

Nuthin' but a G (protein) thang: Insights  
into G protein signaling mechanics from  
sequence and structure

By

Alyssa Dawn Lokits

Dissertation

Submitted to the Faculty of the  
Graduate School of Vanderbilt University  
in partial fulfillment of the requirements  
for the degree of

DOCTOR OF PHILOSOPHY

in

Neuroscience

August 11, 2017

Nashville, Tennessee

Approved:

Vsevolod Gurevich, Ph.D.

Jens Meiler, Ph.D.

Heidi E. Hamm, Ph.D.

Anthony Capra Ph.D.

Annette Beck-Sickinger, Ph.D.

I would like to dedicate this work to my family for keeping me sane and smiling no matter what. To my parents, Kirk and Christine, for being an inspiration and setting me on your shoulders so I could move forward and tackle the next mountain. To my sisters, Rebekah and Miriam, for their unfailing support, words of wisdom, and adventurous spirits. To my brothers, Jeremy, Brandon, Cody, Sean, Jesse, and Kenton, for making me tough while keeping me humble. I could not have succeeded without you all.

## ACKNOWLEDGMENTS

This work would not have been possible without the efforts of so many on my behalf, both within and outside of the lab. First I would like to thank my thesis advisor Heidi Hamm for making me an independent and tenacious researcher and scientist. Being co-advised afforded many new perspectives around research, science, and beyond; therefore, I must thank my secondary advisor, Jens Meiler who consistently stepped up and was there to keep me and my work on track. Thank you for having my back and giving me so many opportunities to travel, study, present, and work abroad throughout my thesis career. Those international collaborations are the capstones of my research experience. I must also thank Peter Stadler and his lab in the Bioinformatics Institute at Leipzig University in Leipzig, Germany. I cannot thank you enough for the wonderful experience. And to Henrike Indrischek for being the best collaborator, teacher and friend I could have hoped for.

I would also like to thank the other members of my thesis committee for their dedication to my success. My committee chair, Seva Gurevich, Tony Capra, and Annette Beck-Sickinger each contributed significant and interdisciplinary perspectives to my projects that I would not have received elsewhere. I would like to recognize the Neuroscience Program cohort, faculty, and staff; you all have been both great friends and colleagues throughout my tenure here. You all have been inspiring and fantastic to work with when I would have otherwise wanted to give up. I would also like to thank the Max Kade Foundation and the University of Leipzig for hosting me and funding my research in Germany.

I have received amazing amounts of assistance and mentorship from various members of both the Hamm and Meiler laboratories. Briefly, I'd like to thank Ali Kaya and James Gilbert for all of their guidance and patience with me in my first few years of the wet lab. I

truly would have accidentally broken everything were it not for you two. I must also thank Susan Young, Kendra Oliver, and Kat Betke for being strong sounding boards and friends throughout this journey. In the Meiler lab, there are many people to recognize. Amanda Duran, Brian Bender, Alberto Cisneros and Darwin Fu, thank you for going through the trenches with me. Knowing you were all right there with me the whole time made life much easier. To my pod mates Ben Mueller, Georg Kuenze, and Rocco Moretti, thank you for listening to all of my off-topic ideas, having non-scientific discussions, and for your amazing ability to answer all of my questions, scientific or otherwise. You made my time in lab much more enjoyable.

They say it takes a village to raise a child; I say it also takes a village to raise a graduate student. Outside of the lab I have some very amazing people behind me. I would like to thank my family for sticking it out with me through all of this, especially my parents who have always been there to lend an ear or give sound advice. Thank you to my brothers and sisters for all of your continued support. I'd also like to recognize my "adopted" family and friends locally who have been instrumental in my success and sanity. Thank you all for being there through the thick and thin. You are my village.

## TABLE OF CONTENTS

	Page
DEDICATION .....	ii
ACKNOWLEDGMENTS.....	iii
LIST OF TABLES .....	vii
LIST OF FIGURES.....	ix
PREFACE AND PERSPECTIVES.....	xiii
Chapter	
1 Introduction: Genomes and G $\alpha$ evolution.....	1
1.1 Introduction.....	1
1.2 Theories of G protein evolution.....	5
1.3 Regulation of G protein signaling across Eukaryota .....	15
1.4 Conclusions.....	22
2 Tracing the evolutionary history of the heterotrimeric G protein G $\alpha$ subunit in Deuterostomia .....	30
2.1 Introduction.....	30
2.2 Materials and Methods.....	36
2.3 Results and Discussion.....	44
2.4 Conclusions.....	69
2.5 Supplemental Information.....	71
3 The Activation of Heterotrimeric G Proteins: Evaluation of Past and Present Mechanistic Models .....	115
3.1 Introduction.....	115
3.2 Proposed Mechanisms of Activation .....	121
3.3 Conclusions.....	128
4 Rosetta Comparative Modeling Protocols .....	132
4.1 RosettaCM protocol.....	132
5 A Survey of Conformational and Energetic Changes in G Protein Signaling .....	138
5.1 Introduction.....	138
5.2 Materials and Methods.....	143
5.3 Results.....	147
5.4 Conclusions and Discussions.....	154
5.5 Protocol Capture .....	159
5.6 Supplemental Information.....	169

6 A conserved phenylalanine as relay between the $\alpha 5$ helix and the GDP binding region of heterotrimeric Gi protein $\alpha$ subunit.....	185
6.1 Introduction.....	185
6.2 Materials and Methods.....	189
6.3 Results .....	194
6.4 Discussion .....	202
6.5 Supplemental Information.....	215
7 A conserved hydrophobic core in G $\alpha$ i1 regulates G protein activation and release from activated receptor .....	222
7.1 Introduction.....	222
7.2 Materials and Methods.....	225
7.3 Results.....	232
7.4 Discussion .....	242
APPENDIX .....	259
A.1 Introduction to PAR4 .....	259
A.2 General Overview of Materials and Methods .....	263
A.3 Results and Discussion.....	271
A.4 Conclusions .....	281
CONCLUDING REMARKS AND FUTURE DIRECTIONS .....	298
BIBLIOGRAPHY .....	309

## LIST OF TABLES

Table	Page
1. <b>1.1</b> Human G $\alpha$ protein sequence similarity .....	24
2. <b>1.2</b> G protein signaling components found across Eukaryota.....	29
3. <b>2.1</b> (pre)GNA- paralog presence before and after the 2R WGD in Vertebrates projected onto a Deuterostome species tree.....	82
4. <b>Supplemental 2.1</b> Species Evaluated.....	96
5. <b>Supplemental 2.2</b> Transcriptome and Expression Data.....	100
6. <b>Supplemental 2.3</b> Sites under positive selection in the branch leading to <i>GNAO.1</i> .....	104
7. <b>Supplemental 2.4</b> Significant results of the branch-site model indicate positive selection in the <i>GNAO.1</i> #1 branch.....	105
8. <b>Supplemental 2.5</b> Retrogenes in Primates.....	114
9. <b>Supplemental 5.1</b> Predicted $\Delta\Delta G$ of the $\alpha 1$ helix across three states of G $\alpha$ signaling – G $\alpha_{i1}$ (GDP) $\beta_1\gamma_1$ , R*-G $\alpha_{i1}$ (empty) $\beta_1\gamma_1$ , and G $\alpha_{i1}$ (GTP).....	169
10. <b>Supplemental 5.2</b> Predicted $\Delta\Delta G$ of the $\alpha 5$ helix across three states of G $\alpha$ signaling – G $\alpha_{i1}$ (GDP) $\beta_1\gamma_1$ , R*-G $\alpha_{i1}$ (empty) $\beta_1\gamma_1$ , and G $\alpha_{i1}$ (GTP).....	170
11. <b>Supplemental 5.3</b> Predicted $\Delta\Delta G$ of the $\alpha F$ helix across three states of G $\alpha$ signaling – G $\alpha_{i1}$ (GDP) $\beta_1\gamma_1$ , R*-G $\alpha_{i1}$ (empty) $\beta_1\gamma_1$ , and G $\alpha_{i1}$ (GTP).....	171
12. <b>Supplemental 5.4</b> Predicted $\Delta\Delta G$ of the P-loop across three states of G $\alpha$ signaling – G $\alpha_{i1}$ (GDP) $\beta_1\gamma_1$ , R*-G $\alpha_{i1}$ (empty) $\beta_1\gamma_1$ , and G $\alpha_{i1}$ (GTP).....	172
13. <b>Supplemental 5.5</b> Predicted $\Delta\Delta G$ of the Switch I region across three states of G $\alpha$ signaling – G $\alpha_{i1}$ (GDP) $\beta_1\gamma_1$ , R*-G $\alpha_{i1}$ (empty) $\beta_1\gamma_1$ , and G $\alpha_{i1}$ (GTP).....	173
14. <b>Supplemental 5.6</b> Predicted $\Delta\Delta G$ of the Switch II region across three states of G $\alpha$ signaling –G $\alpha_{i1}$ (GDP) $\beta_1\gamma_1$ , R*-G $\alpha_{i1}$ (empty) $\beta_1\gamma_1$ , and G $\alpha_{i1}$ (GTP).....	174
15. <b>Supplemental 5.7</b> Predicted $\Delta\Delta G$ of the Switch III region across three states of G $\alpha$ signaling – G $\alpha_{i1}$ (GDP) $\beta_1\gamma_1$ , R*-G $\alpha_{i1}$ (empty) $\beta_1\gamma_1$ , and G $\alpha_{i1}$ (GTP).....	175
16. <b>Supplemental 5.8</b> Predicted $\Delta\Delta G$ of the $\alpha A$ helix across three states of G $\alpha$ signaling – G $\alpha_{i1}$ (GDP) $\beta_1\gamma_1$ , R*-G $\alpha_{i1}$ (empty) $\beta_1\gamma_1$ , and G $\alpha_{i1}$ (GTP).....	176
17. <b>6.1</b> Crystallographic data collection and refinement statistics.....	212
18. <b>6.2</b> G protein alpha subunit $\alpha 1$ helix interface energetic prediction.....	213

19. <b>Supplemental 6.1</b> Average energy scores between interacting residue pairs of the receptor, unbound and bound alpha5 helix and non-alpha5 regions. ....	214
20. <b>7.1</b> Bodipy nucleotide interactions with Gα1 .....	255
21. <b>7.2</b> Crystallographic data collection and refinement statistics.....	256
22. <b>7.3</b> G protein alpha subunit α1 helix interface energetic predictions .....	257
23. <b>A.1</b> Amino acid sequence identity between the human PAR sequences (1-4).....	284
24. <b>A.2</b> 20 Structures selected as templates.....	285
25. <b>A.3</b> Five structures selected as templates .....	286
26. <b>A.4</b> Selected point mutations for binding pocket mapping.....	287



## LIST OF FIGURES

Figure	Page
1. <b>1.1</b> Unikont and Bikont Signaling through G proteins: Distribution of G protein components among eukaryotes .....	23
2. <b>1.2</b> Vertebrate <i>GNA</i> - paralogous gene arrangements.....	25
3. <b>1.3</b> Aligning representative vertebrate protein-coding exon borders of all five major families of the $G\alpha$ subunit.....	26
4. <b>1.4</b> Two theories of G protein evolution.....	27
5. <b>2.1</b> All Deuterostome branches investigated .....	81
6. <b>2.2</b> Maximum Likelihood Tree of ( <i>pre</i> ) <i>GNA</i> - genes.....	84
7. <b>2.3</b> Aligning representative vertebrate protein-coding exon borders of all five major families of the $G\alpha$ subunit.....	85
8. <b>2.4</b> Evolution of the five families of $G\alpha$ .....	86
9. <b>2.5</b> Evolution of $G\alpha\gamma$ .....	88
10. <b>2.6</b> Flexibility of exon-intron borders within the ( <i>pre</i> ) <i>GNA12</i> and <i>GNA13</i> genes .....	90
11. <b>2.7</b> ML tree of the $G\alpha\delta$ family resolves gene relationships.....	91
12. <b>2.8</b> Alternative Splicing of <i>GNAO</i> .....	93
13. <b>2.9</b> Retained exons of <i>GNAO</i> after 3R WGD in Teleosts.....	94
14. <b>2.10</b> Multiple transcripts are possible from the complex locus of <i>GNAS</i> .....	95
15. <b>Supplemental 2.1</b> Primate species investigated for pseudogenes.....	97
16. <b>Supplemental 2.2</b> Implications of alternative exon usage on tertiary structure in lancelet <i>preG\alpha\delta</i> and <i>preG\alpha\gamma</i> .....	101
17. <b>Supplemental 2.3</b> Implications of alternative exon usage on tertiary structure in lancelet <i>preGNAS</i> .....	102
18. <b>Supplemental 2.4</b> Exon structure of <i>GNAI</i> and <i>GNAZ</i> .....	103
19. <b>Supplemental 2.5</b> Sequence frequency logo of <i>GNAO</i> residues that were positively selected on the branch leading to <i>GNAO.1</i> .....	106
20. <b>Supplemental 2.6</b> Exon 3 of <i>GNAS</i> in human .....	107

21. <b>Supplemental 2.7</b> Extension of exon4 of <i>GNAS</i> in placental mammals .....	108
22. <b>Supplemental 2.8</b> DNA- and RNA-binding protein motifs overlapping with the 3' canonical and non-canonical splice sites of intron 3 in <i>GNAS</i> .....	109
23. <b>Supplemental 2.9</b> DNA- and RNA-binding protein motifs overlapping with the extended conserved region around exon 3 in <i>GNAS</i> of 33 placentals .....	110
24. <b>Supplemental 2.10</b> Local exon duplications of <i>GNAQ</i> , <i>GNAIL</i> , and <i>preGNAI</i> .....	111
25. <b>Supplemental 2.11</b> 5' non-canonical splice site pattern of <i>GNAIL</i> intron6 in birds and mammals .....	112
26. <b>Supplemental 2.12</b> DNA- and RNA-binding protein motifs overlapping with the 5' non-canonical splice site of intron6 in <i>GNAIL</i> .....	113
27. <b>3.1</b> G protein signaling in ROS membranes .....	131
28. <b>5.1</b> The G protein signaling cycle .....	161
29. <b>5.2</b> $G\alpha$ primary sequence with secondary and tertiary structure names .....	162
30. <b>5.3</b> Structural representation of predicted $\Delta\Delta G$ of the $\alpha 1$ helix across three states of $G\alpha$ signaling— $G\alpha_{i1}(\text{GDP})\beta_1\gamma_1$ , $R^*-G\alpha_{i1}(\text{empty})\beta_1\gamma_1$ , and $G\alpha_{i1}(\text{GTP})$ .....	163
31. <b>5.4</b> Structural representation of predicted $\Delta\Delta G$ of the $\alpha 5$ helix across three states of $G\alpha$ signaling— $G\alpha_{i1}(\text{GDP})\beta_1\gamma_1$ , $R^*-G\alpha_{i1}(\text{empty})\beta_1\gamma_1$ , and $G\alpha_{i1}(\text{GTP})$ .....	164
32. <b>5.5</b> Pairwise analysis of ROSETTA scores for individual amino acid interactions as a means to monitor information networks across signaling states .....	165
33. <b>5.6</b> Pairwise analysis of individual amino acid interactions across three $G\alpha$ signaling states to investigate changes between residue pairs interacting with the $\alpha 1$ helix and Linker 1 region (K46-I56, & H57-S62, respectfully) .....	166
34. <b>5.7</b> Pairwise analysis of individual amino acid interactions across three $G\alpha$ signaling states to investigate changes between residue pairs interacting with the $\alpha 5$ helix (N331-I343) .....	167
35. <b>Supplemental 5.1</b> Generation of Models.....	168
36. <b>Supplemental 5.2</b> Pairwise Analysis of $G\alpha_{i1}$ : The Switch I Region .....	177
37. <b>Supplemental 5.3</b> Pairwise Analysis of $G\alpha_{i1}$ : The Switch II Region .....	178
38. <b>Supplemental 5.4</b> Pairwise Analysis of $G\alpha_{i1}$ : The Switch III Region.....	179
39. <b>Supplemental 5.5</b> Pairwise Analysis of $G\alpha_{i1}$ : The $\alpha F$ helix .....	180

40. <b>Supplemental 5.6</b> Pairwise Analysis of $G\alpha_{i1}$ : The P-loop.....	181
41. <b>Supplemental 5.7</b> Pairwise Analysis of $G\alpha_{i1}$ : The $\alpha 1$ - $\alpha A$ linker.....	182
42. <b>Supplemental 5.8</b> Pairwise Analysis of $G\alpha_{i1}$ : The $\alpha A$ helix .....	183
43. <b>6.1</b> Heterotrimeric G protein; localization and function $\alpha 5$ helix in G proteins .....	205
44. <b>6.2</b> The effects of $\beta 5$ - $\beta 6$ strands mutations on G protein activation.....	206
45. <b>6.3</b> The effect of F336 residue on G protein activation .....	207
46. <b>6.4</b> Cross-linking of $\alpha 1$ and $\alpha 5$ helices of $G\alpha 1$ HI.....	208
47. <b>6.5</b> The effects of hydrophobic residues around F336 on nucleotide exchange rates .....	209
48. <b>6.6</b> The effect of $MgCl_2$ on $G\alpha i 1$ basal activity .....	210
49. <b>6.7</b> Structural features of GDP bound F336 mutant structures .....	211
50. <b>7.1</b> Heterotrimeric G protein; localization of 4 alanine insertion in $\alpha 5$ helix .....	247
51. <b>7.2</b> Biochemical properties of Ins4A protein.....	248
52. <b>7.3</b> Hypotheses for why Ins4A protein cannot release from the active receptor-G protein complex in presence of guanine nucleotide.....	249
53. <b>7.4</b> Conformational changes at key sites on $G\alpha$ caused by receptor and $GTP\gamma S$ determined using site-directed fluorescent labels .....	250
54. <b>7.5</b> The effect of introducing an FVFD insertion in the $\alpha 5$ helix on $G\alpha i 1$ subunit.....	251
55. <b>7.6</b> Structural features of $GTP\gamma S$ bound Ins4A mutant protein .....	252
56. <b>7.7</b> Structural features of heterotrimeric Ins4A $\beta 1\gamma 1$ mutant protein .....	253
57. <b>7.8</b> Pairwise interaction scores highlight two activation pathways .....	254
58. <b>A.1</b> Activation mechanism of PAR4 .....	283
59. <b>A.2</b> PAR4 and PAR1 sequence alignment.....	288
60. <b>A.3</b> PAR4 Single Template score vs RMSD to PAR1 structure .....	289
61. <b>A.4</b> Multi-Template Comparative Modeling does not accurately reflect extracellular loop (ECL) regions .....	290
62. <b>A.5</b> Score versus RMSD for 20 template comparative modeling.....	291
63. <b>A.6</b> Representative Single Template BMS-3 Docking .....	292

64. <b>A.7</b> Two primary binding modalities for 20 Template docking of BMS-3 .....	293
65. <b>A.8</b> Score versus RMSD for multi-template docking.....	294
66. <b>A.9</b> Representative loop remodeling trials around a docked BMS-3 ligand .....	295
67. <b>A.10</b> Six mutations of PAR4 to PAR1 sequence .....	296

## PREFACE AND PERSPECTIVES

Heterotrimeric G proteins represent an important node in the intracellular signaling network. Ubiquitously expressed in all mammalian cells, heterotrimeric G proteins are responsible for relaying extracellular signals which activate G protein Coupled Receptors (GPCRs) into a message the cell can understand and respond to. With over 800 different GPCRs in the human genome, this represents a significant amount of signal transduction from a wide array of extracellular ligands. Heterotrimeric G proteins, composed of a monomeric  $G\alpha$  subunit and an obligate  $G\beta\gamma$  dimer, must therefore coordinate the propagation and amplification of the signal with a high degree of integrity. This is done through a highly regulated cycle of protein-protein interactions whose affinities for one another alter based on subtle deviations across the  $G\alpha$  structure. Therefore the  $G\alpha$  subunit acts as a regulatory switch protein and effectively manages and maintains signaling across each of the different stages of the cycle based on its structure.

The study of G protein activation through their cognate GPCRs has been the focus of intensive research for decades. However, it is still not known the direct order of events leading to the allosteric  $G\alpha$  activation and subsequent G protein dissociation from the ternary complex upon coupling to an activated receptor. Indeed, recent work has focused on the interaction and some selectivity measures around the  $\alpha 5$  helix of the  $G\alpha$  subunit as it is a primary interface for GPCR coupling. Advances from X-ray crystallography, DEER/EPR, and crosslinking studies (discussed in detail in Chapter 3) have shed much light on this interface and the dynamics around these protein players. However, there still remains no unified understanding of G protein activation and dissociation.

Some open questions in the field less than five years ago were: How does the activated GPCR induce an allosteric change through the  $G\alpha$  subunit's GTPase and nucleotide binding domain to induce nucleotide exchange? How does this activation vary across the five primary  $G\alpha$  families? What does this network of communication need to look like to move the information from the receptor interface all the way to the nucleotide binding pocket? What residue positions confer self-activation of the  $G\alpha$  subunit versus requiring GPCR-GEF activation? Does the helical domain have to open to allow nucleotide exchange? How much must it open? Once the helical domain has opened to allow exchange, how does it close again? What is the order of this closing in regards to subsequent ternary complex dissociation? Does

G $\alpha$  release directly from the complex or does the trimer release from the receptor before G $\beta\gamma$  are freed? After release, how does RGS catalyze hydrolysis of GTP in a coordinated manner? How would this be similar or different from a 7TM-GAP regulator of hydrolysis?

Since posing these initial questions, intensive studies from the Hamm, Meiler, and Stadler labs have resulted in new insight into G protein activation, allosteric communication, and evolution. Specifically, studies focused on G $\alpha$  activation looked at the residue-residue interaction networks underlying allosteric communication across the GPCR-G $\alpha$ -G $\beta\gamma$  complex. These interactions were mapped using *in silico* calculations of predictive energy scores using the software suite, Rosetta (Chapter 4-5). These approximations for energy contributions between side chain and back bone atoms were then translated into predictions for critical residues necessary for propagation of the information from the GPCR interface to the nucleotide binding pocket and across the different G protein signaling states (see Chapter 5-7).

To test these predictions, site-specific point mutations and cross-linking studies were employed to assay for variations in nucleotide exchange in the presence and absence of the receptor and for stabilization of the ternary complex. Membrane binding studies shed light on the ability of the mutant proteins to bind activated receptors and dissociate appropriately. Additional efforts for crystallization of different mutants also assisted in the characterization of these mutations. All of this information was then fed into creating more accurate models of G protein activation.

Further characterization teasing apart the details of the allosteric G protein activation pathway suggested that not only were a collective of conserved residues necessary for signal propagation, but also a conserved hydrophobic core was responsible for transmitting GPCR coupling to the nucleotide binding pocket and helical domain. These studies were foundational in showing how secondary structure elements such as the  $\alpha 5$  helix,  $\alpha 1$  helix and the six  $\beta$ -strands of the GTPase domain communicate across interfaces. All biochemical *in vitro* assays were informed by *in silico* predictive models, while all *in silico* models were improved through iterations of *in vitro* mutations and assays.

Additional characterization of this information flow was carried out across several signaling states of the G protein signaling cycle. These predictive networks are in the process of being validated in order to highlight the power of this computationally inexpensive approach to assisting functional studies of dynamics and function (Chapters 5-7).

Other questions regarding subfamily specificity across this allosteric network were also addressed. Is there a unified activation mechanism across all families of G $\alpha$ ? To answer this, deep multiple sequence alignments (MSAs) were constructed to evaluate the contribution of specific residue positions to structure and function. It was hypothesized that though some variation would be necessary for differences in GPCR and effector selectivity, there would be a conserved mechanism for information propagation across the G $\alpha$  subunit leading to helical domain opening, nucleotide exchange, and subsequent activation.

While interrogating this question by constructing quality sequences for construction of the MSAs, we uncovered a larger problem regarding the curation and annotation of G protein sequences. Though several groups had created tentative theories of G protein evolution, complications in genome sequencing and assembly, limited species knowledge, and in accurate annotations had led to several disparate attempts to understand the evolution of the G $\alpha$  subunit, specifically within Deuterostomes.

To untangle these confounds and to create high quality MSAs, I collaborated with the Stadler laboratory in Leipzig, Germany to create a novel curation and annotation algorithm to aid in the assessing fragmented genome assemblies. Through the use of the ExonMatchSolver, many new G $\alpha$  sequences were uncovered in species that had yet to be studied. These discoveries led to a new proposed theory of G protein evolution in Deuterostomia. Combined with other recent studies in Metazoa, Holozoa, and Opisthokonta, we have charted a new trajectory for when each G $\alpha$  paralog emerged in Opisthokonta evolution.

While outlining the new theory of G $\alpha$  evolution, we have created deep, species diverse, high quality MSA which can now be used to evaluate the sequence-based constraints underlying G $\alpha$  function. Initial efforts to evaluate subfamily specific questions of interaction and selectivity across various receptor and effector proteins will not be included herein but will be the work of future study. Combining these efforts with ongoing structural studies evaluating ensembles of communication interaction maps across the G protein signaling states will lend further credence to our working model of the mechanism of G protein activation.

In addition, efforts to characterize and map different GPCRs will further assist our mechanistic model. The proteinase-activated receptor (PAR) 1 and 4 are both expressed on human platelets and are thrombin receptors. Current efforts to structurally characterize these proteins *in vitro*, *in vivo* and *in silico* will shed light on cleavage-induced activation and the

first steps leading to G protein activation.

Overall, my studies have resulted in new understanding of G protein activation, evolution, and function. As GPCRs represent the targets of roughly half of all therapeutics, increasing our understanding of the intracellular transducing element and the system around these proteins is critical for continued improvement and development of therapeutics. As many diseases are caused by erroneous G protein signaling, study of the mechanism of G protein evolution, activation and signaling remains paramount for the improvement of human health.



## CHAPTER 1

### INTRODUCTION: GENOMES AND G $\alpha$ EVOLUTION

#### 1.1 Introduction

##### **Chapter 1**

Lokits AD, Meiler J, Hamm HE, “Mini-Review: Genomes and G $\alpha$  proteins” *Neuroscience Letters* **2017** in submission

##### **Contribution**

I am first author of this manuscript. I contributed the text: abstract, introduction, review and assessment of all included articles and conclusion. I edited all sections and created the Figure 1.2 as well as Table 1.1. Figures 1.3 and 1.4 were of my creation but previously submitted in a manuscript to the journal of *BMC Evolutionary Biology*. Lokits AD, Meiler J, Hamm HE, “Tracing the evolutionary history of the heterotrimeric G protein G $\alpha$  subunit in Deuterostomia” *BMC Evolutionary Biology* **2017** in submission.

Figures 1.1 was reprinted from <sup>1</sup>. Bradford, W., Buckholz, A., Morton, J., Price, C., Jones, A. M., and Urano, D. (2013) Eukaryotic G protein signaling evolved to require G protein-coupled receptors for activation, *Sci Signal* **6**, ra37. Reprinted with permission from AAAS.

Table 1.2 was reprinted from <sup>2</sup>. de Mendoza, A., Seb e-Pedr os, A., and Ruiz-Trillo, I. (2014) The evolution of the GPCR signaling system in eukaryotes: modularity, conservation, and the transition to metazoan multicellularity, *Genome Biol Evol* **6**, 606-619 by permission of Oxford University Press or the sponsoring society.

##### **Abstract**

The heterotrimeric G protein has evolved to integrate and amplify a wide range of signals into intracellular communication cascades. The G $\alpha$  subunit of the heterotrimeric G protein is a tightly regulated molecular switch, whose subtle structural changes convey alternating affinity for the various protein and nucleotide components it interacts with across the distinct stages of the G protein signaling cycle. Indeed, how the G $\alpha$  subunit is regulated to transition between these stages has evolved to vary significantly across different branches of Eukaryota. The most notable dichotomy is the inverted mechanism of signal regulation found between animal and plant G proteins. Canonically, 7 transmembrane (7TM) receptors, such as the G protein coupled receptors (GPCRs) in

Metazoa (animals), initiate signaling cascades by inducing G $\alpha$  activation to act as the transducing element between an extracellular stimulator and the intracellular response pathway. However, in Archaeplastida (which includes land plants), “self-activating” G $\alpha$  subunits are relied on to initiate intracellular signaling cycles, independent of the 7TM receptor. There is much speculation around which regulatory modality arose first in Eukaryota evolution. In this review, we describe the initial discoveries and studies evaluating the proposed evolution of G $\alpha$  across the major branches of Unikonta (Opisthokonta -- such as animals, fungi, *etc.*, -- and Amoebozoa) and Bikonta (Archaeplastida – such as plants, algae, *etc.*, -- Chromalveolata, Excavata, and Rhizaria), and we survey the current understanding of G protein signaling mechanics across Eukaryota. Looking forward, leveraging knowledge of the G protein’s dynamic system of regulation and control will lead to more targeted, therapeutically relevant moieties able to differentiate and distinguish between human signaling components versus those from pathogenic Fungi, Amoebozoa, and other microbial organisms. The regulatory constraints imposed through evolution can therefore highlight subtle divergences between these conserved signaling networks.

### ***Introduction***

Heterotrimeric guanine nucleotide-binding proteins (G proteins) are critical signal transducing elements which integrate and amplify intracellular signaling cascades. Canonically, these responses are induced by interacting with activated G protein coupled receptors (GPCRs). Therefore, the G proteins act as switch proteins to relay important signals by activating or inhibiting effector proteins and modulating the production of second messengers such as cAMP<sup>3-5</sup>, cGMP<sup>6</sup>, or releasing calcium from intracellular

stores<sup>7, 8</sup>. Composed of a monomeric  $G\alpha$  subunit and an obligate  $G\beta\gamma$  dimer<sup>9</sup>, heterotrimeric G proteins integrate a wide array of signals by coupling to the roughly 800 different GPCRs encoded in the human genome<sup>10</sup>. Their cognate receptors bind and respond to a diverse assortment of ligands ranging from small molecules, proteins, peptides, ions, neurotransmitters, lipids, and even photons of light<sup>9, 11</sup>. With only 16 different  $G\alpha$  subunit proteins encoded in the human genome, G protein signaling has evolved to be a highly regulated and efficient means of signal transduction to incorporate responses from so many receptors. As these transducing elements represent the nexus of many protein interactions, understanding their evolution in order to take advantage of constrained residues which preserve the necessary structure and/or function will be the new frontier for targeted therapeutic approaches which aim to bias signaling and modify signaling sensitivity, rate and amplitude.

As this finely tuned signaling cycle is elegantly poised to initiate, amplify, and terminate intracellular pathways, it is curious that the G protein signaling mechanism appears to have evolved to be differentially regulated between Unikonta and Bikonta<sup>1</sup> (Figure 1.1). Unikonta can be divided into two primary branches, Amoebozoa and Opisthokonta (composed of animals and fungi, *etc.*). Bikonta can be approximated into four major phyla: Archaeplastida (plants, algae, *etc.*), Chromalveolata, Excavata, and Rhizaria. In each of these branches, genome level studies have sought to identify sequences for the various components of G protein signaling<sup>1, 2, 12-14</sup>. The presence and absence of each protein moiety across the different phyla suggests several alternative means of regulation are present outside of the canonical, animal model of G protein signaling. In this review we evaluate the discovery and initial understanding of G protein evolution,

specifically around the  $G\alpha$  subunit, in vertebrates and across the different major phylogenetic branches. In addition, we briefly relay the arguments for and against two predominant theories of G protein evolution and regulation<sup>1,2</sup>.

## 1.2 Theories of G protein evolution

### *Early predictions of G protein evolution*

#### **G $\alpha$ evolution is linked to monomeric G proteins and elongation factors**

The first efforts for sequencing G proteins occurred in the early 1980's. Then, only small peptide fragments were sequenced<sup>15</sup>. Yet despite their size, it was clear that G $\alpha$  had sequence homology to GTPase proteins such as the Ras family (also known as the monomeric G protein family) and Elongation Factors<sup>16</sup>. Between these three protein families, conserved motifs were found which could form a binding pocket for coordinated GDP and GTP binding, and a catalytic site for GTP hydrolysis. In the mid- to late-1980's cloning and sequencing efforts boomed for G proteins. These studies showed that more than the two known G protein G $\alpha$  subunits were present in mammals. Originally, only an Adenylyl Cyclase stimulator (G $\alpha$ s) and inhibitor (G $\alpha$ i) were thought to exist. Advances in cloning and sequencing showed two different G $\alpha$  subunits existed in the rods and cones of the eyes (G $\alpha$ t1 and t2, respectively<sup>17-19</sup>); this expanded search efforts for new G $\alpha$  subunits to other organs of the body. Ultimately 16 G $\alpha$  proteins were found to be expressed in human and other mammalian species<sup>20</sup>.

#### **Evolutionary relationship within G $\alpha$ families based on sequence, expression, and function**

The human G $\alpha$  subunits are divided into 5 primary families (s, i, 12, q, and v)<sup>21, 22</sup>. It was first discovered by sequence similarity that the genes could be related<sup>20</sup> (See Table 1 for sequence similarity). The largest family, G $\alpha$ i, is composed of eight members which share high sequence similarity (>50% between all eight proteins in humans). These members include G $\alpha$ i1, i2, i3, t1, t2, t3, o, and z. G $\alpha$ i1-3 share >85% sequence similarity in humans and have significant overlap in expression and in function<sup>23-25</sup>. G $\alpha$ t1-3, as mentioned previously, all possess more restricted, tissue-specific expression patterns and

are classically known for their importance in different neurosensory perceptions of vision and taste<sup>17-19, 26</sup>. G $\alpha$ o (which stands for “other” due to its unknown function upon its discovery) is expressed abundantly in the brain<sup>27, 28</sup>. G $\alpha$ z, the final member of the G $\alpha$ i family is expressed in the brain and other neuronal tissue types<sup>29</sup> and serves various functions in catecholamine behavioral responses<sup>25</sup>.

Closest to the G $\alpha$ i family in sequence is the G $\alpha$ q family. G $\alpha$ q has four members: G $\alpha$ q, 11, 14, and 15. G $\alpha$ q and 11 are ubiquitously expressed<sup>30</sup> and share a high level of sequence (90% in humans) and functional overlap as all members of the G $\alpha$ q family can activate Phospholipase C<sup>31, 32</sup> but GPCRs do not seem to discriminate between G $\alpha$ q and G $\alpha$ 11<sup>25, 33-35</sup>. G $\alpha$ 14 and 15 are more diverse in sequence (G $\alpha$ 14 is 81% similar to G $\alpha$ q, and G $\alpha$ 15 is 56% similar to G $\alpha$ q) and possess more restricted, tissue-specific expression<sup>36, 37</sup> in liver, kidney, lung (G $\alpha$ 14) and hematopoietic cells (G $\alpha$ 15).

The G $\alpha$ s family is composed of two members, G $\alpha$ s and G $\alpha$ olf. As with the G $\alpha$ i and G $\alpha$ q families, one member, G $\alpha$ s, is ubiquitously expressed in all tissues<sup>25</sup> while the second member, G $\alpha$ olf, possesses more restricted expression in the olfactory tissues and other brain regions<sup>38</sup>. Both can stimulate cAMP production through interaction with all subtypes of Adenylyl Cyclase<sup>25</sup>. The G $\alpha$ s family share 79% sequence similarity to one another and ~40% to the other G $\alpha$  subunits.

The fourth family is composed of G $\alpha$ 12 and 13. Unlike the other families, both members of the G $\alpha$ 12 family appear to be ubiquitously expressed<sup>39</sup>. Based on sequence, this family shares 66% sequence similarity in humans and ~40% to other G $\alpha$  subunits. Indeed G $\alpha$ 12 and 13 have a very specific function within the cell; specifically they have been shown to signal to monomeric G proteins such as Rho to modulate acto-myosin

structures<sup>25, 40</sup>.

### **A new fifth family of G $\alpha$ paralogs shed light on G protein evolution**

A recent advance in the understanding of G protein evolution came from the discovery of a new family of G $\alpha$ <sup>22</sup>. Named G $\alpha$ <sub>v</sub> (the Roman numeral *v* for five), the fifth and final family, G $\alpha$ <sub>v</sub> is composed of only its one namesake member. This protein was discovered exactly 40 years after the first experiments led to the discovery of G proteins<sup>41</sup> and almost 20 years after all members of the other four families had been defined<sup>22</sup>. Its late discovery is primarily due to the wide-spread loss of this gene, *GNAV*, in the genomes of many vertebrate and model organisms. The evolution of G $\alpha$ <sub>v</sub> has been traced through many lineages including: teleost fishes, cartilaginous fishes, cephalochordates (lancelets), echinodermites (sea urchins), beetles, annelids (segmented worms), mollusks and sponges<sup>22, 42</sup>. Though the gene is absent in fruit flies, nematodes, tetrapods, and jawless vertebrates, the G $\alpha$ <sub>v</sub> paralog has played a role in G protein signaling and evolution. The function of G $\alpha$ <sub>v</sub> is unclear though it has a wide tissue distribution in zebrafish<sup>22</sup>.

It was posited that G $\alpha$ <sub>v</sub> was as ancient as the other four classes of G $\alpha$ <sup>22, 42</sup>. Indeed, G $\alpha$ <sub>v</sub> is the most similar in sequence to the G $\alpha$ <sub>i</sub> family (44-47% between coelacanth G $\alpha$ <sub>v</sub> and all human G $\alpha$ <sub>i</sub> family proteins) despite several critical differences in ADP-ribosylation sensitivity<sup>22</sup>, and altered exon border positions between the gene structures<sup>14</sup> (discussed below).

In the original discovery and analysis<sup>22</sup>, the lineage-specific losses were thought to be indicative of a “birth and death” mode<sup>43</sup> of evolution, counter to the previous theory of “concerted evolution”<sup>44</sup> which suggests that multimember genes families evolve together in concert. In the birth and death model, new genetic material is introduced through gene duplication and reinsertion; some of these new genes will be maintained, gain new function

or tissue specificity for subfunctionalization. Others will become inactivated and/or deleted<sup>43</sup>. These findings of a birth and death model of evolution are consistent with current findings surrounding  $G\alpha$  evolution after the two rounds of whole genome duplication (2R WGD) which occurred at the emergence of the vertebrate lineage<sup>14</sup>.

### **Relationship through gene arrangements across G protein families**

Shortly after their initial discovery and sequencing in mammals,  $G\alpha$  encoding genes (*GNA*-) were identified as being a highly conserved “housekeeping genes” due to their lack of a TATA box and high concentration of GC motifs<sup>24</sup>. Many of the  $G\alpha$  genes are arranged in tandem duplication pairs along the mammalian genome<sup>45</sup>. As seen in Figure 1.2, these gene paralogs are arranged in a head-to-tail or head-to-head arrangement. *GNAV*, while not present in mammals also does not appear as a tandem pair in other vertebrate lineages<sup>14</sup>.

The conserved exon-intron structure of all vertebrate *GNA*- genes suggests evolutionary links between the five  $G\alpha$  families (Figure 1.3). The  $G\alpha_i$  family share eight conserved protein coding exons. The only exception is *GNAZ* ( $G\alpha_z$ ), which possesses two exons, the first encoding the sequence of exons 1 through 6 of other  $G\alpha_i$  genes and the second exon aligning with exons 7 and 8. The  $G\alpha_q$  family appears to be the most similar to  $G\alpha_i$  members as it possesses seven protein coding exons. All seven of the exons align with the exon border positions of the  $G\alpha_i$  family with the exception of exon 2, which encompasses the sequences of both exon 2 and 3 of  $G\alpha_i$  (Figure 1.3). *GNAV* shares six exon border positions with the  $G\alpha_i$  and  $G\alpha_q$  families in vertebrates. Again, exon 2 and 3 have mismatched border positions between these families, but exon 7 of  $G\alpha_i$  (exon6 of  $G\alpha_q$ ) is also divided into two exons in *GNAV*.  $G\alpha_i$ ,  $q$ , and  $v$  all share four exon border positions with the *Gas* family. This suggests that  $G\alpha_i$ ,  $q$  and  $v$  are more closely related to one another than to  $G\alpha_s$ <sup>14</sup>. Unlike the other families,  $G\alpha_{12}$  shares no exon border positions



with any other  $G\alpha$  family. Instead, *GNAI2* and *GNAI3* are composed of 4 exons each. Their exon border positions, dissimilar sequence and protein function is indicative of a retrogene that gained introns after being reinserted into the genome before undergoing neofunctionalization<sup>14, 46, 47</sup>.

### **Original theory of evolution**

Using only mammalian sequences, the first major theory on the origins of G protein paralogs, postulated that a progenitor  $G\alpha$  subunit gene, *GNA-*, underwent several duplication events resulting *GNAS*, *GNAI2*, and *GNAI/Q* (Figure 1.4a). *GNAI/Q* underwent a tandem duplication that led to the two side by side genes, *GNAI/Q'*—*GNAI/Q''*. The tandem pair duplicated once more to result in two sets of tandem pairs which differentiated into *GNAI'*—*GNAI''* and *GNAQ'*—*GNAQ''*. In this theory, *GNAS*, *GNAI2*, *GNAQ'*—*GNAQ''* and *GNAI'*—*GNAI''* all duplicated at least once more, though the timing of these duplications are unknown relative to one another. These duplications ultimately created the two family members *GNAS* and *GNAL* ( $G\alpha_s$  and  $G\alpha_{olf}$ ), the two family members *GNAI2* and *GNAI3* ( $G\alpha_{12}$  and  $G\alpha_{13}$ ) as well as the two pairs of genes *GNAQ*—*GNAI4*, *GNAI1*—*GNAI5* which encode the  $G\alpha_q$  family ( $G\alpha_q$ , 14, 11, and 15). For the tandem pair *GNAI'*—*GNAI''*, the duplication resulted in *GNAI*—*GNAT*, and *GNAO* (whose second gene in the pair was partially deleted to allow for an alternatively spliced ending in the  $G\alpha_o$  transcript). In this theory, the *GNAI*—*GNAT* pair duplicated twice more to give rise to the rest of the  $G\alpha_i$  family ( $G\alpha_{i1-3}$ ,  $G\alpha_{t1-3}$ ). In addition, *GNAZ* was believed to arise from a retrotransposition of a  $G\alpha_i$  family member<sup>46</sup>. Some postulation on the timing of these events suggested that the duplications must have occurred prior to the divergence of different mammalian species though no definitive timeline assessment was possible given the limited data set. In addition, *GNAV* had yet to be discovered, and was therefore not

included in the original assessments of G $\alpha$  evolution;

### ***Expanded theories of evolution through chemosensory processing in animals***

#### **Phototransduction**

Focusing on the radiation of phototransduction and the visual system, Nordstrom *et al.* evaluated the evolution of the transducins (Gat1 and Gat2) in both visual rods and cones in the context of the entire visual protein signaling cascade within the retina, effectively launching a series of publications around the evolution of these proteins first in humans<sup>47</sup> then in vertebrates and into chordates with non-deuterostome species as outgroups<sup>48-50</sup>. Using the theory set by Wilkie *et al.* of G $\alpha$  evolution as a springboard<sup>46</sup>, they found that many of the G $\alpha$  subunits were arranged as paralogons within the human genome, meaning that many of the genes necessary for vision had been duplicated as sets of genes in chromosomal regions or “blocks” of linked genes<sup>47</sup>. They proposed that the duplicated genes blocks subsequently gained tissue specificity and underwent subfunctionalization for the development and diversification of proteins specific to the phototransduction system. This is in accordance with the aforementioned hypothesis<sup>46</sup>, in which an ancestral *GNAI*—*GNAT* gene pair were duplicated twice to give rise to three sets of *GNAI*—*GNAT* pairs<sup>47</sup>.

Next, they integrated sequences from vertebrates to invertebrate chordates (tunicates) to further evaluate the evolution of phototransduction genes<sup>48</sup> that may have arisen as a result of whole genome duplication events prior to the emergence of vertebrates<sup>51, 52</sup>. Using the conservation of synteny (gene neighbors) as readouts of a co-evolutionary history between the *Gai* and *Gat* subfamilies, they propose that *Gai* maintained the ancestral, progenitor protein function to inhibit cAMP production with a more ubiquitous expression pattern while the second genes within the pairs, encoding *Gat* proteins, were able to

subfunctionalize with new tissue specificity<sup>48</sup>. In addition, further studies showed that *Gat* did not exist outside of vertebrates, but two *Gat* encoding genes were present in the jawless vertebrate lamprey species<sup>13, 14, 49, 53</sup>. Investigation of *Gai* across invertebrates (fruit fly, sea urchins, lancelet and tunicates) supported that *Gai* emerged prior to the emergence of the vertebrate lineages<sup>13, 14, 49</sup>. However, despite these findings, it was proposed instead that the duplication of *GNAI* into the *GNAI*—*GNAT* pair occurred before the divergence of protostomes and deuterostomes and that *GNAT* was subsequently lost in all invertebrate deuterostomes. It was only suggested as an alternative hypothesis that *GNAT* arose after the emergence of vertebrates, but it was unclear if this coincided with the 2R WGD<sup>49</sup>.

### **Chemosensory perception of taste**

Knowledge of *Gα* subunit evolution has been expanded from phototransduction into the chemosensory and neurosensory organs of taste and smell. Specifically, Oka *et al.* investigated how the nervous system in fish evolved to process external sensory information<sup>54</sup> as the chemosensory property of taste is closely related to food intake behavior. The report from Oka *et al.* suggests a non-redundancy of *Gα* subunits expressed in teleost fishes (zebrafish, pufferfish, medaka, *etc.*) as compared to the redundant expression of several *Gα* subunits in the taste bud and gustatory epithelium tissues seen in mammals<sup>55</sup>. Therefore biased evolution for tissue-specific expression of *Gai2*, *Gaq*, *Ga14*, and *Gat3* in mammals may have occurred. *Gat3* (gustducin) is lost in a lineage-specific manner in teleost fishes despite its role as a primary *Gα* subunit for the mammalian taste of bitter<sup>56</sup>, umami, and sweet flavors<sup>57</sup>; instead only *Gai1* is expressed in the taste cells of teleost fishes. These studies, though not as expansive as those centered on the visual system, represent the first steps to analyzing the evolution of other neurosensory signaling systems in animals involving *Gα*, outside of studies focused on GPCR-specific evolution.

### ***New theory of G $\alpha$ evolution in Opisthokonta***

Since the original cloning and genomic composition of the G proteins were solved, G $\alpha$  subunits from the five primary families have been specifically traced across the animal kingdom<sup>2, 12-14</sup>. Indeed much attention has been paid to G $\alpha$ s and G $\alpha$ i due to their foundational roles as the first G proteins to be discovered and purified<sup>46</sup>. G $\alpha$ t1 and G $\alpha$ t2 have also received attention due to their role in the expansion of the visual phototransduction pathway in vertebrates<sup>47-50</sup>. The evolution of the other G $\alpha$  subunit paralogs is less well studied outside of mammalian phylogeny and model organisms such as *C. elegans*<sup>58</sup>. Indeed, very little focus has been placed on the evolution of the G $\alpha$ q or G $\alpha$ 12 families due to the difficulties of expressing and purifying these proteins. However all five families of G $\alpha$  have critically shaped vertebrate evolution. Therefore understanding how they emerged across evolution before vertebrate and mammalian expansion is paramount to understanding the nature of their function and how that function has been modulated over time.

In addition, the G $\alpha$  subunit underwent a large radiation at the emergence of vertebrate evolution<sup>13, 14, 48, 49</sup>. Following the two rounds of whole genome duplication (2R WGD), new G $\alpha$  genes either underwent neofunctionalization, acted as redundant safeguards for preserving critical functions or were subsequently removed from the genome. Within vertebrates, new paralogs gained signaling and tissue specificity. It is clear from many studies that the G $\alpha$  subunit is the most evolutionarily dynamic component of the G protein signaling cascade, and it is the most susceptible to diversification<sup>1, 2, 13, 14</sup>.

The newest theory of G protein evolution (see Figure 1.4b) posits a progenitor or ancestral G $\alpha$  subunit existed in the last common ancestor of Eukaryota<sup>14</sup>. However instead of this ancestor duplicating several times to give rise to *GNAI2*, *GNAS*, and *GNAI/Q*, it

suggests that the ancestor *GNA*- subunit only duplicated into *GNAS*-like and *GNAI/Q*-like genes prior to the emergence of Opisthokonta. The *GNAI/Q* gene then duplicated and differentiated into the two separate *GNAI* and *GNAQ* genes prior to Holozoa. After duplication and differentiation, it is hypothesized that *GNAQ* was reinserted into the genome and the retrogene *GNAI2* emerged. In addition, the exon/intron structure of *GNAI2* supports the theory of its emergence through reinsertion and subsequent intron reinsertion; however, it is not entirely clear if *GNAQ* or *GNAI* was the progenitor sequence<sup>14</sup>; though Gα12 and Gα13 proteins do share functional selectivity with GPCRs which couple to the Gαq family<sup>59</sup>. Also in Holozoans, *GNAV* emerged through a duplication of *GNAI*<sup>14, 22</sup>. The similarity in gene structure and sequence support this hypothesis.

In the Metazoa lineage, *GNAO* arose through a duplication of *GNAI*. *GNAO* and *GNAI* both possess the conserved eight protein-coding exons indicative of the Gαi family. Also in Metazoa, the *GNAI* and *GNAQ* genes each independently underwent tandem duplications<sup>14</sup>. The exact timing of these duplication events is not known. However this is in contrast to the previous theories of G protein evolution which posit that *GNAI/Q* was tandemly duplicated before an additional duplication which led to the divergence of *GNAI* and *GNAQ*<sup>46</sup>. New evidence of the gene structure, particularly the protein-coding exon/intron structure, supports the theory that *GNAI* and *GNAQ* differentiated as separate genes before their independent tandem duplication events<sup>14</sup>.

At the root of vertebrates, two rounds of whole genome duplication (2R WGD) occurred<sup>51, 52</sup>. As the *GNA*- subunits are considered housekeeping genes<sup>24</sup>, many paralogs were maintained in the vertebrate genome. Specifically, the progenitor *GNAS* duplicated to

give rise to the *GNAS* and *GNAL* members. The progenitor *GNAI2* gene duplicated retained a second member, *GNAI3*. *GNAV* and *GNAO*, though both present before the 2R WGD, did not retain any duplicated copies in the vertebrate lineage. Instead, *GNAO* gained the ability to be alternatively spliced through the mutually exclusive inclusion of its final two exons (exons 7 and 8)<sup>14</sup>. *GNAV* was subsequently deleted, in a lineage specific manner, from Agnatha (jawless vertebrates – ie lampreys) and Sauropsida (amphibians, birds, mammals)<sup>13, 14, 22, 42</sup>.

The  $G\alpha q$  family retained two sets of gene pairs. These diverged into the genes *GNAQ*—*GNAI4* and *GNA11*—*GNAI5*. In the  $G\alpha i$  family, three of the four gene pairs of *GNAI*—*GNAT* were retained in vertebrates. There is some evidence that a fourth *GNAI* gene was maintained in the lamprey lineage alone<sup>14, 54</sup>. In addition, *GNAZ* arose as a retroinsertion of a  $G\alpha i$  family member after the 2R WGD. It has been hypothesized that *GNAZ* arose from *GNAI3*, but this hypothesis has not been validated in current reports<sup>14</sup>.

### 1.3 Regulation of G protein signaling across Eukaryota

#### *Two Primary G protein Signaling Mechanisms*

##### **Canonical 7TM receptor-GEFs and G $\alpha$ signaling**

Since their initial discovery in Metazoa (animals), heterotrimeric G proteins have been shown to couple to a seven transmembrane (7TM) receptor, known as their cognate GPCR, to initiate an intracellular signaling cycle (Figure 1.1b). The G $\alpha$  subunit of the heterotrimer is bound to GDP in its basal, non-signaling state. Upon agonist binding and activation of the GPCR, the membrane-associated heterotrimer couples to the 7TM receptor. This interaction induces a conformational change within the G $\alpha$  subunit that releases GDP in exchange for cytosolically abundant GTP. Therefore, the activated 7TM receptor acts as a guanine nucleotide exchange factor (GEF) and induces G $\alpha$  activation by overcoming the rate-limiting step of nucleotide exchange<sup>9, 60</sup>. Upon GTP binding, the activated G $\alpha$ (GTP) subunit dissociates from the GPCR-trimer complex and initiates downstream signaling through interaction with effector proteins. Along with G $\alpha$ (GTP) activation and dissociation, the G $\beta\gamma$  subunits are also freed in order to initiate their own downstream signaling pathways. The duration of signaling is tied to the hydrolysis of the  $\gamma$  phosphate of GTP which is cleaved through the intrinsic enzymatic ability of G $\alpha$ ; this results in GDP + Pi and therefore an inactive G $\alpha$ (GDP) subunit. G $\alpha$ (GDP) has a higher affinity for the G $\beta\gamma$  subunits than for the downstream effector proteins. The signaling is terminated when G $\alpha$ (GDP) re-associates with G $\beta\gamma$ , and another round of signaling can begin. The rate of GTP hydrolysis in Metazoa is tightly maintained by Regulators of G protein Signaling (RGS) proteins, which stabilize the transition state to encourage hydrolysis. In this way, the RGS protein act as a guanine triphosphatase (GTPase)-activating protein (GAP), which

modifies the integrity of signaling timing and duration by catalyzing hydrolysis and altering the rate of cleavage. In animals, there are no known transmembrane RGS proteins<sup>1, 2, 12</sup>. Instead the soluble RGS proteins function as GAPs and the 7TM GPCRs act as GEFs.

### **7TM receptor-GAPs and G $\alpha$ signaling**

Curiously, this canonical signaling mechanism was found to be inverted in non-animal species (Figure 1.1c), specifically land plants<sup>61</sup> (and presumably in other species of Archaeplastida and different Bikonta groups such as Excavata, Chromalveolata etc.<sup>1</sup>), though the G $\alpha$  subunits were found to be structurally similar to their animal counterparts<sup>62</sup>. Instead of the 7TM receptor acting as a GEF, in many plants, the G $\alpha$  subunit of the heterotrimer can spontaneously exchange nucleotide, swapping GDP for GTP, in order to activate<sup>62, 63</sup>. Nucleotide exchange is, therefore, no longer the rate-limiting step for G protein activation in plants<sup>61, 64</sup>. Instead, it is the hydrolysis of GTP to GDP + P<sub>i</sub> for signal termination which is rate-limiting<sup>61, 62</sup>. The self-activated and dissociated G $\alpha$ (GTP) subunit must interact with a 7TM receptor-RGS protein to catalyze hydrolysis<sup>65, 66</sup>. The receptor is still regulated by ligand binding<sup>67</sup>; however, it is the ligand-free 7TM receptor which inactivates heterotrimer signaling, making the plant 7TM receptor a GAP. When ligand-bound, the 7TM receptor is inhibited from catalyzing GTP hydrolysis (effectively inhibiting the inhibitor). Therefore, any activated G $\alpha$ (GTP) subunits are free to continue signaling and inactive G $\alpha$ (GDP) subunits may continue to exchange nucleotide for GTP spontaneously<sup>62, 65, 68</sup>. The fundamental difference in mechanism lies within the intrinsic G $\alpha$  properties of these two modalities. In one, stimulating the 7TM receptor-GEF induces G $\alpha$  activation and signaling; in the other, ligand binding blocks the GAP activity of the 7TM receptor-RGS on the self-activating G $\alpha$  subunit.



## ***Gα evolution and mechanisms of signal regulation in other Unikonta***

### **Fungi**

Gα subunits were first discovered in Fungi in 1993 from *Neurospora crassa*; this subunit was initially considered a Gαi family member<sup>69</sup>. Now, there are four known, distinct families of Gα genes within Fungi with many subfamily representatives<sup>70, 71</sup>; these paralogs have been renamed *GPAI-4*. Though all group closer to the animal Gαi family than to the other four animal Gα families (~55% sequence similarity to mammalian Gαi), they are distinct from Gαi in many ways<sup>2, 70</sup>. Indeed, these genes may have arisen from the putative ancestral *GNAI/Q* progenitor predicted to exist in Opisthokonta<sup>14</sup>. Their presence in many phyla and subclassifications of Fungi (Ascomycota, Basidiomycota, Mucoromycotina, and Chytridiomycetes) further supports the theory of a Gα subunit present in the last common ancestor of Opisthokonta<sup>1, 12, 70</sup>, though Gα subunits in Fungi and Metazoa (and Holozoa) have greatly diverged. Gα subunits within Fungi have been identified as critical signaling components necessary for growth and development, nutrition and pheromone sensing, mating, and various pathogenesis responses (reviewed in<sup>70, 72-74</sup>). As some Fungi genomes possess both 7TM receptor-GEF and 7TM receptor-GAP genes, (Table 2 from<sup>2</sup>), functional studies are required to ascertain if the mechanisms of Gα signaling relies on self-activation or a 7TM receptor-GEF.

### **Amoebozoa**

Amoebozoa is the sister group of Opisthokonta as both form the branch of Unikonta. Study of Amoebozoa Gα subunits and the cAMP receptor responses began in the late 1980's<sup>75-77</sup>. Eight Gα subunits have been found in *Dictyostelium* genomes<sup>78-80</sup>. Many groups have confirmed previous reports on the presence of Gα subunits and other signaling components (Gβγ<sup>81</sup>, 7TM receptors<sup>82-84</sup>, and RGS proteins<sup>1</sup> as well as finding arrestins<sup>2</sup>) in different species of Amoebozoa (Table 2). Gα appears to be a dynamic protein within this

group as some species have anywhere from one to 30 different G $\alpha$  subunits<sup>12</sup>, all of which are distinct from G $\alpha$  sequences found in Opisthokonta<sup>2</sup>. Like the canonical GPCR-GEF mechanism found in Metazoa, G $\alpha$  subunits in Amoebozoa do not appear to be self-acting as 7TM receptor-GEFs are required for inducing signaling<sup>77</sup>. This suggests that the GEF-regulated G $\alpha$  mechanism of signaling to be a common ancestral modality<sup>79</sup> which must have evolved prior to Unikont emergence.

***G $\alpha$  evolution in Bikonta (Archaeplastida, Chromalveolata, Rhizaria, and Excavata)***

The evolution of heterotrimeric G proteins in Bikonta (Chromalveolata, Archaeplastida, Excavata, and Rhizaria) appears to have differed significantly from the evolution of subunits in Unikonta. Outside of animal and fungal species, the next best studied group of G $\alpha$  subunits comes from land plants (Archaeplastida). Different species of land plants retained various numbers of G $\alpha$  subunits<sup>85</sup>; though it appears, unsurprisingly, that in general, diploid plants have very few G protein while polyploids may have more<sup>86</sup>. The emergence of G $\alpha$  was initially thought to coincide with the acquisition of terrestrial habitats before the discovery of G proteins in aquatic plants<sup>87</sup>; this pushed back the theory of G $\alpha$  emergence to coincide with embryophytic life cycles of “alternation generations” roughly 450 million years ago as both aquatic and land plants may share this cycle. Since then, G $\alpha$  subunits have been found in more “ancient” species, specifically within species of Charophyta green algae<sup>88</sup>; Klebsormidiophyceae (of the Charophyta branch) are the most primitive green plants to date shown to contain G proteins (G $\alpha$  and G $\beta$  no G $\gamma$ ). This suggested that the acquisition of embryophyte life cycles were not correlated with the emergence of G proteins but rather with more ubiquitous signaling behaviors from photoautotrophic organisms such as the appearance of structures cell division<sup>88</sup>.

To date, no solid evidence for GPCRs or 7TM receptor-GEFs have been found in plants (reviewed in <sup>86</sup>). Instead many putative plant GPCR-like proteins have been shown to not possess the 7TM domain topology or the ability to interact with and activate G proteins<sup>86, 89</sup>. Therefore, canonical 7TM receptor-GEF signaling is not suspected to be present in plants; 7TM receptor-RGS proteins appear to be the primary mechanism for G $\alpha$  GAP regulation. It should be noted that though 7TM receptor-RGS components have been found in many plants, a subgroup of monoconts, the cereals, lack functional 7TM receptor-RGS proteins entirely<sup>64</sup>. Instead, in these species, G $\alpha$  is capable of hydrolyzing GTP to GDP in the absence of RGS proteins, though other GAP proteins may exist <sup>64, 86</sup>. Given these varying reports, more studies are required to couple genome-wide searches for 7TM receptor-GEF versus -GAP sequences with functional and structural evidence.

Recently, G $\alpha$ , G $\beta$  and G $\gamma$  subunits, 7TM receptors, RGS proteins, and other components of the G protein signaling cascade have been found in all other branches of Bikonta<sup>1, 2, 12</sup>. However, each of these reports have conflicting views on the presence and gene counts for each of these moieties. In Excavata, Bradford *et al.* suggests there are no 7TM receptors-GEFs at all, only 7TM receptor-GAP sequences<sup>1</sup>. Other studies<sup>2, 12</sup> have also found several hundred 7TM receptors in various branches of Excavata, but with more than half of them corresponding to GPCR-like or 7TM receptor-GEF sequences. These analyses have been expanded into the Chromalveolata and Rhizaria branches with the same dichotomy between their results<sup>1, 2, 12</sup>. Looking towards the G protein, the G $\alpha$  subunit appears to be present in most branches of Bikonta, though there are some lineage specific deletions; again, these deletions are not consistent across all studies. Therefore, conflict remains between the genome-level studies themselves based on the species evaluated, the

sensitivity of their search paradigms, and other metrics within their analyses.

### ***Conflicting views on G $\alpha$ signaling mechanisms across evolution***

#### **7TM receptor-GAPs are ancestral**

Following in this view of Archaeplastida G protein evolution, recent work posited that the 7TM receptor-GAP mechanism of G protein signaling in Bikonta was ancestral to the 7TM receptor-GEF modality found in Unikonta<sup>1</sup>. Indeed, 7TM receptors with carboxy-terminal, intracellular RGS protein motifs have been found in Archaeplastida, Fungi, and the “ancient” branch of Excavata (within genera of *Trichomonas* and *Naegleria*). Barring some potential lateral transfer between *Trichomonas* and Archeplastida, the 7TM receptor-RGS appeared to be a primary component of the self-activating G $\alpha$  signaling modality in these branches<sup>12</sup>. No 7TM receptor-RGS proteins exist in Metazoa<sup>1, 2, 12</sup>. Therefore, 7TM receptor-GEF activity was proposed to be a more evolutionarily recent development within Unikonta and radiated in the lineage of Vertebrata.

#### **7TM receptor-GEFs are ancestral**

This is contrary to work that suggests GPCR-GEF activity was present in the last common ancestor of Eukaryota<sup>2</sup>. Outside of Archaeplastida, it is unclear whether 7TM receptors, like the GPCR, maintained the canonical GEF activity or if the plant-like GAP role is more universal as there is limited functional data across this broad range of species<sup>1, 2</sup>. One recent study evaluating gene presence and absence at the system level of 7TM receptors, G proteins, and other GEF and GAP proteins (Table 2) suggested that some species across the six primary branches of Eukaryota may maintain both 7TM receptor-GEF and 7TM receptor-GAP mechanisms of signaling within the same organism<sup>2</sup>. Therefore, G $\alpha$  may be both “self-activating” or “GEF-regulated” depending on protein expression and microdomain architecture.

In addition, some species possess the canonical GPCR signaling machinery without evidence of any heterotrimeric G protein components, suggesting that GPCR-like, 7TM receptor-GEF signaling may have evolved independently of G proteins in some species. Contrariwise, the reverse was also found where some G $\alpha$  subunits were present in species that did not possess any 7TM components. Taken together, this suggests that there is not “a conserved self-activation” family of G $\alpha$  subunits<sup>2</sup> but rather many variations around the same theme. It has been proposed that self-activation may have arisen through convergent evolution of G $\alpha$  subunits across Bikonta instead of being the primary, ancestral modality. Therefore, canonical 7TM receptor-GEF signaling (as seen in classical animal studies) may be an older mechanism of signaling as all of the components were present in the last common eukaryotic ancestor. Lineage-specific evolution and diversity then promoted the various signaling mechanics seen today<sup>2</sup>. Again, functional studies in species, which retain both “self-activating” and “GEF-regulated” components, are necessary to tease apart the divide between these modalities and ascertain how each mechanism functions across lineages.

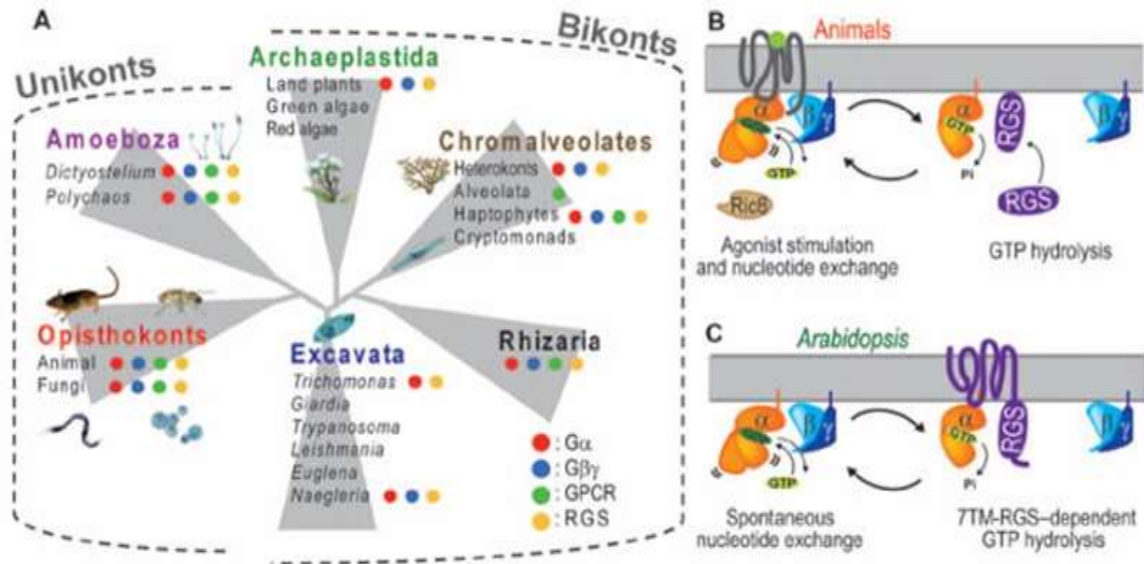
## 1.4 Conclusions

### *Conclusions*

As G $\alpha$  subunits have been found in all six major branches of Eukaryota, despite lineage-specific losses, it can be concluded, at least, that G $\alpha$  was present in the last common ancestor of Eukaryota<sup>1, 2, 12</sup>. In addition, 7TM receptors, RGS proteins, GEFs, GAPs, and other G protein signaling components have been traced across Eukaryota evolution<sup>1, 2, 12</sup>. Each branch tailored the G protein signaling system to their own needs by retaining, modifying or deleting different signaling machinery within the cascade. Functional studies across this wide range of species will be necessary to fully ascertain the evolution of G $\alpha$  self-activation versus GEF-activation signaling mechanisms. However, genome level analysis of the signaling components has allowed for new theories of G protein and 7TM receptor signaling mechanics, and has paved the way to begin critically assessing the functionality of each signaling model.

### *Abbreviations*

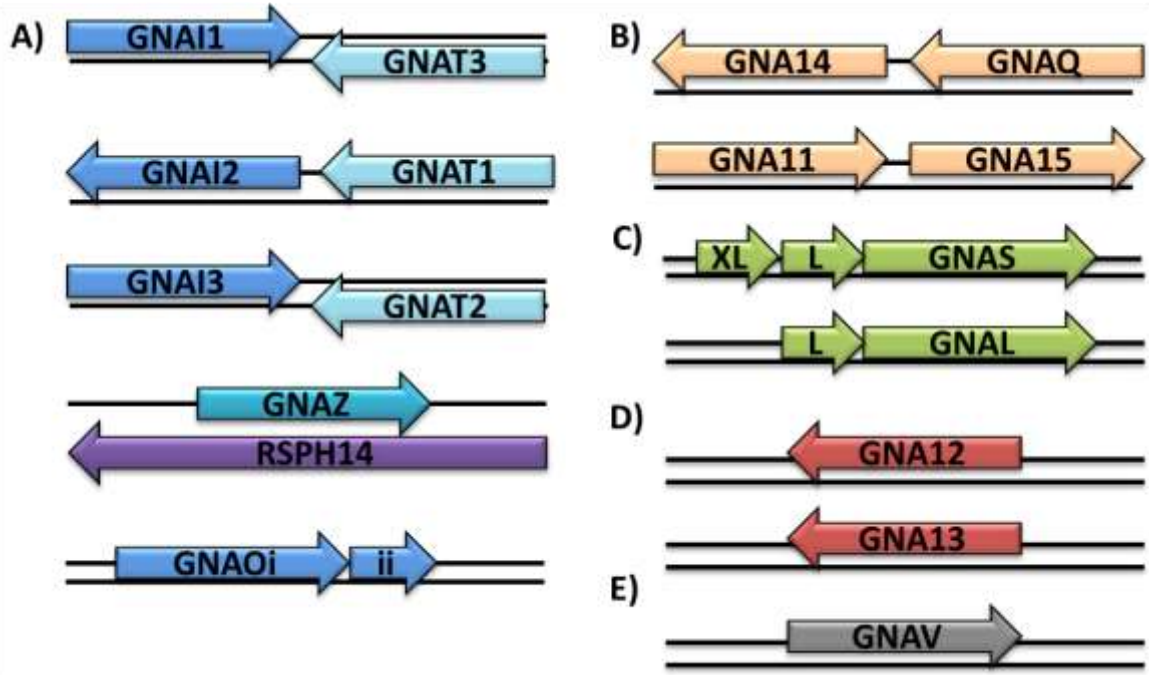
GPCR	- G protein coupled receptor
7TM	- 7 transmembrane (receptor)
GDP	- Guanine diphosphate
GTP	- Guanine triphosphate
cAMP	- cyclic Adenosine monophosphate
cGMP	- cyclic Guanosine monophosphate
GEF	- Guanine nucleotide exchange factor
RGS	- Regulators of G protein Signaling
GTPase	- Guanine triphosphatase
GAP	- GTPase Activating Protein
2R WGD	- Two Rounds of Whole Genome Duplication



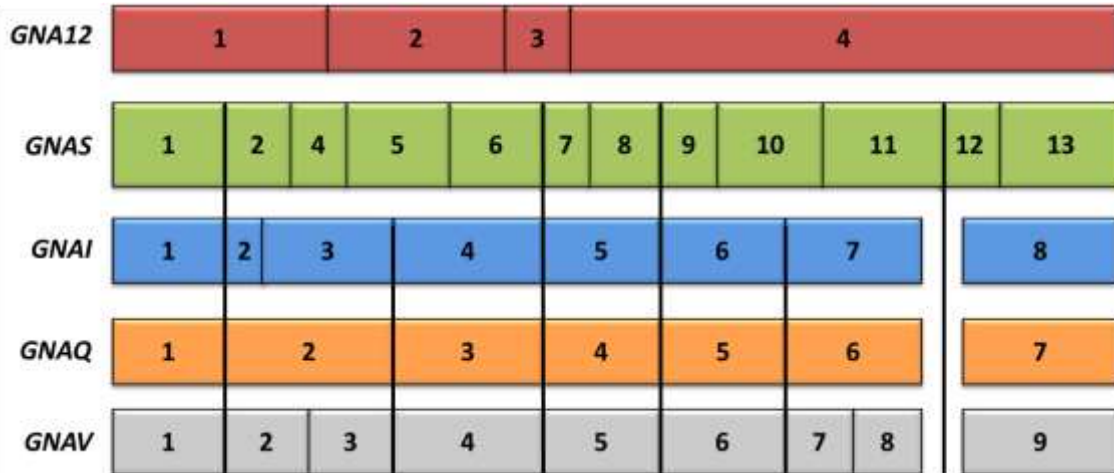
**Figure 1.1: Unikont and Bikont Signaling through G proteins: Distribution of G protein components among eukaryotes.** (A) The indicated taxa are representative genomes. The presence of G protein elements in the indicated species or lineages is represented by red, blue, green, and yellow dots for genes encoding G $\alpha$ , G $\beta\gamma$ , Opisthokont GPCRs, and RGS proteins, respectively. Lack of a dot signifies that those genes were not found. We organized the eukaryotes into six supergroups: Opisthokonta (containing *C. owczarzaki* and *H. sapiens*), Amoebozoa (containing *D. discoideum*), Archaeplastida (containing *A. thaliana*), Excavata (containing *T. vaginalis*), Chromalveolata (containing *E. siliculosus*), and Rhizaria. (B) Regulation of G protein activation in animals. Ligand-bound GPCR accelerates the dissociation of GDP from the G protein  $\alpha$  subunit by changing the orientation of its helical domain. G $\alpha$  hydrolyzes GTP, thereby inactivating itself. GTP hydrolysis is promoted by an RGS or other GAP protein. Nonreceptor GEFs, such as the protein Ric8 (resistance to inhibitors of cholinesterase), act as noncanonical and cytosolic GEFs. (C) Regulatory model of G protein signaling in *Arabidopsis*. The *Arabidopsis* G $\alpha$  protein AtGPA1 rapidly releases its GDP as a result of spontaneous fluctuations between its Ras domain and helical domain. AtGPA1 slowly hydrolyzes its bound GTP; however, the membrane-localized 7TM-RGS protein AtRGS1 constitutively promotes GTP hydrolysis or acts as a GDI. Figures 1.1 was reprinted from <sup>1</sup>. Bradford, W., Buckholz, A., Morton, J., Price, C., Jones, A. M., and Urano, D. (2013) Eukaryotic G protein signaling evolved to require G protein-coupled receptors for activation, *Sci Signal* 6, ra37. Reprinted with permission from AAAS.



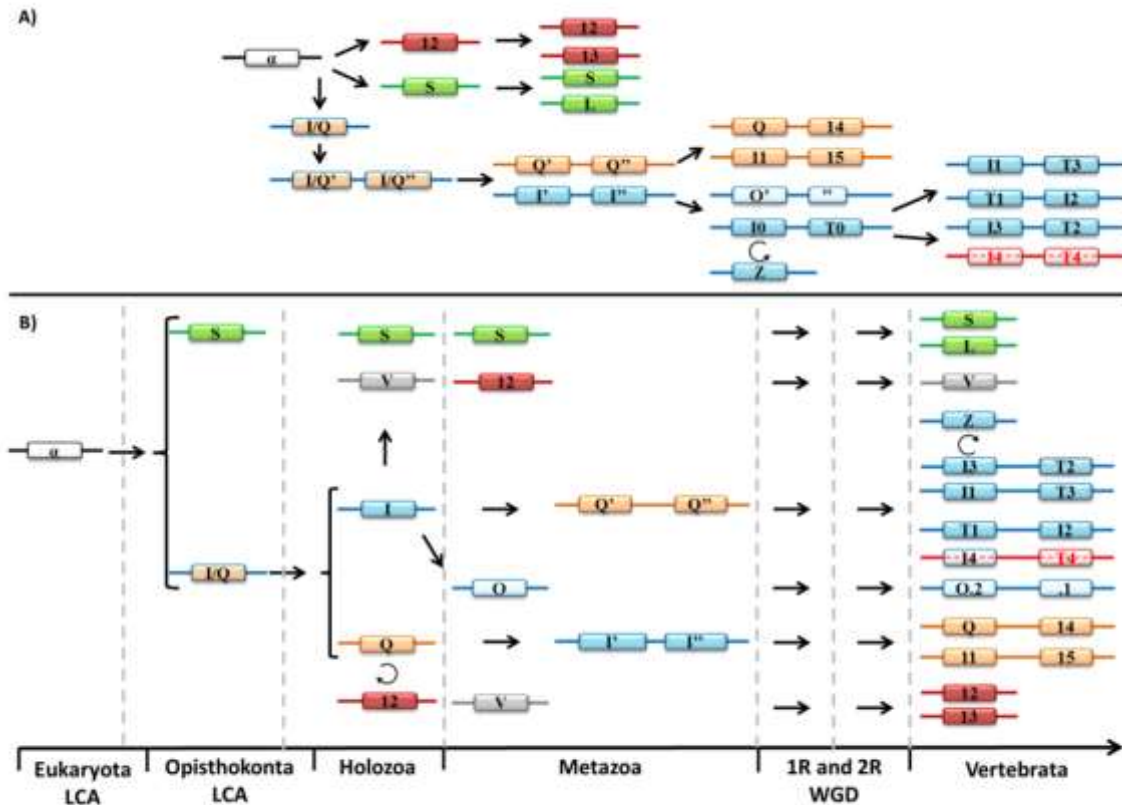




**Figure 1.2: Vertebrate GNA- paralogous gene arrangements.** The *Gai* subfamily consists of 8 paralogs (blue). *Gai* & *Gat* genes (*GNAI* & *GNAT*) are arranged in a head-to-head arrangement (except *GNAT1* & *GNAI2* which are head-to-tail) suggestive of a tandem duplication. *Gaz*'s gene (*GNAZ*) lies within an intron of a protein-coding gene on the opposite strand. *Gao*'s gene (*GNAO*) can be alternatively spliced to create 2 separate mRNA transcripts (*GNAOa* & *GNAOb*). B) The *Gaq* subfamily consists of 4 paralogs (orange). All are arranged in a head-to-tail arrangement suggestive of a tandem duplication. C) The *Gas* subfamily is composed of 2 paralogs (*GNAS* & *GNAL*) which can both be alternatively spliced into extra-long (XL) mRNA (green). D) The *Ga12* subfamily also consists of 2 paralogs (*GNA12* & *GNA13*) (red). E) The newly discovered *Gav* subfamily is composed of one paralog (*GNAV*) (grey) and is only found in a select number of species.

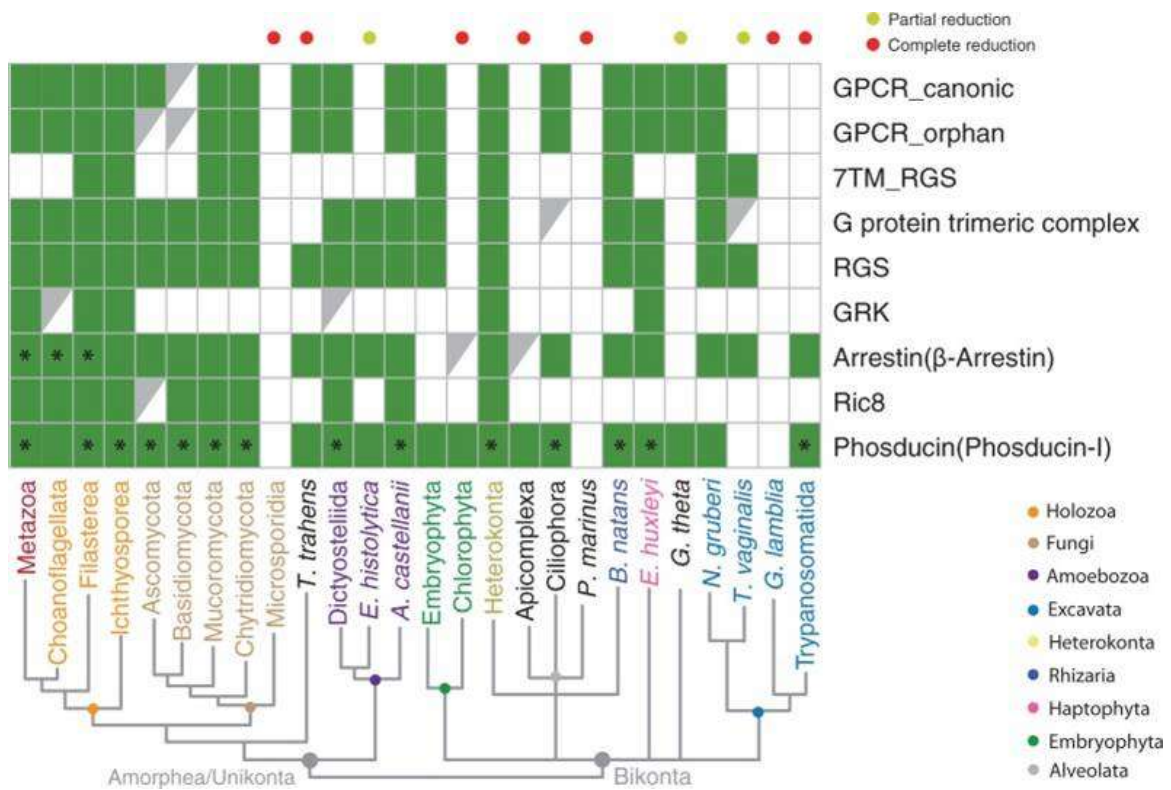


**Figure 1.3: Aligning representative vertebrate protein-coding exon borders of all five major families of the *Ga* subunit.** The highly conserved exon border positions give insight into the evolutionary divisions of *GNA*- genes. All protein-coding exons are represented as boxes which correlate with the curated average exon size (introns removed). *GNAI* and *GNAQ* share many exon borders positions (black lines) and four split codons (not shown) suggesting a closer evolutionary relationship. *GNAV* also shares six exon border positions with *GNAI* and *GNAQ*; this suggests that *Gav* family is related to *Gai* and *Gaq* despite its gene presence in a limited number of species. All three genes share four exon borders positions with *GNAS* (not considering the alternatively spliced exon3 or the extended exon4 of *GNAS* found in placental mammals). The lack of shared exon borders between *GNA12* and the other subfamilies suggests that *GNA12* may have originated as an independent retro-gene which independently gained introns. Figure taken from <sup>14</sup> *Manuscript in Submission*.



**Figure 1.4: Two theories of G protein evolution. A)** Summary of previous theories of  $G\alpha$  evolution without relative timelines<sup>46-48, 50</sup>. An ancestral  $GNA$  ( $\alpha$ -white) underwent a series of duplications before diverging into three primary progenitor families. The progenitor  $GNAI/Q$  tandemly duplicated before undergoing a larger regional or chromosomal duplication. These gene pairs diverged into  $GNAI$ -like (blue) and  $GNAQ$ -like (orange) genes.  $GNAS$  (green),  $GNA12$  (red),  $GNAQ'$ - $GNAQ''$ , and  $GNAI'$ - $GNAI''$  all duplicated to give rise to two copies from each parent.  $GNAI'$ - $GNAI''$  duplicated into  $GNAO'$ - $O''$  (ultimately an alternatively spliced gene) and  $GNAI0$ - $GNAT0$  followed by two more duplications of  $GNAI0$ - $GNAT0$ .  $GNAZ$ , a retrogene of  $GNAI0$ , was reinserted into the genome before the  $GNAI0$ - $GNAT0$  duplications. **B)** New theory of  $G\alpha$  subfamily evolution incorporating current reports<sup>1, 2, 13, 22, 42, 54, 88</sup> with relative timelines included (not fit to scale). An ancestral  $preGNA$  progenitor ( $\alpha$ -white) duplicated into the  $preGNAI/Q$  progenitor and  $preGNAS$ .  $preGNAI/Q$  duplicated into two separate genes that diverged into  $preGNAI$  and  $preGNAQ$ . Then  $preGNAV$  arose from a duplication of  $preGNAI$ .  $preGNAI2$  is a retrogene, possibly of  $preGNAQ$ , though its precise origin is unclear.  $preGNAI$  later duplicated to give rise to  $preGNAO$ . Both  $preGNAI$  and  $preGNAQ$  underwent independent tandem duplication events before the 2R WGD of vertebrates.  $GNAS$ ,  $GNA12$  and  $GNAQ'$ - $GNAQ''$  all retained two copies after the 2R WGD leading to vertebrates, while other hypothetical copies (not shown) were lost immediately after the 2R WGD and are not observed in any extant species.  $GNAI'$ - $GNAI''$  retained three copies of this gene pair after the 2R WGD ( $GNAI4$  remains only in lampreys). Other, lineage-specific deletions occurred for  $GNAV$ ,  $GNAT3$ ,  $GNAI4$ , and  $GNAT4$  as described in the main text.  $GNAO$  gained alternative splicing ability after 2R WGD (O.2-1). The retrogene  $GNAZ$  emerged in the vertebrate lineage from a  $GNAI$  gene. Lineage-specific duplications and retrogenes are not included for clarity. Straight arrows depict duplications

(local, tandem duplications, or WGD), curved arrows depict retrotranspositions. Complete gene names for were simplified for clarity; curated *preGNA*- genes are denoted with “pre-” while “GNA” is removed for clarity in all paralogs. LCA = Last Common Ancestor  
Figure taken from <sup>14</sup> *Manuscript in Submission*.



**Table 1.2: G protein signaling components found across Eukaryota.** Figure reprinted from <sup>2</sup>. de Mendoza, A., Sebé-Pedrós, A., and Ruiz-Trillo, I. (2014) The evolution of the GPCR signaling system in eukaryotes: modularity, conservation, and the transition to metazoan multicellularity, *Genome Biol Evol* 6, 606-619 by permission of Oxford University Press or the sponsoring society.

**TRACING THE EVOLUTIONARY HISTORY OF THE HETEROTRIMERIC G  
PROTEIN  $G\alpha$  SUBUNIT IN DEUTEROSTOMIA**

**2.1 Introduction**

***Chapter 2***

Lokits AD, Indrischek H, Meiler J, Hamm HE, Stadler PF, “Tracing the evolutionary history of the heterotrimeric G protein  $G\alpha$  subunit in Deuterostomia” *BMC Evolutionary Biology* **2017** *in submission*

***Contribution***

I am first author of this manuscript. I contributed to most sections of the text: abstract, introduction/background, methods of species investigated and structural analysis, analysis of intron/exon structure, results, discussion, conclusions, and appendices. I created Figures 2.1-2.10, Table 2.1, and Supplemental Figures 2.1-2.4, 2.10, part of 2.11, and Supplemental Tables 2.1 and 2.2. I edited all sections and all other supplemental figures and tables. I also reviewed all contributions from all other authors.

***Abstract***

**Background:** Heterotrimeric G proteins are fundamental signaling proteins composed of three subunits,  $G\alpha$  and a  $G\beta\gamma$  dimer. The role of  $G\alpha$  as a molecular switch is critical for transmitting and amplifying intracellular signaling cascades initiated by an activated G protein Coupled Receptor (GPCR). Despite their biochemical and therapeutic importance, the study of G protein evolution has been limited to the scope of a few model organisms. Furthermore, only two of the five primary  $G\alpha$  families have been thoroughly investigated outside of mammalian evolution. Therefore our understanding of  $G\alpha$  emergence and evolution across phylogeny remains incomplete.

**Results:** We have computationally identified the presence and absence of every  $G\alpha$  gene ( $GNA$ -) across all major branches of Deuterostomia and evaluated the conservation of the underlying exon-intron structures across these phylogenetic groups. In addition to our curated gene annotations, we have identified nuanced differences in phyla-specific gene

copy numbers, novel paralog duplications and subsequent intron gain and loss events, which substantially alter previous interpretations of G protein evolution in Deuterostomia. We provide evidence of mutually exclusive exon inclusion through alternative splicing of *GNAI*, *GNAQ*, and *GNAS* transcripts in specific lineages and that *GNAO* gained alternative splicing ability co-occurring with the emergence of Vertebrata. Variations in alternative splicing signals and isoforms were found for several paralogs, which coincide with conserved, putative motifs of DNA-/RNA-binding proteins. Our results also indicate that both *GNAI2* and *GNAZ* originated from retrogenes. Within primates, we identified 15 retrotranspositions, many of which have undergone pseudogenization. Most importantly, we find significant deviations from previous findings regarding the presence and absence of individual *GNA*- genes.

**Conclusions:** Our curated annotations allow us to draw more accurate inferences regarding the emergence of all  $G\alpha$  family members across Deuterostomia and to present a new, updated theory of  $G\alpha$  evolution. These observations regarding the evolution of the *GNA*- genes translate into new understanding of the  $G\alpha$  protein family. Leveraging this, our results are critical for gaining new insights into the co-evolution of therapeutically relevant G protein – GPCR signaling pathways and their radiation in Vertebrata.

## ***Background***

### **Evolution of $G\alpha$**

G protein Coupled Receptors (GPCRs) are a highly studied class of receptors due to their integral role in cellular signaling and therefore as therapeutic targets. Their evolution has shaped the chemical and biomolecular signaling systems of eukaryotes<sup>1, 2</sup>. Within this signaling cascade, a transducing element, the heterotrimeric G protein, composed of a monomeric  $\alpha$  and obligate  $\beta\gamma$  dimer, acts as an intracellular relay for activated GPCRs to

convert their message into an amplified signaling cascade. With only 16 paralogs in humans, compared to the 800 GPCR genes, the evolution of heterotrimeric G protein  $\alpha$  subunit has received less attention than their transmembrane protein partners.

Shortly after their initial discovery and sequencing in several mammalian species, the  $G\alpha$  subunit was found to be a highly conserved housekeeping protein<sup>46</sup>. As such, traces of genes encoding heterotrimeric G protein  $\alpha$  subunits (*GNA*-) have been found in almost all major branches of eukaryotes<sup>1, 13, 88</sup> despite the differences in GPCR signaling mechanisms between Unikont and Bikont lineages (see<sup>1</sup>).

Using only mammalian sequences, the first theory of G protein  $\alpha$  evolution posited the relative evolution of four of the five  $G\alpha$  families ( $G\alpha_i$ ,  $G\alpha_q$ ,  $G\alpha_s$  and  $G\alpha_{12}$ ;  $G\alpha_v$  having not yet been discovered)<sup>46</sup>. For their analysis, human, mouse, rat, and bovine sequence similarities and gene arrangements were the foundation for predicting how each of the family and subfamily members arose relative to each other. From this limited data set, conjectures of the timing of evolutionary divergence and lineage-specific variation were not possible.

Focusing on the development and radiation of the visual system, others have evaluated the evolution of transducins (*Gat1* and *Gat2*) and other critical proteins in the visual signal transduction pathway in both rods and cones across Vertebrata and into Chordata lineages<sup>47-50</sup>. However, to our knowledge, there have been no reports focused on studying the evolution of the other three families of  $G\alpha$  in Deuterostomia with the exception of  $G\alpha$  subunits in the fish chemosensory systems<sup>54</sup>, and a more recent, coarse-grained study evaluating paralog counts across Opisthokonta phylogeny<sup>13</sup>.

From these studies and others, we have compared our estimation of when each paralog



emerged within Deuterostomia evolution. We have found significant differences in the timing and number of predicted gene gain and loss events, due to a) differences in methodologies employed while searching for paralogous sequences and constructing phylogenetic trees and b) increased search space through the inclusion of more genomes. In addition to reporting new and manually curated gene annotations, we have also uncovered variations in alternative splicing patterns, non-canonical splice sites (SS), novel intron gain and loss events, primate gene retrotranspositions and subsequent pseudogenization, as well as other nuanced deviations to the gene structure of this family. These data allow us to present an updated view on G protein  $\alpha$  subunit evolution.

***Significance – New Understanding of Paralog Gains and Losses***

Efforts of many genome sequencing consortiums and single laboratories<sup>91-95</sup>, advances in sequencing techniques and assembly algorithms as well as manual curation of public databases<sup>96-100</sup> have resulted in the steady improvement of genome annotation quality and diversity in the species covered. We have taken advantage of these advances in the field to expand our understanding of heterotrimeric G protein evolution. Here, we computationally identified the presence and absence of every G $\alpha$  gene (delineated as “GNA-” following the HUGO convention of gene names<sup>101</sup>) across all major branches of Deuterostomia (where genome assemblies exist) and evaluated the conservation of the exon-intron structure across evolution. We find nuanced differences in alternative splicing signals, alternative isoforms, conservation of putative motifs for DNA-/RNA-binding proteins (DBP/RBP), as well as intron gain and loss events, and thus gain insight into the evolution of the G $\alpha$  protein family. Most importantly, we find significant differences in gene presence and absence as compared to previously published findings<sup>13, 47-49</sup>; this allows us to draw more

accurate inferences on the timing of emergence across individual paralog families.

The strength of this study comes from the inclusion and curation of genes from highly fragmented genome assemblies covering all major branches of Deuterostomia in addition to the genomes of well-studied model organisms (Figure 2.1). Despite improved long-read genome sequencing techniques, computational assembly of accurate whole genome sequences remains a challenge<sup>102</sup>. Roughly 1/3 of all publically available eukaryotic genomes are assembled only on the contig level. Contigs are sets of overlapping nucleotide reads, which are ultimately assembled into larger blocks (scaffolds) and ideally into chromosomes. Long protein-coding genes can span multiple contigs. High sequence similarity between genes due to homology remains challenging when assembling DNA-seq reads into larger scaffolds or when mapping RNA-seq reads to a genome. The ambiguity of these regions can result in chimeric gene annotations where two different genes are presumed to be one. Additional errors can be introduced via automated gene prediction tools which probe the assembly; these tools may not allow for partial gene models or may combine sequence “hits” into merged chimeric gene predictions. For a more thorough examination of these hurdles please see<sup>91, 102</sup>.

The ExonMatchSolver (EMS) algorithm<sup>102</sup> was developed to assist in curating highly fragmented genome assemblies when the query protein family possesses multiple paralogs with low sequence divergence and conserved exon-intron structure. EMS differs from other methodologies by querying for the collective “match” of all paralogous genes of a protein family within an individual genome assembly. As the family of heterotrimeric G proteins contains many paralogs, we used the EMS technique to annotate and disambiguate all possible paralogs of the G $\alpha$  subunit across Deuterostomia. Despite its usefulness, it is of

note that the EMS pipeline does not resolve inversions of exons or significantly altered exon-intron structures. Instead this tool provides contexts for manually resolving such ambiguities in the nucleotide sequences.

Through the use of the EMS pipeline to assist in the curation of the *GNA*- genes across a dense species sampling, we have identified dozens of sequence deviations and inconsistencies within the examined species and paralogs compared to previous works and genome annotations. In this work, we have uncovered many paralogs of *GNA*- not identified by previous methodologies; this is likely due to the use of coarse-grained approaches, which misidentified the presence and absence of genes, and/or due to the reliance on gene trees covering a limited range of species. Along with these newly predicted paralogs and improved gene annotations, we have also found previously unidentified alternative splicing patterns, putative transcription factor binding sites, altered exon-intron borders, and duplications and reinsertions of genes across Deuterostomia phylogeny. The culmination of this information has led us to propose a new theory of the evolution for  $G\alpha$  proteins in regards to the presence of these genes around the multiple whole genome duplication (WGD) events within Vertebrata. We present support for new theories regarding the emergence of each *GNA*- gene in Deuterostomia. We also provide detailed appendices on additional variations to the conserved gene structures, alternative splicing events and other deviations as supplemental material.

## 2.2 Materials and Methods

### *ExonMatchSolver*

Genomes were analyzed for curated annotation within the ExonMatchSolver (EMS) framework according to its *Implementation and Usage*<sup>102</sup> utilizing both paralog-specific, individual translated coding exons (TCE) and full paralog sequences. Briefly, the EMS pipeline utilizes TCEs as the fundamental building blocks for its searches. Therefore we utilized the paralog-specific TCE amino acid (AA) sequences of a close relative to the target species as the query against the target genome. There are 16 *GNA*- genes within humans. As each family was expected to have a conserved exon-intron structure throughout Deuterostomia, the high quality annotations of human *GNA*- genes were utilized as the initial templates. Sister groups of Mammalia were evaluated next, before moving on to more distant families. For each major clade (Sauropsida, Amphibia, Actinopterygii, etc.), curation began within the species assembly with the highest reported sequence coverage, genome quality and level of annotation. This curated sequence was used as a seed TCE query for further analysis within that clade. A minimum of two orthologs were used as individual inputs for the *hmmsearch* when querying each target assembly. In addition to exon border position information, EMS also utilizes full-length protein sequences to annotate orthologous proteins along the target genome assembly via a spliced alignment<sup>102</sup>. A minimum of two orthologs from closely related species were utilized as protein sequence queries for the target spliced alignment.

### ***Data Sources***

A total of 60 species were evaluated; 45 of which were directly assessed through the EMS pipeline for curated gene annotation (see Supplemental Table 2.1); the additional species were utilized for supplemental assays as described (Supplemental Figure 2.1). All queried genomes were obtained from public repositories<sup>92-94, 97, 100, 103</sup>. The latest version of each genome was utilized for all analyses unless otherwise noted (*as of October 2016*). All major phylogenetic clades of Deuterostomia were investigated with the EMS pipeline, including 26 species of Sarcopterygii, or lobed-finned fishes (composed of ten species from Mammalia, eight from Aves, five from non-avian Sauropsida, two from Amphibia and one from Coelacanthiformes). To interrogate the Teleostei-specific third round of whole genome duplication (3R WGD), we included one genome of a non-duplicated Actinopterygii (ray-finned fishes) species and six Teleostei species which evolved after the 3R WGD. To evaluate the influence of the 2R WGD on *GNA*- evolution, we included one Chondrichthyes (cartilaginous fishes), two Agnatha (jawless vertebrates), two Cephalochordata (non-vertebrate chordates), four Urochordata (non-vertebrate chordates), one Hemichordata (non-chordate deuterostomes), and two Echinodermata (non-chordate deuterostomes) species. We included the following outgroups to our analyses: *D. melanogaster* (Arthropoda) and *C. elegans* (Nematoda) as representatives of Protostomia, in addition to *N. vectensis* (Cnidaria) and *T. adhaerens* (Placozoa) as representatives of Metazoa predating the emergence of Bilateria. To reflect the orthology relationship, all *GNA*- genes which predate the radiation of Vertebrata are denoted as *preGNA*- for clarity, as recommended by the HUGO convention of gene names<sup>101</sup>.

