Host-microbiome protein-protein interactions reveal mechanisms in human disease

Authors: Juan Felipe Beltrán[1], Ilana Lauren Brito[1*]

Affiliations

[1] Meinig School of Biomedical Engineering, Cornell University, Ithaca, NY

*Correspondence to: Ilana L. Brito (ibrito@cornell.edu)

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 identify potential mechanisms through which human-associated bacteria impact host health, we leveraged publicly-available interspecies protein-protein interaction (PPI) data to find clusters of microbiome-derived proteins with high sequence identity to known human protein interactors. We observe differential presence of putative human-interacting bacterial genes in metagenomic case-control microbiome studies. In 8 independent case studies, we find evidence that the microbiome broadly targets human immune, oncogenic, apoptotic, and endocrine signaling pathways in relation to IBD, obesity, CRC and T2D diagnoses. This host-centric analysis strategy provides a mechanistic hypothesis-generating platform for any metagenomics cohort study and extensively adds human functional annotation to commensal bacterial proteins.

One-sentence summary

Microbiome-derived proteins are linked to disease-associated human pathways by metagenomic and protein-interaction analyses.

Main Text

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 (*1*) and the lack of functional annotations (*2*) have thwarted efforts to identify the mechanisms by which bacterial genes impact host health. In the cases where functional annotations exist, they tend to reflect the proteins' most granular molecular functions (*e.g.* DNA binding, post-translational modification) rather than their role in biological pathways (*3*) and few, if any, relate to host cell signaling and homeostasis. Associating any commensal bacterial gene and a host pathway has thus far required experimental approaches catered to each gene or gene function (*4*, *5*).

Protein-protein interactions (PPIs) have revealed the mechanisms by which pathogens interact with host tissue through in-depth structural studies of individual proteins (*5–7*), as well as large-scale whole-organism interaction screens (*8*, *9*). We hypothesized that host-microbiome PPIs may underlie health status and could serve to provide additional information, through annotation of human pathways, about the role of bacteria in modulating health. There are already canonical examples of protein-mediated microbe-associated patterns (MAMPs) that directly trigger host-signaling pathways through pattern recognition receptors present on epithelial and immune tissues (*10*), such as flagellin with Toll-like receptor 5 (TLR5). Several recent observations have further underscored a role for commensal-host PPIs in health: a protease secreted by *Enterococcus faecalis* binds incretin hormone glucagon-like peptide 1 (GLP-1), a therapeutic target for type 2 diabetes (T2D) (*11*); and a slew of ubiquitin mimics encoded by both pathogens (*12*) and gut commensals (*13*) play a role in modulating membrane trafficking.

Currently, few experimentally-verified PPIs exist between bacterial and human proteins (roughly 8,000 in the IntAct database (*14*)) and only a handful of these involve proteins pulled from the human gut microbiome. Expanding the commensal-human interaction network through state-of-the-art structural modeling (*15*) is untenable, as there are few sequences homologous to genes found in metagenomes represented in co-crystals from the Protein Data Bank (*16*) (PDB) (Fig. S1, Supplementary Note 1). In the absence of structure and experimental data, sequence identity methods have been used to great effect to infer host-pathogen PPI networks for single pathogens (*17–19*), but such approaches have not yet been applied at the community-level, as would be required for the human gut microbiome.

All pathogen-host interactions are initially implicated in virulence, whereas microbiome-associated disorders tend not to follow Koch's postulates (*20*). To distinguish PPIs that may be associated with health versus disease, we compared community-level PPI profiles in large case-control cohorts of well-established microbiome-associated disorders—namely colorectal cancer (CRC) (*21–24*), T2D (*25*, *26*), inflammatory bowel disease (IBD) (*27*) and obesity (*28*) (Fig. 1A, Table S1). In order to build community-level PPI profiles, we mapped quality-filtered metagenomic sequencing reads from eight studies to a newly constructed database of bacterial human-protein interactors and the bacterial members of their associated UniRef clusters (Fig. S2, Supplementary Methods), which represent homeomorphic protein superfamilies through sequence identity (*29*). Using a normalized feature importance ranking from random forest classifiers trained on each disease cohort (Fig. S3, Supplementary Methods), we find 46,734 commensal bacterial proteins (comprising 579 clusters) associated with disease, by virtue of their putative interactions with 1,145 human proteins.

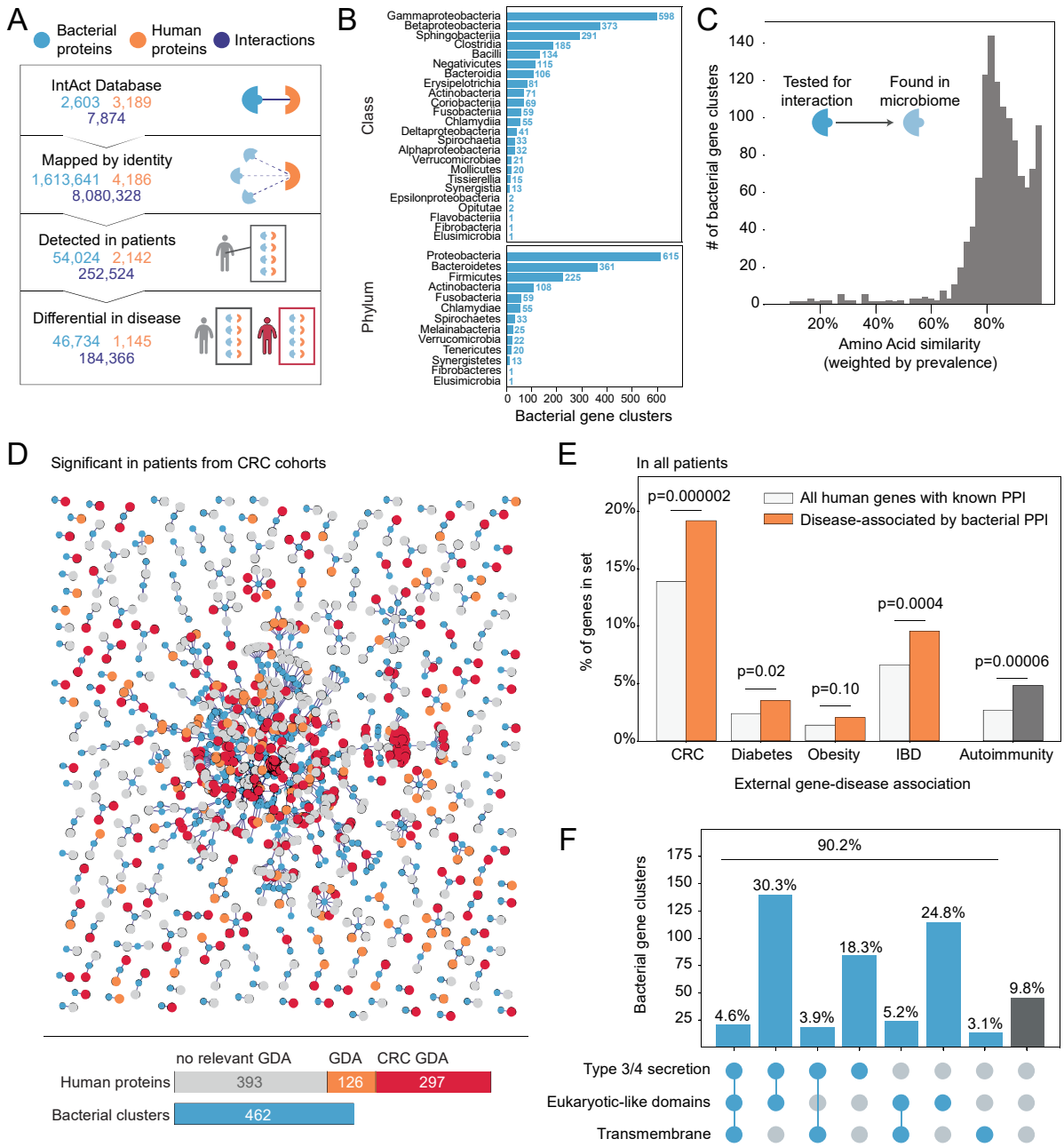


Figure 1. Human proteins differentially targeted by the microbiome in disease are enriched for relevant gene-disease associations. (A) The number of interspecies bacterial proteins (blue), human proteins (orange) and interactions (dark blue) in the IntAct database; those inferred using homology clusters (UniRef); those determined to be present in the gut microbiomes from eight metagenomic studies; and those associated with disease through our metagenomic machine learning approach, comparing prevalence in cases (grey) and control (red). **(B)** The number of bacterial gene clusters that include members from each bacterial phyla (top) and class (bottom). Only genes that were detected in human metagenomes from the eight studies are used in this analysis. Note that most clusters contain proteins from more than one class and phylum (Fig. S4). **(C)** Histograms showing the sequence similarity per bacterial cluster between proteins with experimentally determined human interactions and proteins detected in human microbiomes, normalized according to their prevalence. **(D)** Human proteins implicated in CRC by our method (normalized importance > 0) are plotted with their bacterial interactors (blue), and annotated based on their GDAs in the DisGeNET database to CRC (red) or either diabetes, obesity or IBD (orange). Human proteins without relevant GDAs are colored in gray. **(E)** The proportions of human proteins implicated in disease (normalized importance > 0) compared to all human proteins with experimentally detected PPIs in the IntAct database, according to their GDAs in the DisGeNET database. **(F)** The number of bacterial gene 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).

Interaction does not need to be conserved across homologous proteins in different bacterial species. A key concern is the disproportionate number of bacteria-human PPIs in IntAct derived from high-throughput screens performed on three intracellular pathogens: *Yersinia pestis*, *Francisella tularensis* and *Bacillus anthracis* (*8*). However, we find that patient-detected bacterial clusters are 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 (Fig. 1C and S4). We verified that human microbiome proteins have high amino-acid similarity to experimentally-verified human interactors in the same UniRef cluster (Fig. 1C and S5). Additionally, interspecies bacterial-human protein interface residues, in general, are highly similar, or even identical, between members of the same UniRef cluster (Fig. S6, Supplementary Note 2). Although we appreciate that there will be commensal-human PPIs that are not captured by this approach due to the limited scope of experimental data available, this is the largest and only dataset of microbiome-host associated PPIs.

Surprisingly, the 816 human proteins we associate with CRC via the microbiome contain 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 (Fig. 1D). This represents a larger trend: the 1,145 human interactors are overall enriched for proteins with previously-reported gene-disease associations (GDA) in CRC, T2D, and IBD (Fig. 1E, Table S2), with the exception of obesity, where annotation is generally scarce. In line with mixed etiologies of diseases, we see that GDAs are not disease-cohort specific (Fig. S7). In fact, 36 percent of our genes have more than one GDA for our diseases of interest. We suspected this may extend to autoimmune diseases, which are increasingly studied in the context of the gut microbiome (*30*), and we confirm enrichment of GDAs for autoimmune disorders in the human proteins implicated by our method. This concordance between known disease annotation and disease association through our method increases our confidence that we are capturing relevant molecular heterogeneity underlying microbiome-related disease.

If these bacterial proteins are indeed modulating human health through PPIs, we should expect them to contain signatures of surface localization or secretion. We find that a majority of disease-associated bacterial protein clusters (90.2%) contain proteins that are transmembrane, are secreted by type 3 or type 4 secretion systems, and/or contain eukaryotic-like domains, another marker for secretion (Fig. 1F). The remaining 9.8% may also be adequately localized, but our annotations do not cover all bacterial secretion systems, and it is unclear whether bacterial lysis may result in protein delivery to the host.

One of the major advantages of our work is that through this new interaction network, we vastly improve our ability to annotate host-relevant microbiome functions. 35.8% of our disease-associated bacterial clusters contain no members with annotated microbial pathways in KEGG (Kyoto Encyclopedia of Genes and Genomes) (*31*) (Fig. 2A). Yet, most of these genes can now be annotated according to the pathways of their human targets, obtaining a putative disease-relevant molecular mechanism (Fig. 2B). Interestingly, most of the bacterial clusters with KEGG pathway annotations also gain a secondary human pathway annotation. This dual function is not entirely surprising, as a number of these have orthologs that have been previously identified as bacterial 'moonlighting' proteins, which perform secondary functions in addition to their primary role in the cell (*32*). *Mycoplasma pneumoniae* GroEL and *Streptococcus suis* enolase, a protein involved in glycolysis, bind to both human plasminogen and extra-cellular matrix components (*33*, *34*). *Mycobacterium tuberculosis* DnaK signals to leukocytes causing the release of the chemokines CCL3-5 (*35*). *Streptococcus pyogenes* glyceraldehyde-3-phosphate dehydrogenase (GAPDH), canonically involved in glycolysis, can be shuffled to the cell surface where it plays a role as an adhesin, and can also contribute to human cellular apoptosis (*36*). These examples distinctly illustrate how bacterial housekeeping proteins are used by pathogens to modulate human health. In this study, we uncover commensal proteins that similarly may have 'interspecies moonlighting' functions and appear to be pervasive throughout our indigenous microbiota.

