PROTEIN INTERACTION NETWORK BASED APPROACHES TO CHARACTERIZE PROTEIN FUNCTION, MOLECULARLY PROFILE GENETIC VARIANTS, AND INVESTIGATE MECHANISMS LINKED TO VIRAL-HOST PATHOLOGY IN SARS-COV-2

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Shayne Davis Wierbowski

May 2022

PROTEIN INTERACTION NETWORK BASED APPROACHES TO CHARACTERIZE PROTEIN FUNCTION, MOLECULARLY PROFILE GENETIC VARIANTS, AND INVESTIGATE MECHANISMS LINKED TO VIRAL-HOST PATHOLOGY IN SARS-COV-2

Shayne Davis Wierbowski, Ph. D.

Cornell University 2022

A majority of protein function is mediated through direct binary or complex interactions with other proteins. Therefore, systematic efforts to characterize these protein interaction networks and to structurally resolve their interaction interfaces have provided powerful tools to comprehensively study protein function at a molecular level. For instance, disease mutations are enriched along protein interaction interfaces and network level impacts of disease mutations can elucidate mechanisms of disease in terms of specific interactions affected. The contents of this dissertation describe a range of research efforts I've led or contributed to aimed at broadening the scope of a networks-based approach to human health and disease centered around protein interaction molecular phenotypes. These efforts begin with a systematic effort to provide the resources necessary to functionally characterize rice proteins at high-throughput and direct applications to map the rice protein-protein interactome. The experimental approaches and computational analyses described here can be extended beyond rice and could provide the bases for molecular network characterization in any species. From there, I describe my contributions to validate a mutation library

containing plasmid clones for over 2,000 human population and disease variants. This resource was leveraged to comprehensively measure impacts these variants on protein interaction networks to directly quantify and contextualize the extent of disruptive variants and their relationship with the human genetic background, disease, and overall fitness. Finally, I extend existing machine learning frameworks to predict protein interaction interfaces by applying protein-protein docking to construct full 3D models for these viral-host interactions between SARS-CoV-2 and human proteins. I subsequently perform mutational scanning and binding affinity calculations to predict impacts of molecular perturbations within these interactions. In doing so I explore the utility of structural interactome modelling to investigate the implications of recent evolutionary history, genetic population diversity, and potential drug repurposing on viral-host pathology through the lens of protein interactions. Cumulatively, these efforts have expanded the pace at which systematic molecular profiling of protein interaction networks can be conducted both experimentally and computationally in the Yu lab and the broader scientific community.

BIOGRAPHICAL SKETCH

Shayne Wierbowski was born in Johnson City, New York and raised in Owego, New York. After graduating from Seton Catholic Central High School as valedictorian for the class of 2012, Shayne was admitted as a Presidential Scholar to The University of Scranton to complete his bachelor's degrees. There he pursued Bachelor of Science degrees in Biochemistry, Cell and Molecular Biology (BCMB) and Computer Science. Through participation in the University's Special Jesuit Liberal Arts program (SJLA), Shayne additionally completed a Bachelor of Arts degree in Philosophy. Shayne began his research career at Scranton working in the labs of Drs. Timothy Foley and George Gomez. He also participated in summer research programs in 2014 at Princeton University—during which he studied reproductive aging in *C. elegans* under the mentorship of Dr. Coleen Murphy—and in 2015 at The University of Pittsburgh—during which he constructed benchmarks for and validating protein-ligand molecular docking algorithms under the mentorship of Dr. Carlos Camacho. The later would become the basis for Shayne's undergraduate thesis for Scranton's Honors Program.

After the receiving of his degrees from The University of Scranton *summa cum laude* in 2016, Shayne began graduate studies within the Department of Computational Biology at Cornell University. He is currently completing his sixth year in Dr. Haiyuan Yu's lab and will graduate with a Doctor of Philosophy in Computational Biology in May 2022. During the completion of his graduate work, Shayne's research interests have broadly involved protein-protein interaction networks with a focus on how genetic perturbations can modulate these interactions at a molecular level, and how these molecular modulations contribute to organismal-level phenotypes and disease. An ultimate aim of these interests is in exploring how computational tools and structural modeling can help elucidate this mapping to eventually tailor personalized therapeutic interventions.

ACKNOWLEDGMENTS

I thank all the members of the Yu Lab past and present for their contributions to an excellent lab environment and numerous insightful scientific discussions over the years. I also thank Dr. Haiyuan Yu for being my advisor and for his dedication and commitment to scientific growth and development throughout my Ph.D.

I thank the entire Computational Biology Department at Cornell University, particularly my immediate cohort, Ian Caldas and Juan Felipe Beltran, for providing social support and being an outlet to cope with various stresses of graduate school. I further thank the administrative staff in the Computational Biology Department and the Weill Institute for their immense support to help manage logistics, forms, and deadlines throughout my Ph.D.

I thank and acknowledge receipt of generous gifts from Julian I. Schroeder for providing positive and negative control clones for BiFC in Chapter 2 and from P. H. Wang for providing viral clones for SARS-CoV-1 and SARS-CoV-2 in Chapter 4.

Finally, the duration of my Ph.D and the specific papers reproduced herein would not be possible without financial support from a number of funding agencies including: grants from NIGMS (R01 GM097358, R01 GM104424., R01 GM124559, R01 GM124559, and R01 GM125639), NIDDK (R01 DK115398), NCI (R01 CA167824), NICHD (R01 HD082568), NHGRI (UM1 HG009393 and R01 HG006849), NSF (DBI-1661380), SFARI (575547), HIH (R01AI35270), and a Cornell Rapid Research Response to SARS-CoV-2 Seed Grant. The funders had no role in study design, data collection and analysis, decision to publish or preparation of these works.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 2

The tables included in chapter 2 correspond with Datasets 1-8 of my PNAS paper, "A massively parallel barcoded sequencing pipeline enables generation of the first ORFeome and interactome map for rice" and can most conveniently be accessed online https://doi.org/10.1073/pnas.1918068117. Titles and direct access links are provided below.

Table 8. Manual literature search on top 21 interacting genes.
Table available for download online at:
https://www.pnas.org/doi/suppl/10.1073/pnas.1918068117/suppl_file/pnas.1918068117.sd08.xlsx

Chapter 3

The tables included in chapter 3 correspond in order with Supplementary Table 1-3, Supplementary Data 2-4 and Supplementary Data 7 of the Nature Communications paper I contributed to, "Extensive disruption of protein interactions by genetic variants across the allele frequency spectrum in human populations" and can most conveniently be accessed online https://doi.org/10.1038/s41467-019-11959-3. Titles and direct access links are provided below.

Table 9. Calculation of Functional Missense Mutations for 1000 Genomes Project Phase.
Table available as Supplementary Table 1 in the Supplemental Information here:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-019-11959-3/MediaObjects/41467_2019_11959_MOESM1_ESM.pdf

Table 10. Calculation of Functional Missense Mutations for GoNL.
Table available as Supplementary Table 2 in the Supplemental Information here:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-019-11959-3/MediaObjects/41467_2019_11959_MOESM1_ESM.pdf

Table 11. Calculation of Functional Nonsynonymous Mutations for ESP Phase I.
Table available as Supplementary Table 3 in the Supplemental Information here:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-019-11959-3/MediaObjects/41467_2019_11959_MOESM1_ESM.pdf

Table 12. Interaction perturbation results for ExAC variants.
Table available for download online at:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-019-11959-3/MediaObjects/41467_2019_11959_MOESM6_ESM.xlsx

Table 13. Interaction perturbation results for somatic mutations from COSMIC.
Table available for download online at:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-019-11959-3/MediaObjects/41467_2019_11959_MOESM7_ESM.xlsx

Table 14. Interaction perturbation results for disease-associated mutations from HGMD.
Table available for download online at:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-019-11959-3/MediaObjects/41467_2019_11959_MOESM8_ESM.xlsx

Table 15. List of disruptive variants occurring in drug target-relevant genes.
Table available for download online at:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-019-11959-3/MediaObjects/41467_2019_11959_MOESM11_ESM.xlsx

Chapter 4

The tables included in chapter 4 correspond with Supplementary Tables 1-10 of my Nature Methods paper, "A 3D structural SARS-CoV-2–human interactome to explore genetic and drug perturbations" and can most conveniently be accessed online https://doi.org/10.1038/s41592-021-01318-w. Titles and direct access links are provided below.

Table 16. List of ECLAIR-predicted interface residues.
Table available for download online at:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-021-01318-w/MediaObjects/41592_2021_1318_MOESM2_ESM.xlsx

Table 17. List of guided docking annotated interface residues.
Table available for download online at:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-021-01318-w/MediaObjects/41592_2021_1318_MOESM3_ESM.xlsx

Table 18. List of co-crystal structures and interface annotations from the human–pathogen PDB interaction benchmark.
Table available for download online at:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-021-01318-w/MediaObjects/41592_2021_1318_MOESM4_ESM.xlsx

Table 19. List of human population variants reported by gnomAD.
Table available for download online at:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-021-01318-w/MediaObjects/41592_2021_1318_MOESM5_ESM.xlsx

Table 20. List of sequence divergences between SARS-CoV-1 and SARS-CoV-2.
Table available for download online at:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-021-01318-w/MediaObjects/41592_2021_1318_MOESM6_ESM.xlsx

Table 21. Enrichment for sequence variation on SARS-CoV-2–human interfaces.
Table available for download online at:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-021-01318-w/MediaObjects/41592_2021_1318_MOESM7_ESM.xlsx

Table 22. Enrichment for sequence variation on SARS-CoV-2–human interfaces.

Table available for download online at:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-021-01318-w/MediaObjects/41592_2021_1318_MOESM8_ESM.xlsx

Table 23. Predicted ΔΔG between SARS-CoV-1 and SARS-CoV-2 versions of all docked interactions.
Table available for download online at:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-021-01318-w/MediaObjects/41592_2021_1318_MOESM9_ESM.xlsx

Table 24. Predicted ΔΔG impact of all human population variants at the interface.
Table available for download online at:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-021-01318-w/MediaObjects/41592_2021_1318_MOESM10_ESM.xlsx

Table 25. List of all predicted drug–target binding sites.
Table available for download online at:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-021-01318-w/MediaObjects/41592_2021_1318_MOESM11_ESM.xlsx

CHAPTER 1:
INTRODUCTIONS AND THE VALUE OF PROFILING MOLECULE PHENOTYPES EXPERIMENTALLY AND COMPUTAITONALLY

Context and Personal Contributions

In an earlier review article published in Current Opinion in Systems Biology[1], I make the argument that experimental characterizations of and ability to computationally predict precise molecular phenotypes associated with specific genetic perturbations are critical to inform hypothesis driven research and mechanistic insights surrounding human health and disease. This perspective is fundamental to a wide range of research related to protein-protein interaction networks that has been pursued by me and my colleagues in the Yu lab throughput the duration of my PhD. A restructured and slightly abridged version of this review is provided here to serve as a broadly centralizing focus for the chapters to come. In order these chapters explore the importance of studying molecular phenotypes by: 1) advancing the methodologies available for high-throughput characterization of protein-protein interaction networks by yeast two-hybrid (Y2H) systems (availability of these networks being a prerequisite for extensive molecular profiling), 2) providing bioinformatic support for experimental efforts to systematically map and quantify the molecular perturbations of human population and disease variants on their protein interactions, and 3) applying computational methods to predict protein-protein interaction interfaces (including construction of full 3D models) in the context of the recent SARS-CoV-2 pandemic to facilitate mechanistically-informed hypotheses surrounding recent evolution of the virus, potential impacts of genetic variation on viral-host interactions linked to divergent patient outcomes, and

drug repurposing around the idea of disrupting specific molecular network linked through chemical perturbation. Some sections of this argument were originally co-authored with contributions from Robert Fragoza and Siqi (Charles) Liang. The content of this argument has not been updated to account for any advances in the field between its original 2018 publication and it's reformatting here. Some sections may be outdated.

Introduction

Ever-improving next-generation sequencing technologies have led to the ongoing discovery of tens of millions of DNA variants across diverse human populations[2] and have enabled the identification of tens of thousands of disease-associated mutations[3,4]. Nonetheless, a vast majority of these variants remain uncharacterized and a corresponding understanding of how these unannotated variants may contribute to human disease and traits has yet to materialize[5]. Missense variants are of particular interest to researchers since known disease- and trait-associated mutations have been shown to be enriched in coding regions[6]. Proper interpretation of the functional impact of missense mutations, which dominate exome sequencing datasets, remains a pivotal challenge. Overcoming this challenge will require new tools and approaches that better leverage large-scale sequencing data and that take advantage of newly emerging sources of experimentally assessed functional variant data. We believe operating from a mechanistic perspective aimed at experimentally profiling or computationally predicting the context of a variant's molecular impact will be critical for this task.

Functional prediction algorithms have provided a boon towards the identification and prioritization of disease-associated mutations. Although early

approaches to disease association specifically prioritized rare variants, tools such as SIFT[7-9], PolyPhen-2[9,10], CADD[11], and PROVEAN[12-14] have provided systematic methods for predicting the impact of missense variants. These approaches share a central approach that utilizes principles of population genetics and conservation both within humans and across species as a means of approximating the fitness cost of specific variants. Cumulatively, these methods have been widely used in prior identification of disease-associated mutations[15-20]. However, while these methods continue to persist as invaluable tools for prioritizing coding mutations in disease, annotations from these tools alone do not provide insight into the underlying molecular mechanisms of causal variants. Indeed, no method to-date can effectively identify true risk missense variants for human disease[21,22].

Mutations can perturb cellular activity in multiple ways. In particular, disease-associated missense mutations often function by disrupting protein-protein interactions[23-25], destabilizing protein folding[23,24], or altering transcription factor activity[26,27]. Understanding the molecular mechanisms through which disease-associated mutations function is imperative for developing clinical strategies to treat their corresponding phenotypes and for drug target assessment[28,29]. In spite of this importance, only a single widely used variant annotation algorithm for coding variants, MutPred2[30], currently evaluates the possible mechanisms by which mutations are scored as deleterious may function. This information is critical for developing targeted hypotheses and clinical strategies to target causal mutations. Indeed, despite their widespread use, current algorithms often perform poorly in clinical settings and seldom result in measurable phenotypes; roughly 20% validation rates with poor consistency

between algorithms[31-34]. More precise predictions for deleterious variants and better insights to their corresponding molecular mechanisms may be achieved through improved structural databases to detail where missense mutations physically occur with respect to protein interface residues[35,36].

A guiding principle of precision medicine is to accurately measure clinical and molecular attributes of individual patients so as to tailor personalized therapies based on the outcomes of these measurements[37]. Considering millions of DNA variants segregating in human genomes, and the extraordinary level of allelic heterogeneity found in disease, success of the precision medicine effort hinges not only on the ability to detect disease-causing mutations, but also to understand and properly assess the functional consequences of these mutations. A major challenge, therefore, is to radically accelerate the pace of experimental and computational assessments of the functional impacts of millions of single nucleotide variants (SNVs) uncovered by sequencing efforts. Direct assessments of molecular phenotypes—such as impact on protein stability, enzymatic kinetics, or protein interaction binding affinities by missense mutations—provide a unique and complementary perspective to current methods for detecting causal disease mutations. Integrating molecular phenotype data into fitness-based approaches for identifying deleterious mutations may also provide new insights into how causal mutations mechanistically function and provides a framework for dissecting epistatic relationships that modulate the impact of low penetrance mutations.

A Molecular Phenotype First Perspective on Human Variation

In assessing the impact of human variants, we highlight the importance of distinguishing

three related yet distinct biological concepts: overall fitness, organismal/cellular phenotype, and molecular phenotype (**Figure 1**). Overall fitness refers to the ability of an individual to survive and reproduce. Organismal phenotypes refer to observable features, including disease phenotypes such as diabetes, autism spectrum disorder and cancer, or traits such as height, hair color and blood type. Molecular phenotypes refer to the direct effect of a variant at the molecular level. For example, changes in gene expression, loss of protein stability, changes in enzymatic activity, or modifications to protein-protein, protein-DNA or protein-ligand interaction affinities.

All human genetic variation separates into molecularly inert or molecularly active variants depending on whether or not each variant causes a molecular phenotype. While not all molecular phenotypes contribute directly to observable organismal phenotypes, organismal or cellular phenotypes are largely derived in molecularly active variants; and hence must be directly mediated through one or more molecular phenotypes. Likewise, overall fitness is always rooted in molecular phenotypes since molecular changes modulate the ability of the organism to perform various functions necessary for survival and reproduction. In principle, all organismal phenotypes associate with a fitness value ranging from deleterious, to neutral, to advantageous. While there is a direct relationship between organismal phenotypes and fitness, this relationship is not always clearly defined, particularly in specialized fields of disease research dealing with cancer biology, age related or post-reproductive diseases, and complex diseases with reduced penetrance[38]. In such disease studies, the one-to one correspondence between fitness score and the severity of the organismal phenotype breaks down since clinically deleterious phenotypes can have limited impact on

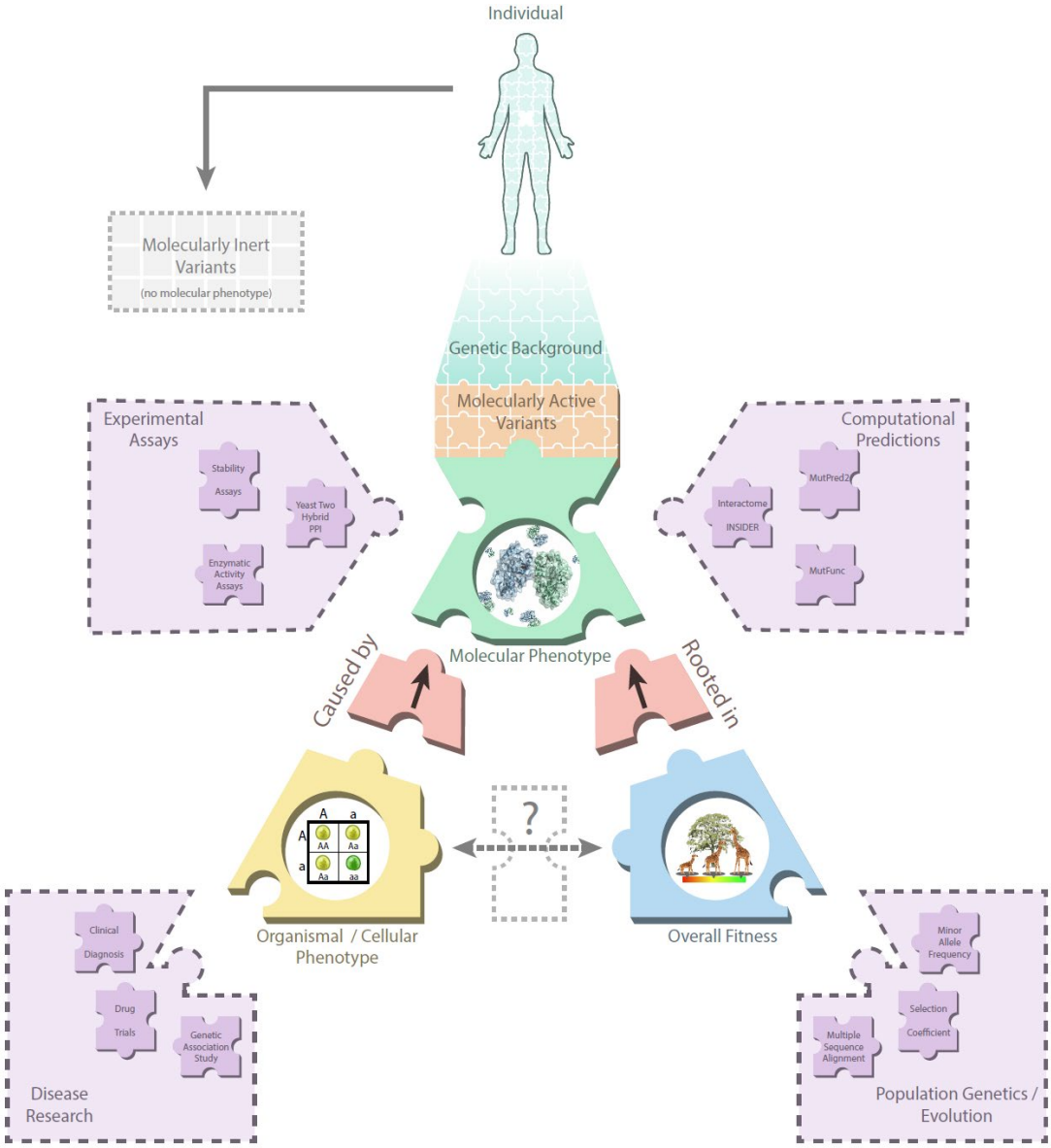


Figure 1. Molecular Phenotype First Perspective to Human Variation
Graphical depiction of the relationship between three related biological concepts associated with human variations: 1) molecular phenotype, 2) organismal/cellular phenotype, and 3) overall fitness. All genetic variation is either molecularly inert or molecularly active. The cumulation of all molecularly active variants—each causing one or more molecular phenotypes—constitutes the unique genetic background of an individual. Molecular phenotypes provide the ultimate link explaining the mechanistic basis for how SNVs manifest in organismal/cellular phenotypes or come to be selected for or against through fitness effects. Although organismal phenotypes, in general, directly relate to overall fitness, weak effect diseases, late onset/post-reproductive diseases, and partially penetrant mutations often confound this relationship. Researchers have various tools to perform direct inquiries into how these three concepts relate to specific molecularly active variants. Human disease research aims to understand organismal/cellular phenotypes while population genetics provides insights into fitness, conservation, and selection.

Researchers investigate molecular phenotypes either through direct experimental assays to observe underlying molecular phenotypes or through computational predictions of putative molecular phenotypes. The ultimate aim is to infer information about one spoke of the triangle through the other two; namely, scientists seek to infer which SNVs are causal disease variants though information about the overall fitness or molecular phenotype effects of the SNV.

reproduction. Molecular phenotypes can be indispensable towards characterizing these cases of ambiguous fitness-to-phenotype relationships.

Molecular phenotypes provide complementary information for identifying causal variants

Whereas most approaches leverage the link between fitness effects and organismal/cellular phenotypes, an alternative framework rooted in molecular phenotypes provides an orthogonal line of support. At least two degrees of separation lie between disease phenotypes caused by particular variants, the fitness effects of these variants, and our ability to discern these effects. By contrast methods aimed at molecular phenotypes directly address the central link. The combination of these two rationally justified, yet conceptually distinct paths connecting SNVs to disease phenotype is expected to culminate in an overall higher degree of accuracy in predicting disease associations. The availability of data and library of tools for assessing molecular phenotypes are currently leagues behind the equivalent datasets for fitness-based approaches. Therefore, it is likely that established conservation and fitness-based methods will remain a valuable step in prioritizing variants, while more direct support from the orthogonal molecular phenotype data should serve as strong confidence in the accuracy of these results.

For instance, a recently developed interaction perturbation framework leveraged annotations of protein-protein interaction (PPI) interface residues[36] alongside PolyPhen-

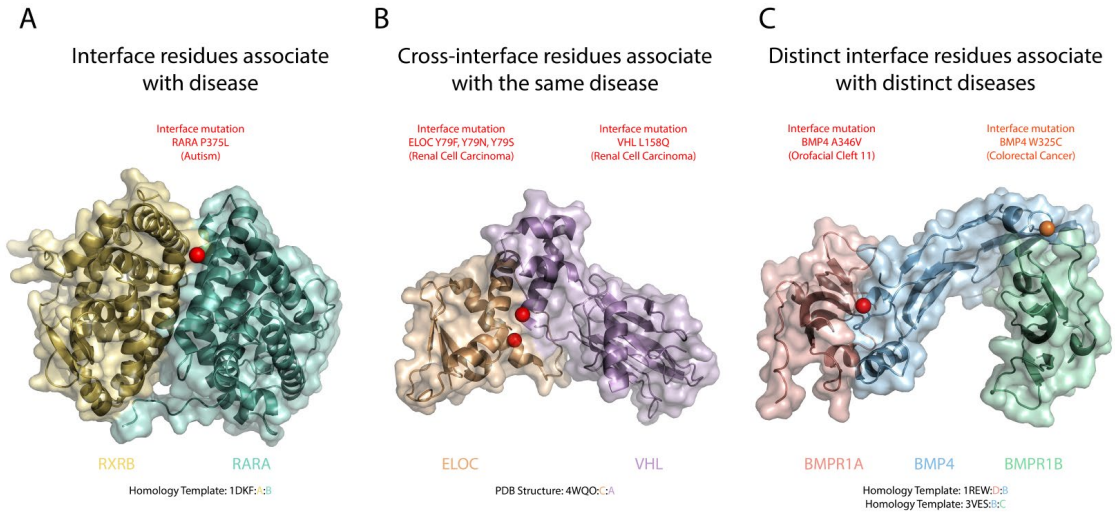


Figure 2. Examples of Interaction Interface Molecular Phenotype Annotation
Molecular phenotypes including the annotation of protein-protein interaction interface residues can inform the mechanism of disease-associated mutations. **a,** Homology model between RARA (template 1DKF:B) and RXRB (template 1DKF:A) used to distinguish a potentially causal mutation from a benign mutation. A *de novo* mutation, P375L, on RARA identified in an autism spectrum disorder-affected individual occurs on an interface residue with RXRB. RARA interface residue mutations were not found in an unaffected sibling. **b,** Homology model between VHL (PDB 4WQO:A) and ELOC (PDB 4WQO:C) demonstrates potential leveraging of molecular phenotypes to identify convergent mechanisms in divergent disease mutations. Variants on both of these proteins associate with the same disease and localize to the same interface. **c,** Homology model between BMP4 (template 1REW:B), BMPR1A (template 1REW:A), and BMPR1B (template 3VES:C) shows hypothesis-driven differentiation of mechanisms of different diseases based on molecular phenotype. Two variants on BMP4, A346V, and W325C, associated with divergent diseases localize to distinct interaction interfaces.

2 scores[39]. Chen and colleagues demonstrated increased accuracy in distinguishing *de novo* risk variants in autism spectrum disorder from benign mutations in unaffected siblings. **Figure 2a** provides a reconstructed example in which a proband PolyPhen-2 mutation scored as "probably damaging", P375L on the protein RARA, occurred on a predicted interface residue. In contrast, a second PolyPhen-2-scored "probably damaging" mutation, R83H on the same RARA protein, was reported in an unaffected individual; however, R83H did not occur on a predicted interaction interface residue. Consequently, despite matching PolyPhen-2 prediction, only the proband P375L mutation was predicted to disrupt the heterodimeric interaction between RARA and

RXRB, a prediction which the authors also validated experimentally. This exemplifies the potential for molecular phenotypes to aid in pinpointing candidate causal variants that are otherwise indistinguishable from molecularly inert variants using fitness-based methods alone.

Leveraging molecular phenotype approaches towards disentangling molecular mechanisms of causal variants

The molecular phenotype framework provides clear potential to investigate the underlying mechanisms behind how variants manifest in disease phenotypes. Since the specific molecular defect associated with a variant often directly relates to the disease phenotype, identification of candidate variants based on molecular phenotype annotations should enable translational studies for disease etiology. The further development of methods to approximate and predict molecular phenotypes will facilitate the development of actional hypotheses to direct future research.

For instance, Chen *et al.* used experimentally derived and computationally predicted annotations of protein interaction interface residues[36] as a predictor for the molecular phenotype, loss of PPI. In addition to distinguishing a true autism risk variant, P375L, from other "probably damaging" variants, the additional knowledge that this variant intersected with the RARA-RXRB interaction interface (**Figure 2a**), led to the testable hypothesis that this variant would disrupt this interaction, and helped to propose a pathway for RARA's involvement in autism spectrum disorder through this interaction[39].

Extending the interface residue approximation for the loss of PPI molecular phenotype facilitates mechanistic inferences in other cases as well. This approach may

be generalized to cases involving variants across both faces of an interface (**Figure 2b**). Corroborating cross-interface evidence may strengthen the hypothesis that disease-associated mutations function through disruption of a specific interaction and helps categorize distinct variants associated with the same disease by similarities in their molecular mechanisms. **Figure 2b** shows a known tumor suppressor gene-encoded protein, VHL[40,41] with a mutation, L158Q, associated with renal cell carcinoma, in complex with an elongation factor, ELOC. The localization of L158Q at the ELOC interface, suggests that the disease may function through disruption of the VHL-ELOC interaction. Moreover, ELOC contains several mutations on the same protein interaction interface, Y79F, Y79N, and Y79S, which are also associated with renal cell carcinoma, solidifying the hypothesis that these cross-interface variants drive a distinct form of renal cell carcinoma through a single shared molecular phenotype.

Understanding the molecular phenotypes caused by certain disease-associated mutations may further elucidate how several mutations on the same gene can associate with different diseases. For instance, two missense mutations found on the protein BMP4, A346V and W325C, are associated with a developmental defect orofacial cleft 11, and colorectal cancer, respectively—two clinically distinct diseases. The homology models provided in **Figure 2c** demonstrate that these variants localize to opposites ends of the BMP4 structure and occur at distinct protein-protein interaction interfaces. These insights suggest these distinct disease phenotypes may manifest through divergent pathways related to the biological functions of their distinctly targeted interaction partners. Indeed, although BMPR1A and BMPR1B are paralogous, previous studies have linked them to unique functions and disease states[42,43].

Cumulatively, these interaction perturbation examples demonstrate how molecular phenotypes contribute to elucidation of disease etiology. We emphasize the potential to explore similar mechanistic hypotheses utilizing molecular phenotypes outside of PPI disruption. Recent studies have highlighted the value of examining other molecular phenotypes, including changes in protein stability[44,45] as well as changes in gene expression level[46,47], to unravel the pathogenic mechanisms of both coding and non-coding mutations.

Expanding the Boundaries of Molecular Phenotype Annotation and Prediction

The incorporation of direct assays for molecular phenotypes and novel computational methods that approximate molecular phenotypes in the continued efforts to identify, prioritize, and understand causal variants in human disease is positioned to provide a truly orthogonal view to the longstanding fitness-based approach. Whereas current variant annotation algorithms rooted in sequencing and fitness approximations have yielded suboptimal specificity, novel methods directed at molecular phenotypes aim to extract complementary molecular insights otherwise unavailable. Towards these ends, researchers have conducted high-throughput assays to directly measure the functional impact of thousands of disease-associated missense mutations on protein-protein interactions[23,24], protein stability[23], and DNA binding[26,27]. Literature curation efforts by the IMEx Consortium have provided protein interaction perturbation data corresponding to nearly 8,000 coding mutations in humans[48]. Continued development of high-throughput approaches—including deep-mutational scanning pipelines capable of probing nearly the entire mutational landscape of targeted proteins[49-52]—will provide an

ever-larger resource of functional mutation data. This data will help elucidate the biochemical and evolutionary properties that differentiate truly damaging mutations from those that are benign.

Despite the impressive scale that high-throughput experimental pipelines have achieved[23,24,50], no experimental pipeline alone can keep pace with the rate of sequence variant discovery, highlighting the need for continued development of computational approaches and variant annotation algorithms. We emphasize a continued need to develop novel computational approaches to directly inform predictions about putative molecular phenotypes. For instance, interaction interface residue annotations provide useful mechanistic insights, but low coverage in experimentally validated structures or homology models has limited their applicability. The recently published Interactome INSIDER resource provides a method to predict interface residues—and consequentially loss of PPI phenotypes—in the absence of structural information[36]. MutPred2 enables a combination of approaches, making predictions both for overall functional effect and prioritized potential mechanisms of action[30]. Recently, Wagih et al have released MutFunc containing precomputed predictions for every possible variant in *H. sapiens*, *S cerevisiae*, and *E. coli*. These predictions include estimates for changes to protein stability, protein interaction interfaces, post translational modifications, and transcription factor binding among other approximations for molecular phenotypes[53].

Throughout the completion of my graduate work, various endeavors researched by myself and colleagues in the Yu lab have sought to expand and improve upon the awareness and availability of comprehensive resources to comprehend molecular phenotypes as they relate to and inform future studies into human disease. Here I

enumerate several of my key contributions and publications to this field with a particular emphasis on the importance of protein-protein interactions. First, I present a systematic effort to construct a rice ORF library and expand the molecular characterization of the rice proteome by mapping the protein-protein interaction network within this library. These efforts broadly cover the essential first steps to enable comprehensive molecular profiling in any species. I then describe my contributions to an extensive effort to generate and screen the molecular impact of over 2,000 human single nucleotide variants on protein stability and interactions. Notably, this constituted one of the only studies to measure the extent of molecular perturbations arising not only from disease mutations but from common population variants as well. These molecularly active variants within seemingly healthy individuals may be crucial to explaining epigenetic phenomena, and this resource generally can provide a training set to inform development of computational methods to predict molecular phenotypes. Finally, I present recent efforts to characterize the protein-protein interaction interfaces for SARS-CoV-2-human interactions. This work extended existing resources within the Yu lab to make residue-level predictions of interaction interfaces to produce full 3D docked models guided by these high-confidence predictions. This is turn facilitated computational prediction of the binding energies and dynamics to contrast viral-host interactions in SARS-CoV-1 against those in SARS-CoV-2 and to explore the mutational landscape around the interfaces for variants capable of modulating these interactions. This structural approach to modeling viral-host pathology through protein-protein interactions further provides an opportunity to explore drug repurposing strategies aimed not at targeting natural human molecular processes, but explicit

molecular interactions critical to viral-host pathology. Broadly, the advances explored throughput my PhD in this realm of widespread predictors for specific molecular phenotypes will prove crucial to prioritizing molecularly informed hypothesis testing and efficient advances in experimental research.

REFERENCES

1 Wierbowski, S. D., Fragoza, R., Liang, S. & Yu, H. Extracting Complementary Insights from Molecular Phenotypes for Prioritization of Disease-Associated Mutations. *Curr Opin Syst Biol* 11, 107-116, doi:10.1016/j.coisb.2018.09.006 (2018).

2 Snyder, M., Du, J. & Gerstein, M. Personal genome sequencing: current approaches and challenges. *Genes Dev* 24, 423-431, doi:24/5/423 [pii] 10.1101/gad.1864110 (2010).

3 The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56-65, doi:http://www.nature.com/nature/journal/v491/n7422/abs/nature11632.html#supplementary-information (2012).

4 Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493, 216-220, doi:nature11690 [pii] 10.1038/nature11690 (2013).

5 Stenson, P. D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome Med* 1, 13, doi:gm13 [pii] 10.1186/gm13 (2009).

6 Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* 106, 9362-9367, doi:10.1073/pnas.0903103106 (2009).

7 Ng, P. C. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research* 31, 3812-3814, doi:10.1093/nar/gkg509 (2003).

8 A Fast Computation of Pairwise Sequence Alignment Scores Between a Protein and a Set of Single-Locus Variants of Another Protein.

9 Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* Chapter 7, Unit7 20, doi:10.1002/0471142905.hg0720s76 (2013).

10 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* 7, 248-249, doi:10.1038/nmeth0410-248 (2010).

11 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46, 310-315, doi:10.1038/ng.2892 (2014).

12 Seifi, M. & Walter, M. A. Accurate prediction of functional, structural, and stability changes in PITX2 mutations using in silico bioinformatics algorithms. *PLoS One* 13, e0195971, doi:10.1371/journal.pone.0195971 (2018).

13 Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* 7, e46688, doi:10.1371/journal.pone.0046688 (2012).

14 Choi, Y. & Chan, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 31, 2745-2747, doi:10.1093/bioinformatics/btv195 (2015).

15 Rosenberg, S. *et al.* A recurrent point mutation in PRKCA is a hallmark of chordoid gliomas. *Nat Commun* 9, 2371, doi:10.1038/s41467-018-04622-w

(2018).
16 Graf, S. *et al.* Identification of rare sequence variation underlying heritable pulmonary arterial hypertension. *Nat Commun* 9, 1416, doi:10.1038/s41467-018-03672-4 (2018).
17 Bhattacharya, S. *et al.* Whole-genome sequencing of Atacama skeleton shows novel mutations linked with dysplasia. *Genome Res* 28, 423-431, doi:10.1101/gr.223693.117 (2018).
18 Tubeleviciute-Aydin, A. *et al.* Rare human Caspase-6-R65W and Caspase-6-G66R variants identify a novel regulatory region of Caspase-6 activity. *Sci Rep* 8, 4428, doi:10.1038/s41598-018-22283-z (2018).
19 Bhatnager, R. & Dang, A. S. Comprehensive in-silico prediction of damage associated SNPs in Human Prolidase gene. *Sci Rep* 8, 9430, doi:10.1038/s41598-018-27789-0 (2018).
20 Cunningham, A. D., Colavin, A., Huang, K. C. & Mochly-Rosen, D. Coupling between Protein Stability and Catalytic Activity Determines Pathogenicity of G6PD Variants. *Cell Rep* 18, 2592-2599, doi:10.1016/j.celrep.2017.02.048 (2017).
21 Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 515, 216-221, doi:10.1038/nature13908 (2014).
22 Geisheker, M. R. *et al.* Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains. *Nat Neurosci* 20, 1043-1051, doi:10.1038/nn.4589 (2017).
23 Sahni, N. *et al.* Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* 161, 647-660, doi:10.1016/j.cell.2015.04.013 (2015).
24 Wei, X. *et al.* A Massively Parallel Pipeline to Clone DNA Variants and Examine Molecular Phenotypes of Human Disease Mutations. *PLOS Genetics* 10, e1004819, doi:10.1371/journal.pgen.1004819 (2014).
25 Zhong, Q. *et al.* Edgetic perturbation models of human inherited disorders. *Molecular Systems Biology* 5, doi:10.1038/msb.2009.80 (2009).
26 Barrera, L. A. *et al.* Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science* 351, 1450-1454, doi:10.1126/science.aad2257 (2016).
27 Fuxman Bass, J. I. *et al.* Human gene-centered transcription factor networks for enhancers and disease variants. *Cell* 161, 661-673, doi:10.1016/j.cell.2015.03.003 (2015).
28 Stefl, S., Nishi, H., Petukh, M., Panchenko, A. R. & Alexov, E. Molecular Mechanisms of Disease-Causing Missense Mutations. *Journal of Molecular Biology* 425, 3919-3936, doi:https://doi.org/10.1016/j.jmb.2013.07.014 (2013).
29 Schenone, M., Dančík, V., Wagner, B. K. & Clemons, P. A. Target identification and mechanism of action in chemical biology and drug discovery. *Nature Chemical Biology* 9, 232, doi:10.1038/nchembio.1199 (2013).
30 Pejaver, V. *et al.* MutPred2: inferring the molecular and phenotypic impact of amino acid variants. *bioRxiv* (2017).
31 Ernst, C. *et al.* Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics. *BMC Medical*

Genomics 11, 35, doi:10.1186/s12920-018-0353-y (2018).
32 Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337, 64-69, doi:10.1126/science.1219240 (2012).
33 Miosge, L. A. *et al.* Comparison of predicted and actual consequences of missense mutations. *Proceedings of the National Academy of Sciences of the United States of America* 112, E5189-E5198, doi:10.1073/pnas.1511585112 (2015).
34 Wang, T. *et al.* Probability of phenotypically detectable protein damage by ENU-induced mutations in the Mutagenetix database. *Nature Communications* 9, 441, doi:10.1038/s41467-017-02806-4 (2018).
35 Wang, X. *et al.* Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature Biotechnology* 30, 159-164, doi:http://www.nature.com/nbt/journal/v30/n2/abs/nbt.2106.html#supplementary-information (2012).
36 Meyer, M. J. *et al.* Interactome INSIDER: a structural interactome browser for genomic studies. *Nature Methods* 15, 107, doi:10.1038/nmeth.4540 https://www.nature.com/articles/nmeth.4540#supplementary-information (2018).
37 Li, Q. *et al.* Variants in TRIM22 That Affect NOD2 Signaling Are Associated With Very-Early-Onset Inflammatory Bowel Disease. *Gastroenterology* 150, 1196-1207, doi:https://doi.org/10.1053/j.gastro.2016.01.031 (2016).
38 Wright, A., Charlesworth, B., Rudan, I., Carothers, A. & Campbell, H. A polygenic basis for late-onset disease. *Trends Genet* 19, 10 (2003).
39 Chen, S. *et al.* An interactome perturbation framework prioritizes damaging missense mutations for developmental disorders. *Nat Genet* 50, 1032-1040, doi:10.1038/s41588-018-0130-z (2018).
40 Sufan, R. I., Jewett, M. A. S. & Ohh, M. The role of von Hippel-Lindau tumor suppressor protein and hypoxia in renal clear cell carcinoma. *American Journal of Physiology-Renal Physiology* 287, F1-F6, doi:10.1152/ajprenal.00424.2003 (2004).
41 Kaelin, W. G. The von Hippel-Lindau Tumor Suppressor Protein: An Update. 435, 371-383, doi:10.1016/s0076-6879(07)35019-2 (2007).
42 Sahni, V. *et al.* BMPR1a and BMPR1b signaling exert opposing effects on gliosis after spinal cord injury. *J Neurosci* 30, 1839-1855, doi:10.1523/JNEUROSCI.4459-09.2010 (2010).
43 Racacho, L. *et al.* Two novel disease-causing variants in BMPR1B are associated with brachydactyly type A1. *Eur J Hum Genet* 23, 1640-1645, doi:10.1038/ejhg.2015.38 (2015).
44 Takano, K. *et al.* An X-linked channelopathy with cardiomegaly due to a CLIC2 mutation enhancing ryanodine receptor channel activity. *Hum Mol Genet* 21, 4497-4507, doi:10.1093/hmg/dds292 (2012).
45 Koczok, K. *et al.* A novel point mutation affecting Asn76 of dystrophin protein leads to dystrophinopathy. *Neuromuscul Disord* 28, 129-136, doi:10.1016/j.nmd.2017.12.003 (2018).
46 Aneichyk, T. *et al.* Dissecting the Causal Mechanism of X-Linked Dystonia-

Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* 172, 897-909 e821, doi:10.1016/j.cell.2018.02.011 (2018).
47 Hua, J. T. *et al.* Risk SNP-Mediated Promoter-Enhancer Switching Drives Prostate Cancer through lncRNA PCAT19. *Cell* 174, 564-575 e518, doi:10.1016/j.cell.2018.06.014 (2018).
48 del-Toro, N. *et al.* Capturing variation impact on molecular interactions: the IMEx Consortium mutations data set. *bioRxiv*, doi:10.1101/346833 (2018).
49 Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nature Methods* 7, 741, doi:10.1038/nmeth.1492 https://www.nature.com/articles/nmeth.1492#supplementary-information (2010).
50 Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nature Methods* 11, 801, doi:10.1038/nmeth.3027 (2014).
51 Starita, L. M. *et al.* Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences* 110, E1263-E1272, doi:10.1073/pnas.1303309110 (2013).
52 Starita, L. M. *et al.* Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* 200, 413-422, doi:10.1534/genetics.115.175802 (2015).
53 Wagih, O. *et al.* Comprehensive variant effect predictions of single nucleotide variants in model organisms. doi:10.1101/313031 (2018).

CHAPTER 2:
A MASSIVELY PARALLEL BARCODED SEQUENCING PIPELINE ENABLES GENERATION OF THE FIRST ORFEOME AND INTERACTOME MAP FOR RICE

Context and Personal Contributions

The following chapter is derived from my first author paper (co-first authored with Tommy Vo) by the same name[54] originally published in The Proceedings of the National Academy of Sciences (PNAS). Full authorship and contributions are provided in the original publication, but the following warrant explicit mentioning. Rice RNA samples for constructing the ORFeome described herein were prepared by Rita Sharma in Dr. Pamela Ronald's lab, and initial clones were generated at the Center for Cancer Systems Biology (primarily led by Drs. Pascal Falter-Braun and Marc Vidal's groups), to be fully sequenced and compiled in the Yu lab. Additional clones were generated at the National Institute of Agrobiologically Sciences through collaborations with Drs. Shoshi Kikuichi and Hiroshi Mizuno. Experimental efforts within the Yu lab were initially led by Tommy Vo, and the "PLATE-seq" sequencing method was developed through combined efforts of Tommy Vo, Xiaomu Wei, and Jin Liang. Final experimental validation and completion of this work was carried out by the lab's experimental technicians Nurten Akturk, Christen Rivera-Erick, and Elnur Shayhidin. Preliminary computational groundwork was laid by Michael Meyer before being further developed and completed by myself.

The main text was authored by me with a partial draft by Tommy Vo as an initial

reference point. The text was further edited, contextualized, and refined through the assistance of experts in the field of plant biology; specifically, Susan McCouch and Gaurav Moghe. Additional methods sections were provided by relevant collaborators and integrated into the manuscript by me. I was responsible for all computational analyses and generation of all figures included in the final manuscript, excepting the conservation analyses shown in **Figure 10** which was completed by Lars Kruse.

Abstract

Systematic mappings of protein interactome networks have provided invaluable functional information for numerous model organisms. Here we develop PCR-mediated Linkage of barcoded Adapters To nucleic acid Elements for sequencing (PLATE-seq) that serves as a general tool to rapidly sequence thousands of DNA elements. We validate its utility by generating the first ORFeome for *Oryza sativa* covering 2,300 genes and constructing a high-quality protein-protein interactome map consisting of 322 interactions between 289 proteins; expanding the known interactions in rice by roughly 50%. Our work paves the way for high-throughput profiling of protein-protein interactions in a wide range of organisms.

Introduction

The genomics revolution has democratized sequencing and structural annotation of genomes, however, assigning functions to predicted genes remains an important unsolved challenge. Identification of protein-protein interactions can help advance functional annotation in sequenced genomes. The first step in systematic, genome-wide mapping of protein-protein interactions involves the construction of a comprehensive

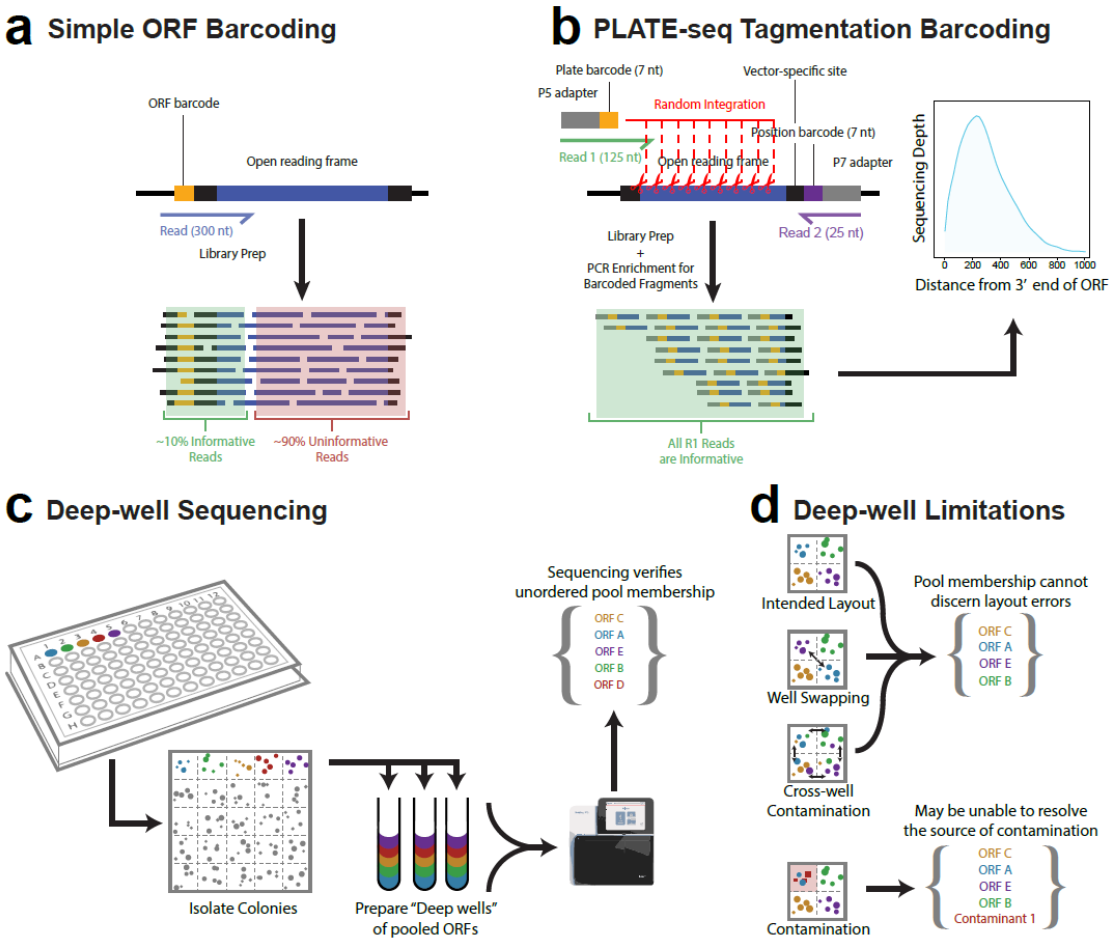


Figure 3. Schematic comparison of other high-throughput sequencing strategies to PLATE-seq.

a, Simple barcoding strategies typically append individual barcodes at the beginning of each ORF. A lengthy vector sequence generally separates the barcode from the beginning of the ORF, so even a long 300 bp read only provides coverage limited to a small portion from the 5' end of the ORF. Moreover, the library prep generates a substantial fraction of uninformative reads with no barcode. **b,** By contrast, the PLATE-seq approach employs Tn5 tagmentation to randomly insert the P5 adapter and plate barcode within the ORF. Thus, even employing a much shorter 150 bp total read, nearly 1,000 bp from the 3' end of the ORF can be sequenced reliably. The kernel density estimate shows the average sequencing depth derived from PLATE-seq results from 94 ORFs included in our human positive control plate. Moreover, all fragments generated through the PLATE-seq library prep contain both barcodes and thus are all informative. **c,** Schematic representation of a deep-well sequencing strategy which pools many wells together to report the overall ORF membership across an entire plate but cannot verify the location of each ORF. **d**, The deep-well sequencing strategy is consequentially unable to differentiate permutations on the same set of ORFs, or verify the source of any contaminants.

set of high-quality open reading frames (ORFeome). To accomplish this, tens of thousands of clones must be sequenced to ensure that only correct, full-length clones are retained. Traditionally, this is achieved through the labor-intensive and cost-prohibitive process of Sanger sequencing of each individual clone. Simple barcoding strategies that append a barcode at the beginning or end of each ORF are problematic because they require thousands of unique barcodes, provide extremely limited ORF coverage, and generate a high fraction of uninformative reads that contain no barcode (**Figure 3a**). A deep-well-pooling approach has recently been used to sequence ORFeome libraries[55,56]; however, these smart pooling approaches cannot accommodate the inclusion of homologous ORFs in one pool, rely on concrete prior knowledge regarding plate layout, and cannot detect potential cross-contamination between wells (**Figure 3c and d**).

