Computational Methods for Functional Interpretation of Diverse Omics Data

by

Sumaiya Nazeen

B.Sc. Engg., Bangladesh University of Engineering and Technology (2011)
S.M., Massachusetts Institute of Technology (2014)

Submitted to the
Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2019

Author . . . .
Department of Electrical Engineering and Computer Science
August 30, 2019

Certified by . . . .
Bonnie Berger
Simons Professor of Mathematics
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Computational Methods for Functional Interpretation of Diverse Omics Data

by

Sumaiya Nazeen

Submitted to the
Department of Electrical Engineering and Computer Science
on August 30, 2019, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Recent technological advances have resulted in an explosive growth of various types of "omics" data, including genomic, transcriptomic, proteomic, and metagenomic data. Functional interpretation of these data is key to elucidating the potential role of different molecular levels (e.g., genome, transcriptome, proteome, metagenome) in human health and disease. However, the massive size and heterogeneity of raw data pose substantial computational and statistical challenges in integrating and interpreting these data. To overcome these challenges, we need sophisticated approaches and scalable analytical frameworks. This thesis outlines two research efforts along these lines. First, we develop a novel three-tiered integrative omics framework for integrating and functionally analyzing heterogeneous omics datasets across a group of co-occurring diseases. We demonstrate the effectiveness of this framework in investigating the shared pathophysiology of autism spectrum disorder (ASD) and its multi-organ-system co-morbid diseases (e.g., inflammatory bowel disease, asthma, muscular dystrophy, cerebral palsy) and uncover a novel innate immunity connection between them. Second, we develop a new end-to-end computational tool, Carnelian, for robust, alignment-free functional profiling of whole metagenome sequencing reads, that is uniquely suited to finding hidden functional trends across diverse data sets in comparative analysis. Carnelian can find shared metabolic pathways, concordant functional dysbioses, and distinguish microbial metabolic function missed by state-of-the-art functional annotation tools. We demonstrate Carnelian's effectiveness on large-scale metagenomic studies of type-2 diabetes, Crohn's disease, Parkinson's disease, and industrialized versus non-industrialized cohorts.

Thesis Supervisor: Bonnie Berger
Title: Simons Professor of Mathematics
Professor of Electrical Engineering and Computer Science

To the wonderful memories of
my beloved maternal and paternal grandmothers,
Jamal Ara Begum and Hayaten Nesha

They always supported me and celebrated every success of mine.
I miss them every day!

Acknowledgments

First and foremost, I would like to immensely thank my thesis supervisor, Bonnie Berger, who has been a constant source of inspiration and encouragement. I am grateful to her for providing me the freedom to explore many topics in computational biology and to pursue my interests while offering technical guidance and advice throughout my years at MIT. Besides being an extremely resourceful mentor, Bonnie is also a compassionate person who always cared for my well-being. When I was diagnosed with a rare health condition during my second year of Ph.D. and had to go through major neurosurgery, she offered her full support. She patiently supported me throughout my surgery and the year-long recovery process, which I am really grateful for. I hope to extend the same level of kindness and compassion towards my students in the future.

I am grateful to Eric Alm and Manolis Kellis for being on my thesis committee. They believed in me, provided guidance, and pointed me to the areas where I could improve to succeed in academia. I want to acknowledge my collaborators, Nathan Palmer, Isaac Kohane, and Yun William Yu; this thesis would not have been possible without their invaluable contributions. Special thanks to Eric Alm, Mathieu Groussin, and Mathilde Poyet for sharing the unpublished gut microbiome data from Bostonian and Baka individuals from the Global Microbiome Conservancy project (`http://microbiomeconservancy.org/`) with us.

I am immensely thankful for my peers in the Berger lab, who have not only helped me grow as a researcher but also helped me persevere through the ups and downs of grad school. In particular, thanks to Po-Ru, George, and Mark for inspiring me to join Berger lab; to Deniz and Jian for helping me get started on research; to Sean, Deniz, Hoon, William, Ariya, Rohit, Noah, Yaron, Sepehr, Ibrahim, Sarah, Tristan, Perry, Youn, Ashwin, Andrew, Ellen, Brian, Max, Lillian, Ben, Rachel, and Alex for all the conversations I had with them about things both research and not research related; and to Patrice for greeting me everyday with warmth, lending a shoulder whenever I needed, and making life smooth for all of us in the lab (you really do keep

the lab from falling apart!).

My graduate experience has been greatly enhanced because of a warm set of people around me. I have enjoyed being involved with the GW6, GWAMIT, and TOC communities, and getting to know so many amazing minds. Thanks to Mie, Yiou, Yi, Becca, Lily, Cindy, and Debbie, with whom I have shared many enjoyable moments throughout grad school. I owe my gratitude to Christina; I have always appreciated your friendship and encouragement; the wisdom you shared with me about life and faith has always inspired me and helped me build confidence whenever I felt low. Thank you, Marty, for all the fun conversations and encouragements; I learned so much about life from you. I have also been blessed with the best roommates and neighbors: Rosalie, Kristi, Bella, Dora, Katherine, Jack, and Dhavala, who helped me have the much needed work-life balance.

I had a great time being part of the Bangladeshi Students Association of MIT and organizing social and cultural events. It was great to have an excellent set of friends in the Boston area, including Ehsan bhai, Nafisa apu, Nasim bhai, Zakia apu, Zubair, Saima, Sabrina, Hasnain bhai, Deeni, Nazia, Urmi, Sujoy bhai, Tapoti apu, Saquib, Mahi, Bristy, Sabrine, Tahsin, Sanzeed, Caleb, Lana, Phei Er, John, Tiara, Jianqiao, Nadim, Takian, Nishit, Deena, Murshid bhai, Zayan, Wasifa, Sourav, Adeeb, Protyasha, Taniya apu, Rumpa apu, Shafqat bhai, Maisha, Usama, and Ayesha. The time spent with all of you will always remain very memorable to me! I am also thankful for my friends from Bangladesh and all over the world made through Fulbright Fellowship. They always made time to chat with me despite being in different time zones. Special thanks to Fatema and Rimpi for patiently listening to all my rants throughout the years and keeping me sane during stressful periods of grad school.

I owe my gratitude to all the members of the EECS Graduate Office, Financial Services, CSAIL Headquarters, and International Students Office for helping me with the logistics of grad school. Special thanks to Janet, Leslie, Alicia, Kathy, and Sylvia for being there for me whenever I needed advice. I am incredibly grateful to my care teams at MIT Medical and MGH, who provided top-notch care for my health concerns. Throughout my Ph.D., I was generously supported by the International

Fulbright Science & Technology Fellowship, MIT's Ludwig Center Molecular Oncology Fellowship, and grants from the National Institute of Health, the Center for Microbiome Informatics and Therapeutics, and F. Thomson Leighton fund.

Thanks also to all the great advisors and mentors I have had over the years outside of MIT. This includes Dr. M Sohel Rahman, Dr. Mohammad Kaykobad, and Dr. Mohammad Eunus Ali at BUET, and Dr. Ishtiaque Ahmed at the University of Toronto, who have always been a source of great advice and encouragement.

I am grateful to my mom (Nilufar Jahan), dad (S M Salim Jahangir), and brothers (Shahriar Salim and Taqi Tahmeed Sakib) for their unwavering support of my research pursuits. Their love and trust in my abilities have been a significant source of inspiration throughout this journey. I would also like to acknowledge my extended family who have always prayed for me, believed in me, and encouraged me to push forward.

Finally, I express my utmost gratitude to my greatest supporter: to the Almighty Allah, who had bestowed this opportunity upon me and helped me persevere through every part of this journey. I am immensely grateful for His innumerable blessings all these years, which have helped me grow not only academically but also spiritually.

"He (Allah) is your Protector; and
excellent is the Protector and excellent is the Helper."
∼ Al Quran 22:78 ∼

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent advances in next-generation sequencing (NGS) technologies have revolutionized research in life sciences. The cost-effective and high throughput nature of these technologies has enabled us to study biological systems in unprecedented detail, generating massive amounts of genomic, transcriptomic, proteomic, epigenomic, metabolomic, and metagenomic data. Each type of omics data on its own typically provides some (associative) evidence of how a particular "ome" (e.g., genome, transcriptome, proteome, metabolome, or metagenome) contributes to a particular phenotype (e.g., disease); but by integrating across omics data, we can identify true causal relationships. However, integrating and interpreting omics data from a functional perspective faces challenges due to its high dimensionality and heterogeneity, the increasing diversity of experimental techniques, the noise in high-throughput measurements, and the nature of the underlying biology [1, 2]. These challenges require intelligent and scalable analytic frameworks. This thesis focuses on designing computational frameworks that (i) integrate heterogeneous omics data at different functional levels to reveal insights about groups of complex diseases which arise together [3], and (ii) enable a comparative functional analysis of large-scale metagenomic data from diverse study populations to uncover hidden functional potential of the microbiome [4, 5].

1.1 Types of omics data

The word "omics" when applied to a molecular term implies a comprehensive assessment of a set of molecules (http://omics.org/). The goal of omics studies is to understand the relationship between the genome and the functioning of cells. To this end, scientists investigate different molecules that play vital roles in the central dogma of life, namely DNA, RNA, and protein. DNA is transcribed into messenger RNA (mRNA) which is translated by the ribosome into polypeptide chains (i.e., sequences of amino acids) which singly or in complexes are known as proteins. Proteins fold into low-energy structures which function as cellular machines. Certain types of RNAs also function as cellular machines. Since proteins are dynamic and interacting molecules, taking proteomic measurements is often challenging. Furthermore, proteins undergo many post-translational modifications, and cannot be readily amplified; therefore, characterizing them is difficult at best [6]. Fortunately, by measuring transcripts of mRNA—the intermediate step between genes and proteins—we can bridge the gap between the genetic code and the functional molecules that run cellular machinery.

In multicellular organisms, nearly every cell has the same genome, thus the same genes. However, there is a wide range of physical, biochemical, and developmental differences observed among various cells and tissues. The expression patterns of genes and the production of specific proteins determine these differences. Many chemical compounds and proteins can attach to DNA and direct such actions as turning genes on or off or controlling the expression of transcripts, and thereby the production of proteins in particular cells. These compounds are collectively known as the epigenome. Moreover, small molecules that are substrates or products of metabolism constitute the metabolome—measurements of which can elucidate the underlying biochemical activity and state of cells or tissues. Environmental factors also play a role in influencing the biochemical processes in cells often through various microorganisms (e.g., virus, bacteria, fungi) that inhabit the host's body (i.e., the microbiome), potentially giving rise to different phenotypes in the host. Recent advances in experimental techniques have allowed us to build assays to deeply investigate every level of this process, resulting in a massive influx of different types of omics data (Box 1.1; [7]).

Box 1.1. Omics Data Types

Genomics is the study of an organism's linear DNA sequence i.e., the genome and its variants, and how these variants associate with diseases, response to treatment, or future patient prognosis. Associated technologies include genotype arrays [8, 9], next-generation sequencing (NGS) for whole genome sequencing [10], and whole exome sequencing (WES) [11].

Proteomics is the study of the proteins in a cell or tissue, their quantity, diversity, and interactions. Mass spectrometry (MS) based approaches [12, 13] are commonly used to investigate the proteome.

Transcriptomics focuses on the study of RNA levels genome-wide, both qualitatively (which transcripts are present, identification of splice sites, RNA editing sites, etc.) and quantitatively (how much of each transcript is expressed). Associated technologies include probe-based arrays [14] and RNA-Seq [15].

Epigenomics is the study of the epigenome which consists of chemical compounds and proteins that can attach to DNA and direct such actions as turning genes on or off or controlling the expression of transcripts and thereby the production of proteins in particular cells. Such data often comes from the genome-wide characterization of reversible modifications of DNA or DNA-associated proteins, such as DNA methylation or histone acetylation [16].

Metabolomics is the large-scale study of metabolites (small molecules that are substrates or products of metabolism, such as amino acids, fatty acids, carbohydrates, etc.) and their interactions within a biological system that elucidates the underlying biochemical activity and state of cells or tissues. MS-based approaches are often used to quantify both relative and targeted small molecule abundances [17].

Metagenomics is the study of the metagenome, i.e., the collection of genetic material from all the microorganisms in a given environment, including bacteria, viruses, and fungi, collectively known as the microbiome. Associated technologies include NGS for 16S ribosomal RNA abundance and whole genome metagenomics quantification [18, 19].

1.2 Challenges in functional interpretation of omics data

Tremendous amounts of omics data have been generated over the past few decades and made available through public repositories such as Sequence Read Archive (SRA) and Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI), European Nucleotide Archive (ENA) and ArrayExpress at the European Bioinformatics Institute (EBI), DNA Data Bank of Japan (DDBJ), Joint Genome Institute (JGI), etc. The goal of functional interpretation of such data is to determine how the individual components (i.e., "omes") work together to produce a particular phenotype (e.g., disease). Naturally, a systems-level understanding of any phenotype requires looking at multiple omics levels in a large number of samples simultaneously in an integrated fashion [1, 20, 21]. However, at present, there is a significant lag in our ability to generate versus integrate and interpret omics data.

Since a single omics study often contains a small number of samples and has limited statistical power, combining information across multiple studies is an intuitive way to increase sensitivity. However, integrating and interpreting omics data across multiple studies face various biological and technical challenges, leading to the possibility of missing potentially valuable insights. These challenges include: (i) inconsistent nomenclature across different databases (e.g. gene, transcript, or protein identifiers), (ii) data generated by different platforms using different protocols (e.g. different array or next-generation sequencing (NGS) platforms, differences in sample preparation, processing pipelines, or study design), (iii) tissue heterogeneity (i.e., data generated from samples from different tissues), (iv) batch effect (i.e., technical differences in sample handling between batches of experiments), (v) size of data versus computational power and storage capacity, etc. All these contribute to the existing shortage of effective and robust frameworks to integrate and analyze omics data across different cell types, tissues, developmental phases, studies, and populations.

Furthermore, when it comes to metagenomic data, these challenges present themselves on a much larger scale. Shotgun sequencing, which has revolutionized metagenomics, presents many additional unique challenges [22]. First, the sheer size of shotgun metagenomic read data sets—which is typically much larger than data from individual genomes, targeted amplicon sequencing of marker genes from microbial communities (e.g., 16S ribosomal RNA), or other meta'omic experiments (e.g., meta-proteomics, meta-metabolomics, etc.)—poses significant computational challenges [22, 23]. Secondly, shotgun reads often come from a mixture of genomes where the genome from which each read comes is unknown and so is the position of the read within the genome. Moreover, the vast majority of the microbial diversity is not represented in any reference database or otherwise characterized in most environments [24–26]. Even for species with sequenced genomes, reference databases do not capture the full collection of genes and functions present across different strains [27–29]. Aside from the reads that cannot be assigned to a taxon or gene or function, it is challenging to generate comparable estimates of abundance from the remaining reads due to a variety of biological and technical biases related to study design, experimental protocols, and bioinformatics pipelines that affect the relationship between true abundance in the community and the number of reads observed for a taxonomic or functional category. These challenges not only complicate answering the question: which taxa and functions are present in a microbial sample, but also interfere with a useful comparison of metagenomic profiles across samples, either within a study or across studies.

1.3 Prior work on these challenges

1.3.1 Study of groups of diseases in the literature

Although different diseases may appear unrelated at an organismal level, it is highly unlikely that they arise completely independently from one another. They often share molecular components so that perturbations causing disease in one organ system can affect another [30, 31]. In neuropsychiatry, where many disorders do not have clear boundaries in terms of their pathophysiology or diagnosis [32, 33], researchers have pursued this line of thinking. Indeed, there is now growing evidence that rare variants ranging from chromosomal abnormalities and copy number variations (CNV) to single nucleotide variations (SNV) have implications for autism spectrum disorder (ASD) and other neuropsychiatric conditions [34–41]. For example, single nucleotide polymorphisms (SNPs), which overlap genes in common molecular pathways, such as calcium channel signaling, are shared between ASD, attention deficit-hyperactivity disorder (ADHD), bipolar disorder (BD), major depressive disorder (MDD), and schizophrenia [42]. Other investigators have integrated genomic data for multiple immune-related diseases, revealing that there are shared molecular commonalities between them [43]. These efforts have mainly focused on developing multinomial models and statistical approaches for integrating genomic and genotype data from multiple genome-wide association studies so that researchers can detect rare variants with increased statistical power. However, efforts looking at groups of diseases have so far remained confined to a single type of omics data, diseases of the same organ system or diseases with a common theme (e.g., immune-related diseases, psychiatric diseases) to avoid dealing with the underlying heterogeneity. Recent studies based on electronic health records and database/literature mining have identified various groups of co-occurring diseases [44–48]. For example, researchers have identified several diseases spanning multiple organ systems that co-occur in ASD patients at a high prevalence rate [44, 45]. These diseases include seizures [49, 50], gastrointestinal disorders [51, 52], ear infections and auditory disorders, sleep disorders [53], muscular dystrophy [54–56], cardiac disorders, psychiatric illnesses [57, 58], etc. Thus, for understanding the molecular basis of a disease like ASD with a complex etiology, there is a need to expand the omics exploration outside of one level and one organ system, in particular, the brain to conditions related to other organ systems that co-occur with it.

1.3.2 Functional profiling of metagenomic reads and comparative functional metagenomics

Functional profiling of metagenomic reads refers to the task of assigning reads to known biological functions (e.g., catalytic action, functional domain categories, genes) and estimating abundances of those functional terms. Traditional whole metagenome functional annotation approaches assemble reads into large contigs, translate them into open reading frames (ORF), and annotate them using protein sequence homology, often using existing alignment tools such as BLAST [59], profile Hidden Markov Models (HMMs), or position-specific weight matrices (PWMs). Such methods include RAST [60], Megan4 [61], MEDUSA [62], Tentacle [63], MOCat2 [64], IMG4 [65], and gene catalogue-based methods [66, 67]. Since assembly is a slow, resource-heavy, and lossy process, annotating translated reads directly via sequence homology or read-mapping is used by another class of tools, including MG-RAST [68], HUMAnN [69], ShotMap [70], Fun4Me [71], mi-faser [72], and HUMAnN2 [73]. However, alignment-based read mapping remains time-consuming when comparing hundreds of samples from different cohorts [74, 75]. HUMAnN2 and mi-faser significantly speed up the alignment step by using a fast protein aligner, DIAMOND [76], and thus can accurately and quickly capture functions from sequences corresponding to known proteins. However, because they often use strict alignment criteria, they are challenged in capturing shared features of functionally similar proteins that are not-so-sequence-similar, multi-domain proteins, and remote homologs [77].

Naturally, predicting function without having characterized a protein experimentally is difficult and runs the risk of false positives. For well-studied populations and environments, there might not be a need to do so. However, when analyzing data from less studied populations and environments—so often the case in metagenomic analysis, a large fraction of reads sequenced do not directly correspond to genes of known species [25, 26]. Thus methods that depend on alignment do not perform as well. Techniques from the field of remote homology detection can be used to explicitly guess at shared functions between an unknown protein and an existing one, but they operate at the level of entire protein sequences, rather than Whole Metagenome Shotgun (WMS) sequencing reads. Hence, functional profiling of metagenomic sequences remains an open problem in metagenomics research [22, 23, 70, 78].

The real power of metagenomic analysis lies in performing comparisons between samples, within a study, or across multiple studies spanning different populations and environments. Detecting differential changes in the functional capacities of microbial communities often requires larger sample sizes than are feasible within a single study. Hence, meta-analyses and comparisons of new and existing cohorts are increasingly becoming popular [67, 79–84]. These efforts have primarily sought to uncover shared taxonomic dysbiosis (i.e., microbial imbalance) between study populations for a given disease. However, these attempts have generally not found shared taxonomic dysbiosis, probably because the healthy microbiomes used as the background differ significantly in taxonomic composition, to begin with. Because different species may fill the same ecological niche, the traditional focus on taxonomy can lose sight of the *functional* relatedness of the microbiomes of two individuals—i.e., commonalities and differences in the functional capabilities of microbial populations [25]. In the large meta-analyses cited above [67, 79–84], there was some attempt to perform functional profiling (in addition to taxonomic profiling), but due to limitations in the study design and methods available, they were unable to find concordant pathways, which one *would* expect from the same disease. Thus, better functional profiling is important to uncovering trends in functional relatedness when comparing study cohorts; this remains an unsolved challenge due to inconsistencies and incompleteness of annotations of microbial genes across reference databases and the lack of comparability of existing relative abundance statistics across samples and studies [23, 78].

1.4 Outline

The remainder of this thesis is organized as two self-contained chapters, parts of which have been previously published and parts, currently under review at a reputed peer-reviewed journal, as well as some concluding remarks.

Chapter 2: Multi-level integrative omics analysis for ASD and its comorbidities

Over the years, ASD has baffled researchers not only with its heterogeneity but also its co-occurrence with many seemingly unrelated diseases of different organ systems (known as comorbidities of ASD). In this chapter, we introduce a three-tiered, statistically robust, meta-analysis approach [3] to capture the shared signals at gene and pathway levels that form the basis of ASD's co-occurrence with other diseases. Integrating heterogeneous transcriptomic data from 53 studies of 12 different diseases at the gene, pathway, and disease levels, our pipeline reveals a novel innate immunity connection between ASD and its comorbidities. Our statistical approach bridges the gap between frequentist and Bayesian statistics. An analysis of this scale for studying ASD and its comorbidities is unheard of as per our knowledge.

Chapter 3: Robust comparative functional metagenomics across diverse study populations

In this chapter, we introduce Carnelian, a novel compositional tool for profiling the metabolic functional potential of a metagenome from whole metagenome sequencing reads, and an end-to-end pipeline that is uniquely suited to finding common functional trends while comparing metagenomic data sets from different study populations [4,5]. Functionally similar proteins often share compositional (gapped k-mer) features in their amino acid sequence, even across species. Carnelian builds on this observation and leverages low-density locality-sensitive hashing [85,86] with a gapped k-mer classifier [87], which is better able to detect the ECs (Enzyme Commission terms that classify proteins by their enzymatic action) present in non-annotated species, while simultaneously avoiding forced spurious labels through training on a negative set. When used with our in-house comprehensive reference protein database focused on comparing metabolic functionality, as opposed to using typical protein databases that contain non-prokaryotic and non-metabolic annotations, and a new read normalization approach, Carnelian produces quantitative summaries of the microbial metabolic functional capacities that are readily comparable across samples. Finally, we present a principled statistical approach for finding shared metabolic pathways using Carnelian-generated functional profiles that enables the discovery of hidden functional trends across diverse study populations. On a variety of simulated and real datasets (both published and unpublished), we demonstrate Carnelian's superior performance in finding shared metabolic pathways, concordant functional dysbioses, and distinguishing Enzyme Commission (EC) terms missed by state-of-the-art metagenomic functional profiling tools.

Chapter 4: Conclusion

Here we summarize the central themes of the previous chapters.

Chapter 2

Multi-level integrative omics analysis for ASD and its comorbidities

Autism spectrum disorder (ASD) is a common neurodevelopmental disorder that tends to co-occur with other diseases, including asthma, inflammatory bowel disease, infections, cerebral palsy, dilated cardiomyopathy, muscular dystrophy, and schizophrenia. However, the molecular basis of this co-occurrence, and whether it is due to a shared component that influences both pathophysiology and environmental triggering of illness, has not been elucidated. To address this, we introduce a three-tiered omics analysis pipeline [3] that functions at the gene, pathway, and disease levels across ASD and its comorbidities. Our pipeline reveals a novel shared innate immune component between ASD and all but three of its comorbidities that were examined. In particular, we find that the Toll-like receptor signaling and the chemokine signaling pathways, which are key pathways in the innate immune response, have the highest shared statistical significance. Moreover, the disease genes that overlap these two innate immunity pathways can be used to classify the cases of ASD and its comorbidities versus controls with at least 70% accuracy. This finding suggests that a neuropsychiatric condition and the majority of its non-brain-related comorbidities share a dysregulated signal that serves as not only a shared genetic basis for the diseases but also as a link to environmental triggers. It also raises the possibility that treatment and prophylaxis used for disorders of innate immunity may be successfully used for ASD patients with immune-related phenotypes.

2.1 Introduction

While at an organismal level, two or more diseases may appear unrelated, at the molecular level, it is unlikely that they arise entirely independently of one another. Studies of the human interactome—the molecular network of physical interactions (e.g., protein-protein, gene, metabolic, regulatory) between biological entities in cells—demonstrate that gene function and regulation are integrated at the level of an organism. Extensive patterns of shared co-occurrences also evidence molecular commonalities between seemingly disparate conditions [30].

