Université de Montréal

Clustering algorithms and shape factor methods to discriminate among small GTPase phenotypes using DIC image analysis

par

Arturo Papaluca

Département de Biochimie
Faculté de Médecine

Mémoire présenté à la Faculté de Médecine
en vue de l'obtention du grade de Maître ès science (M.Sc.)
En Biochimie
Option Génétique Moléculaire

Juillet, 2012

Université de Montréal
Faculté des études supérieures et postdoctorales

Ce mémoire intitulé:

Clustering algorithms and shape factor methods to discriminate among small GTPase phenotypes using DIC image analysis

Présenté par :
Arturo Papaluca

A été évalué par un jury composé des personnes suivantes :

Pascal Chartrand, président-rapporteur
Stephen Michnick, directeur de recherche
Gerardo Ferbeyre, membre du jury

Résumé

Naïvement perçu, le processus d'évolution est une succession d'événements de duplication et de mutations graduelles dans le génome qui mènent à des changements dans les fonctions et les interactions du protéome. La famille des hydrolases de guanosine triphosphate (GTPases) similaire à Ras constitue un bon modèle de travail afin de comprendre ce phénomène fondamental, car cette famille de protéines contient un nombre limité d'éléments qui diffèrent en fonctionnalité et en interactions. Globalement, nous désirons comprendre comment les mutations singulières au niveau des GTPases affectent la morphologie des cellules ainsi que leur degré d'impact sur les populations asynchrones.

Mon travail de maîtrise vise à classifier de manière significative différents phénotypes de la levure *Saccaromyces cerevisiae* via l'analyse de plusieurs critères morphologiques de souches exprimant des GTPases mutées et natives. Notre approche à base de microscopie et d'analyses bioinformatique des images DIC (microscopie d'interférence différentielle de contraste) permet de distinguer les phénotypes propres aux cellules natives et aux mutants. L'emploi de cette méthode a permis une détection automatisée et une caractérisation des phénotypes mutants associés à la sur-expression de GTPases constitutivement actives. Les mutants de GTPases constitutivement actifs Cdc42 Q61L, Rho5 Q91H, Ras1 Q68L et Rsr1 G12V ont été analysés avec succès.

En effet, l'implémentation de différents algorithmes de partitionnement, permet d'analyser des données qui combinent les mesures morphologiques de population native et mutantes. Nos résultats démontrent que l'algorithme Fuzzy *C*-Means performe un partitionnement efficace des cellules natives ou mutantes, où les différents types de cellules sont classifiés en fonction de plusieurs facteurs de formes cellulaires obtenus à partir des images DIC. Cette analyse démontre que les mutations Cdc42 Q61L, Rho5 Q91H, Ras1 Q68L et Rsr1 G12V induisent respectivement des phénotypes amorphe, allongé, rond et large qui sont représentés par des vecteurs de facteurs de forme distincts. Ces distinctions

sont observées avec différentes proportions (morphologie mutante / morphologie native) dans les populations de mutants.

Le développement de nouvelles méthodes automatisées d'analyse morphologique des cellules natives et mutantes s'avère extrêmement utile pour l'étude de la famille des GTPases ainsi que des résidus spécifiques qui dictent leurs fonctions et réseau d'interaction. Nous pouvons maintenant envisager de produire des mutants de GTPases qui inversent leur fonction en ciblant des résidus divergents. La substitution fonctionnelle est ensuite détectée au niveau morphologique grâce à notre nouvelle stratégie quantitative. Ce type d'analyse peut également être transposé à d'autres familles de protéines et contribuer de manière significative au domaine de la biologie évolutive.

Mots-clés : évolution, petite GTPases Ras, morphologie cellulaire, fonction et structure des protéines, réseaux d'interaction protéine-protéine, facteurs de forme cellulaire, algorithmes de partitionnement.

Abstract

Evolution is a gradual process that gives rise to changes in the form of mutations that are reflected at the protein level. We propose that evolution of new pathways occurs by switching binding partners, hence creating new functions. The different functions encountered in a given family of related proteins have emerged from a common ancestor that has been duplicated and mutated to become implicated in new interactions and to gain new functions. In this study, we will use native and constitutive active mutant variants of the Ras-like family of small GTPases as working model, to explore such gene duplications, followed by neo / sub-functionalization. The reason for choosing this family resides in the fact that it is a defined set of proteins with well known functions that are mediated through multiple protein-protein interactions.

The aim of this master is to perform a classification of budding yeast phenotypes using different approaches in order to statistically determine at which level of the population these constitutively active mutations are capable to affect cell morphology. Working with a subset of the Ras-like small GTPases family, we recently developed an approach to catalogue and classify these proteins based on multiple physical and chemical criteria. Using microscopic and bioinformatics methods, we characterized phenotypes associated with over-expression of the native small GTPases of the budding yeast *Saccharomyces cerevisiae*, showing that an established classification is not very clear.

We are interested to investigate how point mutations in small GTPases can affect the cell morphology and their level of impact on asynchronous population. We want to establish a method to determine and quantify mutant and wild type-like phenotypes on these populations using Differential interference contrast microscopy (DIC) images only. As for the first aim of this study, we hypothesize that clustering algorithms can partition mutant cells from wild type cells based on cell shape factor measurements. To prove this hypothesis, we proposed to implement different clustering algorithms to analyze datasets which combines measurements from wild type and respective mutant populations.

We created constitutively active forms of these small GTPases and used Cdc42, Rho5, Ras1 and Rsr1 to validate our results. We observed that Cdc42 Q61L, Rho5 Q91H, Ras1 Q68L and Rsr1 G12V mutations induced characteristic amorphous, clumped/elongated, rounded and discrete large phenotypes respectively. This classification allowed us to define a phenotypical classification related to functions. Phenotype classification of the small GTPases has been confirmed using shape factor formulas accompanied with bioinformatics approaches. These approaches which involved different clustering methods allowed an automated quantitative characterization of the phenotypes of up to 7293 mutant cells.

Sequence alignment of Cdc42 and Rho5 showed 46.1% identity as well as 62.6% for Ras1 and Rsr1 allowing the identification of diverged residues potentially involved in specific functions and protein-protein interactions. Directed mutagenesis and substitution of these sites from one gene to another have been performed in some positions to test for specificity and involvement in morphology changes. In parallel, interactions observed for native and constitutively active mutants Cdc42 and Rho5 will be assayed with protein-fragment complementation assay (PCA). This will enable us to determine whether a high correlation exists between functions switches and binding partner's switches.

We propose to expand this approach to the whole Ras-like small GTPases family and monitor protein-protein interactions and functions at a network scale. This research will confirm whether enrichment or depletion of residues in specific sites induces a switch of function due to switching binding partners. Understanding the mechanism underlying such correlation is important to gain insight in the biological mechanisms underlying the Ras-like small GTPases and other proteins evolution. Such knowledge is of fundamental importance in biomedical and pharmaceutical fields, since Ras-like small GTPases represent important targets for therapeutic interventions and for the evolutionary biology field.

This thesis aims to:

1. Explore the evolution of the Ras-like small GTPases by looking at constitutively active mutants, switched mutants and their impact on the protein-protein interactions network.
2. Solve the issue of quantitative and statistically significant phenotypic analysis using only DIC (Differential interference contrast microscopy) images. We propose a measurement to quantitatively characterize different phenotypes of the budding yeast cells.
3. Apply different clustering methods on the proposed measurement to classify the budding yeast cells according to their phenotypes and choose the best approach to continue future research on DIC images quantification. These analyses are accomplished using the R statistical language.
4. Using the above approach to discriminate between wild type cells and those containing point mutations.

Keywords: Evolution, Ras-like small GTPases, cell morphology, protein function and structure, protein-protein interaction networks, shape factors, clustering algorithms.

Table of Content:

List of Tables:

List of Figures:

Dedication

A Papá y Mamá, quienes amo y me
aman sin dependencia y a mis
queridos hermanos

Acknowledgments

For those in Science and those about to Rock... I salute you!

To my beloved family, without your support, nothing would be possible. I am indebted in every possible way. Mamá & Papá for your undying love, support and care since I came to Canada (this one is for you and there is more to come in the future). To my brothers Oscar & Ulyses, I would like to thank you both for being excellent brothers, Abuela & Tía for whose continued support and prayers I am extremely grateful. My dogs Popi & Tita for being special and unique in the way they were, Ringo, Rocco, Uchy, Harri and Flopy for jumping in front of the camera while skyping! Martina the new member of the Papaluca family for your sweet and cheerful smile and punching the computer when I scream your name, Marcia Brun, thank you for all your love, patience and support (I will remember every day). To my dear friends, Fede Vega and Victor Gaete for long distance support and metallic brotherhood. To all, May the Force be with you.

With regards to my scientific career, the first person I want to thank is Professor Stephen Michnick, for accepting me into his lab and opening the door of science for me (un maestro, una causa, un efecto). You have influenced me tremendously to further continue the exploration and to understand life at the molecular level. My dear colleagues, from whom I learnt everything from scratch in order to perform my experiments and ask big questions! Relax guys, do the experiment! Specifically, I would like to thank Louis-Philippe "Calm down APZ" Bergeron-Sandoval for all the mentorship during this process and try to convert people to get free passports; Abdellali Kelil for the bioinformatics lessons; Benjamin Dubreuil for useful discussions and sharing ideas; Mohan "The mating expert" Malleshaiah for sharing so many good times until the end of his PhD; Durga Sivanesan for useful discussions and taking all my morning heavy jokes; The rest of the team; Diala Abd Rabbo, Jacqueline Kowarzyk, Emmanuelle Tchekanda, Luz Carrillo, Jean-

François Paradis, Emmanuel Levy, Po Hien Ear, Vincent Messier, Eugene Kanshin, Benjamin Ilunga Matala, Sinan Isik, Marcos Rodrigues and Bram Stynen.

Special thanks goes to Louis-Philippe, Abdellali, and Benjamin D. for all the bioinformatics support and useful comments, Durga, Jacqueline, and Diala for feedbacks and again Louis-Philippe for correcting the overall and for proving me with great feedback regarding this project, just relax mate! Cheers guys!!!

In addition, I would like to extend my appreciation and gratitude to Monique Vasseur for guidance and constant support with all my microscopy work. Also, I would like to acknowledge the members of my master thesis jury Dr. Pascal Chartrand and Dr. Gerardo Ferbeyre for accepting to review my work. Prof. Luc DesGroseillers and Sylvie Beauchemin for faculty procedural and administrative guidance to complete this research project.

1. Introduction

1.1. Preamble: Subject Situation

Pathways evolution and new gene functions emerging from genes duplication is of biological and medical importance. Understanding these phenomena could be interesting in certain domains such as drugs development and certain types of medical therapy. Also to target the wanted pathway or engineer a certain cell type in order to accomplish a specific function.

The Ras-like small GTPases family represent important targets for pathway activation and development, since they can regulate a wide variety of cellular functions. These proteins have been the focus of cancer research since the discovery of the isoform *H-Ras p21* mutant in human tumour cells [1]. This discovery highlighted the importance of the implication of proteins from the Ras-like small GTPases family in different types of diseases and different pathways.

Previous work accomplished by Heo *et al.* [2] showed that mutants with effective switch-of-phenotypes is related to functions. The authors developed an algorithm that predicts positions in a pair of protein classes that if exchanged will create mutants with switched functionality. They proved that specificity in a given protein family can be explored by combining genome-wide experimental functional classification with the creation of switch-of-function mutants. However, Heo *et al.* didn't explain how the change in morphology arises as a result of switch-of-function mutants, this observation raises several questions, such as: (i) Will the GTPase activity of the mutated proteins be conserved? (ii) If not how do these mutations affect their targets? (iii) Do these mutations affect the interaction network of the implicated proteins? (iv) How their phenotypes have been quantified in order to statistically differentiate between the various phenotypes?

Proposed questions regarding such observations: (i) Which method can be used or developed to quantify cell phenotypes? (ii) How many mutations are necessary following a gene duplication to exchange functions between two Ras-like small GTPases? (iii) Which residues are relevant for functional specificity in the same small GTPase sub-family? (iv)

What are the possible effects of mutations at the divergent positions? (v) Would the divergent positions give rise to new protein interaction partner connections in order to gain new function and develop a new phenotype?

The aim of this master thesis is to perform a classification of budding yeast phenotypes using different approaches in order to statistically determine at which level of the cellular population, these constitutively active mutations are capable to affect cell morphology. As a starting point, native and constitutively active mutants of small GTPases were over-expressed to see whether there is a change or not in cellular morphology. Using the Ras-like small GTPases family as a working model, an extensive sequence alignment was elaborated in order to create constitutively active mutants. The known human *H-Ras Q61L* constitutively active mutant has been used as a model to align with the rest of the small GTPases of *Homo sapiens, Mus musculus, Drosophila melanogaster, Oryzias latipes, Rattus norvegicus, Gallus gallus, Schizosaccharomyces pombe, Saccharomyces cerevisiae,* and other species. This showed conserved G1 and G3 domains among species. Such domains are known to be involved in giving a constitutive activity by blocking the regulation by GTPase-activating proteins (GAP). In fact, creation of constitutively active mutants of proteins and their over-expression can induce different morphological cell changes which make their classification feasible.

Using microscopic techniques, Shape factor formulas have been used to measure coarse aggregate for concrete using digital image processing [3]. The adaptation and development of these formulas to cellular biology, helped us to define and describe quantitatively cellular shapes related to different phenotypes using various parameters. The main advantage of using the shape factors as a phenotype measurement is that it can be efficiently and easily applied on a large amount of data. All the analysis of this study was accomplished using the R statistical language. This phenotype classification is the first step to further understand if a neo or sub-functionalization process is taking place in order to induce different types of phenotypes. At the same time, this reflects various protein behaviours involved in important cellular events. This phenotype classification is

the first step to further understand if a neo or sub-functionalization process is taking place in order to induce different types of phenotypes. At the same time, this reflects various protein behaviours involved in important cellular events.

The second aim of this study is to explain the neo/sub-functionalization processes of the Ras-like small GTPases by looking at phenotypes linked with functions, signalling processes and protein-protein interaction networks. However, these presented goals are out of the scope of this thesis.

In summary, our phenotype classification method allowed us to observe how proteins from a defined sub-family can have two different roles, belong to different interaction networks and can induce two different phenotypes indicating a possible sub-functionalization process between these proteins. In addition, our applied method allow automating large-scale phenotypes quantification using shape factors and are addressing the issue of direct DIC quantification incognita. Moreover, this approach is opening doors to explore how protein-protein interaction networks can evolve and understand how pathways ends up developing important functions inside the cell.

1.2. Meaning of Evolution

Evolution is the only scientific theory that describes the diversification of life. Evolution theory explains the outstanding similarities among extremely great different forms of life, the development of new life forms, and the changes that occur within populations through inherited traits. Inherited traits are particular distinguishing aspects, including anatomical, biochemical or behavioural characteristics that are passed from one generation to another. The major sources of such variation are mutations, genetic recombination and gene flow (Adapted from Nature Education Knowledge Project). Darwin's Theory *"Evolution is a vital process of life, and speciation is an evolutionary process by which new species comes to life"* (C. Darwin. 1859) gave to evolution a strong, meaningful and controversial concept

that has been defying religious beliefs and point of views by putting in doubt creationism. This concept of evolution is critical point to biology since it helps understand important processes and is the only scientific explanation of life's diversity among species. The theory of evolution has been called the cornerstone of modern biology since it provides a good basis for understanding science. As the prominent scientist Theodosius Dobzhansky stated *"Nothing in biology makes sense except in the light of evolution"* and without it, lots of biological events would remain unclear. Moreover understanding evolution is central for the advancement of biology and evolutionary medicine. In general words, by looking at the molecular point of view, evolution gives rise to changes and speciation which occurs in the form of mutation at the level of genetic sequences, also as a consequence often observed at the protein level. These concepts would lead scientist to develop tools in order to study at the genetic and molecular level, processes such as gene duplication, protein evolution, signalling and pathway development that has been the focus of research in the past decade. If we understand the evolution of a system we can target and engineer specific functions and we can dictate cells behaviour in one way or another and also understand the structure, function and mechanism of that system and we can make predictions about it.

1.3. Gene duplication & protein evolution

In protein evolution, gene duplication is considered to be the most important source of new DNA sequences which are in some cases devoid of selective pressure and other particular cases under selective pressure [4]. For instance, mutations and genomic modifications in these DNA sequence can lead to changes in sequences and switch domains among proteins respectively. These changes result in new protein structure, functions and abundance to just mention among others. It is interesting to know how two proteins from the same family can have totally different functions, how the same function can be conserve, and how they can share the burden of a determined function. Are these previous points affecting the interaction network? Are these previous observations pointing out on

how at the evolution of the protein-protein interaction network of system? These questions can be answered by first starting to understand mechanisms such as gene duplication, divergence and conservation which are the real phenomenon in charge of how eukaryotic signalling proteins can assemble a whole new set of functions and how this ends up implicating the whole network [5].

1.3.1. Molecular Mechanisms of Gene Duplication

The evolutionary biology field tried to address the following questions: (i) Can new genes emerge by duplication of existing ones and how do they acquire new functions after the duplication event? (ii) Can the new duplicate share functions with its ancestral form and which are the possible paths leading to this event? (iii) Which strategies can help us uncover and understand the consequences of gene duplications? It is worthwhile starting this chapter by briefly describing the different mechanisms causing genes duplication. We can cite mainly four major mechanisms orchestrating gene duplications in evolution (Figure 1) [6]:

A) Unequal Crossing-Over or Tandem Duplication: This phenomenon targets gene duplicates that are expected to have the same orientation but an unequal crossing-over. This first mechanism seems to be a common contributor of genetic material in duplicates.

B) Duplicative Transposition: Duplicative DNA Transposition is a mechanism accomplished by mainly two well known pathways: the non allelic homologous recombination (NAHR) and the non homologous end joining (NHEJ), where homologous sequences are used as templates during double-strand break repair.

C) Retrotranposition: In 1991, Brosius *et al.* showed that retrotransposed genes result from the reverse transcription of mRNA into complementary DNA which is then inserted into a new genomic position giving birth to a new duplicate [7]. Some molecular features

about retrotransposition are loss of introns, presence of poly A tracts and presence of flanking short tandem sequences which deviate from the common pattern of this mechanism [8].

D) Polyploidy: Whole-genome duplication result in the retention of two copies of a gene. This process can be considered as duplication event of every single gene of the entire genome including flanking regulatory sequences [6]. It is known that after polyploidization the genes functions are maintained in a different manner from those duplicated by different mechanisms [9, 10].

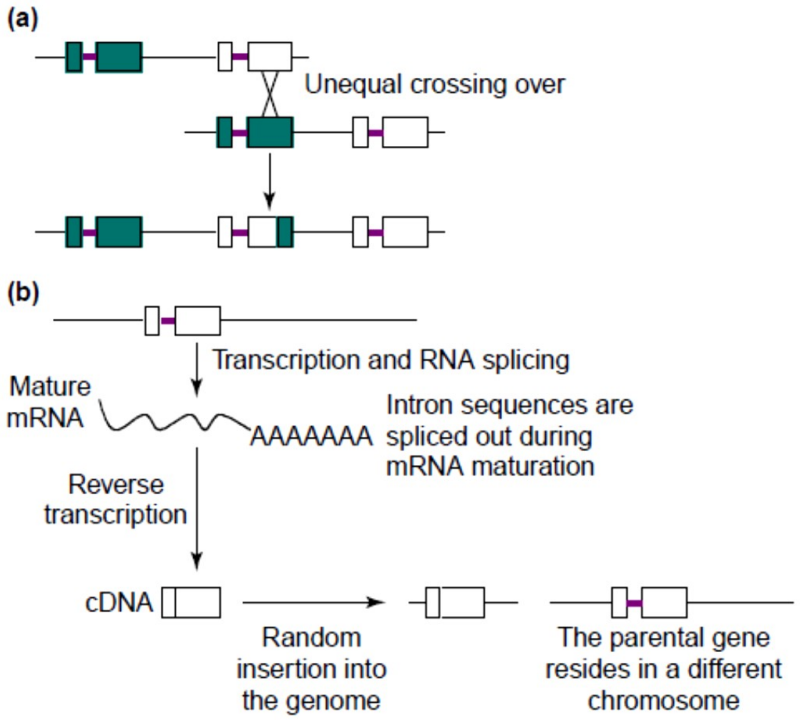


Figure 1: Common mechanisms of gene duplication. a) Unequal crossing over, which results in a recombination event. b) Retrotransposition, which occurs when a mRNA is transcribed into a complementary DNA and inserted into a new genomic position resulting in a gene duplication event. Squares represent exons and bold lines represent introns (Adapted from Zhang, J.Z. [11]).

1.3.2. Models of Gene Duplication

Gene duplication is the principal event that orchestrates evolutionary novelty in eukaryotes, but what are the driving forces for gene duplication, and what are the selection pressures that model the evolving gene pair? Any aspect of genome evolution goes hand to hand with gene duplications among all kinds of species. This is a general way that genes evolve, acquire, develop or share new functions [12]. Understanding these processes is of pivotal interest, since after a gene undergoes duplication, this one can acquire a new truncated function which can segregate in the whole population causing a disease or can confer an adaptive advantage with a new function [13]. In the literature, many aspect and processes of gene duplication and evolution have been well described, but many of these theories have been disproved by experimental approaches [13]. First, we need to address the question of which model can be the most appropriate to understand evolution of gene duplication. Several important models have been proposed. Those models imply the known neo and sub-functionalization, gene conservation and gene loss, which give rise to protein evolution theory.

1.3.2.1. Ohno's Model

Evolution by Gene Duplication, the widely cited theory of Susumu Ohno, was the first to shed light on the evolution of new duplicates, explaining that *"Adaptive mutations accumulate in the new duplicate under no pressure of selection"* [14]. This model is characterized by maintaining a gene duplicate by simply increasing the number of genes coding for a protein (Figure 2) [11]. In an update of Ohno's book "*Evolution by Gene Duplication"*, the authors introduced the notion of "Gene Conservation" where both genes maintain their original functions after duplication. The importance of this mechanism becomes relevant when the same function needs to be duplicated and conserved in order to fulfill certain needs inside the organism. This event is also necessary when there is deficiency in the original or ancestral function and need for higher protein concentrations. The Ohno Model introduced the sub-functionalization theory, where a gene undergoes

duplication accumulating mutations leading one copy to share a burden of the original function with the new copy. Moreover, Ohno's model implies the neo-functionalization after gene duplication and introduces the sub-functionalization theory.

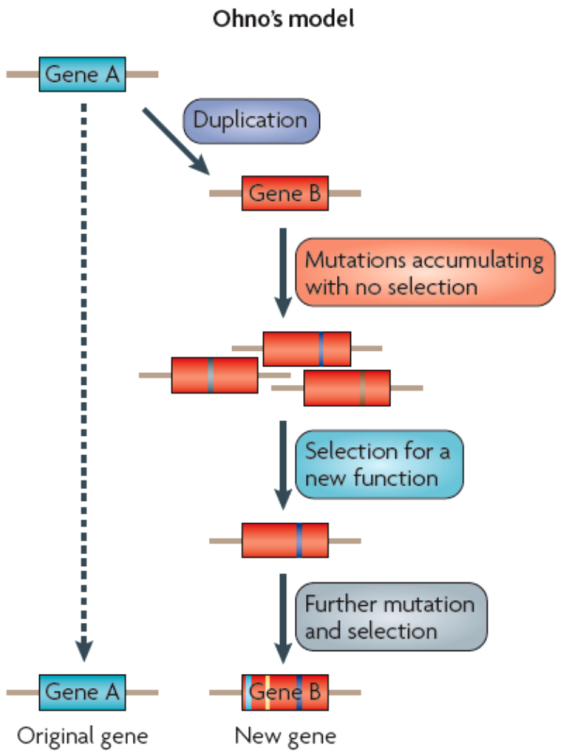


Figure 2: Ohno's model. The new duplicate accumulates mutations under no pressure of selection. This model predicted that gene A undergoes duplication giving birth to gene B. Gene B accumulate mutations under no pressure and with no selection, therefore is selected for a new function by accumulating mutations. While gene A maintains in its original function. (Adapted from Soskine. M and D.F. Tawfik. [5]).

1.3.2.2. Divergence prior to duplication model

This model suggests that while initial levels of new evolving functions are acquired, the original function is maintained. The ancestor gene accumulates mutations under natural selection which leads to be selected for a new function. After duplication, the original gene is maintained (gene A) and new function is acquired by the new duplicate (gene B) (Figure 3). Here, duplication happens after the new duplicated gene is subjected to positive selection [5].

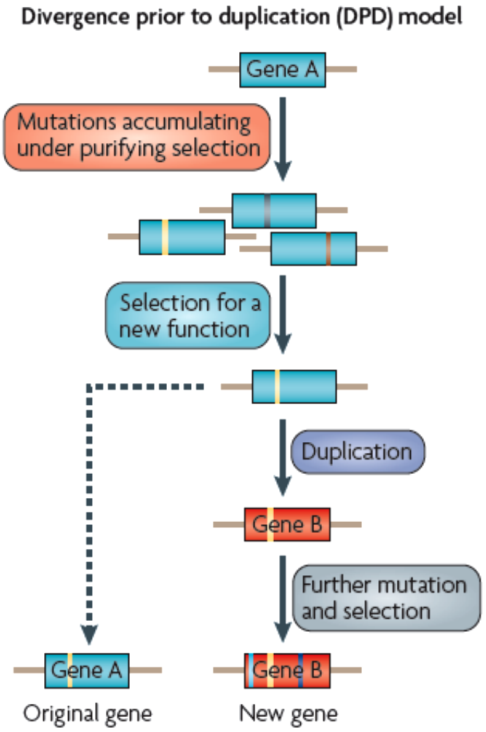


Figure 3: Divergence prior to duplication model (DPD). In the symmetric model the new function appears in the new duplicate and can also appear in the original copy. It proposes that gene A accumulates mutations under pressure of selection which leads to be selected for new function. Gene A undergoes duplication giving birth to gene B, the latter gene accumulates mutations resulting in the development of a new function, while gene A maintains the original function. (Adapted from Soskine. M and D.F. Tawfik. [5]).

1.3.2.3. The Duplication-Degeneration-Complementation Model

The Duplication-Degeneration-Complementation model also known as the sub-functionalization model can be described as a combination of Ohno's and DPD models. This model is often considered as the only model of sub-functionalization in which the maintenance of duplicates does not required adaptive mutations [15]. The two copies may therefore acquire complementary loss-of-function mutations such that both genes are now required to maintain the function of a single ancestral gene (Figure 4).

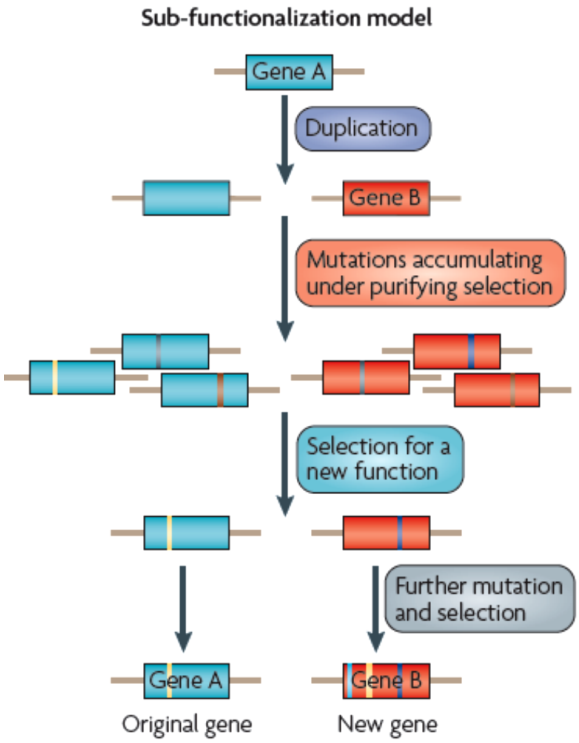


Figure 4: Duplication-Degeneration-Complementation Model (DDC). In this model, one of the two loci results from duplication events, suffer a degenerating mutation and results in loss-of- function. Here gene A undergoes duplication giving birth to gene B. Gene A and B accumulate mutations under pressure of purifying selection, which makes a difference from

the Ohno's model where the duplicates are under no pressure of selection. Gene A and B are selected for new functions where gene A maintains the original function and gene B accumulates mutations and is selected to share a new function (Adapted from Soskine. M and D.F. Tawfik. [5]).

1.3.2.4. Sub/Neo/Non-functionalization

This model explains how genes undergo duplication. Force *et al.* [15] suggested an improved theory of Ohno's model. The authors were mainly interested in the division of expression domains among paralogs, which they claim to be the main genre of sub-functionalization.

Gene's fate after duplication depends on two major phases (Figure 5). Generally, after duplication, genes enter in phase 1 and experience one of the following three alternatives: neo-functionalization, sub-functionalization or non-functionalization (loss-of-function). The latter mechanism differs from the former ones, where the possibility of the loss of genes function was not considered. First, the new duplicate acquire a null mutation in the coding region which drifts to fixation that drive to gene loss (non-functionalization). This event also occurs whether the regulatory region of one duplicate is destroyed. Second, the duplicate may acquire a mutation which confers a totally new function and becomes fixed through Darwinian selection. This selection will lead the duplicate to acquire new mutations on regulatory regions causing a change while conferring a new function (neo-functionalization). Third, each duplicate may experience loss or reduction of expression for different sub-functions cause by degenerative mutations. In this process is known as neutral mutations (sub-functionalization) where no adaptations are formed [14]). In the sub-functionalization process both duplicates can share the burden of a function by fulfilling the requirements of the ancestral function.

