Title

Specificity in Protein-Protein Interactions: High-Throughput Characterization of Rationally Designed and Naturally Evolved Coiled-Coil Networks

Author

Boldridge, William Clifford

Publication Date

2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Specificity in Protein-Protein Interactions: High-Throughput Characterization of Rationally Designed and Naturally Evolved Coiled-Coil Networks

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Biochemistry, Molecular and Structural Biology

by

William Clifford Boldridge

2021

ABSTRACT OF THE DISSERTATION

Specificity in Protein-Protein Interactions:

High-Throughput Characterization of Rationally Designed and Naturally

Evolved Coiled-Coil Networks

by

William Clifford Boldridge

Doctor of Philosophy in Biochemistry, Molecular and Structural Biology

University of California, Los Angeles, 2021

Professor Sriram Kosuri, Chair

As the major effectors of cellular processes, proteins are crucial to all biology. Although proteins are regulated in many fashions, protein-protein interactions are ubiquitous across different classes of proteins. In particular, proteins must interact specifically with certain partners to recapitulate the biology that constitutes life, despite cells containing hundreds of thousands of proteoforms, some fraction of which are highly similar to the intended target. Understanding how specificity in protein-protein interactions occurs has been challenging to investigate because prior techniques were limited to in throughput and ability to pinpoint sequences of interest. We create a high-throughput two-hybrid assay that marries gene synthesis with a next-generation sequencing readout, allowing us to investigate only those interactions of interest with a single experiment providing a quantitative characterization of tens of thousands of

interactions. We use this to first to investigate specificity in designed coiled-coils—small alpha-helical proteins which despite a simple hydrophobic interface exhibit high a high-degree of specificity. After validating our assay on a previously published set of coiled-coils, we iteratively find increasingly large sets of orthogonal proteins, proteins where each on-target interaction is specifically preferred to all off-target interactions. In total we screen more than 26,000 interactions in three experiments, and use our data and improve coiled-coil design algorithms while also finding the largest sets of orthogonal proteins to date. While specificity can be designed with large changes to the protein sequence, nature must come by specificity through the slow tinkering of evolution. To investigate the origins of specificity in nature we characterized a bZip family descended from an ancestral homodimer where the extant paralogs do not heterodimerize. We use ancestral reconstruction to trace protein-protein interactions in the coiled-coil domain across the PAR and E4BP4 family, back to the ancestor of humans and cnidarians. We find specificity does not appear once, but rather eight times across our tree, and while the process begins immediately, the final acquisition of specificity takes substantial time. Finally we find that once interactions are lost they never return, and that there is no direct selection for the acquisition of specificity between paralogs.

The dissertation of William Clifford Boldridge is approved.

James U. Bowie

David S. Eisenberg

Kirk E. Lohmueller

Sriram Kosuri, Committee Chair

University of California, Los Angeles

2021

Dedication

I dedicate this work to my family, David Boldridge, Allison Boldridge, Drew Boldridge, Kristen Reid, and Artemis Reid-Boldridge for their example, encouragement, comradery, support, and love of playtime, respectively.

TABLE OF CONTENTS

List of Figures……vii

Acknowledgements……ix

Vita……xii

Chapter 1

Introduction: Encoding specificity in protein-protein interactions……..1

References……..13

Chapter 2

A Multiplexed Bacterial Two-Hybrid for Rapid Characterization of Protein-Protein Interactions and Iterative Protein Design……..28

References……113

Chapter 3

Comprehensive Experimental Analysis of Functional Diversification Across Seven Hundred Million Years of bZip Evolution……..119

References……159

Chapter 4

Conclusions: The many trajectories to specificity in protein-protein interactions……..166

References……172

LIST OF FIGURES

Figure 2.1 Design and validation of the NGB2H assay……46

Figure 2.2 Large orthogonal sets of coiled-coils from the CCNG1 library……..48

Figure 2.3 Comparison, development and validation of iCipa model……..50

Figure 2.4 The largest orthogonal sets of the CCmax Library…….51

Figure S2.1 Optimization and tuning of the NGB2H system……...89

Figure S2.2 Design of OLS oligonucleotides for libraries used in this work……...91

Figure S2.3 Cloning from OLS oligonucleotides to barcoded X and Y constructs…….92

Figure S2.4 Cloning scheme of the NGB2H system after barcoding……..94

Figure S2.5 Different codon usages of the CC0 Library…….95

Figure S2.6 Indels in the CC0 Library have lower interaction scores than correct…….....96

Figure S2.7 CC0 Library Interaction Scores versus previously published Tms……..97

Figure S2.8 CC1 Library internal controls……..98

Figure S2.9 Isolation of orthogonal coiled-coils from the CC1 Library…….99

Figure S2.10 CCNG1 Library internal controls……101

Figure S2.11 Number of orthogonal interactions by sets with different backbones in the CCNG1 Library…….102

Figure S2.12 CCNG1 Library number of proteins per orthogonal set……...103

Figure S2.13 Effect of variation at the b, c, and f-positions……...104

Figure S2.14 Schematic of heptad shifting…….....105

Figure S2.15 CCmax Library internal controls……..107

Figure S2.16 Correlation of the CC0 Library proteins between different libraries……....109

Figure S2.17 Number of orthogonal proteins per set……..110

Figure S2.18 Ability to predict orthogonality compared across algorithms……...111

Figure S2.19 CCmax Library's agreement with previous models……..112

Figure 3.1 Characterization of 66,000 E4BP4 and PAR protein-protein interactions……134

Figure 3.2 Paralogs gain specificity many times……136

Figure 3.3 Gains of specificity do not necessitate vast rewiring……138

Figure 3.4 Specificity is not driven by direct selection and is permanent……..140

Figure S3.1 Species tree and number of proteins present……153

Figure S3.2 Calculation of classification scores……..154

Figure S3.3 Quality metrics of the NGB2H measurements…….155

Figure S3.4 *k*-means clustering metrics……156

Figure S3.5 Loss of heterodimerization after duplication for paralogs at weak stringency….157

ACKNOWLEDGEMENTS

First and foremost, I must thank my advisor, Sriram Kosuri without whom none of the following work would have been possible; his ideas and imagination were crucial for this research and I will take his enthusiasm for high-throughput systems with me on all my future research projects. I'm also indebted to his generosity and kindness in helping smooth out the rougher corners of graduate student life and helping me grow into a capable independent scientist. Furthermore, I thank my committee members, Jim Bowie, David Eisenberg, and Kirk Lohmueller for their guidance and helping me see the bigger picture.

I must also thank Ajasja Ljubetič and Roman Jerala who were integral to producing Chapter 2. Working closely with Ajasja over the course of several years was a pleasure, and his encouragement to rework some results vastly improved this work. Additionally, I must thank Georg Hochberg and Joe Thornton for their assistance in producing Chapter 3. It's nearly impossible to overstate how valuable Georg was to this process—when I was lost in the miasma that is phylogenetics, he, unasked, offered a collaboration that allowed the project to come to fruition. He has also been an unceasing source of friendship, support and encouragement all while starting his own lab. I am also indebted to Joe, who welcomed me into his own lab to find a community and scientific direction while the Kosuri Lab wound down. In the midst of Covid-19, I am deeply thankful for this.

I would like to thank all the members of the Kosuri Lab for helping me become a better, more focused scientist, while making sure lab was never a chore. First, I'd like to thank the 'New Lab' of Guillaume Urtecho, Calin Plesa and Angus Sidore, who provided levity, guidance, sneaky coffee, and many, many lunch runs. In particular, I must thank Guillaume who read through dozens of my rough drafts over the years and always identified indisputable

improvements. I'd like to thank the other graduate students of my year, Jess Davis and Kim Insigne for providing comradery, insightful feedback and many a board game night. I thank earlier graduate students, Nate Lubock and Eric Jones for showing us that it could be done. Also, thank you Nate for always having an open ear for coding problems and your assistance with all my basic scripts when I was barely able to hack it. I must thank Hwangbeom Kim who got the NGB2H system designed and functioning and who, with Rocky Cheung, made KBBQ a regular lab outing. I am also very thankful for Johnny Lee, who, as an undergraduate researcher, made me grow as a mentor, helped further the work presented here and always had an blinding aura of optimism. I thank our lab managers, Jeff Wang, Danny Cancilla and David Yao for creating the most functional lab environment I've known. I must also thank the current members of the Thornton Lab for the generosity and guidance. In particular, Brian Metzger and Yeonwoo Park for their blazing scientific acumen, Jaeda Patton, Carlos Cortez, Santiago Herrera and Arielle Weinstein for their positivity, bants and researcher solidarity. I also thank Arvind Pillai who's curiosity and ability to find wonder in everything is an inspiration and aspiration.

As science is a journey, each step is dependent on the last. Thus, I thank my undergraduate mentors Shohei and Akiko Koide for teaching me the basics of molecular biology, despite the fact that at the time I had nothing to offer other than enthusiasm. I also thank Yvonne Chen for teaching me organization that became necessary for my high-throughput work and perspective. There's always a few people who add a depth to living despite how little time you may have known them. Thus I thank Duilio Cascio for teaching me how to party as an adult, and Justin Benesch for teaching me the meaning of hospitality.

A wiser man than I once noted that graduate school is by its nature an isolating experience—so it is critical to have friends to get you through. Thus I thank my friends from

Biochemistry cohort, Alex Bradley, Jonelle White, Jess Davis, Mike Leonard, and Chris Kampmeyer for being there, particularly in the rough first year. I thank the my adopted group of MBI graduate students for joining me in holidays, happy hours and Game of Thrones nights—Aaron Van Loon and Sam Edwards, Meaghan Valliere, Dan Ferrer, Joe Bedree and David Gray, Catherine Schweppe, Lisa Golden, Matilde Miranda, and Brenda Molgora. I also have to thank two of the best roommates I could ever ask for: Nathan Majernik and Rich Sportsman, though Rich, you're never getting that keyboard back. I also must thank my family, Allison, Dave and Drew for supporting me, believing in me, and inspiring me throughout the years. I must also thank Artemis for her numerous contributions to my manuscripts, surely some of them will sneak in to the final product. Finally I have to thank my partner Kristen, who put up with four years of long distance and 'I don't know when I'll graduate', but stuck with me, and supported me with love and goofiness all along the way.

VITA

EDUCATION

University of California, Los Angeles

MS in Biochemistry, Structural and Molecular Biology 2016

University of Chicago 2012

BS in Mathematics, BA in Biology

RESEARCH EXPERIENCE

University of California, Los Angeles

Graduate Student Researcher Advisor: Sriram Kosuri 2016-Present

Graduate Student Researcher Advisor: Yvonne Chen 2015-2016

University of Chicago

Undergraduate Researcher Advisor: Shohei Koide 2009-2012

OTHER WORK EXPERIENCE

Epic 2012-2014

Technical Services

Cabot Microelectronics 2008

Research Intern

PUBLICATIONS

Boldridge WC*, Hochberg GKA*, Lee J, Kosuri S, Thornton JW. Comprehensive experimental analysis of functional diversification of seven hundred million years of bZip evolution. *In prep.*

Boldridge WC*, Ljubetic A*, Kim HB, Lubock N, Siladji D, Lee J, Jerala R, Kosuri S. A multiplexed bacterial two-hybrid for rapid characterization of protein-protein interactions and

iterative protein design. *BioRxiv*. 2020

Davis JE, Insigne KD, Jones EM, Hastings QA, **Boldridge WC**, Kosuri S. Dissection of c-AMP response element architecture by using genomic and episomal massively parallel reporter assays. *Cell Systems*. 2020; 11(1): 75-85.

Yasui N, Findlay GM, Gish GD, Hsiung MS, Huang J, Tucholska M, Taylor L, Smith L, **Boldridge WC**, Koide A, Pawson T, Koide S. Directed network wiring identifies a key protein interaction in embryonic stem cell differentiation. *Mol Cell*. 2014; 54(6): 1034-1041.

CONFERENCE PRESENTATIONS

Evolutionary Systems Biology 2020

Massively parallel experimental analysis of the evolution of specificity in bZips across the history of Metazoa, Oral Presentation, Cambridge, UK

Protein society 2018

Iterative High-throughput Isolation of Orthogonally Interacting Coiled-coils, Poster Presentation, Boston, MA

Bioorigami 2017

High-throughput Isolation of Orthogonally Interacting Coiled-coils, Poster Presentation, Ljubjana, Slovenia

Engineering Biology Research Consortium 2017

Isolating Orthogonally Interacting Coiled-coils, Oral Presentation, Chicago, IL

Engineering Biology Research Consortium 2016

Protein-protein Interactions: The Next Generation, Poster Presentation, Pasadena, CA

Chapter 1: Introduction

Encoding specificity in protein-protein interactions

A. Specificity in protein-protein interactions: required but poorly understood

Proteins are the major effectors of the molecular processes that constitute life. Although proteins have myriad different functions—catalysis for enzymes, structural support for keratins, signaling for kinases—all classes of proteins participate in protein-protein interactions (PPIs)[1]. The functional purpose of these PPIs is as diverse as that of the proteins that compose them. Examples include epidermal growth factor receptor dimerization which leads to internalization and signalling[2], linking structural supports together as actin-ARP 2/3 linkages to create lamellipodia[3], regulating a kinase's activation as Cdk2 by cyclin A[4] or creating mechanical motion as myosin does while interacting with actin[5]. Hence, understanding PPIs is crucial to understanding any subfield of molecular biology.

All PPIs share several fundamental properties, namely, affinity or how tightly the proteins are bound, avidity/multivalency or how many interactions are occurring, and specificity or to what extent do the proteins in a PPI interact with other proteins. Significant efforts have tried to understand aspects of PPI strength, by identifying extremely strong PPIs[6,7], generating high-affinity binding partners[8–12], or testing what makes transient, strong PPIs possible[13,14]. Similarly, avidity, though less studied, can be seen largely as a function of PPI strength, oligomer number and local environment[15,16]. Specificity, however, is challenging to investigate as it is an emergent property of a PPI network—proteins only form a specific PPI in relation to other proteins/PPIs. This creates a three-fold challenge: first, specificity is defined by preference for one interaction over others but the magnitude of difference that matters, as well as how to interpret multiple interactions is not well defined; second, even when restricted to dimeric proteins the number of possible PPIs scales exponentially with the size of the protein network;

and third, the only network of proteins that is will provide a comprehensive profile of specificity is a network of all of protein sequence space. Obviously, such an immense number of potential interactions is impossible to investigate[17], so most studies have been limited to a few highly probable interactions[18] leaving the molecular basis of specificity poorly understood.

An approach that compromises between the immense number of potential interactions and possible experimental throughput would focus on portions of interaction space known to exhibit specificity. Fortunately, we know of numerous protein families such as BCL-2s[19–21], coiled-coils[22,23], histidine kinases[24,25], and colicins[26–28] where members have distinct interaction profiles despite their overall high level of homology. Characterizing the interaction profile of proteins that exhibit specificity would let us answer questions about the sequence determinants of specific interactions and, in natural proteins, the processes that produce them. However, to characterize how an interaction profile varies between highly similar proteins first requires measuring interactions against multitudes of partners. Large scale testing of PPIs would clarify each individual residue's contribution to an interaction and how it varies across all interactions.

B. The limitations of technologies to investigate protein-protein interactions

One major obstacle to investigating specificity at such a granular level is the current technology. Gold-standard biochemical techniques such as Surface Plasmon Resonance[29] or Isothermal Calorimetry[30] require purification of each protein and individual testing, which cannot be done on more than a handful of interactions. Higher throughput methods, such as two-hybrid techniques[31] can better investigate specificity but scaling remains problematic. Yeast two-hybrids (Y2H)[32] are the most common system to investigate PPIs at moderate and larger scales,

due to commercially available vectors and the ability to achieve moderate throughput via yeast mating[33]. Because of this there have been a wide variety of techniques to expand the throughput of Y2H, such as pooled mating with mathematical[34] or further mating[34] deconvolution. However, despite these efforts, all large scale PPI experiments (characterizing interactomes), have relied on brute force and vast resources.[35–38]

Although yeast dominate the two-hybrid space, other model systems also have assays to identify PPIs. Most notably, in *E. coli* there are two-hybrids, using subunits of RNA polymerase[39,40], viral repressors[41], or split adenylate cyclase[42]. Despite the advantages of the molecular tools in *E. coli*, freedom from PPIs needing to occur in the nucleus, and a comparative dearth of off-target interaction partners, bacterial two hybrids have remained less popular as they do not easily scale.

In the last fifteen years, DNA sequencing has become orders of magnitude cheaper and it is now possible to sequence hundreds of millions of DNA sequences at once[43,44]. Because of this there has been a concerted effort to link up functional assays with a next-generation sequencing readout[45,46] and we can now efficiently measure phenomena as diverse as promoter activation[47–49], olfactory stimulation[50] and apoptosis[51] in a high-throughput manner. Accordingly, there have been several attempts to attach a next-generation sequencing readout to the Y2H with Barcode Fusion Genetics[52], CrY2H-seq[53], rec-YnH[54], RLL-seq[55], and SynAg[56] mating all allowing multiplexed identification of interacting baits and preys.

However, each high-throughput yeast system has limitations which render it unsuitable for studying specificity at high resolution. CrY2H-seq, rec-YnH and RLL-seq involve sequencing a small portion of the total protein, which for proteins with high sequence identity may not be able to resolve the interaction partners. Furthermore, Barcode Fusion Genetics has a

labor-intensive colony picking and arrayed PCR step which limits its throughput and SynAg mating has lower throughput and occurs outside the cell which may influence the biological relevance of interactions. Moreover, all of these techniques involve all-against-all libraries which prevents the more targeted interrogation of PPIs of interest. Finally, like the yeast two-hybrid, the bacterial two-hybrid has been multiplexed[57], but lack of control over library creation, lack of ability to identify interactions completely, and absence of bioinformatical pipeline limit its use.

All of these high-throughput methods are reliant on either externally provided ORFomes, randomly sheared DNA, or point mutants around a single sequence, any of which dramatically constrains the PPIs that can be investigated. Gene synthesis overcomes most limitations on sequence design. Genes can be synthesized with commercially available pools of oligonucleotides, which provide an economical option for creating tens of thousands of highly divergent variants[58,59]. If the oligonucleotides currently available are too short to synthesize the desired sequences, multiplexed gene assembly techniques[60,61] can be used to create the library of sequences. However, if gene synthesis is used, the entire protein must be sequenced to verify faithful production of the intended sequence. This generally requires lower throughput next-generation sequencing or synthetic long reads as the highest throughput sequencing is currently limited to 300bp[62]. One solution is to use a barcoding scheme, where a short random DNA sequence—a barcode—is cloned into the same plasmid as the protein sequence. Then lower throughput sequencing can be used to read through the protein and the barcode in a single read, after which the barcode can serve to unambiguously identify the protein variant[63] and the barcode can be analyzed on the highest throughput sequencing systems.

C. Coiled-coils as a minimal model system for investigating specificity

There is a wealth of potential model systems for investigating specificity of PPIs as even the simplest protein sequences can exhibit complex behavior. For example, poly-α-amino acids—proteins composed of only a single amino acid type—can take on secondary structure and even alternate between α-helixes and β-sheets[64] and binding interfaces composed of solely of tyrosine and serine can mediate low nanomolar interactions to arbitrary targets[8,65]. One of the simplest protein domains is the coiled-coil domain, which are small alpha helical proteins that supercoil around themselves. First identified in α-keratin by both Crick and Pauling at the dawn of molecular biology[66–68], coiled-coils are determined by a unique amino acid signature, the heptad repeat. The heptad repeat is described as residue positions A-B-C-D-E-F-G, which have the states H-P-P-H-P-P-P where H is a hydrophobic residue and P is a polar residue.[69] The hydrophobic residues at the A- and D-positions form the core of the oligomeric interface, while the polar residues at the E- and G-positions largely determine specificity with salt bridges[70]. Coiled-coils most often dimerize[71] but there are examples of natural coiled-coils that trimerize and tetramerize, and engineers have created non-canonical coiled-coils with up to nonamers[72]. Despite having specificity determined by a handful of residues, dimeric coiled-coils exhibit surprising specificity[22,23], which is critical as they are widespread in the genome and occur in 10% of eukaryotic proteins[73]. In fact, the high-degree of specificity is likely why coiled-coils are so prevalent and why they are so commonly found in transcription factors[74].

As an extremely simple domain, there are many computational methods for describing coiled-coils. Dating back to Crick's parameterized equations[75] for describing coiled-coils mathematically, researchers have created a wealth of tools produced for describing[76–78] coiled-coils and predicting[79–82] interactions, largely based on linear models which are easy to

comprehend. However, while these tools can be used to design a single interaction with high certainty, they are not accurate enough to successfully design more complex behavior such as specificity, without a system that would allow facile characterization of large numbers of interactions. However, with such a system, given their high degree of specificity, widespread application in biology, and rich history and relative predictive capability, coiled-coils provide an ideal model system to investigate the determinants of specificity.

D. There is an unmet need for large sets of specific proteins

The ability to control many PPIs simultaneously is a major challenge for synthetic biology[83,84]. Natural proteins often perform poorly in this context, due to their homology with other endogenous proteins, which has driven efforts to create designed proteins with specific interaction patterns. Because of their simplicity, coiled-coils provide a promising scaffold for building *de novo* designed proteins[85] and there are a variety of tools for further work, including toolkits with K_Ds ranging from low-nanomolar to high-micromolar affinities[86] and different oligomeric states[87]. Accompanying these tools, several groups have tried to program specificity, but these sets have been small such as the four interactions in the PNIC[88] and Crooks-2016[89] sets or eight in the Crooks-2017 set[82], or have numerous off-target interactions, like the SYNZIPs[90].

The lack of specific coiled-coil reagents is currently a limiting factor in numerous applications. Protein origami, analogous to DNA origami, creates large defined nanostructures by using a single-chain polypeptide which has a number of self-interactions that cause it to fold into the desired shape[91]. Protein origami has been able to create tetrahedra[92] and bipyramids[93] but has not yet been able to create larger solids such as octahedra because of a lack of orthogonal

coiled-coils. Similarly, more complex genetic[84] and signaling circuits[94] would be enabled by more orthogonal proteins.

It is not just abstract problems that need orthogonal reagents. Chimeric antigen receptors (CARs), where T-cells are rewired to recognize and kill a target determined by an extracellular antibody fragment, are the most promising cancer therapy to date. To allow CAR T-cells to respond dynamically to their environments, orthogonal coiled-coils have been co-opted to display different domains[95], but without a larger number of orthogonal coiled-coils the possible responses are limited. Likewise, polyketide synthases and non-ribosomal peptide synthetases—both large protein complexes that assemble complex chemicals in a modular manner—can be assembled from component modules which are linked by coiled-coils to produce otherwise unsynthesizable chemicals[96–98]. However, the complexity of potential chemicals is restricted by the lack of orthogonal coiled-coils.

E. Natural development of specificity faces greater hurdles than designed specificity

While we can program designed proteins to be specific with our knowledge of protein structure and function as the only limitation, nature has no such ability. In nature, new proteins most often arise through gene duplication[99] which means a protein will originally share all PPIs with its paralog. How a duplicated paralog changes some interactions, while maintaining others is a crucial question as rewiring PPI networks is implicated as a major source of phenotypic diversity[100]. There has been a dearth of studies that have addressed how PPIs can change over evolutionary time. Though characterization of PPIs within species is common[36,37,101–106], and

there is some work comparing PPIs between extant organisms[23,107–109], understanding the evolutionary causes of PPIs requires characterization of ancestral PPIs.

Characterizing ancestral PPIs is possible due to ancestral sequence reconstruction (ASR) [110,111]. ASR uses extant protein sequences, coupled with models of sequence evolution to infer ancestral sequence states. These sequences can then be experimentally tested to determine the causes of changes in protein function. A handful of previous studies have used ASR to investigate PPIs, and were able to trace phenomena like the emergence of tetramerization in hemoglobin[112] or found a PPI with an Intrinsically Disordered Protein gained affinity over time[113,114]. However, these have used a handful of ancestral proteins, which does not allow insights into the tempo, abruptness, evolutionary process and permeance of changes among PPIs—characterizing such phenomena would require a comprehensive analysis across a family of PPIs.

Experimental analysis of the evolution of PPIs is necessary as nearly every possibility has been suggested for the above phenomena. It has been argued that the tempo of specificity gain must be rapid to escape from paralog interference[115], but a majority of proteins with 30% sequence identity conserve interactions[116] and nearly 50% of paralogous PPIs can compensate for each other[117]. The abruptness of specificity gain is unknown: ancestors of hemoglobin switch from being a dimer to a tetramer with only one substitution[112] but high-throughput studies have shown PPIs developing from many substitutions acting in a semi-additive manner[118]. The evolutionary processes that drive the gain of specificity have never been empirically tested, but positive selection for rewiring interactions has been inferred for several classes of proteins[119,120]. Others, however, have denied that selection necessarily needs a roll in the wiring of PPIs[121]. Finally, the permanence of specificity is unclear: it is known that virus-host interactions are

constantly losing and regaining interactions[122], but it has also been suggested that without such strong selection pressure that sequence space is so vast that proteins would never find their old partners again[123]. Thus, given the broad diversity of opinion, experimental characterization across a PPI family is required to understand how specificity evolves.

F. Investigations into designed and natural specificity

In Chapter 2, a reprint of a submitted manuscript, we first develop the tools to characterize specificity among PPIs at scale. We built, validated, and optimized a novel assay, the Next-Generation Bacterial Two-Hybrid (NGB2H) system which repurposes the adenylate cyclase bacterial two-hybrid[42]. We condense the system to a single plasmid, while adding inducible promoters, and optimized reporters. We address the inability to create purpose-driven designed constructs by using gene synthesis to create library diversity, and unambiguously identify these proteins with a unique 20bp DNA barcode residing in the 3'-UTR of our sfGFP reporter. By mapping our hybrid proteins to a DNA barcode we can use ultrahigh-throughput next-generation sequencing to measure tens of thousands of interactions in a single experiment. We validate this on a previously published set of 256 coiled-coil interactions and find our assay to be highly replicable, internally and externally consistent, with the ability to measure more than 25,000 interactions at once at equal sequencing depth.

We then created a computational framework for rapid prediction of sets of orthogonal coiled-coil interactions. We used this framework to design 8,000 interactions in fifty-five orthogonal subsets using the algorithm designed by Potapov[81]. We tested these interactions with the NGB2H system, and found the largest set of orthogonal coiled-coils to date. As this data set

represented the largest coiled-coil data set, we used it to train a new algorithm, iCipa, for prediction of coiled-coils. iCipa represents a substantial improvement on previous coiled-coil design algorithms, so we used it to design more than 18,000 interactions across 17 subsets, and identified a set of 23 on target interactions which is the largest set of any orthogonal protein to date[123,124].

In Chapter 3, a draft of a future manuscript, we investigate how specificity arises naturally. We used ancestral sequence reconstruction to create a comprehensive map of PPIs across the entire phylogenetic tree of the E4BP4/PAR family of bZip proteins, which dimerize due to their coiled-coil domains. This family has diversified from a single homodimeric ancestor through successive rounds of gene duplications, creating networks of paralogs. We find quite often paralogs that initially interacted with one another rapidly cease to do so, instead losing the ability heterodimerize and sometimes homodimerize, even with relatives containing highly similar interfaces. With our unprecedented phylogenetic resolution we are able to find see these changes gradually accumulate, and that they often do not lead to vast rewiring of the interaction space. Using the phylogenetic structure in our dataset, we dissect the relative roles of natural selection and blind chance in creating this type of self-specificity amongst paralogs. We find that self-specificity evolves at a significant rate by chance alone and that once interactions between paralogs are lost, they are virtually never regained.

We conclude with Chapter 4, where we note further potential uses of the NGB2H system for investigating polymorphic and *de novo* designed proteins. We then suggest further uses of the orthogonal coiled-coils we've designed, and how iterative protein design provides an empirical method for protein engineering, while noting future improvements to iCipa. Finally, we evaluate

how experimental studies of evolution are modifying the some commonly held assumptions and examine how these trends will likely continue.

1. Marsh JA, Teichmann SA. Structure, dynamics, assembly, and evolution of protein complexes. *Annu Rev Biochem*. 2015;84:551-575. doi:10.1146/annurev-biochem-060614-034142
2. Wang Q, Villeneuve G, Wang Z. Control of epidermal growth factor receptor endocytosis by receptor dimerization, rather than receptor kinase activation. *EMBO Rep*. 2005;6(10):942-948. doi:10.1038/sj.embor.7400491
3. Krause M, Gautreau A. Steering cell migration: Lamellipodium dynamics and the regulation of directional persistence. *Nat Rev Mol Cell Biol*. 2014;15(9):577-590. doi:10.1038/nrm3861
4. Jeffrey PD, Russo AA, Polyak K, et al. Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. *Nature*. 1995;376(6538):313-320. doi:10.1038/376313a0
5. Rayment I, Holden HM, Whittaker M, et al. Structure of the actin-myosin complex and its implications for muscle contraction. *Science (80-)*. 1993;261(5117):58-65. doi:10.1126/science.8316858
6. Wallis R, Moore GR, James R, Kleanthous C. Protein-Protein Interactions in Colicin E9 DNase-Immunity Protein Complexes. 1. Diffusion-Controlled Association and Femtomolar Binding for the Cognate Complex. *Biochemistry*. 1995;34(42):13743-13750. doi:10.1021/bi00042a004
7. Schreiber G, Buckle AM, Fersht AR. Stability and function: two constraints in the evolution of barstar and other proteins. *Structure*. 1994;2(10):945-951. doi:10.1016/S0969-2126(94)00096-4

8. Koide A, Gilbreth RN, Esaki K, Tereshko V, Koide S. High-affinity single-domain binding proteins with a binary-code interface. *Proc Natl Acad Sci*. 2007;104(16):6632-6637. doi:10.1073/pnas.0700149104

9. Boder ET, Midelfort KS, Wittrup KD. Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. *Proc Natl Acad Sci U S A*. 2000;97(20):10701-10705. doi:10.1073/pnas.170297297

10. Zahnd C, Spinelli S, Luginbühl B, Amstutz P, Cambillau C, Plückthun A. Directed in Vitro Evolution and Crystallographic Analysis of a Peptide-binding Single Chain Antibody Fragment (scFv) with Low Picomolar Affinity. *J Biol Chem*. 2004;279(18):18870-18877. doi:10.1074/jbc.M309169200

11. Orlova A, Magnusson M, Eriksson TLJ, et al. Tumor imaging using a picomolar affinity HER2 binding Affibody molecule. *Cancer Res*. 2006;66(8):4339-4348. doi:10.1158/0008-5472.CAN-05-3521

12. Votsmeier C, Plittersdorf H, Hesse O, et al. Femtomolar Fab binding affinities to a protein target by alternative CDR residue co-optimization strategies without phage or cell surface display. *MAbs*. 2012;4(3):341-348. doi:10.4161/mabs.19981

13. Nooren IMA, Thornton JM. Structural characterisation and functional significance of transient protein-protein interactions. *J Mol Biol*. 2003;325(5):991-1018. doi:10.1016/S0022-2836(02)01281-0

14. Acuner Ozbabacan SE, Engin HB, Gursoy A, Keskin O. Transient proteinprotein interactions. *Protein Eng Des Sel*. 2011;24(9):635-648. doi:10.1093/protein/gzr025

15. Kane RS. Thermodynamics of multivalent interactions: Influence of the linker. *Langmuir*. 2010;26(11):8636-8640. doi:10.1021/la9047193

16. Erlendsson S, Teilum K. Binding Revisited—Avidity in Cellular Function and Signaling. *Front Mol Biosci*. 2021;7(January):1-13. doi:10.3389/fmolb.2020.615565

17. Currin A, Swainston N, Day PJ, Kell DB. Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem Soc Rev*. 2015;44(5):1172-1239. doi:10.1039/C4CS00351A

18. Schreiber G, Keating AE. Protein binding specificity versus promiscuity. *Curr Opin Struct Biol*. 2011;21(1):50-61. doi:10.1016/j.sbi.2010.10.002

19. Kale J, Osterlund EJ, Andrews DW. BCL-2 family proteins: Changing partners in the dance towards death. *Cell Death Differ*. 2018;25(1):65-80. doi:10.1038/cdd.2017.186

20. Dutta S, Gullá S, Chen TS, Fire E, Grant RA, Keating AE. Determinants of BH3 Binding Specificity for Mcl-1 versus Bcl-xL. *J Mol Biol*. 2010;398(5):747-762. doi:10.1016/j.jmb.2010.03.058

21. Chen L, Willis SN, Wei A, et al. Differential targeting of prosurvival Bcl-2 proteins by their BH3-only ligands allows complementary apoptotic function. *Mol Cell*. 2005;17(3):393-403. doi:10.1016/j.molcel.2004.12.030

22. Newman JRS, Keating AE. Comprehensive identification of human bZIP interactions with coiled-coil arrays. *Science*. 2003;300(5628):2097-2101. doi:10.1126/science.1084648

23. Reinke AW, Baek J, Ashenberg O, Keating AE. Networks of bZIP Protein-Protein Interactions Diversified Over a Billion Years of Evolution. *Science (80-)*. 2013;340(May):730-735. doi:10.1126/science.1233465

24. Capra EJ, Laub MT. Evolution of two-component signal transduction systems. *Annu Rev Microbiol*. 2012;66:325-347. doi:10.1146/annurev-micro-092611-150039

25. Podgornaia AI, Laub MT. Determinants of specificity in two-component signal

transduction. *Curr Opin Microbiol*. 2013;16(2):156-162. doi:10.1016/j.mib.2013.01.004

26. Wojdyla JA, Fleishman SJ, Baker D, Kleanthous C. Structure of the ultra-high-affinity colicin E2 DNase-Im2 complex. *J Mol Biol*. 2012;417(1-2):79-94. doi:10.1016/j.jmb.2012.01.019

27. Meenan NAG, Sharma A, Fleishman SJ, et al. The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. *Proc Natl Acad Sci U S A*. 2010;107(22):10080-10085. doi:10.1073/pnas.0910756107

28. Levin KB, Dym O, Albeck S, et al. Following evolutionary paths to protein-protein interactions with high affinity and selectivity. *Nat Struct Mol Biol*. 2009;16(10):1049-1055. doi:10.1038/nsmb.1670

29. Homola J, Yee SS, Gauglitz G. Surface plasmon resonance sensors: review. *Sensors Actuators, B Chem*. 1999;54(1):3-15. doi:10.1016/S0925-4005(98)00321-9

30. Pierce MM, Raman CS, Nall BT. Isothermal titration calorimetry of protein-protein interactions. *Methods A Companion to Methods Enzymol*. 1999;19(2):213-221. doi:10.1006/meth.1999.0852

31. Stynen B, Tournu H, Tavernier J, Van Dijck P. Diversity in Genetic In Vivo Methods for Protein-Protein Interaction Studies: from the Yeast Two-Hybrid System to the Mammalian Split-Luciferase System. *Microbiol Mol Biol Rev*. 2012;76(2):331-382. doi:10.1128/mmbr.05021-11

32. Fields S, Song O. A novel genetic system to detect protein-protein interactions. *Nature*. 1989;340(6230):245-246. doi:10.1038/340245a0

33. Bendixen C, Gangloff S, Rothstein R. A yeast mating-selection scheme for detection of protein - protein interactions. *Nucleic Acids Res*. 1994;22(9):1778-1779.

doi:10.1093/nar/22.9.1778

34. Zhong J, Zhang H, Stanyon CA, Tromp G, Jr RLF. A Strategy for Constructing Large Protein Interaction Maps Using the Yeast Two-Hybrid System:Regulated Expression Arrays and Two-Phase Mating. *Genome Res*. 2003;13(12):2691-2699. doi:10.1101/gr.1134603

35. Rolland T, Taşan M, Charloteaux B, et al. A proteome-scale map of the human interactome network. *Cell*. 2014;159(5):1212-1226. doi:10.1016/j.cell.2014.10.050

36. Li S, Armstrong CM, Bertin N, et al. A map of the interactome network of the metazoan C. elegans. *Science*. 2004;303(5657):540-543. doi:10.1126/science.1091403

37. Yu H, Braun P, Yildirim MA, et al. High-quality binary protein interaction map of the yeast interactome network. *Science*. 2008;322(5898):104-110. doi:10.1126/science.1158684.High

38. Luck K, Kim DK, Lambourne L, et al. A reference map of the human binary protein interactome. *Nature*. 2020;580(7803):402-408. doi:10.1038/s41586-020-2188-x

39. Dove SL, Joung JK, Hochschild A. Activation of prokaryotic transcription through arbitrary protein-protein contacts. *Nature*. 1997;386(6625):627-630. doi:10.1038/386627a0

40. Dove SL, Hochschild a. Conversion of the omega subunit of Escherichia coli RNA polymerase into a transcriptional activator or an activation target. *Genes Dev*. 1998;12(5):745-754. doi:10.1101/gad.12.5.745

41. Dmitrova M, Younès-Cauet G, Oertel-Buchheit P, Porte D, Schnarr M, Granger-Schnarr M. A new LexA-based genetic system for monitoring and analyzing protein heterodimerization in Escherichia coli. *Mol Gen Genet*. 1998;257(2):205-212.

doi:10.1007/s004380050640

42. Karimova G, Pidoux J, Ullmann a, Ladant D. A bacterial two-hybrid system based on a reconstituted signal transduction pathway. *Proc Natl Acad Sci U S A*. 1998;95(10):5752-5756. doi:10.1073/pnas.95.10.5752

43. Goodwin S, McPherson JD, McCombie WR. Coming of age: Ten years of next-generation sequencing technologies. *Nat Rev Genet*. 2016;17(6):333-351. doi:10.1038/nrg.2016.49

44. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*. 2008;26(10):1135-1145. doi:10.1038/nbt1486

45. Gasperini M, Starita L, Shendure J. The power of multiplexed functional analysis of genetic variants. *Nat Protoc*. 2016;11(10):1782-1787. doi:10.1038/nprot.2016.135

46. Starita LM, Ahituv N, Dunham MJ, et al. Variant Interpretation: Functional Assays to the Rescue. *Am J Hum Genet*. 2017;101(3):315-325. doi:10.1016/j.ajhg.2017.07.014

47. Davis JE, Insigne KD, Jones EM, Hastings QA, Boldridge WC, Kosuri S. Dissection of c-AMP Response Element Architecture by Using Genomic and Episomal Massively Parallel Reporter Assays. *Cell Syst*. 2020;11(1):75-85.e7. doi:10.1016/j.cels.2020.05.011

48. Urtecho G, Tripp AD, Insigne KD, Kim H, Kosuri S. Systematic Dissection of Sequence Elements Controlling σ70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in Escherichia coli. *Biochemistry*. 2019;58(11):1539-1551. doi:10.1021/acs.biochem.7b01069

49. Urtecho G, Insigne KD, Tripp AD, et al. Genome-wide Functional Characterization of Escherichia coli Promoters and Regulatory Elements Responsible for their Function. *bioRxiv*. 2020. doi:10.1101/2020.01.04.894907

50. Jones EM, Jajoo R, Cancilla D, et al. A Scalable, Multiplexed Assay for Decoding GPCR-

Ligand Interactions with RNA Sequencing. *Cell Syst*. 2019;8(3):254-260.e6. doi:10.1016/j.cels.2019.02.009

51. Findlay GM, Daza RM, Martin B, et al. Accurate classification of BRCA1 variants with saturation genome editing. *Nature*. 2018;562(7726):217-222. doi:10.1038/s41586-018-0461-z

52. Yachie N, Petsalaki E, Mellor JC, et al. Pooled-matrix protein interaction screens using Barcode Fusion Genetics. *Mol Syst Biol*. 2016;12(4):863-863. doi:10.15252/msb.20156660

53. Trigg SA, Garza RM, MacWilliams A, et al. CrY2H-seq: A massively multiplexed assay for deep-coverage interactome mapping. *Nat Methods*. 2017;14(8):819-825. doi:10.1038/nmeth.4343

54. Yang JS, Garriga-Canut M, Link N, et al. rec-YnH enables simultaneous many-by-many detection of direct protein–protein and protein–RNA interactions. *Nat Commun*. 2018;9(1). doi:10.1038/s41467-018-06128-x

55. Yang F, Lei Y, Zhou M, et al. Development and application of a recombination-based library versus library highthroughput yeast two-hybrid (RLL-Y2H) screening system. *Nucleic Acids Res*. 2018;46(3):1-12. doi:10.1093/nar/gkx1173

56. Younger D, Berger S, Baker D, Klavins E. High-throughput characterization of protein–protein interactions by reprogramming yeast mating. *Proc Natl Acad Sci U S A*. 2017;114(46):12166-12171. doi:10.1073/pnas.1705867114

57. Andrews SS, Schaefer-Ramadan S, Al-Thani NM, Ahmed I, Mohamoud YA, Malek JA. High-resolution protein–protein interaction mapping using all- versus -all sequencing (AVA-Seq) . *J Biol Chem*. 2019;294(30):11549-11558. doi:10.1074/jbc.ra119.008792

58. Kosuri S, Church GM. Large-scale de novo DNA synthesis: technologies and applications. *Nat Methods*. 2014;11(5):499-507. doi:10.1038/nmeth.2918

59. Hughes RA, Ellington AD. Synthetic DNA synthesis and assembly: Putting the synthetic in synthetic biology. *Cold Spring Harb Perspect Biol*. 2017;9(1). doi:10.1101/cshperspect.a023812

60. Plesa C, Sidore AM, Lubock NB, Zhang D, Kosuri S. Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science (80-)*. 2018;359(6373):343-347. doi:10.1126/science.aao5167

61. Sidore AM, Plesa C, Samson JA, Kosuri S. DropSynth 2.0: high-fidelity multiplexed gene synthesis in emulsions. *bioRxiv*. 2019:740977. doi:10.1101/740977

62. NovaSeq Reagent Kits. https://www.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/novaseq-reagent-kits.html.

63. Buschmann T, Bystrykh L V. Levenshtein error-correcting barcodes for multiplexed DNA sequencing. *BMC Bioinformatics*. 2013;14(1). doi:10.1186/1471-2105-14-272

64. Katchalski-Katzir E. Synthesis, structure and function of poly-α-amino acids - The simplest of protein models. *Cell Mol Life Sci*. 1997;53(10):780-789. doi:10.1007/s000180050099

65. Fellouse FA, Li B, Compaan DM, Peden AA, Hymowitz SG, Sidhu SS. Molecular recognition by a binary code. *J Mol Biol*. 2005;348(5):1153-1162. doi:10.1016/j.jmb.2005.03.041

66. CRICK FHC. Is α-Keratin a Coiled Coil? *Nature*. 1952;170(4334):882-883. doi:10.1038/170882b0

67. Crick FHC. The Fourier transform of a coiled-coil. *Acta Crystallogr*. 1953;6(8):685-689.

doi:10.1107/s0365110x53001952

68. Pauling L, Corey RB. Compound Helical Configurations of Polypeptide Chains: Structure of Proteins of the α-Keratin Type. *Nature*. 1953;171(4341):59-61. doi:10.1038/171059a0

69. Lupas AN, Bassler J. Coiled Coils – A Model System for the 21st Century. *Trends Biochem Sci*. 2017;42(2):130-140. doi:10.1016/j.tibs.2016.10.007

70. Lupas AN, Bassler J, Dunin-Horkawicz S. The Structure and Topology of $α$-Helical Coiled Coils. In: Parry DAD, Squire JM, eds. *Fibrous Proteins: Structures and Mechanisms*. Cham: Springer International Publishing; 2017:95-129. doi:10.1007/978-3-319-49674-0_4

71. Testa OD, Moutevelis E, Woolfson DN. CC+: A relational database of coiled-coil structures. *Nucleic Acids Res*. 2009;37(SUPPL. 1):315-322. doi:10.1093/nar/gkn675

72. Dawson WM, Martin FJO, Rhys GG, Shelley KL, Brady RL, Woolfson DN. Coiled coils 9-to-5: Rational de novo design of α-helical barrels with tunable oligomeric states. *bioRxiv Biochem*. 2021:1-5. http://biorxiv.org/cgi/content/short/2021.01.20.427391v1?rss=1&utm_source=researcher_app&utm_medium=referral&utm_campaign=RESR_MRKT_Researcher_inbound.

