COMPUTATIONAL APPROACHES TO STUDY MICROBES AND MICROBIOMES

CASEY S. GREENE[†]
Systems Pharmacology and Translational Therapeutics
Perelman School of Medicine
University of Pennsylvania
Philadelphia PA 19104, USA
Email: csgreene@upenn.edu

JAMES A. FOSTER
Institute of Bioinformatics and Evolutionary Studies
University of Idaho
Moscow, ID 83844 USA
Email: foster@uidaho.edu

BRUCE A. STANTON
Department of Microbiology and Immunology
The Geisel School of Medicine at Dartmouth
Hanover, NH 03755, USA
Email: Bruce.A.Stanton@dartmouth.edu

DEBORAH A. HOGAN
Department of Microbiology and Immunology
The Geisel School of Medicine at Dartmouth
Hanover, NH 03755, USA
Email: Deborah.A.Hogan@dartmouth.edu

YANA BROMBERG
Biochemistry and Microbiology, School of Environmental and Biological Sciences
Rutgers University
New Brunswick, NJ 08901, USA
Institute for Advanced Study, Technische Universität München
Garching, Germany
Email: yana@bromberglab.org

Technological advances are making large-scale measurements of microbial communities commonplace. These newly acquired datasets are allowing researchers to ask and answer questions about the composition of microbial communities, the roles of members in these communities, and how genes and molecular pathways are regulated in individual community members and communities as a whole to effectively respond to diverse and changing environments. In addition to providing a more comprehensive survey of the microbial world, this new information allows for the development of computational approaches to model the processes underlying microbial systems. We anticipate that the field of computational microbiology will continue to grow rapidly in the coming years. In this manuscript we highlight both areas of particular interest in microbiology as well as computational approaches that begin to address these challenges.

[†] To whom correspondence should be addressed.

1. Introduction

Microbes, including viruses, bacteria, and fungi are the most numerous organisms on earth. Bacteria alone are estimated to equal the biomass of plants on earth.[1]Moreover, they are the key drivers of life on earth by controlling the majority of Earth's biogeochemical fluxes.[2]

Microbial communities also play key roles in human health and disease.[3,4] While the role of microbes underlying certain illnesses has been widely recognized, we are also recognizing their role in normal physiology, and the role that they can play to restore normal physiology. For example, a diet of non-digestible but fermentable carbohydrates given to children affected by the Prader-Willi syndrome has been shown to lead to changes in the gut microbiome structure, contributing to reduction in weight, regardless of the continued presence of the primary driving forces.[5] In a more directed experiment, transplants of fecal microbiota has been used to alleviate chronic *Clostridium difficile* infections.[6,7]

Microbial communities were historically relatively difficult to survey and characterize. The development of fast and inexpensive sequencing methods has dramatically aided in this analysis.[8] We can now readily evaluate and describe communities that we could not easily catalog with other approaches.[9,10] These new experimental platforms are providing the basis of in depth surveys of the microbial components of our world. For example, the human microbiome project (HMP) was designed to catalog human-associated microbial communities,[11] producing an extensive bacterial catalog of over 200 adults.[12]

Many other studies are working towards identifying microbiome features that are important for health or disease. For example, a series of studies have characterized the microbiome in lungs of individuals with conditions such as cystic fibrosis (CF),[13–16] chronic obstructive pulmonary disease (COPD),[17] asthma,[3,18] and in the intestinal tract of individuals with CF[19], and diabetes.[4,20] In some cases it has been possible to identify pathogens and/or the expression of particular genes that are associated with positive or negative outcomes.[19,21] It is the hope that knowledge of the microbiome and gene expression can be leveraged to develop more targeted interventions and preventative treatments.

The wealth of microbial data is generating new challenges as well as new opportunities for computational microbiology. Some predict that genomic data will become the foremost example of big data, outpacing astronomy and other data-intensive fields within the next ten years.[22] Algorithms that address this challenge will transform microbiology, but to do so they will need to be accurate, scalable, and wrapped in software accessible to and usable by biologists.

2. Challenges in Microbiology and Computational Approaches

We discuss existing challenges in microbiology, and highlight computational approaches that address these challenges. We focus primarily on those areas that have been transformed by the wealth of sequencing data now available.

