symmetrical barrels that define the canonical centrosome. However, there is no direct requirement for a centriole-based centrosome for mitosis: Species exist which have a different mitotic MTOCs such as SPBs and NABs in Fungi and Amoebozoa. Moreover many eukaryotes perform mitosis without any visible MTOC at the spindle poles, including most plants and rhizaria. There is even an animal that has evolutionarily lost centriole-based centrosmes completely; the planarian *Schmidtea mediterranea* (Azimzadeh et al., 2012).

Although many animal cells do not require a centriole-based centrosome for mitosis, centrosomes are thought to be required for: ensuring mitotic fidelity, establishing cell polarity, signaling localization, and organizing the cytoskeleton as a whole (Debec et al., 2010). The presence of a centriole-based centrosome has been suggested to be the result of convergent evolution as a basal body separating machine (Debec et al., 2010; Azimzadeh, 2014). However, not all ciliated species use CBB localization to the poles for CBB segregation (many do not have a centriole-based centrosome). This leads to the question of if and how the presence of cilia (and therefore a CBB) affects the cytoskeletal architecture of the mitotic apparatus.

We grouped all species in ***mtoc-explorer.org*** into two groups: those with and without motile cilia, and calculated the MoDI across the collection of MTOC's observed in these two groups (Figure 2.7). The group containing species without motile cilia also contains all species which are non-ciliated. The results show that species with ciliary motility have MTOC's with a MoDI of 0.30. When the requirement for ciliary motility is absent, the observed morphological diversity doubles to 0.60. The presence of a cilium for motility affects the cytoskeletal machinery available for performing mitosis.

In evolutionary biology, there is a term reserved for traits that have evolved due to selection for another trait or a different function: the "spandrel" (Gould and Lewontin, 1979; Pigliucci and Kaplan, 2000) (or "exaptation" (Gould and Vrba, 1982)). In this instance the requirement for ciliary motility restricts the available morphospace of the mitotic apparatus. This would suggest that the

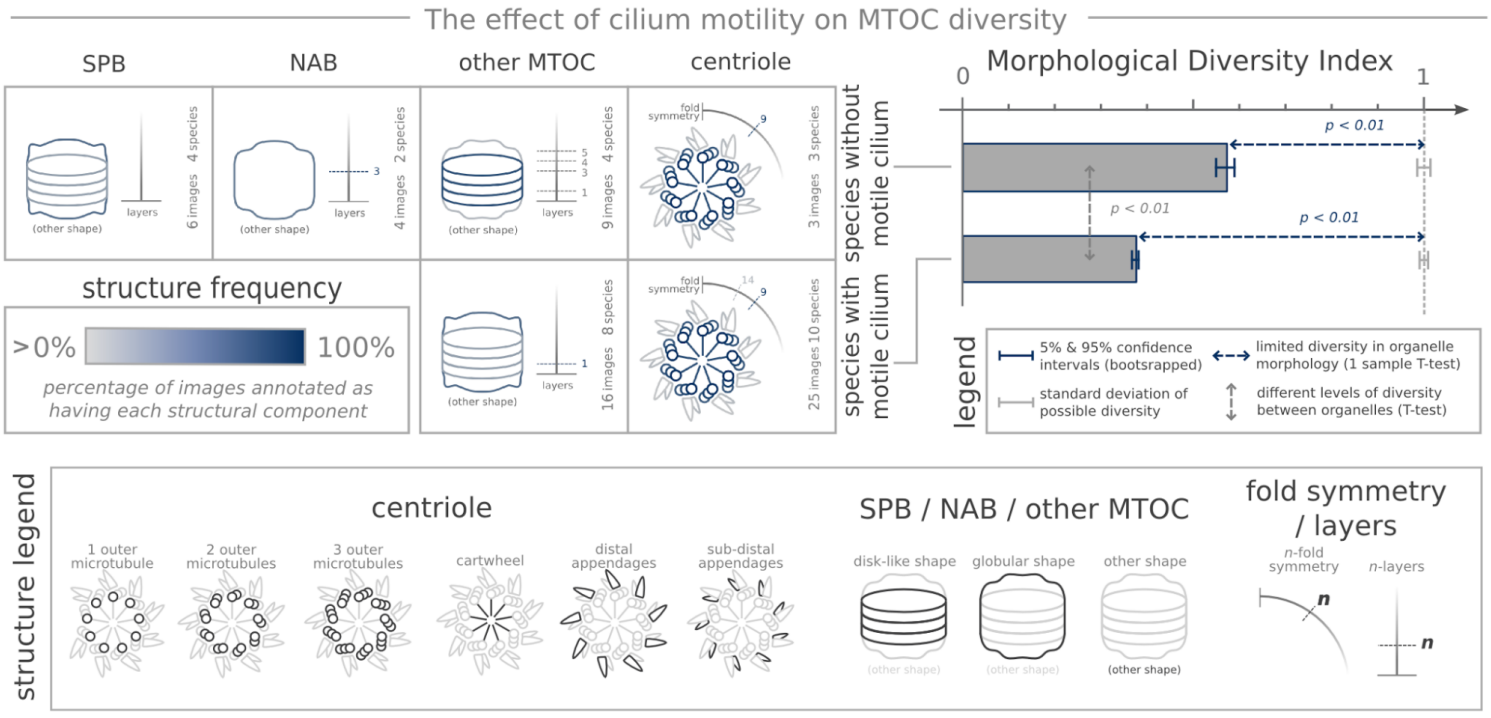



Figure 2.7: **The effect of cilium motility on MTOC morphology.** Morphological diversity of MTOCs in eukaryotes is affected by the requirement for cilium motility: species with motile cilia have morphologically less diverse MTOCs than those without motile cilia. We grouped all images into two groups: Those belonging to species with motile cilia, and those without motile cilia (including non-ciliated species). Species without motile cilia may use any of the 4 types of MTOCs contained in the database (centriole, SPB, NAB and other MTOCs) whereas species with motile cilia only have centrioles or SPBs for their mitotic apparatus. The MoDI of MTOCs from species with motile cilia is also significantly less than that of those without motile cilia, suggesting that the requirement for ciliary motility constrains the structure of a species' MTOC.

morphologically constrained mitotic MTOC in is indeed a cellular spandrel which has been co-opted to novel functions. This is to our knowledge the first time that the existence of a spandrel has been demonstrated quantitatively in organelle evolution. As we expand our knowledge and structural descriptions of organelle morphology and characterize more novel species, we hope to unravel in how far this phenomenon plays a roll in the evolution of cells.

### 2.2.6 Ancestrality versus convergent evolution of cilia and centriole-based centrosomes

Determining the origin (and therefore the age) of an organelle is fundamental to study how it has evolved. Usually a broad distribution across multiple eukaryotic basal groups is used as criteria to establish that an organelle is ancestral to all eukaryotes. This method excludes the possibility of gaining an organelle, and as a consequence the origin of an organelle is always the last common ancestor of all species that have the organelle. However, we know that complex patterns of organelle evolution occur, and allowing for organelle gains and losses can result in different interpretations of ancestral states.

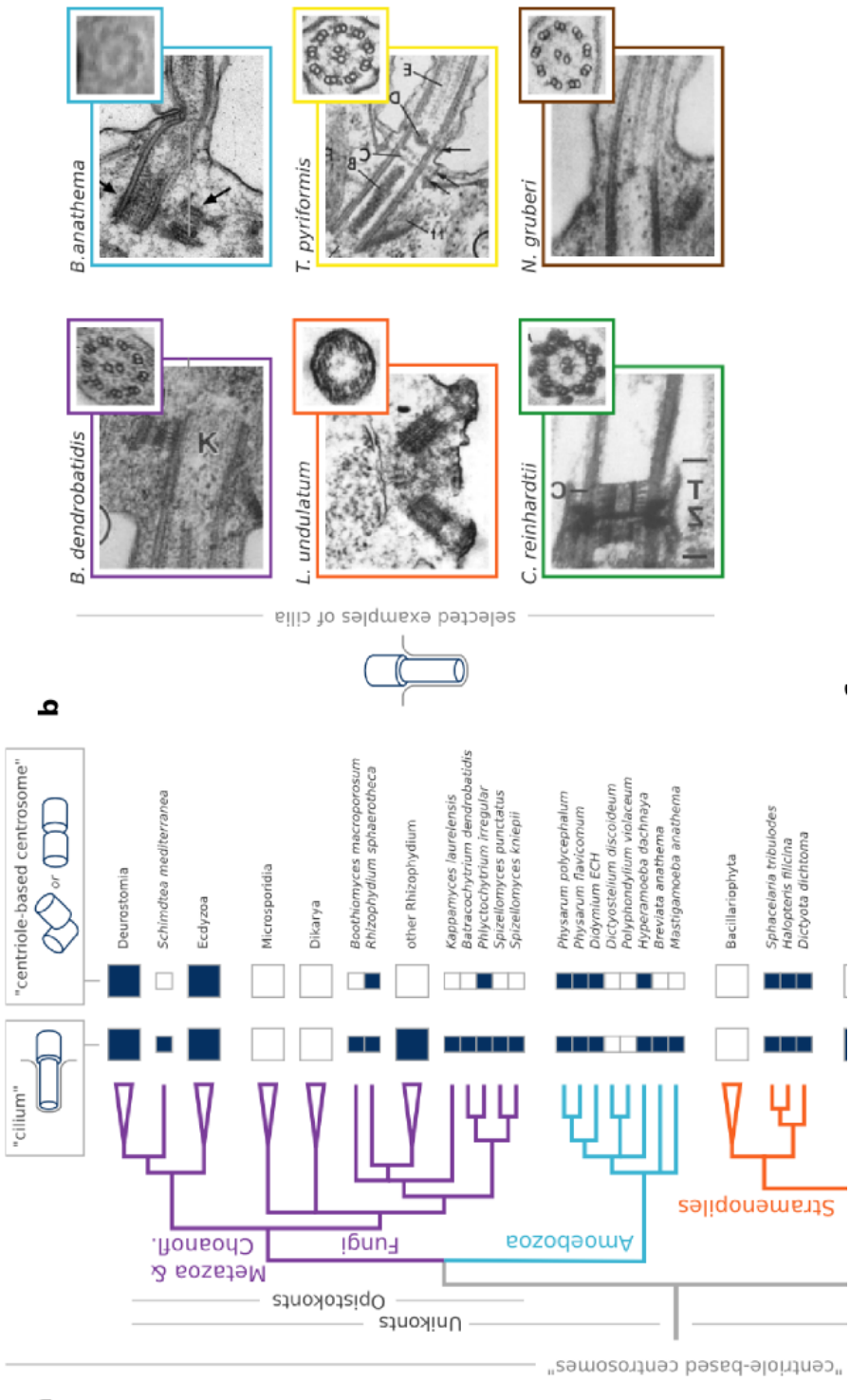
The cilium is present in all major eukaryotic lineages and is widely accepted as being present in the LECA (Carvalho-Santos et al., 2010; Mitchell, 2007). In constrast the centriole-based centrosome, the "other face" of the cilium, is commonly believed to be an evolutionary innovation that occurred in the ophistokont lineage (Bornens, 2012). Figure 2.8A shows the taxonomic distribution of species annotated as having either a CILIUM or a CENTRIOLE BASED CENTROSOME. Both organelles are present in (almost) all major eukaryotic lineages. Figures 2.8B and C show representative examples from mtoc-explorer.org of these structures. To ensure that the organelles we define as centriole-based centrosomes are not simply "Basal Bodies" in the cytoplasm, we specifically selected images annotated as containing POSITION IN THE CELL = AT THE SPINDLE POLES of images with CELL CYCLE STAGE = MITOSIS or MEIOSIS. Suprisingly we observed that centriole-based centrosomes exist in almost all major eukaryotic lineages. A naive interpretation of this observation would immediatly imply that the centriole-based centrosome, like the cilium, was present in the LECA.

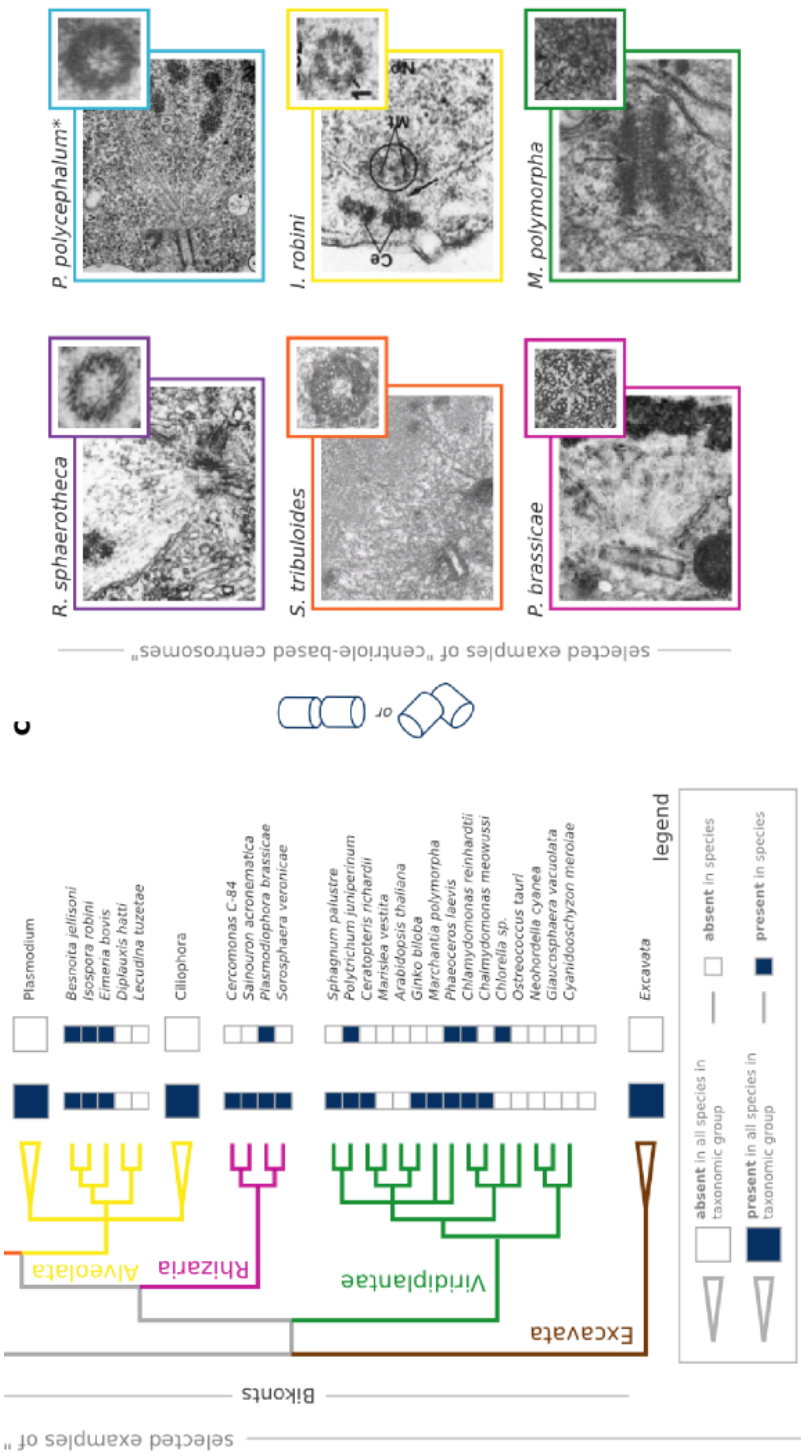
The broad but scattered distribution of these centriole-based centrosomes poses two different evolutionary scenarios: Either they are derived from a

common ancestor, or they evolved by convergent evolution (see also Azimzadeh (2014)). Evolutionary biology offers us objective methods to address such a question. We tested whether the LECA was likely to have had a cilium and a centriole-based centrosome using a Sankoff parsimony (Sankoff, 1975), a method suited to use in the absence of a true species tree (with divergence times), which does not exist for eukaryotes. This approach tests the probability of an ancestral state given a model of evolution, which in this case defines the costs associated with the gain and loss of an organelle. However, as we have no prior knowledge as to the cost of losing or gaining a centriole-based centrosome, we scanned the parameter space for a range of values for each of these. The result is a "Maximum Parsimony Landscape" (Figure 2.9A) , from which we can estimate the probility of ancestrality vs. convergence by determing how many parts of this parameter space favour ancestrality vs. convergence.

The results (Figure 2.10) show that the cilium is likely to have existed in the LECA ($p$(convergence) = 0.14) and that the centriole-based centrosome most probably evolved by convergence (probability of convergence $p$(convergence) = 0.62). As an extra validation, we also tested the probability that the centriole-based centrosome was present in the ancestral metazoan (as is commonly accepted (Bornens, 2012)), resulted in a $p$(convergence) of 0.13 (see supplementary material 2.5.2).

Convergent evolution of morphologically complex organelles may seem like an unlikely scenario. However, if we posit that its not the re-invention of an organelle, but rather convergent evolution of organelle position, this scenario seems more plausible. The plant kingdom serves as an example of how convergent evolution of organelle position may be linked to the generation of gametes. Of the 11 plants in the database, 5 species have sperm, and thus require ciliary motility. Notably the three plants with centriole-based centrosomes (*P. juniperium*, *M. polymorpha*, *P. laevis* and *C. reinhardtii*) have uni- or biflagellated gametets. *Ginko biloba*, on the other hand, does not have centriole-based centrosomes during meiosis. However, its sperm are multiflagellated, and during spermatogenesis has blepharoplasts, which are used to create centrioles "de novo" (Gifford and Larson, 1980). These results strongly suggest that the position of CBBs at the spindle poles during cell division has emerged due to convergent evolution with the need to segregate Basal Bodies


B.anathema
T. pyriformis
N. gruberi
B. dendrobatidis
L. undulatum
C. reinhardtii
selected examples of cilia
b
"centriole-based centrosomes"
Unikonts
Opistokonts
Metazoa & Choanofl.
Fungi
Amoebozoa
Stramenopiles
"cilium"
"centriole-based centrosome"
or
Deurostomia
Schimdtea mediterranea
Ecdyzoa
Microsporidia
Dikarya
Boothiomyces macroporosum
Rhizophydium sphaerotheca
other Rhizophydium
Kappamyces laureiensis
Batracochytrium dendrobatidis
Phlyctochytrium irregular
Spizellomyces punctatus
Spizellomyces kniepii
Physarum polycephalum
Physarum flavicomum
Didymium ECH
Dictyostelium discoideum
Polyphondylium violaceum
Hyperamoeba dachnaya
Breviata anathema
Mastigamoeba anathema
Bacillariophyta
Sphaelaria tribuloides
Halopteris filicina
Dictyota dichtoma
a


P. polycephalum*
I. robini
M. polymorpha
R. sphaerotheca
S. tribuloides
P. brassicae
selected examples of "centriole-based centrosomes"
or
c
Plasmodium
Besnoita jellisoni
Isospora robini
Eimeria bovis
Diplauxis hatti
Lecudina tuzetae
Ciliophora
Cercomonas C-84
Saimouron acronematica
Plasmodiophora brassicae
Sorosphaera veronicae
Sphagnum palustre
Polytrichum juniperinum
Ceratopteris richardii
Marsilea vestita
Arabidopsis thaliana
Ginko biloba
Marchantia polymorpha
Phaeoceros laevis
Chlamydomonas reinhardtii
Chalmydomonas meowussi
Chlorella sp.
Ostreococcus tauri
Neohordella cyanea
Giaucosphaera vacuolata
Cyanidooschyzon merolae
Excavata
legend
absent in all species in taxonomic group
present in all species in taxonomic group
absent in species
present in species
selected examples of "
Bikonts
Excavata
Viridiplantae
Rhizaria
Alveolata

for ciliated species.

## 2.3 Discussion

We have presented a framework to study the evolutionary cell biology of an organelle, focusing on the eukaryotic cilium and centrosome. In contrast to most current approaches to study the evolution of cells, we focused on morphology and function rather than on cataloguing genes. As a community we organised ourselves to produce a unified controlled vocabulary to describe the multiple features on these organelles (mtoc-explorer ontology), compiled an extensive data set of cellular diversity (mtoc-explorer.org) and finally developed methods to study variation and ancestrality. With these tools we revealed a) that the centriole-based-centrosome may be a recurrently evolved organelle b) that different parts of one organelle are under different evolutionary constraint and finally c) that constraints on organelle morphology may originate from functional requirements both within and between organelles.

The mtoc-explorer ontology is the first ontology to allow a complete and comprehensive description of cilium and centrosome morphology. Although the current version is limited to cilia and centrosomes, it provides a template for further development of ontologies for other organelle morphologies. The morphological descriptors we use are cilium/centrosome specific, but many of the terms and relationships are apt for describing generic organelle morphology:

Figure 2.8 *(previous page)*: **Cilia and centriole-based centrosomes across the eukaryotic kingdom.** The cilium and the centriole-based centrosome are observed throughout the entire eukaryotic kingdom. **a)** The taxonomic distribution of centriole-based centrosomes suggests that the presence of a centriole-based centrosome in a species indicates that some of its cell types or life cycles stages are also ciliated. **b)** Examples of cilia across the eukaryotic kingdom. The cilium has long been considered an ancient organelle dating to the Last eukaryotic Common Ancestor (LECA), and frequently observed in all major eukaryotic lineages. **c)** Examples of centriole-based centrosomes across the eukaryotic kingdom. References for EM images: b: *R. sphaerothec* (Powell, 1980), *P. flavicomum* (Aldrich, 1969), and *P. polycephalum* (Gely and Wright, 1986), *S. turbuloides*(original microscopy contributed by Christos Katsaros), *I. robini* (Desser, 1980), *P. brassicae* (Braselton, 1988), (Garber and Aist, 1979), *M. polymorpha* (Moser and Kreitner, 1970). * As there are no documented cross-sections of centrioles from *P. flavicomum* the cross-section shown is from *P. polycephalum* from the same genus.

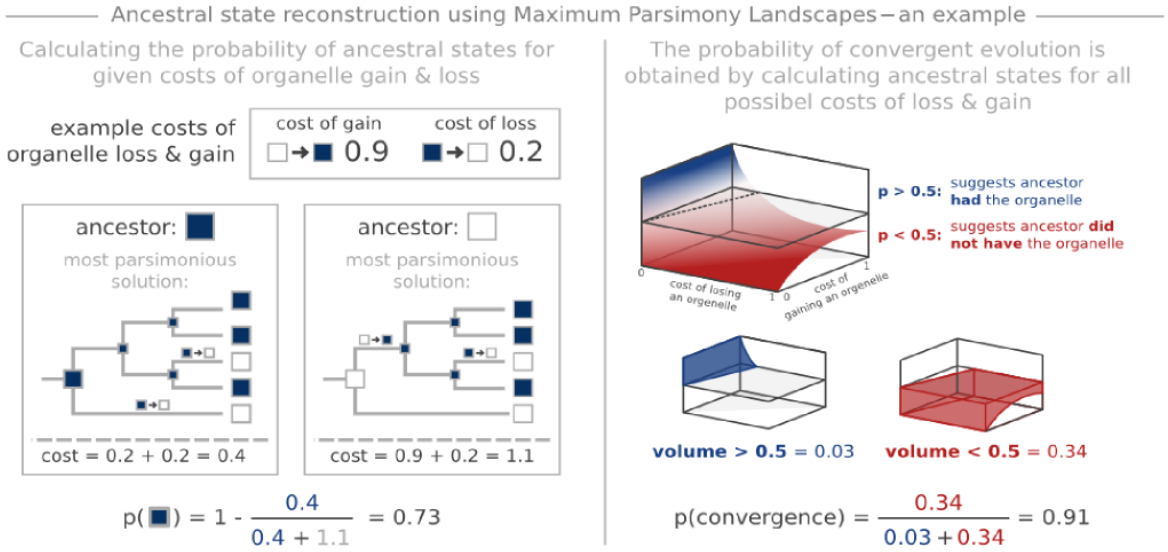



Figure 2.9: **Maximum Parsimony Landscapes.** "Maximum Parsimony Landscapes" are a way to infer ancestrality vs. convergent evolution of an organelle based on its presence/absence profiles in existing species. In this example, Maximum Parsimony is used to calculate the probability of an ancestral state given the evolutionary costs of losing or gaining an organelle (left). When these costs are unknown, we can calculate the probability of ancestral states over all possible combinations of their values (right). A high probability of the ancestror not having an organelle directly supports the likelihood of convergent evolution. The probability of convergence ($p$(convergence)) is defined as the sum of probabilities favouring convergent evolution over the sum of all probabilities.

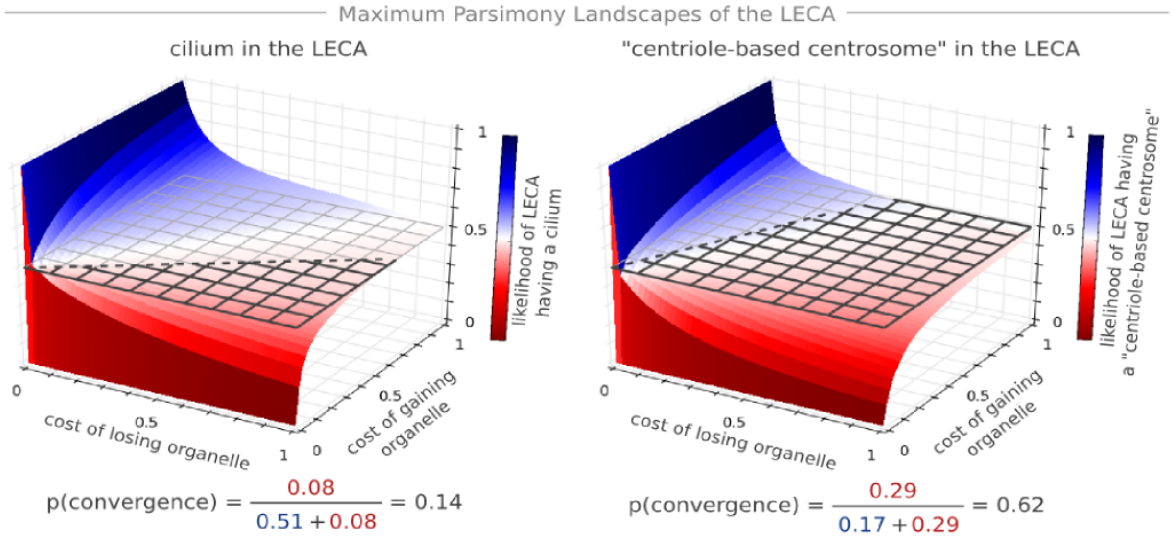



Figure 2.10: **Ancestrality vs. convergent evolution of cilia and centriole-based centrosomes.** Although the Last eukaryotic Common Ancestor (LECA) most likely was ciliated, the centriole-based centrosome is more likely to have evolved by convergent evolution in different species (Azimzadeh, 2014). The probability of convergence of the cilium and the centriole-based centrosome from the LECA. The cilium (left) is considered to date back to the LECA supported by its "Maximum Parsimony Landscape". However, the centriole-based centrosome probably evolved in multiple eukaryotic lineages by convergent evolution.

shape, position, etc. Moreover, the mtoc-explorer ontology also takes into account morphological context such as cell cycle stage, developmental context and tissue types. The field of evolutionary cell biology requires an ontology for cell/organelle morphology that extends beyond protein localization, and is dedicated to characterizing morphological diversity. Similar ontologies already exist in many other fields of biology, and the mtoc-explorer ontology provides such a framework.

The current version of the ontology will soon be made available on the OBO Foundry: the major portal of all "bio-ontologies" (Smith et al., 2007). A similar project with the aim to study the anatomy of the cell: The Subcellular Anatomy of the Cell (SAO) (Larson et al., 2007) has been recognized by and incorporated into the Gene Ontology However this ontology is limited to the ultrastructure of neurons. We hope the mtoc-explorer ontology (where applicable) is also included as part of GO in the near future.

