UNIVERSITY OF CALIFORNIA SAN DIEGO

Characterization of the gut microbial community and *Escherichia coli* in inflammatory bowel disease

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioengineering

by

Xin Fang

Committee in charge:

Professor Bernhard O. Palsson, Chair
Professor Rob Knight
Professor Joseph Pogliano
Professor Larry Smarr
Professor Karsten Zengler

2020

The dissertation of Xin Fang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2020

DEDICATION

To my parents who made all this possible

EPIGRAPH

Science is the acceptance of what works and the rejection of what does not. That needs more courage than we might think

–JACOB BRONOWSKI

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

My journey as a PhD student would not be possible if it was not for the support of my mentors, colleagues, friends and family.

First I'd like to thank Dr. Palsson, my PhD advisor and mentor, who has helped me tremendously during my PhD study. He has not only offered his knowledge and guidance for all my projects, but has also provided me a great platform and resources to work with some of the most talented scientists in the field. He gave me unconditional support when I wanted to work on the gut microbiome projects, and encouraged me to pursue my interest in translational medicine during both my clinical rotation and internship program. I am very grateful to have been a part of the Palsson lab for my PhD training - it has helped me grow so much both professionally and personally.

I would also like to thank Larry, another very important mentor of mine during my PhD program, who has opened the door for me to microbiome science. His passion and curiosity for science really inspired many scientists like myself to answer complex questions using biological data. He has also looked out for me and helped me in so many aspects - I have learned from him how to be a true leader and scientist.

I'd also like to thank my other committee members. I am very honored to have worked with Dr. Knight, who has selflessly shared data generated from his lab, resources, and his expertise on microbiome. He has given me critical feedback, guidance and support in multiple projects and made my gut microbiome projects possible. I want to also thank Dr. Zengler for giving me suggestion on my projects and guidance in professional development, and Dr. Pogliano for his constructive feedback on my projects.

Laurence Yang and Jonathan Monk, my two mentors in the lab who worked with me

closely have taught me so much in research. I am grateful for Laurence's patience when I get started on my first project in graduate school - he has spent so much time teaching me all the basics in research and sharing with me his knowledge selflessly. Jon has worked with me on multiple projects, and he has offered lots of help both in research and writing - he has always been so encouraging and delightful to work with.

I'd like to thank my other mentors, co-workers and collaborators including but not limited to: Yoshiki Vazquez, Brigid Boland, William Sandborn, Sergey Nurk, Anand Sastry, Charles Norsigian, James Yurkovich, Nathan Mih, Erol Kavvas, Yara Seif and Qiyun Zhu

Last but not least, I'd like to thank my families and friends who have been here for me during my PhD career, especially my San Diegan friends: Kai Chen, Jason Zhang, Evan Teng, Rui Guo, Yunke Yang and the badminton crew, who have shared much laughter and provided emotional support for me in the past five years. I'd also like to say a big thank you to my parents, who have being my greatest supporters and role models despite the distance - they have showed me what it takes to be real scientists through their own careers: passion, diligence, and dedication, and they have given me valuable advice as I navigate through life. Lastly, I'd like to thank my fiancee Bin Du, who has been here for me for the past 5 years through ups and downs. His patience, companionship and love has helped me overcome many obstacles along the way and grow significantly.

Chapter 2 in full is a reprint of material submitted for publication: **Fang, X. ***, Vázquez-Baeza, Y. *, Elijah, E., Vargas, F., Ackermann, G., Humphrey, G., Lau, R., Weldon K. C., Sanders, J. G., Panitchpakdi, M., Carpenter, C., Neill, J., Miralles A., Dulai, P., Singh, S., Tsai, M., Swafford, A. D., Smarr, L., Boyle, D. L., Palsson B. O., Chang, J. T., Dorrestein, P. C., Sandborn W. J., Knight, R., Boland, B. S. Gastrointestinal surgery for inflammatory bowel

disease persistently lowers microbiome and metabolome diversity. *Inflammatory Bowel Disease*, Submitted. The dissertation author was one of the primary authors.

Chapter 3 in full is a reprint of material published in: **Fang, X.**, Monk, J. M., Nurk, S., Akseshina, M., Zhu, Q., Gemmell, C., Gianetto-Hill, C., Leung, N., Szubin, R., Sanders, J., Beck, P. L., Li, W., Sandborn, W. J., Gray-Owen, S. D., Knight, R., Allen-Vercoe, E., Palsson, B. O., Smarr, L. (2018). Metagenomics-Based, Strain-Level Analysis of *Escherichia coli* From a Time-Series of Microbiome Samples From a Crohn's Disease Patient. *Frontiers in Microbiology*, 9, 2559. The dissertation author was the primary author.

Chapter 4 in full is a reprint of material published in: **Fang, X.**, Monk, J. M., Mih, N., Du, B., Sastry, A. V., Kavvas, E., Seif, Y., Smarr, L., Palsson, B. O. (2018). *Escherichia coli* B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of intestinal mucosa. *BMC Systems Biology*, 12(1), 66. The dissertation author was the primary author.

Chapter 5 in full is a reprint of material published in: **Fang, X. ***, Sastry, A.*, Mih, N., Kim, D., Tan, J., Yurkovich, J. T., Lloyd, C. J., Gao, Y., Yang, L., Palsson, B. O. (2017). Global transcriptional regulatory network for *Escherichia coli* robustly connects gene expression to transcription factor activities. *Proceedings of the National Academy of Sciences of the United States of America*, 114(38), 10286–10291. The dissertation author was the one of the primary authors.

VITA

2014	Bachelor of Science in Chemical and Biomolecular Engineering, Johns Hopkins University
2014 - 2015	Co-op in Early Stage Drug Development, Genentech
2019.06 - 2019.08	Transnational Bioinformatics Intern, Bristol-Myers Squibb
2020	Doctor of Philosophy in Bioengineering, University of California San Diego

PUBLICATIONS

Fang, X. *, Sastry, A.*, Mih, N., Kim, D., Tan, J., Yurkovich, J. T., Lloyd, C. J., Gao, Y., Yang, L., Palsson, B. O. (2017). Global transcriptional regulatory network for *Escherichia coli* robustly connects gene expression to transcription factor activities. *Proceedings of the National Academy of Sciences of the United States of America*, 114(38), 10286–10291.

Fang, X., Monk, J. M., Mih, N., Du, B., Sastry, A. V., Kavvas, E., Seif, Y., Smarr, L., Palsson, B. O. (2018). *Escherichia coli* B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of intestinal mucosa. *BMC Systems Biology*, 12(1), 66.

Fang, X., Monk, J. M., Nurk, S., Akseshina, M., Zhu, Q., Gemmell, C., Gianetto-Hill, C., Leung, N., Szubin, R., Sanders, J., Beck, P. L., Li, W., Sandborn, W. J., Gray-Owen, S. D., Knight, R., Allen-Vercoe, E., Palsson, B. O., Smarr, L. (2018). Metagenomics-Based, Strain-Level Analysis of *Escherichia coli* From a Time-Series of Microbiome Samples From a Crohn's Disease Patient. *Frontiers in Microbiology*, 9, 2559.

Fang, X. *, Vázquez-Baeza, Y. *, Elijah, E., Vargas, F., Ackermann, G., Humphrey, G., Lau, R., Weldon K. C., Sanders, J. G., Panitchpakdi, M., Carpenter, C., Neill, J., Miralles A., Dulai, P., Singh, S., Tsai, M., Swafford, A. D., Smarr, L., Boyle, D. L., Palsson B. O., Chang, J. T., Dorrestein, P. C., Sandborn W. J., Knight, R., Boland, B. S. Gastrointestinal surgery for inflammatory bowel disease persistently lowers microbiome and metabolome diversity. *Inflammatory Bowel Disease*, Submitted

Fang, X. and Palsson B. O. Reconstructing microorganism *in silico*: from biochemical reactions to pan-genome analysis. *Nature Reviews Microbiology*, Submitted

Norsigian C.*, **Fang, X. ***, Seif, Y., Monk, J. M., Palsson, B. O. A protocol for generating multi-strain genome-scale metabolic models. (2019) *Nat Protocols*, 114(38), 15, 1–14 (2020) doi:10.1038/s41596-019-0254-3

Shiratsubaki, I. S.*, **Fang, X. ***, Souza, O. O., Palsson, B. O., Silber, A. M., Siquieira-Neto, J. L. Genome-scale metabolic models highlight stage-specific differences in essential metabolic pathways in *Trypanosoma cruzi PLOS Neglected Tropical Diseases*, Submitted

Seif, Y., Kavvas, E., Lachance, J. Yurkovich, J. T., Nuccio, S., **Fang, X.**, Catoiu, E., Raffatellu, M.,Palsson, B. O.,Monk, J. M. (2018). Genome-scale metabolic reconstructions of multiple *Salmonella* strains reveal serovar-specific metabolic traits. *Nature Communications*, 9, 3771

Santos-Zavaleta, A., Sánchez-Pérez, M., Salgado, H., Velázquez-Ramírez, D. A., Gama-Castro, S., Tierrafría, Busby, S. JW., Aquino, P., **Fang, X.**, Palsson, B. O., Galagan, J. E., Collado-Vides, J. (2018). A unified resource for transcriptional regulation in *Escherichia coli* K-12 incorporating high-throughput-generated binding data into RegulonDB version 10.0. *BMC Biology*, 16, 91.

Gao, Y., Yurkovich, J., Seo, SW., Kabimoldayev, I., Drager, A., Chen, K., Sastry, A. V., **Fang, X.**, Mih, N., Yang, L., Eichner, J., Cho, BK., Kim, D., Palsson, B. O. (2018) Systematic discovery of uncharacterized transcription factors in *Escherichia coli* K-12 MG1655, *Nucleic Acids Research*, gky752.

Du, B., Olson, C. A., Sastry, A. V., **Fang, X.**, Phaneuf, P. V., Chen, K., Wu, M., Szubin, R., Xu, S., Gao, Y., Hefner, Y., Feist, A. M., Palsson, B. O. (2019) Adaptive laboratory evolution of *Escherichia coli* under acid stress. *Microbiology* mic.0.000867

Du, B., Yang. L., Lloyd. C. J., **Fang, X.**, Palsson, B. O. (2019) Genome-scale model of metabolism and gene expression provides a multi-scale description of acid stress responses in *Escherichia coli*, *PLOS Computational Biology* 15(12): e1007525.

Jensen, C. S., Norsigian, C. J., **Fang, X.**, Nielsen, X. C., Christensen, J. J., Palsson, B. O., Monk, J. M. (2020). Reconstruction and Validation of a Genome-Scale Metabolic Model of *Streptococcus oralis* (iCJ415), a Human Commensal and Opportunistic Pathogen. *Frontiers in genetics*, 11, 116. https://doi.org/10.3389/fgene.2020.00116

Mih, N., Monk, J. M., **Fang, X.**, Catoiu, E., Heckmann, D., Yang, L., Palsson, B. O. (2020) Adaptations of *Escherichia coli* strains to oxidative stress are reflected in properties of their structural proteomes *BMC Bioinformatics* 21, 162 (2020). https://doi.org/10.1186/s12859-020-3505-y

* equal contribution

ABSTRACT OF THE DISSERTATION

Characterization of the gut microbial community and *Escherichia coli* in inflammatory bowel disease

by

Xin Fang

Doctor of Philosophy in Bioengineering

University of California San Diego, 2020

Professor Bernhard O. Palsson, Chair

Dysbiosis of the gut microbiome, including elevated abundance of putative bacterial triggers, such as *Escherichia coli (E. coli)*, is observed in inflammatory bowel disease (IBD). In this dissertation, we characterized the gut microbial community, one of its members - *E. coli*, and their implications in IBD. First, the evaluation of the entire gut microbial community of a cohort of IBD patients suggested that intestinal surgery has a significant impact on gut microbiome, including lowered diversity and stability, changes in bile acid levels and elevated *E. coli* abundance. This result calls for systematic evaluation of IBD treatment and careful consideration of treat-

ment options. We then focused on *E. coli* through extracting *de novo* assemblies of dominant *E. coli* strains from time-series metagenomics data of an IBD patient. Analysis suggest that the *E. coli* community is highly dynamic with changing dominant strains, and certain strain-specific features may be correlated with pathogenicity and disease progression. Third, we characterized the metabolic functions of *E. coli* clinical isolates from IBD patients using comparative genomics analysis and genome-scale models. We identified metabolic genes that are specific to strains in B2 phylogroup that are more prevalent in IBD patients, which potentially enable colonization to human gut. Lastly, we evaluated the most updated transcription regulatory network (TRN) of *E. coli*, as it enables adaptation to various conditions including the inflamed digestive tract of IBD patients. We found that the TRN has robust core functional modules, and has significantly expanded in the past decade, but still has limited coverage, motivating more high-throughput experiments to fill in knowledge gaps. In conclusion, this dissertation broadened the understanding of *E. coli* and gut microbiome in IBD, and provided valuable insight for clinical practice and potential intervention strategies.

Chapter 1

Introduction

1.1 Inflammatory Bowel disease and the dysbiotic gut microbiome

Inflammatory Bowel Disease (IBD) is a complex disease mediated by immune systems, which is usually manifested by chronic inflammation of the digestive tract [1]. Two subtypes of IBD exist - Crohn's Disease (CD) that can affect any part of the digestive tract, but mostly affect distal ileum and colon; and Ulcerative Colitis (UC) that only affects the colon. IBD affects around 1.5 million Americans, 2.2 million people in Europe and even more worldwide. The incidence of IBD was traditionally more common in developed countries in Europe and North America. However, in recent years, an increasing number of IBD cases have appeared in other regions such as Asia, likely due to urbanization and improved hygiene conditions [2].

Various types of treatment have been developed for IBD, yet as of today IBD cannot be cured. It can only be managed to minimize symptoms. Commonly used medications of IBD include anti-inflammatory agents such as corticosteroids, aminosalicylates and other im-

munomodulators [3, 4]. Antibiotics including ciprofloxacin and metronidazole are also effective in killing pathogens that may be responsible for inducing inflammatory responses. Biological agents such as anti-TNF and Integrin-inhibitors have also been proven effective in managing the symptoms. Surgical treatment is a common option for patients that do not respond to medication treatments, which usually involves resection of the inflamed parts of the digestive tract [3, 4].

IBD is an extremely complex disease that has many risk factors, including but not limited to host genetics, environmental factors, lifestyles and gut microbiome [2]. Previous studies have identified more than 200 risk locis in IBD patients that could potentially predispose them to this disease [5]. However, environmental factors as we discussed in an earlier paragraph such as hygiene conditions, as well as lifestyle factors including stress level, smoking and diet were all shown to contribute to IBD development [2]. Previous life experiences including breastfeeding and exposure to antibiotics may also be related to IBD [2].

In this dissertation we focus on the dysbiotic gut microbiome of IBD, as inflammation likely is a result of the dysregulated immune response to the gut flora in a susceptible host. Dysbiosis of the gut microbiome has been consistently observed in both UC and CD patients, which usually involves the reduced bacterial diversity, an increase in Proteobacteria level, and decline in Firmicutes, which is usually a major member of the gut microbiome in healthy individuals [6]. Specifically, *E. coli* - a member of the Proteobacteria group is considered one of the potential bacterial triggers in IBD and usually has an elevated abundance in IBD patients compared to healthy individuals. Specifically, a pathotype of *E. coli* - adherent-invasive *E. coli* (AIEC) has been implicated in IBD. Strains belonging to this pathotype are able to attach to intestinal epithelial cells and survive and replicate within macrophages, yet no unique genetic determinant has been identified in this pathotype [7]. In this dissertation, we aim to deepen our

understanding of the IBD gut microbiome and the implication of *E. coli* in IBD.

1.2 Studying gut microbiome using next-generation sequencing

Two commonly used DNA-based approaches to study microbial communities are based on gene amplicon or marker genes (e.g. 16S rRNA) and shotgun metagenomics data. 16S rRNA studies were used as the primary methodology to study the taxonomic composition of the microbial community in the earlier days of microbiome studies, due to the high cost in sequencing in the past [8]. 16S rRNA are present in all living organisms, and it is a commonly used marker gene because it has both conserved regions that are good candidates for PCR primers and fast-evolving regions that can differentiate between organisms. While 16S data is useful in revealing taxonomic profiles in microbial communities, it has limited taxonomic resolution and potential to characterize functional profiles of the target community [9]. Therefore, researchers have slowly migrated to using whole metagenomics sequencing data.

Shotgun metagenomics data has been gaining its popularity in recent years due to the development of sequencing technology, bioinformatic softwares and drop in sequencing cost. Instead of focusing only on the marker genes, shotgun metagenomics data targets the whole genome, therefore has a higher cost in data generation. When given adequate sequencing depth, it can produce much more detailed information on the microbiome including species or strain level taxonomic profiles, assembly of the whole metagenome and functional annotation of genes and pathways [10]. Metagenomics data can also be used to extract novel genomes of unknown organisms [11]. In two of the projects in this dissertation, we utilized metagenomics data generated from the stool samples from IBD patients.

For organisms with known culture conditions, it is also possible to isolate, culture and

sequence them from the gut microbial community, as did for *E. coli* clinical isolates in IBD in this dissertation [12, 13]. The genome sequences of these isolates are usually more accurate and cheaper to generate than the *de novo* assemblies extracted from metagenomics data. These strains isolated also enable further experimental characterizations such as growth capabilities, invasion abilities, and other phenotypes of interest [12, 13].

More next-generation data types are also on the horizon to be used to study the microbiome. Metatranscriptomics data uses RNA sequencing to delineate the gene expression levels of the organisms in the microbiome. It differs from the metagenomics data, as metatranscriptomics data describes the functional output of the microbiome, while metagenomics data characterizes the functional potentials of the gut microbial community [9]. There have also been attempts in generating long reads or using hybrid approaches instead of traditional short read sequencing to produce metagenomics data. These approaches may have great potential to reduce the assembly error rate, yet challenges still need to be addressed before they can be adopted widely [8].

1.3 Interpreting next-generation sequencing data using systems biology approaches

In addition to applying traditional statistical analysis to the next-generation sequencing data such as diversity analysis, identification of differentially abundant species, genes and pathways, we also incorporated systems-biology approaches into the projects in this dissertation, such as the GEnome-scale Models (GEMs).

Genome-scale network reconstructions are built from curated and systematized knowledge [14, 15] that enables them to quantitatively describe genotype-phenotype relationships. GEMs

are mathematical representations of reconstructed networks that facilitate computation and prediction of multi-scale phenotypes through the optimization of an objective function of interest [16, 17]. GEMs have been successfully implemented for a wide range of applications [18, 19], including metabolic engineering [20], drug development [21], prediction of enzyme functions [22], understanding community interactions [23], and human disease [24, 25].

Flux balance analysis (FBA) is the most widely used [26] approach to characterize GEMS. GEMs are able to simulate metabolic flux while incorporating multiple constraints to ensure the feasibility of a simulated phenotype, such as the metabolic network topology, a steady-state assumption (e.g., the internal metabolites must be produced and consumed in a mass-balanced manner), and other bounds on reaction flux (e.g., nutrient uptake rates, enzyme capacities, protein/gene expression). FBA can identify a single or multiple optimal flux distributions that optimize the objective function in the solution space. FBA and many other GEM analysis methods are available through COBRApy [27] in python or COBRA Toolbox in MATLAB [28].

In this dissertation, we built GEMs for *E. coli* strains of interest from IBD patients, either based on the genome sequences of clinical isolates or the *de novo* assembly from metagenomics data. GEMs of *E. coli* strains were used to generate growth predictions on different nutrient sources, depict the metabolic functions and provide mechanistic insights. GEMs have been shown to be a valuable tool to interpret genomic and metagenomic analysis in the context of known knowledge bases, and provide deeper understanding of the organisms of interest.

GEM-based community modeling workflows and reconstructions are also being developed to understand the complex interactions within gut microbiome. 773 GEMs for human gut bacteria were generated to enable the exploration of microbial community metabolism [29]. A workflow has also been developed to construct personalized gut microbiome community models based on

the metagenomics data [30]. Although this approach was not used in this dissertation, this could be a promising future direction to explore.

Chapter 2

Intestinal surgery impact gut microbiome in inflammatory bowel disease

2.1 Abstract

2.1.1 Background

Many studies have investigated the role of the microbiome in inflammatory bowel disease (IBD), but few have focused on surgery specifically, or its consequences on the metabolome that may differ by surgery type and require longitudinal sampling. Our objective was to characterize and contrast microbiome and metabolome changes following different surgeries for IBD, including ileocolonic resection and colectomy.

2.1.2 Methods

The UC San Diego IBD Biobank was used to prospectively collect 332 stool samples from 129 subjects (50 ulcerative colitis; 79 Crohn's disease). Of these, 21 with Crohn's disease had ileocolonic resections, and 17 had colectomies. We used shotgun metagenomics and untargeted LC/MS/MS metabolomics to characterize the microbiomes and metabolomes of these patients up to 24 months after the initial sampling.

2.1.3 Results

The species diversity and metabolite diversity both differed significantly among groups (species diversity: Mann-Whitney U test P-value = 7.8e-17; metabolomics: P-value =0.0043). *E. coli* in particular expanded dramatically in relative abundance in subjects undergoing surgery. The species profile was better able to classify subjects according to surgery status than the metabolite profile (average precision 0.80 vs 0.68).

2.1.4 Conclusions

Intestinal surgeries reduce the diversity of the gut microbiome and metabolome in IBD patients, and these changes may persist for years. Surgery also further destabilizes the microbiome (but not the metabolome) over time, even relative to the previously established instability in the microbiome of IBD patients. These long-term effects and their consequences for health outcomes need to be studied in prospective longitudinal trials linked to microbiome-involved phenotypes.

2.2 Background

The role of the microbiome [31–33] and metabolome [6, 34] in inflammatory bowel disease (IBD) has been well established, and many studies have reported specific biomarkers of ulcerative colitis (UC), Crohn's disease (CD) at the microbiome [35] or metabolome level [36]. Prior work on microbiome dynamics in IBD includes the intriguing observation that ileal Crohn's patients with resection have especially unstable stool microbiome dynamics [37, 38]; however, prior studies have not investigated the impacts on the microbiome associated with different types of intestinal surgery in detail. Notably, a prospective study of 20 patients with ileal CD undergoing ileocolonic resection showed that there was no change in microbial diversity six months after surgery; however, the microbial community structure was altered in the setting of endoscopic recurrence [39].

Investigating the effect of surgery is important for several reasons. First, an increasing body of evidence on dysbiosis suggests that occasional excursions into deleterious regions of microbiome space are important for triggering adverse events[37]. Second, surgery is seen from a clinical perspective as a way to manage IBD, but is irreversible, and the long-term adverse effects on the microbiome that may worsen disease have not been studied extensively [37, 39]. Third, in general it is not known which therapies for IBD have large versus small effects on the microbiome, and surgery has been especially poorly studied in this respect. Studies of the microbiome in other disease areas, most notably diabetes, have shown significant effects of treatment that can be confounded by consequences of a disease, particularly when the treatment effects are unknown [40].

Although 16S rRNA amplicon analysis has been a very useful tool for revealing microbiome differences and dynamics in IBD [33, 37], there are several important limitations in terms

of taxonomic resolution and insight into function [41]. To overcome these limitations in this study, we perform deep-coverage shotgun metagenomics, allowing species- and strain-level profiling and functional analysis of the microbiome, and untargeted metabolomics with LC/MS/MS (liquid chromatography followed by tandem mass spectrometry), giving a direct chemical readout of the metabolome profile. This combination of techniques allows us to assess the value of these different data types as biomarkers for clinical states, including microbiome volatility and clinical status.

An important clinical consideration is the role and timing of surgery in the treatment of IBD. While surgery is typically reserved as a last resort for medically refractory disease, a randomized trial compared medical therapy to surgery early in the treatment of ileal Crohn's and showed similar outcomes, suggesting that ileocolonic resection could be considered as an alternative to medical therapy early in the course of Crohn's disease [42]. However, the potential long term adverse effects of surgery, especially in terms of impact on the microbiome and/or metabolome, have not yet been fully elucidated.

To address these questions, we investigate the effect of surgical resection in a cohort of 129 subjects with IBD, stratifying by disease subtype and type of surgery.

2.3 Results

Overall demographics for the IBD patient population are shown in Table 2.1. Of 129 patients with IBD, 50 patients have ulcerative colitis, and 79 have Crohn's disease. A total of 332 stool samples were collected (from 18 patients: single sample, from 36 patients: two samples, from 40 patients: three samples, from 23 patients: four samples, from 6 patients: five samples). There is a median disease duration of 8 years, and 95 (73.6%) patients have current or prior TNF

inhibitor exposure. In total, 91 (70.5%) patients have no history of intestinal surgery, 21 with Crohn's disease underwent ileocolonic resection, and 17 including patients with UC and CD have had different types of colectomies. Of the patients who underwent colectomy, 10 with diagnosis of UC underwent subtotal colectomy with ileoanal pouch and 5 out of 10 progressed to develop CD of the pouch, 3 with CD had a subtotal colectomy with ileorectal anastomosis, and 4 with CD underwent total proctocolectomy with end ileostomy. These surgeries occurred a median of 3 years (interquartile range (IQR): 1 - 5.5 years) prior to the baseline stool sample collection.

2.3.1 Surgery lowered alpha diversity in both microbiome and metabolome

Prior intestinal surgery similarly decreases alpha diversity in both UC and CD patients (Fig. 2.1). In patients with UC, alpha diversity is not significantly different than in those with a normal pouch, pouchitis as compared to CD of the pouch. In patients with CD, different types of intestinal surgeries reduce alpha diversity with a trend towards the greatest reduction in samples after a total colectomy with end ileostomy (Kruskal-Wallis Test P value 4.51e-3), and this trend is consistent repeating the analysis using only one sample per patient (Kruskal-Wallis Test P value 2.01e-3). Small sample sizes, however, limit the comparisons among UC and CD surgical subtypes.

Since UC and CD represent a disease spectrum, we combined their analysis and found that ileocolonic resection and colectomy significantly decrease phylogenetic diversity, and in particular colectomy has a large impact on both microbial species diversity (Fig. 2.1C) and molecular diversity (Fig. 2.1D). Alpha diversity for species, as measured by Faith's phylogenetic diversity (PD), is lower in individuals with prior intestinal surgery, and samples from patients who have had a colectomy have the lowest alpha diversity (Kruskal test P value=7.09e-16). Permutational

Table 2.1: Inflammatory Bowel Disease Patient Demographics (N = 129)

	Ulcerative Colitis (N=50)	Crohn's disease (N= 79)
Age (years)		
Median (IQR)	44 (30-59)	36 (27-44)
Sex		
Male, N (%)	25 (50%)	36 (46%)
Female, N (%)	25 (50%)	43 (54%)
Body Mass Index (kg/m2)		
Median (IQR)	23.4 (21.2-28.0)	24.1 (21.3-27.9)
Disease duration (years)		
Median (IQR)	7 (2-12)	9 (4-18)
UC Montreal Classification, N (%)		
Proctitis (E1)	9 (18%)	
Left sided colitis (E2)	10 (20%)	
Extensive colitis (E3)	27 (54%)	
J pouch*	5 (10%)	
Crohn's Disease Location, N (%)		
Ileal (L1)		19 (24%)
Colonic (L2)		19 (24%)
Ileocolonic (L3)		36 (46%)
Crohn's disease of J pouch*		5 (6%)
Crohn's Disease Behavior, N (%)		
Inflammatory (B1)		8 (10%)
Stricturing (B2)		16 (20%)
Fistulizing (B3)		55 (70%)
Surgical Resection, N (%)		
None	45 (90%)	46 (58%)
Ileocolonic	0	21 (27%)
Subtotal colectomy with ileorectal		
anastomosis	0	3 (4%)
Colectomy with end ileostomy	0	4 (5%)
Colectomy with J pouch	5 (10%)	5 (6%)
Smoking, N (%)		
Never	36 (72%)	55 (70%)
Prior	13 (26%)	18 (23%)
Current	1 (2%)	6 (7%)
TNF-inhibitor Exposure, N (%)	30 (60%)	65 (82%)
Biologic use at baseline, N (%)		
TNF-inhibitor use	15 (30%)	42 (53%)
Integrin-inhibitor use	5 (10%)	8 (10%)
p40-inhibitor use	0 (0%)	4 (5%)

multivariate analysis of variance (PERMANOVA) analysis of taxonomic profiles shows that specific surgery type (ileocolonic v. colectomy) explains 9.84% of the variation in the microbiome , followed by disease subtype (7.63%), then antibiotic use (4.69%), then disease activity (3.1%). Other variables such as sex and age had much smaller effect sizes (Table S1). Notably, the number of years since surgery did not affect the overall reduction in alpha diversity (Spearman correlation: $p>0.05$).

Both disease activity and antibiotics use are important potential confounding factors. To account for disease activity, we separated samples into those from patients with active endoscopic disease versus inactive endoscopic disease activity in those with an endoscopic assessment within 3 months of the stool specimen. We find that there were no significant differences in alpha diversity between patients with active versus inactive disease activity. This result suggests that the differences in alpha diversity cannot entirely be explained by disease activity. Furthermore, antibiotic use has been shown to reduce diversity in the short term and long term 38 and may represent another potential confounding factor. While there is significant variation among surgical protocols, at a minimum one dose of multiple intravenous antibiotics is routinely given at the time of surgery. It is difficult, however, to unravel the precise effect of antibiotics during surgery as it is an integral part of the procedure. To examine the effect of antibiotics, we investigated whether current or recent (defined as within 90 days) antibiotic use affected alpha diversity. By stratifying the samples based on both surgery status and current/recent antibiotics use, we found that regardless of surgery status, current/recent antibiotics use consistently decreased alpha diversity (Mann Whitney U test P value = 0.004 [surgery] and P value =0.04 [no surgery]), suggesting that antibiotics administration during surgery may contribute to the reduction in diversity observed in surgery samples.

To assess volatility of the microbiome and metabolome, and the possible effects of surgery on this variability, we performed longitudinal analyses to compare the differences in species and metabolite abundance for samples from patients with and without surgery. We find that the microbiomes of subjects who had prior surgery are much more variable in terms of their overall composition. The boxplots in Fig. 2.1E and Fig. 2.1F show the differences between baseline, 6, 12, and 18 months for these groups, demonstrating that surgery increases microbiome (but, interestingly, not metabolome) volatility.

2.3.2 Surgery affected overall taxonomic, functional, and metabolite profiles

The beta diversity plots (Fig. 2.2A-C, left panels) show dissimilarity among samples, reduced to two dimensions for visualization purposes. These plots show that prior surgery affects overall taxonomic (PERMANOVA P value = 1.0e-3) and functional profiles and metabolite abundances. Although samples from IBD patients who had prior surgery are not identically distributed to samples from patients who did not undergo surgery in these Principal Coordinates Analysis (PCoA) plots, the distributions overlap, so we cannot differentiate between different types of surgery in all PCoA plots.