2.1. *Gene molecular function and process prediction*

While DNA and RNA sequencing has become substantially easier and less costly, the process of understanding the function of genes remains difficult. This process of functional determination has been facilitated by computational algorithms that aim to automatically annotate functions based

on: the gene's nucleic acid sequence; the similarity of the gene's sequence to those with annotated functions;[23] how the gene is expressed;[24] the gene's interaction partners;[25,26] and other features.[27]

While there are many approaches for prediction, there are also many approaches for assessment, and the need for commonly accepted benchmarks has been highlighted as an area of need.[28] Recently, the Critical Assessment of Function Annotation (CAFA) was conducted to address this need.[29] While CAFA represents an important first step, the need for benchmark datasets, particularly those with comprehensive experimental validation and standardized assessment, remains high. This is particularly true in bacterial systems, which have not been well covered by CAFA challenges to date.[29] Ideally microbiologists will be able to both retrieve a best estimate for any gene of interest in an organism, and also receive a well-calibrated confidence score for that prediction.

2.2. *Microbes' molecular functionality and classification*

The overall sum of molecular functionality encoded in the genomes of microbes is representative of both their morphology and physiology – key features in bacterial taxonomic classification. Our interest in microbes is often focused on specific parts of their molecular abilities – their pathogenicity, toxicity and antibiotic resistance (to us and other species, *e.g.* for bio-pesticide purposes), as well as their ability to survive and thrive in extreme environments or with specific or limited nutrient sources (bioremediation and green energy). Thus, classification of microbes that implies similar treatment of similar organisms is important for industrial and clinical applications.

Current taxonomy is guided by evolutionary relationships,[30] which, however, ignores horizontal gene transfer (HGT) and, often, plasmid contributions and, therefore, does not guarantee functional similarity. Recent work[31] has shown the advantages of using microbial genome-guided predictions as proxies for functional comparisons. Microbial functional comparisons, informed by individual organisms' environmental preferences, highlight specific genes and functions responsible for particular environmental adaptations (*e.g.* functional studies of cyanobacteria clades identify sigma factors potentially responsible for salt tolerance).[31] However, despite significant recent efforts[32,33], only a third of the microbial genes (for which sequences are available) are explicitly functionally annotated,[31] and high-throughput experiments exploring temporal relationships between gene expressions are missing for the vast majority of (already fully sequenced) microorganisms, and annotations of molecular pathways are limited. Additionally, any available experimental tests only reflect a portion of overall bacterial functionality, with nearly three hundred tests only accessing 5–20% of the total functional potential.[30] Thus, significant further research is necessary to properly identify, describe, and use microbial functional abilities on a large scale. Within the confines of the current state and speed of the experimental art, computational approaches remain the sole, most significant means for producing new knowledge from existing data (*e.g.* computational studies on co-occurrence of specific functions encoded across genomes of organisms occupying similar environments could inform the necessary molecular pathways).

Microbial molecular functional abilities accurately reflect the environmental challenges faced by the individual subpopulations of microbes (ecotypes). In fact, the environment often has a more

pronounced effect on the microbial genomes than does vertical descent. We further expect that a function-based approach at exploring environmental impact will be even more relevant to the study of entire microbial communities in light of their emergent functionalities (*i.e.* functions that are available to the diversity of microorganisms together occupying a single niche, but not to each individual organism within that niche).

2.3. *Microbes' responses to their environment*

Microbes must respond to their environment to adapt to changing conditions such as nutrient availability, changes in a host, new members of their microbial community and many other factors.[34] Sequencing-based methods allow the transcriptomes of organisms to be measured without the potentially time consuming and costly array-design process that was required in the past.[35] This has allowed for assays of a diverse array of organisms, including many microbes. Such assays readily allow for differential expression analyses, in which genes are ordered by the extent to which they differ between conditions, isolates, or environments. While differential expression analyses play an important role, being able to integrate newly performed experiments into the context of existing data provides a key opportunity.

There are now more than 1.8 million genome-wide assays freely available in repositories such as ArrayExpress[32] and NCBI's Gene Expression Omnibus[33] (GEO). In total, these repositories contain experiments for more than 2000 different organisms (Fig. 1A). More than 150 species had more than 500 assays publicly available as of July 1, 2015 (Fig 1B). We anticipate that the number of organisms with large amounts of transcriptomic data will continue to grow. The transcriptomes of nearly 45,000 single cells were recently sequenced in one experiment,[36,37] surpassing the number of transcriptomes available for many organisms. While such techniques cannot yet be readily applied to bacteria, we expect that new approaches will become available and rapidly expand the diversity and scale of available transcriptomic datasets for microbiological systems.