Indeed, different disorders may share molecular components so that perturbations causing disease in one organ system can affect another [31]. However, since the phenotypes appear so different from each other, medical sub-disciplines address the conditions with sometimes wildly differing treatment protocols. If investigators can uncover the molecular links between seemingly different conditions, the connections may help explain why certain groups of diseases arise together and assist clinicians in their decision-making about the best treatments. Knowledge of shared molecular pathology might also provide therapeutic insights for the repositioning of existing drugs [88].

Such thinking has emerged most recently in the field of neuropsychiatry where, many such illnesses do not have clear boundaries in terms of their pathophysiology or diagnosis [32, 33]. Indeed, there is now growing evidence that rare variants ranging from chromosomal abnormalities and copy number variation (CNV) to single nucleotide variation (SNV) have implications in ASD and other neuropsychiatric conditions [34–41]. For example, single nucleotide polymorphisms (SNP) which overlap genes in common molecular pathways, such as calcium channel signaling, are shared in autism spectrum disorder (ASD), attention deficit-hyperactivity disorder (ADHD), bipolar disorder (BD), major depressive disorder (MDD), and schizophrenia [42]. CNVs, especially the rare ones, can explain a portion of the risk for multiple psychiatric disorders [38, 41]. For example, the 16p11.2 CNV encompassing around 600 kb (chr 16:29.5 - 30.2 Mb) has been implicated in multiple psychiatric disorders with the deletions being associated with ASD, developmental delay and intellectual disability (ID) and duplications being associated with ASD, schizophrenia, BD, and ID [38, 41, 89–93]. However, researchers have observed pathogenic variations in only about 30% of the ASD affected individuals [40, 94–97] and these variations often fail to explain the idiopathic (non-syndromic) ASD cases as well as why ASD affected individuals suffer from many other non-neuropsychiatric conditions.

To complement the evidence of genome-wide pleiotropy across neuropsychiatric diseases, rather than looking at one neurodevelopmental disease (ASD) and comparing it to other seemingly, brain-related diseases, we expand our exploration outside of the brain to conditions related to other organ systems that co-occur with ASD. Recent studies based on electronic health records [44, 45], have identified various comorbidities in ASD, including seizures [49, 50], gastrointestinal disorders [51, 52], ear infections and auditory disorders, developmental disorders, sleep disorders [53], muscular dystrophy [54–56], cardiac disorders and psychiatric illness [57, 58]. We introduce an integrative meta-analysis pipeline to identify the shared pathophysiological component between ASD and eleven other diseases, namely, asthma, bacterial and viral infection, chronic kidney disease, cerebral palsy, dilated cardiomyopathy, ear infection, epilepsy, inflammatory bowel disease, muscular dystrophy, schizophrenia, and upper respiratory infection, that have at least 5% prevalence in ASD patients [44, 45]. We ask the question, "Do these disease states - that are not included in the definition of ASD but co-occur at a significantly high frequency – illuminate dysregulated pathways that are important in ASD?" We reasoned that such pathways might offer previously hidden clues to shared molecular pathology.

Other investigators have integrated genomic data from genome-wide association studies (GWAS) and non-synonymous SNP studies for multiple immune-related diseases, revealing that combining genetic results better identified shared molecular commonalities [43]. We believe that adopting an integrative approach not only at the gene level but also at the biochemical pathway and disease levels will power the results still further.

In our three-tiered, meta-analysis approach, we: (i) look for statistically significant differentially expressed genes in every disease condition; (ii) identify their enrichment in canonical pathways; and (iii) determine the statistical significance of the shared pathways across multiple conditions. We are unaware of any analyses that go from population-based comorbidity clusters of ASD to a multi-level molecular analysis at anywhere near this breadth.

Our results unearth several innate immunity-related pathways - specifically, the Toll-like receptor and chemokine signaling pathways - as significant players in ASD and all but three of its examined comorbidities. Candidate genes in these two pathways significantly overlap in conditions of ASD, asthma, bacterial and viral infection, chronic kidney disease, dilated cardiomyopathy, ear infection, inflammatory bowel disease (IBD), muscular dystrophy, and upper respiratory infection. Candidate genes did not appear to be significantly shared in cerebral palsy, epilepsy, and schizophrenia. Notably, although bacterial and viral infection, respiratory infection, ear infection, IBD, and asthma have well-known connections with the immune system, we demonstrate that innate immunity pathways are shared by ASD and its comorbidities *irrespective* of whether they are immunity-related diseases or not.

Since both Toll-like receptor signaling and chemokine signaling pathways play crucial roles in innate immunity, the results suggest that this first-line defense system (that protects the host from infection by pathogens/environmental triggers) might be involved in ASD and specific comorbidities. If the profiles of genetic susceptibility pathways in relation to environmental triggers can be ascertained, they might help define new treatments such as vaccination [98] or other tolerization therapies [99]. Those might help individuals and families that are at high-risk for ASD to prevent and treat immune-related phenotypes of the illness.

2.2 Results

2.2.1 Overview of the three-tiered integrative omics pipeline

To integrate and interpret heterogeneous transcriptomic data across ASD and eleven of its comorbidities (Table 2.1) at gene and pathway levels, we introduce a novel three-tiered meta-analysis pipeline (Figure 2-1). Our meta-analysis starts at the gene level, in which we first identified the genes that are differentially expressed among cases and controls for a given disease. We then extend this analysis to the pathway level, where we investigate the pathways that are significantly enriched in candidate genes for a given disease. Finally, we identify the pathways that are significant across multiple diseases by newly combining pathway-level results across diseases and performing Bayesian posterior probability analysis of null hypotheses for pathways in each disease as well as in the combined case (Methods).

Using our pipeline, we examined ASD and eleven of its most common comorbidities (Table 2.1). Differential analysis of transcriptomic data using Empirical Bayes method [118] from 53 microarray studies (Appendix A: Table A.1) related to the twelve disease conditions revealed different numbers of genes that show significant differential changes in expression per disease depending on different false discovery rate (FDR) corrections (Table 2.2; `https://tinyurl.com/GenePValues`). We selected the most informative FDR correction test by looking at the accuracy of the classification of cases vs. controls for each disease using the significant genes selected under different FDR corrections. We found the Benjamini-Yekutieli (BY) adjustment as the most informative and accurate—classification accuracy being at least 63% per disease using the genes selected under BY adjustment as features for support vector machine (SVM) classifier (Figure 2-2).

Hypergeometric enrichment analysis on individual pathway gene sets from Kyoto encyclopedia of genes and genomes (KEGG), BioCarta, Reactome, and pathway interaction database (PID) collections, as well as on the combined gene set of all canonical pathways helped us reveal the significantly dysregulated pathways in each of the diseases in question (`https://tinyurl.com/PathPValues`). Combining the p-values per

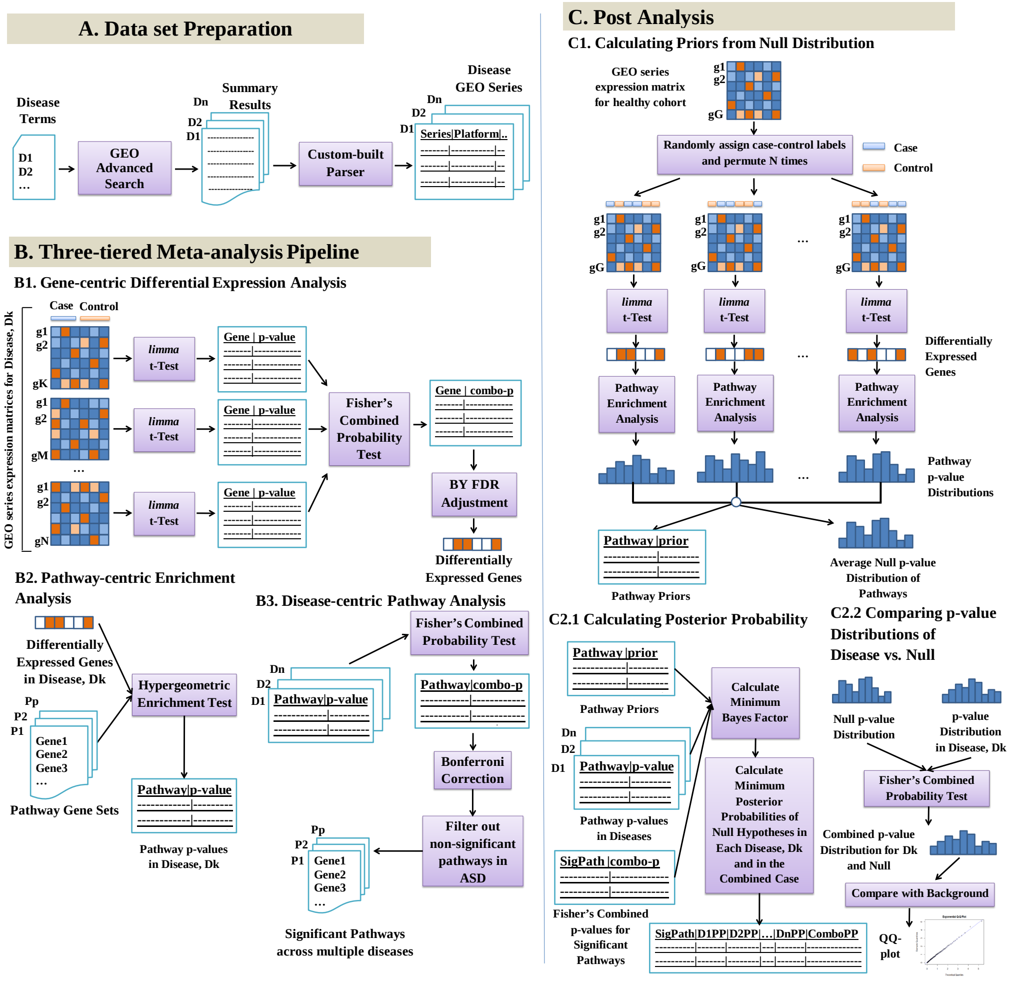

A. Data set Preparation
Disease Terms
D1
D2
…
GEO Advanced Search
Summary Results
Dn
D2
D1
Custom-built Parser
Disease GEO Series
Series|Platform|..
B. Three-tiered Meta-analysis Pipeline
B1. Gene-centric Differential Expression Analysis
GEO series expression matrices for Disease, Dk
Case
Control
g1
g2
gK
gM
gN
limma t-Test
Gene | p-value
Fisher's Combined Probability Test
Gene | combo-p
BY FDR Adjustment
Differentially Expressed Genes
B2. Pathway-centric Enrichment Analysis
Differentially Expressed Genes in Disease, Dk
Pp
P2
P1
Gene1
Gene2
Gene3
…
Pathway Gene Sets
Hypergeometric Enrichment Test
Pathway|p-value
Pathway p-values in Disease, Dk
B3. Disease-centric Pathway Analysis
Fisher's Combined Probability Test
Pathway|combo-p
Bonferroni Correction
Filter out non-significant pathways in ASD
Significant Pathways across multiple diseases
C. Post Analysis
C1. Calculating Priors from Null Distribution
GEO series expression matrix for healthy cohort
gG
Randomly assign case-control labels and permute N times
Case
Control
limma t-Test
Differentially Expressed Genes
Pathway Enrichment Analysis
Pathway p-value Distributions
Pathway|prior
Pathway Priors
Average Null p-value Distribution of Pathways
C2.1 Calculating Posterior Probability
Calculate Minimum Bayes Factor
Pathway p-values in Diseases
SigPath|combo-p
Fisher's Combined p-values for Significant Pathways
Calculate Minimum Posterior Probabilities of Null Hypotheses in Each Disease, Dk and in the Combined Case
SigPath|D1PP|D2PP|…|DnPP|ComboPP
C2.2 Comparing p-value Distributions of Disease vs. Null
Null p-value Distribution
p-value Distribution in Disease, Dk
Fisher's Combined Probability Test
Combined p-value Distribution for Dk and Null
Compare with Background
QQ-plot

Figure 2-1: **Three-tiered integrative omics pipeline.** (A) Data preparation: select GEO Series relevant to ASD and comorbid diseases. (B) Three tiers: 1. For each disease, select significant genes from differential expression analysis of GEO series with a Fisher's combined test with $p < 0.05$ after Benjamini-Yekutieli (BY) FDR adjustment. 2. For each disease, select significant pathways from hypergeometric enrichment analysis with $p < 0.05$. 3. Identify significant shared pathways across diseases using Fisher's combined test with $p < 0.05$ after Bonferroni FDR correction. Exclude the non-significant pathways in ASD. (C) Post analysis: 1. Using the gene expression data from a healthy cohort, generate a null distribution of pathway p-values and calculate prior probabilities of pathways being significant by chance. 2.1. Using the prior probabilities, pathway p- values in each disease, and the Fisher's combined p-values of significant pathways across diseases, calculate minimum Bayes factors and minimum posterior probabilities of null hypotheses for each significant pathway in each disease as well as in the combined case. 2.2. Combine the pathway p-value distribution of each disease with the average null distribution of p-values using Fisher's combined probability test and compare the combined p-value distribution with background chi-squared distribution using QQ-plot for significance. Identify the significant pathways, using the combined p-values, minimum posterior probabilities, and QQ-plots.

pathway across all the diseases using Fisher's combined probability test [119] and correcting for multiple comparisons using Bonferroni correction, we measured the shared significance of pathways across ASD and its comorbidities. Any pathway that had a Bonferroni corrected p-value < 0.05, was termed as 'significant' and pathways that were not significant in ASD, were filtered out.

To confirm that the presence of multiple significant pathways among ASD and its comorbidities was due to shared biology, we estimated minimum Bayes factors (BF) and minimum posterior probabilities of the null hypothesis for each of the significant KEGG pathways in ASD and its comorbidities. The priors for the pathways were estimated from the null distributions of p-values generated by differential expression analysis and pathway analysis performed on permutations of gene expression data of a healthy cohort (GEO Accession: GSE16028). For the significant pathways shared between the diseases, the posterior probabilities of the p-values being significant by chance were always less than 5% (Table 2.3; `https://tinyurl.com/BayesianPosteriorAnalysis`). The quantile-quantile (QQ) plot of combined p-values of pathways across ASD and its comorbidities show marked enrichment of significant p-values indicative of shared disease biology captured by the pathways tested (Figure 2-3(a)). The QQ-plots of hypergeometric p-values of pathways in ASD and its comorbid diseases against theoretical quantiles also show significant enrichment (Figure 2-4). For contrast, we combined pathway p-values from each disease separately with the null $p$value distribution. When the pathway p-value distribution in a disease is combined with null p-value distribution, the QQ-plots do not show much deviation from the background distribution (Figure 2-5), indicating both that there is a lack of shared biology (as expected) and that our analysis does not cause a systematic inflation.

Table 2.1: Comorbidities of Autism Spectrum Disorders

Disease Group	**Clinical Manifestations**	**References**

Multi-system Disorders (Congenital Anomalies, Auditory Disorders, Infections, Gastrointestinal Disorders, Cardiac Disorders etc.)	Asthma	Becker, 2007 [100];
		Doshi-Velez, Ge, and Kohane, 2014 [45];

	Bacterial & Viral Infections	Atladóttir *et al.*, 2010 [101];
		Atladóttir *et al.*, 2012 [102];
		Garbett *et al.*, 2012 [103];
		Hagberg, Gressens, and Mallard, 2012 [104];

	Chronic Kidney Disease	Curatolo *et al.*, 2004 [105]; Loirat *et al.*, 2010 [106]

	Cerebral Palsy	Surén *et al.*, 2012 [107];
		Doshi-Velez, Ge, and Kohane, 2014 [45];

	Dilated Cardiomyopathy	Witchel, Hancox, and Nutt, 2003 [108];
		Bilder *et al.*, 2013 [109];

	Ear Infection/Otitis Media	Konstantareas and Homatidis, 1987 [110];
		Rosenhall *et al.*, 1999 [111];
		Porges *et al.*, 2013 [112];

	Inflammatory Bowel Disease (Crohn's Disease, Ulcerative Colitis)	Horvath *et al.*, 1999 [51];
		Horvath and Perman, 2002 [52];
		Walker *et al.*, 2013 [113]

	Muscular Dystrophy	Wu *et al.*, 2005 [54];
		Hendriksen and Vles, 2008 [55];
		Hinton *et al.*, 2009 [56];
		Kohane *et al.*, 2012 [44];

	Upper Respiratory Infection	Shavelle, Strauss, and Pickett, 2001 [114];
		Porges *et al.*, 2013 [112];
		Bilder *et al.*, 2013 [109];

Seizures	Epilepsy	Mouridsen *et al.*, 1999 [49];
		Tuchman and Rapin, 2002 [50];
		Surén *et al.*, 2012 [107];
		Bilder *et al.*, 2013 [109];

Psychiatric Disorders	Schizophrenia	Morgan, Roy, and Chance, 2003 [57];
		Tabarés-Seisdedos and Rubenstein, 2009 [115];
		Ingason *et al.*, 2011 [116];
		Smoller *et al.*, 2013 [42];
		Murdoch and State, 2013 [117];

Table 2.2: Number of differentially expressed genes selected under different FDR corrections for ASD and its comorbidities with a significance cutoff of p-value < 0.05.

Disease	**Bonferroni**	**BY**	**BH**	**None**
ASD	157	1258	5104	9176
Asthma	238	852	2501	5555
Bacterial & Viral Infection	1613	3630	6016	8183
Chronic Kidney Disease	66	416	3771	12577
Cerebral Palsy	93	220	646	2352
Dilated Cardiomyopathy	146	349	908	3455
Ear Infection/Otitis Media	1629	3867	6708	6708
Epilepsy	5	4	12	2242
Inflammatory bowel Disease	831	2547	4771	6897
Muscular Dystrophy	207	517	1303	3885
Schizophrenia	54	149	508	2881
Upper Respiratory Infection	32	59	172	2664

* Here, BY = Benjamini-Yekutieli, BH = Benjamini-Hochberg, and None = No FDR correction.

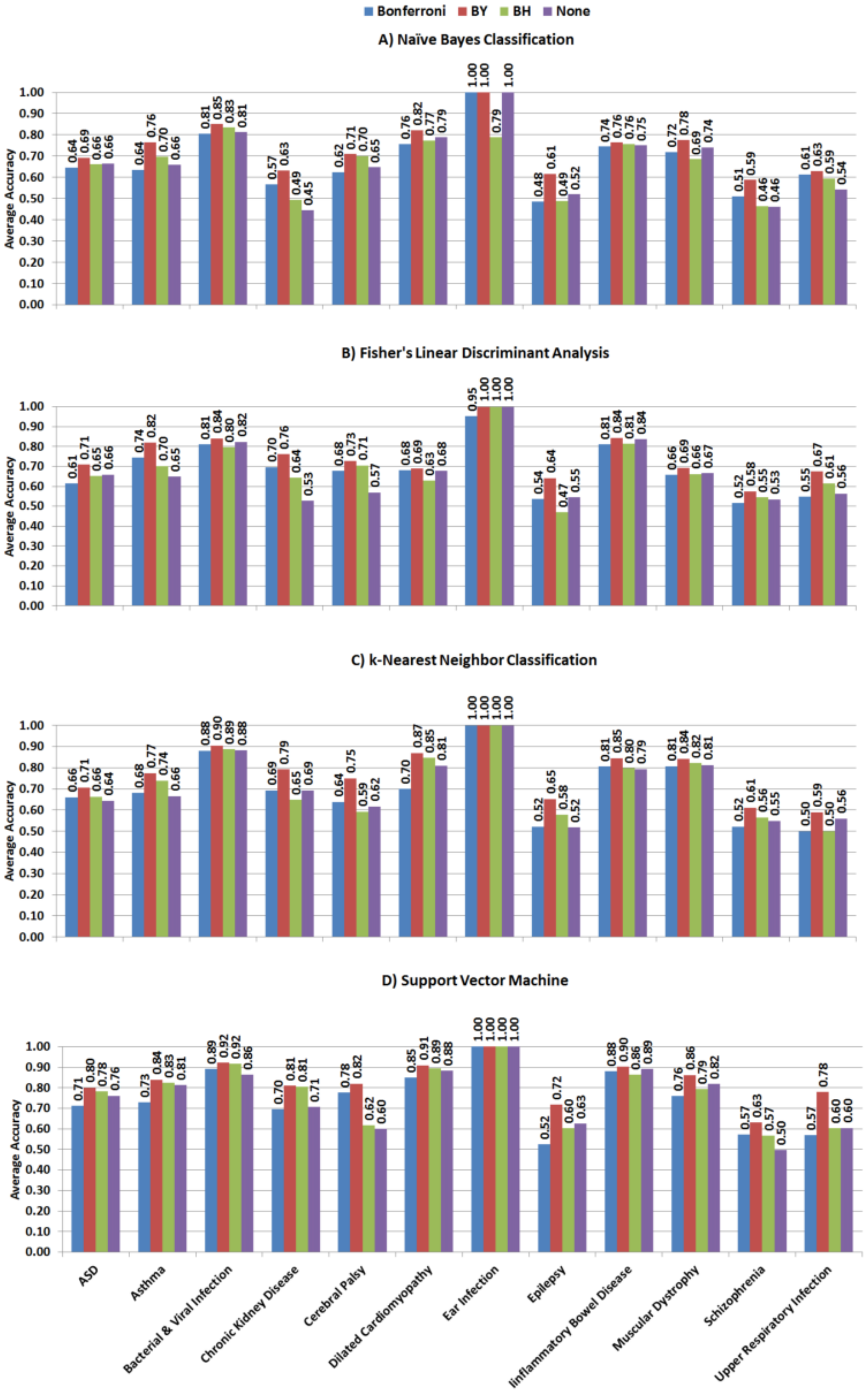
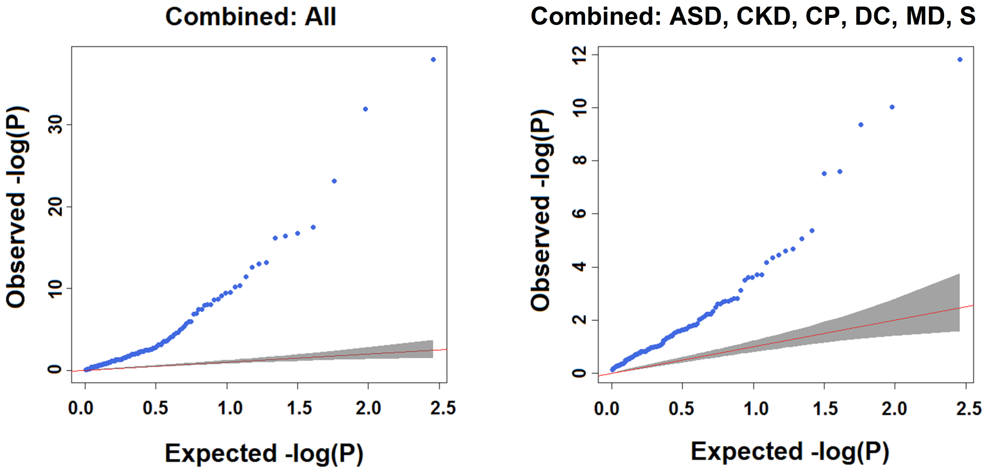


Figure 2-2: **Accuracy of different classification methods for case-control group classification in ASD and its comorbid diseases using genes selected under different false discovery rate (FDR) corrections as features.** (A) Naïve Bayes Classification, (B) Fisher's Linear Discriminant Analysis, (C) k-Nearest Neighbor Classification, and (D) Support Vector Machine.



(a) (b)

Figure 2-3: **Quantile-quantile plots showing Fisher's combined p-value distributions of KEGG pathways across (a) ASD and all its comorbidities, and (b) ASD and its non-immune related comorbidities.** Here, ASD = Autism Spectrum Disorder, CKD = Chronic Kidney Disease, CP = Cerebral Palsy, DC = Dilated Cardiomyopathy, MD = Muscular Dystrophy, and S = Schizophrenia.

Table 2.3: KEGG pathways significantly shared between ASD and its comorbid diseases.

