LINKING HUMAN-GENETIC AND HOST-MICROBIOME ASSOCIATIONS TO MOLECULAR MECHANISMS USING PROTEIN-PROTEIN INTERACTION NETWORKS

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
Juan Felipe Beltrán Lacouture
December 2019

LINKING HUMAN-GENETIC AND HOST-MICROBIOME ASSOCIATIONS TO MOLECULAR MECHANISMS USING PROTEIN-PROTEIN INTERACTION NETWORKS

Juan Felipe Beltrán Lacouture, Ph. D
Cornell University 2019

We continue to better understand human disease through the study of human genetics and the human microbiome. Both fields use cohort study design, where the genes or microbiome communities of patients with a given condition are compared to a group of healthy controls. Systematically performing these comparisons to derive gene-disease associations and microbiome-disease associations is increasingly commonplace, but generating reasonable hypotheses for further study is still a challenge. Overcoming this challenge is vital to understand the causal molecular mechanisms of human disease.

Quality-control and prioritization pipelines for human genetic variants from large-scale studies are hard to build, often involving several computational tools, databases, and tuning parameters. In **Chapter 2**, I present GeMSTONE, an online variant prioritization tool that allows researchers to leverage a large variety of resources to replicate and customize

these pipelines without any computational experience or overhead.

Understanding the location of disease-associated variants relative to protein interaction interfaces can help us understand whether they are likely to disrupt a protein interaction, thereby implicating a discrete molecular phenotype with the disease. In **Chapter 3**, I present Interactome INSIDER, an online resource that expands our network of protein interaction interfaces and performs widespread annotation of disease variants in this context.

Determining the mechanisms behind host-microbiome disease-associations is particularly challenging, as there are few functional annotations for commensal microbiome proteins as well as sparse microbe-human protein interaction networks not involving extreme pathogens. In **Chapter 4**, I build a human-bacteria protein-protein interaction network that is used to detect the differential targeting of human proteins by commensal bacteria in association with disease.

This thesis presents three different examples of mechanistic hypothesis generation from large-scale association studies. I demonstrate how annotation of variants using biological databases, structural interaction networks, and bacteria-human protein interactions can expand our understanding of the likely actors in human disease.

BIOGRAPHICAL SKETCH

Juan Felipe Beltrán Lacouture was born in Bogotá, Colombia, in 1991 and attended the San Carlos School, an all-boys bilingual school under the auspice of Benedictine Monks, until age 12. After six years of marching in height-sorted lines to church, he transitioned to the nature-loving International School Nido de Aguilas in Santiago, Chile. In Chile, Juan Felipe overdeveloped a passion for debate, theater, stand-up comedy, neuroscience, music, astronomy, etymology, and overscheduling. His grades and teacher evaluations strongly reflected the last.

For his undergraduate education, Juan Felipe was elated to be part of the 2010 inaugural class of 146 students at New York University in Abu Dhabi, where he debated between a Theater major or a Neuroscience and Engineering double-major. Juan Felipe ultimately graduated with a Bachelor of Science in Computer Science, having met lifelong friends in the students, faculty, and staff of this new institution.

Under the mentorship of Dr. Godfried Toussaint, Dr. Dennis Shasha, and Dr. Azza Abouzied, Juan published work on rhythm analysis, Alzheimer's biomarkers, and human-computer interaction. After one year in Abu Dhabi working as a research assistant, Juan Felipe moved to Ithaca to

work in Dr. Haiyuan Yu's lab, where he studied protein-protein interaction networks. He was later admitted to the graduate program of Computational Biology at Cornell University, where he found a valuable mentor in Dr. Ilana Brito and studied the human microbiome as a student in her lab.

In the future, Juan Felipe is excited to catalyze the conversation between computer science and biology to understand the molecular mechanisms of human disease. He might also switch back to theater if previous data is any indication.

To my chosen family

ACKNOWLEDGMENTS

If I had to choose two feelings to describe my childhood, they would be loneliness and the sense that American science and opportunity were impossible for me to reach. I have been unbelievably fortunate to find these to be non-issues in my adult life due to the loving intervention of incredible individuals and institutions. This thesis is a product of half a decade of graduate school on the shoulders of twenty years of support.

I want to thank Paula Ottoni and Paula Arismendi, my oldest friends. I would never have been able to pursue this path without their constant support, the shared struggle and celebration, and the sense of family that I felt, for the first time, through both of them. I'm also indebted to Rocio Arismendi and Andrea Jadresic, who shared their space with me and made me feel part of their home. I strive to be a person they can be proud of.

Many thanks to my high school teachers, who went above and beyond to nurture a question-loving, if inattentive, student. Specifically, I would like to thank Kara Wiley, Rick Vezzoli, Linda Duggan. Their patience and love for life is contagious far past their tenure teaching me, and I think of them anytime I get to teach.

I am extremely grateful to John Sexton and His Highness Sheikh Mohammed bin Zayed bin Sultan Al Nahyan for the gift of NYU Abu Dhabi. I owe a profound debt of gratitude to NYUAD as an institution, not only for bankrolling a diverse undergraduate education in Abu Dhabi, New York, London, and Florence, but also bringing together the most spectacular group of faculty and students I've had the pleasure of meeting. Professors Joseph Gelfand and Ingyin Zaw were my first teachers in programming, physics, and stellar academic mentorship. Dean Dave Scicchitano injected scientific fire and wit into his discussions with students with a gusto I can't help but try to emulate. Professor Sana Odeh understood the impact of teaching computer science to those with less privilege, and she will forever have my thanks for teaching me my first line of Python. Finally, I cannot begin to express my thanks to Professors Dennis Shasha and Godfried Toussaint, who, both titans of Computer Science in my eyes, were quick to call me their colleague. I owe every bit of my confidence as an academic to this institution and community.

I am deeply indebted to my all of my college friends for this continued sense of community, even in the face of distance. I am privileged to enjoy the companionship of Cassandra Flores and James Hosken, who both exemplify steadfast friendship and a passionate pursuit of happiness. My life and work are richer for having them as role-models in craft and care. James Lloyd and Layla Al Neyadi are incredible human beings that have filled my past nine years with laughter, and are active participants in the balancing act of taking myself more seriously, and less seriously. I would also like to thank Sophie Feiertag, who manages to be a ball of sunshine even when she feels otherwise. It is rare to find someone so cognizant of other people's stress, and so willing to share her life, family, and holidays with those around her. Brianna Haining, thank you for being a positive, genuine friend and a partner in the journey through graduate school. Stephen Underwood, Eric Johnson, Seung Man Oh, Nick Scoulios, and Katy Blumer: Thank you for sharing your lives and work with me, and sharing in my life and work. I am also very grateful to Jack Dickson, Amelia Kahn, and Maria Paula Soto, who have each been incredible caretakers at different times during my PhD.

I first arrived at Cornell University as an unpaid intern, building skills for my graduate school applications. I would never have been able to come without the support of Florencia and Adam Dolan-Schlamp, who were already pursuing their PhD's at this institution. Flor and Adam gave me a roof over my head and a couch on which to ride out one of the scariest transitions in my life, and I owe them for my life in Ithaca and eventual admission to the PhD program. I deeply treasure my friendship with each of them, and I'm glad to have shared one more adventure with them.

At Cornell, I have been fortunate to find extraordinary friends and mentors. Michael Meyer and Jishnu Das made quick work of teaching me the ropes, wielding patience, impatience, and tireless good humor. I would like to thank Michael in particular for sharing his passion for problem-solving with me, for the late nights of work, Seinfeld script writing, and one awkward bro hug. Michael and Sarai's new home in Connecticut has become a mythical place to me, that I can't help to think of, and smile. I am grateful to have worked with fantastic lab mates throughout, and would like to thank Jin, Ting-Yi, Robert, Siwei, Charles, Ana Maria, Felicia, Albert, Peter, Sarah, Hao, Sharon, Hector, Josh, Rowan, Alyssa, Taylor, and Ian Rose for their fellowship and general chumminess. I would also like to thank Ian Caldas and Shayne Wierbowski for being fantastic cohort-mates and making my

PhD a brighter experience. Additionally, I would like to thank Roman Spektor for an endless number of crash-courses in biology, coffee breaks, workouts, thanksgivings, paper edits, career discussions, and great waffles.

I would like to express my deepest appreciation to my committee, Ilana Brito, Andy Clark, and Giles Hooker, for their advice and support through the PhD process. I'd also like to recognize the great impact that Ilana Brito, Haiyuan Yu, and Andy Clark had in my graduate career, not just through their mentorship an insight, but by creating three unique labs whose members I hope to repeatedly encounter in my career. I'm very grateful for Ilana's passionate drive, which has made my immersion in the microbiome field as intensive as it is enjoyable. I'm also very grateful for the interns I was able to work with – Allison Greenberg, Jeffrey Page, Jay Chia, Jee Won Yang, Adam Ingber, Aaron Rumack, Yaoda Zhou, and Heejung Moon. Working on research with all of these people has brought me great joy. I would also like to thank my brother William, whose progress in his own educational career has been thrilling to follow.

The pursuit of this degree and the resulting thesis has been one of the greatest challenges I've experienced in my life. I am thankful for all the knowledge that has been passed on to me, and the communities that I have been able to access through graduate school. I am grateful that I've been able to live in the beautiful city of Ithaca, and experience its gorgeous gorges and Big Red autumns. In my time here, I am happy to have found my life-partner in Jenny Peet, who has been my rock when I've needed help the most. I could not be more thankful for the loving partner I have found in Jenny. Over three years, we have commuted between Syracuse and Ithaca, then New York City and Ithaca, one weekend at a time. After more than 400 bus-hours each, I'm confident there is no one I would rather share my life with. I'm at the edge of my seat wondering where we will go next, together.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

A common aspiration in genetic association studies is to illuminate the causal chain of events that lead from a given genotype to an observed phenotype. Constant progress in DNA sequencing technology continues to lower the cost and increase the frequency of genetic screens and cohort studies but, paradoxically, the number of association-based hypotheses keeps increasing and the mechanistic understanding of association remains an exceptional event. This struggle is not purely driven by the complexity of the biological systems under study, or the inherent time and effort required to validate a hypothesized mechanism: A key limiter in mechanism discovery is the challenge of prioritizing which hypotheses, out of millions, should be tested first - and how.

Annotating genes and variants with their biological context can help us interpret association results and find the most promising avenues for further study. DNA-level context, like allele frequency in human populations or conservation across species can seed an understanding of the evolutionary pressures at a locus and answer whether the rarity of the variant matches the impact of the phenotype. Additionally, protein-level context opens the door

to the largest variety of functional annotation available: Is this variant in a protein-coding sequence? Does it correspond to an amino acid substitution and what type? In which tissues is the protein expressed[1], and what other molecules does it bind or co-express with?

Several tools have been developed to address quality-control, prioritization, and functional annotation of both variants and genes to aid in answering questions like these. In **Chapter 2** of this thesis I present GeMSTONE, a tool to build custom, reproducible pipelines to process data from genomic cohort studies to identify variants that have a higher likelihood of being directly linked to disease, and that might make the best candidates for experimental validation. A GeMSTONE pipeline can filter out sequencing artifacts, select only variants that follow specific inheritance models given patient pedigrees, and further filter or annotate each variant using allele frequency databases, transcript metadata, functional predictions, conservation scores, gene ontology, disease databases, protein-protein interactions, pathway annotation, and tissue expression. This is all performed through a web interface that requires no programming knowledge, and generates an easily-shareable, versioned "recipe file" that can be used to replicate or modify any previous workflow.

[1] Protein expression is most often measured through RNA transcript quantification as a proxy for protein abundance. Tissue-specific expression and gene coexpression could reasonably be considered to be a third, RNA-level context.

A non-trivial challenge after identifying and annotating high-confidence variants is defining a contained system in which to conduct experimental validation. Even given a specific tissue, protein, and variant of interest, determining which phenotype to measure or which assay to perform is hardly straightforward - individual proteins can be involved in several different pathways, and have a variety of binding partners. A single amino-acid substitution can disrupt the interaction between only a subset of a protein's binding partners, leaving interactions at other sites of the protein intact, and thus disrupt only some of its associated functions (ZHONG *et al.* 2009). This selective disruption is broadly explained through partner-specific protein interaction interfaces: The specific residues of two protein partners that are directly involved in binding. Interaction interfaces are rich areas to search for simple molecular hypotheses, as they have been shown to have functional specificity (JOHNSON and HUMMER 2013) and a close relationship with disease mutations, where those on the same interface are more likely to cause the same disease than those otherwise distributed across a protein (WANG *et al.* 2012). Understanding the protein interaction interfaces that candidate variants may lie on could allow us to pick a specific set of interaction partners and pathway effects to test.

In order to understand genetic variants in the context of protein-protein interaction interfaces we need to expand our current structural

interactome through computational methods, as the experimentally-determined structural interactome is sparse and difficult to expand. This is due to cocrystallization's status as the primary source of experimental interface data, as it provides extremely high-quality information, but at extremely low throughput. In **Chapter 3** of this thesis, I present Interactome INSIDER, a structural interactome resource that combines experimentally determined, homology-modeled, and predicted interaction interfaces for all remaining protein-protein interactions in human and seven model organisms. INSIDER predictions are carried out using a novel machine learning model that utilizes structural and evolutionary features, but can still predict when these features are missing. Previous sequence-only predictions tend to have full-coverage and low performance, while methods requiring that both partners have a structure boast great performance but cover very few protein-interaction pairs in the network. INSIDER's prediction model covers all protein-protein pairs with the high-quality prediction possible given available data.

The resulting human structural interaction network can be accessed through a web interface to interactively explore sets of genetic variants in the context of the structural protein-protein interaction network. Mutations at the interface residues in this network are more likely disruptors of protein interaction, making them viable candidates for *in vitro* molecular follow-up of

novel associated variants. Sets of interacting proteins exhibiting an enrichment of disease-associated variants at their specific interaction interfaces can then be used to pick a molecular pathway or function to further study *in situ* or *in vivo*.

We can see that understanding the protein interaction machinery of the cell can clearly expand our mechanistic understanding of health and homeostasis in human and other organisms. The abundance and variation in members of this network act as dials that tune a variety of cellular functions, spanning metabolic, immune, and differentiation pathways among others. It is then particularly interesting to see the ways in which other organisms can add their own members to our protein interaction networks. *Chlamydia trachomatis* can secrete a tail-specific protease (CT441) that selectively cleaves a human protein, p65, preventing the activation of the NF-κB pathway, and thus preventing apoptosis of the infected cell – allowing the bacterium to replicate (LAD *et al.* 2007). *Helicobacter pylori* secretes a protein, vacuolating cytotoxin (VacA), which binds to the human protein tyrosine phosphatase receptor type Z (Ptprz) triggering a signaling pathway causing severe gastritis (FUJIKAWA *et al.* 2003). A more positive example, secreted protein Amuc_1100 from commensal *Akkermansia muciniphila* binds to human Toll-Like Receptor 2 (TLR2), with observable benefits against obesity, insulin resistance, and gut barrier alteration (PLOVIER *et al.* 2017). In the case of

Amuc_1100 and VacA, these effects persist without the administration of a live bacterium. In fact, VacA both triggers Ptprz signaling and enters human cells, forming vacuoles, when administered intragastrically. These studies are commonplace in pathogens and have increased in scale through experimentally resolved host-pathogen protein-protein interaction networks (DYER *et al.* 2010) as well as the large-scale prediction of structural interaction networks (GUVEN-MAIOROV *et al.* 2019). However, protein-protein interactions between commensal bacteria and their human hosts remain woefully understudied. There are currently no high-throughput interaction studies between commensal and very few computational predictions (COELHO *et al.* 2014).

The human microbiota serves a crucial role in host development and health, and has been implicated in a wide variety of diseases. Large-scale association studies are swiftly moving from taxon-level associations through 16S rRNA gene sequencing to bacterial gene-specific disease associations using whole metagenome shotgun sequencing. In either scenario, a familiar problem surfaces in microbiome studies: Given the heterogeneity of function in both bacterial taxa and the various families of bacterial genes, what steps can be taken to mold these associations into testable mechanistic hypotheses?

In **Chapter 4** of this thesis I present a method to identify human-targeting bacterial proteins associated with disease. To exploit the similarity between commensal bacterial proteins, which tend to lack interaction information, and well-studied exogenous interactions, I build a homology-modeled microbe-human protein-protein interaction network. Given a sequence library of bacterial proteins present in this network, putative bacteria-human interactions can be trivially detected in any metagenomic dataset. I apply this method to nine metagenomic cohort studies for colorectal cancer, inflammatory bowel disease, type 2 diabetes, and obesity, comparing the detected bacteria-human interactions between case/control cohorts and thus measuring the differential targeting of human proteins by bacteria-derived proteins in disease. These differentially targeted human proteins are enriched in pathways involving the immune system, apoptosis, oncogenesis, and endocrine signaling, revealing a complex network of disease-associated protein interactions between human proteins and commensal-microbe proteins. This approach can be applied to any metagenomic cohort study, and has a great potential to characterize the mechanisms of host-microbe interactions, and identifying novel drug targets and therapeutics through molecular, and systems-level, testable hypotheses.

In this thesis I will cover the automation and customization of variant prioritization pipelines (**Chapter 2**), the prediction of interface residues to

understand disease mutations in a full structural interaction network (**Chapter 3**), and the construction of a host-microbe protein-protein interaction network to understand the role of bacterial proteins in human health (**Chapter 4**).

REFERENCES

Coelho E. D., J. P. Arrais, S. Matos, C. Pereira, N. Rosa, *et al.*, 2014 Computational prediction of the human-microbial oral interactome. BMC Syst Biol 8: 24. https://doi.org/10.1186/1752-0509-8-24

Dyer M. D., C. Neff, M. Dufford, C. G. Rivera, D. Shattuck, *et al.*, 2010 The Human-Bacterial Pathogen Protein Interaction Networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis. PLOS ONE 5: e12089. https://doi.org/10.1371/journal.pone.0012089

Fujikawa A., D. Shirasaka, S. Yamamoto, H. Ota, K. Yahiro, *et al.*, 2003 Mice deficient in protein tyrosine phosphatase receptor type Z are resistant to gastric ulcer induction by VacA of Helicobacter pylori. Nat. Genet. 33: 375–381. https://doi.org/10.1038/ng1112

Guven-Maiorov E., C.-J. Tsai, B. Ma, and R. Nussinov, 2019 Interface-Based Structural Prediction of Novel Host-Pathogen Interactions. Methods Mol. Biol. 1851: 317–335. https://doi.org/10.1007/978-1-4939-8736-8_18

Johnson M. E., and G. Hummer, 2013 Interface-Resolved Network of Protein-Protein Interactions. PLOS Computational Biology 9: e1003065. https://doi.org/10.1371/journal.pcbi.1003065

Lad S. P., J. Li, J. da Silva Correia, Q. Pan, S. Gadwal, *et al.*, 2007 Cleavage of p65/RelA of the NF-κB pathway by Chlamydia. Proc Natl Acad Sci U S A 104: 2933–2938. https://doi.org/10.1073/pnas.0608393104

Plovier H., A. Everard, C. Druart, C. Depommier, M. Van Hul, *et al.*, 2017 A purified membrane protein from *Akkermansia muciniphila* or the pasteurized bacterium improves metabolism in obese and diabetic mice. Nature Medicine 23: 107–113. https://doi.org/10.1038/nm.4236

Wang X., X. Wei, B. Thijssen, J. Das, S. M. Lipkin, *et al.*, 2012 Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol 30: 159–164. https://doi.org/10.1038/nbt.2106

Zhong Q., N. Simonis, Q.-R. Li, B. Charloteaux, F. Heuze, *et al.*, 2009 Edgetic perturbation models of human inherited disorders. Mol. Syst. Biol. 5: 321. https://doi.org/10.1038/msb.2009.80

CHAPTER 2

EASY AND REPRODUCIBLE PRIORITIZATION OF GENETIC VARIANTS ASSOCIATED WITH HUMAN DISEASE[1]

ABSTRACT

Integrative analysis of whole-genome/exome-sequencing data has been challenging, especially for the non-programming research community, as it requires simultaneously managing a large number of computational tools. Even computational biologists find it unexpectedly difficult to reproduce results from others or optimize their strategies in an end-to-end workflow. We introduce **Ge**rmline **M**utation **S**coring **T**ool f**O**r **N**ext-generation s**E**quencing data (GeMSTONE), a cloud-based variant prioritization tool with high-level customization and a comprehensive collection of bioinformatics tools and data libraries (**http://gemstone.yulab.org/**). GeMSTONE generates and readily accepts a shareable "recipe" file for each run to either replicate previous results or

[1] Published as: *Chen, S., *Beltrán, J. F., Esteban-Jurado, C., Franch-Expósito, S., Castellví-Bel, S., Lipkin, S., Wei, X., Yu, H. (2017). GeMSTONE: orchestrated prioritization of human germline mutations in the cloud. *Nucleic Acids Research*, *45*(W1), W207–W214.

* These authors contributed equally to the work.

analyze new data with identical parameters, and provides a centralized workflow for prioritizing germline mutations in human disease within a streamlined workflow rather than a pool of program executions.

INTRODUCTION

Next-generation sequencing (NGS) has significantly reduced the cost of obtaining genomic data for increasingly large sample sizes (METZKER 2010), facilitating discovery of causal genes and mutation candidates for various disorders (BOYCOTT *et al.* 2013) and providing sizable genetic variant datasets (NEKRUTENKO AND TAYLOR 2012). As a result, the process of filtering, annotating, and prioritizing variants from large-scale studies has grown in complexity and computational burden. It has become increasingly difficult to organize, maintain, and standardize the variant analysis workflows, increasing the time and monetary investment for less computationally oriented biologists and labs. Some integrative frameworks (REICH *et al.* 2006; GOECKS *et al.* 2010; LUSHBOUGH *et al.* 2010; HALBRITTER *et al.* 2011) have been developed to enhance the reproducibility and accessibility of NGS studies. This same initiative inspired the framework for GeMSTONE: recording all analysis metadata for reproducible computational experiments, specifically focusing on germline mutation prioritization in human disease.

Although other platforms bring together different bioinformatics tools and allow users to schedule their analyses online, none of them are built with an emphasis on streamlined single-run scheduling and automatic fetching of the necessary supplementary public data. Platforms like Galaxy (GOECKS *et al.* 2010), for instance, allow the user to combine many different tools from an impressive catalog, but require the user to reformat their data depending on the particular inputs of the database or tool that they want to add to their analysis. A major design goal in the development of GeMSTONE is the ability to maximize customization for studies in a streamlined workflow rather than a pool of program executions. Within the GeMSTONE interface, databases required by the user-selected tools are pre-loaded, and the user-inputted data will be automatically reshaped to fit query requirements. Therefore, adding an extra layer of analysis to any workflow requires minimal effort.

There is a large research community focusing on genetic variation study relating to human disease (MACKAY *et al.* 2014; EINARSDOTTIR *et al.* 2015; ESTEBAN-JURADO *et al.* 2015; MACKAY *et al.* 2015; RADOVICA-SPALVINA *et al.* 2015; WRIGHT *et al.* 2015; BELLIDO *et al.* 2016; FARLOW *et al.* 2016; MEDEIROS *et al.* 2016; BAILEY *et al.* 2017; COX *et al.* 2017). This community often performs their analysis in-house rather than using any of the currently available tools for variant analysis. GeMSTONE facilitates the

process of integrating and assessing evidence for causal inferences while automating the whole workflow in a reproducible way. Through its design GeMSTONE fills in a significant gap in the online analysis landscape. For the same amount of work, users can customize their workflows to be either fast and cursory, or longer and exploratory. GeMSTONE provides centralized workflows: embedding key features of variant prioritization for DNA sequencing data, focused on but not limited to germline mutations, with a collection of current bioinformatics tools and data libraries in a highly-customizable and reproducible manner. In short, we created GeMSTONE to organize, schedule, document, and reproduce our variant analysis workflows from a single interface.

We show the GeMSTONE workflow is consistent with consensus guidelines for interpreting sequence variants in human disease (MACARTHUR *et al.* 2014; RICHARDS *et al.* 2015) (**Supplementary Table S2.1**). A demo study is fully described and explained as it is designed, scheduled and analyzed through the chained GeMSTONE functionalities (http://gemstone.yulab.org/manual.html); we also demonstrate its feasibility and efficiency in a proof-of-concept case by recapitulating results of a published variant analysis (ESTEBAN-JURADO *et al.* 2015).

MATERIALS & METHODS

GeMSTONE serves as an online variant prioritization framework that leverages 7 popular bioinformatics suites [VT (TAN *et al.* 2015), VCFtools (DANECEK *et al.* 2011), BCFtools (LI *et al.* 2009), SnpEff (CINGOLANI *et al.* 2012), GEMINI (PAILA *et al.* 2013), dbNSFP (LIU *et al.* 2011), and PLINK/SEQ PLINK/SEQ (2014)] in connection to 46 meta-information and prediction resources (**Figure 2.1; Supplementary Table S2.2**) to provide a smooth, customizable workflow for variant analysis.

Users of the GeMSTONE web portal can customize their analyses of genomic data from Variant Call Format (VCF) files by using tools from a range of different classes (**Figure 2.1**). These include

1) variant normalization for unified representation of genetic variants using VT,

2) variant/genotype quality filters on matrices encoded in the VCF file such as QUAL (Phred-scaled quality score), GQ (genotype quality), DP (read depth) and filter status using VCFtools,

3) variant type filters on variant consequence and transcript biotype based on SnpEff annotations,

4) common variant filter on allele frequency in the general population [ExAC (CONSORTIUM 2015), 1000 Genomes (GENOMES PROJECT *et al.* 2012), ESP6500 (FU *et al.* 2013), and TAGC (CARMI *et al.* 2014)],

5) variant function filters on predicted damaging effects [18 methods (e.g., Polyphen-2 (ADZHUBEI *et al.* 2010), SIFT (POLLARD *et al.* 2010), CADD (KIRCHER *et al.* 2014)) complied in dbNSFP (Supplementary Table S2), Rosetta ddG (ROHL *et al.* 2004)] and protein domains [Pfam (FINN *et al.* 2014)],

6) comprehensive annotations (and filters) on gene and gene product attributes [Gene Ontology (ASHBURNER *et al.* 2000)], biological pathways [KEGG (KANEHISA *et al.* 2014), BioCarta, and Reactome (FABREGAT *et al.* 2016) compiled in MSigDB], human disease association [HGMD (STENSON *et al.* 2014), ClinVar (LANDRUM *et al.* 2016), OMIM (AMBERGER *et al.* 2015)] and mouse model knockout phenotypes [MGI (BLAKE *et al.* 2017)], gene-based scores on accumulated mutational damage [GDI (ITAN *et al.* 2015)] and genic intolerance [RVIS (PETROVSKI *et al.* 2015)], gene expression [GTEx (CONSORTIUM 2013), HPA (UHLEN *et al.* 2010)], protein-protein interaction network [IntAct (HERMJAKOB *et al.* 2004), BioGRID (CHATR-ARYAMONTRI *et al.* 2015), and ConcesusPathDB (KAMBUROV *et al.* 2009) compiled in dbNSFP, and HINT (DAS AND YU 2012)], and

7) pathway enrichment analysis using a fisher exact test. Users may also choose to include supplementary files, such as a pedigree (PED)

file for co-segregation analysis, a list of genes for personalized annotation, or a second VCF file with a control cohort for genetic association tests [BURDEN (PLINK/SEQ 2014), Calpha (NEALE *et al.* 2011), vt (PRICE *et al.* 2010), and SKAT (WU *et al.* 2011) implemented in PLINK/SEQ].

All these options come together to provide a holistic filtering, annotation, and prioritization pipeline (**Figure 2.1**).

The customized pipeline task is scheduled for processing on a protected server, alleviating the user's burden to update software, parse data libraries, store large derivative files, and dedicate processing time. The average turnaround time is about 11 minutes for a 1MB VCF input file containing ~13,800 variants under default settings, of which querying up to 18 *in silico* predictions takes a static 8-minute searching through a 76GB dbNSFP database on all chromosomes. Although the processing time will vary depending on the choice of options and the number of concurrent users, GeMSTONE in general can handle a single ~500M VCF input per run within one day. Once the job is finished, the user can log into the GeMSTONE portal to interact with the completed workflow by selectively downloading step-by-step snapshots of their workflow, interactively visualizing their variant statistics, and downloading their recipe (JSON) file, which can be uploaded or shared to replicate or modify the same workflow.

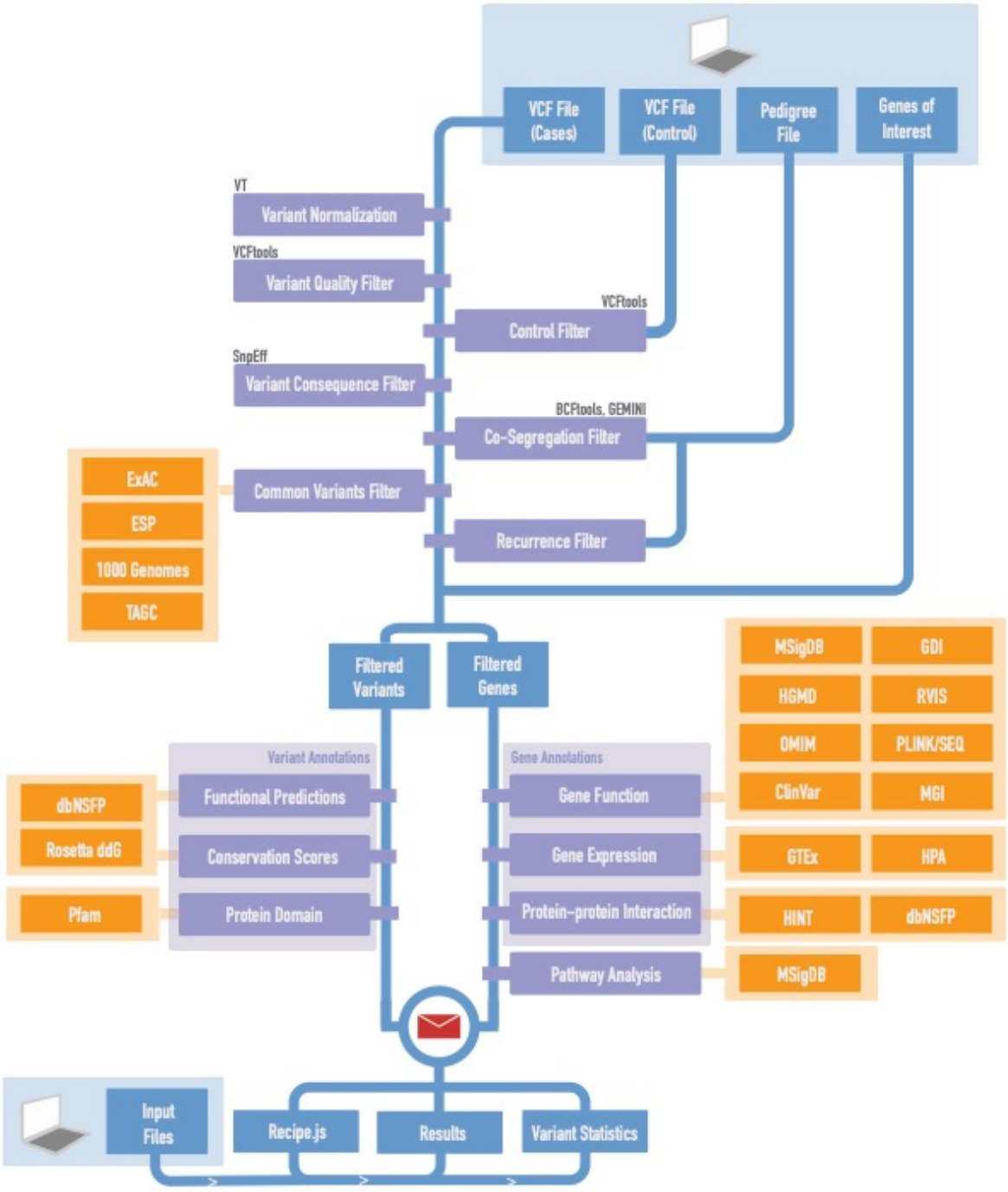


Figure 2.1. GeMSTONE pipeline overview. The schematic represents the GeMSTONE's central analysis pipeline. The fundamental backbone filter cascade can be seen in blue, prioritizing rare and putatively damaging variants, as well as genes of high sequencing quality. Different libraries are grouped in orange, participating in annotation or filtering steps throughout the workflow as indicated.

An essential design to reinforce GeMSTONE's reproducibility function and to ensure the sustainability of our web tool is our rigorous versioning system. We keep in our system static versions of all the external resources where all the tools and datasets that we use for GeMSTONE are loaded onto our server so that it does not go to any external program or server when running. Thus, we are able to ensure backwards-compatibility as we add updated versions of software or new tools. GeMSTONE records the versions of each tool and database used in a job in the recipe file and if users submit a recipe whose workflow uses older software or datasets, they will be prompted on the fly asking whether they want to use the legacy version or the latest version of the resources. GeMSTONE also record the versions in a human-readable summary file for easy access and reference.

One important function for germline mutation prioritization in human disease is GeMSTONE's co-segregation analysis, which provides six common inheritance models (autosomal dominant, autosomal recessive, recessive compound heterozygous (via GEMINI (PAILA *et al.* 2013)), X-linked dominant, X-linked recessive, Y-linked dominant) based on the user-defined pedigree structure in PED file. GeMSTONE screens sample genotypes (using BCFtools (LI *et al.* 2009)) in each family and seeks for variants that are co-segregating with disease status under selected mode of inheritance. Additionally, a recurrence filter constrains the degree to which

co-segregation events are allowed across multiple families and the prevalence of the variants in sporadic samples. We found this option to be seldom implemented by previous web tools yet often recommended by ACMG and AMP (RICHARDS *et al.* 2015). The benefits of this analysis are many-fold: 1) increasing segregation data in families or 2) high mutation frequency affecting multiple sporadic cases suggests stronger evidence for pathogenicity; 3) whereas a upper limit of such recurrence can help eliminate potential false positives in large sample size. This process of user-driven development by which GeMSTONE morphs to the community's needs is the key behind GeMSTONE's ability to grow as a knowledge bank with a robust and updated set of functionalities. Small but necessary prioritizing steps like these, now explicitly documented in the GeMSTONE summary and recipe files, can become an active component of study replication.

Another supporting evidence for disease association comes from *in silico* predictions of variant functional effects. Predictions from different algorithms are considered as a single piece of evidence in sequence interpretation in part due to the underlying similarities in the basis in these software suites (MACARTHUR *et al.* 2014; RICHARDS *et al.* 2015). GeMSTONE's variant functional prediction step allows the user to choose up to 19 different *in silico* predictors (**Supplementary Table S2.2**) with customizable thresholds. More dedicatedly, a 'global deleteriousness filter'

allows users to set a threshold on the number of selected predictors needed for a variant to pass the filter. This set of filters is useful in that it allows users to adjust the stringency of each algorithm while balancing and investigating any inconsistency among different predictions. The availability of these filters and annotations also provide an environment in which users can choose predictive metrics solely based on their relative merit rather than the programming investment that it would take to install, query, and customize them for a study.

Most options within the GeMSTONE workflow can serve a double purpose, acting as either filters or annotations. For the 'global deleteriousness filter' mentioned above, the count of deleterious predictions and their individual scores will be annotated next to each variant, providing information that can be used for variant prioritization without being part of any filter. We also provide the option to combine information across libraries, for example, we allow for known disease gene annotation on candidates to be supplemented with their interaction partners as reported in other databases, asking whether those interactors were previously implicated in the disease of interest. This distribution and coverage of tools (**Figure 2.1**) have never been collected and connected in a centralized workflow before.

By maintaining an updated set of bioinformatics tools for variant analysis, GeMSTONE decreases the barrier to entry for less computationally oriented research groups and establishes a central bioinformatics hub for researchers who study sequence variants implicated in severe familial diseases as well as rare, large-effect risk variants in complex disease. The options offered by the web interface also serve as a way for users to explore and learn about new tools and data sources while providing developers with an overview of the current variant analysis landscape to fill any gaps in the current tool-space. New tools can be easily added to GeMSTONE and presented to the community through the web interface, removing platform-specific barriers.