Here, we develop a massively parallel sequencing approach called PCR-mediated Linkage of barcoded Adapters To nucleic acid Elements for sequencing (PLATE-seq)—a broadly utilizable approach for rapid sequencing of thousands of DNA elements. We validate the utility of PLATE-seq by developing an ORFeome for rice and constructing a high-quality, experimentally validated protein-protein interactome map of this important monocot species. Despite being a staple food for over half of the world's population, and an important model for monocot genomics, empirical annotations currently cover fewer than 5% of genes in the rice genome[57]. Indeed, understanding the function of plant genes has become a major bottleneck for the field of plant biology as a whole.

The majority of plant functional studies have been carried out in the dicot model

organism *Arabidopsis thaliana*[58,59], and to date, only limited characterization has been performed in monocot species. Combatting this disparity in annotation, the development of a comprehensive, high-quality ORFeome for *Oryza sativa* would enable large-scale reverse-proteomics studies—including the systematic mapping of protein-protein interactions—and thus, would significantly expand the functional genomics toolkit for plants.

A full-length cDNA clone library (FLcDNA) has previously been reported in *O. sativa*[60]; however, such libraries are not suitable for high-throughput studies. Specifically, FLcDNA clones contain 5' and 3' UTRs, and therefore are not amenable to C- and N-terminal tagging required for most functional studies (e.g. yeast two-hybrid). Furthermore, these clones were derived from pools of clones rather than single colonies—resulting in contamination of up to 80% of clones. While comprehensive Gateway-compatible ORFeomes amenable to high-throughput cloning and expression analysis have been extensively utilized in model organisms[56,58,61-64], that of *A. thaliana* is currently the only ORFeome available for any plant. Although a handful of functional studies have been completed in rice by first cloning proteins of interest[65,66], the lack of a unified ORFeome is a serious constraint to further advances.

In this study, we produce a fully-sequenced, single-colony-derived, Gateway-cloning-compatible ORFeome for rice; the first for any monocot species. Adapting a proven yeast two-hybrid (Y2H) screening approach[67-69], we leverage the power of PLATE-seq and the ORFeome to systematically generate a high-quality rice protein-protein interaction network. Our work—while contributing a novel pipeline broadly useful for the biological community—expands the known map of the rice interactome,

paves the way for future high-throughput rice biology studies, and provides the a systematic characterization of a monocot genome, an internationally important crop species, and model organism.

Results and Discussion

PLATE-seq achieves robust parallel identification of ORFs comparable to Sanger sequencing

To facilitate the development of the rice and future ORFeomes, we developed PLATE-seq, a massively parallel barcoded sequencing approach to validate identities and locations within complex DNA libraries. Our PLATE-seq methodology (**Figure 4a**) begins with a library of either single or pooled clones arrayed across a 96-well format. Individual PCR amplifications in each well append a unique position-specific barcode and the primary TruSeq sequencing adapter to the DNA product. Samples are then pooled together on a per-plate basis and tagmented by Tn5 transposase. The tagmentation reaction inserts a unique plate-specific barcode and the secondary TruSeq sequencing adapter at a random location within the ORF. Next, universal primers are used in a low-cycle PCR to enrich for clone fragments containing both position- and plate-specific barcodes (**Figure 4b** and **Table 1**). Finally, amplicons from all plates are pooled together and subjected to massively parallel Illumina sequencing. The paired-end sequencing setup generates R2 reads just long enough to span the position-specific barcode and R1 reads maximized to span both the plate-specific barcode and the ORF sequence. Notably, the final low-cycle PCR enrichment ensures that all paired reads contain both barcodes and can be informatively mapped back to their exact source well. Moreover, because Tn5 tagmentation acts at a random position within the ORF

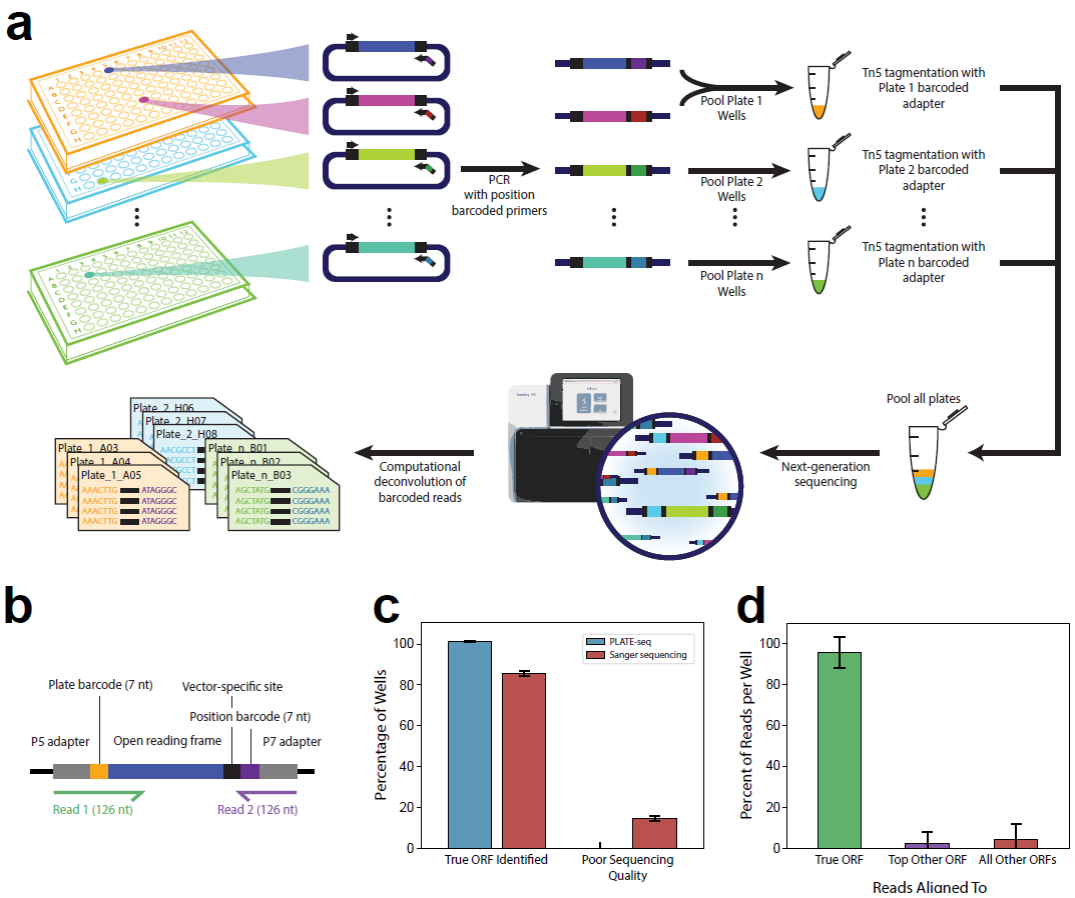


Figure 4. A massively parallel approach to comprehensively index DNA libraries. **a,** A schematic illustration of the PLATE-seq pipeline. **b,** Barcoding design of PLATE-seq output products. **c,** Fraction of ORFs in the human positive control plate that could be correctly identified by either PLATE-seq or Sanger sequencing (n=94). Data are shown as + / - standard deviation. **d,** Fraction of PLATE-seq reads mapping to true human positive control ORFs or other ORFs (n=94). Data are shown as + / - standard deviation.

the fragments generated can theoretically span the entire ORF. However, in practice cluster formation and amplification using current Illumina sequencers become inefficient for prohibitively large fragments. Therefore, we optimize our methods to provide coverage of roughly the last 800-1,000 bp of each ORF (**Figure 3b**).

In order to benchmark the accuracy of PLATE-seq, we implemented the method on a test plate of 94 human ORF clones selected from the sequence-verified human ORFeome 8.1 library[56], demonstrating that PLATE-seq correctly identified the true ORF in 100% of the test cases (**Figure 4c**). Determination of clone identify per-well

was made by aligning the reads to the entire 8.1 reference library to calculate the fraction of reads in each well contributed by each ORF. Importantly the vast majority of PLATE-seq reads aligned to the true ORF with only a minor fraction aligning to an incorrect ORF introduced through experimental contamination or alignment ambiguities (**Figure 4d**). Moreover, the ability of PLATE-seq to detect these minor artifacts demonstrates that it possesses the resolution to discern the relative abundances of multiple clones in a pooled setup. By contrast, Sanger sequencing is ill-suited to handle a pooled setup or resolve contamination errors. Consequently, re-sequencing by Sanger was more prone to sequencing quality failures and was unable to fully reconfirm the identities of all control clones in one attempt (**Figure 4c**).

Although our applications of PLATE-seq were limited to determination of the identity of the ORF(s) in each well, we note that the applications could be extended beyond these. If the reference sequence for the clones being sequenced were unknown, *de novo* sequence assembly could be applied to each set of reads after location deconvolution. To demonstrate this, we input reads from one of the wells of our human control plate into a contig assembly script. The pairwise alignment between our PLATE-seq reconstructed sequence, the Sanger Sequencing result from the same clone, and the true sequence of the human ORF (*BIRC7*) is shown (**Figure 5a**). Our reconstructed sequence perfectly matched the true *BIRC7* clone sequence; the one mismatch recapitulated a synonymous SNP that was reported for the *BIRC7* clone when the human ORFeome 8.1 library was initially released[56]. By contrast, the Sanger Sequencing result only achieved partial coverage and included many sequencing errors (**Figure 5b**). Although it may be possible through repeated trials to achieve Sanger Sequencing

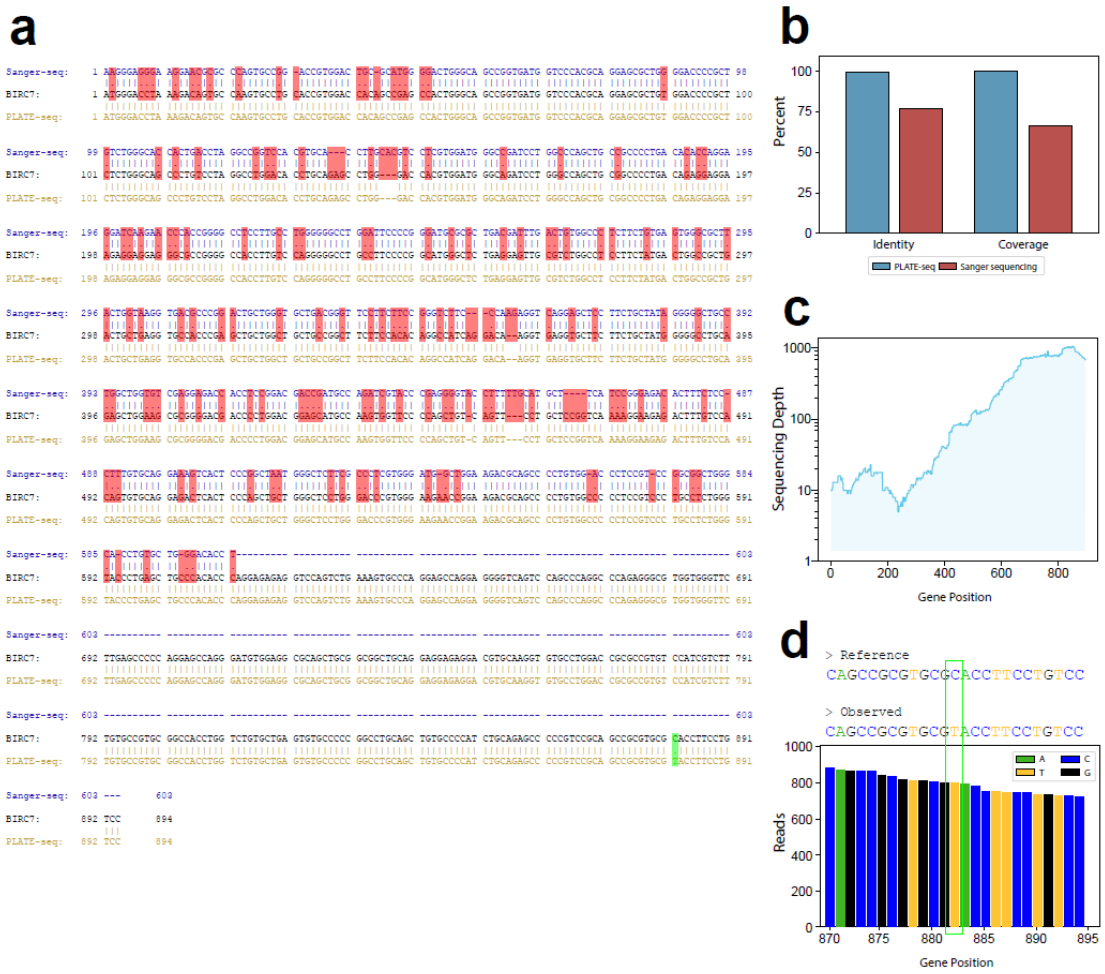


Figure 5. Demonstration of sequence reconstruction and variant calling as alternative applications for PLATE-seq.

a, A contig assembly script was applied to reads from one of the wells of our human control plate containing an ORF for the human gene *BIRC7* (ORF ID 7606 in the human ORFeome 8.1 library). The alignment between the *BIRC7* reference (center), Saner Sequence obtained for this clone (top, blue), and sequence reconstructed from PLATE-seq reads (bottom, yellow) is shown. Disagreements in the alignment are highlighted in red. **b,** Bar plots summarizing the quality of each alignment. For convenience vector backbone sequence from the Sanger and PLATE-seq sequences are trimmed so the alignment only spans coding sequence of *BIRC7*. The Sanger Sequencing result sequenced positions 1-613 (68.6% coverage) with 76.2% identity to the reference sequence. The PLATE-seq result sequenced achieved perfect coverage and percent identity (ignoring one mismatch from a known C882T variant in the clone). **c,** A distribution showing the number of PLATE-seq reads aligned to each position of the *BIRC7* gene. **d,** PLATE-seq reads can be used to identify variants between a reference alignment and the sequenced clone. The stacked read counts for each possible allele observed from all reads aligned to positions 870-894 of *BIRC7* are shown. Highlighted is a C882T variant supported by 784 reads (sufficient coverage to make a definitive call) and which was previously reported for this ORF when the 8.1 library was initially sequenced. This same variant is also highlighted green in panel **a**.

results of equal quality to our PLATE-seq reconstructed sequence, we emphasize that PLATE-seq sequence reconstruction can be applied simultaneously for hundreds of clones from one round of sequencing. Our ability to detect the C882T variant highlights further potential outside of *de novo* sequence reconstruction. By replacing the final step of PLATE-seq with a variant caller it is possible to identify and uniquely assign SNPs among hundreds of copies of the same gene. To demonstrate this, we identify the same C882T variant in *BIRC7* this time by aligning all genes to the known *BIRC7* references and applying a variant caller to the read pileup (**Figure 5d**). We note that using PLATE-seq to call SNPs from a known reference would be more scalable than sequence reconstruction since there would be less need for robust sequencing depth.

A draft rice ORFeome captures 2,300 rice genes across a diverse functional spectrum

Having demonstrated PLATE-seq's capacity for precise parallel-determination of the identities and exact locations of clones within a complex library, we next sought to systematically construct and sequence-verify a first version of the rice ORFeome. To reduce the daunting scale of the full *O. sativa* genome, we initially used RiceNet[70,71] to prioritize 3,269 genes predicted to be most closely associated with a seed set of 89 genes (**Table 2**) that had previously been linked to biotic or abiotic stress tolerance—either through experimental validation or association with validated stress tolerance genes[70,72]. To ensure maximum recovery of these genes, we designed primers for one representative ORF from each gene (**Table 3**) and amplified them using cDNA obtained from 40 combinations of developmental stages and stress exposures (**Figure 6a** and **Table 4**). All cloning was carried out by Gateway recombination-based cloning, to enable versatility for downstream cloning into various Gateway-

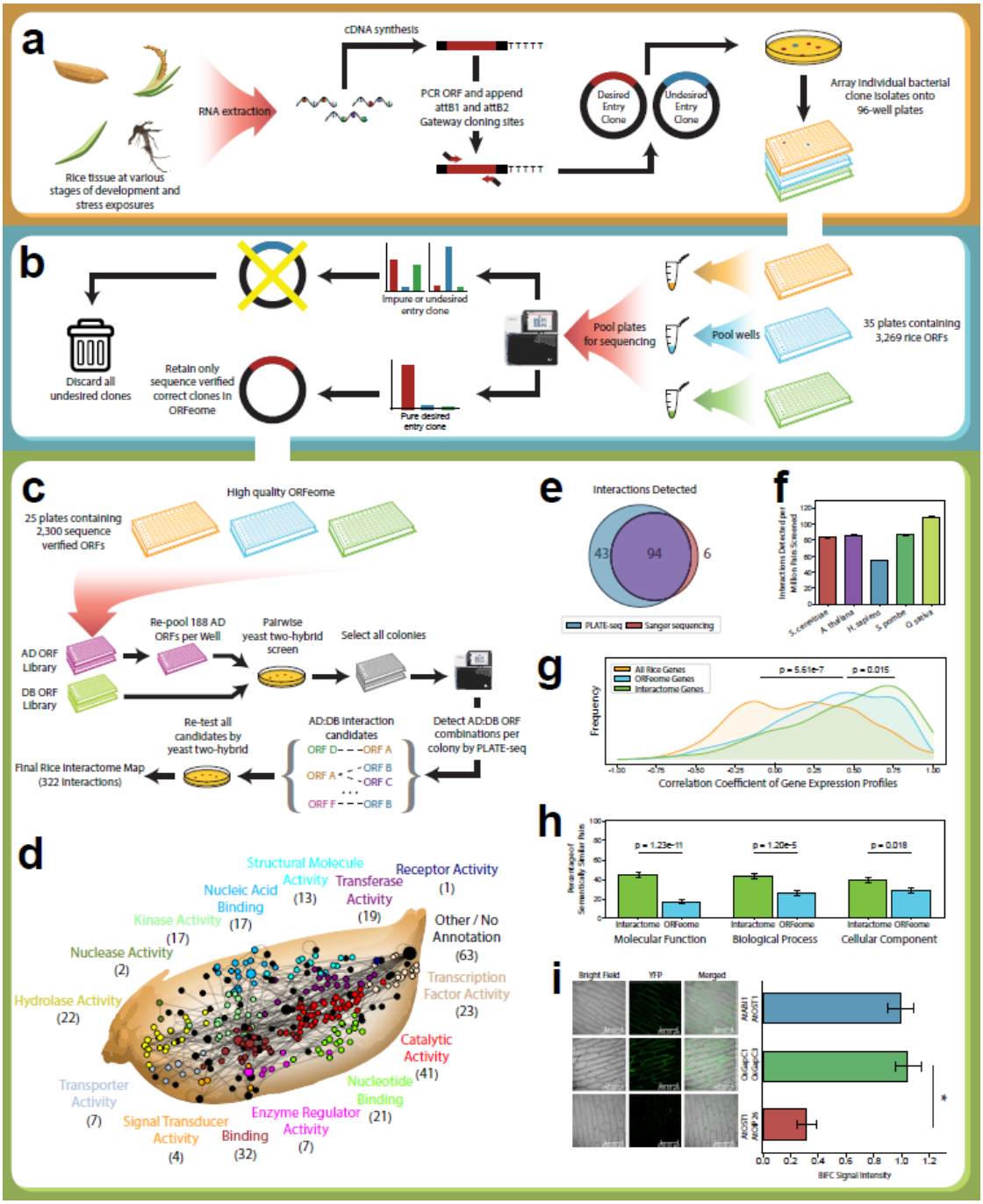


Figure 6. A high-quality ORFeome and binary protein interactome in *O. sativa*.
a, A schematic illustration of the construction of the original rice ORF library. **b,** The raw ORF library was sequence verified by PLATE-seq and only correct clones were retained. **c,** Massively parallel Y2H screening was performed to detect putative Y2H positive interactors, and each interaction was verified by pairwise retesting. **d,** Network representation of the full rice interactome spanning 322 interactions across varied functional annotations. **e,** Comparison of the successful Y2H positive interactor detection rate across 8 plates of putative interactors using either PLATE-seq or Sanger sequencing. **f,** Comparison of the detection rates from previous

high-throughput Y2H interactome screens to our rice interactome. Data are shown as + / - standard error. **g,** Comparison of the distributions of gene co-expression between random rice gene pairs, random pairs sampled from our ORFeome, or our interactome pairs (n=322). Co-expression is reported as Spearman rank correlation coefficients between gene expressions from 11 different rice tissue samples. Expression values were significantly more correlated among interactome pairs compared to random ORFeome pairs (p=0.015 by two-sided Kolmogorov–Smirnov test). However, both interactome pairs and ORFeome pairs were significantly more co-expressed than random genome pairs (p-value=5.61e-7 by two-sided Kolmogorov–Smirnov test). **h,** Comparison of the fraction of detected interactions vs. random ORFeome pairs that share similar molecular function (MF), biological processes (BP), or cellular component (CC) gene ontology annotations. Similar GO annotation is defined as sematic similarity score ≥ 0.75 as reported by GOssTo. Detected interactions were significantly more likely to be similarly annotated among all three classifications (MF, n=236, p=1.23e-11; BP, n=254, p=1.20e-5; CC, n=206, p=0.018; all tests are one-tailed Fisher's Exact Test). Interactions lacking annotation for a specific GO term were excluded from each category. Data are shown as + / - standard error. **i,** Representative BiFC confocal fluorescence images for the positive control (AtABI1-AtOST1), for the rice protein pair encoded by LOC_Os08g03290 (OsGapC1) and LOC_Os02g38920 (OsGapC3), and for negative control (AtOST1-AtOIP26). Average ratios of BiFC signals relative to the AtABI1-AtOST1 positive control. Data are shown as +/- standard deviation. Asterix (*) denotes significance ($p < 0.001$) as ascertained by two-tailed t-test.

compatible expression vectors for functional studies. To prevent contamination of the rice ORFeome with unwanted cloning byproducts (e.g. PCR artifacts), we picked two single colonies per ORF, determined the true identity of every isolate by PLATE-seq, and eliminated any clones that did not align with the intended sequence (**Figure 6b**). As a secondary check that only full-length clones were retained, approximate clone lengths were verified by gel electrophoresis (**Figure 7**).

In total, our final sequence-verified ORFeome provides one representative ORF clone for each of 2,300 rice genes sampled throughout the *O. sativa* genome (**Figure 8a** and **Table 5**). Analysis of the final composition of our ORFeome showed some evidence that the success rate from our cloning method was higher for certain genes. For instance, although the distribution of gene lengths was similar to that of the entire *O. sativa* genome and prioritized gene set, it did show bias towards shorter genes (**Figure 8b**). Moreover, we observed disproportionate representation of highly expressed genes. Although this shift was introduced within our initial prioritized set of

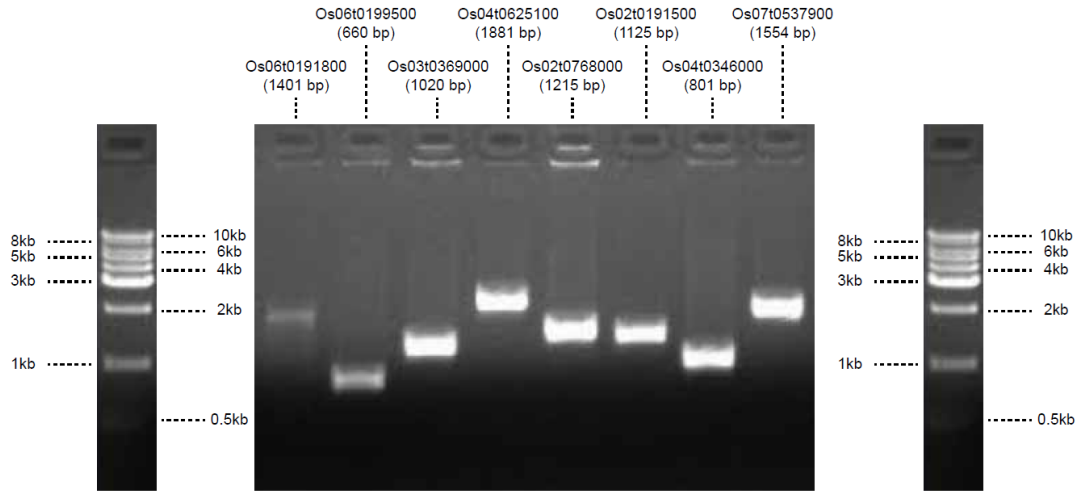


Figure 7. Image of PCR amplicons from a subset of verified *O. sativa* ORFs.
A representative gel in which one row of clones from the final *O. sativa* ORFeome was amplified using primers described in **Supplementary Table 3**. These gels were used as a sanity check to confirm the presence and approximate lengths of the desired ORFs. The DNA molecular weight standard lane has been cropped from the same gel image and reproduced on both sides for easier comparison.

genes, it was further exacerbated in our final ORFeome (**Figure 8c**). These skews are consistent with known consequences of PCR amplification bias[73] and suggest that greater effort may be required in the future when cloning long or lowly expressed genes. Nonetheless, our ORFeome captures a wide diversity of biological processes broadly representative of the functional distributions over the entire *O. sativa* genome (**Figure 8d-f**). We note that we do not observe extensive evidence of functional bias within our ORFeome despite the fact that the seed genes used for gene prioritization came from a specific functional study[72]. High scoring RiceNet predictions should accurately capture true functional associations with our seed set and on their own may have contributed a functional bias. However, because we used a low confidence threshold when prioritizing genes to clone, a large number of less precise predictions may have counteracted this. The clear exception to this came from one of our seed genes,

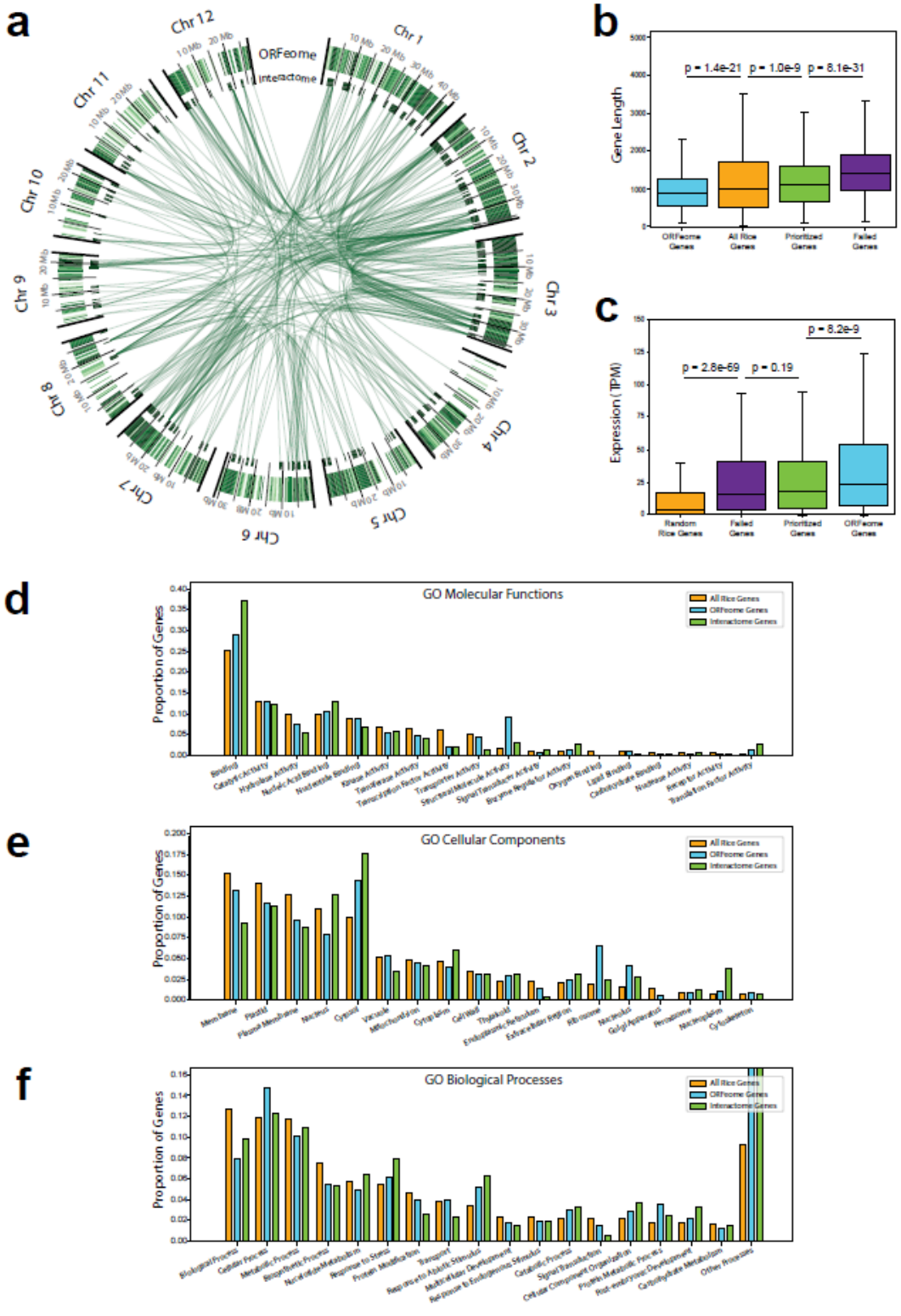

a
ORFeome
Interactome
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
b
Gene Length
p = 1.4e-21
p = 1.0e-9
p = 8.1e-31
ORFeome Genes
All Rice Genes
Prioritized Genes
Failed Genes
c
Expression (TPM)
p = 2.8e-69
p = 0.19
p = 8.2e-9
Random Rice Genes
Failed Genes
Prioritized Genes
ORFeome Genes
d
GO Molecular Functions
Proportion of Genes
All Rice Genes
ORFeome Genes
Interactome Genes
Binding
Catalytic Activity
Hydrolase Activity
Nucleic Acid Binding
Nucleotide Binding
Kinase Activity
Transferase Activity
Translation Factor Activity
Transporter Activity
Structural Molecule Activity
Signal Transducer Activity
Enzyme Regulator Activity
Oxygen Binding
Lipid Binding
Carbohydrate Binding
Nuclease Activity
Receptor Activity
Transcription Factor Activity
e
GO Cellular Components
Proportion of Genes
All Rice Genes
ORFeome Genes
Interactome Genes
Membrane
Plastid
Plasma Membrane
Nucleus
Cytosol
Vacuole
Mitochondrion
Cytoplasm
Golgi
Thylakoid
Endoplasmic Reticulum
Extracellular Region
Ribosome
Nucleolus
Golgi Apparatus
Peroxisome
Nucleoplasm
Cytoskeleton
f
GO Biological Processes
Proportion of Genes
All Rice Genes
ORFeome Genes
Interactome Genes
Biological Process
Cellular Process
Metabolic Process
Biogenesis Process
Nucleotide Metabolism
Response to Stress
Protein Modification
Transport
Response to Abiotic Stimulus
Multicellular Development
Reproduction & Embryonic Development
Catabolic Process
Signal Transduction
Cellular Component Organization
Protein Metabolic Process
Biosynthetic Development
Carbohydrate Metabolism
Other Processes

Figure 8. Summary statistics on the *O. sativa* ORFeome.

a, Circle plot depicting the density of 2,300 ORFeome genes (outer circle) and 289 interactome genes (inner circle) across the *O. sativa* genome. Internal arcs represent interacting genes. In general, genes were evenly sampled across the genome. **b,** Comparison of the gene lengths of all annotated rice genes, 2,300 rice ORFeome genes, the initial prioritized set of 3,269 rice genes, and the set of 969 genes for which we failed to obtain a successful clone. Significant differences were observed between all groups (after down sampling to match the smallest group) as ascertained by two-sided Kolmogorov–Smirnov test. **c.** Comparison of the average expression values among the same four groups. The initial prioritized gene set and failed set were significantly more highly expressed compared to the random subset of genes from the whole genome. To a lesser degree, ORFeome genes were significantly more highly expressed than the prioritized gene set and failed set. All p-values were ascertained by Kruskal-Wallis multiple comparison after down-sampling all groups to match the smallest group. **d-f** Fraction of genes in the rice genome, rice ORFeome, and rice interactome that are represented in GO molecular function, cellular component, or biological process categories, respectively.

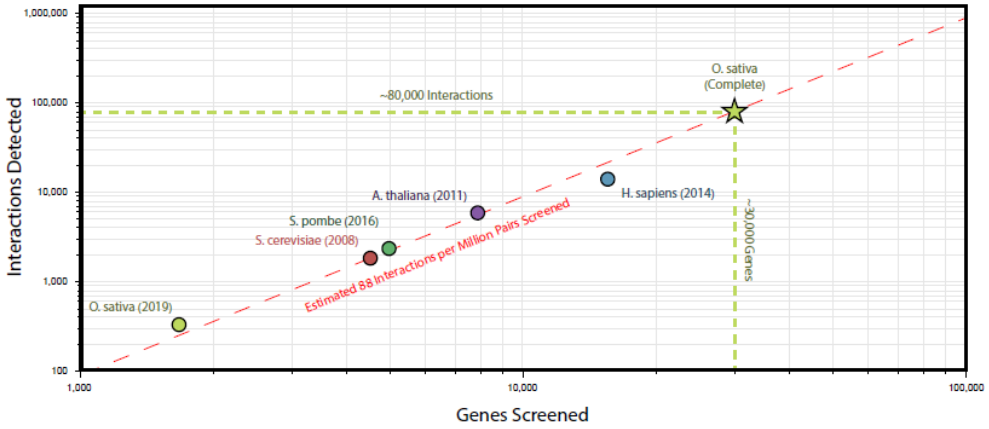
a 60S ribosomal protein L14 (LOC_Os02g40880), which contributed an enrichment for additional ribosomal proteins to our ORFeome (**Figure 8e**). Compared to the other seed genes, this ribosomal gene was a hub for functionally related true interactors that could be predicted with high confidence owing to the high degree of annotation transfer for this highly conserved complex available from other organisms. Thus, despite minor experimental limits, our fully-validated rice ORFeome represents a wide and largely unbiased functional spectrum of the *O. sativa* genome.

A systematic yeast two-hybrid screen reveals 322 rice protein-protein interactions

Because proteins function primarily by physically interacting with each other[74-76], protein interactome networks provide a crucial resource to discover functions associated with protein-coding genes. To date, these interactome maps have been pivotal in uncovering functional relationships between proteins in a wide variety of organisms[59,69,77,78]. Although several resources have applied homology-based annotation transfer to predict *O. sativa* protein-protein interactions[70,71,79-81], a large-scale experimental survey is yet to materialize. A tandem affinity purification method has

been employed to detect rice kinase complex associations[65,82] and a few Y2H studies have probed other specific functional subcomponents of the interactome[66,72,83]. However, these Y2H studies have relied on cDNA libraries that generally produce lower quality Y2H interactome mapping compared to sequence-verified, full-length ORFeome clone libraries[68]. In order to provide a large-scale, experimentally-validated rice interactome map, we tested all pairwise combinations of Y2H-amenable proteins encoded by our rice ORFeome (1,671 x 1,671 ~ 2.7 million protein-protein pairs tested) using the same high-throughput yeast two-hybrid (Y2H) assay we previously used to generate the budding yeast, human, and fission yeast interactome networks[68,69,84] (**Figure 6c**). Previous high-throughput Y2H screening-sequencing approaches identified interaction candidates by screening 188 AD ORFs against one DB ORF at a time but were subject to a bottleneck because each positive colony must be sequenced individually to determine the AD interactors[67-69]. Leveraging PLATE-seq we were able to uncover all protein pair interaction candidates from our Y2H screen in one sequencing step and, subsequently, validate them by pairwise Y2H retest. Our full workflow resulted in a high-quality rice protein interactome network consisting of 322 high-quality interactions between 289 rice proteins (**Figure 6d** and **Table 6**) across a wide span of molecular processes and cellular localizations (**Figure 8d-f**). Notably, sequencing by PLATE-seq boasted higher identification of truly interacting protein pairs when compared to Sanger sequencing (**Figure 6e**), and our overall detection rate was comparable to previous Y2H interactome screens (**Figure 6f** and **Figure 9**).

a



b

Saccharomyces cerevisiae (Yu *et al.* 2008)	*Arabidopsis thaliana* (Arabidopsis Interactome Mapping Consortium 2011)	*Homo sapiens* (Rolland *et al.* 2014)	*Saccharomyces pombe* (Vo *et al.* 2016)	*Oryza sativa* (Wierbowski *et al.* this study)
3,917x5,246 Search Space (20.5 Million Pairs)	8,044x7,771 Search Space (62.5 Million Pairs)	15,517x15,517 Search Space (240.7 Million Pairs)	4,989x4,989 Search Space (24.9 Million Pairs)	1,671x1,671 Search Space (2.80 Million Pairs)
1,809 Interactions Detected	5,664 Interactions Detected	13,944 Interactions Detected	2,278 Interactions Detected	321 Interactions Detected
88.0 Interactions per Million Pairs Screened	90.6 Interactions per Million Pairs Screened	57.9 Interactions per Million Pairs Screened	91.5 Interactions per Million Pairs Screened	115.0 Interactions per Million Pairs Screened

Figure 9. Expanded summary of interaction detection rates among previous high-throughput Y2H interactome screens.
a, Visualization of the number of interactions detected per number of pairwise genes screened across the *S. cerevisiae*, *A. thaliana*, *H. sapiens*, *S. pombe*, and *O. sativa* interactome networks. The red dotted line represents the average detection rate across the five interactomes (88 interactions per million pairs screened). The *O. sativa* star depicts an estimated total number of ~80,000 interactions that could be detected from a completed ORFeome of ~30,000 rice genes. **b,** A complete table comparing the raw size of the search space, number of interactions detected, and interactions detected per million pairs screened across the five interactomes.

To support the biological relevance of our network, we analyzed the quality of our rice interactome map based on the functional relationship between interacting proteins compared to non-interacting proteins. For a physiologically relevant protein-protein interaction to occur, the corresponding genes must be expressed under similar

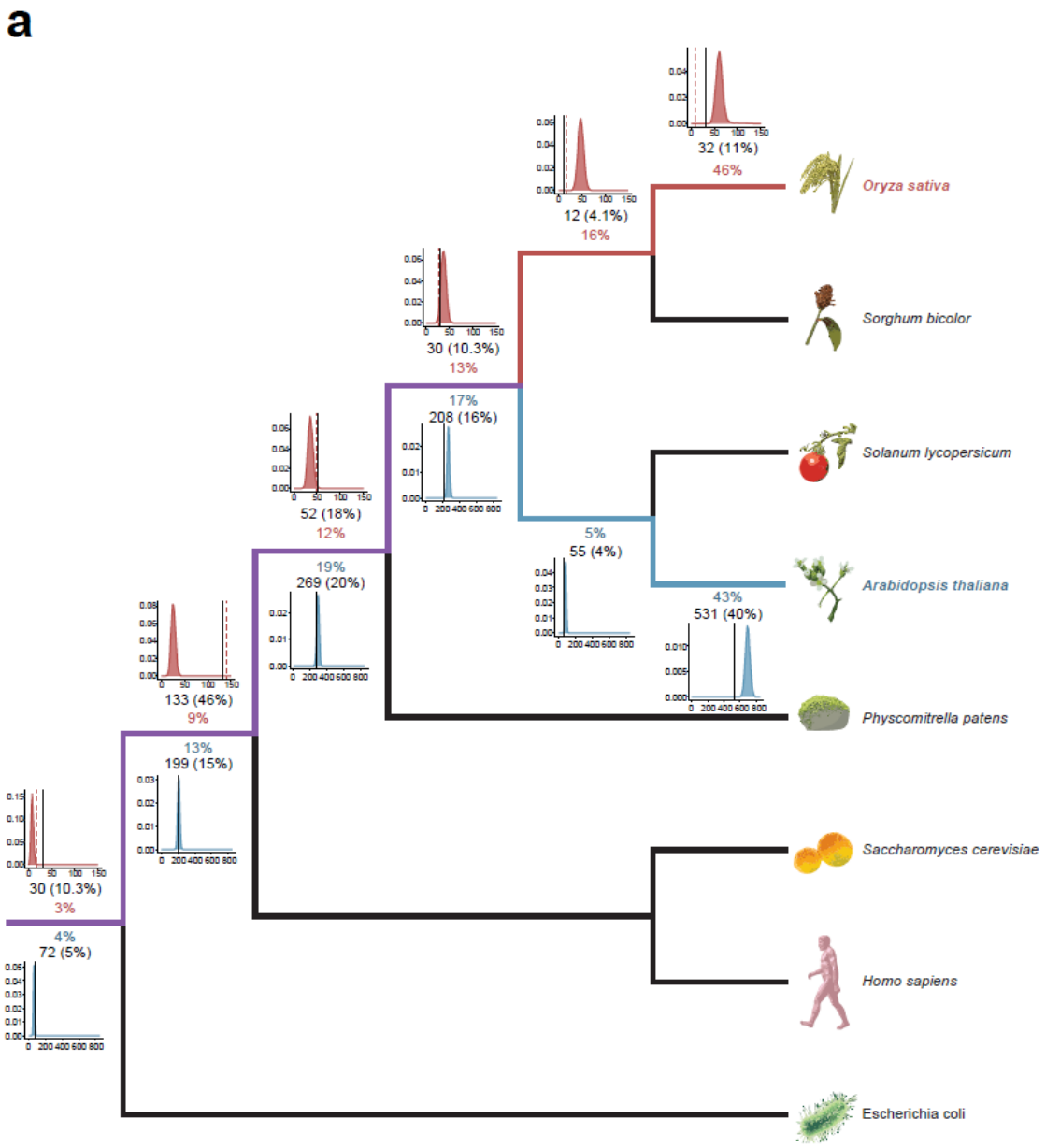


Figure 10. Conservation analysis of interacting genes in rice and *Arabidopsis*.
Phylogenetic tree for eight species included in the conservation analysis of interacting genes reported in *Oryza sativa* (top, red) and *Arabidopsis thaliana* (bottom, blue). All interacting genes (n=289 and n=1,334 for *O. sativa* and *A. thaliana* respectively) were binned into the most ancestral node wherein an orthologous gene could be detected among related species. The histograms at each node represent the distributions of 1,000 bootstrap replicates using randomly sampled non-interacting genes. The mean of each distribution is reported by the colored percentage above or below each distribution. The black lines mark the number of interacting genes conserved at each node in the tree. The exact count of interacting genes is reported in black above or below each distribution). The dotted red line (*O. sativa* only) marks the expected number of genes conserved at each node derived from our ORFeome. A black/red line to the left of the distribution indicates under-representation of interacting genes compared to

background expectation, while a black/red line to the right indicates over-representation. While a statistically significant difference between the actual count of interacting genes and the mean of random expectation was detected at all nodes ($p < 2.2e-16$ as determined by z-test), the effect size of the difference (absolute [% observed - % expected]) was low in *A. thaliana*, compared to *O. sativa*. This finding indicates that the interacting genes in *A. thaliana* interactome show little to no evolutionary bias, while the *O. sativa* interactome and ORFeome are biased towards more conserved, and hence, more widely distributed genes.

spatiotemporal conditions. We demonstrate that interacting genes exhibit higher co-expression compared to random gene pairs selected from our ORFeome or the entire rice genome (**Figure 6g**). However, we do note that gene co-expression is one of the features used in the RiceNet predictions we used to prioritize genes for inclusion in our first draft ORFeome. As a consequence of this selection all pairs within the ORFeome were already significantly co-expressed. We additionally note that our ORFeome captures a high proportion of highly conserved genes (**Figure 10**), potentially also as a consequence of RiceNet prioritization since such genes may borrow evidence from homologs in other organisms. While these caveats must be considered when interpreting our analyses, the high conservation rate among interacting genes highlights the broad applicability of our interactome for high-confidence annotation transfer to other plant species. Because biological pathways involve protein-protein interactions, we also expect interacting protein pairs to be enriched in similar functional annotations. We show that compared to a random sampling of protein pairs from our ORFeome, our interactome map contains a significantly higher proportion of similarly annotated protein pairs across all classifications of gene ontology (GO) terms (**Figure 6h**). Finally, to demonstrate the robustness and accuracy of our Y2H approach, we validated a subset of our interactions through an orthogonal assay. We performed bimolecular fluorescence complementation (BiFC) on a random subset of seven interactions. Six out

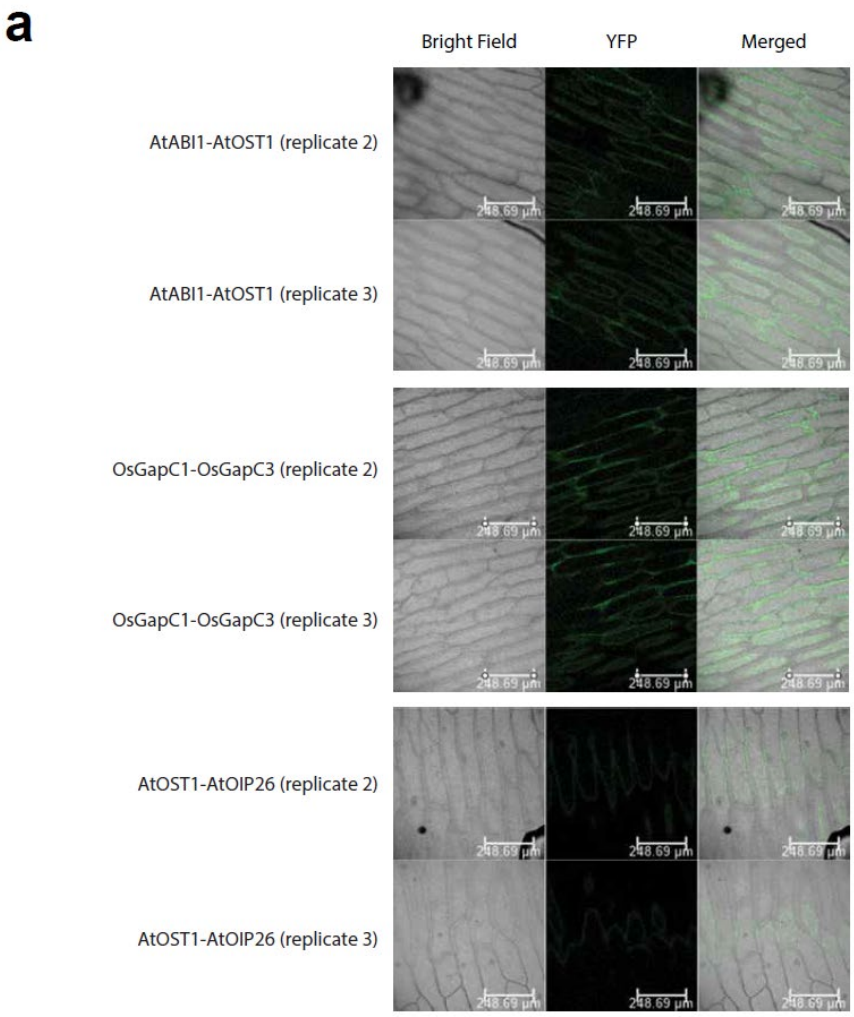


Figure 11. Images of BiFC biological replicates of Fig. 2i.
Confocal fluorescence images of biological replicates for the positive control (AtABI1-AtOST1), for the rice protein pair encoded by LOC_Os08g03290 (OsGapC1) and LOC_Os02g38920 (OsGapC3), and for negative control (AtOST1-AtOIP26).

of the seven interactions (85.7%) were robustly recapitulated (**Table 7**). As a representative example, we confirmed a novel interaction between two predicted glyceraldehyde-3-phosphate dehydrogenases, LOC_Os08g03290 (OsGapC1) and LOC_Os02g38920 (OsGapC3) (**Figure 6i**, **Figure 11**).