73. Liu J, Rost B. Comparing function and structure between entire proteomes. *Protein Sci*. 2001;10(10):1970-1979. doi:10.1110/ps.10101

74. Mason JM, Arndt KM. Coiled coil domains: Stability, specificity, and biological implications. *ChemBioChem*. 2004;5(2):170-176. doi:10.1002/cbic.200300781

75. Crick FHC. The packing of α-helices: simple coiled-coils. *Acta Crystallogr*. 1953;6(8):689-697. doi:10.1107/S0365110X53001964

76. Offer G, Hicks MR, Woolfson DN. Generalized Crick equations for modeling

noncanonical coiled coils. *J Struct Biol*. 2002;137(1-2):41-53. doi:10.1006/jsbi.2002.4448

77. Grigoryan G, Degrado WF. Probing designability via a generalized model of helical bundle geometry. *J Mol Biol*. 2011;405(4):1079-1100. doi:10.1016/j.jmb.2010.08.058

78. Wood CW, Woolfson DN. CCBuilder 2.0: Powerful and accessible coiled-coil modeling. *Protein Sci*. 2018;27(1):103-111. doi:10.1002/pro.3279

79. Mason JM, Schmitz M a, Müller KM, Arndt KM. Semirational design of Jun-Fos coiled coils with increased affinity: Universal implications for leucine zipper prediction and design. *Proc Natl Acad Sci U S A*. 2006;103(24):8989-8994. doi:10.1073/pnas.0509880103

80. Fong J, Keating A, Singh M. Predicting specificity in bZIP coiled-coil protein interactions. *Genome Biol*. 2004;5(2):R11. doi:10.1186/gb-2004-5-2-r11

81. Potapov V, Kaplan JB, Keating AE. Data-Driven Prediction and Design of bZIP Coiled-Coil Interactions. *PLoS Comput Biol*. 2015;11(2):1-28. doi:10.1371/journal.pcbi.1004046

82. Crooks RO, Lathbridge A, Panek AS, Mason JM. Computational Prediction and Design for Creating Iteratively Larger Heterospecific Coiled Coil Sets. *Biochemistry*. 2017;56(11):1573-1584. doi:10.1021/acs.biochem.7b00047

83. Weber W, Fussenegger M. Emerging biomedical applications of synthetic biology. *Nat Rev Genet*. 2012;13(1):21-35. doi:10.1038/nrg3094

84. Chen Z, Kibler RD, Hunt A, et al. De novo design of protein logic gates. *Science (80-)*. 2020;368(6486):78-84. doi:10.1126/science.aay2790

85. Woolfson DN, Bartlett GJ, Burton AJ, et al. De novo protein design: How do we expand into the universe of possible protein structures? *Curr Opin Struct Biol*. 2015;33:16-26. doi:10.1016/j.sbi.2015.05.009

86. Fletcher JM, Harniman RL, Barnes FRH, et al. Self-assembling cages from coiled-coil peptide modules. *Science*. 2013;340(6132):595-599. doi:10.1126/science.1233936

87. Fletcher JM, Boyle AL, Bruning M, et al. A basis set of de novo coiled-Coil peptide oligomers for rational protein design and synthetic biology. *ACS Synth Biol*. 2012;1(6):240-250. doi:10.1021/sb300028q

88. Gradišar H, Jerala R. De novo design of orthogonal peptide pairs forming parallel coiled-coil heterodimers. *J Pept Sci*. 2011;17(2):100-106. doi:10.1002/psc.1331

89. Crooks RO, Baxter D, Panek AS, Lubben AT, Mason JM. Deriving Heterospecific Self-Assembling Protein-Protein Interactions Using a Computational Interactome Screen. *J Mol Biol*. 2016;428(2):385-398. doi:10.1016/j.jmb.2015.11.022

90. Thompson KE, Bashor CJ, Lim WA, Keating AE. SYNZIP Protein Interaction Toolbox: *in Vitro* and *in Vivo* Specifications of Heterospecific Coiled-Coil Interaction Domains. *ACS Synth Biol*. 2012;1(4):118-129. doi:10.1021/sb200015u

91. Lapenta F, Aupič J, Strmšek Ž, Jerala R. Coiled coil protein origami: From modular design principles towards biotechnological applications. *Chem Soc Rev*. 2018;47(10):3530-3542. doi:10.1039/c7cs00822h

92. Gradišar H, Božič S, Doles T, et al. Design of a single-chain polypeptide tetrahedron assembled from coiled-coil segments. *Nat Chem Biol*. 2013;9(6):362-366. doi:10.1038/nchembio.1248

93. Ljubetič A, Lapenta F, Gradišar H, et al. Design of coiled-coil protein-origami cages that self-assemble in vitro and in vivo. *Nat Biotechnol*. 2017;35(11):1094-1101. doi:10.1038/nbt.3994

94. Fink T, Lonzarić J, Praznik A, et al. Design of fast proteolysis-based signaling and logic

circuits in mammalian cells. *Nat Chem Biol*. 2019;15(2):115-122. doi:10.1038/s41589-018-0181-6

95. Cho JH, Collins JJ, Wong WW. Universal Chimeric Antigen Receptors for Multiplexed and Logical Control of T Cell Responses. *Cell*. 2018;173(6):1426-1438.e11. doi:10.1016/j.cell.2018.03.038

96. Klaus M, D'Souza AD, Nivina A, Khosla C, Grininger M. Engineering of Chimeric Polyketide Synthases Using SYNZIP Docking Domains. *ACS Chem Biol*. 2019;14(3):426-433. doi:10.1021/acschembio.8b01060

97. Meinke JL, Simon AJ, Wagner DT, et al. Employing 25-Residue Docking Motifs from Modular Polyketide Synthases as Orthogonal Protein Connectors. *ACS Synth Biol*. 2019;8(9):2017-2024. doi:10.1021/acssynbio.9b00047

98. Bozhueyuek KAJ, Watzel J, Abbood N, Bode HB. Synthetic zippers as an enabling tool for engineering of non-ribosomal peptide synthetases. *bioRxiv*. 2020:1-22. doi:10.1101/2020.05.06.080655

99. Ohno S. *Evolution by Gene Duplication*. Berlin, Heidelberg: Springer Berlin Heidelberg; 1970. doi:10.1007/978-3-642-86659-3

100. Sharma S, Pinkert S, Nagaraju S, et al. Analysis of the human protein interactome and comparison with yeast, worm, and fly interaction datasets. *Nat Genet*. 2006;38(3):285-293. doi:10.1038/1747

101. Bartel PL, Roecklein J a, SenGupta D, Fields S. A protein linkage map of Escherichia coli bacteriophage T7. *Nat Genet*. 1996;12(1):72-77. doi:10.1038/ng0196-72

102. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*.

2001;98(8):4569-4574. doi:10.1073/pnas.061034498

103. Uetz P, Giot L, Cagney G, et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*. 2000;403(6770):623-627. doi:10.1038/35001009

104. Rajagopala S V, Sikorski P, Kumar A, et al. The binary protein-protein interaction landscape of Escherichia coli. *Nat Biotechnol*. 2014;32(3):285-290. doi:10.1038/nbt.2831

105. Kitzman JO, Starita LM, Lo RS, Fields S, Shendure J. Massively parallel single-amino-acid mutagenesis. *Nat Methods*. 2015;12(3):203-206, 4 p following 206. doi:10.1038/nmeth.3223

106. Giot L, Bader JS, Brouwer C, et al. A Protein Interaction Map of Drosophila melanogaster. *Science*. 2003;302(December):1727-1737. doi:10.1126/science.1090289

107. Vo T V., Das J, Meyer MJ, et al. A Proteome-wide Fission Yeast Interactome Reveals Network Evolution Principles from Yeasts to Human. *Cell*. 2016;164(1-2):310-323. doi:10.1016/j.cell.2015.11.037

108. Xin X, Gfeller D, Cheng J, et al. SH3 interactome conserves general function over specific form. *Mol Syst Biol*. 2013;9(652):1-17. doi:10.1038/msb.2013.9

109. Zhong Q, Pevzner SJ, Hao T, et al. An inter-species protein–protein interaction network across vast evolutionary distance. *Mol Syst Biol*. 2016;12(4):865. doi:10.15252/msb.20156484

110. Harms MJ, Thornton JW. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat Rev Genet*. 2013;14(8):559-571. doi:10.1038/nrg3540

111. Hochberg GKA, Thornton JW. Reconstructing Ancient Proteins to Understand the Causes of Structure and Function. *Annu Rev Biophys*. 2017;46:247-269. doi:10.1146/annurev-biophys-070816-033631

112. Pillai AS, Chandler SA, Liu Y, et al. Origin of complexity in haemoglobin evolution. *Nature*. 2020;581(7809):480-485. doi:10.1038/s41586-020-2292-y

113. Hultqvist G, Åberg E, Camilloni C, et al. Emergence and evolution of an interaction between intrinsically disordered proteins. *Elife*. 2017;6:1-25. doi:10.7554/eLife.16059

114. Jemth P, Karlsson E, Vögeli B, et al. Structure and dynamics conspire in the evolution of affinity between intrinsically disordered proteins. *Sci Adv*. 2018;4(10). doi:10.1126/sciadv.aau4130

115. Baker CR, Hanson-Smith V, Johnson AD. Following Gene Duplication, Paralog Interference Constrains Transcriptional Circuit Evolution. *Science (80-)*. 2013;342(6154):104-108. doi:10.1126/science.1240810

116. Levy ED, Erba EB, Robinson C V., Teichmann SA. Assembly reflects evolution of protein complexes. *Nature*. 2008;453(7199):1262-1265. doi:10.1038/nature06942

117. Diss G, Gagnon-Arsenault I, Dion-Coté AM, et al. Gene duplication can impart fragility, not robustness, in the yeast protein interaction network. *Science (80-)*. 2017;355(6325):630-634. doi:10.1126/science.aai7685

118. Podgornaia AI, Laub MT. Pervasive degeneracy and epistasis in a protein-protein interface. *Science (80-)*. 2015;347(6222):673-677. doi:10.1126/science.1257360

119. Beltrao P, Serrano L. Specificity and evolvability in eukaryotic protein interaction networks. *PLoS Comput Biol*. 2007;3(2):0258-0267. doi:10.1371/journal.pcbi.0030025

120. Kim PM, Korbel JO, Gerstein MB. Positive selection at the protein network periphery: Evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci U S A*. 2007;104(51):20274-20279. doi:10.1073/pnas.0710183104

121. Wagner A. How the global structure of protein interaction networks evolves. *Proc R Soc B*

Biol Sci. 2003;270(1514):457-466. doi:10.1098/rspb.2002.2269

122. Franzosa EA, Xia Y. Structural principles within the human-virus protein-protein interaction network. *Proc Natl Acad Sci U S A*. 2011;108(26):10538-10543. doi:10.1073/pnas.1101440108

123. McClune CJ, Alvarez-Buylla A, Voigt CA, Laub MT. Engineering orthogonal signalling pathways reveals the sparse occupancy of sequence space. *Nature*. 2019;574(7780):702-706. doi:10.1038/s41586-019-1639-8

124. Chen Z, Boyken SE, Jia M, et al. Programmable design of orthogonal protein heterodimers. *Nature*. 2019;565(7737):106-111. doi:10.1038/s41586-018-0802-y

Chapter 2

A Multiplexed Bacterial Two-Hybrid for Rapid Characterization of Protein-Protein Interactions and Iterative Protein Design

Title: A Multiplexed Bacterial Two-Hybrid for Rapid Characterization of Protein-Protein Interactions and Iterative Protein Design

Authors: W. Clifford Boldridge[1§], Ajasja Ljubetič[2§], Hwangbeom Kim[1#], Nathan Lubock[1¶], Dániel Szilágyi[3], Jonathan Lee[5°], Andrej Brodnik[3], Roman Jerala[2,4*], Sriram Kosuri[1,6*¶]

Author Affiliations:

[1]Department of Chemistry and Biochemistry, University of California, Los Angeles, CA, USA

[2]Department of Synthetic Biology and Immunology, National Institute of Chemistry, Ljubljana, Slovenia

[3]University of Primorska, Koper, Slovenia.

[4]EN-FIST Centre of Excellence, Ljubljana, Slovenia

[5]Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, CA, USA

[6]UCLA-DOE Institute for Genomics and Proteomics, Molecular Biology Institute, Quantitative and Computational Biology Institute, Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, and Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, CA, USA

§ Authors have contributed equally.

Current address: Samsung Biologics, Incheon, Republic of Korea

¶ Current address: Octant Inc. Emeryville, CA, USA

° Current address: Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

* Correspondence should be addressed to S.K. Email: sri@ucla.edu and R.J. Email: roman.jerala@ki.si

Abstract:

Protein-protein interactions (PPIs) are required for most biological functions as well as applications ranging from drug design to synthetic cell circuits. Understanding and engineering specificity in PPIs is particularly challenging as subtle sequence changes can drastically alter specificity. Coiled-coils are small protein domains that have long served as a simple model for studying the sequence-determinants of specificity and have been used as modular building blocks to build large protein nanostructures and synthetic circuits. Despite their simple rules and long-time use, building large sets of well-behaved orthogonal pairs that can be used together is still challenging because predictions are often inaccurate, and, as the library size increases, it becomes difficult to test predictions at scale. To address these problems, we first developed a method called the Next-Generation Bacterial Two-Hybrid (NGB2H), which combines gene synthesis, a bacterial two-hybrid assay, and a high-throughput next-generation sequencing readout, allowing rapid exploration of interactions of programmed protein libraries in a quantitative and scalable way. After validating the NGB2H system on previously characterized libraries, we designed, built, and tested large sets of orthogonal synthetic coiled-coils. In an iterative set of experiments, we assayed more than 8,000 PPIs, used the dataset to train a novel linear model-based coiled-coil scoring algorithm, and then characterized nearly 18,000 interactions to identify the largest set of orthogonal PPIs to date with twenty-two on-target interactions.

Introduction:

Protein-protein interactions (PPIs) are integral to most biological functions and are required for such diverse processes as cell division, signaling, metabolism and transcription and translation[1]. Our ability to design and build functions and structures as complex as nature is still in its infancy, but is developing with advances in both protein design algorithms and gene synthesis capacities. For example, orthogonal *de novo* designed proteins and the careful reuse of well-characterized orthogonal interactions found in natural systems facilitate building nanoscale superstructures for applications in biology, biological engineering and materials science[2]. Supramolecular protein designs can be created using simple, natural protein families like coiled-coils, which have been used to build numerous designed protein assemblies[3–5]. However, identifying orthogonal natural proteins is difficult, as evolutionarily related proteins often display significant cross-interactions. Another method is to computational design *de novo* proteins; in particular, Rosetta-based designs have produced homodimers[6,7] and heterodimers[8]. However, predicting orthogonal binding and designing large orthogonal sets remains beyond current de-novo design methods[3].

Coiled-coils in particular have many useful characteristics for atomically precise designs of macromolecular structures. They are small, precisely oriented, and numerous sequence-based and computational models exist to describe their properties. First identified at the dawn of molecular biology by both Pauling[9] and Crick[10], coiled-coils are defined by their heptad repeat H-P-P-H-P-P-P (H = hydrophobic residue, P = polar residue). This relatively simple structure has given rise to many computational models to describe coiled-coil interactions, from the parametric Crick equations in 1953[11] to contemporary linear models[12–15]. However, because of their shared similar structure, building large sets of orthogonally interacting coiled-coils, where all on-target interactions occur to the exclusion of all off-target interactions, is still difficult.

Though numerous groups have attempted to create orthogonal sets of coiled-coils, they've been limited in size and still display significant off-target interactions[15,16]. Increasing our ability to design, build and characterize large sets of interacting proteins could help solve this problem by providing empirical data to improve computational models of PPIs. Simultaneously this would vastly increase the number of available orthogonal building blocks for nanoscale structural design allowing the creation of previously unbuildable structures.

Here we combine gene synthesis, a novel assay that allows for multiplexed bimolecular interaction screening, and a computational pipeline to design large libraries of orthogonally interacting coiled-coils. We first built and validated a novel assay, the Next-Generation Bacterial Two-Hybrid (NGB2H) system that has a number of unique advantages over other methodologies for characterizing protein libraries. In particular, the NGB2H system allows for screening of bimolecular interactions without having to test all-against-all libraries, direct large-scale synthesis using oligonucleotide arrays to explore design space, quantitative readouts on the entire library including negative interactions, and allows for understanding low affinity interactions inside the crowded cellular context. We did this iteratively with synthetically-designed libraries increasing in size from 256 interactions to more than 18,000 interactions. From this we identified the largest sets of orthogonal proteins to date and developed an improved coiled-coil design algorithm for future design purposes of this versatile protein domain.

Results:

<u>NGB2H system design</u>

Despite a wealth of techniques to analyze PPIs, there is not currently a method that facilitates high-throughput characterization while analyzing PPIs in formats other than all-against-all or that is able to distinguish between closely related constructs. However, such a system would allow investigations of PPIs within protein families, polymorphic PPIs, and de novo designed PPIs that are currently intractable. Thus, we built a generalizable, scalable bacterial two-hybrid system using a significantly modified version of the *B. pertussis* adenylate cyclase two-hybrid[17] (Figure 2.1A, Supplementary Information Section 1). Briefly, the two-hybrid functions much as in Karimova et al.[17], where interacting hybrid proteins reconstitute adenylate cyclase to produce cAMP which drives reporter gene expression. We measured relative transcription of a uniquely identifying DNA barcode residing in the reporter gene, which serves as a measure for interaction strength. The barcode is mapped to the two fully sequenced hybrid proteins at an early cloning step using high-throughput sequencing when the barcode and proteins are physically adjacent. This unambiguously identifies even highly homologous proteins and separates synthetic errors from programmed designs. Thus, measuring the relative barcode transcription provides a quantitative, massively multiplexed characterization of PPIs with short read sequencing. Because the NGB2H system uses a mapping step, it can use gene synthesis, rather than preconstructed libraries to create diversity, which further frees it from one-against-all or all-against-all testing common in two-hybrids. We made a number of other improvements, including: (1) titratable and inducible control of hybrid protein expression and optimized reporter response on a single plasmid, (2) a background strain with linear cAMP accumulation, (3) a GFP reporter instead of beta-galactosidase for more rapid individual characterization, (4) the use of multiple barcodes per construct to achieve better statistical certainty, and (5) a scarless cloning

scheme that allows for library creation with any designed sequence. (More information Supplementary Information Section 1).

Validation of the NGB2H system

After optimizing the system with single construct GFP measurements (Figure S2.1), we validated the NGB2H system with 256 previously characterized interactions[15], which we call the CC0 Library. The CC0 Library is a set of sixteen *de novo* designed, orthogonal, heterodimeric coiled-coils tested in an all-against-all configuration. The proteins are highly similar, four heptad coiled-coils which vary only at the a-position (Ile/Asn), e-position and g-position (Lys/Glu) (Figure 2.1B). We designed the CC0 Library to be compatible with our system (Figure S2.2A), barcoded and cloned it (Figure S2.3A, S2.4). After inducing the two-hybrid for six hours, we took samples for RNA and DNA extraction to measure interaction strength and normalize for plasmid abundance, respectively. We obtained high-quality measurements for all 256 protein pairs and calculated an Interaction score where

$$\text{Interaction score} = \ln\left(\text{median}\left(\frac{\text{RNA reads per barcode}}{\text{DNA reads per barcode}}\right)\right) \; \forall \text{ barcodes} > 10 \text{ DNA reads in all replicates}$$

The NGB2H assay was highly replicable, with biological replicates having similar Interaction scores (Pearson's $r > 0.98$, $p < 10^{-15}$) with a dynamic range of more than 100-fold (Figure 2.1C).

We checked several internal controls to validate the measurements of the NGB2H assay. First, as the protein code is degenerate, we screened nine different codon usages for each pair of proteins. Different codon usages showed consistent Interaction scores (representative pair Figure 2.1D) with all usages correlating with Pearson's $r > 0.92$ and $p < 10^{-15}$ (Figure S2.5), demonstrating minimal effects from DNA sequence variation and low levels of noise in Interaction scores. We also compared the Interaction scores of protein pairs when attached to the

other half of the two-hybrid, which we call the reciprocal orientation. We found that the CC0 Library has a strong correlation between the primary and reciprocal orientations (Pearson's r = 0.92, $p < 10^{-15}$) (Figure 2.1E), indicating that the biological machinery of the NGB2H system faithfully recapitulates the biochemical interaction. In addition, a portion of our library contained frameshift mutations which should not create functional PPIs. As expected, Interaction scores of constructs with indels cluster at the bottom of the range of correct constructs (Figure S2.6). Lastly, to show that the NGB2H system does not suffer from barcode effects or selection pressure from the repeated cloning steps, we replicated the assay with an independent re-barcoding and re-cloning of the CC0 Library which showed strong correlation with the first iteration's Interaction scores (Pearson's $r > 0.98$, $p < 10^{-15}$) (Figure 2.1F).

Having confirmed the internal consistency of the CC0 Library, we compared it to the previously published results. When compared to circular dichroism data published in Crooks et al.[15], we found the NGB2H system's dynamic range correlated well with melting temperatures greater than 40˚C (Figure 2.1G, 2.1H). Given the differences in technique – *in vivo* versus *in vitro*, interaction strength versus helicity – the correlation between the Interaction score and Tm (Pearson's $r > 0.75$, $p < 10^{-15}$, Figure S2.7) largely validates the NGB2H system. Finally, the NGB2H system needs to be highly scalable. To test its scalability, we downsampled the raw sequencing reads between 10 and 150-fold and found strong agreement with our full dataset even when downsampled 100-fold (Pearson's $r > 0.85$, $p < 10^{-15}$, Figure 2.1I), which implies the ability to accurately screen ~25,000 interactions with an equal number of reads.

<u>Computational design of large sets of orthogonal coiled-coils</u>

To computationally predict large, orthogonal sets of coiled-coils for empirical verification, we built a two-step computational pipeline (Figure 2.2A). In brief, we calculated 16.7 million interaction scores for all four heptad coiled-coils with Ile or Asn at the a-position and Glu or Lys at the e- and g-positions using the scoring model from Potapov et al[13]. We then identified sets with an orthogonality gap—those sets where all on-target interactions had a higher predicted score than all off-target interactions. Though computationally challenging, this is tractable as a variant of the maximum independent set problem[18]. We identified the fifteen largest sets and designed each of them with three different backbones (each containing different residues at the noncontact b, c, and f-positions) to investigate their contribution to coiled-coil stability[19]. We combined these with two sets of controls spanning eleven backbones, for a total of 56 sets containing between 64 to 961 interactions (8,169 interactions overall), which we named the CCNG1 Library. After testing a subset of the CCNG1 Library to validate our in-house designs, (Figure S2.8, S2.9, Supplementary Information Section 8.3), we designed (Figure S2.2C), cloned (Figure S2.3C, S2.4) and ran the NGB2H assay, from which we collected quality data (Figure S2.10) on 8,073 interactions.

Large orthogonal sets in the CCNG1 library

Although we designed our coiled-coils to be orthogonal, the current state-of-the-art design algorithms are relatively inaccurate. Thus, similar to our designs, we reduced the problem to the maximum independent set problem to identify the largest orthogonal subset of each set. We were able to identify a set of orthogonal coiled-coils that contains twelve pairs, which includes four heterodimers and eight homodimers (Figure 2.2B). We have also identified a set of seven heterodimers and three homodimers (Figure 2.2C), that has fewer on-target interactions

(ten versus twelve), but contains more proteins (seventeen versus sixteen) and therefore more total potential interactions (153 versus 136).

We characterized the number of on-target interactions across the CCNG1 Library and found 20 of our 51 sets have more than the seven on-target orthogonal interactions in the state-of-the-art set from Crooks et al., (Figure 2.2D), though many share the same interfacial residues (Figure S2.11). We found that these orthogonal sets were composed of between four and seventeen proteins (Figure S2.12), five of which are larger than the CC0 Library. As the CCNG1 Library represents the first large scale systematic investigation into the effects of variation at the b, c, and f-positions, we sought to understand how these positions influenced interactions. We tested six backbones containing the same interfacial residues as the CC0 Library (Figure S2.13, Supplementary Information Section 8.4) and found that charged backbones led to less specific interaction profiles. To understand the limits of producing orthogonal interactions within highly constrained sequence space, we compared the number of total pairs in each set (the sum of interacting and non-interacting pairs for both orthogonal and non-orthogonal pairs) to the number of pairs (interacting and non-interacting) in the largest orthogonal subset (Figure 2.2E) of each full set. We found that the number of orthogonal pairs appears to increase progressively slower as set size increased, suggesting much larger sets would need to be screened to identify proportionally more orthogonal coiled-coils or, alternatively, the accuracy of prediction algorithms would need to be improved.

Improving the state-of-the-art coil-coiled interaction prediction algorithms

The CCNG1 Library dataset represents the largest dataset of coiled-coil interactions to date. We reasoned that our data could serve as a training set to improve on currently available

models. To benchmark current models, we computed scores using algorithms bCipa[14], Potapov/SVR[13], Fong/SVM[20] and Vinson/CE[12] which are all linear models with features for amino acid pairings. Each algorithm is only weakly predictive of our measured interactions with the bA backbone (Figure 2.3A), as all models have an $R^2 < 0.2$. Notably, each algorithm predicted the strongest interactions well, but also predicted many weak interactions that when measured had high Interaction scores. We built several linear models similar to bCipa which included numerous innovations (Supplementary Information Section 3). First, we trained a model on our data that only included weights for the a-, e- and g-position combinations. We also created versions of this simple model with terms for either consecutive residues in the a-position of the same protein or separate terms for weights at the N-terminal a-position, where fraying may occur (Figure S2.14A). We then expanded these models with a novel scoring technique, which we call heptad shifts (Figure S2.14B). In short, we expect the predominant form of coiled-coil interaction to be the alignment of heptads that has the strongest interaction, which is not necessarily all four heptads aligned from the N-terminus, but could be an interface of three or fewer heptads. All of our heptad shifting scoring algorithms were significantly better than the corresponding non-shifting versions and our N-terminal a-positions weights algorithm was significantly better than both the basic algorithm and the consecutive a-position algorithm (Figure 2.3B). Thus, our final model which we call iCipa uses heptad shifting and terms for the N-terminal a-positions, and it is more predictive of CCNG1 Interaction scores than previous models with an $R^2 = 0.27$ (Figure 2.3C).

iCipa is a linear model, which facilitates interpretation. The weights of iCipa have expected and unexpected characteristics (Figure 2.3D). a-position residues prefer Ile/Ile pairings and tolerate Asn/Asn pairings between proteins and disfavor Ile/Asn pairings as expected. As

expected, e- and g-positions favor salt bridges between Glu/Lys and disfavor Glu/Glu pairings, but counterintuitively, Lys/Lys pairings are acceptable for forming the interface which may be due to non-canonical interactions by long side chain of lysine.

To test the iCipa model, we excluded all the data from the original CC0 Library while we trained the weights. When the scoring functions are normalized and compared (Figure 2.3E), both the Potapov/SVR and bCipa algorithms performed worse in predicting the measured melting points with $R^2 < 0.32$ compared to iCipa, $R^2 = 0.48$—a fifty percent increase in predictive ability. Importantly, the increase in predictive power for iCipa on the CC0 Library demonstrates that iCipa has not been trained on an artifact of the NGB2H system but that the NGB2H system provides high quality data on PPIs which can provide general insights into coiled-coil function.

CCmax Library design and verification

To evaluate iCipa's prediction capabilities, demonstrate the scalability of the NGB2H system, and identify larger orthogonal sets of coiled-coils, we built another library, the CCmax Library. The CCmax Library contains 18,491 interactions of 931 different coiled-coils in fifteen predicted orthogonal sets and seven control sets (Figure 2.4A). The orthogonal sets were designed using our computational framework and scored with one of fifteen variants of iCipa. After designing (Figure S2.2D) and cloning we collected high quality data on 17,983 interactions (Figure S2.15). The CC0 Library was a subset of the CCmax Library and it broadly agreed with its performance in our previous libraries (Figure S2.16).

Orthogonal sets of the CCmax Library

We identified orthogonal sets with sizes of up to twenty-two on-target pairs (Figure 2.4B) and 784 total interactions (Figure S2.17). Five of the sets contained more on-target interactions than any set in the CCNG1 Library, and fifteen contained more on-target interactions than the largest published set[15]. Our largest orthogonal set (Figure 2.4C) contained twenty-two coiled-coil dimers, sixteen homodimers and six heterodimers, which is ten on-target interactions larger than the state-of-the-art set from CCNG1. To characterize the accuracy of different iCipa variants, we subsampled each set (ten proteins, 500 times), and found the largest orthogonal set per subsampling. We found little significant difference between the algorithms (Figure S2.18) suggesting that orthogonality is still challenging to design using current algorithmic accuracy and underscoring the necessity of large scale experimental verification.

Different applications need varying levels of orthogonality; while gene circuits likely need extreme orthogonality, protein origami, which benefits from avidity, is not under such constraints. Thus, we identified the largest orthogonality gap for different numbers of on-target interactions. (Figure 2.4D). As expected, smaller sets had larger gaps, but large orthogonality gaps were identified for sets as large as sixteen on-target interactions. Finally, we compared the CCmax Library's Interaction score with iCipa predictions which show substantial improvement over the CCNG1 Library. iCipa was able to predict Interaction scores with $R^2 = 0.43$ (Figure 2.4E). We attribute the increase in iCipa's power to the use of a single coiled-coil backbone which consists of only alanine residues at the b-, c- and f-positions. The improvement in predictive power appeared in other algorithms to a lesser extent, all of which maintained an $R^2 < 0.28$ (Figure S2.19).

Discussion:

We have developed and validated a novel system for high-throughput identification of PPIs. We built a framework to predict orthogonal coiled-coil interactions and used it to design tens of thousands of interactions which we then assayed with the NGB2H system in a design-build-test cycle. Using the data collected, we improved state-of-the-art coiled-coil interaction prediction algorithms which allowed us to design the largest set of any orthogonal proteins to date with twenty-two on-target interactions. Thus, by iterative design we demonstrate how high-throughput PPI identification can facilitate identification of desired protein function and improve design.

Our work builds on previous high-throughput two-hybrids to create a generalizable system for studying PPIs, that could include both soluble and membrane proteins. By uniting gene synthesis with a mapping step and a barcode readout, our system allows high throughput characterization of any binary PPI. Previous high-throughput studies used highly constrained libraries--either the ORFome[21–24] of one of a handful of reference genomes, targeted single residue mutations which only explore a sliver of sequence space around a primary sequence[25,26] or several randomly sheared coding sequences[27]. Using the capabilities of DNA synthesis broadens the testable sequence space which facilitates investigations of a variety of areas such as protein domains, extant genetic variation, evolutionary trajectories or epistatic effects. Furthermore, for the investigator who is not interested in an all-against-all approach, synthesis allows the explicit pairings of only certain proteins. While we benefited from the short length of our proteins of interest, recent pooled gene synthesis techniques[28,29] can be used to interrogate much larger proteins. Deconvoluting library diversity has also been a challenge for other multiplexed assays. Other multiplexed methods involved picking colonies and sanger sequencing

them[21], mapping the beginning of reading frames to reference genomes[22–24] or manually BLASTing obtained reads[27]. Our explicit mapping step allows for the high-throughput creation of a library to map arbitrary proteins to DNA barcodes, and because it is a separate step it could use long read sequencing to overcome the length limitations of Illumina sequencing. Finally, by using a barcode readout downstream of synthesis and mapping we can measure protein libraries in many formats.

Our improvements to coiled-coil design algorithms represent an important advance for *de novo* protein design. Though coiled-coil interactions have been modelled with diverse approaches, our iCipa algorithm shows clear advantages over existing models. In particular, heptad shifting provides an intuitive, biologically rational addition that can be applied to any future improvements in coiled-coil design. Overall, we found iCipa to be substantially more accurate than other tested algorithms, at least for this limited set of residues tested.

Here, we simultaneously performed a massive characterization of PPIs within a protein family and identified the largest set of orthogonal proteins identified to date. The CCmax Library characterized three times as many interactions than any previous intra-protein family work[13]. From the total of 26,049 interactions we characterized, we found a large number of orthogonal proteins—in sets of up to 12 heterodimers or 22 heterodimers and homodimers. Though orthogonal coiled-coils are particularly needed as the building blocks for protein origami[4,5], they could be substituted for histidine kinases in orthogonal signaling pathways[30], synthetic orthogonal transcriptional logic gates[8,31,32], or for orthogonal cellular localization[33].

Thus, the ability to characterize constructs across highly-diverse sequence space and for the identification of networked properties such as orthogonality, highlights the NGB2H's scalability and generality. Because it can be adapted to any sequence the experimenter desires,

the NGB2H facilitates interrogation of PPIs beyond endogenous interactomes, it can be used to characterize whole protein families, empirically inform protein design, or investigate complex phenomena like epistasis.

Acknowledgements:

We thank the members of the Kosuri and Plesa labs for their feedback on the manuscript and figures. We thank Suhua Feng of the UCLA Broad Stem Cell Research Center and the team of the Technology Center for Genomics and Bioinformatics for performing next-generation sequencing. We thank Octant Inc., the Kruglyak Lab at UCLA, and the Black Lab at UCLA, for use of their next-generation sequencers. We thank Mathew Graf and Will Silkworth for their assistance at the UCLA-DOE Biochemistry Shared Instrumentation Facility. We thank Thomas Kuhlman for kindly providing strain TK310. Finally, we thank Chris Voigt for sharing repressor/promoter sequences with us.

Funding:

The National Institutes of Health (DP2GM114829 to S.K.), Searle Scholars Program (to S.K.), ERASynBio (1445112 to S.K., R.J.) MSCA CC-LEGO 792305 to A.L., Slovenian Research Agency (P4-0176 and J1-9173 to R.J.), ERC project MaCChines to R.J.

Author contributions:

H.K., W.C.B., N.L. and S.K. designed the NGB2H system. A.L., D.S. and R.J. designed the large sets of coiled-coils. H.K., N.L. and W.C.B. designed the oligonucleotide libraries. H.K.

W.C.B. and J.L. performed the experiments. A.L. designed the improved interaction algorithms. W.C.B. and N.L. performed the computational analysis. W.C.B , A.L., R.J. and S.K. analyzed the results and iteratively planned the next steps. W.C.B created the figures. S.K., W.C.B. and A.L. wrote the manuscript with input from all authors.

Competing financial interests:

S.K. is cofounder, CEO and holds equity, N.L is an employee and holds equity, and J.L. was an employee and holds equity in Octant Inc. All other authors declare no competing financial interests.

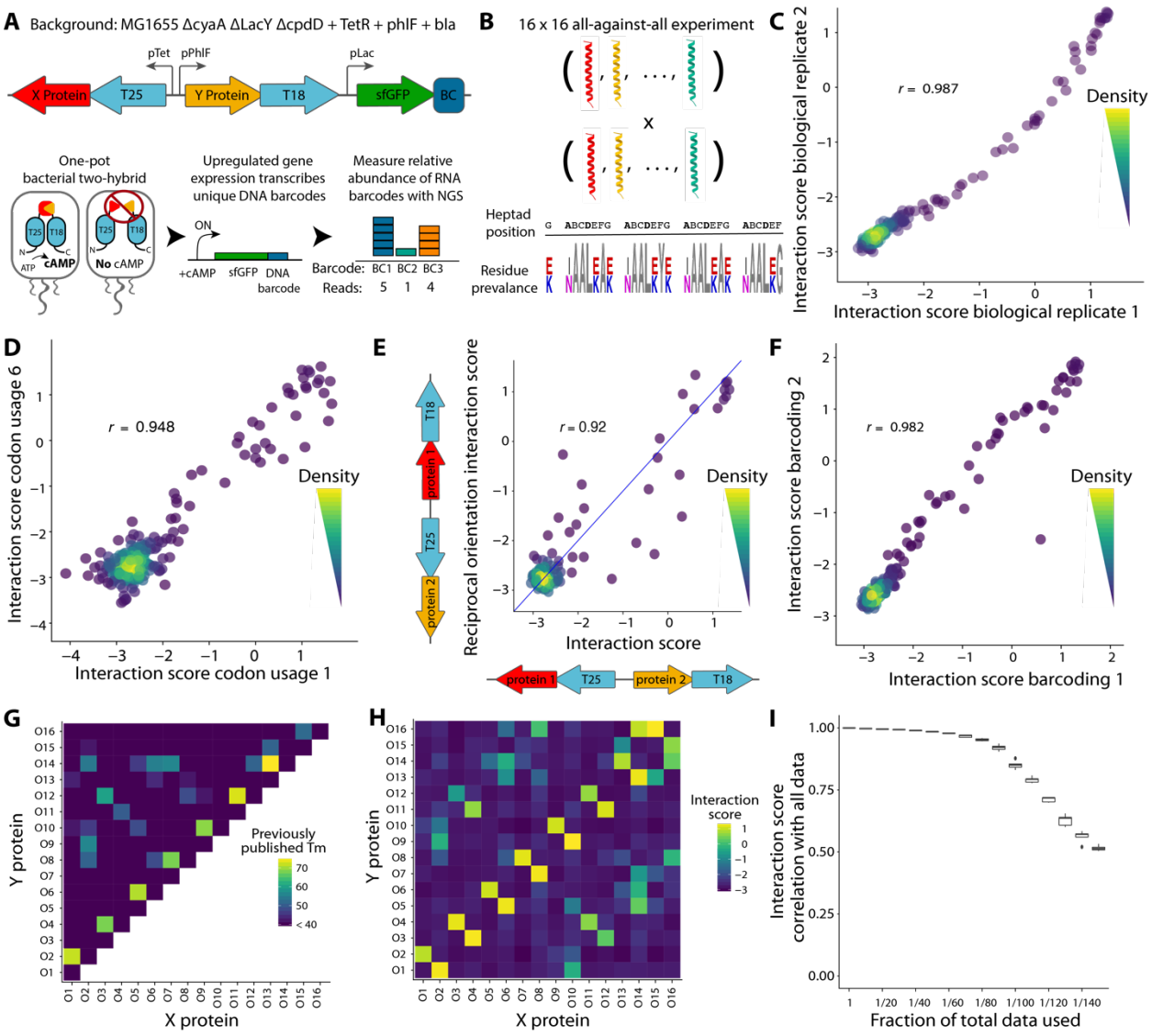

Figure 2.1) Design and validation of the NGB2H assay. A) (Top) Schematic of the NGB2H system reporter construct. T25, T18 - adenylate cyclase halves; BC - unique DNA barcode identifying the protein pair. (Bottom) Workflow of NGB2H system. Interacting proteins reconstitute adenylate cyclase, producing cAMP which drives gene expression of the barcoded sfGFP reporter. Relative barcode abundance is quantified using next generation sequencing (NGS). B) The CC0 Library is composed of 16 coiled-coils tested against one another. (Bottom) Sequence logo representing the diversity represented in the CC0 Library. Residues that vary are

shown in color. C) Interaction scores of CC0 Library members are consistent between biological replicates (Pearson's > 0.98). D) Two different codon usages have consistent interaction scores (Pearson's $r > 0.94$, representative sample). E) Interaction strength is similar (Pearson's $r > 0.92$) regardless of which protein is attached to which half of adenylate cyclase. The blue line represents y = x. F) Interaction scores of separately barcoded, cloned and tested replicates are consistent (Pearson's $r > 0.98$). G) Published circular dichroism (CD) melting point (Tm) data. H) Experimentally determined interaction scores. I) CC0 Library raw data can be subsampled and still correlate well with the full dataset. Boxplots represent the interquartile range of 50 random subsamples of the full data.