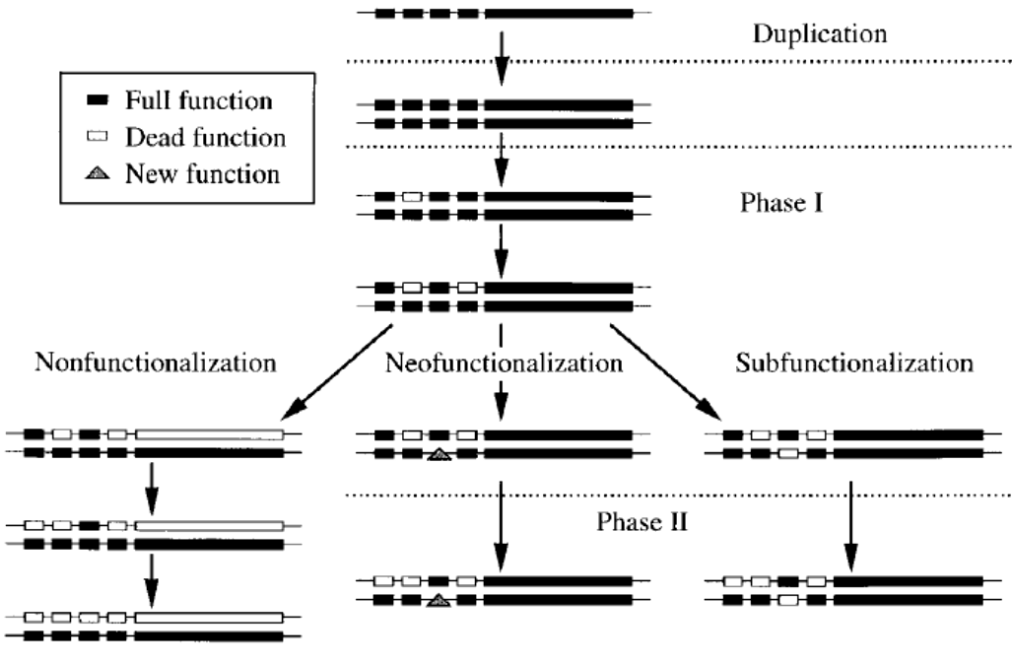


Figure 5: Gene fate after duplication. Notation: Small boxes denote regulatory elements with unique functions. Large boxes denote transcribed regions of a gene. Solid boxes denote intact regions of a gene. Open boxes denote full mutations. Triangles denote the evolution of a new function. (Adapted from Force, A. *et al.* [15]).

Studying these concepts is important to understand the mechanisms underlying the emergence of new phenotypes and how these mechanisms could be regulated in the heart of protein interactions network. Our goal is to show that in a family of proteins, those that were duplicated have evolved their original genetic code and developed new functions and novel structures which can result in phenotypic changes.

1.3.2.5. Understanding how new pathways evolved

The function of a protein is often determined by the identity of the amino acids composing its primary sequence and by various structural constraints. Conserved residues in a dispersed family of proteins are thought to have a direct role in protein structures and functions [4, 16]. In order to understand how proteins achieve specificity, we first need to identify which amino acids are relevant for a given functional class. Second, it's important to determine the number of the proteins residues that need to be mutated in order to exchange functions from one protein into another. Many studies interested in gene duplication and evolution of protein functions were published [15, 17, 18], but none of them have focused on how to identify functional residues. Accordingly, we propose to study the functional differentiation and evolution of the Ras-like small GTPases super family and to examine how novel binding partners evolved and to investigate their influence on the creation of new pathways.

The Ras-like small GTPases protein networks provide an excellent model to demonstrate how new pathways have evolved by making random changes. Therefore, the structures found by evolution depend to some degree, on historical chance and are laden by biochemical details that require special description. The following section describes in details this wide family of proteins in order to understand their specificity and functions.

1.4. Budding yeast as a biological model

The eukaryotic organism *Saccharomyces cerevisiae* represents one of the most studied biological systems and constitutes an ideal model for understanding cellular processes. Used in biochemical and genetics studies, *S. cerevisiae* helped the development of a wide range of biochemical tools which have been designed and optimized to study any specific gene in this organism [19, 20]. Tools such as synthetic genetic array to study the relation between genotype and phenotype, yeast-two hybrid and protein-fragment complementation assay to study protein-protein interaction networks and astrobiology to study survival of different species in the outer space during the *"Living interplanetary flight experiment" (Phobos LIFE. www.interplanetary.org)*. In vivo studies of budding yeast, revealed first evidence of small GTPases by discovering protein sequence similarity of certain proteins with human Ras small GTPases [21].

The approach followed in this research project consists of understanding the development of phenotypes linked to cellular functions involved in this process. This information brings us important clues of which approaches should be taken to start exploring evolution of protein-protein interaction networks using budding yeast as biological model. Furthermore, budding yeast displays a wide range of morphologies such as size and cellular division variants, which helps the development of this research (Figure 6). Moreover, this unicellular organism facilitates the study of cell cycle, cytoskeleton organization, sub cellular organelles, endoplasmic reticulum trafficking, secretion systems, metabolic regulation, signalling processes, chromosome recombination and receptor arrangements.

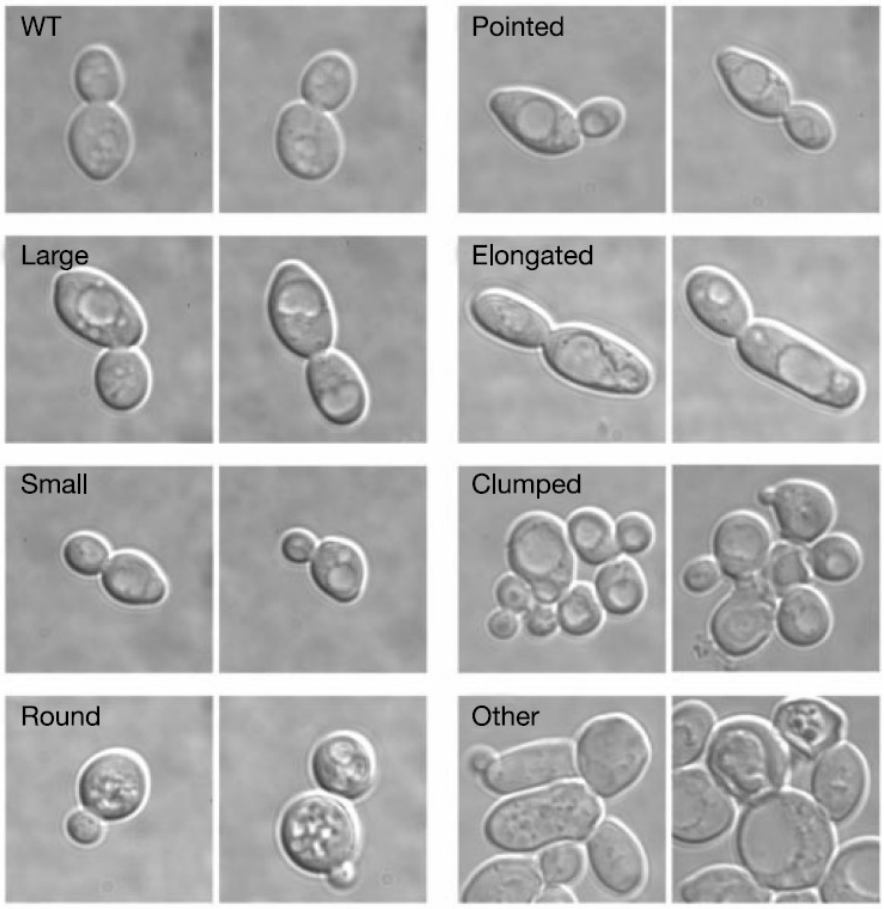


Figure 6 : *Saccharomyces cerevisiae,* a biological model, allowing the visualization of seven distinct phenotypes. These phenotypes serve as a model for classification. (Adapted from Giaever *et al.* [22]).

1.5. Ras-like small GTPases super family

Normal cellular development in multicellular organism is controlled by specific proteins that regulate a variety of complex signalling networks. The Ras-like family of small guanosine triphosphates (small GTPases) are known to be involved in various processes such as cell polarity, polarized growth, membrane trafficking, actin and septin organization and development, cell cycle regulation and cell survival. These proteins are of special interest because they regulate intracellular signal transduction pathways depending on external and internal stimuli. They act as molecular binary switches that are either turned on or off depending on the cell needs. This family of proteins has been well studied in both humans and *Saccharomyces cerevisiae* and showed to have a certain degree of conservation from yeast to humans [23]. Small GTPases serve as an excellent mechanism for the development of similar signalling processes established at levels of functions and structure [24].

1.5.1. How Ras-like small GTPases are regulated?

In *Saccharomyces cerevisiae*, the activity of proteins belonging to the small GTPases family is controlled by the ratio of bound GTP to GDP [25]. They exist in an inactive form (GDP-bound) or an active form (GTP-bound) that activates downstream effectors depending on the signal. They regulate several important biological cascades like the MAPK cascade, cAMP cascade, gene transcription, actin cytoskeleton organization, cell growth, cell cycle progression, bud emergence and pheromone response pathway among others [26]. The switch between active GTP-bound and inactive GDP-bound states is specifically controlled by Guanine nucleotide exchange factors (GEFs). These GEFs stimulate the exchange of GDP into GTP yielding an active form of the protein (Figure 7). This active form is inactivated by the hydrolysis of the GTP, which loses a phosphate group to become inactive GDP state. This inactivation cycle is controlled by the GTPases activating protein (GAPs). It is known that Ras-like small GTPases bind effectors in their

active state in order to activate important signalling pathways. In addition, Rho and Rab small GTPases subfamilies are also regulated by a third class of regulatory protein called Guanosine Nucleotide Dissociation Inhibitors (GDIs) These regulators not only prevent the GDP – GTP exchange cycle, but also maintain the protein in their GDP inactive state and from being localized at the membrane [27] that will be described later on. In addition, these GEFs and GAPs have proved to be important targets for drug design for cancer therapy [28] and rewiring of cellular morphology [29].

1.5.2. The Ras-like small GTPases are grouped in five major branches

Most of the available genomic studies on budding yeast *Saccharomyces cerevisiae* have allowed the classification, distribution and functional diversification of the Ras-like small GTPases. The available genome sequences of many eukaryotes allowed us to analyze this family from an evolutionary perspective. This super family is composed of more than 700 sequences from different species which contributes the understanding of origin, evolution, function and structure [30]. The Ras-like small GTPases is divided into five major branches (Figure 8). These branches are divided into five subfamilies which are Ras, Rho, Rab, Arf and Ran. Previous studies revealed that expansion and evolution of these genes have been related to unique eukaryotic cellular features such as cell division, cell cycle regulation, phagocytosis, apoptosis and signalling regulation. [31]. Each subfamily is involved in several important intracellular tasks. In addition, the comparison of these subfamilies using sequence alignment approaches showed that G domains are conserved among the five subfamilies, an observation that served for creation of constitutively active mutants [32].

1.5.3. Cell morphology: Regulation by Rho and Ras small GTPases

Maintaining the appropriate cell shape and morphology is essential for homeostasis and dynamic processes. *Saccharomyces cerevisiae* undergoes several morphological changes during the cell cycle [27], therefore this family of proteins participates actively in regulating actin cytoskeleton, cell growth, survival and differentiation, which implies regulation of cell cycle. These are fundamental processes that are essential for cellular development [33].

The Rho and Ras small GTPases are the most important subfamilies that particularly regulate cell shape. This regulation happens via the asymmetrical regulation of the actin cytoskeleton which leads to actin localization at sites of growth during budding and bud site selection, ending in normal budding process. Moreover, these subfamilies perform their tasks during the cell cycle through important signalling processes which are mediated by protein-protein interactions [34]. The cytoskeleton plays an important role in determining cell morphology, this process is known to be mainly regulated by Rho and Ras proteins which are known to be mutated in cancerous cells [35]. This chapter describes such processes and the involvement of Rho and Ras small GTPases and how these proteins orchestrate cell morphology behaviour.

1.5.4. Ras-like small GTPases

Ras subfamily was the first discovered proteins of the Ras-like small GTPases superfamily. These proteins are the key regulators of extracellular and cytoplasmic signalling networks that control cell growth, survival and differentiation [28]. An additional information is that in mammalian systems, the Ras subfamily is comprised of 36 proteins that has been studied extensively over the past two decades because of their roles in human cancers [36]. In contrast to mammals, *Saccharomyces cerevisiae* features a Ras subfamily comprised of only 4 members Ras1, Ras2, Rsr1 and Rhb1 which are involved in different

cellular roles. These proteins contain N-terminal portions with a significant homology to mammalian Ras and others subfamilies. As well they feature a C-terminus which includes at the terminal 4 amino acids that constituted the CAAX motif (C is cysteine, A is aliphatic amino acid And X is the C-terminal amino acid). This motif is important for posttranslational modifications that facilitate the association with the membrane. The conserved regions contain boxes known as G1, G2, G3, G4 and G5 which are short sequences of amino acids involved in the recognition of GEFs and GAPs [37].

Briefly, Ras1 is involved in G-protein signalling in the adenylate cycline activating pathway (cAMP) that plays a role in cell proliferation. It is localized at the plasma membrane and is a homolog of mammalian Ras proto-oncogenes [38]. His homolog Ras2 is in charge of the regulating the nitrogen starvation response, sporulation and filamentous growth, farnesyaltion and palmitoylation. Such functions are required for activity and localization to plasma membrane [38]. Following these descriptions, Rsr1 is required for bud site selection, response to mating pheromone and cell fusion. It is localized at the plasma membrane [39]. It has significant similarity to mammalian Rap small GTPases. Finally, Rhb1 which is related to mammalian Rheb and is involved in regulating canavanine resistance and arginine uptake [40]. These are just some functions that Ras proteins are capable to regulate.

1.5.5. GDP-GTP cycle of Ras-like small GTPases

The Ras subfamily of small GTPases feature a GDP-GTP cycle which is similar in other subfamilies such as Ran and Arf with the exception of Rho and Rab which are described in their own sections.

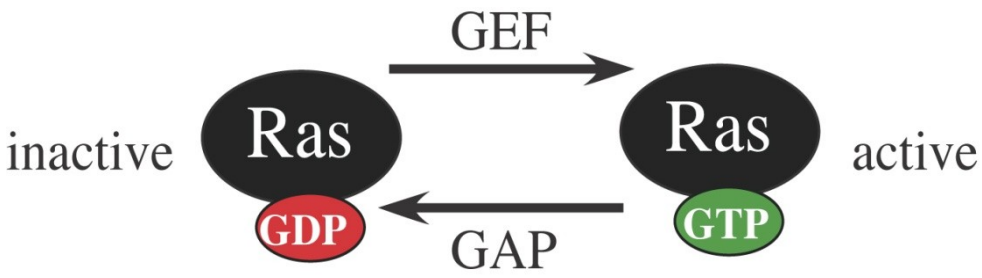


Figure 7: The GDP-GTP cycle of Ras-like small GTPases. Ras proteins are in their active state when bound to a GTP molecule and are inactive when bound to a GDP molecule. Ras activation is controlled by GEFs that stimulates the exchange of GDP into a GTP and Ras inactivation is controlled by GAPs that hydrolyse the GTP into GDP (Adapted from Downward, J. [41]).

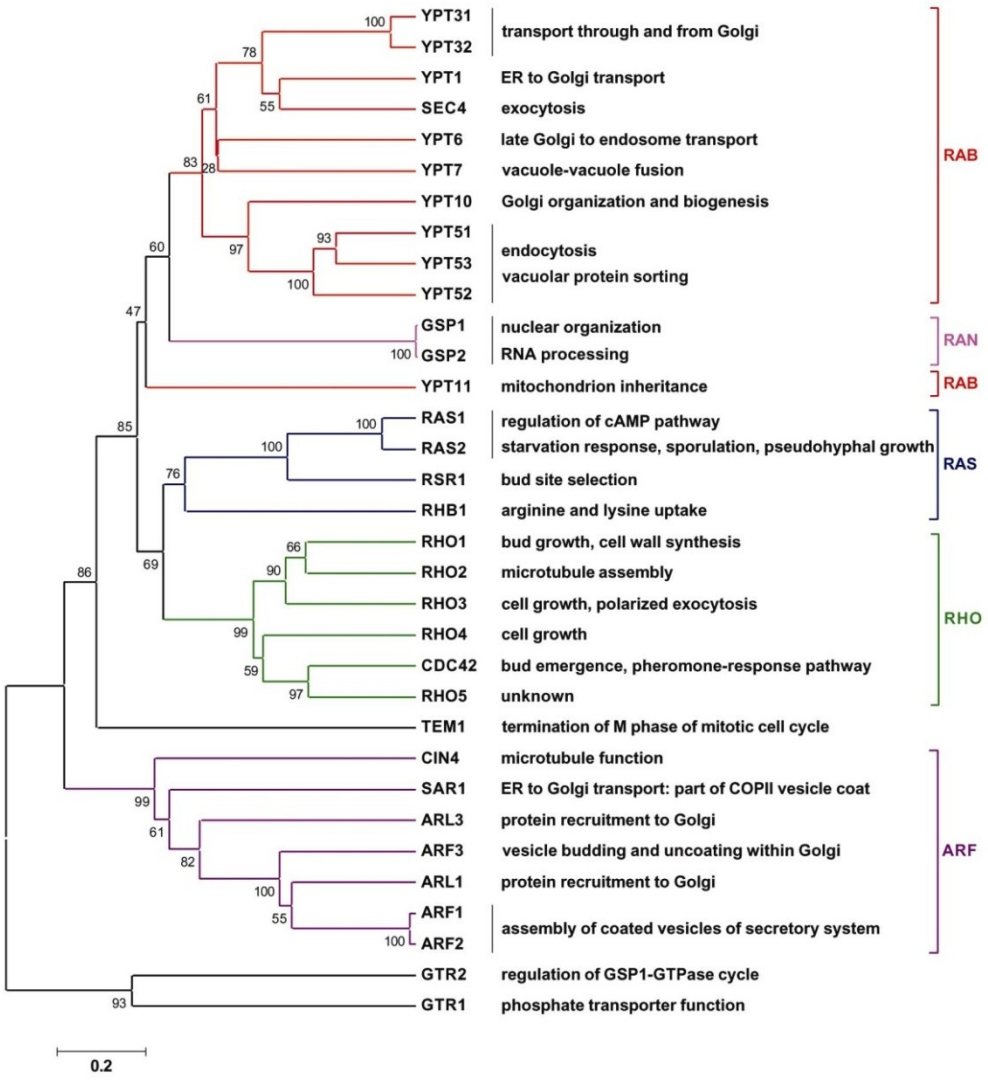


Figure 8: Phylogenetic tree of 34 known Ras-like small GTPases of *Saccharomyces cerevisiae*. Proteins are clustered into subfamilies of Rab, Ran, Ras, Rho and Arf respectively. Clusters appear to be grouped according to cellular functions (Adapted from Garcia-Ranea *et al.* [30]).

1.5.6. Rho-like small GTPases

Through a series of complex biochemical networks, Rho small GTPases controls some of the most important processes of cell morphology. Rho small GTPases are also regulated by GEFs, GAPs and a third class of regulatory protein called Guanosine Nucleotide Dissociation Inhibitors (GDIs). These regulators not only prevent the GDP – GTP exchange cycle, but also maintain the protein in their GDP inactive state and from being localized at the membrane (Figure 9) [27]. This subfamily shares similar roles in signal transduction to Ras small GTPases subfamily and is best characterized for the regulation of actin cytoskeleton organization, morphogenesis and cell shape, movement and polarity and cell cycle progression [42, 43]. Hence, a particular cell shape is adopted and controlled via the asymmetry and organization of signalling pathways which involves the Rho subfamily of small GTPases [9].

The Rho subfamily of small GTPases are present in all eukaryotic systems being conserved from yeast to humans [44]. Such conservation suggests that basic mechanisms involved in cell morphology were conserved during evolution. Such mechanisms have defined the finest jobs in development and maintenance of cell morphology. The dynamics of the actin cytoskeleton are mainly controlled by Rho, which have significant roles during budding process [27, 45]. In addition, Rho subfamily is divided again into two subfamilies which are Rho and Rac and they apparently regulate overlapping pathways [30].

About Rho in mammals, recently Tybulewicz and Henderson reviewed Rho proteins describing their role in the immune response acting as regulators in lymphocytes development, differentiation, activation and migration. Rac1 and Rac2 have demonstrated important roles in B cell development, where loss of *Rac1/Rac2* has for consequence a block of such process. This leads to an arrest of B cell development stage in the spleen, showing that Rac1 and Rac2 are required at the earliest stage of transitional B cells [46, 47]. Moreover, Rac1 is known to be mammalian homolog of budding yeast Rho5p, which

have similar roles in a protein kinase signal transduction pathway (Pkc1p) [48], being one of the reason to work with Rho5 in budding yeast.

A comparison of mammalian and budding yeast Rho small GTPases reveals that in mammals it comprises 23 members which are related in primary structure. Phylogenetic analysis shows in mammals that these proteins cluster into different subfamilies based on sequence similarity [47]. Whereas, in budding yeast is comprised of only 6 proteins; Rho1, Rho2, Rho3, Rho4, Rho5 and Cdc42 (Cell division cycle 42).

A wide range of studies have been made from mammals to yeast regarding Rho behaviour demonstrating important roles in cell integrity. To begin, Rho1 a small GTPase, involved in establishment of cell polarity, regulates protein kinase C (Pkc1p) and cell wall synthesizing 1, 3-beta-glucan synthase (Fks1p and Gscp) for fungal wall biosynthesis, which makes the most important enzyme responsible of the synthesis of cell wall polymer in budding yeast. [49]. Rho1 is essential and required for bud formation and development [50]. It is localized to the area of polarized growth independently of the actin cytoskeleton and is activated by a series of GEFs Rom1, Rom2 and Tus1 [51]. Rho1 GDP/GTP cycle is regulated by GAPs Bem2 and Sac7 which also have a role in negative regulation of the MAPK cell-integrity pathway [52, 53].

Rho1 has 67% identity with human RhoA, and both have similar roles and localized to places of active growth [54]. Recent work by Lee *et al.* showed that Rho1 confers resistance to oxidative responses in budding yeast providing protection. Yeast two hybrid assay showed that Rho1 associates with Ycf1, a vacuolar glutathione S-conjugate transporter, showing biomolecular fluorescent complex on the vacuolar membrane [55]. As well, several works revealed that Rho1 has the most notorious potential implications in control and coordination of three distinct biochemical pathways, which each of them contributes to growth and budding during the cell division cycle.

Rho subfamily also features four non essential small GTPases. Rho2 is involved in the establishment of cell polarity and microtubule assembly. Its disruption does not produce

a detectable phenotype, however, it may regulate aspects of the budding process [49]. Rho3 is involved in the establishment of cell polarity along with Rho2. Its GTPase activity is potentially regulated by the GAP Rgd1p, which is a GAP involved in control of actin cytoskeleton [56]. Rho4 is implicated in the establishment of cell polarity [56].

Moreover, the Rho3 and Rho4 proteins have a slightly difference. They show only 57% similarity in primary structure, but they are functionally related. Evidence about Rho3 and Rho4 tells that they seem to be required after bud formation to maintain cell polarity during the maturation of daughter cells [57]. They are also involved in the regulation of exocytosis and actin polarization [58]. Matsui *et al.* also showed that disruption of Rho3 gene produces viable cells with very poor growth, suggesting to be little essential after all, whereas Rho4 disruption has no significant effect on cell growth, but a double mutant yield cell with growth effects and over-expression of Rho4 is able to rescue Rho3 mutants.

Rho5, one of the main focus for this work, has been originally revealed in silico analysis of the entire yeast genome [30]. It is involved in protein kinase C (Pkc1p)-dependent signal transduction pathway that controls cell integrity. During functional characterization of Rho GAPs Bag7, Lrg1 and Rgd2, Roumanie *et al.*, showed the first evidence of Rho5 in budding yeast using a systematic yeast two hybrid approach. It has been observed that Rho5 showed 46% sequence identity with Cdc42, an essential member of the Rho subfamily in *Schizosaccharomyces pombe,* as well with human homolog Rac1. In addition, Rho5 also shows 45% identity with members of the Rac subfamily, making Rho5 appears as a unique Rho/Rac-like protein in *S. cerevisiae* genome. Another feature that makes Rho5 interesting is that it is the only protein among the six belonging to the Rho subfamily that contains a PEST motif which plays a role in protein stability through ubiquitination process.

Despite that Rho5 was characterized as a non-essential protein in the *S. cerevisiae* genome, several research works have been done demonstrating its importance in development and regulation of cell integrity. In 2002 Schmitz *et al*. showed that Rho5 plays

an important in the protein kinase C (PKC1) dependent signal transduction pathway, where deletion of *rho5* increased its activity. This deletion showed an increased resistance to drugs such as caffeine, calcofluor white and congo red, which indicates activation of the PKC pathway. In contrast, an overlapping activity has risen by over expressing a constitutively active mutant of Rho5 Q91H which renders cells more sensitive to these drugs, suggesting that Rho5 acts as an off-switch for the MAP-kinase cascade, which differentiates between MAPK-dependent and –independent functions of Pkc1 [48]. Other roles that have been assigned for Rho5 are; is necessary for H_2O_2-induced cell death, which goes with reactive oxygen species (ROS) accumulation and DNA fragmentation by over expressing a constitutively active mutant of Rho5p G12V which leads to a ROS accumulation followed by cell death upon H_2O_2 treatment [59]. Also Rho5 binds to Ste50 which is a protein involved in mating response and osmotolerance [60]. The expression of Rho5 Q91H allele in a *ste50* deletion strain is lethal under osmotic stress. These data suggest a role in mediating the osmotic stress response via phosphorylation and ubiquitination [61].

Cell division control 42 or simply Cdc42 is a highly conserved small GTPase essential for establishment and maintenance of cell polarity acting as a molecular switch modulating a wide range of signalling processes. Mutant versions show defects in the organization of actin cytoskeleton and septins which have subtle roles for progression of the cell cycle. Essentially, Cdc42 is required to promote the assembly of bud components at the bud site [62].

This main regulator is known to be involved in processes like actin patch polarization, pheromone response pathway, actin cable nucleation and septin organization which are mostly related with the maintenance of cell morphology [27]. In *S. cerevisiae* Cdc42 is found at the plasma membrane being localized at specific domains coordinating polarized organization of actin cytoskeleton during cell growth [24, 63]. Cdc42 is known to induce specific phenotypes after activation to the GTP state. The Cdc42 Q61L has been reported to induce amorphous/misshapen/elongated phenotype giving his importance in the

regulation and maintenance of cell morphology. In addition, mammalian Cdc42 is known to control the formation of actin bundle containing filopodia at the cellular periphery [64].

These are the major reasons why we choose Rho small GTPases in order to study their phenotypic and protein network evolution. They are very efficient at inducing phenotypes and also inducing malignant cells in humans [26].

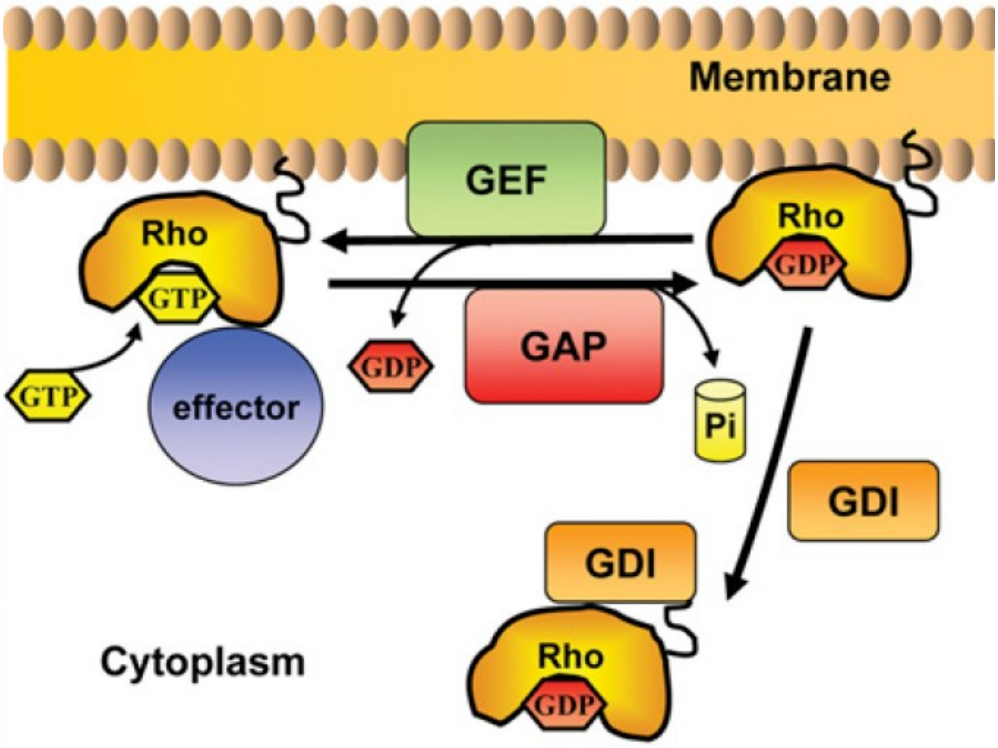


Figure 9: The GDP-GTP cycle of Rho small GTPases. As we can see in the cycle of small GTPases, the footstep from active state to inactive state is managed by GAPs and GEFs. In the case of Rho and Rab subfamilies, these have a third regulator GDIs. These important steps lead these proteins to regulate important intracellular responses like cell polarity, cell shape, cell migration, cell survival, cell proliferation among others which are controlled by different pivotal pathways. (Adapted from Perez *et al.* [27]).

1.5.7. How do vesicles and tubular structures generate cellular transport?

Vesicle trafficking is the movement of molecules inside the cell, from one compartment to another. These processes are known to be involved in novel functions at golgi complex and cilia assembly which implicates cell morphology [65]. Rab and Arf subfamilies of the Ras-like small GTPases are known to be membrane associated proteins. They are molecular switches modulated by GEFs and GAPs effectors. In their active state they promote the traffic and movement of vesicles and organelles in vesicle budding to specific cell compartments [66].

1.5.8. Rab-like small GTPases

This subfamily of Ras GTPases regulates membrane trafficking and intracellular transport. Rab small GTPases constitute the largest family of the known membrane trafficking processes [66]. It is composed of eleven proteins involved in several different functions. These small GTPases have the same activation mechanism but with different roles in their active state. They are molecular switches which in the active GTP bound state promotes trafficking of vesicle budding, vesicle motility, vesicle docking to specific membranes and vesicle fusion across the cell [66].

Rab small GTPases homologs, Ypt31 and Ypt32, are involved in the exocytic pathway. This protein mediates intra-golgi traffic and budding of the post-golgi from the trans-golgi [67]. The Ypt1 protein was one of the first discovered from Rab small GTPases subfamily, was found lying on actin and β-tubulin in the yeast genome [68]. Ypt1 is required for vesicle docking, and in the ER to golgi step of the secretory pathway [69]. The essential Sec4 small GTPase is required during vesicular transport and fusion from golgi to the plasma membrane and at the last steps of exocytic secretion. Sec4 also is located at bud tips and exocytic sites of plasma membrane and post golgi secretory vesicles. Its involvement in cell morphology can be observed at the bud tip in which is capable of

accumulating vesicles at bud tip. Its required for polarized transport of secretory vesicles thereby helping to regulate cell shape [70].

Ypt6p is required for ER to golgi and in cis- to medial-golgi transport. It has been proved to be involved in secretory pathway at elevated temperatures which makes it essential in such processes only at elevated temperatures in budding yeast [71]. Ypt7 is involved in endocytosis. It is known to be localized at the vacuole and is required during vacuole fusion reaction [72]. Protein Ypt10 is the only one of this subfamily that contains a PEST motif which is specific for control of proteolytic enzymes for degradation. Previous studies regarding over-expression of this protein showed an accumulation of golgi like cisternae with budding vesicles proving its involvement in cell shape control [73].