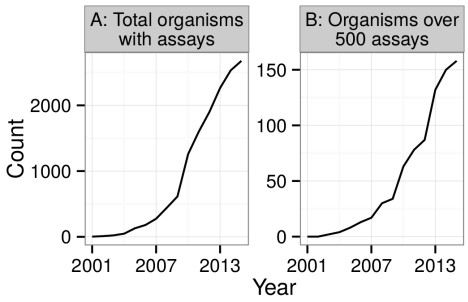

Fig. 1. The number of organisms with (A) genome wide data available and (B) those with more than 500 publicly available genome-wide expression assays. Counts for 2015 include assays from January through July.

We now have the opportunity to integrate and analyze these data to understand how the response to the environment in a specific or newly performed experiment relates to the response observed by others in past experiments. In well-characterized systems, data have been integrated using supervised methodologies that leverage extensively curated knowledgebases.[38,39] For many microbial systems, these knowledge-bases are limited or unavailable. To address this challenge we

need to develop unsupervised algorithms capable of integrating data with diverse information, ideally across multiple platforms. Such algorithms are now being developed,[40] but significant work remains to be done to apply these to large-scale microbial data compendia.

2.4. *Host-microbe and microbe-microbe interactions*

While adaptive immune responses in the host and evasion strategies of the microbe have been extensively studied, we are still discovering new mechanisms of host-microbe interactions. For example, Lee et al. identified genetic variants in a specific bitter taste receptor that were associated with susceptibility to respiratory infections, and that these receptors could be activated by compounds produced by *Pseudomonas aeruginosa*.[41] Subsequent studies have continued to reveal roles for taste receptors in the innate immune response.[42,43] Similar sensing of microbe-produced compounds have been reported in plants.[44] Computational techniques that facilitate the analysis of host genetics in combination with the composition and metagenomic characteristics of microbial communities may continue to identify additional novel mechanisms of host-microbe interactions.

2.5. *Membership in microbial communities*

Microbial communities have now been extensively profiled.[10,45–47] For communities on human hosts, the HMP has provided a large-scale survey across many available surfaces.[12] In addition to this large-scale assessment, numerous surveys have been made of microbial communities in a multitude of specific sites.[48–51] Such analyses have been performed in both healthy individuals and those experiencing a variety of conditions.[52–57]

2.5.1. *Heterogeneity across microbial communities*

Analysis of datasets from the HMP and others has raised numerous questions. For example, the abundance of microbial taxa varies substantially between individuals and body sites, but the relative abundance of metabolic modules within the communities remains consistent.[12,58] Studies of twins have revealed differences in similarity between monozygotic and dizygotic twins, suggesting that an individual's genetics affect his or her microbial communities.[59,60] This observation raises the question: what are the drivers of these differences, and what are their implications for both the community and the host?

2.5.2. *Heterogeneity within microbial populations*

Prior to the advent of large-scale inexpensive sequencing, microbial communities were assessed through sequencing of portions of the 16S ribosomal subunit that provided a family, genus, or species level of resolution.[61] Metagenomic analysis of both mixed species communities and single species populations provides the opportunity to identify genetic heterogeneity at the sub-species level within microbial communities. Such genomic diversification is rapid and common in biofilms[62,63] and chronic disease[64] and needs to be incorporated into our models for microbial communities. For example, traditional microbiological analyses have found that *P. aeruginosa* mutants with increased alginate production or the loss of quorum sensing regulation are commonly selected for in the lungs of individuals with CF, and the appearance of these mutants is associated

alterations in pathways known to be associated with host interactions[65] and have been associated with worse disease outcome[66]. The ability to associate genomic, transcriptomic, and phenotype or outcomes data will position us to understand which environmental factors drive the selection for certain variants and how these variants change the course of host-microbe and microbe-microbe interactions.

3. Conclusions

This is an exciting time in computational microbiology. Both the questions being asked and the experimental methodologies available to answer them are expanding in scope and diversity. We have highlighted a number of areas where we see particular opportunities for computational approaches. We anticipate that addressing these questions will require expertise in microbiology and the development, evaluation, and application of computational systems. The goal of our workshop is to provide a venue that brings these constituencies together.