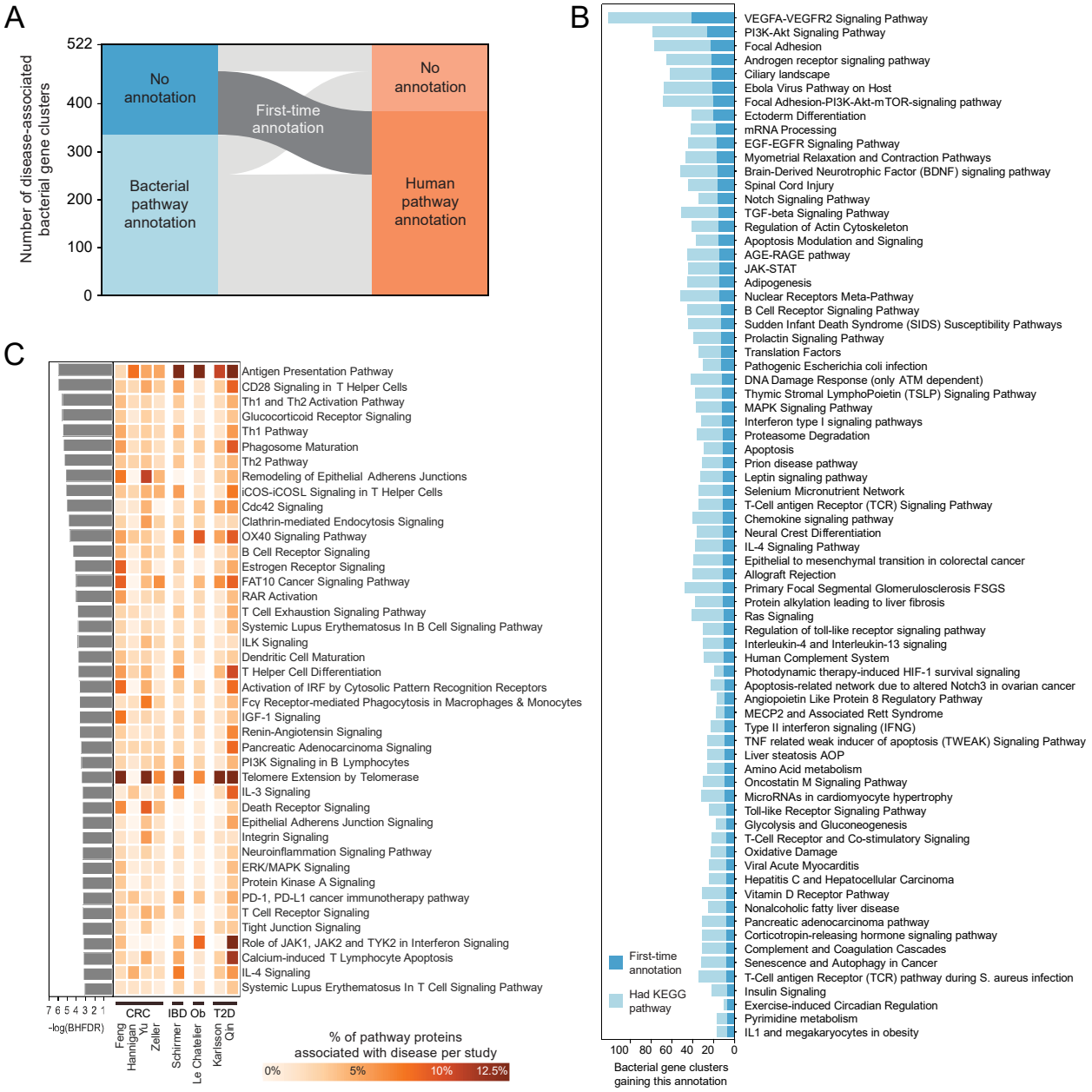


Figure 2. Human pathway annotation can be propagated through interactors to improve bacterial pathway annotation. **(A)** Paired stacked bar plots showing the disease-associated bacterial cluster pathways annotated by KEGG (left) and their inferred pathways according to the human proteins they target (right), as annotated by WikiPathways (*59*). **(B)** Human pathways (annotated using WikiPathways) targeted by bacterial gene clusters detected in human microbiomes from these eight studies. The top 75 human pathways that contribute the most annotations to bacterial clusters detected in the eight metagenomic cohorts that previously lacked KEGG-based annotations are shown. **(C)** Human cellular pathways, enriched in our disease-associated human proteins (with a Benjamini-Hochberg false discovery rate (BHFDR) ≤ 0.05. –log(BHFDR), displayed on the barplot to the left), are colored according to the percent of pathway members differentially targeted in each case-control cohort.

136 In evaluating the statistical significance of recurrent human functional annotations, we performed
137 pathway enrichment analysis on the implicated human proteins and find proteins with established roles in
138 cellular pathways coherent with the pathophysiology of CRC, IBD, obesity and T2D (Fig. 2C), namely
139 those involving immune system, apoptosis, oncogenesis, and endocrine signaling pathways. Though most
140 enriched pathways include human proteins associated with all four diseases, reflecting their associated
141 relative risks (*37–41*), there is heterogeneity in the identity and number of members associated with each
142 study. Far more human proteins from the antigen presentation pathway are associated with T2D, obesity
143 and IBD cohorts' microbiomes than with CRC, perhaps indicating a disease-specific association with this
144 process. We see this again with CRC, in the death receptor signaling pathway and remodeling of
145 epithelial adherens junctions.

146 We see specific examples of known molecular mechanisms for these diseases now implicated with
147 microbiome-host PPIs: We find that DNA fragmentation factor subunit alpha (DFFA) is associated with
148 T2D (in the Qin *et al*. cohort), and is involved in death receptor signaling, an important pathway for the
149 destruction of insulin-producing β-cells (*42*). Collagen alpha-1(I) chain (COL1A1) is also a significant
150 target associated with T2D (in the Karlsson *et al.* cohort), and plays a role in dendritic cell maturation and
151 hepatic fibrosis/hepatic stellate cell activation pathways, capturing known comorbidities between T2D
152 and hepatic steatosis and nonalcoholic steatohepatitis (NASH) (*43*). Proteins associated with CRC
153 spanned expected bacteria-associated pathways, such as the direct sensing of enterotoxins, *e.g.* heat-stable
154 enterotoxin receptor GUCY2C (in the Feng *et al.* and Zeller *et al.* cohorts); but also classical cancer-
155 associated pathways, such as the maintenance of DNA integrity, *e.g.* protection of telomeres protein 1
156 (POT1) (in the Feng *et al.*, Qin *et al.* and Schirmer *et al*. cohorts) and X-ray repair cross-complementing
157 protein 6 (XRCC6) (in the Feng *et al.* and Yu *et al.* cohorts), the latter of which is required for double-
158 strand DNA break repair. Interestingly, actin-related protein 2/3 complex subunit 2 (ARPc2) (associated
159 in the Yu *et al.* and Karlsson *et al.* cohorts) regulates the remodeling of epithelial adherens junctions, a
160 common pathway disrupted in IBD (*44*), CRC (*45*) and, most recently, T2D (*37*). This host-centric
161 annotation is useful beyond large-scale analysis of metagenomic data, as it broadly enables hypothesis-
162 driven research into the protein-mediated mechanisms underlying microbiome impacts on host health.

163 These data suggest a set of discrete protein interactions that induce physiological effects when delivered
164 to the host. Consistent with this idea, we find that indeed many associated human proteins are known drug
165 targets (Table S3). For example, both T2D cohorts' and the obesity cohort's microbiomes independently
166 implicate human protein Rev-ErbA alpha (NR1D1), the target of the drugs GSK4112, SR9009 and
167 SR9011, which inhibit the binding of Rev-ErbA alpha with its natural ligand, heme (Fig. 3A). These
168 drugs have been shown to affect cellular metabolism *in vitro* and affect hyperglycaemia when given to
169 mouse models of metabolic disorder (*46*, *47*). We also find instances where the off-label effects or side
170 effects associated with the drug match our microbiome-driven human protein association. For instance,
171 imatinib mesylate (brand name: Gleevec) has several human binding partners, including macrophage
172 colony-stimulating factor 1 receptor (M-CSF1R) (Fig. 3B), a human protein we associate with CRC (in
173 the Feng *et al.* and Zeller *et al.* cohorts), and platelet-derived growth factor receptor-β (PDGFR-B),
174 associated with obesity and T2D (in the Le Chatelier *et al*. and Qin *et al.* cohorts, respectively). Literature
175 on imatinib supports these findings: although imatinib is best known as a treatment for leukemia, it has
176 been shown to affect glycemic control in patients with T2D (*48*). Furthermore, imatinib can also halt the
177 proliferation of colonic tumor cells and is involved generally in inflammatory pathways, through its
178 inhibition of TNF-alpha production (*49*). Whereas the notion of microbiome-derived metabolites acting as
179 drugs is well-appreciated (*50*, *51*), this work broadens the scope of microbiome-derived drugs to include
180 protein products acting through PPI.

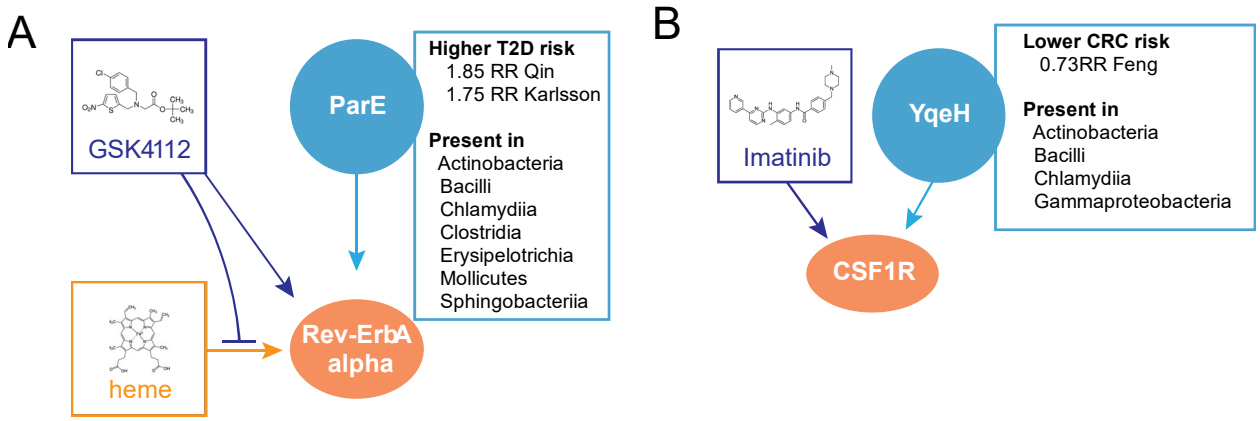


Figure 3. Human proteins targeted by gut commensal proteins include known therapeutic drug targets. (A) RevErbA alpha (NR1D1) binds several human proteins (not shown), DNA (not shown) and heme. GSK4112 competitively binds Rev-ErbA alpha, inhibiting binding with heme. ParE is a microbiome protein present in a diverse range of organisms and has a high relative risk associated with T2D. **(B)** 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.

Here, we reveal for the first time an extensive host-microbiome PPI landscape. To achieve this, we benefit from existing methods in pathogen-host PPI discovery, further informed by community-level PPI profiles of genes differentially detected in human metagenomes. 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. Pinpointing those microbe-derived proteins that interact directly with human proteins will enable the discovery of 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.