RESULTS

As an example of a GeMSTONE use case, we replicated a published analysis of rare pathogenic variants in new predisposition genes for familial colorectal cancer (CRC) (ESTEBAN-JURADO *et al.* 2015). A side-by-side demonstration of the study's workflow and GeMSTONE's re-implementation using the same dataset and prioritization criteria is shown in **Figure 2.2**. The original analyses were conducted in two sequences of prioritization, progressively looking for predisposing mutations with stronger evidence for causality to CRC as they underwent increasingly stringent

criteria (lower allele frequency in general populations; rarer presence among the affected samples; more deleterious molecular impact by *in silico* predictions; more interesting biological functions of the genes and their protein product, e.g. domains and interactions) (ESTEBAN-JURADO *et al.* 2015). While formerly requiring in-house scripting for co-segregation analysis, *in silico* analysis, and a series of gene function annotations querying and parsing several databases, the entirety of each sequence of prioritization pipeline can be performed with a single run through our interactive, lightweight web form using GeMSTONE.

Perhaps the most convenient feature within GeMSTONE is its recipe file generator. The recipe file from any given run can be shared and readily uploaded to our site to modify any part of the filtering and annotation pipeline for more stringent prioritization in a follow-up run. Once uploaded, the recipe file (JSON) will populate the web form dynamically, giving the user the ability to modify the run using the same interface that created it.

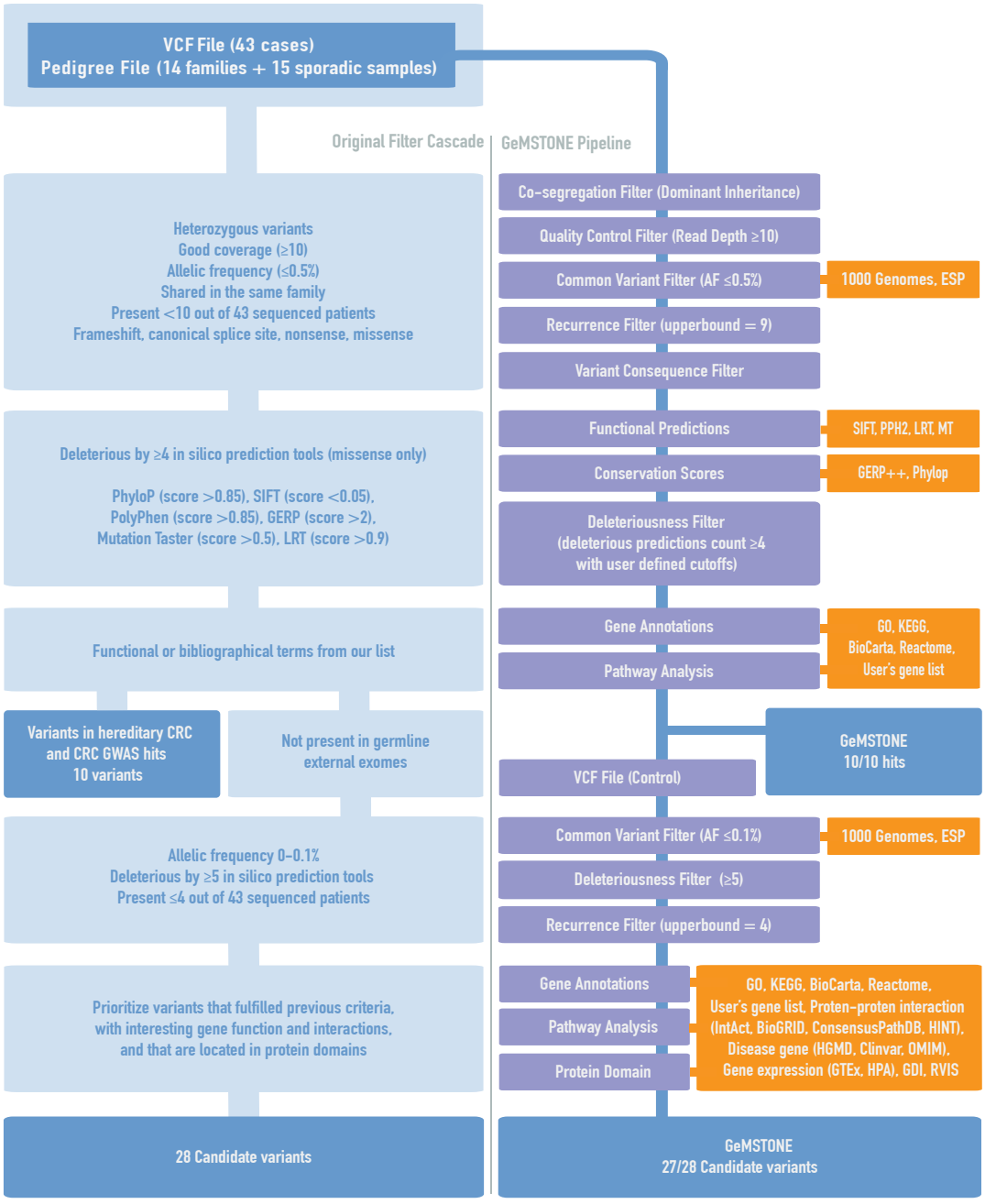


Figure 2.2. Recapitulation of a published CRC study**.** As a proof-of-concept case study, GeMSTONE 1) recapitulated every step in the original Colorectal-Cancer prioritization workflow1, 2) rescuing 27 out of 28 candidate variants from the ~30,000 variants in the raw whole exome sequencing dataset, and 3) hitting all hereditary CRC and CRC GWAS variants.

In our CRC case, we lowered the upper-bound of allele frequency filter from 0.5% to 0.1% [in 1000 Genomes (GENOMES PROJECT *et al.* 2012) and ESP6500 (FU *et al.* 2013)] and recurrence filter from 9 to 4, requiring variants to be present in ≤4 individuals in our data set. Next, we increased the lower-bound of deleteriousness filter from 4 to 5 without changing the user-defined deleterious thresholds of any single predictor [PhyloP (POLLARD *et al.* 2010) score >0.85, SIFT (KUMAR *et al.* 2009) score <0.05, PolyPhen-2 (ADZHUBEI *et al.* 2010) score >0.85, GERP++ (DAVYDOV *et al.* 2010) score >2, Mutation Taster (SCHWARZ *et al.* 2010) score >0.5, LRT (CHUN AND FAY 2009) score >0.9]. Finally, we added variant and gene annotations with interesting gene function, interactions, and locations in protein domains. This workflow leverages a variety of public databases, including Gene Ontology (ASHBURNER *et al.* 2000), KEGG (KANEHISA *et al.* 2014), Reactome (FABREGAT *et al.* 2016), HINT (DAS AND YU 2012), Pfam (FINN *et al.* 2014), and HGMD (STENSON *et al.* 2014), as well as a complementary list of cancer terms collected by the authors. This modified workflow was automatically recorded in a JSON recipe file and packaged with corresponding results and intermediate output files. Through the above two automated runs, GeMSTONE recapitulated every step of the original prioritization workflow. 27 out of 28 candidate variants were rescued (the missing variant was filtered out due to slightly higher allele frequency in a

sub-population database from 1000 Genomes), as well as all hereditary CRC and CRC GWAS variants (ESTEBAN-JURADO *et al.* 2015) (**Figure 2.2**).

DISCUSSION

GeMSTONE provides a code-free portal for variant filtering, annotation, and prioritization, which not only helps standardize genetic variation analyses (**Supplementary Table S1**) but also offers the means to replicate and share computational protocols easily. From a user's perspective, GeMSTONE is a reliable one-stop shop for variant analysis where they can find a collection of tools spanning a broad range of applications through an intuitive, unified user interface subsuming all general-purpose workflows from comparable toolkits (**Figure 2.3**).

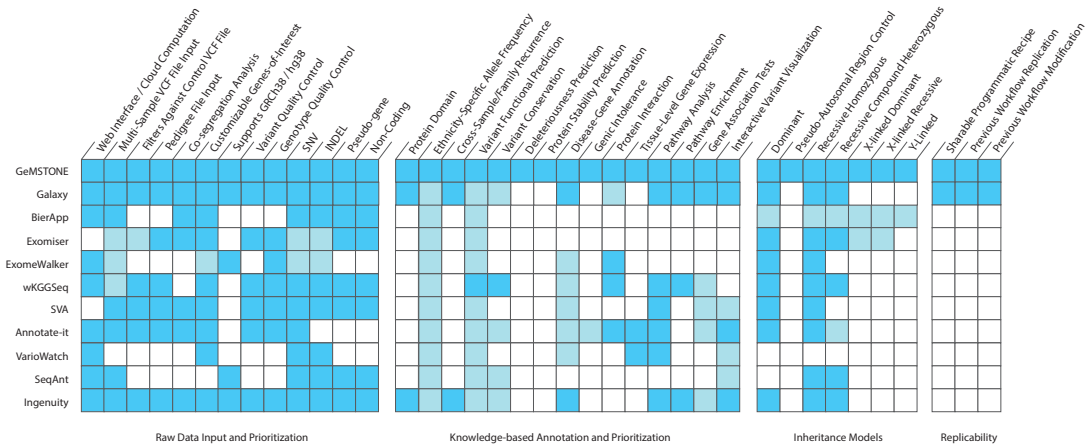


Figure 2.3. Heatmap comparison of GeMSTONE and other variant prioritization tools. This heatmap compares with other tools that have similar objectives on the aspects of (A) raw data inputs and prioritization, (B) knowledge-based annotation from external data resources and libraries, (C) inheritance models for co-segregation analysis, and (D) strategy of reproducibility. Each row represents a different tool, while each column represents a specific feature. Dark blue indicates that a tool has similar capacity for a specific function while light blue indicates that a tool has a similar feature but with less powerful functionality than GeMSTONE (see Discussion).

Although currently most of other variant prioritization tools accept VCF and pedigree files as inputs and can perform routine filtering on quality control and variant consequence (**Figure 2.3A**), GeMSTONE stands out as a more powerful tool by including annotations at the variant, gene, pathway, and network level (**Figure 2.3B**) and co-segregation analysis using different inheritance models for potential germline mutation prioritization (**Figure 2.3C**). We consider certain features in GeMSTONE to be "more powerful" in the aspect of comprehensiveness or/and flexibility: GeMSTONE often provides more comprehensive options for filtering and annotation linking to external resources than others and most of the GeMSTONE options flexibly

allow for annotation or filtering, or both. For a detailed breakdown of our comparison, refer to **Supplementary Table S3**.

A keystone of GeMSTONE is the recipe file (**Figure 2.3D**), which records all workflow parameters in a single file that can be shared and uploaded onto the site to reproduce a previous run. The recipe file can be used to 1) replicate results by rerunning the same workflow on the same dataset, 2) process new data with a known workflow or 3) modify parameters in a known workflow to evaluate study design. This approach has the potential to bring more transparency and openness to the bioinformatics community by enhancing the reproducibility of large-scale genomic studies.

CONCLUSION

GeMSTONE allows for accessible, collaborative, replicable and holistic analysis of genetic variants. First, it seamlessly knits together filters and annotations through different tools with either stringent, study-specific parameters or general best-practice settings. Second, it eliminates the time and space burdens associated with modern variant analysis tools, saving users dozens of gigabytes of potential disk space per run for the same workflow on a medium-sized dataset. Third, it significantly lowers the barrier to entry for traditional biologists by eliminating the installation and scripting sinkholes that may dissuade researchers from pursuing large-scale analysis or

trying new tools. Fourth, it provides a readable, shareable log—both programmatic and human—to allow other researchers to understand and replicate study results given the same starting data. Finally, GeMSTONE encourages the growth of the genomics research community by maintaining and updating a bank of best-practice bioinformatics methods and tools. We expect our GeMSTONE will greatly aid in automating the (re)analysis of genome-wide genetic variation data and enhance the reproducibility of large-scale genomic studies.

AUTHOR INFORMATION

Siwei Chen & Juan Felipe Beltrán

These authors contributed equally to this work.

AUTHOR CONTRIBUTIONS

X.W., S.L. and H.Y. conceived of the GeMSTONE concept. S.C. designed and implemented the GeMSTONE pipeline with supervision and input from H.Y. and X.W. J.F.B. designed and implemented the GeMSTONE web interface. C.E.J., S.F.E. and S.C.B. provided CRC study datasets and the original study design. J.F.B., S.C. and H.Y. wrote the manuscript. All authors reviewed and approved the final manuscript as submitted.

AVAILABILITY OF MATERIALS

Exome sequence data for 43 CRC patients were provided by Esteban-Jurado et al. (ESTEBAN-JURADO *et al.* 2015). The data will be available only upon request.

ETHICS APPROVAL

Ethics approval was not needed for this study.

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

This work was supported by National Institute of General Medical Sciences [R01 GM097358, R01 GM104424, R01 GM108716]; National Cancer Institute [R01 CA167824]; Eunice Kennedy Shriver National Institute of Child Health and Human Development [R01 HD082568]; Simons Foundation Autism Research Initiative [367561 to H.Y.]; and National Human Genome Research Institute [UM1 HG009393]. Funding for open access charge: National Institutes of Health and Simons Foundation Autism Research Initiative.

ACKNOWLEDGEMENTS

IDIBAPS: We are sincerely grateful to the patients and their families for their participation. We are really thankful to the Centre Nacional d'Anàlisi Genòmica and the Biobank of Hospital Clínic–IDIBAPS, Barcelona, for technical help. The work was carried out (in part) at the Esther Koplowitz Centre, Barcelona. CEJ and SFE are supported by a contract from CIBERehd. CIBERehd is funded by the Instituto de Salud Carlos III. This work was supported by grants from Fondo de Investigación Sanitaria/FEDER (14/00173), Fundación Científica de la Asociación Española contra el Cáncer (GCB13131592CAST), COST Action BM1206, Beca Grupo de Trabajo "Oncología" AEG (Asociación Española de Gastroenterología), and Agència de Gestió d'Ajuts Universitaris i de Recerca (Generalitat de Catalunya, 2014SGR255).

BRC: we would appreciate the Cornell University Biotechnology Resource Center for providing facilities' resources and services.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article.

Table S2.1. The table presents how GeMSTONE's centralized workflow aligns to guidelines and consensus recommendations for interpreting variant causality in human [as listed in Table 1 and BOX 2 of MacArthur et al. paper (19)(MACARTHUR *et al.* 2014)].

Table S2.2. The table provides information of all external tools and databases used in GeMSTONE, categorized by their applications. Methods or datasets complied by a software/data library are in italic and listed under the corresponding master resource. Both latest versions, that GeMSTONE uses by default, and older versions available for users' selection are listed below at the time GeMSTONE published. Please refer to GeMSTONE manual page for the latest updates(http://gemstone.yulab.org/manual.html). Sizes of programs and datasets are listed based on their memory allocation on the web server.

Table S2.3. The table details the reasons why other tools were considered less powerful in specific functionalities when compared to GeMSTONE.

REFERENCES

Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova *et al.*, 2010 A method and server for predicting damaging missense mutations. Nat Methods 7**:** 248-249.

Amberger, J. S., C. A. Bocchini, F. Schiettecatte, A. F. Scott and A. Hamosh, 2015 OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic Acids Res 43**:** D789-798.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler *et al.*, 2000 Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25**:** 25-29.

Bailey, J. N., C. Patterson, L. de Nijs, R. M. Duron, V. H. Nguyen *et al.*, 2017 EFHC1 variants in juvenile myoclonic epilepsy: reanalysis according to NHGRI and ACMG guidelines for assigning disease causality. Genet Med 19**:** 144-156.

Bellido, F., M. Pineda, G. Aiza, R. Valdes-Mas, M. Navarro *et al.*, 2016 POLE and POLD1 mutations in 529 kindred with familial colorectal cancer and/or polyposis: review of reported cases and recommendations for genetic testing and surveillance. Genet Med 18**:** 325-332.

Blake, J. A., J. T. Eppig, J. A. Kadin, J. E. Richardson, C. L. Smith *et al.*, 2017 Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. Nucleic Acids Res 45**:** D723-D729.

Boycott, K. M., M. R. Vanstone, D. E. Bulman and A. E. MacKenzie, 2013 Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet 14**:** 681-691.

Carmi, S., K. Y. Hui, E. Kochav, X. Liu, J. Xue *et al.*, 2014 Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. Nat Commun 5**:** 4835.

Chatr-Aryamontri, A., B. J. Breitkreutz, R. Oughtred, L. Boucher, S. Heinicke *et al.*, 2015 The BioGRID interaction database: 2015 update. Nucleic Acids Res 43**:** D470-478.

Chun, S., and J. C. Fay, 2009 Identification of deleterious mutations within three human genomes. Genome Res 19**:** 1553-1561.

Cingolani, P., A. Platts, L. Wang le, M. Coon, T. Nguyen *et al.*, 2012 A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6**:** 80-92.

Consortium, E. A., 2015 Analysis of protein-coding genetic variation in 60,706 humans.

Consortium, G. T., 2013 The Genotype-Tissue Expression (GTEx) project. Nat Genet 45**:** 580-585.

Cox, S. N., F. Pesce, J. S. El-Sayed Moustafa, F. Sallustio, G. Serino *et al.*, 2017 Multiple rare genetic variants co-segregating with familial IgA nephropathy all act within a single immune-related network. J Intern Med 281**:** 189-205.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27**:** 2156-2158.

Das, J., and H. Yu, 2012 HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol 6**:** 92.

Davydov, E. V., D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow *et al.*, 2010 Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6**:** e1001025.

Einarsdottir, E., I. Svensson, F. Darki, M. Peyrard-Janvid, J. M. Lindvall *et al.*, 2015 Mutation in CEP63 co-segregating with developmental dyslexia in a Swedish family. Hum Genet 134**:** 1239-1248.

Esteban-Jurado, C., M. Vila-Casadesus, P. Garre, J. J. Lozano, A. Pristoupilova *et al.*, 2015 Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer. Genet Med 17**:** 131-142.

Fabregat, A., K. Sidiropoulos, P. Garapati, M. Gillespie, K. Hausmann *et al.*, 2016 The Reactome pathway Knowledgebase. Nucleic Acids Res 44**:** D481-487.

Farlow, J. L., L. A. Robak, K. Hetrick, K. Bowling, E. Boerwinkle *et al.*, 2016 Whole-Exome Sequencing in Familial Parkinson Disease. JAMA Neurol 73**:** 68-75.

Finn, R. D., A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt *et al.*, 2014 Pfam: the protein families database. Nucleic Acids Res 42**:** D222-230.

Fu, W., T. D. O'Connor, G. Jun, H. M. Kang, G. Abecasis *et al.*, 2013 Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493**:** 216-220.

Genomes Project, C., G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo *et al.*, 2012 An integrated map of genetic variation from 1,092 human genomes. Nature 491**:** 56-65.

Goecks, J., A. Nekrutenko, J. Taylor and T. Galaxy, 2010 Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11**:** R86.

Halbritter, F., H. J. Vaidya and S. R. Tomlinson, 2011 GeneProf: analysis of high-throughput sequencing experiments. Nat Methods 9**:** 7-8.

Hermjakob, H., L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien *et al.*, 2004 IntAct: an open source molecular interaction database. Nucleic Acids Res 32**:** D452-455.

Itan, Y., L. Shang, B. Boisson, E. Patin, A. Bolze *et al.*, 2015 The human gene damage index as a gene-level approach to prioritizing exome variants. Proc Natl Acad Sci U S A 112**:** 13615-13620.

Kamburov, A., C. Wierling, H. Lehrach and R. Herwig, 2009 ConsensusPathDB--a database for integrating human functional interaction networks. Nucleic Acids Res 37**:** D623-628.

Kanehisa, M., S. Goto, Y. Sato, M. Kawashima, M. Furumichi *et al.*, 2014 Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 42**:** D199-205.

Kircher, M., D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper *et al.*, 2014 A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46**:** 310-315.

Kumar, P., S. Henikoff and P. C. Ng, 2009 Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4**:** 1073-1081.

Landrum, M. J., J. M. Lee, M. Benson, G. Brown, C. Chao *et al.*, 2016 ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 44**:** D862-868.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25**:** 2078-2079.

Liu, X., X. Jian and E. Boerwinkle, 2011 dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 32**:** 894-899.

Lushbough, C., M. K. Bergman, C. J. Lawrence, D. Jennewein and V. Brendel, 2010 BioExtract server--an integrated workflow-enabling system to access and analyze heterogeneous, distributed biomolecular data. IEEE/ACM Trans Comput Biol Bioinform 7**:** 12-24.

MacArthur, D. G., T. A. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure *et al.*, 2014 Guidelines for investigating causality of sequence variants in human disease. Nature 508**:** 469-476.

Mackay, D. S., T. M. Bennett, S. M. Culican and A. Shiels, 2014 Exome sequencing identifies novel and recurrent mutations in GJA8 and CRYGD associated with inherited cataract. Hum Genomics 8**:** 19.

Mackay, D. S., T. M. Bennett and A. Shiels, 2015 Exome Sequencing Identifies a Missense Variant in EFEMP1 Co-Segregating in a Family with Autosomal Dominant Primary Open-Angle Glaucoma. PLoS One 10**:** e0132529.

Medeiros, A. M., A. C. Alves and M. Bourbon, 2016 Mutational analysis of a cohort with clinical diagnosis of familial hypercholesterolemia: considerations for genetic diagnosis improvement. Genet Med 18**:** 316-324.

Metzker, M. L., 2010 Sequencing technologies - the next generation. Nat Rev Genet 11**:** 31-46.

Neale, B. M., M. A. Rivas, B. F. Voight, D. Altshuler, B. Devlin *et al.*, 2011 Testing for an unusual distribution of rare variants. PLoS Genet 7**:** e1001322.

Nekrutenko, A., and J. Taylor, 2012 Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat Rev Genet 13**:** 667-672.

Paila, U., B. A. Chapman, R. Kirchner and A. R. Quinlan, 2013 GEMINI: integrative exploration of genetic variation and genome annotations. PLoS Comput Biol 9**:** e1003153.

Petrovski, S., A. B. Gussow, Q. Wang, M. Halvorsen, Y. Han *et al.*, 2015 The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. PLoS Genet 11**:** e1005492.

PLINK/SEQ, 2014 PLINK/SEQ, pp.

Pollard, K. S., M. J. Hubisz, K. R. Rosenbloom and A. Siepel, 2010 Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20**:** 110-121.

Price, A. L., G. V. Kryukov, P. I. de Bakker, S. M. Purcell, J. Staples *et al.*, 2010 Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86**:** 832-838.

Radovica-Spalvina, I., G. Latkovskis, I. Silamikelis, D. Fridmanis, I. Elbere *et al.*, 2015 Next-generation-sequencing-based identification of familial hypercholesterolemia-related mutations in subjects with increased LDL-C levels in a latvian population. BMC Med Genet 16**:** 86.

Reich, M., T. Liefeld, J. Gould, J. Lerner, P. Tamayo *et al.*, 2006 GenePattern 2.0. Nat Genet 38**:** 500-501.

Richards, S., N. Aziz, S. Bale, D. Bick, S. Das *et al.*, 2015 Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17**:** 405-424.

Rohl, C. A., C. E. Strauss, K. M. Misura and D. Baker, 2004 Protein structure prediction using Rosetta. Methods Enzymol 383**:** 66-93.

Schwarz, J. M., C. Rodelsperger, M. Schuelke and D. Seelow, 2010 MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 7**:** 575-576.

Stenson, P. D., M. Mort, E. V. Ball, K. Shaw, A. Phillips *et al.*, 2014 The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 133**:** 1-9.

Tan, A., G. R. Abecasis and H. M. Kang, 2015 Unified representation of genetic variants. Bioinformatics 31**:** 2202-2204.

Uhlen, M., P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson *et al.*, 2010 Towards a knowledge-based Human Protein Atlas. Nat Biotechnol 28**:** 1248-1250.

Wright, C. F., T. W. Fitzgerald, W. D. Jones, S. Clayton, J. F. McRae *et al.*, 2015 Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet 385**:** 1305-1314.

Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke *et al.*, 2011 Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89**:** 82-93.

CHAPTER 3
CONTEXTUALIZING HUMAN DISEASE MUTATIONS BY BUILDING A STRUCTURAL PROTEIN INTERACTION NETWORK[1]

ABSTRACT

Protein interactions underlie nearly all known cellular function, making knowledge of their binding conformations paramount to understanding the physical workings of the cell. Studying binding conformations has allowed scientists to explore some of the mechanistic underpinnings of disease caused by disruption of protein interactions. However, since experimentally determined interaction structures are only available for a small fraction of the known interactome such inquiry has largely excluded functional genomic studies of the human interactome and broad observations of the inner workings of disease. Here we present Interactome INSIDER, an information center for genomic studies using the first full-interactome map of human interaction interfaces. We applied a new, unified framework to predict protein interaction interfaces for 184,605 protein interactions with previously unresolved interfaces in human and 7 model organisms, including the entire experimentally determined human binary interactome. We find that predicted interfaces share several known functional properties of interfaces, including an enrichment for disease

[1] Published as: *Meyer, M. J., *Beltrán, J. F., *Liang, S., Fragoza, R., Rumack, A., Liang, J., Wei, X., Yu, H. (2018). Interactome INSIDER: a structural interactome browser for genomic studies. *Nature Methods*, *15*, 107.

* These authors contributed equally to the work.

mutations and recurrent cancer mutations, suggesting their applicability to functional genomic studies. We also performed 2,164 *de novo* mutagenesis experiments and show that mutations of predicted interface residues disrupt interactions at a similar rate to known interface residues and at a much higher rate than mutations outside of predicted interfaces. To spur functional genomic studies in the human interactome, Interactome INSIDER (http://interactomeinsider.yulab.org) allows users to explore known population variants, disease mutations, and somatic cancer mutations, or upload their own set of mutations to find enrichment at the level of protein domains, residues, and 3D atomic clustering in known and predicted interaction interfaces.

INTRODUCTION

Protein-protein interactions facilitate much of known cellular function. Recent efforts to experimentally determine protein interactomes in human (ROLLAND *et al.* 2014) and model organisms(YU *et al.* 2008; ARABIDOPSIS INTERACTOME MAPPING 2011; VO *et al.* 2016), in addition to literature curation of small-scale interaction assays(DAS AND YU 2012), have dramatically increased the scale of known interactome networks. Studies of these interactomes have allowed researchers to elucidate how modes of evolution affect the functional fates of paralogs (VO *et al.* 2016) and to examine on a genomic scale network interconnectivities that determine cellular functions and disease states (SAHNI *et al.* 2015).

While simply knowing which proteins interact with each other provides valuable information to spur functional studies, far more specific hypotheses can be tested if the spatial contacts of interacting proteins are known(KIM *et al.* 2006). In the study of human disease, it has been

demonstrated that mutations tend to localize to interaction interfaces and mutations on the same protein may cause clinically distinct diseases by disrupting interactions with different partners (WANG *et al.* 2012; SAHNI *et al.* 2015). However, the binding topologies of interacting proteins (i.e. the relative positions of all atoms in an interaction interface) can only be absolutely determined through resource-intensive X-ray crystallography, NMR, and more recently cryo-EM (KUHLBRANDT 2014) experiments, severely limiting the number of interactions with resolved interaction interfaces.

In order to study protein function on a genomic scale, especially as it relates to human disease, a similarly large-scale set of protein interaction interfaces is needed. Thus far, computational methods, such as docking (HALPERIN *et al.* 2002) and homology modeling (SALI AND BLUNDELL 1993), have been employed to predict the atomic-level bound conformations of interactions whose experimental structures have not yet been determined. Though it is capable of producing high quality interaction models (LENSINK *et al.* 2016), docking remains highly specialized and docked models are not yet available on a large scale. Homology modeling has been used to produce models on a large scale (MOSCA *et al.* 2013), but is only amenable to interactions with structural templates, which comprise <5% of known interactions. Together, co-crystal structures and homology models comprise the currently available pre-calculated sources of structural interactomes, covering only ~6% of all known interactions (**Figure 3.1**).

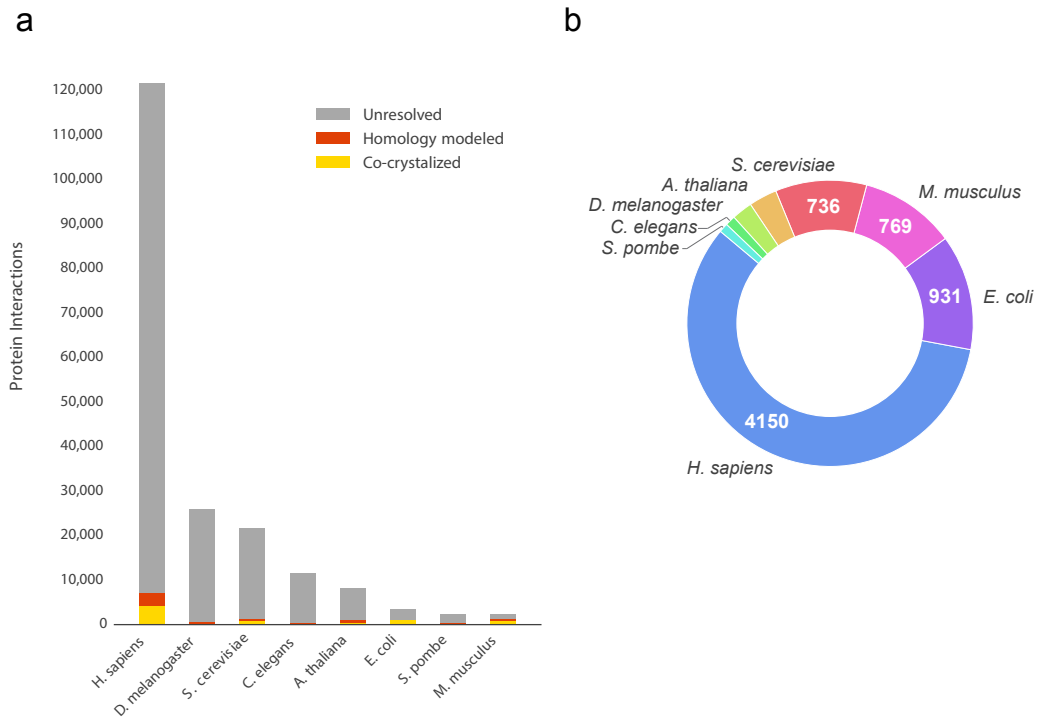


Figure 3.1. The current size of structural interactomes. (a) The sources of pre-computed structural interactomes and their coverage of known high quality binary interactomes. (b) Interactions from the largest 8 interactomes with experimentally solved structures, which can be used to train a classifier.

While we aim to study disease mutations at atomic-resolution when possible, for the ~94% of interactions without structural information, a lower-resolution picture of interfaces can provide crucial information for functional studies, and help to complete structural interactome networks to the best of our current capabilities (VAKSER 2013). For instance, residue-level interaction interfaces, where we know which residues are at the interface, but not their precise structural arrangement, can be a great boon to genomic-scale functional analyses (XIE *et al.* 2014; BRUNK *et al.* 2016), and elucidate common modes of human disease (WANG *et al.* 2012; DAS *et al.* 2014). Therefore, a multi-scale interactome network containing the highest possible resolution of each protein interaction interface can be an

indispensable tool for targeted studies to elucidate pathways and dissect disease mechanisms (WEI *et al.* 2014; VO *et al.* 2016).

Here, we present Interactome INSIDER (**IN**tegrated **S**tructural **I**nteractome and genomic **D**ata brows**ER**), a tool for functional exploration of human disease mutations using the first structurally resolved, multi-scale, proteome-wide human interactome. Interactome INSIDER allows users to find enrichment of disease mutations from popular databases and from user uploads in protein interaction domains, residues, and through atomic 3D clustering in protein interfaces. In order to study disease on a genomic scale, we built an interactome-wide set of protein interaction interfaces by calculating interfaces in experimental co-crystal structures and homology models when available. For the remaining ~94% of interactions, we applied a new, unified framework, ECLAIR (**E**nsemble **C**lassifier **L**earning **A**lgorithm to predict **I**nterface **R**esidues) to predict the interfaces by applying recent advances in partner-specific interface prediction, such as co-evolution- and docking-based feature construction (HOPF *et al.* 2014; HWANG *et al.* 2014). We used ECLAIR to predict protein interaction interfaces in the full human interactome and for 7 highly studied model organisms (*D. melanogaster*, *S. cerevisiae*, *C. elegans*, *A. thaliana*, *E. coli*, *S. pombe*, and *M. musculus*).

Interactome INSIDER (http://interactomeinsider.yulab.org) is deployed as an interactive web server, containing tools for analyzing known and uploaded disease mutations, cancer mutations, and population variants in genome-wide interaction interfaces. Users can also browse predicted interface residues for 184,605 previously un-resolved interactions in human and 7 model organisms, a 15-fold increase over previously known interfaces. Furthermore, for 12,546 interactions with pre-existing sources of structural

evidence (co-crystal structures or homology models), we calculate interface residues and display interactive 3D models. Users can search interaction interfaces for enrichment of disease mutations at the level of protein domains, residues, and 3D atomic clustering in a unified interactome composed of all of these sources. We also include relevant functional annotations, such as deleteriousness predictions (KUMAR *et al.* 2009; ADZHUBEI *et al.* 2010) and biophysical property changes (GRANTHAM 1974; ARTIMO *et al.* 2012) for any proposed mutation or variant that can be viewed in the context of protein and interaction models for a unified functional genomic experience. We anticipate that the marriage of these data sources with our newly predicted full coverage human structural interactome will spur studies of interaction interfaces on a genomic scale.

RESULTS

In order to build Interactome INSIDER, a tool for genome-wide inference in protein interaction interfaces, we first must construct an interactome-wide set of protein interaction interfaces. Due to lack of structural models, we turned to the well-explored field of protein interaction interface prediction to fill in the gaps in interactomes where neither experimentally-determined co-crystal structures nor homology models are available. While there are well-established methods for predicting protein interactions themselves (i.e. whether or not two proteins interact)(ZHANG *et al.* 2012; GARZON *et al.* 2016), we have focused on interactions that have been experimentally determined, but whose interfaces are unknown (**Supplementary Note S3.1**). For this task, there is a rich literature of methods exploring the potential of many structural, evolutionary, and docking-based methods to predict protein interaction interfaces. However,

so far, none of these methods have been used to produce a whole-interactome dataset of protein interaction interfaces (**Supplementary Note S3.2**).

We created ECLAIR, a unified framework for predicting the interface of any interaction, by leveraging several complementary and proven classification features, including both sequence-based biophysical features, and structural features (**Supplementary Note S3.3, Supplementary Figures S3.1-3.2**). Furthermore, ECLAIR uses recently proposed features for predicting binding partner specific interfaces, including co-evolutionary (LOCKLESS AND RANGANATHAN 1999; MORCOS *et al.* 2011) and docking-based metrics (PIERCE *et al.* 2011; HWANG *et al.* 2014). The advantage ECLAIR offers over previous methods is its ability to be applied to any interaction, while using the most informative set of available interactions for that interaction. In order to accomplish this, ECLAIR is structured as an ensemble of 8 independent classifiers, each covering a common case of feature availability (**Supplementary Notes S4-5, Supplementary Figures S3-4**). Because each ECLAIR sub-classifier has been trained and tested using a unified set of known protein interaction interfaces, we were able to benchmark each and show that interfaces can be predicted by the single, top-performing sub-classifier that was trained using the full set of features available for each residue (**Supplementary Note S3.4.2, Supplementary Figure S3.5**). In total, we used ECLAIR to predict the interfaces of 184,605 interactions with previously unknown interfaces, including for 114,504 human interactions (**Supplementary Figure S6**). We supplemented known structural interfaces from co-crystalized proteins and homology models with our predictions to create multi-scale structural interactomes at both the atomic and residue level (**Figure 3.2a**). Finally, in addition to predicting

interaction interfaces in 7 model organisms, we created the first multi-scale proteome-wide structural interactome in human for all 121,575 experimentally-determined binary interactions reported in major databases (SALWINSKI *et al.* 2004; KESHAVA PRASAD *et al.* 2009; TURNER *et al.* 2010; MEWES *et al.* 2011; KERRIEN *et al.* 2012; LICATA *et al.* 2012; CHATR-ARYAMONTRI *et al.* 2015) (4,150 with co-crystal structures, 2,921 with homology models, and 114,504 with ECLAIR predicted interfaces; see Materials and Methods), which we used to explore human disease through our new web tool, Interactome INSIDER.