Pathway	**ASD**	**Asthma**	**INF**	**CKD**	**CP**	**DC**	**EI**	**EP**	**IBD**	**MD**	**S**	**URI**	**Fisher's** p	**Bonferroni corrected** p	**Bayes factor**	**Minimum posterior of null**
Toll-like Receptor Signaling	0.0048	5.52E-06	0.0762	0.0114	0.6550	0.0034	4.28E-16	1	5.93E-05	0.0210	1	1.14E-10	1.1745E-32	1.703E-30	8.78E-31	9.76E-32
Chemokine Signaling	0.0145	0.0003	0.000051	0.2197	0.8628	0.0194	3.21E-10	1	1.37E-06	0.5703	1	8.89E-09	7.0449E-26	1.022E-21	4.01E-22	1.07E-22
NOD-like Receptor Signaling	0.0342	9.02E-05	0.0136	0.0019	0.4760	0.0019	1.99E-08	1	0.0036	0.7335	1	9.04E-05	1.7813E-17	2.583E-15	7.75E-16	3.23E-17
Ribosome	6.49E-13	0.9647	4.84E-10	0.1720	0.6006	1	0.9841	1	0.9460	0.0026	1	1	3.68E-17	5.336E-15	1.68E-15	3.69E-16
Spliceosome	6.70E-05	0.9541	6.39E-06	0.2965	0.3831	0.2746	0.9201	1	1.36E-05	0.5081	0.1721	1	9.9149E-09	1.438E-06	2.38E-07	5.23E-08
Leukocyte Trans-endothelial Migration	0.0023	0.8201	0.0110	0.0797	0.0002	0.8164	0.0974	1	0.1238	7.63E-06	0.5000	1	9.962E-09	1.445E-06	2.40E-07	6.76E-08
Regulation of Actin Cyto-skeleton	0.0234	0.9080	0.2734	0.1131	0.0745	0.0355	0.2280	1	0.2032	5.90E-05	0.1330	1	2.7324E-05	0.003962	0.0004	0.0007
Tight Junction	0.0359	0.5613	0.4111	0.1064	0.0005	0.8542	0.3039	1	0.1900	0.0006	1	1	6.9114E-05	0.010022	0.0010	0.0004

Note: Entries indicating significant p-values are colored in red. The entries with value '1' indicate the case where there was no overlap between the pathway and the disease gene set. Here, ASD = Autism Spectrum Disorder, INF = Bacterial & Viral Infection, CKD = Chronic Kidney Disease, CP = Cerebral Palsy, DC = Dilated Cardiomyopathy, EI = Ear Infection, EP = Epilepsy, IBD = Inflammatory Bowel Disease, MD = Muscular Dystrophy, S = Schizophrenia, URI = Upper Respiratory Infection.

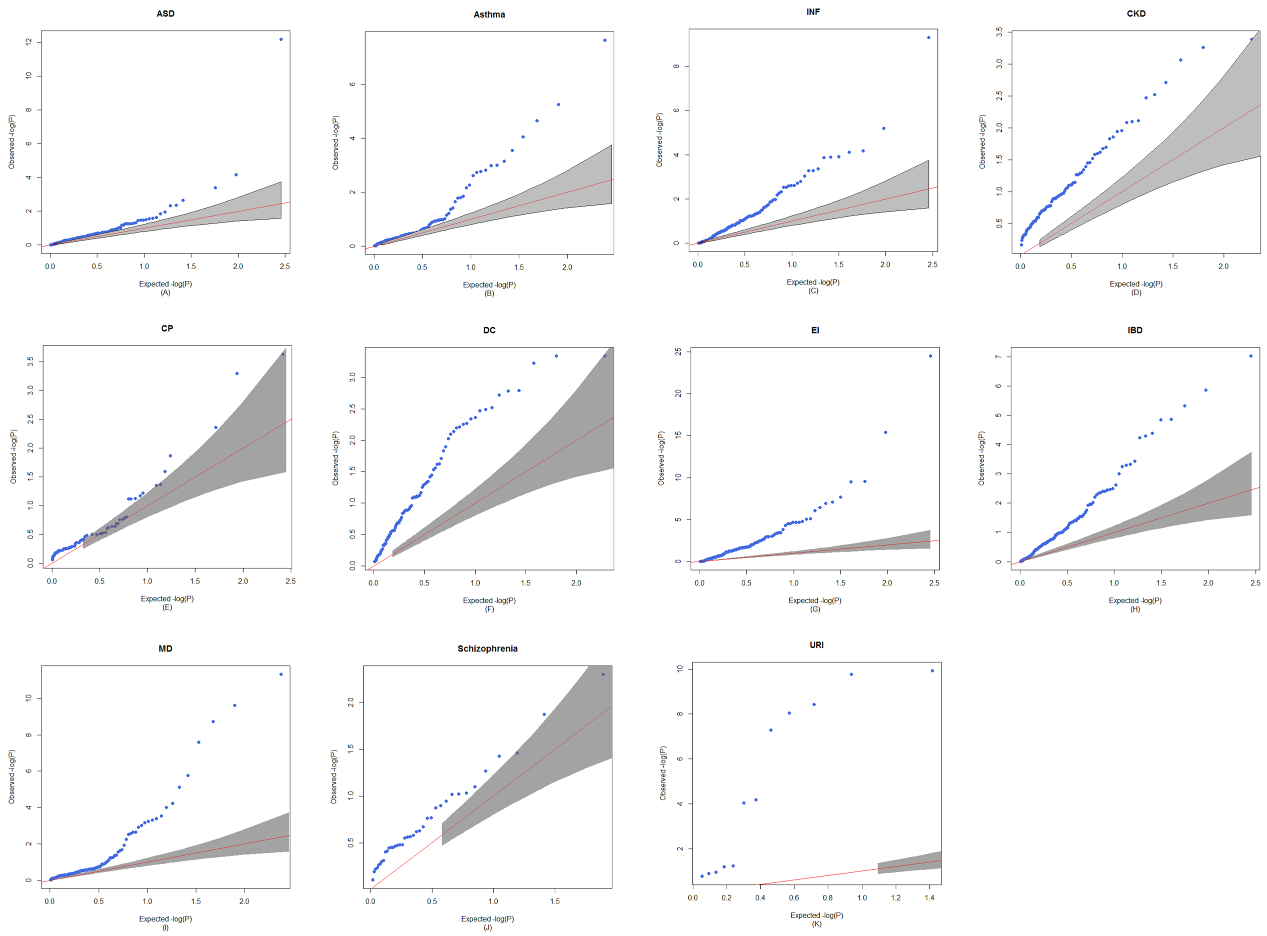


Figure 2-4: **Quantile-quantile plots comparing the distribution of *p*-values of pathways in each of ASD and its comorbidities with theoretical quantiles.** The plots are in log-scale. (A) ASD, (B) Asthma, (C) Bacterial and viral infection, (D) Chronic kidney disease, (E) Cerebral Palsy, (F) Dilated Cardiomyopathy, (G) Ear infection, (H) IBD, (I) Muscular Dystrophy, (J) Schizophrenia, and (K) Upper Respiratory Infection.

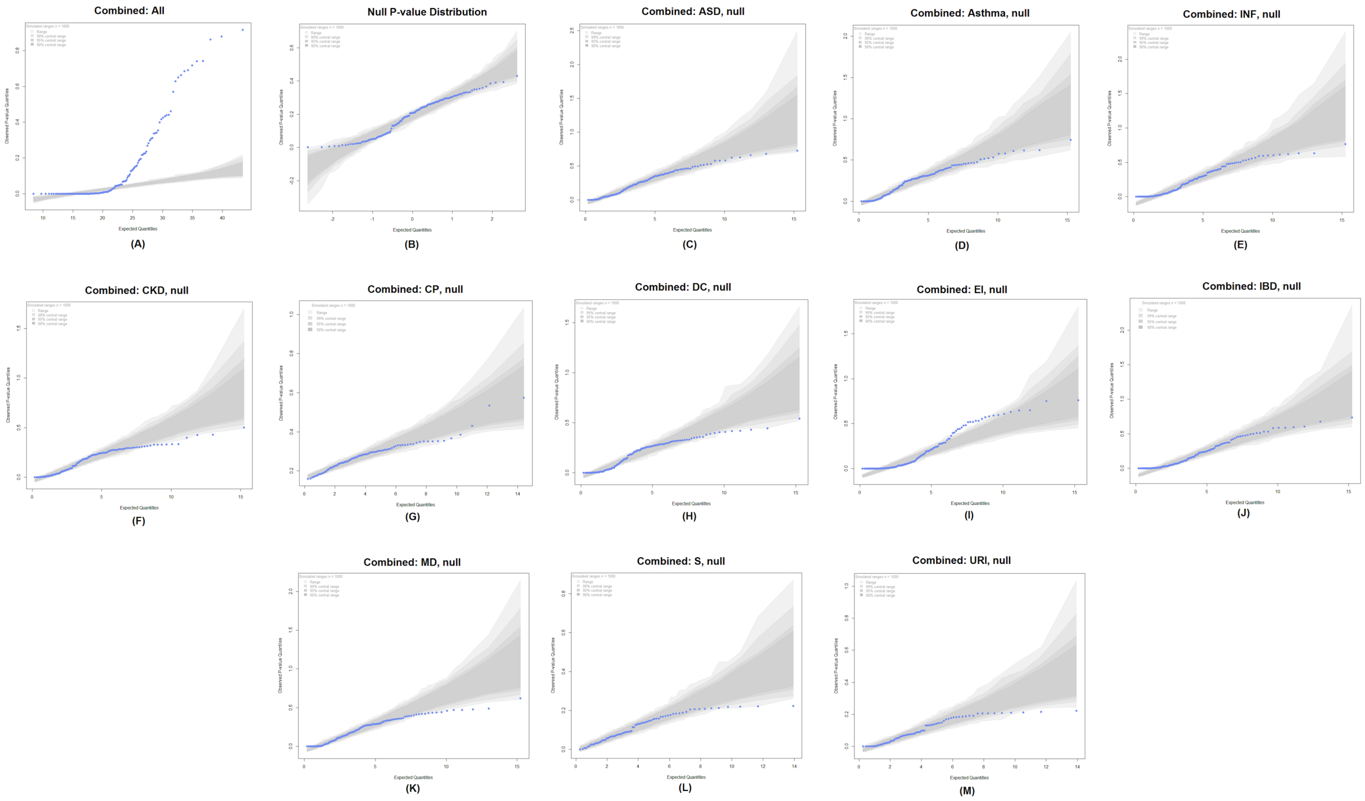


Figure 2-5: **Quantile-quantile plots showing the distribution of combined p-values from each disease with the simulated background p-value distribution of pathways.** The plots are in log-scale. The combined p-values are compared with the theoretical quantiles drawn from appropriate chi-square distributions, and the null distribution is compared with theoretical quantiles from the standard normal distribution. The expected region is colored in gray. (A) Combined p-values of pathways across all diseases (shown for comparison), (B) The simulated background p-value distribution of pathways which we call the “null” p-value distribution, (C) ASD and null, (D) Asthma and null, (E)Bacterial and viral infection and null, (F) Chronic kidney disease and null, (G) Cerebral palsy and null, (H) Dilated cardiomyopathy and null, (I) Ear infection and null, (J) IBD and null, (K) Muscular dystrophy and null, (L) Schizophrenia and null, and (M) Upper respiratory infection and null.

2.2.2 Involvement of innate immunity pathways in ASD and its comorbidities

Our results demonstrated that pathways that are dysregulated across ASD and its comorbidities with the highest statistical significance (i.e., the lowest Bonferroni-corrected combined *p*-value), are all related to innate immune system (Tables 2.3–2.7; full tables at `https://tinyurl.com/PathPValues`.). For the KEGG, BioCarta, and PID gene sets, the Toll-like receptor signaling pathway was found to be the most significant (Tables 2.3, 2.4 and 2.6). For the KEGG database, the top two significant pathways were Toll-like receptor signaling, and chemokine signaling (Table 2.3). The top three significant pathways, revealed from the analysis of Reactome data set, include chemokine receptor signaling, innate immunity, and Toll-like receptor signaling (Table 2.5). When we expanded our aperture of analysis to the gene sets from all canonical pathways, the Toll-like receptor signaling, and chemokine signaling pathways were still found to be the most significantly dysregulated in the disease conditions (Table 2.7). Thus, we primarily focused our attention on these two pathways in ASD and its comorbidities and then, for completeness, extended to other innate immunity KEGG pathways that were found significantly dysregulated (Table 2.3).

Both Toll-like receptor signaling and chemokine signaling pathways are vital pathways in the innate immune response mechanism. Toll-like receptors are the most common pattern recognition receptors that recognize distinct pathogen-associated molecular patterns and participate in the first line of defense against invading pathogens. They also play a significant role in inflammation, immune cell regulation, survival, and proliferation. Toll-like receptors activate various signal transduction pathways which in turn activates expression and synthesis of chemokines which together with cytokines, cell adhesion molecules, and immunoreceptors, orchestrate the early host response to infection and at the same time represent an essential link to the adaptive immune response [120]. Our study revealed that, the KEGG Toll-like receptor signaling pathway, by itself, was significantly dysregulated (with a combined *p*-value of 1.7e-30 after Bonferroni correction) in ASD, asthma, chronic kidney disease, dilated cardiomyopathy, ear infection, inflammatory bowel disease, muscular dystrophy, and upper respiratory infection with the minimum posterior probability of appearing significant by chance being at most 1%. In addition, the KEGG chemokine signaling pathway was found significantly dysregulated (with a combined p-value of 1.02e-21 after Bonferroni correction) in ASD, asthma, bacterial and viral infection, dilated cardiomyopathy, ear infection, inflammatory bowel disease, and upper respiratory infection with the minimum posterior probability of appearing significant by chance being at most 2.4% in each case. These findings indicate the role of immune dysfunction in this wide range of seemingly unconnected disease conditions. Although, there has been some experimental evidence linking abnormal chemokine response to Toll-like receptor ligands associated with autism [121, 122], no study so far linked them to the comorbidities suffered by ASD affected individuals.

When we looked at the other significant KEGG pathways, we found two others involved in innate immunity, namely, NOD-like receptor signaling and leukocyte transendothelial migration pathways. The NOD-like receptor signaling pathway, by itself, was significantly dysregulated (with a combined p-value of 2.6e-15 after Bonferroni correction and a minimum posterior probability of null hypothesis at most 4%) in ASD, asthma, bacterial and viral infection, chronic kidney disease, dilated cardiomyopathy, ear infection, inflammatory bowel disease, and upper respiratory infection. The leukocyte transendothelial migration pathway was significantly dysregulated (with a combined p-value of 1.4e-6 after Bonferroni correction and a minimum posterior probability of null hypothesis at most 1.7%) in ASD, asthma, cerebral palsy, and muscular dystrophy. Some NOD-like receptors recognize certain types of bacterial fragments; others induce caspase-1 activation through the assembly of multi-protein complexes called inflammasomes, which are critical for generating mature pro-inflammatory cytokines in concert with Toll-like receptor signaling pathway. While Toll-like receptor, chemokine, and NOD-like receptor signaling pathways have more to do with the recognition of infectious pathogens and initiating a response, the leukocyte transendothelial migration pathway orchestrates the migration of leukocytes from blood into tissues via a process called diapedesis, which is vital for immune surveillance and inflammation. During this diapedesis of leukocytes, the leukocytes bind to endothelial cell adhesion molecules (CAM) and then migrate across the vascular endothelium to the site of infection. Notably, increased permeability of the blood-brain barrier (BBB) favoring leukocyte migration into the brain tissue has been implicated in ASD before [123], but not as a shared genetic commonality among its comorbidities.

To confirm that, the presence of multiple significant innate immunity-related pathways among ASD and its comorbidities was due to shared biology, we repeated the combined *p*-value analysis excluding the immune system-related diseases – bacterial and viral infection, asthma, inflammatory bowel disease, upper respiratory infection, and ear infection. Innate immunity pathways – Leukocyte transendothelial migration, Toll-like receptor signaling, and NOD-like receptor signaling pathways still appeared among the most significant, dysregulated pathways shared by ASD, cerebral palsy, chronic kidney disease, and muscular dystrophy. The QQ-plot of combined *p*-values of pathways across ASD and its non-immune related comorbidities show marked enrichment of significant *p*-values indicative of shared disease biology of these conditions (Figure 2-3(b)). Table 2.8 shows the most significant KEGG pathways that are shared by ASD and its non-immune related comorbidities. For other pathway gene set collections, the complete lists of Fisher's combined *p*-values per pathway per disease are available at `https://tinyurl.com/NonImmunePathPValues`.

Besides the immune-related ones, Tables 2.3 and 2.8 document several other pathways and gene sets including the ribosome and spliceosome gene sets which have roles in genetic information processing and translation and the actin cytoskeleton regulation pathway which controls various cellular processes like cell motility. Neuronal signal processing and neuron motility have often been associated with ASD [124]. Thus these findings are not surprising. The genes in the tight junction pathway mediate cell adhesion and are thought to constitute the intra-membrane and paracellular diffusion barriers [125]. These findings implicate the involvement of these cellular processes in the shared pathology of ASD and its comorbidities.

Table 2.4: BioCarta pathways significantly shared among ASD and its comorbidities.

Pathway	**ASD**	**Asthma**	**INF**	**CKD**	**CP**	**DC**	**EI**	**IBD**	**MD**	**S**	**URI**	**Fisher's combined** p	**Bonferroni corrected** p
Toll Pathway	0.0094	0.0477	0.0021	0.0038	0.0561	0.0018	6.30E-05	0.0071	0.0426	1	1	4.57E-12	9.77E-10
IL10 Pathway	0.0499	0.0237	0.0007	1	1	1	2.15E-05	0.0090	0.3039	1	0.0003	7.37E-11	1.58E-08
NO2IL12 Pathway	0.0499	0.0237	0.0001	1	1	1	0.0049	0.1208	1	1	1	1.12E-06	0.0002
NKT Pathway	0.0027	0.0216	0.0030	0.2192	0.2607	0.3402	0.0076	0.4268	0.4610	1	0.0395	1.21E-05	0.0026
Caspase Pathway	0.0049	0.2105	0.0252	1	1	1	0.0369	0.0029	1	1	1	2.14E-05	0.0046

* The entries with value '1' indicate the case where there was no overlap between the pathway and the disease gene set. Here, ASD = Autism Spectrum Disorder, INF = Bacterial & Viral Infection, CKD = Chronic Kidney Disease, CP = Cerebral Palsy, DC = Dilated Cardiomyopathy, EI = Ear Infection, EP = Epilepsy, IBD = Inflammatory Bowel Disease, MD = Muscular Dystrophy, S = Schizophrenia, URI = Upper Respiratory Infection.

Table 2.5: Top five Reactome pathways significantly shared among ASD and its comorbidities.

Pathway	**ASD**	**Asthma**	**INF**	**CKD**	**CP**	**DC**	**EI**	**EP**	**IBD**	**MD**	**S**	**URI**	**Fisher's combined** p	**Bonferroni corrected** p
Chemokine receptors bind chemokines	0.0070	0.0012	0.2160	0.0847	0.4479	0.0476	2.21E-07	1	1.10E-07	1	1	1.88E-10	3.41E-22	2.23E-19
Innate immune system	0.0069	0.0003	0.0002	0.0103	0.3276	0.0069	0.0133	1	0.0024	2.97E-05	1	2.54E-06	1.6227E-18	1.06E-15
Toll receptor cascades	0.0060	0.0001	0.0028	0.0185	0.3459	0.0015	3.54E-05	1	0.0071	0.0126	1	0.1512	1.09E-13	7.16E-11
Activated TLR4 signaling	0.0065	0.0024	0.0020	0.0451	0.2505	0.0004	0.0004	1	0.0038	0.0469	1	0.1212	5.59E-12	3.66E-09
Packaging of telomere ends	0.0093	0.8385	1.04E-06	0.0079	0.0883	0.1490	0.0674	1	2.56E-06	0.0803	1	-1	6.86E-12	4.49E-09

[*] The entries with value '1' indicate the case where there was no overlap between the pathway and the disease gene set. Here, ASD = Autism Spectrum Disorder, INF = Bacterial & Viral Infection, CKD = Chronic Kidney Disease, CP = Cerebral Palsy, DC = Dilated Cardiomyopathy, EI = Ear Infection, EP = Epilepsy, IBD = Inflammatory Bowel Disease, MD = Muscular Dystrophy, S = Schizophrenia, URI = Upper Respiratory Infection.

Table 2.6: PID pathways significantly shared among ASD and its comorbidities.

Pathway	**ASD**	**Asthma**	**INF**	**CKD**	**DC**	**EI**	**EP**	**IBD**	**MD**	**S**	**URI**	**Fisher's combined** p	**Bonferroni corrected** p
Toll Endogenous Pathway	1.89E-05	0.0129	2.76E-05	1	0.0274	2.38E-05	0.0005	7.54E-07	0.0968	0.00932	1	1.93E-20	3.79E-18
Integrin1 Pathway	0.0166	0.9185	0.9479	0.4308	0.0311	0.0679	7.43E-05	0.0157	8.14E-11	0.0571	1	5.514E-13	1.084E-10
P53 Downstream Pathway	0.0196	0.0018	0.1358	2.96E-08	0.4158	0.0138	0.1611	0.0016	0.1638	0.0465	1	7.14E-12	1.40E-09
GMCSF Pathway	0.0355	0.0023	0.0006	0.2708	0.3199	0.0155	0.0028	0.1228	0.0426	1	0.0501	2.55E-08	5.00E-06
FCER1 Pathway	0.0116	0.0275	0.0136	0.0019	0.4760	0.0584	0.4085	0.0213	0.0417	1	1	1.91E-06	0.0004
Anthrax Pathway	0.0499	1	0.9106	1	0.1622	0.2163	0.0002	0.0370	0.0489	1	1	8.73E-05	0.0171

[*] The entries with value '1' indicate the case where there was no overlap between the pathway and the disease gene set. Here, ASD = Autism Spectrum Disorder, INF = Bacterial & Viral Infection, CKD = Chronic Kidney Disease, DC = Dilated Cardiomyopathy, EI = Ear Infection, EP = Epilepsy, IBD = Inflammatory Bowel Disease, MD = Muscular Dystrophy, S = Schizophrenia, URI = Upper Respiratory Infection.

Table 2.7: Top 10 MSigDB canonical pathways significantly shared among ASD and its comorbidities.

Pathway	**ASD**	**Asthma**	**INF**	**CKD**	**CP**	**DC**	**EI**	**EP**	**IBD**	**MD**	**S**	**URI**	**Fisher's combined** p	**Bonferroni corrected** p
KEGG Toll like receptor signaling	0.0048	5.52E-06	0.0762	0.0114	0.6550	0.0034	1.96E-08	1	5.93E-05	0.0210	1	1.14E-10	1.00E-25	1.29E-22
Reactome chemokine receptor binds chemokines	0.0070	0.0012	0.2160	0.0847	0.4479	0.0476	2.21E-07	1	1.10E-07	1	1	1.88E-10	3.42E-22	4.40E-19
PID Toll endogenous pathway	1.89E-05	0.0129	2.76E-05	1	0.0274	2.38E-05	0.0005	1	7.54E-07	0.0968	0.0093	1	1.93E-20	2.49E-17
Reactome innate immune system	0.0069	0.0003	0.0002	0.0103	0.3276	0.0069	0.0133	1	0.0025	2.97E-05	1	2.54E-06	1.62E-18	2.09E-15
KEGG chemokine signaling	0.0145	0.0003	0.0005	0.2197	0.8628	0.0194	0.0017	1	1.37E-06	0.5703	1	8.89E-09	5.47E-18	7.06E-15
KEGG ribosome	6.49E-13	0.9647	4.84E-10	0.1720	0.6006	1	1.0000	1	0.9460	0.0026	1	1	3.73E-17	4.81E-14
KEGG NOD-like receptor signaling	0.0341	9.02E-05	0.0136	0.0019	0.4760	0.0019	0.0002	1	0.0036	0.7335	1	9.04E-05	4.54E-14	5.86E-11
Reactome Toll receptor cascades	0.0060	0.0001	0.0028	0.0185	0.3459	0.0015	3.54E-05	1	0.0071	0.0126	1	0.1512	1.09E-13	1.41E-10
PID Integrin1 pathway	0.0166	0.9185	0.9479	0.4308	0.0311	0.0679	7.43E-05	1	0.0157	8.14E-11	0.0571	1	5.51E-13	7.10E-10
Biocarta Toll pathway	0.0094	0.0477	0.0021	0.0038	0.0561	0.0018	6.30E-05	1	0.0071	0.0426	1	1	4.57E-12	5.89E-09

[*] The entries with value '1' indicate the case where there was no overlap between the pathway and the disease gene set. Here, ASD = Autism Spectrum Disorder, INF = Bacterial & Viral Infection, CKD = Chronic Kidney Disease, CP = Cerebral Palsy, DC = Dilated Cardiomyopathy, EI = Ear Infection, EP = Epilepsy, IBD = Inflammatory Bowel Disease, MD = Muscular Dystrophy, S = Schizophrenia, URI = Upper Respiratory Infection.