4. Acknowledgments

C.S.G. is supported in part by the Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative through Grant GBMF4552 and by the National Science Foundation under Grant Number 1458390 to C.S.G. B.A.S. is supported by the CF Foundation, NIH R01-HL074175, NIH P30-GM106394, NIEHS P42-ES007373 and R01-HL074175. D.A.H. is supported by NIH R01-AI091702. Y.B. is supported by the USDA-NIFA grant 1015:0228906 and the Technische Universität München – Institute for Advanced Study Hans Fischer Fellowship, funded by the German Excellence Initiative and the European Union Seventh Framework Programme, grant agreement 291763. Our workshop is supported by NIH P30-GM106394.

References

1. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 6578–83 (1998).

2. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320,** 1034–9 (2008).

3. Dominguez-Bello, M. G. & Blaser, M. J. Asthma: Undoing millions of years of coevolution in early life? *Sci. Transl. Med.* **7,** 307fs39–307fs39 (2015).

4. Semenkovich, C. F., Danska, J., Darsow, T., Dunne, J. L., Huttenhower, C., Insel, R. A., McElvaine, A. T., Ratner, R. E., Shuldiner, A. R. & Blaser, M. J. American Diabetes Association and JDRF Research Symposium: Diabetes and the Microbiome. *Diabetes* (2015). doi:10.2337/db15-0597

5. Zhang, C., Yin, A., Li, H., Wang, R., Wu, G., Shen, J., Zhang, M., Wang, L., Hou, Y., Ouyang, H., Zhang, Y., Zheng, Y., Wang, J., Lv, X., Wang, Y., Zhang, F., Zeng, B., Li, W., Yan, F., *et al.* Dietary modulation of gut microbiota contributes to alleviation of both genetic and simple obesity in children. *EBioMedicine* **2,** 966–82 (2015).

6. Aas, J., Gessert, C. E. & Bakken, J. S. Recurrent Clostridium difficile colitis: case series involving 18 patients treated with donor stool administered via a nasogastric tube. *Clin. Infect. Dis.* **36,** 580–5 (2003).

7. Kassam, Z., Lee, C. H., Yuan, Y. & Hunt, R. H. Fecal microbiota transplantation for Clostridium difficile infection: systematic review and meta-analysis. *Am. J. Gastroenterol.* **108,** 500–8 (2013).

8. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24,** 133–41 (2008).

9. Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R. & Gordon, J. I. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–31 (2006).

10. Dinsdale, E. A., Edwards, R. A., Hall, D., Angly, F., Breitbart, M., Brulc, J. M., Furlan, M., Desnues, C., Haynes, M., Li, L., McDaniel, L., Moran, M. A., Nelson, K. E., Nilsson, C., Olson, R., Paul, J., Brito, B. R., Ruan, Y., Swan, B. K., *et al.* Functional metagenomic profiling of nine biomes. *Nature* **452,** 629–32 (2008).

11. Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R. & Gordon, J. I. The human microbiome project. *Nature* **449,** 804–10 (2007).

12. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–14 (2012).

13. Madan, J. C., Koestler, D. C., Stanton, B. A., Davidson, L., Moulton, L. A., Housman, M. L., Moore, J. H., Guill, M. F., Morrison, H. G., Sogin, M. L., Hampton, T. H., Karagas, M. R., Palumbo, P. E., Foster, J. A., Hibberd, P. L. & O'Toole, G. A. Serial analysis of the gut and respiratory microbiome in cystic fibrosis in infancy: interaction between intestinal and respiratory tracts and impact of nutritional exposures. *MBio* **3,** e00251–12– (2012).

14. Gifford, A. H., Alexandru, D. M., Li, Z., Dorman, D. B., Moulton, L. A., Price, K. E., Hampton, T. H., Sogin, M. L., Zuckerman, J. B., Parker, H. W., Stanton, B. A. & O'Toole, G. A. Iron supplementation does not worsen respiratory health or alter the sputum microbiome in cystic fibrosis. *J. Cyst. Fibros.* **13,** 311–8 (2014).

15. Carmody, L. A., Zhao, J., Kalikin, L. M., LeBar, W., Simon, R. H., Venkataraman, A., Schmidt, T. M., Abdo, Z., Schloss, P. D. & LiPuma, J. J. The daily dynamics of cystic fibrosis airway microbiota during clinical stability and at exacerbation. *Microbiome* **3,** 12 (2015).