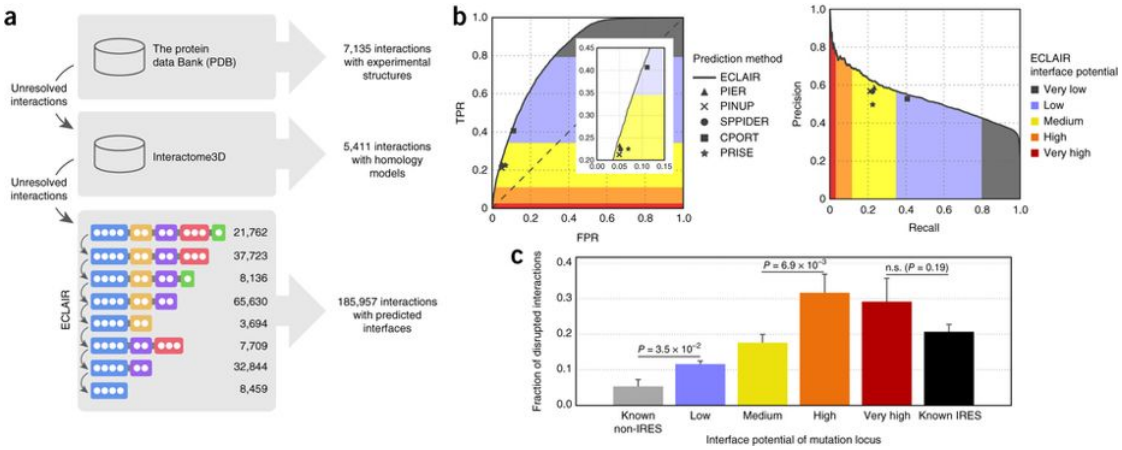


Figure 3.2. ECLAIR prediction results. (a) Workflow for classifying interfaces for all interactions in 8 species. Interactions without experimentally determined or homology modeled interfaces are classified by ECLAIR. (b) ROC and precision-recall curves comparing ECLAIR with other popular interface residue prediction methods. (c) Fraction of interactions disrupted by the introduction of random population variants in known and predicted interfaces. (* denotes significant ($p < 0.05$); *n.s.* denotes not significant by a Z-test)

Comprehensive evaluation of predicted interfaces

In order to use our structural interactomes for functional discovery, we first established that our predictions are of high quality through both machine-learning and biological evaluation. We evaluated the trade-offs between false positive rate and true positive rate, and between precision and

recall for each of the 8 independent sub-classifiers that compose ECLAIR (**Supplementary Figure S3.5**). As expected, we find that as more informative features are added to subsequent classifiers, the areas under the ROC and precision-recall curves increase, justifying the use of classifiers trained on more features for residues where this information is available.

We next compared ECLAIR to several other prediction methods through two independent validations, in order to establish that ECLAIR's performance is comparable to other methods. Due to its ensemble nature, we can then apply ECLAIR to many more interactions than would be possible using each of these methods individually. First, we used several readily available predictors (LIANG *et al.* 2006; KUFAREVA *et al.* 2007; POROLLO AND MELLER 2007; DE VRIES AND BONVIN 2011; JORDAN *et al.* 2012) to predict interfaces for interactions in our testing set. We find that for the set of interactions for which all classifiers can predict, ECLAIR performs as well or slightly better than these methods by measures of precision, recall, true positive rate and false positive rate (**Figure 3.2b**). Furthermore, for this set of predictors and ECLAIR, we also limited our analyses to only known surface residues, showing that all methods have a slightly lower AUROC (since it is more difficult to distinguish interface from non-interface among only surface residues), however ECLAIR still performs as well or better than all tested methods (**Supplementary Figure S3.7**). Finally, we applied ECLAIR to a standard external benchmark set of protein interaction interfaces (HWANG *et al.* 2010) which has been used to evaluate the performance of 10 other interface prediction methods (MAHESHWARI AND BRYLINSKI 2015). We find that ECLAIR outperforms all benchmarked methods in accuracy, and is comparable to the top performers in all other metrics (**Supplementary Table S3.1**). Furthermore, ECLAIR is applicable

to any interaction, while methods in this benchmark rely on single-protein structure inputs, making them less applicable to genome-wide studies.

We also performed >2,000 mutagenesis experiments to measure the rate at which population variants in our predicted interfaces disrupt interactions compared to variants within known co-crystal interfaces and non-interfaces (see Material and Methods). Since it is known that mutations at protein interfaces are more likely to break interactions (WEI *et al.* 2014; SAHNI *et al.* 2015), we hope to show that mutations in our predicted interfaces also break their corresponding interactions at a significantly higher rate than those known to be away from the interface and at similar rates compared to mutations in known interfaces (it is important to note that only ~21% of these mutations at known interfaces disrupt corresponding interactions since we tested population variants randomly selected from the Exome Sequencing Project (FU *et al.* 2013), many of which are believed to be benign). Using our high-throughput mutagenesis yeast two-hybrid assay (WEI *et al.* 2014), we find that the disruption rates for mutations at known interface residues are quite similar to disruption rates for mutations of predicted interface residues (**Figure 3.2c**). Furthermore, even mutations of residues with a 'Low' predicted interface potential are significantly more likely to disrupt interactions than mutations of residues known to be away from the interface. This suggests that there is viable functional signal in ECLAIR predictions, as even interfaces predictions in the 'Low' potential category show some signs of similar functional properties to known interfaces.

Interactome INSIDER, a genomics toolbox for interactome studies

We built Interactome INSIDER, a tool for searching for functionally enriched areas of protein interactomes, and for browsing our multi-scale structural interactome networks. Interactome INSIDER contains all 197,151 protein interactions whose interfaces have been either experimentally determined, homology modeled, or predicted using ECLAIR. Specifically for human, Interactome INSIDER contains interface information for all 121,575 experimentally-determined binary interactions reported in major databases (SALWINSKI *et al.* 2004; KESHAVA PRASAD *et al.* 2009; TURNER *et al.* 2010; MEWES *et al.* 2011; KERRIEN *et al.* 2012; LICATA *et al.* 2012; CHATR-ARYAMONTRI *et al.* 2015). Additionally Interactome INSIDER includes 56,159 disease mutations from HGMD (STENSON *et al.* 2014) and ClinVar (LANDRUM *et al.* 2016) and analyzed 1,300,352 somatic cancer mutations from COSMIC (FORBES *et al.* 2015) to compute their per-disease, pre-calculated enrichment in protein interaction interfaces at the residue level, domain level, and through atomic clustering. Furthermore, the site includes information on >600,000 population variants from the Exome Sequencing Project (FU *et al.* 2013), 1000 Genomes Project (GENOMES PROJECT *et al.* 2015) and more (UNIPROT-CONSORTIUM 2015) (see Materials and Methods). Users can then search Interactome INSIDER by protein to retrieve all interaction partners and their interfaces, or by disease to retrieve all interaction interfaces that are enriched for mutations of that disease. Additionally, users can upload their own set of mutations to find how they are distributed in the interactome and whether they are enriched in any protein interaction interfaces at the residue, domain, and atomic levels (**Figure 3.3**).

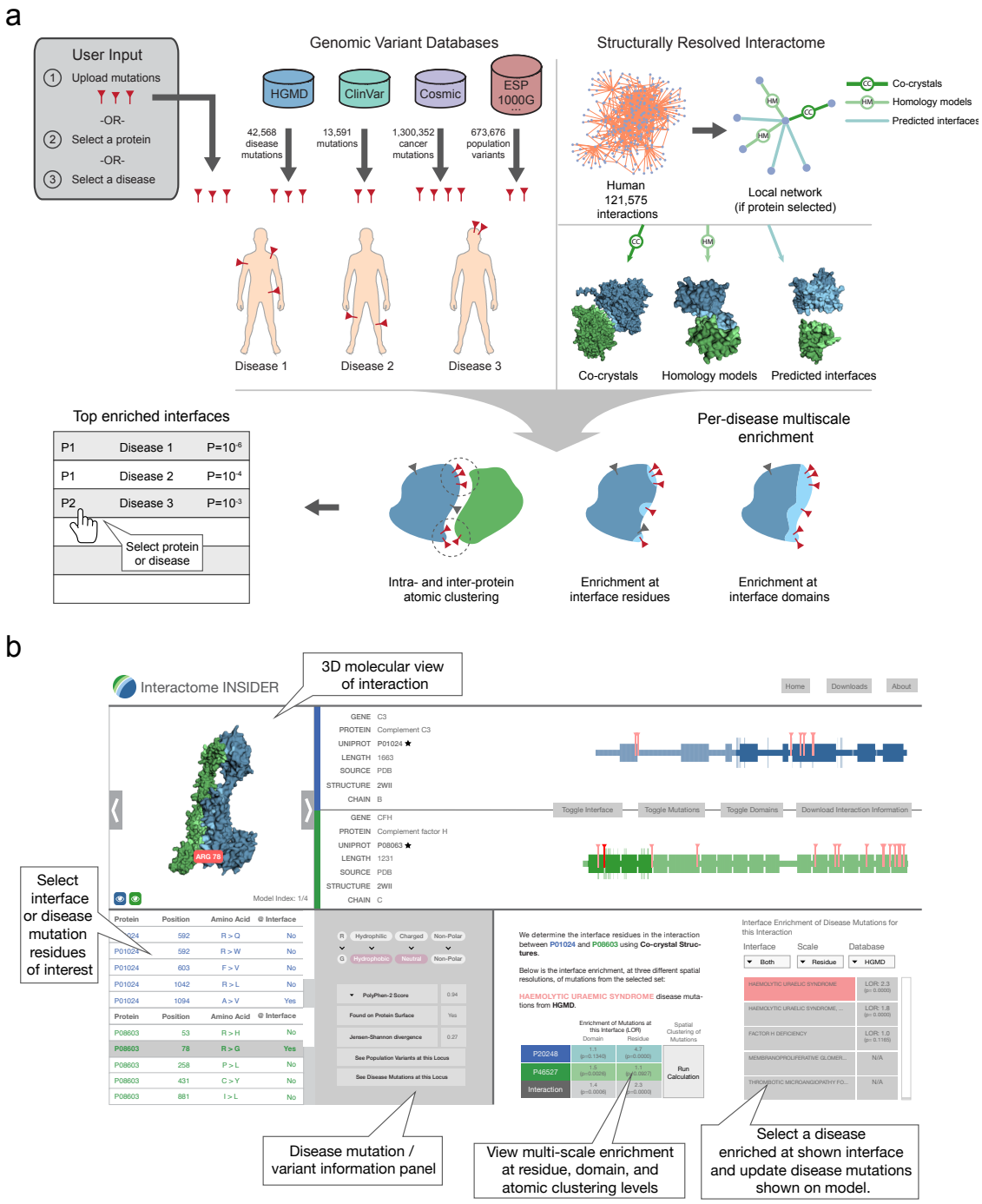


Figure 3.3. Flowchart showing the sources and computational workflow for calculating mutation and variant enrichment using the Interactome INSIDER web interface. Users may submit their own mutations or select sets of known disease and cancer mutations to assess their enrichment in interface domains and residues, or compute 3D atomic clusters of mutations in proteins and across interfaces.

Since our goal is to use Interactome INSIDER to explore protein function, especially disease, in interactomes, we next investigated the functional biological properties of our predicted interaction interfaces. These studies involve measuring functional properties of our de novo predicted interfaces (those without prior experimental evidence) and comparing these measurements to those of known interfaces from co-crystal structures. Importantly, these known properties of interaction interfaces are completely separate of the features used for training ECLAIR, and thus provide an unbiased and functionally relevant means to assess the utility of our predicted interfaces. Showing that our predicted interfaces have many of the same functional properties as known interfaces suggests their applicability to functional genomic studies.

Many studies have probed the link between interactome networks and disease (BARABASI *et al.* 2011; VIDAL *et al.* 2011), and it is well established that disease mutations are enriched at structural interfaces of interacting proteins (WANG *et al.* 2012; KAMBUROV *et al.* 2015; SAHNI *et al.* 2015), suggesting that disruption of binding with one or more partners may contribute to the disease phenotype. Though not all disease mutations will appear at the interfaces of interactions, and can act via other mechanisms, such as destabilizing proteins entirely (WEI *et al.* 2014), their enrichment at interfaces is a statistically significant global trend (WANG *et al.* 2012; SAHNI *et al.* 2015). However, >40% of known missense and nonsense human disease mutations cause alterations to proteins lacking any structurally resolved interaction interfaces. To test whether our predicted interfaces may be useful for the study of disease, and thus help address this knowledge gap, we looked at their positions relative to disease mutations. We find that disease mutations also preferentially occur in our predicted interfaces, at similar rates

to known interface residues occurring in PDB co-crystal structures (**Figure 3.4a**), indicating the viability of using predicted interfaces to study molecular disease mechanisms. Furthermore, each more confident bin of predicted interface residues is more likely to contain disease mutations than the previous, showing that ECLAIR prediction scores are correlated with true protein function.

Similarly, we looked at the locations of somatic cancer mutations from COSMIC in our interface-resolved human interactome. We specifically focused on recurrent cancer mutations as these are known to be more likely than infrequently observed mutations to be functional drivers (HODIS *et al.* 2012; RAPHAEL *et al.* 2014; MEYER *et al.* 2016). We find a marked enrichment of recurrent cancer mutations in our predicted interfaces compared to outside our predicted interfaces (**Figure 3.4b**). Furthermore, the same trend is observed inside and outside of known interfaces from co-crystal structures, suggesting that the functional links between cancer and the potential disruption of protein interactions can be observed within our entire human interface dataset. We also looked at the distribution of population variants, and show that their placement in and out of predicted interfaces matches that of known interfaces, with rarer mutations showing an enrichment in protein interfaces (**Figure 3.4c**). Furthermore, we show that population variants in our predicted interfaces are more likely to be damaging to protein function than variants outside of predicted interfaces, as predicted by PolyPhen-2 (ADZHUBEI *et al.* 2010) (**Figure 3.4d**) and EVmutation (HOPF *et al.* 2017) (**Figure 3.4e**), matching the established trend for experimentally determined interfaces (DAVID *et al.* 2012).

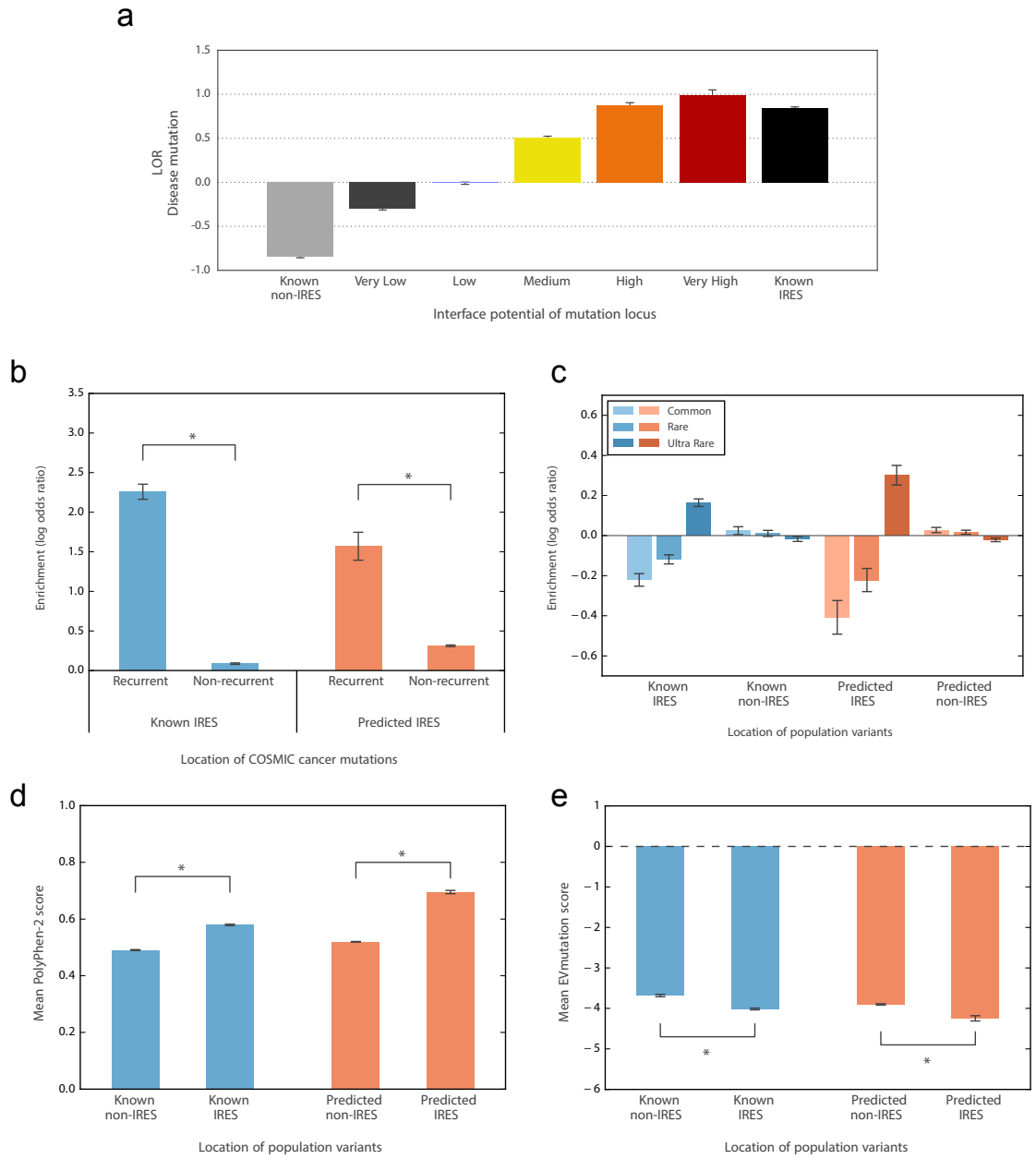


Figure 3.4. Functional properties of predicted interfaces. (a) Enrichment of disease mutations in predicted and known interfaces. (b) Enrichment of recurrent cancer mutations in predicted and known interfaces. (c) Enrichment of rare and common population variants in predicted and known interfaces. (d) Predicted deleteriousness of population variants in known and predicted interfaces (using PolyPhen-2). (e) Predicted effects of population variants in known and predicted interfaces (using EVmutation). (* denotes significant, $p < 0.05$ by a Z-test)

These enrichment analyses and the matched results between our predicted interfaces and known interfaces in co-crystal structures further

confirm the validity of ECLAIR predictions. More importantly, users can take advantage of these enrichment analyses through our Interactome INSIDER web server to better dissect large-scale whole-genome and whole-exome sequencing datasets to help identify novel disease-associated genes and mutations. For instance, if known disease mutations are significantly enriched in a specific interaction interface, this information could be used to complement and further boost the confidence of patient and disease-specific variants in the same interface that have been prioritized by other methods (e.g., co-segregation (MACARTHUR *et al.* 2014), mutation burden (LAWRENCE *et al.* 2013)). This can be particularly helpful to sieve through the large number of variants of unknown significance generated by large-scale sequencing studies. Furthermore, if known disease mutations are enriched in a specific interface of a protein whose involvement in the disease is already understood, this could still suggest its interaction partner's mechanistic involvement in the disease through this specific interface, even if the partner is not yet known to be associated with the disease.

To illustrate the usefulness of Interactome INSIDER, we searched for sub-networks in the human interactome that are enriched for disease mutations associated with a single disease by calculating the enrichment of disease mutations in interaction interfaces interactome-wide, a functionality also available to users via the Interactome INSIDER website. This allowed us to identify the TGF-β/BMP signaling pathway, which is known to be involved in juvenile polyposis syndrome (JPS) (WANG *et al.* 2014), and contains multiple proteins harboring JPS mutations (**Figure 3.5a**).

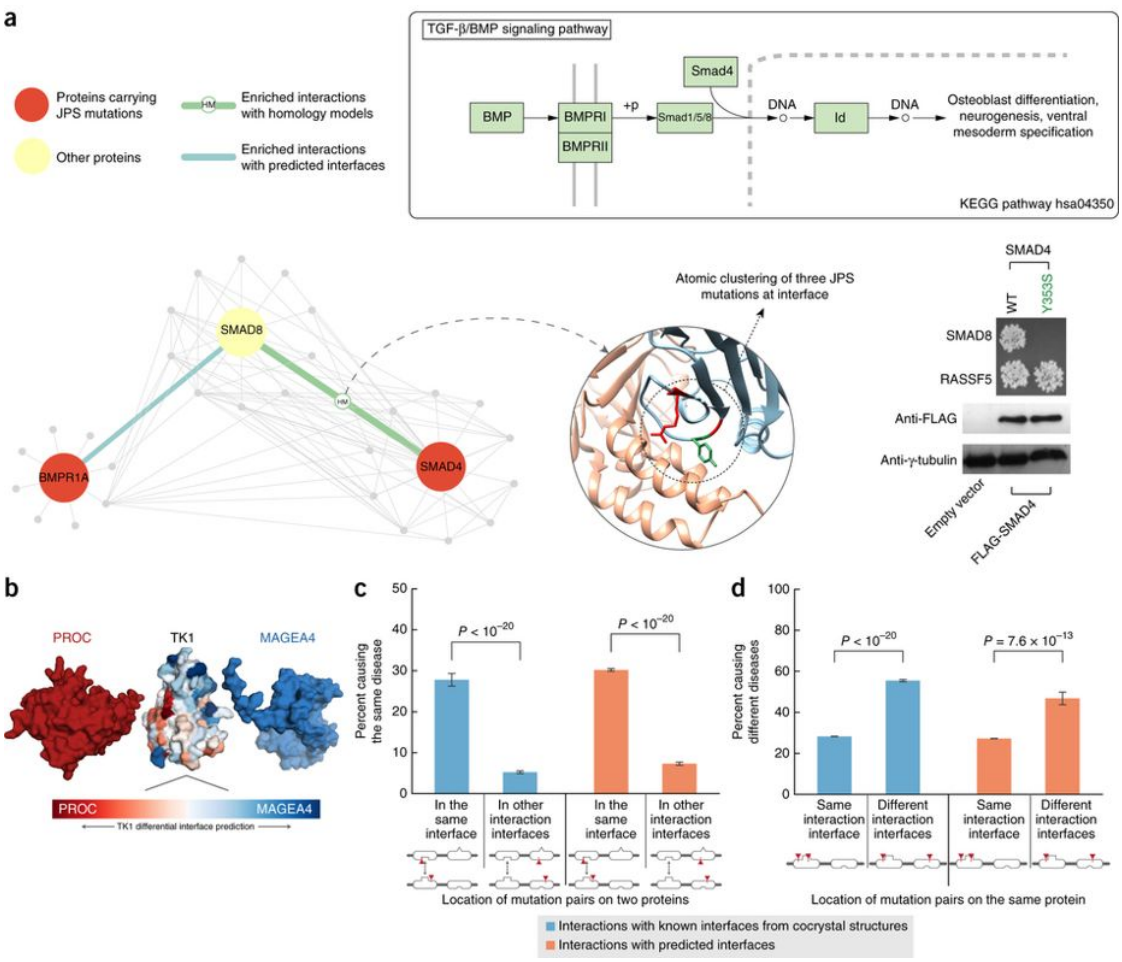


Figure 3.5. (a) The top schematic depicts the TGF-β/BMP signaling pathway. The bottom schematic illustrates that atomic clustering reveals a mutation hotspot for juvenile polyposis syndrome at the interface of SMAD8 and SMAD4. At right, yeast-two-hybrid experiments test the interactions of one of the SMAD4 mutations (Y353S) with SMAD8 and RASSF5. The mutation is not predicted by ECLAIR to be at the SMAD4–RASS5 interface. (b) Superimposed docking results of two different interaction partners with TK1. The differentially predicted interfaces of TK1 with each of its partners correspond with the orientation of the docked poses. (c) The plot shows the fraction of disease mutation pairs in known (blue) or predicted (orange) interfaces that cause the same disease when mutations are on either side of an interaction interface (in different proteins) compared to in other interaction interfaces (that don't facilitate the given interaction). (d) The plot shows the fraction of disease mutation pairs in known (blue) or predicted (orange) interfaces that cause different diseases when mutations are in the same interaction interface compared to in different interaction interfaces (interaction with other proteins is not shown). (Significance determined by two-sided *Z*-test.)

We focused on a specific group of mutations in the SMAD4-SMAD8 interface, which can be found using 3D atomic clustering. Using our

mutagenesis Y2H assay, we were able to test a JPS mutation (SMAD4 Y353S)(ROTH *et al.* 1999), which is at the interface of SMAD4-SMAD8, and show that it breaks this interaction, implicating SMAD8 in JPS. Although SMAD8 (also known as SMAD9) has not been reported to harbor JPS mutations in HGMD (STENSON *et al.* 2014), its involvement in the disease has been suggested (NGEOW *et al.* 2015), showing the ability of Interactome INSIDER to implicate new proteins in disease. Furthermore, Y353S is not predicted by ECLAIR to be at the interface of SMAD4 and another of its binding partners, RASSF4, and indeed, through our Y2H experiment, does not break this interaction, demonstrating the functional insight Interactome INSIDER can provide about differential interfaces and how they might be relevant to understanding the molecular mechanisms of disease.

Disease etiology revealed by partner-specific interfaces

In addition to providing full coverage of interfaces in the human interactome, one major benefit that Interactome INSIDER provides is the ability to interrogate different interfaces for the same protein dependent upon its binding partner. For the study of protein function and disease, this is especially important as a protein may maintain different functional pathways through different interfaces, and disruption of one interface may leave another intact (WANG *et al.* 2012; VO *et al.* 2016). To demonstrate the potential of Interactome INSIDER to tease apart interface-specific disease mutation etiologies, where the same mutation can cause differential effects with two different binding partners, we first investigated an example of differential interface prediction using ECLAIR. Here we highlight an interaction whose predicted interfaces are strongly influenced by a single partner-specific feature, molecular docking. In **Figure 3.5b**, the protein TK1

is shown colored by its docked pose with each of two partners, PROC and MAGEA4. We note that the predicted interface residues on TK1 are drastically different for each partner, and that the areas with elevated interface potential correspond to the position of the two docking results. Even though these interaction interfaces were predicted using features additional to docking, this demonstrates how even a single partner-specific feature can lead to differential interface predictions.

The ability of Interactome INSIDER to reveal interaction partner-specific effects can also be demonstrated as a global trend in our ECLAIR-predicted interfaces. As discussed, this is important because disruptions of different interfaces of the same protein may cause differential disease states; for instance, disruption of one interface may cause a disease while disruption of another may not cause any detriment to protein function. To demonstrate this on a large-scale, we looked at pairs of disease mutations in the human interactome that appear at interaction interfaces. It has been shown that pairs of disease mutations in interacting proteins cause the same disease when located in the interaction interface more often than mutations located in interaction interfaces with separate partners (WANG *et al.* 2012). We performed the same analysis using differential interaction interfaces predicted by ECLAIR and find the same trend—mutation pairs in the interface of two interacting proteins are much more likely to cause the same disease than mutation pairs in other interfaces of the same proteins that do not mediate the given interaction (**Figure 3.5c**). We also find that mutation pairs on the same protein, but in separate interfaces with different binding partners tend to cause different diseases (**Figure 3.5d**). Moreover, this trend is observed in both known and predicted interfaces. This shows that ECLAIR is able to use partner-specific features such as docking and co-

evolution to predict different interfaces depending on the binding partner of a protein that match established trends of pleiotropy and locus heterogeneity in known interfaces (WANG *et al.* 2012). Importantly, this indicates that Interactome INSIDER can be used to form functional hypotheses about the specificity of mutations to specific interactions and molecular pathways.

Using Interactome INSIDER to find sub-networks in the human interactome enriched for disease mutations associated with a single disease, we also uncover a set of interacting proteins known to harbor mutations causal of hypertrophic cardiomyopathy (HCM)(MARON 2002), a disease marked by enlargement of the myocardium heart muscle that can become fatal, and automatically recapitulate the core constituents of a known KEGG pathway related to the same disease (**Figure 3.6**). These proteins were identified by enrichment of disease mutations in their shared interaction interfaces and, in the case of TNNI3-TNNC1, using cross-interface atomic clustering of disease mutation positions in 3D. Access to enrichment and 3D atomic clustering tools for this disease and users' uploaded mutations is available via the Interactome INSIDER web interface.

Interestingly, in addition to identifying known members of the HCM pathway, Interactome INSIDER also identified several additional proteins, including CSRP3, MYOM1, ANKRD and TCAP, which are not part of the known KEGG pathway, but carry HCM mutations enriched at their respective interaction interfaces with members of the pathway. We also identify a protein, TNNT1, which, although it contains no HCM mutations of its own, can be implicated in HCM by interacting with two proteins TPM1 and TNNC1, which are enriched for HCM mutations at their interfaces with TNNT1. Finally, we note that Interactome INSIDER reveals cases of partner-specific interfaces in this pathway. For instance, the known

HCM pathway protein TTN's interface with ACTA1 is enriched for HCM mutations, and ACTA1 mutations are increasingly linked to HCM(D'AMICO *et al.* 2006; DONKERVOORT *et al.* 2015). On the other hand, a separate interface of ACTA1 with its binding partner dystrophin is enriched with mutations causing a distinct disorder, actin myopathy(SPARROW *et al.* 2003). This shows how ACTA1 can play roles in two different diseases through separate interaction interfaces with TTN and dystrophin, and demonstrates Interactome INSIDER's unique ability to discover such cases of differentiable function mirroring differential interfaces.

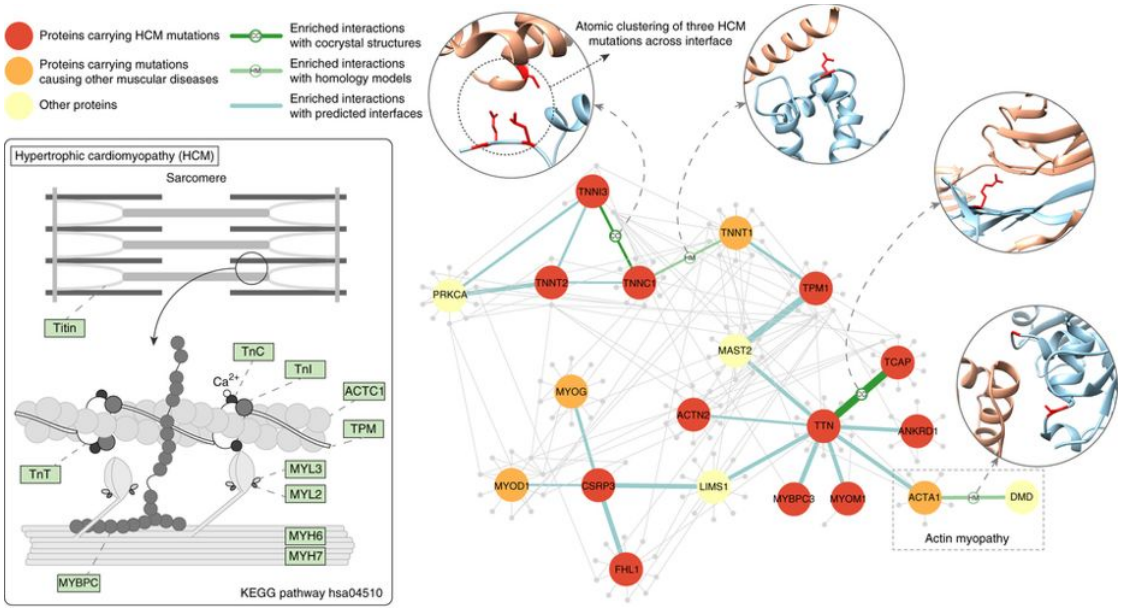


Figure 3.6. The schematic on the left shows the interaction of proteins in the HGM KEGG pathway (hsa04510). On the right is shown a network of KEGG pathway proteins and their structurally resolved interactions from Interactome INSIDER. Proteins that harbor HCM mutations are colored in red. Interfaces are noted for their enrichment of HCM mutations.

DISCUSSION

Interactome INSIDER is an integrative information center for genomics studies in the structural human interactome. By leveraging several

sources of protein interaction interfaces, including experimentally determined co-crystal structures, homology models, and predicted interfaces, Interactome INSIDER allows scientists to probe for functional insights in whole interactomes, and to predict disease etiologies based on network topology and specific structural interfaces at several scales of resolution. Our new interface prediction pipeline, ECLAIR, incorporates many previously validated strategies and features for predicting protein interaction interfaces in whole genomes, allowing Interactome INSIDER to be the first resource to show that predicted interfaces can be used for functional analyses in whole interactomes, especially for the study of human disease.

We anticipate Interactome INSIDER will help to bridge the divide between genomic-scale datasets and structural proteomic analyses, both now and in the future. Now that large-scale sequencing data from many contexts are readily available, for instance from whole-genome/whole-exome population variant studies (FU *et al.* 2013; LEK *et al.* 2016) and cancer studies (FORBES *et al.* 2011; KANDOTH *et al.* 2013), researchers have become increasingly interested in ways to assess the potential functional consequences of variants on a genomic scale (CINGOLANI *et al.* 2012; LAWRENCE *et al.* 2013; TASAN *et al.* 2015; HOFREE *et al.* 2016). For instance, recently we and others have developed methods to predict functional cancer driver mutations by finding hotspots of mutations in the structural proteome (KAMBUROV *et al.* 2015; YANG *et al.* 2015; MEYER *et al.* 2016). With the comprehensive map of protein interfaces presented, we can now go a step further to predict specific etiologies of cancer and disease based on induced biophysical effects (LI *et al.* 2014; KUCUKKAL *et al.* 2015) that may break interactions. Because our interface map is partner-specific, it can also be applied to predict pleiotropic effects, wherein several mutations in a single

protein may affect different pathways depending upon which binding interfaces are mutated (WANG *et al.* 2012). This could be the basis for designing new therapeutics and for rational drug design to selectively target specific protein functional sites (LOUNNAS *et al.* 2013).

The scale of interactomes and functional genomic data in Interactome INSIDER uniquely enables it to be useful for genomic studies. While at least one previous resource, dSysMap (MOSCA *et al.* 2015), is able to display disease mutations in structural interfaces, it is limited to 9,875 human interactions with either co-crystal structures or homology models, severely limiting its applicability to genomic studies by offering the same small slice of the interactome that has been studied extensively. Interactome INSIDER on the other hand contains interfaces for an additional 111,700 human interactions alone, which have never been available before in any repository. Furthermore, unlike dSysMap, Interactome INSIDER contains somatic cancer mutations, population variants, and mutation functional annotations, as well as interfaces for 7 model organisms with potential use in model systems studies, which have proven useful in the study of human disease (AITMAN *et al.* 2011) and for studying molecular evolution (ELLEGREN 2008; VO *et al.* 2016). Thus, we intend Interactome INSIDER be a more broadly applicable resource, with the ability to inform many aspects of genomic studies, from identifying functional regions of proteins, to incorporating orthogonal information about known mutations and functional effects in these regions.

With future increases to the scale of biological databases from which we derive features, we expect that Interactome INSIDER will come to encompass even higher confidence predictions for many more interactions, thereby becoming increasingly applicable to functional studies. This may also

address some limitations of structural databases today. For instance, the PDB is depleted of disordered proteins(PENG *et al.* 2004), and it has been shown that disordered regions can form interfaces(DUNKER *et al.* 2008). Since ECLAIR has not been trained on disordered interfaces, it is unlikely to predict new disordered interfaces. However, the ensemble classifier structure of ECLAIR uniquely positions it to incorporate all newly-available evidence into interface predictions without sacrificing quality or scale, ensuring the highest quality map of interaction interfaces now and in the future. Furthermore, the addition of new variants, especially cancer mutations and population variants from large-scale sequencing studies, will only increase the value of performing systems-level explorations with Interactome INSIDER.