An ontology is a "limited" view of the possible diversity, and any morphological diversity which is not part of an ontology will not be captured. The mtoc-explorer ontology was designed by groups of researchers with common interests and different backgrounds. This increases the likelihood that the ontology reflects a balance between a) objectivity and taxon independence and b) focus on containing the type of terms which are interesting to current research projects. However it is important that an ontology is not static, and we hope to see the ontology grow in the future.

The 100 species dataset created for this project is unique: it covers an extremely broad taxonomic range at a very high morphological resolution. The species were selected to represent the naturally occurring diversity in eukaryotic MTOCs. Nevertheless, we appreciate that the database content and any results generated from it will never truly reflect the natural world. The database content represents only a fraction of species that have ever been examined by EM, there are many more published (and even more unpublished) micrographs that are not included in the database. Furthermore, the number of species studied by EM represents only a fraction of species that exist today, which in turn are only a fraction of all species that have ever existed.

Another major obstacle in working with annotations is that it is impossible to differentiate between "does not exist" and "has never been seen". For

example: many other examples of centriole-based centrosomes may exist outside of opistokonts, which simply have never been analysed under a microscope. This limitation is general to all database projects, including the vast number of genomic sequences databases: our knowledge is limited by what we chose to study and annotate. We look forward to seeing both the ontology and database develop to encompass a more accurate representation of naturally occurring diversity. Moreover we look forward to seeing more species and cell types being characterized morphologically, to keep pace with the ever expanding genomic databases. Although we do not expect the general findings of this paper to change, we look forward to seeing how our understanding of diversity improves as we add more annotated images to the database.

Most of the research on cilium & centrosome morphology from the past few decades has been using EM, and the mtoc-explorer ontology & mtoc-explore.org database have been designed around this. However, there are other forms of microscopy which can equally well contribute to understanding cell and organelle morphology. For example, fluorescence-based microscopy techniques are highly suited for determining the POSITION IN THE CELL. We intend from the ontology and database to be extended to different data formats.

Despite their abundance, community driven annotation projects are notoriously difficult. The model we chose was different than an "open" community project: instead we selected and contacted annotators with requests to submit data. This approach has been shown to work more effectively than, for example, reward based methods (Mazumder et al., 2010). However we found it challenging to find annotators for many of the less well known species, as these are very niche specific. As with many community projects in biology, a large portion of the work ended up being done by a relatively small number of researchers who are typically closely related to the project. Community driven projects typically also require an initial phase of growth largely driven by a small number of people before reaching a critical mass. We hope that ***mtoc-explorer.org*** soon reaches this critical mass in after which the database content & community will continue to grow.

Understanding evolutionary forces and mechanisms in any system will always be an uncertain business: we will never know with absolute certainty why species are the way they are. So far in cell biology most "cell ancestral state"

and "constraints in evolution" publications are essentially "just so stories". Using quantitative methods we can now offer a certain degree of certainty about our predictions. We fully realize the "Morphological Diversity Index" and "Maximum Parsimony Landscapes" are not perfect: both of these make several assumptions about the nature of evolution, and have built in heuristics to cope with the large evolutionary timespans involved. However we wished to set the stage for quantitative ancestral state reconstruction in evolutionary cell biology. If the results still hold in the near future, as our algorithms and datasets improve, is a question that remains to be answered.

In this work we show strong evidence that the cilium is an ancient organelle and that the centriole-based centrosome has emerged multiple times throughout evolution. The antiquity of the cilium has been supported by the presence of genes known to be required for cilium formation (Carvalho-Santos et al., 2010; Carvalho-Santos et al., 2011; Hodges et al., 2010). However we know little about the components specific for centriole-based centrosomes across different kingdoms, and how evolution at the protein level supports our claim for convergent evolution.

The MT cytoskeleton has been evolving for over a billion years, and we show a complex pattern of morphological diversity coupled to function both within and between different organelles. These phenomena are not unique to MTOCs: constraints, spandrels and convergent evolution are concepts that are common to the evolution of entire organisms as well as genetic sequences.

To our knowledge, this is the first attempt made to quantify (absolute) morphological constraints in cell biology. The ubiquitous presence of cilia has resulted in the adoption of centrioles to the mitotic apparatus and are now an integral (required) part of the machinery. Not only has the evolution of the CBB affected the morphology of MTOCs, but in metazoa, it is an example of co-option of an organelle: many metazoan cells do not divide properly without a centrosome.

The results presented in this paper, have been dedicated to studying the naturally occurring diversity in cilia and centrosome morphology. However there are other forms of diversity which are of great interest to the cell biology community: those observed in disease phenotypes and those of transgenetic mutant phenotypes. With this work we have provided an ontology and database

framework which can be extended to study organelle morphology in any scenario, including disease and laboratory constructs.

Evolutionary Cell Biology is still an emerging field of science (Lynch et al., 2014), and in order to see it succeed we need to develop new frameworks with which to enter this new era. In this paper we have presented multiple new novel concepts including: new ways to use the internet & computers to ease and facilitate large collaborative efforts & quantification of data. As well as conceptual frameworks for dealing with this data.

## 2.4 Methods

### 2.4.1 The Taxonomic Tree of Eukaryotes

The NCBI taxonomy database is among the most complete species databases online. As there are no phylogenies of Eukaryotes, we use the NCBI taxonomic tree as the reference tree for all of the work on this paper, and on ***mtoc-explorer.org***.

We used the NCBI taxonomy version 12, with the following changes (to take into account updates to the basal positions of a handful of eukaryotic kingdoms):

- Added the group *Opistokont* (as a child of *Unikonts*) which includes all *Fungi*, *Choanoflagellida*, *Nucleariidae* and *Metazoa*.
- Moved *Amoeboza* to have parent *Unikont* (as a sister group to *Opistokont*).

The "major taxonomic groups" used are those proposed by (Baldauf, 2003a), with the following changes:

- *Discritase* is grouped together with *Excavates*.
- Separate *Opistokonts* to *Metazoa Choanoflagellida Fungi*.

Unless otherwise stated, the above are the taxonomic trees and major eukaryotic branches used throughout the rest of this chapter.

### 2.4.2 The mtoc-ontology

In order to capture the morphological diversity in MTOCs and their derived organelles, we developed an ontology dedicated to capture MTOC morphology.

Formally, an ontology is a directed graph whose nodes are "terms" connected by directed "relationships" that define how the terms are related to each other.

The ***mtoc-explorer.org*** ontology has a single root term: CELL, and has 4 different types of "terms".

- Class: A category describing a collection of a type of structure (for instance, MTOC)
- Structure: A physically observable organelle, or component of an organelle (for instance: CENTRIOLE)
- Property: An observable and measurable descriptor of a Structure.
- Value: A measured value.

In the ***mtoc-explorer.org*** ontology, relationships are defined in the direction parent to the child.

There are 5 types of relationship defined in the mtoc-explorer ontology:

- ***has_instance***: Links a class to the different structures that are examples of the class: the class MTOC ***has_instance*** CENTRIOLE
- ***has_part***: Link 2 structures, of which the child node is a component of the parent structure. For example: CENTRIOLE ***has_part*** CENTRIOLE CARTWHEEL.
- ***has_property***: Connects a structure to a property, for example: CENTRIOLE ***has_property*** FOLD SYMMETRY.
- ***has_value***: Either the relationship between a structure and a value (indicating presence or absence of that structure), or the relationship between a property and a value. For example: the property FOLD SYMMETRY ***has_value*** 9.
- ***associated_with***: a physical association of a structure with another structure (organelle or part of the cell).

The initial ***mtoc-explorer.org*** ontology was initially developed in collaboration with a team of experts from various fields of cilium, centrosomes and MTOC research. New terms were added as the database expanded following the recommendations of contributors. For each recommendation, the mtoc-explorer curators were the final judges on whether a new terms should be added, and where it best fit in the ontology.

**A note on the Gene Ontology**

The best known ontology in cell & molecular biology is the Gene Ontology (Ashburner et al., 2000); an ontology designed for the functional annotation of gene products. Although the mtoc-explorer ontology and the GO have much in common (and indeed components of the mtoc-explorer ontology will be used to extend under-characterized parts of GO) they are also fundamentally different: Whereas the GO (even the Cellular Component part) is designed only to reflect gene function, the mtoc-explorer ontology is designed to describe the morphological diversity of MTOCs and moreover allows for the functional, positional and other characters, independent of any genetic factors. Therefore we were not able to use the Gene Ontology to describe diversity in MTOC morphology.

### 2.4.3 mtoc-explorer.org: The Web Resource & Database

As we were not able to find any existing resources that addressed our needs, we developed ***mtoc-explorer.org*** ourselves based on existing general purpose web frameworks. After creating the website, and implementing image uploading and annotation, we selected a small community of approximately 40 members to upload and contribute data. In order to obtain the data used for this paper, we relied on contributions from a small community of experts to upload an annotate images and a small team of curators to check each contribution. Annotators were selected based on their area of expertise in microtubule based organelles and electron microscopy to represent as wide a possible coverage of different eukaryotes and a diversity of microtubule based structures. This community was responsible for uploading and annotating the (over) 500 images used for analysis in this paper. After annotators have uploaded and annotated their images, and confirmed that the annotation is complete, the image and annotation are passed on to a small team of curators. The curators verify the image quality, annotation correctness, as well as any pending copyright issues. Once the annotation has been verified by the curators, the image and annotation are made available online for the community to see.

The mtoc-explorer website and database are developed and maintained internally, using open-source General Public License, GPL compatible, or

similarly licensed software. The site and database are written in Python 2.6.5 with the Django 1.4.0 web framework using a PostgreSQL 8.4.13 database. Additional use is made of JQuery 1.5.1 on the site. Content is served using Apache 2.2.14 running on Linux/Ubuntu 10.04.4 . The website has been extensively tested on all popular operating systems and browsers.

### 2.4.4 The Morphological Diversity Index

We set out to quantitatively compare the morphological diversity observed in different MTOCs annotated in ***mtoc-explorer.org***. The aim was to able to answer the questions of the type: a) is organelle $X$ constrained? and b) is organelle $Y$ more or less constrained than organelle $Z$?. These questions posed two particular challenges. First is that we have no theoretical framework to estimate the expected diversity of 'unconstrained evolution'. The second is that the ontology has multiple forms of data (boolean, integer, unknown ...) which are not directly comparable. In order to overcome this we developed the Morphological Diversity Index (MoDI), which allows us to ask exactly these types of questions.

#### MoDI theory

The MoDI measures the fraction of observed phenotypes compared to the total number of possible phenotypes. For each component a binay matrix was created, in which each row is a term in the ontology and a value (for instance, fold-symmetry: 9). Each column is an image, and each entry is 1 or 0, depending on whether this image has been annotated with this term and value or not. The number of independent dimensions of variation of the observed phenotype is calculated using Principle Component Analysis (PCA), and finding the number of dimensions required to explain 95% of the variance. The number of dimensions of possible phenotypes is measured in a similar manner, but computed as the average number of dimensions of 1000 random permutations of the original data. A more detailed description of the MoDI is available as supplementary material.

### MoDI experiments

Although the MoDI is generic enough to measure "constraints" in any complex dataset, our intention was to use it to quantify morphological diversity. However the ***mtoc-explorer.org*** ontology has many components that are not strictly morphological in the sense that they do not capture structural attributes of the organelle (for example: position). Therefore we pruned various parts of the ontology before measuring the constraints in MTOCs. The vocabulary terms were filtered to include only structural attributes, removing any terms related to: POSITION IN THE CELL, ORIENTATION, to: MATURATION STAGE, NUMBER [3], and ASSOCIATED WITH.

## 2.4.5 Measuring Diversity

One of the challenges associated with studying diversity is finding ways to display large amounts of heterogeneous data. This is especially the case with the type of diversity we wish to capture, which has boolean and integer attributes which may also be unknown. Take for example the AXONEME which is described by:

RADIAL SPOKES = {YES, NO or UNKNOWN}
CENTRAL MICROTUBULES = {0, 1 or 2}
N-FOLD SYMMETRY = {1, 2 ... N}

In order to obtain an overview of a large part of the diversity in the three major components of the cilium: the axoneme, transition zone and basal body, we focussed on structures that are visible in cross-sections of these structures. Subsequently we filtered for all images annotated as SECTION → CROSS-SECTION for each ciliary component. For all leaf terms we counted the frequency of each annotation. Likewise we can do this calculation for all images limited to a particular taxonomic group.

## 2.4.6 Ancestrality vs. Convergent Evolution

In order to measure the probability of ancestrality vs. convergent evolution of cilia and centriole-based centrosomes, we first needed to obtain a list of all

[3] By NUMBER we refer only to those describing the "number" of organelles, not other numerical values.

species in the database with these structures. Next, we applied **Maximum Parsimony Landscapes** to this dataset.

### Cilia & centiole-based centrosomes

In order to determine the distribution of 'cilia' and 'centriole-based centrosomes' we queried ***mtoc-explorer.org*** for the lists of all species containing images annotated having either of these organelles. For the CILIUM the specific term queried was:

CELL → CILIUM/FLAGELLUM

The centriole-based centrosome is a slightly more complicated issue, as we wanted to ensure that we select only images of *bona fide* mitotic centrioles, and not Basal bodies in the cytoplasm. Therefore we used a more refined search query, selecting only images that show cells in either MITOSIS or MEIOSIS with a CENTRIOLE-BASED CENTROSOME. The following search criteria were used to identify species with centriole-based centrosomes:

CELL → CENTRIOLE-BASED CENTROSOME

AND

IMAGE → IMAGE METADATA CELL CYCLE STAGE → MEIOSIS OR MITOSIS

### Maximum Parsimony Landscapes

Calculating the probability of convergent evolution of cilia and centriole-based centrosomes is a task complicated by two specific factors: The lack of a proper evolution tree (*i.e.* with branch lengths) for Eukaryotes and the lack of a model of evolution (probabilities of gain and loss) of organelles. Hence, we developed "Maximum Parsimony Landscapes": a methods to calculate the probability of convergent evolution from a species' cladogram without any prior assumptions about the model of evolution.

The "Maximum Parsimony Landscape" is an extension of Sankoff parsimony (Sankoff, 1975), which is simply parsimony with a model of evolution (*i.e.* matrix of transition costs). Sankoff Parsimony can be used to calculate the cost of a given evolutionary scenario given: a species' cladogram, an ancestral state, observations in extant species and a cost model. The process of finding the cost of an evolutionary scenario involves finding the most parsimonious (*i.e.* cost

minimizing) combination of state transitions on a tree. For a binary trait (*i.e.* presence ($s = 1$) or absence ($s = 0$) of an organelle) and a given cost of gaining ($\alpha$) and cost of losing ($\beta$) an organelle, the probability of ancestral state $a = s$ is calculated as:

$$p\left(a = s; \alpha, \beta\right) = 1 - \frac{MP\left(\alpha, \beta, a = s\right))}{MP\left(\alpha, \beta, a = s\right)) + MP\left(\alpha, \beta, a \neq s\right))} \tag{2.1}$$

Where $MP\left(\alpha, \beta, a = s\right))$ is the cost of the most parsimonious solution given $\alpha$ $\beta$ and $a = s$ (and likewise for $a \neq s$). Put simply: To calculate the probability of ancestral state $a = 1$ calculate the (maximally parsimonious) cost associated with each ancestral state, and $p(a = 1)$ is (one minus) the cost of $a = 1$ divided by the sum of all costs.

In the case of organelle gain and loss, and if $\alpha$ and $\beta$ are known, equation can directly be used to calculate the probability of convergent evolution. If $p(a = 1)$ is large, this directly implies that the ancestor had the organelle, ruling out convergence. Conversely, a low value for $p(a = 1)$ directly implies that organelle appeared by convergence.

In dealing with the gain and loss of organelles we do not have any prior assumptions about the cost of gain or loss ($\alpha$ or $\beta$). Instead of making arbitrary assumptions, we chose to calculate the probability of each ancestral state for all possible values of $\alpha$ and $\beta$. This results in a "Maximum Parsimony Landscape": a landscape of $p(a = 1)$ values for all possible combinations of $\alpha$ and $\beta$. If all points on this landscape where $p(a = 1) < 0.5$ are taken as evidence for convergent evolution, the probability of convergent evolution without any prior assumptions for $\alpha$ and $\beta$ is given by:

$$p\left(\text{convergence}\right) = \frac{\sum_{\alpha=0}^{1} \sum_{\beta=0}^{1} p\left(s = x\right) \text{ if } p\left(s = x\right) < 0.5}{\sum_{\alpha=0}^{1} \sum_{\beta=0}^{1} p\left(s = x\right)}$$

We calculated "Maximum Parsimony Landscapes" for all species in mtoc-explorer.org for the annotations CILIUM and CENTRIOLE-BASED CENTROSOME as described in section 2.4.6 using an adapted version of the NCBI taxonomy (see section 2.4.1).

## 2.5 Supplementary Material

### 2.5.1 Morphological Diversity Index: Controls

One of the major problems with measuring morphological diversity in a collection of images annotated with an ontology is verifying that that the results are not artifacts. In the case of the MoDI, we deduce that there are two potential sources which could artificially alter the MoDI of an organelle (or organelle compartment): The number of images of that organelle, and the number of ontology terms that exist to describe it.

We took an empirical approach to ensure that the results in section 2.2.4 were not caused by these artifacts. For each organelle, we retrieved the data used to calculate the MoDI, and generated 100 random subsets of 1, 2, 3 ... images, and recalculated the MoDI. In this way it is possible to estimate the effect that the total number of images representing an organelle as on the MoDI.

For example: the central panel of Figure 2.11 shows the controls for the Transition Zone. There are 125 images of Transition Zones in the database, and the MoDI is calculated to 0.58. However, if we take random subsets of images, the MoDI starts to artificially inflate as we simulate having less than 20 images of Transition Zones. In the database there are 118 images of Axonemes. If we look at the estimated Transition Zone MoDI at 118 images, we see an average of 0.589 (thin gray line). However the real MoDI of the Axoneme (top panel) is 0.55. Therefore, the observed value of 0.55 cannot be due to the fact that there are less images of Axonemes than of Transition Zones. A similar method is used to verify that the number of ontology terms for an organelle (Figure 2.12) also does not greatly affect the MoDI.

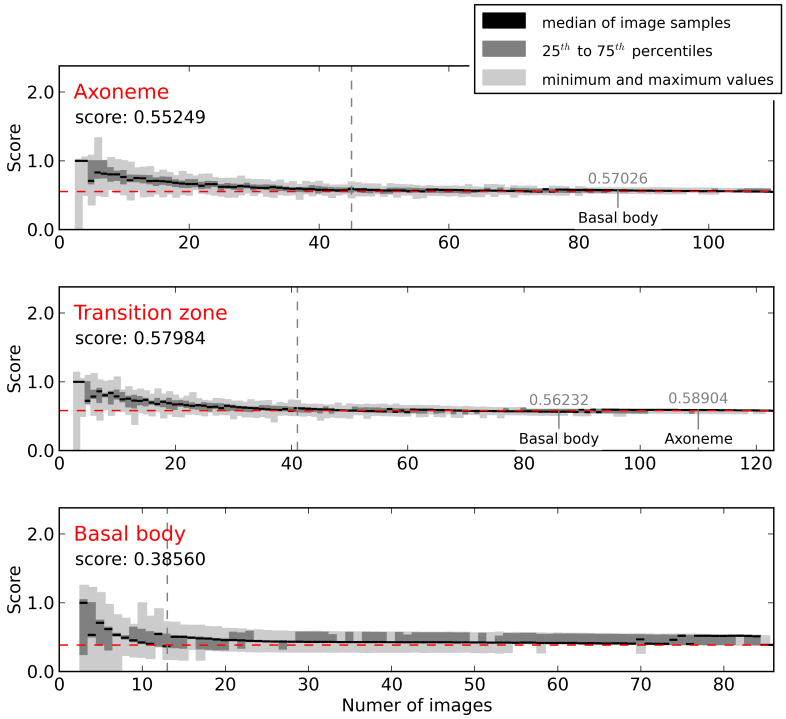



Figure 2.11: **Control for the MoDI: images** The effect of using randomly selected subsets of images to calculate the MoDI of Axonemes, Transition Zones and Basal Bodies (see text for explanation).

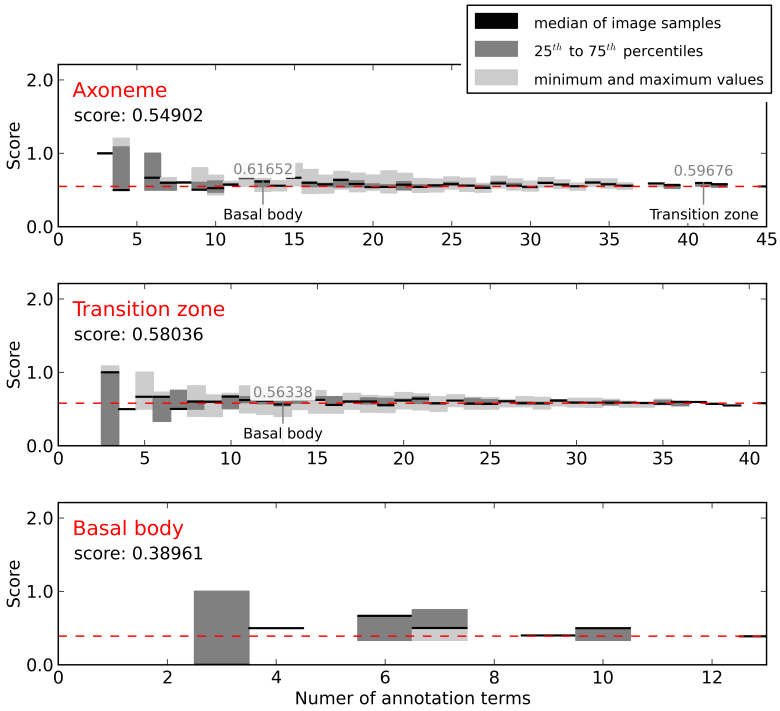



Figure 2.12: **Control for the MoDI: ontology terms** The effect of using randomly selected subsets of ontology terms to calculate the MoDI of Axonemes, Transition Zones and Basal Bodies (see text for explanation).

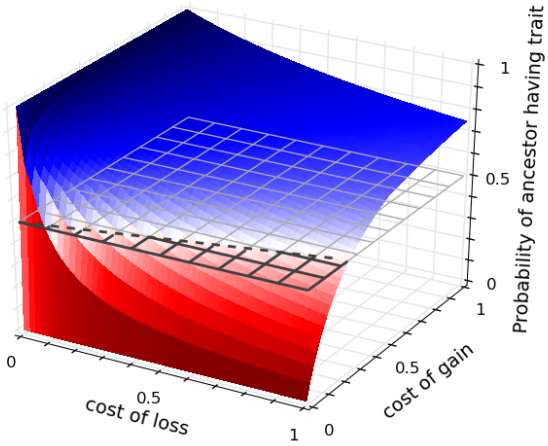



Figure 2.13: **Maximum Parsimony Landscape for the metazoan centriole-based centrosome** The $p$(convergence) for centriole-based centrosomes metazoa is 0.13.

### 2.5.2 Maximum Parsimony Landscapes: Controls

In order to validate using Maximum Parsimony Landscapes as a method to detect convergent evolution, we calculated the $p$(convergence) for the centriole-based centrosome in the branch of metazoa. It is known that the ancestor of metazoa had a centriole-based cilium (Bornens, 2012). The result (Figure 2.13) of $p$(convergence) of 0.13 (out of 1) gives strong support to the ancestrality of the centriole-based centrosome. This lends support to the use of Maximum Parsimony Landscapes to test for convergent evolution.

# Chapter 3

# Associating Genes to Organelles in Eukaryotes Using Phylogenetic Profiling


## Abstract

Recently there has been a growing interest in understanding the evolution of eukaryotes, and specifically at finding the origins of novel functions and novel organelles, and the genes involved with these. Phylogenetic profiling, a technique that uses similarity in protein presence/absence correlation across multiple species, has successfully been used to predict protein functions from known phenotype distributions in prokaryotes. Its use in eukaryotes until now has been quite limited; phylogenetic profiling has been shown to perform perform poorly in eukaryotes. However most studies in eukaryotes thus far have focussed on predicting protein function based on profile similarity to other proteins (not phenotypes). Its use to study eukaryotes, and specifically the evolution of organelles, has never been addressed. We benchmark different orthology prediction methods and profile similarity metrics using the eukaryotic cilium as a "model organelle", since both its phenotypic distribution as well as molecular components have been extensively characterized. Three orthology methods (Reciprocal Best Hits, InParanoid & OrthoMCL) and 3 profile similarity metrics (Hamming, Jaccard and Dollo Reduced Hamming) were tested for their ability to correctly predict cilium related proteins against the Sys-cilia dataset and CilDB proteomes. Although the ability to predict the full ciliary proteome is limited, phylogenetic profiling produces a small number of useful predictions: 50%

Positive Predictive Value for the top ranking 100 predictions. Surprisingly the quality of the predictions does not depend greatly on the orthology prediction method, nor the profile distance metric implemented. Instead the taxonomic range of species used to construct the profile is more important: including species from as many major eukaryotic lineages greatly improves the results. Lastly we use phylogenetic profiling to construct predictors that can be used to predict the presence or absence of cilia with almost 100% accuracy. Our results show that phylogenetic profiling is a viable approach to study the evolution of eukaryotic organelles. Although the ability of phylogenetic profiling to detect the full organellar proteomes is limited, it is a powerful technique for predicting candidate genes to characterize in the "wet lab". In this setting, the performance of "genotype-phenotype" profiling in eukaryotes is on par with the performance of phylogenetic profiling in prokaryotes.

## Publication

This chapter is currently being prepared as a manuscript for publication.

## Author's contributions

I was responsible for conducting the experiments for this chapter, as well as writing the text and creating the figures. This work and writing was done in close collabration with José Pereira-Leal and Mónica Bettencourt-Dias, and with support from Yoan Diekmann.

## 3.1 Introduction

The evolutionary origins of cellular organization has been a topic of great interest for generations, and the evolutionary transition from the Prokaryotic to the Eukaryotic cell plan dubbed one of the great transitions in evolution (Maynard Smith and Szathmáry, 1995). The technological advances that made it possible to sequence whole genomes cheaply and quickly made accessible gene repertoires for species throughout the tree of life, representing the full scope of extant cellular organization. To that extent, the term "evolutionary cell biology" has been used to define an emerging trend of intersecting evolutionary biology with molecular cell biology to understand the evolutionary drivers and mechanisms that underlie the evolution of cellular properties (Lynch et al., 2014; Brodsky et al., 2012).

Many insights have been gained by correlating phenotypes with gene repertoires, for example in discovering new genes associated with organelles (e.g. Avidor-Reiss et al. (2004) and Li et al. (2004)) or cellular differentiation pathways (e.g. Abecasis et al. (2013)). These are examples of the application of a comparative genomics approach termed "phylogenetic profiling" (Marcotte et al., 2000; Pellegrini et al., 1999; Date and Marcotte, 2003), based on the premise that functionally linked proteins co-evolve: if two proteins are functionally related, these proteins (or more precisely their orthologs) are more likely to co-occur or be absent across a set of genomes from different species. Inversely, if two proteins share similar phylogenetic profiles (presence/absence profiles across multiple species) they are more likely to be functionally related. If the function of only one of these proteins is known, this "guilt by association" principle (Aravind, 2000) serves as a prediction for the function of its evolutionary companion.

The mechanistic basis of phylogenetic profiling is that proteins that form complexes, interact, or are part of the same biological pathway tend to be functionally related, and are subject to a common (purifying) selective pressure. A consequence is that besides looking at the evolution of pairs or groups of proteins, we can also identify proteins with phylogenetic profiles similar to that

of a target function (or phenotype). In the context of phylogenetic profiling this is often referred to as "genotype–phenotype" profiling (Slonim et al., 2006), as opposed to "genotype–genotype" profiling.

