MicrobioLink: An integrated computational pipeline to infer functional effects of microbiome-host interactions

Tahila Andrighetti[1,2,3], Leila Gul[1], Tamas Korcsmaros[1,2,$], Padhmanand Sudhakar[1,2,4,$]

[1] Earlham Institute, Norwich Research Park, Norwich, NR4 7UZ, UK
[2] Quadram Institute, Norwich Research Park, Norwich, NR4 7UA, UK
[3] São Paulo University (UNESP), Institute of Biosciences, Botucatu, Brazil.
[4] Department of Chronic Diseases, Metabolism and Ageing, KU Leuven, Leuven, Belgium.

[$] Joint corresponding authors

Corresponding author contact details:

Dr. Padhmanand Sudhakar
Padhmanand.Sudhakar@earlham.ac.uk
Phone : 0044-1603450865
Fax : 0044-1603450021
Address: Earlham Institute, Norwich Research Park, Norwich, NR4 7UZ, UK

Abstract
Microbiome-host interactions are important to nearly all living organisms, and play important roles in health and disease. Understanding these inter-kingdom cross-talks has a huge potential to advance diverse scientific domains such as the clinical and medical sciences, agriculture, and ecology. Detecting such interactions by experimental techniques remain challenging from a cost and feasibility perspective thus hampering large-scale analyses. Computational approaches not only make the inference of microbiome-host interactions viable but also scalable. Here, we present MicrobioLink, a computational pipeline to integrate predicted interactions between microbial and host proteins with host molecular networks. With MicrobioLink users can analyse how microbial proteins in a certain context are influencing cellular processes by interacting with host proteins and affecting downstream pathways, regulating gene expression and metabolism. To show the applicability of the pipeline, we used Crohn's disease metaproteomic data and compared the predictions from the pipeline with those from the healthy condition to demonstrate the achievability of a potential model showing how the influence of microbial proteins can potentially influence disease developments. MicrobioLink is a freely available on GitHub.

Code availability: https://github.com/korcsmarosgroup/HMIpipeline

Introduction

Microbiota-host interactions happen in most organisms, shaping their metabolism and evolution [1–3]. In many ecosystems, microbial communities play important roles in the dynamic interactions with the host [4]. When associated with host organisms, microbial communities influence host physiology and health. The community of microorganisms are indispensable to human life since they modulate and influence immunity and nutrient acquisition. For example, the gastrointestinal microbiome plays a crucial role in nutrient assimilation and energy yield by actively participating in metabolic pathways [9]. Dysbiosis (compositional alterations) of gut microbial communities can induce diseases such as type 2 diabetes, obesity and inflammatory bowel diseases such as Crohn's disease [10, 11]. Infection with

pathogenic microbes are known to cause harm to the host as well as to the microbial community by excluding beneficial species [12].

Therefore, the study of microbiome-host interactions and cross-talks are important in terms of monitoring and remediation of health. Such cross-talks are mediated by molecular interactions between various molecular components expressed by the host and the microbiome. Bacterial molecules, such as metabolites [15], proteins [16, 17, 17–20] and small RNAs [21], can interact with host molecules, and launch an intracellular cascade able to modulate biological processes by affecting key host genes and proteins. Next to metabolite mediated interactions, protein-protein interactions (PPIs) are one of the most relevant and reported types of molecular interactions between the host and microbes [17–20]. However, experimental techniques to probe inter-species PPIs are time consuming and have limitations imposed by cost [2, 14, 16, 22–24]. Hence, there is a dearth of validated inter-species PPIs in publicly available databases and are mostly limited to pathogens. From the inferred microbe-host PPIs, it is possible to build interaction networks to better understand the interactions among the microbial and host components and how these interactions interfere with host metabolism and physiology [25].

There are some publicly available microbe-host interaction prediction based resources and pipelines. Some like PHISTO [26], PATRIC [27], Proteopathogen [28], VirBase [29] and, HPIDB [30] perform PPI predictions but they only include predictions for pathogens [25]. Other tools, which include commensal microbes as COBRA [31], RAVEN [32] and Kbase [33], use metabolomics data instead of proteomics [34]. However, most of the existing resources and pipelines are confined to predicting the direct molecular interactions at the microbe-host interface, and do not infer the downstream effects on functional processes and host signalling pathways.