Our rice interactome map increases the current literature interactome map by 50% and uncovers conserved interactions

Finally, we compared our reported interactions against the previous literature. Using a

curated set of high-quality binary protein-protein interactions[85] compiled from seven primary interaction databases—BioGRID[86], MINT[87], iRefWeb[88], DIP[89], IntAct[90], HPRD[91,92], MIPS[93], and the PDB[94,95]—we uncovered 237 interactions in *O. sativa*. We supplemented this set with an additional 372 interactions from a high-throughput Y2H rice-kinase interactome screen[66] for a total of 609 previously reported interactions. Notably, our additions to the rice interactome map cover a unique search space; among literature interactions only seven were recapitulated by our screen and only about five percent could have theoretically been recapitulated from our ORFeome using an 80% sequence identify cutoff (**Figure 12a and b**). Two of our interactions—one between Elongation factor 1 delta (LOC_Os07g42300) and Elongation factor 1 beta (LOC_Os07g46750), another between a DUF851 domain containing protein (LOC_Os04g49660) and Serine/threonine protein kinases OSK4 (LOC_Os08g37800)—showed near exact sequence identity to a previously reported interaction. Thus, our network greatly expands the known rice protein interactome. We repeat this analysis for the interactomes from four additional organisms (**Figure 12c-f**). For the distantly related organisms—yeast, human, and *E. coli*—we note that for nearly all rice interactions where both interacting proteins had a homolog in the other organism, a homologous interaction was in fact reported; potentially suggesting sampling from a core interactome whose functionality is tightly conserved across species. In *A. thaliana* by contrast although more conserved interactions were detected, there was a larger discrepancy between the number of rice interactions that could have been detected using *Arabidopsis* homologs and the number homologous interaction between those homologs that actually were reported. This may indicate some interaction

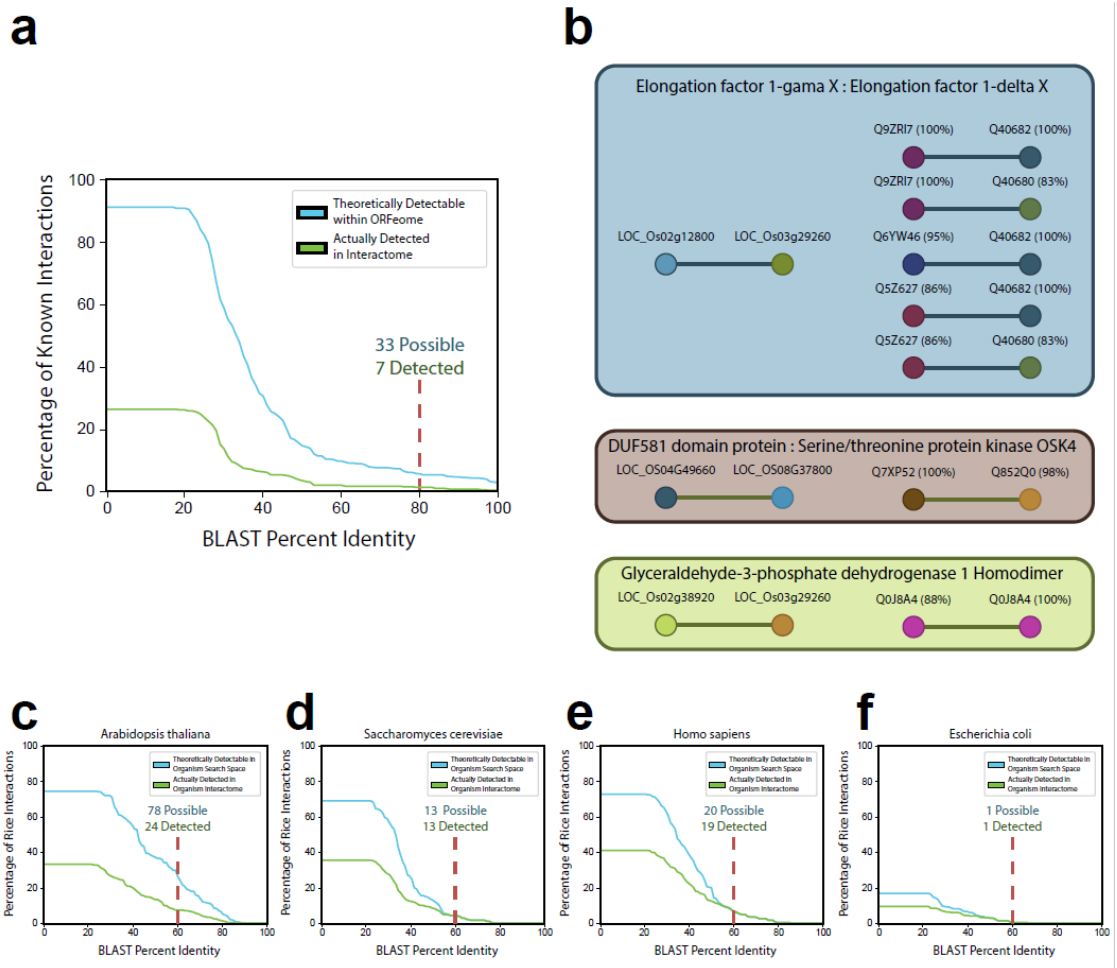


Figure 12. Recapitulation rate of previously reported *O. sativa* interactions by Y2H screen. **a,** All 609 previously reported rice interactions were compared against our ORFeome genes using BLAST to determine the fraction of interactions that could theoretically be recapitulated within our Y2H screen at varying percent identity cutoffs. These BLAST results were intersected with our interactome to determine the fraction of interactions that were actually detected. Using a lenient 80% percent identity 7 out of 33 interactions (21%) were recapitulated by our Y2H screen. **b,** Three of our detected interactions (left) recapitulated seven previously reported interactions (right). Our interactions are labeled using MSU IDs whereas the previously reported interactions are labeled using UniProt IDs. Only the top interaction (Q9ZRI7 and Q40682) is a perfect sequence match against the ORFs used in our interactome, whereas the others are loosely inferred to be close homologs or alternate isoforms. The second interaction (Q7XP52 and Q852Q0) was nearly identical but was originally reported between a related kinase OSK3 rather than OSK4. **c-f,** We repeat this analysis comparing the 322 rice interactions reported here to the binary-high quality interactomes of four additional species (*Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Homo sapiens*, and *Escherichia coli* respectively). Shown are the percentages of rice interactions with homologous interactions reported in the comparison interactome map at various percent identity cutoffs (green lines) and the percentage that could have theoretically been recapitulated—i.e. rice interactions with homologs for both genes in the comparison interactome regardless of if the interaction was reported (blue lines).

re-wiring between the fringe components of the interactomes of rice and *Arabidopsis*. However, we emphasize that at the time our rice interactome map is not large enough to allow any statistically meaningful interpretation of the interactome conservation between species.

To explore the implications of conserved and novel interactions further, we conducted a manual literature search of the top 20 most highly co-expressed interacting genes (SCC $\geq$ 0.8) which yielded evidence supporting the existence of heteromeric protein complexes for the majority of interactions (**Table 8**), providing further confirmation that our method identifies robust protein interactions. Among these, our interactome map shows a physical interaction between the RAD23 DNA repair protein (LOC_Os02g08300) and a component of the 26S proteasome assembly (LOC_Os03g13970). Previous studies in humans and *Arabidopsis* have demonstrated that the ubiquitin receptor RAD23 serves as a link between the nucleotide excision repair and 26S proteasomal degradation pathways[96-98]. We further found that the proteasomal protein LOC_Os03g13970 interacts with a glutaredoxin family protein (LOC_Os04g17050) and a ubiquitin-conjugating enzyme (LOC_Os08g28680). Glutaredoxin proteins, found across bacteria and eukaryotes, have been suggested as candidates for modulating the gate of the 26S proteasomal channel through deglutathionylation of the 20S proteosomal subunit[99,100]. To our knowledge, this interaction has not been reported in plants before. We also detected an interaction between the cytosolic and plastidic versions of fructose-1,6-bisphosphatase responsible for catalyzing the reaction from fructose-1,6-bisphosphate to fructose 6-phosphate during gluconeogenesis and the Calvin cycle in the cytosol and chloroplast respectively.

A previous study in pea (*Pisum sativum*) demonstrated that these proteins co-localize in the nucleus but did not experimentally probe the interaction[101]. Although the functional consequences of this interaction need to be further characterized, these findings suggest this interaction may be conserved between monocots and dicots, highlighting the potential utility of such interactome networks for understanding evolutionary relationships among interacting proteins.

Conclusion

Overall, our work presents a systematic, experimentally-validated advance in the functional annotation of the rice genome. The importance of and effort towards characterization of plant genomes including those of key agricultural species has continued to grow over the years. A recent study has applied a mass spectrometry approach to identify protein complex assemblies broadly conserved throughout the *Viridiplantae* clade to which *O. sativa* belongs[102]. Our annotations alongside others in the literature have critical applications for both basic and translational research that aims to improve the productivity, nutritional value, and climate resilience of this important crop species. Moreover, *O. sativa* is now the first monocot, first agriculturally relevant organism—and indeed the only plant outside of *A. thaliana*—with an ORFeome amenable to high-throughput functional characterization. Thus, our ORFeome and interactome map provide a vital resource to help bridge the ~150 million-year evolutionary gap separating *A. thaliana* from monocot species, including major crop staples such as maize, sorghum, wheat, or barley. We recognize that the work presented herein is limited to interrogating a subset of the ORFeome. The genome of *O. sativa* is

predicted to encode a staggering 30,000 to 50,000 genes[103], dwarfing the number of genes in human and *Arabidopsis*[104,105]. Our reported rice ORFeome currently spans less than 10% of the complete *O. sativa* genome, and as noted above, likely oversamples the most highly expressed genes that are easiest to clone. Moreover, despite matching the recall rate of previous interactomes, and increasing the currently known rice interactome by about 50%, our Y2H screen to date has likely captured less than 1% of the roughly 100,000 protein interactions expected to occur within the proteome as a whole (**Figure 9a**). Nonetheless, our novel PLATE-seq strategy is massively parallel and highly scalable, and thus constitutes a vital tool that will accelerate future high-throughput functional biology studies aimed at filling these gaps.

Methods

PLATE-Seq experimental setup

Plasmid(s) from individual wells of 96-well plates were amplified by PCR using a plasmid-specific forward primer and position-specific reverse primer, consisting of a position-specific barcode and TruSeq 3' sequencing adapter. The reverse primer for Gateway entry clones was comprised of the following: TruSeq 3' adapter (5' GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT 3'), two random bases (5' NN 3'), seven nucleotide long position-specific barcode, and entry-clone specific M13G reverse (5' CAGAGATTTTGAGACAC 3'). The reverse primer for Gateway AD clones was comprised of the following: TruSeq 3' adapter (same as above), three random bases (5' NNN 3'), seven nucleotide long position-specific barcode, barcode to denote the plasmid as an AD-construct (5' CACA 3'), and AD-clone specific reverse

(5’ CAGAGATTTTGAGACAC 3’). The reverse primer for Gateway DB clones was comprised of the following: TruSeq 3’ adapter (same as above), three random bases (5’ NNN 3’), seven nucleotide long position-specific barcode, barcode to denote the plasmid as a DB-construct (5’ CGTC 3’), and DB-clone specific reverse (5’ CAGAGATTTTGAGACAC 3’). All primers can be found in **Table 1**.

Tn5 transposase was purified as described previously[106]. Double-stranded DNA to load into Tn5 enzyme was generated by annealing two oligos: /5Phos/CTGTCTCTTATACACATCT and a plate-specific oligo. Each plate-specific oligo consisted of the following sequences: TruSeq 5’ adapter (5’ TCTTTCCCTACACGACGCTCTTCCGATCT 3’), three random bases (5’ NNN 3’), seven nucleotide long plate-specific barcode, Tn5 mosaic sequence (5’ AGATGTGTATAAGAGACAG 3’). Annealing was performed by heating equimolar ratios of oligos in the presence 50mM NaCl at 95°C for 5 minutes, followed by slow-cooling of the mixture at room temperature. Purified Tn5 enzyme was mixed with the annealed product and kept at room temperature for 30 minutes to allow DNA loading.

Because efficiency of cluster formation and amplification during sequencing is restricted by size, the Tn5 tagmentation was optimized to yield fragments from 300 to 1000 bp appropriate for the Ilumina MiSeq platform. Each tagmentation reaction was performed using pre-loaded Tn5 generated above in HEPES buffer (10mM HEPES-KOH, 5mM $MgCl_2$, 10% v/v DMF) at 55°C for 20 minutes. To stop the reaction, SDS (0.04% final) was then added and incubated at room temperature for 7 minutes. Reaction products were purified by PCR purification columns (Qiagen).

Finally, purified and tagmented products across multiple 96-well plates were

pooled and subjected to low 7-cycle enrichment PCR. Primers used were forward 5' AATGATACGGCGACC ACCGAGATCTACACTCTTTCCCTACACGACGC 3' and reverse 5' CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTG 3'. Next, PCR products were purified using 0.6X AMPure XP beads (Beckman Coulter). Lastly, purified DNA was sequenced paired-end on Illumina MiSeq.

PLATE-seq data analyses

Downstream analysis of PLATE-seq sequencing results was performed in order to determine the identity and positions of all clones sequenced. First, computational deconvolution of the sequencing data was performed to group reads according to their original location. For each paired read, the position- and plate-specific barcodes were identified from R2 and R1 respectively. For pooled sequencing of Y2H interaction candidates, an additional barcode for distinguishing AD-Y clones from DB-X clones was included on R2. After reads were grouped by location, the identifies of any ORFs present in each well were determined through BWA alignment (BWA version 0.7.12-r1039). The reference index was created from a list of predicted *O. sativa* ORF sequences using bwa index [reference]. Alignments were generated using bwa mem -a -t 12 [reference] [query] > [output]. In cases where multiple alignments were reported from one read, either the highest quality alignment was retained, or in cases of a tie, the read count was split equally among all alignments. The final output from the initial deconvolution provided read alignment counts for all ORFs detected per well.

These alignment counts were then processed uniquely depending on the sequencing application. For sequence verification and selection of clones to be included

in the ORFeome, wells were first filtered to eliminate empty wells or wells containing multiple ORFs. Wells containing fewer than 50 reads in total or in which the most prevalent ORF represented fewer than 20% of the aligned reads in the well were removed. Additionally, any wells where the most prevalent ORF was detected in the reverse orientation were removed. The identities of each well were then called based on the most prevalent ORF. In order to retain only the highest quality single-isolates in the final ORFeome, in cases where multiple sequenced clones matched the same ORF, the isolate with the highest quality was retained.

For detection of candidate Y2H interactions in the pooled sequencing setup, criteria were loosened in order to minimize false negatives. ORF alignment counts per-well were obtained as described above and used to define two sets of potentially present ORFs; one for AD-Y ORFs and one DB-X ORFs for. Any ORF that had at least 50 aligned reads in total and represented at least 20% of the well was retained in these sets. To avoid dropping "empty" wells the majority ORF for each AD-Y and DB-X was retained regardless of the total number of aligned reads. All pairwise combinations of these detected AD-Y and DB-X ORFs were reported as putative interactions to be verified independently by follow-up Y2H.

Construction of the O. sativa open reading frame library (ORFeome)

To construct a first draft ORFeome for rice that covered a manageable portion of the *O. sativa* genome, we first used RiceNet[70,71] to prioritize a subset of rice genes to clone. RiceNet leverages a combination of co-expression, domain co-occurrence, protein-protein interaction, genetic interaction, and phylogenetic profile similarity features to report a likelihood that pairs of genes share a functional association. Using a loose

likelihood threshold, we identified 3,269 genes with predicted association with a seed set of 89 genes previously associated with biotic or abiotic stress tolerance[70,72]. The 89 seed genes and predicted associations are reported in **Table 2**.

A single representative ORF was selected for each prioritized gene and the pairs of primers listed in **Table 3** were designed to clone each ORF. To ensure maximum coverage of the prioritized *O. sativa* ORFs, RNA was isolated from a wide range of rice plant parts (e.g. leaf, stem, nodes, roots), at different developmental stages, and at various stress conditions (e.g. light, dark, cold-stress, salt-stress, drought-stress) as described above and detailed in **Table 4**. RNAs were converted to cDNAs and used as templates to append Gateway *attB1* and *attB2* cloning sites to flank the start and stop codons, respectively, using ORF-specific PCR. Amplicons were cloned by Gateway BP reactions into entry vector pDONR223 and transformed into bacterial carrier *DH5α*.

As the BP cloning procedure might inadvertently introduce unwanted PCR artifacts into the entry vectors, we manually picked 2 bacterial transformants per entry clone and verified their identities by PLATE-seq. Only validated clones were retained in the *O. sativa* ORFeome. In cases where duplicate clones were detected, the clone with stronger sequencing evidence was retained.

Plant material, stress treatments, sampling and RNA preparation

We used the rice cultivar Kitaake (*Oryza sativa L. ssp. japonica*) for tissue sampling and RNA preparation. To get the maximum coverage of the transcriptome, tissue samples were collected from different stages of development and in response to biotic and abiotic stress treatments. The developmental tissues including mature leaf, flag leaf, leaf sheath, stem nodes, stem internodes, 0-3 cm panicles, 3-15 cm panicles, mature

panicles before anthesis, developing seeds at 0 days and 15 days after anthesis and, mature seeds were collected from greenhouse-grown plants. Two-week-old seedlings grown separately under light and dark conditions in a growth chamber were also sampled.

For cold stress treatment, two-week-old rice seedlings were exposed to decreasing temperatures from 15°C to 10°C and then 5°C for 24 hours. Leaf tissue was harvested after each treatment. For water deficit stress, two-week-old rice seedlings were gradually subjected to 75, 50 and 25% water deficit stress and samples were collected at each treatment. For salt stress treatment, two-week-old seedlings were subjected to increasing levels of salinity (50 mM, 100 mM and 150 mM) for 24 hours and leaf tissue was collected after every treatment. For submergence stress, three-week-old seedlings in soil containing pots were completely submerged in plastic tanks filled with water and leaf tissue was harvested at 0-, 1- and 6- days post submergence. For pathogen inoculation, plants were grown in pots in a greenhouse for five weeks and then transferred to a growth chamber (14 h daytime period, 28/26°C temperature cycle and 90% humidity, light intensity 100 μmol m−2 s−1). Five to six-week-old plants were inoculated with bacterial suspension (OD600 0.5) of *Xanthomonas Oryzae pv. Oryzae* strain PXO99 (Philippine race 6) using scissors-dip method[107]. The leaf tissues were sampled 0-, 1- and 4-days post inoculation.

All tissue samples were flash frozen in liquid nitrogen and stored at -80°C for RNA extraction. Total RNA was extracted from each tissue sample using TRIzol reagent (Invitrogen, CA), treated with DNase I (Ambion) and purified using Macherey-Nagel Nucleospin RNA II kit (Macherey-Nagel, Duren, Germany) as per manufacturer's

protocol. Purified RNA was quantified using a NanoDrop ND-100 spectrophotometer (Thermo Scientific) and equal quantity of RNA from different developmental stages and stress treatments was pooled in one tube for cDNA synthesis.

Yeast two-hybrid (Y2H) screening to generate the rice interactome

We screened all possible pairs (~2.7 million) of 1,671 O. *sativa* ORFs for interaction by high-throughput yeast two-hybrid (Y2H). Initial interaction screening was performed by testing one DB ORF against mini-pools comprised of 188 AD ORFs at a time. All Y2H positive colonies were collected for sequencing. Pairs of ORFs encoding putative Y2H-positive interactors were identified by PLATE-seq and validated by pair-wise Y2H retest.

Y2H experiments were carried out as previously described by us and other groups[59,67-69,84,108]. In brief, *O. sativa* ORFs in entry vectors pDONR223 were first cloned into pDEST AD and DB destination vectors using Gateway LR reactions to generate N-terminal ORF fusions. We refer to these expression clones as AD-Y and DB-X. All AD-Y and DB-X expression clones were then transformed into Y2H *Saccharomyces cerevisiae* strains *MAT*a Y8800 and *MAT*α Y8930 (genotype: *leu2-3, 112 trp1-901 his3Δ200 ura3-52 gal4Δ gal80Δ GAL2::ADE2 GAL1::HIS3@LYS2 GAL7::lacZ@MET2 cyh2R*), respectively. Next, we screened for autoactivators by individually mating each DB-X strain with a *MAT*a Y8800 strain carrying the empty pDEST AD destination vector. To identify AD autoactivators, we mated each AD-Y strain with a *MAT*α Y8930 strain carrying empty pDEST DB destination vector. After allowing the yeast to mate on yeast extract peptone dextrose (YEPD) (1% yeast extract, 2% bactopeptone, 2% glucose, 0.45mM adenine sulfate) 2% agar plates at 30°C

overnight, yeast were replica plated onto synthetic complete 2% agar plates without leucine and tryptophan (SC+Ade–Leu–Trp+His) and incubated at 30°C overnight to select for diploids with both pDEST AD and DB vector backbones. Finally, diploids were replica plated onto synthetic complete 2% agar plates with 1mM 3-amino-1,2,4-triazole (3AT) and without leucine, tryptophan, and histidine (SC+Ade–Leu–Trp–His+1mM 3AT). Plates were incubated at 30°C for 3-5 days. Any AD-Y or DB-X that grew on SC+Ade–Leu–Trp–His+3AT were scored as autoactivators. We excluded autoactivators from all further screenings.

Thereafter, we performed the first round of testing (called phenotyping I) by mating each unique DB-X with individual mini-pools of 188 unique AD-Y on YEPD 2% agar plates. We selected for diploids by replica plating onto SC+Ade–Leu–Trp+His. To select for positive interactions, we performed the Y2H screening by replica plating the diploids onto SC+Ade–Leu–Trp–His+3AT and incubating at 30°C for 4 days. We used sterile toothpicks to pick and inoculate all positives into liquid SC+Ade–Leu–Trp+His to keep the yeast in the diploid state.

Next, all yeast colonies picked from phenotyping I were individually subjected to another round of Y2H testing called phenotyping II. Here, all picked colonies were spotted directly onto 2% agar plates of SC+Ade–Leu–Trp–His+3AT and SC–Ade–Leu–Trp+His. Plates were incubated at 30°C for 4 days. Positives from this round of screening were picked into liquid SC+Ade–Leu–Trp+His to keep the yeast in the diploid state.

Positive yeast picked from phenotyping II were subject to extraction of plasmid DNA by lysis using zymolyase enzyme (Seigakaku Corporation). Cell and enzyme were

incubated at 37°C for 45 minutes and then at 95°C for 10 minutes. The identities of DB-X and AD-Y were determined by PLATE-seq.

Finally, for every AD-Y and DB-X interaction candidate identified by PLATE-seq, we performed pairwise Y2H testing of each identified pair. This was done by first mating each individual AD-Y with the DB-X putative interaction partner on YEPD plate at 30°C overnight. Then, diploids were selected by replica plating onto SC+Ade–Leu–Trp+His plate and incubating at 30°C overnight. Finally, diploids were selected for interaction-positive cells by replica plating onto SC+Ade–Leu–Trp–His+3AT and SC–Ade–Leu–Trp+His plates and incubating at 30°C for 4-7 days. To identify *de novo* autoactivators (autoactivators that likely arise from accumulation of random mutations during the screening process), we concurrently mated each DB-X with a *MAT*a Y8800 strain carrying the empty pDEST AD destination vector. Afterwards, we followed the same procedure as during the first autoactivator-detection screen. All identified *de novo* autoactivators were removed from the screens. Thus, at the conclusion of the pairwise Y2H phase, we were able to definitively identify and verify all interacting AD-Y and DB-X while controlling for all *de novo* autoactivators.

Bimolecular fluorescence complementation (BiFC)

BiFC assays were performed using onion infiltration. Vectors used were Gateway-compatible constructs pSAT4-DEST-N(1-174)EYFP-C1 (CD3-1089), pSAT4A-DEST-N(1-174)EYFP-N1 (CD3-1080), pSAT5-DEST-C(175-end)EYFP-C1 (CD3-1097) and pSAT5A-DEST-C(175-end)EYFP-N1 (CD3-1096) and were acquired from the *Arabidopsis* Biological Resource Center (ABRC) and described previously[109]. To allow these plasmids to replicate in *Agrobacterium*, the pSa origin-of-replication from

pGREENII-0179 was inserted adjacent to the *E. coli* origin-of-replication.

We selected seven high confidence protein pairs from our yeast two-hybrid screen and cloned each ORF into the four modified pSAT vectors using an LR recombination reaction. After, expression clones were transformed individually into *Agrobacterium* strain GV3101 carrying the pSOUP helper plasmid. Liquid cultures were grown from selected agrobacteria colonies and used to infiltrate onion as described by Xu *et al*.[110]. In brief, the agrobacteria were pelleted by centrifugation and resuspended into the complete resuspension buffer recommended by Xu *et al*. Each culture was then diluted to an OD_{600} of 0.1. For each interaction pair, we then prepared eight mixtures representing all possible interaction orientations by pipetting equal volumes of the appropriate *Agrobacterium* strains into new tubes. Next, we infiltrated approximately 100-200 µL of the *Agrobacterium* mixtures as described by Xu *et al*. We then incubated the samples in the dark at 28ºC for 3-4 days before performing confocal microscopy on epidermal peels taken from the onion samples.

Confocal images were collected on a Leica TCS-SP5 microscope (Leica Microsystems, Exton, PA USA) using a 20X water immersion objective. YFP was excited with the blue argon ion laser (488 nm), and emitted light was collected between 525nm and 595nm. Three images of each sample were taken at random locations on each sample with no changes made to the instrument settings between images or samples except for adjusting the focus. Bright field images were collected simultaneously with the fluorescence images using the transmitted light detector. Images were processed using Leica LAS-AF software (version 3.3.0) and fluorescence was quantified using ImageJ (version 1.51n). Reported values are a ratio of the absolute

fluorescence measured for each sample relative to the positive control. The positive control used for comparison was the AtABI1-AtOST1 strong interaction and the negative control used was AtOST1-AtOIP26 as previously reported[111].

Rice expression analysis

In order to determine whether interacting rice genes exhibited higher co-expression than non-interacting rice genes, we first obtained rice expression data from the Michigan State University Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/expression.shtml)[112]. Only the first 16 expression libraries corresponding to RNA sequencing data derived from tissue at various developmental stages under physiological conditions were used. Gene co-expression was reported as the Spearman correlation coefficient between the expression vectors for two genes. All homo-dimer interactions were removed from our rice interactome for the analysis. An equal number of hetero-dimer pairs were randomly sampled from either the ORFeome or entire *O. sativa* genome for comparison.

Rice GO term semantic similarity analysis

Analysis of the semantic similarities between GO annotations among rice gene pairs was performed using GOssTo[113,114]. GOSlim assignments for ~30,000 *O. sativa* ORFs were obtained using the batch download feature from the Michigan State University Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/downloads_gad.shtml)[112]. The directed acyclic graph (DAG) used to represent the ontology relationship was the core ontology available through The Gene Ontology Resource (http://purl.obolibrary.org/obo/go.obo)[115,116]. In

order to increase coverage across all ontology relationship and thus improve GOssTo performance, the MSU annotations were supplemented with a secondary set of *O. sativa* proteome annotations downloaded through the European Bioinformatics Institute (EMBL-EBI) ftp server (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/2610640.O_sativa_subsp_japonica_Rice.goa)[117]. GOssTo was run with the following command…

java -Xms64G -jar Gossto.jar -calculationdata genewise -calculationtype ism -evidencecodes EXP,IDA,IPI,IMP,IGI,IEP,TAS,IC -goapath [annotations] -hsm Resnik -hsmoutput [out1] -ismoutput [out2] -matrixStyle m -obopath [ontologyDAG] -ontology all -relations part_of,is_a -weightedJaccard True

This command generated pairwise sematic similarity scores between all rice proteins for each the molecular function, biological process, and cellular component GOSlim terms. We retained the final "integrated similarity measure" (ISM) outputs for all further analyses. We compared the distributions of these semantic similarities between our reported rice interactome, and non-interacting protein pairs sampled from our ORFeome. A cutoff of 0.75 was selected to distinguish pairs that were functionally similar from pairs that were not functionally similar.

Conservation analysis

In order to analyze how conserved the captured interactions are across the tree of life we used the OrthoMCL pipeline to identify orthologs of the identified interacting genes from *Sorghum bicolor, Solanum lycopersicum, Arabidopsis thaliana, Physcomitrella patens, Saccharomyces cereviseae, Homo sapiens, and Escherichia coli*[118]. To test

whether interactions are more or less conserved than random expectation, we randomly sampled genes from the rice genome and determined their breadth of conservation across the sampled species. To test for statistical significance, we performed a bootstrap analysis by 1000 sampling replicates and tested if the actual number of captured interacting genes conserved at each node was significantly different from the random expectation using the z-test function in R.

Literature interactome set

In order to compare our novel rice interactions to those previously reported in the literature, we used a curated set of high-quality protein-protein interactions[85] compiled from seven primary interaction databases; BioGRID[86], MINT[87], iRefWeb[88], DIP[89], IntAct[90], HPRD[91,92], MIPS[93], and the PDB[94,95]. The interaction set used was *O. sativa* binary high-quality containing 237 interactions downloaded on May 15, 2019. This interaction set was supplemented with 372 additional interactions from a high-throughput Y2H rice-kinase interactome screen[66] for a total of 609 previously reported interactions.

Literature support for detected interactions

A total of 20 highly co-expressed (SCC ≥ 0.8) interacting gene pairs were manually searched for previous literature evidence supporting the existence of an interaction in other species. As detailed in **Table 7**, in 12 out of 20 interactions searched (60%), previous studied had detected the interaction among homologous genes in other species[96,97,99-101,119-133].

De novo sequence assembly for the human BIRC7 ORF

To explore applications of PLATE-seq beyond clone identification, we used both a custom sequence assembly script and the online version of CAP3[134]. For the custom script a random set of seed reads was selected to begin assembly. These starting contigs were iteratively expanded by aligning all remaining reads to them and incorporating overhanging alignments into the contigs. Intermediate contigs were merged together when sufficient overlap was detected between them. This process was repeated until only one candidate contig remained or until no additional reads could be incorporated into the existing contigs. Final assemblies between our custom script and CAP3 agreed with each other. This method was only applied for demonstrative purposes on the *BIRC7* test case shown in **Figure 5a** and has not be extended to our whole ORFeome.

Identification of variants in human BIRC7 ORF

To explore applications of PLATE-seq beyond clone identification, we applied the sequence analysis scripts from our established CLONE-seq pipeline[54,135] to identify variants relative to the reference sequence. In brief, all reads were aligned to the *BIRC7* reference sequence and all possible mutations were scored based on the ratio of non-reference reads to reference reads at each position, normalized by the sequencing error rate estimated from neighboring positions. This method was only applied for demonstrative purposes to identify the previously reported C882T variant in the *BIRC7* clone shown in **Figure 5d**.

REFERENCES

54 Fragoza, R. *et al.* Extensive disruption of protein interactions by genetic variants across the allele frequency spectrum in human populations. *Nat Commun* 10, 4141, doi:10.1038/s41467-019-11959-3 (2019).

55 Salehi-Ashtiani, K. *et al.* Isoform discovery by targeted cloning, 'deep-well' pooling and parallel sequencing. *Nat Methods* 5, 597-600, doi:10.1038/nmeth.1224 (2008).

56 Yang, X. *et al.* A public genome-scale lentiviral expression library of human ORFs. *Nat Methods* 8, 659-661, doi:10.1038/nmeth.1638 (2011).

57 Sakai, H. *et al.* Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol* 54, e6, doi:10.1093/pcp/pcs183 (2013).

58 Gong, W. *et al.* Genome-wide ORFeome cloning and analysis of Arabidopsis transcription factor genes. *Plant Physiol* 135, 773-782, doi:10.1104/pp.104.042176 (2004).

59 Arabidopsis Interactome Mapping, C. Evidence for network evolution in an Arabidopsis interactome map. *Science* 333, 601-607, doi:10.1126/science.1203877 (2011).

60 Rice Full-Length c, D. N. A. C. *et al.* Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science* 301, 376-379, doi:10.1126/science.1081288 (2003).

61 Reboul, J. *et al.* C. elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat Genet* 34, 35-41, doi:10.1038/ng1140 (2003).

62 Rual, J. F. *et al.* Human ORFeome version 1.1: a platform for reverse proteomics. *Genome Res* 14, 2128-2135, doi:10.1101/gr.2973604 (2004).

63 Gelperin, D. M. *et al.* Biochemical and genetic analysis of the yeast proteome with a movable ORF collection. *Genes Dev* 19, 2816-2826, doi:10.1101/gad.1362105 (2005).

64 Matsuyama, A. *et al.* ORFeome cloning and global analysis of protein localization in the fission yeast Schizosaccharomyces pombe. *Nat Biotechnol* 24, 841-847, doi:10.1038/nbt1222 (2006).

65 Rohila, J. S. *et al.* Protein-protein interactions of tandem affinity purification-tagged protein kinases in rice. *Plant J* 46, 1-13, doi:10.1111/j.1365-313X.2006.02671.x (2006).

66 Ding, X. *et al.* A rice kinase-protein interaction map. *Plant Physiol* 149, 1478-1492, doi:10.1104/pp.108.128298 (2009).

67 Rual, J. F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437, 1173-1178, doi:10.1038/nature04209 (2005).

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

69 Vo, T. V. *et al.* A Proteome-wide Fission Yeast Interactome Reveals Network Evolution Principles from Yeasts to Human. *Cell* 164, 310-323,

doi:10.1016/j.cell.2015.11.037 (2016).
70 Lee, I. *et al.* Genetic dissection of the biotic stress response using a genome-scale gene network for rice. *Proc Natl Acad Sci U S A* 108, 18548-18553, doi:10.1073/pnas.1110384108 (2011).
71 Lee, T. *et al.* RiceNet v2: an improved network prioritization server for rice genes. *Nucleic Acids Res* 43, W122-127, doi:10.1093/nar/gkv253 (2015).
72 Seo, Y. S. *et al.* Towards establishment of a rice stress response interactome. *PLoS Genet* 7, e1002020, doi:10.1371/journal.pgen.1002020 (2011).
73 Krehenwinkel, H. *et al.* Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Sci Rep* 7, 17668, doi:10.1038/s41598-017-17333-x (2017).
74 Vidal, M. A unifying view of 21st century systems biology. *FEBS Lett* 583, 3891-3894, doi:10.1016/j.febslet.2009.11.024 (2009).
75 Robinson, C. V., Sali, A. & Baumeister, W. The molecular sociology of the cell. *Nature* 450, 973-982, doi:10.1038/nature06523 (2007).
76 Barabasi, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 12, 56-68, doi:10.1038/nrg2918 (2011).
77 Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122, 957-968, doi:10.1016/j.cell.2005.08.029 (2005).
78 Vidal, M., Cusick, M. E. & Barabasi, A. L. Interactome networks and human disease. *Cell* 144, 986-998, doi:10.1016/j.cell.2011.02.016 (2011).
79 Snel, B., Lehmann, G., Bork, P. & Huynen, M. A. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* 28, 3442-3444, doi:10.1093/nar/28.18.3442 (2000).
80 Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 47, D607-D613, doi:10.1093/nar/gky1131 (2019).
81 Gu, H., Zhu, P., Jiao, Y., Meng, Y. & Chen, M. PRIN: a predicted rice interactome network. *BMC Bioinformatics* 12, 161, doi:10.1186/1471-2105-12-161 (2011).
82 Rohila, J. S., Chen, M., Cerny, R. & Fromm, M. E. Improved tandem affinity purification tag and methods for isolation of protein heterocomplexes from plants. *Plant J* 38, 172-181, doi:10.1111/j.1365-313X.2004.02031.x (2004).
83 Cooper, B. *et al.* A network of rice genes associated with stress response and seed development. *Proc Natl Acad Sci U S A* 100, 4945-4950, doi:10.1073/pnas.0737574100 (2003).
84 Das, J. *et al.* Cross-species protein interactome mapping reveals species-specific wiring of stress response pathways. *Sci Signal* 6, ra38, doi:10.1126/scisignal.2003350 (2013).
85 Das, J. & Yu, H. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst Biol* 6, 92, doi:10.1186/1752-0509-6-92 (2012).

86 Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 34, D535-539, doi:10.1093/nar/gkj109 (2006).
87 Licata, L. *et al.* MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 40, D857-861, doi:10.1093/nar/gkr930 (2012).
88 Turner, B. *et al.* iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database (Oxford)* 2010, baq023, doi:10.1093/database/baq023 (2010).
89 Xenarios, I. *et al.* DIP: the database of interacting proteins. *Nucleic Acids Res* 28, 289-291, doi:10.1093/nar/28.1.289 (2000).
90 Orchard, S. *et al.* The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* 42, D358-363, doi:10.1093/nar/gkt1115 (2014).
91 Peri, S. *et al.* Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* 13, 2363-2371, doi:10.1101/gr.1680803 (2003).
92 Keshava Prasad, T. S. *et al.* Human Protein Reference Database--2009 update. *Nucleic Acids Res* 37, D767-772, doi:10.1093/nar/gkn892 (2009).
93 Pagel, P. *et al.* The MIPS mammalian protein-protein interaction database. *Bioinformatics* 21, 832-834, doi:10.1093/bioinformatics/bti115 (2005).
94 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* 28, 235-242, doi:10.1093/nar/28.1.235 (2000).
95 Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* 10, 980, doi:10.1038/nsb1203-980 (2003).
96 Schauber, C. *et al.* Rad23 links DNA repair to the ubiquitin/proteasome pathway. *Nature* 391, 715-718, doi:10.1038/35661 (1998).
97 Dantuma, N. P., Heinen, C. & Hoogstraten, D. The ubiquitin receptor Rad23: at the crossroads of nucleotide excision repair and proteasomal degradation. *DNA Repair (Amst)* 8, 449-460, doi:10.1016/j.dnarep.2009.01.005 (2009).
98 Farmer, L. M. *et al.* The RAD23 family provides an essential connection between the 26S proteasome and ubiquitylated proteins in Arabidopsis. *Plant Cell* 22, 124-142, doi:10.1105/tpc.109.072660 (2010).
99 Silva, G. M. *et al.* Role of glutaredoxin 2 and cytosolic thioredoxins in cysteinyl-based redox modification of the 20S proteasome. *FEBS J* 275, 2942-2955, doi:10.1111/j.1742-4658.2008.06441.x (2008).
100 Demasi, M. *et al.* Redox regulation of the proteasome via S-glutathionylation. *Redox Biol* 2, 44-51, doi:10.1016/j.redox.2013.12.003 (2013).
101 Anderson, L. E., Yousefzai, R., Ringenberg, M. R. & Carol, A. A. Both chloroplastic and cytosolic fructose bisphosphatase isozymes are present in the pea leaf nucleus. *Plant Science* 166, 721-730, doi:10.1016/j.plantsci.2003.11.008 (2004).
102 McWhite, C. D. *et al.* A pan-plant protein complex map reveals deep conservation and novel assemblies. *Cell*, doi:10.1101/815837 ((In Press)).
103 Goff, S. A. *et al.* A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* 296, 92-100, doi:10.1126/science.1068275 (2002).
104 Venter, J. C. *et al.* The sequence of the human genome. *Science* 291, 1304-1351,

doi:10.1126/science.1058040 (2001).
105 Arabidopsis Genome, I. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408, 796-815, doi:10.1038/35048692 (2000).
106 Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res* 24, 2033-2040, doi:10.1101/gr.177881.114 (2014).
107 Kauffman, H. E., Reddy, A.P.K., Hsieh, S.P.Y. and Merca, S.D. An improved technique for evaluating resistance of rice varieties to Xanthomonas oryzae. *Plant Disease Reporter* 57, 537-541 (1973).
108 Li, S. *et al.* A map of the interactome network of the metazoan C. elegans. *Science* 303, 540-543, doi:10.1126/science.1091403 (2004).
109 Williams, B., Kabbage, M., Britt, R. & Dickman, M. B. AtBAG7, an Arabidopsis Bcl-2-associated athanogene, resides in the endoplasmic reticulum and is involved in the unfolded protein response. *Proc Natl Acad Sci U S A* 107, 6088-6093, doi:10.1073/pnas.0912670107 (2010).
110 Xu, K. *et al.* A rapid, highly efficient and economical method of Agrobacterium-mediated in planta transient transformation in living onion epidermis. *PLoS One* 9, e83556, doi:10.1371/journal.pone.0083556 (2014).
111 Waadt, R. *et al.* Identification of Open Stomata1-Interacting Proteins Reveals Interactions with Sucrose Non-fermenting1-Related Protein Kinases2 and with Type 2A Protein Phosphatases That Function in Abscisic Acid Responses. *Plant Physiol* 169, 760-779, doi:10.1104/pp.15.00575 (2015).
112 Kawahara, Y. *et al.* Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice (N Y)* 6, 4, doi:10.1186/1939-8433-6-4 (2013).
113 Caniza, H. *et al.* GOssTo: a stand-alone application and a web tool for calculating semantic similarities on the Gene Ontology. *Bioinformatics* 30, 2235-2236, doi:10.1093/bioinformatics/btu144 (2014).
114 Yang, H., Nepusz, T. & Paccanaro, A. Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. *Bioinformatics* 28, 1383-1389, doi:10.1093/bioinformatics/bts129 (2012).
115 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25-29, doi:10.1038/75556 (2000).
116 The Gene Ontology, C. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* 47, D330-D338, doi:10.1093/nar/gky1055 (2019).
117 Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* 47, W636-W641, doi:10.1093/nar/gkz268 (2019).
118 Li, L., Stoeckert, C. J., Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13, 2178-2189, doi:10.1101/gr.1224503 (2003).
119 Sanders, J., Brandsma, M., Janssen, G. M. C., Dijk, J. & Moller, W. Immunofluorescence studies of human fibroblasts demonstrate the presence of the complex of elongation factor-1 beta gamma delta in the endoplasmic reticulum. *Journal of Cell Science* 109, 1113-1117 (1996).

120 Kim, K. K., Kim, R. & Kim, S. H. Crystal structure of a small heat-shock protein. *Nature* 394, 595-599, doi:10.1038/29106 (1998).
121 Tanabe, N., Yoshimura, K., Kimura, A., Yabuta, Y. & Shigeoka, S. Differential expression of alternatively spliced mRNAs of Arabidopsis SR protein homologs, atSR30 and atSR45a, in response to environmental stress. *Plant Cell Physiol* 48, 1036-1049, doi:10.1093/pcp/pcm069 (2007).
122 Golovkin, M. & Reddy, A. S. An SC35-like protein and a novel serine/arginine-rich protein interact with Arabidopsis U1-70K protein. *J Biol Chem* 274, 36428-36438, doi:10.1074/jbc.274.51.36428 (1999).
123 Riegler, H. *et al.* Crystal structure and functional characterization of a glucosamine-6-phosphate N-acetyltransferase from Arabidopsis thaliana. *Biochem J* 443, 427-437, doi:10.1042/BJ20112071 (2012).
124 Lee, S. J. & Baserga, S. J. Imp3p and Imp4p, two specific components of the U3 small nucleolar ribonucleoprotein that are essential for pre-18S rRNA processing. *Mol Cell Biol* 19, 5441-5452, doi:10.1128/mcb.19.8.5441 (1999).
125 Sa-Moura, B. *et al.* Mpp10 represents a platform for the interaction of multiple factors within the 90S pre-ribosome. *PLoS One* 12, e0183272, doi:10.1371/journal.pone.0183272 (2017).
126 Huang, X. Y. *et al.* CYCLIN-DEPENDENT KINASE G1 is associated with the spliceosome to regulate CALLOSE SYNTHASE5 splicing and pollen wall formation in Arabidopsis. *Plant Cell* 25, 637-648, doi:10.1105/tpc.112.107896 (2013).
127 Lorkovic, Z. J., Lehner, R., Forstner, C. & Barta, A. Evolutionary conservation of minor U12-type spliceosome between plants and humans. *RNA* 11, 1095-1107, doi:10.1261/rna.2440305 (2005).
128 Beatrix, B., Sakai, H. & Wiedmann, M. The alpha and beta subunit of the nascent polypeptide-associated complex have distinct functions. *J Biol Chem* 275, 37838-37845, doi:10.1074/jbc.M006368200 (2000).
129 Couvreur, B. *et al.* Eubacterial HslV and HslU subunits homologs in primordial eukaryotes. *Mol Biol Evol* 19, 2110-2117, doi:10.1093/oxfordjournals.molbev.a004036 (2002).
130 Song, H. K. *et al.* Isolation and characterization of the prokaryotic proteasome homolog HslVU (ClpQY) from Thermotoga maritima and the crystal structure of HslV. *Biophysical Chemistry* 100, 437-452, doi:10.1016/s0301-4622(02)00297-1 (2002).
131 Kwon, A.-R., Kessler, B. M., Overkleeft, H. S. & McKay, D. B. Structure and Reactivity of an Asymmetric Complex between HslV and I-domain Deleted HslU, a Prokaryotic Homolog of the Eukaryotic Proteasome. *Journal of Molecular Biology* 330, 185-195, doi:10.1016/s0022-2836(03)00580-1 (2003).
132 Dantuma, N. P. & Lindsten, K. Stressing the ubiquitin-proteasome system. *Cardiovasc Res* 85, 263-271, doi:10.1093/cvr/cvp255 (2010).
133 Dantuma, N. P. & Bott, L. C. The ubiquitin-proteasome system in neurodegenerative diseases: precipitating factor, yet part of the solution. *Front Mol Neurosci* 7, 70, doi:10.3389/fnmol.2014.00070 (2014).
134 Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome*

Res 9, 868-877, doi:10.1101/gr.9.9.868 (1999).
135 Wei, X. *et al.* A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. *PLoS Genet* 10, e1004819, doi:10.1371/journal.pgen.1004819 (2014).

CHAPTER 3:
EXTENSIVE DISRUPTION OF PROTEIN INTERACTIONS BY GENETIC VARIANTS ACROSS THE ALLELE FREQUENCY SPECTRUM IN HUMAN POPULATIONS

Context and Personal Contributions

The following chapter is derived from a Nature Communications article by the same name[54] originally authored by Robert Fragoza and Jishnu Das (Reproduced with permission from Springer Nature). Full authorship and contributions are provided in the original publication, but the following warrant explicit mentioning. This text was primarily written by Robert Fragoza who is also responsible for leading this paper and completing most of the experiments related to its completion. Most of the computational analyses and figures were generated by Jishnu Das or Siqi (Charles) Liang.

