Title
Evolutionary constraints on the sequence of Ras

Author
Bandaru, Pradeep

Publication Date
2017

Peer reviewed|Thesis/dissertation

Evolutionary constraints on the sequence of Ras

By

Pradeep Bandaru

A dissertation submitted in partial satisfaction of the requirements for the degree of

Doctor in Philosophy

in

Molecular and Cell Biology

in the Graduate Division of the

University of California, Berkeley

Committee in Charge:

Professor John Kuriyan, Chair

Professor Tanja Kortemme

Professor Michael Eisen

Professor Phillip Geissler

Fall 2017

Dedication

"As the heat of a fire reduces wood to ashes, the fire of knowledge burns to ashes all karma."

Acknowledgements

When I first started graduate school in 2012, I expected a relatively breezy path from classes and rotations to publishing and graduating. Suffice to say, reality did not fall in line with my expectations. But at the end, everything worked out as well as it could have and more. Along the way, I made some absolutely great friends and did some very exciting science, and I couldn't be more thankful for the experience.

I first have to thank my family for being there for me throughout every step of graduate school. My parents have been exceptionally supportive and optimistic over the past five years, and they kept me going even when I wanted to quit. One of the main reasons I kept pushing through graduate school was my dad, a fellow scientist. I realized at a young age that science isn't always the easiest career path as a relatively unknown commodity, and he would have loved to study at a place as scientifically vibrant as UC Berkeley. Though he didn't have that opportunity, I always tried to remember that I'm doing this in part for him. My dad has always been an invaluable scientific sounding board, and I hope that he could live vicariously through my experience. My future wife, Sayuri Khandavilli, has also been exceptionally supportive and have constantly pushed me to realize my full potential, for which I am truly grateful. I'm blessed to be the newest Ph.D. in my wonderfully loving family.

I fortuitously joined the Kuriyan lab at UC Berkeley near the end of 2013, and I quickly realized that I wasn't just joining any other lab, but that I was becoming part of a large scientific family. Jeff Iwig first showed me how to prep proteins and he absolutely crushed my ego at every possible chance, but I learned how to take it all in stride. Because of Jeff, I'm the reigning Kuriyan lab taco-eating champion (20 tacos in one sitting), and I'm sure my title will stand the test of time. Steven Jacques, Tiago Barros, Aaron Cantor, and Meg Stratton helped me settle into the lab and made my experience that much more enjoyable, particularly as a shell-shocked new graduate student. Neel Shah taught me pretty much every technique there is in protein chemistry, and I can't possibly downplay his impact on how I think about science. I'm glad Neel put up with my persistent moodiness without much complaint, which probably means that he'll be able to handle just about anything his future graduate students throw at him. Laura Nocka and Ethan McSpadden have been absolutely great friends through my time in the lab, and we've definitely made some great memories over the years.

I've been exceptionally fortunate to have two amazing co-mentors, Rama Ranganathan and Tanja Kortemme. Some of the first papers that drew me into science were Rama's, particularly those on the epistasis and co-evolution of amino acids in proteins. For reasons still unknown to me, I strayed from these interests early on in graduate school, but I quickly regained my senses in the Kuriyan lab where I was lucky enough to take regular trips down to Dallas to interact with Rama and his lab. Rama has such an elegant and profound way of thinking, which somehow gets even better over a few drinks, and he always challenged me to develop generalizable principles about protein evolution over the course of my research. Also, driving far beyond the speed limit through the side roads of Dallas in Rama's Porsche while talking science was a truly near-religious experience. Likewise, Tanja has always pushed me to think about proteins from the perspective of dynamics and design,

and I've benefitted greatly from her meticulous attention to detail and scientific rigor. I also wouldn't have transitioned so smoothly from UCSF to UC Berkeley without Tanja, and I couldn't be more grateful for her mentorship.

Words can't quite convey how thankful I am to have been mentored by John Kuriyan. To this date, I'm not quite sure why John decided to take a chance on me, especially at a time when I was down on my luck and it felt like nobody really believed in me. During this time, I even had the gall to ask John to work on a project that was completely outside of the purview of his lab, which I must admit was a complete shot in the dark. It really says a lot about John that he not only accepted me into his lab, but gave me everything I could possibly need to focus on the science that most excited me and nothing else. John's clarity, intensity, and insight are second to none, and I've learned more about science from him than I can possibly describe. But just as importantly, John taught me how to be professional, responsible, and accountable when I needed it the most. Simply put, I owe John the world and I couldn't be more grateful for everything he's done for me.

Table of Contents

Abstract

Evolutionary constraints on the sequence of Ras

by

Pradeep Bandaru

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor John Kuriyan, Chair

Ras proteins are highly conserved signaling molecules that exhibit regulated, nucleotide-dependent switching between an active GTP-bound state that transduces signals by binding to effector proteins, and an inactive GDP-bound state that cannot bind to effectors. The high conservation of Ras requires mechanistic explanation, especially given that proteins are generally robust to mutation, a concept that was first established from early structural and phylogenetic analysis of hemoglobin by Max Perutz and John Kendrew.

During my thesis research, I adapted a two-hybrid selection system to analyze how mutations affect the functional cycle of human H-Ras, with the ultimate goal of understanding the constraints on the sequence of Ras that give rise to its high evolutionary conservation. My strategy was to isolate just the minimal biochemical network that defines this cycle, comprising Ras, its effector Raf, a GTPase accelerating protein (GAP), and a guanine-nucleotide exchange factor (GEF). Using this selection system, I analyzed the sensitivity of every residue in Ras to mutation in the context of this network, while excluding the effects of the membrane and additional regulatory factors. This approach provided an opportunity, for the first time, to use deep mutational scanning approaches to study how local regulatory networks influence the mutational sensitivity and phenotypic plasticity of key signaling molecules.

I found that Ras exhibits global sensitivity to mutation when regulated by a GAP and a GEF, effectively displaying global constraints that result in the majority of mutations leading to a modest decrease in Ras function. In the absence of regulators, Ras shows considerable tolerance to mutation, as seen previously in saturation mutagenesis experiments on other proteins, where the distribution of mutational effects shifted to be largely neutral. Surprisingly, the analysis of Ras in the absence of regulators also revealed allosteric hotspots of activating mutations in residues that restrain Ras dynamics and promote the inactive, GDP-bound state. This showed that structural fold of Ras is intrinsically capable of accommodating sequence changes that, in evolution, could lead to the acquisition of new function, but could also lead to unwanted Ras activation in disease.

Indeed, oncogenic mutations that disturb the switching mechanism of Ras result in aberrant signaling and cancer, highlighted by the fact that Ras is one of the most important proto-oncogenes in the human genome.

Altogether, my research shows that the local regulatory network places a stringent constraint on the sequence of Ras, and also creates the potential conditions in which it is susceptible to activating mutations. This extended previous observations that mutational sensitivity in proteins is strongly dependent on the selective conditions in which the protein operates. From a practical perspective, though small molecule inhibitors of Ras have yet to achieve clinical relevance despite a concerted effort to obtain such inhibitors, my research also provides a roadmap of allosteric hotspots of Ras activation that can be exploited to design novel cancer therapeutics.

Chapter 1: Introduction

Protein structure is remarkably robust to mutation

Proteins have a traceable evolutionary history that can be revealed by comparing their three-dimensional structures and their primary amino acid sequences. Among the first proteins to have their evolutionary histories extensively characterized were the globins, which laid ground for the notion that proteins are remarkably robust to changes in their primary sequence. Globin proteins were the first proteins whose structures were determined by x-ray crystallography, and are aptly named for their compact, globular protein fold which essentially forms a box around a central heme group that enables globins to bind and store oxygen. The structure of the globin fold is distinctive in that it is typically comprised of eight connected alpha-helices that surround the heme group. At the center of the heme group is an iron atom that binds oxygen, and this iron atom is coordinated by a porphyrin ring and histidine side chains that extend from the adjacent helices of the protein structure. Myoglobin and hemoglobin are the epitome of an exquisitely designed molecular oxygen-transporting system. Evolution has optimized the architecture of the globin fold to store oxygen with high affinity, to readily release oxygen, and to transport waste products, namely carbon dioxide, upon the completion of cellular respiration.

The determination of the sequence of horse hemoglobin in the early 1960s, which was among the first proteins to be sequenced (Braunitzer et al., 1961), laid the groundwork for the notion that proteins are robust to changes in primary sequence. Max Perutz and John Kendrew observed striking similarities between the three-dimensional structures of horse hemoglobin and sperm whale myoglobin, despite the sequence identity between myoglobin and the α- and β- chain of hemoglobin being approximately 30% (Kendrew et al., 1954; Lesk and Chothia, 1980; Perutz et al., 1960). Upon this observation, Perutz remarked:

> How does this arise? It is scarcely conceivable that a three-dimensional template forces the chain to take up this fold. More probably, the chain, once it is synthesized and provided with a haem group around which it can coil, takes up this configuration spontaneously, as the only one which satisfies the stereochemical requirements of its amino acid sequence. This suggests the occurrence of similar sequences throughout this group of proteins, despite their marked differences in amino acid content. This seems all the more likely, since their structural similarity suggests that they have developed from a common genetic precursor.

Subsequent structures of hemoglobin orthologs from a variety of animals, many of which were only approximately 10-20% identical since their divergence in metazoan evolution 600-700 million years ago, further confirmed the ability of vastly divergent globin sequences to adopt similar structures (Figure 1). Perutz's surprise at this ability of proteins does not go without recognition, as it is not evident that such a low level of similarity between globin orthologs should lead to the conservation of three-dimensional structure, especially given the length of a typical globin protein (Sander and Schneider, 1991). Only

two residues are preserved without substitution at the same positions in the polypeptide chain of hemoglobin, one of which is the histidine residue that coordinates the iron atom of the heme group, suggesting that the structurally similar globin orthologs are indeed functionally similar. Thus, much of the difference between the globins is likely neutral sequence change, where evolution has found multiple paths to designing globin variants that are robust to sequence variation.

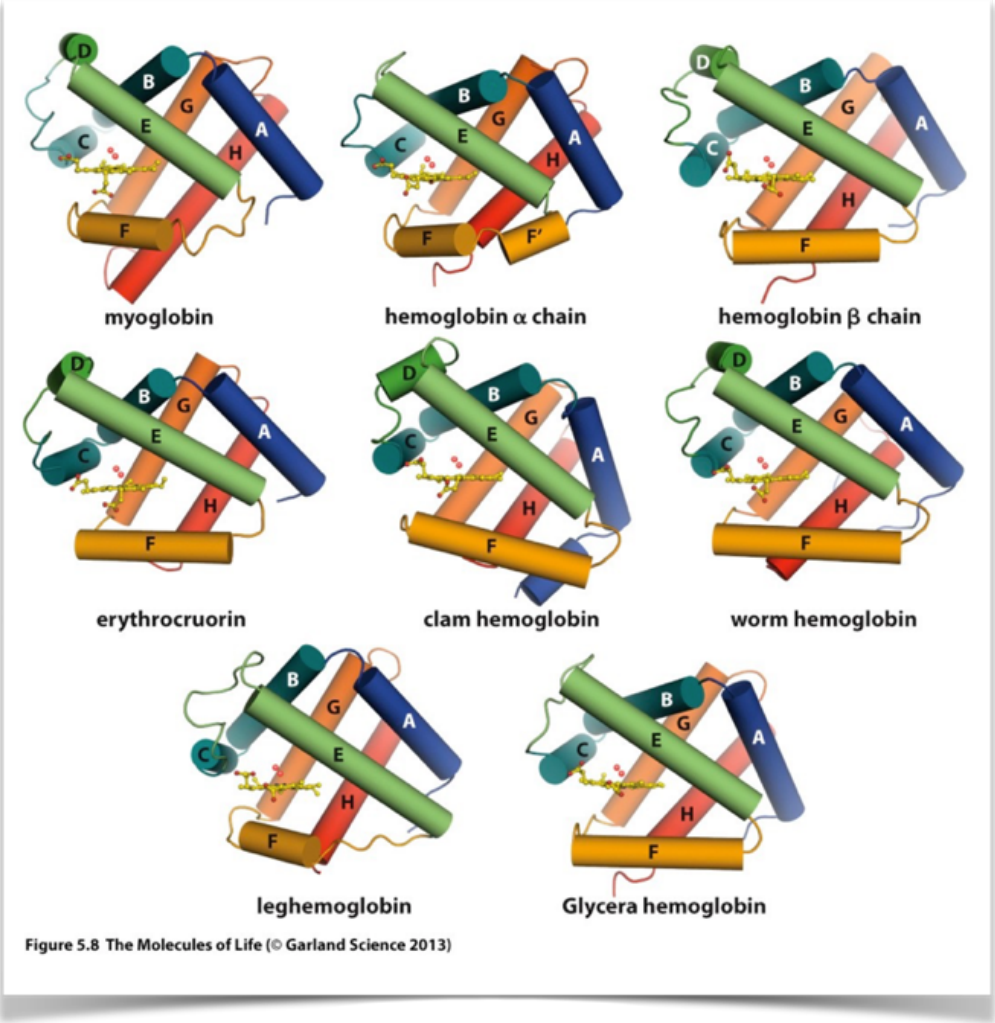

Figure 1. Sequence variation among globin orthologs. Globins are often 10-20% similar over the course of evolution. Despite extensive sequence variation in the globin family, the protein fold is remarkably conserved, suggesting that the protein is indeed robust to mutation. Individual residues that are involved in heme coordination are conserved (not shown), though the rest of the sequence is largely variable. The structural diagrams were adapted from The Molecules of Life, First Edition, and were initially based on structures from (Dickerson and Geis, 1983).

Not all residues in proteins are created equal

The phylogenetic analysis of sequences of various protein families suggests that proteins are tolerant of random mutations throughout their sequence. However, the function of some proteins is significantly altered by mutations at specific positions, suggesting a potentially broad distribution of functional effects in proteins where mutations at some residues are more impactful than others. The idea of a spectrum of functional effects in proteins elicited by mutation was observed from the comparison of globin sequences by Zuckerkandl and Pauling:

> There is no reason to expect that the extent of functional change in a polypeptide chain is proportional to the number of amino acid substitutions in the chain. Many such substitutions may lead to relatively little functional change, whereas at other times the replacement of one single amino acid residue by another may lead to a radical functional change. [...] It is the type rather than the number of amino acid substitutions that is decisive (Zuckerkandl and Pauling, 1965).

The importance of certain residues over others in proteins is directly observed in disease. Linus Pauling suggested in 1949 that sickle-cell anemia, a disease that produces abnormally shaped and rigid red blood cells, arose from molecular defects in hemoglobin (Pauling and Itano, 1949). It was later shown by Vernon Ingram that replacement of two glutamic acid residues by valine in each β-chain of the hemoglobin tetramer was sufficient to produce the sickling phenotype (Ingram, 1956). Subsequent structural studies showed that these substitutions disrupted lateral contacts along the β-chain and caused the defective hemoglobin molecules, termed hemoglobin S, to aggregate in cells (Bookchin and Nagel, 1974; Dean and Schechter, 1978a, 1978b, 1978c; Noguchi and Schechter, 1981).

Studying the distribution of mutational effects in proteins

That a seemingly innocuous single point mutation can disrupt a network of amino acid interactions throughout the β-chain of hemoglobin highlights the nontriviality of predicting the effects of mutations from the protein structure alone. The basis of the distribution of functional effects in a protein is rooted in the thermodynamic coupling between interacting amino acids. Early studies attempted to infer this energetic coupling from sequence alignments of protein families, under the assumption that interactions between amino acids in proteins may be reflected in their correlated evolution (coevolution) in sequences spanning the protein family (Göbel et al., 1994; Lichtarge et al., 1996; Lockless and Ranganathan, 1999; Neher, 1994). Statistical analysis of sequence alignments predicted interacting amino acids, and these interactions are crucial to the overall function of the protein, as demonstrated for PDZ domains, serine proteases, small GTPases, and many other protein families (Halabi et al., 2009; Hatley et al., 2003; Shulman et al., 2004; Smock et al., 2010). These analyses have revealed several generalizable principles that may begin to explain protein structure and function. For example, interacting amino acids often form a sparse, yet spatially contiguous network throughout the entirety of the protein structure for any given protein family, even connecting residues

at distal ends of the structure, whereas the remaining amino acids appear to be statistically independent from each other. The structural architecture of coevolving residues is shown in Figure 2 for the PDZ domain (McLaughlin et al., 2012). This observation suggests that most residues in a protein evolve independently while others coevolve. In retrospect, the coevolution of functionally relevant residues in proteins was again predicted from the analysis of globin sequences by Zuckerkandl and Pauling:

> Of course the two aspects are not unrelated, since the functional effect of a given single substitution will frequently depend on the presence or absence of a number of other substitutions (Zuckerkandl and Pauling, 1965).

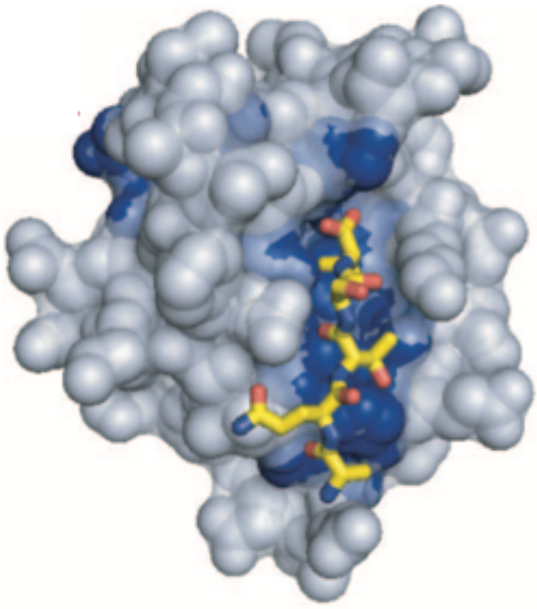

Figure 2. Co-evolving residues in the PDZ domain. Co-evolving residues in the PDZ domain (blue spheres) comprise a sparse network of residues built around the ligand-binding pocket and extending to a distal surface site. The allosteric network that connects the surface site to the ligand-binding pocket traverses through the central hydrophobic core of the PDZ domain. The ligand is shown in yellow sticks. Figure was adapted from (McLaughlin et al., 2012).

An experimental approach for determining the pattern of interactions between amino acids in a protein is the thermodynamic double mutant cycle (Carter et al., 1984; Hidalgo and MacKinnon, 1995; Horovitz and Fersht, 1990). In this approach, the thermodynamic coupling between two residues in a protein is determined by making a mutation to the residues in question, independently and in combination. If mutations x and y at positions i and j act independently, the energetic effect of the double mutation ΔG_{ij}^{xy} is the sum of the effects of each single mutant ($\Delta G_i^x + \Delta G_j^y$). The thermodynamic

coupling between the two amino acids in the protein is expressed as the difference between the effects of the single and the double mutations ($\Delta\Delta G_{ij}^{xy}$):

$$\Delta\Delta G_{ij}^{xy} = (\Delta G_i^x + \Delta G_j^y) - \Delta G_{ij}^{xy}$$

The robustness of a protein to mutation is rooted in the thermodynamic coupling between all of its amino acids. Quantitatively, the thermodynamic coupling measures the degree of cooperativity between amino acids in a protein, which is effectively the parameter that can be statistically estimated from alignments of protein sequences (Lockless and Ranganathan, 1999), and can be extended beyond double mutations to higher order mutations. However, while measuring the coupling between two residues in a protein has proven remarkably effective at understanding the functional roles of amino acids in proteins, it is often practically difficult to execute for multiple residue pairs. Thus, the traditional, first-order approach to studying mutational robustness in proteins has relied upon assessing the energetic effects of single mutations in proteins.

The robustness of proteins to mutation was experimentally demonstrated in a study of four residues that comprise the majority of the interface between the bacterial histidine kinases, PhoQ and PhoP (Podgornaia and Laub, 2015). Studies of protein interfaces have revealed that hotspot residues concentrate the majority of the binding free energy of protein-protein interactions (Clackson and Wells, 1995). For example, the interface between human growth hormone and its cell-surface receptor is quite large at 1,300 Å^2, though typical for binding interfaces between proteins (Davies et al., 1990; Janin and Chothia, 1990), and involves the interaction of about 30 sidechains from each protein (de Vos et al., 1992). Despite the extensive number of sidechain interactions at the interface, mutagenesis of interfacial residues showed that the majority of the binding free energy is determined by two tryptophan residues within a buried hydrophobic patch, and the remainder of the binding free energy is determined by less important hydrophilic contact residues (Clackson and Wells, 1995). For histidine kinases involved in bacterial two component signaling systems, four residues comprise the interface in one histidine kinase, PhoQ. These residues were subject to saturation mutagenesis individually and in combination, generating all 20^4 (160,000) possible combinations of PhoQ variants. The library of PhoQ variants was screened and selected for functional variants that maintained the ability to phosphorylate the partner histidine kinase, PhoP. It was found that the interface was quite degenerate and permissive of a large number of substitutions. Approximately 21% of single point mutants were tolerated, which is consistent with prior studies that have shown that protein interfaces are indeed robust to mutation (DePristo et al., 2005; Harms and Thornton, 2013). Interestingly, the space of higher order mutations was less degenerate than expected (nearly less than 1% of all possible mutations were functional), suggesting that epistatic interactions between residues at the interface constrains mutational space much greater than previously appreciated, and may reflect a fundamental evolutionary constraint this manifested in the limited diversity of natural protein sequences. Furthermore, the energetic contributions of interacting amino acids in proteins are not immediately obvious from examination of crystal structures alone, which

suggests that the mutational robustness of proteins will indeed require the systematic study of higher order mutational effects.

Mutagenesis has commonly been employed to study the distribution of functional effects in many protein model systems. These studies have relied upon making many mutations in regions of the protein, and coupling the effect of the mutation to a functional or energetic readout. Along these lines, an early landmark study concerning the mutational tolerance of proteins showed that λ-repressor, a phage transcriptional repressor, can tolerate many different mutations within the hydrophobic core (Lim and Sauer, 1989). In this study, mutations were made to the DNA sequence of the hydrophobic core of λ-repressor by cassette mutagenesis which allowed single, double, and higher-order mutations to be made to the protein. These protein variants were introduced into a phage lacking a functional λ-repressor protein, and the survival of the phage depends on the expression of an active and functional λ-repressor variant. This screen identified several mutations that led to phage survival, many of which were hydrophobic substitutions, suggesting that the hydrophobic core of the protein was largely tolerant of mutation. However, an *in vitro* measurement of the transfer free energy of several variants identified in the functional screen suggested that there were further constraints on the nature of the substitutions. Namely, the amino acids substitutions that were tolerated were subject to steric and volumetric constraints that did not disrupt the folding and packing of the hydrophobic core. Thus, this study was the first of its kind to demonstrate the robustness of proteins to mutation, and to illuminate the constraints on the nature of tolerated amino acid substitutions by measuring the energetic effects of mutations.

Understanding the interactions between every amino acid in a protein is required to understand its energetic architecture in addition to assessing the mutational robustness of the entire protein. While targeted cassette mutagenesis was an early approach to study these properties, more recently, the individual function of every mutation at every amino acid of a protein has been measured experimentally. Recent advances have exploited the ability of DNA synthesis technologies to create large libraries of protein variants, typically saturation mutagenesis libraries, and coupled these libraries to a functional selection assay for the protein of interest. In another technological advance, high-throughput DNA sequencing is used to accurately count the frequency of each protein variant before and after selection, subsequently enabling a quantitative measure of the tolerance of the entire protein to mutation. Studies using deep sequencing have reinforced the concept that proteins are remarkably tolerant of mutation while retaining the ability to fold and function (Bershtein et al., 2006; Fowler and Fields, 2014; Podgornaia and Laub, 2015; Roscoe et al., 2013; Tripathi and Varadarajan, 2014). In particular, "two-hybrid" selection systems, combined with deep sequencing, have been utilized to determine the effects of mutations on the binding affinities of interacting proteins (Dove et al., 1997; Joung et al., 2000). This strategy has been used to probe the sensitivity to mutation of a PDZ domain-peptide interaction (McLaughlin et al., 2012). Residues in the PDZ domain are highly tolerant of mutation, except for those that govern folding, stability, and energetic interactions with the peptide target.

Are highly conserved proteins less robust to mutation?

The view that proteins are tolerant of mutation contrasts with the observation that many metazoan proteins are extraordinarily conserved in their sequences. Mutagenesis studies have typically focused on those that exhibit considerable sequence variation over evolution, such as PDZ domains. That proteins with considerable sequence variation over evolution are tolerant of mutation is not necessarily surprising, as evolution has recapitulated an in vitro mutation and selection experiment over many millions of years. We can then infer that highly evolutionarily conserved proteins are exquisitely sensitive to mutation, directly contrasting with the established mutational robustness of proteins. This concept was tested using ubiquitin as a model protein system, where the fitness of many ubiquitin variants was tested in vitro and in cells (Leung et al., 2016; Roscoe et al., 2013). Ubiquitin is found in all three branches of life, and in eukaryotes its sequence is virtually identical from protists to metazoans, suggesting that novel functions for ubiquitin did not arise in eukaryotic evolution (Zuin et al., 2014) (Figure 3). In light of its conservation, ubiquitin was indeed found to be sensitive to mutation throughout its sequence, particularly in the hydrophobic core of the protein and at interfaces that mediate binding to protein targets in cells, suggesting that binding is an important determinant of ubiquitin function (Leung et al., 2016; Roscoe et al., 2013). Sequence variation that is tolerated in ubiquitin generally preserves its ability to fold and bind its protein targets. Thus, the evolutionary conservation of a protein is directly indicative of its mutational robustness, and mutations that are tolerated are additionally subject to constraints that preserve protein function.

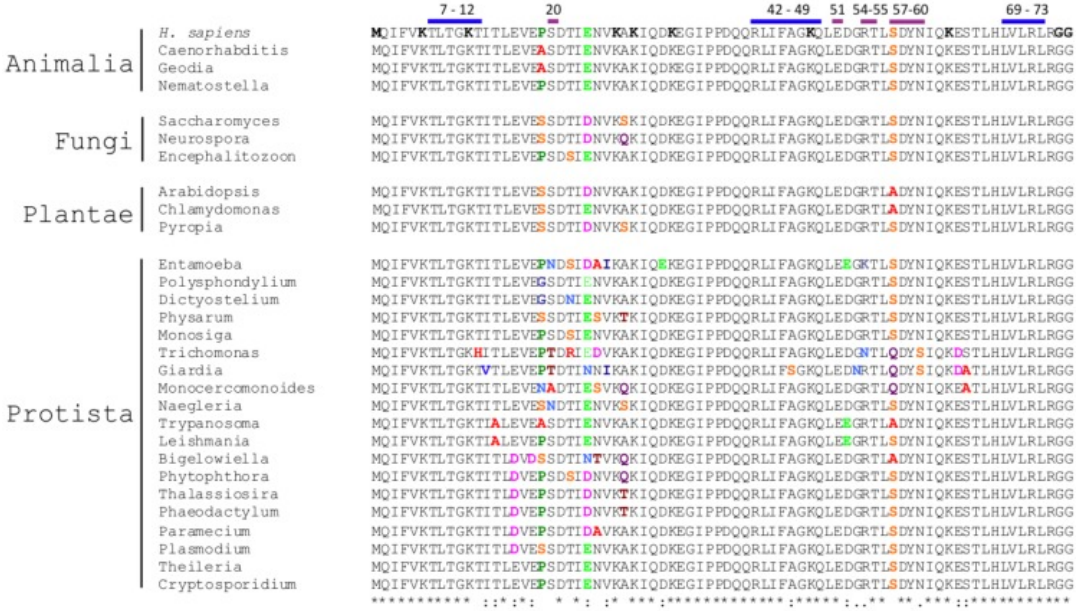

Figure 3. Alignment of metazoan ubiquitin orthologs. Ubiquitin shows a high level of sequence conservation over the course of metazoan evolution (grey

residues). Sequence variation is highlighted in different colored residues. Sequence alignment was adapted from (Zuin et al., 2014).

The constraints on the substitutions to protein sequence have largely been dissected for simple binding proteins such as the PDZ domain and ubiquitin. However, interesting test cases for the mutational robustness of proteins are multifunctional enzymes such as signaling proteins. In cells, signaling proteins typically bind to multiple partners and perform enzymatic reactions that are often highly regulated by auxiliary proteins. These proteins, including kinases and small GTPases, are central nodes in signaling networks and often act as molecular switches that control the response to cellular stimuli. Signaling proteins are often highly conserved in metazoans, and the origin of the conservation of signaling proteins may stem from their essentiality in cells. It has been posited that "dispensable" proteins that do not contribute significantly to the overall fitness of an organism evolve at a faster rate, and are thus less conserved, in comparison to essential proteins which evolve at slower rates (Hirsh and Fraser, 2001). Furthermore, studies have also suggested that the rate of a protein's evolution decreases as the number of molecular interactions in which it participates increases (Dickerson, 1971; Fraser et al., 2002; Wilson et al., 1977). Though protein expression level may bias the strength of these correlations, these general relationships may contribute to the overall conservation of signaling proteins.

Deconstructing the constraints on the sequence of Ras

Consider, for example, the Ras proteins, which are prototypical members of a superfamily of membrane-associated small G-proteins that control an extraordinary range of cellular processes (Bourne et al., 1991; Rojas et al., 2012). Ras proteins are capable of producing a diversity of signaling outputs by binding to effector proteins, most importantly to the Raf family of kinases, which activate the MAP kinase pathway (Wittinghofer and Nassar, 1996; Wellbrock et al., 2004). Ras proteins cycle between an active GTP-bound state that transduces signals by binding to effector proteins, and an inactive GDP-bound state that cannot bind to effectors (Figure 4). The sequences of the three principal isoforms of Ras (H-Ras, K-Ras and N-Ras) are almost invariant in vertebrates, and Ras proteins are highly conserved across all metazoans (Figure 5 shows a phylogenetic tree based on alignment of H-Ras orthologs).

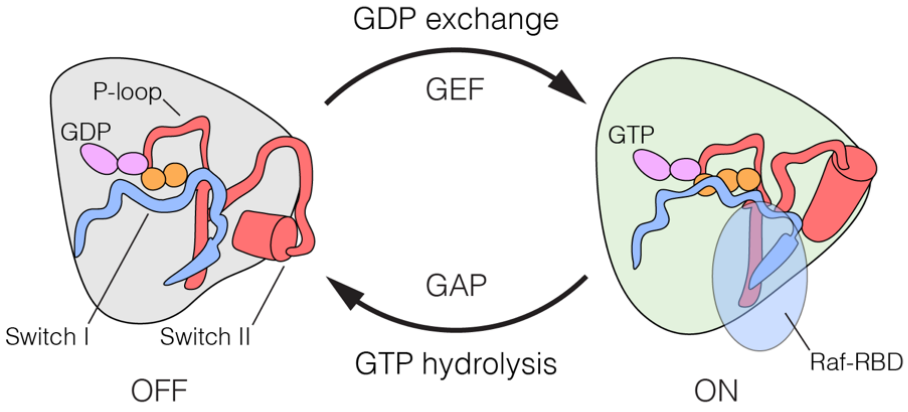

Figure 4. The Ras switching cycle. Ras cycles between an active, GTP-bound state and an inactive, GDP-bound state. Ras•GTP binds to effector proteins, such as Raf kinase, which binds to Switch I. The intrinsic hydrolysis of GTP is slow, unless catalyzed by a GTPase activating protein (GAP) which binds to Switch I. Intrinsic GDP release is also a slow process, unless facilitated by a guanine nucleotide exchange factor (GEF) which binds to both Switch I and Switch II.

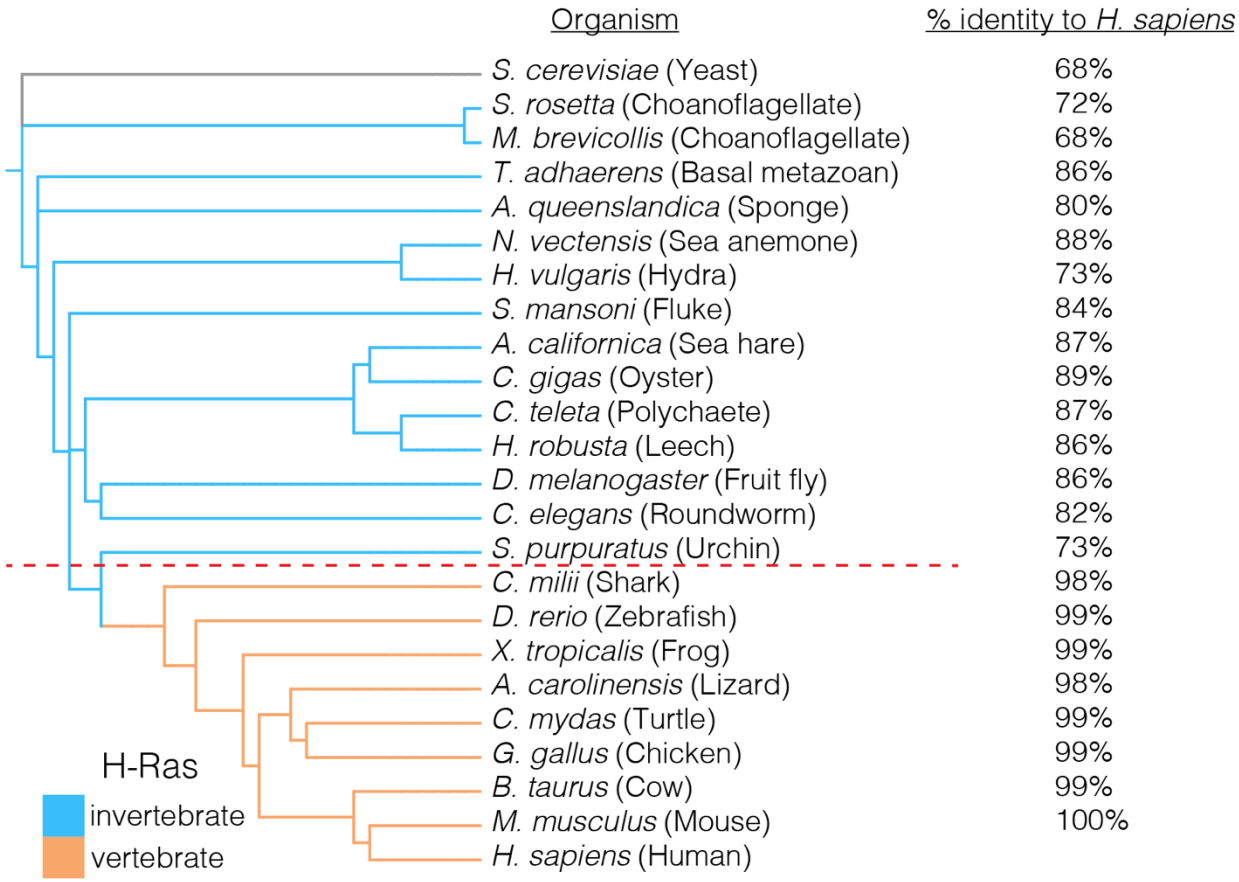

Figure 5. Conservation of the GTPase domains of H-Ras. H-Ras displays a high level of sequence conservation in metazoans. Sequences of H-Ras orthologs that were most similar to human H-Ras were used in this analysis. Invertebrate Ras sequences, such as those from a choanoflagellate or a sponge, are greater than ~70% identical to human H-Ras. The H-Ras sequence is nearly invariant in vertebrates, where greater than 98% sequence identity to human H-Ras is observed. Yeast Ras is an outgroup in the phylogenetic analysis, and plays a different functional role in comparison to metazoan H-Ras.