MATERIALS & METHODS

Interaction datasets

We compiled binary protein interactions available for *H. sapiens*, *D. melanogaster*, *S. cerevisiae*, *C. elegans*, *A. thaliana*, *E. coli*, *S. pombe*, and *M. musculus* from 7 primary interaction databases. These databases include IMEx(ORCHARD *et al.* 2012) partners DIP (SALWINSKI *et al.* 2004), IntAct (KERRIEN *et al.* 2012), and MINT (LICATA *et al.* 2012), IMEx observer BioGRID (CHATR-ARYAMONTRI *et al.* 2015), and additional sources iRefWeb (TURNER *et al.* 2010), HPRD (KESHAVA PRASAD *et al.* 2009), and MIPS (MEWES *et al.* 2011). Furthermore, iRefWeb combines interaction data from BIND (ALFARANO *et al.* 2005), CORUM (RUEPP *et al.* 2010), MPact (GULDENER *et al.* 2006), OPHID (BROWN AND JURISICA 2005), and MPPI (PAGEL *et al.* 2005). We filtered these interactions using the PSI-MI (HERMJAKOB *et al.* 2004) evidence codes of assays that can determine

experimental binary interactions (**Supplementary Table S3.2**), as these are interactions where proteins are known to share a direct binding interface that we can then predict (DAS AND YU 2012). In total, we curated 197,151 interactions in these 8 species including the full experimentally determined binary interactome in human (121,575 interactions) (**Supplementary Note S3.1**). Those interactions with known interface residues based on available co-crystal structures in the Protein Data Bank (PDB)(BERMAN 2000) were set aside for use in training and testing the classifier. Interactions without known interface residues comprise the set for which we make predictions.

Testing and training sets for interface residue prediction

For those interactions with known co-crystal structures in the PDB, we calculate interface residues for their specific binding partners. To identify UniProt protein sequences in the PDB, we use SIFTS (VELANKAR *et al.* 2013), which provides a mapping of PDB-indexed residues to UniProt-indexed residues (UNIPROT-CONSORTIUM 2015). For each interaction and representative co-crystal structure, interface residues are calculated by assessing the change in solvent accessible surface area of the proteins in complex and apart using NACCESS(LEE AND RICHARDS 1971). Any residue that is at the surface of a protein (≥15% exposed surface) and whose solvent accessible surface area (SASA) decreases by ≥1.0 $Å^2$ in complex is considered to be at the interface. We aggregate interface residues across all available structures in the PDB for a given interaction, wherein a residue is considered to be at the interface of the interaction if it has been calculated to be at the interface in one or more co-crystal structures of that interaction (all other residues are considered to be away from the interface). In building our final training and testing sets, we only consider interactions for which

aggregated co-crystal structures have combined to cover at least 50% of UniProt residues for both interacting proteins.

The training and testing sets each include a random selection of 400 interactions with known co-crystal structures, of which 200 are heterodimers and 200 are homodimers (**Supplementary Table S3.3**). To ensure an unbiased performance evaluation, we disallowed any homologous interactions (i.e. interactions whose structures could be used as templates for homology modeling) between the training and testing set. We also disallowed repeated proteins between the two sets to avoid simply reporting a remembered shared interface between a protein and multiple binding partners, thereby artificially elevating the performance of our classifier on the testing set.

Hyperparameter optimization with TPE

In order to train our ensemble of classifiers that comprise ECLAIR, we used the tree-structured Parzen estimator approach (TPE) (BERGSTRA *et al.* 2011), a Bayesian method for optimizing hyperparameters for machine learning algorithms. TPE models the probability distribution $p(x|y)$ of hyperparameters given evaluated loss from a defined objective function, $L(x)$. We selected the following loss function to minimize based on classical hyperparameter inputs and residue window sizes:

$$L(\theta, w) = 1 - \min_{n \in \{1,2,3\}} \{AUROC_{\theta,w,n}\}$$

where x is comprised of θ, a set of hyperparameters, and w, a set of residue window sizes. The evaluation metric, $AUROC_n$, is the area under the roc curve for the n^{th} left-out evaluation fold in a three-fold cross-validation scheme. We then used TPE to randomly sample an initial uniform distribution of each of our hyperparameters and window sizes and evaluate

the loss function for each random set of inputs. TPE then replaces this initial distribution with a new distribution built on the results from regions of the sampled distribution that minimize *L(x)*:

$$p(x|y) = \begin{cases} l(x) & if\ y < y^* \\ g(x) & if\ y \geq y^* \end{cases}$$

where y^* is a quantile γ of the observed y values so that $p(y < y^*) = \gamma$. Importantly, y^* is guaranteed to be greater than the minimum observed loss, so that some points are used to build *l(x)*. TPE then chooses candidate hyperparameters to sample as those representing the greatest expected improvement, *EI*, according to the expression:

$$EI_{y^*}(x) = \frac{\gamma y^* l(x) - l(x)\int_{-\infty}^{y^*} yp(y)dy}{\gamma l(x) + (1-\gamma)g(x)} \propto \left(\gamma + \frac{g(x)}{l(x)}(1-\gamma)\right)^{-1}$$

In order to maximize *EI*, the algorithm picks points *x* with high probability under *l(x)* and low probability under *g(x)*. Each iteration of the algorithm returns x^*, the next set of hyperparameters to sample, with the greatest *EI* based on previously sampled points.

Training the classifier

The ECLAIR classifier was trained in three stages, using a custom wrapper of the scikit-learn (PEDREGOSA *et al.* 2011) random forest (BREIMAN 2001) classifier to allow for use of TPE to search both algorithm hyperparameters and residue window sizes simultaneously. In all cross-validations performed, we allowed TPE to search the following hyperparameters, beginning with uniform distributions of the indicated ranges: (1) minimum samples per leaf (0-1000), (2) maximum fraction of features per tree (0-1), and (3) split criterion (entropy or Gini diversity index). The number of estimators (decision trees) in each random forest was

fixed at either 200 for training the feature selection classifiers, or 500 for training the full ensemble. We also allowed TPE to search over residue window sizes (± 0-5 residues for a total window of up to 11 residues, centered on the residue of interest). This was achieved by allowing extra features for neighboring residues to be included at the time of algorithm initialization.

In the first stage of training, cross-validation using TPE was performed on classifiers trained using only features from 1 of the 5 feature categories. The feature or set of features from each category with the minimum loss was selected to represent that category in building the ensemble classifier (**Supplementary Table S4**). In the second stage, the ensemble classifier was built of 8 random forest classifiers, each trained on different subsets of feature categories, and hyperparameters and window sizes were again chosen using cross-validation and TPE (**Supplementary Table S5**). In the final stage, following performance measurement on the testing set, the 8 sub-classifiers were retrained using the full set of 3,447 interactions with at least 50% UniProt residue coverage in the PDB, using the same hyperparameters and window sizes found in the previous step.

Evaluating the ensemble

After training and optimizing using only the training set, we predicted interface residues in a completely orthogonal testing set. For each sub-classifier of the ensemble, all residues in the testing set that could be predicted (given the full set of necessary features or a superset) were ranked according to their raw prediction scores to produce ROC and precision-recall plots.

Benchmarking against other methods

Interfaces for interactions in our testing set were computed using several popular interface prediction methods (LIANG *et al.* 2006; KUFAREVA *et al.* 2007; POROLLO AND MELLER 2007; DE VRIES AND BONVIN 2011; JORDAN *et al.* 2012). We compiled a set of representative protein structures from the PDB for each protein in our testing set, selecting the structure with the highest UniProt residue content based on SIFTS and excluding any PDB structures of interacting protein pairs from our testing set. We then evaluated the precision, recall, and false positive rate for proteins that were able to be classified by all methods. These represent point estimates of these metrics for the external methods with binary prediction scores.

We also compared ECLAIR to 10 popular methods for interface prediction by predicting interfaces in a standard benchmark set of protein complexes (HWANG *et al.* 2010) (**Supplementary Table S3.1**). Here, we followed the experimental procedures laid out by Maheshwari *et al.* (MAHESHWARI AND BRYLINSKI 2015), and excluded complexes in which the receptor is <50 or >600 amino acids, where the interface is made up of <20 residues, or where multiple interfaces are present.

Predicting new interfaces

We retrained the ensemble using all available co-crystal structures, including those from both testing and training sets, a standard machine learning practice that makes maximal use of labeled data (WITTEN *et al.* 2016). Using this fully trained ensemble of classifiers, we predicted interface residues for the remaining 184,605 interactions not resolved by either PDB structures or homology models. Sub-classifiers were ordered based on the number and information content of features used in their training. Each

residue was then predicted by only the top-ranking classifier of the ensemble trained on the full set or a subset of available features for that residue.

Interface enrichment and 3D atomic clustering

Interface domain enrichment, residue enrichment, and 3D atomic clustering can be calculated through the Interactome INSIDER web interface. For enrichments presented in this study, we accessed all disease mutations from the Human Gene Mutation Database (HGMD)(STENSON *et al.* 2014) and ClinVar (LANDRUM *et al.* 2016), recurrent cancer mutations appearing ≥ 6 times in COSMIC (FORBES *et al.* 2015), and population variants from the Exome Sequencing Project (FU *et al.* 2013) to compute the log odds ratio:

$$LOR = ln\left(\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}\right)$$

where p_1 is the probability of a mutation or variant being at the interface and p_2 is the probability of any residue being at the interface. We computed the log odds ratio for residues in each of the interface prediction potential categories. We also computed the log odds ratio for interactions with known interfaces from PDB co-crystal structures, defined as all known interface residues from NACCESS calculations and all residues in Pfam(PUNTA *et al.* 2012) domains with ≥ 5 interface residues. For the disease mutation enrichment analysis (**Figure 3.4a**, we used all disease mutations available from HGMD, and the following numbers of mutations occurred in each category: 10,196 Very Low, 10,547 Low, 2,970 Medium, 1,135 High, and 305 Very High. We also computed enrichment of 18,638 mutations in known interfaces and 17,760 mutations in known non-interfaces (from co-crystal structure evidence).

To perform 3D atomic clustering of amino acid loci of interest, we used an established method (MEYER *et al.* 2016) for clustering and empirical *p*-value calculation and applied it to multi-protein clustering, wherein clusters can occur across an interaction interface. Here, we perform complete-linkage clustering (SØRENSEN 1948) in the shared 3D space of both proteins, and iteratively, and randomly rearrange mutations in each protein to produce an empirical null distribution of cluster sizes.

Mutagenesis validation experiments

We performed mutagenesis experiments in which we introduced random human population variants from the Exome Sequencing Project(FU *et al.* 2013) into known and predicted interfaces. We randomly selected mutations of predicted interface residues in each of the top four ECLAIR categories (Low – Very High). As positive and negative controls, we also selected random mutations of known interface and non-interface residues in co-crystal structures in the PDB. The selected mutations were then introduced into the proteins according to our previously published Clone-seq pipeline (WEI *et al.* 2014) and their impact (either disrupting or maintaining the interaction) was assessed using our yeast two-hybrid assay (**Supplementary Note S3.6**). In this manner, we tested the impact of 2,164 mutations: 1,664 in our predicted interfaces and 500 in known interface and non-interface residues from co-crystal structures. In **Figure 3.2c**, we report the fraction of tested interface residue mutations that caused a disruption of the given interaction for each of the interface residue bins.

Web server

Interactome INSIDER is deployed as an interactive web server (http://interactomeinsider.yulab.org) containing known and predicted interfaces for 197,151 protein interactions in 8 species, as well as variants and functional annotations mapped relative to the residues in the human proteome. For each interaction, the most reliable, high-resolution model is presented, i.e. co-crystal structures are always displayed in lieu of homology models, and all remaining unresolved interactions are predicted by our ECLAIR classifier. Co-crystal structures are derived from the PDB, with extraneous chains removed for each interaction, and homology models are computed by MODELLER (SALI AND BLUNDELL 1993) and downloaded from Interactome3D (MOSCA *et al.* 2013). For both types of structural model, we computed all residues at the interface over all available models, and allow users to view any model from which a unique interface residue has been calculated. For predicted interfaces, a non-redundant set of single protein models are shown when available, with locations of predicted interface residues indicated. In total, the resource contains 7,135 interactions with co-crystal structures, 5,411 with homology models, and 184,605 with predicted interfaces.

Interactome INSIDER also includes pre-calculated enrichment of mutations derived from several sources: 56,159 disease mutations from HGMD (STENSON *et al.* 2014) and ClinVar (LANDRUM *et al.* 2016) and 1,300,352 somatic cancer mutations from COSMIC (FORBES *et al.* 2015). It also includes 194,396 population variants from the 1000 Genomes Project (GENOMES PROJECT *et al.* 2015), 425,115 from the Exome Sequencing Project (FU *et al.* 2013), and 54,165 catalogued by UniProt (UNIPROT-CONSORTIUM 2015). Predictions of deleteriousness for all variants and any

user-submitted variants within the curated interactomes are obtained from PolyPhen-2 (ADZHUBEI *et al.* 2010) and SIFT (KUMAR *et al.* 2009), and biophysical property change guides (i.e. polar to non-polar, hydrophobic to hydrophilic) are also displayed for convenience. Mutation and variant enrichment analyses can be triggered by the user for existing variants or for user-submitted sets within interacting protein domains, residues, and 3D clustering using the atomic coordinates of structures when available.

AUTHOR INFORMATION

Michael J Meyer, Juan Felipe Beltrán & Siqi Liang

These authors contributed equally to this work.

AUTHOR CONTRIBUTIONS

M.J.M., J.F.B., S.L., and H.Y. conceived the study. H.Y. oversaw all aspects of the study. M.J.M., J.F.B., S.L., and A.R. performed computational analyses. M.J.M. and J.F.B. designed ECLAIR. J.F.B. designed the web interface. R.F., J.L., and X.W. performed laboratory experiments. M.J.M. wrote the manuscript with input from J.F.B., S.L., and H.Y. All authors edited and approved of the final manuscript.

COMPETING INTERESTS

The authors declare no competing financial interests.

ACKNOWLEDGEMENTS

The authors would like to thank G. Hooker, D. Bindel, and K. Weinberger for helpful discussions and J. VanEe for technical support. This work was supported by National Institute of General Medical Sciences grants (R01

GM097358, R01 GM104424, R01 GM124559); National Cancer Institute grant (R01 CA167824); Eunice Kennedy Shriver National Institute of Child Health and Human Development grant (R01 HD082568); National Human Genome Research Institute grant (UM1 HG009393); National Science Foundation grant (DBI-1661380); and Simons Foundation Autism Research Initiative grant (367561) to H.Y.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article.

Supplementary Note S3.1: Curating interactomes

Supplementary Note S3.2: Current methods for interaction interface prediction

Supplementary Note S3.3: Feature selection and engineering

Supplementary Note S3.4: Constructing the ECLAIR ensemble classifier

Supplementary Note S3.5: Training classifiers with incomplete data

Supplementary Note S3.6: Mutagenesis experiment methods

Supplementary Figure S3.1. Features for predicting protein interaction interfaces. (a) A schematic showing the five feature categories from which feature sets are optimized to train ECLAIR. (b) The portions of high-quality binary interactomes for which each feature type is available. (c) Feature aggregation strategies employed for combining multiple points of evidence

into single co-evolution- and structure-based features. For co-evolution, we select the top co-evolved residue, the mean of features for the top 10 co-evolved residues, or the mean over all co-evolved residues in the partner protein. For proteins with multiple structures, we take the mean, minimum, or maximum SASA over all available structures.

Supplementary Figure S3.2. Balance between testing/training and prediction sets of sequence- and structure-based feature depths. (a) Sources (PDB or ModBase) and number of structures used to calculate solvent-accessible surface area. (b) Number of homologous sequences used to calculate evolutionary features. (c) Sources of docked models for calculating docking-based features.

Supplementary Figure S3.3. A comparison of two methods for handling missing data in classification: (1) imputation and (2) an ensemble of fully-trained classifiers. During training, imputation must fill in gaps in feature coverage, whereas an ensemble trains independent classifiers on each feature-availability scenario. Since structural feature coverage is highly correlated with the existence of known interface residues in training, imputation will fail to predict interface residues outside of regions with structural feature coverage (red). An ensemble will predict interface residues based only on the features available and will not be biased by the missing structural feature.

Supplementary Figure S3.4. Training and optimizing the ECLAIR classifier. (a) Training the ECLAIR classifier. (b) Four methods for optimizing machine learning algorithm hyperparameters, showing the order of trials and granularity of hyperparameter sampling spaces for optimizing two hyperparameters. (c) Cross-validation strategy using TPE to optimize hyperparameters and window sizes for both feature selection and ensemble classifier training. (d) Cross-validation results using TPE trials to select top performing feature or set of features (in red) in each feature category. (e) Comparison of four hyperparameter optimization methods' performance (top panel) and hyperparameter and residue window sampling patterns (bottom panels) on one of the eight sub-classifiers of the ECLAIR ensemble.

Supplementary Figure S3.5. Performance of ECLAIR sub-classifiers on testing set. (a) Receiver operating characteristic (ROC) curves for each sub-classifier. (b) Precision-recall curves for each sub-classifier. (c) Distribution of raw prediction scores for each sub-classifier. For all panels, sub-classifiers plotted in blue used only sequence-based features; sub-classifiers in red used additional structure-based features. (d) Raw prediction scores compared to actual probabilities of residues in each bin to be at the interface.

Supplementary Figure S3.6. (a) Number of residues predicted in each prediction confidence category. (b) Cumulative distribution of interactions

with ≥ n residues classified as interface for each of the highest interface potential categories.

Supplementary Figure S3.7. ROC and precision-recall curves comparing ECLAIR with other popular interface residue prediction methods. Here, only known surface residues were used in benchmarking all methods.

Supplementary Table S3.1. Comparison of ECLAIR using docking benchmark 4.0

Supplementary Table S3.2. PSI-MI binary evidence codes

Supplementary Table S3.3. Training and Testing Sets

Supplementary Table S3.4. Feature Selection

Supplementary Table S3.5. Full sub-classifier training

Supplementary Table S3.6. Comparison of ECLAIR performance with and without co-evolution

Supplementary Table S3.7. ECLAIR prediction category performance using docking benchmark 4.0

Supplementary Table S3.8. Initially-trained ECLAIR vs. fully-trained ECLAIR performance

REFERENCES

Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova *et al.*, 2010 A method and server for predicting damaging missense mutations. Nat Methods 7**:** 248-249.

Aitman, T. J., C. Boone, G. A. Churchill, M. O. Hengartner, T. F. Mackay *et al.*, 2011 The future of model organisms in human disease research. Nat Rev Genet 12**:** 575-582.

Alfarano, C., C. E. Andrade, K. Anthony, N. Bahroos, M. Bajec *et al.*, 2005 The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res 33**:** D418-424.

Arabidopsis Interactome Mapping, C., 2011 Evidence for network evolution in an Arabidopsis interactome map. Science 333**:** 601-607.

Artimo, P., M. Jonnalagedda, K. Arnold, D. Baratin, G. Csardi *et al.*, 2012 ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res 40**:** W597-603.

Barabasi, A. L., N. Gulbahce and J. Loscalzo, 2011 Network medicine: a network-based approach to human disease. Nat Rev Genet 12**:** 56-68.

Bergstra, J. S., R. Bardenet, Y. Bengio and B. Kégl, 2011 Algorithms for hyper-parameter optimization, pp. 2546-2554 in *Advances in Neural Information Processing Systems*.

Berman, H. M., 2000 The Protein Data Bank. Nucleic Acids Research 28.

Breiman, L., 2001 Random Forests. Machine Learning 45**:** 5-32.

Brown, K. R., and I. Jurisica, 2005 Online predicted human interaction database. Bioinformatics 21**:** 2076-2082.

Brunk, E., N. Mih, J. Monk, Z. Zhang, E. J. O'Brien *et al.*, 2016 Systems biology of the structural proteome. BMC Syst Biol 10**:** 26.

Chatr-Aryamontri, A., B. J. Breitkreutz, R. Oughtred, L. Boucher, S. Heinicke *et al.*, 2015 The BioGRID interaction database: 2015 update. Nucleic Acids Res 43**:** D470-478.

Cingolani, P., A. Platts, L. Wang le, M. Coon, T. Nguyen *et al.*, 2012 A program for annotating and predicting the effects of single nucleotide

polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6**:** 80-92.

D'Amico, A., C. Graziano, G. Pacileo, S. Petrini, K. J. Nowak *et al.*, 2006 Fatal hypertrophic cardiomyopathy and nemaline myopathy associated with ACTA1 K336E mutation. Neuromuscul Disord 16**:** 548-552.

Das, J., R. Fragoza, H. R. Lee, N. A. Cordero, Y. Guo *et al.*, 2014 Exploring mechanisms of human disease through structurally resolved protein interactome networks. Mol Biosyst 10**:** 9-17.

Das, J., and H. Yu, 2012 HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol 6**:** 92.

David, A., R. Razali, M. N. Wass and M. J. Sternberg, 2012 Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. Hum Mutat 33**:** 359-363.

de Vries, S. J., and A. M. Bonvin, 2011 CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. PLoS One 6**:** e17695.

Donkervoort, S., M. Yang, M. Leach, L. Medne, S. Yum *et al.*, 2015 Cardiomyopathy in patients with ACTA1-myopathy. Abstracts/Neuromuscular Disorders 25**:** S184-S316.

Dunker, A. K., C. J. Oldfield, J. Meng, P. Romero, J. Y. Yang *et al.*, 2008 The unfoldomics decade: an update on intrinsically disordered proteins. BMC Genomics 9 Suppl 2**:** S1.

Ellegren, H., 2008 Comparative genomics and the study of evolution by natural selection. Mol Ecol 17**:** 4586-4596.

Forbes, S., N. Bindal, S. Bamford, C. Cole, C. Kok *et al.*, 2011 COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Research 39**:** 50.

Forbes, S. A., D. Beare, P. Gunasekaran, K. Leung, N. Bindal *et al.*, 2015 COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 43**:** D805-811.

Fu, W., T. O'Connor, G. Jun, H. Kang, G. Abecasis *et al.*, 2013 Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493**:** 216-220.

Garzon, J. I., L. Deng, D. Murray, S. Shapira, D. Petrey *et al.*, 2016 A computational interactome and functional annotation for the human proteome. Elife 5.

Genomes Project, C., A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison *et al.*, 2015 A global reference for human genetic variation. Nature 526**:** 68-74.

Grantham, R., 1974 Amino acid difference formula to help explain protein evolution. Science 185**:** 862-864.

Guldener, U., M. Munsterkotter, M. Oesterheld, P. Pagel, A. Ruepp *et al.*, 2006 MPact: the MIPS protein interaction resource on yeast. Nucleic Acids Res 34**:** D436-441.

Halperin, I., B. Ma, H. Wolfson and R. Nussinov, 2002 Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins 47**:** 409-443.

Hermjakob, H., L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski *et al.*, 2004 The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. Nat Biotechnol 22**:** 177-183.

Hodis, E., I. Watson, G. Kryukov, S. Arold, M. Imielinski *et al.*, 2012 A landscape of driver mutations in melanoma. Cell 150**:** 251-263.

Hofree, M., H. Carter, J. F. Kreisberg, S. Bandyopadhyay, P. S. Mischel *et al.*, 2016 Challenges in identifying cancer genes by analysis of exome sequencing data. Nat Commun 7**:** 12096.

Hopf, T. A., J. B. Ingraham, F. J. Poelwijk, C. P. Scharfe, M. Springer *et al.*, 2017 Mutation effects predicted from sequence co-variation. Nat Biotechnol 35**:** 128-135.

Hopf, T. A., C. P. Scharfe, J. P. Rodrigues, A. G. Green, O. Kohlbacher *et al.*, 2014 Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife 3.

Hwang, H., T. Vreven, J. Janin and Z. Weng, 2010 Protein-protein docking benchmark version 4.0. Proteins 78**:** 3111-3114.

Hwang, H., T. Vreven and Z. Weng, 2014 Binding interface prediction by combining protein-protein docking results. Proteins 82**:** 57-66.

Jordan, R. A., Y. El-Manzalawy, D. Dobbs and V. Honavar, 2012 Predicting protein-protein interface residues using local surface structural similarity. BMC Bioinformatics 13**:** 41.

Kamburov, A., M. S. Lawrence, P. Polak, I. Leshchiner, K. Lage *et al.*, 2015 Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc Natl Acad Sci U S A 112**:** E5486-5495.

Kandoth, C., M. D. McLellan, F. Vandin, K. Ye, B. Niu *et al.*, 2013 Mutational landscape and significance across 12 major cancer types. Nature 502**:** 333-339.

Kerrien, S., B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter *et al.*, 2012 The IntAct molecular interaction database in 2012. Nucleic Acids Res 40**:** D841-846.

Keshava Prasad, T. S., R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar *et al.*, 2009 Human Protein Reference Database--2009 update. Nucleic Acids Res 37**:** D767-772.

Kim, P. M., L. J. Lu, Y. Xia and M. B. Gerstein, 2006 Relating three-dimensional structures to protein networks provides evolutionary insights. Science 314**:** 1938-1941.

Kucukkal, T. G., M. Petukh, L. Li and E. Alexov, 2015 Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. Curr Opin Struct Biol 32**:** 18-24.

Kufareva, I., L. Budagyan, E. Raush, M. Totrov and R. Abagyan, 2007 PIER: protein interface recognition for structural proteomics. Proteins 67**:** 400-417.

Kuhlbrandt, W., 2014 Cryo-EM enters a new era. Elife 3**:** e03678.

Kumar, P., S. Henikoff and P. Ng, 2009 Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols 4**:** 1073-1081.

Landrum, M. J., J. M. Lee, M. Benson, G. Brown, C. Chao *et al.*, 2016 ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 44**:** D862-868.

Lawrence, M., P. Stojanov, P. Polak, G. Kryukov, K. Cibulskis *et al.*, 2013 Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499**:** 214-218.

Lee, B., and F. M. Richards, 1971 The interpretation of protein structures: estimation of static accessibility. J Mol Biol 55**:** 379-400.

Lek, M., K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks *et al.*, 2016 Analysis of protein-coding genetic variation in 60,706 humans. Nature 536**:** 285-291.

Lensink, M. F., S. Velankar, A. Kryshtafovych, S. Y. Huang, D. Schneidman-Duhovny *et al.*, 2016 Prediction of homo- and hetero-protein complexes by protein docking and template-based modeling: a CASP-CAPRI experiment. Proteins.

Li, M., M. Petukh, E. Alexov and A. R. Panchenko, 2014 Predicting the Impact of Missense Mutations on Protein-Protein Binding Affinity. J Chem Theory Comput 10**:** 1770-1780.

Liang, S., C. Zhang, S. Liu and Y. Zhou, 2006 Protein binding site prediction using an empirical scoring function. Nucleic Acids Res 34**:** 3698-3707.

Licata, L., L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli *et al.*, 2012 MINT, the molecular interaction database: 2012 update. Nucleic Acids Res 40**:** D857-861.

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

Lounnas, V., T. Ritschel, J. Kelder, R. McGuire, R. P. Bywater *et al.*, 2013 Current progress in Structure-Based Rational Drug Design marks a new mindset in drug discovery. Comput Struct Biotechnol J 5**:** e201302011.

MacArthur, D. G., T. A. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure *et al.*, 2014 Guidelines for investigating causality of sequence variants in human disease. Nature 508**:** 469-476.

Maheshwari, S., and M. Brylinski, 2015 Predicting protein interface residues using easily accessible on-line resources. Brief Bioinform 16**:** 1025-1034.

Maron, B. J., 2002 Hypertrophic cardiomyopathy: a systematic review. JAMA 287**:** 1308-1320.

Mewes, H. W., A. Ruepp, F. Theis, T. Rattei, M. Walter *et al.*, 2011 MIPS: curated databases and comprehensive secondary data resources in 2010. Nucleic Acids Res 39**:** D220-224.

Meyer, M. J., R. Lapcevic, A. E. Romero, M. Yoon, J. Das *et al.*, 2016 mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome. Hum Mutat 37**:** 447-456.

Morcos, F., A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks *et al.*, 2011 Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A 108**:** E1293-1301.

Mosca, R., A. Céol and P. Aloy, 2013 Interactome3D: adding structural details to protein networks. Nature methods 10**:** 47-53.

Mosca, R., J. Tenorio-Laranga, R. Olivella, V. Alcalde, A. Ceol *et al.*, 2015 dSysMap: exploring the edgetic role of disease mutations. Nat Methods 12**:** 167-168.

Ngeow, J., W. Yu, L. Yehia, F. Niazi, J. Chen *et al.*, 2015 Exome Sequencing Reveals Germline SMAD9 Mutation That Reduces Phosphatase and Tensin Homolog Expression and Is Associated With Hamartomatous Polyposis and Gastrointestinal Ganglioneuromas. Gastroenterology 149**:** 886-889 e885.

Orchard, S., S. Kerrien, S. Abbani, B. Aranda, J. Bhate *et al.*, 2012 Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nat Methods 9**:** 345-350.

Pagel, P., S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach *et al.*, 2005 The MIPS mammalian protein-protein interaction database. Bioinformatics 21**:** 832-834.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion *et al.*, 2011 Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12**:** 2825-2830.

Peng, K., Z. Obradovic and S. Vucetic, 2004 Exploring bias in the Protein Data Bank using contrast classifiers. Pac Symp Biocomput**:** 435-446.

Pierce, B. G., Y. Hourai and Z. Weng, 2011 Accelerating protein docking in ZDOCK using an advanced 3D convolution library. PLoS One 6**:** e24657.

Porollo, A., and J. Meller, 2007 Prediction-based fingerprints of protein-protein interactions. Proteins 66**:** 630-645.

Punta, M., P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate *et al.*, 2012 The Pfam protein families database. Nucleic Acids Res 40**:** D290-301.

Raphael, B. J., J. R. Dobson, L. Oesper and F. Vandin, 2014 Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. Genome Med 6**:** 5.

Rolland, T., M. Tasan, B. Charloteaux, S. J. Pevzner, Q. Zhong *et al.*, 2014 A proteome-scale map of the human interactome network. Cell 159**:** 1212-1226.

Roth, S., P. Sistonen, R. Salovaara, A. Hemminki, A. Loukola *et al.*, 1999 SMAD genes in juvenile polyposis. Genes Chromosomes Cancer 26**:** 54-61.

Ruepp, A., B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach *et al.*, 2010 CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Res 38**:** D497-501.

Sahni, N., S. Yi, M. Taipale, J. I. Fuxman Bass, J. Coulombe-Huntington *et al.*, 2015 Widespread macromolecular interaction perturbations in human genetic disorders. Cell 161**:** 647-660.

Sali, A., and T. L. Blundell, 1993 Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234**:** 779-815.

Salwinski, L., C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie *et al.*, 2004 The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 32**:** D449-451.

Sørensen, T., 1948 A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol. Skr. 5**:** 1-34.

Sparrow, J. C., K. J. Nowak, H. J. Durling, A. H. Beggs, C. Wallgren-Pettersson *et al.*, 2003 Muscle disease caused by mutations in the skeletal muscle alpha-actin gene (ACTA1). Neuromuscul Disord 13**:** 519-531.

Stenson, P. D., M. Mort, E. V. Ball, K. Shaw, A. Phillips *et al.*, 2014 The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 133**:** 1-9.

Tasan, M., G. Musso, T. Hao, M. Vidal, C. A. MacRae *et al.*, 2015 Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. Nat Methods 12**:** 154-159.

Turner, B., S. Razick, A. L. Turinsky, J. Vlasblom, E. K. Crowdy *et al.*, 2010 iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. Database (Oxford) 2010**:** baq023.

UniProt-Consortium, 2015 UniProt: a hub for protein information. Nucleic Acids Res 43**:** D204-212.

Vakser, I. A., 2013 Low-resolution structural modeling of protein interactome. Curr Opin Struct Biol 23**:** 198-205.

Velankar, S., J. Dana, J. Jacobsen, G. van Ginkel, P. Gane *et al.*, 2013 SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Research 41**:** 9.

Vidal, M., M. E. Cusick and A. L. Barabasi, 2011 Interactome networks and human disease. Cell 144**:** 986-998.

Vo, T. V., J. Das, M. J. Meyer, N. A. Cordero, N. Akturk *et al.*, 2016 A Proteome-wide Fission Yeast Interactome Reveals Network Evolution Principles from Yeasts to Human. Cell 164**:** 310-323.

Wang, R. N., J. Green, Z. Wang, Y. Deng, M. Qiao *et al.*, 2014 Bone Morphogenetic Protein (BMP) signaling in development and human diseases. Genes Dis 1**:** 87-105.

Wang, X., X. Wei, B. Thijssen, J. Das, S. M. Lipkin *et al.*, 2012 Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol 30**:** 159-164.

Wei, X., J. Das, R. Fragoza, J. Liang, F. M. Bastos de Oliveira *et al.*, 2014 A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. PLoS Genet 10**:** e1004819.

Witten, I. H., E. Frank, M. A. Hall and C. J. Pal, 2016 *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier Science.

Xie, L., X. Ge, H. Tan, L. Xie, Y. Zhang *et al.*, 2014 Towards structural systems pharmacology to study complex diseases and personalized medicine. PLoS Comput Biol 10**:** e1003554.

Yang, F., E. Petsalaki, T. Rolland, D. E. Hill, M. Vidal *et al.*, 2015 Protein domain-level landscape of cancer-type-specific somatic mutations. PLoS Comput Biol 11**:** e1004147.

Yu, H., P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan *et al.*, 2008 High-quality binary protein interaction map of the yeast interactome network. Science 322**:** 104-110.

Zhang, Q. C., D. Petrey, L. Deng, L. Qiang, Y. Shi *et al.*, 2012 Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature 490**:** 556-560.

CHAPTER 4

CONSTRUCTING A HUMAN-MICROBE PROTEIN-PROTEIN INTERACTION NETWORK TO IDENTIFY MECHANISMS BEHIND MICROBIOME-ASSOCIATED DISEASE[1]

ABSTRACT

Host-microbe interactions are crucial for normal physiological and immune system development and are implicated in a wide variety of diseases, including inflammatory bowel disease (IBD), obesity, colorectal cancer (CRC), and type 2 diabetes (T2D). Despite large-scale case-control studies aimed at identifying microbial taxa or specific genes involved in pathogeneses, the mechanisms linking them to disease have thus far remained elusive. To better identify potential mechanisms linking human-associated bacteria with host health, we leveraged publicly-available interspecies protein-protein interaction (PPI) data to identify clusters of homologous microbiome-derived proteins that bind human proteins. By detecting human-interacting bacterial genes in metagenomic case-control microbiome studies and applying a tailored machine learning algorithm, we

[1] Provisional patents have been filed for both the process described in this paper, and the therapeutic/diagnostic protein candidates found through this process through Cornell University. Inventors: Ilana Brito and Juan Felipe Beltrán.

are able to identify bacterial-human PPIs strongly linked with disease. In 9 independent case studies, we discover the microbiome broadly targets human immune, oncogenic, apoptotic, and endocrine signaling pathways, among others. This host-centric analysis strategy illuminates human pathways targeted by the commensal microbiota, provides a mechanistic hypothesis-generating platform for any metagenomics cohort study, and extensively annotates bacterial proteins with novel host-relevant functions.

INTRODUCTION

Metagenomic case-control studies of the human gut microbiome have implicated bacterial genes in a myriad of diseases. Yet, the sheer diversity of genes within the microbiome and the lack of functional annotations have thwarted efforts to identify the mechanisms by which these bacterial genes impact host health. In the cases where functional annotations exist, they tend to refer to molecular function (*e.g.* DNA binding, post-translational modification) rather than their role in biological pathways (LLOYD-PRICE *et al.* 2017), and fewer even relate to host cell signaling and homeostasis. Obtaining a clearer idea of the health impacts of each gene has thus far required experimental approaches catered to each gene or gene function (NEŠIĆ *et al.* 2014; PLOVIER *et al.* 2017).