To identify the specific taxa that contribute to these overall microbiome differences, we used the compositionally-aware method ANCOM [43] to identify the top ten species that differentiate between surgery and non-surgery samples. Of all species identified as differentially abundant, potential pathogens such as *Klebsiella pneumoniae, Enterococcus faecium*, and *E. coli*, as well as *Veillonella atypica*, a known oral bacterium, have higher relative abundance in the surgery group compared to the non-surgery group. In contrast, butyrate producers such as *Eubacterium rectale, Faecalibacterium prausnitzii, Eubacterium Eligens*, and *Roseburia inulin-*

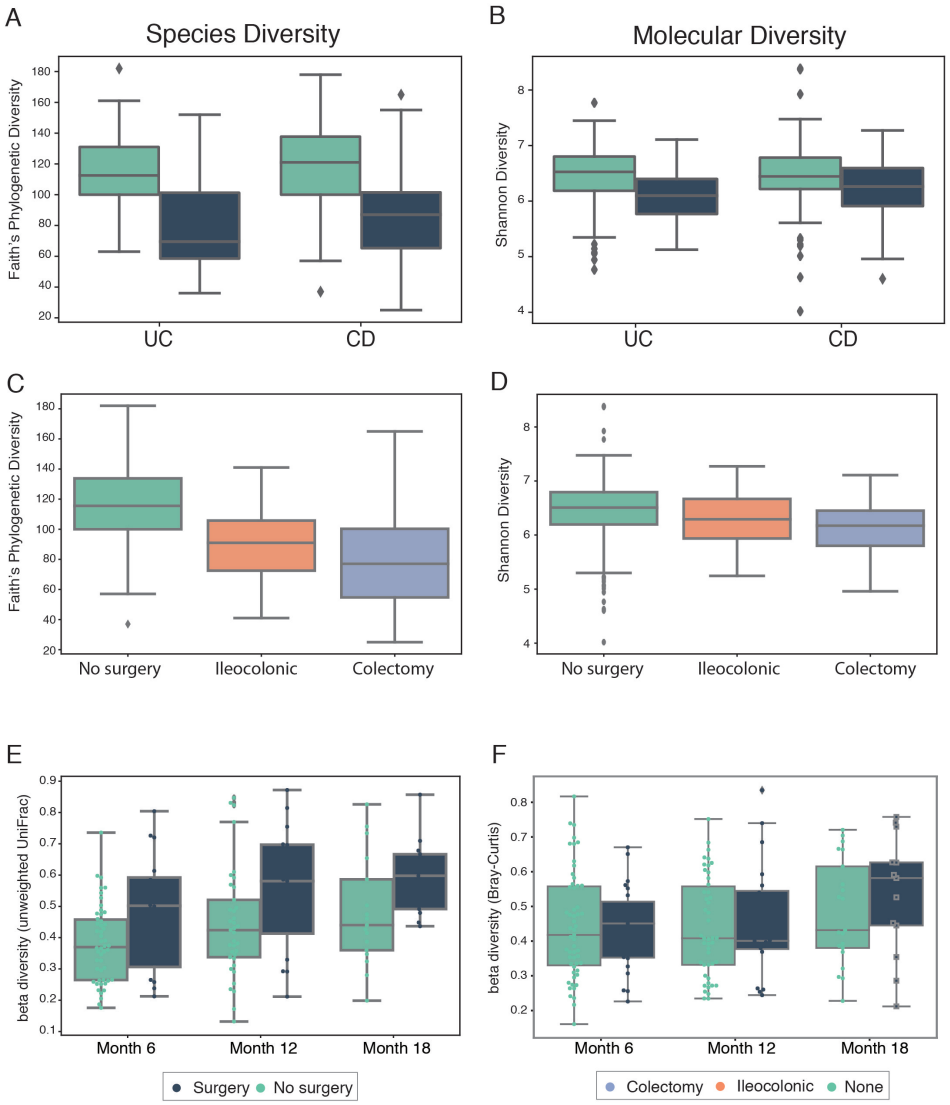


Figure 2.1: Comparison of alpha diversity and stability between surgery and non surgery groups. (A) Phylogenetic diversity (metric: Faith) of species abundance for UC and CD samples. (B) Molecular diversity (metric: Shannon) of molecular intensity evenness for UC and CD samples. (C) Phylogenetic diversity (metric: Faith) of species abundance for samples with different types of surgery. (D) Molecular diversity (metric: Shannon) of molecular intensity evenness for samples with different types of surgery. (E) Phylogenetic volatility (metric: unweighted UniFrac) comparing each follow-up sample to the baseline time point. F) Molecular volatility (metric: Bray-Curtis) comparing each follow-up sample to the baseline time point.

ivorans, and gut symbionts including *Roseburia hominis* and *Ruminococcus obeum* have lower relative abundance in surgery samples than non-surgery samples. Repeating these analyses at the pathway level, nitrate reduction is significantly elevated in surgery samples, consistent with previous studies that have shown that nitrate respiration occurs in the inflamed gut and promotes the growth of pathogens such as *E. coli* [44].

We used an unbiased approach to examine differences in the fecal metabolome. Combining all samples with prior surgery shows a significant decrease of Shannon diversity in UC (p=0.027, Mann-Whitney U test) but not in CD (p=0.100, Mann-Whitne U test). Specifically, the metabolomics from the CD samples with prior surgery cluster together with the lowest relative evenness in those who underwent total colectomy with ileostomy. Several metabolites are differentially abundant in individuals with prior surgery, most of these were bile acids. Only cholic acid and amino-2-ethoxybenzene are more abundant in the setting of prior surgery. Both tyrosine and glutamic acid are less abundant in subjects with surgery. A co-occurrence analysis using a neural-network approach [45] on all samples and subgroups of the samples does not reveal any strong associations between particular microbes and metabolites or pathways.

2.3.3 Higher abundance of primary bile acids detected for surgery samples

Overall, primary bile acids are increased for subjects that underwent surgery (in both UC and CD), whereas secondary bile acids are not significantly different (Fig. 2.3 A-D). Primary biliary acids (BA), including cholic acid and chenodeoxycholic acid, are produced by the liver, then gut microbiota are responsible for deconjugation of BA to generate secondary BA. Fecal primary BA have been shown to be enriched in IBD with relative depletion of secondary BA in Crohn's disease [6, 46, 47]. Resection of the terminal ileum as is the case of ileocolonic resections

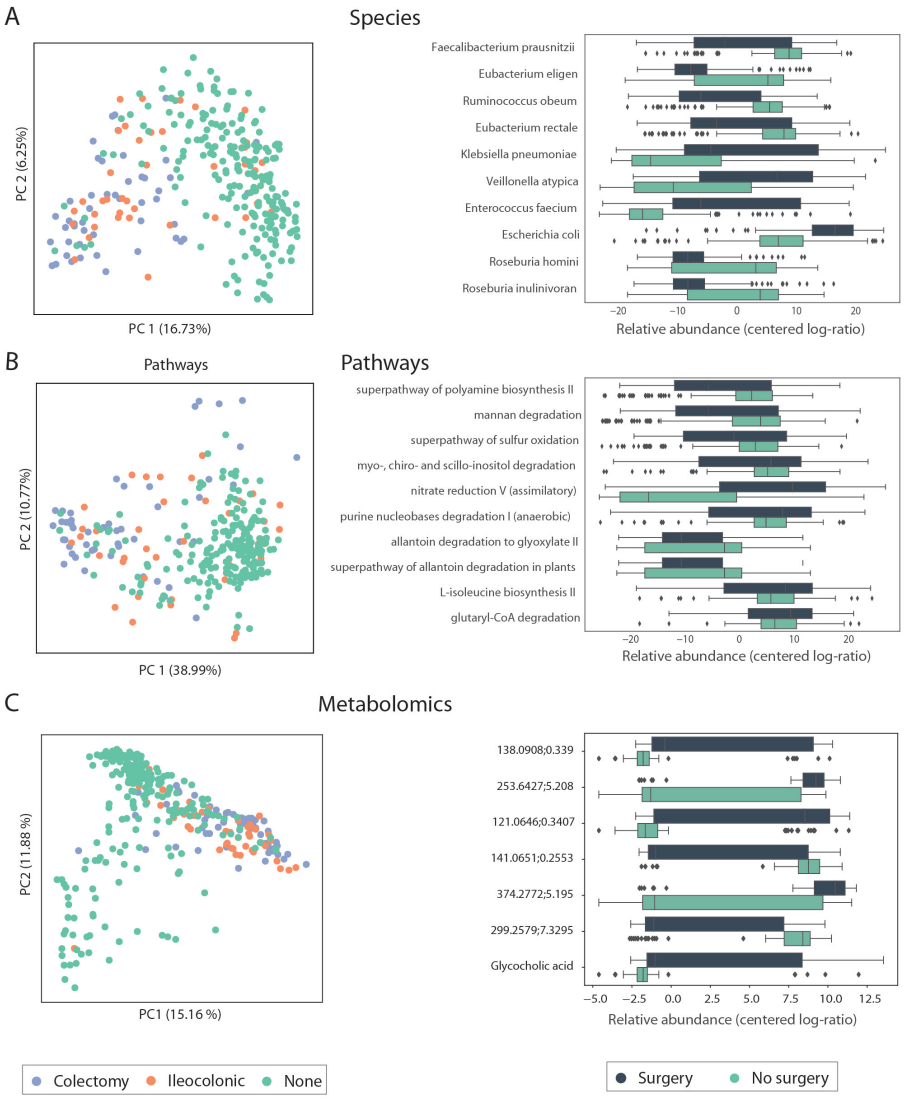


Figure 2.2: Comparison of surgery types vs non-surgery for taxonomic, functional and metabolomics profiles. (A) PCoA plot of species abundance (metric: unweighted unifrac) labeled by surgery subtypes, and top ten differentiating species between surgery and non-surgery samples. (B) PCoA plot of pathway abundance (metric: Bray-Curtis) labeled by surgery subtypes, and top ten differentiating pathways between surgery and non-surgery groups (C) PCoA plot of metabolomics abundance (metric: Bray-Curtis) labeled by surgery subtypes, and top seven differentiating metabolomics between surgery and non-surgery groups.

and most colectomies may reduce reabsorption of primary BA and increase the concentration of BA in the colon. We detected 21 distinct BA, including 14 primary BA, predominantly cholic acid and chenodeoxycolic acid, and 7 secondary BA, but few conjugated BA. Notably, in CD, primary BA are increased in patients with prior ileocolonic resection (Fig. 2.3A). There is a non-significant trend towards lower secondary bile acids in surgery samples without any specific signal based on surgery subtype. In UC, there is a similar trend towards increased primary bile acids in those with colectomy and J pouch; however, there are no significant changes in secondary bile acids stratified by prior surgery, though these analyses are limited by small sample sizes in subgroups.

2.3.4 Elevated *E. coli* relative abundance observed in surgery samples

Given the observed importance of *E. coli* in IBD [48] and its dominance of the overall community patterns, we performed a pan-genome analysis of this species specifically. Re-plotting the PCoA plot and labeling by *E. coli* abundance, we examined the association between prior intestinal surgery and *E. coli* abundance. Examining UC and CD, we clearly observe that non-surgery samples have markedly lower *E. coli* abundance as compared to surgery samples (Fig. 2.4A). Specifically, samples from all IBD patients who underwent colectomy have the highest abundance of *E. coli*, and samples from patients without prior surgery have the lowest (Fig. 4B). In CD, relative *E. coli* abundance is higher in patients with prior surgery with the highest abundance occurring in patients with a colectomy with an end ileostomy (Kruskal-Wallis test P value = 2.39 e-2). In UC, similar non-significant trends exist when samples were analyzed by UC post surgery subtype. The *E. coli* level also has a negative association with alpha diversity (Spearman correlation -0.26, P value = 6.47e-6). We then ran PanPhlan on samples with >1%

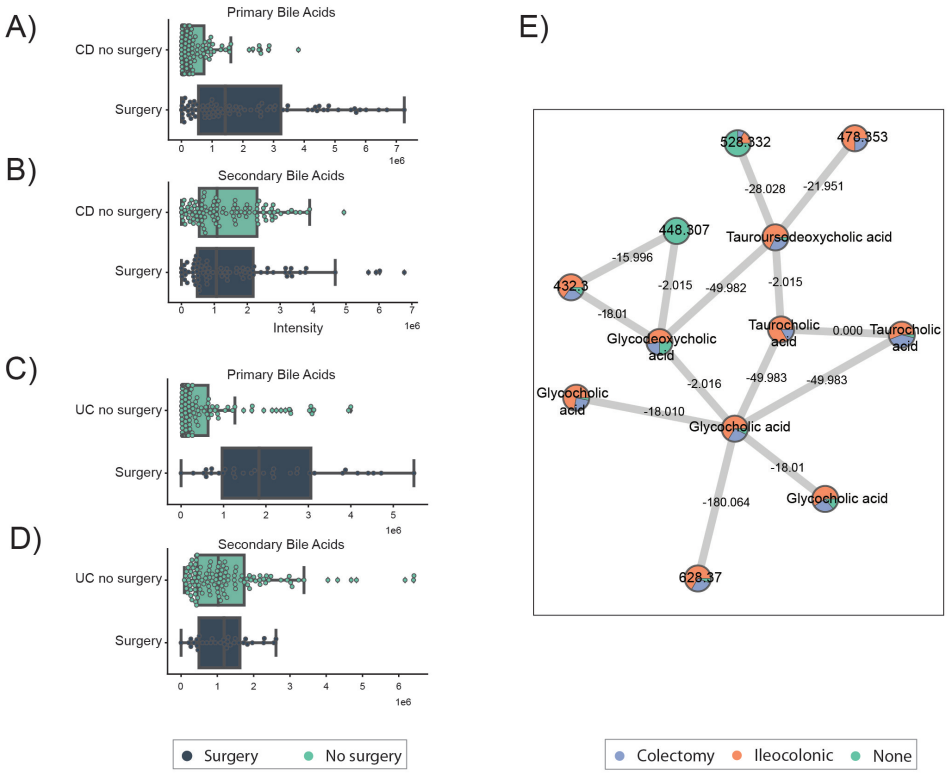


Figure 2.3: Metabolomic analysis of the primary and secondary bile acids identified in the cohort. (A) Primary bile acids for subjects with Crohn's disease, (B) Secondary bile acids for subjects with Crohn's disease, (C) Primary bile acids for subjects with Ulcerative Colitis, (D) Secondary bile acids for subjects with Ulcerative colitis (E) Molecular network of the metabolomics data, each node represents a metabolite and the edges represent the cosine similarity between the metabolite pairs.

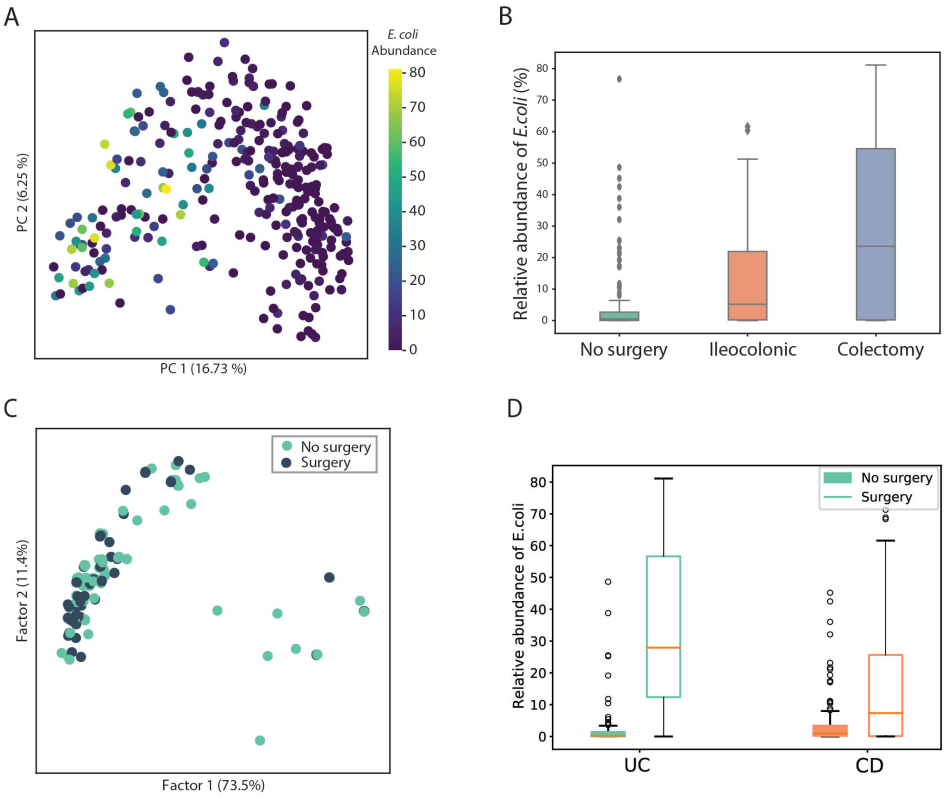


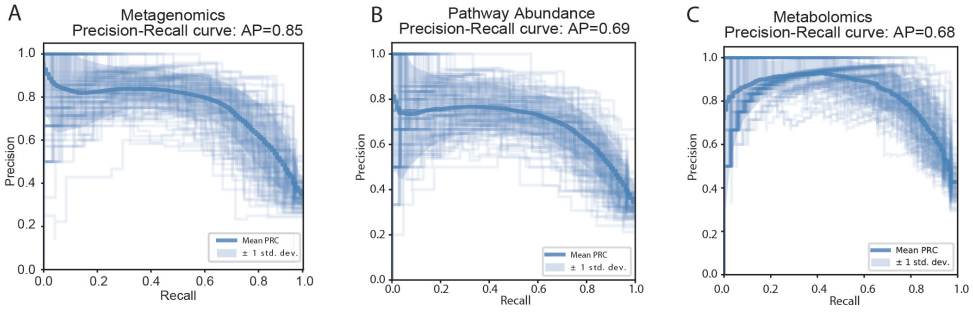
Figure 2.4: Comparison of *E. coli* abundance and characteristics of dominant *E. coli* strains between surgery and non-surgery samples. (A) PCoA plot of taxonomic profile labeled by *E. coli* abundance. (B) Relative abundance of *E. coli* in no surgery, ileocolonic, and colectomy samples. (C) MCA plot that describes the similarity of dominant *E. coli* strains from surgery and non-surgery samples in terms of metabolic and virulent functions. (D) Relative abundance of *E. coli* stratified by UC and CD.

E. coli abundance, and obtained genomic contents of dominant *E. coli* strains in 147 samples. We then constructed metabolic networks, and identified the presence/absence of adherent-invasive *E. coli* -associated virulence genes. Multiple correspondence analysis (MCA) analysis on the matrix describing metabolic reaction and virulence factor content in the 147 strains (at one strain per sample, Fig. 2.4C) suggests that *E. coli* strains from surgery as compared to no prior surgery samples have similar metabolic and virulence functions.

2.3.5 Taxonomic profiles differentiate surgery status better than metabolic or functional profiles

Finally, we tested the ability of these different data layers to discriminate between patients who had or had not previously undergone surgery. Specifically, we trained Random Forest classifiers to classify each sample according to whether it came from a patient who had undergone surgery, using the relative abundance tables for bacterial species, pathways and metabolites described above. We split the samples by subject to ensure that subjects in the training samples do not overlap with the test dataset. We randomly split the data (70% train, 30% split), tested and trained the model 100 times, yielding an average precision of 0.80 (Fig. 2.5A). Species adding the greatest weight to the classifier include: *Ruminococcus obeum, Clostridium asparagiforme, Faecalibacterium prausnitzii, Escherichia coli*, and *Bacteroides ovatus*. Repeating this analysis at the pathway and the metabolite level, we found that pathway abundance and metabolomics data yielded substantially worse classifiers (average precision of 0.69 and 0.68, Fig. 2.5B, Fig. 2.5C). The top five contributing pathway features involve carbon and energy metabolism, while the top five contributing metabolite features are all unidentified, except cholic acid.

We did not identify differences based on response to TNF-inhibitors with and without stratifying based on prior surgery ($p>0.05$ via Mann-Whitney U). We did not have sufficient statistical power to stratify by response to vedolizumab or by subsequent need for surgery after the timepoints sampled in this study; however, these topics are of intense clinical interest and would be valuable to explore in an adequately powered prospective longitudinal study.



	Top features
1	*Ruminococcus obeum*
2	*Faecalibacterium prausnitzii*
3	*Clostridium asparagiforme*
4	*Escherichia coli*
5	*Bacteroides ovatus*

	Top features
1	*GLUCUROCAT-PWY: beta-D-glucuronide and D-glucuronate degradation*
2	*PWY-5177: glutaryl-CoA degradation*
3	*PWY66-399: gluconeogenesis III*
4	*PWY0-1241: ADP-L-glycero-beta -D-manno-heptose biosynthesis*
5	*PWY-7383: anaerobic energy metabolism*

	Top features
1	*Unidentified 253.6427:5.208 m/z*
2	*Unidentified 431.2768:5.1948 m/z*
3	*Unidentified 160.1329:0.257 m/z*
4	*Cholic Acid*
5	*Unidentified 374.2772:5.195 m/z*

Figure 2.5: Random forest classifier to differentiate surgery from non-surgery samples using species, pathway and metabolite abundance. (A) Precision-recall curve of species classifier and the top five features. (B) Precision-recall curve of pathway classifier and the top five features. (C) Precision-recall curve of metabolite classifier and the top five features.

2.4 Discussion

Our study has shown that surgery has a large, persistent effect in decreasing the alpha diversity (i.e., the diversity within each sample) both of the microbiome and of the metabolome. Additionally, we have demonstrated that the instability of the microbiome is increased following surgery, but the instability of the metabolome appears to be unchanged. Not all surgeries are the same, and colectomy has a larger effect as compared to ileocolonic resection. This observation is interesting because in previous work, we showed that ileal CD with resection led to unstable microbiome dynamics [37], but it appears that colectomy has an even larger effect. Type of surgery explained more variation in the microbiome data than by any other variable (9.84%), followed by disease subtype (7.63%), then antibiotic use (4.69%), then disease activity (3.1%).

The large effect size of surgery on the microbiome and metabolome indicates that surgery is a key variable that must be controlled for in studies of the microbiome and IBD that seek to

assess smaller effect-size variables, such as sex-specific or age-specific factors. Although diet has a large effect, particularly in studies of IBD, it was unfortunately not assessed in this cohort to date. Integrating dietary assessment into future studies will be of considerable importance.

This study also reinforces the value of collecting multi-omics data, as the metagenomic and metabolomic data provide different views into IBD, providing a concordant view of alpha and beta diversity changes and a discordant view of instability with surgery. Additional data layers such as the metatranscriptome and the metaproteome, as were collected in iHMP [49], may be useful for further untangling these relationships, as will more extensive host immune phenotyping and other host profiling. Intriguingly, the metagenomic data, and the pathway information generated from it, were better able to detect changes associated with surgery than the metabolome data, suggesting that integration of capabilities over a longer period rather than immediate readout of current state may be most important for explaining phenomena associated with surgery in IBD.

As an observational study, there were potential confounding factors that were difficult to measure and control for in our analyses. Specifically, antibiotics are known to have a significant and potentially lasting effect on the microbiome [50, 51], and they are routinely given during colorectal surgery, potentially contributing to the reduction in microbial diversity on top of surgery itself. While we tried to control for disease activity using recent endoscopic disease activity as a surrogate, stool samples were not all collected at the time of endoscopy, and additional stool specimens were collected between endoscopic assessments. Alternative biomarkers, such as fecal calprotectin, would improve our ability to correct for disease activity, but were not measured. In this study we did not have samples both pre- and post-surgery from most patients. Furthermore, there remains the possibility that patients with more severe disease and corresponding micro-

biome changes are more likely to have received surgery and have more pronounced reductions in alpha diversity. In addition, small sample sizes within disease subtypes and specific surgeries limit our analyses and conclusion due to lack of statistical power. This point needs to be addressed in future work with a prospective longitudinal study design. Such studies can be guided by our recent work on determining the appropriate sampling interval and number of samples required to characterize IBD dynamics [38].

The results of this study expand upon what is known about the microbiome and its central role in the pathogenesis of disease recurrence in CD patients after ileocolonic resection [50] and in UC patients after colectomy with pouchitis, showing reduction in the microbial diversity that persists for years. However, the durable effect on reduction of diversity with surgery, especially in the longer term, has not previously been well characterized in many studies where it represents a significant potential confounder. Furthermore, current research strategies are working to harness the microbiome into both diagnostic and treatment strategies, and post-surgical patients represent a population that may particularly benefit from approaches such as fecal microbial transplantation or other targetted means of modifying the microbiome. One such recent study demonstrated the short term alterations in the microbiome of patients with ileal Crohn's disease who underwent ileocolonic resection, identifying bacterial species that may aid with diagnosis and prediction of recurrence [39]. Further mechanistic studies, as well as more detailed targeted biomarker discovery efforts, are required to understand the clinical effects of the reduction in microbiome and metabolome diversity or the increase in microbiome instability, and to develop inexpensive assays that allow clinicians to predict or explain relapse.

2.5 Conclusion

In this study, the collection and analysis of metagenomics and metabolomics data of an IBD cohort suggest that intestinal surgeries may have long-term effect on the gut microbiome, including reduced diversity of the microbes and metabolites, and further increased the instability in the gut microbiome of IBD patients. These long-term consequences of intestinal surgery need to be taken into consideration and evaluated carefully in future IBD microbiome studies.

2.6 Methods

2.6.1 Recruitment

Patients with a diagnosis of Crohn's disease or ulcerative colitis who were seen at the University of California, San Diego at the Inflammatory Bowel Disease Center were prospectively recruited and consented into the UCSD IBD Biobank. Diagnosis was confirmed by an IBD specialist. The study participants provided written informed consent, and the study was approved by the Institutional Review Board at the University of California, San Diego.

2.6.2 Specimen Collection:

Participants collected samples at home in Covidien 2450SA stool specimen containers, then refrigerated samples for transport in a cooler. Samples were returned within 72 hours, aliquoted and frozen at -80°C until DNA isolation.

2.6.3 UCSD Inflammatory Bowel Disease Biobank:

Each patient's clinical phenotype was assessed by an IBD specialist to define disease subtype (UC or CD), location, and phenotype based on Montreal disease classification for UC and CD (Table 2.1) [52]. Clinical and endoscopic data were collected prospectively, and disease phenotypes were confirmed by two IBD specialist physicians. Stool samples for each subject were collected approximately every 6 months.

2.6.4 Shotgun metagenomic data collection and profiling:

DNA was extracted with the Qiagen MagAttract PowerSoil DNA kit as previously described [53] and constructed the shotgun metagenomics libraries using 100 ng DNA from each sample. DNA was sheared to fragment sizes of 300 bp and input to the TruSeq Nano library-prep kit. Amplified libraries were then pooled and sequenced using HiSeq 4000 platform.

For the sequenced reads, we trimmed the adaptors and performed quality-filtering using atropos1.1.21 [54] (default parameters) and filtered out host reads using Bowtie2 2.3.0 [55]. After filtering out low-quality samples we were working with 300 metagenomics samples. The average number of reads per sample after quality filtering is 32,103,916 reads. The taxonomic profiles were generated using MetaPhlan2 2.7.7 [56] (default parameters) and the functional profiles were generated with HUMAnN2 0.11.2 [57] (default parameters).

We investigated the profiles of dominant *E. coli* strains using PanPhlan [58] for 147 samples with *E. coli* abundance >1% using the "ecoli16" database downloaded from the PanPhlan website. We then constructed pan-reactome matrix that describes the metabolic capability of the dominating *E. coli* strain in each sample following the method outlined in a previous study [13]. We also performed multiple correspondence analysis on the pan-reactome matrix using python

package mca [59] with Benzecri correction, with the parameter of TOL set to 1e-9.

2.6.5 Untargeted metabolomics profiling + data processing:

Sample Extraction Chemically cleaned stainless sterile beads were added to 100 mg - 50 mg of human fecal samples along with 50% methanol (spiked with 2 uM sulfamethazine) at a volume ratio of 1 mg per sample to 10 ul extraction solvent followed by a tissue homogenization using a Qiagen TissueLyzer II for 5 min at 25 Hz. Samples were centrifuged for 15 minutes at 14,000 rpm and 400 ul of the resulting supernatant were transferred to a 96 well deep-well plate and dried via a centrifugal low-pressure system (SpeedVac Plus, Savant) and stored at -80°C until mass spectrometry analysis. Samples were resuspended with 130 uL 50% methanol (spiked with 1 uM of sulfadimethoxine) and sonicated for 5 minutes. Following centrifugation at 14,000 RPM for 15 minutes, 100 uL of supernatant were transferred to a new shallow-well 96 well plate. The 96-well plate was then diluted 20 fold.

Data Acquisition The fecal samples were analyzed using an ultra-high performance liquid chromatography (Ultimate 3000, Thermo) coupled to a quadrupole time-of-flight mass spectrometer (maXis Impact, Bruker). Chromatographic separation was accomplished using a Kinetex C18 1.7 uM, 100 Å, 2.1 mm by 50 mm column (Phenomenex) maintained at 40°C during separation. 5 uL of extract was injected per sample. Mobile phase composition was: A, LC-MS grade water with 0.1 % formic acid (v/v) and B, LC-MS grade acetonitrile with 0.1 % formic acid (v/v). The chromatographic elution gradient parameters were the following: 0.0 - 1.0 min, 5% B; 1.0 - 9.0 min, 100% B; 9.0 - 10 min, 100% B. An MS1 scan from 50-1500 at 3 Hz was followed by MS2 scans. The heated electrospray ionization parameters were the following: drying gas, 9.0 L min-1; dry gas temperature 200 °C; capillary voltage, 3.5 kV; end plate offset, -0.5 kV; and

nebulizer, 2.0 bar. Hexakis (2,2-difluoroexthoxy)phosphazene, lock mass standard, was added to the ionization source.

Data Processing The acquired qTOF files (.d) were exported using DataAnalysis (Bruker) as .mzXML files after lock mass correction. Feature finding was performed on MS1 data in MZmine2 [60], producing a data matrix of MS1 features (i.e m/z and retention time) and associated peak area. MS2 data were analyzed using GNPS (estimated false discovery rate used is 0.005 at the settings) [61].

2.6.6 Statistical Analyses:

The following description summarizes the main software packages used in this analysis. For reading and writing data we used scikit-bio 0.5.5, the BIOM format 2.1.7 [62] and QIIME 2 version 2019.1 [63]. Data visualization was done using Seaborn 0.9.0 [64], Matplotlib 3.0.3 [65], and QIIME2. The machine learning, and linear algebra tasks were performed using scikit-learn 0.20.2 [66], SciPy 1.2.1 [67], Pandas 0.24.2 [68], and NumPy 1.16.2 [69]. A detailed description of the individual steps has been published as a collection of Jupyter notebooks: https://github.com/knightlab-analyses/ibd-surgery

This survey was represented by three contingency matrices: one for the taxonomic profile, one for the functional profiles and one for the untargeted metabolomics. Alpha diversity calculations for the metabolomics and metagenomics matrices were performed using scikit-bio. For metabolomics we used Shannon's index. For the taxonomic profiles, we used Faith's phylogenetic diversity [70] based on the NCBI taxonomy of the represented bacteria. In both cases low-quality samples were removed from analyses. Similarly for beta diversity calculations, we used the Bray-Curtis distance for the metabolomics matrix, and the unweighted UniFrac [71] matrix for the

taxonomic profiles as implemented in SciPy and scikit-bio. The differentially abundant features were estimated using ANCOM [43] as implemented in scikit-bio and QIIME2.

For the evaluation of a model to classify samples according to the surgery status, we used a Random Forests classifier [72], and a Precision-Recall curve for each data matrix. The performance of the classifier was evaluated using the average precision over 100 independent iterations. At each iteration, the subjects were exclusively split in a training and a test set. The Precision-Recall curve was selected to account for the class imbalance (70% of the subjects did not undergo surgery).

2.6.7 Data Availability

The metagenomic sequencing data has been deposited to the European Bioinformatics Institute, and the untargeted metabolomics has been deposited to the MASSIVE repository (MSV000082221). In addition, the full dataset including sample metadata has been made public on Qiita [73] https://qiita.ucsd.edu/study/description/11546. Accession number on EBI will be available soon as the submission is being processed.

Acknowledgments

We would like to thank Tara Schwartz and Andre Matti for assistance with sample processing and Daniel Freed for his contributions to the project. This work was supported by Crohn's and Colitis Foundation Career Development Award and UCSD 1KL2TR001444 (B.S.B), Microbial Science Initiative Graduate Research Fellowship and Seed Grant by UC San Diego Center for Microbiome Innovation, Novo Nordisk Foundation Center for Biosustainability and the Technical University of Denmark (NNF10CC1016517), Janssen Human Microbiome Initia-

tive, NIDDK-funded San Diego Digestive Diseases Research Center (P30 DK120515), Clinical and Translational Science Awards grant (UL1- TR-001442).

Chapter 2 in full is a reprint of material published in: **Fang, X. ***, Vázquez-Baeza, Y. *, Elijah, E., Vargas, F., Ackermann, G., Humphrey, G., Lau, R., Weldon K. C., Sanders, J. G., Panitchpakdi, M., Carpenter, C., Neill, J., Miralles A., Dulai, P., Singh, S., Tsai, M., Swafford, A. D., Smarr, L., Boyle, D. L., Palsson B. O., Chang, J. T., Dorrestein, P. C., Sandborn W. J., Knight, R., Boland, B. S. Gastrointestinal surgery for inflammatory bowel disease persistently lowers microbiome and metabolome diversity. *Inflammatory Bowel Disease*, Submitted. The dissertation author was one of the primary authors.