Here, we introduce MicrobioLink - a pipeline to analyse microbiome-host interactions at a cellular level using network approaches. There are two central phases in the workflow of MicrobioLink: the prediction of interactions between microbial and host proteins and the integration of network and omics datasets for the downstream analysis. The first phase is built on the structural features of the interacting proteins. The second phase uses publicly available host multilayered networks containing experimentally

determined molecular interactions, we prioritize the microbiome-host interactions that converge to and potentially effect a selected set of user-provided gene/protein list. The user-provided gene/protein list can be compiled either from *a priori* knowledge derived from phenotypic observations or contextual data obtained from -omic experiments such as gene expression, proteomics etc. This key feature of MicrobioLink helps end-users to expand the list of experimentally verifiable hypotheses postulated from the inferred indirect effects of microbial proteins on the host and prioritize pathways and signalling chains based on various criteria (such as chain length, host protein localization, contextual expression).

Pipeline description

As a first step in the workflow (Figure 1a), bacterial and host protein lists for the interaction prediction are compiled. The bacterial proteins can be obtained from annotated and/or experimentally derived bacterial proteomes or metaproteomes in the case of communities. As for the host proteins, the list is compiled based on their localization depending on whether or not the interactions between the bacterial and host proteins are spatially possible. Localization based filtering can be applied also to bacterial proteins depending on the context. For example, in the case of microbes that are extracellular to the host, host proteins are confined to those present at the extracellular matrix or plasma membrane, because these are the locations where they are more prone to interact with the bacterial proteins. In this case, bacterial proteins are narrowed down to extracellular or secreted or membrane bound subsets. The bacterial proteins can be provided by the user, either from experiments,publicly available datasets or bioinformatic predictions. Resources which are currently used in the MicrobioLink pipeline to compile the human proteins are ComPPI [35], MatrixDB [36] and Human Protein Atlas (HPA) [37]. The user can also provide their own pre-compiled lists of proteins.

The next step is the interaction prediction. In this step, we compare the annotated domains and motifs from the compiled bacterial and host proteins with the databases DOMINE [38] and ELM [39], which respectively contain high-confidence lists of domain-domain and domain-motif pairs already

known to interact [18, 40, 41]. This procedure helps determine which bacterial proteins *via* their respective domains interact with host receptor proteins. For quality control purposes, the interactions are then filtered to retain only those interactions which are sterically possible [42, 43]. This is performed by excluding interactions involving motifs outside disordered interactions (using IUPRED [44]) and/or within globular domains (from PFAM [45], InterPro [46]). Some studies such as [18–20] have already successfully used this structure-based approach to predict PPIs between microbial and host proteins, and validated the PPIs experimentally.

To trace the possible paths from the host receptor proteins potentially modulated by the bacterial proteins, directed multilayered networks containing protein-protein interactions (PPIs) and transcriptional regulatory interactions (TRIs) are compiled from publicly available molecular interaction databases such as HTRIdb (Bovolenta et al. 2012), TRUSST (Han et al. 2018) and OmniPath (Türei et al. 2016). By compiling the networks, it is possible to obtain potential chains/pathways affected by the bacterial proteins. Subsequently, the user can provide a list of host proteins of interest, which can be important proteins for the studied conditions or related cellular processes. This data is provided by the user based on GWAS, literature search or -omic expression datasets. Such proteins can be subsequently used to to prioritize the molecular chains, retaining only those involving the selected host proteins of interest.

Then, the chains are filtered by their specificity: chains connected with bacterial proteins from both conditions are excluded, keeping only chains specific from one of the conditions. Additionally, if the selected host target genes at the end of the chains are from a transcriptomic dataset, then the immediate upstream neighbour of the selected host target genes needs to be a transcription factor, so only chains with TRIs between them are kept.

Finally, the obtained model can be visualized by a multilayered network which starts with the bacterial protein-host receptor interactions, and ultimately reaching the selected host target genes through PPIs and TRIs. A pipeline summary is presented in Figure 1.a.