We utilized the Ensembl genome browser<sup>97, 98</sup> and NCBI's genome and assembly browser<sup>100</sup> for our starting queries as these databases contain easily accessible and high

quality genome annotations. To validate gene gain and loss events, we evaluated the transcriptome shotgun assembly (TSA) sequence database, expression sequence tag (EST) database, and UniGene databases, accessed through NCBI<sup>100, 103-105</sup>, using amino acid-based (*tblastn*) search queries. It is important to note that tissue-specific expression of some paralogs may hinder sequence validation through this approach. Synteny information (co-localization with neighboring genes) was also utilized in evaluating paralog assignments and gene loss, when available, through the Ensembl and NCBI genome browsers. The species tree that was used for mapping gene gain and loss events (Figure 2.1) is based on screening of recent literature and the consensus therein<sup>91, 106-108</sup>.

### ***Reconstruction of Gene Trees***

In order to build phylogenetic maximum likelihood (ML) trees on the nucleotide and amino acid level using RAxML protocols<sup>109, 110</sup>, exonic, protein-coding sequences of interest were aligned using ClustalOmega<sup>90</sup> or MUSCLE<sup>110</sup>, edited with the Jalview alignment editor<sup>111</sup> and handed over to RAxML<sup>112</sup>. The appropriate amino acid or nucleotide substitution model for each tree was determined through Prottest<sup>113</sup> and additional tree parameter optimizations were conducted through preliminary rounds of ML searches comparing the different models of rate heterogeneity available in RAxML (Gamma, CAT, and a variable heuristics optimization<sup>109, 114</sup>). Random starting trees were also employed for initial independent ML tree searches to determine if random starting trees improved topology search space over a maximum parsimony starting tree. After optimizing the substitution model with the best model of among-site versus per-site heterogeneity rates and starting tree, the ML trees were compared for their diversity across tree topology. The strength of the phylogenetic signal was assessed through comparison of

the best likelihoods, and pairwise-Robinson Fould (RF) distance calculations were conducted across all independent searches. Production runs calculated support values for all ML trees and utilized *bootstopping* for all bootstrap replicates to decrease computational time. Bootstrapped replicates were summarized into Extended Majority Rule Consensus Trees and reported with bootstrap (BS) values as supplemental files. Pairwise-RF distance calculations across topologies as well as a Shimodaira and Hasegawa test were used to confirm that differences between likelihoods were not significant before summarizing into consensus trees. In addition to inferring support values through bootstrapping, the approximate likelihood-ratio test was also utilized for statistical validation.

#### ***Investigation of Protein-binding Motifs within DNA/RNA sequences***

Centrimo<sup>115</sup> was used to perform a local (positional) enrichment analysis of *in vivo* and *in vitro* DNA- and RNA-binding protein (DPB/RBP) motifs from the following databases: Ray 2013<sup>116</sup>, Jolma 2013<sup>117</sup>, Jaspar Core database 2014<sup>118</sup>, BS Uniprot<sup>96</sup> mouse. Centrimo evaluates absolute enrichment of a motif by performing a binomial test to determine whether the best match motif counts at a specific position are significantly different from a uniform motif distribution. Centrimo was also run in differential mode to conduct a Fisher's exact test to determine positional motif enrichment in a primary sequence set in comparison to a control set.

First, the potential overlap of all conserved non-canonical splice sites (SS) (the 5' 'GC' SS of intron6 in *GNAII*, and the 3' 'TG' SS of intron3 in *GNAS*) with DBP/RBP motifs were interrogated by testing differential motif enrichment in the nucleotide sequence surrounding the SS (full length exon sequence and 40 nt of the intronic sequence). All

primary orthologous sequences conserved the non-canonical SS, while the control set contained sequences with the canonical SS at the orthologous position. Second, the positional enrichment of potential DBP/RBP motifs were investigated within exon3 of *GNAS* and the surrounding conserved region by performing an absolute, local enrichment test. Homologous sequences were extracted from an additional 27 placental mammals from the Ensembl webserver<sup>97</sup> to form a total dataset of 33 species.

### ***Detection of Retrogenes in Primates***

The longest protein-coding isoform of each human *GNA*- gene was blasted against the human genome. Sequence matches overlapping annotated retrogenes were extracted at the nucleotide level via the Ensembl webserver<sup>97</sup> (*GNAI2P2*, *GNAI2P1*, *GNAQP1*, *GSI-124K5.9*, *RP11-611O2.6*, *AC010975.2*, *RP11-100N3.2*). 11 target primate genomes (Supplemental Figure 2.1) were then queried using these human *GNA*- pseudogene annotations. Primate retrogenes were retrieved as single blast hits with the following settings: *blastn*; *e-value* cutoff:  $10^{-5}$ ; match/mismatch: 1, -3; and opening/extension: 5, 2. Additional synteny (gene co-localization) information was also considered when identifying potential retrogenes. In cases with short scaffold lengths and no available synteny information, full length parent genes were re-blasted against the putative target loci. Loci that retrieved multiple, subsequent sequence matches were then excluded. A single sequence match was considered to be an individual exon of a multi-exon paralog if it covered less than 50% of the query sequence. Cases of 30-50% query coverage were manually inspected to identify exon borders.

Conserved open reading frames (ORFs) between orthologous retrogenes that showed similarity to the multi-exon paralog were interrogated. These potential ORFs within the



retrogene loci (Blast hit +/- 300 nt) were identified with ORF Finder<sup>119</sup> and similarity to the parent protein confirmed by blast (bl2seq -n blastp). Then new ORFs with coding potential that were not similar to the parent protein sequence were investigated. For this purpose, the retrogene loci were aligned with ClustalOmega<sup>90</sup> and coding potential was accessed with RNAcode<sup>120</sup> probing at least four different reference species. Sequence hits were reported if the region was conserved in all primates and contained at least one methionine as a possible initiation codon for translation.

#### ***Detection of Natural Selection in GNAO***

The branch-site model implemented in CODEML in the PAML package<sup>121</sup> was utilized for the identification of residues within branches under positive selection. A parameter to estimate positive selection is the ratio of non-synonymous substitutions/non-synonymous sites (dN) per synonymous substitutions/synonymous sites (dS). A dN/dS ratio = 1 indicates neutral selection, > 1 positive selection and << 1 purifying selection. The likelihood ratio test was performed to select the better of the following two hypotheses for each scenario in the branch site model: fixing a fraction of residues in class 2a and 2b to be under neutral selection (dN/dS =  $w = 1$ ) in the foreground branch (H0) vs. free estimation of the dN/dS ratio of this residue fraction in the foreground branch (H1) corresponding to the classical branch site model. Significance was tested using the  $\chi^2$  distribution. To exclude possible biases from codon model choice or shifts in GC content, three different codon models were applied (Codon table, F3X4 and F1X4) and were assessed for consistency between results. Residues under positive selection were identified by Bayes Empirical Bayes (BEB) analysis<sup>122</sup>. Before assaying for positive selection, the respective alignments were tested for the presence of recombination with the RDP4 software<sup>123</sup> (linear

sequence = TRUE, Disentangle overlapping events = TRUE). All recombination tests results were not significant (default values used). To obtain estimates of the robustness of model parameters, we performed 100x bootstrapping with the *codeml\_sba* software for those branch-site tests that returned significant<sup>124, 125</sup>.

A phylogenetic tree was constructed for the concatenation of exons7 and 8 of all *GNAOs* including Cephalochordata and Vertebrata (excluding Teleostei and Agnatha) and evaluated with two different foreground branches: the ancestral branch of *GNAO.1* and *GNAO.2* after the exon duplication, but preceding speciation of Vertebrata, respectively (see Figure 2.9). The respective nucleotide sequences were aligned with MASCE v1.01b<sup>126</sup>. Sequences with missing data in these exons were excluded. The divergence of this alignment is not ideal (tree length 15.7 in H0, F3X4). However, as high divergence would lead to a loss of power rather than an increase in the rate of false positives in the test<sup>127</sup>, the divergence is not considered to be deleterious to the analysis. Positive selection and differences in selection pressure were also tested in the foreground branch of a gene tree composed of *GNAO(a,b).1s* and *GNAOa.2* sequences including exons7 and 8 of Actinopterygii (ray-finned fishes). Foreground branches were defined as the branches after the 3R WGD and before Teleostei speciation (ancestral branches of *GNAOa.1*, *b.1* and *a.2*, respectively, see Figure 2.9).

### ***Computational Modeling of Tertiary Structures***

Available crystal structures of  $G\alpha$  subunits and structural models based on crystal structures were utilized to map exon sequence positions onto tertiary folds. Though all structures and models utilize mammalian sequences, the highly conserved tertiary structure of  $G\alpha$  supports that the relative exon position mappings are maintained across all phyla.

The crystal structures of Gαq bound to PLCβ3 and RGS8 were utilized (PDB ID 4QJ3<sup>128</sup> and 5DO9<sup>129</sup>, respectively). The active monomer of Gαs (PDB ID 1AZT<sup>130</sup>) was used in addition to the crystal structure of Gαi bound to Gβγ (PDB ID 1GP2<sup>131</sup>) and to RGS4 (PDB ID 1AGR<sup>132</sup>). Comparative models of Gαo (human *GNAO.1* transcript variant) and Gαs (human sequence without exon3 or extended exon4) were constructed from previous modeling studies of the ternary complex<sup>133</sup> (activated GPCR bound to Gαi and Gβγ) by replacing Gαi side chain residues with either Gαo or Gαs sequence while maintaining backbone atom coordinates. After threading these sequences, model hybridization continued with optimizing fragment insertions, and relieving chain breaks through the comparative modeling RosettaCM protocol<sup>134</sup>. The relaxed and optimized structural models were then utilized for further exon sequence mapping based on conserved sequence positions. All crystal structures and models were visualized with Pymol<sup>135</sup>.

## 2.3 Results and Discussion

### ***Gα Paralog Evolution before the 2R WGD of Vertebrata*** ***preGNA- genes before the 2R WGD***

The branch of Deuterostomia underwent multiple rounds of whole genome duplication (WGD)<sup>51, 136-138</sup>. These events allowed for increased gene number and sequence diversity. First, we sought to identify which *GNA*- paralogs were present before the two rounds (2R) of WGD that occurred before the diversification of the Vertebrata lineage. Therefore, we primarily limited our study to species of Deuterostomia. We included two Protostomia species (*C. elegans* and *D. melanogaster*), one Cnidaria (*N. vectensis*) and one Placozoa species (*T. adhaerens*) as outgroups. To clarify the orthology relationship we use the following gene names to refer to the progenitor representatives of the *Gα* families before the Vertebrata radiation: *preGNAI*, *preGNAO*, *preGNAQ*, *preGNAS*, *preGNAV*, *preGNA12*.

We identified gene sequences for all five primary families (i, q, v, s and 12) in Cnidaria and four in Placozoa (Only *N. vectensis* maintained *preGNAV*) (Table 2.1). We and others found evidence of *preGNAO*-like sequences in Protostomia and Placozoa but not within Cnidaria<sup>13, 22</sup>. Overall, we found evidence of six *GNA*- paralogs before the diversification of Deuterostomia. Therefore, we conclude that the five known primary families of the *Gα* subunit existed before the split of Bilateria into Deuterostomia and Protostomia though species-specific deletions exist. Moreover, the *Gαi* family was represented by two members, *preGNAI* and *preGNAO*. Lineage-specific tandem duplication events in *T. adhaerens* and *N. vectensis* are discussed in Appendix A.i.

Multiple duplication events occurred in *C. elegans* resulting in over 20 copies of *GNA*-like genes (named *GPA*- in *C. elegans*). We included previously annotated *GPA*s. However, only four genes appear to be similar to the five primary  $G\alpha$  families of Vertebrata; the rest cluster into two separate branches on the ML tree (black subtrees Figure 2.2). *GPA-4* and *GPA-16* are sequentially similar to *preGNAI*, though their exon border positions differ from the conserved eight protein-coding exons found within this family. They nest within the *preGNAI* branch with bootstrap values (BS) of 55. *GOA* and *GPA-12* may be orthologs of *preGNAO* and *preGNA12*, respectively, despite both genes possessing altered exon positions relative to the other non-Deuterostomia genes included; both possess moderate BS values (85 and 86) and form separate monophyletic groups with other *preGNAO* and *preGNA12* members, respectively.

More specifically within Deuterostomia, we investigated two species within Echinodermata, one Hemichordata, two Cephalochordata species, and four Urochordata species. These species diverged before the 2R WGD of Vertebrata, providing a clear starting point before the radiation of this gene family. Within each of these phyla we verified the existence of at least the six established paralogs. Exceptions were found within Urochordata, as we find a lineage-specific loss of *preGNAO* and *preGNAV* at the base of this phylum. A putative gene fragment, found only within *B. schlosseri*, groups with *preGNAV* (BS value 66). Due to limited data, it is unclear if this sequence represents a protein-coding gene or a pseudogene (Table 2.1 and Figure 2.2).

In addition, each phylum interrogated maintained their own number of local gene duplications and/or retrotranspositions for the different primary  $G\alpha$  families (see Appendix A.i). For example, two copies of *preGNAS* were found in all investigated species of

Echinodermata and Hemichordata, suggesting a local duplication of this gene occurred before the split of these branches (Ambulacraria). Both copies of *preGNAS* (*a* and *b*) maintained their exon-intron structure after duplication. In Urochordata, we find evidence of multiple independent retrotranspositions of *preGNAI* into different regions of the genome. These paralogs are encoded by a single exon, characteristic of a retrogene. All examples of these gene duplications are expanded upon in the Appendix A.i. To our knowledge, we are the first to report evidence of these duplications and the existence of these retrogenes. We further validated the presence of these independent gene duplication and retrogene events through interrogation of additional transcriptome and expression data when available (Supplemental Table 2.2).

#### **The (pre)Gai, q, and v Families form a Monophyletic Group within Ga**

We uncovered the evolutionary relationship of the different families by reconstructing phylogenetic trees based on amino acid and nucleotide sequences and by using the conservation of exon-intron structure as an additional evolutionary signal. *preGNAI*, *preGNAQ*, and *preGNAV* share six exon borders and four split codons (codons encoded across two exons) in comparison to the other families suggesting a common origin for these three families (Figure 2.3). Only four major exon borders are shared between these three genes and *preGNAS*.

Focusing on the Gai and Gaq families, it was theorized by Wilkie *et al.* that a progenitor gene to *GNAI* and *GNAQ* (denoted here as *preGNAI/Q*) was tandemly duplicated (*preGNAI/Q*'-*preGNAI/Q*'') and then underwent a larger chromosomal or regional duplication which ultimately led to the *preGNAI*'-*preGNAI*'' and *preGNAQ*'-*preGNAQ*'' gene pair arrangements<sup>46</sup> (Figure 2.4a). Indeed, many others also noted the similar exon-intron organization between paralogs of the Gai and Gaq families; taken

together, this strongly suggests a shared ancestral tandem duplication between these families<sup>20, 47, 54, 55</sup>. Our genomic data of the exon lengths, positions of exon borders, split codons shared across two exons, conserved synteny mapping (gene co-localization) and sequence similarities also support the hypothesis of a tandem duplication event and a regional duplication event of a *preGNAI/Q* progenitor. However, we propose a different timeline for the tandem duplication(s) and the regional duplication relative to each other.

The individual *preGNAI* and *preGNAQ* genes in non-Deuterostomia species investigated show that the exon-intron border positions are conserved in Cnidaria and Placozoa. As seen in Figure 2.3, these two families (and their subfamily members in Vertebrata) maintain their own family-specific exon borders and split codons; *preGNAI* contains eight exons while *preGNAQ* is composed of seven protein-coding exons (excluding variations within Protostomia). In addition, *preGNAI* and *preGNAQ* are not arranged in tandem within the investigated Protostomia and Cnidaria species. If the tandem duplication was ancestral to the regional duplication and differentiation of *preGNAI* and *preGNAQ*, this would require independent intron gain and loss events within exon2/3 of *preGNAI* and exon2 of *preGNAQ* as well as independent lineage-specific losses of one of the gene copies in both *preGNAI'* and *preGNAQ'* gene pairs in the lineages which evolved after the divergence of *preGNAI/Q* into separate genes. Therefore, we conclude that it is highly unlikely that the tandem duplication occurred before the duplication and divergence of *preGNAI* and *preGNAQ* into separate genes.

Instead of *preGNAI/Q* tandemly duplicating into *preGNAI/Q'-preGNAI/Q''*, we propose that *preGNAI/Q* duplicated and diverged into the separate genes of *preGNAI* and *preGNAQ* before the emergence of Metazoa. *preGNAI* and *preGNAQ* then underwent two

independent tandem duplications preceding the 2R WGD events at the emergence of Vertebrata; this gave rise to the *preGNAI'-preGNAI''* and *preGNAQ'-preGNAQ''* paralog pairs that retained their tandem orientation (Figure 2.4). These genes are also referred to as *GNAI0-GNAT0* and *GNAQ/11-GNA14/15*, respectively.

No confirmed tandem duplications of *preGNAQ* were found in the investigated species prior to the 2R WGD of Vertebrata. This suggests that though lineage-specific duplications may have occurred, *preGNAQ* tandemly duplicated into the *preGNAQ'-preGNAQ''* pair at the root of the Vertebrata lineage prior to the 2R WGD events. This progenitor pair then duplicated twice and retained the two gene pairs *GNAQ-GNA14* and *GNA11-GNA15* in Vertebrata.

We identified tandem duplications of *preGNAI* into what could be the progenitor *preGNAI'-preGNAI''* arrangements in Placozoa and Hemichordata. The gene pairs are both arranged in head to head orientations similar to those found in the two of the *GNAI* and *GNAT* gene pairs of Vertebrata. The placozoan *preGNAI* duplications (GIa\_Tadhaerens and GIb\_Tadhaerens) both group within the *preGNAI* subtree with medium BS values (43). Within Hemichordata, one gene copy (GIa\_AcornWorm) groups with the *preGNAI* subtree while the other forms the root of the *GNAT* subtree (GIb\_AcornWorm) (Figure 2.2). Though this grouping suggests that the gene pair may be a *preGNAI0-preGNAT0* set, the low BS value (14) prevents this conclusion. All other identified *preGNAI* duplicates found within Echinodermata and non-vertebrate Chordata are not in a tandem arrangement; however, their small contig sizes prohibit thorough examination of conserved synteny. Overall, this suggests that the tandem duplication of *preGNAI* could have occurred prior to the emergence of Deuterostomia, but our



annotations are not sufficient for further speculation without including more sequences and synteny information.

### **Independent Duplications of *preGNAI* led to the Emergence of *preGNAV* and *preGNAO***

We further expand on the hypothesis set by Wilkie *et al.* by including *Gav* into our analysis<sup>46</sup>. Discovered in 2009<sup>22</sup> *Gav* represents what some suggest is the fifth and final family of the G protein  $\alpha$  subunit in animals<sup>42</sup>. Though *Gav* has been uniquely identified as a separate family by its exon-intron orientation, its gene structure also provides a history linked to the *Gai* and *Gaq* families. In comparison to (*pre*)*GNAI* genes, exon7 is split into exon7 and 8 in the *GNAV* of Vertebrata, and the nucleotide lengths of exons2 and 3 are also altered between the two families (Figure 2.5a). From our analysis, we find that the split exon7 and 8 of (*pre*)*GNAV* does not exist outside of Vertebrata and Cephalochordata. Indeed, in Echinodermata and Hemichordata, we find an exon-intron structure of *preGNAV* closely akin to *preGNAI* and *preGNAQ* (Figure 2.5b). As we and others<sup>13, 22</sup> find no evidence of full-length *GNAV* sequences in the Agnatha (jawless vertebrates), or in any of the four Urochordata lineages investigated, the exact timing of this change in exon-intron structure remains unknown. Nevertheless, it is tempting to speculate that this represents the ancestral exon-intron structure of *GNAV* while the additional intron was gained in the ancestor of Cephalochordata and Vertebrata. Intron gains are an unsurprising addition to gene structures, given the usefulness of introns for elevated transcript accumulation, maturation, and splicing of protein-encoding genes<sup>139-144</sup>.

To determine if *preGNAV* was derived from a duplication of either *preGNAI* or *preGNAQ*, we built ML trees of all (non-)Vertebrata *preGNAI*, *Q*, and *V* sequences. Interrogation of the sequence divergence between these families suggests that *preGNAV* is

more closely related to *preGNAI* (and *preGNAO*) than to *preGNAQ* (Figure 2.5c). We hypothesize that *preGNAV* originated from an ancestral duplication of *preGNAI* before the emergence of Metazoa as we and others<sup>13, 22</sup> have found this paralog across Metazoa lineages (Protostomia, Deuterostomia, and Cnidaria) as well as reports of its existence outside of Metazoa (Choanoflagellata and Filasterea)<sup>13</sup>.

Unlike *preGNAV*, *preGNAO* conserves the positions of the eight exon borders and the shared split codons that are hallmark of the G $\alpha$  family. As noted before, *preGNAO* is present in Placozoa, Protostomia and in Deuterostomia<sup>13</sup>. This suggests that *preGNAI* duplicated before the emergence of Bilateria, to give rise to the *preGNAO* gene (Figure 2.4b). Indeed Krishnan *et al.* found evidence of *preGNAO*-like sequences in several species of Metazoa (*Amphimedon queenslandica* and *Mnemiopsis leidyi*)<sup>13</sup>.

### ***preGNAI2* Originated from a Retrotransposition**

The (*pre*)*GNAI2* gene shares no exon border positions or split codons across exons with any of the other members of the G $\alpha$  family (Figure 2.3). Instead, its exon-intron structure hints that *preGNAI2* originated from a retrotransposition (Figure 2.4b). The ML tree (Figure 2.2) suggests *preGNAI2* may have originated from a retrotransposition of a *preGNAQ* sequence, but more sequences outside of Metazoa are required to interrogate this origin. Evidence of *preGNAI2*-like sequences have been found in several branches of Holozoa<sup>13</sup>. After the retrotransposition, we hypothesize that introns were gained for ease of transcription.

Indeed, across the different branches of Vertebrata and non-vertebrate Deuterostomia we observe flexibility of the exon-intron structure within this family as introns were gained at different positions along the gene in different species (Figure 2.6a-d). The same is true

after the duplication of *preGNAI2* (into *GNAI2* and *GNAI3*) coinciding with the 2R WGD. The *GNAI3* paralog is conserved across Vertebrata, but we see altered exon-intron border positions between species which arose before and after the 3R WGD of Teleostei (Figure 2.6e-g) (the 3R WGD is discussed below). Intron gains have been found to promote gene expression, transcript maturity, accumulation, and processing<sup>139-144</sup>. The flexibility of the exon-intron structure within the  $G\alpha 12$  family across species, its lack of similarity to the other family members' exon-intron structures, and its diversity in function<sup>25</sup>, all suggest the possibility of *preGNAI2* being the product of a reinsertion event and subsequent neofunctionalization before the lineage of Metazoa arose.

#### **Gas is Related to Gai/q**

Excluding retrogenes, all *preGNA*- genes (*preGNAV*, *preGNAI*, *preGNAQ*, and *preGNAS*) shared at least four exon border positions and three split codons (codons encoded across two exons) before the emergence of Metazoa. This suggests that *preGNAI/Q* and *preGNAS* may have arisen as a result of a gene duplication event from a common ancestor (Figure 2.4). *preGNAS*-like sequences have been putatively seen in Choanoflagellata and unicellular Holozoa lineages<sup>13</sup>. Further analysis is required to ascertain the exact evolutionary relationship between the Gas and Gai/q families; however, we see that (*pre*)*GNAV* and (*pre*)*GNAI* form a monophyletic group while (*pre*)*GNAS* clusters outside of this branch on the ML tree (Figure 2.2).

#### **Individual Exon Duplications of *preGNAI/Q* and *preGNAS* in Lancelets**

Prior to the 2R WGD of Vertebrata, many paralogs underwent independent, local, single exon duplication events. Our findings are expanded upon in Appendix A.ii. Briefly, exon6 of *preGNAI* in Cephalochordata is duplicated and available for alternative splicing as well as exon5 of *preGNAQ* in the same lancelet species investigated (Supplemental

Figure 2.2a-b). As this exon corresponds to two homologous sequence regions between these paralogs, we investigated whether this was the result of two independent exon duplications in both paralogs or whether the exon was duplicated in the ancestral *preGNAI/Q* but subsequently lost in all lineages except Cephalochordata. To test this hypothesis, we built ML trees on the nucleotide level composed of the exonic sequences in question (Supplemental Figure 2.2e).

Our findings point to two independent exon duplications within lancelet *preGNAI* and *preGNAQ* respectively. The exon duplications allow these paralogs to alternatively include either of the two different exons sequences within an mRNA transcript. These exons correspond to protein sequences in critical regions of the tertiary fold necessary for protein-protein interactions. Such interfaces are necessary for binding the G $\beta\gamma$  subunits, Regulators of G protein signaling (RGS), Phospholipase C (PLC) and other downstream effector proteins (Supplemental Figure 2.2c-d). Therefore, these two independent exon duplication events may have allowed for the evolution of new functionality by increasing sequence diversity within Cephalochordata.

Additionally, Cephalochordata *preGNAS* can also be alternatively spliced giving rise to two transcripts that differ in the sequence of their final two exons (Supplemental Figure 2.3a). This alternative splicing event introduces changes to the protein sequence when translated which could alter GPCR interaction and subsequent G protein activation (Supplemental Figure 2.3b).

### **Deviations from Previous Publications**

One of the most important results from this study is the updated view of (*pre*)*GNA*-evolution, specifically regarding paralog emergence and loss events across non-vertebrate Deuterostomia species. The genes annotated through the EMS pipeline show a more

thorough and consistent picture of when each paralog emerged relative to one another. The increased genome availability, in addition to our fine-tuned approach, has led us to discover more accurate records of gene gain/loss events and gene duplications. We also sought to confirm our paralog counts with available transcriptome and expression data<sup>103, 104</sup>. Though transcriptome and expression data are not publicly available for all species with sequenced genomes, we validated many of our paralog annotations. This additional information is summarized in Supplemental Table 2.2.

### ***Gα Paralog Evolution after the Vertebrate 2R WGD***

#### **Paralog Gains and Losses**

After a whole genome duplication event, new genetic material will either be maintained (if evolving under purifying or positive selection pressures) or will vanish into the genomic background (if evolving under neutral selection)<sup>145</sup>. Duplicated genes that are maintained may gain new functions or subfunctionalize through mutations in the protein-coding sequence. Temporal and spatial expression patterns may be altered through changes in regulatory regions of the gene. Changes may be maintained to compensate for dosage effects, or serve as a failsafe against the accumulation of deleterious mutations<sup>146-148</sup>. It was estimated that after the 2R WGD of Vertebrata only 20-25% of the duplicated genetic material was retained within genomes<sup>52, 136</sup>. Genes with a low rate of amino acid substitution are more likely to be retained after a WGD<sup>149</sup>, as are genes involved in the nervous system<sup>150</sup> or cellular signaling<sup>151</sup>. The Gα subunit is considered a housekeeping gene due to its pivotal role in transducing and amplifying signaling cascades in all cells. Many paralogs are ubiquitously expressed (Gαs, 12, 13, q, i2) in mammalian tissues, and all but Gα14 and Gα15 are expressed in the brain or neurosensory tissues<sup>25</sup>. Therefore, the duplicated and retained *GNA*- genes (Table 2.1b-c) are expected to evolve under strong

purifying pressure to prevent the gain of deleterious mutations. Many duplicated *Ga* paralogs that were retained after the 2R WGD gained new functions, interaction partners, tissue specificity and/or new cellular signaling properties<sup>25, 49</sup>.

### **The Radiation of *Gai***

The *Gai* family expanded in Vertebrata to include *GNAI1-4*, *GNAT1-4*, and *GNAZ*, in addition to *GNAO*. *GNAT4* and *GNAI4* were quickly deleted. A ML tree built on the nucleotide level further supports the emergence of these paralogs from the 2R WGD in Vertebrata, and shows the pattern of *GNAI0-GNAT0* duplication by resolving *GNAI2* as the outgroup of the *Gai* subfamily and *GNAT1* as outgroup of the *Gat* subfamily when excluding lamprey sequences (Figure 2.7). These outgroups support the hypothesis of the individual *Gai* and *Gat* subfamily members emerging through the tandem duplication of *preGNAI* followed by two consecutive whole genome duplications. The tree constructed in the current study has a different tree topology than those constructed with amino acid sequences by Lagman *et al.*<sup>49</sup> and Krishnan *et al.*<sup>13</sup>. This tree topology is in accordance with the arrangement of *GNAI2* and *GNAT1* as neighbors, which resolves the inconclusiveness of previous studies.

We found no evidence of the proposed *GNAT*-like progenitor gene<sup>50</sup> in the Chordate lineage (*preGNAT0*) prior to Vertebrata divergence; this is in accordance with previous findings<sup>49</sup>. In addition, we identified a putative *preGNAT0* sequence within the Hemichordata lineage (denoted G1b\_AcornWorm), that is positioned in a head to tail arrangement with a *preGNAI* gene (G1a\_AcornWorm). Nevertheless, the low BS support (14) of G1b\_AcornWorm with the split of the Vertebrata *GNAT* subtree prevents the conclusion that this duplication is a 1:1 ortholog to *GNAT0*.

*GNAT3*, which is situated adjacent to *GNAI1* in a head to head orientation, is lost in a

lineage-specific manner in Amphibia and Actinopterygii (ray-finned fishes) as reported previously<sup>47, 54</sup> and confirmed by the current study. The conserved synteny regions around *GNAI1* are maintained, revealing that this loss of *GNAT3* is local and not connected to additional rearrangements. The fourth *GNAI-GNAT* gene pair (*GNAI4-GNAT4*) was predicted to be immediately lost subsequent to the 2R WGD<sup>48</sup>; synteny mapping in humans show a conserved fourth set of genes surrounding the region where the *GNAI4-GNAT4* pair was initially situated after duplication and then presumably deleted<sup>48</sup>.

However, we found nucleotide sequence evidence for all four paralogs of *GNAI* in the Agnatha lineage in both lamprey species investigated, which may correspond to the four copies originating from duplications of the *GNAI0-GNAT0* gene pair. All four *GNAI* genes have the same eight protein-coding exon structure with conserved border positions, and the ML tree shows *GNAI1-4* all clustering close to the root of the Gnathostomata (jawed vertebrate) *GNAI* subtree. Synteny mapping supports the expected head to tail orientation of the *GNAT1-GNAI2* pair and the head to head orientation of *GNAI3-GNAT2*. In addition, *GNAI1* synteny supports the loss of *GNAT3* by maintaining conserved flanking gene orthologs. While a fourth copy of *GNAI* (*GNAI4*) has been briefly described previously in lampreys<sup>54</sup>, the lack of clear synteny information prevents further validation of its origin in the Vertebrata ancestor. Though the conservation of exon border positions, split codons, and nucleotide sequence support the assignment of this paralog to the *Gai* subfamily, evidence of conserved gene neighbors are needed to ascertain if this paralog is the product of an independent duplication or if it is a product of the 2R WGD. There is no evidence of orthologs to the lamprey-specific *GNAI4* in other Vertebrata lineages.

In addition, we found no evidence of *GNAT4* in any Vertebrata lineage. Previous

groups proposed a fourth member of *GNAT* was present in lamprey (denoted *GNAX* or *GNAT4*)<sup>50</sup>. We find that though this gene (denoted here as *GNATI*) is situated close to the root of the phylogenetic tree (Figure 2.7), its sequence, in addition to its synteny with *GNAI2*, suggests it is an ortholog of *GNATI*, not a novel *GNAT* gene or the missing fourth member. With the data considered in the current study, we cannot resolve whether lamprey *GNAI1-3* and *GNATI-3* represent 1:1 orthologs to human *GNAI1-3* and *GNATI-3*, respectively. This reflects the current debate about the exact timing of the 2R WGD relative to the divergence of lampreys and possible lamprey-specific (whole) genome duplications<sup>93, 152</sup>.

### **Gaz**

We show that, contrary to previous theories<sup>47</sup>, *GNAZ* emerged after the 2R WGD through the duplication of a *Gai* family member. We found no substantial evidence of *preGNAZ*-like sequences in non-vertebrate Deuterostomia. The exon-intron structure of *GNAZ* largely deviates from the exon-intron structure of other *Gai* family members (Supplemental Figure 2.4). *GNAZ* is located on the opposite strand within an intron of the *RSPH14* gene. We hypothesize that *GNAZ* emerged through retrotransposition into this position and subsequently gained one intron. This resulted in the conserved two protein-coding exon gene structure. Appendix B.i discusses further analysis done to investigate whether the intron of *GNAZ* carries signatures of insertion mediated by a retrotransposon mechanism; however, no conservation of these residues was found in *GNAZ*.

We identified full-length *GNAZ* genes in all Vertebrate species evaluated (including ghostshark), as well as partial genes (due to small contig size) in both lamprey species - contrary to previous reports<sup>13</sup>. The ML tree composed of all five primary families (Figure 2.2) shows *GNAZ* grouping tightly within the *Gai* family; this further suggests *GNAZ*



originated from a retrotransposition of a Gai family member.

Two non-Vertebrata *GNA*-like sequences (*B. schlosseri* and *T. adhaerens*) are seen on the ML tree to group with the *GNAZ* branch. Both genes in question possess a gene structure that is highly similar to *preGNAI*. Thus, we conclude that these are not 1:1 orthologs of a putative *preGNAZ*.

### **Gao**

Though *preGNAO* emerged before the 2R WGD, we do not find evidence of additional *GNAO* gene copies being retained in Vertebrata after the whole genome duplications (with the exception of Teleostei after the 3R WGD, discussed below). Instead we observe a local duplication that gave rise to two mutually exclusive exons (7.2-8.2 and 7.1-8.1) that are conserved in all major Vertebrata clades (Figure 2.8a).

The resulting two *Gao* isoforms likely show functional differences as the final two exons of *GNAO* map to regions of the tertiary *Gao* protein structure (Figure 2.8b) which have been shown to be necessary for receptor-G protein interaction<sup>153, 154</sup>, receptor selectivity, and subsequent G protein activation<sup>133, 155-157</sup>. *GNAO.1* evolved slightly faster after the duplication in comparison to *GNAO.2* as indicated by a longer ancestral branch (Figure 2.9). This is in accordance with results from the natural selection analysis. This points to signs of positive selection ( $wFG = 613 \pm 428$ ) acting on roughly 10% of the residues on the *GNAO.1* branch after duplication (1-p0-p1, see Supplemental Table 2.3). Given this small percentage of residues, the exact estimate of selection pressure, 'w', in the foreground branch is uncertain. In addition, 88% ( $\pm 9.9\%$ ) of all residues are under strong purifying selection ( $w0 = 0.017 \pm 0.004$ ). Ten residues which were identified to be positively selected differ systematically between *GNAO.1* and *GNAO.2*; the amino acids are conserved in *GNAO.2* in comparison to the non-vertebrate Deuterostomia *preGNAO*

(Supplemental Table 2.3, Supplemental Table 2.4, and Supplemental Figure 2.5).

### **Gaq**

Three of the four known family members (prior to *Gav* discovery) were previously predicted to be situated on large blocks of duplicated genetic material<sup>47</sup>. We systematically validated that *preGNAQ* duplicates (*GNAQ*, *14*, *11* and *15*) were present in all Vertebrata. The head to tail arrangement of the gene pairs *GNAQ-GNA14* and *GNA11-GNA15* is conserved in all investigated species. As described above, we hypothesized that *preGNAQ* underwent tandem duplication at the root of the Vertebrata lineage giving rise to the *preGNAQ'-preGNAQ''* progenitor gene pair prior to the 2R WGD. As seen in the ML trees, *GNAQ* and *GNA11* are very closely related while *GNA14* and *GNA15* though diverged, group together.

*GNA14* and *15* have gained sequence divergence, tissue expression specificity and new functionality, while *GNAQ* and *11* appear to be ubiquitously expressed in mammalian tissues and are involved in a high level of redundant cellular signaling processes<sup>25</sup>. We see two lineage-specific losses of *GNA15* in Coelacanthiformes as well as in Neoaves (supported by loss in all six investigated neoavian species). No evidence of *GNA15* pseudogenes were found proximal to *GNA11* within those species pointing to a complete loss of the gene. Additional EST and TSA data support the loss of *GNA15* in Neoaves and Coelacanthiformes (Supplemental Table 2.2).

### **Gas**

During the 2R WGD, *preGNAS* duplicated to give rise to *GNAS* and *GNAL* (*Gαolf*)<sup>47</sup>; *GNAL* developed tissue-specific expression and functional specificity within the olfactory bulb and various neuronal tissues<sup>25</sup>. We found a species-specific loss of *GNAL* in the genome of the green anole lizard. However, when validating this putative loss with

transcriptome and expression data, we found evidence of *GNAL* expression within lizard TSA and EST data<sup>103, 104</sup> (Supplemental Table 2.2c-d). We thus conclude that *GNAL* must be encoded within the genome of the green anole lizard though it is not represented within the investigated genome assembly. Such issues have been previously reported and may be due to problems during scaffold assembly and coverage during sequencing<sup>158</sup>.

In all investigated Vertebrata genomes, we show that *GNAS* possesses an upstream alternative first exon, extra-long exon (XL-exon) (Figure 2.10a), which is similar in sequence to the 3' sequence of exon1<sup>159</sup>. *GNAL* also possesses a homologous alternative, longer upstream exon, suggesting that this alternative exon sequence existed before the 2R WGD. The XL-exon appears to be absent in non-vertebrate Deuterostomia. Nevertheless, we are careful to speculate about the exact timing of its emergence due to 1) the significant variability in XL-exon's length and its 5' sequence which make homology searches challenging, 2) the highly fragmented quality of the non-vertebrate genome assemblies utilized which hinder even highly refined searches with the EMS pipeline. We were unable to confirm the presence or absence of the XL-exon in *preGNAS* before the 2R WGD. Transcriptome and expression data searches were also uninformative.

In addition to the XL-exon, an extra-extra-long exon (XXL-exon) has been reported upstream of *GNAS* in human and rodent species<sup>160</sup>. Due to its variability in size (approximately ranging from 1400 nt to 2300 nt) and vast sequence divergence, the XXL-exon was not investigated here. Conservation of imprinting<sup>161, 162</sup> and the gene promoter, which is shared with four other upstream genes<sup>163, 164</sup>, were not the subject of this study. For excellent reports on the complex *GNAS* gene structure in Mammalia, please see<sup>160, 165,</sup>

<sup>166</sup>.

As another peculiarity, *GNAS* possesses a cassette exon, exon3, which can be skipped during splicing<sup>167, 168</sup> (Figure 2.10a). The inclusion of exon3 adds 15 AA to the Gas protein (14 AA encoded by this exon plus one AA encoded by a split codon shared with exon4). When mapped onto the tertiary protein structure, the amino acid region encoded by exon3, extends a flexible linker between  $\alpha$ -helix1 of the enzymatic GTPase domain and  $\alpha$ -helixA of the helical domain (Figure 2.10b). This region may be important for G protein activation and nucleotide exchange<sup>157, 169</sup>.

The cassette exon3 of *GNAS* appears to be a very “recent” evolutionary invention as we only find it conserved in Placentalia (placental mammals) but not in other Vertebrata. Interrogation of available transcriptome and expression data confirmed that there is no evidence of exon3 existence outside of this branch (Supplemental Table 2.2). The intron between exon2 and 4 is large (~43,000-72,000 nt) in non-placental Sarcopterygii, while the homologous region becomes much smaller (~6,000-9,000 nt) after emergence of exon3.

We searched for sequences similar to exon3 in other species of Mammalia to elucidate the possible origin of this new exon. We could not find sequence similarity to human proteins from UniProt KB<sup>96</sup> or the NCBI database<sup>103</sup> or to the intronic region between exon2 and exon4 in 14 Sarcopterygii (lobed-finned fishes) when querying with the amino acid and nucleotide sequence of exon3, respectively. Within Placentalia, a highly conserved sequence stretch of roughly 75 nt is situated upstream and 25 nt downstream of exon3, bookending the exon (Supplemental Figure 2.6). Appendix B.ii discusses predicted motifs for DBPs and RBPs which may be present within this sequence stretch.

The emergence of exon3 in Placentalia also co-occurs with the ability of exon4 to be extended by three nt (Figure 2.10, Supplemental Figure 2.7). This extension is mediated by

a well-documented non-canonical SS ‘TG’ situated 3 nt upstream of the canonical SS ‘AG’<sup>168</sup>. The ‘TG’ splice recognition pattern shifts the SS to allow the nucleotides ‘CAG’ to be included within the exon. Due to the position of the splice junctions and the existence of a split codon, a Serine (S) is encoded by the extension (Supplemental Figure 2.7). Therefore, one codon is split across exon2 and exon3 or shared with exon4 resulting in four different isoforms around this exon junction variation: exon2-E-exon3-G-exon4, exon2-E-exon3-GS-exon4, exon2-D-exon4, exon2-DS-exon4.

We found no evidence of an extended exon4 outside of Placentalia in any genome interrogated. Therefore, we conclude that exon3 and the extension of exon4 co-occurred in the ancestor of Placentalia after the split from Marsupialia (marsupials). The expression of all four possible variations of transcripts with the inclusion/exclusion of exon3 and the possible extension of exon4 is supported by transcriptome and expression data.

Pyne *et al.* speculated that the additional amino acid arisen from the exon extension could promote phosphorylation<sup>170</sup>. We did not find any evidence for posttranslational modifications at this or neighboring positions in UniProt KB<sup>96</sup> or the PhosphoSite database<sup>171</sup>. Amino acids encoded by exon3 and the exon4 extension are situated in a flexible linker region between the GTPase domain and the helical domain of the G protein. This region is unresolved in all crystal structures of the Gas subunit (Figure 2.10b).

### **Gα12**

*preGNA12* was duplicated to give rise to *GNA12* and *GNA13* in Vertebrata during the 2R WGD. Both paralogs are present in all Vertebrata genomes investigated with the exception of *GNA12* which appears to be lost within the interrogated species from Amphibia (*X. tropicalis* and *X. laevis*). Available EST data also support a loss of *GNA12* (Supplemental Table 2.2) though *GNA13* is present in both species.

### **Gav**

*GNAV* was the most recently discovered member of the *GNA*- genes<sup>22</sup> due to the widespread loss of this paralog. *GNAV* was lost independently twice within Vertebrata: at the base of Tetrapoda and at the base of Agnatha. Any *preGNAV* gene duplications were not retained after the 2R WGD. Prior to the 2R WGD, *preGNAV* gained an intron dividing exon7 into two (Figure 2.5a). This gene structure is maintained in all species of Vertebrata where the paralog is present (ghostshark, coelacanth, gar and Teleostei).

### **Individual Exon Duplications in *GNAI*, *GNAQ*, and *GNA11***

We found additional duplications of exon4 in *GNAQ* and *GNA11* in some species of Vertebrata. The two different exon4 sequences (4.1 or 4.2, respectively) can be included into the processed transcript via alternative splicing in coelacanth and gar *GNAQ*. *GNA11* can mutually exclude exon4.1 or 4.2 in coelacanth, gar, and Teleostei (except for the second copy of zebrafish and cod which possess only the .1 variant - *GNA11b.1*) (Supplemental Figure 2.10a). Surprisingly, the homologous sequence of *preGNAI*, encoded by exon5, can also be alternatively spliced in Urochordata. As this duplicated exon5 is only present for *preGNAI* outside of Vertebrata, these genes appear to have undergone independent local exon duplication events. This indicates an especially high susceptibility for this region to be retained after local exon duplication. The protein segment encoded by these exon sequences mediates the interaction of G $\alpha$  with the G $\beta\gamma$  subunits as shown from the overlay of these exon positions onto the tertiary protein structures (Supplemental Figure 2.10b). Such interaction is necessary for G protein heterotrimer formation<sup>131, 172</sup>, interaction with the GPCR<sup>172, 173</sup>, and ultimately signal cessation and complex reformation<sup>173</sup>. For further analysis of these exons, please see Appendix B.iii.

### **Non-canonical Splice Sites of *GNA11***

Flanking most exons are highly conserved SS sequence patterns which direct the

binding of the splicing machinery and thus mediate the removal of introns out of the primary RNA transcripts<sup>174</sup>. The canonical SS ‘GT’ is found immediately downstream of the transcribed exon (5’ SS of the intron) while ‘AG’ is found upstream of an exon (3’ intron SS) in 98.93% of all splicing events in Vertebrata<sup>175</sup>. We found conservation of these canonical ‘GT-AG’ splicing patterns for all of the exon sequences annotated with two exceptions. The first is the alternative upstream SS of exon4 in *GNAS* in Placentalia which has been discussed above.

In addition, we found the highly conserved 5’ non-canonical SS ‘GC’ in intron6 of *GNAIL* in most species of Sauropsida and Mammalia (Supplemental Figure 2.11). ‘GC-AG’ represents 0.89% of all splicing events, making it the most common SS in Vertebrata after ‘GT-AG’<sup>175</sup>. This non-canonical SS is present in neither species of Amphibia investigated (*X. tropicalis* and *X. laevis*) nor in alligator. All fishes investigated possess the canonical ‘GT’ 5’ SS for intron6; however, this region is unresolved in the coelacanth genome preventing dating of exact origin of this non-canonical SS.

The emergence of the non-canonical SS in *GNAIL* co-occurs with the conservation of the extended ‘GC’ SS consensus motif: ‘AAG’ (exonic) and ‘GCAAGT’ (intronic) with one substitution in the exonic region indicated in bold<sup>176</sup>. These nucleotides are not conserved in Deuterostomia possessing the canonical SS (Supplemental Figure 2.11). It can be excluded that the non-canonical SS is involved in the skipping of exon6 as no such isoform is supported by EST or TSA data. The conservation of the extended ‘GC-AG’ SS consensus motif thus promotes splicing of exon6, and it is not involved in alternative splicing.

The observed switch from a canonical to a non-canonical SS and its systematic

conservation is surprising. Therefore, we evaluated potential selective pressures acting on nucleotides surrounding the SS, e.g. to maintain binding sites of RBPs or DBPs requiring its strict conservation in Mammalia and Sauropsida. Non-canonical SS may also regulate tissue-specific expression and alternative splicing efficiency<sup>175</sup>; *GNAII* has widely distributed mammalian tissue expression<sup>25</sup> and no functional alternative transcripts were found for this gene.

To evaluate other potential selection pressures present in this region, we compared the nucleotide sequences surrounding the non-canonical SS to species possessing the canonical SS and scanned for local enrichment of DBP and RBP motifs. We uncovered several potential transcription factor binding sites (Supplemental Figure 2.12a) and RBP motifs (Supplemental Figure 2.12b) that overlap with the respective non-canonical SS region. These binding motifs are strictly conserved in all Mammalia and Neoavies genomes yet are not conserved in the four control species with canonical SS. The only exception being the RBP motif for FXR1; this motif seems to be shifted, making binding as equally likely in comparison to the control. The other DBPs all have a reduced binding probability in the control species. Although these motifs show an interesting distribution across the positive and control set, none of the binding motifs are seen more often in the positive than in the control set (Fisher's exact test). Therefore, experimental validation is necessary to infer the roles of these *cis*-regulatory factors in transcription and splicing of *GNAII*.

### **Retrogenes in Primates**

We find that *GNA*- genes have also been subjected to repeated retrotransposition during very recent evolutionary history, specifically during the evolution of primates and specific suborders within (Supplemental Table 2.5). We detected at least 15 *GNA*-retrotransposition events in primates that led to the insertion of a retrotransposed copy on a



non-parental chromosome. The retrotranspositions occurred in members of four out of the five *Gα* families, with most events in the *Gαs* family (six). Additionally, the *GNA11* retrogene *GS1-124K5.9* was tandemly duplicated on two subsequent occasions as indicated by the location of these retrogenes in proximity to their parent retrogene. This duplication was followed by a rearrangement in the ancestor of Cercopithecoidea (old world monkeys), while both copies are located in a head to tail arrangement immediately adjacent within gorilla. Seven of the 15 retrotranspositions are species-specific and limited to the marmoset and tarsier-lineages. Most of these species-specific retrotranspositions still show high sequence similarity to the AA sequence of their parent genes with single ORFs of the retrotransposon aligning to 60-300 AA of the parent. Surprisingly, the gorilla-specific copy of *GS1-124K5.9* conserves more than 80% of the full-length ORF of the parent gene with 99.34% sequence identity to the protein sequence. All other retrotranspositions were detectable in at least eight of the twelve investigated primate genomes; the only exception being independent retrogene3 which was only detected in two Cercopithecoidea species (macaque and baboon).

Most of the primate retrogenes degraded into pseudo-retrogenes conserving several short ORFs that are still similar to the parent genes. *GS1-124K5.9* and *GNAQP1* are interesting examples of pseudo-retrogenes which conserve a homologous region longer than 40 AA with high similarity to the parent protein across all Catarrhini. *GNAQP1* covers an amino acid stretch homologous to AA 1-105 of *GNAQ* within all Hominoidea (apes). Interestingly, promoters are annotated directly upstream of *GNAQP1* and *GS1-124K5.9* on the same strand in human (Ensembl<sup>97</sup>). Pseudo-retrogenes can also be a source for novel peptides. Within *GNAI2P1* and *GS1-124K5.9*, we detected short ORFs with protein-coding

potential conserved in all investigated primates (see Methods). These ORFs do not show any sequence similarity to the parent protein sequences and are thus candidates for short, novel peptides<sup>177</sup>.

Our strategy to identify retrogenes is conservative as we queried the primate genomes with only the annotated *GNA*- retrogenes from human rather than querying with the full length parent genes. Therefore, species-specific retrotranspositions or retrotranspositions that did not conserve the same region of the parent gene as the annotated human retrogene were not captured. The 15 retrogenes captured here should be considered the lower boundary rather than an exact count of retrogenes. Instead the high frequency of retrotranspositions in the evolutionary history of *GNA*- genes is exemplified in the primate lineage.

This observation is in accordance with findings that correlate retrotransposition with the expression level of the parent gene in germ line tissue<sup>178, 179</sup>. Most members of the *Ga* family are housekeeping proteins that are known to have widely distributed or ubiquitous expression patterns throughout the body<sup>25</sup>. In addition, the activity of retrotransposable elements is known to be high in primates<sup>180</sup>. These events in Mammalia are mediated by LINE1 retrotransposons, which can recognize processed and polyadenylated mRNA transcripts. These mRNA are subsequently reverse transcribed and inserted into the DNA through the activity of an endonuclease. Retrotransposons usually must gain new regulatory elements unless they possess a downstream alternative ORF. The existence of an upstream promotor in the *GNAQPI* and GS1-124K5.9 pseudogenes in human, together with a high conservation of the ORFs, makes these candidates for functional retrogenes.

### ***Ga. Paralogs after the 3R WGD in Teleostei***

#### **Paralog Gains and Losses**

In addition to the Vertebrata 2R WGD<sup>51, 136</sup> a third round of whole genome duplication (3R WGD) occurred at the base of Teleostei<sup>138, 181, 182</sup>. It is estimated that over 75% of the genes which arose from the 3R WGD were subsequently lost<sup>181, 182</sup>. The paralog gains and losses obtained from the EMS are summarized in Table 2.1. We confirmed and updated the paralog counts reported by Oka *et al.*<sup>54</sup>. Briefly, we find two copies of *GNAI1*, *GNAI2*, *GNAL*, *GNAI1*, and *GNAI4* in all Teleostei. *GNAV*, *GNAS*, *GNAQ* all have two copies present in Euteleostei, but only one copy remains in zebrafish. *GNAO* and *GNAI3* also have two copies, though there are lineage-specific deletions in pufferfish and Atlantic cod, respectively. Only one copy is maintained after the 3R WGD for *GNAI3*, *GNAZ*, *GNAT1*, and *GNAT2*. *GNAI2* also has one copy retained in Euteleostei, but two copies are present in zebrafish. It appears that zebrafish *GNAI5* underwent several duplications resulting in an arrangement of four *GNAI5* paralogs<sup>54</sup> situated on the same chromosome next to each other with otherwise conserved synteny. At least three of the four copies are expressed as confirmed by EST and TSA data. *GNAT3* is deleted in all Actinopterygii. Of the paralogs that are retained, we find variations in the positions of intron-exon borders (*GNAI2* and *GNAI3*) and variations in alternative splicing patterns (*GNAO*, *GNAI1*, *GNAQ*) as discussed in other sections.

#### ***GNAO Alternative Splicing in Teleostei***

Two copies of *GNAO* were retained after the 3R WGD (except within pufferfish). In zebrafish, Japanese medaka and stickleback both mutually exclusive exons (exon7.2-8.2 and exon7.1-8.1) were retained in one copy (referred to as gene copy “a” - *GNAOa.1* and *GNAOa.2*). The other gene copy (*GNAOb*) lost one pair of exons7-8 immediately following the 3R WGD. In pufferfish, we see a lineage-specific deletion of the complete

*GNAOa* copy (Figure 2.9a).

To determine which copies of the exon sequences were retained in these paralogs (either variant .1 or .2), we created a ML tree of the nucleotide sequences for *GNAO*'s exon7 and exon8 across all phylogenetic branches evaluated. We see that the alternatively spliced exons7 and 8 of *GNAOa* possess both the .1 and the .2 transcript variants while all of the .1 sequence variants are conserved within *GNAOb*. Thus, we resolve that the .2 exon pair of *GNAOb* was lost at the base of Teleostei and that *GNAOa.2* was lost independently in Atlantic cod. In our selection analysis, we did not detect any residues under positive selection in any of the ancestral branches tested (*GNAOb.1*, *GNAOa.1* and *GNAOa.2*). While all residues of exons 7.1 and 8.1 are under strong purifying selection in both a and b copies ( $w = 0.0075$ ), the selection pressure is slightly released with about 6% of residues evolving under neutral selection in the ancestral branch leading to *GNAOa.2*. This might also reflect the released pressure that ultimately led to the loss of *GNAOb.2* in all Teleostei.

## 2.4 Conclusions

### *Conclusions*

The EMS is a powerful and novel method for the assistance of sequence annotation from highly fragmented genome assemblies<sup>102</sup>. Through its use, we have exhaustively searched through dozens of genomes to identify the presence and absence of paralogous genes within the G $\alpha$  family and exon-intron border rearrangements. We computationally annotated genes for all subfamilies of the G $\alpha$  subunit of heterotrimeric G proteins across Deuterostomia phylogeny with the EMS pipeline and collected four species outgroups from Metazoa. These sequences are from a representative 49 species from all major clades (where genome assemblies allow) and give insight into the evolution of the subfamilies beyond that of model organisms.

We found significant deviations from previous literature on the presence and absence of *GNA*- paralogs. Our updated report allows us to refine the theories surrounding G $\alpha$  evolution. Briefly, we propose a G $\alpha$  progenitor gene (*preGNA*-) duplicated into *preGNAS* and *preGNAI/Q*. *preGNAI/Q* duplicated and diverged into *preGNAI* and *preGNAQ*. *preGNAI2* originated as a retrogene which was reinserted into the genome as an intronless sequence and subsequently gained introns. *preGNAV* appears to be a duplication of *preGNAI*. All five primary families are predicted to have differentiated and evolved prior to the emergence of Metazoa. *preGNAI* and *preGNAQ* each underwent independent tandem duplications prior to the 2R WGD. Tandem duplication of *preGNAQ* occurred at the root of the Vertebrata lineage. *preGNAI* may have tandemly duplicated prior to the split of Chordata and non-chordate Deuterostomia, though more support is needed to validate this timeline.

In addition to the major findings within this manuscript, we also uncovered previously unknown variance in gene duplications, the conservation of alternative splicing patterns, exon duplications/insertions, non-canonical SS, conserved DBP and RBP motifs, and traced back the emergence of primate retrogenes. Each of these variants are expanded upon in the appendices. In addition, our curated sequences have been made available for use as the basis of future annotations, sequencing efforts, and as seed inputs for developing biological questions surrounding the Gα family.

### ***Abbreviations***

EMS	- ExonMatchSolver
GPCR	- G protein Coupled Receptor
TCE	- Translated Coding Exon
2R WGD	- 2 <sup>nd</sup> (and 1 <sup>st</sup> ) Round of Whole Genome Duplication in the Vertebrata ancestor
3R WGD	- 3 <sup>rd</sup> Round of Whole Genome Duplication in the Teleostei ancestor
SS	- Splice Site
EST	- Expressed Sequence Tags
TSA	- Transcriptome Shotgun Assembly
BS	- Bootstrap
AA	- Amino Acid
nt	- Nucleotide
ML	- Maximum Likelihood
RGS	- Regulator of G protein Signaling
PLCβ	- Phospholipase Cβ
XL	- Extra-long exon1 ( <i>GNAS</i> and <i>GNAL</i> )
XXL	- Extra-extra-long exon1 ( <i>GNAS</i> )
DBP	- DNA binding protein
RBP	- RNA binding protein
ORF	- Open reading frame

## 2.5 Supplemental Information

### *Appendix A.i – Lineage-specific duplications across Metazoa shed light on gene flexibility and duplication integrity.*

Of note, there are other lineage-specific tandem duplications found for *preGNAI* within Placozoa (*T. adhaerens*). We found evidence of *preGNAI* tandemly duplicating into three copies (copy *a* and *b* are side by side, and the third ‘*c*’ copy lies ~116,000 nucleotides downstream). All three copies of *preGNAI* maintain the same exon-intron structure (eight protein-coding exons with five split codons). As mentioned before, the *preGNAIa* and *b* copies group within the *preGNAI* subtree as an independent branch, while copy ‘*c*’ forms the base of the *GNAZ* tree. Despite the location of *preGNAIc* on the ML tree, it is unlikely that it is a progenitor to the vertebrate *GNAZ* due to its absence in all other non-vertebrate Metazoa lineages investigated. In addition, *GNAZ* genes are situated within the intron of *RSPH14* genes in Vertebrata. Introns of *RSPH14*-like sequences found in all non-vertebrate Deuterostomia branches did not possess traces of this *preGNAIc* gene or any other *preGNAZ*-like gene. Taken together this suggests that *preGNAIc* is the result of an independent, local gene duplication event which occurred within *T. adhaerens*. Therefore, we term the third copy of *preGNAI* in Placozoa as *preGNAIc* and not *preGNAZ*.

The putative fourth copy of *preGNAI* in *T. adhaerens* was identified as *preGNAO* which lies on a different scaffold roughly 750,000 nt upstream of *preGNAS*. *preGNAI2* is also tandemly duplicated into two adjacent genes. These gene copies are arranged in a head to tail orientation on the same scaffold roughly 3,000 nt apart.

*preGNAI* duplications in Cnidaria, *N. vectensis*, are not tandem, but rather are located on separate scaffolds of either 46,000 or 37,000 base pairs in size. The multiple gene copies all appear to be lineage-specific duplications, as the ML tree shows both *a* and *b*

copies forming their own separate branch, independent of the *T. adhaerens* tandemly duplicated genes. As they all maintain the conserved exon-intron structures specific to *GNAI*, these data support our hypothesis of *preGNAI* and *preGNAQ* differentiating into separate genes before the emergence of Holozoa.

Only one possible duplication event of *preGNAQ* was found within the investigated non-Deuterostomia species. A gene fragment in *B. schlosseri* was found (missing 4.5 out of 7 exons) which groups within the *GNAI5* subtree (*Gaq* family). Due to the lack of available synteny, transcriptome and expression data, it is inconclusive whether this gene is a pseudogene or a full length *preGNA*- gene.

Two copies of *preGNAS* genes were identified within *D. melanogaster* (denoted *GalphaS* and *GalphaF*). These genes do not share synteny and are located on chromosomes 2R and 3L, respectively. In addition, *GalphaS* only shares seven exon border positions with the gene structure found for *preGNAS*; *GalphaF* maintains only three.

Two copies of *preGNAS* were found in both species of Echinodermata investigated and in Hemichordata suggesting a local duplication of this gene occurred within non-chordate Deuterostomia which maintained its exon-intron structure. We created a ML tree of all sequences found through the EMS pipeline (Figure 2.2). We find that these *preGNAS* are situated at the root of the *GNAS/GNAL* branch of Vertebrata. Nevertheless, the *preGNAS* do not form a monophyletic group. In addition, none of *preGNAS* duplications appear to share synteny, though lack of large contig size for all paralogs prevents a thorough analysis of gene neighbors.

The duplication of *preGNAI2*, found in *P. miniata* (bat star starfish), also does not appear to be tandem, but rather an independent gene duplication. It does not appear to be a



progenitor to the Vertebrata *GNAI3* as it groups tightly to *preGNAI2* genes found in Echinodermata and other non-vertebrate *preGNAI2* genes.

Two *preGNAV*-like sequence fragments were found within *S. kowalevskii* and *B. belcheri* (Hemichordata and Cephalochordata, respectively). The sequence fragments maintain some conserved exon border positions indicative of *preGNAV*. However, the small contig size and missing data prevented identification of start codons within the sequences. Therefore, it is unclear if these fragments are true protein-coding genes which were not fully assembled or they represent pseudogene remnants of a parent *preGNAV* duplication. Within the ML tree, these two gene fragments are situated between the *(pre)GNAV* and *(pre)GNAS* subtrees.

In Urochordata, we find evidence of multiple independent retrotranspositions which led to the reinsertion of *preGNAI* into different regions of the genome as an intronless gene or as fragments which maintained some exon border positions. In two species (*C. intestinalis* and *C. savignyi*), several different reinsertions were found that group within the *preGNAI* branch. Synteny mapping was unsuccessful in distinguishing gene neighbors around these two paralogs. In *B. schlosseri* other gene duplications and fragments were found; these genes are sequentially distinct from those found in the *C. intestinalis* and *C. savignyi* species. Two nest within the *preGNAI* branch, one appears to be a fragment of a *preGNAV*-like gene, while the fourth gene, though it maintains some *preGNAI/GNAV*-like exon border positions groups betwixt the *GNAV* and *GNAS* subtrees.

#### ***Appendix A.ii – Individual, local exon duplication across phyla and preGNA- paralogs.***

In addition to the local, full-length duplications of *(pre)GNA*- genes in different branches, we also found evidence of smaller duplications involving individual exons within

some of the *(pre)GNA*- genes. These exon duplications gave rise to alternative transcripts with different mutually exclusive exons.

In *preGNAI*, we found evidence of exon6 being duplicated, while in *preGNAQ* exon5 was duplicated in both species of Cephalochordata (*B. floridae* and *B. belcheri*) (Supplemental Figure 2.2a-b). Though exon6 of *preGNAI* and exon5 of *preGNAQ* correspond to homologous sequence regions, we did not find evidence of this exon duplication arising before the emergence of separate *preGNAI* and *preGNAQ* genes (pre-Metazoa divergence). Instead it appears that both alternative splicing events arose independently at the base of the Cephalochordata lineage. All tests for recombination/gene conversion were negative.

If this exon duplication had occurred before the divergence of *preGNAI* and *preGNAQ* (within *preGNAI/Q*), all other non-Cephalochordata Metazoa (non-lancelet) lineages would have each independently lost their ability to alternatively splice this exon sequence in both families. This seems unlikely, as it would require independent losses of both, the duplicated exon6 in *preGNAI* and the duplicated exon5 in *preGNAQ* across all other Metazoa branches.

To test this unlikely scenario, we built ML trees of the non-vertebrate Deuterostomia *preGNAI* and *preGNAQ* nucleotide sequences which corresponded to the mutually exclusive exons, exon6 or exon5, respectively. If the duplication of this exon occurred in *preGNAI/Q*, we expect the mutually exclusive exons (.1 and .2) of *preGNAI* and *preGNAQ* to be more akin to each other across species than to their own family, *(pre)GNAI* or *(pre)GNAQ*, respectively. Instead, we see that each of the mutually exclusive exons is more closely related to the other members of its own family (Supplemental Figure 2.2e). This

suggests that *preGNAI* duplicated its exon5 independently of the *preGNAQ* duplication of exon6, and both occurred within the lancelet lineage.

We then compared the number of per site nucleotide substitutions that arose since the split of *preGNAI* and *preGNAQ* until the speciation of both lancelets to create two sets of sequences: exons 5/6 vs. all other exons. The average nucleotide substitution rate of exons 5/6 is roughly equal to the substitution rate for all other exons (0.6 vs. 0.57). In contrast, the average rate of nucleotide substitution is higher for the branches leading to *preGNAQ* exon5.1 and *preGNAI* exon6.1, respectively, than for the other exons (0.69, 0.65). This suggests an increased substitution rate in the branches leading to *preGNAQ* exon5.1 and *preGNAI* exon6.1 after the exon duplication.

The lancelet lineage appears to have undergone several such local exon duplication events. We found evidence of alternative splicing of the final two exons of *preGNAS* in both species of lancelets investigated (Supplemental Figure 2.3). The second of the exon pairs (12.2 and 13.2) encodes additional three nucleotides resulting in the extension of Gas by one amino acid within the C-terminus. Though the sequences have diverged, they still maintain several highly conserved motifs and high sequence similarity (80-81% at the amino acid level, 88-89% at the nucleotide level).

These exons encode sequences of the  $\alpha 5$  helix, which is important for GPCR interaction and specificity<sup>154, 183</sup>. The ability to alternatively encode two different C-terminal exons for these transcripts may impact the diversity of receptors with which the preGas subunit may interact. In addition, movement of the secondary structural element, the  $\alpha 5$  helix, has been shown to be necessary for subsequent G protein activation after coupling with the receptor<sup>133, 169</sup>. The resulting different protein isoforms may therefore

have different abilities to bind GPCRs, respond to and undergo the necessary conformational changes to activate as their  $\alpha 5$  and  $\alpha 4$  helices and  $\beta 6$  strand differ in sequence. Supplemental Figure 2.3b shows these exon borders mapped to available tertiary structural models of a G $\alpha$ s protein. These two exons (dark green) overlay with regions necessary for receptor interaction and subsequent G protein activation necessary for signal propagation.

***Appendix B.i – The intron of GNAZ does not show traces of a transposon insertion mechanism.***

In order to clarify the origin of GNAZ's intron, we checked whether this intron could have originated from the insertion of a transposon. This mechanism often leaves traces within the exonic sequence in the form of a conserved 'AG' as last nucleotides of the upstream exon and 'GT' as first nucleotides of the downstream exon<sup>184, 185</sup>. The transposon preferentially cuts downstream of the 'AGGT' consensus sequence and then inserts into this genomic position. Two of the nucleotides of the consensus sequence then become part of the intron on either side of the transposon sequence resulting in the following intron sequence: 'GT-transposon-AG'. We evaluated the conservation of these residues in all GNAZ; however, none were found to be conserved. Several alternative mechanisms for gaining introns exist, e.g. intron transposition and intronization. These alternative possibilities were not evaluated due to the sparse species sampling (and thus the extensive evolutionary distances) within this position of the tree. Therefore, the origin of the exon-intron structure of GNAZ remains an open question.

***Appendix B.ii – Conservation of nucleotides flanking exon3 and the 5' end of exon4 in GNAS overlap with DNA-/RNA-binding protein motifs.***

Interestingly, not only the sequence of exon3 of GNAS, but also the surrounding

intronic sequences (3' 75 nt of intron2 and 5' 25 nt of intron3) are conserved in Placentalia (placental mammals) (Supplemental Figure 2.6). A similar pattern of conservation is observed for the 3' 20 nt of intron3 adjacent to exon4 in Placentalia (Supplemental Figure 2.7). In contrast, there are no conserved regions within the 5' end of intron2. The conserved genomic footprints suggested external pressures were constraining the nucleotides surrounding exons3 and 4.

We tested for local enrichment of DBP and RBP motifs at these three SS including the conserved nucleotides of the introns. Near the 3' SS of intron3, five DBP motifs as well as seven RBP motifs are locally enriched within the intronic, conserved nucleotide region or overlapping with the SS in Placentalia in comparison to a uniform background distribution (Supplemental Figure 2.8). The binding sites of the transcription factors Gata3 and 4, which are involved in myogenesis<sup>186</sup> partially overlap with the non-canonical SS. Moreover, the 'TG' non-canonical SS is part of the consensus binding motif of *Rbm24*, a RBP that is known to play a role in myogenic differentiation<sup>187</sup>.

The recognition of a 3' 'TG' SS by the U2 spliceosome is highly unusual (0.016% of U2 SS)<sup>175</sup>, but well documented for *GNAS*. A previous study showed that the usage of the 'TG' SS is promoted by the splicing factor SF2/ASF that has been suggested to bind within exon3 of *GNAS*<sup>188</sup>. SF2/ASF has an antagonistic relationship with another splicing factor, hnRNPA1, which is also suggested to bind in exon3. Our current study confirms this functional connection of exon3 and the 'TG' SS by phylogenetic co-occurrence. Investigation of RBP and DBP sites within exon3 and the surrounding, conserved intronic sequence suggest the conservation of the SF2/ASF binding site (SRSF1) within exon3 in 31 out of 33 species of Placentalia and conservation of the hnRNPA1 binding site in 32

Placentalia (Supplemental Figure 2.9c). The hnRNPA1 binding site is situated in the conserved intronic region upstream of exon3. The 3' SS region of intron2 and the 5' SS region of intron3 harbor roughly 30 DBP motifs and seven RBP motifs that are locally enriched in the reported region in all 33 investigated species of Placentalia (Supplemental Figure 2.9a-b). We additionally tested for motif enrichment in the whole region encompassing exon2, intron2, exon3, intron3 (when available) and exon4 in Placentalia in comparison to non-placental Sarcopterygii. No DBP or RBP motifs were significantly enriched.

The conservation of 100-300 intronic nucleotides surrounding cassette exons has been observed previously and used as a predictor for alternative splicing levels leading to the inclusion or exclusion of the respective exons in several large-scale studies<sup>189, 190</sup>. Nevertheless, Wainberg *et al.* noticed that there is only little overlap of over-represented 6-mers from these conserved, intronic regions with known RBP motifs<sup>190</sup>. A full mechanistic explanation of the observed conservation pattern is the focus of current research.

***Appendix B.iii – Local exon duplications add sequence variety and potential functional divergence for GNAQ, GNA11 and preGNAI.***

We found exon duplications in exon4 of *GNAQ* and *GNA11* which allow the inclusion of either exon4.1 or exon4.2 during alternative splicing (Supplemental Figure 2.10a). This region of the gene, when mapped to tertiary structure, is important for protein-protein interaction of the  $G\alpha$  subunit with the  $G\beta\gamma$  subunits and multiple downstream-signaling effector proteins such as the RGS or PLC proteins (Supplemental Figure 2.10b). The ability to alternatively splice this region, and increase the sequence diversity of the  $G\alpha_q$  and  $G\alpha_{11}$  proteins could alter which  $G\beta\gamma$  subunits bind or which downstream signaling cascades are initiated by these  $G\alpha$  subunits.

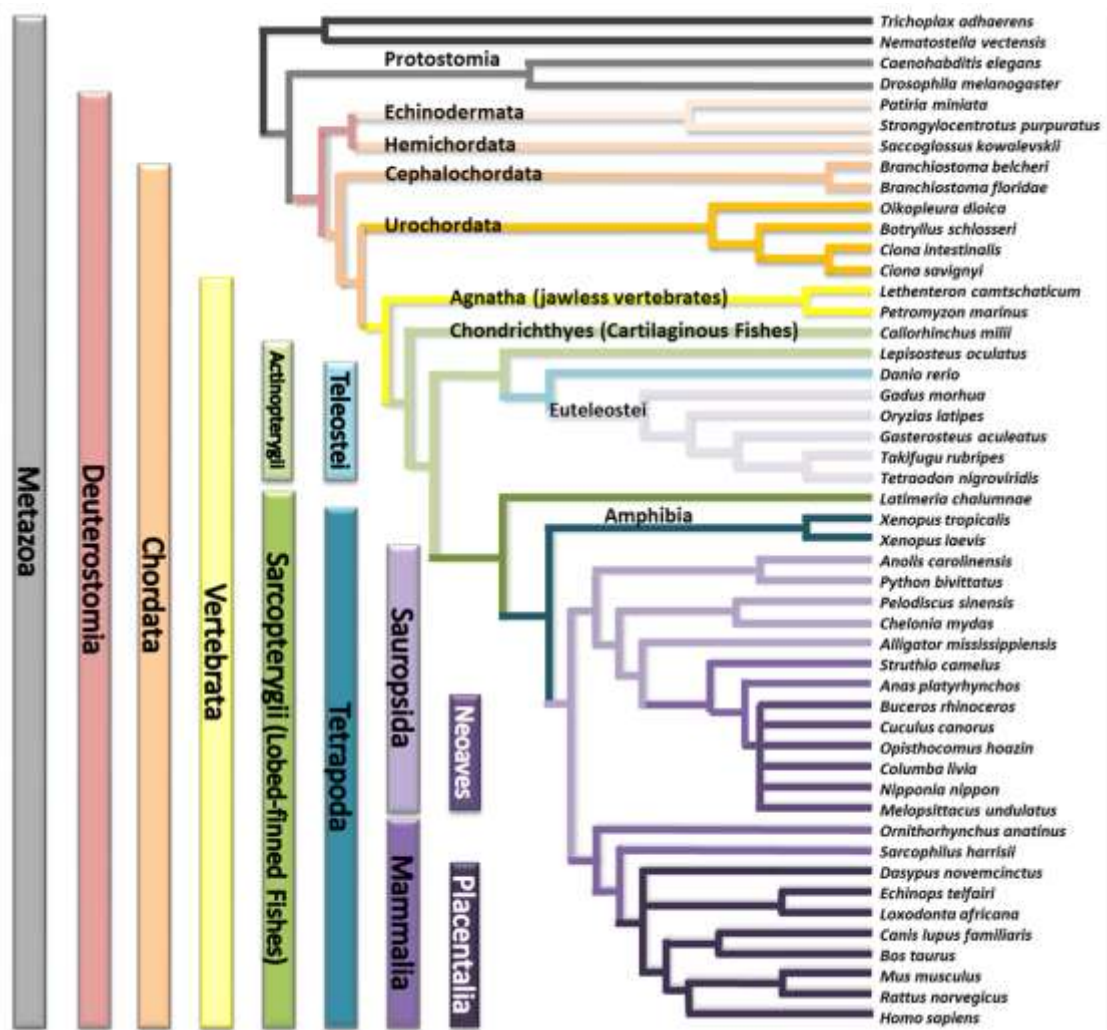
As this exon duplication is present in coelacanth and gar for both paralogs, we propose that this duplication occurred before the 2R WGD in *preGNAQ*, but was subsequently deleted in the other Vertebrata lineages of *GNAQ* and *GNA11* (e.g. within the Agnatha lineage). Upon 3R WGD, at the base of Teleostei, *GNAQ* lost one variant of exon4; therefore, no Teleostei *GNAQ* exon4 duplications exist. However, both *GNA11* exon variants were retained in one gene copy of zebrafish and cod (*GNA11a.1,2* and *GNA11b.1*) and in both gene copies of medaka, stickleback and pufferfish (*GNA11a.1,2* and *GNA11b.1,2*).

Interestingly, we also see the homologous exon duplicated in *preGNAI* of Urochordata. This is replicated across all four species investigated, implying that the duplication took place in the common ancestor of all Urochordata. As this local exon duplication appears only in *preGNAI* before the 2R WGD, and in *GNAQ* and *GNA11* after the 2R WGD, we conclude that these represent two independent, local duplication events and are not ancestral variants of *preGNAI/Q*.

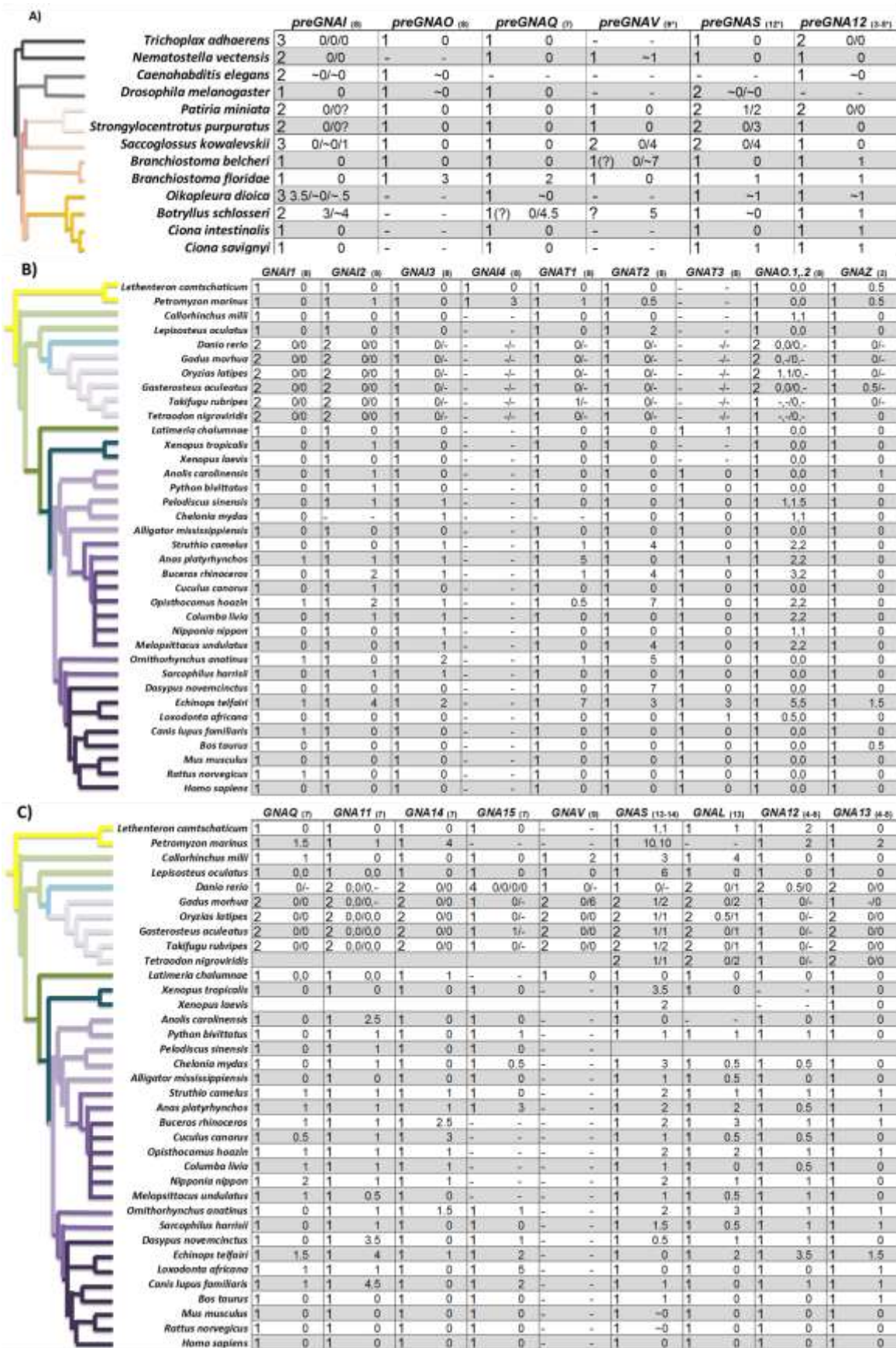
To test this hypothesis, we constructed ML trees of nucleotide sequences from exon5 of (*pre*)*GNAI* and exon4 of (*pre*)*GNAQ* and *GNA11* from Deuterostomia (excluding tetrapod sequences, Supplemental Figure 2.10c). As discussed with the local exon duplications found in the Cephalochordata lineage, it is expected that if the exon duplication occurred before the gene duplication and divergence of *preGNAI/Q* into *preGNAI* and *preGNAQ*, the two exon variants, .1 and .2, would be more similar within their exon variant group than to their subfamily counterparts. Instead, *preGNAI* variants are independent nodes outside of the *Gaq* family and are not nested within any other branch. We find that the ancestral branch of *preGNAQ* orthologs from Urochordata, Cephalochordata, and

Hemichordata bifurcates into two main branches composed of *GNAQ* and *GNA11*; one subtree branches into the .1 variant while the other branches to become the .2 variant cluster. This further supports our hypothesis that the local exon duplication occurred before the 2R WGD which resulted in *GNAQ* and *GNA11*.

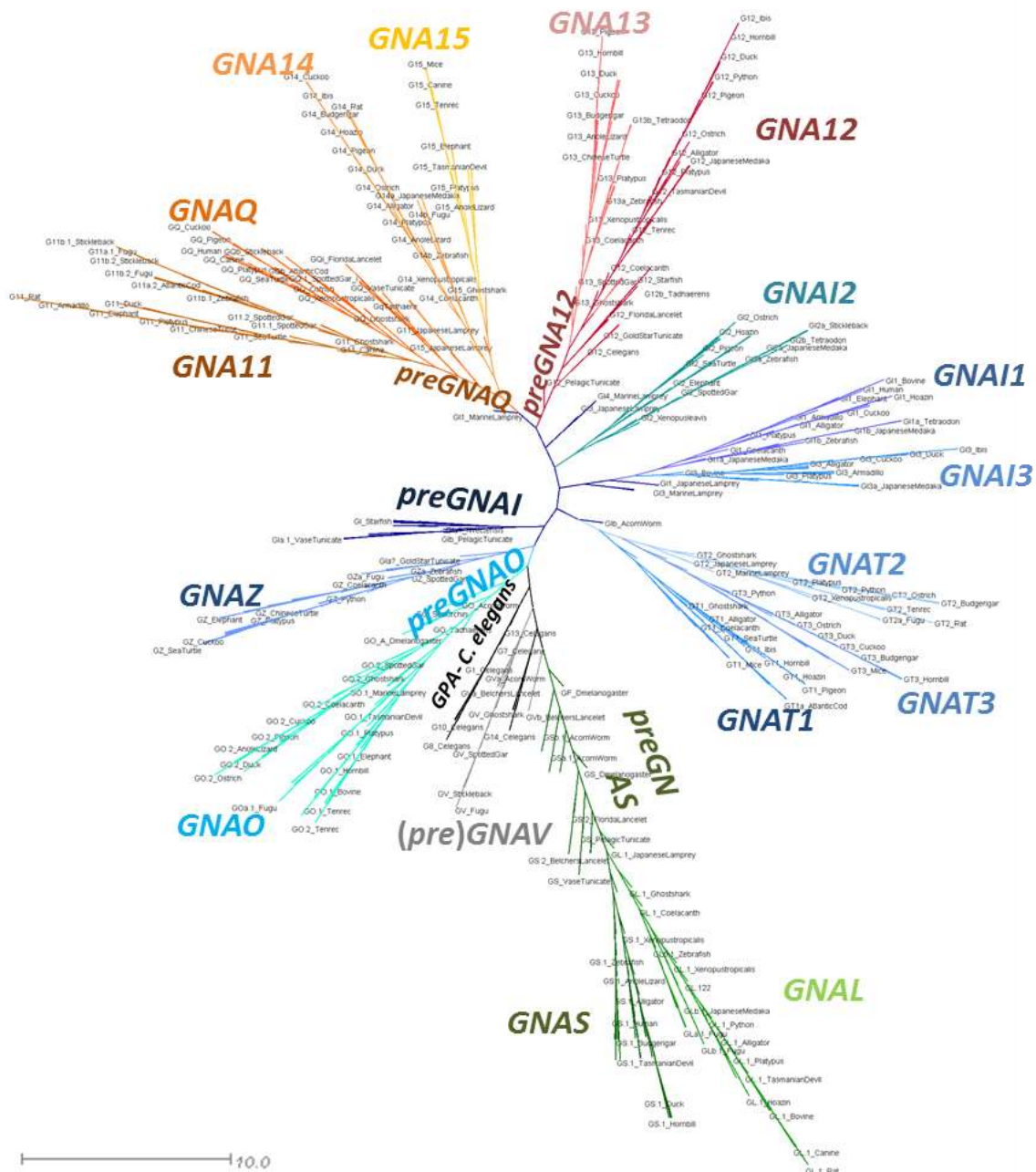




**Figure 2.1: All Deuterostome branches investigated.** 45 species of Deuterostomes were evaluated through the EMS pipeline. The Latin names and clades for each species are provided. Protostomes and Deuterostomes together form the group of Bilateria. Echinodermites and Hemichordates form the group of Ambulacraria. The Protostome and Non-Bilaterian outgroups include *D. melanogaster* (Arthropods), *C. elegans* (Nematodes), *N. vectensis* (Cnidaria), and *T. adhaerens* (Placozoans).



**Table 2.1: pre)GNA- paralog presence before and after the 2R WGD in Vertebrates projected onto a Deuterostome species tree.** A) Sequence evidence of the six preGNA- genes present in non-vertebrate Deuterostomes; two Protostome species, one Cnidarian, and one Placozoan species were included as outgroups (black and grey branches). These genes encode preGai, o, q, v, s, and 12. The first number denotes the number of genes found. Small numbers denote the number of exons missing after curating the annotation as compared to the expected exon counts per phyla (specified at the top of the column). “/” separates multiple paralog gene copies (a, b, c, d). “;” indicate multiple transcripts variants exist which include different exons (.1 or .2), “~” indicate altered and/or erroneous exon borders as compared to other members within the same phylum. “?” indicate unclear paralog assignments due to missing exon data. B) Sequence evidence of individual paralogs after the radiation of vertebrates for the Gai family. Note only one species of pufferfish, turtle, and frog were interrogated if no ambiguity existed. C) Continuation of B for genes encoding Gai, v, s, and 12 families. Note: exonXL was not included in preGNAS exon counts for a total of 12 exons, GNAS includes exonXL for 13 exons, GNAS in placental mammals possess 14 possible exons. GNAL possesses 13 exons for the alternatively spliced long and short exon1, preGNAV possess 8 exons except in Cephalochordates while GNAV is encoded by 9 exons. GNAZ possess 2 exons. \*preGNA12, \*GNA12, and \*GNA13 exon counts vary across phyla, please refer to Figure 6 for details.

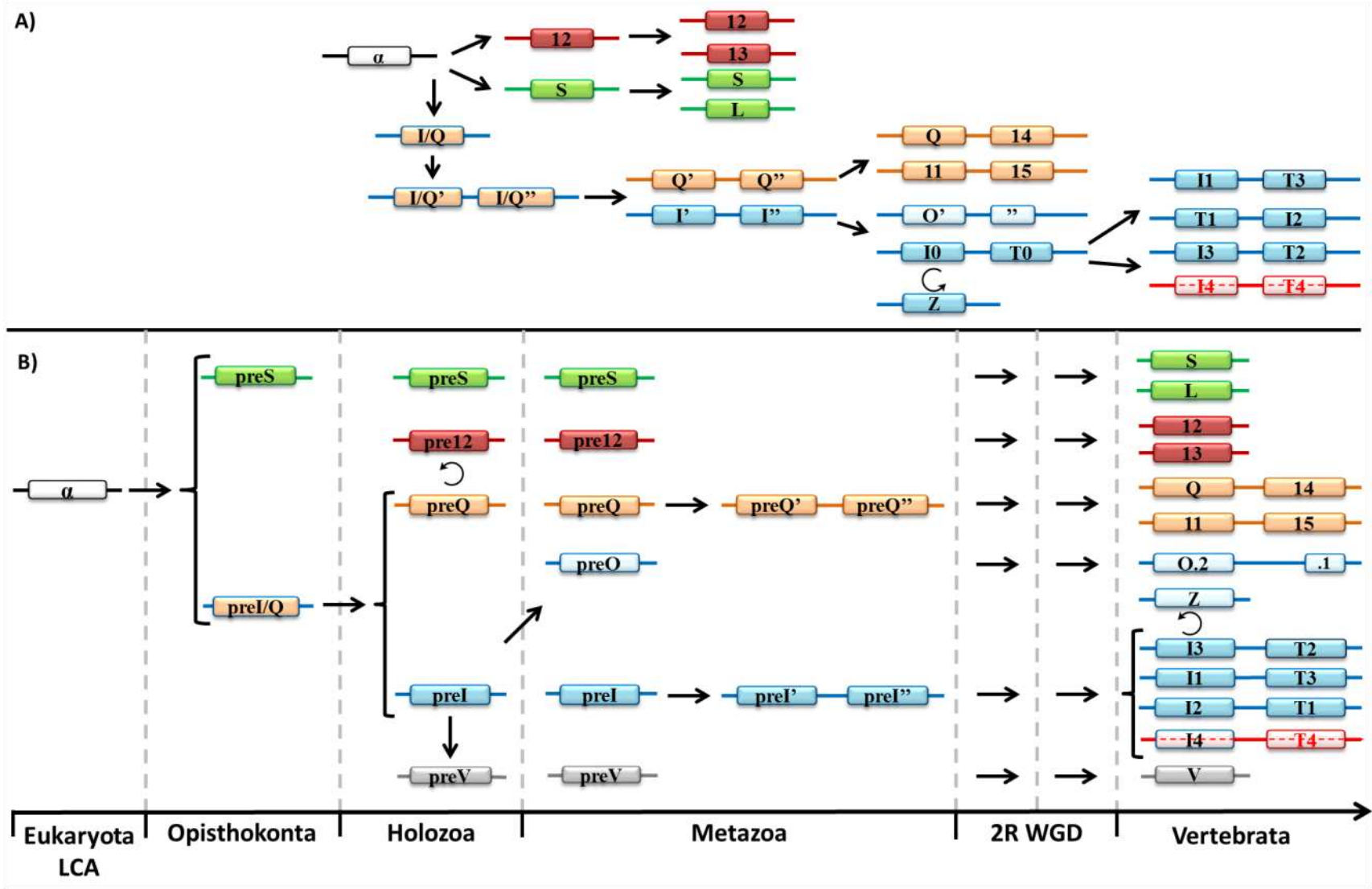


**Figure 2.2: Maximum Likelihood Tree of (*pre*)GNA- genes.** ML tree built with all paralogs and sequences evaluated. The tree is also included as separate file with BS values in NEXML format File1. See Supplemental Table 2.1 for taxonomic group.



**Figure 2.3: Aligning representative vertebrate protein-coding exon borders of all five major families of the G $\alpha$  subunit.** The highly conserved exon border positions give insight into the evolutionary divisions of *GNA*- genes. All protein-coding exons are represented as boxes which correlate with the curated average exon size (introns removed). *GNAI* and *GNAQ* share many exon borders positions (black lines) and four split codons (not shown) suggesting a closer evolutionary relationship. *GNAV* also shares six exon border positions with *GNAI* and *GNAQ*; this suggests that *Gav* family is related to *Gai* and *Gaq* despite its gene presence in a limited number of species. All three genes share four exon borders positions with *GNAS* (not considering the alternatively spliced exon3 or the extended exon4 of *GNAS* found in placental mammals). The lack of shared exon borders between *GNA12* and the other subfamilies suggests that *GNA12* may have originated as an independent retro-gene which independently gained introns.



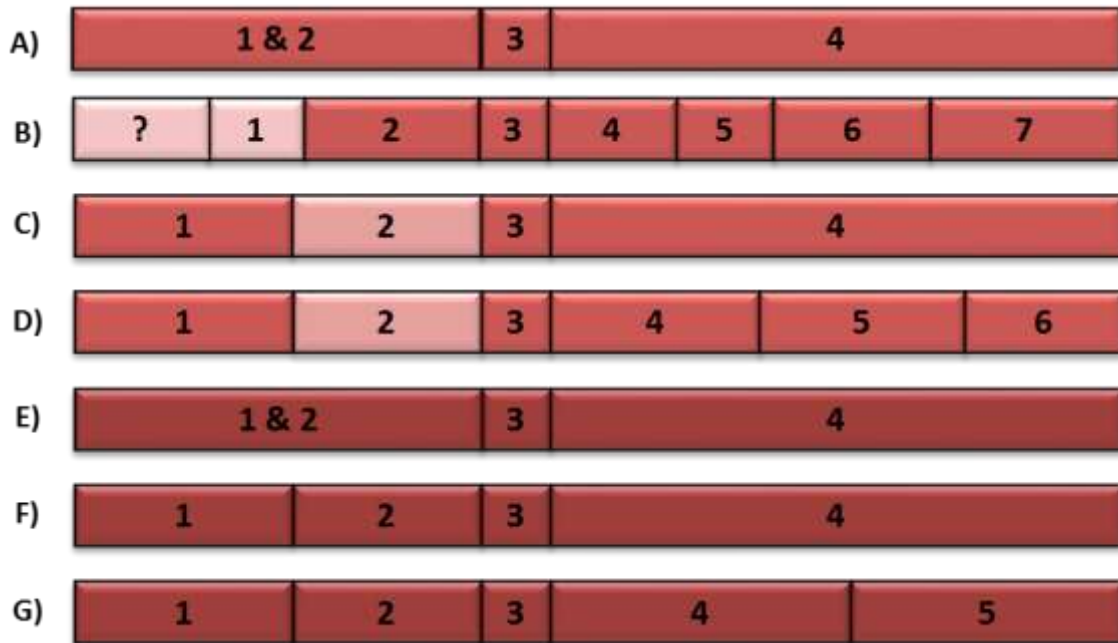


**Figure 2.4: Evolution of the five families of Ga.** **A)** Summary of previous theories of Ga evolution without relative timelines<sup>46-48, 50</sup>. An ancestral *GNA* ( $\alpha$ -white) underwent a series of duplications before diverging into three primary progenitor families. The progenitor *GNAI/Q* tandemly duplicated before undergoing a larger regional or chromosomal duplication. These gene pairs diverged into *GNAI*-like (blue) and *GNAQ*-like (orange) genes. *GNAS* (green), *GNA12* (red), *GNAQ'*-*GNAQ''*, and *GNAI'*-*GNAI''* all duplicated to give rise to two copies from each parent. *GNAI'*-*GNAI''* duplicated into *GNAO'-O''* (ultimately an alternatively spliced gene) and *GNAI0-GNAT0* followed by two more duplications of *GNAI0-GNAT0*. *GNAZ*, a retrogene of *GNAI0*, was reinserted into the genome before the *GNAI0-GNAT0* duplications. **B)** New theory of Ga subfamily evolution incorporating current reports<sup>1, 2, 13, 22, 42, 54, 88</sup> with relative timelines included (not fit to scale). An ancestral *preGNA* progenitor ( $\alpha$ -white) duplicated into the *preGNAI/Q* progenitor and *preGNAS*. *preGNAI/Q* duplicated into two separate genes that diverged into *preGNAI* and *preGNAQ*. Then *preGNAV* arose from a duplication of *preGNAI*. *preGNA12* is a retrogene, possibly of *preGNAQ*, though its precise origin is unclear. *preGNAI* later duplicated to give rise to *preGNAO*. Both *preGNAI* and *preGNAQ* underwent independent tandem duplication events before the 2R WGD of vertebrates. *GNAS*, *GNA12* and *GNAQ'*-*GNAQ''* all retained two copies after the 2R WGD leading to vertebrates, while other hypothetical copies (not shown) were lost immediately after the 2R WGD and are not observed in any extant species. *GNAI'*-*GNAI''* retained three copies of this gene pair after the 2R WGD (*GNAI4* remains only in lampreys). Other, lineage-specific deletions occurred for *GNAV*, *GNAT3*, *GNAI4*, and *GNAT4* as described in the main text. *GNAO* gained alternative splicing ability after 2R WGD (O.2-.1). The retrogene *GNAZ* emerged in the vertebrate lineage from a *GNAI* gene. Lineage-specific duplications and retrogenes are not included for clarity. Straight arrows depict duplications (local, tandem duplications, or WGD), curved arrows depict retrotranspositions. Complete gene names for were simplified for clarity; curated *preGNA*- genes are denoted with “pre-” while “GNA” is removed for clarity in all paralogs. LCA = Last Common Ancestor





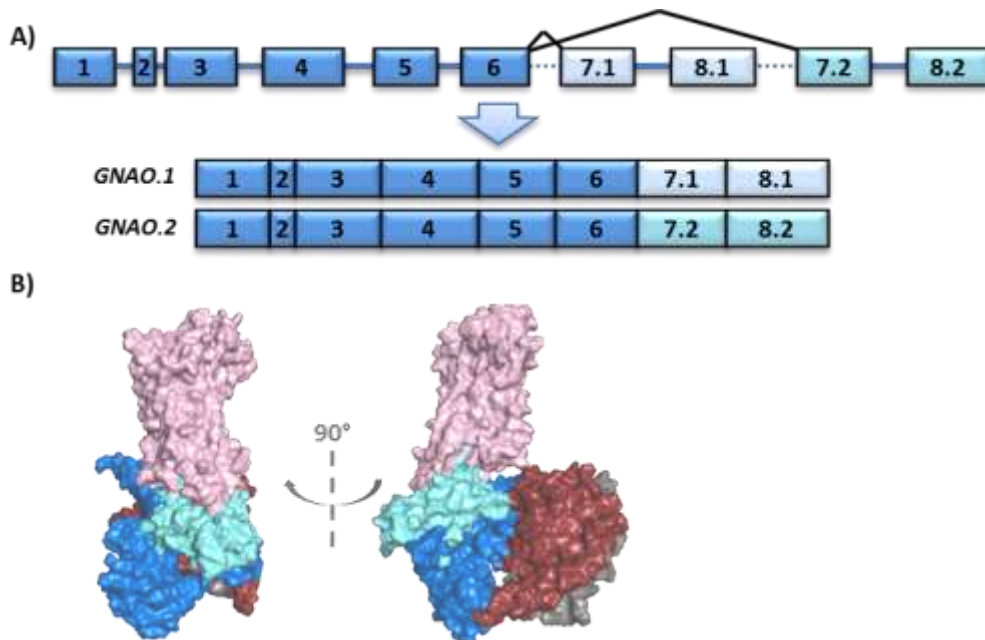
**Figure 2.5: Evolution of *Gav*.** A) A schematic representation of the exon-intron structure of jawed-vertebrate and Cephalochordate (*pre*)*GNAV* genes with 9 protein-coding exons (grey boxes). Box sizes roughly correlate with exon size, while line lengths do not correlate to intron size. B) The exon-intron structure of Ambulacraria *GNAV* genes (Hemichordates and Echinodermates). This *preGNAV* has no intron to divide exon7 and 8, making its exon-intron structure closely akin to (*pre*)*GNAI* (blue boxes) and (*pre*)*GNAQ* (orange boxes) exon-intron structures. This may represent an ancestral exon-intron structure of *preGNAV*. C) The ML tree of all Deuterostome (*pre*)*GNAI*, (*pre*)*GNAQ*, and (*pre*)*GNAV* genes resolves (*pre*)*GNAV* (grey) nesting within (*pre*)*GNAI* (blue) while (*pre*)*GNAQ* (orange) clusters into a distal branch.