Most successful applications and developments of phylogenetic profiling fall in the domain of genotype–genotype profiling in Prokaryotes. Several large scale benchmarks show that it is possible to predict protein localization (Marcotte et al., 2000), protein–protein interactions (PPIs) (Sun et al., 2005; Zhou et al., 2006) and proteins that form part of the same biological pathway (Pellegrini et al., 1999; Date and Marcotte, 2003; Enault et al., 2003). Prokaryotic genomes are also very suitable for genotype–phenotype profiling, and have been analyzed to predict proteins involved in endosporulation, gram negativity, motility and oxygen requirement (Slonim et al., 2006), pathogenicity (Huynen et al., 1997) and hyperthermophily (Makarova et al., 2003; Jim et al., 2004). Lastly, Jim et al. (2004) have used phylogenetic profiling to predict proteins related to two bacterial organelles: flagella and pili.

Detection of homologues for specific gene families or gene sets has been frequently used to infer the presence of an organelle in a given species. The most famous example is perhaps the demonstration that *Giardia lamblia* is not a Golgi-less early Eukaryote, but in fact a derived Eukaryote that had lost a morphologically distinct Golgi but still retained a gene complement indicative of the presence of a Golgi function. It is also frequently used to identify the molecular components associated with an organelle and to infer evolutionary pathways in organelles, supporting views as proposed by Dacks and Field for the evolution of the endomembrane system by duplication from an ancestral core (Dacks and Field, 2007a), or of a stepwise evolution with the addition of taxon specific components (Carvalho-Santos et al., 2010). However, we do not fully understand the limitations associated with correlating phenotype and genotype in cellular evolution, in other words, we have not benchmarked phylogenetic profiling in the study of organelle evolution. Phylogenetic profiling has seen limited success in Eukaryotes. One of the reasons is that the complex evolution of proteins in eukaryotes make orthology detection more difficult than in prokaryotes (Chen et al., 2007). Nevertheless, previous studies have shown that it is possible to predict protein complexes (Barker et al., 2007), PPIs (Kensche et al., 2008; Barker and Pagel, 2005; Ruano-Rubio et al., 2009),

shared biological pathways (Kensche et al., 2008; Ruano-Rubio et al., 2009) and common Gene Ontology annotations (Singh and Wall, 2008), demonstrating that the phylogenetic signal is strong enough for genotype–genotype predictions.

To our knowledge no one has addressed how well genotype–phenotype phylogenetic profiling performs in eukaryotes. Here, we assess the ability of phylogenetic profiling to detect genotype–phenotype correlations in eukaryotic organelles. In order to do this we require an organelle to use as a gold standard to benchmark and validate phylogenetic profiling. For our purposes such a gold standard should a) be present in most major eukaryotic branches, b) be phenotypically well annotated in a range of model and non-model species and c) have a well characterized and high confidence list of molecular components. The cilium is a microtubule based organelle present in all major eukaryotic lineages (Figure 3.1), that was present in the first eukaryote, and has been lost multiple independent times throughout evolution (see Carvalho-Santos et al. (2011) for a comprehensive review). It has served as a model organelle for morphological and evolutionary studies of cells for over one hundred years (Haimo and Rosenbaum, 1981), and its presence and absence has been thoroughly annotated in a wide range of model and non-model organisms. Its central role as a model organelle for cellular phenotyping is exemplified by **mtoc-explorer.org**, an online resource containing ultrastructural annotations of cilia in over 100 species. Its molecular components (of which a few hundred exist) have also been extensively studied and characterized in the lab and various high throughput studies. Its role as a model organelle is highlighted by the existence of Sys-cilia, a "gold standard of known ciliary components" (Dam et al., 2013) for benchmarking and validating high throughput and systems biology approaches. Likewise CilDB, a "knowledgebase for centrosomes and cilia" (Arnaiz et al., 2009) has become a central reference point for obtaining high throughput proteomics studies from different species.

In this chapter we first review different implementations of phylogenetic profiling and its application to predicting protein function from phenotype profiles. Next we benchmark various orthology detection methods and profile similarity metrics. We also examine the effect of choosing different species (both number of species as well as taxonomic distribution) on the performance of phylogenetic profiling. Finally, we present an approach able to predict whether

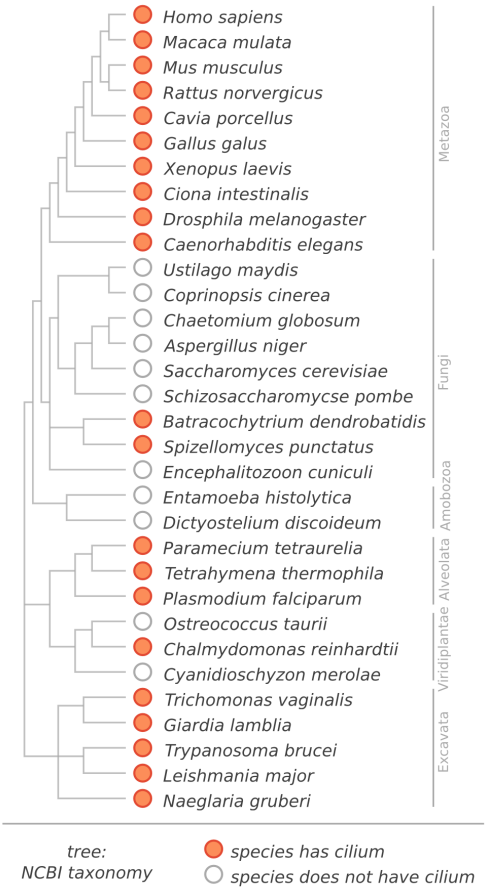



Figure 3.1: **The presence/absence profile of the eukaryotic cilium.** The eukaryotic cilium is present in all major eukaryotic lineages, is ancestral to the Last Eukaryotic Common Ancestor (LECA), and has been lost multiple independent times throughout evolution. The presence/absence profile of this organelle will be used as the target phenotype in this study. These 32 species will be used in this chapter to benchmark the performance of phylogenetic profiling eukaryotic of organelles. Unfortunately at the time of this writing we were unable to find any Rhizaria or Heterokonts for which both phenotype data as well as full genome sequences are available.

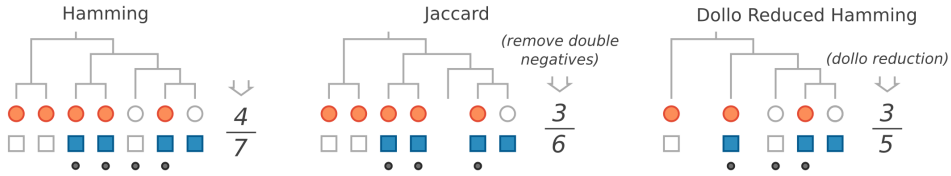



Figure 3.2: **An overview of phylogenetic profiling similarity metrics.** Examples of how the three profile similarity metrics investigated in this chapter are calculated. Dollo Reduced Hamming is the only phylogeny aware method.

or not a species is ciliated based on its genotype. In each of these sections we provide practical advice and "rules of thumb" for the use of phylogenetic profiling in eukaryotes.

### 3.1.1 A Primer on Phylogenetic Profiling

In its most basic form phylogenetic profiling is a comparison of two binary vectors. In genotype–genotype profiling each vector represents the presence/absence profile of a query protein's orthologs across different species. In genotype–phenotype profiling one vector represents the presence/absence profile of the target phenotype, and the other is the query protein's orthologs. The similarity between the two profiles is used to determine if two proteins are functionally related (in the case of genotype–genotype profiling) or if a protein participates in creating the phenotype (in the case of genotype–phenotype profiling). There are two technical factors that influence the effectiveness of phylogenetic profiling: a) The method used to detect orthologs between different species and b) the metric used to measure similarity between two profiles.

Accurate orthology detection is a crucial step in phylogenetic profiling: Incorrectly predicted or missing orthologs result in erroneous profiles that are detrimental for the quality of the results. However, orthology prediction is a challenging problem in eukaryotes, especially for large numbers of species where computational power limits the use of more accurate methods. We will be focussing on 3 methods that have been particularly useful in eukaryotes: Reciprocal Best Hits (RBH), InParanoid and OrthoMCL (Chen et al., 2007; Altenhoff and Dessimoz, 2009). RBH is simplest of the three, and assigns two proteins as orthologs if they are each other's highest scoring hits in an all-vs-all

pairwise BLAST between the full proteomes of two species. InParanoid (Remm et al., 2001) is based on RBH, and uses a self-vs-self BLAST of each proteome to filter out paralogs. OrthoMCL (Li et al., 2003) is also based on RBH, removes paralogous sequences, and subsequently uses Markov Clustering to identify robust groups of orthologs.

The second major factor affecting predictions made by phylogenetic profiling is the the metric used to measure similarity between profiles. We will be looking at 3 representatives of 2 types of methods: naive (co-occurrence based) and phylogeny aware (Ruano-Rubio et al., 2009; Kensche et al., 2008) (see Figure 3.2). The (normalized) Hamming similarity between two binary vectors is the fraction of positions in which the two vectors are the same. A variant of Hamming similarity is Jaccard similarity (Jaccard, 1912), in which only positions where one or both vectors are positive are counted. Neither of these methods take the species phylogeny into account, and multiple studies have shown improvements by using phylogeny aware methods both in prokaryotes (Cokus et al., 2007; Zhou et al., 2006) and eukaryotes (Barker and Pagel, 2005; Barker et al., 2007). Phylogeny aware methods compensate for the phylogenetic signal from closely related species, and emphasize independent gain or loss co-occurrences. We examine one phylogeny aware metric: Dollo Reduced Hamming, which is the (normalized) Hamming similarity between two vectors after the species tree has been collapsed to contain only monophyletic gain and/or loss co-occurrences (similar to the "Dollo-overall" method proposed in Barker *et al* (Barker et al., 2007)).

A handful of other methods have been proposed with the aim of increasing the efficacy of phylogenetic profiling, which we will not be testing in this chapter. The Wruns method (Cokus et al., 2007) was developed to take into account independent losses and gains. However as this approach is a simplified heuristic to take account phylogeny and is sensitive to the (arbitrary) ordering of species in the tree. Methods which use continuous values in the profile vectors (for instance BLAST score) have shown improvements in prokaryotes (Enault et al., 2003; Date and Marcotte, 2003). Other methods which have performed well in eukaryotes include those using group sizes by gene family (Ruano-Rubio et al., 2009) or domain composition (Lingner et al., 2010). However, as these methods do not work for individual proteins, they are not explored in this chapter.

Maximum Likelihood based methods have also been used for phylogenetic profiling, and greatly improve the predictive power in eukaryotes (Barker and Pagel, 2005; Barker et al., 2007). However Maximum Likelihood methods require an evolutionary tree with branch lengths, which does not exist for eukaryotes, and therefore cannot be used to study the evolution of eukaryotes as a whole.

## 3.2 Results and Discussion

### 3.2.1 Predicting the proteome of the eukaryotic cilium

One of the major challenges in systems biology is to predict the set of proteins required for the biogenesis and function of an organelle (its proteome). Given the presence/absence profile of an organelle across various species and the full genome of a reference species, phylogenetic profiling should be able to identify which proteins are required for that phenotype. We set out to benchmark the 3 different orthology detection methods and 3 similarity metrics (Figure 3.2) for their ability to detect the full cilium proteome using Humans as a reference species. See materials and methods (section 3.4.3) and (Figure 3.9) for a description of how we validate predictions.

In order to assess the efficacy of phylogenetic profiling organelles in eukaryotes, we benchmarked predictions against SysCilia, a manually curated database of 303 Human proteins known to be required for cilia biogenesis and function (Dam et al., 2013). We used 3 orthology prediction methods to predict orthologs of all Human proteins across 32 species (Figure 3.1). Subsequently we used each of the 3 profile similarity metrics, and measured how well the combination of orthology detection method and profile similarity metric was able to retrieve proteins in the SysCilia dataset.

The ability to capture an organelles proteome is best summarized by the Receiver Operator Characteristic (ROC) curve: The fraction of ciliary proteins retrieved (sensitivity) vs. the fraction of correctly rejected non-ciliary proteins (specificity) at different values of similarity cutoff. The Area Under the Curve (AUC) provides a measure for the quality of a predictor that is independent of a specific parameter value, and captures the tradeoff between identifying as many proteins as possible correctly whilst avoiding false positive predictions.

Another measure of interest is the sensitivity at specificity = 0.95 (Sens 95): the fraction of proteins identified (vs. all proteins in the cilium proteome) whilst correctly rejecting 95% of non-cilium related proteins.

No combination of orthology detection method or similarity metric is substantially better than others in predicting the cilium proteome (Figure 3.3). RBH performs marginally better than other orthology detection methods, and Hamming similarity outperforms Dollo Reduced Hamming and Jaccard. However, none of the methods perform well at detecting all 303 proteins required for cilia formation and function, reaching full detection (sensitivity = 1) at very low specificity values.

Until now phylogenetic profiling in eukaryotes has focussed on genotype–genotype predictions, and has proven to perform poorly. In a study in Fungi Kensche *et al* (Kensche et al., 2008) obtained AUC values for PPIs and shared biological pathways of approximately 0.55 and 0.60 respectively, whereas here we observe AUCs as high as 0.77 In a pan-eukaryotic study using 53 species Ruano-Rubio (Ruano-Rubio et al., 2009) obtained Sens 95 values below 0.30 (not including gene group size based methods) in PPI's and shared biological pathways, whereas we observed values around 0.40. Using phylogenetic profiling to predict the (bacterial) flagellar proteome, Jim et al. (2004) obtained Sens 95 values of (approximately) 0.55. The results obtained in this analysis are show an improvement when compared to genotype–genotype profiling in eukaryotes, although are not as good as genotype–phenotype profiling in prokaryotes.

The fact that genotype–phenotype prediction in eukaryotes performs substantially better than phylogenetic profiling PPIs and shared biological pathways likely comes from two sources. The first is related to the difference between genotype–genotype and genotype–phenotype profiling in a setting where orthology detection is problematic: The overall score in genotype–genotype profiling is based on an all-vs-all profile comparison between all proteins from the reference organism's proteome, and each erroneous profile will contribute multiple times to lowering the score. In genotype–phenotype profiling, each erroneous profile contributes only once to lowering the score. The second source is likely due to the fact that the cilium profile (Figure 3.1) is optimal for phylogenetic profiling. PPI and shared biological pathways may have profiles that are either all absent, all present, or devoid of phylogenetic signal, thus hindering the power of

phylogenetic profiling. It has already been shown that removing proteins with few orthologs in a profile can greatly enhance the performance of phylogenetic profiling (Lin et al., 2013; Zhou et al., 2006). Likewise Jim et al. (2004) also found that genotype–phenotype profiling bacterial organelles works best if the target phenotype is neither rare nor common. This topic is further explored in section 3.2.5 of the results.

Retrieving the full proteome of an organelle remains a challenge in genotype–phenotype profiling in eukaryotes. It is possible to retrieve approximately 40% of the cilium proteome (120 proteins) at a specificity rate of 0.95, although this set will also contain over 1000 (5% of the non-ciliary Human genome) false positives as well. However, it is worthwhile to note that no improvements are obtained from using methods more complicated than Hamming similarity and InParanoid (or RBH).

### 3.2.2 Predicting candidate genes based on known phenotypic distributions

One of the applications of phylogenetic profiling (and indeed comparative genomics in general) is predicting (new) proteins associated with a particular phenotype. These predictions can for example be taken to the "wet lab" for experimental validation. We examined the predictive power and top ranking predictions for each of the orthology detection methods and profile similarity measure from the previous section.

In a setting in which we are testing for candidate genes, there are two major values of importance: the Positive Predictive Value (PPV, the fraction of predictions that are correct) and the total number of predictions. The number of missed predictions (false negatives) is of less importance, and the focus is on obtaining a high number of predictions with a high PPV.

Figure 3.3 shows the PPVs for the top ranked predictions (up to 200) for each combination of orthology detection method and similarity metric. The major differences result from the choice of orthology detection method: InParanoid produces few yet high confidence predictions, OrthoMCL produces many but low confidence predictions, and RBH inhabits the Goldilocks zone in between.

In Figure 3.4 we show the PPV and number of predictions for the most stringent similarity cutoffs (above 0.80). The general trends are that Hamming

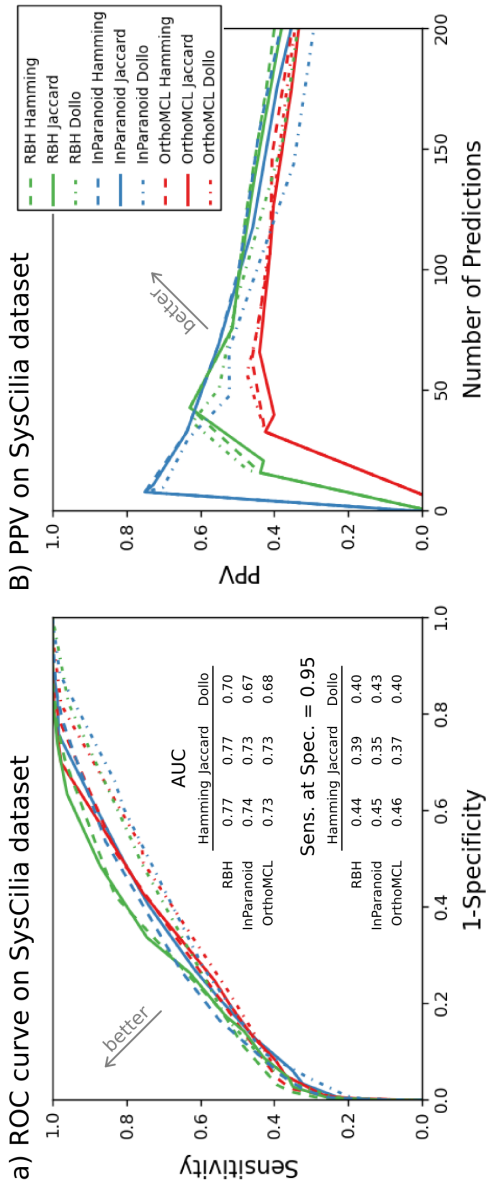


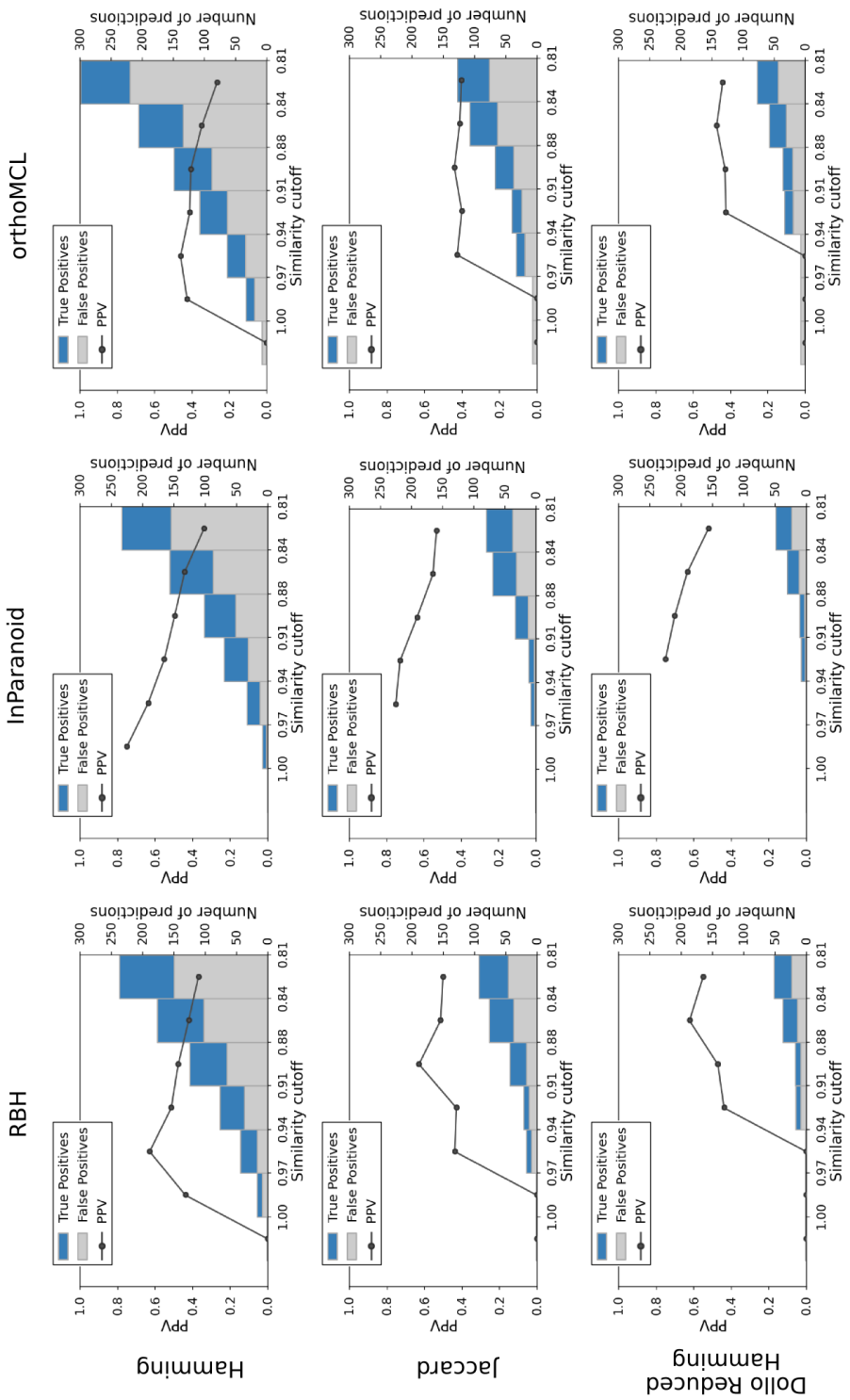

Figure 3.3: **ROC curve and PPV for predicting Human proteins contained in the SysCilia dataset.** The SysCilia dataset is a manually curated dataset of 303 Human proteins required for cilia, which we used to benchmark 3 orthology detection methods and 3 profile similarity metrics. A) Most combinations of methods perform similarly for retrieving the full set of 303 proteins, although Dollo Reduced Hamming similarity performs notably worse than Hamming and Jaccard. B) The main differences in PPV are due to different orthology detection methods: OrthoMCL correctly predicts most proteins, but at low confidence values. InParanoid predicts the least, although at high confidence values.

similarity results in more predictions than either Dollo Reduced Hamming or Jaccard similarity, whilst not sacrificing PPV. As suggested by Figure 3.3, InParanoid is the prefered orthology detection method, showing high PPV values at high confidence levels without sacrificing the total number of predictions. Using this combination of methods the PPV of the top 70 predictions is 0.55, which is higher than those obtained for genotype–phenotype profiling of the (bacterial) flagellum (0.4 for the top 60 predictions) (Jim et al., 2004). The PPV is also much higher than that obtained in a 3 species comparative genomics study of the eukaryotic cilium, in which only 56 of 688 predictions were verified ciliary components (Li et al., 2004).

The fact that that using Hamming similarity and InParanoid results in the best PPV is surprising: each of these are conceptually simpler and computationally cheaper than some of the other methods tested. The possible reason that OrthoMCL is outperformed is that OrthoMCL is less sensitive — but more specific — than either RBH or InParanoid (Chen et al., 2007), and thus may miss many orthology predictions resulting in incorrect orthology profiles. Likewise, Dollo Reduction improves predictions for genotype–genotype profiling in eukaryotes (Barker and Pagel, 2005; Barker et al., 2007), but decreases performance in our benchmarks. This suggests that evidence from multiple species, even if closely related, strengthens phylogenetic profiling predictions. All in all, it appears that the phylogenetic profiling is able to predict small number of proteins with very high confidence, which has already been observed in eukaryotes and prokaryotes. When using phylogenetic profiling as a hypothesis generation tool for wet lab experiments, we suggest using InParanoid for orthology detection and Hamming similarity. Although the total number of predictions may be low, a PPV of 0.50 (at 100 predictions) directly implies that 50% of the predictions taken to the "wet lab" will turn out to be functionally related to the phenotype under investigation.

### 3.2.3 Consistency between distance metrics

We have examined 3 different similarity metrics, which fall into two distinct types: Naive (Hamming and Jaccard) and phylogeny aware (Dollo Reduced Hamming). Although the previous sections show that Hamming outperforms Jaccard and Dollo Reduced Hamming in terms of PPV, it is possible that the


RBH
InParanoid
orthoMCL
Hamming
Jaccard
Dollo Reduced Hamming
True Positives
False Positives
PPV
Number of predictions
Similarity cutoff
1.00
0.97
0.94
0.91
0.88
0.84
0.81

these two similarity metrics predict other proteins correctly.

In order to verify if different similarity metrics predict different sets of proteins, we compared the top ranking predictions of each. We measured the overlap between each of the similarity metrics on the SysCilia dataset with the similarity cutoff set to the lowest value for which at least 100 proteins were predicted. The results are only shown for InParanoid orthology predictions, although similar results were observed with RBH and OrthoMCL.

Figure 3.5a shows the complete overlap between proteins predicted by each method, and 5b shows only the true positives shared between each method. This is summarized in 3.5c, which shows the PPV for each method and combination of methods. All of the top 101 predictions using Hamming are also predicted using either Dollo Reduced Hamming or Jaccard, and usually both. The same is true when only true positives are counted 3.5b, and almost all true positives predicted using Dollo Reduced Hamming and Jaccard are also predicted using Hamming similarity. As a result, using Hamming similarity (or any combination of methods that include Hamming similarity) results in the highest PPVs. Moreover, when predictions are used from the combination of all three metrics, there is no substantial improvement in PPV compared to using Hamming similarity alone.

In conclusion, there is little benefit associated to using different distance metrics other than Hamming: neither Dollo Reduced Hamming nor Jaccard predict proteins not predicted using Hamming, although filtering those predicted using the Hamming distance with Jaccard, may lead to a slight improvement in predictions.

Figure 3.4 *(previous page)*: **Positive Predictive Value for different orthology detection methods and profile similarity measured on the SysCilia dataset.** Positive Predictive Value (left axis) and total number of predictions (right axis) are shown for each combination of orthology detection method and profile similarity metric for all predictions with a similarity above 0.80. The highest number of predictions with a high PPV are obtained using Hamming similarity, and in general InParanoid and RBH result in better predictions than OrthoMCL.

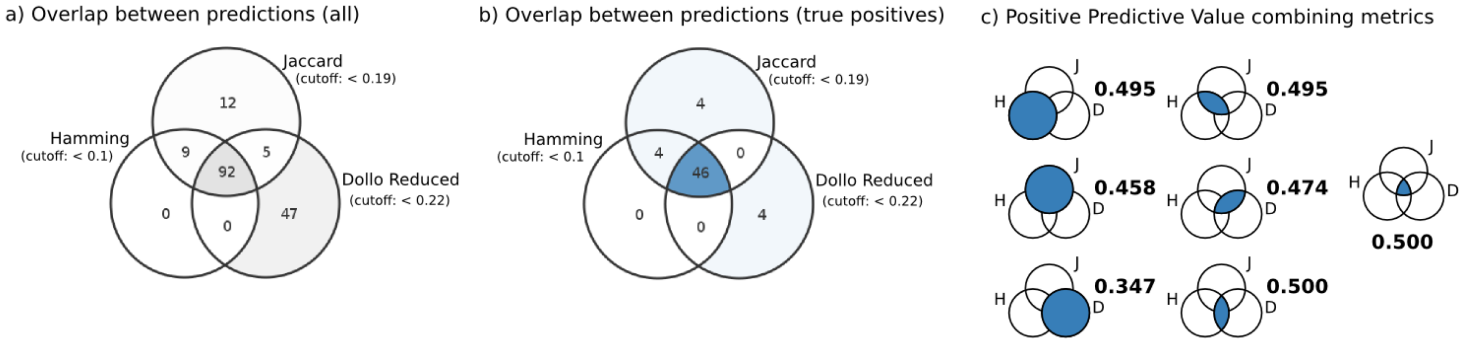



Figure 3.5: **Prediction consistency between different profile similarity metrics.** We examined the overlap in predictions for Hamming, Jaccard and Dollo Reduced Hamming similarity measures. **a)** The consistency between predictions without taking into account false and true positives. All top ranking predictions using Hamming similarity are also predicted by either Jaccard or Dollo Reduced Hamming, and usually predicted by both. **b)** The consistency between predictions limited to true positive predictions. The three different similarity measures almost always predict the same proteins correctly. **c)** The PPV values obtained when using one or a combination of predictions generated with different similarity measures. Any combination including either Hamming or Jaccard similarity achieves a PPV of near 0.50. The large number of false positive predictions produced by Dollo (compare Figure A and B) result in Dollo Reduced Hamming similarity to perform substantially worse.