References

1. J. Li, H. Jia, X. Cai, H. Zhong, Q. Feng, S. Sunagawa, M. Arumugam, J. R. Kultima, E. Prifti, T. Nielsen, A. S. Juncker, C. Manichanh, B. Chen, W. Zhang, F. Levenez, J. Wang, X. Xu, L. Xiao, S. Liang, D. Zhang, Z. Zhang, W. Chen, H. Zhao, J. Y. Al-Aama, S. Edris, H. Yang, J. Wang, T. Hansen, H. B. Nielsen, S. Brunak, K. Kristiansen, F. Guarner, O. Pedersen, J. Doré, S. D. Ehrlich, M. Consortium, N. Pons, E. L. Chatelier, J.-M. Batto, S. Kennedy, F. Haimet, Y. Winogradski, E. Pelletier, D. LePaslier, F. Artiguenave, T. Bruls, J. Weissenbach, K. Turner, J. Parkhill, M. Antolin, F. Casellas, N. Borruel, E. Varela, A. Torrejon, G. Denariaz, M. Derrien, J. E. T. van H. Vlieg, P. Viega, R. Oozeer, J. Knoll, M. Rescigno, C. Brechot, C. M'Rini, A. Mérieux, T. Yamada, S. Tims, E. G. Zoetendal, M. Kleerebezem, W. M. de Vos, A. Cultrone, M. Leclerc, C. Juste, E. Guedon, C. Delorme, S. Layec, G. Khaci, M. van de Guchte, G. Vandemeulebrouck, A. Jamet, R. Dervyn, N. Sanchez, H. Blottière, E. Maguin, P. Renault, J. Tap, D. R. Mende, P. Bork, J. Wang, An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).

2. R. Joice, K. Yasuda, A. Shafquat, X. C. Morgan, C. Huttenhower, Determining microbial products and identifying molecular targets in the human microbiome. *Cell Metab.* **20**, 731–741 (2014).

3. J. Lloyd-Price, A. Mahurkar, G. Rahnavard, J. Crabtree, J. Orvis, A. B. Hall, A. Brady, H. H. Creasy, C. McCracken, M. G. Giglio, D. McDonald, E. A. Franzosa, R. Knight, O. White, C. Huttenhower, Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature*. **550**, 61–66 (2017).

4. H. Plovier, A. Everard, C. Druart, C. Depommier, M. Van Hul, L. Geurts, J. Chilloux, N. Ottman, T. Duparc, L. Lichtenstein, A. Myridakis, N. M. Delzenne, J. Klievink, A. Bhattacharjee, K. C. H. van der Ark, S. Aalvink, L. O. Martinez, M.-E. Dumas, D. Maiter, A. Loumaye, M. P. Hermans, J.-P. Thissen, C. Belzer, W. M. de Vos, P. D. Cani, A purified membrane protein from *Akkermansia muciniphila* or the pasteurized bacterium improves metabolism in obese and diabetic mice. *Nat. Med.* **23**, 107–113 (2017).

5. D. Nešić, L. Buti, X. Lu, C. E. Stebbins, Structure of the Helicobacter pylori CagA oncoprotein bound to the human tumor suppressor ASPP2. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 1562–1567 (2014).

6. E. Guven-Maiorov, C.-J. Tsai, R. Nussinov, Structural host-microbiota interaction networks. *PLoS Comput. Biol.* **13**, e1005579 (2017).

7. C. Hamiaux, A. van Eerde, C. Parsot, J. Broos, B. W. Dijkstra, Structural mimicry for vinculin activation by IpaA, a virulence factor of Shigella flexneri. *EMBO Rep.* **7**, 794–799 (2006).

8. M. D. Dyer, C. Neff, M. Dufford, C. G. Rivera, D. Shattuck, J. Bassaganya-Riera, T. M. Murali, B. W. Sobral, The human-bacterial pathogen protein interaction networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis. *PloS One*. **5**, e12089 (2010).

9. P. S. Shah, N. Link, G. M. Jang, P. P. Sharp, T. Zhu, D. L. Swaney, J. R. Johnson, J. Von Dollen, H. R. Ramage, L. Satkamp, B. Newton, R. Hüttenhain, M. J. Petit, T. Baum, A. Everitt, O. Laufman, M. Tassetto, M. Shales, E. Stevenson, G. N. Iglesias, L. Shokat, S. Tripathi, V. Balasubramaniam, L. G. Webb, S. Aguirre, A. J. Willsey, A. Garcia-Sastre, K. S. Pollard, S. Cherry, A. V. Gamarnik, I. Marazzi, J. Taunton, A. Fernandez-Sesma, H. J. Bellen, R. Andino, N. J. Krogan, Comparative Flavivirus-Host Protein Interaction Mapping Reveals Mechanisms of Dengue and Zika Virus Pathogenesis. *Cell*. **175**, 1931-1945.e18 (2018).

244 10. A. P. Bhavsar, J. A. Guttman, B. B. Finlay, Manipulation of host-cell pathways by bacterial
245 pathogens. *Nature*. **449**, 827–834 (2007).

246 11. S. L. LeValley, C. Tomaro-Duchesneau, R. A. Britton, Degradation of the Incretin Hormone
247 Glucagon-Like Peptide-1 (GLP-1) by Enterococcus faecalis Metalloprotease GelE. *mSphere*. **5**
248 (2020), doi:10.1128/mSphere.00585-19.

249 12. E. Guven-Maiorov, C.-J. Tsai, B. Ma, R. Nussinov, Prediction of Host-Pathogen Interactions for
250 Helicobacter pylori by Interface Mimicry and Implications to Gastric Cancer. *J. Mol. Biol.* **429**,
251 3925–3941 (2017).

252 13. L. Stewart, J. D M Edgar, G. Blakely, S. Patrick, Antigenic mimicry of ubiquitin by the gut
253 bacterium Bacteroides fragilis: a potential link with autoimmune disease. *Clin. Exp. Immunol.* **194**,
254 153–165 (2018).

255 14. S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G.
256 Chavali, C. Chen, N. del-Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli,
257 S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R. C. Lovering, B. Meldal, A. N.
258 Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B.
259 Roechert, A. Stutz, M. Tognolli, K. van Roey, G. Cesareni, H. Hermjakob, The MIntAct project--
260 IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*
261 **42**, D358-363 (2014).

262 15. E. Guven-Maiorov, C.-J. Tsai, B. Ma, R. Nussinov, Interface-Based Structural Prediction of Novel
263 Host-Pathogen Interactions. *Methods Mol. Biol. Clifton NJ*. **1851**, 317–335 (2019).

264 16. S. K. Burley, H. M. Berman, G. J. Kleywegt, J. L. Markley, H. Nakamura, S. Velankar, Protein
265 Data Bank (PDB): The Single Global Macromolecular Structure Archive. *Methods Mol. Biol.*
266 *Clifton NJ*. **1607**, 627–641 (2017).

267 17. T. Huo, W. Liu, Y. Guo, C. Yang, J. Lin, Z. Rao, Prediction of host - pathogen protein interactions
268 between Mycobacterium tuberculosis and Homo sapiens using sequence motifs. *BMC*
269 *Bioinformatics*. **16**, 100 (2015).

270 18. R. Sen, L. Nayak, R. K. De, A review on host-pathogen interactions: classification and prediction.
271 *Eur. J. Clin. Microbiol. Infect. Dis. Off. Publ. Eur. Soc. Clin. Microbiol.* **35**, 1581–1599 (2016).

272 19. F.-E. Eid, M. ElHefnawi, L. S. Heath, DeNovo: virus-host sequence-based protein-protein
273 interaction prediction. *Bioinforma. Oxf. Engl.* **32**, 1144–1150 (2016).

274 20. A. L. Byrd, J. A. Segre, Adapting Koch's postulates. *Science*. **351**, 224–226 (2016).

275 21. Q. Feng, S. Liang, H. Jia, A. Stadlmayr, L. Tang, Z. Lan, D. Zhang, H. Xia, X. Xu, Z. Jie, L. Su, X.
276 Li, X. Li, J. Li, L. Xiao, U. Huber-Schönauer, D. Niederseer, X. Xu, J. Y. Al-Aama, H. Yang, J.
277 Wang, K. Kristiansen, M. Arumugam, H. Tilg, C. Datz, J. Wang, Gut microbiome development
278 along the colorectal adenoma-carcinoma sequence. *Nat. Commun.* **6**, 6528 (2015).

279 22. G. D. Hannigan, M. B. Duhaime, M. T. Ruffin, C. C. Koumpouras, P. D. Schloss, Diagnostic
280 Potential and Interactive Dynamics of the Colorectal Cancer Virome. *mBio*. **9** (2018),
281 doi:10.1128/mBio.02248-18.

282 23. J. Yu, Q. Feng, S. H. Wong, D. Zhang, Q. Y. Liang, Y. Qin, L. Tang, H. Zhao, J. Stenvang, Y. Li,
283 X. Wang, X. Xu, N. Chen, W. K. K. Wu, J. Al-Aama, H. J. Nielsen, P. Kiilerich, B. A. H. Jensen, T.
284 O. Yau, Z. Lan, H. Jia, J. Li, L. Xiao, T. Y. T. Lam, S. C. Ng, A. S.-L. Cheng, V. W.-S. Wong, F.
285 K. L. Chan, X. Xu, H. Yang, L. Madsen, C. Datz, H. Tilg, J. Wang, N. Brünner, K. Kristiansen, M.
286 Arumugam, J. J.-Y. Sung, J. Wang, Metagenomic analysis of faecal microbiome as a tool towards
287 targeted non-invasive biomarkers for colorectal cancer. *Gut*. **66**, 70–78 (2017).

288 24. G. Zeller, J. Tap, A. Y. Voigt, S. Sunagawa, J. R. Kultima, P. I. Costea, A. Amiot, J. Böhm, F.
289 Brunetti, N. Habermann, R. Hercog, M. Koch, A. Luciani, D. R. Mende, M. A. Schneider, P.
290 Schrotz-King, C. Tournigand, J. Tran Van Nhieu, T. Yamada, J. Zimmermann, V. Benes, M. Kloor,
291 C. M. Ulrich, M. von Knebel Doeberitz, I. Sobhani, P. Bork, Potential of fecal microbiota for early-
292 stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).

293 25. F. H. Karlsson, V. Tremaroli, I. Nookaew, G. Bergström, C. J. Behre, B. Fagerberg, J. Nielsen, F.
294 Bäckhed, Gut metagenome in European women with normal, impaired and diabetic glucose control.
295 *Nature*. **498**, 99–103 (2013).

296 26. J. Qin, Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen, Y. Peng, D.
297 Zhang, Z. Jie, W. Wu, Y. Qin, W. Xue, J. Li, L. Han, D. Lu, P. Wu, Y. Dai, X. Sun, Z. Li, A. Tang,
298 S. Zhong, X. Li, W. Chen, R. Xu, M. Wang, Q. Feng, M. Gong, J. Yu, Y. Zhang, M. Zhang, T.
299 Hansen, G. Sanchez, J. Raes, G. Falony, S. Okuda, M. Almeida, E. LeChatelier, P. Renault, N.
300 Pons, J.-M. Batto, Z. Zhang, H. Chen, R. Yang, W. Zheng, S. Li, H. Yang, J. Wang, S. D. Ehrlich,
301 R. Nielsen, O. Pedersen, K. Kristiansen, J. Wang, A metagenome-wide association study of gut
302 microbiota in type 2 diabetes. *Nature*. **490**, 55–60 (2012).

303 27. M. Schirmer, E. A. Franzosa, J. Lloyd-Price, L. J. McIver, R. Schwager, T. W. Poon, A. N.
304 Ananthakrishnan, E. Andrews, G. Barron, K. Lake, M. Prasad, J. Sauk, B. Stevens, R. G. Wilson, J.
305 Braun, L. A. Denson, S. Kugathasan, D. P. B. McGovern, H. Vlamakis, R. J. Xavier, C.
306 Huttenhower, Dynamics of metatranscription in the inflammatory bowel disease gut microbiome.
307 *Nat. Microbiol.* **3**, 337–346 (2018).

308 28. E. Le Chatelier, T. Nielsen, J. Qin, E. Prifti, F. Hildebrand, G. Falony, M. Almeida, M. Arumugam,
309 J.-M. Batto, S. Kennedy, P. Leonard, J. Li, K. Burgdorf, N. Grarup, T. Jørgensen, I. Brandslund, H.
310 B. Nielsen, A. S. Juncker, M. Bertalan, F. Levenez, N. Pons, S. Rasmussen, S. Sunagawa, J. Tap, S.
311 Tims, E. G. Zoetendal, S. Brunak, K. Clément, J. Doré, M. Kleerebezem, K. Kristiansen, P. Renault,
312 T. Sicheritz-Ponten, W. M. de Vos, J.-D. Zucker, J. Raes, T. Hansen, MetaHIT consortium, P. Bork,
313 J. Wang, S. D. Ehrlich, O. Pedersen, Richness of human gut microbiome correlates with metabolic
314 markers. *Nature*. **500**, 541–546 (2013).