Chapter 3

Metagenomics-based, strain-level analysis of *Escherichia coli* in a Crohn's disease patient

3.1 Abstract

Dysbiosis of the gut microbiome, including elevated abundance of putative leading bacterial triggers such as *E. coli* in inflammatory bowel disease (IBD) patients, is of great interest. To date, most *E. coli* studies in IBD patients are focused on clinical isolates, overlooking their relative abundances and turnover over time. Metagenomics-based studies, on the other hand, are less focused on strain-level investigations. Here, using recently developed bioinformatic tools, we analyzed the abundance and properties of specific *E. coli* strains in a Crohn's disease (CD) patient longitudinally, while also considering the composition of the entire community over time. In this report, we conducted a pilot study on metagenomic-based, strain-level analysis of a time-series of *E. coli* strains in a left-sided CD patient, who exhibited sustained levels of *E. coli* greater than 100X healthy controls. We: 1) mapped out the composition of the gut microbiome

over time, particularly the presence of *E. coli* strains, and found that the abundance and dominance of specific *E. coli* strains in the community varied over time; 2) performed strain-level *de novo* assemblies of seven dominant *E. coli* strains, and illustrated disparity between these strains in both phylogenetic origin and genomic content; 3) observed that strain ST1 (recovered during peak inflammation) is highly similar to known pathogenic AIEC strains NC101 and LF82 in both virulence factors and metabolic functions, while other strains (ST2-ST7) that were collected during more stable states displayed diverse characteristics; 4) isolated, sequenced, experimentally characterized ST1, and confirmed the accuracy of the *de novo* assembly; and 5) assessed growth capability of ST1 with a newly reconstructed genome-scale metabolic model of the strain, and showed its potential to use substrates found abundantly in the human gut to outcompete other microbes. In conclusion, inflammation status (assessed by the blood C-reactive protein and stool calprotectin) is likely correlated with the abundance of a subgroup of *E. coli* strains with specific traits. Therefore, strain-level time-series analysis of dominant *E. coli* strains in a CD patient is highly informative, and motivates a study of a larger cohort of IBD patients.

3.2 Background

Dysbiosis of the gut microbiome in inflammatory bowel disease (IBD) patients is associated with reduced bacterial diversity, an increase in relative abundance of Proteobacteria [74], and decline in Firmicutes [74]. Specifically, *E. coli* is considered one of the potential causes of IBD formation and progression [75, 76]. One specific pathotype, adherent-invasive *E. coli* (AIEC), which is able to attach to intestinal epithelial cells and survive and replicate within macrophages, has been implicated in intestinal inflammation [77, 78]. Members of this pathotype, as well as other IBD-associated *E. coli* isolates, mainly belong to phylogroup B2 [79], carrying a diverse

set of virulence factors and displaying distinct metabolic phenotypes [7, 12]. However, no unique genetic determinant has been identified for this group [80].

Previous studies on *E. coli* in IBD mainly focused on clinical isolates extracted from intestinal biopsy and fecal samples, which are then cultured and experimentally characterized [80–83]. However, most of these studies did not take into consideration other factors including composition of the gut microbiome and dynamics of the community. Recently, with the drop in sequencing costs, metagenomics data has become a popular source of information with which to investigate the composition [35], function [32, 84] and dynamics [37, 85] of the IBD microbiome. However, these studies lack a detailed characterization of the *E. coli* community. They generally only examine the relative abundance of *E. coli*, overlooking the strain-level composition and strain-specific traits of the *E. coli* community, yet previous study has already showed genetic diversity and temporal variation in the *E. coli* population [86].

Fortunately, strain-level analysis of metagenomics data has been made possible with recently developed bioinformatics tools, including MIDAS, that characterizes strain-level variation [11], DESMAN, that allows *de novo* extraction of strains [41], among other strain-level population genomics tools [87–89]. Additionally, tools developed for genome-level analysis, such as genome-scale metabolic models (GEMs), enable comprehensive strain-level analysis. GEMs are reconstructions of the metabolic network of strains that are subsequently converted to computable mathematical models, allowing mapping between the genetic basis and phenotypic metabolic functions [90]. Due to the versatile genomic content of *E. coli* [91], strain-level GEM analysis has proven to be essential and informative [92].

Here, we conducted a pilot study on one IBD patient, specifically a patient with left-sided Crohn's disease (CD), and performed metagenomics-based, strain-level analysis of the patient's

time-series *E. coli* community. We not only examined the composition of the gut microbiome, relative abundance of *E. coli*, and community dynamics, but also performed strain-level analysis to identify, assemble, and characterize the dominant *E. coli* strains at different time points, followed by experimental validation.

3.3 Results

3.3.1 Time-series stool samples were collected and sequenced for three years

We studied 27 time-series stool samples (named TP1 - TP27 as shown in Fig. 3.1) collected from a 69 year-old male CD patient, who was diagnosed with colonic CD at the age of 63 with inflammation confined to his sigmoid colon. These samples were collected over a period of three years between 2011 and 2014, covering both stable and inflamed states [93]. We generated metagenomics data for each sample collected, and recorded detailed metadata including body mass index (BMI), blood C-reactive protein (CRP) level, fecal calprotectin level, and other biomarker measurements during this period. During the three years, this patient took Ciprofloxacin, Metronidazole, and Prednisone daily in February 2012, and also used Lialda and Uceris from June to November 2013. BMI was recorded for all samples and ranged between 23.6 and 25.9 (Fig. 3.1). High-Sensitivity CRP (hs-CRP) level, which is indicative of inflammation level, was measured for 18/27 samples, and fluctuated between 2.4 and 27.1 mg/L (Fig. 3.1). Fecal calprotectin showed a trend similar to blood hs-CRP level, with significant variation. In particular, blood and fecal inflammation levels were the highest when the first sample was collected, with hs-CRP peaked at 27.1 mg/L, while the normal range of hs-CRP for healthy controls is $\leq$ 1 mg/L [94], and with Calprotectin peaking at 2500, over 50x the upper limit for healthy

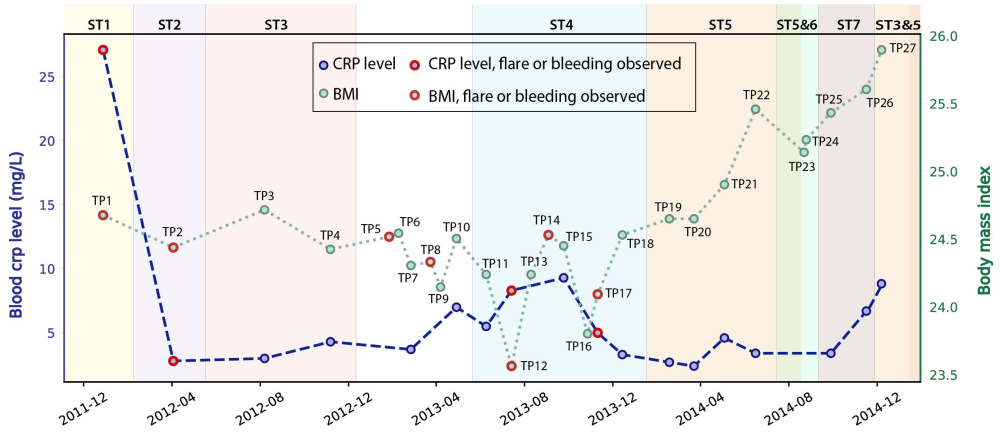


Figure 3.1: Blood hs-CRP level and BMI of the patient fluctuated during the three years of this study. hs-CRP only available for 18 samples. Samples collected during bleeding or flare are labeled in red. The dominant *E. coli* strain varied for different time points (discussed in the next paragraph), and are labeled by different background colors.

controls. Therefore, we aimed to explore the relationship between inflammation status and gut microbes, especially with the *E. coli* community in the gut microbiome.

3.3.2 Composition of the gut microbiome and *E. coli* community changed over time

Analysis of the gut microbiome composition and richness indicates that the gut microbial community of this patient was dysbiotic, and highly dynamic during the three years of this study. We performed taxonomy assignment for the metagenomic samples using MetaPhlan2 [88], and calculated the alpha and beta diversity of the 27 samples. Compared to the gut microbiome of healthy controls that are mostly dominated by Firmicutes ((49-76%) and Bacteroidetes (16-23%) [74] with a minor component of Proteobacteria (median=1%) [95], this patient had an elevated

level of Proteobacteria ranging from 1.09% to 55.3%, and a reduced level of Firmicutes between 22.3% and 49.1%. We also found enterobacteria phages K1E (accession: NC_007637.1) and K1-5 (accession: NC_008152.1) in TP1, which are not shown in MetaPhlan2 results in (Fig. 3.2). We also performed principal coordinate analysis (PCoA) on the beta diversity calculated to evaluate the dissimilarity between samples.

In particular, we characterized the *E. coli* community in the gut microbiome, since *E. coli* is considered one of the leading bacterial triggers in IBD [75]. The relative abundance of *E. coli* in this patient ranges from 0.1% to 42.6%, which was abnormally high (as much as 400x) compared to that of the healthy controls ($\leq$0.1% in the healthy cohort [96], but consistent with elevated *E. coli* abundance observed in previous IBD studies [74]. During the three years of study, the *E. coli* level remained relatively high, except for the first four months of 2013, during which TP5-TP10 were collected (highlighted in red in Fig. 3.2B). The inflammation level during this particular period did not show significant differences compared to other time points. Interestingly, the relative abundance of *E. coli* did not necessarily correlate with inflammation level in all samples. For example, TP2 has the highest *E. coli* relative abundance of 42.6%, yet it only has a hs-CRP level of 2.8 mg/L (1/10 of the hs-CRP level for TP1). Since *E. coli* is a highly versatile species with an open pan-genome [97], it is possible that only a subset of *E. coli* strains with certain pathogenic features contribute to disease progression in IBD. Therefore, we further investigated the strain-level composition for the *E. coli* community in the 21 samples that have $\geq$5% *E. coli* relative abundance (highlighted in green in Fig. 3.2B). Six samples (TP5-TP10) were excluded from further *E. coli* studies due to their scarcity of *E. coli*.

Single-nucleotide variants (SNV) analysis on the selected 21 samples suggests that the *E. coli* community was dominated by a single strain in most samples, and the dominant strain

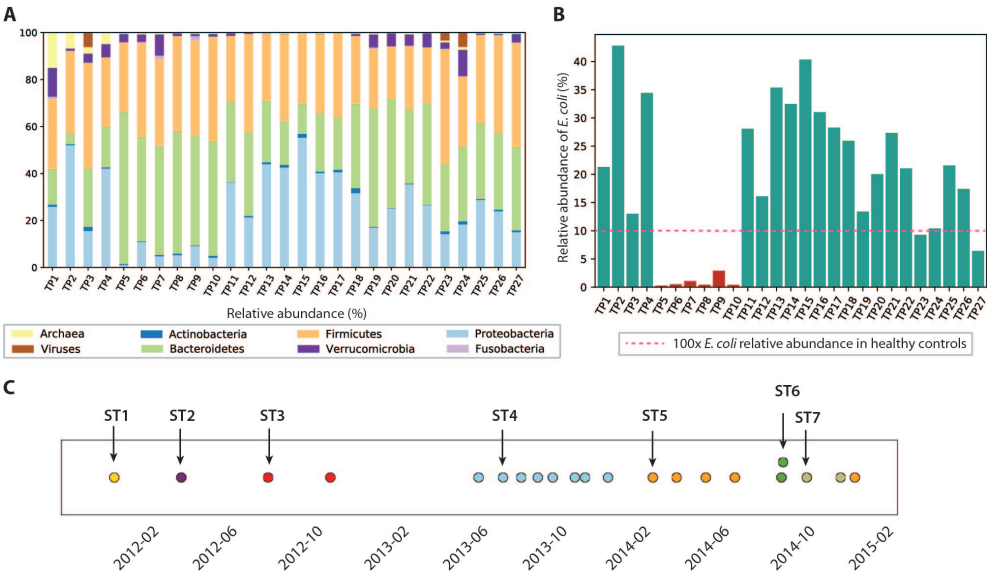


Figure 3.2: Composition of the gut microbiome and the *E. coli* community is dynamic. (A) Relative abundance of microbes at phyla level. (B) Relative abundance of *E. coli* in this patient. *E. coli* relative abundance is <0.1% in the healthy cohort. (C) Dominant strains of the *E. coli* community identified in 21/27 samples. Colors represents different dominant strains. Arrows highlight the samples we selected for further analysis on dominant strains.

switched over time. SNV frequencies for *E. coli* species were detected by MIDAS [11]. Most SNV frequencies are close to 0 or 1, implying that a single strain was typically dominating the *E. coli* community at a given point in time. This result is consistent with a finding in a previous study that a single strain dominates most species in the gut microbiome [88].

Positions of the detected SNVs across multiple samples also suggest that the dominant *E. coli* strain changed over time (see Methods) (Fig. 3.2C), potentially due to alterations in diet, microbiome ecological structure, and environment (including the components of the human immune system). In the 21 samples with higher *E. coli* relative abundance, we identified a total of seven dominant strains (some of them abundant in several time points). To further characterize the dominant strains and understand their association with inflammation, we then focused on the highlighted samples in Fig. 3.2C that contain the seven dominant strains.

3.3.3 Dominant *E. coli* strains assembled and computationally characterized

We attempted to recover genome sequences of the seven dominant *E. coli* strains from the selected samples. Draft assemblies of the dominant strains (named ST1-ST7) were obtained by *de novo* metagenomic assembly and binning of individual samples (see Methods), followed by functional annotation using Prokka [98]. Numbers of protein coding genes in the resulting annotations range from 4,411 to 5,213 (Table 3.1). In addition, we performed phylogenetic analysis using PhyloPhlan [99] to infer the phylogroup of each assembly. Although previous studies have shown that strains in B2 and D phylogroups are more frequently found in IBD patients [100], the seven dominant strains in this patient have diverse phylogenetic origins and are predicted to span phylogroups B2, E, D, B1, and A. In particular, ST1 and ST5 likely belong to phylogroup B2, which contains most of AIEC strains. In addition, we have also assigned the sequence types of the dominant strains using the *de novo* assemblies and BacWGSTdb [101]. The dominant strains are reported to have different sequence types (Table 3.1). Specifically, sequence type 95, 69 and 131 are predominant in extraintestinal pathogenic *E. coli* strains [102].

To further explore the diversity of the selected strains, we constructed a pan-genome for the seven assemblies and found significant variation between strains. We built the pan-genome with Roary [103] using a threshold of 80% for gene similarity (see Methods). We identified a total of 8,459 orthologues, of which only 37.7% are core genes shared between all strains. Among the rest of the accessory genes, 39.9% are unique to only one strain, highlighting the diversity of the seven strains. To further explore the variation between strains, we next investigated the genomic features and metabolic functions of the dominant strains.

Table 3.1: Characteristics of the seven dominant strains recovered from metagenomic samples.

Name	Time	Number of CDS	Inferred Phylogroup	Sequence Type
ST1	2011/12/28	5134	B2	95
ST2	2012/04/03	5213	E	1629
ST3	2012/08/07	4591	D	69
ST4	2013/07/14	4618	B1	58
ST5	2014/03/23	4498	B2	131
ST6	2014/08/25	4411	A	409
ST7	2014/09/28	4487	B1	1727

3.3.4 The analysis of recovered strains reveals a diversity of virulence factors

We examined the distribution of virulence factors in the seven assemblies. For comparison, we included two well-studied AIEC strains, NC101 [104, 105], associated with inducing colon-cancer [106], and LF82, an *E. coli* strain associated with right-sided ileal CD patients [107–109]. In addition, we included the widely studied commensal strain K-12 MG1655 as a well-defined reference strain. We first mapped the seven genome assemblies and three reference strains to a curated virulence factor database VFDB [110] using BLAST [111] with a threshold of 80% sequence similarity. This procedure identified a total of 164 virulence factors amongst the ten strains. Many of these virulence factors are involved in functions that are previously implicated in pathophysiology in IBD, including iron-acquisition [112], adhesion [113], secretion systems [114], and capsule synthesis [115]. We observed that strains in phylogroup B2 (NC101, LF82, ST1, ST5) generally have more virulence factors compared to the other strains, and have more virulence factors in common.

3.3.5 Presence/absence of 57 known IBD-associated virulence factors in the recovered strains

We next focused on 57 genes that have been associated with pathogenicity in IBD patients from previous studies. We collected the genes and their sequences from literature, and mapped

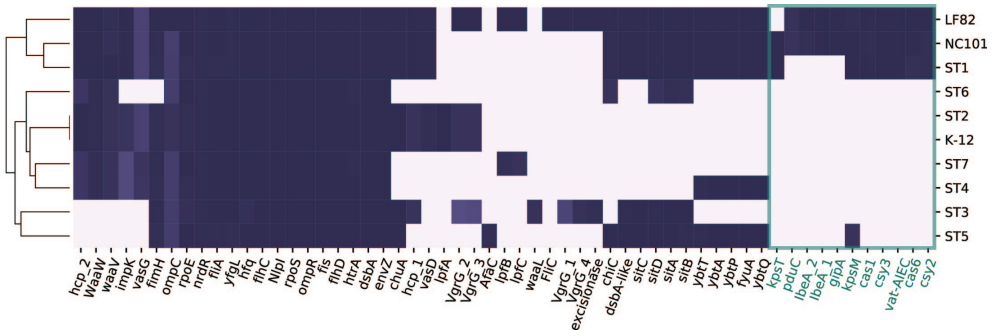


Figure 3.3: Distribution of 57 genes that were implicated in AIEC pathogenesis in ten strains. Genes unique to ST1, NC101 and LF82 are involved in various functions including capsule synthesis (*kpsT* [115], mucins protease (*vat-AIEC* [116]), CRISPR-associated genes (*cys3, cas6, cys2, and cas1* [117]), invasion (*ibeA* and its variant [118]), phage encoded VFs (*gipA* [119]), and propanediol utilization (*pduC* [112]).

them against the ten strains using BLAST [111]. Interestingly, only ST1 clustered with the representative AIEC strains LF82 and NC101, while ST5 did not share as many genes with the selected pathogenic strains (Fig. 3.3). This could potentially explain why ST1 correlated with high inflammation level, while hs-CRP was only 2.4 mg/L when ST5 was collected. We found a set of genes that are unique or more prevalent in ST1, LF82, and NC101 that differentiate them from other strains (highlighted in Fig. 3.3). Besides IBD-associated virulence factors, we also found that similar to NC101, ST1 also harbors the polyketide synthase (*pks*) genotoxic island that was shown to induce colorectal cancer [106].

3.3.6 Metabolic networks differentiate ST1 and AIEC strains from other dominant strains collected during periods of low inflammation

Besides virulence factors, we also delineated the differences in metabolism between strains. We built draft metabolic networks for seven assemblies and the three reference strains based on the previously published multi-strain genome-scale metabolic models (GEMs) [92] (see Methods). For the ten metabolic networks reconstructed, there are 3,077 metabolic reactions in total, among

which 302 are accessory reactions missing from at least one strain, and 2,775 core reactions that are present in all strains.

To investigate the discrepancy in metabolic functions between these strains, we created a pan-reactome for these ten strains (see Methods). We then performed multiple correspondence analysis (MCA) on the pan-reactome matrix formed by absence/presence calls for these reactions, which has been shown to effectively classify reactomes [120]. We then focused on factor 1 and factor 2 (Fig. 3.4A), since they explained a total of 84% variance (67.1% and 16.9%, respectively).

The plot of factor 1 vs. factor 2 (Fig. 3.4A) shows that TP1 is very similar to NC101 and LF82 in terms of metabolic functions, while strains isolated from other time points displayed diverse characteristics (Fig. 3.4A). We observed that factor 1 separated B2 strains from non-B2 strains, while factor 2 separated TP5 and K12 from the other strains. We further investigated the 50 reactions that have the greatest contribution to factor 1 and 2, and plotted their functional distribution (Fig. 3.4B). Many of the top contributing reactions in factor 1 are involved in alternative carbon metabolism, cofactor biosynthesis, and transport reactions. Further analysis showed B2 and non-B2 strains have distinct reactions involved in carbon utilization and metabolite transport, suggesting that B2 strains and non-B2 strains may be adapted to different microenvironments and nutrient substrates.

For the top contributing reactions in factor 2, although some are also involved in carbon metabolism and transport reactions, more than half of the reactions are engaged in lipopolysaccharide (LPS) biosynthesis and recycling. Additional analysis showed that TP5 and K12 have a unique set of reactions involved in LPS synthesis compared to the other eight strains. Previous studies showed that endotoxicity of LPS produced by intestinal microbiota plays a vital role in the development of intestinal colitis [121]. Thus, the difference we observed in LPS biosynthe-

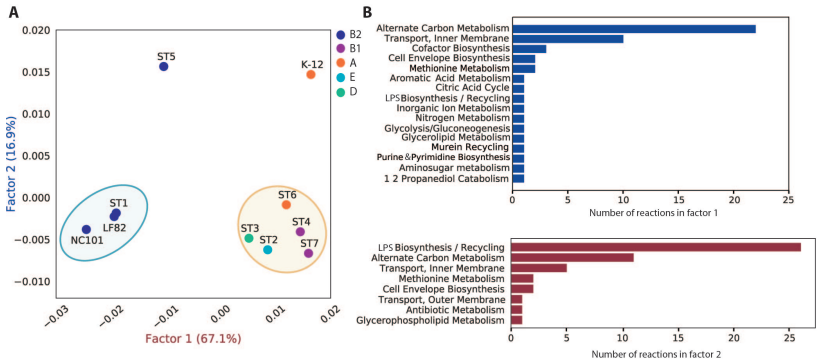


Figure 3.4: MCA analysis of pan-reactome for ten strains. (A) Visualization of factor 1 and factor 2 of MCA results. (B) Functional distribution of important reactions in factor 1 and factor 2.

sis may correlate with host inflammation status, and needs to be experimentally studied in the future.

MCA analysis of the pan-reactome showed similarity in metabolic functions between ST1 and AIEC strains LF82 and NC101, suggesting that *E. coli* strains associated with intestinal inflammation in IBD patients may share certain metabolic capabilities. However, because we only obtained *de novo* assemblies that are incomplete, we could not construct accurate GEMs to further evaluate their growth capabilities. To verify our results and enable accurate GEM simulation of the most interesting ST1 strain, we proceeded with its experimental isolation, sequencing, and characterization.

3.3.7 ST1 isolation and characterization

Since ST1 was present in high abundance during peak inflammation and showed the closest resemblance to known AIEC strains, we proceeded to isolate ST1 from the stool sample and characterize it experimentally. Its identity was confirmed with SNV analysis (see Methods). This strain, which we named CG1MAC was sequenced and assembled to give a 5,169,659 bp

genome with 4,916 coding regions, of which 4,905 genes were present in the ST1 assembly. The accuracy of the ST1 assembly, compared to CG1MAC, is 95.5%. Additional genomic analysis showed that CG1MAC is closely-related to 3_2_53FAA (sharing 4837/4916 ORFs), an *E. coli* strain previously isolated from the inflamed left-sided descending colon of a 52-year-old male CD patient, and is part of the HMP reference genome collection with the strain identification number HM-38 [122]. We note the similarity in gender, age, and colon inflammation site with our patient. Additionally, the serotype of CG1MAC was experimentally determined to be O2:H7 by the National Microbiology Laboratory in Canada. Phylogenetic analyses suggest that CG1MAC is evolutionarily closely related to AIEC and uropathogenic (UPEC) strains in phylogroup B2.

To examine whether CG1MAC exhibited AIEC characteristics, we conducted adhesion and invasion assays. Experimental results showed that CG1MAC is able to adhere well to the intestinal epithelial cell line Caco-2, but does not invade THP-1 macrophages, unlike the representative AIEC strain LF82. CG1MAC was engulfed at a low level and showed poor survival intracellularly.

3.3.8 Growth capability of CG1MAC is predicted to be similar to that of AIEC strains

We built a draft genome-scale model (GEM) for CG1MAC based on its genome sequence and previously published *E. coli* models [92] (see Methods). The GEM for the CG1MAC strain contains 1,581 genes, 2,913 metabolic reactions, and 2,115 metabolites. We then predicted the growth capability of CG1MAC, along with three draft reference models K-12, LF82, and NC101 that were reconstructed following the same procedure.

Growth simulation results on various nutrient sources indicate that CG1MAC is similar

to AIEC strains in terms of growth capability. Growth predictions suggest the four strains (CG1MAC, K-12, LF82, and NC101) have distinct metabolic capabilities, as their predicted growth ability differs for 35 substrates (Fig. 3.5A). The predicted growth phenotype displayed by CG1MAC is similar to LF82 and NC101, as they share the ability to utilize a subset of six substrates (labeled in orange in Fig. 3.5A), but not K-12. Among the six identified substrates, some are found abundantly in the intestine, including cellobiose, a derivative of an insoluble dietary fiber cellulose [123, 124], as well as monosaccharides derived from intestinal mucosa: N-acetyl-D-galactosamine (GalNAc) and N-acetyl-D-galactosamine 1-phosphate (GalNAc 1P) [125]. The ability to utilize deoxyribose, on the other hand, suggests pathogenicity of CG1MAC and NC101. A previous study showed that the capability to metabolize deoxyribose is associated with the pathogenic potential of intestinal and extraintestinal *E. coli* strains, as this ability increases their competitiveness [126]. Deoxyribose availability also promotes host colonization of the intestine by pathogenic *E. coli* strains [127]. The remaining two substrates, 3-phospho-D-glycerate (3PG) and 2-phospho-D-glycerate (2PG), are important intermediates in glycolysis [128], and precursors for amino acid biosynthesis [129]. The ability to directly uptake these substrates potentially enables NC101 and CG1MAC to generate energy more efficiently, thus likely to outcompete other microbes. We have also identified reactions that enable growth on the above six substrates, that are missing from K-12 (labeled in red in Fig. 3.5A). We observed that K-12 lacks transporters for all six substrates, as well as some downstream enzymes. We also performed experimental growth experiments for model validation.

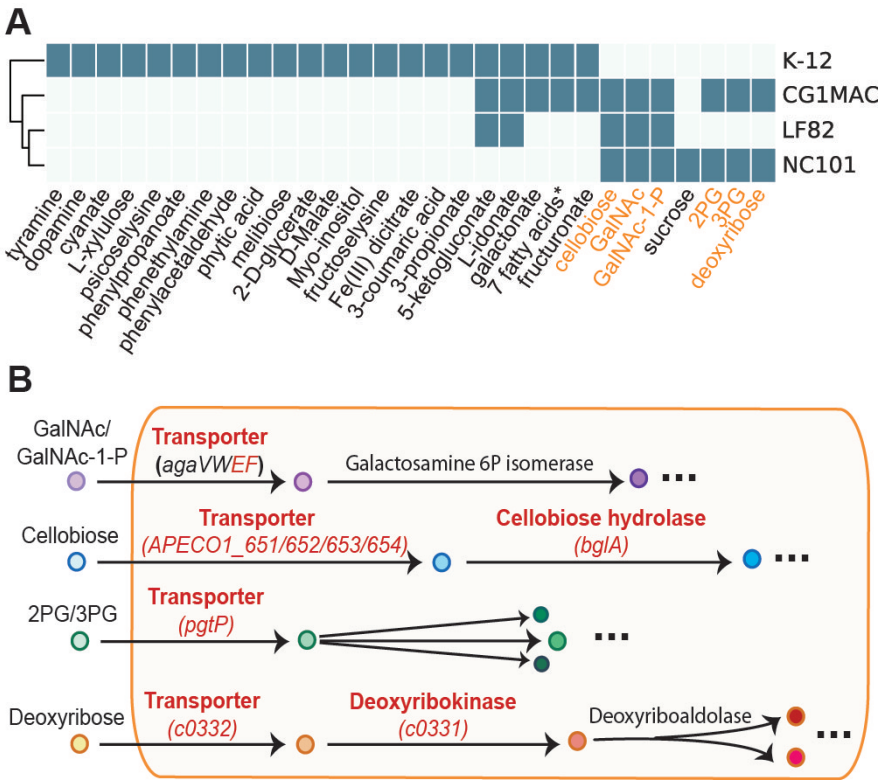


Figure 3.5: Simulation results of four GEMs. (A) Growth capabilities on various nutrient sources can be used to differentiate between strains. (B) The key pathways involved in the capability to catabolize the six highlighted substrates. Enzymes in red are missing from *E. coli* K-12.

3.4 Discussion

In this study we performed metagenomic-based, strain-level analysis of *E. coli* in a time-series of stool samples from a CD patient. The key findings are as follows: 1) The *E. coli* community was highly dynamic in this patient, with different relative abundance and dominant strains at different time points. 2) We were able to extract strain-level *de novo* assemblies of seven dominant strains from metagenomics data, and showed large variation in genomic content among strains using a pan-genome analysis. 3) Comparative genomic analysis and metabolic network reconstruction suggest ST1 (isolated during peak inflammation) resembles known AIEC reference strains NC101 and LF82, while other strains collected during stable states displayed diverse characteristics. 4) To assess the accuracy of de novo assemblies from metagenomics data, we isolated ST1 (named CG1MAC) from the stool sample, sequenced and experimentally characterized it. 5) We then built a complete genome-scale metabolic model of CG1MAC and assessed its growth capability.

Detailed time-series data not only showed intestinal dysbiosis of this patient, but also revealed the dynamics of his gut microbiome at strain level. Although recent studies have already shown dramatic fluctuations in both composition and function of the gut microbiome of IBD patients [37, 85], and linked it to disease development [130], they only focused on species level evaluations. In this study, however, we presented strain-level dynamics of the *E. coli* community: not only did relative abundance of *E. coli* vary over time, we also identified seven strains that dominated the *E. coli* community at different time points, which are later shown to have diverse gene contents and phylogenetic origins by *de novo* assemblies.

Strain-level analysis of the dominant *E. coli* strains and their correlation with metadata led us to hypothesize that only certain *E. coli* strains with specific features contribute to IBD

progression. Comparative genomic analysis and metabolic network reconstructions suggest similarity in both virulence factors and metabolic functions between ST1 (collected during peak inflammation) and known pathogenic IBD isolates NC101 and LF82. Evidence from literature suggests that the AIEC pathotype, to which both LF82 and NC101 belong, is implicated in IBD. However, we isolated and experimentally characterized ST1 (later named CG1MAC), and found that it does not display AIEC phenotypes. Interestingly, previous studies focused on clinical isolates have also isolated non-AIEC strains from IBD patients, as well as AIEC strains from healthy controls [80]. These results suggest that strains capable of eliciting an inflammatory response in IBD patients may share certain features, but they may not necessarily belong to the AIEC pathotype. Although this result needs to be further verified, both experimentally and in a larger cohort, it illustrates the importance of strain-level evaluations of gut microbiome. Another aspect that needs to be taken into consideration in future study is the association between the strain-specific features and the subtypes of IBD (ileal CD, colonic CD and ulcerative colitis), as research has shown that the three subtypes are genetically determined and may be triggered by different external factors [131].

Moreover, with the sequence of CG1MAC, we confirmed the validity of the *de novo* assemblies, and characterized the growth capability of CG1MAC with an accurate GEM. Strain-level *de novo* assemblies have not been widely adopted in microbiome studies, but we illustrated the potential and feasibility of such analysis, as the ST1 assembly accurately captures 95.5% of the actual genome content. On the other hand, another powerful tool - GEMs, allowed us to predict that: CG1MAC, along with NC101 and LF82, are able to utilize substrates that are either abundant in the human gut (including cellobiose and mucus glycan), or substrates that potentially enable them to outcompete other strains such as deoxyribose [126, 127].

Additionally, medication also plays an important role in gut microbiome composition. Antibiotics including Ciprofloxacin and Metronidazole have been shown to lower bacterial diversity and decrease abundance of enterobacteria [132], while corticosteroid such as Prednisone and Uceris may contribute to substantial shift in gut microbiota [133]. Additionally, this patient has also taken mesalamine (Lialda) that has been shown to decrease abundance of Escherichia and Shigella [32]. We observe in this patient that after taking Ciprofloxacin, Metronidazole and Prednisone in February 2012, the CRP level dropped dramatically, while the alpha diversity also decreased. After taking Uceris and Liada in 2013 from June to November, no more bleeding or flare was observed. However, more data and experiments are needed to obtain a comprehensive understanding of the impact of medications on microbiome structure and disease progression.