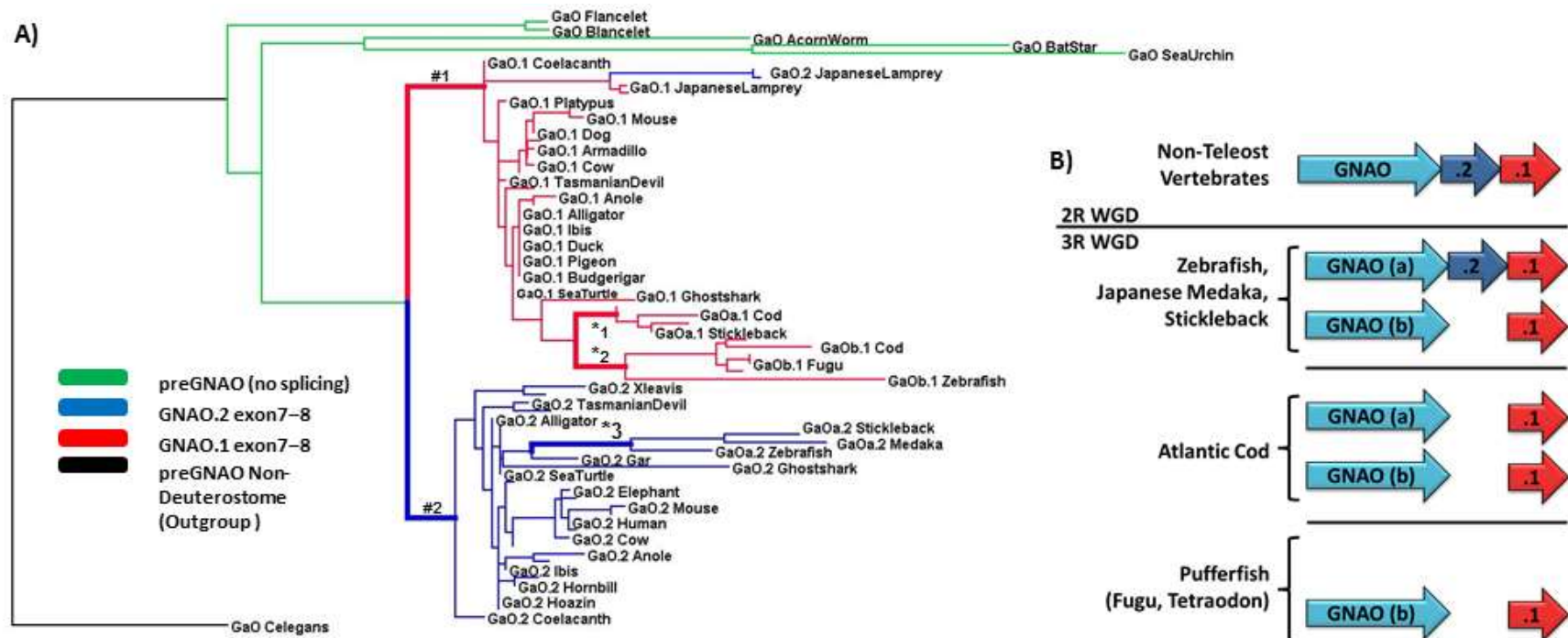
**Figure 2.6: Flexibility of exon-intron borders within the (*pre*)*GNA12* and *GNA13* genes.** The positions of (*pre*)*GNA12* and *GNA13* exon borders (represented boxes) change across phylogeny. Box lengths correlate with average curated exon lengths (introns removed). A) *preGNA12* (red) has three protein-coding exons in Placozoans, Cnidarians, Echinodermates, Hemichordates, and Cephalochordates. B) In Urochordates, the first exon of *preGNA12* is divided into at least two exons while the final exon is divided into four exons. As the 5' sequence is unresolved, more exons may be present (pink with ?). C) *GNA12* exon-intron structure in jawed vertebrates (excluding euteleosts). The exon sequences upstream of exon3 are not resolved in either jawless vertebrate (lamprey) species investigated. The 5' end of exon2 is extended by nine nt (pink) in all jawed vertebrates including euteleosts. D) *GNA12* exon-intron structure in euteleosts (after 3R WGD but not in zebrafish) E) *GNA13* (dark red) exon-intron structure in jawless vertebrates and cartilaginous fish. *GNA13* arose after the 2R WGD that occurred before the emergence of vertebrates. Note that the exon border positions are identical to the *GNA12* from (A). F) *GNA13* exon-intron structure in lobe-finned fishes. The exon positions are identical to *GNA12* in jawed vertebrates (except euteleosts) (C). The *GNA13* sequence is extended by one split codon between exon1 and 2 and six nucleotides within exon2 (not shown). G) *GNA13* exon border positions of euteleosts. The split codon and extended exon2 sequences are maintained.



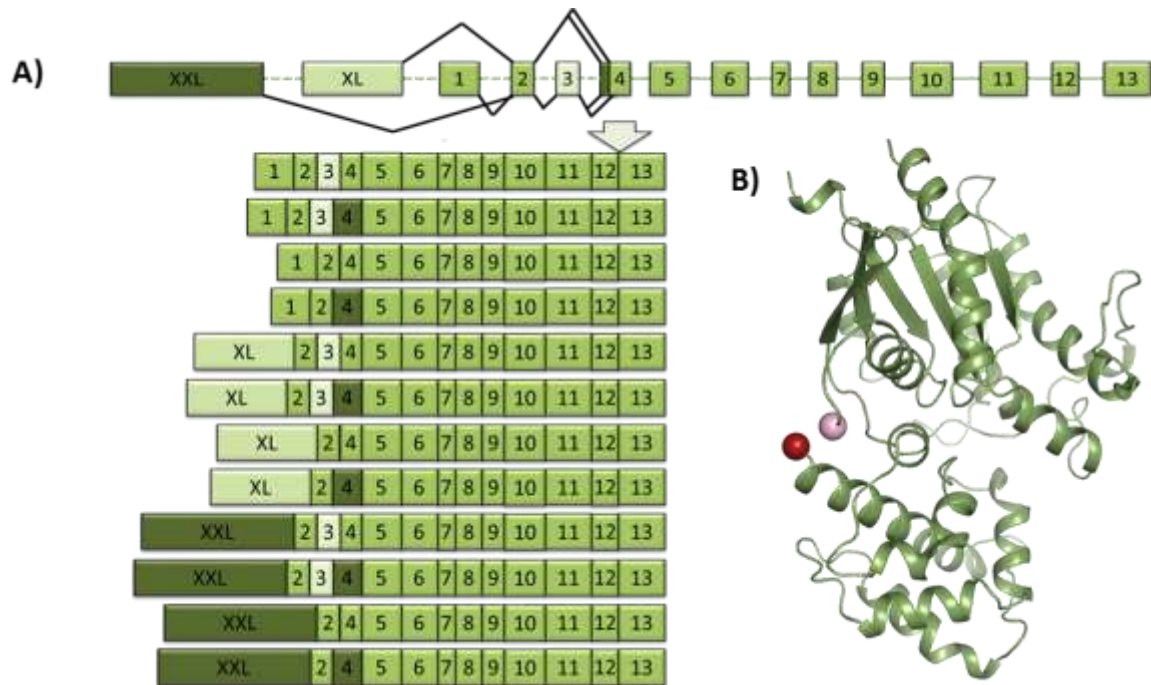
**Figure 2.7: ML tree of the Gai family resolves gene relationships.** A) ML tree built with all protein-coding sequences found within the (pre)Gai family (*preGNAI*, *GNAI1-4*, *GNAT1-3*, (*pre*)*GNAO*, and *GNAZ*) in all Deuterostome lineages evaluated. All lamprey branches are denoted in pink. The outgroups of the *GNAI* and *GNAT* genes have high bootstrap supports (*GNAT1* -100 and *GNAI2* - 92). The two *GNAT* lamprey genes form a monophyletic group with *GNAT1* and *GNAT2* of jawed vertebrates, as do *GNAO* and *GNAZ* lamprey genes with their respective subtrees. B) *GNAI1* and *GNAI3* are situated next to *GNAT2* and *GNAT3*, respectively, within the genome. These pairs are the result of one duplication event. *GNAI2* and *GNAI4* are situated next to *GNAT1* and *GNAT4*, respectively; they also arose from one duplication event. *GNAI4* and *GNAT4* (red) are not observed in the genomic data investigated (except for *GNAI4* in lamprey).



**Figure 2.8: Alternative Splicing of *GNAO*.** A) The vertebrate *GNAO* gene has two transcripts (.1 light blue and .2 cyan) that arise from mutually exclusive splicing of its final exon pair: exon7 and 8. Note that exon lengths correlated with box lengths while lines do not correlate with intron size. B) Tertiary structural model of the heterotrimeric G protein. G $\alpha$  (blue) and the heterotrimer G $\beta$  subunits (crimson/grey) coupled to a GPCR (pink). The two mutually exclusive exons encode regions necessary for coupling to active GPCRs and subsequently activating the G protein itself. The differences in sequences may influence coupling affinity and activation efficiency.



**Figure 2.9: Retained exons of *GNAO* after 3R WGD in Teleosts.** A) A ML tree of exon7 and 8 nucleotide sequence indicates which exon pairs were retained across different teleosts. Note that though this tree is displayed as being rooted by the *C. elegans* *GNAO* sequence, it is an unrooted tree by definition and only shown to be “rooted” for clarity. Branches tested for positive selection are marked by ‘#’ and ‘\*’. B) After the 3R WGD, only one gene copy of *GNAO* (named copy ‘a’) maintained two sets of the mutually exclusive exon7-8 endings (variant ‘.2’ – blue, variant ‘.1’ – red). In Atlantic cod, both gene copies possess only one set of the final exons which was identified as the ‘.1’ variant. In both species of pufferfish, only the ‘b’ copy of *GNAO* was retained with the ‘.1’ exon variant.



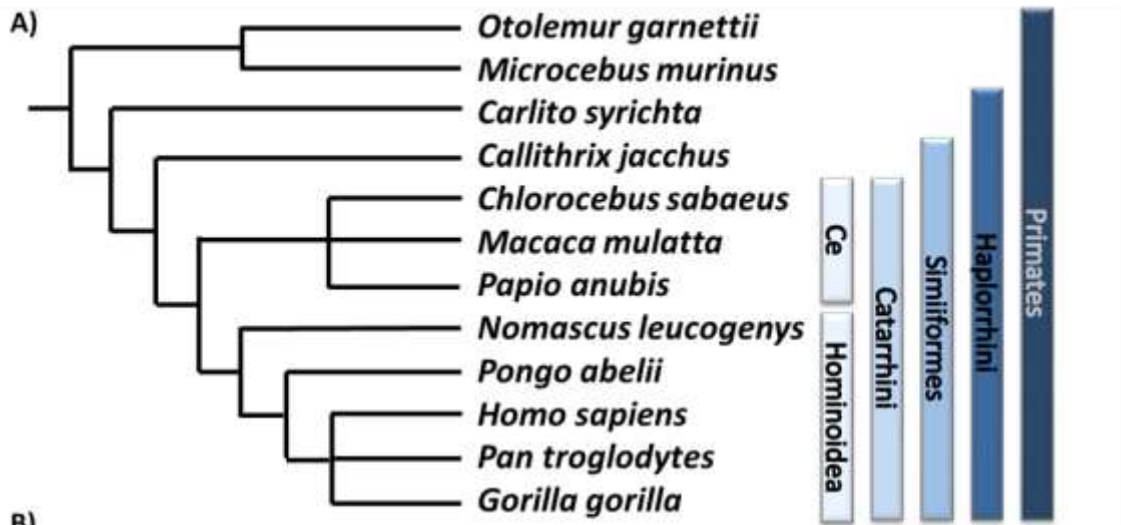
**Figure 2.10: Multiple transcripts are possible from the complex locus of *GNAS*.** A) Different mRNA transcripts can be produced from the *GNAS* locus through alternative splicing. The XXL exon, though not examined herein, can be alternatively included into the transcript in exchange for exon1 or the XLexon. In addition, placental mammals possess a cassette exon3 (light green) which can be included or excluded within the transcript; a non-canonical SS can also give rise to an extended exon4 (dark green) in the same species. Box lengths correlate with average curated exon lengths (introns lines do not). B) Crystallographic tertiary structure of mammalian Gas (PDB ID 1AZT<sup>130</sup>) missing exon3. The C-terminus of exon2 (pink sphere) and N-terminus of exon4 (red sphere) are shown.



Major Phylogenetic Branch	Species	Genome Assembly	Assembly Accession ID
Mammalia – Euarchontoglires (Primates)	Human ( <i>Homo sapiens</i> )	GRCh38.p7	GCA_000001405.22
Mammalia – Euarchontoglires (Glires)	Mouse ( <i>Mus musculus</i> )	GRCh38.p4	GCA_000001635.6
Mammalia – Euarchontoglires (Glires)	Rat ( <i>Rattus norvegicus</i> )	Rnor_6.0	GCA_000001895.4
Mammalia – Laurasiatheria	Bovine ( <i>Bos taurus</i> )	UMD3.1	GCA_000003055.3
Mammalia – Laurasiatheria	Canine ( <i>Canis lupus familiaris</i> )	CanFam3.1	GCA_000002285.2
Mammalia – Afrotheria	Tenrec ( <i>Echinops telfairi</i> )	TENREC	GCA_000313985.1
Mammalia – Afrotheria	Elephant ( <i>Loxodonta africana</i> )	Lorafr3.0	GCA_000001905.1
Mammalia – Xenarthra	Armadillo ( <i>Dasypus novemcinctus</i> )	Dasnov3.0	GCA_000208855.2
Mammalia – Metatheria	Tasmanian devil ( <i>Sarcophilus harrisii</i> )	Devil_refv7.0	GCA_000189315.1
Mammalia – Monotremata	Platypus ( <i>Ornithorhynchus anatinus</i> )	OANA5	GCF_000002275.2
Aves – Palaeognathae	Ostrich ( <i>Struthio camelus australis</i> )	ASM69896v1	GCA_000698965.1
Aves – Neognathae (Galloanserae)	Duck ( <i>Anas platyrhynchos</i> )	BGI_duck_1.0	GCA_000355885.1
Aves – Neognathae (Pallaciiformes)	Budgerigar ( <i>Melopsittacus undulatus</i> )	Melopsittacus_undulatus_6.3	GCA_000238935.1
Aves – Neognathae (Bucerotiformes)	Hornbill ( <i>Buceros rhinoceros</i> )	ASM71030v1	GCA_000710305.1
Aves – Neognathae (Pelicaniformes)	Ibis ( <i>Nipponia nippon</i> )	ASM70822v1	GCF_000708225.1
Aves – Neognathae (Opisthocomiformes)	Haazin ( <i>Opisthocomus hoazin</i> )	ASM89207v1	GCA_000692075.1
Aves – Neognathae (Cuculiformes)	Cuckoo ( <i>Cuculus canorus</i> )	ASM70932v1	GCA_000709325.1
Aves – Neognathae (Columbiformes)	Pigeon ( <i>Columba livia</i> )	Cliv_1.0	GCA_000337935.1
Nonavian Sauropsida – Lepidosauria (Iguanidae)	Anole Lizard ( <i>Anolis carolinensis</i> )	AnoCar2.0	GCA_000090745.1
Nonavian Sauropsida – Archosauria (Crocodylia)	Alligator ( <i>Alligator mississippiensis</i> )	ASM28112v4	GCA_000281125.4
Nonavian Sauropsida – Lepidosauria (Serpentes)	Burmese Python ( <i>Python bivittatus</i> )	Python_molurus_bivittatus-5.0.2	GCA_000186305.2
Nonavian Sauropsida – Archelosauria (Trionychia)	Chinese Softshell Turtle ( <i>Pelodiscus sinensis</i> )	PeISin_1.0	GCA_000230535.1
Nonavian Sauropsida – Archelosauria (Durocryptodira)	Green Sea Turtle ( <i>Chelonia mydas</i> )	ChelMyd_1.0	GCA_000344595.1
Amphibia – Anura	Tropical Clawed Frog ( <i>Xenopus tropicalis</i> )	XJGI4.2	GCA_000004195.1
Amphibia – Anura	African Clawed Frog ( <i>Xenopus laevis</i> )	XLv80	GCA_001663975.1
Basal Sarcopterygii – Coelacanthiformes	Coelacanth ( <i>Latimeria chalumnae</i> )	LatCha1	GCA_000225785.1
Basal Actinopterygii	Spotted Gar ( <i>Lepisosteus oculatus</i> )	LepOcu1	GCA_000242895.1
Teleostei – Ostariophysi	Zebrafish ( <i>Danio rerio</i> )	GRCz10	GCA_000002035.3
Teleostei – Gadiformes	Atlantic Cod ( <i>Gadus morhua</i> )	gadMor1	GCA_000231765.1
Teleostei – Belontiiformes	Japanese Medaka ( <i>Oryzias latipes</i> )	HdrR	GCA_000313675.1
Teleostei – Perciformes	Stickleback ( <i>Gasterosteus aculeatus</i> )	BROAD S1	GCA_000180675.1
Teleostei – Tetraodontiformes	Fugu ( <i>Takifugu rubripes</i> )	FUGU4.0 & 5.0	GCA_000180615.2
Teleostei – Tetraodontiformes	Tetraodon ( <i>Tetraodon nigroviridis</i> )	TETRAODONB.0	GCA_000180735.1
Chondrichthyes	Ghostshark ( <i>Callorhynchus milii</i> )	Callorhynchus_milii_3.1.3	GCA_000165045.2
Cyclostomata	Marine lamprey ( <i>Petromyzon marinus</i> )	Pmarinus_7.0	GCA_000148955.1
Cyclostomata	Japanese lamprey ( <i>Lethenteron japonicum</i> )	LeJap1.0	GCA_000466285.1
Non-Vertebrate Chordata – Cephalochordata	Florida Lancelet ( <i>Branchiostoma floridae</i> )	Version 2	GCA_000003815.1
Non-Vertebrate Chordata – Cephalochordata	Belcher's Lancelet ( <i>Branchiostoma belcheri</i> )	Diploidv18	GCA_001625405.1
Urochordata – Tunicata (Ascidiacea)	Vase Tunicata ( <i>Ciona intestinalis</i> )	KH	GCA_000224145.1
Urochordata – Tunicata (Ascidiacea)	Sea Squirt ( <i>Ciona savignyi</i> )	CSAV 2.0	GCA_000149265.1
Urochordata – Tunicata (Appendicularia)	Pelagic Tunicata ( <i>Oikopleura dioica</i> )	ASM20953v1	GCA_000209535.1
Urochordata – Tunicata (Ascidiacea)	Gold Star Tunicata ( <i>Botryllus schlosseri</i> )	355a-chromosome-assembly	GCA_000444245.1
Hemichordata	Acorn worm ( <i>Saccoglossus kowalevskii</i> )	Skow_1.1	GCA_000003605.1
Echinodermata	Sea Urchin ( <i>Strongylocentrotus purpuratus</i> )	Spur_4.2	GCA_000002235.3
Echinodermata	Starfish ( <i>Patiria miniata</i> )	Pmin_1.0	GCA_000285935.1
Protostomia – Arthropoda	Fruit Fly ( <i>Drosophila melanogaster</i> )	BDGP6	GCA_000001215.4
Protostomia – Nematoda	Roundworm ( <i>Caenorhabditis elegans</i> )	WBcel235	GCA_000002985.3
Non-Bilateria – Cnidaria	Sea Anemone ( <i>Nematostella vectensis</i> )	ASM20922v1	GCA_000209225.1
Non-Bilateria – Placozoa	Trichoplax adhaerens	ASM15027v1	GCA_000150275.1

**Supplemental Table 2.1: Species Evaluated.** All major branches of Deuterostomes were investigated using the EMS pipeline (where sequenced genomes exist). Four species were also included from Metazoan lineages (Protostomes and non-Bilateria) to act as outgroups. Column1 – Description of phylogenetic branch. Column2 – Common name (*Genus species*). Column3 – Genome assembly used. Column4 – Accession number for genome assembly, when available.

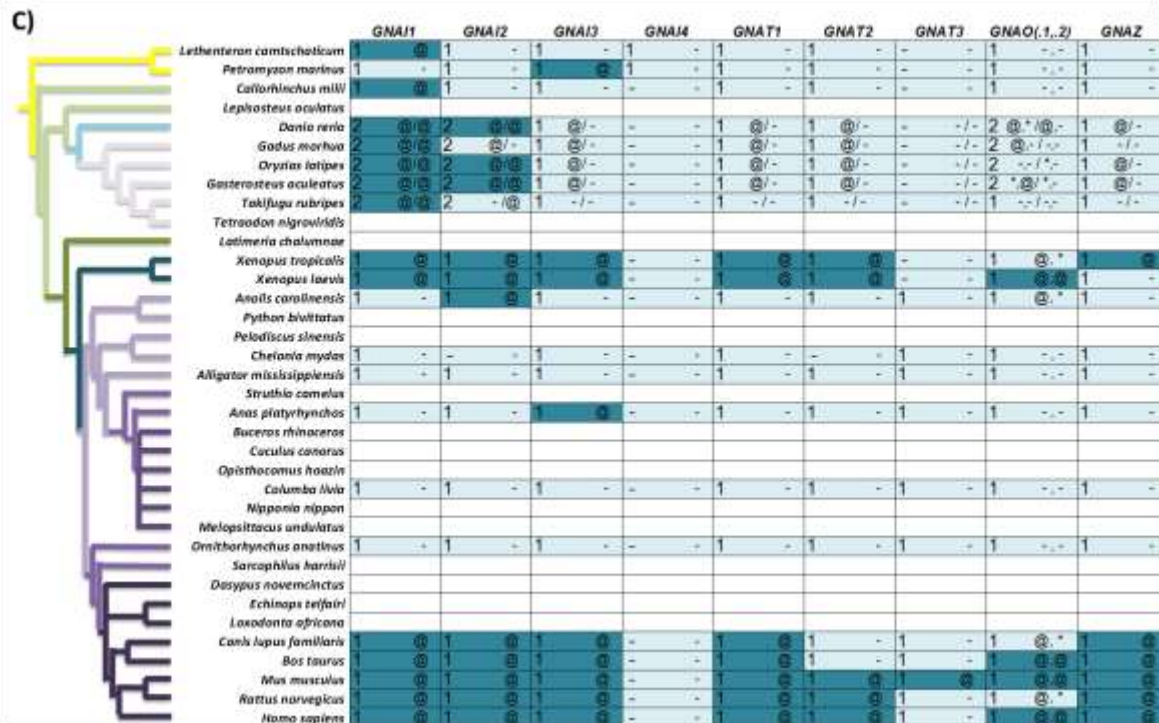




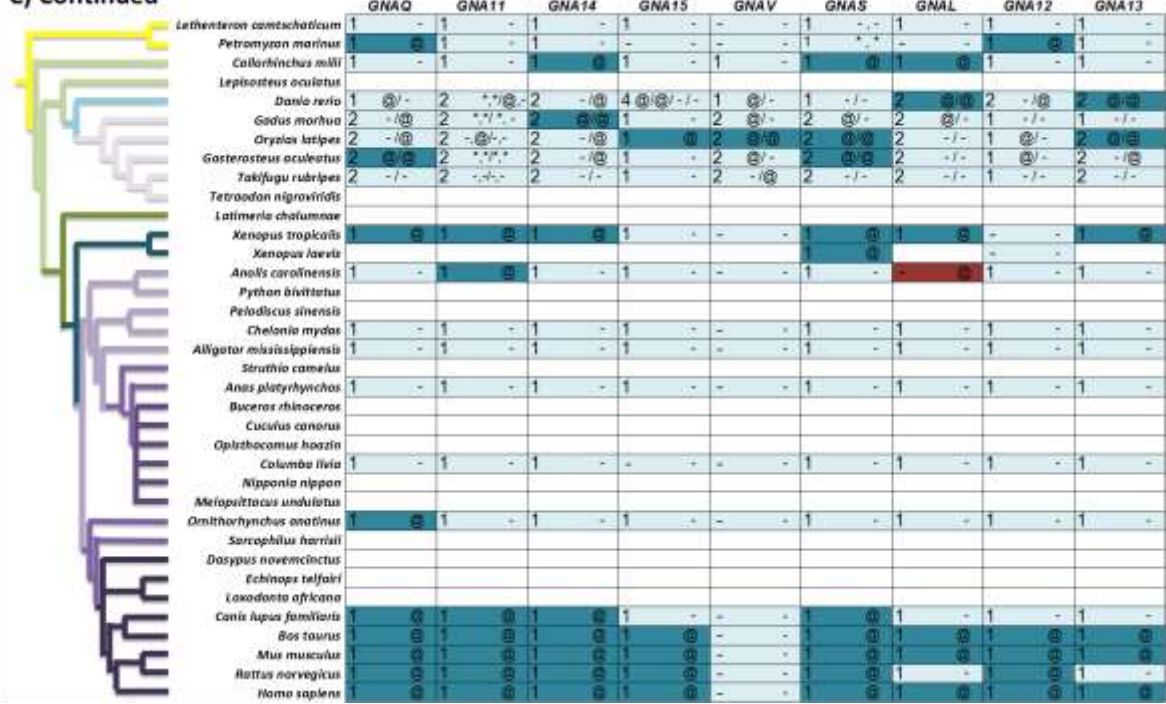
B)

Species	Genome Assembly	Assembly Accession ID
Chimp ( <i>Pan troglodytes</i> )	CHIMP2.1.4	GCA_000001515.4
Gorilla ( <i>Gorilla gorilla gorilla</i> )	gorGor3.1	GCA_000151905.1
Orangutan ( <i>Pongo abelii</i> )	PPYG2	-
Gibbon ( <i>Nomascus leucogenys</i> )	Nleu1.0	GCA_000146795.1
Macaque ( <i>Macaca mulatta</i> )	Mmul_8.0.1	GCA_000772875.3
Olive baboon ( <i>Papio anubis</i> )	PapAnu2.0	GCA_000264685.1
Vervet-AGM ( <i>Chlorocebus sabaeus</i> )	ChiSab1.1	GCA_000409795.2
Marmoset ( <i>Callithrix jacchus</i> )	C_jacchus3.2.1	GCA_000004665.1
Tarsier ( <i>Tarsius syrichta</i> )	tarSyr1	-
Bushbaby ( <i>Otolemur garnettii</i> )	OtoGar3	GCA_000181295.3
Mouse lemur ( <i>Microcebus murinus</i> )	Mmur_2.0	GCA_000165445.2

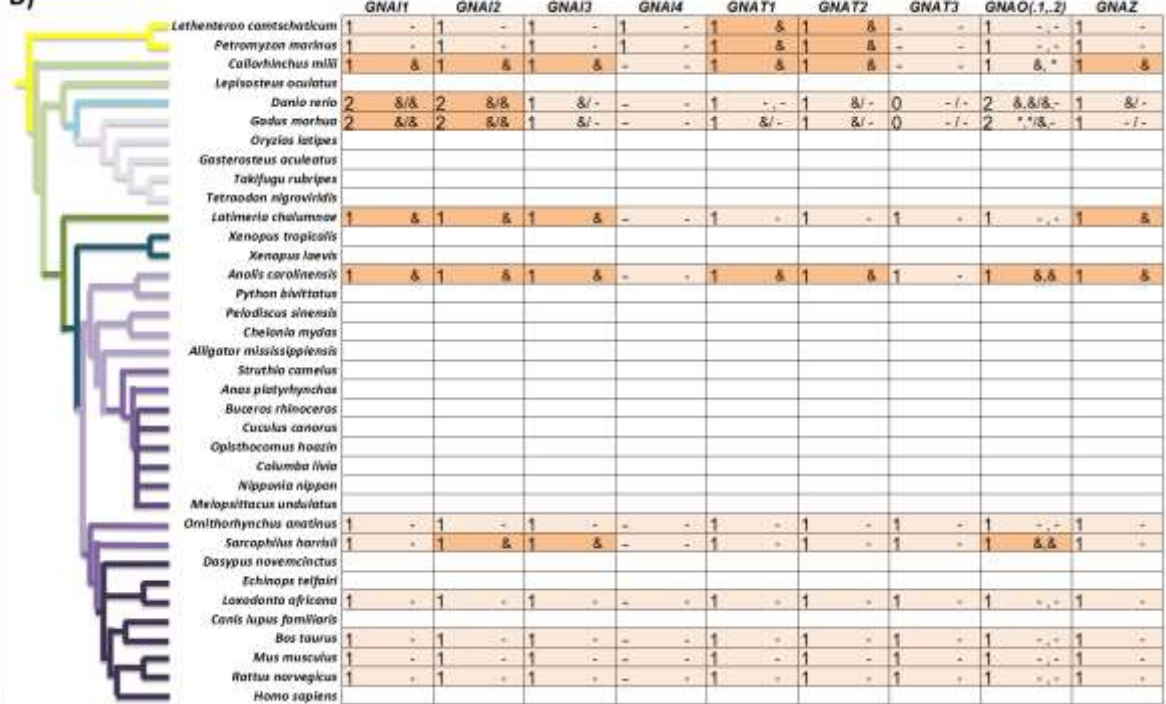
**Supplemental Figure 2.1: Primate species investigated for pseudogenes.** The existence of *GNA*-pseudogenes was investigated within human and 11 other primate species. A) Primate species investigated. The Latin names and clades for each species are provided. Ce – Cercopithecoidea. B) Column1 – Common name (*Genus species*). Column2 – Genome assembly used. Column3 – Accession number for genome assembly.



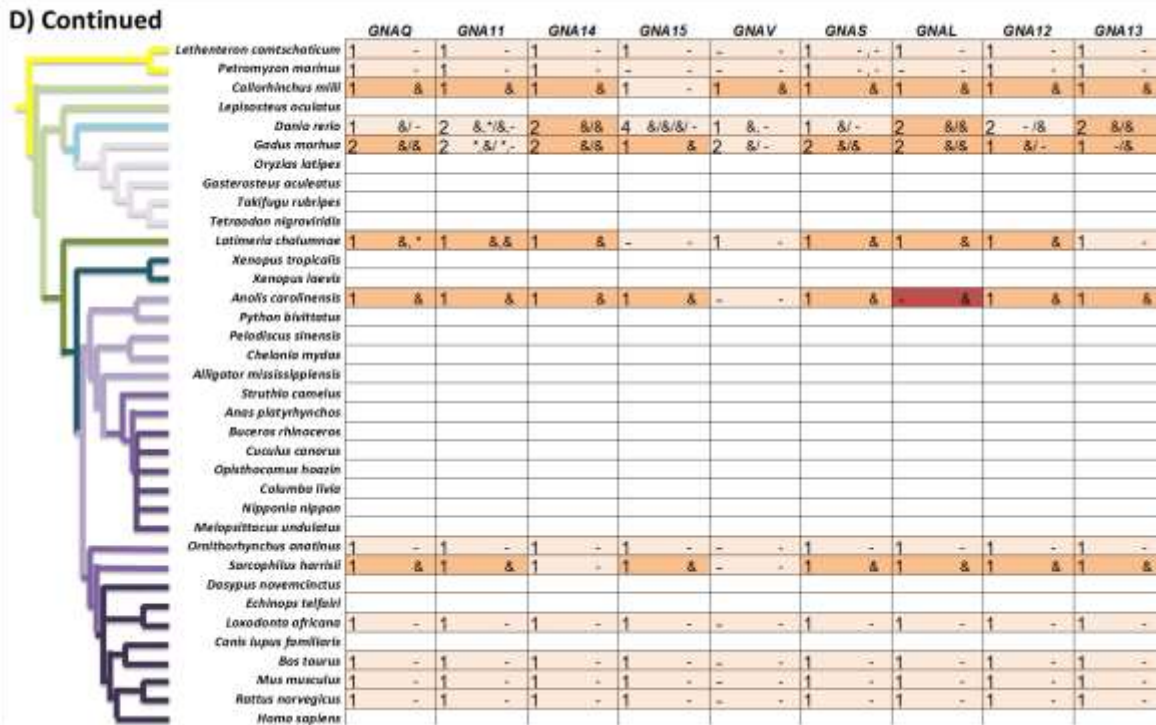
C) Continued



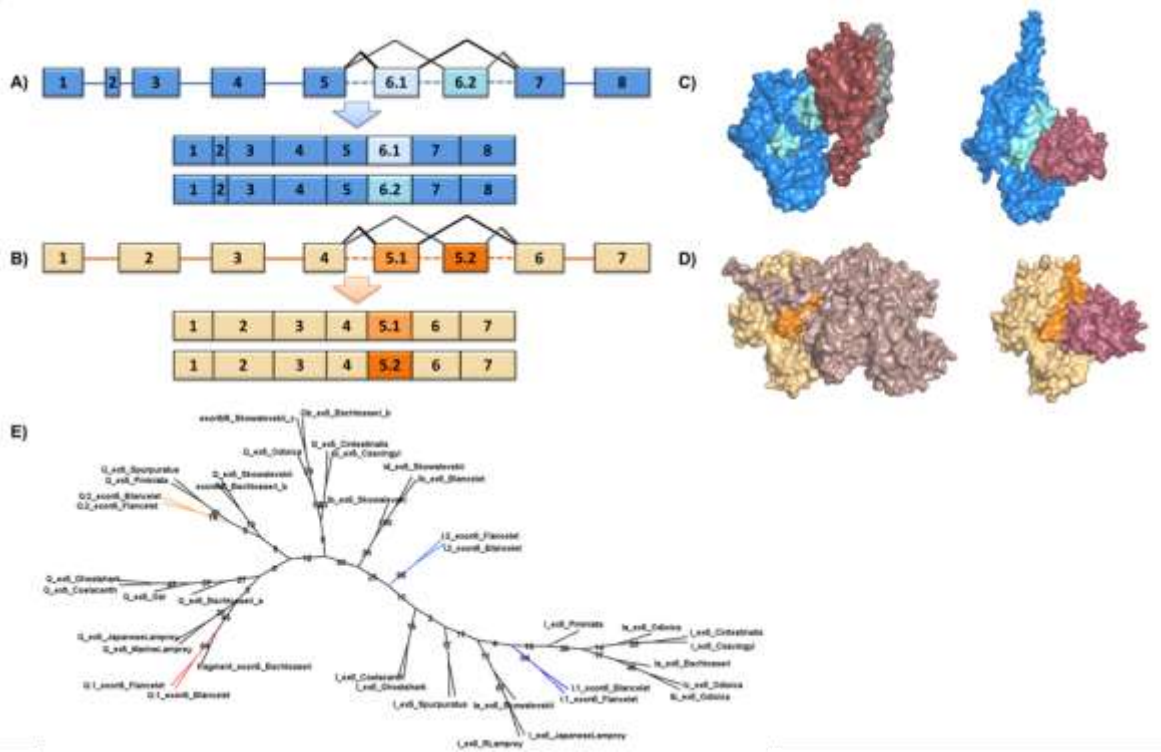
D)



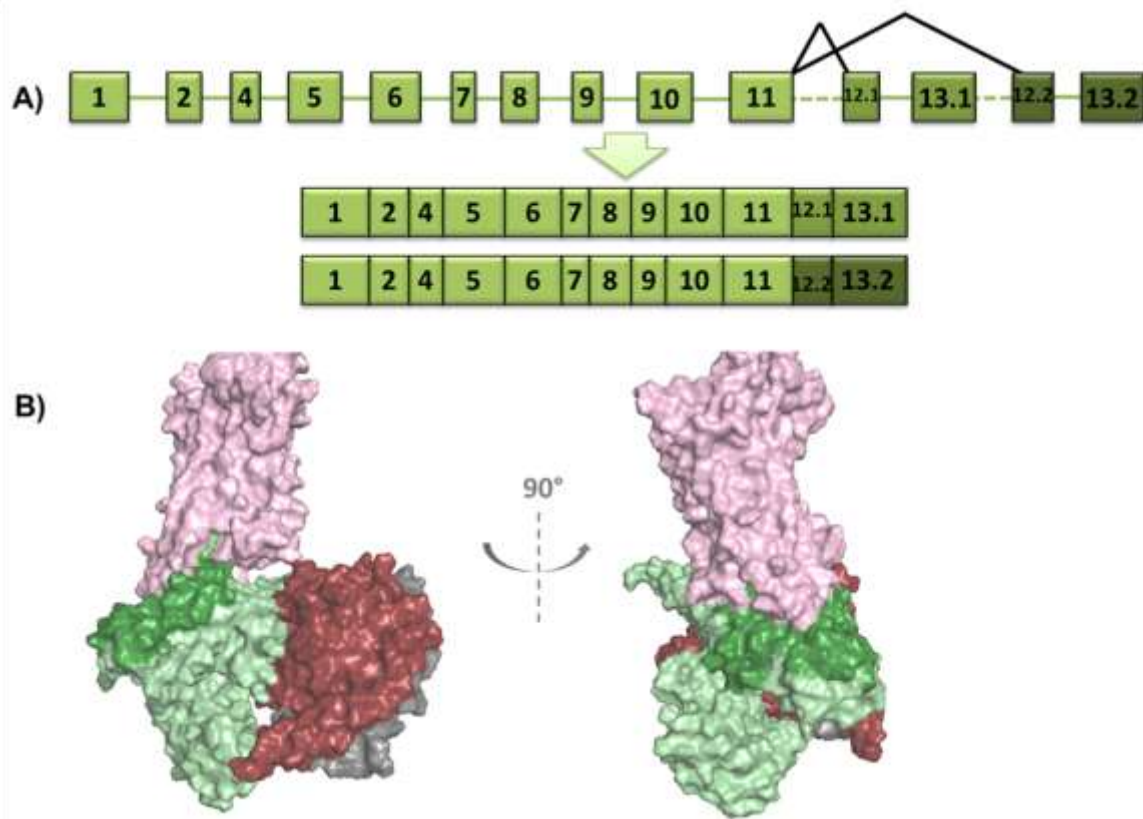




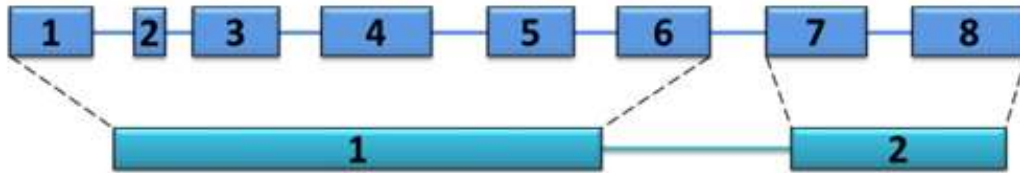
**Supplemental Table 2.2: Transcriptome and Expression Data.** All gene sequences were validated by blasting against Expressed Sequence Tags (EST) and/or Transcriptome Shotgun Assembly (TSA) data when available<sup>103, 104</sup>. The tables show which species and paralogs were validated. The first number indicates the number of genes found per family (same as Table 1); the smaller characters represent EST/TSA data for each paralog (see Figure 2). “@” indicates that a full-length or partial expression read fragment was found, “&” indicates a full-length or partial transcriptome read, “-” indicates no EST/TSA support was found. “/” separates multiple paralog gene copies (a, b, c, d), “,” indicate multiple transcript variants exist which include different exons (.1,.2), “ \* ” indicates EST/TSA data did not include exon sequences for respective alternative transcripts (.1,.2). Dark blue/orange boxes indicate all paralogs were validated by partial or full EST/TSA hits, light blue/orange boxes indicate no reads were found to support that paralog. White boxes indicate that no EST or TSA data were available for analysis. Red boxes indicate EST and TSA data were found without sequence evidence for the gene present within the genome assembly. A) Non-Vertebrate Deuterostome EST data. B) Non-Vertebrate Deuterostome TSA data. C) Vertebrate EST data. D) Vertebrate TSA data.



**Supplemental Figure 2.2: Implications of alternative exon usage on tertiary structure in lancelet *preGai* and *preGaq*.** A) Mutually exclusive inclusion of lancelet exon6.1 and 6.2 in *preGNAI* (blue) yields two different transcripts during alternative splicing. Representative box lengths correlate with the average curated exon lengths (intron lines do not). B) Mutually exclusive inclusion of lancelet exon5.1 and 5.2 in *preGNAQ* (beige) also yields two different transcripts during alternative splicing. C) Splice variant exon borders mapped onto two *Gai* crystal structures (PDB IDs 1GP2<sup>131</sup> and 1AGR<sup>132</sup>, respectively). The sequence encoded by exon6 (light blue) influences the interface between the G $\beta\gamma$  subunits of the heterotrimer (crimson/grey - left) and downstream effector protein partners such as the RGS protein (purple - right). D) Splice variant exon borders mapped onto two *Gaq* crystal structures. The sequence encoded by exon5 (orange) influences the protein interfaces between effector proteins such as PLC (lavender - left) and RGS (purple - right) (PDB IDs 4QJ3<sup>128</sup> and 5DO9<sup>129</sup>). E) ML tree of (*pre*)*GNAI/GNAQ* exons indicates both duplications were independent.



**Supplemental Figure 2.3: Implications of alternative exon usage on tertiary structure in lancelet *preGNAS*.** A) Alternative splicing of lancelets exon12 and 13 in *preGNAS* (green) yields two different mutually exclusive transcripts. B) Splice variant exon borders (dark green) mapped onto a *Gas* structural model bound to the G protein  $\beta\gamma$  subunits (crimson/grey) and a GPCR (pink) respectively and rotated 90°.



**Supplemental Figure 2.4: Exon structure of *GNAI* and *GNAZ*.** Most members of the *Gai* family have a conserved gene structure with 8 protein-coding exons, similar exon lengths, and five conserved split codons shared across exons. The relative exons lengths of *GNAI* genes are represented by dark blue boxes. *GNAZ* only possesses two protein-coding exons (light blue). The first *GNAZ* exon sequence maps to exons 1-6 of *GNAI*, while the second *GNAZ* exon position maps to exons 7 and 8 of *GNAI*. This exon-intron structure is indicative of a retrotransposition. The intron sequence may have been reinserted later into the gene to promote transcription.

<b>Residue Position (Human)</b>	<b>F1X4</b>	<b>F3X4</b>	<b>Codon Table</b>	<b>Sum</b>
<b>249 K</b>	0.978	0.635	0.549	2.162
<b>259 W</b>	1.000	0.998	0.980	2.978
<b>261 T</b>	0.968	0.596	0.638	2.202
<b>276 E</b>	0.989	0.808	0.507	2.304
<b>298 T</b>	0.693	0.962	0.970	2.625
<b>299 E</b>	0.973	0.628	0.459	2.060
<b>313 K</b>	0.981	0.999	0.997	2.977
<b>321 T</b>	0.999	0.997	0.988	2.984
<b>334 F</b>	0.975	0.881	0.709	2.565
<b>342 V</b>	0.972	0.442	0.409	1.823

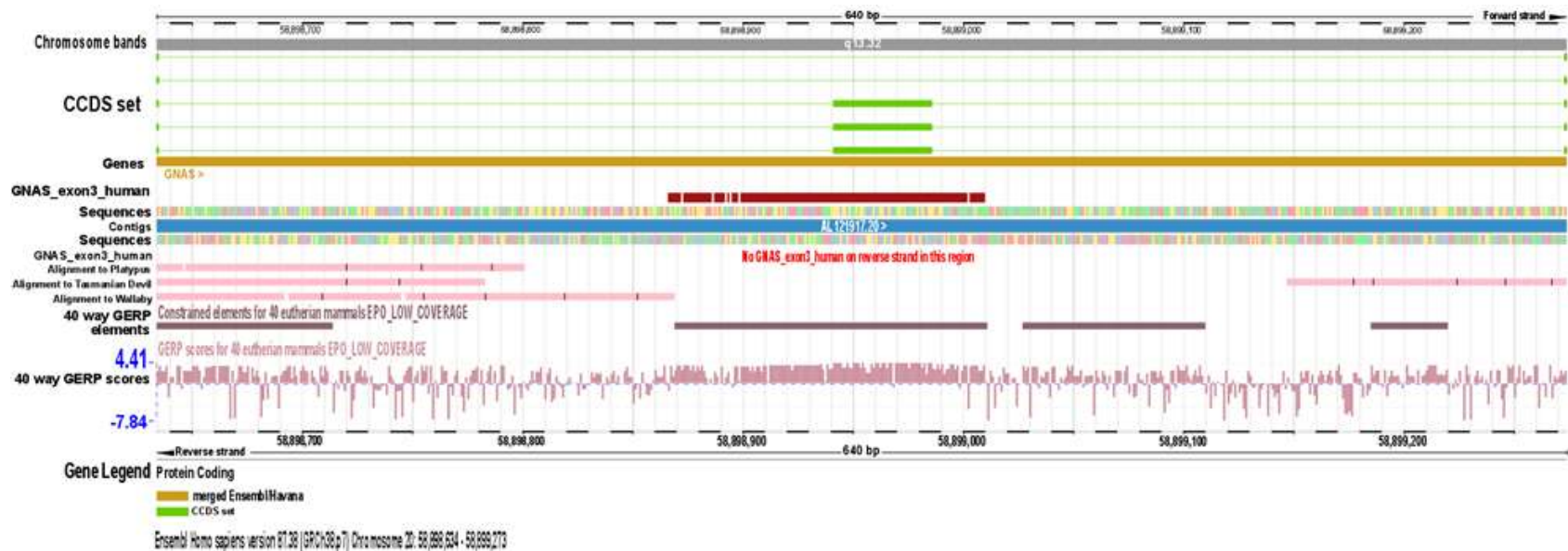
**Supplemental Table 2.3: Sites under positive selection in the branch leading to *GNAO1*.** Data is given for those residues that have a BEB probability for being in class 2a (sites under positive selection) for branch #1 (Figure 9) > 90% in at least one of the tested codon models (F1X4, F3X4, Codon table). The probabilities > 90% are marked in red. The identity and numbering of the residues in respect to the full length protein sequence in human are given in column 1.



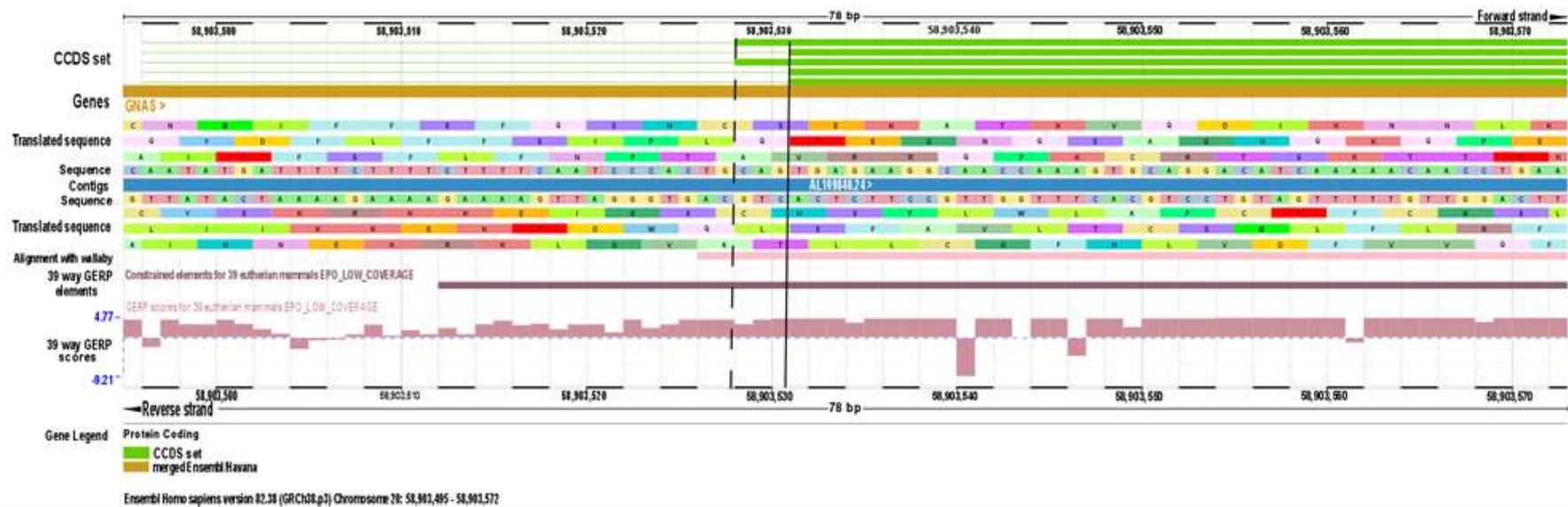
<b>Model</b>	<b>LR</b>	<b>P-value</b>	<b>Significance</b>	
BS, F1X4	19.979	$7.8 \times 10^{-6}$	***	
BS, F3X4	7.301	$6.8 \times 10^{-3}$	**	
BS, Codon Table	6.802	$9.1 \times 10^{-3}$	**	
<b>Model Parameters</b>	<b>Mean</b>	<b><math>\sigma</math></b>	<b>Q<sub>1</sub></b>	<b>Q<sub>2</sub></b>
p0	0.883	0.995	0.861	0.892
p1	0.017	0.014	0.008	0.015
w0	0.017	0.004	0.013	0.016
wFG	613.040	428.657	163.727	999

**Supplemental Table 2.4: Significant results of the branch-site model indicate positive selection in the *GNAO.1* #1 branch.** The result of the likelihood ratio test was compared to a  $\chi^2$  distribution with following significance levels \*\* < 0.01, \*\*\* < 0.001 for each codon model tested (F1X4, F3X4, codon table) in the #1 branch of *GNAO.1* (marked in Figure 9). All other tested branches (#2, \*1, \*2, and \*3) were not significant. Robustness of the parameter inferences (p0, p1, w0, wFG) was accessed by bootstrapping. BS = Branch-Site, LR = Likelihood Ratio,  $\sigma$  = standard deviation, Q<sub>1</sub> = First Quantile (25<sup>th</sup> percentile), Q<sub>2</sub> = Second Quantile (75<sup>th</sup> percentile).

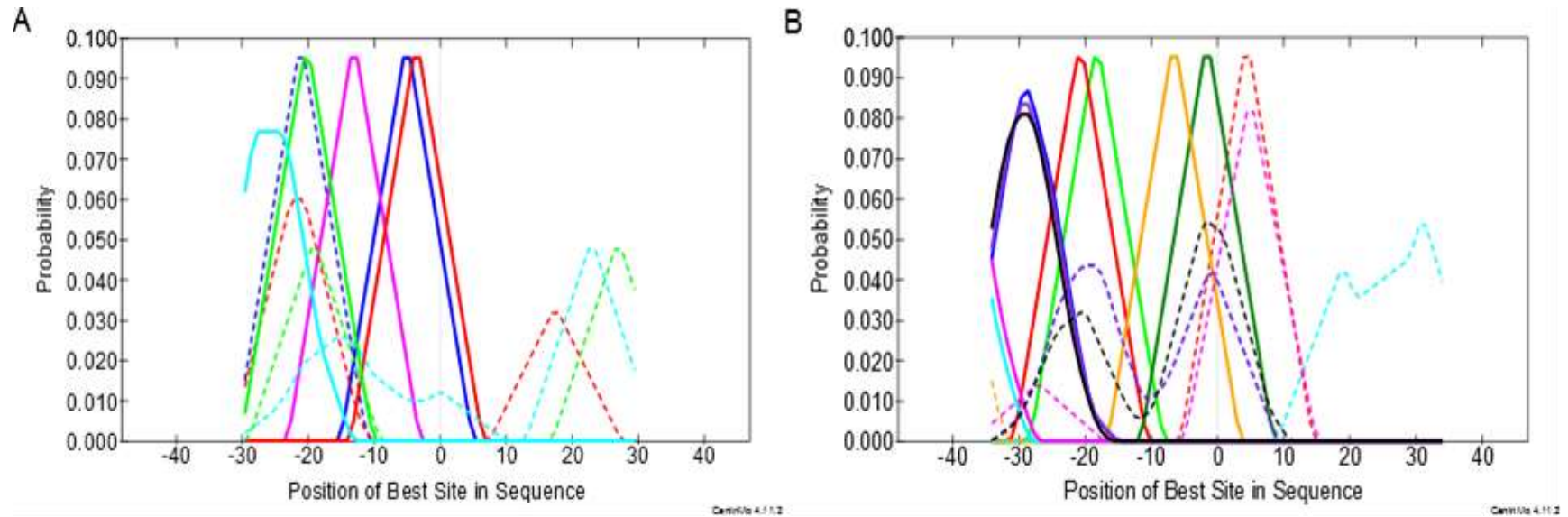




**Supplemental Figure 2.6: Exon 3 of *GNAS* in human.** Expression of exon3 is supported by CCDS data. A region ~75 nt upstream and 25 nt downstream of the exon boundaries shows high levels of conservation in placentals. The same region is not conserved in non-placental mammals (platypus, wallaby and Tasmanian devil) as no BLASTz hits are retrieved (pink boxes). The figure was created with the Ensembl webserver<sup>97</sup>. Bp - Basepair, CCDS - consensus coding sequence, GERP - Genomic Evolutionary Rate Profiling.

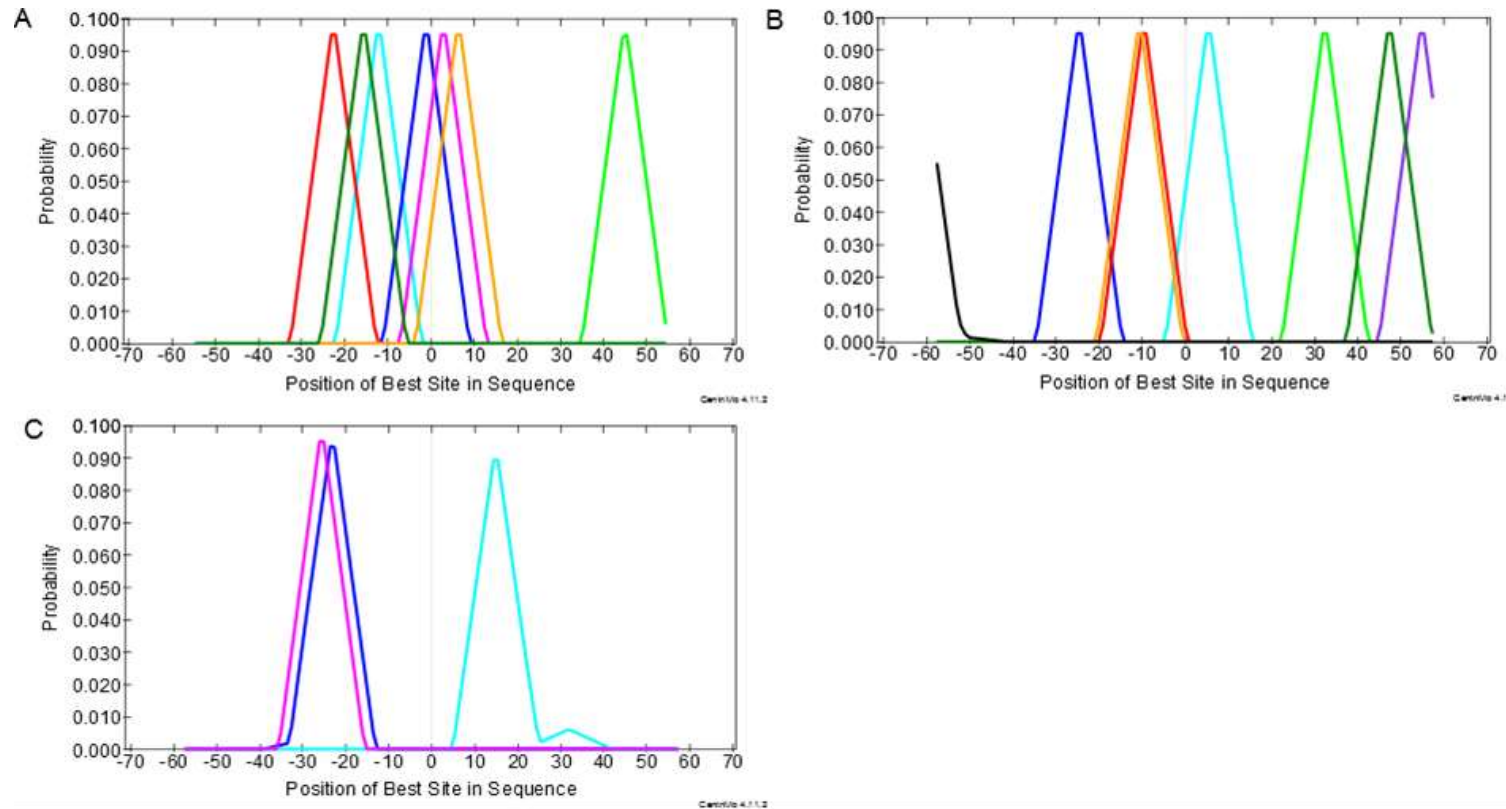


**Supplemental Figure 2.7: Extension of exon4 of *GNAS* in placental mammals.** The 3' genomic region of intron3 and 5' region of exon4 for human. There are two alternative 3' SS of intron3 that can be used to include exon4 into different isoforms of *GNAS*. The first is mediated by a 'TG' (SS) (exon border - dotted line) while the second is mediated by a canonical 'AG' SS (exon border - solid line). The upper most line of the translated sequence track represents the reading frame of exon4. The track with 40 way GERP elements/40 way GERP scores shows a region with high conservation in 40 different placentals (eutherian mammals) that exceed the exon-intron boundary. The homologous region, as retrieved from the BLASTz alignment, is shorter in the non-placental wallaby gene (pink box). The figure was created using the Ensembl webserver<sup>97</sup>. Bp - Basepair, CCDS - consensus coding sequence, GERP - Genomic Evolutionary Rate Profiling.

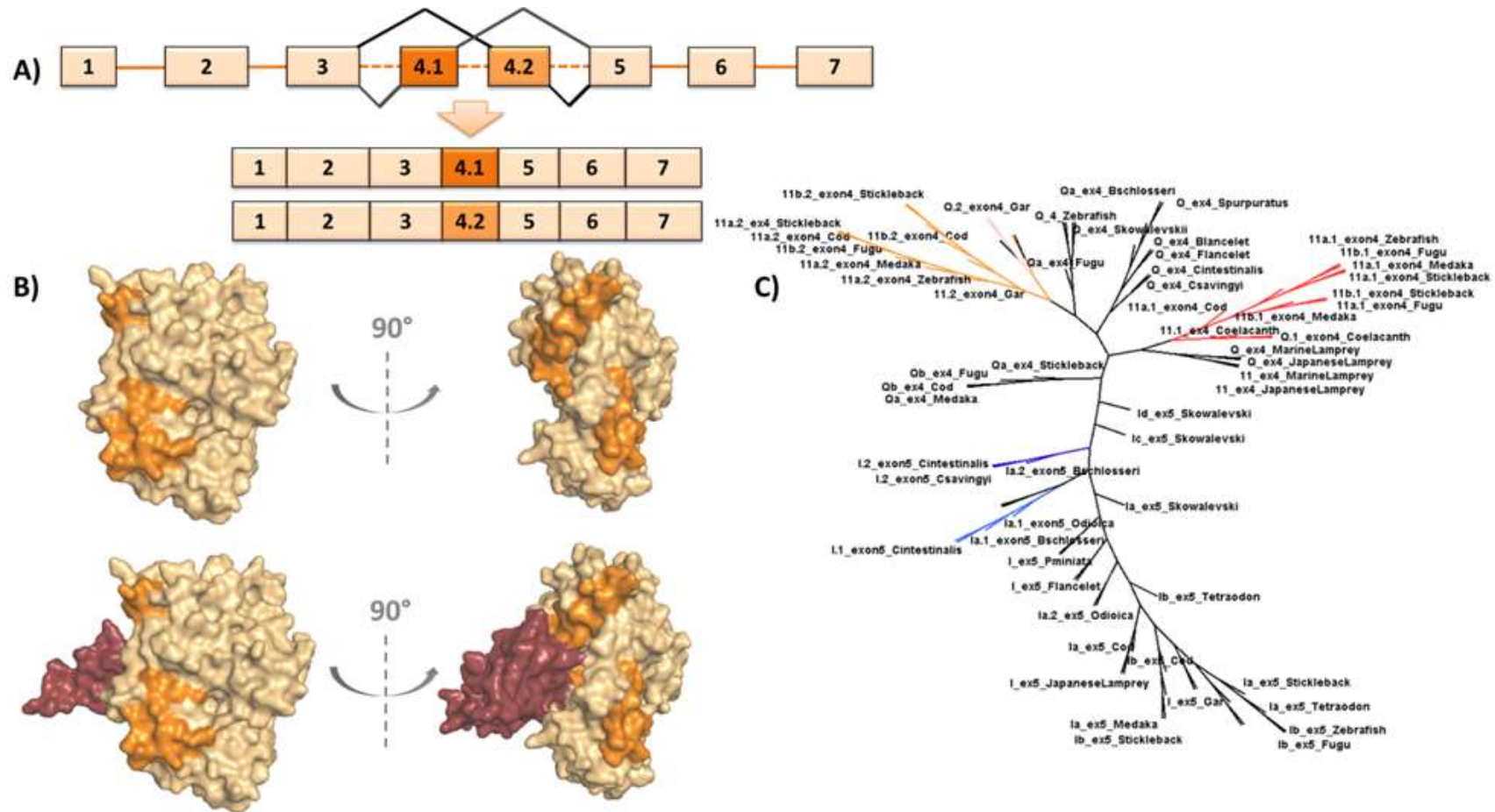


**Supplemental Figure 2.8: DNA- and RNA-binding protein motifs overlapping with the 3' canonical and non-canonical splice sites of intron 3 in *GNAS*.** All included motifs are predicted to occur in the positive set (for six placentals), but not at the same position in the control set (eight non-placental lobe-finned fish). Note, that some motifs occur in the control set, but at a different position than in the positive set, e.g. *Gata4* or *RIN*. The shown motifs overlap with the conserved intronic region upstream of exon 4. A) Local enrichment of known DNA-binding protein motifs (DBP) in comparison to a uniform distribution of motifs (E-value < 1). *Gata4* (blue), *Mybl1\_secondary* (pink), *GATA3\_full* (red), *Sox4\_secondary* (green), *FOXP1* (turquoise). B) Local enrichment of known RNA-binding protein motifs in comparison to a uniform distribution of motifs (E-value < 1). *PCBP1* (light green), *RIN* (red), *Tb\_0230* (dark green), *Rbm24* (orange), *SHEP* (turquoise), *U2AF2* (dark blue), *RBM47* (pink), *U2AF50* (purple), *Tv\_0257* (black). The non-canonical splice site is located at position -7. Sequence positions < -7 belong to intron 2 while positions > -7 belong to exon 4. The y-axis indicates the probability of a DBP/RBP binding centrally at the indicated position for the positive set (solid lines) and for the control (dotted lines). None of the motifs occur more often at a specific position in the positive set than in the control set (Fisher's exact test). The figure was created with Centrimo<sup>115</sup>.

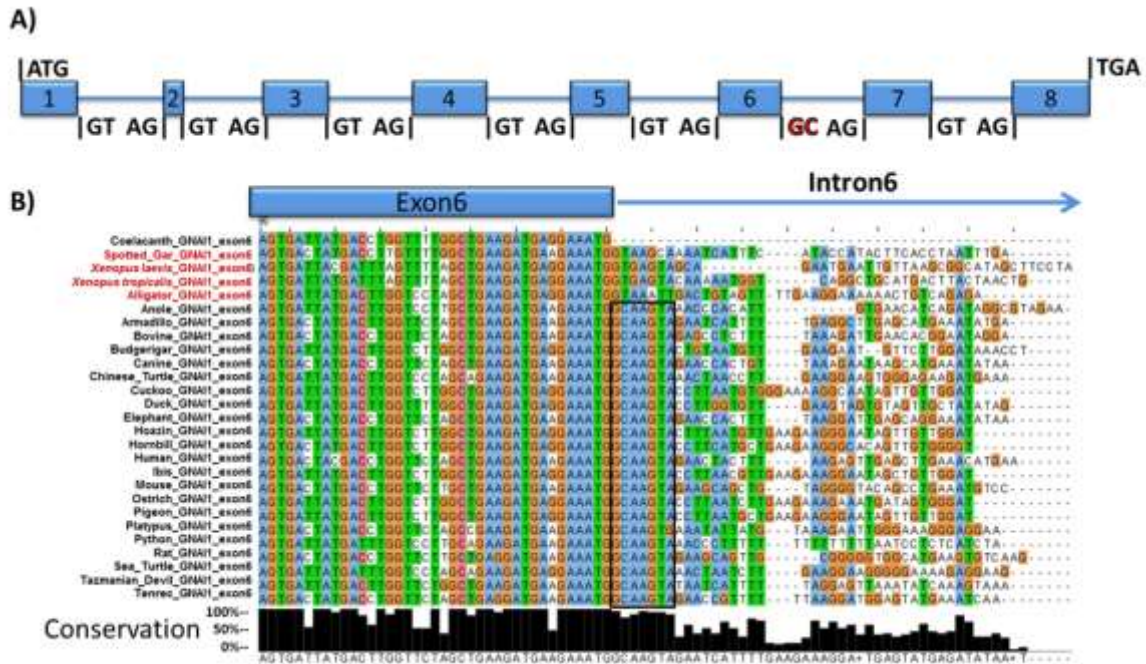




**Supplemental Figure 2.9: DNA- and RNA-binding protein motifs overlapping with the extended conserved region around exon 3 in *GNAS* of 33 placentals.** Exon 3 is located at positions 0-46 on the x-axis. A) Local enrichment of known DNA-binding protein (DBP) motifs in comparison to a uniform motif distribution. 30 motifs are enriched in the reported region with a E-value < 0.0001 in all investigated placentals; only a subset of these is shown for clarity: Gfi1 (light blue), Hltf (dark blue), EGR1 (pink), MZF1\_5-13 (light green), En1 (red), E2F4 (orange), Hoxc9 (dark green). B) Local enrichment of known RNA-binding protein (RBP) motifs in comparison to a uniform motif distribution. Eight motifs are enriched in the reported region with an E-value < 0.0001 in all investigated placentals. TARDBP (purple), DAZAP1 (dark blue), PPRC1 (light blue), SRSF9 (light green), SRSF10 (red), CNOT4 (orange), PCBP1 (dark green), BRUNL6 (black). Note that the SRSF9 binding site is located within the exon and does not overlap with either splice site. C) Local enrichment of RBP sites predicted by Pollard *et al.*<sup>188</sup>. The respective motifs do not occur in all investigated placentals as indicated by a lower probability. SRSF2 (dark blue), SRSF1 (light blue), HNRNPA1 (pink). The 3' AG SS is located at position 0 along the x-axis. The y-axis indicates the probability of a DBP/RBP motif being located centrally at this position. The figure was created with Centrimo<sup>115</sup>.

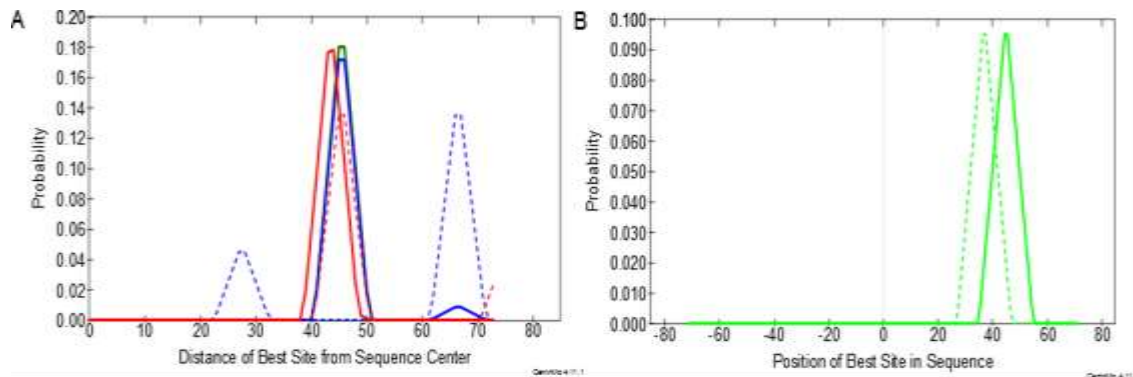


**Supplemental Figure 2.10: Local exon duplications of *GNAQ*, *GNA11*, and *preGNAI*.** A) Alternative splicing of two mutually exclusive exon4 of *GNAQ* and *GNA11* result in two different RNA transcripts represented. Box lengths correlate with average curated exon lengths (intron line lengths do not correspond to intron lengths). B) Tertiary crystal structure of mammalian Gαq (taupe) with exon4 (orange) borders mapped with RGS protein interaction removed (top) and with RGS present (bottom - ruby) (PDBID 5D09<sup>129</sup>). Alternatively spliced exon4 provides sequence diversity for critical protein-protein interfaces such as the RGS protein (purple). C) ML trees of nucleotide sequences from exon4 of *GNAQ/GNA11* and exon5 of *GNAI* across basal Chordates.



**Supplemental Figure 2.11: 5' non-canonical splice site pattern of *GNAI1* intron6 in birds and mammals.** A) Schematic representation of the primary transcript sequence of the *GNAI1* gene in birds and mammals with the start and stop codons as well as the SS explicitly shown. Possible untranslated regions (UTRs) are not shown. The representative exons (boxes) are drawn to approximate scale with their nucleotide length while introns (lines) are not drawn to scale. B) 5' SS of intron6 in *GNAI1* of lobe-finned fishes and spotted gar. The first seven nt of intron6 are highly conserved in all mammals and birds (black box), while they vary in alligator, frogs and spotted gar (species marked in red). The intron sequence, and thus SS, is unknown for coelacanth. The first two nt of the boxed region constitute the SS pattern GC/GT. The figure was produced with the Jalview alignment viewer<sup>171</sup>.





**Supplemental Figure 2.12: DNA- and RNA-binding protein motifs overlapping with the 5' non-canonical splice site of intron6 in *GNAII*.** A) Local enrichment of known DNA-binding protein (DBP) motifs in comparison to a uniform motif distribution are shown for lobe-finned fish with 'GC' SS (positive set) versus lobe-finned fish and spotted gar with 'GT' splice site (SS) (control). The shown motifs are either present in all species of the positive set and in none of the controls (PRDM1\_full, FXR1) or follow this rule with at most one exception. Mafk\_secondary UP0004\_2 (light blue), NFIX\_full\_3 (dark blue), PRDM1\_full (green), STAT2:STAT1 (pink, behind green). B) Local enrichment of known RNA-binding protein (RBP) motifs in comparison to a uniform motif distribution. FXR1 (lime green). The SS is located at position 45 along the x-axis. Sequence positions <45 correspond to exon6, while positions >45 correspond to intron6. The y-axis indicates the probability of a DBP/RBP motif present centrally at the indicated position for the positive set (solid line) and the control set (dotted line). None of the motifs occurs surprisingly more often at a specific position in the positive set than in the control set (Fisher's exact test). The figure was created with Centrimo<sup>115</sup>.

Retrogene Name	Location	Retrogene Synteny	Parent Gene	Parent Family	Parent Location	LCA	RNA code	Promotor	Conserved ORF > 40 AA
<i>GNAI2P2</i>	Chr 10	POLR3A	<i>GNAI2</i>	<i>GNAI</i>	Chr 3	Primates	No	No	No
<i>GNAI2P1</i>	Chr 12	ATF7IP	<i>GNAI2</i>	<i>GNAI</i>	Chr 3	Simiiformes	Yes, +	No	No
independent retrogene 1	Chr 13 (Cja)	GOLGA7	<i>GNAI2</i>	<i>GNAI</i>	Chr 15 (Cja)	Cja	-	-	ORF: AA 241-355
<b><i>GNAQP1</i></b>	Chr 2	PLEKH2	<i>GNAQ</i>	<i>GNAQ/11</i>	Chr 9	Catarrhini	Yes, -	Yes	ORF: AA 59-100
independent retrogene 2	Scaffold_131918 (Csy)	-	<i>GNAQ/GNA11</i>	<i>GNAQ/11</i>	GeneScaffold_5240 (Csy) -	Csy	-	-	ORF: AA 47-110
independent retrogene 3	AHZD01183494.1 (Pan), JSUE03219333.1 (Mmu)	-	<i>GNAQ</i>	<i>GNAQ/11</i>	Chr 1 (Mmu, Pan)	Papionini	-	-	No
<i>GS1-124K5.9</i>	Chr 7	TPST1, KCTD7	<i>GNA11</i>	<i>GNAQ/11</i>	Chr 19	Catarrhini	Yes, +	Yes	ORF: AA 312-359
<b><i>GS1-124K5.9 duplication 1</i></b>	<b>Chr 3 (Cercopithecinae)</b>	<b>TMEM248</b>	<b><i>GS1-124K5.9</i></b>	<b><i>GNAQ/11</i></b>	<b>Chr 3 (Cercopithecinae)</b>	<b>Cercopithecinae</b>	-	-	<b>No</b>
<b><i>GS1-124K5.9 duplication 2</i></b>	<b>Chr 7 (Ggo)</b>	<b>TPST1, <i>GS1-124K5.9</i></b>	<b><i>GS1-124K5.9</i></b>	<b><i>GNAQ/11</i></b>	<b>Chr 7 (Ggo)</b>	<b>Ggo</b>	-	-	<b>ORF: AA 59-359</b>
<i>RP11-611O2.6</i>	Chr 12	CPM (intronic)	<i>GNA12</i>	<i>GNAI</i>	Chr 1	Simiiformes	Yes, -	No	No
<i>AC010975.2</i>	Chr 2	CTNNA2 (intronic)	<i>GNA13</i>	<i>GNA12/13</i>	Chr 17	Primates	Yes, -	No	No
<i>RP11-100N8.2</i>	Chr 11	OR9G4	<i>GNAS</i>	<i>GNAS</i>	Chr 20	Simiiformes	No	No	No
independent retrogene 4	Chr 3 (Cja)	KCTD8	<i>GNAS</i>	<i>GNAS</i>	Chr 5 (Cja)	Cja	-	-	ORF: AA 689-815
independent retrogene 5	Chr 2 (Cja)	MTX3, CMYA5	<i>GNAS</i>	<i>GNAS</i>	Chr 5 (Cja)	Cja	-	-	ORF: AA 703-828
independent retrogene 6	Chr 11 (Cja)	-	<i>GNAS</i>	<i>GNAS</i>	Chr 5 (Cja)	Cja	-	-	ORF: AA 703-784
independent retrogene 7	Chr 1 (Cja)	<i>U6snRNA</i>	<i>GNAS</i>	<i>GNAS</i>	Chr 5 (Cja)	Cja	-	-	No
independent retrogene 8	Scaffold_5972 (Csy)	-	<i>GNAS</i>	<i>GNAS</i>	GeneScaffold_8406 (Csy)	Csy	-	-	No

**Supplemental Table 2.5: Retrogenes in Primates.** The table summarizes the properties of *GNA*- retrogenes found in primates. Two retrogenes (highlighted in bold) are the result of independent duplications of an existing retrogene. All other retrogenes are the result of a retrotransposition event. The retrogene name, location, location of the parent and the proximity to a promotor are given for human unless specified differently in parenthesis. The retrogene is situated next to the gene specified in the synteny column for the phylogenetic group given in the column ‘LCA’ (last common ancestor). Requiring conservation within the complete phylogenetic group, the coding potential of the respective region was evaluated with RNAcode<sup>120</sup> (+: methionine contained in open reading frame, ORF; -: no methionine in ORF). Conserved ORFs that are similar to the parent ORF were detected via *blastn* with the human parent gene as query. Cja – *Callithrix jacchus*, Ggo – *Gorilla gorilla*, Csy – *Tarsius syrichta*, Mmu – *Macaca mulatta*, Pan – *Pongo abelii*.

## CHAPTER 3

### THE ACTIVATION OF HETEROTRIMERIC G PROTEINS: EVALUATION OF THE PAST AND PRESENT MECHANISTIC MODELS

#### 3.1 Introduction

##### *Chapter 3*

Lokits AD, “The Activation of Heterotrimeric G proteins: Evaluation of Past and Present Mechanistic Models” *Vanderbilt Reviews Neuroscience* **2014**

##### *Contribution*

I am the sole author of this manuscript. I contributed all text: abstract, introduction, reviewed all articles and conclusion. I edited all texts and created the figure as modified from *Neuroscience*, Purves, 4<sup>th</sup> Ed.

##### *Abstract*

Cell signaling through G protein coupled receptors (GPCRs) represents one of the most modulatory and regulatory communicative mechanisms in the nervous system. Extracellular signals in the form of neural peptides, hormones, neurotransmitters, small molecules, and ligands are converted into neural responses through GPCR interactions with heterotrimeric G proteins. Since their discovery, GPCRs have become the largest therapeutic target for a wide range of human pathologies including diseases such as Schizophrenia, autism, neuropathic pain, sleep/wake disorders, and bipolar disorder. However, despite their importance, the mechanism of G protein activation through their cognate receptors for signal transduction has not been fully elucidated. This review looks to introduce the basic structure and function of heterotrimeric G proteins, the challenges present in the field which have hindered our understanding, and summarize the different proposed mechanisms of activation leading to intracellular signaling cascades. Lastly, our current working model of activation will be presented as a unified mode of understanding

these dynamics while predicting where the field looks for future drug discovery and research.

### ***Introduction***

Cell signaling is a fundamental process required in all living organisms for development and homeostasis. In neurons, rapid cell to cell signaling via chemical synapses requires ligand-gated ion channels which quickly activate and alter the membrane potential through their ability to directly fluctuate local ion concentrations<sup>192</sup>. Slower or more modulatory signaling is often mediated through G protein coupled receptors (GPCRs) which require more time to alter the membrane potential as they are not ion channels themselves, but interact with transducing proteins to communicate their signal and propagate an intracellular response<sup>193</sup>. GPCR-mediated ion channels along the post synaptic membrane are therefore acted upon through a course of protein-protein interactions. With around 800 different types of GPCRs, this modulatory class of cellular receptors accounts for 2% of the expressed protein population<sup>194</sup>. This diversity of receptor subtype provides an additional layer of neuron and synapse specific signaling.

G protein coupled receptors (GPCRs) are the largest and most diverse class of membrane receptors in eukaryotes<sup>10</sup>. In the nervous system, GPCRs can be present both pre-and post- and even peri- synaptically to differentially influence neuronal communication<sup>195-198</sup>. This allows for modulation on all sides of the “message” being transmitted. Likewise, their ability to bind many different forms of ligands makes these receptors responsible for an estimated 40% of all signaling within the cell<sup>173</sup>. This significant role in cell to cell communication in conjunction with their transmembrane locations makes GPCRs a prominent therapeutic target<sup>199</sup>. Though roughly half of current

therapeutics act upon GPCRs, a consensus model of the mechanism of G protein activation via their receptors does not exist. Such a model would provide better understanding of modulator neuronal communication as well as increase the current understanding of cell signaling on a broader scale.

### ***G Protein Signaling Cycle***

As implied by their name, G protein-coupled receptors interact with intracellular, heterotrimeric complexes, called G proteins, in order to transduce their signal into a cellular response. This signal exchange occurs, as depicted in Figure 3.1 with the Rhodopsin GPCR, when an external ligand activates the receptor allowing for association of the heterotrimer. Composed of three subunits, the  $G\alpha$  and functional  $G\beta\gamma$  dimer subunits become activated after binding to an activated receptor ( $R^*$ ); this interaction catalyzes the  $G\alpha$  subunit to exchange GDP for GTP. This requires a conformational change that rearranges the binding affinities within the complex, promoting the  $G\alpha$  subunit to dissociate from the receptor and its  $G\beta\gamma$  subunits. The activated and dissociated  $G\alpha$  may then interact with and initiate signaling cascades with its effector proteins while in the GTP-bound state. However, as most  $G\alpha$  subunits possess intrinsic enzymatic activity, the  $\gamma$  phosphate of the GTP is ultimately hydrolyzed resulting in a GDP-bound  $G\alpha$  subunit. This conformation, once again, maintains a higher affinity for its  $\beta\gamma$  pair than its effector proteins allowing for the termination of the signal and the recycling of the transducer signaling machinery at the membrane interface for future signaling events coupled to the receptor<sup>11, 173, 193</sup>. The present work seeks to synthesize the abundance of literature surrounding this activation process and the current mechanistic models of the allosteric interaction networks required to modulate the  $G\alpha$  subunit for activation.

### ***Difficulties to elucidate the Structure and Mechanics***

It has long been known that the ligand binding site in the archetypical, Class A rhodopsin GPCR, required for Phototransduction in vision, is approximately 40 Angstroms away from its intracellular loops which bind to the  $G\alpha$  subunit<sup>200-202</sup>. This interface of the receptor-G protein is 30 Angstroms away from the  $G\alpha$  subunit's nucleotide binding pocket<sup>11, 173</sup>. Therefore, the extracellular ligand must induce an allosteric conformational change some 70 Angstroms from its binding site; a dynamic conformational change must propagate across the complex to release the basal GDP nucleotide in exchange for GTP. This has led to multiple proposed  $G\alpha$  activation mechanisms; each model attempts to understand how the securely lodged nucleotide becomes expelled from within the GTPase and helical domains of  $G\alpha$ .

However there have been many practical and technical difficulties hindering the elucidation of this mechanism. Among them is the inherent GPCR-G protein cross talk which makes studying only one G protein subtype difficult as multiple  $G\alpha$  subunits have been shown to interact with the same receptor but elicit a different cellular response<sup>203</sup>. Likewise it is still not understood which of the various possible  $\beta\gamma$  combinations<sup>204</sup> and  $\alpha/\beta\gamma$  combinations<sup>21, 205</sup> are possible *in vitro* and *in vivo*. Therefore all mutational studies for G protein activation are severely limited in scope as they only focus on a handful of known G protein  $\alpha$ ,  $\beta$ , and  $\gamma$  combinations. Confounding this limitation in the first messenger signal components, there is cross talk in the downstream signaling effector molecules activated by  $G\alpha$  and  $\beta\gamma$ . This creates caveats in using indicators such as cAMP accumulation as measures of G protein activation as it can be activated by both the  $G\alpha$  subunit and  $\beta\gamma$  subunits of different subfamilies. Technical difficulties with expressing and purifying individual subunits and receptors in large enough quantities for experimentation

has led to the use of many different cell lines for the various protein components. Even once these challenges are met, many studies have been limited by the presence of endogenous signaling machinery present in the experimental cell lines.

In conjunction with this, crystallization of these proteins in their various states has been a challenging enterprise as these complexes are composed of highly flexible domains, very transient conformational switching states, and a lipid membrane-bound receptor. Even with recent advances in crystallographic techniques, nanodiscs, and stabilizing nanobodies, crystal structures only provide a “snapshot” of one or more of the various conformations the protein complexes may undergo<sup>206, 207</sup>. Instead, to probe the various physiologically relevant conformations the complex is capable of maintaining across the entire signaling cycle, a more dynamic approach is required to elucidate the modulatory and inherently flexible states these proteins undergo during the signaling cycle.

### ***Current Structural Knowledge***

The elucidation of atomic resolution crystallographic and NMR structures has been paramount to the advances in conformational dynamics of G protein activation. The GDP-bound crystal structures<sup>131, 208</sup> of the heterotrimeric G protein as well as the dissociated dimeric subunits of G $\beta\gamma$  complexes<sup>209</sup> were key to understanding the basal, unbound state of the signaling complex. Crystallization of the activated G $\alpha$  subunits in the presence of various GTP analogues provided the structural snapshots of the final stage of activation<sup>210-212</sup>; likewise structures of G $\alpha$  and G $\beta\gamma$  subunits with their downstream effectors<sup>213</sup> have also led to many interesting discoveries for protein-protein interfaces, signal propagation, and insights into causes for various disease states<sup>214</sup>.

Wall and colleagues solved the first structure of a G-protein heterotrimer, which

provided the structural basis for G $\alpha$ -G $\beta$  interactions<sup>131</sup>. Crystal structures of G $\alpha_i$ / $\beta_1\gamma_1$  and G $\alpha_i$ / $\beta_1\gamma_2$  show two sites of interaction for the G $\alpha$  and G $\beta$  subunits. These include the hydrophobic pocket of the Switch I and II regions as well as a portion of the G $\alpha$   $\alpha$ N helix<sup>131, 208</sup>. However there is no crystallographic evidence of any G $\alpha$ -G $\gamma$  interaction. The  $\alpha$ N helix (amino terminus of the G $\alpha$  subunit) is in close proximity to the carboxyl terminal tail of G $\gamma$ . The  $\alpha$ N helix seems to be unstructured when in the monomeric, GDP-bound state, but it becomes helical when bound to G $\beta\gamma$ . Addition of a lipid modification (myristolation) allows it to maintain a  $\beta\gamma$ -independent helical nature, associate with the membrane and the activated receptor<sup>193, 203</sup>.

Biochemical and mutational studies have further mapped areas around this helix thought to be important for receptor interaction and G protein activation<sup>215-218</sup>. From these studies, the G $\alpha$  subunit's amino terminus and carboxyl terminus seem to be crucial for R\* interaction<sup>215, 219-221</sup>. Several studies have mapped regions of interest for conformational propagation from the binding interface of the receptor to the nucleotide binding site. Of these around the quinone ring<sup>214</sup>, the  $\alpha$ 5 helix<sup>153, 155, 222</sup>, and the G $\beta$  subunit<sup>203, 216</sup> have been identified as important structural moieties for protein interaction and/or function. Likewise for nucleotide exchange and activation of the G $\alpha$  subunit, initial interaction with R\* requires binding of the  $\beta\gamma$  subunits to the G $\alpha$ <sup>172, 204</sup>. The G $\beta\gamma$  subunits have also been shown to provide receptor selectivity with their interaction with G $\alpha$ <sup>30</sup>.



### 3.2 Proposed Mechanisms of Activation

#### *The Lever Arm*

It has been hypothesized that the activated receptor must act “at a distance” to induce nucleotide exchange based on the aforementioned crystallographic knowledge of GPCRs and G proteins. Looking at monomeric G proteins of the Ras or Elongation Factor (EF) subfamilies related to the trimeric complex has provided the foundation of insight into a means of disrupting nucleotide stability in the presence of their exchange factors. Though crystallization of the R\*-G $\alpha_{(\text{empty})}\beta\gamma$  had not been solved in 1998, the 3D structure of bacterial elongation factors Tu and Ts in the empty nucleotide transition state had been solved<sup>223, 224</sup>. From these structures the Bourne lab developed the lever arm model of G $\alpha$  activation based on the Tu/Ts interface and its similarity to the G $\alpha$ -G $\beta$  interface<sup>208</sup>.

From these structural comparisons, they proposed that similar to the three pronged “comprehensive attack” the EF-Ts imbued upon EF-Tu for nucleotide destabilization, the G $\beta$  subunit’s action on the binding pocket may be analogous. This included directly expelling the magnesium ion from the pocket, interrupting the loops around this region, and destabilizing the guanine ring of GDP. The overall structural means of catalysis was through the rotation, or tilting, of the G $\beta$  subunit relative to the G $\alpha$  inter-domain interface. This alteration forces the “lip” occluding the nucleotide pocket to open, creating an exit route<sup>214</sup>.

Similar to the EF-Tu/EF-Ts mechanism, the lever arm hypothesis also included the rotation of loop and secondary structure elements around the binding pocket in a molecular “tug-of-war” for the binding of GTP versus recoupling to the G $\beta$  subunit. The rotation of G $\alpha$ ’s  $\alpha 2$  helix is consistent with the basal and activated G $\alpha$  crystal structures as well as

kinetic studies evaluating the affinity and irreversibility of the creation and termination of the stable ternary complex<sup>225</sup>.

The lever arm model was qualitatively supported using mutant  $G\alpha$ - $G\beta$  with an altered conformation of interaction<sup>226</sup>. These experiments showed that shortening of the  $\alpha$ N helix of  $G\alpha_s$  by four residues near its amino terminus resulted in increased spontaneous nucleotide exchange by  $G\beta\gamma$ ; this was thought to be due to a tilted interface between the subunits which altered the exit route “lip” interaction of  $G\beta$  resulting in increased nucleotide exchange. Likewise further biochemical studies mutating residues along the  $G\alpha$ - $G\beta$  interface showed increased instability of this region leading to altered rates of receptor-mediated nucleotide exchange but not heterotrimer formation<sup>203, 205, 226</sup>. Studies also support the hypothesis of the  $\alpha$ N helix interacting with and moving upon receptor binding<sup>216, 227</sup>.

Though the lever arm model outlines a potential mechanism of information propagation across the complex for activation, it inherently possesses several flaws. Small GTPases do not have the long  $\alpha$ N terminal helix yet still possess the same level or higher rate of nucleotide exchange<sup>193</sup>. Small GTPases also do not require the coordination of a magnesium ion for binding and hydrolysis<sup>228</sup>. Likewise speculation of the thermo-stability of the complex suggested that pulling of  $G\beta$  subunit alone may not be sufficient to force nucleotide exit<sup>229</sup>. Furthermore, the lever arm model did not account for putative  $G\gamma$  interactions with the receptor and  $G\alpha$  subunit<sup>218</sup>. Reports that mutations to the carboxyl terminus of  $G\gamma$  had been shown to increase receptor-mediated nucleotide exchange resulted in revisions to the lever arm model to account for the subunits role in stability, selectivity and activation<sup>218, 230</sup>.

### ***The Gear Shift***

To address some of the issues present in the lever arm model, Chabre and colleagues modified the activation mechanism to still include the G $\beta\gamma$  subunits, but removed the torqueing, or prying, motion of the G $\beta$  subunit<sup>229</sup>. This change was due to the packing differences seen in the Switch II (binding interface of G $\alpha$ -G $\beta$ ) region between GTP- and GDP-bound G $\alpha$  subunits. In the GDP-bound form, Switch II was loose and allowed for a water-filled channel; however it was ordered and densely packed upon GTP binding. Likewise, a disordered interface seemed to be important for GDP instability upon G $\beta$  interaction between the Switch I and II elements. Instead of pulling on the G $\beta$  subunit to pry open the pocket and allow nucleotide displacement, Chabre suggests that G $\beta$  rotates closer to the nucleotide binding region during the exchange. This allows the G $\gamma$  subunit to interact with the helical domain of the G $\alpha$  subunit. The amino terminus of the G $\gamma$  subunit could then alter and shift the helical domain of the G $\alpha$  subunit.

The gear-shift model received its name for the presence of three “gears”. These gears are composed of (1) the activated receptor’s interaction with the carboxyl and amino terminal helices of G $\alpha$  and the carboxyl terminus of G $\gamma$ , (2) the G $\beta$  subunit interacting with the GTP-binding domain of G $\alpha$  in the basal, GDP-bound state, and (3) the G $\gamma$  subunit interacting with the helical domain of G $\alpha$ . These gears cooperate to imbue conformational alterations across the complex to facilitate GDP instability and release. The G $\gamma$  subunit acts as the conformational “shifter” as it is suggested to physically coordinate each of the gears to promote helical domain opening.

A major component of this model is the speculative interaction of the G $\alpha$  helical domain directly with the G $\gamma$  subunit. Several studies have suggested that the helical domain

must open for nucleotide exchange to occur<sup>210</sup>, and G $\alpha$  proteins in which the helical domain was entirely removed possess increased spontaneous nucleotide dissociation<sup>231</sup>. Chabre proposed that the G $\gamma$  subunit hooks the G $\alpha$  helical domain subsequent to G $\alpha$ -G $\beta$  interface tightening. This is not without biochemical validity as several studies suggested G $\alpha$  and G $\gamma$  could interact<sup>218, 230, 232-235</sup> though there was no direct crystallographic or NMR data to support such an interface. In the original description of the gear-shift model, the G $\gamma$  subunit displaces the helical domain as a “rigid body” away from the nucleotide binding region. This model also suggests that the amino terminus of the G $\gamma$  subunit might also play a role in the efficiency of the exchange process by allowing an additional level of specificity with combinatorial subunit composition<sup>232, 233</sup>.

The gear-shift model is similar to the lever arm model in that the G $\beta\gamma$  subunits play a pivotal role in inducing activation. However, unlike the lever arm model, the  $\alpha$ N helix is proposed to move in the opposite direction to allow for a tightening of packing at the G $\alpha$ -G $\beta$  interface<sup>229</sup>. Further support for this model came through observation of similarity between the heterotrimeric G protein  $\beta$  subunits and guanine nucleotide exchange factors associated with the monomeric G protein Arf family (subfamily of the monomeric Ras superfamily). In heterotrimeric G proteins, G $\beta$  has been suggested to play a crucial and active role in GDP exchange for GTP<sup>203, 216</sup>. It has been predicted to alter its relative conformation upon GDP release when the ternary complex is in its transient nucleotide free state<sup>236</sup>. This is similar to the mechanism of activation of small monomeric G proteins and their guanine nucleotide exchange factors (GEFs)<sup>237-240</sup>. For their activation, the GEFs of the Arf family of proteins alter the conformation of the Switch II regions<sup>241, 242</sup>. The G $\beta$  subunit could parallel this mechanism of activation by tightening its interaction during the

free complex<sup>229</sup>.

### ***The G $\alpha$ Unified Model***

Several site-directed spin label studies from the Hubbell and Hamm laboratories investigated the interface between the G $\beta\gamma$  subunits and the G $\alpha$ 's Switch I, II and  $\alpha$ N helix. Though these studies found displacement of the Switch regions and disordered loop conformations that destabilized nucleotide binding, their results did not directly align with either the gear-shift of the lever arm models. Instead these displacements contradicted the larger conformational changes predicted. Likewise, each of the studies suggested that the G $\beta\gamma$  subunits rotated perpendicular to the membrane as opposed to parallel, in contrast with both of the aforementioned models<sup>157, 243, 244</sup>. Though these data may potentially be due to the introduced cysteine mutations or the spin labels introduced into the system, these studies led to the search for a new model of G $\alpha$  activation.

Other experimental findings also conflicted with both models. Primarily these studies investigated the G $\alpha$  carboxyl terminal helix ( $\alpha$ 5). This region of the G $\alpha$  subunit has been shown to interact with activated GPCRs through much mutational, fluorescence, NMR, EPR/DEER and crystallographic structural studies<sup>156, 157, 245-250</sup>. Composed of an  $\alpha$  helix, this secondary structural element has been shown to move as a rigid body, connecting the receptor to the nucleotide binding domain<sup>251</sup>. This conformational change was suggested to be transmitting activation to the binding pocket via G $\alpha$ 's  $\alpha$ 1 and  $\beta$ 2/3 strands<sup>222, 252</sup>. Mutations to this helix and the  $\beta$ 6- $\alpha$ 5 loop led to spontaneous nucleotide release, indicating that this region is important for the signal propagation from the receptor to the GDP binding site<sup>155, 222, 253-255</sup>. Mutational studies and molecular dynamics indicated that the  $\alpha$ 1 helix,  $\beta$ 2,  $\beta$ 3, and  $\beta$ 6 strands may be the midpoints to transmit the conformational change

from the  $\alpha 5$  helix to the helical domain<sup>216, 222, 252</sup>. Still other studies indicated that mutations along  $\alpha 5$  resulted in decreased affinity for GDP and increased spontaneous release<sup>254, 255</sup>. Male patient with precocious puberty were also found to have similar  $\alpha 5$  mutations in their  $G\alpha_s$  proteins<sup>253</sup>. In summation, these reports suggest that the  $G\alpha$  subunit could act as its own “microdomain” for activation independent of  $G\beta\gamma$  movement.