Given that protein structures are tolerant to mutation, why are the sequences of Ras proteins (referred to collectively as "Ras") so highly conserved? In particular, to what extent does the high degree of sequence conservation arise from the necessity to maintain the

GTPase cycle that switches signaling on and off? The inherent sensitivity of Ras to activation by mutation is highlighted by the fact that it is an important oncogene (Prior et al., 2012). The signaling activity of Ras can be increased by mutations that prolong the GTP-bound state by reducing the rate of GTP hydrolysis, or that increase the rate at which GDP bound to Ras is replaced by GTP. The switch between the GDP- and GTP-bound states is accompanied by correlated conformational changes distributed throughout the structure of Ras (Grant et al., 2009; Guterres et al., 2015). Two segments of Ras, referred to as Switch I and Switch II, undergo substantial rearrangements when GTP is replaced by GDP, as illustrated schematically in Figure 4 (for a comprehensive review, see (Vetter and Wittinghofer, 2001)). Effector proteins, such as Raf and phosphatidyl inositol 3' kinase, transduce signals from Ras•GTP by binding to Switch I, and the conformation of the switch regions is also important for binding to GAPs and GEFs. We expect that the switching, enzymatic, and binding functions of Ras will impose stringent constraints on the sequence, but a systematic study of the nature of these constraints is lacking.

In this thesis, I present an analysis of the constraints on the primary sequence of the GTPase domain of Ras that arise as a consequence of its function as a molecular switch that is turned off and on by GAPs and GEFs. To do this, I adapted the saturation point-mutagenesis and bacterial two-hybrid selection strategy that had been used to study sequence variation in PDZ domains (Raman et al., 2016). In this system, the binding of Ras•GTP to the Ras-binding domain (RBD) of C-Raf is coupled to the transcription of an antibiotic resistance factor, which enables the fitness of mutant forms of Ras to be evaluated through their effect on bacterial growth in the presence of an antibiotic. Co-expression of a GAP and a GEF enables the effect of regulators to be evaluated. The two-hybrid system omits some of the constraints that Ras normally operates under in a mammalian cell, such as interaction with the membrane (Abankwa et al., 2008; Mazhab-Jafari et al., 2015). Even without these constraints, the experiments show that the necessity of maintaining a properly regulated switching function severely constrains the sequence of Ras relative to that of proteins with a simpler binding function, such as PDZ domains or lambda repressor.

I find that for regulated wild-type H-Ras, in the presence of a GAP and a GEF, most mutations reduce the function of Ras to a mild extent, with the more severe effects being associated with mutations to residues that are critical for binding to GTP, the Raf-RBD and the regulators, and for the formation of the hydrophobic core. Strikingly, a distinct pattern of mutational sensitivity emerges for unregulated wild-type Ras, in the absence of the GAP and the GEF. Under these conditions, most mutations shift to being near-neutral and a set of gain-of-function mutations at hotspot residues appears, where almost any change to the wild-type residue results in increased activity. I show that these hotspots of activating mutations, which form a spatially contiguous network distributed around the GTPase domain, correspond to residues that reduce spontaneous switching through allosteric coupling to the GTP binding site. Mutations that activate Ras inappropriately would normally be eliminated by natural selection, and the combination of inactivating and activating mutations places constraints on nearly all of the residues in Ras. I also carried out a saturation mutagenesis experiment in the background of the oncogenic G12V mutation, and found that the presence of this mutation reduces the effect of mutations that

activate wild-type H-Ras, indicative of coupling between the G12V mutation and the broadly distributed hotspot residues. Finally, I present an evolutionary analysis that, along with the mutational data, indicates that allosteric control in Ras was modified during the emergence of the vertebrate lineage. These data show that mutational sensitivity in Ras is strongly context-dependent on it regulatory network, a finding that can explain the high sequence conservation of Ras orthologs.

References

Abankwa, D., Hanzal-Bayer, M., Ariotti, N., Plowman, S.J., Gorfe, A.A., Parton, R.G., McCammon, J.A., and Hancock, J.F. (2008). A novel switch region regulates H-ras membrane orientation and signal output. EMBO J *27*, 727–735.

Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N., and Tawfik, D.S. (2006). Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. Nature *444*, 929–932.

Bookchin, R.M., and Nagel, R.L. (1974). Interactions between human hemoglobins: sickling and related phenomena. Semin Hematol *11*, 577–595.

Braunitzer, G., Gehring-Mueller, R., Hilschmann, N., Hilse, K., Hobom, G., Rudloff, V., and Wittmann-Liebold, B. (1961). [The structure of normal adult human hemoglobins]. Hoppe-Seyler's Z Physiol Chem *325*, 283–286.

Carter, P.J., Winter, G., Wilkinson, A.J., and Fersht, A.R. (1984). The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (Bacillus stearothermophilus). Cell *38*, 835–840.

Clackson, T., and Wells, J.A. (1995). A hot spot of binding energy in a hormone-receptor interface. Science *267*, 383–386.

Davies, D.R., Padlan, E.A., and Sheriff, S. (1990). Antibody-antigen complexes. Annu Rev Biochem *59*, 439–473.

Dean, J., and Schechter, A.N. (1978a). Sickle-cell anemia: molecular and cellular bases of therapeutic approaches (second of three parts). N Engl J Med *299*, 804–811.

Dean, J., and Schechter, A.N. (1978b). Sickle-cell anemia: molecular and cellular bases of therapeutic approaches (first of three parts). N Engl J Med *299*, 752–763.

Dean, J., and Schechter, A.N. (1978c). Sickle-cell anemia: molecular and cellular bases of therapeutic approaches (third of three parts). N Engl J Med *299*, 863–870.

DePristo, M.A., Weinreich, D.M. and Hartl, D.L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Reviews. Genetics* 6(9), 678–687.

Dickerson, R.E. (1971). The structures of cytochrome c and the rates of molecular evolution. J Mol Evol *1*, 26–45.

Dickerson, R.E. and Geis, I. 1983. Hemoglobin: Structure, Function, Evolution, and

Pathology (The Benjamin/Cummings series in the life sciences). Benjamin.

Dove, S.L., Joung, J.K., and Hochschild, A. (1997). Activation of prokaryotic transcription through arbitrary protein-protein contacts. Nature *386*, 627–630.

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nat Methods *11*, 801–807.

Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C., and Feldman, M.W. (2002). Evolutionary rate in the protein interaction network. Science *296*, 750–752.

Göbel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. Proteins *18*, 309–317.

Grant, B.J., Gorfe, A.A., and McCammon, J.A. (2009). Ras conformational switching: simulating nucleotide-dependent conformational transitions with accelerated molecular dynamics. PLoS Comput Biol *5*, e1000325.

Guterres, H., Ma, B., Nussinov, R., and Mattos, C. (2015). Ras Isoforms Conformational Clustering and Community Networks Studies: Simulating Ras with Accelerated Molecular Dynamics. FASEB J.

Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. Cell *138*, 774–786.

Harms, M.J. and Thornton, J.W. (2013). Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nature Reviews. Genetics* 14(8), 559–571.

Hatley, M.E., Lockless, S.W., Gibson, S.K., Gilman, A.G., and Ranganathan, R. (2003). Allosteric determinants in guanine nucleotide-binding proteins. Proc Natl Acad Sci U S A *100*, 14445–14450.

Hidalgo, P., and MacKinnon, R. (1995). Revealing the architecture of a K+ channel pore through mutant cycles with a peptide inhibitor. Science *268*, 307–310.

Hirsh, A.E., and Fraser, H.B. (2001). Protein dispensability and rate of evolution. Nature *411*, 1046–1049.

Horovitz, A., and Fersht, A.R. (1990). Strategy for analysing the co-operativity of intramolecular interactions in peptides and proteins. J Mol Biol *214*, 613–617.

Ingram, V.M. (1956). A Specific Chemical Difference Between the Globins of Normal Human and Sickle-Cell Anæmia Hæmoglobin. Nature *178*, 792–794.

Janin, J., and Chothia, C. (1990). The structure of protein-protein recognition sites. J Biol Chem *265*, 16027–16030.

Joung, J.K., Ramm, E.I., and Pabo, C.O. (2000). A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. Proc Natl Acad Sci U S A *97*, 7382–7387.

Kendrew, J.C., Parrish, R.G., Marrack, J.R., and Orlans, E.S. (1954). The species specificity of myoglobin. Nature *174*, 946–949.

Lesk, A.M., and Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J Mol Biol *136*, 225–270.

Leung, I., Dekel, A., Shifman, J.M., and Sidhu, S.S. (2016). Saturation scanning of ubiquitin variants reveals a common hot spot for binding to USP2 and USP21. Proc Natl Acad Sci U S A *113*, 8705–8710.

Lichtarge, O., Bourne, H.R., and Cohen, F.E. (1996). An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol *257*, 342–358.

Lim, W.A., and Sauer, R.T. (1989). Alternative packing arrangements in the hydrophobic core of lambda repressor. Nature *339*, 31–36.

Lockless, S.W., and Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. Science *286*, 295–299.

Mazhab-Jafari, M.T., Marshall, C.B., Smith, M.J., Gasmi-Seabrook, G.M.C., Stathopulos, P.B., Inagaki, F., Kay, L.E., Neel, B.G., and Ikura, M. (2015). Oncogenic and RASopathy-associated K-RAS mutations relieve membrane-dependent occlusion of the effector-binding site. Proc Natl Acad Sci U S A *112*, 6625–6630.

McLaughlin, R.N., Poelwijk, F.J., Raman, A., Gosal, W.S., and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. Nature *491*, 138–142.

Neher, E. (1994). How frequent are correlated changes in families of protein sequences? Proc Natl Acad Sci U S A *91*, 98–102.

Noguchi, C.T., and Schechter, A.N. (1981). The intracellular polymerization of sickle hemoglobin and its relevance to sickle cell disease. Blood *58*, 1057–1068.

Pauling, L., and Itano, H.A. (1949). Sickle cell anemia a molecular disease. Science *110*, 543–548.

Perutz, M.F., Rossmann, M.G., Cullis, A.F., Muirhead, H., Will, G., and North, A.C. (1960). Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. Nature *185*, 416–422.

Podgornaia, A.I., and Laub, M.T. (2015). Protein evolution. Pervasive degeneracy and epistasis in a protein-protein interface. Science *347*, 673–677.

Prior, I.A., Lewis, P.D., and Mattos, C. (2012). A comprehensive survey of Ras mutations in cancer. Cancer Res *72*, 2457–2467.

Raman, A.S., White, K.I., and Ranganathan, R. (2016). Origins of Allostery and Evolvability in Proteins: A Case Study. Cell.

Roscoe, B.P., Thayer, K.M., Zeldovich, K.B., Fushman, D., and Bolon, D.N.A. (2013). Analyses of the effects of all ubiquitin point mutants on yeast growth rate. J Mol Biol *425*, 1363–1377.

Sander, C., and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins *9*, 56–68.

Shulman, A.I., Larson, C., Mangelsdorf, D.J., and Ranganathan, R. (2004). Structural determinants of allosteric ligand activation in RXR heterodimers. Cell *116*, 417–429.

Smock, R.G., Rivoire, O., Russ, W.P., Swain, J.F., Leibler, S., Ranganathan, R., and Gierasch, L.M. (2010). An interdomain sector mediating allostery in Hsp70 molecular chaperones. Mol Syst Biol *6*, 414.

Tripathi, A., and Varadarajan, R. (2014). Residue specific contributions to stability and activity inferred from saturation mutagenesis and deep sequencing. Curr Opin Struct Biol *24*, 63–71.

Vetter, I.R., and Wittinghofer, A. (2001). The guanine nucleotide-binding switch in three dimensions. Science *294*, 1299–1304.

de Vos, A.M., Ultsch, M., and Kossiakoff, A.A. (1992). Human growth hormone and extracellular domain of its receptor: crystal structure of the complex. Science *255*, 306–312.

Wilson, A.C., Carlson, S.S., and White, T.J. (1977). Biochemical evolution. Annu Rev Biochem *46*, 573–639.

Zuckerkandl, E., and Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In Evolving Genes and Proteins, (Elsevier), pp. 97–166.

Zuin, A., Isasa, M., and Crosas, B. (2014). Ubiquitin signaling: extreme conservation as a

source of diversity. Cells *3*, 690–701.

Chapter 2: Development of a high-throughput assay for Ras function

Introduction

In order to generally assess the mutational robustness of Ras, I needed to design an assay that could quantitatively measure the functional effect of many mutations to the protein sequence. Such an assay would allow for the measurement of thousands of mutant forms of Ras, comprising a saturation point-mutagenesis library of the entire protein. The functional effects of Ras mutations would be correlated to a measurable biophysical parameter that reflects an aspect of Ras function. Furthermore, the assay should as closely as possible recapitulate all of the constraints on Ras function, such as regulation by auxiliary enzymes. Given the need to tractably introduce thousands of point mutations and measure their effects on Ras function in the presence of co-expressed regulatory proteins, I needed to work in an organism that allowed for the simultaneous expression of multiple proteins and for the effects of mutations to Ras to be coupled to the fitness of the organism itself. This chapter will detail the design of the selection system for Ras, as well as creation of the saturation point-mutagenesis library of Ras variants. The design of the selection system was performed in close consultation with Bill Russ, Arjun Raman, and Rama Ranganathan. The design of the saturation point-mutagenesis library of Ras was done in close consultation with Neel Shah in the Kuriyan lab.

Requirements for an assay for Ras function

With the goal of being able to study the function of all amino acids in Ras, I wanted to design an assay that met the following criteria:

1. The measure of function should be reflective of the known constraints imposed on Ras. This measure should correlate to the fitness of the organism in which the assay is being performed.
2. The assay should reproducibly measure the effects of mutations to Ras, in the presence and absence of regulatory proteins.
3. The assay should be high-throughput, and capable of measuring thousands of individual Ras mutants in order to sample enough of the sequence space to then make global statements about the mutational robustness of Ras.

General methods to study protein function and protein-protein interactions

Many methods exist for screening for Ras function and for generally studying protein-protein interactions. Each method is limited in its own way, either by throughput, for the need to work with purified proteins, or for the lack of precise quantification. For example, fluorescence polarization has been used to measure the interaction of GTP loaded Ras with a Ras-binding domain from Raf kinase (Raf-RBD) (Guzmán et al., 2014). Isothermal titration calorimetry has also been used to measure this interaction (Kiel et al., 2009; Rudolph et al., 2001; Wohlgemuth et al., 2005). Though these methods provide a precise quantification of the binding affinity between the two proteins, they would require the purification of thousands of Ras variants *in vitro*, an undertaking that would be

incredibly laborious and time-consuming. In this work, I used isothermal titration calorimetry to measure the affinity of approximately 30 Ras mutants for the Raf-RBD, but even this scale of measurements took over a month to complete. Furthermore, these methods do not necessarily provide a sensible readout of the interaction in the presence of regulation by a GTPase accelerating protein (GAP) and guanine-nucleotide exchange factor (GEF). In fact, an *in vitro* assay has been developed to monitor the output of the entire Ras regulatory cycle (Coyle and Lim, 2016), and though this assay allows for the measurement of the regulated cycling of Ras between a GTP- and GDP-bound state, cycling would have to be measured for a long duration of time and many variants of Ras would again have to be expressed and purified to make global conclusions.

Initially, a mammalian cell selection assay was proposed to be used to measure the function of Ras variants. This assay is based on the work of Erica Kovacs in the Kuriyan lab, who developed a platform to measure the downstream effects of activation of the guanine nucleotide exchange factor SOS when activated by mitogen stimulation. Conceptually, a variant of Ras could be transformed into HEK293 cells, and the activation of Ras by guanine nucleotide exchange factors could be measured by stimulating the cells with a mitogen such as epidermal growth factor. The downstream signaling effects of the Ras variant could, in theory, be measured by using an antibody for the MAP kinase ERK. I initially tried to develop this approach for a potentially large library of Ras variants whose effects would be read out by cell sorting, but several technical hurdles were encountered, such as the fact that mammalian cells would take up multiple Ras variants per cell using transient plasmid transfection. Instead, we turned to different organisms in which to measure the function of Ras mutations. I will later describe the development of a novel mammalian cell selection assay that has been adapted to measure Ras function.

I next explored the possibility of studying Ras function in the context of a yeast two-hybrid assay, which is a platform to detect the interaction of two proteins in a yeast nucleus (Fields and Song, 1989; Legrain and Selig, 2000). This assay works by fusing the GAL4 N-terminal DNA-binding domain to one protein of interest, while also fusing the GAL4 C-terminal transcription-activating domain to the other protein of interest. If the two proteins interact, the two domains of GAL4 are brought into close proximity, subsequently activating transcription of a reporter gene (Fields and Song, 1989). This assay enables screening a large number of potential interactors of a protein of interest, and has been previously used against large cDNA libraries (Legrain et al., 2001). Upon inducing the expression of the fusion proteins, the entire library is plated onto selective media, and surviving colonies in theory contain a viable interaction partner of the protein of interest. However, this assay is susceptible to false positive interactions that can lead to cell growth. Furthermore, the function of an individual protein is only scored as either permitting cell growth or not, and thus, there is no nuanced measure of protein function in the yeast two-hybrid assay. Finally, the presence of endogenous Ras in yeast, as well as endogenous GAP and GEF proteins, can convolutes the ultimate readout of a yeast two-hybrid assay. However, I do use a yeast assay for Ras function later in this work to validate the effects of Ras mutations in a completely orthogonal manner to the selection system to be further detailed (Sass et al., 1986).

The bacterial two-hybrid assay to study protein-protein interactions

Because the two-hybrid assay allows for a high-throughput screen of protein variants, I decided to build an analogous assay for Ras which provided a measure of Ras function that was linked to binding to the Raf-RBD in the presence and absence of the components of the Ras regulatory cycle. I decided to develop a two-hybrid system in the context of *E. coli*, analogous to a system developed by our collaborators in the Ranganathan lab at UTSW (McLaughlin et al., 2012; Raman et al., 2016). Bacteria do not contain Ras-superfamily G-proteins and their associated downstream effectors such as Raf-RBD, and also do not contain GAP or GEF proteins that modulate the cycling of Ras, thereby eliminating the potential convolution of the results of a screen of Ras variants. Furthermore, bacteria are able to take up one unique Ras variant per cell, and allow for the introduction of multiple plasmids harboring components of the regulatory cycle, so they are easily genetically manipulated in a controllable manner for the purposes of my work.

The original bacterial two-hybrid system in *E. coli* couples a protein-protein interaction of interest to a transcriptional readout (Dove et al., 1997; Joung et al., 2000). The assay requires fusion of one protein to a DNA-binding domain and another protein to one of the subunits of RNA polymerase. In the presence of a plasmid containing the appropriate DNA-binding motif upstream of a reporter gene for antibiotic resistance, activation of transcription occurs if the two proteins of interest directly interact. The general principle of transcriptional activation by promoter binding and interaction with RNA polymerase is also used by bacterial transcriptional activators (Busby and Ebright, 1994). Typically, bacterial-two hybrid assays use the phage λ-cI protein for its DNA-binding domain, the N-terminal domain of the α-subunit of RNA polymerase, and β-galactosidase or β-lactamase as the reporter gene (Dove and Hochschild, 2004).

The endogenous phage-λ promoter, P_{RM}, contains three operators for λ-cI, two upstream of and one overlapping with the polymerase binding site (Li et al., 1994). The N-terminal domain of λ-cI dimerizes, and binds to any of its three operators, but has the highest affinity for O_R1, and O_R2 is occupied after O_R1 binding (Li et al., 1994). λ-cI thus occupies both O_R1 and O_R2, which are approximately 20 base pairs apart. λ-cI binding activates transcription when fused to a protein that interacts directly with RNA polymerase (Dove and Hochschild, 2004). In the version of the bacterial two-hybrid that I use, which was modified in the Ranganathan lab, O_R3, which overlaps with the promoter and thus represses transcription when occupied, has been mutated to abolish binding and repression of transcription from P_{RM}. Furthermore, the N-terminal domain of λ-cI also directly interacts with the α-subunit of RNA polymerase and activates transcription independent of the λ-cI fusion when bound to O_R2 (Kuldell and Hochschild, 1994). α-subunit-independent transcriptional activation is mediated through two amino acids revealed through mutagenesis experiments performed by the Ranganathan lab. The Ranganathan lab found that the E34P mutation significantly reduced P_{RM} expression of the reporter gene in the absence of an interaction and therefore increased the range of reporter gene levels between wild-type and weakly-binding mutants of their proteins of interest (McLaughlin et al., 2012). I adopted this version of the bacterial two-hybrid system previously developed

in the Ranganathan lab for expression of Ras in order to monitor its interaction with the Raf-RBD.

A bacterial two-hybrid assay for Ras function

The genetic components of the two-hybrid assay exist on three plasmids (Figure 1). The pZS22 plasmid expresses the C-Raf-RBD fused to the λ-cI protein under the control of an IPTG-inducible promoter, $P_L$lacO1, and contains a low copy number origin of replication, pSC101, and a trimethoprim resistance cassette. The pZA31 plasmid expresses H-Ras (residues 2-166), fused to the N-terminal domain of the α-subunit of E. coli RNA polymerase. This plasmid contains a doxycycline sensitive promoter, the low copy number p15A origin of replication, and a kanamycin resistance cassette. The final plasmid contains the P_{RM} promoter, with a mutated O_R3, upstream of the chloramphenicol resistance gene chloramphenicol acyltransferase (CAT), the medium copy number ColE1 origin, and the ampicillin resistance gene, β-lactamase. The GAP (the catalytic domain of p120 RasGAP) and the GEF (the catalytic domain of RasGRP1) were optionally expressed on the same plasmid architecture. The GAP was expressed on the same transcript as the Raf-RBD, separated by an internal ribosome binding site, and the GEF was expressed under the control of the $P_L$lacO1 promoter introduced into the CAT plasmid.

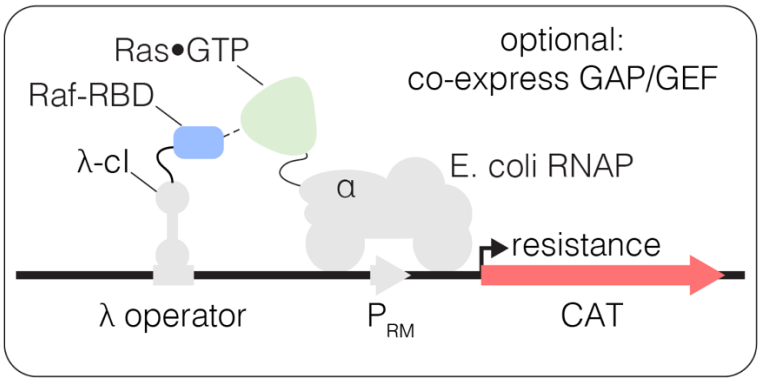

Figure 1. The bacterial two-hybrid system. The bacterial two-hybrid system couples the Ras•GTP:Raf-RBD interaction to the production of an antibiotic resistance factor. The Ras variant library, the Raf-RBD, and the antibiotic resistance factor are encoded on three inducible plasmids. The GAP and GEF can also be co-expressed in the bacterial two-hybrid system. After protein expression, a fraction of the cells is removed and the plasmids encoding the Ras variant library are isolated and deep sequenced to count the frequency of each variant before antibiotic selection. The remainder of cells are subject to antibiotic selection with chloramphenicol and the plasmids encoding the Ras variant library are isolated and deep sequenced to count the frequency of each variant after antibiotic selection. The counts of each variant before and after selection are used to calculate the enrichment of each Ras variant.

In order to decrease background reporter gene expression, the fusion protein plasmids are maintained in a repressed state by the constitutive expression of the lac- and tet-repressors. The genes for these repressors, lacI and tetR, were integrated into the genome of the *E. coli* K12 derivative MC4100 to create the MC4100-Z1 cell line (Lutz and Bujard, 1997). In the absence of IPTG and doxycycline, expression of the fusion proteins and any resulting CAT should remain low. Upon inducing protein expression, Ras interacts with the Raf-RBD, leading to the production of CAT. Chloramphenicol resistance is measure as an enhancement in the growth rate, which is a true measure of cellular fitness. An advantage to a growth rate measurement is that the exponential nature of bacterial growth provides a stringent selection, where beneficial mutations are rapidly enriched exponentially over deleterious mutations. Prior iterations of the bacterial two-hybrid assay successfully used eGFP fluorescence as a selective marker, suggesting a modularity in the choice of a reporter system (McLaughlin et al., 2012).

Creation of a saturation point-mutagenesis library of Ras variants

In this thesis, I will describe the use of the bacterial two-hybrid system to carry out four experiments, each of which involved analysis of a complete set of single-site mutations for the relevant Ras protein, expressed together with Raf-RBD. In the first experiment, variants of wild-type Ras were expressed in the presence of the GAP and the GEF (a condition we refer to as "*regulated-Ras*"). In the next experiment, Ras variants were expressed with the GAP, but without the GEF ("*attenuated-Ras*"). In the third experiment, variants of wild-type Ras were expressed without the GAP and the GEF ("*unregulated-Ras*"). Finally, I examined the effects of an oncogenic mutation by expressing variants of Ras-G12V without the GAP and the GEF ("*Ras-G12V*").

For these experiments, I created two saturation point-mutagenesis libraries of Ras variants, one for wild-type Ras and one for Ras-G12V, by site-directed mutagenic PCR as described (McLaughlin et al., 2012), using primers containing randomized bases at each codon. The mutagenic primers contain NNS codons, where N represents a mixture of A/T/G/C and S represents a mixture of G/C, in the nucleotide sequence of Ras. In this way, I created libraries in which the amino acid sequence of Ras is randomized one position at a time. The libraries were then cloned into the bacterial two-hybrid expression vector and deep-sequenced to ensure that each variant was represented in roughly equal proportion. *E. coli* cells were transformed with the libraries, screened in the bacterial two-hybrid system, and deep-sequenced before and after selection in order to calculate the fitness effects of mutations in Ras.

Quantifying the functional effects of Ras mutations

Chloramphenicol concentration and selection times were chosen to give maximal growth differences between oncogenic Ras mutants (e.g. G12V, Q61L) and mutants that are impaired in their ability to bind to Raf (e.g. T35S, D38N) (Chuang et al., 1994) (Figure 2). Identical selection conditions were used for the four point-mutagenesis experiments. The ability of the two-hybrid system to report on Ras activity is illustrated in (Figure 2A-B), which shows that the growth of bacteria expressing wild-type Ras decreases when an

inactivating mutation is made to the GEF. In this case, Ras•GTP is hydrolyzed to Ras•GDP by the GAP, reducing CAT expression.

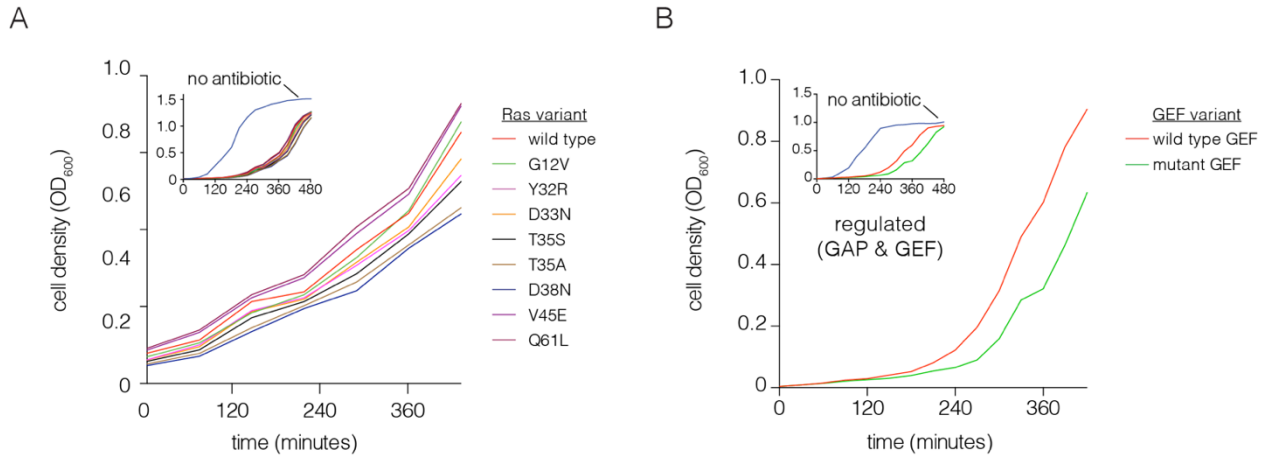

Figure 2. Growth rates in the bacterial two-hybrid system. A. Growth rate differences are observed for Ras variants that perturb Raf-RBD binding affinity in the presence of chloramphenicol. Cells rapidly grow to high density when no chloramphenicol is added. Oncogenic mutations exhibit faster growth due to increased CAT production (G12V, Q61L). The other mutations shown here weaken Ras•GTP:Raf-RBD binding affinity and lead to decreased growth rates. These growth curves were measured without a GAP or GEF. **B.** Growth rate differences are observed with co-expressed p120 RasGAP (GAP) and RasGRP1 (GEF). Cells grow to high density without chloramphenicol. In the presence of an antibiotic and co-expression of GAP and GEF, a mutation in the GEF (F929A) that impairs binding to Ras reduces the growth rate.

Expression of the full regulatory cycle in the bacterial two-hybrid system modulates the concentration of active Ras•GTP available to bind the Raf-RBD. Removal of these regulatory components subsequently leads to a change in the concentration of Ras•GTP, and thus, we developed a theoretical framework to understand these shifts in [Ras•GTP] as the bacterial two-hybrid experiment is performed. This theoretical framework is based on the equilibrium binding of Ras•GTP and Raf-RBD, which leads to the formation of a complex that transmits signal through the MAP Kinase pathway in native signaling systems. The formation of Ras•GTP:RBD complex is driven out of equilibrium by the GAP-mediated hydrolysis of Ras•GTP to Ras•GDP and the GEF-mediated exchange of Ras•GDP to Ras•GTP. Figure 3 depicts these processes in the Ras regulatory network. Production of Ras•GTP and Ras•GDP in the Ras regulatory cycle is given by the following equations:

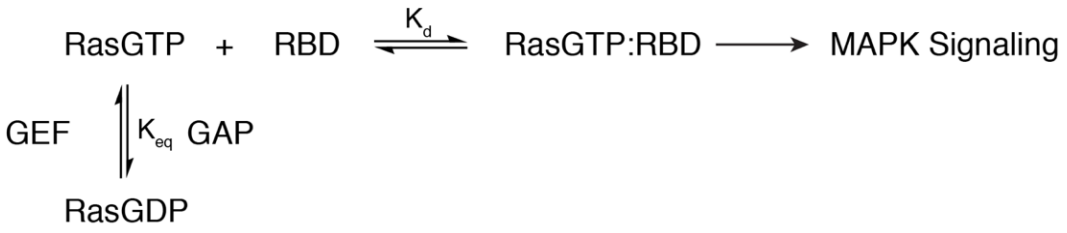

Figure 3. A schematic of the Ras regulatory cycle. The formation of the active Ras•GTP:Raf-RBD complex is driven out of equilibrium by irreversible GTP hydrolysis and exchange by a GAP and GEF, respectively.

$$(1) \qquad \frac{d[RasGTP]}{dt} = k_{GEF}[GEF][RasGDP] - k_{GAP}[GAP][RasGTP] - k_{int}[RasGTP] = -\frac{d[RasGDP]}{dt}$$

$$(2) \qquad [Ras_{tot}] = [RasGTP] + [RasGDP]$$

$[Ras_{tot}]$ is total Ras concentration in the regulatory network, where the concentration of $[RasGTP{:}RBD]$ complex is assumed to be small, and the enzymatic activities of GAP and GEF are much faster than changes in their concentrations. At steady state:

$$(3) \qquad \frac{d[RasGTP]}{dt} = -\frac{d[RasGDP]}{dt} = 0$$

$$(4) \qquad [RasGTP] = \frac{k_{GEF}[GEF][RasGDP]}{k_{GAP}[GAP] + k_{int}} = \frac{k_{ex}}{k_{hyd}}[RasGDP] = K_{eq}[RasGDP]$$

Exchange rate k_{ex} is equal to $k_{GEF}[GEF]$ and k_{hyd} is equal to the sum of GAP-mediated and intrinsic GTP hydrolysis. K_{eq} is the rate constant for the interconversion between Ras•GTP and Ras•GDP. From Equations 2 and 4, the steady state Ras activation level is written as:

$$(5) \qquad \frac{[RasGTP]}{[Ras_{tot}]} = [RasGTP_{SS}] = \frac{1}{1+\frac{1}{K_{eq}}} = \frac{K_{eq}}{1+K_{eq}} = \gamma$$

Cell growth in the bacterial two-hybrid (B2H) is dependent on the production of an antibiotic resistance factor upon the interaction of a Ras variant and a Ras Binding Domain (RBD). The bacterial two-hybrid system reports on the exponential growth of cells containing an individual Ras mutant over time under antibiotic selection. The number of cells N containing a Ras variant x over selection time t is:

$$(6) \qquad N_x(t) = N_{0,x}e^{k_x t}$$

Where $N_{0,x}$ is the initial number of cells before selection containing mutant x and k_x is the growth rate of mutant x. We also define an analogous expression for the number of cells containing wild-type Ras (*WT)* over time:

$$N_{WT}(t) = N_{0,WT}e^{k_{WT}t} \quad (7)$$

Where $N_{0,WT}$ and k_{WT} are the initial number of cells before selection and growth rate of wild-type Ras, respectively. The bacterial two-hybrid selection experiments are performed for a constant, fixed selection time *t*, which can therefore be viewed as a constant scaling factor. The number of Ras variants in the pooled selection experiment are counted with high accuracy with next generation sequencing.