### 3.2.4 Profiling non-model organsims

In the context of evolutionary studies, it is often interesting to study model organisms in different branches of the eukaryotic tree (and cilia are no exception to this (Holst and Wiemer, 2010)). We wanted to quantify how well phylogenetic profiling works in different branches of the eukaryotic kingdom.

As there are no gold standards for cilium related proteins in non model organisms, and we did not want to limit our search to those for which orthologs of the SysCilia dataset can be detected, we chose to assess the efficacy of phylogenetic profiling in 4 different species using results from high throughput proteomics. Proteomics data are available from CilDB for a vast number of species (Arnaiz et al., 2009). We selected 4 species: *Homo sapiens* (mammals), *Chlamydomonas reinhadtii* (plants), *Tetrahymena thermophila* (ciliates) and *Trypanosoma bruceii* (excacates). The result is a set of 825 mammal, 912 cilate, 269 plant and 500 excavate proteins (2506 total). We collected the ciliary proteomes of each of these species, and examined the ability of different orthology detection methods and profile similarity metrics to retrieve them.

The overall result (Figure 3.6) is similar to that observed using the SysCilia dataset: Dollo Reduction is substantially worse than Hamming and Jaccard in retrieving proteomes (Figure 3.6). OrthoMCL performs marginally better in these non classical model organisms. Again, the major difference in PPV lies in orthology detection method, and again InParanoid results in the highest PPV values, whereas OrthoMCL results in the highest number of predictions (Figure 3.6b). Figure 3.6c shows the PPV for top ranking predictions on a per-species basis. The predictions for *H. sapiens* and *C. reinhardtii* are decent for small number of proteins and high confidence predictions. However, for all non-human species, the overall efficacy of phylogenetic profiling appears to be limited.

The same general rules of thumb for genotype–phenotype profiling in eukaryotes stated earlier also holds for non-model organisms: Inparanoid (or RBH) is the most suitable orthology detection method, and Hamming (or Jaccard) outperform Dollo Reduced Hamming. However, the results suggest that phylogenetic profiling may also be challenging in certain non-model organisms.

### 3.2.5 Species selection

One of the major factors affecting the predictions generated by phylogenetic profiling is the set of species used. Various previous studies (most notably Sun et al. (2005) in prokaryotes) have found that using more species and a broader taxonomic range increases predictive power. Also, although never explicitly tested, one can hypothesise that the number of independent gains/losses of an organelle could strongly influence the quality of phylogenetic profiling predictions.

We investigated the effect of species selection on phylogenetic profiling using the SysCilia dataset as a benchmark. We generated 100 random selections of 1, 2, 3... etc. species, and for each subset of species checked the effect of a) number of species, b) number of major eukaryotic groups (Baldauf, 2003a), see also Figure 3.1) and c) number of independent losses of cilia (in the final tree after species selection). InParanoid orthology detection and Hamming similarity were used for all calculations.

Figure 3.7 shows the ability to retrieve the SysCilia dataset using randomly selected subsets of species. In all cases, increasing the number of species, taxonomic range and number of loss events increase the AUC. The effect of adding new species is strongest when the number of species, major eukaryotic groups, or loss events is low. Figure 3.7 shows PPVs obtained when the Hamming similarity cutoff is set at 0.9. Adding extra species to the analysis is important for low numbers of species, however as the total number of species increases, the contribution of each added species decreases. The same is not observed for total number of major taxonomic groups nor total number of organelle losses: including more species from distant lineages and including more losses has a large effect on the quality of the results. Note that the 'steps' observed using Hamming similarity with different number of species is due to the fact that there are only a discrete number of possible values for Hamming distance between two binary profiles of a finite length. Setting a fixed cutoff for predictions (in this case 0.9) means that different sets of discrete Hamming similarity are selected as positive predictions. Transitioning from (for example) 9 to 10 species means taking into account Hamming similarities of only 1, and of 1 and 0.9 respectively, thereby decreasing the overall PPV.

The general trends found for PP in prokaryotes (Sun et al., 2005) also

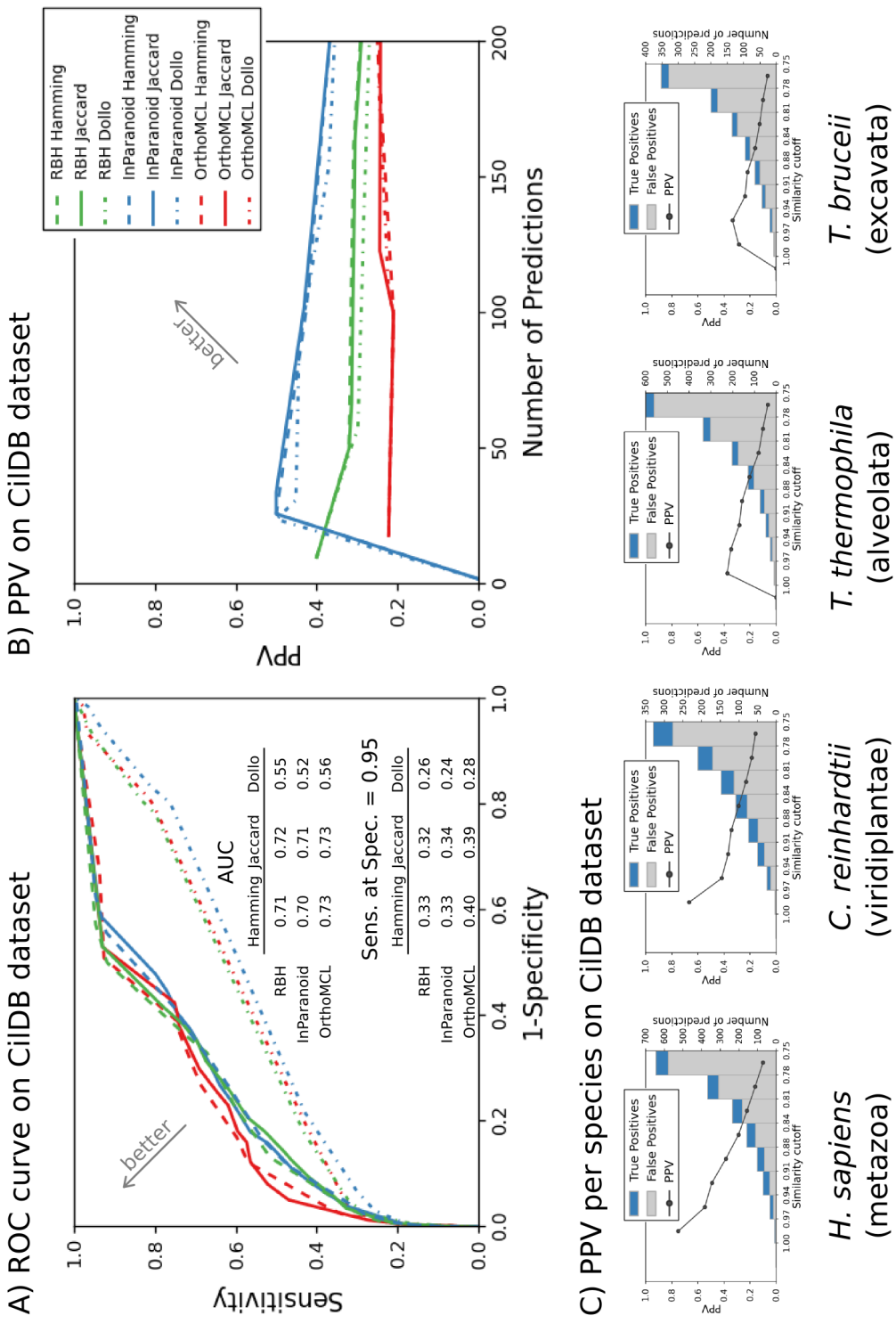
holds for organelle profiling in eukaryotes: Increasing the number of species and taxon range are important contributing factors to the performance of PP. If we measure the ability to predict full organelle proteomes using the AUC, Figure 3.7 suggests that at least a handful of species should be included, although beyond this the contribution of each added species decreases rapidly. The results the for predicting candidate genes (measured by PPV), however, suggest that the taxonomic distribution of species is the most important factor: Whereas the PPV levels off for increasing number of species and loss events, the biggest increase in PPV is observed when species from all major eukaryotic groups are included.

As a general guideline, these results suggest that simply adding more species alone is not enough, and that emphasis should be placed on selecting taxonomically distant species from as many major branches of the eukaryotic tree as possible. Likewise at least a few loss events should be included in the final set of species. In practice, this will limit the phenotypes that can be studied using phylogenetic profiling: Any phenotype which is monophyletic or only present in a closely related groups of species may prove challenging or even impossible.

### 3.2.6 Predicting phenotype from genotype

One of the major outstanding challenges in biology is to predict the phenotype of an organism based on its genome (Lingner et al., 2010), especially in the emerging field of evolutionary cell biology. In practice, when trying to determine if a species has a particular organelle, the presence or absence of a handful of well conserved proteins is used. For example, the presence of peroxisomes can be predicted based on the presence of 8 conserved genes (Schlüter et al., 2009). However, such molecular markers are not known for all organelles. It is even possible that proteins which are known to be required for an organelle's function cannot be used due to orthology detection problems, as is the case with the eukaryotic cilium (Carvalho-Santos et al., 2010).

We set out to determine if phylogenetic profiling can be used to select proteins which can be used as accurate predictors for the absence or presence of cilia. We generated 100 random subsets of 1, 2, 3 species (as in section 5), and used these as a "training set" to select proteins with phylogenetic profiles


A) ROC curve on CilDB dataset
Sensitivity
1-Specificity
better
AUC
Hamming Jaccard Dollo
RBH 0.71 0.72 0.55
InParanoid 0.70 0.71 0.52
OrthoMCL 0.73 0.73 0.56
Sens. at Spec. = 0.95
Hamming Jaccard Dollo
RBH 0.33 0.32 0.26
InParanoid 0.33 0.34 0.24
OrthoMCL 0.40 0.39 0.28
B) PPV on CilDB dataset
PPV
Number of Predictions
better
RBH Hamming
RBH Jaccard
RBH Dollo
InParanoid Hamming
InParanoid Jaccard
InParanoid Dollo
OrthoMCL Hamming
OrthoMCL Jaccard
OrthoMCL Dollo
C) PPV per species on CilDB dataset
True Positives
False Positives
PPV
Similarity cutoff
Number of predictions
H. sapiens (metazoa)
C. reinhardtii (viridiplantae)
T. thermophila (alveolata)
T. bruceii (excavata)

similar to that of the cilium. These proteins were used to predict the presence (or absence) of cilia in the remaining 31, 30, 29, ... species (the "test set"). For each subset of species we selected proteins with similarity of 0.85 to the cilium profile from the "training set". Species in the "test set" were predicted to be ciliated if they had 50% of the proteins selected from the "training set". We tested multiple different values for these cutoffs, and determined that these were the optimal values for this setting.

The overall ability to predict the presence of the cilium based on genotype is very high, reaching near 100% accuracy (Figure 3.8). Once again increasing the number of species, especially the taxonomic range and number of losses is important. Note that the apparently high accuracy of 0.55 observed for low number of species is an artifact of the naive classifier: if all species are predicted to have cilium, 55% of these predictions will be correct.

Predicting phenotype from genotype is not only possible, but a comparatively simple task when techniques from phylogenetic profiling are used. The main reason for this is that phylogenetic profiling will select proteins to use as phenotype predictors that a) correlate well with the phenotype and b) behave well in orthology prediction. This finding is very promising for the field of evolutionary cell biology: we now have the potential to make confident inferences about the organellar composition of a species based solely on its genome.

## 3.3 Conclusion

We showed that phylogenetic profiling of organelles in eukaryotes is a useful but limited predictor of organelles proteome, but performs quite well as a predictor

Figure 3.6 *(previous page)*: **Phylogenetic profiling non-classical model organisms with proteomics data from CilDB.** We benchmarked phylogenetic profiling using proteins obtained from proteomics experiments across 4 different branches of the eukaryotic kingdom: Mammals (*H. sapiens*), plants (*C. reinhadtii*), ciliates (*T. thermophila*) and excavates (*T. brucei*). A) ROC curve and Positive Predictive Value (B) summarizing the ability to predict whole proteomes across all 4 species for different orthology detection methods and similarity metrics. The predictions obtained using InParanoid and Hamming similarity give the best results, as with the SysCilia dataset (Figures 3.3 & 3.4). C) Positive Predictive Value of the top ranking proteins for each species using InParanoid for orthology detection and Hamming similarity.

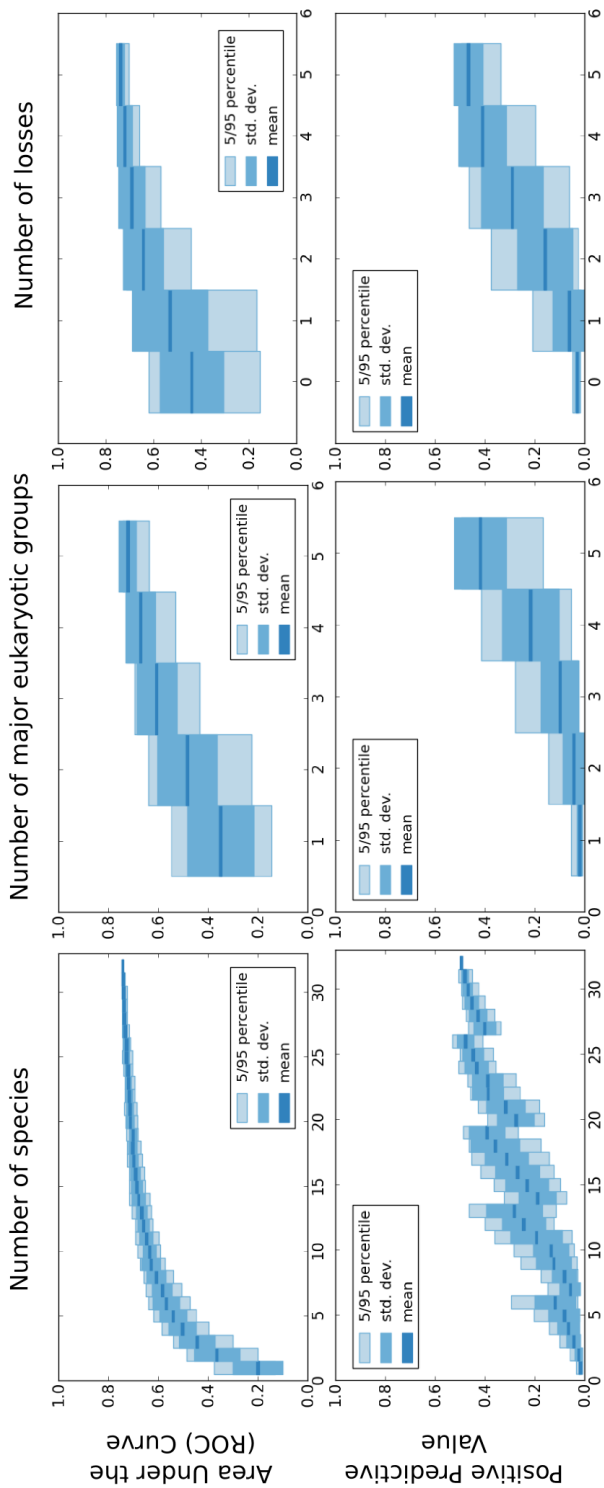



Figure 3.7: **The effect of species selection benchmarked against the SysCilia dataset.** The AUC and PPV obtained when random subsets of species are selected from the original dataset and validated with the SysCilia dataset. InParanoid was used for orthology detection, Hamming distances for phylogenetic profile similarity, and Positive Predictive Value was tested using a cutoff of 0.1. For predicting whole proteomes, the AUC (top) shows that increasing the number of species as well as major taxonomic groups (see Figure 3.1) and instances of loss are all important factors. For predicting candidate genes with high confidence, the PPV (bottom) suggests that increasing the number of species is effective, although better predictions are obtained from using species with a broad taxonomic range. Note: the "steps" observed in PPV for varying numbers of species are an artifact of the limited number of possible Hamming distances for a given number of species (see text for more details).

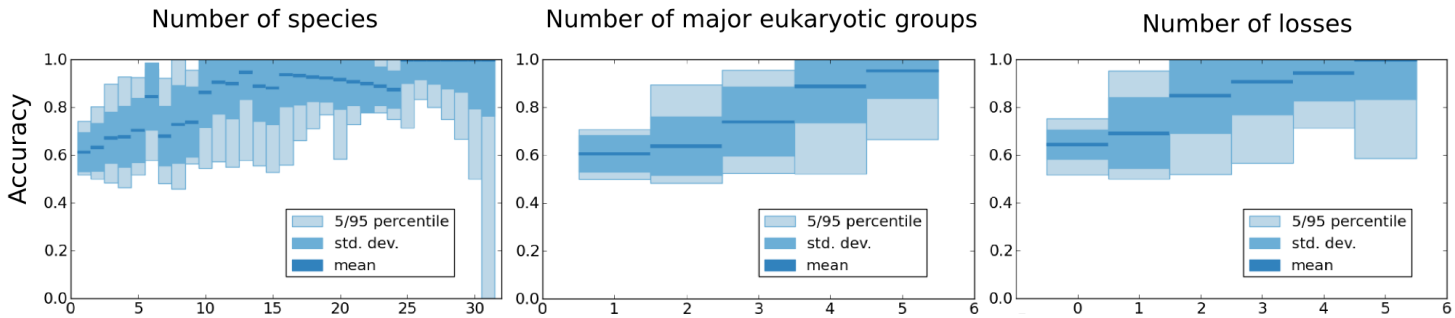



Figure 3.8: **Predicting the presence or absence of cilia using genome content.** We randomly selected subsets of species from the original dataset, selected all proteins with a Hamming similarity above 0.85, and used these to predict the presence or absence of cilia in the non-selected species. A species was predicted as having a cilium if 50% of the proteins selected had an ortholog (via InParanoid) in that species. Using this method its is possible to correctly predict the presence or absence of cilia with near 100% accuracy. As is the case with genotype–genotype predictions, increasing the number of species is important, as is selecting species from diverse taxa representing multiple independent losses of the cilium.

for a low number of highly specific proteins. Competing methodologies fare equally well and rather than the number of species, the nature of species chosen appears to be most relevant factor in improve overall efficacy.

We approached the usefulness of phylogenetic profiling eukaryotes for genotype–phenotype predictions. Our results show that there is a big tradeoff between sensitivity and specificity. The best positive predictive value we observed was of 50%, which means that in every two predictions of the nature "gene X is part of organelle Y in species Z", one is correct. This suggests that this is a good approach to identify candidate genes for further testing in the laboratory. However, for this level specificity very few genes are predicted to be associated to the organelle (typically tens, up to a few hundred). By accepting more false positive predictions, for example setting specificity at 95%, we can predict up to 40% of the organelle proteome, but with a positive predictive value of about 12

To our surprise, we found that the set of species used in the analysis is more important than using complex orthology detection methods or similarity measures. The use of InParanoid to detect orthologs and Hamming similarity to compare profiles almost always results in the highest quality predictions, even if only marginally better than the other approaches tested. More emphasis should instead be placed on selecting species from different major eukaryotic lineages, preferably representing multiple independent phenotypic states, which in this chapter represents cilium loss events.

Finally, phylogenetic profiling can be used to construct sets of proteins which can be used to accurately predict phenotypes in a newly sequenced organism. This finding is very reassuring for the field of evolutionary cell biology: we can make confident inferences about the organellar composition of a species based solely on its genome.

Our expectation is that rather than sequencing more genomes, phenotypic characterisation of sequenced species is most likely to improve the performance of phylogenetic profiling in the study of organellar evolution. It is unclear to us whether the inability to predict the full proteome of an organelle by mapping orthologues is a result of technical artifacts, i.e. low sensitivity/specificity of the orthology detection methods, or instead it represents taxon- and species-specificity of organellar components. The fact that different methods tested

here, know to have different sensitivities, have similar performance, suggests that the latter may be the dominant reason. In our experience, when we studied the evolution of the assembly pathways of the animal centriole/basal body, where we invested a significant effort in very sensitive methods and manual data analysis beyond automated orthology mapping, we concluded that regulatory components tended to be animal-specific (Carvalho-Santos et al., 2011). While our view is that the limitations of phylogenetic profiling reflect biology rather than artifact, this still needs further investigation.

Our results show that genotype–phenotype phylogenetic profiling is a viable approach to study the evolution of organelles, or at least, the eukaryotic cilium. Although the ability of phylogenetic profiling to detect the full organellar proteomes is limited, it is a powerful technique for predicting candidate genes to characterize in the "wet lab". In this setting, the performance of genotype–phenotype profiling in eukaryotes is on par with the performance of phylogenetic profiling in prokaryotes. Phylogenetic profiling has existed for over 25 years, and despite many successes, has never been used to study the evolution of eukaryotic cells. We feel that this is a missed opportunity, and hope that this study informs the further application of phylogenetic profiling in the study of eukaryotic cell evolution.

## 3.4 Materials and Methods

### 3.4.1 Sequence and Phenotype databases

We selected the 32 species (Figure 3.1) based on the availability of both sequence and phenotype annotations. The full predicted proteomes were obtained from Superfamily (Wilson et al., 2009b) version 1.75. Phenotype annotations were obtained from ***mtoc-explorer.org***, which contains annotated EM images of cilia for many species covering all major eukaryotic branches.

### 3.4.2 Orthology Detection

Since no existing databases contained pairwise orthology predictions for the 32 species we selected for this analysis, we computed RBH, InParanoid (Remm et al., 2001) and OrthoMCL (Fischer et al., 2011) on in house equipment. RBH

were calculated using BLAST+ (with default parameter values) InParanoid orthologs were calculated using the "inparanoid" program with default values as made available by the authors. OrthoMCL was implemented using the OrthoMCL pipeline made available on GitHub (pipeline, 2015)

### 3.4.3 Validation datasets

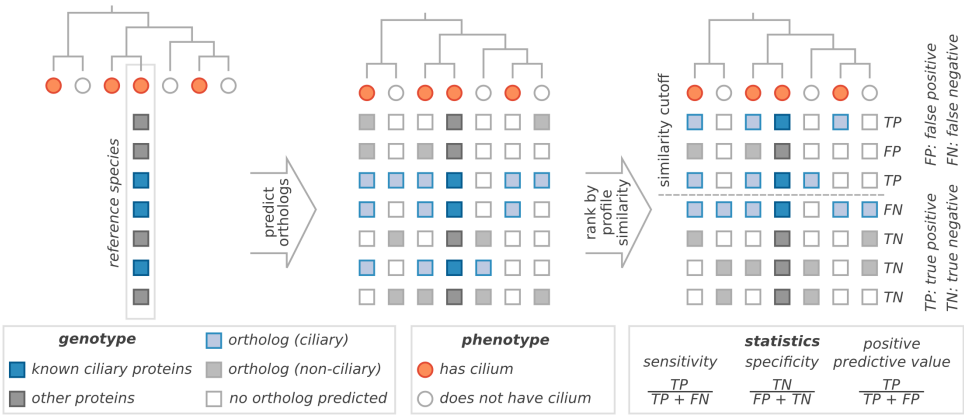



Figure 3.9: **Validation of genotype–phenotype phylogenetic profiling.** Genotype–phenotype profiling consists of 2 major steps: 1) Detecting orthologs for all proteins in a reference species and 2) Computing the similarity of each orthology profile to the target phenotype profile. Validation is done using a set of proteins known to be involved in the phenotype, and verifying how many of these are correctly identified.

Proteins required for cilia formation and function in Humans were obtained from SysCilia (Dam et al., 2013). Proteomic data was downloaded from CilDB (Arnaiz et al., 2009) for the following studies: *Tetrahymena thermophila* (Smith et al., 2005), *Homo sapiens* (Ostrowski, 2002), *Trypanospma bruceii* (Broadhead et al., 2006) and *Chamydomonas reinhardtii* (Keller et al., 2005; Pazour et al., 2005) at 'low', 'medium' and 'high' confidences.

### 3.4.4 Statistics

The following statistics were used to benchmark and validate the predictions generated using phylogenetic profiling:

True positives (TP) is the total number of proteins predicted correctly, from the reference set of proteins (i.e. in SysCilia or one of the CilDB datasets), and

false positives (FP) is the total number of protein incorrectly predicted to be associated with the phenotype. True negatives are proteins that are not part of the reference set that fall below the cutoff threshold (i.e. are not predicted), and false negatives are proteins that are not predicted, but are part of the reference set.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

$$\text{Positive Predictive Value} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

### 3.4.5 Predicting the presence and absence of cilia

To test the ability to predict the presence or absence of cilia, we created random subsets of 1, 2, 3...etc species, and selected the set proteins predicted by phylogenetic profiling at a cutoff of 0.85. Subsequently we checked the presence of these proteins in the remaining set of species, and if 50% or more of them were present in a species, it was predicted to be ciliated. Performance was measured as the total number of correct predictions of phenotype divided by the total number of species predicted from:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

# Chapter 4

# The Evolution of Rab GTPases


## Abstract

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.

## Publication

*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's contribution

Most of the this chapter was the work of Dr. Yoan Diekmann conducted as part of his PhD at Computational Genomics Laboratory at the Instituto Gulebenkian de Ciência. Dr. Diekmann created a pipeline to classify Rabs in to families, and ran this pipeline on 247 eukaryotic genomes. My contribution to this project was in developing and setting up *RabDB.org*, a web resource which made the data and the pipeline presented in the paper available to the public. On *RabDB.org* users can browse the Rab family assignments across the eukaryotic kingdom, and also submit protein sequences to be classified by the *Rabifier*.

## 4.1 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 (Aridor and Hannan, 2000; Bon, 2002; Seabra et al., 2002; Mitra et al., 2011), as well as cancer (Agarwal et al., 2009; Akavia et al., 2010; Chia and Tang, 2009; Cheng et al., 2004). Furthermore, protein trafficking pathways are frequently exploited by human pathogens to gain entry and survive within host cells (Weber et al., 2009; Bhavsar et al., 2007; Frey and Robatzek, 2009; Brumell and Scidmore, 2007).

The endomembrane system accounts for a large fraction of the protein coding sequences in eukaryotic genomes (Brighouse et al., 2010), 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 (Avidor-Reiss et al., 2004), whereas the presence of taxon-specific pathways allowed the identification of Fosmidomycin as a potential antimalarial drug (Jomaa et al., 1999). 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 (Dacks and Field, 2007b) including most types of extant molecular components (Jékely, 2003). These are believed to have expanded by duplication and specialisation giving rise to the full diversity of organelles and trafficking pathways observed today (see (Dacks and Field, 2007b) 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 (Stenmark, 2009). Most recently the authors of (Miserey-Lenkei et al., 2010) also linked Rabs to membrane fission. They recruit molecular motors to organelles and transport-vesicles, coordinate intracellular signalling with membrane trafficking, organise distinct sub-domains within membranous organelles and play a critical role in the definition of organelle identity (recently reviewed in reference (Grosshans et al., 2006)). 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 (Dacks and Field, 2007b; Gurkan et al., 2007), suggesting that it provided the primary diversification element in the evolution of trafficking (Gurkan et al., 2007). 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 (Haubruck et al., 1989). 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 (Pereira-Leal, 2008), which have been used as a basis for detailed manual analysis of the Rab families in a variety of organisms (Abbal et al., 2008; Pereira-Leal, 2008; Bright et al., 2010; Lal et al., 2005; Saito-Nakano et al., 2010; Saito-

Nakano et al., 2005; Rutherford and Moore, 2002; Ackers et al., 2005; Quevillon et al., 2003). 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 (Valencia et al., 1991), 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.* (Lal et al., 2005; Saito-Nakano et al., 2010)). 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 ***RabDB.org***.