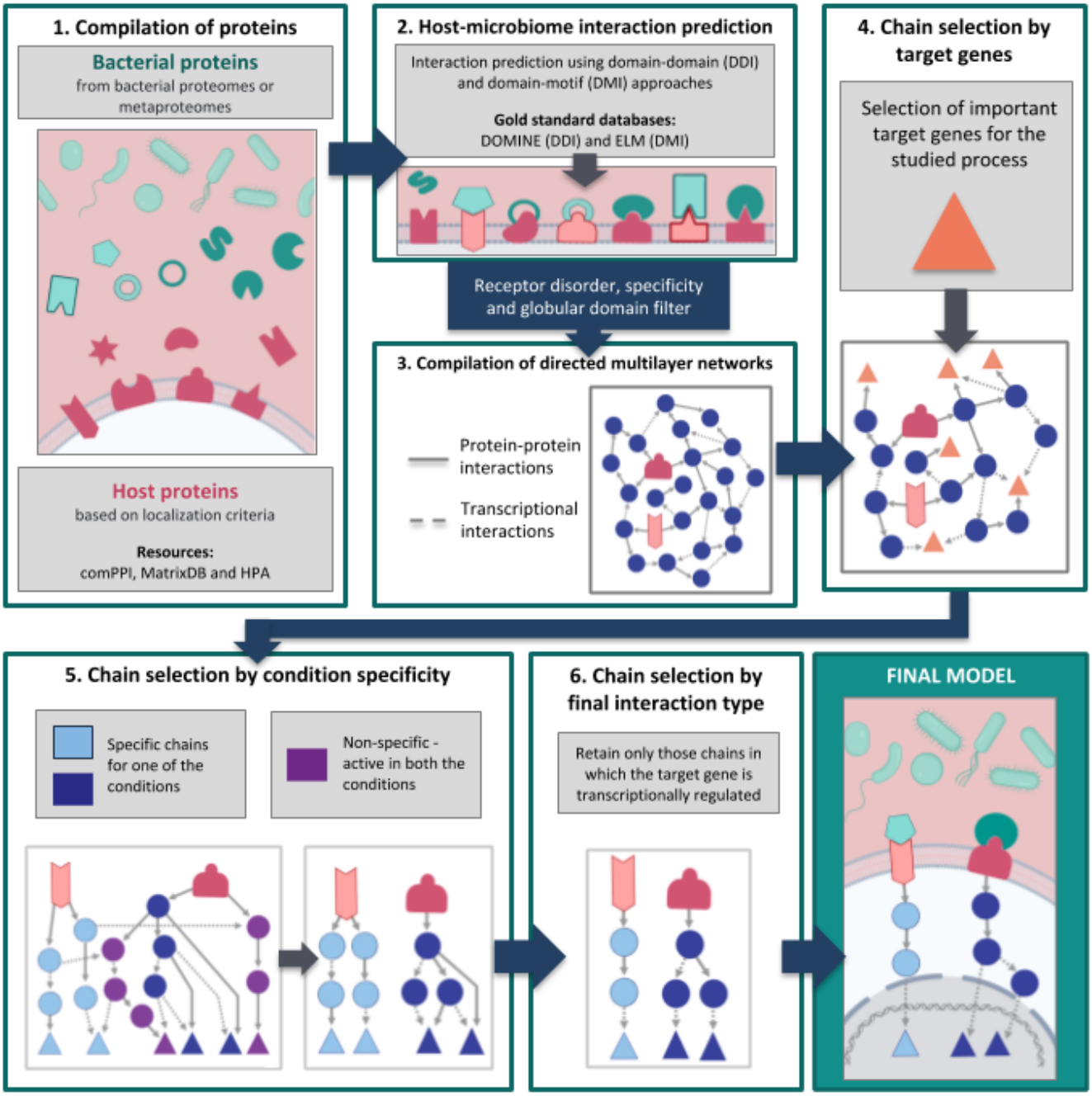
Case study

As an example of demonstrating the applicability of the MicrobioLink pipeline, we performed a study comparing how different are the microbiome-host interactions in patients with Crohn's disease (CD) (a disease characterized by microbial dysbiosis) and from healthy individuals. We then use these network-derived differences to generate hypotheses which can potentially explain the role of the microbiota proteins either in disease development or progression. To perform this study, we first obtained the differentially abundance microbial proteins in healthy individuals in comparison to CD subjects from a Swedish twin study [47]. In total, 1221 microbial proteins from healthy individuals and 865 protein from CD patients were obtained. Then, we compiled human proteins located in the extracellular matrix and cellular membrane, resulting in a total of 3601 proteins.

After compiling the protein lists, we performed the interaction prediction between microbial and human proteins, resulting in 1921 predicted interactions: 433 from healthy and 1488 from CD. The predicted interactions were then refined by passing them through a disorder and structural filtering step in order to eliminate physically and sterically improbable interactions. Subsequently, we selected autophagy genes as potential target genes in the network, since autophagy is one of the known dysregulated cellular processes in CD. This would also enable us to throw light into understanding the potential pathways by which autophagy could be modulated by microbial proteins.

In the next phase, we compiled signaling networks from the publicly available datasets described in the Pipeline description section. These compiled networks contained molecular signaling chains which potentially mediate the signal transmission from the receptors to the autophagy genes. The initial network is depicted in Supplementary Figure 1. In this figure, we observed that there were many human proteins that shared connections with both CD and healthy bacterial proteins. To retain a specificity as high as possible, we filtered the proteins by excluding nodes connected with both conditions, which then resulted in the network shown in Supplementary Figure 2.

In the final filtering step, we retained only chains where the last interaction is a TRIs, excluding chains with PPIs in that position. Then, we obtained a model showing how bacterial proteins can potentially modulate autophagy in Crohn's disease. The final model is depicted in Figure 1b.