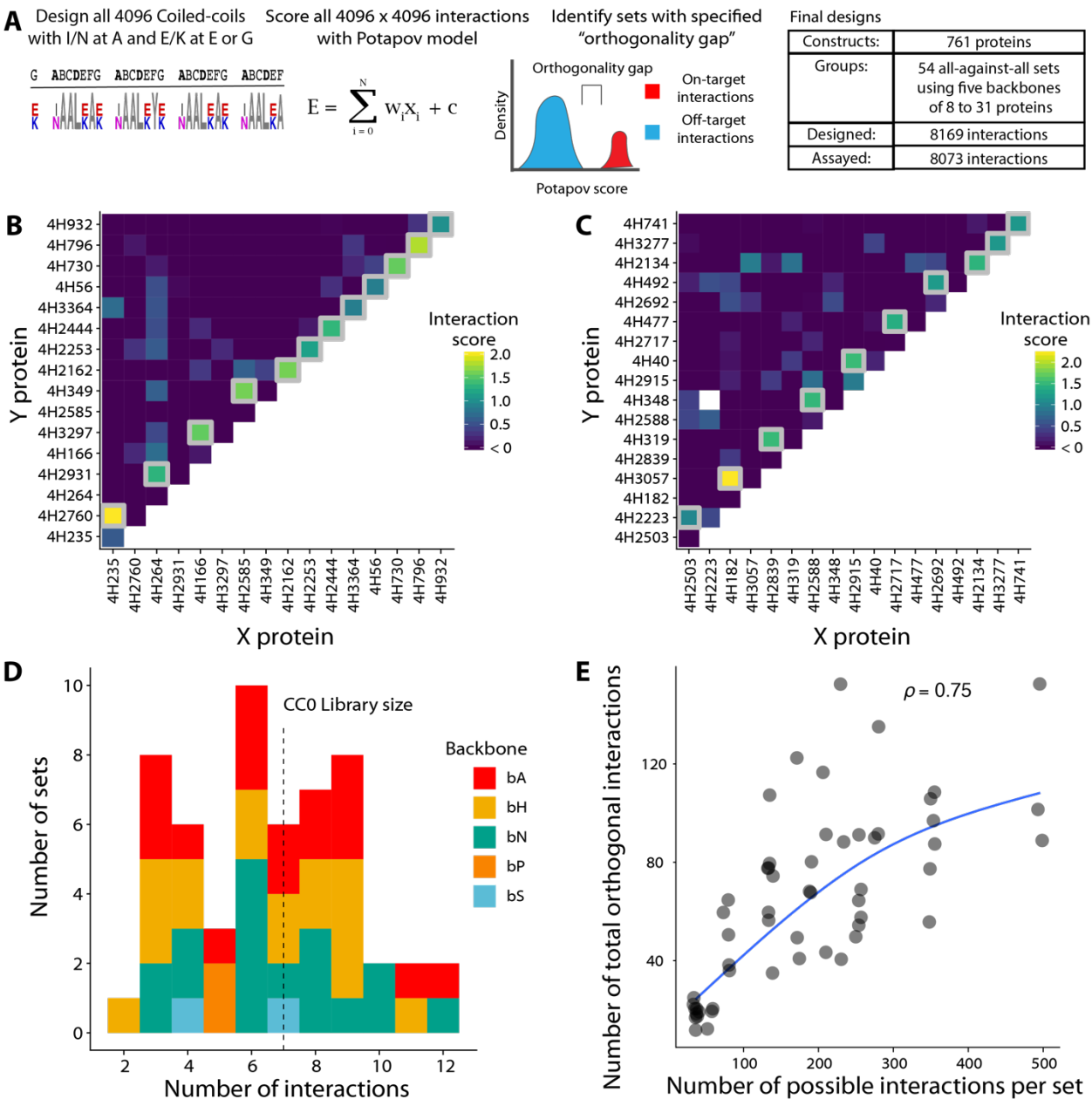

Figure 2.2) Large orthogonal sets of coiled-coils from the CCNG1 library. A) Schematic of the CCNG1 Library design. All 4-heptad coiled-coils with variation at positions a, e, and g were scored for interactions with the model from Popatov et al., and sets of coiled-coils with large orthogonality gaps were identified. In total we designed and tested 54 sets of orthogonal coiled-coils. B) The set of coiled-coils with the largest number of on-target orthogonal interactions (12 on-target interactions). Grey boxes identify on-target interactions. C) Set of orthogonal coiled-

coils with the largest number of total (on-target and non-interacting) interactions (153 total interactions). Grey boxes represent on-target interactions. D) Number of interactions per set of coiled-coils. Dashed line represents the number of on-target orthogonal interactions in the CC0 library. Colors show different backbones used, while the interfacial residues stayed the same. E) Set size increases more rapidly than the number of pairs in the largest orthogonal subset. Blue line represents a spline with degree two.

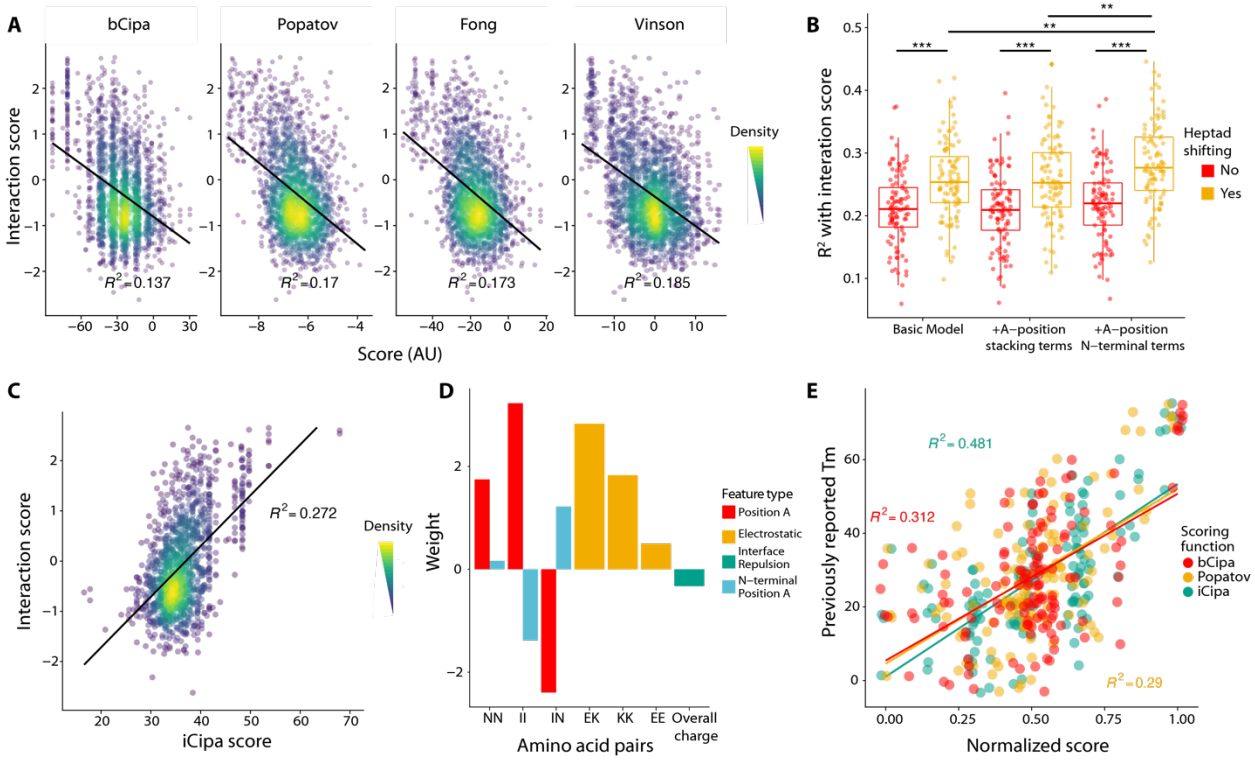

Figure 2.3) Comparison, development and validation of iCipa model. A) Previous models of coiled-coil interactions are only mildly predictive of interactions in the CCNG1 Library. All R^2's are Interaction Score as predicted by the scoring algorithms. The black line represents a linear model of Interaction scores predicted by the different algorithms. B) Coefficient of determination of Interaction scores with different iCipa candidates. Each point represents a subsample of ten percent of the total data. *** = $p < 0.001$, ** = $p < 0.01$, two-tailed t-test. C) iCipa is more predictive of Interaction score ($R^2 > 0.27$) than previous models, shown in A. Black line represents a linear model of Interaction scores as predicted by iCipa scores D) Weights for the iCipa model. Each weight scores a single type pair of amino acid between the two interacting coiled-coils. E) iCipa is more predictive of previously published CC0 Library melting points than the bCipa or Potapov algorithm. Individual dots represent previously reported melting points compared with the normalized score from one of three scoring algorithms.

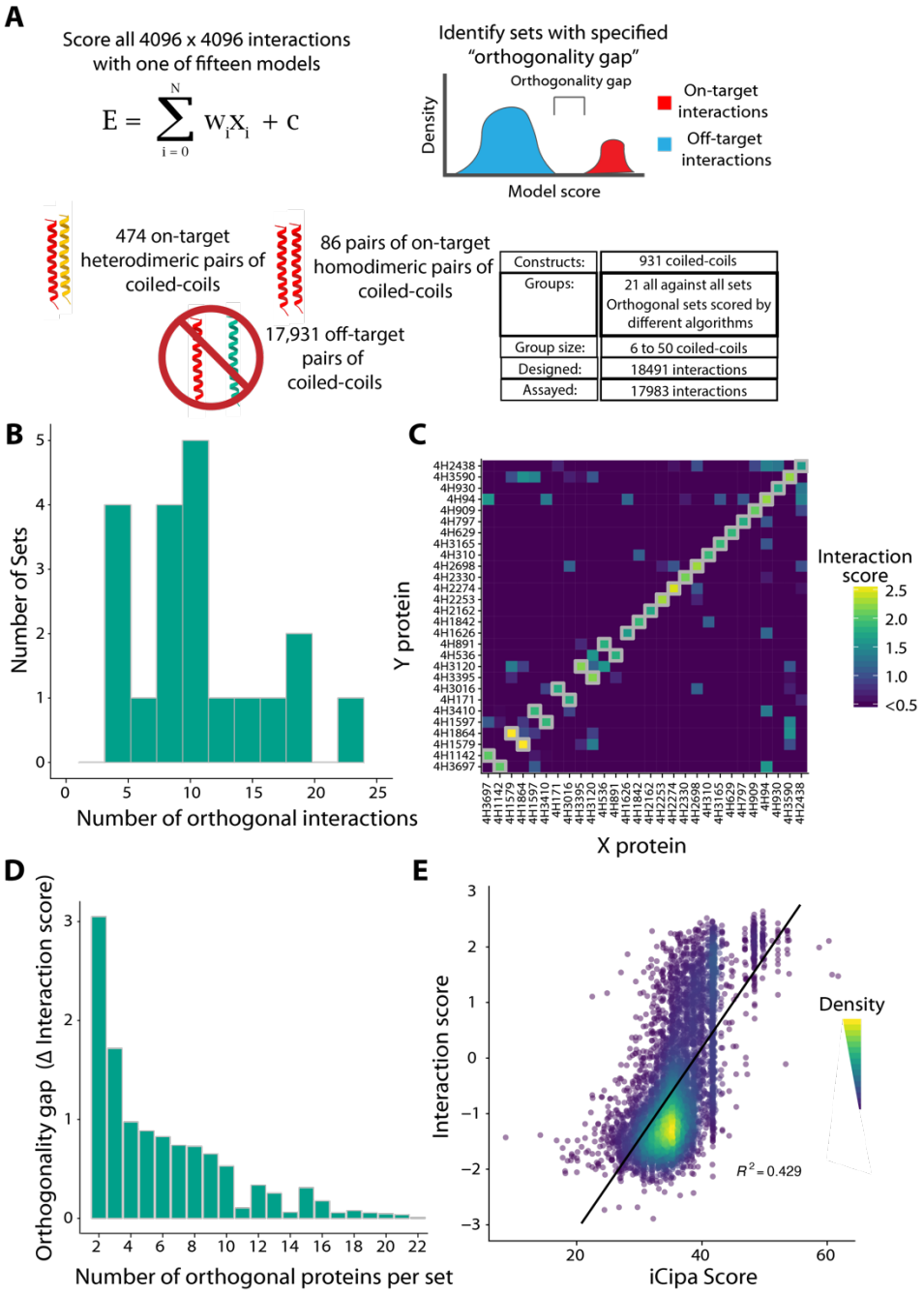

Figure 2.4) The largest orthogonal sets of the CCmax Library. A) Design of the CCmax library. Using seven different algorithms, each possible interaction was scored and sets of orthogonal interactions with an orthogonality gap were identified. In total 21 sets comprised of 18,491 interactions were analyzed. B) Number of on-target orthogonal interactions per set. Between 3 and 22 orthogonal interactions were obtained per set. C) The largest orthogonal set

contains 22 on-target interactions across 784 tested pairs. Grey boxes represent on-target interactions. Unboxed squares represent non-interactions D) The largest orthogonality gap per number on-target interactions in a set. The orthogonality gap is the difference between the weakest on-target Interaction score and strongest off-target Interaction Score. E) iCipa's agreement with the Interaction score ($R^2 = 0.429$) . The black line is a linear model predicting Interaction scores from iCipa predictions.

Methods:

Oligonucleotide designs:

Libraries were designed as shown in Figure S2.2. Though the CC0 and CC1 Libraries were assembled from two oligonucleotides and the CCNG1 and CCmax Libraries from one oligonucleotide, they followed the same overall assembly logic. In brief, each library was flanked with two orthogonal 15bp primers[34] for amplification from the OLS pool. Interior to the flanking primers were type IIS restriction enzyme sites to facilitate scarless cloning, and the complete coiled-coil sequence. The CC0 and CC1 Libraries contain extra type IIS sites and flanking 15bp primers to allow linking and amplification of the X and Y halves of the two-hybrid. A complete description of each design is listed in Supplementary Information Section 2 and all oligonucleotides and all proteins used are available upon correspondence with the author.

Orthogonal coiled-coil interaction prediction:

To predict orthogonal coiled-coils, we generated all 4,096 possible four heptad coiled-coil with asparagine or isoleucine at the a-position and glutamic acid or lysine at the e- and g-positions and scored 16.7 million interactions in an all-on-all design using the Potapov algorithm (CCNG1 Library) or our iCipa candidate algorithms (CCmax Library). Calculating orthogonality is a challenging problem that scales in exponential time with the number of possible binding partners. We used a maximal clique algorithm to identify sets of orthogonal coiled-coils where all on-target interactions have a higher score than all off-target interactions and it runs in dozens seconds on a standard laptop. Full code can be found at:

https://github.com/dancsi/DiplomaThesis

Construct and library cloning:

Each library was cloned in a similar manner, with slight differences in methods to attach a random DNA barcode to the OLS pools. After the 20bp of random DNA was attached with PCR to the 3' end of the X and Y construct (Figure S2.3), constructs were sequenced in bulk on a MiSeq to identify it and a specific X and Y (below). After barcode mapping, the T25 and T18 + GFP halves were cloned in sequentially with type IIS restriction enzymes for scarless cloning (Figure S2.4). All enzymes and polymerases came from NEB. A complete description for how each library was cloned can be found in the Supplementary Information Section 4 and oligonucleotides used for cloning are listed upon request with the author.

Mapping random barcodes:

Once random barcodes were attached and cloned, constructs were sequenced on an Illumina MiSeq to identify the X and Y proteins which each barcode was connected to. DNA containing the X and Y proteins, and the barcode were amplified as a linear fragment, and Illumina's P5 and P7 adapters attached. Constructs were sequenced with a v3 300 cycle paired end kit (Illumina TG-142-3003), with custom primers spiked into the Illumina primers. Sequences were demultiplexed, and mapped with a BBtools pipeline and consensus building custom script. Full descriptions of how each library was mapped can be found in the Supplementary Information Section 5 and scripts can be found at https://github.com/cliff-b/ortho-ccs.

Strains used:

All NGB2H experiments were run in TK310[35] carrying pSK34. TK310 is a previously published MG1655 derivative with deletions in cpdA, lacY and cyaA, which give it a large linear response

range to cAMP. pSK34 contains repressors for both the phlF and Tet promoters to maintain repression of the two-hybrid proteins. CB216 is a NEB5α derivative with pSK34 integrated genomically and only used for cloning. All plasmids used for basic cloning are listed upon request with the author and available at Addgene.

NGB2H assay execution:

Glycerol stocks of each library were thawed, and 100uL were grown up overnight in 100mL MOPS EZ Rich Defined Media (Teknova M2105) with kanamycin (Teknova K2125) and carbenicillin (Teknova C2130). For time course studies, a glycerol stock containing a library of constitutive GFP constructs was also thawed, and 100uL was inoculated into 10mL of MOPS EZ Rich Defined Media with kanamycin and carbenicillin and grown overnight. The next morning 1mL of the GFP library was added to the 100mL of library culture. After mixing GFP and experimental libraries, 1mL of overnight culture was added to a fresh culture of 100mL MOPS EZ Rich Defined Media with carbenicillin and kanamycin and the inducers for two hybrid expression: 5ng/mL anhydrotetracycline, 1.5uM 2,4-Diacylphlorolglucinol and 100uM IPTG, done twice for biological replicates, except where indicated (Supplemental Information Section 6). Flasks were placed in a 37C degree shaker for six hours. Samples were pulled after 6 hours and placed on an ice slurry to quickly cool for 15 minutes after which cells were spun down for RNA and DNA extraction.

RNA and DNA preparation for barcode sequencing

Samples of RNA were prepared with Qiagen RNeasy kits (Qiagen 74106, or 75144) according to manufacturer's instructions, with on column DNase digestion (Qiagen 79254) and concentrated

with RNeasy MinElute Cleanup kit (Qiagen 74204). RNA was reverse transcribed with Superscript IV (ThermoFisher 18090050) with a modified protocol such that 25ug of input RNA was used, the extension step ran for 1h at 55C, and 1uL of RNase A was added in the RNA removal step. Each sample was transcribed with a specific primer, often oSK193 or oSK194, that attached the i7 index and P7 sequencing primer. Samples of DNA were prepared with Qiagen Plasmid Plus Maxi kits (Qiagen 12963) according to the manufacturer's instructions. RNA samples were verified to contain very low levels of DNA (< 1:1000) by qPCR (Kapa Biotechnology KK4601) with oSK199 and oSK200, which was repeated with a high-fidelity PCR for a low number of cycles to keep samples in the exponential amplification phase. DNA samples were similarly quantified with qPCR and amplified for low cycles to attach P5 and P7 and multiplexing indices. Amplified samples were then quantified on an Agilent Tapestation 2200 with D1000 screentape (Agilent 5067-5582), verified to be monodispersed and mixed in equimolar quantities. Complete details for RNA and DNA preparation can be found in the Supplementary Information Section 7.

Barcode sequencing

Pooled RNA and DNA barcodes from each experiment were sequenced with various cores and startups at UCLA. The CC1 and CCNG1 Libraries were sequenced on a Hiseq 2500 while the CCmax and CC0 libraries were sequenced on a Nextseq 550. Samples were diluted and mixed with 5-20% phiX control v3 (Illumina FC-110-3001) and sequenced with oSK326 for read 1 and oSK324 for the index read.

Barcode Counting:

We used a custom bash script to count DNA barcodes from barcode sequencing. After demultiplexing into reads from RNA or DNA samples, reads were truncated to the 20bp containing the barcode and unique sequences counted. Barcode counts were then processed with Starcode (v1.3), to condense barcodes within a levenshtein distance of one to remove sequencing errors and tallied again. Full processing scripts can be found at https://github.com/cliff-b/ortho-ccs.

Interaction quantification

Barcode count files were imported into R where they were merged with the mapping file to provide the protein pair identified with each barcode. Barcodes corresponding to the same construct were summarized (dplyr 0.7.4) and total counts of RNA barcodes and DNA barcodes per protein pair were obtained. For our analysis we used Interactions scores calculated as $\ln(\text{median}(\frac{\text{RNA reads per barcode}}{\text{DNA reads per barcode}}))$ for barcodes that had >10 reads in all DNA samples. Interactions for all libraries are reported upon correspondence with the author.

Orthogonal set identification

Orthogonal sets were identified for the CCNG1 and CCmax libraries. Briefly, we wrote a script, maximal.py that took the Interaction scores for each set and built a graph with interactions forming the edges between proteins. Finding the maximum independent set of the line graph of this graph gave us the largest orthogonal set of interactions, are available with correspondence with the author The largest sets for different numbers of on-target interactions are listed upon request with the author. The full script for orthogonal set identification can be found at https://github.com/cliff-b/ortho-ccs.

Supplementary Information

1. System design

1.1 NGB2H system design

We decided to use the *B. pertussis* adenylate cyclase bacterial two-hybrid[17], as the starting point for our high-throughput two-hybrid to access the genetic tools available in *E. coli*. To make a quantitative, inducible, multiplexed system we made numerous changes to Karimova et al.'s two-hybrid. First, the two halves of adenylate cyclase, T18 and T25, are traditionally expressed on separate plasmids, which is incompatible with pooled library screens in *E. coli*. Second, both T18 and T25 were under control of a lacUV5 promoter which creates a positive feedback loop because it is cAMP inducible[36]. Though the authors note that a thresholding effect is useful for their work, this prohibits a quantitative PPI screen. Third, the presence of exogenous cAMP confers a growth advantage except during transformation when it is toxic (unpublished data) which could lead to highly biased libraries. To simultaneously maintain plasmid stability with both halves of adenylate cyclase on the same plasmid, remove the positive feedback loop, and ameliorate cAMP growth effects, and, we placed T18 and T25 under separate inducible promoters, pPhlF and pTet[37], respectively. Though pPhlF and pTet did not originally have the same induced expression levels, by tuning their ribosome binding sites we were able to equalize their expression. Finally, as T18 and T25 were on the same plasmid as the reporter, we were concerned about efficient termination, so all transcripts ended with two strong terminators[38].

In addition to our modifications to adenylate cyclase we built a new reporter to create a multiplexed assay with the genetic elements necessary for an Illumina-based readout. Most significantly, we replaced LacZ with sfGFP to allow efficient measurements in a standard plate reader of single interactions. This allowed for us to characterize kinetics of the NGB2H system,

quickly optimize genetic elements and test single constructs. We also changed the promoter to a variant of pLac[39] which had the greatest fold induction (Figure S2.1B, S2.1C), to increase the system's dynamic range. To enable Illumina sequencing of reporter expression we placed a Truseq adaptor, a placeholder for a DNA barcode, and reverse transcription primers in the 3' untranslated region of sfGFP. In addition to our episomal changes, we replaced the original strains with TK310, an MG1655 derivative with deletions in adenylate cyclase, lactose permease and cAMP-phosphodiesterase which gives the Lac promoter a linear response to cAMP[35], enabling direct comparison between interactions. Finally, for studies in TK310 we included a plasmid with the PhlF and Tet repressors, which we call pSK34.

2. Library composition and design

2.1 CC0 Library design

Jody Mason and colleagues used circular dichroism to characterize a set 16x16 of orthogonal coiled-coils[15] which we used for validation of the NGB2H system. To create the CC0 library, which consisted of all 256 interactions, we used a custom python script difflength-lib-gen.py with mason.fasta to specify the protein sequences. This script created 230bp oligonucleotides encoding either the X protein or Y protein (Figure S2.2A). We noticed that the chip amplification of some proteins would fail for unknown reasons with certain primers so we designed a redundant system of five subpools containing the same protein encoding DNA. The X containing oligonucleotides were designed to contain one of five sets of flanking subpool primers: oSK528-oSK532 for the forward primers and oSK609-oSK613 for the reverse primers. Immediately 3' to the forward subpool primer was a group forward subpool primer, oSK608 which was the same across subpools. Downstream of the primers was the reverse complement of

one of three codon usages for each CC0 coiled-coil, followed by a BspQI site for scarless cloning into the NGB2H system. Finally, the BspQI site was followed by a BbsI site that would allow ligation with oligonucleotides containing the Y protein and a spacer to make the oligonucleotide 230bp.

The oligonucleotides containing the Y protein were similarly designed. The Y containing oligonucleotides were flanked with one of five sets of subpool primers oSK533-oSK537 for the forward primer and oSK614-oSK618 for the reverse primers. Immediately 3' to the forward primer is a spacer to make the oligonucleotides 230bp, followed by a BbsI site to allow ligation with the X containing oligonucleotides. Downstream of the BbsI site was a BtsαI site to allow scarless cloning of the Y oligonucleotides into the NGB2H system followed by one of three codon usages for the CC0 coiled-coils. At the end of the coiled-coil protein coding sequence we placed a BsaI site to allow insertion of the T18 half of the *B. pertussis* two-hybrid. Finally, 5' adjacent to the reverse subpool primer, we placed a group subpool reverse primer oSK687. The five subpools with three codon usages of sixteen proteins in two orientations for a total of 480 oligonucleotides were ordered as an OLS pool from Agilent's high-fidelity process.

2.2 CC1 Library composition

We designed a library of twenty coiled-coils based on our own designs, which we call the CC1 Library (Figure S2.9A). Half of the CC1 Library, P# set, has previously been partially characterized[40]. We expected the P# set to be orthogonal based on Glu/Lys bonding from the E- and G-positions and Ile/Ile or Asn/Asn complementarity at the A-position. The other half of the CC1 Library are backbone variants (changes at the B, C, and F-positions) of four and six coiled-coils, the P#SC# set, and P#mS set respectively. As the B, C, and F-positions are thought to

contribute little to coiled-coil specificity, we expected the P#mS and P#SC# constructs to behave like the corresponding P# constructs.

2.3 CC1 Library design

The proteins were converted into oligonucleotides for either the X or Y half of the two-hybrid, with the accessory sequences for cloning into the NGB2H system (Figure S2.2B) with a custom script, codon_opt_hb.py. Briefly, the protein sequences were reverse translated into codon optimized DNA sequences and the sequence was checked for restriction enzyme sites used in downstream cloning steps and replaced any found.

We composed the X oligonucleotides from 5' to 3' thusly: a forward OLS amplification primer, stop codon, the reverse complement of the protein coding sequence, a BspQI site for scarless cloning, a junk spacer region, a reverse OLS amplification primer and another spacer to extend the oligonucleotides to a uniform 230bp. We used different amplification primers for different codon usages: oSK529 with oSK699, oSK700 with oSK701 or oSK702 with oSK703.

We composed the Y oligonucleotides from 5' to 3' thusly: a forward OLS amplification primer, a junk spacer region, a BtsαI site to facilitate scarless cloning, the coding sequence of the protein, a BsaI site for scarless cloning, a reverse OLS amplification primer and a spacer base to make the oligonucleotides a uniform 230bp. We used different amplification primers for different codon usages: oSK570 with oSK704, oSK705 with oSK706 and oSK708 with oSK709. The twenty one proteins, with three codon usages in two orientations were ordered as 126 oligonucleotides from Agilent as an OLS pool.

2.4 CCNG1 Library composition

The CCNG1 Library was composed of four heptad coiled-coils with either one or two Asn in the A-position with the other A-positions being Ile. E- and G-positions varied between Glu and Lys with no limits on abundance. All D-positions were Leu to improve dimerization. With these constraints, our orthogonal computational framework (below) identified fifteen potential orthogonal sets to which we added previously tested designs, the P# set from the CC1 Library and the CC0 Library. We included six different backbone compositions (Figure S2.13A)--three across the experimental sets, two more in both control sets, and one in just the CC0 Library . These backbones were intended to vary stability and helical propensity: bA-largely alanine, bN-largely glutamine and bH-glutamic acid and lysine, bS-largely serine and glutamine and bP-largely lysine and glutamic acid, as well as the published backbone on the CC0 Library. The CCNG1 Library encoded 132 on-target homodimers, 429 on-target heterodimers and 7,608 off-target interactions.

2.5 CCNG1 Library design

To create the CCNG1 Library we used a custom python script, lib-gen.py which produced 230bp oligonucleotides containing each pair of proteins specified in 170307 all-8k R8000.pairs with the protein coding sequence in 170307 all-8k R8000.fasta. This script created an oligonucleotide that started with oSK229 followed by the reverse complement of one of three codon usages of the X protein (Figure S2.2C). Downstream of the X protein was a BspQI site to allow scarless cloning of the X protein and a BtsαI site to allow scarless cloning of the Y protein. The Y protein followed the BtsαI site in one of three codon usages before terminating in a BsaI site to allow scarless cloning of the T18 half of the NGB2H assay. Finally, at the 3’ end, it contained the

reverse complement of oSK232. The script checked each protein coding region for restriction enzyme sites, remaking those that contained a reserved enzyme site, and added one nucleotide to bring the constructs to a uniform 230bp. The 24,483 oligonucleotides were ordered from Agilent as an OLS pool.

2.6 CCmax Library composition

The CCmax Library was largely composed under the same constraints as the CCNG1 Library. Library members were four heptad coiled-coils with the A-positions varying between Ile and Asn, the D-position invariantly Leu and the E- and G-positions varying between Glu and Lys. The backbones were invariantly Alanine. We also included two sets of anti-parallel coiled-coils under the analogous constraints. Our computational pipeline used fifteen different iCipa candidates as scoring functions and we took the largest orthogonal set produced by each. We also included five control sets that had been previously evaluated, including the CC0 Library and P# set from the CC1 Library. It also included two large sets that systematically varied each A-, E- and G- position though but were not predicted to be orthogonal and thus excluded from the analysis. The CCmax Library encoded 474 on-target heterodimers, 86 pairs of on-target homodimers and 17,931 off-target interactions for a total of 18,491 interactions.

2.7 CCmax Library oligo design

To create the CCmax Library we used a custom python script, lib-gen2.py, which functioned very similarly to lib-gen.py. Briefly, it took all pairs of proteins in 18491_flipped_zipped_final_pairs.pairs and encoded them into DNA from proteins sequences in 18491_flipped_zipped_final_pairs.fasta into 230bp oligonucleotides with the functional DNA

sequences needed for cloning into the NGB2H assay (Figure S2.2D). This created oligonucleotides flanked with oSK470 and oSK471 for subpool amplification. Immediately 5' to oSK470 was the reverse complement of the X protein in one of three codon usages, all of which were optimized to the frequency which they naturally occur in E. coli. To facilitate scarless cloning the X protein was followed by a BspQI site. Likewise, for scarless cloning of the Y protein we placed a BtsαI site immediately downstream of the BspQI site and N-terminal to Y protein coding sequence. Finally, we added a BsaI site C-terminal to the Y protein coding sequence. We repeated this process three times to create three different codon usages and then ordered the full set of 55,473 oligonucleotides as an OLS pool from Agilent using their high-fidelity process.

2.8 GFP Library design

The GFP Library was designed as a library of barcodes that remain constant despite difference in library composition, similar to ERCC for RNA-seq[41]. We created a library of constitutive GFP constructs spanning several orders of magnitude in expression, that contained a unique DNA barcode and the flanking sequences for amplification and sequencing with our other libraries. We used previously published expression data[42] from which we selected constitutive promoters that would give a wide range in expression. In order from low to high expression we chose J23117, pAPFAB277, pAPFAB69, pAPFAB48, pAPFAB70 and pAPFAB101. Each of these promoters was attached to sfGFP and in the 3' UTR of the sfGFP we inserted our Truseq adaptor and a 7bp random barcode.

3. Computational framework and model refinement

3.1 Orthogonal computational framework

We built a computational framework to calculate predicted interactions for four heptad coiled-coils, for which all programs can be found at https://github.com/dancsi/DiplomaThesis. To limit the search space, we examined a subset of four heptad coiled coils where the A-positions varied between Ile and Asn and the E- and G-positions varied between Lys and Glu, for a total of 4096 possible sequences (2*2*2)^4. Using the algorithm described in Potapov et al.,[13] we scored all 16.7 million potential interactions. To make this computationally tractable we reimplemented the algorithm in the C for a 1000x fold increase in speed in fastscore.exe. After scoring interactions we sought to identify sets of orthogonal interactions, those where all on-target interactions had higher interaction scores than all off-target interactions. The identification of orthogonal interactions can be reduced to the maximum clique problem[18] which we implemented in Solver.exe. Solver.exe uses score cutoffs for on-target and off-target interactions to enforce orthogonality. Based on experimental data of the P# and CC0 sets we searched for sets with a difference of 1.0 (arbitrary units) between the scoring cutoffs. Finally, to obtain the multiple sets of the CCNG1 Library we varied the threshold for non-interacting pairs from -7.5 to -9 in increments of 0.05.

3.2 Model refinement from CCNG1 Library

We used the CC0 Library subset in the CCNG1 Library as a calibration curve to convert interaction scores to Tms to be comparable to other coiled-coil interaction prediction algorithms (Fig S11B). We found that the prediction power of the algorithm increases significantly if we allow coiled-coils to bind in the heptad alignment that maximizes the ΔG. In calculating the off-

target pairs, most algorithms assume the coiled-coil interacts with all four heptads but this does not necessarily reflect reality as a three heptad alignments may be more energetically favorable (Figure S2.14). To develop an algorithm that incorporates heptad shifts, we first extracted the features of all possible interactions without permitting shifts. Using this model we have scored three heptad shifts at positions -7, 0, and +7. We then changed the alignment to the lowest scoring position and retrained the model. After five iterations the model parameters converged on a fixed point.

We fit several groups of parameters in a series of models. First we incorporated features known to be important to coiled-coil interaction by counting the number of aligned residues between pairs of coiled-coils that were Asn/Asn, Asn/Ile, or Ile/Ile at the A-position, and Glu/Glu, Glu/Lys, or Lys/Lys between E- and G-positions. We also included a term for total charge between the two coiled-coils. As heptad shifting means fewer Leu in the D-position are interacting, we also scored all parameter groups with and without a term for the D-position. We then tested two separate parameter groups, one that looked at the effect of consecutive A-positions and one that looked at the most N-terminal A-position. Our consecutive A-position parameter group compared Ile/Ile, Asn/Asn, Asn/Ile and Ile/Asn in consecutive heptads, while our N-terminal parameter group looked at Ile/Ile, Asn/Asn and Asn/Ile pairs at the first heptad. Of note, the first heptad was scored again in the central suite. All models were fit using the Ridge regression algorithm of the sklearn library. Interactions were weighted by the number of different barcode counts, and on-target pairs upweighted 10 fold. We found that our term for repulsion to be beneficial to our core model, but scoring Leu interactions did not matter. Our consecutive A-position parameter group did not improve the core model's predictive capabilities but the N-

terminal group significantly did (Figure 2.3B). Thus our best model, which we call iCipa, is the core model with N-terminal A-position parameters (Figure 2.3D).

Our model comes with several limitations. In order to limit the design space, the model only uses Glu/Lys at the E- and G-positions and Asn/Ile at the A positions while the D-position is expected to be a Leu and B-, C- and F-positions are expected to be Ala. As the backbone is expected to be Ala, the helical propensity of all proteins is very expected to be very high. Because of this, our model does not include a term for helical propensity as it was not predictive of interaction strength.

4. Library Cloning

For all cloning steps the reagents purchased from the following vendors unless otherwise noted. All reactions were performed according to the manufacturer's instructions unless otherwise noted. All restriction enzymes, phosphatase and ligase were purchased from NEB and DNA polymerase was NEBNext Q5 Hotstart HiFi PCR Master Mix (NEB M0543L) unless otherwise noted. All qPCR was performed using KAPA SYBR Fast 2x Master Mix (Kapa Biotechnology KK4601). All nucleic acid prep was purchased from Qiagen (Qiagen Plasmid Plus Maxi Kit 12963), though DNA cleanup and gel extraction kits were from Zymo (Zymo D4014 and Zymo D4008).

4.1 CC0 Library cloning

The 10pM high-fidelity OLS pool containing the lyophilized CC0 Library was resuspended in 25uL EB. This was diluted 1:20 in ddH2O and used as template for qPCR using the subpool primers oSK528-oSK532 and oSK609-oSK613 for X oligonucleotides and oSK533-oSK537

and oSK614-oSK618 for Y oligonucleotides. All subpools exhibited exponential amplification through 20 cycles, so high-fidelity PCR was performed in triplicate for 18 cycles. Replicates were pooled and digested at 1ug scale with BbsI-HF for two hours. Digested samples were run on a 4% agarose gel and the band containing the X or Y-protein was extracted. The entire extracted product was ligated with its subpool partner (ie X subpool-1 with Y subpool-1) with T4 ligase overnight at 50ng scale. Ligated samples were cleaned up and qPCR with the group pool primers, oSK608 and oSK687 established exponential amplification through eight cycles, with the exception of group pool one which showed no amplification and was excluded from downstream steps. High-fidelity PCR was performed for six cycles and cleaned up. Each group pool's concentration was analyzed with an Agilent Tapestation 2200 (Agilent 5067-5582) and equimolar fractions were pooled and diluted 100 fold. qPCR of the group pool samples with primers oSK689 to attach an AscI site and oSK690 to attach a BsaI site, random 20bp barcode, and EcoRI site, showed exponential amplification through ten cycles. High-fidelity PCR was performed in sextuplicat of which three were pooled to make the primary barcoding and three were pooled to make a replicate barcoding (Figure 2.1F). The barcoded products were run on a 3% agarose gel, extracted and digested at 2ug scale with AscI and EcoRI-HF for two hours. Freshly prepared pSK33 was likewise digested with AscI, EcoRI-HF and rSAP for two hours, run on a 1% agarose gel and extracted. Digested plasmid and barcoded product were ligated with T4 ligase for two hours at 250ng scale. Sample was cleaned up into 6uL ddH2O, and 1uL electroporated into NEB 5-alpha (NEB C2989K). After 35 minutes recovery in SOC samples were plated on LB agar + Kanamycin and grown overnight at 37C. Of the 2.4 million transformants, ~40,000 were scraped off the overnight plates and grown up in 150mL LB + Kanamycin at 30C overnight which was then purified and used for cloning the T25 segment.

The T25 insert was cloned as follows. The T25 segment was amplified from pSK59 with oSK694 and oSK695. PCR product was run on a 1% agarose gel and extracted. The insert was digested with BspQI for two hours at 4ug scale. Product was cleaned up and digested with BtsɑI for two hours. Plasmid from the above cloning step, for both barcodings, was digested with BspQI for three hours at 5ug scale, cleaned up and digested with BtsɑI overnight. Plasmid was then cooled to 37C and rSAP was added for 30 minutes. Dephosphorylated plasmid was run on a 1% agarose gel and extracted. The digested plasmids and T25 were ligated with T4 ligase at 250ng scale for six hours before transformation into freshly prepared electrocompetent CB216. Cells were recovered in SOC for 45 minutes and plated on LB agar + Kanamycin + Carbenicillin. The transformation was repeated three separate times for a total of ~100,000 colonies. These were scraped from the plates, diluted to OD 0.02 and grown up in 150mL LB + Kanamycin + Carbenicillin at 30C overnight which was then purified and used for cloning the T18 segment.

The T18 insert was cloned as follows. The T18 segment was amplified from pSK59 with oSK698 and oSK202. PCR product was run on a 1% agarose gel and a band corresponding to the expected 1715bp was extracted and digested with BsaI-HFv2 at 3ug scale for two hours. The vectors containing both barcodings were digested with BsaI-HFv2 and rSAP for two hours at 4ug scale. Digested plasmids were run on a 1% agarose gel and extracted. Plasmids and the T18 insert were ligated at 200ng scale with T4 ligase overnight. Ligation products were cleaned up and electroporated into freshly prepared electrocompetent CB216. Cells were recovered in SOC for 35 minutes and plated on LB agar + Kanamycin + Carbenicillin. Approximately 180,000 colonies were obtained for each barcoding. These were scraped from the plates diluted to OD 0.02 and grown up in 150mL LB + Kanamycin + Carbenicillin at 30C overnight. DNA was

extracted, run on a 1% agarose gel, extracted, and electroporated at low concentration into previously frozen electrocompetent pSK34. Samples were recovered for 35 minutes in SOC at 37C and plated onto LB agar + Kanamycin + Carbencillin. Cells were scraped off the plates, diluted to OD 0.02 and grown up in 150mL LB + Kanamycin + Carbenicillin at 30C overnight. Glycerol stocks of overnight culture were prepared and stored at -80C. For downstream experiments one tube was fully thawed and subsequently discarded.