My contributions to this paper were in ensuring high sequence fidelity of the clones for all missense single nucleotide variants (SNVs) profiled in this study. I was responsible for developing an updated next-generation sequencing pipeline to confirm the success of designed mutagenesis experiments and prune clones that introduced inadvertent off target mutations or insertions. The previous sequencing readout for our massively parallel mutagenesis pipeline is described within Wei *et al*. 2014[24], and had been discovered to miss certain off target mutations: most particularly, failure to detect rare duplications of the mutagenesis primer sequence during cloning. My new implementation of this "Clone-seq" pipeline was used to retroactively reassess all mutant clones that had been generated in the Yu lab to date, and to complete the initial screen on the last mutagenesis batches for this and future projects. I authored the methods section titled "Identifying successfully mutated clones" which details for this new sequencing readout in addition to the section titled "Defining duplicate genes and

functionally similar proteins" which describes one of the supplemental analyses to investigate whether disruptive variants could be compensated for by homologous genes (see **Figure 19**). Beyond contributions explicitly described in this paper, my updates to the "Clone-seq" pipeline included curation of all of the mutagenesis attempts, successful mutant clone library, and subsequent experimental profiling results into a comprehensive relational database to ensure this data remains organized and accessible.

Abstract

Each human genome carries tens of thousands of coding variants. The extent to which this variation is functional and the mechanisms by which they exert their influence remains largely unexplored. To address this gap, we leveraged the ExAC database of 60,706 human exomes to investigate experimentally the impact of 2,009 missense single nucleotide variants (SNVs) across 2,185 protein-protein interactions, generating interaction profiles for 4,797 SNV-interaction pairs, of which 421 of these SNVs are segregating at > 1% allele frequency in the human population. Surprisingly, we find that interaction-disruptive SNVs are prevalent across both rare and common allele frequencies. Furthermore, these results suggest that 10.5% of missense variants carried per individual are disruptive, a much higher proportion than previously reported; this indicates that each individual's genetic makeup may be significantly more complex than expected. Notably, disruptive variants also occur at elevated proportions in disease-associated genes and are enriched at conserved genomic loci, signifying their potential phenotypic relevance. Finally, we demonstrate that candidate disease-associated mutations can be identified through shared interaction perturbations between variants of interest and known disease mutations. Overall, our interactome perturbation study serves as an important framework for providing mechanistic insights and contextual

information to interpret the impact of coding variation on protein function genome-wide, which is crucial for dissecting complex genotype-to-phenotype relationships.

Introduction

Recent explosive population growth has generated an excess of rare genetic variation segregating in human populations that likely plays a key role in the individual genetic burden of complex disease risk[32,136-140]. In agreement with this paradigm, large-scale whole-genome and whole-exome sequencing efforts have reported an excess of genetic variation in human genomes segregating at very low allele frequencies[3,32,138,140-142]. In particular, rare coding single nucleotide variants (SNVs) have been predicted to disproportionately impact protein function[3,32,143] in human genomes; however, methods and metrics for estimating the functionality of coding SNVs vary widely, and there is no consensus estimate for the number of functional variants per individual[144]. As such, a direct assessment of the functional impact of coding SNVs could prove indispensable to furthering our understanding on how segregating genetic variation influences complex traits and human disease.

Biological processes are likely regulated through intricate networks of protein and macromolecular interactions, as opposed to single proteins acting independently[145,146]. Researchers have accordingly identified a large number of mutations that disrupt these interactions; however, most of these perturbations correspond to synthetic mutations from scanning mutagenesis experiments[147-149], the vast majority of which do not occur naturally in human populations. For example, the SKEMPI database comprehensively collected the impact of 3,047 mutations on protein binding events published in the literature[150], only seven of which are listed as human population variants in ExAC[142]. Efforts to examine the impact of disease-associated mutations on protein function[23,24,27] are also limited because most of these mutations are very rare and consequently only impact a small number of individuals. The evolutionary context in which all genomic variants evolve is largely missing from such studies as a result.

In order to acquire a more representative understanding of the functional impact of human population variants on protein function, we leveraged the ExAC dataset of coding variants from 60,706 human exomes[142] to systematically evaluate the impact of 2,009 missense SNVs, 811 of which are segregating at minor allele frequency (MAF) > 0.1% in the human population, across 2,185 protein-protein interactions. We find that disruptive SNVs are strongly enriched at conserved protein loci and occur more prevalently at lower allele frequencies, underscoring the functional importance of disruptive variants uncovered by our assays. Moreover, we also determined that on average 10.5% of coding SNVs carried per individual are expected to impact protein-protein interactions, a rate much higher than indicated by previous reports[3,32,143]. Unexpectedly, while we observe an enrichment of functional SNVs at rare allele frequencies in agreement with previous literature[3,32,138,143], we also find that 9.6% of tested common variants with MAF > 10% perturbed protein interactions, indicating that many common variants are also functional[151,152].

Cellular and organism-level phenotypes stem from macromolecular perturbations[23,146]. Furthermore, the genetic background of an individual and its influence on complex traits and disease is determined by the cumulative impact of functional variation, including disruptive SNVs[153] (**Figure 13a**). Hence, experimental measurements of the molecular-level impacts of each disruptive SNV is the imperative first step toward advancing our mechanistic understanding of the genetic background of each individual and their differences across the population.

Results

Disruptive SNV's occur extensively across broad MAF ranges

Alterations to protein-protein interactions can have deleterious consequences to fitness[154], particularly in human genetic disease[23,25,35]. As such, coding variation at interaction interfaces is mostly rare[155] and subject to evolutionary constraint[156,157]. In contrast, common variation is expected to be largely neutral and therefore unlikely to be extensively functional[158-160]. Nonetheless, notable exceptions exist, including APOE-

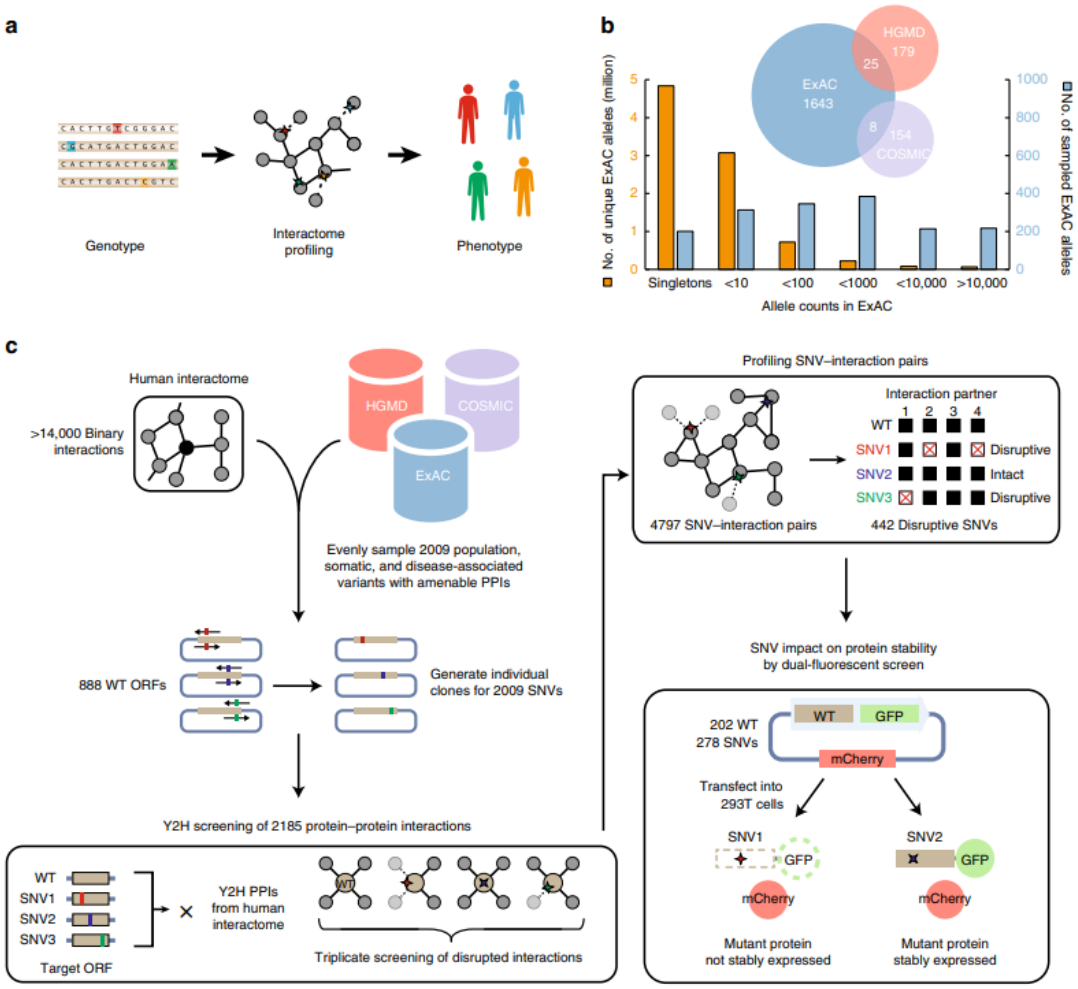


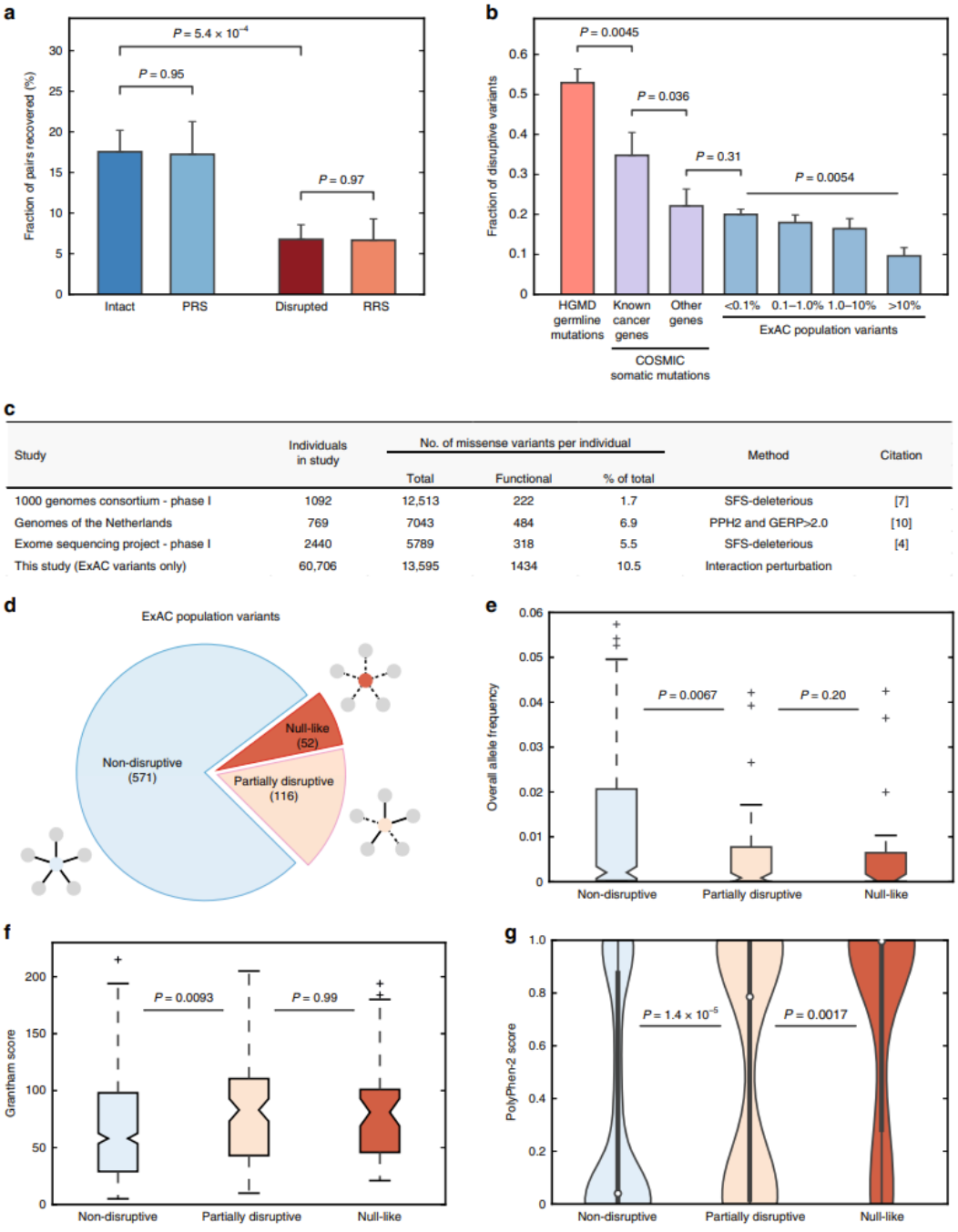
Figure 13. A pipeline for surveying the impact of 2,009 SNVs on protein-protein interactions.
a, Phenotypic consequences of coding variants in human genotypes can be interpreted as products of protein-protein interaction perturbations in the interactome. **b,** Over half of all unique missense variants in ExAC are singletons. To avoid oversampling very rare variants from ExAC, 1,676 ExAC variants were selected across a wide range of allele frequencies. 204 disease-associated mutations listed in HGMD and 162 cancer somatic mutations from COSMIC were also examined. **c,** Pipeline for testing the functional impact of 2,009 SNVs on protein interactions and stability impact of 278 population variants by dual-fluorescence screen.

epsilon 4, a risk-associated allele for Alzheimer's disease[161-163] (MAF = 18.4%), and the P12A polymorphism (MAF = 11.0%) of PPARG, which increases risk for type 2 diabetes[164,165]. Indeed, the extent to which MAF indicates whether an allele is disruptive to protein interactions remains largely unexplored. Hence, to systematically identify functionally relevant SNVs across rare to common allele frequencies, we constructed a

resource of sequence-verified single-colony clones for 2,009 SNVs derived from three major databases: 1,676 variants from ExAC[142], 204 Mendelian disease-associated mutations from HGMD[166], and 162 somatic mutations in cancer from COSMIC[167]. To avoid oversampling rare variants which dominate ExAC, we randomly selected alleles in ExAC across defined MAF bins ranging from singletons to very common alleles (**Figure 13b**; Methods).

Upon constructing this resource, we then performed yeast two-hybrid (Y2H) experiments to measure the impact of these 2,009 missense SNVs across 2,185 human protein-protein interactions. In this manner, we identified 442 interaction-disrupting SNVs, including 298 disruptive ExAC variants, comprising a network of 4,797 SNV-interaction pairs. We further validated the quality of our SNV-interaction network by performing Protein Complementation Assay (PCA)[168] in human 293T cells to retest a representative subset of ~400 disrupted and non-disrupted SNV-interactions pairs from our ExAC subset. SNV-disrupted interactions retested at a rate approximate to a negative reference set comprising randomly selected ORF pairs whereas non-disrupted interactions retested at a rate statistically indistinguishable from a positive reference set of literature-established protein interactions[169,170] (**Figure 14a**, **Figure 15a**). Our result remained unchanged when we removed interactions corresponding to highly-disruptive SNVs (**Figure 15b**). Taken together, our PCA retest demonstrated the reproducibility and validated the quality of our Y2H-generated SNV-interaction network.

To examine the influence of allele frequency on disruptive variants, we partitioned our tested ExAC variants across four allele frequency bins, ranging from very rare (MAF $\leq$ 0.1%) to very common (MAF > 10%) alleles and then calculated the fraction of variants that disrupted one or more protein interactions per MAF bin. We found that the fraction of disruptive variants decreased inversely with increasing allele frequency (P = 0.0054 by chi-square test, **Figure 14b**), which agrees with



c

Study	Individuals in study	No. of missense variants per individual			Method	Citation
		Total	Functional	% of total		
1000 genomes consortium - phase I	1092	12,513	222	1.7	SFS-deleterious	[7]
Genomes of the Netherlands	769	7043	484	6.9	PPH2 and GERP>2.0	[10]
Exome sequencing project - phase I	2440	5789	318	5.5	SFS-deleterious	[4]
This study (ExAC variants only)	60,706	13,595	1434	10.5	Interaction perturbation	

Figure 14. The probability of observing a disruptive allele is inversely related to the allele frequency.

a, Fraction of protein pairs recovered by PCA for disrupted and intact interactions in comparison to positive and random reference sets (PRS and RRS). *P* values by one-tailed *Z*-test between disrupted and intact interactions. *P* values by two-tailed *Z*-test for all other comparisons. **b,** Fraction of disruptive variants in ExAC (blue) across four allele frequency ranges (i) < 0.1%, (ii) 0.1 – 1.0%, (iii) 1.0 – 10%, and (iv) > 100%. *P* value by chi-square test. Fraction of disruptive

somatic mutations in COSMIC (purple) in known cancer-affiliated genes or other genes and fraction of disruptive germline disease-associated genes from HGMD (red) are also shown. *P* values by one-tailed *Z*-test. **c,** Reported number of functional missense variants per individual genome varies extensively across different studies. **d,** ExAC variants tested against ≥ 2 interactions further partitioned into three disruption categories. Distribution of **e,** allele frequency, **f,** Grantham scores, and **g,** PolyPhen-2 scores across three disruption categories. Error bars in **a,** and **b,** indicate +SE of proportion. Thick black bars in **g,** are the interquartile range, white dots display the median, and extended thin black lines represent 95% confidence intervals. *P* values in **e,** and **g,** by one-tailed *U*-test. *P* values in **f,** by two-tailed *U*-test. See also **Tables 9-14** and **Figure 15**.

expectations[3,32,136]; however, we note that 9.6% of very common variants (MAF > 10%) were still disruptive. Considering that the majority of SNVs found in an individual genome are common[171], this elevated proportion may indicate that disruptive coding variation is markedly widespread across populations. To investigate this more closely, we weighted these MAF-stratified disruption rates by their expected proportions within a typical human genome using the site frequency spectrum for missense variants in ExAC (Methods). In this manner, we determined that given an average of 13,595 missense variants per genome, 1,434 (10.5% ± 1.8%) are expected to disrupt protein interactions, a figure notably higher than indicated by previous estimates (**Figure 14c**, **Tables 9-11**). We note, however, that the extent to which interaction disruptions result in cellular phenotypes, particularly for common variants, remains undetermined. Regardless, our results demonstrate that many variants show some degree of functionality, at least within the context of our interaction assays; as such, the genetic background in each individual genome might be far more complex than expected.

To add further context to our disruption rate analysis, we also determined the fraction of cancer-associated somatic mutations that disrupt interactions and found that 34.8% of somatic mutations located in genes with established roles in cancer progression were disruptive (**Figure 14b**; Methods). Notably, this fraction decreased significantly to 22.1% for somatic mutations located in all other genes ($P = 0.036$ by one-tailed *Z*-test), a figure comparable to the 20.0% disruption rate observed for very

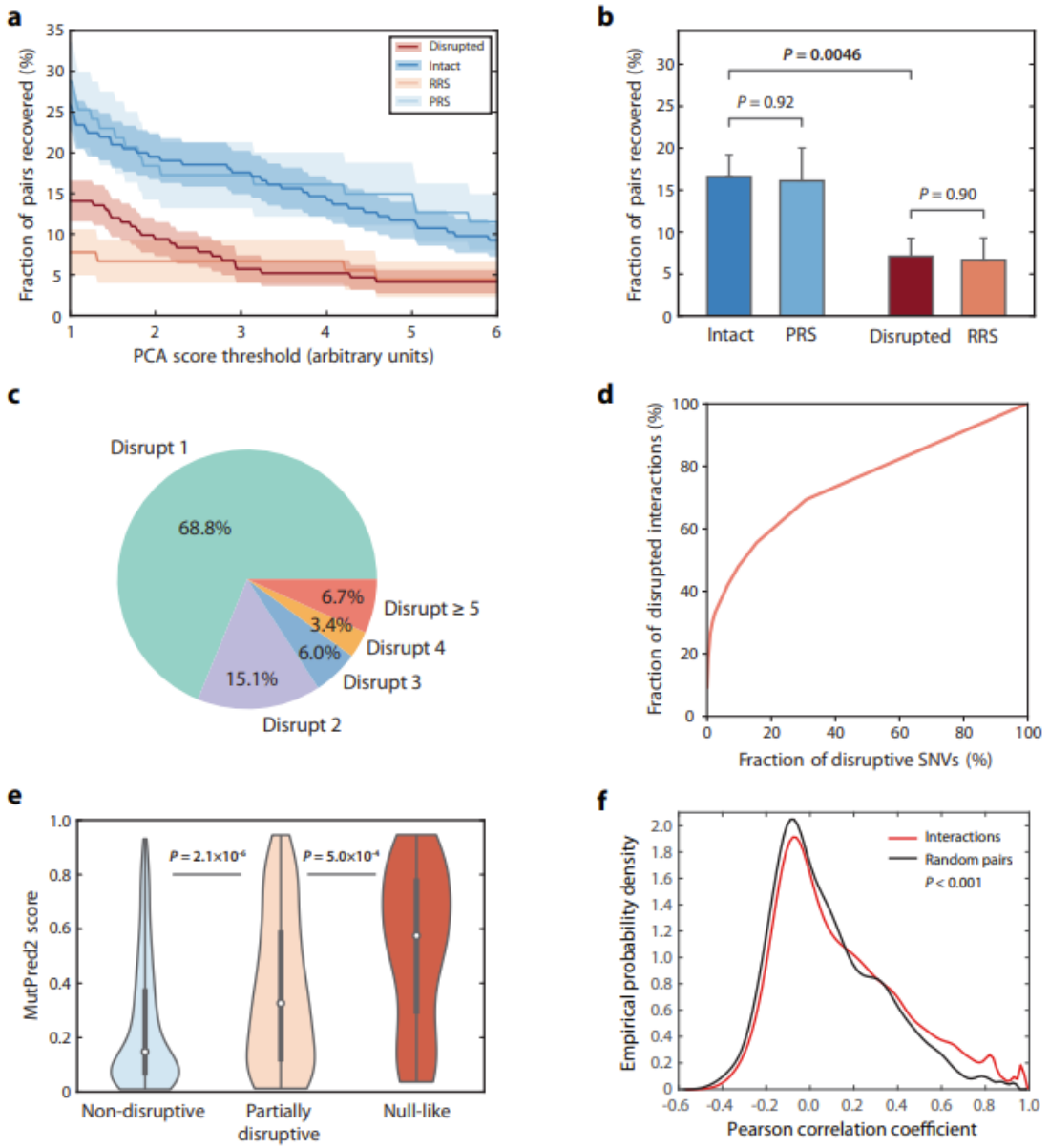


Figure 15. Distribution and reproducibility of disrupted and non-disrupted SNV-interaction pairs.

a, Fraction of protein pairs recovered by PCA across increasingly stringent PCA scoring thresholds. SE of proportion is demarcated by shading. **b,** Fraction of protein pairs recovered by PCA for disrupted and intact interactions in comparison to positive and random reference sets (PRS and RRS). Interactions corresponding to SNVs found on overrepresented bait proteins (bait has >20 interaction partners) were removed. P values by one-tailed Z-test between disrupted and intact interactions. P values by two-tailed Z-test for all other comparisons. **c,** Fraction of disruptive variants (n = 298) categorized by number of disrupted interaction partners. **d,** Cumulative distribution function plotting the fraction of disruptive variants against the total fraction of interactions perturbed. **e,** Distribution of MutPred2 scores across three disruption categories. Thick black bars are the interquartile range, white dots display the median, and extended thin black lines represent 95% confidence intervals. P values by one-tailed U-test.

f, Co-expression of protein abundance levels for protein interaction pairs used in this study. Interacting protein pairs were significantly more likely to be co-expressed than random protein pairs in tissue and cell data from the Human Proteome Map. P value by two-sided KS test.

rare (MAF ≤ 0.1%) ExAC alleles. In contrast, 52.9% of tested HGMD disease-associated mutations were disruptive (**Figure 14b**). Collectively, these trends in disruption rate suggest that driver mutations in oncogenesis may often function by perturbing interactions, as is the case for disease-associated mutations. Therefore, prioritizing disruptive somatic mutations through our interaction perturbation approach may be an effective means to identify potential driver genes and mutations.

The extent to which a mutation is disruptive can also be categorized by measuring the fraction of corresponding protein interactions disrupted by a particular variant. Accordingly, we first grouped each of our 298 disruptive variants by the number of interactions they perturb (**Figure 15c**). We observed that 205 of our tested SNVs disrupted only a single interaction (68.8%) while a small fraction of variants (6.7%) disrupted five or more interactions, suggesting that disruptive mutations tend to perturb specific subsets of protein function as opposed to perturbing protein function as a whole. Examining the distribution of disruptive variants across the number of interactions perturbed revealed a similar trend (**Figure 15d**).

Next, for proteins tested against multiple interaction partners, ExAC variants that leave all interactions intact were categorized as non-disruptive, variants that disrupt a subset of interaction partners were categorized as partially disruptive, and variants that disrupt all tested protein interactions were categorized as null-like (**Figure 14d**). Across these three categories, the median allele frequency for tested variants in ExAC decreased significantly from 0.21% for non-disruptive variants to 0.085% for partially disruptive variants ($P = 0.0067$ by one-tailed U-test) then nominally to 0.034% for null-like variants (**Figure 14e**), suggesting that partially disruptive and null-like variants are potentially deleterious. Furthermore, we also find that Grantham scores, a biochemical

measure quantifying the dissimilarity between amino acid residues[172], for partially disruptive and null-like variants are significantly higher in comparison to non-disruptive variants (**Figure 14f**). Moreover, conservation-based functional prediction algorithms, including PolyPhen-2[159] and MutPred2[30], show significant increases in the likelihood that a variant is deleterious across non-disruptive, partially disruptive, and null-like disruption categories (**Figure 14g**, **Figure 15e**). Taken together, these results show that disruptive variants follow expected patterns of selective constraint and conservation that are characteristic of damaging mutations and imply that these disruptive variants may be functionally relevant in cells.

Coding variants seldom result in unstable protein expression

Mutations can disrupt interactions through local perturbations to specific interaction interfaces or by destabilizing protein folding as a whole[25]. To distinguish between these two distinct mechanisms, we developed a dual fluorescence screening assay to survey the impact of interaction-disruptive variants on protein folding. To set up our dual fluorescent screen, we cloned a subset of wild-type ORFs that are stably expressed when tagged with GFP, as well as their corresponding ExAC variants, into a custom GFP-tag expression vector that co-expresses an untagged mCherry control (Methods). We then transfected wild-type and mutant ORFs into 293T cells to test for mutation-induced changes to protein expression in 96-well plate formats (**Figure 13c**). GFP expression levels for transfected wild-type and mutant samples, normalized with respect to mCherry expression levels, were then used to calculate stability scores for all wild-type and mutant proteins (**Figure 16a**). In this manner, we determined the impact of 278 ExAC variants on protein folding from which we grouped these variants across stable, moderately stable, and unstable protein expression categories (Methods). We note that our stable, moderately stable, and unstable demarcations corresponded well with

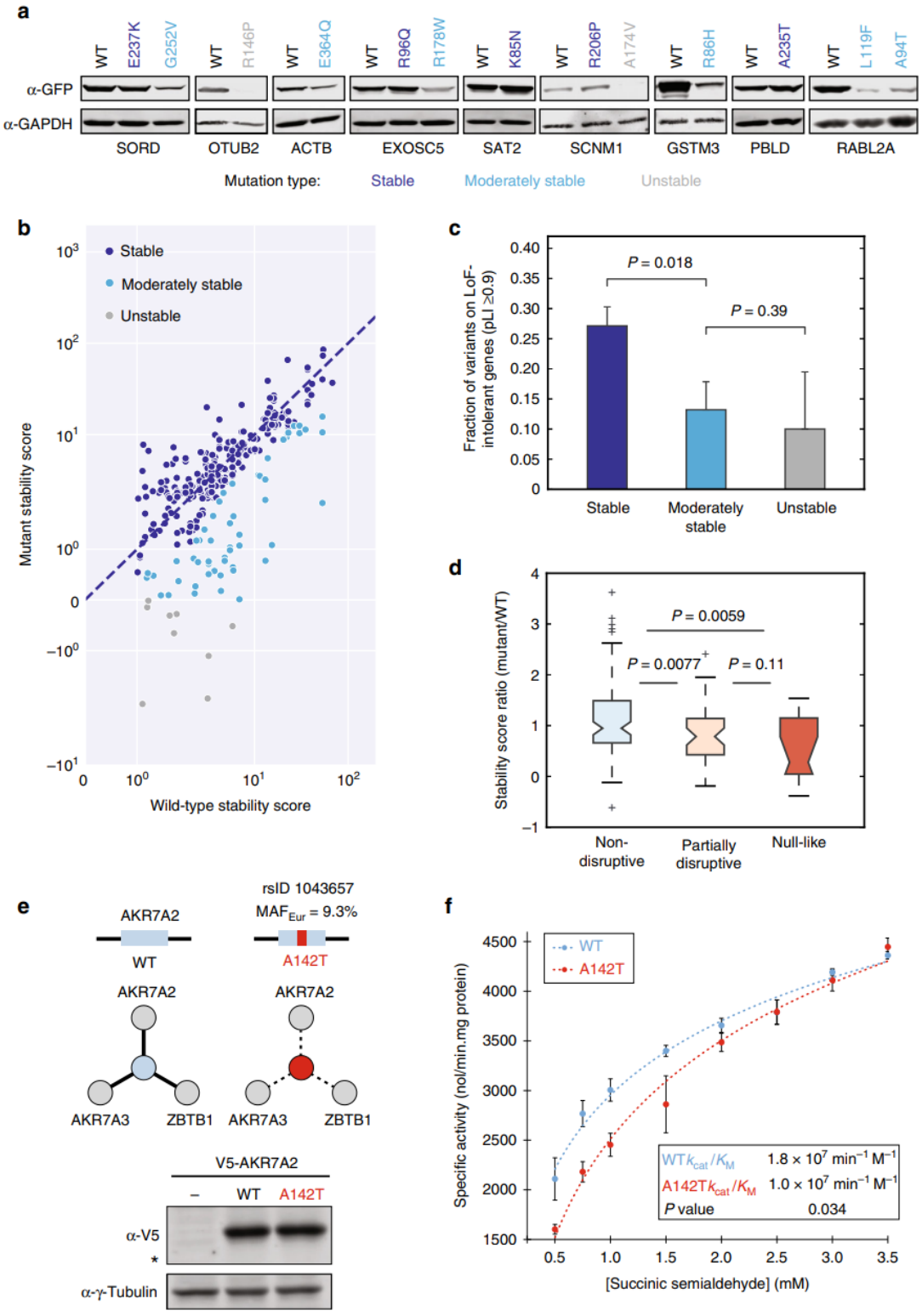

a
WT
E237K
G252V
WT
R146P
WT
E364Q
WT
R96Q
R178W
WT
K85N
WT
R206P
A174V
WT
R86H
WT
A235T
WT
L119F
A94T
α-GFP
α-GAPDH
SORD
OTUB2
ACTB
EXOSC5
SAT2
SCNM1
GSTM3
PBLD
RABL2A
Mutation type:
Stable
Moderately stable
Unstable
b
Stable
Moderately stable
Unstable
Mutant stability score
Wild-type stability score
c
P = 0.018
P = 0.39
Fraction of variants on LoF-intolerant genes (pLI ≥0.9)
Stable
Moderately stable
Unstable
d
P = 0.0059
P = 0.0077
P = 0.11
Stability score ratio (mutant/WT)
Non-disruptive
Partially disruptive
Null-like
e
AKR7A2
WT
rsID 1043657
MAF Eur = 9.3%
A142T
AKR7A2
AKR7A2
AKR7A3
ZBTB1
AKR7A3
ZBTB1
V5-AKR7A2
–
WT
A142T
α-V5
*
α-γ-Tubulin
f
WT
A142T
Specific activity (nol/min.mg protein)
WT kcat/KM
1.8 × 10^7 min^−1 M^−1
A142T kcat/KM
1.0 × 10^7 min^−1 M^−1
P value
0.034
[Succinic semialdehyde] (mM)

Figure 16. Disruptive population variants seldom result in unstable protein expression. **a,** DUAL-FLOU protein stability scores for 278 wild-type:variant pairs. **b,** Western blots for representative wild-type:variant pairs across three stability categories detected using α-GFP. α-GAPDH was used as a loading control. **c,** Fraction of variants residing in LoF-intolerant genes (pLI ≥ 0.9) for stable (*n* = 199), moderately stable (*n* = 53), and unstable (*n* = 10) protein stability categories. **d,** Ratio of mutant-to-wild-type stability score corresponding to non-disruptive (*n* = 103), partially disruptive (*n* = 45), and null-like variants (*n* = 12). **e,** Distribution of interaction-disruptive ExAC variants across three stability categories. **f,** Diagram of interactions disrupted by null-like AKR7A2_A142T variant. Cellular expression levels of V5-tagged AKR7A2 was measured by Western blot using α-V5. α-γ-Tubulin was used as a loading control. **g,** *In vitro* specific activities of purified recombinant AKR7A2 wild-type and A142T using succinic semialdehyde substrate. Fitted curves (dashed lines) are shown for wild-type and A142T. *P* value by one-tailed *t*-test. Error bars indicate ±SE of mean at eight different substrate concentrations. Error bars in **c,** and **d,** indicate +SE of proportion. *P* values in **c,** and **d,** by one-tailed *U*-test. See also **Figure 17**.

western blot intensity (**Figure 16b**).

Mutations that destabilize protein folding should abolish the function of the harboring protein and may likely be depleted within genes that are sensitive to loss-of-function (LoF) mutations as a result. Accordingly, we examined the fraction of variants that occur on genes with pLI ≥ 0.9, a threshold used to define genes that are intolerant to LoF mutations[142], and found that the fraction of variants in LoF-intolerant genes decreased significantly from 27.1% to 13.2% for stable and moderately stable variants, respectively ($P = 0.018$ by one-tailed *U*-test; **Figure 16c**). Protein-destabilizing variants also tend to be rare; we observed that median allele frequency decreased from 0.064% for stable protein variants to 0.021% for moderately stable and unstable variants combined ($P = 0.019$ by one-tailed *U*-test; **Figure 17a**), implying that the destabilized variants uncovered by our protein stability assay are functionally consequential and selectively constrained as a result.

We next investigated the correspondence between protein stability and interaction-disruptive phenotypes by comparing the distribution of stability scores across tested variants from non-disruptive, partially disruptive, and null-like categories. We found that the ratio of mutant-to-wild-type stability scores is significantly lower for partially disruptive variants than non-disruptive ($P = 0.0077$ by one-tailed *U*-test) and

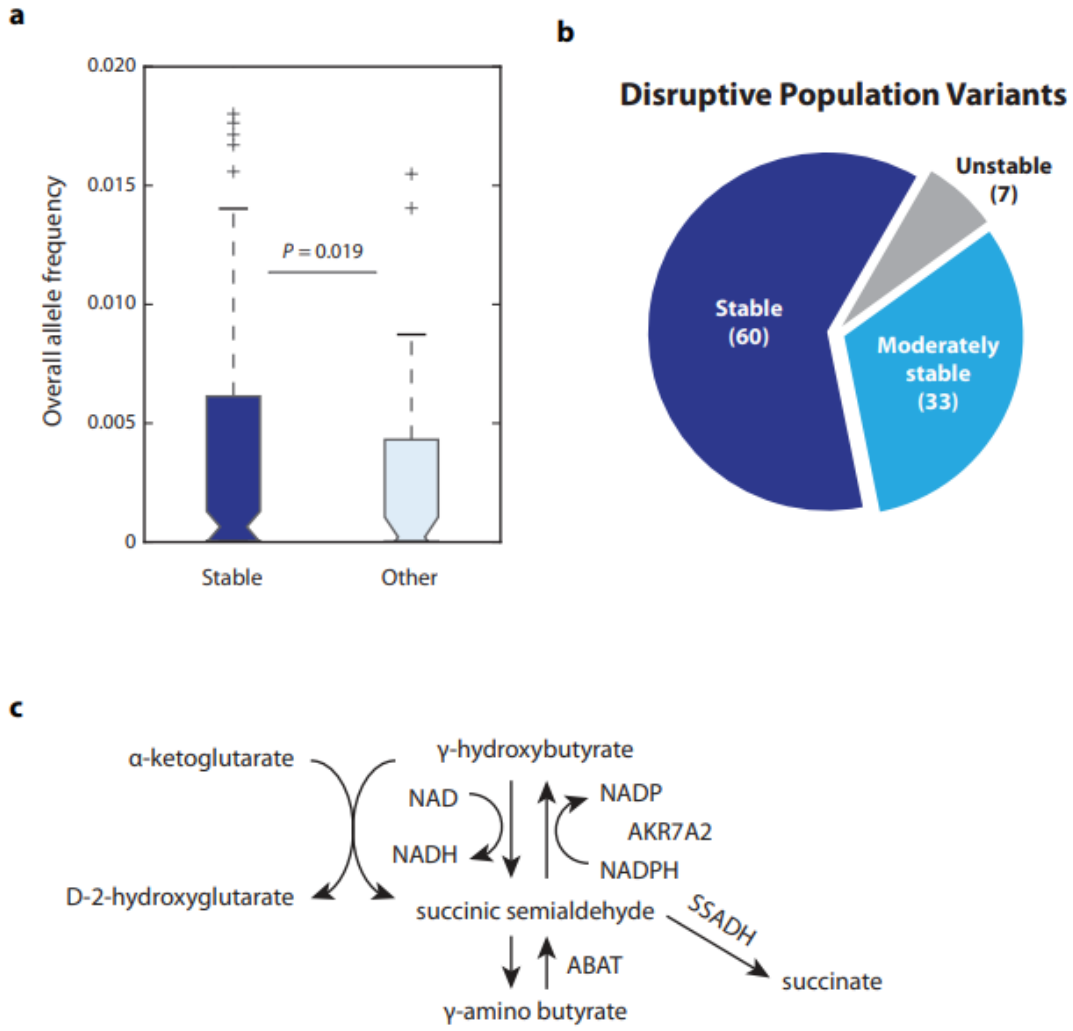


Figure 17. Protein-destabilizing variants are selectively constrained and do not fully account for interaction perturbation phenotypes.
a, Distribution of allele frequencies for variants categorized as stable (n = 214) or other (n = 64). Other was constructed by combining moderately stable and unstable variants. P values by one-tailed U-test. **b,** Distribution of interaction-disruptive ExAC variants across three stability categories. **c,** γ-hydroxybutyrate metabolism pathways involving AKR7A2, ABAT, and SSADH.

nominally reduced for null-like variants in comparison to partially disruptive variants (**Figure 16d**). While destabilized protein expression certainly influences protein interaction perturbations, we note that only seven cases (7%) in which an interaction-disruptive variant resulted in unstable mutant protein expression were found (**Figure**

16e). As such, we conclude that most disruptive variants function by inducing local structural perturbations that disrupt specific protein interactions as opposed to destabilizing protein stability as a whole, which agrees with previous studies on disease-associated mutations[23,24]. These results further highlight the importance of dissecting specific interaction disruptions induced by SNVs.

To demonstrate that stably expressed, disruptive variants can be functionally relevant even at common allele frequencies, we characterized a null-like, common variant, A142T (MAF_{Eur} = 9.3%), on the protein AKR7A2 (**Figure 16f**). AKR7A2 is an NADPH-dependent aldo1-keto reductase that catalyzes the reduction of succinic semialdehyde (SSA) to gamma-hydroxybutyrate (GHB), an important reaction in the degradation pathway for the inhibitory neurotransmitter GABA[173]. Since AKR7A2 is a dimer in solution and A142T disrupts an AKR7A2 interaction with itself, we hypothesized that this mutation might also impact AKR7A2 enzymatic activity. As such, we purified recombinant wild-type and mutant AKR7A2 protein to test for changes in NADPH-dependent turnover of SSA (Methods). Accordingly, we found that k_{cat}/K_M decreased from 1.8×10^7 $min^{-1} \cdot M^{-1}$ for wild-type protein to 1.0×10^7 $min^{-1} \cdot M^{-1}$ for AKR7A2_A142T (P = 0.035 by one-tailed t-test, **Figure 16g**). In addition to impacting SSA turnover, the A142T mutation is reported to significantly decrease the *in vitro* metabolism of both doxorubicin and daunorubicin by AKR7A2, which could have important implications in cancer therapy[174]. Moreover, missense mutations that impair ABAT and SSADH activity, enzymes immediately upstream of AKR7A2 (**Figure 17b**), can result in severe human neurological disorders[175-177]. Hence, we postulate that AKR7A2_A142T may indeed be functionally relevant in genetic backgrounds with lowered ABAT or SSADH activity.

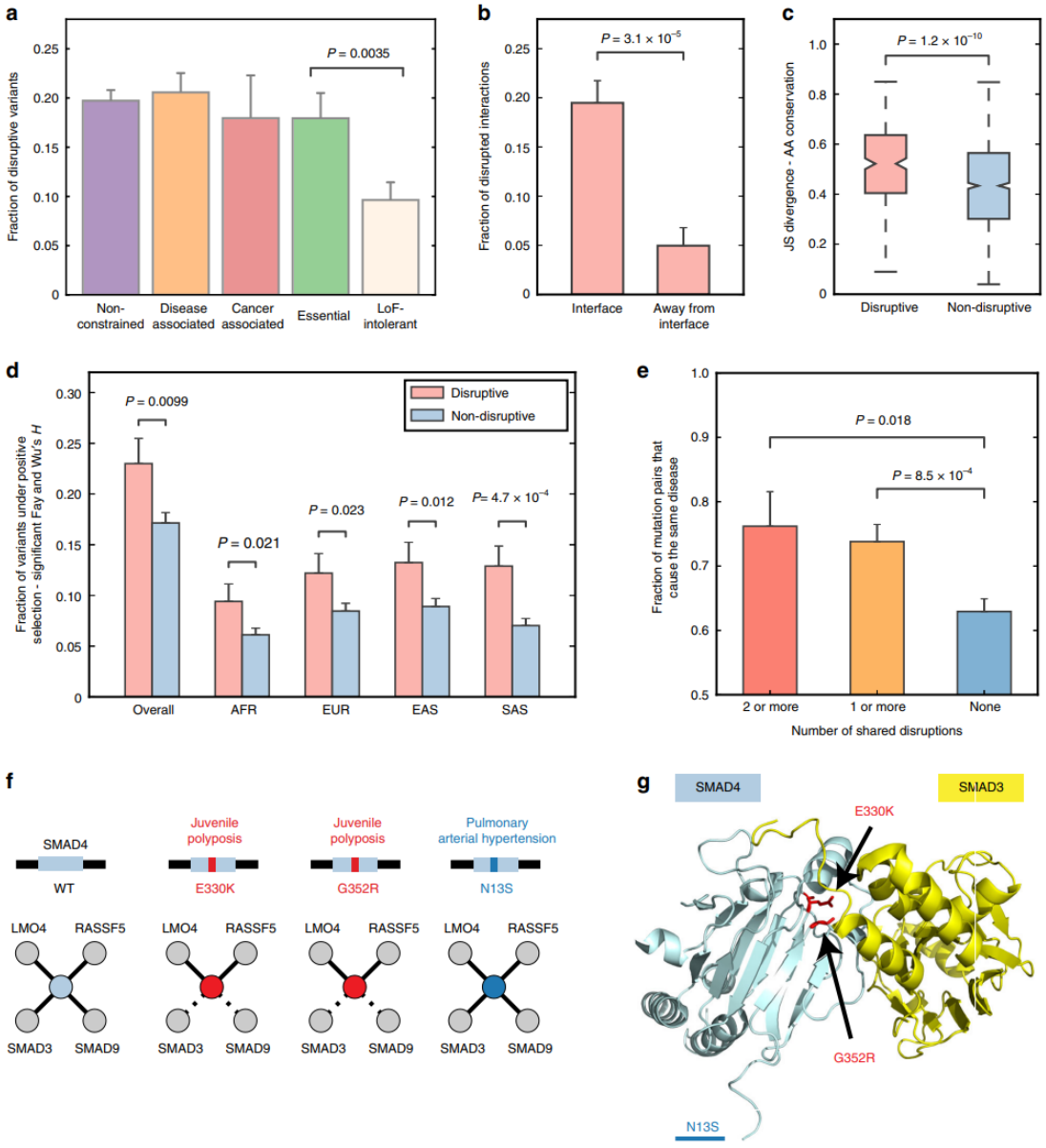


Figure 18. Disruptive variants occur in important gene groups and at conserved genomic sites.

a, Fraction of disruptive variants that occur in non-constrained (*n* = 1,349), disease-associated (*n* = 423), cancer-associated (*n* = 78), essential (*n* = 223), or LoF-intolerant genes (*n* = 270). **b,** Fraction of interactions disrupted by variants that occur on interface residues or interface domains (*n* = 307) in comparison to interactions disrupted by variants that occur away from interaction interfaces (*n* = 41). **c,** Distribution of Jensen-Shannon divergence scores for amino acid residues at sites corresponding to disruptive and non-disruptive variants. Larger scores indicate more conserved sites. **d,** Fraction of disruptive variants found in genomic regions where Fay and Wu's *H* is significant measured across four different population groups and across overall population. **e,** Fraction of mutations pairs that lead to the same disease for germline mutations that share two or more disrupted interactions (*n* = 42), share one or more disrupted

interactions (n = 271), or do not share disrupted interactions (n = 599). **f,** Schematic of interaction disruption profiles for SMAD4 disease-associated mutations E330K, G352R, and N13S. Corresponding disease names are labeled. **g,** Co-crystal structure of SMAD4-SMAD3 interacting proteins (PDB ID: 1U7F). Disease-associated mutations are labeled. Structure covers SMAD4 residues 315-546 and therefore N13S mutation is not represented on this structure. Error bars in **a, b, d,** and **e,** indicate +SE of proportion. P values in **a, b, d,** and **e,** by one-tailed Z-test. P value in **c,** by one-tailed U-test. * $P < 0.05$. See also **Figure 20**.

Disruptive variants are widespread in disease-relevant genes

We next investigated how disruptive variants are distributed across different gene categories and protein functional sites. We observed comparable enrichment for disruptive variants across disease-associated, cancer-associated, and essential gene sets (**Figure 18a**; Methods); this enrichment was also comparable to the fraction of disruptive variants found across all genes tested in our SNV-interaction network, excluding highly constrained LoF-intolerant genes (pLI ≥ 0.9) which were significantly depleted for disruptive variants in comparison to other gene sets (**Figure 18a**). LoF-intolerant genes correspond well with haploinsufficient genes[142] in which a single mutant copy of a gene is enough to be deleterious. Such genes would be highly sensitive to disruptive variants as a result, potentially explaining the lower fraction of disruptive variants observed in such genes. Notably, duplicate or functionally similar genes can compensate for corresponding proteins impacted by a disruptive mutation. However, we found no enrichment for disruptive variants within a published set of duplicate genes[178] in comparison to non-disruptive variants (**Figure 19a**), nor within a custom-generated set of sequence-conserved, functionally similar proteins (**Figure 19b**; Methods). In contrast, a sizable proportion of the disruptive variants in our SNV-interaction network occur in genes relevant to human disease and traits, warranting further exploration into their potential impact.