From the ratio of Equations 6 and 7, the number of cells containing a Ras variant *x* can be expressed relative to the number of cells containing wild-type Ras:

$$\frac{N_x(t)}{N_{WT}(t)} = \frac{N_{0,x}(t)}{N_{0,WT}(t)}e^{(k_x - k_{WT})t} \quad (8)$$

The log ratio of the rearranged Equation 8 can be defined as the relative enrichment (ΔE) of a Ras variant *x* in the selected ($N_x(t)$) versus starting ($N_{0,x}(t)$) population of variants, relative to wild-type (Equation 9):

$$\Delta E_x = ln\left[\frac{N_x(t)}{N_{WT}(t)}\right] - ln\left[\frac{N_{0,x}(t)}{N_{0,WT}(t)}\right] = ln\left(e^{(k_x - k_{WT})t}\right) = (k_x - k_{WT})t \quad (9)$$

Mutations that show no functional effect give a relative enrichment in the selected population that is identical to wild-type ($\Delta E_x \approx 0$), and deviations provide a quantitative measure of the functional effect of each mutation on the $[RasGTP_{SS}{:}\,RBD]$ interaction.

Assume that production of the antibiotic resistance factor and growth rate *k* in the bacterial two-hybrid system is proportional to the amount of $[RasGTPP{:}\,RBD]$ complex formed at steady state:

$$k \propto [RasGTP{:}\,RBD] \quad (10)$$

Growth rates are calculated directly from the bacterial two-hybrid sequencing data. Since pre- ($N_{0,x}(t)$) and post-selection ($N_x(t)$) counts for each mutant *x* are not on the same scale, the log ratio of counts defined in Equation 9 leads to unbounded positive and negative growth rates. A scaling factor α is defined as the value of $\frac{N_x(t)}{N_{0,x}(t)}$ that minimally bounds growth rates at 0. In practice, α is set to the average of the bottom 1 percent of ratios stored in a sorted array:

$$Sorted\ array\ of\ \frac{N_x(t)}{N_{0,x}(t)} \rightarrow \begin{bmatrix} \frac{N_x(t)}{N_{0,x}(t)} \\ \vdots \end{bmatrix} \xrightarrow{\substack{\alpha = average\ of \\ bottom\ 1\% \\ of\ ratios}} \begin{bmatrix} \frac{\left(\frac{N_x(t)}{N_{0,x}(t)}\right)}{\alpha} \\ \vdots \end{bmatrix} \quad (11)$$

Thus, the log ratio of counts scaled by α is equal to the growth rate of mutant *x* in the range $[0, \infty)$:

$$(12) \qquad \begin{bmatrix} ln\left(\frac{\left(\frac{N_x(t)}{N_{0,x}(t)}\right)}{\alpha}\right) \\ \vdots \end{bmatrix} = k_x \in [0, \infty)$$

The dependency between growth rate and the amount of $[RasGTP{:}\,RBD]$ complex formed at steady state can be subsequently expressed as the fractional saturation (f_b) of $[RasGTP{:}\,RBD]$ complex:

$$(13) \qquad f_b = \frac{[RasGTP_{SS}{:}RBD]}{[RasGTP_{ss}]+[RasGTP_{ss}{:}RBD]}$$

Furthermore, the apparent dissociation constant for the $[RasGTP{:}\,RBD]$ complex is expressed as:

$$(14) \qquad K_{D,app} = \frac{[RasGTP_{SS}][RBD]}{[RasGTP_{SS}{:}RBD]}$$

Combining Equations 13 and 14 leads to an equivalent expression for fractional saturation:

$$(15) \qquad f_b = \frac{\frac{[RasGTP_{SS}]}{K_{D,app}}}{1+\frac{[RasGTP_{SS}]}{K_{D,app}}}$$

From Equation 10, the amount of $[RasGTP{:}\,RBD]$ formed at steady state is equal to the fractional saturation multiplied by the maximal amount of $[RasGTP{:}\,RBD]$ complex:

$$(16) \qquad [RassGTP{:}\,RBD] = f_b \times [RasGTP{:}\,RBD]_{max} = f_b \times \theta$$

In Equation 16, θ is a constant representing the maximal amount of $[R\text{?}sGTP{:}\,RBD]$ complex. Combining Equation 10 and Equation 16 leads to an expression for growth rate in terms of fractional saturation of $[RasGTP{:}\,RBD]$ complex:

$$(17) \qquad \frac{k}{\theta} = f_b$$

In Equation 17, θ is a normalization that bounds growth rate in the bacterial two-hybrid system between $[0, 1]$. In practice, θ is the average of the top 5 percent of growth rates stored in the array from Equation 12:

$$(18)\qquad \begin{bmatrix} \frac{\left(\frac{N_X(t)}{N_{0,X}(t)}\right)}{\alpha} \\ \vdots \end{bmatrix} \xrightarrow{\theta = average\ of\ top\ 1\%\ of\ growth\ rates} \begin{bmatrix} \frac{ln\left(\frac{\left(\frac{N_{\square}(t)}{N_{0,X}(t)}\right)}{\alpha}\right)}{\theta} \\ \vdots \end{bmatrix} = \begin{bmatrix} f_b \\ \vdots \end{bmatrix} \in [0,1]$$

From Equation 15, the dimensionless quantity $\frac{[RasGTP_{ss}]}{K_{D,app}}$ is equal to:

$$(19)\qquad \frac{[RasGTP_{ss}]}{K_{D,app}} = \frac{f_b}{1-f_b}$$

Equation 19 can be used to calculate $\frac{[RasGTP_{ss}]}{K_{D,app}}$ directly from the array storing fractional saturation values in Equation 18:

$$(20)\qquad \begin{bmatrix} f_b \\ \vdots \end{bmatrix} \xrightarrow[\frac{[RasGTP_{ss}]}{K_{D,app}}]{Calculate} \begin{bmatrix} \frac{f_b}{1-f_b} \\ \vdots \end{bmatrix}$$

Finally, Equation 15 can be expressed in terms of the scaling factor γ from Equation 5:

$$(21)\qquad {f'}_b = \frac{\frac{[RasGTP_{ss}]}{K_{D,app}}\gamma}{1+\frac{[RasGTP_{ss}]}{K_{D,app}}\gamma}$$

The recalculated values of scaled fractional saturation (${f'}_b$) are stored in a new array:

$$(22)\qquad \begin{bmatrix} \frac{f_b}{1-f_b} \\ \vdots \end{bmatrix} \xrightarrow{Scale\ by\ \gamma} \begin{bmatrix} \gamma\left(\frac{f_b}{1-f_b}\right) \\ \vdots \end{bmatrix} \xrightarrow{Calculate\ f'_b} \begin{bmatrix} {f'}_b \\ \vdots \end{bmatrix}$$

Thus, the scaling factor γ effectively modulates $[RasGTP_{ss}]$ through GAP and GEF activity along a "universal" binding isotherm that describes the $[RasGTP_{SS}{:}\,RBD]$ binding equilibrium (Figure 4).

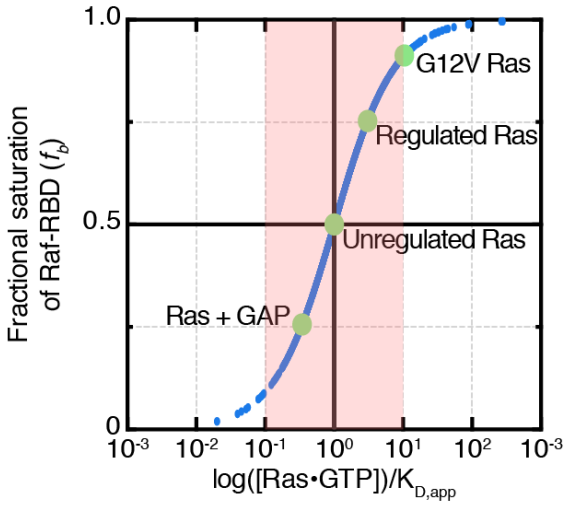

Figure 4. Scaling the bacterial two-hybrid enrichments. A "universal" binding isotherm of concentration of $[RasGTP_{ss}]$ vs. fractional saturation of the $[RasGTP_{SS}{:}\,RBD]$ complex.

In the bacterial two-hybrid system, fractional saturation of the $[RasGTP_{SS}{:}\,RBD]$ complex is high and maximal growth rates are achieved. The latter statement is an assumption partly substantiated by the observation in Figure 5 that growth rates maximally plateau at the inducer concentration chosen for all bacterial two-hybrid experiments.

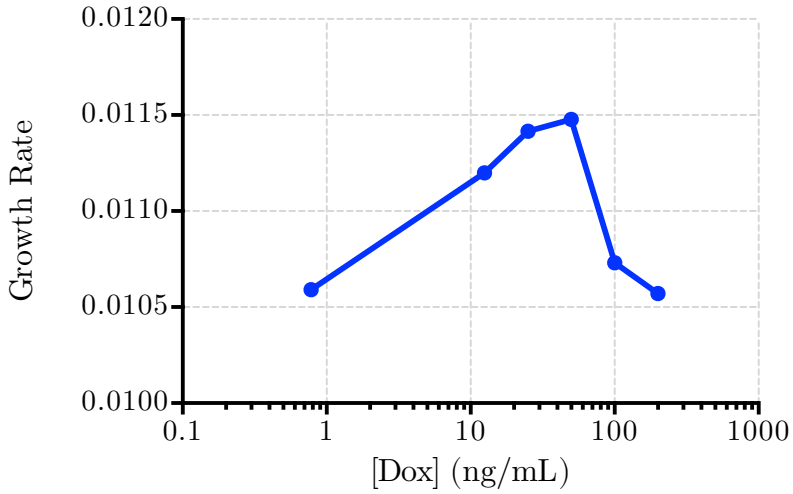

Figure 5. Growth rates in the bacterial two-hybrid system as a function of Ras expression. Growth rate of G12V Ras (blue) in the bacterial two-hybrid system as a function of doxycycline inducer concentration (ng/mL). Growth rate is maximal at the concentration used for all B2H experiments (50 ng/mL).

Fractional saturation is calculated according to Equation 11 through Equation 22 for each mutant x and wild-type *(WT)* and multiplied by the normalization constant θ to calculate a scaled growth rate k':

$$(23) \qquad f'_{b,x} * \theta = k'_x$$

$$(24) \qquad f'_{b,WT} * \theta = k'_{WT}$$

The difference in growth rates for each mutant x and wild-type *(WT)* from Equation 9 is:

$$(25) \qquad \Delta E_x = ln\left[\frac{N_x(t)}{N_{WT}(t)}\right] - ln\left[\frac{N_{0,x}(t)}{N_{0,WT}(t)}\right] = (k'_x - k'_{WT})t = \theta \times (f'_{b,x} - f'_{b,WT})$$

Equation 21 can be substituted into Equation 25 and multiplied through by $K_{D,app}$ to give the final form of relative enrichment:

$$(26) \qquad \Delta E_x = \theta \times \left(\frac{[RasGTP_{SS,x}]\gamma}{K_{D,app;x}+[RasGTP_{SS,x}]\gamma} - \frac{[RasGTP_{SS,WT}]\gamma}{K_{D,app;WT}+[RasGTP_{SS,WT}]\gamma}\right)$$

Thus, we have created a theoretical framework to scale the relative enrichment values derived from sequencing the bacterial two-hybrid screened Ras variants. This scaling enables the quantitative comparison of experiments in which either the full regulatory cycle is expressed, or when components of the regulatory cycle are removed.

The enrichment of a Ras variant is proportional to the RasGTP:Raf-RBD binding affinity

We assume that the fitness of a Ras variant is directly related to the concentration of the Ras•GTP:Raf-RBD complex. ΔE_x values less than or greater than zero indicate a decrease or increase, respectively, in the steady-state concentration of Ras•GTP:Raf-RBD. All ΔE_x values reported here represent averages of two independent experiments, and the data were highly reproducible between duplicate experiments (Figure 6).

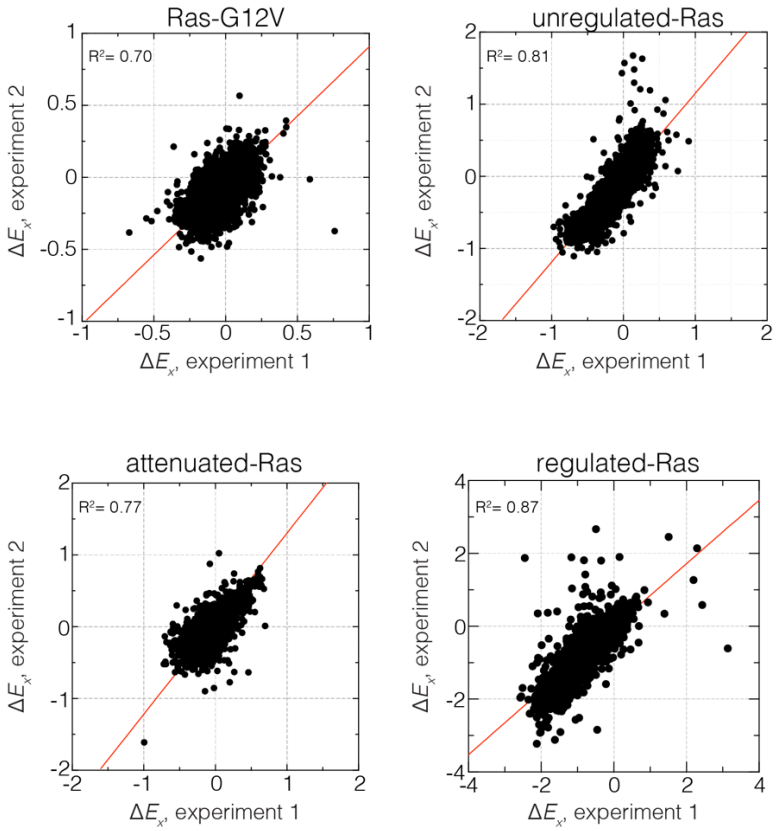

Figure 6. Reproducibility of the bacterial two-hybrid experiments. Correlation of relative enrichment values (ΔE_x) for all four conditions in the presence and absence of regulation for two independent experiments. Reported ΔE_x values are averaged over both experiments.

We compared the relative fitness values, derived from two-hybrid experiments without the GAP or the GEF, to changes in the binding free energy for the interaction of Raf-RBD and Ras•GTP, measured using isothermal titration calorimetry (ITC). The ITC experiments were done for 25 Ras variants loaded with the non-hydrolyzable GTP analog, GMP-PNP. Complete loading of Ras with GMP-PNP was verified by HPLC analysis (Figure 7). The ITC data are consistent with earlier measurements of the affinity of the interaction between Raf-RBD and Ras variants (Kiel et al., 2009; Rudolph et al., 2001; Wohlgemuth et al., 2005). Wild-type Ras binds to Raf-RBD with a dissociation constant (K_D) of ~170 nM, corresponding to a binding free energy ($\Delta G_{binding}$) of -38 kJ/mol. Many mutations in Switch I, such as D33N, T35S, and D38N, lead to a greater than 100-fold increase in the value of K_D relative to wild-type, corresponding to values of $\Delta\Delta G_{binding}$ that range from 3 to 10 kJ/mol. There is a good correlation between the ΔE_x values and the corresponding values of $\Delta\Delta G_{binding}$ for mutations that have strong positive or negative effects on Raf-RBD binding (Figure 8). For mutations with near wild-type binding, there is a greater variance of relative fitness, a result that might arise due to effects on GTP hydrolysis and exchange rates, which do not enter into the thermodynamic measurement.

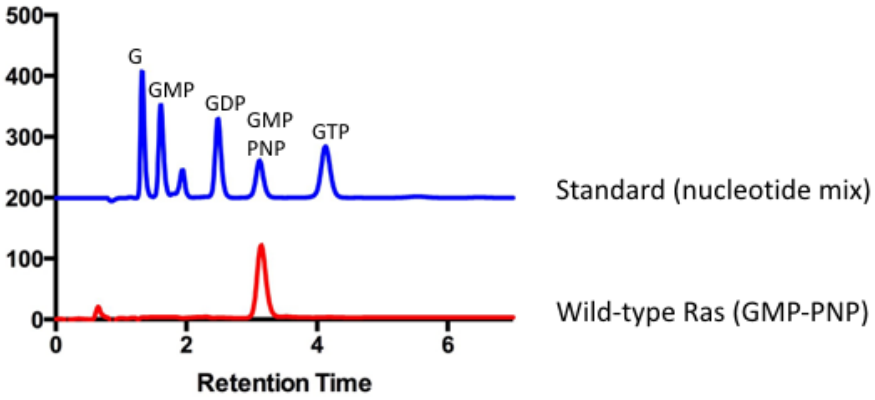

Figure 7. Validating GMP-PNP loading of Ras variants by HPLC. Ion-pair HPLC was used to confirm GMP-PNP loading of Ras variants according to Eberth and Ahmadian (Eberth and Ahmadian, 2009). A mix of nucleotides (GMP, GDP, GMP-PNP, and GTP) was first injected as a reference to assess retention times of each nucleotide. Next, GMP-PNP loaded Ras was injected to confirm that the peak to the bound nucleotide corresponded to the peak obtained from the reference.

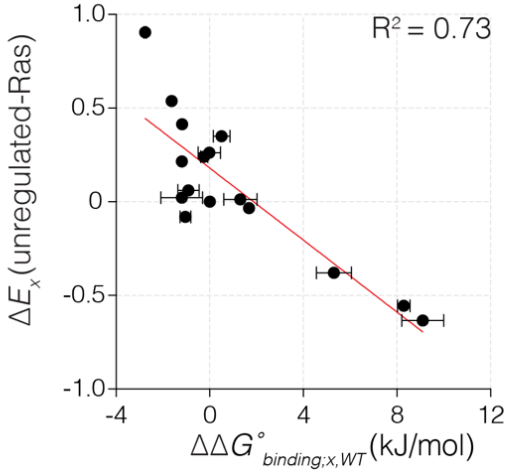

Figure 8. Correlation of enrichment with Ras•GTP:Raf-RBD binding affinity. *In vitro* validation of the bacterial two-hybrid system. The enrichment of individual Ras variants is approximately proportional to the change in Ras•GTP:Raf-RBD binding free energy upon mutation. Binding free energy of individual Ras mutants was measured by isothermal titration calorimetry, where error bars represent the standard deviation from three experiments, and relative enrichment values (ΔE_x) are derived from wild-type Ras binding to Raf-RBD in the unregulated-Ras experiment.

For the saturation mutagenesis experiments, we found that the range of relative enrichment values (ΔE_x) in two of the experiments (attenuated-Ras and Ras-G12V) was

different, when compared to the other two (unregulated-Ras and regulated-Ras). Although all four experiments were carried out using the same nominal conditions, the strength of the antibiotic selection appears to be somewhat variable from experiment to experiment. The differences in effective chloramphenicol concentration are expected to result in linear scaling of the relative enrichment values (ΔE_x) in one dataset with respect to the other. We applied scale factors of 1.2 and 0.7, respectively, to the ΔE_x values for the attenuated-Ras and Ras-G12V datasets. The scale factors were calculated such that the variance of ΔE_x values from the two datasets were equal for the Val 112 reference residue in the hydrophobic core (Figure 9).

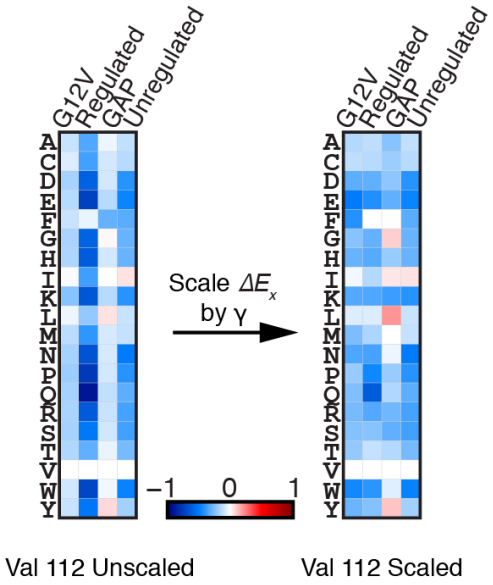

Figure 9. Scaling enrichment data for Val 112. Scaling enrichment values from the bacterial two-hybrid experiments for a reference residue in the hydrophobic core, Val 112.

Summary

This chapter has described the considerations that were taken into account when developing a selection assay for Ras function. Ultimately, I designed a bacterial two-hybrid selection assay that allowed for the fitness of several thousand Ras variants to be determined in a quantitative manner. This assay also allowed for the modular expression of components of the Ras regulatory cycle, enabling the mutational robustness of Ras to be determined in the presence and absence of regulation by auxiliary proteins such as a GAP and a GEF. I created and screened a saturation point-mutagenesis library of Ras variants in this selection assay. The data from the bacterial two-hybrid selection assay was scaled based on a competitive binding model of the Ras regulatory cycle, in order to allow the data to be quantitatively compared across different experimental conditions. In sum, the bacterial two-hybrid system is an effective way to measure the fitness of a saturation point-

mutagenesis library of Ras variants, enabling the assessment of mutational robustness across multiple conditions of selection.

Methods

Bacterial two-hybrid selection assay

We utilized a modified version of a bacterial two-hybrid system in which transcription of chloramphenicol acetyltransferase (pZERM1-CAT plasmid, pRM+ promoter, ampicillin resistant) is made dependent on the binding between each human H-Ras variant (fused to the N-terminal domain of the *E. coli* RNA polymerase α-subunit on the pZA31 plasmid, doxycycline inducible, kanamycin resistant) and the RBD from human C-Raf kinase (fused to the bacteriophage λ-cI DNA binding domain on the pZS22 plasmid, IPTG inducible, trimethoprim resistant). p120 RasGAP was expressed from the pZS22 plasmid, separated from the Raf-RBD by an internal ribosome binding site, and RasGRP1 was expressed from the pZERM1 plasmid on an IPTG-inducible cassette independent of the reporter CAT gene.

Electrocompetent MC4100-Z1 cells containing pZERM1-CAT and pZS22 plasmids were transformed with 100 ng of the pZA31-RNA polymerase α-Ras variant library, recovered for 2 hours in LB media, and grown to saturation overnight in LB media containing 20 μg/mL trimethoprim, 50 μg/mL kanamycin, and 100 μg/mL ampicillin. 20 μL of saturated culture was diluted to an optical density (OD) at 600 nm of 0.001 for a 2-hour outgrowth in 30 mL LB at the same antibiotic concentrations. Cells were diluted to an OD of 0.0001 and induced with 50 ng/mL doxycycline and 100 μM IPTG for 3 hours in 60 mL of LB + antibiotics. After induction, ~50 mL of culture was reserved for deep-sequencing of the pre-selection population. Selection cultures were started with the remainder of the induction cultures at an OD of 0.001 in 60 mL of LB + antibiotics + inducer with 75 μg/mL chloramphenicol for 7 hours, ensuring that the OD remained below 0.1 over the course of the experiment. Both pre- and post-selection cultures were subject to plasmid DNA isolation, PCR amplification of the Ras variant library, and standard preparation for Illumina Mi-Seq paired-end sequencing (Bentley et al., 2008).

Construction of the Ras variant libraries

Comprehensive single-mutant libraries were constructed using oligonucleotide-directed mutagenesis of human H-Ras. To mutate each residue in wild-type and G12V H-Ras (residues 2-166), two mutagenic oligonucleotides (one sense, one antisense) were synthesized (Integrated DNA Technologies), with sequence complementary to 15 base pairs on either side of the targeted codon. For the targeted codon, the oligonucleotides contain NNS codons, in which an N is a mixture of A, T, G, and C and S is a mixture of G and C. This biased randomization results in 32 possible codons with all 20 amino acids sampled – a significant decrease in library complexity, and a significant reduction in stop-codon frequency, without loss of amino acid complexity (McLaughlin et al., 2012; Raman et al., 2016). One round of PCR was carried out with and an antisense or sense oligonucleotide flanking the Ras gene. A second round of PCR was carried out using a combination of the first-round products and both flanking primers, and produced the full-length double-

stranded product containing a single degenerate codon. For the 165 randomized positions, we carried out 165*2 first-round reactions and 165 second-round reactions, for a total of 495 reactions. All reactions yielded single intense bands on agarose gels, and these products were purified using standard gel extraction protocols. PCR product concentrations were measured using Picogreen (ThermoFisher), pooled in an equimolar ratio, gel purified, digested with Bsu36I and XhoI, and ligated with T4 DNA ligase overnight with the manufacturer's recommended protocol into the pZA31 expression vector. The ligation product was purified, eluted into 10 μL ddH2O, and transformed into 100 μL electrocompetent TOP10 *E. coli* cells (Thermo Fisher). We limited our Illumina sequencing to the use of 300 base paired-end reads, rather than longer reads, to ensure high-quality base calls. Since these read lengths cannot cover the full Ras gene, our scanning mutagenesis library was generated as three separate sub-libraries, covering mutations at residues 2-56, 57-111, and 112-166. Each sub-library was independently subject to the bacterial two-hybrid assay, amplification, and sample preparation for MiSeq sequencing. Samples from each sub-library were uniquely barcoded so that all samples from a single experiment were analyzed on a single MiSeq chip.

Protein expression and purification

Human H-Ras variants were cloned into a pET-28b expression vector modified with a 5' His6-tag and a TEV protease site coding sequence, and variants were generated by Quikchange site-directed mutagenesis (Agilent).

Hexahistidine-tagged recombinant human H-Ras variants (residues 1-166) were transformed into 50 μL *Escherichia coli* (BL21 (DE3)) in the pET-28b expression vector. After bacterial growth to an OD of 0.5 in Terrific Broth containing 50 mg/mL kanamycin at 37°C, induction was carried out at 18°C with 0.5 mM IPTG, and growth was continued at 18°C for 18 hours. The bacteria were pelleted by centrifugation and the pellet was either stored at -80°C or used freshly for subsequent purification steps.

The pellet was resuspended in lysis buffer (50 mM Tris, 500 mM NaCl, 20 mM imidazole, 5% glycerol, pH 8.0) containing protease inhibitor cocktail (Roche) and 2 mM β-mercaptoethanol (BME). The cells were lysed by sonication, and the cell debris was removed by centrifugation. The supernatant was applied to a 5 mL nickel column (HisTrap FF, GE Healthcare) and the column was washed with 50 mL lysis buffer and 50 mL low salt buffer (50 mM Tris, 50 mM NaCl, 20 mM imidazole, 5% glycerol, pH 8.0). The nickel column was then connected directly to an ion-exchange chromatography column (HiTrap Q FF, GE Healthcare) and the protein was eluted with elution buffer (50 mM Tris, 50 mM NaCl, 250 mM imidazole, 5% glycerol, pH 8.0). The protein was then eluted from the ion-exchange column in a salt gradient from 50 mM to 1M NaCl to give the partially purified protein, usually in the following buffer: 50 mM Tris, ~230 mM NaCl, 5% glycerol, pH 8.0. 100 μM GDP and 1 mM TCEP was added to the protein solution and the hexahistidine tag was then cleaved overnight using hexahistidine-tagged TEV protease (1 mg TEV per 25 mg crude Ras). The protein solution was run over a nickel column and washed with 35 mL low salt buffer to separate the protein from the cleaved hexahistidine tag and TEV protease. The protein was concentrated to ~40 mg/mL and either frozen at -80°C or immediately

used for GMP-PNP loading. The purification protocol for *S. rosetta* Ras was identical to that of human H-Ras.

Ras was loaded with GMP-PNP using a protocol modified from Eberth and Ahmadian (Eberth and Ahmadian, 2009). A tenfold molar excess of GMP-PNP was added to ~10 mg crude Ras and bound nucleotide was cleaved by 10 U bovine alkaline phosphatase (Sigma) and 1 mM $ZnCl_2$ at 25°C for 1 hour. Free nucleotide and alkaline phosphatase were removed by size exclusion chromatography using a Superdex 75 column (10/300 GL, GE Healthcare) in the following buffer: 40 mM HEPES, 150 mM NaCl, 4 mM $MgCl_2$, 1 mM TCEP, 5% glycerol, pH 7.4. The protein was concentrated to ~40 mg/mL for further use.

The purification protocol for human p120 RasGAP was identical to that for Ras, except during the ion-exchange chromatography step, where the protein was eluted in a salt gradient from 10 to 250 mM NaCl. The purification procedure for human RasGRP1 was also similar to that of Ras, except during the size exclusion chromatography step, where the protein was purified using a Superdex 200 column (26/600, GE Healthcare).

Hexahistidine-tagged, SUMO-fused recombinant human C-Raf-RBD (residues 55-131) was transformed into *Escherichia coli* (BL21 (DE3)) in the pET-SUMO expression vector. Cells were grown to an OD of 0.5 in Terrific Broth containing 50 mg/mL kanamycin at 37°C and induction was carried out at 37°C with 0.5 mM IPTG for 7 hours. Cells were lysed by sonication and affinity purified on a nickel column, using a protocol similar to that for Ras. Protein was eluted from the nickel column in the following elution buffer: 50 mM Tris, 50 mM NaCl, 250 mM imidazole, 5% glycerol, pH 7.5. The protein was immediately desalted into elution buffer without imidazole and SUMO was removed by the addition of ULP1 protease (Thermo Fisher) overnight with 1 mM TCEP. The protein was run over a nickel column to remove hexahistidine-tagged SUMO and the cleaved protein was bound to an ion-exchange chromatography column (HiTrap SP HP, GE Healthcare). The protein was eluted in a salt gradient from 50 to 1 M NaCl and Raf-RBD eluted at 300 mM NaCl. The Raf-RBD was further purified by size exclusion chromatography, concentrated, and stored in the following buffer: 40 mM HEPES, 150 mM NaCl, 1 mM TCEP, 5% glycerol, pH 7.4.

Determining nucleotide loading state of Ras variants by HPLC

HPLC was used to determine the loading state of Ras variants after loading with GMP-PNP. A reverse-phase C18 analytical column (Targa) and a precolumn were equilibrated in a buffer containing 100 mM KH_2PO_4, 10 mM tetrabutylammonium bromide, and 7.5% acetonitrile. A mixture of pure nucleotides, containing GMP, GDP, GTP, and GMP-PNP was first injected as a reference to determine the elution time of each species. Ras variants after GMP-PNP loading were next injected and bound nucleotide retention time was compared to the reference injection to verify that the variants were loaded with GMP-PNP.

Raf-RBD binding assays

Ras•GTP:Raf-RBD binding assays were performed on a MicroCal-autoITC 200 instrument (GE Healthcare). GMP-PNP loaded Ras and Raf-RBD samples were diluted in a binding buffer containing 40 mM HEPES, 150 mM NaCl, 4 mM $MgCl_2$, 1 mM TCEP, 5% glycerol, pH 7.4. Final Ras and Raf-RBD concentrations were typically 100 μM and 10 μM,

respectively. Each set of ITC experiments included three samples: 420 μL of Raf-RBD in the cell, 140 μL Ras in the syringe, and 400 μL of binding buffer. All samples were stored at 4°C before the titration experiments.

The ITC experiments were performed at 25°C. An initial injection of 0.5 μL was excluded from data analysis, followed by 15 injections of 2.6 μL each, separated by 420 s with a filter period of 5 s. The protein solution was stirred at 500 rpm over the course of the titration. Titration curves were fit with a one-site binding model, where the three fitting parameters are stoichiometry, N, association constant, K_A, and binding enthalpy, ΔH. The binding entropy is then calculated using the formula: $\Delta S = (\Delta H + RT^{*}ln(K_A))/T$.

References

Busby, S., and Ebright, R.H. (1994). Promoter structure, promoter recognition, and transcription activation in prokaryotes. Cell *79*, 743–746.

Chuang, E., Barnard, D., Hettich, L., Zhang, X.F., Avruch, J., and Marshall, M.S. (1994). Critical binding and regulatory interactions between Ras and Raf occur through a small, stable N-terminal domain of Raf and specific Ras effector residues. Mol Cell Biol *14*, 5318–5325.

Coyle, S.M., and Lim, W.A. (2016). Mapping the functional versatility and fragility of Ras GTPase signaling circuits through in vitro network reconstitution. Elife *5*.

Dove, S.L., and Hochschild, A. (2004). A bacterial two-hybrid system based on transcription activation. Methods Mol Biol *261*, 231–246.

Dove, S.L., Joung, J.K., and Hochschild, A. (1997). Activation of prokaryotic transcription through arbitrary protein-protein contacts. Nature *386*, 627–630.

Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. Nature *340*, 245–246.

Guzmán, C., Šolman, M., Ligabue, A., Blaževitš, O., Andrade, D.M., Reymond, L., Eggeling, C., and Abankwa, D. (2014). The efficacy of Raf kinase recruitment to the GTPase H-ras depends on H-ras membrane conformer-specific nanoclustering. J Biol Chem *289*, 9519–9533.