4.2 CC1 Library cloning

The 10pM OLS pool containing the CC1 Library was resuspended in 30uL EB. Each respective subpool—the X oligonucleotides with codon usages 1, 2 or 3 or the Y oligonucleotides with codon usages 1, 2, or 3—were amplified with their flanking primers with KAPA Real-time Library Amplification (KAPA Biosystems KK2702) via qPCR (oSK529 and oSK699, oSK700 and oSK701, oSK702 and oSK703 for X oligonucleotides and oSK570 and oSK704, oSK705 and oSK706, oSK708 and oSK709 for Y oligonucleotides). Using a 1:10 dilution of the OLS pool in ddH2O we found that the pools all exhibited exponential amplification through twenty-five cycles so amplification was repeated for twenty cycles. The products were cleaned up and qPCR was performed again with primers to attach a 30bp annealing region for the X oligonucleotides or a 30bp annealing region and the reverse primer from the first amplification for the Y oligonucleotides. This was done first with qPCR with KAPA Real-time Library Amplification which suggested exponential amplification through 10 cycles. It was then repeated with KAPA HiFI HotStart Ready for 10 cycles in quadruplicate, which were then purified and pooled. We then prepared a primerless PCR with mixtures of X and Y oligonucleotides of different codon usages. 100ng of X and Y were mixed into a 50uL KAPA HiFi HotStart Ready

Mix reaction, with ten cycles of denaturation, annealing of the 30bp complementary regions, and extension. These samples were run on a 2.5% agarose gel and bands corresponding to the expected length of 400bp were extracted. Mixed samples were then amplified with biotinylated primers, oSK712, oSK713, or oSK714 and oSK715, oSK716 or oSK717 to attach restriction enzyme sites and barcodes using qPCR. Samples showed exponential amplification through fifteen cycles. The PCR was then repeated with KAPA HiFi HotStart Ready Mix for twelve cycles after which they were cleaned up and digested with AscI and EcoRI-HF at 0.4ug scale for two hours. Samples were heat inactivated and cleaned up with Dynabeads (ThermoFisher 65306) to remove undigested product. Likewise, backbone pSK33 was purified, PCR amplified with biotinylated primers oSK718 and oSK719 (KAPA Biosystems KK2602), digested with AscI and EcoRI-HF at 10ug scale for two hours and cleaned up with Dynabeads. The barcoded CC1 X and Y proteins were ligated into the digested vector with T4 Ligase at 100ng scale for an hour. Samples were cleaned up and electroporated into freshly prepared electrocompetent pSK34. After recovery in SOC media for an hour, cells were plated on LB agar + Kanamycin + Carbenicillin in serial 10-fold dilutions. The next day colonies were counted and the transformation was repeated while taking only 441,000 transformants (100x library coverage) from the SOC media. This was inoculated into 400mL of LB +Carbenicillin + Kanamycin at 30C overnight and DNA extracted for cloning the T25 insert.

The T25 insert was cloned into the barcoded X and Y proteins. It was amplified from pSK59 with biotinylated primers oSK203 and oSK204 with KAPA HiFi HotStart Ready Mix. 1.4ug was digested with BspQI and BtsαI in NEB Cutsmart buffer at 55C for one hour. Undigested product was removed with Dynabeads, and purified again. 30ug of barcoded plasmid product was digested with 10uL of BspQI and 10uL BtsαI in 200uL NEB Cutsmart for an hour.

Product was run on a 1% agarose gel and the corresponding band extracted before dephosporylation with rSAP at 60uL scale for 30 minutes. 400ng of vector was ligated with T4 ligase for one hour, cleaned up, drop dialyzed and electroporated into pSK34 in two separate reactions. Colony PCR showed ~230k transformants, which were grown up in 200mL LB + Kanamycin + Carbenicillin overnight and plasmid was extracted for cloning in the T18 insert.

The T18 insert was amplified from pSK59 with biotinylated primers oSK201 and oSK202 with KAPA HiFi HotStart Ready Mix. The sample was gel extracted and 4ug were digested with BsaI-HF. The plasmid described above was digested with BsaI-HF for an hour. BsaI was heat inactivated and the digested plasmid was gel extracted before being treated with rSAP for half an hour. Dephoporylated plasmid was cleaned up and 200ng were ligated to the T18 insert at 1:3 ratio with T4 ligase for an hour. Sample was cleaned up and electroporated into freshly prepared electrocompetent pSK34. Approximately 5 million colonies were obtained which were grown up in 200mL LB + Kanamycin + Carbenicillin overnight at 30C. Glycerol stocks were made and stored at -80C. For downstream experiments one tube was fully thawed and subsequently discarded.

4.3 CCNG1 Library cloning

The CCNG1 Library was ordered as part of a 10pM OLS pool and resuspended in 25uL EB. As each interacting pair fit on one oligonucleotide early cloning steps were significantly simplified. The oligonucleotides from the pool were amplified in bulk with oSK229 and oSK232 for all three codon usages. qPCR showed exponential amplification through 16 cycles so a high fidelity PCR was then amplified for 14 cycles in quadruplicate. Samples were pooled and a band of 230bp was run on a 3.5% agarose gel and extracted. Purified sample was diluted 100x and re-

amplified for eight cycles in quadruplicate to attach an AscI site (oSK358) and a second BsaI site, 20bp random barcode, and an EcoRI site (oSK359). PCR product was again pooled, run on a 3.5% agarose gel, and extracted for digest with AscI and EcoRI-HF for two hours. Low-copy plasmid pSK33 was purified from 200mL of LB + Kan and 5ug was digested with AscI and EcoRI-HF and rSAP for two hours. The plasmid digest was then run on a 1% agarose gel and the linearized fragment was extracted. 250ng of the linearized plasmid and 1:3 ratio of barcoded OLS product were ligated with T7 ligase for three hours. The ligation product was cleaned up and eluted in 6uL of ddH2O. Electrocompetent NEB 5-alpha cells (NEB C2989K) were transformed with 1uL of the ligation product, cells were recovered for 35 minutes in SOC and plated on large LB + Kan plates. Two million colonies were obtained, and colony PCR showed 23/24 containing the insert. 1.2 million colonies were scraped, diluted to OD 0.02 and grown up in 200mL of LB + Kan for DNA purification.

The T25 insert was cloned as follows: it was amplified from pSK59 with oSK360 and oSK361. The sample was gel purified and digested at 3ug scale sequentially with BspQI and BtsαI for four hours each. 5ug of the plasmid from the previous step was digested with BspQI and BtsαI, for five hours and overnight, respectively, before a 30 minute dephosphorylation with rSAP. The previously purified plasmid and the T25 segment were ligated with T7 ligase for four hours and cleaned up before transformation into freshly prepared electrocompetent pSK34. After electroporation, cells were recovered for 35 minutes and plated onto large LB + Kan + Carb plates. Colonies were scraped, diluted to OD 0.02 in 200mL of LB + Kan + Carb and grown to saturation and plasmid was purified.

The T18-sfGFP insert was cloned as follows: it was amplified from pSK59 with oSK201 and oSK202. The sample was run on a 1% agarose gel and purified and 4ug of it were digested

with BsaI-HF for two hours. Likewise, the previously purified plasmid was digested BsaI-HF at 5ug scale for two hours with rSAP, run on a 1% agarose gel and gel extracted. 250ng of plasmid was ligated with the T18-sfGFP insert at a 3:1 ratio with T7 ligase for two hours at room temperature. The sample was cleaned up and electroporated into freshly prepared pSK34. Cells were recovered in SOC for 35 minutes at 37C and plated onto large LB + Kanamycin + Carbenicillin plates. Plates were grown overnight, and all two million cells were scraped diluted to OD 0.02 and grown overnight in LB + Kanamycin + Carbenicillin. Glycerol stocks were made from overnight culture and frozen at -80C. For downstream experiments one tube was fully thawed and subsequently discarded.

4.4 CCmax Library cloning

The CCmax Library was ordered as a 10pM OLS pool and resuspended in 20uL of EB. qPCR with oligonucleotides oSK470 and oSK471 showed exponential exponential amplification through ten cycles, so a high-fidelity PCR was repeated for eight cycles. The amplified product was cleaned up and diluted. Reamplification with qPCR using oSK472 to attach an AscI site and oSK473 to attach a BsaI site, the random DNA barcode and EcoRI site showed exponential amplification through twelve cycles so we performed a high-fidelity PCR for eight cycles with in triplicate. Samples were pooled and run on a 3% agarose gel. A band corresponding to the expected 290bp was extracted, and digested with AscI and EcoRI-HF for two hours at 1ug scale. pSK33 was grown up in 200mL of LB + Kanamycin, purified, and digested at 3ug scale with AscI, EcoRI-HF and rSAP. Digested product was run on a 1% agarose gel and extracted. 250ng of digested pSK33 and a 3:1 ratio of the insert were ligated with T7 ligase for 3 hours. The sample was cleaned up, and 1uL was electroporated into NEB 5-alpha cells (NEB C2989K).

Cells were recovered for 35 minutes in SOC media and plated on to LB agar + Kanamycin. Approximately four million transformants were obtained of which 1.2 million were scraped from the plate, and diluted to OD 0.02 in 150mL of LB + Kanamycin and grown up overnight at 37C and DNA was purified.

The T25 insert was cloned as follows: it was amplified from pSK179 with oSK474 and oSK475. The sample was run on a 1.5% agarose gel and then was digested at 4ug scale sequentially with BspQI and BtsαI for four hours each. 5ug of the previously purified plasmid was also digested with BspQI and BtsαI, for three hours and five, respectively, before a 30 minute dephosphorylation with rSAP. The digested plasmid and the T25 insert were ligated with T4 ligase for four hours and cleaned up before transformation into freshly prepared electrocompetent pSK34. After electroporation, cells were recovered for 35 minutes and plated onto LB agar + Kanamycin + Carbenicillin plates. After three successive transformations 500,000 transformants were obtained all of which were scraped from the plates, diluted to OD 0.02, grown up in 200 mL of LB + Kanamycin + Carbenicillin overnight at 37C, and DNA was purified.

The T18 insert was cloned as follows: it was amplified from pSK168 with oSK476 and oSK477. The sample was run on 1% agarose gel, extracted and digested at 3ug scale with BsaI-HF. The previously purified plasmid was digested with BsaI-HF and rSAP for four hours at 5ug scale. Digested product was run on a 1% agarose gel and extracted. 250ng of vector was ligated at 3:1 ratio with the T18 insert with T4 ligase for two hours. The ligation product was purified and electroporated into freshly prepared electrocompetent pSK34. Cells were recovered for 35 minutes in SOC at 37C and plated on LB agar + Kanamycin + Carbenicillin. Plates were scraped diluted to OD 0.02 and grown in 200mL LB + Kanamycin + Carbenicillin to saturation and

glycerol stocks were stored at -80C. For downstream experiments one tube was fully thawed and subsequently discarded.

4.5 GFP Library cloning

Each design was synthesized de novo from oligonucleotides ordered from IDT. Ribosome binding sites were synthesized in oSK218-222 and barcodes were attached from oSK217. After cloning into backbone pSK33, we chose 10 colonies from each ribosome binding site, and sequenced their barcodes. Sequenced colonies were inoculated into a deep well plate of LB + Kan + Carb, grown to saturation, pooled and frozen in glycerol stocks at -80C. For downstream experiments one tube was fully thawed and subsequently discarded.

4.6 Individual construct cloning

DNA for the NGB2H system was ordered from Addgene or DNA 2.0. For the bacterial two-hybrid, we used pEB1029 and pEB1030 (Addgene 22066 and 22067)[43]. The pSC101 origin was drawn from pZS-123 (Addgene 26598)[44]. Most of our functional features were drawn from the work of the Voigt lab and iGEM. Inducible promoters for both pTet and pPhlF[37], strong terminators[38], and ribozyme elements upstream of our open reading frames[45] were all synthesized *de novo*. Linkers between *B. Pertussis* proteins and the proteins of interest were drawn from the original B2H with the T18 linker having an GTG extension in the center[17]. sfGFP was drawn from a previously published strain[42]. DNA for all subsequent plasmids was assembled with standard molecular biology techniques, namely Gibson assembly[46] and restriction enzyme digestion/ligation. Individual constructs were sequenced verified before experimental use.

5. Library mapping:

All mapping steps used KAPA SYBR Fast qPCR (Kapa Biosciences KK4601), NEBNext Q5 Hotstart HiFi PCR Master Mix (NEB M0543L) for high-fidelity PCR, an Agilent Tapestation with D5000 ScreenTape (Agilent 5067-5583) for DNA quantification, PhiX sequencing control v3 (Illumina FC-110-3001) as a control library and an Illumina Miseq with a v3 600 cycle paired end kit (Illumina TG-142-3003) for sequencing unless otherwise noted.

5.1 CC0 Library mapping

After cloning the CC0 Library proteins and barcodes into pSK33 we sequenced the barcode through the proteins on a MiSeq to provide a mapping function between the two. The amplicon containing the CC0 Library proteins and barcode was amplified with oSK696 and oSK193 or oSK194 for the different replicate barcodings. This attached P5, P7 and a Nextera lowplex index to allow demultiplexing of the barcoding replicates. qPCR showed exponential amplification for 15 cycles for both samples, so a high-fidelity PCR was repeated for 12 cycles in triplicate. Samples were pooled and a band corresponding to the expected size of 391bp was extracted from a 2% agarose gel. The sample concentrations were measured on an Agilent Tapestation 2200 and found to be monodispersed with a length of 405bp. The samples were mixed in equimolar amounts with 15% PhiX control, diluted to 14pM and loaded into the Illumina Miseq kit. We used three separate custom primers, oSK696 for the forward read, oSK323 for the reverse read and oSK324 for the index read. We obtained 26,359,427 reads which mapped to

137,213 unique barcodes with a correct X or Y protein for the first barcoding replicate and 96,129 unique barcodes with a correct X or Y protein in the second barcoding replicate.

5.2 CC1 Library mapping

The CC1 Library mapping step was performed similar to the CC0 Library mapping. After cloning the CC1 Library and random barcode in pSK33, we amplified constructs with either oSK691, oSK692 or oSK693 and oSK193 to attach Illumina adaptor P5 and a Nextera lowplex I7 and Illumina adaptor P7 respectively. qPCR showed exponential amplification through 14 cycles. PCR was repeated with high-fidelity polymerase, KAPA HiFi HotStart ReadyMix (KAPA KK2602) and the samples were cleaned up and pooled. Concentration of the pooled sample was measured with KAPA Library Quantification Kit Illumina Systems (Kapa Biosystems KK4824) and found to be 26nM. This was diluted to 12pM, mixed with 5% phiX control, and run on a Miseq with a pool of oSK709, oSK710 and oSK711 for the forward read, oSK324 for the index read and oSK323 for the reverse read. 8,060,843 reads passed filter which mapped 1,166,860 unique barcodes mapping to one or both proteins from the CC1 Library.

5.3 CCNG1 Library mapping

Similar to the CC0 Library mapping, after cloning the barcoded CCNG1 Library into pSK33 we sequenced the barcode through the proteins with a MiSeq. The amplicon containing the CCNG1 Library was amplified using oligonucleotides oSK366 and oSK193 to attach P5, Nextera lowplex index N702 and P7. qPCR showed exponential amplification through eleven cycles so samples were amplified for high-fidelity PCR for nine cycles in triplicate. Samples were pooled and a band corresponding to the expected size of 376bp was gel extracted from a 1.5%

agarose gel. The sample concentration was quantified on an Agilent Tapestation 2200 and found to be monodispersed and of approximately 376bp. The sample was diluted to 20pM and mixed with 10% PhiX control, loaded into the Miseq kit, and sequenced with custom primers oSK367 for the forward read, oSK323 for the reverse read and oSK324 for the index read. 27,255,659 reads passed filter which mapped 1,121,668 unique barcodes to one or both proteins from the CCNG1 Library.

5.4 CCmax Library mapping

Similar to the CC0 Library mapping, after cloning the CCmax Library into pSK33, sequenced through the proteins and the barcode with a Miseq. The amplicon containing the CCmax proteins and barcode was amplified with oSK513 and oSK193. This was first done with qPCR which showed exponential amplification for 1ng through 15 cycles. High-fidelity PCR was repeated in triplicate for nine cycles, samples were pooled and run on a 1.5% agarose gel and a band corresponding to the expected 380bp was extracted. Sample concentration was quantified with an Agilent Tapestation 2200 and was found to be monodispersed and approximately 384bp. The sample was mixed with 10% PhiX control and 18pM was loaded into an Miseq kit. To the MiSeq Kit we added custom primers oSK514 for the forward read, oSK323 for the reverse read and oSK324 for the index read. 20,710,707, reads passed filter which mapped to 1,629,936 unique barcodes mapping to one or both proteins in the CCmax Library.

5.5 Mapping script

To map DNA barcodes to a unique interaction we used a custom Makefile that chained several programs from the BBTools suite[47] with our own custom script. From the raw paired fastq files

we used BBduk (v38.32) to trim adapter sequences and remove contaminants. Forward and reverse reads were then merged with BBmerge (v38.32) with strictness set to maxloose. Merged reads with levenshtein distance three or less were then condensed with Starcode (v1.3) to remove sequencing errors. From the condensed reads, barcodes were mapped with a custom python script. Briefly, this script called the last twenty bases per sequence the barcode and discarded those barcodes within levenshtein distance one of each other. It then removed those barcodes that had proteins that were more than 5% different from each other, under the assumption that these were contaminants. The variants were then mapped to the DNA corresponding to the protein coding regions using BBMap (v38.32) of the X and Y proteins sequentially. Sequencing data that mapped with no errors to a reference sequence of X or Y were then joined together by barcode and this text file was used to identify barcodes from barcode sequencing steps.

6. NGB2H experimental conditions and validation

6.1 CC0 Library NGB2H assay

Glycerol stocks of both barcodings of the CC0 Library were thawed, and 100uL were grown up overnight in 100mL EZ Rich Defined Media (Teknova M2105) with Kanamycin (Teknova K2125) and Carbenicillin (Teknova C2130). After overnight growth 1mL of the GFP Library was added to the 100mL of CC0 Library culture and 1mL of the GFP Library was added to a fresh culture of 100mL EZ Rich Defined Medium + Carbenicillin + Kanamycin with 25ng/mL anhydrotetracycline (aTC), 1uM 2,4-Diacylphlorolglucinol (DAPG) and 100uM Isopropyl B-D-1-thiogalactopyranoside (IPTG) in biological replicates for each barcoding. Flasks were incubated in a 37C shaker for six hours. Samples were pulled at 6h and placed on an ice slurry for 15 minutes, spun down for nucleic acid extraction and flash frozen. As reported in the main

text we obtained high quality verification of the CC0 Library from its internal controls. Briefly, we obtained 57,541 barcodes providing quantitative measurements of interaction strength for all 256 protein pairs. The assay was highly replicable with biological replicates having similar Interaction Scores (Pearson's $r > 0.98$, $p < 10^{-15}$) (Figure 2.1C). Different codon usages showed consistent Interaction scores with all usages correlating with Pearson's $r > 0.92$ and $p < 10^{-15}$ (Figure S2.5). The CC0 Library has a strong correlation between the primary and reciprocal orientations (Pearson's $r = 0.92$, $p < 10^{-15}$) (Figure 2.1E). The Interaction Scores for constructs with indels was much less than full length perfect constructs (Figure S2.6). When compared to the published Tms of the CC0 Library the Interaction Scores correlated with Pearson's $r > 0.73$, $p < 10^{-15}$ (Figure S2.7). Finally, when the assay is replicated with an independent re-barcoding and re-cloning of the library, we found very strong replication with the previous CC0 Library's Interaction Scores (Pearson's $r > 0.98$, $p < 10^{-15}$) (Figure 2.1F).

6.2 CC1 Library NGB2H assay

Glycerol stocks were thawed and we grew the CC1 Library overnight in EZ Rich Defined Medium + Kanamycin + Carbenicillin. We mixed 1:100 of the GFP Library with the CC1 Library and diluted it to OD 0.001 in a fresh EZ Rich media + antibiotics with 5ng aTC and 5uM DAPG at 30C to induce the library. The library was grown in a time-course experiment, where we took samples of RNA and DNA at 0h, 0.5h, 1h, 2h and 4h which were spun down and flash frozen for nucleic acid preparation. In total we linked 385,078 different barcodes 400 different PPIs. Our internal controls validated that the library performed as expected. After normalization to a constitutive GFP Library, Interaction Scores were minimal at the beginning of the assay, but they increased monotonically over four hours of induction (Figure S2.8A). In addition,

comparison of six different codon usages in the CC1 Library showed high replicability ($r > 0.89$, $p < 10^{-15}$, Figure S2.8B). Finally, compared to full length constructs, indels had a markedly reduced Interaction Score (Figure S2.8C) and the reciprocal orientation of each protein were similar ($r > 0.85$, $p < 10^{-15}$, Figure S2.8D).

6.3 CCNG1 Library NGB2H assay

After overnight growth in EZ Rich media + Carbenicillin + Kanamycin , we mixed in a constitutive GFP Library to the CCNG1 Library at a 1:100 ratio and took RNA and DNA at 0h. We induced a 1:100 dilution of the library for 6h with 25ng/mL aTC, 15uM DAPG and 100uM IPTG in EZ Rich media + antibiotics at 37C in biological replicates. We obtained 76 million reads across 164,778 barcodes which gave quality data on 8,073 interactions. Our internal controls for the CCNG1 Library behaved as expected. The Interaction Score of constructs in the CCNG1 Library strongly correlated between biological replicates (Pearson's $r > 0.95$, $p < 10^{-15}$, Figure S2.10A); the CC0 Library subset of the CCNG1 Library correlated strongly with published Tms (Pearson's $r > 0.79$, $p < 10^{-15}$, Figure S2.10B) and with our previous experiments of the CC0 Library (Pearson's $r > 0.94$, $p < 10^{-15}$, Figure S2.16). Furthermore, different codon usages for the same construct gave similar Interaction Scores (Pearson's $r > 0.75$, $p < 10^{-15}$, Figure S2.10C), and indels had significantly Interaction scores than full length constructs (Figure S2.10D).

6.4 CCmax Library NGB2H assay

After overnight growth in EZ Rich media + Kanamycin + Carbenicillin, we mixed the CCmax Library with the GFP Library in 100:1 ratio. We diluted the library by a 1:100 ratio in fresh EZ

Rich media with 15ng/mL aTC, 5uM DAPG and 100uM IPTG in biological replicates for 6h at 37C. We obtained 346,733 barcodes mapping to 17,731 constructs and collected high quality data on 17,320 interactions. We found that biological replicates strongly correlated (Pearson's $r = 0.973$, $p < 10^{-15}$, Figure S2.15A), as did different codon usages with (Pearson's $r > 0.92$, $p < 10^{-15}$, Figure S2.15B). As expected full length perfect constructs had higher Interaction Scores than those containing indels in the X or Y protein (Figure S2.15D). We again included the CC0 Library found that the published Tms correlated with our Interaction Score with (Pearson's $r = 0.876$, $p < 10^{-15}$, Figure S2.15C), and that correlated well with the CC0 proteins in other libraries (Pearson's $r > 0.84$, $p < 10^{-15}$, Figure S2.16). Finally, the reciprocal orientation of the proteins in the CCmax Library largely agrees with their primary one, with (Pearson's $r = 0.835$, $p < 10^{-15}$, Figure S2.15E).

7. Barcode sequencing preparation

All nucleic acid preparation for barcode sequencing was done with Qiagen kits. For RNA prep we used RNeasy Midi (Qiagen 74106) or RNeasy Minipreps (Qiagen 75144) with on column DNase digestion (Qiagen 79254) and concentrated with RNeasy MinElute Cleanup kit (Qiagen 74204). DNA was prepped with QIAprep spin Minipreps (Qiagen 27106) or Plasmid Plus Maxiprep (Qiagen 12963), though DNA cleanup and gel extraction was performed with Zymo kits (Zymo D4014 and Zymo D4008). As for previous steps, we used NEBNext Q5 Hotstart HiFi PCR Master Mix (NEB M0543L) for high-fidelity PCR and KAPA SYBR Fast 2x Master Mix (Kapa Biotechnology KK4601) for qPCR. All samples were quantified with Agilent D5000 Screentape (Agilent 5067-5582).

7.1 RNA Barcode Sequencing Preparation

Cell pellets containing either 1 x 10^10 cells or 1 x 10^9 cells were thawed and purified with midi or mini scale respectively with on column DNase digestion. RNA was concentrated and purified RNA was subject to primer specific reverse transcription (oSK193-oSK198 and oSK210-oSK215) to create cDNA of transcripts containing DNA barcodes. We used the reverse transcription step to attach Nextera lowplex I7 indexes and P7 Illumina sequence adaptors to our barcodes which allowed multiplexed sequencing of different times and conditions. Reverse transcription was performed with Superscript IV (ThermoFisher 18090050) with the following changes to the manufacturer's protocol. Instead of 5ug of RNA we used 22.5ug of RNA, concentrated to 11uL, the reverse transcription step at 55C was allowed to go for an hour rather than fifteen minutes, and 1uL RNAase A (Qiagen 19101) was spiked in with RNase H for twenty minutes. After reverse transcription samples were amplified with qPCR with oligonucleotides oSK199 and oSK200 to attach Illumina sequencing adapter P5. We compared the cDNAs with no-RT controls to check for DNA contamination in the RNA which invariably showed less than 1:1000 ratio of DNA to RNA. The qPCR showed exponential amplification through 20 cycles. We then repeated the PCR in triplicate for 12-15 cycles. Replicate PCRs were pooled, run on a 3% agarose gel and extracted. DNA concentration was measured on an Agilent Tapestation 2200 with D1000 screentape (Agilent 5067-5582), and equimolar fractions pooled with DNA barcodes.

7.2 DNA Barcode Sequencing Preparation

Cell pellets containing either 10mL or 50mL of culture were thawed and plasmid DNA was extracted with mini or maxi scale respectively. Barcodes were amplified with qPCR using

oSK199 and oSK193-198 or oSK210-215 to attach Illumina sequencing adaptor P5, and a Nextera lowplex I7 index and Illumina sequencing adaptor P7. We used qPCR to guide exponential amplification of the barcodes which normally showed exponential amplification through 13-16 cycles. We repeated the PCRs for 10-12 cycles in triplicate. Replicate PCR samples were pooled, run on a 3% agarose gel, and extracted. DNA concentration was measured on an Agilent Tapestation 2200 and equimolar fractions were pooled with the RNA barcodes.

8. NGB2H small scale results:

8.1 Plate reader measurements:

Strains used in plate reader assays were grown up overnight in MOPS EZ Rich Defined Media (Teknova M2105) with kanamycin (Teknova K2125) and carbenicillin (Teknova C2130) in a 37C degree shaker. The next evening these cultures were diluted 1:100 in 100uL fresh MOPS EZ Rich Defined Media with kanamycin, carbenicillin and 100uM IPTG and the indicated inducers in a 96 well, flat bottom plate (Corning 0720090). The plate was then incubated in a Tecan M1000 Plate Reader at 37C overnight. Optical Density (OD600) and GFP fluorescence (excitation 488nm, emission 508nm) were taken every half an hour after 3 minutes of 1mm orbital shaking. Data was collected for a minimum of fourteen hours but normally reached saturation by eight hours.

8.2 Single construct optimization and benchmarking

We found the NGB2H system behaved as expected with sfGFP fluorescence dependent on a pair of interacting proteins and both anhydrotetracycline (aTC) to induce pTet and 1,4-Diacylphloroglucinol (DAPG) to induce pPhlF (Figure S2.1A). Lacking either inducer or

assaying a pair of non-interacting proteins yielded only low levels of sfGFP fluorescence. We reasoned that basal sfGFP expression would correspond to more noise in our multiplexed experiments, as non-interacting barcodes would constitute a larger proportion of all sequenced barcodes. Thus, we empirically optimized the signal to noise ratio of induction by testing a panel of Jun/Fos constructs where pPhlF and pTet had varying ribosome binding sites and sfGFP was driven by several pLac variants. We selected a construct that gave 96x signal of induced/uninduced sfGFP fluorescence, called pSK59 (Figure S2.1B). Although our overall signal strength was weaker with the PhlF RBS variant we selected compared to some constructs assayed, there was extremely little sfGFP fluorescence in our uninduced sample (Figure S2.1C) which we reasoned would lead to higher signal in the multiplexed assay.

To evaluate the quantitative range of our assay we analyzed a previously published set of coiled-coils with K_ds ranging from 10^{-6} to 10^{-10} M,[48] that as well as an additional construct with an inferred $K_d < 10^{-6}$. Measuring sfGFP fluorescence, we found that our system can detect weak interactions, as low as 10^{-6} (Figure S2.1D), however, it lacks the power to discriminate between medium (10^{-7}) to high (10^{-10}) K_ds. In contrast with a previous study using the standard *B. pertussis* adenylate cyclase two-hybrid system, our modified system enables us to achieve quantitative measurements in agreement with published K_ds[49].

8.3 CC1 Library results

The designs from the CC1 Library were expected to be orthogonal within each backbone subset. The expected orthogonal design of the coiled-coils was largely recapitulated in our results (Figure S2.9A), with the P#s having only one strong interaction, P3/P12, which was unexpected. The P#mS and P#SC# backbones also exhibited the expected orthogonal pattern within their

subsets. As P3mS-P8mS and P5SC1-P6SC2 contain the same interfacial residues as their P# counterparts we expected them to cross react with the corresponding P# partners. We found this to be the case with the P#mS having the same reaction pattern (Figure S2.9A bottom, grey) with the P# set as with itself. The P#SC# set also cross reacted as expected but had the unexpected reaction of P6SC#/P7 (Figure S2.9A, bottom orange).

We were curious if our interaction profiles matched what would be expected from our designs. We defined a favorable electrostatic interaction as an E- or G-position having a Glu or Lys forming a salt bridge with a Lys or Glu, respectively, in the corresponding position on the partner protein. Likewise, we defined an Ile/Ile or Asn/Asn pair as having isoleucines or asparagines in the A-position of a given heptad for both proteins. We found that our strong interactions highly favored having all eight possible salt bridges, as nearly all interactions with eight salt-bridges have a higher Interaction score than all other interactions (Figure S2.9B, top). We found the identity of the residue at position A to be less determinative. Though most constructs with a high Interaction score had four pairs of Ile/Ile or Asn/Asn at position A, there were many constructs with four pairs of Ile/Ile or Asn/Asn that did not have a high Interaction score (Figure S2.9B, lower). Taken as a whole, these designs largely functioned as expected: Glu/Lys pairings were far more preferable to Glu/Glu or Lys/Lys pairings and Ile/Ile and Asn/Asn pairs, rather than Ile/Asn pairs, were necessary but not sufficient to create an interaction.

8.4 Effects of backbone variation in the CCNG1 Library

Though the B-, C- and F-position residues are thought to modulate binding affinity, the CCNG1 Library is the first large dataset that systematically tests the effects of different

backbones. A subset of the CCNG1 Library contains the interfacial residues of the CC0 Library on six different backbones, which vary from nearly exclusively small non-polar residues to nearly exclusively large charged residues (Figure S2.13A). To understand how the different backbones affect specificity we divided interactions into on-target and off-target groups where the on-target group had an interface with published Tms greater than 60C (Figure S2.13B). On-target interactions invariably had high interaction scores and no significant difference was noted between the backbones, though the bH backbone did have higher variance than the rest. Off-target interactions, however, unexpectedly showed that the original backbone had lower Interaction scores than the other backbones. To investigate this further we compared each protein, as determined by the interfacial residues, against the previously published Tm (Figure S2.13C). We found that all backbones except the original were strongly shifted to higher interaction scores. Although further tests are needed to understand why there is a global shift to higher interaction scores with other backbones--especially given that highly helical amino acids such as alanine are expected to produce the strongest coiled-coil interactions--one possible hypothesis is that the presence of glutamine in these backbones can facilitate low strength off-target interactions.

When broken down into categories of, small and large underestimates and overestimates, a clear pattern emerged that every backbone had an over representation of interaction scores that were higher than the original backbone, and many were much higher than expected (>1 interaction score more). Conversely, the bA backbone had only a handful of interactions below the expected strength and none that were strongly so. Taken together this suggests the need for care when using backbones with polar residues as unpredicted effects may occur, particularly at the expected lower range.

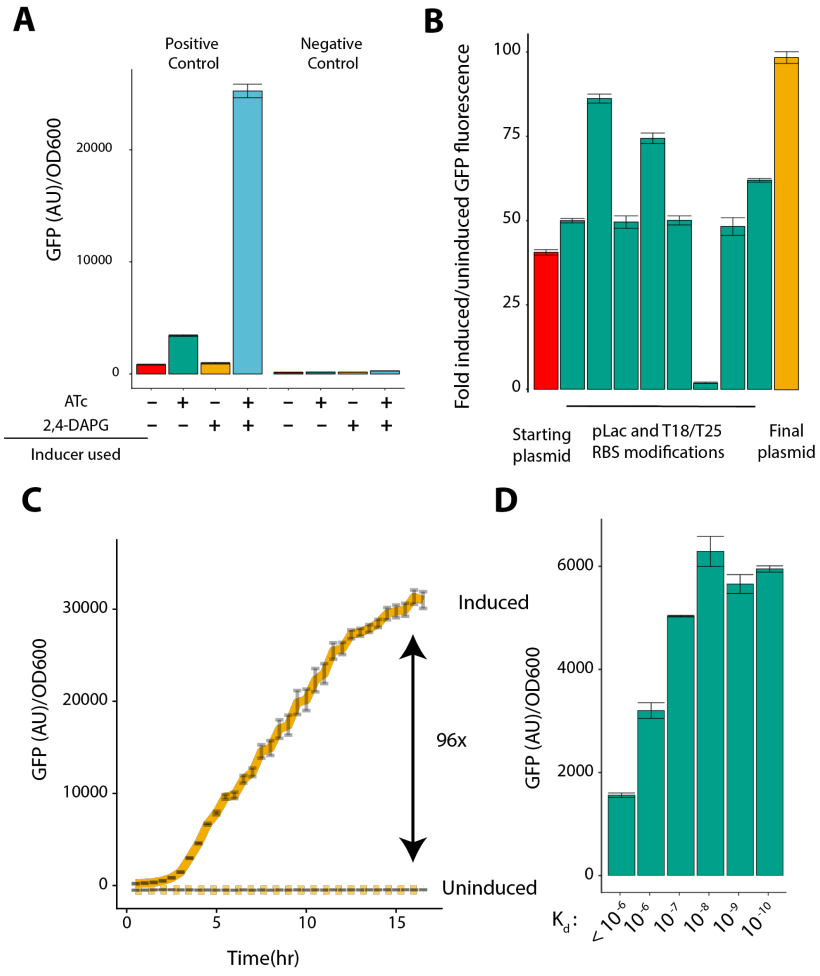

Figure S2.1) Optimization and tuning of the NGB2H system: A) High GFP expression requires positively interacting proteins and both inducers, ATc and DAPG. Error bars are standard deviation of three technical replicates B-C) Optimizations of the promoters and RBSes to find the maximal signal to noise ratio between induced and uninduced samples. B) The ratio of (Induced GFP/OD fluorescence)/(Uninduced GFP/OD fluorescence) for Jun/Fos constructs with RBS and promoter variations. Error bars represent standard deviation of three replicates. C) The Final plasmid induced or uninduced over 16hr. Samples were taken every 30 minutes. Error

bars represent standard deviation of three replicates D) A panel of previously characterized proteins shows GFP/OD depends on K_d with maximal expression occurring at K_d's stronger than 10^{-7} M. Error bars represent standard deviation of three replicates.

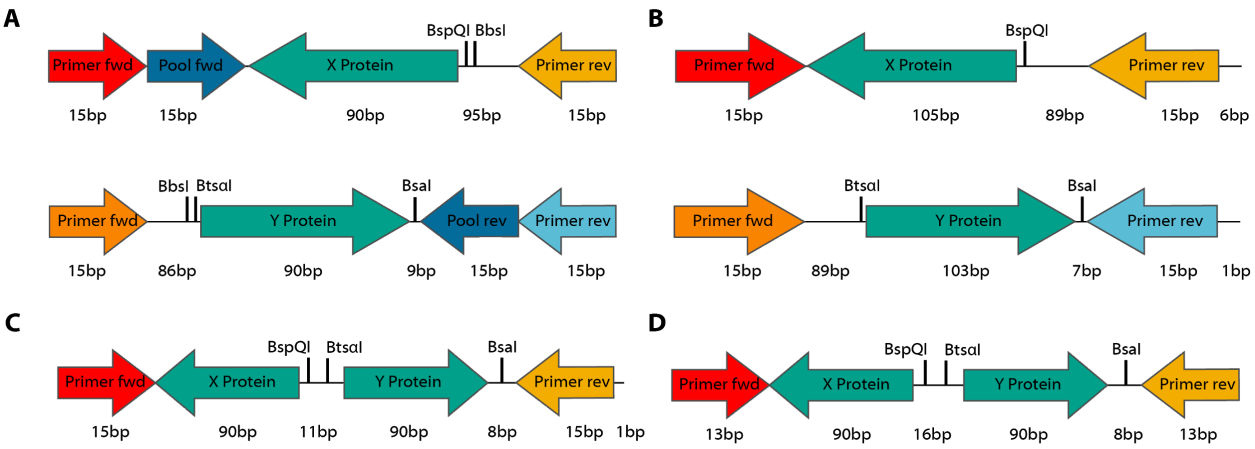

Figure S2.2) Design of OLS oligonucleotides for libraries used in this work: With minor variation each oligonucleotide consists of primers to amplify it out of the OLS pool, the coding sequence and several restriction enzyme sites. Numbers below the constructs represent how CC0 Library oligonucleotides are divided into those with the X protein and those with the Y protein. B) CC1 Library oligonucleotides are divided into those coding the X protein and those coding the Y protein. C) CCNG1 Library oligonucleotides contain both the X and Y protein on a single oligonucleotide. D) CCmax Library oligonucleotides contain both the X and Y protein on a single oligonucleotide.

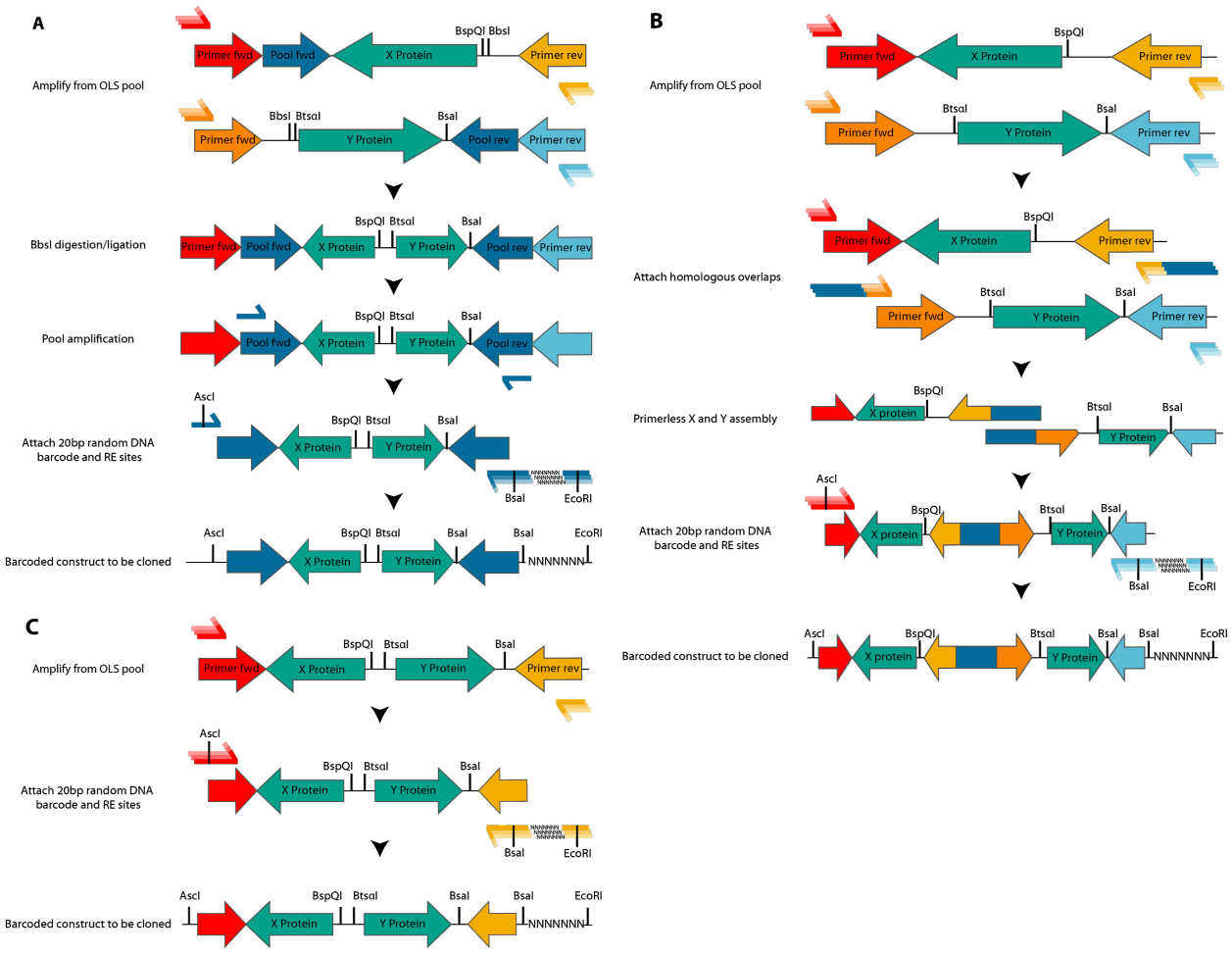

Figure S2.3) Cloning from OLS oligonucleotides to barcoded X and Y constructs: A) CC0 Library barcoding schema. Five OLS pools for both the X and Y oligonucleotides are amplified with OLS primers. All oligonucleotides are digested with BbsI and ligated in pairs, before amplification with pool primers and mixing. Finally, restriction enzyme sites for cloning into the vector and the barcode are attached. B) CC1 Library barcoding schema. After amplification of the OLS pool, matching overlaps are attached which are stitched together with overlap PCR. Finally, the barcode and restriction enzyme sites for cloning into the vector are attached. C)

CCNG1 and CCmax Library barcoding schema. Oligonucleotides are amplified from the pool and then restriction enzyme sites for cloning into the vector and barcode are attached.