The structural and genomic loci at which a disruptive variant occurs is strongly indicative of the functional relevance of the mutation. Similar to disease-associated mutations[24], we found that variants located at the interaction interface disrupted

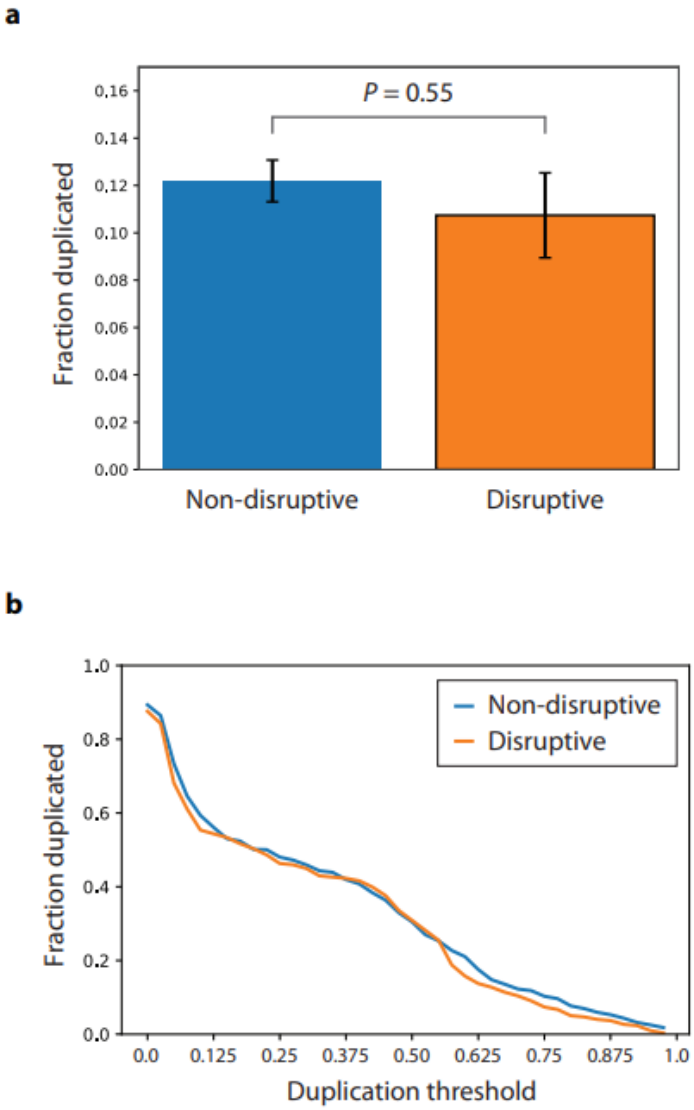


Figure 19. Disruptive variants are not biased towards redundant genes.
a, Genes harboring non-disruptive and disruptive variants were intersected with genes found in the Duplicated Genes Database. For non-disruptive and disruptive variants, the fraction of genes that overlap with genes listed in this database were plotted. Error bars indicate ±SE of proportion. P values by two-tailed Z-test. **b,** Sets of sequence-conserved, functionally similar proteins were generated at increasingly stringent thresholds for defining gene duplication. Proteins harboring non-disruptive and disruptive variants were intersected with these sets and the fraction of duplicate proteins was plotted at different duplication thresholds. A higher duplication threshold indicates a more stringent cutoff criteria for defining functionally similar proteins.

interactions significantly more often than variants located away from the interface (19.2% and 5.0%, respectively; $P = 3.9\times10^{-5}$ by one-tailed *Z*-test, **Figure 18b**; Methods). Protein sites corresponding to disruptive variants were also found to be substantially more conserved than those for non-disruptive variants ($P = 1.2\times10^{-10}$ by one-tailed *U*-test, **Figure 18c**). Moreover, purifying selection may also be more specific to disruptive variants at conserved protein sites than non-disruptive variants at equally conserved sites. To demonstrate this, we binned disruptive and non-disruptive variants by their corresponding Jensen-Shannon Divergence (JSD) scores, an amino acid-based metric for conservation, and then compared their mean allele frequency per JSD scoring bin (Methods). We found that while allele frequencies for both disruptive and non-disruptive variants were somewhat comparable at low JSD conservation scores, allele frequency for disruptive variants strongly decreased across increasingly stringent JSD cutoffs in comparison to non-disruptive variants (**Figure 20a**). A similar pattern was also observed using a genomic, as opposed to an amino acid-based, measure for conservation, phyloP[179] (**Figure 20b**). Therefore, in addition to frequently occurring in disease-relevant genes, disruptive variants also frequently occur at functionally important sites in these genes, further implying that a significant fraction of these disruptive variants may be phenotypically relevant.

In addition to exploring the relationship between conservation and disruptive variation, we also investigated whether disruptive variants tend to occur at genomic regions under positive selection. We applied a test of positive selection based on the distribution of allele frequency around a variant using whole-genome sequencing data from Phase 3 of the 1000 Genomes Project[180] (Methods). We observed that genomic regions with disruptive variants exhibit a significant signature of positive selection more often than those with non-disruptive variants. This is the case both within 1000 Genomes continental population groups and globally (**Figure 18d**). This result may

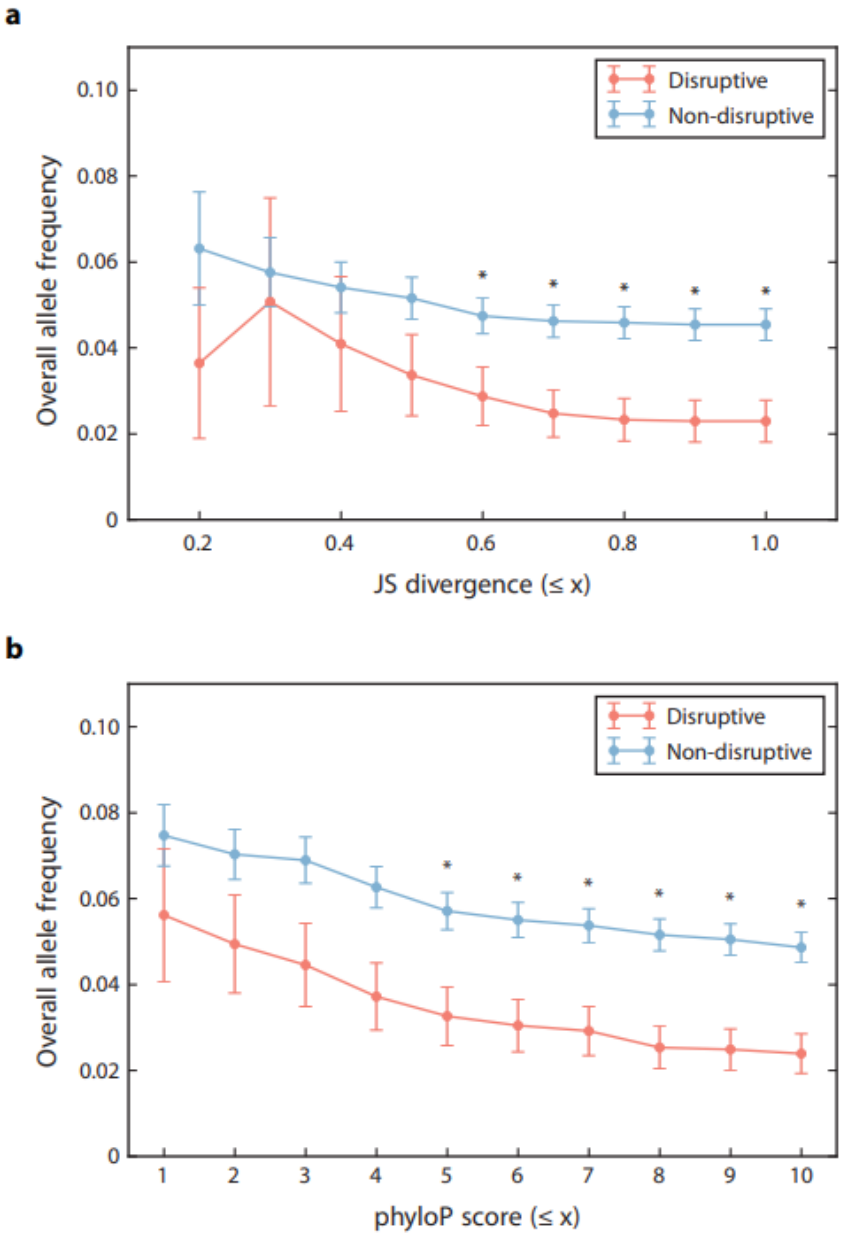


Figure 20. Purifying selection may be stronger for disruptive variants at conserved protein sites.
a, Relationship between conservation and allele frequency for disruptive and non-disruptive variants examined across increasingly stringent cutoffs for JS divergence scores. Error bars indicate ±SE of mean. **b,** Relationship between conservation and overall allele frequency for disruptive and non-disruptive variants examined across increasingly stringent phyloP scores. Error bars indicate ±SE of mean. P values by one-tailed Z-test. * $P < 0.05$

point to the functional importance of some of the disruptive variants identified here. Therefore, this result also facilitates molecular interpretation of positive selection signals, both in terms of interaction perturbations and by investigating the functions of interacting proteins lost and gained in the presence of these disruptive variants.

Identifying phenotypic SNVs via matching disruption profiles

Previous studies have shown that disease-associated mutations often function by perturbing specific protein-protein interactions[23-25]. We therefore investigated whether a disruptive population variant with the same interaction impact as a known disease-associated mutation could also result in disease. To do this, we first examined whether pairs of disease-associated mutations that occur on the same gene tend to result in the same interaction perturbations (Methods). We found that pairs of disease-associated mutations that share at least one or more disrupted interactions resulted in the same disease significantly more often than mutations that did not share any disrupted interactions (0.738 to 0.630, respectively; $P = 8.5\times10^{-4}$ by one-tailed *Z*-test, **Figure 18e**). This trend persisted when mutation pairs shared two or more disrupted interactions in comparison to no shared disrupted interactions (0.760 to 0.630, respectively; $P = 0.018$ by one-tailed *Z*-test, **Figure 18e**). This result therefore suggests that shared interaction disruption profiles may be an informative approach to prioritizing candidate disease-associated mutations.

To demonstrate how pairs of disease-associated mutations on the same gene with matching disruption profiles can result in the same disease, we highlight three disease-associated mutations on SMAD4 (**Figure 18f**), a crucial protein in the TGFβ/SMAD signaling pathway. Two mutations on SMAD4, E330K and G352R, are associated with juvenile polyposis[181,182] while a third mutation, N13S, results in a clinically distinct disease, pulmonary arterial hypertension[183]. We observed that E330K and G352R

cluster together in three dimensional space near the SMAD4-SMAD3 interaction interface (**Figure 18g**). N13S, in contrast, appears positioned away from E330K and G352R near the N-terminus of SMAD4. In agreement with the proximal clustering of E330K and G352R near the SMAD4-SMAD3 interaction interface, both mutations disrupted the SMAD4 interaction with SMAD3 in addition to disrupting the SMAD4-SMAD9 interaction (**Figure 18f**). These SMAD protein disruption results agree with previous evidence implicating the TGFβ/SMAD signaling pathway in the formation of juvenile polyposis[184,185]. In contrast, the N13S mutation left SMAD4 interactions with SMAD3 and SMAD9 intact, which agrees with a previous study that found no evidence that N13S alters SMAD-mediated signaling[183].

With this example as a template, we then explored cases in which both an ExAC variant and a known disease-associated mutation shared the same disruption profile with the goal of determining whether the population variant exhibited evidence of the same disease phenotype. To do this, we tested two mutations with matching disruption profiles on the protein PSPH (**Figure 21a**): (i) T152I, a rare variant (MAF = 0.10%) in ExAC that disrupts an interaction with itself and (ii) D32N, which also disrupts an interaction with itself and causes phosphoserine phosphatase deficiency in a compound heterozygous individual with two deleterious PSPH mutations[186]. An additional PSPH non-disruptive rare variant, T149M, was included as a control. Since PSPH exists as a dimer in solution and can aggregate when mutations that interfere with dimerization are introduced[187], we reasoned that mutations that disrupt this dimerization may also reduce PSPH enzymatic activity. We therefore purified recombinant wild-type, D32N, T152I, and T149M PSPH proteins and measured for changes in phosphatase activity for PSPH mutants relative to wild-type using a malachite green assay. Our *in vitro* assays revealed that T152I significantly reduced PSPH phosphatase activity to 59.2% ± 4.3% (P = 0.0010 by one-tailed t-test), which nearly matched the D32N reduction in activity

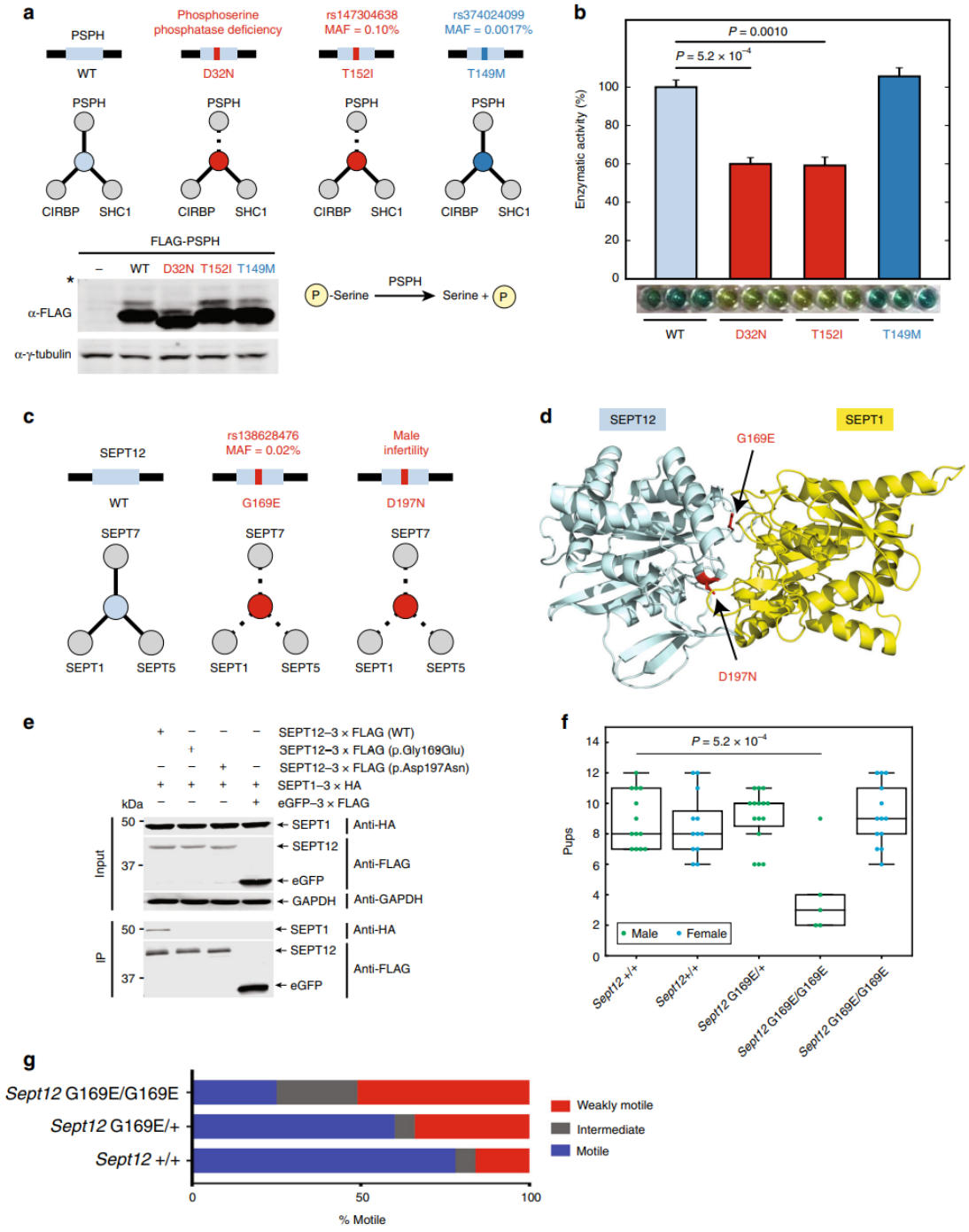


Figure 21. Prioritizing candidate disease-associated mutations through shared disruption profiles.

a, Schematic of interaction disruption profiles for disease-associated mutation D32N and rare variants T152I and T149M. Stable expression of FLAG-tagged wild-type and mutant PSPH proteins was validated by Western blot using α-FLAG. α-γ-Tubulin was used as a loading control. A brief diagram of PSPH phosphatase activity is shown. **b,** Enzymatic activity of

purified recombinant wild-type and mutant PSPH proteins using phosphoserine substrate was measured *in vitro* using a malachite green assay performed in triplicate. Enzymatic activities for PSPH mutants are shown in proportion to wild-type activity. Error bars indicate +SE of mean. * $P < 0.01$. *P* value by one-tailed *t*-test. **c,** Schematic of interaction disruption profiles for SEPT12 rare variant G169E and disease-associated mutation D197N. **d,** Homology model of SEPT12-SEPT1 interaction. PDB ID 5CYO chains A and B used as template. Disruptive mutations on interaction interface are labeled. **e,** Disruption of SEPT12 interaction with SEPT1 by G169E and D197N was validated by co-IP. SEPT12 bait proteins were detected using α-FLAG. SEPT1 prey was detected using α-HA. α-GAPDH was used as a loading control. **f,** Fertility tests of two-to-six month old WT (n=2 males, avg=8.9 ± 0.51; n=2 females, avg=8.6 ± 0.61) and *Sept12*$^{G169E/G169E}$ (n=3 males, avg=4.0 ± 1.3; n=2 females, avg=9.2 ± 0.57) mice bred to age-matched controls. Litter sizes were recorded. Blue = males. Red = females. All comparisons are not significant except for male WT vs male *Sept12*$^{G169E/G169E}$ ($P = 0.00052$; by two-tailed *t*-test). **g,** Assessment of sperm motility of WT (n=2, sperm=166), *Sept12*$^{G169E/+}$ (n=4, sperm=484), and *Sept12*$^{G169E/G169E}$ (n=3, sperm=416) mice.

(60.0% ± 3.3%, $P = 6.6\times10^{-4}$ by one-tailed *t*-test compared to wild-type). In contrast, T149M showed no significant change in enzymatic activity relative to wild-type ($P = 0.19$ by one-tailed *t*-test, **Figure 21b**). Because phosphoserine phosphatase deficiency is a recessively inherited condition[186], our findings suggest that T152I may lead to the same disease phenotype in homozygous or compound heterozygous individuals.

To further demonstrate how potentially physiologically-relevant mutations can be identified using shared disruption profiles, we also characterized a pair of disruptive mutations on the GTPase, SEPT12: a rare variant not known to associate with any disease phenotypes, G169E (MAF = 0.02%), and D197N, an infertility-causing mutation in men[188]. Both mutations perturbed interactions with SEPT7 and SEPT2 subgroup proteins, SEPT1 and SEPT5 (**Figure 21c**). These perturbations are particularly relevant because SEPT12 is known to interact with other septin proteins found in the SEPT2, SEPT6, and SEPT7 protein subgroups to form a filamentous structure at the sperm annulus[189-191]. Moreover, the infertility-causing mutation SEPT12_D197N, which was previously shown to perturb interactions with these same septin subgroup proteins, resulted in a disorganized sperm annulus and poor sperm motility in a mouse model for D197N[191]. Lastly, using homology modeling, we observed that both G169E and D197N mutations occur at SEPT12 interaction interface

residues with SEPT1 (**Figure 21d**) and confirmed that both mutations disrupt the SEPT12-SEPT1 interaction without reducing protein stability in 293T cells (**Figure 21e**). These results demonstrate that these mutations function by specifically perturbing SEPT12 protein-protein interactions as opposed to disrupting SEPT12 stability as a whole.

We then investigated whether these matching SEPT12 molecular phenotypes result in corresponding organismal phenotypes by generating *Sept12*^{G169E} mice using a CRISPR-editing approach[192]. We found that homozygous *Sept12*^{G169E} males were subfertile in comparison to wild-type males (**Figure 21f**). Notably, sperm from homozygous *Sept12*^{G169E} males exhibited poor motility (**Figure 21g**), a phenotype also reported for *Sept12*^{D197N} male mice[191]. These observations of poor sperm motility and subfertility in mice suggest that SEPT12_G169E may deleteriously impact fertility in men homozygous for this mutation, although we note that no individuals homozygous for SEPT12_G169E have been reported in ExAC. Taken together with our *in vitro* data, these results also demonstrate how shared disruption profiles can be used to prioritize candidate disease-associated mutations.

Discussion

Disentangling the phenotypic impact of functional missense mutations from benign mutations has proven to be uniquely challenging[33,34,193-195]. Conventions for determining which missense mutations are functional vary widely[3,32,143], as do their genome-wide estimates for the number of functional coding mutations per individual (**Figure 14c**). These inconsistencies are problematic since accurate measurements of the impact of SNVs on protein functions are essential for generating concrete hypotheses about disease etiology based on molecular mechanisms[35]. Therefore, in the absence of

a consensus metric for assessing the functional impact of missense mutations across a large set of proteins, we directly measured the impact of 1,676 missense ExAC-listed population variants (811 with MAF > 0.1%) across 4,109 protein-variant interaction pairs and identified 298 disruptive variants affecting 669 human protein interactions. In this manner, we have constructed an unbiased resource to examine the relationships between the population genetic and evolutionary characteristics of SNVs and their functional impact genome-wide.

By weighing our measured disruption rates against their expected proportions per individual genome, we further determined that 10.5% of missense variants per individual are expected to be disruptive. It should be noted that, like any high-throughput assay, Y2H cannot detect all interactions of a given protein. If we were able to detect more interactions, we would likely discover more interaction disruptions. Therefore, this 10.5% figure represents only a lower-bound estimate for the number of disruptive missense variants per individual. Furthermore, considering that interaction perturbations are just one way in which mutations can perturb protein function, genome-wide surveys for other types of activities (e.g., enzymatic activities, transcription factors binding to DNA, etc) may reveal that functional variants, at least at the molecular-level, are even more widespread than suggested here. Finally, we note that literature-curated sources are not appropriate for reproducing the analyses presented here because of their strong biases to synthetic and very rare mutations. Even literature-curated mutations listed at appreciable allele frequencies may be inappropriate since such mutations are often selected because of their known disease associations. For example, a recent study comprehensively collected the impact of 7,955 mutations on human protein interactions published in the literature[48]; however, only 161 of these mutations were without disease annotations and listed in ExAC, of which 49 occurred at appreciable frequencies (MAF > 0.1%).

The genetic and genomic context in which a variant occurs is crucial for properly interpreting the functional impact a disruptive mutation may have. While haplosufficiency likely mitigates the impact of numerous disruptive variants, an individual already harboring one disruptive variant becomes sensitized to the consequences of subsequent mutations in the same gene or pathway. For example, we identified a null-like, common variant, A142T, on the protein AKR7A2 that significantly reduces enzymatic activity relative to wild-type (**Figure 16g**). This mutation alone likely has a minimal impact on fitness; however, mutations to enzymes immediately upstream to AKR7A2, particularly ABAT and SSADH (**Figure 17b**), can result in severe neurological disorders[175-177]. Co-occurrence of AKR7A2_A142T with similarly disruptive mutations in ABAT or SSADH could therefore result in a neurological disorder that would not otherwise occur in an individual harboring only a single disruptive mutation.

Such relationships are frequent in complex disease, including cancer and heart disease, which unlike Mendelian mutations, require multiple mutations on more than one gene to cause a disorder. Each disease-associated mutation in complex disease therefore contributes a certain measure of disease risk that can be quantified by a GWAS, and some authors consider these effects to be approximately additive across loci[196]. Measuring how one mutation modulates the impact of another is challenging; however, measuring which mutations are individually functional is a crucial first step. Hence, we anticipate that our SNV-interaction network will serve as a pivotal framework for defining the epistatic relationships that modulate the impact of disruptive variants, particularly for partially penetrant variants that only result in disease in certain genetic backgrounds.

The results of our study may have important implications in related fields such as pharmacogenomics and toxicogenomics. Disruptive SNVs on enzymes may alter the

metabolic kinetics of impacted enzymes, while SNVs on transporters and targets of drugs may lead to changes in the pharmacokinetic and pharmacodynamic properties of their corresponding proteins. For example, the D816H/V mutations on the receptor tyrosine kinase, KIT, confers resistance to imatinib and sunitinib by shifting the conformational equilibrium of KIT[197]. As a potential resource to pharmacogenomics and toxicogenomics, we provide a table of all disruptive SNVs that may be relevant to drug action (**Table 15**; Methods).

Several methods to experimentally measure the impact of coding mutations at large scales have been recently reported[24,52,198,199]. The depth of proteins, variants, and interactions presented here complements these previous methods well. For example, Fields and Shendure developed a massively parallel single-amino-acid mutagenesis pipeline, named PALS, that can generate nearly all potential singleton mutations possible for a particular gene of interest[198]. This impressive depth makes PALS an excellent method for studying extensive variation in a single protein, most notably *TP53*[198] and *BRCA1*[200] but remains to be optimized for studying variation across a large set of unique genes. In contrast, our mutagenesis approach allowed us to survey >2,000 mutations across 847 unique genes. Similarly, while Y2H is widely used for characterizing the impact of mutations on protein-protein interactions[23,24], several derivatives for detecting perturbations by Y2H exist. For instance, Stelzl and colleagues developed the Int-Seq platform for probing protein-protein interaction disruptions using a Reverse Two-Hybrid (R2H) approach[199]. While this R2H approach increases assay sensitivity, a R2H reference interactome is not yet available, limiting the coverage of this approach to a handful of interactions.

Interaction perturbations constitute only a particular subset of the variety of ways in which mutations can impair protein function. Large-scale surveys assessing other modes of altering protein function are needed. For example, Lehner and colleagues

used a deep mutational scanning pipeline[50] to measure the impact of mutations on alternative splicing[201]. Continued efforts to survey all potential manners in which molecular-level perturbations can alter cellular and organismal phenotypes are needed to properly understand the impact of mutations on human health. Although our experimental framework was not designed to find potentially causal variants driving GWAS phenotypes (**Figure 22**; Methods), experimental frameworks that can differentiate functional variants from those that are non-functional will be key to identifying causal variants in common disease. Towards this goal, the genetic, protein interaction, and population-level insights presented here may represent a pivotal step forward to an improved understanding of the evolutionary forces that shape the human genome and protein function.

Methods

Selecting SNVs from ExAC, HGMD, and COSMIC databases

Population variants encoding for missense mutations were selected from ExAC release 0.3.1[142]. Unless a specific subpopulation is listed, all reported allele frequencies and allele frequency-derived calculations refer to allele frequency across all ExAC populations. Disease-associated missense mutations were obtained from HGMD (Public release version, 2014). Cancer-associated somatic missense mutations were selected from COSMIC version 84. For all three datasets, we required that (i) mutations reside on genes in either hORFeome v8.1[202] or v5.1[203], (ii) corresponded with one or more high-throughput Y2H-testable protein-protein interactions[170,204-206], and, (iii) for ExAC variants, achieved a PASS filter status. We mapped each RefSeq transcript from ExAC to an appropriate ORF in our library by looking at the top BLAXTX candidate with an E-value $\leq$ 0.001. We verified that this was a representative ORF for our mutation by

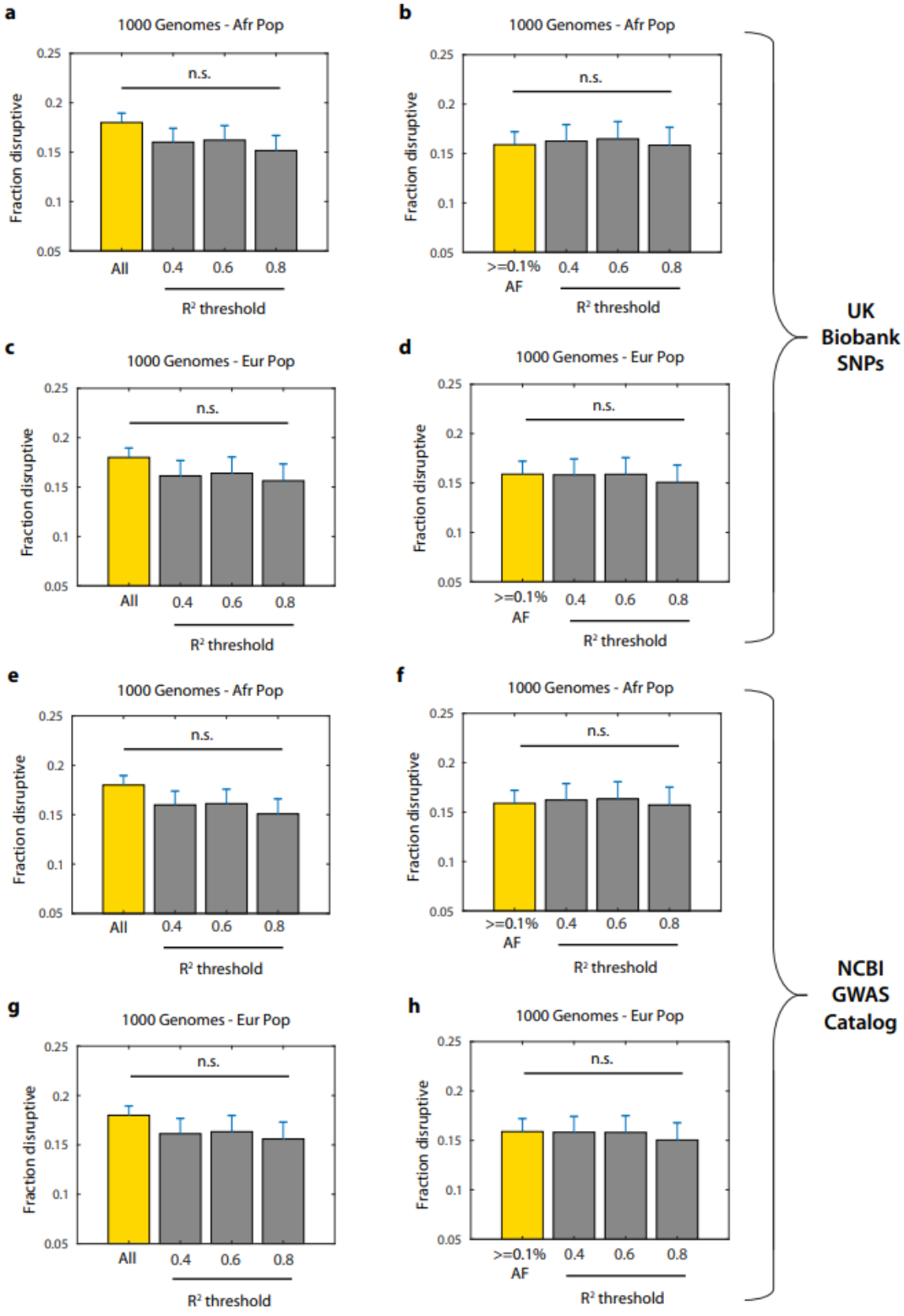

a
1000 Genomes - Afr Pop
Fraction disruptive
n.s.
All
0.4
0.6
0.8
R² threshold
b
1000 Genomes - Afr Pop
Fraction disruptive
n.s.
>=0.1% AF
0.4
0.6
0.8
R² threshold
c
1000 Genomes - Eur Pop
Fraction disruptive
n.s.
All
0.4
0.6
0.8
R² threshold
d
1000 Genomes - Eur Pop
Fraction disruptive
n.s.
>=0.1% AF
0.4
0.6
0.8
R² threshold
UK Biobank SNPs
e
1000 Genomes - Afr Pop
Fraction disruptive
n.s.
All
0.4
0.6
0.8
R² threshold
f
1000 Genomes - Afr Pop
Fraction disruptive
n.s.
>=0.1% AF
0.4
0.6
0.8
R² threshold
g
1000 Genomes - Eur Pop
Fraction disruptive
n.s.
All
0.4
0.6
0.8
R² threshold
h
1000 Genomes - Eur Pop
Fraction disruptive
n.s.
>=0.1% AF
0.4
0.6
0.8
R² threshold
NCBI GWAS Catalog

Figure 22. Disruptive variants show no bias towards GWAS phenotypes.
a, Fraction of disruptive variants for variants found in the 1000 Genomes phase 3 Afr population at all allele frequencies is plotted (yellow). Fraction of disruptive variants for variants for variants in LD with GWAS SNPs listed in the UK biobank at R2 thresholds of ≥ 0.4, ≥ 0.6, and ≥ 0.8 are also plotted (grey). **b,** Fraction of disruptive variants for variants found in the 1000 Genomes phase 3 Afr population at AF ≥ 0.1% is plotted (yellow). Fraction of disruptive variants for variants for variants in LD with GWAS SNPs listed in the UK biobank at R2 thresholds of ≥ 0.4, ≥ 0.6, and ≥ 0.8 are also plotted (grey). **c,** Same analysis as **a** but restricted to 1000 Genomes phase 3 Eur population. **d,** Same analysis as **b** but restricted to 1000 Genomes phase 3 Eur population. **e-h,** Same analyses as **a-d** but for GWAS SNPs listed in the NCBI GWAS Catalog. Error bars indicate +SE of proportion. P values by two-tailed Z-test. n.s. = not significant.

performing EDNAFULL matrix pairwise alignment using EMBOSS Stretcher. Valid representative ORFs had to be identical within a 31 amino acid window centered on the position of interest for mutagenesis. Beyond local identity, ORFs were required to have more than 95% global identity, or be an exact subset of the transcript, spanning at least a third of the query transcript.

Since over half of all variants in ExAC are singletons, to avoid oversampling rare alleles, we selected between 200-400 variants across six mutually exclusive allele count bins of 1, <10, <100, <1,000, < 10,000, and >10,000 for a total of 1676 ExAC alleles (**Figure 13b**). In each bin, we randomly selected variants on genes with Y2H-testable interactions. To minimize gene bias, we selected an average of two variants per gene. 204 HGMD mutations listed as DM (disease-causing mutations) were selected in accordance to criteria detailed in [20] but expanded to test across all amenable Y2H protein-protein interactions. 162 COSMIC mutations among 110 different genes with available hORFeome clones were also tested across all amenable Y2H protein-protein interactions. Genes listed in the Cancer Gene Census (v84) and listed as a Tier 1 known drivers in IntOGen (2016.5) were designated as *Known cancer genes*. Genes not listed in the Cancer Gene Census and not listed as a driver in IntOGen were designated as *Other genes* (**Figure 14b**).

Large-scale cloning of SNVs through Clone-seq pipeline

Single colony-derived mutant clones were constructed using a previously described methodology termed Clone-seq[24], a high-throughput mutagenesis and next-generation sequencing platform. In brief, wild-type clones were picked from hORFeome clones and served as templates for site-directed mutagenesis performed in 96-well plates using site-specific mutagenesis primers (Eurofins). To minimize sequencing artifacts, PCR was limited to 18 cycles using Phusion polymerase (NEB, M0530). PCR products were digested overnight with *Dpn*I (NEB, R0176) then transformed into competent bacteria cells to isolate single colonies. Up to four colonies per individual mutagenesis reaction were then hand-picked and arrayed into 96-well plates and incubated for 21 hrs at 37°C under constant vibration. After incubation, glycerol stocks were generated; clones were then pooled into independent bacterial pools. An additional maxiprepped bacterial pool comprising only wild-type DNA templates corresponding to each mutagenesis PCR reaction was also prepared. Maxiprepped clonal DNA from each bacterial pool was then combined through multiplexing (NEB, E7335) and sequenced in a single 1x75 single-end Illumina NextSeq run. Properly mutated clones which differed from their sequenced wild-type templates only by the desired single base-pair mutation – and nowhere else – were then identified by next-generation sequencing analysis and recovered from their corresponding single colony glycerol stocks.

Identifying successfully mutated clones

After de-multiplexing, mapped reads corresponding to the generated pools (wildtype plus up to four mutant pools) were mapped to genes of interest using the BWA mem algorithm (bwa mem -a -t 12 <reference> <reads>). In order to detect both the desired variant as well as undesired off-target mutations, we first obtained the read counts for each allele (A, T, C, G, insertion, or deletion) for all positions in the clones. Using these

read counts we calculated the score for a given position, pos, containing a mutation from the wildtype allele, WT, to a mutant allele, Mut, as follows:

$$\text{Score(WT, pos, Mut)} = \frac{\text{Observed}_{\text{Mut, pos}}}{\text{Expected}_{\text{Mut,pos}}}$$

where $\text{Observed}_{\text{Mut,pos}}$ is the observed fraction of reads at position pos matching allele Mut and $\text{Expected}_{\text{Mut,pos}}$ is the fraction of reads at position pos matching allele Mut that we would expect to see if it the mutation in question had indeed occurred. We define this fraction as:

$$\text{Expected}_{\text{Mut, pos}} = \frac{1}{\text{TotalMutations}} + (\text{TotalMutations}-1)*\text{SeqErr(pos)} - (\text{Alleles}-1)*\text{SeqErr(pos)}$$

where TotalMutations is the total number of mutants attempted for a particular ORF (i.e. the number of copies of the ORF included in the pool), SeqErr(pos) refers to the inherent sequencing error, and Alleles is the total number of alleles.

To explain further, assuming that all clones for a particular gene contribute a similar number of reads, we expect that if one of the clones for a gene contains a mutation to the Mut allele at position pos, we should see $\frac{1}{\text{TotalMutations}}$ fraction of the reads match the Mut allele. Due to sequencing errors, we expect the true fraction observed to deviate slightly from this base fraction. We first add a term for the fraction of Mut alleles that we expect to see as a result of sequencing errors in the other non-mutant clones for the gene. Second, we subtract a term for sequencing errors in the mutant clone converting the Mut allele to any of the $(\text{Alleles} - 1)$ other alleles. We define the sequencing error as the average fraction of non-WT bases observed in the ten closest positions that were not targeted for mutagenesis.

Based on comparisons to Sanger sequencing results, we set a threshold of

Score(WT, pos, Mut)≥0.5 to call true mutations. In identifying successful instances of site-directed mutagenesis, we first checked for the presence of the desired mutation using this score threshold. Using the scores for all other positions along the clone, we then screened each successful mutant for the presence of any other unwanted mutations that may have been introduced as PCR artifacts. Any clones containing unwanted mutations were removed, and the remaining clones were sorted using a combination of their desired mutation score, maximum undesired mutation score, sequencing coverage, and sequencing quality.

Calculating proportion of functional mutations exome-wide

The total number of missense variants in ExAC release 0.3.1, diagrammed in **Figure 13b**, was determined by summing the adjusted allele count found in the ExAC database for all variants annotated as *missense_variant* in at least one transcript. The number of functional mutations was calculated by multiplying the mean disruption rate per individual by the total number of missense variants in ExAC.

The total number of missense variants in the 1000 Genomes Consortium – Phase I, Genomes of the Netherlands, and Exome Sequencing Project – Phase I were obtained from [7], [10], and [4], respectively. Calculations for the number of functional missense mutations from each source are annotated in **Tables 9-11**. We note that the number of functional mutations by mutation type was not reported for ESP variants in [4]. As such, functional nonsynonymous mutations, including nonsense variants, were instead reported for ESP – Phase I. We expect the proportions of functional missense variants for ESP, 5.5% and 10.0% using conservative and liberal criteria counts listed in [4], respectively, (**Table 11**), to be slight overestimates as a result.

Profiling disrupted protein interactions through Y2H

Clone-seq-identified mutant clones were transferred into Y2H vectors pDEST-AD and pDEST-DB by Gateway LR reactions then transformed into *MAT*a Y8800 and *MAT*α Y8930, respectively. All DB-ORF *MAT*α transformants, including wild-type ORFs, were then mated against corresponding wild-type (WT) and mutant AD-ORF *MAT*a transformants in a pairwise orientation using automated 96-well procedures to inoculate AD-ORF and DB-ORF yeast cultures followed by mating on YEPD agar plates. All DB-ORF yeast cultures were also mated against *MAT*a yeast transformed with empty pDEST-AD vector to screen for autoactivators. After overnight incubation at 30°C, yeast were replica-plated onto selective Synthetic Complete agar media lacking leucine and tryptophan (SC-Leu-Trp) to select for mated diploid yeast then incubated again overnight at 30°C. Diploid yeast were then replica-plated onto SC-Leu-Trp agar plates also lacking histidine and supplemented with 1 mM of 3-amino-1,2,4-triazole (SC-Leu-Trp-His+3AT) as well as SC-Leu-Trp agar plates lacking adenine (SC-Leu-Trp-Ade). After overnight incubation at 30°C, plates were replica-cleaned and incubated again for three days at 30°C.

Disrupted protein-protein interactions were identified as follows: (1) mutated protein reduces growth by at least 50% relative to wild-type interaction as benchmarked by twofold serial dilution experiments, (2) neither wild-type or mutant DB-ORFs are autoactivators, (3) reduced growth phenotype reproduces across three screens. A mutation was scored as disruptive if one or more corresponding protein-protein interactions were disrupted and was scored as non-disrupted if otherwise. Mutations tested against two or more interactions partners were further categorized as non-disruptive, partially disruptive, and null-like if no tested interactions were perturbed, some tested interactions were perturbed, or all tested interactions were perturbed, respectively. PSPH interactions with CIRBP and SHC1 were detected using PCA. No

significant change in PCA signal intensity was detected between any wild-type and mutant PSPH interaction with CIRBP and SHC1 and therefore all mutant PSPH interactions with CIRBP and SHC1 were scored as non-disrupted. Interaction disruption data for all tested ExAC variants, COSMIC somatic mutations, and HGMD disease-associated mutations can be found in **Tables 12-14**, respectively.

DUAL-FLUO assay to measure SNV impact on protein stability

In order to screen for variants that destabilize protein expression, we first screened for stably expressed GFP-tagged wild-type proteins. To do this, we transferred wild-type ORFs into pDEST-DUAL by Gateway LR reactions. HEK293T cells (ATCC, CRL-3216) were then seeded onto black 96-well flat-bottom dishes (Costar, 3603). HEK293T cells were maintained in complete DMEM medium supplemented with 10% FBS. All cell incubation steps were performed at 37°C under air with 5% CO_2. Cells were grown to 60% confluency then co-transfected using 150 ng sample DNA in pDEST-DUAL and 1.0 µL of 1 mg/mL PEI (Polysciences Inc, 23966) mixed thoroughly with 20 µL OptiMEM (Gibco, 31985-062). Four replicates of empty pDEST-DUAL and four replicates of empty pcDNA-DEST47 were also transfected per 96-well plate as positive controls for mCherry expression and negative controls for GFP expression, respectively. After 72 hrs incubation, stably expressed wild-type GFP-tagged proteins were identified using a Tecan M1000 plate reader. Samples that resulted in GFP and mCherry expression significantly above background were confirmed by automated fluorescence microscopy using an ImageXpress system. In this manner, we identified 202 wild-type genes corresponding to 278 ExAC variants. Single clones for ExAC variants were then transferred into pDEST-DUAL by Gateway LR reactions for further screening.

Wild-type and mutant ORF pairs in pDEST-DUAL were transfected into 293T cells in the same fashion as described for our first wild-type screen, including eight total

pDEST-DUAL and pcDNA-DEST47 controls per plate. Mutant ORFs corresponding to a particular wild-type ORF were always partitioned onto the same plate. After 72 hrs incubation, GFP and mCherry fluorescence readings using a Tecan M1000 plate reader were measured for all samples and imaged by automated fluorescence microscopy using an ImageXpress system. Mutant proteins were then processed into stable, moderately stable and unstable categories of protein expression as follows: if the ratio between mutant and wild-type stability scores fell below 0.5, indicating that the mutant protein is still expressed but at markedly reduced levels, we categorized the mutant protein as moderately stable. If mutant protein expression dropped below plate reader detection thresholds, as indicated by a mutant stability score < 0, we instead categorized the mutant protein as unstable. Mutant proteins above both thresholds are scored as stable.

Retesting disrupted and non-disrupted interactions by PCA

To confirm that variant-disrupted protein-protein interactions are reproducible across a different assay, we systemically selected a subset of Y2H-tested mutant protein interactions for retesting by PCA. Bait ORFs in pDONR223 for disruptive and non-disruptive variants were transferred into F1 Venus fragments while prey ORFs for corresponding interaction partners were transferred into F2 Venus fragments using Gateway LR reactions for a total of 192 Y2H-disrupted mutant interaction pairs and 205 non-disrupted Y2H mutant interaction pairs. Bait and prey ORF pairs from both sets were then randomly scrambled across 87 PRS and 90 RRS ORF pairs previously described in [41, 42] to minimize detection bias across different 96-well plates. As a quality control measure, interaction pairs in which either a bait or prey ORF did not amplify by PCR using F1 Venus- or F2 Venus-specific primers, respectively, were removed from PCA analysis.

To perform PCA, HEK293T cells (ATCC, CRL-3216) were seeded onto black

96-well flat-bottom dishes (Costar, 3603). HEK293T cells were maintained in complete DMEM medium supplemented with 10% FBS and incubated at 37°C under air with 5% CO_2. Cells were grown to 60-70% confluency then co-transfected using 100 ng bait vector plus 100 ng prey vector with 1.0 µL of 1 mg/mL PEI (Polysciences Inc, 23966) mixed thoroughly with 20 µL OptiMEM (Gibco, 31985-062) per transfection. After 72 hrs incubation at 37°C, a Tecan M1000 plate reader was used to measure PCA fluorescence (excitation = 514 nm; excitation = 527 nm) for all samples. A manually-adjusted gain was applied to ensure all measurements were performed within a linear range. Detection thresholds were selected such that ORF pairs resulting in a signal greater than the threshold were scored as *detected* while scores that fell below the threshold were scored as *undetected*. The fraction of recovered pairs represents the proportion of ORF pairs that scored above a given threshold over the total set of ORF pairs tested per category.

Constructing vectors for DUAL-FLUO screen and Western blot

Gateway LR reactions were used to transfer ORFs into mammalian expression vectors. The pDEST-DUAL vector for our dual-fluorescence screen was constructed by inserting an mCherry cassette independently driven by a minCMV promoter into pcDNA-DEST47 (Invitrogen, 12281-010), which features a C-terminal GFP tag. PSPH wild-type, D32N, T152I, and T149M were transferred into a pQCXIP (ClonTech, 631516) vector modified to include a Gateway cassette featuring a C-terminal 3×FLAG tag. SEPT12 wild-type, G169E, and D197N were transferred also into this same modified pQCXIP 3×FLAG vector. SEPT1 was transferred into a modified pcDNA3.1 (Invitrogen, V79020) vector featuring a C-terminal 3×HA tag. AKR7A2 wild-type and A142T were transferred into pcDNA-DEST40, which includes a V5 tag.

Cell culture for Western blotting

HEK293T cells (ATCC, CRL-3216) were maintained in complete DMEM medium supplemented with 10% FBS and incubated at 37°C under air with 5% CO_2. Cells were grown in 6-well dishes to 70-80% confluency then transfected using 2 μg of vector with 10 μL of 1mg/mL PEI (Polysciences Inc, 23966) mixed thoroughly with 150 μL OptiMEM (Gibco, 31985-062). After 24 hrs incubation, cells were gently washed three times in 1x PBS and then resuspended in 200 μL cell lysis buffer [10 mM Tris-Cl pH 8.0, 137mM NaCl, 1% Triton X-100, 10% glycerol, 2 mM EDTA, and 1x EDTA-free Complete Protease Inhibitor tablet (Roche)] and incubated on ice for 30 min. Extracts were cleared by centrifugation for 10 mins at 16,000xg at 4°C. Samples were then treated in 6x SDS protein loading buffer (10% SDS, 1 M Tris-Cl pH 6.8, 50% glycerol, 10% β-mercaptoethanol, 0.03% Bromophenol blue) and subjected to SDS-PAGE. Proteins were then transferred from gels onto PVDF (Amersham) membranes. Anti-FLAG (Sigma, F1804) at 1:3000, anti-V5 (Invitrogen, R960-25) at 1:5000, anti-HA (Sigma, H3663) at 1:3000, anti-GFP (SCBT, sc-9996) at 1:1000, anti-GAPDH (Proteintech, 60004-1-Ig) at 1:3000, and anti-γ-Tubulin (Sigma, T5192) at 1:3000 dilutions were used for immunoblotting analyses.

Protein purification of recombinant PSPH and AKR7A2

Gene-specific primers were used to clone *Bam*HI and *Xho*I restriction endonuclease digestion sites onto the 5' and 3' ends, respectively, of ORFs for wild-type, D32N, T152I, and T149M clones of PSPH by PCR. PCR products as well as a pET28a-based, custom generated pET-6xHis-SUMO expression vector were then digested overnight using *Bam*HI (NEB, R3136) and *Xho*I (NEB, R0146) restriction endonucleases. All digested products were cleaned up by gel extraction. PCR products were then ligated into double-digested pET-6xHis-SUMO vector by 10 μL T4 ligase (NEB, M0202) reactions using a 3:1 ratio of insert to template incubated for 30 min at RT. Ligated

products were then transformed into competent cells and plated to isolate single colonies. Properly ligated colonies were validated by colony PCR. Colony PCR-validated pET-6xHis-SUMO PSPH constructs were then transformed into Rosetta strain competent bacteria cells (Novagen, 71401-3).

To purify recombinant wild-type and mutant PSPH proteins, single colonies of transformed Rosetta strain bacteria were inoculated overnight for use as starter cultures. Starter cultures were used to inoculate 1.0 L LB media including kanamycin and chloramphenicol and incubated for 2-4 hrs at 37°C, shaking at 250 rpm until OD600 = 0.6. 200 μL of 1 M IPTG was then added to induce protein expression. Induced cultures were incubated for 18 hrs at 18°C, shaking at 250 rpm. After incubation, cultures were centrifuged at 4,000xg for 20 min at 4°C. Supernatant was discarded and pellet was resuspended in 35 mL Resuspension Buffer (500mM NaCl, 50mM Tris-base pH 8.0) on ice. Unless stated otherwise, all steps moving forward were performed on ice or at 4°C. Resuspended pellet was sonicated to lyse cells and then centrifuged at 16,000xg for 45 min. Supernatant was then run through a column prewashed with Wash Buffer (20mM NaCl, 20mM Tris pH 7.5) and loaded with Cobalt agarose beads (GoldBio, H-310) for purification of 6x His-tagged protein. Purified samples bound to Cobalt beads were then treated overnight with lab-purified Ulp1 protease for SUMO tag cleavage. Afterwards, samples were again run through a column prewashed with Resuspension Buffer and eluted samples were collected. Lastly, purified protein samples were fractionated by FPLC and samples lacking detectable SUMO expression by Coomassie gel were used for experiments.