Joung, J.K., Ramm, E.I., and Pabo, C.O. (2000). A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. Proc Natl Acad Sci U S A *97*, 7382–7387.

Kiel, C., Filchtinski, D., Spoerner, M., Schreiber, G., Kalbitzer, H.R., and Herrmann, C. (2009). Improved binding of raf to Ras.GDP is correlated with biological activity. J Biol Chem *284*, 31893–31902.

Kuldell, N., and Hochschild, A. (1994). Amino acid substitutions in the -35 recognition motif of sigma 70 that result in defects in phage lambda repressor-stimulated transcription. J Bacteriol *176*, 2991–2998.

Legrain, P., and Selig, L. (2000). Genome-wide protein interaction maps using two-hybrid systems. FEBS Lett *480*, 32–36.

Legrain, P., Wojcik, J., and Gauthier, J.M. (2001). Protein--protein interaction maps: a lead

towards cellular functions. Trends Genet *17*, 346–352.

Li, M., Moyle, H., and Susskind, M.M. (1994). Target of the transcriptional activation function of phage lambda cI protein. Science *263*, 75–77.

Lutz, R., and Bujard, H. (1997). Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. Nucleic Acids Res *25*, 1203–1210.

McLaughlin, R.N., Poelwijk, F.J., Raman, A., Gosal, W.S., and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. Nature *491*, 138–142.

Raman, A.S., White, K.I., and Ranganathan, R. (2016). Origins of allostery and evolvability in proteins: A case study. Cell *166*, 468–480.

Rudolph, M.G., Linnemann, T., Grunewald, P., Wittinghofer, A., Vetter, I.R., and Herrmann, C. (2001). Thermodynamics of Ras/effector and Cdc42/effector interactions probed by isothermal titration calorimetry. J Biol Chem *276*, 23914–23921.

Sass, P., Field, J., Nikawa, J., Toda, T., and Wigler, M. (1986). Cloning and characterization of the high-affinity cAMP phosphodiesterase of Saccharomyces cerevisiae. Proc Natl Acad Sci U S A *83*, 9303–9307.

Wohlgemuth, S., Kiel, C., Krämer, A., Serrano, L., Wittinghofer, F., and Herrmann, C. (2005). Recognizing and defining true Ras binding domains I: biochemical analysis. J Mol Biol *348*, 741–758.

Chapter 3: The mutational robustness of Ras

Introduction

In this chapter, I will describe the mutational sensitivity of Ras in the presence of different components of the Ras regulatory cycle, which effectively serve to change the selective pressure imposed upon the Ras variants in each experiment. I used the bacterial two-hybrid system to carry out four experiments, each of which involved analysis of a complete set of single-site mutations for the relevant Ras protein, expressed together with Raf-RBD. In the first experiment, variants of wild-type Ras were expressed in the presence of the GAP and the GEF (a condition we refer to as "*regulated-Ras*"). In the next experiment, Ras variants were expressed with the GAP, but without the GEF ("*attenuated-Ras*"). In the third experiment, variants of wild-type Ras were expressed without the GAP and the GEF ("*unregulated-Ras*"). Finally, I examined the effects of an oncogenic mutation by expressing variants of Ras-G12V without the GAP and the GEF ("*Ras-G12V*").

Ras exhibits a global sensitivity to mutation when regulated by a GAP and a GEF

The values of the relative fitness, ΔE_x, for every possible substitution in the sequence of Ras are shown in Figure 1A for the regulated-Ras experiment. The mode of the distribution is centered at a negative value of ΔE_x, with most mutations resulting in a slight loss of function, with relatively few mutations resulting in a near-neutral effect or in a gain of function (Figure 1B). This is qualitatively different from the results of saturation point-mutagenesis of binding proteins such as a PDZ domain, and enzymes such as TEM-1 β-lactamase, where the distributions of fitness effects show a primary mode centered close to zero (McLaughlin et al., 2012; Stiffler et al., 2015). Our results suggest that regulation by the GAP and the GEF constrains Ras in a state where variations almost anywhere in the protein cause reduction of fitness. In principle, even a small loss of function can lead to effective purifying selection in evolving populations, and so the global constraint imposed by the regulators might explain the high degree of conservation in Ras sequences.

To identify the residues that are particularly sensitive to mutation, we averaged the fitness values for all the mutations for each residue. The distribution of residue-averaged ΔE_x values was fit to a normal distribution, and we define residues as highly sensitive to mutation if the corresponding residue-averaged ΔE_x value deviates more than 1σ from the mean value (Figure 1C). The highly sensitive sites are displayed on the structure of Ras in Figure 1D. These sites cluster within the hydrophobic core, and around the nucleotide binding site and the regions of Switch I and Switch II that interact with Raf-RBD, the GAP and the GEF. Most sites on the surface of Ras are less sensitive to mutation, except where surface residues are involved in interactions that are likely to stabilize the structure, as shown for an ion-pairing network involving sensitive residues that are surface exposed (Figure 1E).

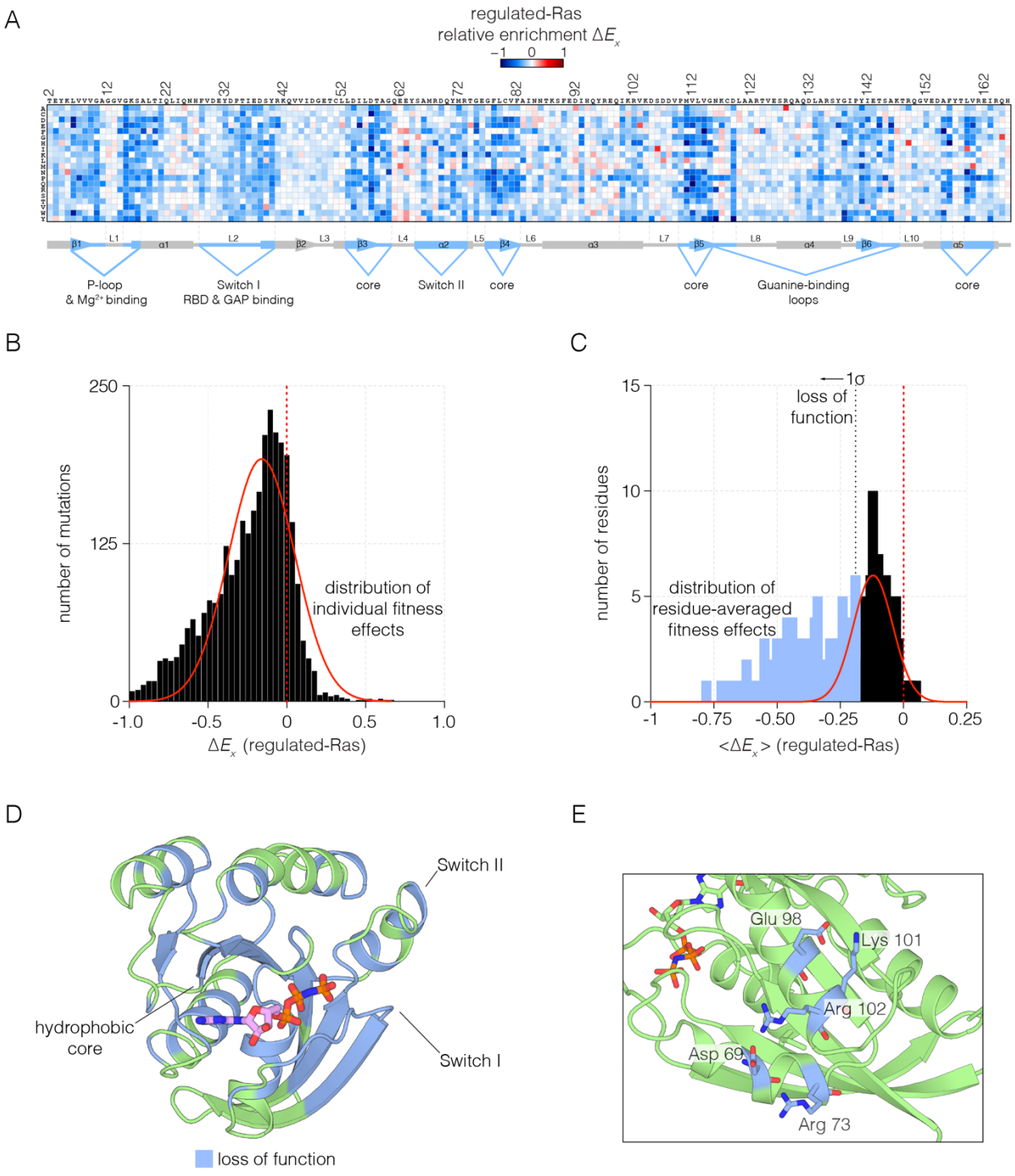

Figure 1. Mutational tolerance of Ras in the regulated-Ras experiment. A. The results of the regulated-Ras experiment are shown in the form of a 165x20 matrix. Each row of the matrix represents one of the 20 amino acids, and each column shows one of the residues of Ras, from 2 to 166. Each entry in the matrix represents, in color-coded form, the value of the relative enrichment for the corresponding mutation ΔE_x. All data are normalized to the wild type Ras reference sequence,

which has a relative enrichment value of zero. The numerical values of ΔE_x are provided in the supplementary data. **B-C.** Distribution of individual fitness effects (ΔE_x; left) and residue-averaged fitness effects (<ΔE_x>; right). Residues with a significant loss of function effect on Ras (<1σ from the mean) are indicated on the residue-averaged histogram. **D.** Mapping the residues that lead to a significant loss of function onto the tertiary structure of Ras. These positions include the hydrophobic core, as well as residues involved in GDP/GTP and Raf-RBD binding. **E.** Additional sites of mutational sensitivity include surface residues involved in ion-pairing networks that stabilize the GTPase fold.

As expected, mutations in Switch I that have been identified previously to impair Raf-RBD binding also compromise Ras function in our assay (Spoerner et al., 2001). The Ras•GTP:Raf-RBD binding interface includes interactions made by Glu 31, Asp 33, Glu 37, Asp 38, and Ser 39 in H-Ras. Nearly all mutations to these residues result in a loss of function, indicative of a highly constrained binding interface (Figure 2A). We note that since Raf-RBD and GAP have an overlapping interface, mutations in Switch I also impede GAP binding (Scheffzek et al., 1997; Smith and Ikura, 2014).

Many of the residues in Ras that coordinate the nucleotide, such as Lys 16, Thr 35, and Asp 119, are highly sensitive to mutation. Some of these mutations may cause protein instability, such as those at Asp 119. Interestingly, a few residues that interact with the nucleotide are not very sensitive to mutation, such as Lys 117 and Lys 147, and many mutations at these sites lead to a gain of Ras function. This may be due to the fact that the affinity of Ras for GTP is much higher than is required to saturate the active site (Bourne et al., 1991). Thus, residues that have only a modest effect on affinity do not alter the extent of GTP loading appreciably, given the high cellular concentrations of GTP (Bennett et al., 2009) (Figure 2B). Mutations that reduce GTP affinity may lead to fast-cycling Ras variants, which are able to more quickly replace GDP by GTP.

The binding and activity of the GEF (RasGRP1) imposes an additional constraint on the Ras sequence. One component of the interface between Ras and GEFs involves Switch II (Boriack-Sjodin et al., 1998; Iwig et al., 2013), where mutations to residues, including Asp 69, Tyr 71 and Arg 73, impair GEF binding. The binding footprint of the GEF extends beyond Switch II to residues 101-104, and mutations to residues, such as Arg 102, in this region cause a similar loss in Ras function. Not all mutations to residues in the GEF binding interface cause a loss of function. For example, Tyr 64 is crucial for GEF binding, but this residue is located in Switch II, near Gln 61, which is important for GTP hydrolysis. Thus, mutation of Tyr 64 and other residues in its vicinity leads to a complex pattern of mutational sensitivity in the regulated-Ras experiment (Figure 2C).

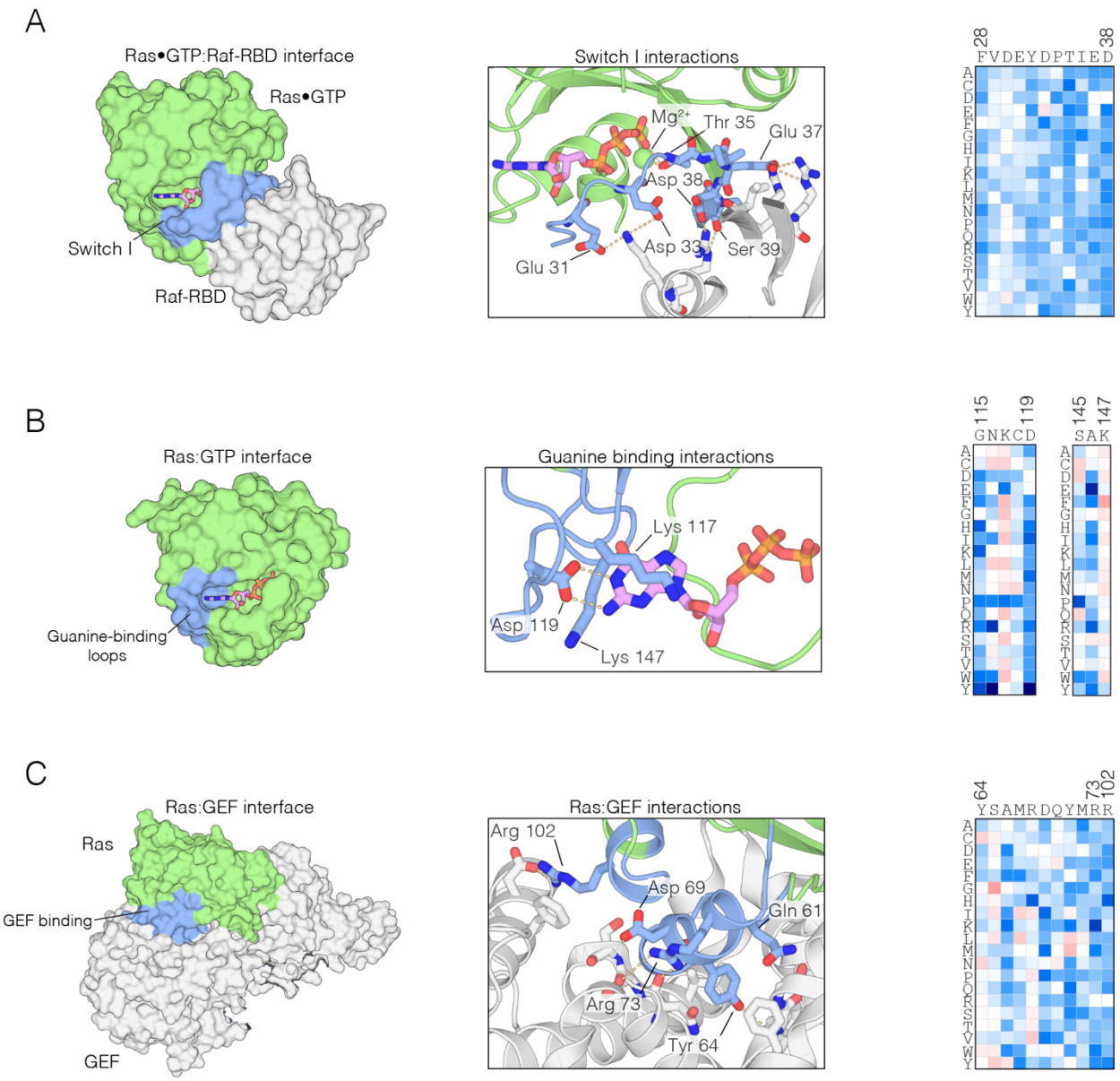

Figure 2. Mutational sensitivity of structural elements involved in nucleotide and effector binding. A. The Ras•GTP:Raf-RBD complex. Mutations to Ras that affect Raf-RBD binding result in a loss-of-function. Critical interactions at the interface (blue) involving residues Glu 31, Asp 33, Thr 35, Glu 37, Asp 38, and Ser 39 are sensitive to mutation, where nearly every possible substitution is disfavored. Loss of function mutations at Thr 35 impact Mg^{2+} binding. The mutational data shown in this and the other two panels are for the regulated-Ras experiment **B.** The Ras:GTP complex. Mutations to Ras that affect nucleotide binding result in a loss-of-function. Asp 119 is sensitive to mutation and critical for recognition of the guanine base, whereas Lys 117 and Lys 147 are auxiliary residues that are less sensitive to mutation. **C.** The Ras-GEF complex. Mutations to Ras that perturb GEF binding result in a loss-of-function, since Ras is predominantly in the GDP-bound form in

the bacterial two-hybrid system when a GAP is co-expressed. Critical interactions at the interface (blue) are sensitive to mutation, although some residues, such as Tyr 64, display a complex pattern of mutational sensitivity due proximity to Gln 61 which is responsible for GTP hydrolysis.

We tested the effect of a selected set mutations on GAP activity by using a fluorescent protein sensor for the detection of inorganic phosphate released by GTP hydrolysis, using purified mutant Ras proteins (Brune et al., 1994). Mutations in Switch I lowered GAP-catalyzed GTP hydrolysis rates, as expected (Figure 3A). We also tested the effect of Ras mutations on GEF activity, using an assay in which the release of labeled GTP or GDP from Ras is monitored as a function of time (Eberth and Ahmadian, 2009). Mutations in Switch II that altered fitness also affected GEF stimulated GDP exchange rates (Figure 3B). Raw data are available online at jkweb.berkeley.edu.

We used a yeast assay for the activity of human H-Ras to verify that a selected subset of the mutations that lead to decreased fitness in the two-hybrid assay do indeed lead to a loss of function. Transformation and expression of human H-Ras in the yeast S. cerevisiae renders the cells sensitive to heat shock. Mutations that affect the ability of Ras to bind nucleotide reduce Ras activity, restoring protection from heat shock, and consequently result in increased cell growth after heat shock (Sass et al., 1986). Using this assay, we verified that several mutations in the nucleotide binding regions that lead to decreased fitness also lead to increased yeast growth after heat shock. These results confirm that these mutations compromise the function of Ras (Figure 3C).

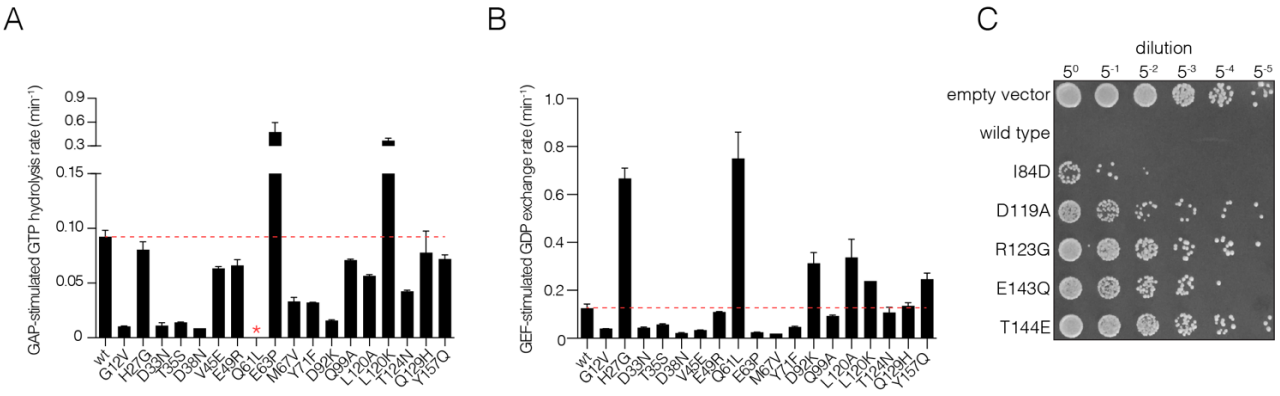

Figure 3. Validation of mutational effects through *in vitro* biochemical measurements of GAP- and GEF- stimulated nucleotide hydrolysis and exchange rates and a yeast growth assay. A. GAP-stimulated GTP hydrolysis rates for a panel of Ras mutants, as measured by a fluorescent sensor for inorganic phosphate release upon hydrolysis of GTP to GDP. Mutations in Switch I (residues 28-40) directly affect GAP binding and lead to reduced GTP hydrolysis rate. **B.** GEF-stimulated GDP exchange rates for a panel of Ras mutants, as measured by fluorescent-nucleotide release. Mutants in Switch II (residues 59-76) directly affect GEF binding and lead to reduced GDP release rates. **C.** *S. cerevisiae* Ras assay for mutations that result in a loss-of-function in the bacterial two-hybrid system. Yeast

cells are transformed with Ras mutants and are assayed for growth after heat shock. Colony growth is measured upon plating five-fold serial dilutions of yeast cultures after heat shock for various Ras mutations. Wild-type Ras renders the yeast cells highly sensitive to heat shock. Mutants that destabilize Ras by impairing nucleotide binding lessens sensitivity to heat shock and result in increased cell growth.

In summary, in the context of its natural regulatory network, Ras displays global mutational sensitivity at nearly all positions. The effects include, but extend well beyond, regions known to mediate the basic biochemical activities of Ras, a result that highlights the value of deep mutational scanning in uncovering unexpected functional constraints in a mechanistically unbiased manner.

The presence of the GAP without the GEF leads to strong activation by oncogenic mutations

When Ras mutants are expressed in the presence of the GAP, but without the GEF (the attenuated-Ras experiment), we observe greater sensitivity to mutation across most of the structure (Figure 4A). It is likely that the presence of the GAP, without the compensating effect of the GEF, reduces the concentration of Ras•GTP to well below the dissociation constant for binding to Raf-RBD. In this situation, the effects of mutations that have a small deleterious effect are amplified, as a small extent of binding is further weakened by lower Ras•GTP concentration. In addition, the presence of the GAP without the GEF unmasks several strong gain-of-function mutations (Figure 4B). These mutations are much more strongly activating than in the regulated-Ras experiment, and nearly any substitution to residues involved in guanine or phosphate-binding and GTP hydrolysis lead to a gain of function, presumably because these interfere with the functioning of the GAP (Figure 4C).

In human cancers where Ras is mutated, 98% of the mutations occur at Gly 12, Gly 13 and Gln 61 (Hobbs et al., 2016; Prior et al., 2012). The first two residues are in the phosphate-binding P loop of Ras, and Gln 61 plays a role in the catalysis of GTP hydrolysis, by positioning a water molecule for attack on the terminal phosphate group (Pai et al., 1990; Scheffzek et al., 1997). Under the conditions of the regulated-Ras experiment, mutations of Gly 12 and Gly 13 do not result in strong activation, and there is only a mild activating effect for mutations of Gln 61. In contrast, for the attenuated-Ras experiment, mutations at all three oncogenic sites are strongly activating (Figure 4D). Mutations of Gly 12 result in decreased GAP efficiency, and mutations at positions 13 and 61 affect GTP hydrolysis as well as the intrinsic rate of GDP-GTP exchange, as shown previously (Mazhab-Jafari et al., 2015).

There are five other sites in Ras where almost any substitution leads to strong activation in the presence of the GAP without the GEF: Tyr 64, Lys 117, Leu 120, Ser 145 and Lys 147. Tyr 64, located near Gln 61, is located at the interface with the GAP (Scheffzek et al., 1997) (Figure 4E). The other four residues are involved in binding the guanine base of GTP, and mutation at these sites increases the intrinsic rate of nucleotide exchange (see below). The COSMIC database of somatic mutations in cancer includes mutations at some of these sites in Ras, such as Lys 117 and Leu 120, although these mutations are rare compared to the three principal sites of oncogenic mutation (Bamford et al., 2004).

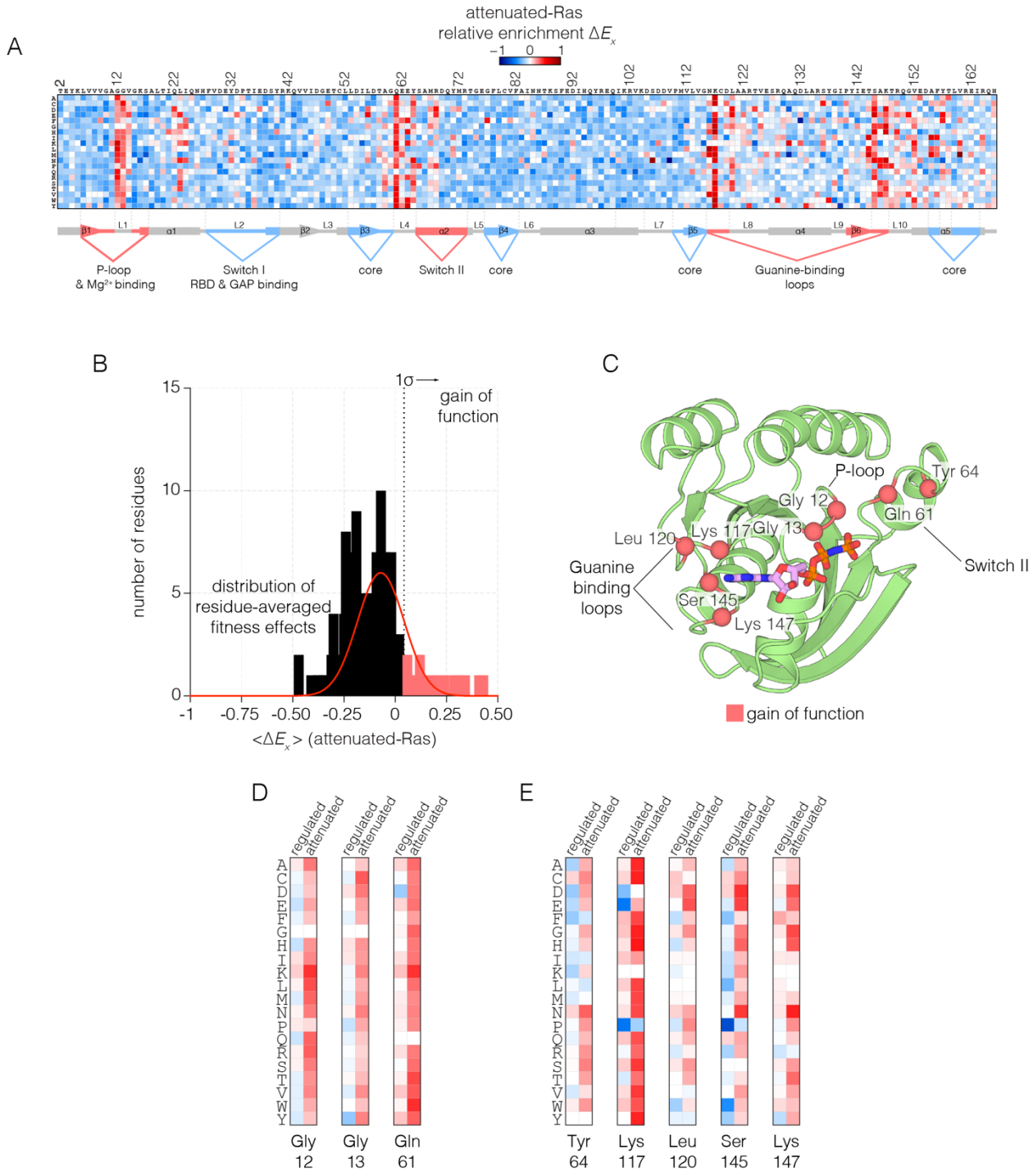

Figure 4. Mutational tolerance of Ras in the attenuated-Ras experiment. A. Relative enrichment values (ΔE_x) are shown in matrix form as in Figure 1a, for the attenuated-Ras experiment. In this experiment, Ras is expressed without the GEF, and shows a muted pattern of mutational sensitivity. Mutations at residues known to be mutated in human cancer (e.g. Gly 12, Gly 13, Gln 61, Lys 117) show a strong gain of function in this context. **B.** Distribution of residue-averaged relative

enrichment values. Residues with a significant gain of function effect on Ras (>1σ from the mean) are indicated on the histogram. **C.** Mapping the residues that lead to a gain of function onto the tertiary structure of Ras. These positions span the P-loop, Switch II, and guanine binding loops of Ras and are directly involved in nucleotide coordination and hydrolysis. **D-E.** Comparison of relative enrichment values for mutations at selected residues in the regulated-Ras and attenuated-Ras experiments. Substitutions at residues that commonly mutated in cancer, such as Gly 12, Gly 13, and Gln 61, are exclusively gain of function in the attenuated-Ras experiment.

Hotspots of activating mutations in unregulated wild-type Ras

In our bacterial two-hybrid system, we have chosen experimental conditions such that wild-type Ras, without the GAP and the GEF, exhibits measurable growth in the presence of antibiotic selection (the unregulated-Ras experiment). Under these conditions, the balance of activation is set by the intrinsic rates of GTP hydrolysis and nucleotide exchange, and the expression level of Ras. The distribution of fitness effects for unregulated-Ras is remarkably distinct from that for regulated-Ras. Instead of global purifying selection (Figure 1), the distribution now displays a primary mode centered around zero, meaning that most mutations are near-neutral (see Figure 1B). This result demonstrates that in the absence of its local regulatory network, Ras behaves somewhat like other proteins studied through deep mutational scanning – generally tolerant to mutation, with a subset of positions showing significant functional effects.

An unexpected result is that for unregulated-Ras, there are many mutations, distributed throughout the structure, that result in *increased* fitness with respect to wild-type Ras (Figure 5A). For the majority of these residues, almost *any* mutation that alters the wild-type residue increases the activity of Ras in the unregulated-Ras experiment. We refer to residues where multiple mutations lead to activation as hotspots, and a subset of these residues were identified in the attenuated-Ras experiment (see Figure 4A).

Many of the hotspots of activation in the unregulated-Ras experiment involve residues for which activating mutations have not been previously identified in human Ras, and which do not lead to activation in the presence of the regulators (Figure 5B shows a scatter plot comparing the fitness effects of all mutations in the regulated-Ras and unregulated-Ras experiments). These conditionally-activating mutations are located further away from the active site than the mutations that activate in both experiments (Figure 5B). One example is Tyr 157, a residue located in helix α5, about 20 Å from the nucleotide-binding site and the switch regions. As can be seen in Figure 5A, replacement of Tyr 157 by 17 of the 19 possible alternatives results in increased function in the unregulated-Ras experiment. Another example is Gln 99, located in helix α3, on the side of Ras opposite to the location of Tyr 157, and also about 20 Å from the GTP-binding site. Replacement of Gln 99 by any residue, other than glycine and proline, leads to increased fitness in the unregulated-Ras experiment.

These hotspots of activation most likely correspond to residues that in the wild-type sequence play a critical role in maintaining the intrinsic rate of GTP hydrolysis, or

suppressing the intrinsic rate of GDP release. We tested the effects on intrinsic GTP hydrolysis and nucleotide release for a subset of these mutations, using purified proteins and *in vitro* assays. Some of the mutations decrease the rate of hydrolysis, as seen for Q61L, E63P, and Q99A. Mutations at the base of Switch I (H27G), in helix α5 (Y157A), and in distal regions of the guanine binding loops (L120A), increase the rate of intrinsic nucleotide release (Figure 6).

Taken together, mutational scanning in regulated and unregulated Ras reveal a profound context dependence of the distribution of fitness effects. In the context of its regulatory network, Ras is subject to global purifying selection. However, without regulation, Ras displays a distinct character – general tolerance to mutation and the capacity for activating mutations that act allosterically to destabilize the inactive state.

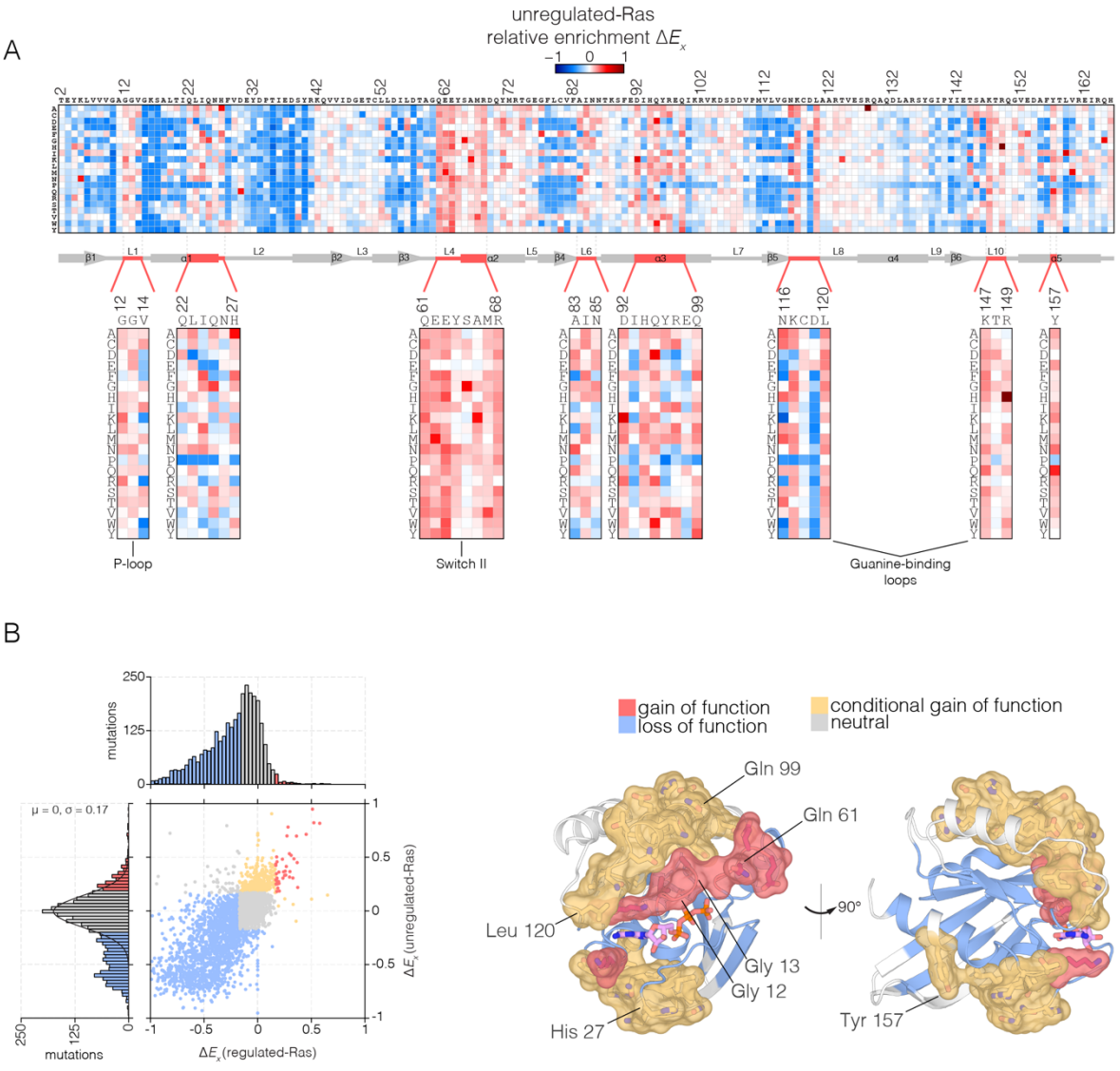

Figure 5. Mutational tolerance of Ras in the unregulated-Ras experiment. A. Relative enrichment values (ΔE_x) are shown in matrix form as in Figure 1a, for the

unregulated-Ras experiment. In this experiment, Ras is expressed without the GAP and the GEF, and reveals hotspots of activating mutations. Mutations at residues known to be mutated in human cancer (e.g. Gly 12, Gly 13, Gln 61, Lys 117) show a strong gain of function in this context. **B.** A scatter plot of relative enrichment values (ΔE_x) for the regulated-Ras and unregulated-Ras experiments. The distribution of relative enrichments values for each experiment are also shown. Loss of function mutations are shown in blue, and neutral mutations are shown in grey. Gain of function mutations in both experiments are shown in red, and mutations that are only gain of function in the unregulated-Ras experiment, but are neutral in the regulated-Ras experiment, are shown in yellow (conditional gain of function). **C.** The spatial distribution of gain of function (red) and conditional gain of function (yellow) mutations on the three-dimensional structure of Ras. Residues that contain a majority of gain of function mutations from the scatter plot in (b) are colored red, and residues that contained a majority of conditional gain of function mutations are colored in yellow.