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

## 4.2 Results and Discussion

### 4.2.1 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 4.0. 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 4.4 and (Diekmann et al., 2011).

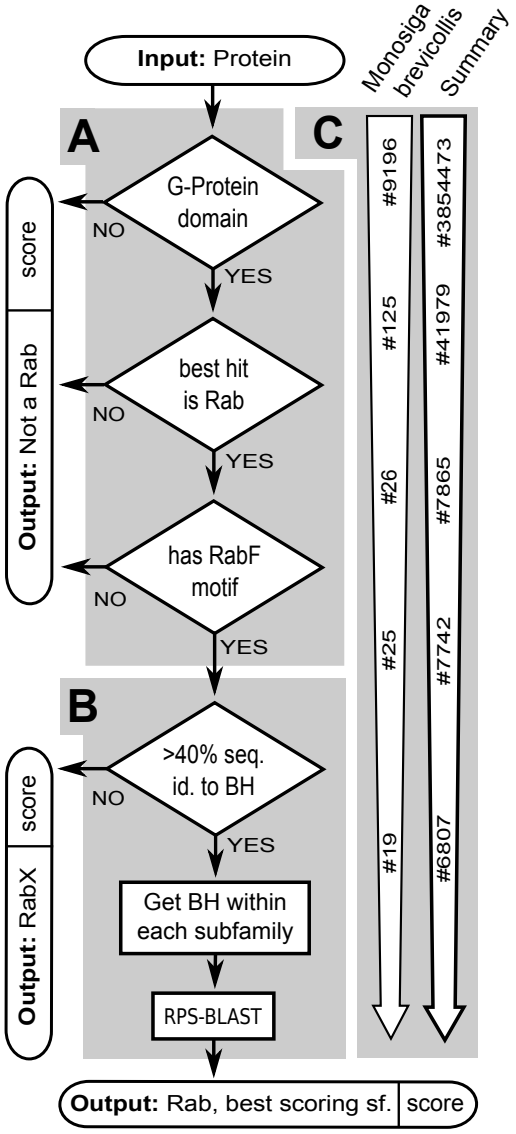
Phase 1 (Figure 4.0A), which classifies protein sequences to the Rab family, proceeds in three stages. First, we check that the protein has a G-protein family domain. As the presence of such a domain can be decided with near certainty, this step drastically reduces the number of candidate Rabs while not excluding any real Rab. In order to do so, we align the sequence against a profile Hidden Markov Model (HMMs) (Eddy, 1996) describing the known GTPase structures, as provided by the Superfamily database (Gough and Chothia, 2002). Secondly, we search for local sequence similarity by performing a BLASTp (Altschul et al., 1990) 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 4.0C, where the number of sequences 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 (Pereira-Leal and Seabra, 2000). 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 4.0B) 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 (Diekmann et al., 2011), and has previously been employed for example in reference (Saito-Nakano et al., 2005). 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 (Altschul et al., 1997)). 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 input 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 4.4 and Supplementary Methods Text S1.

### 4.2.2 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).


Input: Protein
Monosiga brevicollis
Summary
A
G-Protein domain
NO
YES
score
Output: Not a Rab
best hit is Rab
NO
YES
has RabF motif
NO
YES
B
>40% seq. id. to BH
NO
YES
score
Output: RabX
Get BH within each subfamily
RPS-BLAST
C
#9196
#125
#26
#25
#19
#3854473
#41979
#7865
#7742
#6807
Output: Rab, best scoring sf.
score

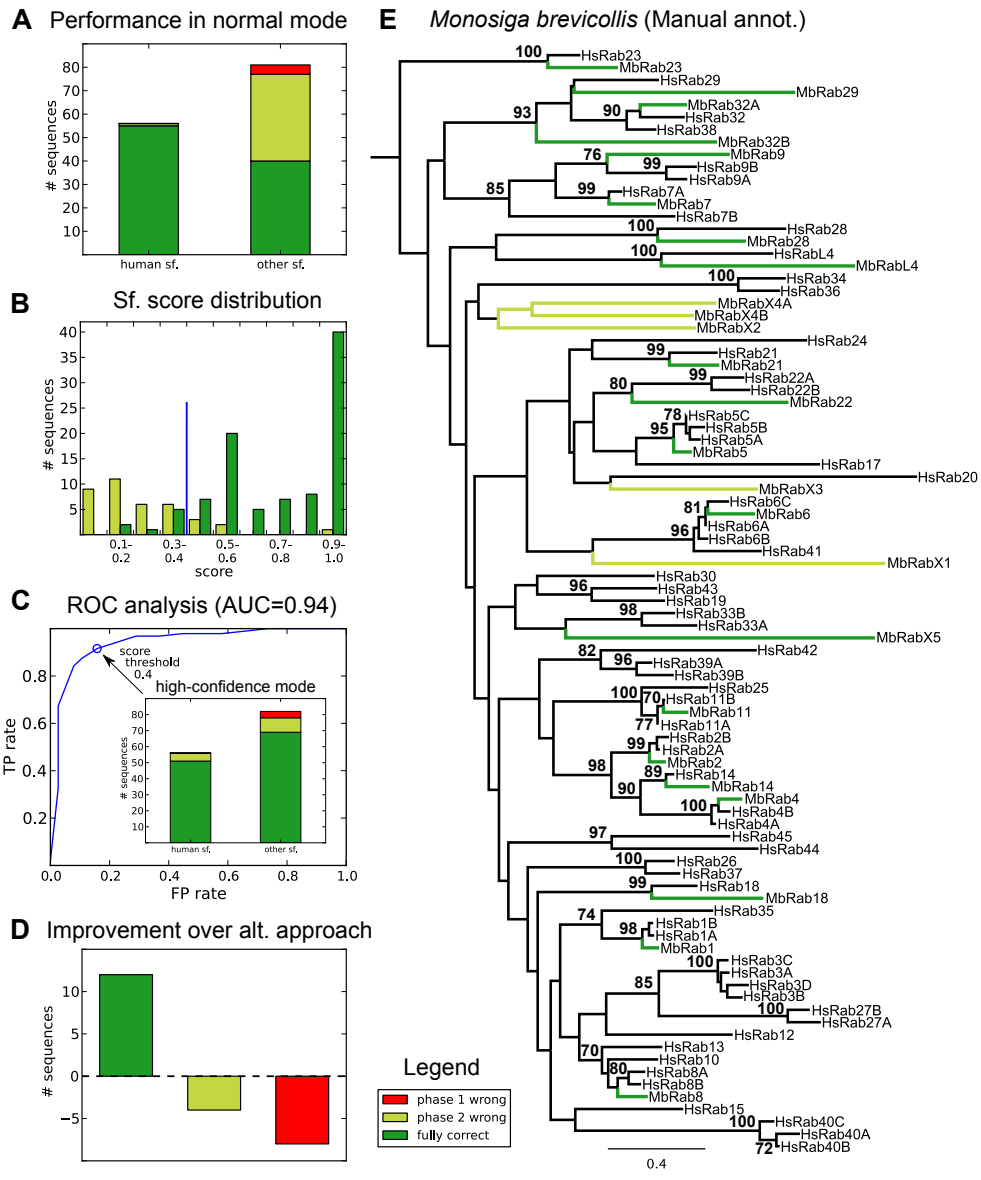
In the absence of a suitable validation dataset, we opted to validate the Rabifier against the manually curated Rab families of three organisms representing distinct worst case scenarios for the Rabifier (Figure 4.0A-C, see Table S1 for a list of all sequences used). This ensures that the validation is meaningful, as it provides a strict lower bound on the expected performance in every day use. First, we chose the Excavate *Trypanosoma brucei* (Ackers et al., 2005), 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* (Saito-Nakano et al., 2005), 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 4.0E), 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 4.0A). 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

Figure 4.0 *(previous page)*: *Flowchart of the Rabifier*—(A) Identification- and (B) classification-procedure implemented by the Rabifier, see Section 4.2 for details on the two phases. Panel (C) shows descriptive statistics from the application of the Rabifier to 247 genomes in the Superfamily database (Wilson et al., 2009a), and details about *M. brevicollis*. *Abbreviations:* best [1]BLAST hit ([1]BH) (Altschul et al., 1990), Rab family motif (RabF) (Pereira-Leal and Seabra, 2000), reverse [1]Ψ-BLAST ([1]RPS-BLAST) (Altschul et al., 1997), subfamily (sf.), Rab not classified to any subfamily within our internal reference set (RabX)

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 (Ackers et al., 2005) 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 4.4). 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 (Saito-Nakano et al., 2005), 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 4.0A). Concretely, for those sequences which were correctly identified as Rabs, we checked 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. brevicollis*, 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 4.0B 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 4.0C. This machine learning technique allows to summarise and quantify the classification performance for all thresholds (Area Under the Curve (AUC) (Hanley and McNeil, 1982), here 0.94), and enables to objectively choose a threshold providing an optimal


A Performance in normal mode
# sequences
human sf.
other sf.
B Sf. score distribution
# sequences
score
C ROC analysis (AUC=0.94)
TP rate
FP rate
score threshold 0.4
high-confidence mode
D Improvement over alt. approach
# sequences
E Monosiga brevicollis (Manual annot.)
Legend
phase 1 wrong
phase 2 wrong
fully correct
0.4

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 4.0C: 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.

Figure 4.0 *(previous 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* (Ackers et al., 2005), *Entamoeba histolytica* (Saito-Nakano et al., 2005) 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 (Fawcett, 2006) analysis shown in (C), which plots the true positive rate against the false positive rate for each possible confidence threshold (Fawcett, 2006) and provides a combined measure of the accuracy of a classifier (Area under the curve, small[1]AUC (Hanley and McNeil, 1982)). 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.)

### 4.2.3 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 (Marchler-Bauer et al., 2011), 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 4.0D (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.

### 4.2.4 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 4.1A). 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 ***RabDB.org*** (Figure 4.1A) together with basic browsing and visualisation tools. Our database is built on top of the Superfamily database (Wilson et al., 2009a) (September 2009 release), which allows us to follow its release cycle and include predictions for all newly sequenced genomes contained therein. Figure 4.1B 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 ***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 ***RabDB.org***.

### 4.2.5 New hypothetical subfamilies

As can be noticed from Figure 4.1B, the Rabifier detected a large number of Rabs not belonging to any subfamily represented in our reference set, *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

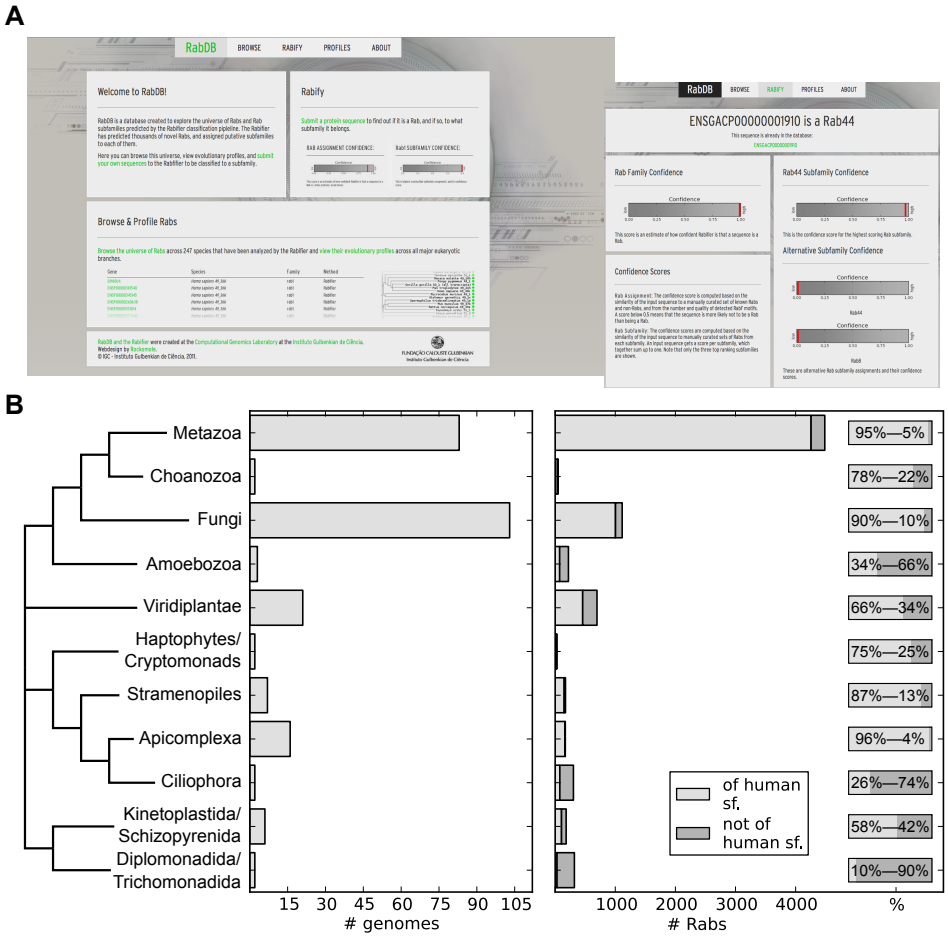



Figure 4.1: *Resources we make available*—(A) Snapshots of the database ***RabDB.org*** which provides public access to the results of the Rabifier applied to the Superfamily database (Wilson et al., 2009a) and the online version of the Rabifier. (B) Statistics of the current content of ***RabDB.org*** 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 (Baldauf, 2003b)) of the eukaryotic taxa is derived from (Burki et al., 2008).

several hypothetical Rab subfamilies (see Section 4.4 for details). The result of this procedure is shown in Figure 4.2, which details the amount of hypothetical subfamilies according to the breadth of their occurrence (see Figure (Diekmann et al., 2011) for an overview of the amount of Rabs falling into each of these classes). We integrated these new subfamilies both in our database, where

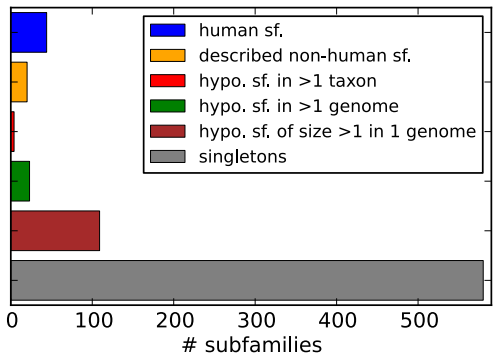



Figure 4.2: *Rab subfamilies in our dataset*—Number of different Rab subfamilies found in our dataset. Human sf. are shown in blue, and other known sf. in orange. The last four categories are hypothetical subfamilies we propose in the context of this paper (see Section 4.4 for details on the procedure): 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 ***RabDB.org***. *Abbreviations:* hypothetical (hypo.), subfamily (sf.)

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.

### 4.2.6 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 4.2). Expansion of certain protein families has been found to correlate with organismal complexity (Vogel and Chothia, 2006). 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.

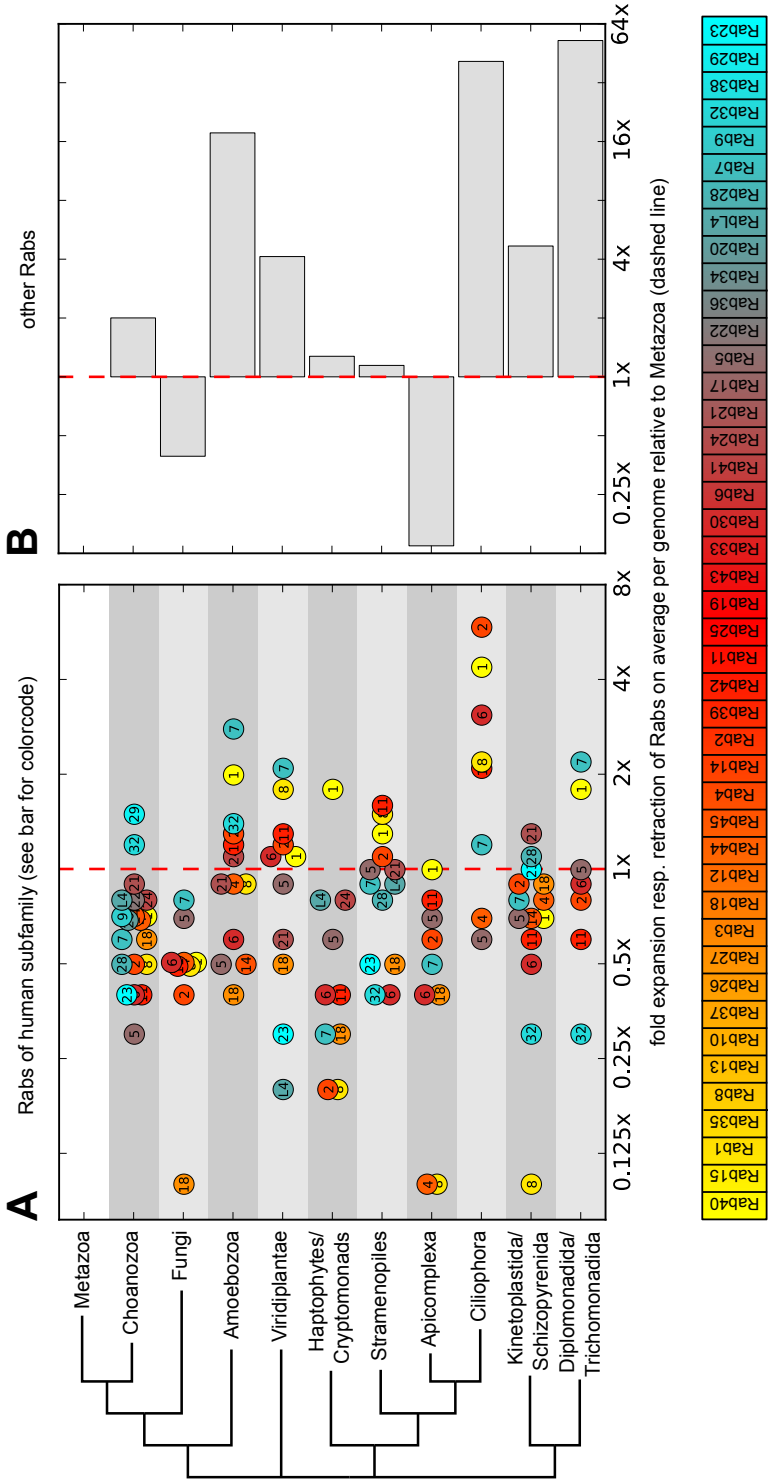
We observe a tremendous heterogeneity in the sizes of Rab repertoires, ranging from five to several hundreds of Rabs in *Encephalitozoon cuniculi* and *Trichomonas vaginalis* respectively. Genomic analyses have shown a general trend for more and larger families in bigger genomes (Jordan et al., 2001; Pushker et al., 2004). 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 (Diekmann et al., 2011)). However, due to the current bias in fully sequenced genomes towards Opisthokonts (compare Figure 4.1B), 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 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 (Moya et al., 2008) and in the obligate intracellular parasitic Microsporidia (Moya et al., 2008). Another example of reduction of Rab repertoires is observed in the fungal branch, as we reported previously (Pereira-Leal, 2008) 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 (Pereira-Leal and Seabra, 2000; Saito-Nakano et al., 2005). We observe large expansions in DiplomonadidaTrichomonadida, Ciliophora and Amoebozoa. Much of these expansions are accounted for by species-specific subfamilies (see Figure 4.2). 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 4.2 reveals that for those Rabs that can be classified, 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 (Pereira-Leal, 2008; Field et al., 1998). 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.

### 4.2.7 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 4.3) and determined their likely time of origin (Figure 4.3). For a detailed analysis of fungal Rabs see (Pereira-Leal, 2008). 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 (Pelkmans et al., 2004; Poteryaev et al., 2010; Kauppi et al., 2002; Mesa et al., 2001; Barbieri et al., 2000). 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


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
RabL4
Rab28
Rab7
Rab9
Rab32
Rab38
Rab29
Rab23
colorcode for human rab subfamilies, ordered by phylogeny (see Materials and Methods)

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 4.3. 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 (Dacks and Field, 2007b), Rab18 (Rutherford and Moore, 2002; Pereira-Leal and Seabra, 2001), Rab21 (Saito-Nakano et al., 2005; Fritz-Laylin et al., 2010) as well as Rab23 and 28 (Ackers et al., 2005) 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 (Diekmann et al., 2011)). 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 (Ackers et al., 2005; Quevillon et al., 2003), and free living fungi frequently have less (Pereira-Leal, 2008). It is remarkable that with every new analysis the LECA appears to become increasingly more complex (Koonin, 2010). 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

Figure 4.2 *(previous page)*: *Rab subfamily expansions relative to Metazoa in a dataset of 247 genomes*—For each of the eukaryotic taxa (as derived from (Burki et al., 2008)), (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 ***RabDB.org***. *Abbreviations:* subfamily (sf.)

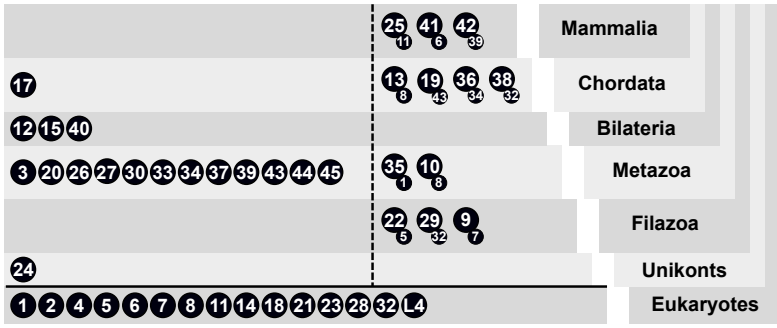



Figure 4.3: *Summary of evolutionary age and duplication origin of human subfamilies*— Each level represents a nested evolutionary stage from the small[1]LECA to humans (derived from (Burki et al., 2008; Shalchian-Tabrizi et al., 2008)) with one circle per human subfamily. Those subfamilies for which 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.

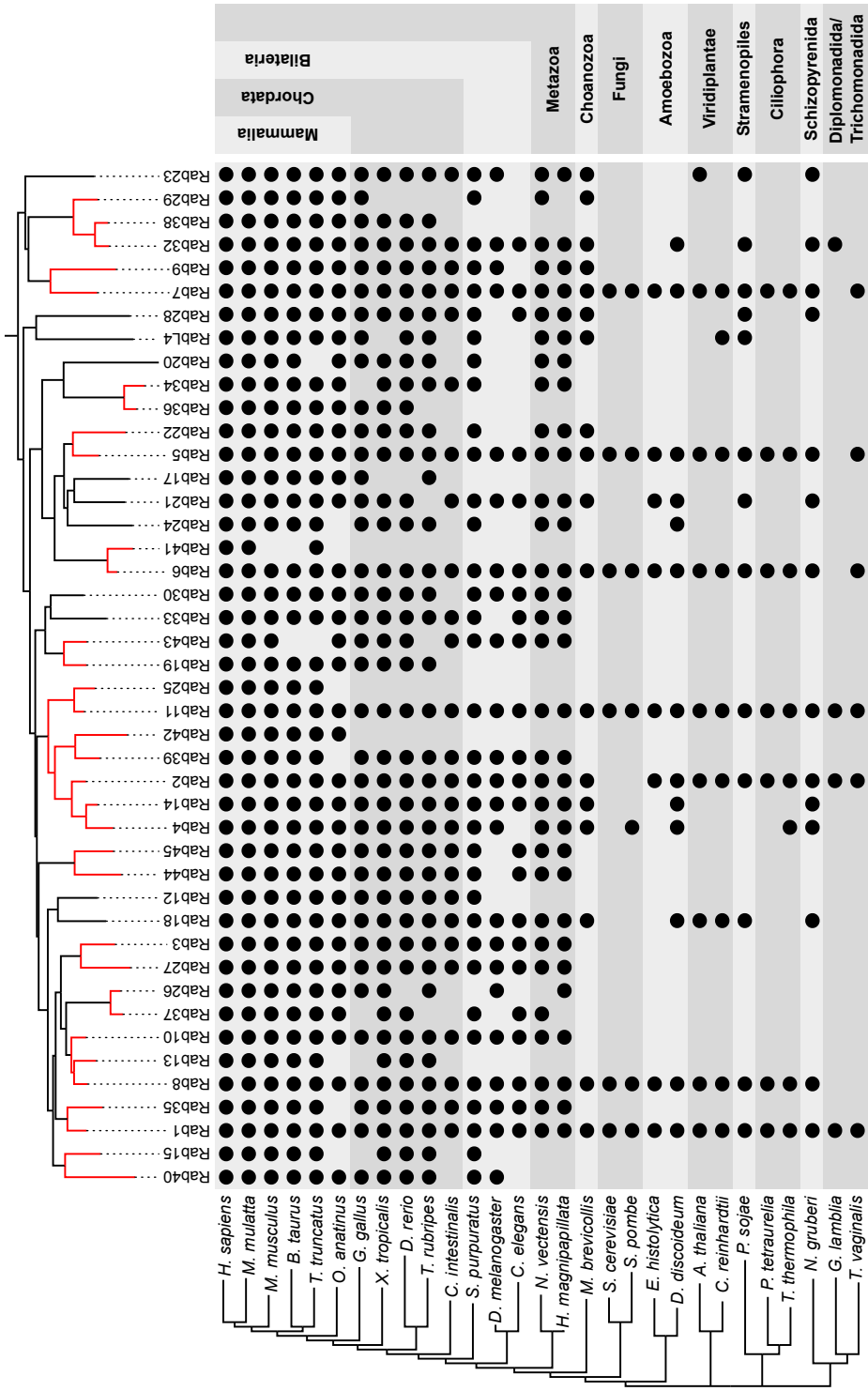
Flagella Transport (IFT) machinery (Qin et al., 2007). Flagella are believed to be ancestral characters, present in the LECA (**CarvalhoSantos2010**; Hodges et al., 2010). Rab32 regulates transport to the pigmentedsecretory granules (Wasmeier et al., 2006), an animal-specific function, but it has also been claimed to have a mitochondria-related function (Alto et al., 2002; Bui et al., 2010). The known function of Rab14 in phagosome maturation and a recycling step at the TGN (Kyei et al., 2006; Proikas-Cezanne et al., 2006) is less clearly ancestral, but it may lend support for a phagotrophic LECA as previously proposed (Cavalier-Smith, 2002).

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 (Dacks and Field, 2007b).

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

The emergence of multicellularity is one of the major transitions in evolution (Smith and Szathmáry, 1997), which happened independently multiple times (see (Rokas, 2008) 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 sequenced (King et al., 2008), and in the context of the validation of the Rabifier we conducted a detailed analysis of its Rab family. The phylogenetic tree in Figure 4.0E reveals a relatively large Rab family with nearly no subfamily expansions (see also Figure 4.2), *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* (Pereira-Leal and Seabra, 2001), suggesting that larger subfamilies observed in mammals represent taxon-specific duplications. Secondly, we observe several organism-specific Rabs, which we labeled MbRabX.