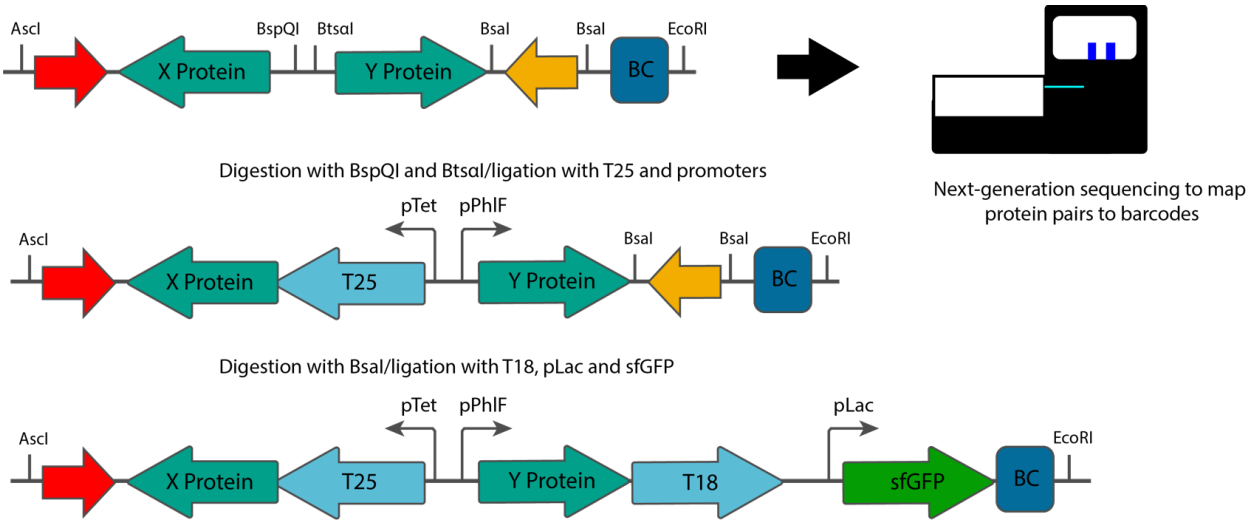

Figure S2.4) Cloning scheme of the NGB2H system after barcoding: After barcoding, the X and Y proteins are sequenced through the barcode using an Illumina MiSeq to identify each barcode's corresponding protein pair. After mapping, the T25 section and inducible promoters are cloned into the mapped plasmid. After the T25 section is cloned, the T18 section with sfGFP is cloned into the T25 containing plasmid.

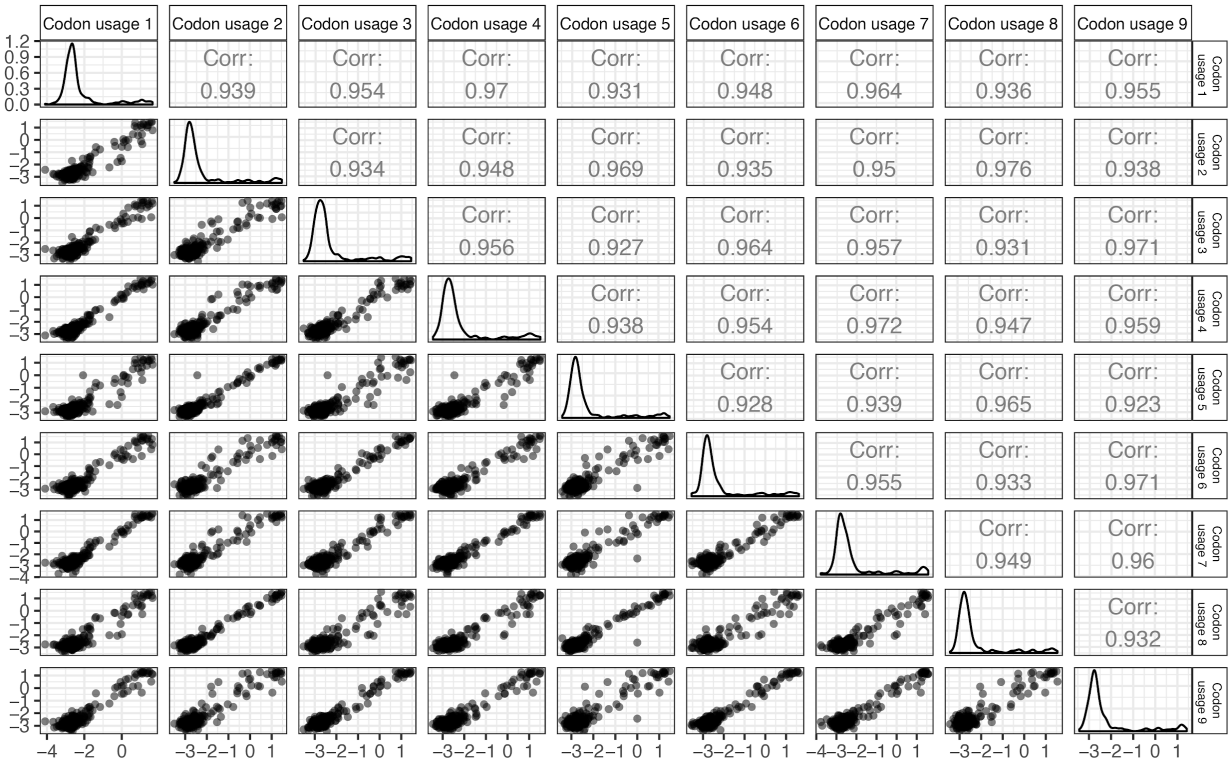

Figure S2.5) Different codon usages of the CC0 Library: Different codon usages in the CC0 Library produce similar Interaction Scores. All nine codon usages from the CC0 Library show high replicability with Pearson's $r > 0.92$ in all pairwise interactions and a mean replicability of $r > 0.949$.

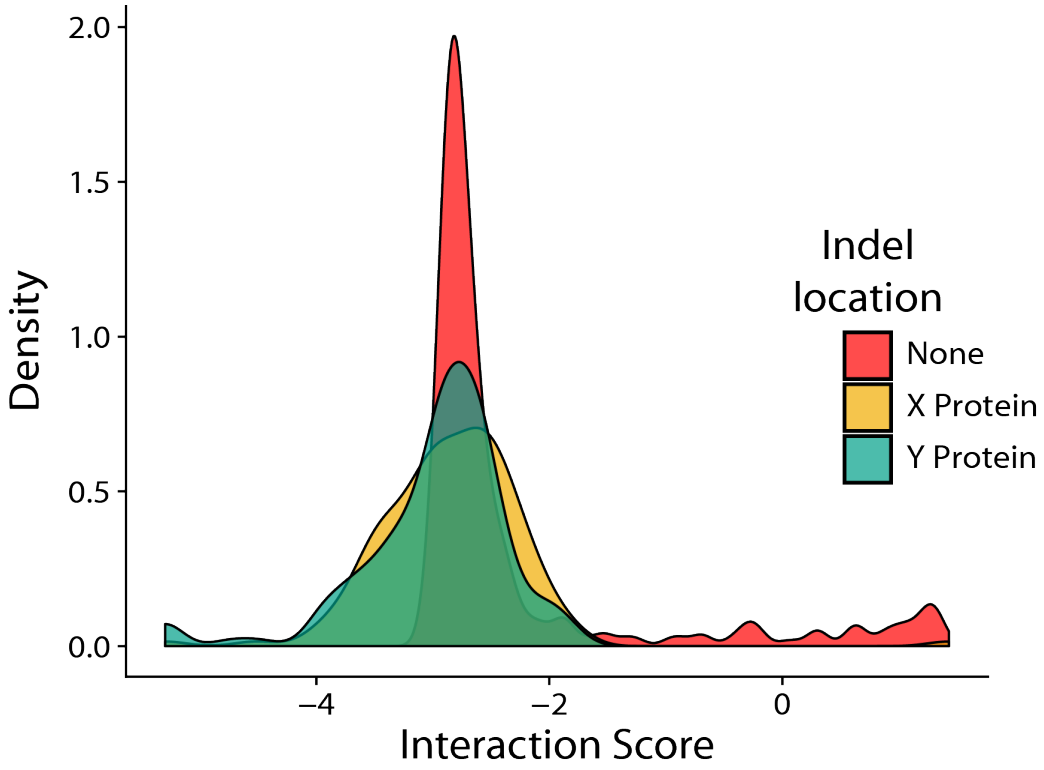

Figure S2.6) Indels in the CC0 Library have lower interaction scores than correct sequences: Constructs with insertions or deletions in the X or Y protein invariably have an Interaction Score of less than -1.8. Constructs without indels, however have some Interaction scores as great as .8.

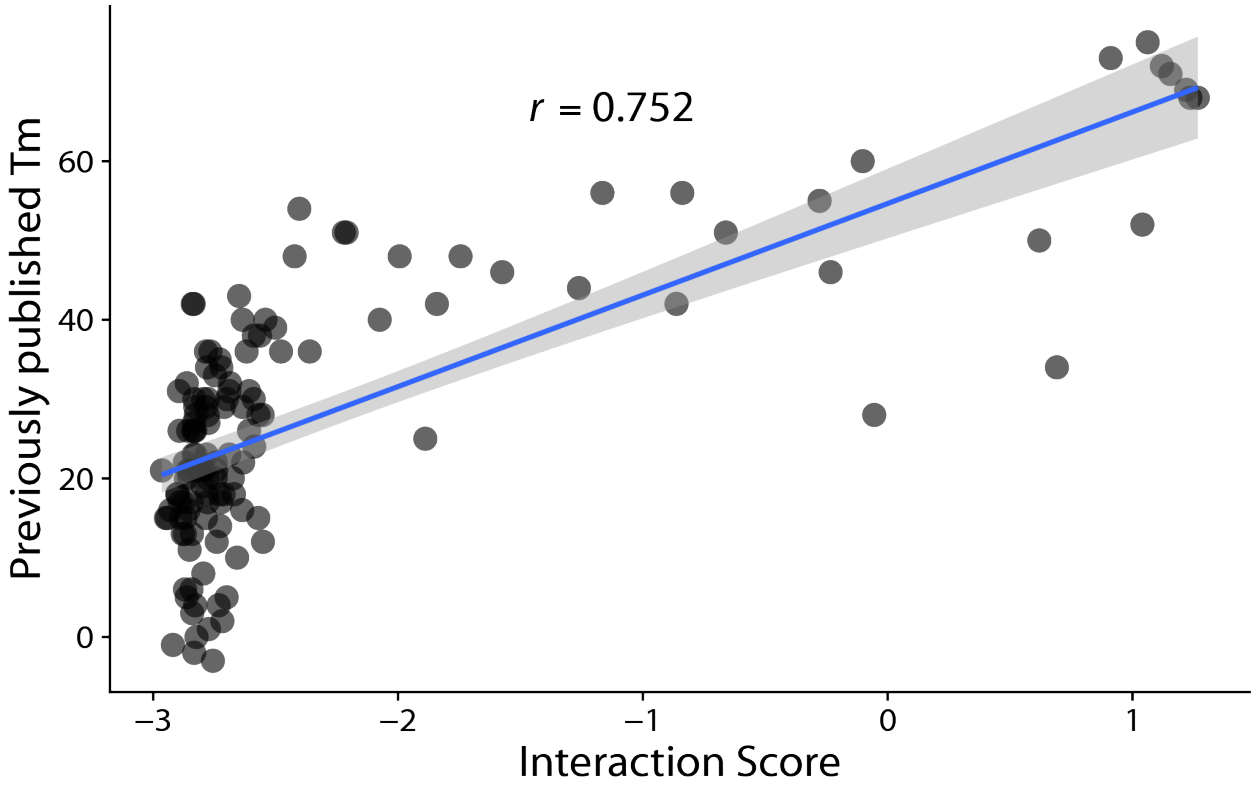

Figure S2.7) CC0 Library Interaction Scores versus previously published Tms: The Tms measured by circular dichroism correlate well with the Interaction Score measured in the CC0 Library with Pearson's $r > 0.75$. Tms > 40C were well distinguished by the NGB2H assay. Blue line represents a linear model fit to the data, with standard error as the gray shading.

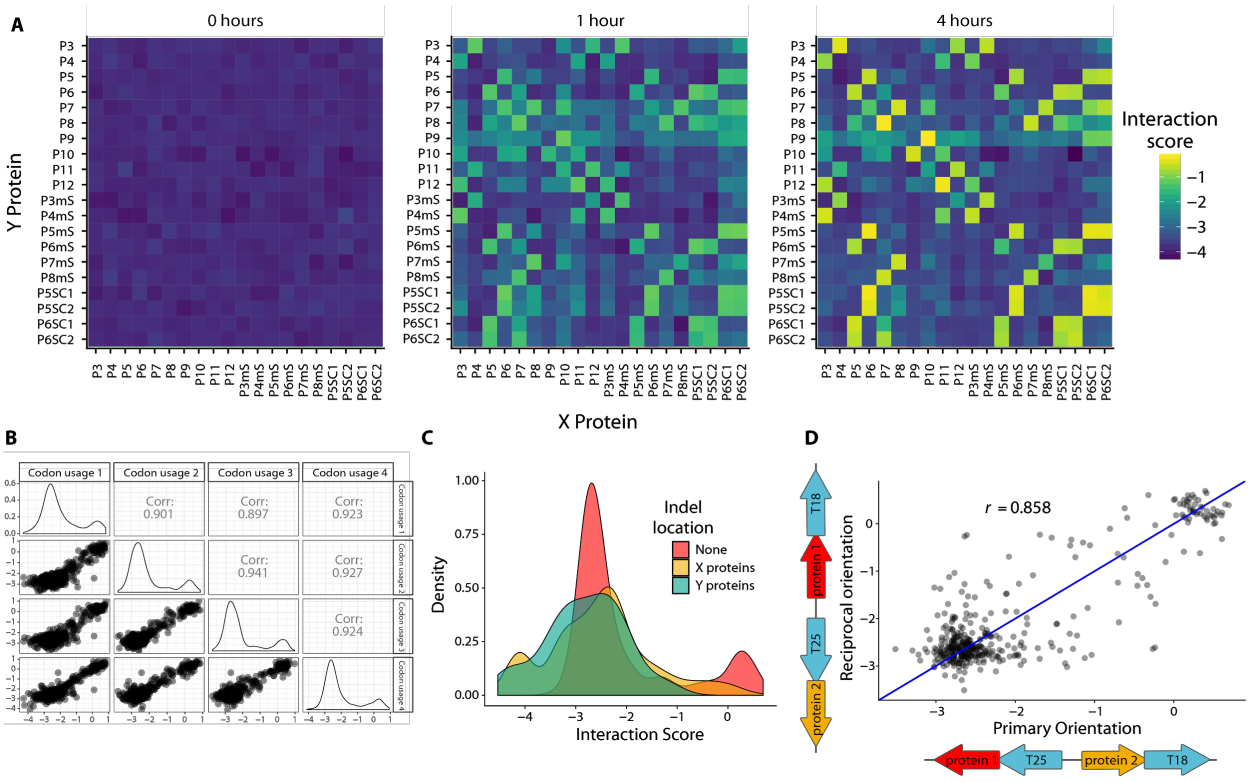

Figure S2.8) CC1 Library internal controls: A) A timecourse experiment of the CC1 Library shows increasing Interaction Scores over four hours, with the strongest signal coming at four hours. Interaction scores were normalized across time with the constitutive GFP Library. B) Different codon usages for the CC1 Library replicate with Pearson's R > 0.89 for all pairwise interactions and a mean of Pearson's $r = 0.918$ C) The Interaction Score for constructs with indels in the CC1 Library is lower than for those without indels. D) The CC1 Library constructs give similar ($r > 0.85$) Interaction Scores for protein pairs attached

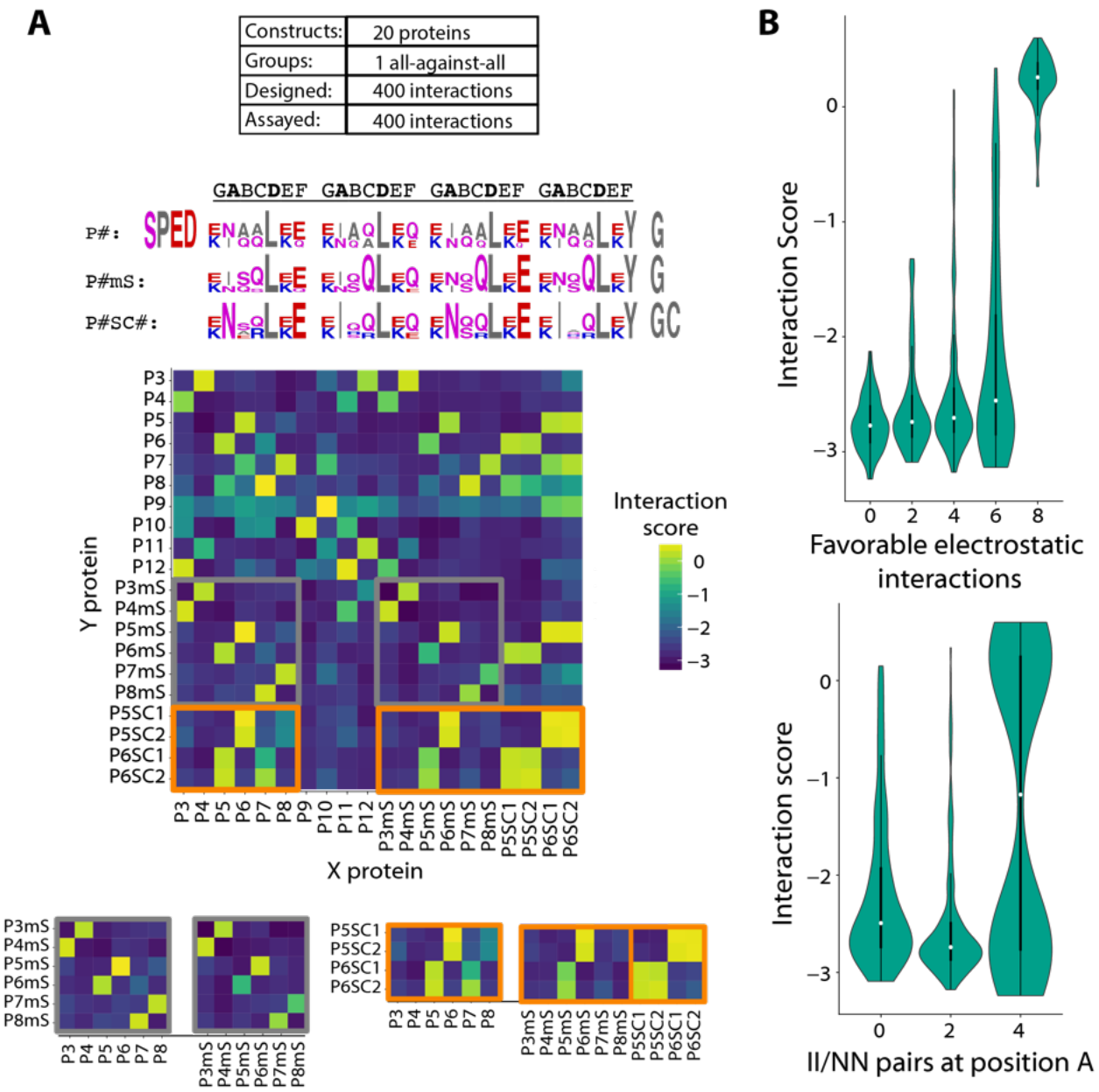

Figure S2.9) Isolation of orthogonal coiled-coils from the CC1 Library. A) (Upper) Expected outcome for the CC1 Library and logo illustrating the library diversity. Residues are colored by their functional group Red: positively charged, Blue: negatively charged, purple: polar, grey: non-polar or other (Lower) Interaction scores for entire CC1 Library. The inset in gray shows similarities between library P# and P#mS library members. The inset in orange shows similarities between P#, P#mS and P#SC#. B) (Upper) The Interaction score is highly dependent

on the number of electrostatic interactions. White dot represents the median. (Lower) Interaction score is dependent on the residue pairing in the A position, but this is not determinative.

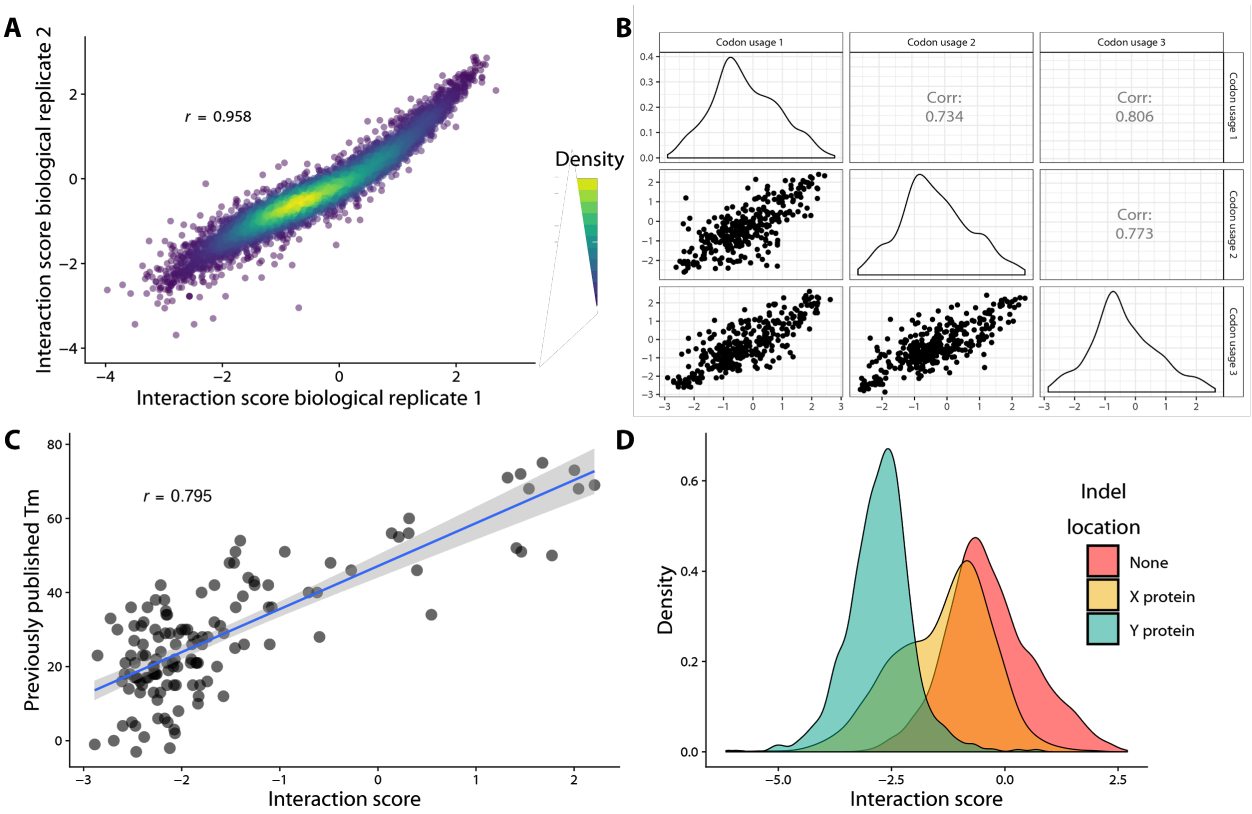

Figure S2.10) CCNG1 Library internal controls: A) The CCNG1 Library replicates in biological replicates with Pearson's $r > 0.95$ ($p < 10^{-15}$). B) Different codon usages in the CCNG1 Library correlate with each other with $r > 0.73$ ($p < 10^{-15}$) for all pairwise comparisons (mean = 0.77). For this analysis only constructs with ten barcodes were used. C) The CC0 Library was included in the CCNG1 Library, and correlates with previously published Tms with Pearson's $r > 0.79$ ($p < 10^{-15}$). Blue line represents a linear model fit to the Interaction scores; grey shading represents the standard error) Indels in the CCNG1 Library show decreased Interaction scores compared to constructs without indels.

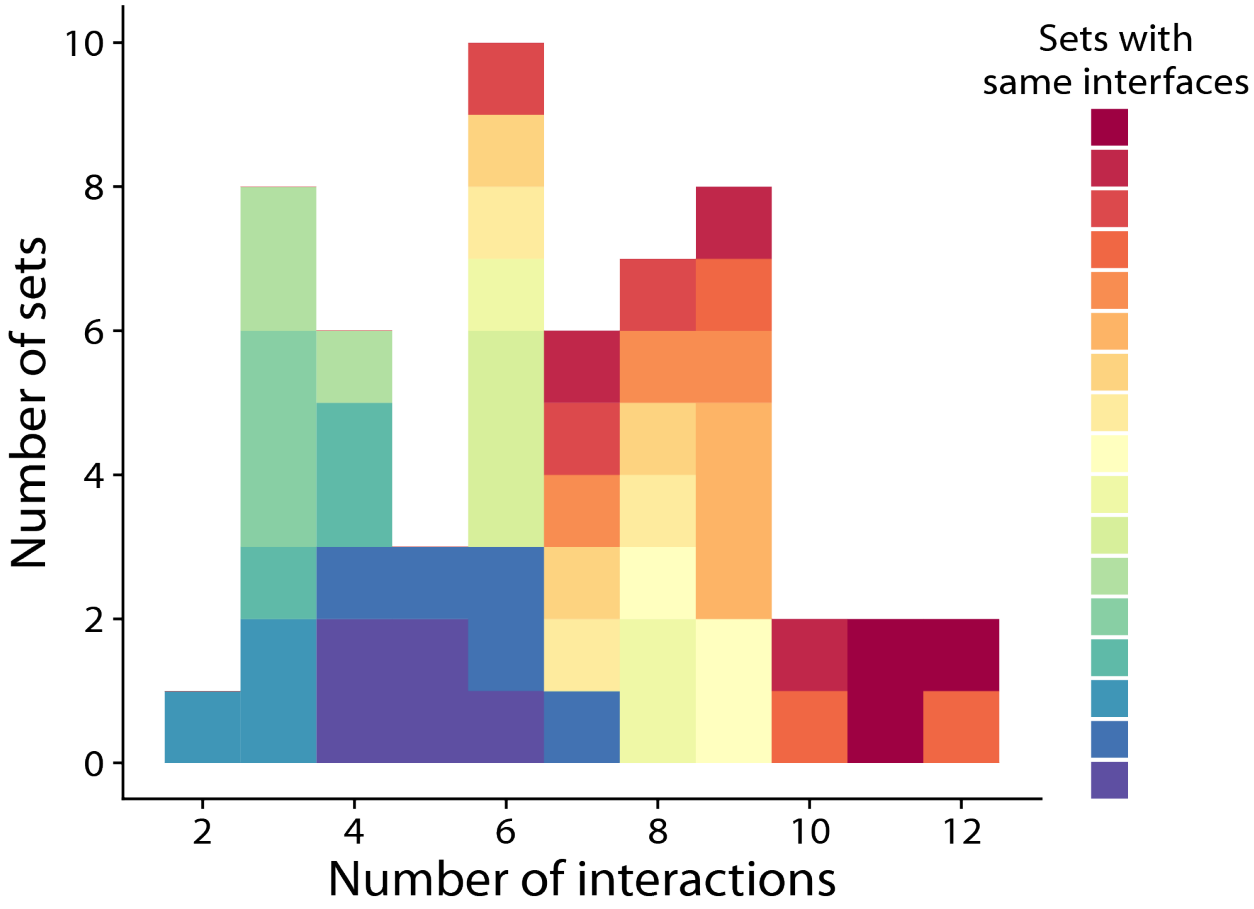

Figure S2.11) Number of orthogonal interactions by sets with different backbones in the CCNG1 Library: Sets with different backbones but the same interfacial residues generally contained similar numbers of orthogonal interactions.

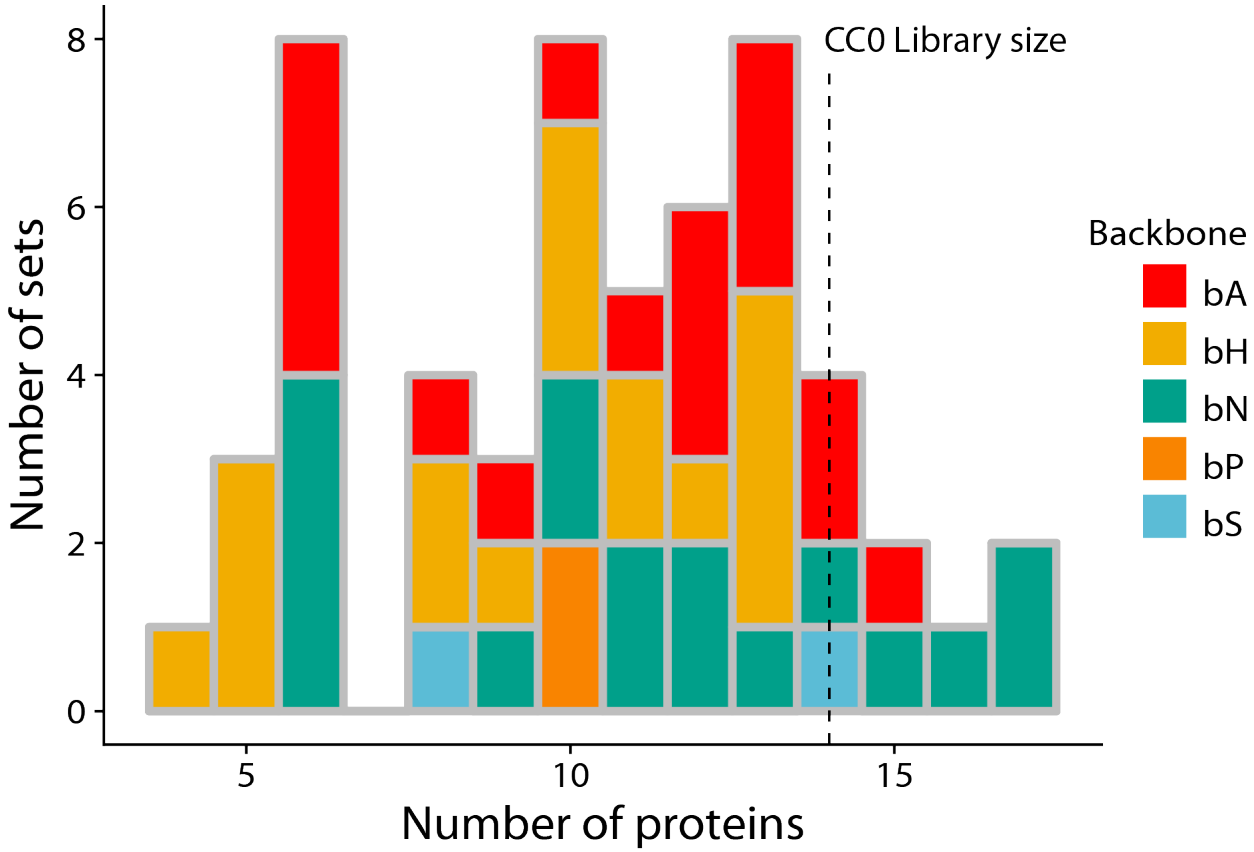

Figure S2.12) CCNG1 Library number of proteins per orthogonal set: Sets in the CCNG1 Library contained between four and seventeen distinct proteins. Five of these contain more than the fourteen in the CC0 Library. Different backbones contained the same interfacial patterns.

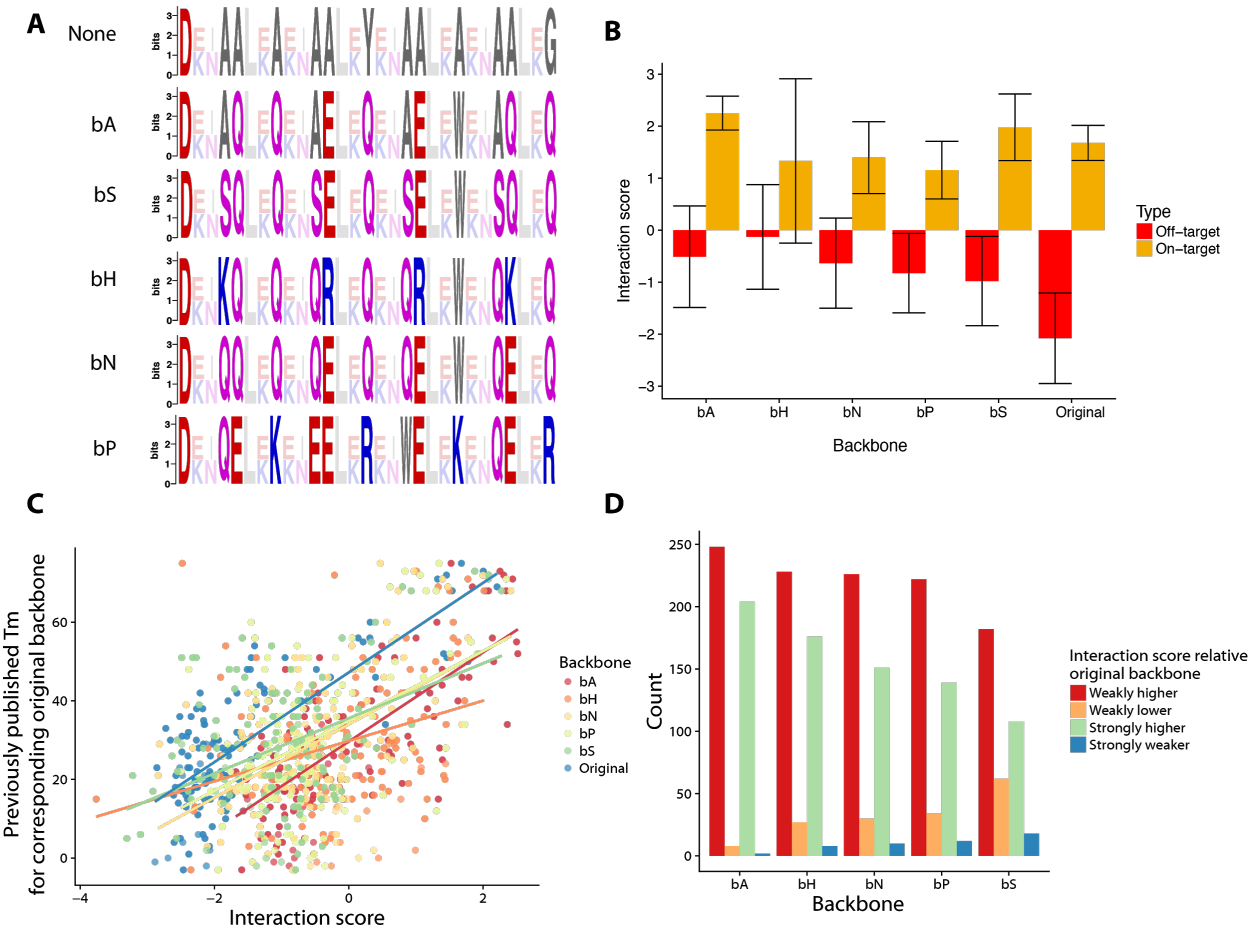

Figure S2.13) Effect of variation at the b, c, and f-positions: A) Sequence logos representing the different backbones. Amino acids are colored according to their type. Red: negative, Blue: positive, Purple: polar, Grey: non-polar. Non-transparent residues are backbone residues. B) Mean Interaction scores for different backbones. On-target is defined as the eight interactions with predicted Tm = 72C with bCipa; off-target interactions are all other interactions. Error bars are standard deviation. C) The Interaction scores from each backbone compared to the Tm of proteins that share the same interfacial residues D) Counts of Interactions scores for different backbones above or below the Interaction score for the original backbone. Strongly higher/lower is defined as an interaction score greater than ±1.

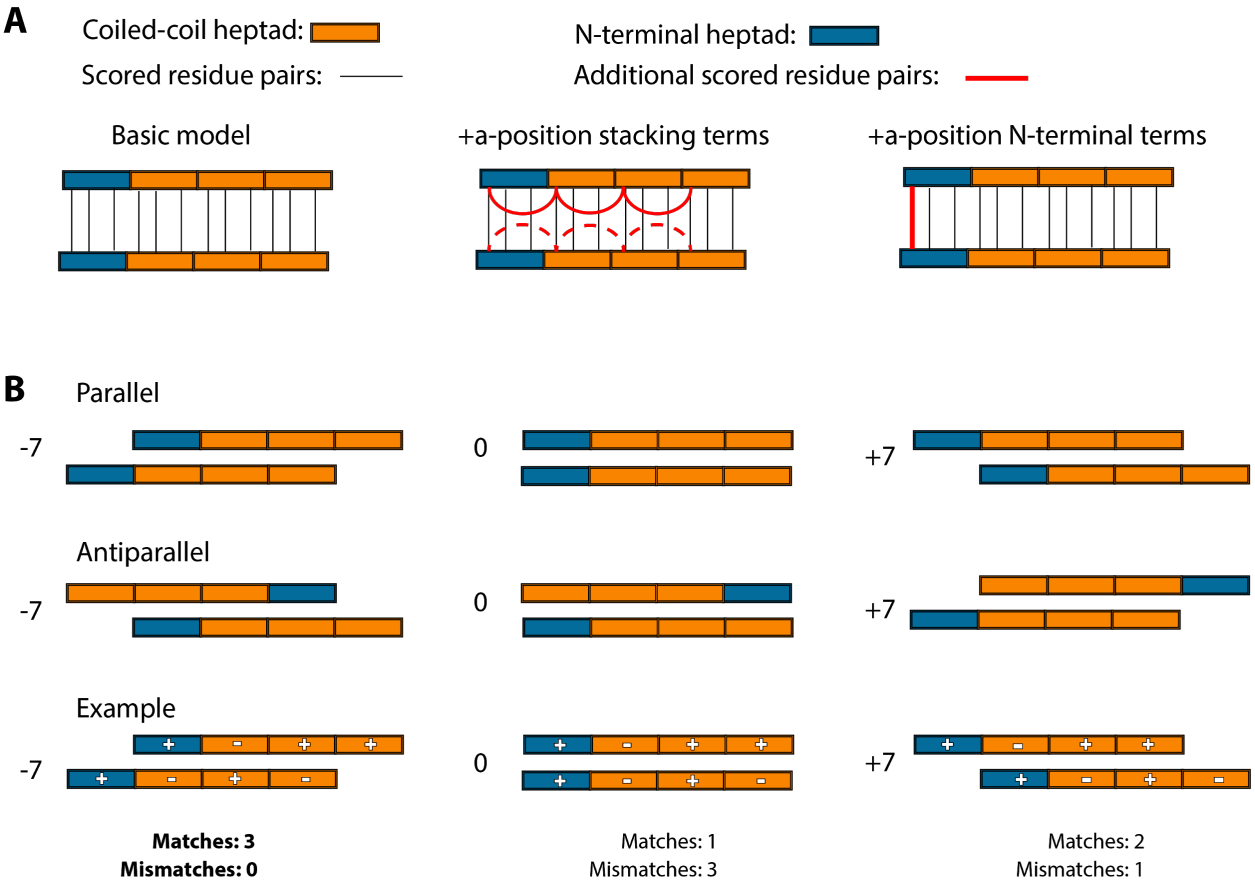

Figure S2.14) Schematic of heptad shifting: A) Different model variations tried for iCipa. The basic model only scores a-, e- and g-positions. The +a position stacking terms model scores consecutive residues in the a-position while the +a-position N-terminal terms model includes separate weights for the first a-position. B) All iCipa candidates score interactions with heptad shifting, that is moving up or down seven residues in an interaction. From left to right shows progressive heptad shifts of the bottom coiled-coil with respect to the top coiled coil for both parallel and antiparallel coiled-coils. (Bottom row) As an example illustrating how heptad shifting is scored, each heptad is given a plus sign or a minus sign, the combination of which is considered a match. In the -7 position all three heptads match giving a high score. In the 0 and 7

positions the heptads have fewer matches and more mismatches so the -7 position would be chosen as the orientation to score. Note though, iCipa calculates individual residues rather than entire heptads at a time.