Table 2.8: KEGG pathways significantly shared among ASD and its non-immune-related comorbidities.

Pathway	**ASD**	**CKD**	**CP**	**DC**	**EP**	**MD**	**S**	**Fisher's combined *p*-value**	**Bonferroni corrected *p*-value**
Ribosome	6.49E-13	0.1720	0.6006	1.0000	1.0000	0.0026	1.0000	1.50E-12	2.18E-10
Leukocyte Transendothelial Migration	0.0023	0.0797	0.0002	0.8164	1.0000	7.63E-06	0.5000	2.97E-08	4.30E-06
Regulation of Actin Cytoskeleton	0.0234	0.1132	0.0745	0.0355	1.0000	5.90E-05	0.1330	4.22E-06	6.12E-04
Tight Junction	0.0359	0.1064	0.0005	0.8542	1.0000	0.0006	1.0000	8.77E-06	1.27E-03
Toll-like Receptor Signaling	0.0048	0.0114	0.6550	0.0034	1.0000	0.0210	1.0000	2.03E-05	2.94E-03
NOD-like Receptor Signaling	0.0342	0.0019	0.4760	0.0019	1.0000	0.7335	1.0000	1.93E-04	2.79E-02
Oxidative Phosphorylation	0.0004	0.6844	0.7558	1.0000	1.0000	0.0023	1.0000	3.06E-04	4.44E-02

[*] Entries indicating significant p-values are colored in red. The entries with value '1' indicate the case where there was no overlap between the pathway and the disease gene set. Here, ASD = Autism Spectrum Disorder, CKD = Chronic Kidney Disease, CP = Cerebral Palsy, DC = Dilated Cardiomyopathy, EP = Epilepsy, MD = Muscular Dystrophy, S = Schizophrenia.

2.2.3 Disease–innate immunity pathway overlap at gene level

To examine the shared innate immunity KEGG pathways through a finer lens, we examined the genes that overlapped with them (Table 2.9). Although these pathways have broad involvement in a variety of diseases, a small number of genes in these pathways appear dysregulated most often in ASD and its comorbidities. Thus, we gave a closer look at the genes that are shared by ASD and at least one of its comorbid conditions.

In the Toll-like receptor signaling pathway, as shown in Figure 2-6(a), commonly shared, differentially expressed genes included CD14 and LY96 (also known as MD-2), responsible for mediating the lipopolysaccharide response which itself has been shown to create an autism-like phenotype in murine model systems [126], but has never been linked to the shared biology of ASD, cerebral palsy, dilated cardiomyopathy, muscular dystrophy, IBD. The widely-expressed Toll-like receptors, primarily, TLR1, TLR2, and TLR9 mediate recognition of foreign substances, including infectious pathogens and regulation of subsequent cytokine production required for the immune response. Although these genes have been known to be involved in immunity-related conditions, they have not been implicated in the co-occurrence of such conditions in ASD patients. Other genes involved were: CCL4, also known as Macrophage inflammatory protein 1β (MIP-1β) which is the most upregulated chemokine in natural killer cells of children with autism [127]; MAPK21, a gene upstream of the MAP-kinases that mediates multiple intracellular and extracellular signals; JUN (a subunit of transcription factor, AP-1) that regulates gene expression in response to a variety of stimuli, including cytokines, growth factors, stress, and bacterial and viral infections; SPP1 (also known as OPN), a cytokine that upregulates expression of interferon-γ (IFN-γ) which itself has been implicated in ASD and other diseases characterized by social dysfunction [128]; and TBK1, a gene that can mediate NFκB activation in response to specific growth factors and is often considered as a therapeutic target for inflammatory diseases.

In the chemokine pathway, as shown in Figure 2-6(b), the commonly shared genes include the chemokines (e.g., CCL4 which had altered expression levels in asthma and ear infection) and MAP-kinases (e.g., MAP2K1 which had altered expression levels in ASD, dilated cardiomyopathy, ear infection, and muscular dystrophy). The HCK gene, which belongs to the Src family of tyrosine kinases, showed altered expression levels in ASD, asthma, inflammatory bowel dis-ease, ear infection, bacterial and viral infection, and muscular dystrophy. Considering HCK's role in microglia and macrophages in controlling proliferation and cell survival [129], this finding is not surprising. The JAK2 which is dysregulated in ASD and its multiple immune-related comorbidities regulates STAT3 activity which in turn transduces IL-6 signals and increased Interleukin-6 (IL-6) in the maternal serum has been known in altering fetal brain development impairing social behaviors in the offspring [130, 131]. The alpha and beta subunits of G-proteins, dysregulated in ASD, asthma, IBD, and bacterial and viral infections, are essential signaling molecules which are often considered to have weak links to several brain conditions. The RAP1B gene, a member of the RAS family, regulates multiple cellular processes including cell adhesion, growth and differentiation, and integrin-mediated cell signaling. This protein also plays a role in regulating outside-in signaling in platelets and G protein-coupled receptor signaling. Thus, it can be of importance.

In the NOD-like receptor signaling pathway, the genes NOD1 and NOD2 drive the activation of NFκB and MAPK, production of cytokines, and apoptosis. The BIRC2 and BIRC3 genes (which had altered expressions in ASD, asthma, ear infection, and bacterial and viral infections) are members of the inhibitor-of-apoptosis protein family and are key regulators of NOD1 and NOD2 innate immunity signaling. In the leukocyte transendothelial migration pathway, the TXK gene which is a non-receptor tyrosine kinase (with altered expression in ASD, ear infection, IBD, and bacterial and viral infections), specifically regulates interferon-γ gene transcription and the development, function, and differentiation of conventional T-cells and nonconventional NKT-cells. Mutation of TXK gene has been identified to be a segregating factor for many neurodevelopmental disorders, including ASD, bipolar disorder, and intellectual disabilities [132].

Table 2.9: Differentially expressed genes in ASD and its comorbidities overlapping with innate immunity pathways.

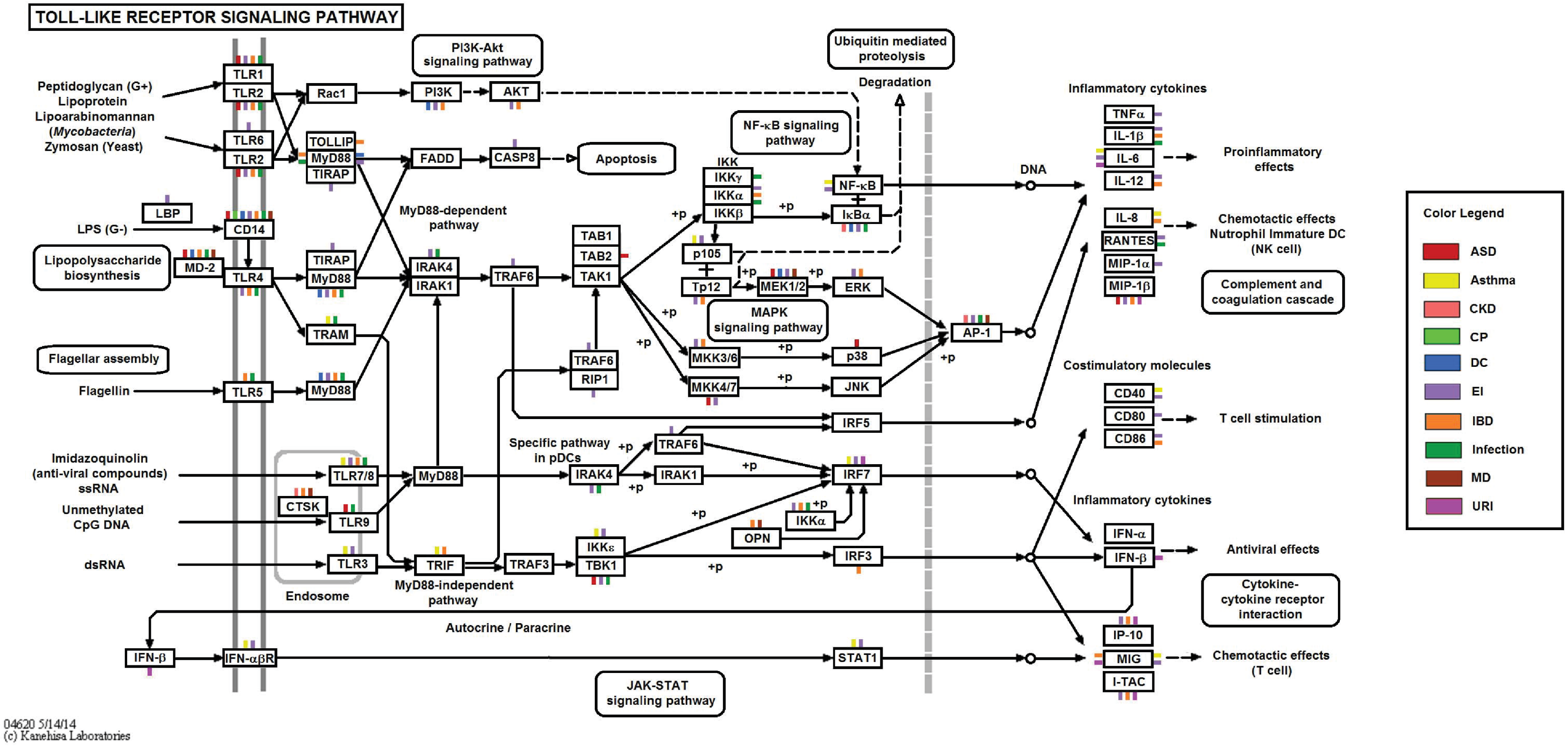
	Toll-like receptor signaling	**Chemokine signaling**	**NOD-like receptor signaling**	**Leukocyte transendothelial migration**
ASD	TLR9, MAP2K4, CCL4, LY96, CD14, TAB2, MAP2K2, MAPK13, MAP2K1, TBK1, TLR1, TLR2	CCL4, JAK2, GRK7, CCL17, CCL21, CCL22, GNB3, GNAI2, CCR2, CXCR3, GNAI3, CCR10, ADCY6, PREX1, HCK, MAP2K1, RAP1B	BIRC3, MAPK13, SUGT1, PSTPIP1, PYCARD, TAB2, BIRC2	TXK, NCF2, JAM2, GNAI2, GNAI3, CLDN23, ACTN3, ICAM1, ACTN1, MAPK13, CD99, RAP1B, CLDN14, MSN
Asthma	STAT1, IKBKE, NFKB1, RELA, TLR7, TICAM1, IL8, IFNAR1, IFNAR2, TICAM2, CD40, CXCL9, TLR3, IL6, IRF7	STAT1, CCL2, GNB4, JAK2, CCL20, NFKB1, RELA, CXCL5, XCR1, PLCB1, CXCL1, PRKCD, HCK, IL8, CCL1, CXCL2, CXCL9, LYN	CXCL1, RIPK2, BIRC3, CCL2, IL8, CASP5, NFKB1, RELA, CXCL2, IL6	TXK, ACTN2, ICAM1
CKD	JUN, CTSK, NFKBIA, FOS	CCL17, NFKBIA, CXCR6	NFKBIA, HSP90AA1, NLRC4, HSP90AB1	CLDN16, ACTN4, CLDN9
CP	CD14	CCL2	CCL2	JAM3, MMP2, VCAM1, ACTN4, ACTG1, MSN, CTNNA3
DC	MYD88, LY96, CD14, NFKBIA, MAP2K1, PIK3R1	CCL2, CCL11, CCL8, NFKBIA, MAP2K1, CCR1, PIK3R1	RIPK2, CCL2, CCL11, CCL8, NFKBIA	PIK3R1
EI	JUN, CD86, STAT1, CCL3, MYD88, CCL5, CCL4, LBP, TLR6, MAP3K8, CD14, IKBKE, NFKB1, NFKBIA, PIK3R5, TLR5, ITIRAP, RELA, TOLLIP, CXCL11, TLR7, TLR8, CXCL10, CASP8, TNF, IL12B, MAP2K3, MAP2K1, MAP2K6, MAPK3, IL1B, CD40, TBK1, CXCL9, TLR3, TLR4, TLR1, FOS, TLR2, IL6, IRF7, PIK3R2	STAT3, STAT1, STAT2, CCL2, CCL3, CCL5, CCL4, CX3CR1, CCL11, CCL7, CXCL14, JAK3, JAK2, CCL17, CCL20, CCL19, CCL22, NFKB1, NFKBIB, NFKBIA, PIK3R5, RELA, GNG7, FGR, GNG11, GNGT2, XCL1, CXCL5, ADCY3, CXCL11, ADCY2, CXCL10, CXCL1, PLCB3, CXCR5, CXCR2, GNG8, HCK, MAP2K1, CCR5, CCR7, MAPK3, CXCL16, CCR1, CXCL13, CXCL2, CXCL3, CXCL9, LYN, PIK3R2	CASP8, CXCL1, RIPK2, TNF, BIRC3, CCL2, CCL5, CCL11, CCL7, MEFV, CASP1, TNFAIP3, MAPK3, NFKB1, NFKBIB, NFKBIA, IL18, RELA, IL1B, NLRP3, BIRC2, CXCL2, IL6	TXK, NCF4, VCAM1, PIK3R5, CLDN23, CLDN10, CLDN8, MYL9, CLDN5, ICAM1, ACTN4, CLDN19, CLDN22, RASSF5, CLDN7, CLDN4, PIK3R2
EP	–	–	–	–

(continued on next page)

Table 2.9: *cont.* Differentially expressed genes in ASD and its comorbidities overlapping with innate immunity pathways.

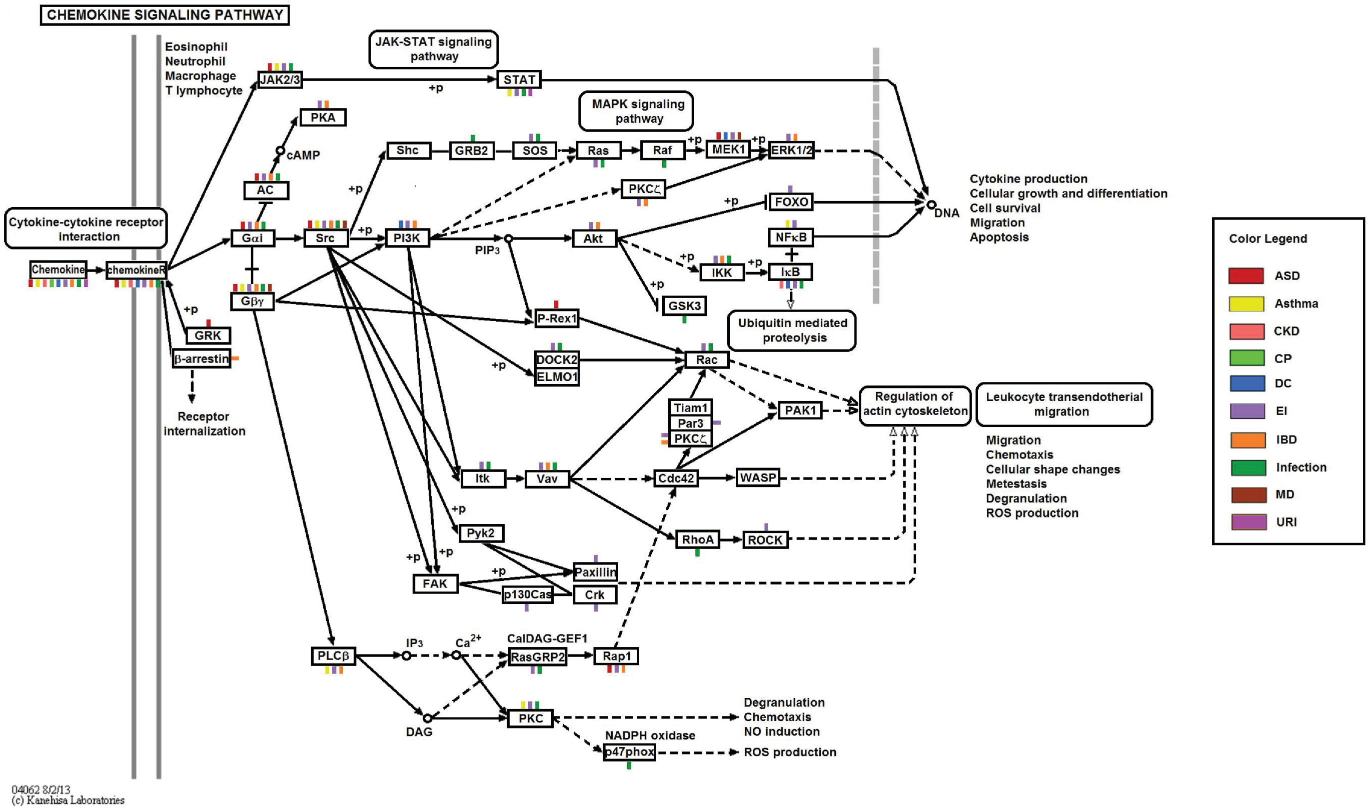
	Toll-like receptor signaling	**Chemokine signaling**	**NOD-like receptor signaling**	**Leukocyte transendothelial migration**
IBD	CD86, MYD88, CCL4, LY96, MAP3K8, AKT1, CD14, CTSK, SPP1, TOLLIP, CXCL11, TLR8, CXCL10, TICAM1, CHUK, IL8, MAP2K3, IL12A, MAPK3, IRF3, IL1B, PIK3CA, CXCL9, TLR4, TLR1, TLR2	CCL2, CCL4, CCL11, CCL7, AKT1, CCL18, ARRB2, CCL20, GNG5, GNB3, GNB2, CCL24, PRKX, GNG10, GNAI2, GNG11, XCL1, CXCL11, CXCL6, CCR10, CXCL10, ADCY6, CHUK, CXCL1, PLCB3, CXCR2, CXCR1, HCK, IL8, ADCY4, PRKCZ, MAPK3, CCR1, XCL2, CXCL13, RAP1A, PF4, CXCL2, CXCL3, PF4V1, PIK3CA, CXCL9, PPBP, VAV3, LYN	CHUK, CXCL1, BIRC3, CCL2, CARD6, CCL11, CCL7, IL8, CASP5, CASP1, MAPK3, IL1B, NLRP1, CXCL2, HSP90AB1	TXK, ITGA4, NCF4, MMP9, NCF2, MYL12A, THY1, GNAI2, MYL5, RHOH, CLDN8, MYL9, CLDN15, MYL12B, RAP1A, PIK3CA, MSN, VAV3
INF	TLR9, MYD88, CCL5, LY96, CD14, NFKBIA, TLR5, TLR8, CHUK, IRAK4, MAP2K7, IRF3, IKBKG, TICAM2, IL1B, TBK1, TLR4, TLR1, FOS, TLR2, IRF7	STAT3, STAT2, CCL5, DOCK2, JAK2, VAV2, NRAS, NFKBIA, GNG7, GNG5, GNB2, GNG10, GNG11, ADCY3, CXCR3, CCL4L1, GNAI3, GNB1, SOS2, CHUK, RAF1, RHOA, CXCR2, CXCR1, PRKCD, HCK, RAC2, RASGRP2, ADCY4, CCR3, CCR4, CXCR6, CCR7, GRB2, IKBKG, HRAS, GSK3B, CCR1, ITK, NCF1, PPBP, LYN	CHUK, RIPK2, CCL5, NOD1, CARD6, CCL8, CARD8, CASP5, NOD2, NFKBIA, IKBKG, PYCARD, NLRC4, IL1B, BIRC2	TXK, ITGAM, NCF4, MMP9, NCF2, VAV1, VASP, MYL12A, VAV2, ITGB2, CTNNA1, GNAI3, EZR, PLCG1, RHOH, PRKCA, ESAM, RAC2, CD99, ITK, NCF1, CYBA, CYBB, MYL5, RHOA
MD	LY96, CD14, CTSK, SPP1, MAP2K1, FOS	GNG10, HCK, MAP2K1, GNG12	PYCARD	JAM3, MMP2, NCF2, VCAM1, ITGB2, JAM2, MYL5, CD99, ACTG1, MYL12B, MSN, CYBA
S	–	–	–	CD99
URI	CCL4, CXCL11, CXCL10, IFNB1, CXCL9, IL6, IRF7	STAT2, CCL2, CCL4, CCL7, CXCL11, CXCL10, CXCL9	CCL2, CCL7, IL6	–

* Here, ASD = Autism Spectrum Disorder, CKD = Chronic Kidney Disease, CP = Cerebral Palsy, DC = Dilated Cardiomyopathy, EI = Ear Infection, EP = Epilepsy, IBD = Inflammatory Bowel Disease, Infection = Bacterial & Viral Infection, MD = Muscular Dystrophy, S = Schizophrenia, and URI = Upper Respiratory Infection.



(a)

Figure 2-6: **Innate immunity pathways color-tagged by comorbidity findings. (a) Toll-like receptor signaling.** Genes were mapped onto the pathway using the "user data mapping tool" from KEGG [133, 134]. Genes are represented by rectangular boxes on KEGG pathways. We put color tags on a gene to indicate in which diseases it is differentially expressed. Sometimes a set of genes are mapped onto a single box. In that case, the color tags on that box represent the union set of all diseases in which those genes are differentially expressed. Here, ASD = Autism Spectrum Disorder, CKD = Chronic Kidney Disease, CP = Cerebral Palsy, DC = Dilated Cardiomyopathy, EI = Ear Infection, IBD = Inflammatory Bowel Disease, Infection = Bacterial & Viral Infection, MD = Muscular Dystrophy, and URI = Upper Respiratory Infection.



(b)

Figure 2-6: **Innate immunity pathways color-tagged by comorbidity findings. (b) Chemokine signaling.** Genes were mapped onto the pathway using the "user data mapping tool" from KEGG [133, 134]. Genes are represented by rectangular boxes on KEGG pathways. We put color tags on a gene to indicate in which diseases it is differentially expressed. Sometimes a set of genes are mapped onto a single box. In that case, the color tags on that box represent the union set of all diseases in which those genes are differentially expressed. Here, ASD = Autism Spectrum Disorder, CKD = Chronic Kidney Disease, CP = Cerebral Palsy, DC = Dilated Cardiomyopathy, EI = Ear Infection, IBD = Inflammatory Bowel Disease, Infection = Bacterial & Viral Infection, MD = Muscular Dystrophy, and URI = Upper Respiratory Infection.

2.2.4 Discriminatory power of innate immunity pathway genes

The innate immunity pathway genes that overlapped the chemokine signaling and Toll-like receptor signaling pathways could accurately discriminate cases vs. controls to an extent in three-fold cross-validation SVM classification. We could achieve an average accuracy of at least 70% (Figure 2-7). For the baseline, we performed the same classification using the same number of randomly selected genes which do not overlap with these pathways. With randomly selected genes, the classification accuracy was much lower. This result suggests that the genes which have altered expressions in the diseases examined and are present in these innate immunity pathways were sufficient to distinguish the disease states from the controls partially. When we included the overlapped genes in NOD-like receptor signaling and transendothelial migration pathways in this analysis, the classification accuracy was at least 65% (Figure 2-8) which was still better than the randomly selected non-immune genes case. A recent functional genomic study showed that immune/inflammation-related genes could provide reasonable accuracy in the diagnostic classification of male infants and toddlers with ASD [135].

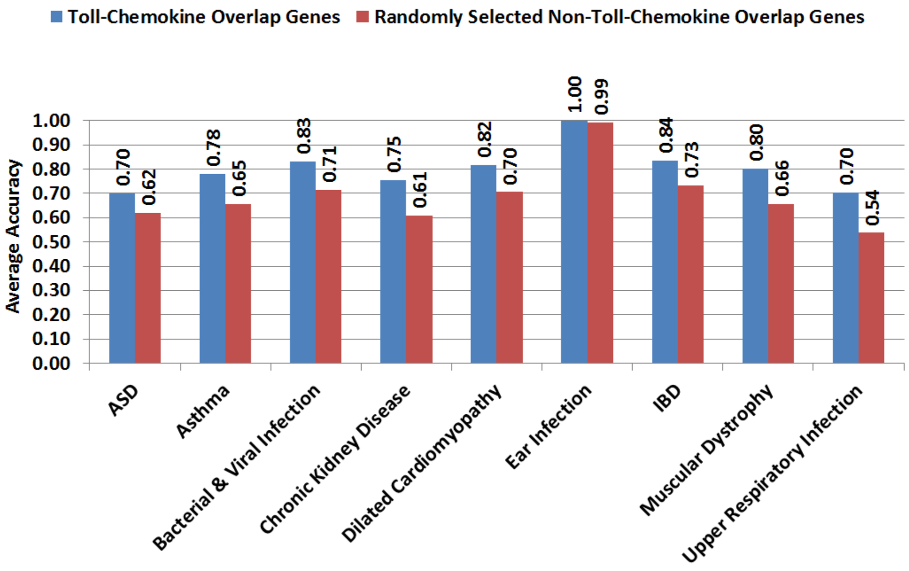


Figure 2-7: Classification of cases vs. controls in different disease cohorts using overlapped genes in the Toll-like receptor and chemokine signaling pathways from KEGG. Classification performance using randomly selected disease genes that do not overlap in the innate immunity pathways was used as a baseline.