The  $\alpha 5$  helix was also shown to undergo a rotation, insertion and displacement upon receptor interaction, using a five-glycine linker at base of the  $\alpha 5$  helix that “decoupled” the receptor from the nucleotide binding pocket resulting in decreased receptor-mediated nucleotide exchange<sup>156</sup>. This decoupling linker did not drastically alter the intrinsic basal rate of nucleotide exchange. These results suggested that movement of the  $\alpha 5$  helix upon receptor interaction is necessary for propagation of the conformational change from the receptor binding interface to the nucleotide binding pocket<sup>133, 256, 257</sup>. Therefore, the new unified model of G protein activation is founded on the principle that receptor interaction with the  $\alpha 5$  helix is sufficient for nucleotide exchange. Work done by the Hamm, Hubbell, and Meiler laboratories suggest that perturbing this region leads to interaction of the highly conserved TCAT motif of the  $\beta 6$ - $\alpha 5$  loop. This loop directly interacts with guanine ring of GDP allowing for instability in that region<sup>156</sup>.

Lastly, the unified model of  $G\alpha$  activation has been corroborated by recent crystallographic advances in field. In 2012, Kobilka and collaborators solved the nucleotide-free ternary complex in which a ligand-bound GPCR was coupled to its cognate heterotrimeric complex during the transition of GDP for GTP exchange<sup>207</sup>. In this crystal structure, the gear-shift model was shown to be highly unlikely as the  $G\gamma$  subunit was not shown to interact with the  $G\alpha$  subunit. Likewise, further studies on this complex

corroborated the rotation of the  $\alpha 5$  helix upon receptor interaction and potential information propagation to the nucleotide binding pocket<sup>258</sup>.

### 3.3 Conclusions

#### *Conclusions*

##### **The Mechanism of G protein Activation Impacts Drug Discovery**

Understanding the fine-tuned mechanics of the GPCR-G protein system lends itself to more focused drug targets. As GPCRs and G proteins play a “slow” and modulatory role in neuronal signaling, targeting GPCRs through the use of positive and negative allosteric modulators (PAMs/NAMs) has been shown to augment current therapeutic strategies in order to lessen required dosing and/or modulate GPCR function without direct agonism/antagonism. Strong support for the utility of GPCR modulators in attenuating vagrant signaling cascades can be seen in a range of neurological dysfunctions. Research on diseases such as Parkinson’s<sup>259</sup>, addiction<sup>260</sup>, psychosis<sup>261</sup>, Fragile X Syndrome<sup>262</sup>, and memory deficits<sup>263</sup> are all turning to modulators of GPCRs for better medicinal or research compounds.

Roughly 30-40% of approved drugs currently on the market bind extracellularly to G protein coupled receptors<sup>199</sup>. However, understanding GPCR-G protein interactions and their mechanics can lead to novel intracellular medicinal targets. With 800 different GPCRs encoded in the human genome and 21 G $\alpha$  genes<sup>194</sup>, creating small molecules which enhance or inhibit GPCR-G protein interaction may provide a novel means of directing cell signaling in disrupted neural circuits. Enhancing or attenuating receptor-G protein specific intracellular dynamics could shift erroneous receptor conformations or over activated signaling pathways.

Developing interaction-specific GPCR-G protein signaling modulators may also open doors for treatments which synergize with current neurological therapeutics. As many GPCRs couple to multiple G $\alpha$  subtypes, preferentially enhancing or inhibiting one



interaction over the other may lead to significantly less therapeutic off-target effects when dosed congruently with a modulator of that GPCR. In Schizophrenia, modulators of metabotropic Glutamate Receptors II/III (mGluR<sub>2/3</sub>) in Phase II clinical trials have shown promise in attenuating anxiety and hallucinations<sup>264</sup>. Yet with chronic dosing, there are some reports of negative side effects involved with prolonged disruption of the Hypothalamic-Pituitary-Adrenal (HPA) axis. As mGluR<sub>2/3</sub> can couple to several G $\alpha$  subunit subfamilies<sup>265</sup> there is an opportunity to investigate GPCR-G protein specific molecules to alleviate some of the off-target effects on steroid and hormone production. However, without a working understanding of receptor-G protein interaction and activation, knowledge-based creation and testing of such ligands is unattainable.

#### **Future Studies of G protein Structure and Dynamics**

Though the unified model has not fully been tested and is not expected to completely explain G protein activation, it is in consensus with the current literature in the field. Furthermore the unified G $\alpha$  model also allows for other “routes” of information flow across the G $\alpha$  subunit to the nucleotide pocket. For example, the molecular trigger model suggest that a conserved binding site between the  $\alpha 5$  and  $\beta 2/3$  loop (R in DRY motif) is responsible for the G $\alpha$  family’s conformational changes upon R\* binding<sup>266</sup>. This model utilizes a different “door” or direction of information flow across the G $\alpha$  subunit, connecting the receptor to the nucleotide binding pocket across a different face of the G $\alpha$  subunit.

This activation model also opens the field to analyze different questions about GPCR-G protein interactions. Do receptors and G proteins which form pre-coupled complexes utilize different routes of activation? Do GPCR dimers, such as the Class C GPCRs, activate in a similar mode as the monomeric GPCRs? Do different G proteins possess multiple means of activation as an additional level of signaling selectivity? How does the helical domain

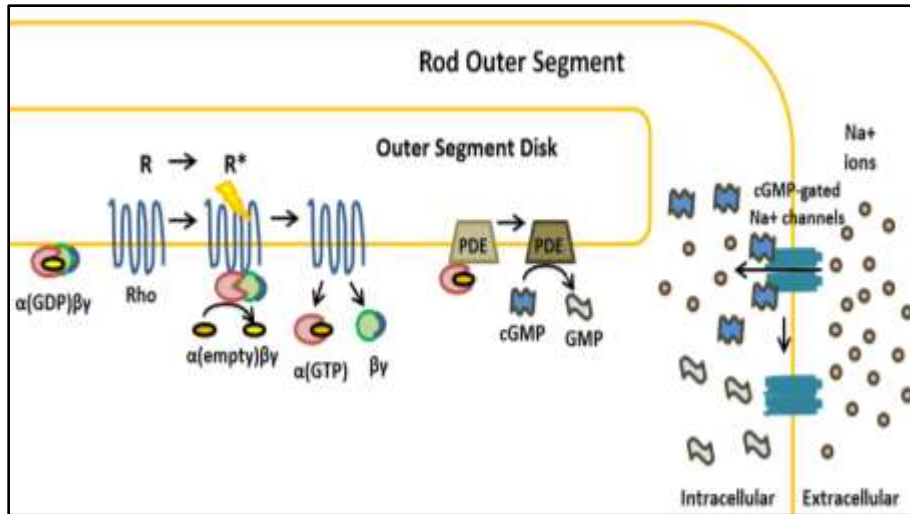
return to a closed conformation upon GTP binding in order to promote complex dissociation? All of these questions will continue to challenge the field of G protein signaling indicating that there is still much to learn about heterotrimeric G proteins and their cognate receptors. Furthermore, elucidation of each of these questions within the neural circuitry will provide researchers with more targeted and specific routes of therapeutic intervention with fewer off-target effects. Understanding GPCR-G protein interaction is paramount to understanding neuronal modulation, signal integration, and aberrant circuitry.

#### ***Abbreviations***

GPCR	- G protein Coupled Receptor
EF	- Elongation Factor
GEF	- Guanine nucleotide Exchange Factor
PAM/NAM	- Positive/Negative Allosteric Modulator
HPA axis	- Hypothalamic Pituitary Adrenal axis

#### ***Acknowledgements***

The author would like to thank Courtney Bricker for her critical edits to the manuscript as well as Professors Bruce Cater, Randy Blakely, Roger Colbran, Seva Gurevich, and Aurelio Galli for their review of the manuscript.



**Figure 3.1: G protein signaling in ROS membranes.** The most studied GPCR, Rhodopsin, becomes activated ( $R^*$ ) by the energy of a single photon.  $R^*$  undergoes a change in structure allowing it to bind to its trimer ( $G\alpha\beta\gamma$ ) and catalyze nucleotide exchange via an unknown mechanism.  $G\alpha(GTP)$  then dissociates to interact with downstream effectors such as Phosphodiesterase (PDE) which hydrolyzes cGMP to GMP. Decreased levels of cGMP leads to the closure of cGMP-gated  $Na^+$  channels. With cations inhibited from entering the photoreceptor, the membrane hyperpolarizes. Adapted from *Neuroscience, Purves, 4<sup>th</sup> Ed.*

## ROSETTA COMPARATIVE MODELING PROTOCOLS

### 4.1 RosettaCM protocol

#### *Chapter 4*

Bender BJ, Cisneros III A, Duran AM, Finn JA, Fu D, Lokits AD, Mueller BK, Sangha AK, Sauer MF, Sevy AM, Sliwoski G, Sheehan JH, DiMaio F, Moretti R, Meiler J, “Molecular modeling with Rosetta3 and RosettaScripts” *Biochemistry* **2016** 55 (34), pp 4748–4763

This section is reprinted with permission from Biochemistry. This is an open access article published under an ACS AuthorChoice License, which permits copying and redistribution of the article or any adaptations for non-commercial purposes.

#### *Contribution*

I am a co-first author of this manuscript. I contributed the texts for the section of comparative modeling, the introduction and parts of the conclusion. I also edited the manuscript and the contributions from the other authors. The protocol capture/workshop tutorial was created by Greg Sliwoski and edited by me. The protocol capture can be found in <sup>267</sup>

#### *Abstract*

Previously, we published an article providing an overview of the Rosetta suite of bio-macromolecular modeling software and a series of step-by-step tutorials<sup>268</sup>. The overwhelming positive response we received to this publication motivates us to here share the next iteration of these tutorials which feature *de novo* folding, comparative modeling, loop construction, protein docking, small molecule docking, and protein design. This updated and expanded set of tutorials is needed, as since 2010 Rosetta has been fully recoded into an object-oriented protein modeling program Rosetta3. Notable improvements include a substantially improved energy function, an XML-like language termed “RosettaScripts” for flexibly specifying modeling task, new analysis tools, the addition of the TopologyBroker to control conformational sampling, and support for multiple

templates in comparative modeling. Rosetta's ability to model systems with symmetric proteins, membrane proteins, non-canonical amino acids and RNA has also been greatly expanded and improved.

### ***Comparative Modeling***

Comparative modeling differs from *de novo* methods in that it utilizes a known protein structure as the starting scaffold or template for structural prediction. If the template structure is a homologous protein, one speaks often of 'homology modeling'. Comparative modeling is a useful strategy for predicting protein structure and function when experimental methods fail or would be too resource-intensive to employ. It increases the probability of obtaining realistic conformational predictions, especially when the target, or desired protein, is greater than 150 amino acids in length and/or adopts a complex tertiary fold. However, it requires that a related, often homologous, structure has been determined experimentally; this is termed the template. Ideally the sequence identity between the target and the template is above 30%, although proteins with lower sequence identity may still be used for comparative modeling when their tertiary fold is conserved.

The latter case will be examined within the tutorial provided with the Supporting Information. This tutorial, **rosetta\_cm**, outlines the basic steps necessary for comparative modeling in Rosetta. The tutorial focuses on the use of RosettaRelax and RosettaMembrane, as well as information on how to implement basic restraints into the system. This tutorial is intended to function as a skeleton protocol for a comparative modeling problem. It does not encompass all possible modifications necessary for application to a real user-defined problem. Over the past several years, comparative modeling in Rosetta has incorporated many improvements, specifically the use of multiple

templates and a specific low-resolution scoring functions<sup>134</sup>. Previously published protocols of comparative modeling with Rosetta suggested using multiple templates to obtain diversity and flexibility<sup>269</sup>. However, models were built on individual templates. The new RosettaCM protocol allows for integration of multiple templates with *de novo* fragments into a single structural model of the protein<sup>134</sup>. Hence, this multi-template, multi-staged protocol samples a broader structural landscape and can select well-scoring sub-templates for different regions of the protein to be modeled.

A highly detailed description of RosettaCM design, sampling and scoring has previously been published<sup>134</sup>. Users are encouraged to refer to this manuscript for a comprehensive assessment of RosettaCM applications, considerations, and caveats. Herein we will briefly describe features of RosettaCM as they apply to the protocol presented.

### **Starting Templates**

Before utilizing RosettaCM, starting templates must be identified through remote homolog detection methods such as PSIBLAST<sup>270</sup>. When homologs are not found using sequence-based methods, 3-D fold recognition software may be used to obtain suitable templates. As with other modeling software, RosettaCM performance improves with higher sequence similarity and identity.

### **Three Stages of Multi-Template Comparative Modeling**

Multi-template RosettaCM is a three-staged process in which the best scoring model from each stage is utilized as the input for the following step. The output of stage one is a full-length, assembled model that is generally correct in topology. However, segment boundaries where templates are mended can be sub-optimal in geometry and energetically frustrated. To resolve these energetic frustrations and to explore the conformational space around this starting model, stage two of RosettaCM iteratively improves local

environments through a series of fragment insertions, side chain rotamer sampling, and gradient-based energy minimization of the entire structure using a RosettaCM-specific low-resolution energy function. The best model from this cycle is then moved to stage three for a final round of all-atom refinement that improves side chain geometries, backbone conformations, and packing density before converging on a final output model.

### **Modeling Loops**

In previous Rosetta comparative modeling protocols a user-defined, “loop” closure step was required to remove chain breaks, reconcile long unstructured coils, or rebuild regions of low sequence similarity (all of which are defined as “loops” within the Rosetta framework). Two different algorithms are available: Cyclic Coordinate Descent (CCD) and Kinematic Loop Closure (KIC). Briefly, CCD quickly closes roughly 99% of loops utilizing a robotics-inspired iterative approach to manipulate dihedral angles of three residue backbone atoms between user-specified C-terminal and N-terminal anchor points. The second loop-building algorithm, KIC, explicitly determines all possible combinations of torsion angles within the defined segment using polynomial resultants<sup>271</sup>. While being slower than CCD, KIC determines more accurate loop structures, provided the anchor points are optimally set. Both algorithms within Rosetta can be used in conjunction with fragments derived from the PDB to build regions of missing electron density, poor homology, or backbone gaps.

Unlike the single template loop building application, comparative modeling with multiple templates closes chain breaks and rebuilds loops internally during stage two. *De novo* fragment insertions are encouraged in regions of weak backbone geometry while template-based fragment insertions anneal chain-breaks and regions of low electron density. Additional smoothing occurs with the RosettaCM-specific scoring function. This

internal step removes the need for additional loop closures by the user. However, it is encouraged for the user to critically examine all output models to validate structural accuracy.

### ***Conclusions***

The Rosetta software suite represents a compilation of computational tools aimed at obtaining physically-relevant structural models of proteins, RNA, and small molecule interactions. Herein we presented a general outline of updated Rosetta applications, protocols, frameworks and functionalities with the aim of improving user success. All protocols are generalizable and can be applied to an extended list of biological queries that other structure-determining methods may not be able address.

Improvements to the variety of Rosetta interfaces (RosettaScript, PyRosetta and many web interfaces) allow the user a high degree of flexibility and personalization for each specific structural problem, as well as providing a previously unavailable entry point for novice users.

The current, default Rosetta score function (*talaris2014*) has been optimized and improved with new score terms as well as reweighted knowledge- and physics-based potentials. Rosetta also incorporates a new release of the Dunbrack rotamer library<sup>272</sup>.

*De novo* structure prediction has greatly improved with the implementation of the TopologyBroker which was developed to create consensus sampling which satisfies all user-requested constraints without requiring additional code development for each unique system. Recent progress in comparative modeling applications have broadened the conformational search space possible by incorporating multiple starting templates. Protocols for protein-protein docking now include flexibility to modularize the coarse-grained and high-resolution modes of RosettaDock, giving the user more freedom to



incorporate additional features in the docking process while narrowing the computational search space. Improvements in protein-small molecule docking utilizes an improved *Transform* algorithm which increases both the speed and quality of this tool in obtaining more native-like conformations. Likewise, the flexibility in incorporating experimentally-derived constraints for most protocols has also greatly improved. In order to tackle the challenge of the inverse folding problem, new implementations of multi-state design permit users to optimize sequences while considering several structures simultaneously.

Continuous developments in Rosetta have increased its utility by adding functionality to model proteins embedded in the membrane, expansion into non-traditional protein modeling by adding non-canonical amino acids, non-canonical backbones, and nucleic acids, as well as adding the ability to model ever-larger proteins by the addition of symmetry.

#### ***Abbreviations***

XML	- eXtensible Markup Language
RosettaCM	- RosettaComparativeModeling
PDB	- Protein Databank
CCD	- Cyclic Coordinate Descent
KIC	- Kinematic Loop Closure

## CHAPTER 5

### A SURVEY OF CONFORMATIONAL AND ENERGETIC CHANGES IN G PROTEIN SIGNALING

#### 5.1 Introduction

##### *Chapter 5*

Lokits AD, Koehler Leman J, Kitko KE, Alexander NS, Hamm HE, Meiler J, “ A survey of conformational and energetic changes in G protein signaling” *AIMS Biophysics* **2015** 2(4): 613-631

This section is reprinted with permission from AIMS Biophysics © 2015, Jens Meiler et al. licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License

##### *Contribution*

I am first author of this manuscript. I contributed all energetics data, ran the analyses, wrote several of the scripts for analysis, created and/or modified most structural models, created all figures and tables, wrote the manuscript, and wrote the protocol capture. I also reviewed all other author contributions and modifications.

##### *Abstract*

Cell signaling is a fundamental process for all living organisms. G protein-coupled receptors (GPCRs) are a large and diverse group of transmembrane receptors which convert extracellular signals into intracellular responses primarily via coupling to heterotrimeric G proteins. In order to integrate the range of very diverse extracellular signals into a message the cell can recognize and respond to, conformational changes occur that rewire the interactions between the receptor and heterotrimer in a specific and coordinated manner. By interrogating the energetics of these interactions within the individual proteins and across protein-protein interfaces, a communication network between amino acids involved in conformational changes for signaling, is created. To construct this mapping of pairwise interactions *in silico*, we analyzed the Rhodopsin GPCR coupled to a  $G\alpha_{i1}\beta_1\gamma_1$  heterotrimer. The structure of this G protein complex was modeled in

the receptor-bound and unbound heterotrimeric states as well as the activated, monomeric  $G\alpha(GTP)$  state. From these tertiary structural models, we computed the average pairwise residue-residue interactions and interface energies across ten models of each state using the ROSETTA modeling software suite. Here we disseminate a comprehensive survey of all critical interactions and create intra-protein network communication maps. These networks represent nodes of interaction necessary for G protein activation.

### ***Introduction***

G protein-coupled receptors (GPCRs) are the largest and most diverse class of membrane receptors in eukaryotes <sup>11</sup>; they bind many different types of ligands to initiate an array of intracellular signaling cascades. GPCRs primarily interact with membrane associated, heterotrimeric complexes called G proteins in order to transduce their extracellular signal into a cellular response. The three subunits,  $G\alpha$ ,  $\beta$ , and  $\gamma$ , undergo conformational changes to interact with different protein binding partners along their signaling cycle in order to transmit the appropriate messages within the cell <sup>11, 228</sup>.

The most dynamic changes in structure and affinity can be seen in the  $G\alpha$  subunit, which mitigates each step of the complex's signaling dynamics and function <sup>154, 210, 244</sup>. The affinity of the  $G\alpha$  subunit to each of its different binding partners is determined by the structural changes it undergoes within the signaling cycle <sup>273</sup>. Therefore one can think of the  $G\alpha$  subunit as the control center of this signal transducing machinery as it preferentially interacts with different proteins, complexes, and small molecules via conformational changes of its own structure to propagate the information to other signaling moieties within the cell (Figure 5.1A).

### **G protein signaling cycle**

In its inactive state, the  $G\alpha$  subunit has a high affinity for the nucleotide GDP,

possesses a closed helical domain, and interacts with the G $\beta\gamma$  subunits. Upon interaction with an activated GPCR, the G $\alpha$  subunit undergoes conformational changes to accommodate binding the receptor (Figure 5.1B)<sup>154, 156, 250</sup>. This includes the rigid body rotation of its  $\alpha 5$  helix up and into the receptor (Figure 5.2), as it moves along the hydrophobic  $\beta$ -sheets surrounding it to create new interactions sites within the GTPase domain and to the helical domain<sup>133</sup>. This rotation signals to the rest of the complex through an altered interaction network that the G $\alpha$  subunit is bound to the receptor. In this receptor-bound conformation, the G $\alpha$  subunit's affinity for GDP is drastically reduced as its flexible helical domain opens to allow nucleotide escape<sup>256</sup>. Upon GDP release, the G $\alpha$  subunit has a high affinity for GTP, though the nucleotides only differ in the addition of a single phosphate group. Once GTP is present in the binding pocket, the G $\alpha$  subunit once again alters its conformation and affinity for both the activated receptor and the G $\beta\gamma$  subunits bound to it. Subsequent dissociation of the G $\alpha$  subunit from this complex frees G $\alpha$ , as well as G $\beta\gamma$ , to interact with downstream signaling effector proteins and regulator molecules in order to continue the signaling cascade<sup>203</sup>. In this GTP-bound, active conformation, the G $\alpha$  subunit possesses different binding interfaces to interact with various effector moieties<sup>210, 211</sup>. The intrinsic enzymatic ability of G $\alpha$  hydrolyzes GTP back to GDP<sup>228</sup>. The rate of hydrolysis can be altered by interacting with accessory proteins which alter the enzyme's catalytic efficiency<sup>193</sup>. Upon cleavage of the  $\gamma$  phosphate group, the G $\alpha$  subunit structure returns to its basal state where its propensity to complex with G $\beta\gamma$  is once again higher than its affinity to interact with other signaling moieties; the reunion of the heterotrimer allows the signaling cycle to terminate or for the complex to begin additional rounds of signaling<sup>11</sup>.

### **Significance**

Current progress in crystallization of GPCRs has greatly aided in our understanding of the G protein's role within the ternary complex model. Recent work from the Kobilka laboratory has provided the first glimpse of an activated GPCR, the  $\beta$ 2-Adrenergic receptor, in complex with a  $G\alpha_s$  heterotrimeric G protein<sup>154</sup>. However, the experimental structure does not provide information on the energetic interactions between amino acids critical for the signaling process. What are the energetic contributions of interactions, broken and newly made, that move the signal from the receptor to the nucleotide binding site? Such an analysis is complicated as the experimental structure presents a static image of interactions in a dynamic system. Crystal structures alone cannot show the conformational dynamics the  $G\alpha$  subunit must continue to undergo to propagate information to the rest of the complex. Further, the use of nanobodies, mutations, and various crystallization aids can alter physiologically relevant conformations of the protein to achieve the most energetically stable interactions for crystal formation.

To better understand the modulatory process the  $G\alpha$  subunit undergoes to propagate its signaling information, an energetic analysis of these conformational changes was performed. We introduce a new pairwise, residue-residue assessment of protein side chain and backbone interactions to describe tertiary topology. Using the available crystallographic structures of each conformation the  $G\alpha$  subunit progresses through during different signaling states, we have created interaction network "maps". Specifically, we have chosen to investigate the heterotrimeric G protein  $\alpha$  subunit in its basal, receptor-unbound  $G\alpha_{i1}(GDP)\beta\gamma$  state, the receptor-bound  $R^*-G\alpha_{i1}(\text{empty})\beta\gamma$  state, and the activated, monomeric  $G\alpha_{i1}(GTP)$  state using the protein software suite, ROSETTA (Figure 5.1B).

Understanding the mechanism of cellular signaling is a crucial step in understanding

the biology of any living organism. This article analyzes changes in conformational and structural information by evaluating the predicted energy of interactions required to maintain function of the G $\alpha$  subunit before, during, and after binding with the membrane-bound receptor. The ROSETTA protein modeling software allows interrogation of intra-protein and inter-protein interactions on the amino acid level. Using an established comparative modeling protocol <sup>274, 275</sup> and binding interface analysis <sup>276, 277</sup>, we have created the first comprehensive framework for interrogation of pairwise amino acid interactions across each of the signaling states. This analysis has allowed us to create predictive communication maps between interacting side chain pairs throughout the G $\alpha$  structure as the conformational shifts propagate.

## 5.2 Materials and Methods

### *Models*

To create interaction networks within the different signaling states of the G protein  $\alpha$  subunit, we have combined several methodologies. Using previously published comparative models of the GPCR- $G\alpha_{i1}$  heterotrimeric proteins<sup>133</sup>, we have created an ensemble of structures for both the basal  $G\alpha_{i1}(\text{GDP})\beta\gamma$  and the receptor-bound  $R^*-G\alpha_{i1}(\text{empty})\beta\gamma$  states. Likewise, we have utilized the available crystal structures of activated, monomeric  $G\alpha_{i1}$  for a similar analysis (PDBIDs: 1GIA, 1GIL). Each structure of activated  $G\alpha_{i1}$  was energy-minimized in the presence of its GTP-analogue. To ensure a robust sampling of the backbone and side chain conformational space consistent with low energy, 500 models were created based on a  $G\alpha_i$  crystal structure (PDBID 1GIA). As more extensive sampling with 1000 poses was not shown to greatly increase model quality, generation of 500 models was used for all other structures. This is consistent with the findings of previous protocols<sup>133</sup>. Of these models, the ten lowest scoring models by ROSETTA score were shown to cover the spread of structural flexibility without allowing for larger structural deviations (Supplemental Figure 5.1). These ten models were employed for further analysis. For all analyses herein, each model possessed the appropriate nucleotide for the given signaling state during all calculations.

### *44G*

From these initial models we then probed for intra- and intermolecular interaction

energies using the ROSETTA computer modeling software suite. Three signaling states of the G protein  $\alpha$  subunit were addressed:  $G\alpha_{i1}(\text{GDP})\beta\gamma$ ,  $R^*-G\alpha_{i1}(\text{empty})\beta\gamma$ , and  $G\alpha_{i1}(\text{GTP})$  (Figure 5.1B). For each state, the binding interface energy ( $\Delta\Delta G$ ) was calculated for various key inter-protein interfaces across the complex and within the GTPase and the helical domains of the  $G\alpha$  subunit. Regions for analysis were selected for their roles as protein-protein interfaces or for their apparent role in maintaining protein stability within each conformational state. Specifically, key secondary structure elements (Figure 5.2) were evaluated for their ability to contribute to overall protein stability by calculating the changes in free energy before and after removal from the structure. Note that all energies are given in ROSETTA Energy Units (REUs) and include predicted contributions of van der Waals interactions, desolvation effects, hydrogen bonds, and electrostatics. While the ROSETTA-predicted energy has been shown to correlate with the free energy in kcal/mol<sup>276, 277</sup>, it is important to highlight that inaccuracies in the structural models and simplifications in the ROSETTA energy function lead to deviations between predicted and experimentally observed energies. Furthermore, the internal energy of small molecules is assumed to be unaltered upon binding to the protein; the energy measurements herein reflect energy perturbations induced by the ligand when binding to the protein. All  $\Delta\Delta G$  results are reported as the absolute value of REU scores for consistency with previously published data<sup>133</sup>.

### ***Pairwise interaction score analysis***

Each of the three signaling states of the  $G\alpha_{i1}$  subunit were then interrogated at the amino acid level utilizing ROSETTA's pairwise score breakdown assessment. This feature calculates the interaction score for each possible amino acid pair. Note, that while this



score is also measured in ROSETTA Energy Units (REUs) it is *not* a free energy in the thermodynamic sense. We therefore call these values consistently ‘interaction scores’. However, this analysis allows for intra-molecular probing of information flow across signaling states while creating a network of stabilizing amino acid interactions. A protocol capture for this application has been validated externally and is available for public use within the ROSETTA framework. Herein, we apply this method to the G protein  $\alpha_{i1}$  subunit to highlight the method’s effectiveness in predicting relevant structural nuances. Each of the signaling states of  $G\alpha_{i1}$  were assessed by averaging the per-residue contribution of the top ten lowest scoring models. The appropriate nucleotides and subunits were present throughout all calculations.

#### ***Pairwise interaction score calculation***

Pairwise interaction scores were calculated using the ROSETTA software suite. The per residue score breakdown was calculated on ten comparative models which were created as previously described <sup>133</sup>.

```
/residue_energy_breakdown.linuxgccrelease -database /rosetta/main/database/ -in:files:s  
<list individual pdbs> -output:prefix <output file name> -  
restore_pre_talaris_2013_behavior
```

#### **Protocols for pairwise interaction score analysis**

Average per residue interaction pairs were calculated across ten models per signaling state in MATLAB using the following script:

```
file_1 = 'model_1.xlsm';  
[~,~, raw_1] = xlsread(file_1);  
model_1 = zeros(5223,3);  
model_1(:,1) = cell2mat(raw_1(1:end,3));  
model_1(:,2) = cell2mat(raw_1(1:end,5));  
model_1(:,3) = cell2mat(raw_1(1:end,26));  
new_matrix_1 = nan(354,354);  
for i = 1:size(model_1)  
    new_matrix_1(model_1(i,1),model_1(i,2))= model_1(i,3);  
end
```

Continued for all models analyzed, then average scores across all models:

```
ave_matrix = nan(354,354);  
for ii = 1:354  
  for jj = 1:354  
    ave_matrix(ii,jj) = mean([new_matrix_1(ii,jj) new_matrix_2(ii,jj) new_matrix_3(ii,jj)  
etc.]);  
  end  
end  
g = ave_matrix(~isnan(ave_matrix));  
[i,j] = ind2sub(size(ave_matrix), find(~isnan(ave_matrix)));  
fin = [i,j,g];
```

### 5.3 Results

#### *Estimating free energy changes across protein-protein interfaces*

Predicting free energy changes across protein-protein interfaces has been a staple in understanding the dynamics and kinetics of protein-protein interaction<sup>276, 278, 279</sup>. Used as a measure of binding efficiency,  $\Delta\Delta G$  estimates are a useful means of probing the thermodynamic stability of a protein interface in the bound and unbound states<sup>133, 279, 280</sup>. For our purposes, we utilized this measure to assess the energetic contribution secondary structure elements possessed along intra-protein interfaces between the helical and GTPase domains as well as for inter-domain stability.

For our calculations, specific secondary structure elements (Figure 5.2) were evaluated for their ability to contribute to overall protein stability by calculating the changes in free energy before and after their removal from the subunit structure. For all calculations, the appropriate nucleotides were present. The top ten lowest scoring models for the  $G\alpha_{i1}(\text{GDP})\beta\gamma$ ,  $R^*-G\alpha_{i1}(\text{empty})\beta\gamma$ , and  $G\alpha_{i1}(\text{GTP})$  states were each assessed, and their ROSETTA scores were averaged.

#### *GDP vs. GTP-bound models*

The  $G\alpha$  subunit possesses similar energy in both the basal,  $G\alpha_{i1}(\text{GDP})\beta\gamma$ , and activated,  $G\alpha_{i1}(\text{GTP})$ , states. This is expected as the two states differ only in the addition of a  $\gamma$ -phosphate ion. Though the  $G\beta\gamma$  subunits were present for the basal calculations of the trimer, they do not significantly alter  $G\alpha$ 's energetics when evaluating regions such as the  $\alpha 1$ ,  $\alpha 5$ , or  $\alpha F$  helices (Figures 5.3–5.4, and Supplemental Tables 5.1–5.3). When evaluating these regions, the resulting energies highlight a consistency between these two states suggesting that any structural changes within these regions begin and end with similar

energies of interaction.

Noteworthy alterations in energies are seen around the nucleotide binding pocket and residues involved in stabilizing the G $\beta$  interface between the basal and activated states. Examination of the P-loop and the variable Switch regions (I–III) (Figure 5.2) indicate more subtle  $\Delta\Delta G$  changes across these regions (Supplemental Tables 5.4–5.7). In the basal, trimeric state, the G $\beta$  subunit organizes the loop regions into a binding interface. In its absence, the activated monomeric models do not show significant changes as seen in ROSETTA energy scores overall, though specific amino acid positions are reported to modulate.

#### ***Receptor-induced conformational changes***

In contrast to the basal and activated states, the R\*-G $\alpha_{i1}$ (empty) $\beta\gamma$  models show a stark transition in the communication network across the secondary structure elements of the G $\alpha$  subunit. During this phase of the signaling cycle, the G $\alpha$  subunit undergoes a large conformational change which can be seen in the shifting of energetics around the  $\alpha 1$  helix, the  $\alpha 5$  helix (Figures 5.3–5.4, Supplemental Tables 5.1–5.2), and regions involved with nucleotide stability, namely the P-loop and the variable Switch (I–III) regions (Supplemental Table 5.4–5.7). It is during this stage of the signaling cycle that the receptor induces activation, the helical domain is opened, and the guanine nucleotide is allowed to exchange. The results from our models are consistent with experimental studies of these structural changes<sup>133, 207, 256</sup>.

#### ***Predicting pairwise residue-residue contributions to protein stability***

To interrogate the conformational changes that must occur at the amino acid network level between the signaling states, we devised a new application for the Rosetta modeling software's per-residue assessment of predicted interactions (publically available); this

application allowed us to evaluate individual amino acid contributions to stability and function. For each of the three signaling states, the top ten models were assessed for each amino acid pair contribution to stability. The average score across the ten models was then plotted for each state (Figures 5.5–5.7, Supplemental Figures 5.2–5.8).

To evaluate which interactions were made and broken between the different signaling states, we compared the basal, heterotrimeric  $G\alpha_{i1}(GDP)\beta_1\gamma_1$  model scores to the receptor-bound,  $R^* G\alpha_{i1}(\text{empty})\beta_1\gamma_1$  and to the monomeric,  $G\alpha_{i1}(GTP)$  active state (Figure 5.5). From this calculation we show the variability of the Switch regions, as interactions are lost, or are diminished during receptor binding (red, above the diagonal) and remade in the active state (blue, below the diagonal). Some of this variability may be due to the loss of  $\beta_1\gamma_1$  binding upon activation.

The predicted opened and closed conformations of the helical domain are also recognized when evaluating residue-residue interactions across states. The receptor binding induces structural rearrangements that ultimately lead to helical domain opening<sup>133, 154, 207, 256</sup>; therefore, the upper matrix, above the diagonal, indicates the helical domain must break contacts for activation (red). The basal and active  $G\alpha_{i1}$  subunits possess very similar secondary structure and tertiary fold. Therefore, fewer interactions are lost or diminished between the two states (below the diagonal). New interactions or more favorable interactions (blue) must be made to accommodate the GTP nucleotide and the lack of  $\beta_1\gamma_1$  subunits.

However, this broad representation does not do justice to the nuanced alterations of residue-residue interaction. In addition the overall number of intraprotein  $G\alpha$  interactions is not expected to change across the different signaling states as all secondary structure

elements and the global tertiary fold is maintained. Though there are technically fewer intraprotein contacts when the helical domain is opened during the receptor bound state, these differences are subtracted from interaction scores that are present in the GDP-bound trimer; the result is a change in magnitude from interaction to no interaction that is recorded in this matrix. Additionally changes in magnitude for the pairwise interactions in the range of  $-0.5$  to  $0.5$  REU were removed to highlight more significant contributions.

### ***The Switch Regions***

Across the three signaling states, a pattern of interaction emerges. As seen with the  $\Delta\Delta G$  calculations (Figures 5.3–5.4, Supplemental Tables 5.1–5.8), the basal and activated  $G\alpha$  subunits maintain similar amino acid interactions. However, the  $\gamma$  phosphate group present in the activated  $G\alpha$  monomer leads to shifts in the communication networks of the Switch I-III regions (Supplemental Figures 5.2–5.4). As implied by their name, these regions have been shown to alter their conformation in the presence of GTP instead of GDP in crystal structures<sup>212, 281</sup>.

The largest shift among these three elements is shown in the Switch II region (Supplemental Figure 5.3). This is expected as the Switch II region also interfaces with the  $G\beta$  subunit. Present in this analysis during the trimeric basal and receptor-bound states, the pairwise interaction map of  $G\beta$  with the Switch II region maintains interaction similarity and therefore structural similarity between these two states. Key differences can be seen around the Switch I and  $\beta 2$  elements with which the Switch II region interacts in the GDP-bound state, but not during the receptor bound state.

Alterations in the communication network along the  $G\beta$  subunit itself were not the primary focus of the current study; however, the  $G\beta/G\gamma$  subunits were present for the

analysis of the basal and receptor-bound states. Therefore, they are included as interaction partners along the corresponding interface residues. Interestingly, the G $\beta$  subunit does show an altered conformation network between the receptor-unbound and bound states suggesting some flexibility between the two G protein subunits. This modulation of the G $\alpha$  Switch II region does not seem to show similar intra-protein interaction flexibility within the G $\alpha$  subunit itself, but rather it highlights relevant changes within the heterotrimeric complex which may contribute to the mechanisms of receptor induced activation.

#### ***Rearrangements from nucleotide exchange***

With this detailed analysis, we show that more elements other than the Switch regions possess an altered communication network. Subtle changes in the  $\alpha 1$  helix,  $\alpha 5$  helix,  $\alpha F$  helix and P-loop highlight structural alterations induced by the nucleotide (Figures 5.6–5.7, Supplemental Figures 5.5–5.6). Specifically, the GDP-bound versus GTP-bound G $\alpha$  subunit show altered network intensities along the  $\beta 6$ - $\alpha 5$  loop within the highly conserved TCAT motif (residues 326–329 in Figure 5.2B). Changes in the P-loop are also much more dynamic than we had originally predicted (Supplemental Figure 5.6). Interactions with the residues 147–150 of the  $\alpha D$ - $\alpha E$  loop in the basal state are not recovered in the active state. Likewise Switch II and III, and  $\beta 4$  interact with variable degrees of binding intensity (as defined by ROSETTA Energy Units) with the P-loop suggesting more dynamic structural rearrangements in this region.

#### ***Receptor-induced network changes***

As expected, the receptor-bound heterotrimer possesses an altered interaction network indicative of altered structure. These conformational changes are highlighted in interaction shifts along the  $\alpha 1$  helix and the  $\alpha 5$  helix (Figures 5.6–5.7) as these secondary structure

elements move to make transient connections. Connections are also lost between Switch I and the Switch II/ $\beta$ 3 interface during receptor binding, which are recovered upon  $G\alpha$  activation and dissociation.

On the backside of the  $G\alpha$  subunit, the P-loop, which has also been implicated in nucleotide stability and possible mechanisms of release<sup>133, 169, 252</sup> shows a drastic structural rearrangement and transition during receptor binding (Supplemental Figure 5.6). As observed above, the P-loop possesses a structural alteration resulting in a loss of interaction with the  $\alpha$ D- $\alpha$ E loop that is not present during receptor binding nor is it recovered post-dissociation in the monomeric, active state.

The linker 1 region connecting the helical domain to the GTPase domain via the  $\alpha$ 1 to  $\alpha$ A helices also possesses a shift in conformation (Figure 5.6, Supplemental Figure 5.7). This element was hypothesized to be an important mechanistic feature to allow domain opening for nucleotide escape<sup>169, 256</sup>. However some movement is expected as it does not possess any secondary structure elements.

### ***The Helical Domain as a rigid body***

Interesting secondary structural elements within the helical domain, such as the  $\alpha$ A helix, do not drastically alter their interaction networks across the three signaling states. This suggests that these elements move together while maintaining a similar tertiary fold (Supplemental Table 8, Supplemental Figure 5.8). These results are in agreement with DEER, EPR, NMR and crystallographic data<sup>154, 252, 256</sup>, which suggests the helical domain moves as rigid body away from the nucleotide binding pocket<sup>154, 207, 256</sup>.



## 5.4 Conclusions and Discussion

The heterotrimeric G protein undergoes dynamic changes in its structure and its binding affinity throughout the stages of the signaling cycle. We utilized structural models of these conformational states to analyze the energetic contributions that stabilize intra- and inter-molecular interactions that define these states, specifically within the  $G\alpha_{i1}$  subunit. This new analysis application predicts key amino acids to be nodes within the information network that propagate the signal across the complex upon interaction with the receptor.

Utilizing the ROSETTA software suite, we computed energy values for residue interactions along different binding interfaces. This benchmarked computational technique has been shown to provide useful insight in the following studies<sup>276, 277, 282</sup>. Likewise, ROSETTA was used to compute pairwise interactions between individual amino acids within the  $G\alpha$  subunit of the heterotrimeric G protein. This technique allowed us to compare the predicted thermodynamically stabilizing interactions between the basal, receptor-bound and activated conformations of the  $G\alpha$  subunit. Through this analysis we were able to detect intra-protein differences in amino acid interaction networks important for propagating conformational changes.

In a previous analysis<sup>133</sup> the  $G\alpha_{i1}$  subunit was evaluated in the basal, GDP-bound trimeric state and in the receptor-bound state through the use of  $\Delta\Delta G$  analysis. Our current study expands on this progress by also including the activated monomeric state for comparison of energy contributions made by key secondary structure elements to evaluate critical regions for G protein activation. In addition we have modeled all three signaling states to evaluate changes in residue pair contributions during signaling.

### ***GDP- vs GTP-bound models***

From this analysis, we have highlighted the similarity of the  $G\alpha_{i1}$  GDP- versus GTP-bound structures. By excising specific structural elements, a broad map of protein stability can be painted. Regions important for interfacing with other proteins, such as the  $\alpha 5$  helix and the Switch II domain show the most altered energy changes between these two states (Figure 5.4, Supplemental Table 5.2, 5.6). This is expected as the binding partners contribute to the relative energy of the system and inhibit interface flexibility. Regions not involved in protein-protein interactions or large structural rearrangements, such as the  $\alpha A$  helix (Supplemental Table 8), remain more energetically stable and consistent across the different models in the GDP and GTP-bound  $G\alpha$  subunit. This result is in agreement with other structural studies that suggest the helical domain moves as a rigid body throughout G protein activation <sup>157, 207, 244, 256</sup>.

### ***Receptor-induced activation***

The  $\Delta\Delta G$  calculations serve to highlight the role of key secondary structure elements as well as specific non-structured linker regions in G protein activation via receptor coupling,  $R^*-G\alpha_{i1}(\text{empty})\beta\gamma$ . During this structural transition state in which  $G\alpha$  must undergo a dynamic conformational change, the  $\Delta\Delta G$  analysis shows a shift in interaction partners for the  $\alpha 1$  helix,  $\alpha 5$  helix, and P-loop (Figures 5.3–5.4, Supplemental Tables 5.1–5.2, 5.4). This conformation must therefore propagate from the receptor to the helical domain of the  $G\alpha$  subunit in order to disrupt binding of GDP. Each of these elements has been implicated in the mechanism of nucleotide escape and G protein activation <sup>169, 212, 252, 283</sup>. From this analysis alone, however, no direct conclusions could be made on the order or dynamics of conformational propagation across the subunit.

### ***Residue-Residue changes within the network***

To better address this, a more detailed analysis of the structural differences was performed. The informational network mapping through the per-residue pairwise analysis highlighted subtle changes in G protein side chains induced by the  $\gamma$  phosphate group of the nucleotide. These altered interaction scores are indicative of altered structures which may prove to be important for interaction with downstream signaling and regulator moieties. However, we do not predict that all changes seen between these two states contribute to effector selectivity and interaction, as some of the altered network must be involved in maintaining the stability of the new structure without contributing to function.

Our pairwise analysis provides insight into possible routes of this information flow from the receptor to the nucleotide binding pocket. Through examination of the  $\alpha 1$  helix, the  $\alpha 5$  helix, and the P-loop, extreme displacement of the interaction pairs predicts the importance of these structural elements in allowing nucleotide exchange and G protein activation (Figures 5.6–5.7, Supplemental Figure 5.6). To further test and validate these predictions, additional experiments must be performed to further elucidate the mechanism of G protein activation.

From these analyses, we have created full, downloadable interaction matrices of our results to provide further understanding of G protein structure and modulation (Supplemental Material). By including pairwise score information across several signaling states, we hope this data will prompt new and unique questions on G protein activation and its signaling mechanics through investigation of these interactive communication maps. The values represent averaged relative interaction scores within these protein complexes as derived from comparative modeling based on published crystal structures. Future studies

will be required to investigate the true predictive power of these results *in vitro*.

### ***Method development***

The use of  $\Delta\Delta G$  calculations in evaluating protein-protein interfaces has long been an important application within ROSETTA<sup>133, 276, 277, 282</sup>. We utilized this analysis not only for evaluating changes along known protein-protein interfaces, but also along key secondary structure elements thought to be important for propagating conformational changes across the protein subunit or necessary for stability. By mapping the  $\Delta\Delta G$  of critical structures across multiple models, we were able to compare relative energy contributions as described by the ROSETTA score function for multiple structural snapshots.

One of the primary purposes for the creation of these energy calculations was to apply and validate a new method of interaction analysis available in the ROSETTA modeling software suite. Here we introduce a new methodology for evaluating the communication networks underlying three dimensional protein topology. By evaluating the residue-residue contributions to protein structure, we have created a technique to map interaction partners necessary for structural stability and conformation transmission. The ROSETTA score term for each contributing residue pair provides a roadmap for amino acid interactions necessary for both structure and function. This pairwise analysis also highlights key nodes of information flow when calculated across multiple protein structural states. The protocol utilized herein has been externally validated and made available for academic and public use with the ROSETTA software suite.

### ***Downloadable Communication Maps***

From these analyses, we have created downloadable interaction matrices available as supplementary material. They combine secondary structure stability with individual ROSETTA scores of interactions on a residue-residue level. This novel perspective has

allowed us to begin to probe regionally specific interactions required for GPCR-G protein interaction, residues required to propagate intra-domain conformational changes, and stabilize the basal, receptor-bound, and activated  $G\alpha$  states. The download also possesses general features about the regional selection such as secondary structure elements, relative evolutionary conservation, amino acid composition etc. as specific to the  $G\alpha_{i1}$  subunit sequence.

***Abbreviations***

GDP	- Guanosine diphosphate
GPCR	- G Protein Coupled Receptor
GTP	- Guanosine triphosphate
GTP $\gamma$ S	- Guanosine 5'-[ $\gamma$ -thio]triphosphate
P-loop	- phosphate binding loop
REU	- ROSETTA Energy Units
r.m.s.d	- root mean square deviation
$\Delta\Delta G$	- delta, delta G binding interface energy

## 5.5 Protocol Capture

For further breakdown of all computational methods utilized herein, please refer to the companion Protocol Capture. All *in silico* methods and calculations were graciously verified by an external reviewer, Dr. J. Koehler Leman, Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, MD.

### ***Protocol capture***

All steps are carried out within the `/pairwise_energy_protocol_capture/` folder.

### **I. Preparing Target Proteins/Small Molecules**

**I.1** Prepare input PDB files. This includes removing unnecessary chains, waters, and small molecules. Cleaned PDB files are then renumbered using the `renumber_pdb.py` script. Our protein example, 1GIA begins with residue 34. We renumber the amino acids to ensure this position is maintained.

```
/path/to/Rosetta/tools/protein_tools/scripts/clean_pdb.py
1GIA A
    (output file: 1GIA_A.pdb)
```

```
/ path/to/Rosetta/tools/protein_tools/scripts/renumber_pdb.py
1GIA_A.pdb -n 34 1GIA_renum.pdb
    (output file: 1GIA_renum.pdb)
```

**I.2** Prepare the `.params` file. Target proteins are relaxed to relieve any small clashes induced by crystallization conditions. Small molecules necessary for protein structure/function are described in a `.params` file during the relax protocol.

- i.) Download any relevant small molecules included within the target protein's structure from the Protein Database Save in `.sdf` format. Next add hydrogens to the small molecule and save in `.mol` format. One of the easiest ways to do this is by opening the `sm_molecule.sdf` file in Pymol.

#### Example commands in Pymol:

```
PyMOL> load GSP.sdf, discrete=0
PyMOL> h_add
save as GSP.mol
    (output file: GSP.mol)
```

- ii.) Generate the `.params` file for each small molecule using the `molfile_to_params.py` script:

```
/path/to/Rosetta/main/source/src/python/apps/public/m
olfile_to_params.py GSP.mol -n gsp
(output file: gsp.params)
```

**I.3** Relax protein in presence of small molecules if necessary. Note\* Using the `clean_pdb.py` script will remove all non-protein molecules. Therefore, inclusion of small molecules will require their manual addition back into the target protein file. Note\*\* for production runs 1000 models is suggested. Only 15 will be created herein.

```
/path/to/Rosetta/main/source/bin/relax.default.linuxgccreleas
e -database /path/to/Rosetta/database -s 1GIA_renum.pdb -
extra_res_file gsp.params -nstruct 15
(output files: 1GIA_renum_0001.pdb - 1GIA_renum_0500.pdb)
```

**I.4** Sort relaxed models for top ten scoring poses. These top poses will be used for calculating Pairwise Energies. Top models included in this protocol have been renamed for simplification.

```
grep pose *.pdb | sort -nk 18 | awk 'FS=":"{print$1}' | head
> best_10_1GIA.ls
(output file: best_10_1GIA.ls)
```

**I.5** Calculate Pairwise Amino Acid Scores using Rosetta's `residue_energy_breakdown` tool. Note\* the output file will contain pairwise energy scores for every amino acid and a “onebody” term. For the current protocol, the `onebody` lines can be ignored. Of interest are the pairwise scores for every possible pair of amino acids listed following the `position::onebody` lines. Within these lines, amino acid position 1 will be in column 3. The amino acid position 2 it is being calculated with will be in column 5. The `total_score` between these two moieties will either be within the 18<sup>th</sup> or 21<sup>st</sup> column of the output depending on the Rosetta version used. Within this protocol, column 18 represents the pairwise `total_score`

```
/path/to/Rosetta/main/source/bin/residue_energy_breakdown.lin
uxgccrelease -database /path/to/Rosetta/main/database -
in:file:1 best_10_1GIA.ls -restore_pre_talaris_2013_behavior
(output file: default.out)
```

**I.6** Open the output file of the `residue_energy_breakdown`. Divide the position to position energy results by individual models. Discard output lines which include amino acid positions compared to a `onebody`. Save only lines of amino acid `position::amino acid position` comparisons. Save these files as individual models `.txt` files. Note\* the first

lines for each model will individually compare each amino acid position ( $n$ ) to the onebody. For each model input, the initial  $n$  lines of the output can therefore be deleted. Within each of the position::position lines are tab delimited columns for each of the Rosetta\_energy score terms. Columns 3, 5, and 18 are relevant for this protocol. See provided output examples for reference.

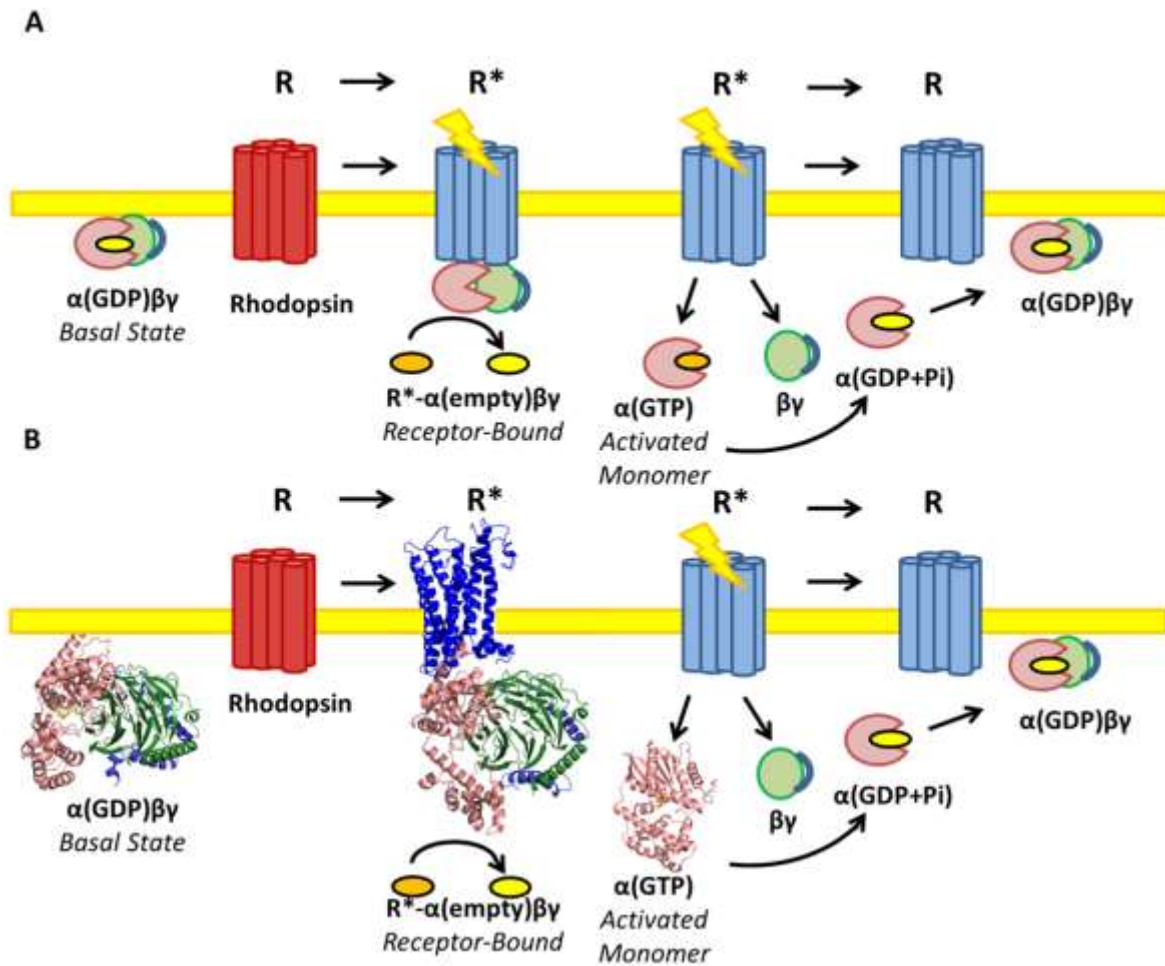
(output files: **1GIA\_model\_1.txt** - **1GIA\_model\_10.txt**)

**I.7** For this protocol, results were visualized in both Excel (tab delimited) and in MATLAB. A MATLAB script (`pos_energy_matrix.m`) was also used to average results from multiple models. For each model, column 3 represents an amino acid position. Column 5 represents all possible amino acid positions the initial residue interacts with. Within the same line, column 18 (or 21 depending on the version of Rosetta) represents the predicted energy of interaction between those two amino acid pairs. The following MATLAB script averages Rosetta's predictive energy across all interacting amino acid pairs. Note\* this script takes Excel files as inputs and extracts columns 3, 5, and 18 (or 21) from multiple models to average the predicted pairwise energies. The final averaged positional pair energy matrix is then formulated into a figure.

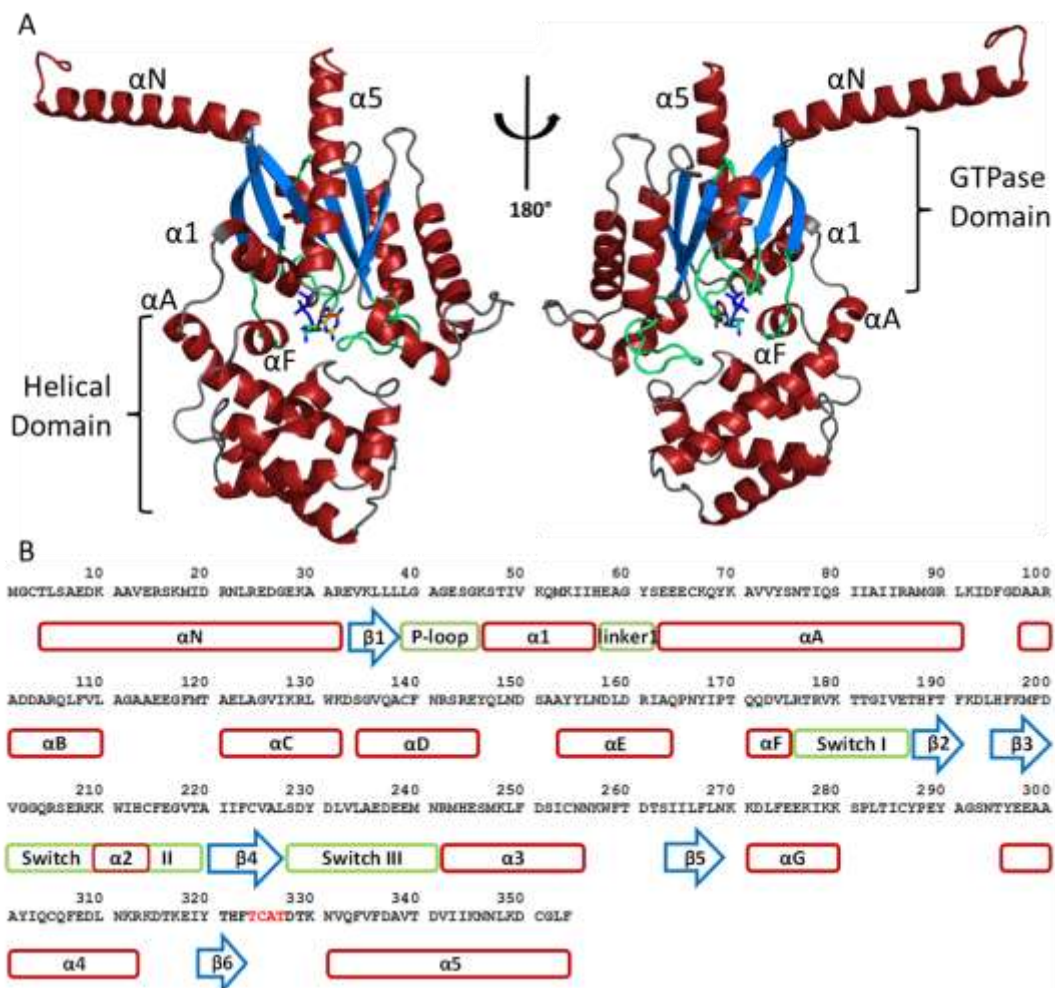
MATLAB Command Line:

`/pairwise_energy_protocol_capture/scripts/pos_energy_matrix.m`  
(output file: **ave\_energy\_1GIA.mat**)

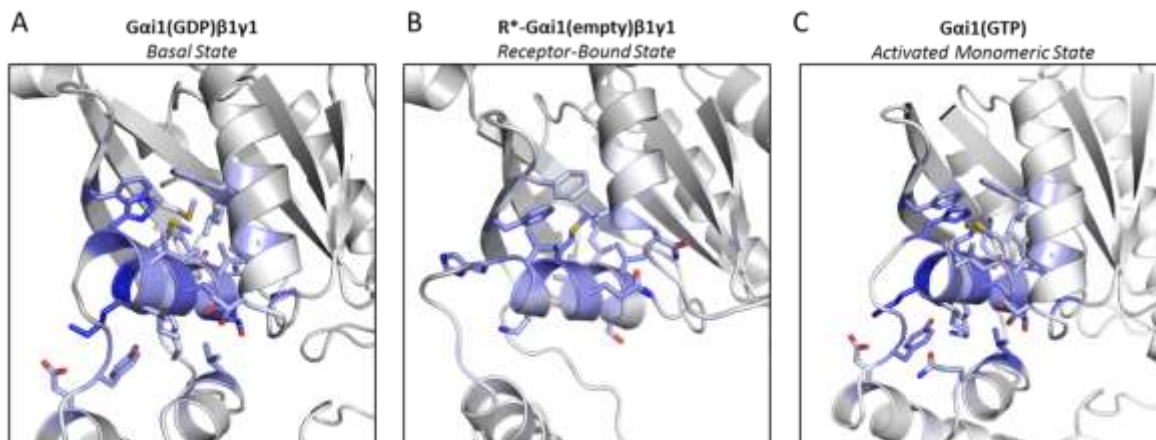




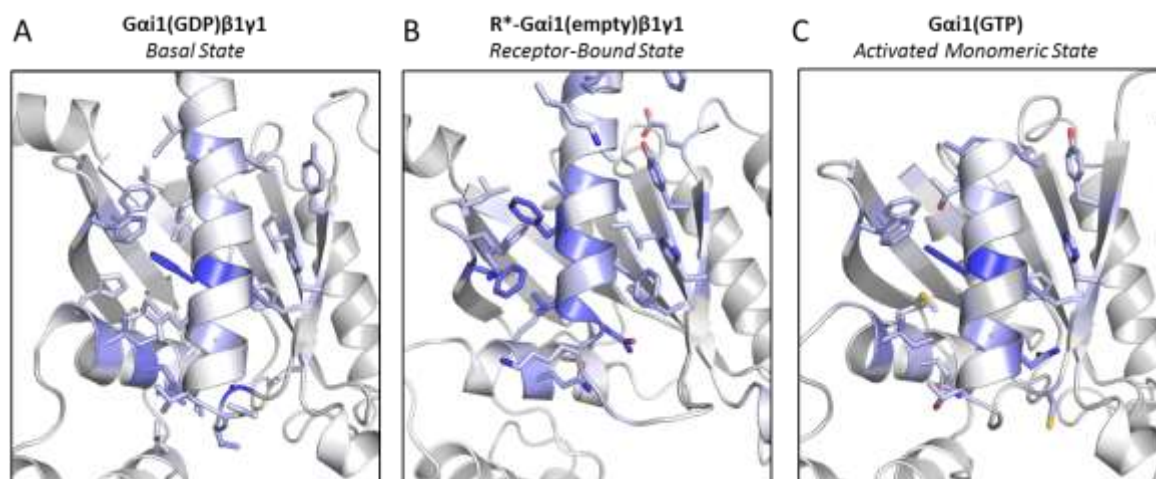
**Figure 5.1.** G protein coupled receptors (GPCRs) typically signal through interaction with membrane-associated heterotrimeric G proteins. These proteins become activated via the exchange of GDP for GTP, induced by the activated receptor ( $\text{R}^*$ ). Upon this nucleotide exchange, the heterotrimer dissociates into the monomer  $\text{G}\alpha(\text{GTP})$  and  $\text{G}\beta\gamma$  dimer which may then interact with downstream signal effector proteins (not shown for clarity) to propagate and amplify intracellular signaling. The cycle is complete when  $\text{G}\alpha$  hydrolyzes GTP to GDP + Pi which allows the trimer to reassemble into the basal, non-signaling state. A) Linear schematic of the G protein signaling cycle. B) ROSETTA-derived structural representations of the three  $\text{G}\alpha$  states examined herein;  $\text{G}\alpha_{i1}(\text{GDP})\beta_1\gamma_1$ ,  $\text{R}^*-\text{G}\alpha_{i1}(\text{empty})\beta_1\gamma_1$ , and  $\text{G}\alpha_{i1}(\text{GTP})$ .



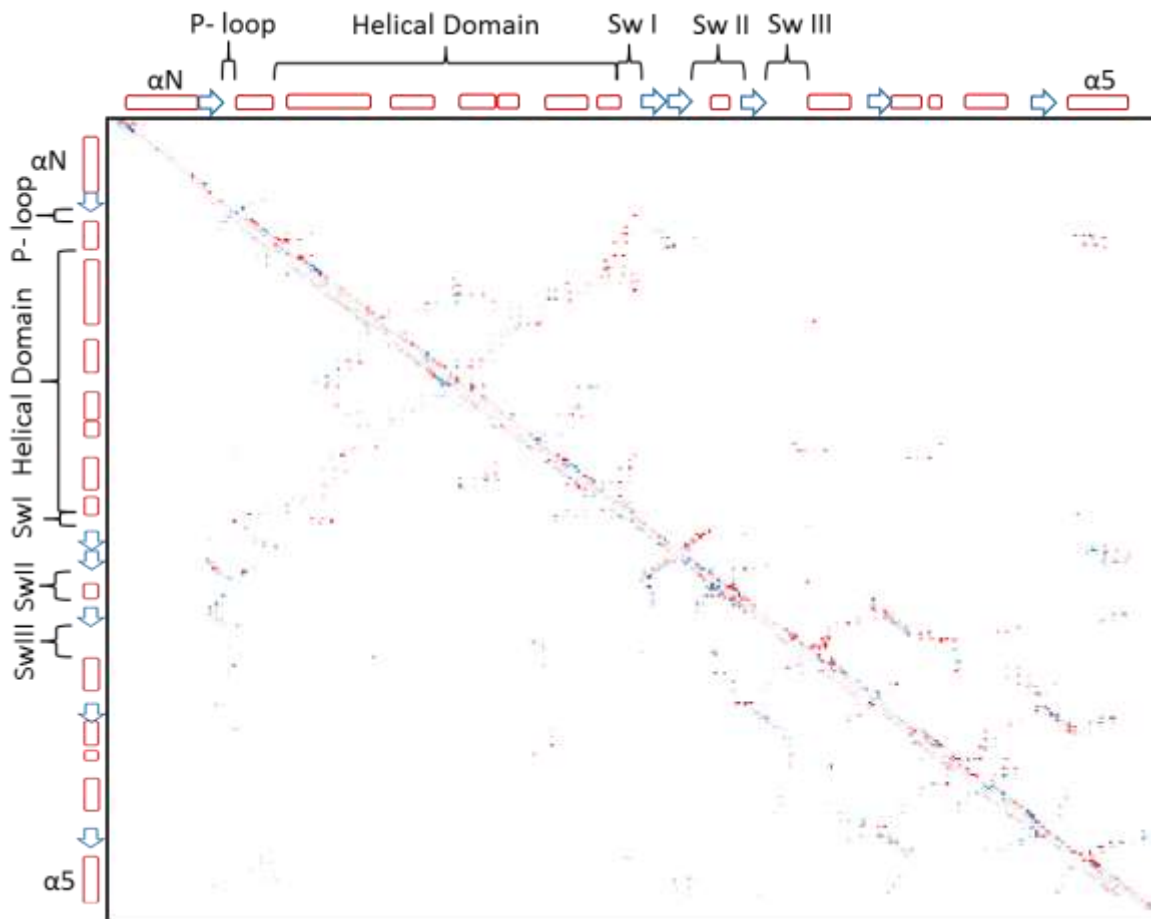
**Figure 5.2.** A) Representative  $G\alpha_{i1}$ (GDP) structure in the basal state ( $\beta_1\gamma_1$  removed for clarity) and rotated 180°. Basal nucleotide (GDP) is depicted in dark blue. B) The sequence is derived from rat  $G\alpha_{i1}$ . Secondary structure elements (red—helices, blue—sheets, and green—critical loop regions) are labeled as described in <sup>210</sup>.



**Figure 5.3. Structural representation of predicted  $\Delta\Delta G$  of the  $\alpha 1$  helix across three states of  $G\alpha$  signaling— $G\alpha_{i1}(\text{GDP})\beta 1\gamma 1$ ,  $R^*\text{-}G\alpha_{i1}(\text{empty})\beta 1\gamma 1$ , and  $G\alpha_{i1}(\text{GTP})$ .** The  $\alpha 1$  helix was defined as residues G45-E58 based on rat  $G\alpha i 1$  sequence. All calculations were averaged across the top ten scoring models for each state. Values reported here represent the absolute values of Rosetta Energy Units (REUs). REUs above 0.5 are considered significant. Residue contributions to the interface are color coded to indicate a greater contribution to stability. Lighter blue indicate a lower REU value relative to the darker shades.

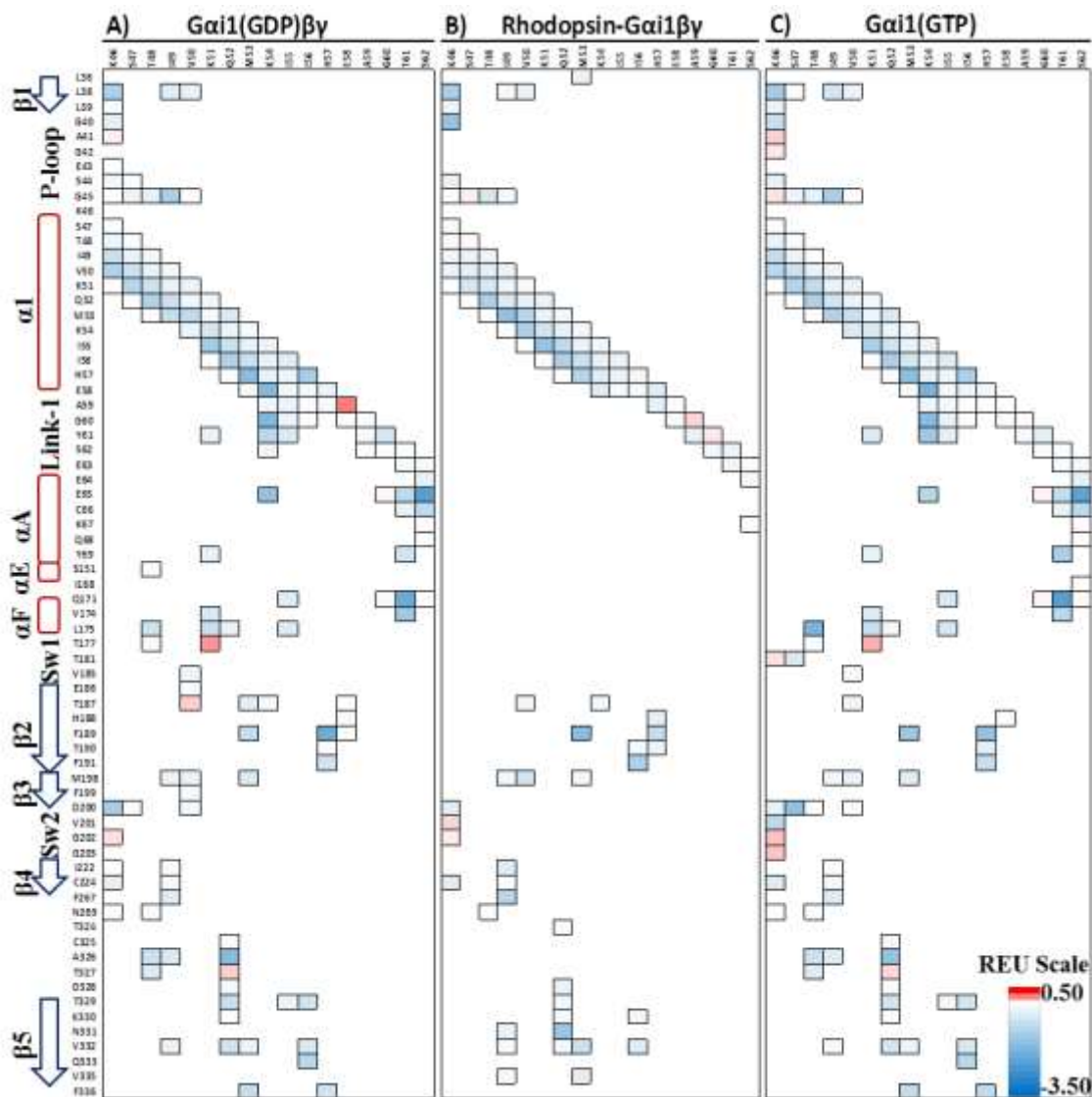


**Figure 5.4. Structural representation of predicted  $\Delta\Delta G$  of the  $\alpha_5$  helix across three states of  $G\alpha$  signaling— $G\alpha_{i1}(\text{GDP})\beta_1\gamma_1$ ,  $R^*\text{-}G\alpha_{i1}(\text{empty})\beta_1\gamma_1$ , and  $G\alpha_{i1}(\text{GTP})$ .** The  $\alpha_5$  helix was extended to include part of the TCAT motif and is defined as residues C325-F354 based on rat  $G\alpha_1$  sequence. All calculations were averaged across the top ten scoring models for each state. Values reported here represent the absolute values of ROSETTA Energy Units (REUs). REUs above 0.5 are considered significant. Residue contributions to the interface are color coded to indicate a greater contribution to stability. Lighter blue indicate a lower REU value relative to the darker shades. \*Note:  $G\alpha_{i1}(\text{GTP})$  crystal structure only extends to residue I343 preventing analysis of the  $\alpha_5$  helix beyond this residue in the activated, monomeric state.

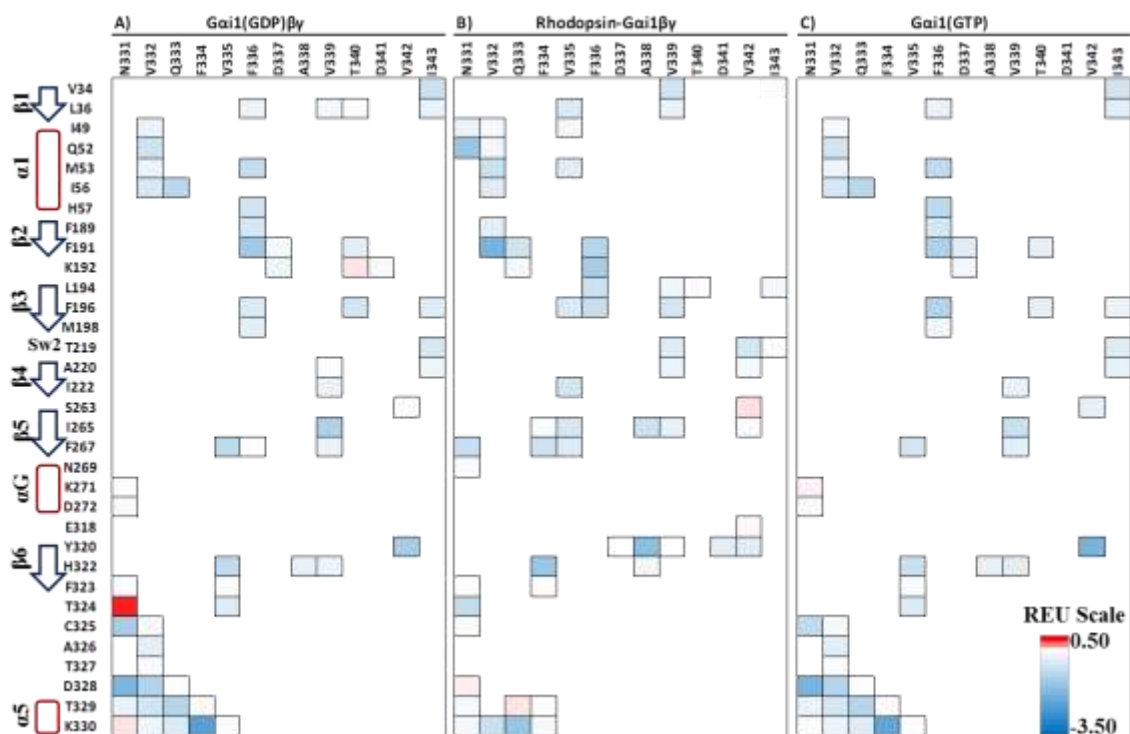


**Figure 5.5. Pairwise analysis of ROSETTA scores for individual amino acid interactions as a means to monitor information networks across signaling states.** These interaction networks represent side chain and backbone atom contributions to the stability and functionality of the protein structure. These matrices compare across the protein signaling states to investigate predicted interaction (and therefore structural) changes between residue pairs. The x- and y-axes represent each residue position of the  $G\alpha$  subunit compared across all other possible residue positions. Above the diagonal depicts the score difference (in REU) between the basal, heterotrimeric  $G\alpha_{i1}(GDP)\beta_1\gamma_1$  structure and the receptor-bound complex,  $R^*G\alpha_{i1}(\text{empty})\beta_1\gamma_1$ . The lower matrix below the diagonal depicts the score difference (in REU) between the basal, heterotrimeric  $G\alpha_{i1}(GDP)\beta_1\gamma_1$  structure and the activated  $G\alpha_{i1}(GTP)$  monomeric subunit after dissociation from Rhodopsin and  $\beta_1\gamma_1$ . Only the  $G\alpha$  subunit's amino acid contributions are shown for clarity. Residue-residue interactions in the range of  $-0.5$  to  $0.5$  REU were removed to highlight more significant differences in contributions. Stabilizing residue interactions are depicted in red while a predicted loss of interaction scores are shown in blue. Note\* The crystal structures used for the monomeric  $G\alpha_{i1}(GTP)$  models lack residues 1–34, and 343–354. These residues are therefore removed from the analysis.



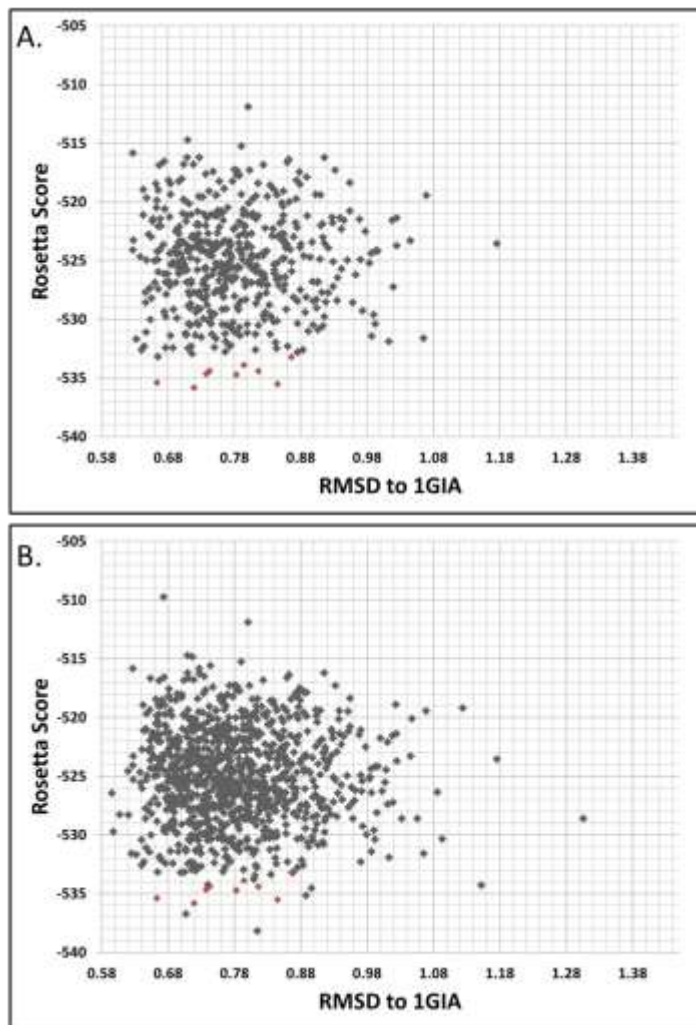


**Figure 5.6. Pairwise analysis of individual amino acid interactions across three  $G\alpha$  signaling states to investigate changes between residue pairs interacting with the  $\alpha_1$  helix and Linker 1 region (K46-I56, & H57-S62, respectively).** A) The heterotrimeric  $G\alpha_i(GDP)\beta_1\gamma_1$  structure in the basal state. B) The heterotrimeric  $G\alpha_i(\text{empty})\beta_1\gamma_1$  subunits interacting with the GPCR, Rhodopsin. C) The activated  $G\alpha_i(GTP)$  monomeric subunit after dissociation from Rhodopsin and  $\beta_1\gamma_1$ . Only the  $G\alpha$  subunit's pairwise amino acid contributions are shown for A–C. REUs for individual pairs are color coded ranging from more stable predicted interaction scores (minimum  $-2.32$ ) in blue to positive, repulsive scores terms (maximum  $0.26$ ) in red.



**Figure 5.7. Pairwise analysis of individual amino acid interactions across three  $G\alpha$  signaling states to investigate changes between residue pairs interacting with the  $\alpha 5$  helix (N331-I343).** A) The heterotrimeric  $G\alpha_{i1}(GDP)\beta_1\gamma_1$  structure in the basal state. B) The heterotrimeric  $G\alpha_{i1}(\text{empty})\beta_1\gamma_1$  subunits interacting with the GPCR, Rhodopsin. C) The activated  $G\alpha_{i1}(GTP)$  monomeric subunit after dissociation from Rhodopsin and  $\beta_1\gamma_1$ . Only the  $G\alpha$  subunit's amino acid contributions are shown for A–C. REUs for individual pairs are color coded ranging from more stable predicted interaction scores (minimum  $-2.30$ ) in blue to positive, repulsive score terms (maximum  $0.31$ ) in red. \*Note:  $G\alpha_{i1}(GTP)$  crystal structure lacks residues 344–354 preventing analysis of the full carboxyl terminus.

## 5.6 Supplemental Information




**Supplemental Figure 5.1: Generation of Models.** To determine the minimum number of models required to represent conformational flexibility of these structural states, an active G $\alpha$ i crystal structure (PDBID 1GIA) was used to generate both 500 and 1000 energy-minimized models using ROSETTA. A) 500 models were generated for 1GIA. Each model was then compared to the crystal structure to determine structural variability by RMSD calculation to C alpha atoms (grey dots). The ten lowest scoring models by Rosetta Energy Units and with lower RMSD values (red dots) covered the conformational variability without allowing for larger structural deviations of the pose. B) Generation of 1000 models (grey dots) for 1GIA show similar a sampling pattern to the 500 model protocol. The ten lowest scoring models from the initial 500 model screen are shown in red. Though two models were produced with lower energy scores, the computational time, resources used, and the degree of energy improvement between these models showed that 500 models represented parameters for the minimum number of models required to generate conformational flexibility.



Alpha 1 Interface (resi 45-58)					Alpha 1 Interface (resi 45-58)				
region	residue	$\alpha(\text{GDP})\beta\gamma$	$\text{R}^*\alpha\beta\gamma$	$\alpha(\text{GTP})$	region	residue	$\alpha(\text{GDP})\beta\gamma$	$\text{R}^*\alpha\beta\gamma$	$\alpha(\text{GTP})$
$\beta 1$	L38	0.9	0.8	1.0	linker 1	Y61			1.4
P loop	G40		0.7		$\alpha A$	E65	0.8		0.6
$\alpha 1$	K46	1.1	1.6	1.9	$\alpha F$	Q171			0.6
$\alpha 1$	S47	1.0	0.7	1.1	$\alpha F$	L175	1.1		1.7
$\alpha 1$	T48	1.8		1.4	$\beta 2$	F189	1.4	1.4	1.5
$\alpha 1$	I49	1.0	1.1	1.0	$\beta 2$	F191		0.6	
$\alpha 1$	V50		0.8		$\beta 3$	M198	0.5		
$\alpha 1$	K51	0.8	0.5	0.9	$\beta 3$	D200	0.8		1.0
$\alpha 1$	Q52	1.7	1.1	1.4	TCAT	A326	1.6		1.5
$\alpha 1$	M53	1.3	1.6	1.4	$\alpha 5$	T329	0.8		0.7
$\alpha 1$	K54	2.5	0.9	2.4	$\alpha 5$	N331		1.0	
$\alpha 1$	I55	1.0		1.0	$\alpha 5$	V332	0.9	0.7	0.8
$\alpha 1$	I56	1.1	1.2	1.1	$\alpha 5$	Q333			0.5
$\alpha 1$	H57	1.7	1.2	1.8	$\alpha 5$	F336	0.7		0.9
linker 1	G60			0.9					

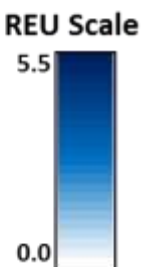
**Supplemental Table 5.1: Predicted  $\Delta\Delta G$  of the  $\alpha 1$  helix across three states of  $G\alpha$  signaling -  $G\alpha_{ii}(\text{GDP})\beta_1\gamma_1$ ,  $\text{R}^*-G\alpha_{ii}(\text{empty})\beta_1\gamma_1$ , and  $G\alpha_{ii}(\text{GTP})$ .** The  $\alpha 1$  helix was defined as residues G45-E58 based on rat  $G\alpha_{i1}$  sequence. All calculations were averaged across the lowest ten scoring models for each state. Values reported here represent the absolute values of ROSETTA Energy Units (REUs). REUs above 0.5 are considered significant. Residue contributions to the interface are color coded to indicate a greater contribution to stability. Lighter blue indicate a lower REU value relative to the darker shades.

Alpha 5 Interface (resi 325-354)					Alpha 5 Interface (resi 325-354)				
region	residue	$\alpha(\text{GDP})\beta\gamma$	$R^*\alpha\beta\gamma$	$\alpha(\text{GTP})$	region	residue	$\alpha(\text{GDP})\beta\gamma$	$R^*\alpha\beta\gamma$	$\alpha(\text{GTP})$
$\alpha 1$	T48	0.8			$\alpha 5$	V332	0.9	1.8	1.0
$\alpha 1$	Q52	1.5	1.0	0.7	$\alpha 5$	Q333			
$\alpha 1$	M53	0.5	0.6	0.5	$\alpha 5$	G334		1.0	
$\alpha 1$	I56	1.1		1.1	$\alpha 5$	V335	1.1	1.6	0.9
$\beta 2$	F191	0.9	1.8	0.9	$\alpha 5$	F336	2.2	1.9	2.5
$\beta 2$ - $\beta 3$	K192		0.8		$\alpha 5$	A338		1.2	
$\beta 3$	L194		0.5		$\alpha 5$	V339	1.0	1.4	1.0
$\beta 3$	F196	0.8	1.0	0.8	$\alpha 5$	V342	0.6	0.5	1.2
$\beta 5$	I265	0.6	0.8		$\alpha 5$	I343	1.1		1.1
$\beta 5$	F267	0.6	0.9		$\alpha 5$	I344		0.9	
$\alpha 4$ - $\beta 6$	E318		0.8		$\alpha 5$	N347		0.8	
$\beta 6$	Y320	0.6	1.2	1.0	$\alpha 5$	L348		0.8	
$\beta 6$	H322	0.7	0.8	0.8	$\alpha 5$	K349		0.7	
TCAT	A326	2.4			$\alpha 5$	D350		1.8	
TCAT	T327	1.1			$\alpha 5$	C351		0.6	
$\alpha 5$	T329	0.8		0.7	$\alpha 5$	L353		1.4	
$\alpha 5$	K330		0.6		$\alpha 5$	F354		1.3	
$\alpha 5$	N331		1.7						




**Supplemental Table 5.2: Predicted  $\Delta\Delta G$  of the  $\alpha 5$  helix across three states of  $G\alpha$  signaling -  $G\alpha_{i1}(\text{GDP})\beta_1\gamma_1$ ,  $R^*G\alpha_{i1}(\text{empty})\beta_1\gamma_1$ , and  $G\alpha_{i1}(\text{GTP})$ .** The  $\alpha 5$  helix was extended to include part of the TCAT motif and is defined as residues C325-F354 based on rat  $G\alpha i1$  sequence. All calculations were performed and presented as described in *Supplemental Table 5.1*. \*Note:  $G\alpha_{i1}(\text{GTP})$  crystal structure only extends to residue I343 preventing analysis of the  $\alpha 5$  helix beyond this residue in the activated, monomeric state.

AlphaF Interface Bound (resi 170-175)				
region	residue	$\alpha(\text{GDP})\beta\gamma$	$\text{R}^*\alpha\beta\gamma$	$\alpha(\text{GTP})$
$\alpha 1$	T48			0.8
$\alpha 1$	K51	0.6		0.6
$\alpha 1$	I55			0.5
linker 1	Y61			1.7
$\alpha A$	Y69	0.9	0.8	1.0
$\alpha E$	Y154		0.8	
$\alpha E$	Y155			0.6
$\alpha E$	R161	1.3	0.7	1.5
$\alpha F$	Q171	1.4		1.5
$\alpha F$	Q172		1.1	0.7
$\alpha F$	D173	3.8	2.5	3.8
$\alpha F$	V174	2.5	1.5	2.3



**Supplemental Table 5.3: Predicted  $\Delta\Delta G$  of the  $\alpha F$  helix across three states of  $G\alpha$  signaling -  $G\alpha_{i1}(\text{GDP})\beta_1\gamma_1$ ,  $\text{R}^*\text{-}G\alpha_{i1}(\text{empty})\beta_1\gamma_1$ , and  $G\alpha_{i1}(\text{GTP})$ .** The  $\alpha F$  helix is defined as residues T170-L175 based on rat  $G\alpha_{i1}$  sequence. All calculations were performed and presented as described in *Supplemental Table 5.1*.

P-loop Interface (resi 40-45)				
region	residue	$\alpha(\text{GDP})\beta\gamma$	$\text{R}^*\alpha\beta\gamma$	$\alpha(\text{GTP})$
P loop	A41		0.9	0.7
P loop	G42		0.7	
P loop	E43	1.6	1.2	2.5
P loop	S44	2.3	2.3	2.0
$\alpha 1$	I49	0.6		0.7
Switch I	R178	0.8		
$\beta 4$	A226		0.8	
$\alpha 3$	R242		1.5	1.2
$\beta 5\text{-}\alpha\text{G}$	N269	0.7		0.7




**Supplemental Table 5.4: Predicted  $\Delta\Delta\text{G}$  of the P-loop across three states of  $\text{G}\alpha$  signaling -  $\text{G}\alpha_{i1}(\text{GDP})\beta_1\gamma_1$ ,  $\text{R}^*\text{-G}\alpha_{i1}(\text{empty})\beta_1\gamma_1$ , and  $\text{G}\alpha_{i1}(\text{GTP})$ .** The P-loop is defined as residues G40-G45 based on rat  $\text{G}\alpha_{i1}$  sequence. All calculations were performed and presented as described in *Supplemental Table 5.1*.

Switch 1 Interface (resi 176-184)					Switch 1 Interface (resi 176-184)				
region	residue	$\alpha(\text{GDP})\beta\gamma$	$\text{R}^*\alpha\beta\gamma$	$\alpha(\text{GTP})$	region	residue	$\alpha(\text{GDP})\beta\gamma$	$\text{R}^*\alpha\beta\gamma$	$\alpha(\text{GTP})$
P loop	E43	0.7		1.0	Switch I	G183	1.2	0.6	1.5
$\alpha\text{A}$	V72		0.6		Switch I	I184	3.0		1.5
$\alpha\text{A}$	N76	1.3		0.7	$\beta\text{2}$	V185	2.1		2.2
$\alpha\text{E}$	S151		0.7		$\beta\text{2}$	E186	2.0		1.5
$\alpha\text{E}$	Y154		0.8		$\beta\text{3}$	M198	1.9		2.0
$\alpha\text{F}$	Q172		0.7		$\beta\text{3}$	F199	0.7		0.7
$\alpha\text{F}$	D173		0.5		$\beta\text{3}$	D200	2.2		2.6
Switch I	T177		1.0		Switch II	V201	0.6		0.9
Switch I	R178	2.4	1.0	2.3	Switch II	G202			1.1
Switch I	V179		0.6		Switch II	G203	1.2		
Switch I	T181			0.8	Switch II	E207			1.2
Switch I	T182	1.3			$\alpha\text{2}$	K210			0.6

**Supplemental Table 5.5: Predicted  $\Delta\Delta\text{G}$  of the Switch I region across three states of  $\text{G}\alpha$  signaling -  $\text{G}\alpha_{\text{il}}(\text{GDP})\beta_1\gamma_1$ ,  $\text{R}^*\text{-G}\alpha_{\text{il}}(\text{empty})\beta_1\gamma_1$ , and  $\text{G}\alpha_{\text{il}}(\text{GTP})$ . Switch I is defined as residues R176-I184 based on rat  $\text{G}\alpha_{\text{i1}}$  sequence. All calculations were performed and presented as described in *Supplemental Table 5.1*.**

Switch 2 Interface Unbound (resi 200-220)					Switch 2 Interface Unbound (resi 200-220)				
region	residue	$\alpha(\text{GDP})\beta\gamma$	$R^*\alpha\beta\gamma$	$\alpha(\text{GTP})$	region	residue	$\alpha(\text{GDP})\beta\gamma$	$R^*\alpha\beta\gamma$	$\alpha(\text{GTP})$
$\beta 1$	K35	1.5	1.4	2.4	Switch II	S206	1.2		
$\beta 1$	L36	0.7			Switch II	E207	0.6	0.9	1.3
$\beta 1$	L37	1.7	1.8	1.5	$\alpha 2$	R208			1.6
$\beta 1$	L38	1.1	1.0	1.4	$\alpha 2$	K210	1.7	1.5	0.6
$\beta 1$	L39			1.4	$\alpha 2$	W211	1.6	1.4	2.3
$\alpha 1$	K46	0.6			$\alpha 2$	I212	0.6		
$\alpha 1$	S47			0.8	$\alpha 2$	H213	1.6	2.4	
Switch I	T181			0.6	$\alpha 2$	C214	0.8	0.8	
Switch I	G183	1.0		1.5	$\alpha 2$	F215	2.3	3.0	1.7
Switch I	I184	1.2		1.5	Switch II	E216	1.4	0.9	0.7
$\beta 2$	V185	1.9		2.1	Switch II	G217	0.6		
$\beta 2$	E186			0.6	Switch II	V218	2.3	1.6	2.6
$\beta 3$	K197			0.6	Switch II	T219		1.0	
$\beta 3$	D200	3.6		3.7	Switch III	E236	0.7		0.8
Switch II	V201	0.9	1.8	1.7	$\alpha 3$	E245			1.7
Switch II	G202	0.6	0.5	1.3	$\alpha 3-\beta 5$	W258	2.6	2.5	1.1
Switch II	G203	1.2	0.7	0.9	$\alpha 3-\beta 5$	F259	1.0	0.8	1.0
Switch II	Q204	1.4	1.1	0.7	$\alpha 3-\beta 5$	T262	0.8		0.7
Switch II	R205	2.0	0.6	1.7	$\beta 5$	S263			0.6



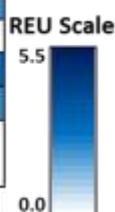
**Supplemental Table 5.6: Predicted  $\Delta\Delta G$  of the Switch II region across three states of  $G\alpha$  signaling -  $G\alpha_{i1}(\text{GDP})\beta_1\gamma_1$ ,  $R^*-G\alpha_{i1}(\text{empty})\beta_1\gamma_1$ , and  $G\alpha_{i1}(\text{GTP})$ .** Switch II is defined as residues N200-A220 based on rat  $G\alpha_{i1}$  sequence. All calculations were performed and presented as described in *Supplemental Table 5.1*.