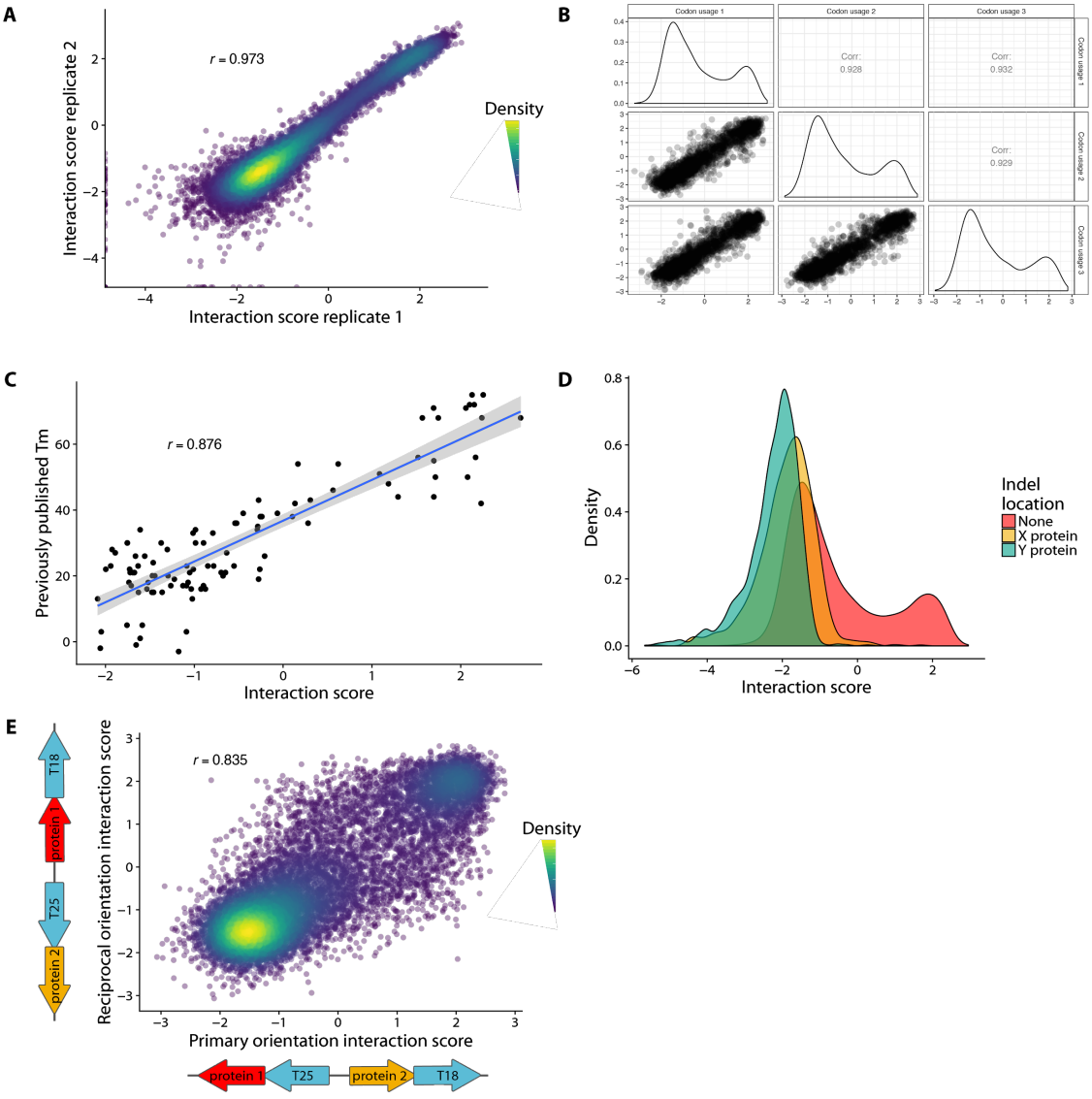

Figure S2.15) CCmax Library internal controls: A) Interaction scores of variants in the CCmax Library correlate strongly between biological replicates (Pearson's $r > 0.97$, $p < 10^{-15}$). B) Different codon usages of the CCmax Library have similar Interaction scores with Pearson's $r > 0.92$ and $p < 10^{-15}$ for all pairwise comparisons. C) The CCmax Library contained the CC0 Library. When our Interaction scores are compared to the previously published Tms they correlate well with Pearson's $r > 0.87$, $p < 10^{-15}$. D) Correct constructs from the CCmax Library

have a higher interaction score than those produced with indels. E) The reciprocal orientations of the CCmax Library have similar Interaction scores, and correlate with Pearson's $r > 0.83$, $p < 10^{-15}$.

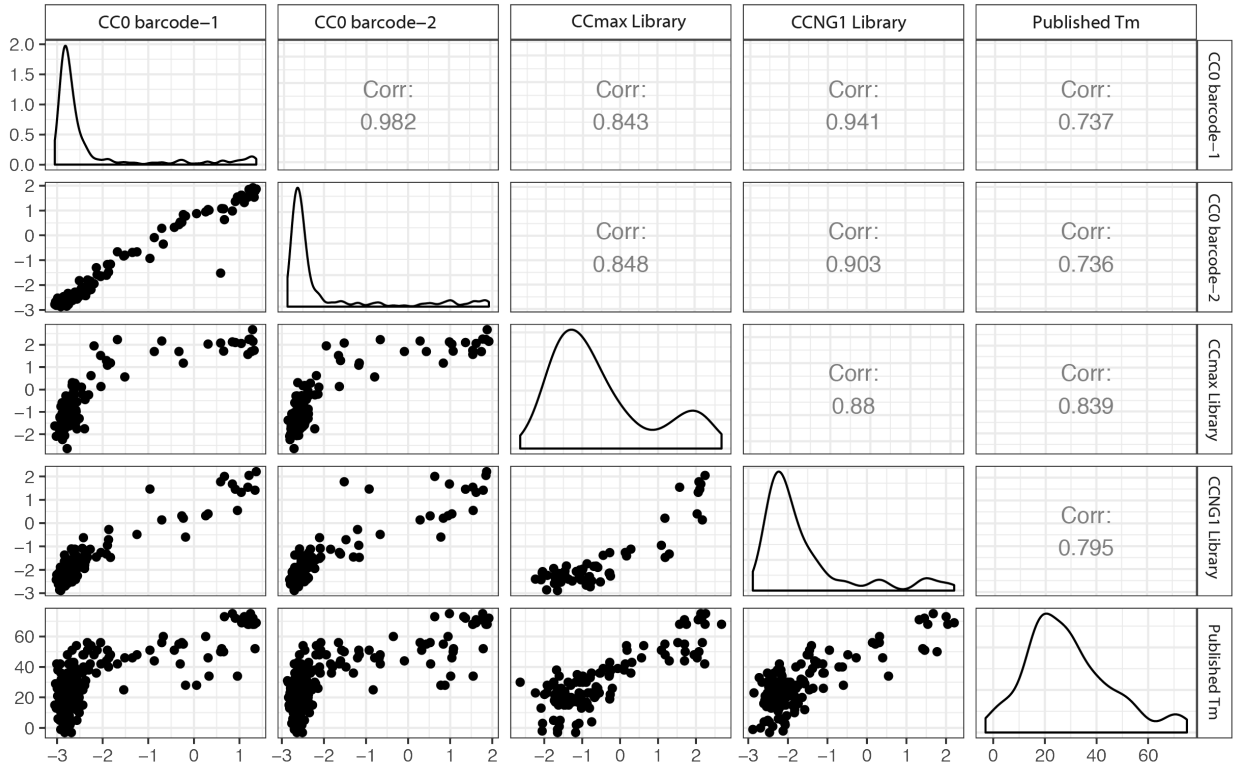

Figure S2.16) Correlation of the CC0 Library proteins between different libraries: The CC0 Library was a subset of all libraries except CC1. Comparing how it performed in all libraries shows strong agreement between sets with Pearson's $r > 0.84$, $p < 10^{-15}$ between all libraries and Pearson's $r > 0.73$, $p < 10^{-15}$ for all libraries with the previously published Tms.

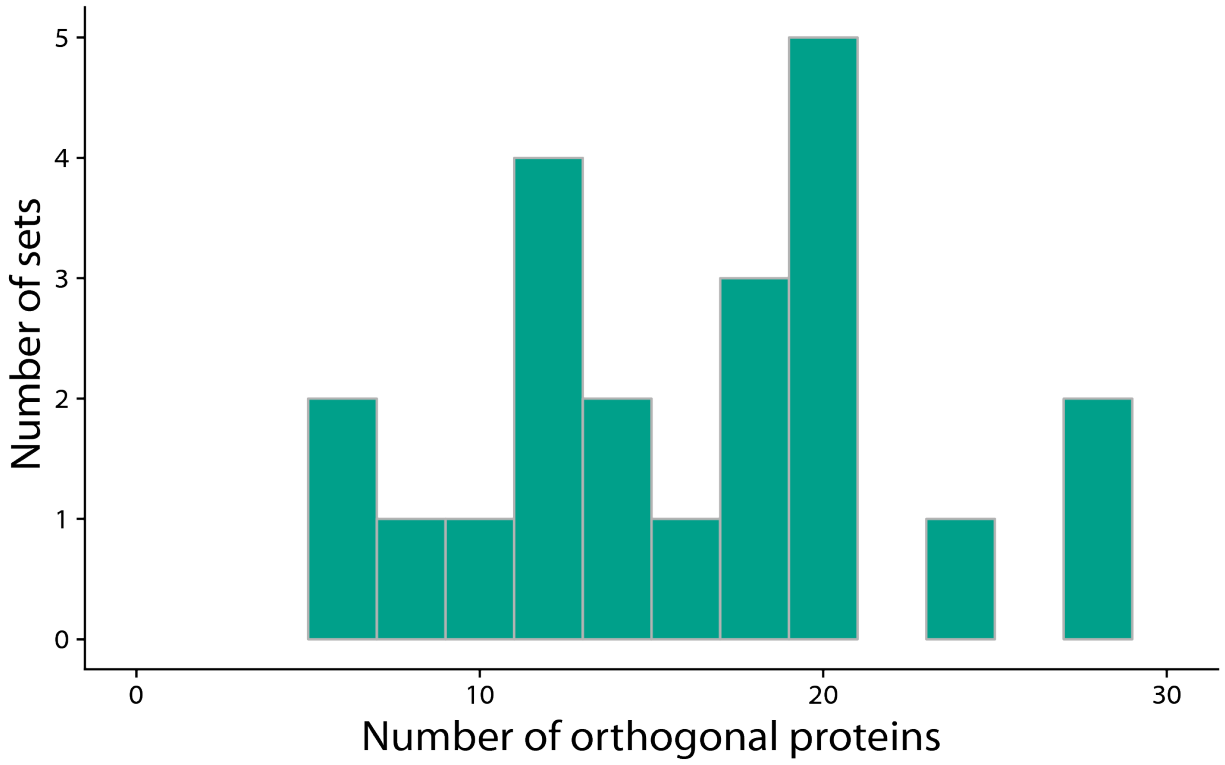

Figure S2.17) Number of orthogonal proteins per set: The CCmax Library had orthogonal sets that contained the most orthogonal proteins of any group of orthogonal proteins to date. Sets contained between 6 and 28 proteins or between 36 and 784 total interactions.

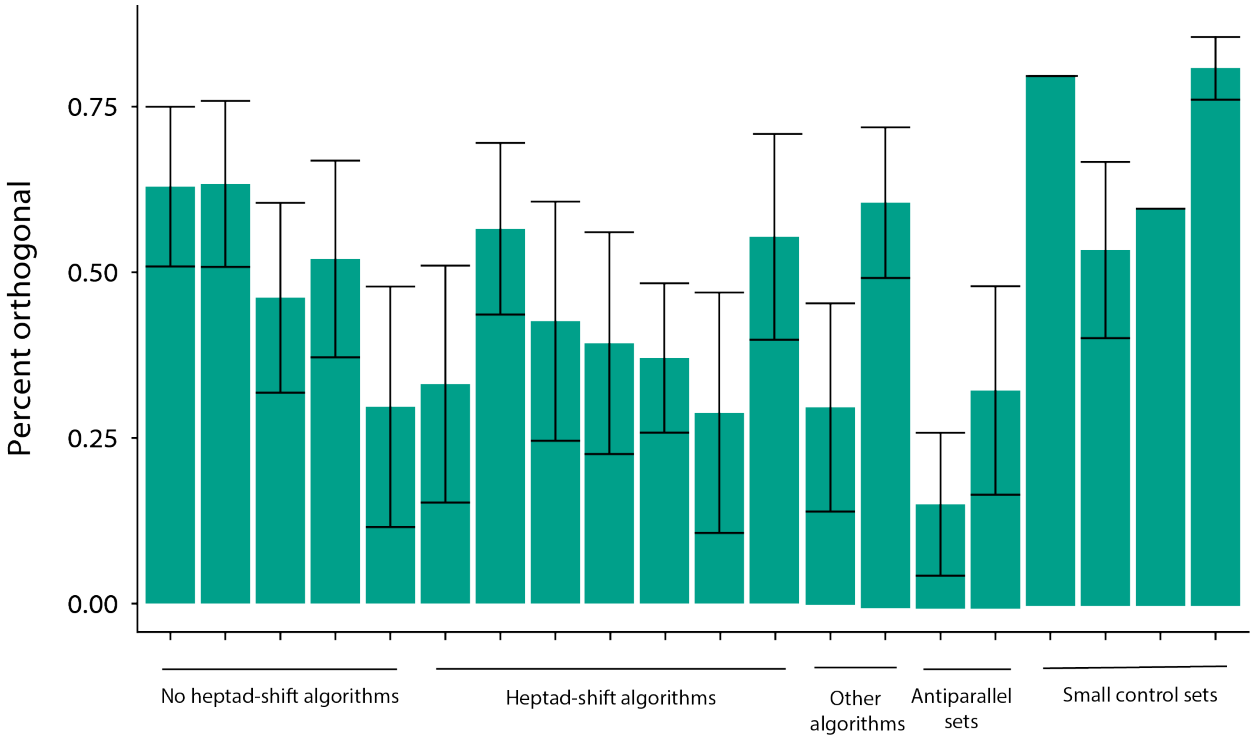

Figure S2.18) Ability to predict orthogonality compared across algorithms: Sets of proteins from the CCmax Library that were designed with different algorithms were randomly subsampled to subsets of ten proteins and the largest orthogonal group was identified. Subsampling was repeated 500 times. Error bars are standard deviation.

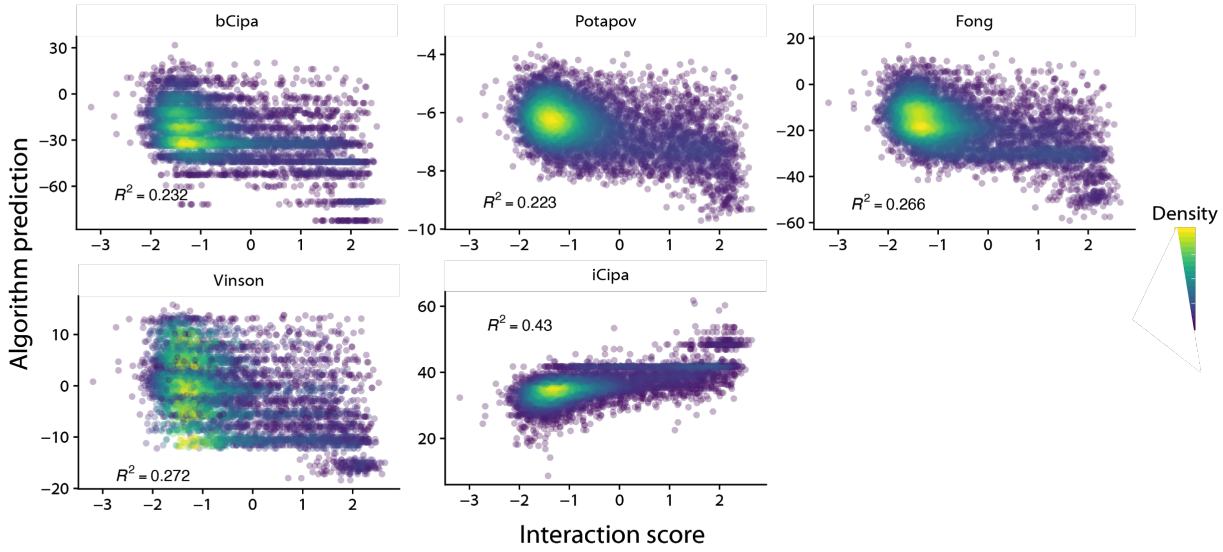

Figure S2.19) CCmax Library's agreement with previous models: Interaction scores from the CCmax Library correlate poorly with previous models. All previous models predict Interaction scores with a coefficient of determination less than 0.28, but iCipa predicts Interaction scores with $R^2 = 0.43$.

References:

1. Vidal M, Cusick ME, Barabási A-L. Interactome networks and human disease. *Cell*. 2011;144(6):986-998. doi:10.1016/j.cell.2011.02.016
2. Huang P-S, Boyken SE, Baker D. The coming of age of de novo protein design. *Nature*. 2016;537(7620):320-327. doi:10.1038/nature19946
3. Ljubetič A, Gradišar H, Jerala R. Advances in design of protein folds and assemblies. *Curr Opin Chem Biol*. 2017;40:65-71. doi:10.1016/j.cbpa.2017.06.020
4. Ljubetič A, Lapenta F, Gradišar H, et al. Design of coiled-coil protein-origami cages that self-assemble in vitro and in vivo. *Nat Biotechnol*. 2017;35(11):1094-1101. doi:10.1038/nbt.3994
5. Gradišar H, Božič S, Doles T, et al. Design of a single-chain polypeptide tetrahedron assembled from coiled-coil segments. *Nat Chem Biol*. 2013;9(6):362-366. doi:10.1038/nchembio.1248
6. Boyken SE, Chen Z, Groves B, et al. De novo design of protein homo-oligomers with modular hydrogen-bond network – mediated specificity. 2016;399(1999):69-72.
7. Fallas JA, Ueda G, Sheffler W, et al. Computational design of self-assembling cyclic protein homo-oligomers. *Nat Chem*. 2017;9(4):353-360. doi:10.1038/nchem.2673
8. Chen Z, Boyken SE, Jia M, et al. Programmable design of orthogonal protein heterodimers. *Nature*. 2019;565(7737):106-111. doi:10.1038/s41586-018-0802-y
9. Pauling L, Corey RB. Compound Helical Configurations of Polypeptide Chains: Structure of Proteins of the α-Keratin Type. *Nature*. 1953;171(4341):59-61. doi:10.1038/171059a0
10. Crick FHC. The packing of α-helices: simple coiled-coils. *Acta Crystallogr*. 1953;6(8):689-697. doi:10.1107/S0365110X53001964

11. Crick FHC. The Fourier transform of a coiled-coil. *Acta Crystallogr*. 1953;6(8):685-689. doi:10.1107/s0365110x53001952

12. Acharya A, Rishi V, Vinson C. Stability of 100 homo and heterotypic coiled-coil a-a′ pairs for ten amino acids (A, L, I, V, N, K, S, T, E, and R). *Biochemistry*. 2006;45(38):11324-11332. doi:10.1021/bi060822u

13. Potapov V, Kaplan JB, Keating AE. Data-Driven Prediction and Design of bZIP Coiled-Coil Interactions. *PLoS Comput Biol*. 2015;11(2):1-28. doi:10.1371/journal.pcbi.1004046

14. Mason JM, Schmitz M a, Müller KM, Arndt KM. Semirational design of Jun-Fos coiled coils with increased affinity: Universal implications for leucine zipper prediction and design. *Proc Natl Acad Sci U S A*. 2006;103(24):8989-8994. doi:10.1073/pnas.0509880103

15. Crooks RO, Lathbridge A, Panek AS, Mason JM. Computational Prediction and Design for Creating Iteratively Larger Heterospecific Coiled Coil Sets. *Biochemistry*. 2017;56(11):1573-1584. doi:10.1021/acs.biochem.7b00047

16. Thompson KE, Bashor CJ, Lim WA, Keating AE. SYNZIP Protein Interaction Toolbox: *in Vitro* and *in Vivo* Specifications of Heterospecific Coiled-Coil Interaction Domains. *ACS Synth Biol*. 2012;1(4):118-129. doi:10.1021/sb200015u

17. Karimova G, Pidoux J, Ullmann a, Ladant D. A bacterial two-hybrid system based on a reconstituted signal transduction pathway. *Proc Natl Acad Sci U S A*. 1998;95(10):5752-5756. doi:10.1073/pnas.95.10.5752

18. Brodnik A, Palangetić M, Siladi D, Jovičić V. Construction of orthogonal CC-sets. *Inform*. 2019;43(1):19-22. doi:10.31449/inf.v43i1.2693

19. Drobnak I, Gradišar H, Ljubetič A, Merljak E, Jerala R. Modulation of Coiled-Coil Dimer

Stability through Surface Residues while Preserving Pairing Specificity. *J Am Chem Soc*. 2017;139(24):8229-8236. doi:10.1021/jacs.7b01690

20. Fong J, Keating A, Singh M. Predicting specificity in bZIP coiled-coil protein interactions. *Genome Biol*. 2004;5(2):R11. doi:10.1186/gb-2004-5-2-r11

21. Yachie N, Petsalaki E, Mellor JC, et al. Pooled-matrix protein interaction screens using Barcode Fusion Genetics. *Mol Syst Biol*. 2016;12(4):863-863. doi:10.15252/msb.20156660

22. Trigg SA, Garza RM, MacWilliams A, et al. CrY2H-seq: A massively multiplexed assay for deep-coverage interactome mapping. *Nat Methods*. 2017;14(8):819-825. doi:10.1038/nmeth.4343

23. Yang JS, Garriga-Canut M, Link N, et al. rec-YnH enables simultaneous many-by-many detection of direct protein–protein and protein–RNA interactions. *Nat Commun*. 2018;9(1). doi:10.1038/s41467-018-06128-x

24. Yang F, Lei Y, Zhou M, et al. Development and application of a recombination-based library versus library highthroughput yeast two-hybrid (RLL-Y2H) screening system. *Nucleic Acids Res*. 2018;46(3):1-12. doi:10.1093/nar/gkx1173

25. Younger D, Berger S, Baker D, Klavins E. High-throughput characterization of protein–protein interactions by reprogramming yeast mating. *Proc Natl Acad Sci U S A*. 2017;114(46):12166-12171. doi:10.1073/pnas.1705867114

26. Diss G, Lehner B. The genetic landscape of a physical interaction. *Elife*. 2018;7:1-31. doi:10.7554/eLife.32472

27. Andrews SS, Schaefer-Ramadan S, Al-Thani NM, Ahmed I, Mohamoud YA, Malek JA. High-resolution protein–protein interaction mapping using all- versus -all sequencing

(AVA-Seq) . *J Biol Chem*. 2019;294(30):11549-11558. doi:10.1074/jbc.ra119.008792

28. Plesa C, Sidore AM, Lubock NB, Zhang D, Kosuri S. Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science (80-)*. 2018;359(6373):343-347. doi:10.1126/science.aao5167

29. Sidore AM, Plesa C, Samson JA, Lubock NB, Kosuri S. DropSynth 2.0: high-fidelity multiplexed gene synthesis in emulsions. *Nucleic Acids Res*. 2020;48(16):e95. doi:10.1093/nar/gkaa600

30. McClune CJ, Alvarez-Buylla A, Voigt CA, Laub MT. Engineering orthogonal signalling pathways reveals the sparse occupancy of sequence space. *Nature*. 2019;574(7780):702-706. doi:10.1038/s41586-019-1639-8

31. Chen Z, Kibler RD, Hunt A, et al. De novo design of protein logic gates. *Science (80-)*. 2020;368(6486):78-84. doi:10.1126/science.aay2790

32. Fink T, Lonzarić J, Praznik A, et al. Design of fast proteolysis-based signaling and logic circuits in mammalian cells. *Nat Chem Biol*. 2019;15(2):115-122. doi:10.1038/s41589-018-0181-6

33. Lebar T, Lainšček D, Merljak E, Aupič J, Jerala R. A tunable orthogonal coiled-coil interaction toolbox for engineering mammalian cells. *Nat Chem Biol*. 2020;16(5):513-519. doi:10.1038/s41589-019-0443-y

34. Kosuri S, Eroshenko N, Leproust EM, et al. Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat Biotechnol*. 2010;28(12):1295-1299. doi:10.1038/nbt.1716

35. Kuhlman T, Zhang Z, Saier MH, Hwa T. Combinatorial transcriptional control of the lactose operon of Escherichia coli. *Proc Natl Acad Sci U S A*. 2007;104(14):6043-6048.

doi:10.1073/pnas.0606717104

36. Battesti A, Bouveret E. The bacterial two-hybrid system based on adenylate cyclase reconstitution in Escherichia coli. *Methods*. 2012;58(4):325-334. doi:10.1016/j.ymeth.2012.07.018

37. Stanton BC, Nielsen A a K, Tamsir A, Clancy K, Peterson T, Voigt C a. Genomic mining of prokaryotic repressors for orthogonal logic gates. *Nat Chem Biol*. 2014;10(2):99-105. doi:10.1038/nchembio.1411

38. Chen Y-J, Liu P, Nielsen AAK, et al. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nat Methods*. 2013;10(7):659-664. doi:10.1038/nmeth.2515

39. Badran AH, Guzov VM, Huai Q, et al. Continuous evolution of Bacillus thuringiensis toxins overcomes insect resistance. *Nature*. 2016;533(7601):58-63. doi:10.1038/nature17938

40. Gradišar H, Jerala R. De novo design of orthogonal peptide pairs forming parallel coiled-coil heterodimers. *J Pept Sci*. 2011;17(2):100-106. doi:10.1002/psc.1331

41. Islam S, Kjallquist U, Moliner A, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res*. 2011;21(7):1160-1167. doi:10.1101/gr.110882.110

42. Kosuri S, Goodman DB, Cambray G, et al. Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proc Natl Acad Sci*. 2013;110(34):14024-14029. doi:10.1073/pnas.1301301110

43. Battesti A, Bouveret E. Improvement of bacterial two-hybrid vectors for detection of fusion proteins and transfer to pBAD-tandem affinity purification, calmodulin binding

peptide, or 6-histidine tag vectors. *Proteomics*. 2008;8(22):4768-4771. doi:10.1002/pmic.200800270

44. Cox R, Dunlop MJ, Elowitz MB. A synthetic three-color scaffold for monitoring genetic regulation and noise. *J Biol Eng*. 2010;4(1):10. doi:10.1186/1754-1611-4-10

45. Lou C, Stanton B, Chen YJ, Munsky B, Voigt CA. Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nat Biotechnol*. 2012;30(11):1137-1142. doi:10.1038/nbt.2401

46. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods*. 2009;6(5):343-345. doi:10.1038/nmeth.1318

47. Bushnell B. BBMap.

48. Thomas F, Boyle AL, Burton AJ, Woolfson DN. A set of de novo designed parallel heterodimeric coiled coils with quantified dissociation constants in the micromolar to sub-nanomolar regime. *J Am Chem Soc*. 2013;135(13):5161-5166. doi:10.1021/ja312310g

49. Smith AJ, Thomas F, Shoemark D, Woolfson DN, Savery NJ. Guiding Biomolecular Interactions in Cells Using de Novo Protein-Protein Interfaces. *ACS Synth Biol*. 2019;8(6):1284-1293. doi:10.1021/acssynbio.8b00501

Chapter 3

Comprehensive Experimental Analysis of Functional Diversification Across Seven Hundred Million Years of bZip Evolution

Title: Comprehensive Experimental Analysis of Functional Diversification Across Seven Hundred Million Years of bZip Evolution

Authors: W. Clifford Boldridge[1§], Georg K. A. Hochberg[2§], Jonathan Lee[3], Sriram Kosuri[1,4], Joseph W. Thornton[2,5*]

Author Affiliations:

[1] Department of Chemistry and Biochemistry, University of California, Los Angeles, CA, USA

[2] Department of Ecology and Evolution, University of Chicago, IL, USA

[3] Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, CA, USA

[4] UCLA-DOE Institute for Genomics and Proteomics, Molecular Biology Institute, Quantitative and Computational Biology Institute, Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, and Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, CA, USA

[5] Department of Human Genetics, University of Chicago, IL, USA

§ Authors have contributed equally

* Correspondence should be addressed to J.W.T. Email: joet1@uchicago.edu

Abstract:

Changes in protein-protein interactions (PPIs) are a major cause of phenotypic diversity. How the tinkering of evolution can lead to rewiring PPIs is less clear. Although gene duplication and subsequent divergence are well established theoretically, we lack detailed experimental knowledge to show how PPIs can change over the course of evolution. The dearth of experiments has led to a wide array of competing theories about the tempo, precision and evolutionary processes driving changes in PPIs. Here, we examine a simple PPI network of two paralogs, descended from an ancestral homodimer, which do not heterodimerize. We used ancestral sequence reconstruction and a high-throughput two-hybrid system to empirically characterize more than 65,000 interactions in the PAR/E4BP4 family of bZips proteins. We find specificity is acquired by more than half the duplications on our tree in a rapid, though not immediate, process. However, perhaps because of the speed with which it occurs, loss of interaction with a partner paralog does not necessitate broad rewiring of other unrelated interactions. We find the acquisition of specificity is permanent and even subsequent gains of specificity maintain the prior acquisitions of specificity. Finally, we develop a novel empirical test which finds that the loss of interactions between newly duplicated proteins is not driven by selection, supporting the hypothesis that after duplication proteins are released from evolutionary constraint.

Introduction:

To explain diversity in organismal phenotypes requires explaining diversity of function in homologous proteins. As most proteins are regulated to some extent by protein-protein interactions (PPIs), one method of creating functional diversity is manipulation of protein-protein

interaction networks[1]. Though remarkably stable globally[2,3], small changes in PPI networks have been linked to evolutionary adaptation[4] and disease[5–7]. While gene duplication is commonly thought[8] to add to genomes and subsequent interactions, there must be a balance with loss of interactions to maintain the constant properties of protein interaction networks[2,9]. Our ability to understand how these changes occur is severely limited because up to now we have either characterized PPIs within species[10–16] or between a handful of species[17–19].

Moreover, to study the evolutionary dynamics of how protein-protein networks change we must look at the ancestral states. Through statistical inference of ancestral sequence states, ancestral sequence reconstruction (ASR) allows characterization of ancestral properties and functional determinants[20,21] but its application to understanding functional diversity in PPIs has been limited. Previously, works using ASR to investigate PPIs have characterized a handful of proteins[22–25] and moderate sized protein networks[26,27]. However, even for simple networks of PPIs, we lack a high-resolution characterization which is necessary to understand the tempo, mechanism and processes underlying evolution.

Though many proteins undergo some change of interactions in the course of evolution, the most drastic phenotypic effects can be linked to transcription factors[28]. bZips are a class of dimerizing transcription factors, that have diversified from twelve bZips[29] in the ancestor of metazoan to fifty-three[17] in humans while maintaining strict patterns of dimeric specificity between member proteins[30]. Though substantial changes between species in the network of bZip interactions have been documented[17,31,32], there has been no dissection of the genetic causes of these changes.

Here we test interactions across the PAR and E4BP4 families of bZips. As bZips, the PAR/E4BP4 family are transcription factors, and they have been implicated in myriad processes

including circadian rhythm[33], haemopoietic development[34] and metabolite detoxification[35]. Notably, PAR/E4BP4 proteins all bind the same DNA sequence, but which protein is bound can have opposing effects on gene expression[36]. As a family descended from an ancestral homodimer that does not co-assemble, the PAR/E4BP4 family provides a simple network of interactions to investigate how evolution can drive acquisition of novel specificity (Figure 3.1A). We use ancestral sequence reconstruction to characterize the evolutionary dynamics of PAR/E4BP4 back to the human-cnidarian ancestor. With a high-throughput two-hybrid system we characterize more than 65,000 potential interactions and find that specificity—where the ancestor homodimerizes but two descendant proteins stop heterodimerizing—arises from most duplication events. We find the process of gaining specificity starts immediately, but takes some time to resolve and specific proteins do not necessarily have a drastic rewiring of interaction profiles. To characterize the evolutionary mechanism of specificity acquisition, we develop an experimental test of selection and find no evidence for direct selection as the process driving the loss of interactions between paralogs. Finally, the loss of all heterodimeric interactions is permanent.

Results:

We inferred a phylogeny of 171 PAR and E4BP4 proteins using the highly conserved bZip domain (Figure 3.1B, see Methods). All extant PAR and E4BP4 homologs are descended from a protein that underwent duplication in the ancestor of humans and cnidarians, thus we sampled all major deuterostome, protostome and cnidarian clades. In total, we sampled 52 species (Figure 3.S1), including six species in a poriferan outgroup. Subsequent duplications occurred

throughout history and our tree contained a total of 21 duplications, each of which could potentially lead to non-interacting daughter proteins. The extent to which this is the case is unknown, however, as previous work has only been characterized interactions in a handful of species[17]. We performed ancestral sequence reconstruction (see Methods) on this phylogeny, and inferred all nodes to identify 258 unique sequences across our tree.

Massive experimental analysis of PAR/E4BP4 PPIs

To characterize the 66,548 potential protein-protein interactions on this tree we used the Next-generation bacterial two-hybrid (NGB2H) system, which allows rapid multiplexed measurement of tens of thousands of researcher-designed interactions in a single experiment[37]. In brief, the NGB2H system (Figure 3.1C) consists of an inducible bacterial two hybrid which replaces the standard colorimetric reporter readout with a transcribed DNA barcode that uniquely identifies the hybrid protein pair (the mapping of DNA barcode to hybrid proteins occurs with next generation sequencing at an early cloning step). Hybrid proteins that interact produce cAMP, which drives expression of a Lac promoter with a reporter gene that contains the DNA barcode in the 3' UTR. Relative barcode abundance can then be obtained by next-generation sequencing to quantify interaction strength (see Methods).

We obtained high-quality measurements on 65,892 distinct interactions and calculated two primary measurements for each protein pair: the Interaction Score (IS), a numeric measure of relative gene expression and a Classification Score (CS), a statistical measure relative to protein pairs containing indels that classifies each pair as an interaction, a weak interaction or a non-interaction (see Methods, Figure 3.S2). The NGB2H system performed well according to a variety of metrics (Figure 3.S3), including highly correlated biological replicates, strong

agreement the published melting temperatures of a previously characterized library of coiled-coils[38] and clear signal when hierarchically clustered. Classification produced 10,193 interactions, 10,700 weak interactions and 29,488 non-interactions, when mapped on to the tree. To understand the underlying data structure better, we used *k*-means clustering on the raw Interaction Scores. Using the seven clusters (Figure 3.S4) we identified several major phylogenetic groups when plotted on the tree (Figure 3.1D). Notably, each of three protostome PAR clades were majority clustered exclusively (purple: HLF, red, blue), and the majority of E4BP4 formed one cluster (orange), though this did not include cnidarian E4BP4 or some arthropod E4BP4 (gray nodes were not measured). As expected there was little overlap between E4BP4 and PAR, with the exception of cnidarian homologs, for which the majority clustered in our broadest group (green).

Specificity is commonly acquired

We sought understand how specificity—the loss of heterodimeric interactions between descendants of a homodimer—could occur in our data set, so we looked at pairs of time matched paralogs (Figure 3.2A, Top). Specifically we took homodimerizing ancestors that underwent duplication and identified when descendant paralogs lost those interactions for two stringency levels: when descendant paralogs have weak interactions and when they have non-interactions. Although there are numerous mechanisms by which proteins could stop interacting—pseudogenization, different temporal or tissue specific expression profiles, or a biochemical loss of interaction—it is only in the same species that the loss of an interaction could matter. Thus, we looked at 518 paralogous interactions, in both extant and extinct species. When visualized on our phylogeny (Figure 3.2A, Bottom), these interactions again show strong phylogenetic signal,

with many interactions within clades, but few interactions between clades. Unexpectedly, we found that even duplication nodes with relatively few characterized descendant paralogs contained diversity in descendant interaction types (pie graphs).

To better understand where and how specificity was achieved, we plotted the evolutionary path along branches of our tree from the duplication node to the first non-interacting daughter paralogs (Figure 3.2B) for non-interacting stringency (Figure 3.S5 for weak interacting stringency). Overall we found that, either eight of the sixteen duplication nodes for which we characterized descendant paralogs acquired specificity (non-interaction stringency) or eleven of sixteen possible duplication events acquired specificity (weak interaction stringency) Though we anticipated our most ancestral duplication event would acquire specificity, we found that specificity was gained in duplication events across our tree. In fact, with the exception of cnidarian PAR duplications, all duplications for which we measured more than three descendant paralogs acquired specificity at one of our stringency levels, suggesting this is a highly common occurrence even within a protein subfamily. We also found that with the exception of one (non-interaction stringency) or three (weak interaction stringency) cases, specificity was gained prior to currently extant species, implying this is a constantly occurring process.

Specificity is quickly, though not immediately acquired

To characterize the tempo of specificity acquisition we characterized three temporal metrics on our tree. First, we calculated the number of paralog pairs that had descended from a duplication event (Figure 3.2C). This showed that specificity at non-interaction stringency overwhelming took more than one descendant paralog node, which means that specificity takes a non-zero amount of time to appear, and implies our tree has the resolution necessary to identify

when these changes occurred . However, at the weak stringency level, specificity was overwhelming gained by the first descendant paralog. Taken together this implies that the weakening of heterodimeric interactions starts soon after duplication but takes significant time to fully resolve. Characterizing the branch lengths it took to gain specificity provided a complementary result (Figure 3.2D). While specificity at weak interaction stringency had been gained with all branch lengths < 0.9 (substitutions per site), three gains of specificity at non-interaction stringency had branch lengths > 0.9 (substitutions per site), indicating the progressive nature of gaining specificity. As the branch lengths over which specificity was gained were relatively short compared to the length of our tree where the median distance between paralogs is branch length = 2.55 (substitutions per site), and because most of duplication nodes had a limited number of descendant nodes (median seven descendant paralogous interactions), it was unclear if the time to gain specificity is due to sampling the available descendants or if it occurs faster or slower than expected. Thus, to determine if this was faster than would occur from a random selection of descendant paralogs on our tree, we performed a permutation test on all branches that gained specificity (Figure 3.2E). We found for both weak interaction stringency and non-interaction stringency that the branch length average was significantly longer than the average of the branches that did gain specificity ($p < 0.01$ weak interaction stringency, $p = 0.02$ non-interaction stringency). Thus, though specificity is not gained immediately, it is faster than would occur from chance alone.

Homodimer loss is often concomitant with specificity gain

Our data suggests an unexpected mechanism of specificity gain, that is a weakening of homodimeric interactions. Although ancestrally the PAR/E4BP4 family has homodimerized, we

found that while acquiring specificity many proteins lost this interaction as well (Figure 3.2F). Notably, while at weak interaction stringency a plurality of paralogs still homodimerize, the majority of paralogs pairs at non-interaction stringency have lost one homodimer. This implies a mechanism where one paralog loses its self-interaction which facilitates the loss interaction with the highly similar paralog. Importantly, these are not non-functional proteins; even in the most extreme example where neither protein homodimerizes, the descendant paralogs interact with more than eight percent and weakly interact with eighteen percent of proteins on our tree (Figure 3.2G).

Vast rewiring is not required for specificity gain

We next sought to understand the how targeted the gain of specificity is—is the only interaction lost the partner paralog, or do the proteins interact with substantially different sets of partners. To do so, we calculated a metric, which we call Interaction Correlation, using the two paralogs which gain specificity and took the Pearson correlation of each partner's Interactions Scores with all other proteins assayed. This provides a characterization of the sequence space that is most likely to interact with these proteins, while also allowing highly differential measurements. We found that acquisition of specificity did not necessarily entail drastic rewiring of interactions (Figure 3.3A). Notably, eight specific paralogs had correlations greater than $r = 0.5$ at weak interaction stringency and three at non-interaction stringency, while our median Interaction Correlation was $r = 0.05$, which implies interactions can gain specificity without changing many other interactions. We also noted that though there was an inverse relationship between branch length and Interaction Correlation, it was not absolute, and increasing the stringency of specificity did not always lead to a decrease in Interaction Correlation. Taken

together this further suggests that specificity can arise without dramatically changing off-target interactions.