We hypothesized that host-microbiome protein-protein interactions may underlie health status and could serve to provide additional information, through annotation of human pathways, about the role of bacteria in modulating health. Protein-protein interactions (PPIs) have revealed the mechanisms by which pathogens interact with host tissue through in-depth structural studies of individual proteins (HAMIAUX *et al.* 2006; NEŠIĆ *et al.* 2014; GUVEN-MAIOROV *et al.* 2017b), as well as large-scale whole-organism interaction screens(DYER *et al.* 2010; SHAH *et al.* 2018). Although there are canonical microbe-associated patterns (MAMPs) that directly trigger host-signaling pathways through pattern recognition receptors present on epithelial and immune tissues(BHAVSAR *et al.* 2007), such as flagellin with Toll-like receptor 5 (TLR5), several recent observations have further underscored a role for commensal-host PPIs in health: An integrase encoded by several *Bacteroides* species binds human islet-specific glucose-6-phosphatase-catalytic-subunit-related protein (IGRP) thereby protecting against colitis (HEBBANDI NANJUNDAPPA *et al.* 2017); a protease secreted by *Enterococcus faecalis* binds incretin hormone glucagon-like peptide 1 (GLP-1), a therapeutic target for type 2 diabetes (T2D)(LEVALLEY *et al.* 2019); and a slew of ubiquitin mimics encoded by both pathogens (GUVEN-MAIOROV *et al.* 2017a) and gut commensals (STEWART *et al.* 2018) play a role in modulating membrane trafficking.

In the absence of experimental data, *in silico* homology modeling has been used to great effect to inform pathophysiology using inferred host-pathogen PPI networks (SEN *et al.* 2016; GUVEN-MAIOROV *et al.* 2017a, 2019), but such approaches have not yet been applied to the human gut microbiome. Here, we leverage roughly 8,000 experimentally-verified binary inter-species PPIs from the IMEx Consortium members, as curated in the publicly-available IntAct database (ORCHARD *et al.* 2014) (**Figure 4.1A**), to gain insight into host-microbiome interactions. By propagating interactions to all bacterial proteins sharing the same UniRef homology clusters (SUZEK *et al.* 2015), we expanded the set of human-microbe PPIs to include over 1.6 million bacterial proteins and 4,186 human proteins, comprising more than 8 million interspecies interactions (**Figure 4.1A, Supplementary Figure S4.1**).

Focusing on diseases where abundant information links microbiota with disease phenotypes and where large case-control cohorts exist—namely colorectal cancer (CRC)(ZELLER *et al.* 2014; FENG *et al.* 2015; YU *et al.* 2017; HANNIGAN *et al.* 2018), T2D(QIN *et al.* 2012; KARLSSON *et al.* 2013), inflammatory bowel disease (IBD)(NIELSEN *et al.* 2014; SCHIRMER *et al.* 2018) and obesity(LE CHATELIER *et al.* 2013) (Supplementary Table 1)—we then mapped quality-filtered metagenomic sequencing reads from nine case-control study cohorts to our database of bacterial human-protein interactors.

RESULTS

Using stringent detection criteria, we find roughly 255,000 potential human-bacterial interactions across the human microbiome. Inferred bacterial interactors found in the human microbiome have strong homology with proteins with experimentally-verified human interaction data (**Figure 4.1B, Supplementary Figure S4.2**). We applied a random forest machine learning algorithm to differentiate between cases and controls in each study based on binary detection vectors of both bacterial protein clusters and targeted human proteins. We calculated a balance-aware forest-based feature importance metric (ALTMANN *et al.* 2010; KURSA and RUDNICKI 2010) to rank the disease-association of each bacterial or human protein relative to their detection frequency, hereby called 'zboost'.

We noticed that a disproportionate number of bacteria-human PPIs in IntAct were derived from high-throughput screens performed on three intracellular pathogens: *Yersinia pestis*, *Francisella tularensis* and *Bacillus anthracis* (DYER *et al.* 2010). Nevertheless, we find that patient-detected bacterial clusters are taxonomically diverse, not biased towards the originating classes of these three pathogens—Bacilli or Gammaproteobacteria—and rather, reflect the breadth of taxa typically associated with human gut microbiomes (**Figure 4.1C, Supplementary Figure S4.3**).

Figure 4.1. Identifying human-interacting bacterial proteins within the gut microbiomes of T2D, obesity, IBD and CRC cohorts reveals enrichment for disease-associated pathways in human cells.

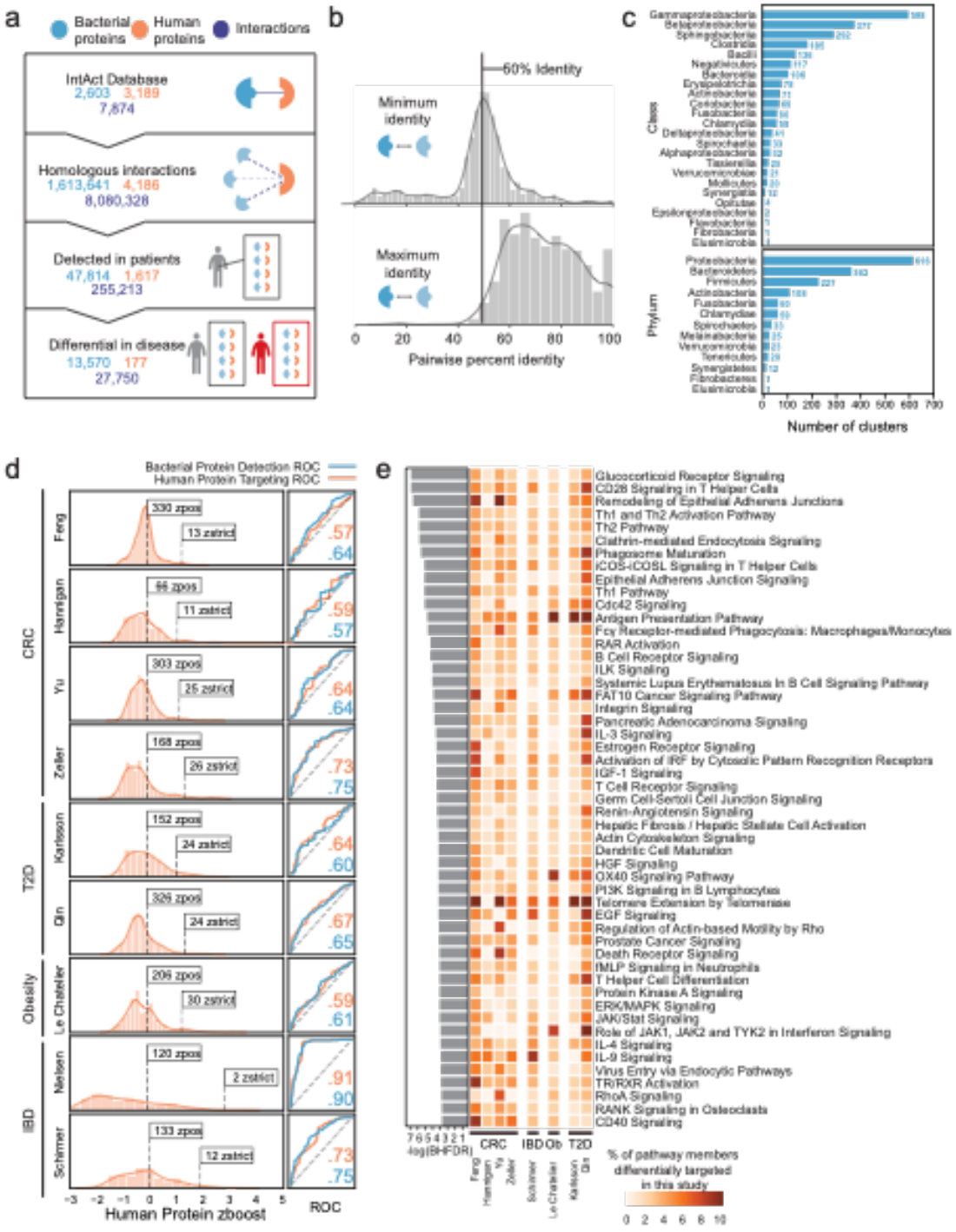
(a) The number of interspecies bacterial proteins (blue), human proteins (orange) and interactions (dark blue) in the IntAct database; those inferred using homology clusters (UniRef); those determined to be present in the gut microbiomes from nine metagenomic studies; and those deemed important (zboost greater than zstrict, the magnitude of the minimum zboost) through our comparative metagenomic machine learning approach. If we use the zpos cutoff (zboost greater than zero), we find 40,663 important bacterial proteins (comprising 582 protein homology clusters), 1,156 important human proteins and 149,045 interactions between them. For zstrict, the bacterial proteins comprise 128 protein homology clusters.

(b) Histograms showing the maximum and minimum percent identity per bacterial cluster between bacterial proteins with experimental verification and proteins detected in human microbiomes. The histograms are annotated with a gaussian kernel density estimate of the distribution.

(c) The number of bacterial clusters that include members from each bacterial phyla and class. Note that most clusters contains proteins from more than one class and phylum.

(d) Distributions of human proteins targeted in the gut microbiomes associated with each study according to their zboost scores (left). Numbers of proteins with zboost scores over zpos and zstrict are noted. Receiver-operator characteristic (ROC) curves for our random forests predictions for each dataset (right) based on bacterial (blue) proteins or their human interactors (orange), along with their corresponding AUC values.

(e) Human cellular pathways overrepresented in the $zpos_{hum}$ subset (Benjamini-Hochberg false discovery rate (BHFDR) ≤ 0.05). –log(BHFDR) of each pathway is displayed on the barplot to the left. The heatmap is colored according to the percent of pathway members differentially targeted in each case-control cohort.


a
Bacterial proteins
Human proteins
Interactions
IntAct Database
2,603 3,189
7,874
Homologous interactions
1,613,641 4,186
8,080,328
Detected in patients
47,814 1,617
255,213
Differential in disease
13,570 177
27,750
b
60% Identity
Minimum identity
Maximum identity
Pairwise percent identity
c
Class
Phylum
Number of clusters
d
Bacterial Protein Detection ROC
Human Protein Targeting ROC
CRC
T2D
Obesity
IBD
Feng
Hannigan
Yu
Zeller
Karlsson
Qin
Le Chatelier
Nielsen
Schirmer
Human Protein zboost
ROC
e
Glucocorticoid Receptor Signaling
CD28 Signaling in T Helper Cells
Remodeling of Epithelial Adherens Junctions
Th1 and Th2 Activation Pathway
Th2 Pathway
Clathrin-mediated Endocytosis Signaling
Phagosome Maturation
iCOS-iCOSL Signaling in T Helper Cells
Epithelial Adherens Junction Signaling
Th1 Pathway
Cdc42 Signaling
Antigen Presentation Pathway
Fcγ Receptor-mediated Phagocytosis: Macrophages/Monocytes
RAR Activation
B Cell Receptor Signaling
ILK Signaling
Systemic Lupus Erythematosus in B Cell Signaling Pathway
FAT10 Cancer Signaling Pathway
Integrin Signaling
Pancreatic Adenocarcinoma Signaling
IL-3 Signaling
Estrogen Receptor Signaling
Activation of IRF by Cytosolic Pattern Recognition Receptors
IGF-1 Signaling
T Cell Receptor Signaling
Germ Cell-Sertoli Cell Junction Signaling
Renin-Angiotensin Signaling
Hepatic Fibrosis / Hepatic Stellate Cell Activation
Actin Cytoskeleton Signaling
Dendritic Cell Maturation
HGF Signaling
OX40 Signaling Pathway
PI3K Signaling in B Lymphocytes
Telomere Extension by Telomerase
EGF Signaling
Regulation of Actin-based Motility by Rho
Prostate Cancer Signaling
Death Receptor Signaling
fMLP Signaling in Neutrophils
T Helper Cell Differentiation
Protein Kinase A Signaling
ERK/MAPK Signaling
JAK/Stat Signaling
Role of JAK1, JAK2 and TYK2 in Interferon Signaling
IL-4 Signaling
IL-9 Signaling
Virus Entry via Endocytic Pathways
TR/RXR Activation
RhoA Signaling
RANK Signaling in Osteoclasts
CD40 Signaling
-log(BH-FDR)
CRC
IBD
Ob
T2D
% of pathway members differentially targeted in this study

Overall, we are able to reasonably predict disease based on the detection of either bacterial or human interactors (**Figure 4.1D**). Interestingly, our approach showed greater predictive capability in some datasets over others, even for the same disease. We suspect this variation may be due to the wide range of etiologies that give rise to these diseases, as is the case for CRC, which can be driven by germ-line mutation, immune status, diet and environmental factors (WEITZ *et al.* 2005). Taking these studies together, the variation between the detected human interactors across participants could not be explained purely by the health status of the individual, the specific cohort, or any other available characteristic associated with the samples (**Supplementary Figure S4.5**). The only exception was one IBD study where ethnicity correlated with disease status, and was therefore excluded from the remainder of our analyses.

We identify subsets of important bacterial interactors and their human targets that are predictive of disease (**Figure 4.1D; Supplementary Figure S4.4**). We applied two thresholds to generate protein sets for analysis: those with zboosts greater than 0 ($zpos_{bact}$ and $zpos_{hum}$) and those with zboosts greater than the magnitude of the minimum zboost ($zstrict_{bact}$ and $zstrict_{hum}$). Within the larger human subsets ($zpos_{hum}$), we find proteins with established roles in cellular pathways coherent with the pathophysiology of CRC, IBD, obesity and T2D. For example, we find that DNA fragmentation factor

subunit alpha (DFFA) is important in T2D (in the Qin *et al.* cohort), and is involved in death receptor signaling, an important pathway for the destruction of insulin-producing β-cells (SIA and HÄNNINEN 2006). Collagen alpha-1(I) chain (COL1A1) is also a significant target associated with T2D (in the Karlsson *et al.* cohort), and plays a role in dendritic cell maturation and hepatic fibrosis/hepatic stellate cell activation pathways, capturing known comorbidities between T2D and hepatic steatosis and nonalcoholic steatohepatitis (NASH)(RICHARD and LINGVAY 2011). Proteins important in CRC studies spanned expected bacteria-associated pathways, such as the direct sensing of enterotoxins, *e.g.* heat-stable enterotoxin receptor GUCY2C (in the Feng *et al.* and Zeller *et al.* cohorts); but also classical cancer-associated pathways, such as the maintenance of DNA integrity, *e.g.* protection of telomeres protein 1 (POT1) (in the Feng *et al.* and also the Qin *et al.* cohorts) and X-ray repair cross-complementing protein 6 (XRCC6) (in the Feng *et al.* and Yu *et al.* cohorts), the latter of which is required for double-strand DNA break repair. We also find common targeting of human pathways across diseases that speak to their known shared etiologies and symptoms, for instance, actin-related protein 2/3 complex subunit 2 (ARPc2) (in the Yu *et al.*, Schirmer *et al.* and Karlsson *et al.* cohorts), a protein involved in remodeling epithelial adherens junctions, a process

strongly associated with IBD(FRANKE *et al.* 2008), CRC (DAULAGALA *et al.* 2019) and, most recently, T2D(KANG *et al.* 2019).

These examples are illustrative of a larger trend of disease-associations driven by host-microbiome interactions. A more robust statistical analysis for overall pathway enrichment in the $zpos_{hum}$ subset confirms significant enrichment in pathways involving the immune system, apoptosis, oncogenesis, and endocrine signaling, among others (**Figure 4.1E**). Although we see significant overlap in the pathways targeted across diseases, which may reflect their associated relative risks(JURJUS *et al.* 2016; DE KORT *et al.* 2017; STIDHAM and HIGGINS 2018; KANG *et al.* 2019; JESS *et al.* 2019), there is some disease-specificity. For example, more human proteins in the antigen presentation pathway are differentially targeted in T2D and obesity cohorts than elsewhere. In the CRC cohorts, more $zpos_{hum}$ proteins target the CD40 signaling, RANK signaling in osteoclasts, and TR/RXR activation pathways than other studies.

We next sought to determine whether the human protein interactors that we associated with disease were enriched for relevant previously-reported gene-disease associations (GDA). We find many human targets associated with microbiome-related disorders, such as CRC, diabetes, autoimmune disease, obesity and IBD (**Figure 4.2A**). Although none of the cohorts we studied focused on the larger spectrum of autoimmune disease,

these disorders are increasingly studied in the context of the gut microbiome(GIANCHECCHI and FIERABRACCI 2019), and therefore we included them in our analysis.

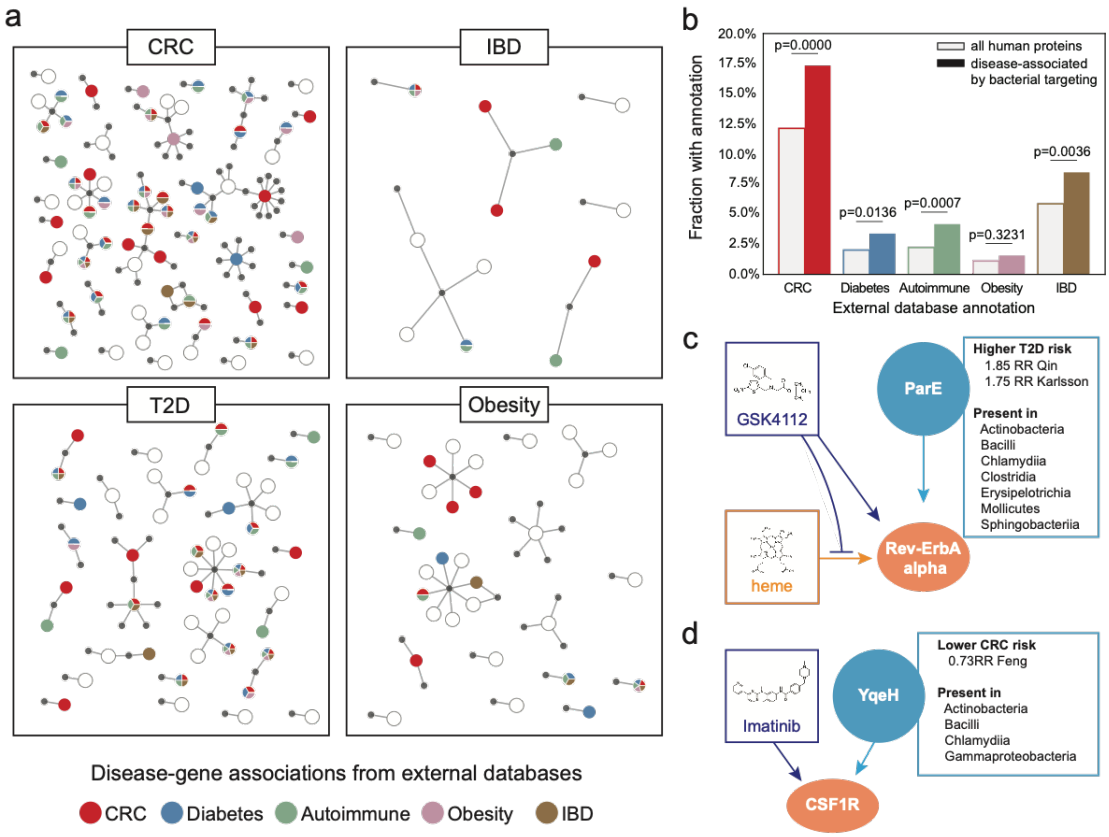


Figure 4.2. Human proteins differentially targeted by the microbiome in disease are enriched for particular gene-disease associations and contain known therapeutic drug targets. **(a)** Important human proteins ($zstrict_{hum}$) are plotted with their bacterial partners (gray), according to their disease-gene associations in the DisGeNet database: CRC (red), diabetes (blue), autoimmune disease (green), obesity (mauve) and IBD (brown). **(b)** Bar chart comparing the proportions of human proteins with disease-gene associations in important human proteins ($zpos_{hum}$) targeted by microbiomes and all human proteins in DisGeNet. **(c)** RevErbA alpha (NR1D1) binds several human proteins (not shown), DNA (not shown) and heme. GSK4112 competitively binds RevErbA alpha, inhibiting binding with heme. ParE is a microbiome protein present in a diverse range of organisms and has a high relative risk associated with T2D. **(d)** Macrophage colony stimulating factor 1 receptor (CSF1R) is targeted by imatinib, among other drugs, as well as the uncharacterized bacterial protein YqeH, a protein that has a low relative risk associated with CRC.

Interestingly, our disease annotations were ubiquitous (82.5% of the $zstrict_{hum}$ subset had at least one GDA), but were not strictly isolated to the matching metagenomic cohort's condition (**Figure 4.2A, Supplementary Table S4.2**). Across the larger $zpos_{hum}$ subset, GDAs for these microbiome-associated disorders were enriched overall, with the exception of obesity, where annotation is generally scarce (**Figure 4.2B**). Surprisingly, in the CRC cohorts were a number of previously identified CRC-associated genetic loci, including well-known cancer genes: tumor protein p53, epidermal growth factor receptor (EGFR), matrix metalloprotease 2 (MMP2), and insulin-like growth factor-binding protein 3 (IGFBP3), among others.

Our data suggest many molecular mechanisms that might be regulating human cellular functions through bacterial proteins. In order to better contextualize these mechanisms, we asked whether any of the human proteins in our dataset were already known to be drug targets. Using the Probes & Drugs database(SKUTA *et al.* 2017), we find many $zstrict_{hum}$ proteins are targeted by drugs (**Supplementary Table S4.3**). In many cases, those drugs are known to either treat or affect the pathogenesis of the microbiomes of patients with those diseases. For example, in both T2D cohorts, we found elevated zboost scores associated with human protein RevErbA alpha (NR1D1), the target of the drugs GSK4112, SR9009 and SR9011, which inhibit the binding of RevErbA alpha with its natural ligand,

heme (**Figure 4.2C**). These drugs have been shown to affect cellular metabolism *in vitro* and affect hyperglycaemia when given to mouse models of metabolic disorder (SOLT *et al.* 2012; VIEIRA *et al.* 2013).

We also find instances where our analysis of a particular disease cohort is consistent not with the therapeutic purpose of a drug targeted by human interactors in that microbiome, but with off-label effects or side effects associated with the drug. For example, we find that imatinib mesylate (brand name: Gleevec), has several human binding partners, including macrophage colony-stimulating factor 1 receptor (M-CSF1R) (**Figure 4.2D**), an important target found in CRC (in the Feng *et al.* cohort), and platelet-derived growth factor receptor-β (PDGFR-B), an important target found in the obesity and T2D (in the Le Chatelier *et al.* and Qin cohorts, respectively). Literature on imatinib supports these findings: although imatinib is best known as a treatment for leukemia, it has been shown to affect glycemic control in patients with T2D(CHOI *et al.* 2016). Furthermore, imatinib can also halt the proliferation of colonic tumor cells and is involved generally in inflammatory pathways, through its inhibition of TNF-alpha production(WOLF *et al.* 2005).

One of the major advantages of our work is that through homology mapping, we vastly improve our overall ability to annotate host-relevant microbiome functions. When we annotated the microbial pathways using

KEGG (Kyoto Encyclopedia of Genes and Genomes)(KANEHISA *et al.* 2017), we found that 41.2% of the $zpos_{bact}$ protein clusters found in human microbiomes lacked any pathway information (**Figure 4.3A**). Yet, these genes can now be annotated according to the pathways of their human targets, obtaining a putative disease-relevant molecular mechanism (**Figure 4.3A, B**). This host-centric annotation is useful beyond large-scale analysis of metagenomic data, but it broadly enables hypothesis-driven research into how these microbial proteins impact host health.

We examined the means by which bacterial proteins may be interacting with host proteins and found that a majority of bacterial protein clusters (90.2% of $zpos_{bact}$) contain proteins that are transmembrane, are secreted by type 3 or type 4 secretion systems, and/or contain eukaryotic-like domains (**Figure 4.3C**), another marker for secretion.

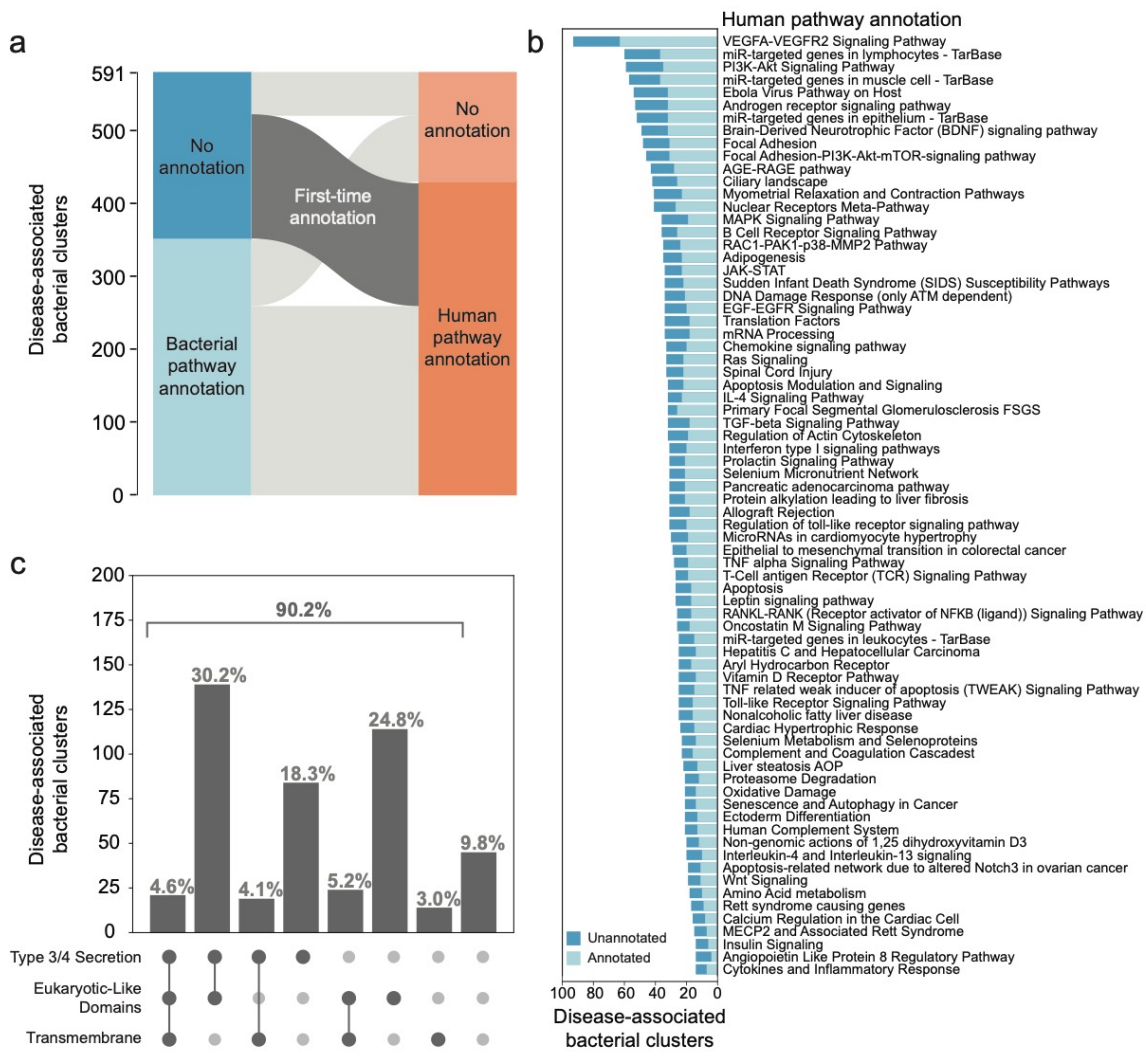


Figure 4.3. Human pathway annotation can be transferred across interactors to improve bacterial pathway annotation.

(a) Paired stacked bar plots showing the disease-associated bacterial cluster pathways annotated by KEGG (left) and their inferred pathways according to the human proteins they target (right), as annotated by WikiPathways(SLENTER *et al.* 2018).

(b) Human pathways (annotated using WikiPathways) targeted by disease-associated bacterial clusters. The 75 human pathways with the most previously unannotated bacterial targeters (annotated using KEGG) are shown.

(c) The number of $zpos_{bact}$ clusters plotted according to their transmembrane and secretion predictions, *i.e.* type 3 or type 4 secretion systems (T3SS or T4SS), and/or the presence of eukaryotic-like domains (ELDs).

Of particular interest were bacterial proteins in this subset that have well-known core functions, *e.g.* protein chaperones DnaK and GroL, RNA polymerases RpoB and RpoC, and canonical glycolysis enzymes, among others. A number of these proteins have been previously identified as 'moonlighting' proteins, which perform secondary functions in addition to their primary role in the cell (HENDERSON 2014). *Mycoplasma pneumoniae* DnaK and enolase, a protein involved in glycolosis, from a number of pathogens, bind to both human plasminogen and extra-cellular matrix components(HENDERSON and MARTIN 2013; HAGEMANN *et al.* 2017). *Mycobacterium tuberculosis* DnaK signals to leukocytes causing the release of the chemokines CCL3-5(LEHNER *et al.* 2000). *Streptococcus pyogenes* glyceraldehyde-3-phosphate dehydrogenase (GAPDH), another protein involved in glycosylis, can be shuffled to the cell surface where it plays a role as an adhesin, and can also contribute to human cellular apoptosis(SEIDLER and SEIDLER 2013). These examples widely illustrate how bacterial housekeeping proteins are used by pathogens to modulate human health. In this study, we uncover commensal proteins that have 'interspecies moonlighting' functions, which are not constrained to pathogenic organisms, but are pervasive throughout our indigenous microbiota.

DISCUSSION

Here, we reveal for the first time an extensive host-microbiome PPI landscape. This work highlights the myriad host mechanisms targeted by the gut microbiome and the extent to which these mechanisms are targeted across microbiome-related disorders. However, this network is far from complete. Few of the interaction studies on which this interaction network is based were performed on commensal bacteria and therefore, we may be missing interactions specific to our intimately associated bacteria. In addition to large-scale PPI studies involving commensal bacteria and their hosts, further in-depth studies will be needed to fully characterize these mechanisms, such as whether these bacterial proteins activate or inhibit their human protein interactors' pathways.

This platform enables a high-throughput glimpse into the mechanisms by which microbes impact host tissue, allowing for mechanistic inference and hypothesis generation from any metagenomic dataset. Much as recent studies have uncovered the mechanistic roles of commensal-derived small molecules in disease(DONIA and FISCHBACH 2015), we shed light on a greater role for commensal-derived proteins. By focusing on proteins, our methods connect pharmacology, human genetic variation and microbiome diversity through tangible mechanisms, owing to the large amount of existing data on human proteins. Pinpointing those microbe-derived proteins

that interact directly with human proteins will pave the way for novel diagnostics and therapeutics for microbiome-driven diseases, more nuanced definitions of the host-relevant functional differences between bacterial strains, and a deeper understanding of the co-evolution of humans and other organisms with their commensal microbiota.

MATERIALS & METHODS

Building a putative bacteria-human protein-protein interaction network

Interactions were downloaded from the IntAct database [August 2018]. Only interactions with evidence codes that indicated binary, experimental determination of the interaction between UniProt identifiers with non-matching taxa were preserved, thereby excluding co-complex associations, small molecule interactions, and predicted interactions. This resulted in a set of 296,103 interspecies PPIs. Interspecies protein interactors were mapped to their UniRef sequence clusters at the 100%, 90%, and 50% identity-to-seed levels, which are publicly available through the UniProt web service. Given two UniRef homology clusters with a known PPI between their members, we map that interaction to all combinations of members from the two clusters. We perform this mapping at all levels of homology (and their combinations). From this large list of putative PPIs, we store only

interactions between bacterial proteins and reviewed SwissProt human proteins. The latter step avoids the over-annotation of human isoforms or homologs, or non-verified human proteins. Overall, we generate 8,808,328 bacteria-human PPIs involving 1,613,641 bacterial proteins and 4,186 reviewed human proteins. This corresponds to 18,097 interactions between 33,123 UniRef clusters containing bacterial proteins and the aforementioned 4,186 reviewed human proteins.

Detection of human-targeting proteins in metagenomic shotgun sequencing data

Reads from nine metagenomic studies (Supplementary Table 1) were downloaded from the Sequence Read Archive (SRA) using fasterq-dump. Reads belonging to more than one replicate from the same patient were concatenated and treated as a single run. Reads were then dereplicated using prinseq (v0.20.2) and trimmed using trimmomatic (v0.36) with the following parameters:

```
Dereplication
perl prinseq-lite.pl -fastq {1} -fastq2 {2} \
        -derep 12345 -out_format 3 -no_qual_header \
        -out_good {3} -out_bad {4};

{1,2} Refer to paired read input files
{3,4} Refer to output filepaths
```

```
Trimming
java -Xmx8g -jar trimmomatic-0.36.jar \
        PE -threads 5 {1}
        ILLUMINACLIP:{2}:2:30:10:8:true \
        SLIDINGWINDOW:4:15 LEADING:3 TRAILING:3
MINLEN:50

{1} Refer to input files
{2} Is the path to a fasta file of Nextera TruSeq
adapters
```

Paired reads were combined into a single file and aligned to a protein library of all 1,613,641 human-interacting bacterial proteins generated above. This read-to-protein alignment was performed using blastx through the diamond command line tool (v0.9.24.125). Read alignments were filtered to only consider results with an identity of at least 90% and no gaps. Bacterial proteins were considered detected with sufficient depth and coverage: more than 10 reads across 95% of the protein sequence, excluding 10 amino acids at each terminus. We assign any bacterial protein detection to its corresponding UniRef homology cluster. Human-interacting bacterial clusters are marked as either 'detected' or 'not detected' for each patient in each study. For each patient, we also generate a file of human proteins that are targeted by their detected bacterial proteins based on our bacteria-human PPI network.

Prioritization of disease-associated bacterial protein clusters and human targets

Each patient from each study can be represented as either (a) a binary vector of detected bacterial protein clusters or (b) a binary vector of targeted human proteins. We removed human proteins that were considered redundant based on the same exact bacterial protein partners in our database. We used either the bacterial protein clusters or the human interactors to separate case and control cohorts using a random forest machine learning algorithm. Though classically we could simply extract the average Gini coefficient of the trained random forest and use that as a proxy for feature importance, binary labels introduce a complication: the balance of each feature can limit or inflate the Gini coefficients for that feature. In order to avoid for this complication, we use an empirical normalization method similar to previous work in the field, that we call zboost.

zboost algorithm

(1) Fit a random forest with 100 estimators (X=protein detection per patient , y=case/control labels), then extract and store the average Gini coefficient for each feature ($Gini^{P}_{real}$)

(2) For each feature, generate a random binary vector with similar balance, where each protein detection in each patient is a Bernoulli trial where:

$$P(1) = overall\ patient\ detection\ rate\ for\ that\ protein$$

(3) Fit a random forest with 300 estimators (X=random protein detection per patient, y=case/control labels), then extract and store the average Gini coefficient for each feature ($Gini^{P}_{rand}$)

(4) Calculate the zboost of each feature as:

$$zboost^{P} = \frac{avg(Gini^{P}_{real}) - avg(Gini^{P}_{rand})}{\max\left(std(Gini^{P}_{real}), std(Gini^{P}_{rand})\right)}$$

(5) Calculate the width of the zboost distribution, given as:

$$width = \max(zboost) - \min(zboost)$$

(6) Repeat 1-5 until the width of the zboost distribution does not increase for 200 iterations.

An extra step of filtering is applied to avoid uninformative proteins (too rarely detected, or ubiquitous): any proteins where the minimum value in their expected contingency matrix (case/control vs. detected/not-detected) is less than 5 is removed from consideration. We apply the zboost algorithm to both the bacterial-protein and human-protein binary representations of each patient for all 9 studies. Additionally, we measure our performance on

these tasks by training a separate random forest, with 200 estimators, using 5-fold cross-validation.

This work was implemented and applied to our datasets using python (v3.7.3), pandas(MCKINNEY 2010) (v0.25.1) and the scikit-learn(PEDREGOSA *et al.*) library (v0.21.3). We used two thresholds when conducting analysis on the resulting data: zpos, where zboosts must be greater than zero, and zstrict, where zboost must be greater than the absolute value of the lowest zboost in that learning task.

The high performance in the Nielsen *et al.* study, along with the lack of significant proteins using zboost (only two passing the zstrict threshold), led to the exclusion of this study from further analysis, as we believe that the signal is driven by the demographic differences between cases and controls in this particular study. No available metadata explained the variation in the other metagenomic studies.