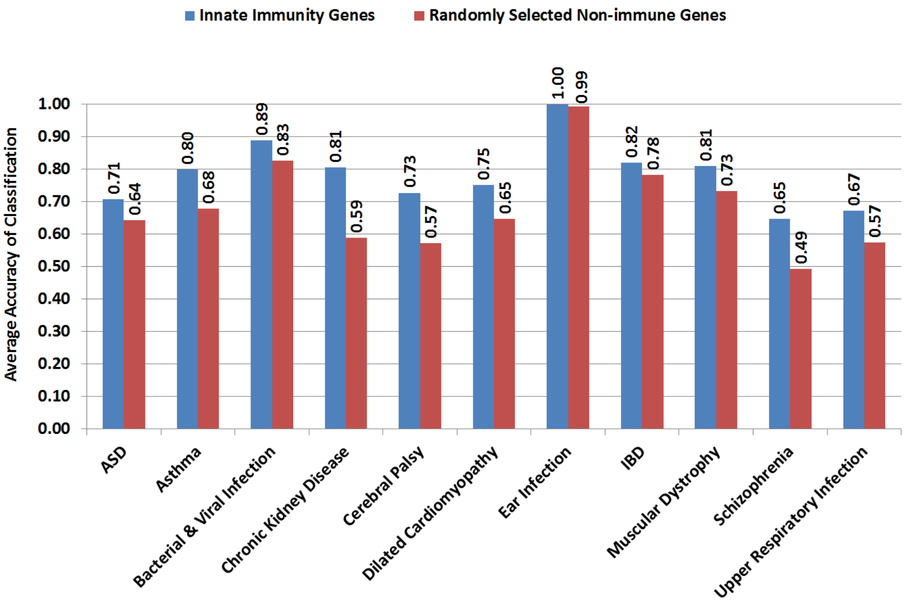


Figure 2-8: Classification of cases vs. controls in different disease cohorts using overlapped genes in the innate immunity pathways from KEGG. Classification performance using randomly selected disease genes that do not overlap in the innate immunity pathways was used as a baseline.

2.3 Methods

Here, we describe the steps of our three-tiered meta-analysis pipeline.

2.3.1 Gene-centric transcriptomic analysis per disease

Using the GEOquery package [136] from Bioconductor in R, we downloaded the gene expression data for each disease in gene matrix transposed (GMT) format from the gene expression omnibus (GEO). We removed 'NA' values from the data and log-normalized the expression values for subsequent analysis. Then, we performed differential expression analysis on each dataset using an Empirical Bayes model [118]. The model was implemented using the limma package [118] from Bioconductor in R, and p-value for each gene in each study was obtained from limma t-test.

To determine the degree of correlation between the differential expression analyses p-values of datasets selected under each disease, we calculated the pairwise Pearson correlation coefficient of p-values. Considering Pearson correlation coefficient of at least 0.30 with $p < 0.05$ as significant, we found that the p-values are not significantly correlated (Figure 2-9). This lack of correlation allowed us to use Fisher's combined probability test to calculate combined p-values for the genes in each disease condition. We used Fisher's combined probability test as follows:

$$P \sim \chi^2 = -2 \sum_{i=1}^{k} ln(p_i)$$

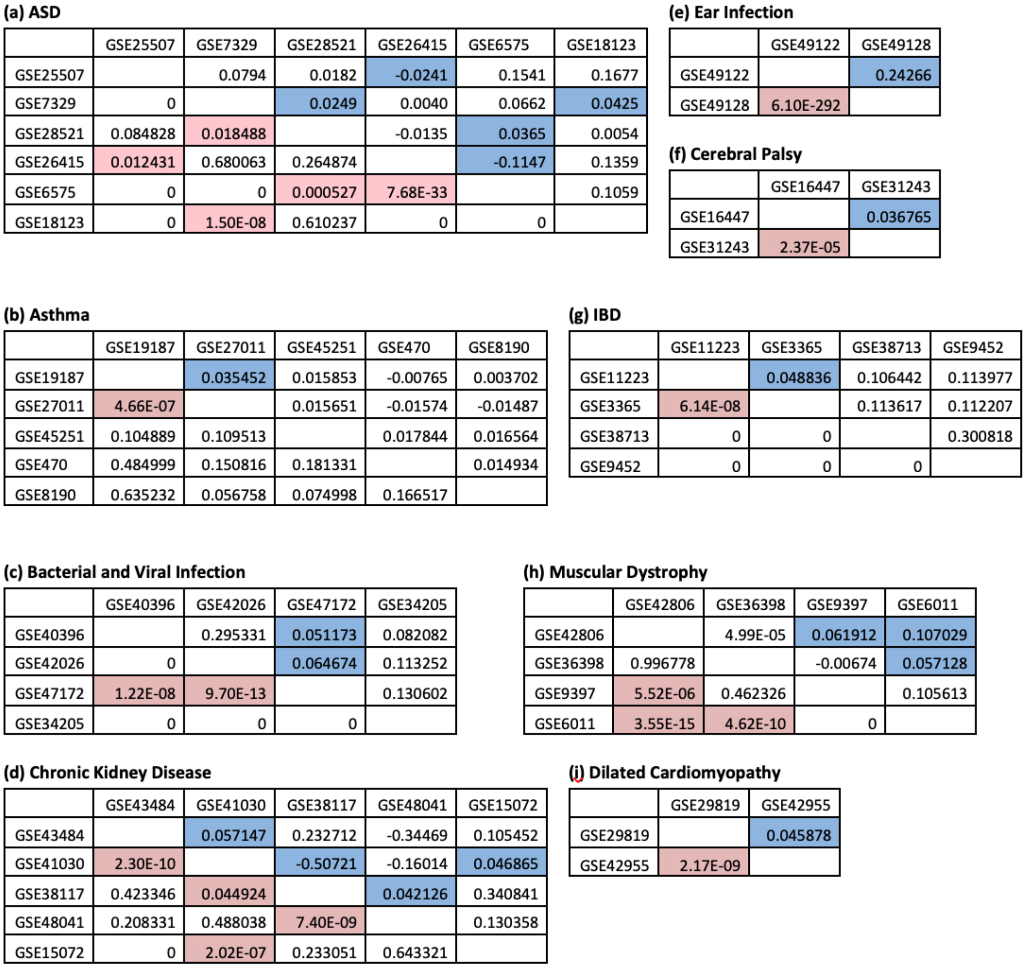
p_i is the p-value of test i; χ^2 is the Chi-squared distribution; k is the number of tests; and P is the adjusted p-value ($p < 0.05$ was considered significant).

Selection of FDR correction test for multiple comparisons

To adjust the combined p-values, we considered different FDR corrections (i.e., Bonferroni, Benjamini-Yekutieli (BY), and Benjamini-Hochberg (BH)). We also considered the 'no correction' case for completeness. We selected the most informative one, based on the level of accuracy we could achieve in classifying cases of a particular disease, versus controls, using the genes selected under a specific test with a significance cutoff of $p < 0.05$. We tested the accuracy of case-control classification for each of the 53 disease datasets using four different classification methods, namely, naïve Bayes method (NB), Fisher's linear discriminant analysis (FLDA), k nearest neighbor (KNN), and support vector machine (SVM) (Appendix A). The set of significant genes selected under different FDR corrections were considered as features of the classification methods. We performed three-fold cross-validation and calculated the average accuracy. We selected the FDR correction test that produced the best average accuracy in each disease (Figure 2-2).

2.3.2 Pathway-centric enrichment analysis per disease

From the disease-level gene-centric expression analysis, we obtained a list of significant genes per disease. For each disease, we then performed a hypergeometric enrichment



(a) ASD

	GSE25507	GSE7329	GSE28521	GSE26415	GSE6575	GSE18123
GSE25507		0.0794	0.0182	-0.0241	0.1541	0.1677
GSE7329	0		0.0249	0.0040	0.0662	0.0425
GSE28521	0.084828	0.018488		-0.0135	0.0365	0.0054
GSE26415	0.012431	0.680063	0.264874		-0.1147	0.1359
GSE6575	0	0	0.000527	7.68E-33		0.1059
GSE18123	0	1.50E-08	0.610237	0	0	

(b) Asthma

	GSE19187	GSE27011	GSE45251	GSE470	GSE8190
GSE19187		0.035452	0.015853	-0.00765	0.003702
GSE27011	4.66E-07		0.015651	-0.01574	-0.01487
GSE45251	0.104889	0.109513		0.017844	0.016564
GSE470	0.484999	0.150816	0.181331		0.014934
GSE8190	0.635232	0.056758	0.074998	0.166517	

(c) Bacterial and Viral Infection

	GSE40396	GSE42026	GSE47172	GSE34205
GSE40396		0.295331	0.051173	0.082082
GSE42026	0		0.064674	0.113252
GSE47172	1.22E-08	9.70E-13		0.130602
GSE34205	0	0	0	

(d) Chronic Kidney Disease

	GSE43484	GSE41030	GSE38117	GSE48041	GSE15072
GSE43484		0.057147	0.232712	-0.34469	0.105452
GSE41030	2.30E-10		-0.50721	-0.16014	0.046865
GSE38117	0.423346	0.044924		0.042126	0.340841
GSE48041	0.208331	0.488038	7.40E-09		0.130358
GSE15072	0	2.02E-07	0.233051	0.643321	

(e) Ear Infection

	GSE49122	GSE49128
GSE49122		0.24266
GSE49128	6.10E-292	

(f) Cerebral Palsy

	GSE16447	GSE31243
GSE16447		0.036765
GSE31243	2.37E-05	

(g) IBD

	GSE11223	GSE3365	GSE38713	GSE9452
GSE11223		0.048836	0.106442	0.113977
GSE3365	6.14E-08		0.113617	0.112207
GSE38713	0	0		0.300818
GSE9452	0	0	0	

(h) Muscular Dystrophy

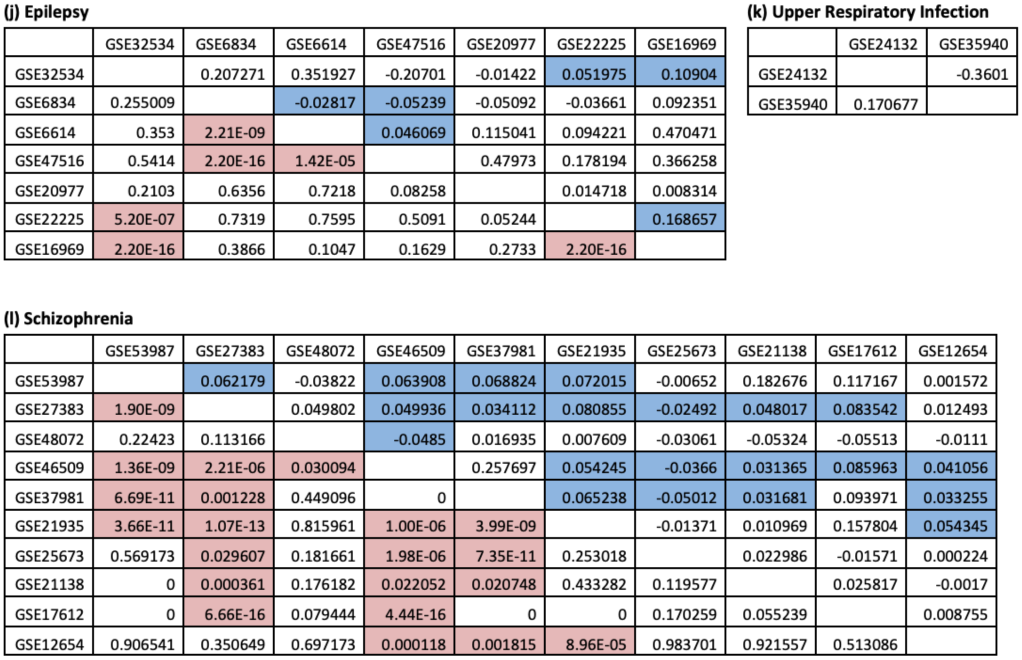
	GSE42806	GSE36398	GSE9397	GSE6011
GSE42806		4.99E-05	0.061912	0.107029
GSE36398	0.996778		-0.00674	0.057128
GSE9397	5.52E-06	0.462326		0.105613
GSE6011	3.55E-15	4.62E-10	0	

(i) Dilated Cardiomyopathy

	GSE29819	GSE42955
GSE29819		0.045878
GSE42955	2.17E-09	

Figure 2-9: **Independence of the p-values of genes across selected GEO series for ASD and its comorbidities.** Pairwise Pearson correlation of p-value distributions of differentially expressed genes from the selected GEO series was determined. The upper-right triangle shows the pairwise correlation coefficients, and the lower-left triangle shows the corresponding significance p-values in each matrix. Two distributions were considered independent if the pairwise correlation coefficient was < 0.30 or the significance p-value was > 0.05. The cells with significant p-values are marked pink, and the cells with corresponding correlations are marked in blue. No blue cell contains a value ≥ 0.30 satisfying the desired independence assumption of gene p-value distributions. (a)-(i) ASD, Asthma, Bacterial & Viral Infection, Chronic Kidney Disease, Ear Infection, Cerebral Palsy, IBD, Muscular Dystrophy, Dilated Cardiomyopathy

test for each pathway. This test uses the hypergeometric distribution to calculate the statistical significance of k or more significant disease genes, out of n total genes, appearing in a specific pathway gene set. It helps identify whether or not the specific



(j) Epilepsy

	GSE32534	GSE6834	GSE6614	GSE47516	GSE20977	GSE22225	GSE16969
GSE32534		0.207271	0.351927	-0.20701	-0.01422	0.051975	0.10904
GSE6834	0.255009		-0.02817	-0.05239	-0.05092	-0.03661	0.092351
GSE6614	0.353	2.21E-09		0.046069	0.115041	0.094221	0.470471
GSE47516	0.5414	2.20E-16	1.42E-05		0.47973	0.178194	0.366258
GSE20977	0.2103	0.6356	0.7218	0.08258		0.014718	0.008314
GSE22225	5.20E-07	0.7319	0.7595	0.5091	0.05244		0.168657
GSE16969	2.20E-16	0.3866	0.1047	0.1629	0.2733	2.20E-16	

(k) Upper Respiratory Infection

	GSE24132	GSE35940
GSE24132		-0.3601
GSE35940	0.170677	

(l) Schizophrenia

	GSE53987	GSE27383	GSE48072	GSE46509	GSE37981	GSE21935	GSE25673	GSE21138	GSE17612	GSE12654
GSE53987		0.062179	-0.03822	0.063908	0.068824	0.072015	-0.00652	0.182676	0.117167	0.001572
GSE27383	1.90E-09		0.049802	0.049936	0.034112	0.080855	-0.02492	0.048017	0.083542	0.012493
GSE48072	0.22423	0.113166		-0.0485	0.016935	0.007609	-0.03061	-0.05324	-0.05513	-0.0111
GSE46509	1.36E-09	2.21E-06	0.030094		0.257697	0.054245	-0.0366	0.031365	0.085963	0.041056
GSE37981	6.69E-11	0.001228	0.449096	0		0.065238	-0.05012	0.031681	0.093971	0.033255
GSE21935	3.66E-11	1.07E-13	0.815961	1.00E-06	3.99E-09		-0.01371	0.010969	0.157804	0.054345
GSE25673	0.569173	0.029607	0.181661	1.98E-06	7.35E-11	0.253018		0.022986	-0.01571	0.000224
GSE21138	0	0.000361	0.176182	0.022052	0.020748	0.433282	0.119577		0.025817	-0.0017
GSE17612	0	6.66E-16	0.079444	4.44E-16	0	0	0.170259	0.055239		0.008755
GSE12654	0.906541	0.350649	0.697173	0.000118	0.001815	8.96E-05	0.983701	0.921557	0.513086	

Figure 2-9: *cont.* **Independence of the p-values of genes across selected GEO series for ASD and its comorbidities.** Pairwise Pearson correlation of p-value distributions of differentially expressed genes from the selected GEO series was determined. The upper-right triangle shows the pairwise correlation coefficients, and the lower-left triangle shows the corresponding significance p-values in each matrix. Two distributions were considered independent if the pairwise correlation coefficient was < 0.30 or the significance p-value was > 0.05. The cells with significant p-values are marked pink, and the cells with corresponding correlations are marked in blue. Two distributions were considered independent if the pairwise correlation coefficient was < 0.30 or the significance p-value of the Pearson correlation was > 0.05. No blue cell contains a value ≥ 0.30 satisfying the desired independence assumption of gene p-value distributions. (j)-(l) Epilepsy, Upper Respiratory Infection, Schizophrenia

disease gene set is over-represented in a particular pathway, by providing a p-value per pathway per disease.

2.3.3 Across-disease shared significance analysis of pathways

Once we obtained the p-values for the pathways per disease, first we calculated the pairwise Pearson correlation of pathway p-values across diseases (Table 2.10). Since the distributions were not significantly correlated (Pearson correlation coefficient $<$ 0.30 with p-value < 0.05), we safely assumed the distributions to be independent.

Next, we calculated the combined p-value of each pathway across all the diseases using Fisher's combined probability test. We corrected for multiple comparisons using Bonferroni correction. We defined a significance threshold of adjusted p-value < 0.05 and called any pathway that passed this threshold, 'significant.' We restricted our results to the pathways that appeared significant in ASD.

Table 2.10: **Independence of pathway p-values in ASD and its comorbidities.** Pairwise Pearson correlation (with significance) of hypergeometric p-values of KEGG pathways in ASD and its comorbidities, as well as the null data set, were determined. The p-value distributions of two diseases were considered independent of each other if their pairwise correlation coefficient was < 0.30 or the significance p-value of Pearson correlation was > 0.05.

	ASD	Asthma	INF	CKD	CP	DC	EI	IBD	MD	S	URI	NULL
ASD		0.0042	1.0000	0.0829	0.0453	0.0770	0.0053	9.70E-02	4.99E-03	0.037704	0.006864	0.009253
Asthma	0.9506		0.2090	0.0261	0.0120	0.0234	0.2512	1.79E-01	1.00E-01	0.153183	0.138483	0.169088
INF	0.0000	0.0090		0.0885	0.2234	0.0424	0.2392	2.04E-01	4.99E-03	0.037704	0.006864	0.118191
CKD	0.1604	0.7890	0.2931		0.2825	0.1034	0.2785	2.46E-01	7.63E-02	0.33307	0.069892	0.059191
CP	0.4435	0.9090	0.0207	0.0101		0.1473	0.1490	8.04E-02	9.92E-02	0.177432	0.021374	0.231012
DC	0.1928	0.7996	0.6115	0.2915	0.1812		0.1502	3.42E-02	8.14E-02	0.315967	0.056339	0.094819
EI	0.9292	0.0109	0.0041	0.0008	0.2252	0.1421		1.18E-01	3.75E-03	0.028345	0.032568	0.250611
IBD	0.1003	0.0261	0.0027	0.0029	0.4106	0.6824	0.1618		6.98E-02	0.081573	0.097878	0.056727
MD	0.9328	0.0895	0.9328	0.1965	0.0929	0.1681	0.9495	2.38E-01		0.158418	0.004895	0.211801
S	0.5239	0.0092	0.5239	0.0000	0.0025	0.0000	0.9495	1.67E-01	7.07E-03		0.03701	0.297515
URI	0.9077	0.0187	0.9077	0.2371	0.7180	0.3407	0.5820	9.74E-02	9.34E-01	0.531602		0.067065
NULL	0.9165	0.0649	0.1643	0.5730	0.0580	0.3633	0.0027	5.07E-01	2.25E-02	0.000323	0.427778	

[1] The upper-right triangle of each table contains the Pearson correlation coefficients, and the lower-left triangle contains the corresponding p-values. None of the cells in the upper triangle has a correlation coefficient greater than or equal to 0.3 for which the corresponding $p < 0.05$; thus, the independence assumption for data sets is satisfied. Here, ASD = Autism Spectrum Disorder, CKD = Chronic Kidney Disease, CP = Cerebral Palsy, DC = Dilated Cardiomyopathy, EI = Ear Infection, IBD = Inflammatory Bowel Disease, Infection = Bacterial & Viral Infection, MD = Muscular Dystrophy, and URI = Upper Respiratory Infection.

2.3.4 Bayesian posterior analysis of null hypothesis

Following Goodman's approach [137], we derive the formula for estimating a lower bound on the posterior probability of the null hypothesis as follows. Let,

H_0: null hypothesis

H_1: alternative hypothesis

D: observed data

Following Bayes theorem, we can express the posterior probability of the null hypothesis as follows.

$$P(H_0|D) = \frac{P(D|H_0) \times P(H_0)}{P(D)} \tag{2.1}$$

where,

$P(H_0|D)$: posterior probability of the null hypothesis being true given the observed data

$P(D|H_0)$: probability of observing the data when the null hypothesis is true ($\approx$ p-value from the frequentist approach)

$P(H_0)$: prior probability of the null hypothesis

$P(D)$: marginal probability of observing the data

Similarly, the posterior probability of the alternative hypothesis is given by,

$$P(H_1|D) = \frac{P(D|H_1) \times P(H_1)}{P(D)} \tag{2.2}$$

where,

$P(H_1|D)$: posterior probability of the alternative hypothesis being true given the observed data

$P(D|H_1)$: probability of observing the data when the alternative hypothesis is true

$P(H_1)$: prior probability of the alternative hypothesis

Dividing Equation (2.1) by Equation (2.2), we get

$$\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(D|H_0)}{P(D|H_1)} \times \frac{P(H_0)}{P(H_1)} \tag{2.3}$$

By definition,

$$P(H_1|D) = 1 - P(H_0|D) \tag{2.4}$$

$$P(H_1) = 1 - P(H_0) \tag{2.5}$$

Substituting Equations (2.4) and (2.5) into Equation (2.3), we get

$$\begin{aligned}
&\frac{P(H_0|D)}{1-P(H_0|D)} = \frac{P(D|H_0)}{P(D|H_1)} \times \frac{P(H_0)}{1-P(H_0)} \\
\implies &\frac{P(H_0|D)}{1-P(H_0|D)} = BF \times \frac{q}{1-q} \\
\implies &\frac{1-P(H_0|D)}{P(H_0|D)} = \left(\frac{BF \times q}{1-q}\right)^{-1} \\
\implies &\frac{1}{P(H_0|D)} - 1 = \left(\frac{BF \times q}{1-q}\right)^{-1}
\end{aligned}$$

$$\therefore P(H_0|D) = \left(1 + \left(\frac{BF \times q}{1-q}\right)^{-1}\right)^{-1} \tag{2.6}$$

Here, Bayes Factor, $BF = \frac{P(D|H_0)}{P(D|H_1)}$ and $P(H_0) = q$.

To estimate the prior probability of pathways, we selected a publicly available GEO study of 109 gene expression profiles of blood drawn from healthy individuals enrolled at a single site (GEO Accession: GSE16028). We assigned case-control labels randomly to the samples and performed differential expression analysis using R package limma. We selected differentially expressed genes using uncorrected p-values (< 0.05) because after BY correction, none of the genes remained significant. On the list of significant genes, we performed hypergeometric enrichment analysis to obtain pathway p-value distribution. We repeated this process for one hundred times to obtain hundred null p-value distributions. We calculated the prior for each pathway by looking at how many times the pathway appeared significant (p-value < 0.05) during these hundred runs. We took an average of the hundred distributions to obtain the null p-value distribution.

In case of pathway p-values, the null hypothesis is that p-values are uniformly distributed, and the alternative hypothesis is smaller p-values are more likely than larger p values. Following the approach of Sellke, Bayarri, and Berger [138], we estimated the minimum Bayes factors using the following formula:

$$BF = \begin{cases} -ep\log(p), \text{if } p < \frac{1}{e} \\ 1, \text{otherwise} \end{cases}$$

where e is Euler's constant.

For calculating minimum Bayes factors for χ^2-distributed test statistics we used Johnson's formula [139]:

$$BF = \begin{cases} (\frac{v}{x})^{-\frac{v}{2}} \exp(-\frac{x-v}{2}), \text{for } x > v \\ 1, \text{otherwise} \end{cases}$$

where x is the chi-square statistic which gave rise to the observed p-value and v is the degrees of freedom.