The homologues Ypt51, Ypt52 and Ypt53 small GTPases are required for endocytic transport, for sorting hydrolases and vacuolar proteins. They are detected at the mitochondria level and are known to be mammalian Rab5 homologs [74]. Last, Ypt11 mediates distribution of mitochondria and endoplasmic reticuli to daughter cells at the moment of budding [75]. These are known roles regarding this subfamily, showing its importance in vesicle trafficking and cellular morphology behaviour.

1.5.9. Arf-like small GTPases

ADP-ribosylation factor (ARF) is a subfamily best characterized to be regulators of membrane trafficking, organelle structure and intracellular transport. Functions of these proteins are GDP-GTP cycle dependant. An interesting observation regarding Arf's GEFs and GAPs is that these GEFs contain a conserved Sec7 domain that catalyses hydrolysis of the GTP-bound state having intrinsic GTP hydrolysis activity, while the GAPs contain a conserved zinc-finger GAP catalytic domain [65]. These features facilitated the identification of Arf regulators from yeast to human.

Chromosome instability is mainly controlled by the Arf-like small GTPase Cin4, where its major function is observed at the folding of beta-tubulin [76]. Sar1 component of COPII coat of vesicles is required for transport vesicle formation between ER to golgi. The main function of Sar1 is to recruit COPII to the membrane mediated by its GEF Sec12 [77]. Related to previous roles, Arl3 is required for recruitment of Arl1 to the golgi [78]. Arl1 soluble small GTPase involved in regulation of membrane traffic and intracellular control of potassium influx. This role has been shown by creating a null Arl1 strain [79].

Homologs Arf1 and Arf2 are involved in regulation of coated vesicles formation in intracellular trafficking inside the golgi [80]. While, Arf2 gene encodes one of the three ADP ribosylation factors in the *S. cerevisiae.* These two proteins share a 96% identity in sequences. Mutant versions of Arf2 show no morphological changes whereas deletion of both Arf1 and Arf2 show no viability for the cell [80]. Not being a homolog but with similar name, Arf3 is involved in development of cell polarity, making it important for morphogenesis processes [81].

1.5.10. Ran-like small GTPases

This Ras-related nuclear small GTPases regulates nucleocytoplasmic transport of macromolecules, RNA, proteins and the organization of the spindle apparatus during mitosis [82]. Comprise only two proteins Gsp1 and Gsp2. It has also been shown that Ran small GTPases GDP/GTP cycling controls DNA replication and nuclear envelope gathering being important during cell cycle development [83]. In budding yeast have been characterized two major small GTPases which are homolog in sequences and involved in similar cellular tasks. Gsp1 represents an essential protein involved in the maintenance of nuclear organization and RNA transport and processing [84]. His homolog Gsp2 is not required for viability and is also involved in the maintenance of nuclear organization and RNA processing [84].

1.5.11. Ras-like small GTPases and their influence in cancer

Among the first discovered proteins involved in cancer development, Ras proteins were the ones indentified to be involved in regulating cell growth and differentiation. These are the most important steps for cellular development in order to achieve a well structural organization [41]. About 20 % of human tumours are known to express mutant versions of Ras-like small GTPases. These point mutations turn a normal protein into a constitutively active form contributing enormously to aberrant phenotypes. These aberrations imply deregulation of the cell cycle, apoptosis and growth control [41, 85]. As mentioned before, point mutations tend to compromise GTPase activity preventing GAPs from promoting hydrolysis of active GTP and therefore leading to accumulation of active Ras. Hence, this accumulation leads to cancerous cells.

Aberrant Ras signalling in cancer are actually caused by: (i) Mutations, where the majority of Ras aberrant activation in tumour cells are accounted by mutations in codons 12, 13 and 61 which represent the G1 and G3 domains respectively [86]. These mutations yield a marked and disrupted signalling through Ras pathways. (ii) GAP deletion, where tumour cells can be activated by the deletion of GAPs, meaning that Ras GTPases will remain active. These GAPs proteins are pivotal for the regulation of the GDP–GTP cycle and the lack of them ends in a not regulated protein. (iii) Growth-factor-receptor activation, where the over-expression of growth factor receptor tyrosine kinases, are common in tumour cells, where Ras-mediated growth phenotype signalling pathway are turned on. Also (iv) Mutation or amplification of Ras effectors, where serine/threonine kinase BRAF is frequently activated by mutations in human tumours. This occurs in a very limited number of residues in the kinase domains which results in cascade-like kinase activation [41, 87]. These four deregulation in signalling processes found in cancer cells help us understand the rational of designing new cancer therapies in order to target Ras pathways.

1.6. Divergent amino acid positions involved in cell morphology

A previous study has focused on the conservation of these Ras small GTPases, identifying potential residues associated with specific function [2]. Heo *et al.* showed that over-expression of constitutively active mutants of mammalian small GTPases induced different cell morphologies that fell into nine distinct classes. They were linked with functions taking as a constraint morphology classification. They generated constitutively active mutants based on the k-Ras derived oncogenic glutamine (G) to leucine (L) mutation at position 61 found in the third GTP binding region [88]. In order to predict relevant amino acids positions, they developed an algorithm that predicted residues positions which can be exchanged to create switched-of-function mutants and the number of them necessary to achieve it. Moreover, Frankel *et al.* was showing that transcriptional enhancers have caused morphological evolution by single nucleotides exchanges in *Drosophila melanogaster*. They showed that single nucleotide substitution has very little effect on morphology but the combination of them has large impact causing significant differences in morphology [89].

We will implement a similar approach exploring the relation of protein networks; whether they evolved by influence of single switched point-mutations or the combination of them. With these observations we want to focus on the evolution of protein-protein interaction networks in budding yeast.

1.7. Shape Factors and morphologic properties

It is known that in budding yeast, cell cycle is a series of processes that orchestrate cell development and division, which is linked to cell morphology regulation [90]. Cells have evolved specializing in different shapes which help them to accomplish tasks more efficiently and make them differentiate from normal to abnormal cells [33]. The major challenge of this project is to quantify and classify the effect that constitutively active mutants might have on cellular shape. We want to know at which level these point

mutations are able to affect morphological characterization of budding yeast cells? We tried to study these questions in cells over-expressing such mutants and by using cellular size and shape as major features in order to describe cells morphology after division.

Why cell shapes are important? Shapes provide important information that characterize cellular morphology and classify cells accordingly. However, there are many ways to measure the cell size and identify its shape. For our study, we chose shape factors which are mainly used in cement and concrete research [3]. Here, shape factors were developed and adapted to cellular phenotypes which represents a number of values describing cellular size and shape. It is calculated by a simple formula which takes into account two parameters per formula: circularity shape factor (area and perimeter), elliptical shape factor (length and breadth), aspect ratio shape factor (inner radius and outer radius) and elongation shape factor (centroid coordinates in a two dimensional space (X and Y)). Shape factors are often affected by the shape of the object being independent of its dimensions [3, 91].

Previous works were investigating the possibility of quantifying differential interference contrast (hereafter DIC) images without using fluorophores [92]. In other words, previous studies tried to automate DIC quantification [93]. So we addressed this issue by using shape factor formulas using DIC. DIC technique, in addition of being a straightforward approach, also extracts the morphology of the whole cell as it is. In general, DIC microscopy should be considered as a complementary method, but in our study, we are taking DIC as main method to fully investigate budding yeast morphology.

1.8. Cluster analysis algorithms

Cluster analysis consists of grouping a collection of objects that share identity or similarity in one or a combination of several parameters. These parameters could refer to any measurement that describes the objects of interest (i.e. a set of colors, object type or a

set of measures). And therefore the major goal of a cluster analysis is to arrange these objects according to a given parameter. Elements of the same cluster share more similarity than those in a different cluster. There are different clustering algorithm methods like Hierarchical, K-Means and Fuzzy *C*-Means.

Since cluster analysis takes into account a set of measures as parameter, we thought of applying this method on the shape factors (Section 1.7.). Based on this information cluster analysis will help us to group cells according to their morphology profile into different cluster-phenotypes. This will allow us to investigate whether the population of wild type and mutants could be distinguished using statistical measurements.

As well, looking at budding yeast phenotypes, we know that we can find cells expressing several cell shapes simultaneously (Figure 6). Our goal is to analyse and discriminate these morphologies. To that end, we need at least three methods that can be applied on the same data and analyse them according to different stringency criteria. The three chosen methods are described below.

1.8.1. Hierarchical clustering algorithm

Introduced by Johnson (1967) [94], Hierarchical clustering creates a set of nested clusters that can be represented as a hierarchical tree (Figure 10). This method is known to create a hard partition, meaning that each object, the cell phenotypes in our case, could belong only to one cluster. This method is based on two factors: the distance and the grouping criteria, where similarity is defined on the basis of the distance between two samples in m-dimensional space. The default is Euclidean distance, which is the length of a straight line between two points in a Euclidean space, and Ward's method as a grouping criteria, which assumes that each cluster is represented by its centroid and aims to minimize the sum of square errors (SSE) between the elements of each cluster and its centroid [95]. Often this method is represented by the hierarchical tree as an output graphic (Figure 10).

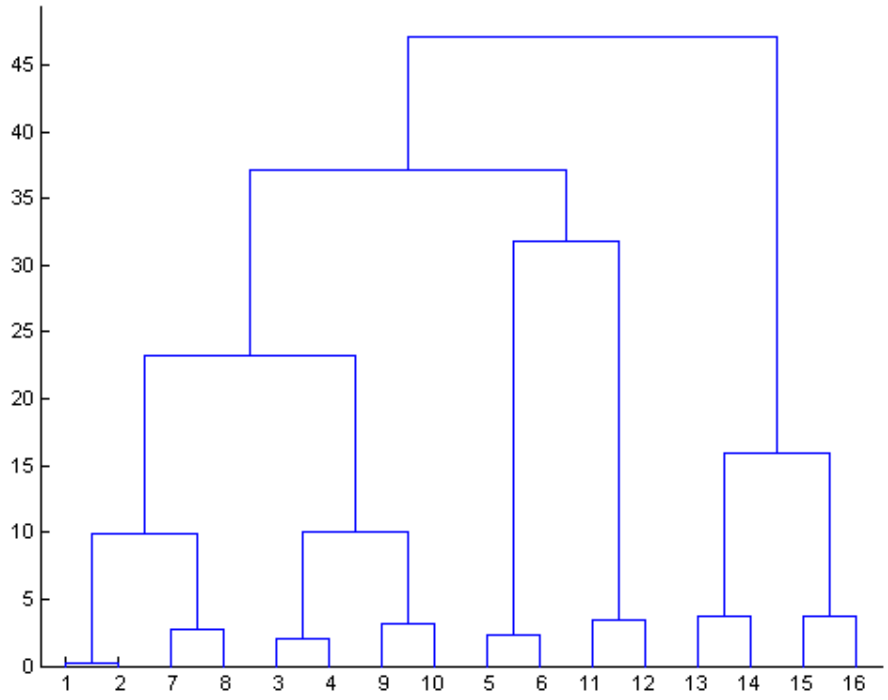


Figure 10: Hierarchical binary tree structure or dendrogram illustrates the result of a hierarchical clustering analysis. The height of the bars indicates how similar the clustered objects are.

1.8.2. *K*-Means clustering algorithm

A cluster analysis method which aims to partition *N* observations into *K* clusters in which each object belongs to the same cluster with the closest mean, was first introduced by MacQueen (1967) [96]. This is a simple algorithm, since it repeatedly calculates the similarity of each point to each centroid grouping the similar cell phenotypes together (Figure 11). This algorithm is also known as hard partition clustering where each cell belongs only to one cluster [95]. Moreover, in the clustering field there is no agreement for choosing the number of clusters (*K*) and most scientists tend to choose it arbitrarily, and there is no general solution in order to define the most optimal number of cluster when compute *K*-Means algorithm. The best approach to take is by trying just different *K* values according to a given criteria. This method has criteria and the advantage that it can be used to predict the number of clusters by calculating the within cluster sum of squares (variance) and the between cluster sum of squares (covariance), which gives an approximate *K* value to compute the algorithm [97].

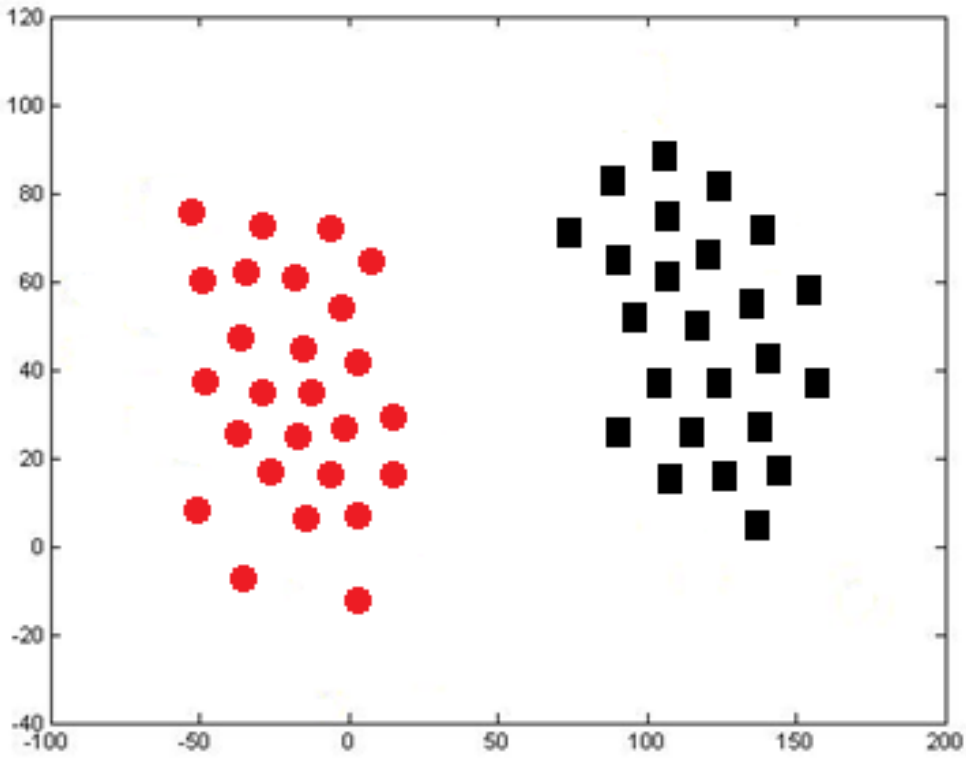


Figure 11: *K*-Means dotplot representation. Here *K*= 2, meaning that two defined clusters exist. The data separation in *K*-Means depends on the clusters with the nearest mean. *K*-Means clustering algorithm shows a clear-cut distribution of objects.

1.8.3. Fuzzy *C*-Means clustering algorithm

The fuzzy logic or probabilistic logic is a reasoning that deals with the approximate rather than with the exact. To a reasonable extend, computer scientists have merged fuzzy logic with data mining techniques in order to concept the soft and permissive computing of data, where the data have a certain degree of freedom to belong to different groups. Fuzzy *C*-Means is a clustering algorithm which combines these needs, the *K*-Means algorithm with the fuzzy logic. Its a method of cluster analysis which performs a soft partition clustering which was first developed by Dunn (1973) [98] and improved by Bezdek (1981) [99]. In other words, like in fuzzy logic objects can belong to more than one cluster (Figure 12). Here, every object is assigned a membership weight that varies between 0 and 1. In Fuzzy *C*-Means, the centroid of a cluster is calculated as being the mean of all measured points, weighted by their degree of membership to one cluster or another different cluster. The degree of membership in a certain cluster is related to the inverse of the distance to the cluster. This is the soft version of *K*-Means clustering algorithm.

Fuzzy *C*-Means is known to be a remarkable clustering algorithm to compute biological data. Fuzzy *C*-Means clustering algorithm is use in gene expression time-course data in order to monitor time points of the budding yeast cell cycle [100]. Fuzzy *C*-Means is relevant assigning genes to clusters in gene expression data in gradual manner. These features increase his robustness to noise and excellent performance in such type of dataset. Fuzzy *C*-Means is also applied for analysis of medical diagnostic systems and in the studies of geographical field.

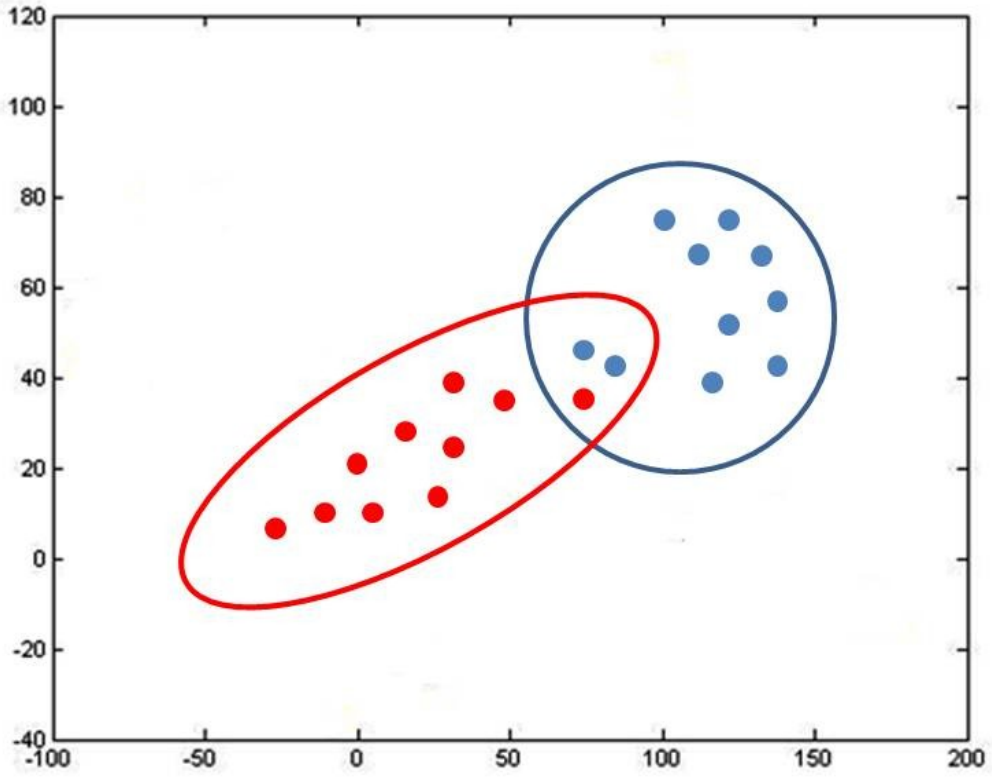


Figure 12: Fuzzy *C*-Means dotplot. In this method each object (red and blue dots) has the degree to belong to both clusters. This criterion is base on the fuzzy logic of data mining.

1.8.4. Clustering Quality Measure

To highlight the significance of clustering algorithm methods and significance of each single cluster, we propose to use the Clustering Quality Measure (CQ_m). The CQ_m quantifies the quality of a clustering algorithm by quantifying the percentage of correctly clustered cells based on their classification (wild type or mutant in our case) [101]. This formula will allow us to judge correctly the significance of each clustering method and choose the best among them.

1.9. Hypothesis

We are interested to investigate how point mutations in small GTPases can affect the cell morphology and their level of impact on asynchronous population. We want to establish a method to determine and quantify mutant and wild type-like phenotypes on these populations using DIC images only. We hypothesize that clustering algorithms can partition mutant cells from wild type cells based on cell shape factor measurements. To prove this hypothesis, we proposed to implement different clustering algorithms to analyze datasets which combines measurements from wild type and respective mutant populations Cdc42 Q61L, Rho5 Q91H, Ras1 Q65L and Rsr1 G12V.

1.10. Specific aims

1. Development of a method that can statistically determine the quantification of each phenotype, establish the effects of constitutively active mutations in shape and size of cells population and demonstrate the utility of this method.

2. Engineer interaction networks and function to comprehend the mechanisms of neo/sub-functionalization of the Ras-like small GTPases family by taking as a constraint phenotype classification.

3. Understanding the mechanisms that improve Ras-like small GTPases specificity in yeast requires an interaction (*in vivo*) map and a protein interactions database to visualize physical interactions for these proteins. In order to test whether the switch of function leads to switch of binding partners, we will first create mutations to identify specific residues that are involved in specific functions linked with phenotypes. Furthermore, we will perform protein-protein interactions experiments and determine whether these switch of function positions lead to neo/sub-functionalization of these proteins and how new pathways evolved.

This project is divided in three specific goals (present and future):

(i) Develop a method for cell quantification, benchmark phenotypes linked to constitutively active mutants Ras-like small GTPases in budding yeast,

(ii) Identify specific residues that are involved in defining specific phenotypes and interaction partners,

(iii) Study interaction network and function to understand mechanisms of evolution of the Ras-like small GTPases.

2. Materials and methods

The current work involved simple steps from growing budding yeast samples in basic media, introduction of point mutations using basic mutagenesis protocols and analysing the changes in cell population. These changes were analysed using bioinformatics tools which enable us to visualize differences in cellular populations.

2.1. Budding yeast media

Rich yeast media (YPD) containing 1% Bacto yeast extract, 2% Peptone and 2% glucose (Bioshop. Burlington, Ontario, Canada) was used for normal growing conditions. Yeast synthetic complete drop-out media (SC) containing 0.67% Bacto yeast nitrogen base without amino acids and 2% glucose (Bioshop. Burlington, Ontario, Canada) was used for overexpression of mutant forms of Ras-like small GTPases. For YPD and SC media agar plates, was added 2% Bacto agar (Bioshop. Burlington, On, Canada). SC-ura media was used for growth of Yeast Open Reading Frame (ORF) expression vectors (Gateway ORF Collection, Open Biosystems).

2.2. *E. coli* media

Escherichia coli DH5α strain was used for transformation of recombinant DNA (PCR products) For growth of DH5α strain plates with LB (Luria-Bertani media) 1% tryptone, 0.5% yeast extract, 1% NaCl, pH 7.0 and ampicillin (1 mg 1000X amp) media (Bioshop. Burlington, On, Canada) were used.

2.3. Strain and plasmids

E. coli ampicillin resistant selection marker DH5α (80*lac*ZM15 *lac*ZYA-argF U169 *deo*R *rec*A1 *end*A1 *hsd*R17 (rk-, mk+) *pho*A *sup*E44 □ –*thi*-1 *gyr*A96 *rel*A1) strain was used for site-directed mutagenesis. For isolation of genomic DNA, *S. cereviciae* yeast strain of choice for this study was the *BY4741* MAT A (his3Δ1 leu2Δ0 met15Δ0 ura3Δ0) (see

transformation procedures section). *BG1805* URA3 (multicopy 2 micron) destination vector (Thermo Scientific Open Biosystems, Invitrogen) which Yeast ORFs are controlled under *GAL1* promoter by induction with galactose (Figure 10) was used. A special feature of this vector is that it contains a His6 tag which is easily replaced by PCA (Protein fragment complementation assay) for later screening for protein-protein interaction studies [102].

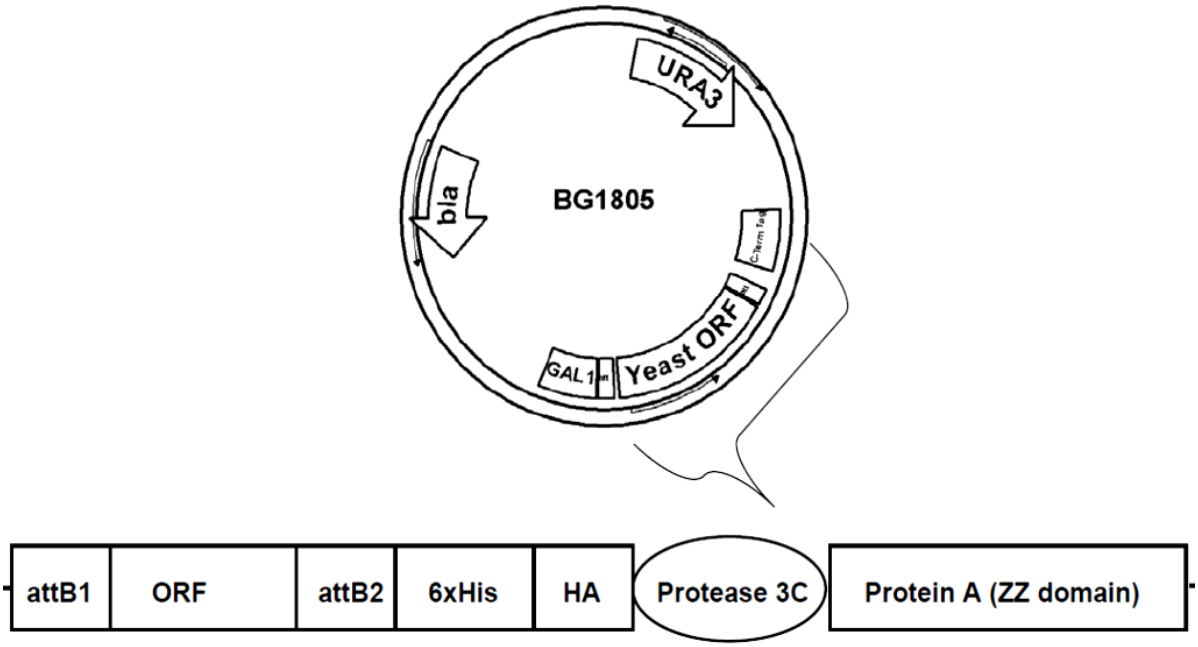


Figure 13: *BG1805* vector construction scheme lacking and ORF. Each ORF replaces the Gateway vector and vector's size can vary depending of the length of the gene's ORF. *BG1805* carries the GAL1 promoter which is inducible by galactose. This was used for expression of the whole Ras-like small GTPases super family.

2.4. Transformation procedures

Budding yeast BY4741 competent cells were prepared according to Knop *et al*. 1999 protocol [103]. Yeast cells were inoculated from previous fresh culture and grown overnight until OD_{600} 0.5-0.7 in YPD 100 ml culture. The next day, cells were split in 50 ml falcon tubes and harvested by centrifugation at 1500 rpm for 5 min at room temperature. Pellets were washed with 5 ml sterile water, centrifuge at 1500 rpm for 5 min, washed with 5 ml of sterile sorbitol buffer (1 M sorbitol, 1mM EDTA, 10 mM Tris pH 8.0, 100 mM LiOAc) and centrifuged again at 1500 rpm for 5 min. SORB buffer was removed by aspiration and resuspended in 360 µl SORB and 40 µl of salmon sperm DNA (10 mg/ml) (Sigma, Oakville, Ontario, Canada) and aliquot in 1.5 ml microtubes and stored at -80 °C.

BY4741 yeast cells transformation was performed by mixing 10 µl of frozen competent cells with 1 µl of plasmid DNA (100 ng) in a 1.5 ml microtube for each protein of the Super family of Ras-like small, GTPases (Gateway ORF Collection, Thermo Scientific Open Biosystems, Invitrogen). 100 µl of Plate solution (40 % PEG 4000, 100 mM Lithium acetate, 10 mM Tris pH 7.5, and 0.4 mM EDTA pH 8.0) was added to each transformation. 10 µl of Dimethylsulfoxide (DMSO) (Fisher Scientific, Fairlawn, NJ, U.S.) was added and mixed by vortexing during 5 seconds followed by incubation at 42 °C for a 15 min heat shock. Cells were pelleted by centrifugation at 2000 rpm for 3 min, supernatant was discarded and the pellet was resuspended in 100 µl of sterile water and plated on 10 cm Petri-dishes. Cells were incubated at 30 °C for 3 days.

E. coli (DH5α) $CaCl_2$ competent cell stock were prepared by streaking out -80 °C frozen stock of DH5α and incubated on LB-only plates at 37 °C overnight. A single colony was picked and inoculated in 5 ml LB media incubating overnight at 300 rpm. Next day cells were transferred into 200 ml LB and grew at 37 °C at 300 rpm until absorbance of OD_{600} 0.6 was achieved; cells were put on ice and chilled at 0 °C for 15 min. Cells were split in 50 ml sterile centrifuge tubes and spun at 4000 rpm for 10 min at 4 °C. Supernatant were discarded and cells were resuspended with 15 ml sterile ice-cold 0.1 M $CaCl_2$ by gentle pipetting and leaving on ice for 15 min and spun again for 10 min at 4000 rpm at 4

°C. After washing cells were resuspended in 4 ml sterile ice-cold 0.1 M $CaCl_2$/15 % glycerol, 100 µl of competent cells were aliquoted in 1.5 ml sterile and labelled centrifuge microtubes for storage at -80 °C.

Recombinant DNA (PCR products) were heat shock transformed into DH5α competent cells by transferring 10 µl of DNA into 100 µl of competent cells, mixed and incubated on ice for 30 min. Followed by heat shock in 42 °C water bath for 45 sec, cells were placed on ice for 1 min and added 1 ml of sterile LB medium and incubated for 1 hr at 37 °C. Cells were spun at 4000 rpm for 2 min and supernatant were removed by aspiration. Bacteria were resuspended in the remaining LB medium by gently pipetting up and down. Cells were transferred and plated into LB + ampicillin plate and incubated overnight at 37 °C. A single colony was picked up next day for diagnostic PCR and later DNA extraction (Bio Basic Inc., Markham, Ontario, Canada).

2.5. Over expression of wild type Ras-like small GTPases

In order to benchmark phenotypes, overexpression of wild type and constitutively active mutants of small GTPases were done. *BG1805* expression vector (Thermo Scientific Open Biosystems, Invitrogen) which was transformed into *BY4741 S. cerevisiae* strain as previously described. Overexpression of small GTPases was performed by incubating at 30 °C 1 ml of overnight growth culture in growth media SD-Ura, 2% raffinose. Overnight culture was diluted in 20 ml SD-Ura, 2% raffinose, incubated starting at OD_{600} 0.3 at 30 °C until reached OD_{600} 0.5 – 0.7. Induction of cell morphology were performed adding galactose (Bioshop. Burlington, On, Canada) until final concentration of 2%. Cells were harvested for 6 hours. After, cells were placed in 96-well optical quality with clear bottom plates. Microscope procedures are described below.

2.6. Sequence alignment to create constitutively active mutants

Protein sequence alignment is the process of finding the best matching between the sequences by inserting "gaps" in the appropriate positions in each sequence, so that the positions where the sequences have identical or similar residues are aligned. The alignment aims to identify regions of similarity that might reveal significant patterns of functional, structural, or evolutionary significance in a given set of protein sequences. Previous studies in mammalian systems have been shown that small GTPases contain guanine nucleotide binding sequences which are involved in generating constitutive activity. This activity have been demonstrated in the G3 domain which derived from oncogenic K-Ras Glutamine to Leucine mutation at position 61 [88]. We took mammalian small GTPases which contain a well conserved G1 and G3 domain and we aligned them against *S. cerevisiae* and other species in order to visualize if the G3 domain was conserved.