Switch 3 Interface (resi 226-242)				
region	residue	$\alpha(\text{GDP})\beta\gamma$	$\text{R}^*\alpha\beta\gamma$	$\alpha(\text{GTP})$
P loop	E43		1.0	1.3
P loop	S44		0.8	
$\alpha\text{A}$	R90	0.5		
$\alpha\text{D}-\alpha\text{E}$	R144	0.5		0.8
$\alpha\text{D}-\alpha\text{E}$	Q147	1.2		0.7
Switch II	R205	0.7		1.2
Switch III	L227	4.3		4.2
Switch III	S228			0.8
Switch III	D229	0.8		0.6
Switch III	Y230	3.7		3.3
Switch III	D231	1.1		1.5
Switch III	V233	0.7		0.6
Switch III	E236	0.8		0.9

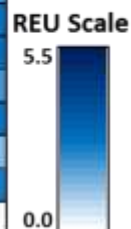
  

Switch 3 Interface (resi 226-242)				
region	residue	$\alpha(\text{GDP})\beta\gamma$	$\text{R}^*\alpha\beta\gamma$	$\alpha(\text{GTP})$
Switch III	E238	0.6		
Switch III	N241	2.4		2.3
$\alpha\text{3}$	H244	1.1	1.4	0.9
$\alpha\text{3}$	E245	1.3		2.1
$\alpha\text{3}$	S246	0.6	0.7	
$\beta\text{5}$	L268	0.6	0.8	0.7
$\beta\text{5}-\alpha\text{G}$	N269	1.2	0.9	1.2
$\beta\text{5}-\alpha\text{G}$	K270		1.0	
$\alpha\text{G}$	F274	1.6	0.9	1.8
$\alpha\text{G}-\alpha\text{4}$	S281	1.5		1.2
$\alpha\text{G}-\alpha\text{4}$	L283	0.7		
$\alpha\text{4}$	I303	0.6		

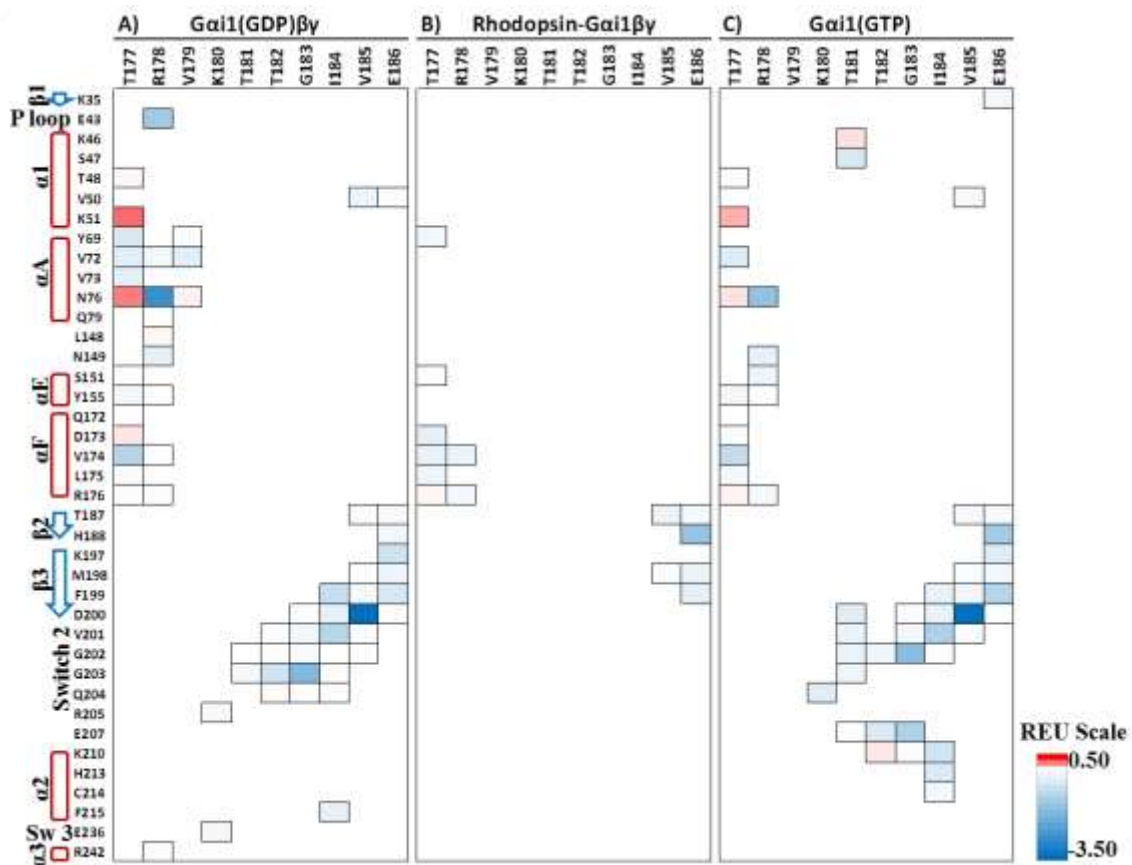
**Supplemental Table 5.7: Predicted  $\Delta\Delta\text{G}$  of the Switch III region across three states of  $\text{G}\alpha$  signaling -  $\text{G}\alpha_{\text{ii}}(\text{GDP})\beta_1\gamma_1$ ,  $\text{R}^*\text{-G}\alpha_{\text{ii}}(\text{empty})\beta_1\gamma_1$ , and  $\text{G}\alpha_{\text{ii}}(\text{GTP})$ .** Switch III is defined as residues A226-R242 based on rat  $\text{G}\alpha_{\text{i1}}$  sequence. All calculations were performed and presented as described in *Supplemental Table 5.1*.

Alpha A Interface (resi 62-91)					Alpha A Interface (resi 62-91)				
region	residue	$\alpha(\text{GDP})\beta\gamma$	$R^*\alpha\beta\gamma$	$\alpha(\text{GTP})$	region	residue	$\alpha(\text{GDP})\beta\gamma$	$R^*\alpha\beta\gamma$	$\alpha(\text{GTP})$
$\alpha 1$	K54	1.0		0.6	$\alpha B$	L107	1.4	1.2	0.9
$\alpha A$	E63			0.6	$\alpha B$	F108	1.1	0.7	0.6
$\alpha A$	E64		0.7		$\alpha B$	A111			0.5
$\alpha A$	E65	1.6	0.8	1.2	$\alpha B-\alpha C$	A114			0.6
$\alpha A$	C66	1.1		1.0	$\alpha B-\alpha C$	G117	0.8	0.6	
$\alpha A$	K67	0.7		0.8	$\alpha B-\alpha C$	F118		1.5	
$\alpha A$	Y69	1.8	1.8	2.0	$\alpha B-\alpha C$	M119	1.1		1.2
$\alpha A$	K70	1.1	0.7	1.1	$\alpha C$	I127	1.0	0.7	1.0
$\alpha A$	V72		0.6		$\alpha C$	L130	0.8	0.7	0.6
$\alpha A$	V73	1.8	1.2	1.3	$\alpha D$	V136	1.0	0.8	0.8
$\alpha A$	Y74	2.3	1.9	3.0	$\alpha D$	C139		0.6	
$\alpha A$	N76	1.7	1.1	1.2	$\alpha D$	R142		0.6	
$\alpha A$	T77	1.1	0.7	0.9	$\alpha D-\alpha E$	Y146	1.4	1.2	1.2
$\alpha A$	I78	1.0	1.3	1.1	$\alpha D-\alpha E$	L148	0.8	0.8	0.8
$\alpha A$	Q79		0.6		$\alpha D-\alpha E$	N149	1.4	0.8	0.7
$\alpha A$	S80	1.4	1.0	0.9	$\alpha E$	S151		0.8	
$\alpha A$	I81	2.3	1.9	2.1	$\alpha E$	Y155	1.4	1.0	1.2
$\alpha A$	I82	1.1	1.0	0.9	$\alpha E$	L156	1.1	0.6	0.9
$\alpha A$	A83	0.6	0.6		$\alpha E$	I162	1.2	1.1	1.0
$\alpha A$	I84	1.7	1.7	1.8	$\alpha E-\alpha F$	Y167	1.1	0.8	0.8
$\alpha A$	I85	1.9	1.6	1.1	$\alpha E-\alpha F$	I168	1.0		1.1
$\alpha A$	R86		0.7		$\alpha E-\alpha F$	P169	1.3	0.8	1.3
$\alpha A$	A87	0.5	0.6		$\alpha F$	D173	0.9		0.8
$\alpha A$	M88	3.2	2.3	2.9	$\alpha F$	V174	1.1	0.7	1.1
$\alpha A$	R90	0.7			Switch I	T177		0.6	0.5
$\alpha A-\alpha B$	I93	1.7	1.3		Switch I	R178	1.6		1.0
$\alpha A-\alpha B$	F95	1.1	0.7	1.0	Switch III	E238	0.5		

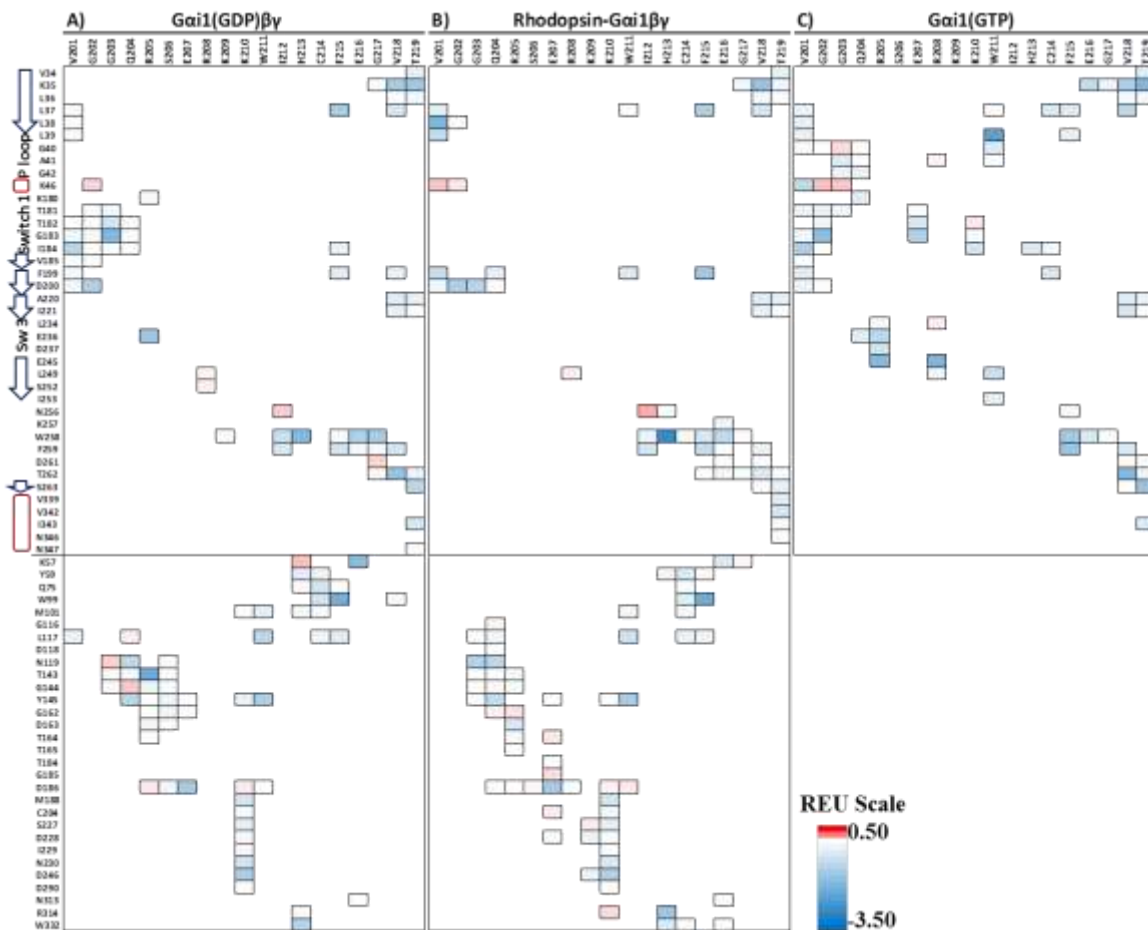


**Supplemental Table 5.8: Predicted  $\Delta\Delta G$  of the  $\alpha A$  helix across three states of  $G\alpha$  signaling -  $G\alpha_{ii}(\text{GDP})\beta_1\gamma_1$ ,  $R^*-G\alpha_{ii}(\text{empty})\beta_1\gamma_1$ , and  $G\alpha_{ii}(\text{GTP})$ .** The  $\alpha A$  helix is defined as residues S62-L91 based on rat  $G\alpha_{i1}$  sequence. All calculations were performed and presented as described in Supplemental Table 5.1.

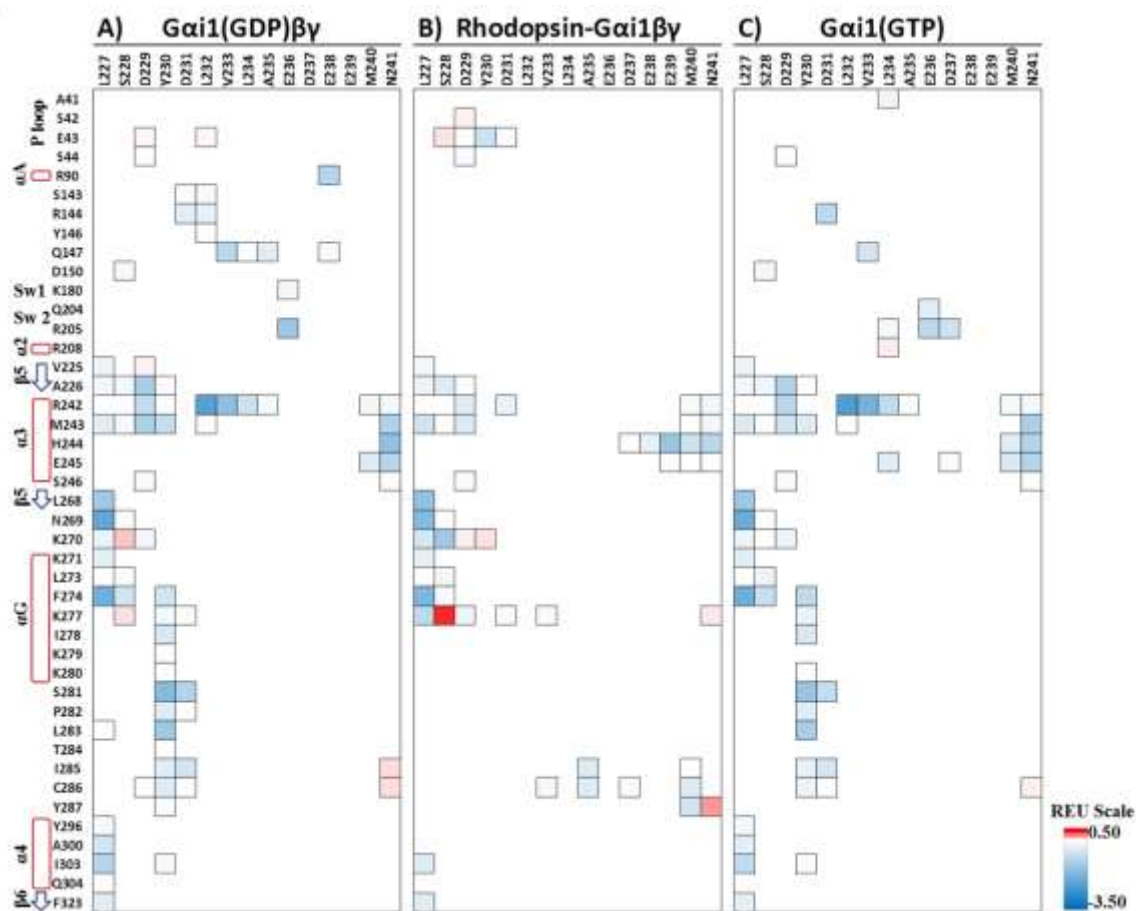




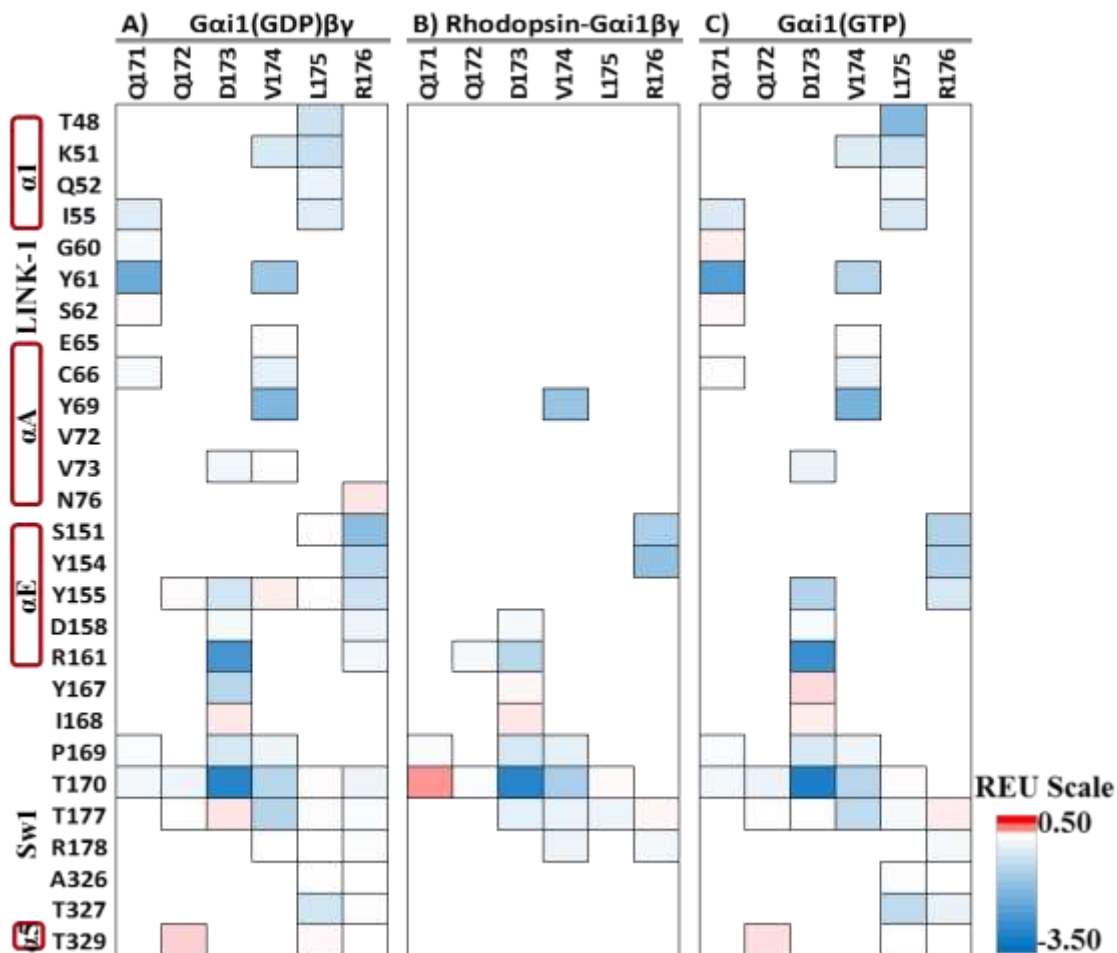
**Supplemental Figure 5.2: Pairwise Analysis of  $G\alpha_{i1}$ : The Switch I Region** Pairwise analysis of individual amino acid interactions across three  $G\alpha_{i1}$  signaling states to investigate energy changes between residue pairs interacting with the Switch I region (T177-E186). A) The heterotrimeric  $G\alpha_{i1}(\text{GDP})\beta_1\gamma_1$  structure in the basal state. B) The heterotrimeric  $G\alpha_{i1}(\text{empty})\beta_1\gamma_1$  subunits interacting with the GPCR, Rhodopsin. C) The activated  $G\alpha_{i1}(\text{GTP})$  monomeric subunit after dissociation from Rhodopsin and  $\beta_1\gamma_1$ . Only the  $G\alpha$  subunit's amino acid contributions are shown for A-C. REUs for individual pairs are color coded ranging from more stable predicted energy scores (*minimum* -4.03) in blue to positive, clashing energy terms (*maximum* 0.21) in red.



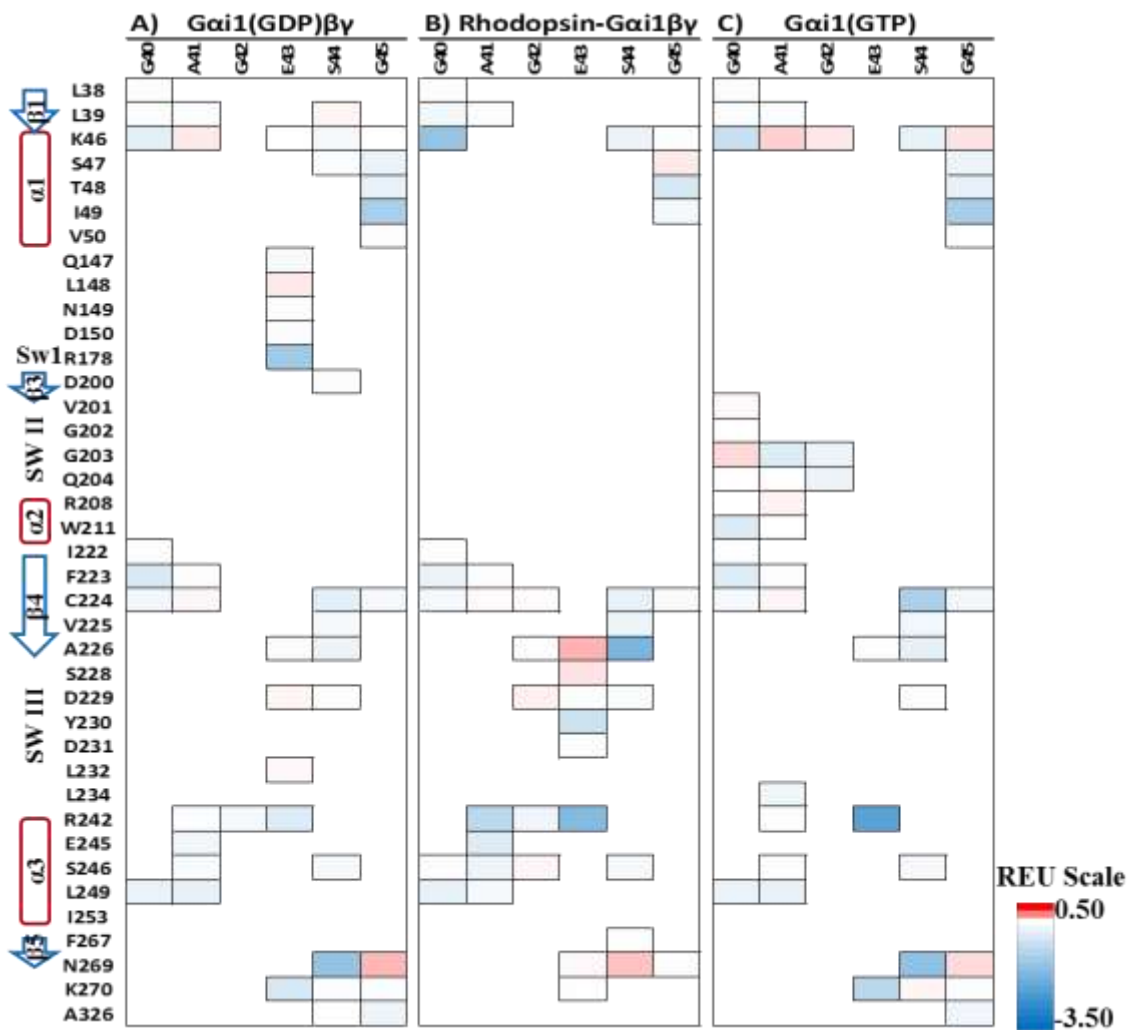
**Supplemental Figure 5.3: Pairwise Analysis of  $G\alpha_{i1}$ : The Switch II Region** Pairwise analysis of individual amino acid interactions across three  $G\alpha$  signaling states to investigate energy changes between residue pairs interacting with the Switch II region (V201-T219). A) The heterotrimeric  $G\alpha_{i1}(\text{GDP})\beta_1\gamma_1$  structure in the basal state. B) The heterotrimeric  $G\alpha_{i1}(\text{empty})\beta_1\gamma_1$  subunits interacting with the GPCR, Rhodopsin. C) The activated  $G\alpha_{i1}(\text{GTP})$  monomeric subunit after dissociation from Rhodopsin and  $\beta_1\gamma_1$ . The  $G\alpha$  amino acid contributions are shown for A-C. A&B also include  $G\beta$  interactions with the  $G\alpha$  subunit. REUs for individual pairs are color coded ranging from more stable predicted energy scores (*minimum* -2.79) in blue to positive, clashing energy terms (*maximum* 0.12) in red.



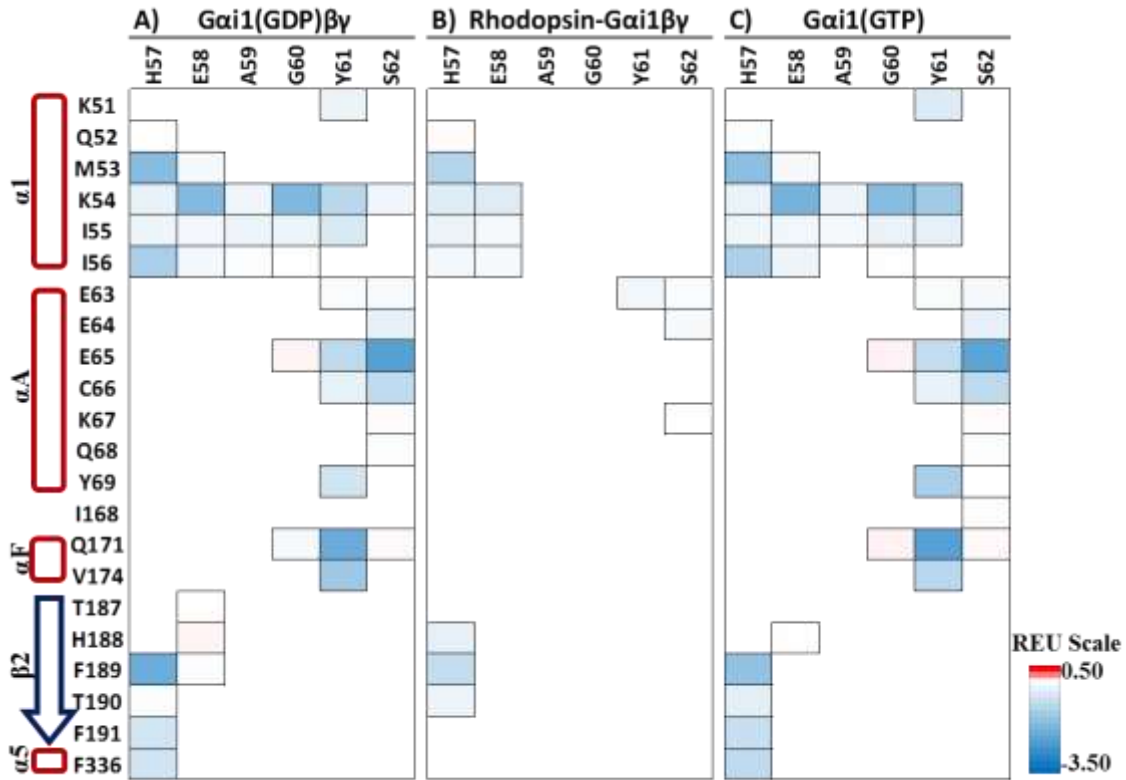
**Supplemental Figure 5.4: Pairwise Analysis of  $G\alpha_{i1}$ : The Switch III Region** Pairwise analysis of individual amino acid interactions across three  $G\alpha$  signaling states to investigate energy changes between residue pairs interacting with the Switch III region (L227-N241). A) The heterotrimeric  $G\alpha_{i1}(GDP)\beta_1\gamma_1$  structure in the basal state. B) The heterotrimeric  $G\alpha_{i1}(\text{empty})\beta_1\gamma_1$  subunits interacting with the GPCR, Rhodopsin. C) The activated  $G\alpha_{i1}(GTP)$  monomeric subunit after dissociation from Rhodopsin and  $\beta_1\gamma_1$ . The  $G\alpha$  amino acid contributions are shown for A-C. REUs for individual pairs are color coded ranging from more stable predicted energy scores (*minimum* -2.39) in blue to positive, clashing energy terms (*maximum* 0.30) in red.



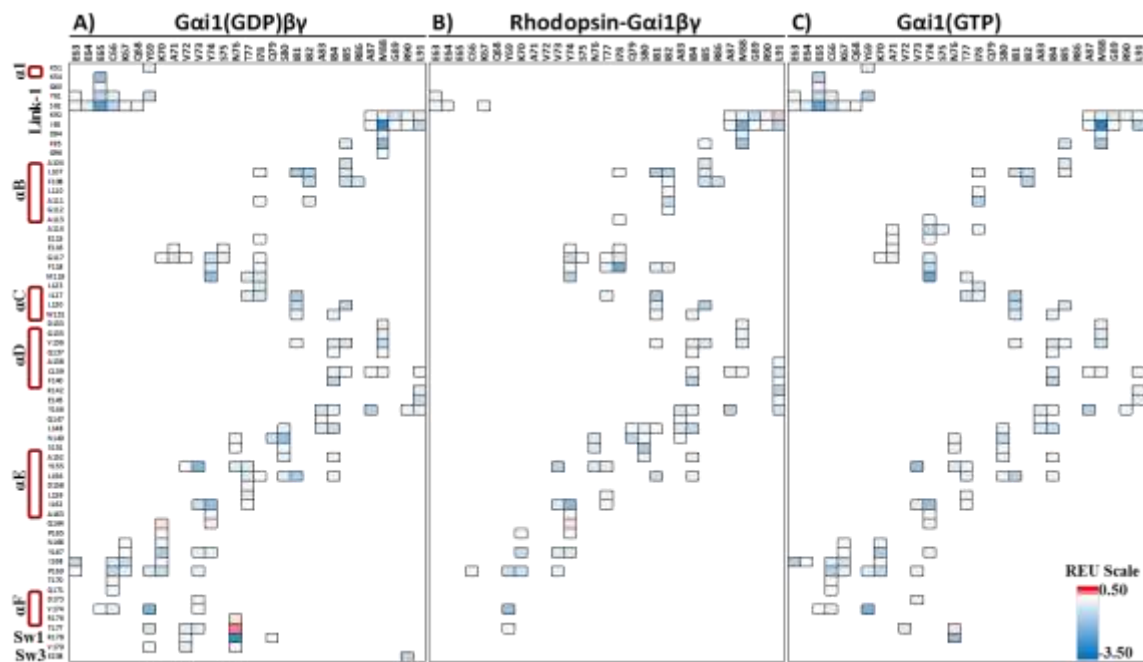
**Supplemental Figure 5.5: Pairwise Analysis of  $G\alpha_{i1}$ : The  $\alpha F$  helix** Pairwise analysis of individual amino acid interactions across three  $G\alpha$  signaling states to investigate energy changes between residue pairs interacting with the  $\alpha F$  helix (Q171-R176). A) The heterotrimeric  $G\alpha_{i1}(\text{GDP})\beta_1\gamma_1$  structure in the basal state. B) The heterotrimeric  $G\alpha_{i1}(\text{empty})\beta_1\gamma_1$  subunits interacting with the GPCR, Rhodopsin. C) The activated  $G\alpha_{i1}(\text{GTP})$  monomeric subunit after dissociation from Rhodopsin and  $\beta_1\gamma_1$ . Only the  $G\alpha$  subunit's amino acid contributions are shown for A-C. REUs for individual pairs are color coded ranging from more stable predicted energy scores (*minimum* -3.11) in blue to positive, clashing energy terms (*maximum* 0.15) in red.



**Supplemental Figure 5.6: Pairwise Analysis of  $G\alpha_{i1}$ : The P-loop** Pairwise analysis of individual amino acid interactions across three  $G\alpha$  signaling states to investigate energy changes between residue pairs interacting with the P-loop (G40-G45). A) The heterotrimeric  $G\alpha_{i1}(\text{GDP})\beta_1\gamma_1$  structure in the basal state. B) The heterotrimeric  $G\alpha_{i1}(\text{empty})\beta_1\gamma_1$  subunits interacting with the GPCR, Rhodopsin. C) The activated  $G\alpha_{i1}(\text{GTP})$  monomeric subunit after dissociation from Rhodopsin and  $\beta_1\gamma_1$ . Only the  $G\alpha$  subunit's amino acid contributions are shown for A-C. REUs for individual pairs are color coded ranging from more stable predicted energy scores (*minimum* -2.24) in blue to positive, clashing energy terms (*maximum* 0.10) in red.



**Supplemental Figure 5.7: Pairwise Analysis of  $G\alpha_{i1}$ : The  $\alpha_1$ - $\alpha_A$  linker** Pairwise analysis of individual amino acid interactions across three  $G\alpha$  signaling states to investigate energy changes between residue pairs interacting with the Linker 1 region (H57-S62). A) The heterotrimeric  $G\alpha_{i1}(\text{GDP})\beta_1\gamma_1$  structure in the basal state. B) The heterotrimeric  $G\alpha_{i1}(\text{empty})\beta_1\gamma_1$  subunits interacting with the GPCR, Rhodopsin. C) The activated  $G\alpha_{i1}(\text{GTP})$  monomeric subunit after dissociation from Rhodopsin and  $\beta_1\gamma_1$ . Only the  $G\alpha$  subunit's amino acid contributions are shown for A-C. REUs for individual pairs are color coded ranging from more stable predicted energy scores (*minimum* -2.32) in blue to positive, clashing energy terms (*maximum* 0.03) in red.



**Supplemental Figure 5.8: Pairwise Analysis of  $G\alpha_{i1}$ : The  $\alpha A$  helix** Pairwise analysis of individual amino acid interactions across three  $G\alpha$  signaling states to investigate energy changes between residue pairs interacting with the  $\alpha A$  helix (E63-L91). A) The heterotrimeric  $G\alpha_{i1}(GDP)\beta_1\gamma_1$  structure in the basal state. B) The heterotrimeric  $G\alpha_{i1}(\text{empty})\beta_1\gamma_1$  subunits interacting with the GPCR, Rhodopsin. C) The activated  $G\alpha_{i1}(GTP)$  monomeric subunit after dissociation from Rhodopsin and  $\beta_1\gamma_1$ . Only the  $G\alpha$  subunit's amino acid contributions are shown for A-C. REUs for individual pairs are color coded ranging from more stable predicted energy scores (*minimum* -3.04) in blue to positive, clashing energy terms (*maximum* 0.18) in red.

## CHAPTER 6

### A CONSERVED PHENYLALANINE AS A RELAY BETWEEN THE $\alpha 5$ HELIX AND THE GDP BINDING REGION OF HETEROTRIMERIC $G_i$ PROTEIN $\alpha$ SUBUNIT

#### 6.1 Introduction

##### *Chapter 6*

This research was originally published in the Journal of Biological Chemistry. Kaya AI, Lokits AD, Gilbert JA, Iverson TM, Meiler J, Hamm HE, “A conserved phenylalanine as relay between the  $\alpha 5$  helix and the GDP binding region of heterotrimeric  $G_i$  protein  $\alpha$  subunit” *Journal of Biological Chemistry* **2014** 29;289(35):24475-87 © the American Society for Biochemistry and Molecular Biology.

##### *Contribution*

I am the second author of this manuscript. I predicted the energetic contributions of several of the amino acids leading to the study of F336. I contributed the analysis and text surrounding this energetic analysis and the analysis of amino acid conservation. I created Table 6.2, and Supplemental Table 6.1. I also created all Supplemental Movies 6.1-6.4. I also reviewed and edited all text from the other authors.

##### *Abstract*

G protein activation by G protein coupled receptors (GPCRs) is one of the critical steps for many cellular signal transduction pathways. Previously, we and other groups reported that the alpha 5 ( $\alpha 5$ ) helix in the G protein alpha subunit plays a major role during this activation process. However, the precise signaling pathway between the  $\alpha 5$  helix and the GDP binding pocket remains elusive. Here, using structural, biochemical and computational techniques, we probed different residues around the  $\alpha 5$  helix for their role in signaling. Our data showed that perturbing the F336 ( $\alpha 5$ ) residue disturbs hydrophobic interactions with the  $\beta 2$ - $\beta 3$  strands and  $\alpha 1$  helix, leading to high basal nucleotide exchange. However, mutations in  $\beta$  strands  $\beta 5$  and  $\beta 6$  do not perturb G protein activation. We have highlighted critical residues that leverage F336 as a relay. Conformational changes are transmitted starting from F336 via  $\beta 2$ - $\beta 3$ / $\alpha 1$  to Switch I and the P-loop, decreasing the stability of the GDP binding pocket and triggering nucleotide release. When the  $\alpha 1$  and  $\alpha 5$



helices were cross-linked, inhibiting the receptor-mediated displacement of the C-terminal  $\alpha 5$  helix, mutation of F336 still leads to high basal exchange rates. This suggests that unlike receptor mediated activation, helix 5 rotation and translocation is not necessary for GDP release from the  $\alpha$  subunit. Rather, destabilization of the backdoor region of the  $G\alpha$  subunit is sufficient for triggering the activation process.

### ***Introduction***

Heterotrimeric G proteins play a critical role as molecular switch proteins that couple the activation of cell surface receptors, G protein coupled receptors (GPCRs), to different intracellular effector proteins mediating intracellular responses. Therefore, G proteins have a crucial role in defining the specificity and temporal characteristics of many different cellular responses <sup>11, 284-287</sup>.

Several structural and biophysical studies have proposed the conformation of the receptor in its active state and have identified potential receptor mediated mechanisms for G protein activation and GDP release <sup>156, 200, 202, 246, 283, 288-293</sup>. Two well-studied receptor mediated G protein activation routes have been hypothesized. In the first, the binding of the GPCR to the C-terminus (CT) of  $G\alpha$  is thought to trigger conformational changes that can be transmitted via rotation of the  $\alpha 5$  helix of  $G\alpha$  to the  $\beta 6$ - $\alpha 5$  turn on the purine ring of the GDP (Figure 6.1) <sup>156, 207, 286, 294, 295</sup>. In the second proposed mechanism, the GPCR is thought to take advantage of  $G\beta\gamma$  as a nucleotide exchange factor in order to disrupt the phosphate interactions of the nucleotide binding pocket via destabilization of switch (SW) I-II regions through perturbing  $\alpha 5$  interaction with the  $\beta 2$ - $\beta 3$  strands (Figure 6.1) <sup>157, 214, 296-299</sup>.

In 2011, Kobilka and colleagues provided an important missing piece of the puzzle in

the receptor mediated G protein activation cycle by determining the structure of the  $\beta_2$ -adrenergic receptor - Gs heterotrimer complex ( $\beta_2$ AR-Gs) structure <sup>295</sup>. This groundbreaking study detailed the receptor - G protein (R-G) interaction and G protein activation. This structure represents the end point in the signal transduction step. The signaling route by which an active receptor interacts with an inactive G protein and causes conformational changes that lead to the final high-affinity complex of a receptor with its cognate G protein and GDP release is still unknown.

To address the conformational dynamics underlying nucleotide release from the  $G\alpha$  subunit, we recently generated a predictive computational model of the energy of receptor activation with the goal of understanding conformational changes and connections between potential key residues during G protein activation <sup>133</sup>. In this model of the rhodopsin -  $G\alpha\beta\gamma$  complex, it was suggested that the  $\alpha 5$  helix is the most critical region for G protein stability and activation, and is consistent with previous studies <sup>156, 291-293, 300</sup>. The  $\alpha 5$  helix is protected and surrounded with primarily hydrophobic interactions within six beta strands ( $\beta 1$ - $\beta 6$ ) and one alpha helix ( $\alpha 1$ ) (Figure 6.1C & D). Energetic analysis predicted that residues F191, F196 in  $\beta 2$ - $\beta 3$ ; I265, F267 in  $\beta 5$ ; Y320, H322 in  $\beta 6$  strands; Q52 and M53 in the  $\alpha 1$  helix are making critical interaction with the  $\alpha 5$  helix in both basal and receptor mediated G protein activation <sup>133</sup>. These key residues might either be important for the overall structural integrity of the GTPase domain during the activation process, or they may be directly involved in activation.

In order to identify the residue-residue interactions that are critical for activation as a part the signaling pathway, we systematically tested the effects of these residue-residue interactions on G protein activation. Residues were examined using biochemical,

computational, and structural approaches in both basal and receptor bound states. In this study, recombinant G $\alpha$ 1 was used for all experiments instead of visual G protein, given that G $\alpha$ i is a very close homolog of G $\alpha$ t yet much more easily expressed in *E.coli*. Our data showed that single mutations in the  $\beta$ 5 and  $\beta$ 6 strands that face the  $\alpha$ 5 helix were not able to break hydrophobic interactions and trigger GDP release from G protein in both receptor bound and unbound states. In the receptor bound state, using pairwise coupling energy analysis, we predicted that the  $\alpha$ 5 rotation compensates the effect of  $\beta$ 5- $\beta$ 6 mutations on protein activation.

However, the hydrophobic interactions on the opposite side of the  $\alpha$ 5 helix were predicted to directly affect G protein function. Energetic analysis predicted that phenylalanine 336 (F336) is the most critical residue in the  $\alpha$ 5 helix; it creates a hydrophobic hotspot of G protein activation, consistent with previous studies<sup>292, 301, 302</sup>. The amplitude of this effect was correlated with decreasing hydrophobicity of the side chain. Experimentally tracing the hydrophobic interactions around the F336 residue together with computational analysis provided evidence for a dynamic interplay between F336, the  $\beta$ 2 and  $\beta$ 3 strands, and the  $\alpha$ 1 helix on the G protein activation route.

## 6.2 Materials and Methods

### ***Materials***

The TSKgel G2000SW column, GDP, and guanosine 5'-O-(3-thiotriphosphate) tetralithium salt (GTP $\gamma$ S) were purchased from Sigma. All other reagents and chemicals were of the highest available purity.

### ***Rosetta interface energy calculations***

Interface energies were computed following the Rosetta  $\Delta\Delta$ G protocol previously described<sup>133</sup>. Briefly: we leveraged the previously published ensembles of ten structures of the G-protein in the basal state and receptor bound state. Residue-residue interactions across  $\alpha$ 1 helix/GTPase domain interface were evaluated by measuring energetic perturbations when computationally removing the  $\alpha$ 1 helices from the models. The  $\alpha$ 1 helix was defined as residues 45 to 58. For all analyses, GDP remained fixed within the nucleotide binding pocket. The  $\Delta\Delta$ G value is reported as an average over the ten structural models in Rosetta Energy Units (REU). Absolute values larger than 0.5 REU are considered to be significant. Using the standard deviation over the ten structures a Z-score was computed. The total  $\Delta\Delta$ G-value across the interface is calculated as the sum of individual residue contributions.

### ***Rosetta pairwise binding energy calculation***

Average energies between pairwise interacting residues were computed using Rosetta's per residue energy-breakdown protocol. The energy between all possible pairs of interacting amino acid residues within the G-protein were calculated across the previously

published ensembles of ten structures<sup>133</sup>. These energies between all residues pairs was then averaged across the ten models in both the receptor bound and basal state. Predicted energy values are reported in Rosetta Energy Units (REU) and considered significant if greater than 0.5 REU.

### ***Preparation of urea washed ROS membranes and Gβ1γ1***

Urea washed ROS membranes and Gβ1γ1 were prepared from bovine retina as described previously<sup>303, 304</sup>.

### ***Construction, expression and purification of proteins***

Briefly, the pSV277 expression vector encoding Gai1 with N-terminal His-tag served as the template for introducing individual mutant substitutions using the QuickChange system (Stratagene). All mutations were confirmed by DNA sequencing (DNA Sequencing Facility, Vanderbilt University). The mutant constructs were then expressed and purified as previously described<sup>305</sup>. The purified proteins were cleaved with thrombin (Sigma, 0.5 U/mg final concentration) for 16 hr at 4 °C in order to remove the N-terminal His-tag. The sample was then loaded onto a Ni-NTA column to separate the protein from the cleaved His-tag and any uncleaved fraction. For further purification, the protein solution was loaded onto size-exclusion column (TSKgel G2000SW) that was equilibrated in buffer A [50 mM Tris-HCl (pH 7.4), 150 mM NaCl, 2 mM MgCl<sub>2</sub>, 40 μM GDP (or 1 μM GTPγS), 2 mM DTT and 100 μM PMSF]. SDS-PAGE was used to test the purity of the proteins. Protein concentrations were determined by Bradford assay<sup>306</sup>.

### ***Nucleotide-exchange assay***

The basal rate of GTPγS binding was determined by monitoring the relative increase in the intrinsic tryptophan (W211) fluorescence (λ<sub>ex</sub> 290 nm, λ<sub>em</sub> 340 nm) of Gai1 (200 nM)

in buffer containing 50 mM Tris (pH 7.2), 100 mM NaCl and different amounts of MgCl<sub>2</sub> for 60 min at 25 °C after the addition of 10 mM GTPγS. Receptor mediated nucleotide exchange was determined with Gβ1γ1 (400 nM) in the presence of 50 nM rhodopsin at 21 °C for 60 min after the addition of GTPγS. The data were normalized to the baseline and maximum fluorescence and then fit to the exponential association equation ( $Y = Y_{\max} * (1 - e^{-kt})$ ), to calculate the rate constant (k) as previously described<sup>156</sup>.

#### ***Intrinsic Trp fluorescence assay with AIF***

Intrinsic tryptophan (W211) fluorescence upon AIF<sub>4</sub><sup>-</sup> activation, relative to emission in the GDP bound state of G protein alpha subunit, was monitored as previously described<sup>307</sup>. Data represent the averages from 6-8 experiments.

#### ***Trypsin digestion and analysis***

2 μg Gαi1 were incubated in buffer containing 50 mM Tris (pH 7.5), 100 mM NaCl, 20 μM GDP and different amounts of MgCl<sub>2</sub> (0.5, 1, 2 mM). 10 mM NaF and 50 μM AlCl<sub>3</sub> were added to samples, then incubated for 2 min at 25 °C. One microliter of a 1 mg/ml TPCK trypsin solution was added and incubated on ice for 25 min. The reaction was stopped by adding 2.5 μl of termination solution (10 mg/ml aprotinin, 10 mM PMSF). Subsequently, samples were boiled with Laemmli sample buffer for 5 min, and run on a 12.5% SDS-polyacrylamide gel, stained with Coomassie Blue and quantified by densitometry (Multimager, Bio-Rad)<sup>303, 308, 309</sup>.

#### ***Cross-Linking***

An expression vector encoding Gαi1 with six amino acid substitutions at solvent exposed cysteines (Gαi1 HI) and an internal His6 tag between residues Met119 and Thr120 served as the template for introducing individual cysteine substitutions using the

QuikChange system (Stratagene) as describe above. The bifunctional cross-linking reagent Bis-maleimidoethane (BMOE, Pierce Biotechnology) was incubated in a 2:1 molar ratio with G $\alpha$ i1 HI as previously described <sup>256</sup>. The concentrated, cross-linked monomeric protein was then purified by size exclusion chromatography on a calibrated G2000SW column. Calibration was performed under the same conditions as purification, using a broad range of molecular weight standards (Biorad) <sup>256</sup>.

### ***Membrane binding assay***

The ability of mutant G $\alpha$  subunits to bind rhodopsin in urea-washed ROS membranes was determined as previously described <sup>156</sup>. Each sample was evaluated by comparison of the amount of G $\alpha$ i1 subunit within the pellet (P) or supernatant (S) to the total amount of G $\alpha$ i1 subunit (P+S) in both treatments expressed as a percentage of the total G $\alpha$ i1 protein. Data represents the average of three experiments.

### ***Protein crystallization, data collection and structure determination***

Purified GDP bound G $\alpha$  subunits were exchanged into crystallization buffer (50 mM EPPS (pH 8.0), 1 mM EDTA, 2 mM MgCl<sub>2</sub>, 5 mM DTT, 1 mM GDP) using a size exclusion chromatography column. Appropriate fractions were pooled as described above and SDS-PAGE was used to assess to test the purity of the proteins. Crystals were grown by the hanging drop vapor diffusion method at 18 °C by equilibration against a reservoir solution containing 2.0-2.3M (NH<sub>4</sub>)<sub>2</sub>SO<sub>3</sub> and 100mM sodium acetate (pH 5.9-6.4). Proteins (10 mg/ml) were mixed 1:2.5 ratio with reservoir solution and crystals appeared after 14-18 days with in the space group I4. A similar strategy was used to grow crystals G $\alpha$ i1-GTP $\gamma$ S proteins. Proteins were incubated with 10  $\mu$ M GTP $\gamma$ S for 30 min on ice and then storage buffer replaced the crystallization solution containing 50  $\mu$ M GTP $\gamma$ S instead of GDP. G $\alpha$ i1-GTP $\gamma$ S samples crystallized in the space group P3<sub>2</sub>21. Crystals were cryo-protected

prior to data collection by briefly soaking in stabilization solution containing 18% Glycerol and 2.4 M  $(\text{NH}_4)_2\text{SO}_3$  for ~30 s and cryo cooled by immersion in liquid nitrogen.

Data sets were collected at the LS-CAT (21-ID-G) of the Advanced Photon Source (APS) at Argonne National Laboratory at -180 °C using a wavelength of 0.98 Å on a MAR CCD detector. Data were processed and scaled using the HKL2000, CCP4 and Phenix suites<sup>310-312</sup>. Crystallographic data processing and refinement statistics are reported in **Table 1**. Criteria for data cutoff were a combination of  $R_{\text{sym}}$  and  $I/\sigma$  which both rose to unacceptable levels if the resolution were extended by  $G_{\alpha}$ . The structures of  $G_{\alpha}1$ -GDP and  $G_{\alpha}1$ -GTP $\gamma$ S complexes were determined by molecular replacement using 1GDD (WT  $G_{\alpha}1$ -GDP)<sup>313</sup> and 1GIA ( $G_{\alpha}1$ -GTP $\gamma$ S·Mg<sup>2+</sup>)<sup>314</sup> as search models for Phaser-MR in the Phenix suite<sup>312</sup>. Since 1GDD and 1GIA preceded the requirement for deposition of structural factors R-free reflections were randomly selected for F336C variant and was the same as F336Y. As a result, the free R is of limited utility. Model building was performed in Coot<sup>315</sup> using composite omit maps calculated in Phenix<sup>312</sup> to minimize model bias. Refinement conducted by both CNS<sup>316</sup> and Phenix, final refinements done by Phenix suite. In the final model, the regions corresponding to amino acids 1-8 and 203-211 in F336C-GDP, and 1-8, 202-217 and 233-240 in F336Y-GDP are not included. Similarly, in the GTP $\gamma$ S bound structures, amino acids 1-32 and 349-354 are not included due to lack of electron density. Structural superpositions were performed using Superpose for the  $C_{\alpha}$  carbon backbone in the CCP4 suite<sup>317, 318</sup>. All structural images were made with PyMOL (PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.) unless otherwise indicated.



### 6.3 Results

In this study, our strategy was to test residues around the  $\alpha 5$  helix that were previously identified as critical for the function of this helix during G protein activation. Residues were examined using biochemical, computational, and structural approaches in both basal and receptor bound state.

#### *The effects of $\beta 5$ - $\beta 6$ strand mutants on G protein activation*

In our previous study, we proposed four residues that face the  $\alpha 5$  helix in  $\beta 5$  (I265, F267) and  $\beta 6$  (Y320, H322) <sup>133</sup>. Any one of these might be critical for  $\alpha 5$  helix stability and therefore the G protein activation (Figure 6.2A) <sup>133</sup>. To test the effect of these residues on G protein function, we evaluated nucleotide exchange rates after introduction of site directed mutations. Basal and receptor mediated nucleotide exchange rates of mutants were determined by monitoring the relative increase in the intrinsic tryptophan (W211) fluorescence of *Gai1*. All of the mutants showed similar nucleotide exchange rates compare to WT *Gai1* in both receptor bound and unbound states (Figure 6.2B). The simplest way to explain this data would be that those residues do not play a major role in G protein activation or that a single mutation is not enough to disturb the  $\alpha 5$  helix for GDP release. However, when we computed pairwise residue interactions, we identified interesting details for receptor mediated activation. In the basal state, I265, F267, Y320 and H322 were interacting hydrophobically with V339, V335, V342 and V335, respectively, within the  $\alpha 5$  helix. After receptor interaction and  $\alpha 5$  helix rotation, the same residues in  $\beta 5$  and  $\beta 6$  were predicted to hydrophobically interact with new sets of residues in the  $\alpha 5$  helix that were previously pointing toward solvent and not involved in binding in the basal state. Specifically I265, F267, Y320 and H322 started to interact with A338, N331, A338 and

F334 respectively (Figure 6.2C&D, See Supplemental Table 6.1 and Movies 6.1-6.3 for full data). The  $\alpha 5$  helix can glide along this hydrophobic surface during its rotation. These calculations thus suggested how new interactions on the rotated the  $\alpha 5$  helix can possibly compensate for the effect of single mutations in  $\beta 5$  and  $\beta 6$  strands during receptor mediated G protein activation.

### ***The effects of F336 mutants on G protein activation***

To test the role of interactions with the opposite site of the  $\alpha 5$  helix post-rotation, we focused on one specific residue in the  $\alpha 5$  helix, phenylalanine 336 (F336). F336 is one of the highly conserved residues in the  $G\alpha$  protein family as well as the small GTPases. The side chain faces the  $\beta 1$ ,  $\beta 2$ , and  $\beta 3$  strands as well as the  $\alpha 1$  helix, which creates one of the conserved hydrophobic clusters in the  $G\alpha$  subunit. Our previous energetic study predicted that F336 is the most critical residue for both basal and receptor mediated G protein activation within the  $\alpha 5$  helix (Figure 6.3A&B)<sup>133</sup>. To test the effect of mutating this residue, we substituted F336 with residues with decreasing hydrophobicity. All of the F336 mutants displayed increased basal exchange rates compared to WT (Figure 6.3C). Furthermore, a strong correlation was identified between the hydrophobicity of this residue and basal activity (Figure 6.3E). The fastest nucleotide exchange rate was detected for F336Y. However, in receptor mediated activation, nucleotide exchange rates were decreased compared to WT without any correlation with hydrophobicity (Figure 6.3D&F, Supplemental Movie 6.2&6.4). This result is consistent with a rotation of  $\alpha 5$  leading to a new surface-exposed location of F336 during  $\alpha 5$  helix rotation and translation caused by interaction with the receptor<sup>133</sup>. Overall these data suggest that F336 is one of the critical control points that regulate GDP release during G protein activation.

### ***The effects of F336 mutations on $\beta$ 6- $\alpha$ 5 loop; Cross-linking $\alpha$ 1 and $\alpha$ 5 helices***

The most obvious connection between the  $\alpha$ 5 helix and the nucleotide binding pocket is the  $\beta$ 6- $\alpha$ 5 loop. Perturbation of the  $\alpha$ 5 helix during receptor-mediated activation would disturb the interaction between the  $\beta$ 6- $\alpha$ 5 loop and the guanine ring of the nucleotide, leading to destabilization of the GDP in its binding pocket and domain opening of the  $\alpha$  subunit. To test the effect of F336 mutations on this loop, we cross-linked  $\alpha$ 1 to  $\alpha$ 5 to minimize the disruption of its interactions with the guanine ring by translocation toward the receptor. Cross-linking (XL) was performed between I56C-T329C residues on a cysteine depleted G $\alpha$ i1 (G $\alpha$ i1 HI) protein (Figure 6.4A). Without cross-linking, G $\alpha$ i1 HI I56C-T329C showed higher basal nucleotide exchange rates compared with the G $\alpha$ i1 HI protein (Figure 6.4B, black bars). Moreover, as expected, substitution of F336 for C on G $\alpha$ i HI I53C-T329C further increased the protein's activity. After cross-linking, the nucleotide exchange rate of cross-linked G $\alpha$ i1 XL HI I56C-T329C was decreased as compared to uncross-linked proteins, demonstrating the stabilizing effect of the cross-linking. Substitution of F336C on cross-linked G $\alpha$ i1 HI I56C-T329C increased basal protein activation as compared to the uncross-linked G $\alpha$ i1 HI I56C-T329C-F336C mutant (Figure 6.4B, black bars). This indicates that perturbation of F336 can trigger the activation mechanism without translocation of  $\alpha$ 5 toward the receptor and disruption of  $\beta$ 6- $\alpha$ 5 loop region.

Since receptor-mediated activation causes both a rotation of the  $\alpha$ 5 helix as well as an uncoiling of one turn of helix, we expected the cross-linked G $\alpha$  would be resistant to receptor-mediated activation. This is indeed what was found in both cross-linked proteins (Figure 6.4B, grey bars). This result might be caused by the reduced capability of cross-linked G $\alpha$  to interact with either G $\beta\gamma$  subunits or the receptor. To test the first possibility,

we measured the basal nucleotide exchange rates of  $G\alpha$  mutants in the presence or absence of  $G\beta\gamma$  subunits (Figure 6.4C). The results showed that basal nucleotide exchange rates decreased on both cross-linked and uncross-linked mutant  $G\alpha$  proteins in the presence of the  $G\beta\gamma$  subunit, like the WT protein. This suggested that cross-linked  $G\alpha$  subunits were still capable of interacting with  $G\beta\gamma$  subunits. To test the receptor binding capability of mutant  $G\alpha 1$  subunits, we determined the effect of cross-linking on the membrane association of the G protein with light-activated rhodopsin, a measure of the formation of the high-affinity R-G complex. As expected, cross-linking between  $\alpha 1$  and  $\alpha 5$  impaired this membrane binding (Figure 6.4D&E), consistent with a lack of ability of the cross-linked  $\alpha 5$  helix to translocate towards the receptor and the decreased nucleotide exchange rates. Overall, the cross-linking data suggest that perturbation of F336 triggers GDP release through destabilization of SW I-II regions via perturbing the  $\alpha 5$  helix interactions along the  $\alpha 1$  helix and  $\beta 2$ - $\beta 3$  strands, rather than disrupting the  $\beta 6$ - $\alpha 5$  loop region.

#### ***Hydrophobic interactions around F336: $\alpha 1$ helix interface binding energy and G protein activation***

Previous data suggested that F336's interaction with the  $\alpha 1$  helix and  $\beta 2$ - $\beta 3$  strands might be crucial for domain opening as the  $\alpha 1$  helix is positioned at the interface of the  $G\alpha$ -GTPase domain and the helical domain<sup>299</sup>. In addition, the  $\alpha 1$  helix and  $\beta 2$ - $\beta 3$  strands interact with the P-loop and SWI-II, respectively. To probe the effects of hydrophobic interactions around F336 with the  $\alpha 1$  helix, we computed interaction energies for all residues within the  $\alpha 1$  helix in both basal and receptor bound states of the heterotrimeric  $G\alpha\beta\gamma$  using our established protocol<sup>133</sup>. These  $\Delta\Delta G$  values probed for a potential network of intramolecular interactions which could propagate the conformational changes necessary for G protein activation and nucleotide exchange.  $\Delta\Delta G$  calculations predicted the

importance of  $\pi$ - $\pi$  interactions between the aromatic rings of F189 and H57 in the  $\beta$ 2 strand and  $\alpha$ 1 helix, respectively (Figure 6.5A, Table 6.2). This pairwise interaction couples with F336 on the  $\alpha$ 5 helix. Other predicted stabilizing interactions between  $\alpha$ 1 (Q52 and I56) keep the  $\alpha$ 5 helix (T329) fixed in the receptor unbound state; receptor interaction triggers unwinding of a turn of the  $\alpha$ 5 helix, disturbing this interaction (Figure 6.5B, Table 6.2). On the face of  $\alpha$ 1, in contact with the helical domain, residues (K51, K54, I55, Y61, and L175) on both the  $\alpha$ 1 and  $\alpha$ F helices assist to secure the helical domain in a “closed” GDP-bound conformation. The total interaction energy was approximately 25.4 Rosetta Energy Units (REUs). In the basal state, the  $\alpha$ 1 helix was predicted to interact favorably with  $\beta$ 2- $\beta$ 3 (F189, M198 and D200; 3.59 REU),  $\alpha$ 5 (V332, F336; REU 2.44) and helical domain (E65, L175; 1.84 REU) (Table 6.2). In the receptor bound state, the  $\alpha$ 1 helix was predicted to interact favorably with  $\alpha$ 5 (N331, V332; 2.1 REU) and as expected, the overall interaction was calculated as lower than the unbound state (Table 6.2).

To test our computational results, we mutated two residues that are predicted to stabilize the  $\alpha$ 1- $\alpha$ 5 interaction (F189 and F191). In the basal state, F189C increased nucleotide exchange 5-fold, while F191C showed no change relative to WT  $G\alpha$ 1 (Figure 6.5C). We prepared double and triple mutants with M53C and F196C mutants which we had previously tested<sup>133</sup>. Double mutants (M53C-F189C and F189C-F196C) exhibited similar basal activation and a triple mutant (M53C-F189C-F196C) showed an even higher basal exchange rate compared to the F336C  $G\alpha$ 1 mutant protein (Figure 6.5C). In receptor mediated activation of exchange, there was again a pattern of only modest inhibition, with F191C showing the largest decrease (Figure 6.5D) consistent with previously predicted  $\alpha$ 5<sup>133</sup> and  $\alpha$ 1 interface binding energy calculations.

### ***Perturbation of phosphate site of nucleotide binding region with F336 mutants***

To determine if the hydrophobic pocket around F336 was necessary to control the local order of the phosphate binding region of GDP, we used the sensitive monitor of  $Mg^{2+}$  binding into this region. Three different strategies were used to investigate the influence of Gai1 mutants on  $Mg^{2+}$  binding to this region: a)  $[Mg^{2+}]$  effects on the kinetics of nucleotide exchange, b)  $AlF_4^-$  binding, and c) trypsin digestion of Gai1 in the presence of different concentrations of  $Mg^{2+}$ . The results showed that the high nucleotide exchange rates of the mutants could be decreased in elevated  $Mg^{2+}$  concentrations (Figure 6.6A&B), suggesting that these mutations had allosteric effects on the phosphate binding region that could be overcome with higher  $Mg^{2+}$  concentration. The highest decrease in the rate of exchange, as a function of increasing concentrations of  $Mg^{2+}$ , was observed for the F336Y mutant, which showed the fastest exchange rate in the presence of low  $Mg^{2+}$  concentrations (Figure 6.3C).

To investigate the order of the  $Mg^{2+}$  binding region in the presence of GDP, the  $AlF_4^-$  binding assay was used. In this assay, changes in intrinsic tryptophan fluorescence rates of Gai1 were measured upon  $AlF_4^-$  addition in the presence of different  $MgCl_2$  concentrations.  $Mg^{2+}$  is necessary for  $AlF_4^-$  binding and generation of the active or transition state. Thus, this assay reflects both  $AlF_4^-$  and  $Mg^{2+}$  coordination in that region without nucleotide exchange. All mutations showed destabilization effects that were overcome with increasing  $Mg^{2+}$  concentration. The  $EC_{50}$  for  $Mg^{2+}$  stabilization of  $AlF_4^-$  binding for F336M, F336C, F336A and F336Y was increased by 1.4, 2.1, 2.8 and 3.1 fold, respectively, over the WT Gai1 under the same experimental conditions (Figure 6.6C). In addition to the  $\alpha 5$  helix mutants, the M53C-F189C-F196C mutant also exhibited

statistically significant increased EC<sub>50</sub> (Figure 6.6C).

The sensitivity of the Gα1 mutants to the trypsin digestion assay is a complementary assay to show the subtle changes in local order at the trypsin digestion site at R208 in the presence of varying Mg<sup>2+</sup> concentrations. After activation by either GDP·AlF<sub>4</sub><sup>-</sup> or GTPγS, Gα1 yields a ~34 kDa fragment following trypsin digestion. All high nucleotide exchange mutants had reduced stability as assayed by decreased 34kDa fragment in the presence of low Mg<sup>2+</sup> concentrations compared to the WT Gα1 subunit (Figure 6.6D).

#### ***Structural features of the x-ray structures of the F336C and F336Y mutants***

To probe the structural basis for the increased rates of nucleotide exchange observed in the F336 mutants, the crystal structures of the F336C and F336Y variants of the Gα1 subunit were determined in both the GDP and the GTPγS-bound states. The data collection and refinement statistics are summarized in Table 6.1. The mutations in the protein were confirmed by the crystal structure, where electron density at position 336 corresponded to either cysteine or tyrosine (Figure 6.7A&B). The structures of the GDP-bound form of F336C and F336Y Gα1 were refined to 2.0 and 2.4 Å resolution, respectively. Both GTPγS bound structures were refined to a resolution of 2.0 Å. The GDP- and GTPγS-bound structures for F336C and F336Y were determined in space groups identical to those of the WT Gα1 structures. Neither mutant showed significant structural differences compared to WT Gα1. Even with the F336 mutations in the α5 helix, the crystal structures showed the same localization and similar average *B* (temperature) factors around F336 region relative to those of WT Gα1 structures (Figure 6.7C). The effects of F336 mutations on the β2-β3 strands and β2-β3 loop were minimal (Figure 6.7D&E). Overall the root-mean square deviation (r.m.s.d) between WT Gα1-GDP with F336C and F336Y

G $\alpha$ i1-GDP was 0.42 Å and 0.36 Å (310 C $\alpha$  atoms aligned out of 324 total), respectively, whereas it was 0.31 and 0.29 Å (304 C $\alpha$  atoms aligned out of 315) for their GTP $\gamma$ S-bound structures.



## 6.4 Discussion

The  $\alpha 5$  helix of the  $G\alpha$  subunit is a critical region for both the receptor-mediated and basal activity<sup>156, 284, 292, 293, 295</sup>. It is encircled by hydrophobic interactions from six beta strands ( $\beta 1$ - $\beta 6$ ) and the  $\alpha 1$  helix ( $\alpha 1$ ). In the current study, we tested residues around the  $\alpha 5$  helix that we predicted as critical for the function of this helix during G protein activation in our previous studies. We highlight information flow within the G protein, starting from the  $\alpha 5$  helix to the GDP binding site of  $G\alpha$  using biochemical, structural and computational approaches.

Our previous study predicted that F336 within the  $\alpha 5$  helix is an important amino acid for both the active and inactive states<sup>133</sup>; a finding consistent with other studies<sup>292, 301</sup>. Mutation of this residue resulted in constitutive activity in both monomeric and heterotrimeric G proteins<sup>283, 292, 302, 319</sup>. It is also known that in small GTPases, structural perturbation of that region through mutation causes increased guanine nucleotide turnover that can lead to several diseases; these include Noonan, Cardio-faciocutaneous and Costello syndromes<sup>319-321</sup>.

In contrast to strong constitutive G protein activation, in this study, we did not observe drastic differences in the crystal structures of either GDP or GTP $\gamma$ S bound F336 mutants. Like another highly constitutively active G protein mutant,  $G\alpha i 1$  A326S<sup>254</sup>, F336 mutants showed similar structural features compared to WT  $G\alpha i 1$ . The guanine nucleotide provides a number of stabilizing interactions to the protein, perhaps inhibiting our ability to visualize subtle allosteric changes in the protein. In addition, other residues in the  $\alpha 5$  helix and  $\beta$ -strands may contribute in holding this region intact during the crystallization process.

How does the perturbation at F336 connect to the GDP binding region which is  $\sim 16$  Å

removed? F336 is a part of a highly conserved hydrophobic core in the  $G\alpha$  subunit. The effect of F336 mutations on basal G protein activation is correlated with the hydrophobicity of this region (Figure 6.3C&E). Once the receptor contacts the  $\alpha 5$  helix and causes its rotation and displacement into the receptor binding site, this F336 is now in a hydrophilic environment. We propose that breaking the hydrophobic core is a key event in perturbing GDP binding<sup>133</sup>. Interestingly, we did not observe any effects of the hydrophilic mutants on receptor mediated activation; this is likely due to the new solvent exposed site which prevents these side chains from contacting anything other than solvent upon receptor binding (Figure 6.3D&F).

To trace the hydrophobic interactions and to discern a possible interaction network from the F336 residue to the GDP binding site, we computed binding energies of different regions in the  $G\alpha$  subunit by using different Rosetta algorithms. Adding to our previous calculations ( $\alpha 5$  helix: $G\alpha$  interface binding energy,<sup>133</sup>), we predicted that the F336 side chain is mostly coupled with M53 ( $\alpha 1$ ), I56 ( $\alpha 1$ ), F189 ( $\beta 2$ ), F191 ( $\beta 2$ ), F196 ( $\beta 3$ ), V332 ( $\alpha 5$ ), Q333 ( $\alpha 5$ ), V339 ( $\alpha 5$ ) and T340 ( $\alpha 5$ ). Thus the effects of F336 are not solely local and not coupled to a single residue, but rather might be part of a distributed network of interactions in which the activation is coupled to changes in regions dispersed across both domains of the  $G\alpha$  subunit. F336 is likely making direct hydrophobic contacts with F191 and M53. It potentially communicates with F189 via two paths.

The first is through residues M53-H57-F191 which interact with F189 through a  $\pi$ - $\pi$  interaction between residues H57 and F189 (Figure 6.5A). This is consistent with one of our previous studies<sup>322</sup> in which the constitutively active I56C( $\alpha 1$ )-Q333C( $\alpha 5$ ) double mutant of  $G\alpha 1$  made a spontaneous disulphide bond between the  $\alpha 1$  and  $\alpha 5$  helices. This

structure showed significant rearrangement of side chain residues H57, F189, F191, and F332 and disturbed  $\pi$ - $\pi$  interaction between H57 and F189.

The second path begins from the direction of F196, which interacts with F336 via F191 and T340 residues. These observations indicate that the perturbation effects of F336 spread with complex interactions via the  $\alpha$ 1 helix and  $\beta$ 2- $\beta$ 3 strands. These interactions also spread to the  $Mg^{2+}$  ion and the nucleotide binding region (Figures 6.5 & 6.6) as evidenced by our nucleotide exchange data combined with the perturbations seen in the  $Mg^{2+}$  and  $AlF_4^-$  assays which supports previous studies<sup>246, 283</sup>.

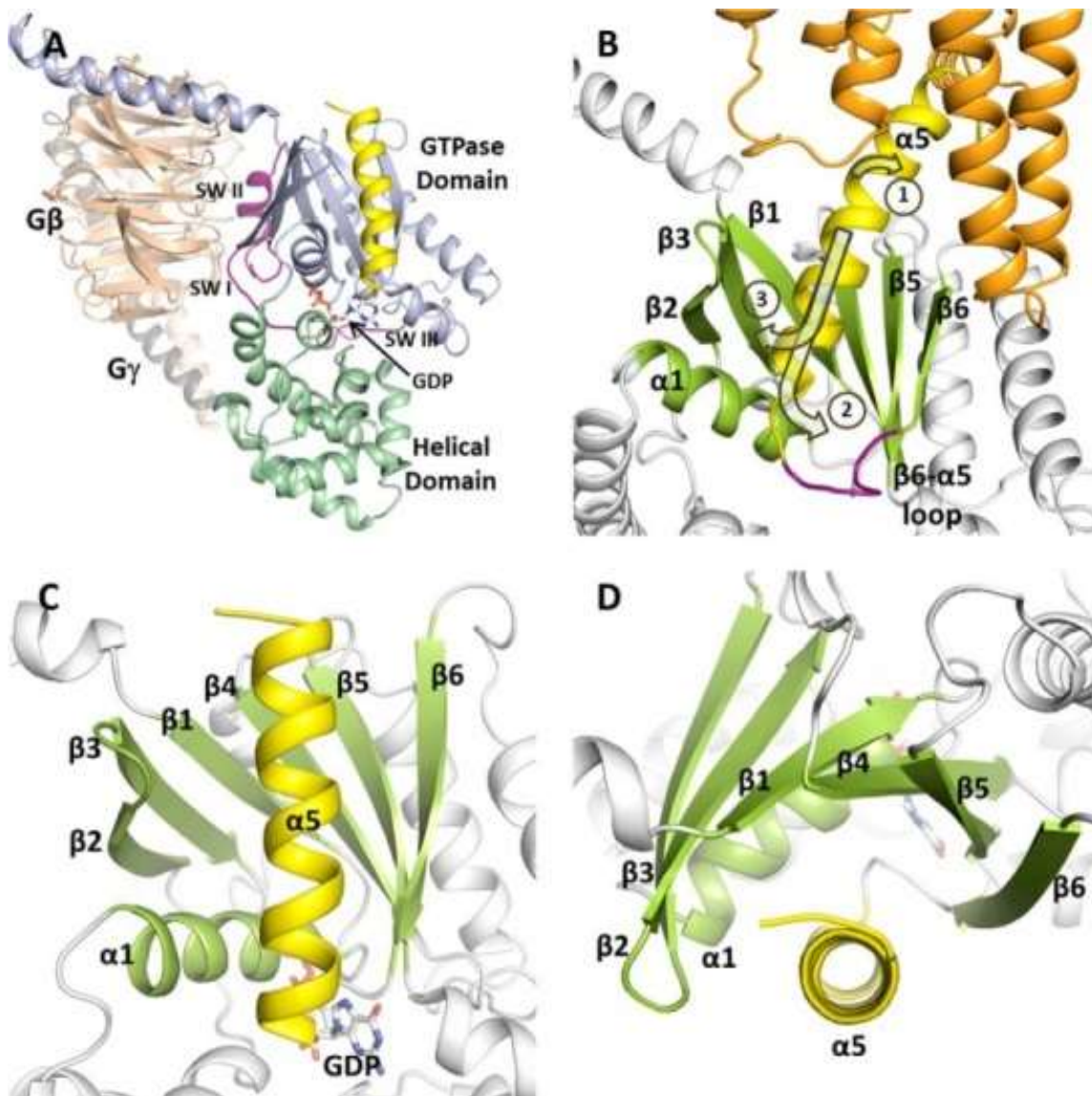
We also tested the effects of residues within the  $\beta$ 5- $\beta$ 6 strands (I265 ( $\beta$ 5), F267 ( $\beta$ 5), Y320 ( $\beta$ 6), H322 ( $\beta$ 6)) interacting with the other side of the  $\alpha$ 5 helix on G protein activation. We observed no major effects from the mutations either in the basal or receptor mediated exchange assays. These data suggest how new interactions on the rotated  $\alpha$ 5 helix can compensate for the effects of single mutations in the  $\beta$ 5 and  $\beta$ 6 strands during receptor mediated G protein activation. It also strongly suggests that the activation route goes through the other side of the protein (the  $\beta$ 1- $\beta$ 3/ $\alpha$ 1 to Switch I, P-loop,  $Mg^{2+}$  binding and GDP binding site), consistent with previously published findings<sup>301</sup>. In addition, after restricting the C-terminal rotation and translocation by cross-linking the  $\alpha$ 1 and  $\alpha$ 5 helices, F336 mutants can still induce increased basal nucleotide exchange (Figure 6.4). This observation indicates that G proteins do not need a large displacement of  $\alpha$ 5 for basal state activation; rather, perturbing the  $\beta$ 2- $\beta$ 3 and  $\alpha$ 1 regions are sufficient.

In summary, our study used a predictive energetic analysis to pinpoint information flow through  $G\alpha$  from receptor interaction to triggering of GDP release. We highlighted the hydrophobic interactions around F336 as a key for stability of GDP binding, as well as

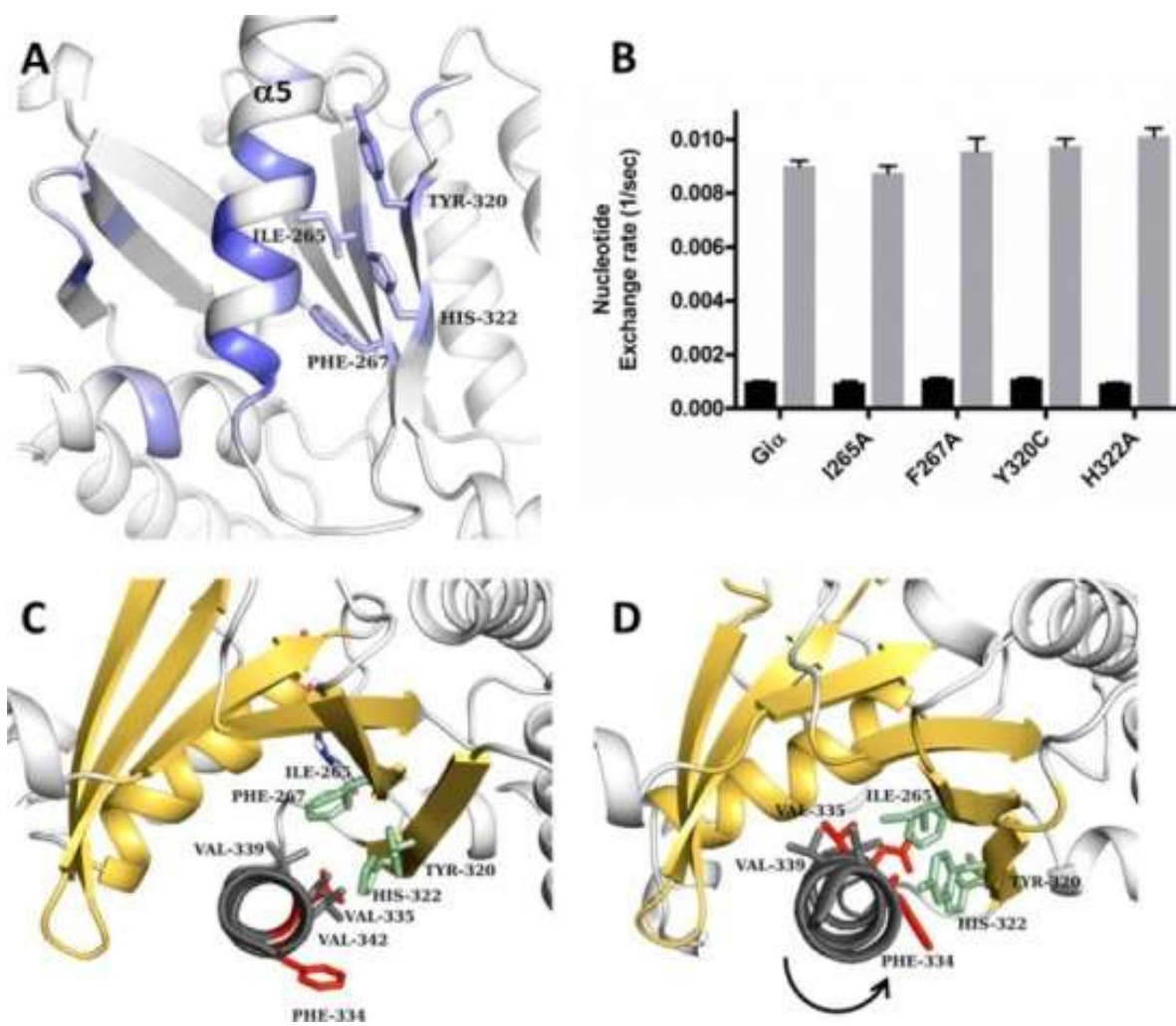
removal of these hydrophobic interactions by receptor-mediated helical rotation to trigger GDP release. We suggested the route of information flow triggers through the  $\alpha 5$  helix,  $\beta 2$ - $\beta 3$  strands and the  $\alpha 1$  helix using energetic analysis and mutagenesis. We also showed that the dynamics of the  $Mg^{2+}$  and  $\beta$ -phosphate binding area of GDP are perturbed by mutagenesis of this conserved residue. The  $\beta 5$ - $\beta 6$  residues which face the  $\alpha 5$  helix are likely important structurally rather than functionally according to our analysis. Thus, our data suggest that after the initial interaction of the G protein with the receptor and CT rotation, disruption of a conserved hydrophobic network around F336 engages both  $\beta 1$ - $\beta 3$  and  $\alpha 1$  to Switch I and the P-loop which decreases the stability of the GDP binding pocket and triggers nucleotide release.

#### ***Abbreviations***

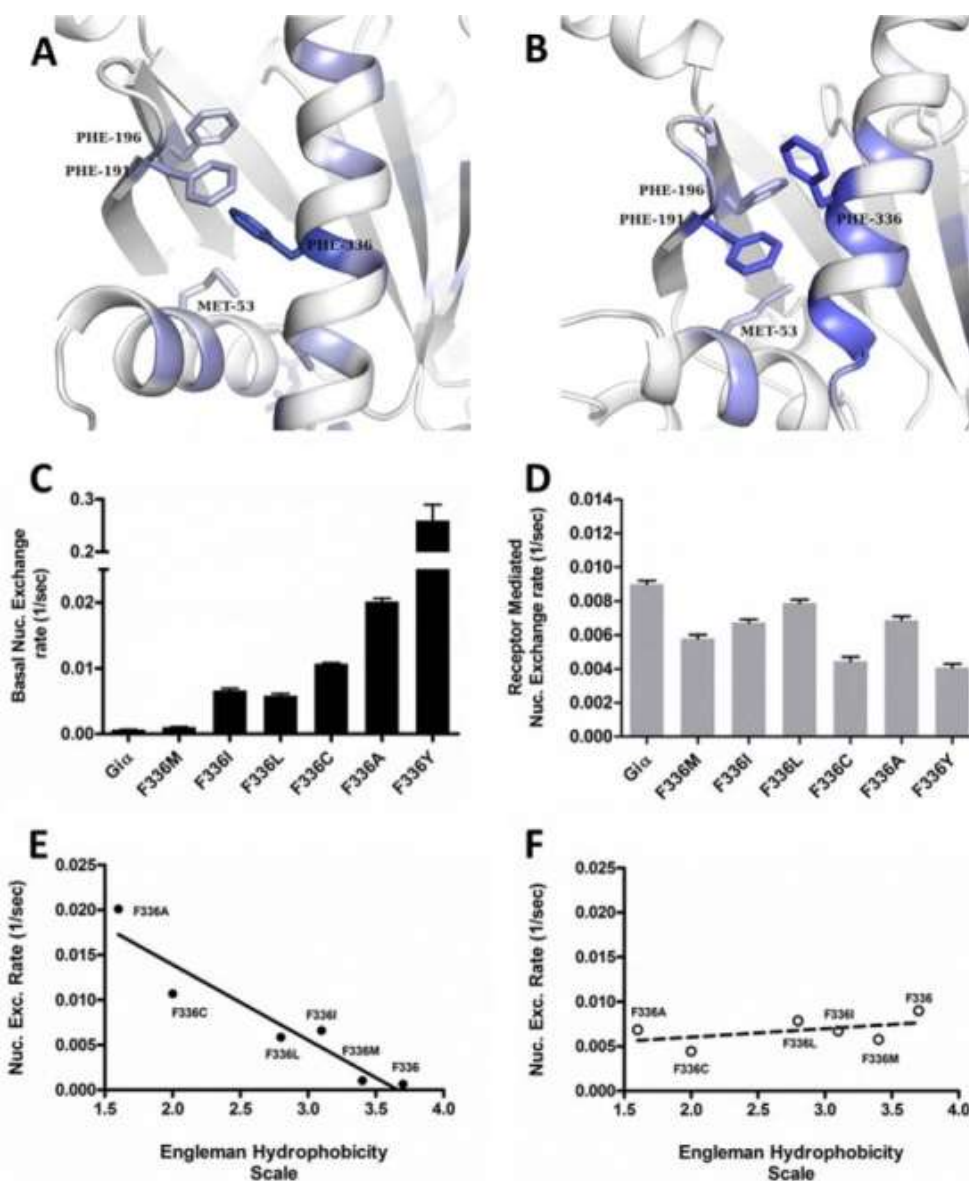
GPCR	- G protein-coupled receptor
GDP	- guanosine diphosphate
GTP $\gamma$ S	- guanosine 5'-[ $\gamma$ -thio]triphosphate
G $\alpha_{i1}$ HI	- G $\alpha_{i1}$ Hexa I
REU	- Rosetta energy unit(s)
EPPS	- 4-(2-hydroxyethyl)-1-piperazinepropanesulfonic acid.



**FIGURE 6.1. Heterotrimeric G protein; localization and function  $\alpha 5$  helix in G proteins.** (A) Ribbon model of heterotrimeric G protein ( $G_i\alpha\beta\gamma$ , PDB entry, 1GP2). The  $G\alpha$  subunit is composed of nucleotide binding (GTPase domain, light blue) and helical domains (green). The  $\alpha 5$  helix and switch regions are colored yellow and purple, respectively. GDP is shown as sticks. (B) Receptor (orange) mediated G protein activation routes. The binding of the GPCR to the C-terminus (CT) of  $G\alpha$  is thought to trigger conformational changes that can be transmitted via rotation of the  $\alpha 5$  helix (black, arrow 1) of  $G\alpha$  to the  $\beta 6$ - $\alpha 5$  loop (purple, arrow 2) that binds the purine ring of the GDP. In the second route, disruption of the phosphate interactions with the nucleotide binding pocket via destabilization of SW I-II regions through perturbing  $\alpha 5$  interaction with the  $\beta 2$ - $\beta 3$  strands (arrow 3). Rhodopsin –  $G_i$  complex model adapted from Alexander et.al. <sup>133</sup>. (C and D) The  $\alpha 5$  helix is one of the most critical regions for G protein stability and activation. (A and B) The  $\alpha 5$  helix (yellow) is protected by six beta strands ( $\beta 1$ - $\beta 6$ ) and one  $\alpha$  helix ( $\alpha 1$ ) (green). The structure is adapted from the crystal structure of the  $G_i$  heterotrimer (PDB entry, 1GP2).

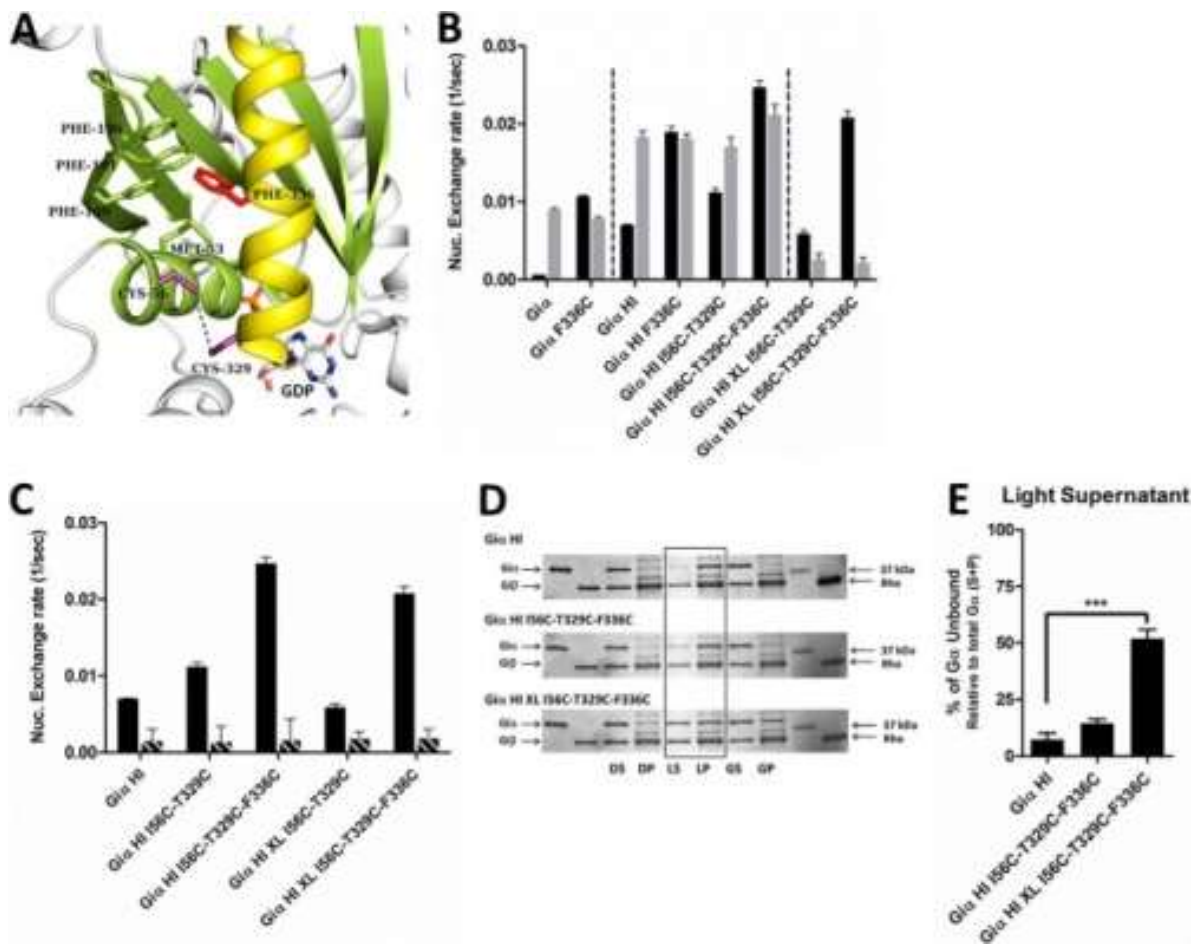


**FIGURE 6.2. The effects of  $\beta 5$ - $\beta 6$  strands mutations on G protein activation.** (A) Rosetta energy analysis of the interface between the  $\alpha 5$  helix (black) and the GTPase domain in the receptor bound state. Residues are colored by the interaction energy as reported in REU, or Rosetta Energy Units (dark blue, the most attractive). Calculations adapted from <sup>133</sup>. (B) Basal (black bars) and receptor (grey bars) mediated nucleotide exchange rates for the  $\beta 5$  strand (I265A and F267A) and  $\beta 6$  strand (Y320C and H322A) mutations in *Gai1* proteins. Nucleotide exchange was monitored by measuring the enhancement in intrinsic tryptophan (W211) fluorescence (ex 290 nm, em 340 nm) as a function of time after addition of GTP $\gamma$ S <sup>323</sup>. (C) Most favorable interactions between the  $\alpha 5$  helix (V335, V339 and V342),  $\beta 5$  strand (I265 and F267), and  $\beta 6$  strand (T320 and H322) interface in the basal state. (D) After receptor interaction and  $\alpha 5$  helix rotation (arrow), the same residues in  $\beta 5$  and  $\beta 6$  were hydrophobically interacting with new residues in the  $\alpha 5$  helix (red labeled). Please see Supplemental Table and Supplemental Movie 1-3 for full interactions in both receptor bound and unbound states.



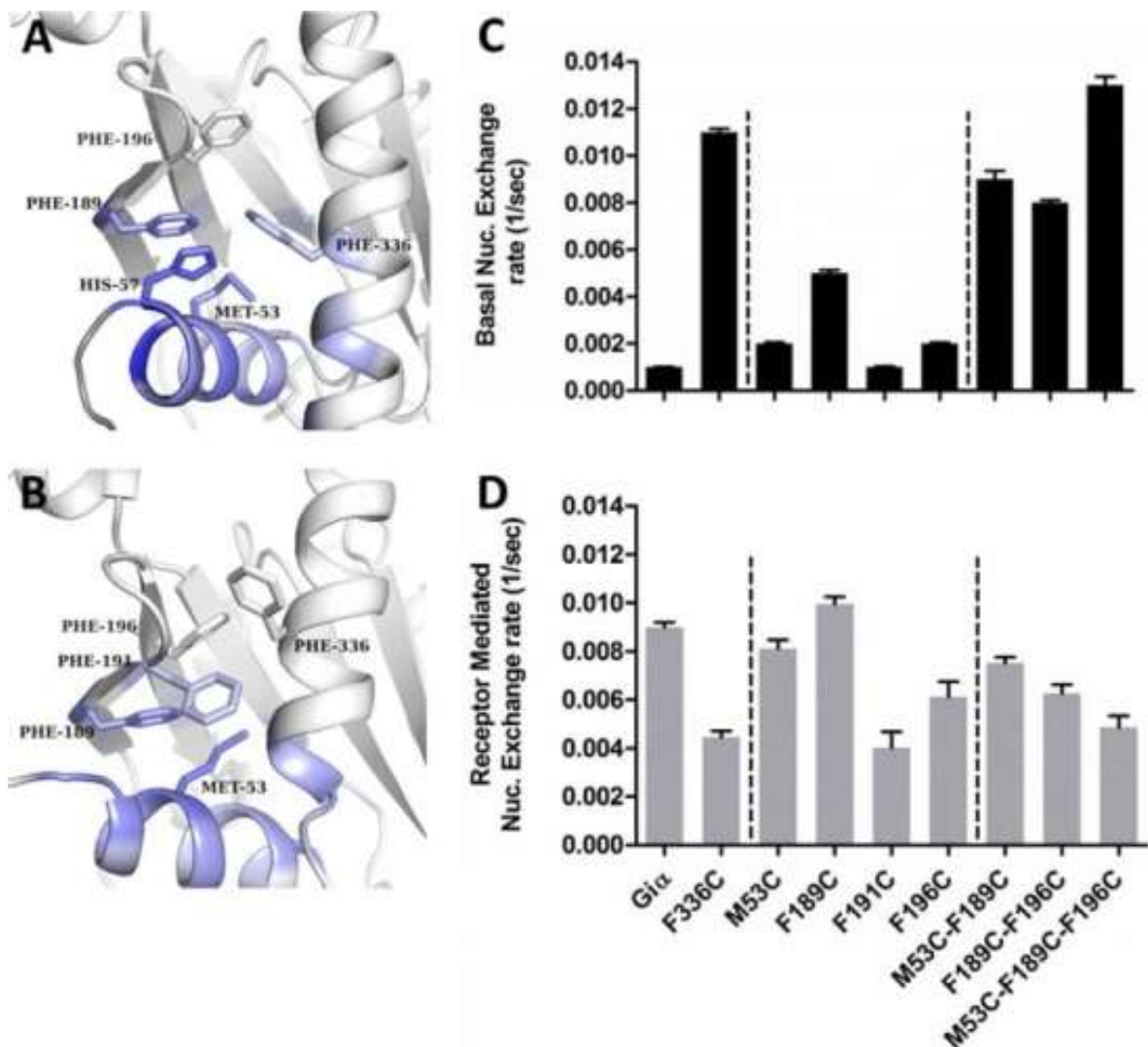
**FIGURE 6.3. The effect of F336 residue on G protein activation.** Rosetta energy analysis of the interface between the  $\alpha 5$  helix and GTPase domain in the basal state (A) and receptor bound state (B). Residues are colored by the interaction energy REU (dark blue, the most attractive). Calculations adapted from <sup>133</sup>. Basal (C) and Receptor mediated (D) nucleotide exchange rates of G $\alpha$ 1 F336 mutants. The data were normalized to the baseline and maximum fluorescence and then fit to the exponential association equation ( $Y = Y_{max} * (1 - e^{-kt})$ ), to calculate rate constant (k). Data were collected at 21 C $^{\circ}$  for 60 min. Results represent the mean  $\pm$  SEM values of at least three independent experiments. Correlation between nucleotide exchange rates and hydrophobicity identity of the amino acids in basal (E) and receptor bound (F) state. Engelman Scale was used during comparison and correlation coefficients were calculated with or without F336Y mutant data. The Pearson correlation in the basal state with F336Y is 0.9358; without F336Y, it is 0.9945. In receptor mediated state Pearson correlation with F336Y is 0.6992, without F336Y is 0.4861.



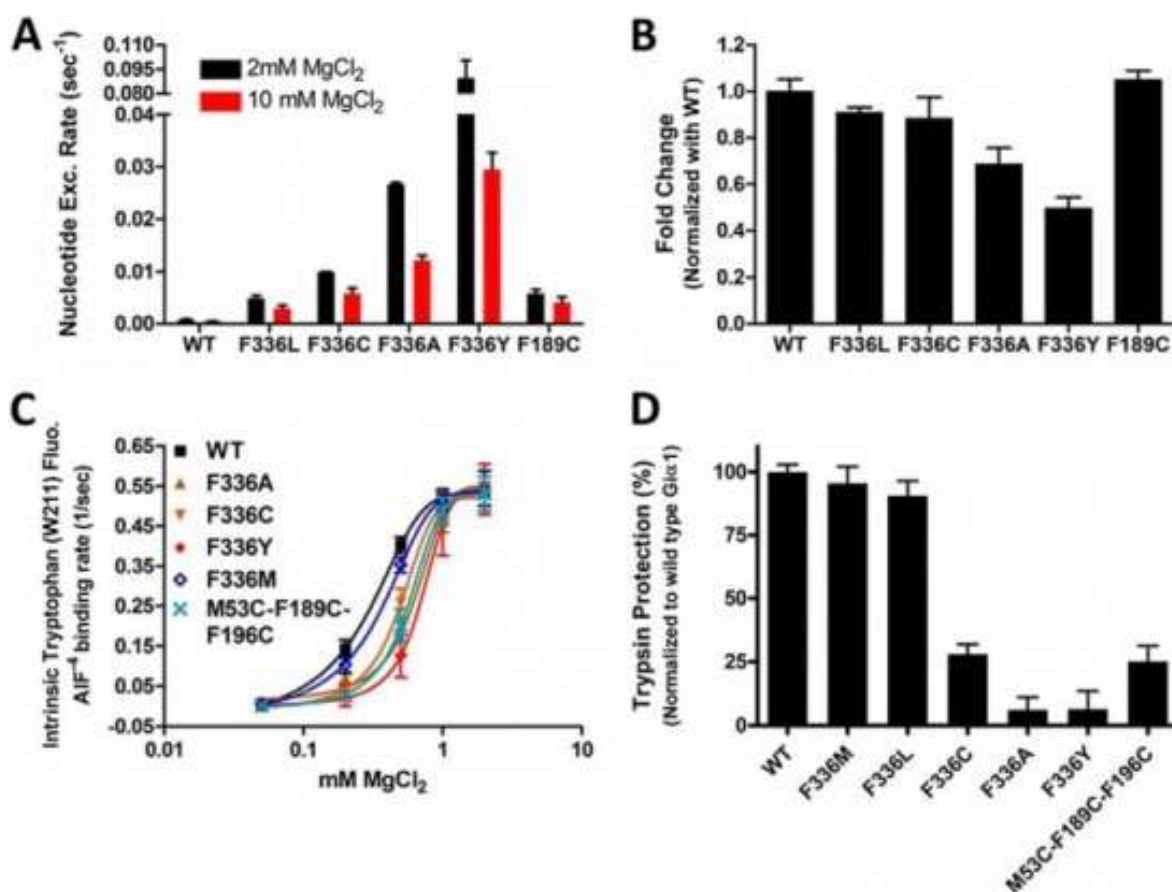


**FIGURE 6.4. Cross-linking of  $\alpha 1$  and  $\alpha 5$  helices of *Gai1* HI.** (A) Cartoon representation of cross-linking (XL) region. Cross-linking was performed between I56C ( $\alpha 1$ ) and T329C ( $\alpha 5$ ) (purple) residues on a cysteine depleted *Gai1* (*Gai1* HI) protein. F336 ( $\alpha 5$ ) residue is colored red, F189 ( $\beta 2$ ), F191 ( $\beta 2$ ), F196 ( $\beta 3$ ) and M53 ( $\alpha 1$ ) residues are colored green. The  $\alpha 5$  helix is colored yellow, the  $\beta 1$ - $\beta 6$  strands and  $\alpha 1$  helix are colored green. (B) Basal (black bars) and receptor (grey bars) mediated nucleotide exchange rates for cross-linked *Gai1* HI proteins. (C) Basal nucleotide exchange rates in the presence of  $G\beta\gamma$  subunit.  $G\alpha$  (black bars),  $G\alpha\beta\gamma$  (shaded black bars). (D) Membrane binding of wild type and mutant *Gai1* HI proteins. Assay was performed as described in method section. DS, supernatant from dark sample; DP, pellet fraction from dark sample; LS, supernatant from light sample; LP, pellet from light sample; GS, supernatant from light- and  $GTP\gamma S$ -activated sample; GP, pellet from light- and  $GTP\gamma S$ -activated sample; XL, cross-linked sample. (E) Densitometric quantification of supernatant from light samples. Each sample from SDS-PAGE (section d) was evaluated by comparison of the amount of *Gai1* subunits in pellet (P) or supernatant (S) to the total amount of *Gai1* subunits (P+S) in both treatments and expressed as a percentage of the total *Gai1* protein. Data represents the average of three independent experiments.