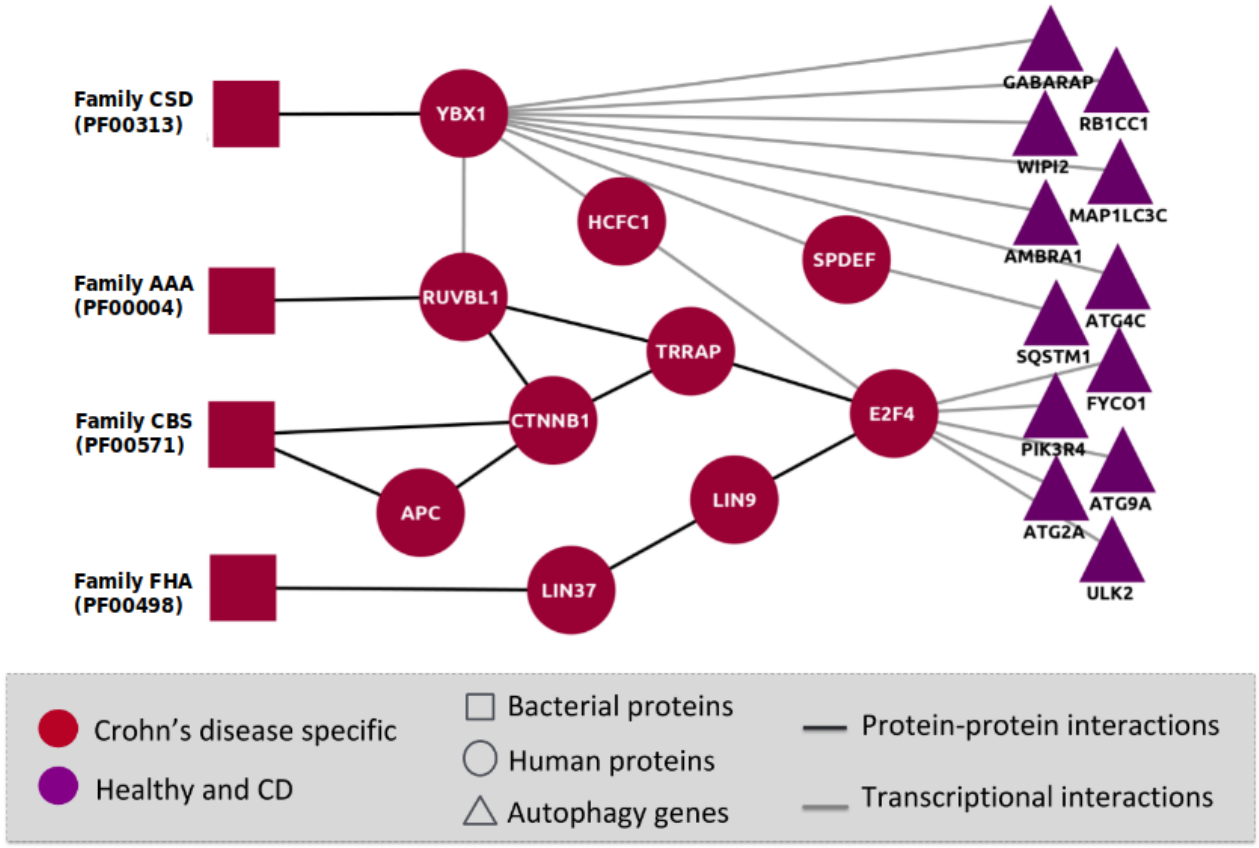


Figure 1: a) Graphical representation of the workflow; b) An example of a network model obtained from the workflow.

Conclusion

Microbiome-host interactions have considerable impact on host phenotypes and their understanding is crucial in making advances in various scientific fields such as medicine, agriculture, environmental conservation etc. However, there are technological and methodological boundaries that challenge researchers to investigate such interactions in a cost-effective and easy-to-use way. Therefore, the development of computation-enabled bioinformatic pipelines, as MicrobioLink, can pave the way for researchers in the community to understand the potential microbe-host interactions and the modulated host processes. Due to the different possibilities in terms of all the interactions and pathways possible between the set of microbial and host proteins, MicrobioLink will help prioritize on a selected number of pathways

based on specific criteria such as contextual expression data etc. MicrobioLink builds on existing methods, which were used separately and in a nonstandardized way before, to predict microbiome-host interactions and extends them with novel additional functionalities such as inferring the functional relevance of the host proteins potentially modulated by the microbial proteins. This inference is achieved by embedding the host proteins within molecular interaction networks and relevant -omic datasets.