2.7. Jalview 2: Sequence alignment software

Ras-like small GTPases sequence alignment was performed using Jalview 2 [104] which is a multiple sequence alignment editor that allows to visualize conserved and divergent amino acids positions that might be involved in a functional class and be significant for functional specificity. The Small GTPases multiple sequence alignment allows to identify the G1 and G3 domains involved in nucleotide binding conferring a constitutive activity to these proteins.

2.8. Perform site-directed mutagenesis to induce constitutively active mutants

Conserved residues of Ras small GTPases between yeast and human were identify and performed site-directed mutagenesis to induce constitutively active mutants of these proteins. Gene-specific primers for site-directed mutagenesis and prediction of functional

sites were designed using A Plasmid Editor Software (ApE, M. Wayne Davis, http://biologylabs.utah.edu/jorgensen/wayned/ape/). The PCR reactions were performed using the Quickchange protocol [105] and standard condition for the Arf, Rab, Ran, Ras and Rho subfamilies. *Pfu ultra* polymerase (Invitrogene) and Bio-Basic site-directed mutagenesis kit was used. Nucleotide sequences were verified by sequencing. In addition, special conditions were considered for the creation of some constitutively active mutants. For all the Ras small GTPases were the G3 motif was conserved, were aligned against human and other species [88] of small GTPases and considered for creation of CAMs. For Sar1 D32V and Rsr1 G12V were the G3 motif it is not conserved among species, the G1 motif was considered [88, 106] (Figure 17).

Table 1: Primers used to create constitutively active mutants. Point mutations are enlightened in red.

Proteins	Sense Primers	Anti-Sense Primers
Ypt32 Q72L	TTGGGACACGGCAGGTCTAGAACGTTACAGGGCCATCACG	CGTGATGGCCCTGTAACGTTCTAGACCTGCCGTGTCCCAA
Rho1 Q68L	ATGGGATACCGCTGGTCTAGAAGATTATGATAGACTAAGA	TCTTAGTCTATCATAATCTTCTAGACCAGCGGTATCCCAT
Rho2 Q65L	GCTCTGGGATACAGCGGGACTAGAGGAATATGAACGTTTA	TAAACGTTCATATTCCTCTAGTCCCGCTGTATCCCAGAGC
Cin4 Q73L	CTATGGGACATTGGGGGGCTACGCACATTAAGGCCATTTT	AAAATGGCCTTAATGTGCGTAGCCCCCCAATGTCCCATAG
Sar1 D32V	ACTACTTTTCTTGGGTTTGGTTAATGCCGGTAAGACCACA	TGTGGTCTTACCGGCATTAACCAAACCCAAGAAAAGTAGT
Cdc42 Q61L	GGCCGGTCTAGAAGATTACGATCGATTGAGACCCTTGTCA	CAATCGATCGTAATCTTCTAGACCGGCCGTATCAAACAAACCTAAC
Ras2 Q68L	GGATACTGCAGGGCTGGAAGAATACTCTGCTATGAGGGAAC	GTTCCCTCATAGCAGAGTATTCTTCCAGCCCTGCAGTATCC
Rhb1 Q74L	CTAGATACTGCAGGCCTAGATGAAGTTTCTCTATTAAACATTAAATCGT	ACGATTTAATGTTTAATAGAGAAACTTCATCTAGGCCTGCAGTATCTAG
Rsr1 G12V	GTAGTATTGGGTGCTGTTGGTGTCGGTAAATCCTGCTTAACCG	CGGTTAAGCAGGATTTACCGACACCAACAGCACCCAATACTAC
Rho3 Q74L	GTGGGATACTGCGGGCCTAGAGGAATTTGACAGGTTACGATCCTTG	CAAGGATCGTAACCTGTCAAATTCCTCTAGGCCCGCAGTATCCCAC
Rho4 Q131L	CATTATGGGACACTGCCGGCCTAGAAGAGTATAGTAGACTTAGACCGCTTTC	GAAAGCGGTCTAAGTCTACTATACTCTTCTAGGCCGGCAGTGTCCCATAATG
Ypt1 Q67L	GGGACACTGCAGGTCTAGAACGTTTCCGTACTATCACTTCATC	GATGAAGTGATAGTACGGAAACGTTCTAGACCTGCAGTGTCCC
Ypt6 Q69L	GGGATACAGCAGGTCTGGAAAGATTTAGATCATTAATACCTTCATATATCAGA	TCTGATATATGAAGGTATTAATGATCTAAATCTTTCCAGACCTGCTGTACCC
Ypt7 Q68L	GGGATACTGCTGGACTGGAACGTTTCCAATCACTGGGTGT	ACACCCAGTGATTGGAAACGTTCCAGTCCAGCAGTATCCC
Ypt10 Q69L	GGGACACGGCGGGTCTGGAACGGTATAAATCACTGGTGCC	GGCACCAGTGATTTATACCGTTCCAGACCCGCCGTGTCCC
Ypt51 Q66L	GGGACACTGCTGGGCTAGAGAGATTTGCATCTTTAGCACCTATG	CATAGGTGCTAAAGATGCAAATCTCTCTAGCCCAGCAGTGTCCC
Ypt11 Q232L	GTGGGACACTGCGGGACTAGAACGGTACCAAAACGCAATCATTCC	GGAATGATTGCGTTTTGGTACCGTTCTAGTCCCGCAGTGTCCCAC

Arl3 Q78L	GGGATGTAGGTGGTCTAGAATCACTGAGATCAATGTGGTCCG	CGGACCACATTGATCTCAGTGATTCTAGACCACCTACATCCC
Arf1 Q71L	GGGATGTCGGTGGACTAGACAGAATTAGATCTCTATGGAGACAC	GTGTCTCCATAGAGATCTAATTCTGTCTAGTCCACCGACATCCC
Arf2 Q71L	GGGACGTCGGTGGACTAGACAGGATTAGATCTTTATGGAGACAC	GTGTCTCCATAAAGATCTAATCCTGTCTAGTCCACCGACGTCCC
Rho5 Q91H	GGGACACTGCAGGACACGAAGATTACGATCGTTTAAGACCGTTATG	CATAACGGTCTTAAACGATCGTAATCTTCGTGTCCTGCAGTGTCCC
Tem1 Q79L	CGATTTAGGCGGACTAAGAGAATTCATCATCAACATGCTCCC	TGGGAGCATGTTGATGAATTCTCTTAGTCCGCCTAAATC
Gsp1 Q71L	GGGATACTGCCGGCCTAGAAAAATTCGGTGGTGGTTTAAGAGC	GTCTCTTAAACCACCGAATTTTTCTAGGCCGGCAGTATCCC
Sec4 Q79L	TTGGGATACCGCTGGTCTAGAACGTTTCCGGACAATCAC	GGTGATTGTCCGGAAACGTTCTAGACCAGCGGTATCCCA
Gtr2 Q66L	GGAGCTTCCCGGGCTGCTAAATTACTTTGAACCGAGTTATGATTC	GAATCATAACTCGGTTCAAAGTAATTTAGCAGCCCGGGAAGCTCC
Arl1 Q72L	GGGATCTTGGTGGTCTAACAAGTATCAGGCCCTACTGGAGG	CCTCCAGTAGGGCCTGATACTTGTTAGACCACCAAGATCCC

Table 2: Primers used to create switch-of-function mutants. Point mutations are enlightened in red.

Proteins	Sense Primers	Anti-Sense Primers
Cdc42 S41A	CGACTATGTTCCAACAGTGTTCGATAACTATTCTGTGACTGTGATGATTGGTGATGAACCATATACGTTAGG	GGTTCATCACCAATCATCACAGTCACAGAATAGTTATCGAACACTGTTGGAACATAGTCGGCTG
Rho5 A41S	GTTCCAACGGTTTTTGATAATTATGCTACTACGATAGCTATCCCGAACGGAACTGC	GCAGTTCCGTTCGGGATAGCTATCGTAGTAGCATAATTATCAAAAACCGTTGGAAC
Rho5 L186K	GTCTCAAGAGGAAATAGATGAAAAGGTACAAAGATGTGGGTTTATGGGCTATACC	GCCCATAAACCCACATCTTTGTACCTTTTCATCTATTTCCTCTTGAGACACGTAATCGG
Cdc42 K186L	GTGGCCGCCTTGGAGCCTCCTGTTATCAAGAAAAGTGCAAAATGTGCAATTTTGTAG	CTACAAAATTGCACATTTTGCACTTTTCTTGATAACAGGAGGCTCCAAGGCGGCCAC
Rho5 P164K	GTGACCTAAGAGATGATAAGGCAACTCAGAAAAAATTGCAGGAAGCAAAC	GTTTGCTTCCTGCAATTTTTTCTGAGTTGCCTTATCATCTCTTAGGTCAC
Cdc42 K123P	ATTGATCTAAGGGATGACCCGGTAATCATCGAGAAGTTGCAAAGACAAAG	CTTTGTCTTTGCAACTTCTCGATGATTACCGGGTCATCCCTTAGATCAT

2.9. Differential interference contrast (DIC) and fluorescence staining microscopic imagery

To classify and identify the effects of the mutations on phenotypes. The following procedure was used. This procedure has been adapted and modified from Saito *et al.* [107] for fluorescence microscopy. For brightfield procedures, protocol previously described by Malleshaiah *et al.* [108] was used.

For cell morphology analysis, Differential interface contrast microscopy (DIC) and florescence based staining microscopy were used. Cells were grown overnight in Synthetic complete *–Ura* to make pre-culture. From the pre-culture, a log-phase culture was started from OD_{600} 0.05 up to OD_{600} 0.1 at 30ºC with shaking.

For culture and fixation, 20 ml of liquid media were dispensed into 100 ml sterile Erlenmeyer flask and cells were grown for 8 hours and then transferred into culture flask and incubated at 25°C in water bath overnight. Cells were transferred to disposal centrifuge tube and treated at 0.1 OD_{600} with 37% formaldehyde solution and 1M K-Pi buffer. Cells were collected by centrifugation (3000 rpm/5 min) and suspended in 37% formaldehyde, 1M K-Pi and distilled water for a later incubation at 25°C for 45 min. Again cells were collected by centrifugation and resuspended in 1 ml PBS for cell staining procedure.

For fluorescence staining, cells were transferred and suspended in PBS buffer in 1.5 ml centrifuge microtubes. After centrifugation (3000 rpm/1 min), cells were resuspended in PBS buffer. After two washing steps with PBS buffer, cells were incubated at 4ºC overnight in the dark in PBS buffer, 200 units/ml rhodamin-phalloidin (# 00027 Biotium, Inc. Hayward, CA, U.S.) and 10% triton X-100 for actin staining. After overnight incubation, cells were washed with PBS and P-buffer separately, resuspended and incubated at room temperature in P-buffer and 1 mg/ml FITC-ConA (Sigma-Aldrich, Oakville, ON, Canada) for cell membrane staining. Cells were washed again with P-buffer and collected by centrifugation, cells were drop in a mix of DAPI (AAT Bioquest, Inc.

Sunnyvale, CA, U.S.) with mounting solution for nucleus staining on slide glass. DAPI staining is normally performed after all other staining. Please bear in mind that fluorescence staining was only used to corroborate cell morphologies.

Preparation of microscopy plates for inverted microscope and DIC image analysis, a 96-well optical quality with clear bottom plates (NUNC) were used. For cells attachment to the wells, Concanavalin A (ConA # C-2631 Sigma-Aldrich, Oakville, ON, Canada) was used as a binding agent and each well was loaded with 0.1% of ConA solution at room temperature for 15 min. ConA was removed and wells were washed with sterile water. ConA was activated with 20 mM $CaCl_2$ and 20 mM $MnSO_4$ solution and incubated for 15 min at room temperature and washed once again with sterile water. Cells were added to wells and incubated for 10 min at room temperature allowing for cells attachment to wells.

For image acquisition by microscopy, after treatment, cells were observed in DIC and acquired using a NIKON eclipse TE2000-U right microscope connected to a CoolSNAP-fx CCD camera (Photometrics, Pleasanton, CA, USA) using 60X DIC H Plan APO oil objective using Metamorph-Image Analysis software for quantification (Molecular Devices, Downingtown, PA, USA). 349 pictures in total were acquired to obtain 7293 x 24 measures (Table 3) corresponding to each parameter to study the cellular phenotypes and calculate the shape factors (Section 2.11.). These parameters were used to calculate all the shape factors. For fluorescence imaging DAPI, FITC and Cy3 filters were used for the nucleus, cell wall and actin staining respectively and were analysed using Metamorph® image analysis software.

Table 3 : Set of total number of measured cellular phenotypes corresponding with the number of pictures acquired. Mean were calculated taking in account all four shape factors (elliptical shape factor, elongation shape factor, circularity shape factor and aspect ratio shape factor) in order to have a total average of measured wild type and mutants.

Strain	Quantity of measured cells	Number of pictures acquired at 60X	Shape factor mean
Wild type 1	1194	45	0.8102
Wild type 2	1218	61	0.7512
Cdc42 Q61L	1224	63	0.4181
Rho5 Q91H	1222	56	0.4909
Ras1 Q68L	1209	52	0.6382
Rsr1 G12V	1226	72	0.6450
Total of measured cells along with acquired pictures	7293	349	

2.10. Use of a budding yeast expression tool to observe changes in phenotypes

BG1805 Gateway expression vector was heat shock transformed into *BY4741 Saccharomyces cerevisiae* strains using frozen competent cells. Transformed cells were incubated for two days in petri dishes with YEPD (Yeast Extract Peptone Dextrose) media. Change in the protein concentration and protein activity were made through over expression of wild type and constitutively active forms of Ras-like small GTPases, using the *BG1805* yeast expression vector that contains an ampicillin resistance gene and amplifies in *E. coli.* The *BG1805* vector contains a Gal promoter and is cloned into the *BY4741* strain of *S. cerevisiae*.

2.11. Budding yeast cells measurements

To evaluate the shape of budding yeast cells and classify phenotypes, we used shape factors as a method for measuring cells. Cells were previous treated as described above. Shape factor formulas are normalized in a range from 0 to 1. These shape factor values are understood as 1 which is a perfect round or the ideal case of a perfect shape which can be compared to wild type cells due to its rounded form. Values less than 1 (<1) are the deviation or the irregular form of an object, in this case we can represent as mutant cells. This range of measurements for budding yeast are hypothetically represented and compared with objects (Figure 14).

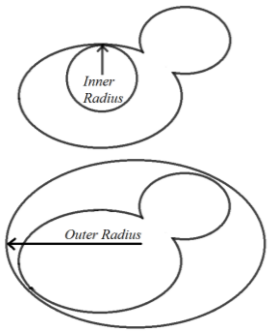
2.11.1 Shape factors formulas

Shape factors are measure index to describe different aspects of objects. However, different fields adopt different terms in order to describe the shape of the object [3]. To conduct image analysis and quantification of cellular phenotypes, shape factors were developed and adapted for cellular phenotypes (Section 1.7.) and were calculated using the following formulas:

2.11.1.1. Aspect ratio ($ARSF$):

The most common shape factor is the aspect ratio, a function of the inner radius and the outer radius orthogonal to it.

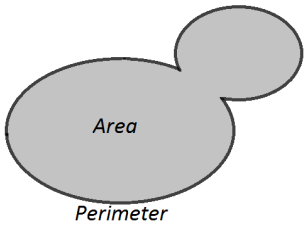
$$ARSF = \frac{Inner\ Radius}{Outer\ Radius}$$



2.11.1.2. Circularity shape factor (CSF):

Another very common shape factor is the circularity, a function of the perimeter (P) and the area (A). The reciprocal of the circularity equation is also used, such that CSF varies from one for a circle to infinity.

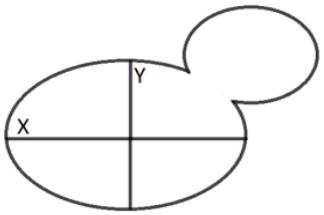
$$CSF = \frac{4\pi A}{P2}$$



2.11.1.3. Elongation shape factor (ESF):

The less-common elongation shape factor is defined as the square root of the ratio of the two second moments *X/Y* (i_n) of the particle around its principal axes.

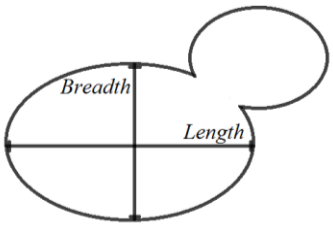
$$ESF = \sqrt{\frac{y\ axis}{x\ axis}}$$



2.11.1.4. Elliptical shape factor ($ELSF$):

Is the ratio of the object's breadth to its length. By definition, the length/breadth give values greater than one. Since the above shape factor formulas are normalized between 0 and 1, the elliptical shape factor formula were inverted as breadth/length in order to obtain values inside the range of 0 to 1 in order to have all shape factor measures normalized and observe their variation in cellular phenotypes.

$$ELSF = \frac{Breadth}{Length}$$



2.11.2. Shape factor descriptors values

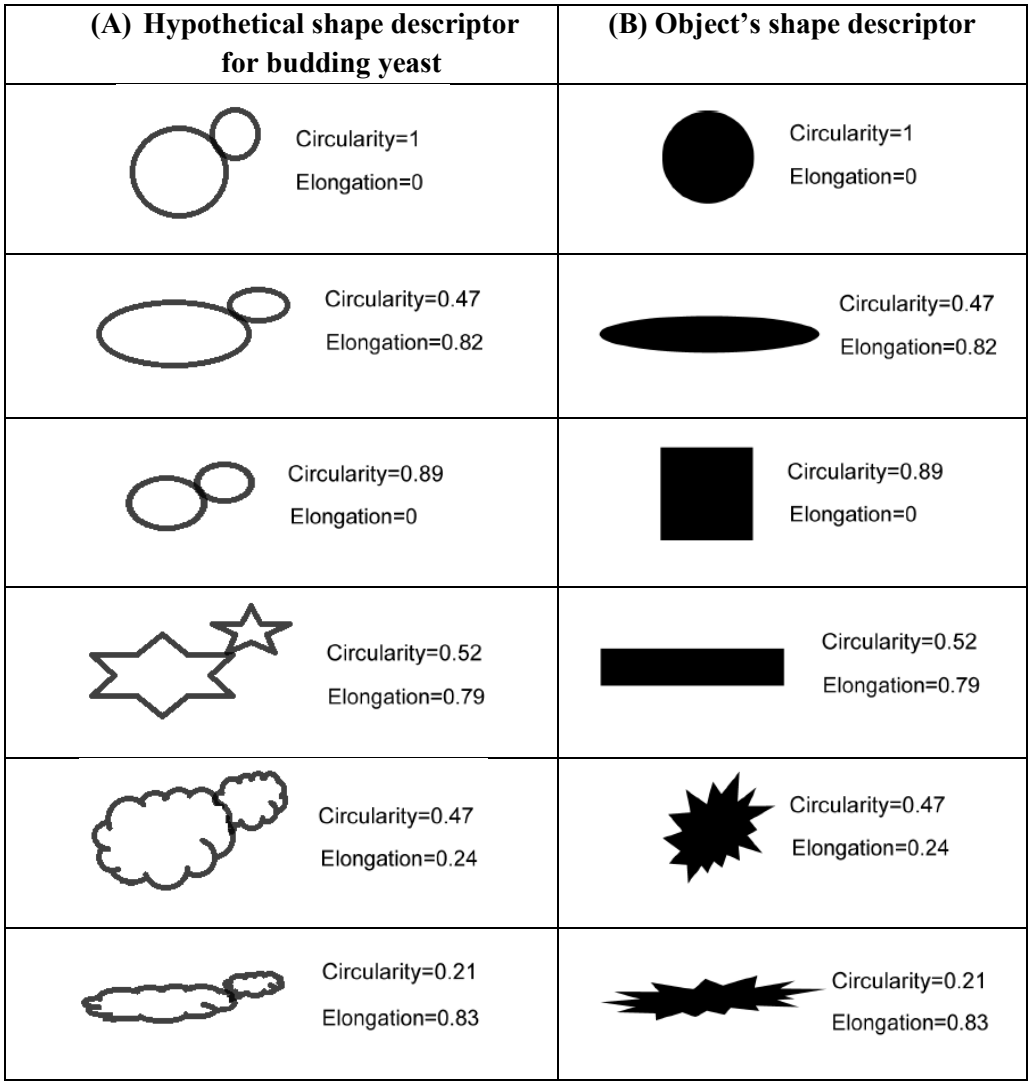
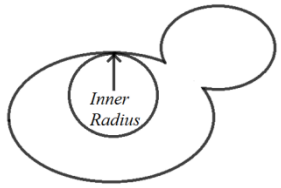
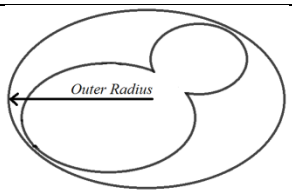
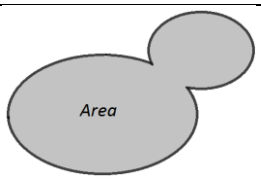
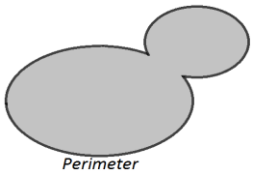
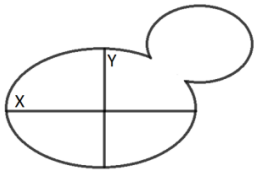
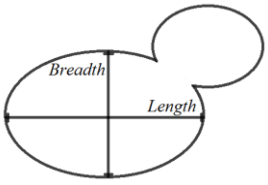


Figure 14: Illustration of Shape factor descriptor (A) Hypothetical shape descriptor for budding yeast cells compared with (B) object's shape descriptor. Values go in a range of 0 to 1. Values = 1 (perfect shape) < 1 (deformed shape). (Source: Malvern Instruments (www.malvern.co.uk)).

2.11.2. Measurement parameters

The following table explains each measurement parameter used to calculate shape factor formulas along with the shapes to help understand the values.

Table 4: Description of parameters used in shape factor formulas.

Parameter	Description	Shape
Inner radius	The distance from the centroid to the nearest point along the object's edge.	
Outer radius	The distance from the centroid to the farthest point along the object's edge.	
Area	The area of the object in calibrated units. Represent the area of the entire object.	
Perimeter	The distance around the edge of the object, measuring from the mid-points of each pixel that defines its border.	
Centroid X and Y	The point that represents the center of mass of the object. The X and Y coordinate of the centroid of the object.	
Length	The span of the longest chord through the object.	
Breadth	The caliper width of the object, perpendicular to the longest chord.	

2.12. Clustering methods

Shape factor measurements were obtained with Metamorph®. Upon completion, measurements results were saved on plain text file for statistical analysis and processing with R statistical language [109]. Phenotype classification between wild type and mutants were analysed using Hierarchical, *K*-Means and Fuzzy *C*-Means clustering algorithms whose were described above (Section 1.8.).

2.12.1. Clustering analysis packages and parameters using R

R Statistical Language provides with a big variety of functions for clustering algorithms developed which are applicable to solve a wide range of biological problems. All the shape factor measures were analyzed using different packages in R. The packages used in this study are; *Stats (hClust)* for Hierarchical analysis, *Stats (kmeans)* for the *K*-Means analysis and *Mfuzz* for Fuzzy *C*-Means analysis [110].

2.12.1.1. Data processing

Prior to analyze the data with clustering algorithms of all combinations of one wild type and one mutant, the data files from each wild type and mutant strains were concatenated in a single matrix. The distinct clustering algorithms (Hierarchical, *K*-Means and Fuzzy *C*-Means) were then applied on six subsets of all mutant strain with wild type (1 and 2), wild type 2 were used as a control.

2.12.1.2. Perform Hierarchical clustering

Hierarchical clustering is performed using the *hClust (stats)* function. This function performs the hierarchical clustering using a set of dissimilarities for the *n* objects being clustered using *K* (number of clusters) value. Cutree function was used to specify the number of clusters *K.* (i.e., the *K* value was set as 4). At start, each shape factor method is

assigned to its own cluster and then the algorithm proceeds iteratively, at each stage joining the two most similar clusters, continuing until there is just a single cluster. At each step of the algorithm, distances between clusters are recomputed using the Lance–Williams (1966) dissimilarity formula (i.e., commonly used with hierarchical clustering) [111]. The first dissimilarities (i.e., between each pair of shape factor measures) is computed with Euclidean distances. To find clusters of similar and dissimilar cellular shape factors, distances between clusters were measured using and Ward's methods which aim to minimize the variance within a cluster with similar shape factor measures.

Procedure:

1. Start by assigning each item to a cluster, so that if you have *n* items, you now have *n* clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size *n*.

2.12.1.3. Perform *K*-Means clustering

The *K*-Means clustering is performed with the *kmeans (stats)* function. The *K*-Means clustering perform on a matrix set with shape factors measures. The *K*-Means partition the shape factor measures into *K* groups such that the sum of squares from points to the defined cluster centres is minimized. In R, the *K*-Means algorithm uses the Hartigan and Wong (1979) algorithm method [112] by default. The *kmeans* function conforms a user-specific number of clusters, such that the within-cluster sum of squares (variance) and between-cluster sum of squares (covariance) from these centres is minimized, based on Euclidian distance. The numbers of cluster were calculated with the within-cluster sum of squares and the between-cluster sum of squares formula which give an approximate *K* value in order to choose *K* (number of clusters) (Figure 24). Classical multidimensional

scaling (mds = cmdscale) of the matrix were used to return the best-fitting *k*-dimensional representation of the shape factor data.

Procedure:

1. Place *K* points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the *K* centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated

2.12.1.4. Perform Fuzzy *C*-Means clustering

The Fuzzy *C*-Means is a clustering algorithm which allows shape factor measures to belong to two or more clusters. It is based on minimizing the degree of fuzzification *(m)* (i.e., *m* = values greater than 1) and the degree of membership. Fuzzy *C*-Means is performed using *cmeans (Mfuzz)* function on the same data matrix containing the shape factors measures. Fuzzy *C*-Means perform clustering using Euclidean distances where the mean square error is computed and the algorithm stops when the maximum number of iterations reached its default value (i.e., default value is 100). The data given is the shape factor measures which is clustered by generalized version of the Fuzzy *C*-Means algorithm, which use either a fixed-point or an on-line heuristic for minimizing the objective function;

$$sum\_i\ sum\_j\ w\_i\ u\_\{ij\}\hat{}m\ d\_\{ij\},$$

Where *w_i* is the weight of observation *i*, *u_{ij}* is the membership of observation *i* in cluster *j*, and *d_{ij}* is the distance (dissimilarity) between observation *i* and center *j* [110]. The dissimilarities used are the sums of squares (Euclidean distances) of the object differences.

Fuzzy *C*-means algorithm organized the shape factor data in 4 clusters (i.e., number of clusters $c = 4$) and a degree of fuzzification (m) greater than 1 ($m = 1.25$) [100, 113, 114].

Procedure:

1. Each data point belongs to two clusters to different degrees of membership and place two cluster centres.
2. Assign a fuzzy membership to each data point depending on the distance.
3. Compute the centre of each class; recalculate the positions of number of clusters *(c)* to its centroids.
4. Iteration stops when is a termination criteria between 0 and 1. As well, when there are not visible changes and each data point belongs to two clusters to a degree of fuzzification ($m = 1.25$).

2.12.1.5. Determination of the number of clusters for *K*-Means and Fuzzy *C*-Means

The number of clusters (K) for *K*-Means and (C) for Fuzzy *C*-Means were calculated using the within-cluster sum of squares (variance) and between-cluster sum of squares (covariance) (Figure 24). We determined the number of clusters by maximizing the covariance and minimizing the variance and we are able to select a K that is a trade-off between the within and the between cluster sum of squares when the intersection is form giving and approximate value of K (i.e., 2, 3, 4, 6) (Figure 24).

2.13. Identify residues that are involved in specific functions

Using a multiple sequence alignment and switch of function position based algorithm [2], specific residues that involve a specific function linked to a specific phenotype will be identified. In order to identify switch of function positions, we are using an algorithm that measures divergence distances using conservation matrices as amino acid identity, charge conservation and aromatic amino acid conservation [2, 16]. This approach will allow us to classify potential residues that will be mutated for the following switch of function study.

Figure 15 : Sequence alignment comparison between Cdc42 and Rho5. This alignment have been adapted from Heo *et al*.[2]. Such alignment shows a clear separation between conserved and divergent residues. Among the diverge residues we chose positions S41A, K123P and K186L for Cdc42p and A41S, P164K and L186K for Rho5p. These positions were mutated in order to switch residues between Cdc42p and Rho5p.

3. Results

3.1. Assembling a set of small GTPases

The Ras-like small GTPases super-family is a rich functional set of proteins which serves as working model to explore their influences on cellular morphology. They are subject of intense investigation since constitutively active mutants of Ras are able to induced marked phenotypes and are considered critical for cellular morphogenesis [88].

In this study we sought to answer how many distinct functions can proteins have in the same family and how these can be classified. In order to answer this question, budding yeast *Saccharomyces cerevisiae,* was used to classify phenotypes. As described above, the *BG1805* expression vector was used (Thermo Scientific Open Biosystems, Invitrogen) to express the protein of interest. To study the phenotypes and explore whether a classification can be done or not, two approaches were carried on; first, the whole set of native small GTPases were over-expressed in order to observe whether one can or cannot induce changes on cell morphology. This might provide a hint if there is a link between over-expressing native small GTPases with their function and therefore cell morphology.

However, we did not observe any changes in cellular morphologies as a result of over-expressing small GTPases (Figure 16). This suggest that over-expression of native small GTPases does not implicate a change in cell morphology. This result can be due to an imbalance in regulatory networks when small GTPases are over-expressed and that are quickly regulated to the normal state [115]. To obtain marked phenotypes and study the influence of small GTPases on cell morphology, constitutively active mutants of small GTPases were created.