Wild-type and mutant A142T recombinant proteins were prepared in the same manner as PSPH except for the following changes: (1) AKR7A2 gene-specific primers were used for PCR, followed by *Eco*RI (NEB, R3101) and *Xho*I (NEB, R0146) double digestion of PCR product and pET-6xHis-SUMO vector; (2) after induction with 200

μL of 1 M IPTG, cultures were incubated for 5 hrs at 37°C, shaking at 250 rpm.

Phosphatase activity measurements for PSPH variants

Wild-type and mutant PSPH activity were measured using a malachite green assay as follows: Malachite Green Reagent Stock was prepared by combining 30 mL Malachite Green (Sigma, M9636) with 20 mL 4.2% ammonium molybdate (Sigma, 277908) / 4M HCL and mixing for > 30 min. Malachite Green Reagent Stock was filtered through a 0.2 μm filter unit and stored at 4°C. Malachite Green Working Reagent was then prepared by adding Tween-20 to a final concentration of 0.01% in Malachite Green Reagent Stock. Using a 96-well plate (Costar, 3696), A_{620} for sodium phosphate in Malachite Green Working Reagent at concentrations of 10, 15, 20, 25, 30, 35, and 40 μM at pH 7.4 was then measured using a Tecan M1000 plate reader to generate a standard curve. Next, 100 ng of purified recombinant PSPH protein was added to 20 μL total of Assay Buffer (30 mM HEPES at pH 7.4, 1 mM EGTA, 1 mM $MgCl_2$ and 100 μM phosphoserine) and mixed with 80 μL Malachite Green. Negative controls lacking recombinant protein or phosphoserine substrate were also included. After plate incubation at 37°C for 5 min, A_{620} was measured for all samples. Percent change in phosphatase activity for mutant PSPH was measured as the ratio of mean mutant PSPH activity to mean wild-type PSPH enzymatic activity over three replicates.

Defining duplicate genes and functionally similar proteins

Duplicate genes were obtained from the Duplicated Genes Database[178] which lists 3,543 duplicate genes across 945 different gene groups. In order to compare the robustness of duplicate gene definitions across many different cutoffs, we additionally defined our own metric for protein similarity by running a BLAST of the human proteome against itself and eliminating all pairs of proteins with less than 40% sequence identity. The

remaining pairs were scored using a weighted combination of the pair's percent identity and the coverage with respect to each protein. In **Figure 19b**, we flexibly defined duplicate genes as all pairs of genes whose score met a minimal duplication threshold tested across all valid ranges (where 0 for Duplication Threshold represents no appreciable similarity and 1 represents perfect identity). Score is calculated as:

$$\text{Score}= \alpha*\text{PercentIdentity}*\text{Coverage}_{\text{Avg}}+(1-\alpha)*\text{Coverage}_{\text{Avg}}$$

where α=0.95 and $\text{Coverage}_{\text{Avg}}$ is the average coverage between both proteins.

Enrichment of disruptive mutations on interaction interfaces

We examined the positions of ExAC variant residues relative to protein-protein interaction interfaces. *On interface* was defined as either at an interface residue or in the interface domain, while *away from interface* was defined as neither at an interface residue nor in the interface domain. Interface residues and domains were defined as previously described in the AtomInt[207] and Instruct[208] databases. The fraction of interactions disrupted by variants *on the interface* or *away from the interface* was then calculated.

Metrics for evolutionary site conservation at variant sites

Jensen-Shannon Divergence (JSD) scores were obtained as previously described in Interactome INSIDER[36]. phyloP scores were obtained using the Table Browser of the UCSC Genome Browser and inputting the hg19 coordinates for each tested variant. To measure the average global allele frequency across different JSD or phyloP scores, cutoff scores of 0.2, 0.3, … , 1.0 were applied and the global allele frequencies per tested ExAC variant were averaged cumulatively across each cutoff score.

Signals of positive selection for disruptive alleles

Fay and Wu's *H* was calculated genome-wide with 1 kb sliding windows using the 1000 Genomes Phase 3 dataset[180]. Analyses were conducted in the merged global population as well as in AFR, EUR, EAS, and SAS populations individually. Genomic regions with a Fay and Wu's *H* statistic at or below the 5th percentile were considered statistically significant. Among all variants that occurred in regions with a measurable Fay and Wu's *H* statistic, the number of disruptive variants that occurred in regions with a significant *H* statistic was recorded.

Comparing disruption profiles for disease-associated SNVs

Interaction perturbation data for disease-associated mutations measured here were combined with interaction perturbation data from [19] and then filtered for mutations listed as DM in HGMD (Public release version 2017), resulting in interaction perturbation data for 495 mutations. Mutation pairs were deemed to cause the same disease if strings for their corresponding disease phenotypes listed in HGMD were equal. Mutations were compared pairwise and had to share at least one interaction in common in order to be compared. If one or more interactions were found in common, mutation pairs were categorized by either sharing two or more disrupted interaction in common, one or more disrupted interactions in common, or no disrupted interactions in common.

Calculating LD between ExAC-tested variants and GWAS SNPs

To examine whether ExAC variants that are in strong linkage disequilibrium (LD) with GWAS SNPs are more likely to be disruptive, we first extracted all SNPs associated with phenotypes in the UK Biobank GWAS Atlas[209]. We then calculated R^2 values

between all ExAC variants in our dataset and the UK Biobank GWAS SNPs, using 1000 Genomes Phase 3 data[210].

ExAC variants in strong LD with GWAS SNPs had a disruption rate that was not significantly different from the overall disruption rate (**Figure 22a**). Results were robust across multiple R^2 thresholds using either African or European allele frequencies (**Figure 22a**, **Figure 22c**). Since most GWAS SNPs occur at MAF ≥ 0.1% and a sizable fraction of our tested variants are rare, we also repeated our analysis restricted for variants at MAF ≥ 0.1% but still found no significant trend (**Figure 22b**, **Figure 22d**). As a control, we also repeated these same analyses using SNPs from the NCBI GWAS Catalog[211] and found the exact same trends as those for the UK Biobank GWAS Atlas (**Figure 22e-h**).

Developing a dataset of drug-relevant disruptive SNVs

To generate a dataset of disruptive SNVs potentially relevant to pharmacogenomics and toxicogenomics, we intersected our dataset with four sets of genes: all human enzymes, drug-metabolizing enzymes, drug targets, and drug transporters. The list of all human enzyme genes was obtained from HumanCyc version 21.5[212], while the lists of drug-related genes were obtained from DrugBank version 5.1.2[213]. Among the SNVs that we tested, 350 were on enzymes, and 84 of them disrupted at least one interaction. More specifically, 54 SNVs were tested on drug-metabolizing enzymes and 12 of them were disruptive. In addition, 227 SNVs were tested on drug targets, 66 of which disrupted at least one interaction. Lastly, five SNVs were tested on drug transporters and three of them were disruptive.

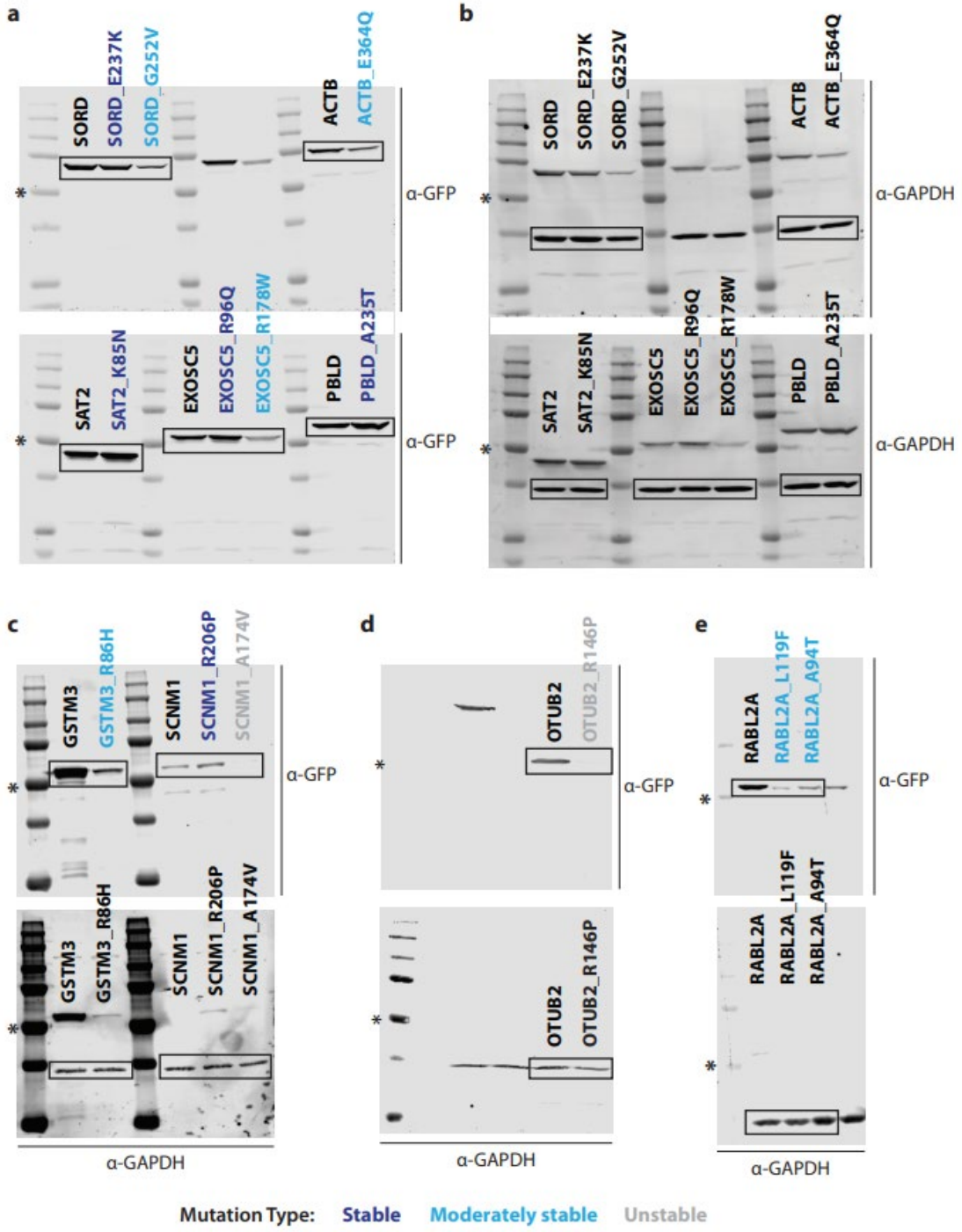


Figure 23. Uncropped Western blots for stable, moderately stable, and unstable GFP expression examples in Figure 16a.

a, Westerns for wild-type and corresponding mutant proteins detected by α-GFP. **b,** α-GAPDH controls for westerns for wild-type and corresponding mutant proteins detected in **a**. **c,** Upper: Westerns for wild-type and corresponding mutant proteins detected by α-GFP. Lower: α-GAPDH controls for western in upper. **d,** Upper: Westerns for wild-type and corresponding mutant

proteins detected by α-GFP. 50 kDa marker assigned using ladder from lower. Lower: α-GAPDH controls for western in upper. (e) Upper: Westerns for wild-type and corresponding mutant proteins detected by α-GFP. Lower: α-GAPDH controls for western in upper. In **a-e**, stable, partially stable, and unstable mutations are labeled in blue, cyan, and gray, respectively; all α-GAPDH controls were detected using stripped membranes. Bands unrelated to this project are not boxed and were not used in any analyses. Bands corresponding to α-GFP and α-GAPDH examples used in **Figure 16a** are enclosed in black boxes. * indicates 50 kDa marker.

Generation of Mice Using CRISPR-Cas9 Genome Editing

A brief diagram of the CRISPR/Cas9 genome editing strategy and mice genotyping is provided in **Figure 23**. Optimal guide sequences were selected using online software at mit.crispr.edu. To generate the sgRNA, we used a previously published PCR overlap method[214-217]. Briefly, overlapping PCR primers, together encoding the T7 promoter, 20-nucleotide guide sequence, and RNA secondary structure sequence, were ordered from IDT. The DNA template was reverse-transcribed using Ambion MEGAshortscript T7 Transcription Kit (cat#AM1354) and resulting sgRNA was purified using Qiagen MinElute columns (cat#28004). For pronuclear injection, the sgRNA (50 ng/μL), ssODN (50 ng/μL, IDT Ultramer Service), and Cas9 mRNA (25 ng/μL, TriLink) were co-injected into zygotes (F1 hybrids between strains FVB/NJ and B6(Cg)-*Tyr*$^{c-2J}$/J) then transferred into the oviduct of pseudopregnant females. Founders carrying at least one copy of the desired alteration were identified and backcrossed into FVB/NJ. Initial phenotyping was done after one backcross generation and additional phenotyping was done with mice backcrossed at least two or more generations.

Genotyping Sept12 mice

For genotyping, we collected toes from 8-14 day old mice and created a crude DNA lysate as previously described[218]. PCR, using the following two primers: 5'-GAGATGGGATGACAGGACTATTG-3' and 5'-GTGGATGAGTGAGGGAAGAAAG-3', was performed using EconoTaq and associated PCR reagents (Lucigen) with 3μL of crude DNA lysate. The PCR cycle used

for *Sept12*^{G169E} was: 95°C for 5min, 30 cycles of 95°C for 30sec, 64°C for 30sec, 72°C for 30sec, and final elongation at 72°C for 5min. To distinguish between WT and G169E, PCR amplicons were digested by *Msc*I to yield WT fragments of 180 and 138bp, whereas the G169E allele remains uncut.

Fertility Test

Wild-type, heterozygous, and homozygous males and females were bred to wild-type counterparts starting at 2 months until 7-15 months of age. The litter size and sex of pups were recorded.

Sperm Motility

Both epididymides were harvested from adult males, washed in PBS, and placed in a puddle of *in vitro* fertilization media (Cook Medical). A slit was cut along each epididymis and sperm were allowed to swim out for 2min at 37°C. Next, 10 μL of sperm was moved to a glass slide for motility assessment.

REFERENCES

3 The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56-65, doi:http://www.nature.com/nature/journal/v491/n7422/abs/nature11632.html#supplementary-information (2012).

23 Sahni, N. *et al.* Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* 161, 647-660, doi:10.1016/j.cell.2015.04.013 (2015).

24 Wei, X. *et al.* A Massively Parallel Pipeline to Clone DNA Variants and Examine Molecular Phenotypes of Human Disease Mutations. *PLOS Genetics* 10, e1004819, doi:10.1371/journal.pgen.1004819 (2014).

25 Zhong, Q. *et al.* Edgetic perturbation models of human inherited disorders. *Molecular Systems Biology* 5, doi:10.1038/msb.2009.80 (2009).

27 Fuxman Bass, J. I. *et al.* Human gene-centered transcription factor networks for enhancers and disease variants. *Cell* 161, 661-673, doi:10.1016/j.cell.2015.03.003 (2015).

30 Pejaver, V. *et al.* MutPred2: inferring the molecular and phenotypic impact of amino acid variants. *bioRxiv* (2017).

32 Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337, 64-69, doi:10.1126/science.1219240 (2012).

33 Miosge, L. A. *et al.* Comparison of predicted and actual consequences of missense mutations. *Proceedings of the National Academy of Sciences of the United States of America* 112, E5189-E5198, doi:10.1073/pnas.1511585112 (2015).

34 Wang, T. *et al.* Probability of phenotypically detectable protein damage by ENU-induced mutations in the Mutagenetix database. *Nature Communications* 9, 441, doi:10.1038/s41467-017-02806-4 (2018).

35 Wang, X. *et al.* Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature Biotechnology* 30, 159-164, doi:http://www.nature.com/nbt/journal/v30/n2/abs/nbt.2106.html#supplementary-information (2012).

36 Meyer, M. J. *et al.* Interactome INSIDER: a structural interactome browser for genomic studies. *Nature Methods* 15, 107, doi:10.1038/nmeth.4540 https://www.nature.com/articles/nmeth.4540#supplementary-information (2018).

48 del-Toro, N. *et al.* Capturing variation impact on molecular interactions: the IMEx Consortium mutations data set. *bioRxiv*, doi:10.1101/346833 (2018).

50 Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nature Methods* 11, 801, doi:10.1038/nmeth.3027 (2014).

52 Starita, L. M. *et al.* Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* 200, 413-422, doi:10.1534/genetics.115.175802 (2015).

54 Fragoza, R. *et al.* Extensive disruption of protein interactions by genetic variants across the allele frequency spectrum in human populations. *Nat Commun* 10,

4141, doi:10.1038/s41467-019-11959-3 (2019).
136 Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336, 740-743, doi:10.1126/science.1217283 (2012).
137 Gazave, E., Chang, D., Clark, A. G. & Keinan, A. Population growth inflates the per-individual number of deleterious mutations and reduces their mean effect. *Genetics* 195, 969-978, doi:10.1534/genetics.113.153973 (2013).
138 Nelson, M. R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337, 100-104, doi:10.1126/science.1217876 (2012).
139 Coventry, A. *et al.* Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications* 1, 131, doi:10.1038/ncomms1130
https://www.nature.com/articles/ncomms1130#supplementary-information (2010).
140 Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493, 216-220, doi:http://www.nature.com/nature/journal/v493/n7431/abs/nature11690.html#supplementary-information (2013).
141 The UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* 526, 82-90, doi:10.1038/nature14962
http://www.nature.com/nature/journal/v526/n7571/abs/nature14962.html#supplementary-information (2015).
142 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285-291, doi:10.1038/nature19057
http://www.nature.com/nature/journal/v536/n7616/abs/nature19057.html#supplementary-information (2016).
143 The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics* 46, 818-825, doi:10.1038/ng.3021
http://www.nature.com/ng/journal/v46/n8/abs/ng.3021.html#supplementary-information (2014).
144 Henn, B. M., Botigué, L. R., Bustamante, C. D., Clark, A. G. & Gravel, S. Estimating Mutation Load in Human Genomes. *Nature Reviews Genetics* 16, 333-343, doi:10.1038/nrg3931 (2015).
145 Vidal, M. A biological atlas of functional maps. *Cell* 104, 333-339, doi:10.1016/S0092-8674(01)00221-5 (2001).
146 Vidal, M., Cusick, Michael E. & Barabási, A.-L. Interactome networks and human disease. *Cell* 144, 986-998 (2011).
147 Goldman, E. R., Dall'Acqua, W., Braden, B. C. & Mariuzza, R. A. Analysis of binding interactions in an idiotope-antiidiotope protein-protein complex by double mutant cycles. *Biochemistry* 36, 49-56, doi:10.1021/bi961769k (1997).
148 Radisky, E. S., Kwan, G., Karen Lu, C. J. & Koshland, D. E., Jr. Binding, proteolytic, and crystallographic analyses of mutations at the protease-inhibitor interface of the subtilisin BPN'/chymotrypsin inhibitor 2 complex. *Biochemistry* 43, 13648-13656, doi:10.1021/bi048797k (2004).

149 Keeble, A. H. *et al.* Experimental and computational analyses of the energetic basis for dual recognition of immunity proteins by colicin endonucleases. *Journal of Molecular Biology* 379, 745-759, doi:10.1016/j.jmb.2008.03.055 (2008).
150 Moal, I. H. & Fernández-Recio, J. SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics* 28, 2600-2607, doi:10.1093/bioinformatics/bts489 (2012).
151 Manolio, T. A., Brooks, L. D. & Collins, F. S. A HapMap harvest of insights into the genetics of common disease. *The Journal of Clinical Investigation* 118, 1590-1605, doi:10.1172/JCI34772 (2008).
152 Gibson, G. Rare and Common Variants: Twenty arguments. *Nature Reviews Genetics* 13, 135-145, doi:10.1038/nrg3118 (2011).
153 Chow, C. Y. Bringing genetic background into focus. *Nature Reviews Genetics* 17, 63, doi:10.1038/nrg.2015.9 (2015).
154 Schoenrock, A. *et al.* Evolution of protein-protein interaction networks in yeast. *PLOS ONE* 12, e0171920, doi:10.1371/journal.pone.0171920 (2017).
155 Khurana, E. *et al.* Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. *Science* 342, doi:10.1126/science.1235587 (2013).
156 Guharoy, M. & Chakrabarti, P. Conservation and relative importance of residues across protein-protein interfaces. *Proceedings of the National Academy of Sciences of the United States of America* 102, 15447-15452, doi:10.1073/pnas.0505425102 (2005).
157 Mintseris, J. & Weng, Z. Structure, function, and evolution of transient and obligate protein–protein interactions. *Proceedings of the National Academy of Sciences of the United States of America* 102, 10930-10935, doi:10.1073/pnas.0502667102 (2005).
158 Maher, M. C., Uricchio, L. H., Torgerson, D. G. & Hernandez, R. D. Population genetics of rare variants and complex diseases. *Human Heredity* 74, 118-128, doi:10.1159/000346826 (2012).
159 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* 7, 248-249, doi:http://www.nature.com/nmeth/journal/v7/n4/suppinfo/nmeth0410-248_S1.html (2010).
160 Gorlov, I. P., Gorlova, O. Y., Sunyaev, S. R., Spitz, M. R. & Amos, C. I. Shifting Paradigm of Association Studies: Value of Rare Single-Nucleotide Polymorphisms. *American Journal of Human Genetics* 82, 100-112, doi:10.1016/j.ajhg.2007.09.006 (2008).
161 Strittmatter, W. J. *et al.* Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proceedings of the National Academy of Sciences of the United States of America* 90, 1977-1981 (1993).
162 Deary, I. J. *et al.* Cognitive change and the APOE ε4 allele. *Nature* 418, 932, doi:10.1038/418932a (2002).
163 Corder, E. *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of

Alzheimer's disease in late onset families. *Science* 261, 921-923, doi:10.1126/science.8346443 (1993).

164 Robitaille, J., Després, J. P., Pérusse, L. & Vohl, M. C. The PPAR-gamma P12A polymorphism modulates the relationship between dietary fat intake and components of the metabolic syndrome: results from the Québec Family Study. *Clinical Genetics* 63, 109-116, doi:doi:10.1034/j.1399-0004.2003.00026.x (2003).

165 Florez, J. C. *et al.* Effects of the type 2 diabetes-associated PPARG P12A polymorphism on progression to diabetes and response to troglitazone. *The Journal of Clinical Endocrinology and Metabolism* 92, 1502-1509, doi:10.1210/jc.2006-2275 (2007).

166 Stenson, P. D. *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics* 136, 665-677, doi:10.1007/s00439-017-1779-6 (2017).

167 Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research* 39, D945-D950, doi:10.1093/nar/gkq929 (2011).

168 Das, J. *et al.* Exploring mechanisms of human disease through structurally resolved protein interactome networks. *Molecular BioSystems* 10, 9-17, doi:10.1039/C3MB70225A (2014).

169 Braun, P. *et al.* An experimentally derived confidence score for binary protein-protein interactions. *Nature Methods* 6, 91-97, doi:http://www.nature.com/nmeth/journal/v6/n1/suppinfo/nmeth.1281_S1.html (2009).

170 Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nature Methods* 6, 83-90, doi:http://www.nature.com/nmeth/journal/v6/n1/suppinfo/nmeth.1280_S1.html (2009).

171 Fu, W., O'Connor, T. D. & Akey, J. M. Genetic architecture of quantitative traits and complex diseases. *Current Opinion in Genetics & Development* 23, 678-683, doi:https://doi.org/10.1016/j.gde.2013.10.008 (2013).

172 Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* 185, 862 (1974).

173 Lyon, R. C., Johnston, S. M., Watson, D. G., McGarvie, G. & Ellis, E. M. Synthesis and catabolism of γ-hydroxybutyrate in SH-SY5Y human neuroblastoma cells: role of the aldo-keto reductase AKR7A2. *Journal of Biological Chemistry* 282, 25986-25992, doi:10.1074/jbc.M702465200 (2007).

174 Bains, O. S., Grigliatti, T. A., Reid, R. E. & Riggs, K. W. Naturally occurring variants of human aldo-keto reductases with reduced *in vitro* metabolism of daunorubicin and doxorubicin. *Journal of Pharmacology and Experimental Therapeutics* 335, 533 (2010).

175 Medina-Kauwe, L. K., Nyhan, W. L., Gibson, K. M. & Tobin, A. J. Identification of a familial mutation associated with GABA-transaminase deficiency disease. *Neurobiology of Disease* 5, 89-96,

doi:https://doi.org/10.1006/nbdi.1998.0184 (1998).
176 Tsuji, M. *et al.* A new case of GABA transaminase deficiency facilitated by proton MR spectroscopy. *Journal of Inherited Metabolic Disease* 33, 85-90, doi:10.1007/s10545-009-9022-9 (2010).
177 Akaboshi, S. *et al.* Mutational spectrum of the succinate semialdehyde dehydrogenase (ALDH5A1) gene and functional analysis of 27 novel disease-causing mutations in patients with SSADH deficiency. *Human Mutation* 22, 442-450, doi:10.1002/humu.10288 (2003).
178 Ouedraogo, M. *et al.* The Duplicated Genes Database: Identification and Functional Annotation of Co-Localised Duplicated Genes across Genomes. *PLOS ONE* 7, e50653, doi:10.1371/journal.pone.0050653 (2012).
179 Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research* 20, 110-121, doi:10.1101/gr.097857.109 (2010).
180 The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 526, 68, doi:10.1038/nature15393 https://www.nature.com/articles/nature15393#supplementary-information (2015).
181 Gallione, C. *et al.* Overlapping spectra of SMAD4 mutations in juvenile polyposis (JP) and JP–HHT syndrome. *American Journal of Medical Genetics Part A* 152A, 333-339, doi:10.1002/ajmg.a.33206 (2010).
182 Sayed, M. G. *et al.* GermlineSMAD4 orBMPRIA mutations and phenotype of juvenile polyposis. *Annals of Surgical Oncology* 9, 901-906, doi:10.1007/bf02557528 (2002).
183 Nasim, M. T. *et al.* Molecular genetic characterization of SMAD signaling molecules in pulmonary arterial hypertension. *Human Mutation* 32, 1385-1389, doi:doi:10.1002/humu.21605 (2011).
184 Jung, B., Staudacher, J. J. & Beauchamp, D. Transforming Growth Factor β Superfamily Signaling in Development of Colorectal Cancer. *Gastroenterology* 152, 36-52, doi:10.1053/j.gastro.2016.10.015 (2017).
185 Massagué, J. TGFβ in Cancer. *Cell* 134, 215-230, doi:https://doi.org/10.1016/j.cell.2008.07.001 (2008).
186 Veiga-da-Cunha, M. *et al.* Mutations responsible for 3-phosphoserine phosphatase deficiency. *European Journal of Human Genetics* 12, 163-166 (2003).
187 Kim, H.-Y. *et al.* Molecular Basis for the Local Conformational Rearrangement of Human Phosphoserine Phosphatase. *Journal of Biological Chemistry* 277, 46651-46658, doi:10.1074/jbc.M204866200 (2002).
188 Kuo, Y.-C. *et al.* SEPT12 mutations cause male infertility with defective sperm annulus. *Human Mutation* 33, 710-719, doi:10.1002/humu.22028 (2012).
189 Mostowy, S. & Cossart, P. Septins: the fourth component of the cytoskeleton. *Nature Reviews Molecular Cell Biology* 13, 183, doi:10.1038/nrm3284 https://www.nature.com/articles/nrm3284#supplementary-information (2012).
190 Sellin, M. E., Stenmark, S. & Gullberg, M. Cell type–specific expression of SEPT3-homology subgroup members controls the subunit number of heteromeric septin complexes. *Molecular Biology of the Cell* 25, 1594-1607,

doi:10.1091/mbc.e13-09-0553 (2014).
191 Kuo, Y.-C. *et al.* SEPT12 orchestrates the formation of mammalian sperm annulus by organizing core octameric complexes with other SEPT proteins. *Journal of Cell Science* 128, 923-934, doi:10.1242/jcs.158998 (2015).
192 Singh, P. & Schimenti, J. C. The genetics of human infertility by functional interrogation of SNPs in mice. *Proceedings of the National Academy of Sciences* 112, 10431-10436, doi:10.1073/pnas.1506974112 (2015).
193 Dorfman, R. *et al.* Do common in silico tools predict the clinical consequences of amino-acid substitutions in the CFTR gene? *Clinical Genetics* 77, 464-473, doi:10.1111/j.1399-0004.2009.01351.x (2010).
194 Masica, D. L. *et al.* Predicting survival in head and neck squamous cell carcinoma from TP53 mutation. *Human Genetics* 134, 497-507, doi:10.1007/s00439-014-1470-0 (2015).
195 Cassa, C. A., Tong, M. Y. & Jordan, D. M. Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. *Human Mutation* 34, 1216-1220, doi:10.1002/humu.22375 (2013).
196 Visscher, P. M. & Goddard, M. E. From R.A. Fisher's 1918 Paper to GWAS a Century Later. *Genetics* 211, 1125-1130, doi:10.1534/genetics.118.301594 (2019).
197 Gajiwala, K. S. *et al.* KIT kinase mutants show unique mechanisms of drug resistance to imatinib and sunitinib in gastrointestinal stromal tumor patients. *Proceedings of the National Academy of Sciences* 106, 1542-1547, doi:10.1073/pnas.0812413106 (2009).
198 Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nature Methods* 12, 203, doi:10.1038/nmeth.3223
https://www.nature.com/articles/nmeth.3223#supplementary-information (2015).
199 Woodsmith, J. *et al.* Protein interaction perturbation profiling at amino-acid resolution. *Nature Methods* 14, 1213, doi:10.1038/nmeth.4464
https://www.nature.com/articles/nmeth.4464#supplementary-information (2017).
200 Findlay, G. M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* 562, 217-222, doi:10.1038/s41586-018-0461-z (2018).
201 Julien, P., Miñana, B., Baeza-Centurion, P., Valcárcel, J. & Lehner, B. The complete local genotype–phenotype landscape for the alternative splicing of a human exon. *Nature Communications* 7, 11558, doi:10.1038/ncomms11558
https://www.nature.com/articles/ncomms11558#supplementary-information (2016).
202 Yang, X. *et al.* A public genome-scale lentiviral expression library of human ORFs. *Nature Methods* 8, 659-661, doi:10.1038/nmeth.1638 (2011).
203 The MGC Project Team. The completion of the Mammalian Gene Collection (MGC). *Genome Research* 19, 2324-2333, doi:10.1101/gr.095976.109 (2009).
204 Rual, J.-F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437, 1173-1178, doi:http://www.nature.com/nature/journal/v437/n7062/suppinfo/nature04209_S1.html (2005).
205 Yu, H. *et al.* Next-generation sequencing to generate interactome datasets.

Nature Methods 8, 478-480, doi:http://www.nature.com/nmeth/journal/v8/n6/abs/nmeth.1597.html#supplementary-information (2011).
206 Rolland, T. *et al.* A proteome-scale map of the human interactome network. *Cell* 159, 1212-1226, doi:10.1016/j.cell.2014.10.050 (2014).
207 Das, J. *et al.* Elucidating common structural features of human pathogenic variations using large-scale atomic-resolution protein networks. *Human Mutation* 35, 585-593, doi:doi:10.1002/humu.22534 (2014).
208 Meyer, M. J., Das, J., Wang, X. & Yu, H. INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics*, doi:10.1093/bioinformatics/btt181 (2013).
209 Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nature Genetics* 50, 1593-1599, doi:10.1038/s41588-018-0248-z (2018).
210 Arnold, M., Raffler, J., Pfeufer, A., Suhre, K. & Kastenmüller, G. SNiPA: an interactive, genetic variant-centered annotation browser. *Bioinformatics (Oxford, England)* 31, 1334-1336, doi:10.1093/bioinformatics/btu779 (2015).
211 MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic acids research* 45, D896-D901, doi:10.1093/nar/gkw1133 (2017).
212 Romero, P. *et al.* Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology* 6, R2, doi:10.1186/gb-2004-6-1-r2 (2004).
213 Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research* 46, D1074-D1082, doi:10.1093/nar/gkx1037 (2018).
214 Varshney, G. K. *et al.* High-throughput gene targeting and phenotyping in zebrafish using CRISPR/Cas9. *Genome Research* 25, 1030-1042, doi:10.1101/gr.186379.114 (2015).
215 Singh, P., Schimenti, J. C. & Bolcun-Filas, E. A Mouse Geneticist's Practical Guide to CRISPR Applications. *Genetics* 199, 1-15, doi:10.1534/genetics.114.169771 (2015).
216 Gagnon, J. A. *et al.* Efficient Mutagenesis by Cas9 Protein-Mediated Oligonucleotide Insertion and Large-Scale Assessment of Single-Guide RNAs. *PLOS ONE* 9, e98186, doi:10.1371/journal.pone.0098186 (2014).
217 Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* 8, 2281, doi:10.1038/nprot.2013.143
https://www.nature.com/articles/nprot.2013.143#supplementary-information (2013).
218 Truett, G. E. *et al.* Preparation of PCR-Quality Mouse Genomic DNA with Hot Sodium Hydroxide and Tris (HotSHOT). *BioTechniques* 29, 52-54, doi:10.2144/00291bm09 (2000).

CHAPTER 4:

A 3D STRUCTURAL SARS-COV-2–HUMAN INTERACTOME TO EXPLORE GENETIC AND DRUG PERTURBATIONS

Context and Personal Contributions

The following chapter is derived from my first author Nature Methods paper by the same name[219] (Reproduced with permission from Springer Nature). The project was originally conceived of through discussions between myself and Dr. Haiyuan Yu, and I went on to independently develop these discussions into the final manuscript. In particular, through this project I initiated new computational efforts in the Yu lab to transition from sequence-level prediction and annotation of protein-protein interaction interfaces to generation of full 3D docked protein-protein interaction models informed by biophysical, sequence conservation, and structural features. This transition enabled direct biophysical analysis to computationally predict the effects of mutations on the binding affinities of protein-protein interactions, never before explored in the Yu lab.

I was responsible for the original design, writing, analyses and generation of figures for this manuscript. Additional contributions from collaborators were as follows. Siqi (Charles) Liang was responsible for the initial implementation of the online 3D-SARS2 interactome browser, maintenance of which was later handled by Shagun Gupta. You Chen completed the analysis of enrichment of GWAS reported phenotypic variants within human genes and their protein interfaces interacting with SARS-CoV-2. Yuan Liu completed all Y2H, IP-MS, and mutagenesis for the experimental validations of our predictions. Nicole Andre and Drs. Steven Lipkin and Gary Whittaker provided expert perspectives for virology that were essential for framing the introduction to this

manuscript and for interpreting and contextualizing our results.

Abstract

Emergence of new viral agents is driven by evolution of interactions between viral proteins and host targets. For instance, increased infectivity of SARS-CoV-2 compared to SARS-CoV-1 arose in part through rapid evolution along the interface between the Spike protein and its human receptor ACE2, leading to increased binding affinity. To facilitate broader exploration of how pathogen-host interactions might impact transmission and virulence in the ongoing COVID-19 pandemic, we performed state-of-the-art interface prediction followed by molecular docking to construct a 3D structural interactome between SARS-CoV-2 and human. We additionally carried out downstream meta-analyses to investigate enrichment of sequence divergence between SARS-CoV-1 and SARS-CoV-2 or human population variants along viral-human protein interaction interfaces, predict changes in binding affinity by these mutations/variants, and further prioritize drug repurposing candidates predicted to competitively bind human targets. We believe this resource (http://3D-SARS2.yulab.org) will aid in development and testing of informed hypotheses for SARS-CoV-2 etiology and treatments.

Introduction

The ongoing global COVID-19 pandemic has resulted in over 210 million SARS-CoV-2 infections and over 4.4 million deaths worldwide[220]. The coronavirus family of enveloped viruses causes respiratory and enteric tract infections in avian and mammalian hosts[221]. Seven well characterized human coronaviruses[222-224] exhibit

symptoms ranging from mild respiratory illness to severe pneumonia and acute respiratory distress syndrome (ARDS). These coronaviruses are either highly transmissible yet generally not highly pathogenic (e.g. HCoV-229E, HCoV-OC43) or highly pathogenic but poorly transmissible (SARS-CoV-1 and MERS-CoV). Unique from these, SARS-CoV-2 is both highly transmissible and capable of causing severe disease with infectivity and pathogenesis differing between individuals[225,226]. While ~25-35% of infected individuals experience only mild or minimal symptoms, ~1-2% of infected patients die primarily from severe respiratory failure and ARDS[227,228]. Differences in morbidity, hospitalization, and mortality among different ethnic groups[229-234] are not fully explained by cardiometabolic, socioeconomic, or behavioral factors, suggesting a role for human genetic variation in SARS-CoV-2 pathogenicity. Insights into the evolution of SARS-CoV-2, its elevated transmission relative to SARS-CoV-1, and dynamic range of symptoms have been key areas of interest. These traits are likely driven by molecular mechanisms of pathology including interactions between the virus and its host, but specific causes are yet to be fully characterized.

Networks of protein-protein interactions between pathogens and their hosts provide one avenue to understand mechanisms of infection and pathology. Viral-human interactome maps have been compiled for SARS-CoV-1[235], HIV[236], Ebola virus[237], and Dengue and Zika viruses[238] among others. Recent, affinity-purification mass-spectrometry experiments on 29 SARS-CoV-2 proteins identified 332 viral-human interactions[239]. Inter-species interactions contribute to disease progression by facilitating pathogen entry into host cells[240-245], inhibiting host response proteins and pathways[246-248], and hijacking cell signaling or metabolism to accelerate cellular—and

consequentially viral—replication[249-251]. Structures and dynamics of these interactions can provide insights into their roles. For instance, the viral-human binding interface between poxvirus chemokine inhibitor vCCI and human MIP-1β is shown to occlude domains vital to chemokine homodimerization, receptor binding, and interactions with GAG, thus explaining poxvirus' inhibitory effect on chemokine signaling[248]. Additionally, the dynamics of a herpesvirus cyclin and human cdk2 interaction induce a conformational change on cdk2 that matches its interaction with human cyclin A leading to dysregulated cell cycle progression[250].

Because protein-protein interactions mediate the majority of protein function[74-76], targeted disruption by small molecule inhibitors that compete for the same binding site provide a precise toolkit to modulate cellular function[74,76,252-254]. For instance, BCL-2 inhibitors that displace bound anti-apoptotic BCL-X interactors can treat chronic lymphocytic leukemia pathogenesis[255,256]. This approach can be particularly effective in viral networks and several potent inhibitors of key interactions have been developed. Disruption of viral complexes involved in viral replication has been successful in vaccinia virus[257] and human papilloma virus therapies[258,259]. Specifically, disruption of viral-host protein-protein interactions involved in early viral infection is an important therapeutic strategy. Discovery that a population variant in the membrane protein CCR5 conferred resistance to HIV-1 by disrupting its interaction with the viral envelope glycoprotein led to the development of Maraviroc as an FDA approved treatment for HIV-1 that functions by blocking the interface for this interaction[242,260].

Here we apply a full-interactome modeling framework to construct a 3D structural interactome between SARS-CoV-2 and human proteins. Our framework first

applies our previous ECLAIR framework[261] to identify interface residues for the whole SARS-CoV-2-human interactome and leverages these predictions to guide atomic-resolution interface modeling and docking in HADDOCK[262,263]. We additionally carried out in-silico scanning mutagenesis in PyRosetta[264] to predict the impact of mutations on interaction binding affinity and explored the overlap between protein-protein and protein-drug binding sites. All results from our 3D structural interactome are provided as a user-friendly web server allowing exploration of individual interactions or bulk download and analysis of the whole dataset. We further explore the utility of our 3D interactome modeling approach in identifying key interactions undergoing evolution along viral protein interfaces, highlighting population variants on human interfaces that could modulate the strength of viral-host interactions to confer protection from or susceptibility to COVID-19, and prioritizing drug candidates predicted to bind competitively at viral-human interaction interfaces, some of which could potentially be used for therapeutic purposes. Cumulatively these predictions and analyses are intended as a resource to facilitate investigation and further characterization of SARS-CoV-2-human interactions.

Results

Enrichment of variation on the spike-ACE2 binding interface

We highlight the utility of computational and structural approaches to model the SARS-CoV-2-human interactome, from the interaction between the SARS-CoV-2 spike protein (S) and human angiotensin-converting enzyme 2 (ACE2) (**Figure 24a**). This interaction mediates viral entry into human cells[222] and is among the only viral-human

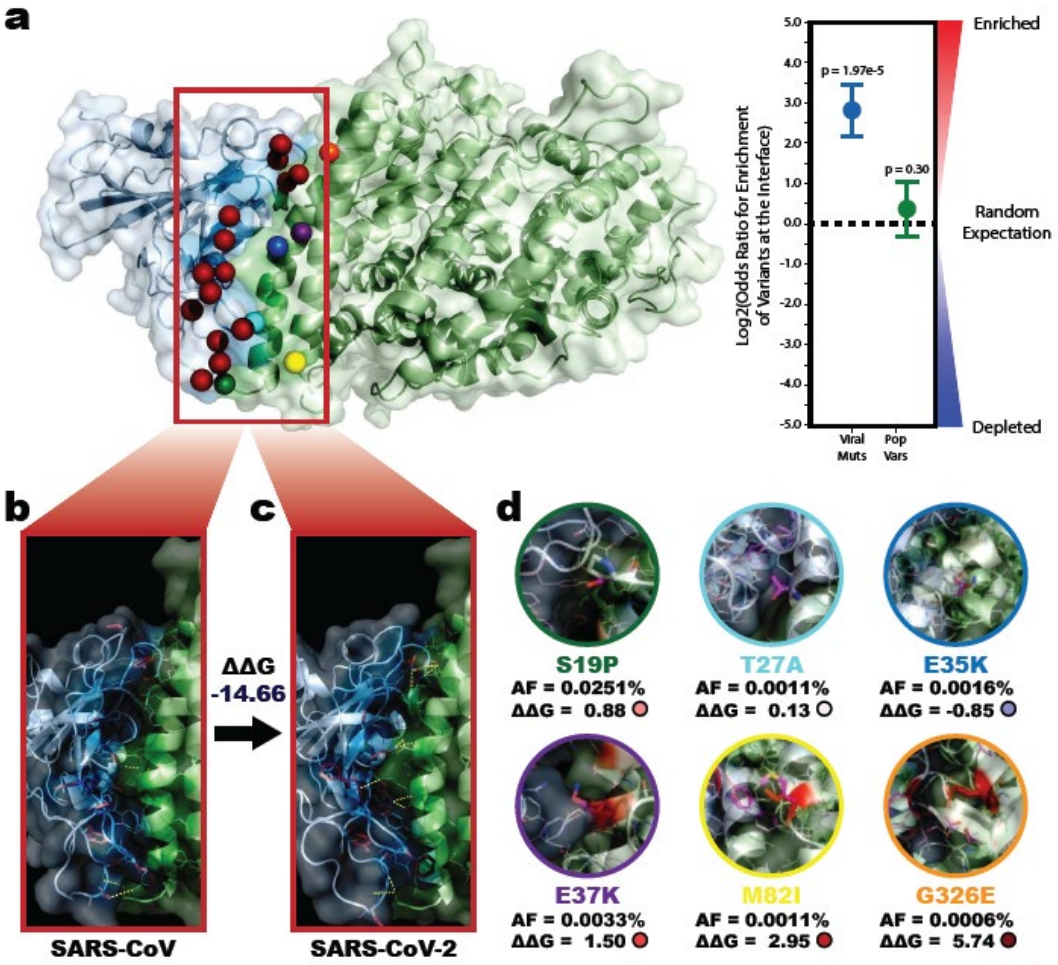


Figure 24. Enrichment and predicted impact of divergences between SARS-CoV-1 and SARS-CoV-2 along the S-ACE2 interface.
a, Co-crystal structure of the interaction between SARS-CoV-2 Spike protein (S) with human ACE2 (PDB 6LZG). All 15 sequence divergences between SARS-CoV-1 and SARS-CoV-2 Spike protein interfaces are highlighted as red spheres while all 6 population variants on the ACE2 protein interface are highlighted as green (ACE2_S19P), cyan (ACE2_T27A), blue (ACE2_E35K), purple (ACE2_E37K), yellow (ACE2_M82I), and orange (ACE2_G326E) spheres. Enrichment of these variants on the interface are reported for SARS-CoV-2 (Log2OR=2.82, p=1.97e-5 by two-sided z-test) and human (Log2OR=0.38, p=0.30 by two-sided z-test) shown to the right. Data presented as Log2OR ± SE. **b, c,** Expanded interface views for the SARS-CoV-1 S-ACE2 structure (PDB 6CS2) and SARS-CoV-2 S-ACE2 structure (PDB 6LZG). Sequence divergences are highlighted as red sticks. Inter-protein polar contacts that contribute to stabilizing the interaction are shown as yellow dashed lines. The negative predicted change in binding affinity (ΔΔG=-14.66 Rosetta Energy Units (REU)) indicates the interaction is more stable (lower energy) in the SARS-CoV-2 version of the interaction. **d,** Predicted impact each ACE2 population variant. Mutated structures superimposed over the wildtype structure (magenta). The mutated residue is shown as sticks. Residues contributing to the overall change in binding energy are colored from blue (decreased ΔΔG) to white (no change) to red (increased ΔΔG). The gnomAD reported allele frequency and predicted ΔΔG for each mutation are reported.

interactions solved in both SARS-CoV-1[265] and SARS-CoV-2[266-268]. Recent sequence divergences of the S protein are highly enriched at the S-ACE2 interaction interface (**Figure 24a**; Log2OddsRatio=2.82, p=1.97e-5), indicating functional evolution around this interaction. We predicted the impact of these mutations on the binding affinity (ΔΔG) between the SARS-CoV-1 and SARS-CoV-2 versions of the S-ACE2 interaction using the Rosetta energy function[269] (**Figure 24b and c**). The negative ΔΔG value of -14.66 Rosetta Energy Units (REU) indicates an increased binding affinity using the SARS-CoV-2 S protein driven by better optimized solvation and hydrogen bonding potential fulfillment. Our result is consistent with the hypothesis that increased stability of the S-ACE2 interaction contributes to the elevated transmission of SARS-CoV-2[270]. Experimental kinetics assays have confirmed that compared to SARS-CoV-1, SARS-CoV-2 S protein binds ACE2 with 10-20-fold higher affinity[271] supporting the conclusions from our computational modeling.