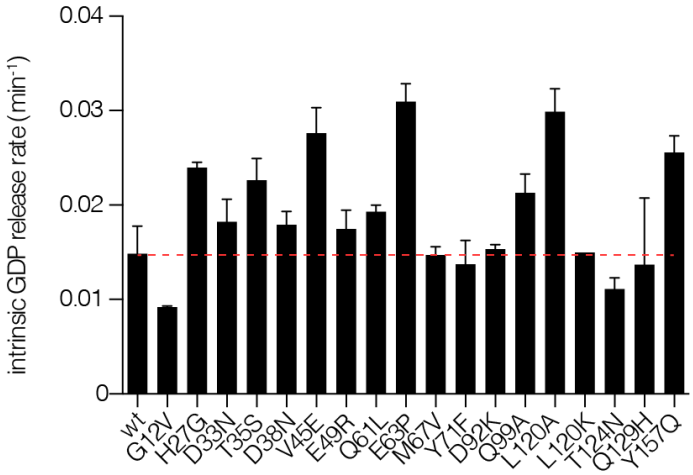

Figure 6. Intrinsic nucleotide release rates. Intrinsic GDP release rates for a panel of Ras mutants, as measured by fluorescent nucleotide release. H27G, L120A, and Y157A mutants lead to increased GDP release rates.

The oncogenic G12V mutation attenuates the effects of mutations that activate Ras in the wild-type background

We carried out a saturation mutagenesis screen in the G12V background, without a GAP or a GEF (the Ras-G12V experiment), the results of which are shown in Figure 7A. The introduction of the G12V mutation into Ras markedly reduces the strongly activating effects of the hotspot mutations seen in the unregulated-Ras experiment (Figure 7B). This effect is seen for both residues that increase nucleotide exchange when mutated (such as Lys 117, Leu 120 and Lys 147; see Figure 7C), and residues that decrease the hydrolysis rate when mutated (such as Gly 13, Gln 61, and Gln 99) (Baker et al., 2013; Janakiraman et al., 2010).

The oncogenic G12V mutation is only mildly activating in the unregulated Ras experiment (see Figure 5A). The G12V mutation has two opposing effects on the activity of unregulated Ras. One is a decrease in the intrinsic rate of GTP hydrolysis with respect to wild type, which is an activating effect in terms of Ras signaling. The phosphate-detection assay we use to measure GTP hydrolysis is not very sensitive when the rate of hydrolysis is low, and our data indicate only small reduction in hydrolysis rate for G12V versus wild-type. A much more sensitive NMR assay demonstrated, however, that the G12V mutation reduces the intrinsic hydrolysis rate substantially, to 10% of the wild-type rate (Mazhab-Jafari et al., 2015). The reduction in hydrolysis rate is counter-balanced by a reduction in the intrinsic rate of nucleotide exchange, by about 30% relative to wild-type in our assay (see Figure 6), with a similar reduction seen by NMR (Mazhab-Jafari et al., 2015).

Since Ras-G12V has a suppressed rate of GTP hydrolysis, we presume that mutations that further suppress this rate do not have a discernible effect on activity. Likewise, the decreased nucleotide exchange rate for Ras-G12V reduces the effect of mutations that tend to increase this rate. The epistasis between the effects of the oncogenic mutations and other mutations distributed across the structure of Ras will be an interesting line for future inquiry.

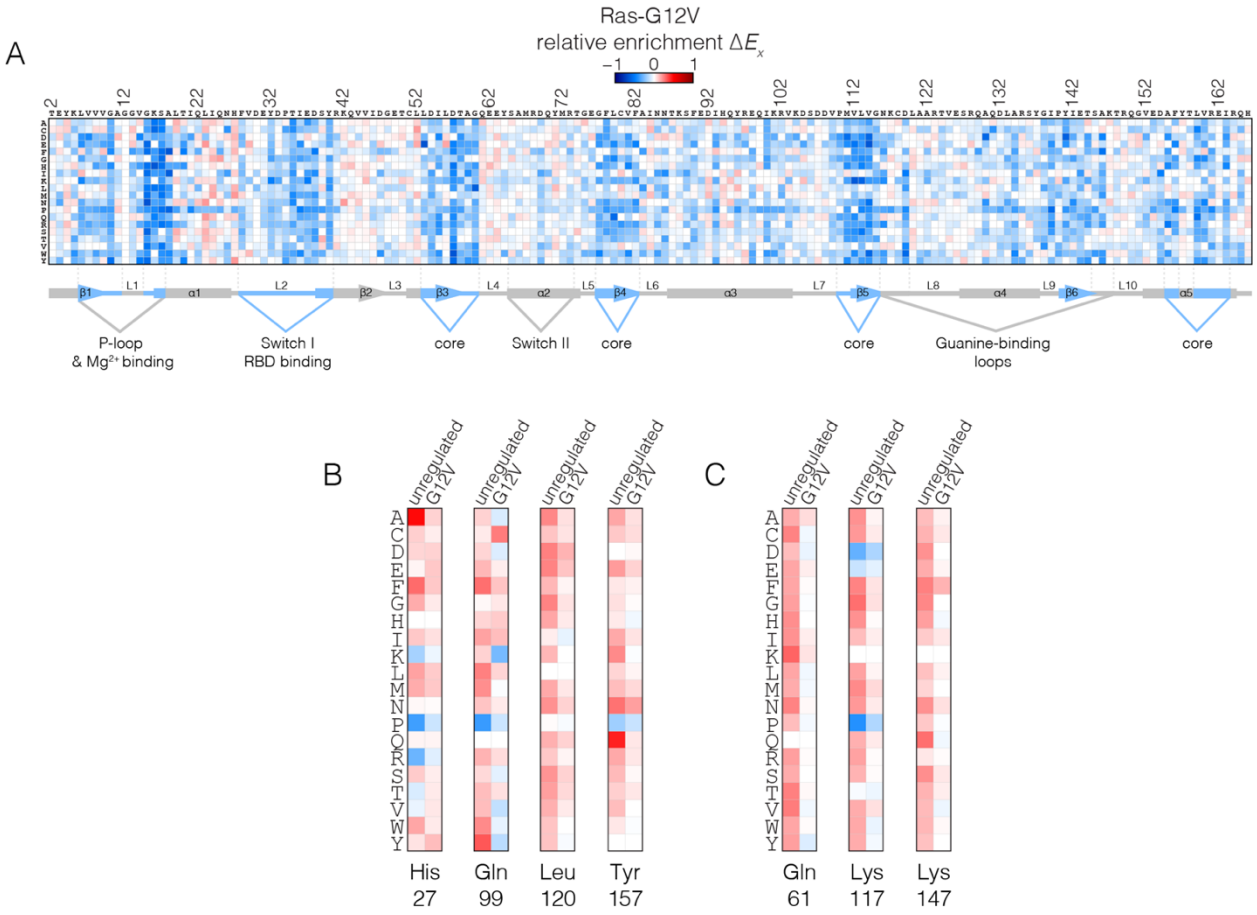

Figure 7. Mutational tolerance of Ras in the Ras-G12V experiment. A. Relative enrichment values (ΔE_x) are shown in matrix form as in Figure 1a, for the Ras-G12V experiment, in the absence of the GAP and the GEF. In this experiment, Ras shows a muted pattern of mutational sensitivity when compared to the wild-type unregulated-Ras experiment. Mutations at residues known to be mutated in human cancer (e.g. Gly 13, Gln 61, Lys 117) do not show a strong gain of function in this context. Mutations at Gly 12 are not included in the data. **B-C.** Comparison of relative enrichment values for mutations at selected residues in the unregulated-Ras and Ras-G12V experiments. Substitutions at hotspot residues and residues that are commonly mutated in cancer, such as Gln 61, Lys 117, and Lys 147, result in a gain of function in the unregulated-Ras experiment, but have an attenuated effect in the Ras-G12V experiment.

The mutational robustness of Ras in mammalian cells

A recent development in this line of investigation has been to assess the mutational robustness of Ras in mammalian cells. Ras is an endogenous signaling molecule in mammalian cells, and is of course crucial for cell proliferation in response to extracellular mitogenic cues. In collaboration with Neel Shah, we have adapted an existing mammalian cell selection platform to introduce the saturation point-mutagenesis libraries of Ras into mammalian cells and assess its global sensitivity to mutation. The selection system we have used is the Ba/F3 cell line, which is a cell line whose growth is dependent on the presence of the cytokine interleukin-3. Many studies over the past few decades have shown that oncogenes are capable of transforming Ba/F3 cells, leading to cell growth that is independent of IL3 addition (Palacios and Steinmetz, 1985). This feature of the Ba/F3 system enables a selection that is conceptually analogous to the bacterial selection system described earlier in this work, and particularly allows for the identification of transforming variants in many protein systems of interest such as BCR/Abl and Ras (Daley and Baltimore, 1988; Hoover et al., 2001; Mathey-Prevot et al., 1986). We have used this assay to monitor the transforming capacity of Ras variants, where the basic premise of the selection involves monitoring Ba/F3 growth in the absence of IL3, which should correspond to transforming Ras variants driving cell growth. The results of the screen are shown below in Figure 8:

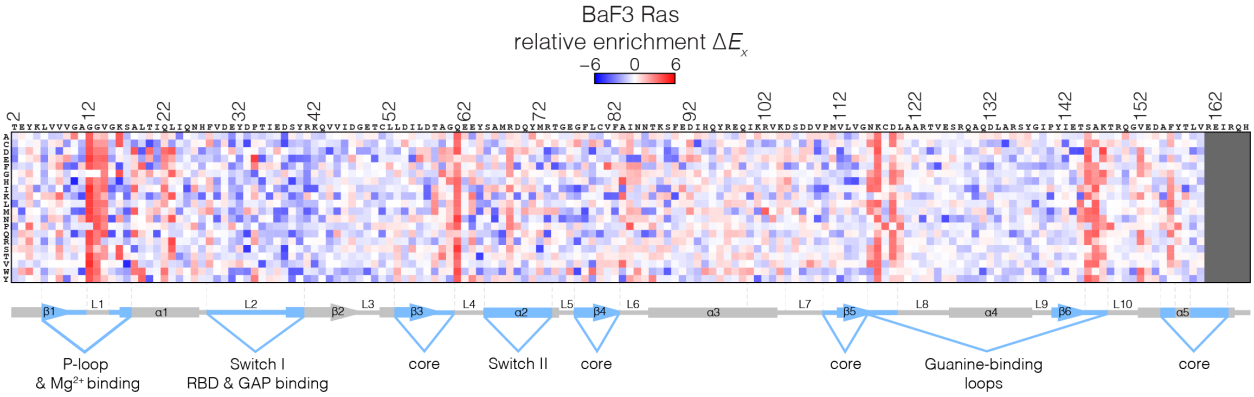

Figure 8. Mutational tolerance of Ras in the Ba/F3 experiment. Relative enrichment values (ΔE_x) are shown in matrix form as in Figure 1a, for the unregulated-Ras experiment. In this experiment, Ras is expressed in Ba/F3 cells which express all known components of MAP Kinase signaling, and reveals hotspots of activating mutations. Mutations at residues known to be mutated in human cancer (e.g. Gly 12, Gly 13, Gln 61, Lys 117) show a strong gain of function in this context. The last five residues are not shown, since the cloning strategy removed these residues from the selection experiment.

The data from this experiment show a strong enrichment of classically oncogenic mutations at Gly 12, Gly 13, Gln 61, Lys 117, and Lys 147. In human cancers where Ras is mutated, 98% of the mutations occur at Gly 12, Gly 13 and Gln 61 (Hobbs et al., 2016; Prior et al., 2012). Again, these residues play a role in the catalysis of GTP hydrolysis (Pai et al., 1990; Scheffzek et al., 1997), and also result in decreased GAP efficiency as shown previously (Mazhab-Jafari et al., 2015). The other sites where almost any substitution leads to strong activation, Tyr 64, Lys 117, Leu 120, Ser 145 and Lys 147, increase the intrinsic rate of nucleotide exchange as determined in earlier screens of the attenuated-Ras experiment and by *in vitro* measurements, and also correspond to cancer mutations in the COSMIC database (Bamford et al., 2004). Interestingly, residues that corresponded to activating hotspots in the unregulated-Ras experiment, such as His 27, Gln 99, and Tyr 157, are not highly enriched in the Ba/F3 screen. This discrepancy might arise from the intrinsic "set point" of the Ba/F3 cells, as these cells might have a basal level of GAP activity that could affect the Ras•GTP to Ras•GDP ratio. In support of this notion, the correlation between the residue-averaged enrichments between the Ba/F3 experiment and bacterial experiments suggests a basal GAP activity, as the attenuated-Ras experiment is most similar to the Ba/F3 experiment (Figure 9). These results are consistent with the resting state of mammalian cells where the MAPK pathway has not been stimulated. It would be interesting to treat these cells with GAP inhibitors or to activate MAPK signaling through the addition of mitogens such as epidermal growth factor.

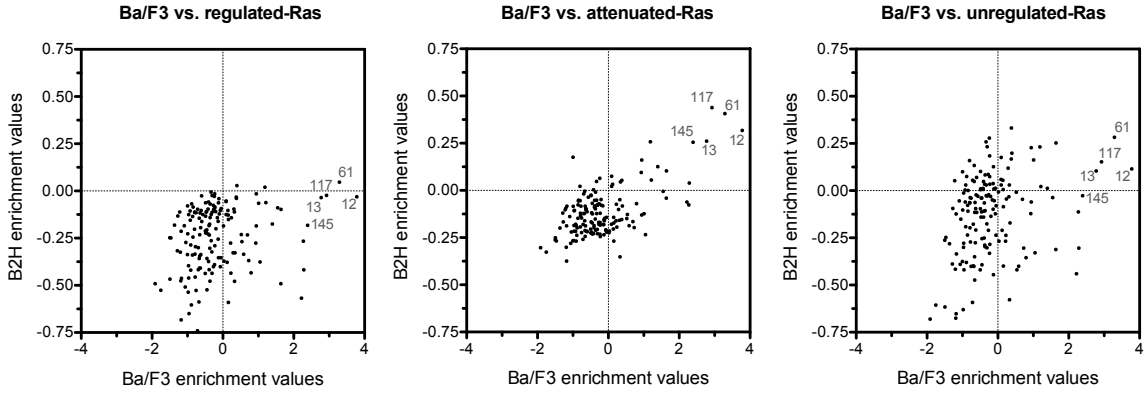

Figure 9. Correlation between Ba/F3 and bacterial Ras screens. Residue-averaged enrichment values for Ba/F3 experiment plotted against values from bacterial two-hybrid screens for the regulated-, attenuated-, and unregulated-Ras

experiments. The correlation appears to be strongest for the attenuated-Ras experiment, suggesting that the Ba/F3 cells represent a state of basal GAP activity without GEF stimulation.

Summary

In this chapter, I described the global sensitivity to mutation exhibited in Ras. The mutational robustness of Ras is context dependent, and is largely affected by the presence of its regulators, the GAP and the GEF. When both regulators are present, Ras exhibits a global sensitivity to mutation, where most mutations throughout the protein sequence lead to a loss of Ras function. When the GEF is removed, and only the GAP is present, many mutations at residues that are commonly mutated in human cancers lead to a dramatic gain of function. When both the GAP and GEF are removed, many gain-of-function mutations are observed, again at residues that are mutated in cancers but also at residues that appear to act as hotspots of Ras activation. Indeed, mutations at these residues lead to a decrease in intrinsic GTP hydrolysis rate or an increase in intrinsic nucleotide exchange rate. Furthermore, these residues are one shell away from active site residues, and form a contiguous allosteric network around the nucleotide binding site of Ras. When the global mutational sensitivity of a highly prevalent oncogenic variant of Ras is interrogated, Ras G12V, hotspot mutations do not elicit a gain-of-function effect, suggesting an epistatic coupling between the Gly 12 residue and other residues in the protein structure. In sum, the mutational robustness of Ras is determined by its need to be regulated by a GAP and a GEF, as well as its suppression of intrinsic, spontaneous activation by hotspot residues that are distributed throughout the protein structure. From the perspective of evolution, it appears that dual selective pressures analogous to positive and negative selection in the presence and absence of regulators, respectively, contribute to the evolutionary conservation of the primary sequence of Ras. This concept will be explored further, later on in this thesis.

Methods

In vitro GTP hydrolysis assay

GTP hydrolysis assays were performed with recombinant *E. coli* phosphate-binding protein labeled with the MDCC fluorophore (phosphate sensor, Thermo Fisher). First, each Ras mutant was loaded with GTP by incubating protein with a ten-fold molar excess of EDTA in the presence of a ten-fold molar excess of GTP. The loading reaction was performed on ice for at least one hour and the reaction was quenched by the addition of a twenty-fold molar excess of $MgCl_2$. The reaction was then buffer exchanged using a small scale NAP-5 column (GE Healthcare) in order to remove excess nucleotide and salt.

For the GTP hydrolysis assay, 0.5 μM Ras was mixed with 0.5 μM phosphate sensor and 0.25 μM p120 RasGAP in the following buffer: 40 mM HEPES, 150 mM NaCl, 4 mM $MgCl_2$, 1mM TCEP, 5% glycerol. The reaction progress curve was monitored on a Tecan fluorescence plate reader with measurements taken every 15 seconds. Data were fit to a single exponential for quantification of GAP-stimulated GTP hydrolysis rates.

In vitro nucleotide exchange assay

GDP release assays were performed with Ras bound to mant-dGDP (Axxora Biosciences), loaded with the protocol described previously. 1.5 μM Ras was mixed with 2.5 μM GEF and 3.5 μM GDP in solution in the following buffer: 40 mM HEPES, 150 mM NaCl, 4 mM $MgCl_2$, 1 mM TCEP, 5% glycerol. Nucleotide exchange rates were monitored on a Tecan fluorescence plate reader with measurements taken every 15 seconds. Data were fit to a single exponential decay curve for quantification of GEF-mediated GDP exchange rates.

References

Baker, R., Wilkerson, E.M., Sumita, K., Isom, D.G., Sasaki, A.T., Dohlman, H.G., and Campbell, S.L. (2013). Differences in the regulation of K-Ras and H-Ras isoforms by monoubiquitination. J Biol Chem *288*, 36856–36862.

Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P.A., Stratton, M.R., et al. (2004). The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer *91*, 355–358.

Bennett, B.D., Kimball, E.H., Gao, M., Osterhout, R., Van Dien, S.J., and Rabinowitz, J.D. (2009). Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. Nat Chem Biol *5*, 593–599.

Boriack-Sjodin, P.A., Margarit, S.M., Bar-Sagi, D., and Kuriyan, J. (1998). The structural basis of the activation of Ras by Sos. Nature *394*, 337–343.

Bourne, H.R., Sanders, D.A., and McCormick, F. (1991). The GTPase superfamily: conserved structure and molecular mechanism. Nature *349*, 117–127.

Brune, M., Hunter, J.L., Corrie, J.E., and Webb, M.R. (1994). Direct, real-time measurement of rapid inorganic phosphate release using a novel fluorescent probe and its application to actomyosin subfragment 1 ATPase. Biochemistry *33*, 8262–8271.

Daley, G.Q., and Baltimore, D. (1988). Transformation of an interleukin 3-dependent hematopoietic cell line by the chronic myelogenous leukemia-specific P210bcr/abl protein. Proc Natl Acad Sci U S A *85*, 9312–9316.

Eberth, A., and Ahmadian, M.R. (2009). In vitro GEF and GAP assays. Curr Protoc Cell Biol *Chapter 14*, Unit 14.9.

Hobbs, G.A., Der, C.J., and Rossman, K.L. (2016). RAS isoforms and mutations in cancer at a glance. J Cell Sci *129*, 1287–1292.

Hoover, R.R., Gerlach, M.J., Koh, E.Y., and Daley, G.Q. (2001). Cooperative and redundant effects of STAT5 and Ras signaling in BCR/ABL transformed hematopoietic cells. Oncogene *20*, 5826–5835.

Iwig, J.S., Vercoulen, Y., Das, R., Barros, T., Limnander, A., Che, Y., Pelton, J.G., Wemmer, D.E., Roose, J.P., and Kuriyan, J. (2013). Structural analysis of autoinhibition in the Ras-specific exchange factor RasGRP1. Elife *2*, e00813.

Janakiraman, M., Vakiani, E., Zeng, Z., Pratilas, C.A., Taylor, B.S., Chitale, D., Halilovic, E.,

Wilson, M., Huberman, K., Ricarte Filho, J.C., et al. (2010). Genomic and biological characterization of exon 4 KRAS mutations in human cancer. Cancer Res *70*, 5901–5911.

Mathey-Prevot, B., Nabel, G., Palacios, R., and Baltimore, D. (1986). Abelson virus abrogation of interleukin-3 dependence in a lymphoid cell line. Mol Cell Biol *6*, 4133–4135.

Mazhab-Jafari, M.T., Marshall, C.B., Smith, M.J., Gasmi-Seabrook, G.M.C., Stathopulos, P.B., Inagaki, F., Kay, L.E., Neel, B.G., and Ikura, M. (2015). Oncogenic and RASopathy-associated K-RAS mutations relieve membrane-dependent occlusion of the effector-binding site. Proc Natl Acad Sci U S A *112*, 6625–6630.

McLaughlin, R.N., Poelwijk, F.J., Raman, A., Gosal, W.S., and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. Nature *491*, 138–142.

Pai, E.F., Krengel, U., Petsko, G.A., Goody, R.S., Kabsch, W., and Wittinghofer, A. (1990). Refined crystal structure of the triphosphate conformation of H-ras p21 at 1.35 A resolution: implications for the mechanism of GTP hydrolysis. EMBO J *9*, 2351–2359.

Palacios, R., and Steinmetz, M. (1985). Il-3-dependent mouse clones that express B-220 surface antigen, contain Ig genes in germ-line configuration, and generate B lymphocytes in vivo. Cell *41*, 727–734.

Prior, I.A., Lewis, P.D., and Mattos, C. (2012). A comprehensive survey of Ras mutations in cancer. Cancer Res *72*, 2457–2467.

Sass, P., Field, J., Nikawa, J., Toda, T., and Wigler, M. (1986). Cloning and characterization of the high-affinity cAMP phosphodiesterase of Saccharomyces cerevisiae. Proc Natl Acad Sci U S A *83*, 9303–9307.

Scheffzek, K., Ahmadian, M.R., Kabsch, W., Wiesmüller, L., Lautwein, A., Schmitz, F., and Wittinghofer, A. (1997). The Ras-RasGAP complex: structural basis for GTPase activation and its loss in oncogenic Ras mutants. Science *277*, 333–338.

Smith, M.J., and Ikura, M. (2014). Integrated RAS signaling defined by parallel NMR detection of effectors and regulators. Nat Chem Biol *10*, 223–230.

Spoerner, M., Herrmann, C., Vetter, I.R., Kalbitzer, H.R., and Wittinghofer, A. (2001). Dynamic properties of the Ras switch I region and its importance for binding to effectors. Proc Natl Acad Sci U S A *98*, 4944–4949.

Stiffler, M.A., Hekstra, D.R., and Ranganathan, R. (2015). Evolvability as a function of purifying selection in TEM-1 β-lactamase. Cell *160*, 882–892.

Chapter 4: The conformational dynamics of Ras

Introduction

In this chapter, I will describe the conformational dynamics of Ras, as assessed computationally by molecular dynamics simulations. Over the decades of research on Ras and Ras-superfamily GTPases, a large body of work has amassed on the conformational dynamics of these proteins. I have leveraged these preceding studies to examine the dynamics of wild-type H-Ras, and compared these computational simulations to those in which a mutation has been made at a hotspot residue identified from the experimental bacterial two-hybrid screens. I have examined the simulations visually through structural superpositions of backbone and sidechain fluctuations, as is traditional for the analysis of molecular dynamics simulations. Additionally, in collaboration with Moitrayee Bhattacharyya, I have adapted a sidechain contact analysis algorithm to identify stable residue contacts over the course of the simulations, both for wild-type and hotspot mutant simulations. These analyses have revealed the effects of mutations on the stability of Ras, as gauged by changes in networks of interacting sidechains, which will be discussed in this chapter.

Hotspot residues dampen the conformational dynamics of wild-type Ras

To address the structural mechanism of activating mutations in unregulated Ras, we ran molecular dynamics simulations of wild-type H-Ras and H-Ras with four hotspot mutations, introduced individually (H27A, Q99A, L120A, and Y157A). His 27 is located at the beginning of Switch I (see Chapter 3). Gln 99 is in helix α3, and an indication of its possible importance comes from the fact that it interacts with Switch II. Leu 120 is in the guanine-binding loop, with the sidechain in van der Waals contact with the edge of the nucleotide base. Tyr 157 is in helix α5, and the loop leading into the N-terminal portion of helix α5 packs against the guanine base. Mutations at each of these sites result in increased nucleotide release, and mutation at one of them (Q99A) also shows suppression of GTP hydrolysis (see Chapter 3; note, however, that our assay for GTP hydrolysis does not provide a precise measurement of the suppression of the low intrinsic rate of GTP hydrolysis by Ras, as discussed earlier).

Molecular dynamics trajectories, each totaling 600 nanoseconds (ns), were generated for wild-type Ras and each mutant. In each case, we began with the crystal structures of wild-type H-Ras bound to GDP or GTP (PDB codes 5P21 and 4Q21, respectively (Milburn et al. 1990; Pai et al. 1990), replacing the appropriate sidechains by alanine for each simulation of a mutant protein. Some of the trajectories were generated as one contiguous 600 ns simulation, while others were generated in two independent blocks of 300 ns each, and then merged for analysis.

In each of the simulations of the mutant proteins bound to GTP, we discern changes in the dynamics of the switch regions, the guanine-binding loops and helix α3, which abuts Switch II. A superposition of structures sampled from the GTP-bound simulations is shown in Figure 1. The overall structure of Ras is very stable in each of the simulations, as can be seen in views of the superimposed structures facing the side of Ras that is opposite to the switch regions (middle diagrams in each of the panels in Figure 1 and Figure 2, with helices

α4 and α5 facing the viewer). Excluding the switch regions, the r.m.s. fluctuation in Cα positions from the mean structure is very small (less than 1 Å) over the course of each of the trajectories, and the r.m.s. deviation from the crystal structure is also small (less than 1 Å). Comparison of the GDP-bound simulations was less informative, since the increased dynamics of the switch regions in these simulations made it difficult to identify the effects of the mutations.

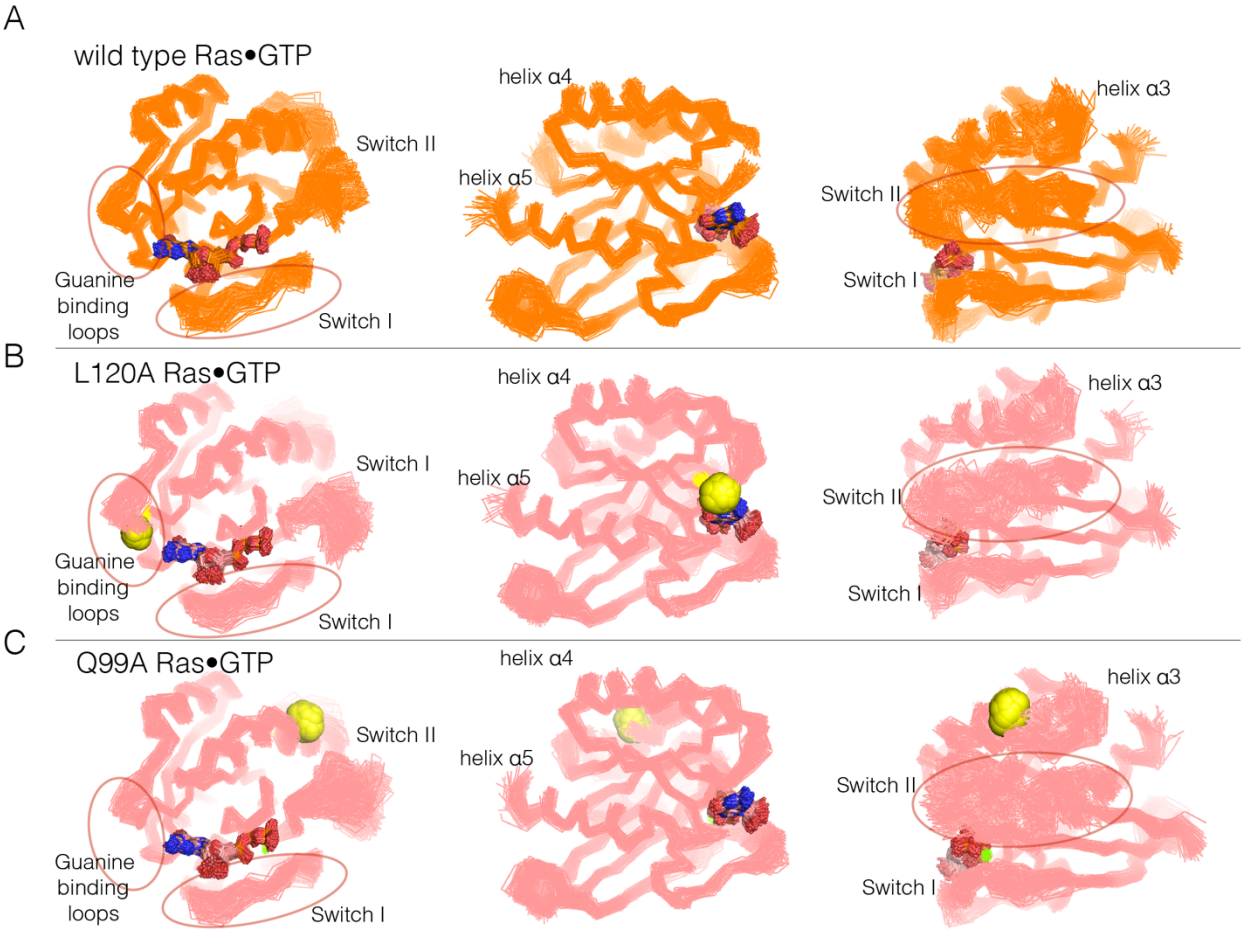

Figure 1. Superposition of structures from molecular dynamics simulations of GTP-bound forms of Ras. A. Wild-type Ras•GTP. The diagrams show the superposition of 30 structures sampled evenly from two 300 ns trajectories. The diagram on the left shows a canonical view of Ras•GTP. The two other diagrams show two orthogonal views. **B.** As in **A.**, for Ras•GTP L120A, shown as yellow spheres. **C.** As in **A.**, for Ras•GTP Q99A, shown as yellow spheres.