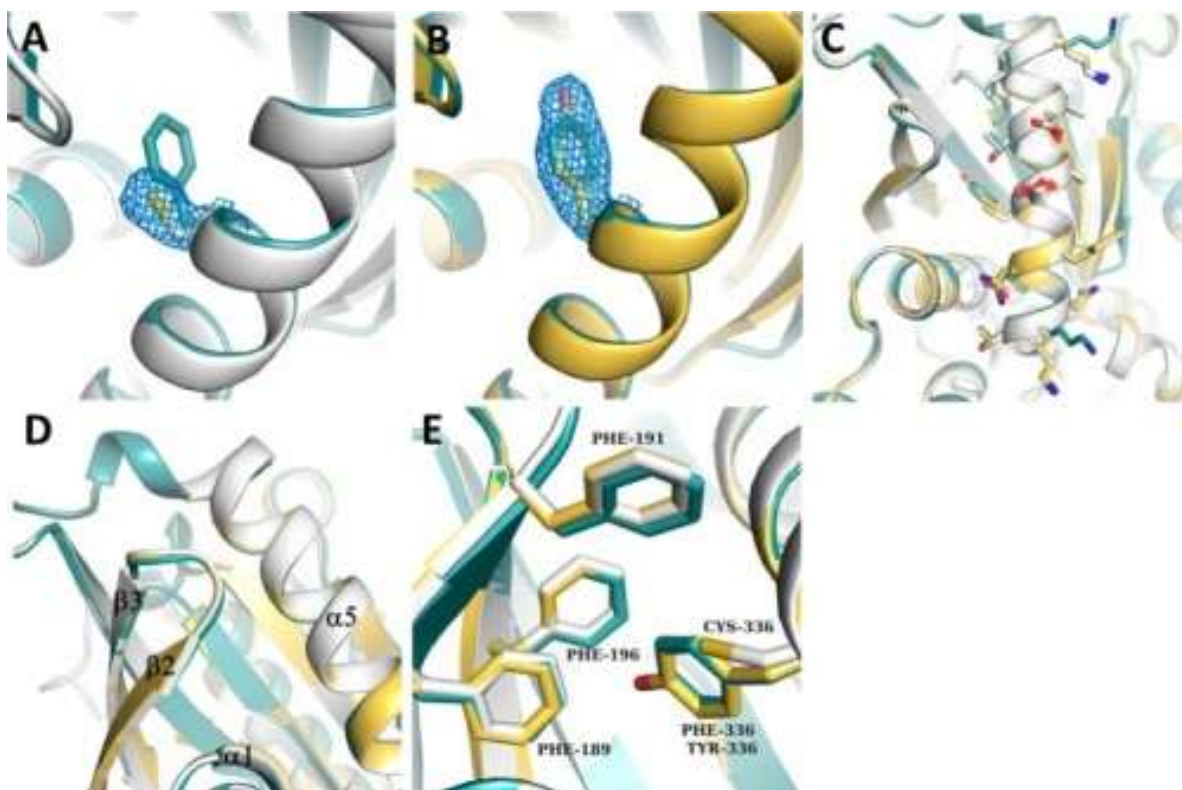




**FIGURE 6.5. The effects of hydrophobic residues around F336 on nucleotide exchange rates.** Rosetta energetic analysis of the interface between  $\alpha 1$  helix and GTPase domain in the basal state (A) and receptor bound state (B). Residues are colored by the interaction energy in REU (dark blue, the most attractive). Basal (C) and receptor mediated (D) nucleotide exchange rates of single, double and triple mutants within the  $\beta 2$ - $\beta 3$  strands and  $\alpha 1$  helix as determined by monitoring intrinsic tryptophan (W211) fluorescence changes upon addition of GTP $\gamma$ S. Data were collected at 21 °C for 45 min. Results represent the mean  $\pm$  SEM values of at least three independent experiments.



**FIGURE 6.6. The effect of MgCl<sub>2</sub> on Gai1 basal activity.** (A) Basal nucleotide exchange in the presence 2 mM and 10 mM MgCl<sub>2</sub> concentrations. (B) Changes in the nucleotide exchange rate in the presence of different MgCl<sub>2</sub> concentrations. Fold change calculated from (A) and normalized with Gai1 (WT) data. (C) Rates of intrinsic tryptophan fluorescence changes in Gai1 upon aluminum fluoride (AlF<sub>4</sub><sup>-</sup>) addition in the presence of different MgCl<sub>2</sub> concentrations (0.1-2 mM). Intensity of tryptophan signal were monitored (ex: 290 nm, em: 340 nm) at 21 °C for 10 min before and after the addition of AlF<sub>4</sub><sup>-</sup> (10 mM NaF and 50 μM AlCl<sub>3</sub>). The data were calculated as described above and rate constants plotted against MgCl<sub>2</sub> concentrations. (D) Trypsin digestion and analysis of Gai1 protein subunit. The densitometric measurement of proteolytic fragments in the presence of GDP - AlF<sub>4</sub><sup>-</sup> + 0.5 mM MgCl<sub>2</sub>. Results normalized with WT Gai1 data and fragments quantified by densitometry (Multimager, Bio-Rad). Results represent the mean ± SEM values of at least 6-8 independent experiments.



**FIGURE 6.7. Structural features of GDP bound F336 mutant structures.** Electron density for the F336C (A) and F336Y (B) side chains in the GDP bound state of  $G\alpha 1$ . Corresponding regions in GDP-bound WT  $G\alpha 1$  (PDB entry, 1GDD<sup>313</sup>; teal) are superposed. Difference electron density is from a  $|F_o| - |F_c|$  omit map calculated after the removal of residue 330 to 340 and contoured to  $3\sigma$  around the omitted side chain. (C) Comparison of the  $\alpha 5$  helix between F336C-GDP (white), F336Y-GDP (yellow) and WT  $G\alpha 1$ -GDP (PDB entry, 1GDD, teal). (D) Overview of the  $\beta 2$ - $\beta 3$  strands and  $\beta 2$ - $\beta 3$  loop. (E) Comparison of relative localization of F189 ( $\beta 2$ ), F191 ( $\beta 2$ ), F196 ( $\beta 3$ ) and F336 ( $\alpha 5$ ) residues between F336C-GDP (white), F336Y-GDP (yellow) and WT- $G\alpha 1$  (PDB entry, 1GDD, teal) structures.

	F336C-GDP	F336C-GTP $\gamma$ S	F336Y-GDP	F336Y-GTP $\gamma$ S
<b>Data Collection and Processing<sup>a</sup></b>				
Beamline	21-ID-G	21-ID-G	21-ID-G	21-ID-G
Space groups	I4	P3 <sub>2</sub> 21	I4	P3 <sub>2</sub> 21
Cell Dimensions: a, b, c (Å)	121.1, 121.1, 68.18	79.2, 79.2, 107.9	121.5, 121.5, 68.2	79.3, 79.3, 105.1
$\alpha$ , $\beta$ , $\gamma$ (degrees)	90, 90, 90	90, 90, 120	90, 90, 90	90, 90, 120
Resolution (Å)	34-2.1 (2.18-2.1)	31-2.0 (2.07-2.0)	20-2.4 (2.5-2.4)	42-2.0 (2.07-2.0)
Total Reflections	255,402	307,412	177,466	437,402
Unique Reflections	28,903	26,186	19,617	26,483
R <sub>sym</sub> <sup>b</sup> (%)	5.3 (37.9)	10.1 (44.7)	6.2 (32)	10.2 (44.6)
R <sub>pim</sub> <sup>c</sup> (%)	2.9 (23.2)	5.2 (23.5)	3.3 (18.4)	4.7 (20.7)
$\langle I \rangle / \langle \sigma \rangle$	19.9 (2.6)	13.5 (3.1)	19.3 (3.46)	17.5 (3.9)
Completeness (%)	99.6 (99.5)	100 (100)	99.3 (99)	100 (100)
<b>Refinement Statistics</b>				
R <sub>work</sub> <sup>d</sup> (%)	18.8	16.4	18.2	16.9
R <sub>free</sub> (%)	21.8	20.8	23.2	20.6
RMS deviations				
Bond (Å)	0.008	0.007	0.008	0.007
Angle (°)	1.029	0.981	1.011	1.009
Ramachandran statistics <sup>e</sup>				
Favored (%)	98.5	99.06	98.11	98.42
Allowed (%)	1.5	0.94	1.89	1.58
Outliers (%)	0.0	0.0	0.0	0.0

<sup>a</sup>Numbers in parentheses indicate statistics for the highest shell.

<sup>b</sup> $\sigma_{\text{int}} = \sum |I_{\text{obs}} - \langle I \rangle| / \sum |I_{\text{obs}}|$  where  $I$  is intensity,  $I_{\text{obs}}$  is the  $i$ th measurement, and  $\langle I \rangle$  is the weighted mean of  $I$ .

<sup>c</sup> $\sigma_{\text{pim}} = \sum_{hkl} \sqrt{[1/(N-1)] \sum_i |I_{\text{obs}}(hkl) - \overline{I(hkl)}|^2} / \sum_{hkl} \sum_i I_{\text{obs}}(hkl)$  where  $I$  is running over the number of independent observations of reflection  $hkl$  and  $N$  is representing the number of replicate observations.

<sup>d</sup> $\sigma_{\text{Rwork}} = \sum ||F_{\text{obs}}| - |F_{\text{calc}}|| / \sum |F_{\text{obs}}|$  where  $F_{\text{obs}}$  and  $F_{\text{calc}}$  are the observed and calculated structure factor amplitudes.  $\sigma_{\text{Rfree}}$  is the same as  $\sigma_{\text{Rwork}}$  for a set of data omitted from the refinement.

<sup>e</sup>Ramachandran analysis from MOLPROBITY<sup>324</sup>.

**Table 6.1. Crystallographic data collection and refinement statistics.**

Free G $\alpha$					Receptor – G $\alpha$ complex				
Entity	Amino acid	Energy in REU	Std. dev	Z-score	Entity	Amino acid	Energy in REU	Std. dev	Z-score
$\beta$ 1	L038	0.87	$\pm$ 0.04	22.75	$\beta$ 1	L038	0.78	$\pm$ 0.16	4.85
$\alpha$ 1	K046	1.14	$\pm$ 0.28	4.01	$\beta$ 1	G040	0.72	$\pm$ 0.34	2.12
$\alpha$ 1	S047	0.95	$\pm$ 0.04	21.55	$\alpha$ 1	K046	1.58	$\pm$ 0.41	3.88
$\alpha$ 1	T048	1.82	$\pm$ 0.05	38.27	$\alpha$ 1	S047	0.71	$\pm$ 0.14	5.01
$\alpha$ 1	I049	0.99	$\pm$ 0.09	11.42	$\alpha$ 1	I049	1.11	$\pm$ 0.1	10.77
$\alpha$ 1	K051	0.82	$\pm$ 0.1	8.10	$\alpha$ 1	V050	0.75	$\pm$ 0.16	4.61
$\alpha$ 1	Q052	1.65	$\pm$ 0.05	34.32	$\alpha$ 1	K051	0.52	$\pm$ 0.34	1.55
$\alpha$ 1	M053	1.33	$\pm$ 0.11	12.04	$\alpha$ 1	Q052	1.08	$\pm$ 0.13	8.03
$\alpha$ 1	K054	2.49	$\pm$ 0.07	38.00	$\alpha$ 1	M053	1.62	$\pm$ 0.15	10.72
$\alpha$ 1	I055	1.03	$\pm$ 0.16	6.53	$\alpha$ 1	K054	0.94	$\pm$ 0.45	2.11
$\alpha$ 1	I056	1.07	$\pm$ 0.03	32.77	$\alpha$ 1	I056	1.18	$\pm$ 0.2	5.85
$\alpha$ 1	H057	1.73	$\pm$ 0.08	22.03	$\alpha$ 1	H057	1.20	$\pm$ 0.57	2.12
Helical	E065	0.78	$\pm$ 0.11	6.96	$\beta$ 2	F189	1.43	$\pm$ 0.17	8.56
Helical	L175	1.06	$\pm$ 0.08	13.10	$\beta$ 2	F191	0.55	$\pm$ 0.08	6.60
$\beta$ 2	F189	1.41	$\pm$ 0.09	15.72	$\alpha$ 5	N331	1.02	$\pm$ 0.04	23.43
$\beta$ 3	M198	0.50	$\pm$ 0.12	4.28	$\alpha$ 5	V332	0.68	$\pm$ 0.08	8.11
$\beta$ 3	D200	0.81	$\pm$ 0.33	2.45	GDP		0.70	$\pm$ 0.19	3.79
$\beta$ 6- $\alpha$ 5	A326	1.62	$\pm$ 0.04	41.26	GDP	cumulative	0.70		
$\beta$ 6- $\alpha$ 5	T329	0.82	$\pm$ 0.02	41.22	$\alpha$ 1	cumulative	10.69		
$\alpha$ 5	V332	0.85	$\pm$ 0.03	31.85	$\alpha$ 5	cumulative	1.70		
$\alpha$ 5	F336	0.72	$\pm$ 0.05	15.83	$\beta$ -	cumulative	3.48		
GDP		0.95	$\pm$ 0.13	7.32	strands				
GDP	cumulative	0.95			overall	cumulative	16.57		
$\alpha$ 1	cumulative	15.02							
Helical	cumulative	1.84							
$\beta$ 6- $\alpha$ 5	cumulative	2.44							
$\alpha$ 5	cumulative	1.57							
$\beta$ -	cumulative	3.59							
strands									
overall	cumulative	25.43							

Table 6.2. G protein alpha subunit  $\alpha$ 1 helix interface energetic prediction

## 6.5 Supplemental Information

**SUPPLEMENTARY TABLE 6.1.** Average energy scores between interacting residue pairs of the receptor-unbound and bound alpha5 helix (x-axis) and non-alpha5 regions (y-axis). Average energies between pairwise interacting residues were computed using Rosetta's per residue energy-breakdown protocol. The energy between all possible pairs of interacting amino acid residues within the G-protein were calculated across the previously published ensembles of ten structures<sup>133</sup>. These energies between all residues pairs was then averaged across the ten models in both the receptor bound and basal state.

	T329 ( $\alpha$ )		K330 ( $\alpha$ )		N331 ( $\alpha$ )		V332 ( $\alpha$ )		Q333 ( $\alpha$ )	
	Unbound	Bound	Unbound	Bound	Unbound	Bound	Unbound	Bound	Unbound	Bound
V34 ( $\beta$ 1)	-	-	-	-	-	-	-	-	-	-
L36 ( $\beta$ 1)	-	-	-	-	-	-	-	-	-	-
I49 ( $\alpha$ 1)	-	-	-	-	-	-0.2776	-0.2848	-0.0851	-	-
Q52 ( $\alpha$ 1)	-0.7536	-0.1637	-0.0005	-0.0941	-	-1.3867	-0.6477	-0.1549	-	-
M53 ( $\alpha$ 1)	-	-	-	-	-	-	-0.2292	-0.7078	-	-
I55 ( $\alpha$ 1)	-0.2232	-	-	-	-	-	-	-	-	-
I56 ( $\alpha$ 1)	-0.6288	-	-	-0.17	-	-	-0.538	-0.4219	-0.9294	-
H57 ( $\alpha$ 1)	-	-	-	-	-	-	-	-	-	-
Q172 ( $\alpha$ F)	0.0654	-	-	-	-	-	-	-	-	-
L175 ( $\alpha$ F)	0.0109	-	-	-	-	-	-	-	-	-
F189 ( $\beta$ 2)	-	-	-	-	-	-	-	-0.3583	-	-0.642
F191 ( $\beta$ 2)	-	-	-	-0.152	-	-	-	-1.813	-	-0.119
K192	-	-	-	-0.2594	-	-	-	-	-	-
L194 ( $\beta$ 3)	-	-	-	-	-	-	-	-	-	-
F196 ( $\beta$ 3)	-	-	-	-	-	-	-	-	-	-
M198 ( $\beta$ 3)	-	-	-	-	-	-	-	-	-	-
T219	-	-	-	-	-	-	-	-	-	-
A220 ( $\beta$ 4)	-	-	-	-	-	-	-	-	-	-
I222 ( $\beta$ 4)	-	-	-	-	-	-	-	-	-	-
D261	-	-	-	-	-	-	-	-	-	-
S263 ( $\beta$ 5)	-	-	-	-	-	-	-	-	-	-
I265 ( $\beta$ 5)	-	-	-	-	-	-	-	-	-	-
F267 ( $\beta$ 5)	-	-	-	-	-	-0.7487	-	-	-	-
N269 ( $\beta$ 5)	-	-	-	-	-	-0.0743	-	-	-	-
K271 ( $\alpha$ G)	-	-	-	-	-0.0047	-	-	-	-	-
D272 ( $\alpha$ G)	-	-	-	-	-0.0898	-	-	-	-	-
D315	-	-	-	-	-	-	-	-	-	-
T316	-	-	-	-	-	-	-	-	-	-
K317	-	-	-	-	-	-	-	-	-	-
E318	-	-	-	-	-	-	-	-	-	-
I319 ( $\beta$ 6)	-	-	-	-	-	-	-	-	-	-
T320 ( $\beta$ 6)	-	-	-	-	-	-	-	-	-	-
H322 ( $\beta$ 6)	-	-	-	-	-	-	-	-	-	-
F323 ( $\beta$ 6)	-	-	-	-	-0.0462	0.0027	-	-	-	-
T324	-	-0.1044	-	-	0.3071	-0.8013	-	-	-	-
C325	0.0046	-	-	-	-1.2066	0.006	-0.1078	-	-	-
A326	0.0002	-0.2351	-	-	-	-	-0.348	-	-	-
T327	-0.1073	0.8301	-	-	-	-	-0.0561	-	-	-
D328	-0.0634	-	-0.4477	-0.1653	-1.7808	0.0216	-1.041	-	0.0054	-

	F334 ( $\alpha$ )		V335 ( $\alpha$ )		F336 ( $\alpha$ )		D337 ( $\alpha$ )		A338 ( $\alpha$ )	
	Unbound	Bound	Unbound	Bound	Unbound	Bound	Unbound	Bound	Unbound	Bound
V34 ( $\beta$ 1)	-	-	-	-	-	-	-	-	-	-
L36 ( $\beta$ 1)	-	-	-	-0.4941	-0.2545	-	-	-	-	-
I49 ( $\alpha$ 1)	-	-	-	-0.1414	-	-	-	-	-	-
Q52 ( $\alpha$ 1)	-	-	-	-	-	-	-	-	-	-
M53 ( $\alpha$ 1)	-	-	-	-0.4012	-0.784	-	-	-	-	-
I55 ( $\alpha$ 1)	-	-	-	-	-	-	-	-	-	-
I56 ( $\alpha$ 1)	-	-	-	-	-	-	-	-	-	-
H57 ( $\alpha$ 1)	-	-	-	-	-0.644	-	-	-	-	-
Q172 ( $\alpha$ F)	-	-	-	-	-	-	-	-	-	-
L175 ( $\alpha$ F)	-	-	-	-	-	-	-	-	-	-
F189 ( $\beta$ 2)	-	-	-	-	-0.5458	-	-	-	-	-
F191 ( $\beta$ 2)	-	-	-	-	-1.2928	-1.003	-0.132	-	-	-
K192	-	-	-	-	-	-1.2049	-0.1929	-	-	-
L194 ( $\beta$ 3)	-	-	-	-	-	-0.7166	-	-	-	-
F196 ( $\beta$ 3)	-	-	-	-0.4847	-0.4841	-0.785	-	-	-	-
M198 ( $\beta$ 3)	-	-	-	-	-0.364	-	-	-	-	-
T219	-	-	-	-	-	-	-	-	-	-
A220 ( $\beta$ 4)	-	-	-	-	-	-	-	-	-	-
I222 ( $\beta$ 4)	-	-	-	-0.6073	-	-	-	-	-	-
D261	-	-	-	-	-	-	-	-	-	-
S263 ( $\beta$ 5)	-	-	-	-	-	-	-	-	-	-
I265 ( $\beta$ 5)	-	-0.058	-	-0.5413	-	-	-	-	-	-0.6674
F267 ( $\beta$ 5)	-	-0.6571	-0.8787	-0.4508	-0.0286	-	-	-	-	-
N269 ( $\beta$ 5)	-	-	-	-	-	-	-	-	-	-
K271 ( $\alpha$ G)	-	-	-	-	-	-	-	-	-	-
D272 ( $\alpha$ G)	-	-	-	-	-	-	-	-	-	-
D315	-	-	-	-	-	-	-	-	-	-
T316	-	-	-	-	-	-	-	-	-	-
K317	-	-	-	-	-	-	-	-	-	-
E318	-	-	-	-	-	-	-	-	-	-
I319 ( $\beta$ 6)	-	-	-	-	-	-	-	-	-	-
T320 ( $\beta$ 6)	-	-	-	-	-	-	-	-0.0003	-	-1.5442
H322 ( $\beta$ 6)	-	-1.3476	-0.8542	-	-	-	-	-	-0.3368	-0.1518
F323 ( $\beta$ 6)	-	-0.003	-0.0501	-	-	-	-	-	-	-
T324	-	-	-0.4425	-	-	-	-	-	-	-
C325	-	-	-	-	-	-	-	-	-	-
A326	-	-	-	-	-	-	-	-	-	-
T327	-	-	-	-	-	-	-	-	-	-
D328	-	-	-	-	-	-	-	-	-	-



	V339 ( $\alpha$ )		T340 ( $\alpha$ )		D341 ( $\alpha$ )		V342 ( $\alpha$ )		B43 ( $\alpha$ )	
	Unbound	Bound	Unbound	Bound	Unbound	Bound	Unbound	Bound	Unbound	Bound
V34 ( $\beta$ 1)	-	-0.6416	-	-	-	-	-	-	-0.7295	-0.0246
L36 ( $\beta$ 1)	-0.1373	-0.2778	-0.0352	-	-	-	-	-	-0.282	-
I49 ( $\alpha$ 1)	-	-	-	-	-	-	-	-	-	-
Q52 ( $\alpha$ 1)	-	-	-	-	-	-	-	-	-	-
M53 ( $\alpha$ 1)	-	-	-	-	-	-	-	-	-	-
I55 ( $\alpha$ 1)	-	-	-	-	-	-	-	-	-	-
I56 ( $\alpha$ 1)	-	-	-	-	-	-	-	-	-	-
H57 ( $\alpha$ 1)	-	-	-	-	-	-	-	-	-	-
Q172 ( $\alpha$ F)	-	-	-	-	-	-	-	-	-	-
L175 ( $\alpha$ F)	-	-	-	-	-	-	-	-	-	-
F189 ( $\beta$ 2)	-	-	-	-	-	-	-	-	-	-
F191 ( $\beta$ 2)	-	-	-0.3792	-	-	-	-	-	-	-
K192	-	-	0.037	-	-0.1	-	-	-	-	-
L194 ( $\beta$ 3)	-	-0.1942	-	0.0048	-	-	-	-	-	-0.1278
F196 ( $\beta$ 3)	-	-0.5994	-0.5965	-	-	-	-	-	-0.4264	-
M198 ( $\beta$ 3)	-	-	-	-	-	-	-	-	-	-
T219	-	-0.4324	-	-	-	-	-	-0.581	-0.5615	-0.0372
A220 ( $\beta$ 4)	-0.0274	-0.2908	-	-	-	-	-	-0.1144	-0.2219	-
I222 ( $\beta$ 4)	-0.2659	-	-	-	-	-	-	-	-	-
D261	-	-	-	-	-	-	-	-	-	-
S263 ( $\beta$ 5)	-	-	-	-	-	-	-0.0026	0.0407	-	-
I265 ( $\beta$ 5)	-1.0743	-0.3327	-	-	-	-	-	-0.026	-	-
F267 ( $\beta$ 5)	-0.2845	-	-	-	-	-	-	-	-	-
N269 ( $\beta$ 5)	-	-	-	-	-	-	-	-	-	-
K271 ( $\alpha$ G)	-	-	-	-	-	-	-	-	-	-
D272 ( $\alpha$ G)	-	-	-	-	-	-	-	-	-	-
D315	-	-	-	-	-	-	-	-	-	-
T316	-	-	-	-	-	-	-	-	-	-
K317	-	-	-	-	-	-	-	-	-	-
E318	-	-	-	-	-	-	-	0.0108	-	-
B319 ( $\beta$ 6)	-	-	-	-	-	-	-	-	-	-
T320 ( $\beta$ 6)	-	-0.0017	-	-	-	-0.3851	-1.2335	-0.406	-	-
H322 ( $\beta$ 6)	-0.2085	-	-	-	-	-	-	-	-	-
F323 ( $\beta$ 6)	-	-	-	-	-	-	-	-	-	-
T324	-	-	-	-	-	-	-	-	-	-
C325	-	-	-	-	-	-	-	-	-	-
A326	-	-	-	-	-	-	-	-	-	-
T327	-	-	-	-	-	-	-	-	-	-
D328	-	-	-	-	-	-	-	-	-	-

	I344 ( $\alpha$ )		K345 ( $\alpha$ )		N346 ( $\alpha$ )		N347 ( $\alpha$ )		K349 ( $\alpha$ )	
	Unbound	Bound	Unbound	Bound	Unbound	Bound	Unbound	Bound	Unbound	Bound
V34 ( $\beta$ 1)	-	-	-	-	-	-	0.014	-	-	-
L36 ( $\beta$ 1)	-	-	-	-	-	-	-	-	-	-
I49 ( $\alpha$ 1)	-	-	-	-	-	-	-	-	-	-
Q52 ( $\alpha$ 1)	-	-	-	-	-	-	-	-	-	-
M53 ( $\alpha$ 1)	-	-	-	-	-	-	-	-	-	-
I55 ( $\alpha$ 1)	-	-	-	-	-	-	-	-	-	-
I56 ( $\alpha$ 1)	-	-	-	-	-	-	-	-	-	-
H57 ( $\alpha$ 1)	-	-	-	-	-	-	-	-	-	-
Q172 ( $\alpha$ F)	-	-	-	-	-	-	-	-	-	-
L175 ( $\alpha$ F)	-	-	-	-	-	-	-	-	-	-
F189 ( $\beta$ 2)	-	-	-	-	-	-	-	-	-	-
F191 ( $\beta$ 2)	-	-	-	-	-	-	-	-	-	-
K192	-0.1979	-	-	-	-	-	-	-	-	-
L194 ( $\beta$ 3)	-0.358	-	-	-	-	-	0.01	-	-	-
F196 ( $\beta$ 3)	-	-	-	-	-	-	-	-	-	-
M198 ( $\beta$ 3)	-	-	-	-	-	-	-	-	-	-
T219	-	-	-	-	-	-0.0282	-0.0519	-	-	-
A220 ( $\beta$ 4)	-	-	-	-	-	-	-	-	-	-
I222 ( $\beta$ 4)	-	-	-	-	-	-	-	-	-	-
D261	-	-	-	-	-	0.0087	-	-	-	0.0533
S263 ( $\beta$ 5)	-	-	-	0.0143	-0.1594	-0.0834	-	-	-	0.0084
I265 ( $\beta$ 5)	-	-	-	-	-	-	-	-	-	-
F267 ( $\beta$ 5)	-	-	-	-	-	-	-	-	-	-
N269 ( $\beta$ 5)	-	-	-	-	-	-	-	-	-	-
K271 ( $\alpha$ G)	-	-	-	-	-	-	-	-	-	-
D272 ( $\alpha$ G)	-	-	-	-	-	-	-	-	-	-
D315	-	-	-	-	-	-	-	-	-	-0.4082
T316	-	-	-	-	-	-	-	-	0.0216	-
K317	-	-	-	0.0095	0.0118	-	-	-	-	-
E318	-	-	-	-0.0187	-	-0.0418	-	-	-	-1.0178
I319 ( $\beta$ 6)	-	-	-	0.0166	-	-	-	-	-	-
T320 ( $\beta$ 6)	-	-	-	-0.1343	-	-	-	-	-	-
H322 ( $\beta$ 6)	-	-	-	-	-	-	-	-	-	-
F323 ( $\beta$ 6)	-	-	-	-	-	-	-	-	-	-
T324	-	-	-	-	-	-	-	-	-	-
C325	-	-	-	-	-	-	-	-	-	-
A326	-	-	-	-	-	-	-	-	-	-
T327	-	-	-	-	-	-	-	-	-	-
D328	-	-	-	-	-	-	-	-	-	-

	C350 (a)		F354 (a)	
	Unbound	Bound	Unbound	Bound
V34 ( $\beta$ 1)	-	-	-	-
L36 ( $\beta$ 1)	-	-	-	-
I49 ( $\alpha$ 1)	-	-	-	-
Q52 ( $\alpha$ 1)	-	-	-	-
M53 ( $\alpha$ 1)	-	-	-	-
I55 ( $\alpha$ 1)	-	-	-	-
I56 ( $\alpha$ 1)	-	-	-	-
H57 ( $\alpha$ 1)	-	-	-	-
Q172 ( $\alpha$ F)	-	-	-	-
L175 ( $\alpha$ F)	-	-	-	-
F189 ( $\beta$ 2)	-	-	-	-
F191 ( $\beta$ 2)	-	-	-	-
K192	-	-	-	-
L194 ( $\beta$ 3)	-	-	-	-
F196 ( $\beta$ 3)	-	-	-	-
M198 ( $\beta$ 3)	-	-	-	-
T219	-	-	-	-
A220 ( $\beta$ 4)	-	-	-	-
I222 ( $\beta$ 4)	-	-	-	-
D261	-	-	-	-
S263 ( $\beta$ 5)	-	-	-	-
I265 ( $\beta$ 5)	-	-	-	-
F267 ( $\beta$ 5)	-	-	-	-
N269 ( $\beta$ 5)	-	-	-	-
K271 ( $\alpha$ G)	-	-	-	-
D272 ( $\alpha$ G)	-	-	-	-
D315	-	-	-	-
T316	-	-	-	-
K317	-0.2477	-	-	-1.2155
E318	-	-	-	-0.3277
I319 ( $\beta$ 6)	-	-	-	-
T320 ( $\beta$ 6)	-	-	-	-
H322 ( $\beta$ 6)	-	-	-	-
F323 ( $\beta$ 6)	-	-	-	-
T324	-	-	-	-
C325	-	-	-	-
A326	-	-	-	-
T327	-	-	-	-
D328	-	-	-	-

**Movies Available here: <http://www.jbc.org/content/289/35/24475/suppl/DC1>**

**SUPPLEMENTARY MOVIE 6.1.** A representative morph of the interaction between the  $\alpha 5$  helix and  $\beta 5$ - $\beta 6$  strands as the alpha subunit transitions from the basal to receptor-bound state. Upon interaction with the activated receptor, the  $\alpha 5$  helix rotates to allow new side chain interactions along the  $\beta$  strands. This rotation effectively alternates one set of hydrophobic side chains on the helix for another, allowing the rotation. Colors of morph are representative of the energies of interaction with darker blue corresponding to a more stable interaction (energy). Secondary structure elements are shown in white while  $\beta$  strands 1-6 are denoted in teal. The receptor,  $G\beta$ , and  $G\gamma$  subunits are deleted for clarity. The initial side chain colors reflect the energy of interaction calculated for the basal state with key stabilizing side chains labeled. These side chain colors shift mid-morph to reflect the calculated energy in the receptor-bound state.

**SUPPLEMENTARY MOVIE 6.2.** A representative morph of the interaction between the  $\alpha 5$  helix and surrounding  $\beta$  strands as the  $G\alpha$  subunit of the heterotrimer transitions from the basal to receptor-bound state. Upon interaction with the activated receptor, the  $\alpha 5$  helix rotates to allow new side chain interactions along the  $\beta$  strands. This rotation effectively alternates one set of hydrophobic side chains on the helix for another, allowing rotation. The color code is same as Supplementary Movie 1.

**SUPPLEMENTARY MOVIE 6.3.** A representative morph of the interaction between the  $\alpha 5$  helix and surrounding  $\beta$  strands looking down the  $\alpha 5$  helix from the perspective of the receptor as the  $G\alpha$  subunit transitions from the basal to receptor-bound state. The color code is same as Supplementary Movie 1.

**SUPPLEMENTARY MOVIE 6.4.** A representative morph of the interaction between the  $\alpha 5$  helix and  $\beta 2$ - $\beta 3$  strands as the  $G\alpha$  subunit transitions from the basal to receptor-bound state. The color code is same as Supplementary Movie 1.

**A CONSERVED HYDROPHOBIC CORE IN GAI1 REGULATES G PROTEIN  
ACTIVATION AND RELEASE FROM ACTIVATED RECEPTOR**

**7.1 Introduction**

***Chapter 7***

This research was originally published in the Journal of Biological Chemistry. Kaya AI, Lokits AD, Gilbert JA, Iverson TM, Meiler J, Hamm HE, “A Conserved Hydrophobic core in G $\alpha$ i1 regulates G protein activation and release from activated receptor” *Journal of Biological Chemistry* **2016** 9;291(37):19674-86 © the American Society for Biochemistry and Molecular Biology.

***Contribution***

I am the second author of this manuscript. I contributed the Rosetta Interface Calculations and the Pairwise Energy Calculation analyses. I contributed all data and text for these sections and created Figure 7.8 and Table 7. I also reviewed and edited all texts contributions from the other authors.

***Abstract***

G protein coupled receptor (GPCR) mediated heterotrimeric G protein activation is a major mode of signal transduction in the cell. Previously, we and other groups reported that the alpha 5 ( $\alpha$ 5) helix of G $\alpha$ i1, especially the hydrophobic interactions in this region, plays a key role during nucleotide release and G protein activation. To further investigate the effect of this hydrophobic core, we disrupted it in G $\alpha$ i1 by inserting 4 alanine amino acids into the  $\alpha$ 5 helix between residues Q333 and F334 (Ins4A). This extends the length of the  $\alpha$ 5 helix without disturbing the  $\beta$ 6- $\alpha$ 5 loop interactions. This mutant has high basal nucleotide exchange activity, yet no receptor-mediated activation of nucleotide exchange. By using structural approaches, we show that this mutant loses critical hydrophobic interactions leading to significant rearrangements of side chain residues H57, F189, F191, and F336; it also disturbs the rotation of the  $\alpha$ 5 helix, and the  $\pi$ - $\pi$  interaction between H57 and F189. In addition, the insertion mutant abolishes G protein release from the activated receptor after

nucleotide binding. Our biochemical and computational data indicate that the interactions between  $\alpha 5$ ,  $\alpha 1$  and  $\beta 2$ - $\beta 3$  are not only vital for GDP release during G protein activation, but they are also necessary for proper GTP binding (or GDP re-binding). Thus, our studies suggest that this hydrophobic interface is critical for accurate rearrangement of the  $\alpha 5$  helix for G protein release from the receptor after GTP binding.

### ***Introduction***

Heterotrimeric G proteins, composed of  $\alpha$ ,  $\beta$ , and  $\gamma$  subunits, act as a molecular switches that turn on intracellular signaling cascades in response to the activation of G protein-coupled receptors (GPCRs) by extracellular stimuli. Therefore, G proteins have a critical role in many different cellular responses<sup>284-287, 325, 326</sup>.

The  $G\alpha$  subunit binds GDP and forms a tight complex with the  $G\beta\gamma$  subunits. Activated GPCRs can catalyze the exchange of GDP for GTP, which leads to the dissociation of the receptor-G protein complex into isolated receptor,  $G\alpha$  and  $G\beta\gamma$  subunits. Both the  $G\alpha$  and  $G\beta\gamma$  subunits can then stimulate or inhibit downstream effectors. Signal propagation ceases after the  $G\alpha$  subunit hydrolyzes GTP, returns to the inactive state, and re-binds to the  $G\beta\gamma$  subunit, regenerating the GDP-bound heterotrimeric state.

Previous studies showed that the activated receptor directly interacts with the G protein by binding to  $G\alpha$ 's C-terminal  $\alpha 5$  helix, inducing a rigid body rotation and translation that pulls this helix into a hydrophobic pocket on the receptor<sup>156, 295</sup>. This leads to the rearrangement of the interfaces between helices  $\alpha 5$ ,  $\alpha 1$  and the  $\beta 2$ - $\beta 3$  strands, and between  $\alpha 5$  and the  $\beta 6$ - $\alpha 5$  loop<sup>156, 286, 293, 301, 327</sup>. Residue F336 in the  $\alpha 5$  helix is highly conserved in small<sup>319, 328</sup> and large GTPases<sup>329</sup> in both the animal and plant kingdoms<sup>1, 62, 302, 330</sup>. Our *in silico* results predicted that F336 is the most energetically important residue both in

maintaining the basal state, and in promoting the receptor bound conformation (6). Our proposed mechanism involves F336 acting as a relay to transmit conformational changes via strands  $\beta 2$  and  $\beta 3$  and helix  $\alpha 1$  to the phosphate-binding loop (5,6). These studies are supported by recently published computational studies<sup>133, 327, 331</sup>. Another critical computational paper from Dror *et al.* used molecular dynamic simulations to suggest that the key events in receptor-mediated G protein activation and GDP release are due to the structural rearrangements of the  $\beta 6$ - $\alpha 5$  loop. This is one of the two identified signal transmission pathways from the receptor to the GDP binding site<sup>332</sup>.

To critically examine the roles of these two possible routes of communication with the nucleotide binding site, we inserted a 4-amino acid linker into the  $\alpha 5$  helix of G $\alpha 1$  between residues Q333 and F334. This insert should disrupt the hydrophobic core (F336, H57, F189, and F191), and mimic the receptor-bound state, while leaving the  $\beta 6$ - $\alpha 5$  loop interactions intact (Figure 7.1A and B). Mutant G $\alpha 1$  subunits were analyzed for their ability to interact with light-activated rhodopsin (R\*), to exchange nucleotides in both the basal and receptor-bound states, and for the structural changes mediated by this insertion. In this study, G $\alpha 1$  was used to replace the visual G protein found in rods, G $\alpha t 1$ . G $\alpha 1$  shows very close homology with G $\alpha t 1$ , is activated by rhodopsin as well as G $\alpha t 1$ <sup>305</sup> and is much more easily expressed in *E. coli*.

Our findings support the role of the hydrophobic interaction between  $\alpha 5$ , the  $\beta 2$ - $\beta 3$  strands and the  $\alpha 1$  helix during activation and nucleotide release. We also uncovered an unexpected dependence on these hydrophobic interactions for promoting G protein release from the receptor-G protein complex.

## 7.2 Materials and Methods

### **Materials**

The TSKgel G2000SW and G3000SW columns, GDP, and guanosine 5'-O-(3-thiotriphosphate) tetralithium salt (GTP $\gamma$ S) were purchased from Sigma. Bodipy GDP and GTP $\gamma$ S were purchased from ThermoFisher Scientific. All other reagents and chemicals were of the highest available purity.

### **Construction, expression and purification of proteins**

In this study, recombinant Gai1 was used for all experiments instead of visual Ga protein (Gat), given that Gai is a very close homolog of Gat, yet is more easily expressed in *E. coli*. Briefly, the pSV277 expression vector encoding Gai1 with an N-terminal His-tag served as the template for introducing amino acid insertions between residues Q333 and F334 by using the QuikChange system (Stratagene). The 4 Ala insertion (Ins4A-Gai1) mutant used primers 5` GTA ACG GAC GTC ATC GCA GCA GCA GCA ATA AAG AAT AAC C 3` (forward) and 5` G GTT ATT CTT TAT TGC TGC TGC TGC GAT GAC GTC CGT TAC 3` (reverse). The Phe-Val-Phe-Asp insertion (Ins4X-Gai1) mutant used primers 5` CG AAG AAT GTG CAG TTT GTG TTC GAT TTT GTG TTC GAT GC 3` (forward) and 5` GC ATC GAA CAC AAA ATC GAA CAC AAA CTG CAC ATT CTT CG 3` (reverse). All mutations were confirmed by DNA sequencing (GenHunter Corporation). The wild type (WT) and the mutant constructs were expressed and purified as previously described<sup>305</sup>. The purified proteins were cleaved with thrombin (Sigma, 0.5 U/mg final concentration) for 16 hours at 4 °C in order to remove the N-terminal His-tag.



The samples were then loaded onto a Ni-NTA column to separate the proteins from the cleaved His-tag and any uncleaved fraction. For further purification, the protein solutions were loaded onto a size-exclusion chromatography (SEC) column (TSKgel G3000SW) that was equilibrated in buffer A [50 mM Tris-HCl (pH 7.4), 100 mM NaCl, 1 mM MgCl<sub>2</sub>, 20 μM GDP (or 1 μM GTPγS), 1 mM DTT and 100 μM PMSF]. SDS-PAGE was used to test the purity of the proteins. Urea-washed rod outer segment membranes (ROS) containing dark-adapted rhodopsin and Gβ1γ1 subunits were prepared as previously described<sup>303, 333</sup>. Protein concentrations were determined spectroscopically<sup>333</sup> and by Bradford assay<sup>306</sup>.

#### ***Preparation of urea-washed ROS membranes and Gβ1γ1***

Urea-washed ROS membranes and Gβ1γ1 were prepared from bovine retina as described previously<sup>303, 304</sup>.

#### ***Nucleotide-exchange assays***

The basal rate of GTPγS binding was determined by monitoring the relative increase in the intrinsic tryptophan (W211) fluorescence ( $\lambda_{\text{ex}}=290$  nm,  $\lambda_{\text{em}}=340$  nm) of Gαi1 (200 nM) in buffer containing 50 mM Tris (pH 7.2), 100 mM NaCl and 1mM MgCl<sub>2</sub> for at least 60 min at 21 °C after the addition of 10 μM GTPγS. Receptor-mediated nucleotide exchange was determined with Gβ1γ1 (400 nM) in the presence of 50 nM rhodopsin at 21 °C for 60 min after the addition of GTPγS. The data were normalized to the baseline and maximum fluorescence and then fit to the exponential association equation ( $Y = Y_{\text{max}} * (1 - e^{-kt})$ ), to calculate the rate constant (k) as previously described<sup>334</sup>. For nucleotide-exchange experiments with Bodipy nucleotides, the fluorophore was monitored at  $\lambda_{\text{ex}}=490$  nm and  $\lambda_{\text{em}}=510$  nm with 5 nm slit widths as described<sup>335</sup>. All experiments were performed in a buffer containing 50 mM Tris (pH 7.2), 100 mM NaCl and 1mM MgCl<sub>2</sub> and 1mM DTT at

21 °C. To measure the GDP release from the G protein, the G $\alpha$ 1 subunit was incubated with Bodipy-GDP in the absence of unlabeled GDP, G $\beta\gamma$  subunit or receptor for 90 min at room temperature to exchange GDP with Bodipy nucleotide. After 1.5 hours, a two-fold excess of G $\beta\gamma$  was added and incubated for 15 min to suppress the nucleotide exchange. Bodipy-GDP bound heterotrimeric G protein was recorded as the basal signal. After 2.5 min, light activated receptor was added to the quartz cuvette. To measure the Bodipy-GTP $\gamma$ S binding, heterotrimeric G proteins were incubated with in buffer containing 50 mM Tris (pH 7.2), 100 mM NaCl and 1mM MgCl<sub>2</sub> in the presence of labeled GTP $\gamma$ S to obtain the basal signal; After 2.5 min, activated receptor was added to initiate the exchange reaction. The kinetic data were plotted and fit to a one phase association function. Data represent the averages from 8-10 experiments.

### ***Protein Labeling***

A cysteine-reduced G $\alpha$ 1 protein (C3S, C66A, C214S, C305S, C325I, C351I) was labeled as described previously using a 10 molar fold excess of Alexa Fluor 594C5-maleimide (A1) (Invitrogen), with a labeling time of 3-5 hours in in 50 mM Tris (pH 7.5), 100 mM NaCl, 1 mM MgCl<sub>2</sub> and 20  $\mu$ M GDP<sup>334</sup>. Proteins were purified via size-exclusion purification and the fractions were screened by intrinsic Trp fluorescence to ensure the functional integrity of the labeled proteins. Labeling efficiency was determined from comparison of A<sub>580</sub> to protein concentration, as determined by Bradford, and found to be between 0.5-0.75 mol label/mol protein, depending on the location of the residue<sup>334, 335</sup>.

### ***Membrane binding assay***

The ability of mutant G $\alpha$ 1 subunits to bind R\* in urea-washed ROS membranes was determined as previously described<sup>156</sup>. Each sample was evaluated by comparison of the

amount of G $\alpha$ i1 subunit within the pellet (P) or supernatant (S) to the total amount of G $\alpha$ i1 subunit (P+S) in both treatments expressed as a percentage of the total G $\alpha$ i1 protein. Data represents the average of at least five experiments.

***Protein crystallization, data collection and structure determination***

Purified GTP $\gamma$ S-bound Ins4A-G $\alpha$ i1 subunits were exchanged into crystallization buffer (50 mM HEPES (pH 8.0), 1 mM EDTA, 10 mM MgSO<sub>4</sub>, 5 mM dithiothreitol (DTT), 20  $\mu$ M GTP $\gamma$ S) using a TSKgel G3000SW SEC column. Appropriate fractions were pooled as described above, and SDS-PAGE was used to assess the purity of the proteins. Crystals were grown using the hanging drop vapor diffusion method at 21 °C by equilibrating a 1:1 ratio of protein (10 mg/mL in crystallization buffer) and reservoir solution (12-16% polyethylene glycol (PEG) 2000 monomethyl ether, 18% 2-propanol and 100 mM MES (pH 6.0)) against a reservoir solution. Crystals appeared after 15 days and grew in the primitive monoclinic space group P2<sub>1</sub>.

For Ins4A-G $\alpha$ i1 $\beta$ 1 $\gamma$ 1 crystallization, separately purified and concentrated Ins4A-G $\alpha$ i1 and WT G $\beta$ 1 $\gamma$ 1 subunits were mixed in a 1:1 molar ratio and incubated for 30 min at 25 °C. The heterotrimeric G protein complex was purified away from uncomplexed subunits using a G3000SW SEC column equilibrated with buffer containing 20 mM HEPES (pH7.5), 150 mM NaCl, 1 mM EDTA, 5 mM DTT, and 200  $\mu$ M GDP. Appropriate fractions were pooled and the post translational palmitoylation of the G $\beta$ 1 $\gamma$ 1 subunit was removed by incubating with 10 units of endoproteinase Lys-C in 50 mM Tris (pH7.5, 150 mM NaCl) for 24 hours at 4 °C<sup>208</sup>. The protein complex was subjected to an additional step of size exclusion chromatography using a G3000SW column, as described above. Fractions were analyzed by SDS-PAGE to provide a guide to appropriate pooling of the purified

heterotrimer. Heterotrimeric complex crystallized using the hanging drop vapor diffusion method at 21 °C by equilibrating the protein (10 mg/mL in 20 mM HEPES (pH7.5), 150 mM NaCl, 200  $\mu$ M GDP, 1mM EDTA) in a 1:1 ratio with reservoir solution against a reservoir solution containing 19-24% PEG 8000, 1-5% 2-propanolol, 1% OG, 100 mM HEPES (pH 7.0) and 100 mM NaOAc (pH 6.4). Crystals appeared after 5 days and grew in the primitive tetragonal space group P4<sub>3</sub>.

Both Ins4A-G $\alpha$ i1-GTP $\gamma$ S and Ins4A-G $\alpha$ i1 $\beta$ 1 $\gamma$ 1 crystals were cryo-protected prior to data collection by briefly soaking in stabilization solution containing 18% glycerol and cryo cooled by plunging into liquid nitrogen. Data sets were collected at the LS-CAT (21-ID-G) of the Advanced Photon Source (APS) at Argonne National Laboratory at -180 °C using a wavelength of 0.98 Å on a MAR CCD detector. Data were scaled using the HKL2000<sup>336</sup>, truncated and converted using CCP4<sup>311</sup> and processed using Phenix suites<sup>312</sup>. Crystallographic data collection and refinement statistics are reported in Table 7.2. Criteria for data cutoffs were a combination of R<sub>sym</sub> and I/ $\sigma$  which both rose to unacceptable levels if the resolution were extended for either dataset. The structures of Ins4A-G $\alpha$ i1-GTP $\gamma$ S·Mg<sup>+2</sup> and Ins4A-G $\alpha$ i1 $\beta$ 1 $\gamma$ 1-GDP complexes were determined by molecular replacement using 1GIA (WT G $\alpha$ i1-GTP $\gamma$ S ·Mg<sup>+2</sup>)<sup>212</sup> and 1GP2 (WT G $\alpha$ i1 $\beta$ 1 $\gamma$ 2-GDP)<sup>337</sup> as search models for Phaser-MR<sup>338</sup> in the Phenix suite<sup>312</sup>. Since PDB entries 1GIA and 1GP2 were deposited prior to the requirement for deposition of structural factors, R-free reflections were randomly selected for Ins4A-G $\alpha$ i1-GTP $\gamma$ S·Mg<sup>+2</sup> and Ins4A-G $\alpha$ i1 $\beta$ 1 $\gamma$ 1-GDP. As a result, the free-R is of limited utility. Model building was performed in Coot<sup>315</sup> using composite omit maps calculated in Phenix<sup>312</sup> to minimize model bias. Refinement was conducted using both Refmac<sup>339</sup> and Phenix<sup>312</sup>, with the final

rounds of refinement performed using Phenix <sup>312</sup>. In the final model, the regions corresponding to amino acids 1-33 and 348-354 (correspond to WT numbers) in Ins4A-G $\alpha$ 1-GTP $\gamma$ S·Mg<sup>+2</sup> are not included. Similarly in the Ins4A-G $\alpha$ 1 $\beta$ 1 $\gamma$ 1-GDP structure, residue numbers 1-6 and 346-354 in the G $\alpha$ 1 subunit; 1 and 129-132 in the G $\beta$ 1 subunit; 1-9 and 66-74 in the G $\gamma$ 1 subunit are not included due to the lack of interpretable electron density. Structural superimpositions were performed using C $\alpha$  atoms and the program Superpose in the CCP4 suite <sup>317, 318</sup>. All structural figures were made using PyMOL (PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.) unless otherwise indicated.

### ***Rosetta interface energy calculations***

Interface energies were computed following the Rosetta  $\Delta\Delta$ G protocol previously described <sup>277, 340</sup>. Briefly, we re-relaxed the previously published ensembles of ten structures of the G protein in the basal state and receptor bound state using a DualSpace relax Rosetta protocol <sup>341</sup> for consistency between the mutant structures and the *in silico* models <sup>340</sup>. The G $\gamma$  subunit of the receptor-bound models were truncated along both the N- and C- termini to match the available crystal density of the mutant structures. For the Ins4A insertion protein, all residues present with crystallographic density were included in the analyses. The Ins4A-G $\alpha$ 1 and Ins4A-G $\alpha$ 1 $\beta$ 1 $\gamma$ 1 structures were relaxed using the same DualSpace relax protocol in Rosetta. Residue-residue interactions across the  $\alpha$ 1 helix/GTPase domain interface were evaluated by measuring the changes in energetic perturbations when computationally removing the  $\alpha$ 1 helices from the models. The  $\alpha$ 1 helix was defined as residues G45 to E58. For all analyses, GDP · Mg<sup>2+</sup> or GTP $\gamma$ S remained positioned within the nucleotide binding pocket. The predicted  $\Delta\Delta$ G value is reported as an

average over the ten best structural models in Rosetta Energy Units (REU). Absolute values larger than 0.5 REU are considered to be significant. Using the standard deviation over the ten structures, a Z-score was computed. The total  $\Delta\Delta G$  value across the interface is calculated as the sum of individual residue contributions.

### ***Rosetta pairwise interaction score calculations***

Average interaction scores between pairwise interacting residues were computed using Rosetta's per residue energy-breakdown protocol as previously described<sup>340</sup>. The strength between all possible pairs of interacting amino acid residues within the G protein were calculated across the previously published ensembles of ten structures after an initial energy minimization using the DualSpace relaxation protocol<sup>341</sup>. Mutant crystal structures of Ins4A-G $\alpha$ 1 and Ins4A-G $\alpha$ 1 $\beta$ 1 $\gamma$ 1 also underwent an initial round of DualSpace relax using Rosetta to relieve minor energetic clashes, in both torsional and Cartesian space, induced by crystallization before calculations were conducted. The resulting predicted interaction scores, between all residue pairs, was then averaged across the top ten scoring models (as assessed by total Rosetta energy) in the basal state, receptor-bound state, and the activated monomeric state. Predicted values are reported in Rosetta Energy Units (REU) and considered significant if greater than 0.5 REU. While these scores are also reported in REUs, they are *not* free energies in the thermodynamic sense. We therefore call these values pairwise 'interaction scores' for intra-molecular probing of information flow.

### 7.3 Results

#### *Biochemical characterization and functional properties of Ins4A-Gai1 protein*

To examine the two activation routes of G protein activation, we inserted 4 alanines between residues Q333 and F334 of the  $\alpha 5$  helix, with this variant termed Ins4A (Figure 7.1A and 1B). This insertion is proposed to perturb the interactions between the critical F336 and both the  $\alpha 1$  helix and the  $\beta 2$ - $\beta 3$  strands while leaving the  $\beta 6$ - $\alpha 5$  loop intact. We tested how this insertion, which should mimic the rotation of the  $\alpha 5$  helix toward the receptor in the R\*-G $\alpha\beta\gamma$  complex, affects both the critical structural interactions between  $\alpha 5$  and  $\alpha 1$  and  $\beta 1$ - $\beta 3$  and the functions of basal and receptor-mediated nucleotide exchange rates.

Ins4A displayed a highly increased basal exchange rate as monitored by the relative increase in the intrinsic tryptophan (W211) fluorescence of Gai1 compared to WT protein (Figure 7.2A, light grey). However, in receptor mediated activation, the Ins4A mutant showed a significantly decreased nucleotide exchange rate compared to WT (dark grey).

One potential explanation for these data would be that the mutant does not interact with the receptor properly. To test this idea, we conducted a membrane binding assay with light-activated rhodopsin in rod outer segments (ROS). The data show normal levels of Ins4A interaction with R\* and the capability to bind ROS membrane as well as the WT protein (Figure 7.2B). However, the addition of the GTP $\gamma$ S non-hydrolyzable nucleotide analog does not induce disassociation of the complex even at high concentration (0.5  $\mu$ M) (Figure 7.2B, black arrows). We repeated this experiment in the presence of 1 mM GDP, and once again the mutant did not release from the ROS membrane. Densitometric calculations of membrane binding show that the mutant is not responsive to nucleotide (Figure 7.2C).

Accordingly, an alternative possibility is that the Ins4A mutant might not properly dock

its C-terminus to R\* to transmit the activation signal to the nucleotide binding region. Using extra-Meta II (eMII) to measure the high affinity state of the receptor shows that there is normal eMII induced by increasing concentrations of heterotrimeric Gi binding (Figure 7.2D), implying normal interaction between the  $\alpha 5$  C-terminal helix and active receptor. Thus the ability of Gi to induce a high affinity state was similar between WT and Ins4A mutant (Figure 7.2D). To confirm the nucleotide sensitivity in the membrane binding experiment (Fig. 2B), the eMII assay was repeated in the presence of a high concentration of GDP (0.5mM). Even this high concentration of GDP did not inhibit eMII in Ins4A, though it did effectively inhibit it in the WT protein (compare Figure 7.2D and E). This result confirms the membrane binding results and also shows that the C-terminus of Ins4A properly interacts with and induces the high affinity state of R\* similar to WT.

#### ***Guanine nucleotide interactions with Ins4A protein***

There are several scenarios that might explain how the Ins4A protein could bind the receptor with similar affinity to WT, yet lack receptor-mediated nucleotide exchange or nucleotide-dependent membrane release activity (Figure 7.3A). 1) The helical domain opening does not take place properly so GDP cannot release. 2) The  $\beta 6$ - $\alpha 5$  loop does not properly trigger GDP release as suggested by *Dror et al.*<sup>332</sup>. 3) GDP can release normally but GDP, GTP or GTP $\gamma$ S cannot rebind to the empty nucleotide-binding pocket. 4) Nucleotide exchange happens normally, but the G protein cannot release from the receptor.

To distinguish between these possibilities, we measured receptor-mediated GDP release and GTP $\gamma$ S binding using Bodipy-labeled nucleotides. To measure GDP release from the G protein, the G $\alpha$  subunit was incubated with Bodipy-GDP, then G $\beta\gamma$  was added as described in the methods section. After 2 min, light-activated rhodopsin was added (Fig. 3B, first



arrow). The data show that WT *Gai1*- $\beta 1\gamma 1$  releases labeled GDP very quickly after interaction with R\* (Figure 7.3B, black circles), while Ins4A- $\beta 1\gamma 1$  releases GDP almost 100-fold more slowly (Figure 7.3B, grey trace, Table 7.2). The bodipy-GDP dissociation rate constants were calculated to be  $\sim 3.52 \text{ min}^{-1}$  and  $0.042 \text{ min}^{-1}$  for WT and the Ins4A mutant, respectively. To test if GDP was still able to access the nucleotide binding region, we added excess unlabeled GDP and monitored Bodipy-GDP release. Unlabeled GDP can compete with the Bodipy nucleotide (Figure 7.3B, second arrow). Bodipy-GDP release was faster in the presence of unlabeled GDP (dissociation rate,  $\sim 0.755 \text{ min}^{-1}$ ); this is likely due to the affinity difference between these two GDP nucleotides.

GTP $\gamma$ S binding was also monitored by using Bodipy-GTP $\gamma$ S (Figure 7.3C). Like GDP release, the Ins4A insertion mutant also affects GTP $\gamma$ S binding. The data show that labeled GTP $\gamma$ S interaction with the mutant was approximately 30-fold slower than with the WT protein (Table 7.1); the binding rate reflects GDP release as well as labeled GTP $\gamma$ S interaction. The GTP $\gamma$ S binding rate constants were calculated to be  $0.913 \text{ min}^{-1}$  and  $0.031 \text{ min}^{-1}$  for WT and the Ins4A insertion mutant, respectively. These results indicate that the insertion of an extra helical turn in  $\alpha 5$  dramatically affects receptor-mediated GDP release; however, GDP can still be released from the nucleotide binding pocket, and both GDP and GTP $\gamma$ S can access the pocket.

***Examination of conformational changes in functionally important regions mediated by receptor and GTP $\gamma$ S.***

In order to examine local environmental changes within specific regions of the  $G\alpha$  subunit, we used a *Gai1* protein lacking six solvent-exposed cysteines as a background for the introduction of cysteine residues at sites of interest. We selected three positions in the

G $\alpha$ i1 subunit that are critical for G protein function<sup>156</sup>. L273 (L296 in G $\alpha$ s) is a sensor of the presence of the guanine ring of guanine nucleotides, K349 (R389 in G $\alpha$ s) is a sensor of receptor binding, and K330 (E370 in G $\alpha$ s) senses rotation and disorder in  $\alpha$ 5 in the presence of R\*<sup>295</sup> (Figure 7.4A and 7.4C). These positions were mutated to Cys and labeled with the Alexa Fluor 594C5-maleimide probe. The fluorescent intensity was measured after a 40 min incubation with either GDP, GTP $\gamma$ S, receptor or receptor plus GTP $\gamma$ S. Each result was normalized to the fluorescence of its wild type G protein (Figure 7.4D, black bars). To determine the relative changes in those regions in the basal state, we compared the fluorescence intensity in GDP- and GTP $\gamma$ S-bound states.

The fold change in emission intensity of Ins4A in the presence of GDP (black bars) or GTP $\gamma$ S (grey bars) with the indicated labeled residues, as compared to the environment of the same labeled residue in WT are shown in Figure 7.4B. The extreme C terminal region (K349) showed relatively low fluorescence intensity compared with the WT protein in both GDP and GTP $\gamma$ S bound states, which indicates a highly polar environment. This highly polar environment might be due to the more exposed location induced by the extra 4 alanine residues in the  $\alpha$ 5 helix. Other mutants were similar to WT.

Next, we evaluated the conformational changes of the same regions in the heterotrimeric G protein (black bars) in the presence of active receptor (grey bars) and after addition of GTP $\gamma$ S (black shaded bars) (Figure 7.4D). The decreased emission intensity from labeled L273 upon receptor activation indicates an increased polar environment for the probe in both WT and the insertion mutant, consistent with the effect of nucleotide release from the binding pocket after receptor interaction. After GTP $\gamma$ S incubation, the fluorescence intensity came back to its GDP bound level in both proteins, indicating

nucleotide binding and domain closing.

Residue K330 is located at the beginning of the  $\alpha 5$  helix; it senses rotation of the helix<sup>156</sup> and disorder in presence of active receptor<sup>295</sup>. The local environment of this residue indicated low solvent exposure in both WT and Ins4A after receptor interaction, indicating that it establishes new contact interactions that were absent in the heterotrimeric structure (Figure 7.4D). These results are consistent with previous EPR studies<sup>156</sup>. However, unlike WT, the mutant fluorescence intensity did not fully return to its heterotrimeric state after GTP $\gamma$ S incubation, indicating a perturbation in this region.

The extreme C terminus of G $\alpha$  is disordered or absent in most crystal structures of isolated G $\alpha$  or the G $\alpha\beta\gamma$  heterotrimer<sup>208, 212, 307, 337</sup>. It is a known receptor contact site that undergoes a receptor- mediated conformational change. Comparison of the fluorescence intensity of the Alexa Fluor label inserted in the C-terminal region at K349 in wild type versus the Ins4A-G $\alpha$ i1 $\beta$ 1 $\gamma$ 1 suggests that this residue is in a similar environment before receptor activation. Upon binding to the light-activated rhodopsin, the fold change in intensity indicates an immobilization of the probe for both wild type and Ins4A-G $\alpha$ i1 $\beta$ 1 $\gamma$ 1, consistent with the expected interactions at the receptor-G protein interface (Figure 7.4D, right panel). As expected, the strong fluorescence intensity of K349 disappeared in the GTP $\gamma$ S-bound WT G protein (Figure 7.4D), while in the mutant the signal did not change, consistent with the membrane binding and eMII results (Figure 7.2B).

***Amino acid identity and the hydrophobic core is important for rearrangement of the  $\alpha 5$  helix after nucleotide binding***

To test if the functional properties of the Ins4A protein are due to the longer  $\alpha 5$  helix, or due to disruption of the hydrophobic core, we replaced the four alanine insertion of Ins4A with a duplication of the four adjacent wild type residues (from F334 to D337),

termining the variant Ins4X (Figure 7.1B). This change reestablishes the hydrophobic core around F336 ( $\alpha 5$ ) in the presence of an insertion while altering the length of  $\alpha 5$  to be the same as Ins4A. To investigate the function of the Ins4X protein, we evaluated its nucleotide exchange rates and membrane binding properties. Unlike Ins4A, Ins4X exhibited basal and receptor-mediated nucleotide exchange rates (Figure 7.5A) and membrane binding (Figure 7.5B) similar to wild type  $G\alpha i1\beta 1\gamma 1$ . The Ins4X protein dissociated from ROS membrane after incubation with active receptor and  $GTP\gamma S$ , similar to WT (Figure 7.2B). As shown in Figure 7.5B, unlike Ins4A, Ins4X released from the ROS membrane completely after incubation with nucleotide. This result suggests that the effect of Ins4A on G protein function is not due to the increase in length of the C terminus. Instead, it suggests that the amino acid identity and the establishment of the hydrophobic core play critical roles for proper rearrangement of the  $\alpha 5$  helix and  $G\alpha$  subunit release from the receptor after nucleotide binding.

#### ***X-ray structures of the Ins4A mutant***

To probe the structural basis for the biochemical properties of the Ins4A variant, the crystal structure in the  $GTP\gamma S$ -bound state was determined at 2.7 Å resolution (Table 2). After insertion of the four alanines between Q333 and F334, the  $\alpha 5$  helix rotates  $\sim 60^\circ$  starting from the insertion point (Figure 7.6A, labeled with red). This rotation relocates F336 to a position similar to that observed for the homologous residue (F376) in the  $\beta_2$ -adrenergic receptor ( $\beta 2AR$ )-Gs complex structure<sup>295</sup> (Figure 7.6B, compare WT- $GTP\gamma S$ , brown, Ins4A, cyan,  $\beta 2AR$ -Gs, green (PDB entry 3SN6<sup>295</sup>). Though attempted, we could not crystallize either the GDP-bound or nucleotide-free Ins4A protein.

The 4 Ala insert completely repositions the network of interactions between F336 ( $\alpha 5$ ),

F189, F191, F196 ( $\beta 2$ - $\beta 3$ ) and H57 ( $\alpha 1$ ). It also disturbs the  $\pi$ - $\pi$  interaction between H57 and F189 (Figure 7.6C and 7.6D). In Ins4A, almost the entire  $\beta 2$ - $\beta 3$  strands move away from the  $\alpha 5$  helix compared to the WT structure (Figure 7.6C and 7.6D). The relative C $\alpha$  distances between insertion mutant and WT proteins in F189, F191, K192 and F196 are 1.5, 2.3, 3.8 and 1 Å, respectively, while the overall root-mean square (RMS) deviation between WT Gai1 and Ins4A was 0.79 Å (304 C $\alpha$  atoms aligned totally).

Crystallized Ins4A has GTP $\gamma$ S bound and the guanine nucleotide holds the GTPase, and helical domains together in the structure. Therefore, we did not expect to see any significant differences between the WT and mutant structure in the nucleotide contact regions. However, we identified an interesting feature in this structure. In the structure of Ins4A, the side chain of H57 (localized on the end of the  $\alpha 1$  helix) flips from pointing inside to outside of the core, probably due to the lost network of interactions between F336 ( $\alpha 5$ ), F189 ( $\beta 2$ ) F191 ( $\beta 2$ ). The relative C $\alpha$  distances in the  $\alpha 1$  helix, residues I55, I56, H57, and E58 are 0.5, 0.9, 1.3 and 1.2 Å, respectively, between the Ins4A insertion mutant and the WT protein, with the end of the  $\alpha 1$  helix moving away from  $\alpha 5$ . This structural rearrangement of the end of the  $\alpha 1$  helix and H57 were predicted in our Rho-Gi complex model (Figure 7.6F, grey) <sup>133</sup>. The  $\beta 2$ AR-G $\alpha$ s complex structure (Figure 7.6E), is lacking the end of the  $\alpha 1$  helix.

#### ***Structural features of Ins4A- $\beta 1 \gamma 1$ mutant***

In the  $\beta 2$ AR-Gs complex crystal structure <sup>295</sup>, the G $\beta 1 \gamma 2$  subunit does not make any contact with the receptor and does not undergo statistically significant conformational changes upon complex formation; though because of the low resolution of that structure, some real changes might not have been statistically significant. To evaluate any possible

role of the G $\beta\gamma$  subunits in the biochemical properties seen in the Ins4A mutant, we determined the crystal structure of the heterotrimeric Ins4A- $\beta 1\gamma 1$  mutant in the GDP bound state to 1.9 Å resolution (Table 7.2 Figure 7.7A). The Ins4A- $\beta 1\gamma 1$  structure shows a similar  $\alpha 5$  helix rotation pattern as the isolated Ins4A bound to GTP $\gamma$ S (Figure 7.7A, teal). However, there was no dramatic displacement of the  $\alpha 1$  helix and  $\beta 2$ - $\beta 3$  regions (Figure 7.7A). The relative C $\alpha$  distances between mutant (Figure 7.7A, teal) and WT (yellow) heterotrimeric structures in H57, F189, F191, K192 and F196 residues are 0.5, 0.5, 1.4, 1.4 and 0.6 Å, respectively. This might be due to the effect of the crystal packing. Figure 7.7B and 7.7C shows that the  $\alpha 5$  helix,  $\beta 2$ - $\beta 3$  strands and  $\alpha 1$  helix interact with a symmetric molecule of the G $\beta 1\gamma 1$  subunit which might block or limit the displacement of the  $\beta 2$ - $\beta 3$  strands and  $\alpha 1$  helix. The Ins4A- $\beta 1\gamma 1$  heterotrimeric structure also shows significant differences at the  $\alpha N$  (Figure 7.7D) and  $\alpha 2$  helices and the G $\beta 1\gamma 1$  subunits (Figure 7.7E) compared to the WT structure. There is not any direct interaction between the 4 Ala insertion site and these regions. Therefore, these structural differences might be allosteric effects of the insertion region. Another possibility is that all three heterotrimeric structures, Gat $\beta 1\gamma 1$ , Gai1 $\beta 1\gamma 2$  and Ins4A-Gai1 $\beta 1\gamma 1$ , contain the same G $\beta$  but different G $\alpha$  and G $\gamma$  subtype combinations, which might affect the heterotrimeric structures in specific regions. The RMS deviation of the G $\alpha$  subunit and heterotrimeric structure between WT- $\beta 1\gamma 2$  (PDB entity, 1GP2<sup>337</sup>) and Ins4A- $\beta 1\gamma 1$  was 0.82 Å and 1.2 Å (with a total of 329 and 697 C $\alpha$  atoms aligned respectively).

#### ***The effect of the 4 alanine insertion on $\alpha 1$ helix interface binding energy***

To investigate the effect of the extra helical turn of  $\alpha 5$  on the  $\alpha 1$  helix computationally, we calculated interaction energy scores for all residues within the  $\alpha 1$  helix in both the Gai1

monomer and heterotrimeric G $\alpha$ 1 $\beta$ 1 $\gamma$ 1 proteins using an established protocol<sup>133</sup> (Table 7.3). These  $\Delta\Delta G$  values probed for a potential network of intramolecular interactions which could propagate the conformational changes necessary for G protein activation and nucleotide exchange. The  $\Delta\Delta G$  calculations predicted and support the crystallographic data.

We did not see any major differences at the N terminus of the  $\alpha$ 1 helix, K46-I49, compared with the WT protein structure. However, starting from V50, significant differences were identified between mutant and WT proteins. The predicted  $\Delta\Delta G$  values of Q52, M53, I56 and H57 residues, which play a major role in interaction with and stabilization of the  $\alpha$ 5 helix in GDP-bound state, were decreased compared to WT<sup>133, 327, 331, 342</sup>. The total interaction energy score was approximately 4 Rosetta Energy Units (REUs) in the Ins4A compared to 5.5 REUs in the WT protein (Table 7.3).

There are two critical stabilizing routes between the  $\alpha$ 1 and  $\alpha$ 5 helices in the GDP-bound state. To look at the individual residue-residue interactions, and distinguish between these two pathways, we used Rosetta to predict the network energy scores between all amino acid pairs in our structural models and protein crystals. The first route is between Q52 ( $\alpha$ 1) and I56 ( $\alpha$ 1) with T329 ( $\alpha$ 5). Previously, Kapoor *et al.* showed that the T329A mutation causes high G $\alpha$ 1 activity<sup>283</sup>. The pairwise interaction scores were calculated between Q52 ( $\alpha$ 1) - T329 ( $\alpha$ 5) and I56 ( $\alpha$ 1) - T329 ( $\alpha$ 5) as 0.5 and 0.2 REU, respectively. The second pathway is between M53 ( $\alpha$ 1) and H57 ( $\alpha$ 1) with V332 ( $\alpha$ 5) F336 ( $\alpha$ 5), a part of the hydrophobic core between  $\alpha$ 5,  $\alpha$ 1 and  $\beta$ 2- $\beta$ 3 strands (Figure 7.8). The structural rearrangement at the end of the  $\alpha$ 1 helix also affects linker 1 and the beginning of the  $\alpha$ A helix. The  $\Delta\Delta G$  values calculated at G60 (linker1) decreased from 0.7 to under 0.5; at Y61

it is changed from 1.5 to 1.0 REU compared to WT protein. In E65 ( $\alpha$ A), it increased from 0.5 to 0.8 REU, as it approaches linker1. In the heterotrimeric structures, we observed a similar pattern between Ins4A and WT, but with smaller margins.



## 7.4 Discussions

Two receptor-mediated G protein activation routes have been hypothesized. In the first, binding of the receptor to the C-terminus of G $\alpha$  is thought to trigger conformational changes that can be transmitted to the nucleotide-binding pocket via outward rotation and translation of the  $\alpha 5$  helix and distortion of the  $\beta 6$ - $\alpha 5$  loop, a key site of interaction with the guanine ring<sup>294, 295, 298, 332, 343</sup>. In the second pathway, the receptor-dependent  $\alpha 5$  rotation and translation destabilizes the hydrophobic interactions between the  $\alpha 5$  and  $\alpha 1$  helices and the  $\beta 2$ - $\beta 3$  strands, which weaken both phosphate and purine binding sites of nucleotide<sup>283, 301, 327, 331, 342</sup>. In the two proposed activation pathways, the extreme C-terminus of the  $\alpha 5$  helix facilitates both receptor-G protein interaction and G protein activation<sup>284, 293, 295, 323, 344</sup>. To separate these two pathways and to further investigate the effect of the hydrophobic core between  $\alpha 5$ ,  $\alpha 1$  and  $\beta 2$ - $\beta 3$  strands, we inserted a 4 Ala linker between Q333 and F334 in the  $\alpha 5$  helix.

Our data show that the Ins4A mutant caused high basal nucleotide exchange, as anticipated from previous studies<sup>293, 301, 327</sup>. The Ins4A-GTP $\gamma$ S crystal structure showed that, starting from Q333, the  $\alpha 5$  helix is displaced by an extra helical turn, which partially mimics the effect of the receptor on the G protein. Indeed, F336 of the  $\alpha 5$  helix, which we previously showed was a critical residue for forming a hydrophobic core in the G $\alpha$  subunit, is localized at a similar position as it is in the  $\beta 2$ AR-Gs complex structure.

The  $\alpha 5$  helix is protected and surrounded with mostly hydrophobic interactions by six beta strands ( $\beta 1$ - $\beta 6$ ) and one alpha helix ( $\alpha 1$ ). The effects of  $\alpha 5$  helix rotation on the  $\beta$  strands are clearly observed in the Ins4A structure compared to WT protein. The relative positions of the  $\beta 5$  and  $\beta 6$  strands are not affected by the rotation, and these two strands

almost perfectly superimpose with the WT structure. However, there are significant and progressive differences in the  $\beta$ -strands amino-terminal to  $\beta 4$ . This is most dramatically observed in the  $\beta 2$ - $\beta 3$  strands. This rotation completely repositions the network of interactions between F336 ( $\alpha 5$ ), F189, F191, F196 ( $\beta 2$ - $\beta 3$ ) and M53, H57 ( $\alpha 1$ ), including disturbing the  $\pi$ - $\pi$  interaction between H57 and F189. The conformational changes in this region mimic the receptor-bound state<sup>258, 301, 327, 331, 342</sup>. This result supports the second route of G protein activation (see above) which was proposed in our previous study<sup>327</sup> and was recently supported by Flock *et al.* and Sun *et al.* via using evolutionary analysis and alanine scanning approaches, respectively<sup>331, 342</sup>.

In the  $\beta 2$ AR-Gs complex structure, the  $\alpha 1$  helix, starting from M53 (M60 in Gas), is not ordered<sup>295</sup>. In the Rhodopsin-Gi complex model, it was predicted that the end of the  $\alpha 1$  helix would move away from  $\alpha 5$ , and most of the residues (from Q52 to H57) would lose contact with the  $\alpha 5$  helix after GDP release and helical domain opening<sup>133</sup>. The Ins4A mutant structure confirmed this prediction, even though we could only crystallize the GTP $\gamma$ S-bound state, which holds the GTPase and helical domains together. Given that it is GTP $\gamma$ S-bound, significant differences between WT and mutant structures in the nucleotide contact regions, such as P- and  $\beta 6$ - $\alpha 5$ -loops, were not expected. However, it appears that the reorganization between the  $\alpha 5$  helix and  $\beta 2$ - $\beta 3$  strands is enough to trigger the  $\alpha 1$  rearrangement even though the  $\beta 6$ - $\alpha 5$  loop and first helical turn of the  $\alpha 5$  helix are still intact.

In contrast to its high constitutive activity in the basal state, the Ins4A mutant showed very little receptor-mediated nucleotide exchange activity. This, we believe, is due to the effect of the G $\beta\gamma$  subunit. In the basal state, without G $\beta\gamma$ , the G $\alpha$  subunit does not require a

large displacement of  $\alpha 5$  and the  $\beta 6$ - $\alpha 5$  loop to release GDP from the binding pocket. Perturbation of  $\beta 2$ - $\beta 3$ ,  $\alpha 1$  and the  $Mg^{+2}$  binding regions is sufficient to trigger GDP release<sup>283, 301, 327</sup>. However, in the heterotrimeric G protein, the  $G\beta\gamma$  subunit interacts with Switch II and the phosphate binding region, reducing the dynamics of this region.  $G\beta\gamma$  binding significantly limits nucleotide exchange of the G protein in the absence of receptor<sup>283, 327, 345</sup>. When the receptor interacts with the heterotrimeric G protein, it rotates the  $\alpha 5$  helix and initiates the nucleotide release mechanism.

The Ins4A mutant shows similar receptor binding capability compared to WT protein. The nucleotide binding and release data show that the Ala insertion significantly affects the G protein nucleotide interaction. The heterotrimeric Ins4A mutant releases GDP almost 100-fold more slowly than WT in the presence of activated receptor. Comparison with the  $GTP\gamma S$  binding kinetics allows us to conclude that GDP release is the main affected event. However, even with a very slow nucleotide exchange rate,  $GTP\gamma S$  can still access the nucleotide binding pocket. However, release from the receptor-G protein complex is abolished even in the presence of high concentrations of either GDP or  $GTP\gamma S$ . This indicates that disrupting the hydrophobic core not only affects nucleotide interaction with the  $G\alpha$  subunit in the receptor G protein complex but also that G protein can no longer release from the receptor complex.

How does the heterotrimeric G protein bind normally to the receptor, interact with the nucleotide but not release from the receptor complex? The  $\beta 6$ - $\alpha 5$  loop directly interacts with the guanine ring of the nucleotide, and it is the only direct way to connect the nucleotide binding region to the receptor through the  $\alpha 5$  helix. Within the  $\beta 6$ - $\alpha 5$  loop resides a conserved TCAT motif that mediates key contacts with the guanine ring of GDP

that are believed to stabilize the binding of GDP within  $G\alpha$ . Indeed, mutations within this region result in enhanced spontaneous nucleotide exchange rates<sup>254, 346, 347</sup>.

Thus, receptor contacts to the  $G\alpha$  C terminus communicates structural changes through the  $\alpha 5$  helix which may modulate the conformation of the  $\beta 6$ - $\alpha 5$  loop, ultimately resulting in the release of GDP. The N-terminus of the  $\alpha 5$  helix is unfolded in the  $\beta 2AR$ -Gs complex structure<sup>295</sup>. Recently Dror *et al.* suggested that the structural rearrangements in the  $\beta 6$ - $\alpha 5$  loop are the key events in G protein activation and GDP release<sup>332</sup>. To examine environmental changes around this region, we fluorescently labeled residues L273 ( $\alpha G$ ) and K330 ( $\alpha 5$ ) and showed that the N terminus of the  $\alpha 5$  helix does not properly refold in the presence of nucleotide. This result indicates that either the 4 alanine insertion creates a buffer due to the extra length of the  $\alpha 5$  helix between the receptor and nucleotide binding site of  $G\alpha$  subunit, or it disturbs the nucleotide-dependent rearrangement of the N terminus of the  $\alpha 5$  helix (residues 368-371 in  $G\alpha s$  and 328-331 in  $G\alpha i 1$ ).

To address the question of whether slow nucleotide exchange and receptor release is caused by the increased length of the C terminal helix, the repeated set of 4 WT amino acid residues were inserted back into the same region (Ins4X, Figure 7.1B). This restores the hydrophobic core around F336 ( $\alpha 5$ ), while maintaining the longer  $\alpha 5$  helix. Notably, Ins4X showed similar basal and receptor-mediated nucleotide exchange activity to WT. In addition, it recovered its receptor release activity after guanine nucleotide incubation. This is consistent with previous studies<sup>293, 343</sup>; Natochin *et al.* showed that an 11 amino acid insertion above the hydrophobic core region (between I343 and I344 residues) did not affect G protein-receptor binding and G protein activation compared to the WT protein<sup>343</sup>. This indicates that it is not the length of the C-terminus, but rather maintaining the

hydrophobic core interactions that is critical to complete the receptor-mediated G protein activation cycle.

In summary, the Ins4A crystal structure showed how  $\alpha 5$  rotation significantly changes the conformation of  $\beta 2$ - $\beta 3$  and the  $\alpha 1$  helix. F336 is likely making direct hydrophobic contacts with F191 and M53, and it may also communicate with F189 indirectly. Residues M53-H57-F191 interact with F189 through a  $\pi$ - $\pi$  interaction between residues H57 and F189. In addition Q52, I55 and I56 in the  $\alpha 1$  helix also interact with T329 and Q333 in the  $\alpha 5$  helix.

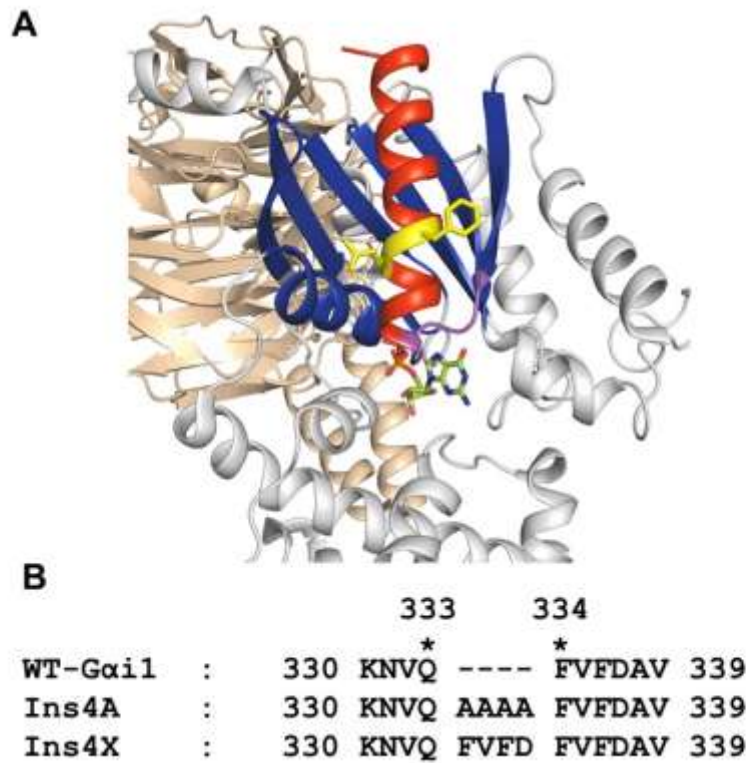
This network not only plays a major role during G protein activation, but it also influences proper rearrangement of the N terminus of the  $\alpha 5$  helix to allow release of  $G\alpha$  from the activated receptor after nucleotide binding. Thus, this study highlights changes through the G protein for receptor-mediated GDP release and G protein activation, but also the reverse communication from GDP binding to release of the G protein from the activated receptor. How G proteins influence the ligand binding of receptor, leading to a high affinity ligand binding, and in the case of rhodopsin, MetaII stabilization, is currently unknown. This study provides the first clue that rearrangement of the N-terminus of the  $\alpha 5$  helix and re-engagement of the hydrophobic core are important elements of that signaling back to the receptor.

This mechanism might be generalizable for many receptor-G protein combinations; indeed, all residues of this hydrophobic core and the N-terminus of the  $\alpha 5$  helix are highly conserved in heterotrimeric G proteins. Our results support and experimentally demonstrate that the structural rearrangements of this region complete the G protein activation cycle. Although different receptor-mediated G protein activation models are presented as

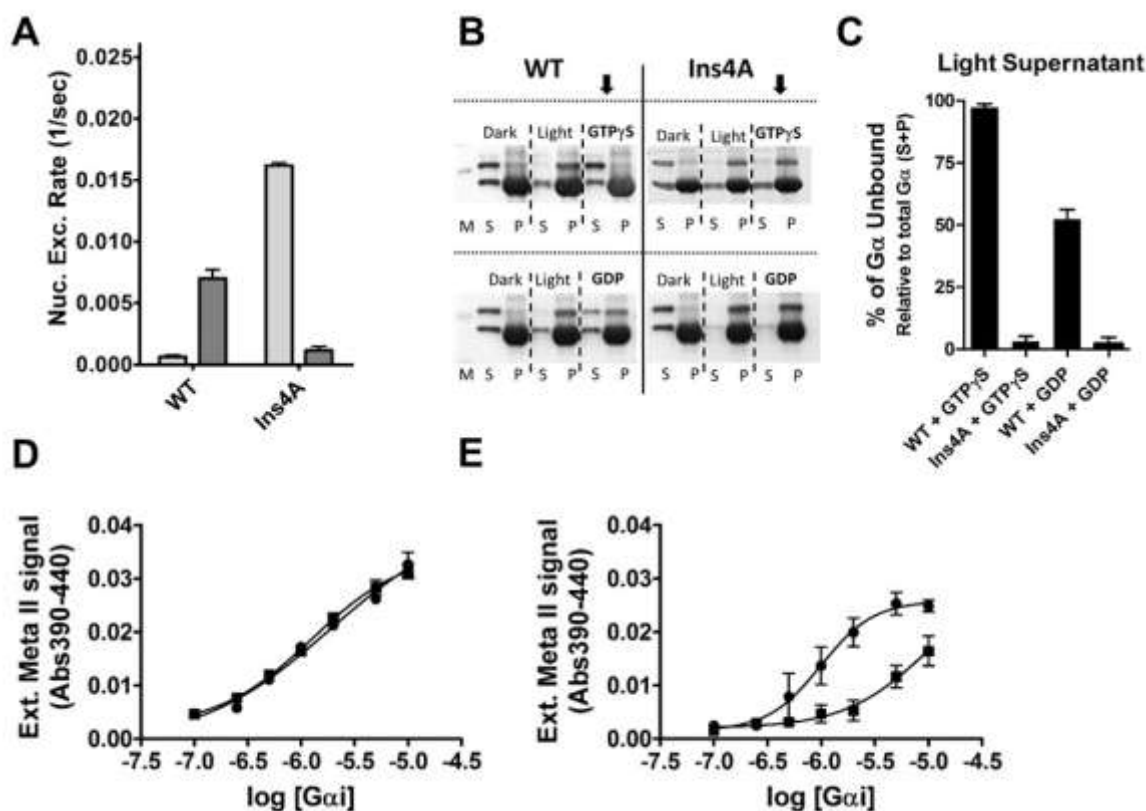
opposing mechanisms<sup>156, 287, 293, 297, 327, 331, 332, 342</sup>, they may play complementary roles in the overall action of activated receptors. However, further studies are needed to identify the sequence of events involved in receptor-mediated G protein activation in molecular detail.

### ***Abbreviations***

APS	- Advanced Photon Source
GDP	- Guanosine diphosphate
GPCR	- G Protein Coupled Receptor
GTP $\gamma$ S	- Guanosine 5'-[ $\gamma$ -thio]triphosphate
LS-CAT	- Life Sciences Collaborative Access Team
P-loop	- phosphate binding loop
REU	- Rosetta Energy Units
r.m.s	- root mean square
ROS	- rod outer segment

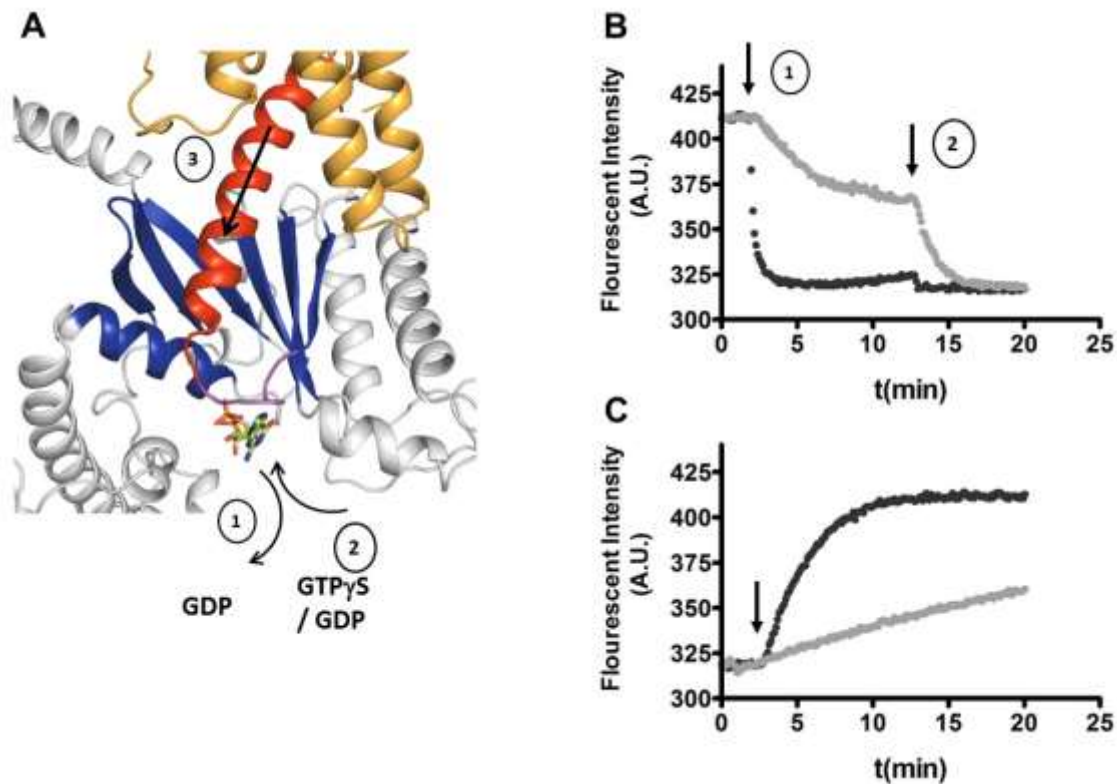


**FIGURE 7.1. Heterotrimeric G protein; localization of 4 alanine insertion in  $\alpha 5$  helix.** (A) Ribbon representation of heterotrimeric G protein ( $G\alpha\beta\gamma$ , PDB entry, 1GP2<sup>337</sup>). The  $G\alpha$  subunit is composed of nucleotide binding (blue) and helical (white) domains. The  $\alpha 5$  helix (red) is a critical region for G protein stability and activation. This helix directly interacts with six  $\beta$ -strands ( $\beta 1$ - $\beta 6$ ) and one  $\alpha$ -helix ( $\alpha 1$ ) (blue). Four amino acids were inserted between Q333 and F334 (yellow) in the  $\alpha 5$  helix. (B) Amino acid sequences and names of insertion mutants developed in this study.

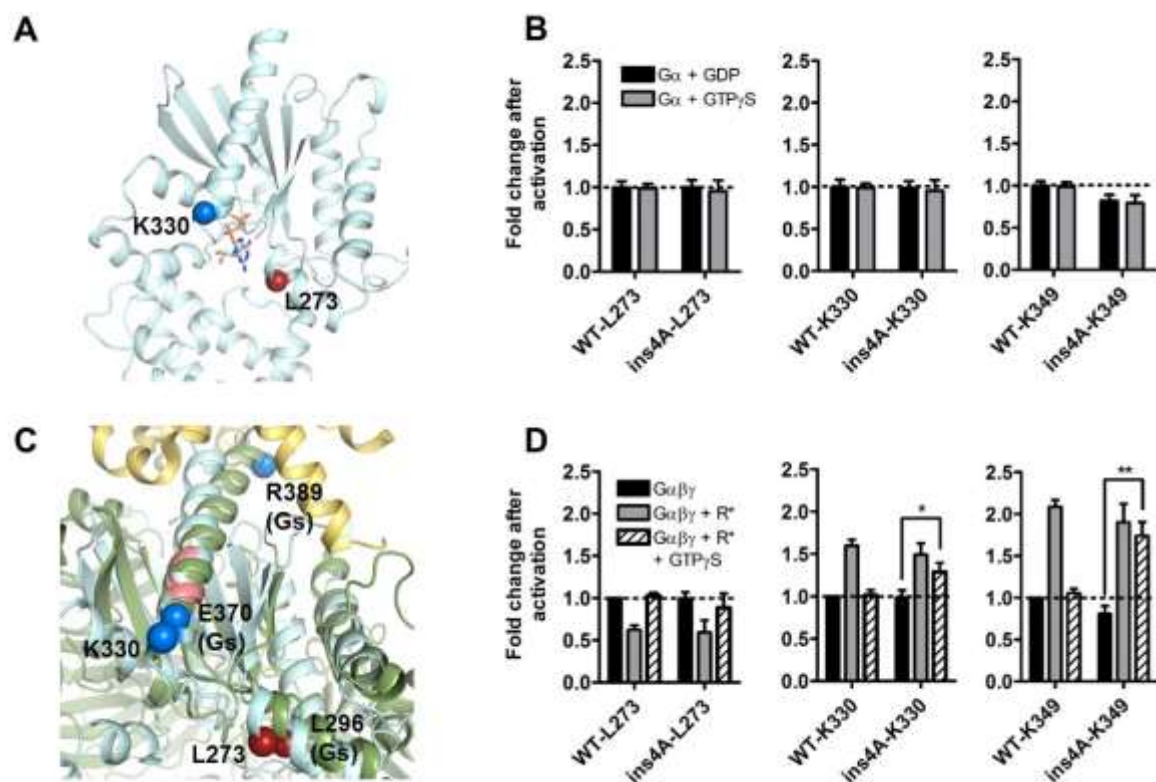


**FIGURE 7.2. Biochemical properties of Ins4A protein.** (A) Basal (grey bars) and receptor-mediated (dark grey bars) nucleotide exchange rates for wild-type and 4 alanine insertion (Ins4A) mutations in Gαi1 proteins. Nucleotide exchange was monitored by measuring the enhancement in intrinsic tryptophan (W211) fluorescence ( $\lambda_{ex}=290$  nm,  $\lambda_{em}=340$  nm) as a function of time after addition of GTPγS<sup>32,3</sup>. The data were normalized to the baseline and maximum fluorescence and then fit to the exponential association equation ( $Y = Y_{max} * (1 - e^{-kt})$ ), to calculate rate constant (k). Data were collected at 21 °C for 90 min. Results represent the mean  $\pm$  SEM values of at least three independent experiments. (B) Membrane binding of wild type and mutant Gαi1 proteins. The assay was performed as described in the methods section. Dark, from dark sample; Light, from light activated sample; GTPγS or GDP, from light activated and nucleotide incubated samples. S, supernatant; P, pellet. (C) Densitometric quantification of supernatant from light supernatant samples. Each sample from SDS-PAGE (section b) was evaluated by comparison of the amount of Gαi1 subunits in pellet (P) or supernatant (S) to the total amount of Gαi1 subunits (P+S) in both treatments and expressed as a percentage of the total Gαi1 protein. Data represent the average of three independent experiments. (D) Concentration-response curves of Meta-rhodopsin II (MII) signal stabilized by WT-Giα1 (black square) and Ins4A (black circle). (E) Concentration-response curves of MII signal stabilized by WT-Giα1 (black square) and Ins4A (black circle) in the presence of 0.5 mM GDP. The EC<sub>50</sub> value of WT-Giα1 and Ins4A protein for rhodopsin in ROS membranes was  $9.43 \pm 0.13$  μM and  $0.99 \pm 0.02$  μM, respectively. Solid curves are best fits from a four parameter logistic equation. Results are mean  $\pm$  S.E.M. from of at least three independent experiments.