A wide range in severity of and susceptibility to SARS-CoV-2 exists between individuals[225,226,272]. Genetic predisposition hypotheses explaining this range include both expression regulating and protein-coding variants [273,274]. For instance, an RNA-sequencing analysis suggested higher expression of ACE2 in Asian males could facilitate viral entry and explain increased susceptibility among this population[275]. Alternatively, missense population variants in ACE2 could strengthen or weaken the S-ACE2 interaction, thereby modulating susceptibility to infection. We used a mutation scanning pipeline in PyRosetta[276,277] to predict the impact of six missense variants reported in gnomAD[278] that occur on the S-ACE2 interface (**Figure 24d**). The three variants with the largest predicted impact on S-ACE2 binding affinity—ACE2_E37K

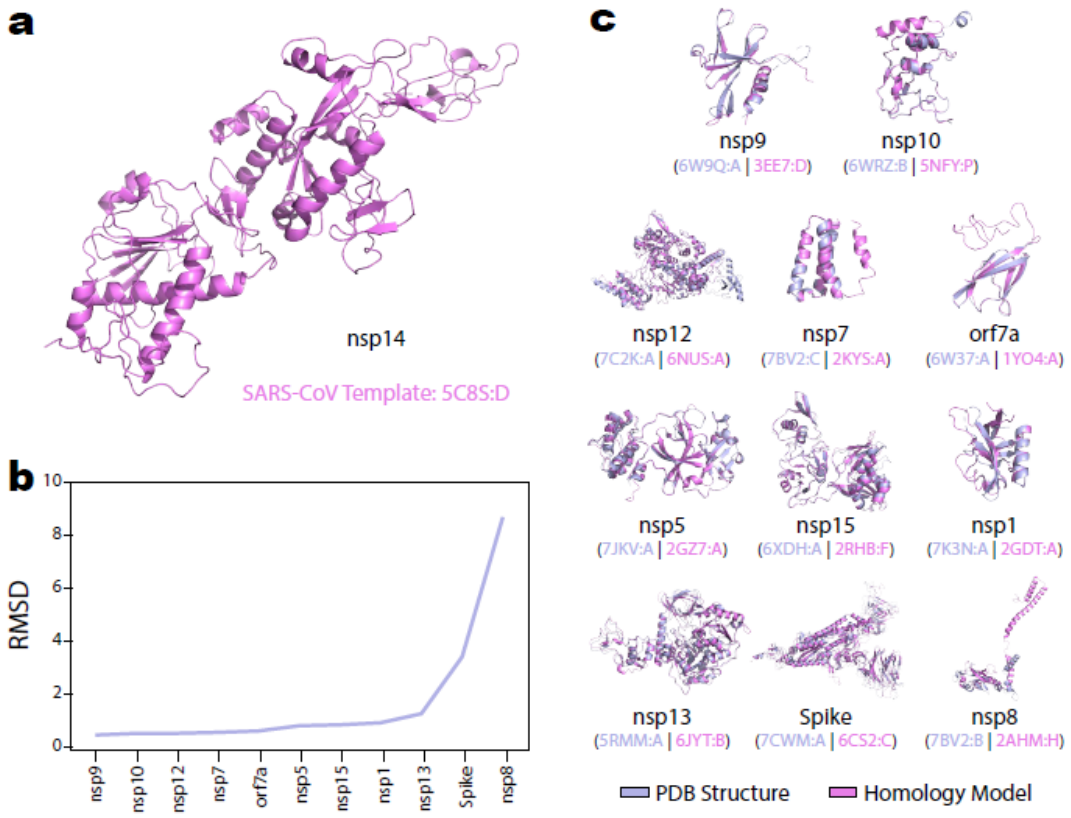


Figure 25. Homology modeling for SARS-CoV-2 proteins
a, Homology models for SARS-CoV-2 nsp14 modeled from a high-quality template for from SARS-CoV-1 nsp14 (PDB 5C8S:D). The nsp14 homology model was retained and used in downstream computational predictions. **b,** Quality assessment on 11 SARS-CoV-2 models generated using the same method as the nsp14 model. For these 11 proteins solved crystal-structures for the SARS-CoV-2 protein were deposited into the PDB during submission and revision of this manuscript and validated the quality of the homology modelling. Assessment is based on the on root-mean-square deviation (RMSD) following alignment of the homology model and PDB structure using PyMol. **c,** Visual representation of the alignment between all homology models (magenta) against their available PDB structure (light blue). PDB IDs and chains used for both the homology template and the reference PDB structure are indicated.

(ΔΔG=1.50), ACE2_M82I (ΔΔG=2.95), and ACE2_G326E (ΔΔG=5.74)—were consistent with previous experimental screens identifying them as putative protective variants exhibiting decreased binding of ACE2 to S[279,280]. Our results highlight utility for a 3D structural interactome modeling approach in identifying interactions and mutations important for viral infection, pathogenesis, and transmission.

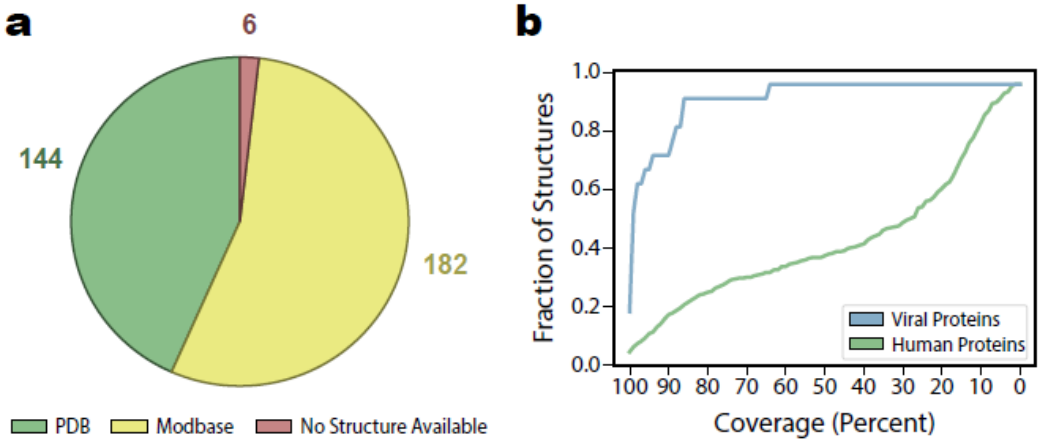


Figure 26. Source and coverage of available protein structures
a, Breakdown of the source of all structures available for the 332 human interactors of SARS-CoV-2 as being either a experimentally solved structure from the Protein Data Bank (n=144) a homology model from Modbase (n=182), or no available structure (n=6). **b,** Analysis of the coverage of all available structures for both human (green) and viral (blue) proteins. The fraction of structures retained with coverage greater than or equal to a range of coverage thresholds is shown. For our purposes, all available structures were used for solvent accessibility feature calculations for ECLAIR predictions, but structures were only retained for docking if either 1) total coverage was at least 33% of 2) the structure covered at least one high-confidence interface prediction from ECLAIR.

Constructing the 3D structural SARS-CoV-2-human interactome

To facilitate similar investigation and hypothesis development at the full interactome scale, we next compiled a comprehensive 3D structural interactome between SARS-CoV-2 and human proteins based on 332 viral-human interactions uncovered in an early interactome screen by Gordon *et al.*[239]. First, we modeled SARS-CoV-2 proteins supplementing solved structures from the Protein Data Bank (PDB)[94] (16 of 29 proteins) with homology derived from SARS-CoV-1 templates (12 of 29 proteins). Homology models added one new structure for nsp14 (**Figure 25a**) while comparison against the available SARS-CoV-2 PDB structures from the remaining 11 validated the quality of our modeling approach (**Figure 25b-c**). For human interactors all models were obtained from the PDB or Modbase[281] (**Figure 26a**). We then predicted the interface residues for each interaction using our ECLAIR framework[261]. In total, our pipeline identified 679

interface residues across 21 SARS-CoV-2 proteins with an average 18.23 residues per interface and 5,790 across 189 human proteins with an average 17.4 residues per interface.

To provide structural interaction models for visualization and downstream analysis we performed guided docking in HADDOCK[262,263] using our high-confidence ECLAIR-predicted interface residues as restraints to refine the search space. To avoid potential biases in interface identification from docking low coverage models (**Figure 26b**), we only performed docking for 138 out of 332 interactions for which either 1) at least 33% of the full-length proteins were covered by available structures, or 2) available structures included at least one high-confidence ECLAIR prediction to use as docking restraint. In total we report 1,248 docked interface residues across 15 SARS-CoV-2 proteins with an average 33.4 residues per interface and 4,604 across 138 human proteins with an average 32.4 residues per interface. For all analyses, docked interface annotations were prioritized over initial ECLAIR predictions. The full interface annotations from our ECLAIR and docking predictions are available in **Table 16** and T**able 17**, respectively.

Benchmarking ECLAIR and guided docking predictions

Our specific applications of ECLAIR—for interspecies interactions—and HADDOCK—performing data-driven docking with computational rather than experimental priors—are unique from those these tools were previously validated for. To ensure the robustness and quality of these methods for our interface prediction task, we constructed a comprehensive human-pathogen PDB benchmark set consisting of 509 interactions between a human protein and a viral or bacterial interactor (**Figure 27a**).

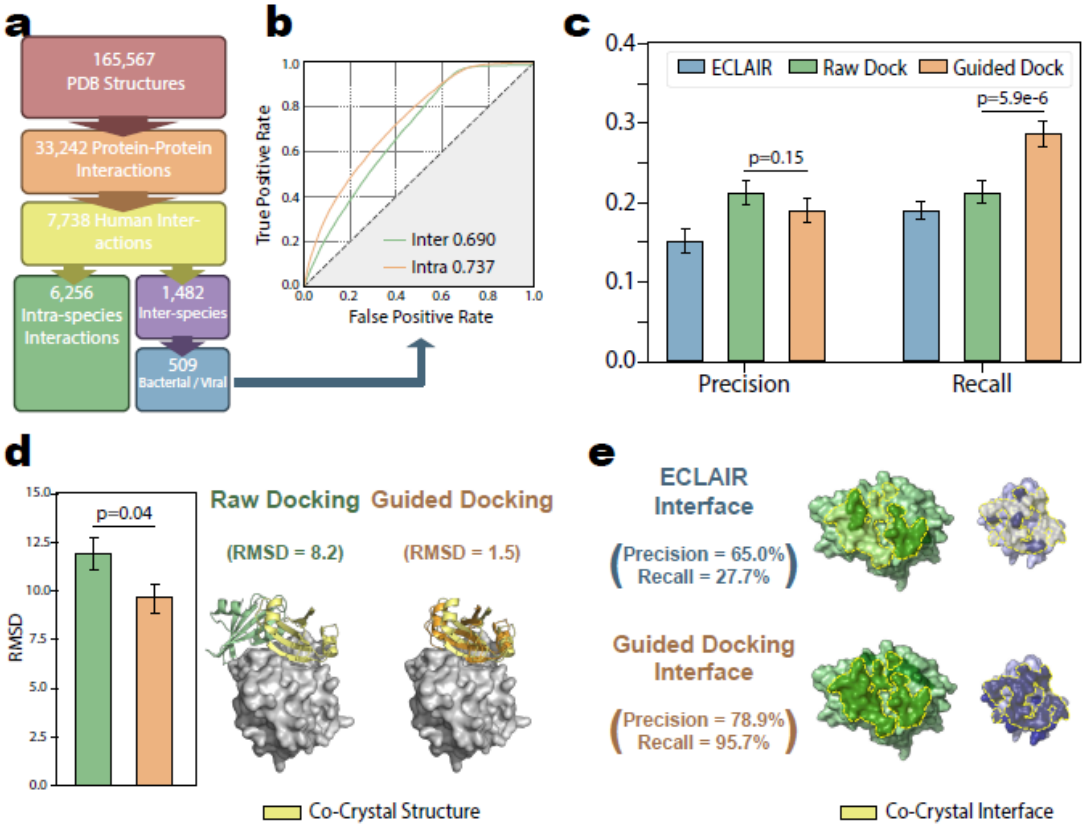


Figure 27. Validation of ECLAIR and Guided Docking Performance.
a, Steps taken to parse the Protein Data Bank and construct our human-pathogen PDB benchmark set. **b,** Comparison of ECLAIR performance on intra-species interactions (n=200 human-human interactions) against inter-species interactions (n=509 human-pathogen interactions). Area under the receiver operating characteristic (AUROC) evaluation indicates considerable predictive power is achieved in both tasks, (intra-species AUROC=0.737, inter-species AUROC=0.690). **c,** Comparison of final interface predictions across all residues in 153 dockable human-pathogen interactions using either ECLAIR (precision=0.15, recall=0.19), a raw docking HADDOCK protocol (precision=0.21, recall=0.21), or our guided docking HADDOCK protocol implementing ECLAIR predictions as restraints (precision=0.19, recall=0.29). Recall from guided docking significantly outperformed raw docking method (p=5.88e-6 by two-sided two proportion z-test) without sacrificing precision (p=0.15 by two-sided two proportion z-test). Data presented as precision or recall ± SD as estimated by 1000-fold bootstrapping sampling 153 interactions and interface predicitons with replacement each iteration. **d,** Distributions of root-mean-square deviation (RMSD) between the top-scored raw or guided docking output and the co-crystal structure (n=153 dockable human-pathogen interactions). Interior boxplots represent the distribution quartiles with whiskers representing the most extreme non-outlier values. Average RMSD from the guided docking (average RMSD=9.45) was significantly lower than the raw docking (average RMSD=11.79) based on two-sided t-test (p=0.04). To the right, an example where the guided docking accurately identifies the correct interaction orientation missed by the raw docking (Human protein shown as gray surface, raw docking, guided docking, and co-crystal structure viral protein shown as green, orange, and yellow cartoon respectively). **e,** Example showing a best case scenario where a few true interface residues predicted by ECLAIR (top, recall=27.7%) are successfully

expanded to identify the rest of the interface by the guided docking (bottom, recall=95.7%). Human and viral proteins shown to the left in green and to the right in blue respectively. Residues identified as interface in each approach are darkened. True interface from the co-crystal structure outlined and shaded in yellow.

The full list of interactions in this benchmark set alongside the PDB sources plus true and predicted interfaces are provided in **Table 18**.

To validate ECLAIR's applicability to inter-species interactions, we compared its published performance the test set of 200 human-human interactions to its performance on our human-pathogen PDB benchmark set. Both tasks achieved comparable performance (ROC AUC=0.69 vs. 0.74), although the intra-species task slightly outperformed inter-species (**Figure 27b**). We note feature availability between sets—for instance, co-evolution features can only be calculated for intra-species interactions—may confound direct comparisons between different interaction sets. Overall, the evaluation of our benchmark conclusively shows that ECLAIR retains predictive power for inter-species interactions.

To evaluate the benefit of using ECLAIR predicted interfaces as restraints in HADDOCK docking, we compared our ECLAIR data driven protocol against a raw protocol with no restraints. From the original 509 inter-species interactions, 153 fit our criteria for docking. We compared interface annotations from each protocol based on precision and recall (**Figure 27c**). Overall interface quality was comparable between both raw and guided protocols (precision=0.21 vs 0.19, p=0.15), however, the guided docking better recovered the total interface (recall=0.21 vs 0.29, p=5.88e-6). Previous evaluation on the HADDOCK framework confirms accurate interface predictions can be achieved even if the precise binding orientation is not recovered. While our main evaluation of interest is correct identification of interface residues, by evaluating the

RMSD between docked and reference structures, we further demonstrate that the guided docking better recapitulated the true co-crystal structures (**Figure 27d**; average RMSD=9.45 vs. 11.79, p=0.04).

Our aim in performing guided docking based on ECLAIR predicted interfaces was to produce atomic-resolution structures that reflected our residue level predictions for use in downstream analyses. However, we also hypothesized that docking would be effective in expanding accurate interface annotations to nearby residues if ECLAIR only identified a few high-confidence interface residues (**Figure 27e**). Comparison of the precision and recall between ECLAIR and our guided docking (**Figure 27c**) is consistent with this hypothesis and clearly demonstrates improvement in our guided docking approach over both raw docking and ECLAIR predictions.

Depletion of human disease mutation at SARS-CoV-2 interfaces

We explored evidence of interface-specific variation by mapping gnomAD[278] reported human population variants (**Table 19**) and sequence divergences between SARS-CoV-1 and SARS-CoV-2 (**Table 20**) onto predicted interfaces. Conserved residues generally cluster along protein-protein interfaces[282], and an analysis of SARS-CoV-2 structure and evolution similarly concluded highly conserved surface residues likely drove protein-protein interactions[283]. Consistent with these prior studies, we observed significant interactome-wide depletion for both viral and human variation along predicted interfaces comparable to that observed along solved human-human interfaces (**Figure 28a**).

Nonetheless, considering each interaction individually, we identify 11 interaction interfaces enriched for human population variants (**Figure 28b**), and 4

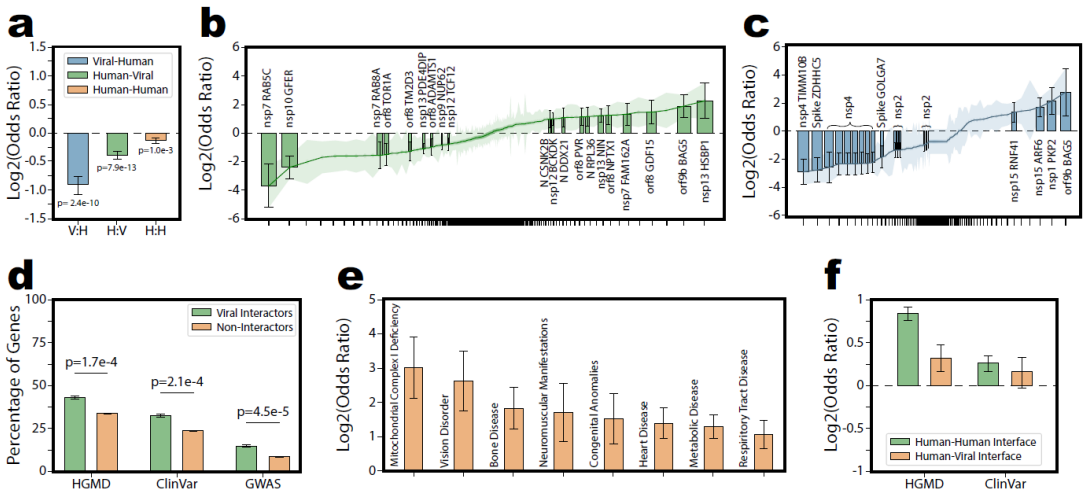


Figure 28. Enrichment of sequence divergences and disease mutations across all SARS-CoV-2-Human interaction interfaces.

a, Enrichment across 332 human genes interacting with SARS-CoV-2 for viral sequence divergence or human population variants along viral-human (V:H, Log2OR=-0.91, p=2.41e-10 by two-sided z-test) human-viral (H:V, Log2OR=-0.38, p=7.92e-13 by two-sided z-test) or human-human (H:H, Log2OR=-0.14, p=9.98e-4 by two-sided z-test) interfaces. Data presented as Log2OR ± SE. **b, c,** Individual enrichments (sorted from most depleted to most enriched) for human population variants and viral sequence divergences respectively on all 332 SARS-CoV-2-human interaction interfaces. Interfaces with statistically significant Log2OR (by two-sided z-test) are labeled and shown as bars, the remainder plotted as a line. Data presented as Log2OR ± SE. Clusters of SARS-CoV-2 enrichments involving the nsp4 interactions with (IDE, NUP210, DNAJC11, TIMM29, TIMM9, and TIMM10) and nsp2 interactions with (GIGYF2, FKBP15, WASHC4, EIF4E2, POR, and SLC27A2) were labeled as a group for legibility. **d,** Percentage of human genes that interact with (green, n=332) or do not interact with (orange, n=20,018) SARS-CoV-2 that contain disease annotations in HGDM (Log2OR=0.57, p=1.70e-4 by two-sided z-test), ClinVar (Log2OR=0.64, p=1.05e-4 by two-sided z-test), and GWAS (Log2OR=0.89, p=4.54e-5 by two-sided z-test) respectively. Genes targeted by SARS-CoV-2 proteins were significantly more likely to harbor disease mutations than non-interactors by log odds enrichment test. Data presented as percentage ± SE. **e,** Sample of individual disease terms enriched in human genes targeted by SARS-CoV-2. Full results reported in **Supplemental Table 6**. Data presented as Log2OR ± SE. **f,** Comparison of the enrichment of HGDM or ClinVar annotated mutations on human-vial interfaces or human-human interfaces for 332 genes interacting with SARS-CoV-2. Disease mutations enriched on human-human interfaces (HGMD, Log2OR=0.84, p<1e-20 by two-sided z-test; ClinVar, Log2OR=0.25, p=2.9e-3 by two-sided z-test), while human-viral interface show no or marginal enrichment (HGMD, Log2OR=0.31, p=0.048 by two-sided z-test; ClinVar, Log2OR=0.15, p=0.39 by two-sided z-test). GWAS category excluded from this analysis because most lead GWAS SNPs occur in non-coding regions. Data presented as Log2OR ± SE.

enriched for recent viral sequence divergences (**Figure 28c**). **Table 21** provides the log odds enrichments for each interface. Similar to the S-ACE2 interface, a high degree of variation on these viral interfaces may indicate recent functional evolution around specific viral-human interactions. Because human evolution is slower, enrichment of population variants along the human interfaces is unlikely to be a selective response to the virus. Rather, interfaces with high population variation may represent edges in the interactome most prone to modulation by existing variation between individuals or populations. Alternatively, enrichment and depletion of variation along the human-viral interfaces could help distinguish viral proteins that bind along existing—likely conserved—human-human interfaces from those that bind using novel interfaces—unlikely to be under selective pressure.

To further explore the functional importance of variations within human interactors of SARS-CoV-2, we considered phenotypic associations reported in HGMD[284], ClinVar[285] or the NHGRI-EBI GWAS Catalog[286]. Interactors of SARS-CoV-2 were enriched for phenotypic variants from each database (**Figure 28d**). Notably, several of the individual disease categories enriched among interactors, were consistent with SARS-CoV-2 comorbidities including heart disease, respiratory tract disease, and metabolic disease[287,288] (**Figure 28e**; **Table 22**). Disruption of native protein-protein interactions is one mechanism of disease pathology, and disease mutations are known to be enriched along protein interfaces[289,290]. Variants on predicted human-viral interfaces matched allele frequency distributions of variants off the interfaces, but were considered overall to be more deleterious by SIFT[291] and PolyPhen[10] (**Figure 29**). However, while we show annotated disease mutations were

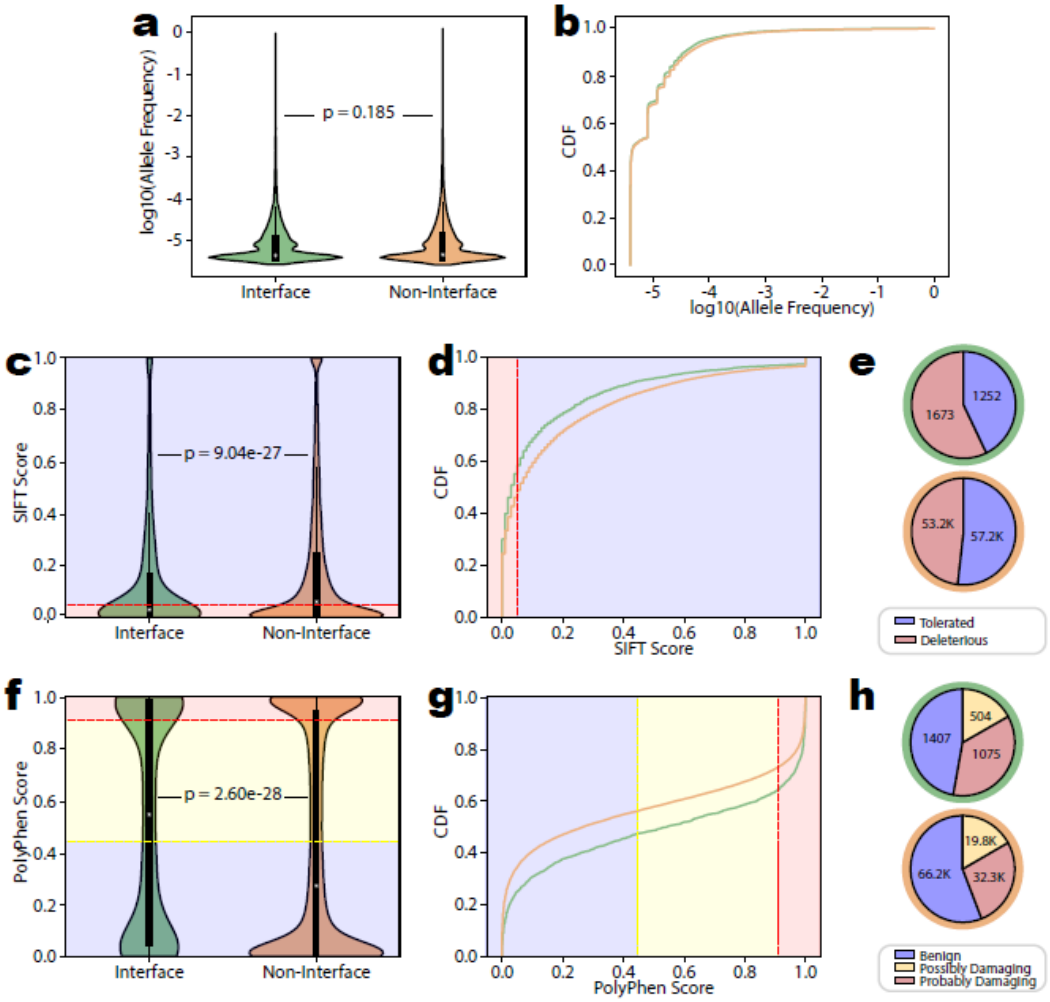


Figure 29. Summary of human population variant frequency and deleteriousness
a, b, Summary of allele frequency for human population variants either on (n=2,925) or off (n=118,042) the predicted human-viral interface presented as either a raw distribution or a cumulative density respectively. Variants in either category had roughly identical allele frequency distributions. Interior boxplots represent the distribution quartiles with whiskers representing the most extreme non-outlier values. **c, d**, Equivalent plots to a and b for the distribution of the SIFT deleteriousness scores for the same human population variant sets. Plots are colored based on the split between SIFT tolerated and deleterious categories. Population variants on the interface were significantly more likely to be classified deleterious by two-sample Kolmogorov-Smirnov test. **e,** Pie chart breakdown of SIFT categories. Pie chart outlines distinguish interface (green) from non-interface (orange). **f, g,** Equivalent plots to **a** and **b** for the distribution of the PolyPhen deleteriousness scores for the same human population variant sets. Plots are colored based on the split between PolyPhen benign, possibly damaging, and probably damaging categories. Population variants on the interface were significantly more likely to be classified deleterious by two-sample Kolmogorov-Smirnov test. **h,** Pie chart breakdown of PolyPhen categories as in **e**. All p-values based on two-sided two-sample Kolmogorov-Smirnov test.

significantly enriched along known human-human interfaces, enrichment was drastically reduced (HGMD) or insignificant (ClinVar) on human-viral interfaces (**Figure 28f**). This is likely because mutations that disrupt human-viral interfaces would not disrupt natural cell function, and hence would be unlikely to manifest as disease phenotypes. Our finding that disease mutations and viral proteins affect human proteins at distinct sites is consistent with a two-hit hypothesis of comorbidities whereby proteins whose function is already affected by genetic background may be further compromised by viral infection.

Changes in binding affinity between SARS-CoV-1 and SARS-CoV-2

Using a PyRosetta pipeline[264,276,277] we predicted the impact of sequence divergences between SARS-CoV-2 and SARS-CoV-1 on the binding energy (ΔΔG) of 138 viral-human interactions amenable to docking. Although the binding energy for most interactions was unchanged, we note that the divergence from SARS-CoV-1 to SARS-CoV-2 was biased towards a decreased binding energy (i.e. more stable interaction) (**Figure 30a**; **Table 23**). The significant outliers in these ΔΔG predictions may help pinpoint key differences between the viral-human interactomes of SARS-CoV-1 and SARS-CoV-2.

To further explore and validate the biological relevance of these predicted changes, we performed yeast two-hybrid (Y2H) screens to test 30 human interactors against both SARS-CoV-1 and SARS-CoV-2 baits. Our Y2H experiments reconstituted 6 of these interactions (20%) using the SARS-CoV-2 bait. Extensive prior studies across many species and hundreds of well-validated interactions show inherent limits in assay sensitivity for all high-throughput interaction assays (detection rates span 15-

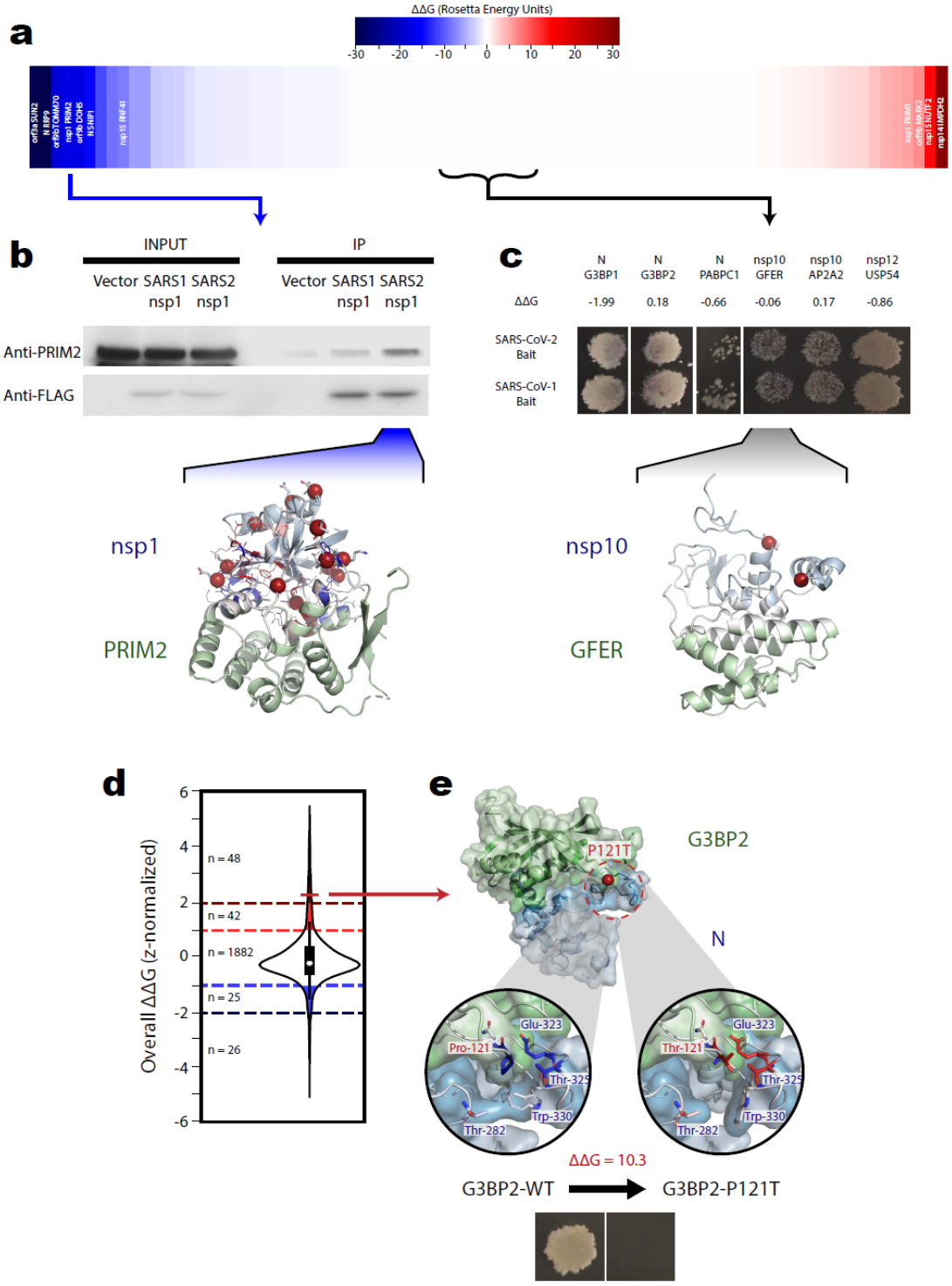


Figure 30. Predicted impact of sequence divergences on the binding affinity of SARS-CoV-2-Human interactions.

a, Predicted impact of SARS-CoV-1 to SARS-CoV-2 sequence divergences on binding affinity from docked structure for 83 applicable SARS-CoV-2-human interactions sorted

from largest decrease (most stabilized relative to SARS-CoV-2) to largest increase (most destabilized relative to SARS-CoV-1) (mean=-0.57 REU, std=5.78 REU). Interaction labels shown wherever predicted ΔΔG exceeds mean ± 1 std. **b,** Representative cropped western blots (among 3 replicates) from co-immunoprecipitation (co-IP) comparing the interaction between human PRIM2 with SARS-CoV-1 or SARS-CoV-2 nsp1. More efficient PRIM2 pull down with SARS-CoV-2 bait validates the PRIM2-nsp1 ΔΔG prediction (ΔΔG=-17.3 REU, z-score=-2.9). Shown below, docked structure for PRIM2 with SARS-CoV-2 nps1 (green and blue cartoon respectively). SARS-CoV-1 to SARS-CoV-2 sequence divergences represented as spheres. Interface residues colored relative to overall ΔΔG contribution ranging from blue (more stabilizing in SARS-CoV-2) to white (little impact on ΔΔG), to red (more stabilizing in SARS-CoV-1). Residue side chains shown as sticks in regions with high local ΔΔG. **c,** Representative Y2H results (among 3 replicates) confirming that 6 interactions with no predicted ΔΔG values can be detected using either SARS-CoV-2 or SASR-CoV viral protein as bait. The docked structure (visualized as in **b**) for human GFER and SARS-CoV-2 nsp10 (ΔΔG=-0.06) shown to highlight that sequence divergences in these 6 interactions did not localize near the interface. **d,** Distribution of the predicted changes in binding affinity from scanning mutagenesis for all 2,023 human population variants on SARS-CoV-2-human interfaces. Values were z-score normalized across for each residue type and on each interface. Shaded regions indicate putative interface binding energy hotspots annotated as strongly disruptive (z-score ≥ 2, 48 total variants), disruptive (1 ≤ z-score < 2, 42 total variants), stabilizing (-2 < z-score ≤ -1, 25 total variants), or strongly stabilizing (z-score ≤ -2, 26 total variants). Interior boxplot represents the distribution quartiles with whiskers representing the most extreme non-outlier values. **e,** Docked structure between SARS-CoV-2 N protein and human G3BP2, alongside expanded interface views comparing the wildtype interface (left) with a predicted strongly disruptive (ΔΔG=10.3 REU, z-score=2.3) population variant, G3BP2_P121T (right). Shown below, yeast two-hybrid results confirmed that the G3BP2_P121T variant completely disrupts the G3BP2-N interaction.

25%)[68,69,292,293]. This is due in part to inability to match native expression, proper folding, or post-translational modifications under assay conditions. Our 20% reproducibility rate—in line with expected sensitivity of the Y2H system—indicating good quality of the published interactome. In each of the 6 reproduced interactions we predicted no changes in binding affinity between SARS-CoV-2 and SARS-CoV-1. Consistent with this prediction, each interaction was also detected using the SARS-CoV-1 bait (**Figure 30c**). Docked models for these interactions suggest sequence divergences between SARS-CoV-1 and SARS-CoV-2 occurred away from the interface and would be unlikely to affect binding (**Figure 30c**).

We additionally performed co-immunoprecipitation (co-IP) assays for the interaction between human DNA Primase Subunit 2 (PRIM2) and SARS-CoV-2 nsp1

(**Figure 30b**; predicted ΔΔG=-17.3 REU). Several deviations in nsp1 were predicted to cumulatively stabilize this interaction near the edges of its interface. Results from the co-IP validated our prediction showing that SARS-CoV-2 nsp1 was more effective at pulling down human PRIM2 than was SARS-CoV-1 nsp1. Moreover, a follow-up quantitative mass spectrometry comparison of SARS-CoV-2, SARS-CoV-1, and MERS-CoV by Gordon et al.[294] included 5 interactions we predicted to be more stable in SARS-CoV-2. Consistent with our predictions 3 of these (RNF41-nsp15, PRIM2-nsp1, and SNIP1-N) showed interaction preferences for the SARS-CoV-2 protein. Specifically, the interaction between RNF41 and nsp15 was exclusively detected in SARS-CoV-2. Overall, these independent experimental results together with our co-IP result thoroughly validate the accuracy of our 3D interactome modelling approach and demonstrate its utility in identifying functional differences between SARS-CoV-1 and SARS-CoV-2.

Impact of population variants on binding affinity

We hypothesized the dynamic range of patient responses and symptoms reported for SARS-CoV-2 infection can be explained in part by missense variations and their impact on viral-human interactions. This is consistent with previous reports that up to 10.5% of missense population variants can disrupt native protein-protein interactions[54] and that underlying genetic variation can explain up to 15% of variation in patient response and viral load in other viruses including HIV[295]. To explore this hypothesis we employed a previously benchmarked scanning mutagenesis protocol provided through PyRosetta[264,276] to identify candidate binding energy hotspot mutations for all docked interfaces. Out of 2,023 population variants on eligible interfaces, we identify 90 (4.4%)

as predicted disruptive hotspots, and 51 (2.5%) as predicted stabilizing hotspots (**Figure 30d**).

To validate our predictions for the impact of population variants, we generated a Ras GTPase-activating protein-binding protein 2 (G3BP2) variant, G3BP2_P121T (RS ID=rs1185000405) using site-directed mutagenesis as described previously[135]. We annotated this variant as strongly disruptive (predicted ΔΔG=10.3 REU) and had confirmed earlier the interaction between N and wildtype G3BP2 could be recapitulated using Y2H. Comparing the Y2H results between wildtype and mutant G3BP2 confirmed complete disruption of the G3BP2-N protein interaction by G3BP2_P121T (**Figure 30e**). Analysis of the docked models, suggests this disruption is driven by steric clashes between the mutated residue in G3BP2 and Glu-323 and Thr-325 of the N protein. The unfavorable polar interaction and steric bulk from the hydroxyl side chain of the threonine variant was also predicted to induce a rotation in the Trp-330 of N disrupting hydrophobic interaction with Trp-282.

G3BP2 is implicated in cardiovascular diseases[296], potentially linking this interaction to known comorbidities. Moreover, G3BP2 alongside G3BP1 is an important target in viral etiology; sequestration of both proteins by SARS-CoV-2 N protein results in an inhibition of stress granule formation and suppression of host innate immune responses[297,298]. Therefore, the existence of naturally occurring variation disrupting this interaction is of particular interest. Although the G3BP2_P121T variant is rare (AF=0.00043%), it may affect SARS-CoV-2 progression in roughly 30,000 individuals who carry it worldwide. Overall, our computational and experimental work concretely shows that human population variants can modulate the SARS-CoV-2-

human interactome network and that our interface and energy modelling predictions can help identify such variants. The full predicted impact of all 2,023 population variants along SARS-CoV-2 interaction interfaces is provided in **Table 24** and may inform future studies investigating genetic contribution to COVID-19.

Comparing binding sites of drugs and SARS-CoV-2 proteins

Drugs that directly interfere with viral-host interactions—for instance by competing for the same binding site—could provide promising clinical leads to target viral infection or replication. On this basis we consider potential for our 3D interactome modelling approach to inform drug repurposing strategies. We aimed to further prioritize a current candidate set including 76 expert-reviewed drugs targeting one or more of the 332 identified human interactors of SARS-CoV-2[239] based on potential for competitive binding. We performed protein-ligand docking using smina[299] to identify drug binding sites for 30 out of 76 candidate drug-target pairs that have available human receptor structures **Table 25**. Sima is a fork of the widely used AutoDock Vina, competes competitively in pose prediction challenges[299], as is validated by us to robustly identify the true binding site from the full protein surface on a published benchmark set of 4,399 experimentally solved protein-ligand complexes (**Figure 31a**)[300].

We compared the overlap of predicted drug binding sites with the corresponding docked viral-human interaction interface for 16 cases with both predictions available. Overall drug binding sites were significantly enriched at the interaction interface compared to the rest of the protein surface (**Figure 31b**; Log2OddsRatio=1.38, p=2.1e-7). Individually, we further prioritize 8 drugs that exhibited significant overlap between the drug- and viral-protein-binding sites (**Figure 31c**), several of which have been

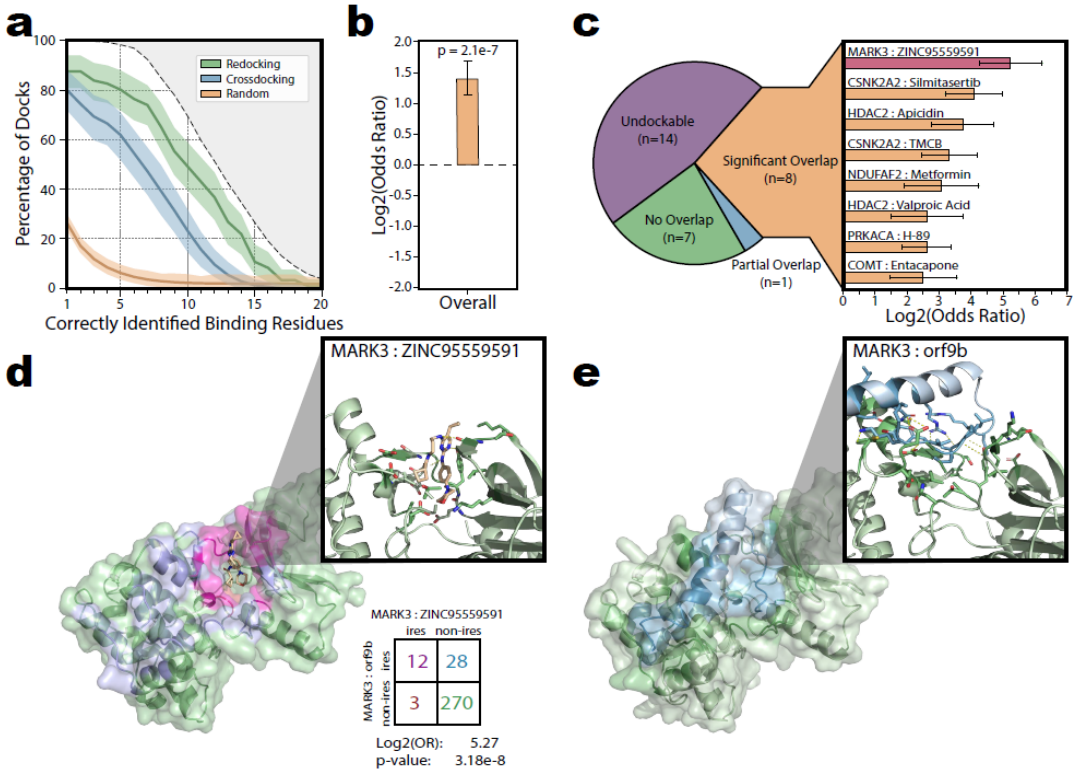


Figure 31. Drug Docking and Prioritization of SARS-CoV-2-Human Interaction Inhibitors.

a, Validation of smina's ability to identify the correct binding site from the full protein surface based on 4,399 drug-ligand pairs across 95 protein targets. Docking was carried out either by redocking each ligand back into its native protein structure, or cross-docking each ligand into a representative receptor structure. Baseline performance expectation derived from random selection of surface patches matching the size of the correct binding site is shown for comparison. Each line and shaded area indicates the percentage of docks that correctly identify X binding site residues ± SD as estimated by 1000-fold bootstrapping sampling 95 drug-target pairs with replacement each iteration. The gray shaded area at the top indicates the maximum fraction of docks whose true binding sites contain at least X residues. **b,** Protein-protein and protein-drug binding sites pooled across 16 applicable drug-target pairs were significantly enriched (Log2OR=1.38, p=2.1e-7 by two-sided z-test). Data presented as Log2OR ± SE. **c,** Individual breakdown of the overlap between the each of the protein-protein and protein-drug binding sites as either undockable (i.e. no protein-protein docked structure available for comparison; 14 total), no overlap (7 total), partial overlap (1 total) or significant overlap (8 total). The individual log2(Odds Ratios) for each of the significant drug-target pairs are shown. Data presented as Log2OR ± SE. The MARK3-ZINC95559591 pair (shown in **d**) is highlighted red. **d,** Docked structure for ZINC95559591 bound to human MARK3. MARK3 surface is colored either green (non-interface, n=270), blue (orf9b interface, n=28), red (ZINC95559591 interface, n=3), or magenta (shared interface, n=12). Cut-out display highlights the MARK3- ZINC95559591 binding site. Polar contacts between MARK3 and ZINC95559591 shown as dashed lines. **e,** Corresponding docked structure for SARS-CoV-2 orf9b bound to human MARK3.

explored by recent independent studies. A retroactive association study identified prior treatment with metformin as an independent factor associated with reduced mortality in diabetic patients[301], although a precise mechanism was not explored at the time. Ongoing phase 2 and phase 4 clinical trials are being conducted or planned for silmitasertib and valproic acid respectively[302,303].

As an example, we highlight orf9b-MARK3 interaction whose interface we predicted could be blocked by ZINC95559591 (MRT-68601 hydrochloride) (**Figure 31d and e**). MARK3 is a serine / threonine protein kinase involved in microtubule organization with implicated roles in modulating gene expression by activating histone deacetylation proteins. Our models suggest both ZINC95559591 and orf9b bind and make several polar contacts with MARK3 (e.g. one with Tyr-134) near its active ATP site. Consistent with its known role as an inhibitor of MARK3[304] our model shows ZINC95559591 binds deep within the ATP active site of MARK3. By contrast the N-terminal tail of orf9b forms looser contact only entering the periphery of the active pocket. Therefore, we suspect ZINC95559591 may outcompete orf9b for this pocket; thus making it a prime candidate to explore targeted disruption of SARS-CoV-2-human protein-protein interactions through drug repurposing.

While this example fits our criteria for prioritized drug repurposing and competitive binding, it does raise further questions to consider. Namely, the functional role of a SARS-CoV-2-human interaction—whether the viral protein co-opts vs. disrupts native human protein function or if interaction is part of an immune response against the virus—is needed to inform potential clinical utility of drug repurposing. Since both orf9b and ZINC95559591 bind within the same MARK3 active site, both

may induce an inhibitory effect and ZINC95559591 could be counterproductive; even if it outcompetes orf9b, it may replace a harmful viral inhibitor with a more potent chemical one. In this scenario, exploration of the predicted binding sites of SARS-CoV-2 proteins could still help uncover an inhibitory role in viral etiology. Moreover, it may be possible to design analogs of inhibitor drugs that retain high binding affinity to their receptor but lose their inhibitor activity. Therefore, while these factors may complicate the prospects of drug repurposing, we are optimistic that our 3D interactome modelling approach can facilitate understanding of viral mechanisms and may aid development of new treatments.

The SARS-CoV-2-human 3D structural interactome web server

We constructed the SARS-CoV-2-human 3D interactome web server (http://3D-SARS2.yulab.org) to provide our computational predictions and modeling as a comprehensive resource to the public. All results and analyses described herein are directly available for bulk download or users can quickly navigation through the reported interactome to see a summary of our analyses for specific interactions of interest (**Figure 32**).

The interface comparison panel (**Figure 32 top left**) visualizes the interface annotation along a linear sequence and provides comparison against all other known or predicted interfaces from the same protein. This comparison may reveal biologically meaningful insights about the interface overlap and possible competition between viral and human interactors.

The mutations panel (**Figure 32 top right**) presents information on variation

Figure 32. 3D-SARS2 Structural Interactome Browser Overview.
Overview of the main results page for exploring a given interaction in our 3D-SARS2 structural interactome browser. The main display contains information for both the SARS-CoV-2 and human proteins including structural displays for either the docked or single crystal structures as well as a table summarizing the interface residues for both proteins. Interface residues are colored dark blue and dark green for the viral and human proteins respectively. By default the page will display the docked structure if available. The display can be toggled between docked structures and single structures using the button in the bottom middle. When single structures display is selected residues will instead be colored based on the initial ECLAIR interface definition. Four categories of expandable panels containing additional analyses are provided. **upper left,** The interface view shows a linear representation of the protein sequence with interface residues annotated in dark blue or dark green. Interfaces for other interactors of the protein are shown underneath for easy comparison. **upper right,** The mutations panel summarizes either human population variants or viral sequence divergences on the protein. Mutations on the interface are labeled. **lower left,** The ΔΔG information panel summarizes the results from in-silico mutagenesis scanning along the interface. Results for each mutation are z-score normalized relative to the rest of the interface and colored on a blue (negative ΔΔG, stabilizing) to yellow (minimal impact) to red (positive ΔΔG, destabilizing). The heatmap can be filtered to only show values corresponding to known mutations on the interface. **lower right,** The candidate drugs panel shows docking information for any known drug targets of the human protein.

within each interaction partner; divergences from the SARS-CoV-1 or gnomAD population variants. We provide a log odds enrichment or depletion of variation along the interface which can help highlight interactions undergoing functional evolution for further characterization.

For interactions amenable to docking, the ΔΔG Information panel (**Figure 32 lower left**) compiles the predicted impact of all possible mutations across the docked interface on binding affinity. Individual mutations are colored by their z-score normalized ΔΔG prediction and can be toggled to only show the impacts of known variants. On the viral side, a cumulative ΔΔG value compares binding affinity between the SARS-CoV-1 and SARS-CoV-2 versions of the protein.

Finally, the drug panel (**Figure 32 lower right**) describes any drugs known to target human proteins and provides information for each drug alongside display options for visualizing predicted binding conformations. The overlap between the drug binding site and interface with the viral protein is reported.

The SARS-CoV-2 human 3D structural interactome web server currently includes 332 viral-human interactions reported by Gordon *et al*.[239]. We will continue support for the web server with periodic updates as additional interactome screens between SARS-CoV-2 and human are published. As we update, a navigation option to select between the current or previous stable releases of the web server will be provided.