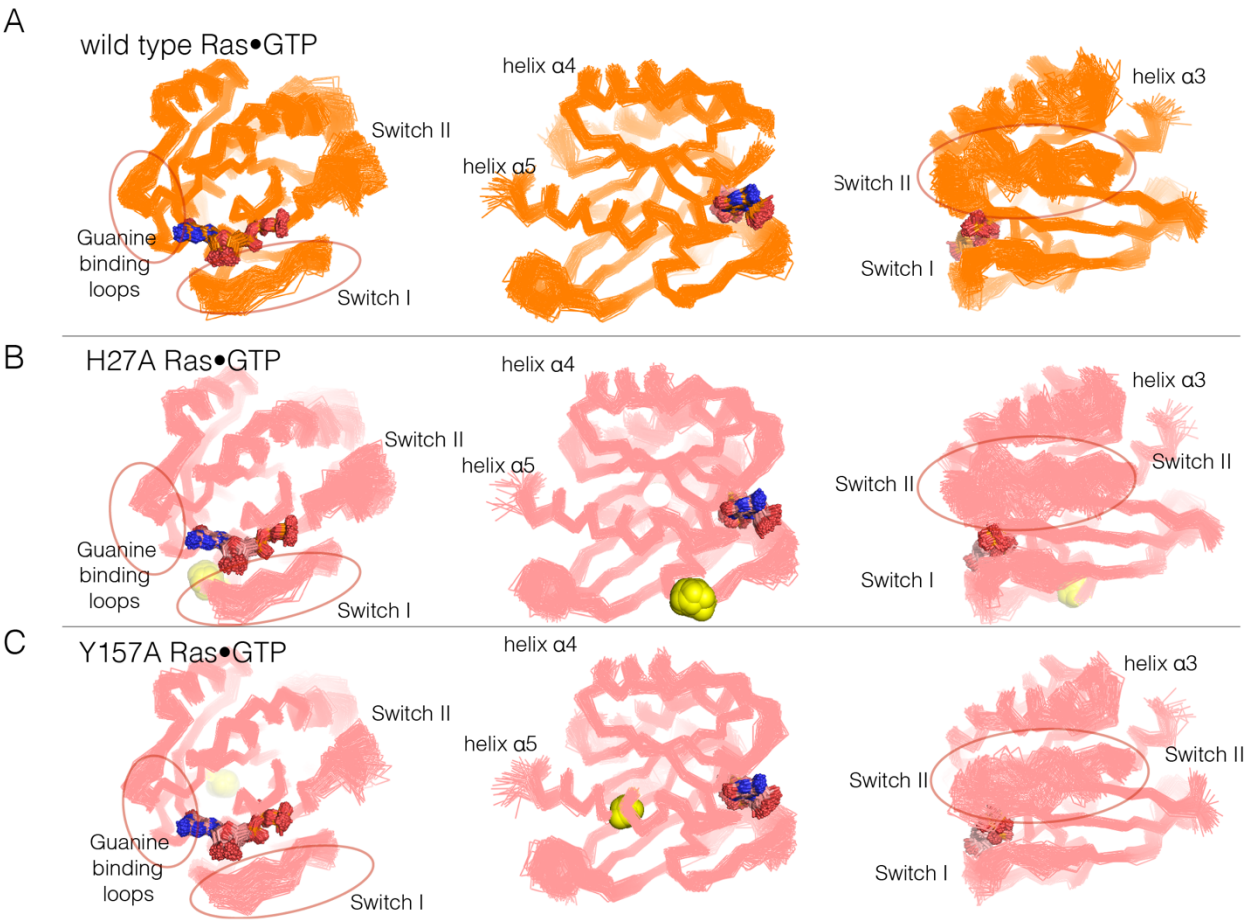

Figure 2. Superposition of structures from GTP-bound simulations of wild type, H27A, and Y157A mutants. A. Wild-type Ras. The diagrams show the superposition of 300 structures sampled evenly from two 300 ns trajectories. The diagram on the left shows a canonical view of Ras•GTP. The two other diagrams show two orthogonal views. **B.** As in **A.**, for Ras H27A, shown as yellow spheres. **C.** As in **A.**, for Ras Y157A, shown as yellow spheres.

Switch II is very mobile in the simulation of wild-type Ras•GTP, with r.m.s. fluctuations in Cα positions of ~1.2 Å when the structures are superimposed without using Switch I and II. But in all four simulations of the mutant structures, there are increased conformational fluctuations of Switch II (left and right diagrams in each of the panels in Figure 1; the r.m.s. fluctuations in Cα positions range from ~1.4 Å to ~1.7 Å (Figure 3)). This is particularly striking since the four mutations are at diametrically opposite ends of the structure, and the changes in dynamics indicate allosteric coupling between distant points in Ras and the Switch II helix. There are also changes in the dynamics of helix α3, which is structurally coupled to Switch II. Alterations in the guanine-binding loops and Switch I are also evident in the trajectories, but are more difficult to appreciate in the structural superpositions.

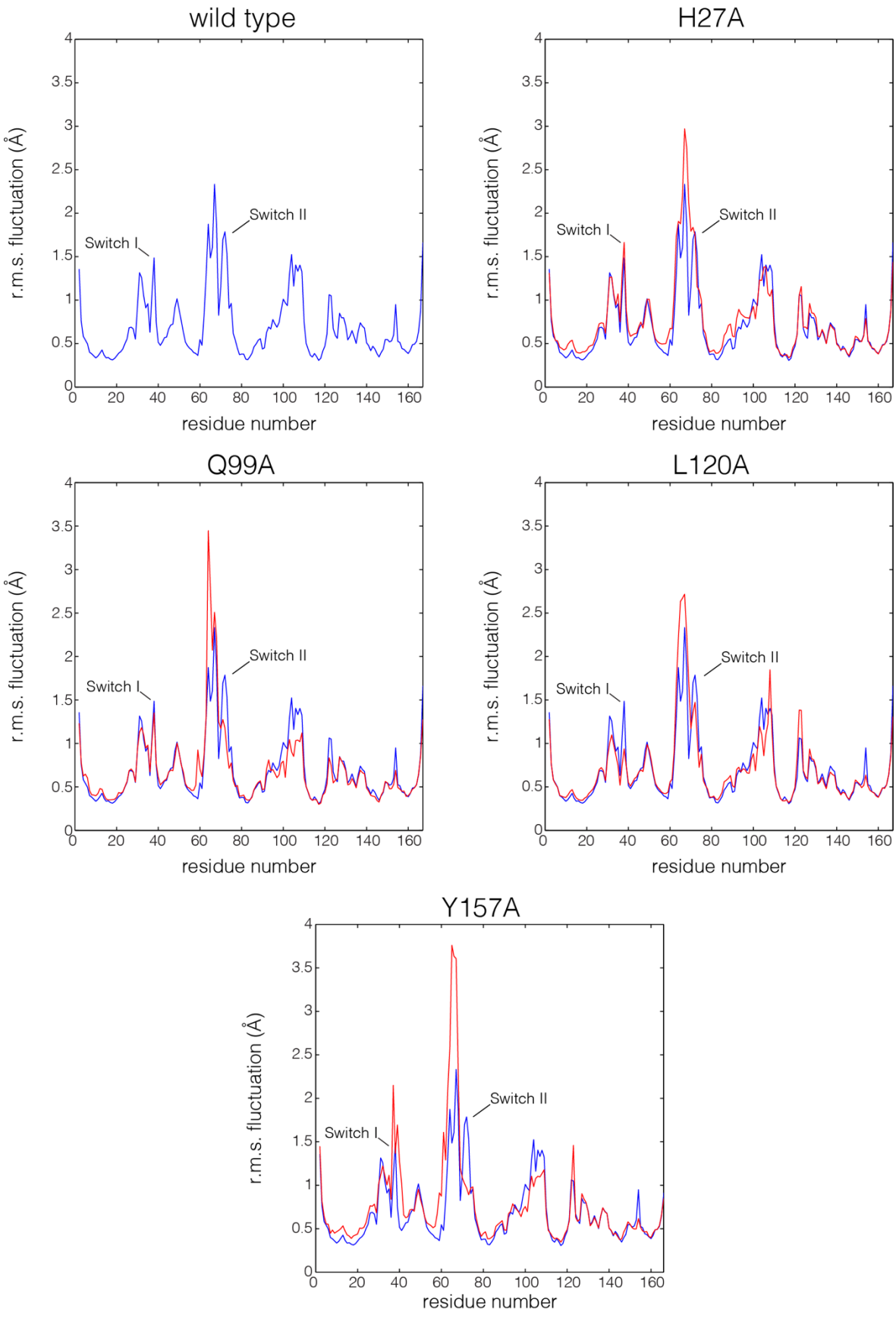
wild type
H27A
Q99A
L120A
Y157A
r.m.s. fluctuation (Å)
residue number
Switch I
Switch II
0
0.5
1
1.5
2
2.5
3
3.5
4
0
20
40
60
80
100
120
140
160

Figure 3. Flexibility in Ras measured by root mean squared fluctuations of Cα atoms. Fluctuations in Ras are measured by the root mean square (r.m.s.) fluctuations of Cα atoms in Ras over the course of the 600 ns molecular dynamics trajectories. The r.m.s. fluctuations for wild type Ras are depicted in blue, and each Ras mutant is depicted in red. Switch II shows greater mobility in every mutant simulation in comparison to wild type Ras.

The molecular dynamics trajectories suggest that there is a loosening of the structure of Ras when His 27, Gln 99, Leu 120, and Tyr 157 are mutated. To assess the effects of mutations quantitatively, we carried out an analysis of interaction networks in the simulations, as has been done previously for a Gln 61 mutant of Ras (Fetics et al. 2015). We used a different method from that used in the previous simulation study, and we determined the strength of a noncovalent interaction between sidechains by counting the number of atom pairs between residues within a 4.5 Å cutoff, normalized by the sizes of the interacting residues (Brinda and Vishveshwara 2005; Ghosh and Vishveshwara 2007; Bhattacharyya et al. 2013; Bhattacharyya et al. 2016). Interactions between sidechains that are present in more than a threshold fraction of instantaneous structures in the trajectory are identified as stable interactions. We chose 50% as the threshold value for definition of a stable contact, and analyzed each trajectory by drawing lines between the Cα atoms of the corresponding residues in stable contacts, as shown for the simulation of wild-type Ras•GTP. The stability of the wild-type Ras simulation is reflected in the large number of stable contacts that anchor all of the neighboring structural elements, throughout the structure (Figure 4).

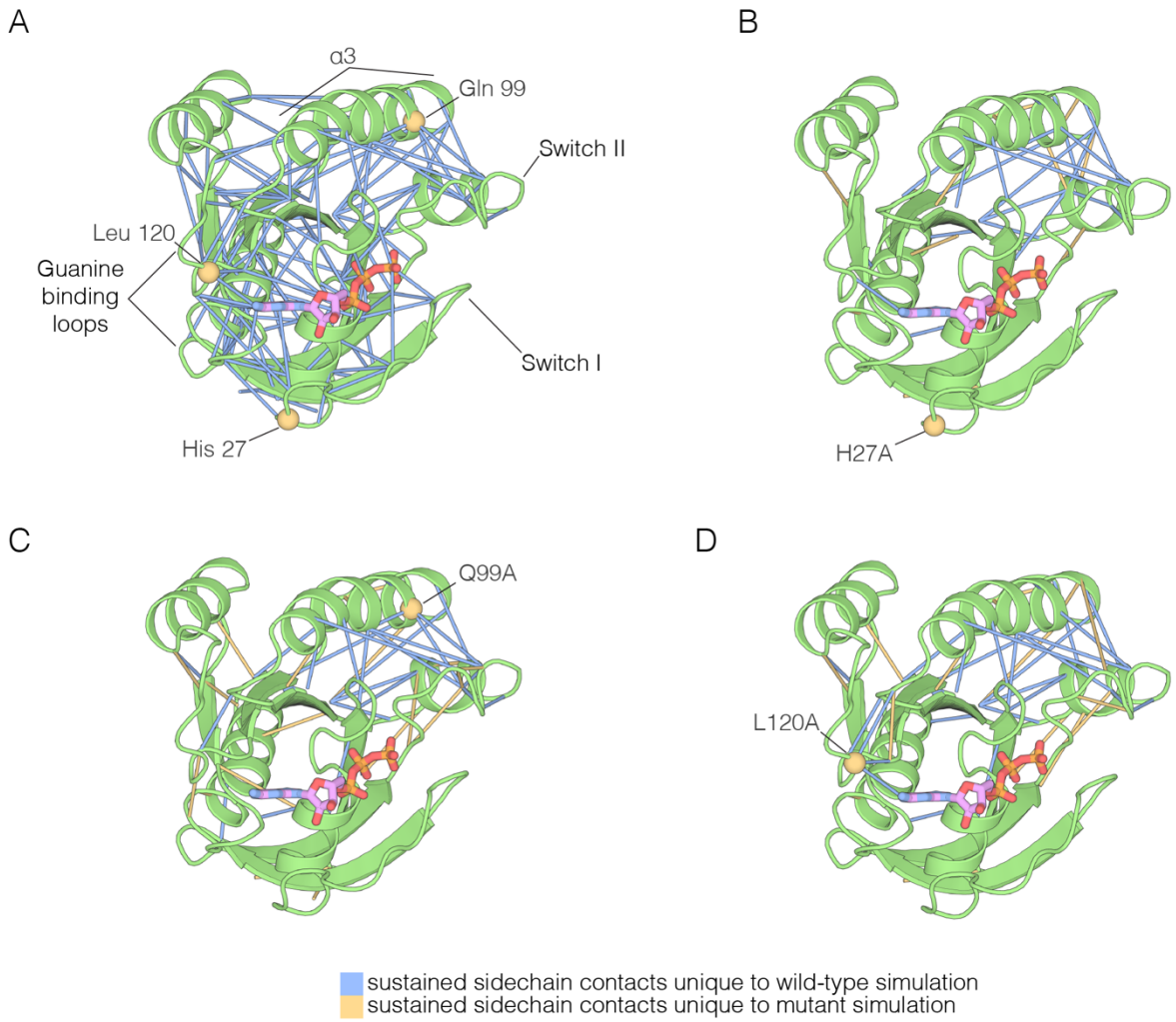

Figure 4. The impact of hotspot mutations on structural flexibility in Ras. A. Sustained sidechain-sidechain contacts in the simulation of wild-type Ras•GTP. We used the network analysis tool of Vishveshwara and co-workers (Bhattacharyya et al., 2013) to analyze the molecular dynamics trajectories for Ras. Strong interactions between sidechains, as defined by Bhattacharyya et al., 2013, that are present in more than a threshold fraction of 50% of the instantaneous structures in the trajectory are identified by blue lines. These lines are drawn between the Cα atoms of the corresponding residues. **B.** For simulations of Ras•GTP H27A, contacts that are present in the wild type simulation but not in the H27A simulation are drawn in blue, and contacts that are present in the H27A simulation but not in the wild type simulation are drawn in yellow. **C**. As in **B**., for Ras•GTP Q99A. **D.** As in **B**., for Ras•GTP L120A.

Global analysis of sidechain contacts in molecular dynamics simulations of Ras

The analysis of stable sidechain contacts shows dramatically different results for the H27A, Q99A and L120A simulations (Figure 4B-D). In each of these trajectories, there are far fewer contacts that persist for longer that the 50% threshold we used in the analysis. Switch II and helix α3 remain tightly connected to each other, but the rest of the structure is no longer stably coupled. We emphasize that the loss of stable contacts in the simulations of the mutants does not reflect a general breakdown in the integrity of the structure, but is due instead to a subtle decrease in the maintenance of stable contacts. This can be seen in Figure 5, which illustrates the interaction between Tyr 137, in helix α4, and His 94, in helix α3. These two residues form a hydrogen-bonding interaction in the crystal structure (Ting et al. 2015) (Figure 5). In the simulations, His 94 switches between different conformations, but maintains contact with Tyr 137 more frequently in the wild-type simulation than in the L120A simulation (Figure 5B-C). The simulation results for the Y157A mutant are different than those for the other three. The trajectory for the Y157A mutant does not show a decrease in stable sidechain contacts when compared to the wild-type simulation, although there is evidence for a loosening of the coordination of the nucleotide in both GTP- and GDP-bound simulations (Figure 6).

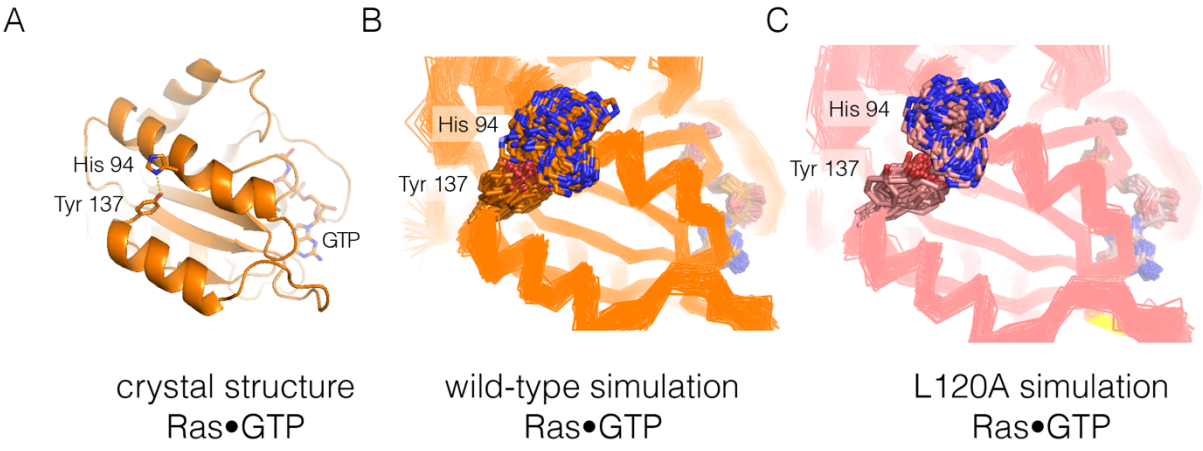

Figure 5. Residue contacts in Ras. A. The crystal structure of wild-type Ras•GTP with the sidechains of His 94 and Tyr 137 shown as sticks. **B-C.** Shown here are 30 structures sampled from two 600 ns simulations, one for wild-type Ras (**B**) and one for the L120A mutant (**C**). The sidechains of His 94 and Tyr 137 are shown as sticks, with the Cα trace of the protein shown as a thin line.

Figure 6. Dynamics of the Y157A mutation. Simulation of wild-type Ras (left) and the Y157A mutation (right) shows an increase in dynamics of Switch I (as highlighted by the red reference residue, Tyr 32) and the guanine moiety of the nucleotide, suggesting that the nucleotide is more readily released upon mutation to Tyr 157.

In the crystal structure of Ras•GTP, Switch II is positioned such that the sidechain of Gln 61 is not far from the catalytically important conformation that is stabilized by the binding of Ras-GAP. The increased fluctuations seen in Switch II in simulations of mutant Ras tend to move Gln 61 away from catalytically competent conformations, and would therefore be expected to reduce the intrinsic rate of GTP hydrolysis, leading to increased signaling activity. Destabilization of interactions between residues in the guanine binding loops and the guanine base of the nucleotide, or between the P-loop and the phosphate groups of the nucleotide, would increase the intrinsic GDP release rate, leading to activation by the replacement of GDP by GTP. Thus, the simulations support the idea that the hotspot residues analyzed in this study dampen structural fluctuations that would otherwise disrupt the functional cycle.

Dynamic cross correlation analysis of Ras dynamics

An alternative method to quantify the extent of coupling in protein structures involves calculating the cross-correlations of atomic displacements across the entire molecular dynamics trajectory. Though we did not use this method to quantify the extent of molecular motions that were observed in the molecular dynamics simulations, I include this section of the thesis for completion. These correlations were calculated using the Bio3D suite in R. The correlation data is computed from a numeric matrix of Cartesian coordinates, with a row per structure or frame of the simulation, and the atomic displacements are relative to a reference structure that is typically the average structure over the entire trajectory. The covariance matrix of all atomic fluctuations is informative of the extent to which atomic fluctuations are correlated with one another, and this matrix is often termed the dynamic cross correlation matrix (DCCM). Positively correlated parts of the protein structure often move together in tandem over the simulations, whereas negatively correlated parts of the protein structure move in opposite directions reminiscent of a collective "breathing" motion. The first proteins to be studied with DCCM analysis

included cytochrome c and ribonuclease A with two different substrates (McCammon 1984; Brünger et al. 1985).

DCCM analysis of Ras and activating hotspot mutations revealed extensive coupling throughout the protein structure. Wild-type Ras exhibits anti-correlations between diametrically opposed parts of the protein structure, such as between the α1 helix at the base of Switch I and the α3 helix, shown as blue lines between Cα atoms in the protein structure (Figure 7A). Interestingly, many significant positive correlations are not observed in this analysis, which may be due to the overall rigidity of the Ras structure. Additional anti-correlations exist between the α2 helix of Switch II and the α3 helices (Figure 7A). Taken together, this suggests that these regions of the protein are coupled in a manner that allows the protein to "relax", and it is tempting to speculate that this overall relaxation impacts the functional properties of nucleotide release and hydrolysis. Mutations such as H27A and Q99A appear to greatly exacerbate these relaxations, and in conjunction with the network analysis of sidechain interactions, appears to be consistent with an overall loosening of the protein structure that enables nucleotide to be more readily released (Figure 7B). The L120A mutation introduces new anti-correlations between the α1 helix and the α4 helix, and again, anti-correlations between the α1 helix at the base of Switch I may also impact the process of intrinsic nucleotide release. Finally, the DCCM analysis for the Y157A mutation does not appear to identify significant correlated motions in the protein structure, potentially suggesting that the effects of this mutation may be subtler than reflected in a global analysis of protein motions (Figure 7B). It will be interesting to follow up studies on the Y157A mutation using hydrogen exchange coupled to NMR, for a fine grained picture of the dynamics of this Ras mutant.

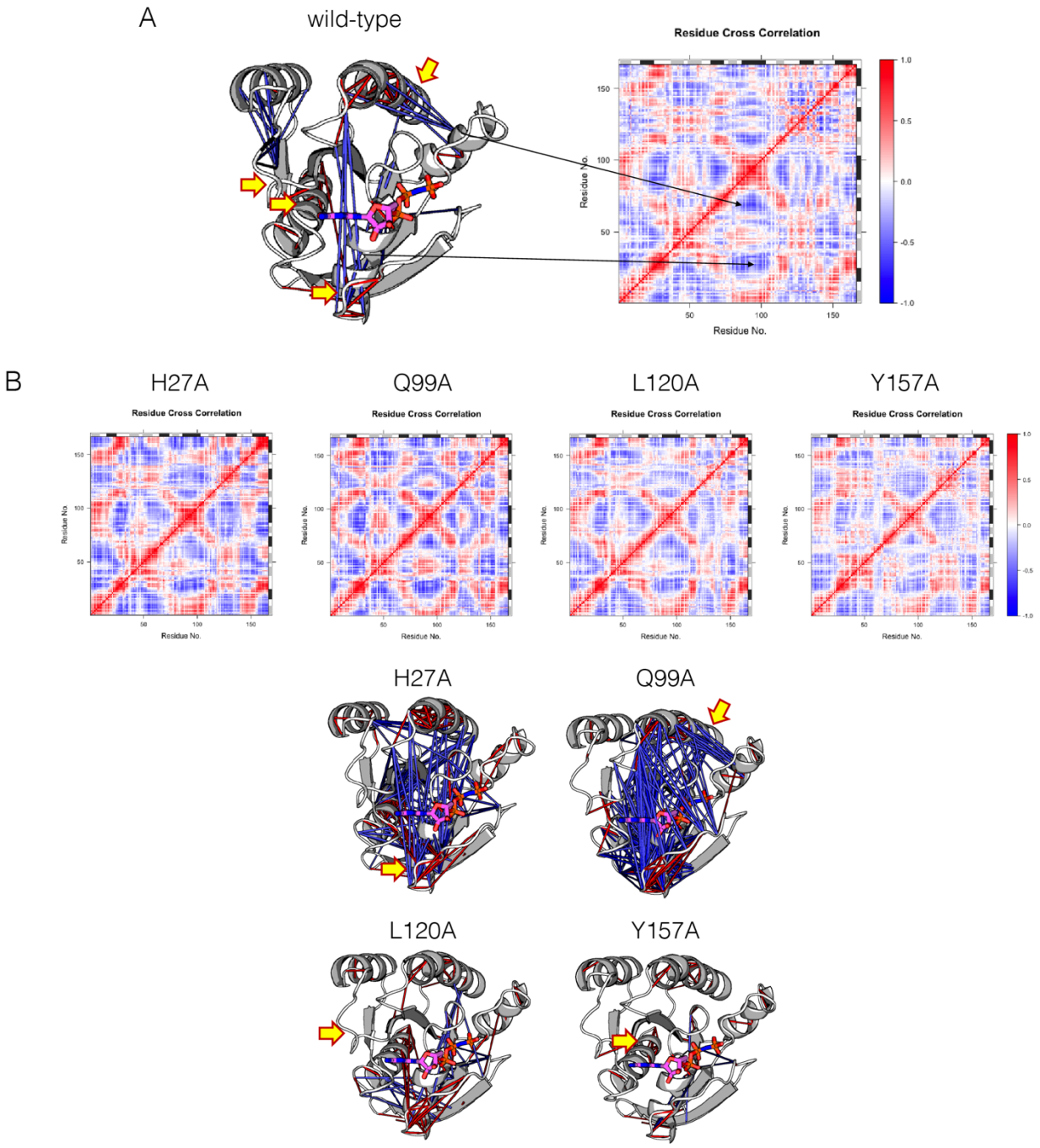

Figure 7. Dynamic cross correlation analysis of activating Ras mutations. A. Wild-type Ras exhibits significant anti-correlations between the α1 and α3 helices (blue lines), but does not exhibit a large number of positively correlated regions of the structure (red lines). **B.** DCCM analysis of activating Ras mutations shows extensive anti-correlated motions in the simulations.

Summary

Ras, like all proteins, is inherently dynamic and its dynamics affect crucial functional properties such as its binding to Raf-RBD in its GTP-bound state, its ability to hydrolyze GTP, and its ability to release bound nucleotide. Computational simulations of wild-type Ras have demonstrated these functional properties in my studies, as well as in the large body of literature on Ras dynamics over the past decades. My experimental studies on Ras mutations in the bacterial two-hybrid assay, described in Chapter 3, identified several residues that affected the properties of intrinsic GTP hydrolysis and intrinsic nucleotide release. These residues appeared to exert their effects allosterically, as they were largely one shell away from active site residues that contact the nucleotide directly. In this chapter, we have shown that mutations to these residues exert their allosteric effects by directly impacting the dynamics of the structural regions in Ras that control the functional properties of nucleotide exchange and release. Furthermore, we show that these residues act as crucial "latches" that serve to hold the overall structure rigid, as revealed by a network analysis of sidechain motions computed from the molecular dynamics simulations. Mutations to these residues consequently explain the gain of function that was observed in the experimental screens, thus providing a structural context to a set of newly identified residues that allosterically control Ras function.

Methods

Molecular dynamics simulations

Simulations of Ras in this study were carried out using the AMBER force field (Case et al., 2016) combined with the ff99SB-ILBN backbone correction (Lindorff-Larsen et al., 2010) and the TIP3P water model (Mahoney and Jorgensen 2000). The simulated systems were solvated in a water box with a minimum of 12.5 Å separation from any protein atom and the box edges, and residue protonation states corresponded to pH 7. The protein backbone atoms were restrained to their initial positions using a harmonic potential with a force constant of 1 kcal mol^{-1} $Å^2$ for 5 ns as an equilibration step and restraints were subsequently removed. Simulations were performed in the NPT ensemble with T = 300K and p = 1 bar. Water molecules and all bond lengths were constrained using M-SHAKE (Kräutler et al. 2001). Van der Waals and short-range electrostatic interactions were cut off at 12.5 Å for the simulations. Long-range electrostatic energies were calculated using particle-mesh Ewald summation. The simulation time step was 2 fs for the equilibration stage and 2 fs for production simulations. Every simulation in the study was performed in duplicate. Network analysis of sidechain-sidechain interactions was carried out using the PSN-Ensemble software (Bhattacharyya et al. 2013).

References

Bhattacharyya, M., Bhat, C.R. and Vishveshwara, S. 2013. An automated approach to network features of protein structure ensembles. *Protein Science* 22(10), pp. 1399–1416.

Bhattacharyya, M., Ghosh, S. and Vishveshwara, S. 2016. Protein Structure and Function: Looking through the Network of Side-Chain Interactions. *Current Protein & Peptide Science* 17(1), pp. 4–25.

Brinda, K.V. and Vishveshwara, S. 2005. A network representation of protein structures: implications for protein stability. *Biophysical Journal* 89(6), pp. 4159–4170.

Brünger, A.T., Brooks, C.L. and Karplus, M. 1985. Active site dynamics of ribonuclease. *Proceedings of the National Academy of Sciences of the United States of America* 82(24), pp. 8458–8462.

Fetics, S.K., Guterres, H., Kearney, B.M., Buhrman, G., Ma, B., Nussinov, R. and Mattos, C. 2015. Allosteric effects of the oncogenic RasQ61L mutant on Raf-RBD. *Structure* 23(3), pp. 505–516.

Ghosh, A. and Vishveshwara, S. 2007. A study of communication pathways in methionyl-tRNA synthetase by molecular dynamics simulations and structure network analysis. *Proceedings of the National Academy of Sciences of the United States of America* 104(40), pp. 15711–15716.

Kräutler, V., van Gunsteren, Wilfred F. and Hünenberger, P.H. 2001. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of Computational Chemistry*. Available at: http://onlinelibrary.wiley.com/doi/10.1002/1096-987X(20010415)22:5%3C501::AID-JCC1021%3E3.0.CO;2-V/abstract.

Mahoney, M.W. and Jorgensen, W.L. 2000. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.* 112(20), p. 8910.

McCammon, J.A. 1984. Protein dynamics. *Reports on Progress in Physics* 47(1), pp. 1–46.

Milburn, M.V., Tong, L., deVos, A.M., Brünger, A., Yamaizumi, Z., Nishimura, S. and Kim, S.H. 1990. Molecular switch for signal transduction: structural differences between active and inactive forms of protooncogenic ras proteins. *Science* 247(4945), pp. 939–945.

Pai, E.F., Krengel, U., Petsko, G.A., Goody, R.S., Kabsch, W. and Wittinghofer, A. 1990. Refined crystal structure of the triphosphate conformation of H-ras p21 at 1.35 A resolution: implications for the mechanism of GTP hydrolysis. *The EMBO Journal* 9(8), pp. 2351–2359.

Ting, P.Y., Johnson, C.W., Fang, C., Cao, X., Graeber, T.G., Mattos, C. and Colicelli, J. 2015. Tyrosine phosphorylation of RAS by ABL allosterically enhances effector binding. *The FASEB Journal* 29(9), pp. 3750–3761.

Chapter 5: Natural sequence variation in the evolution of Ras

Introduction

Evolution has explored a broad span of the primary sequence of Ras, and extant Ras sequences in phylogenetic databases that have been collected by genome sequencing is representative of sequence variation that has persisted throughout the processes of mutation and natural selection. As such, sequence databases provide complementary data to that collected from saturation mutagenesis experiments, and represent sequences that have often accrued multiple mutations over the course of many millions of years of evolution. In this chapter, I will present an analysis of natural sequence variation in the evolution of Ras in metazoans. I compare the space of natural sequence variation to the chemically accessible space of mutations that was collected from screening the saturation point-mutagenesis library of Ras. Analysis of natural sequence variation in Ras sequences from invertebrates, and a computationally reconstructed ancestral Ras sequence, revealed regions of sequence space that tolerate mutation, which are also tolerant of mutation in the experimental screen. This variation is largely not present in Ras sequences from vertebrates, suggesting an evolutionarily recent innovation that further constrains Ras sequence space in vertebrates. Finally, I will discuss the emergence of a regulatory factor that may have imposed this constraint, drawing from existing structural information as well as structure determination of a Ras ortholog. The work described in this chapter was done in collaboration with John P. Barton, Joshua Cofsky, Yasushi Kondo, and Arup Chakraborty.

The regulation of Ras is likely to have been altered in the vertebrate lineage

We carried out an evolutionary analysis of Ras proteins by aligning sequences from 72 species and generating an evolutionary tree (Figure 1A). This analysis identifies four regions that are the principal sites of divergence between invertebrate and vertebrate Ras sequences, denoted Variable Regions 1 (spanning residues 48 to 52 in human H-Ras), 2 (residues 82 to 101), 3 (residues 121 to 142) and 4 (residues 149 to 155). These variable regions are highlighted in an alignment of the sequence of human H-Ras with Ras from a sponge (*Amphimedon queenslandica*) and from a choanoflagellate (*Salpingoeca rosetta)* (Richter and King 2013) (Figure 1B).

In order to capture the essential features that distinguish human H-Ras from invertebrate Ras, we used the evolutionary tree to construct a hypothetical sequence corresponding to the root of the metazoan lineage (Ashkenazy et al. 2012). There are 48 differences between the sequences of H-Ras and the hypothetical protein at the base of the metazoan lineage (Figure 1C). Of these, 30 represent residues that would activate H-Ras if the wild-type residue were replaced by the residue in the ancestral sequence, based on mutational data in the unregulated-Ras experiment (Figure 1D). Eight differences correspond to neutral substitutions, and nine substitutions would decrease function if introduced into H-Ras. We also examined the effects of these substitutions in the data for regulated-Ras in the bacterial two-hybrid experiments. In this case, the trend is the opposite. Eight of the substitutions activate H-Ras, three are neutral, and 36 lead to a decrease in function (not shown). Similar results are obtained by comparing the sequence

of human H-Ras to that of extant invertebrate Ras sequences (see Figure 2, in which the sequence of a choanoflagellate Ras is compared to that of human H-Ras, along with the fitness effects of the substitutions as seen in the unregulated Ras experiment).