We used the prior probability distribution drawn from the simulated background data set and the minimum Bayes factor to estimate the minimum posterior probability of the null hypothesis for the pathways. The simulated null p-value distributions and the priors for all KEGG pathways, the minimum Bayes factors, and the minimum posterior probabilities of null hypotheses are available at `https://tinyurl.com/BayesianPosteriorAnalysis`.

2.3.5 Data set selection

Gene expression data sets

We selected eleven disease conditions which co-occur most commonly in ASD patients. Each of these diseases has at least 5% prevalence in ASD patients [45]. The prevalence of a comorbid condition can be defined in two ways – (i) percentage of ASD patients having a comorbid disease, and (ii) percentage of patients of a comorbid disease having ASD [44]. The diseases which satisfy either of these criteria include asthma, bacterial and viral infection, cerebral palsy, chronic kidney disease, dilated cardiomyopathy, ear infection/otitis media, epilepsy, IBD, muscular dystrophy, schizophrenia, and upper respiratory infection. Table 2.1 shows the disease groups along with the literature references.

To identify publicly available studies relevant to these comorbidities, we performed an extensive literature search on the gene expression omnibus (GEO) of the National Center for Biotechnology Information (NCBI) [140, 141]. Using the advanced search tool provided by GEO, we searched series data sets from studies that performed expression profiling by microarray on either human or mouse. The search results were parsed using a custom-built parser. It identified 1329 GEO studies for ASD and eleven of its comorbidities, which were publicly available since 2002. We verified the search results by hand to remove false positives. From the hand-curated results, we retained only those series that corresponded to case-control studies and had complete, gene annotations supplied by either NCBI or the submitter. We investigated case-control studies to have matched controls for the disease cases as well as to reduce noise. We have made sure that we have at least 30 samples under each disease. For each selected GEO series, accession identifier, as well as abridged study details including the organism, tissue type, platform, and the number of samples, is provided in Appendix A: Table A.1. To remove the potential of biases that could arise from using gene expression datasets from different array platforms, tissues, and species, we avoided combining the actual measurements of expression values across platforms, tissues, and diseases. Instead, we performed differential expression analysis on each study separately and then combined the *p*-values only.

Pathway gene sets

We collected 1320 curated pathway gene sets, including those from the KEGG pathways [133, 134]; Reactome pathways [142, 143]; BioCarta pathways [144]; PID pathways [145]; SigmaAldrich gene sets; Signaling Gateway gene sets, Signal Transduction KE gene sets; and SuperArray gene sets from the Molecular Signatures Database (MSigDb) version 4.0 [146]. The gene sets were downloaded in Gene Matrix Transposed (GMT) format. Of the available gene sets, we used those that were expert-curated: C2:CP (canonical pathways); C2:CP-BioCarta (BioCarta gene sets); C2:CP-

KEGG (KEGG gene sets); C2:CP-Reactome (Reactome gene sets); and PID (Pathway Interaction Database gene sets extracted from C2). From the KEGG collection, we excluded the disease and drug-related gene sets. After excluding too large (> 300 genes) and too small (< 10 genes) gene sets, 1261, 146, 211, 629, and 196 gene sets remained in these categories, respectively.

2.4 Discussion

This study bridges previous electronic health record-based analyses of the comorbidities of large populations of individuals with ASD and the gene expression profiles of each of these comorbid diseases as well as ASD against their respective control cases. We have identified that the most significantly and consistently dysregulated pathways shared by these diseases are the innate immunity signaling pathways. For most of these disorders, the genes in these pathways can discriminate the cases from their controls with moderate accuracy, providing further evidence of the extent of the dysregulation in these pathways.

In contrast to traditional approaches that look at a group of disorders of the same organ system, we have focused on ASD and its comorbidities which often occur in different organ systems intending to find their shared genetics. It would have been ideal if we could perform the study on a sufficiently large cohort of ASD patients having enough representatives of all the comorbid diseases, but in practice, such a study is currently infeasible due to cost constraints and patient availability. Thus, to perform this study with existing datasets for ASD and its comorbidities, we make use of the power of statistics and computation. First, we look at the genetic makeup of patients of ASD and its comorbid diseases separately and then find the genetic commonalities between them. Some of the microarray studies we looked at have small sample sizes which give rise to the possibility of poor random error estimates and inaccurate statistical tests for differential expression. For this reason, we selected limma t-statistics, an Empirical Bayes method [118] which is reportedly one of the most effective methods for differential expression analysis even for data sets with small sample sizes [147].

To find the combined significance of the pathways across multiple diseases we used Fisher's combined probability test [119], because, it gives a single test of significance for multiple not-so-correlated tests of significance performed on very heterogeneous datasets. When the individual tests do not appear as significant, yet have a combined effect, Fisher's combined p-value can indicate whether the probability of the combined effect is on the whole lower than would often have been obtained by chance. Notably, a significant statistic from Fisher's test implies that the pathway is involved in the biology of at least one of the diseases. Thus, to ensure that the combined significant statistic is due to shared biology of multiple diseases we calculate minimum Bayes factors and minimum posterior probabilities of significance by chance for each significant pathway and also compare the combined p-value distributions of diseases and null data set using QQ-plots. We draw our conclusions using a combination of the p-values and the posteriors to avoid any systematic bias inherent to the methods used.

As expected in case of a neurological disease, the pathways that are most significantly dysregulated in ASD are often the pathways involved in neuronal signaling and development, synapse function, and chromatin regulation [40]. Similarly, for immunity-related diseases like, asthma, inflammatory bowel disease, and various infections, the role of innate immunity pathways is well documented in individual studies [148–154]. Despite some controversy, in the last fifteen years, experimental evidence has also pointed in the direction of dysregulated immunological signaling in at least some subsets of individuals with autism. This evidence includes findings of abnormal chemokine response to Toll-like receptor ligands associated with autism in experimental studies [121, 122], differential gene and protein expression in the central nervous system and peripheral blood of patients with ASD [58, 103, 121, 155–161]. Many reports suggest alteration of activation, amount, distribution of microglia, a representative immune cell in the brain, and its autophagy to be involved in ASD [162–165]. A recent study implicates adaptive immune dysfunction, in particular, disruption of IFN-γ signaling driven anti-pathogen response to be related to ASD and other diseases characterized by social dysfunction [128]. However, the fact that dysregulation of innate immunity pathways connects ASD with some of its non-immune related comorbidities (e.g., chronic kidney disease, cerebral palsy, muscular dystrophy), is rather intriguing.

That the innate immunity pathways are shared between ASD and the other comorbid states does not mean that a disorder in these pathways can characterize all cases of ASD. For example, in our previous work we have shown that although on average, the gene expression profile of children with ASD shows dysregulated innate immunity signaling, this is a reflection of a smaller number of individuals with ASD who are outliers in this pathway [166]. With our growing understanding of the heterogeneity of ASD and the characterization of ASD populations with distinct comorbidity associations [45], the integrative analysis we describe here may therefore implicate a subset of individuals with ASD with innate immune dysregulation that is either the result of genetic vulnerabilities [167] or particular exogenous stimuli such as infections or disordered microbiome ecology [168].

Although it is tempting to consider that innate immunity signaling is primarily driven by external environmental stimuli such as infection, we have to recognize that different organs may repurpose the same signaling mechanisms for different purposes. For example, 21% of the genes described in the KEGG long term potentiation pathway (one of the mechanisms underlying synaptic plasticity), overlap with the genes in the Gene Ontology's collection of "immune genes." It may be, as suggested by extensive epidemiological studies that sometimes the disorder is in the signaling system and at other times, it is because of an external stimulus. Specifically, nationally-scaled studies have demonstrated an increased frequency of autoimmune diseases in the parents of children with ASD [169], increased levels of gestational C-reactive protein in mothers of children with ASD [170], and an increased frequency of ASD in children after pregnancies complicated by infection [101, 102]. Some early studies also suggest the infectious exposure may be directly from the gastrointestinal microbiome [171–175], which also can engage the innate immune system. The success of treatment and prophylaxis for disorders of innate immunity in some of the diseases that are comorbid with ASD raises the possibility that similar treatments may also be successful for subsets of those with ASD.

2.5 Conclusions and future directions

Over the years, autism spectrum disorder (ASD) has baffled researchers not only with its heterogeneity but also its co-occurrence with many seemingly unrelated diseases of different organ systems. In this chapter, we introduced a three-tiered integrative omics analysis approach to capture the shared genetic and pathway-level signals that form the basis of the co-occurrence of ASD with other diseases. For ASD and 11 of its most frequently occurring comorbidities, we extracted significant differentially expressed genes, measured their enrichment in canonical pathways, and determined the pathways that are significantly shared by the diseases in question. Our pipeline can integrate transcriptomic and genetic data from heterogeneous sources in a statistically principled way. An analysis of this scale for studying ASD and its comorbidities is unheard of as per our knowledge. Our results revealed the involvement of two disrupted innate immunity pathways—Toll-like receptor signaling and chemokine signaling—in ASD and several of its comorbidities irrespective of whether they are immune-related diseases or not. We also showed that the disease genes that overlapped with these pathways could discriminate between patients and controls in each disease with at least 70 % accuracy, further proving their importance. As innate immunity pathways are imperative in orchestrating the first line-of-defense mechanism against infection-causing pathogens and environmental triggers, their involvement in ASD and its comorbidities can be thought of as the missing genetic link for environmental factors in the pathophysiology of ASD. This mindset also raises the possibility that successful treatments for innate immunity disorders may help ASD patients.

Our finding from this multi-level integrative omics study not only proves the importance of looking at different omics levels to understand a complex disease, but also motivates us to extend our efforts to other groups of seemingly unrelated diseases, including but not limited to type-2 diabetes, hypertension, Parkinson's disease, and dementia in aging populations. Furthermore, while genome-wide association studies have successfully identified genetic risk factors for complex diseases in Caucasian populations, much work still needs to be done in populations of non-Caucasian origin. Extending efforts to investigate the differential etiologies of comorbid diseases across diverse populations is particularly crucial in this era of globalization. As populations of countries become ever more admixed, clinical protocols must adapt to account for the increased genome-phenome variability across ethnicities. Thus, a reasonable future direction is to perform integrative omics studies to investigate the shared pathophysiology of groups of comorbid diseases and the population-level stratification of these comorbidities. Such a study would give us a representative understanding of the genetic and environmental risk factors of different diseases in the overall world population.

2.6 Availability of data and source code

All microarray expression studies included in this analysis are publicly available via GEO website [176]. The accession ID for each study is provided in Appendix A: Table A.1. All the pathway gene sets used for the analysis are publicly available on MSigDB website [177]. All calculations were performed in R version 2.15.1. Some pre- and post-processing were performed in Python version 2.7.6. The source code and instruction for performing the analysis are licensed under the terms of the MIT License (`https://opensource.org/licenses/MIT`) and are available from `https://github.com/snz20/3TierMA` (DOI:10.5281/zenodo.159288).

Chapter 3

Robust comparative functional metagenomics across diverse study populations

Microbial populations exhibit functional changes in response to different ambient environments. Though whole metagenome sequencing promises enough raw data to study those changes, existing functional read annotation tools are limited in their ability to compare microbial metabolic function across samples and studies directly. We introduce Carnelian, an end-to-end pipeline for metabolic functional profiling uniquely suited to finding common functional trends across diverse data sets. Carnelian can find shared metabolic pathways, concordant functional dysbioses, and distinguish Enzyme Commission (EC) terms missed by state-of-the-art functional profiling tools. We demonstrate Carnelian's effectiveness on large-scale metagenomic studies of type-2 diabetes and Crohn's disease, Parkinson's disease, and industrialized/non-industrialized cohorts.

3.1 Background

Recent advances in Next-Generation Sequencing (NGS) technologies and large-scale national and international efforts [66, 178] have generated unprecedented amounts of microbial genomic data; the NIH's National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI), and the Joint Genome Institute (JGI) currently host an order of magnitude more shotgun metagenomic data than they did 10 years ago [23]. Many methods have been developed for the initial analyses of this data—assembly, taxonomic binning, and functional profiling of metagenomic reads [22,23,179] in order to enable comparing the taxonomic and functional profiles of microbial communities. Here, we turn our attention to the discovery of trends in microbial metabolic function across diverse populations (different nations or geographical boundaries) concerning health and disease.

Hundreds of recent studies have demonstrated associations between the human microbiome and disease, including Crohn's disease [180], obesity [181], type-2 diabetes (T2D) [182, 183], colorectal cancer [184], Parkinson's disease (PD) [185], and even Autism Spectrum Disorder (ASD), which has been found to have an innate immunity component [3, 186]. Many efforts have sought to uncover shared taxonomic dysbiosis (i.e., microbial imbalance) between study populations for a given disease; however, these attempts have generally not found shared taxonomic dysbiosis, probably because the background healthy microbiomes differ significantly in taxonomic composition to begin with [79, 82–84]. Because different species may fill the same ecological niche, the traditional focus on taxonomy can lose sight of the *functional* relatedness of the microbiomes of two individuals—i.e., commonalities and differences in the functional capabilities of microbial populations [25]. For example, while most strains of lactobacilli exhibit galactosidase activity, that particular functionality can also be partially substituted for by many taxonomically distinct strains of bifidobacteria and bacteriodes [187]. In the large meta-analyses cited above [79, 82–84], there was some attempt to perform functional profiling (in addition to taxonomic profiling), but due to limitations in the study design and methods available, they were unable to find concordant pathways, which one *would* expect from the same disease. Thus, better functional profiling is essential to uncovering trends in functional relatedness when comparing study cohorts; this remains an unsolved challenge due to inconsistencies and incompleteness of annotations of microbial genes across reference databases and the lack of comparability of existing relative abundance statistics across samples and studies [23, 78].

An essential first step for uncovering functional trends in microbiomes is functional profiling of metagenomic reads, the task of assigning reads to known biological function (e.g., catalytic action, functional domain categories, genes) and estimating abundances of those functional terms. Traditional whole metagenome functional annotation approaches assemble reads into large contigs and annotate them using sequence homology, often using existing alignment tools such as BLAST [59], profile Hidden Markov Models (HMMs) or position-specific weight matrices (PWMs). Such methods include RAST [60], Megan4 [61], MEDUSA [62], Tentacle [63], MOCat2 [64], IMG4 [65], and gene catalogue-based methods [66, 67]. Since assembly is a slow, resource-heavy, and lossy process, annotating reads directly via sequence homology or read-mapping is used by another class of tools, including MG-RAST [68], HUMAnN [69], ShotMap [70], Fun4Me [71], mi-faser [72], and HUMAnN2 [73]. However, alignment-based read mapping remains time-consuming when comparing hundreds of samples from different disease conditions [74, 75]. HUMAnN2 and mi-faser significantly speed up the alignment step by using a fast protein aligner, DIAMOND [76], and thus can accurately and quickly capture functions from sequences corresponding to known proteins. However, because they are based on alignment, they are challenged in capturing shared features of functionally similar proteins that are not-so-sequence-similar, multi-domain proteins, and remote homologs.

Naturally, predicting function without having characterized a protein experimentally is difficult and runs the risk of false positives. For well-studied populations, there is little need to do so. However, when analyzing data from less studied populations—so often the case in metagenomic analysis, a significant fraction of reads sequenced do not directly correspond to proteins of known species [25, 26]. Thus methods that depend on alignment do not perform as well. We observe this problem when studying the non-industrialized Baka population (Results). Techniques from the field of remote homology detection can be used to explicitly guess at shared functions between an unknown protein and an existing one, but they operate at the level of entire protein sequences, rather than Whole Metagenome Shotgun (WMS) sequencing reads.

Alternately, k-mer based taxonomic binning methods have shown great utility compared to read-alignment approaches in assigning reads to taxonomic units [86, 87, 188, 189]. Importantly, they can be trained to directly classify WMS reads by function, even when the read itself comes from a protein that is not in existing databases. Using these techniques, we pursue the intuition that we can, for example, predict that reads correspond to a particular enzymatic function (e.g., galactosidase activity) even when the training set does not include the protein from which those reads were taken, but only for distantly related proteins (Section 3.2.8). Importantly, the design of these classification tools allows us to easily construct negative examples during training time to control the false positive rate while still allowing labeling of reads for which alignment is insufficient. Our work thus newly repurposes gapped k-mer binning techniques to directly perform efficient and accurate *functional* binning, which performs much better than existing functional profilers based on either alignment or assembly for *analyzing functional relatedness across diverse microbiomes.*

To this end, we introduce Carnelian, a compositional tool for metabolic functional profiling of whole metagenome sequencing reads, and an end-to-end pipeline that is uniquely suited to finding common functional trends across metagenomic data sets from different study populations. The pipeline we present is better suited for “comparative functional metagenomics" for three reasons. First, Carnelian makes use of a gapped k-mer classifier [85, 86], which is better able to detect the ECs (Enzyme Commission terms that classify proteins by their enzymatic action) present in non-annotated species, while simultaneously avoiding forced spurious labels through training on a negative set. Second, we build a comprehensive database focused on comparing metabolic functionality, as opposed to using typical protein databases that contain non-prokaryotic and non-metabolic annotations. Third, we present a principled statistical significance analysis for finding shared metabolic pathways using the results of EC-detection.

We demonstrate Carnelian's effectiveness through analyses of several real published and unpublished data sets. First, we compare geographically separated study cohorts of type-2 diabetes (T2D) and Crohn's disease (CD). Several of today's state-of-the-art functional annotation tools, including mi-faser, HUMANn2 (translated search), and Kraken2 (protein search) were unable to find concordant functional dysbioses between healthy and diseased microbiomes, which one would expect given that the same disease should have similar effects on different study populations. Importantly, Carnelian alone is able to find those expected concordant functional dysbioses. Next, we find that Carnelian-identified EC terms can classify patients vs. controls consistently, with higher accuracy than existing tools across T2D, CD, and Parkinson's disease (PD); this finding suggests that the additional Carnelian classifications are not spurious. Next, using a combination of published and unpublished data sets, we further demonstrate Carnelian's effectiveness on geographically and dietarily diverse healthy microbiomes of industrialized individuals from the United States (Boston: new data set) and non-industrialized communities from Cameroon (Baka ethnicity: new data set), Ethiopia (Gimbichu region) [26], and Madagascar (Betsimisaraka and Tsimihety ethnicities) [26]. Unlike existing methods, Carnelian was able to uncover the expected pathway-level similarities in core metabolic function between healthy individuals from each of those communities. Lastly, on a Parkinson's disease case-control metagenomic read data set, we show that Carnelian uniquely finds several hallmarks of Parkinson's disease in the patient microbiomes. For all these experiments, Carnelian, mi-faser, HUMAnN2, and Kraken2 were run with Carnelian's curated reference database to ensure an unbiased comparison.

Carnelian is robust to sequencing technology biases and is equally applicable to non-human metagenomic data sets where it can find meaningful biological patterns. In benchmarking experiments, Carnelian achieves higher sensitivity and F1-score than current state-of-the-art alignment-based tools: mi-faser [72] and HUMAnN2 [73] (translated search) as well as a fast alignment-free k-mer based tool: Kraken2 [188] (protein search)—all run with the same reference database. On a synthetic human gut metagenomic data set of 5 million reads (150 bp, single-ended), Carnelian requires $\sim$ 16 minutes using 16 CPU cores—this is roughly 2x faster than mi-faser ($\sim$ 29 minutes) and similar to HUMAnN2's translated search ($\sim$ 18 minutes) on the same number of CPUs on the same machine (a 40-core machine with 320 GB RAM, each core was Intel Xeon CPU E5-2695 v2 @ 2.40GHz). As new data is being collected from all over the world (e.g., our Cameroon data), we expect Carnelian to be an essential tool in analyzing functional similarities and differences.

3.2 Results

3.2.1 Overview of Carnelian

We present Carnelian, a novel gapped k-mer based functional profiler, and an end-to-end pipeline for comparative functional metagenomic studies using WMS reads from diverse study populations. Our pipeline enables the comparison of functional summaries of WMS data by designing more consistently annotated reference databases of microbial proteins, building a functional annotation tool better suited for assigning functions to reads that are not readily alignable to known proteins, and generating comparable abundance statistics across samples and studies (Figure 3-1).

WMS data comes from a mixture of many different organisms and can encode 100x more unique genes than are present in just the human genome [67]. Only a fraction of these genes has known functional annotations in existing databases. Even of those genes with annotations, many of the annotations are computationally predicted and therefore less reliable. We are also primarily interested in microbial functions that can influence host health, such as the production of metabolites, extracellular enzymes, or immunostimulatory surface structures [190]. Thus, we constructed our gold standard reference database with curated prokaryotic proteins that have verified unique and complete EC labels which provide a direct mapping to KEGG metabolic pathways for our later analyses. Our curated database consists of 7,884 prokaryotic proteins with 2,010 unique EC labels and is provided on our website (`http://carnelian.csail.mit.edu`).

Another important characteristic of metagenomic data is that the reads sequenced often come from non-annotated species; without a known reference, taxonomic read

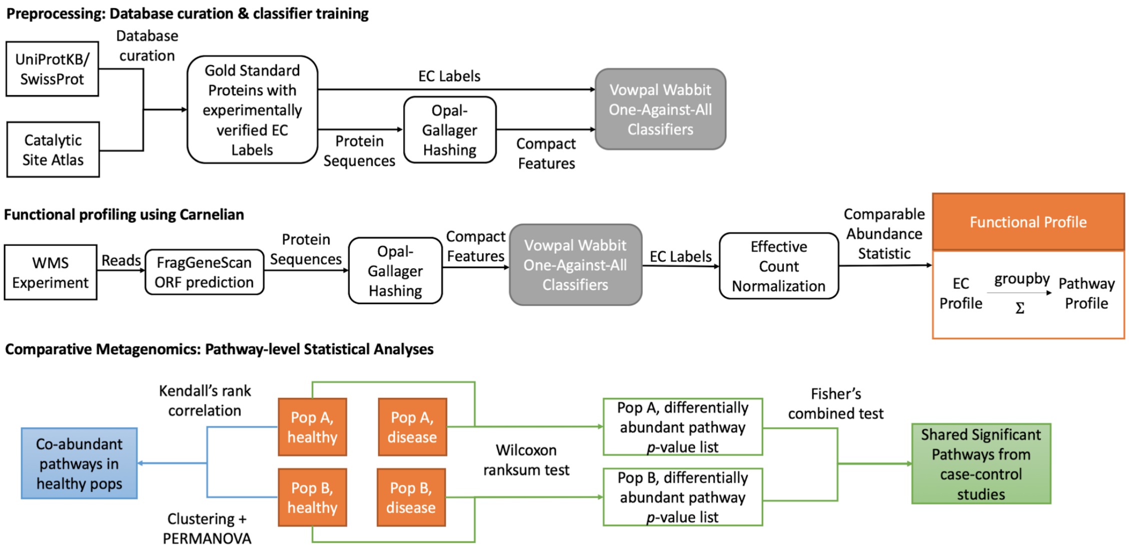

Preprocessing: Database curation & classifier training
UniProtKB/ SwissProt
Catalytic Site Atlas
Database curation
Gold Standard Proteins with experimentally verified EC Labels
EC Labels
Protein Sequences
Opal-Gallager Hashing
Compact Features
Vowpal Wabbit One-Against-All Classifiers
Functional profiling using Carnelian
WMS Experiment
Reads
FragGeneScan ORF prediction
Protein Sequences
Opal-Gallager Hashing
Compact Features
Vowpal Wabbit One-Against-All Classifiers
EC Labels
Effective Count Normalization
Comparable Abundance Statistic
Functional Profile
EC Profile
groupby
Σ
Pathway Profile
Comparative Metagenomics: Pathway-level Statistical Analyses
Kendall's rank correlation
Co-abundant pathways in healthy pops
Clustering + PERMANOVA
Pop A, healthy
Pop A, disease
Pop B, healthy
Pop B, disease
Wilcoxon ranksum test
Pop A, differentially abundant pathway p-value list
Pop B, differentially abundant pathway p-value list
Fisher's combined test
Shared Significant Pathways from case-control studies

Figure 3-1: **Comparative functional metagenomics with Carnelian.** *Preprocessing.* We build a gold standard database by combining reviewed prokaryotic proteins with complete Enzyme Commission (EC) labels and evidence of existence from UniProtKB/SwissProt with curated prokaryotic catalytic residues with complete EC labels from the Catalytic Site Atlas. Carnelian first represents gold standard proteins in a compact feature space using Opal-Gallager hashing. Then it trains a set of one-against-all (OAA) classifiers (implemented using the Vowpal Wabbit framework) using the compact feature representation of those proteins as well as negative samples based off of randomly shuffled sequences generated by HMMER. *Functional Profiling.* To functionally profile WMS reads from an experiment, Carnelian first performs probabilistic ORF prediction using FragGeneScan. Next, the ORFs are represented in a compact feature space using the same Opal-Gallager hashing technique. The trained OAA classifier ensemble is then used to classify the ORFs into appropriate EC bins. Abundance estimates of ECs are computed from the raw ORF counts in the EC bins by normalizing against effective protein length per EC bin and a per million scaling factor. Pathway profiles (Orange) are computed by grouping the ECs into metabolic pathways and summing the abundance estimates. *Comparative Metagenomics.* We start from pathway profiles (Orange) of different populations and conditions. (Blue) Functional relatedness of healthy microbiomes across different populations is assessed by co-abundance pathway analysis. Pathway co-abundance estimates are quantified by Kendall's rank correlation. Co-abundance clusters are determined by Ward-Linkage hierarchical clustering, and the PERMANOVA test is used to determine if the centroids of those clusters differ between Populations A and B. (Green) Functional trends analysis across different case-control cohorts of a disease is performed using differential abundance analysis by Wilcoxon rank-sum test and shared significance analysis by Fisher's combined probability test.

classifiers are limited in their annotation ability. Luckily, related proteins that share a function also share compositional (gapped k-mer) features in their amino acid sequence, even across species. Leveraging this intuition, the Carnelian pipeline uses probabilistic ORF detection to enable the application of a compositional gapped classifier ensemble on the full amino-acid sequence; this classifier ensemble is better able to bin proteins present in non-annotated species. More precisely, Carnelian first detects all possible ORFs from the input reads using FragGeneScan [191], which probabilistically detects the coding part(s) of the reads and translates them to the best possible ORFs. Then Carnelian encodes the ORFs into a low-dimensional compact feature space using Opal-Gallager hashes [85, 86]. Once so encoded, these ORFs are annotated by Carnelian's classifier ensemble, a set of one-against-all support vector machines. The classifier ensemble is trained with functionally annotated gold standard proteins represented in the same compact feature space, and with negative samples based off of randomly shuffling in human sequences generated via HMMER [192]. The training is performed in an online fashion (i.e., we load only one input sequence in memory at a time), making incremental training of Carnelian's classifier ensemble easy when new annotations become available.