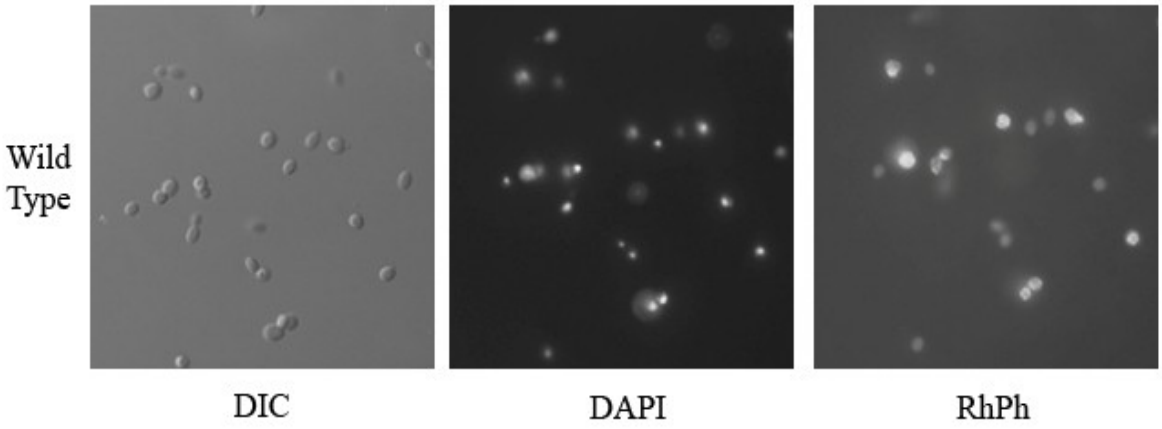


Figure 16: Over-expression of native small GTPases in *S. cerevisiae BY4741* strain. Cells were imaged at 60 X H-plan Apo oil in right microscope after 6 hours induction under Gal promoter using Differential interference contrast (DIC) and DAPI, FITC, Cy3 filters for fluorescence imagery. DIC imaging was performed to show the overall cellular morphology. DAPI and Rhodmaine phalloidin (RhPh) staining were used to show the nucleus and actin staining respetively.

3.2. Creation of constitutively active mutants and morphological profiling

Since over-expression of native forms of small GTPases did not produce any impact on cell morphology, a second approach was utilized. Previous work by Heo *et al.* [2] demonstrated in a mammalian system that constitutively active mutants were capable of inducing changes in cell morphology. Based on these observations, constitutively active mutants of small GTPases were created in budding yeast. This approach was developed as follows:

-Identify residues that are to be mutated using multiple sequence alignment of small GTPases protein sequences in budding yeast with other species (Section 3.2.1)

-Create constitutively active mutants and over-express these mutant small GTPases (Section 3.2.2)

-Classify the phenotypes with DIC microscopy (Section 3.2.3)

3.2.1. Identify residues that are to be mutated using multiple sequence alignment of small GTPases protein sequences in budding yeast with other species

A multiple sequence alignment which includes sequences from *Homo sapiens, Mus musculus, Drosophila melanogaster, Oryzias latipes, Rattus norvegicus, Schizosaccharomyces pombe* and *Saccharomyces cerevisiae* as the main species, was performed using Jaview2 (Section 2.7.). This alignment allows identification of residues among species in the G1/G3 domains (Figure 17). The G1/G3 domains show a conservation of glycine (G12) and glutamine (Q61) which provides essential GTPase and nucleotide exchange activity. These mutated domains confer a constitutive activity to the

proteins preventing regulation by GAPs. [32, 88]. These two domains were chosen in order to modify the activity of small GTPases. While these morphological results give significant input into cellular functions of these Ras-like small GTPases, it also can provide an insight into the sub network of protein-protein interactions.

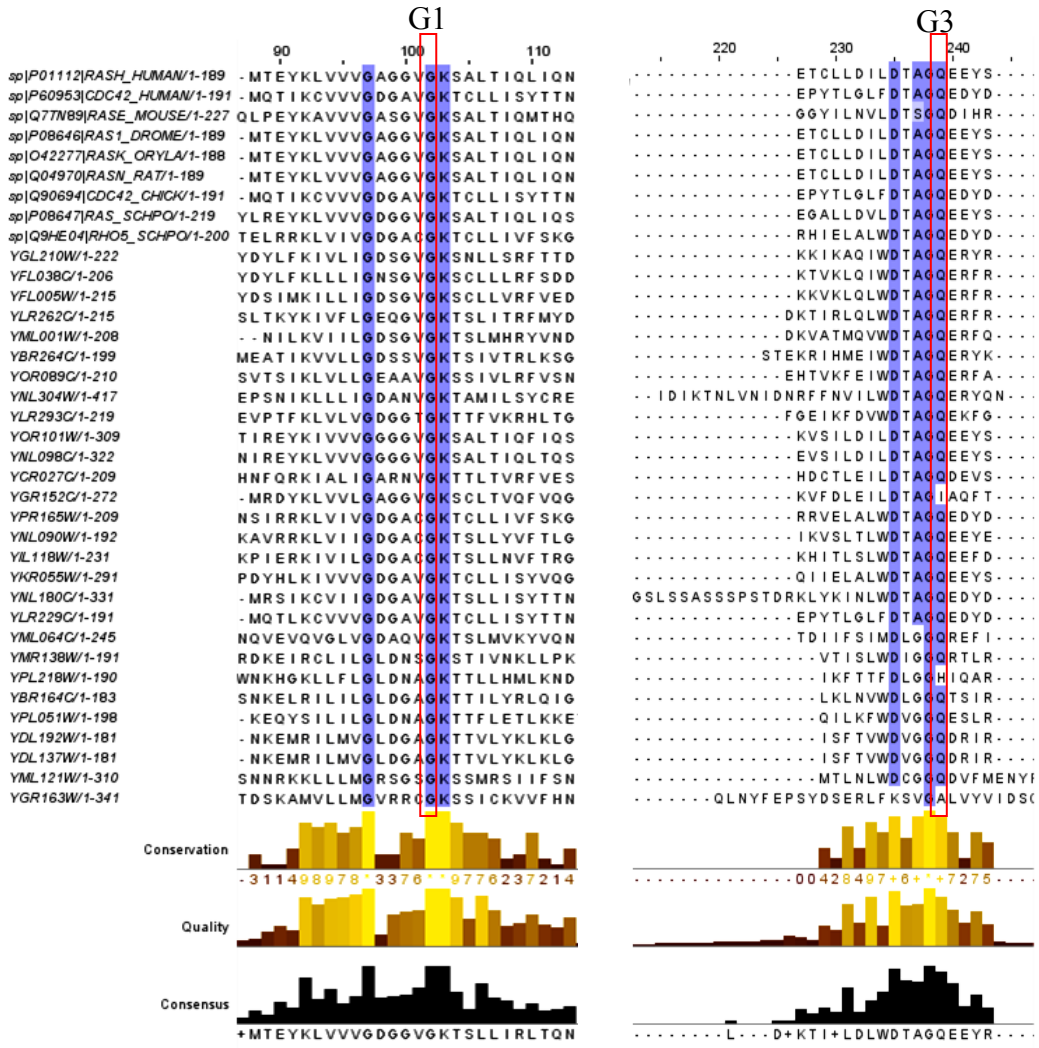


Figure 17: G1/G3 sequence alignment of small GTPases**.** Alignments corresponds to *Homo sapiens, Mus musculus, Drosophila melanogaster, Oryzias latipes, Rattus norvegicus, Gallus gallus, Schizosaccharomyces pombe,* and whole super family corresponding to *Saccharomyces cerevisie* as main specie. The conserved residues in G1 and G3 domains are identified by blue highlight column for all species. The G1 domains feature the G12 position and G3 contains the Q61 positions, together these domains are involved in constitutive activity of proteins [88]. Residues are framed in red.

3.2.2. Create constitutively active mutants and express these mutant small GTPases

Experimental classification of phenotypes is a preferable strategy in some cases over sequence homology classification to distinguish among functional classes [2]. Based on this, point mutations in G1/G3 domains were introduced using Quickchange protocol (Section 2.8). The G1/G3 domains are known to be responsible for specific interaction with GDP and GTP and thereby GTPase activity [116]. These domains can be easily replaced by valine (V) at position 12 in G1 domain and by leucine (Q) or histidine (H) at position 61 in G3 domain in order to confer a constitutive activation to the protein [88]. These substitutions allow creating constitutively active forms or mutants of small GTPases. Therefore, these mutants enable specify expression and differentiation cellular programs as indicated by phenotypes.

To test this notion, Cdc42 and Rho5 from the Rho-like small GTPases subfamily were chosen to visualize the induced phenotypes. Cdc42 and Rho5 are particularly known to regulate cell polarity, actin and septin organization and cell morphology integrity respectively [27, 48]. These functions are completely disrupted, when these proteins are expressed as constitutively active mutants. Expression of Cdc42 Q61L results in cells expressing amorphous/starfish-type morphology and expression of Rho5 Q91H results in elongated/clumped-type cells (Figure 18). Apart from belonging to the same subfamily, these two proteins are capable of inducing two different phenotypes. They show a 46.1% identity in DNA sequences and are involved in activating two different pathways like the MAPK and the osmotic stress response for Cdc42 [24] and Rho5 [61] respectively. The functionality of these pathways is accomplished by known physical interactions which have been extensively studied and reported in the literature. [117]. Similar approach was performed with Ras1 and Rsr1 which belong to the Ras-like small GTPases and have 62.6% identity. Here, we also observed morphological changes to be induced by these constitutively active forms and allows visualizing effects on cell morphology in this

subfamily (Figure 19). Percentage of identity among sequences for small GTPases was calculated with LaLign web server (http://www.ch.embnet.org).

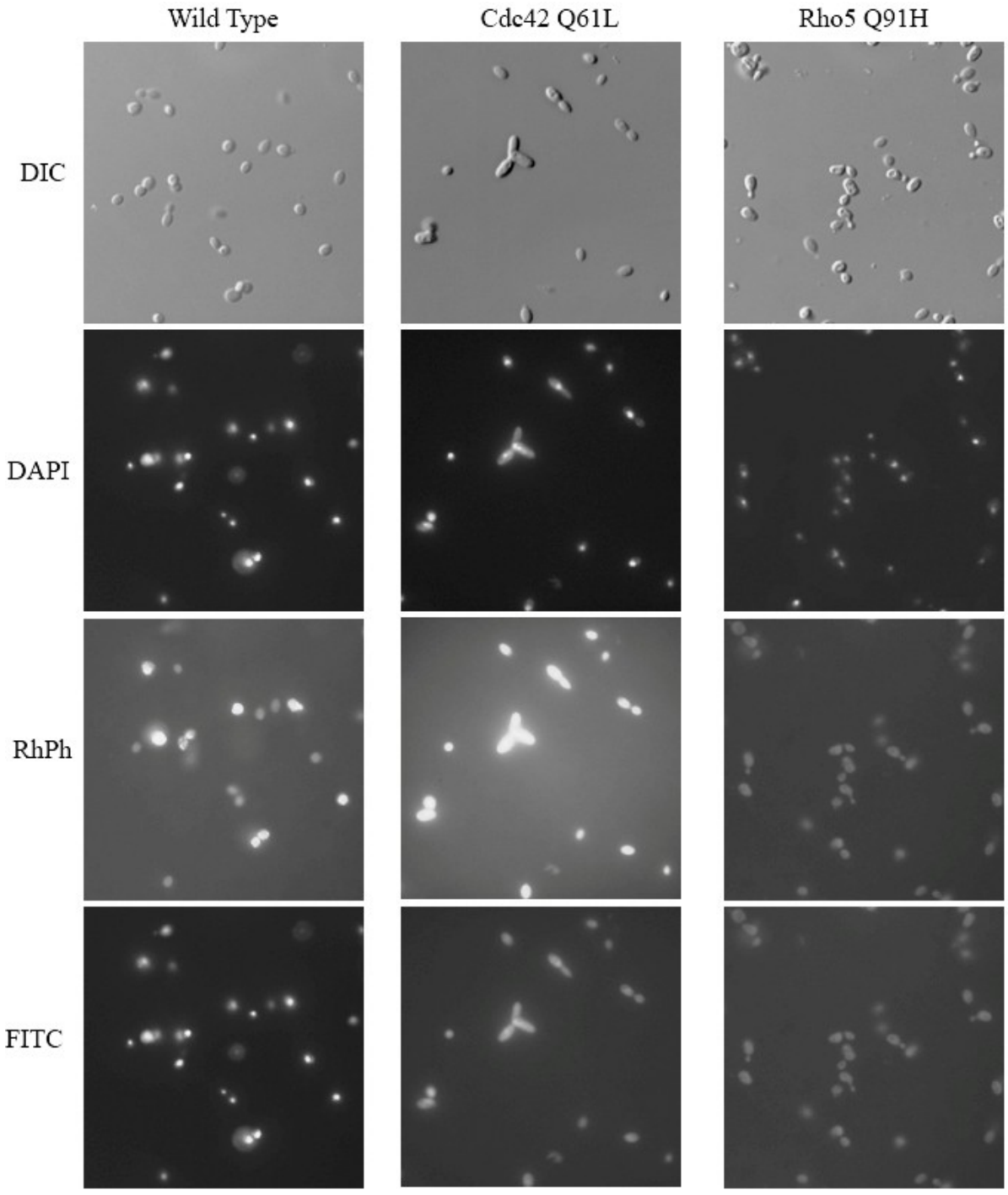

Wild Type
Cdc42 Q61L
Rho5 Q91H
DIC
DAPI
RhPh
FITC

Figure 18: Images of constitutively active Cdc42 (amorphous), Rho5 (elongated/clumped) and wild type cells. Mutants and wt *BY4741* cells were acquired using differential interference contrast (DIC) and DAPI, FITC, Cy3 filters for fluorescence imagery, 100 X H-plan Apo oil in right microscope after 6 hours induction under Gal promoter. Site-directed mutations Cdc42p Q61L and Rho5p Q91H were based on human k-Ras Q61L. DAPI and Rhodamine phalloidin (RhPh) were used for nucleus and actin staining respectively.

3.2.3. Classify the phenotypes with DIC microscopy

Creation of constitutively active mutants and induction of phenotypes were done. To question, whether the resulting phenotypes are really due to the expression of mutant proteins or whether they are artifacts due to the galactose background; the cells were stained using 4',6-diamidino-2-phenylindole (DAPI) for nucleus, Rhodamine phalloidin (RhPh) for actin and Fluorescein isothiocyanate (FITC) for cell wall staining (Section 2.9.). Ending this procedure, mutant cells were perfectly observed and respectively classified. This classification was linked to disrupted functions due to point mutations in G1/G3 domains.

To study the influence of constitutively active mutations on cell morphology, DIC and fluorescence images were acquired following the described protocol (Section 2.9.). These images show difference in phenotypes that were caused by disruption of cellular functions in charge of the maintenance of cell morphology (Figure 19). Mutant version of Cdc42 Q61L fails to form a bud under normal condition. Even though activity of this protein is turn into constitutive active, it is known that DNA replication, nuclear division and increase in cell mass continues, and disruption of these events results in an amorphous/elongated/starfish-like mutant phenotype [24]. It is widely known that among the Rho-like small GTPases Cdc42 and Rho1 are the most studied ones because of their well establish roles in cell morphology [27]. Moreover they show a 58.0% sequence identity. Regardless of these observations, we choose Rho5 since the induced phenotype is easily distinguishable and Rho5 feature an effector domain that is very similar to Cdc42 [118]. On the other hand Rho1 did not express a well define phenotype. Rho5 Q91H induces an elongated/clumped-like phenotype. This phenotype it is due to the deregulation of the cell integrity pathway where Rho5 plays a pivotal role [48]. It is important to know that Rho5 mutants seems to show slightly morphological differences depending on the strain used for the experiment [24], here *BY4741* was used.

Following the observations with Rho-like small GTPases we also assayed proteins from the Ras-like small GTPases subfamily as mention before (Section 1.5.4.). Mutant

version of Ras1 Q68L expresses an enlarged/round phenotype with presence of big vacuole (Figure 19). Ras1 Q68L is abnormally rounded and displays an arrest G1/S phase transition. This abnormality is due to disruption of adenylyl cyclase activating pathway that was previously reported [24, 34]. This subfamily also feature Ras2, Rsr1 and Rhb1 (Section 1), and hence Rsr1 was chosen as it is involved in cell morphology regulation as well. Constitutively active mutant Rsr1 G12V shows a very discrete large phenotype that is very difficult to distinguish from wild type cells (Figure 19) and more specific features are required to discriminate these phenotypes. In an effort to distinguish phenotypes in an unbiased manner, shape factors were utilized as a method for morphological classification. The phenotype classification was based on a previous classification described by Giaever *et al*. [22] (Figure 6).

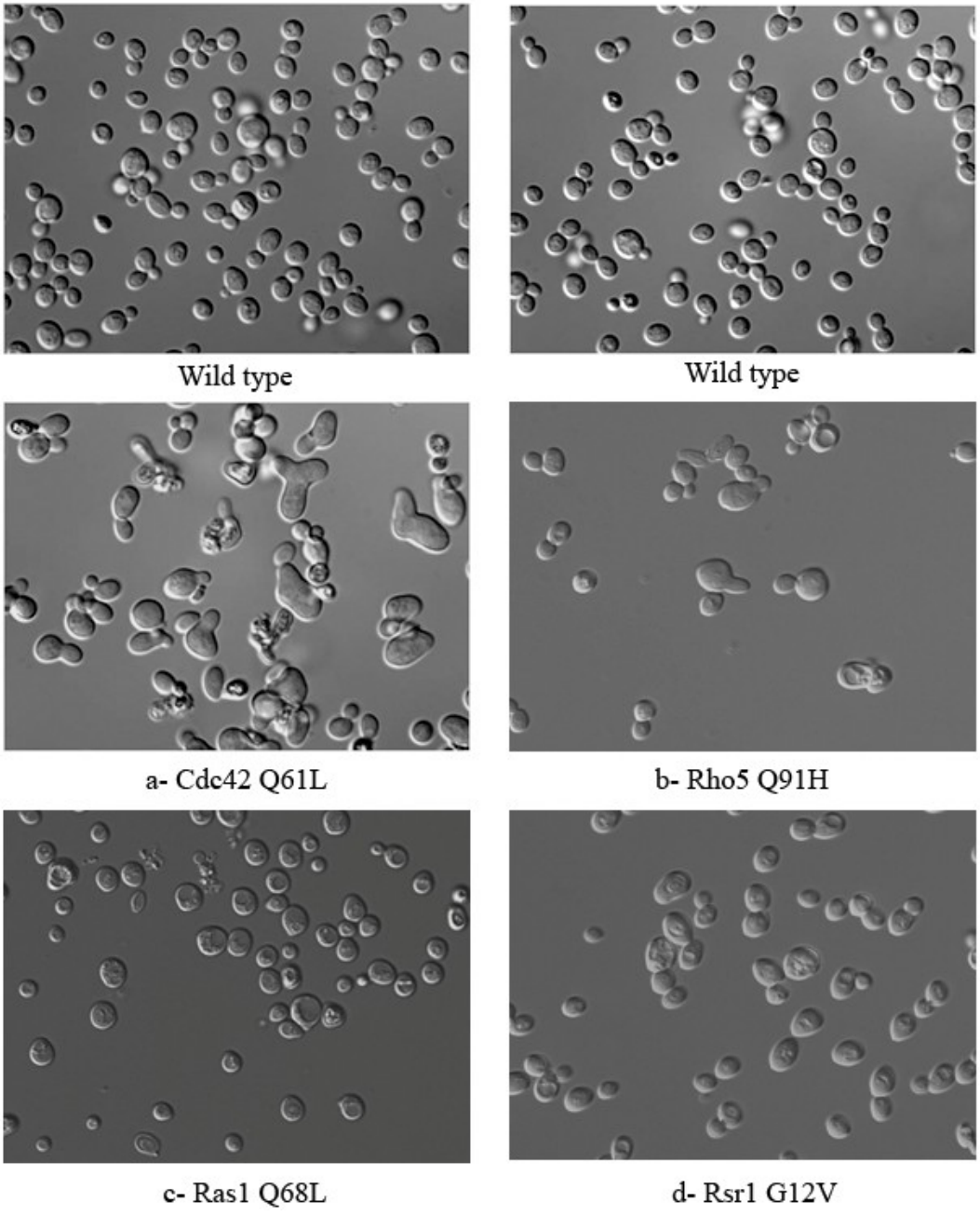


Figure 19 : Morphologies of cells expressing small GTPases with point mutations in comparison with cells expressing native small GTPases. Morphotypes were induced under 2% galactose for 6 hours. DIC images were obtained at 60 X H-plan Apo oil using inverted microscope.

3.3. Shape factor of cells allows a benchmarking of phenotypes

Phenotypic classification is an important step in order to study the switching of phenotypes and evolution of protein-protein interaction networks. In this step, a given numerical value has been assigned to each phenotypes and to this end, Metamorph® Image Analysis Software was used to analyse the acquire images. This was followed by analysis with different clustering methods using R statistical language [109]. Therefore the following steps were carried out:

-1st analysis: Manual quantification (Section 3.3.1)

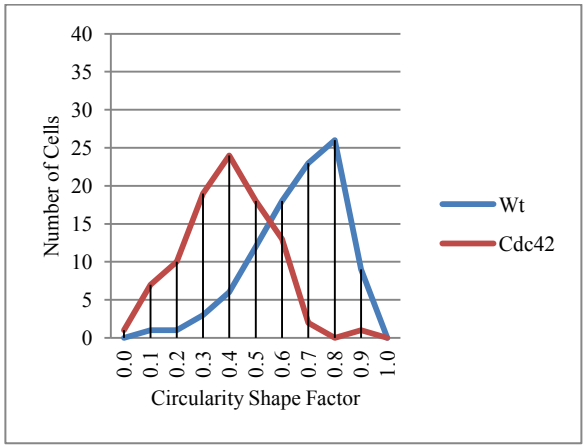
-2nd analysis: Simultaneous use of shape factors and macro design (Section 3.3.2)

-3rd analysis: Clustering methods applied using R statistical language (Section 3.3.3)

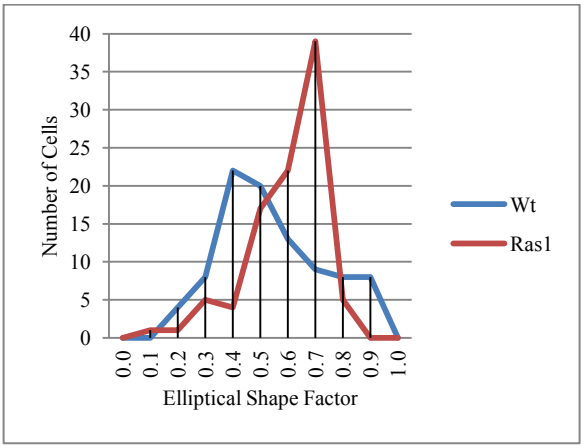
3.3.1. 1st analysis: Manual quantification

A total of 24 parameters were obtained by measuring cell shapes, which includes pixel area, area, average intensity, total intensity, radial dispersion, perimeter, centroid X, centroid Y, orientation, width, height, length, breadth, inner radius, outer radius, mean radius, equivalence radius, equivalence sphere, texture different moment, circularity shape factor (CSF) and elliptical shape factor (ELSF). These last two (CSF and ELSF) were already calculated by Metamorph®. From these parameters elongation shape factor (ESF) and aspect ratio shape factor (ARSF) were calculated by applying their respective formulas (Section 2.11.1.), since ESF and ARSF were not included in Metamorph®. Data is exported to Excel file where the next analyses were performed. After obtaining these measurements, shape factor indices were chosen because they are normalized formulas ranging from 0 to 1, which makes it more appropriate and defined for analysis.

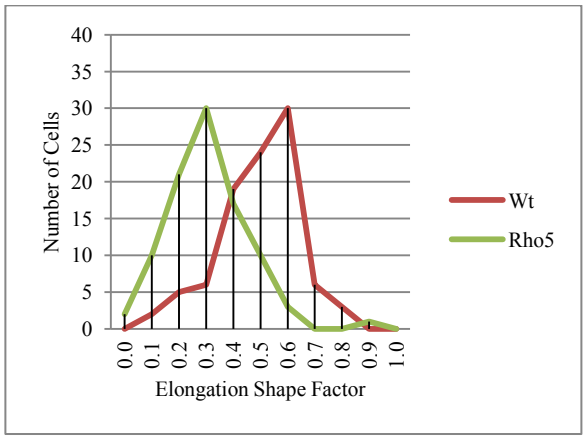
We hypothesized that shape factors could be assigned to a respective cell phenotype depending on the shape. In order to prove this rationale, each observed phenotypes were assigned to one shape factor manually. This approach was based on previous observation where the shape factors are exposed as shape descriptors (Figure 14). One hundred wild type cells and one hundred mutant cells were manually assigned according to the phenotype and the shape factor value. Cdc42 Q61L, Rho5 Q91H, Ras1 Q68L and Rsr1 G12V were plotted and a shift in cell populations among wild type and mutants were observed, base on these (Figure 20). We resumed the range of characteristic shape factor measures for each mutant strain in (Table 5).



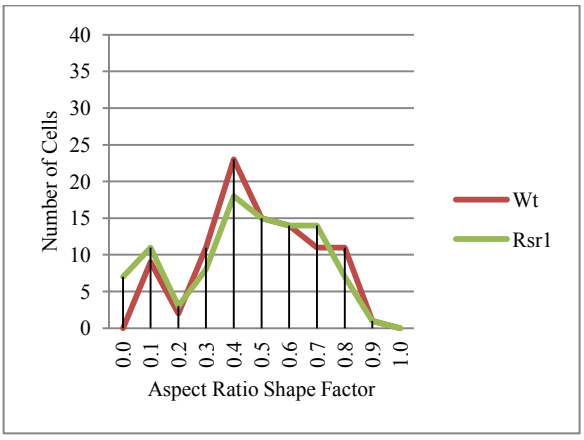
A- Wt vs Cdc42 Q61L



C- Wt vs Ras1 Q68L



B- Wt vs Rho5 Q91H



D- Wt vs Rsr1 G12V

Figure 20: Comparison of tendency curves of shape factors used with 4 different mutants. Each panel displays normalized shape factors and manual quantification of total 200 cells, in each case including wild type (Wt) and mutant, A) Wt vs Cdc42 Q61L, B) Wt vs Rho5 Q91H, C) Wt vs Ras1 Q68L, and D) Wt vs Rsr1 G12V.

Table 5: Range of characteristic shape factor measures. For each mutant displaying a phenotype, a shape factor and shape factor value was assigned.

Protein	Shape Factor	Wild Type	Mutant
Cdc42 Q61L	CSF	0.8<CSF>=1	CSF<0.8
Rho5 Q91H	ESF	0.7<ESF<=1	ESF<0.7
Ras1 Q68L	ELSF	0.6<ELSF<=1	0.5<ELSF>1
Rsr1 G12V	ARSF	0.65<ARSF<=1	0.6<ARSF>1

Expression of Cdc42 Q61L (Figure 20 A, Table 5) and Rho5 Q91H (Figure 20 B, Table 5), show similar patterns as these two mutants have a tendency to be under 0.8 and 0.7, respectively. For Cdc42, circularity shape factor (CSF) formula was used, which measures the area and perimeter of the cell. Cdc42 expresses an amorphous/elongated shape, thus measurements results in values below 0.8. Elongation shape factor (ESF) was assigned to Rho5 which express elongated/clumped phenotype. This formula calculates the X and Y axis of the cell, which for Rho5 mutants is measuring below 0.75. These values suggest that the phenotypes resulting from expression of Cdc42 and Rho5 variants are distinguishable from wild type cells.

To analyze images from expression of Ras1p Q68L (Figure 20 C, Table 5), elliptical shape factor (ELSF) was assigned. It simply calculates the breadth and the length of the cell. Ras1 shows enlarged phenotypes placing values between 0.5 and 1, being similar to wild type with values between 0.6 and 1. The phenotypes resulting from expression of Rsr1 G12V (Figure 20 D, Table 5) phenotypes were calculated using the ARSF formula which measures the inner and outer radius. In this particular case, these shape factors failed to distinguish wild type cells. Cell morphology and shape factor thresholds for Cdc42, Rho5, Ras1 and Rsr1 can be observed in (Figure 19, Table 5) respectively.

In summary, this first analysis was not highly efficient, mainly due to the little small number of cells that were picked for analysing one hundred cells for each wild type and mutants. These quantities of cells were not sufficient for a statistical analysis, since, at which percentage this point mutation is capable of impact on budding yeast population, it is important to know. Also, manual threshold is not the best approach to pursue as it may also cause a bias in the results.

3.3.2. 2nd analysis: Simultaneous use of shape factors and macro design

Analysis of the images using manual threshold identified the need to increase the number of cells imaged and secondly the need to create a method that automated the analysis of the phenotypes. *Visual Basic Editor in Excel* (VBE) was used for macros development and 7293 cells were analysed. This number of cells will give an idea of, at which percentage of the population, point mutations are capable to have an effect. In addition, after acquiring the data from 7293 cells, a manual quantification will be time consuming and will confuse the results.

Therefore a macro was written to measure the phenotypes of cells using shape factors to discriminate them through organization of the data in groups based on "shape factor thresholds". Moreover, this method might discriminate such phenotypes from a pool of cells expressing various mutants. To test this strategy, we set up a pool of wild type cells harboring empty *BG1805* vector, cells expressing wild type forms of Cdc42, Rho5, Ras1 and Rsr1 and mutant cells expressing Cdc42 Q61L, Rho5 Q91H, Ras1 Q68L and Rsr1 G12V. Shape factor formulas were calculated however, no discrimination whatsoever was done (Figure 21). This suggested that a data classifier is required. In order to overcome this, a macro was written using *Visual Basic Editor* (VBE) based on the threshold values given by Malvern® (Figure 14). The function of this macro is to arrange the data in clusters. Each cluster identification (Cluster ID) was set from 1 to 4 depending on values shown on (Table 6). This was applied for each mutant, for example, cluster 1; measures from 0.1 to

0.39 cluster 2; measures from 0.4 to 0.59, cluster 3; measure from 0.6 to 0.79 and cluster 4; measures from 0.8 to 1. Normal cells were expected to fall into cluster 4 since measures closer to 1 are wild type with rounded shapes. Every other measure under 0.79 was corresponding to mutants. Mutant's measures tend to be away from 1 because they show a phenotype with deformed shape. However, using this type of clustering, we were not able to group the data in 4 clusters, but into 3 clusters. Therefore, the classification still required more clarity (Figure 22).

Table 6: Values for phenotype classification using macro.

Phenotype	**ARSF**	**ESF**	**CSF**	**ELSF**
Wild Type	0.8<SF>=1	0.8<SF>=1	0.8<SF>=1	0.8<SF>=1
Cdc42 Q61L (amorphous)	0.1<SF<0.47	0.1<SF<0.47	0.1<SF<0.47	0.1<SF<0.47
Rho5 Q91H (elongated/clumped)	0.1<SF<0.82	0.1<SF<0.82	0.1<SF<0.82	0.1<SF<0.82
Ras1 Q68L (rounded)	0.1<SF<0.89	0.1<SF<0.89	0.1<SF<0.89	0.1<SF<0.89
Rsr1 G12V (discrete large)	0.1<SF<0.79	0.1<SF<0.79	0.1<SF<0.79	0.1<SF<0.79

In summary the following were encountered on 2nd analysis:

a) The macro alone is not powerful enough to quantify and classify a set of 7293 measures per 21 parameters.
b) A manual threshold is not the best approach to take. More variables are needed.
c) An unclear distinction between mutants and wild type.

To this end, different statistical analysis and clustering methods were applied to the data to discriminate the phenotypes clearly.