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

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.* (Lal et al., 2005; Saito-Nakano et al., 2005; Ackers et al., 2005; Pereira-Leal and Seabra, 2001)). 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 (Kauppi et al., 2002; Mesa et al., 2001; Ganley et al., 2004; Rodriguez-Gabin et al., 2001). Surprisingly, the genome of *M. brevicollis* codes for proteins previously believed to be specific to multicellular organisms, for example Cadherins (King et al., 2008; Abedin and King, 2010). In animals, trafficking of the cell adhesion molecules Integrins and Cadherins is regulated by Rab4, 5, 11, 21 and 25 (Roberts et al., 2001; Powelka et al., 2004; Pellinen et al., 2006; Caswell et al., 2007), and Rab5 and 7 (Kimura et al., 2006; Frasa et al., 2010), 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 4.3). Surveying the currently known functions of these animal-specific subfamilies suggests roles mainly in regulated secretion (Rab3 (Khvotchev et al., 2003; Rupnik et al., 2007; Schlüter et al., 2002; Tsuboi and Fukuda, 2006), Rab26 (Yoshie et al., 2000), Rab27 (Tsuboi and Fukuda, 2006; Barral et al., 2002; Futter, 2006; Tolmachova et al., 2007), Rab33 (Tsuboi and Fukuda, 2006), Rab37 (Tsuboi and Fukuda, 2006; Masuda et al., 2000), Rab39 (Becker et al., 2009)), trafficking from (Rab10 (Schuck et al., 2007)) and to the Golgi (Rab43 (Dejgaard et al., 2008)) and more generally localisation at

Figure 4.3 *(previous 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 (Burki et al., 2008; Ponting, 2008; Springer and Murphy, 2007; Eliáš, 2010) together with the naming of the partially nested monophyletic groups on the right.

the Golgi (Rab30 (Leeuw et al., 1998; Sinka et al., 2008; Thomas et al., 2009), Rab33 (Valsdottir et al., 2001), Rab34 (Goldenberg et al., 2007), Rab43 (Haas et al., 2007)). Hence, our analysis suggests that the appearance of animals cooccurred with an important expansion and specialisation of the secretory pathway.

### 4.2.9 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 (Innan and Kondrashov, 2010). 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 (Innan and Kondrashov, 2010), 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 4.3 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 4.4) by means of PCR (see Section 4.4). We also analysed publicly available

microarrays (Figures (Diekmann et al., 2011) and S5 from reference (Diekmann et al., 2011)) which overall corroborate the trends described in the following.

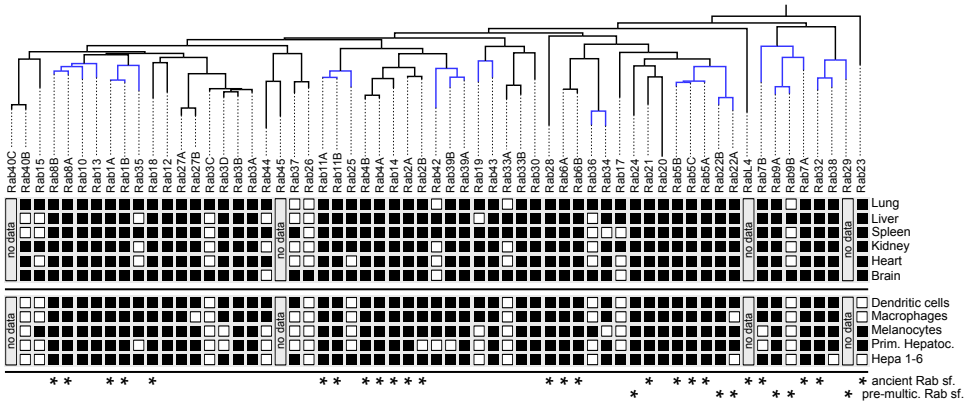



Figure 4.4: *Increasing tissue specificity in expression of derived Rabs in mice*—Summary of small[1]PCR experiments establishing expression (black squares) or lack thereof (white squares) of mouse Rabs in six tissues and five mouse cell lines. Stars on the bottom indicate subfamilies which we found already present in small[1]LECA, and that predate the evolution of multicellularity (see Figure 4.3). 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 4.3 for a summary of which Rabs have a clear origin). *Abbreviations:* subfamily (sf.), primary Hepatocytes (Prim. Hepatoc.), multicellularity (multic.), last eukaryotic common ancestor (small[1]LECA)

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 (Freilich et al., 2006). Second, for the derived subfamilies in which a clear directionality of duplication could be established (see Figure 4.3), 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.

Overall, these observations are strong indications that derived subfamilies are retained for a new tissue-specific functions, different from or at least comple-

menting the progenitor ones. Thus, our results support a neo-functionalisation model explaining the retention of novel Rab sub-families 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 (Diekmann et al., 2011)). 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 (Gurkan et al., 2005) (see also Figure (Diekmann et al., 2011)), 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 small[1]PCR are functionally significant.

## 4.3 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 coexist 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 (Pereira-Leal and Seabra, 2001; Colicelli, 2004; Wennerberg et al., 2005) and Figures S4A-C in reference (Diekmann et al., 2011)). We thus need to derive objective criteria that define a Rab subfamily which go beyond the clearly outdated yet still useful sequence identity cutoff (Pereira-Leal and Seabra, 2000). 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 (Young et al., 2010). 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 (Marques et al., 2008; Byun-McKay and Geeta, 2007).

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.

## 4.4 Materials and Methods

### 4.4.1 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.

### 4.4.2 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) (Pereira-Leal and Seabra, 2000), five subfamilies have been newly discovered (41 to 45/RasEF) (Schwartz et al., 2007). 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 (Joseph, 2006), and are frequently considered to be members of the Rab family (Colicelli, 2004; Schwartz et al., 2007). This view is supported by our own phylogenetic analysis (see tree in Figure S3 in reference (Diekmann et al., 2011)), although without strong bootstrap support. Due to the distinct function and localisation (Joseph, 2006) 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 (Dishinger et al., 2010), and they may be included in the future.

RabL2 proteins were already mentioned in reference (Pereira-Leal and Seabra, 2000) where it is concluded that they are not Rabs, amongst others due to non-conforming RabF motifs. In reference (Colicelli, 2004), RabL2s are said to cluster together with Rans, which we do not include in our analysis. The tree of human GTPases shown in reference (Wennerberg et al., 2005) suggests that RabL2 proteins branch of Rhos at an early stage. Finally, our own tree of human GTPases (Figure S3 in reference (Diekmann et al., 2011)) 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 (Colicelli, 2004), whereas in reference (Wennerberg et al., 2005) 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 (Pereira-Leal and Seabra, 2000).

The last case is RabL4, which all (Colicelli, 2004; Wennerberg et al., 2005;

Schwartz et al., 2007) 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 (Colicelli, 2004; Schwartz et al., 2007) 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 (Diekmann et al., 2011)). In mouse, RabL4 is not classified as being monophyletic with Rab28 (see Figure S4 C in reference (Diekmann et al., 2011)).

### 4.4.3 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 4.0A), the profile HMMs representing the G-protein family domain are either run manually using Perl scripts (as of June 2010) provided by Superfamily (Gough and Chothia, 2002) and HMMER 2.3.2 (Eddy, 1996), or in the case the sequences have been retrieved from the Superfamily database (Wilson et al., 2009a) 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 (Altschul et al., 1990) 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 ***RabDB.org*** with redundancy removed using CDHit (at a 90% sequence identity threshold) (Li and Godzik, 2006). 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) (Rutherford and Moore, 2002), yeast (yptA, ypt10, ypt11), *Drosophila melanogaster* (DmRabX1-X6, DmRab9D, DmRab9F) and *C. elegans* (CeRabY6) (Pereira-Leal and Seabra, 2001). 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 ***RabDB.org***. The last stage of the first phase is performed using the Motif Alignment & Search Tool (MAST) (motif finding threshold 0.0005) (Bailey and Gribskov, 1998) from the MEME-suite (Bailey and Elkan, 1994), with probabilistic representations of the motifs 'IGVDF', 'KLQIW', 'RFxxxT', 'YYRGA', 'LVYDIT' (Pereira-Leal and Seabra, 2000) as input generated on our reference database of Rabs beforehand using MEME.

In the second phase (Figure 4.0B), RPS-BLAST queries (Altschul et al., 1997) 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.

### 4.4.4 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% (Pereira-Leal and Seabra, 2000). In order to resolve the potential conflicts caused by sequences that belong to several clusters at the same time, we applied MCL (Dongen, 2000) (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.

### 4.4.5 Phylogenetic trees

All phylogenetic trees of Rabs and GTPases presented in this article have been generated with PhyML (Guindon and Gascuel, 2003), which implements a Maximum Likelihood probabilistic model, using standard parameters and 100 bootstraps. Alignments were performed with MAFFT (Katoh and Toh, 2008), and manually edited to remove sites with deletions using Jalview (Waterhouse et al., 2009). 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 4.2, Figure 4.3) has been generated by removing isoforms and keeping the longest branch as representative of the corresponding subfamily.

### 4.4.6 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 $\mu$g/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 $\mu$g/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 $\mu$gml 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.

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

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 (Gonçalves et al., 2007) 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 $\mu$M, 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.

# Chapter 5

# Discussion

## 5.1 This thesis, a brief summary

In this thesis we have studied the evolution of the eukaryotic cell from a purely morphological perspective (chapter 2), by comparing morphology to genotype (chapter 3), and based on sequence (chapter 4). In each case we started with a question about the evolution of cells and organelles, only to discover that the methods or data we required did not yet exist. Thus we implemented existing (and sometimes novel) bioinformatics systems to fill this knowledge gap.

The major body of this work focussed on the morphological evolution of MTOCs in eukaryotes (chapter 2). By using a comparative approach to cell biology we were able to create a unique resource which quantifies cilium and centrosome diversity as has never been done before. Using this data I show that cells evolves the same way as the rest of biology. I believe that we should not be surprised by this: there is no reason to assume that evolution works differently at different levels of biology. However, in this project we were able to show that this is the case quantitatively, as opposed to simply postulating. I presented a metric to measure absolute levels of constraint in morphology (the MoDI), as well as a method to calculated the probability of convergent evolution in the absence of a species tree with divergence times. Although both of these metrics were used here to study "comparative cell biology" they can be applied to any biological system.

In chapter 3 we used an existing technique (phylogenetic profiling) to build

predictors for gene function. Although phylogenetic profiling is a technique which is said not to work well in eukaryotes. However, we show that by adding more species from different branches of the eukaryotic tree can greatly enhance the quality of the results. Lastly, we show that it is possible to use phylogenetic profiling to select genes that allow one to predict the presence or absence of an organelle based on its genome.

In chapter 4 we show how creating a new bioinformatics pipeline can result in novel insights to the evolution of diversity in protein families. Although Rab GTPases are a complex family of proteins, identifying and classifying Rabs based on their amino acid sequences turns out to be a trivial task. After trivializing a complex task using a bioinformatics pipeline we can now analyse the entire trafficking machinery present in an organism based on its genome.

The concept of studying cells from an evolutionary perspective using "comparative cell biology" clearly works, has provided some novel insights into how evolution operates at the level of the cell. Each of these projects shows how using bioinformatics approaches allows one to simultaneous study a large amount of species from the entire euakryotic kingdom.

## 5.2 Bioinformatics for comparative cell biology

Bioinformatics is an incredibly young discipline compared to the field of biology, and even cell biology. Yet, in this short time, computers have become an indispensable part of the way we work, especially when dealing with large amounts of complex data.

The bioinformatics approaches we have developed as part of this thesis (mostly those in chapter 2 & 4) are examples where we systematise (and automate) a process typically done manually. One of the advantages of systematization is that it removes the ambiguity caused by independent researchers describing biological in their own way. The mtoc-ontology defines a formal ontology to describe diversity in a single unified language, and the Rabifier outputs unambiguous and clear Rab family assignments using a single nomenclature scheme. The process of automation also allows for more work to be done in less time: using the Rabifier it was possible to annotate 247 eukaryotic genomes in less than a few days. In chapter 2 our goal was to analyze a broad

range of species in incredible structural detail, a task which is near impossible for a single human. By creating a web-resource and annotation pipeline, it became possible to harness the power of the community, and create a database containing the combined knowledge of over 40 experts.

One of the major goals in this thesis was to provide quantitative measures to study cell biology: The Morhplogical Diversity Index, Maximum Parsimony Landscapes and the Rabifier. As discussed in the previous paragraph, automated systems may be less accurate in their calculations or predictions than a detailed manual analysis. However, the advantage of using computational methods is that each measurement is associated with a degree of error, or a confidence. One striking example is the convergent evolution of the centriole-based centrosome (section 2.2.6): Although the results favour a convergent evolution scenario, the 62% confidence suggests that more evidence is required.

Another caveat of using bioinformatics approaches to annotate or classify biological entities is that they are limited by what has been seen before. For example, the Rabifier pipeline is unable to detect novel families of Rabs, and at most can classify a sequence as an "unknown Rab". Another pertinent example is the mtoc-ontology: the possible annotations created an ontology are largely limited by what is possible in the ontology. Although the mtoc-ontology can be extended in some instances (for example, adding a new $n$-fold symmetry), it is not possible to add a completely new MTOC.

Bioinformatics and "comparative cell biology" are both extremely valuable and complementary approaches to studying cell biology. In any process which is systematised or automated, there is typically a trade-off in the amount of data we are able to process and the accuracy of the results. However, we have also seen that in "comparative cell biology" the quality of the results increases with the number of species included in the study. Possibly the most effective (or at least efficient) approach gain a global understanding of the evolution of cells we need to sacrifice some accuracy in the individual datapoints. As with every application of computational techniques, it is important to consider the limitations of the tools being used, and this is also true for bioinformatics in cell biology.

## 5.3 Future directions

The concept of studying the evolution of cells has been gaining interest in the evolutionary, molecular & cell biology communities under the name "evolutionary cell biology" (Lynch et al., 2014; Brodsky et al., 2012). In this thesis we have seen different ways in which both existing and novel bioinformatics approaches can be used to further our understanding of basic biology. As we continue on this quest there will be many opportunities for computational approaches to play a role.

One of the tennets of cell theory is that cell is the atomic unit of life. It is clear that we can use cells to study biology, and that in turn we can use evolution to study cells. This work shows that from a morphological perspective, cells evolve along similar principles as classically studied model organisms from the plant and animal world. Also, we can study the evolution both of cellular components and functions by looking at Rabs: a functionally well classified family of proteins.

One of the initial goals of these projects was to identify genes directly associated with morphological diversity in cilia & centrosome morphology. After obtaining a database of MTOCs morphology across all eukaryotes (chapter 2, we had planned to use genotype–phenotype phylogenetic profiling (chapter 3) to which genes are associated with which phenotypes. For example, we might then ask: "which genes are required for stellate fibers?" or "are there any genes specific to 9-fold symmetry?". Unfortunately this proved to be impossible: The overlap between species for which we have a complete morphological description of MTOCs and those for which the complete genome has been sequenced is incredibly low (and almost completely metazoan). However, as chapter 3 shows, phenotype–genotype predictions in eukaryotes work (at least for organelle presence and absence). Whether this is also the case for diversity in organelle shape and context, still remains to be seen. I see this as a very strong motivator to increase our genomic knowledge and to allocate more resources to sequencing species beyond model organisms.

In chapter 2 (as well as 4) we created databases with the intent to serve as a central point to collect and share biological knowledge. One of the great features of bioinformatics is the ability to integrate data across different resources. We

would like to see the data in ***mtoc-explorer.org*** extend to (and become part of) other resources: The image collection annotated on ***mtoc-explorer.org*** may become part of other image repositories (for instance "The Cell Image Library"), and the mtoc-ontology has many terms which may be part of the Gene Ontology. All of the projects in this thesis have given us a glimpse of cellular diversity in species as the naturally occur. We might stand to much more about constraints, morphology and function in cell biology by addressing similar questions in knockouts and knockdown experiments in model systems. This will allow us to ask questions about the nature of "naturally occurring" morphospace of organelles vs. that of perturbed cells. Similarly we may ask if there certain families of Rabs which are more coupled certain knockout phenotypes?

All of these projects have focussed on naturally occurring morphological, functional and genomic diversity in "healthy representatives" of different species. Disease is also a phenomenon that often occurs at the level of the cell, and with the coming age of translational and personalized medicine, we can envision a place for cell biology in clinical research. For instance, we may catalogue the diversity in MTOCs in cancer cells, or aberrant Rab networks in neural disorders. Lastly, of course, this data on disease phenotypes could be directly mapped to the diversity observed in existing model systems as well as naturally occurring species.

The emerging picture is a full and complete understanding of the biology of the atomic unit of life: the cell. Part of this will involve the using "comparative cell biology" techniques, as well as an "evolutionary cell biology" perspective on how cells operate.

In the long term future we envision a triad composed of disease phenotypes, work in model organisms, and naturally occurring variation across the tree of life, all connected with the cell as the central focal point. This is just the beginning: our understanding of the cell as the atomic unit of life has just started, and bioinformatics will be along for the ride, in the drivers seat.

# Bibliography

Abbal, Philippe, Martine Pradal, Lisa Muniz, François-Xavier Sauvage, Philippe Chatelet, Takashi Ueda, and Catherine Tesniere (2008). "Molecular characterization and expression analysis of the Rab GTPase family in Vitis vinifera reveal the specific expression of a VvRabA protein". *Journal of Experimental Botany* 59.9, pp. 2403–2416.

Abecasis, Ana B., Mónica Serrano, Renato Alves, Leonor Quintais, José B. Pereira-Leal, and Adriano O. Henriques (2013). "A genomic signature and the identification of new sporulation genes". *Journal of Bacteriology* 195.9, pp. 2101–2115. ISSN: 00219193. DOI: 10.1128/JB.02110-12.

Abedin, Monika and Nicole King (2010). "Diverse evolutionary paths to cell adhesion". *Trends in Cell Biology* 20.12, pp. 734–742.

Ackers, John P, Vivek Dhir, and Mark C Field (2005). "A bioinformatic analysis of the RAB genes of Trypanosoma brucei". *Molecular and Biochemical Parasitology* 141.1, pp. 89–97.

Adl, Sina M et al. (2012). "The revised classification of eukaryotes." *The Journal of eukaryotic microbiology* 59.5, pp. 429–93. ISSN: 1550-7408. DOI: 10.1111/j.1550-7408.2012.00644.x.

Agarwal, Roshan, Igor Jurisica, Gordon B Mills, and Kwai Wa Cheng (2009). "The emerging role of the RAB25 small GTPase in cancer". *Traffic* 10.11, pp. 1561–1568.

Aikawa, M. (1966). "The fine structure of the erythrocytic stages of three avian malarial parasites, Plasmodium fallax, P. lophurae, and P. cathemerium." *American Journal of Tropical Medicine and Hygiene* 15.4, pp. 449–471. ISSN: 00029637.

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

Aldrich, H. C. (1968). "The development of flagella in swarm cells of the myxomycete Physarum flavicomum." *Journal of general microbiology* 50.2, pp. 217–222. ISSN: 0022-1287. DOI: 10.1099/00221287-50-2-217.

Aldrich, H C (1969). "The ultrastructure of mitosis in myxamoebae and plasmodia of Physarum flavicomum." *American journal of botany* 56.3, pp. 290–9. ISSN: 0002-9122.

Allen, R. D. (1968). "A reinvestigation of cross-sections of cilia." *Journal of Cell Biology* 37.3, pp. 825–831. ISSN: 00219525. DOI: 10.1083/jcb.37.3.825.

Altenhoff, Adrian M and Christophe Dessimoz (2009). "Phylogenetic and functional assessment of orthologs inference projects and methods." *PLoS computational biology* 5.1, e1000262. ISSN: 1553-7358.

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

Altschul, S F, W Gish, W Miller, E W Myers, and D J Lipman (1990). "Basic local alignment search tool." *Journal of molecular biology* 215, pp. 403–410. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(05)80360-2.

Altschul, Stephen F, Thomas L Madden, Alejandro A Schäffer, J Zhang, Z Zhang, W Miller, and David J Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". *Nucleic Acids Research* 25.17, pp. 3389–3402.

Antonovics, J and P H van Tienderen (1991). "Ontoecogenophyloconstraints? The chaos of constraint terminology." *Trends in ecology & evolution* 6.5, pp. 166–8. ISSN: 0169-5347. DOI: 10.1016/0169-5347(91)90059-7.

Aravind, L. (2000). *Guilt by association: Contextual information in genome analysis*. DOI: 10.1101/gr.10.8.1074.

Archer, F L and D N Wheatley (1971). "Cilia in cell-cultured fibroblasts. II. Incidence in mitotic and post-mitotic BHK 21-C13 fibroblasts." *Journal of anatomy* 109.Pt 2, pp. 277–292. ISSN: 0021-8782.

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

Arnaiz, Olivier, Agata Malinowska, Catherine Klotz, Linda Sperling, Michal Dadlez, France Koll, and Jean Cohen (2009). "Cildb: a knowledgebase for centrosomes and cilia." *Database : the journal of biological databases and curation* 2009, bap022. ISSN: 1758-0463. DOI: 10.1093/database/bap022.

Arnold, S J (1992). "Constraints on phenotypic evolution." *The American naturalist* 140 Suppl, S85–107. ISSN: 0003-0147. DOI: 10.1086/285398.

Ashburner, M et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." *Nature genetics* 25.1, pp. 25–9. ISSN: 1061-4036. DOI: 10.1038/75556.

Avidor-Reiss, Tomer, Andreia M Maer, Edmund Koundakjian, Andrey Polyanovsky, Thomas Keil, Shankar Subramaniam, and Charles S Zuker (2004). "Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis." *Cell* 117.4, pp. 527–39. ISSN: 0092-8674.

Azimzadeh, Juliette (2014). "Exploring the evolutionary history of centrosomes." *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 369.1650. ISSN: 1471-2970. DOI: 10.1098/rstb.2013.0453.

Azimzadeh, Juliette and Michel Bornens (2007). "Structure and duplication of the centrosome." *Journal of cell science* 120.Pt 13, pp. 2139–42. ISSN: 0021-9533. DOI: 10.1242/jcs.005231.

Azimzadeh, Juliette, Mei Lie Wong, Diane Miller Downhour, Alejandro Sánchez Alvarado, and Wallace F Marshall (2012). "Centrosome Loss in the Evolution of Planarians." *Science (New York, N.Y.)* 461. ISSN: 1095-9203. DOI: 10.1126/science.1214457.

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

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

Baldauf, S L (2003a). "The deep roots of eukaryotes." *Science (New York, N.Y.)* 300.5626, pp. 1703–6. ISSN: 1095-9203. DOI: 10.1126/science.1085544.

Baldauf, Sandra L (2003b). "Phylogeny for the faint of heart: a tutorial". *Trends in Genetics : TIG* 19.6, pp. 345–351.

Balhoff, James P, Wasila M Dahdul, Cartik R Kothari, Hilmar Lapp, John G Lundberg, Paula Mabee, Peter E Midford, Monte Westerfield, and Todd J Vision (2010). "Phenex: ontological annotation of phenotypic diversity." *PloS one* 5.5, e10500. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0010500.

Barbieri, M A, R L Roberts, A Gumusboga, H Highfield, C Alvarez-Dominguez, A Wells, and P D Stahl (2000). "Epidermal growth factor and membrane

trafficking. EGF receptor activation of endocytosis requires Rab5a". *The Journal of Cell Biology* 151.3, pp. 539–550.

Bard, Jonathan, Seung Y Rhee, and Michael Ashburner (2005). "An ontology for cell types." *Genome biology* 6.2, R21. ISSN: 1465-6914. DOI: 10.1186/gb-2005-6-2-r21.

Barker, Daniel and Mark Pagel (2005). "Predicting functional gene links from phylogenetic-statistical analyses of whole genomes." *PLoS computational biology* 1.1, e3. ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.0010003.

Barker, Daniel, Andrew Meade, Mark Pagel, and Mark Page (2007). "Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes." *Bioinformatics (Oxford, England)* 23.1, pp. 14–20. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btl558.

Barral, Duarte C et al. (2002). "Functional redundancy of Rab27 proteins and the pathogenesis of Griscelli syndrome". *The Journal of clinical investigation* 110.2, pp. 247–257.

Becker, Christine E, Emma M Creagh, and Luke A J O'Neill (2009). "Rab39a binds caspase-1 and is required for caspase-1-dependent interleukin-1beta secretion". *The Journal of biological chemistry* 284.50, pp. 34531–34537.

Beneden, Edouard van and Adolphe Neyt (1887). "Nouvelle recherches sur la fécondation et la division mitosique chez l'Ascaride mégalocéphale". *Bull. Acad. Roy. Belg. III* 14, p. 215.

Bhavsar, Amit P, Julian A Guttman, and B Brett Finlay (2007). "Manipulation of host-cell pathways by bacterial pathogens". *Nature* 449.7164, pp. 827–834.

Biggins, Sue and Matthew D Welch (2014). "Editorial overview: Cell architecture: Cellular organization and function". *Current Opinion in Cell Biology*, pp. 1–3. ISSN: 09550674. DOI: 10.1016/j.ceb.2013.12.008.

Bisby, F a (2000). "The quiet revolution: biodiversity informatics and the internet." *Science (New York, N.Y.)* 289.5488, pp. 2309–2312. ISSN: 00368075. DOI: 10.1126/science.289.5488.2309.

Blake, Judith (2004). "Bio-ontologies-fast and furious." *Nature biotechnology* 22.6, pp. 773–4. ISSN: 1087-0156. DOI: 10.1038/nbt0604-773.

Bon, Mr (2002). "Traffic jams II: an update of diseases of intracellular transport". *Traffic* 3.11, pp. 781–790.

Bornens, Michel (2012). "The Centrosome in Cells and Organisms". *Science* 335.6067, pp. 422–426. ISSN: 1095-9203. DOI: 10.1126/science.1209037.

Boveri, Theodor (1887). "Ueber den Antheil des Spermatozoon an der Teilung des Eies". *Sitzungsber. Ges. Morph. Physiol.* 3, pp. 151–164.

Bowes, Jeff B., Kevin a. Snyder, Erik Segerdell, Chris J. Jarabek, Kenan Azam, Aaron M. Zorn, and Peter D. Vize (2009). "Xenbase: Gene expression and improved integration". *Nucleic Acids Research* 38.November 2009, pp. 607–612. ISSN: 03051048. DOI: 10.1093/nar/gkp953.

Bradford, Yvonne et al. (2011). "ZFIN: enhancements and updates to the Zebrafish Model Organism Database." *Nucleic acids research* 39.Database issue, pp. D822–9. ISSN: 1362-4962. DOI: 10.1093/nar/gkq1077.

Braselton, James P. (1988). "10. Karyology and systematics of Plasmodiophoromucytes". *Viruses with Fungal Vectors*, pp. 139–152.

Bricheux, Geneviève, Gérard Coffe, and Guy Brugerolle (2007). "Identification of a new protein in the centrosome-like "atractophore" of Trichomonas vaginalis". *Molecular and Biochemical Parasitology* 153.2, pp. 133–140. ISSN: 01666851. DOI: 10.1016/j.molbiopara.2007.02.011.

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

Bright, Lydia J, Nichole Kambesis, Scott Brent Nelson, Byeongmoon Jeong, and Aaron P Turkewitz (2010). "Comprehensive analysis reveals dynamic and evolutionary plasticity of Rab GTPases and membrane traffic in Tetrahymena thermophila". *PLoS Genetics* 6.10, e1001155.

Broadhead, Richard et al. (2006). "Flagellar motility is required for the viability of the bloodstream trypanosome." *Nature* 440.7081, pp. 224–7. ISSN: 1476-4687. DOI: 10.1038/nature04541.

Brodsky, Frances M, Mukund Thattai, and Satyajit Mayor (2012). "Evolutionary cell biology: Lessons from diversity." *Nature cell biology* 14.7, p. 651. ISSN: 1476-4679. DOI: 10.1038/ncb2539.

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

Bui, Michael et al. (2010). "Rab32 modulates apoptosis onset and mitochondria-associated membrane (MAM) properties". *The Journal of biological chemistry* 285.41, pp. 31590–31602.