315 29. C. H. Wu, A. Nikolskaya, H. Huang, L.-S. L. Yeh, D. A. Natale, C. R. Vinayaka, Z.-Z. Hu, R.
316 Mazumder, S. Kumar, P. Kourtesis, R. S. Ledley, B. E. Suzek, L. Arminski, Y. Chen, J. Zhang, J. L.
317 Cardenas, S. Chung, J. Castro-Alvear, G. Dinkov, W. C. Barker, PIRSF: family classification
318 system at the Protein Information Resource. *Nucleic Acids Res.* **32**, D112-114 (2004).

319 30. E. Gianchecchi, A. Fierabracci, Recent Advances on Microbiota Involvement in the Pathogenesis of
320 Autoimmunity. *Int. J. Mol. Sci.* **20** (2019), doi:10.3390/ijms20020283.

321 31. M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: new perspectives on
322 genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).

323 32. B. Henderson, An overview of protein moonlighting in bacterial infection. *Biochem. Soc. Trans.* **42**,
324 1720–1727 (2014).

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

328 34. B. Henderson, A. Martin, Bacterial moonlighting proteins and bacterial virulence. *Curr. Top.*
329 *Microbiol. Immunol.* **358**, 155–213 (2013).

330 35. T. Lehner, L. A. Bergmeier, Y. Wang, L. Tao, M. Sing, R. Spallek, R. van der Zee, Heat shock
331 proteins generate beta-chemokines which function as innate adjuvants enhancing adaptive
332 immunity. *Eur. J. Immunol.* **30**, 594–603 (2000).

333 36. K. A. Seidler, N. W. Seidler, Role of extracellular GAPDH in Streptococcus pyogenes virulence.
334 *Mo. Med.* **110**, 236–240 (2013).

335 37. E. A. Kang, K. Han, J. Chun, H. Soh, S. Park, J. P. Im, J. S. Kim, Increased Risk of Diabetes in
336 Inflammatory Bowel Disease Patients: A Nationwide Population-based Study in Korea. *J. Clin.*
337 *Med.* **8** (2019), doi:10.3390/jcm8030343.

338 38. A. Jurjus, A. Eid, S. Al Kattar, M. N. Zeenny, A. Gerges-Geagea, H. Haydar, A. Hilal, D. Oueidat,
339 M. Matar, J. Tawilah, I. H. Hussein, P. Schembri-Wismayer, F. Cappello, G. Tomasello, A. Leone,
340 R. A. Jurjus, Inflammatory bowel disease, colorectal cancer and type 2 diabetes mellitus: The links.
341 *BBA Clin.* **5**, 16–24 (2016).

342 39. T. Jess, B. W. Jensen, M. Andersson, M. Villumsen, K. H. Allin, Inflammatory Bowel Disease
343 Increases Risk of Type 2 Diabetes in a Nationwide Cohort Study. *Clin. Gastroenterol. Hepatol. Off.*
344 *Clin. Pract. J. Am. Gastroenterol. Assoc.* (2019), doi:10.1016/j.cgh.2019.07.052.

345 40. R. W. Stidham, P. D. R. Higgins, Colorectal Cancer in Inflammatory Bowel Disease. *Clin. Colon*
346 *Rectal Surg.* **31**, 168–178 (2018).

347 41. S. de Kort, A. A. M. Masclee, S. Sanduleanu, M. P. Weijenberg, M. P. P. van Herk-Sukel, N. J. J.
348 Oldenhof, J. P. W. van den Bergh, H. R. Haak, M. L. Janssen-Heijnen, Higher risk of colorectal
349 cancer in patients with newly diagnosed diabetes mellitus before the age of colorectal cancer
350 screening initiation. *Sci. Rep.* **7**, 46527 (2017).

351 42. C. Sia, A. Hänninen, Apoptosis in autoimmune diabetes: the fate of beta-cells in the cleft between
352 life and death. *Rev. Diabet. Stud. RDS*. **3**, 39–46 (2006).

353 43. J. Richard, I. Lingvay, Hepatic steatosis and Type 2 diabetes: current and future treatment
354 considerations. *Expert Rev. Cardiovasc. Ther.* **9**, 321–328 (2011).

355 44. A. Franke, T. Balschun, T. H. Karlsen, J. Sventoraityte, S. Nikolaus, G. Mayr, F. S. Domingues, M.
356 Albrecht, M. Nothnagel, D. Ellinghaus, C. Sina, C. M. Onnie, R. K. Weersma, P. C. F. Stokkers, C.
357 Wijmenga, M. Gazouli, D. Strachan, W. L. McArdle, S. Vermeire, P. Rutgeerts, P. Rosenstiel, M.
358 Krawczak, M. H. Vatn, IBSEN study group, C. G. Mathew, S. Schreiber, Sequence variants in
359 IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. *Nat. Genet.* **40**,
360 1319–1323 (2008).

361 45. A. C. Daulagala, M. C. Bridges, A. Kourtidis, E-cadherin Beyond Structure: A Signaling Hub in
362 Colon Homeostasis and Disease. *Int. J. Mol. Sci.* **20** (2019), doi:10.3390/ijms20112756.

363 46. E. Vieira, L. Marroquí, A. L. C. Figueroa, B. Merino, R. Fernandez-Ruiz, A. Nadal, T. P. Burris, R.
364 Gomis, I. Quesada, Involvement of the clock gene Rev-erb alpha in the regulation of glucagon
365 secretion in pancreatic alpha-cells. *PloS One*. **8**, e69939 (2013).

366 47. L. A. Solt, Y. Wang, S. Banerjee, T. Hughes, D. J. Kojetin, T. Lundasen, Y. Shin, J. Liu, M. D.
367 Cameron, R. Noel, S.-H. Yoo, J. S. Takahashi, A. A. Butler, T. M. Kamenecka, T. P. Burris,
368 Regulation of circadian behaviour and metabolism by synthetic REV-ERB agonists. *Nature*. **485**,
369 62–68 (2012).

370 48. S.-S. Choi, E.-S. Kim, J.-E. Jung, D. P. Marciano, A. Jo, J. Y. Koo, S. Y. Choi, Y. R. Yang, H.-J.
371 Jang, E.-K. Kim, J. Park, H. M. Kwon, I. H. Lee, S. B. Park, K.-J. Myung, P.-G. Suh, P. R. Griffin,
372 J. H. Choi, PPARγ Antagonist Gleevec Improves Insulin Sensitivity and Promotes the Browning of
373 White Adipose Tissue. *Diabetes*. **65**, 829–839 (2016).

374 49. A. M. Wolf, D. Wolf, H. Rumpold, S. Ludwiczek, B. Enrich, G. Gastl, G. Weiss, H. Tilg, The
375 kinase inhibitor imatinib mesylate inhibits TNF-{alpha} production in vitro and prevents TNF-
376 dependent acute hepatic inflammation. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 13622–13627 (2005).

377 50. M. S. Donia, M. A. Fischbach, Small molecules from the human microbiota. *Science*. **349**, 1254766
378 (2015).

379 51. M. R. Wilson, Y. Jiang, P. W. Villalta, A. Stornetta, P. D. Boudreau, A. Carrá, C. A. Brennan, E.
380 Chun, L. Ngo, L. D. Samson, B. P. Engelward, W. S. Garrett, S. Balbo, E. P. Balskus, The human
381 gut bacterial genotoxin colibactin alkylates DNA. *Science*. **363** (2019),
382 doi:10.1126/science.aar7785.

383 **Acknowledgments**

384 We wish to acknowledge members of the Brito lab, Indrayudh Ghosal, Giles Hooker and Andy Clark for
385 their thoughtful comments. Funding: Ilana Brito is a Pew Scholar in Biomedical Sciences, a Packard
386 Foundation Fellow, recipient of a Packard Foundation Fellowship, and a Sloan Foundation Research
387 Fellow. Author contributions: J.F.B and I.L.B. conceptualized and designed the study and co-wrote the
388 manuscript. Competing Interests: Provisional patents have been filed for both the process and
389 therapeutic/diagnostic protein candidates found herein by Cornell University. Inventors: I.L.B. and J.F.B.
390 Data and materials availability: All data used for this paper is publicly available and described in
391 Supplemental Table 1. Processed detection matrices and code to calculate normalized feature importance
392 from random forest are available as supplementary material.

Methods

Building a putative bacteria-human protein-protein interaction (PPI) network

Interactions were downloaded from the IntAct database (*14*) [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 UniRef50 sequence clusters (*52*). UniRef50 Clusters are calculated every week and are publicly available through the UniProt web service. UniRef50 clusters are named after their seed sequence, which has at least 50% sequence identity to all other members in the cluster. Additionally, all members in the cluster have at least 80% sequence identity to the seed sequence. Given a UniRef cluster with an experimentally determined PPI with a human protein, all bacterial members of the cluster are labeled as putative interactors. Human proteins that have not been verified by the SwissProt curating platform are filtered out of the final interaction network. 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 eight metagenomic studies (Table S1) 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 (*53*) 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.

Identity, similarity, and conservation measurements

The sequence identity constraints imposed from a UniRef cluster's seed on all other member sequences don't explicitly provide any information about the sequence identity, or similarity, between other pairs of sequences in the cluster. We compute pairwise alignments in order to understand how appropriate our annotation mapping is between proteins experimentally-verified to interact with human proteins and bacterial members of the same UniRef cluster that were detected in metagenomic samples. Experimentally-verified interactors are aligned to their metagenome-detected UniRef cluster members using the Smith-Waterman local alignment algorithm with a BLOSUM62 matrix via python's parasail (*54*) library (v.1.1.17). Amino acid identity is calculated as the number of identical matches in the pairwise alignment, divided by the length of the experimentally-verified interactor. Amino acid similarity is likewise calculated as the number of matches in the pairwise alignment that represent frequent substitutions (non-negative BLOSUM62 scores), divided by the length of the experimentally-verified interactor.

Each cluster has a different number of bacterial members, and thus, comparisons, so we need to summarize the bacterial identity and similarity metrics per cluster. We represent the identity between experimentally-verified and metagenomic-detected bacterial protein sequences for each cluster as mean, median, or a weighted average (Fig. S5). Specifically, we calculate the weighted average of identity/similarity as:

$$weighted\%match(cluster) = \sum_{member \in cluster} \frac{Prevalence_{member} \times \%match}{Prevalence_{cluster}}$$

Where $Prevalence_{member}$ is the percent of patients where the bacterial sequence was detected, $Prevalence_{cluster}$ is the percent of patients where any bacterial sequence from the cluster was detected, and %match is either the identity or similarity between the member and the experimentally-verified protein interactor.

When necessary, we constructed multiple sequence alignments using only the experimentally-verified interactor sequence and all the metagenome-detected members of its homology cluster in order to quantify amino acid conservation at each site. We calculated the Jensen-Shannon divergence using the code provided by Capra *et al*. (*55*) with a window size of 3.

Prioritization of disease-associated bacterial protein clusters and human targets

In order to identify heterogeneity in the prevalence of bacteria-human PPIs, we preprocessed the data into two detection matrices. Each patient from each study is represented in two feature spaces: (a) a binary vector of detected bacterial gene clusters or (b) a binary vector of putatively targeted human proteins. Human proteins were considered redundant if they shared all the same bacterial protein partners in our database, as their "detection" is, by definition, perfectly correlated in this design, and were treated as a single feature. Additionally, we build a contingency table based on the case/control balance of the dataset and the prevalence of each bacterial gene cluster or human protein. Features with an expected count of less than 5 in any cell of the contingency table were also filtered out as being under- or over-detected.

We use these processed matrices to train a random forest machine learning classifier on the task of separating case and control patients and, after verifying that they achieve reasonable performance on the task using leave-one-out cross-validation (Fig. S3), we extract the feature importance from the classifiers. Having used the scikit-learn (*56*) implementation of the random forest algorithm, feature importance corresponds to the average Gini impurity of the feature in all splits that it was involved in. Gini feature importance is a powerful prioritization tool, as it can capture the multivariate feature importance (whereas simple metrics like log-odds ratio and corrected chi-squared statistics only capture univariate feature

importance). However, it has been noted that in sparse, binary decision tasks like our own (*57*, *58*), this feature importance can be can overestimate the importance of features based on their prevalence alone.

To obtain a normalized Gini feature importance, we perform a Monte Carlo estimate of the expected Gini importance for each feature given the prevalence of all features in that dataset. On each iteration of the simulation, we generate a random null feature matrix using a Bernoulli binary generator where:

$$P(null[patient][feature] = 1) \; = \; Prevalence(feature)$$

We train two random forests on the disease labels for each patient, using either the real matrix or the null matrix. Both real and null Gini feature importances are extracted for each feature and aggregated across iterations of the simulation. The normalized Gini importance for each feature is expressed as a z-score:

$$normGini_f = \frac{\overline{Gini_f^{real}} - \overline{Gini_f^{null}}}{\max(\sigma(Gini_f^{real}), \sigma(Gini_f^{null}))}$$

The simulation is repeated until the distance between the maximum and minimum normalized Gini importance converges (at least 200 iterations of holding equal value). Code and preprocessed detection matrices for each of the studies are provided in the Auxiliary Supplementary Materials.