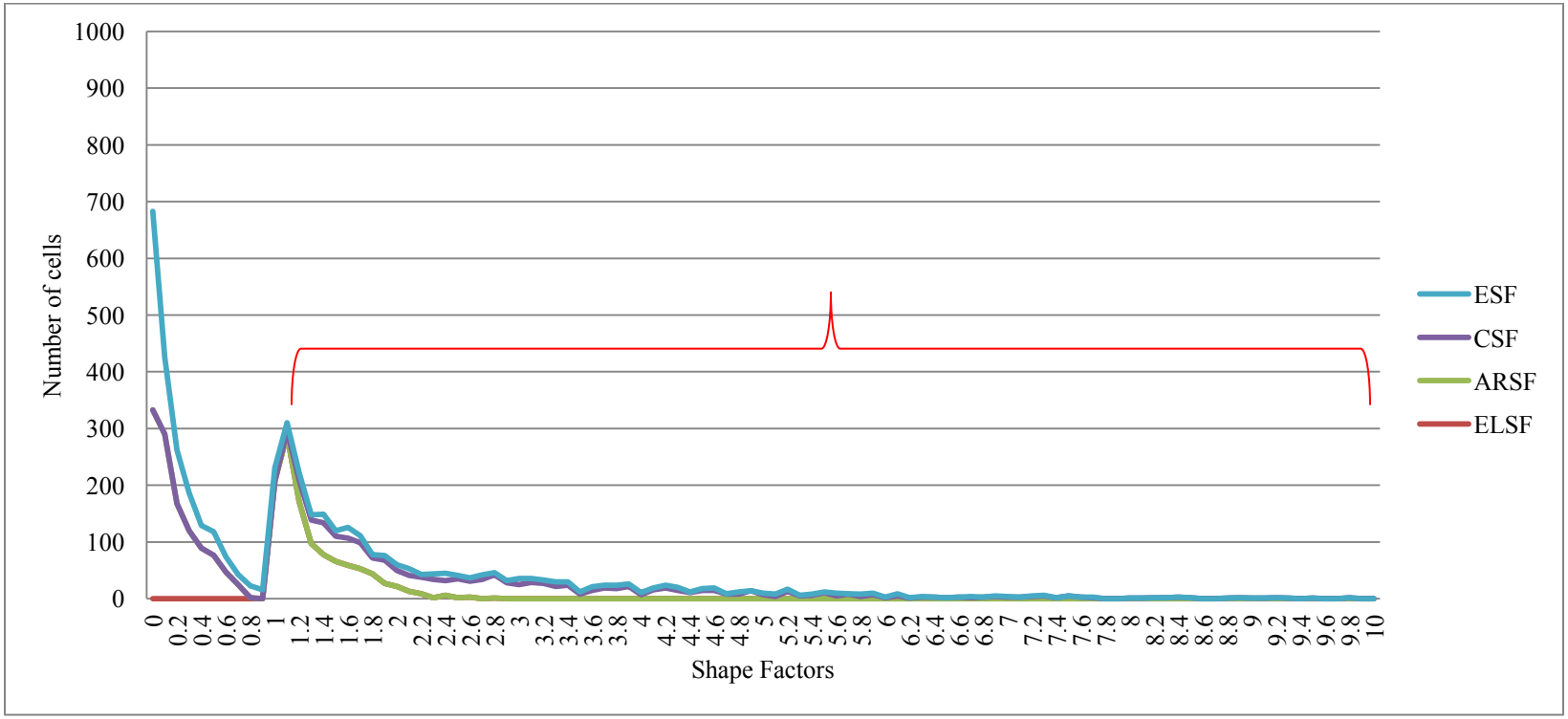


Figure 21: Simultaneous use of different shape factors in a pool of wild type and mutants of 7293 cells. Results higher than shape factor 1, measurements are considered as cell aggregates as is shown between red brackets. ESF (elongation shape factor), CSF (circularity shape factor), ARSF (aspect ratio shape factor), ELSF (elliptical shape factor).

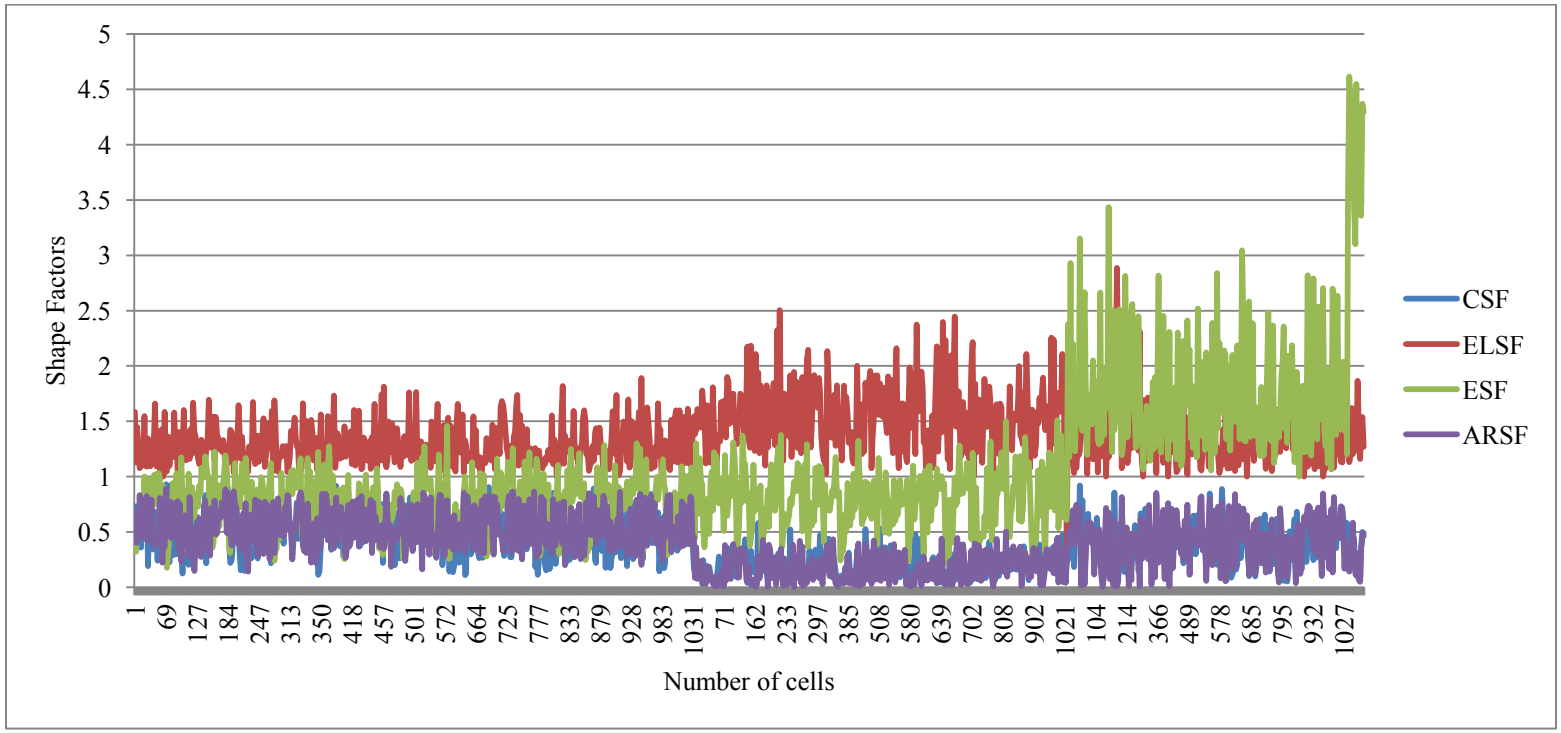


Figure 22: Quantification of 7293 cells using macro. Cells are distributed into three groups. Groups 1, 2 and 3 represent the distribution of wild type, mutants and cell aggregates. These measures were obtained calculating the ESF (elongation shape factor), CSF (circularity shape factor), ARSF (aspect ratio shape factor), ELSF (elliptical shape factor) for each cell.

3.3.3. 3rd Analysis: Clustering methods applied using R statistical language

Next, R statistical analysis was used and three different clustering methods [109] were applied to the data obtained from 7293 cells. Each of these clustering algorithms, was controlled under clustering quality measures (CQ_m) which gives significance to each computed cluster and each clustering method [101] (Section 1.8.4.) Note that $CQ_m < 75\%$ were considered to be poor in significance. CQ_m values are based on calculation of the total CQ_m, which is explained later. The following approach was taken to analyse the data obtained from measures of 7293 cells:

-Hierarchical clustering algorithm (hard clustering) (Section 3.3.3.1)

-*K*-Means clustering algorithm (hard clustering) (Section 3.3.3.2)

-Fuzzy *C*-means clustering algorithm (soft clustering) (Section 3.3.3.3)

3.3.3.1. Hierarchical clustering algorithm

The hierarchical method seeks to build a hierarchy using shape factors. Shape factor measures were normalized between 0 and 1. Distances between two clusters were measured with Euclidean distances that is combined with Ward's algorithm [110]. It builds a hierarchical clustering of wild type and mutant cells. In addition, this method takes into account the difference between two samples that is directly based on changes at the level of shape factor measurements. With this method, a hierarchical tree for cellular phenotypes was created; as well the statistical significance of different populations was calculated. However, after applying this algorithm to shape factors, no major classification was observed (Figure 23 A). In fact, hierarchical clustering takes every cellular phenotype as the same since it is not able to cluster into different groups (wild type and mutant). The CQ_m was calculated for all clusters and only Ras1 and Rsr1 showed significance with 82%

and 85%, respectively (Figure 23 B). This clustering result suggests that hierarchical clustering is not suitable for phenotypic discrimination; however, it provides us with information that other algorithms should use for such purposes.

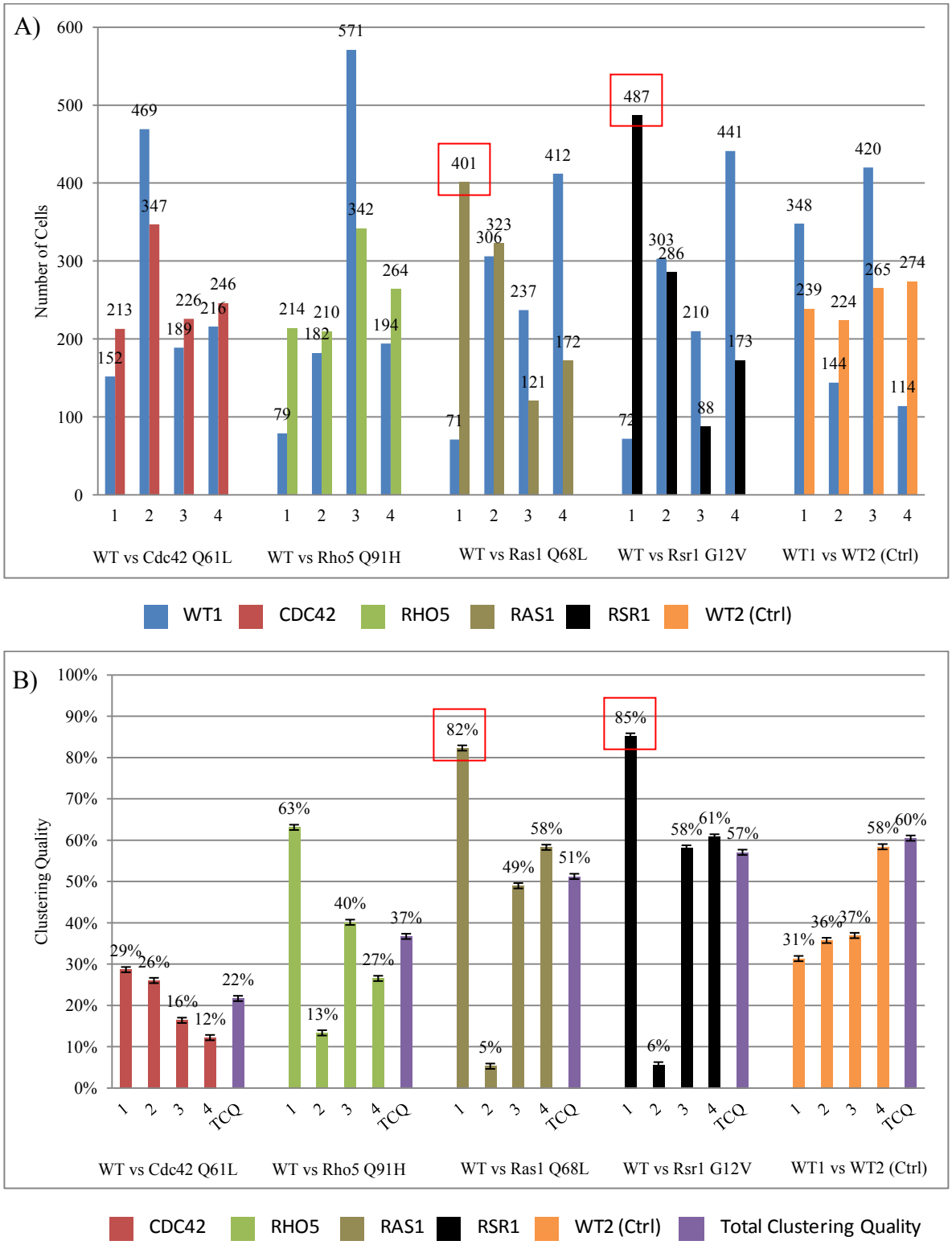

A)
Number of Cells
600
500
400
300
200
100
0
152
213
469
347
189
226
216
246
79
214
182
210
571
342
194
264
71
401
306
323
237
121
412
172
72
487
303
286
210
88
441
173
348
239
144
224
265
420
114
274
1 2 3 4
WT vs Cdc42 Q61L
WT vs Rho5 Q91H
WT vs Ras1 Q68L
WT vs Rsr1 G12V
WT1 vs WT2 (Ctrl)
WT1
CDC42
RHO5
RAS1
RSR1
WT2 (Ctrl)
B)
Clustering Quality
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
29%
26%
16%
12%
22%
63%
13%
40%
27%
37%
82%
5%
49%
58%
51%
85%
6%
58%
61%
57%
31%
36%
37%
58%
60%
1 2 3 4 TCQ
WT vs Cdc42 Q61L
WT vs Rho5 Q91H
WT vs Ras1 Q68L
WT vs Rsr1 G12V
WT1 vs WT2 (Ctrl)
CDC42
RHO5
RAS1
RSR1
WT2 (Ctrl)
Total Clustering Quality

Figure 23: Hierarchical clustering results represented in histograms A) Clustering were performed using Euclidean distances in combination with Ward's method, where the distance of the cell to its centroid is minimized. These two parameters defines group of cells by hard clustering (one cell belongs only to one cluster) B) clustering quality represented in percentages. Only clusters of Ras1 and Rsr1 shows high percentage. Numbers of clusters were chosen arbitrary and are represented by 1, 2, 3 and 4. Total clustering quality (TCQ) for each mutant was calculated to obtain the average TCQ of the four clusters. Significant clusters are framed in red. Wild type 1 and wild type 2 serve as a control representing the homogeneity of the population where clustering cannot be efficient.

3.3.3.2. *K*-Means clustering algorithm

Next *K*-Means algorithm was applied to the data. This method aims to partition *n* observation in *K* clusters (Section 1.8.2.) and could cluster the closest shape factor measures (wild type together and mutants together) by calculating their nearest mean. A key step in *K*-means clustering first is to choose the best number of clusters (*K*). For that purpose, *K* value was determined by applying the within and the between cluster sum of squares that is calculated by taking into account the shape factor values in this particular case (Figure 24). The *K*-Means algorithm performed better than hierarchical clustering for our phenotypic classification (Figure 25 A). Indeed, the CQ_m values show that almost all the clusters have a significant percentage and confirm that *K*-Means is a good method to discriminate phenotypes (Figure 24 B). For instance, Cdc42 show two significant clusters while Rho5 show 3 significant clusters and Ras1 with Rsr1 display 4 significant clusters.

Since *K*-Means is a hard clustering method, these results suggest that this approach might gave a better clustering output if is considered a more flexible version. Therefore *C*-Means algorithm can be applied. Basic steps to perform *K*-Means are as follows, a) choose the best number of *K* and b) perform *K*-Means clustering.

3.3.3.3.1. Choosing the number of clusters

To choose the best number of clusters, the between cluster sum of squares and the within cluster sum of squares have been calculated which resulted in a value above 2 where the blue and red lines intersected (Figure 24). These values gave us a hint that the optimal number of clusters (*K*) is higher than 2. We also decided that a pair number of clusters was a good strategy for our bipartite comparison between each respective mutant strain and the wild type strain. So we tested *K* values of 2, 4, 6 and 8. *K* values above 4 generated empty clusters with little quantity of elements and this confirmed that *K*=4 is the best numbers of clusters to analyze our data.

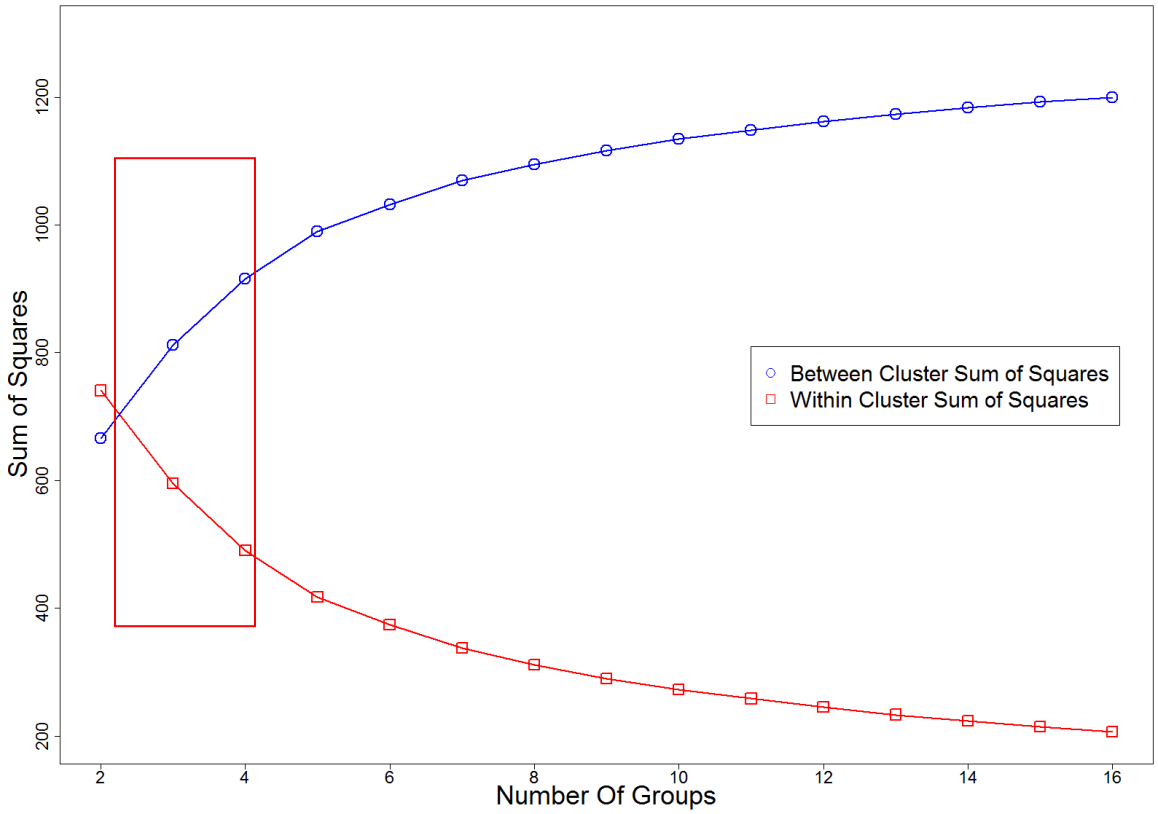


Figure 24: Calculation of the between cluster sum of squares (blue) and the within cluster sum of squares (red). An intersection of the blue and red lines can be observed above 2 suggesting that the number of clusters must be chose between 3 and 6.

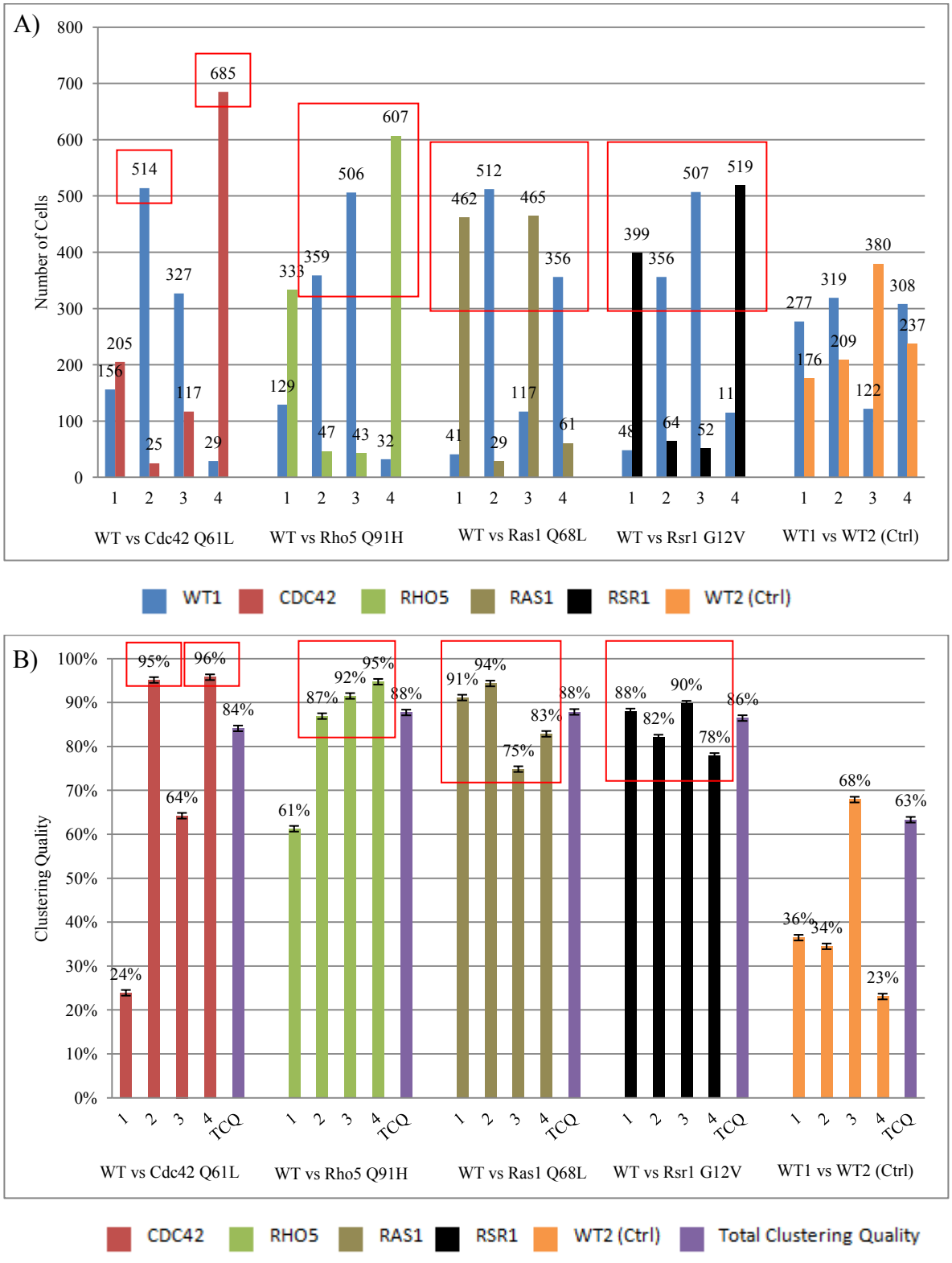

A)
Number of Cells
800
700
600
500
400
300
200
100
0
685
514
607
506
512
462
465
399
507
519
356
359
333
327
205
156
129
117
25
29
47
43
32
41
29
117
61
48
64
52
11
277
319
380
308
237
176
209
122
1 2 3 4
WT vs Cdc42 Q61L
WT vs Rho5 Q91H
WT vs Ras1 Q68L
WT vs Rsr1 G12V
WT1 vs WT2 (Ctrl)
WT1
CDC42
RHO5
RAS1
RSR1
WT2 (Ctrl)
B)
Clustering Quality
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
95%
96%
84%
64%
24%
87%
92%
95%
88%
61%
91%
94%
88%
83%
75%
88%
90%
86%
82%
78%
68%
63%
36%
34%
23%
1 2 3 4 TCQ
WT vs Cdc42 Q61L
WT vs Rho5 Q91H
WT vs Ras1 Q68L
WT vs Rsr1 G12V
WT1 vs WT2 (Ctrl)
CDC42
RHO5
RAS1
RSR1
WT2 (Ctrl)
Total Clustering Quality

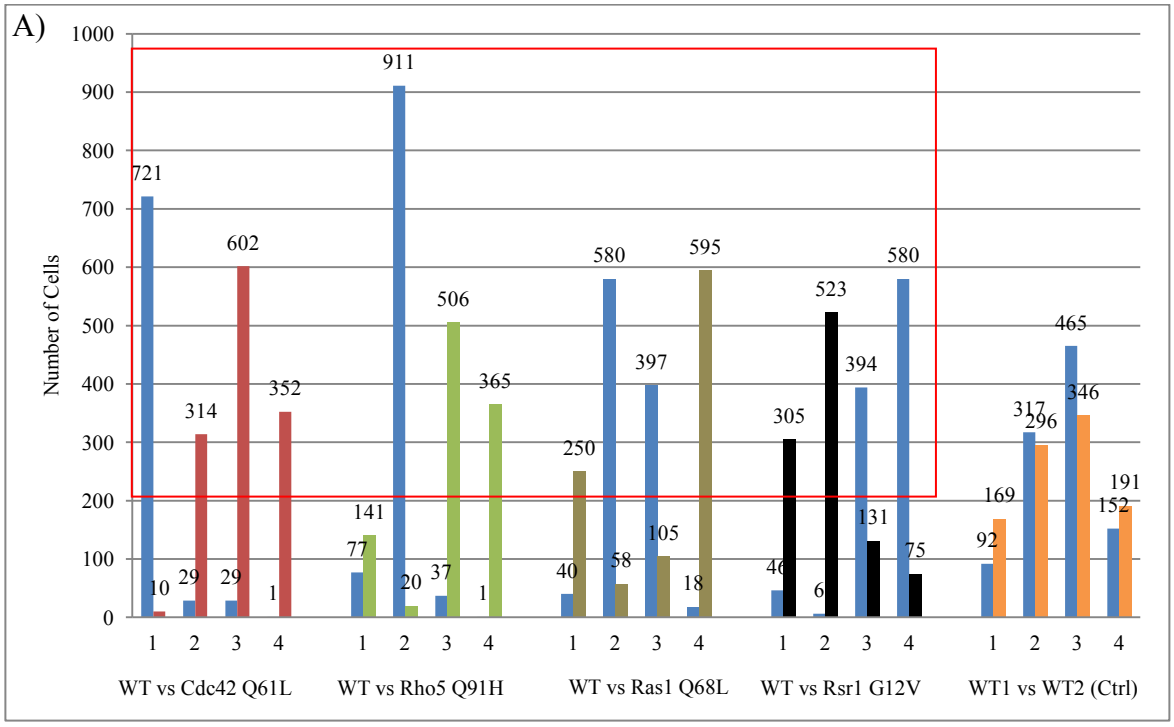
Figure 25: *K*-Means clustering results represented in histograms A) in all four mutants the clustering performed by *K*-Means gave a better separation. Cdc42 show two well defined clusters and Rho5 three well defined clusters, whereas for Ras1 and Rsr1 the separation is efficient in all four clusters. The number of clusters *K* for *K*-Means were previously calculated and giving *K*= 4 as the best number of clusters, represented by 1, 2, 3 and 4 on the x axis. The significance for all clusters corresponding to Cdc42, Rho5, Ras1 and Rsr1 are calculated by clustering quality measure (CQ_m) shown in percentages in chart B). Total clustering quality (TCQ) for each mutant was calculated to obtain the average TCQ of the four clusters. Significant clusters are framed in red. Wild type 1 and wild type 2 serve as a control where *K*-Means algorithm cannot group due to homogeneity of the population.

3.3.3.3. Fuzzy *C*-Means clustering algorithm

Since hierarchical and *K*-Means algorithms are classified as hard clustering, a more permissive clustering method is required. For this purpose, Fuzzy *C*-Means clustering algorithm was chosen to group the shape factor data.

It is logical to consider that asynchronous cells populations display many types of phenotypes and that wild type and mutant strains can share similar sub-phenotypes. This means that a wild type population can contain "mutant-like" cells and a mutant population can feature "wild type-like" cells. To test this idea, Fuzzy *C*-Means clustering provides a flexible way to cluster phenotypes, where one cell can belong to more than one cluster (Section 1.8.3.). Using this algorithm, each shape factor measure including wild type and mutant were assigned a membership degree. Using degree of membership, whether wild type or mutant, these cells can belong to more than one cluster which provides a nice and natural distribution of these features. With Fuzzy *C*-Means, can be observe that cells naturally segregate to their respective groups (Figure 26 A).

Fuzzy *C*-Means CQ_m results were greater than 75% and even higher than 80% which is strongly significant for phenotypes classification using such method (Figure 26 B). These results suggest that Fuzzy *C*-Means is the best clustering method for data composed of shape factor measures of cellular phenotypes.