Specificity evolves gradually

The gain of specificity is clearly a gradual phenomenon, with changes accumulating slowly on both descendant branches. This gradualism is apparent when we characterize the Interaction Correlation between proteins from duplication events and their descendent specific paralogs. We find that specificity gaining nodes are significantly more similar to the ancestral duplicated protein than its paralog at the point they gain specificity, as characterized by our Interaction Correlation (Figure 3.3B-C) with $p = 0.05$ at weak interaction stringency and $p = 0.09$ at non-interaction stringency (Wilcoxon two-sided test). This gradualism can also be seen by calculating the interaction scores at the points at which specificity are gained and the duplicated ancestor (Figure 3.3D-E), as it represents an approximately halfway point between the paralogs at which specificity was gained. At weak interaction stringency (Figure 3.3D) a majority of specific paralogs are able to interact with the their ancestral duplication event, and even those paralogs that do not interact with their duplication event still maintain at least weak interactions—that is at no point do we see weaker interactions appear on branches between specific paralogs than what the interaction between the specific paralogs themselves is. This is similar at non-interaction stringency, where a majority of specific paralogs interact with the ancestral duplication protein, and no paralog pair has more than one protein not interacting with the ancestral node

Ancestral proteins have a broader set of interactions than their descendants

It is often thought that ancestral proteins are more promiscuous[25,39], though this has recently been contested[40]. Because of its size, our data set offers unique insight into changes in specificity. Importantly, we do not test random proteins to determine specificity, as the majority of partner proteins would not be accessible by evolution. While other paths may have been possible, all proteins characterized were are evolutionarily accessible. When the number of interactions are quantified, we find that there is a slight increase in interactions for the ancestral duplication nodes compared to specific paralogs (Figure 3.3F-G) at weak interaction stringency, which increases at non-interaction stringency. This slight advantage seems likely due to the bifurcating nature of phylogenetic trees, which places ancestral nodes closer to all other nodes.

There is no direct selection for specificity after duplication

Up to this point we have strictly measured paralogs, as the proteins which evolution must act upon. However we also characterized many orthologs in our data, which exist in separate organisms and therefore can't have direct selection acting on them (Figure 3.4A). Crucially, this allows a comparison between those proteins which could undergo direct selection (paralogs) and those that cannot (orthologs), as a novel, biochemically informed selection test. Given that purely inferential selection tests[41] have been refuted by recent biochemical analysis[42], our method provides direct experimental evidence for the presence or absence of selection. To measure the effects of selection on our tree, we took phylogenetically independent samples (see Methods) of both orthologs and paralogs binned by branch length. We then characterized the fraction interacting for both weak interaction stringency (Figure 3.4B) and non-interaction stringency (Figure 3.4C). In both cases we found that percent of interactions between paralogs were not significantly different from the interactions between orthologs with $p = 0.94$ at non-interaction

stringency and p = 0.46 at weak interaction stringency (Wilcoxon signed-rank test, two-sided). Thus we reject the hypothesis that direct selection is acting on these bZips after duplication. This largely fits within the Duplication-Degeneration-Complementation (DDC) paradigm[43], where after duplication selection is relaxed and proteins slowly accumulate mutations that eventually lead to descendants retaining a subset of the ancestral functions. In this case, however, the new functions are the same as the ancestral function (dimerization) but the potential interactions of the descendant genes has been partitioned to exclude each other.

Acquisition of specificity is permanent

Given that neutral evolution does not necessarily keep proteins from interacting after specificity has been gained, it is an open question how permanent acquisitions of specificity would be. To quantify this we identified all paralogs that descended after specificity was acquired and found that interactions can be regained at weak interaction stringency (Figure 3.4D). However, at non-interaction stringency descendants very rarely are able to reform interactions (Figure 3.4D) with zero paralogs having full interactions. This was true even when non-paralogous interactions were consider post-specificity gain, with only a handful of interactions present. Together, this shows that specificity gain becomes increasingly strict, with marginal losses of heterodimeric capacity able to be rescued while total loss of heterodimeric capacity is irreversible.

The lack of selection and the rarity of interactions returning post specificity gain presents a model where even proteins under many constraints, such as bZips, are able to evolve specificity in non-overlapping ways (Figure 3.4E). Particularly, the interaction space (the group of proteins a given protein interacts with) is disjoint for proteins after specificity has been

gained, regardless of new duplications opening up new spaces. This is similar to other work which has noted that sequence space is sparsely populated[44].

Discussion:

Here we performed the first comprehensive characterization of the evolutionary development of a PPI. With a nearly complete map of the interactions among the E4BP4/PAR family back to the vertebrate/cnidarian ancestor we provide an in depth examination of how a homodimeric protein can gain diversity in function among its descendants. Unexpectedly, we found specificity occurs in a majority of duplication events, implying that it may not be a relatively rare event, but rather a bias from lack of characterization. The gain of specificity is a gradual process with many contributing substitutions, that starts occurring immediately post-duplication. We found that once specificity was achieved it was functionally irreversible, no matter where on the tree it occurred. Finally, we developed a novel experimental measure to test for selection by comparing ortholog and paralog rates of interaction. Using this, we were unable to find any selection leading to specificity gain.

The fact that we see specificity arise multiple times across our tree, but never find subsequent gains of specificity interacting with the descendants of previous gains of specificity implies that there are many different ways to achieve specificity, without overlapping in interaction space. Although it has previously been noted that interaction space is sparsely populated[44], our work shows that it is not just all of interaction space that is sparsely populated, but that the portion that is evolutionarily accessible is also sparsely populated. Coupled with the lack of selection in our study and the speed with which proteins gain specificity, this suggests

that even for highly constrained protein interaction space is so vast that once an interaction is lost it cannot be found again.

The pressures on a genome post-gene duplication has engendered much speculation. Simple subfunctionalization or neofunctionalization may not be possible when the duplicated protein homodimerizes, and instead paralog interference may result which must be resolved[45]. The pervasiveness of specificity among descendant paralogs in our study suggests that paralog interference maybe an issue, but find a reliance on protein sequence to do so. Although this disagrees with the predominant thought of gene regulatory networks regulating duplicated gene[46], and we cannot rule out regulatory effects, clearly PPIs in the E4BP4/PAR family are avoiding conflicts.

Finally, here we developed a system that opens up a new way to test how novel biochemical phenomena arise. Rather than testing a few extant orthologs we were able to test proteins across a phylogeny in a high-throughput manner, allowing us to ascertain insights in to the evolutionary process that would otherwise be unobtainable. We note that this approach could be used to study any other protein property that has been adapted to a high-throughput assay for gaining evolutionary insights into functions such as RNA or DNA binding or activation, apoptosis or fluorescence[47].

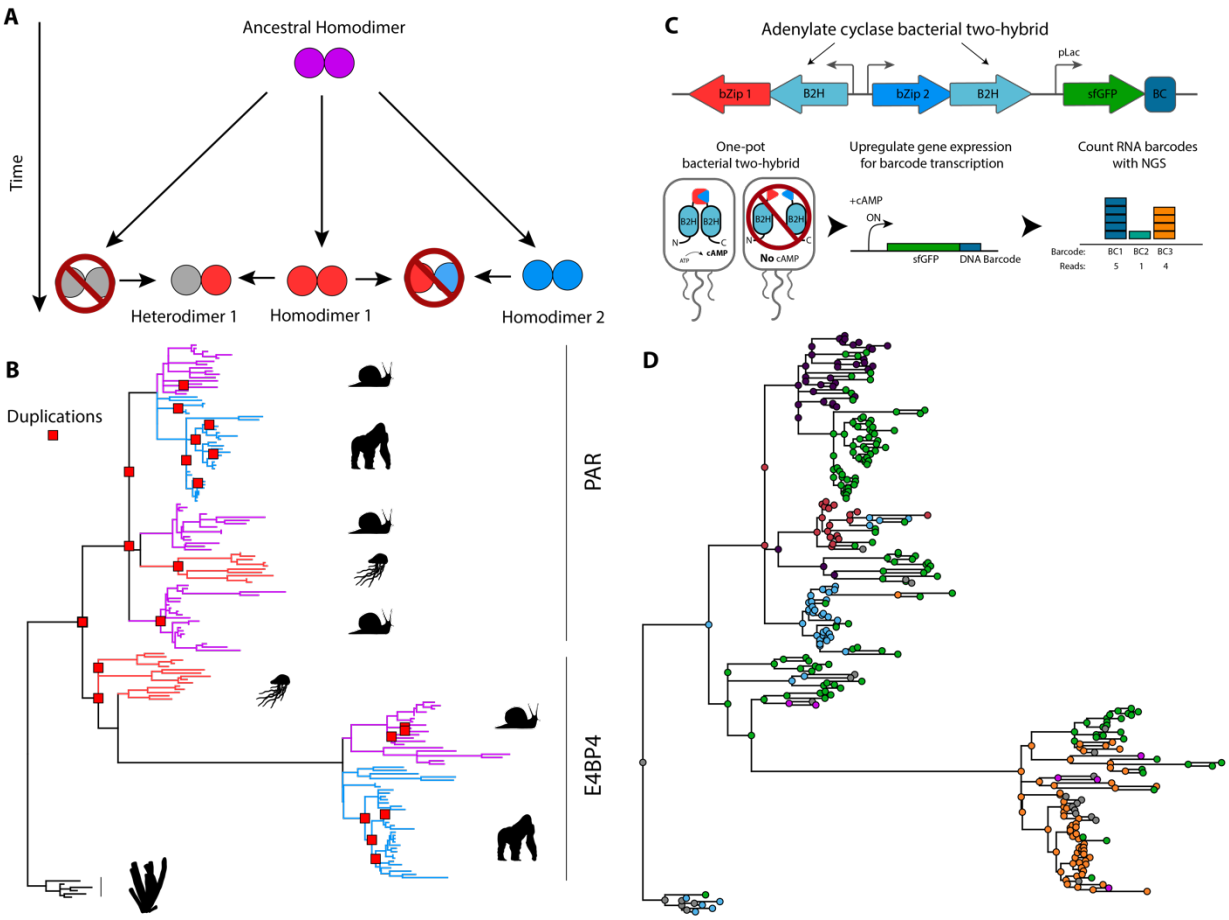

Figure 3.1) Characterization of 66,000 E4BP4 and PAR protein-protein interactions: A) Schematic of a hypothetical protein family, descended from an ancestral homodimer. Extant sequences (lower row) display a diversity of functions including a losses of homodimerization and heterodimerization. B) Phylogeny of E4BP4 and PAR, going back to their last common ancestor of vertebrates and cnidaria. Major classes are shown in silhouette for deuterostome (gorilla), protostome (snail), cnidaria (jellyfish) and outgroup porifera (sponge). C) Schematic of the NGB2H system. Top: illustration of the major components of the bacterial two-hybrid (B2H), with sfGFP reporter and DNA barcode (BC). Bottom: interacting hybrid proteins reconstitute adenylate cyclase activity and produce cAMP which drives expression of pLac and the corresponding DNA barcode which is then quantified through next-generation sequencing.

D) K-means clustering on Interaction Scores plotted on the phylogeny. Seven clusters, illustrated by different colored dots, broadly correspond to several major clades.

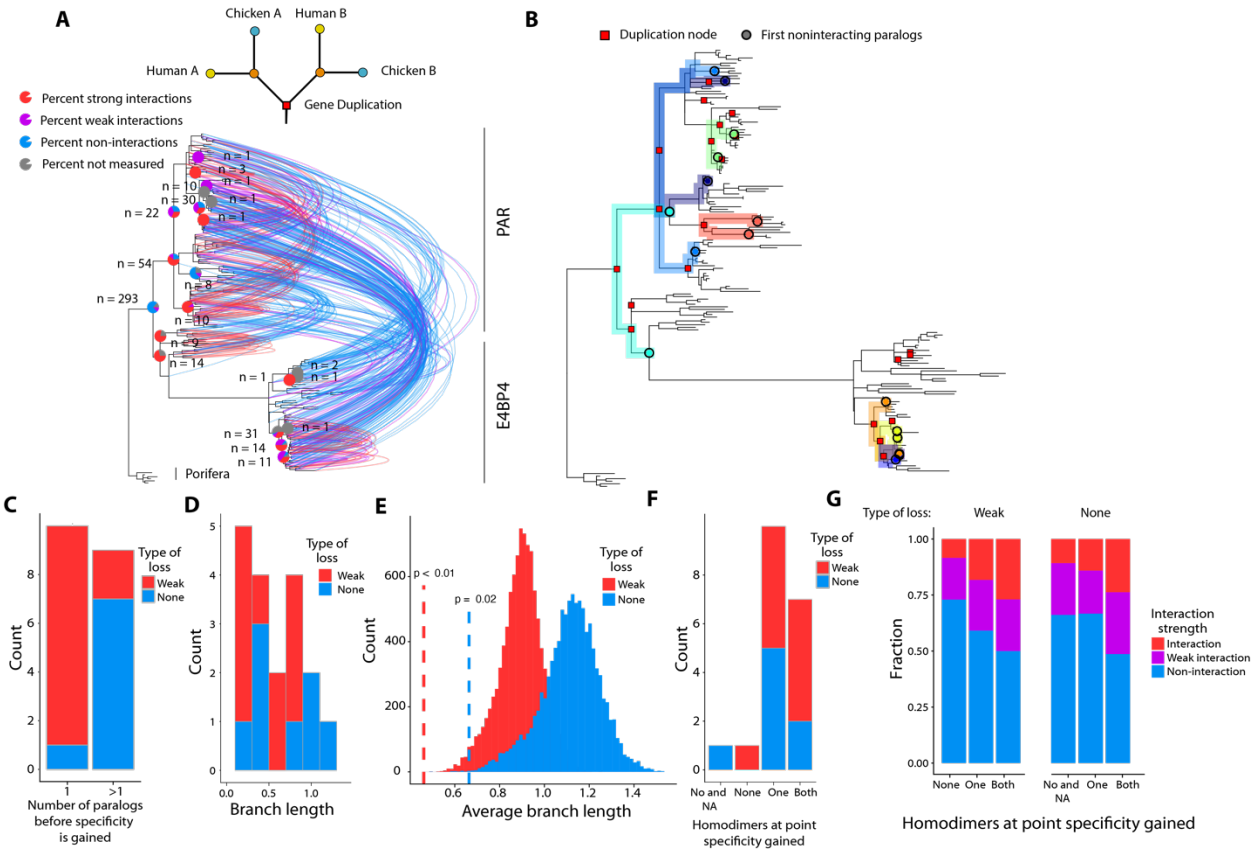

Figure 3.2) Paralogs gain specificity many times: A) (Top) Schematic illustrating paralogs examined in this study. Diagram includes both extant (yellow, teal) and extinct (orange) paralogs. (Bottom) Phylogeny of illustrating interactions between paralogs. Each arc is colored to represent the classification of each interaction. Pie charts at each duplication node quantify the interactions of descendant paralogs and list the number of interactions in the pie chart. B) Traces showing the evolutionary path from duplication to the first point at which specificity is achieved. Each color represents a unique duplication (red square) that evolves (colored rectangles overlaying branches) to the first point at which paralogs no longer interact (filled circles). C) Number of paralogs required for specificity to be gained. D) Total summed branch lengths between paralogs that have gained specificity. E) Permutation tests showing the distribution of averages of randomly sampled paralogs from duplications that have gained specificity. Dotted

lines indicate the measured average of nodes where specificity was gained. F) Homodimerization of descendant paralogs at the point specificity was gained. Colors represent by different levels of stringency for loss. NA was not measured. G) Fraction of interaction classifications with all measured proteins by the homodimers at the point specificity was gained at weak interaction stringency (left) and non-interaction stringency (right).

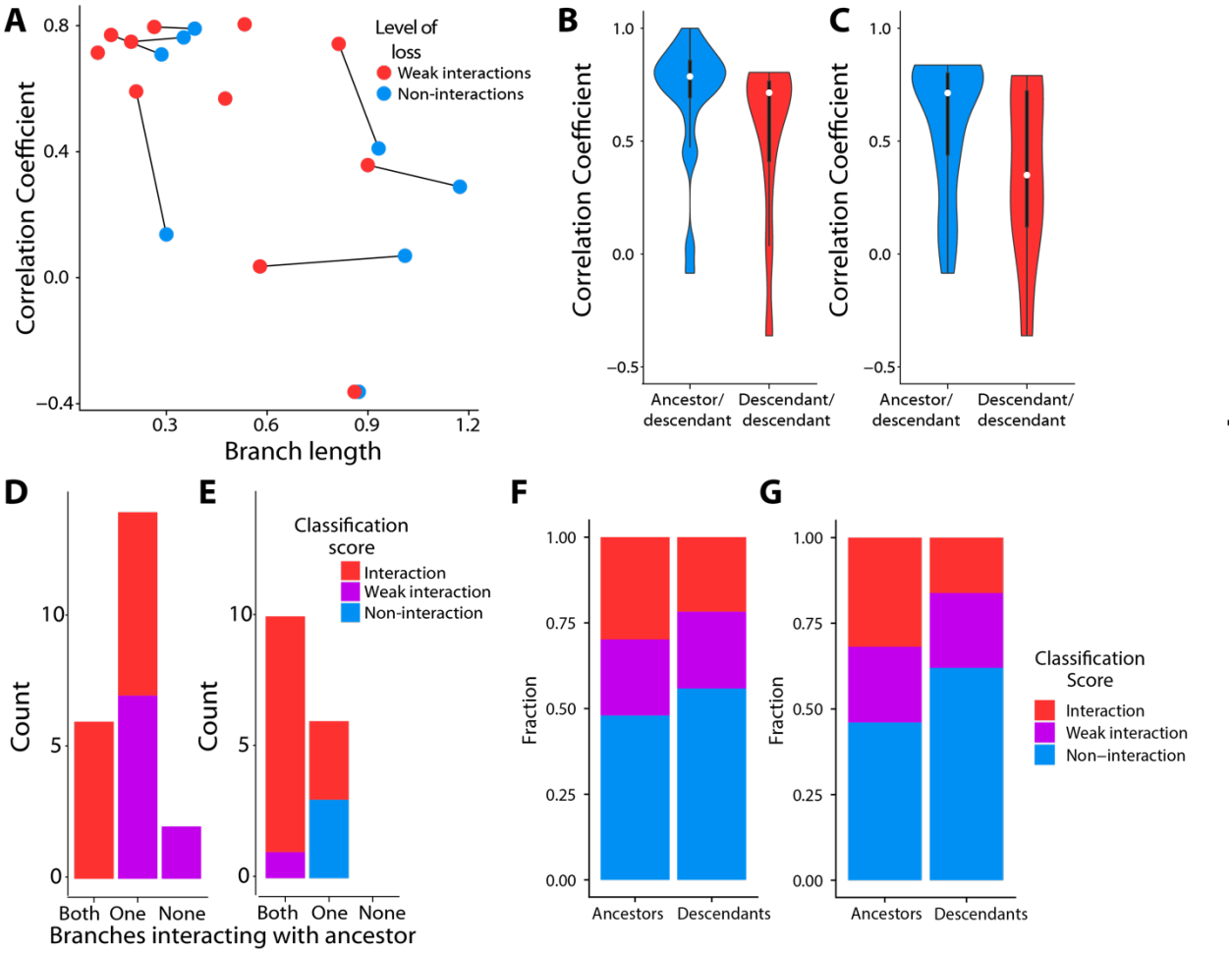

Figure 3.3) Gains of specificity do not necessitate vast rewiring: A) Pearson's correlation coefficients between interaction score of proteins at the point specificity is gained. Lines connect different stringencies of specificity gains from the same duplication event. B-C) Pearson's correlations of between ancestral duplications and descendants that have gained specificity. B) Weak interaction stringency for specificity gains. C) Non-interactions stringency for specificity gains. D-E) Interactions between points where specificity was gained and the duplication node. Colors indicate the interaction strength. D) Weak interaction stringency for specificity gains. E) Non-interaction stringency for specificity gains. F-G) Composition of all interactions for

ancestor duplication nodes and descendants that have gained specificity F) Weak interaction stringency for specificity gains, G) Non-interaction stringency for specificity gains.

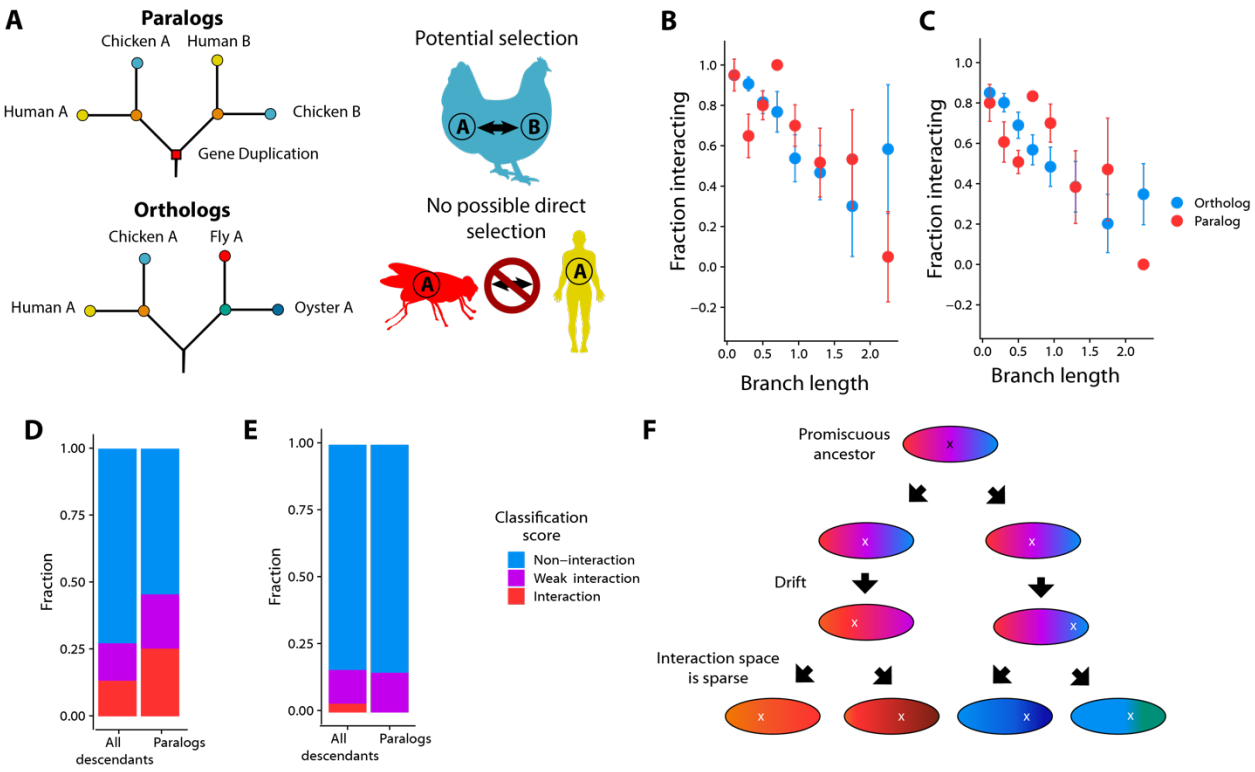

Figure 3.4) Specificity is not driven by direct selection and leads to an open ended space: A) Schematic of the differences between orthologs and paralogs. Paralogs (Top) exist in the same organism and interactions can be selected for or against. Orthologs (Bottom) exist in separate organisms and interactions between them cannot be directly selected for. B-C) Odds of orthologs or paralogs interacting depending on branch length. B) Weak interactions are lost interactions. C) Non-interactions are lost interactions. D-E) Fraction of descendants in each class after specificity has been gained D) At weak interaction stringency E) At non-interaction stringency F) Potential mechanism for divergence and specificity gain. X's denote a protein descending along a tree (Black: prior to duplication, White post duplication). Ovals denote the protein's potential interaction partners. Different colors denote different interaction partners.

Methods:

Sequences data and alignment

We initially collected 236 bZip sequences from across metazoa in the PAR, E4BP4 and closely related paralog families. These sequences were aligned with MUSCLE[48], and the data alignment curated by hand. We used RaxML[49] to infer the best fit evolutionary model on this tree using the PROTGAMMAAUTO feature, and found it to be LG[50] with ML base frequencies. We then inferred the tree using PhyML[51] using the LG model, a 4 category gamma model and the amino acid frequencies calculated in RaxML. aLRT[52] support values and bootstrap replicates were calculated in PhyML. TBE[53] support values were calculated using BOOSTER. For the ancestral sequence reconstruction, we next trimmed the alignment down to just the bZip domain and adjacent DNA binding domain. We then constrained the species relationships within each paralog to conform to a unified species phylogeny. This was necessary in order for us to be able to time particular ancestral paralogs relative to each other so that we could infer groups of ancestral paralogs that would have existed in the same ancestral genomes. To do this we removed taxa until we arrived at a species phylogeny on which all groupings are supported by literature available at the time the tree was constructed in 2018. It is possible that our constraining created some incorrect relationships, if there are extensive histories of duplication and lineage specific losses, but we saw no evidence for this on our larger ML phylogeny. After pruning, we had 95 PAR homologs, 70 E4BP4 homologs and 6 poriferan orthologs across 52 species which included 6 poriferans, 6 cnidarians, 21 protostomes and 19 deuterostomes.

For ancestral sequence reconstruction we first optimized the branch-lengths in PAML[54], and then calculated ancestral sequences also in PAML. Posterior probabilities of each position were for

each protein were calculated using the LAZARUS python wrapper for PAML. The maximum a posterior sequences for each node and the sequences of each tip are available upon correspondence with the author.

Gene tree reconciliation

To reconcile the gene tree with our species phylogeny, we used NOTUNG[55]. This program assigns the most parsimonious history of gene duplications and losses to each gene, based on the species tree we defined. The number of losses are almost certainly an overestimate, because we did not always include all paralogs from every organisms. Particular paralogs were excluded if they had unstable positions on the phylogeny, or if their sequence was not complete, or if we could not unambiguously assign the exon structure for sequences that derived from DNA data. The number of losses should thus not be interpreted as meaningful on our reconciled gene tree.

Timing of ancestral proteins

To find pairs of ancestral proteins in our dataset that would have existed in the same genome, we used our species phylogeny. Nodes on our gene phylogeny that correspond to the same speciation event on the species phylogeny represent proteins that would have existed in the same genome.

Oligonucleotide design

Every full-length sequence (258 unique sequences) on the tree was designed to fit oligonucleotides that would be cloned into each half of the two-hybrid (X and Y halves). For those to be cloned into the X position sequences consisted of one of five 15bp forward flanking

subpool primer, a 15 bp forward pooled primer, 128 bp reverse complement of the coding sequence of protein sequence, a BspQI site for scarless cloning with the T25 fragment, a BbsI site for ligating to corresponding Y oligonucleotides, a 34bp spacer and one of five 15 bp reverse flanking subpool primer for a total of 230 nucleotides. Likewise, oligonucleotides encoding sequences to be cloned into the Y position consisted of one of five 15 bp subpool primer, a 25 bp spacer, a BbsI site for ligating to corresponding X oligonucleotides, a BtsaI site for scarless cloning with the T25 fragment and promoter, 128 bp of protein coding sequence, a BsaI site for scarless cloning of the T18 fragment and two-hybrid reporter, a reverse pooled primer and one of five reverse subpool primers. Each oligonucleotide is encoded in one of three codon usages for each bZip. A total of 258 sequences * 2 orientations * 3 codon usages * 5 subpools = 7740 unique oligonucleotides were ordered as an OLS pool from Agilent using their high-fidelity process. All protein coding oligonucleotide sequences are available upon correspondence with the author.

Reagents used

All reagents used for cloning and NGB2H assay were purchased from the following vendors. All qPCR was performed using KAPA SYBR Fast 2x Master Mix (KAPA KK4601). Colony PCR for characterization of cloning efficiency was performed with Apex TAQ Red Master Mix 2x (Genesee Scientific, 42-138). All other PCR, was performed using high-fidelity polymerase NEBNext Q5 Hotstart HiFi PCR Master Mix (NEB M0542L). Restriction enzymes, ligase and phosphatase for cloning were all ordered from NEB (BbsI-HF R3539L, AscI R0558L, EcoRI-HF R3101L, BspQI R0712L, BstαI R0667L, BsaI-HFv2 R3733L, High concentration T4 Ligase M0202M, rSAP M0371L). Nucleic acids were prepared with kits from Qiagen: vector with

Qiagen Plasmid Plus Maxi Kit (Qiagen 12963), DNA from the NGB2H assay with QIAprep Spin Miniprep Kit (Qiagen 27106), RNA from the NGB2H assay with RNeasy Mini Kit (Qiagen 74106) with on column DNase digestion (Qiagen 79254) and concentrated with RNeasy MinElute Cleanup Kit (Qiagen 74204). Both PCR and gel DNA cleanup were purified with Zymo (Zymo D4014 and D4008) and the removal of biotinylated nucleic acid products was performed using Dynabeads M-270 Strepavidin (Themofisher 65305). Reverse transcription was performed with Superscript IV (ThermoFisher 18090050) with the addition of RNase A (Qiagen 19101). All DNA samples were confirmed to be monodispersed on an Agilent Tapestation 2200 using D1000 screentape (Agilent 5067-5582). Cloning was performed into NEB 5-alpha (NEB C2987I) or custom strains, as indicated.

Library cloning

10pM High-fidelity OLS pools were resuspended in 25uL EB. Samples were diluted 20x in ddH2O and used as template for qPCR with subpool primers oSK538-oSK542 and oSK619-oSK623 for X oligonucleotides and oSK543-oSK547 and oSK624-oSK628 for Y oligonucleotides. All subpool samples showed robust exponential amplification with Cqs between 6 and 13 cycles. A high-fidelity PCR was performed in triplicate for each subpool, for a number of cycles which maintained exponential amplification to avoid potential biases in the library composition. Replicates were pooled, cleaned up and digested with BbsI-HF at 1ug scale for 3 hours. Digestions were run on a 4% agarose gel, and the band containing the protein coding sequence was extracted. Matching subpools (ie, subpool-1 for the X-containing oligonucleotides and subpool-1 for the Y-containing oligonucleotides) were mixed and ligated with T4 DNA Ligase overnight at 50ng scale. Samples were cleaned up and qPCR performed with a 1:5

dilution using subpool primers oSK603 and oSK682. All subpools, exhibited exponential amplification, though for fewer than 12 cycles. The PCR was repeated in triplicate with high fidelity polymerase for two cycles fewer than the Cq in the qPCR. The triplicates were pooled and ran on 3% gel and cleaned up. Each subpool was quantified on a tapestation and equimolar fractions of each were pooled and diluted 100-fold. qPCR of the merged subpools was performed with oSK358 and oSK727 to attach restriction enzyme sites (AscI to the n-terminus and BsaI and EcoRI to the c-terminus) and a 20bp random DNA barcode to the X-Y containing constructs. Samples exhibited exponential amplification for nine cycles, so a high fidelity PCR was performed in triplicate for seven cycles and pooled. This barcoded product was run on a 3% agarose gel and extracted. Samples were then digested with AscI and EcoRI-HF for four hours at 3ug scale. Freshly prepared pSK33 was also digested with AscI and EcoRI-HF, with the addition of rSAP, for three hours, before being ran on a 1% gel and the digested band extracted. pSK33 and the barcoded insert were ligated with T4 DNA Ligase at 400ng scale at RT for two hours. Samples were cleaned up into 6uL ddH2O and 1uL was electroporated into 25uL NEB 5-alpha. After 35 minutes recovery in 1mL SOC, samples were plated on LB agar + Kanamycin plates and grown overnight at 37C. In the morning colony PCR was performed on 16 colonies with oSK191 and oSK120 and showed 16/16 containing the insert. Approximately 30 million colonies were obtained of which 6 million were scraped from the plates, diluted to OD 0.02 in 150mL of LB + Kanamycin and grown overnight at 30C. This was then purified and used as backbone vector for cloning the T25 segment and mapping the barcodes to the hybrid proteins.

To clone the T25 segment, the backbone vector was diluted to 0.5ng/uL and amplified in qPCR with biotinylated oligonucleotides oSK720 and oSK721. The vector showed exponential amplification through 14 cycles, so a high-fidelity PCR was repeated in triplicate for twelve

cycles. The vector was cleaned up, pooled and digested with BspQI at 10ug scale overnight, before being cleaned up and digested at 10ug scale with BstαI for six hours with the addition of rSAP for the final hour. The digested vector was then purified with Dynabeads to remove undigested product, before being cleaned up again. The T25 segment was prepared by high-fidelity PCR of template pSK59 with oligonucleotides oSK694 and oSK695. Sample was run on a 1.5% agarose gel, cleaned up and digested with BspQI at 5ug scale for 6 hours. Sample was then digested with BstαI for six hours at 5ug scale before being cleaned up with Dynabeads to remove undigested product. Insert was cleaned up again and ligated with the vector at 4ug scale using T4 DNA Ligase at RT overnight. Sample was concentrated to into 6uL ddH2O and 1uL was transformed into freshly prepared electrocompetent pSK34 in four separate transformations. Transformations were resuspended in 1mL SOC and incubated at 37C for 35 minutes before being plated on LB agar + Kanamycin + Carbenicillin overnight at 37C. Approximately 4 million colonies were obtained, and colony PCR was performed on 32 colonies of which 31/32 had the correct insert. All four million colonies were scraped and resuspended in 150 mL LB + Kanamycin + Carbenicillin at OD 0.02 and grown overnight at 30C. This was then purified and used as backbone for cloning the T18 segment.

To clone the T18 segment the backbone vector was diluted to 0.5ng/uL and qPCR was performed with biotinylated oligonucleotides oSK753 and oSK754. The vector showed exponential amplification for sixteen cycles so a high-fidelity PCR was repeated in triplicate for fourteen cycles. Sample was cleaned up and pooled and digested at 5ug scale with BsaI-HFv2 for four hours with the addition of rSAP for the last half hour. Digested vector was purified with Dynabeads to remove undigested product and cleaned up again. The T18 segment was prepared by high-fidelity PCR using pSK59 as the template and oligonucleotides oSK698 and oSK202.

PCR product was run on a 1.5% agarose gel, extracted and digested with BsaI-HFv2 for three hours at 3ug scale. Digested insert was purified with Dynabeads to remove undigested product, before being cleaned up again and ligated with the digested vector at 500ng scale using T4 DNA ligase for two hours at room temperature. This was purified in 6uL ddH2O and 1uL was electroporated into 25uL of freshly prepared electrocompetent pSK34. The transformation recovered in 1mL SOC at 37C for 35 minutes. Sample was plated on LB agar + Kanamycin + Carbenicillin and grown up overnight at 37C. Approximately 10 million colonies were obtained and colony PCR showed 16/16 colonies contained the insert. All 10 million colonies were scraped, diluted to OD 0.02 in LB + Kanamycin + Carbenicillin and grown overnight at 30C. We created glycerol stocks from this overnight culture, stored at -80C and for downstream experiments one glycerol stock was thawed and subsequently discarded. Finally, to confirm that only one unique plasmid was present per cell, we sanger sequenced the barcodes of 20 colonies all of which had only one construct, and none of which shared a barcode, suggesting a relatively even dispersal of construct representation.

Library mapping

After the barcodes were attached to the hybrid proteins, but before the T25 segment was cloned in, we used an Illumina MiSeq to read through the hybrid proteins and the barcode in a single read. To map the proteins, we first performed qPCR using oSK752 and oSK193 on the vector, which showed exponential amplification through 12 cycles. A high-fidelity PCR was repeated in triplicate, samples were run on a 2% gel, extracted and pooled. Samples were quantified and shown to be monodispersed on an Agilent Tapestation, and requantified and shown to be free from salt contamination on a nanodrop. Samples were loaded into Illlumina MiSeq V3 PE 600

kit (Illumina MS-102-3003) at 18pM with a 10% spike-in of PhiX Sequencing Control (Illumina, FC-110-3001) with primers oSK751 for read 1, oSK324 for the index read, and oSK323 for read 2. Two runs were completed for 54M paired-end reads which we used our custom mapping pipeline to link barcodes to those constructs where both proteins perfectly matched reference sequences. In brief, we used BBTools to filter high-quality, non-PhiX reads, and merge the paired ends at which point we used Starcode to remove PCR errors. We then used a custom python script that identified barcodes and coding proteins, while discarding those barcodes that were too close in sequence space or chimeric. Coding proteins were then mapped to reference sequences using BBMap. In total we mapped to 3.6M unique barcodes with errorless hybrid-proteins. Each barcode had a read depth between one and 461 with a mean of 7.7 reads. Each of 66,559 protein pairs had between one and 203 barcodes uniquely identifying it with a mean of 54.5 barcodes per construct.

NGB2H Assay

After cloning, we performed the NGB2H assay. Similar to previous work, glycerol stocks of the library were thawed, and 100uL grown up overnight in 100mL EZ Rich Defined Media (Teknova M2105) with Kanamycin and Carbenicillin at 30C. We also grew up a several microliters of the previously published CC0 Library in 10mL of EZ Rich Defined Media with Kanamycin and Carbenicillin at 30C. The next morning, we inoculated two biological replicates of 100mL of EZ Rich Defined Media with Kanamycin and Carbenicillin with inducers 10ng/mL Anhydrotetracycline and 2.5uM 2,4-Diacylphlorolglucinol with 100uM Isopropyl B-D-1-thiogalactopyranoside (IPTG) with a 0.2% spike-in of the CC0 Library. Samples were incubated with shaking at 37C for six hours before being placed on an ice slurry for 15 minutes and

samples taken for RNA and DNA processing. These samples were then flash frozen and stored at -80C.

Barcode preparation

Samples were thawed for both DNA and RNA extraction. RNA was extracted with on-column DNase digestion, and concentrated to ~2ug/mL. Samples were subject to primer specific RT-PCR using Superscript IV with oligonucleotide oSK193 or oSK194, for separate biological replicates, to attach Nextera lowplex I7 indexes and P7 sequence adapters. RT-PCR was performed with the following modifications to the protocol: instead of 5ug of RNA, we used 22.5ug in 11uL and the reverse transcript step at 55C was 1 hour instead of 15 minutes, and we added 1uL of RNase A to the RNase H digestion step. After reverse transcription, samples were qPCR'd with oligonucleotides oSK730 or oSK731 and oSK200 to attach the P5 sequencing adapter and I5 indices, and no-RT controls confirmed a lack of DNA contamination. Samples showed exponential amplification through 13 cycles, so a high-fidelity PCR was repeated in triplicate for 11 cycles, samples were run on a 3% gel, extracted and pooled. DNA samples were similarly extracted, and used for qPCR with oSK732 or oSK733 and oSK195 or oSK196 for separate biological replicates to attach sequencing adapters and Nextera I5 and I7 lowplex indices. Samples showed exponential amplification through 16 cycles, so a high-fidelity PCR was repeated in triplicate for 14 cycles. Samples were run on a 3% gel, extracted and pooled.

Barcode sequencing

After purification samples were quantified on an Agilent Tapestation, which showed them to be monodispersed and concentration was confirmed on a Nanodrop. Biological replicates from both

the RNA and DNA were pooled in equimolar fractions, and sequenced on a Nextseq 500 with a Nextseq 500/550 High Output Kit v2.5 (75 cycles) (Illumina 0024906) at 1.8pM with 15% spike-in of PhiX Sequencing Control (Illumina, FC-110-3001) for 40 cycles. Primers oSK326 for read 1, oSK324 for index read 1 and oSK742 for index read 2 were spiked in to the corresponding Illumina primer wells. ~480M reads were obtained which passed filter and demultiplexed as expected.

Barcode processing and data quality

Barcodes were stripped to the first 20bp, counted and single base errors were removed with Starcode. 633,570 barcodes identified sequence-perfect hybrid proteins with more than 10 reads in the DNA samples of both replicates, quantifying the interactions of 65,892 unique protein pairs, with median depth of nine barcodes. For each protein pair we calculated an Interaction Score defined as the median(RNA counts for both replicates/DNA counts for both replicates) for each barcode. Biological replicates showed strong agreement between interaction scores, with Pearson's $r = 0.83$ (Figure 3.S3A). We suspect the correlation between replicates would have been higher had the Interaction scores not been dominated by low values near the bottom of our dynamic range. Our previously validated CC0 Library performed as expected. Though admittedly undersequenced, when using the alternative interaction score of sum(RNA barcode counts)/sum(DNA barcode counts) for those with > 8 DNA barcodes in each replicate, we found a strong correlation with the previously published Tms[38] with Pearson $r = 0.79$ (Figure 3.S3B) which is on par with our previous uses of the CC0 Library[37]. Finally when examining Interaction Scores of indels in either the X or Y-hybrid protein, we found that the indels had 95% of Interaction Scores < 0.2, compared to 0.85 for sequence-perfect pairs (Figure 3.S2A).

Pair classification

To classify each pair we compared the distribution Interaction Scores of individual barcodes of a protein pair to the distribution of indels in the X-proteins (specifically those 1bp indels that occur in the first five amino acids). The X-proteins with indels still have complete adenylate cyclase proteins, and thus should have the background level of reporter activation seen in a cell without an interaction—as the X-protein indels are assumed to be non-interacting. To compare the distribution of barcodes to the distribution of indels we performed a one-side Mann-Whitney-Wilcoxon test to test if the barcodes of protein-pair were drawn from the indel distribution. Those proteins with a p-value < 0.05 were classified as interactions, while those with greater p-values were classified as non-interactions. No correction was performed for multiple testing as non-interactions are of equal importance as interactions for our analysis. For heterodimers, we tested both orientations—that is a protein can be a hybridized to the X half or the Y half of the two-hybrid and our assay can distinguish between these options—of interactions for the vast majority of pairs. Often the classification of the two orientations disagreed. We found that this group constituted its own class as the p-values for the orientation that were classified as an interaction were significantly higher when compared to those where both orientations were classified as an interaction. Similarly, the orientation that was classified as a non-interaction had significantly lower p-values than those pairs where both orientations were classified as non-interactions (Figure 3.S2B). Moreover, when analyzed phylogenetically, the majority of these intermediate pairs appeared between interactions in both orientations and non-interactions in both orientations. Taken as a whole, we believe this indicates a true intermediate class of weak interactions, and use three classes of strong interactions (both orientations are interacting), weak

interactions (orientations disagree about whether it is an interaction), and non-interactions (both orientations are non-interacting).