We also recognize some limitations of this approach that need to be addressed going forward: 1) The accuracy of *de novo* assembly at strain-level from metagenomics data needs to be carefully evaluated. Our study showed that such assembly does not capture the genome sequence at 100% accuracy, and such analysis is only possible for samples with high read coverage of *E. coli*. However, with metagenomics analytics tools being rapidly developed, the quality and feasibility of *de novo* assembly at the strain level are expected to be improved in the future. 2) We only examined metagenomics data, not gene expression level in this study. By including metatranscriptomics in the future, one should be able to describe functional states of microbes more accurately. 3) This workflow only allow us to examine the the dominant strains at each time point, while *E. coli* strains of lower abundance are not taken into consideration. Therefore, genetic variation in the *E. coli* community at each time point is not characterized. 4) Other factors that contribute to IBD need to be taken into consideration. Association between characteristics of *E. coli* strains and other elements such as host genomics, diet, and their microbial neighbors will

likely add valuable insights to future analyses. Overall, we believe performing such an analysis on a large cohort of IBD patients will greatly enrich our knowledge of IBD and the gut microbiome.

3.5 Conclusion

In this study, we observed the dominant *E. coli* strain in this patient varied over time. Particularly, the dominant strains isolated during peak inflammation is most similar to known pathogenic strains implicated in IBD, while other strains collected during more stable states have diverse properties. Overall, this pilot study illustrates that a strain-level analysis of *E. coli* from a time-series of stool samples can be very productive. The approach we utilized in this study not only captures the structure and dynamics of the entire microbiome community, but also allows a detailed evaluation of *E. coli* at the strain level. Due to decreasing sequencing cost, and fewer experimental procedures involved, this approach should also enable rapid and large-scale analyses in the future.

3.6 Methods

3.6.1 Metagenomics data generation

DNA was extracted from primary fecal samples using the MoBio PowerMag extraction kit (Qiagen Inc). Shotgun metagenomic libraries were prepared and sequenced at the sequencing core facility at the Institute for Genomic Medicine at UCSD. Briefly, libraries were constructed from each sample using 200 ng of extracted DNA, sheared to a target fragment size of approximately 400 bp using a Covaris E220 sonicator, and input to the TruSeq Nano PCR-based library prep kit (Illumina Inc), with samples individually indexed using dual 8 bp barcoded adapters.

Amplified libraries were then pooled and sequenced on a HiSeq4000 instrument. Sequenced reads were trimmed of adapter sequences and quality-filtered using Skewer [134] (end-quality trimming parameter of Phred 15 and a minimum length setting of 100 bp after trimming) and cutadapt [135] v1.15 (parameters -m 36 -q 20 -a ATCGGAAGAGCACACGTCTGAACTCCAGTCAC, -AATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT). Trimmed sequences were then filtered of human-derived reads using Bowtie2 [55] under the 'very-sensitive' setting, only retaining read pairs for which neither pair mapped to the human reference.

3.6.2 Metagenomics data analysis

Taxonomic profiles of the metagenomics data were evaluated using MetaPhlan2 [88] with default parameters. We extracted *E. coli* relative abundance from the result and compared it across samples. In addition to MetaPhlan2 analysis, we also performed additional analysis to confirm the presence of bacteriophages in a sample using Bowtie2 by mapping sequencing reads to genome sequences of the two phages (Langmead and Salzberg, 2012).

Alpha and beta diversity of the metagenomics data were calculated with the python package skbio [136]. We utilized the previously calculated taxonomy profiles at species level as the input OTU tables for diversity calculation. We calculated alpha diversity using the metric "observed_otus" and beta diversity using the metric "braycurtis". We then performed principal coordinate analysis, and plotted the PC1,2,3 using the same python package skbio.

3.6.3 Characterization of the dominant *E. coli* strains using single nucleotide variant (SNV) frequencies

First, MIDAS pipeline (database v 1.2) [11] was used with defaults to call genome-wide SNVs for all abundant species within individual samples. SNVs frequencies information for *E. coli* 58110 (representative genome for *E. coli* species in the MIDAS database) was merged across samples. Minor allele frequencies at particular genomic sites were calculated across all samples (positions chosen by MIDAS, columns reordered with respect to their hierarchical clustering). The heatmap suggests that *E. coli* population within most of metagenomic samples was dominated by single strain (only values close to 0 or 1 are observed in respective rows) that changed over time.

We then performed refined computational analysis of the SNV frequencies to confirm this hypothesis and identify samples with the same dominant *E. coli* strain. For each metagenomic sample, MIDAS pipeline with parameter "–species_id Escherichia_coli_58110" was used to compute per-base coverage and SNV frequencies for the *E. coli* reference. To avoid various artifacts, we then discarded sites with aberrant coverage as follows: positions with coverage 0, positions with coverage less than twice the median across the remaining sites, and positions with coverage falling within low/high 10% of the coverage values across the remaining sites.

To test whether the *E. coli* population within the sample is dominated by a single strain we then analyzed variant allele frequencies at the remaining positions. Specifically, we considered population as dominated if less than 0.05% of the positions had a minor variant frequency exceeding 10%. All but two samples (TP23 and TP27) satisfied this condition.

We further attempted to divide the remaining 19 samples into groups dominated by the same strain. We define the similarity between the pair of samples as a fraction of positions

in which the major variants matched (only the sites retained in the analysis of both samples were considered). Single-linkage clustering with a 99.9% threshold was used to obtain 7 groups of samples each corresponding to a particular *E. coli* strain. A single sample has been chosen within each group to attempt the reconstruction of the strain genome via *de novo* assembly (see section "Assembling dominant *E. coli* strains from metagenomics data").

3.6.4 Assembling dominant *E. coli* strains from metagenomics data

metaSPAdes assembler v3.11.1 with default parameters has been used to perform *de novo* assembly of 7 individual metagenomic samples (12/28/2011; 4/3/2012; 8/7/2012; 7/14/2013; 3/23/2014; 8/25/2014; 9/28/2014).

Resulting scaffolds and their coverage depths (average 56-mer coverage reported by metaSPAdes) were provided as input to MaxBin2 [137]. Each sample contained a bin annotated as *E. coli* with an estimated completeness exceeding 97% (as reported by CheckM [138]), which was used as a draft assembly for the downstream analysis. We have also considered including smaller bins annotated as *E. coli* by MaxBin2, but it has resulted in sharp increase of the contamination level (as reported by CheckM).

3.6.5 Phylogenetic analysis and pan-genome construction of the seven assemblies

We first annotated the assemblies using Prokka [98] with default parameters. The output files from Prokka were then used to perform phylogenetics analysis and pan-genome reconstruction. To perform phylogenetic analysis using PhyloPhlan [99], we utilized the protein FASTA files ending in ".faa" from Prokka output, and constructed the phylogenetic trees with 110 other

E. coli strains with known phylogroups to infer phylogroup of each assembly. To construct pan-genome of the seven assemblies, we used Roary [103] that takes input files ending in ".gff" , which contain the master annotation in GFF3 format produced by Prokka. We set the parameter "minimum percentage identity for blastp" to 80.

3.6.6 Virulence factor analysis

We mapped genome assemblies of the dominant strains against two sets of virulence factor references. The first set is the curated virulence factors collected from the VFDB database [110]. The second set is 57 genes identified from literature that are associated with AIEC strains, which are implicated in IBD. These genes are mainly identified and collected according to the review paper by Palmela et al.[78]. Note that the 57 genes contain variants of genes that perform the same functions. We used BLAST [111] to map the assemblies to the references and considered genes to be present when the sequence similarity is greater than 80%.

3.6.7 Metabolic network reconstruction and pan-reactome matrix analysis

The draft metabolic reconstructions of *E. coli* strains are created based on a previous multi-strain *E. coli* study [92]. We first created an *E. coli* pan model that combines all the genes, reactions, and metabolite in the 55 *E. coli* models reconstructed by Monk [92]. To incorporate the most recent update in *E. coli* reconstruction, we also added the content of the latest K-12 model iML1515 [139] to the pan model. Since all included *E. coli* strains span various pathotypes and phylogenetic origins, the pan model created is considered a comprehensive representation of metabolic functions in *E. coli* strains, as well as a good starting point for metabolic network reconstruction. We then mapped the sequences of strains of interest to all the genes in the pan

model using BLAST [140], and set a threshold of 80% for both gene similarity and alignment length, in order for a gene to be considered present in the strains. The missing genes and their correlated reactions and metabolites in each strain are removed from the pan model to create strain-specific metabolic network reconstructions. The metabolic network was reconstructed using the python package COBRApy [141].

To compare the metabolic networks of the 7 dominant strains and 3 reference strains, we then created a binary matrix of size 10 by 3,077 that records the presence and absence of each reaction in all 10 strains. To determine the similarity in metabolic functions in 10 strains, we performed MCA analysis using python package mca [59] with Benzecri correction, with the parameter of TOL set to 1e-9. To extract the important reactions in factor 1 and factor 2, we identified the top 50 reactions that has the highest contribution to these two factors.

3.6.8 Isolation of bacterial Strains: CG1MAC and 3_2_53FAA

In order to isolate CG1MAC from the stool sample, we diluted the sample in saline and plated dilutions on McConkey agar to select for *E. coli* isolates. All obtained isolates were picked, and gDNA was extracted using a Qiagen stool mini kit. To verify the isolates that identified with the predicted genotype of the target strain, four genes were used, *fyuA, vasD, xerD, gsp*, to which we designed PCR primers based on sequence data from the *de novo* assembly of ST1. Comparative analysis showed that these genes were present in the metagenomic dataset obtained from the originating stool sample, and are more prevalent in IBD-associated *E. coli* strains. There were 40 strains obtained and screened by PCR in this way, and all were found to positively identify with the ST1 assembly. Of the clones, one was selected for further analysis, and named CG1MAC.

Strain 3_2_53FAA was isolated from an inflamed biopsy sample from the descending colon of a 52 year old male left-sided CD patient in a Calgary, Canada clinic in 2007. The patient had an initial diagnosis of ulcerative colitis which was later changed to Crohn's colitis (ileal biopsies were normal). Strain 3_2_53FAA was placed into the Human Microbiome Project reference genome collection as HM-38, and as such was genome sequenced by the Broad Institute (GenBank assembly accession number GCA_000157115.2).

Both CG1MAC and 3_2_53FAA were serotyped by The National Microbiology Laboratory (Public Health Agency of Canada) at Guelph, Ontario

3.6.9 Bacterial genome sequence

We sequenced the genome of isolated *E. coli* strain CG1MAC. First, we isolated and purified gDNA from pelleted cells using the Macherey-Nagel NucleoSpin Tissue Kit (Catalog number 740952.50) following the manufacturer's protocol, including RNAse treatment. Second, we prepared a genomic DNA library using a KAPA HyperPlus Library Preparation Kit (catalog number KK8514) incorporating dual indices during the PCR amplification step, and checking quality with TapeStation. Eventually we pooled the library and sequenced using the Illumina HIseq 4000 instrument with paired-end and 100/100 reads settings.

We used SPAdes [142] to assemble the high quality reads with default parameters. The assembled genome has been submitted to NCBI with accession number QLAC00000000.

3.6.10 Confirmation of CG1MAC isolate identity with SNV analysis

We used genome-wide single nucleotide variant (SNV) frequencies analysis to verify that: 1) the population of *E. coli* in the TP1 metagenomic sample is dominated by a single subpopu-

lation; 2) Dominant subpopulation is represented by isolated CG1MAC strain.

Both TP1 and CG1MAC isolate reads were processed by MIDAS pipeline [11] with parameter –species_id Escherichia_coli_58110 to compute coverage and SNV frequencies for metagenomic and isolate sequencing reads against *E. coli* reference included in its database.

First we demonstrate that *E. coli* population in TP1 is likely dominated by a single strain. To avoid various artifacts we ignored positions with coverage falling within low/high 10% of the coverage values across all covered positions of the reference. Out of 2.85 million remaining sites only 181 had major allele frequency (MAF) not exceeding 90% (in comparison, CG1MAC sample had 96 of such positions), suggesting that a single strain accounted for the lion's share of *E. coli* population. Then we compared the predicted genotypes of the CG1MAC isolate and dominant *E. coli* strain in TP1. Only sites with MAF $\geq$ 90% and coverage falling within 10th and 90th percentiles in both samples were considered. While they cover 59% of the reference genome (total 2.48Mb), no differences were observed between the major alleles of the two samples, reliably indicating that the CG1MAC isolate originates from the dominant subpopulation.

3.6.11 Curation of the CG1MAC model

First, we created the draft metabolic reconstruction of CG1MAC following the procedure described in section 5.6. We then performed additional curation to improve the accuracy of the draft model. We annotated the genome of CG1MAC with RAST [143] and identified metabolic genes using Enzyme Commision (EC) numbers. We then identified 413 metabolic genes not included in the pan model, and looked into the reactions associated with them in Uniprot database [144] regarding their annotation score and experimental evidence. Among all identified reactions, we only added six to the model based on the following filtering criteria: 1) Have a complete EC

number with four numbers; 2) Not involved in DNA/RNA modification, as suggested by the established GEM reconstruction protocol [145]; 3) experimentally proven to be present in *E. coli*; 4) have a defined reaction with specificity; 5) do not duplicate with existing reactions. The majority of the identified reactions are already present in the model, as their encoding genes are variants of existing genes in the model. We then added the new reactions to the CG1MAC model and the 3 reference models whenever appropriate, to ensure the growth simulation performed on these four models is accurate. Finally, we performed the manual curation step for CG1MAC model following the established protocol [145]. Because the 55 existing models that the reconstruction was based on were already manually curated, we focused on curating newly added reaction. We removed reactions and metabolites in the wrong compartment, added in subsystem of new reactions, ensured the new reactions were mass/charge balanced, and checked gene-protein-reaction (gpr) of newly-added reactions.

3.6.12 Adhesion and Invasion assays on Caco-2 and THP-1 cells

To determine bacterial invasion in epithelial cells and survival in macrophages Caco-2, cells were maintained in DMEM + 10% FBS (Invitrogen). THP-1 cells were maintained in RPMI + 10% FBS (Invitrogen) in 5% CO2 humidified atmosphere at 37°C. Differentiation of THP-1 cells was achieved by treatment with 5ng/ml of PMA (Sigma-Aldrich) for 2 days. Cells were allowed to recuperate in normal media for 1 day before assay was performed.

Adhesion, invasion and survival assay were performed as described in [146]. Briefly, cell invasion analyses were carried out in Caco-2 cells cultured in DMEM without antibiotics, and maintained in 5% CO_2 and 37°C. Cell monolayers were infected with *E. coli* strains at multiplicity of infection (MOI) of 100, for 2h at 37°C. After the infection period, cells were washed with 3

x PBS and placed in fresh medium supplemented with gentamicin (50 µgml), incubated for 1 h at 37 °C, and lysed with 0.1% Triton-X-100PBS. Lysate serial dilutions were plated on LB agar (Invitrogen) and incubated at 37°C overnight. Cell adhesion analysis was also carried out in Caco-2 cells using similar infection conditions as described for invasion assays, but omitting the gentamicin treatment. Differentiated THP-1 cells were infected with *E. coli* strains (MOI = 100) for 2h at 37°C. Cells were then washed in PBS and placed in fresh medium supplemented with gentamicin (50 µg/ml). Intracellular bacterial content was determined at 1 and 24 hours post infection at 37°C and the ratio between bacterial content at each period was determined.

3.6.13 In Silico Growth Simulations

Growth simulation for CG1MAC, K-12, LF82 and NC101 were performed using COBRApy. We simulated growth in M9 minimal media, with the lower bound of exchange reactions for the following substrate set to -1000: Ca^{2+} , Cl^{-} , CO^{2}, Co^{2+}, Cu^{2+}, Fe^{2+}, Fe^{3+}, H^{+}, H_2O, K^{+}, Mg^{2+}, Mn^{2+}, $MoO_4{}^{2}$, Na^{+}, Ni^{2+}, $SeO_4{}^{2}-$, $SeO_3{}^{2}+$, and Zn^{2+}. Moreover, the default carbon, nitrogen, sulfur and phosphate sources are glucose, $NH_4{}^{-}$, $SO_4{}^{2}$, $HPO_4{}^{2}$. These reactions have lower bounds set to -1000. Another essential substrate is cob(I)alamin, for which the exchange reaction has a lower bound of -0.01. We evaluated if sole carbon, nitrogen, sulfur or phosphate substrate supported growth. To do so, we set the lower bound of the exchange reaction of the default substrate to 0, and added sole substrate by setting the lower bound of exchange reaction to -10. Additionally, we have simulated growth under aerobic condition by setting the lower bound of oxygen uptake to -10.

3.6.14 Growth experiments

The *E. coli* strains K12 and CG1MAC were grown in modified M9 media with the main carbon, nitrogen, or sulfur source replaced. For tests involving the replacement of the carbon source, glucose was omitted from the M9 media and 0.022 moles/L of the new carbon source was added in its place. For the nitrogen source replacement tests, NH_4Cl was replaced with 0.019 moles/L of the new nitrogen source and for the sulfur source replacement tests, $HPO_4 \bullet 7$ H_2O was replaced with 0.001 moles/L of the new sulfur source. For the media used in both the nitrogen and sulfur replacement tests, the glucose concentration was increased to 0.004 g/mL.

Freshly-cultured single colonies of each strain were selected after overnight incubation on blood agar plates and individually diluted in 5 mL of basal M9 media (no carbon, nitrogen or sulfur source) and 100 µL of each diluted strain was used to inoculate 5 mL of modified M9 medium containing the test carbon, nitrogen, or sulfur source. Inoculated tubes were incubated for 24 hours at 37°C with orbital shaking at 200 rpm to pre-expose cells to each metabolite. Following incubation, cells were pelleted at 5000 rpm for ten minutes and resuspended in 200 µL PBS buffer, whereupon the the optical density at 555nm was recorded. This value was used to normalize the amount of each culture that was added to 5 mL of the appropriate test medium in a glass test tube. Sample test tubes were incubated for 48 hours at 37°C with orbital shaking at 200 rpm, and 100 µL samples of these tubes were used to measure the optical density at 555nm which was recorded every 24hrs using a Wallac Victor 3 plate reader.

Metabolites used were: glucose (Fisher Scientific); melibiose, phenylacetaldehyde, trans-3-Hydroxycinnamic acid, oxaloacetic acid, dopamine hydrochloride, 2-Deoxy-D-ribose, iron (III) citrate, taurine, threonine and sodium thiosulfate (from Sigma Aldrich); cellobiose and D-arabinose (from Fluka); 3-(3-hydroxy-phenyl) propionate, D-arabinose, choline chloride, D-(+)-

galactose, 3-hydroxyphenylacetic acid, tyramine and methyl-4-hydroxyphenylacetate (from Alfa Aesar), and sucrose (from Bioshop).

3.6.15 Ethics statement

Patient that had the stool samples collected is consented under two protocols: HRPP #141853 American Gut Project and HRPP #150275 Evaluating the Human Microbiome. Both protocols were approved by University of California San Diego's Human Research Protection Program (HRPP). Written informed consent on dissemination of the result and scientific publication are also included in the approved protocols, and was obtained from the patient.

The patient in which strain 3_2_53FAA was isolated was recruited and consented through the Intestinal Inflammation Tissue Bank at University of Calgary and this study was approved through the Conjoint Health Research Ethics Board of the University of Calgary (Project Numbers; REB14-2429 and REB14-2430).

Acknowledgments

We thank Dr. Roger Johnson at the National Microbiology Laboratory in Guelph, Ontario, for serotyping expertise. This research is supported by Microbial Science Initiative Graduate Research Fellowship (UCSD Center for Microbiome Innovation), NIH Grant 1-U01-AI124316-01, Novo Nordisk Foundation Center for Biosustainability and the Technical University of Denmark (grant number NNF10CC1016517).

Chapter 3 in full is a reprint of material published in: **Fang, X.**, Monk, J. M., Nurk, S., Akseshina, M., Zhu, Q., Gemmell, C., Gianetto-Hill, C., Leung, N., Szubin, R., Sanders, J., Beck, P. L., Li, W., Sandborn, W. J., Gray-Owen, S. D., Knight, R., Allen-Vercoe, E., Palsson,

B. O., Smarr, L. (2018). Metagenomics-Based, Strain-Level Analysis of *Escherichia coli* From a Time-Series of Microbiome Samples From a Crohn's Disease Patient. *Frontiers in Microbiology*, 9, 2559. The dissertation author was the primary author.

Chapter 4

Escherichia coli B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities

4.1 Abstract

4.1.1 Background

Escherichia coli is considered a leading bacterial trigger of inflammatory bowel disease (IBD). *E. coli* isolates from IBD patients primarily belong to phylogroup B2. Previous studies have focused on broad comparative genomic analysis of *E. coli* B2 isolates, and identified virulence factors that allow B2 strains to reside within human intestinal mucosa. Metabolic capabilities of *E. coli* strains have been shown to be related to their colonization site, but remain unexplored in IBD-associated strains.

4.1.2 Results

In this study, we utilized pan-genome analysis and genome-scale models (GEMs) of metabolism to study metabolic capabilities of IBD-associated *E. coli* B2 strains. The study yielded three results: i) Pan-genome analysis of 110 *E. coli* strains (including 53 isolates from IBD studies) revealed discriminating metabolic genes between B2 strains and other strains; ii) Both comparative genomic analysis and GEMs suggested that B2 strains have an advantage in degrading and utilizing sugars derived from mucus glycan, and iii) GEMs revealed distinct metabolic features in B2 strains that potentially allow them to utilize energy more efficiently. For example, B2 strains lack the enzymes to degrade amadori products, but instead rely on neighboring bacteria to convert these substrates into a more readily usable and potentially less sought after product.

4.1.3 Conclusions

Taken together, these results suggest that the metabolic capabilities of B2 strains vary significantly from those of other strains, enabling B2 strains to colonize intestinal mucosa. The results from this study motivate a broad experimental assessment of the nutritional effects on *E. coli* B2 pathophysiology in IBD patients.

4.2 Background

Alteration of the composition of the gut microbial community has been implicated in inflammatory bowel disease (IBD) [7]. Several studies have shown that the abundance of *E. coli* in the gut microbiome of IBD patients is higher compared to healthy subjects [7, 78, 147]. In comparison with healthy controls, *E. coli* isolates from IBD patients mainly belong to B2 and D

phylogroups, including extraintestinal pathogenic *E. coli* strains (ExPEC) [148]. In particular, a specific *E. coli* pathotype, adherent-invasive *E. coli* (AIEC), has been shown to be a leading bacterial trigger of IBD [149]. AIEC strains mostly belong to B2 phylogroups [78]. They are able to adhere to intestinal epithelial cells and survive and replicate within macrophages, yet the specific genetic determinants of this pathotype are still unknown [150].

In recent years, several comparative studies were performed on *E. coli* isolates to understand their pathogenicity in IBD [82, 150, 151]. Additionally, a few specific *E. coli* strains associated with IBD have been characterized in detail, including LF82 [109], UM146 [152], and NRG857c [153], all of which are in phylogroup B2. Most of these studies have focused on comparative phenotypic assays and genome analysis such as virulence factor determination. A previous study has shown that strains in B2 phylogroup possess certain virulence factors including adherence genes, that allow them to persist within the human intestine, while strains in A and B1 phylogroups are primarily transient *E. coli* strains [154]. However, the systems biology of IBD-related *E. coli* strains, such as metabolic network reconstructions that elucidate nutrient niches, remains unexplored.

Genome-scale models (GEMs) represent a mathematical framework that enables a mechanistic description of metabolic functions and how they relate to physiological properties. GEMs have been used extensively to contextualize multi-omics data as well as to understand the genetic basis of phenotypic functions [18, 26, 90, 155]. The metabolism of *E. coli* strains has been studied extensively, enabling the development of GEMs for a wide range of *E. coli* strains. Recent studies have shown that strain-specific GEMs are necessary to capture the variation in metabolic capabilities in different strains [156], as the *E. coli* pan-genome is estimated to have more than 45,000 genes [97].

In this study, we analyzed the metabolic capabilities of B2 *E. coli* strains prevalent in IBD patients using pan-genome analysis and genome-scale metabolic models. We look at a large set of *E. coli* strains from IBD patients and healthy controls, as well as strains from other origins, to see if we could identify any common metabolic patterns associated with IBD pathophysiology in B2 strains. We showed that specific metabolic capabilities of the B2 group allow them to colonize intestinal mucus and become resident *E. coli* strains in the human gut.

4.3 Results

4.3.1 Strain collection studied

We collected available genomes of *E. coli* isolates from previous IBD studies - 53 *E. coli* strains (22 AIEC, 31 non-AIEC), most of which were isolated from intestinal biopsies of both IBD patients and healthy subjects. 52 of the 53 strains belong to B2 groups; however these studies did not include many genome sequences in other phylogroups from healthy controls [150]. Thus, we set out to compare these isolates with 57 other *E. coli* strains including commensal strains and those that exhibit extra-intestinal and intra-intestinal (InPEC) pathotypes. Of the 57 other *E. coli* strains, 14 strains belong to phylogroup B2, and the other strains span various phylogroups.

4.3.2 Strains in B2 phylogroup contain distinct metabolic genes compared to strains in other phylogroups

To identify important metabolic features in B2 *E. coli* strains, we first constructed the pan-genome from the 110 strains, including 53 strains isolated from both IBD patients and healthy controls. A pan-genome for the 110 strains was built using CD-HIT [157] with 80%

similarity setting (see Methods). We found an open pan genome with 16,091 orthologous genes, among which 2,979 are metabolic genes annotated by Enzyme Commission (EC) numbers. Out of all the metabolic genes identified, only 1,081 clusters are conserved across 110 strains. We then further investigated the distribution of the 1,898 accessory metabolic genes in 110 strains.

We found that most B2 strains have distinct metabolic genes compared to strains in other phylogroups (Fig. 4.1A). Metabolic genes highlighted in the red box in Fig. 4.1A are missing from most B2 strains, while genes highlighted by the orange box are more prevalent in B2 strains (present in <15% non-B2 strains and >80% B2 strains). We then selected the 100 most differentiating metabolic genes between B2 strains and strains in other phylogroups using the SelectKBest function from scikit-learn package [158] (see Methods). Of the selected genes, 53 genes are more prevalent in B2 strains and encode various functions including energy production, amino acid metabolism, carbohydrate metabolism, and metal binding. GO enrichment analysis [159] suggested that these genes are only enriched for tricarboxylic acid (TCA) cycle (False discovery rate (FDR) adjusted p-value = 3.89×10^{-2}). Upon further investigation, we found that B2 strains possess an extra set of *sucABCD* genes that share ~50% sequence identity with the original *sucABCD* genes present in all strains. These four genes encode the important enzymes in the TCA cycle: alpha-ketoglutarate dehydrogenase (*sucAB*) and succinyl coenzyme A synthetase (*sucCD*) [160]. Experiments are needed to characterize the function and importance of these gene variants in B2 strains. The remaining 47 metabolic genes that are primarily absent from B2 strains are enriched for folic acid catabolism (FDR adjusted p-value = 4.52×10^{-2}), 3-phenylpropionate catabolism (FDR adjusted p-value = 7.65×10^{-4}) and putrescine catabolism (FDR adjusted p-value = 1.97×10^{-3}). To explore the relationship between the metabolic functions and nutrient niches, we further investigated specific metabolic genes.

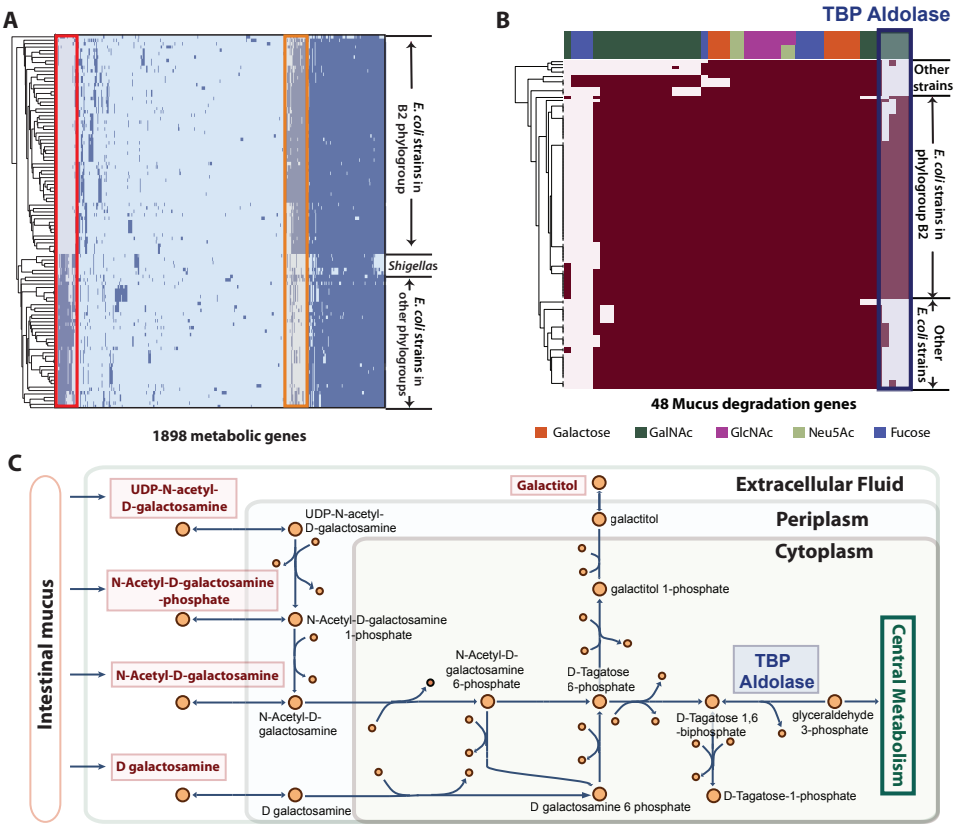


Figure 4.1: Pan-genome analysis shows B2 strains contain distinct metabolic genes. (A) 110 strains are clustered by the presence and absence of 1898 accessory metabolic genes. Genes in the red box are primarily absent from B2 strains, while genes in the orange box are more prevalent in B2 strains. (B) Presence and absence of genes involved in mucus degradation in 110 *E. coli* strains (genes are colored based on their functions in monosaccharides degradation). The four highlighted genes annotated as tagatose bisphosphate (TBP) aldolase are more prevalent in B2 strains. (C) Metabolic pathways of degradation of five nutrient sources involve TBP aldolase

4.3.3 IBD isolates and other ExPEC strains in the B2 phylogroup contain unique metabolic genes that enable them to utilize mucus glycan

We focused on elucidating the metabolic genes that allow *E. coli* strains of the B2 group to reside within intestinal mucosa. Glycans of the intestinal mucus can be utilized as a source of carbon and energy by intestinal microbiota, and depletion of mucus is associated with Crohn's disease [161]. Research has shown that commensals are mostly involved in cleavage of glycans into monosaccharides, while pathogens such as *E. coli* utilize the five monosaccharides released by commensals: fucose, galactose, N-acetylgalactosamine (GalNAc), N-acetylglucosamine (GlcNAc), and N-acetylneuraminic acid (Neu5Ac) [125]. Therefore, we performed comparative analysis on 48 genes involved in mucus degradation among the 110 strains. These genes were identified from a previous study on degradation of mucin glycans [125] (see Methods). The resulting heatmap (Fig. 4.1B) illustrated that although many genes have similar distribution among all 110 strains, four genes that are involved in tagatose 1,6-bisphosphate (TBP) aldolase are more prevalent in the B2 phylogroup (highlighted in Fig. 4.1B). These genes are also present in a few D strains, which was expected since both B2 and D strains are commonly found in IBD patients [148]. TBP aldolase converts TBP to dihydroxyacetone phosphate and glyceraldehyde-3-phosphate that is subsequently fed into central metabolism (Fig. 4.1C). Two of the four genes are identified to be variants of known TBP aldolase subunit GatY, while the other two genes are annotated to be TBP aldolase and related Type B Class II aldolases, but have not been well-characterized.