This iterative procedure is a convenient way to generate a null feature importance for comparison, but also provides a very robust measurement of $Gini^{real}$. Random trees from a random forest act as independent estimators: Given the same data, calculating the average importance of N forests with E estimators is equivalent to the importance on a single forest with N × E estimators. For the real Gini feature importance calculation, our final estimates are equivalent to a forest with I × E trees, where I is the number of iterations at convergence. The entire procedure is analogous to iteratively increasing the number of trained trees in a random forest (and its paired null model) with a step-size E (in our case, E=100) until normalized feature importance converges. An example is provided in the Auxiliary Supplementary Materials.

Most of the normalized feature importances across studies fall at or below zero, indicating that their Gini feature importance is not higher than would be expected in the null model (Fig. S8). This provides a convenient cutoff (normalized Gini > 0) to prioritize a set of proteins, as human proteins with positive normalized Gini importance capture proteins with large log-odds ratio magnitudes and rescues candidates that would've been missed through univariate analysis.

Human pathway annotation and enrichment analysis

Human pathway annotation was performed using the mygene python library. Specifically, we queried pathway annotations from WikiPathways (*59*), filtering out pathways from TarBase, as they specifically only include miRNA interaction annotation.

We performed pathway enrichment analysis using QIAGEN's Ingenuity Pathway Analysis (IPA) (*60*) tool. All human proteins with a normalized feature importance greater than zero were uploaded as UniProt identifiers into the desktop interface and submitted to their webserver for Core Enrichment Analysis was conducted only on human tissue and cell lines and 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%.

We additionally annotated all human proteins with any known drug targets from the probes-and-drugs database (*61*) (04.2019 database dump), which aggregates drug-target interactions from the largest drug-target databases.

558 **Human gene-disease association**
559 Disease annotations were extracted from all of GDAs from DisGeNET (*62*) (v.6.0). Lacking a simple
560 hierarchy of disease, we binned similar disease terms into the 5 larger categories relevant to our study.
561 Human protein identifiers were mapped to their Entrez gene ID's using the UniProt batch mapping
562 resource and then annotated with these 5 labels:
563
564 CRC: Adenocarcinoma of large intestine, Hereditary non-polyposis colorectal cancer syndrome,
565 Hereditary nonpolyposis colorectal carcinoma, Malignant neoplasm of colon stage IV, Malignant
566 neoplasm of sigmoid colon, Malignant tumor of colon, Microsatellite instability-high colorectal
567 cancer,
568
569 Diabetes: Brittle diabetes, Familial central diabetes insipidus, Fibrocalculous pancreatic diabetes,
570 Gastroparesis due to diabetes mellitus, Insulin resistance in diabetes, Insulin-dependent but
571 ketosis-resistant diabetes, Insulin-dependent diabetes mellitus secretory diarrhea syndrome,
572 Insulin-resistant diabetes mellitus, Insulin-resistant diabetes mellitus at puberty, Latent
573 autoimmune diabetes mellitus in adult, Macroalbuminuric diabetic nephropathy, Maturity onset
574 diabetes mellitus in young, Maturity-onset diabetes of the young, type 10, Maturity-onset diabetes
575 of the young, type 11, Microalbuminuric diabetic nephropathy, Moderate nonproliferative
576 diabetic retinopathy, Monogenic diabetes, Neonatal diabetes mellitus, Neonatal insulin-dependent
577 diabetes mellitus, Non-insulin-dependent diabetes mellitus with unspecified complications,
578 Nonproliferative diabetic retinopathy, Other specified diabetes mellitus, Other specified diabetes
579 mellitus with unspecified complications, Pancreatic disorders (not diabetes), Partial nephrogenic
580 diabetes insipidus, Prediabetes syndrome, Proliferative diabetic retinopathy, Renal cysts and
581 diabetes syndrome, Severe nonproliferative diabetic retinopathy, Transient neonatal diabetes
582 mellitus, Type 2 diabetes mellitus in nonobese, Type 2 diabetes mellitus in obese, Type 2
583 diabetes mellitus with acanthosis nigricans, Visually threatening diabetic retinopathy, diabetes
584 (mellitus) due to autoimmune process, diabetes (mellitus) due to immune mediated pancreatic
585 islet beta-cell destruction, diabetes mellitus risk, idiopathic diabetes (mellitus), postprocedural
586 diabetes mellitus, secondary diabetes mellitus NEC
587
588 Autoimmune: Addison's disease due to autoimmunity, Adult form of celiac disease, Aneurysm of
589 celiac artery, Ankylosing spondylitis, Ankylosing spondylitis and other inflammatory
590 spondylopathies, Arteriovenous fistulas of celiac and mesenteric vessels, Blood autoimmune
591 disorders, Bullous systemic lupus erythematosus, Chilblain lupus 1, Dianzani autoimmune
592 lymphoproliferative syndrome, Dilatation of celiac artery, Hyperthyroidism, Nonautoimmune,
593 Latent autoimmune diabetes mellitus in adult, Maternal autoimmune disease, Multiple sclerosis in
594 children, Neonatal Systemic lupus erythematosus, Subacute cutaneous lupus, Systemic lupus
595 erythematosus encephalitis, Venous varicosities of celiac and mesenteric vessels, Warm
596 autoimmune hemolytic anemia, diabetes (mellitus) due to autoimmune process, lupus cutaneous,
597 lupus erythematodes
598
599 Obesity: Abdominal obesity metabolic syndrome, Adult-onset obesity, Aplasia/Hypoplasia of the
600 earlobes, Childhood-onset truncal obesity, Constitutional obesity, Familial obesity, Generalized
601 obesity, Gross obesity, Hyperplastic obesity, Hypertrophic obesity, Hypoplastic olfactory lobes,
602 Hypothalamic obesity, Moderate obesity, Overweight and obesity, Overweight or obesity,
603 Prominent globes, Simple obesity, Type 2 diabetes mellitus in nonobese, Type 2 diabetes mellitus
604 in obese
605
606 IBD: Acute and chronic colitis, Acute colitis, Allergic colitis, Amebic colitis, Chronic colitis,
607 Chronic ulcerative colitis, Crohn Disease, Crohn's disease of large bowel, Crohn's disease of the
608 ileum, Cytomegaloviral colitis, Distal colitis, Enterocolitis, Enterocolitis infectious, Eosinophilic

609 colitis, Food-protein induced enterocolitis syndrome, Hemorrhagic colitis, IIeocolitis, Infectious
610 colitis, Left sided colitis, Necrotizing Enterocolitis, Necrotizing enterocolitis in fetus OR
611 newborn, Neonatal necrotizing enterocolitis, Non-specific colitis, Pancolitis, Pediatric Crohn's
612 disease, Pediatric ulcerative colitis, Perianal Crohn's disease, Typhlocolitis, Ulcerative colitis in
613 remission, Ulcerative colitis quiescent
614
615 We additionally downloaded all human proteins involved in protein-protein interactions from the IntAct
616 database and annotated them in the same manner in order to compare label frequencies.
617
618 **Bacterial pathway, secretion, and taxonomy annotation**
619 We submitted all bacterial protein sequences that were detected in human metagenomes to the
620 KofamKOALA (*63*) KEGG orthology search resource. We additionally submitted our bacterial sequences
621 to EffectiveDB (*64*) in order to obtain predictions for EffectiveT3 (type 3 secretion based on signal
622 peptide), T4SEpre (type 4 secretion based on composition in C-terminus), EffectiveCCBD (type 3
623 secretion based on chaperone binding sites), and EffectiveELD (predicts secretion based on eukaryotic-
624 like domains). We used the single default cutoffs for T4SEpre, EffectiveCCBD, and EffectiveELD, and
625 chose the 'sensitive' cutoff (0.95) rather than the 'selective' (0.9999) cutoff for EffectiveT3.
626 Transmembrane proteins or signal peptides were predicted using TMHMM (*65*) (v.2.0c), with a threshold
627 of 19 or more expected number of amino acids in transmembrane helices.
628
629 Bacterial taxonomy information was extracted from NCBI. UniProt identifiers and annotations were
630 downloaded using UniProt SPARQL endpoint.
631
632 **Statistics**
633 For Fig. 1D, we quantify the difference between the human proteins implicated in disease by our method
634 (StudySet) and all human proteins that have available protein interaction information (NullSet) by
635 comparing the proportion of these sets that have certain gene-disease associations. To do so we perform a
636 chi squared test (dof=1): The total number of proteins in these sets is 13,698 (NullSet) and 767
637 (StudySet). The breakdown of chi squared statistics and p-values can be found in Supplementary Table 6.

638 **Supplementary Notes**

639 **Supplementary Note 1. Structural data available for these microbiome-human PPIs**

640 Interaction network studies have increasingly moved towards structural interaction networks (*66*). These
641 networks represent not only the group of binary PPIs that have been detected, but also the partner-specific
642 interfaces on which these interactions occur. In the absence of resolved structural data for a given
643 bacterial-human PPI, structural PPI data of homologous proteins can be used to identify potential protein
644 interfaces.

645 We measured the extent to which structural interfaces could be used to infer gut commensal-human
646 protein-protein interaction by using DIAMOND (*64*) to query all amino acid sequences submitted to the
647 PDB for any templates that might match bacterial or human proteins in our putative interactor library. Out
648 of the 732 bacterial gene clusters that contain both members with experimentally-verified PPIs and were
649 detected in human gut metagenomic sequences, 596 have BLASTP matches to a sequence in the PDB. A
650 low-quality filter for at least 50% identity and 50% query coverage further lowers this set to 478 bacterial
651 gene clusters. The same process and cutoffs detect PDB matches for 837 of the 2,140 human proteins in
652 our interaction network. The overlap of these two sets reveals 20 cocrystal structures that can provide
653 interface information for only 18 protein pairs including 15 bacterial gene-cluster proteins and 8 human
654 proteins (Fig. S1).

655 In order to identify interface residues between each pair of chains in the 20 cocrystal structures, we first
656 use NACCESS (*67*) to calculate the solvent accessibility of each residue in each chain. Chains with an
657 accessible surface area of 15 Å or more are considered surface residues. We then calculate the change in
658 accessible surface area for each residue when other chains in the same crystal structures are introduced.
659 Residues which have a change in solvent accessible surface area above 1 Å are determined to be interface
660 residues (*68*).

661 While we identify interface residues in all 18 protein pairs (Table S4), 12 of these cases involve large
662 complexes where the human protein and bacterial protein match domains on more than one chain, and
663 sometimes the same chain (Table S5). Determining interface residues for two proteins with multiple
664 matches can complicate analysis, as they can result in multiple interfaces for the same protein partner. For
665 example, in PDB 2b3y, both the human protein IREB2 and the bacterial proteins from the Aconitate
666 hydratase cluster match domains in chains A and B. This would cause IREB2's interface residues to
667 contain interface residues from two sources in the same crystal structure. There are, however, 6 cases in
668 which the human protein and bacterial proteins match their respective chains exclusively. We highlight
669 one example in which there are uniquely mapped chains, where 1p0s chains H and E match human
670 coagulation factor X and bacterial Ecotin, respectively (Fig. S10). Through this analysis, we demonstrate
671 the power of sequence homology searches in structural databases to confirm bacteria-human PPIs and
672 characterize their interfaces, but find that there are currently not enough representative sequences to do
673 structural prediction at a large scale for the commensal human microbiome.

674 **Supplementary Note 2. Conservation of interface residues in bacterial members of UniRef50**
675 **Clusters with human interactors in the PDB**

676 Functional annotations are commonly propagated between members of the same UniRef50 cluster (*52*,
677 *69*), yet it is not clear whether this intra-cluster conservation of function applies to exogenous interaction.
678 To validate whether this is generally the case, we analyzed the conservation of interface residues across
679 bacterial members of the same UniRef cluster for all bacteria-human protein-protein interactions
680 submitted to the Protein Data Bank (PDB).

681 Using UniProt's SPARQL API, we compiled a list of all PDB structures which contain both human
682 proteins and bacterial proteins (751 structures as of January 2020), the UniRef50 cluster identifier for the
683 bacterial protein, and all protein sequences in the corresponding cluster that also originate from bacterial

684 proteomes (68,434 unique bacterial protein sequences as of January 2020). Using Clustal Omega, we then
685 generated multiple sequence alignments for all the members of each UniRef50 clusters, excluding any
686 duplicated sequences. We calculated interface residues on all pairs of chains in each of the 751 structures
687 and measured the BLOSUM62 similarity between bacterial interface residues and their corresponding
688 amino acids in their respective UniRef50 cluster MSA.