Using MicrobioLink is possible to evaluate microbiome-host interactions for many different conditions and organisms. It can be used either for evaluate how single microorganism interfere in the host processes or entire microbiomes. Also, it can be used for many types of hosts, being plants, animals or fungi and enables the evaluation for either commensal and pathogen interactions, enabling to understand the environment dynamics to develop models from this interactions.

Code availability

Codes of the MicrobioLink pipeline (in Python) are available at https://github.com/korcsmarosgroup/HMIpipeline.

Acknowledgement

TK was supported by a fellowship in computational biology at Earlham Institute (Norwich, UK) in partnership with the Quadram Institute (Norwich, UK), and strategically supported by Biotechnological and Biosciences Research Council, UK (BB/J004529/1 and BB/P016774/1). LG was supported by the BBSRC Norwich Research Park Biosciences Doctoral Training Partnership (grant BB/M011216/1). PS was supported by the ERC Advanced Grant (ERC-2015-AdG-694679).

References

1. Barrett KE, Wu GD. Influence of the microbiota on host physiology - moving beyond the gut. J Physiol (Lond). 2017;595:433–5. doi:10.1113/JP273451.

2. Weimer BC, Chen P, Desai PT, Chen D, Shah J. Whole Cell Cross-Linking to Discover Host-Microbe Protein Cognate Receptor/Ligand Pairs. Front Microbiol. 2018;9:1585. doi:10.3389/fmicb.2018.01585.

3. Simon J-C, Marchesi JR, Mougel C, Selosse M-A. Host-microbiota interactions: from holobiont theory to analysis. Microbiome. 2019;7:5. doi:10.1186/s40168-019-0619-4.

4. Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. PLoS Comput Biol. 2010;6:e1000667. doi:10.1371/journal.pcbi.1000667.

5. Fitzpatrick CR, Copeland J, Wang PW, Guttman DS, Kotanen PM, Johnson MTJ. Assembly and ecological function of the root microbiome across angiosperm plant species. Proc Natl Acad Sci USA. 2018;115:E1157–65. doi:10.1073/pnas.1717617115.

6. Morgavi DP, Kelly WJ, Janssen PH, Attwood GT. Rumen microbial (meta)genomics and its application to ruminant production. Animal. 2013;7 Suppl 1:184–201. doi:10.1017/S1751731112000419.

7. Di Bella JM, Bao Y, Gloor GB, Burton JP, Reid G. High throughput sequencing methods and analysis for microbiome research. J Microbiol Methods. 2013;95:401–14. doi:10.1016/j.mimet.2013.08.011.

8. Abdallah F, Mijouin L, Pichon C. Skin immune landscape: inside and outside the organism. Mediators Inflamm. 2017;2017:5095293. doi:10.1155/2017/5095293.

9. Krajmalnik-Brown R, Ilhan Z-E, Kang D-W, DiBaise JK. Effects of gut microbes on nutrient absorption and energy regulation. Nutr Clin Pract. 2012;27:201–14. doi:10.1177/0884533611436116.

10. Devaraj S, Hemarajata P, Versalovic J. The human gut microbiome and body metabolism: implications for obesity and diabetes. Clin Chem.

2013;59:617–28. doi:10.1373/clinchem.2012.187617.

11. Loh G, Blaut M. Role of commensal gut bacteria in inflammatory bowel diseases. Gut Microbes. 2012;3:544–55. doi:10.4161/gmic.22156.

12. Pickard JM, Zeng MY, Caruso R, Núñez G. Gut microbiota: Role in pathogen colonization, immune responses, and inflammatory disease. Immunol Rev. 2017;279:70–89. doi:10.1111/imr.12567.

13. Khater M, de la Escosura-Muñiz A, Merkoçi A. Biosensors for plant pathogen detection. Biosens Bioelectron. 2017;93:72–86. doi:10.1016/j.bios.2016.09.091.

14. Foley SL, Johnson TJ, Ricke SC, Nayak R, Danzeisen J. Salmonella pathogenicity and host adaptation in chicken-associated serovars. Microbiol Mol Biol Rev. 2013;77:582–607. doi:10.1128/MMBR.00015-13.

15. Heinken A, Sahoo S, Fleming RMT, Thiele I. Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. Gut Microbes. 2013;4:28–40. doi:10.4161/gmic.22370.

16. Nourani E, Khunjush F, Durmuş S. Computational approaches for prediction of pathogen-host protein-protein interactions. Front Microbiol. 2015;6:94. doi:10.3389/fmicb.2015.00094.

17. Sudhakar P, Jacomin AC, Hautefort I, Samavedam S, Fatemian K, Ari E, et al. Targeted interplay between bacterial pathogens and host autophagy. Autophagy. Autophagy.