Burki, Fabien, Kamran Shalchian-Tabrizi, and Jan Pawlowski (2008). "Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes". *Biology letters* 4.4, pp. 366–369.

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

Carvalho-Santos, Zita, Pedro Machado, Pedro Branco, Filipe Tavares-Cadete, Ana Rodrigues-Martins, José B Pereira-Leal, and Mónica Bettencourt-Dias (2010). "Stepwise evolution of the centriole-assembly pathway." *Journal of cell science* 123.Pt 9, pp. 1414–26. ISSN: 1477-9137. DOI: 10.1242/jcs.064931.

Carvalho-Santos, Zita, Juliette Azimzadeh, José B. Pereira-Leal, and Mónica Bettencourt-Dias (2011). "Evolution: Tracing the origins of centrioles, cilia, and flagella." *The Journal of cell biology* 194.2, pp. 165–75. ISSN: 1540-8140. DOI: 10.1083/jcb.201011152.

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

Cavalier-Smith, Thomas (2002). "The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa". *International journal of systematic and evolutionary microbiology* 52.2, pp. 297–354.

Cavalier-Smith, Thomas, Rhodri Lewis, Ema E. Chao, Brian Oates, and David Bass (2008). "Morphology and Phylogeny of Sainouron acronematica sp. n. and the Ultrastructural Unity of Cercozoa". *Protist* 159.4, pp. 591–620. ISSN: 14344610. DOI: 10.1016/j.protis.2008.04.002.

Chen, Feng, Aaron J Mackey, Jeroen K Vermunt, and David S Roos (2007). "Assessing performance of orthology detection strategies applied to eukaryotic genomes." *PloS one* 2.4, e383. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0000383.

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

Cherry, J Michael et al. (2012). "Saccharomyces Genome Database: the genomics resource of budding yeast." *Nucleic acids research* 40.Database issue, pp. D700–5. ISSN: 1362-4962. DOI: 10.1093/nar/gkr1029.

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

Cokus, Shawn, Sayaka Mizutani, and Matteo Pellegrini (2007). "An improved method for identifying functionally linked proteins using phylogenetic profiles." *BMC bioinformatics* 8 Suppl 4, S7. ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-S4-S7.

Colicelli, John (2004). "Human RAS superfamily proteins and related GTPases". *Science's STKE* 2004.250, RE13.

Costa, Marta, Simon Reeve, Gary Grumbling, and David Osumi-Sutherland (2013). "The Drosophila anatomy ontology." *Journal of biomedical semantics* 4, p. 32. ISSN: 2041-1480. DOI: 10.1186/2041-1480-4-32.

Dacks, Joel B and Mark C Field (2007a). "Evolution of the eukaryotic membrane-trafficking system: origin, tempo and mode." *Journal of cell science* 120.Pt 17, pp. 2977–2985. ISSN: 0021-9533. DOI: 10.1242/jcs.013250.

Dacks, Joel B and Mark C Field (2007b). "Evolution of the eukaryotic membrane-trafficking system: origin, tempo and mode". *Journal of Cell Science* 120.17, pp. 2977–2985.

Dahdul, Wasila M et al. (2010). "Evolutionary characters, phenotypes and ontologies: curating data from the systematic biology literature." *PloS one* 5.5, e10708. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0010708.

Dam, Teunis Jp van, Gabrielle Wheway, Gisela G Slaats, Martijn a Huynen, and Rachel H Giles (2013). "The SYSCILIA gold standard (SCGSv1) of known ciliary components and its applications within a systems biology consortium." *Cilia* 2.1, p. 7. ISSN: 2046-2530. DOI: 10.1186/2046-2530-2-7.

Date, Shailesh V and Edward M Marcotte (2003). "Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages." *Nature biotechnology* 21.9, pp. 1055–62. ISSN: 1087-0156. DOI: 10.1038/nbt861.

Daunderer, C, M Schliwa, and R Gräf (1999). "Dictyostelium discoideum: a promising centrosome model system." *Biology of the cell / under the auspices of the European Cell Biology Organization* 91.4-5, pp. 313–20. ISSN: 0248-4900.

Deans, Andrew R, Matthew J Yoder, and James P Balhoff (2012). "Time to change how we describe biodiversity." *Trends in ecology & evolution* 27.2, pp. 78–84. ISSN: 0169-5347. DOI: 10.1016/j.tree.2011.11.007.

Debec, Alain, William Sullivan, and Monica Bettencourt-Dias (2010). "Centrioles: active players or passengers during mitosis?" *Cellular and molecular life sciences : CMLS* 67.13, pp. 2173–94. ISSN: 1420-9071. DOI: 10.1007/s00018-010-0323-9.

Dejgaard, Selma Y et al. (2008). "Rab18 and Rab43 have key roles in ER-Golgi trafficking". *Journal of Cell Science* 121.Pt 16, pp. 2768–2781.

Desser, SS (1980). "An ultrastructural study of the asexual development of a presumed Isospora sp. in mononuclear, phagocytic cells of the evening grosbeak (Hesperiphona vespertina)". *The Journal of Parasitology* 66.4, pp. 601–612.

Diekmann, Yoan and José B Pereira-Leal (2013). "Evolution of intracellular compartmentalization." *The Biochemical journal* 449.2, pp. 319–31. ISSN: 1470-8728. DOI: 10.1042/BJ20120957.

Diekmann, Yoan, Elsa Seixas, Marc Gouw, Filipe Tavares-Cadete, Miguel C. Seabra, and José B. Pereira-Leal (2011). "Thousands of Rab GTPases for the cell biologist". *PLoS Computational Biology* 7.10. ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1002217.

Dingemans, K. P. (1969). "THE RELATION BETWEEN CILIA AND MITOSES IN THE MOUSE ADENOHYPOPHYSIS". *The Journal of Cell Biology* 43.2, pp. 361–367. ISSN: 0021-9525. DOI: 10.1083/jcb.43.2.361.

Dingle, A. D. and C. Fulton (1966). "Development of the flagellar apparatus of Naegleria." *Journal of Cell Biology* 31.1, pp. 43–54. ISSN: 00219525. DOI: 10.1083/jcb.31.1.43.

Dippell, R V (1968). "The development of basal bodies in paramecium." *Proceedings of the National Academy of Sciences of the United States of America* 61.2, pp. 461–468. ISSN: 0027-8424. DOI: 10.1073/pnas.61.2.461.

Dishinger, John F et al. (2010). "Ciliary entry of the kinesin-2 motor KIF17 is regulated by importin-beta2 and RanGTP". *Nature Cell Biology* 12.7, pp. 703–710.

Dongen, Stijn van (2000). "A cluster algorithm for graphs". *Technical report* INS-R0010, pp. 1–42.

Dress, Andreas et al. (2008). *Anatomy Ontologies for Bioinformatics*. Vol. 6. ISBN: 978-1-84628-884-5. DOI: 10.1007/978-1-84628-885-2.

Eddy, Sean R (1996). "Hidden Markov models". *Current Opinion in Structural Biology* 6.3, pp. 361–365.

Eliáš, Marek (2010). "Patterns and processes in the evolution of the eukaryotic endomembrane system". *Molecular Membrane Biology* 27.8, pp. 469–489.

Enault, F., K. Suhre, C. Abergel, O. Poirot, and J.-M. Claverie (2003). "Annotation of bacterial genomes using improved phylogenomic profiles". *Bioinformatics* 19.Suppl 1, pp. i105–i107. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btg1013.

Fawcett, D W and K R Porter (1954). "A study of the fine structure of ciliated epithelia". *Journal of Morphology* 94.2, pp. 221–281. ISSN: 0362-2525. DOI: 10.1002/jmor.1050940202.

Fawcett, Tom (2006). "An introduction to ROC analysis". *Pattern Recognition Letters* 27, pp. 861–874.

Field, H, M Farjah, A Pal, Keith Gull, and Mark C Field (1998). "Complexity of trypanosomatid endocytosis pathways revealed by Rab4 and Rab5 isoforms in Trypanosoma brucei". *The Journal of biological chemistry* 273.48, pp. 32102–32110.

Fischer, Steve, Brian P. Brunk, Feng Chen, Xin Gao, Omar S. Harb, John B. Iodice, Dhanasekaran Shanmugam, David S. Roos, and Christian J. Stoeckert (2011). "Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups". *Current Protocols in Bioinformatics*, pp. 1–19. ISSN: 19343396. DOI: 10.1002/0471250953.bi0612s35.

Frasa, Marieke A M et al. (2010). "Armus is a Rac1 effector that inactivates Rab7 and regulates E-cadherin degradation". *Current Biology* 20.3, pp. 198–208.

Freilich, Shiri, Tim Massingham, Eric Blanc, Leon Goldovsky, and Janet M Thornton (2006). "Relating tissue specialization to the differentiation of expression of singleton and duplicate mouse proteins". *Genome Biology* 7.10, R89.

Frey, Nicolas Frei dit and Silke Robatzek (2009). "Trafficking vesicles: pro or contra pathogens?" *Current Opinion in Plant Biology* 12.4, pp. 437–443.

Fritz-Laylin, Lillian K et al. (2010). "The genome of Naegleria gruberi illuminates early eukaryotic versatility". *Cell* 140.5, pp. 631–642.

Fulton, Chandler and Allan D Dingle (1971). "Basal bodies, but not centrioles, in Naegleria." *The Journal of cell biology* 51.3, pp. 826–36. ISSN: 0021-9525.

Futter, Clare E (2006). "The molecular regulation of organelle transport in mammalian retinal pigment epithelial cells". *Pigment cell research* 19.2, pp. 104–111.

Gadelha, Catarina, Bill Wickstead, Paul G McKean, and Keith Gull (2006). "Basal body and flagellum mutants reveal a rotational constraint of the central pair microtubules in the axonemes of trypanosomes." *Journal of cell science* 119.Pt 12, pp. 2405–13. ISSN: 0021-9533. DOI: 10.1242/jcs.02969.

Ganley, Ian G, Kate Carroll, Lenka Bittova, and Suzanne R Pfeffer (2004). "Rab9 GTPase regulates late endosome size and requires effector interaction for its stability". *Molecular Biology of the Cell* 15.12, pp. 5420–5430.

Garber, R C and J R Aist (1979). "The ultrastructure of mitosis in Plasmodiophora brassicae (Plasmodiophorales)." *Journal of cell science* 40, pp. 89–110. ISSN: 0021-9533.

Geimer, Stefan and Michael Melkonian (2004). "The ultrastructure of the Chlamydomonas reinhardtii basal apparatus: identification of an early marker of radial asymmetry inherent in the basal body." *Journal of cell science* 117.Pt 13, pp. 2663–74. ISSN: 0021-9533. DOI: 10.1242/jcs.01120.

Gely, C and M Wright (1986). "The centriole cycle in the amoebae of the myxomycetePhysarum polycephalum". *Protoplasma* 132.1-2, pp. 23–31. ISSN: 0033-183X. DOI: 10.1007/BF01275786.

Getty, Thomas (2000). "A constrained view of constraints". *Trends in Ecology & Evolution* 15.6, p. 249. ISSN: 01695347. DOI: 10.1016/S0169-5347(00)01865-6.

Gifford, Ernest M. and Susan Larson (1980). "Developmental Features of the Spermatogenous Cell in Ginkgo biloba". *American Journal of Botany* 67.1, p. 119. ISSN: 00029122. DOI: 10.2307/2442543.

Gkoutos, Georgios V, Eain C J Green, Ann-Marie Mallon, John M Hancock, and Duncan Davidson (2005). "Using ontologies to describe mouse phenotypes." *Genome biology* 6.1, R8. ISSN: 1465-6914. DOI: 10.1186/gb-2004-6-1-r8.

Goetz, Sarah C and Kathryn V Anderson (2010). "The primary cilium: a signalling centre during vertebrate development." *Nature reviews. Genetics* 11.5, pp. 331–344. ISSN: 1471-0056. DOI: 10.1038/nrg2774.

Goldenberg, Neil M, Sergio Grinstein, and Mel Silverman (2007). "Golgi-bound Rab34 is a novel member of the secretory pathway". *Molecular Biology of the Cell* 18.12, pp. 4762–4771.

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

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

Gould, S. J. and R. C. Lewontin (1979). "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme". *Proceedings of the Royal Society B: Biological Sciences* 205.1161, pp. 581–598. ISSN: 0962-8452. DOI: 10.1098/rspb.1979.0086.

Gould, SJ and ES Vrba (1982). "Exaptation-a missing term in the science of form". *Paleobiology* 8.1, pp. 5–15.

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

Guindon, Stéphane and Olivier Gascuel (2003). "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood". *Systematic Biology* 52.5, pp. 696–704.

Gurkan, Cemal, Hilmar Lapp, Christelle Alory, Andrew I Su, John B Hogenesch, and William E Balch (2005). "Large-scale profiling of Rab GTPase trafficking networks: the membrome". *Molecular Biology of the Cell* 16.8, pp. 3847–3864.

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

Haas, Alexander K, Shin-ichiro Yoshimura, David J Stephens, Christian Preisinger, Evelyn Fuchs, and Francis A Barr (2007). "Analysis of GTPase-activating proteins: Rab1 and Rab43 are key Rabs required to maintain a functional Golgi complex in human cells". *Journal of Cell Science* 120.17, pp. 2997–3010.

Haeckel, Ernst (1904). *Kunstformen der Natur*. Leipzig und Wien: Bibliographisches Institut.

Haimo, Leah T and Joel L Rosenbaum (1981). "Cilia, flagella, and microtubules." *The Journal of cell biology* 91.3 Pt 2, 125s–130s. ISSN: 0021-9525.

Hall, Brian K. (2008). "From Marshalling Yards to Landscapes to Triangles to Morphospace". *Evolutionary Biology* 35.2, pp. 97–99. ISSN: 0071-3260. DOI: 10.1007/s11692-008-9021-z.

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

Haubruck, H, R Prange, C Vorgias, and D Gallwitz (1989). "The ras-related mouse ypt1 protein can functionally replace the YPT1 gene product in yeast". *The EMBO Journal* 8.5, pp. 1427–1432.

Hayamizu, Terry F, Michael N Wicks, Duncan R Davidson, Albert Burger, Martin Ringwald, and Richard A Baldock (2013). "EMAP/EMAPA ontology of mouse developmental anatomy: 2013 update." *Journal of biomedical semantics* 4.1, p. 15. ISSN: 2041-1480. DOI: 10.1186/2041-1480-4-15.

Heath, I. Brent and W. Marshall Darley (1972). "OBSERVATIONS ON THE ULTRASTRUCTURE OF THE MALE GAMETES OF BIDDULPHIA LEVIS EHR." *Journal of Phycology* 8.1, pp. 51–59. ISSN: 0022-3646. DOI: 10.1111/j.1529-8817.1972.tb04001.x.

Hodges, Matthew E, Nicole Scheumann, Bill Wickstead, Jane A Langdale, and Keith Gull (2010). "Reconstructing the evolutionary history of the centriole from protein components". *Journal of Cell Science* 123.9, pp. 1407–1413.

Hoffman, John C. and Kevin C. Vaughn (1995). *Using the Developing Spermatogenous Cells of Ceratopteris to Unlock the Mysteries of the Plant Cytoskeleton*. DOI: 10.1086/297256.

Holst, Elke and Anita Wiemer (2010). "Zur Unterrepräsentanz von Frauen in Spitzengremien der Wirtschaft: Ursachen und Handlungsansätze".

Howe, Doug and Seung Yon (2008). "The future of biocuration". *Nature* 455.September, pp. 47–50. ISSN: 14764687. DOI: 10.1038/455047a.

Hunter, Amy, Matthew H. Kaufman, Angus McKay, Richard Baldock, Martin W. Simmen, and Jonathan B L Bard (2003). "An ontology of human developmental anatomy". *Journal of Anatomy* 203, pp. 347–355. ISSN: 00218782. DOI: 10.1046/j.1469-7580.2003.00224.x.

Huynen, M A, Yolande Diaz-Lazcoz, and Peer Bork (1997). *Differential genome display*. DOI: 10.1016/S0168-9525(97)01255-9.

Innan, Hideki and Fyodor A Kondrashov (2010). "The evolution of gene duplications: classifying and distinguishing between models". *Nature Reviews Genetics* 11.2, pp. 97–108.

Jaccard, P (1912). "The distribution of the flora in the alphine zone". *The New Phytologist* XI.2, pp. 37–50. ISSN: 1469-8137. DOI: 10.1111/j.1469-8137.1912.tb05611.x.

Jékely, G (2007). *Eukaryotic Membranes and Cytoskeleton*. Vol. 607. Advances in Experimental Medicine and Biology. New York, NY: Springer New York. ISBN: 9780387740201. DOI: 10.1007/978-0-387-74021-8.

Jékely, Gáspár (2003). "Small GTPases and the evolution of the eukaryotic cell". *BioEssays* 25.11, pp. 1129–1138.

Jékely, Gáspár and Detlev Arendt (2006). "Evolution of intraflagellar transport from coated vesicles and autogenous origin of the eukaryotic cilium". *BioEssays* 28, pp. 191–198. ISSN: 02659247. DOI: 10.1002/bies.20369.

Jim, Kam, Kush Parmar, Mona Singh, and Saeed Tavazoie (2004). "A cross-genomic approach for systematic mapping of phenotypic traits to genes." *Genome research* 14.1, pp. 109–15. ISSN: 1088-9051. DOI: 10.1101/gr.1586704.

Jin, Ke et al. (2012). "PhenoM: a database of morphological phenotypes caused by mutation of essential genes in Saccharomyces cerevisiae." *Nucleic acids research* 40.Database issue, pp. D687–94. ISSN: 1362-4962. DOI: 10.1093/nar/gkr827.

Jomaa, H et al. (1999). "Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs". *Science* 285.5433, pp. 1573–1576.

Jordan, I King, Kira S Makarova, J L Spouge, Yuri I Wolf, and Eugene V Koonin (2001). "Lineage-specific gene expansions in bacterial and archaeal genomes". *Genome Research* 11.4, pp. 555–565.

Joseph, Jomon (2006). "Ran at a glance". *Journal of Cell Science* 119.Pt 17, pp. 3481–3484.

Karpinka, J. B., J. D. Fortriede, K. a. Burns, C. James-Zorn, V. G. Ponferrada, J. Lee, K. Karimi, a. M. Zorn, and P. D. Vize (2014). "Xenbase, the Xenopus model organism database; new virtualized system, data types and genomes". *Nucleic Acids Research* 43, pp. D756–D763. ISSN: 0305-1048. DOI: 10.1093/nar/gku956.

Katoh, Kazutaka and Hiroyuki Toh (2008). "Recent developments in the MAFFT multiple sequence alignment program". *Briefings in Bioinformatics* 9.4, pp. 286–298.

Kauppi, Maria, Anne Simonsen, Bjørn Bremnes, Amandio Vieira, Judy Callaghan, Harald Stenmark, and Vesa M Olkkonen (2002). "The small

GTPase Rab22 interacts with EEA1 and controls endosomal membrane trafficking". *Journal of Cell Science* 115.5, pp. 899–911.

Keller, Lani C, Edwin P Romijn, Ivan Zamora, John R Yates, and Wallace F Marshall (2005). "Proteomic analysis of isolated chlamydomonas centrioles reveals orthologs of ciliary-disease genes." *Current biology : CB* 15.12, pp. 1090–8. ISSN: 0960-9822. DOI: 10.1016/j.cub.2005.05.024.

Kensche, Philip R, Vera van Noort, Bas E Dutilh, and Martijn a Huynen (2008). "Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution." *Journal of the Royal Society, Interface / the Royal Society* 5.19, pp. 151–70. ISSN: 1742-5689. DOI: 10.1098/rsif.2007.1047.

Khvotchev, Mikhail V, Mindong Ren, Shigeo Takamori, Reinhard Jahn, and Thomas C Südhof (2003). "Divergent functions of neuronal Rab11b in Ca2+-regulated versus constitutive exocytosis". *The Journal of neuroscience* 23.33, pp. 10531–10539.

Kilmartin, John V (2014). "Lessons from yeast: the spindle pole body and the centrosome." *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 369.1650. ISSN: 1471-2970. DOI: 10.1098/rstb.2013.0456.

Kimura, Toshihiro, Toshiaki Sakisaka, Takeshi Baba, Tomohiro Yamada, and Yoshimi Takai (2006). "Involvement of the Ras-Ras-activated Rab5 guanine nucleotide exchange factor RIN2-Rab5 pathway in the hepatocyte growth factor-induced endocytosis of E-cadherin". *Journal of Biological Chemistry* 281.15, pp. 10598–10609.

King, Nicole et al. (2008). "The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans". *Nature* 451.7180, pp. 783–788.

Konno, Alu, Maiko Kaizu, Kohji Hotta, Takeo Horie, Yasunori Sasakura, Kazuho Ikeo, and Kazuo Inaba (2010). "Distribution and structural diversity of cilia in tadpole larvae of the ascidian Ciona intestinalis". *Developmental Biology* 337.1, pp. 42–62. ISSN: 00121606. DOI: 10.1016/j.ydbio.2009.10.012.

Koonce, M P, P M Grissom, and J R McIntosh (1992). "Dynein from Dictyostelium: primary structure comparisons between a cytoplasmic motor enzyme and flagellar dynein." *The Journal of cell biology* 119.6, pp. 1597–604. ISSN: 0021-9525.

Koonin, Eugene V (2010). "The incredible expanding ancestor of eukaryotes". *Cell* 140.5, pp. 606–608.

Kyei, George B, Isabelle Vergne, Jennifer Chua, Esteban Roberts, James Harris, Jagath R Junutula, and Vojo Deretic (2006). "Rab14 is critical for maintenance of Mycobacterium tuberculosis phagosome maturation arrest". *The EMBO Journal* 25.22, pp. 5250–5259.

Lal, Kalpana, Mark C Field, Jane M Carlton, Jim Warwicker, and Robert P Hirt (2005). "Identification of a very large Rab GTPase family in the parasitic protozoan Trichomonas vaginalis". *Molecular and Biochemical Parasitology* 143.2, pp. 226–235.

Larson, Stephen D, Lisa L Fong, Amarnath Gupta, Christopher Condit, William J Bug, and Maryann E Martone (2007). "A formal ontology of subcellular neuroanatomy." *Frontiers in neuroinformatics* 1.November, p. 3. ISSN: 1662-5196. DOI: 10.3389/neuro.11.003.2007.

Lee, R. Y N and Paul W. Sternberg (2003). "Building a cell and anatomy ontology of Caenorhabditis elegans". *Comparative and Functional Genomics* 4.1, pp. 121–126. ISSN: 15316912. DOI: 10.1002/cfg.248.

Leeuw, H P de, P M Koster, J Calafat, H Janssen, A J van Zonneveld, J A van Mourik, and J Voorberg (1998). "Small GTP-binding proteins in human endothelial cells". *British journal of haematology* 103.1, pp. 15–19.

Li, Jin Billy et al. (2004). "Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene." *Cell* 117.4, pp. 541–52. ISSN: 0092-8674.

Li, Li, Christian J. Stoeckert, and David S Roos (2003). "OrthoMCL: Identification of ortholog groups for eukaryotic genomes". *Genome Research* 13, pp. 2178–2189. ISSN: 10889051. DOI: 10.1101/gr.1224503.

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

Lin, Tzu-Wen, Jian-Wei Wu, and Darby Tien-Hao Chang (2013). "Combining phylogenetic profiling-based and machine learning-based techniques to predict functional related proteins." *PloS one* 8.9, e75940. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0075940.

Lingner, Thomas, Stefanie Mühlhausen, Toni Gabaldón, Cedric Notredame, and Peter Meinicke (2010). "Predicting phenotypic traits of prokaryotes from protein domain frequencies." *BMC bioinformatics* 11, p. 481. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-481.

Longcore, Joyce E, Allan P Pessier, and Donald K Nichols (1999). "Batrachochytrium Dendrobatidis gen. et sp. nov., a Chytrid Pathogenic to Amphibians". *Mycologia* 91.2, pp. 219–227. ISSN: 00275514. DOI: 10.2307/3761366.

Lynch, Michael, Mark C Field, Holly V Goodson, Harmit S Malik, José B Pereira-Leal, David S Roos, Aaron P Turkewitz, and Shelley Sazer (2014). "Evolutionary cell biology: Two origins, one objective". 111.48. DOI: 10.1073/pnas.1415861111.

Makarova, Kira S, Yuri I Wolf, and Eugene V Koonin (2003). "Potential genomic determinants of hyperthermophily." *Trends in genetics : TIG* 19.4, pp. 172–6. ISSN: 0168-9525. DOI: 10.1016/S0168-9525(03)00047-7.

Manton, I. and B. Clarke (1952). "An electron microscope study of the spermatozoid of sphagnum". *Journal of Experimental Botany* 3.3, pp. 265–275. ISSN: 00220957. DOI: 10.1093/jxb/3.3.265.

Manton, I, K Kowallik, H. A. von Stosch, and H a Von Stosch (1969). "Observations on the fine structure and development of the spindle at mitosis and meiosis in a marine centric diatom (Lithodesmium undulatum). I. Preliminary survey of mitosis in spermatogonia." *Journal of microscopy* 89.3, pp. 295–320. ISSN: 00222720. DOI: 10.1111/j.1365-2818.1969.tb00678.x.

Manton, I, K Kowallik, and H. A. von Stosch (1970). "Observations on the fine structure and development of the spindle at mitosis and meiosis in a marine centric diatom (Lithodesmium undulatum). 3. The later stages of meiosis I in male gametogenesis." *J Cell Sci* 6.1, pp. 131–157.

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

Marcotte, E M, I Xenarios, a M van Der Bliek, and D Eisenberg (2000). "Localizing proteins in the cell from their phylogenetic profiles." *Proceedings of the National Academy of Sciences of the United States of America* 97.22, pp. 12115–20. ISSN: 0027-8424. DOI: 10.1073/pnas.220399497.

Margulis, Lynn (1980). "Undulipodia, flagella and cilia". *Biosystems* 12.1-2, pp. 105–108. ISSN: 03032647. DOI: 10.1016/0303-2647(80)90041-6.

Marques, Ana Claudia, Nicolas Vinckenbosch, David Brawand, and Henrik Kaessmann (2008). "Functional diversification of duplicate genes through subcellular adaptation of encoded proteins". *Genome Biology* 9.3, R54.

Marshall, Wallace F et al. (2012). "What determines cell size?" *BMC biology* 10, p. 101. ISSN: 1741-7007. DOI: 10.1186/1741-7007-10-101.

Martone, Maryann E, Amarnath Gupta, Mona Wong, Xufei Qian, Gina Sosinsky, Bertram Ludäscher, and Mark H Ellisman (2002). "A cell-centered database for electron tomographic data." *Journal of structural biology* 138.1-2, pp. 145–55. ISSN: 1047-8477.

Masuda, E S et al. (2000). "Rab37 is a novel mast cell specific GTPase localized to secretory granules". *FEBS Letters* 470.1, pp. 61–64.

Maynard Smith, John and Eörs Szathmáry (1995). *The Major Transitions in Evolution*. Oxford, England: Oxford University Press. ISBN: 0-19-850294-X.

Mazumder, Raja, Darren a Natale, Jessica Anne Ecalnir Julio, Lai-Su Yeh, and Cathy H Wu (2010). "Community annotation in biology." *Biology direct* 5, p. 12. ISSN: 1745-6150. DOI: 10.1186/1745-6150-5-12.

Mazzarello, P (1999). "A unifying concept: the history of cell theory." *Nature cell biology* 1.1, E13–E15. ISSN: 1465-7392. DOI: 10.1038/8964.

Mencarelli, C., P. Lupetti, and R. Dallai (2008). *Chapter 4 New Insights into the Cell Biology of Insect Axonemes*. DOI: 10.1016/S1937-6448(08)00804-6.

Mesa, R, C Salomón, M Roggero, P D Stahl, and L S Mayorga (2001). "Rab22a affects the morphology and function of the endocytic pathway". *Journal of Cell Science* 114.22, pp. 4041–4049.