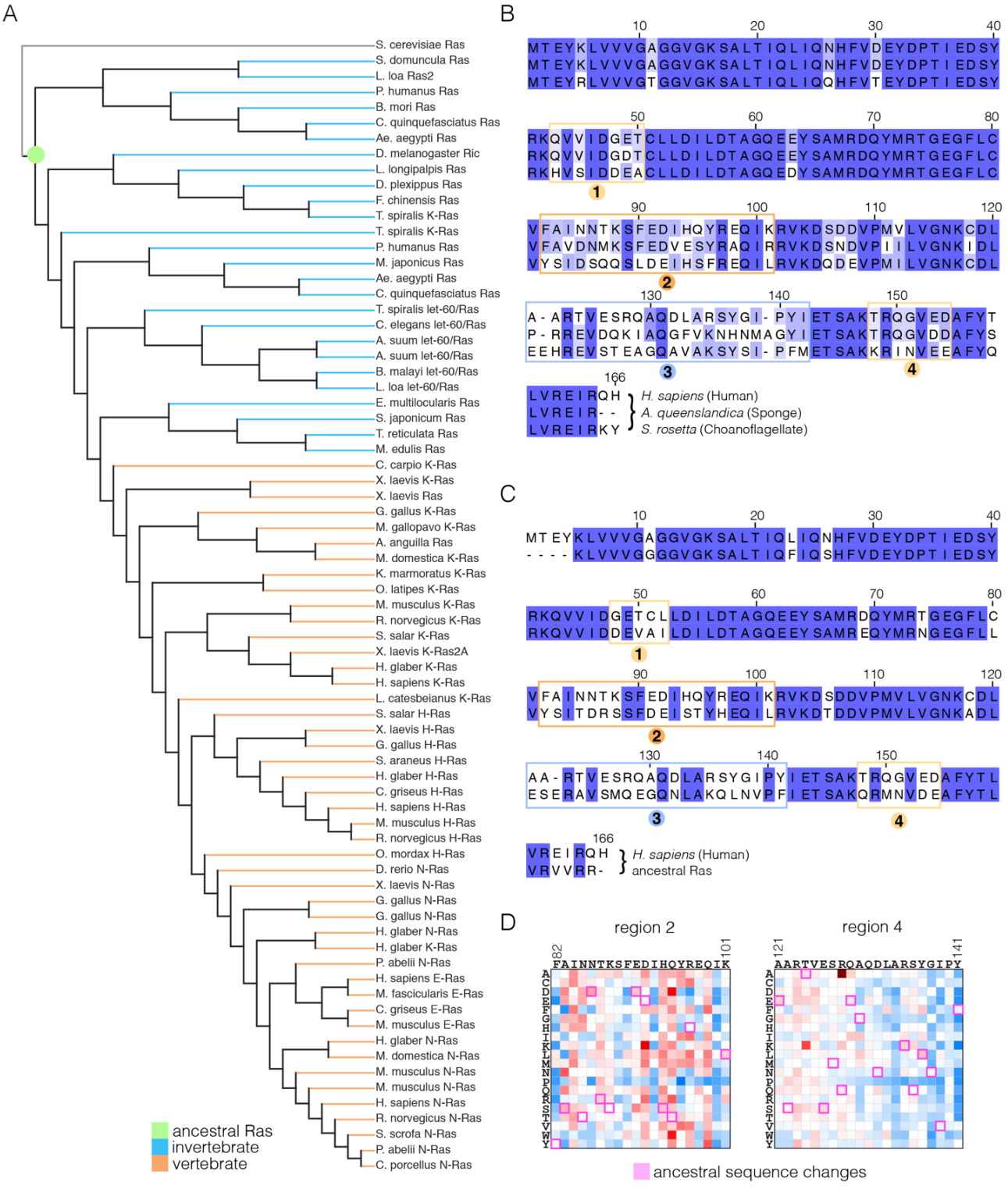

Figure 1. Sequence variation in Ras. A. An evolutionary tree from an alignment of 72 extant Ras sequences from invertebrates (blue) and vertebrates (orange). The hypothetical ancestral sequence at the base of the tree is highlighted in green. **B.** Sequence alignments of Ras from choanoflagellate (*S. rosetta*) and sponge (*A. queenslandica*) are 72% and 80% identical to human Ras, respectively. While the sequences are largely identical, there are four principal regions of sequence divergence. These regions correspond to residues 45-50, helix α3, helix α4, and residues 148-154 in human H-Ras. **C.** Comparison of the sequence of human H-Ras to the ancestral sequence from the base of the metazoan lineage reveals similar regions of sequence variation. *S. rosetta* Ras was not used in the alignment of Ras sequences to generate the tree shown in (**A**). **D.** The substitutions in human H-Ras that are present in the ancestral sequence are compared to the mutational data from the unregulated-Ras experiment. There are 48 differences between the sequences of the hypothetical ancestral protein and human H-Ras. Of these, 30 represent residues that would activate unregulated human H-Ras if the wild-type residue were replaced by the residue in the ancestral sequence. Eight differences correspond to neutral substitutions, and nine substitutions would decrease function if introduced into human H-Ras.

To summarize these findings, the ancestral Ras protein contains residues that, when substituted into human H-Ras, are activating in the absence of regulators and detrimental in the presence of a GAP and GEF. This suggests that as the sequence of Ras evolved into the vertebrate lineage, it underwent changes in the mechanisms that prevent unregulated activation, and a co-evolved dependence on the vertebrate GAPs and GEFs. Such co-evolution is consistent with the strong dependence of Ras mutations on the presence or absence of its associated regulatory proteins.

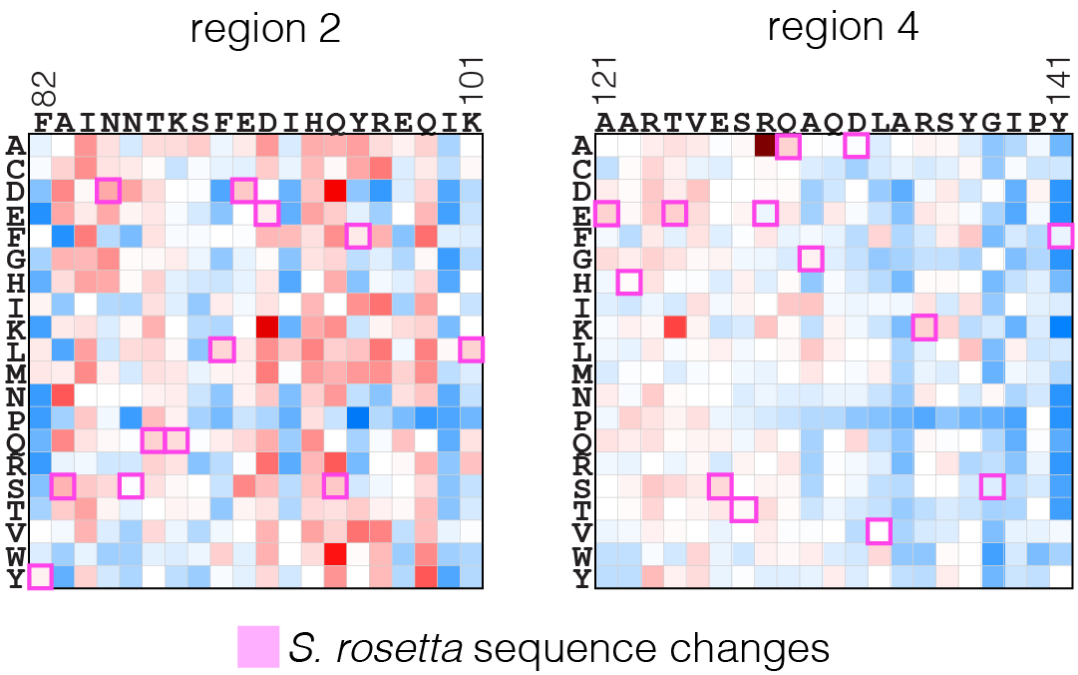

Figure 2. Effect of substitutions in human H-Ras that are present in *S. rosetta* Ras. The substitutions in human H-Ras that are present in the *S. rosetta* Ras sequence are compared to the mutational data from the unregulated-Ras experiment.

The variable regions of Ras are involved in interactions with the Ras activator Son-of-Sevenless

The Ras-specific nucleotide exchange factor son-of-sevenless (SOS) is unusual in having two binding sites for Ras. One is the canonical binding site that mediates nucleotide exchange (Boriack-Sjodin et al. 1998). The other is an allosteric site unique to SOS, which is specific for Ras•GTP, and SOS has low activity unless Ras•GTP is bound at this site (Margarit et al. 2003; Sondermann et al. 2004) (Figure 3). The allosteric binding of Ras•GTP to SOS converts SOS to a highly processive enzyme at the membrane, and this positive feedback loop in Ras activation is necessary for the generation of a sharp response in signaling by the T cell receptor (Das et al. 2009; Gureasko et al. 2008; Iversen et al. 2014). We find that the four Variable Regions correspond to elements in Ras that are correlated with the engagement of SOS at the SOS allosteric site.

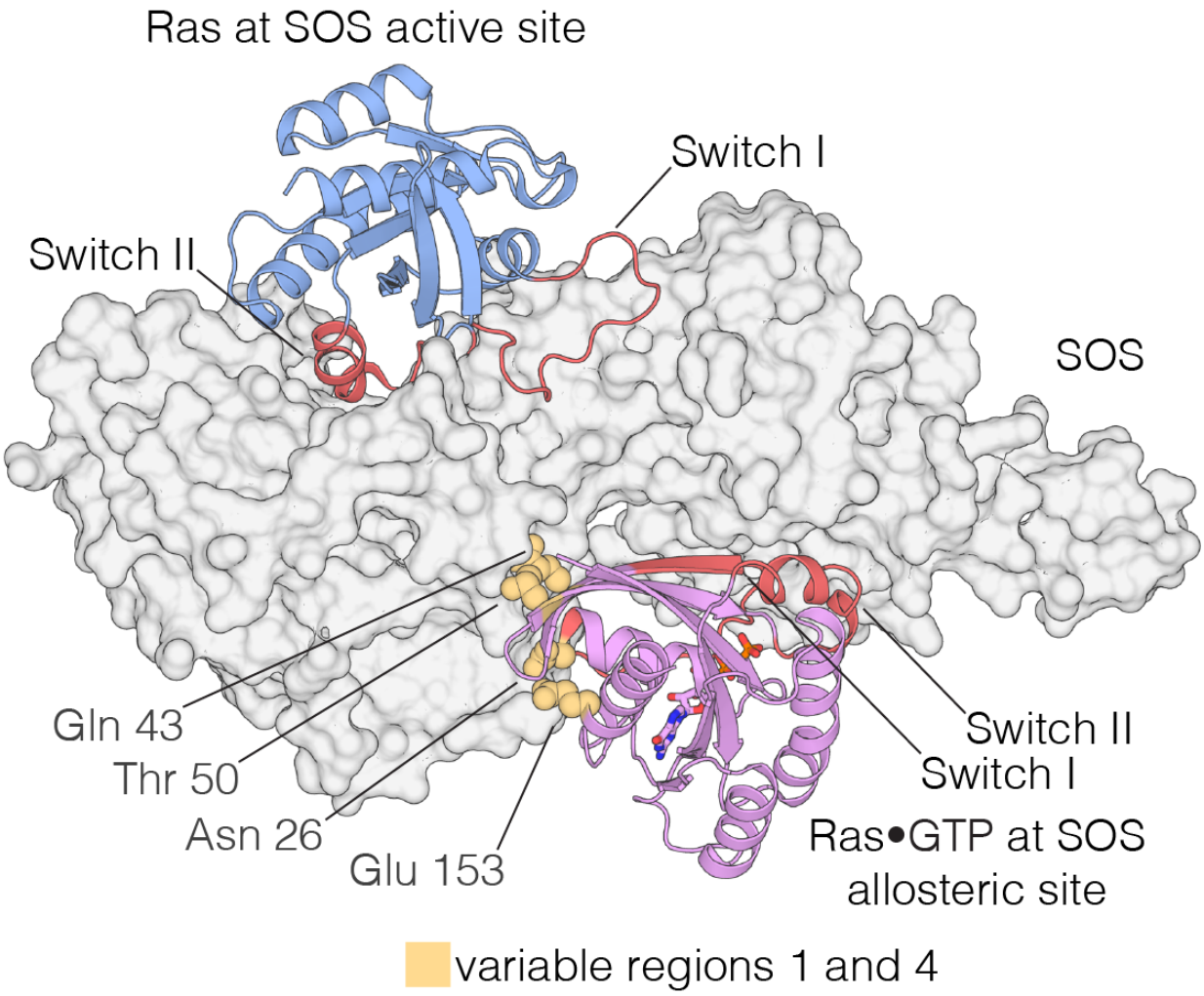

Figure 3. Interaction of Ras with the two Ras-binding sites of SOS. Structure of the Ras:SOS complex (PDB code: 1NVV). Two molecules of H-Ras are bound to the allosteric and active site of SOS. The allosteric site of SOS is bound by Ras•GTP, whereas the active site of SOS is bound by nucleotide-free Ras. Switch I and Switch II of Ras (red) are responsible for engaging both sites in SOS. Additionally, residues in variable regions 1 and 4 are involved in the binding of Ras to the allosteric site of SOS, including Asn 26, Gln 43, Thr 50, and Glu 153 (yellow).

Variable Regions 1 and 4 in human H-Ras interact directly with SOS at the allosteric site (Figure 3), and are altered in *S. rosetta* Ras and in the ancestral sequence (Figure 1). We expressed and purified Ras from the choanoflagellate *S. rosetta*. Human SOS stimulates nucleotide exchange with *S. rosetta* Ras and H-Ras with comparable rates, but only when GDP bound to Ras is exchanged for GDP in solution (Figure 4A). When the exchange reaction is carried out with excess GTP in solution, the reaction shows a marked acceleration in rate for human H-Ras, due to the allosteric stimulation of SOS by Ras•GTP. In contrast, *S. rosetta* Ras is deficient in allosterically stimulating the activity of human SOS (Figure 4B).

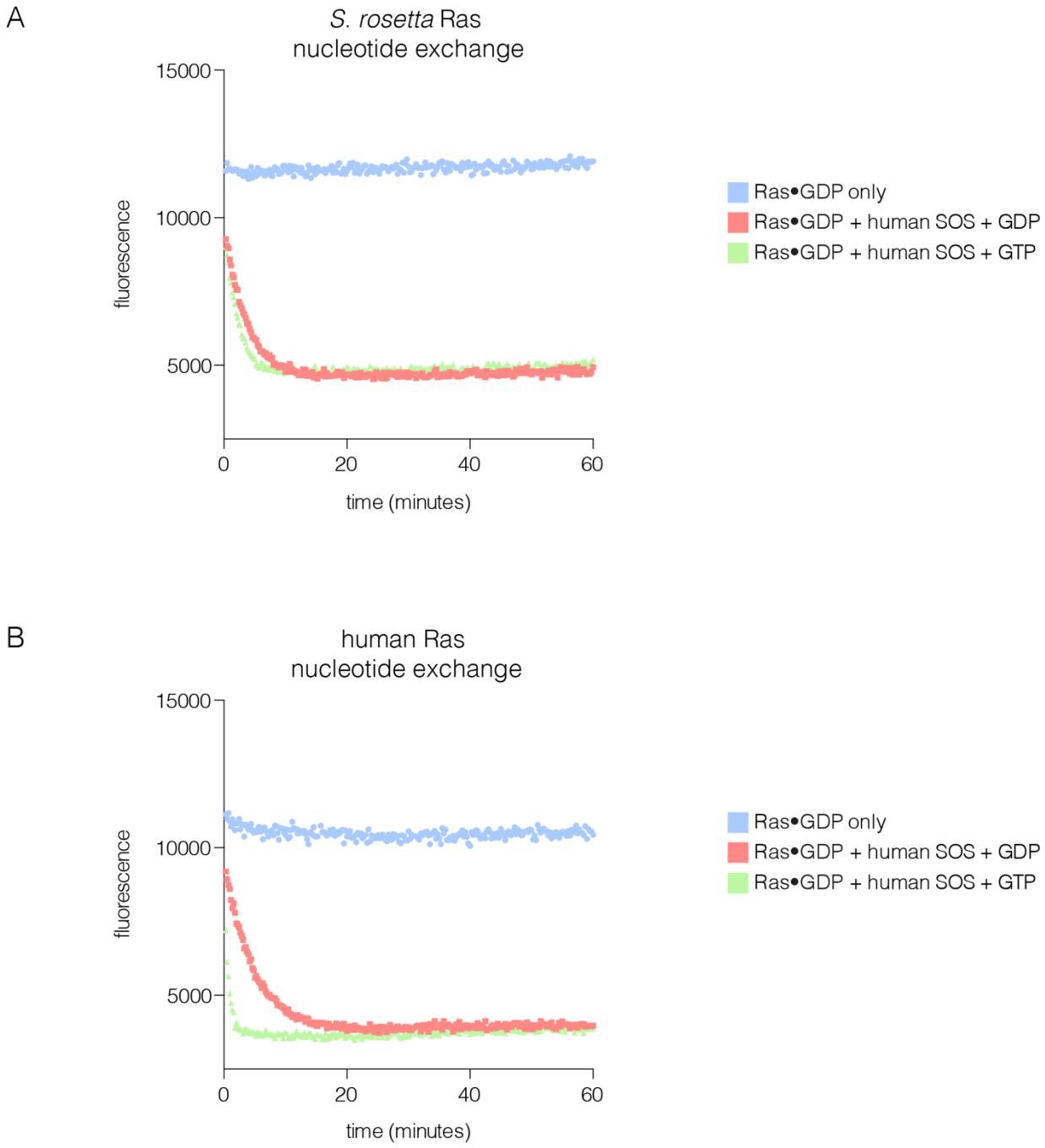

Figure 4. SOS-stimulated nucleotide exchange of *S. rosetta* and human Ras. **A.** Human SOS stimulates nucleotide exchange for *S. rosetta* Ras•GDP in the presence of excess GDP in solution, however, no increase in exchange rate is observed for *S. rosetta* Ras•GDP in the presence of excess GTP in solution (red and green curves, respectively). Intrinsic nucleotide release rate for *S. rosetta* Ras•GDP is shown in blue. **B.** Human SOS stimulates nucleotide exchange for human Ras•GDP in the presence of excess GDP in solution and an increase in exchange rate

is observed for human Ras•GDP in the presence of excess GTP in solution (red and green curves, respectively). This increase in exchange rate is indicative of human SOS activation by binding of Ras•GTP to the allosteric site. Intrinsic nucleotide release rate for human Ras•GDP is shown in blue.

Mattos and colleagues have described alternative conformations of Ras•GTP involving Variable Regions 2 and 3, and also Switch II (Figure 5A; the two states are shown schematically in Figure 5C) (Buhrman et al. 2010; Holzapfel et al. 2012). The two conformational states have been denoted "R" (assumed to be competent for GTP hydrolysis, although this has not been established definitively) and "T" (assumed to be inactive in terms of GTP hydrolysis). A physiologically relevant allosteric effector that controls the R to T transition has not yet been identified (Buhrman et al. 2010; O'Connor and Kovrigin 2012). The two Ras-binding sites in SOS differ in their recognition of the T and R states. SOS binds Ras in the T state conformation at the active site, considering just the conformation of helices α3 and α4, and Switch II (Switch I is displaced during the exchange process) (Figure 6A). In contrast, SOS recognizes Ras in the R state at the allosteric site (Figure 6B).

We found that the transition between the R and T states can be triggered by one of the hotspot mutations identified in our study, L120A (Figure 5B). We determined the structure of wild-type Ras•GTP and the L120A variant at room temperature, both crystallized with the same crystallographic lattice, and found that wild-type Ras•GTP is in the R state, whereas the L120A mutant is in the T state (there is evidence in the electron density for a minor population of the R state, with an occupancy of 20% or less).

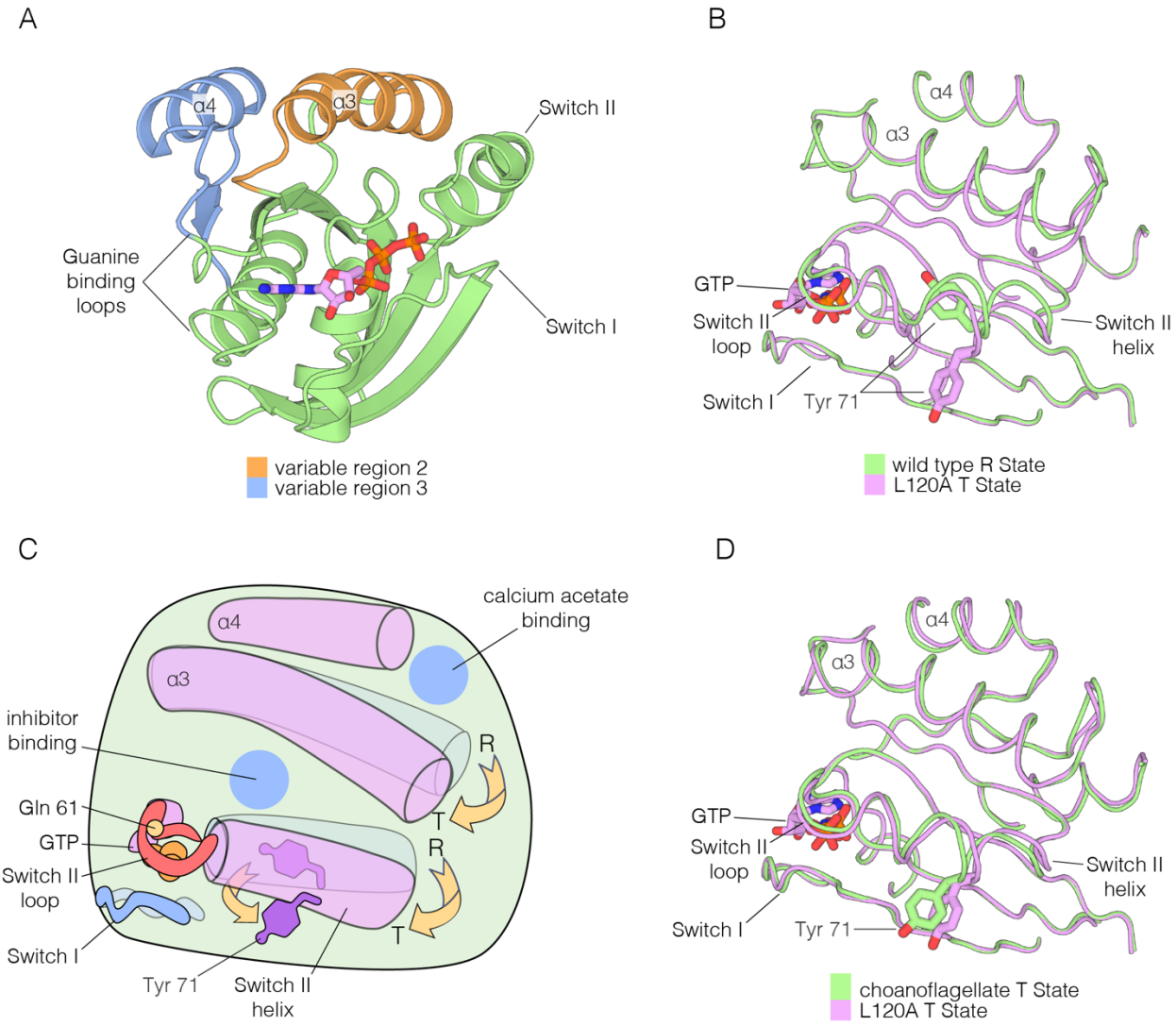

Figure 5. Comparison of the R and T states in Ras. A. Variable Regions 2 and 3 highlighted on the structure of Ras•GTP. Variable Region 2 comprises helix α3 and the preceding loop, and Variable Region 3 comprises helix α4 and the preceding loop that is partially involved in guanine binding. **B.** Room temperature crystal structure of wild-type Ras adopts the R state as defined by Mattos and colleagues (Buhrman et al. 2010; Holzapfel et al. 2012), where the sidechain of Tyr 71 is buried in the hydrophobic core of the protein. Crystal structure of the L120A latch mutant shows a rotation of Tyr 71, whereby the sidechain is exposed. Additionally, the interface between helix α3 and Switch II undergoes a conformational change. **C.** Schematic of the conformational transition between the R state and T state as defined by Mattos and colleagues (Buhrman et al. 2010; Holzapfel et al. 2012). Helix α3 and the C-terminal helix of Switch II shift downwards in the transition from the R state to the T state, leading to a rotation of Tyr 71 from a buried to an exposed conformation. **D.** Comparison of the choanoflagellate *S. rosetta* crystal structure to

the L120A structure. *S. rosetta* Ras adopts the T state, similar to L120A, where Tyr 71 is rotated outwards and exposed.

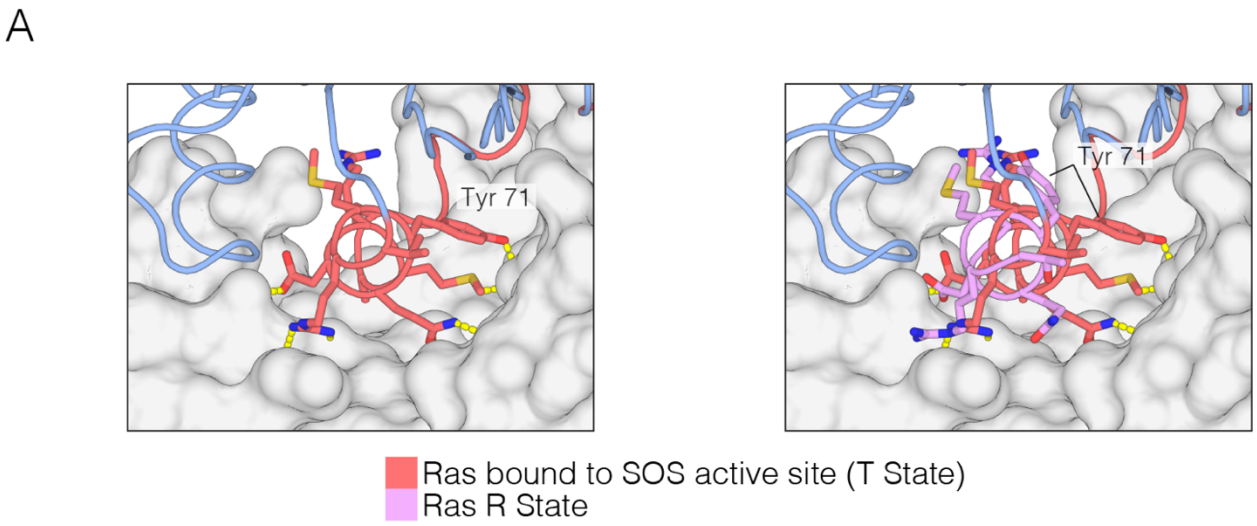

Figure 6. R and T states of Ras bound to SOS. A. Ras in the T state bound to the active site of SOS (left). Many sidechain contacts are present between the Switch II helix (red) and SOS (grey), including the sidechain of Tyr 71. Ras in the R state is incompatible with binding to the active site of SOS (right), and the sidechain of Tyr 71, in addition to others, does not form hydrogen bonding necessary for SOS binding. **B.** Ras in the R state bound to the allosteric site of SOS (left). The Switch II helix conformation is compatible with allosteric SOS binding. Ras in the T state is incompatible with binding to the allosteric site of SOS (right), and the sidechain of Tyr 71 clashes with SOS, thus leading to an incompatible binding interface.

There are no structures for an invertebrate Ras protein in the protein databank. We crystallized Ras from the choanoflagellate *S. rosetta*, and determined structures bound to GDP and GTP at resolutions of 1.85 Å and 1.6 Å, respectively (Figure 7). The conformation of Switch II and helix α3 in *S. rosetta* Ras•GTP is very close to that of the T state structure of human Ras (Figure 5D), suggesting that the balance between the R and T states of Ras•GTP may be altered in invertebrate Ras with respect to human Ras.

Comparison of human Ras to a non-vertebrate Ras and to Rap

The sequence divergence between invertebrate and vertebrate forms of Ras contrasts with the situation for its close relative Rap. Ras and Rap share 57% sequence identity between the human proteins. The sequence of Rap is more highly conserved than that of Ras throughout the metazoan lineage and the choanoflagellates, with 92% identity between Rap-1B in humans and the Rap homolog in *S. rosetta* (Figure 8). The differences between the human and *S. rosetta* Rap sequences are not localized to specific regions, as they are for Ras, pointing to the emergence of a unique selective pressure on Ras that led to the divergence of the four variable regions. Our data suggest that this divergence is correlated with the emergence of a positive feedback loop in the activation of Ras by SOS. This conformational balance may also affect interactions of Ras with as yet unknown signaling partners.

Data Collection					
	Human WT Ras.GMPPNP (RT)	Human L120A Ras.GMPPNP (LT)	Human L120A Ras.GMPPNP (RT)	Choano Ras.GTP (LT)	Choano Ras.GDP (LT)
Temperature	277K	100K	277K	100K	100K
Wavelength (Å)	0.88557	1.000	0.88557	1.11587	1.000
Space group	H32	H32	H32	$P2_1$	$P\ 4_3\ 2_1\ 2$
Cell dimensions a, b, c (Å) α, β, γ °	90.0 90.0 136.7 90 90 120	87.96, 87.96, 133.1 90.0, 90.0 120.0	90.0 90.0 136.0 90 90 120	38.6, 119.2, 38.6 90.0, 117.3, 90.0,	67.2, 67.2, 184.7 90, 90, 90
Resolution (Å) (high res shell)	45.58 - 1.65 (1.68 - 1.65)	50-1.25 (1.27-1.25)	45.32 - 1.35 (1.37 - 1.35)	39.70 - 1.60 (1.63 - 1.60)	47.53 - 1.85 (1.89 - 1.85)
Rmeas (%) Rpim (%)	17.9 (327.2) 4.0 (75.6)	4.5 (59.4) 1.0 (30.8)	9.9 (297.2) 3.2 (94.2)	9.8 (94.4) 4.7 (55.7)	9.3 (143.7) 2.7 (51.1)
$I/\sigma(I)$	12.6 (1.2)	36.9 (2.4)	16.2 (1.2)	9.5 (1.4)	12.5 (2.0)
$CC_{1/2}$	0.999 (0.514)	1.00 (0.790)	0.999 (0.565)	0.997 (0.428)	0.995 (0.923)
Completeness (%)	100 (100)	99.0 (92.0)	100 (100)	99.9 (99.8)	100.0 (100.0)
Multiplicity	19.8 (18.6)	18.7 (4.6)	19.7 (20.0)	4.5 (3.8)	9.9 (8.8)
Wilson B factor	17.4	11.5	14.6	12.1	25.7
Refinement					
Reflections used	25866 (2838)	54265 (2483)	46639 (4612)	40548 (2634)	37071 (2786)
Rfree reflections	1254 (121)	2731 (123)	2725 (132)	2065 (137)	1837 (135)
R_{work}/R_{free}	14.16/16.84 (25.26/28.43)	14.11/ 15.24 (23.49/25.27)	14.50/15.36 (27.20/26.87)	22.85/27.98 (30.00/36.71) 16.64/20.58* (26/29)*	21.62/24.90 (35.8/35.6)
Twinning fraction				0.47 for l, –k, h	
Number of non hydrogen atoms	1695	1782	1728	3010	3026
Protein	1530	1570	1578	2754	2778
Ligands	36	46	36	68	70
Average B factors					
Overall	28.39	21.65	26.7	20.0	41.9
Protein	27.80	20.15	26.4	18.8	41.7
Ligand	18.99	22.40	15.6	12.4	42.3
Solvent	38.0	35.62	34.7	22.8	45.5
Root mean square deviation from ideality					
Bonds (Å)	0.012	0.011	0.009	0.011	0.006
Angles (Å)	1.17	1.12	1.052	1.23	0.885
Ramachandran Statictics					
Favored (%)	98.8	96.95	97.6	98.8	98.8
Disallowed (%)	0.0	0.61	0.0	0.0	0.3
MolProbity clash score	3.57	4.7 2	2.21	3.25	3.02

Figure 7. Crystallographic data collection and refinement statistics.

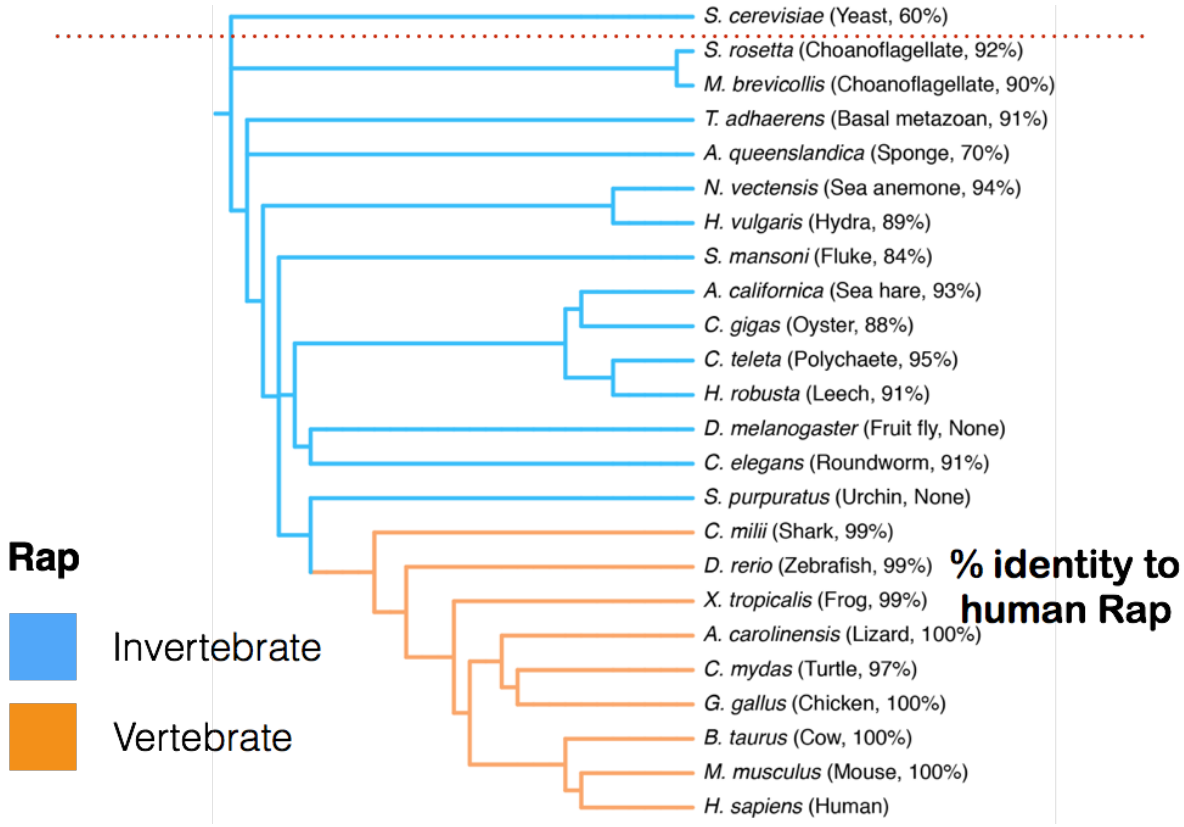

Figure 8. Comparison of Rap sequences. Note that Rap is also highly conserved across the metazoan lineage. Invertebrate Rap sequences, such as choanoflagellate or sponge, are greater than ~90% identical to human Rap. Only the sequence of the GTPase domain (residues 1 to 166 in human Ras) was used in this analysis.