Relative abundance statistics output from standard functional profiling tools are not directly comparable across samples and studies; to address this problem, Carnelian borrows from transcriptomic normalization practices. From input WMS reads, Carnelian constructs a functional vector containing effective read counts per EC label (i.e., read counts normalized against effective protein length per EC label and a per million scaling factor that takes into account the effect of the lengths of proteins with other EC labels on the relative abundance of a particular EC label) (Methods). This normalization step is similar to the "transcripts per million" (TPM) counts used for quantifying transcript abundances from RNA-seq data [158]. The sum of Carnelian's effective read counts thus remains constant across all samples, unlike the raw read counts and reads per kilobase (RPK) measures used by existing functional annotation tools (e.g., HUMAnN2). This normalization makes sample profiles directly comparable to each other across experiments performed with different sequencing depths

(Methods). These EC profiles are used to quantify KEGG metabolic pathways for comparative analysis of different study populations.

3.2.2 Revealing concordant functional dysbiosis across geographically separated disease cohorts

Comparing healthy and diseased microbiomes is key to understanding their effect on host biology, enabling clinical diagnoses and informed therapeutics [193, 194]. While taxonomic dysbiosis (i.e., alteration of the species-level composition of the microbiome) in the patient population is often geography-specific and not generalizable [83, 183, 195], we instead looked at functional dysbiosis. As expected, functional dysbiosis is indeed more generalizable in type-2 diabetes and Crohn's disease data sets we studied, but only when we used Carnelian as opposed to other methods for the analysis.

We quantified the metabolic functional capacity of the gut microbiomes of patients and controls in two large-scale T2D data sets [182,183], and two CD data sets [178,196] at enzyme and pathway levels. Our results revealed concordant functional dysbioses between geographically separated disease cohorts—13 common metabolic pathways between Chinese and European T2D patient microbiomes and eight common pathways between US and Swedish CD patient microbiomes (Table 3.1).

For the T2D cohorts, we generated the EC profiles of preprocessed fecal samples from Chinese and European individuals using Carnelian and determined the differentially abundant ECs between patients and controls using a cutoff of Wilcoxon rank-sum test p-value < 0.05 after Benjamini-Hochberg (BH) correction and absolute log fold change > 0.33. In both Chinese and European cohorts, Carnelian reported reduced levels of several glycosyltransferases (e.g. 2.4.1.1, 2.4.1.7, 2.4.1.15) and abundance of several carbon-oxygen lyases (e.g. 4.2.1.120, 4.2.1.20, 4.2.1.42) in the T2D gut (Tables 3.2 and 3.3). At the pathway level, it found 30 significantly altered metabolic pathways in the Chinese T2D patients (BH-corrected Wilcoxon rank-sum test p-value < 0.05 and absolute log fold change > 0.11) and 36 pathways altered be-

Table 3.1: **Shared functional dysbiosis between two type-2 diabetes (T2D) cohorts and two Crohn's disease (CD) cohorts.** **(a)** Common pathways between Chinese and European T2D cohorts which have significantly altered read abundances. We found 13 shared pathways of which 12 are highly relevant to T2D; these pathways are significant in individual cohorts (BH-corrected Wilcoxon rank-sum test p-value < 0.05) as well as in Fisher's combined test at p-value < 0.05 cutoff. On the other hand, mi-faser finds only the photosynthesis pathway and Kraken2 finds the photosynthesis and aflatoxin biosynthesis pathways to be commonly disrupted between both the cohorts; with HUMAnN2-profiles, no overlap at the pathway level was found (Tables 3.12–3.17). **(b)** Common pathways between the US and Swedish CD cohorts which have significantly altered read abundances. We identify shared dysbiosis in 8 pathways between the two study cohorts; these pathways are significant in individual cohorts as well as in Fisher's combined test at p-value < 0.05 cutoff. On the other hand, only Kraken2 finds the beta-alanine metabolism pathway to be commonly disrupted between both the cohorts; with mi-faser- and HUMAnN2-profiles, no overlap at the pathway level was found (Tables 3.24, 3.25, 3.28, 3.29, 3.32 and 3.33). Here, SB: significant in both the studies, NB: detected but not significant in both the studies, SC: significant in Chinese cohort only, SE: significant in European cohort only, SU: significant in the US cohort only, SS: significant in the Swedish cohort only.

(a) Common pathways between Chinese and European T2D cohorts

ID	Pathway	Carnelian	mi-faser	HUMAnN2	Kraken2	Fisher's p
						(Carnelian)
00030	Pentose phosphate pathway	SB	NB	NB	NB	6.59E-03
00040	Pentose and gluconerate interconversions	SB	NB	NB	NB	9.88E-03
00051	Fructose and mannose metabolism	SB	SE	NB	NB	4.94E-04
00052	Galactose metabolism	SB	NB	NB	NB	4.71E-03
00061	Fatty acid biosynthesis	SB	SC	NB	SC	6.56E-03
00190	Oxidative phosphorylation	SB	SE	SC	SE	4.97E-04
00250	Alanine, aspartate and glutamate metabolism	SB	NB	NB	NB	1.48E-04
00290	Valine, leucine and isoleucine biosynthesis	SB	SE	NB	NB	1.68E-05
00590	Arachidonic acid metabolism	SB	NB	NB	NB	2.11E-03
00600	Sphingolipid metabolism	SB	SE	NB	SC	8.86E-05
00730	Thiamine metabolism	SB	NB	NB	NB	2.62E-03
00983	Drug metabolism - other enzymes	SB	NB	NB	NB	2.62E-03
00195	Photosynthesis	SB	SB	SC	SB	2.74E-03
00254	Aflatoxin biosynthesis	SC	SC	NB	SB	1.03E-02

(b) Common pathways between US and Swedish CD cohorts

ID	Pathway	Carnelian	mi-faser	HUMAnN2	Kraken2	Fisher's p
						(Carnelian)
00500	Starch and sucrose metabolism	SB	NB	SS	SS	4.91E-06
00620	Pyruvate metabolism	SB	NB	NB	SS	4.05E-04
00640	Propanoate metabolism	SB	NB	NB	NB	9.04E-03
00290	Valine, leucine and isoleucine biosynthesis	SB	SS	NB	SS	5.03E-03
00450	Selenocompound metabolism	SB	NB	NB	NB	8.95E-03
00460	Cyanoamino acid metabolism	SB	NB	SS	SS	8.33E-05
00513	Various types of N-glycan biosynthesis	SB	NB	NB	NB	5.79E-03
00710	Carbon fixation in photosynthetic organisms	SB	NB	NB	SS	1.09E-05
00410	Beta-alanine metabolism	NB	SS	NB	SB	5.79E-01

tween European T2D patients and individuals with normal glucose tolerance (NGT) (Tables 3.4 and 3.5). Notably, 13 of these pathways are significantly shared between both patient cohorts (Fisher's combined p-value < 0.05) and highly relevant to T2D (Table 3.1(a)). For example, we observed significant depletion of reads in several carbohydrate metabolism pathways, such as the pentose phosphate pathway, pentose and glucuronate interconversions, fructose and mannose metabolism, galactose metabolism in patient guts compared to controls in both cohorts (Tables 3.4 and 3.5). Across these two cohorts, we also observed a higher rate of oxidative phosphorylation in the patient gut—a finding that is in agreement with the original studies [182,183]. Additionally, in each of the patient cohorts, we found significantly lower read abundances in several vitamin-B metabolism pathways (e.g. thiamine metabolism) compared to the healthy gut. Notably, EC- and Pathway-level results from mi-faser, HUMAnN2, and Kraken2 were unable to uncover shared pathways of relevance between the two cohorts (Tables 3.6–3.17).

Carnelian-generated EC profiles of the Crohn's disease cohorts revealed a shift in the metabolic functionality of the patient gut microbiome compared to the control gut microbiome as indicated by lower read abundances in several essential enzymes and pathways. The most significantly variable ECs between patients and controls (Wilcoxon rank-sum test p-value < 0.05 after BH correction and absolute log fold change > 0.58) in both the US and Swedish cohorts include several hexosyltransferases (2.4.1.-), oxidoreductases acting on aldehyde group (1.2.7.-), glycosidases (3.2.1.-), and hydrolyases (4.2.1.-), which are key players in different carbohydrate metabolism pathways (Tables 3.18 and 3.19). Many of these enzymes were not found by other methods. We also observed a decrease in the relative abundance of several enzymes, including aminobutyraldehyde dehydrogenase (1.2.1.19), acetylornithinase (3.5.1.16), lysine decarboxylase (4.1.1.18), and 5-carboxymethyl-2-hydroxymuconic acid isomerase (5.3.3.10). These enzymes play crucial roles in the metabolism of several essential amino acids, including arginine, proline, lysine, and tyrosine. Thus, this finding might indicate a lower rate of microbial absorption of such amino acids from the diet. Several enzymes involved in vitamin B metabolism such as pyridoxine phosphatase (3.1.3.74), dihydroxy-acid dehydratase (4.2.1.9), phosphomethylpyrimidine synthase (4.1.99.17), etc. were also found to be depleted in the CD gut; of the methods we compared against, only Carnelian was able to uncover these findings (Tables 3.18, 3.19, 3.22, 3.23, 3.26, 3.27, 3.30 and 3.31).

At the pathway-level, we found 25 significantly altered metabolic pathways in the guts of CD patients from the US (BH-corrected Wilcoxon rank-sum test p-value < 0.05 and absolute log fold change > 0.11) and 35 pathways altered between Swedish CD patients and healthy individuals (Tables 3.20 and 3.21). Notably, eight of these pathways are significantly shared between both patient cohorts (Fisher's combined p-value < 0.05) and seven of them are highly relevant to Crohn's disease (Table 3.1(b)). For example, we observed significant depletion of reads in three carbohydrate metabolism pathways, namely, starch and sucrose metabolism, pyruvate metabolism, and propanoate metabolism in patient guts compared to the controls in both the cohorts (Tables 3.20 and 3.21). In both data sets, we also observed lower abundance of reads in valine, leucine and isoleucine (essential amino acids) biosynthesis and cyanoamino acid metabolism pathways in CD patients. We further observed a lower abundance of reads in the selenocompound metabolism and various N-glycan biosynthesis pathways in the CD guts compared to the normal individuals in both cohorts. Although non-specific to CD, the reduced read abundance in carbon fixation pathway might be indicative of the imbalance of energy homeostasis in the patient gut. Importantly, mi-faser and HUMAnN2 found no shared pathways of relevance between the two cohorts and Kraken2 found shared dysbiosis in only the beta-alanine metabolism pathway. EC- and pathway-level results from mi-faser, HUMAnN2, and Kraken2 can be found in Tables 3.22–3.33.

Table 3.2: Significant differentially abundant ECs identified by Carnelian in the T2D-Qin data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.33.

EC	**log Fold Change**	**adjusted *p*-value**	**EC**	**log Fold Change**	**adjusted *p*-value**
5.4.99.62	0.36	0.0426	1.8.2.3	0.33	0.032
3.6.1.23	-0.92	0.0457	1.1.1.28	-0.35	0.0088
2.4.2.2	-0.34	0.0348	3.2.1.52	-0.33	0.0242
2.4.2.6	0.36	0.0168	2.4.1.1	-0.43	0.0165
1.1.1.100	-0.33	0.0021	1.4.1.24	0.35	0.0179
2.7.8.35	0.42	0.017	1.4.1.4	-0.37	0.0004
3.7.1.8	0.37	0.0044	2.6.1.84	0.33	0.0065
2.7.2.4	0.39	0.0375	2.7.7.61	0.34	0.0033
1.13.11.27	0.35	0.0117	4.2.1.120	0.39	0.005
4.1.1.33	0.34	0.0015	5.4.2.11	-0.38	0.0031
4.2.1.20	0.35	0.0036	4.3.1.15	0.34	0.0493
2.7.1.220	0.36	0.0051	1.3.1.70	0.34	0.0003
1.1.1.408	0.35	0.0155	1.13.11.6	0.5	0.0002
1.12.2.1	0.34	0.031	4.2.1.147	0.35	0.0053
1.8.4.14	0.35	0.0432	2.4.1.7	-0.5	0.0067
1.17.7.4	-0.45	0.0149	5.4.3.2	0.41	0.0068
1.17.7.3	-0.34	0.0059	3.1.3.85	0.39	0.0424

Table 3.3: Significant differentially abundant ECs identified by Carnelian in the T2D-Karlsson data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.33.

EC	**T2D-NGT**		**IGT-NGT**	
	logFC	**adj *p*-value**	**logFC**	**adj *p*-value**
2.4.1.329	-0.35	0.013	-0.12	0.4543
4.1.1.101*	0.35	0.0059	0.32	0.0047
4.2.99.20	-0.61	0.0136	-0.59	0.018
3.1.3.8	-0.38	0.0136	-0.05	0.864
2.6.1.113*	0.41	0.0309	0.26	0.0188
3.2.2.23	-0.57	0.0211	-0.55	0.2546
4.2.2.n2	-0.73	0.008	-0.18	0.2417
4.1.1.79	-0.34	0.0083	-0.09	0.4543
1.14.13.127	-0.53	0.0069	-0.49	0.0483
3.5.1.5	0.41	0.0026	0.25	0.1976
3.1.3.12*	-0.45	0.0326	-0.59	0.02
1.12.98.1	-0.4	0.0072	-0.13	0.3454
6.2.1.44	-0.36	0.0082	-0.15	0.2417
2.3.2.21	-0.52	0.0266	-0.45	0.2546
5.4.99.20	-0.49	0.0398	-0.52	0.1142
2.1.1.10	0.53	0.004	0.24	0.0837
5.3.1.22	-0.71	0.0091	-0.91	0.0809
1.5.1.36	-0.36	0.0326	-0.07	0.9132
4.2.1.42	-0.36	0.0483	-0.31	0.3868
1.1.1.251	-0.38	0.0007	-0.31	0.0242
4.1.2.48*	-0.45	0.0101	-0.53	0.002
6.3.2.33	-0.49	0.0013	-0.23	0.1179
1.8.98.1	-0.36	0.0428	-0.14	0.3575
6.3.2.36	-0.38	0.0091	-0.08	0.6347
2.3.1.5	0.37	0.0233	0.08	0.5279
2.8.4.1	-0.34	0.0144	-0.11	0.5028
1.2.99.7	0.34	0.0089	0.23	0.007
2.4.1.15*	-0.53	0.0184	-0.55	0.0242
6.2.1.3	-0.34	0.0009	-0.22	0.0258
4.2.1.119*	0.4	0.0015	0.62	0.0006
3.4.13.9*	-0.7	0.0016	-0.58	0.0316
3.1.4.16	-0.35	0.0039	-0.12	0.1949
3.1.4.12	0.34	0.0266	0.12	0.5485
2.1.1.90	-0.42	0.0266	-0.18	0.2386
3.2.1.80	0.38	0.0018	0.33	0.0069
2.4.1.182	-0.34	0.0059	-0.27	0.053
1.13.11.3	0.38	0.0002	0.19	0.0138

[*] ECs marked with '*' are significantly variable between impaired glucose tolerance (IGT) and normal glucose tolerance (NGT) individuals as well.

Table 3.4: Pathways identified as significantly variable between T2D patients and healthy controls in the T2D-Qin data set using Carnelian-generated functional profiles. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value $<$ 0.05 and abs (log fold change) $>$ 0.11.

Category	**Name**	**logFC**	**Adjusted *p*-value**
C	Glycolysis / Gluconeogenesis	-0.21	0.0333
C	Citrate cycle (TCA cycle)	-0.20	0.0478
C	Pentose phosphate pathway	-0.22	0.0393
C	Pentose and glucuronate interconversions	-0.23	0.0361
C	Fructose and mannose metabolism	-0.21	0.0465
C	Galactose metabolism	-0.22	0.0376
C	Ascorbate and aldarate metabolism	-0.25	0.0165
L	Fatty acid biosynthesis	0.21	0.0262
L	Fatty acid elongation	0.22	0.0372
L	Steroid biosynthesis	0.34	0.0019
E	Oxidative phosphorylation	0.35	0.0035
E	Photosynthesis	0.18	0.0448
N	Pyrimidine metabolism	0.18	0.0432
AA	Alanine, aspartate and glutamate metabolism	0.18	0.0018
SM	Aflatoxin biosynthesis	-0.19	0.0197
AA	Valine, leucine and isoleucine biosynthesis	0.25	0.0007
AA	Arginine and proline metabolism	0.23	0.0034
AA	Tyrosine metabolism	0.19	0.0325
AA	Phenylalanine metabolism	0.20	0.0366
AA	Glutathione metabolism	0.28	0.0068
C	Amino sugar and nucleotide sugar metabolism	-0.23	0.0217
L	Arachidonic acid metabolism	0.27	0.0071
L	Sphingolipid metabolism	0.25	0.0221
G	Glycosphingolipid biosynthesis - globo and isoglobo series	0.28	0.0144
C	Pyruvate metabolism	-0.19	0.0083
V	Thiamine metabolism	-0.18	0.0002
V	Vitamin B6 metabolism	-0.14	0.0054
V	Nicotinate and nicotinamide metabolism	-0.23	0.0034
V	Biotin metabolism	-0.17	0.0002
X	Drug metabolism - other enzymes	0.20	0.0350

[*] Here, C = Carbohydrate Metabolism; L = Lipid Metabolism; E = Energy Metabolism; N = Nucleotide Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); SM = Biosynthesis of Secondary Metabolites; G = Glycan Biosynthesis and Metabolism; V = Metabolism of Co-factors and Vitamins; X = Xenobiotics Biodegradation and Metabolism.

Table 3.5: Pathways identified as significantly variable between T2D patients and normal glucose tolerance (NGT) individuals in the T2D-Karlsson data set using Carnelian-generated functional profiles. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.11.

Category	**Name**	**T2D-NGT**		**IGT-NGT**	
		logFC	**adj *p*-value**	**logFC**	**adj *p*-value**
C	Pentose phosphate pathway	-0.20	0.0219	-0.08	0.0268
C	Pentose and glucuronate interconversions	-0.21	0.0357	-0.15	0.8701
C	Fructose and mannose metabolism	-0.24	0.0010	-0.13	0.1573
C	Galactose metabolism	-0.22	0.0147	-0.06	0.0094
L	Fatty acid biosynthesis	0.24	0.0309	0.11	0.5331
E	Oxidative phosphorylation	0.28	0.0128	0.12	0.8274
E	Photosynthesis	0.35	0.0067	0.15	0.6797
AA	Arginine biosynthesis	0.26	0.0069	0.13	0.2546
AA	Alanine, aspartate and glutamate metabolism	0.24	0.0069	0.11	0.6126
AA	Glycine, serine and threonine metabolism	0.21	0.0207	0.06	0.4638
AA	Valine, leucine and isoleucine biosynthesis	0.27	0.0016	0.14	0.2481
AA	Lysine biosynthesis	0.24	0.0147	0.12	0.1766
SM	Carbapenem biosynthesis	0.28	0.0117	0.14	0.3181
AA	Histidine metabolism	0.21	0.0345	0.07	0.0456
AA	Phenylalanine, tyrosine and tryptophan biosynthesis	0.26	0.0047	0.16	0.0492
SM	Phenazine biosynthesis	0.28	0.0191	0.18	0.0492
AA	Selenocompound metabolism	0.21	0.0017	0.11	0.0539
AA	Cyanoamino acid metabolism	0.31	0.0091	0.08	0.4831
AA	D-Arginine and D-ornithine metabolism	0.14	0.0492	0.09	0.7318
G	Other glycan degradation	0.22	0.0215	0.08	0.4879
L	Glycerolipid metabolism	0.25	0.0002	0.16	0.0316
L	Arachidonic acid metabolism	0.26	0.0314	0.22	0.0144
L	Sphingolipid metabolism	0.26	0.0332	0.09	0.6402
L	Glycosphingolipid biosynthesis - ganglio series	0.21	0.0130	0.05	0.7912
V	Thiamine metabolism	-0.22	0.0377	-0.15	0.7673
GI	Aminoacyl-tRNA biosynthesis	0.15	0.0428	0.10	0.7792
X	Drug metabolism - other enzymes	0.24	0.0082	0.12	0.5695
X	Atrazine degradation	0.47	0.0031	0.39	0.1843
V	Retinol metabolism	0.28	0.0297	0.17	0.3414
C	C5-Branched dibasic acid metabolism	0.27	0.0017	0.13	0.1161
T	Terpenoid backbone biosynthesis	0.24	0.0083	0.08	0.8517
E	Nitrogen metabolism	0.25	0.0063	0.10	0.3534
E	Sulfur metabolism	0.22	0.0037	0.10	0.1505
X	Polycyclic aromatic hydrocarbon degradation	0.57	0.0002	0.38	0.0138
X	Chloroalkane and chloroalkene degradation	0.24	0.0276	0.10	0.7201
X	Naphthalene degradation	0.30	0.0371	0.16	0.2031

[*] Here, C = Carbohydrate Metabolism; L = Lipid Metabolism; E = Energy Metabolism; N = Nucleotide Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); SM = Biosynthesis of Secondary Metabolites; G = Glycan Biosynthesis and Metabolism; V = Metabolism of Co-factors and Vitamins; X = Xenobiotics Biodegradation and Metabolism; GI = Genetic Information Processing; T = Metabolism of Terpenoids and Polyketides.

Table 3.6: Significant differentially abundant ECs identified by mi-faser in the T2D-Qin data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.33.