Identity measurements

For each bacterial cluster, we compared the sequence identity between the bacterial proteins with experimentally verified interactions with human proteins and the bacterial proteins detected in human microbiomes from the same UniRef cluster. Original interactors and their homologs were aligned using Smith-Waterman local alignment with a BLOSUM62 matrix via

python's parasail library (v.1.1.17). The identity was calculated as the number of exact matches in the alignment, divided by the total number of alignment columns. Note that this denominator results in an under-estimation of the identity relative to UniRef's cluster identities.

Pathway enrichment analysis, disease and functional annotations

We performed pathway enrichment analysis using QIAGEN's Ingenuity Pathway Analysis (IPA)(QIAGEN) tool. All $zpos_{hum}$ proteins were uploaded as UniProt identifiers into the interface. Core Enrichment Analysis was conducted on all human tissue and cell lines from all data sources under IPA's stringent evidence filter. Pathways were considered enriched if they had both a -log(p-value) ≥ 1.3 and a Benjamini-Hochberg False Discovery Rate less or equal to 5%.

Disease annotations were extracted from all of gene-disease associations from DisGeNet (v.6.0). Lacking a simple hierarchy of disease, we binned similar disease terms into the 5 larger categories:

> CRC: Adenocarcinoma of large intestine, Hereditary non-polyposis colorectal cancer syndrome, Hereditary nonpolyposis colorectal carcinoma, Malignant neoplasm of colon stage IV, Malignant neoplasm of sigmoid colon, Malignant tumor of colon, Microsatellite instability-high colorectal cancer,
>
> Diabetes: Brittle diabetes, Familial central diabetes insipidus, Fibrocalculous pancreatic diabetes, Gastroparesis due to diabetes mellitus, Insulin resistance in diabetes, Insulin-dependent but ketosis-

resistant diabetes, Insulin-dependent diabetes mellitus secretory diarrhea syndrome, Insulin-resistant diabetes mellitus, Insulin-resistant diabetes mellitus at puberty, Latent autoimmune diabetes mellitus in adult, Macroalbuminuric diabetic nephropathy, Maturity onset diabetes mellitus in young, Maturity-onset diabetes of the young, type 10, Maturity-onset diabetes of the young, type 11, Microalbuminuric diabetic nephropathy, Moderate nonproliferative diabetic retinopathy, Monogenic diabetes, Neonatal diabetes mellitus, Neonatal insulin-dependent diabetes mellitus, Non-insulin-dependent diabetes mellitus with unspecified complications, Nonproliferative diabetic retinopathy, Other specified diabetes mellitus, Other specified diabetes mellitus with unspecified complications, Pancreatic disorders (not diabetes), Partial nephrogenic diabetes insipidus, Prediabetes syndrome, Proliferative diabetic retinopathy, Renal cysts and diabetes syndrome, Severe nonproliferative diabetic retinopathy, Transient neonatal diabetes mellitus, Type 2 diabetes mellitus in nonobese, Type 2 diabetes mellitus in obese, Type 2 diabetes mellitus with acanthosis nigricans, Visually threatening diabetic retinopathy, diabetes (mellitus) due to autoimmune process, diabetes (mellitus) due to immune mediated pancreatic islet beta-cell destruction, diabetes mellitus risk, idiopathic diabetes (mellitus), postprocedural diabetes mellitus, secondary diabetes mellitus NEC

Autoimmune: Addison's disease due to autoimmunity, Adult form of celiac disease, Aneurysm of celiac artery, Ankylosing spondylitis, Ankylosing spondylitis and other inflammatory spondylopathies, Arteriovenous fistulas of celiac and mesenteric vessels, Blood autoimmune disorders, Bullous systemic lupus erythematosus, Chilblain lupus 1, Dianzani autoimmune lymphoproliferative syndrome, Dilatation of celiac artery, Hyperthyroidism, Nonautoimmune, Latent autoimmune diabetes mellitus in adult, Maternal autoimmune disease, Multiple sclerosis in children, Neonatal Systemic lupus erythematosus, Subacute cutaneous lupus, Systemic lupus erythematosus encephalitis, Venous varicosities of celiac and mesenteric vessels, Warm autoimmune hemolytic anemia, diabetes (mellitus) due to autoimmune process, lupus cutaneous, lupus erythematodes

Obesity: Abdominal obesity metabolic syndrome, Adult-onset obesity, Aplasia/Hypoplasia of the earlobes, Childhood-onset truncal obesity, Constitutional obesity, Familial obesity, Generalized obesity, Gross

obesity, Hyperplastic obesity, Hypertrophic obesity, Hypoplastic olfactory lobes, Hypothalamic obesity, Moderate obesity, Overweight and obesity, Overweight or obesity, Prominent globes, Simple obesity, Type 2 diabetes mellitus in nonobese, Type 2 diabetes mellitus in obese

IBD: Acute and chronic colitis, Acute colitis, Allergic colitis, Amebic colitis, Chronic colitis, Chronic ulcerative colitis, Crohn Disease, Crohn's disease of large bowel, Crohn's disease of the ileum, Cytomegaloviral colitis, Distal colitis, Enterocolitis, Enterocolitis infectious, Eosinophilic colitis, Food-protein induced enterocolitis syndrome, Hemorrhagic colitis, IIeocolitis, Infectious colitis, Left sided colitis, Necrotizing Enterocolitis, Necrotizing enterocolitis in fetus OR newborn, Neonatal necrotizing enterocolitis, Non-specific colitis, Pancolitis, Pediatric Crohn's disease, Pediatric ulcerative colitis, Perianal Crohn's disease, Typhlocolitis, Ulcerative colitis in remission, Ulcerative colitis quiescent

We annotated bacterial protein clusters with their corresponding KEGG pathways by blasting all detected bacterial proteins of interest against the KEGG prokaryotes peptide file using blastp. Results had an identity ≥43.9% and e-values below 0.00067.

Human pathway annotation was performed using the mygene python library. Specifically, we queried pathway annotations from Wikipathways.

We submitted our bacterial sequences to EffectiveDB(EICHINGER *et al.* 2016) in order to obtain predictions for EffectiveT3 (type 3 secretion based on signal peptide), T4SEpre (type 4 secretion based on composition in C-terminus), EffectiveCCBD (type 3 secretion based on chaperone binding sites), and EffectiveELD (predicts secretion based on eukaryotic-like

domains). We used the single default cutoffs for T4SEpre, EffectiveCCBD, and EffectiveELD, and chose the 'sensitive' cutoff (0.95) rather than the 'selective' (0.9999) cutoff for EffectiveT3. Transmembrane proteins or signal peptides were predicted using TMHMM(KROGH *et al.* 2001) (v.2.0c), with a threshold of 19 or more expected number of amino acids in transmembrane helices.

Drug target information was extracted from probes-and-drugs (04.2019 database dump). Bacterial taxonomy information was extracted from NCBI. UniProt identifiers and annotations were downloaded using UniProt SPARQL endpoint.

Statistics

For Figure 4.2b, p-values for the difference in the proportion of DisGeNet proteins and disease-associated proteins was calculated through a chi squared test (dof=1): The total number of DisGeNet and disease-associated proteins is 17,549 and 767 respectively. For the labels CRC, Diabetes, Autoimmunity, Obesity, and IBD we find {2128, 355, 420, 195, and 1,029} DisGeNet genes against {133, 26, 34, 12, and 65} disease-associated genes respectively. This corresponds to a chi squared statistic of 2.2e-5 (p=0.000022) for CRC, 1.4e-2 (p=0.013619) for Diabetes, 5.9e-4

(p=0.000587) for Autoimmunity, 0.3 (p=323067) for Obesity, and 3.6e-3 (p=0.003627) for IBD.

SUPPORTING INFORMATION

Additional supporting information may be found in the appendix.

Supplementary Figure S4.1. An outline of our homology mapping procedure and alignment. Depiction of the interaction network inference and protein detection pipeline. Note that only bacterial proteins found to be human-interactors through the mapping procedure are used as candidates for detection in metagenomic studies.

Supplementary Figure S4.2. Pairwise identity between proteins found in the human microbiome and those with experimentally verified interaction. Histogram showing the percent identity between all bacterial proteins with experimental verification and their corresponding detected proteins in human microbiomes. This histogram is annotated with a gaussian kernel density estimate of the distribution.

Supplementary Figure S4.3. Taxonomic diversity in bacterial clusters detected in patients. Histogram showing the number of species, genera,

families, orders, classes and phyla for bacterial clusters with members detected in human microbiomes.

Supplementary Figure S4.4. Human protein interactors according to their zboost scores and log odds ratio. Volcano plots of the human protein interactors present in each study according to their zboost scores and log odds ratios in each case-control cohort study.

Supplementary Figure S4.5. Clustering of cases and controls is not due to disease status, study or metadata, except for ethnicity in Nielsen *et al.*
(A) Principal components analysis of detected human protein interactors for samples, according to study.
(B) Principal components analysis of detected human protein interactors for all samples in nine metagenomic studies colored by disease status according to study. Controls are all colored together in blue.
(C) Principal components analysis of detected human protein interactors in each study, separated by controls (blue) and cases (orange).

Supplementary Table S4.1. Metagenomic studies used in this research. For each study, we list its focus, the labels in the cohort study, the patient count for each of the labels, how we grouped cases and controls, the number

of detected bacterial clusters and inferred human interactors, and the number of important bacterial and human proteins, passing each of our thresholds: zpos (zboost greater than zero) and zstrict (zboost greater than the magnitude of the minimum zboost).

Supplementary Table S4.2. Important human interactors that have known gene-disease associations. Listed are the important $zstrict_{hum}$ proteins with gene-disease associations in DisGeNet, along with the study in which they are found to be important.

Supplementary Table S4.3. Important human interactors that are known drug targets. For each human protein in the $zsig_{hum}$ subset, we list the drug interactor and the study in which it was found to be important.

REFERENCES

Altmann A., L. Tolosi, O. Sander, and T. Lengauer, 2010 Permutation importance: a corrected feature importance measure. Bioinformatics 26: 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Bhavsar A. P., J. A. Guttman, and B. B. Finlay, 2007 Manipulation of host-cell pathways by bacterial pathogens. Nature 449: 827–834. https://doi.org/10.1038/nature06247

Choi S.-S., E.-S. Kim, J.-E. Jung, D. P. Marciano, A. Jo, *et al.*, 2016 PPARγ Antagonist Gleevec Improves Insulin Sensitivity and Promotes the Browning of White Adipose Tissue. Diabetes 65: 829–839. https://doi.org/10.2337/db15-1382

Daulagala A. C., M. C. Bridges, and A. Kourtidis, 2019 E-cadherin Beyond Structure: A Signaling Hub in Colon Homeostasis and Disease. Int J Mol Sci 20. https://doi.org/10.3390/ijms20112756

Donia M. S., and M. A. Fischbach, 2015 HUMAN MICROBIOTA. Small molecules from the human microbiota. Science 349: 1254766. https://doi.org/10.1126/science.1254766

Dyer M. D., C. Neff, M. Dufford, C. G. Rivera, D. Shattuck, *et al.*, 2010 The human-bacterial pathogen protein interaction networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis. PLoS ONE 5: e12089. https://doi.org/10.1371/journal.pone.0012089

Eichinger V., T. Nussbaumer, A. Platzer, M.-A. Jehl, R. Arnold, *et al.*, 2016 EffectiveDB--updates and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI secretion systems. Nucleic Acids Res. 44: D669-674. https://doi.org/10.1093/nar/gkv1269

Feng Q., S. Liang, H. Jia, A. Stadlmayr, L. Tang, *et al.*, 2015 Gut microbiome development along the colorectal adenoma-carcinoma sequence. Nat Commun 6: 6528. https://doi.org/10.1038/ncomms7528

Franke A., T. Balschun, T. H. Karlsen, J. Sventoraityte, S. Nikolaus, *et al.*, 2008 Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. Nat. Genet. 40: 1319–1323. https://doi.org/10.1038/ng.221

Gianchecchi E., and A. Fierabracci, 2019 Recent Advances on Microbiota Involvement in the Pathogenesis of Autoimmunity. Int J Mol Sci 20. https://doi.org/10.3390/ijms20020283

Guven-Maiorov E., C.-J. Tsai, B. Ma, and R. Nussinov, 2017a Prediction of Host-Pathogen Interactions for Helicobacter pylori by Interface Mimicry and Implications to Gastric Cancer. J. Mol. Biol. 429: 3925–3941. https://doi.org/10.1016/j.jmb.2017.10.023

Guven-Maiorov E., C.-J. Tsai, and R. Nussinov, 2017b Structural host-microbiota interaction networks. PLoS Comput. Biol. 13: e1005579. https://doi.org/10.1371/journal.pcbi.1005579

Guven-Maiorov E., C.-J. Tsai, B. Ma, and R. Nussinov, 2019 Interface-Based Structural Prediction of Novel Host-Pathogen Interactions. Methods Mol. Biol. 1851: 317–335. https://doi.org/10.1007/978-1-4939-8736-8_18

Hagemann L., A. Gründel, E. Jacobs, and R. Dumke, 2017 The surface-displayed chaperones GroEL and DnaK of Mycoplasma pneumoniae interact with human plasminogen and components of the extracellular matrix. Pathog Dis 75. https://doi.org/10.1093/femspd/ftx017

Hamiaux C., A. van Eerde, C. Parsot, J. Broos, and B. W. Dijkstra, 2006 Structural mimicry for vinculin activation by IpaA, a virulence factor of Shigella flexneri. EMBO Rep. 7: 794–799. https://doi.org/10.1038/sj.embor.7400753

Hannigan G. D., M. B. Duhaime, M. T. Ruffin, C. C. Koumpouras, and P. D. Schloss, 2018 Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome. MBio 9. https://doi.org/10.1128/mBio.02248-18

Hebbandi Nanjundappa R., F. Ronchi, J. Wang, X. Clemente-Casares, J. Yamanouchi, *et al.*, 2017 A Gut Microbial Mimic that Hijacks Diabetogenic Autoreactivity to Suppress Colitis. Cell 171: 655-667.e17. https://doi.org/10.1016/j.cell.2017.09.022

Henderson B., and A. Martin, 2013 Bacterial moonlighting proteins and bacterial virulence. Curr. Top. Microbiol. Immunol. 358: 155–213. https://doi.org/10.1007/82_2011_188

Henderson B., 2014 An overview of protein moonlighting in bacterial infection. Biochem. Soc. Trans. 42: 1720–1727. https://doi.org/10.1042/BST20140236

Jess T., B. W. Jensen, M. Andersson, M. Villumsen, and K. H. Allin, 2019 Inflammatory Bowel Disease Increases Risk of Type 2 Diabetes in a Nationwide Cohort Study. Clin. Gastroenterol. Hepatol. https://doi.org/10.1016/j.cgh.2019.07.052

Jurjus A., A. Eid, S. Al Kattar, M. N. Zeenny, A. Gerges-Geagea, *et al.*, 2016 Inflammatory bowel disease, colorectal cancer and type 2 diabetes mellitus: The links. BBA Clin 5: 16–24. https://doi.org/10.1016/j.bbacli.2015.11.002

Kanehisa M., M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, 2017 KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45: D353–D361. https://doi.org/10.1093/nar/gkw1092

Kang E. A., K. Han, J. Chun, H. Soh, S. Park, *et al.*, 2019 Increased Risk of Diabetes in Inflammatory Bowel Disease Patients: A Nationwide Population-based Study in Korea. J Clin Med 8. https://doi.org/10.3390/jcm8030343

Karlsson F. H., V. Tremaroli, I. Nookaew, G. Bergström, C. J. Behre, *et al.*, 2013 Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature 498: 99–103. https://doi.org/10.1038/nature12198

Kort S. de, A. A. M. Masclee, S. Sanduleanu, M. P. Weijenberg, M. P. P. van Herk-Sukel, *et al.*, 2017 Higher risk of colorectal cancer in patients with newly diagnosed diabetes mellitus before the age of colorectal cancer screening initiation. Sci Rep 7: 46527. https://doi.org/10.1038/srep46527

Krogh A., B. Larsson, G. von Heijne, and E. L. Sonnhammer, 2001 Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305: 567–580. https://doi.org/10.1006/jmbi.2000.4315

Kursa M. B., and W. R. Rudnicki, 2010 Feature Selection with the Boruta Package. Journal of Statistical Software 36: 1–13. https://doi.org/10.18637/jss.v036.i11

Le Chatelier E., T. Nielsen, J. Qin, E. Prifti, F. Hildebrand, *et al.*, 2013 Richness of human gut microbiome correlates with metabolic markers. Nature 500: 541–546. https://doi.org/10.1038/nature12506

Lehner T., L. A. Bergmeier, Y. Wang, L. Tao, M. Sing, *et al.*, 2000 Heat shock proteins generate beta-chemokines which function as innate adjuvants enhancing adaptive immunity. Eur. J. Immunol. 30: 594–603. https://doi.org/10.1002/1521-4141(200002)30:2<594::AID-IMMU594>3.0.CO;2-1

LeValley S. L., C. Tomaro-Duchesneau, and R. A. Britton, 2019 Degradation of the incretin hormone Glucagon-Like Peptide-1 (GLP-1) by Enterococcus faecalis metalloprotease GelE. bioRxiv 732495. https://doi.org/10.1101/732495

Lloyd-Price J., A. Mahurkar, G. Rahnavard, J. Crabtree, J. Orvis, *et al.*, 2017 Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 550: 61–66. https://doi.org/10.1038/nature23889

McKinney W., 2010 Data Structures for Statistical Computing in Python. 6.

Nešić D., L. Buti, X. Lu, and C. E. Stebbins, 2014 Structure of the Helicobacter pylori CagA oncoprotein bound to the human tumor suppressor ASPP2. Proc. Natl. Acad. Sci. U.S.A. 111: 1562–1567. https://doi.org/10.1073/pnas.1320631111

Nielsen H. B., M. Almeida, A. S. Juncker, S. Rasmussen, J. Li, *et al.*, 2014 Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol. 32: 822–828. https://doi.org/10.1038/nbt.2939

Orchard S., M. Ammari, B. Aranda, L. Breuza, L. Briganti, *et al.*, 2014 The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 42: D358-363. https://doi.org/10.1093/nar/gkt1115

Pedregosa F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, *et al.*, Scikit-learn: Machine Learning in Python. MACHINE LEARNING IN PYTHON 6.

Plovier H., A. Everard, C. Druart, C. Depommier, M. Van Hul, *et al.*, 2017 A purified membrane protein from *Akkermansia muciniphila* or the pasteurized bacterium improves metabolism in obese and diabetic

mice. Nature Medicine 23: 107–113. https://doi.org/10.1038/nm.4236

Qiagen, Ingenuity Pathway Analysis. QIAGEN Bioinformatics.

Qin J., Y. Li, Z. Cai, S. Li, J. Zhu, *et al.*, 2012 A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490: 55–60. https://doi.org/10.1038/nature11450

Richard J., and I. Lingvay, 2011 Hepatic steatosis and Type 2 diabetes: current and future treatment considerations. Expert Rev Cardiovasc Ther 9: 321–328. https://doi.org/10.1586/erc.11.15

Schirmer M., E. A. Franzosa, J. Lloyd-Price, L. J. McIver, R. Schwager, *et al.*, 2018 Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. Nat Microbiol 3: 337–346. https://doi.org/10.1038/s41564-017-0089-z

Seidler K. A., and N. W. Seidler, 2013 Role of extracellular GAPDH in Streptococcus pyogenes virulence. Mo Med 110: 236–240.

Sen R., L. Nayak, and R. K. De, 2016 A review on host-pathogen interactions: classification and prediction. Eur. J. Clin. Microbiol. Infect. Dis. 35: 1581–1599. https://doi.org/10.1007/s10096-016-2716-7

Shah P. S., N. Link, G. M. Jang, P. P. Sharp, T. Zhu, *et al.*, 2018 Comparative Flavivirus-Host Protein Interaction Mapping Reveals Mechanisms of Dengue and Zika Virus Pathogenesis. Cell 175: 1931-1945.e18. https://doi.org/10.1016/j.cell.2018.11.028

Sia C., and A. Hänninen, 2006 Apoptosis in autoimmune diabetes: the fate of beta-cells in the cleft between life and death. Rev Diabet Stud 3: 39–46. https://doi.org/10.1900/RDS.2006.3.39

Skuta C., M. Popr, T. Muller, J. Jindrich, M. Kahle, *et al.*, 2017 Probes &Drugs portal: an interactive, open data resource for chemical biology. Nat. Methods 14: 759–760. https://doi.org/10.1038/nmeth.4365

Slenter D. N., M. Kutmon, K. Hanspers, A. Riutta, J. Windsor, *et al.*, 2018 WikiPathways: a multifaceted pathway database bridging

metabolomics to other omics research. Nucleic Acids Res. 46: D661–D667. https://doi.org/10.1093/nar/gkx1064

Solt L. A., Y. Wang, S. Banerjee, T. Hughes, D. J. Kojetin, *et al.*, 2012 Regulation of circadian behaviour and metabolism by synthetic REV-ERB agonists. Nature 485: 62–68. https://doi.org/10.1038/nature11030

Stewart L., J. D M Edgar, G. Blakely, and S. Patrick, 2018 Antigenic mimicry of ubiquitin by the gut bacterium Bacteroides fragilis: a potential link with autoimmune disease. Clin. Exp. Immunol. 194: 153–165. https://doi.org/10.1111/cei.13195

Stidham R. W., and P. D. R. Higgins, 2018 Colorectal Cancer in Inflammatory Bowel Disease. Clin Colon Rectal Surg 31: 168–178. https://doi.org/10.1055/s-0037-1602237

Suzek B. E., Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, *et al.*, 2015 UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31: 926–932. https://doi.org/10.1093/bioinformatics/btu739

Vieira E., L. Marroquí, A. L. C. Figueroa, B. Merino, R. Fernandez-Ruiz, *et al.*, 2013 Involvement of the clock gene Rev-erb alpha in the regulation of glucagon secretion in pancreatic alpha-cells. PLoS ONE 8: e69939. https://doi.org/10.1371/journal.pone.0069939

Weitz J., M. Koch, J. Debus, T. Höhler, P. R. Galle, *et al.*, 2005 Colorectal cancer. Lancet 365: 153–165. https://doi.org/10.1016/S0140-6736(05)17706-X

Wolf A. M., D. Wolf, H. Rumpold, S. Ludwiczek, B. Enrich, *et al.*, 2005 The kinase inhibitor imatinib mesylate inhibits TNF-{alpha} production in vitro and prevents TNF-dependent acute hepatic inflammation. Proc. Natl. Acad. Sci. U.S.A. 102: 13622–13627. https://doi.org/10.1073/pnas.0501758102

Yu J., Q. Feng, S. H. Wong, D. Zhang, Q. Y. Liang, *et al.*, 2017 Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut 66: 70–78. https://doi.org/10.1136/gutjnl-2015-309800

Zeller G., J. Tap, A. Y. Voigt, S. Sunagawa, J. R. Kultima, *et al.*, 2014 Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol. Syst. Biol. 10: 766. https://doi.org/10.15252/msb.20145645

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

In this thesis, I demonstrated three different methods to shift results from association studies towards hypotheses to test molecular mechanisms. In **Chapter 2** I designed an online hub for the customization of quality-control and annotation pipelines to find high-quality candidate variants for further study in genetic association studies. In **Chapter 3** I assemble structural protein interaction networks through experimental, homology, and inferred interaction interfaces to identify disease mutations that might act through the disruption of specific protein-protein interactions. Finally, in **Chapter 4** I generate a homology-based bacteria-human protein-protein interaction network to detect the disease-associated targeting of human proteins by the commensal microbiome.

Prioritizing genetic variants for further study

The quality-control and annotation process in variant prioritization pipelines can be extremely complex, and requires the coordination of several bioinformatic tools. The order in which they are applied, the thresholds used for filtering, and the versions of the software used at each step in the analysis

are all variables that can result in different downstream sets of variants. Replicating a prioritization pipeline from published research given all raw data is then difficult. Even when proper documentation is provided, with the appropriate version numbers and a clearly laid-out computational protocol, a researcher still requires the computational resources and know-how to be able to implement this type of pipeline. An orthogonal problem involves researchers looking to prioritize variants with little previous experience in the field. Understanding which tools are available, how they interact with each other, and what further options for annotations they might have is an often-time-consuming process that also contributes to the heterogeneity of computational variant prioritization protocols.

In **Chapter 2** I presented GeMSTONE, a central hub for variant prioritization motivated by the need to systematically generate these pipelines in a reproducible, intuitive way, without the need to have access to computational expertise or resources. Outsourcing the computational burden and documentation task of these pipelines to an external server allows researchers to prioritize variants in a more systematic, reproducible, and customizable way.

GeMSTONE provides quality-control functionality to avoid prioritization of sequencing artifacts, and annotations that allow researchers to select candidate variants that are more likely to be associated with their

disease of interest. However, it is important to remember that using these annotations for variant prioritization is just a heuristic to find likely consequential variants in the genome, and is in no way guaranteed to explain the underlying mechanisms driving disease. Variant prioritization pipelines are currently best suited for monogenic trait identification and work is still needed to capture polygenic effects systematically. Despite the wide variety of resources aggregated in this work, most of the annotation databases and tools leveraged by GeMSTONE are only relevant to protein-coding sequences. Further annotation with regulatory element and non-coding RNA databases could greatly enhance the ability of these customized pipelines to find consequential variants. Additionally, there is still work to be done in order to implicate specific disease pathways in the output of these pipelines. The current version of GeMSTONE can filter or annotate using from gene ontology or disease databases, but only provides gene set enrichment analysis from KEGG, BioCarta, and Reactome pathway databases. Expanding the functionality of GeMSTONE to perform enrichment analogies on both gene ontology and disease databases would increase its usability. Additionally, adopting protein interaction network clustering methods can help identify subnetworks that are enriched for variant-containing genes and greatly improve the functional interpretability of prioritized variants. Finally, this type of resource is only useful as long as it is

constantly updated with the latest available data. Keeping up with the publication of new allele frequency databases, variant function scoring tools, and gene annotation resources is a challenging problem, and it is unlikely to have an automated solution.

Contextualizing genetic variation in structural protein-protein interaction networks

Understanding the pathway membership and binding partners of a protein can help us understand its role in the cell and the potential consequences that variation in this protein's sequence might have. Specifically, we can use structural protein-protein interaction networks to characterize the specific surfaces that are responsible for the binding of different proteins with the same partner. These binding surfaces can be altered by variation, and consequently only disrupt a protein's ability to bind to some of its partners, affecting only some of its associated pathways and functions. This high-resolution understanding of the potential effects of variation on the protein-protein interaction network of an organism is extremely useful to understand the intricacies of molecular mechanisms, but most protein-protein interactions lack this partner-specific binding surface information.

In **Chapter 3**, I brought together existing experimentally-determined and homology-modeled protein interaction interfaces, as well as newly-generated computational interface predictions in an online resource called Interactome INSIDER. By annotating all available protein-protein interactions with their known or likely interfaces, we are able to implicate disease mutations in the disruption of specific protein-protein interactions.

This network, however, is far from complete. Although computational prediction and homology modeling are powerful tools, they do not offer the high-resolution, high-confidence interface information that can be learned from co-crystallization of protein partners. Further, both of the computational methods rely on already-existing data in order to estimate where interfaces might lie. Binding pairs involving membrane-protein partners, for example, are challenging to crystallize, and are thus less represented in co-crystal databases, less likely to be homology modelled, and their computational predictions are less likely to be correct (or correctly evaluated). Although INSIDER is currently a static resource, there is a lot to be gained by updating the experimental interfaces, homology models, and computational predictions periodically when new experimental information is released. Additionally, it is important to note that the protein-protein interaction network itself is not complete. Expanding the INSIDER interface in order to be able to predict interfaces for arbitrary protein-protein

pairs or increasing the size of the network periodically as mentioned above, would greatly increase the long-term impact of this tool. Finally, the performance of the underlying machine learning algorithm could be improved by using new models for interface prediction. Graph Convolutional Neural Networks have shown recent success at predicting protein interaction interfaces, and might serve as a good alternative to the models in ECLAIR that require protein structure. It is also still not clear that windowed approaches are the best fit for protein-protein interaction prediction – there might be higher-quality predictions available through the use of models in the Recurrent Neural Network family, where sequence information does not have to be preprocessed into windows, but is taken as a continuous input. Beyond these improvements to Interactome INSIDER as it currently exists – there is protein-relevant biological context still missing from this picture. Additional annotation with post-translational modification sites, DNA and RNA binding surfaces, and other functional sites that are well-characterized in proteins can provide a clearer picture of the functional consequence of protein-coding variants.

Understanding the role of exogenous proteins on human health

The human gut hosts complex communities of bacteria which interact with their hosts to influence human health. Understanding the associations

between changes in the human microbiome and human disease outcomes is a very challenging problem: closely related bacterial taxa might have very different influences on human health, while horizontal gene transfer allows for unrelated bacterial species to share the same genetic elements. Overall, human-associated bacteria are functionally heterogenous, and their genomes and proteins are not well annotated with molecular function or pathway information.

One approach to characterize the interaction between bacteria and their human hosts is to look at exogenous protein-protein interactions, that is, interactions between bacterial proteins and human proteins. This area of study has found success in pathogens, with large-scale interaction studies where *Bacillus anthracis*, *Francisella Tularensis*, and *Yersinia pestis* genomes have been screened against a large variety of human proteins to generate organism specific bacteria-human protein-protein interactions. In one rare example, computational methods were applied to a commensal community, rather than individual pathogenic organisms, specifically in the oral microbiome (COELHO *et al.* 2014). Despite the potential in this work, no large-scale characterization of commensal human microbiomes has been performed, computationally or experimentally.

In **Chapter 4** I presented a protein-protein interaction network between bacterial and human proteins, inferred from homology, that I use to

characterize potential differential targeting of human proteins by microbe-derived proteins from the human gut microbiome. This disease association between human disease and the bacterial targeting of specific human proteins opens the door to a host-centric understanding of the molecular pathways involved in the human microbiome's role in diseases like colorectal cancer, type 2 diabetes, inflammatory bowel disease, and obesity. This work also holds a potential boon for the functional characterization of bacterial proteins whose function had thus far remained unknown.

It is important to remember, however, that his network of homologous interactions is largely informed by experimentally-resolved protein interactions in pathogens. In **Chapter 4** I demonstrate that there is little phylogenetic bias in the bacterial proteins detected in patients, however, some health-relevant interactions might involve proteins from commensal bacteria that lack a homolog in the pathogenic organisms that have been studied. Further experimental assays on commensal-human protein-protein interactions would greatly expand the scope of this work and provide a fuller picture of protein-driven host-bacteria protein interactions. Additionally, it is important to note that high-throughput interaction assays can have biases in the types of interactions that they can capture. As was the case in **Chapter 3**, membrane proteins are likely under-represented in our homology-modeled or experimentally-resolved bacteria-human protein interaction networks.

This is particularly worrying as we know bacterial secreted proteins and membrane-bound human receptors (e.g. TLR2, TL5, or G-protein coupled receptors) interact to control human signaling pathways, and finding more examples of these types of interactions could be very impactful. Finally, protein interactions, though they are an exciting new avenue to understand human-bacteria interactions, do not capture the full picture. Paired metagenomic and host-genome samples can help us better understand the relationship between variants in human proteins targeted by bacteria and the types of taxa that can inhabit that human gut. Additionally, pairing these metagenomic experiments with host-RNAseq experiments can help us understand the changes in expression that might be induced by these bacteria-human interactions, particularly when the human protein targeted is a transcription factor. Finally, understanding the contribution of bacteria to human disease will likely require the aggregation of several effects arising from the cross-species interactome, the gut metabolome, and biofilm composition and organization.

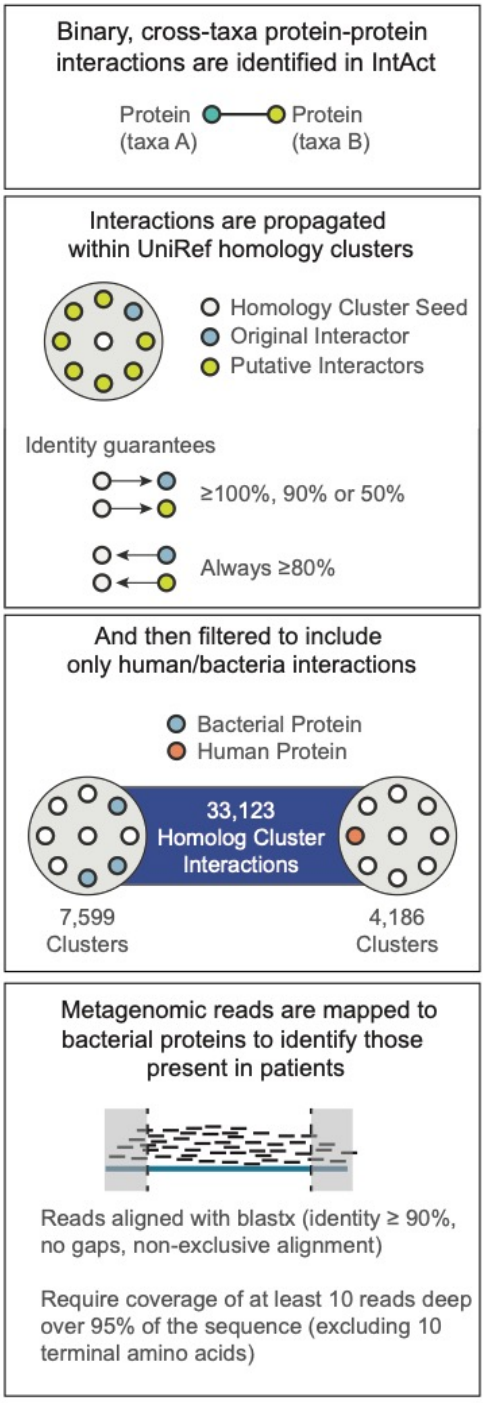
Together, these resources can be integrated to better understand human health in terms of both endogenous and exogenous protein-protein interactions. When seeking to understand the molecular mechanisms that explain the association between a human variant and disease, we must systematically query the functional pathways, protein interactors, interaction

surfaces, and bacterial-protein partners of that given human gene's protein product.

REFERENCES

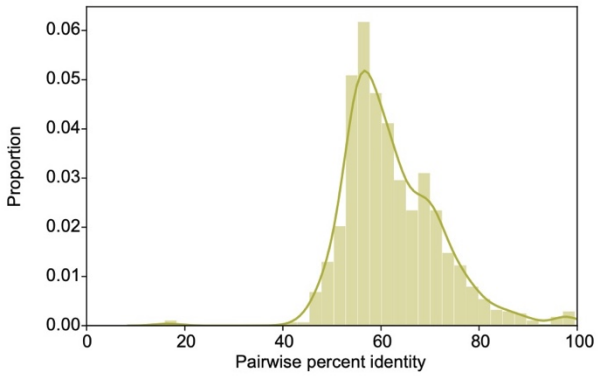
Coelho E. D., J. P. Arrais, S. Matos, C. Pereira, N. Rosa, *et al.*, 2014 Computational prediction of the human-microbial oral interactome. BMC Syst Biol 8: 24. https://doi.org/10.1186/1752-0509-8-24

APPENDIX

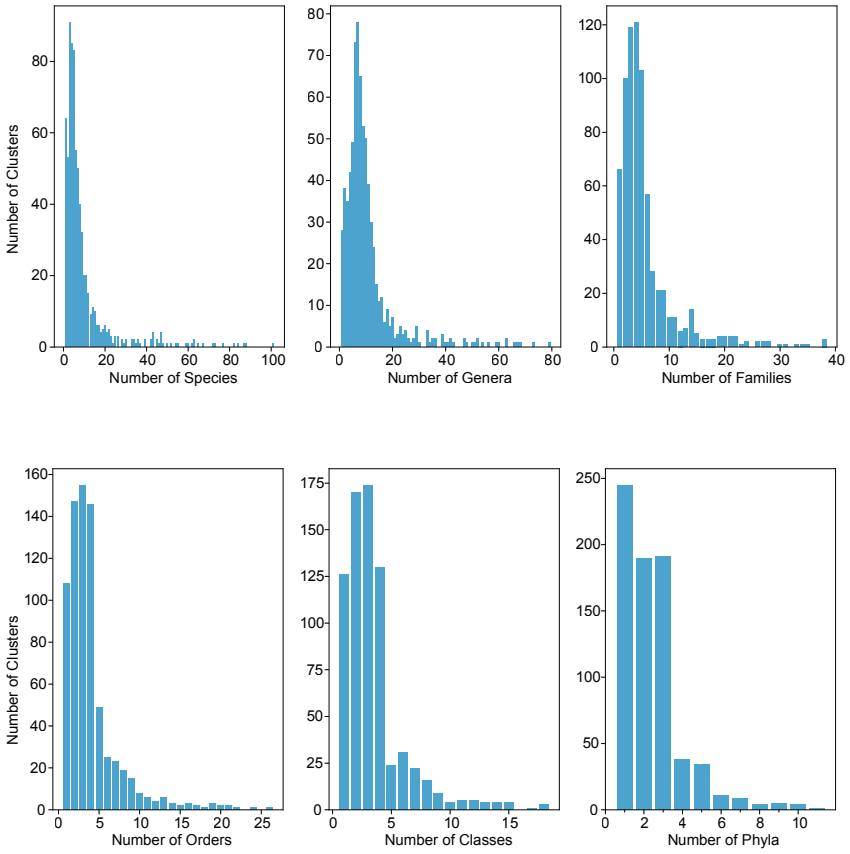


Supplementary Figure S4.1. An outline of our homology mapping procedure and alignment. Depiction of the interaction network inference and protein detection pipeline.