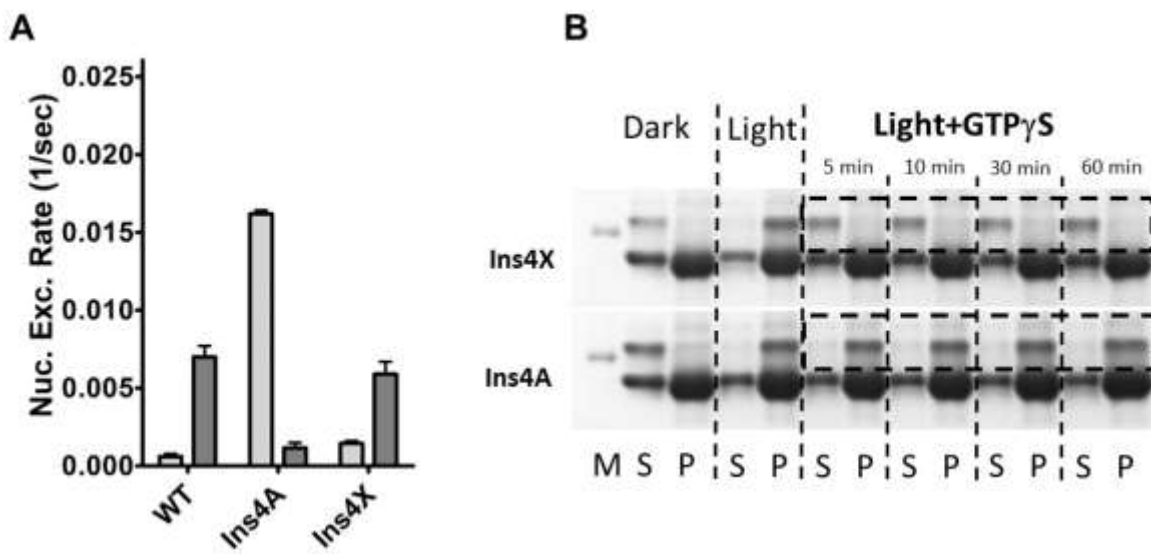




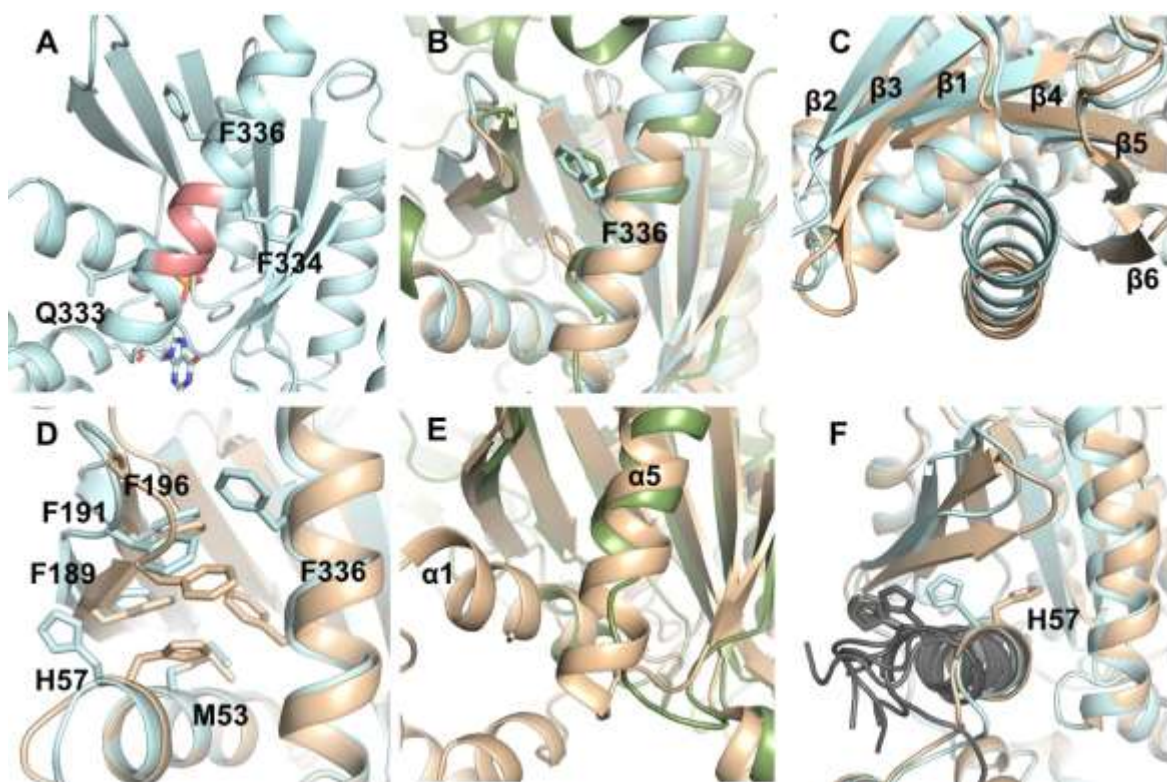
**FIGURE 7.3. Hypotheses for why Ins4A protein cannot release from the active receptor-G protein complex in presence of guanine nucleotide.** (A) Cartoon representation of possible scenarios to explain biochemical data of the Ins4A proteins: 1- Domain opening does not take place properly so GDP cannot release or the domain is able to open, but the  $\beta 6\alpha 5$  loop does not properly trigger GDP release, 2- GDP can release similar to WT-Gi $\alpha 1$  but GDP, GTP or GTP $\gamma$ S cannot bind the nucleotide binding pocket, 3- Exchange of nucleotide happens normally, but the Ins4A protein cannot release from the receptor. (B) Bodipy-GDP release from WT-Gi $\alpha 1$  (black circle) and Ins4A (grey circle). First and second arrows indicate the start of G protein incubation with light activated Rhodopsin and addition of unlabeled GDP, respectively. (C) Bodipy-GTP $\gamma$ S binding of WT-Gi $\alpha 1$  (black circle) and Ins4A (grey circle). Arrow indicates addition of light activated Rhodopsin. The fluorophore of Bodipy nucleotides was monitored at  $\lambda_{ex}$  490 nm and  $\lambda_{em}$  510 nm. The kinetic data were plotted and fit to a one-phase association function. Data are from a representative experiment that was repeated 8-10 times.



**FIGURE 7.4. Conformational changes at key sites on  $G\alpha$  caused by receptor and  $GTP\gamma S$  determined using site-directed fluorescent labels.** L273 (L296 in  $G\alpha_s$ ) residue is a sensor of the presence of the guanine ring of GDP, K349 (R389 in  $G\alpha_s$ ) is a sensor of receptor binding and K330 (E370 in  $G\alpha_s$ ) is a sensor of rotation and disorder in the presence of active receptor<sup>323</sup>. (A) Cartoon representation of labeled residues in the Ins4A- $G\alpha i1$ -GDP protein. (B) Fold change in emission intensity of  $G\alpha i1$  proteins in the presence of GDP (black bars) or  $GTP\gamma S$  (grey bars) in the presence of the indicated labeled residues, as compared to the environment of the same labeled residue in WT GDP bound state. (C) Comparison of labeled residues between Ins4A (cyan) and  $\beta 2$  adrenergic receptor- $G\alpha_s$  complex structure (PDB entry 3SN6,<sup>295</sup>, green). (D) Fold change emission intensity of  $G\alpha i1\beta 1\gamma 1$  (black bars) in the presence or absence of light activated rhodopsin (grey bars) and  $GTP\gamma S$  (white bars). Data are the average of at least three independent experiments (\*  $p < 0.05$ ; \*\*  $p < 0.01$ ).

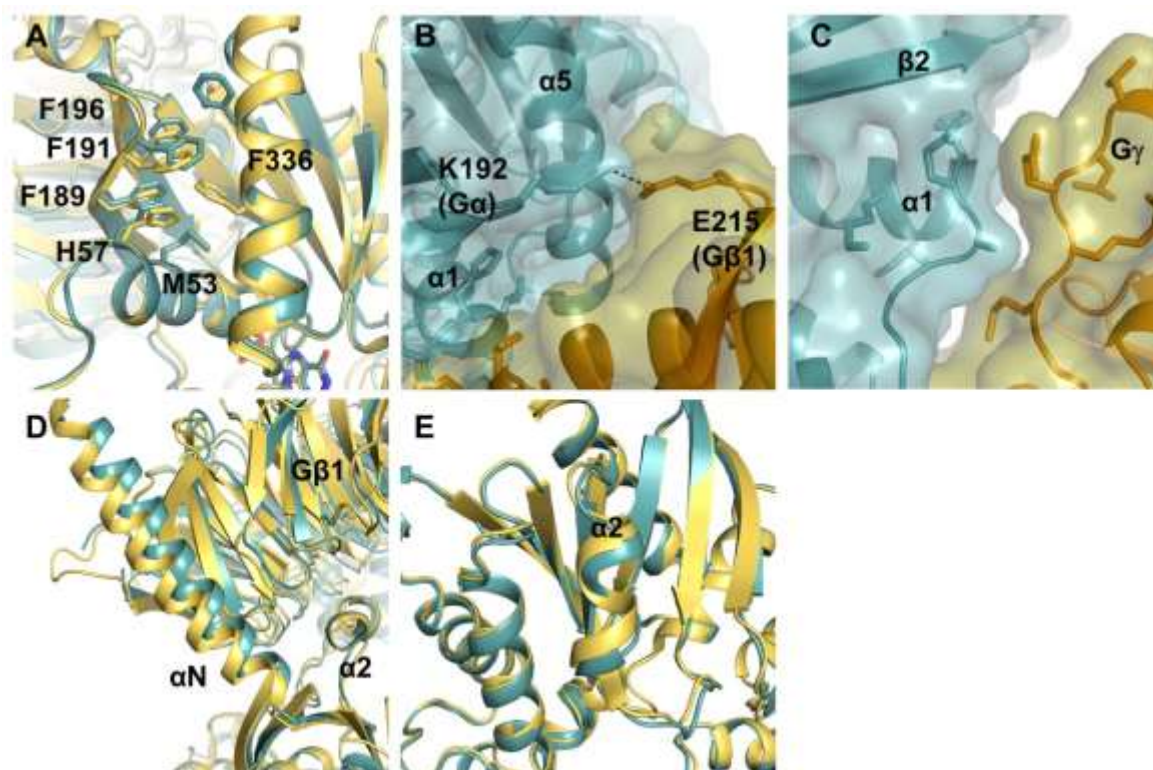


**FIGURE 7.5. The effect of introducing an FVFD insertion in the  $\alpha 5$  helix on G $\alpha$ 1 subunit.** (A) Basal (grey bars) and receptor-mediated (dark grey bars) nucleotide exchange rates for WT, Ins4A and FVFD insertion mutation (Ins4X) in G $\alpha$ 1 proteins. Nucleotide exchange was monitored by measuring the enhancement in intrinsic tryptophan (W211) fluorescence as a function of time after addition of GTP $\gamma$ S<sup>323</sup>. Data were collected at 21 C° for 90 min and represent the mean  $\pm$  SEM values of at least three independent experiments. (B) Membrane binding of Ins4A and Ins4X proteins. Dark, from dark sample; Light, from light activated sample; GTP $\gamma$ S or GDP, from light activated and nucleotide incubated samples. S, supernatant; P, pellet. Data represent the average of three independent experiments.

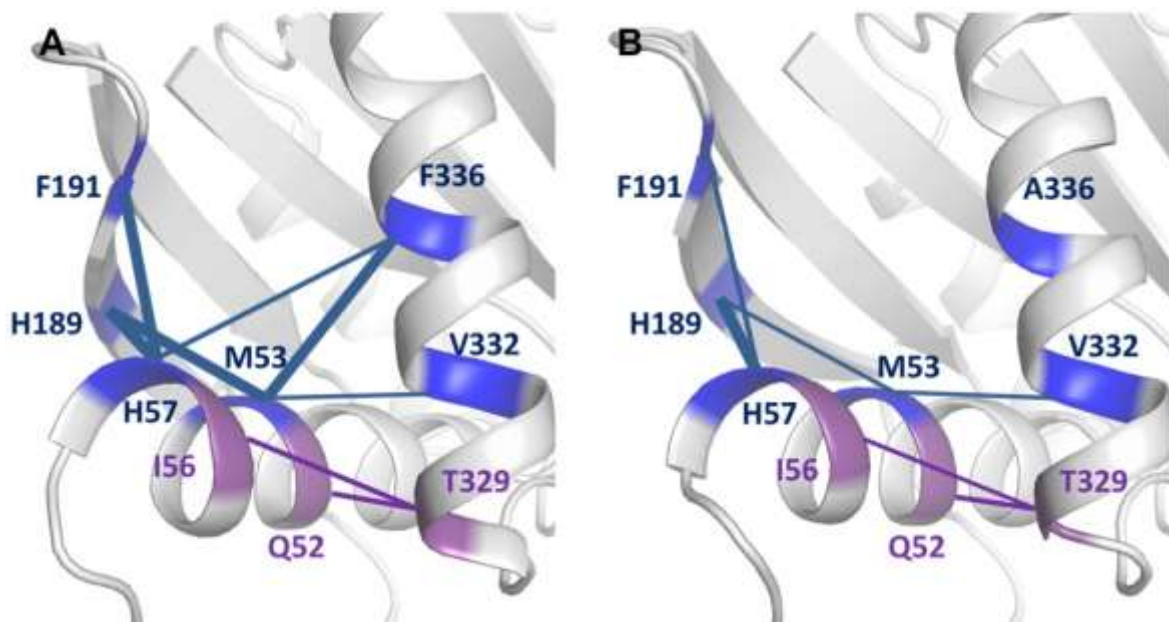


**FIGURE 7.6. Structural features of GTP $\gamma$ S bound Ins4A mutant protein.** (A) Cartoon representation of Q333, F334 and F336 residues in the  $\alpha 5$  helix of Ins4A. The 4 alanine insertion region is represented in salmon color. (B) Comparison of the  $\alpha 5$  helix and F336 residue location between WT-Gi $\alpha 1$ (PDB entry 1GIA<sup>212</sup>, brown), Ins4A (cyan) and  $\beta 2$ AR-G $\alpha s$  complex structure (PDB entry 3SN6<sup>295</sup>, green). (C) Comparison of the  $\alpha 5$  helix and  $\beta 1$ - $\beta 6$  strands between WT-Gi $\alpha 1$ (brown) and Ins4A (cyan) (D) The effect of  $\alpha 5$  helix rotation and the connection between F336 and the  $\beta 2$ - $\beta 3$  strands and the  $\alpha 1$  helix. Comparison of  $\alpha 5$ ,  $\beta 2$ - $\beta 3$  and  $\alpha 1$  regions between WT-Gi $\alpha 1$  (brown) and Ins4A (cyan). This structure shows significant rearrangement of side chains in H57, F189, F191, F196 and disturbed  $\pi$ - $\pi$  interaction between H57 and F189. (E) Comparison of  $\beta 2$ AR-G $\alpha s$  complex structure (green) and WT-Gi $\alpha 1\beta 1\gamma 2$  (brown) (F) Relative position of residue H57 and the  $\alpha 1$  helix between WT-Gi $\alpha 1$  (brown), Ins4A (cyan) and rhodopsin-G protein model (grey)<sup>133</sup>.





**FIGURE 7.7. Structural features of heterotrimeric Ins4Aβ1γ1 mutant protein.** (A) Comparison of the  $\alpha 5$ ,  $\beta 2$ - $\beta 3$  and  $\alpha 1$  regions between WT-Gi $\alpha 1\beta 1\gamma 2$  (PDB entry, 1GP2, yellow) and Ins4A $\beta 1\gamma 1$  (teal). The relative C $\alpha$  distances between mutant and WT heterotrimeric structure protein in H57, F189, F191, K192 and F196 are 0.5, 0.5, 1.4, 1.4 and 0.6 Å, respectively. (B) The interaction between K192 in the G $\alpha$  subunit,  $\beta 2$ - $\beta 3$  loop (teal) and E215 in the G $\beta 1$  subunit in a symmetry molecule (orange). (C) The contact between the G $\alpha$  subunit  $\alpha 1$  helix and G $\gamma 1$  in a symmetry molecule (orange). Surface representation in teal and orange for mutant and WT structures, respectively. (D) Comparison of the  $\alpha N$  and (E)  $\alpha 2$  helices between Ins4A $\beta 1\gamma 1$  (teal) and WT-Gi $\alpha 1\beta 1\gamma 2$  (PDN entry, 1GP2<sup>337</sup>, yellow) protein.



**FIGURE 7.8. Pairwise interaction scores highlight two activation pathways.** There are two critical stabilizing routes between the  $\alpha 1$  and  $\alpha 5$  helices in the GDP-bound state. The first (purple) is between Q52 ( $\alpha 1$ ), I56 ( $\alpha 1$ ) and T329 ( $\alpha 5$ ). The second pathway (blue) connects M53 ( $\alpha 1$ ), H57 ( $\alpha 1$ ), V332 ( $\alpha 5$ ), and F336 ( $\alpha 5$ ). (A) WT-Gi $\alpha 1$  (PDB entry, 1GIA<sup>212</sup>) maintains both networks in the GDP-bound state. (B) The Ins4A mutant loses the hydrophobic core between  $\alpha 5$ ,  $\alpha 1$  and the  $\beta 2$ - $\beta 3$  strands.

	<b>Bodipy-GDP Dissociation (k, min<sup>-1</sup> ± S.E.M)</b>	<b>Bodipy-GTPγS Binding (k, min<sup>-1</sup> ± S.E.M)</b>
<b>WT</b>	3.521 ± 0.095	0.913 ± 0.045
<b>Ins4A</b>	0.042 ± 0.001	0.031 ± 0.007

**Table 7.1. Bodipy nucleotide interactions with *Gai1***

	<b>Ins4A-Gai1-GTP<math>\gamma</math>S·Mg<sup>+2</sup></b>	<b>Ins4A-Gai1<math>\beta</math>1<math>\gamma</math>1-GDP</b>
<b>PDB accession code</b>	5KDL	5KDO
<b>Data Collection and Processing<sup>a</sup></b>		
Beamline	21-ID-F	21-ID-G
Space groups	P2 <sub>1</sub>	P4 <sub>3</sub>
Cell Dimensions: a, b, c (Å)	61.7, 77.3, 73.1	84.65, 84.65, 130.09
$\alpha$ , $\beta$ , $\gamma$ (degrees)	90, 99.8, 90	90, 90, 90
Resolution (Å)	50.0-2.8 (2.8-2.7)	40.25-1.90 (1.97-1.90)
Total Reflections	217,114	1,504,669
Unique Reflections	19,135	71,960
R <sub>sym</sub> <sup>b</sup> (%)	8.8 (51.7)	7.4 (123.4)
R <sub>pim</sub> <sup>c</sup> (%)	5.6 (33.6)	4.3 (54.0)
<I>/< $\sigma$ >	14.4 (2.04)	17.4 (1.4)
Completeness (%)	99.5 (97.4)	100 (100)
<b>Refinement Statistics</b>		
R <sub>work</sub> <sup>d</sup> (%)	20.9	18.21
R <sub>free</sub> (%)	26.4	20.79
RMS deviations		
Bond (Å)	0.002	0.007
Angle (°)	0.522	0.977
Ramachandran statistics <sup>e</sup>		
Favored (%)	99.2	98.07
Allowed (%)	0.8	1.93
Outliers (%)	0.0	0.0

<sup>a</sup>Numbers in parentheses indicate statistics for the highest shell.

<sup>b</sup> $\sigma_{\text{sym}} = \sum |I_{\text{obs}} - \langle I \rangle| / \sum |I_{\text{obs}}|$  where  $I_{\text{obs}}$  is intensity,  $I_{\text{obs}}$  is the  $i$ th measurement, and  $\langle I \rangle$  is the weighted mean of  $I$ .

<sup>c</sup> $R_{\text{pim}} = \sum_{hkl} \sqrt{[1/(N-1)] \sum_i |I_{\text{obs}}(hkl) - \overline{I}(hkl)| / \sum_{hkl} \sum_i I_{\text{obs}}(hkl)}$  where  $I$  is running over the number of independent observations of reflection  $hkl$  and  $N$  is representing the number of replicate observations.

<sup>d</sup> $R_{\text{work}} = \sum ||F_{\text{obs}}| - |F_{\text{calc}}|| / \sum |F_{\text{obs}}|$  where  $F_{\text{obs}}$  and  $F_{\text{calc}}$  are the observed and calculated structure factor amplitudes.  $R_{\text{free}}$  is the same as  $R_{\text{work}}$  for a set of data omitted from the refinement.

<sup>e</sup>Ramachandran analysis from MOLPROBITY<sup>324</sup>.

**Table 7.2. Crystallographic data collection and refinement statistics.**



		<i>Ins4A-GTP<math>\gamma</math>S</i>		WT-GTP $\gamma$ S		<i>Ins4A<math>\beta</math>1<math>\gamma</math>1</i>		WT- $\beta$ 1 $\gamma$ 2	
Entity	Amino acid	Energy in REU	Z-score	Energy in REU	Z-score	Energy in REU	Z-score	Energy in REU	Z-score
$\beta$ 1	L38	1.1	75	1.2	281	1.4	53	2.0	17
Ploop	G40	0.8	103						
$\alpha$ 1	K46	1.3	53	1.1	89	1.9	6	1.9	35
$\alpha$ 1	S47	0.7	61						
$\alpha$ 1	T48	0.9	52	0.8	154	0.7	27	0.9	27
$\alpha$ 1	I49	1.0	38	1.0	179	1.1	192	1.0	10
$\alpha$ 1	V50			0.5	235	0.9	118		
$\alpha$ 1	Q52	1.2	25	1.5	34	1.7	266	1.6	25
$\alpha$ 1	M53	0.5	6	1.4	462	1.0	116	1.4	14
$\alpha$ 1	K54	1.8	122	2.3	101	2.2	441	2.5	24
$\alpha$ 1	I55	1.4	170	1.0	263	1.0	158	1.0	10
$\alpha$ 1	I56	1.0	85	1.1	57			1.1	35
$\alpha$ 1	H57	1.3	28	1.6	133	1.5	127	1.7	34
linker 1	G60			0.7	71	0.9	339	0.9	67
linker 1	Y61	1.0	125	1.5	70	0.8	208	1.0	9
$\alpha$ A	E65	0.8	104	0.5	9	0.8	384	0.8	52
$\alpha$ F	L175	0.9	51	1.0	279	1.0	104	1.0	23
$\beta$ 2	F189	1.0	13	1.4	114	1.5	229	1.4	28
$\beta$ 2	F191								
$\beta$ 3	M198					0.5	73	0.5	6
$\beta$ 3	D200	0.8	84			1.0	3	1.0	24
$\beta$ 6- $\alpha$ 5	A326	1.3	53	1.3	88	1.3	124	1.6	38
$\beta$ 6- $\alpha$ 5	T329	0.5	21	0.8	100	0.8	346	0.8	16
$\alpha$ 5	V332	0.6	11	0.7	189	0.7	397	0.8	35
$\alpha$ 5	F336			0.9	736			0.8	24

**Table 7.3. G protein alpha subunit  $\alpha$ 1 helix interface energetic predictions**

## COMPARATIVE MODELING OF PAR4 USING ROSETTACM AND *IN SILICO* POINT MUTATIONS

### A.1 Introduction to PAR4

#### *PAR signaling*

Proteinase-activated receptors (PARs) are a class of G protein-coupled receptors (GPCRs) which can induce signaling through the G protein intracellular signaling cascade. PAR proteins are unique GPCRs in that they require proteinase cleavage for activation<sup>348</sup>. The amino terminus of PAR is exposed to the extracellular space (Figure A.1). It can be bound and cleaved by the coagulation proteinase thrombin. Upon cleavage, an encrypted “tethered ligand” is exposed along the new amino terminus. This tethered ligand can then act to mediate the receptor’s activation. This unique mechanism of activation can also be induced by proteases other than thrombin such as trypsin<sup>349</sup> and cathepsin G<sup>350</sup>.

There are four PAR proteins encoded in the human genome, named PAR1-4. All four sequences are less than 38% similar to one another at the amino acid level (See Table A.1). Indeed, each have a unique tethered ligand sequence and various affinities to thrombin cleavage and activation. The PAR family primarily signals through the G $\alpha$ q and G $\alpha$ 12 families, though there is some evidence that a G $\alpha$ i heterotrimer may also interact<sup>351, 352</sup>. Specifically for PAR4, G $\alpha$ q and G $\alpha$ 13 seem to be the primary modality in platelets.

#### *Ambiguity in PAR4 function*

In human platelets, two PAR proteins are present, PAR4 and PAR1. While PAR4 has a relatively low affinity for thrombin, PAR1 is a high affinity thrombin receptor. This disparity in affinity therefore translates into differential roles of activation between the two

platelet receptors and a dichotomy in their mechanism of response. However, the exact function of PAR4 signaling is not known, though it has been implicated to play a role in hemostasis and thrombosis as seen from mice knockout studies<sup>353, 354</sup>. This result is confounded by the lack of PAR1 expression in mouse platelets, preventing a clear depiction of how PAR4 behaves in the presence of the high affinity thrombin receptor PAR1. Therefore the role of PAR4 in conjunction with PAR1 using primate studies and/or human platelets remains largely uncharted.

One major impediment to disambiguating the role of PAR4 within platelets stems from the lack of small molecule agonists or antagonists selective for PAR4 over PAR1. This deficit is primarily due to the unique nature of this class of GPCRs whose natural ligand requires protease cleavage to be exposed. Use of only the activating peptide (AP) fragment for further ligand development and binding studies is hindered by the altered nature of the ligand when tethered to the receptor versus being free in solution<sup>355</sup>. This difference confounds *in vitro* signaling studies aimed at mapping the binding pocket.

Though extensive pharmacological and medicinal chemistry efforts have tried to create a selective, soluble, antagonist, the limited structural insight into this system has prevented progress. A recent advance came in the form of a x-ray crystallographic PAR1 structure bound to a selective antagonist<sup>356</sup>. Though at a reasonably high resolution for a membrane protein, the amino terminus is truncated, preventing analysis of the tethered ligand within the binding pocket. Other considerations for improved crystallization conditions, stabilization, and the use of fusion proteins also cloud the ability of the structure to provide insight into the mechanism of signaling.

Other groups have worked to model PAR2 using computational approaches. Their

methods started with comparative modeling of PAR2 using the known crystal structures of bovine rhodopsin and the human nociception receptor<sup>357</sup>. Next *in silico* docking was used to further refine understanding of the binding pocket. Their *in silico* methods have resulted in the refinement of several selective pharmacological compounds which bind with relatively high affinities *in vitro* and *in vivo*. These models were then used for future screens in an iterative approach to virtually screening compounds for further biochemical analysis<sup>358, 359</sup>.

#### ***Pursuing a model of PAR4***

Following in this vein, we have capitalized on the recently crystalized PAR1 structure bound to its selective agonist as a starting template for *in silico* studies of PAR4. Though PAR4 and PAR1 have a sequence similarity at the very limits for possible comparative modeling techniques (their sequence identity is 31% and the suggested limit is 30%), new advances in computational comparative modeling suggest an alternative route for probing conserved structural space. RosettaCM, an application within the software suite Rosetta, can use both single template and multiple template starting structures to thread target sequences into an aligned structural space<sup>267</sup>. We used the human PAR4 sequences as are target sequence for several different methods of structural modeling within RosettaCM.

First, using the solved PAR1 crystal structure, we used a single template modeling approach. Here we utilized the backbone atoms' coordinates as the starting points for the PAR4 structure. This was done by threading the PAR4 sequence onto the PAR1 structure using an alignment. Next we utilized the multi-template modality of comparative modeling within Rosetta to use both a five template starting point and a 20 template trajectory. This was guided by the idea that all GPCRs maintain a conserved seven transmembrane helical

bundle with several conserved motifs across class A GPCRs. Using both the five and 20 template schemes would allow for interrogation of the structural space around more than that of the single PAR1 crystal structure.

***Validating a model of PAR4 with in silico and in vitro pharmacology and mutations***

In order to validate the computational modeling efforts, *in silico* ligand docking was performed using small molecule compounds currently under development. A recent PAR4 antagonist has been obtained from Bristol-Myers-Squibb. The compound (named BMS-3 herein) was used as a starting scaffold for further chemistry efforts to optimize the scaffold, as well as drug metabolism and pharmacokinetic properties. *In silico* this compound was used as a control for docking studies in the structural models. In addition, point mutations around the ligand binding pocket were selected based on the structural model and docking poses. These mutations were created both *in silico* and *in vitro* to ascertain the validity of the computational model and map the binding pocket itself.

## A.2 General Overview of Materials and Methods

### *Crystal structure templates*

GPCR structures deposited within the Protein Databank (PDB)<sup>360</sup> were queried for their sequence similarity to PAR4, resolution, and sequence coverage by electron density. Where duplicate structures of the same protein existed, the structures of the highest resolution were utilized resulting in 20 potential templates (Table A.2). Sequences were extracted from the template PDBs and aligned using ClustalOmega<sup>90</sup>. All non-GPCR molecules included for crystallization (T4 Lysozyme, nanobodies, co-factors, *etc.*) were removed before alignment. All GPCRs possessed <30% sequence identity to the human PAR4 sequence with the exception of PAR1 which has ~32% sequence identity. As 30% represents the “cut-off” value for successful comparative modeling templates<sup>269</sup>, three different approaches for starting templates were utilized to maximize the structural search space. The first was the standard, single template comparative modeling protocol<sup>134, 269</sup> which utilized the human PAR1 crystal structure bound to Vorapaxar (PDBID 3VW7)<sup>356</sup>. Two multi-template comparative modeling strategies were also used with either five or 20 starting template structures (Table A.2 & A.3, respectively). The five starting templates were selected for their structural similarity to PAR1 as assessed by all atom RMSD (root mean squared deviation) and transmembrane atom RMSD (Table A.3).

### *Sequence similarity & SSE alignments*

Starting from the sequence alignment obtained from ClustalOmega<sup>90</sup>, secondary structure elements (SSE) were individually aligned across all five or 20 starting templates with the PAR4 sequence using MUSTANG<sup>361</sup>. PSIPRED<sup>362</sup> and OCTOPUS<sup>363</sup> were used to predict SSE and membrane spanning regions, respectively, of the human PAR4

sequence. Sequence alignments and SSE alignments were then manually evaluated and adjusted using Jalview<sup>364</sup> to ensure minimal gaps in transmembrane (TM) helices and conserved residues within the TM regions utilizing the Bissantz conservation numbering schema as a reference<sup>365</sup>. All alignments were then used for the threading step of comparative modeling.

### ***RosettaCM Protocols and Sampling***

Rosetta Comparative Modeling (RosettaCM) was performed according to Combs *et al.* for single template modeling<sup>269</sup> and Song *et al.* and Bender *et al.* for multi-template comparative modeling<sup>134, 267</sup>. See Chapter 4 for a detailed description of methods and general protocol<sup>267</sup>. Briefly, for single template comparative modeling, after selecting the appropriate starting templates and aligning sequence and secondary structure elements, the three dimensional coordinates of the template backbone structure was threaded with the target sequence based on the alignment information. Gaps within the alignment are considered “loops” for future refinement with *de novo* atom generation within Rosetta or fragment insertions composed of 3-mer and 9-mer fragments derived from the PDB. The Robetta webserver<sup>366</sup> was used for fragment generation. The remaining gaps between the template and target, flexible non-structured regions, and regions with low confidence in the alignment were re-evaluated and improved using an external loop modeling protocols.

RosettaCM with multiple templates begins in a similar manner as single template comparative modeling. Five or 20 starting GPCR crystal structures were utilized as templates. After template selection, sequence and SSE alignments were utilized to create threaded models of the PAR4 target sequence onto the backbone coordinates of each template. The threaded model from each template was then aligned into a global coordinate

frame. Hybridized models were created in an iterative manner moving between *de novo* fragment insertions and recombination of threaded-template segment insertions. Conformational search space was then expanded by randomly sampling *de novo* backbone fragment insertions or template segment insertions and energy minimizing the models to improve geometry of the backbone. This method effectively models any gap regions that would be present within the backbone when using a single template modeling protocol. Side chains were added last and optimized using a more rigorous scoring function for full-atom refinement.

It was suggested by Raman *et al.* that more aggressive iterative stochastic rebuilds and sampling are required when there is lower sequence identity between the template and target<sup>275</sup>. Preliminary runs of 500 models were created for the single template CM for several different sequence/structure alignments, 1500 for the five template alignments, and 3000 for the 20 template alignments. Each of the three methods possessed two to five different sequence/structure alignments as starting inputs for the preliminary runs. The top input alignments for each method were then carried forward for production runs as score vs. RMSD plots suggested that more sampling may result in model convergence.

#### **Clustering and Top Models**

Calibur<sup>367</sup> was used for clustering models. The ten best models by RMSD to PAR1 and Rosetta Score in the top cluster were used for further loop modeling and refinement.

#### **Model Refinement and Disulfide Constraints**

After model hybridization across the different template paradigms, a DualSpace relax protocol within Rosetta was utilized to relax each model and relieve any remaining atom clashes. Disulfide constraints were utilized to ensure proper bonding between the conserved cysteine residues 149 and 228.



### ***Loop Remodeling***

Single template comparative modeling requires an external loop closure step to remodel flexible, non-structured loop regions and for any gaps, or chain breaks, which remain after template hybridization and fragment insertion and refinement. Chain breaks result from unsatisfied peptide bonds due to template-target alignment mismatch and missing residues in target sequences versus the template protein backbone. These flexible regions and gaps, also called loops in RosettaCM parlance, were solved using CCD (cyclic coordinate descent) for initial loop closure<sup>368, 369</sup> and KIC (kinematic loop closure) as the all-atom phase of loop modeling refinement<sup>271</sup>.

Multi-template comparative modeling does not require external loop closure steps as the internal *de novo* fragment insertion and template segment insertion and annealing protocols resolve all chain breaks and smooth fragment/template edges<sup>134, 267</sup>. However, CCD and KIC were used in later refinement steps of the multi-template modeling approaches to relieve clashes around the ligand binding pocket in the presence and absence of ligand (discussed below).

ECL2 was predicted to be over 20 amino acids long in PAR4. Loops greater than 15 amino acids are not easily closed by any computational software. Therefore, ECL2 was modeled as a whole and as three separate sections: 1) the N-terminal half of the loop, up to the base of the  $\beta$ -hairpin, 2) the  $\beta$ -hairpin using Paired Distance Constraints, and 3) the C-terminal half of ECL2 after the  $\beta$ -hairpin.

During loop modeling and remodeling, disulfide constraints were utilized to ensure proper bonding between the conserved Cysteine residues 149 and 228, and Paired Distance Constraints were utilized to obtain the putative  $\beta$ -sheets that were seen in PAR1 and others

of the input templates. Preference was given to well-scoring models that recovered these SSE, though models without the hairpin would be moved forward to refine with additional distance constraints.

### ***Molecular Docking***

#### **Ligand Prep**

In order to sample more conformational diversity, conformers were generated for the BMS3 compound using MOE<sup>370</sup>. These poses were then sampled during the following docking simulations after initial placement within the PAR4 models.

#### **RosettaDock**

Docking simulations were run after placing the BMS-3 ligand into an initial starting pose within the binding pocket. The ligand (and its conformers) were then able to sample several degrees of freedom around the initial starting position. The conformers could sample up to five Angstroms of translation and a full 360° of rotation. After an initial low resolution docking trial, a high resolution docking phase optimized ligand placement and chemistry within the pocket. A final minimization step was used to relieve any remaining clashes between the protein and the small molecule. A least 2,000-3,000 models were created for each docking run.

#### **Molecular Docking with loop reconstruction**

In addition to favoring the putative  $\beta$ -hairpin, ECL2 also maintains an extended region of non-structured loops before and after the hairpin. These loop regions, in addition to ECL3, are highly flexible in the absence of ligand. Therefore, their starting conformation during ligand docking could influence the Monte Carlo sampling. RosettaDock utilizes stochastic pose sampling that is not necessarily physiologically relevant; it is not a “trajectory” in the same way as a molecular dynamics simulation. Therefore, the initial starting pose of the ligand and the loops may result in confounding clashes that are

challenging to overcome without extensive sampling and computational resources.

To overcome this confound during docking, the loop regions of both ECL2 and ECL3 were both remodeled in the presence and absence of BMS-3 after initial docking trials gave relative ligand placement predictions. Initially, with all loop backbone atoms present, ligand docking occurred as described previously. Then ECL2 was remodeled in two sections around the docked ligand. Rosetta loop modeling cannot accurately recover loop conformations for loop regions that are greater than 15 amino acids. The ECL2 region of PAR4, subtracting the  $\beta$ -hairpin is 20 amino acids in length (roughly 30 amino acids with the hairpin). Therefore, the hairpin region was considered “fixed” for the purposes of loop remodeling as were the TM regions. The remaining 20 amino acids of ECL2 were divided in half and re-built from the carboxyl terminus of the hairpin or the amino terminus of TM 5. ECL3 is less than 10 amino acids in length. It was rebuilt in conjunction with ECL2 as the conformation of one loop region would affect the conformation of the other.

Loop modeling around the top BMS-3 ligand poses continued as described above using an initial 3-mer and 9-mer fragment insertion followed with a low resolution centroid phase of loop closure using CCD and a high resolution, all atom phase of refined loop modeling using KIC. Final models were relaxed and energy minimized after loop closure to remove any remaining clashes between the protein and ligand.

### **Model validation**

After docking with and without loop remodeling around the ligand, the top models by Rosetta Score were clustered using Calibur<sup>367</sup>. The best scoring models from the top three clusters were then used for additional iterations of loop modeling and ligand docking in a cyclic fashion. In order to compare structural diversity, the PAR1 crystal structure was energy minimized within Rosetta and used to compute RMSD (root mean squared

deviation) between all of the models. In addition, the top model from each cluster was also used to compare cluster structural diversity across each round of BMS-3 docking and loop modeling.

Interface energy scores between the BMS-3 ligand and the protein pocket were also computed using Rosetta. This computes the interaction score between proximal atoms of the ligand and the binding pocket and is useful for filtering for clashes and steric hindrance.

The conserved disulfide between TM3 and ECL2 was maintained for all modeling, docking, and loop remodeling runs. In addition, as PAR4 is a GPCR, all modeling, docking, and loop remodeling trials were done in the presence of an implicit membrane within Rosetta. The membrane spanning region was determined from 1) the PAR1 crystal structure and conserved TM residues between it and PAR4, 2) the OCTOPUS membrane topology predictor<sup>363</sup> and 3) PSIPRED v.3 protein secondary structure prediction tool<sup>362</sup>.

Despite the 31% overall sequence similarity between PAR4 and PAR1, the TM regions of both proteins share many conserved residues. In addition, the core binding pocket also possesses many overlapping amino acids. This improved confidence in the modeling approaches which resulted in TM regions and binding pockets with lower RMSDs compared to the PAR1 structures.

### ***Site Specific Point Mutations in silico***

#### **Point Mutations**

Point mutations of PAR4 into PAR1 sequence were created *in silico* to mimic *in vitro* studies and map the binding pocket around the BMS-3 and Vorapaxar ligands. Six residues within ECL2 and ECL3 were mutated to PAR1 sequence (See Table A.4) in the top comparative models. All individual mutations were incorporated using fixed backbone coordinates within Rosetta. An all-atom relax was then used to resolve any clashes induced

by the point mutation. Models combining all six point mutations were also created using the aforementioned protocol of a fixed backbone, side chain atom replacement, then all-atom relax. The disulfide and membrane spanning regions were maintained through all experiments.

### **Docking**

The BMS-3 (PAR1 selective antagonist) and Vorapaxar (PAR1 selective antagonist) were then re-docked into the mutant models (without loop re-modeling) to evaluate the effects of the single and group mutations on the modeled PAR4 binding pockets.

### ***Site Specific Point Mutations in vitro***

Single point mutations were assayed *in vitro* to validate the binding site of the computational models and better map the pocket. PAR4 human sequence was expressed along with a GFP fusion protein. These biochemical studies are in progress to evaluate the residues within the binding pocket necessary for ligand interaction and selectivity between the two proteins.

### A.3 Results and Discussion

#### *Construction of Comparative Models*

Comparative modeling utilizes solved crystal structures as templates for transferring backbone coordinates to unsolved structural protein models. It relies on conserved tertiary structure and has been shown to perform with relatively higher accuracy the higher the sequence similarity between the template and target<sup>269</sup>. However, for all templates utilized herein, sequence identity was between 18-32% to human PAR4 (Table A.2). Therefore, multiple comparative modeling strategies were employed to cover a broad swath of conformational search space.

#### **Single template**

For single template comparative modeling, the sequence of the target human PAR4 was treaded onto the backbone coordinates of the structural template, PAR1 (PDBID 3VW7)<sup>356</sup>, and energy minimized. The PAR1 crystal structure was solved at a 2.2 Angstrom resolution bound to the antagonist Vorapaxar. The electron density of crystal structure begins with the residue D91 on the extracellular side exposing a flexible N-terminal region, but not the tethered peptide ligand (a.a. 42-47 SFLLRN) which was cleaved prior to crystallization between amino acids 85 and 86. The final nine residues of the C-terminus are also not resolved in the crystal structure. Therefore, the PAR4 structural model was created by a sequence and predicted structural alignment between residues 91 to 416 of PAR1. In addition, PAR1 was crystalized with a T4 lysozyme insertion within intracellular loop 3 (ICL3). This insertion resulted in the deletion of residue V207. The lysozyme was removed

before modeling, and the deletion of V207 was not suspected to affect the overall composition of the model or the binding pocket as sits in ICL3.

As the PAR4-PAR1 sequence alignment is very close to the threshold cutoff for comparative modeling with RosettaCM<sup>269</sup>, we evaluated the predicted secondary structural elements (SSE) and predicted transmembrane spanning regions of the PAR4 sequence using PSIPRED<sup>362</sup> and OCTOPUS<sup>363</sup>, respectively. Figure A.2 shows an overlay of the PAR1-PAR4 alignment and the predicted SSE and transmembrane regions compared to the SSE of the PAR1 crystal structure.

### **Multi template**

Improvements in comparative modeling within Rosetta suggest the use of multiple starting templates to increase conformational search space when sequence similarity is low between the target and templates, but the overall tertiary fold is thought to be conserved. Therefore, we created comparative models of human PAR4 using either 20 or five starting structures of solved GPCRs (Table A.2 & A.3). The sequence and SSE elements of these templates were aligned to ensure no gaps were found between the PAR4 sequence and TM regions, as gap closures in these regions are more deleterious to the integrity of the protein models. All comparative models derived from the 20 or five templates were clustered and evaluated by their Rosetta score, RMSD to the best scoring model and to the PAR1 structure, compliance with the disulfide constraints, recapitulation of the  $\beta$ -hairpin within ECL2, and quality of all loop regions.

### ***Comparative Models Quality between the Single and Multi-Template Approaches***

#### **Single Template**

Comparative models using the single template approach resulted in favorable models by RMSD (See Figure A.3). Within loop modeling, disulfide constraints were utilized to

ensure the conserved disulfide bond between residues C149 and C228 was maintained. No distance constraints were required to recapitulate the  $\beta$ -hairpins when using the PAR1 template alone as input. However, 18 prolines are encoded within the PAR4 sequence, 12 of which are found within the TM helical elements. This explains the odd geometries found in the helical regions as prolines are known to kink and bend  $\alpha$ -helices.

### **Multi-Template**

ECL2 was predicted to be over 20 residues long based on PSIPRED and OCTOPUS results. No current computational techniques have been shown to accurately model loop regions greater than 15 amino acids. Therefore capturing this loop region was challenging. All models from both the 20 template and five template approaches maintained the conserved disulfide bond between TM3 and ECL2. However, roughly 80% of the 20 template models did not capture the  $\beta$ -hairpin of ECL2. Instead,  $\alpha$ -helical elements were found in many of the extended loops (Figure A.4a). This is due, in part, to the nature of RosettaCM and its scoring function, as it is known to prefer helical regions over  $\beta$ -strand pairing.

In addition, the tops of the helical bundles exposed to the extracellular regions were expanded outward, away from the bundle core. The extracellular loops were situated within the pocket, occluding binding of any ligand (Figure A.4b). This is likely due to Rosetta's need to satisfy as many bonds as possible for improved scoring. Flexible loop regions were therefore filling space that was physiologically less probable due to the constraints of the scoring and sampling of Rosetta. Instead, the binding pocket must be able to accommodate the tethered ligand of PAR4 which is exposed and available after thrombin cleavage. Our models were not created in the presence of a tethered ligand as little is understood about the three dimensional peptide interactions of this receptor. Therefore, initial modeling efforts



using 20 starting templates, regardless of the input alignment modifications did not result in physiologically relevant loop poses. This is reflected in the Rosetta score versus RMSD plots (Figure A.5) which compare the RMSD of all residues and the RMSD of only the TM regions.

Another potential weakness in these models was the diversity of the GPCR starting templates utilized. We hypothesized that the sheer expanse of conformational search space was too large to be sampled thoroughly by Rosetta when using 20 starting templates without more extensive model generation and computational resources.

Therefore, we utilized a more focused modeling strategy using the five starting templates with the highest RMSD to the PAR1 structure. We hypothesized that the PAR1 structure would be the closest protein fold to PAR4, specifically in the TM regions as PAR4 and PAR1 both share a high number of conserved residues within these helices. However, it was important to broaden the search space beyond that of the biased single template modeling. Therefore, the five template comparative modeling approach focused the search space around templates which were of highest crystal resolution, and were closest in RMSD to the PAR1 starting structure (Table A.3); two of the starting templates (3VW7 and 4N4H) maintained  $\beta$ -hairpins in ECL2. These five templates were then aligned by sequence and SSE with PAR4.

### ***Docking the selective antagonist BMS-3 into PAR4 models***

#### **Single Template**

The binding pose of Vorapaxar within the PAR1 crystal structure is fairly shallow for a GPCR. The PAR1 selective antagonist anchors itself between TM6 and 7 along its carbamate end, its tricyclic core interacts with ECL2 while its fluorophenyl tail interacts with TM4 and 5. The structural overlay of Vorapaxar and BMS-3 suggests that BMS-3

may have a similar binding pose in PAR4 as Vorapaxar in PAR1.

Single template ligand docking resulted in several large clusters of ligand poses (Figure A.6). The indole core of BMS-3 roughly aligned with the heterocyclic core of VPX in several of the top clusters. However, the tails of BMS-3 primarily preferred to interact with ECL2 and 3 rather than rest between the TM helices. Other clusters of BMS-3 poses occupied a deeper binding pose than the Vorapaxar ligand, preferring to stay within the helical core as opposed to interacting between the helices. These results appear to be due to the bulky, hydrophobic side chains (primarily leucines and tyrosines) found in the analogous regions of PAR4 which occlude interaction between the helices, but not in PAR1 which is less tightly packed (primarily with alanines). Another major difference between the PAR4 single template models and the PAR1 structure is the position of the helical bundle tops before the loop regions. In the PAR1 structure, TM6 and 7 each open away from the core, giving more space for ligand binding. In the PAR4 models, the top of TM6 is packed 2.2-4.0 Angstroms tighter into the core while TM7 is 2.3-4.7 Angstroms closer. As the helices act as anchor points for the loop regions, ECL3 of PAR4 takes on a very different conformation across models, deviating away from the PAR1 structure. These loop conformations and helical packing interactions make recovery of the Vorapaxar binding pose challenging. Further biochemical mapping of the PAR4 binding pocket is necessary to ascertain which side chains are interacting with the ligand and which are stabilizing the pocket.

### **Multi-Template**

One of the greatest challenges in comparative modeling is the accurate construction of flexible loop regions. Presently, there are no computational methods available to accurately recapitulate unstructured regions above 15 amino acids in length. As the ECL2 region is

predicted to be over 20 amino acids long according to both PSIPRED<sup>362</sup> and OCTOPUS<sup>363</sup> (Figure A.2), we utilized an iterative loop modeling approach for the multi-template modeling and docking as BMS-3 was hypothesized to bind shallowly within PAR4 similar to Vorapaxar in PAR1.

For the initial models created without the ligand present, ECL2 and 3 were highly unstructured and preferentially occluded the binding pocket. It was suspected that this was due to the scoring function within Rosetta and the need to satisfy tighter packing interactions within the loops and the protein core during the comparative modeling step. Therefore, though the BMS-3 compound was placed in similar pose as Vorapaxar, the ligand preferred to bind more deeply within the helical core (Figure A.7 & A.8) and away from the extracellular loops. As opposed to the single template models, the tops of the helical bundles spread farther away from the protein core than in PAR1. Therefore shallow ligand docking into the pocket required more interaction with the flexible loop regions and less with the TM helices; this is contrary to what is seen with PAR1 and Vorapaxar whose ends interact with side chains between TM4-5 and TM6-7. We found that BMS-3 binding poses that were shallower, sterically clashed with the loops, and were scored unfavorably.

To overcome this challenge, an iterative docking and loop remodeling approach was therefore utilized. We hypothesized that the PAR4 protein assumes the  $\beta$ -hairpin structures seen in many GPCRs, including its paralog, PAR1. Therefore to iteratively model ECL2 and dock BMS-3, the extracellular loop region was broken down around the hairpin and conserved disulfide bond. Smaller loops were modeled in between “fixed” anchor points defined as the top of the fourth TM helix and the start of one side of the hairpin for one loop, and the cysteine involved in the disulfide (to TM3) at the other end of the hairpin to a

proline at the top of TM helix6. This effectively divided ECL2 into 3 regions, the pre-hairpin segment, the structured hairpin, and the post-hairpin segment.

In order to remodel the three segments, the top models of the 20-template comparative modeling approach with the top three ligand binding poses were used to cover more conformational diversity. Though it was hypothesized that BMS-3 binds in the same manner as Vorapaxar, we selected an ensemble of models for the top three binding modes and remodeled the initial loop segment across all decoys.

The initial segment, from an anchor point on TM4 to the base of the hairpin was only six amino acids long. Keeping the rest of the protein fixed, we found little variance in the this loop segment when rebuilding in the presence of the ligand as this section does not interact with BMS-3 in any of the best-scoring binding poses. However, when evaluating the region C-terminal of the hairpin to the top of TM5, there was much flexibility in the loop region in the presence of the ligand. Specifically, the  $\alpha$ -helix of TM5 was extended in many of the models as well as additions of helices throughout the models (Figure A.9). Depending on which pose was modeled for the ligand, Rosetta scored ECL2 interaction with the ligand as favorably as it did the ligand interacting solely in the core of the protein without any ECL2 interaction. The same was true of remodeled ECL3 in the presence of docked BMS-3. Therefore further biochemical mapping of the binding pocket is required to ascertain where the BMS-3 compound interacts within the large PAR4 binding pocket.

#### ***Site Specific Point Mutations in silico***

To aid in mapping the PAR4 binding pocket, six point mutations were selected for characterization *in vitro*; these site specific point mutations would then be characterized for their ability to interact with the PAR4 selective antagonist BMS-3 versus the PAR1

selective antagonist, Vorapaxar. Thrombin activation and the response to the soluble, activated PAR4 or PAR1 peptides would also assist in the characterization of the peptide binding pocket versus the small molecule binding pose.

To compliment these studies, we recapitulated these mutations *in silico* to assess the predictability of our structural models and the accuracy of the docking trials. As the multi-template docking protocol which used 20 templates appeared to recapitulate binding poses similar to Vorapaxar binding, the top models from this approach were utilized for the six individual point mutations as well as a combination mutant trial containing all six mutations. Figure A.10 shows the location of all 6 mutations as compared to the WT model.

For each point mutant, the PAR4 sequence was swapped with the aligning PAR1 sequence (Table A.4). The mutations were selected from ECL2 and ECL3 for their interaction with Vorapaxar or their potential contact with the computationally docked BMS-3. These residues were also selected for their location in non-conserved regions between the two proteins. After the fixed backbone mutation and protein relax, BMS-3 and or Vorapaxar was re-docked into the mutant proteins.

The A231L mutation within ECL2 drastically affected the binding pose of BMS3. The insertion of the larger hydrophobic chain pushed the ligand deeper into the binding pocket or required a rotation of the ligand core down and away from the extracellular loops suggesting that a flexible alanine at this position would be important for shallow docking and interaction with ECL2. In addition, Vorapaxar binding to the mutant protein was almost fully recovered in all best scoring models suggesting that the leucine in position 262 (or 231 in PAR4) may play an important role in selectivity between PAR1 and PAR4.

Strangely, the Q232L mutation had a large effect on BMS-3 binding. The glutamine does not interact with BMS-3; however, the mutation to the hydrophobic leucine resulted in all of the top models shifting their binding pose to interact with the mutant side chain. These results are confounded by the propensity of the ligand to move outside of the protein into what should be the lipid membrane. RosettaDock uses an implicit membrane and should score these poses less favorably. These confounds were also seen in all trials of Vorapaxar docking where most of the best scoring trials resulted in the ligands outside of the protein and in the membrane. More docking trials must be run around this mutant to fully understand how these new interactions with leucine were more favorable than the penalties in Rosetta that should prevent the ligand from “exiting” the protein.

The proline to leucine change at the top of TM6 into ECL3 (P310L) prevented recovery of the initial BMS-3 binding pose. The methyl tail of BMS-3 could no longer interact between the TM helices and preferred to bind farther into the core of the protein. Partial recovery of Vorapaxar was found in less than a third of the top scoring docking trials. The carbamate tail preferentially interacted with TM6 and 7; however, the tricyclic core rotated to interact with ECL2 in many other models to accommodate the TM interactions. This suggests that position 310 relays some ligand selectivity to the PAR protein; however, it is coordinated across several other positions within the binding pocket.

The second proline mutant in ECL2 (P312H) also prevented recovery of the initial BMS-3 binding pose. Though no specific interactions or contact were found between either the proline or the histidine mutant, the proline residues acts as an anchor point to kink ECL3 down towards the pocket. In the presence of the histidine, all of the best scoring models possessed a more flexible and “open” ECL3 away from the binding pocket toward

ECL2 with C $\alpha$  distances of 4.2-4.8 Angstroms. Indeed, the mutant models lacking the second proline at residue 312 maintained an extra turn in TM helix 6; this allowed the BMS-3 compound to move up and out of the pocket to maintain its interactions with other residues of ECL3 while making new contacts within ECL2. The shifted ECL3 also allowed Vorapaxar to sample more conformational space making recovery of Vorapaxar binding limited to only 20% of the top models. Therefore, though proline 312 is not predicted to directly interact with BMS-3 (nor H342 to directly interact with Vorapaxar), this position acts a critical anchor point at the top of TM6 and dictates the flexibility and conformation of ECL3 which ultimately affects the shallow ligand binding pockets of PAR1 and 4.

A314S and W315T both appeared to have little effect on the binding of BMS-3 as the ECL3 loop still accommodated the 5-5 ring system, though BMS-3 maintained several new binding poses to satisfy more hydrogen bond interactions in the A314S mutant. W315T was not predicted to directly interact with BMS-3, and the threonine side chains did not disrupt the flexible loop region. Vorapaxar binding was recovered in a fraction of the scoring models of A314S and W315T, though the scores were less favorable than for BMS-3 binding; this is expected as Vorapaxar should not bind favorably to PAR4 unless the mutant residues improved binding. Therefore A314 and W315 of PAR4 and S344 and T345 of PAR1 do not appear to be critical residues for the binding of either ligand. Instead, residues within ECL2 and the N-terminal region of ECL3 appear to have more influence on the shallow ligand binding pocket of PAR1 and 4.

## A.4 Conclusions

There are two proteinase-activated receptors on human platelets, the high affinity thrombin receptor PAR1 which is responsible for quickly responding to thrombin, and the low affinity PAR4 protein. The function of PAR4 as a low-affinity thrombin receptor has been studied to a limited degree as no high affinity, selective antagonists previously existed for PAR4. *In vivo* assessment of its function in platelets was limited to knockout studies in mice. Confounding these results were the lack of PAR1 co-expression in these studies. Therefore, new tool compounds are required to interrogate this receptor's function in human or primate platelets.

RosettaCM was employed to create an initial approximation of the PAR4 structure based on the recently solved PAR1 crystal structure<sup>356</sup>. The PAR1 and PAR4 sequences share a low degree of sequence similarity which made the computational modeling efforts challenging. Therefore a three-pronged modeling effort was utilized which incorporated three different starting trajectories for the transmembrane receptor. This included a single template modeling approach based solely on the human PAR1 homolog of PAR4 which shared some ligand similarity and couple to similar G proteins. The other two approaches assessed more conformational diversity by incorporating either 20 or five Class A GPCRs as starting templates for modeling. These templates also possessed low sequence similarity to the human PAR4 sequence, but maintained the overall conserved 7TM receptor fold. Preference was giving to receptors which also possessed a  $\beta$ -hairpin within ECL2. These



two multi-template approaches allowed for sampling of more conformational diversity outside of the initial PAR1 starting structure. The best scoring models from each of these three trajectories were then moved forward for further characterization and to guide novel ligand synthesis (not shown).

We also took advantage of the newly patented BMS-3 compound which was shown to possess PAR4 selectivity over PAR1 in *in vitro* assays. This made BMS-3 an excellent candidate for docking trials in the comparative modeling decoys to assess the models' validity and further assist biochemical efforts in characterizing the binding pocket. In addition, the PAR1 selective antagonist, Vorapaxar, maintained a very tight alignment with the BMS-3 compound when overlaid. This suggested that perhaps Vorapaxar and BMS-3 shared a similar binding pose within the receptors with a few critical amino acids or small molecule functional group deviations to account for their binding selectivity between the two proteins. Our docking efforts suggest that BMS-3 can assume a binding pose similar to Vorapaxar, though there are many well-scoring binding poses available to the ligand deeper within the core of the PAR4 pocket.

These subtle differences in the binding pocket and binding modalities were further assayed through both *in vitro* and *in silico* site-specific point mutations and ligand redocking trials. Though the biochemical assays are on-going and not discussed here, we found that of the six mutants selected along ECL2 and 3, only four of them altered ligand binding. Two of the four did so without physically interacting with the ligand; rather, they altered the flexibility of the loop regions by acting as immobile anchor points at the tops of TM6 and into ECL3.

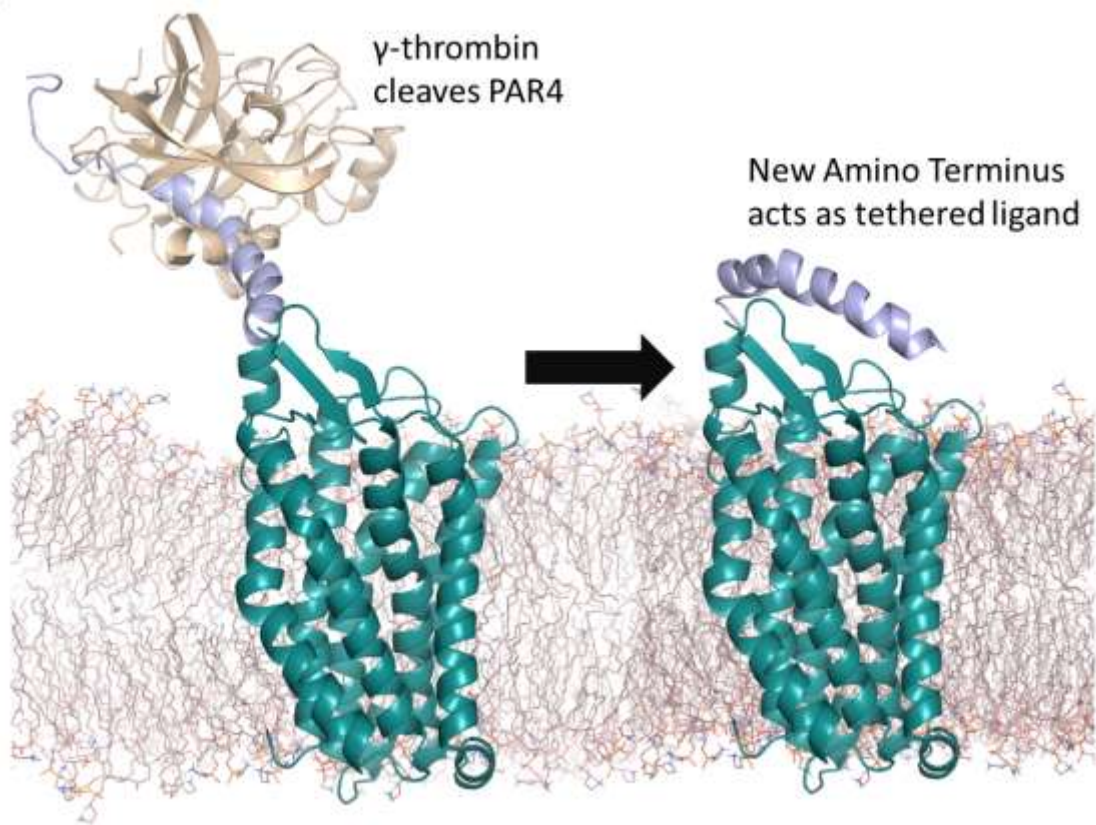
The protein modeling and small molecule docking efforts reported herein were

performed entirely *in silico* though they were guided with *in vitro* data. In parallel, we have assayed human PAR1 and PAR4 in platelets in the presence of thrombin, activated peptide, and a library of putative small molecule inhibitors. Our approach to modeling was cyclic and iterative between the pharmacology and medical chemistry efforts. Not discussed were the results of the novel ligand developments using a diverse series of starting scaffolds derived from the BMS-3 compound via a scaffold-hopping approach. In addition library screenings of the Vanderbilt chemical libraries were also applied to obtain novel scaffolds which may preferentially interact with PAR4 over PAR1.

These pharmacology and synthetic chemistry efforts are ongoing. In addition, several of these scaffolds with favorable binding kinetics will be pursued for radiolabeling. The use of a hot ligand in conjunction with the site-specific mutations would be an invaluable tool for biochemically characterization of the ligand binding pocket. In addition, collaborations with the Kobilka lab are underway to crystalize the structure of PAR4 bound to BMS-3.

#### **Abbreviations**

PAR1 & 4	-- Protease Activated Receptor1 & 4
GPCR	-- G protein Coupled Receptors
PDB	-- Protein DataBank
SSE	-- Secondary Structure Elements
TM	-- Transmembrane
RosettaCM	-- Rosetta Comparative Modeling
ICL	-- Intracellular Loop
ECL	-- Extracellular Loop
a.a.	-- Amino Acid



**Figure A.1: Activation mechanism of PAR4.** Protease-activated receptors (PAR), such as PAR4, are a class of GPCRs whose extracellular N-terminus can be cleaved via a thrombin protease. After cleavage, the newly exposed amino terminus acts as a “tethered ligand” which binds within the receptor and mediate activation

<b>PAR1</b>				
<b>PAR2</b>	37.37			
<b>PAR3</b>	32.23	33.04		
<b>PAR4</b>	30.98	32.68	34.28	
	<b>PAR1</b>	<b>PAR2</b>	<b>PAR3</b>	<b>PAR4</b>

**Table A.1: Amino acid sequence identity between the human PAR sequences (1-4).** Uniprot<sup>96</sup> Entry P25116, P55085, O00254, and Q96RI0, respectively. Alignments performed in ClustalOmega<sup>90</sup> using the default settings.

<u>PDBID</u>	<u>Protein Family</u>	<u>Crystal Resolution</u>	<u>Sequence Identity</u>
3vw7	PAR1	2.2	32.11
1u19	Rhodopsin	2.2	20.54
3eml	A2AR	2.6	25.18
3odu	Chemokine R	2.5	25.00
3rze	Histamine H1R	3.1	22.94
3v2w	Lipid Receptor	3.4	23.36
4djh	K-Opioid R	2.9	24.32
4ea3	Nocicep/Orphan	3.0	28.03
3pbl	D3 DAR	2.9	25.36
3uon	M2 AChR	3.0	25.87
4daj	M3 AChR	3.4	23.77
4dkl	M-Opioid R	2.8	25.50
4n6h	$\Delta$ -Opioid R	1.8	24.06
4iar	5HT 1b	2.7	22.26
4ib4	5HT 2b	2.7	23.38
3sn6	$\beta$ 2AR	2.9	21.59
4bvn	$\beta$ 1AR	2.1	21.38
4phu	GPR40	2.3	26.80
4grv	Neurotensin R	2.8	22.07
4rnb	Orexin R	2.5	23.34

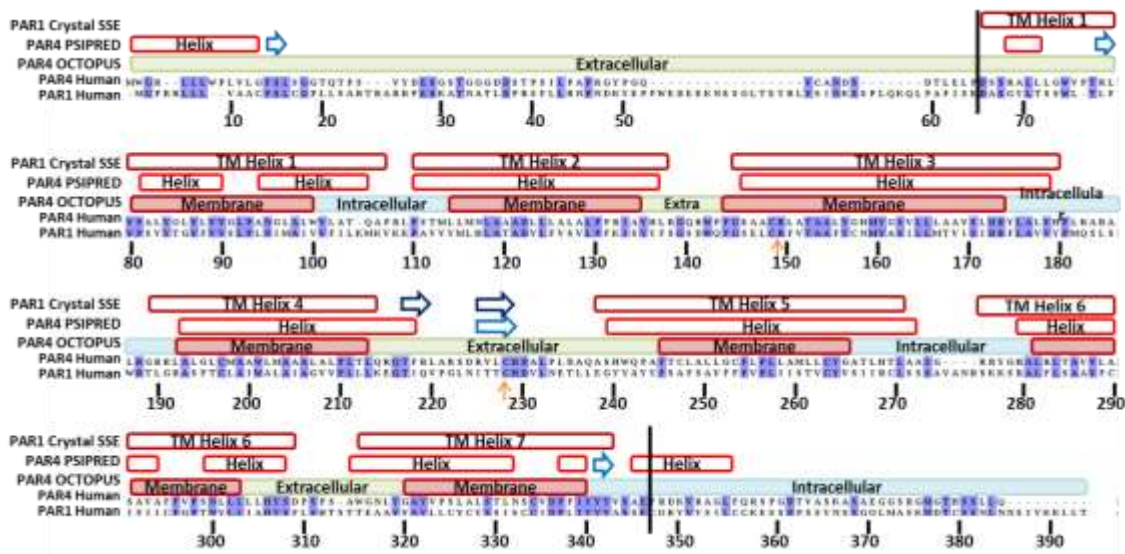
**Table A.2: 20 Structures selected as templates.** Crystal structures of GPCRs were curated from the PDB to act as templates for comparative modeling experiments. Structures were selected for their resolution, lack of mutations and quality of electron density in critical regions of the protein. Sequence similarity between all templates and the target sequence was below the threshold for single template comparative modeling except for PAR1.

<u>PDBID</u>	<u>Protein Family</u>	<u>Crystal Resolution</u>	<u>Sequence Identity to PAR4</u>	<u>All Atom RMSD to 3VW7</u>	<u>TM RMSD to 3VW7</u>
3vw7	PAR1	2.2	32.1	-	-
3odu	Chemokine R	2.5	25.0	3.3	1.2
4ea3	Nocicep/Orphan	3.0	28.0	2.3	1.9
4n6h	$\Delta$ -Opioid R	1.8	24.0	2.3	1.7
4phu	GPR40	2.3	26.8	1.6	1.2

**Table A.3: Five structures selected as templates.** To refine the search space around the PAR4 sequence, five crystal structures were selected from the initial 20. These templates were selected based on their structural similarity to the PAR1 crystal structure as a whole and their conserved transmembrane (TM) regions. As only two of the five maintained  $\beta$ -hairpins in their ECL2, these regions were “dis-aligned” for all alignment inputs that did not contain hairpins to improve SSE recovery.

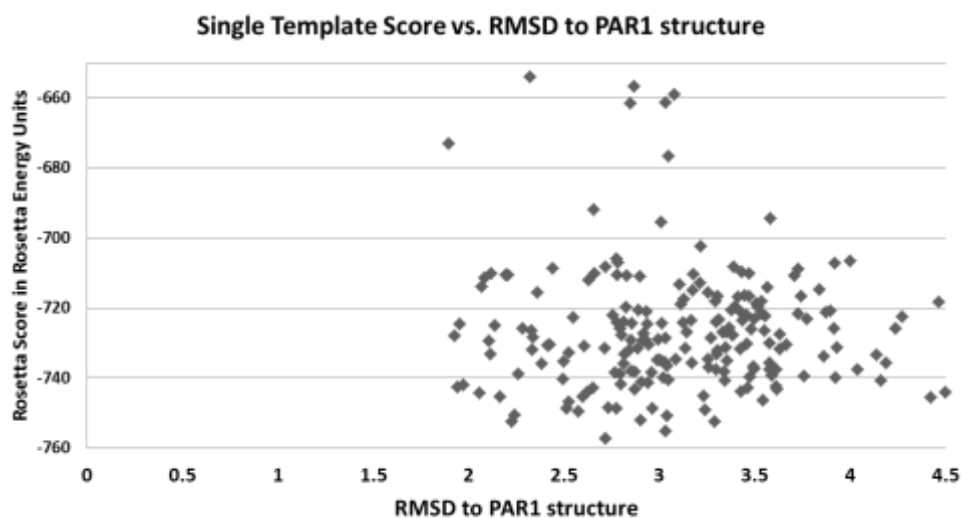
<b>PAR4</b>	<b>PAR1</b>	<b>SSE</b>
<b>A231</b>	L262	ECL2
<b>Q232</b>	L263	ECL2
<b>P310</b>	L340	ECL3
<b>P312</b>	H342	ECL3
<b>A314</b>	S344	ECL3
<b>W315</b>	T345	ECL3

**Table A.4: Selected point mutations for binding pocket mapping.** PAR4 sequences were mutated to the corresponding PAR1 amino acid. Column 1 – Residues and amino acid sequences from human PAR4 to be mutated to the amino acid from column 2. Column 2 – Residues and amino acid identities of human PAR1. Column 3 – Location of residue along PAR1 crystal structure and the predicted region of PAR4.

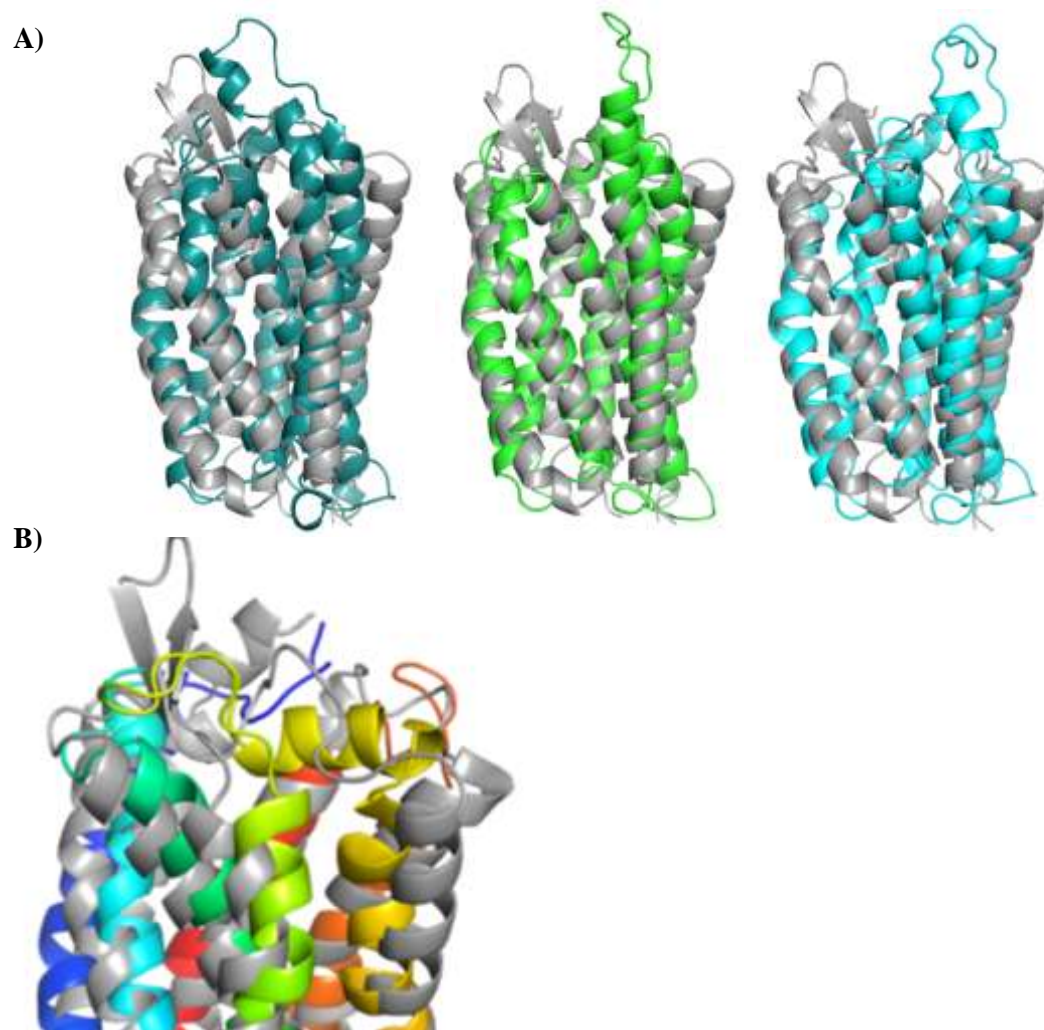


**Figure A.2: PAR4 and PAR1 sequence alignment.** The full length human PAR4 and PAR1 sequences were aligned using ClustalOmega<sup>90</sup> with conserved residues highlighted in indigo. PAR1 crystal structure secondary structure elements (SSE) are depicted as red transmembrane (TM) helices (1-7) and blue beta strands. The truncated N- and C- terminus of the PAR1 crystal structure are denoted by black horizontal lines. PAR4 secondary structure element predictions and transmembrane spanning predictions from PSIPRED<sup>362</sup> and OCTOPUS<sup>363</sup> are shown. Conserved cysteine residues involved in a disulfide bond are denoted by orange arrows. Residue numbering refers to the PAR4 sequence.

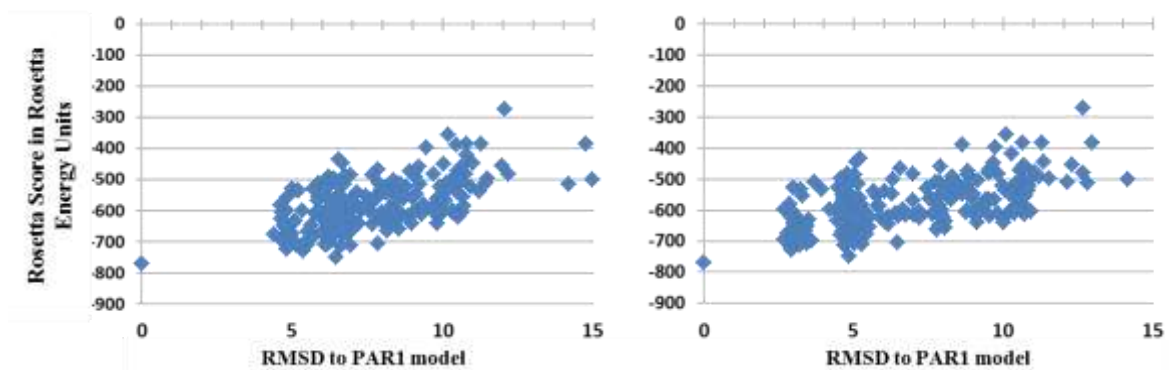




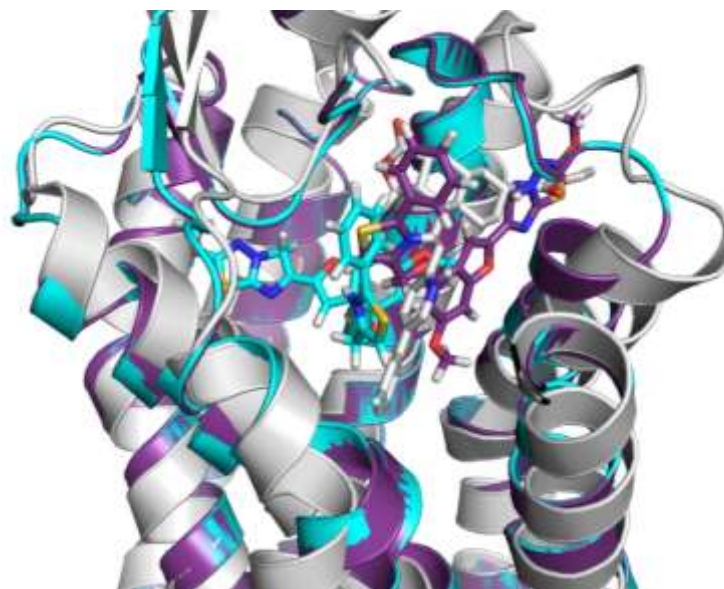
**Figure A.3: PAR4 Single Template score vs RMSD to PAR1 structure.** 250 preliminary decoys were created of the PAR4 structure threaded onto the PAR1 backbone template. Initial score versus RMSD plot shows that a high structural similarity to the starting template.



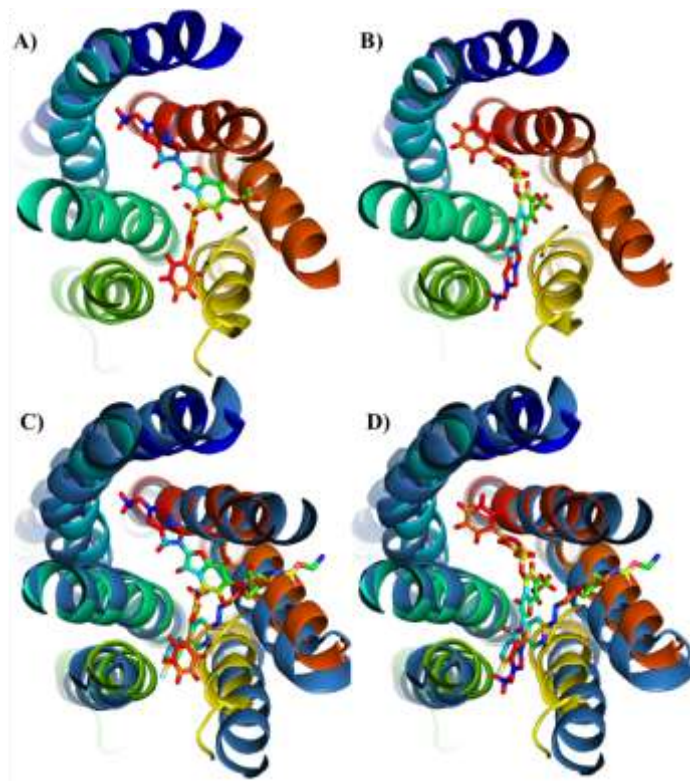
**Figure A.4. Multi-Template Comparative Modeling does not accurately reflect extracellular loop (ECL) regions.** A) Three top models as representatives of the 20 template modeling approach. RosettaCM prefers helices in extracellular loops despite a third of the input templates maintaining  $\beta$ -hairpins within their ECL2 regions. B) Representative model with ECL2 and ECL3 occlude binding pocket. Grey – relaxed PAR1 structure. Color – representative top PAR4 multi-template comparative models by Rosetta Score.



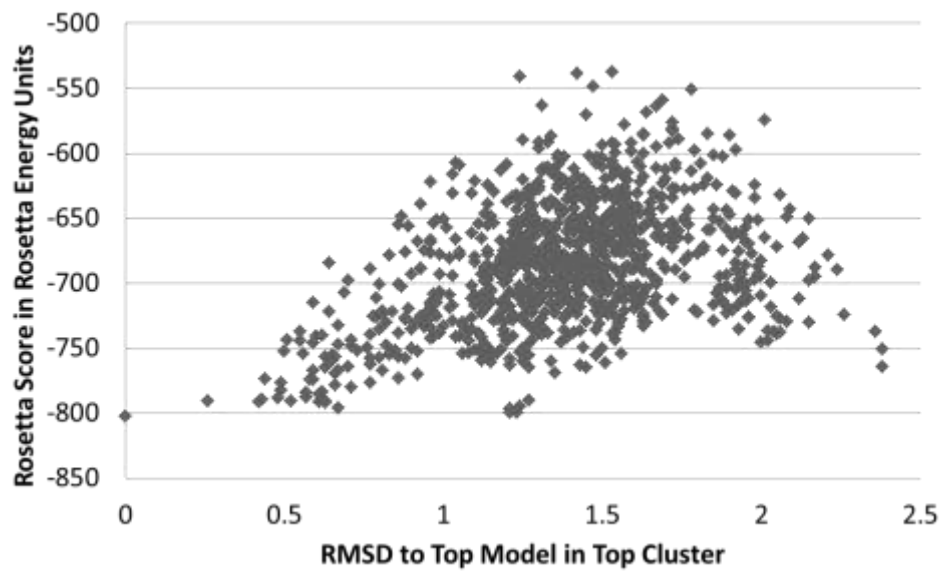
**Figure A.5. Score versus RMSD for 20 template comparative modeling.** A) All-residue RMSD of preliminary runs of multi-template comparative modeling as compared to an energy minimized PAR1 model. B) RMSD of transmembrane (TM) residues only.



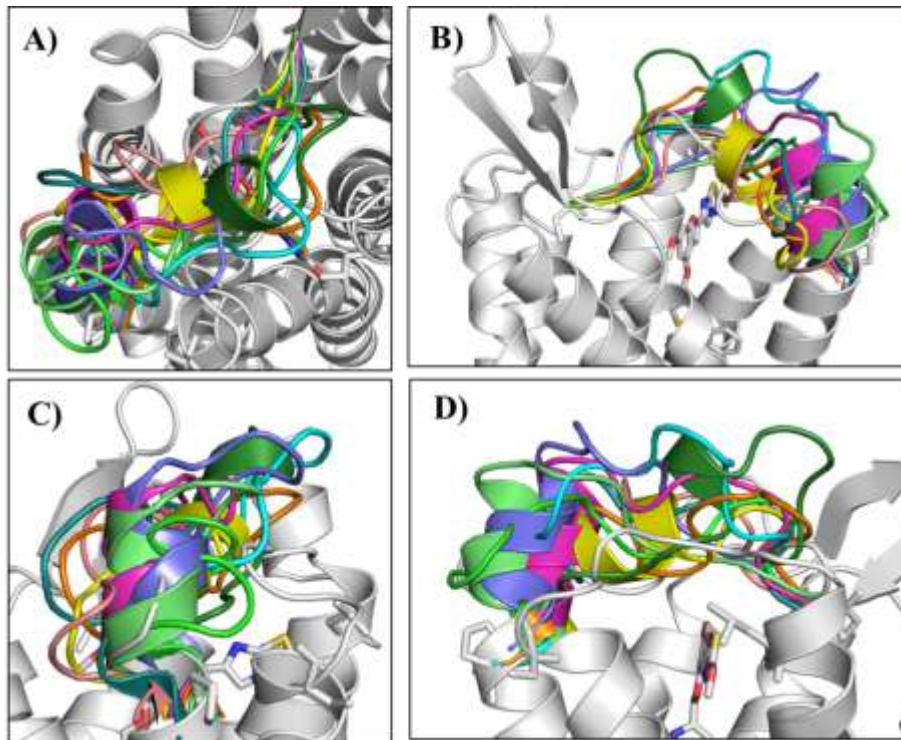
**Figure A.6: Representative Single Template BMS-3 Docking.** Single template docking of BMS-3 resulted in several different binding poses away from the PAR1-Vorapaxar pose (white). BMS-3 preferentially interacted deeper in the protein core (cyan) or with the extracellular loops (purple). The C-terminal section of ECL2 and the top of TM5 were removed for clarity. TM6 and TM7 of the single template models pack tighter than PAR1.



**Figure A.7: Two primary binding modalities for 20 Template docking of BMS-3.** BMS-3 preferentially bound across the protein core of PAR4 as opposed to interacting with the extracellular loops. A) Primary binding pose of BMS-3 found in largest cluster across the core of PAR4. B) Secondary binding pose found in largest cluster. TM1- navy, TM2- light blue, TM3- cyan, TM4- green, TM5- yellow, TM6- orange, TM7- red. ECL2 and 3 are removed for clarity. C) Overlay of A) with PAR1 structure (blue) and Vorapaxar which binds between TM4-5 and TM6-7. D) Overlay of B) with PAR1 structure (blue) and Vorapaxar which binds between TM4-5 and TM6-7. ECL2 and 3 are removed for clarity.

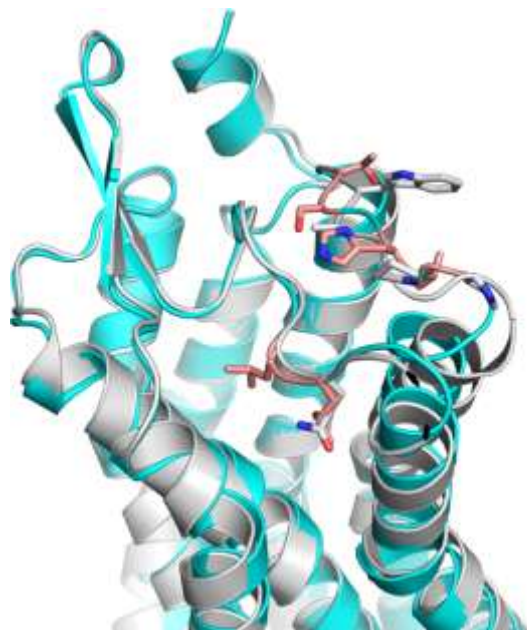


**Figure A.8. Score versus RMSD for multi-template docking.** A) RMSD of preliminary runs of multi-template docking trials of BMS-3 as compared to the top decoy in the top cluster.



**Figure A.9: Representative loop remodeling trials around a docked BMS-3 ligand.** ECL2 was remodeled from the cysteine involved in the disulfide to TM3 to the proline at the top of TM6. A) Top view of the best model of PAR4 with BMS-3 docked (white). Loop structures (colored) show some propensity to form helices and/or interact with ECL3. B) Side view of A. C) Side view of B rotated 90° to the left. D) Side view of C rotated 90° to the left.





**Figure A.10: Six mutations of PAR4 to PAR1 sequence.** PAR4 models with WT sequence (white) were individual mutated to PAR1 sequence for a total of six individual point mutants (blue model, mutations in pink). A combination of all six mutations was also made to better map the ligand binding pocket. BMS-3 and Vorapaxar were then re-docked into the mutant models to evaluate the predictive ability of the models.



## CONCLUDING REMARKS AND FUTURE DIRECTIONS

### *G protein Evolution*

Heterotrimeric G proteins are critically important elements for the transduction of extracellular signals into intracellular responses. Their evolution across eukaryotes highlights their necessity as all branches of Eukaryota possess different components of G protein-mediated signaling. Across this evolution, each organism has modified these components to fit its own unique environment, providing a rich diversity for sequences, novel interactions, signal regulation, and downstream cascade responses. Though there is still much to understand, it is clear that G proteins were present in the last common eukaryotic ancestor and have shaped intracellular signaling significantly across evolution.

Improved long-read sequencing techniques and genome assembly algorithms have greatly aided in our understanding of genetics and evolution. Future sequencing efforts focusing on more basal organisms from species such as non-terrestrial plants and algae to species in other branches like excavates, rhizarians, and chromalveolates will continue to broaden our knowledge of G protein evolution across these organisms.

Future development of curation algorithms will be necessary to accurately annotate, reflect and assess G protein sequences across evolution. Presently, there are several new genome curation and annotation software suites available. Improvement to the ExonMatchSolver, used herein, should focus on incorporating accurate gene models and pseudogene search criteria to assist in the differentiation of functional versus inert genes. In addition, curation of tandem gene pairs when on a single scaffold is still difficult to assess using the ExonMatchSolver. Instead development of multi-hit, multi-gene analysis could be incorporated to resolve these clashes which are currently treated as redundant, and therefore irrelevant, data matches.

Application specific future work should evaluate the transition between GEF-mediated versus spontaneous self-activation of the G protein  $\alpha$  subunit, as it remains a large open question in the field. These two largely conflicting views on which system of signal regulation arose first are confounding, as both camps have found evidence of the necessary signaling components in different “ancient” species of excavates. Therefore, future functional studies, in conjunction with deeper species genome sequencing efforts, are required to ascertain if the  $G\alpha$  subunits from these species are capable of unassisted nucleotide exchange or if a 7TM receptor, like a GPCR, is required. These functional studies will also give insight on the dynamics of the system through analysis of the rates of nucleotide exchange, nucleotide hydrolysis, heterotrimer dissociation versus association, intracellular localization patterns, and preferential protein-protein couplings.

### ***G protein structure***

The G protein  $\alpha$  subunit appears to have maintained its tertiary fold across evolution. Similar to small or monomeric G proteins such as the Ras/Rho and elongation factors families, the  $\alpha$  subunit of the heterotrimeric G protein possesses a nucleotide binding domain with catalytic ability to hydrolyze GTP. This GTPase domain maintains several conserved motifs across the different families of G proteins for facilitating nucleotide interaction. The kinetics of these interactions are modified by subtle changes in amino acid characteristics around these motifs.

Specifically within the heterotrimeric  $G\alpha$  subunit, in addition to variants within the GTPase domain, an evolutionarily recent helical domain alters the binding and association rates of the subunit by occluding the binding pocket in different signaling states. This helical domain, named for its array of helical bundles, remains in a closed position around

the nucleotide binding pocket when the  $G\alpha$  subunit is in its basal state. Its opening is induced, in Metazoans at least, through the activation of a GPCR which couples to the GTPase domain and initiates an allosteric conformational change through the protein to release the helical domain away from the pocket. This clam-shell like opening of the helical domain away from the GTPase domain then returns to a closed state upon the binding of a new nucleotide. This structural rearrangement is critical to the activation of the G protein and downstream signaling cascades.

Therefore, to understand this structural rearrangement and the order of events leading to G protein activation, one must understand the underlying residue characteristics across the communication network of the protein. The helical domain of the  $G\alpha$  subunit is by far the most flexible in terms of sequence identity though its tertiary structure has been shown to be maintained in both plants and animal crystal structures. Yet this helical region possesses the most sequence variability across different species of eukaryotes and even across the subfamilies of  $G\alpha$  within Metazoans. Therefore, understanding the sequence underlying the conserved tertiary structure is paramount to coupling the function of the protein with the intrinsic dynamics of the system.

The structural rearrangements are not made on a per-residue basis, but through residue-residue interactions which propagate across the protein allosterically between the different signaling states. The specific structure and nuanced conformations of the  $G\alpha$  subunit strictly dictates its affinity for its numerous protein partners along the signaling cycle. Therefore, understanding how the  $G\alpha$  subunit moves and transitions between these different conformations at the per-residue level will shed light on the mechanism of G protein activation and function.

In order to interrogate these networks herein, Rosetta Energy scores were used as a predictive means to evaluate residue-residue interactions. This approach lends significant power as it is computationally very inexpensive to run these analyses. Focused on the  $G\alpha 1$  paralog, we assessed the variance of the  $\alpha$  subunit across three different signaling states. However, there are more possible signaling states and transition states between the selected basal, ternary, and activated  $G\alpha$  states. Future work could evaluate the complexes of interaction such as the  $RGS+G\alpha(GTP)$  and  $RGS+G\alpha(GDP+Pi)$  states, the inactive  $G\alpha$  subunit after hydrolysis, and  $G\alpha(GTP)+effector$  complexes.

Confounding these future directions are the dynamics of the G protein signaling system; there are many structural transition states the subunits must pass through to reach the next signaling state along the cycle. Therefore, one primary limitation of the current study was the lack of dynamics assayed within our system. To partially overcome this challenge, I have modeled the G protein subunit across multiple signaling states, and within each of these states, I have created ensembles of models that are more representative of the conformational flexibility the protein possesses. However, though these ensembles capture some degree of conformational flexibility, only the average residue-residue energetics were evaluated herein. Therefore, this diversity was reduced to a single representative number which mutes and partially nullifies the goal of evaluating more conformation. Future approaches will require more evaluation of the standard deviations of these ensembles to assess how flexible these communication networks truly are.

In addition, this work focuses solely on the two body interactions along the backbone and side chain atoms of each residue. To gain a fuller understanding of how this communication map propagates information, two, three and four body interaction maps are

needed. The two body interactions oversimplify the system into a 2D coordinate plane. Evaluation of three and four body interactions, though not possible in the current Rosetta framework, would shed light on the 3D interaction coordinates and how the communication maps actually function within a network or web of residue-residue interactions.

As stated before, these three body and four body interaction maps are not possible in the current Rosetta framework, nor are there plans to develop an interface for such analyses. Instead, a more realistic future direction for this work would be to pursue reanalysis of these communication maps and  $\Delta\Delta G$  calculations with accurate reference energies reflective of the unfolded state of the protein and to more accurately take into account the orbitals around these residues using an updated score function. At present, the Rosetta score function gives some preference to interactions with more hydrogen bonds available and inaccurately reflects certain atomic orbitals; thus the Rosetta score function biases the network maps away from the interaction energies of certain amino acids. When calculating interaction energy maps, and more specifically for  $\Delta\Delta G$  calculations, evaluating the unfolded energy landscape of the protein on a per residue basis would alleviate some of the biases seen in the present implementation and results. In addition, updating these calculations using a more recent score function which takes into account the correct orbital positions may change how the network is created and scored.

### ***G protein subfamilies***

In addition to updating the score function used to analyze the communication networks across the  $G\alpha$  subunit, it is equally important to assess the variances of networks between different G protein subfamilies. There are five known  $G\alpha$  subunit families in vertebrates, four of which are present in the human genome ( $G\alpha_i$ , q, s, and 12). These four families are

subdivided by sequence and downstream effector function. There are 16 different paralogs total in humans, eight are classified in the G $\alpha$ i family, four in G $\alpha$ q, and two each in G $\alpha$ s and G $\alpha$ 12. Representative members of these families have been crystalized and show a high degree of tertiary structure conservation. Indeed, there is very little structural deviation at first glance between these subfamilies.

In order to create a more nuanced functional picture of these families, per-residue calculations of Rosetta energy scores could be created to map the variances across these proteins' communication maps. These different families preferentially couple to different receptors, different downstream effectors, and exchange and catalyze nucleotide hydrolysis at different rates. Therefore, there should be predictive differences in their communication maps along the  $\alpha$ 5 and  $\alpha$ N helices for GPCR interaction and selectivity, along their switch regions, and within their GTPase/helical bundle interfaces, respectively.

Future studies should create ensembles of the five different G protein families across their specific signaling states (basal trimer, the ternary complex, activated monomer, activated monomer bound to RGS or effectors, inactivated monomer etc.). These conformational models should then be assessed for their intrinsic communication network maps. Deviations in these networks between families would shed light on which positions of the G protein are necessary for GPCR interaction, and which contain information for selectivity of interaction.

In conjunction, the exact order of G $\alpha$  activation and subsequent dissociation from the ternary complex remains unknown. At present, biochemical studies using crosslinking may shed light on which regions of the proteins are responsible for propagating information across the complex. However, the use of *in silico* thermodynamic predictions through

Rosetta could answer targeted questions about which residues within the helical domain are necessary for helical domain opening in response to GPCR binding. Future biochemical and computational studies should iteratively evaluate which residues, and residue positions across G protein families, are responsible for the integrity of this message and further relay it to the GPCR and G $\beta\gamma$  subunits.  $\Delta\Delta G$  and Rosetta energy scores could be used to predict these nuanced variances which in turn could be biochemically validated through site-specific point mutations and crosslinking studies.

Other questions which could be answered through such a study include analysis of the rate of GTP hydrolysis. Specifically it is not known how the different regulators of G protein signaling assist in the catalysis and timing of nucleotide hydrolysis coupled with subcellular localization and signaling integrity. This is especially true in the eye where various regulators and interaction partners ensure G $\alpha$  signaling only when specific proteins are present. *In silico* work modeling the nucleotide binding pocket in the presence and absence of Mg<sup>2+</sup>, RGS proteins, soluble GAP proteins, 7TM-GAPS, different effector proteins, and other modifiers of GTP hydrolysis would assist in understanding this highly regulated system.

### ***Combining G protein evolution and structure***

Each of the aforementioned future directions look at either the sequences of the G $\alpha$  subunit to understand mechanism and regulation, or the structural determinants underlying the thermodynamics of the system. A more encompassing future scope for these studies should include elements of both sequence and structural constraints to assess G protein activation and function.

Sequence conservation is the most common and straightforward method employed to

evaluate the significance of a residue at a given position. However, conservation of an individual amino acid ignores the evolutionary constraints imbued on the rest of the sequence to maintain the integrity of the given amino acid. Therefore, more advanced mathematical modelings of evolutionary constraints look at co-evolving pairs of residues, statistically coupled residues, and directly coupled residues to evaluate how a protein's sequence was maintained and/or modified. These different analyses (among many others) provide “scores” for different networks of amino acids across many positions of a multiple sequence alignment. These scores in turn, can act as constraints when evaluating the communication networks of protein function and give insight into which positions of a protein sequence were maintained for structure or function and which were more amenable to modification and change.

Both structure-based and sequence-based methods which aim to evaluate amino acid significance for function contain errors. These errors come from innumerable sources such as limited knowledge-based approximations, limitations of sample size, inaccuracies in the input data curation, and the use of simplified, heuristics or shortcuts required to decrease computational time. In order to overcome these challenges, future work should focus on incorporating data from many structure- and sequence-derived methodologies in order to filter for false positives and amplify true positives between the different metrics. Incorporation of multiple powerful computational methods lends more predictive probability to identifying critical residues necessary for maintaining structure and function.