18. Sudhakar P, Jacomin A-C, Hautefort I, Samavedam S, Fatemian K, Ari E, et al. Targeted interplay between bacterial pathogens and host autophagy. Autophagy. 2019;:1–14. doi:10.1080/15548627.2019.1590519.

19. Evans P, Dampier W, Ungar L, Tozeren A. Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. BMC Med Genomics. 2009;2:27. doi:10.1186/1755-8794-2-27.

20. Wojcik J, Schächter V. Protein-protein interaction map inference using interacting domain profile pairs. Bioinformatics. 2001;17 Suppl 1:S296-305. doi:10.1093/bioinformatics/17.suppl_1.S296.

21. Katiyar-Agarwal S, Jin H. Role of small RNAs in host-microbe interactions. Annu Rev Phytopathol. 2010;48:225–46. doi:10.1146/annurev-phyto-073009-114457.

22. Dyer MD, Murali TM, Sobral BW. The landscape of human proteins interacting with viruses and other pathogens. PLoS Pathog. 2008;4:e32. doi:10.1371/journal.ppat.0040032.

23. Dyer MD, Neff C, Dufford M, Rivera CG, Shattuck D, Bassaganya-Riera J, et al. The human-bacterial pathogen protein interaction networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis. PLoS ONE. 2010;5:e12089. doi:10.1371/journal.pone.0012089.

24. Gould AL, Zhang V, Lamberti L, Jones EW, Obadia B, Korasidis N, et al. Microbiome interactions shape host fitness. Proc Natl Acad Sci USA. 2018;115:E11951–60. doi:10.1073/pnas.1809349115.

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

26. Durmuş Tekir S, Çakır T, Ardiç E, Sayılırbaş AS, Konuk G, Konuk M, et al. PHISTO: pathogen-host interaction search tool. Bioinformatics. 2013;29:1357–8. doi:10.1093/bioinformatics/btt137.

27. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 2014;42 Database issue:D581-91. doi:10.1093/nar/gkt1099.

28. Vialás V, Nogales-Cadenas R, Nombela C, Pascual-Montano A, Gil C. Proteopathogen, a protein database for studying Candida albicans--host interaction. Proteomics. 2009;9:4664–8. doi:10.1002/pmic.200900023.

29. Li Y, Wang C, Miao Z, Bi X, Wu D, Jin N, et al. ViRBase: a resource for virus-host ncRNA-associated interactions. Nucleic Acids Res. 2015;43 Database issue:D578-82. doi:10.1093/nar/gku903.

30. Kumar R, Nanduri B. HPIDB--a unified resource for host-pathogen interactions. BMC Bioinformatics. 2010;11 Suppl 6:S16. doi:10.1186/1471-

2105-11-S6-S16.

31. Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nat Protoc. 2019;14:639–702. doi:10.1038/s41596-018-0098-2.

32. Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J. The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum. PLoS Comput Biol. 2013;9:e1002980. doi:10.1371/journal.pcbi.1002980.

33. Arkin AP, Stevens RL, Cottingham RW, Maslov S, Henry CS, Dehal P, et al. The DOE systems biology knowledgebase (kbase). BioRxiv. 2016. doi:10.1101/096354.

34. Sen P, Orešič M. Metabolic modeling of human gut microbiota on a genome scale: an overview. Metabolites. 2019;9. doi:10.3390/metabo9020022.

35. Veres DV, Gyurkó DM, Thaler B, Szalay KZ, Fazekas D, Korcsmáros T, et al. ComPPI: a cellular compartment-specific database for protein-protein interaction network analysis. Nucleic Acids Res. 2015;43 Database issue:D485-93. doi:10.1093/nar/gku1007.

36. Launay G, Salza R, Multedo D, Thierry-Mieg N, Ricard-Blum S. MatrixDB, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities. Nucleic Acids Res. 2015;43 Database issue:D321-7. doi:10.1093/nar/gku1091.

37. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015;347:1260419. doi:10.1126/science.1260419.

38. Yellaboina S, Tasneem A, Zaykin DV, Raghavachari B, Jothi R. DOMINE: a comprehensive collection of known and predicted domain-domain interactions. Nucleic Acids Res. 2011;39 Database issue:D730-5. doi:10.1093/nar/gkq1229.

39. Gouw M, Michael S, Sámano-Sánchez H, Kumar M, Zeke A, Lang B, et

al. The eukaryotic linear motif resource - 2018 update. Nucleic Acids Res. 2018;46:D428–34. doi:10.1093/nar/gkx1077.

40. Bader S, Kühner S, Gavin A-C. Interaction networks for systems biology. FEBS Lett. 2008;582:1220–4. doi:10.1016/j.febslet.2008.02.015.