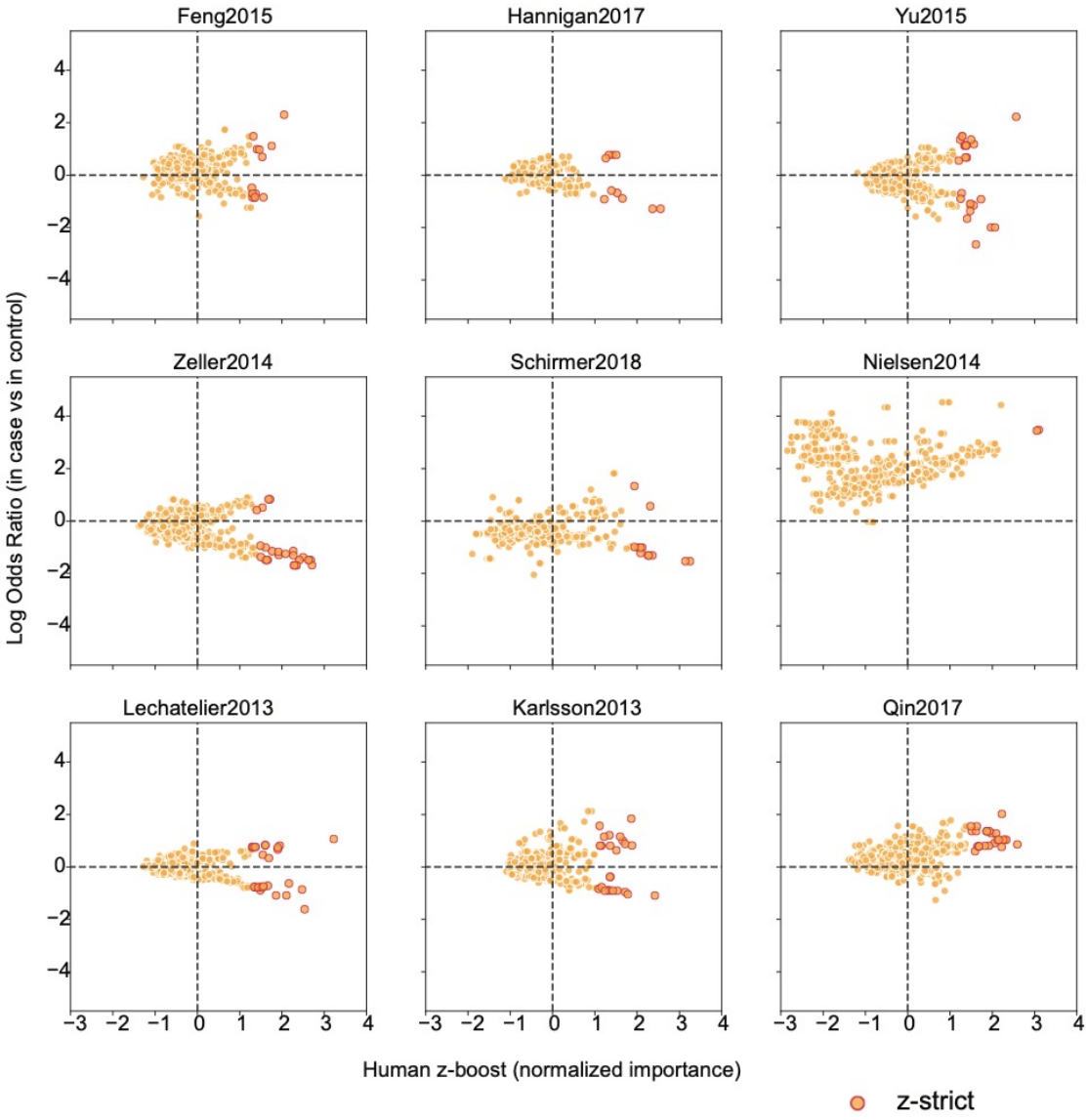
Note that only bacterial proteins found to be human-interactors through the mapping procedure are used as candidates for detection in metagenomic studies.



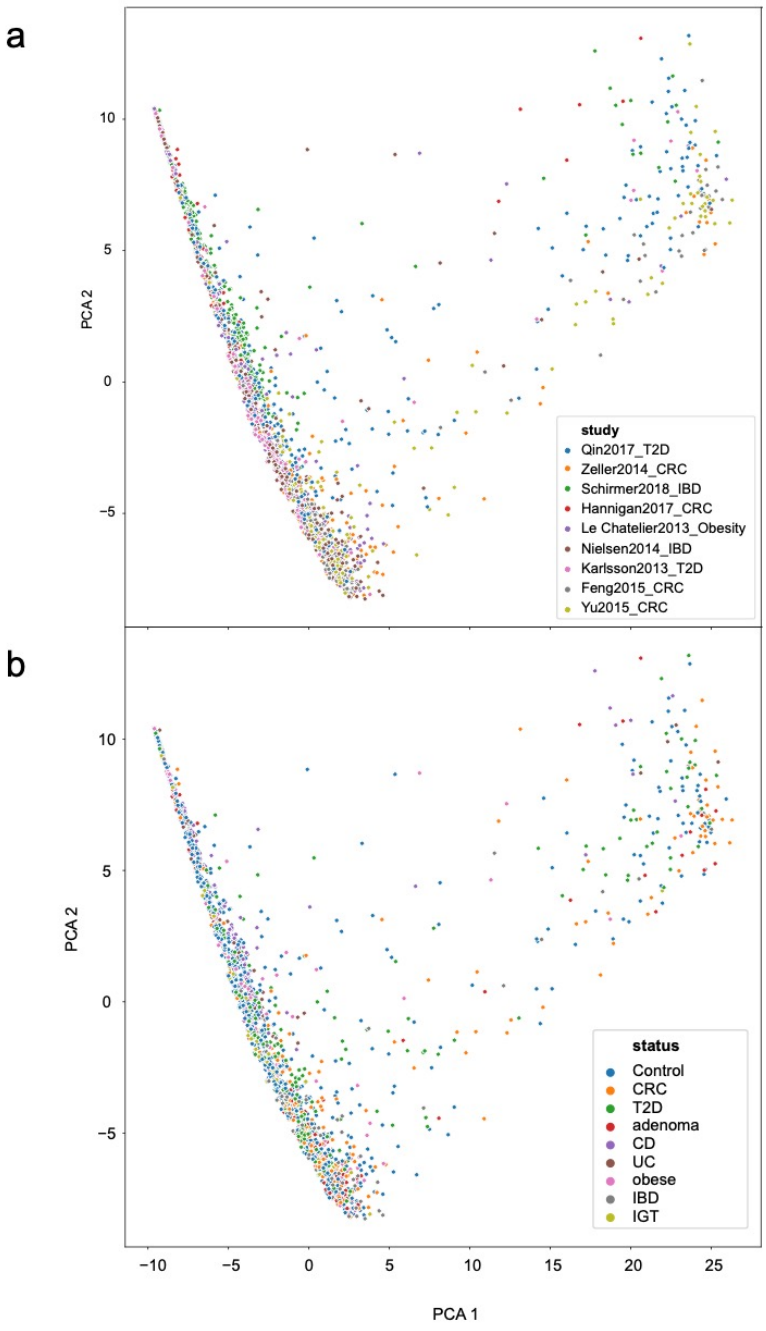
Supplementary Figure S4.2. Pairwise identity between proteins found in the human microbiome and those with experimentally verified interaction. Histogram showing the percent identity between all bacterial proteins with experimental verification and their corresponding detected proteins in human microbiomes. This histogram is annotated with a gaussian kernel density estimate of the distribution.



Supplementary Figure S4.3. Taxonomic diversity in bacterial clusters detected in patients. Histogram showing the number of species, genera, families, orders, classes and phyla for bacterial clusters with members detected in human microbiomes.



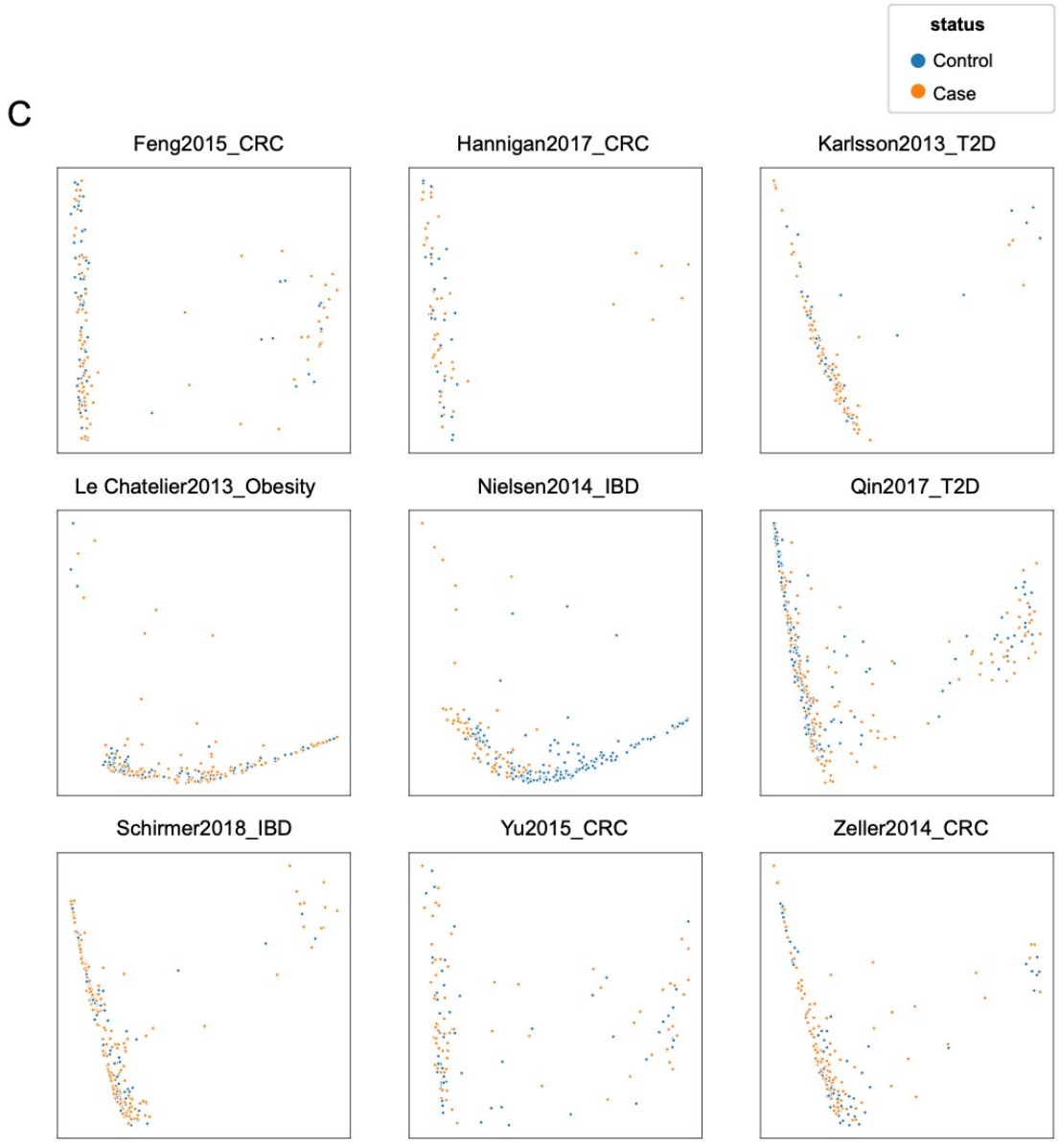
Supplementary Figure S4.4. Human protein interactors according to their zboost scores and log odds ratio. Volcano plots of the human protein interactors present in each study according to their zboost scores and log odds ratios in each case-control cohort study.



Supplementary Figure S4.5. Clustering of cases and controls is not due to disease status, study or metadata, except for ethnicity in Nielsen *et al.*
(A) Principal components analysis of detected human protein interactors for samples, according to study.

(B) Principal components analysis of detected human protein interactors for all samples in nine metagenomic studies colored by disease status according to study. Controls are all colored together in blue.
(C) Principal components analysis of detected human protein interactors in each study, separated by controls (blue) and cases (orange).

Figure S4.5 (continued)



Supplementary Table S4.1. Metagenomic studies used in this research.
For each study, we list its focus, the labels in the cohort study, the patient count for each of the labels, how we grouped cases and controls, the number of detected bacterial clusters and inferred human interactors, and the number of important bacterial and human proteins, passing each of our thresholds: zpos (zboost greater than zero) and zstrict (zboost greater than the magnitude of the minimum zboost).

Index	**Author**	**Focus**	**Labels**	**Patient Counts**
1	Feng et al.	Colorectal Cancer	CRC/adenoma/Control	46/47/61
2	Hannigan et al.		CRC/adenoma/Control	26/24/27
3	Yu et al.		CRC/Control	75/53
4	Zeller et al.		CRC/adenoma/Control	91/42/66
5	Karlsson et al.	Type 2 Diabetes	T2D/IGT/Control	53/49/43
6	Qin et al.		T2D/Control	171/174
7	Le Chatelier et al.	Obesity	Obese/Control	156/101
8	Nielsen et al.	Inflammatory Bowel Disease	IBD/Control	81/190
9	Schirmer et al.		Crohn's Disease/Ulcerative Colitis/ Control	142/77/54

Index	**Case Labels**	**Control Labels**	**Bacterial Clusters Detected (all/zpos/zstrict)**	**Human Interactors Inferred (all/zpos/zstrict)**
1	CRC, adenoma	Control	542/188/12	1171/424/15
2	CRC, adenoma	Control	56/16/5	272/102/11
3	CRC	Control	563/184/19	1271/375/30
4	CRC, adenoma	Control	342/85/18	895/210/34
5	T2D, IGT	Control	132/53/10	451/193/26
6	T2D	Control	685/156/17	1529/426/36
7	Obese	Control	190/65/18	733/266/33
8	IBD	Control	181/43/2	674/141/2
9	CD, UC	Control	96/35/12	420/173/12

Supplementary Table S4.2. Important human interactors that have known gene-disease associations. Listed are the important $zstrict_{hum}$ proteins with gene-disease associations in DisGeNet, along with the study in which they are found to be important.

substudy	humgene	DisGeNet_curated_disease
Feng2015_CRC_CRC|adenoma	ABLIM3	Liver Cirrhosis, Experimental
	BRWD3	Intellectual Disability; MENTAL RETARDATION, X-LINKED 93 (disorder)
	CSF1R	Acute Myeloid Leukemia (AML-M2); Acute Myeloid Leukemia, M1; Benign Neoplasm; Breast Carcinoma; Childhood Ataxia with Central Nervous System Hypomyelinization; Hematologic Neoplasms; Hereditary Diffuse Leukoencephalopathy with Spheroids; Intellectual Disability; Leukemia, Myelocytic, Acute; Leukodystrophy; Leukoencephalopathies; Liver Cirrhosis, Experimental; MYELODYSPLASTIC SYNDROME; Malignant Neoplasms; Malignant neoplasm of breast; Mammary Neoplasms; Mammary Neoplasms, Human; Neoplasm Metastasis; Neoplasms; Parkinson Disease; Squamous cell carcinoma of lung
	CXCR5	Biliary Cirrhosis, Secondary; Biliary cirrhosis; Malignant neoplasm of breast; Primary biliary cirrhosis
	DDX3X	Corpus Callosum, Agenesis of, with Facial Anomalies and Robin Sequence; Degenerative polyarthritis; Epileptic encephalopathy; Intellectual Disability; MENTAL RETARDATION, X-LINKED 102; Malignant mesothelioma; Malignant neoplasm of breast; Neurodevelopmental Disorders; Osteoarthrosis Deformans; Precursor T-Cell Lymphoblastic Leukemia-Lymphoma; T-Cell Lymphoma
	DDX3Y	Male sterility due to Y-chromosome deletions; Partial chromosome Y deletion;

		Spermatogenic Failure, Nonobstructive, Y-Linked
	MBNL1	MYOTONIC DYSTROPHY 1; Myotonia; Myotonic Phenomenon; Percussion Myotonia; Schizophrenia
	PHIP	Colorectal Cancer; Intellectual Disability; Mental Retardation, Psychosocial; Mental deficiency; Profound Mental Retardation; Sicca Syndrome; Sjogren's Syndrome
	PITPNM3	Cone-Rod Dystrophy 5; Retinitis Pigmentosa
	POGZ	Intellectual Disability; Microcephaly; Neurodevelopmental Disorders; WHITE-SUTTON SYNDROME
	TLE5	Myocardial Ischemia
	TNXB	Cakut; Congenital aneurysm of ascending aorta; Ehlers-Danlos Syndrome; Ehlers-Danlos syndrome caused by tenascin-X deficiency; Ehlers-Danlos syndrome, type 3 (disorder); Osteogenesis Imperfecta; Schizophrenia; Squamous cell carcinoma of esophagus; VESICOURETERAL REFLUX 8; Vesico-Ureteral Reflux
Hannigan2017_CRC_CRC|adenoma	ANXA2	Acute Myeloid Leukemia (AML-M2); Acute Myeloid Leukemia, M1; Animal Mammary Neoplasms; Chemical and Drug Induced Liver Injury; Chemically-Induced Liver Toxicity; Drug-Induced Acute Liver Injury; Drug-Induced Liver Disease; Hepatitis, Drug-Induced; Hepatitis, Toxic; Leukemia, Myelocytic, Acute; Liver Cirrhosis, Experimental; Liver neoplasms; Lung Neoplasms; Malignant mesothelioma; Malignant neoplasm of liver; Malignant neoplasm of lung; Malignant neoplasm of mouth; Mammary

		Carcinoma, Animal; Mouth Neoplasms; Neoplasm Invasiveness; Osteoporosis; Osteoporosis, Age-Related; Osteoporosis, Senile; Post-Traumatic Osteoporosis; Squamous cell carcinoma; Weight Gain
	ARFGEF2	Epileptic encephalopathy; Heterotopia, Periventricular, Autosomal Recessive; Intellectual Disability; Malformations of Cortical Development; Malignant neoplasm of breast; Microcephaly; Periventricular Nodular Heterotopia
	CLNS1A	melanoma
	EIF2B1	Adenocarcinoma of lung (disorder); Ataxias, Hereditary; Childhood Ataxia with Central Nervous System Hypomyelinization; Epileptic encephalopathy; Leukodystrophy; OVARIOLEUKODYSTROPHY
	PIK3R1	ACTIVATED PI3K-DELTA SYNDROME; AGAMMAGLOBULINEMIA 7, AUTOSOMAL RECESSIVE; Abnormality of the cornea; Adenocarcinoma of large intestine; African Burkitt's lymphoma; Agammaglobulinemia, non-Bruton type; Alcoholic Intoxication, Chronic; Anaplastic carcinoma; Animal Mammary Neoplasms; Burkitt Lymphoma; Carcinoma; Carcinoma, Spindle-Cell; Carcinomatosis; Glioblastoma Multiforme; Glioma; IMMUNODEFICIENCY 36; Insulin Resistance; Insulin Sensitivity; Intellectual Disability; Malignant Neoplasms; Malignant neoplasm of ovary; Malignant neoplasm of prostate; Mammary Carcinoma, Animal; Mammary

		Neoplasms, Experimental; Neoplasm of uncertain or unknown behavior of ovary; Ovarian Carcinoma; Prostatic Neoplasms; SHORT syndrome; Undifferentiated carcinoma; ovarian neoplasm
	PIK3R2	Carcinoma in situ of endometrium; Cerebrovascular Disorders; Endometrial Carcinoma; Endometrial adenocarcinoma; Epileptic encephalopathy; Hydrocephalus; Intellectual Disability; MEGALENCEPHALY-POLYMICROGYRIA-POLYDACTYLY-HYDROCEPHALUS SYNDROME 1; Malformations of Cortical Development; Malignant neoplasm of endometrium; Malignant neoplasm of prostate; Megalanecephaly Polymicrogyria-Polydactyly Hydrocephalus Syndrome; Polydactyly; Prostatic Neoplasms
	PSMB4	HIV Coinfection; HIV Infections; Major Depressive Disorder; Unipolar Depression
	TAF1	Adenocarcinoma of large intestine; Dystonia 3, Torsion, X-Linked; Intellectual Disability; MENTAL RETARDATION, X-LINKED, SYNDROMIC 33; Metastatic melanoma; Parkinson Disease
	TKT	Alcoholic Intoxication, Chronic; Anaplastic carcinoma; Animal Mammary Neoplasms; Anoxemia; Anoxia; Carcinoma; Carcinoma, Spindle-Cell; Carcinomatosis; Hypoxemia; Hypoxia; Mammary Carcinoma, Animal; Mammary Neoplasms, Experimental; Neoplasm Invasiveness; SHORT STATURE, DEVELOPMENTAL DELAY, AND CONGENITAL

		HEART DEFECTS; Undifferentiated carcinoma; Wernicke Encephalopathy
Karlsson2013_T2D_T2D|IGT	CEP89	Mitochondrial Diseases
	COL1A1	Abortion, Tubal; Aneurysm, Dissecting; Aortic Aneurysm, Thoracic; Aortic Valve Insufficiency; Autoimmune Diseases; Calcinosis; Cholangitis; Cholestasis, Extrahepatic; Cirrhosis; Congenital aneurysm of ascending aorta; Cortical Congenital Hyperostosis; Dermatofibrosarcoma Protuberans; Dissection of aorta; Dissection, Blood Vessel; EHLERS-DANLOS SYNDROME, ARTHROCHALASIA TYPE; Ehlers-Danlos Syndrome; Ehlers-Danlos syndrome classic type; Ehlers-Danlos syndrome type 1; Ehlers-Danlos syndrome type 2; Ehlers-Danlos syndrome vascular-like type; Fibrosis; Fibrosis, Liver; Heart valve disease; Hypertensive disease; Keloid; Left Ventricular Hypertrophy; Liver Cirrhosis; Liver Cirrhosis, Experimental; Lobstein Disease; Microcalcification; Nephrogenic Fibrosing Dermopathy; Nephrotic Syndrome; Oral Submucous Fibrosis; Osteogenesis Imperfecta; Osteogenesis imperfecta type III (disorder); Osteogenesis imperfecta type IV (disorder); Osteogenesis imperfecta, dominant perinatal lethal; Osteoporosis; Osteoporosis, Age-Related; Osteoporosis, Senile; Post-Traumatic Osteoporosis; Spontaneous abortion; Tumoral calcinosis

	COL1A2	Abortion, Tubal; Aortic Aneurysm, Thoracic; Cirrhosis; Congenital aneurysm of ascending aorta; Degenerative polyarthritis; EDS VIIB; EHLERS-DANLOS SYNDROME, ARTHROCHALASIA TYPE; Ehlers-Danlos Syndrome; Ehlers-Danlos syndrome, cardiac valvular form; Fibrosis; Fibrosis, Liver; Heart valve disease; Intellectual Disability; Liver Cirrhosis; Liver Cirrhosis, Experimental; Lobstein Disease; Oral Submucous Fibrosis; Osteoarthrosis Deformans; Osteogenesis Imperfecta; Osteogenesis imperfecta type III (disorder); Osteogenesis imperfecta type IV (disorder); Osteogenesis imperfecta, dominant perinatal lethal; Osteoporosis; Osteoporosis, Age-Related; Osteoporosis, Senile; Post-Traumatic Osteoporosis; Spontaneous abortion; Systemic Scleroderma

	CTNNB1	Aberrant Crypt Foci; Adamantinous Craniopharyngioma; Adenocarcinoma; Adenocarcinoma of large intestine; Adenocarcinoma of lung (disorder); Adenocarcinoma, Basal Cell; Adenocarcinoma, Oxyphilic; Adenocarcinoma, Tubular; Adenoma; Adenoma, Basal Cell; Adenoma, Microcystic; Adenoma, Monomorphic; Adenoma, Trabecular; Adrenal Cancer; Adrenal Cortical Adenoma; Adrenal Gland Neoplasms; Adrenocortical carcinoma; Adult Craniopharyngioma; Adult Hepatocellular Carcinoma; Adult Medulloblastoma; Bilateral Wilms Tumor; Brain Neoplasms; Breast Carcinoma; Carcinoma in situ of endometrium; Carcinoma, Cribriform; Carcinoma, Granular Cell; Cecal Neoplasms; Chemical and Drug Induced Liver Injury; Chemically-Induced Liver Toxicity; Childhood Hepatocellular Carcinoma; Childhood Medulloblastoma; Colon Carcinoma; Colonic Neoplasms; Colorectal Cancer; Colorectal Carcinoma; Colorectal Neoplasms; Craniofacial Abnormalities; Craniopharyngioma; Craniopharyngioma, Child; Cutaneous Melanoma; Desmoplastic Medulloblastoma; Disease Exacerbation; Drug-Induced Acute Liver Injury; Drug-Induced Liver Disease; Drug-Induced Stevens Johnson Syndrome; EXUDATIVE VITREORETINOPATHY 7; Endometrial Carcinoma; Epithelial ovarian cancer; Experimental Hepatoma; Exudative

		vitreoretinopathy 1; Familial Exudative Vitreoretinopathy; Fibromatosis, Aggressive; Fibrosis, Liver; Follicular adenoma; Gastric Adenocarcinoma; Hemangiosarcoma; Hepatitis, Drug-Induced; Hepatitis, Toxic; Hepatoblastoma; Hepatocellular Adenoma; Hepatoma, Morris; Hepatoma, Novikoff; Intellectual Disability; Intestinal Cancer; Intestinal Neoplasms; Liver Cirrhosis; Liver Neoplasms, Experimental; Liver carcinoma; Liver neoplasms; Lung Neoplasms; Lymphoma, Lymphocytic, Intermediate; MENTAL RETARDATION, AUTOSOMAL DOMINANT 19; Malignant Neoplasms; Malignant lymphoma, lymphocytic, intermediate differentiation, diffuse; Malignant mesothelioma; Malignant neoplasm of brain; Malignant neoplasm of breast; Malignant neoplasm of cecum; Malignant neoplasm of endometrium; Malignant neoplasm of liver; Malignant neoplasm of lung; Malignant neoplasm of ovary; Malignant neoplasm of pancreas; Malignant neoplasm of prostate; Malignant tumor of colon; Mammary Neoplasms; Mammary Neoplasms, Human; Medulloblastoma; Medulloblastoma with extensive nodularity; Medullomyoblastoma; Melanotic medulloblastoma; Multiple congenital anomalies; Muscular Atrophy; Mycoplasma-Induced Stevens-Johnson Syndrome; Neoplasm Invasiveness; Neoplasm Metastasis; Neoplasm of uncertain or unknown behavior of ovary;

		Nephroblastoma; Nerve Degeneration; Neurodevelopmental Disorders; Neurogenic Muscular Atrophy; No-Reflow Phenomenon; Ovarian Carcinoma; Pancreatic Neoplasm; Papillary adenoma; Papillary craniopharyngioma; Parathyroid Adenoma; Peritoneal Neoplasms; Peritoneal Surface Malignancy; Pilomatrixoma; Primary microcephaly; Prostatic Neoplasms; Salivary Gland Neoplasms; Schizophrenia; Squamous cell carcinoma of esophagus; Stevens-Johnson Syndrome; Stevens-Johnson Syndrome Toxic Epidermal Necrolysis Spectrum; Toxic Epidermal Necrolysis; Uterine Cervical Neoplasm; Vascular calcification; cervical cancer; melanoma; ovarian neoplasm
	DEFA3	Acute Promyelocytic Leukemia
	HMG20A	Diabetes Mellitus, Non-Insulin-Dependent
	HSP90B1	Bipolar Disorder; Malignant neoplasm of prostate; Prostatic Neoplasms
	HSPB8	CHARCOT-MARIE-TOOTH DISEASE, AXONAL, TYPE 2L (disorder); Charcot-Marie-Tooth Disease; Charcot-Marie-Tooth disease, Type 2I; Distal Hereditary Motor Neuropathy, Type II; Distal Muscular Dystrophies; Glioblastoma Multiforme; NEURONOPATHY, DISTAL HEREDITARY MOTOR, TYPE IIA; Spinal muscular atrophy, Jerash type

	MED12	Adrenocortical carcinoma; Aortic Aneurysm, Thoracic; Arthrogryposis; Blepharophimosis syndrome Ohdo type; Breast Carcinoma; Congenital aneurysm of ascending aorta; Depressive disorder; Ehlers-Danlos Syndrome; Epileptic encephalopathy; FG SYNDROME 2; FG SYNDROME 3; FG SYNDROME 4 (disorder); FG syndrome; Fibroadenoma; Intellectual Disability; Lujan Fryns syndrome; Major Depressive Disorder; Malignant Cystosarcoma Phyllodes; Malignant neoplasm of breast; Malignant neoplasm of prostate; Mammary Neoplasms; Mammary Neoplasms, Human; Mental Depression; Mental Retardation, X-Linked 1; Nonorganic psychosis; Ohdo syndrome, Maat-Kievit-Brunner type; Phyllodes Tumor; Prostatic Neoplasms; Psychotic Disorders; Schizophrenia; Unipolar Depression
	MYLK	AORTIC ANEURYSM, FAMILIAL THORACIC 7; Acute Lung Injury; Allergic Reaction; Aortic Aneurysm, Thoracic; Brain Edema; Cerebral Edema; Cholera Infantum; Congenital aneurysm of ascending aorta; Cytotoxic Brain Edema; Cytotoxic Cerebral Edema; Ehlers-Danlos Syndrome; Experimental Lung Inflammation; Functional Gastrointestinal Disorders; Gastrointestinal Diseases; Glaucoma; Glioma; Hypercholesterolemia; Hypersensitivity; Lobar Pneumonia; Malignant Glioma; Marfan Syndrome; Megacystis microcolon intestinal

		hypoperistalsis syndrome; Neoplasm Invasiveness; Neoplasm Metastasis; Ovarian Mucinous Adenocarcinoma; Pneumonia; Pneumonitis; Vascular Diseases; Vasogenic Brain Edema; Vasogenic Cerebral Edema; mixed gliomas
	NR1D1	Bipolar Disorder; Depressive disorder; Fatty Liver; Seasonal Affective Disorder; Steatohepatitis
	PCNA	ATAXIA-TELANGIECTASIA-LIKE DISORDER; ATAXIA-TELANGIECTASIA-LIKE DISORDER 2; Adenocarcinoma; Adenocarcinoma, Basal Cell; Adenocarcinoma, Oxyphilic; Adenocarcinoma, Tubular; Benign neoplasm of brain, unspecified; Brain Neoplasms; Brain Tumor, Primary; Carcinoma, Cribriform; Carcinoma, Granular Cell; Colonic Neoplasms; Focal glomerulosclerosis; Hepatoblastoma; Hyalinosis, Segmental Glomerular; Ischemia; Lung Neoplasms; Malignant neoplasm of brain; Malignant neoplasm of lung; Malignant tumor of colon; Neoplasms, Experimental; Neoplasms, Intracranial; Primary malignant neoplasm of brain; Psoriasis; Pustulosis of Palms and Soles; Radiation Injuries, Experimental; Recurrent Brain Neoplasm
	PEX5	Adrenoleukodystrophy, Neonatal; Arthrogryposis; Cataract; Cholestasis; Cholestasis in newborn; Hydrops Fetalis; Infantile Refsum Disease (disorder); Intellectual Disability; Malformations of Cortical Development; PEROXISOME BIOGENESIS DISORDER 2A (ZELLWEGER); PEROXISOME

		BIOGENESIS DISORDER 2B; PEROXISOME BIOGENESIS DISORDER, COMPLEMENTATION GROUP 2; Peroxisomal Disorders; RHIZOMELIC CHONDRODYSPLASIA PUNCTATA, TYPE 5; Zellweger Syndrome
	TNXB	Cakut; Congenital aneurysm of ascending aorta; Ehlers-Danlos Syndrome; Ehlers-Danlos syndrome caused by tenascin-X deficiency; Ehlers-Danlos syndrome, type 3 (disorder); Osteogenesis Imperfecta; Schizophrenia; Squamous cell carcinoma of esophagus; VESICOURETERAL REFLUX 8; Vesico-Ureteral Reflux
	TPPP	Depression in children
	USP9X	Choanal Atresia; Intellectual Disability; MENTAL RETARDATION, X-LINKED 99; MENTAL RETARDATION, X-LINKED 99, SYNDROMIC, FEMALE-RESTRICTED; Mental Retardation, X-Linked 1; Polydactyly
	VCP	AMYOTROPHIC LATERAL SCLEROSIS 14 WITH OR WITHOUT FRONTOTEMPORAL DEMENTIA; Amyotrophic Lateral Sclerosis; Arthrogryposis; Behavioral variant of frontotemporal dementia; CHARCOT-MARIE-TOOTH DISEASE, AXONAL, TYPE 2Y; Charcot-Marie-Tooth Disease; Congenital myopathy (disorder); Distal Muscular Dystrophies; Drug-Induced Stevens Johnson Syndrome; Frontotemporal Dementia With Motor Neuron Disease; INCLUSION BODY

		MYOPATHY WITH EARLY-ONSET PAGET DISEASE AND FRONTOTEMPORAL DEMENTIA; Muscular Dystrophies, Limb-Girdle; Mycoplasma-Induced Stevens-Johnson Syndrome; Primary Progressive Nonfluent Aphasia; Stevens-Johnson Syndrome; Stevens-Johnson Syndrome Toxic Epidermal Necrolysis Spectrum; Toxic Epidermal Necrolysis
	ZNF277	Malignant neoplasm of breast
Lechatelier2013_Obesity_obese vs lean	BLOC1S6	HERMANSKY-PUDLAK SYNDROME 9; Hermanski-Pudlak Syndrome
	EEF2	Adenocarcinoma; Adenocarcinoma, Basal Cell; Adenocarcinoma, Oxyphilic; Adenocarcinoma, Tubular; Breast Carcinoma; Carcinoma, Cribriform; Carcinoma, Granular Cell; Chromophobe Renal Cell Carcinoma; Collecting Duct Carcinoma of the Kidney; Conventional (Clear Cell) Renal Cell Carcinoma; Degenerative polyarthritis; Lung Neoplasms; Malignant neoplasm of breast; Malignant neoplasm of lung; Mammary Neoplasms; Mammary Neoplasms, Human; Neoplasm Invasiveness; Neoplasm Metastasis; Osteoarthrosis Deformans; Papillary Renal Cell Carcinoma; Renal Cell Carcinoma; SPINOCEREBELLAR ATAXIA 26; Sarcomatoid Renal Cell Carcinoma
	FYB1	Dermatitis, Allergic Contact; Thrombocytopenia 3
	GOPC	melanoma
	ITSN2	Breast Carcinoma; Malignant neoplasm of breast; Malignant neoplasm of prostate; Mammary Neoplasms; Mammary Neoplasms,