Mezey, Jason G. and David Houle (2005). "THE DIMENSIONALITY OF GENETIC VARIATION FOR WING SHAPE IN DROSOPHILA MELANOGASTER". *Evolution* 59.5, p. 1027. ISSN: 0014-3820. DOI: 10.1554/04-491.

Miserey-Lenkei, Stéphanie, G Chalancon, Sabine Bardin, E Formstecher, Bruno Goud, and Arnaud Echard (2010). "Rab and actomyosin-dependent fission of transport vesicles at the Golgi complex". *Nature Cell Biology* 12.7, pp. 645–654.

Mitchell, David R (2007). "The evolution of eukaryotic cilia and flagella as motile and sensory organelles." *Advances in experimental medicine and biology* 607, pp. 130–40. ISSN: 0065-2598. DOI: 10.1007/978-0-387-74021-8_11.

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

Moore, Jason H. (2007). *Bioinformatics*. DOI: 10.1002/jcp.21218.

Morgan, T. H. (1901). "Regeneration of Proportionate Structures in Stentor". *Biological Bulletin* 2.6, p. 311. ISSN: 00063185. DOI: 10.2307/1535709.

Moser, J W and G L Kreitner (1970). "Centrosome structure in Anthoceros laevis and Marchantia polymorpha." *The Journal of cell biology* 44.2, pp. 454–8. ISSN: 0021-9525.

Mowbrey, Kevin and Joel B. Dacks (2009). "Evolution and diversity of the Golgi body". *FEBS Letters* 583.23, pp. 3738–3745. ISSN: 00145793. DOI: 10.1016/j.febslet.2009.10.025.

Moya, Andrés, Juli Peretó, Rosario Gil, and Amparo Latorre (2008). "Learning how to live together: genomic insights into prokaryote-animal symbioses". *Nature Reviews Genetics* 9.3, pp. 218–229.

Mungall, Chris, Melissa Haendel, Georgios Gkoutos, and Suzanna Lewis (2009). *Uberon: towards a comprehensive multi-species anatomy ontology*. DOI: 10.1038/npre.2009.3592.1.

Mungall, Christopher J, Georgios V Gkoutos, Cynthia L Smith, Melissa a Haendel, Suzanna E Lewis, and Michael Ashburner (2010). "Integrating phenotype ontologies across multiple species." *Genome biology* 11.1, R2. ISSN: 1465-6906. DOI: 10.1186/gb-2010-11-1-r2.

Mungall, Christopher J, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, and Melissa a Haendel (2012). "Uberon, an integrative multi-species anatomy ontology." *Genome biology* 13.1, R5. ISSN: 1465-6914. DOI: 10.1186/gb-2012-13-1-r5.

Neerincx, Pieter B T and Jack a M Leunissen (2005). "Evolution of web services in bioinformatics." *Briefings in bioinformatics* 6.2, pp. 178–188. ISSN: 1467-5463. DOI: 10.1093/bib/6.2.178.

O'Leary, Maureen a. and Seth Kaufman (2011). "MorphoBank: Phylophenomics in the "cloud"". *Cladistics* 27, pp. 529–537. ISSN: 07483007. DOI: 10.1111/j.1096-0031.2011.00355.x.

Olson, L. W. and M. S. Fuller (1968). "Ultrastructural evidence for the biflagellate origin of the uniflagellate fungal zoospore". *Archiv fur Mikrobiologie* 62.3, pp. 237–250. ISSN: 03028933. DOI: 10.1007/BF00413894.

Orloff, David N, Janet H Iwasa, Maryann E Martone, Mark H Ellisman, and Caroline M Kane (2013). "The cell: an image library-CCDB: a curated repository of microscopy data." *Nucleic acids research* 41.Database issue, pp. D1241–50. ISSN: 1362-4962. DOI: 10.1093/nar/gks1257.

Ostrowski, L. E. (2002). "A Proteomic Analysis of Human Cilia: Identification of Novel Components". *Molecular & Cellular Proteomics* 1.6, pp. 451–465. ISSN: 15359476. DOI: 10.1074/mcp.M200037-MCP200.

Osumi-Sutherland, David, Steven J Marygold, Gillian H Millburn, Peter a McQuilton, Laura Ponting, Raymund Stefancsik, Kathleen Falls, Nicholas H Brown, and Georgios V Gkoutos (2013). "The Drosophila phenotype ontology." *Journal of biomedical semantics* 4, p. 30. ISSN: 2041-1480. DOI: 10.1186/2041-1480-4-30.

Parr, Cynthia S et al. (2014a). "The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth." *Biodiversity Data Journal* 2, e1079. ISSN: 1314-2828. DOI: 10.3897/BDJ.2.e1079.

Parr, Cynthia S., Nathan Wilson, Katja Schulz, Patrick Leary, Jennifer Hammock, Jeremy Rice, and Robert J. Corrigan Jr. (2014b). "TraitBank: Practical semantics for organism attribute data".

Pavlicev, Mihaela, Günter P. Wagner, and James M. Cheverud (2009). "Measuring Evolutionary Constraints Through the Dimensionality of the Phenotype: Adjusted Bootstrap Method to Estimate Rank of Phenotypic Covariance Matrices". *Evolutionary Biology* 36.3, pp. 339–353. ISSN: 0071-3260. DOI: 10.1007/s11692-009-9066-7.

Pazour, Gregory J, Nathan Agrin, John Leszyk, and George B Witman (2005). "Proteomic analysis of a eukaryotic cilium." *The Journal of cell biology* 170.1, pp. 103–13. ISSN: 0021-9525. DOI: 10.1083/jcb.200504008.

Pelkmans, Lucas, Thomas Bürli, Marino Zerial, and Ari Helenius (2004). "Caveolin-stabilized membrane domains as multifunctional transport and sorting devices in endocytic membrane traffic". *Cell* 118.6, pp. 767–780.

Pellegrini, M, E M Marcotte, M J Thompson, D Eisenberg, and T O Yeates (1999). "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." *Proceedings of the National Academy of Sciences of the United States of America* 96.8, pp. 4285–8. ISSN: 0027-8424.

Pelletier, Laurence, Eileen O'Toole, Anne Schwager, Anthony a Hyman, and Thomas Müller-Reichert (2006). "Centriole assembly in Caenorhabditis elegans." *Nature* 444.7119, pp. 619–23. ISSN: 1476-4687. DOI: 10.1038/nature05318.

Pellinen, Teijo, Antti Arjonen, Karoliina Vuoriluoto, Katja Kallio, Jack A M Fransen, and Johanna Ivaska (2006). "Small GTPase Rab21 regulates cell adhesion and controls endosomal traffic of beta1-integrins". *The Journal of Cell Biology* 173.5, pp. 767–780.

Pereira-Leal, J B and M C Seabra (2001). "Evolution of the Rab family of small GTP-binding proteins". *Journal of Molecular Biology* 313.4, pp. 889–901.

Pereira-Leal, José B (2008). "The Ypt/Rab family and the evolution of trafficking in fungi". *Traffic* 9.1, pp. 27–38.

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

Perkins, L A, E M Hedgecock, J N Thomson, and J G Culotti (1986). "Mutant sensory cilia in the nematode Caenorhabditis elegans." *Developmental biology* 117.2, pp. 456–487. ISSN: 00121606. DOI: 10.1016/0012-1606(86)90314-3.

Pigliucci, I and I Kaplan (2000). "The fall and rise of Dr Pangloss: adaptationism and the Spandrels paper 20 years later." *Trends in ecology & evolution* 15.2, pp. 66–70. ISSN: 0169-5347.

Pigliucci, Massimo (2007). "Finding the way in phenotypic space: the origin and maintenance of constraints on organismal form." *Annals of botany* 100.3, pp. 433–8. ISSN: 0305-7364. DOI: 10.1093/aob/mcm069.

Ponting, Chris P (2008). "The functional repertoires of metazoan genomes". *Nature Reviews Genetics* 9.9, pp. 689–698.

Poteryaev, Dmitry, Sunando Datta, Karin Ackema, Marino Zerial, and Anne Spang (2010). "Identification of the switch in early-to-late endosome transition". *Cell* 141.3, pp. 497–508.

Powelka, Aimee M, Jianlan Sun, Jian Li, Minggeng Gao, Leslie M Shaw, Arnoud Sonnenberg, and Victor W Hsu (2004). "Stimulation-dependent recycling of integrin $\beta$1 regulated by ARF6 and Rab11". *Traffic* 5.1, pp. 20–36.

Powell, MJ (1980). "Mitosis in the aquatic fungus Rhizophydium spherotheca (Chytridiales)". *American Journal of Botany* 67.6, pp. 839–853.

Proikas-Cezanne, Tassula, Anja Gaugel, Tancred Frickey, and Alfred Nordheim (2006). "Rab14 is part of the early endosomal clathrin-coated TGN microdomain". *FEBS Letters* 580.22, pp. 5241–5246.

Pushker, Ravindra, Alex Mira, and Francisco Rodríguez-Valera (2004). "Comparative genomics of gene-family size in closely related bacteria". *Genome Biology* 5.4, R27.

Qin, Hongmin, Zhaohui Wang, Dennis Diener, and Joel Rosenbaum (2007). "Intraflagellar transport protein 27 is a small G protein involved in cell-cycle control". *Current Biology* 17.3, pp. 193–202.

Quevillon, Emmanuel, Tobias Spielmann, Karima Brahimi, Debasish Chattopadhyay, Edouard Yeramian, and Gordon Langsley (2003). "The Plasmodium falciparum family of Rab GTPases". *Gene* 306, pp. 13–25.

Ramírez, Martín J et al. (2007). "Linking of digital images to phylogenetic data matrices using a morphological ontology." *Systematic biology* 56.2, pp. 283–94. ISSN: 1063-5157. DOI: 10.1080/10635150701313848.

Raup, D M and A Michelson (1965). "Theoretical Morphology of the Coiled Shell." *Science (New York, N.Y.)* 147.3663, pp. 1294–5. ISSN: 0036-8075. DOI: 10.1126/science.147.3663.1294.

Raup, DM (1966). "Geometric analysis of shell coiling: general problems". *Journal of Paleontology*.

Remm, M, C E Storm, and E L Sonnhammer (2001). "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons." *Journal of molecular biology* 314, pp. 1041–1052. ISSN: 0022-2836. DOI: 10.1006/jmbi.2000.5197.

Roberts, M, S Barry, A Woods, P van der Sluijs, and J Norman (2001). "PDGF-regulated rab4-dependent recycling of alphavbeta3 integrin from early endosomes is necessary for cell adhesion and spreading". *Current Biology* 11.18, pp. 1392–1402.

Robinow, C. F. (1966). "A FIBER APPARATUS IN THE NUCLEUS OF THE YEAST CELL". *The Journal of Cell Biology* 29.1, pp. 129–151. ISSN: 0021-9525. DOI: 10.1083/jcb.29.1.129.

Rodriguez-Gabin, A G, M Cammer, G Almazan, M Charron, and J N Larocca (2001). "Role of rRAB22b, an oligodendrocyte protein, in regulation of transport of vesicles from trans Golgi to endocytic compartments". *Journal of neuroscience research* 66.6, pp. 1149–1160.

Rokas, Antonis (2008). "The molecular origins of multicellular transitions". *Current Opinion in Genetics & Development* 18.6, pp. 472–478.

Roncaglia, Paola, Maryann E Martone, David P Hill, Tanya Z Berardini, Rebecca E Foulger, Fahim T Imam, Harold Drabkin, Christopher J Mungall, and Jane Lomax (2013). "The Gene Ontology (GO) Cellular Component Ontology: integration with SAO (Subcellular Anatomy Ontology) and other recent developments." *Journal of biomedical semantics* 4.1, p. 20. ISSN: 2041-1480. DOI: 10.1186/2041-1480-4-20.

Roos, U P (1975). "Mitosis in the cellular slime mold Polysphondylium violaceum." *The Journal of cell biology* 64.2, pp. 480–91. ISSN: 0021-9525.

Ropstorf, P, N Hulsmann, and K Hausmann (1994). "Comparative Fine-Structural Investigations of Interphase and Mitotic Nuclei of Vampyrellid Filose Amebas". *Journal of Eukaryotic Microbiology* 41, 18–30 ST – Comparative Fine–Structural Investigat. ISSN: 1066-5234. DOI: 10.1111/j.1550-7408.1994.tb05930.x.

Ruano-Rubio, Valentín, Olivier Poch, and Julie D Thompson (2009). "Comparison of eukaryotic phylogenetic profiling approaches using species tree aware methods." *BMC bioinformatics* 10, p. 383. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-383.

Rupnik, M et al. (2007). "Distinct role of Rab3A and Rab3B in secretory activity of rat melanotrophs". *American Journal of Physiology, Cell Physiology* 292.1, pp. C98–105.

Rutherford, Stephen and Ian Moore (2002). "The Arabidopsis Rab GTPase family: another enigma variation". *Current Opinion in Plant Biology* 5.6, pp. 518–528.

Saito-Nakano, Yumiko, Brendan J Loftus, Neil Hall, and Tomoyoshi Nozaki (2005). "The diversity of Rab GTPases in Entamoeba histolytica". *Experimental parasitology* 110.3, pp. 244–252.

Saito-Nakano, Yumiko, Tohru Nakahara, Kentaro Nakano, Tomoyoshi Nozaki, and Osamu Numata (2010). "Marked amplification and diversification of products of ras genes from rat brain, Rab GTPases, in the ciliates Tetrahymena thermophila and Paramecium tetraurelia". *The Journal of Eukaryotic Microbiology* 57.5, pp. 389–399.

Saito, Taro L, Miwaka Ohtani, Hiroshi Sawai, Fumi Sano, Ayaka Saka, Daisuke Watanabe, Masashi Yukawa, Yoshikazu Ohya, and Shinichi Morishita (2004). "SCMD: Saccharomyces cerevisiae Morphological Database." *Nucleic acids research* 32.Database issue, pp. D319–22. ISSN: 1362-4962. DOI: 10.1093/nar/gkh113.

Sanders, M. A. (1989). "Centrin-mediated microtubule severing during flagellar excision in Chlamydomonas reinhardtii". *The Journal of Cell Biology* 108.5, pp. 1751–1760. ISSN: 0021-9525. DOI: 10.1083/jcb.108.5.1751.

Sankoff, David (1975). "Minimal Mutation Trees of Sequences". *SIAM Journal on Applied Mathematics* 28.1, pp. 35–42. ISSN: 0036-1399. DOI: 10.1137/0128004.

Sansom, Roger (2008). "The nature of developmental constraints and the difference-maker argument for externalism". *Biology & Philosophy* 24.4, pp. 441–459. ISSN: 0169-3867. DOI: 10.1007/s10539-008-9121-2.

Scheer, Ulrich (2014). "Historical roots of centrosome research: discovery of Boveri's microscope slides in Würzburg." *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 369.1650. ISSN: 1471-2970. DOI: 10.1098/rstb.2013.0469.

Schlüter, Agatha, Stéphane Fourcade, Raymond Ripp, Jean Louis Mandel, Olivier Poch, and Aurora Pujol (2006). "The evolutionary origin of peroxisomes: An ER-peroxisome connection". *Molecular Biology and Evolution* 23.4, pp. 838–845. ISSN: 07374038. DOI: 10.1093/molbev/msj103.

Schlüter, Agatha, Alejandro Real-Chicharro, Toni Gabaldón, Francisca Sánchez-Jiménez, and Aurora Pujol (2009). "PeroxisomeDB 2.0: An integrative view of the global peroxisomal metabolome". *Nucleic Acids Research* 38.SUPPL.1, pp. 800–805. ISSN: 03051048. DOI: 10.1093/nar/gkp935.

Schlüter, Oliver M, Mikhail V Khvotchev, Reinhard Jahn, and Thomas C Südhof (2002). "Localization versus function of Rab3 proteins—Evidence for a common regulatory role in controlling fusion". *The Journal of biological chemistry* 277.43, pp. 40919–40929.

Schrevel, Joseph and C Besse (1975). "A functional flagella with a 6 + 0 pattern". *The Journal of cell biology* 66.3, pp. 492–507. ISSN: 0021-9525. DOI: 10.1083/jcb.66.3.492.

Schuck, Sebastian, Mathias J Gerl, Agnes Ang, Aki Manninen, Patrick Keller, Ira Mellman, and Kai Simons (2007). "Rab10 is involved in basolateral transport in polarized Madin-Darby canine kidney cells". *Traffic* 8.1, pp. 47–60.

Schwartz, Samantha L, Canhong Cao, Olena Pylypenko, Alexey Rak, and Angela Wandinger-Ness (2007). "Rab GTPases at a glance". *Journal of Cell Science* 120.Pt 22, pp. 3905–3910.

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

Segerdell, Erik, Jeff B Bowes, Nicolas Pollet, and Peter D Vize (2008). "An ontology for Xenopus anatomy and development." *BMC developmental biology* 8, p. 92. ISSN: 1471-213X. DOI: 10.1186/1471-213X-8-92.

Shalchian-Tabrizi, Kamran, Marianne A Minge, Mari Espelund, Russell Orr, Torgeir Ruden, Kjetill S Jakobsen, and Thomas Cavalier-Smith (2008). "Multigene phylogeny of choanozoa and the origin of animals". *PLoS ONE* 3.5, e2098.

Shanahan, Timothy (2008). "Why don't zebras have machine guns? Adaptation, selection, and constraints in evolutionary theory." *Studies in history and philosophy of biological and biomedical sciences* 39.1, pp. 135–46. ISSN: 1369-8486. DOI: 10.1016/j.shpsc.2007.12.008.

Singh, Saurav and Dennis P DP Wall (2008). "Testing the accuracy of eukaryotic phylogenetic profiles for prediction of biological function." *Evolutionary bioinformatics online* 4, pp. 217–23. ISSN: 1176-9343.

Singla, Veena and Jeremy F Reiter (2006). "The primary cilium as the cell's antenna: signaling at a sensory organelle." *Science (New York, N.Y.)* 313.5787, pp. 629–633. ISSN: 0036-8075. DOI: 10.1126/science.1124534.

Sinka, Rita, Alison K Gillingham, Vangelis Kondylis, and Sean Munro (2008). "Golgi coiled-coil proteins contain multiple binding sites for Rab family G proteins". *The Journal of Cell Biology* 183.4, pp. 607–615.

Slonim, Noam, Olivier Elemento, and Saeed Tavazoie (2006). "Ab initio genotype-phenotype association reveals intrinsic modularity in genetic networks." *Molecular systems biology* 2, p. 2006.0005. ISSN: 1744-4292. DOI: 10.1038/msb4100047.

Smith, Barry et al. (2007). "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration." *Nature biotechnology* 25.11, pp. 1251–5. ISSN: 1087-0156. DOI: 10.1038/nbt1346.

Smith, J. Maynard, R. Burian, S. Kauffman, P. Alberch, J. Campbell, B. Goodwin, R. Lande, D. Raup, and L. Wolpert (1985). "Developmental Constraints and Evolution: A Perspective from the Mountain Lake Conference on Development and Evolution". *The Quarterly Review of Biology* 60.3, p. 265. ISSN: 0033-5770. DOI: 10.1086/414425.

Smith, Jeffrey C, Julian G B Northey, Jyoti Garg, Ronald E Pearlman, and K W Michael Siu (2005). "Robust method for proteome analysis by MS/MS using an entire translated genome: demonstration on the ciliome of Tetrahymena thermophila." *Journal of proteome research* 4.3, pp. 909–19. ISSN: 1535-3893. DOI: 10.1021/pr050013h.

Smith, John Maynard and Eörs Szathmáry (1997). *The Major Transitions in Evolution*. Oxford University Press.

Sorokin, Sergei (1962). "SERGEI SOROKIN, M.D. From the Department of Anatomy, Harvard Medical School, Boston, Massachusetts". 10.

Spang, Anja et al. (2015). "Complex archaea that bridge the gap between prokaryotes and eukaryotes". *Nature*. ISSN: 0028-0836. DOI: 10.1038/nature14447.

Sprague, Judy et al. (2006). "The Zebrafish Information Network: the zebrafish model organism database." *Nucleic acids research* 34.Database issue, pp. D581–5. ISSN: 1362-4962. DOI: 10.1093/nar/gkj086.

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

St Pierre, Susan E, Laura Ponting, Raymund Stefancsik, Peter McQuilton, Susan E. St. Pierre, Laura Ponting, Raymund Stefancsik, and Peter McQuilton (2014). "FlyBase 102–advanced approaches to interrogating FlyBase." *Nucleic acids research* 42.Database issue, pp. D780–8. ISSN: 1362-4962. DOI: 10.1093/nar/gkt1092.

Stenmark, Harald (2009). "Rab GTPases as coordinators of vesicle traffic". *Nature Reviews Molecular Cell Biology* 10.8, pp. 513–525.

Sugden, A and E Pennisi (2000). *Diversity digitized.* DOI: 10.1126/science.289.5488.2305.

Sun, Jingchun, Jinlin Xu, Zhen Liu, Qi Liu, Aimin Zhao, Tieliu Shi, and Yixue Li (2005). "Refined phylogenetic profiles method for predicting protein-protein interactions." *Bioinformatics (Oxford, England)* 21.16, pp. 3409–15. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti532.

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

Tippit, David H., J. D. Pickett Heaps, and Jeremy D. Pickett-Heaps (1977). "Mitosis in the pennate diatom Surirella ovalis". *Journal of Cell Biology* 73.3, pp. 705–727. ISSN: 00219525. DOI: 10.1083/jcb.73.3.705.

Tolmachova, Tanya, Magnus Abrink, Clare E Futter, Kalwant S Authi, and Miguel C Seabra (2007). "Rab27b regulates number and secretion of platelet dense granules". *Proceedings of the National Academy of Sciences of the United States of America* 104.14, pp. 5872–5877.

Tsuboi, Takashi and Mitsunori Fukuda (2006). "Rab3A and Rab27A cooperatively regulate the docking step of dense-core vesicle exocytosis in PC12 cells". *Journal of Cell Science* 119.11, pp. 2196–2203.

Ueda, M, M Schliwa, and U Euteneuer (1999). "Unusual centrosome cycle in Dictyostelium: correlation of dynamic behavior and structural changes." *Molecular biology of the cell* 10.1, pp. 151–60. ISSN: 1059-1524.

Valencia, Alfonso, P Chardin, Alfred Wittinghofer, and Chris Sander (1991). "The Ras protein family: evolutionary tree and role of conserved amino acids". *Biochemistry* 30.19, pp. 4637–4648.

Valsdottir, R, H Hashimoto, K Ashman, T Koda, B Storrie, and Tommy Nilsson (2001). "Identification of rabaptin-5, rabex-5, and GM130 as putative effectors of rab33b, a regulator of retrograde traffic between the Golgi apparatus and ER". *FEBS Letters* 508.2, pp. 201–209.

Vogel, Christine and Cyrus Chothia (2006). "Protein family expansions and biological complexity". *PLoS Computational Biology* 2.5, e48.

Vogt, Lars (2008). "Learning from Linnaeus : towards developing the foundation for a general structure concept for morphology". *Zootaxa* 152, pp. 123 –152.

Vogt, Lars, Thomas Bartolomaeus, and Gonzalo Giribet (2009). "The linguistic problem of morphology: structure versus homology and the standardization of morphological data". *Cladistics* 26.3, pp. 301–325. ISSN: 07483007. DOI: 10.1111/j.1096-0031.2009.00286.x.

Wasmeier, Christina, Maryse Romao, Lynn Plowright, Dorothy C Bennett, Graça Raposo, and Miguel C Seabra (2006). "Rab38 and Rab32 control post-Golgi trafficking of melanogenic enzymes." *The Journal of Cell Biology* 175.2, pp. 271–281.

Waterhouse, Andrew M, James B Procter, David M A Martin, Michele Clamp, and Geoffrey J Barton (2009). "Jalview Version 2—a multiple sequence alignment editor and analysis workbench". *Bioinformatics* 25.9, pp. 1189–1191.

Weber, Stefan S, Curdin Ragaz, and Hubert Hilbi (2009). "Pathogen trafficking pathways and host phosphoinositide metabolism". *Molecular Microbiology* 71.6, pp. 1341–1352.

Wennerberg, Krister, Kent L Rossman, and Channing J Der (2005). "The Ras Superfamily at a Glance". *Journal of Cell Science* 118.5, pp. 843–846.

Wheatley, Denys N (2005). "Landmarks in the first hundred years of primary (9+0) cilium research." *Cell biology international* 29.5, pp. 333–9. ISSN: 1065-6995. DOI: 10.1016/j.cellbi.2005.03.001.

Wilson, Derek, Ralph Pethica, Yiduo Zhou, Charles Talbot, Christine Vogel, Martin Madera, Cyrus Chothia, and Julian Gough (2009a).

"SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny". *Nucleic Acids Research* 37.Database issue, pp. D380–6.

Wilson, Derek, Ralph Pethica, Yiduo Zhou, Charles Talbot, Christine Vogel, Martin Madera, Cyrus Chothia, and Julian Gough (2009b). "SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny." *Nucleic acids research* 37.Database issue, pp. D380–6. ISSN: 1362-4962. DOI: 10.1093/nar/gkn762.

Wood, R L (1979). "The fine structure of the hypostome and mouth of hydra. II. Transmission electron microscopy." *Cell and tissue research* 199.2, pp. 319–338. ISSN: 0302766X.

Yoder, Matthew J., István Mikó, Katja C. Seltmann, Matthew A. Bertone, and Andrew R. Deans (2010). "A gross anatomy ontology for hymenoptera". *PLoS ONE* 5.12, pp. 1435–1439. ISSN: 19326203. DOI: 10.1371/journal.pone.0015991.

Yoon, Hwan Su, Jeremiah D. Hackett, Claudia Ciniglia, Gabriele Pinto, and Debashish Bhattacharya (2004). "A Molecular Timeline for the Origin of Photosynthetic Eukaryotes". *Molecular Biology and Evolution* 21.5, pp. 809–818. ISSN: 07374038. DOI: 10.1093/molbev/msh075.

Yoshie, S, A Imai, T Nashida, and H Shimomura (2000). "Expression, characterization, and localization of Rab26, a low molecular weight GTP-binding protein, in the rat parotid gland". *Histochemistry and cell biology* 113.4, pp. 259–263.

Young, Joanne, Julie Ménétrey, and Bruno Goud (2010). "RAB6C is a retrogene that encodes a centrosomal protein involved in cell cycle progression". *Journal of Molecular Biology* 397.1, pp. 69–88.

Zhou, Yun, Rui Wang, Li Li, Xuefeng Xia, and Zhirong Sun (2006). "Inferring Functional Linkages between Proteins from Evolutionary Scenarios". *Journal of Molecular Biology* 359, pp. 1150–1159. ISSN: 00222836. DOI: 10.1016/j.jmb.2006.04.011.

digimorph (2015). *DigiMorph*. URL: www.digimorph.org.

ema (2015). *E Mouse Atlas*. URL: http://www.emouseatlas.org.

eol (2015). *The Encyclopedia of Life*. URL: http://www.eol.org.

library, cell image (2015). *The Cell: Image Library*. URL: http://www.cellimagelibrary.org/.

morphDbase (2015). *MorphDBase*. URL: www.morphdbase.de.

morphbank (2015). *Morphbank :: Biological Imaging*. URL: http://www.morphbank.net.

morphobank (2015). *MorphoBank*. URL: www.morphobank.org.

pato (2015). *PATO homepage on Obofoundry*. URL: obofoundry.org/wiki/index.php/PATO:MainPage.

pipeline, orthomcl (2015). *Orthomcl Pipeline*. URL: https://github.com/apetkau/orthomcl-pipeline.

ITQB-UNL | Av. da República, 2780-157 Oeiras, Portugal
Tel (+351) 214 469 100 | Fax (+351) 214 411 277

www.itqb.unl.pt