Discussion

Our 3D SARS-CoV-2-human interactome provides a comprehensive resource to supplement ongoing and future investigations into COVID-19. The analyses provided

and discussed throughput highlight potential applications of these predictions to inform structure-based hypotheses regarding the roles of individual interactions and prioritize further functional characterization of evolutionarily relevant interactions, causal links connecting population variation with differences in response to infection, and drug candidates that may interfere with interaction-mediated disease pathology. Our observation that perturbation from underlying disease mutations and viral protein binding occur at distinct sites on human proteins may warrant further investigation into whether the combined role of these two sources of perturbation is clinically relevant to mechanisms of comorbidities.

Although we have experimentally validated several of our predictions, we emphasize that further experimental characterization should be conducted to corroborate any hypotheses derived from individual predictions. Moreover, these predictions are not without limitation. Interface predictions may not be applicable to some published human targets identified by mass spectrometry[239] if they represent indirect complex associations rather than direct binary interactions[68]. Further, while structural coverage from SARS-CoV-2 proteins was robust, per-residue coverage of the human proteome is less complete (**Figure 26**). Though we only performed molecular docking for low coverage structures when strong prior ECLAIR interface restraints were available, coverage restrictions can nonetheless introduce bias and may prohibit identification of true interface residues. Recent advances in protein-folding predictions[305-307] may ameliorate this restriction in the future. In the meantime, initial ECLAIR interface annotations—not susceptible to structural coverage limitations—may provide orthogonal value to docked models.

Additionally, we caution that direct quantitative interpretation of Rosetta-predicted ΔΔG values is often difficult. In particular, relative importance of scoring function terms may differ between proteins and interactions of varying sizes and compositions. For these reasons, we only evaluate normalized predictions to compare the relative qualitative differences from our scanning mutagenesis results. Moreover, because mutated structure optimization focuses only on side-chain repacking, our analysis is limited to mutations at or near the interface where side-chain repacking can have a direct effect. We expect mutations that significantly impact binding affinity through refolding or other allosteric effects exist but cannot be captured by our method.

Importantly, users can tailor use of our raw predictions to their own interests; thus expanding upon the concepts and applications our analyses explore. For instance, we limited investigation of druggable interactions to repurposing known drugs that overlap and might disrupt viral-host interactions which we hypothesized would elicit the most promising clinical responses. However, this approach reduces the scope of the SARS-CoV-2-human interactome to only a few interactions that already have known drug candidates. An alternative application could prioritize candidate druggable interfaces throughout the whole SARS-CoV-2-human interactome by overlapping our interface annotations with predictions of druggable protein surfaces using recent deep-learning approaches[308] with the aim of designing novel protein-protein interaction inhibitors.

Overall, we believe our 3D structural SARS-CoV-2-human interactome web server (http://3D-SARS2.yulab.org) will prove to be a key resource in informing hypothesis-driven exploration of the mechanisms of SARS-CoV-2 pathology and host

response. The scope, and potential impacts of our webserver will continue to grow as we incorporate the results of ongoing and future interactome screens between SARS-CoV-2 and human. Finally, we note our 3D structural interactome framework can be rapidly deployed to analyze future viruses.

Methods

Generation and validation of SARS-CoV-2 homology models

Homology-based modeling of all 29 SARS-CoV-2 proteins was performed in Modeller[309] using a multiple template modeling procedure consistent with previous high-profile homology modelling resources[310]. In brief, candidate template structures for each query protein were selected by running BLAST[311] against all sequences in the Protein Data Bank (PDB)[94] retaining only templates with at least 30% identify. Remaining templates were ranked using a weighted combination of percent identity and coverage described previously[310]. The final set of overlapping templates to use was first seeded with the top ranked template with additional templates being added iteratively if: 1) overall coverage increase from the template was at least 10%, and 2) percent identify of the new template was no less than 25% the identity of the initial seed template. Query-template Pairwise alignments were generated in Modeller using default settings and were manually trimmed to remove large gaps (≥5 gaps in a 10 residue window). Finally, modelling was carried out using the Modeller automodel function.

This approach generated homology models for 18 out of 29 proteins. Based on manual inspection of the template quality and sources, homology models were further filtered to 12 models for which a high-quality template from a SARS-CoV-1 homolog

was available. Moreover, during revision of this manuscript, newly deposited PDB structures for many SARS-CoV-2 proteins (https://rcsb.org/covid19) allowed independent validation of homology model quality based on the root-mean-square deviation (RMSD) following alignment and refinement in PyMol[312]. Visual representations these alignments between modelled and solved structures are provided in **Figure 25**. For all analyses SARS-CoV-2 PDB structures were prioritized where available, and only the homology model for nsp14 was retained.

Interface prediction using ECLAIR

Interface predictions for all 332 interactions reported by Gordon *et al.*[239] were made in two phases. In phase one, we leveraged our previously validated ECLAIR frameworkd[261] to perform initial residue-level predictions across all interactions. ECLAIR compiles five sets of features; biophysical, conservation, coevolution, structural, and docking. In brief, biophysical features are compiled using a windowed average of several ExPASy ProtScales[313], conservation features are derived from the Jensen-Shannon divergence[314,315] from known homologs for each protein, coevolution features between interacting proteins are derived from direct coupling analysis (DCA)[316] and statistical coupling analysis (SCA)[317] among paired homologs, structural features are obtained by calculating the solvent accessible surface area of available PDB[94] or ModBase[281] models using NACCESS[318], and docking features are the average inter-chain distance and surface occlusion per residue from a consensus of independent Zdock[319] trials.

Slight alterations were made to accommodate SARS-CoV-2-human predictions. First, construction of multiple sequence alignment (MSA) for SCA and DCA

calculations require at least 50 species containing homologs of both interacting proteins. Therefore, co-evolution features could not be calculated for inter-species interactions. Second, MSAs for conservation features typically only allow one homolog per species. Because viral species classifications are less precise and are often subdivided into unique strains (and because all higher-order ECLAIR classifiers require protein conservation features) we modified the MSAs for viral proteins to include homologs from various strains in a single species. The initial prediction results from ECLAIR are provided in **Table 16**.

Interface Prediction Using Guided HADDOCK Docking

Interface predictions for all 332 interactions reported by Gordon *et al.*[239] were made in two phases. In phase two, we leveraged high-confidence interface predictions from ECLAIR to perform guided docking in HADDOCK[262,263]. For a thorough introduction to protein-protein docking in HADDOCK, see https://www.bonvinlab.org/education/HADDOCK-protein-protein-basic/.

In brief, HADDOCK is designed to perform data-driven docking using (traditionally experimentally derived) priors about the interface. These data (e.g. scanning mutagenesis) often indicate sets of residues involved in the interface but no pairwise information linking interface residues between each protein. These residues (termed active residues) are used in conjunction with any neighboring surface residues (termed passive residues) to drive rigid body docking, by introducing a scoring penalty for any active residue on one protein not in proximity of an active or passive residue on the other. This approach is formalized as a set of ambiguous interaction restraints (AIR)

that evaluate the distances of each active residue to the active or passive residues on the other protein. The approach ensures experimental priors about interface composition are enforced, but leaves the exact orientation and pairing of residues flexible to HADDOCK's energy based scoring function.

To incorporate computational interface predictions from ECLAIR we use the standard HADDOCK protein-protein docking framework. Active residues are encoded as all high-confidence ECLAIR predictions at the surface (≥15% SASA). Passive residues are identified as all surface residues (≥40% SASA) within 6 Å of an active residue. For definition of surface residues, the 15% SASA cutoff is for consistency with our definition of interface residues, while the 40% SASA cutoff is for consistency with the typical recommendation in HADDOCK. All SASA calculations were carried out using NACCESS[318] and neighboring residues were selected using PyMol[312]. Following HADDOCK recommendations to reduce computational burden from using many restraints, we defined our AIR using only the alpha carbons and increased the upper distance limit for from 2 Å to 3 Å. All other HADDOCK run parameters were left at the default. In total 1000 rigid body docking trials were performed, and the top 200 scored orientations were retained for subsequent iterations refinement and analysis.

For each interaction we identified available PDB or homology model structures to determine whether the interaction should be eligible for docking. Previous benchmark evaluations show the HADDOCK performs using homology models, but that performance drops off for models produced from low sequence identity templates[320]. In all cases PDB models were prioritized over homology models. We next evaluated risks of using low coverage structures for protein-protein docking; using structure fragments

that completely exclude the true interface residues will produce false interface predictions. We aimed to minimize this risk while maximizing the dockable interactome by setting two conditions for determining structure eligibility. First, protein structures covering at least 33% of the total protein length were considered sufficiently large for docking. Second, protein structures at least 50 residues in length and containing at least one high-confidence ECLAIR predicted interface residue to use as an active residue were made eligible. Inclusion of an ECLAIR-defined active residue gives us reasonable confidence that part of the interface is covered, and therefore, true docked interface predictions should be possible. When multiple structures were available for one protein, ranking was based on the sum of ECLAIR scores for all residues covered by each structure; we always selected the available structure most likely to include the true interface.

In total we performed guided HADDOCK docking on 138 out of 332 interactions. The remaining 194 interactions did not have reliable 3D models for both interactors. The top scored docked conformation from each HADDOCK run was retained. The final docked interface annotations are provided in **Table 17**.

Definition of interface residues

We annotate interface residues from atomic resolution docked models, using an established definition for interface residues[261]. The solvent accessible surface area (SASA) for both bound and unbound docked structures was calculated using NACCESS[318]. We define as interface residue, any residue that is both 1) at the surface of a protein (defined as ≥ 15% relative accessibility) and 2) in contact with the interacting chain (defined by a ≥ 1.0 $Å^2$ decrease in absolute accessibility).

Human-Pathogen co-crystal structure benchmark set

We constructed a benchmark set of experimentally determined co-crystal structures to evaluate the performance of both our ECLAIR and guided HADDOCK docking interface predictions on inter-species interactions (**Figure 27a**). First, we parsed 165,567 PDB structures, identified all interacting chains by interface residue calculation, and mapped PDB chains to UniProt protein IDs using SIFT[291] to identify a total of 33,242 unique protein-protein interactions. Using taxonomic lineages from UniProt we filtered this set to 7,738 interactions involving human proteins, of which 6,256 represented human-human intra-species interactions, and 1,482 represented inter-species interactions between human and some other species. Finally, to provide the most relevant set of interactions that would biologically similar to SARS-CoV-2-human interactions, we only considered interactions between human and viral proteins (346) or between human and bacterial proteins (163). We refer to this collective set of 509 co-crystal structures as our human-pathogen PDB benchmark set. The full list of structures and interface annotations for this benchmark set is provided in **Table 18**.

To validate performance of ECLAIR predictions on the human-pathogen PDB benchmark, ECLAIR predictions were run as described above for SARS-CoV-2-human interactions. Evaluation of raw prediction probabilities was done by area under the receiver operating characteristic curve (AUROC) in python using scikit-learn and was compared against ECLAIR's original test set containing 200 intra-species interactions[261]. Precision and recall metrics were calculated based on ECLAIR's binary definition for high-confidence vs. non-interface predictions.

To validate HADDOCK guided docking performance using our human-

pathogen PDB benchmark, we compared performance with a raw HADDOCK docking protocol. Guided docking was performed as described for SARS-CoV-2-human interactions. No PDB protein chains from the human-pathogen benchmark were used during docking. For raw HADDOCK docking no experimental constraints (AIR) were provided and the ranair and surfrest parameters in the run.cns were set to true. Using these parameters, each rigid dock generates one random AIR between one surface residue from each protein A and B which is used to ensure the two protein chains slide together during docking. Overall performance of protocols was evaluated based on precision and recall of the true interface (**Figure 27c**). Secondary evaluation of was done based on root-mean-squared deviation (RMSD) in PyMol before refinement between the docked and co-crystal structures (**Figure 27d**). When multiple co-crystal structures were used to define the interfaces, the RMSD was reported as the average RMSD against all co-crystal structures.

Compilation of sequence variation sets

For analysis of genetic variation that may impact the viral-human interactome, two sets of mutations were compiled; 1) viral mutations, and 2) human population variants.

For viral mutations, we identified sequence divergences between SARS-CoV-1 and SARS-CoV-2 versions of each protein based on alignment. Representative sequences for 16 SARS-CoV-1 proteins were obtained from UniProt (Proteome ID UP000000354)[321,322]. Sequences for 29 SARS-CoV-2 proteins were reported by Gordon *et al.*[239] and based on genbank accession MN985325[323,324]. Notably, UniProt accessions for the SARS-CoV-1 proteome report two sequences for the uncleaved ORF1a and ORF1a-b which correspond to NSP1 through NSP16 in SARS-CoV-2. Sequence

divergences were reported after pairwise Needleman Wench alignment[325,326] (using Blosum62 scoring matrix, gap open penalty of 10 and gap extension penalty of 0.5) between the corresponding protein sequences from each species. A total of 1,003 missense variants were detected among 23 SARS-CoV-2 proteins. No suitable alignment form a SARS-CoV-1 sequence was available for orf3b, orf8, or orf10. Additionally, orf7b, nsp3, and nsp16 were excluded because they were not involved in any viral-human interactions. The full list of SARS-CoV-2 mutations is reported in **Table 20**.

We obtained human population variants for all 332 human proteins interacting with SARS-CoV-2 proteins from gnomAD[278]. We used gnomAD's graphQL API to run programmatic queries to fetch all missense variants per gene. Details on performing gnomAD queries in this manner are available in the gnomad-api github page (https://github.com/broadinstitute/gnomad-browser/tree/master/projects/gnomad-api). We used the Ensembl Variant Effect Predictor (VEP)[327] to map gnomAD DNA-level SNPs to equivalent protein-level UniProt annotations. After VEP mapping, variants were parsed to ensure the reported reference amino acid and position agree with the UniProt sequence and roughly 4.4.6% of variants that did not match were dropped from our dataset because they could not reliably be mapped to UniProt coordinates. In total 127,528 human population variants were curated. The full list of human population variants from GnomAD is reported in **Table 19**.

Log odds enrichment calculations

To determine enrichment or depletion, odds ratios were calculated as described previously[328]…

$$OR = \frac{a\ /\ c}{b\ /\ d}$$

Where, a, b, c, and d describe values in a contingency table between case and exposure criteria. For a particular application, where we are interested in the enrichment of viral mutations or human populations variants (case: Variant vs. NonVariant) along predicted interaction interfaces (exposure: Interface vs. NonInterface), we would have…

$$a = \ Number\ of\ Variant\ Interface\ Residues$$

$$b = \ Number\ of\ NonVariant\ Interface\ Residues$$

$$c = \ Number\ of\ Variant\ NonInterface\ Residues$$

$$d = Number\ of\ NonVariant\ NonInterface\ Residues$$

Statistical tests for enrichment or depletion were performed by calculating the z-statistic and corresponding two-sided p-value for the odds ratio (unadjusted for multiple hypothesis testing)…

$$z = \frac{\ln OR}{\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}$$

All reported odds ratios were $\log_2$ transformed to maintain interpretable symmetry between enriched and depleted values. To avoid arbitrary odds ratio inflation or depletion from missing data, in all cases where the interface residues were predicted by molecular docking, the odds ratio was altered to only account for positions that were included in the structural models used for docking.

Curation of disease associated variants

To explore whether human proteins interacting with SARS-CoV-2 proteins were enriched for disease or trait associated variants, three datasets were curated; the Human

Gene Mutation Database (HGMD)[284], ClinVar[285], and the NHGRI-EBI GWAS Catalog[286]. Disease annotations for HGMD and ClinVar were downloaded directly from these resources and mapped to UniProt. To calculate enrichment of individual disease terms, we reconstructed the disease ontology from NCBI MedGen term relationships (https://ftp.ncbi.nlm.nih.gov/pub/medgen/MGREL.RRF.gz) and propagated counts up through all parent nodes up to a singular root node. Significant terms were reported as the most general term with no more significant ancestor term (**Table 22, sheet 1**). Raw enrichment values for all terms are also provided (**Table 22, sheet 2**).

For curation of disease and trait associations from NHGRI-EBI GWAS Catalog (http://www.ebi.ac.uk/gwas/)[286], lead SNPs (p-value<5e-8) for all diseases/traits were retrieved on June 16, 2020. Proxy SNPs in high linkage disequilibrium (LD) (Parameters: $R^2 > 0.8$; pop: "ALL") for individual lead SNPs were obtained through programmatic queries to the LDproxy API[329], which used phase 3 haplotype data from the 1000 Genomes Project as reference for calculating pairwise metrics of LD. Both lead SNPs and proxy SNPs were filtered to only retain missense variants.

In-silico scanning mutagenesis and ΔΔG estimation

To explore importance of each SARS-CoV-2-human interface residue and the impact of all possible mutations along the interface, we performed in-silico scanning mutagenesis. We use a setup provided by the PyRosetta documentation (https://graylab.jhu.edu/pyrosetta/downloads/scripts/demo/D090_Ala_scan.py) designed around an approach previously benchmarked to correctly identify nearly 80% of interface hotspot mutations[276]. For consistency, we replaced the PyRosetta implementation's definition of interface residues (≤ 8.0 Å away from partner chain),

with our definition described above.

We encourage reference to the original well-documented demo for details, but in brief, we consider all interface residue positions and, begin by estimating the wildtype binding energy for the interaction. The complex state energy is calculated following a PackRotamersMover operation to optimize the side-chains of residues within 8.0 Å of the interface residue to be mutated. The chains are separated 500.0 Å to eliminate any interchain energy contributions and energy for the unbound state is calculated the same way. The difference between these two values provides the binding energy for the wildtype structure.

$$\Delta G_{WT} = E_{complex} - E_{unbound}$$

To estimate the binding energy for all 19 amino acid mutations possible at the given position, each mutation is made iteratively, and the ΔG_{Mut} is as above using the mutated structures. Finally, the change in binding energy from each mutation is the difference between these two binding energies.

$$\Delta\Delta G = \Delta G_{Mut} - \Delta G_{WT}$$

The scoring function used for these calculations is as described previously[276] using the following weights; fa_atr=0.44, fa_rep=0.07, fa_sol=1.0, hbond_bb_sc=0.5, hbond_sc=1.0. To account stochasticity of the PackRotamersMover optimization between trials, all ΔΔG values are reported from an average of 10 independent trials. To test whether an a mutation had a significantly non-zero impact on binding energy, a two-sided z-test between the 10 independent trials was performed. To account for average impact of other same amino acid mutations at other positions along the interface, each average ΔΔG was z-normalized relative to the rest of the interface and outliers were

called at ≥ 1 standard deviation away from the mean. Mutations that passed both criteria were identified as significant interface binding affinity hotspots. No adjustments were made for multiple hypothesis corrections.

Predicting ΔΔG from SARS-CoV-1 and SARS-CoV-2 divergences

Estimates of the overall impact of the cumulative set of mutations between SARS-CoV-1 and SARS-CoV-2 were made based on the in-silico mutagenesis framework modified to introduce multiple mutations at a time. We generated interaction models using the SARS-CoV-1 protein by applying all amino acid substitutions between the two viruses to initial docked models containing the SARS-CoV-2 protein. A minority of mutations that comprised insertions or deletions could not be modelled under this framework. The ΔΔG calculation here was identical to the single mutation ΔΔG described above, except that side-chain rotamer optimization involved all residues within 8.0 Å of any of the mutated residues. The ΔΔG were calculated considering the SARS-CoV-1 as the wildtype such that a negative ΔΔG indicates the interaction is more stable (lower binding energy) in the SARS-CoV-2 version of the interaction compared to the SARS-CoV-1 version of the interaction…

$$\Delta\Delta G = \Delta G_{SARSCoV2} - \Delta G_{SARSCoV1}$$

To account for stochasticity between trials for these predictions (which notably had a larger impact likely due to the decreased constraints on rotamer optimization in these cases), this set of ΔΔG values was reported as an average of 50 trials. Significant outliers for overall binding affinity change from SARS-CoV-1 to SARS-CoV-2 were called based on similar criteria to the individual mutations, except the z-score normalization was performed relative to all other interactions.

Protein-ligand docking using smina

To further prioritize 76 previously reported candidate drugs targeting human proteins in the SARS-CoV-2-human interactome[239], we performed protein-ligand docking for, 30 interaction-drug pairs (involving 25 unique drugs) that were amenable to docking. For docking, we excluded any human protein targets whose structures were below 33% coverage. To prep for docking, 3D structures for all ligands were first generated using Open Babel[330] and the command:

obabel -:"[SMILES_STRING]" --gen3d -opdb -O [OUT_FILE] -d

Protein-ligand docking was executed using smina[299] with the following parameters. The autobox_ligand option was turned on and centered around the receptor PDB file with an autobox_add border size of 10 Å. To increase the number of independent stochastic sampling trajectories and increase the likelihood of identifying a global minimum, the exhaustiveness was set to 40and the num_modes was set to retain the top 1000 ranked models. To reduce real wall time each docking process was run using 5 CPU cores (no impact on net CPU time). The final smina command used was as follows:

smina -r [RECEPTOR] -l [LIGAND] --autobox_ligand [RECEPTOR] --autobox_add 10 -o [OUT_FILE] --exhaustiveness 40 --num_modes 1000 --cpu 5 --seed [SEED]

Each protein-ligand docking command was repeated 10 times (essentially the same as one trial with exhaustiveness set to 400) with a unique seed in order to saturate the ligand binding search space as thoroughly as possible. We note that a single run with exhaustiveness ranging from 30-50 is considered sufficient for most applicaitons[299]. To retain candidate poses covering different low-energy binding sites, a final set of up to 10 of the best scoring poses with centers at least 1 Å away from one another was

selected. Results described in this manuscript are reported based the top ranked pose. Protein residues involved in drug binding sites were annotated using the same criteria used to define interface residues. The Record Type for all ligand atoms was first manually changed from HETATM to ATOM because NACCESS otherwise excluded ligand atoms from the solvent accessible surface area calculations.

Validation of smina docking to identify drug binding sites

Past evaluation of smina show competitive performance across numerous Community Structure-Activity Resources (CSAR)[299,300]. However, tradition docking evaluation tasks, focus on sampling and correctly scoring docked conformations within a single known binding site and may frequently restrict the docking space to a few angstroms bounding box around the known ligand conformation. The focus is on recovering precisely how a ligand orients within a binding site rather than identifying the binding site from the whole protein surface.

Because this performance metric may not provide sufficient confidence in smina's ability to identify a binding site from scratch (our application in this manunscript) we re-benchmarked smina's performance using an established drug docking benchmark set containing 4,399 protein-ligand complexes representing 95 protein targets[300]. We defined true ligand binding site residues from the available crystal structure and evaluated the fraction correctly recovered by smina's top-ranked dock across the full protein surface.

Docking was performed as above and evaluated based on both redocking—ligand docked back into the exact receptor structure it came from—and crossdocking—ligand docked into an alternate conformation of the receptor it came from—conditions.

Because the conformation of the binding pocket from an alternate receptor may not perfectly accommodate the ligand, crossdocking is considered more difficult, but also more representative of real conditions when making new predictions.

To provide a reference for whether smina selectively recovered the true binding site we calculated a baseline random expectation. Artificial binding sites were defined by selecting a single surface residue and its N nearest neighbors where N is the number of binding site residues in the true binding site. The average recovery of the true binding site from all such artificial binding sites was used as the null expectation for each drug-target pair.

Construction of plasmids for Y2H and co-IP

Clones of all human proteins tested were picked from the hORFeome 8.1 library[56]. Clones for all SARS-CoV-1 and SARS-CoV-2 proteins tested were designed to match GenBank entries AY357076 and MN908947 respectively. To construct plasmids for testing by Y2H viral genes were PCR amplified and cloned into PDEST-AD and PDEST-DB vectors (for Y2H). For co-immunoprecipitation (co-IP) Gateway LR reactions were used to transfer bait SARS-CoV-2 nsp1 protein into a pQXIP (ClonTech, 631516) vector modified to include a Gateway cassette featuring a carboxy-terminal 3×FLAG.

Yeast two-hybrid (Y2H) screens

Y2H experiments were carried out as previously described[54,68,84] in order to 1) confirm that SARS-CoV-2-human interactions previously detected by immunoprecipitation mass-spectrometry (IP-MS) could be recapitulated in Y2H, 2) compare the occurrence

of interactions using SARS-CoV-1 vs. SARS-CoV-2 viral baits, and 3) profile the disruption of SARS-CoV-2-human interactions by human population variants. In brief human and viral clones were transferred into Y2H vectors pDEST-AD and pDEST-DB by Gateway LR reactions then transformed into *MAT*a Y8800 and *MAT*α Y8930, respectively. For comparisons of interest, the viral-human interactions were screened in both orientations; namely viral DB-ORF MATα transformants were mated against corresponding human AD-ORF *MAT*a transformants and vice versa. All DB-ORF yeast cultures were also mated against *MAT*a yeast transformed with empty pDEST-AD vector to screen for autoactivators. Mated transformants were incubated overnight at 30 °C, before being plated onto selective Synthetic Complete agar media lacking leucine and tryptophan (SC-Leu-Trp) to select for mated diploid yeast. After another overnight incubation at at 30 °C, diploid yeast were plated onto two sets of SC-Leu-Trp agar selection plates; one lacking histidine and supplemented with 1 mM of 3-amino-1,2,4-triazole (SC-Leu-Trp-His+3AT), the other lacking adenine (SC-Leu-Trp-Ade). After overnight incubation at 30 °C, plates were replica-cleaned and incubated again for three days at 30 °C for final interaction calling.

Cell culture, co-immunoprecipitation and western blotting

HEK 293T cells (ATCC, CRL-3216) were maintained in complete DMEM medium supplemented with 10% FBS. Cells were seeded onto 6-well dishes and incubated until 70–80% confluency. Cells were then transfected with 1 µg of either empty vector, SARS-CoV-1 nsp1 or SARS-CoV-2 nsp1, respectively, and combined with 10 µl of 1 mg ml−1 PEI (Polysciences, 23966) and 150 µl OptiMEM (Gibco, 31985-062). After 24 h incubation, cells were gently washed three times in 1×PBS and then resuspended

in 200 µl cell lysis buffer (10 mM Tris-Cl pH 8.0, 137 mM NaCl, 1% Triton X-100, 10% glycerol, 2 mM EDTA and 1×EDTA-free Complete Protease Inhibitor tablet (Roche)) and incubated on ice for 30 min. Extracts were cleared by centrifugation for 10 min at 16,000g at 4 °C. For co-immunoprecipitation, 100 µl cell lysate per sample was incubated with 5 µl EZ view Red Anti-FLAG M2 Affinity Gel (Sigma, F2426) for 2 h at 4 °C under gentle rotation. After incubation, bound proteins were washed three times in cell lysis buffer and then eluted in 50 µl elution buffer (10 mM Tris-Cl pH 8.0, 1% SDS) at 65 °C for 10 min. Cell lysates and co-immunoprecipitated samples were then treated in 6×SDS protein loading buffer (10% SDS, 1 M TrisCl pH 6.8, 50% glycerol, 10% β-mercaptoethanol, 0.03% bromophenol blue) and subjected to SDS–PAGE. Proteins were then transferred from gels onto PVDF (Amersham) membranes. Anti-FLAG (Sigma, F1804) and anti-PRIM2 (abcam, ab241990) at 1:3,000 dilutions were used for immunoblotting analysis.

Cloning human population variants through site-directed mutagenesis

Generation of mutant clones containing human population variants was done using site-directed mutagenesis as described previously[135]. In brief, WT G3BP2 was picked from the hORFeome 8.1 library[56] and used as a template for site-directed mutagenesis. Site-specific mutagenesis primers (Eurofins) for mutagenesis were designed using the webtool primer.yulab.org. To minimize sequencing artifacts, PCR was limited to 18 cycles using Phusion polymerase (NEB, M0530). PCR products were digested overnight with DpnI (NEB, R0176) then transformed into competent bacteria cells to isolate single colonies. To confirm successful mutagenesis single colonies were then hand-picked, incubated for 21 h at 37 °C under constant vibration, and submitted for

Sanger sequencing to ensure the desired single base-pair mutation—an no other mutations—had been introduced.

REFERENCES

10 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* 7, 248-249, doi:10.1038/nmeth0410-248 (2010).

54 Fragoza, R. *et al.* Extensive disruption of protein interactions by genetic variants across the allele frequency spectrum in human populations. *Nat Commun* 10, 4141, doi:10.1038/s41467-019-11959-3 (2019).

56 Yang, X. *et al.* A public genome-scale lentiviral expression library of human ORFs. *Nat Methods* 8, 659-661, doi:10.1038/nmeth.1638 (2011).

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

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

74 Vidal, M. A unifying view of 21st century systems biology. *FEBS Lett* 583, 3891-3894, doi:10.1016/j.febslet.2009.11.024 (2009).

75 Robinson, C. V., Sali, A. & Baumeister, W. The molecular sociology of the cell. *Nature* 450, 973-982, doi:10.1038/nature06523 (2007).

76 Barabasi, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 12, 56-68, doi:10.1038/nrg2918 (2011).

84 Das, J. *et al.* Cross-species protein interactome mapping reveals species-specific wiring of stress response pathways. *Sci Signal* 6, ra38, doi:10.1126/scisignal.2003350 (2013).

94 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* 28, 235-242, doi:10.1093/nar/28.1.235 (2000).

135 Wei, X. *et al.* A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. *PLoS Genet* 10, e1004819, doi:10.1371/journal.pgen.1004819 (2014).

219 Wierbowski, S. D. *et al.* A 3D structural SARS-CoV-2-human interactome to explore genetic and drug perturbations. *Nat Methods* 18, 1477-1488, doi:10.1038/s41592-021-01318-w (2021).

220 *COVID-19 Dashboard by the Center for Systems Science and Engineering at Johns Hopkins University*, <https://coronavirus.jhu.edu/map.html> (2020).

221 Fehr, A. R. & Perlman, S. Coronaviruses: an overview of their replication and pathogenesis. *Methods Mol Biol* 1282, 1-23, doi:10.1007/978-1-4939-2438-7_1 (2015).

222 Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579, 270-273, doi:10.1038/s41586-020-2012-7 (2020).

223 McIntosh, K. & Perlman, S. Coronaviruses, Including Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS). *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases*, 1928-1936.e1922, doi:10.1016/B978-1-4557-4801-3.00157-0 (2015).

224 Zhou, H. *et al.* A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. *Curr Biol* 30, 2196-2203 e2193, doi:10.1016/j.cub.2020.05.023 (2020).
225 Gupta, A. *et al.* Extrapulmonary manifestations of COVID-19. *Nat Med* 26, 1017-1032, doi:10.1038/s41591-020-0968-3 (2020).
226 Wang, D. *et al.* Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA* 323, 1061-1069, doi:10.1001/jama.2020.1585 (2020).
227 Yang, X. *et al.* Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *Lancet Respir Med* 8, 475-481, doi:10.1016/S2213-2600(20)30079-5 (2020).
228 Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* 395, 1054-1062, doi:10.1016/S0140-6736(20)30566-3 (2020).
229 Palaiodimos, L. *et al.* Severe obesity, increasing age and male sex are independently associated with worse in-hospital outcomes, and higher in-hospital mortality, in a cohort of patients with COVID-19 in the Bronx, New York. *Metabolism* 108, 154262, doi:10.1016/j.metabol.2020.154262 (2020).
230 Ferdinand, K. C. & Nasser, S. A. African-American COVID-19 Mortality: A Sentinel Event. *J Am Coll Cardiol* 75, 2746-2748, doi:10.1016/j.jacc.2020.04.040 (2020).
231 Killerby, M. E. *et al.* Characteristics Associated with Hospitalization Among Patients with COVID-19 - Metropolitan Atlanta, Georgia, March-April 2020. *MMWR Morb Mortal Wkly Rep* 69, 790-794, doi:10.15585/mmwr.mm6925e1 (2020).
232 Raisi-Estabragh, Z. *et al.* Greater risk of severe COVID-19 in Black, Asian and Minority Ethnic populations is not explained by cardiometabolic, socioeconomic or behavioural factors, or by 25(OH)-vitamin D status: study of 1326 cases from the UK Biobank. *J Public Health (Oxf)*, doi:10.1093/pubmed/fdaa095 (2020).
233 Moore, J. T. *et al.* Disparities in Incidence of COVID-19 Among Underrepresented Racial/Ethnic Groups in Counties Identified as Hotspots During June 5-18, 2020 - 22 States, February-June 2020. *MMWR Morb Mortal Wkly Rep* 69, 1122-1126, doi:10.15585/mmwr.mm6933e1 (2020).
234 Mahajan, U. V. & Larkins-Pettigrew, M. Racial demographics and COVID-19 confirmed cases and deaths: a correlational analysis of 2886 US counties. *J Public Health (Oxf)* 42, 445-447, doi:10.1093/pubmed/fdaa070 (2020).
235 Pfefferle, S. *et al.* The SARS-coronavirus-host interactome: identification of cyclophilins as target for pan-coronavirus inhibitors. *PLoS Pathog* 7, e1002331, doi:10.1371/journal.ppat.1002331 (2011).
236 Jager, S. *et al.* Global landscape of HIV-human protein complexes. *Nature* 481, 365-370, doi:10.1038/nature10719 (2011).
237 Batra, J. *et al.* Protein Interaction Mapping Identifies RBBP6 as a Negative Regulator of Ebola Virus Replication. *Cell* 175, 1917-1930 e1913,

doi:10.1016/j.cell.2018.08.044 (2018).
238 Shah, P. S. *et al.* Comparative Flavivirus-Host Protein Interaction Mapping Reveals Mechanisms of Dengue and Zika Virus Pathogenesis. *Cell* 175, 1931-1945 e1918, doi:10.1016/j.cell.2018.11.028 (2018).
239 Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 583, 459-468, doi:10.1038/s41586-020-2286-9 (2020).
240 Niemann, H. H. *et al.* Structure of the human receptor tyrosine kinase met in complex with the Listeria invasion protein InlB. *Cell* 130, 235-246, doi:10.1016/j.cell.2007.05.037 (2007).
241 Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* 181, 271-280 e278, doi:10.1016/j.cell.2020.02.052 (2020).
242 Xu, G. G., Guo, J. & Wu, Y. Chemokine receptor CCR5 antagonist maraviroc: medicinal chemistry and clinical applications. *Curr Top Med Chem* 14, 1504-1514, doi:10.2174/1568026614666140827143745 (2014).
243 Hayouka, Z. *et al.* Inhibiting HIV-1 integrase by shifting its oligomerization equilibrium. *Proc Natl Acad Sci U S A* 104, 8316-8321, doi:10.1073/pnas.0700781104 (2007).
244 Peat, T. S. *et al.* Small molecule inhibitors of the LEDGF site of human immunodeficiency virus integrase identified by fragment screening and structure based design. *PLoS One* 7, e40147, doi:10.1371/journal.pone.0040147 (2012).
245 Maginnis, M. S. Virus-Receptor Interactions: The Key to Cellular Invasion. *J Mol Biol* 430, 2590-2611, doi:10.1016/j.jmb.2018.06.024 (2018).
246 Daczkowski, C. M. *et al.* Structural Insights into the Interaction of Coronavirus Papain-Like Proteases and Interferon-Stimulated Gene Product 15 from Different Species. *J Mol Biol* 429, 1661-1683, doi:10.1016/j.jmb.2017.04.011 (2017).
247 Yao, J. *et al.* Mechanism of inhibition of retromer transport by the bacterial effector RidL. *Proc Natl Acad Sci U S A* 115, E1446-E1454, doi:10.1073/pnas.1717383115 (2018).
248 Zhang, L. *et al.* Solution structure of the complex between poxvirus-encoded CC chemokine inhibitor vCCI and human MIP-1beta. *Proc Natl Acad Sci U S A* 103, 13985-13990, doi:10.1073/pnas.0602142103 (2006).
249 Jonker, H. R., Wechselberger, R. W., Boelens, R., Folkers, G. E. & Kaptein, R. Structural properties of the promiscuous VP16 activation domain. *Biochemistry* 44, 827-839, doi:10.1021/bi0482912 (2005).
250 Card, G. L., Knowles, P., Laman, H., Jones, N. & McDonald, N. Q. Crystal structure of a gamma-herpesvirus cyclin-cdk complex. *EMBO J* 19, 2877-2888, doi:10.1093/emboj/19.12.2877 (2000).
251 Smith, M., Honce, R. & Schultz-Cherry, S. Metabolic Syndrome and Viral Pathogenesis: Lessons from Influenza and Coronaviruses. *J Virol* 94, doi:10.1128/JVI.00665-20 (2020).
252 Scott, D. E., Bayly, A. R., Abell, C. & Skidmore, J. Small molecules, big targets:

drug discovery faces the protein-protein interaction challenge. *Nat Rev Drug Discov* 15, 533-550, doi:10.1038/nrd.2016.29 (2016).

253 Arkin, M. R., Tang, Y. & Wells, J. A. Small-molecule inhibitors of protein-protein interactions: progressing toward the reality. *Chem Biol* 21, 1102-1114, doi:10.1016/j.chembiol.2014.09.001 (2014).

254 Rooklin, D., Wang, C., Katigbak, J., Arora, P. S. & Zhang, Y. AlphaSpace: Fragment-Centric Topographical Mapping To Target Protein-Protein Interaction Interfaces. *J Chem Inf Model* 55, 1585-1599, doi:10.1021/acs.jcim.5b00103 (2015).

255 Lampson, B. L. & Davids, M. S. The Development and Current Use of BCL-2 Inhibitors for the Treatment of Chronic Lymphocytic Leukemia. *Curr Hematol Malig Rep* 12, 11-19, doi:10.1007/s11899-017-0359-0 (2017).

256 *VENCLEXTA combination regimens for CLL work through 2 distinct cytotoxic mechanisms of action*, <https://www.venclextahcp.com/cll/venclexta-efficacy/mechanism-of-action.html> (2019).

257 Schormann, N. *et al.* Identification of protein-protein interaction inhibitors targeting vaccinia virus processivity factor for development of antiviral agents. *Antimicrob Agents Chemother* 55, 5054-5062, doi:10.1128/AAC.00278-11 (2011).

258 White, P. W. *et al.* Inhibition of human papillomavirus DNA replication by small molecule antagonists of the E1-E2 protein interaction. *J Biol Chem* 278, 26765-26772, doi:10.1074/jbc.M303608200 (2003).

259 Goudreau, N. *et al.* Optimization and determination of the absolute configuration of a series of potent inhibitors of human papillomavirus type-11 E1-E2 protein-protein interaction: a combined medicinal chemistry, NMR and computational chemistry approach. *Bioorg Med Chem* 15, 2690-2700, doi:10.1016/j.bmc.2007.01.036 (2007).

260 Brito, A. F. & Pinney, J. W. Protein-Protein Interactions in Virus-Host Systems. *Front Microbiol* 8, 1557, doi:10.3389/fmicb.2017.01557 (2017).

261 Meyer, M. J. *et al.* Interactome INSIDER: a structural interactome browser for genomic studies. *Nat Methods* 15, 107-114, doi:10.1038/nmeth.4540 (2018).

262 Dominguez, C., Boelens, R. & Bonvin, A. M. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* 125, 1731-1737, doi:10.1021/ja026939x (2003).

263 van Zundert, G. C. P. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol* 428, 720-725, doi:10.1016/j.jmb.2015.09.014 (2016).

264 Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* 26, 689-691, doi:10.1093/bioinformatics/btq007 (2010).

265 Kirchdoerfer, R. N. *et al.* Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. *Sci Rep* 8, 15701, doi:10.1038/s41598-018-34171-7 (2018).

266 Wang, Q. *et al.* Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. *Cell* 181, 894-904 e899, doi:10.1016/j.cell.2020.03.045 (2020).

267 Wrobel, A. G. *et al.* SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. *Nat Struct Mol Biol* 27, 763-767, doi:10.1038/s41594-020-0468-7 (2020).
268 Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* 181, 281-292 e286, doi:10.1016/j.cell.2020.02.058 (2020).
269 Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J Chem Theory Comput* 13, 3031-3048, doi:10.1021/acs.jctc.7b00125 (2017).
270 Shang, J. *et al.* Structural basis of receptor recognition by SARS-CoV-2. *Nature* 581, 221-224, doi:10.1038/s41586-020-2179-y (2020).
271 Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 367, 1260-1263, doi:10.1126/science.abb2507 (2020).
272 Jordan, R. E. & Adab, P. Who is most likely to be infected with SARS-CoV-2? *The Lancet Infectious Diseases* 20, 995-996, doi:10.1016/s1473-3099(20)30395-9 (2020).
273 Cao, Y. *et al.* Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations. *Cell Discov* 6, 11, doi:10.1038/s41421-020-0147-1 (2020).
274 Darbeheshti, F. & Rezaei, N. Genetic predisposition models to COVID-19 infection. *Med Hypotheses* 142, 109818, doi:10.1016/j.mehy.2020.109818 (2020).
275 Zhao, Y. *et al.*, doi:10.1101/2020.01.26.919985 (2020).
276 Kortemme, T. & Baker, D. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A* 99, 14116-14121, doi:10.1073/pnas.202485799 (2002).
277 Shulman-Peleg, A., Shatsky, M., Nussinov, R. & Wolfson, H. J. Spatial chemical conservation of hot spot interactions in protein-protein complexes. *BMC Biol* 5, 43, doi:10.1186/1741-7007-5-43 (2007).
278 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434-443, doi:10.1038/s41586-020-2308-7 (2020).
279 Stawiski, E. W. *et al.* Human ACE2 receptor polymorphisms predict SARS-CoV-2 susceptibility. *bioRxiv*, doi:10.1101/2020.04.07.024752 (2020).
280 Procko, E. The sequence of human ACE2 is suboptimal for binding the S spike protein of SARS coronavirus 2. *bioRxiv*, doi:10.1101/2020.03.16.994236 (2020).
281 Pieper, U. *et al.* ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* 42, D336-346, doi:10.1093/nar/gkt1144 (2014).
282 Guharoy, M. & Chakrabarti, P. Conserved residue clusters at protein-protein interfaces and their use in binding site identification. *BMC Bioinformatics* 11, 286, doi:10.1186/1471-2105-11-286 (2010).
283 Gupta, R. *et al.* SARS-CoV2 (COVID-19) Structural/Evolution Dynamicome: Insights into functional evolution and human genomics. *bioRxiv*,

doi:10.1101/2020.05.15.098616 (2020).
284 Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* 21, 577-581, doi:10.1002/humu.10212 (2003).
285 Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 46, D1062-D1067, doi:10.1093/nar/gkx1153 (2018).
286 Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 47, D1005-D1012, doi:10.1093/nar/gky1120 (2019).
287 Killerby, M. E. *et al.* Characteristics Associated with Hospitalization Among Patients with COVID-19 - Metropolitan Atlanta, Georgia, March-April 2020. *MMWR Morb Mortal Wkly Rep* 69, 790-794, doi:10.15585/mmwr.mm6925e1 (2020).
288 Yang, J. *et al.* Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis. *Int J Infect Dis* 94, 91-95, doi:10.1016/j.ijid.2020.03.017 (2020).
289 Sahni, N. *et al.* Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* 161, 647-660, doi:10.1016/j.cell.2015.04.013 (2015).
290 Wang, X. *et al.* Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat Biotechnol* 30, 159-164, doi:10.1038/nbt.2106 (2012).
291 Sim, N. L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* 40, W452-457, doi:10.1093/nar/gks539 (2012).
292 Braun, P. *et al.* An experimentally derived confidence score for binary protein-protein interactions. *Nat Methods* 6, 91-97, doi:10.1038/nmeth.1281 (2009).
293 Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* 580, 402-408, doi:10.1038/s41586-020-2188-x (2020).
294 Gordon, D. E. *et al.* Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. *Science* 370, doi:10.1126/science.abe9403 (2020).
295 Fellay, J. *et al.* A whole-genome association study of major determinants for host control of HIV-1. *Science* 317, 944-947, doi:10.1126/science.1143767 (2007).
296 Hong, H. Q. *et al.* G3BP2 is involved in isoproterenol-induced cardiac hypertrophy through activating the NF-kappaB signaling pathway. *Acta Pharmacol Sin* 39, 184-194, doi:10.1038/aps.2017.58 (2018).
297 Lu, S. *et al.* The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. *Nat Commun* 12, 502, doi:10.1038/s41467-020-20768-y (2021).
298 Nabeel-Shah, S. *et al.* Nucleus-specific linker histones Hho1 and Mlh1 form distinct protein interactions during growth, starvation and development in Tetrahymena thermophila. *Sci Rep* 10, 168, doi:10.1038/s41598-019-56867-0 (2020).
299 Koes, D. R., Baumgartner, M. P. & Camacho, C. J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J Chem Inf*

Model 53, 1893-1904, doi:10.1021/ci300604z (2013).
300 Wierbowski, S. D., Wingert, B. M., Zheng, J. & Camacho, C. J. Cross-docking benchmark for automated pose and ranking prediction of ligand binding. *Protein Sci* 29, 298-305, doi:10.1002/pro.3784 (2020).
301 Crouse, A. B. *et al.* Metformin Use Is Associated With Reduced Mortality in a Diverse Population With COVID-19 and Diabetes. *Front Endocrinol (Lausanne)* 11, 600439, doi:10.3389/fendo.2020.600439 (2020).
302 *Silmitasertib (CX-4945) in Patients With Severe Coronavirus Disease 2019 (COVID-19) (CX4945)*, <https://clinicaltrials.gov/ct2/show/NCT04668209> (2020).
303 *Valproate Alone or in Combination With Quetiapine for Severe COVID-19 Pneumonia With Agitated Delirium*, <https://clinicaltrials.gov/ct2/show/NCT04513314> (2020).
304 McIver, E. G. *et al.* Synthesis and structure-activity relationships of a novel series of pyrimidines as potent inhibitors of TBK1/IKKepsilon kinases. *Bioorg Med Chem Lett* 22, 7169-7173, doi:10.1016/j.bmcl.2012.09.063 (2012).
305 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*, doi:10.1038/s41586-021-03819-2 (2021).
306 Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature*, doi:10.1038/s41586-021-03828-1 (2021).
307 Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373, 871-876, doi:10.1126/science.abj8754 (2021).
308 Gainza, P. *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods* 17, 184-192, doi:10.1038/s41592-019-0666-6 (2020).
309 Eswar, N. *et al.* Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics* Chapter 5, Unit-5 6, doi:10.1002/0471250953.bi0506s15 (2006).
310 Mosca, R., Ceol, A. & Aloy, P. Interactome3D: adding structural details to protein networks. *Nat Methods* 10, 47-53, doi:10.1038/nmeth.2289 (2013).
311 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* 215, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).
312
313 Gasteiger, E. *et al.* Protein Identification and Analysis Tools on the ExPASy Server. *(In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press (2005)*, 571-607 (2005).
314 Lin, J. H. Divergence Measures Based on the Shannon Entropy. *Ieee Transactions on Information Theory* 37, 145-151, doi:Doi 10.1109/18.61115 (1991).
315 Capra, J. A. & Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* 23, 1875-1882, doi:10.1093/bioinformatics/btm270 (2007).
316 Morcos, F., Hwa, T., Onuchic, J. N. & Weigt, M. Direct coupling analysis for

protein contact prediction. *Methods Mol Biol* 1137, 55-70, doi:10.1007/978-1-4939-0366-5_5 (2014).
317 Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286, 295-299, doi:10.1126/science.286.5438.295 (1999).
318 Lee, B. & Richards, F. M. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* 55, 379-400, doi:10.1016/0022-2836(71)90324-x (1971).
319 Pierce, B. G., Hourai, Y. & Weng, Z. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PLoS One* 6, e24657, doi:10.1371/journal.pone.0024657 (2011).
320 Rodrigues, J. P. *et al.* Defining the limits of homology modeling in information-driven protein docking. *Proteins* 81, 2119-2128, doi:10.1002/prot.24382 (2013).
321 UniProt, C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 47, D506-D515, doi:10.1093/nar/gky1049 (2019).
322 He, R. *et al.* Analysis of multimerization of the SARS coronavirus nucleocapsid protein. *Biochem Biophys Res Commun* 316, 476-483, doi:10.1016/j.bbrc.2004.02.074 (2004).
323 Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265-269, doi:10.1038/s41586-020-2008-3 (2020).
324 Chan, J. F. *et al.* Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg Microbes Infect* 9, 221-236, doi:10.1080/22221751.2020.1719902 (2020).
325 Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48, 443-453, doi:10.1016/0022-2836(70)90057-4 (1970).
326 Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 89, 10915-10919, doi:10.1073/pnas.89.22.10915 (1992).
327 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* 17, 122, doi:10.1186/s13059-016-0974-4 (2016).
328 Szumilas, M. Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry* 19, 227-229 (2010).
329 Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* 31, 3555-3557, doi:10.1093/bioinformatics/btv402 (2015).
330 O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J Cheminform* 3, 33, doi:10.1186/1758-2946-3-33 (2011).