16. LiPuma, J. J. Assessing Airway Microbiota in Cystic Fibrosis: What More Should Be Done? *J. Clin. Microbiol.* **53,** 2006–2007 (2015).

17. Garcia-Nuñez, M., Millares, L., Pomares, X., Ferrari, R., Pérez-Brocal, V., Gallego, M., Espasa, M., Moya, A. & Monsó, E. Severity-related changes of bronchial microbiome in chronic obstructive pulmonary disease. *J. Clin. Microbiol.* **52,** 4217–23 (2014).

18. Huang, Y. J., Nariya, S., Harris, J. M., Lynch, S. V, Choy, D. F., Arron, J. R. & Boushey, H. The airway microbiome in patients with severe asthma: Associations with disease features and severity. *J. Allergy Clin. Immunol.* (2015). doi:10.1016/j.jaci.2015.05.044

19. Hoen, A. G., Li, J., Moulton, L. A., O'Toole, G. A., Housman, M. L., Koestler, D. C., Guill, M. F., Moore, J. H., Hibberd, P. L., Morrison, H. G., Sogin, M. L., Karagas, M. R. & Madan, J. C. Associations between Gut Microbial Colonization in Early Life and Respiratory Outcomes in Cystic Fibrosis. *J. Pediatr.* **167,** 138–147.e3 (2015).

20. Delzenne, N. M., Cani, P. D., Everard, A., Neyrinck, A. M. & Bindels, L. B. Gut microorganisms as promising targets for the management of type 2 diabetes. *Diabetologia* **58,** 2206–17 (2015).

21. Filkins, L. M., Hampton, T. H., Gifford, A. H., Gross, M. J., Hogan, D. A., Sogin, M. L., Morrison, H. G., Paster, B. J. & O'Toole, G. A. Prevalence of Streptococci and Increased Polymicrobial Diversity Associated with Cystic Fibrosis Patient Stability. *J. Bacteriol.* **194,** 4709–4717 (2012).

22. Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S. & Robinson, G. E. Big Data: Astronomical or Genomical? *PLOS Biol.* **13,** e1002195 (2015).

23. Wilson, C. A., Kreychman, J. & Gerstein, M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297,** 233–49 (2000).

24. Greene, C. S. & Troyanskaya, O. G. PILGRM: an interactive data-driven discovery platform for expert biologists. *Nucleic Acids Res.* **39,** W368–W374 (2011).

25. Marcotte, E. M., Pellegrini, M., Ng, H.-L., Rice, D. W., Yeates, T. O. & Eisenberg, D. Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. *Science (80-.).* **285,** 751–753 (1999).

26. Vazquez, A., Flammini, A., Maritan, A. & Vespignani, A. Global protein function prediction from protein-protein interaction networks. *Nat. Biotechnol.* **21,** 697–700 (2003).

27. Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B. & Botstein, D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc. Natl. Acad. Sci. U. S. A.* **100,** 8348–53 (2003).

28. Murali, T. M., Wu, C.-J. & Kasif, S. The art of gene function prediction. *Nat. Biotechnol.* **24,** 1474–5; author reply 1475–6 (2006).

29. Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., Pandey, G., Yunes, J. M., Talwalkar, A. S., Repo, S., Souza, M. L., Piovesan, D., Casadio, R., Wang, Z., Cheng, J., *et al.* A large-scale evaluation of computational protein function prediction. *Nat. Methods* **10,** 221–7 (2013).

30. Garrity, G., Boone, D. R. & Castenholz, R. W. *Bergey's Manual of Systematic Bacteriology Volume 1: The Archaea and the Deeply Branching and Phototrophic Bacteria*. (Springer, 2001). doi:10.1007/0-387-29298-5

31. Zhu, C., Delmont, T. O., Vogel, T. M. & Bromberg, Y. Functional Basis of Microorganism Classification. *PLoS Comput. Biol.* **11,** e1004472 (2015).

32. Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Ison, J., Keays, M., Kurbatova, N., Malone, J., Mani, R., Mupo, A., Pedro Pereira, R., Pilicheva, E., Rung, J., Sharma, A., Tang, Y. A., *et al.* ArrayExpress update--trends in database growth and links to data analysis tools. *Nucleic Acids Res.* **41,** D987–90 (2013).

33. Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Muertter, R. N., Holko, M., Ayanbule, O., Yefanov, A. & Soboleva, A. NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic Acids Res.* **39,** D1005–10 (2011).

34. Dey, N., Wagner, V. E., Blanton, L. V., Cheng, J., Fontana, L., Haque, R., Ahmed, T. & Gordon, J. I. Regulators of Gut Motility Revealed by a Gnotobiotic Model of Diet-Microbiome Interactions Related to Travel. *Cell* **163,** 95–107 (2015).

35. Ekblom, R. & Galindo, J. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity (Edinb).* **107,** 1–15 (2011).

36. Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A. & McCarroll, S. A. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–1214 (2015).

37. Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A. & Kirschner, M. W. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161,** 1187–1201 (2015).

38. Wong, A. K., Park, C. Y., Greene, C. S., Bongo, L. A., Guan, Y. & Troyanskaya, O. G. IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. *Nucleic Acids Res.* **40,** W484–90 (2012).

39. Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., Zhang, R., Hartmann, B. M., Zaslavsky, E., Sealfon, S. C., Chasman, D. I., FitzGerald, G. A., Dolinski, K., Grosser, T. & Troyanskaya, O. G. Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47,** 569–576 (2015).

40. Tan, J., Ung, M., Cheng, C. & Greene, C. S. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. *Pacific Symp. Biocomput.* 132–43 (2015).

41. Lee, R. J., Xiong, G., Kofonow, J. M., Chen, B., Lysenko, A., Jiang, P., Abraham, V., Doghramji, L., Adappa, N. D., Palmer, J. N., Kennedy, D. W., Beauchamp, G. K., Doulias, P.-T., Ischiropoulos, H., Kreindler, J. L., Reed, D. R. & Cohen, N. A. T2R38 taste receptor polymorphisms underlie susceptibility to upper respiratory infection. *J. Clin. Invest.* **122,** 4145–59 (2012).

42. Lee, R. J., Kofonow, J. M., Rosen, P. L., Siebert, A. P., Chen, B., Doghramji, L., Xiong, G., Adappa, N. D., Palmer, J. N., Kennedy, D. W., Kreindler, J. L., Margolskee, R. F. & Cohen, N. A. Bitter and sweet taste receptors regulate human upper respiratory innate immunity. *J. Clin. Invest.* **124,** 1393–405 (2014).

43. Lee, R. J. & Cohen, N. A. Role of the bitter taste receptor T2R38 in upper respiratory infection and chronic rhinosinusitis. *Curr. Opin. Allergy Clin. Immunol.* **15,** 14–20 (2015).

44. Hartmann, A., Rothballer, M., Hense, B. A. & Schröder, P. Bacterial quorum sensing compounds are important modulators of microbe-plant interactions. *Front. Plant Sci.* **5,** 131 (2014).

45. Xie, W., Wang, F., Guo, L., Chen, Z., Sievert, S. M., Meng, J., Huang, G., Li, Y., Yan, Q., Wu, S., Wang, X., Chen, S., He, G., Xiao, X. & Xu, A. Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries. *ISME J.* **5,** 414–26 (2011).

46. Mackelprang, R., Waldrop, M. P., DeAngelis, K. M., David, M. M., Chavarria, K. L., Blazewicz, S. J., Rubin, E. M. & Jansson, J. K. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* **480,** 368–71 (2011).

47. Pope, P. B., Mackenzie, A. K., Gregor, I., Smith, W., Sundset, M. A., McHardy, A. C., Morrison, M. & Eijsink, V. G. H. Metagenomics of the Svalbard reindeer rumen microbiome reveals abundance of polysaccharide utilization loci. *PLoS One* **7,** e38571 (2012).

48. Segata, N., Haake, S. K., Mannon, P., Lemon, K. P., Waldron, L., Gevers, D., Huttenhower, C. & Izard, J. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol.* **13,** R42 (2012).

49. Willner, D., Haynes, M. R., Furlan, M., Hanson, N., Kirby, B., Lim, Y. W., Rainey, P. B., Schmieder, R., Youle, M., Conrad, D. & Rohwer, F. Case studies of the spatial heterogeneity of DNA viruses in the cystic fibrosis lung. *Am. J. Respir. Cell Mol. Biol.* **46,** 127–31 (2012).