689

690 Despite the small number of PPIs in our dataset that have representatives in the PDB (Fig. S1), examining
691 bacteria-human PPI co-crystal structures supports transfer of interaction among UniRef50 cluster
692 members. We find that there is high amino acid sequence identity and similarity between interface
693 residues in bacteria-human cocrystal structures and other bacterial members of the same cluster (Fig.S6).
694 We additionally calculate the Jensen-Shannon divergence on the columns of the MSA containing
695 interface residues and find that they are well-conserved (Fig.S6). Overall, we find evidence that interface
696 residues with a human protein interactor tend to be maintained between bacterial members of the same
697 UniRef50 cluster.

698 **Supplementary References**

699 52. B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, UniProt Consortium, UniRef clusters:
700 a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinforma.*
701 *Oxf. Engl.* **31**, 926–932 (2015).

702 53. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat.*
703 *Methods*. **12**, 59–60 (2015).

704 54. J. Daily, Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments.
705 *BMC Bioinformatics*. **17**, 81 (2016).

706 55. J. A. Capra, M. Singh, Predicting functionally important residues from sequence conservation.
707 *Bioinforma. Oxf. Engl.* **23**, 1875–1882 (2007).

708 56. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P.
709 Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M.
710 Perrot, É. Duchesnay, *J. Mach. Learn. Res.*, in press.

711 57. M. B. Kursa, W. R. Rudnicki, Feature Selection with the Boruta Package. *J. Stat. Softw.* **36**, 1–13
712 (2010).

713 58. A. Altmann, L. Tolosi, O. Sander, T. Lengauer, Permutation importance: a corrected feature
714 importance measure. *Bioinformatics*. **26**, 1340–1347 (2010).

715 59. D. N. Slenter, M. Kutmon, K. Hanspers, A. Riutta, J. Windsor, N. Nunes, J. Mélius, E. Cirillo, S. L.
716 Coort, D. Digles, F. Ehrhart, P. Giesbertz, M. Kalafati, M. Martens, R. Miller, K. Nishida, L.
717 Rieswijk, A. Waagmeester, L. M. T. Eijssen, C. T. Evelo, A. R. Pico, E. L. Willighagen,
718 WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research.
719 *Nucleic Acids Res.* **46**, D661–D667 (2018).

720 60. Data were analyzed through the use of IPA (QIAGEN Inc.,
721 https://www.qiagenbioinformatics.com/products/ingenuitypathway-analysis).

722 61. C. Skuta, M. Popr, T. Muller, J. Jindrich, M. Kahle, D. Sedlak, D. Svozil, P. Bartunek, Probes &
723 Drugs portal: an interactive, open data resource for chemical biology. *Nat. Methods*. **14**, 759–760
724 (2017).

725 62. J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J.
726 García-García, F. Sanz, L. I. Furlong, DisGeNET: a comprehensive platform integrating
727 information on human disease-associated genes and variants. *Nucleic Acids Res.* **45**, D833–D839
728 (2017).

729 63. T. Aramaki, R. Blanc-Mathieu, H. Endo, K. Ohkubo, M. Kanehisa, S. Goto, H. Ogata,
730 KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold.
731 *Bioinforma. Oxf. Engl.* (2019), doi:10.1093/bioinformatics/btz859.

732 64. V. Eichinger, T. Nussbaumer, A. Platzer, M.-A. Jehl, R. Arnold, T. Rattei, EffectiveDB--updates
733 and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI
734 secretion systems. *Nucleic Acids Res.* **44**, D669-674 (2016).

735 65. A. Krogh, B. Larsson, G. von Heijne, E. L. Sonnhammer, Predicting transmembrane protein
736 topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–
737 580 (2001).

738 66. R. Arnold, K. Boonen, M. G. F. Sun, P. M. Kim, Computational analysis of interactomes: Current
739 and future perspectives for bioinformatics approaches to model the host–pathogen interaction space.
740 *Methods*. **57**, 508–518 (2012).

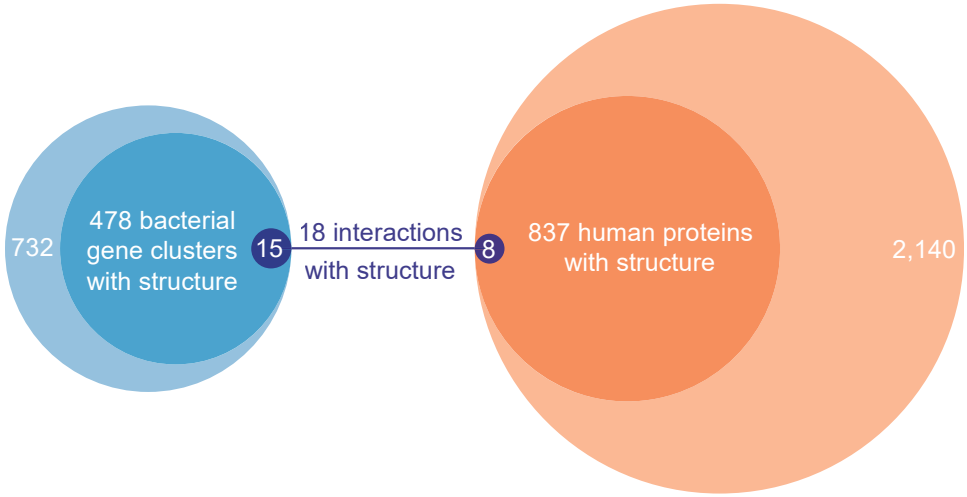
741 67. S. J. Hubbard, J. M. Thornton, *Naccess* (Computer Program, Department of Biochemistry and
742 Molecular Biology, University College London, 1993).

743 68. M. J. Meyer, J. F. Beltrán, S. Liang, R. Fragoza, A. Rumack, J. Liang, X. Wei, H. Yu, Interactome
744 INSIDER: a structural interactome browser for genomic studies. *Nat. Methods*. **15**, 107–114 (2018).

745 69. B. T. Sherman, D. W. Huang, Q. Tan, Y. Guo, S. Bour, D. Liu, R. Stephens, M. W. Baseler, H. C.
746 Lane, R. A. Lempicki, DAVID Knowledgebase: a gene-centered database integrating heterogeneous
747 gene annotation resources to facilitate high-throughput gene functional analysis. *BMC*
748 *Bioinformatics*. **8**, 426 (2007).

Supplementary Figures

Figure S1. Few bacterial-human interaction sequences populate the Protein Data Bank.
A Venn diagram describing the number of detected bacterial clusters and human proteins in the eight metagenomic cohorts that have any matching structure (using BLASTp) in the PDB and whether their structures appear on the same PDB cocrystal structure.



755 **Figure S2. An outline of our homology mapping procedure and alignment.**
756 Depiction of the interaction network inference and protein detection pipeline. Note that only bacterial proteins found to be
757 human-interactors through the mapping procedure are used as candidates for detection in metagenomic studies.

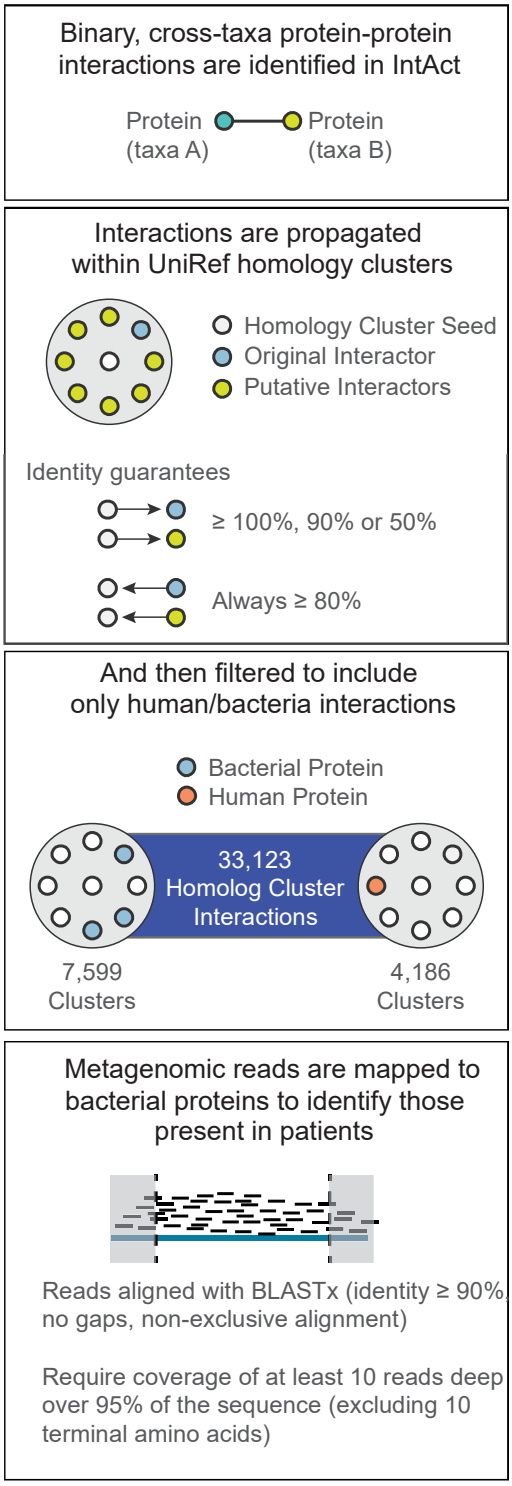
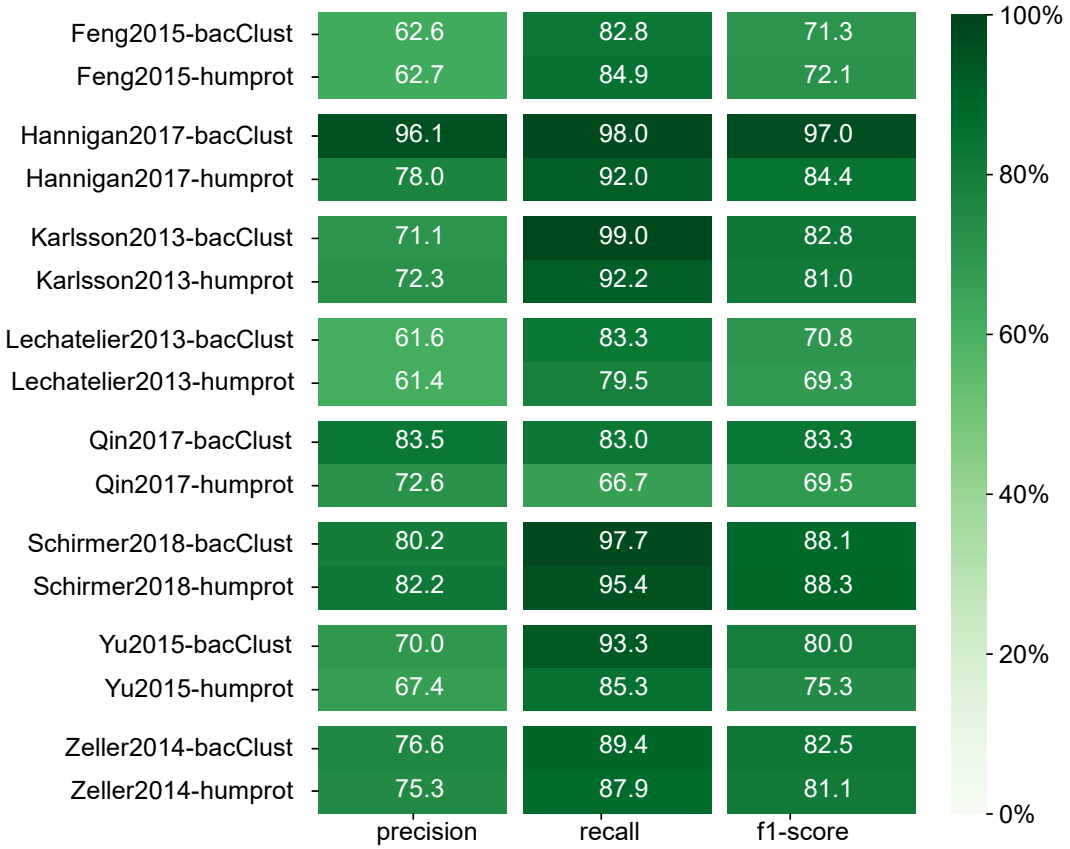
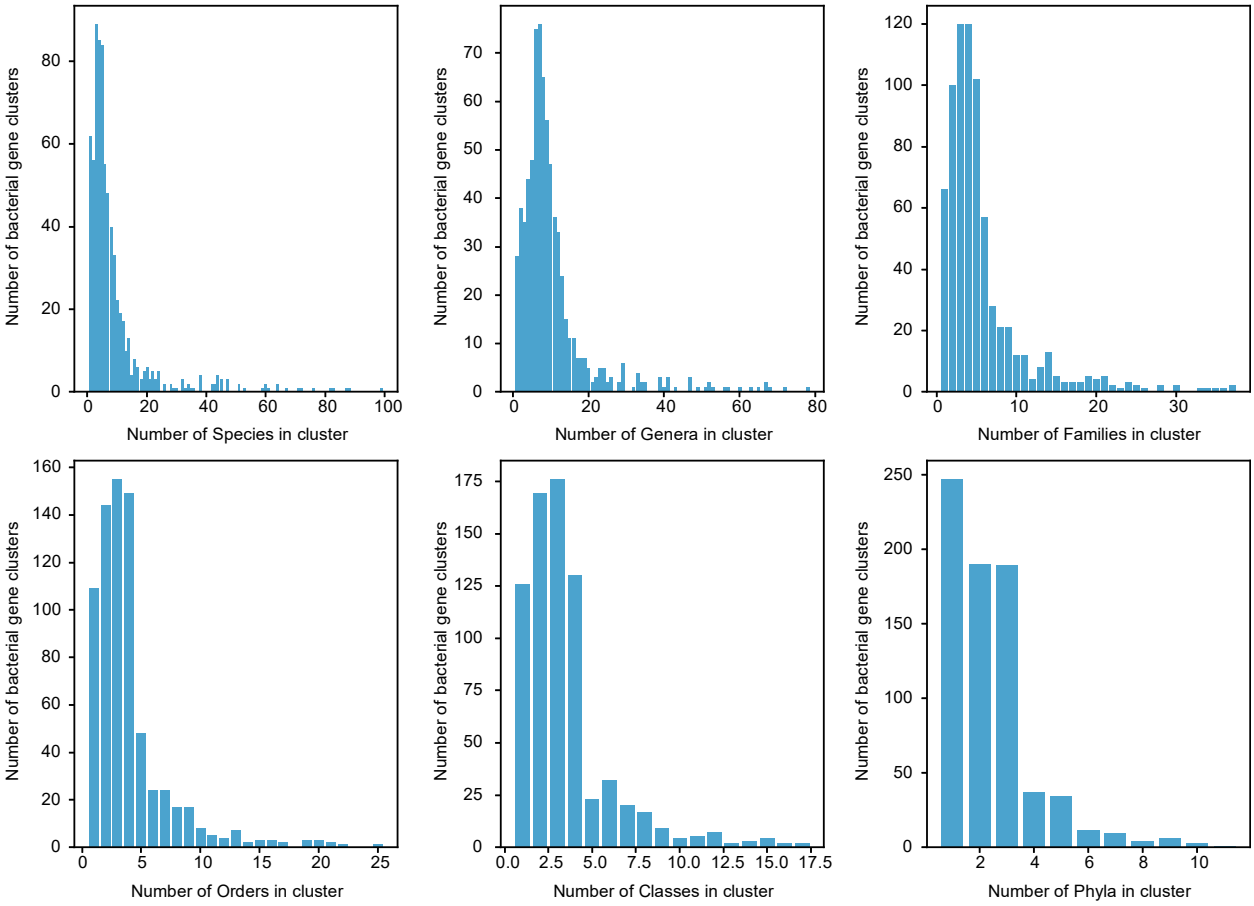