Data analysis

The classified data was used for all downstream analyses. Analysis was done in Python (v3.8) and R (v4.0.3). Phylogenetic figures were generated in Python while all other figures were generated in R. Images were assembled in Adobe Illustrator (v25.1).

Phylogenetic independence

To calculate independent phylogenetic intervals we used a sampling procedure to ensure branches of the tree were not counted more than once. We first isolated those protein pairs that were paralogs or orthologs where the last common ancestor homodimerized. We then binned samples based on the branch length between the two proteins. Binned samples went through a Monte Carlo process where in each iteration a random sample was taken and all protein pairs in the bin that shared branches with the sample pair were removed. This continued until there were no more proteins in the bin at which point the average interaction classification was computed. Averages were computed for all bins and the process was repeated 20 times to collect error measurements.

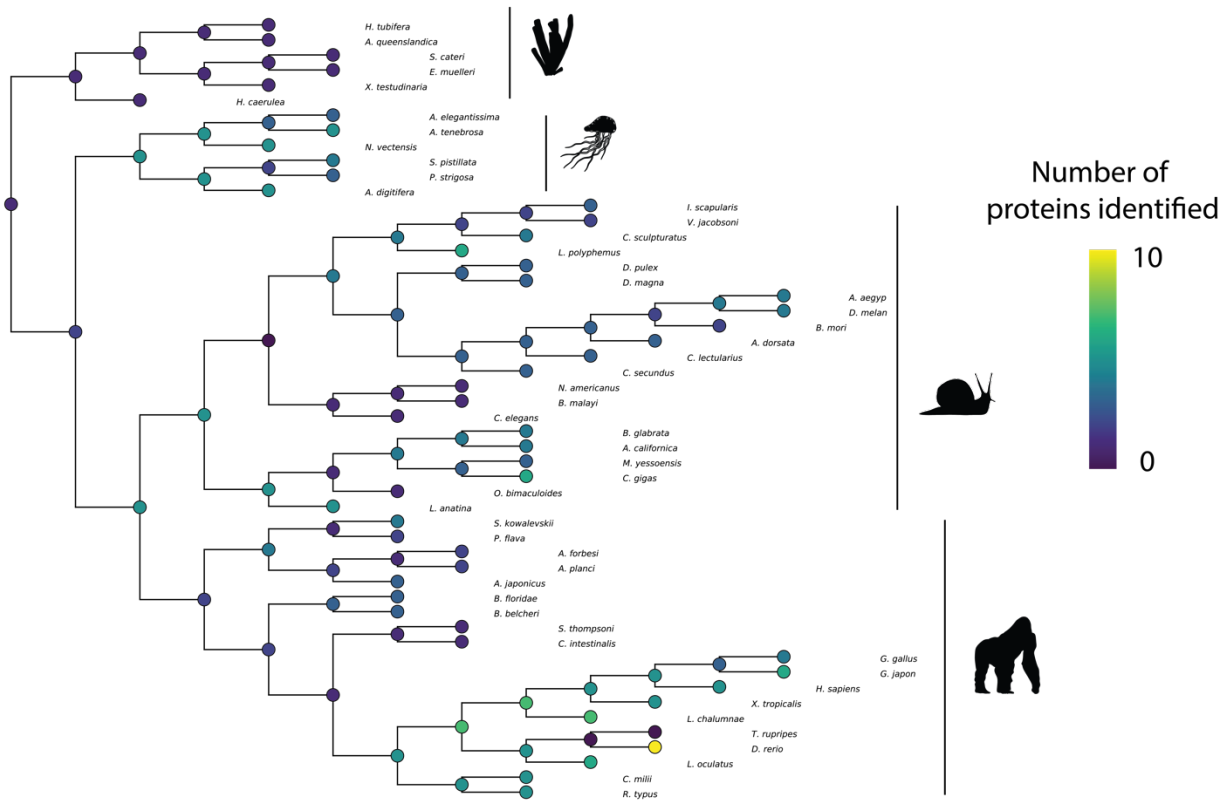

Figure 3.S1) Species tree and number of proteins present: Phylogeny containing the species used in this study. The major clades, porifera (sponge), cnidaria (jellyfish), protostomia (snail), and deuterostomia (gorilla) are marked by silhouettes. Each node is colored by the number of proteins in the species it that were included in this study.

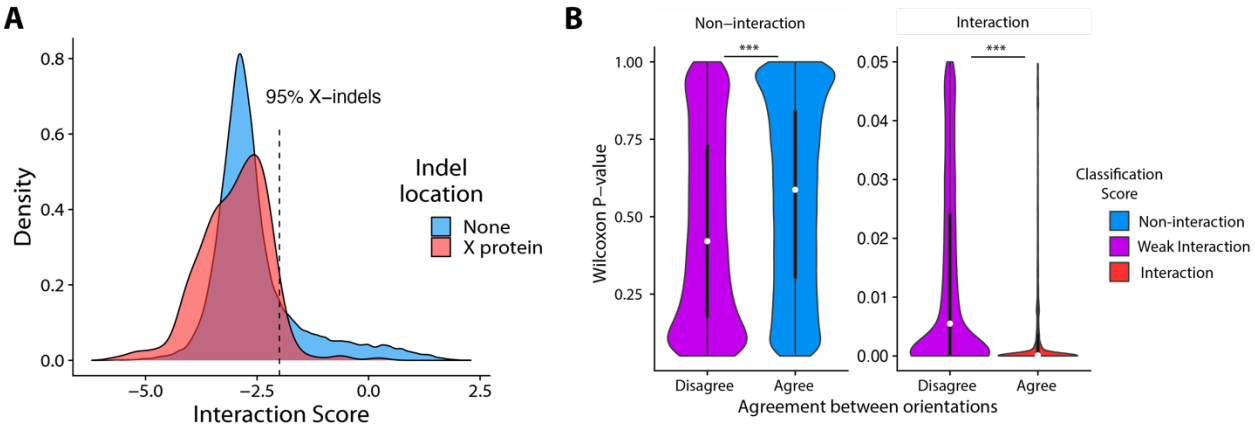

Figure 3.S2) Calculation of classification scores: A) Distribution of Interaction Scores for protein pairs with an indel in the X protein (Red), plotted against the Interaction Scores for perfectly constructed protein pairs. Dashed line indicates the significance level used. B) Distribution of P-values by Classification Score before the two orientations of a protein pair are collapsed into a single measurement. *** = $p < 10^{-160}$ Wilcoxon two-sided test.

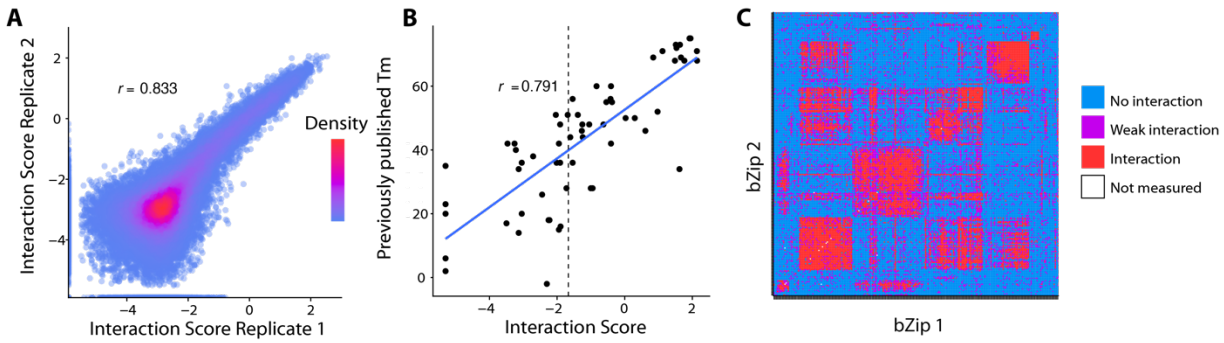

Figure 3.S3 Quality metrics of the NGB2H measurements: A) Interaction Scores of biological replicates of the NGB2H assay on the library of all proteins included in this study. Replicates correlate well with Pearson's $r = 0.833$, $p < 10^{-16}$ B) Interaction Scores of CC0 Library spike-in against their published melting temperatures. Dashed line represents the 95% cut-off of the indels used for classification. C) Hierarchical clustering of Interaction Scores collected. Individual squares represents the interaction of proteins listed along the axes. Each interaction is colored by interaction classification.

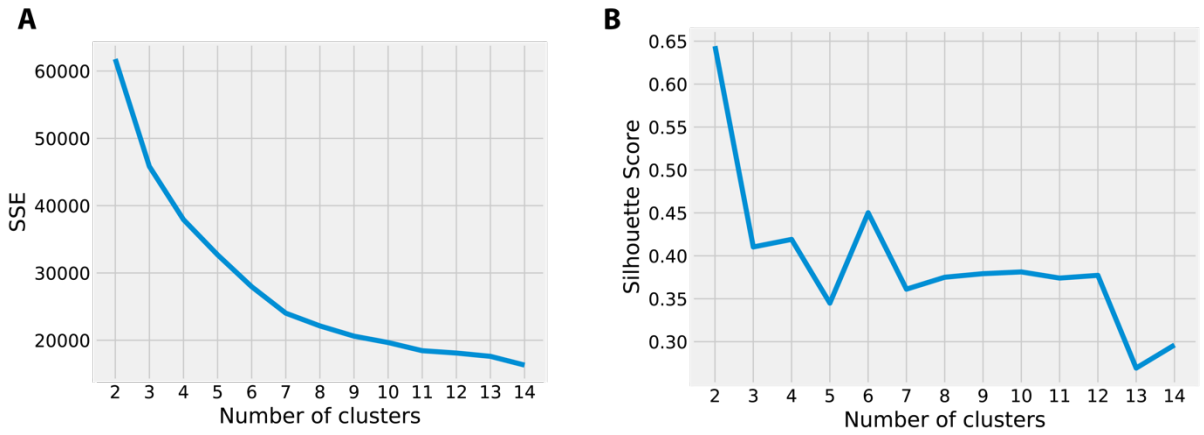

Figure 3.S4) *k*-means clustering metrics: A) Elbow plot of sum of squares error by number of clusters used. The graph has two potential elbows, at three clusters and seven clusters. B) Silhouette coefficient averages by number of clusters. Highest coefficient is at two clusters because of pseudocounts on missing data. Local maxima at six clusters suggested data was best clustered at six or seven clusters (as determined by SSE).

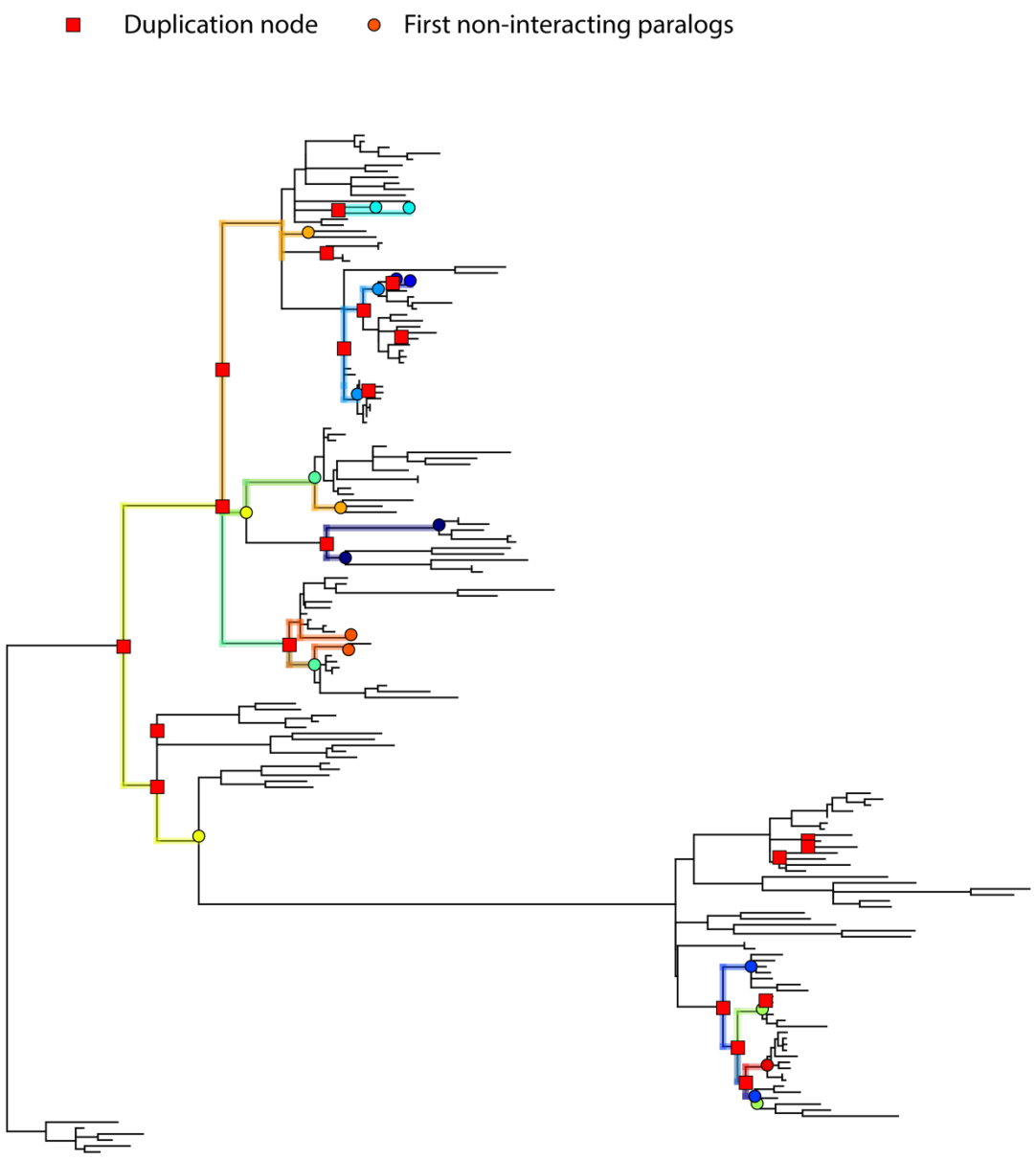

Figure 3.S5) Loss of heterodimerization after duplication for paralogs at weak stringency: Phylogeny illustrating the losses of heterodimerization where a weak interaction is considered lost. Red squares indicate duplication nodes, and those which have descendent paralogs which

have gained specificity are outlined in (a unique) color from the duplication node to the paralogs which no longer interact.

References:

1. Sharma S, Pinkert S, Nagaraju S, et al. Analysis of the human protein interactome and comparison with yeast, worm, and fly interaction datasets. *Nat Genet*. 2006;38(3):285-293. doi:10.1038/1747
2. Wagner A. How the global structure of protein interaction networks evolves. *Proc R Soc B Biol Sci*. 2003;270(1514):457-466. doi:10.1098/rspb.2002.2269
3. Beltrao P, Serrano L. Specificity and evolvability in eukaryotic protein interaction networks. *PLoS Comput Biol*. 2007;3(2):0258-0267. doi:10.1371/journal.pcbi.0030025
4. Kim PM, Korbel JO, Gerstein MB. Positive selection at the protein network periphery: Evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci U S A*. 2007;104(51):20274-20279. doi:10.1073/pnas.0710183104
5. Sahni N, Yi S, Taipale M, et al. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*. 2015;161(3):647-660. doi:10.1016/j.cell.2015.04.013
6. Yates CM, Sternberg MJE. The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein-protein interactions. *J Mol Biol*. 2013;425(21):3949-3963. doi:10.1016/j.jmb.2013.07.012
7. Kumar A, Butler BM, Kumar S, Ozkan SB. Integration of structural dynamics and molecular evolution via protein interaction networks: A new era in genomic medicine. *Curr Opin Struct Biol*. 2015;35:135-142. doi:10.1016/j.sbi.2015.11.002
8. Ohno S. *Evolution by Gene Duplication*. Berlin, Heidelberg: Springer Berlin Heidelberg; 1970. doi:10.1007/978-3-642-86659-3
9. Pastor-Satorras R, Smith E, Solé R V. Evolving protein interaction networks through gene duplication. *J Theor Biol*. 2003;222(2):199-210. doi:10.1016/S0022-5193(03)00028-6

10. Stelzl U, Worm U, Lalowski M, et al. A human protein-protein interaction network: A resource for annotating the proteome. *Cell*. 2005;122(6):957-968. doi:10.1016/j.cell.2005.08.029

11. Rolland T, Taşan M, Charloteaux B, et al. A proteome-scale map of the human interactome network. *Cell*. 2014;159(5):1212-1226. doi:10.1016/j.cell.2014.10.050

12. Trigg SA, Garza RM, MacWilliams A, et al. CrY2H-seq: A massively multiplexed assay for deep-coverage interactome mapping. *Nat Methods*. 2017;14(8):819-825. doi:10.1038/nmeth.4343

13. Bartel PL, Roecklein J a, SenGupta D, Fields S. A protein linkage map of Escherichia coli bacteriophage T7. *Nat Genet*. 1996;12(1):72-77. doi:10.1038/ng0196-72

14. Yu H, Braun P, Yildirim MA, et al. High-quality binary protein interaction map of the yeast interactome network. *Science*. 2008;322(5898):104-110. doi:10.1126/science.1158684.High

15. Giot L, Bader JS, Brouwer C, et al. A Protein Interaction Map of Drosophila melanogaster. *Science*. 2003;302(December):1727-1737. doi:10.1126/science.1090289

16. Vo T V., Das J, Meyer MJ, et al. A Proteome-wide Fission Yeast Interactome Reveals Network Evolution Principles from Yeasts to Human. *Cell*. 2016;164(1-2):310-323. doi:10.1016/j.cell.2015.11.037

17. Reinke AW, Baek J, Ashenberg O, Keating AE. Networks of bZIP Protein-Protein Interactions Diversified Over a Billion Years of Evolution. *Science (80-)*. 2013;340(May):730-735. doi:10.1126/science.1233465

18. Xin X, Gfeller D, Cheng J, et al. SH3 interactome conserves general function over specific form. *Mol Syst Biol*. 2013;9(652):1-17. doi:10.1038/msb.2013.9

19. Zhong Q, Pevzner SJ, Hao T, et al. An inter-species protein–protein interaction network across vast evolutionary distance. *Mol Syst Biol*. 2016;12(4):865. doi:10.15252/msb.20156484

20. Hochberg GKA, Thornton JW. Reconstructing Ancient Proteins to Understand the Causes of Structure and Function. *Annu Rev Biophys*. 2017;46(1):247-269. doi:10.1146/annurev-biophys-070816-033631

21. Harms MJ, Thornton JW. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat Rev Genet*. 2013;14(8):559-571. doi:10.1038/nrg3540

22. Laursen L, Čalyševa J, Gibson TJ, Jemth P. Divergent Evolution of a Protein-Protein Interaction Revealed through Ancestral Sequence Reconstruction and Resurrection. *Mol Biol Evol*. 2021;38(1):152-167. doi:10.1093/molbev/msaa198

23. Hultqvist G, Åberg E, Camilloni C, et al. Emergence and evolution of an interaction between intrinsically disordered proteins. *Elife*. 2017;6:1-25. doi:10.7554/eLife.16059

24. Jemth P, Karlsson E, Vögeli B, et al. Structure and dynamics conspire in the evolution of affinity between intrinsically disordered proteins. *Sci Adv*. 2018;4(10). doi:10.1126/sciadv.aau4130

25. Wheeler LC, Anderson JA, Morrison AJ, Wong CE, Harms MJ. Conservation of Specificity in Two Low-Specificity Proteins. *Biochemistry*. 2018;57(5):684-695. doi:10.1021/acs.biochem.7b01086

26. Zhang Z, Coenen H, Ruelens P, et al. Resurrected protein interaction networks reveal the innovation potential of ancient whole-genome duplication. *Plant Cell*. 2018;30(11):2741-2760. doi:10.1105/tpc.18.00409

27. Alhindi T, Zhang Z, Ruelens P, et al. Protein interaction evolution from promiscuity to

specificity with reduced flexibility in an increasingly complex network. *Sci Rep*. 2017;7(February):1-15. doi:10.1038/srep44948

28. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: Function, expression and evolution. *Nat Rev Genet*. 2009;10(4):252-263. doi:10.1038/nrg2538

29. Jindrich K, Degnan BM. The diversification of the basic leucine zipper family in eukaryotes correlates with the evolution of multicellularity Genome evolution and evolutionary systems biology. *BMC Evol Biol*. 2016;16(1):1-12. doi:10.1186/s12862-016-0598-z

30. Newman JRS, Keating AE. Comprehensive identification of human bZIP interactions with coiled-coil arrays. *Science*. 2003;300(5628):2097-2101. doi:10.1126/science.1084648

31. Amoutzias GD, Veron AS, Weiner J, et al. One billion years of bZIP transcription factor evolution: Conservation and change in dimerization and DNA-binding site specificity. *Mol Biol Evol*. 2007;24(3):827-835. doi:10.1093/molbev/msl211

32. Pinney JW, Amoutzias GD, Rattray M, Robertson DL. Reconstruction of ancestral protein interaction networks for the bZIP transcription factors. *Proc Natl Acad Sci U S A*. 2007;104(51):20449-20453. doi:10.1073/pnas.0706339104

33. Mitsui S. Antagonistic role of E4BP4 and PAR proteins in the circadian oscillatory mechanism. *Genes Dev*. 2001;15(8):995-1006. doi:10.1101/gad.873501

34. Male V, Nisoli I, Gascoyne DM, Brady HJM. E4BP4 : an unexpected player in the immune response. *Trends Immunol*. 2012;33(2):98-102. doi:10.1016/j.it.2011.10.002

35. Gachon F, Olela FF, Schaad O, Descombes P, Schibler U. The circadian PAR-domain basic leucine zipper transcription factors DBP, TEF, and HLF modulate basal and

inducible xenobiotic detoxification. *Cell Metab*. 2006;4(1):25-36. doi:10.1016/j.cmet.2006.04.015

36. Cowell IG. E4BP4 / NFIL3 , a PAR-related bZIP factor with many roles. 2002;(15):1023-1029. doi:10.1002/bies.10176

37. Boldridge WC, Ljubetič A, Kim H, et al. A multiplexed bacterial two-hybrid for rapid characterization of protein-protein interactions and iterative protein design. *bioRxiv*. 2020. doi:10.1101/2020.11.12.377184

38. Crooks RO, Lathbridge A, Panek AS, Mason JM. Computational Prediction and Design for Creating Iteratively Larger Heterospecific Coiled Coil Sets. *Biochemistry*. 2017;56(11):1573-1584. doi:10.1021/acs.biochem.7b00047

39. Khersonsky O, Tawfik DS. Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annu Rev Biochem*. 2010;79:471-505. doi:10.1146/annurev-biochem-030409-143718

40. Wheeler LC, Harms MJ. Were ancestral proteins less specific? *bioRxiv*. 2020:1-35. doi:10.1101/2020.05.27.120261

41. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature*. 1991;351(6328):652-654. doi:10.1038/351652a0

42. Siddiq MA, Loehlin DW, Montooth KL, Thornton JW. Experimental test and refutation of a classic case of molecular adaptation in Drosophila melanogaster. *Nat Ecol Evol*. 2017;1(2):1-6. doi:10.1038/s41559-016-0025

43. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*. 1999;151(4):1531-1545. doi:10101175

44. McClune CJ, Alvarez-Buylla A, Voigt CA, Laub MT. Engineering orthogonal signalling pathways reveals the sparse occupancy of sequence space. *Nature*. 2019;574(7780):702-706. doi:10.1038/s41586-019-1639-8

45. Baker CR, Hanson-Smith V, Johnson AD. Following Gene Duplication, Paralog Interference Constrains Transcriptional Circuit Evolution. *Science (80-)*. 2013;342(6154):104-108. doi:10.1126/science.1240810

46. Carroll SB. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell*. 2008;134(1):25-36. doi:10.1016/j.cell.2008.06.030

47. Kinney JB, McCandlish DM. Massively Parallel Assays and Quantitative Sequence-Function Relationships. *Annu Rev Genomics Hum Genet*. 2019;20:99-127. doi:10.1146/annurev-genom-083118-014845

48. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792-1797. doi:10.1093/nar/gkh340

49. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30(9):1312-1313. doi:10.1093/bioinformatics/btu033

50. Le SQ, Gascuel O. An improved general amino acid replacement matrix. *Mol Biol Evol*. 2008;25(7):1307-1320. doi:10.1093/molbev/msn067

51. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol*. 2010;59(3):307-321. doi:10.1093/sysbio/syq010

52. Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol*. 2006;55(4):539-552.

doi:10.1080/10635150600755453

53. Ehman EC, Johnson GB, Villanueva-meyer JE, et al. Renewing Felsenstein's phylogenetic bootstrap in the era of Big Data. *Nature*. 2018;556(7702):452-456.

54. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007;24(8):1586-1591. doi:10.1093/molbev/msm088

55. Chen K, Durand D, Farach-Colton M. NOTUNG: A program for dating gene duplications and optimizing gene family trees. *J Comput Biol*. 2000;7(3-4):429-447. doi:10.1089/106652700750050871

Chapter 4: Conclusions

The many trajectories to specificity in protein-protein interactions

In this thesis we have attempted to understand several aspects of specificity in PPIs. As it previously was quite difficult to investigate specificity, we first built a generalizable, scalable, high-throughput system to measure PPIs. The NGB2H system allows a wide variety of previous inaccessible questions to be addressed. Specifically, the use of gene synthesis coupled with a mapping step to unambiguously identify the hybrid protein sequences allows any protein to be assayed for interactions, unlike previous high-throughput PPI systems[1–6]. We used the NGB2H system to identify the largest sets of orthogonal proteins currently known, a reagent class that has a wide variety of uses in synthetic biology. These proteins were used in a design-build-test cycle to improve coiled-coil design algorithms, in an iterative fashion which should inform future *de novo* design. We then used the NGB2H system to investigate the evolution of specificity, which we found occurred very often in the PAR/E4BP4 family of bZips. Moreover we found the evolution of specificity to have several unique properties: it occurs gradually, it is a permanent gain, and it is not driven by direct selection. Direct measurement of these properties requires high-throughput characterization of ancestral proteins, and allow experimental revision of received evolutionary theory.

A. Potential applications of the NGB2H system

Although the interactomes of species have been characterized, there has been very little work on how polymorphisms can effect PPIs. The few studies that do exist of PPIs that don't use genome consensus sequences only serve to highlight how important it is to study polymorphisms. The first studies on variants of PPIs have characterized a few thousand mendelian[7] and developmental[8] disease mutants, and found perturbations were quite common. Interestingly,

these studies have found a substantial fraction of mutations will inhibit one interaction but not others, which emphasizes the need for targeted studies rather than gross effects such as gene knockouts to better understand the pleiotropic nature of most proteins[9]. Though one expects that characterizing polymorphisms that are not disease linked would capture fewer changes relative to the consensus genome, in light of the omnigenic model, characterizing the effects of all changes to PPIs in a quantitative manner is clearly necessary[10].

Because it uses gene synthesis to generate library diversity, the NGB2H system allows facile testing of synthetic proteins in PPIs. Given the veritable flood of *de novo* designed proteins which often engage in PPIs[11,12], there is a need for a system that can test sequence-similar synthetic constructs and a high-throughput system is particularly useful given the low percentage of successful designs[13]. Moreover, high-throughput studies have allowed actual learning from de novo design, such as finding pure alpha helical domains to be the most stable among microprotein topologies[14] and tradeoffs between stability and fluorescence in beta-barrel proteins[15]. Use of the NGB2H system could allow deep learning on de novo designed proteins that have modular hydrogen bond networks analogous to Watson-Crick base pairing, or easy characterization of binding strength of antigen-targeting proteins[16,17].

B. The future of engineered coiled-coils

Coiled-coils have a bright future as the nanoscale building blocks. Although DNA origami has long dominated the creation of nanoscale structures, proteins are becoming more popular as their design becomes more tractable[18]. Symmetric protein structures have used both *de novo* designed proteins[19–22] and repurposed oligomeric proteins[23–25] but are limited to platonic solids. Coiled-

coils, however, are able to build a wide variety of structures such as triangles, tetrahedra, and four sided pyramids[26–28] or large spherical cages[29]. The expanded set of orthogonal proteins that we have produced in this work will open up a wide range of previously uncreatable topologies. This is particularly exciting given the developing applications. Of note, coiled-coil nanostructures easily endocytosed, both in cell cultures[30] and in mice[28] which allows for *in vivo* uses. As such they could be functionalized with enzymes for improved localized catalysis, fluorophores for imaging, or loaded with cargo for intercellular release.

Our improvements to coiled-coil design offer some insight into the limits of protein design with a simple linear model. Most of our improvement in prediction did not come from reweighting compared to previous linear models of coiled-coils[31,32] even with our dramatically increased training set, but rather from heptad shifting—where we choose the heptad alignment between proteins that has the strongest interaction. Further improvements will likely use a similar approach and be driven by considerations outside of the individual residue pairings that determine iCipa.

Currently, the main weakness of iCipa is that it only functions for an extremely narrow set of amino acids, I and N at the A-position and E and K at the E- and G-positions. Fortunately, heptad shifting is likely generalizable to a more complex model than we built here, and a more general model of coiled-coil interactions should be able to incorporate heptad shifting with minimal changes.

C. High-throughput empirical studies of evolution

The study of evolution has long been driven by inference—as we cannot know what happened in the course of evolution we are left to mine sequences and morphologies for signatures of selection, adaptation, constraints, or gene flow. However, this can lead to biases in approach. One of the preeminent evolutionists of the second half of the twentieth century, Ernst Mayr, argued for an 'adaptationist program' saying, "Can one deduce the probability of causation by selection? Yes, by showing that possession of the respective feature would be favored by selection ... When one selectionist explanation of a feature has been discredited , the evolutionist must test other possible adaptationist solutions before he can resign."[33] This favors selective explanations in several ways, but has been broadly influential in shaping modern evolutionary thought[34,35].

Ancestral sequence reconstruction (ASR), though not free from its own form of inference, provides an important crosscheck on misguided attribution of some feature to a given evolutionary force[36]. Importantly, ASR has found numerous examples that suggest that acquisition of novel features can be driven by chance. For example the McDonald-Kreitman test[37], which compares the ratio of fixed non-synonymous to synonymous substitutions to the ratio of polymorphic non-synonymous to synonymous substitutions to provide a measure of adaption, has been used to find pervasive positive selection across genomes[38–40]. The origin of the McDonald-Kreitman test was to demonstrate the adaptive nature of *D. melanogaster's* ability to metabolize alcohol from rotting fruits relative to other Drosophila species, however, when the ancestral enzyme was inferred and K_m and K_{cat} compared to extant melanogaster's alcohol dehydrogenase, no differences were observed[41]. Similarly, the presence of widespread oligomerization has been attributed to positive selection[42], but ASR has demonstrated that this could easily occur neutrally and become entrenched by purifying selection[43]. In this work we

also showed non-adaptive processes could give rise to complexity. One speculates that the role of neutral processes in evolution has been underappreciated[44], and that ASR based experiments could provide a remedy for this. High-throughput ASR experiments, in particular, have a promising future answering some of the most pressing questions in evolution. Indeed we are already seeing experimental replay of the tape of life to measure the contributions of chance and contingency[45] and characterization of the evolutionary trajectories life did not take[46]. Future experiments could characterize entire evolutionary trajectories, to understand the dynamics of chance, contingency, entrenchment, selection and drift across time.

References:

1. Yachie N, Petsalaki E, Mellor JC, et al. Pooled-matrix protein interaction screens using Barcode Fusion Genetics. *Mol Syst Biol*. 2016;12(4):863-863. doi:10.15252/msb.20156660
2. Trigg SA, Garza RM, MacWilliams A, et al. CrY2H-seq: A massively multiplexed assay for deep-coverage interactome mapping. *Nat Methods*. 2017;14(8):819-825. doi:10.1038/nmeth.4343
3. Yang F, Lei Y, Zhou M, et al. Development and application of a recombination-based library versus library highthroughput yeast two-hybrid (RLL-Y2H) screening system. *Nucleic Acids Res*. 2018;46(3):1-12. doi:10.1093/nar/gkx1173
4. Yang JS, Garriga-Canut M, Link N, et al. rec-YnH enables simultaneous many-by-many detection of direct protein–protein and protein–RNA interactions. *Nat Commun*. 2018;9(1). doi:10.1038/s41467-018-06128-x
5. Andrews SS, Schaefer-Ramadan S, Al-Thani NM, Ahmed I, Mohamoud YA, Malek JA. High-resolution protein–protein interaction mapping using all- versus -all sequencing (AVA-Seq) . *J Biol Chem*. 2019;294(30):11549-11558. doi:10.1074/jbc.ra119.008792
6. Younger D, Berger S, Baker D, Klavins E. High-throughput characterization of protein–protein interactions by reprogramming yeast mating. *Proc Natl Acad Sci U S A*. 2017;114(46):12166-12171. doi:10.1073/pnas.1705867114
7. Sahni N, Yi S, Taipale M, et al. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*. 2015;161(3):647-660. doi:10.1016/j.cell.2015.04.013
8. Chen S, Fragoza R, Klei L, et al. An interactome perturbation framework prioritizes damaging missense mutations for developmental disorders. *Nat Genet*. 2018;50(7):1032-

1040. doi:10.1038/s41588-018-0130-z

9. Wagner GP, Zhang J. The pleiotropic structure of the genotype-phenotype map: The evolvability of complex organisms. *Nat Rev Genet*. 2011;12(3):204-213. doi:10.1038/nrg2949

10. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*. 2017;169(7):1177-1186. doi:10.1016/j.cell.2017.05.038

11. Huang P-S, Boyken SE, Baker D. The coming of age of de novo protein design. *Nature*. 2016;537(7620):320-327. doi:10.1038/nature19946

12. Beesley JL, Woolfson DN. The de novo design of α-helical peptides for supramolecular self-assembly. *Curr Opin Biotechnol*. 2019;58:175-182. doi:10.1016/j.copbio.2019.03.017

13. Ljubetič A, Gradišar H, Jerala R. Advances in design of protein folds and assemblies. *Curr Opin Chem Biol*. 2017;40:65-71. doi:10.1016/j.cbpa.2017.06.020

14. Rocklin GJ, Chidyausiku TM, Goreshnik I, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science (80-)*. 2017;357(6347):168-175. doi:10.1126/science.aan0693

15. Dou J, Vorobieva AA, Sheffler W, et al. De novo design of a fluorescence-activating β-barrel. *Nature*. 2018;561(7724):485-491. doi:10.1038/s41586-018-0509-0

16. Tinberg CE, Khare SD, Dou J, et al. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*. 2013;501(7466):212-216. doi:10.1038/nature12443

17. Chevalier A, Silva DA, Rocklin GJ, et al. Massively parallel de novo protein design for targeted therapeutics. *Nature*. 2017;550(7674):74-79. doi:10.1038/nature23912

18. Cannon KA, Ochoa JM, Yeates TO. High-symmetry protein assemblies: patterns and emerging applications. *Curr Opin Struct Biol*. 2019;55:77-84.

doi:10.1016/j.sbi.2019.03.008

19. Hsia Y, Bale JB, Gonen S, et al. Design of a hyperstable 60-subunit protein icosahedron. *Nature*. 2016;535(7610):1-12. doi:10.1038/nature18010

20. Bale JB, Gonen S, Liu Y, et al. Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science (80-)*. 2016;353(6297):389-394. doi:10.1126/science.aaf8818

21. King NP, Sheffler W, Sawaya MR, et al. Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy. *Science (80-)*. 2012;336(6085):1171-1174. doi:10.1126/science.1219364

22. King NP, Bale JB, Sheffler W, et al. Accurate design of co-assembling multi-component protein nanomaterials. *Nature*. 2014;510(7503):103-108. doi:10.1038/nature13404

23. Padilla JE, Colovos C, Yeates TO. Nanohedra: using symmetry to design self assembling protein cages, layers, crystals, and filaments. *Proc Natl Acad Sci U S A*. 2001;98(5):2217-2221. doi:10.1073/pnas.041614998

24. Lai Y-T, Reading E, Hura GL, et al. Structure of a designed protein cage that self-assembles into a highly porous cube. *Nat Chem*. 2014;6(12):1065-1071. doi:10.1038/nchem.2107

25. Cannon KA, Nguyen VN, Morgan C, Yeates TO. Design and Characterization of an Icosahedral Protein Cage Formed by a Double-Fusion Protein Containing Three Distinct Symmetry Elements. *ACS Synth Biol*. 2020;9(3):517-524. doi:10.1021/acssynbio.9b00392

26. Božič Abram S, Gradišar H, Aupič J, Round AR, Jerala R. Triangular in Vivo Self-Assembling Coiled-Coil Protein Origami . *ACS Chem Biol*. 2021. doi:10.1021/acschembio.0c00812

27. Gradišar H, Božič S, Doles T, et al. Design of a single-chain polypeptide tetrahedron assembled from coiled-coil segments. *Nat Chem Biol*. 2013;9(6):362-366. doi:10.1038/nchembio.1248

28. Ljubetič A, Lapenta F, Gradišar H, et al. Design of coiled-coil protein-origami cages that self-assemble in vitro and in vivo. *Nat Biotechnol*. 2017;35(11):1094-1101. doi:10.1038/nbt.3994

29. Fletcher JM, Harniman RL, Barnes FRH, et al. Self-assembling cages from coiled-coil peptide modules. *Science*. 2013;340(6132):595-599. doi:10.1126/science.1233936

30. Beesley JL, Baum HE, Hodgson LR, Verkade P, Banting GS, Woolfson DN. Modifying Self-Assembled Peptide Cages to Control Internalization into Mammalian Cells. *Nano Lett*. 2018;18(9):5933-5937. doi:10.1021/acs.nanolett.8b02633

31. Potapov V, Kaplan JB, Keating AE. Data-Driven Prediction and Design of bZIP Coiled-Coil Interactions. *PLoS Comput Biol*. 2015;11(2):1-28. doi:10.1371/journal.pcbi.1004046

32. Mason JM, Schmitz M a, Müller KM, Arndt KM. Semirational design of Jun-Fos coiled coils with increased affinity: Universal implications for leucine zipper prediction and design. *Proc Natl Acad Sci U S A*. 2006;103(24):8989-8994. doi:10.1073/pnas.0509880103

33. Mayr E. How to Carry Out the Adaptationist Program? *Am Nat*. 1983;121(3):324-334. doi:10.1086/284064

34. Kern AD, Hahn MW. The neutral theory in light of natural selection. *Mol Biol Evol*. 2018;35(6):1366-1371. doi:10.1093/molbev/msy092

35. Yang Z, Bielawski JR. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol*. 2000;15(12):496-503. doi:10.1016/S0169-5347(00)01994-7

36. Hochberg GKA, Thornton JW. Reconstructing Ancient Proteins to Understand the Causes of Structure and Function. *Annu Rev Biophys*. 2017;46(1):247-269. doi:10.1146/annurev-biophys-070816-033631

37. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature*. 1991;351(6328):652-654. doi:10.1038/351652a0

38. Smith NGC, Eyre-Walker A. Adaptive protein evolution in Drosophila. *Nature*. 2002;415(6875):1022-1024. doi:10.1038/4151022a

39. Bustamante CD, Fledel-Alon A, Williamson S, et al. Natural selection on protein-coding genes in the human genome. *Nature*. 2005;437(7062):1153-1157. doi:10.1038/nature04240

40. Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genet*. 2010;6(1). doi:10.1371/journal.pgen.1000825

41. Siddiq MA, Loehlin DW, Montooth KL, Thornton JW. Experimental test and refutation of a classic case of molecular adaptation in Drosophila melanogaster. *Nat Ecol Evol*. 2017;1(2):1-6. doi:10.1038/s41559-016-0025

42. Goodsell DS, Olson AJ. Structural Symmetry and Protein Function. *Annu Rev Biophys Biomol Struct*. 2000;29(1):105-153. doi:10.1146/annurev.biophys.29.1.105

43. Hochberg GKA, Liu Y, Marklund EG, Metzger BPH, Laganowsky A, Thornton JW. A hydrophobic ratchet entrenches molecular complexes. *Nature*. 2020;588(7838):503-508. doi:10.1038/s41586-020-3021-2

44. Lynch M. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci*. 2007;104(Supplement 1):8597-8604. doi:10.1073/pnas.0702207104

45. Xie VC, Pu J, Metzger BPH, Thornton JW, Dickinson BC. Chance, contingency, and necessity in the experimental evolution of ancestral proteins. *bioRxiv*. 2020:2020.08.29.273581. https://doi.org/10.1101/2020.08.29.273581.

46. Starr TN, Picton LK, Thornton JW. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature*. 2017;549(7672):409-413. doi:10.1038/nature23902