50. Zhou, X., Brown, C. J., Abdo, Z., Davis, C. C., Hansmann, M. A., Joyce, P., Foster, J. A. & Forney, L. J. Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women. *ISME J.* **1,** 121–33 (2007).

51. Hunt, K. M., Foster, J. A., Forney, L. J., Schütte, U. M. E., Beck, D. L., Abdo, Z., Fox, L. K., Williams, J. E., McGuire, M. K. & McGuire, M. A. Characterization of the Diversity and Temporal Stability of Bacterial Communities in Human Milk. *PLoS One* **6,** e21313 (2011).

52. Erb-Downward, J. R., Thompson, D. L., Han, M. K., Freeman, C. M., McCloskey, L., Schmidt, L. A., Young, V. B., Toews, G. B., Curtis, J. L., Sundaram, B., Martinez, F. J. & Huffnagle, G. B. Analysis of the lung microbiome in the 'healthy' smoker and in COPD. *PLoS One* **6,** e16384 (2011).

53. Spencer, M. D., Hamp, T. J., Reid, R. W., Fischer, L. M., Zeisel, S. H. & Fodor, A. A. Association between composition of the human gastrointestinal microbiome and development of fatty liver with choline deficiency. *Gastroenterology* **140,** 976–86 (2011).

54. Larsen, N., Vogensen, F. K., van den Berg, F. W. J., Nielsen, D. S., Andreasen, A. S., Pedersen, B. K., Al-Soud, W. A., Sørensen, S. J., Hansen, L. H. & Jakobsen, M. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS One* **5,** e9085 (2010).

55. Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490,** 55–60 (2012).

56. Greenblum, S., Turnbaugh, P. J. & Borenstein, E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 594–9 (2012).

57. Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V, Reyes, J. A., Shah, S. A., LeLeiko, N., Snapper, S. B., Bousvaros, A., Korzenik, J., Sands, B. E., Xavier, R. J. & Huttenhower, C. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* **13,** R79 (2012).

58. Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M., Henrissat, B., White, O., Kelley, S. T., Methé, B., Schloss, P. D., Gevers, D., Mitreva, M. & Huttenhower, C. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* **8,** e1002358 (2012).

59. Goodrich, J. K., Waters, J. L., Poole, A. C., Sutter, J. L., Koren, O., Blekhman, R., Beaumont, M., Van Treuren, W., Knight, R., Bell, J. T., Spector, T. D., Clark, A. G. & Ley, R. E. Human Genetics Shape the Gut Microbiome. *Cell* **159,** 789–799 (2014).

60. Hampton, T. H., Green, D. M., Cutting, G. R., Morrison, H. G., Sogin, M. L., Gifford, A. H., Stanton, B. A. & O'Toole, G. A. The microbiome in pediatric cystic fibrosis patients: the role of shared environment suggests a window of intervention. *Microbiome* **2,** 14 (2014).

61. Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L. & Pace, N. R. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl. Acad. Sci.* **82,** 6955–6959 (1985).

62. Traverse, C. C., Mayo-Smith, L. M., Poltak, S. R. & Cooper, V. S. Tangled bank of experimentally evolved Burkholderia biofilms reflects selection during chronic infections. *Proc. Natl. Acad. Sci. U. S. A.* **110,** E250–9 (2013).

63. Penterman, J., Nguyen, D., Anderson, E., Staudinger, B. J., Greenberg, E. P., Lam, J. S. & Singh, P. K. Rapid Evolution of Culture-Impaired Bacteria during Adaptation to Biofilm Growth. *Cell Rep.* **6,** 293–300 (2014).

64. Sousa, A. M. & Pereira, M. O. Pseudomonas aeruginosa Diversification during Infection Development in Cystic Fibrosis Lungs-A Review. *Pathog. (Basel, Switzerland)* **3,** 680–703 (2014).

65. Hammond, J. H., Dolben, E. F., Smith, T. J., Bhuju, S. & Hogan, D. A. Links between Anr and quorum sensing in Pseudomonas aeruginosa biofilms. *J. Bacteriol.* JB.00182–15– (2015). doi:10.1128/JB.00182-15

66. Hoffman, L. R., Kulasekara, H. D., Emerson, J., Houston, L. S., Burns, J. L., Ramsey, B. W. & Miller, S. I. Pseudomonas aeruginosa lasR mutants are associated with cystic fibrosis lung disease progression. *J. Cyst. Fibros.* **8,** 66–70 (2009).