		Human; Neoplasm Recurrence, Local; Prostatic Neoplasms; Sicca Syndrome; Sjogren's Syndrome
	MAP4K1	Metastatic melanoma
	NFE2L2	Acute Kidney Insufficiency; Acute Lung Injury; Acute kidney injury; Adenocarcinoma of lung (disorder); Adverse reaction to drug; Carcinogenesis; Carcinoma, Pancreatic Ductal; Chemical and Drug Induced Liver Injury; Chemically-Induced Liver Toxicity; Cholera Infantum; Cholestasis; Chromosome Breakage; Chromosome Breaks; Congestive heart failure; Dermatitis, Allergic Contact; Diabetic Nephropathy; Drug toxicity; Drug-Induced Acute Liver Injury; Drug-Induced Liver Disease; Encephalopathy, Toxic; Endometrial Carcinoma; Endometrial Neoplasms; Experimental Autoimmune Encephalomyelitis; Fatty Liver; Fibrosis, Liver; Functional Gastrointestinal Disorders; Gastrointestinal Diseases; Gonadotropin-Resistant Ovary Syndrome; Hamman-Rich syndrome; Heart Decompensation; Heart Failure, Right-Sided; Heart failure; Hepatitis, Drug-Induced; Hepatitis, Toxic; Hyperglycemia; Hyperglycemia, Postprandial; Hypergonadotropic Ovarian Failure, X-Linked; Hyperplasia; Keratoma; Keratosis; Keratosis Blennorrhagica; Kidney Diseases; Kidney Failure, Acute; Left-Sided Heart Failure; Liver Cirrhosis; Liver Cirrhosis, Experimental; Liver carcinoma; Liver neoplasms; Malignant neoplasm of liver; Malignant neoplasm of prostate; Malignant neoplasm of skin; Myocardial Failure; Necrosis;

		Neurotoxicity Syndromes; Nodular glomerulosclerosis; Non-Small Cell Lung Carcinoma; Non-alcoholic Fatty Liver Disease; Nonalcoholic Steatohepatitis; Ovarian Failure, Premature; Pathological accumulation of air in tissues; Prostatic Neoplasms; Pulmonary Fibrosis; Skin Neoplasms; Squamous cell carcinoma; Squamous cell carcinoma of esophagus; Squamous cell carcinoma of the head and neck; Steatohepatitis; Toxic Encephalitis; Vascular System Injuries
	PEX26	Adrenoleukodystrophy; Adrenoleukodystrophy, Neonatal; Adrenomyeloneuropathy; Amelogenesis Imperfecta; Arthrogryposis; Cataract; Cholestasis; Cholestasis in newborn; Hydrops Fetalis; Infantile Refsum Disease (disorder); Intellectual Disability; Leukodystrophy; Malformations of Cortical Development; PEROXISOME BIOGENESIS DISORDER 7A (ZELLWEGER); PEROXISOME BIOGENESIS DISORDER 7B; PEROXISOME BIOGENESIS DISORDER, COMPLEMENTATION GROUP 8; PEROXISOME BIOGENESIS DISORDER, COMPLEMENTATION GROUP A; Peroxisomal Disorders; Peroxisome biogenesis disorders; Zellweger Spectrum; Zellweger Syndrome; Zellweger-Like Syndrome
	PHKG2	Glycogen Storage Disease IXC; Intellectual Disability; Ketotic hypoglycemia
	PRDX4	Dermatitis, Allergic Contact; Diffuse Large B-Cell Lymphoma; Disease Exacerbation;

		Hyperglycemia; Hyperglycemia, Postprandial; Impaired glucose tolerance; Intellectual Disability
	PSME2	HIV Coinfection; HIV Infections; Liver Cirrhosis, Experimental; Squamous cell carcinoma of esophagus
	RTN2	Intellectual Disability; SPASTIC PARAPLEGIA 12, AUTOSOMAL DOMINANT (disorder); Spastic Paraplegia, Hereditary
	SMARCE1	Adenoid Cystic Carcinoma; COFFIN-SIRIS SYNDROME 5; Coffin-Siris syndrome; Familial meningioma; Intellectual Disability; Meningioma; Meningiomas, Multiple
	STAB1	Bipolar Disorder; Juvenile arthritis; Juvenile psoriatic arthritis; Juvenile-Onset Still Disease; Manic; Manic mood
	SYNCRIP	Intellectual Disability; leukemia
	TRAF5	Diffuse Large B-Cell Lymphoma
Nielsen2014_IBD_IBD	UXT	Malignant neoplasm of prostate; Prostatic Neoplasms
Qin2017_T2D_T2D	ABLIM2	Pancreatic Ductal Adenocarcinoma
	AFG1L	Bipolar Disorder
	ANXA7	Liver neoplasms; Malignant neoplasm of liver
	CASP4	Liver Cirrhosis, Experimental; Schizophrenia
	DDB1	Disease Exacerbation; Hereditary Diffuse Gastric Cancer; Malignant neoplasm of stomach; Stomach Neoplasms
	DNMT3A	Acute Myeloid Leukemia (AML-M2); Acute Myeloid Leukemia, M1; Acute Promyelocytic Leukemia; Acute monocytic leukemia; Angioimmunoblastic Lymphadenopathy; Autism Spectrum Disorders; Breast Carcinoma; Clear-cell metastatic

		renal cell carcinoma; Congenital anemia; Craniofacial Abnormalities; Crohn Disease; Crohn's disease of large bowel; Crohn's disease of the ileum; Cytopenia; Facies; Granulomatous Slack Skin; Growth Disorders; IIeocolitis; Intellectual Disability; Juvenile Myelomonocytic Leukemia; Leukemia, Myelocytic, Acute; Leukemia, Myelomonocytic, Chronic; Lung Neoplasms; Lymphoma, T-Cell, Cutaneous; Malignant neoplasm of breast; Malignant neoplasm of lung; Mammary Neoplasms; Mammary Neoplasms, Human; Mental Retardation, Psychosocial; Mental deficiency; Neoplasm Recurrence, Local; Peripheral T-Cell Lymphoma; Profound Mental Retardation; Regional enteritis; Sezary Syndrome; TATTON-BROWN-RAHMAN SYNDROME; cocaine use
	GLS	Abnormality of the cornea; Abortion, Tubal; Congenital Disorders of Glycosylation; Degenerative polyarthritis; Liver Cirrhosis, Experimental; Osteoarthrosis Deformans; Schizophrenia; Spontaneous abortion
	HCAR2	Bipolar Disorder; Flushing; Malignant neoplasm of skin; Psoriasis; Pustulosis of Palms and Soles; Schizophrenia; Skin Neoplasms; Squamous cell carcinoma
	HCAR3	Bipolar Disorder; Malignant neoplasm of skin; Schizophrenia; Skin Neoplasms; Squamous cell carcinoma

	HMOX1	Acute Confusional Senile Dementia; Acute Kidney Insufficiency; Acute kidney injury; Adrenoleukodystrophy; Adrenomyeloneuropathy; Adult Learning Disorders; Adverse reaction to drug; Aggressive Systemic Mastocytosis; Alloxan Diabetes; Alzheimer Disease, Early Onset; Alzheimer Disease, Late Onset; Alzheimer's Disease; Alzheimer's Disease, Focal Onset; Anemia, Hemolytic; Anemia, Hemolytic, Acquired; Anemia, Microangiopathic; Arterial Diseases, Common Carotid; Asthma; Blood Coagulation Disorders; Breast Carcinoma; Cardiac Hypertrophy; Cardiomegaly; Carotid Artery Diseases; Carotid Atherosclerosis; Centriacinar Emphysema; Cerebral Hemorrhage; Chemical and Drug Induced Liver Injury; Chemically-Induced Liver Toxicity; Chronic Airflow Obstruction; Chronic Lung Injury; Chronic Obstructive Airway Disease; Cirrhosis; Colitis; Compensatory Hyperinsulinemia; Complex Partial Status Epilepticus; Congestive heart failure; Contact Dermatitis; Contact hypersensitivity; Coronary Arteriosclerosis; Coronary Artery Disease; Degenerative Diseases, Central Nervous System; Degenerative Diseases, Spinal Cord; Developmental Academic Disorder; Diabetes Mellitus, Experimental; Diabetes Mellitus, Non-Insulin-Dependent; Diabetic Angiopathies; Drug toxicity; Drug-Induced Acute Liver Injury; Drug-Induced Liver Disease; Endogenous Hyperinsulinism; Exogenous Hyperinsulinism;

		Experimental Lung Inflammation; External Carotid Artery Diseases; Extravascular Hemolysis; Familial Alzheimer Disease (FAD); Fibrosis; Focal Emphysema; Gastroparesis; Grand Mal Status Epilepticus; Growth Disorders; Hamman-Rich syndrome; Heart Decompensation; Heart Failure, Right-Sided; Heart failure; Heme Oxygenase 1 Deficiency; Hemolysis (disorder); Hepatitis; Hepatitis, Drug-Induced; Hepatitis, Toxic; Hepatorenal Syndrome; Hereditary Diffuse Gastric Cancer; Hyperinsulinism; Hyperplasia; Hypertensive disease; Indolent Systemic Mastocytosis; Inflammation; Injury wounds; Insulin Resistance; Insulin Sensitivity; Internal Carotid Artery Diseases; Intravascular hemolysis; Iron Metabolism Disorders; Ischemia; Kidney Failure, Acute; Kidney Failure, Chronic; Learning Disabilities; Learning Disorders; Learning Disturbance; Left-Sided Heart Failure; Leishmaniasis, Visceral; Liver Cirrhosis, Experimental; Liver Dysfunction; Liver diseases; Liver neoplasms; Lobar Pneumonia; Lung Injury; Lung Neoplasms; Malignant neoplasm of breast; Malignant neoplasm of liver; Malignant neoplasm of lung; Malignant neoplasm of prostate; Malignant neoplasm of stomach; Mammary Neoplasms; Mammary Neoplasms, Experimental; Mammary Neoplasms, Human; Mastocytosis, Systemic; Microangiopathic hemolytic anemia; Microangiopathy, Diabetic; Myocardial Failure; Myocardial Ischemia; Neoplasm Invasiveness;

		Neurodegenerative Disorders; Non-Convulsive Status Epilepticus; Obesity; Panacinar Emphysema; Pancreatic Diseases; Parkinson Disease; Petit mal status; Pneumonia; Pneumonitis; Pre-Eclampsia; Presenile dementia; Prostatic Neoplasms; Pulmonary Emphysema; Pulmonary Fibrosis; Reperfusion Injury; Research-Related Injuries; Retinal Diseases; SPINOCEREBELLAR ATAXIA 17; Schizophrenia; Simple Partial Status Epilepticus; Status Epilepticus; Status Epilepticus, Subclinical; Stomach Neoplasms; Streptozotocin Diabetes; Thrombosis; Thrombus; Traumatic injury; Vascular System Injuries; Wounds and Injuries
	KIF2A	CORTICAL DYSPLASIA, COMPLEX, WITH OTHER BRAIN MALFORMATIONS 3; Cortical Dysplasia; Epileptic encephalopathy; Intellectual Disability; Malformations of Cortical Development; Microcephaly; Microlissencephaly; Schizophrenia; Severe Congenital Microcephaly
	NR1D1	Bipolar Disorder; Depressive disorder; Fatty Liver; Seasonal Affective Disorder; Steatohepatitis
	PYGB	Adenoid Cystic Carcinoma; Malignant neoplasm of salivary gland; Myocardial Ischemia; Salivary Gland Neoplasms
	SHOC2	Female Pseudo-Turner Syndrome; Hair Diseases; Hydrops Fetalis; Hypertrophic Cardiomyopathy; Intellectual Disability; NOONAN SYNDROME-LIKE DISORDER

		WITH LOOSE ANAGEN HAIR 1; Noonan Syndrome; Noonan syndrome-like disorder with loose anagen hair; Noonan-Like Syndrome With Loose Anagen Hair; Turner Syndrome, Male
	SMAD3	Aneurysm, Dissecting; Aortic Aneurysm; Aortic Aneurysm, Thoracic; Arthrogryposis; Cerebrovascular Disorders; Cocaine Abuse; Cocaine Dependence; Cocaine-Related Disorders; Colorectal Cancer; Congenital aneurysm of ascending aorta; Craniofacial Abnormalities; Crohn Disease; Crohn's disease of large bowel; Crohn's disease of the ileum; Degenerative polyarthritis; Dissection of aorta; Dissection, Blood Vessel; Ehlers-Danlos Syndrome; Fibroid Tumor; Fibrosis, Liver; IIeocolitis; Intellectual Disability; Juvenile arthritis; Juvenile psoriatic arthritis; Juvenile-Onset Still Disease; LOEYS-DIETZ SYNDROME 3; Left Ventricle Remodeling; Liver Cirrhosis; Liver Cirrhosis, Alcoholic; Loeys-Dietz Aortic Aneurysm Syndrome; Loeys-Dietz Syndrome; Marfan Syndrome; Osteoarthrosis Deformans; Regional enteritis; Uterine Cancer; Uterine Fibroids; Uterine Neoplasms; Ventricular Remodeling
	STMN1	Anaplastic carcinoma; Animal Mammary Neoplasms; Breast Carcinoma; Carcinoma; Carcinoma, Spindle-Cell; Carcinomatosis; Depressive disorder; Glioma; Malignant Glioma; Malignant neoplasm of breast; Malignant neoplasm of prostate; Mammary Carcinoma, Animal; Mammary Neoplasms;

		Mammary Neoplasms, Experimental; Mammary Neoplasms, Human; Mental Depression; Prostatic Neoplasms; Undifferentiated carcinoma; mixed gliomas
	SYNGAP1	Aura; Autistic Disorder; Awakening Epilepsy; Epilepsy; Epilepsy, Cryptogenic; Epileptic encephalopathy; Intellectual Disability; Mental Retardation, Autosomal Dominant 5; Mental Retardation, Psychosocial; Mental deficiency; Neurodevelopmental Disorders; Profound Mental Retardation; Schizophrenia
	TAF9	HIV Coinfection; HIV Infections
	TMBIM4	Corpus Luteum Cyst; Ovarian Cysts
	TNKS2	Intellectual Disability
Schirmer2018_IBD_CD|UC	CEP89	Mitochondrial Diseases
	CLNS1A	melanoma
	CTSD	AMYOTROPHIC LATERAL SCLEROSIS 1; Amyotrophic Lateral Sclerosis, Sporadic; Chromophobe Renal Cell Carcinoma; Collecting Duct Carcinoma of the Kidney; Conventional (Clear Cell) Renal Cell Carcinoma; Degenerative polyarthritis; Epileptic encephalopathy; Intellectual Disability; Kidney Diseases; Liver carcinoma; Malignant neoplasm of prostate; NEURONAL CEROID LIPOFUSCINOSIS DUE TO CATHEPSIN D DEFICIENCY; Neoplasm Invasiveness; Neuronal Ceroid Lipofuscinosis, Congenital; Osteoarthrosis Deformans; Papillary Renal Cell Carcinoma; Prostatic Neoplasms; Renal Cell Carcinoma; Rheumatoid Arthritis; Sarcomatoid Renal Cell Carcinoma; Weight Gain

	HNRNPR	Breast Carcinoma; Malignant neoplasm of breast; Mammary Neoplasms; Mammary Neoplasms, Human
	TKT	Alcoholic Intoxication, Chronic; Anaplastic carcinoma; Animal Mammary Neoplasms; Anoxemia; Anoxia; Carcinoma; Carcinoma, Spindle-Cell; Carcinomatosis; Hypoxemia; Hypoxia; Mammary Carcinoma, Animal; Mammary Neoplasms, Experimental; Neoplasm Invasiveness; SHORT STATURE, DEVELOPMENTAL DELAY, AND CONGENITAL HEART DEFECTS; Undifferentiated carcinoma; Wernicke Encephalopathy
	VCP	AMYOTROPHIC LATERAL SCLEROSIS 14 WITH OR WITHOUT FRONTOTEMPORAL DEMENTIA; Amyotrophic Lateral Sclerosis; Arthrogryposis; Behavioral variant of frontotemporal dementia; CHARCOT-MARIE-TOOTH DISEASE, AXONAL, TYPE 2Y; Charcot-Marie-Tooth Disease; Congenital myopathy (disorder); Distal Muscular Dystrophies; Drug-Induced Stevens Johnson Syndrome; Frontotemporal Dementia With Motor Neuron Disease; INCLUSION BODY MYOPATHY WITH EARLY-ONSET PAGET DISEASE AND FRONTOTEMPORAL DEMENTIA; Muscular Dystrophies, Limb-Girdle; Mycoplasma-Induced Stevens-Johnson Syndrome; Primary Progressive Nonfluent Aphasia; Stevens-Johnson Syndrome; Stevens-Johnson Syndrome Toxic

		Epidermal Necrolysis Spectrum; Toxic Epidermal Necrolysis
Yu2015_CRC_CRC	ARPC2	Ulcerative Colitis
	BRWD3	Intellectual Disability; MENTAL RETARDATION, X-LINKED 93 (disorder)
	CD34	Major Depressive Disorder; Neoplastic Cell Transformation; Nonorganic psychosis; Psychotic Disorders; Unipolar Depression
	CEP85L	Malignant neoplasm of breast
	CXCR5	Biliary Cirrhosis, Secondary; Biliary cirrhosis; Malignant neoplasm of breast; Primary biliary cirrhosis
	EIF4A1	HIV Coinfection; HIV Infections
	EIF4A3	Intellectual Disability; Richieri Costa Pereira syndrome
	G3BP2	Malignant neoplasm of breast
	GCC1	Cannabis Abuse; Cannabis Dependence; Cannabis-Related Disorder; Cocaine Abuse; Cocaine Dependence; Cocaine-Related Disorders; Hashish Abuse; Marijuana Abuse; Phencyclidine Abuse; Phencyclidine-Related Disorders
	ITGB4	Abortion, Tubal; Adenoid Cystic Carcinoma; Adult junctional epidermolysis bullosa (disorder); Amelogenesis Imperfecta; Anhydrotic Ectodermal Dysplasias; Aplasia Cutis Congenita; Astrocytosis; EPIDERMOLYSIS BULLOSA, JUNCTIONAL, LOCALISATA VARIANT (disorder); Ectodermal Dysplasia; Epidermolysis Bullosa; Epidermolysis Bullosa Herpetiformis Dowling-Meara; Epidermolysis Bullosa Progressiva; Epidermolysis Bullosa Simplex;

		Epidermolysis Bullosa Simplex Kobner; Epidermolysis Bullosa Simplex With Pyloric Atresia; Epidermolysis bullosa inversa dystrophica; Epidermolysis bullosa with pyloric atresia; Gliosis; Herlitz Disease; Hidrotic Ectodermal Dysplasia; JEB-I; Junctional Epidermolysis Bullosa; Malignant neoplasm of salivary gland; Non-Small Cell Lung Carcinoma; Pyloric Atresia; Salivary Gland Neoplasms; Spontaneous abortion; Weber-Cockayne Syndrome
	NSUN2	Dubowitz syndrome; Intellectual Disability; MENTAL RETARDATION, AUTOSOMAL RECESSIVE 5
	PABPC1	Bladder Neoplasm; Breast Carcinoma; Carcinoma, Transitional Cell; Malignant neoplasm of breast; Malignant neoplasm of urinary bladder; Mammary Neoplasms; Mammary Neoplasms, Human
	PCK1	Alloxan Diabetes; Congestive heart failure; Diabetes Mellitus, Experimental; Heart Decompensation; Heart Failure, Right-Sided; Heart failure; Left-Sided Heart Failure; Liver Cirrhosis, Experimental; Liver carcinoma; Myocardial Failure; Obesity; Phosphoenolpyruvate carboxykinase deficiency; Schizophrenia; Streptozotocin Diabetes
	PCK2	Anoxemia; Anoxia; Hypoxemia; Hypoxia; Liver Cirrhosis, Experimental; Phosphoenolpyruvate carboxykinase 2 deficiency; Phosphoenolpyruvate carboxykinase deficiency
	PDZK1	Adenoid Cystic Carcinoma; Breast Carcinoma; Malignant neoplasm of

		breast; Malignant neoplasm of prostate; Malignant neoplasm of salivary gland; Mammary Neoplasms; Mammary Neoplasms, Human; Prostatic Neoplasms; Salivary Gland Neoplasms; Schizophrenia
	PEPD	Deficiency of prolidase; Diabetes Mellitus, Non-Insulin-Dependent; Hepatomegaly; Infection; Intellectual Disability; Organophosphate poisoning; Organophosphorus Poisoning; Organothiophosphate Poisoning; Organothiophosphonate Poisoning; Respiratory Tract Diseases; Skin Ulcer; Splenomegaly
	PEX5	Adrenoleukodystrophy, Neonatal; Arthrogryposis; Cataract; Cholestasis; Cholestasis in newborn; Hydrops Fetalis; Infantile Refsum Disease (disorder); Intellectual Disability; Malformations of Cortical Development; PEROXISOME BIOGENESIS DISORDER 2A (ZELLWEGER); PEROXISOME BIOGENESIS DISORDER 2B; PEROXISOME BIOGENESIS DISORDER, COMPLEMENTATION GROUP 2; Peroxisomal Disorders; RHIZOMELIC CHONDRODYSPLASIA PUNCTATA, TYPE 5; Zellweger Syndrome
	PGK1	Chromophobe Renal Cell Carcinoma; Collecting Duct Carcinoma of the Kidney; Conventional (Clear Cell) Renal Cell Carcinoma; Deficiency of phosphoglycerate kinase; Intellectual Disability; Ketotic hypoglycemia; Liver carcinoma; Papillary Renal Cell Carcinoma; Phosphoglycerate Kinase 1

		Deficiency; Renal Cell Carcinoma; Rhabdomyolysis; Sarcomatoid Renal Cell Carcinoma
	PHIP	Colorectal Cancer; Intellectual Disability; Mental Retardation, Psychosocial; Mental deficiency; Profound Mental Retardation; Sicca Syndrome; Sjogren's Syndrome
	SERPINH1	Embolic Infarction, Middle Cerebral Artery; Fibrosis, Liver; Infarction, Middle Cerebral Artery; Ischemia; Left Middle Cerebral Artery Infarction; Liver Cirrhosis; Liver Cirrhosis, Experimental; Middle Cerebral Artery Embolus; Middle Cerebral Artery Occlusion; Middle Cerebral Artery Syndrome; Middle Cerebral Artery Thrombosis; OSTEOGENESIS IMPERFECTA, TYPE X; Osteogenesis Imperfecta; Osteogenesis imperfecta type III (disorder); Preterm premature rupture of membranes (disorder); Respiratory Distress Syndrome, Adult; Right Middle Cerebral Artery Infarction; Thrombotic Infarction, Middle Cerebral Artery
	SREBF1	Alloxan Diabetes; Diabetes Mellitus, Experimental; Fatty Liver; Insulin Resistance; Insulin Sensitivity; Kidney Failure, Chronic; Liver Cirrhosis, Experimental; Liver carcinoma; Non-alcoholic Fatty Liver Disease; Nonalcoholic Steatohepatitis; Obesity; Schizophrenia; Steatohepatitis; Streptozotocin Diabetes
	VTN	Crohn Disease; Crohn's disease of large bowel; Crohn's disease of the ileum; Fibrosis, Liver; IIeocolitis; Liver Cirrhosis; Regional enteritis

Zeller2014_CRC_CRC|adenoma	ATRN	Liver Cirrhosis, Experimental; Nerve Degeneration
	BCAP31	Contiguous Abcd1-Dxs1375e Deletion Syndrome; DEAFNESS, DYSTONIA, AND CEREBRAL HYPOMYELINATION; Intellectual Disability; Leukodystrophy
	BRWD3	Intellectual Disability; MENTAL RETARDATION, X-LINKED 93 (disorder)
	CD6	Multiple Sclerosis; Multiple Sclerosis, Acute Fulminating
	CXCR5	Biliary Cirrhosis, Secondary; Biliary cirrhosis; Malignant neoplasm of breast; Primary biliary cirrhosis
	DEFA3	Acute Promyelocytic Leukemia
	EIF2S3	Arthrogryposis; Epileptic encephalopathy; Intellectual Disability; MENTAL RETARDATION, EPILEPTIC SEIZURES, HYPOGONADISM AND HYPOGENITALISM, MICROCEPHALY, AND OBESITY (disorder); Microcephaly
	EPAS1	Autosome Abnormalities; Benign neoplasm of adrenal gland; Benign neoplasm of aortic body and other paraganglia; Chromophobe Renal Cell Carcinoma; Chromosome Aberrations; Collecting Duct Carcinoma of the Kidney; Conventional (Clear Cell) Renal Cell Carcinoma; Erythrocytosis, Familial, 4; Familial erythrocytosis; Malignant Adrenal Medulla Neoplasm; Malignant neoplasm of aortic body and other paraganglia; Neoplastic Cell Transformation; Papillary Renal Cell Carcinoma; Renal Cell Carcinoma; Sarcomatoid Renal Cell Carcinoma; Secondary polycythemia

	FUT8	Schizophrenia; Schizophrenia and related disorders
	GLUL	Alcohol withdrawal syndrome; Alcoholic Intoxication, Chronic; Brain Diseases, Metabolic, Inborn; Brain Diseases, Metabolic, Inherited; Central Nervous System Inborn Metabolic Diseases; Depressive disorder; Epileptic encephalopathy; Experimental Hepatoma; Fulminant Hepatic Failure with Cerebral Edema; Glutamine deficiency, congenital; Hepatic Coma; Hepatic Encephalopathy; Hepatic Stupor; Hepatoma, Morris; Hepatoma, Novikoff; Intellectual Disability; Liver Neoplasms, Experimental; Liver carcinoma; Mental Depression; Mood Disorders; Obesity; Psychotic Disorders; Schizophrenia
	MADD	Intellectual Disability
	MYH10	Focal glomerulosclerosis; Hyalinosis, Segmental Glomerular; Intellectual Disability
	MYH9	Alport Syndrome; Alport Syndrome, Autosomal Dominant; Alport Syndrome, Autosomal Recessive; Alport Syndrome, X-Linked; Breast Carcinoma; COCHLEOSACCULAR DEGENERATION; Carcinoma, Lobular; Cataract; Congenital anemia; Cytopenia; DEAFNESS, AUTOSOMAL DOMINANT 17; Deafness, autosomal dominant nonsyndromic sensorineural 17; Epstein syndrome (disorder); Familial hematuria; Fechtner syndrome (disorder); Focal glomerulosclerosis; Hemorrhagic hereditary nephritis; Hereditary nephritis; Hyalinosis, Segmental Glomerular; Intellectual Disability; Kidney Failure, Chronic;

		MACROTHROMBOCYTOPENIA AND PROGRESSIVE SENSORINEURAL DEAFNESS; Malignant neoplasm of breast; Mammary Neoplasms; Mammary Neoplasms, Human; May-Hegglin anomaly; Neoplasm Invasiveness; Nodular fasciitis; Renal hypertension; SEBASTIAN SYNDROME; Schizophrenia
	PAH	Bipolar Disorder; Classical phenylketonuria; Epileptic encephalopathy; Hyperphenylalaninaemia; Hyperphenylalaninemia, Non-Phenylketonuric; Hyperphenylalaninemia, Non-Pku Mild; Intellectual Disability; Phenylketonuria II; Phenylketonuria, Maternal; Phenylketonurias; Psychotic Disorders; Schizophrenia
	PEPD	Deficiency of prolidase; Diabetes Mellitus, Non-Insulin-Dependent; Hepatomegaly; Infection; Intellectual Disability; Organophosphate poisoning; Organophosphorus Poisoning; Organothiophosphate Poisoning; Organothiophosphonate Poisoning; Respiratory Tract Diseases; Skin Ulcer; Splenomegaly
	PGM1	Colorectal Cancer; Colorectal Carcinoma; Colorectal Neoplasms; Congenital Disorders of Glycosylation; Glycogen Storage Disease XIV; Intellectual Disability; Ketotic hypoglycemia; Myocardial Ischemia; Rhabdomyolysis
	PHIP	Colorectal Cancer; Intellectual Disability; Mental Retardation, Psychosocial; Mental deficiency; Profound Mental Retardation; Sicca Syndrome; Sjogren's Syndrome

	PIK3R5	ATAXIA-OCULOMOTOR APRAXIA 3; Ataxias, Hereditary; Colorectal Cancer; SPINOCEREBELLAR ATAXIA, AUTOSOMAL RECESSIVE 1
	RAB11A	Chloracne
	SLC30A9	BIRK-LANDAU-PEREZ SYNDROME; Intellectual Disability
	SPHK2	Age-Related Memory Disorders; Embolic Infarction, Middle Cerebral Artery; Infarction, Middle Cerebral Artery; Left Middle Cerebral Artery Infarction; Memory Disorder, Semantic; Memory Disorder, Spatial; Memory Disorders; Memory Loss; Memory impairment; Middle Cerebral Artery Embolus; Middle Cerebral Artery Occlusion; Middle Cerebral Artery Syndrome; Middle Cerebral Artery Thrombosis; Right Middle Cerebral Artery Infarction; Thrombotic Infarction, Middle Cerebral Artery
	ULK1	Adenocarcinoma of lung (disorder); Ovarian Mucinous Adenocarcinoma

Supplementary Table S4.3. Important human interactors that are known drug targets. For each human protein in the $zsig_{hum}$ subset, we list the drug interactor and the study in which it was found to be important.

Human Protein	Significant Substudies (z-strict)	Drug Targeters
BRD7	Lechatelier (Obesity)	BI-9564
		LP99
		TP-472
CASP4	Qin (T2D)	EMRICASAN
		M826
		VRT-043198
		casp 4 inhib
CSF1R	Feng (CRC)	IMATINIB
		AC710
		AC710 (Mesylate)
		AZD6495
		BLZ945
		CC-223
		CEDIRANIB
		CRENOLANIB
		DOVITINIB
		ENMD-2076
		FOSTAMATINIB
		GR-389988
		GTP 14564
		GW2580
		ILORASERTIB
		JNJ-28312141
		JNJ-40346527
		Ki20227
		LINIFANIB
		OSI-930
		PAZOPANIB
		Pazopanib (Hydrochloride)
		QUIZARTINIB
		QUIZARTINIB DIHYDROCHLORIDE

		SUNITINIB
		SUNITINIB MALATE
		TANDUTINIB
		TG-02
		VATALANIB
		Vatalanib (PTK787) 2HCl
		Vatalanib succinate
		cerdulatinib
		pexidartinib
CTSD	Schirmer (IBD)	GRASSYSTATIN A
		compound 1 [PMID: 10498202]
		compound 3 [PMID: 8410973]
DDB1	Qin (T2D)	CC-5013 hydrochloride
		LENALIDOMIDE
		Lenalidomide (hemihydrate)
		Lenalidomide-d5
		POMALIDOMIDE
		Pomalidomide-d5
		THALIDOMIDE
DNMT3A	Qin (T2D)	AZACITIDINE
		DECITABINE
EEF2	Lechatelier (Obesity)	esketamine
GLS	Qin (T2D)	CB-839
HCAR2	Qin (T2D)	(R)-3-HYDROXYBUTANOATE
		5-methyl nicotinic acid
		ACIFRAN
		ACIPIMOX
		CINNAMIC ACID
		GSK256073
		L-??Hydroxybutyric acid
		MK 1903
		MK-6892
		NIACIN
		PD051995
		SCH-900271
		butyric acid
		compound (+)17a [PMID: 20363624]
		compound 21 [PMID: 21185185]

		compound 2g [PMID: 19309152]
		compound 8f [PMID: 20615702]
		homonicotinic acid
		inositol nicotinate
		monomethyl fumarate
		p-Coumaric Acid
		compound 42 [PMID: 22420767]
		compound 9n [PMID: 18752940]
		CLOTRIMAZOLE
		MK-0354
		compound 1q [PMID: 18029181]
HCAR3	Qin (T2D)	2-hydroxyoctanoic acid
		3-hydroxyoctanoic acid
		ACIFRAN
		D-TRYPTOPHAN
		D-kynurenine
		D-phenylalanine
		IPBT-5CA
		L-phenylalanine
		L-tryptophan
		NIACIN
		PD047831
		PD048724
		compound 5b [PMID: 17358052]
		compound 6o [PMID: 19524438]
		inositol nicotinate
HMOX1	Qin (T2D)	compound 1 [PMID: 16821802]
HSP90B1	Karlsson (T2D)	GELDANAMYCIN
		Grp94 inhibitor 54
		PU-WS13
		SEMAPIMOD
		RIFABUTIN
KDM7A	Lechatelier (Obesity), Yu (CRC)	daminozide
MAP4K1	Lechatelier (Obesity)	FOSTAMATINIB
MYLK	Karlsson (T2D)	FOSTAMATINIB
		RKI-1447
NFE2L2	Lechatelier (Obesity)	ML385

		bardoxolone methyl
NPEPPS	Yu (CRC)	ANTAQ
		TOSEDOSTAT
NR1D1	Karlsson (T2D), Qin (T2D)	GSK4112
		SR9009
		SR9011
		SR9011 (hydrochloride)
		heme
		SR8278
PAH	Zeller (CRC)	SAPROPTERIN DIHYDROCHLORIDE
		sapropterin
		DROXIDOPA
		FENCLONINE
		NOREPINEPHRINE
PCNA	Karlsson (T2D)	LIOTHYRONINE
PHKG2	Lechatelier (Obesity)	STAUROSPORINE
		compound 2c [PMID: 24900538]
PIK3R1	Hannigan (CRC)	isoprenaline
		APITOLISIB
		AZD6482
		BEZ235 (Tosylate)
		BUPARLISIB
		CH5132799
		COPANLISIB
		COPANLISIB HYDROCHLORIDE
		DACTOLISIB
		GDC-0941
		GDC-0941 (dimethanesulfonate)
		GEDATOLISIB
		GSK1059615
		NVP-BEZ235 (hydrochloride)
		NVP-BGT226
		NVP-BKM120 (Hydrochloride)
		OMIPALISIB
		PF-04691502
		PILARALISIB
		PX 866
		RECILISIB

		RG7422
		SONOLISIB
		TASELISIB
		VS-5584
		voxtalisib
PIK3R2	Hannigan (CRC)	isoprenaline
		APITOLISIB
		AZD6482
		BEZ235 (Tosylate)
		BUPARLISIB
		CH5132799
		COPANLISIB
		COPANLISIB HYDROCHLORIDE
		DACTOLISIB
		GDC-0941
		GDC-0941 (dimethanesulfonate)
		GEDATOLISIB
		GSK1059615
		NVP-BEZ235 (hydrochloride)
		NVP-BGT226
		NVP-BKM120 (Hydrochloride)
		OMIPALISIB
		PF-04691502
		PILARALISIB
		PX 866
		RECILISIB
		RG7422
		SONOLISIB
		TASELISIB
		VS-5584
		voxtalisib
PIK3R5	Zeller (CRC)	APITOLISIB
		AZD6482
		BEZ235 (Tosylate)
		BUPARLISIB
		CH5132799
		COPANLISIB
		COPANLISIB HYDROCHLORIDE

		DACTOLISIB
		GDC-0941
		GDC-0941 (dimethanesulfonate)
		GEDATOLISIB
		GSK1059615
		NVP-BEZ235 (hydrochloride)
		NVP-BGT226
		NVP-BKM120 (Hydrochloride)
		OMIPALISIB
		PF-04691502
		PILARALISIB
		PX 866
		RECILISIB
		RG7422
		SONOLISIB
		TASELISIB
		VS-5584
		voxtalisib
PSMB4	Hannigan (CRC)	BORTEZOMIB
		CARFILZOMIB
		Carfilzomib (PR-171)
		Carfilzomib-d8
		MARIZOMIB
		MLN9708
		OPROZOMIB
		Oprozomib (ONX 0912)
SPHK2	Zeller (CRC)	ABC294640
		PF-543
		ROMe
		SKI II
		SLC4101431
		compound 27d [PMID: 28231433]
SREBF1	Yu (CRC)	DOCONEXENT
TAF1	Hannigan (CRC)	BAY-299
ULK1	Zeller (CRC)	FOSTAMATINIB
		compound R-16 [PMID: 21967808]