41. Korcsmaros T, Dunai ZA, Vellai T, Csermely P. Teaching the bioinformatics of signaling networks: an integrated approach to facilitate multi-disciplinary learning. Brief Bioinformatics. 2013;14:618–32. doi:10.1093/bib/bbt024.

42. Dinkel H, Van Roey K, Michael S, Davey NE, Weatheritt RJ, Born D, et al. The eukaryotic linear motif resource ELM: 10 years and counting. Nucleic Acids Res. 2014;42 Database issue:D259-66. doi:10.1093/nar/gkt1047.

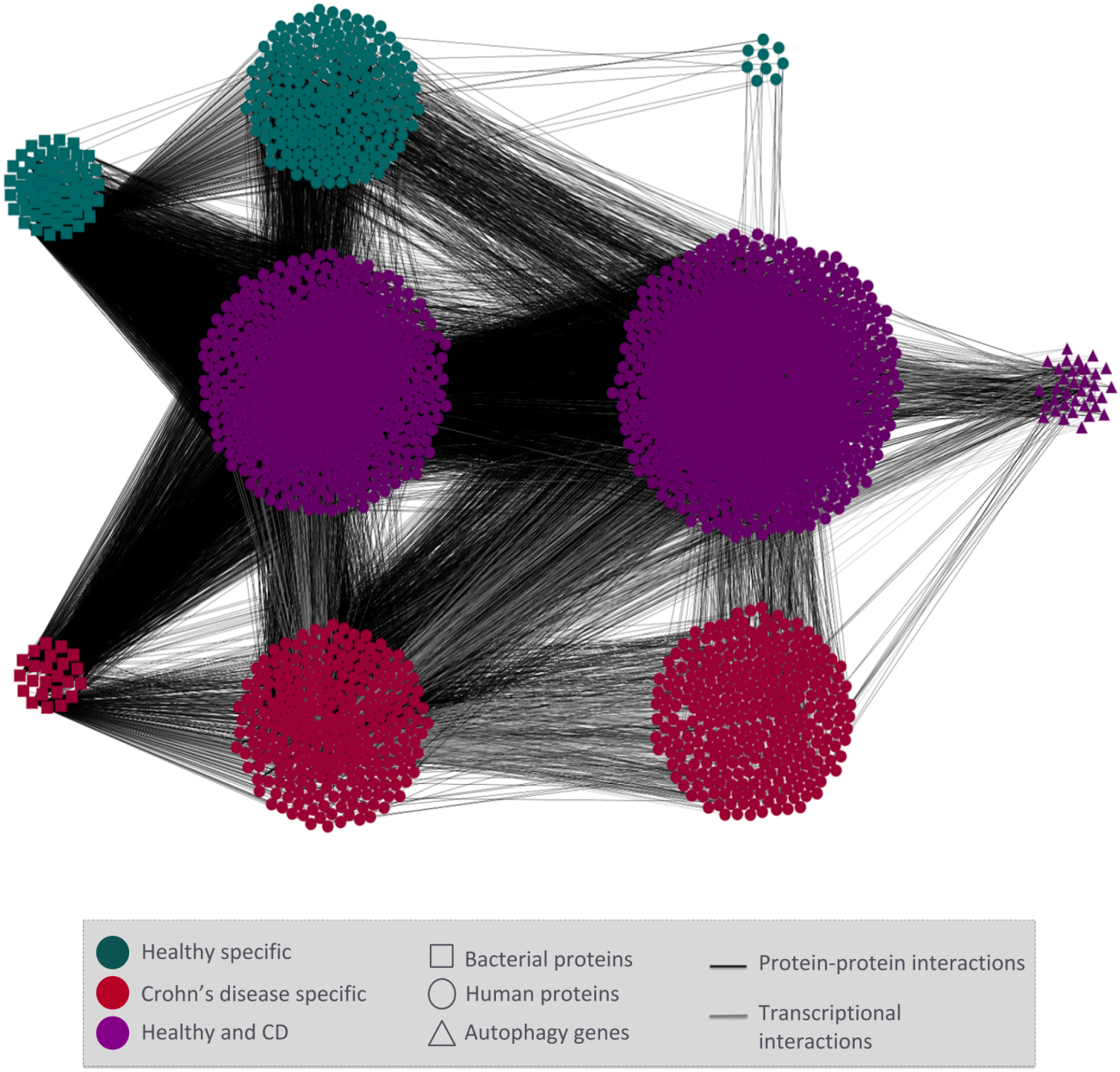
43. Weatheritt RJ, Jehl P, Dinkel H, Gibson TJ. iELM--a web server to explore short linear motif-mediated interactions. Nucleic Acids Res. 2012;40 Web Server issue:W364-9. doi:10.1093/nar/gks444.