Figure S3. Performance metrics.
A heatmap of precision, recall, and F1-scores for random forests with 5000 estimators, evaluated using leave-one-out cross-validation on each of the eight studies. Performances are listed for both the bacterial and human representations of the metagenomic sample. The bacterial representation lists all the bacterial genes detected in a patient that share a UniRef cluster with an experimentally-verified human-protein interactor. The vector of human proteins represents all the human proteins which might be targeted by the bacterial genes found in each metagenome.



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



767 **Figure S5. Pairwise identity and similarity between proteins found in the human microbiome and those with**
768 **experimentally verified interaction.**
769 Histogram showing the percent identity and similarity between bacterial proteins with experimental verification and their
770 corresponding detected proteins in human microbiomes in the same UniRef cluster. Three aggregation methods are used to
771 estimate each metric at a cluster level: median, average, and an average weighted by prevalence.

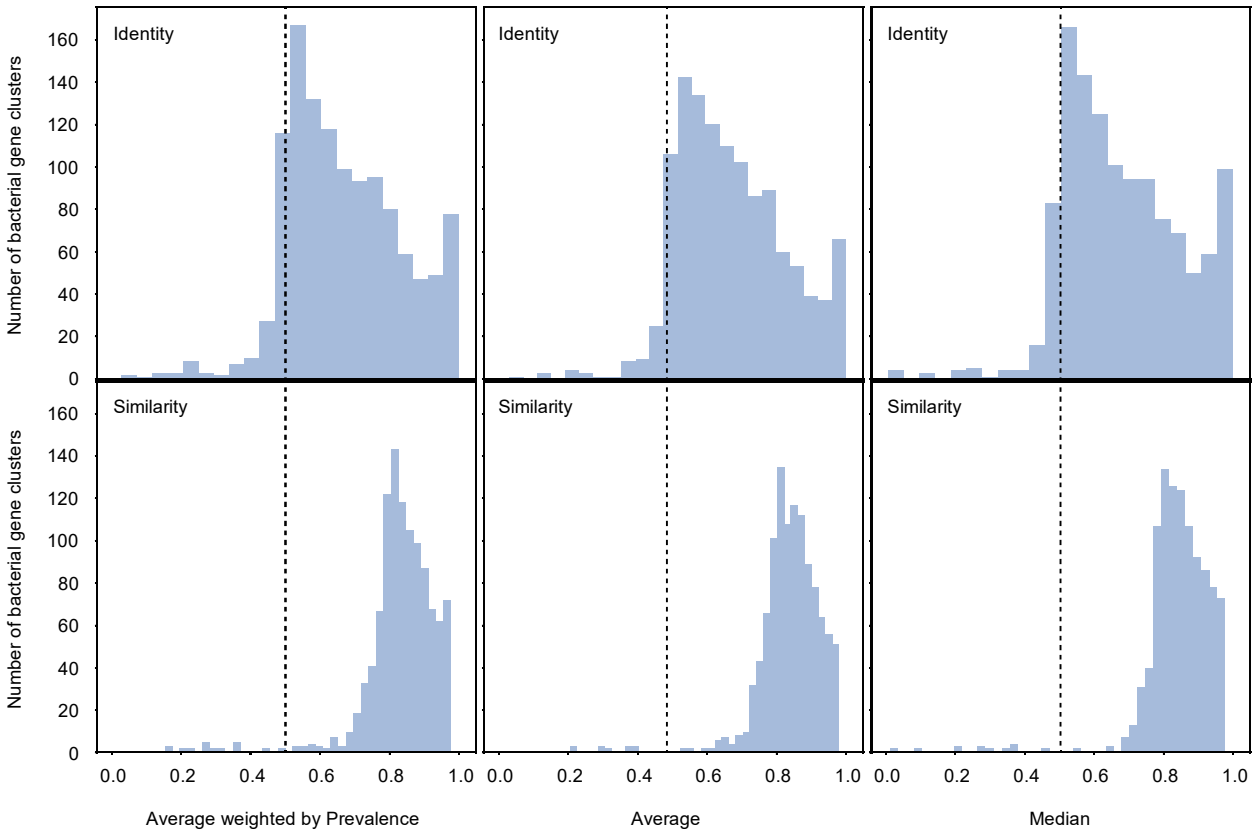


Figure S6. Interface similarity between bacterial proteins within a UniRef cluster.
Similarity, identity, and Jensen-Shannon divergence of interface residues across all bacterial members of the same UniRef cluster sourced from all cocrystal structures with human and bacterial interactors and no filtering based on our datasets.

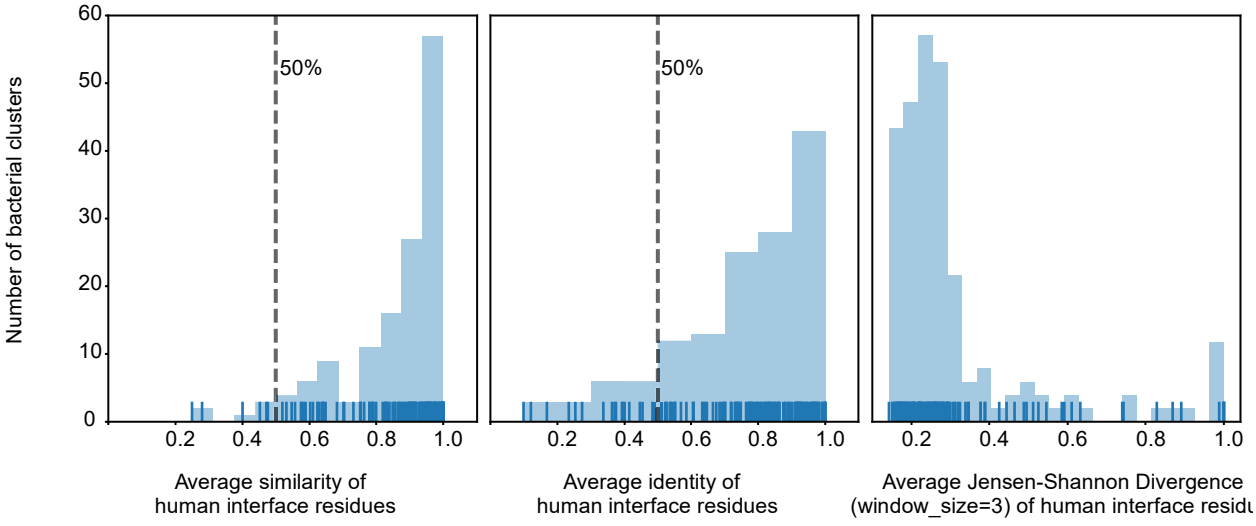


Figure S7. Previous gene-disease associations for human interactors in our dataset.
The number of human interactors (with normalized feature importance greater than 0) according to their GDAs for CRC, T2D, obesity, IBD and autoimmunity.

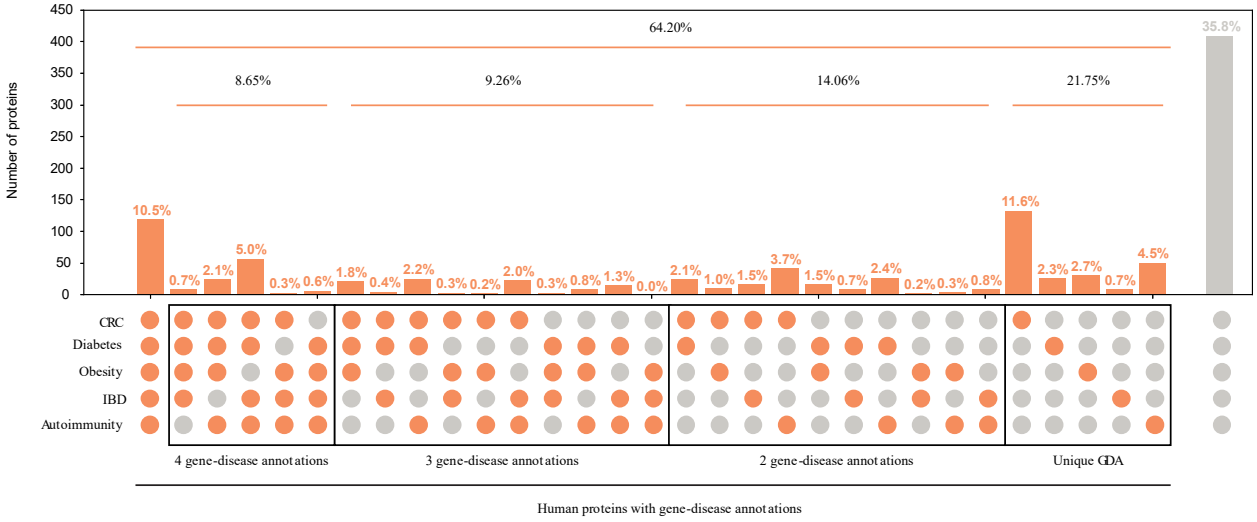


Figure S8. Human protein interactors according to their normalized feature importance and log odds ratio. Volcano plots of the human protein interactors according to their normalized feature importance and log odds ratios in each case-control cohort study.