We then performed structural analysis to confirm the substrates and functions of the four genes annotated as TBP aldolases. We obtained homology models for the four proteins and compared them against the crystallized structure of the known TBP aldolase [162] and fructose-

1,6-bisphosphate (FBP) aldolase [163], since these two enzymes are highly similar. The models were found to be more structurally similar to the known TBP aldolase, rather than the FBP aldolase. This conclusion mainly arose due to an extended sequence of amino acids in FBP aldolase compared to TBP aldolase. This sequence extends the α10 loop - α11 arm [162] that results in the main differentiating feature between the enzymes' monomer subunits. Additionally, differences in the substrate binding sites lead to steric restrictions in FBP aldolase that constrain its substrate to be highly specific for FBP. All four predicted TBP aldolases contain different sets of residues, suggesting that they have the potential to greatly alter these steric restrictions and allow a wider range of substrates (including TBP) to enter the binding site.

The presence of these additional TBP aldolases potentially gives B2 strains an advantage to thrive in intestinal mucosa, as TBP aldolase is an important enzyme that is involved in the degradation of GalNAc and its derivatives [125], as well as galactitol (Fig. 4.1C). These B2 strains are likely to be more efficient in breaking down these nutrient sources produced from mucus glycan, thus having an advantage to survive in intestinal mucosa. Based on these observations of differentiating metabolic features, we next utilized genome-scale models to obtain a systems-level understanding of the metabolic capabilities of B2 and other strains.

4.3.4 Reconstruction of draft genome-scale metabolic models for 110 strains

GEMs can be used to systematically determine the metabolic capabilities of a strain [155]. We built GEMs of the 110 strains by mapping their genomes to a pan-metabolic model that contains all the reactions and genes collected from a previous *E. coli* multi-strain study [156] (see Methods). We first identified 2,485 core metabolic reactions that are present in all 110 GEMs, and 441 accessory reactions that are absent from at least one GEM. Functional distribution of

pan and core reactions indicates that most accessory reactions are involved in transport processes, carbon metabolism and cell envelope biosynthesis (Fig. 4.2A), suggesting that these strains are adapted to their own nutrient niches. Transporters in bacteria are adapted to their environment in order to best utilize the nutrients available [164]. Moreover, some accessory reactions in the category of cell envelope biosynthesis are involved in the synthesis of lipopolysaccharides (LPS), molecules also known as endotoxins, that contribute to the pathogenicity of *E. coli* strains. The toxic portion of LPS, lipid A, induces a release of host proinflammatory cytokines and causes infection within the host [165]. These models illustrate potential variation in LPS components, which could directly correlate with host inflammatory state in IBD patients.

We specifically examined the distribution of reactions in B2 and non-B2 strains. We investigated the 26 reactions that are unique to B2 strains, and identified three reactions that exist in more than 80% of B2 strains: manganese ATP-binding cassette (ABC) transporter, arabino-3-hexulose-6-phosphate isomerase, and reversible dihydrolipoamide dehydrogenase (Fig. 4.2B). Strains in both B2 and non-B2 groups are able to transport manganese via permease, while only B2 strains are able to transport manganese via ABC transporter. Knowledge of the other two enzymes is limited, and are thus potential experimental targets. In addition, three other transport reactions involved in the uptake of phosphoenolpyruvate, D-glycerate 2-phosphate, and D-glycerate 3-phosphate are also present in more than 30% of B2 strains. This is due to the presence of the *pgtP* gene, originally found in *Salmonella*, which is responsible for phosphoglycerate transport [166]. Thirteen reactions are missing from all B2 models, but were later found to be uncommon in non-B2 strains, as well. To further elucidate the differences in metabolic functions between B2 and non-B2 strains, we simulated growth of these strains on a variety of nutrient sources.

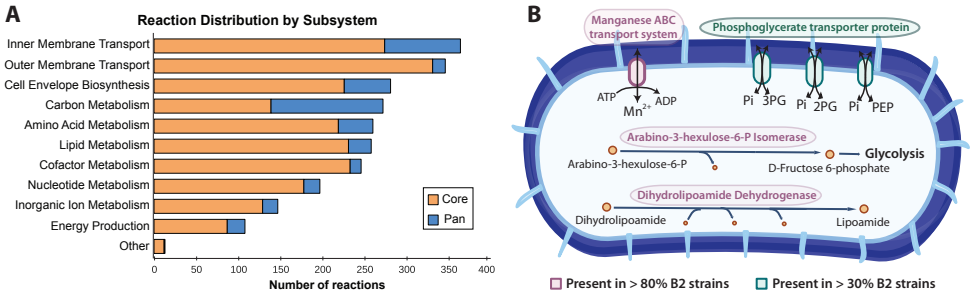


Figure 4.2: Reactions distribution in 110 GEMs. (A) Distribution of pan and core reactions in different systems for 110 *E. coli* models. (B) Unique reactions in models of B2 strains. Reactions present in more than 80% and 30% of B2 models are shown

4.3.5 Comparative analysis of GEMs highlights metabolic capabilities unique to B2 *E. coli* strains

Growth simulation of GEMs predicted that strains in the B2 group, including 52 isolates from IBD studies, share similar metabolic capabilities (Fig. 4.3A), regardless of the IBD status of their hosts. Growth simulations were performed for 649 substrates under aerobic conditions, as research has shown that aerobic respiration is required for *E. coli* to colonize the mouse intestine [167]. B2 strains isolated in IBD studies displayed distinct metabolic capabilities compared to other InPEC strains, including Enterotoxigenic *E. coli*, Enteropathogenic *E. coli*, and Enteroaggregative *E. coli*, but are similar to ExPEC strains in B2 groups such as Uropathogenic *E. coli* strains. This result is interesting since a subset of AIEC and other InPEC strains all colonize epithelial cells in the small intestine [7, 168] and thus likely share a preferred microenvironment, yet they display distinct metabolic capabilities. Specifically, B2 strains were predicted to be unable to grow on certain substrates, including psicoselysine, fructoselysine, meliobiose, cyanate, phenylpropanoate and L-Xylulose (Table 4.1)

We then investigated the most differentiating nutrient sources between B2 and non-B2 strains: fructoselysine and psicosylysine, also known as amadori products, that are abundantly

Table 4.1: Growth substrates that differentiate *E. coli* strains in B2 phylogroup from other strains.

	Phylogroup B2			Other phylogroups		
Growth Substrates	**AIEC IBD (%)**	**Commensal IBD (%)**	**ExPEC (%)**	**InPEC (%)**	**Commensal (%)**	**Shigella (%)**
Fructoselysine	0	3.2	0	90.9	69.2	87.5
Psicoselysine	0	3.2	0	90.9	69.5	87.5
Melibiose	4.6	6.5	33.3	81.8	57.7	100
L-Xylulose	4.6	6.5	33.3	45.5	69.2	12.5
Cyanate	4.6	6.5	33.3	90.9	65.4	0
Phenylpropanoate	4.6	6.5	33.3	90.91	65.4	12.5
Xanthosine 5'-phosphate	77.3	93.6	77.8	45.5	38.5	0
Xanthosine	77.3	93.6	77.8	45.5	38.5	0

formed in heated food and decomposed by microorganisms in the large intestine [169]. Further investigation using GEMs suggested that both the fructoselysine transporter and *frl* operon, including fructoselysine 6-kinase and fructoselysine 6-phosphate deglycase, are missing from *E. coli* strains in phylogroup B2, resulting in their inability to metabolize fructoselysine and psicoselysine. This result is consistent with experimental data describing growth of mutant *E. coli* strains on fructoselysine [170]. It is possible that B2 *E. coli* strains do not use these substrates directly, but instead use their derivatives produced by other organisms. Research has shown that *Intestinimonas AF211* and related bacteria that are abundantly present in colonic samples are able to convert amadori products into butyrate [171], a substrate that could be metabolized by all *E. coli* strains in the B2 group, while 50% of the non-B2 strains failed to do so (Fig. 4.3B). This could potentially explain the lack of degradation enzymes for fructoselysine and psicoselysine in B2 strains: by dispensing these enzymes, B2 strains could rely on neighboring bacteria to convert these substrates into a more readily usable and potentially less sought after product. Additionally, butyrate plays an important role in maintaining intestinal homeostasis and has therapeutic potential for IBD patients [172]. The elevated abundance of B2 strains in IBD patients and their capability to metabolize butyrate could potentially be related to the decreased concentration of butyrate in feces of IBD patients [173] and inflammation.

Moreover, model simulations showed strains in phylogroup B2 differ from other strains

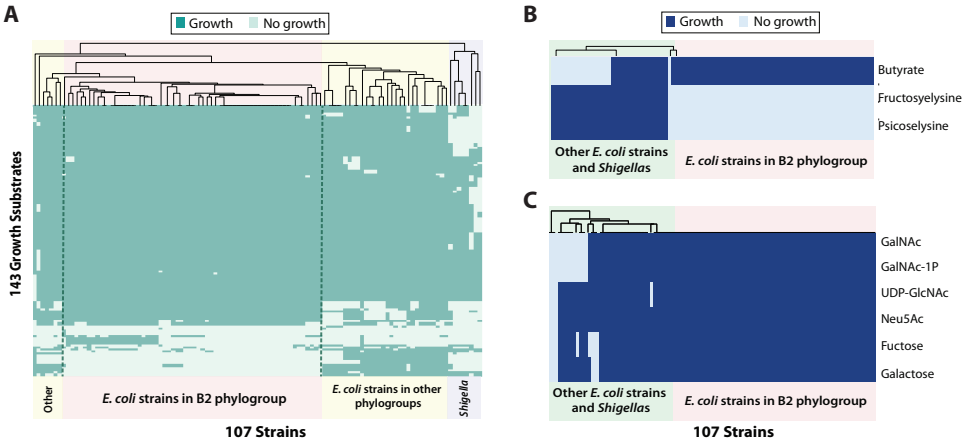


Figure 4.3: Simulated growth capabilities of 107 GEMs on various nutrient sources. (A) 107 strains are clustered by simulated growth capabilities on 143 differentiating nutrient sources. (B) Simulated growth on monosaccharides and their derivatives from mucus glycan. (C) Simulated growth on butyrate, fructoselysine and psicoselysine

in their ability to catabolize mucus monosaccharides. We examined the simulated growth capabilities of *E. coli* strains on five monosaccharides and their derivatives that are released from intestinal mucus glycan by commensals. Simulated growth results suggest that 100% of the B2 strains can utilize all tested monosaccharides as their sole carbon source, while only 65% of the strains from other phylogroups can utilize all six substrates tested (Fig. 4.3C).

4.4 Discussion

In this study, we delineated the specific metabolic capabilities of *E. coli* B2 strains that are found to be prevalent in IBD patients. Our study used pan-genome analysis of metabolic genes and the growth capabilities they confer. The study yielded three results: i) pan-genome analysis of 110 *E. coli* strains (including 53 isolates from IBD studies) revealed discriminating metabolic genes between B2 strains and other strains; ii) both comparative genomic analysis and GEMs suggested that B2 strains have an advantage in degrading and utilizing sugars derived from

mucus glycan, and iii) B2 strains display distinct metabolic features, such as their inability to catabolize fructoselysine and psicoselysine, but instead are able to utilize the derivatives produced by neighboring bacteria.

Pan-genome analysis of metabolic genes in 110 strains revealed that B2 strains have distinct metabolic genes. We identified genes that are unique to or more prevalent in B2 strains, including an extra copy of *sucABCD* variant that encodes two important enzymes in the TCA cycle. The importance and function of identified genes need to be experimentally characterized in future studies.

To evaluate the metabolic capabilities of these 110 strains on a systems level, we constructed draft models of 110 strains and examined their *in silico* growth capabilities. B2 strains showed differentiating growth capabilities on certain substrates (Table 4.1), including amadori products fructoselysine and psicoselysine, potentially because they are able to utilize a derivative of amadori products - butyrate, produced by their neighbouring bacteria.

Both pan-genome and GEM analysis showed that B2 strains have potential advantages that allow them to reside within the human intestinal mucosa. In addition to existing TBP aldolases, GatY and KbaY, that are involved in degrading mucus glycan component, B2 strains contain four extra variants of TBP aldolases, suggesting a potential advantage in utilizing intestinal mucus. Growth simulation with GEMs also suggested that all B2 strains are able to utilize all tested monosaccharides derived from mucus glycan, while 35% of other strains failed to do so.

Although we were able to identify common features among B2 strains, we could not further differentiate subgroups within B2 strains (e.g. AIEC strains versus non-AIEC strains, IBD isolates versus non-IBD isolates). To separate subgroups of B2 strains, we explored diverse datasets (e.g. a growth capability matrix and reaction content matrix generated from GEMs,

gene presence/absence matrix generated from pan-genome) using various methods including feature selection method, supervised and unsupervised clustering methods. Such attempts were not entirely successful due to the following reasons: 1. AIEC strains were shown to be a heterogeneous pathotype that displays different genotypes, as shown in previous studies [150]. Therefore, classification of AIEC strains based solely on genomic information remains challenging. 2. Other factors that affect IBD disease state were not taken into account in this analysis, including environmental conditions, host genetics and other microbial community members. A broader approach that takes these factors into consideration could provide valuable insight to the role of *E. coli* strains in IBD patients. 3. Our study utilized only genome sequences of *E. coli* strains, which only delineates the genetic potentials, but not functional states of these strains. Gene expression levels are unavailable for these strains, making it difficult for us to differentiate subgroups of B2 strains (e.g. isolates from healthy controls versus IBD patients). However, we hypothesize that the genes identified here that are unique to B2 strains may be upregulated and used during active IBD. This hypothesis remains to be tested, however. Thus, while we did not observe differences in genetic potential between subgroups of B2 strains, gene expression data would likely help differentiate IBD patient isolates from healthy control isolates based on the different functional states they are in.

4.5 Conclusion

Taken together, these results suggest that the metabolic capabilities of B2 strains vary from those of other strains, enabling them to colonize intestinal mucosa. The results from this study motivate a broad experimental assessment of the ability of B2 *E. coli* strains to utilize different substrates, and further investigations in if they confer growth rate advantages under

simulated intestinal conditions. If these strain-specific growth advantages are confirmed *in vitro*, the nutritional effects on *E. coli* B2 pathophysiology in IBD patients should be rigorously evaluated.

4.6 Methods

4.6.1 Bacterial genome sequences

We collected 76 genome sequences (including 39 AIEC strains) from various publications [83, 112, 117, 150, 152, 153, 174]. We recorded their associated metadata: IBD status of originating patient, anatomic site of collection, serotype, phylotype, and other relevant information where available. For comparison, we utilized genome sequences of 57 other *E. coli* strains that span various pathotypes as well as commensal strains, most of which are collected from a previous multi-strain *E. coli* study [156]. The quality of the genome sequences varied since they originated from multiple publications. Therefore, we calculated N50 scores of each genome sequence, and only performed analysis on 110 *E. coli* strains (including 53 IBD-associated strains) that have a N50 score greater than 200,000.

4.6.2 Pan-genome construction and analysis

We first annotated 110 *E. coli* genome sequences and aligned them against each other using CD-HIT [157] with the cutoff for "align average" set to 80%, so that genes with 80% or more sequence similarity are grouped together. We utilized the PATRIC database [175] to extract our sequences and gene calls. All annotations in this resource are called using the same pipeline that consists of assembly with SPADES [142] and annotation with RAST [143]. RAST annotation has also provided EC numbers that allow us to identify metabolic genes. With the alignment, we

created a binary matrix that describes the presence or absence of each gene in the strains. We extracted only metabolic genes with enzyme commision numbers. We then performed feature selection using SelectKBest function from the scikit-learn package [158] to select the top features that differentiate B2 and non-B2 strains.

4.6.3 Analysis of genes involved in mucus degradation

Genes that are involved in degrading the five monosaccharides derived from mucus were primarily identified from a previous study by Ravcheev and Thiele [125]. Gene sequences of the identified genes were collected from the supplementary file of the aforementioned paper. Additional genes involved in galactose metabolism were identified and added based on gene annotation and known pathways. Genome sequences of 110 strains were blasted against 48 identified genes with a threshold of 80% sequence similarity using BLAST [140].

4.6.4 Protein structural analysis of TBP aldolase

To inspect the possible functions of the additional four predicted class II TBP aldolases in B2 strains, we carried out a comparative analysis of each enzyme's predicted protein structure. Homology models were obtained from two modeling pipelines (SWISS-MODEL [176] and I-TASSER [177]) in order to compare results from different modeling approaches. Models were compared to the only crystallized structure of TBP aldolase (PDB ID: 1GVF [162]) and a structure of FBP aldolase (PDB ID: 1B57 [163]), which are both bound to a substrate analog of the natural substrate of TBP as well as the cations required for catalysis. Important residues for catalysis were gathered from Hall et al. [162] for comparison in all models. The two sets of homology models were found to be very similar in overall structure and location of these

important residues, and as a result the reported results do not differ between the generated models. For visualization, VMD [178] was used along with the MultiSeq plugin [179] to structurally superimpose all models.

4.6.5 Draft model reconstruction of other *E. coli* strains

We first created an *E. coli* pan model that combines all the genes, reactions, and metabolites in the 55 *E. coli* models reconstructed by Monk et al [156]. In addition, in order to incorporate any novel metabolic functions in the 110 strains that are absent from the previously-built *E. coli* models, we identified 340 metabolic genes in the constructed pan-genome that are absent from the previously studied 55 strains. However, the majority of the 340 genes are variants of existing genes, and only 96 genes may encode new functions. For these 96 genes, we utilized Uniprot [180] database to identify associated reactions, with the following criteria to select the reactions to include: 1) Not involved in DNA/RNA modification, as suggested by the established GEM reconstruction protocol [14]; 2) experimentally proven to be present in *E. coli*; 3) have a defined reaction with specificity; 4) do not duplicate with existing reactions in the 55 GEMs. In the end, we only identified five new metabolic reactions that fulfill all above requirements, mainly because these strains are not as well studied compared to the previous 55 strains, and little experimental evidence was found for the majority of the investigated metabolic functions. We then added these new reactions to the pan model created from the previous 55 models. To create strain-specific draft models, we mapped the 110 *E. coli* genome sequences to all the genes in the pan model using BLAST [140], and set a homology threshold of 80% for a gene to be considered present in the strain. The missing genes and their correlated reactions and metabolites in each strain were removed from the pan model to create strain-specific draft models.

4.6.6 *In silico* growth simulations

Growth simulation for *E. coli* draft models were performed using COBRApy [27]. We used M9 minimal media with the lower bound of exchange reactions for the following substrate set to -1000: Ca^{2+} , Cl^{-} , CO^{2}, Co^{2+}, Cu^{2+}, Fe^{2+}, Fe^{3+}, H^{+}, H^{2}O, K^{+}, Mg^{2+}, Mn^{2+}, MoO^{42}, Na^{+}, Ni^{2+}, $SeO_4{}^{2}-$, $SeO_3{}^{2}+$, and Zn^{2+}. Another essential substrate is cob(I)alamin, for which the exchange reaction has a lower bound of -0.01. In addition, the default carbon source is glucose with default lower bound set to be -20, while the default nitrogen, sulfur and phosphate sources are $NH_4{}^{-}$, $SO_4{}^{2}$, $HPO_4{}^{2}$ with the lower bounds all set to be -1000. We evaluated if sole carbon, nitrogen, sulfur or phosphate substrates supported growth by setting the lower bound of the exchange reaction of the default substrate to 0, and added sole substrates by setting the lower bound of exchange reaction to -10. We simulated growth under aerobic conditions with the lower bound of the oxygen exchange reaction set to -20. If the simulated growth rate is greater than 1% of the original growth rate (when all default nutrient sources are used), the strain is considered to be able to grow under the tested condition.

Among all 110 strains tested, three draft GEMs were not able to simulate growth on the majority of the substrates, potentially due to auxotrophy: *E.coli* str K-12 substr DH10B, *E.coli* O111 H-str 11128, *E.coli* NA114. These strains were therefore excluded from the following growth capability analysis.

We used SelectKBest function in scikit-learn package [158] to select the top 10 growth substrates that differentiate B2 and non-B2 strains, with the score function set to "f_classif". We then summarized the percentage strains in each pathotype that could utilize these substrates in Table 4.1. Note that in Table 4.1 we classified pathotypes to B2 group and non-B2 group, but with a few exceptions in both groups: i.e. non-B2 strains in the ExPEC group and B2 strains in

the commensal group.

Abbreviations

IBD: Inflammatory bowel disease; GEM: Genome-scale model; AIEC: Adherent-invasive *E. coli*; ExPEC: Extra-intestinal *E. coli*; InPEC: Intra-intestinal *E. coli*; EC: Enzyme Commission; TCA: Tricarboxylic acid; FDR: False Discovery Rate; GalNAc: N-acetylgalactosamine; GlcNAc: N-acetylglucosamine; Neu5Ac: N-acetylneuraminic acid; TBP: Tagatose 1,6-bisphosphate; FBP: Fructose-1,6-bisphosphate; LPS: Lipopolysaccharides; ABC: ATP-binding cassette.

Acknowledgments

This research is supported by Microbial Science Initiative Graduate Research Fellowship (UCSD Center for Microbiome Innovation), NIH Grant 1-U01-AI124316-01, Novo Nordisk Foundation Center for Biosustainability and the Technical University of Denmark (grant number NNF10CC1016517).

Chapter 4 in full is a reprint of material published in: **Fang, X.**, Monk, J. M., Mih, N., Du, B., Sastry, A. V., Kavvas, E., Seif, Y., Smarr, L., Palsson, B. O. (2018). *Escherichia coli* B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of intestinal mucosa. *BMC Systems Biology*, 12(1), 66. The dissertation author was the primary author.

Chapter 5

A systems-level evaluation of transcriptional regulatory network for *Escherichia coli*

5.1 Abstract

Transcriptional regulatory networks (TRNs) have been studied intensely for over 50 years. Yet even for the *E. coli* TRN–probably the best characterized TRN–several questions remain. Here we address three questions: (i) how complete is our knowledge of the *E. coli* TRN, (ii) how well can we predict gene expression using this TRN, and (iii) how robust is our understanding of the TRN? First, we reconstructed a high-confidence TRN (hiTRN) consisting of 147 transcription factors (TFs) regulating 1,538 transcription units (TUs) encoding 1,764 genes. The 3,797 high-confidence regulatory interactions were collected from published, validated chromatin immunoprecipitation (ChIP) data and RegulonDB. For 21 different TF knockouts, up to 63% of the differentially expressed genes were traced to the knocked out TF through regulatory cascades. Second, we trained supervised machine learning algorithms to predict the expression of 1,364

TUs given TF activities using 441 samples. The models accurately predicted condition-specific expression for 86% (1,174/1,364) of the TUs, while 193 TUs (14%) were predicted better than random TRNs. Third, we identified ten regulatory modules whose definitions were robust against changes to the TRN or expression compendium. Using surrogate variable analysis, we also identified three unmodeled factors that systematically influenced gene expression. Our computational workflow comprehensively characterizes the predictive capabilities and systems-level functions of an organism's TRN from disparate data types.

5.2 Background

Transcriptional regulatory network (TRN) plays a major role in enabling an organism to modulate expression of thousands of genes in response to environmental and genetic perturbations [181]. *Escherichia coli*'s TRN is probably the most extensively studied in any organism. Yet, the structure of even this TRN is still subject to considerable uncertainty, seriously limiting its utility for predicting gene expression or for interpreting disparate data sets. Indeed, over a decade ago, a combined metabolic and regulatory network model of *E. coli* could explain only 15% of differential gene expression in response to the major environmental change of oxygen deprivation [182]. While much progress has been made since then [183–185], predicting global gene expression remains a fundamental challenge [186].

A global TRN can consist of regulatory interactions determined from a variety of data sources. These include direct and indirect experimental evidence or computational predictions [187]. For the latter, reducing false positive interactions remains challenging—the state of the art achieves 60% precision [188, 189]. In recent years, improved chromatin immunoprecipitation (ChIP) methods have enabled precise characterization of transcription factor (TF) binding

sites. Combining ChIP with transcriptomics for TF KO strains have yielded high-confidence regulatory interactions in the conditions studied. Such ChIP studies have now accumulated for over a dozen major transcription factors, with each study increasing the number of known binding sites of a TF by 74% to 400% [190–193].

Here, we use this critical mass of data to perform a rigorous assessment of the latest high-confidence TRN (hiTRN) of *E. coli* that is devoid of uncertain regulatory interactions. We further examine this TRN's ability to explain differential gene expression in response to genetic and environmental perturbations. The hiTRN is reconstructed using published ChIP data for 15 TFs added to only high-confidence regulatory interactions from RegulonDB [187]. We assess our hiTRN using transcriptomics compendia [194–196] and multiple computational approaches: unsupervised and supervised machine learning, mutual information analysis, network topology analysis, integer programming, and community detection (Figure. 5.1).

5.3 Results

5.3.1 The coverage of the TRN has expanded, but remains incomplete.

Our reconstructed hiTRN consisted of 147 TFs, 1,538 TUs, and 1,764 genes regulated by 3,797 high-confidence regulatory interactions. We assessed the coverage of the hiTRN for explaining differential gene expression in three ways.

Completeness: To assess hiTRN's completeness, we computed what fraction of the 1,764 genes whose expression changed across 154 experimental conditions could be directly explained with the hiTRN (see Methods). On average, 27% of differentially expressed genes (DEGs) in the set of 1,764 genes considered were enriched for at least one regulon, and 20% or more of DEGs

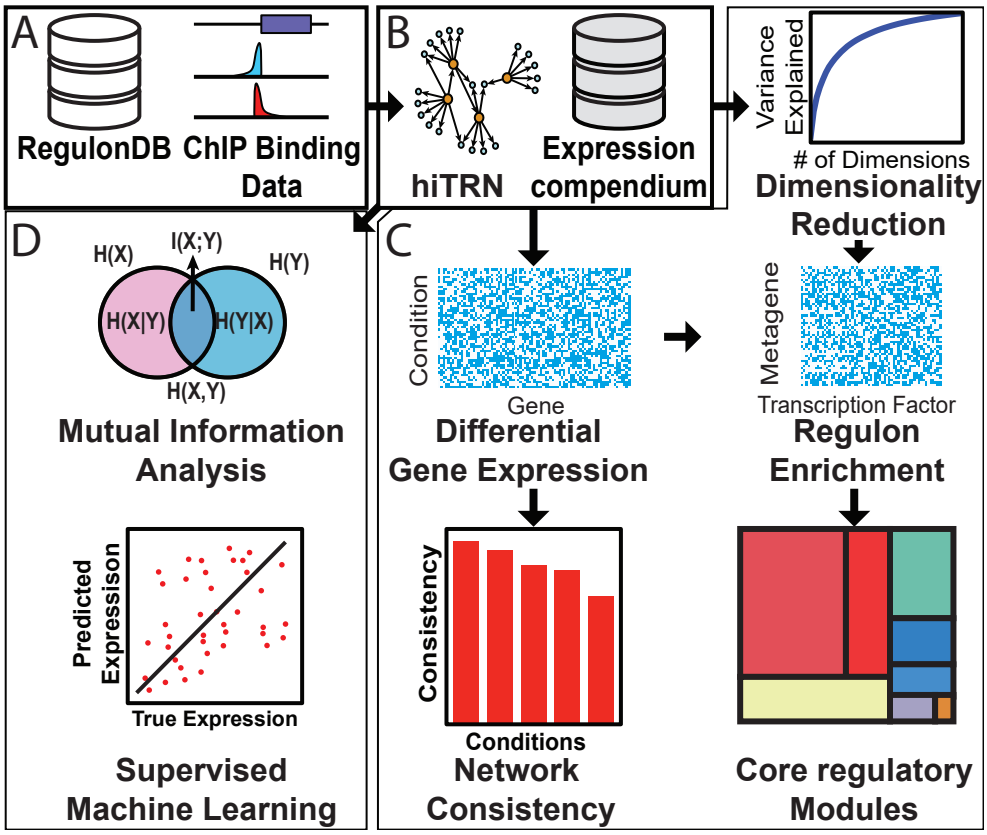


Figure 5.1: Overview of our workflow. (A) RegulonDB [187] and additional published ChIP data were combined to reconstruct the hiTRN. (B) Using our hiTRN, we analyzed transcriptome shifts in expression compendia (EcoMAC [194], *E. coli* Expression 2 [195], COLOMBOS [196]). (C) We evaluated the completeness of our knowledge of our hiTRN on the basis of network consistency, dimensionality reduction, and detection of stable regulatory modules. (D) We assessed the ability of our hiTRN to quantitatively predict gene expression using mutual information and regression.

were enriched for at least one regulon in 57% of the conditions.

Genetic perturbations: We then assessed whether differential gene expression could be traced through regulatory paths in our hiTRN to a knocked out TF gene. We investigated 21 different sets of single or double TF knockouts. We found that 0% to 63% of DEGs in the TRN were successfully traced to the knocked out TF through one or more regulatory paths (Fig. 5.2B). We could best explain DEGs in the TRN (50% to 63%) for experiments $\Delta arcA \Delta fnr$ and $\Delta purR$ (with adenine), while experiments $\Delta narP \Delta oxyR$ and $\Delta soxS \Delta purR$ (without adenine) were explained poorly. For comparison, we have also extracted 9 additional TF knockout experiments from the COLOMBOS dataset[196] and performed the same analysis. For the 4 experiments that have DEGs identified, we found around 56% of the DEGs in the TRN can be successfully traced to the knocked out TF.

Regulatory bias: We then evaluated whether the regulatory bias (activation or inhibition) assigned to each regulatory interaction was consistent with DEGs given a TF knockout. This "sign consistency" analysis was conducted by formulating the hiTRN as an influence graph [197], with edge signs reflecting activation (+) or repression (-). Overall, total sign consistency accounting for both differentially expressed and non-differentially expressed genes was 51–99% consistent with the TRN (Fig. 5.2A). The highest consistency was observed for local TFs such as *narL*, *narP*, *dnaA* and *purR*. Consistency was low for *arcA/fnr* under both conditions. We found a negative correlation between the number of genes regulated (both directly and indirectly) by a TF and the sign consistency between the TRN and experimental data (Pearson $r = -0.875$, $P = 2.10 \times 10^{-7}$). We also found a significant negative correlation between the longest regulatory path length of a TF and consistency. For the TF KO experiments from the COLOMBOS dataset, the overall consistency was similar (59% to 99%).

Not all DEGs could be traced to the deleted TF (Fig. 5.2B). Unreachable DEGs may indicate missing regulatory interactions, while low sign consistency suggests that additional factors influencing regulatory bias need to be explicitly modeled (e.g., effect of adenine on *PurR* binding [191]). Additionally, some differential expression may have been due to growth rate-dependent global regulation [185, 198] (growth rates ranged from 2% to 37% of the wild-type, where these meta-data were available).

Overall, the coverage of our hiTRN for DEGs varied across experimental conditions, with an average coverage of 26%, which was significant (permutation test, $P = 3.91 \times 10^{-4}$) and up to 63%. The low DEG coverage for individual TF knockout experiments reflects the highly interwoven nature of many regulons. Thus, achieving 100% DEG coverage will likely require precise reconstruction of individual regulons including mechanisms beyond TRN topology and regulatory bias.