Future work focused on G protein activation should evaluate metrics such as direct coupling analyses, statistical coupling analyses, conservation, and metrics such as mutual information to use as different lenses with different degrees of accuracy and sensitivity to



answer the same questions: which residue positions are necessary versus sufficient for G protein activation across all G $\alpha$  families and within individual subfamilies? Which positions convey the ability to interact with a 7TM receptor, and which convey selectivity? Which positions along the GTPase and helical domain interfaces are preferred for self-activation versus GEF-regulated? Which residue positions have been constrained across evolution to maintain effector interaction and selectivity? Which are constrained to induce nucleotide hydrolysis and which positions could potentially modify this rate?

These predictions, from multiple sequence-based algorithms, should then be coupled with the aforementioned structural studies evaluating G $\alpha$  structural dynamics across signaling states and within specific subfamilies. The Rosetta modeling, sampling, and scoring methodologies employed herein use a combination of physics and knowledge-based potentials. Intrinsic limitations in the databases used to derive each of the knowledge-based terms, limitations in modeling heuristics, stochastic Monte Carlo sampling, and approximations of atomic details all prevent Rosetta from flawlessly recapitulating nature. As Rosetta does not reflect a physiologically relevant trajectory of possible structural conformations, careful design and evaluation of input and output models is necessary. Even so, these limitations result in baseline level of error and insensitivity which can only be overcome through the coupling of multiple structure- and sequence-based methods.

At present, there is some EPR and DEER information available for certain families and mutants of the G $\alpha$  subunit. Future efforts should incorporate these constraints into more accurate models of G protein activation. In addition, much is known about the interface of the GPCR with the G $\alpha$  subunit through mutational and crystallographic studies. These

constraints should be used to improve our current understanding of the high affinity ternary complex.

Additional biochemical mutants and crosslinking assays should also be fed into this model in an iterative cycle. The predictive power of the residue-residue energy maps must now be evaluated *in vitro* through selective mutation and crosslinking studies. These should be coupled with the sequence-based studies which will highlight additional residue pairs or confirm structure-derived pairs predicted as necessary to maintain either structure or function. Some of these studies are currently underway. However, larger scale point mutations, paired mutations, and reversal mutations of residue-residue interactions previously reported as critical would shed more light on the validity of our computational approach. In addition, there are several crystallographic efforts being made to crystalize the various mutants at positions calculated to be critical. The results of these studies should be fed back into the computational modeling efforts as restraints for the next round of calculations and predictions.

Expanding these *in vitro* studies into the different G $\alpha$  subfamilies remains a challenge. Future collaborations with external labs with the technology and expertise to express and purify these subunits will be necessary for the successful integration of the computations into working models of G protein activation. At present, collaboration with the Sondek labs for G $\alpha_q$  represent the first steps towards a unified, cross-family model. Future efforts linking their mutational data with the predictive power of different bioinformatics approaches will be necessary. To date, the Capra lab has developed several such methods evaluating the presence and propensity of specific mutations to cluster across proteins based on databases of cancerous and benign mutations. Future work linking these

collaborations with our own interdisciplinary work will result in a more holistic picture of G protein mechanics.

## BIBLIOGRAPHY

- [1] Bradford, W., Buckholz, A., Morton, J., Price, C., Jones, A. M., and Urano, D. (2013) Eukaryotic G protein signaling evolved to require G protein-coupled receptors for activation, *Sci Signal* 6, ra37.
- [2] de Mendoza, A., Sebé-Pedrós, A., and Ruiz-Trillo, I. (2014) The evolution of the GPCR signaling system in eukaryotes: modularity, conservation, and the transition to metazoan multicellularity, *Genome Biol Evol* 6, 606-619.
- [3] Cassel, D., and Selinger, Z. (1977) Mechanism of adenylyl cyclase activation by cholera toxin: inhibition of GTP hydrolysis at the regulatory site, *Proceedings of the National Academy of Sciences of the United States of America* 74, 3307-3311.
- [4] Brandt, D. R., Asano, T., Pedersen, S. E., and Ross, E. M. (1983) Reconstitution of catecholamine-stimulated guanosinetriphosphatase activity, *Biochemistry* 22, 4357-4362.
- [5] Hildebrandt, J. D., Sekura, R. D., Codina, J., Iyengar, R., Manclark, C. R., and Birnbaumer, L. (1983) Stimulation and inhibition of adenylyl cyclases mediated by distinct regulatory proteins, *Nature* 302, 706-709.
- [6] Miki, N., Keirns, J. J., Marcus, F. R., Freeman, J., and Bitensky, M. W. (1973) Regulation of cyclic nucleotide concentrations in photoreceptors: an ATP-dependent stimulation of cyclic nucleotide phosphodiesterase by light, *Proceedings of the National Academy of Sciences of the United States of America* 70, 3820-3824.
- [7] HOKIN, M. R., and HOKIN, L. E. (1953) Enzyme secretion and the incorporation of P32 into phospholipides of pancreas slices, *The Journal of biological chemistry* 203, 967-977.
- [8] Putney, J. W. (1986) A model for receptor-regulated calcium entry, *Cell Calcium* 7, 1-12.
- [9] Gilman, A. G. (1987) G proteins: transducers of receptor-generated signals, *Annu Rev Biochem* 56, 615-649.
- [10] Fredriksson, R., Lagerström, M. C., Lundin, L. G., and Schiöth, H. B. (2003) The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints, *Mol Pharmacol* 63, 1256-1272.
- [11] Hamm, H. E. (1998) The Many Faces of G Protein Signaling, *Journal of Biological Chemistry* 273, 669-672.
- [12] Anantharaman, V., Abhiman, S., de Souza, R. F., and Aravind, L. (2011) Comparative genomics uncovers novel structural and functional features of the heterotrimeric GTPase signaling system, *Gene* 475, 63-78.
- [13] Krishnan, A., Mustafa, A., Almén, M. S., Fredriksson, R., Williams, M. J., and Schiöth, H. B. (2015) Evolutionary hierarchy of vertebrate-like heterotrimeric G protein families, *Molecular phylogenetics and evolution* 91, 27-40.

- [14] Lokits, A. D., Indrischek, H., Meiler, J., Hamm, H. E., Stadler, P. F. (2017) Tracing the evolution of the heterotrimeric G protein  $\alpha$  subunit in Deuterostomes *BMC Evolutionary Biology*, *Manuscript in submission*.
- [15] Hurley, J. B., Simon, M. I., Teplow, D. B., Robishaw, J. D., and Gilman, A. G. (1984) Homologies between signal transducing G proteins and ras gene products, *Science* 226, 860-862.
- [16] Halliday, K. R. (1983) Regional homology in GTP-binding proto-oncogene products and elongation factors, *J Cyclic Nucleotide Protein Phosphor Res* 9, 435-448.
- [17] Tanabe, T., Nukada, T., Nishikawa, Y., Sugimoto, K., Suzuki, H., Takahashi, H., Noda, M., Haga, T., Ichiyama, A., and Kangawa, K. (1985) Primary structure of the alpha-subunit of transducin and its relationship to ras proteins, *Nature* 315, 242-245.
- [18] Lochrie, M. A., Hurley, J. B., and Simon, M. I. (1985) Sequence of the alpha subunit of photoreceptor G protein: homologies between transducin, ras, and elongation factors, *Science* 228, 96-99.
- [19] Yatsunami, K., and Khorana, H. G. (1985) GTPase of bovine rod outer segments: the amino acid sequence of the alpha subunit as derived from the cDNA sequence, *Proceedings of the National Academy of Sciences of the United States of America* 82, 4316-4320.
- [20] Downes, G. B., and Gautam, N. (1999) The G protein subunit gene families, *Genomics* 62, 544-552.
- [21] Simon, M. I., Strathmann, M. P., and Gautam, N. (1991) Diversity of G proteins in signal transduction, *Science* 252, 802-808.
- [22] Oka, Y., Saraiva, L. R., Kwan, Y. Y., and Korsching, S. I. (2009) The fifth class of G $\alpha$  proteins, *Proceedings of the National Academy of Sciences of the United States of America* 106, 1484-1489.
- [23] Suki, W. N., Abramowitz, J., Mattera, R., Codina, J., and Birnbaumer, L. (1987) The human genome encodes at least three non-allelic G proteins with alpha i-type subunits, *FEBS Lett* 220, 187-192.
- [24] Itoh, H., Toyama, R., Kozasa, T., Tsukamoto, T., Matsuoka, M., and Kaziro, Y. (1988) Presence of three distinct molecular species of G $\alpha$ i protein alpha subunit. Structure of rat cDNAs and human genomic DNAs, *The Journal of biological chemistry* 263, 6656-6664.
- [25] Wettschureck, N., and Offermanns, S. (2005) Mammalian G proteins and their cell type specific functions, *Physiol Rev* 85, 1159-1204.
- [26] McLaughlin, S. K., McKinnon, P. J., and Margolskee, R. F. (1992) Gustducin is a taste-cell-specific G protein closely related to the transducins, *Nature* 357, 563-569.
- [27] Hsu, W. H., Rudolph, U., Sanford, J., Bertrand, P., Olate, J., Nelson, C., Moss, L. G., Boyd, A. E., Codina, J., and Birnbaumer, L. (1990) Molecular cloning of a novel splice variant of the

- alpha subunit of the mammalian Go protein, *The Journal of biological chemistry* 265, 11220-11226.
- [28] Strathmann, M., Wilkie, T. M., and Simon, M. I. (1990) Alternative splicing produces transcripts encoding two forms of the alpha subunit of GTP-binding protein Go, *Proceedings of the National Academy of Sciences of the United States of America* 87, 6477-6481.
- [29] Fong, H. K., Yoshimoto, K. K., Eversole-Cire, P., and Simon, M. I. (1988) Identification of a GTP-binding protein alpha subunit that lacks an apparent ADP-ribosylation site for pertussis toxin, *Proceedings of the National Academy of Sciences of the United States of America* 85, 3066-3070.
- [30] Strathmann, M., and Simon, M. I. (1990) G protein diversity: a distinct class of alpha subunits is present in vertebrates and invertebrates, *Proceedings of the National Academy of Sciences of the United States of America* 87, 9113-9117.
- [31] Exton, J. H. (1996) Regulation of phosphoinositide phospholipases by hormones, neurotransmitters, and other agonists linked to G proteins, *Annu Rev Pharmacol Toxicol* 36, 481-509.
- [32] Rhee, S. G. (2001) Regulation of phosphoinositide-specific phospholipase C, *Annu Rev Biochem* 70, 281-312.
- [33] Offermanns, S., Heiler, E., Spicher, K., and Schultz, G. (1994) Gq and G11 are concurrently activated by bombesin and vasopressin in Swiss 3T3 cells, *FEBS Lett* 349, 201-204.
- [34] Wu, D., Katz, A., Lee, C. H., and Simon, M. I. (1992) Activation of phospholipase C by alpha 1-adrenergic receptors is mediated by the alpha subunits of Gq family, *The Journal of biological chemistry* 267, 25798-25802.
- [35] Xu, X., Croy, J. T., Zeng, W., Zhao, L., Davignon, I., Popov, S., Yu, K., Jiang, H., Offermanns, S., Muallem, S., and Wilkie, T. M. (1998) Promiscuous coupling of receptors to Gq class alpha subunits and effector proteins in pancreatic and submandibular gland cells, *The Journal of biological chemistry* 273, 27275-27279.
- [36] Nakamura, F., Ogata, K., Shiozaki, K., Kameyama, K., Ohara, K., Haga, T., and Nukada, T. (1991) Identification of two novel GTP-binding protein alpha-subunits that lack apparent ADP-ribosylation sites for pertussis toxin, *The Journal of biological chemistry* 266, 12676-12681.
- [37] Wilkie, T. M., Scherle, P. A., Strathmann, M. P., Slepak, V. Z., and Simon, M. I. (1991) Characterization of G-protein alpha subunits in the Gq class: expression in murine tissues and in stromal and hematopoietic cell lines, *Proceedings of the National Academy of Sciences of the United States of America* 88, 10049-10053.
- [38] Jones, D. T., and Reed, R. R. (1989) Golf: an olfactory neuron specific-G protein involved in odorant signal transduction, *Science* 244, 790-795.

- [39] Strathmann, M. P., and Simon, M. I. (1991) G alpha 12 and G alpha 13 subunits define a fourth class of G protein alpha subunits, *Proceedings of the National Academy of Sciences of the United States of America* 88, 5582-5586.
- [40] Buhl, A. M., Johnson, N. L., Dhanasekaran, N., and Johnson, G. L. (1995) G alpha 12 and G alpha 13 stimulate Rho-dependent stress fiber formation and focal adhesion assembly, *The Journal of biological chemistry* 270, 24631-24634.
- [41] Birnbaumer, L., and Rodbell, M. (1969) Adenyl cyclase in fat cells. II. Hormone receptors, *The Journal of biological chemistry* 244, 3477-3482.
- [42] Oka, Y., and Korsching, S. I. (2009) The fifth element in animal Galpha protein evolution, *Commun Integr Biol* 2, 227-229.
- [43] Nei, M., and Rooney, A. P. (2005) Concerted and birth-and-death evolution of multigene families, *Annu Rev Genet* 39, 121-152.
- [44] Ohta, T. (1980) *Evolution and Variation of Multigene Families*, 1 ed., Springer-Verlag Berlin Heidelberg, Berlin Heidelberg New York.
- [45] Hurowitz, E. H., Melnyk, J. M., Chen, Y. J., Kouros-Mehr, H., Simon, M. I., and Shizuya, H. (2000) Genomic characterization of the human heterotrimeric G protein alpha, beta, and gamma subunit genes, *DNA Res* 7, 111-120.
- [46] Wilkie, T. M., Gilbert, D. J., Olsen, A. S., Chen, X. N., Amatruda, T. T., Korenberg, J. R., Trask, B. J., de Jong, P., Reed, R. R., and Simon, M. I. (1992) Evolution of the mammalian G protein alpha subunit multigene family, *Nat Genet* 1, 85-91.
- [47] Nordström, K., Larsson, T. A., and Larhammar, D. (2004) Extensive duplications of phototransduction genes in early vertebrate evolution correlate with block (chromosome) duplications, *Genomics* 83, 852-872.
- [48] Larhammar, D., Nordström, K., and Larsson, T. A. (2009) Evolution of vertebrate rod and cone phototransduction genes, *Philos Trans R Soc Lond B Biol Sci* 364, 2867-2880.
- [49] Lagman, D., Sundström, G., Ocampo Daza, D., Abalo, X. M., and Larhammar, D. (2012) Expansion of transducin subunit gene families in early vertebrate tetraploidizations, *Genomics* 100, 203-211.
- [50] Lamb, T. D., Patel, H., Chuah, A., Natoli, R. C., Davies, W. I., Hart, N. S., Collin, S. P., and Hunt, D. M. (2016) Evolution of Vertebrate Phototransduction: Cascade Activation, *Mol Biol Evol* 33, 2064-2087.
- [51] Ohno, S. (1999) Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999, *Semin Cell Dev Biol* 10, 517-522.
- [52] Putnam, N., Butts, T., Ferrier, D., Furlong, R., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J., Benito-Gutierrez, E., Dubchak, I., Garcia-Fernandez, J., Gibson-Brown, J., Grigoriev, I., Horton, A., de Jong, P., Jurka, J., Kapitonov, V., Kohara, Y., Kuroki, Y., Lindquist, E., Lucas, S., Osoegawa, K., Pennacchio, L., Salamov, A., Satou, Y., Sauka-Spengler, T., Schmutz, J., Shin-I, T.,

- Toyoda, A., Bronner-Fraser, M., Fujiyama, A., Holland, L., Holland, P., Satoh, N., and Rokhsar, D. (2008) The amphioxus genome and the evolution of the chordate karyotype, *Nature* 453, 1064-U1063.
- [53] Muradov, H., Kerov, V., Boyd, K. K., and Artemyev, N. O. (2008) Unique transducins expressed in long and short photoreceptors of lamprey *Petromyzon marinus*, *Vision Res* 48, 2302-2308.
- [54] Oka, Y., and Korsching, S. I. (2011) Shared and unique G alpha proteins in the zebrafish versus mammalian senses of taste and smell, *Chem Senses* 36, 357-365.
- [55] Ohmoto, M., Okada, S., Nakamura, S., Abe, K., and Matsumoto, I. (2011) Mutually exclusive expression of  $G\alpha_{i1}$  and  $G\alpha_{i4}$  reveals diversification of taste receptor cells in zebrafish, *J Comp Neurol* 519, 1616-1629.
- [56] Mueller, K. L., Hoon, M. A., Erlenbach, I., Chandrashekar, J., Zuker, C. S., and Ryba, N. J. (2005) The receptors and coding logic for bitter taste, *Nature* 434, 225-229.
- [57] Zhao, G. Q., Zhang, Y., Hoon, M. A., Chandrashekar, J., Erlenbach, I., Ryba, N. J., and Zuker, C. S. (2003) The receptors for mammalian sweet and umami taste, *Cell* 115, 255-266.
- [58] Jansen, G., Thijssen, K. L., Werner, P., van der Horst, M., Hazendonk, E., and Plasterk, R. H. (1999) The complete family of genes encoding G proteins of *Caenorhabditis elegans*, *Nat Genet* 21, 414-419.
- [59] Dhanasekaran, N., and Dermott, J. M. (1996) Signaling by the G12 class of G proteins, *Cell Signal* 8, 235-245.
- [60] Ferguson, K. M., Higashijima, T., Smigel, M. D., and Gilman, A. G. (1986) The influence of bound GDP on the kinetics of guanine nucleotide binding to G proteins, *The Journal of biological chemistry* 261, 7393-7399.
- [61] Johnston, C. A., Taylor, J. P., Gao, Y., Kimple, A. J., Grigston, J. C., Chen, J. G., Siderovski, D. P., Jones, A. M., and Willard, F. S. (2007) GTPase acceleration as the rate-limiting step in Arabidopsis G protein-coupled sugar signaling, *Proceedings of the National Academy of Sciences of the United States of America* 104, 17317-17322.
- [62] Jones, J. C., Duffy, J. W., Machius, M., Temple, B. R., Dohlman, H. G., and Jones, A. M. (2011) The crystal structure of a self-activating G protein alpha subunit reveals its distinct mechanism of signal initiation, *Sci Signal* 4, ra8.
- [63] Jones, J. C., Jones, A. M., Temple, B. R., and Dohlman, H. G. (2012) Differences in intradomain and interdomain motion confer distinct activation properties to structurally similar  $G\alpha$  proteins, *Proceedings of the National Academy of Sciences of the United States of America* 109, 7275-7279.
- [64] Urano, D., Jones, J. C., Wang, H., Matthews, M., Bradford, W., Bennetzen, J. L., and Jones, A. M. (2012) G protein activation without a GEF in the plant kingdom, *PLoS Genet* 8, e1002756.



- [65] Chen, J. G., Willard, F. S., Huang, J., Liang, J., Chasse, S. A., Jones, A. M., and Siderovski, D. P. (2003) A seven-transmembrane RGS protein that modulates plant cell proliferation, *Science* 301, 1728-1731.
- [66] Urano, D., Phan, N., Jones, J. C., Yang, J., Huang, J., Grigston, J., Taylor, J. P., and Jones, A. M. (2012) Endocytosis of the seven-transmembrane RGS1 protein activates G-protein-coupled signalling in Arabidopsis, *Nat Cell Biol* 14, 1079-1088.
- [67] Kimple, A. J., Bosch, D. E., Giguère, P. M., and Siderovski, D. P. (2011) Regulators of G-protein signaling and their  $G\alpha$  substrates: promises and challenges in their use as drug discovery targets, *Pharmacological reviews* 63, 728-749.
- [68] Booker, K. S., Schwarz, J., Garrett, M. B., and Jones, A. M. (2010) Glucose attenuation of auxin-mediated bimodality in lateral root formation is partly coupled by the heterotrimeric G protein complex, *PLoS One* 5.
- [69] Turner, G. E., and Borkovich, K. A. (1993) Identification of a G protein alpha subunit from *Neurospora crassa* that is a member of the  $G_i$  family, *The Journal of biological chemistry* 268, 14805-14811.
- [70] Bölker, M. (1998) Sex and crime: heterotrimeric G proteins in fungal mating and pathogenesis, *Fungal Genet Biol* 25, 143-156.
- [71] Borkovich, K. A. (1996) *Signal Transduction Pathways and Heterotrimeric G proteins*, Springer Berlin Heidelberg.
- [72] Li, L., Wright, S. J., Krystofova, S., Park, G., and Borkovich, K. A. (2007) Heterotrimeric G protein signaling in filamentous fungi, *Annu Rev Microbiol* 61, 423-452.
- [73] Bardwell, L. (2005) A walk-through of the yeast mating pheromone response pathway, *Peptides* 26, 339-350.
- [74] Kays, A. M., and Borkovich, K. A. (2004) *Signal Transduction Pathways Mediated by Heterotrimeric G proteins*, Springer Berlin Heidelberg.
- [75] Firtel, R. A., van Haastert, P. J., Kimmel, A. R., and Devreotes, P. N. (1989) G protein linked signal transduction pathways in development: dictyostelium as an experimental system, *Cell* 58, 235-239.
- [76] Devreotes, P. (1989) Dictyostelium discoideum: a model system for cell-cell interactions in development, *Science* 245, 1054-1058.
- [77] Kumagai, A., Pupillo, M., Gundersen, R., Miake-Lye, R., Devreotes, P. N., and Firtel, R. A. (1989) Regulation and function of G alpha protein subunits in Dictyostelium, *Cell* 57, 265-275.
- [78] Pupillo, M., Kumagai, A., Pitt, G. S., Firtel, R. A., and Devreotes, P. N. (1989) Multiple alpha subunits of guanine nucleotide-binding proteins in Dictyostelium, *Proceedings of the National Academy of Sciences of the United States of America* 86, 4892-4896.

- [79] Hadwiger, J. A., Wilkie, T. M., Strathmann, M., and Firtel, R. A. (1991) Identification of Dictyostelium G alpha genes expressed during multicellular development, *Proceedings of the National Academy of Sciences of the United States of America* 88, 8213-8217.
- [80] Wu, L. J., and Devreotes, P. N. (1991) Dictyostelium transiently expresses eight distinct G-protein alpha-subunits during its developmental program, *Biochem Biophys Res Commun* 179, 1141-1147.
- [81] Lilly, P., Wu, L., Welker, D. L., and Devreotes, P. N. (1993) A G-protein beta-subunit is essential for Dictyostelium development, *Genes Dev* 7, 986-995.
- [82] Klein, P. S., Sun, T. J., Saxe, C. L., Kimmel, A. R., Johnson, R. L., and Devreotes, P. N. (1988) A chemoattractant receptor controls development in Dictyostelium discoideum, *Science* 241, 1467-1472.
- [83] Saxe, C. L., Ginsburg, G. T., Louis, J. M., Johnson, R., Devreotes, P. N., and Kimmel, A. R. (1993) CAR2, a prestalk cAMP receptor required for normal tip formation and late development of Dictyostelium discoideum, *Genes Dev* 7, 262-272.
- [84] Johnson, R. L., Saxe, C. L., Gollop, R., Kimmel, A. R., and Devreotes, P. N. (1993) Identification and targeted gene disruption of cAR3, a cAMP receptor subtype expressed during multicellular stages of Dictyostelium development, *Genes Dev* 7, 273-282.
- [85] Pandey, S. (2011) More (G-proteins) please! Identification of an elaborate network of G-proteins in soybean, *Plant Signal Behav* 6, 780-782.
- [86] Urano, D., Chen, J. G., Botella, J. R., and Jones, A. M. (2013) Heterotrimeric G protein signalling in the plant kingdom, *Open Biol* 3, 120186.
- [87] Hackenberg, D., Sakayama, H., Nishiyama, T., and Pandey, S. (2013) Characterization of the heterotrimeric G-protein complex and its regulator from the green alga Chara braunii expands the evolutionary breadth of plant G-protein signaling, *Plant Physiol* 163, 1510-1517.
- [88] Hackenberg, D., and Pandey, S. (2014) Heterotrimeric G-proteins in green algae. An early innovation in the evolution of the plant lineage, *Plant Signal Behav* 9, e28457.
- [89] Taddese, B., Upton, G. J., Bailey, G. R., Jordan, S. R., Abdulla, N. Y., Reeves, P. J., and Reynolds, C. A. (2014) Do plants contain g protein-coupled receptors?, *Plant Physiol* 164, 287-307.
- [90] Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, *Mol Syst Biol* 7, 539.
- [91] Koepfli, K. P., Paten, B., O'Brien, S. J., and Scientists, G. K. C. o. (2015) The Genome 10K Project: a way forward, *Annu Rev Anim Biosci* 3, 57-111.
- [92] Sodergren, E., and Weinstock, G. M., and Davidson, E. H., and Cameron, R. A., and Gibbs, R. A., and Angerer, R. C., and Angerer, L. M., and Arnone, M. I., and Burgess, D. R., and

Burke, R. D., and Coffman, J. A., and Dean, M., and Elphick, M. R., and Etensohn, C. A., and Foltz, K. R., and Hamdoun, A., and Hynes, R. O., and Klein, W. H., and Marzluff, W., and McClay, D. R., and Morris, R. L., and Mushegian, A., and Rast, J. P., and Smith, L. C., and Thorndyke, M. C., and Vacquier, V. D., and Wessel, G. M., and Wray, G., and Zhang, L., and Elsik, C. G., and Ermolaeva, O., and Hlavina, W., and Hofmann, G., and Kitts, P., and Landrum, M. J., and Mackey, A. J., and Maglott, D., and Panopoulou, G., and Poustka, A. J., and Pruitt, K., and Sapojnikov, V., and Song, X., and Souvorov, A., and Solovyev, V., and Wei, Z., and Whittaker, C. A., and Worley, K., and Durbin, K. J., and Shen, Y., and Fedrigo, O., and Garfield, D., and Haygood, R., and Primus, A., and Satija, R., and Severson, T., and Gonzalez-Garay, M. L., and Jackson, A. R., and Milosavljevic, A., and Tong, M., and Killian, C. E., and Livingston, B. T., and Wilt, F. H., and Adams, N., and Bellé, R., and Carbonneau, S., and Cheung, R., and Cormier, P., and Cosson, B., and Croce, J., and Fernandez-Guerra, A., and Genevière, A. M., and Goel, M., and Kelkar, H., and Morales, J., and Mulner-Lorillon, O., and Robertson, A. J., and Goldstone, J. V., and Cole, B., and Epel, D., and Gold, B., and Hahn, M. E., and Howard-Ashby, M., and Scally, M., and Stegeman, J. J., and Allgood, E. L., and Cool, J., and Judkins, K. M., and McCafferty, S. S., and Musante, A. M., and Obar, R. A., and Rawson, A. P., and Rossetti, B. J., and Gibbons, I. R., and Hoffman, M. P., and Leone, A., and Istrail, S., and Materna, S. C., and Samanta, M. P., and Stolc, V., and Tongprasit, W., and Tu, Q., and Bergeron, K. F., and Brandhorst, B. P., and Whittle, J., and Berney, K., and Bottjer, D. J., and Calestani, C., and Peterson, K., and Chow, E., and Yuan, Q. A., and Elhaik, E., and Graur, D., and Reese, J. T., and Bosdet, I., and Heesun, S., and Marra, M. A., and Schein, J., and Anderson, M. K., and Brockton, V., and Buckley, K. M., and Cohen, A. H., and Fugmann, S. D., and Hibino, T., and Loza-Coll, M., and Majeske, A. J., and Messier, C., and Nair, S. V., and Pancer, Z., and Terwilliger, D. P., and Agca, C., and Arboleda, E., and Chen, N., and Churcher, A. M., and Hallböök, F., and Humphrey, G. W., and Idris, M. M., and Kiyama, T., and Liang, S., and Mellott, D., and Mu, X., and Murray, G., and Olinski, R. P., and Raible, F., and Rowe, M., and Taylor, J. S., and Tessmar-Raible, K., and Wang, D., and Wilson, K. H., and Yaguchi, S., and Gaasterland, T., and Galindo, B. E., and Gunaratne, H. J., and Juliano, C., and Kinukawa, M., and Moy, G. W., and Neill, A. T., and Nomura, M., and Raisch, M., and Reade, A., and Roux, M. M., and Song, J. L., and Su, Y. H., and Townley, I. K., and Voronina, E., and Wong, J. L., and Amore, G., and Branno, M., and Brown, E. R., and Cavalieri, V., and Duboc, V., and Duloquin, L., and Flytzanis, C., and Gache, C., and Lapraz, F., and Lepage, T., and Locascio, A., and Martinez, P., and Matassi, G., and Matranga, V., and Range, R., and Rizzo, F., and Röttinger, E., and Beane, W., and Bradham, C., and Byrum, C., and Glenn, T., and Hussain, S., and Manning, G., and Miranda, E., and Thomason, R., and Walton, K., and Wikramanayake, A., and Wu, S. Y., and Xu, R., and Brown, C. T., and Chen, L., and Gray, R. F., and Lee, P. Y., and Nam, J., and Oliveri, P., and Smith, J., and Muzny, D., and Bell, S., and Chacko, J., and Cree, A., and Curry, S., and Davis, C., and Dinh, H., and Dugan-Rocha, S., and Fowler, J., and Gill, R., and Hamilton, C., and Hernandez, J., and Hines, S., and Hume, J., and Jackson, L., and Jolivet, A., and Kovar, C., and Lee, S., and Lewis, L., and Miner, G., and Morgan, M., and Nazareth, L. V., and Okwuonu, G., and Parker, D., and Pu, L. L., and Thorn, R., and Wright, R., and Consortium, S. U. G. S. (2006) The genome of the sea urchin *Strongylocentrotus purpuratus*, *Science* 314, 941-952.

- [93] Mehta, T. K., Ravi, V., Yamasaki, S., Lee, A. P., Lian, M. M., Tay, B. H., Tohari, S., Yanai, S., Tay, A., Brenner, S., and Venkatesh, B. (2013) Evidence for at least six Hox clusters in the Japanese lamprey (*Lethenteron japonicum*), *Proceedings of the National Academy of Sciences of the United States of America* 110, 16044-16049.

- [94] Karpinka, J. B., Fortriede, J. D., Burns, K. A., James-Zorn, C., Ponferrada, V. G., Lee, J., Karimi, K., Zorn, A. M., and Vize, P. D. (2015) Xenbase, the *Xenopus* model organism database; new virtualized system, data types and genomes, *Nucleic Acids Res* 43, D756-763.
- [95] James-Zorn, C., Ponferrada, V. G., Burns, K. A., Fortriede, J. D., Lotay, V. S., Liu, Y., Brad Karpinka, J., Karimi, K., Zorn, A. M., and Vize, P. D. (2015) Xenbase: Core features, data acquisition, and data processing, *Genesis* 53, 486-497.
- [96] Consortium, U. (2015) UniProt: a hub for protein information, *Nucleic Acids Res* 43, D204-212.
- [97] Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Birney, E., Harrow, J., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S. J., Cunningham, F., Aken, B. L., Zerbino, D. R., and Flicek, P. (2016) Ensembl 2016, *Nucleic Acids Res* 44, D710-716.
- [98] Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T., Howe, K., Kähäri, A., Kokocinski, F., Martin, F. J., Murphy, D. N., Nag, R., Ruffier, M., Schuster, M., Tang, Y. A., Vogel, J. H., White, S., Zadissa, A., Flicek, P., and Searle, S. M. (2016) The Ensembl gene annotation system, *Database (Oxford)* 2016.
- [99] Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J., Searle, S. M., Amode, R., Brent, S., Spooner, W., Kulesha, E., Yates, A., and Flicek, P. (2016) Ensembl comparative genomics resources, *Database (Oxford)* 2016.
- [100] O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufu, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvermin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation, *Nucleic Acids Res* 44, D733-745.
- [101] Gray, K. A., Yates, B., Seal, R. L., Wright, M. W., and Bruford, E. A. (2015) Genenames.org: the HGNC resources in 2015, *Nucleic Acids Res* 43, D1079-1085.
- [102] Indrischek, H., Wieseke, N., Stadler, P. F., and Prohaska, S. J. (2016) The paralog-to-contig assignment problem: high quality gene models from fragmented assemblies, *Algorithms Mol Biol* 11, 1.

- [103] Wheeler, D. L., Church, D. M., Federhen, S., Lash, A. E., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E., Tatusova, T. A., and Wagner, L. (2003) Database resources of the National Center for Biotechnology, *Nucleic Acids Res* 31, 28-33.
- [104] Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. (1993) dbEST--database for "expressed sequence tags", *Nat Genet* 4, 332-333.
- [105] Ouellette, B. F., and Boguski, M. S. (1997) Database divisions and homology search files: a guide for the perplexed, *Genome Res* 7, 952-955.
- [106] Delsuc, F., Brinkmann, H., Chourrout, D., and Philippe, H. (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates, *Nature* 439, 965-968.
- [107] Jarvis, E. D., Mirarab, S., and Aberer, A. J., and Li, B., and Houde, P., and Li, C., and Ho, S. Y., and Faircloth, B. C., and Nabholz, B., and Howard, J. T., and Suh, A., and Weber, C. C., and da Fonseca, R. R., and Li, J., and Zhang, F., and Li, H., and Zhou, L., and Narula, N., and Liu, L., and Ganapathy, G., and Boussau, B., and Bayzid, M. S., and Zavidovych, V., and Subramanian, S., and Gabaldón, T., and Capella-Gutiérrez, S., and Huerta-Cepas, J., and Rekepalli, B., and Munch, K., and Schierup, M., and Lindow, B., and Warren, W. C., and Ray, D., and Green, R. E., and Bruford, M. W., and Zhan, X., and Dixon, A., and Li, S., and Li, N., and Huang, Y., and Derryberry, E. P., and Bertelsen, M. F., and Sheldon, F. H., and Brumfield, R. T., and Mello, C. V., and Lovell, P. V., and Wirthlin, M., and Schneider, M. P., and Prosdocimi, F., and Samaniego, J. A., and Vargas Velazquez, A. M., and Alfaro-Núñez, A., and Campos, P. F., and Petersen, B., and Sicheritz-Ponten, T., and Pas, A., and Bailey, T., and Scofield, P., and Bunce, M., and Lambert, D. M., and Zhou, Q., and Perelman, P., and Driskell, A. C., and Shapiro, B., and Xiong, Z., and Zeng, Y., and Liu, S., and Li, Z., and Liu, B., and Wu, K., and Xiao, J., and Yinqi, X., and Zheng, Q., and Zhang, Y., and Yang, H., and Wang, J., and Smeds, L., and Rheindt, F. E., and Braun, M., and Fjeldsa, J., and Orlando, L., and Barker, F. K., and Jönsson, K. A., and Johnson, W., and Koepfli, K. P., and O'Brien, S., and Haussler, D., and Ryder, O. A., and Rahbek, C., and Willerslev, E., and Graves, G. R., and Glenn, T. C., and McCormack, J., and Burt, D., and Ellegren, H., and Alström, P., and Edwards, S. V., and Stamatakis, A., and Mindell, D. P., and Cracraft, J., and Braun, E. L., and Warnow, T., and Jun, W., and Gilbert, M. T., and Zhang, G. (2014) Whole-genome analyses resolve early branches in the tree of life of modern birds, *Science* 346, 1320-1331.
- [108] Crawford, N. G., Parham, J. F., Sellas, A. B., Faircloth, B. C., Glenn, T. C., Papenfuss, T. J., Henderson, J. B., Hansen, M. H., and Simison, W. B. (2015) A phylogenomic analysis of turtles, *Molecular phylogenetics and evolution* 83, 250-257.
- [109] Stamatakis, A. (2015) Using RAxML to Infer Phylogenies, *Curr Protoc Bioinformatics* 51, 6.14.11-14.
- [110] Edgar, R. C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity, *BMC Bioinformatics* 5, 113.
- [111] Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., and Barton, G. J. (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench, *Bioinformatics* 25, 1189-1191.

- [112] Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, *Bioinformatics* 30, 1312-1313.
- [113] Abascal, F., Zardoya, R., and Posada, D. (2005) ProtTest: selection of best-fit models of protein evolution, *Bioinformatics* 21, 2104-2105.
- [114] Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, *Bioinformatics* 22, 2688-2690.
- [115] Bailey, T. L., and Machanick, P. (2012) Inferring direct DNA binding from ChIP-seq, *Nucleic Acids Res* 40, e128.
- [116] Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., Na, H., Irimia, M., Matzat, L. H., Dale, R. K., Smith, S. A., Yarosh, C. A., Kelly, S. M., Nabet, B., Mecnas, D., Li, W., Laishram, R. S., Qiao, M., Lipshitz, H. D., Piano, F., Corbett, A. H., Carstens, R. P., Frey, B. J., Anderson, R. A., Lynch, K. W., Penalva, L. O., Lei, E. P., Fraser, A. G., Blencowe, B. J., Morris, Q. D., and Hughes, T. R. (2013) A compendium of RNA-binding motifs for decoding gene regulation, *Nature* 499, 172-177.
- [117] Jolma, A., Yan, J., Whittington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T., and Taipale, J. (2013) DNA-binding specificities of human transcription factors, *Cell* 152, 327-339.
- [118] Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C. Y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., and Wasserman, W. W. (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles, *Nucleic Acids Res* 42, D142-147.
- [119] Stothard, P. (2000) The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences, *Biotechniques* 28, 1102, 1104.
- [120] Washietl, S., Findeiss, S., Müller, S. A., Kalkhof, S., von Bergen, M., Hofacker, I. L., Stadler, P. F., and Goldman, N. (2011) RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data, *RNA* 17, 578-594.
- [121] Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood, *Mol Biol Evol* 24, 1586-1591.
- [122] Yang, Z., Wong, W. S., and Nielsen, R. (2005) Bayes empirical bayes inference of amino acid sites under positive selection, *Mol Biol Evol* 22, 1107-1118.
- [123] Martin, D. P., Murrell, B., Golden, M., Khoosal, A., and Muhire, B. (2015) RDP4: Detection and analysis of recombination patterns in virus genomes, *Virus Evol* 1, vev003.
- [124] Bielawski, J. P., Baker, J. L., and Mingrone, J. (2016) Inference of Episodic Changes in Natural Selection Acting on Protein Coding Sequences via CODEML, *Curr Protoc Bioinformatics* 54, 6.15.11-16.15.32.

- [125] Mingrone, J., Susko, E., and Bielawski, J. (2016) Smoothed Bootstrap Aggregation for Assessing Selection Pressure at Amino Acid Sites, *Mol Biol Evol* 33, 2976-2989.
- [126] Ranwez, V., Harispe, S., Delsuc, F., and Douzery, E. J. (2011) MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons, *PLoS One* 6, e22594.
- [127] Gharib, W. H., and Robinson-Rechavi, M. (2013) The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC, *Mol Biol Evol* 30, 1675-1686.
- [128] Lyon, A. M., Begley, J. A., Manett, T. D., and Tesmer, J. J. (2014) Molecular mechanisms of phospholipase C  $\beta$ 3 autoinhibition, *Structure* 22, 1844-1854.
- [129] Taylor, V. G., Bommarito, P. A., and Tesmer, J. J. (2016) Structure of the Regulator of G Protein Signaling 8 (RGS8)-G $\alpha$ q Complex: MOLECULAR BASIS FOR G $\alpha$  SELECTIVITY, *The Journal of biological chemistry* 291, 5138-5145.
- [130] Sunahara, R. K., Tesmer, J. J., Gilman, A. G., and Sprang, S. R. (1997) Crystal structure of the adenylyl cyclase activator G $\alpha$ s, *Science* 278, 1943-1947.
- [131] Wall, M. A., Coleman, D. E., Lee, E., Iñiguez-Lluhi, J. A., Posner, B. A., Gilman, A. G., and Sprang, S. R. (1995) The structure of the G protein heterotrimer Gi  $\alpha$ 1  $\beta$ 1  $\gamma$ 2, *Cell* 83, 1047-1058.
- [132] Tesmer, J. J., Berman, D. M., Gilman, A. G., and Sprang, S. R. (1997) Structure of RGS4 bound to AlF $_4^-$ -activated G(i  $\alpha$ 1): stabilization of the transition state for GTP hydrolysis, *Cell* 89, 251-261.
- [133] Alexander, N. S., Preininger, A. M., Kaya, A. I., Stein, R. A., Hamm, H. E., and Meiler, J. (2014) Energetic analysis of the rhodopsin-G-protein complex links the  $\alpha$ 5 helix to GDP release, *Nature structural & molecular biology* 21, 56-63.
- [134] Song, Y., DiMaio, F., Wang, R. Y., Kim, D., Miles, C., Brunette, T., Thompson, J., and Baker, D. (2013) High-resolution comparative modeling with RosettaCM, *Structure* 21, 1735-1742.
- [135] The PyMOL Molecular Graphics System, Version 1.8 Schroedinger, LLC.
- [136] Holland, P. W., Garcia-Fernàndez, J., Williams, N. A., and Sidow, A. (1994) Gene duplications and the origins of vertebrate development, *Dev Suppl*, 125-133.
- [137] Spring, J. (1997) Vertebrate evolution by interspecific hybridisation--are we polyploid?, *FEBS Lett* 400, 2-8.
- [138] Meyer, A., and Van de Peer, Y. (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD), *Bioessays* 27, 937-945.
- [139] Hamer, D. H., and Leder, P. (1979) Splicing and the formation of stable RNA, *Cell* 18, 1299-1302.

- [140] Cullen, B. R., Kopchick, J. J., and Stacey, D. W. (1982) Effect of intron size on splicing efficiency in retroviral transcripts, *Nucleic Acids Res* 10, 6177-6190.
- [141] Chung, S., and Perry, R. P. (1989) Importance of introns for expression of mouse ribosomal protein gene rpL32, *Molecular and cellular biology* 9, 2075-2082.
- [142] Rose, A. B., and Beliakoff, J. A. (2000) Intron-mediated enhancement of gene expression independent of unique intron sequences and splicing, *Plant Physiol* 122, 535-542.
- [143] Palmiter, R. D., Sandgren, E. P., Avarbock, M. R., Allen, D. D., and Brinster, R. L. (1991) Heterologous introns can enhance expression of transgenes in mice, *Proceedings of the National Academy of Sciences of the United States of America* 88, 478-482.
- [144] Nott, A., Meislin, S. H., and Moore, M. J. (2003) A quantitative analysis of intron effects on mammalian gene expression, *RNA* 9, 607-617.
- [145] Innan, H., and Kondrashov, F. (2010) The evolution of gene duplications: classifying and distinguishing between models, *Nat Rev Genet* 11, 97-108.
- [146] Lynch, M., and Conery, J. S. (2000) The evolutionary fate and consequences of duplicate genes, *Science* 290, 1151-1155.
- [147] Conant, G. C., and Wolfe, K. H. (2008) Turning a hobby into a job: how duplicated genes find new functions, *Nat Rev Genet* 9, 938-950.
- [148] Van de Peer, Y., Maere, S., and Meyer, A. (2009) The evolutionary significance of ancient genome duplications, *Nat Rev Genet* 10, 725-732.
- [149] Davis, J. C., and Petrov, D. A. (2004) Preferential duplication of conserved proteins in eukaryotic genomes, *PLoS Biol* 2, E55.
- [150] Julien Roux, J. L., Marc Robinson-Rechavi (2016) Selective constraints on coding sequences of nervous system genes are a major determinant of duplicate gene retention in vertebrates, *bioRxiv*, PrePrint.
- [151] Brunet, F. G., Roest Crolius, H., Paris, M., Aury, J. M., Gibert, P., Jaillon, O., Laudet, V., and Robinson-Rechavi, M. (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes, *Mol Biol Evol* 23, 1808-1816.
- [152] Smith, J. J., and Keinath, M. C. (2015) The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications, *Genome Res* 25, 1081-1090.
- [153] Hamm, H. E., Deretic, D., Arendt, A., Hargrave, P. A., Koenig, B., and Hofmann, K. P. (1988) Site of G protein binding to rhodopsin mapped with synthetic peptides from the alpha subunit, *Science* 241, 832-835.
- [154] Rasmussen, S. G., DeVree, B. T., Zou, Y., Kruse, A. C., Chung, K. Y., Kobilka, T. S., Thian, F. S., Chae, P. S., Pardon, E., Calinski, D., Mathiesen, J. M., Shah, S. T., Lyons, J. A., Caffrey, M., Gellman, S. H., Steyaert, J., Skiniotis, G., Weis, W. I., Sunahara, R. K., and Kobilka, B. K. (2011) Crystal structure of the beta2 adrenergic receptor-Gs protein complex, *Nature* 477, 549-555.



- [155] Marin, E. P., Krishna, A. G., and Sakmar, T. P. (2002) Disruption of the alpha5 helix of transducin impairs rhodopsin-catalyzed nucleotide exchange, *Biochemistry* 41, 6988-6994.
- [156] Oldham, W. M., Van Eps, N., Preininger, A. M., Hubbell, W. L., and Hamm, H. E. (2006) Mechanism of the receptor-catalyzed activation of heterotrimeric G proteins, *Nature structural & molecular biology* 13, 772-777.
- [157] Oldham, W. M., Van Eps, N., Preininger, A. M., Hubbell, W. L., and Hamm, H. E. (2007) Mapping allosteric connections from the receptor to the nucleotide-binding pocket of heterotrimeric G proteins, *Proceedings of the National Academy of Sciences of the United States of America* 104, 7927-7932.
- [158] Indrischek, H., Prohaska, S. J., Gurevich, V. V., Gurevich, E., and Stadler, P. F. (2016 In Review) Uncovering missing pieces: Duplication and deletion history of arrestins in Deuterostomes, *BMC Evolutionary Biology*.
- [159] Kehlenbach, R. H., Matthey, J., and Huttner, W. B. (1994) XL alpha s is a new type of G protein, *Nature* 372, 804-809.
- [160] Abramowitz, J., Grenet, D., Birnbaumer, M., Torres, H. N., and Birnbaumer, L. (2004) XLalphas, the extra-long form of the alpha-subunit of the Gs G protein, is significantly longer than suspected, and so is its companion Alex, *Proceedings of the National Academy of Sciences of the United States of America* 101, 8366-8371.
- [161] Hayward, B. E., and Bonthron, D. T. (2000) An imprinted antisense transcript at the human GNAS1 locus, *Hum Mol Genet* 9, 835-841.
- [162] Hayward, B. E., Moran, V., Strain, L., and Bonthron, D. T. (1998) Bidirectional imprinting of a single gene: GNAS1 encodes maternally, paternally, and biallelically derived proteins, *Proceedings of the National Academy of Sciences of the United States of America* 95, 15475-15480.
- [163] Wroe, S. F., Kelsey, G., Skinner, J. A., Bodle, D., Ball, S. T., Beechey, C. V., Peters, J., and Williamson, C. M. (2000) An imprinted transcript, antisense to Nesp, adds complexity to the cluster of imprinted genes at the mouse Gnas locus, *Proceedings of the National Academy of Sciences of the United States of America* 97, 3342-3346.
- [164] Peters, J., Wroe, S. F., Wells, C. A., Miller, H. J., Bodle, D., Beechey, C. V., Williamson, C. M., and Kelsey, G. (1999) A cluster of oppositely imprinted transcripts at the Gnas locus in the distal imprinting region of mouse chromosome 2, *Proceedings of the National Academy of Sciences of the United States of America* 96, 3830-3835.
- [165] Klemke, M., Pasolli, H. A., Kehlenbach, R. H., Offermanns, S., Schultz, G., and Huttner, W. B. (2000) Characterization of the extra-large G protein alpha-subunit XLalphas. II. Signal transduction properties, *The Journal of biological chemistry* 275, 33633-33640.
- [166] Pasolli, H. A., Klemke, M., Kehlenbach, R. H., Wang, Y., and Huttner, W. B. (2000) Characterization of the extra-large G protein alpha-subunit XLalphas. I. Tissue distribution and subcellular localization, *The Journal of biological chemistry* 275, 33622-33632.

- [167] Bray, P., Carter, A., Simons, C., Guo, V., Puckett, C., Kamholz, J., Spiegel, A., and Nirenberg, M. (1986) Human cDNA clones for four species of G alpha s signal transduction protein, *Proceedings of the National Academy of Sciences of the United States of America* 83, 8893-8897.
- [168] Kozasa, T., Itoh, H., Tsukamoto, T., and Kaziro, Y. (1988) Isolation and characterization of the human Gs alpha gene, *Proceedings of the National Academy of Sciences of the United States of America* 85, 2081-2085.
- [169] Kaya, A. I., Lokits, A. D., Gilbert, J. A., Iverson, T. M., Meiler, J., and Hamm, H. E. (2014) A conserved phenylalanine as relay between the  $\alpha 5$  helix and the GDP binding region of heterotrimeric Gi protein  $\alpha$  subunit, *The Journal of biological chemistry*.
- [170] Pyne, N. J., Freissmuth, M., and Pyne, S. (1992) Phosphorylation of the recombinant spliced variants of the alpha-sub-unit of the stimulatory guanine-nucleotide binding regulatory protein (Gs) by the catalytic sub-unit of protein kinase A, *Biochem Biophys Res Commun* 186, 1081-1086.
- [171] Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E. (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations, *Nucleic Acids Res* 43, D512-520.
- [172] Onrust, R., Herzmark, P., Chi, P., Garcia, P. D., Lichtarge, O., Kingsley, C., and Bourne, H. R. (1997) Receptor and betagamma binding sites in the alpha subunit of the retinal G protein transducin, *Science* 275, 381-384.
- [173] Bourne, H. R. (1997) How receptors talk to trimeric G proteins, *Curr Opin Cell Biol* 9, 134-142.
- [174] Aebi, M., Hornig, H., Padgett, R. A., Reiser, J., and Weissmann, C. (1986) Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA, *Cell* 47, 555-565.
- [175] Parada, G. E., Munita, R., Cerda, C. A., and Gysling, K. (2014) A comprehensive survey of non-canonical splice sites in the human transcriptome, *Nucleic Acids Res* 42, 10564-10578.
- [176] Thanaraj, T. A., and Clark, F. (2001) Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions, *Nucleic Acids Res* 29, 2581-2593.
- [177] Andrews, S. J., and Rothnagel, J. A. (2014) Emerging evidence for functional peptides encoded by short open reading frames, *Nat Rev Genet* 15, 193-204.
- [178] Kaessmann, H., Vinckenbosch, N., and Long, M. (2009) RNA-based gene duplication: mechanistic and evolutionary insights, *Nat Rev Genet* 10, 19-31.
- [179] Zhang, Z., Carriero, N., and Gerstein, M. (2004) Comparative analysis of processed pseudogenes in the mouse and human genomes, *Trends Genet* 20, 62-67.
- [180] Marques, A. C., Dupanloup, I., Vinckenbosch, N., Reymond, A., and Kaessmann, H. (2005) Emergence of young human genes after a burst of retroposition in primates, *PLoS Biol* 3, e357.

- [181] Jaillon, O., Aury, J. M., Brunet, F., Petit, J. L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, V., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Biémont, C., Skalli, Z., Cattolico, L., Poulain, J., De Berardinis, V., Cruaud, C., Duprat, S., Brottier, P., Coutanceau, J. P., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K. J., McEwan, P., Bosak, S., Kellis, M., Volff, J. N., Guigó, R., Zody, M. C., Mesirov, J., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., Laudet, V., Schachter, V., Quétier, F., Saurin, W., Scarpelli, C., Wincker, P., Lander, E. S., Weissenbach, J., and Roest Crolius, H. (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype, *Nature* 431, 946-957.
- [182] Woods, I. G., Wilson, C., Friedlander, B., Chang, P., Reyes, D. K., Nix, R., Kelly, P. D., Chu, F., Postlethwait, J. H., and Talbot, W. S. (2005) The zebrafish gene map defines ancestral vertebrate chromosomes, *Genome Res* 15, 1307-1314.
- [183] Oldham, W. M., and Hamm, H. E. (2007) How do Receptors Activate G Proteins?, 74, 67-93.
- [184] Yenerall, P., and Zhou, L. (2012) Identifying the mechanisms of intron gain: progress and trends, *Biol Direct* 7, 29.
- [185] Roy, S. W., and Irimia, M. (2009) Mystery of intron gain: new data and new models, *Trends Genet* 25, 67-73.
- [186] Amsen, D., Antov, A., Jankovic, D., Sher, A., Radtke, F., Souabni, A., Busslinger, M., McCright, B., Gridley, T., and Flavell, R. A. (2007) Direct regulation of Gata3 expression determines the T helper differentiation potential of Notch, *Immunity* 27, 89-99.
- [187] Jin, D., Hidaka, K., Shirai, M., and Morisaki, T. (2010) RNA-binding motif protein 24 regulates myogenin expression and promotes myogenic differentiation, *Genes Cells* 15, 1158-1167.
- [188] Pollard, A. J., Krainer, A. R., Robson, S. C., and Europe-Finner, G. N. (2002) Alternative splicing of the adenylyl cyclase stimulatory G-protein alpha(s) is regulated by SF2/ASF and heterogeneous nuclear ribonucleoprotein A1 (hnRNPA1) and involves the use of an unusual TG 3'-splice Site, *The Journal of biological chemistry* 277, 15241-15251.
- [189] Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B. J., and Frey, B. J. (2010) Deciphering the splicing code, *Nature* 465, 53-59.
- [190] Wainberg, M., Alipanahi, B., and Frey, B. (2016) Does conservation account for splicing patterns?, *BMC Genomics* 17, 787.
- [191] Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004) WebLogo: a sequence logo generator, *Genome Res* 14, 1188-1190.
- [192] Kandel, E. R., and Squire, L. R. (2000) Neuroscience: breaking down scientific barriers to the study of brain and mind, *Science* 290, 1113-1120.

- [193] Sprang, S. R. (1997) G protein mechanisms: insights from structural analysis, *Annu Rev Biochem* 66, 639-678.
- [194] Bjarnadóttir, T. K., Gloriam, D. E., Hellstrand, S. H., Kristiansson, H., Fredriksson, R., and Schiöth, H. B. (2006) Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse, *Genomics* 88, 263-273.
- [195] Shigemoto, R., Kinoshita, A., Wada, E., Nomura, S., Ohishi, H., Takada, M., Flor, P. J., Neki, A., Abe, T., Nakanishi, S., and Mizuno, N. (1997) Differential presynaptic localization of metabotropic glutamate receptor subtypes in the rat hippocampus, *J Neurosci* 17, 7503-7522.
- [196] Corti, C., Aldegheri, L., Somogyi, P., and Ferraguti, F. (2002) Distribution and synaptic localisation of the metabotropic glutamate receptor 4 (mGluR4) in the rodent CNS, *Neuroscience* 110, 403-420.
- [197] Schoch, S., and Gundelfinger, E. D. (2006) Molecular organization of the presynaptic active zone, *Cell Tissue Res* 326, 379-391.
- [198] Zhai, R. G., and Bellen, H. J. (2004) The architecture of the active zone in the presynaptic nerve terminal, *Physiology (Bethesda)* 19, 262-270.
- [199] Hopkins, A. L., and Groom, C. R. (2002) The druggable genome, *Nat Rev Drug Discov* 1, 727-730.
- [200] Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., Le Trong, I., Teller, D. C., Okada, T., Stenkamp, R. E., Yamamoto, M., and Miyano, M. (2000) Crystal structure of rhodopsin: A G protein-coupled receptor, *Science* 289, 739-745.
- [201] Cherezov, V., Rosenbaum, D. M., Hanson, M. A., Rasmussen, S. G., Thian, F. S., Kobilka, T. S., Choi, H. J., Kuhn, P., Weis, W. I., Kobilka, B. K., and Stevens, R. C. (2007) High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor, *Science* 318, 1258-1265.
- [202] Rasmussen, S. G., Choi, H. J., Rosenbaum, D. M., Kobilka, T. S., Thian, F. S., Edwards, P. C., Burghammer, M., Ratnala, V. R., Sanishvili, R., Fischetti, R. F., Schertler, G. F., Weis, W. I., and Kobilka, B. K. (2007) Crystal structure of the human beta2 adrenergic G-protein-coupled receptor, *Nature* 450, 383-387.
- [203] Ford, C. E., Skiba, N. P., Bae, H., Daaka, Y., Reuveny, E., Shekter, L. R., Rosal, R., Weng, G., Yang, C. S., Iyengar, R., Miller, R. J., Jan, L. Y., Lefkowitz, R. J., and Hamm, H. E. (1998) Molecular basis for interactions of G protein betagamma subunits with effectors, *Science* 280, 1271-1274.
- [204] Fletcher, J. E., Lindorfer, M. A., DeFilippo, J. M., Yasuda, H., Guilford, M., and Garrison, J. C. (1998) The G protein beta5 subunit interacts selectively with the Gq alpha subunit, *The Journal of biological chemistry* 273, 636-644.
- [205] Schmidt, C. J., Thomas, T. C., Levine, M. A., and Neer, E. J. (1992) Specificity of G protein beta and gamma subunit interactions, *The Journal of biological chemistry* 267, 13807-13810.

- [206] Pardon, E., Laeremans, T., Triest, S., Rasmussen, S. G., Wohlkonig, A., Ruf, A., Muyldermans, S., Hol, W. G., Kobilka, B. K., and Steyaert, J. (2014) A general protocol for the generation of Nanobodies for structural biology, *Nature protocols* 9, 674-693.
- [207] Westfield, G. H., Rasmussen, S. G., Su, M., Dutta, S., DeVree, B. T., Chung, K. Y., Calinski, D., Velez-Ruiz, G., Oleskie, A. N., Pardon, E., Chae, P. S., Liu, T., Li, S., Woods, V. L., Jr., Steyaert, J., Kobilka, B. K., Sunahara, R. K., and Skiniotis, G. (2011) Structural flexibility of the G alpha s alpha-helical domain in the beta2-adrenoceptor Gs complex, *Proceedings of the National Academy of Sciences of the United States of America* 108, 16086-16091.
- [208] Lambright, D. G., Sondek, J., Bohm, A., Skiba, N. P., Hamm, H. E., and Sigler, P. B. (1996) The 2.0 Å crystal structure of a heterotrimeric G protein, *Nature* 379, 311-319.
- [209] Clapham, D. E., and Neer, E. J. (1997) G protein beta gamma subunits, *Annu Rev Pharmacol Toxicol* 37, 167-203.
- [210] Noel, J. P., Hamm, H. E., and Sigler, P. B. (1993) The 2.2 Å crystal structure of transducin-alpha complexed with GTP gamma S, *Nature* 366, 654-663.
- [211] Sondek, J., Lambright, D. G., Noel, J. P., Hamm, H. E., and Sigler, P. B. (1994) GTPase mechanism of Gproteins from the 1.7-Å crystal structure of transducin alpha-GDP-AIF-4, *Nature* 372, 276-279.
- [212] Coleman, D. E., Berghuis, A. M., Lee, E., Linder, M. E., Gilman, A. G., and Sprang, S. R. (1994) Structures of active conformations of Gi alpha 1 and the mechanism of GTP hydrolysis, *Science* 265, 1405-1412.
- [213] Lyon, A. M., Dutta, S., Boguth, C. A., Skiniotis, G., and Tesmer, J. J. (2013) Full-length Galpha(q)-phospholipase C-beta3 structure reveals interfaces of the C-terminal coiled-coil domain, *Nature structural & molecular biology* 20, 355-362.
- [214] Iiri, T., Farfel, Z., and Bourne, H. R. (1998) G-protein diseases furnish a model for the turn-on switch, *Nature* 394, 35-38.
- [215] Martin, E. L., Rens-Domiano, S., Schatz, P. J., and Hamm, H. E. (1996) Potent peptide analogues of a G protein receptor-binding region obtained with a combinatorial library, *The Journal of biological chemistry* 271, 361-366.
- [216] Onrust, R. (1997) Receptor and beta gamma Binding Sites in the alpha Subunit of the Retinal G Protein Transducin, *Science* 275, 381-384.
- [217] Muradov, K. G., and Artemyev, N. O. (2000) Coupling between the N- and C-terminal domains influences transducin-alpha intrinsic GDP/GTP exchange, *Biochemistry* 39, 3937-3942.
- [218] Azpiazu, I., and Gautam, N. (2001) G protein gamma subunit interaction with a receptor regulates receptor-stimulated nucleotide exchange, *The Journal of biological chemistry* 276, 41742-41747.

- [219] Mazzoni, M. R., Malinski, J. A., and Hamm, H. E. (1991) Structural analysis of rod GTP-binding protein, Gt. Limited proteolytic digestion pattern of Gt with four proteases defines monoclonal antibody epitope, *The Journal of biological chemistry* 266, 14072-14081.
- [220] Denker, B. M., Schmidt, C. J., and Neer, E. J. (1992) Promotion of the GTP-liganded state of the Go alpha protein by deletion of the C terminus, *The Journal of biological chemistry* 267, 9998-10002.
- [221] Denker, B. M., Boutin, P. M., and Neer, E. J. (1995) Interactions between the amino- and carboxyl-terminal regions of G alpha subunits: analysis of mutated G alpha o/G alpha i2 chimeras, *Biochemistry* 34, 5544-5553.
- [222] Marin, E. P., Krishna, A. G., and Sakmar, T. P. (2001) Rapid activation of transducin by mutations distant from the nucleotide-binding site: evidence for a mechanistic model of receptor-catalyzed nucleotide exchange by G proteins, *The Journal of biological chemistry* 276, 27400-27405.
- [223] Kawashima, T., Berthet-Colominas, C., Wulff, M., Cusack, S., and Leberman, R. (1996) The structure of the Escherichia coli EF-Tu.EF-Ts complex at 2.5 Å resolution, *Nature* 379, 511-518.
- [224] Wang, Y., Jiang, Y., Meyering-Voss, M., Sprinzl, M., and Sigler, P. B. (1997) Crystal structure of the EF-Tu.EF-Ts complex from Thermus thermophilus, *Nat Struct Biol* 4, 650-656.
- [225] Perez, D. M., and Karnik, S. S. (2005) Multiple signaling states of G-protein-coupled receptors, *Pharmacological reviews* 57, 147-161.
- [226] Rondard, P., Iiri, T., Srinivasan, S., Meng, E., Fujita, T., and Bourne, H. R. (2001) Mutant G protein alpha subunit activated by Gbeta gamma: a model for receptor activation?, *Proceedings of the National Academy of Sciences of the United States of America* 98, 6150-6155.
- [227] Blahos, J., Fischer, T., Brabet, I., Stauffer, D., Rovelli, G., Bockaert, J., and Pin, J. P. (2001) A novel site on the Galpha -protein that recognizes heptahelical receptors, *The Journal of biological chemistry* 276, 3262-3269.
- [228] Sprang, S. R. (1997) G proteins, effectors and GAPs: structure and mechanism, *Curr Opin Struct Biol* 7, 849-856.
- [229] Cherfils, J., and Chabre, M. (2003) Activation of G-protein Galpha subunits by receptors through Galpha-Gbeta and Galpha-Ggamma interactions, *Trends Biochem Sci* 28, 13-17.
- [230] Hou, Y., Azpiazu, I., Smrcka, A., and Gautam, N. (2000) Selective role of G protein gamma subunits in receptor interaction, *The Journal of biological chemistry* 275, 38961-38964.
- [231] Markby, D. W., Onrust, R., and Bourne, H. R. (1993) Separate GTP binding and GTPase activating domains of a G alpha subunit, *Science* 262, 1895-1901.

- [232] Kleuss, C., Scherübl, H., Hescheler, J., Schultz, G., and Wittig, B. (1993) Selectivity in signal transduction determined by gamma subunits of heterotrimeric G proteins, *Science* 259, 832-834.
- [233] Kisselev, O., and Gautam, N. (1993) Specific interaction with rhodopsin is dependent on the gamma subunit type in a G protein, *The Journal of biological chemistry* 268, 24519-24522.
- [234] Rahmatullah, M., Ginnan, R., and Robishaw, J. D. (1995) Specificity of G protein alpha-gamma subunit interactions. N-terminal 15 amino acids of gamma subunit specifies interaction with alpha subunit, *The Journal of biological chemistry* 270, 2946-2951.
- [235] Rahmatullah, M., and Robishaw, J. D. (1994) Direct interaction of the alpha and gamma subunits of the G proteins. Purification and analysis by limited proteolysis, *The Journal of biological chemistry* 269, 3574-3580.
- [236] Bornancin, F., Pfister, C., and Chabre, M. (1989) The transitory complex between photoexcited rhodopsin and transducin. Reciprocal interaction between the retinal site in rhodopsin and the nucleotide site in transducin, *Eur J Biochem* 184, 687-698.
- [237] 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.
- [238] Goldberg, J. (1998) Structural basis for activation of ARF GTPase: mechanisms of guanine nucleotide exchange and GTP-myristoyl switching, *Cell* 95, 237-248.
- [239] Worthylake, D. K., Rossman, K. L., and Sondek, J. (2000) Crystal structure of Rac1 in complex with the guanine nucleotide exchange region of Tiam1, *Nature* 408, 682-688.
- [240] Renault, L., Kuhlmann, J., Henkel, A., and Wittinghofer, A. (2001) Structural basis for guanine nucleotide exchange on Ran by the regulator of chromosome condensation (RCC1), *Cell* 105, 245-255.
- [241] Cherfils, J., and Chardin, P. (1999) GEFs: structural basis for their activation of small GTP-binding proteins, *Trends Biochem Sci* 24, 306-311.
- [242] Pasqualato, S., Renault, L., and Cherfils, J. (2002) Arf, Arl, Arp and Sar proteins: a family of GTP-binding proteins with a structural device for 'front-back' communication, *EMBO Rep* 3, 1035-1041.
- [243] Medkova, M., Preininger, A. M., Yu, N. J., Hubbell, W. L., and Hamm, H. E. (2002) Conformational changes in the amino-terminal helix of the G protein alpha(i1) following dissociation from Gbetagamma subunit and activation, *Biochemistry* 41, 9962-9972.
- [244] Van Eps, N., Oldham, W. M., Hamm, H. E., and Hubbell, W. L. (2006) Structural and dynamical changes in an alpha-subunit of a heterotrimeric G protein along the activation pathway, *Proceedings of the National Academy of Sciences of the United States of America* 103, 16194-16199.
- [245] Altenbach, C., Kusnetzow, A. K., Ernst, O. P., Hofmann, K. P., and Hubbell, W. L. (2008) High-resolution distance mapping in rhodopsin reveals the pattern of helix movement due

- to activation, *Proceedings of the National Academy of Sciences of the United States of America* 105, 7439-7444.
- [246] Scheerer, P., Park, J. H., Hildebrand, P. W., Kim, Y. J., Krauss, N., Choe, H. W., Hofmann, K. P., and Ernst, O. P. (2008) Crystal structure of opsin in its G-protein-interacting conformation, *Nature* 455, 497-502.
- [247] Johnston, C. A., and Siderovski, D. P. (2007) Structural basis for nucleotide exchange on G alpha i subunits and receptor coupling specificity, *Proceedings of the National Academy of Sciences of the United States of America* 104, 2001-2006.
- [248] Dratz, E. A., Furstenau, J. E., Lambert, C. G., Thireault, D. L., Rarick, H., Schepers, T., Pakhlevaniants, S., and Hamm, H. E. (1993) NMR structure of a receptor-bound G-protein peptide, *Nature* 363, 276-281.
- [249] Kisselev, O. G., Kao, J., Ponder, J. W., Fann, Y. C., Gautam, N., and Marshall, G. R. (1998) Light-activated rhodopsin induces structural binding motif in G protein alpha subunit, *Proceedings of the National Academy of Sciences of the United States of America* 95, 4270-4275.
- [250] Yang, C. S., Skiba, N. P., Mazzoni, M. R., and Hamm, H. E. (1999) Conformational changes at the carboxyl terminus of Galpha occur during G protein activation, *The Journal of biological chemistry* 274, 2379-2385.
- [251] Natochin, M., Moussaif, M., and Artemyev, N. O. (2001) Probing the mechanism of rhodopsin-catalyzed transducin activation, *J Neurochem* 77, 202-210.
- [252] Ceruso, M. A., Periole, X., and Weinstein, H. (2004) Molecular dynamics simulations of transducin: interdomain and front to back communication in activation and nucleotide exchange, *Journal of molecular biology* 338, 469-481.
- [253] Iiri, T., Herzmark, P., Nakamoto, J. M., van Dop, C., and Bourne, H. R. (1994) Rapid GDP release from Gs alpha in patients with gain and loss of endocrine function, *Nature* 371, 164-168.
- [254] Posner, B. A., Mixon, M. B., Wall, M. A., Sprang, S. R., and Gilman, A. G. (1998) The A326S mutant of Galpha1 as an approximation of the receptor-bound state, *The Journal of biological chemistry* 273, 21752-21758.
- [255] Thomas, T. C., Schmidt, C. J., and Neer, E. J. (1993) G-protein alpha o subunit: mutation of conserved cysteines identifies a subunit contact surface and alters GDP affinity, *Proceedings of the National Academy of Sciences of the United States of America* 90, 10295-10299.
- [256] Van Eps, N., Preininger, A. M., Alexander, N., Kaya, A. I., Meier, S., Meiler, J., Hamm, H. E., and Hubbell, W. L. (2011) Interaction of a G protein with an activated receptor opens the interdomain interface in the alpha subunit, *Proceedings of the National Academy of Sciences of the United States of America* 108, 9420-9424.
- [257] Preininger, A. M., Funk, M. A., Oldham, W. M., Meier, S. M., Johnston, C. A., Adhikary, S., Kimple, A. J., Siderovski, D. P., Hamm, H. E., and Iverson, T. M. (2009) Helix dipole



movement and conformational variability contribute to allosteric GDP release in Galphai subunits, *Biochemistry* 48, 2630-2642.

- [258] Chung, K. Y., Rasmussen, S. G., Liu, T., Li, S., DeVree, B. T., Chae, P. S., Calinski, D., Kobilka, B. K., Woods, V. L., Jr., and Sunahara, R. K. (2011) Conformational changes in the G protein Gs induced by the beta2 adrenergic receptor, *Nature* 477, 611-615.
- [259] Niswender, C. M., Johnson, K. A., Weaver, C. D., Jones, C. K., Xiang, Z., Luo, Q., Rodriguez, A. L., Marlo, J. E., de Paulis, T., Thompson, A. D., Days, E. L., Nalywajko, T., Austin, C. A., Williams, M. B., Ayala, J. E., Williams, R., Lindsley, C. W., and Conn, P. J. (2008) Discovery, characterization, and antiparkinsonian effect of novel positive allosteric modulators of metabotropic glutamate receptor 4, *Mol Pharmacol* 74, 1345-1358.
- [260] Dhanya, R. P., Sidique, S., Sheffler, D. J., Nickols, H. H., Herath, A., Yang, L., Dahl, R., Ardecky, R., Semenova, S., Markou, A., Conn, P. J., and Cosford, N. D. (2011) Design and synthesis of an orally active metabotropic glutamate receptor subtype-2 (mGluR2) positive allosteric modulator (PAM) that decreases cocaine self-administration in rats, *J Med Chem* 54, 342-353.
- [261] Zhou, Y., Manka, J. T., Rodriguez, A. L., Weaver, C. D., Days, E. L., Vinson, P. N., Jadhav, S., Hermann, E. J., Jones, C. K., Conn, P. J., Lindsley, C. W., and Stauffer, S. R. (2010) Discovery of N-Aryl Piperazines as Selective mGlu(5) Potentiators with Efficacy in a Rodent Model Predictive of Anti-Psychotic Activity, *ACS Med Chem Lett* 1, 433-438.
- [262] Dölen, G., Osterweil, E., Rao, B. S., Smith, G. B., Auerbach, B. D., Chattarji, S., and Bear, M. F. (2007) Correction of fragile X syndrome in mice, *Neuron* 56, 955-962.
- [263] Herrold, A. A., Voigt, R. M., and Napier, T. C. (2013) mGluR5 is necessary for maintenance of methamphetamine-induced associative learning, *Eur Neuropsychopharmacol* 23, 691-696.
- [264] Patil, S. T., Zhang, L., Martenyi, F., Lowe, S. L., Jackson, K. A., Andreev, B. V., Avedisova, A. S., Bardenstein, L. M., Gurovich, I. Y., Morozova, M. A., Mosolov, S. N., Neznanov, N. G., Reznik, A. M., Smulevich, A. B., Tochilov, V. A., Johnson, B. G., Monn, J. A., and Schoepp, D. D. (2007) Activation of mGlu2/3 receptors as a new approach to treat schizophrenia: a randomized Phase 2 clinical trial, *Nat Med* 13, 1102-1107.
- [265] Swanson, C. J., Bures, M., Johnson, M. P., Linden, A. M., Monn, J. A., and Schoepp, D. D. (2005) Metabotropic glutamate receptors as novel targets for anxiety and stress disorders, *Nat Rev Drug Discov* 4, 131-144.
- [266] Oliveira, L., Paiva, A. C., and Vriend, G. (1999) A low resolution model for the interaction of G proteins with G protein-coupled receptors, *Protein Eng* 12, 1087-1095.
- [267] Bender, B. J., Cisneros, A., Duran, A. M., Finn, J. A., Fu, D., Lokits, A. D., Mueller, B. K., Sangha, A. K., Sauer, M. F., Sevy, A. M., Sliwoski, G., Sheehan, J. H., DiMaio, F., Meiler, J., and Moretti, R. (2016) Protocols for Molecular Modeling with Rosetta3 and RosettaScripts, *Biochemistry*.

- [268] Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H., and Meiler, J. (2010) Practically useful: what the Rosetta protein modeling suite can do for you, *Biochemistry* 49, 2987-2998.
- [269] Combs, S. A., Deluca, S. L., Deluca, S. H., Lemmon, G. H., Nannemann, D. P., Nguyen, E. D., Willis, J. R., Sheehan, J. H., and Meiler, J. (2013) Small-molecule ligand docking into comparative models with Rosetta, *Nature protocols* 8, 1277-1298.
- [270] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res* 25, 3389-3402.
- [271] Mandell, D. J., Coutsias, E. A., and Kortemme, T. (2009) Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling, *Nature methods* 6, 551-552.
- [272] Shapovalov, M. V., and Dunbrack, R. L., Jr. (2011) A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions, *Structure* 19, 844-858.
- [273] Cabrera-Vera, T. M., Vanhauwe, J., Thomas, T. O., Medkova, M., Preininger, A., Mazzoni, M. R., and Hamm, H. E. (2003) Insights into G protein structure, function, and regulation, *Endocr Rev* 24, 765-781.
- [274] Chivian, D., and Baker, D. (2006) Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection, *Nucleic Acids Res* 34, e112.
- [275] Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., Kellogg, E., DiMaio, F., Lange, O., Kinch, L., Sheffler, W., Kim, B. H., Das, R., Grishin, N. V., and Baker, D. (2009) Structure prediction for CASP8 with all-atom refinement using Rosetta, *Proteins* 77 Suppl 9, 89-99.
- [276] Kortemme, T., and Baker, D. (2002) A simple physical model for binding energy hot spots in protein-protein complexes, *Proceedings of the National Academy of Sciences of the United States of America* 99, 14116-14121.
- [277] Kellogg, E. H., Leaver-Fay, A., and Baker, D. (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability, *Proteins* 79, 830-838.
- [278] DeLano, W. L. (2002) Unraveling hot spots in binding interfaces: progress and challenges, *Curr Opin Struct Biol* 12, 14-20.
- [279] Lazaridis, T., and Karplus, M. (1999) Effective energy function for proteins in solution, *Proteins* 35, 133-152.
- [280] Lazaridis, T., and Karplus, M. (1999) Discrimination of the native from misfolded protein models with an energy function including implicit solvation, *Journal of molecular biology* 288, 477-487.

- [281] Mixon, M. B., Lee, E., Coleman, D. E., Berghuis, A. M., Gilman, A. G., and Sprang, S. R. (1995) Tertiary and quaternary structural changes in Gi alpha 1 induced by GTP hydrolysis, *Science* 270, 954-960.
- [282] Sharabi, O., Shirian, J., and Shifman, J. M. (2013) Predicting affinity- and specificity-enhancing mutations at protein-protein interfaces, *Biochem Soc Trans* 41, 1166-1169.
- [283] Kapoor, N., Menon, S. T., Chauhan, R., Sachdev, P., and Sakmar, T. P. (2009) Structural evidence for a sequential release mechanism for activation of heterotrimeric G proteins, *Journal of molecular biology* 393, 882-897.
- [284] Sprang, S. R. (1997) G Protein Mechanisms: Insights from Structural Analysis, *Annual Review of Biochemistry* 66, 639-678.
- [285] Higashijima, T., Ferguson, K. M., Smigel, M. D., and Gilman, A. G. (1987) The Effect of GTP and Mg<sup>2+</sup> on the GTPase Activity and the Fluorescent Properties of G<sub>o</sub>, *Journal of Biological Chemistry* 262, 757-761.
- [286] Oldham, W. M., and Hamm, H. E. (2007) How do receptors activate G proteins?, *Adv Protein Chem* 74, 67-93.
- [287] Onrust, R., Herzmark, P., Chi, P., Garcia, P. D., Lichtarge, O., Kingsley, C., and Bourne, H. R. (1997) Receptor and  $\beta\gamma$  Binding Sites in the  $\alpha$  Subunit of the Retinal G Protein Transducin, *Science* 275, 381-384.
- [288] Rosenbaum, D. M., Rasmussen, S. G., and Kobilka, B. K. (2009) The structure and function of G-protein-coupled receptors, *Nature* 459, 356-363.
- [289] Okada, T., Fujiyoshi, Y., Silow, M., Navarro, J., Landau, E. M., and Shichida, Y. (2002) Functional role of internal water molecules in rhodopsin revealed by X-ray crystallography, *Proceedings of the National Academy of Sciences of the United States of America* 99, 5982-5987.
- [290] Ernst, O. P., Bieri, C., Vogel, H., and Hofmann, K. P. (2000) Intrinsic biophysical monitors of transducin activation: fluorescence, UV-visible spectroscopy, light scattering, and evanescent field techniques, *Methods Enzymol* 315, 471-489.
- [291] Hamm, H. E., Deretic, D., Arendt, A., Hargrave, P. A., Koenig, B., and Hofmann, K. P. (1988) Site of G Protein Binding to Rhodopsin Mapped with Synthetic Peptides from the  $\alpha$  Subunit, *Science* 241, 832-835.
- [292] Marin, E. P., Krishna, A. G., and Sakmar, T. P. (2001) Rapid Activation of Transducin by Mutations Distant from the Nucleotide-Binding Site. Evidence for a Mechanistic Model of Receptor-Catalyzed Nucleotide Exchange by G Proteins, *Journal of Biological Chemistry* 276, 27400-27405.
- [293] Marin, E. P., Krishna, A. G., and Sakmar, T. P. (2002) Disruption of the  $\alpha 5$  Helix of Transducin Impairs Rhodopsin-Catalyzed Nucleotide Exchange, *Biochemistry* 41, 6988-6994.

- [294] Bourne, H. R. (1997) How receptors talk to trimeric G proteins, *Current Opinion in Cell Biology* 9, 134-142.
- [295] Rasmussen, S. G., Devree, B. T., Zou, Y., Kruse, A. C., Chung, K. Y., Kobilka, T. S., Thian, F. S., Chae, P. S., Pardon, E., Calinski, D., Mathiesen, J. M., Shah, S. T., Lyons, J. A., Caffrey, M., Gellman, S. H., Steyaert, J., Skiniotis, G., Weis, W. I., Sunahara, R. K., and Kobilka, B. K. (2011) Crystal structure of the beta(2) adrenergic receptor-Gs protein complex, *Nature* 477, 549-555.
- [296] Abdulaev, N. G., Ngo, T., Zhang, C., Dinh, A., Brabazon, D. M., Ridge, K. D., and Marino, J. P. (2005) Heterotrimeric G-protein alpha-subunit adopts a "preactivated" conformation when associated with betagamma-subunits, *The Journal of biological chemistry* 280, 38071-38080.
- [297] Cherfils, J., and Chabre, M. (2003) Activation of G-protein G $\alpha$  subunits by receptors through G $\alpha$ -G $\beta$  and G $\alpha$ -G $\gamma$  interactions, *Trends in Biochemical Sciences* 28, 13-17.
- [298] Rondard, P., Iiri, T., Srinivasan, S., Meng, E., Fujita, T., and Bourne, H. R. (2001) Mutant G protein  $\alpha$  subunit activated by G $\beta\gamma$ : A model for receptor activation?, *Proceedings of the National Academy of Sciences of the United States of America* 98, 6150-6155.
- [299] Louet, M., Martinez, J., and Floquet, N. (2012) GDP release preferentially occurs on the phosphate side in heterotrimeric G-proteins, *PLoS computational biology* 8, e1002595.
- [300] Marin, E. P., Krishna, A. G., Archambault, V., Simuni, E., Fu, W. Y., and Sakmar, T. P. (2001) The Function of Interdomain Interactions in Controlling Nucleotide Exchange Rates in Transducin, *Journal of Biological Chemistry* 276, 23873-23880.
- [301] Ceruso, M. A., Periole, X., and Weinstein, H. (2004) Molecular Dynamics Simulations of Transducin: Interdomain and Front to Back Communication in Activation and Nucleotide Exchange, *Journal of Molecular Biology* 338, 469-481.
- [302] Johnston, C. A., Willard, M. D., Kimple, A. J., Siderovski, D. P., and Willard, F. S. (2008) A sweet cycle for Arabidopsis G-proteins: Recent discoveries and controversies in plant G-protein signal transduction, *Plant signaling & behavior* 3, 1067-1076.
- [303] Mazzoni, M. R., Malinski, J. A., and Hamm, H. E. (1991) Structural Analysis of Rod GTP-binding Protein, G $_i$ . Limited Proteolytic Digestion Pattern of G $_i$  with Four Proteases Defines Monoclonal Antibody Epitope., *Journal of Biological Chemistry* 266, 14072-14081.
- [304] Thaker, T. M., Kaya, A. I., Preininger, A. M., Hamm, H. E., and Iverson, T. M. (2012) Allosteric mechanisms of G protein-Coupled Receptor signaling: a structural perspective, *Methods in molecular biology* 796, 133-174.
- [305] Medkova, M., Preininger, A. M., Yu, N. J., Hubbell, W. L., and Hamm, H. E. (2002) Conformational Changes in the Amino-Terminal Helix of the G Protein  $\alpha_{i1}$  Following Dissociation from G $\beta\gamma$  Subunit and Activation, *Biochemistry* 41, 9962-9972.

- [306] Bradford, M. M. (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding, *Anal Biochem* 72, 248-254.
- [307] Coleman, D. E., and Sprang, S. R. (1998) Crystal Structures of the G Protein  $G_{i\alpha 1}$  Complexed with GDP and  $Mg^{2+}$ : A Crystallographic Titration Experiment, *Biochemistry* 37, 14376-14385.
- [308] Schmidt, C. J., Thomas, T. C., Levine, M. A., and Neer, E. J. (1992) Specificity of G Protein  $\beta$  and  $\gamma$  Subunit Interactions, *Journal of Biological Chemistry* 267, 13807-13810.
- [309] Linder, M. E., Middleton, P., Hepler, J. R., Taussig, R., Gilman, A. G., and Mumby, S. M. (1993) Lipid modifications of G proteins: alpha subunits are palmitoylated, *Proc Natl Acad Sci U S A* 90, 3675-3679.
- [310] Otwinowski, Z., and Minor, W. (1997) Processing of X-ray Diffraction Data Collected in Oscillation Mode, In *Macromolecular Crystallography Part A* (Carter, C. W., Jr., and Sweet, R. M., Eds.), pp 307-326, Academic Press, New York.
- [311] Potterton, E., Briggs, P., Turkenburg, M., and Dodson, E. (2003) A graphical user interface to the CCP4 program suite, *Acta Crystallographica Section D* 59, 1131-1137.
- [312] Adams, P. D., Afonine, P. V., Bunkoczi, 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 Section D* 66, 213-221.
- [313] Mixon, M. B., Lee, E., Coleman, D. E., Berghuis, A. M., Gilman, A. G., and Sprang, S. R. (1995) Tertiary and Quaternary Structural Changes in  $G_{i\alpha 1}$  Induced by GTP Hydrolysis, *Science* 270, 954-960.
- [314] Coleman, D. E., Berghuis, A. M., Lee, E., Linder, M. E., Gilman, A. G., and Sprang, S. R. (1994) Structures of Active Conformations of  $G_{i\alpha 1}$  and the Mechanism of GTP Hydrolysis, *Science* 265, 1405-1412.
- [315] Emsley, P., and Cowtan, K. (2004) Coot: model-building tools for molecular graphics, *Acta Crystallographica Section D* 60, 2126-2132.
- [316] Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination, *Acta Crystallogr D Biol Crystallogr* 54, 905-921.
- [317] 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., 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 Crystallogr D Biol Crystallogr* 67, 235-242.

- [318] Krissinel, E., and Henrick, K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions, *Acta Crystallogr D Biol Crystallogr* 60, 2256-2268.
- [319] Quilliam, L. A., Zhong, S., Rabun, K. M., Carpenter, J. W., South, T. L., Der, C. J., and Campbell-Burk, S. (1995) Biological and structural characterization of a Ras transforming mutation at the phenylalanine-156 residue, which is conserved in all members of the Ras superfamily, *Proc Natl Acad Sci U S A* 92, 1272-1276.
- [320] Sovik, O., Schubbert, S., Houge, G., Steine, S. J., Norgard, G., Engelsen, B., Njolstad, P. R., Shannon, K., and Molven, A. (2007) De novo HRAS and KRAS mutations in two siblings with short stature and neuro-cardio-facio-cutaneous features, *Journal of medical genetics* 44, e84.
- [321] Schubbert, S., Bollag, G., Lyubynska, N., Nguyen, H., Kratz, C. P., Zenker, M., Niemeyer, C. M., Molven, A., and Shannon, K. (2007) Biochemical and functional characterization of germ line KRAS mutations, *Molecular and cellular biology* 27, 7765-7770.
- [322] Preininger, A., Funk, M., Meier, S., Oldham, W., Johnston, C., Adhikary, S., Kimple, A., Siderovski, D., Hamm, H., and Iverson, T. (2009) Helix dipole movement and conformational variability contribute to allosteric GDP release in Gi subunits, *Biochemistry* 48, 2630-2642.
- [323] Oldham, W. M., and Hamm, H. E. (2006) Structural basis of function in heterotrimeric G proteins, *Quarterly reviews of biophysics* 39, 117-166.
- [324] Chen, V. B., Arendall, W. B., 3rd, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography, *Acta Crystallogr D Biol Crystallogr* 66, 12-21.
- [325] Johnston, C. A., and Siderovski, D. P. (2007) Receptor-mediated activation of heterotrimeric G-proteins: current structural insights, *Mol Pharmacol* 72, 219-230.
- [326] Kimple, A. J., Bosch, D. E., Giguere, P. M., and Siderovski, D. P. (2011) Regulators of G-protein signaling and their Galpha substrates: promises and challenges in their use as drug discovery targets, *Pharmacol Rev* 63, 728-749.
- [327] Kaya, A. I., Lokits, A. D., Gilbert, J. A., Iverson, T. M., Meiler, J., and Hamm, H. E. (2014) A conserved phenylalanine as a relay between the alpha5 helix and the GDP binding region of heterotrimeric Gi protein alpha subunit, *J Biol Chem* 289, 24475-24487.
- [328] Valencia, A., Chardin, P., Wittinghofer, A., and Sander, C. (1991) The ras protein family: evolutionary tree and role of conserved amino acids, *Biochemistry* 30, 4637-4648.
- [329] Wilkie, T. M., Gilbert, D. J., Olsen, A. S., Chen, X. N., Amatruda, T. T., Korenberg, J. R., Trask, B. J., de Jong, P., Reed, R. R., Simon, M. I., and et al. (1992) Evolution of the mammalian G protein alpha subunit multigene family, *Nat Genet* 1, 85-91.

- [330] Ma, H., Yanofsky, M. F., and Meyerowitz, E. M. (1990) Molecular cloning and characterization of GPA1, a G protein alpha subunit gene from *Arabidopsis thaliana*, *Proc Natl Acad Sci U S A* 87, 3821-3825.
- [331] Flock, T., Ravarani, C. N., Sun, D., Venkatakrisnan, A. J., Kayikci, M., Tate, C. G., Veprintsev, D. B., and Babu, M. M. (2015) Universal allosteric mechanism for Galpha activation by GPCRs, *Nature* 524, 173-179.
- [332] Dror, R. O., Mildorf, T. J., Hilger, D., Manglik, A., Borhani, D. W., Arlow, D. H., Philippsen, A., Villanueva, N., Yang, Z., Lerch, M. T., Hubbell, W. L., Kobilka, B. K., Sunahara, R. K., and Shaw, D. E. (2015) SIGNAL TRANSDUCTION. Structural basis for nucleotide exchange in heterotrimeric G proteins, *Science* 348, 1361-1365.
- [333] Kaya, A. I., Thaker, T. M., Preininger, A. M., Iverson, T. M., and Hamm, H. E. (2011) Coupling efficiency of rhodopsin and transducin in bicelles, *Biochemistry* 50, 3193-3203.
- [334] Preininger, A. M., Kaya, A. I., Gilbert, J. A., 3rd, Busenlehner, L. S., Armstrong, R. N., and Hamm, H. E. (2012) Myristoylation exerts direct and allosteric effects on Galpha conformation and dynamics in solution, *Biochemistry* 51, 1911-1924.
- [335] Hamm, H. E., Kaya, A. I., Gilbert, J. A., 3rd, and Preininger, A. M. (2013) Linking receptor activation to changes in Sw I and II of Galpha proteins, *J Struct Biol* 184, 63-74.
- [336] Otwinowski, Z., and Minor, W. (1997) [20] Processing of X-ray diffraction data collected in oscillation mode, In *Methods in Enzymology* (Charles W. Carter, Jr., Ed.), pp 307-326, Academic Press.
- [337] Wall, M. A., Coleman, D. E., Lee, E., Iniguez-Lluhi, J. A., Posner, B. A., Gilman, A. G., and Sprang, S. R. (1995) The Structure of the G Protein Heterotrimer  $G_{i\alpha 1}\beta_1\gamma_2$ , *Cell* 83, 1047-1058.
- [338] McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. (2007) Phaser crystallographic software, *J Appl Crystallogr* 40, 658-674.
- [339] Murshudov, G. N., Vagin, A. A., and Dodson, E. J. (1997) Refinement of macromolecular structures by the maximum-likelihood method, *Acta Crystallogr D Biol Crystallogr* 53, 240-255.
- [340] Lokits, A. D., Leman, J. K., Kitko, K. E., Alexander, N. S., Hamm, H. E., and Meiler, J. (2015) A survey of conformational and energetic changes in G protein signaling, *AIMS Biophysics* 2, 630-648.
- [341] Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E., and Baker, D. (2014) Relaxation of backbone bond geometry improves protein energy landscape modeling, *Protein Sci* 23, 47-55.
- [342] Sun, D., Flock, T., Deupi, X., Maeda, S., Matkovic, M., Mendieta, S., Mayer, D., Dawson, R. J., Schertler, G. F., Babu, M. M., and Veprintsev, D. B. (2015) Probing Galpha1 protein activation at single-amino acid resolution, *Nat Struct Mol Biol* 22, 686-694.

- [343] Natochin, M., Moussaif, M., and Artemyev, N. O. (2001) Probing the mechanism of rhodopsin-catalyzed transducin activation, *Journal of Neurochemistry* 77, 202-210.
- [344] Oldham, W. M., and Hamm, H. E. (2008) Heterotrimeric G protein activation by G-protein-coupled receptors, *Nature reviews. Molecular cell biology* 9, 60-71.
- [345] Higashijima, T., Ferguson, K. M., Sternweis, P. C., Smigel, M. D., and Gilman, A. G. (1987) Effects of Mg<sup>2+</sup> and the  $\beta\gamma$ -Subunit Complex on the Interactions of Guanine Nucleotides with G proteins, *Journal of Biological Chemistry* 262, 762-766.
- [346] Thomas, T. C., Schmidt, C. J., and Neer, E. J. (1993) G-protein  $\alpha_o$  subunit: Mutation of conserved cysteines identifies a subunit contact surface and alters GDP affinity, *Proceedings of the National Academy of Sciences of the United States of America* 90, 10295-10298.
- [347] Iiri, T., Herzmark, P., Nakamoto, J. M., van Dop, C., and Bourne, H. R. (1994) Rapid GDP release from G<sub>s $\alpha$</sub>  in patients with gain and loss of endocrine function, *Nature* 371, 164-168.
- [348] Vu, T. K., Hung, D. T., Wheaton, V. I., and Coughlin, S. R. (1991) Molecular cloning of a functional thrombin receptor reveals a novel proteolytic mechanism of receptor activation, *Cell* 64, 1057-1068.
- [349] Cottrell, G. S., Amadesi, S., Grady, E. F., and Bunnett, N. W. (2004) Trypsin IV, a novel agonist of protease-activated receptors 2 and 4, *The Journal of biological chemistry* 279, 13532-13539.
- [350] Sambrano, G. R., Huang, W., Faruqi, T., Mahrus, S., Craik, C., and Coughlin, S. R. (2000) Cathepsin G activates protease-activated receptor-4 in human platelets, *The Journal of biological chemistry* 275, 6819-6823.
- [351] Zhao, P., Metcalf, M., and Bunnett, N. W. (2014) Biased signaling of protease-activated receptors, *Front Endocrinol (Lausanne)* 5, 67.
- [352] Canto, I., Soh, U. J., and Trejo, J. (2012) Allosteric modulation of protease-activated receptor signaling, *Mini Rev Med Chem* 12, 804-811.
- [353] Mao, Y., Zhang, M., Tuma, R. F., and Kunapuli, S. P. (2010) Deficiency of PAR4 attenuates cerebral ischemia/reperfusion injury in mice, *J Cereb Blood Flow Metab* 30, 1044-1052.
- [354] Sambrano, G. R., Weiss, E. J., Zheng, Y. W., Huang, W., and Coughlin, S. R. (2001) Role of thrombin signalling in platelets in haemostasis and thrombosis, *Nature* 413, 74-78.
- [355] Hollenberg, M. D., Saifeddine, M., Sandhu, S., Houle, S., and Vergnolle, N. (2004) Proteinase-activated receptor-4: evaluation of tethered ligand-derived peptides as probes for receptor function and as inflammatory agonists in vivo, *Br J Pharmacol* 143, 443-454.
- [356] Zhang, C., Srinivasan, Y., Arlow, D. H., Fung, J. J., Palmer, D., Zheng, Y., Green, H. F., Pandey, A., Dror, R. O., Shaw, D. E., Weis, W. I., Coughlin, S. R., and Kobilka, B. K. (2012) High-resolution crystal structure of human protease-activated receptor 1, *Nature* 492, 387-392.



- [357] Perry, S. R., Xu, W., Wirija, A., Lim, J., Yau, M. K., Stoermer, M. J., Lucke, A. J., and Fairlie, D. P. (2015) Three Homology Models of PAR2 Derived from Different Templates: Application to Antagonist Discovery, *J Chem Inf Model* 55, 1181-1191.
- [358] Yau, M. K., Suen, J. Y., Xu, W., Lim, J., Liu, L., Adams, M. N., He, Y., Hooper, J. D., Reid, R. C., and Fairlie, D. P. (2016) Potent Small Agonists of Protease Activated Receptor 2, *ACS Med Chem Lett* 7, 105-110.
- [359] Yau, M. K., Liu, L., Lim, J., Lohman, R. J., Cotterell, A. J., Suen, J. Y., Vesey, D. A., Reid, R. C., and Fairlie, D. P. (2016) Benzylamide antagonists of protease activated receptor 2 with anti-inflammatory activity, *Bioorg Med Chem Lett* 26, 986-991.
- [360] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank, *Nucleic Acids Res* 28, 235-242.
- [361] Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J., and Lesk, A. M. (2006) MUSTANG: a multiple structural alignment algorithm, *Proteins* 64, 559-574.
- [362] McGuffin, L. J., Bryson, K., and Jones, D. T. (2000) The PSIPRED protein structure prediction server, *Bioinformatics* 16, 404-405.
- [363] Viklund, H., and Elofsson, A. (2008) OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar, *Bioinformatics* 24, 1662-1668.
- [365] Bissantz, C., Logean, A., and Rognan, D. (2004) High-throughput modeling of human G-protein coupled receptors: amino acid sequence alignment, three-dimensional model building, and receptor library screening, *J Chem Inf Comput Sci* 44, 1162-1176.
- [366] Kim, D. E., Chivian, D., and Baker, D. (2004) Protein structure prediction and analysis using the Robetta server, *Nucleic Acids Res* 32, W526-531.
- [367] Li, S. C., and Ng, Y. K. (2010) Calibur: a tool for clustering large numbers of protein decoys, *BMC Bioinformatics* 11, 25.
- [368] Wang, C., Bradley, P., and Baker, D. (2007) Protein-protein docking with backbone flexibility, *Journal of molecular biology* 373, 503-519.
- [369] Canutescu, A. A., and Dunbrack, R. L. (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure, *Protein Sci* 12, 963-972.
- [370] Molecular Operating Environment (MOE), C. C. G. I., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2017.