Summary

In this chapter, I have described a comparison between extant sequences of Ras in the metazoan lineage and the experimentally derived accessible sequence space from the screen of the saturation point-mutagenesis library of Ras variants. This comparison revealed four variable regions of sequence space, all of which were also present in a computational reconstruction of a hypothetical ancestral Ras protein that preceded both invertebrate and vertebrate Ras sequences. Sequence variation in these four regions was largely present in invertebrate Ras sequences, and vertebrate Ras sequences were largely conserved in these four regions. We speculated that a new mode of regulation had emerged during the course of Ras evolution, and showed that residues in the four variable regions are involved in the recognition of the guanine nucleotide exchange factor SOS. Furthermore, structural studies of a choanoflagellate Ras have revealed a role for residues in the variable regions of sequence space in potentially controlling an allosteric transition that dictates whether Ras and SOS can mutually activate each other. Interestingly, an activating hotspot mutation in Ras, L120A, appears to trigger this conformational

transition. Further functional and biochemical experiments will be needed to study this conformational transition in Ras, and it will also be interesting to study these effects on the membrane, especially with Ras and SOS orthologs.

Methods

Ancestral sequence reconstruction

We obtained an initial set of 21,243 Ras superfamily (PF00071) sequences from the Pfam protein families database (Finn et al. 2016). Out of the 8,988 metazoan sequences in the alignment, we selected a set of 72 Ras sequences for evolutionary analysis, with yeast Ras added as an outgroup. The corresponding maximum likelihood phylogenetic tree and sequence alignment were jointly inferred with PASTA (Mirarab et al. 2015). We used four iterative alignment and phylogenetic inference steps. RAxML (Stamatakis 2014) was used as the tree estimator with the PROTCATLGF model, which employs the LG amino acid substitution matrix (Le and Gascuel, 2008) and optimizes site-specific substitution rates. PASTA likelihood scores indicated convergence of the alignment. Ancestral sequences were then reconstructed via maximum likelihood with FastML (Ashkenazy et al. 2012) using the same substitution matrix, with the phylogenetic tree and alignment from PASTA as input. Most residues in the hypothetical metazoan ancestral sequence were reconstructed by FastML with high confidence (median probability of reconstructed amino acids was 97%; only 10 of 178 residues had probabilities below 50%).

We also tested our results on the effect of substitutions in the reconstructed ancestral Ras sequence for robustness to uncertainty in the reconstructed sequence. Averaging over the probability of each reconstructed residue at each site, we found that the average number of substitutions in the ancestral sequence relative to human H-Ras was 48.5 (compared with 48 for the most likely sequence). The average number of substitutions that would increase/have little effect on/decrease unregulated Ras function was 30.5/9.0/9.0, respectively, in good agreement with 30/8/9 for the most likely sequence. Similarly, for regulated Ras the average numbers were 9.2/2.9/36.4, compared with 8/3/36 for the most likely sequence. Locations of likely substitutions were also consistent with the patterns displayed in the most likely ancestral sequence shown in Figure 8C. Thus, we find that uncertainty in the amino acid composition of the reconstructed ancestral sequence does not affect our findings about the predicted effects of substitutions relative to the human H-Ras sequence.

Crystallization and structure determination of Ras variants

The *S. rosetta* cDNA library was a kind gift from Nicole King (Fairclough et al., 2013). The coding region for the catalytic domain of *S. rosetta* Ras residues 1-167 was amplified from the cDNA library and cloned into pProEX HTb expression vector using the NcoI – XhoI restriction sites. Human H-Ras variants were cloned into a pET-28b expression vector modified with a 5' His6-tag and a TEV protease site coding sequence, and variants were generated by Quikchange site-directed mutagenesis (Agilent).

Crystals of wild-type human H-Ras bound to GMP-PNP were grown by mixing 1 µL of protein solution (17 mg/mL) with 1 µL of reservoir buffer (16% PEG3500, 100 mM calcium acetate, pH 7.5). Human H-Ras mutant L120A bound to GMP-PNP was crystallized by

mixing 1 µL of protein solution (40 mg/mL) with 1 µL of reservoir buffer (17% PEG8000, 200 mM calcium acetate, 100 mM MES/NaOH, pH 6.0). Drops were allowed to equilibrate in a sitting drop tray over 500 µL reservoir solution for one week at 20°C.

Crystals of *S. rosetta* Ras with GDP were obtained by mixing 2 µL of 30 mg/mL protein with 2 µL of the reservoir buffer (20% PEG3350, 100 mM Tris-HCl pH9.0, 100 mM $CaCl_2$) at 20°C. Crystals were cryo-protected by adding the cryo-protection buffer (20% PEG3350, 100 mM Tris-HCl pH 9.0, 100 mM $CaCl_2$, 15% glycerol) in one step, and flash frozen in liquid nitrogen.

The GMP-PNP bound *S. rosetta* Ras was crystallized by mixing 1 µL of 18 mg/mL protein with 1 µL of the reservoir buffer (15% PEG8000, 100 mM Bis-Tris pH 5.5, 200 mM magnesium acetate, 50 mM $CaCl_2$) at 20°C. The best crystals were obtained by streak-seeding. Crystals were added the cryo-protection buffer containing (15% PEG8000, 100 mM Bis-Tris pH 5.5, 200 mM magnesium acetate, 50 mM $CaCl_2$, 12.5 % ethylene glycol) before being flash-frozen in liquid nitrogen.

Preliminary diffraction data for room-temperature wild-type human H-Ras were collected at the Stanford Synchrotron Radiation Lightsource, beamline 9-2. Final diffraction data presented herein were collected at the Advanced Light Source (ALS), beamline 8.3.1. For low temperature data, crystals were cryo-protected with reservoir buffer supplemented with 20% glycerol and flash cooled with liquid nitrogen. Low temperature data were collected at 100 K on ALS beamline 8.2.1. For room temperature data, crystals were looped directly from the sitting drop tray into polyimide mounts (MiTeGen MicroMounts), covered with a polyester sleeve containing a small amount of reservoir solution in the tip, and immediately placed in a cryostream at 277 K and irradiated with a defocused beam of 14 keV X-rays. Data were integrated with XDS (Kabsch 2010) and scaled and reduced to the asymmetric unit with Aimless in the CCP4 suite (Winn et al. 2011). Initial structures for both wild-type and mutant Ras were produced with phases from a structure of calcium acetate-bound Ras in the same space group (PDB code 3K8Y; (Buhrman et al. 2010)). The model was rebuilt using Coot (Emsley and Cowtan 2004) and iteratively refined with Phenix (Adams et al. 2010). These programs were provided through SBGRID (Morin et al. 2013). Positive difference electron density around the modeled magnesium ion associated with bound GMP-PNP was observed in both L120A structures. Considerable anomalous signal was also observed at this position. The ion position was therefore refined using Phenix with both metals present, with variable occupancy. The observed density was best accounted for with 70% calcium ion and 30% magnesium ion.

Diffraction data for *S. rosetta* Ras were collected at ALS beamline 8.2.1 and 8.3.1. Data were integrated in XDS (Kabsch 2010) and scaled with Scala (Evans 2006) or Aimless. For GDP-bound *S. rosetta* Ras, a molecular replacement solution was found in Phaser (McCoy et al. 2007) using the H-Ras structure (PDB ID: 5P21; (Pai et al. 1990)). The structural model was rebuilt using Coot (Emsley and Cowtan 2004) and refined by Refmac (Murshudov et al. 2011) and Phenix. The model was used to find a molecular replacement solution for GMP-PNP-bound *S. rosetta* Ras. The data were processed in space group $C222_1$ and molecular replacement was successful, but the model could not be refined. We used Zanuda (Lebedev and Isupov 2014) to help resolve the space group as $P2_1$. The crystal was pseudo-merohedrally twinned (twin fraction estimated at 0.47) (Hamdane et al. 2009). The

structure was successfully rebuilt and refined in $P2_1$ with Coot and Phenix, without accounting for the apparent twinning. Table 1 reports the final model R-factors calculated in Phenix without accounting for twinning and for the same model (no further refinement) calculated with Refmac using the twin operator l, -k, h.

References

Adams, P.D., Afonine, P.V., Bunkóczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.-W., Kapral, G.J., Grosse-Kunstleve, R.W., McCoy, A.J., Moriarty, N.W., Oeffner, R., Read, R.J., Richardson, D.C., Richardson, J.S., Terwilliger, T.C. and Zwart, P.H. 2010. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica. Sect. D, Biological Crystallography* 66(Pt 2), pp. 213–221.

Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O. and Pupko, T. 2012. FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research* 40(Web Server issue), pp. W580–4.

Boriack-Sjodin, P.A., Margarit, S.M., Bar-Sagi, D. and Kuriyan, J. 1998. The structural basis of the activation of Ras by Sos. *Nature* 394(6691), pp. 337–343.

Buhrman, G., Holzapfel, G., Fetics, S. and Mattos, C. 2010. Allosteric modulation of Ras positions Q61 for a direct role in catalysis. *Proceedings of the National Academy of Sciences of the United States of America* 107(11), pp. 4931–4936.

Das, J., Ho, M., Zikherman, J., Govern, C., Yang, M., Weiss, A., Chakraborty, A.K. and Roose, J.P. 2009. Digital signaling and hysteresis characterize ras activation in lymphoid cells. *Cell* 136(2), pp. 337–351.

Emsley, P. and Cowtan, K. 2004. Coot: model-building tools for molecular graphics. *Acta Crystallographica. Sect. D, Biological Crystallography* 60(Pt 12 Pt 1), pp. 2126–2132.

Evans, P. 2006. Scaling and assessment of data quality. *Acta Crystallographica. Sect. D, Biological Crystallography* 62(Pt 1), pp. 72–82.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J. and Bateman, A. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* 44(D1), pp. D279–85.

Gureasko, J., Galush, W.J., Boykevisch, S., Sondermann, H., Bar-Sagi, D., Groves, J.T. and Kuriyan, J. 2008. Membrane-dependent signal integration by the Ras activator Son of sevenless. *Nature Structural & Molecular Biology* 15(5), pp. 452–461.

Hamdane, D., Lechauve, C., Marden, M.C. and Golinelli-Pimpaneau, B. 2009. Pseudo-merohedral twinning in monoclinic crystals of wild-type human brain neuroglobin. *Acta Crystallographica. Sect. D, Biological Crystallography* 65(Pt 4), pp. 388–392.

Holzapfel, G., Buhrman, G. and Mattos, C. 2012. Shift in the equilibrium between on and

off states of the allosteric switch in Ras-GppNHp affected by small molecules and bulk solvent composition. *Biochemistry* 51(31), pp. 6114–6126.

Iversen, L., Tu, H.-L., Lin, W.-C., Christensen, S.M., Abel, S.M., Iwig, J., Wu, H.-J., Gureasko, J., Rhodes, C., Petit, R.S., Hansen, S.D., Thill, P., Yu, C.-H., Stamou, D., Chakraborty, A.K., Kuriyan, J. and Groves, J.T. 2014. Molecular kinetics. Ras activation by SOS: allosteric regulation by altered fluctuation dynamics. *Science* 345(6192), pp. 50–54.

Kabsch, W. 2010. XDS. *Acta Crystallographica. Sect. D, Biological Crystallography* 66(Pt 2), pp. 125–132.

Lebedev, A.A. and Isupov, M.N. 2014. Space-group and origin ambiguity in macromolecular structures with pseudo-symmetry and its treatment with the program Zanuda. *Acta Crystallographica. Sect. D, Biological Crystallography* 70(Pt 9), pp. 2430–2443.

Margarit, S.M., Sondermann, H., Hall, B.E., Nagar, B., Hoelz, A., Pirruccello, M., Bar-Sagi, D. and Kuriyan, J. 2003. Structural evidence for feedback activation by Ras.GTP of the Ras-specific nucleotide exchange factor SOS. *Cell* 112(5), pp. 685–695.

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C. and Read, R.J. 2007. Phaser crystallographic software. *Journal of Applied Crystallography* 40(Pt 4), pp. 658–674.

Mirarab, S., Nguyen, N., Guo, S., Wang, L.-S., Kim, J. and Warnow, T. 2015. PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences. *Journal of Computational Biology* 22(5), pp. 377–386.

Morin, A., Eisenbraun, B., Key, J., Sanschagrin, P.C., Timony, M.A., Ottaviano, M. and Sliz, P. 2013. Collaboration gets the most out of software. *eLife* 2, p. e01456.

Murshudov, G.N., Skubák, P., Lebedev, A.A., Pannu, N.S., Steiner, R.A., Nicholls, R.A., Winn, M.D., Long, F. and Vagin, A.A. 2011. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica. Sect. D, Biological Crystallography* 67(Pt 4), pp. 355–367.

O'Connor, C. and Kovrigin, E.L. 2012. Characterization of the second ion-binding site in the G domain of H-Ras. *Biochemistry* 51(48), pp. 9638–9646.

Pai, E.F., Krengel, U., Petsko, G.A., Goody, R.S., Kabsch, W. and Wittinghofer, A. 1990. Refined crystal structure of the triphosphate conformation of H-ras p21 at 1.35 A resolution: implications for the mechanism of GTP hydrolysis. *The EMBO Journal* 9(8), pp. 2351–2359.

Richter, D.J. and King, N. 2013. The genomic and cellular foundations of animal origins. *Annual Review of Genetics* 47, pp. 509–537.

Sondermann, H., Soisson, S.M., Boykevisch, S., Yang, S.-S., Bar-Sagi, D. and Kuriyan, J. 2004. Structural analysis of autoinhibition in the Ras activator Son of sevenless. *Cell* 119(3), pp. 393–405.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9), pp. 1312–1313.

Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G.W., McCoy, A., McNicholas, S.J., Murshudov, G.N., Pannu, N.S., Potterton, E.A., Powell, H.R., Read, R.J., Vagin, A. and Wilson, K.S. 2011. Overview of the CCP4 suite and current developments. *Acta Crystallographica. Sect. D, Biological Crystallography* 67(Pt 4), pp. 235–242.

Chapter 6: Concluding remarks

The mutational robustness of Ras

In this thesis I have adapted a bacterial two-hybrid system to analyze how mutations affect the functional cycle of human H-Ras. Our strategy was to isolate just the minimal biochemical network that defines this cycle, comprising Ras, its effector Raf, and the GAP and the GEF. Thus, we can analyze the sensitivity of Ras to mutation in the context of this network, while excluding the effects of the membrane and additional regulatory factors. This approach provides an opportunity, for the first time, to use deep mutational scanning approaches to study how local regulatory networks influence the mutational sensitivity and phenotypic plasticity of key signaling molecules.

In the absence of regulators, the distribution of fitness effects of mutations in Ras shows considerable tolerance to mutation, as seen previously in saturation mutagenesis experiments on other proteins (McLaughlin et al. 2012; Soskine and Tawfik 2010; Boucher et al. 2014; Podgornaia and Laub 2015)). However, in the presence of a GAP and a GEF, Ras displays global constraints, whereby the majority of mutations lead to a modest decrease in function. The data for attenuated-Ras, in the presence of the GAP alone, are similar to that for fully regulated Ras, except that under these conditions Ras is very sensitive to mutations that disrupt the GTPase activity, which cause a dramatic gain of function. Taken together, our data show that the local regulatory network places a stringent constraint on the sequence of Ras, and also creates the potential conditions in which it is susceptible to activating mutations. These data extend previous observations that mutational sensitivity in proteins is strongly dependent on the selective conditions in which the protein operates (Stiffler et al. 2015). Figure 1 depicts the overall mutational sensitivity of Ras, when both gain-of-function and loss-of-function mutations are taken into account from each selection experiment. A deeper understanding of the relationship between protein structure and function will emerge as saturation mutagenesis experiments are done in more complete biological contexts.

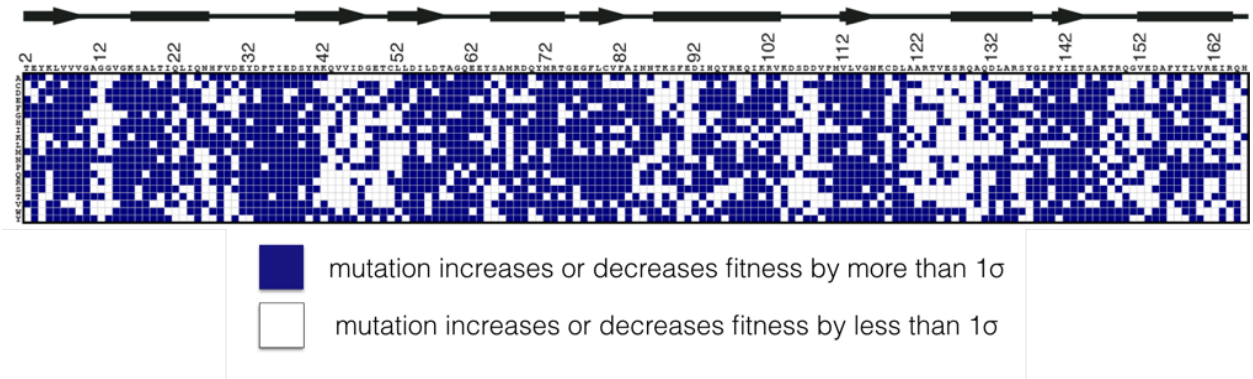

Figure 1. The mutational robustness of Ras. Statistically significant mutations that either increase or decrease the fitness of Ras are depicted in blue across the entire primary sequence. Most of the protein is sensitive to mutation, and importantly, mutational sensitivity is context-dependent based on the presence or absence of components of the regulatory cycle.

Extant sequence variation of Ras proteins

The structural fold of Ras is intrinsically capable of accommodating sequence changes that lead to the acquisition of new function. For example, paralogous members of the Ras superfamily show a high degree of structural conservation within the core elements of the GTPase domain despite considerable sequence divergence (Rojas et al. 2012). In contrast, orthologous members of the family (such as the Ras proteins) show high sequence conservation across metazoan evolution. We propose that a critical aspect of function of these proteins is the tuning of the GTPase cycle such that switching between active and inactive states and responsiveness to regulators is matched to its specific biological context. More generally, we suggest that high conservation in orthologs of a protein family may arise from the pressures imposed by the dynamics of local biochemical environments.

Evolution of the interdependent activation of Ras and SOS

Ras itself has undergone a further specialization in the transition from invertebrates to vertebrates. Our analysis indicates that this specialization is associated with the acquisition, or modification, of allosteric control, as indicated by the fact that the majority of these sequence changes involve residues that lead to activation when mutated in human Ras. One such control mechanism is the ability of Ras to allosterically activate its own activator, SOS, a property that may have emerged in higher vertebrates. Additional regulation, involving interactions at the membrane (Abankwa et al. 2008; Mazhab-Jafari et al. 2015), the potential dimerization of Ras (Muratcioglu et al. 2015), and interaction with different effector proteins, may also play a role in the emergence of the new levels of allosteric control.

An evolutionary innovation in metazoans is the coupling of Ras activation to the phosphorylation of tyrosine residues on cell-surface receptors. This occurs through the SH2-mediated recruitment to these receptors of a Ras-specific guanine nucleotide exchange factor (GEF), Son-of-Sevenless (SOS) (Lowenstein et al. 1992; Gale et al. 1993; Buday and Downward 1993; Egan et al. 1993; Rozakis-Adcock et al. 1993). There are several classes of Ras-specific GEFs in mammalian cells, and a key difference between SOS and other Ras-specific GEFs is that the latter are regulated by soluble second messengers, such as calcium, cyclic AMP (cAMP), and diacylglycerol, whereas SOS is not (Bos et al. 2007). Instead, SOS is recruited to the membrane by tyrosine phosphorylation of receptors or scaffold proteins and, in addition, the activation of SOS requires the binding of Ras•GTP to an allosteric site (Margarit et al. 2003).

SOS is the only Ras-specific GEF that is known to be allosterically activated by Ras•GTP, pointing to an evolutionary divergence between SOS and other Ras-specific GEFs. Insight into the unique dependence of SOS on Ras may be gained from analyzing the sequence conservation of Ras in metazoans, relative to other Ras family GTPases. The three principal isoforms of Ras (H-Ras, K-Ras and N-Ras) are generally highly conserved in metazoans (Rojas et al. 2012; Johnson et al. 2017), though less so in invertebrates compared to vertebrates (Bandaru et al. 2017). The sequence of Ras from the choanoflagellate *S. rosetta*, thought to be the closest living relative of metazoans, is 72% identical to human Ras, whereas Ras is nearly invariant (>90% identical to human) in vertebrates. In comparison, the sequence of the Ras-family GTPase, Rap, in *S. rosetta* is 92% identical to

that of human Rap, indicating that the selective pressures on Rap did not change as much in the transition into the vertebrate lineage. It is intriguing to note that the high conservation of Ras in vertebrates can possibly be traced to interactions with SOS, which may have been a factor that constrained the evolution of Ras (Bandaru et al. 2017).

Buffering gain-of-function mutations in natural signaling systems

An interesting dichotomy exists between the number of known oncogenic mutations in Ras and the number of activating mutations observed in the data. One explanation for this dichotomy stems from the notion that evolution has imposed negative selective pressure on the primary sequence of Ras, and perhaps many other metazoan signaling proteins, in their basal, inactive states. In the case of Ras, this evolutionary design principle prevents activation at residues containing gain-of-function mutations through reduction of nucleotide hydrolysis rates and increase of nucleotide release rates. The conservation of these regions of primary sequence is significantly higher in Ras orthologs than in its paralogs, suggesting that Ras has evolved a unique autoinhibitory mechanism to tightly regulate its self-activation in comparison to related small GTPases.

Another explanation for this dichotomy is that natural signaling systems can suppress spontaneous Ras activation is through regulation by the GAP and GEF. The suppression of spontaneous Ras activation is demonstrated in the case of nucleotide release rate: unless the rate of intrinsic nucleotide release dramatically increases the rate of GEF mediated nucleotide exchange, the rate increase of intrinsic nucleotide release will largely be unnoticed. This provides a conceptual explanation for the transforming effect of Lys 117 Ras mutants, where both intrinsic nucleotide release and GEF-mediated exchange are significantly increased. If activating mutations in Ras led to unwanted enzymatic activity and functional effects, then regulation may have evolved as a general mechanism to buffer the effects of activating mutations (Soskine and Tawfik 2010).

Implications for cancer

Oncogenic mutations that disturb the switching mechanisms of Ras result in aberrant signaling and cancer, highlighted by the fact that Ras is one of the most important proto-oncogenes in the human genome (Young et al. 2009; Prior et al. 2012). An important result from our experiments that merits further study is that oncogenic mutations, such as G12V, are conditionally neutral (see Chapter 3). That is, such mutations are minimally disruptive in the presence of the regulators, which may allow these mutations to appear and float in the population, manifesting a dramatic gain of function only upon perturbation to the regulators of the Ras switching cycle. Thus, normal physiology may provide evolutionary paths to acquiring neutral mutations that, with a second hit such as loss of regulation, can display pathological activities. Such mutations could conceptually be analogous to "latent driver" mutations that emerge prior to and during tumor evolution (Nussinov and Tsai 2015).

The extensive experience with cancer drugs that target protein kinases has shown that resistance to these drugs emerges quickly, often through mutation at sites of allosteric communication with the drug-binding site (Azam et al. 2003; Shah et al. 2002; Branford et al. 2009). Small molecule inhibitors of Ras have yet to achieve clinical relevance, but there

is a concerted effort underway to obtain such inhibitors (McCormick 2016). Experience with kinase inhibitors suggests that the clinical introduction of Ras inhibitors will be followed by the rapid acquisition of resistance mutations in patients, necessitating further cycles of inhibitor development to overcome resistance. Our analysis of the mutational sensitivity of Ras-G12V has indicated that there are epistatic interactions that reduce the sensitivity to mutation of other elements of the structure, such as Switch II. The mapping of such epistatic interactions for other oncogenic mutations will provide an important conceptual framework for understanding the response of Ras to inhibition by small molecules.

Future directions

As the technology around the high-throughput interrogation of protein structure-function relationships matures, several lines of future research directions appear compelling, particularly for signaling proteins. First, the ability to synthesize entire genes in a large scale for low cost enables entire protein phylogenies to be created and screened. This would be particularly interesting for Ras, as the sequence constraints on extant Ras sequences can be studied in the context of the bacterial two-hybrid screen. Along these lines, the mutational landscape of the hypothetical ancestral Ras protein can be studied, and the expectation would be that this protein is more susceptible to gain-of-function mutations than human Ras.

Second, epistasis between residues in Ras can be studied by means of creating double mutant libraries to be screened in the bacterial two-hybrid assay. The contiguous spatial distribution of residues susceptible to conditional gain-of-function mutations around the active site of Ras strongly suggests cooperativity among these residues. The extent of this epistasis can be determined by double mutant libraries of a stretch of Ras containing these residues, such as the $\alpha 3$ helix. Furthermore, the additivity of mutational effects can be studied in Ras, and this may be particularly interesting for "neutral" regions of Ras that appear not to elicit a functional effect upon making single mutations. It would be particularly interesting to study epistatic effects in the context of a mammalian cell selection system such as Ba/F3 cells. A potential platform for engineering higher order mutations into Ras, while still maintaining the modular bacterial two-hybrid system, involves designing a variation of the PACE system, which enables the rapid creation of large libraries of proteins that would be otherwise challenging to design (Badran et al. 2016).

Third, the work in this thesis has implications for protein design, particularly in the design of molecular switches. In fact, orthogonal G-protein and effector pairs have been the targets of computational protein design in prior studies (Kapp et al. 2012). Most protein design algorithms optimize the packing of amino acids in proteins, and often are unable to take into account long range allosteric effects when designing or optimizing functional properties of proteins (Bradley et al. 2005; Kuhlman et al. 2003; Schymkowitz et al. 2005). A recent development in computational protein design with Rosetta, termed "multi-state design", involves taking into account constraints imposed by binding interfaces that are involved multiple protein-protein interactions in order to encode these functions into a newly designed protein (Leaver-Fay et al. 2011; Warszawski et al. 2014). These design algorithms in theory can also take into account accessible mutational space as determined

by saturation mutagenesis experiments, which would be a step forward in the field of computational protein design.

References

Abankwa, D., Hanzal-Bayer, M., Ariotti, N., Plowman, S.J., Gorfe, A.A., Parton, R.G., McCammon, J.A. and Hancock, J.F. 2008. A novel switch region regulates H-ras membrane orientation and signal output. *The EMBO Journal* 27(5), pp. 727–735.

Azam, M., Latek, R.R. and Daley, G.Q. 2003. Mechanisms of autoinhibition and STI-571/imatinib resistance revealed by mutagenesis of BCR-ABL. *Cell* 112(6), pp. 831–843.

Badran, A.H., Guzov, V.M., Huai, Q., Kemp, M.M., Vishwanath, P., Kain, W., Nance, A.M., Evdokimov, A., Moshiri, F., Turner, K.H., Wang, P., Malvar, T. and Liu, D.R. 2016. Continuous evolution of Bacillus thuringiensis toxins overcomes insect resistance. *Nature* 533(7601), pp. 58–63.

Bandaru, P., Shah, N.H., Bhattacharyya, M., Barton, J.P., Kondo, Y., Cofsky, J.C., Gee, C.L., Chakraborty, A.K., Kortemme, T., Ranganathan, R. and Kuriyan, J. 2017. Deconstruction of the Ras switching cycle through saturation mutagenesis. *eLife* 6.

Bos, J.L., Rehmann, H. and Wittinghofer, A. 2007. GEFs and GAPs: critical elements in the control of small G proteins. *Cell* 129(5), pp. 865–877.

Boucher, J.I., Cote, P., Flynn, J., Jiang, L., Laban, A., Mishra, P., Roscoe, B.P. and Bolon, D.N.A. 2014. Viewing protein fitness landscapes through a next-gen lens. *Genetics* 198(2), pp. 461–471.

Bradley, P., Misura, K.M.S. and Baker, D. 2005. Toward high-resolution de novo structure prediction for small proteins. *Science* 309(5742), pp. 1868–1871.

Branford, S., Melo, J.V. and Hughes, T.P. 2009. Selecting optimal second-line tyrosine kinase inhibitor therapy for chronic myeloid leukemia patients after imatinib failure: does the BCR-ABL mutation status really matter? *Blood* 114(27), pp. 5426–5435.

Buday, L. and Downward, J. 1993. Epidermal growth factor regulates p21ras through the formation of a complex of receptor, Grb2 adapter protein, and Sos nucleotide exchange factor. *Cell* 73(3), pp. 611–620.

Egan, S.E., Giddings, B.W., Brooks, M.W., Buday, L., Sizeland, A.M. and Weinberg, R.A. 1993. Association of Sos Ras exchange protein with Grb2 is implicated in tyrosine kinase signal transduction and transformation. *Nature* 363(6424), pp. 45–51.

Gale, N.W., Kaplan, S., Lowenstein, E.J., Schlessinger, J. and Bar-Sagi, D. 1993. Grb2 mediates the EGF-dependent activation of guanine nucleotide exchange on Ras. *Nature* 363(6424), pp. 88–92.

Johnson, C.W., Reid, D., Parker, J.A., Salter, S., Knihtila, R., Kuzmic, P. and Mattos, C. 2017. The small GTPases K-Ras, N-Ras and H-Ras have distinct biochemical properties determined by allosteric effects. *The Journal of Biological Chemistry*.
Kapp, G.T., Liu, S., Stein, A., Wong, D.T., Reményi, A., Yeh, B.J., Fraser, J.S., Taunton, J., Lim, W.A. and Kortemme, T. 2012. Control of protein signaling using a computationally designed GTPase/GEF orthogonal pair. *Proceedings of the National Academy of Sciences of the United States of America* 109(14), pp. 5277–5282.

Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L. and Baker, D. 2003. Design of a novel globular protein fold with atomic-level accuracy. *Science* 302(5649), pp. 1364–1368.

Leaver-Fay, A., Jacak, R., Stranges, P.B. and Kuhlman, B. 2011. A generic program for multistate protein design. *Plos One* 6(7), p. e20937.

Lowenstein, E.J., Daly, R.J., Batzer, A.G., Li, W., Margolis, B., Lammers, R., Ullrich, A., Skolnik, E.Y., Bar-Sagi, D. and Schlessinger, J. 1992. The SH2 and SH3 domain-containing protein GRB2 links receptor tyrosine kinases to ras signaling. *Cell* 70(3), pp. 431–442.

Margarit, S.M., Sondermann, H., Hall, B.E., Nagar, B., Hoelz, A., Pirruccello, M., Bar-Sagi, D. and Kuriyan, J. 2003. Structural evidence for feedback activation by Ras.GTP of the Ras-specific nucleotide exchange factor SOS. *Cell* 112(5), pp. 685–695.

Mazhab-Jafari, M.T., Marshall, C.B., Smith, M.J., Gasmi-Seabrook, G.M.C., Stathopulos, P.B., Inagaki, F., Kay, L.E., Neel, B.G. and Ikura, M. 2015. Oncogenic and RASopathy-associated K-RAS mutations relieve membrane-dependent occlusion of the effector-binding site. *Proceedings of the National Academy of Sciences of the United States of America* 112(21), pp. 6625–6630.

McCormick, F. 2016. K-Ras protein as a drug target. *Journal of Molecular Medicine* 94(3), pp. 253–258.

McLaughlin, R.N., Poelwijk, F.J., Raman, A., Gosal, W.S. and Ranganathan, R. 2012. The spatial architecture of protein function and adaptation. *Nature* 491(7422), pp. 138–142.

Muratcioglu, S., Chavan, T.S., Freed, B.C., Jang, H., Khavrutskii, L., Freed, R.N., Dyba, M.A., Stefanisko, K., Tarasov, S.G., Gursoy, A., Keskin, O., Tarasova, N.I., Gaponenko, V. and Nussinov, R. 2015. GTP-Dependent K-Ras Dimerization. *Structure* 23(7), pp. 1325–1335.

Nussinov, R. and Tsai, C.-J. 2015. “Latent drivers” expand the cancer mutational landscape. *Current Opinion in Structural Biology* 32, pp. 25–32.

Podgornaia, A.I. and Laub, M.T. 2015. Protein evolution. Pervasive degeneracy and epistasis

in a protein-protein interface. *Science* 347(6222), pp. 673–677.

Prior, I.A., Lewis, P.D. and Mattos, C. 2012. A comprehensive survey of Ras mutations in cancer. *Cancer Research* 72(10), pp. 2457–2467.

Rojas, A.M., Fuentes, G., Rausell, A. and Valencia, A. 2012. The Ras protein superfamily: evolutionary tree and role of conserved amino acids. *The Journal of Cell Biology* 196(2), pp. 189–201.

Rozakis-Adcock, M., Fernley, R., Wade, J., Pawson, T. and Bowtell, D. 1993. The SH2 and SH3 domains of mammalian Grb2 couple the EGF receptor to the Ras activator mSos1. *Nature* 363(6424), pp. 83–85.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. and Serrano, L. 2005. The FoldX web server: an online force field. *Nucleic Acids Research* 33(Web Server issue), pp. W382–8.

Shah, N.P., Nicoll, J.M., Nagar, B., Gorre, M.E., Paquette, R.L., Kuriyan, J. and Sawyers, C.L. 2002. Multiple BCR-ABL kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia. *Cancer Cell* 2(2), pp. 117–125.

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

Stiffler, M.A., Hekstra, D.R. and Ranganathan, R. 2015. Evolvability as a function of purifying selection in TEM-1 β-lactamase. *Cell* 160(5), pp. 882–892.

Warszawski, S., Netzer, R., Tawfik, D.S. and Fleishman, S.J. 2014. A “fuzzy”-logic language for encoding multiple physical traits in biomolecules. *Journal of Molecular Biology* 426(24), pp. 4125–4138.

Young, A., Lyons, J., Miller, A.L., Phan, V.T., Alarcón, I.R. and McCormick, F. 2009. Ras signaling and therapies. *Advances in cancer research* 102, pp. 1–17.