5.3.2 The hiTRN is consistent with major modes of changes in the entire transcriptome

We evaluated our hiTRN in the context of transcriptomics data consisting of 4,189 genes $\times$ 441 samples. Transcriptomics data are difficult to interpret, in part because they are high-dimensional and noisy. We thus used non-negative matrix factorization (NMF) [199] to identify major modes (i.e., important features) of transcriptome changes across conditions. NMF identifies cohesive subsystems from complex expression data set by reducing thousands of genes into several dozen metagenes, which represent the major modes [199–201]. A metagene is a linear combination of the genes whose expression changes are correlated across conditions. Using NMF, we reduced the dimensionality of the expression data from 4,189 genes $\times$ 441 samples to

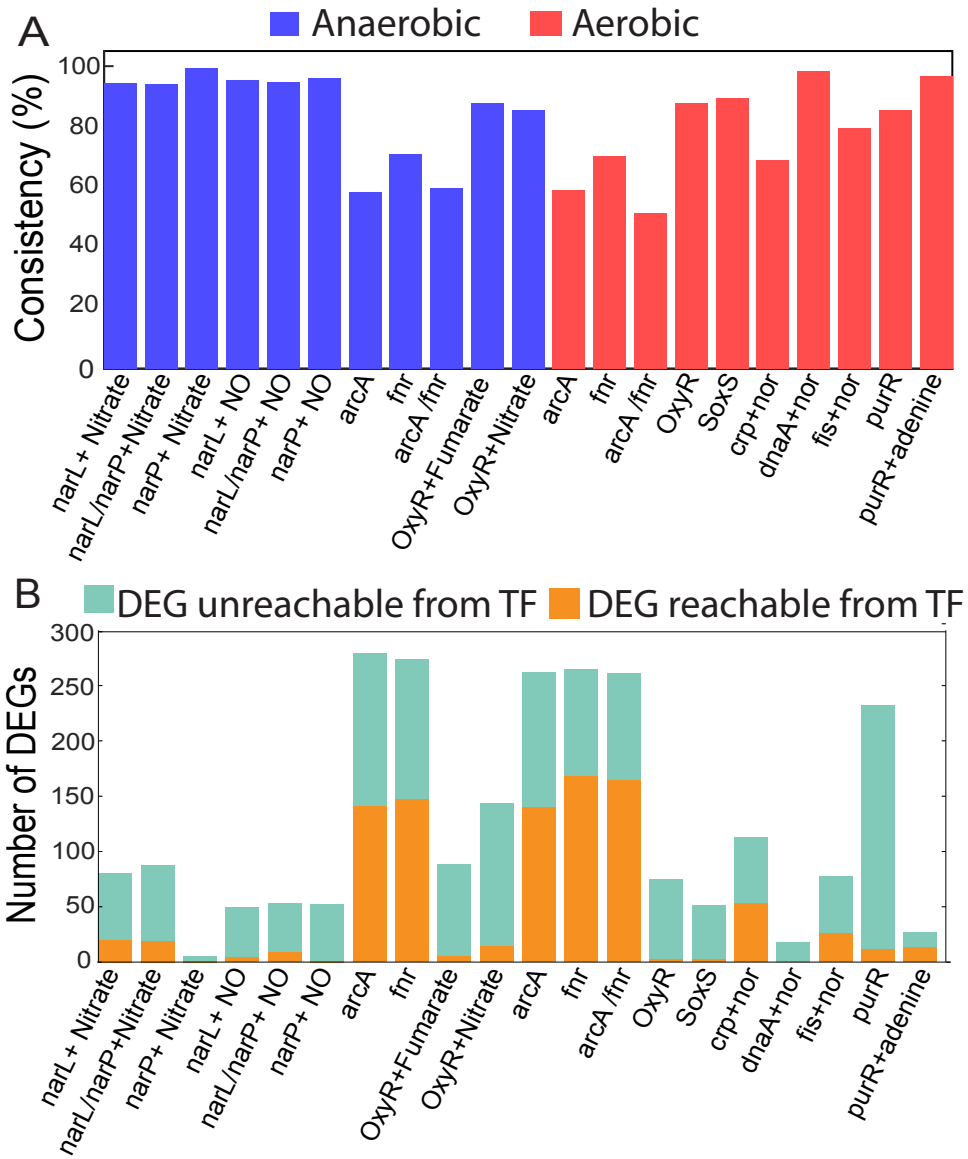


Figure 5.2: Consistency of hiTRN with observed differential gene expression in wild type cells and in strains with a deleted TF gene. (A) Consistency of our TRN with observed differential and non-differential gene expression accounting for regulatory bias (i.e., sign consistency). (B) Reachability (existence of contiguous regulatory paths in the TRN) from deleted TFs to DEGs in the TRN.

40 metagenes × 441 samples. We confirmed using Principal Component Analysis (PCA) that 40 dimensions sufficiently explained (88%) of variance in the expression data.

We then characterized the relationship between regulons and metagenes. To do so, we identified regulons that were enriched for genes that were determined to be major contributers within each metagene (i.e., genes that had large coefficients within a metagene). We found that all metagenes were enriched for at least one regulon (Fig. 5.3).

Furthermore, metagenes tended to be enriched for hiTRN-regulons that shared related functions (Fig. 5.3). For example, some stress response TFs (*rcsB, gadE, gadX* and *gadW*) were enriched simultaneously for several metagenes. This result is consistent with the hiTRN structure, which causally links these TFs: $rcsB \rightarrow gadX \rightarrow gadW \rightarrow gadE$ [187, 202]. Overlapping regulons were also co-enriched in the same metagene including *narL* and *narP*, or *nrdR* and *dnaA*.

These results demonstrate coverage and coherency in the hiTRN, and consistency with high-dimensional transcriptomics data, albeit at a more coarse-grained level than the reachability and regulatory bias described above.

5.3.3 Robust regulatory modules were identified

Next, we evaluated if the organization of TFs in our hiTRN was consistent with the functional organization of regulons as derived from the transcriptome. We thus developed a computational pipeline to identify clusters of TFs (or modules) that were significantly and strongly co-enriched across conditions. We identified ten co-enriched modules (Fig. 5.4). Six modules in particular represented core biological functions. For example, module 6 includes TFs and toxin-antitoxin pairs associated with multiple stress responses. Since the *E. coli* TRN is still

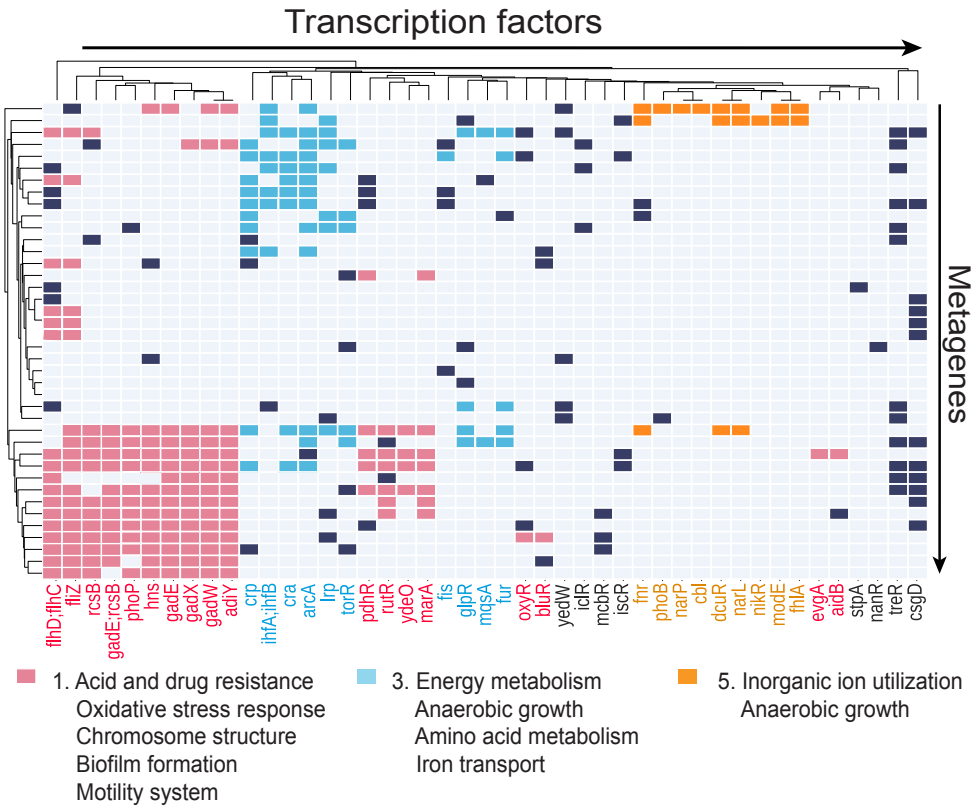


Figure 5.3: Regulon enrichment and functions of metagenes. Colors indicate functions of metagenes based on enriched regulons. Functionally-related regulons are enriched in the same metagenes. Note only TFs in modules 1, 3, 5 are shown here. A full heatmap can be found in the supplement of the original paper)

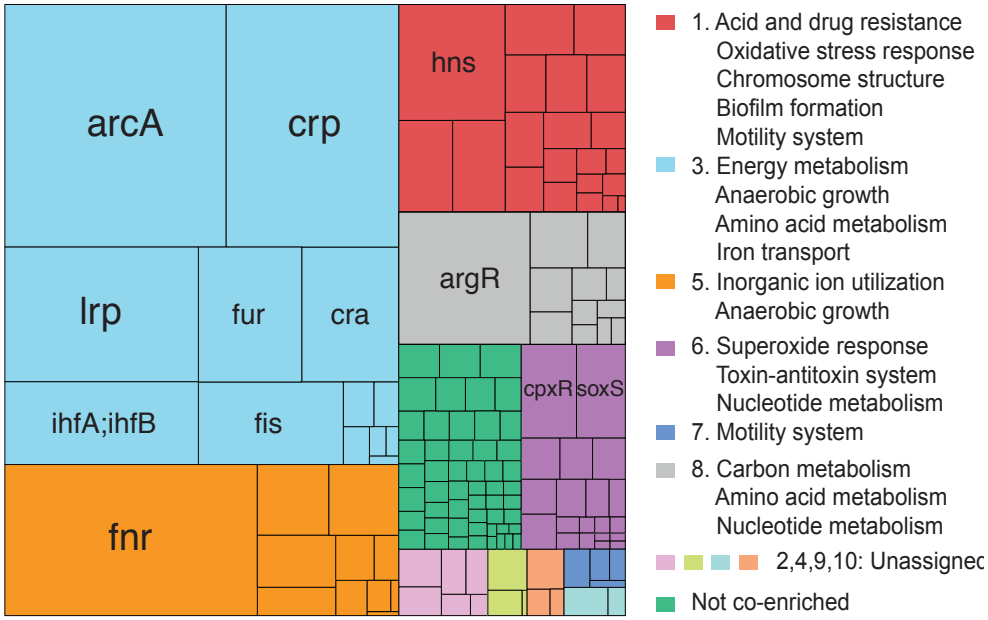


Figure 5.4: Ten functional regulatory modules for 147 TFs. Size of rectangles are proportional to the size of regulons (i.e., the number of regulated genes). The overlap between regulons is not shown in this figure. Modules fully defined in Dataset S2 in original paper

expanding, we evaluated whether these modules would remain stable as new regulatory interactions are incorporated into the TRN. We randomly added up to 60 regulons and used our pipeline to identify new modules, which were compared against the original modules. The average Jaccard index between modules ranged from 0.34 to 0.81, and normalized variation of information (VI) [203] ranged from 0.025 to 0.34. Module stability depended on which regulons were added, since even when 34 new TFs and 1,290 regulatory interactions were added, the clusters could be stable (Jaccard index=0.55, normalized VI=0.14). The modules were also relatively consistent when we used a different expression compendium, COLOMBOS [196]. The normalized VI for the modules identified between the two compendia was 0.17, which was significant (permutation test, $P < 10^{-4}$).

5.3.4 Regulatory modules identified have broad implications

Next, we examined evolutionary characteristics of the regulatory modules at the DNA sequence and protein structure levels.

Evolutionary conservation: We hypothesized that modules associated with vital functions would be conserved across species, while organism-specific responses would not. We thus computed conservation of the 147 TFs in our hiTRN across Enterobacteriaceae, γ-protebacteria, β-proteobacteria, α-proteobacteria, and δ-proteobacteria. We found two conserved modules (7 & 10), involved with motility, metal ion uptake, and DNA damage response. We also found one significantly less-conserved module (1), primarily involved with various stress responses, including acid stress. This result was consistent with availability of alternative pH stress response systems or alternative regulators of conserved proton consumption or generation genes [202].

Binding motifs: We investigated the sequence homology of DNA binding motifs of the 147 TFs. We found that the TFs in modules 1, 2, 3, 8, and 9 shared more similar binding motifs within the module compared to those in other modules (Mann-Whitney-Wilcoxon test $P < 0.05$).

Protein structure: We explored the structural similarity between TFs in each module to identify if a structure-function relationship existed within the modules. We aligned annotated DNA-binding domains and the full structures of TF pairs using the FATCAT structural alignment tool [204]. The TFs in modules 4, 5, and 8 showed significantly higher structural similarity in both cases to TFs within the module as compared to TFs in other modules. Specifically, module 8 was enriched for the periplasmic binding protein-like I domain [205].

5.3.5 The expression of most TUs can be predicted quantitatively from TF expression

We next asked whether gene expression could be quantitatively predicted as a function of TF expression levels across varying conditions and genetic perturbations. Eight potential model structures were explored to identify the best modeling procedure: four multiple linear regression models, and four support vector regressors (SVRs). The model structures are similar to those described in previous literature [194, 206–209].

Multivariate Linear Regression: We used multivariate linear regression to predict the average expression of genes grouped by TUs from RegulonDB [187] (see Methods). 1,364 of the 1,538 identified TUs were measured in EcoMAC. We tested both TFs and sigma factors as regressors [210], and included a cooperativity term that allowed for bilinear interactions between TFs for each case (Fig. 5.5A). Using an F-test of overall significance [211], we determined that in 77% (1045/1364) of TU-specific bilinear models, TFs significantly improved the fit of the models compared to intercept-only models under a false discovery rate (FDR) < 0.05. However, sigma factors alone significantly improved the fit of 91% (999/1093) of TUs with known sigma factors, highlighting the strong influence of sigma factors on TU expression.

Non-linear Interactions: To better account for nonlinear regulatory interactions, we next trained support vector regressors (SVRs) with linear and Gaussian kernels, both with and without sigma factors. Not only did the Gaussian kernel SVR with sigma factors fit the training data better than the best performing linear model ($P < .001$), the SVR significantly improved the predictive power of the model when applied to the testing data ($P < .001$) (Fig. 5.5A). The coefficient of determination (R^2) of the SVR with sigma factors on the training data was correlated with the number of known TFs (Pearson $R = 0.41$, $P < .001$) (Fig. 5.5B). Thus, as we

discover more about the structure of the TRN, the predictive power of the TRN should increase. The regression was performed on COLOMBOS using the hiTRN, and on both COLOMBOS and EcoMAC using only strong interactions from RegulonDB, providing highly similar results. Using this new model, we also tested whether our predictions captured condition-specific effects by shuffling the TU expression profiles (predicted outputs) while maintaining the order of the TF expression profiles (features). 86% (1174/1364) of TUs yielded significant differences between the shuffled expression profile regression and the original regression for the SVR (FDR-adjusted $P < 0.05$)

5.3.6 Sensitivity of TU expression to TRN topology

We next evaluated whether certain TUs were predicted more accurately using our hiTRN than random TRNs. We identified 193/1364 TUs (14%) that were predicted significantly better than random TRNs for the best SVR (FDR-adjusted $P < 0.05$). The random TRNs preserved the hiTRN's distribution of the number of TFs regulating a TU. Also, TFs sharing high mutual information with known regulators of a TU were excluded.

Mutual Information Analysis: We next investigated why predicting the expression of 86% of TUs was apparently insensitive to the exact TRN. Based on mutual information (MI), we found that only 28% (39/137) of measured TFs shared significantly higher MI with genes inside as compared to genes outside their regulons (FDR < 0.05). However, the average MI between genes within each regulatory module was significantly higher than the average MI between genes that did not share a regulatory module. Furthermore, expression profiles of many TFs shared high MI with other TFs.

Therefore, the expression of most TUs could be predicted from the expression of many

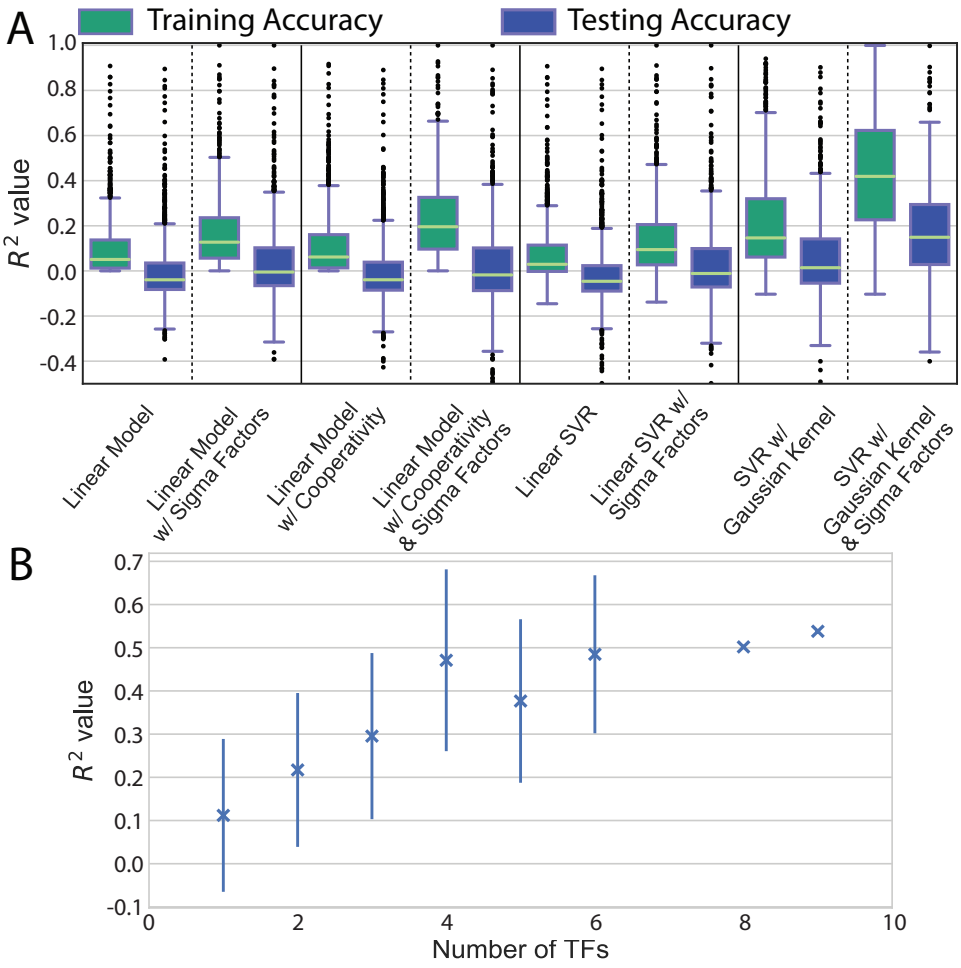


Figure 5.5: Accuracy of expression predictions on training and held-out testing transcription units. (A) R^2 (coefficient of determination) of predicted expression profile vs. true expression profile using various regression models. (B) R^2 value of the testing dataset predicted by a gaussian kernel SVR, grouped by number of known TFs. Error bars indicate standard deviation for groups with >3 observations.

alternate TFs, even if there was no evidence of TF binding. This result reflects known issues with non-identifiability of the TRN from expression profiles [212]. This issue is also encountered in other organisms [206] and reinforces the need for high-confidence regulatory interactions.

5.3.7 Other cellular processes influence gene expression

Unmeasured factors can affect gene expression, from batch effects pertaining to the experimental procedure, to unmodeled cellular processes. Such systemic variables are not represented in the hiTRN and may pose difficulties for reconciling observed differential expression. Surrogate variable analysis (SVA)[213] determines the effect of such unmodeled variables directly from expression data. We applied SVA to EcoMAC, using the best model (i.e., SVR with sigma factors) to predict the primary expression signatures. We identified three surrogate variables, which were enriched in data from a single laboratory. These variables imply that unaccounted variation stemmed from this data source, potentially related to the use of a substrate not present in other sources (LB + glycerol). Presence of three surrogate variables was consistent with clustering of data in the first two principal components in PCA analysis.

5.4 Discussion

In this study, we answered three questions concerning the scope and gaps in our knowledge of the *E. coli* TRN and our ability to predict gene expression.

5.4.1 (i) The hiTRN explains many causal connections between differential gene expression and TF activation

We found that across 21 different TF knockouts, up to 63% (26% on average) of the resulting differentially expressed genes in the TRN were traced back to the knocked out TF through regulatory cascades in our hiTRN. Compared to 15% coverage reported in an earlier assessment in 2004 [182], the new result represents an increase in coverage ranging from 70% to 320%. Additionally, we found that when accounting for network topology and regulatory bias (i.e., inhibition or activation) [197], our hiTRN explained 51–99% of differentially and non-differentially expressed genes. Some of the unexplained differential expression was potentially related to unmodeled variables that systematically influenced gene expression. We identified three such surrogate variables, one of which corresponded to a single data source having a distinct media condition. Unmodeled variables may also be explained by systematic variation in other biological processes including growth rate [214].

5.4.2 (ii) We can predict expression for 86% of TUs but only 14% of TUs are unambiguously linked to their direct regulators

We could predict expression for 86% (1174/1364) of TUs significantly better than shuffled expression profiles (FDR-adjusted $P < 0.05$). We found 193 TUs (14%) whose expression was predicted significantly better than random TRNs (FDR-adjusted $P < 0.05$), indicating the critical importance of having a well-defined TRN for these TUs. This progress resulted from having high-resolution measurements for TF binding sites and knowing the occupancy of these sites in a context-specific manner. Thus, designing experiments to define high-confidence regulatory interactions should be prioritized when characterizing the TRN of an organism for which data

are scarce.

With the advances in ChIP-exo and transcriptomics methods, a much more comprehensive understanding of the TRN can be achieved in the near term. Furthermore, given recent progress in understanding dormant TF-DNA binding events [215, 216], it will be important to investigate diverse conditions for expanding the repertoire of high-confidence interactions, which involves observing a proximal effect of binding on gene expression.

5.4.3 (iii) We robustly understand global TRN function within a limited scope

We identified ten regulatory modules representing core biological functions from two expression compendia using the hiTRN. These modules showed evolutionary conservation at the DNA sequence and protein structure levels, and overlapped with previously identified clusters [217]. Furthermore, the modules were consistent when the hiTRN was perturbed by adding random regulatory interactions from up to 60 regulons.

Taken together, these results indicate that core TRN functions are understood robustly and gene expression can be predicted. To grow the scope of the hiTRN, new high precision ChIP experiments *en masse* directed at unconfirmed TFs or TFs regulating uncharacterized genes are expected to greatly enhance the scope of understanding of *E. coli*'s TRN, and can do so in the near term. Analyzing disparate data using *in silico* models is likely to be important for guiding us through the selection and execution of the most informative experiments to fill gaps in our understanding, and to design experiments to test its robustness.

5.5 Methods

5.5.1 High-confidence regulatory network reconstruction

To reconstruct the high-confidence transcriptional regulatory network (hiTRN), we combined strong evidence interactions from RegulonDB 9.4[187] according to the RegulonDB Evidence Classification [218], with TF KO-validated ChIP-based interactions for 15 regulons from literature: *arcA* and *fnr* [190, 219, 220], *argR* [221, 222], *trpR, lrp* [222], *fur* [193], *gadEWX* [202], *oxyR, soxRS* [223], *purR* [191], *crp* [224] and *cra* [214]. The regulatory direction (+ or −) was preserved from the original study. Both directions were added if the direction was uncertain.

5.5.2 Expression compendium preparation

Experimental conditions from EcoMAC[194] were filtered to exclude non-relevant conditions as in Yang et al. [225] resulting in expression profiles for 4189 genes × 444 samples. Three of these samples (wild-type *E. coli* MG1655 grown aerobically in M9 media with glucose) were used as a reference.

5.5.3 Non-negative Matrix Factorization (NMF)

We performed NMF using sklearn with 'nnsvd' initialization [66]. The top genes accounting for 15% of each metagene's weight were used for regulon enrichment. We compared NMF with singular value decomposition to support our choice of 40 metagenes [200]. We also used non-smooth NMF (nsNMF) to identify sparse metagenes [226] and removed genes from each metagene having coefficients below 0.001 (attributable to numerical error). Since NMF solves a nonconvex optimization problem and requires multiple runs to ensure global optimality, we used two methods, by Kim and Tidor [200], and Wu et al. [201], to confirm that our NMF

decomposition was stable.

5.5.4 Regulatory module identification

We compiled a network of TFs that were co-enriched in a metagene, from 100 runs each of NMF and nsNMF [226]. We kept only 522 TF pairs that were strongly co-enriched (Jaccard index>0.18) and significant (permutation test, $P < 0.05$, from 100,000 random networks sampled from the observed frequency of co-enriched TFs). We then identified modules using multi-level modularity optimization [227]. The modularity coefficient of 0.483 was above the recommended cutoff of 0.3 to indicate community structure by Clauset et al. [228]. The functional labels of the modules were assigned using DAVID [229] functional annotation followed by manual curation.

5.5.5 Differentially-expressed gene (DEG) Identification

DEGs were identified using the R package limma in Bioconductor [230], with thresholds of $|\log_2(\text{Fold change}| > 1$ and FDR-adjusted $P < 0.05$. Three samples were used as the reference: wild type MG1655 grown in M9 with glucose as carbon source under aerobic conditions. The resulting 441 samples of expression profiles relative to the reference corresponded to 174 experimental conditions. Of these conditions, 166 showed significant differential expression. In 162 of these 166 conditions, at least one regulon was enriched for DEGs.

5.5.6 Network-expression consistency analysis

We determined the consistency of DEGs with the hiTRN for 21 TF knockout experiments. Network reachability was performed using igraph in Python [227], and sign consistency using SigNetTrainer in Matlab [197].

5.5.7 Expression profile regression

We used supervised machine learning (multiple linear regression and support vector regression) to predict log-fold change in expression of 1,364 TUs having at least 1 known regulator. We compared eight model structures with features including known regulators of each TU, cooperation/competition terms for all pairs of TFs, and known sigma factors. Models were evaluated using a stratified 10-fold cross validation to reduce overfitting (See Methods). We determined whether our models captured condition-specific effects by comparing them against models trained on 1,000 randomly shuffled TU profiles, while maintaining the order of regulator expression profiles. We further determined the significance of the TRN for predicting expression by comparing models trained on the known TRN against those trained on 1,000 random TRNs having random TFs assigned to each TU, preserving the distribution of regulators per TU. TFs having high MI with known TFs were not randomly assigned to the TU.

5.5.8 Information analysis

We computed mutual information (MI) between TFs and target genes using the NPEET Python package [231]. As in Faith et al.[189], we compared this MI to a background distribution of MI scores using the Wilcoxon rank-sum test ($\alpha = 0.05$).

Acknowledgments

We thank Daniel Zielinski for valuable discussions. This work was funded by the National Institute of General Medical Sciences of the National Institutes of Health (awards U01GM102098 and R01GM057089), the US Department of Energy (DE-SC0008701), the National Science Foundation Graduate Research Fellowship (DGE-1144086), and the Novo Nordisk Foundation

[NNF10CC1016517].

Chapter 5 in full is a reprint of material published in: **Fang, X. ***, Sastry, A.*, Mih, N., Kim, D., Tan, J., Yurkovich, J. T., Lloyd, C. J., Gao, Y., Yang, L., Palsson, B. O. (2017). Global transcriptional regulatory network for *Escherichia coli* robustly connects gene expression to transcription factor activities. *Proceedings of the National Academy of Sciences of the United States of America*, 114(38), 10286–10291. The dissertation author was the one of the primary authors.

Chapter 6

Conclusion

In this era where biological data are being generated at an unprecedented fast pace and low cost, it is especially important to be able to extract valuable information from the mass data being produced. As illustrated in this dissertation, next generation sequencing data, combined with systems-biology approaches, enable understanding of how genotypes map to phenotypes of interest, including human diseases. This dissertation focuses on multiple cross-discipline projects that utilizes systems biology frameworks and methodologies to understand the roles of gut microbiome and *E. coli* in IBD. We interpreted various data types including genome sequences, metagenomics and transcriptomics using different approaches including genomics analysis, genome-scale modeling, statistical analysis and machine learning methods. Through these studies we obtained a deeper understanding of the role of *E. coli* and gut microbiome in IBD, which shed light into development of intervention strategies and clinical decisions.

In the first chapter of this dissertation, we started with evaluating the entire gut microbial community of 127 IBD patients using time-series metagenomics and metabolomics, and revealed that intestinal surgeries may have long-term effect on the gut microbiome. We observed

striking differences in the gut communities of patients who have undergone intestinal surgery versus patients without intestinal surgery history. Specifically, in patients who have been treated with surgeries, we found that their gut microbiome presented significantly lowered species and metabolite diversity, lower stability, changes in bile acid levels and a dramatic expansion in one of the potential bacterial triggers - *E. coli*. Most of the above changes are generally considered "unhealthy" and potentially drive the gut microbiome more dysbiotic given our current knowledge. We were even able to use random forest classifiers to predict surgery history using taxonomy profiles, further emphasizing the differences in gut microbiome.

Surgery is a commonly used method to manage IBD, yet this is also irreversible and may affect the patients' life significantly. So it is important to understand its long-term consequences, including its impact on the patients' health and their gut microbiome, which is partly revealed by this study. However, limitations exist for this study as this is an observational study, including the short timeframe of this study (2 years), and the heterogeneity of the patient population. Better designed futures studies are needed to address these important questions. In this study, both metagenomics and metabolomics data are collected for analysis. While we generated valuable results from each type of data, we found it challenging to integrate these two data types together effectively, possibly due to too many confounding factors, but this also calls for development of methods to interpret multi-omics data and more carefully designed projects in the future. This study also poses a potential question for other gut microbiome studies related to human disease - how much of the changes in gut microbiome we observe are due to the disease, and how much are contributed by the disease treatment? Our project motivates such complicating factors to be carefully evaluated and carefully considered during study design and result interpretation.

Starting from the second chapter of this dissertation, we zoomed into the one the members

of the IBD gut microbiome - *E. coli*. Not only was it found to be a species that is differentially abundant in patients with/without surgery history as we discussed in chapter 1, it has long been considered a potential bacterial trigger in IBD and has been extensively studied. In the project described in chapter 2, we studied the *E. coli* community and its correlation with disease progression using time-series metagenomics data collected from a Crohn's disease patient. We identified seven *E. coli* strains that dominated at different time points, highlighting that the gut microbiome is usually highly dynamic. Most interestly, we found these strains to have drastically different gene content - including virulence capability and metabolic functions. Specifically, the dominant strains isolated from peak inflammation is the most similar to the strains in known AIEC pathotype that is implicated in IBD, while strains isolated during stable time are more similar to other commensal strains. These results suggest that strains-specific features of *E. coli* strains in IBD may be potentially related to disease progression, and motivates similar studies on a larger cohort. If certain features can be consistently identified across patients, they can be followed up with experiments for verification, and will likely provide much needed mechanistic insight for IBD and its intervention strategy.

This study was performed as a pilot study to showcase the importance of strain level analysis. When the project was getting started, the strain-level analysis of gut microbiome using metagenomics data was gaining its popularity. With this project, we showed that great potential exists for strain-level analysis for gut microbiomes, especially for species with open genomes, such as *E. coli*. Of course such high resolution analysis requires sequencing data with great depth - but with the development of the sequencing technology and drop in cost, we are optimistic that depth of sequencing data will only increase in the future, which will further facilitate such strain-level analysis.

Following the strain-level analysis of *E. coli* in chapter 2, chapter 3 focused on the characterization of metabolic capabilities of *E. coli* strains using available genome sequences of clinical isolates from IBD patients. Through pan-genome analysis and genome-scale modeling, we found that strains in B2 phylogroup that are known to be more prevalent in IBD patients have genes that are specific to them that enable utilization of mucus glycan as growth substrate, therefore potentially give them advantage in colonization of the digestive tract. Growth simulation using metabolic models also confirmed they have different growth capability that differentiates them from strains from other phylogroups.

This study again emphasizes the potential of strain- level importance, as well as the value of genome-scale modeling as it allows direct mapping of genotype to phenotype and mechanistic understanding. Our study has shown that metabolic capabilities of *E. coli* may just be as important as its virulence factors, as metabolic functions allow them to adjust to thrive in different environmental niches. We were able to identify the genetic basis of the predicted difference in metabolism, which enables straightforward experimental design to verify the hypothesis when needed - genes of interest can be knocked out and we can evaluate the changes in growth capabilities on certain growth substrates. A step forward would be to place strains with different metabolic capabilities in animal models (either healthy or IBD-induced) and observe their performances.

Lastly, we evaluated the transcriptional regulatory network (TRN) in *E. coli*, as TRN is responsible for regulating gene expression level in response to different environments, including the inflamed human digestive tract in *E. coli*. Analyzing a larging transcriptomics dataset against the most up-to-date TRN revealed that TRN in *E. coli* has expanded significantly in the past 15 year due to the emergence of ChiP-Seq technology, but the coverage of TRN is still limited.

We observe that only a portion of the differentially expressed genes across conditions can be explained by the known TRN and prediction of gene expression level using TRN is far from ideal. It is a bit surprising given that *E. coli* has the most well-studied TRN but knowledge gaps still exist, which calls for more high-throughput experimental in the future.