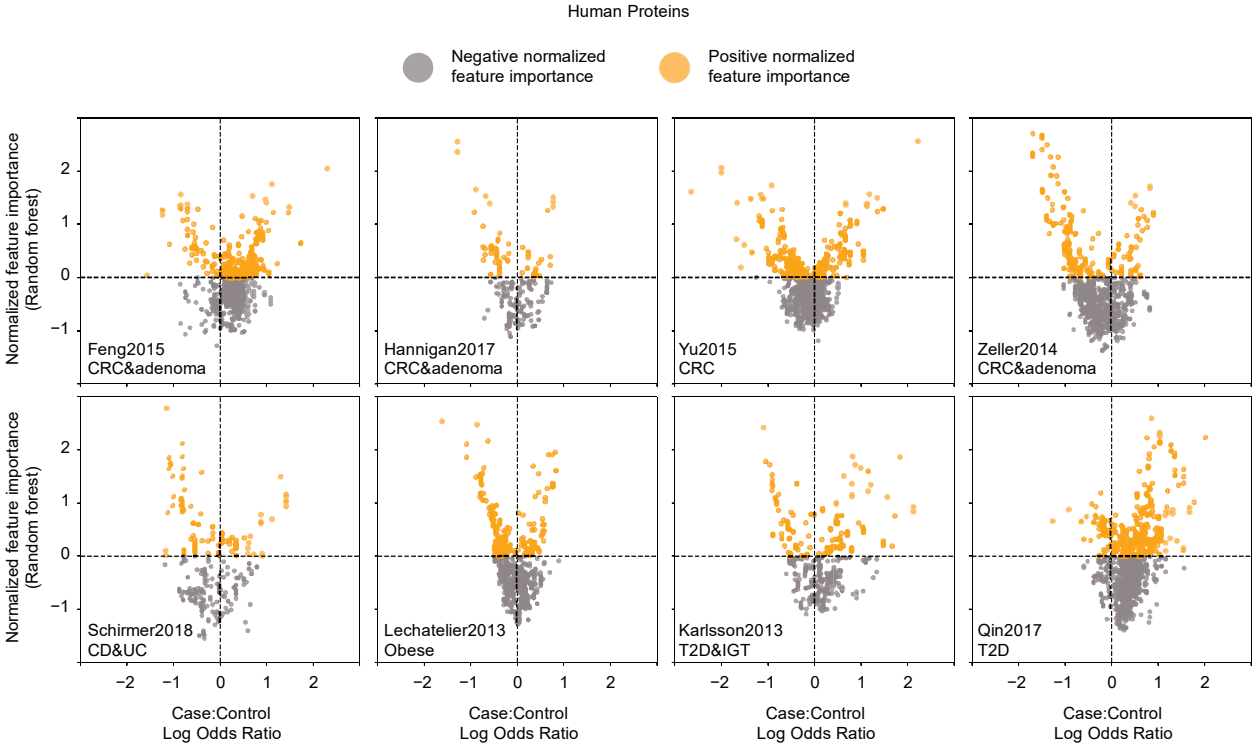


Figure S9. Clustering of cases and controls is not due to disease status or study.
(A) Principal components analysis of patients by their detected human protein interactors, colored by study and label. **(B)** Principal components analysis of detected human protein interactors for all samples in eight metagenomic studies colored by disease status according to study. Controls are all colored together in blue.

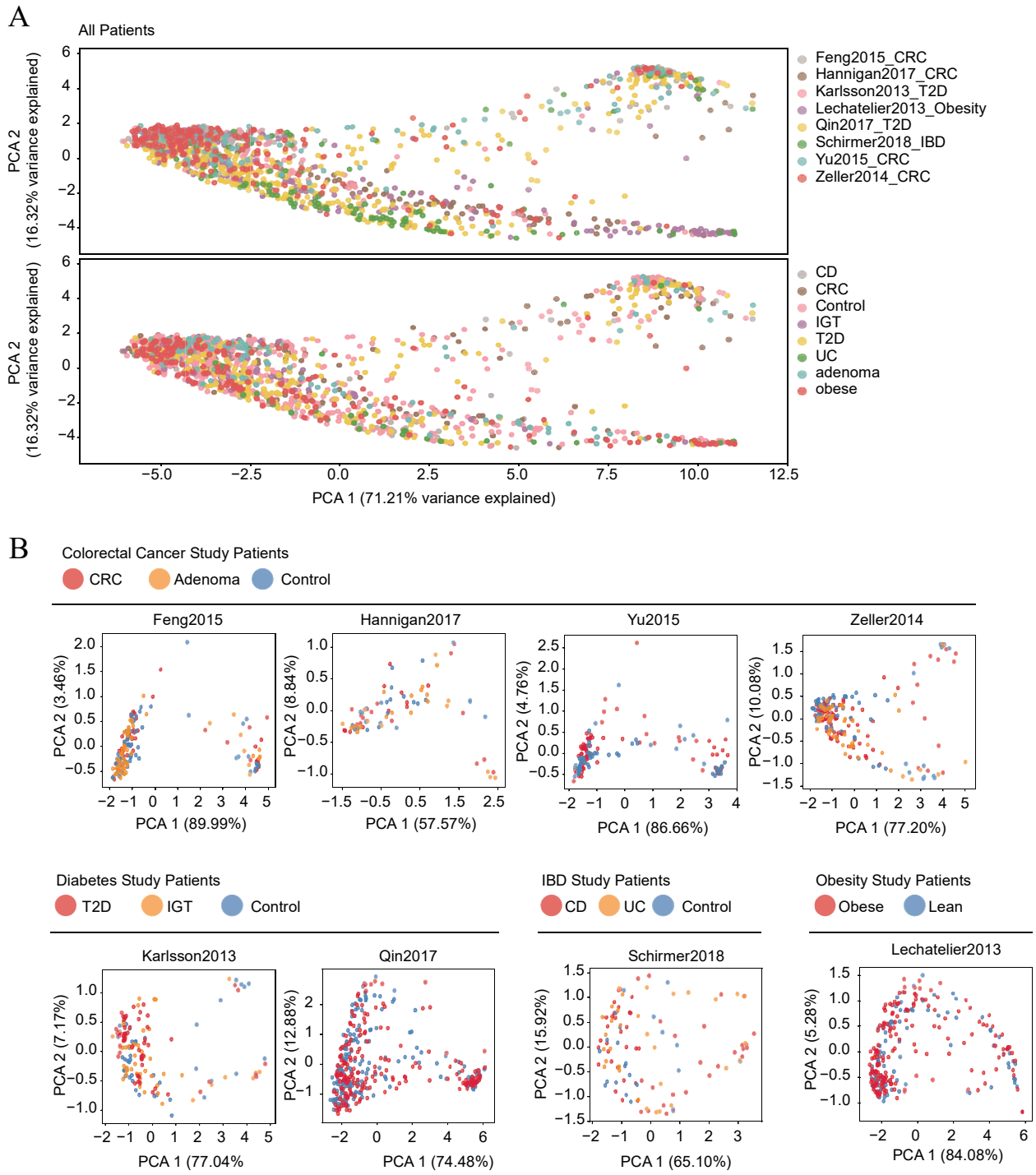
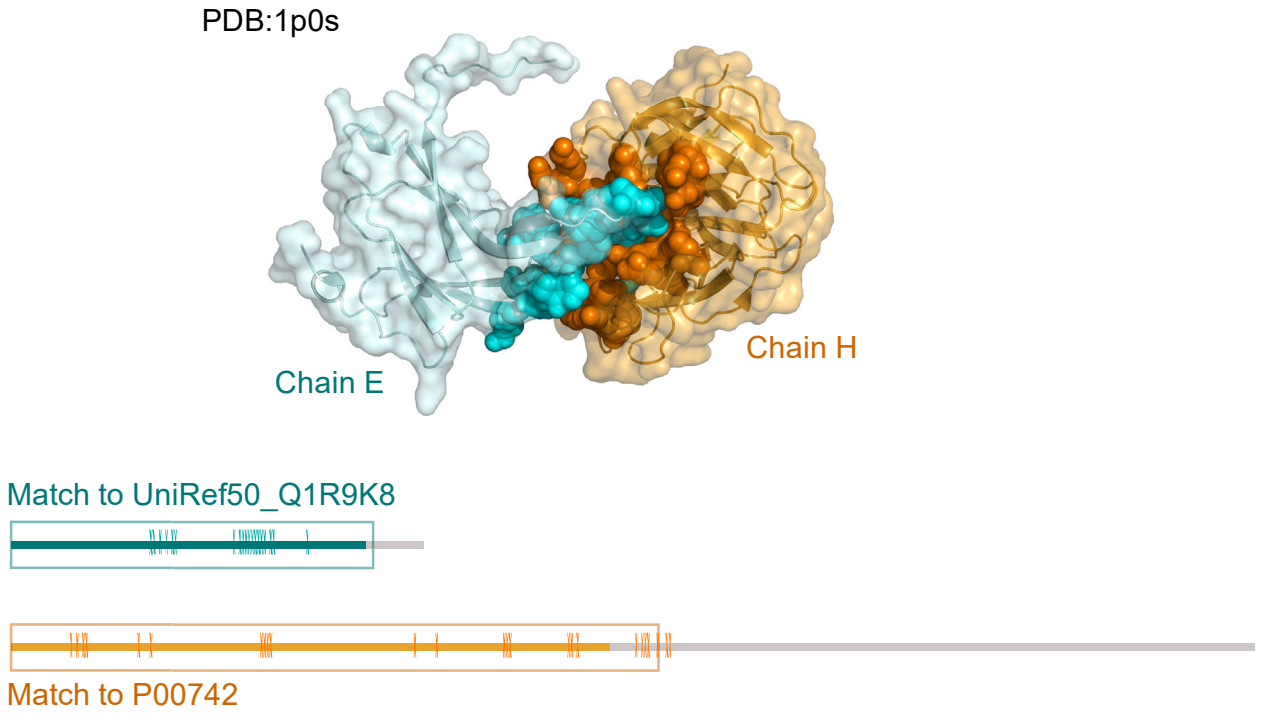


Figure S10. Cocrystal structure of blood coagulation factor Xa in complex with Ecotin M84R. Cluster Uniref50_Q1R9K8 contains several bacterial ecotins detected in human metagenomes. Using BLAST, we found high-quality matches between members of this cluster and the structure 1p0s:E (Ecotin precursor M84R) in the PDB (identity of 97.2%, eval=1e-75). Our putative interactor to this cluster, coagulation factor X (P00742) likewise matched structure 1p0s:H (coagulation factor X precursor) (identity of 100%, eval=3.8 e-150). Chain E is shown in blue, and chain H in orange, with their interface residues highlighted as spheres. The linear model of both proteins is shown underneath. The linear model's colored areas indicate the part of the proteins that were crystallized in this PDB, while the greyed-out areas indicate non-crystallized spans. The squares indicate the range of the BLAST match between our query proteins and the PDB reference sequences. Finally, ticks on the linear model indicate the location of interface residues as detected in this model. There are currently not enough published structures to perform this analysis on all interactions involving detected bacterial genes (Fig. S1, Tables S4 and S5).



Supplementary Tables

Table S1. Metagenomic studies used in this research.
For each study, we list its disease 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 with normalized feature importance greater than 0.

Table S2. Human interactors that have known gene-disease associations.
Listed are the disease-associated human proteins (with normalized feature importance greater than 0) with GDAs in DisGeNET, along with the study in which they are found to be important.

Table S3. Human interactors that are known drug targets.
For each disease-associated human protein (with normalized feature importance greater than 0), we list the drug interactor and the study in which it was found to be important.

Table S4. Interface residues from PDB chain pairs matching human and bacterial interactors in our dataset.
All pairs of detected bacterial proteins and human proteins in the eight metagenomic datasets that have BLASTp matches to two different chains within the same PDB cocrystal structure (totaling 15 bacterial protein clusters and 8 human proteins). Listed are the BLAST readouts for both matches, as well as the interface residues for each chain at the PDB index, PDB sequence, and UniProt sequence mappings.

Table S5. Cocrystal structures representing interactions in our set.
A summary of the PDB chain-pairs (presented in Table S4) that can be used as representatives to identify interface residues for interactions in our set. We annotate each interaction by whether the bacterial and human proteins match non-overlapping pairs of chains.

Table S6. Gene-disease association comparison statistics.
The set sizes, fractions, chi-squared statistics and p-values used to generate Fig. 1E.