A)
Number of Cells
1000
900
800
700
600
500
400
300
200
100
0
721
10
29
314
29
602
1
352
77
141
911
20
37
506
1
365
40
250
580
58
397
105
595
18
46
305
6
523
394
131
580
75
92
169
317
296
465
346
152
191
1 2 3 4
WT vs Cdc42 Q61L
WT vs Rho5 Q91H
WT vs Ras1 Q68L
WT vs Rsr1 G12V
WT1 vs WT2 (Ctrl)
WT1
CDC42
RHO5
RAS1
RSR1
WT2 (Ctrl)

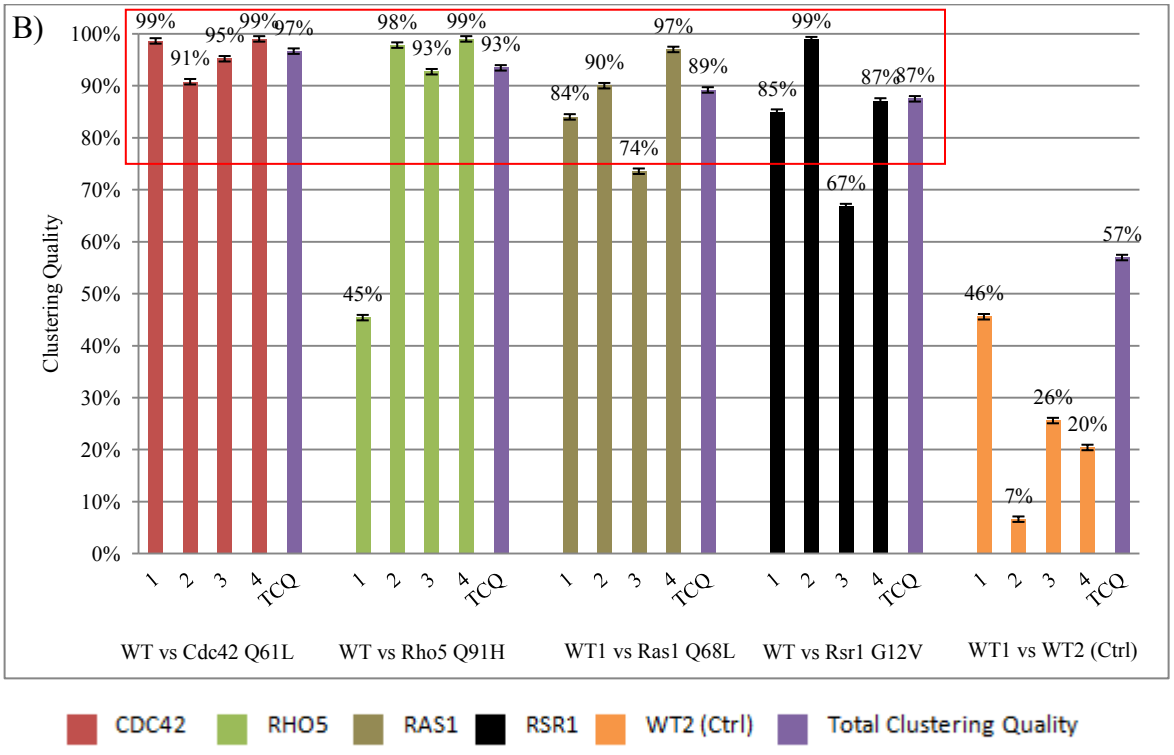

B)
Clustering Quality
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
99%
91%
95%
99%
97%
45%
98%
93%
99%
93%
84%
90%
74%
97%
89%
85%
99%
67%
87%
87%
46%
7%
26%
20%
57%
1 2 3 4 TCQ
WT vs Cdc42 Q61L
WT vs Rho5 Q91H
WT1 vs Ras1 Q68L
WT vs Rsr1 G12V
WT1 vs WT2 (Ctrl)
CDC42
RHO5
RAS1
RSR1
WT2 (Ctrl)
Total Clustering Quality

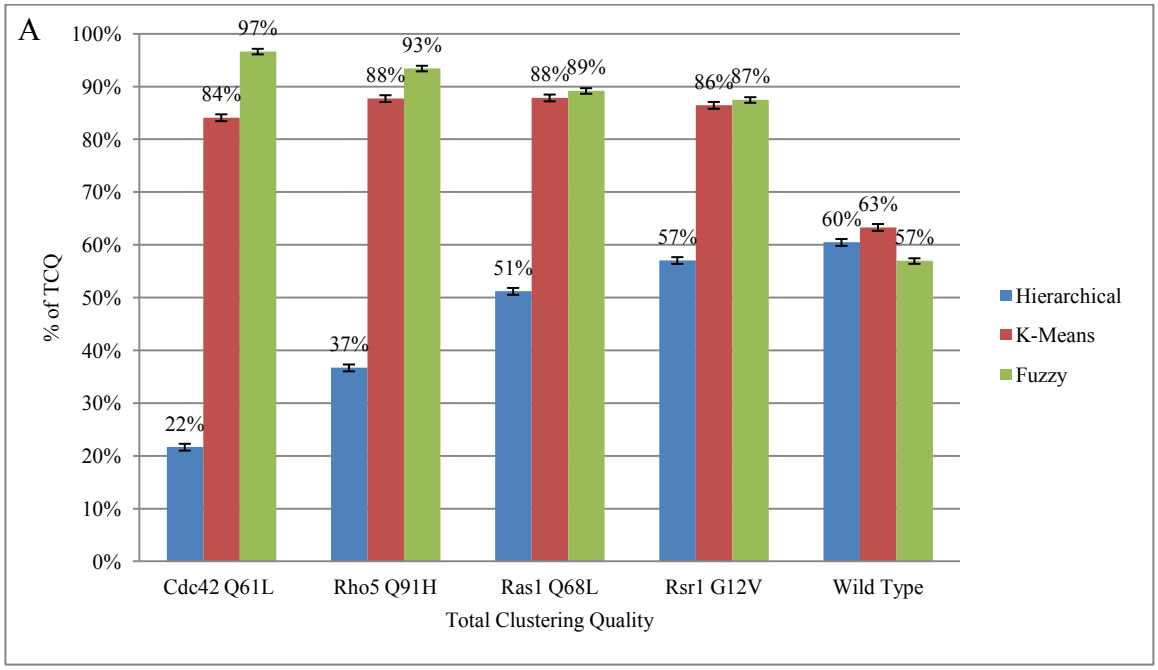
Figure 26: Identification of clusters of phenotypes using Fuzzy *C*-Means algorithm. A) Cdc42, Rho5, Ras1 and Rsr1 resulted in four well separated clusters showing the efficiency of Fuzzy *C*-Means in grouping shape factor measures obtained from distinct cellular phenotypes. As well, to test the efficiency of Fuzzy *C*-Means the clustering quality measure (CQ_m) were calculated B) showing > 75% of clustering quality in all clusters with exception of cluster 1 in Rho5 and cluster 3 in Rsr1. Fuzzy *C*-Means parameters were $c = 4$ (number of clusters) and $m = 1.25$ (degree of fuzzification). Number if clusters c is represented by 1, 2, 3, 4 on the x axis with the total clustering quality (TCQ), which was calculated for each mutant to obtain the average TCQ of the four clusters. Significant clusters are framed in red. Wild type 1 and wild type 2 serve as a control where Fuzzy *C*-Means algorithm cannot group due to homogeneity of the population. This homogeneity is shown by close number of cells in all four clusters.

3.4. Fuzzy *C*-Means outperforms Hierarchical and *K*-Means clustering and demonstrate the unique value of mutant phenotype

For each analysis (mutants Cdc42 Q61L, Rho5 Q91H, Ras1 Q65L and Rsr1 G12V respectively) CQ_m were calculated in order to check the average quality for each clustering method. Cdc42 Q61L, Rho5 Q91H, Ras1 Q65L and Rsr1 G12V show high percentage of CQ_m for Fuzzy *C*-Means and suggest that there is a clear discrimination between wild type and mutants cells. For the negative control analysis, comparison of wild type 1 versus wild type 2 cells, all three clustering methods gave similar CQ_m % which was expected since all wild type population display the same phenotype and are more homogeneous (Figure 27 A). Total average CQ_m was calculated for each clustering method (Figure 27 B) and allows discriminate the mutant strain.

After demonstrating with a CQ_m value of 85% that Fuzzy *C*-Means clustering outperforms Hierarchical clustering and *K*-Means clustering, we asked how significant and unique are these mutant phenotypes on the cellular population and what is the percentage of impact in it?. To answer such question, we also calculated the unique value for each Cdc42 Q61L, Rho5 Q91H, Ras1 Q65L and Rsr1 G12V and demonstrated the percentage of each mutant expressing unique phenotype distinct from wild type and the impact they had on cell population. We found that Cdc42 Q61L had 92% of unique mutant-like phenotypes and 8% of wild type-like, Rho5 Q91H had 82% unique mutant-like phenotype and 16% wild type-like phenotype, Ras1 Q65L had 80% mutant-like phenotype with 20% wild type-like phenotype and Rsr1 G12V gave 61% mutant-like phenotype with 31% wild type –like phenotype (Figure 28). These results suggested that Rho-like small GTPases Cdc42 Q61L and Rho5 Q91H had a major impact on the population since these two proteins are mainly involved in regulating cellular morphology processes like cell polarity and cell wall assembly. On the other hand, Ras-like small GTPases Ras1 Q65L and Rsr1 G12V had less

impact than Rho-like small GTPases Cdc42 and Rho5. This could be due to the fact that Ras1 and Rsr1 have more impact in other cellular responses than cellular morphogenesis.


A
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
% of TCQ
22%
84%
97%
37%
88%
93%
51%
88%
89%
57%
86%
87%
60%
63%
57%
Cdc42 Q61L
Rho5 Q91H
Ras1 Q68L
Rsr1 G12V
Wild Type
Total Clustering Quality
Hierarchical
K-Means
Fuzzy

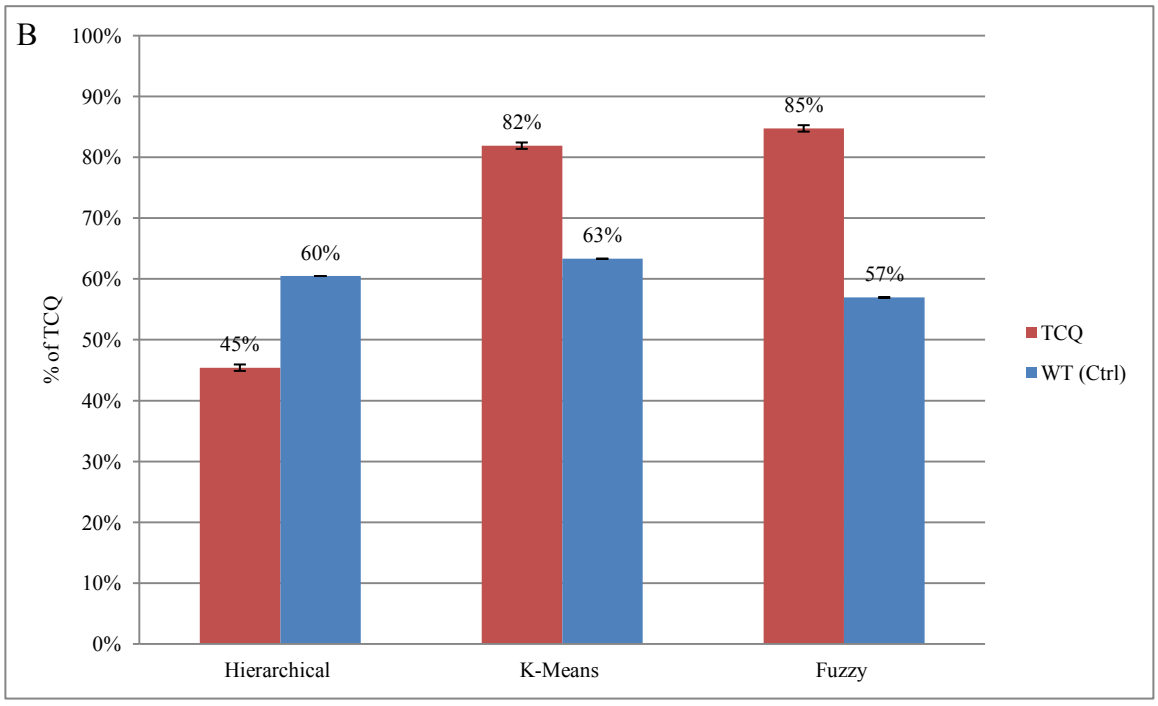

B
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
% of TCQ
45%
60%
82%
63%
85%
57%
Hierarchical
K-Means
Fuzzy
TCQ
WT (Ctrl)

Figure 27: A) Average CQ_m for each mutant population were calculated for all three clustering methods. Fuzzy *C*-Means outperforms Hierarchical and *K*-Means clustering. B) Total average CQ_m applied to each clustering method. Fuzzy *C*-Means outperforms the other two methods with an overall CQ_m = 85%. Wild type maintains a CQ_m value within 57%, 60% and 63%. The clustering methods show small variation whereas wild type does not show variation in the data.

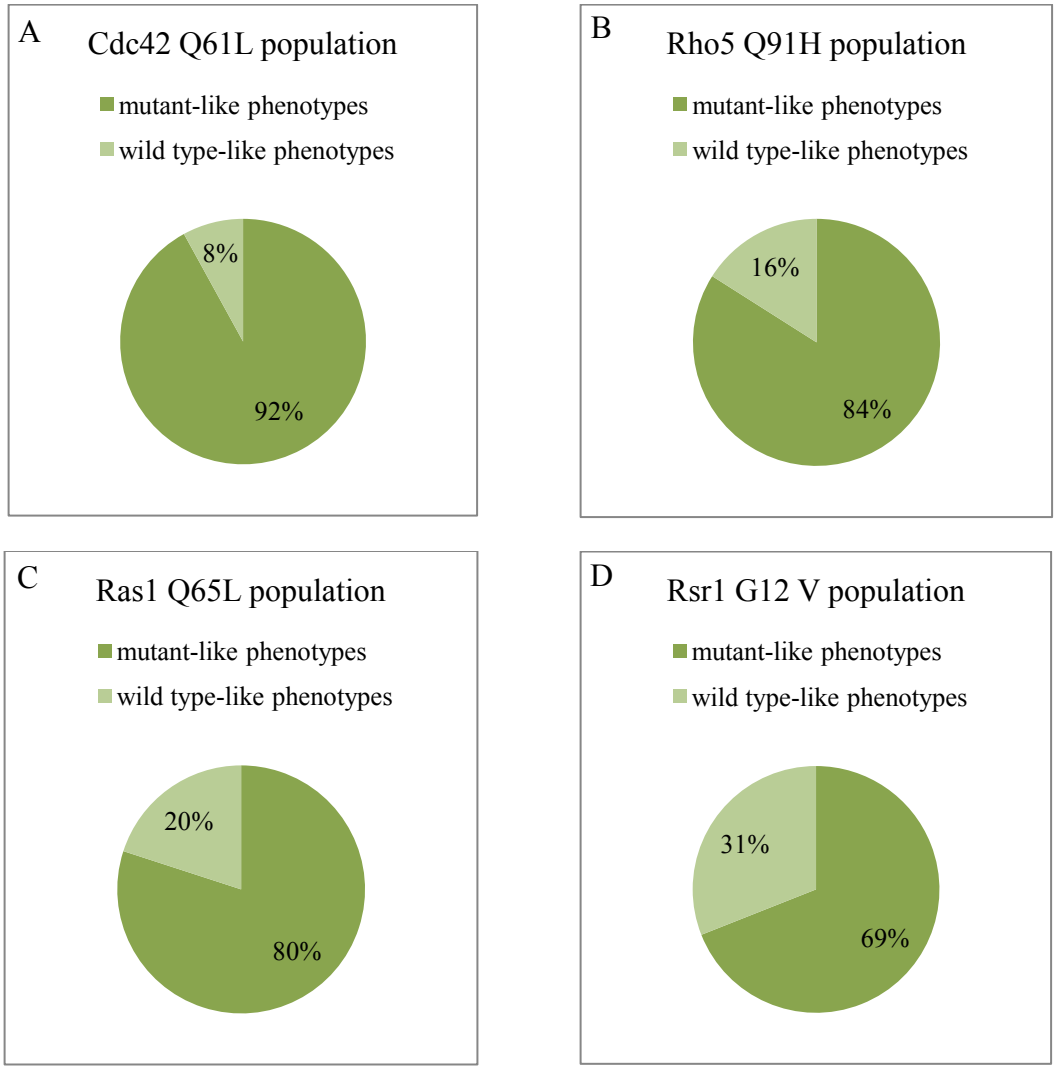


Figure 28: Percentage of unique cellular phenotypes. Cdc42 Q61L (A) amorphous phenotype, Rho5 Q91H (B) elongated/clumped phenotype, Ras1 Q68L (C) rounded phenotype and Rsr1 G12V (D) large phenotype are unique in comparison with wild type population after morphological analysis.

3.5. Data randomization

Clustering results were randomized in order to observe whether there is variation and to calculate the mean and standard deviation for each mutant where the CQ_m was calculated (Table 7). This was performed to observe whether the clustering is not happening by chance. The results show a reduction in bias, which indicates that this is not the case.

Table 7: Randomization results from each clustering method with their respective mean and standard deviation

		Cdc42	**Rho5**	**Ras1**	**Rsr1**	**WildType**	**Mean CQ_m**
Hierarchical	**CQ_m**	97.75%	93.65%	89.62%	87.00%	57.00%	85.00%
	Mean	51.49%	51.52%	51.55%	51.50%	100%	61.21%
	Std.Dev	0.0063	0.0065	0.0068	0.0066	0	0.0053
***K*-Means**	**CQ_m**	84.11%	87.48%	87.48%	86.46%	63.00%	81.71%
	Mean	51.5%	51.53%	51.50%	51.50%	100%	61.21%
	Std.Dev	0.0064	0.0064	0.0065	0.0065	0	0.0051
Fuzzy *C*-Means	**CQ_m**	22.00%	37.00%	51.00%	57.00%	60.00%	45.40%
	Mean	51.52%	51.46%	51.54%	51.48%	100%	61.20
	Std.Dev	0.0066	0.0066	0.0063	0.0062	0	0.0051

3.6. Divergent mutations do not result in pronounced phenotypic changes

After establishing a large scale quantification method, next we tested the hypothesis of switching functions by switching interaction partners (Section 1.9.). Single positions were switched between Cdc42 and Rho5. The initial test did not render any pronounced changes in switching phenotypes, however, when positions S41A, K123P and K186L for Cdc42 and A41S, P164K and L186K for Rho5, were exchanged, disruption in cell shape were observed (Figure 29).

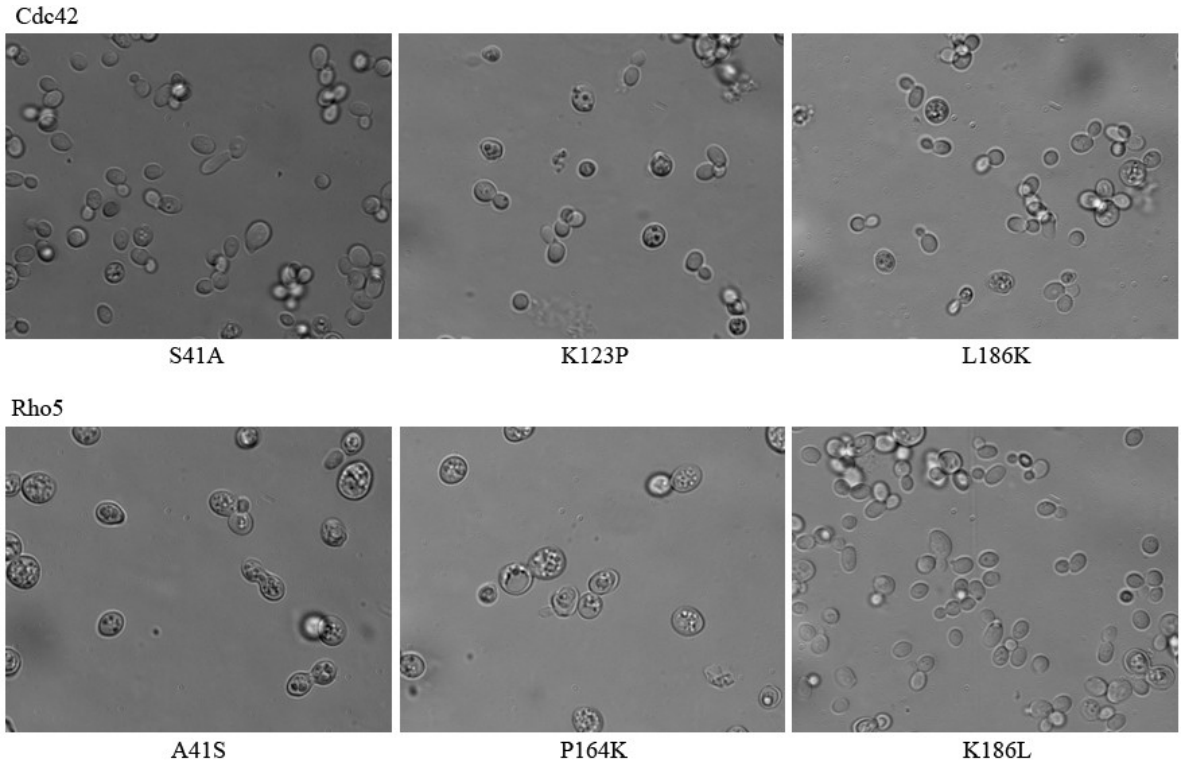


Figure 29: DIC images of Cdc42p and Rho5p switched-mutants. Point mutations of switched positions affect only cell morphology partially.

4. Discussion and conclusion

4.1. Small GTPases and their influence in cell morphology

In budding yeast *Saccharomyces cerevisiae*, Ras-like small GTPases plays a major role in regulating signal transduction pathways, which involves at least five different MAPKs. These signals have output responses like cell wall construction, morphological switch and mating response, which are mainly involved in cell morphology [119]. It is known that disruptions of such responses are caused by point mutations in G1/G3 regulatory domains [88]. Previous work done by Heo *et al.*, showed that over-expression of constitutively active mutants of these protein were capable of disrupting cell morphology, thus enabling a clear phenotypic classification in mammalian systems [2]. This classification directed the hypothesis of the current study that replacing single divergent positions between two homologous proteins should result in phenotypic and protein interaction network switch thus a similar approach was taken in *Saccharomyces cerevisiae*. However to arrive to the point of testing this hypothesis, first we needed to establish a method to classify small GTPases using the budding yeast *Saccharomyces cerevisiae* as model system.

To study how the Ras-like small GTPases can affect budding yeast cell morphology, constitutively active mutants were created and over-expressed in comparison with native forms. Native forms of these proteins were over-expressed (Figure 16) and was found that over-expression of native small GTPases did not induce any difference in phenotype, leading to hypothesize that over-expression of these proteins, may be rapidly controlled by their regulators, thereby, preventing abnormalities in cell morphogenesis. However, one can think that the cellular population containing over-expressed version of these native forms of small GTPases are not synchronize during the cell cycle and can be further explore by identifying subpopulations that might express different cell phenotypes at different stages of the cell cycle. For instance, doing FACS experiments with several different markers for morphology identification and native small GTPases abundance, whether will be in G1, S, G2 or M stages of the cell cycle [120].

This observation led to construct small GTPases constitutive active mutants. The GDP/GTP cycle of small GTPases are in charge of managing these proteins by switching them on and off depending on extracellular stimuli [26]. This cycle can be locked by replacing glycine (G) per valine (V) on G1 domain and glutamine (Q) per leucine (L) or histidine (H) on G3 domain of the small GTPases, hence, conferring a constitutive activity to these proteins called constitutively active mutants. Here, we constructed the phenotypic diversity within Ras-like small GTPases by introducing point mutations in the G1 and G3 domain of their sequences respectively. After over-expressing these constitutively active forms, was found that they were capable of inducing various types of phenotypes from the wild type cells, and conferred to the cells a complete change of shape (Figure 19). This change can not only be due to the constitutive activity, but also possibly by the disruption or change in the protein-protein interaction networks, which are in charge of orchestrating the appropriate cell morphology regulation components.

Cdc42 and Rho5 were chosen as candidates small GTPases. These two proteins belong to Rho-like small GTPases subfamily in which both possess different roles and are involved in different sub network of interaction partners. This is a key point in order to further explore functional switching of interaction partners. Cdc42 is known to be the master regulator to polarize growth process in mammals and *S. cerevisiae* [27]. Cdc42 is necessary for promoting and regulating cell polarization, and in order to obtain a normal cell shape Cdc42 must be restricted temporally until activation is needed. Similar to other enzymes, Cdc42 function is regulated and requires the ability to interact with important effectors to hydrolyse GTP [62]. Disruption of these processes results in amorphous/elongated phenotype as consequence of Cdc42 Q61L mutant (Figure 19 A). Cdc42 Q61L is able to polarize in G1-arrested cells thus inducing such changes in the morphology of the cell [121]. We can also think that mutant Cdc42 Q61L is able to cluster on the plasma membrane at random points in more than one site, thus, resulting in such phenotype too. A similar behaviour was observed by Rho5 Q91H. Rho5 is a not an essential small GTPase, however, it is involved in several cell responses that involve cell

wall integrity linked to protein kinase C (Pkc1p) dependent signal transduction pathway [48]. Over-expression of the constitutively active mutant Rho5 Q91H resulted in clumped/elongated phenotypes (Figure 19 B). This phenotype might be caused by cell wall integrity disruption. Rho5 constitutive active mutant not only displays clumped/elongated phenotypes, but also cell cycle arrest-like phenotypes. This result could be due to Rho5 known interaction with Sic1, which is an inhibitor of the Cdc28 kinase complex that regulates the G1/S phase transition by preventing premature S phase and ensuring genomic integrity [122].

Ras1 and Rsr1 were the two other homologs proteins chosen for experimental effort. They belong to the Ras-like small GTPases subfamily in which both have different tasks and are involved as well in different sub network of interaction partners. Ras1 is related to the oncogene Ras found in mammalian systems [123] and is known to play a role in the adenylate cycline activating pathway (cAMP), which is involved in controlling cell proliferation [38]. Our results shows that constitutively active Ras1 Q65L affected morphology of the cell by expressing enlarged phenotypes with cell cycle arrested cells (Figure 19 C). Ras1 interacts with Bcy1 [124], a regulatory subunit of the Protein kinase A (PKA) complex that regulates cell cycle and disruption of Ras1 processes also results in disruption of cell morphology. This might be due to the disruption of the interaction between Ras1 and Bcy1.

Rsr1 G12V mutant was mutated in the G1 domain since the G3 domain shows no conservation with others small GTPases (Figure 17). Rsr1 is required for bud site selection and it is involved in morphological responses by interacting with itself and the Cdc42 during bud-site selection and cell polarity events. Mutation in Rsr1 G12V resulted in a discrete large phenotype. One can think that this morphological observation could be a result of Rsr1 disrupted interaction with Cdc42 [125], which together regulate bud site and polarity processes. Indeed, over-expression of the Rsr1 G12V mutant binds constitutively to Cdc42 and Cdc24 proteins which are both required for establishment and regulation of cell polarity [126].

It is possible that these events are affecting the cell shape behaviour as a result of change in protein-protein interaction sub-network in charge of regulating cell morphogenesis. These hypotheses could be further investigated by designing protein-protein interaction experiments in presence of Cdc42 Q61L, Rho5 Q91H, Ras1 Q65L and Rsr1 G12V constitutive active mutants. Such experiment can be assayed at different temperatures and conditions using protein-fragment complementation assay (PCA). We can further explore if mutant version of these small GTPases can arrive to scramble whole dynamics of protein-protein interaction networks involved in morphological responses.

4.2. Use of the shape factor formulas along with clustering methods

Morphogenesis is a key feature of every multicellular organism and quantification of this could play a pivotal role to study how external insults can impact an entire population of cells. One main criticism to the work of Heo *et al.* (2003) [2] is that they did not show the degree of impact of constitutively active mutations on the entire cell population, and they did not present a clear method of automated cell quantification. Hence, important points were induced to develop this approach. Here, in this study we addressed these issues by creating a method to measure phenotypes using DIC imaging, followed by a method to establish quantitative and statistically significant phenotypic changes in a cell population, and finally developing a method to discriminate between discrete phenotypes among wild type and mutants.

While performing microscopy of the mutants using the protocol by Saito *et al.* (2005) [107], some issues with the fluorescence was encountered, mainly photobleaching. In order to simplify the procedure and reduce the time necessary for using fluorescence procedure, differential interference contrast (DIC) imaging was decided to use. DIC imaging has several advantages, to mention some, the procedure to take pictures is fast and

straight forward and moreover, DIC images allows me to observe and extract the real morphology of the cells. The only issue faced with DIC imaging was the automated quantification which required being address by microscopists as an alternative technique to whole fluorescence techniques.

To measure cell phenotypes, shape factors were used whose parameters are provided by Metamorph® image analysis software. These formulas are composed of two parameters that enable a more punctual view regarding classification of phenotypes. Therefore, after DIC pictures were obtained and measured, data files were exported as Excel files; however, it was challenging to work with large amount of data (7293 measures per 28 columns including shape factor calculations).

As a preliminary approach, the data was quantified and we calculated the shape factors of 200 cells (wild type and mutant) manually. A shift in populations between wild type and mutant was observed (Figure 20), but this amount of cells was not statistically significant and this approach was not optimal. Therefore, next was used simultaneously all the shape factors to see whether this alone will help discrimination of phenotypes by looking at the shape factor measures. However, no discrimination was possible so a macro was developed using *Visual Basic Editor in Excel* (VBE). This macro was programmed based on unique preliminary thresholds to automatically select, organize and quantify shape factor measures. This approach did not yield discrimination among phenotypes because the macro was programmed on manual thresholds and this approach was not suited enough for our purposes. In addition, the macro was not powerful enough to quantify and classify the amount of data.

Based on all these short comings to quantify and discriminate cell phenotypes, another approach was taken. This one was based on bioinformatics and statistical tools, which were more suited for such amount of data. Three different clustering algorithms were applied and each of them were carefully compared and analysed. The starting point of the whole bioinformatics approach was to build a hierarchy of cell phenotypes by looking at

their shapes and measures which were accomplished by using shape factor formulas. We hypothesize that clustering algorithms can partition mutant cells from wild type cells based on cell shape factor measurements. To prove this hypothesis, we assayed three different clustering methods in order to compare which can be more suitable to analyze with this kind of data. Foremost, hierarchical clustering was tested giving only two significant clusters for Ras1 Q65L and Rsr1 G12V with CQ_m of 82% and 85% respectively, and the rest of the cells expressing proteins were less clusterized. Unquestionably hierarchical clustering was not efficient to classify our data. The intrinsic nature of hierarchical clustering is to partition the data based on one common ancestor for each feature and our results suggest the absence of such common ancestor in our shape factor measurements for mutant and wild type strains. In addition hierarchical clustering is based in similarities between objects taking into account the concept of the common ancestor and work perfectly in phylogeny and evolution studies.

Next we tried clustering the phenotypes by using *K*-Means algorithm, in which we have the choice of the number of clusters (*K*). We calculated the number of clusters (*K*) and *K*= 2, 4, 6 and 8 were assayed. This led us to choose *K*=4 since above this value empty clusters were generated. These empty clusters were no primary shape factor assigned to them Hence, *K*-Means with *K*=4 were applied, which resulted in thirteen significant clusters that include all Cdc42 Q61L, Rho5 Q91H, Ras1 Q65L and Rsr1 G12V. The phenotype classification was improved but some arguments were still missing. Such arguments would be based on the separation and merging ability of the *K*-Means clustering which there is a lack of fuzziness to analyse biological data. The hierarchical and *K*-Means clustering are considered as hard partition clustering, meaning that with these methods one cell phenotype belongs only to one cluster.