In all projects presented here, our efforts to study *E. coli*, gut microbiome and their roles in IBD using a combination of next generation sequencing data and systems biology methods have turned out to be very fruitful, as we produced multiple testable hypotheses, generated valuable insight and constructed useful workflows. From this work, we predict that analysis in two particular directions will be especially fruitful. First direction is the strain-level analysis either using metagenomics data or genome sequences of isolates from the gut, as many species have significant variation across strains that would not be visible through species level analysis. And the bloom in sequencing data will only make such analysis easier and more accessible to more researchers. Additionally, we believe that studies that intersect both systems biology and gut microbiome will be promising , as more and more systems biology approaches have be developed in recent years to study the gut microbiome, include the construction of genome-scale models for common gut microbes, development of community models, and personalized gut microbial models using metagenomics data. These analyses differ from traditional gut microbiome analysis that mainly focuses on statistical analysis and correlations, but aim to identify the underlying mechanisms of phenotypes of interest, which is becoming increasingly important in the field of gut microbiome.

Bibliography

1. Cosnes, J., Gower-Rousseau, C., Seksik, P. & Cortot, A. Epidemiology and natural history of inflammatory bowel diseases. en. *Gastroenterology* **140,** 1785–1794 (May 2011).

2. Ananthakrishnan, A. N. Epidemiology and risk factors for IBD. en. *Nat. Rev. Gastroenterol. Hepatol.* **12,** 205–217 (Apr. 2015).

3. Seyedian, S. S., Nokhostin, F. & Malamir, M. D. A review of the diagnosis, prevention, and treatment methods of inflammatory bowel disease. en. *J. Med. Life* **12,** 113–122 (Apr. 2019).

4. Fakhoury, M., Negrulj, R., Mooranian, A. & Al-Salami, H. Inflammatory bowel disease: clinical aspects and treatments. en. *J. Inflamm. Res.* **7,** 113–120 (June 2014).

5. Momozawa, Y., Dmitrieva, J., Théâtre, E., Deffontaine, V., Rahmouni, S., Charloteaux, B., Crins, F., Docampo, E., Elansary, M., Gori, A.-S., Lecut, C., Mariman, R., Mni, M., Oury, C., Altukhov, I., Alexeev, D., Aulchenko, Y., Amininejad, L., Bouma, G., Hoentjen, F., Löwenberg, M., Oldenburg, B., Pierik, M. J., Vander Meulen-de Jong, A. E., Janneke van der Woude, C., Visschedijk, M. C., International IBD Genetics Consortium, Lathrop, M., Hugot, J.-P., Weersma, R. K., De Vos, M., Franchimont, D., Vermeire, S., Kubo, M., Louis, E. & Georges, M. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. en. *Nat. Commun.* **9,** 2427 (June 2018).

6. Franzosa, E. A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H. J., Reinker, S., Vatanen, T., Hall, A. B., Mallick, H., McIver, L. J., Sauk, J. S., Wilson, R. G., Stevens, B. W., Scott, J. M., Pierce, K., Deik, A. A., Bullock, K., Imhann, F., Porter, J. A., Zhernakova, A., Fu, J., Weersma, R. K., Wijmenga, C., Clish, C. B., Vlamakis, H., Huttenhower, C. & Xavier, R. J. Gut microbiome structure and metabolic activity in inflammatory bowel disease. en. *Nat Microbiol* **4,** 293–305 (Feb. 2019).

7. Martinez-Medina, M. & Garcia-Gil, L. J. Escherichia coli in chronic inflammatory bowel diseases: An update on adherent invasive Escherichia coli pathogenicity. en. *World J. Gastrointest. Pathophysiol.* **5,** 213–227 (Aug. 2014).

8. Bharti, R. & Grimm, D. G. Current challenges and best-practice protocols for microbiome analysis. en. *Brief. Bioinform.* (Dec. 2019).

9. Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L.-I., McDonald, D., Melnik, A. V., Morton, J. T., Navas, J.,

Quinn, R. A., Sanders, J. G., Swafford, A. D., Thompson, L. R., Tripathi, A., Xu, Z. Z., Zaneveld, J. R., Zhu, Q., Caporaso, J. G. & Dorrestein, P. C. Best practices for analysing microbiomes. en. *Nat. Rev. Microbiol.* **16,** 410–422 (July 2018).

10. Kuczynski, J., Lauber, C. L., Walters, W. A., Parfrey, L. W., Clemente, J. C., Gevers, D. & Knight, R. Experimental and analytical tools for studying the human microbiome. en. *Nat. Rev. Genet.* **13,** 47–58 (Dec. 2011).

11. Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K. S. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* **26,** 1612–1625 (Nov. 2016).

12. Fang, X., Monk, J. M., Mih, N., Du, B., Sastry, A. V., Kavvas, E., Seif, Y., Smarr, L. & Palsson, B. O. Escherichia coli B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of intestinal mucosa. en. *BMC Syst. Biol.* **12,** 66 (June 2018).

13. Fang, X., Monk, J. M., Nurk, S., Akseshina, M., Zhu, Q., Gemmell, C., Gianetto-Hill, C., Leung, N., Szubin, R., Sanders, J., Beck, P. L., Li, W., Sandborn, W. J., Gray-Owen, S. D., Knight, R., Allen-Vercoe, E., Palsson, B. O. & Smarr, L. Metagenomics-Based, Strain-Level Analysis of Escherichia coli From a Time-Series of Microbiome Samples From a Crohn's Disease Patient. en. *Front. Microbiol.* **9,** 2559 (Oct. 2018).

14. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. en. *Nat. Protoc.* **5,** 93–121 (Jan. 2010).

15. Norsigian, C. J., Fang, X., Seif, Y., Monk, J. M. & Palsson, B. O. A workflow for generating multi-strain genome-scale metabolic models of prokaryotes. en. *Nat. Protoc.* (Dec. 2019).

16. Price, N. D., Reed, J. L. & Palsson, B. Ø. Genome-scale models of microbial cells: evaluating the consequences of constraints. en. *Nat. Rev. Microbiol.* **2,** 886–897 (Nov. 2004).

17. Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* (2012).

18. Feist, A. M. & Palsson, B. Ø. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. en. *Nat. Biotechnol.* **26,** 659–667 (June 2008).

19. Bordbar, A., Monk, J. M., King, Z. A. & Palsson, B. O. Constraint-based models predict metabolic and associated cellular functions. en. *Nat. Rev. Genet.* **15,** 107–120 (Feb. 2014).

20. Guan, N., Du, B., Li, J., Shin, H.-D., Chen, R. R., Du, G., Chen, J. & Liu, L. Comparative genomics and transcriptomics analysis-guided metabolic engineering of Propionibacterium acidipropionici for improved propionic acid production. *Biotechnol. Bioeng.* **115,** 483–494 (2018).

21. Folger, O., Jerby, L., Frezza, C., Gottlieb, E., Ruppin, E. & Shlomi, T. Predicting selective drug targets in cancer through metabolic networks. en. *Mol. Syst. Biol.* **7,** 501 (June 2011).

22. Guzmán, G. I., Utrilla, J., Nurk, S., Brunk, E., Monk, J. M., Ebrahim, A., Palsson, B. O. & Feist, A. M. Model-driven discovery of underground metabolic functions in Escherichia coli. en. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 929–934 (Jan. 2015).

23. Kumar, M., Ji, B., Zengler, K. & Nielsen, J. Modelling approaches for studying the microbiome. en. *Nat Microbiol* **4,** 1253–1267 (Aug. 2019).

24. Mardinoglu, A., Agren, R., Kampf, C., Asplund, A., Uhlen, M. & Nielsen, J. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. en. *Nat. Commun.* **5,** 3083 (2014).

25. Mardinoglu, A. & Nielsen, J. Systems medicine and metabolic modelling. en. *J. Intern. Med.* **271,** 142–154 (Feb. 2012).

26. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? en. *Nat. Biotechnol.* **28,** 245–248 (Mar. 2010).

27. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Systems Biology* **7,** 74. ISSN: 1752-0509. `https://doi.org/10.1186/1752-0509-7-74` (Aug. 2013).

28. Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., Heinken, A., Haraldsdóttir, H. S., Wachowiak, J., Keating, S. M., Vlasov, V., Magnusdóttir, S., Ng, C. Y., Preciat, G., Žagare, A., Chan, S. H. J., Aurich, M. K., Clancy, C. M., Modamio, J., Sauls, J. T., Noronha, A., Bordbar, A., Cousins, B., El Assal, D. C., Valcarcel, L. V., Apaolaza, I., Ghaderi, S., Ahookhosh, M., Ben Guebila, M., Kostromins, A., Sompairac, N., Le, H. M., Ma, D., Sun, Y., Wang, L., Yurkovich, J. T., Oliveira, M. A. P., Vuong, P. T., El Assal, L. P., Kuperstein, I., Zinovyev, A., Hinton, H. S., Bryant, W. A., Aragón Artacho, F. J., Planes, F. J., Stalidzans, E., Maass, A., Vempala, S., Hucka, M., Saunders, M. A., Maranas, C. D., Lewis, N. E., Sauter, T., Palsson, B. Ø., Thiele, I. & Fleming, R. M. T. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. en. *Nat. Protoc.* **14,** 639–702 (Mar. 2019).

29. Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., Greenhalgh, K., Jäger, C., Baginska, J., Wilmes, P., Fleming, R. M. T. & Thiele, I. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. en. *Nat. Biotechnol.* **35,** 81–89 (Jan. 2017).

30. Magnúsdóttir, S. & Thiele, I. Modeling metabolism of the human gut microbiome. en. *Curr. Opin. Biotechnol.* **51,** 90–96 (June 2018).

31. Frank, D. N., St Amand, A. L., Feldman, R. A., Boedeker, E. C., Harpaz, N. & Pace, N. R. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. en. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 13780–13785 (Aug. 2007).

32. 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 (Apr. 2012).

33. Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., Schwager, E., Knights, D., Song, S. J., Yassour, M., Morgan, X. C., Kostic, A. D., Luo, C., González, A., McDonald, D., Haberman, Y., Walters, T., Baker, S., Rosh, J., Stephens, M., Heyman, M., Markowitz, J., Baldassano, R., Griffiths, A., Sylvester, F., Mack, D., Kim, S., Crandall, W., Hyams, J., Huttenhower, C., Knight, R. & Xavier, R. J. The treatment-naive microbiome in new-onset Crohn's disease. en. *Cell Host Microbe* **15,** 382–392 (Mar. 2014).

34. Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., Andrews, E., Ajami, N. J., Bonham, K. S., Brislawn, C. J., Casero, D., Courtney, H., Gonzalez, A., Graeber, T. G., Hall, A. B., Lake, K., Landers, C. J., Mallick, H., Plichta, D. R., Prasad, M., Rahnavard, G., Sauk, J., Shungin, D., Vázquez-Baeza, Y., White 3rd, R. A., IBDMDB Investigators, Braun, J., Denson, L. A., Jansson, J. K., Knight, R., Kugathasan, S., McGovern, D. P. B., Petrosino, J. F., Stappenbeck, T. S., Winter, H. S., Clish, C. B., Franzosa, E. A., Vlamakis, H., Xavier, R. J. & Huttenhower, C. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. en. *Nature* **569,** 655–662 (May 2019).

35. Pascal, V., Pozuelo, M., Borruel, N., Casellas, F., Campos, D., Santiago, A., Martinez, X., Varela, E., Sarrabayrouse, G., Machiels, K., Vermeire, S., Sokol, H., Guarner, F. & Manichanh, C. A microbial signature for Crohn's disease. *Gut* **66,** 813–822 (May 2017).

36. Jansson, J., Willing, B., Lucio, M., Fekete, A., Dicksved, J., Halfvarson, J., Tysk, C. & Schmitt-Kopplin, P. Metabolomics reveals metabolic biomarkers of Crohn's disease. en. *PLoS One* **4,** e6386 (July 2009).

37. Halfvarson, J., Brislawn, C. J., Lamendella, R., Vázquez-Baeza, Y., Walters, W. A., Bramer, L. M., D'Amato, M., Bonfiglio, F., McDonald, D., Gonzalez, A., McClure, E. E., Dunklebarger, M. F., Knight, R. & Jansson, J. K. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol* **2,** 17004 (Feb. 2017).

38. Vázquez-Baeza, Y., Gonzalez, A., Xu, Z. Z., Washburne, A., Herfarth, H. H., Sartor, R. B. & Knight, R. Guiding longitudinal sampling in IBD cohorts. en. *Gut* **67,** 1743–1745 (Sept. 2018).

39. Mondot, S., Lepage, P., Seksik, P., Allez, M., Tréton, X., Bouhnik, Y., Colombel, J. F., Leclerc, M., Pochart, P., Doré, J., Marteau, P. & GETAID. Structural robustness of the gut mucosal microbiota is associated with Crohn's disease remission after surgery. en. *Gut* **65,** 954–962 (June 2016).

40. Baxter, N. T., Lesniak, N. A., Sinani, H., Schloss, P. D. & Koropatkin, N. M. The Glucoamylase Inhibitor Acarbose Has a Diet-Dependent and Reversible Effect on the Murine Gut Microbiome. en. *mSphere* **4** (Feb. 2019).

41. Quince, C., Delmont, T. O., Raguideau, S., Alneberg, J., Darling, A. E., Collins, G. & Eren, A. M. DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol.* **18,** 181 (Sept. 2017).

42. Ponsioen, C. Y., de Groof, E. J., Eshuis, E. J., Gardenbroek, T. J., Bossuyt, P. M. M., Hart, A., Warusavitarne, J., Buskens, C. J., van Bodegraven, A. A., Brink, M. A., Consten, E. C. J., van Wagensveld, B. A., Rijk, M. C. M., Crolla, R. M. P. H., Noomen, C. G., Houdijk, A. P. J., Mallant, R. C., Boom, M., Marsman, W. A., Stockmann, H. B., Mol, B., de Groof, A. J., Stokkers, P. C., D'Haens, G. R., Bemelman, W. A. & LIR!C study group. Laparoscopic ileocaecal resection versus infliximab for terminal ileitis in Crohn's disease: a randomised controlled, open-label, multicentre trial. en. *Lancet Gastroenterol Hepatol* **2,** 785–792 (Nov. 2017).

43. Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R. & Peddada, S. D. Analysis of composition of microbiomes: a novel method for studying microbial composition. en. *Microb. Ecol. Health Dis.* **26,** 27663 (May 2015).

44. Hughes, E. R., Winter, M. G., Duerkop, B. A., Spiga, L., Furtado de Carvalho, T., Zhu, W., Gillis, C. C., Büttner, L., Smoot, M. P., Behrendt, C. L., Cherry, S., Santos, R. L., Hooper, L. V. & Winter, S. E. Microbial Respiration and Formate Oxidation as Metabolic Signatures of Inflammation-Associated Dysbiosis. en. *Cell Host Microbe* **21,** 208–219 (Feb. 2017).

45. Morton, J. T., Aksenov, A. A., Nothias, L. F., Foulds, J. R., Quinn, R. A., Badri, M. H., Swenson, T. L., Van Goethem, M. W., Northen, T. R., Vazquez-Baeza, Y., Wang, M., Bokulich, N. A., Watters, A., Song, S. J., Bonneau, R., Dorrestein, P. C. & Knight, R. Learning representations of microbe–metabolite interactions. *Nat. Methods* **16,** 1306–1314 (Dec. 2019).

46. Hakala, K., Vuoristo, M., Luukkonen, P., Järvinen, H. J. & Miettinen, T. A. Impaired absorption of cholesterol and bile acids in patients with an ileoanal anastomosis. en. *Gut* **41,** 771–777 (Dec. 1997).

47. Miettinen, T. A. The role of bile salts in diarrhoea of patients with ulcerative colitis. en. *Gut* **12,** 632–635 (Aug. 1971).

48. Palmela, C., Chevarin, C., Xu, Z., Torres, J., Sevrin, G., Hirten, R., Barnich, N., Ng, S. C. & Colombel, J.-F. Adherent-invasive Escherichia coli in inflammatory bowel disease. en. *Gut* **67,** 574–587 (Mar. 2018).

49. Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project. en. *Nature* **569,** 641–648 (May 2019).

50. Rutgeerts, P., Goboes, K., Peeters, M., Hiele, M., Penninckx, F., Aerts, R., Kerremans, R. & Vantrappen, G. Effect of faecal stream diversion on recurrence of Crohn's disease in the neoterminal ileum. en. *Lancet* **338,** 771–774 (Sept. 1991).

51. Blaser, M. J. Antibiotic use and its consequences for the normal microbiome. en. *Science* **352,** 544–545 (Apr. 2016).

52. Satsangi, J., Silverberg, M. S., Vermeire, S. & Colombel, J.-F. The Montreal classification of inflammatory bowel disease: controversies, consensus, and implications. en. *Gut* **55,** 749–753 (June 2006).

53. Marotz, C., Amir, A., Humphrey, G., Gaffney, J., Gogul, G. & Knight, R. DNA extraction for streamlined metagenomics of diverse environmental samples. en. *Biotechniques* **62,** 290–293 (June 2017).

54. Didion, J. P. & Collins, F. S. *Atropos: specific, sensitive, and speedy trimming of sequencing reads*

55. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (Mar. 2012).

56. Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C. & Segata, N. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12,** 902–903 (Oct. 2015).

57. Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., Knight, R., Caporaso, J. G., Segata, N. & Huttenhower, C. Species-level functional profiling of metagenomes and metatranscriptomes. en. *Nat. Methods* **15,** 962–968 (Nov. 2018).

58. Scholz, M., Ward, D. V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., Truong, D. T., Tett, A., Morrow, A. L. & Segata, N. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. en. *Nat. Methods* **13,** 435–438 (May 2016).

59. *mca* `https://pypi.org/project/mca/`. Accessed: 2018-5-4. 2018.

60. Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. en. *BMC Bioinformatics* **11,** 395 (July 2010).

61. Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapono, C. A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A. V., Meehan, M. J., Liu, W.-T., Crüsemann, M., Boudreau, P. D., Esquenazi, E., Sandoval-Calderón, M., Kersten, R. D., Pace, L. A., Quinn, R. A., Duncan, K. R., Hsu, C.-C., Floros, D. J., Gavilan, R. G., Kleigrewe, K., Northen, T., Dutton, R. J., Parrot, D., Carlson, E. E., Aigle, B., Michelsen, C. F., Jelsbak, L., Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B. T., Gerwick, L., Liaw, C.-C., Yang, Y.-L., Humpf, H.-U., Maansson, M., Keyzers, R. A., Sims, A. C., Johnson, A. R., Sidebottom, A. M., Sedio, B. E., Klitgaard, A., Larson, C. B., P, C. A. B., Torres-Mendoza, D., Gonzalez, D. J., Silva, D. B., Marques, L. M., Demarque, D. P., Pociute, E., O'Neill, E. C., Briand, E., Helfrich, E. J. N., Granatosky, E. A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J. J., Zeng, Y., Vorholt, J. A., Kurita, K. L., Charusanti, P., McPhail, K. L., Nielsen, K. F., Vuong, L., Elfeki, M., Traxler, M. F., Engene, N., Koyama, N., Vining, O. B., Baric, R., Silva, R. R., Mascuch, S. J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P. G., Dai, J., Neupane, R., Gurr, J., Rodrıguez,

A. M. C., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B. M., Almaliti, J., Allard, P.-M., Phapale, P., Nothias, L.-F., Alexandrov, T., Litaudon, M., Wolfender, J.-L., Kyle, J. E., Metz, T. O., Peryea, T., Nguyen, D.-T., VanLeer, D., Shinn, P., Jadhav, A., Müller, R., Waters, K. M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P. R., Palsson, B. O., Pogliano, K., Linington, R. G., Gutiérrez, M., Lopes, N. P., Gerwick, W. H., Moore, B. S., Dorrestein, P. C. & Bandeira, N. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. en. *Nat. Biotechnol.* **34,** 828–837 (Aug. 2016).

62. McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., Knight, R. & Caporaso, J. G. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. en. *Gigascience* **1,** 7 (July 2012).

63. Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodrıguez, A. M., Chase, J., Cope, E. K., Da Silva, R., Diener, C., Dorrestein, P. C., Douglas, G. M., Durall, D. M., Duvallet, C., Edwardson, C. F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J. M., Gibbons, S. M., Gibson, D. L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G. A., Janssen, S., Jarmusch, A. K., Jiang, L., Kaehler, B. D., Kang, K. B., Keefe, C. R., Keim, P., Kelley, S. T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M. G. I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B. D., McDonald, D., McIver, L. J., Melnik, A. V., Metcalf, J. L., Morgan, S. C., Morton, J. T., Naimey, A. T., Navas-Molina, J. A., Nothias, L. F., Orchanian, S. B., Pearson, T., Peoples, S. L., Petras, D., Preuss, M. L., Pruesse, E., Rasmussen, L. B., Rivers, A., Robeson 2nd, M. S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S. J., Spear, J. R., Swafford, A. D., Thompson, L. R., Torres, P. J., Trinh, P., Tripathi, A., Turnbaugh, P. J., Ul-Hasan, S., van der Hooft, J. J. J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K. C., Williamson, C. H. D., Willis, A. D., Xu, Z. Z., Zaneveld, J. R., Zhang, Y., Zhu, Q., Knight, R. & Caporaso, J. G. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. en. *Nat. Biotechnol.* **37,** 852–857 (Aug. 2019).

64. Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Ostblom, J., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., de Ruiter, J., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., Meyer, K., Miles, A., Ram, Y., Brunner, T., Yarkoni, T., Williams, M. L., Evans, C., Fitzgerald, C., Brian & Qalieh, A. *mwaskom/seaborn: v0.9.0 (July 2018)* July 2018.

65. Hunter, J. D. *Matplotlib: A 2D Graphics Environment* 2007.

66. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D.,

Brucher, M., Perrot, M. & Duchesnay, É. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).

67. Jones, E., Oliphant, T. & Peterson, P. SciPy: Open source scientific tools for Python (2001).

68. McKinney, W. *Data structures for statistical computing in python* in *Proceedings of the 9th Python in Science Conference* **445** (2010), 51–56.

69. Oliphant, T. E. *A guide to NumPy* (Trelgol Publishing USA, 2006).

70. Faith, D. P. Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* **61,** 1–10 (Jan. 1992).

71. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. en. *Appl. Environ. Microbiol.* **71,** 8228–8235 (Dec. 2005).

72. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (Oct. 2001).

73. Gonzalez, A., Navas-Molina, J. A., Kosciolek, T., McDonald, D., Vázquez-Baeza, Y., Ackermann, G., DeReus, J., Janssen, S., Swafford, A. D., Orchanian, S. B., Sanders, J. G., Shorenstein, J., Holste, H., Petrus, S., Robbins-Pianka, A., Brislawn, C. J., Wang, M., Rideout, J. R., Bolyen, E., Dillon, M., Caporaso, J. G., Dorrestein, P. C. & Knight, R. Qiita: rapid, web-enabled microbiome meta-analysis. en. *Nat. Methods* **15,** 796–798 (Oct. 2018).

74. Matsuoka, K. & Kanai, T. The gut microbiota and inflammatory bowel disease. *Semin. Immunopathol.* **37,** 47–55 (Jan. 2015).

75. Rhodes, J. M. The role of Escherichia coli in inflammatory bowel disease. *Gut* **56,** 610–612 (May 2007).

76. Sasaki, M., Sitaraman, S. V., Babbin, B. A., Gerner-Smidt, P., Ribot, E. M., Garrett, N., Alpern, J. A., Akyildiz, A., Theiss, A. L., Nusrat, A. & Klapproth, J.-M. A. Invasive Escherichia coli are a feature of Crohn's disease. *Lab. Invest.* **87,** 1042–1054 (Oct. 2007).

77. Darfeuille-Michaud, A., Neut, C., Barnich, N., Lederman, E., Di Martino, P., Desreumaux, P., Gambiez, L., Joly, B., Cortot, A. & Colombel, J. F. Presence of adherent Escherichia coli strains in ileal mucosa of patients with Crohn's disease. *Gastroenterology* **115,** 1405–1413 (Dec. 1998).

78. Palmela, C., Chevarin, C., Xu, Z., Torres, J., Sevrin, G., Hirten, R., Barnich, N., Ng, S. C. & Colombel, J.-F. Adherent-invasive Escherichia coli in inflammatory bowel disease. en. *Gut* (Nov. 2017).

79. Petersen, A. M., Nielsen, E. M., Litrup, E., Brynskov, J., Mirsepasi, H. & Krogfelt, K. A. A phylogenetic group of Escherichia coli associated with active left-sided inflammatory bowel disease. *BMC Microbiol.* **9,** 171 (Aug. 2009).

80. O'Brien, C. L., Bringer, M.-A., Holt, K. E., Gordon, D. M., Dubois, A. L., Barnich, N., Darfeuille-Michaud, A. & Pavli, P. Comparative genomics of Crohn's disease-associated adherent-invasive Escherichia coli. *Gut* (Apr. 2016).

81. Eaves-Pyles, T., Allen, C. A., Taormina, J., Swidsinski, A., Tutt, C. B., Jezek, G. E., Islas-Islas, M. & Torres, A. G. Escherichia coli isolated from a Crohn's disease patient adheres, invades, and induces inflammatory responses in polarized intestinal epithelial cells. *Int. J. Med. Microbiol.* **298,** 397–409 (July 2008).

82. Vejborg, R. M., Hancock, V., Petersen, A. M., Krogfelt, K. A. & Klemm, P. Comparative genomics of Escherichia coli isolated from patients with inflammatory bowel disease. en. *BMC Genomics* **12,** 316 (June 2011).

83. Desilets, M., Deng, X., Deng, X., Rao, C., Ensminger, A. W., Krause, D. O., Sherman, P. M. & Gray-Owen, S. D. Genome-based Definition of an Inflammatory Bowel Disease-associated Adherent-Invasive Escherichia coli Pathovar. en. *Inflamm. Bowel Dis.* **22,** 1–12 (Jan. 2016).

84. Ni, J., Wu, G. D., Albenberg, L. & Tomov, V. T. Gut microbiota and IBD: causation or correlation? *Nat. Rev. Gastroenterol. Hepatol.* **14,** 573–584 (Oct. 2017).

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

86. Caugant, D. A., Levin, B. R. & Selander, R. K. Genetic diversity and temporal variation in the E. coli population of a human host. *Genetics* **98,** 467–490 (July 1981).

87. Luo, C., Knight, R., Siljander, H., Knip, M., Xavier, R. J. & Gevers, D. ConStrains identifies microbial strains in metagenomic datasets. *Nat. Biotechnol.* **33,** 1045–1052 (Oct. 2015).

88. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27,** 626–638 (Apr. 2017).

89. Fischer, M., Strauch, B. & Renard, B. Y. Abundance estimation and differential testing on strain level in metagenomics data. *Bioinformatics* **33,** i124–i132 (July 2017).

90. McCloskey, D., Palsson, B. Ø. & Feist, A. M. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. en. *Mol. Syst. Biol.* **9,** 661 (Jan. 2013).

91. Rasko, D. A., Rosovitz, M. J., Myers, G. S. A., Mongodin, E. F., Fricke, W. F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N. R., Chaudhuri, R., Henderson, I. R., Sperandio, V. & Ravel, J. The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *J. Bacteriol.* **190,** 6881–6893 (Oct. 2008).

92. Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., Feist, A. M. & Palsson, B. Ø. Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. en. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 20338–20343 (Dec. 2013).

93. Wu, S., Li, W., Smarr, L., Nelson, K., Yooseph, S. & Torralba, M. *Large memory high performance computing enables comparison across human gut microbiome of patients with autoimmune diseases and healthy subjects* in *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery* (ACM, July 2013), 25.

94. Mosli, M. H., Zou, G., Garg, S. K., Feagan, S. G., MacDonald, J. K., Chande, N., Sandborn, W. J. & Feagan, B. G. C-Reactive Protein, Fecal Calprotectin, and Stool Lactoferrin for Detection of Endoscopic Activity in Symptomatic Inflammatory Bowel Disease Patients: A Systematic Review and Meta-Analysis. *Am. J. Gastroenterol.* **110,** 802–19, quiz 820 (June 2015).

95. Bradley, P. H. & Pollard, K. S. Proteobacteria explain significant functional variability in the human gut microbiome. *Microbiome* **5,** 36 (Mar. 2017).

96. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (June 2012).

97. Snipen, L., Almøy, T. & Ussery, D. W. Microbial comparative pan-genomics using binomial mixture models. en. *BMC Genomics* **10,** 385 (Aug. 2009).

98. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–2069 (July 2014).

99. Segata, N., Bornigen, D., Morgan, X. C. & Huttenhower, C. *PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat. Commun. 4: 2304* 2013.

100. Kotlowski, R., Bernstein, C. N., Sepehri, S. & Krause, D. O. High prevalence of Escherichia coli belonging to the B2+D phylogenetic group in inflammatory bowel disease. *Gut* **56,** 669–675 (May 2007).

101. Ruan, Z. & Feng, Y. BacWGSTdb, a database for genotyping and source tracking bacterial pathogens. *Nucleic Acids Res.* **44,** D682–7 (Jan. 2016).

102. Doumith, M., Day, M., Ciesielczuk, H., Hope, R., Underwood, A., Reynolds, R., Wain, J., Livermore, D. M. & Woodford, N. Rapid identification of major Escherichia coli sequence types causing urinary tract and bloodstream infections. *J. Clin. Microbiol.* **53,** 160–166 (Jan. 2015).

103. Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A. & Parkhill, J. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31,** 3691–3693 (Nov. 2015).

104. Allen-Vercoe, E. & Jobin, C. Fusobacterium and Enterobacteriaceae: important players for CRC? *Immunol. Lett.* **162,** 54–61 (Dec. 2014).

105. Ellermann, M., Huh, E. Y., Liu, B., Carroll, I. M., Tamayo, R. & Sartor, R. B. Adherent-Invasive Escherichia coli Production of Cellulose Influences Iron-Induced Bacterial Aggregation, Phagocytosis, and Induction of Colitis. en. *Infect. Immun.* **83,** 4068–4080 (Oct. 2015).

106. Arthur, J. C., Perez-Chanona, E., Mühlbauer, M., Tomkovich, S., Uronis, J. M., Fan, T.-J., Campbell, B. J., Abujamel, T., Dogan, B., Rogers, A. B., Rhodes, J. M., Stintzi, A., Simpson, K. W., Hansen, J. J., Keku, T. O., Fodor, A. A. & Jobin, C. Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* **338,** 120–123 (Oct. 2012).

107. Darfeuille-Michaud, A. Adherent-invasive Escherichia coli: a putative new E. coli pathotype associated with Crohn's disease. *Int. J. Med. Microbiol.* **292,** 185–193 (Sept. 2002).

108. Darfeuille-Michaud, A., Boudeau, J., Bulois, P., Neut, C., Glasser, A.-L., Barnich, N., Bringer, M.-A., Swidsinski, A., Beaugerie, L. & Colombel, J.-F. High prevalence of adherent-invasive Escherichia coli associated with ileal mucosa in Crohn's disease. *Gastroenterology* **127,** 412–421 (Aug. 2004).

109. Miquel, S., Peyretaillade, E., Claret, L., de Vallée, A., Dossat, C., Vacherie, B., Zineb, E. H., Segurens, B., Barbe, V., Sauvanet, P., Neut, C., Colombel, J.-F., Medigue, C., Mojica, F. J. M., Peyret, P., Bonnet, R. & Darfeuille-Michaud, A. Complete genome sequence of Crohn's disease-associated adherent-invasive E. coli strain LF82. en. *PLoS One* **5** (Sept. 2010).

110. Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y. & Jin, Q. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.* **33,** D325–8 (Jan. 2005).

111. Boratyn, G. M., Schäffer, A. A., Agarwala, R., Altschul, S. F., Lipman, D. J. & Madden, T. L. Domain enhanced lookup time accelerated BLAST. *Biol. Direct* **7,** 12 (Apr. 2012).

112. Dogan, B., Suzuki, H., Herlekar, D., Sartor, R. B., Campbell, B. J., Roberts, C. L., Stewart, K., Scherl, E. J., Araz, Y., Bitar, P. P., Lefébure, T., Chandler, B., Schukken, Y. H., Stanhope, M. J. & Simpson, K. W. Inflammation-associated adherent-invasive Escherichia coli are enriched in pathways for use of propanediol and iron and M-cell translocation. en. *Inflamm. Bowel Dis.* **20,** 1919–1932 (Nov. 2014).