44. Dosztányi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics. 2005;21:3433–4. doi:10.1093/bioinformatics/bti541.

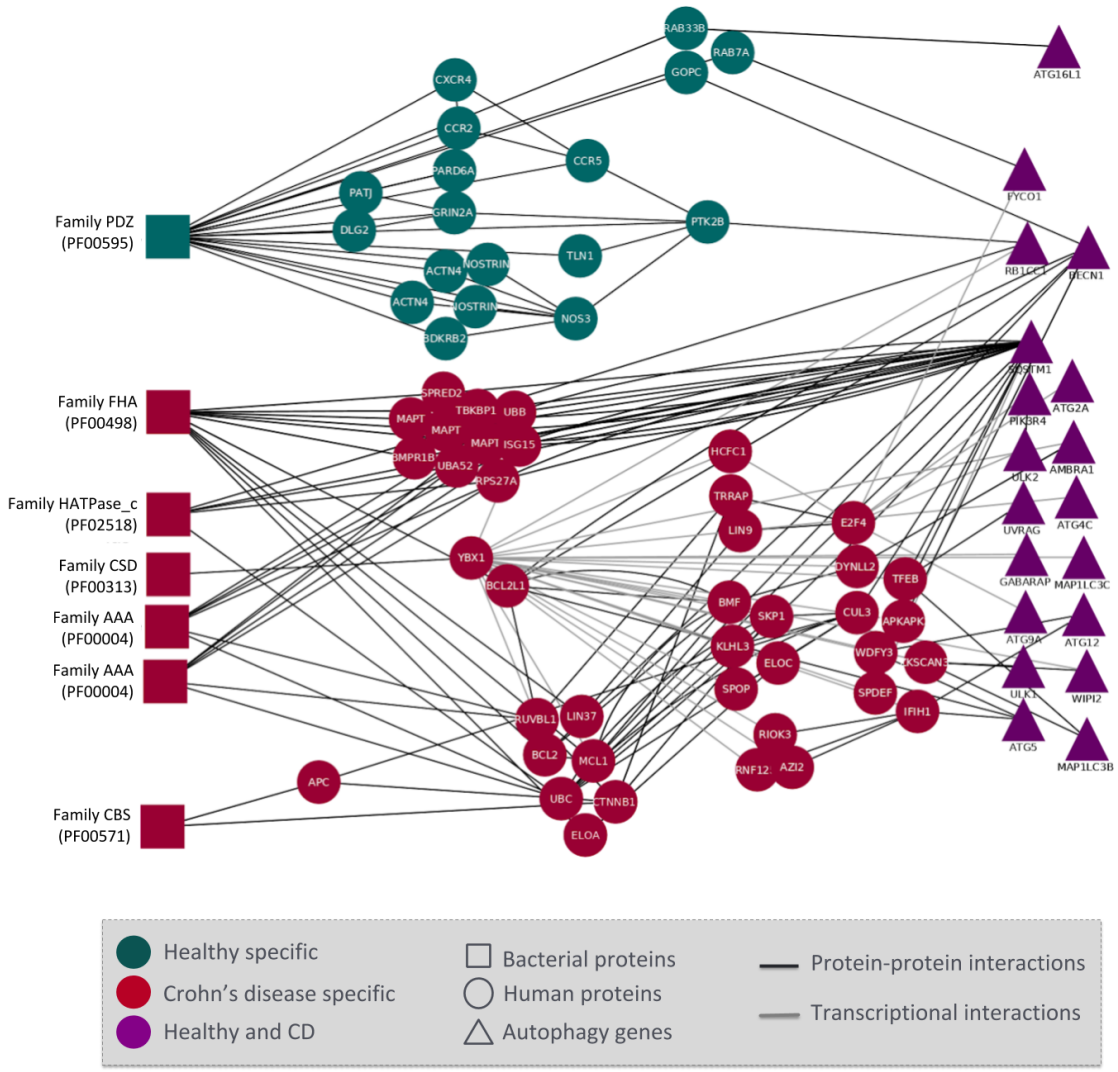
45. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47:D427–32. doi:10.1093/nar/gky995.

46. Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res. 2019;47:D351–60. doi:10.1093/nar/gky1100.

47. Erickson AR, Cantarel BL, Lamendella R, Darzi Y, Mongodin EF, Pan C, et al. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn’s disease. PLoS ONE. 2012;7:e49138. doi:10.1371/journal.pone.0049138.



Supplementary figure 1: The network obtained by tracing the signalling chains from the human receptors (predicted to interact with the bacterial proteins) to the autophagy proteins.



Supplementary figure 2: The network obtained after filtering the initial network (in Supplementary figure 1) according to the various specificities.