Nevertheless, in cell morphology it is observed that in a population of wild type cells there will always be a small percentage of cells that will be “mutant-like” and a mutant cells population will always contain a small percentage of “wild type-like” cells. Therefore, this logical thinking led to implement a more permissive clustering algorithm

knows as Fuzzy *C*-Means (soft clustering). Like in fuzzy logic, in this method one cell can belong to more than one cluster while satisfying previous conditions. After calculating the data with Fuzzy *C*-Means, all clusters were showing statistical significance of more than 95% which is better than results obtained with Hierarchical clustering and *K*-means clustering. With Fuzzy *C*-Means, clusters within wild type and mutants were far more defined and significant. Fuzzy *C*-Means clustering improved the partition of the data because it is able to differentiate different measures of cell phenotypes based on their features. Fuzzy *C*-Means features an outstanding degree of fuzziness. We can conclude that Fuzzy *C*-Means clustering is among the best algorithm used to analyze biological data.

To validate such clustering results, clustering quality measure (CQ_m) was calculated, which is a statistical measure [101]. This essential formula led to choose the best clustering method which was Fuzzy *C*-Means, with which we obtained an overall significant CQ_m of 85% outperforming Hierarchical with CQ_m 45% and *K*-Means with CQ_m 82% (Figure 27, B). Interestingly, in the wild type (WT1/WT2) control the CQ_m of 60% was calculated for Hierarchical, CQ_m 63% for *K*-Means and CQ_m 57% for Fuzzy *C*-Means, which demonstrate the accuracy of these methods by not being able to discriminate a homogeneous population. These observations and results helped to establish a method that can discriminate cell phenotypes and at the same time show statistical significance of the constitutive active mutations in a large population of cells.

Looking at the composition of asynchronous cellular population, we wanted to know at which level these point mutations were able to impact having a quantitative significance on the population. We calculate the unique value for each mutant strain Cdc42 Q61L, Rho5 Q91H, Ras1 Q65L and Rsr1 G12V and demonstrated the percentage of each mutant expressing unique phenotype distinct from wild type and the impact on cell population. Cdc42 Q61L had 92% of unique mutant-like phenotypes with an 8% of wild type-like phenotypes, Rho5 Q91H had 82% unique mutant-like phenotype and 16% wild type-like phenotype, Ras1 Q65L had 80% mutant-like phenotype with 20% wild type-like phenotype and Rsr1 G12V gave 61% mutant-like phenotype with 31% wild type –like

phenotype (Figure 28). After obtaining these results, with support from the literature Rho-like small GTPases Cdc42 Q61L and Rho5 Q91H show major changes on the population. Being expected since Rho-like small GTPases are mainly in charge of regulating major morphological processes, whereas Ras-like small GTPases Ras1 Q65L and Rsr1 G12V expressed phenotypes which had less change than those expressing Cdc42 Q61L and Rho5 Q91H. This may be because, Ras1 and Rsr1 are more involved in other cellular responses than being specific to cellular morphogenesis responses.

Moreover looking at the composition of an asynchronous cellular population, for this study we choose the *BG1805* 2 micron based vector. This could generate some phenotypic variability in cells expressing each mutant since they might have a tendency to not keep the same copy number per cell. There is a point that various vector copies could induce various expression levels which affect phenotypes differently inside a given population. This variability may be reflected in our results through heterogeneous phenotypes in mutant populations; however, we can also observe low percentage of phenotype heterogeneity in wild type populations. Also, a homogeneous mutant population does not reflect the natural phenotypic distribution of cellular populations. Therefore, we think that our approach can deal with heterogeneous population phenotypes while confirming the robustness of the algorithm.

In our study, we think that the distribution of vector copy numbers per cell is similar between all the strains, given the same budding yeast background, vector backbone and relatively similar length of small GTPases gene. Regarding stability of the vector, a future approach would be to test vector stability by insertion and deletions of different length of DNA fragments. This could lead to compare different strains containing small GTPases and distinct vector copy numbers per cell.

4.3. Evolution of cell morphology

The phenotypic diversity of *Saccharomyces cerevisiae* that we observe today arose as evolutionary systems continually sampled new phenotypes that resisted ever changing selective pressures. This phenotypic diversification is driven by variations in the regulatory network that dictate cells to form multicellular patterns and structures. Such structures are mainly induced by variations in their protein sequences known as point mutations. The aim of this study was to establish a path that can help to understand what underlies evolution of cell morphology and beyond. One main question was how cells started to acquire different phenotypes and if the different models of gene duplication [4] have a potential to impact cell morphology. It is already known that changes in developmental genes makes cell morphology evolve and little is understood on how single-nucleotide substitution are able to affect morphology [89]. The gene duplication phenomenon is caused by events such as neo/sub-functionalization, which also leads one to think that they will be strongly involved in cell morphology.

The overall aim of this research was to switch phenotypes between two related proteins; however, due to lack of time, that goal was not completely achieved. We hypothesize that new functions or developmental small GTPase pathways might evolve by switch of function through gain or loss of their interaction partners. In a given set of related proteins, the observed distinct functions have emerge from a common ancestor and have subsequently been mutated to gain new interactions and new functions. We are interested to investigate how many mutations are necessary to exchange functions between two Ras-like small GTPases and what are the effects of these mutations. Such phenomena might be explained by the mutation of domains in duplicated genes and the number of such mutation that could govern specific functions (sub-functionalization), the creation of a new domain to achieve new functions (neo-functionalization) and the destruction of a domain that ends up with a loss of function; The previous scenarios might result in a modification in the activity and the interaction network involving the Ras-like small GTPases. However the evolutionary mechanisms that involve gene duplication and protein trade-off remain

unclear. To prove this hypothesis, we establish the level that a mutation affects the whole cell population and how we to statistically quantify these events. Cell morphologies have been affected by single point mutations exchanged between Cdc42 and Rho5 (Figure 29). These single substitutions were performed on divergent positions leading us to think that development of different phenotypic traits of cell shapes might involve amino acid substitutions during evolution. As well, such substitutions not only could affect morphology but protein-protein interaction networks underlying it.

These observations are of outstanding interest for evolutionary biology, as they could be the basis on which one can begin to comprehend how cellular functions differ from each other and how single but related proteins are involved and capable of regulating different signalling pathways. In addition, this could be extended to understand how proteins evolved to contribute to different morphologies that are related to the function of the cell and accomplish their function accurately.

4.4. Concluding remarks

We show that significant phenotypic changes induced by point mutations can be quantitative establish by using bioinformatics tools that enable us to process large amount of data. Moreover, such data were obtained from microscopy measures using shape factor formulas that explicitly describe the phenotype of each mutant.

Furthermore, by comparing and quantify the small GTPases phenotypes we systematically deduce the quantitative impact that these small GTPases have on the cellular population. These results are opening doors to further explore at which level these small GTPases are more significant to orchestrate cellular morphology and which ones have direct impact and which ones have indirect impact on cell morphology.

4.5. Perspectives and future approach

The methods established in this study can help to discriminate between cellular phenotypes. At the moment this approach is under optimization in order to be more sensitive to discriminate mutant cells and simultaneously classify more than two mutants using only DIC images and shape factors. With these methods, it is now easier to quantify cell population taking in account the statistical significance and the use of only DIC images.

Moreover, the number of divergent positions to create switches and the combination of them will be increased in the Ras-like small GTPases. It is the current thinking that identification of divergent positions which implies evolution of cell morphology will provide a framework for engineering novel protein-protein interactions and functions. These aims might lead to the identification and control of morphological changes caused by external insults. In addition, it will help to validate the gain or loss of functions that are linked to the gain or loss of interaction partners under specific cellular responses to understand how networks have evolved to result in a particular phenotype.

This work provides an insight into how one should design and calculate the experimental strategies to visualize functional protein evolution, as well as monitor protein-protein interactions and functions at a network scale. With this, we can further explore the aspects of adaptive dynamics of cellular morphologic pathways and study protein-protein interaction networks changes in present of insults like point mutations and observe if the fitness of such networks were depending on point mutation to evolve during time to result in a particular phenotype.

5. Bibliography

1. Wong-Staal, F., et al., *Three distinct genes in human DNA related to the transforming genes of mammalian sarcoma retroviruses.* Science, 1981. **213**(4504): p. 226-8.

2. Heo, W.D. and T. Meyer, *Switch-of-function mutants based on morphology classification of Ras superfamily small GTPases.* Cell, 2003. **113**(3): p. 315-28.

3. Mora, C.F. and A.K.H. Kwan, *Sphericity, shape factor, and convexity measurement of coarse aggregate for concrete using digital image processing.* Cement and Concrete Research, 2000. **30**(3): p. 351-358.

4. Soskine, M. and D.S. Tawfik, *Mutational effects and the evolution of new protein functions.* Nature Reviews Genetics, 2010. **11**(8): p. 572-82.

5. Soskine, M. and D.S. Tawfik, *Mutational effects and the evolution of new protein functions.* Nature Reviews Genetics, 2010. **11**(8): p. 572-582.

6. Hahn, M.W., *Distinguishing among evolutionary models for the maintenance of gene duplicates.* J Hered, 2009. **100**(5): p. 605-17.

7. Brosius, J., *Retroposons--seeds of evolution.* Science, 1991. **251**(4995): p. 753.

8. Long, M., *Evolution of novel genes.* Curr Opin Genet Dev, 2001. **11**(6): p. 673-80.

9. Conant, G.C. and A. Wagner, *Asymmetric sequence divergence of duplicate genes.* Genome Res, 2003. **13**(9): p. 2052-8.

10. Conant, G.C. and K.H. Wolfe, *Turning a hobby into a job: how duplicated genes find new functions.* Nature Reviews Genetics, 2008. **9**(12): p. 938-50.

11. Zhang, J.Z., *Evolution by gene duplication: an update.* Trends in Ecology & Evolution, 2003. **18**(6): p. 292-298.

12. Kondrashov, F.A., et al., *Selection in the evolution of gene duplications.* Genome Biol, 2002. **3**(2): p. RESEARCH0008.

13. Singleton, A.B., et al., *alpha-Synuclein locus triplication causes Parkinson's disease.* Science, 2003. **302**(5646): p. 841.

14. Ohno, S., *Evolution by gene duplication*1970, Berlin, New York,: Springer-Verlag. xv, 160 p.

15. Force, A., et al., *Preservation of duplicate genes by complementary, degenerative mutations.* Genetics, 1999. **151**(4): p. 1531-45.

16. Casari, G., C. Sander, and A. Valencia, *A method to predict functional residues in proteins.* Nat Struct Biol, 1995. **2**(2): p. 171-8.

17. Innan, H. and F. Kondrashov, *The evolution of gene duplications: classifying and distinguishing between models.* Nature Reviews Genetics, 2010. **11**(2): p. 97-108.

18. Wuster, A., et al., *Spial: analysis of subtype-specific features in multiple sequence alignments of proteins.* Bioinformatics, 2010. **26**(22): p. 2906-7.

19. Botstein, D., S.A. Chervitz, and J.M. Cherry, *Yeast as a model organism.* Science, 1997. **277**(5330): p. 1259-60.

20. Botstein, D. and G.R. Fink, *Yeast: an experimental organism for modern biology.* Science, 1988. **240**(4858): p. 1439-43.

21. Gallwitz, D., C. Donath, and C. Sander, *A yeast gene encoding a protein homologous to the human c-has/bas proto-oncogene product.* Nature, 1983. **306**(5944): p. 704-7.

22. Giaever, G., et al., *Functional profiling of the Saccharomyces cerevisiae genome.* Nature, 2002. **418**(6896): p. 387-91.

23. Bourne, H.R., D.A. Sanders, and F. McCormick, *The GTPase superfamily: a conserved switch for diverse cell functions.* Nature, 1990. **348**(6297): p. 125-32.

24. Park, H.O. and E. Bi, *Central roles of small GTPases in the development of cell polarity in yeast and beyond.* Microbiol Mol Biol Rev, 2007. **71**(1): p. 48-96.

25. Campbell, S.L., et al., *Increasing complexity of Ras signaling.* Oncogene, 1998. **17**(11 Reviews): p. 1395-413.

26. Bar-Sagi, D. and A. Hall, *Ras and Rho GTPases: a family reunion.* Cell, 2000. **103**(2): p. 227-38.

27. Perez, P. and S.A. Rincon, *Rho GTPases: regulation of cell polarity and growth in yeasts.* Biochem J, 2010. **426**(3): p. 243-53.

28. Vigil, D., et al., *Ras superfamily GEFs and GAPs: validated and tractable targets for cancer therapy?* Nat Rev Cancer, 2010. **10**(12): p. 842-57.

29. Yeh, B.J., et al., *Rewiring cellular morphology pathways with synthetic guanine nucleotide exchange factors.* Nature, 2007. **447**(7144): p. 596-600.

30. Garcia-Ranea, J.A. and A. Valencia, *Distribution and functional diversification of the ras superfamily in Saccharomyces cerevisiae.* FEBS Lett, 1998. **434**(3): p. 219-25.

31. van Dam, T.J., J.L. Bos, and B. Snel, *Evolution of the Ras-like small GTPases and their regulators.* Small Gtpases, 2011. **2**(1): p. 4-16.

32. Vetter, I.R. and A. Wittinghofer, *The guanine nucleotide-binding switch in three dimensions.* Science, 2001. **294**(5545): p. 1299-304.

33. Madden, K. and M. Snyder, *Cell polarity and morphogenesis in budding yeast.* Annu Rev Microbiol, 1998. **52**: p. 687-744.

34. Alberghina, L., et al., *Cell growth and cell cycle in Saccharomyces cerevisiae: Basic regulatory design and protein-protein interaction network.* Biotechnol Adv, 2011.

35. Hall, A., *The cytoskeleton and cancer.* Cancer Metastasis Rev, 2009. **28**(1-2): p. 5-14.

36. Repasky, G.A., E.J. Chenette, and C.J. Der, *Renewing the conspiracy theory debate: does Raf function alone to mediate Ras oncogenesis?* Trends Cell Biol, 2004. **14**(11): p. 639-47.

37. Anand, B., S.K. Verma, and B. Prakash, *Structural stabilization of GTP-binding domains in circularly permuted GTPases: implications for RNA binding.* Nucleic Acids Res, 2006. **34**(8): p. 2196-205.

38. Fujiyama, A. and F. Tamanoi, *Processing and fatty acid acylation of RAS1 and RAS2 proteins in Saccharomyces cerevisiae.* Proc Natl Acad Sci U S A, 1986. **83**(5): p. 1266-70.

39. Chant, J. and I. Herskowitz, *Genetic control of bud site selection in yeast by a set of gene products that constitute a morphogenetic pathway.* Cell, 1991. **65**(7): p. 1203-12.

40. Urano, J., et al., *The Saccharomyces cerevisiae Rheb G-protein is involved in regulating canavanine resistance and arginine uptake.* J Biol Chem, 2000. **275**(15): p. 11198-206.

41. Downward, J., *Targeting RAS signalling pathways in cancer therapy.* Nat Rev Cancer, 2003. **3**(1): p. 11-22.

42. Jaffe, A.B. and A. Hall, *Rho GTPases: biochemistry and biology.* Annu Rev Cell Dev Biol, 2005. **21**: p. 247-69.

43. Etienne-Manneville, S. and A. Hall, *Rho GTPases in cell biology.* Nature, 2002. **420**(6916): p. 629-35.

44. Rojas, A.M., et al., *The Ras protein superfamily: evolutionary tree and role of conserved amino acids.* J Cell Biol, 2012. **196**(2): p. 189-201.

45. Bussey, H., *Cell shape determination: a pivotal role for Rho.* Science, 1996. **272**(5259): p. 224-5.

46. Walmsley, M.J., et al., *Critical roles for Rac1 and Rac2 GTPases in B cell development and signaling.* Science, 2003. **302**(5644): p. 459-62.

47. Tybulewicz, V.L.J. and R.B. Henderson, *Rho family GTPases and their regulators in lymphocytes.* Nature Reviews Immunology, 2009. **9**(9): p. 630-644.

48. Schmitz, H.P., et al., *Rho5p downregulates the yeast cell integrity pathway.* J Cell Sci, 2002. **115**(Pt 15): p. 3139-48.

49. Madaule, P., R. Axel, and A.M. Myers, *Characterization of two members of the rho gene family from the yeast Saccharomyces cerevisiae.* Proc Natl Acad Sci U S A, 1987. **84**(3): p. 779-83.

50. Bussey, H., *Rho returns: its targets in focal adhesions.* Science, 1996. **273**(5272): p. 203.

51. Ozaki, K., et al., *Rom1p and Rom2p are GDP/GTP exchange proteins (GEPs) for the Rho1p small GTP binding protein in Saccharomyces cerevisiae.* EMBO J, 1996. **15**(9): p. 2196-207.

52. Schmidt, A. and A. Hall, *Guanine nucleotide exchange factors for Rho GTPases: turning on the switch.* Genes Dev, 2002. **16**(13): p. 1587-609.

53. Martin, H., et al., *Regulatory mechanisms for modulation of signaling through the cell integrity Slt2-mediated pathway in Saccharomyces cerevisiae.* J Biol Chem, 2000. **275**(2): p. 1511-9.

54. Arellano, M., A. Duran, and P. Perez, *Localisation of the Schizosaccharomyces pombe rho1p GTPase and its involvement in the organisation of the actin cytoskeleton.* J Cell Sci, 1997. **110 (Pt 20)**: p. 2547-55.

55. Lee, M.E., et al., *The Rho1 GTPase Acts Together With a Vacuolar Glutathione S-conjugate Transporter to Protect Yeast Cells from Oxidative Stress.* Genetics, 2011.

56. Roumanie, O., et al., *Evidence for the genetic interaction between the actin-binding protein Vrp1 and the RhoGAP Rgd1 mediated through Rho3p and Rho4p in Saccharomyces cerevisiae.* Mol Microbiol, 2000. **36**(6): p. 1403-14.

57. Matsui, Y. and A. Toh-e, *Isolation and characterization of two novel ras superfamily genes in Saccharomyces cerevisiae.* Gene, 1992. **114**(1): p. 43-9.

58. Adamo, J.E., G. Rossi, and P. Brennwald, *The Rho GTPase Rho3 has a direct role in exocytosis that is distinct from its role in actin polarity.* Mol Biol Cell, 1999. **10**(12): p. 4121-33.

59. Singh, K., P.J. Kang, and H.O. Park, *The Rho5 GTPase is necessary for oxidant-induced cell death in budding yeast.* Proc Natl Acad Sci U S A, 2008. **105**(5): p. 1522-7.

60. Ramezani-Rad, M., *The role of adaptor protein Ste50-dependent regulation of the MAPKKK Ste11 in multiple signalling pathways of yeast.* Curr Genet, 2003. **43**(3): p. 161-70.

61. Annan, R.B., et al., *Rho5p is involved in mediating the osmotic stress response in Saccharomyces cerevisiae, and its activity is regulated via Msi1p and Npr1p by phosphorylation and ubiquitination.* Eukaryot Cell, 2008. **7**(9): p. 1441-9.

62. Etienne-Manneville, S., *Cdc42--the centre of polarity.* J Cell Sci, 2004. **117**(Pt 8): p. 1291-300.

63. Ziman, M., et al., *Subcellular localization of Cdc42p, a Saccharomyces cerevisiae GTP-binding protein involved in the control of cell polarity.* Mol Biol Cell, 1993. **4**(12): p. 1307-16.

64. Krugmann, S., et al., *Cdc42 induces filopodia by promoting the formation of an IRSp53:Mena complex.* Curr Biol, 2001. **11**(21): p. 1645-55.

65. Donaldson, J.G. and C.L. Jackson, *ARF family G proteins and their regulators: roles in membrane transport, development and disease.* Nat Rev Mol Cell Biol, 2011. **12**(6): p. 362-75.

66. Zerial, M. and H. McBride, *Rab proteins as membrane organizers.* Nat Rev Mol Cell Biol, 2001. **2**(2): p. 107-17.

67. Benli, M., et al., *Two GTPase isoforms, Ypt31p and Ypt32p, are essential for Golgi function in yeast.* EMBO J, 1996. **15**(23): p. 6460-75.

68. Botstein, D., et al., *Diverse biological functions of small GTP-binding proteins in yeast.* Cold Spring Harb Symp Quant Biol, 1988. **53 Pt 2**: p. 629-36.

69. Bialek-Wyrzykowska, U., et al., *Low levels of Ypt protein prenylation cause vesicle polarization defects and thermosensitive growth that can be suppressed by genes involved in cell wall maintenance.* Mol Microbiol, 2000. **35**(6): p. 1295-311.

70. Elkind, N.B., C. Walch-Solimena, and P.J. Novick, *The role of the COOH terminus of Sec2p in the transport of post-Golgi vesicles.* J Cell Biol, 2000. **149**(1): p. 95-110.

71. Li, B. and J.R. Warner, *Mutation of the Rab6 homologue of Saccharomyces cerevisiae, YPT6, inhibits both early Golgi function and ribosome biosynthesis.* J Biol Chem, 1996. **271**(28): p. 16813-9.

72. Haas, A., et al., *The GTPase Ypt7p of Saccharomyces cerevisiae is required on both partner vacuoles for the homotypic fusion step of vacuole inheritance.* EMBO J, 1995. **14**(21): p. 5258-70.

73. Louvet, O., et al., *Characterization of the ORF YBR264c in Saccharomyces cerevisiae, which encodes a new yeast Ypt that is degraded by a proteasome-dependent mechanism.* Mol Gen Genet, 1999. **261**(4-5): p. 589-600.

74. Singer-Kruger, B., et al., *Role of three rab5-like GTPases, Ypt51p, Ypt52p, and Ypt53p, in the endocytic and vacuolar protein sorting pathways of yeast.* J Cell Biol, 1994. **125**(2): p. 283-98.

75. Itoh, T., et al., *Complex formation with Ypt11p, a rab-type small GTPase, is essential to facilitate the function of Myo2p, a class V myosin, in mitochondrial distribution in Saccharomyces cerevisiae.* Mol Cell Biol, 2002. **22**(22): p. 7744-57.

76. Hoyt, M.A., T. Stearns, and D. Botstein, *Chromosome instability mutants of Saccharomyces cerevisiae that are defective in microtubule-mediated processes.* Mol Cell Biol, 1990. **10**(1): p. 223-34.

77. Hughes, H. and D.J. Stephens, *Assembly, organization, and function of the COPII coat.* Histochem Cell Biol, 2008. **129**(2): p. 129-51.

78. Behnia, R., et al., *Targeting of the Arf-like GTPase Arl3p to the Golgi requires N-terminal acetylation and the membrane protein Sys1p.* Nat Cell Biol, 2004. **6**(5): p. 405-13.

79. Munson, A.M., et al., *Yeast ARL1 encodes a regulator of K+ influx.* J Cell Sci, 2004. **117**(Pt 11): p. 2309-20.

80. Stearns, T., et al., *ADP-ribosylation factor is functionally and physically associated with the Golgi complex.* Proc Natl Acad Sci U S A, 1990. **87**(3): p. 1238-42.

81. Huang, C.F., et al., *Role for Arf3p in development of polarity, but not endocytosis, in Saccharomyces cerevisiae.* Mol Biol Cell, 2003. **14**(9): p. 3834-47.

82. Weis, K., *Regulating access to the genome: nucleocytoplasmic transport throughout the cell cycle.* Cell, 2003. **112**(4): p. 441-51.

83. Li, H.Y., K. Cao, and Y. Zheng, *Ran in the spindle checkpoint: a new function for a versatile GTPase.* Trends Cell Biol, 2003. **13**(11): p. 553-7.

84. Belhumeur, P., et al., *GSP1 and GSP2, genetic suppressors of the prp20-1 mutant in Saccharomyces cerevisiae: GTP-binding proteins involved in the maintenance of nuclear organization.* Mol Cell Biol, 1993. **13**(4): p. 2152-61.

85. Shields, J.M., et al., *Understanding Ras: 'it ain't over 'til it's over'.* Trends Cell Biol, 2000. **10**(4): p. 147-54.

86. Bos, J.L., *ras oncogenes in human cancer: a review.* Cancer Res, 1989. **49**(17): p. 4682-9.

87. Overmeyer, J.H. and W.A. Maltese, *Death pathways triggered by activated Ras in cancer cells.* Front Biosci, 2011. **16**: p. 1693-713.

88. Kaziro, Y., et al., *Structure and function of signal-transducing GTP-binding proteins.* Annu Rev Biochem, 1991. **60**: p. 349-400.

89. Frankel, N., et al., *Morphological evolution caused by many subtle-effect substitutions in regulatory DNA.* Nature, 2011. **474**(7353): p. 598-603.

90. Weinert, T.A., G.L. Kiser, and L.H. Hartwell, *Mitotic checkpoint genes in budding yeast and the dependence of mitosis on DNA replication and repair.* Genes Dev, 1994. **8**(6): p. 652-65.

91. Russ, J.C., *The image processing handbook*. 4th ed2002, Boca Raton, FL: CRC Press. 732 p.

92. Selinummi, J., et al., *Bright field microscopy as an alternative to whole cell fluorescence in automated analysis of macrophage images.* PLoS One, 2009. **4**(10): p. e7497.

93. Heise, B., A. Sonnleitner, and E.P. Klement, *DIC image reconstruction on large cell scans.* Microsc Res Tech, 2005. **66**(6): p. 312-20.

94. Johnson, S.C., *Hierarchical Clustering Schemes.* Psychometrika, 1967. **2**: p. 241-254.

95. Tan, P.-N., M. Steinbach, and V. Kumar, *Introduction to data mining*. 1st ed2006, Boston: Pearson Addison Wesley. xxi, 769 p.

96. MacQueen, J.B., *Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley, University of California Press, 1967. **1**: p. 281-297.

97. Crawley, M.J., *The R book*2007, Chichester, England ; Hoboken, N.J.: Wiley. viii, 942.

98. Dunn, J.C., *A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact WellSeparated Clusters.* Cybernetics and Systems, 1973. **3** (3): p. 32 — 57.

99. Bezdek, J.C., *Pattern recognition with fuzzy objective function algorithms.* Advanced Applications in Pattern Recognition.

100. Futschik, M.E. and B. Carlisle, *Noise-robust soft clustering of gene expression time-course data.* J Bioinform Comput Biol, 2005. **3**(4): p. 965-88.

101. Kelil, A., et al., *CLUSS: clustering of protein sequences based on a new similarity measure.* BMC Bioinformatics, 2007. **8**: p. 286.

102. Michnick, S.W., et al., *A toolkit of protein-fragment complementation assays for studying and dissecting large-scale and dynamic protein-protein interactions in living cells.* Methods Enzymol, 2010. **470**: p. 335-68.

103. Knop, M., et al., *Epitope tagging of yeast genes using a PCR-based strategy: more tags and improved practical routines.* Yeast, 1999. **15**(10B): p. 963-72.

104. Waterhouse, A.M., et al., *Jalview Version 2--a multiple sequence alignment editor and analysis workbench.* Bioinformatics, 2009. **25**(9): p. 1189-91.

105. Evans, P.M. and C. Liu, *SiteFind: a software tool for introducing a restriction site as a marker for successful site-directed mutagenesis.* BMC Mol Biol, 2005. **6**: p. 22.

106. Fenwick, C., et al., *A subclass of Ras proteins that regulate the degradation of IkappaB.* Science, 2000. **287**(5454): p. 869-73.

107. Saito, T.L., et al., *Data mining tools for the Saccharomyces cerevisiae morphological database.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W753-7.

108. Malleshaiah, M.K., et al., *The scaffold protein Ste5 directly controls a switch-like mating decision in yeast.* Nature, 2010. **465**(7294): p. 101-5.

109. Venables, W.N., D.M. Smith, and R Development Core Team., *An introduction to R : notes on R: a programming environment for data analysis and graphics, version 1.4.1*2002, Bristol: Network Theory. vi, 139 p.

110. Crawley, M.J., *The R book*2007, Chichester, England ; Hoboken, N.J.: Wiley. viii, 942 p.

111. Lance, G.N., and W.T. Williams, *A General Theory of Classifactory Sorting Strategies, I. Hierarchical Systems.* Computer J., 1966. **9**: p. 373–380.

112. Hartigan, J.A., *Clustering algorithms*1975, New York ; Toronto: Wiley. xiii, 351.

113. Nikhil R. Pal, J.C.B., and Richard J. Hathaway., *Sequential Competitive Learning and the Fuzzy c-Means Clustering Algorithms.* Neural Networks, 1996. **9**(5): p. 787-796.

114. Bezdek, J.C., *Pattern recognition with fuzzy objective function algorithms*. Advanced Applications in Pattern Recognition1981, New York: Plenum Press. xv, 256 p.

115. Sopko, R., et al., *Mapping pathways and phenotypes by systematic gene overexpression.* Mol Cell, 2006. **21**(3): p. 319-30.

116. Takai, Y., T. Sasaki, and T. Matozaki, *Small GTP-binding proteins.* Physiol Rev, 2001. **81**(1): p. 153-208.

117. Drees, B.L., et al., *A protein interaction map for cell polarity development.* J Cell Biol, 2001. **154**(3): p. 549-71.

118. Roumanie, O., et al., *Functional characterization of the Bag7, Lrg1 and Rgd2 RhoGAP proteins from Saccharomyces cerevisiae.* FEBS Lett, 2001. **506**(2): p. 149-56.

119. Saito, H. and K. Tatebayashi, *Regulation of the osmoregulatory HOG MAPK cascade in yeast.* J Biochem, 2004. **136**(3): p. 267-72.

120. Porro, D., et al., *Identification of different daughter and parent subpopulations in an asynchronously growing Saccharomyces cerevisiae population.* Res Microbiol, 1997. **148**(3): p. 205-15.

121. Wedlich-Soldner, R., et al., *Spontaneous cell polarization through actomyosin-based delivery of the Cdc42 GTPase.* Science, 2003. **299**(5610): p. 1231-5.

122. Schwob, E., et al., *The B-type cyclin kinase inhibitor p40SIC1 controls the G1 to S transition in S. cerevisiae.* Cell, 1994. **79**(2): p. 233-44.

123. Marshall, M.S., et al., *Regulatory function of the Saccharomyces cerevisiae RAS C-terminus.* Mol Cell Biol, 1987. **7**(7): p. 2309-15.

124. Cannon, J.F. and K. Tatchell, *Characterization of Saccharomyces cerevisiae genes encoding subunits of cyclic AMP-dependent protein kinase.* Mol Cell Biol, 1987. **7**(8): p. 2653-63.

125. Kang, P.J., et al., *The Rsr1/Bud1 GTPase interacts with itself and the Cdc42 GTPase during bud-site selection and polarity establishment in budding yeast.* Mol Biol Cell, 2010. **21**(17): p. 3007-16.

126. Fujimura-Kamada, K., T. Hirai, and K. Tanaka, *Essential role of the NH2-terminal region of Cdc24 guanine nucleotide exchange factor in its initial polarized localization in Saccharomyces cerevisiae.* Eukaryot Cell, 2012. **11**(1): p. 2-15.