113. Barnich, N., Carvalho, F. A., Glasser, A.-L., Darcha, C., Jantscheff, P., Allez, M., Peeters, H., Bommelaer, G., Desreumaux, P., Colombel, J.-F. & Darfeuille-Michaud, A. CEACAM6 acts as a receptor for adherent-invasive E. coli, supporting ileal mucosa colonization in Crohn disease. *J. Clin. Invest.* **117,** 1566–1574 (June 2007).

114. Nash, J. H., Villegas, A., Kropinski, A. M., Aguilar-Valenzuela, R., Konczy, P., Mascarenhas, M., Ziebell, K., Torres, A. G., Karmali, M. A. & Coombes, B. K. Genome sequence of adherent-invasive Escherichia coli and comparative genomic analysis with other E. coli pathotypes. *BMC Genomics* **11,** 667 (Nov. 2010).

115. Martinez-Medina, M., Mora, A., Blanco, M., López, C., Alonso, M. P., Bonacorsi, S., Nicolas-Chanoine, M.-H., Darfeuille-Michaud, A., Garcia-Gil, J. & Blanco, J. Similarity and divergence among adherent-invasive Escherichia coli and extraintestinal pathogenic E. coli strains. *J. Clin. Microbiol.* **47,** 3968–3979 (Dec. 2009).

116. Gibold, L., Garenaux, E., Dalmasso, G., Gallucci, C., Cia, D., Mottet-Auselo, B., Faıs, T., Darfeuille-Michaud, A., Nguyen, H. T. T., Barnich, N., Bonnet, R. & Delmas, J. The Vat-AIEC protease promotes crossing of the intestinal mucus layer by Crohn's disease-associated Escherichia coli. *Cell. Microbiol.* **18,** 617–631 (May 2016).

117. Zhang, Y., Rowehl, L., Krumsiek, J. M., Orner, E. P., Shaikh, N., Tarr, P. I., Sodergren, E., Weinstock, G. M., Boedeker, E. C., Xiong, X., Parkinson, J., Frank, D. N., Li, E. & Gathungu, G. Identification of Candidate Adherent-Invasive E. coli Signature Transcripts by Genomic/Transcriptomic Analysis. en. *PLoS One* **10,** e0130902 (June 2015).

118. Cieza, R. J., Hu, J., Ross, B. N., Sbrana, E. & Torres, A. G. The IbeA invasin of adherent-invasive Escherichia coli mediates interaction with intestinal epithelia and macrophages. *Infect. Immun.* **83,** 1904–1918 (May 2015).

119. Vazeille, E., Chassaing, B., Buisson, A., Dubois, A., de Vallée, A., Billard, E., Neut, C., Bommelaer, G., Colombel, J.-F., Barnich, N., Darfeuille-Michaud, A. & Bringer, M.-A. GipA Factor Supports Colonization of Peyer's Patches by Crohn's Disease-associated Escherichia Coli. *Inflamm. Bowel Dis.* **22,** 68–81 (Jan. 2016).

120. Monk, J., Nogales, J. & Palsson, B. O. Optimizing genome-scale network reconstructions. *Nat. Biotechnol.* **32,** 447–452 (May 2014).

121. Gronbach, K., Flade, I., Holst, O., Lindner, B., Ruscheweyh, H. J., Wittmann, A., Menz, S., Schwiertz, A., Adam, P., Stecher, B., Josenhans, C., Suerbaum, S., Gruber, A. D., Kulik, A., Huson, D., Autenrieth, I. B. & Frick, J.-S. Endotoxicity of lipopolysaccharide as a determinant of T-cell-mediated colitis induction in mice. *Gastroenterology* **146,** 765–775 (Mar. 2014).

122. Human Microbiome Jumpstart Reference Strains Consortium, Nelson, K. E., Weinstock, G. M., Highlander, S. K., Worley, K. C., Creasy, H. H., Wortman, J. R., Rusch, D. B., Mitreva, M., Sodergren, E., Chinwalla, A. T., Feldgarden, M., Gevers, D., Haas, B. J., Madupu, R., Ward, D. V., Birren, B. W., Gibbs, R. A., Methe, B., Petrosino, J. F., Strausberg, R. L., Sutton, G. G., White, O. R., Wilson, R. K., Durkin, S., Giglio, M. G., Gujja, S., Howarth, C., Kodira, C. D., Kyrpides, N., Mehta, T., Muzny, D. M., Pearson, M., Pepin, K., Pati, A., Qin, X., Yandava, C., Zeng, Q., Zhang, L., Berlin, A. M., Chen, L., Hepburn, T. A., Johnson, J., McCorrison, J., Miller, J., Minx, P., Nusbaum, C., Russ, C., Sykes, S. M., Tomlinson, C. M., Young, S., Warren, W. C., Badger, J., Crabtree, J., Markowitz, V. M., Orvis, J., Cree, A., Ferriera, S., Fulton, L. L., Fulton, R. S., Gillis, M., Hemphill, L. D., Joshi, V., Kovar, C., Torralba, M., Wetterstrand, K. A., Abouellleil, A., Wollam, A. M., Buhay, C. J., Ding, Y., Dugan, S., FitzGerald, M. G., Holder, M., Hostetler, J., Clifton, S. W., Allen-Vercoe, E., Earl, A. M., Farmer, C. N., Liolios, K., Surette, M. G., Xu, Q., Pohl, C., Wilczek-Boney, K. & Zhu, D. A catalog of reference genomes from the human microbiome. *Science* **328,** 994–999 (May 2010).

123. Cummings, J. H. Cellulose and the human gut. *Gut* **25,** 805–810 (Aug. 1984).

124. Cocinero, E. J., Gamblin, D. P., Davis, B. G. & Simons, J. P. The building blocks of cellulose: the intrinsic conformational structures of cellobiose, its epimer, lactose, and their singly hydrated complexes. *J. Am. Chem. Soc.* **131,** 11117–11123 (Aug. 2009).

125. Ravcheev, D. A. & Thiele, I. Comparative Genomic Analysis of the Human Gut Microbiome Reveals a Broad Distribution of Metabolic Pathways for the Degradation of Host-Synthetized Mucin Glycans and Utilization of Mucin-Derived Monosaccharides. en. *Front. Genet.* **8,** 111 (Aug. 2017).

126. Bernier-Fébreau, C., du Merle, L., Turlin, E., Labas, V., Ordonez, J., Gilles, A.-M. & Le Bouguénec, C. Use of deoxyribose by intestinal and extraintestinal pathogenic Escherichia coli strains: a metabolic adaptation involved in competitiveness. *Infect. Immun.* **72,** 6151–6156 (Oct. 2004).

127. Martinez-Jéhanne, V., du Merle, L., Bernier-Fébreau, C., Usein, C., Gassama-Sow, A., Wane, A.-A., Gouali, M., Damian, M., Aıdara-Kane, A., Germani, Y., Fontanet, A., Coddeville, B., Guérardel, Y. & Le Bouguénec, C. Role of deoxyribose catabolism in colonization of the murine intestine by pathogenic Escherichia coli strains. *Infect. Immun.* **77,** 1442–1450 (Apr. 2009).

128. Neidhardt, F. C. & Curtiss, R. *Escherichia coli and Salmonella: cellular and molecular biology* (ASM press Washington, DC: 1999).

129. Kaleta, C., Schäuble, S., Rinas, U. & Schuster, S. Metabolic costs of amino acid and protein production in Escherichia coli. *Biotechnol. J.* **8,** 1105–1114 (Sept. 2013).

130. Sharpton, T., Lyalina, S., Luong, J., Pham, J., Deal, E. M., Armour, C., Gaulke, C., Sanjabi, S. & Pollard, K. S. Development of Inflammatory Bowel Disease Is Linked to a Longitudinal Restructuring of the Gut Metagenome in Mice. *mSystems* **2** (Sept. 2017).

131. Cleynen, I., Boucher, G., Jostins, L., Schumm, L. P., Zeissig, S., Ahmad, T., Andersen, V., Andrews, J. M., Annese, V., Brand, S., Brant, S. R., Cho, J. H., Daly, M. J., Dubinsky, M., Duerr, R. H., Ferguson, L. R., Franke, A., Gearry, R. B., Goyette, P., Hakonarson, H., Halfvarson, J., Hov, J. R., Huang, H., Kennedy, N. A., Kupcinskas, L., Lawrance, I. C., Lee, J. C., Satsangi, J., Schreiber, S., Théâtre, E., van der Meulen-de Jong, A. E., Weersma, R. K., Wilson, D. C., International Inflammatory Bowel Disease Genetics Consortium, Parkes, M., Vermeire, S., Rioux, J. D., Mansfield, J., Silverberg, M. S., Radford-Smith, G., McGovern, D. P. B., Barrett, J. C. & Lees, C. W. Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. *Lancet* **387,** 156–167 (Jan. 2016).

132. Langdon, A., Crook, N. & Dantas, G. The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation. *Genome Med.* **8,** 39 (Apr. 2016).

133. Huang, E. Y., Inoue, T., Leone, V. A., Dalal, S., Touw, K., Wang, Y., Musch, M. W., Theriault, B., Higuchi, K., Donovan, S., Gilbert, J. & Chang, E. B. Using corticosteroids

to reshape the gut microbiome: implications for inflammatory bowel diseases. *Inflamm. Bowel Dis.* **21,** 963–972 (May 2015).

134. Jiang, H., Lei, R., Ding, S.-W. & Zhu, S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* **15,** 182 (June 2014).

135. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (May 2011).

136. *scikit-bio* `http://scikit-bio.org/`. Accessed: 2018-5-3. 2018.

137. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32,** 605–607 (Feb. 2016).

138. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25,** 1043–1055 (July 2015).

139. Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., Nomura, W., Zhang, Z., Mori, H., Feist, A. M. & Palsson, B. O. iML1515, a knowledgebase that computes Escherichia coli traits. *Nat. Biotechnol.* **35,** 904–908 (Oct. 2017).

140. McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. en. *Nucleic Acids Res.* **32,** W20–5 (July 2004).

141. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst. Biol.* **7,** 74 (Aug. 2013).

142. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A. & Pevzner, P. A. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. en. *J. Comput. Biol.* **19,** 455–477 (May 2012).

143. Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A. & Zagnitko, O. The RAST Server: rapid annotations using subsystems technology. en. *BMC Genomics* **9,** 75 (Feb. 2008).

144. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (Jan. 2017).

145. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* **5,** 93–121 (Jan. 2010).

146. Negroni, A., Costanzo, M., Vitali, R., Superti, F., Bertuccini, L., Tinari, A., Minelli, F., Di Nardo, G., Nuti, F., Pierdomenico, M., Cucchiara, S. & Stronati, L. Characterization

of adherent-invasive Escherichia coli isolated from pediatric patients with inflammatory bowel disease. en. *Inflamm. Bowel Dis.* **18,** 913–924 (May 2012).

147. Sartor, R. B. & Mazmanian, S. K. Intestinal microbes in inflammatory bowel diseases. *The American Journal of Gastroenterology Supplements* **1,** 15 (2012).

148. Kotlowski, R., Bernstein, C. N., Sepehri, S. & Krause, D. O. High prevalence of Escherichia coli belonging to the B2+D phylogenetic group in inflammatory bowel disease. en. *Gut* **56,** 669–675 (May 2007).

149. Martinez-Medina, M., Aldeguer, X., Lopez-Siles, M., González-Huix, F., López-Oliu, C., Dahbi, G., Blanco, J. E., Blanco, J., Garcia-Gil, L. J. & Darfeuille-Michaud, A. Molecular diversity of Escherichia coli in the human gut: new ecological evidence supporting the role of adherent-invasive E. coli (AIEC) in Crohn's disease. en. *Inflamm. Bowel Dis.* **15,** 872–882 (June 2009).

150. O'brien, C. L., Bringer, M.-A., Holt, K. E., Gordon, D. M., Dubois, A. L., Barnich, N., Darfeuille-Michaud, A. & Pavli, P. Comparative genomics of Crohn9s disease-associated adherent-invasive Escherichia coli. *Gut,* gutjnl–2015 (2016).

151. Conte, M. P., Longhi, C., Marazzato, M., Conte, A. L., Aleandri, M., Lepanto, M. S., Zagaglia, C., Nicoletti, M., Aloi, M., Totino, V., Palamara, A. T. & Schippa, S. Adherent-invasive Escherichia coli (AIEC) in pediatric Crohn's disease patients: phenotypic and genetic pathogenic features. en. *BMC Res. Notes* **7,** 748 (Oct. 2014).

152. Krause, D. O., Little, A. C., Dowd, S. E. & Bernstein, C. N. Complete genome sequence of adherent invasive Escherichia coli UM146 isolated from Ileal Crohn's disease biopsy tissue. en. *J. Bacteriol.* **193,** 583 (Jan. 2011).

153. Nash, J. H., Villegas, A., Kropinski, A. M., Aguilar-Valenzuela, R., Konczy, P., Mascarenhas, M., Ziebell, K., Torres, A. G., Karmali, M. A. & Coombes, B. K. Genome sequence of adherent-invasive Escherichia coli and comparative genomic analysis with other E. coli pathotypes. en. *BMC Genomics* **11,** 667 (Nov. 2010).

154. Nowrouzian, F. L., Adlerberth, I. & Wold, A. E. Enhanced persistence in the colonic microbiota of Escherichia coli strains belonging to phylogenetic group B2: role of virulence factors and adherence to colonic cells. en. *Microbes Infect.* **8,** 834–840 (Mar. 2006).

155. O'Brien, E. J., Monk, J. M. & Palsson, B. O. Using Genome-scale Models to Predict Biological Capabilities. en. *Cell* **161,** 971–987 (May 2015).

156. Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., Feist, A. M. & Palsson, B. Ø. Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. en. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 20338–20343 (Dec. 2013).

157. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. en. *Bioinformatics* **28,** 3150–3152 (Dec. 2012).

158. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).

159. Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D. & Thomas, P. D. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. en. *Nucleic Acids Res.* **45,** D183–D189 (Jan. 2017).

160. Park, S. J., Chao, G. & Gunsalus, R. P. Aerobic regulation of the sucABCD genes of Escherichia coli, which encode alpha-ketoglutarate dehydrogenase and succinyl coenzyme A synthetase: roles of ArcA, Fnr, and the upstream sdhCDAB promoter. en. *J. Bacteriol.* **179,** 4138–4142 (July 1997).

161. Mahowald, M. A., Rey, F. E., Seedorf, H., Turnbaugh, P. J., Fulton, R. S., Wollam, A., Shah, N., Wang, C., Magrini, V., Wilson, R. K., Cantarel, B. L., Coutinho, P. M., Henrissat, B., Crock, L. W., Russell, A., Verberkmoes, N. C., Hettich, R. L. & Gordon, J. I. Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla. en. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 5859–5864 (Apr. 2009).

162. Hall, D. R., Bond, C. S., Leonard, G. A., Watt, C. I., Berry, A. & Hunter, W. N. Structure of tagatose-1,6-bisphosphate aldolase. Insight into chiral discrimination, mechanism, and specificity of class II aldolases. en. *J. Biol. Chem.* **277,** 22018–22024 (June 2002).

163. Hall, D. R., Leonard, G. A., Reed, C. D., Watt, C. I., Berry, A. & Hunter, W. N. The crystal structure of Escherichia coli class II fructose-1, 6-bisphosphate aldolase in complex with phosphoglycolohydroxamate reveals details of mechanism and specificity. en. *J. Mol. Biol.* **287,** 383–394 (Mar. 1999).

164. Ferenci, T. Adaptation to life at micromolar nutrient levels: the regulation of Escherichia coli glucose transport by endoinduction and cAMP. en. *FEMS Microbiol. Rev.* **18,** 301–317 (July 1996).

165. Wilson, J. W., Schurr, M. J., LeBlanc, C. L., Ramamurthy, R., Buchanan, K. L. & Nickerson, C. A. Mechanisms of bacterial pathogenicity. en. *Postgrad. Med. J.* **78,** 216–224 (Apr. 2002).

166. Niu, S., Jiang, S. Q. & Hong, J. Salmonella typhimurium pgtB mutants conferring constitutive expression of phosphoglycerate transporter pgtP independent of pgtC. en. *J. Bacteriol.* **177,** 4297–4302 (Aug. 1995).

167. Jones, S. A., Chowdhury, F. Z., Fabich, A. J., Anderson, A., Schreiner, D. M., House, A. L., Autieri, S. M., Leatham, M. P., Lins, J. J., Jorgensen, M., Cohen, P. S. & Conway, T. Respiration of Escherichia coli in the mouse intestine. en. *Infect. Immun.* **75,** 4891–4899 (Oct. 2007).

168. Arenas-Hernández, M. M. P., Martınez-Laguna, Y. & Torres, A. G. Clinical implications of enteroadherent Escherichia coli. en. *Curr. Gastroenterol. Rep.* **14,** 386–394 (Oct. 2012).

169. Erbersdobler, H. F. & Faist, V. Metabolic transit of Amadori products. en. *Nahrung* **45,** 177–181 (June 2001).

170. Wiame, E. & Van Schaftingen, E. Fructoselysine 3-epimerase, an enzyme involved in the metabolism of the unusual Amadori compound psicoselysine in Escherichia coli. en. *Biochem. J* **378,** 1047–1052 (Mar. 2004).

171. Bui, T. P. N., Ritari, J., Boeren, S., de Waard, P., Plugge, C. M. & de Vos, W. M. Production of butyrate from lysine and the Amadori product fructoselysine by a human gut commensal. en. *Nat. Commun.* **6,** 10062 (Dec. 2015).

172. Geirnaert, A., Calatayud, M., Grootaert, C., Laukens, D., Devriese, S., Smagghe, G., De Vos, M., Boon, N. & Van de Wiele, T. Butyrate-producing bacteria supplemented in vitro to Crohn's disease patient microbiota increased butyrate production and enhanced intestinal epithelial barrier integrity. en. *Sci. Rep.* **7,** 11450 (Sept. 2017).

173. Walsh, C. J., Guinane, C. M., Hill, C., Ross, R. P., O'Toole, P. W. & Cotter, P. D. In silico identification of bacteriocin gene clusters in the gastrointestinal tract, based on the Human Microbiome Project's reference genome database. en. *BMC Microbiol.* **15,** 183 (Sept. 2015).

174. Clarke, D. J., Chaudhuri, R. R., Martin, H. M., Campbell, B. J., Rhodes, J. M., Constantinidou, C., Pallen, M. J., Loman, N. J., Cunningham, A. F., Browning, D. F. & Henderson, I. R. Complete genome sequence of the Crohn's disease-associated adherent-invasive Escherichia coli strain HM605. en. *J. Bacteriol.* **193,** 4540 (Sept. 2011).

175. Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., Gillespie, J. J., Gough, R., Hix, D., Kenyon, R., Machi, D., Mao, C., Nordberg, E. K., Olson, R., Overbeek, R., Pusch, G. D., Shukla, M., Schulman, J., Stevens, R. L., Sullivan, D. E., Vonstein, V., Warren, A., Will, R., Wilson, M. J. C., Yoo, H. S., Zhang, C., Zhang, Y. & Sobral, B. W. PATRIC, the bacterial bioinformatics database and analysis resource. en. *Nucleic Acids Res.* **42,** D581–91 (Jan. 2014).

176. Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Gallo Cassarino, T., Bertoni, M., Bordoli, L. & Schwede, T. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. en. *Nucleic Acids Res.* **42,** W252–8 (July 2014).

177. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. en. *Nat. Protoc.* **5,** 725–738 (Apr. 2010).

178. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. en. *J. Mol. Graph.* **14,** 33–8, 27–8 (Feb. 1996).

179. Roberts, E., Eargle, J., Wright, D. & Luthey-Schulten, Z. MultiSeq: unifying sequence and structure data for evolutionary analysis. en. *BMC Bioinformatics* **7,** 382 (Aug. 2006).

180. The UniProt Consortium. UniProt: the universal protein knowledgebase. en. *Nucleic Acids Res.* **45,** D158–D169 (Jan. 2017).

181. Martınez-Antonio, A., Janga, S. C. & Thieffry, D. Functional organisation of Escherichia coli transcriptional regulatory network. en. *J. Mol. Biol.* **381,** 238–247 (Aug. 2008).

182. Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J. & Palsson, B. O. *Integrating high-throughput and computational data elucidates bacterial networks* 2004.

183. Chandrasekaran, S. & Price, N. D. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. en. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 17845–17850 (Oct. 2010).

184. Rustad, T. R., Minch, K. J., Ma, S., Winkler, J. K., Hobbs, S., Hickey, M., Brabant, W., Turkarslan, S., Price, N. D., Baliga, N. S. & Sherman, D. R. Mapping and manipulating the Mycobacterium tuberculosis transcriptome using a transcription factor overexpression-derived regulatory network. en. *Genome Biol.* **15,** 502 (2014).

185. Kochanowski, K., Gerosa, L., Brunner, S. F., Christodoulou, D., Nikolaev, Y. V. & Sauer, U. Few regulatory metabolites coordinate expression of central metabolic genes in Escherichia coli. en. *Mol. Syst. Biol.* **13,** 903 (Jan. 2017).

186. Browning, D. F. & Busby, S. J. W. Local and global regulation of transcription initiation in bacteria. en. *Nat. Rev. Microbiol.* **14,** 638–650 (Oct. 2016).

187. Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muñiz-Rascado, L., Garcıa-Sotelo, J. S., Alquicira-Hernández, K., Martınez-Flores, I., Pannier, L., Castro-Mondragón, J. A., Medina-Rivera, A., Solano-Lira, H., Bonavides-Martınez, C., Pérez-Rueda, E., Alquicira-Hernández, S., Porrón-Sotelo, L., López-Fuentes, A., Hernández-Koutoucheva, A., Moral-Chávez, V. D., Rinaldi, F. & Collado-Vides, J. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.* **44,** D133–D143 (Jan. 2016).

188. Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., The DREAM5 Consortium, Kellis, M., Collins, J. J. & Stolovitzky, G. Wisdom of crowds for robust gene network inference. *Nat. Methods* **9,** 796 (July 2012).

189. Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J. & Gardner, T. S. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. en. *PLoS Biol.* **5,** e8 (Jan. 2007).

190. Federowicz, S., Kim, D., Ebrahim, A., Lerman, J., Nagarajan, H., Cho, B.-K., Zengler, K. & Palsson, B. Determining the control circuitry of redox metabolism at the genome-scale. en. *PLoS Genet.* **10,** e1004264 (Apr. 2014).

191. Cho, B.-K., Federowicz, S. A., Embree, M., Park, Y.-S., Kim, D. & Palsson, B. Ø. The PurR regulon in Escherichia coli K-12 MG1655. en. *Nucleic Acids Res.* **39,** 6456–6464 (Aug. 2011).

192. Cho, B.-K., Barrett, C. L., Knight, E. M., Park, Y. S. & Palsson, B. Ø. Genome-scale reconstruction of the Lrp regulatory network in Escherichia coli. en. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 19462–19467 (Dec. 2008).

193. Seo, S. W., Kim, D., Latif, H., O'Brien, E. J., Szubin, R. & Palsson, B. O. Deciphering Fur transcriptional regulatory network highlights its complex role beyond iron metabolism in Escherichia coli. en. *Nat. Commun.* **5,** 4910 (Sept. 2014).

194. Carrera, J., Estrela, R., Luo, J., Rai, N., Tsoukalas, A. & Tagkopoulos, I. An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of Escherichia coli. en. *Mol. Syst. Biol.* **10,** 735 (July 2014).

195. Lewis, N. E., Cho, B.-K., Knight, E. M. & Palsson, B. O. Gene expression profiling and the use of genome-scale in silico models of Escherichia coli for analysis: providing context for content. *J. Bacteriol.* **191,** 3437–3444 (June 2009).

196. Moretto, M., Sonego, P., Dierckxsens, N., Brilli, M., Bianco, L., Ledezma-Tejeida, D., Gama-Castro, S., Galardini, M., Romualdi, C., Laukens, K., Collado-Vides, J., Meysman, P. & Engelen, K. COLOMBOS v3.0: leveraging gene expression compendia for cross-species analyses. *Nucleic Acids Res.* **44,** D620–D623 (Jan. 2016).

197. Melas, I. N., Samaga, R., Alexopoulos, L. G. & Klamt, S. Detecting and removing inconsistencies between experimental data and signaling network topologies using integer linear programming on interaction graphs. en. *PLoS Comput. Biol.* **9,** e1003204 (Sept. 2013).

198. Berthoumieux, S., de Jong, H., Baptist, G., Pinel, C., Ranquet, C., Ropers, D. & Geiselmann, J. Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. en. *Mol. Syst. Biol.* **9,** 634 (2013).

199. Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. en. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 4164–4169 (Mar. 2004).

200. Kim, P. M. & Tidor, B. Subsystem identification through dimensionality reduction of large-scale gene expression data. en. *Genome Res.* **13,** 1706–1718 (July 2003).

201. Wu, S., Joseph, A., Hammonds, A. S., Celniker, S. E., Yu, B. & Frise, E. Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. en. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 4290–4295 (Apr. 2016).

202. Seo, S. W., Kim, D., O'Brien, E. J., Szubin, R. & Palsson, B. O. Decoding genome-wide GadEWX-transcriptional regulatory networks reveals multifaceted cellular responses to acid stress in Escherichia coli. en. *Nat. Commun.* **6,** 7970 (Aug. 2015).

203. Meilă, M. *Comparing Clusterings by the Variation of Information* in *Learning Theory and Kernel Machines* (Springer Berlin Heidelberg, 2003), 173–187.

204. Ye, Y. & Godzik, A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. en. *Bioinformatics* **19 Suppl 2,** ii246–55 (Oct. 2003).

205. Madan Babu, M. & Teichmann, S. A. Functional determinants of transcription factors in Escherichia coli: protein families and binding sites. *Trends Genet.* **19,** 75–79 (2003).

206. Galagan, J. E., Minch, K., Peterson, M., Lyubetskaya, A., Azizi, E., Sweet, L., Gomes, A., Rustad, T., Dolganov, G., Glotova, I., Abeel, T., Mahwinney, C., Kennedy, A. D., Allard, R., Brabant, W., Krueger, A., Jaini, S., Honda, B., Yu, W.-H., Hickey, M. J., Zucker, J., Garay, C., Weiner, B., Sisk, P., Stolte, C., Winkler, J. K., Van de Peer, Y., Iazzetti, P., Camacho, D., Dreyfuss, J., Liu, Y., Dorhoi, A., Mollenkopf, H.-J., Drogaris, P., Lamontagne, J., Zhou, Y., Piquenot, J., Park, S. T., Raman, S., Kaufmann, S. H. E., Mohney, R. P., Chelsky, D., Moody, D. B., Sherman, D. R. & Schoolnik, G. K. The Mycobacterium tuberculosis regulatory network and hypoxia. en. *Nature* **499,** 178–183 (July 2013).

207. Carrera, J., Rodrigo, G. & Jaramillo, A. Model-based redesign of global transcription regulation. en. *Nucleic Acids Res.* **37,** e38 (Apr. 2009).

208. Gustafsson, M. & Hörnquist, M. *Gene Expression Prediction by Soft Integration and the Elastic Net—Best Performance of the DREAM3 Gene Expression Challenge* 2010.

209. Gardner, T. S. *Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling* 2003.

210. Cho, B.-K., Kim, D., Knight, E. M., Zengler, K. & Palsson, B. O. Genome-scale reconstruction of the sigma factor network in Escherichia coli: topology and functional states. en. *BMC Biol.* **12,** 4 (Jan. 2014).

211. Seber, G. A. F. & Lee, A. J. *Linear Regression Analysis* (Wiley, 2012).

212. Arrieta-Ortiz, M. L., Hafemeister, C., Bate, A. R., Chu, T., Greenfield, A., Shuster, B., Barry, S. N., Gallitto, M., Liu, B., Kacmarczyk, T., Santoriello, F., Chen, J., Rodrigues, C. D. A., Sato, T., Rudner, D. Z., Driks, A., Bonneau, R. & Eichenberger, P. An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. en. *Mol. Syst. Biol.* **11,** 839 (Nov. 2015).

213. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. en. *PLoS Genet.* **3,** 1724–1735 (Sept. 2007).

214. Kim, M., Rai, N., Zorraquino, V. & Tagkopoulos, I. Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli. en. *Nat. Commun.* **7,** 13090 (Oct. 2016).

215. Ishihama, A. Prokaryotic genome regulation: a revolutionary paradigm. en. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.* **88,** 485–508 (2012).

216. Minch, K. J., Rustad, T. R., Peterson, E. J., Winkler, J., Reiss, D. J., Ma, S., Hickey, M., Brabant, W., Morrison, B., Turkarslan, S., Mawhinney, C., Galagan, J. E., Price, N. D., Baliga, N. S. & Sherman, D. R. *The DNA-binding network of Mycobacterium tuberculosis. Nat Commun 6: 5829* 2015.

217. Brooks, A. N., Reiss, D. J., Allard, A., Wu, W.-J., Salvanha, D. M., Plaisier, C. L., Chandrasekaran, S., Pan, M., Kaur, A. & Baliga, N. S. A system-level model for the microbial regulatory genome. *Mol. Syst. Biol.* **10** (2014).

218. Weiss, V., Medina-Rivera, A., Huerta, A. M., Santos-Zavaleta, A., Salgado, H., Morett, E. & Collado-Vides, J. Evidence classification of high-throughput protocols and confidence integration in RegulonDB. en. *Database* **2013,** bas059 (Jan. 2013).

219. Myers, K. S., Yan, H., Ong, I. M., Chung, D., Liang, K., Tran, F., Keleş, S., Landick, R. & Kiley, P. J. Genome-scale analysis of escherichia coli FNR reveals complex features of transcription factor binding. en. *PLoS Genet.* **9,** e1003565 (June 2013).

220. Park, D. M., Akhtar, M. S., Ansari, A. Z., Landick, R. & Kiley, P. J. The bacterial response regulator ArcA uses a diverse binding site architecture to regulate carbon oxidation globally. en. *PLoS Genet.* **9,** e1003839 (Oct. 2013).

221. Cho, S., Cho, Y.-B., Kang, T. J., Kim, S. C., Palsson, B. & Cho, B.-K. The architecture of ArgR-DNA complexes at the genome-scale in Escherichia coli. en. *Nucleic Acids Res.* **43,** 3079–3088 (Mar. 2015).

222. Cho, B.-K., Federowicz, S., Park, Y.-S., Zengler, K. & Palsson, B. Ø. Deciphering the transcriptional regulatory logic of amino acid metabolism. en. *Nat. Chem. Biol.* **8,** 65–71 (Jan. 2012).

223. Seo, S. W., Kim, D., Szubin, R. & Palsson, B. O. Genome-wide Reconstruction of OxyR and SoxRS Transcriptional Regulatory Networks under Oxidative Stress in Escherichia coli K-12 MG1655. en. *Cell Rep.* **12,** 1289–1299 (Aug. 2015).

224. Latif, H., Haythem, L., Stephen, F., Ali, E., Janna, T., Richard, S., Jose, U., Karsten, Z. & Bernhard, P. *ChIP-exo interrogation of Crp, DNA, and RNAP holoenzyme interactions* tech. rep. (2016).

225. Yang, L., Laurence, Y., Justin, T., O'Brien, E. J., Monk, J. M., Donghyuk, K., Li, H. J., Pep, C., Ali, E., Lloyd, C. J., Yurkovich, J. T., Bin, D., Andreas, D., Alex, T., Yuekai, S., Saunders, M. A. & Palsson, B. O. Systems biology definition of the core proteome of metabolism and expression is consistent with high-throughput data. *Proceedings of the National Academy of Sciences* **112,** 10810–10815 (2015).

226. Pascual-Montano, A., Carazo, J. M., Kochi, K., Lehmann, D. & Pascual-Marqui, R. D. Nonsmooth nonnegative matrix factorization (nsNMF). en. *IEEE Trans. Pattern Anal. Mach. Intell.* **28,** 403–415 (Mar. 2006).

227. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Systems* **1695,** 1–9 (2006).

228. Clauset, A., Newman, M. E. J. & Moore, C. Finding community structure in very large networks. en. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **70,** 066111 (Dec. 2004).

229. Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Clifford Lane, H. & Lempicki, R. A. *DAVID: Database for Annotation, Visualization, and Integrated Discovery* 2003.

230. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. & Smyth, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. en. *Nucleic Acids Res.* **43,** e47 (Apr. 2015).

231. Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. en. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **69,** 066138 (June 2004).