EC	**logFC**	**adjusted *p*-value**
2.7.1.58	0.89	0.0379
6.2.1.13	0.34	0.0188
1.2.3.3	0.77	0.0176
2.7.1.113	0.85	0.0293
4.1.1.48	0.39	0.0444
6.1.2.1	0.38	0.04
4.2.1.162	-0.39	0.012
2.5.1.88	0.67	0.0012
1.1.3.48	1.01	0.0213
5.4.99.61	-0.67	0.0256
2.4.2.6	0.83	0.0059
2.6.1.34	0.35	0.0103
2.6.1.39	0.56	0.0023
1.1.1.215	0.97	0.0349
4.1.99.1	0.35	0.0221
2.7.8.36	0.37	0.0077
2.7.8.38	0.82	0.0101
2.4.2.45	-1.37	0.0022
2.1.1.264	-0.86	0.0052
5.3.99.11	-2.83	0.016
5.1.1.13	1	0.0053
1.1.1.310	0.7	0.0075
2.6.1.17	0.63	0.0439
2.7.1.95	1.88	0.0211
1.3.8.2	1.27	0.0003
1.13.11.27	-4.84	0.0326
3.2.1.11	-1.58	0.0243
4.1.1.86	-0.44	0.0127
4.1.1.33	5.97	0.0043
4.2.1.5	1.53	0.0282
1.3.1.12	-1.16	0.0295
2.1.1.289	-0.37	0.0174
2.7.1.220	-0.72	0.0406
3.5.3.8	-3.34	0.0082
3.5.3.1	0.39	0.0351
2.4.1.288	2.49	0.0095
5.1.99.1	0.45	0.0027
2.7.1.162	0.57	0.0452
5.4.2.8	-0.52	0.0484
5.1.3.23	0.92	0.048
3.2.1.136	0.49	0.047
1.1.1.28	0.36	0.0023
1.3.1.101	-0.52	0.0327
2.4.1.8	0.44	0.0102
6.3.1.12	1.1	0.0004
1.1.1.377	0.63	0.0029
1.1.1.371	0.54	0.0002
4.2.1.120	0.69	0.0012
3.2.1.89	-0.51	0.0279
2.3.1.245	0.47	0.0047
4.3.1.14	0.35	0.0386
2.7.1.76	1.05	0.0108
3.2.1.99	-0.82	0.0274
3.1.3.90	5.54	0.0219
2.4.1.345	0.72	0.0041
5.1.99.1	0.45	0.0027
2.7.1.162	0.57	0.0452
5.4.2.8	-0.52	0.0484
5.1.3.23	0.92	0.048
3.2.1.136	0.49	0.047
1.1.1.28	0.36	0.0023

Table 3.7: Significant differentially abundant ECs identified by HUMAnN2 in the T2D-Qin data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.33.

EC	**log Fold Change**	**adjusted *p*-value**
4.6.1.12	-0.39	0.0054
4.4.1.25	1.74	0.0307
4.1.1.48	0.74	0.0096
6.1.2.1	1.94	0.027
2.4.1.329	-0.4	0.0038
1.6.5.8	-0.78	0.0187
3.1.3.3	-0.39	0.0065
4.1.99.2	-0.68	0.013
1.1.1.310	2.92	0.0356
1.1.1.304	-0.85	0.0331
2.7.1.95	1.88	0.0423
2.7.1.205	-0.85	0.0026
3.2.1.18	-0.64	0.0288
4.1.1.86	-1.19	0.0219
1.12.5.1	0.42	0.007
2.3.1.30	-0.35	0.0119
3.1.3.73	-0.62	0.0277
2.4.1.282	-1.51	0.0189
3.4.24.3	2.35	0.0102
1.1.1.39	0.39	0.0365
1.2.1.92	3.52	0.012
5.4.2.8	-0.75	0.0295
2.3.3.3	0.66	0.015
5.1.3.23	1.47	0.0176
3.2.1.135	0.96	0.0095
3.3.1.1	0.34	0.0291
1.97.1.2	1.14	0.034
4.2.1.119	1.14	0.0152
6.3.1.12	2.6	0.0069
2.1.1.228	-0.36	0.043
2.4.99.16	1.61	0.0414
4.3.3.6	-0.33	0.0261
2.7.1.76	0.86	0.0197
3.6.3.42	1.09	0.0367
1.3.7.5	-1.51	0.0072
3.1.3.83	0.44	0.0144

Table 3.8: Significant differentially abundant ECs identified by Kraken2 in the T2D-Qin data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.33.

EC	**log Fold Change**	**adjusted *p*-value**	**EC**	**log Fold Change**	**adjusted *p*-value**
2.7.7.13	0.60	0.0433	3.4.24.75	-1.66	0.0205
1.7.7.2	0.77	0.0079	1.1.1.336	0.50	0.0432
1.13.11.73	-0.41	0.0120	1.1.1.338	-5.93	0.0453
2.4.1.329	-0.37	0.0055	1.14.13.92	-3.91	0.0441
6.3.2.49	1.21	0.0482	2.3.3.10	-1.43	0.0435
2.4.2.2	-0.41	0.0371	2.7.1.39	-0.60	0.0334
2.4.2.6	6.11	0.0424	3.4.11.15	1.32	0.0167
3.5.1.93	-0.58	0.0356	1.13.11.2	1.72	0.0399
1.7.2.1	0.34	0.0092	1.15.1.2	-0.57	0.0462
2.6.1.39	0.80	0.0280	1.1.1.28	0.94	0.0320
2.5.1.96	1.71	0.0438	1.97.1.2	1.14	0.0137
6.3.2.n2	0.94	0.0212	3.4.23.42	-1.26	0.0172
2.4.1.250	-0.41	0.0285	3.2.1.1	0.36	0.0189
2.4.1.332	-1.27	0.0032	4.99.1.3	-1.01	0.0149
2.4.1.247	-0.41	0.0380	2.1.1.80	-0.57	0.0257
1.17.8.1	-0.42	0.0124	1.3.1.101	-0.35	0.0075
5.1.1.13	0.57	0.0488	1.3.1.54	-1.56	0.0111
2.7.4.2	-0.93	0.0089	1.5.99.13	-0.53	0.0274
1.14.16.1	2.74	0.0083	4.2.1.120	0.96	0.0210
3.6.4.9	1.64	0.0080	3.2.1.89	-0.60	0.0247
4.2.2.22	3.56	0.0225	3.4.19.1	-3.72	0.0343
3.2.2.3	0.61	0.0229	1.13.11.9	2.00	0.0491
1.3.3.11	-0.90	0.0478	2.5.1.68	-0.63	0.0205
2.6.1.77	0.90	0.0470	4.1.1.39	0.64	0.0256
3.5.1.44	-0.99	0.0162			

Table 3.9: Significant differentially abundant ECs identified by mi-faser in the T2D-Karlsson data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.33.

EC	**T2D-NGT**		**IGT-NGT**		**EC**	**T2D-NGT**		**IGT-NGT**	
	logFC	**adj *p*-value**	**logFC**	**adj *p*-value**		**logFC**	**adj *p*-value**	**logFC**	**adj *p*-value**
3.1.22.4	-0.86	0.0159	-1.09	0.0037	1.1.1.311	0.77	0.0001	0.76	0.0002
4.6.1.16	-1.45	0.0188	-0.22	0.6484	1.12.98.2	-0.92	0.0273	-0.35	0.1814
2.5.1.77	-1.11	0.0494	-0.38	0.2929	1.1.1.87	-0.39	0.0391	-0.05	0.9070
2.3.1.169	0.84	0.0099	0.42	0.5426	2.8.3.12	0.56	0.0133	-0.12	0.9438
1.2.7.4	0.36	0.0443	0.20	0.5416	2.1.1.10	1.33	0.0044	0.75	0.2677
1.2.7.8	-0.38	0.0398	-0.22	0.1310	4.2.1.40	-0.53	0.0364	-0.50	0.0744
2.1.1.217	1.56	0.0451	1.25	0.4678	2.4.1.54	1.60	0.0061	0.86	0.2544
3.5.4.10	-0.79	0.0098	-0.04	0.2118	2.2.1.10	-1.01	0.0498	-0.59	0.0523
1.1.1.136	-0.61	0.0488	-0.43	0.0458	1.12.7.2	-0.49	0.0027	-0.19	0.1095
1.1.1.302	-0.35	0.0234	-0.17	0.2101	2.8.1.10	0.52	0.0001	0.12	0.2059
2.7.7.23	-0.34	0.0475	-0.45	0.0895	2.7.4.26	-1.21	0.0051	-0.34	0.0689
4.4.1.19	-1.08	0.0295	-0.26	0.2198	2.1.1.246	-1.52	0.0404	-0.25	0.3333
4.2.2.6	-2.56	0.0386	-8.00	0.0149	1.1.1.405	-0.44	0.0246	-0.21	0.5110
4.1.1.15	-0.40	0.0323	-0.68	0.0015	2.1.1.171	-0.52	0.0433	-1.00	0.0018
5.4.99.60	0.45	0.0124	-0.05	0.8942	2.5.1.114	-1.03	0.0099	-0.33	0.2767
2.4.1.11	-1.27	0.0068	-1.35	0.0434	3.2.1.70	0.47	0.0451	1.28	0.0035
2.4.1.19	-0.92	0.0276	-1.40	0.0035	1.1.1.39	0.34	0.0163	-0.03	0.7722
2.4.1.329	-1.07	0.0031	-0.16	0.6834	3.4.13.22	-1.88	0.0478	-1.46	0.3035
5.1.2.1	0.44	0.0420	0.54	0.0659	1.2.99.7	0.63	0.0014	0.34	0.0636
3.2.1.185	0.63	0.0026	0.48	0.2105	3.1.21.3	-0.40	0.0017	-0.27	0.0451
4.1.1.101	0.96	0.0127	0.72	0.3697	5.1.3.21	1.76	0.0448	1.34	0.2146
3.5.4.27	-1.11	0.0184	-0.47	0.1579	2.7.7.1	-0.97	0.0060	-0.17	0.1514
2.6.1.109	-1.27	0.0042	-0.51	0.0805	1.5.98.1	-0.92	0.0124	-0.24	0.1579
1.1.1.107	1.28	0.0002	0.53	0.0424	1.5.98.2	-0.70	0.0298	-0.13	0.1898
2.6.1.34	-0.88	0.0057	-0.18	0.0971	4.2.3.153	-0.97	0.0370	-0.32	0.1561
4.2.1.36	0.50	0.0051	0.68	0.0591	3.1.3.45	-0.93	0.0487	-2.17	0.0001
2.7.8.35	1.65	0.0063	1.11	0.0363	2.1.1.86	-0.87	0.0430	-0.34	0.1944
3.5.4.39	-1.10	0.0098	-0.41	0.1587	2.4.1.8	0.60	0.0147	0.26	0.5785
2.4.2.48	-0.76	0.0331	-0.03	0.2412	2.4.1.5	2.44	0.0418	1.88	0.7413
3.2.2.20	-0.35	0.0357	-0.21	0.1055	1.4.1.24	-0.80	0.0131	-0.26	0.1445
1.2.1.22	-0.76	0.0315	-0.35	0.1693	1.4.1.1	0.43	0.0177	0.28	0.0895
4.1.99.14	-1.15	0.0258	-0.35	0.4157	3.1.4.16	-0.34	0.0133	-0.37	0.0268
2.7.1.85	0.52	0.0359	-1.39	0.8280	2.5.1.41	-1.12	0.0481	-0.44	0.2051
4.1.1.79	-0.95	0.0071	-0.40	0.0967	1.5.99.15	-0.40	0.0494	-0.21	0.2767
4.1.1.75	0.67	0.0230	-0.13	0.7452	3.4.21.62	0.45	0.0382	0.19	0.3887
2.3.1.136	2.65	0.0386	-0.12	0.6655	2.1.1.90	-0.95	0.0202	-0.11	0.3653
2.1.1.74	-0.34	0.0405	-0.05	0.4831	2.1.1.98	-1.09	0.0104	-0.84	0.0409
2.4.2.4	-0.99	0.0041	-1.35	0.0009	3.2.1.80	0.85	0.0008	0.80	0.0008
2.4.2.54	-0.95	0.0122	-0.22	0.1337	2.5.1.120	2.64	0.0149	0.56	0.6458
1.3.4.1	-0.80	0.0108	-0.30	0.0657	1.1.1.385	0.91	0.0145	-0.83	0.5544
3.4.16.4	-0.34	0.0371	-0.44	0.0520	2.7.7.73	0.53	0.0003	-0.06	0.6726
3.5.1.2	-0.55	0.0160	-0.84	0.0012	2.7.7.72	-1.07	0.0428	-1.41	0.0059
3.5.1.5	0.51	0.0047	0.37	0.0520	5.3.2.8	0.64	0.0225	-0.35	0.8501
1.5.1.49	-0.99	0.0245	-0.04	0.2489	2.4.1.7	0.47	0.0125	0.34	0.1477
1.5.1.40	-0.98	0.0410	-0.64	0.1291	4.1.1.31	-0.82	0.0205	-0.88	0.0188
3.6.4.9	-0.83	0.0159	-0.21	0.1579	2.1.1.206	-0.98	0.0108	-0.25	0.2505
2.4.2.29	-0.34	0.0015	-0.06	0.2031	6.3.4.19	-0.69	0.0389	-1.19	0.0035
1.1.1.261	-0.79	0.0286	-0.35	0.0895					

[*] Here, IGT = impaired glucose tolerance; NGT = normal glucose tolerance.

Table 3.10: Significant differentially abundant ECs identified by HUMAnN2 in the T2D-Karlsson data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test *p*-value < 0.05 and abs (log fold change) > 0.33.

EC	**T2D-NGT**		**IGT-NGT**	
	logFC	**adj *p*-value**	**logFC**	**adj *p*-value**
1.2.7.8	-0.77	0.0294	-0.43	0.1691
2.5.1.86	-1.8	0.0186	-1.82	0.0318
5.1.2.1	0.46	0.0459	1.02	0.021
4.1.1.101	2.42	0.0135	1.36	0.1014
3.5.4.27	-0.74	0.0082	0.03	0.246
2.3.1.n4	1.34	0.0055	1.62	0.1079
2.7.14.1	-1.71	0.0319	-0.31	0.1862
2.3.1.136	1.88	0.0386	-0.08	0.9506
4.2.1.28	0.64	0.0075	-0.17	0.3897
5.1.1.13	-0.54	0.0317	-1.74	0.0003
2.7.1.121	-1.5	0.0142	-0.51	0.0028
2.7.4.8	-0.54	0.0298	-0.98	0.0011
3.5.1.5	0.63	0.0122	0.65	0.0705
3.1.3.15	-3.55	0.0402	-0.4	0.8988
2.7.1.205	-0.93	0.0025	-1.16	0.0019
4.1.2.57	-1.44	0.033	-1.44	0.0661
1.12.98.2	-0.49	0.0232	0.36	0.429
1.12.98.1	-0.46	0.0391	0.23	0.3172
1.1.1.87	-1.55	0.0076	-0.11	0.3338
2.8.3.12	2.25	0.0249	2.69	0.0148
4.1.99.17	0.67	0.0348	0.09	0.9407
1.5.1.36	5.31	0.0334	5	0.1246
6.1.1.11	-0.37	0.0314	-0.24	0.0783
1.3.7.11	-0.84	0.0144	0.36	0.9021
2.3.1.30	-0.47	0.0376	-0.55	0.033
3.4.21.107	-0.4	0.0483	-0.65	0.0002
2.1.1.192	-0.51	0.0093	-0.35	0.0169
2.4.2.17	-1.4	0.0089	-0.41	0.2987
3.4.13.22	-2.29	0.0465	-2.03	0.035
2.8.4.3	-0.56	0.0136	-0.13	0.2612
1.5.98.1	-0.39	0.0148	0.16	0.2677
6.2.1.3	-0.61	0.0164	-0.49	0.0301
2.4.1.5	2.61	0.0021	2	0.0077
1.8.1.8	-0.5	0.0422	-1.33	0.0001
3.2.1.80	2.05	0.0162	1.34	0.9552
4.3.1.3	-0.98	0.016	-0.38	0.7561
2.7.1.6	-0.46	0.0309	0.08	0.9689
2.7.1.76	3.62	0.0372	3.59	0.0187
3.2.1.97	1.55	0.0063	1.71	0.4388
1.1.1.65	-1.17	0.0321	-2.33	0.0030
2.4.1.7	0.89	0.0356	0.60	0.2126

* Here, IGT = impaired glucose tolerance; NGT = normal glucose tolerance.

Table 3.11: Significant differentially abundant ECs identified by Kraken2 in the T2D-Karlsson data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.33.

EC	**T2D-NGT**		**IGT-NGT**		**EC**	**T2D-NGT**		**IGT-NGT**	
	logFC	**adj *p*-value**	**logFC**	**adj *p*-value**		**logFC**	**adj *p*-value**	**logFC**	**adj *p*-value**
2.5.1.77	-0.94	0.0141	-0.35	0.173	1.16.3.1	-0.9	0.0385	-2.07	0.0057
1.7.1.4	-1.52	0.0387	-0.44	0.1002	1.1.1.333	-3.18	0.0128	-4.73	0.0019
3.4.14.13	0.49	0.0463	0.34	0.3417	2.7.4.26	-1.02	0.0028	-0.45	0.074
1.7.7.1	-1.72	0.0399	-0.1	0.438	2.1.1.198	-0.59	0.0106	-0.45	0.0214
1.14.14.12	2.76	0.004	2.67	0.0319	3.4.13.22	-0.96	0.0492	-0.75	0.3439
1.14.99.3	-0.48	0.0339	-0.51	0.0513	2.8.4.1	-0.69	0.0045	-0.12	0.6181
1.1.1.136	-0.79	0.0063	-0.37	0.0284	1.2.99.7	0.6	0.0017	0.51	0.0055
4.2.1.162	0.55	0.002	0.22	0.0837	2.4.1.9	0.91	0.0293	0.1	0.9686
1.3.99.28	-0.98	0.0117	-0.06	0.0197	3.4.11.10	-1.84	0.0105	-0.88	0.2528
2.4.1.11	-0.69	0.0015	-0.57	0.0743	4.1.2.17	-0.99	0.0131	-1.28	0.0331
5.1.2.1	0.51	0.0147	0.56	0.1236	2.7.7.1	-1.09	0.0189	-0.42	0.1802
3.5.4.27	-1	0.0179	-0.31	0.2367	2.7.7.47	-0.69	0.0433	-3.56	0.0066
3.5.4.29	-1.12	0.0383	-0.56	0.2187	1.5.98.1	-1.02	0.0077	-0.27	0.2805
2.5.1.96	-1.19	0.0097	0.23	0.1779	3.8.1.7	-1	0.0032	-0.17	0.5086
2.1.1.63	-0.63	0.0272	-1.06	0.0005	4.2.3.153	-0.9	0.0078	-0.12	0.288
4.1.1.65	-0.73	0.0169	-1.48	0.0139	6.2.1.3	-0.46	0.0398	-0.58	0.0046
4.2.1.36	0.49	0.0036	0.16	0.4157	3.1.1.48	1.88	0.0463	0.55	0.4962
2.4.2.48	-0.44	0.0299	0.03	0.7862	1.14.13.154	2.29	0.0093	-0.23	0.6692
2.7.14.1	-0.57	0.0091	-0.19	0.2513	2.1.1.86	-0.54	0.0072	-0.11	0.5695
1.2.1.22	-0.5	0.0252	-0.64	0.0381	1.8.1.14	0.56	0.003	0.36	0.0161
4.1.1.79	-0.36	0.0265	0	0.4826	6.5.1.1	0.44	0.006	0.73	0.0293
6.3.4.21	-0.52	0.0384	-0.68	0.0058	3.1.4.17	0.95	0.0386	0.49	0.2786
2.4.2.4	-0.57	0.003	-0.32	0.0634	2.6.1.84	0.77	0.0054	0.61	0.328
1.3.4.1	-0.86	0.0453	-0.28	0.2245	3.4.21.62	0.49	0.0215	0.2	0.4385
1.2.1.38	0.42	0.0009	0.11	0.4543	2.1.1.98	-0.84	0.0078	-0.45	0.1586
3.5.1.5	0.5	0.0045	0.38	0.1005	3.5.1.32	0.65	0.0126	0.15	0.7188
6.3.4.12	1.14	0.03	0.27	0.7834	1.9.3.1	-0.52	0.0147	-0.28	0.2468
3.6.4.9	-0.86	0.0213	-0.07	0.2528	3.4.21.72	-0.94	0.0013	-0.53	0.0266
3.2.1.91	0.81	0.0442	-0.25	0.618	1.4.99.1	0.47	0.0125	-0.31	0.5855
1.12.98.1	-0.8	0.0075	-0.02	0.7156	1.4.99.5	1.77	0.0396	1.07	0.0861
1.1.1.87	-0.65	0.0055	-0.26	0.3335	2.5.1.113	0.38	0.0443	0.24	0.2233
4.1.99.17	0.43	0.0009	0.06	0.8824	1.13.11.5	1.58	0.0131	0.29	0.9321
2.1.1.10	1.46	0.0064	0.87	0.2528	3.5.1.110	-1.33	0.0142	-1.62	0.003
1.5.1.36	2.63	0.0066	2.5	0.1205	3.4.11.1	0.63	0.037	0.68	0.0731
2.7.1.167	0.73	0.0098	0.22	0.6314	4.2.1.148	1.74	0.0266	-1.84	0.8681
3.4.21.96	-1.34	0.0091	-0.28	0.0399	1.1.1.65	-1.45	0.0257	-1.93	0.0687
6.3.2.4	-0.35	0.0136	-0.11	0.2355	1.1.1.69	-0.77	0.0419	-0.48	0.0399
1.12.7.2	-0.49	0.0133	-0.21	0.1157	2.4.1.7	0.54	0.0224	0.37	0.0559
3.1.3.70	-0.94	0.037	-1.6	0.0112	5.4.1.4	-1.02	0.0042	-0.62	0.0466
2.7.1.175	1.4	0.0132	0.71	0.2943	4.1.1.38	-1.65	0.0292	-1.57	0.086

* Here, IGT = impaired glucose tolerance; NGT = normal glucose tolerance.

Table 3.12: Pathways identified as significantly variable between T2D patients and healthy controls in the T2D-Qin data set using mi-faser-generated functional profiles. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.11.

Category	**Name**	**logFC**	**Adjusted *p*-value**
L	Fatty acid biosynthesis	-0.15	0.000009
E	Photosynthesis	-0.21	0.000456
SM	Aflatoxin biosynthesis	-0.24	0.017628
AA	D-Arginine and D-ornithine metabolism	0.21	0.003276
AA	Glutathione metabolism	0.20	0.002753
G	Various types of N-glycan biosynthesis	0.22	0.030179
T	Biosynthesis of 12-, 14- and 16-membered macrolides	1.02	0.038230
G	Lipoarabinomannan (LAM) biosynthesis	0.78	0.001068
L	Glycosphingolipid biosynthesis - globo and isoglobo series	0.18	0.010605
X	Aminobenzoate degradation	0.23	0.006568
V	Riboflavin metabolism	-0.13	0.006686
V	Vitamin B6 metabolism	-0.21	0.011154
V	Lipoic acid metabolism	0.18	0.037381
T	Carotenoid biosynthesis	0.62	0.000006
SM	Flavone and flavonol biosynthesis	-0.33	0.043905
T	Insect hormone biosynthesis	0.27	0.012885
SM	Biosynthesis of secondary metabolites - unclassified	-0.19	0.001662
T	Biosynthesis of ansamycins	-0.26	0.000619

[*] Here, C = Carbohydrate Metabolism; L = Lipid Metabolism; E = Energy Metabolism; N = Nucleotide Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); SM = Biosynthesis of Secondary Metabolites; G = Glycan Biosynthesis and Metabolism; V = Metabolism of Co-factors and Vitamins; X = Xenobiotics Biodegradation and Metabolism; GI = Genetic Information Processing; T = Metabolism of Terpenoids and Polyketides.

Table 3.13: Pathways identified as significantly variable between T2D patients and normal glucose tolerance (NGT) individuals in the T2D-Karlsson data set using mi-faser-generated functional profiles. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.11.

Category	**Name**	**T2D-NGT**		**IGT-NGT**	
		logFC	**adj *p*-value**	**logFC**	**adj *p*-value**
E	Oxidative phosphorylation	0.12	0.0031	0.06	0.3296
E	Photosynthesis	0.17	0.0215	0.12	0.3335
AA	Arginine biosynthesis	0.12	0.0125	0.09	0.0756
SM	Monobactam biosynthesis	0.11	0.0048	0.06	0.0756
AA	Valine, leucine and isoleucine biosynthesis	0.11	0.0078	0.05	0.1817
AA	Phenylalanine, tyrosine and tryptophan biosynthesis	0.14	0.0034	0.10	0.0501
AA	Cyanoamino acid metabolism	0.16	0.0139	-0.01	0.6627
G	Other glycan degradation	0.24	0.0082	0.16	0.5485
G	Glycosaminoglycan degradation	0.19	0.0320	0.10	0.9689
L	Glycerolipid metabolism	0.19	0.0002	0.11	0.0796
G	Arabinogalactan biosynthesis - Mycobacterium	1.57	0.0069	0.82	0.1062
L	Sphingolipid metabolism	0.22	0.0287	0.15	0.5909
L	Glycosphingolipid biosynthesis - ganglio series	0.23	0.0101	0.15	0.5963
X	Atrazine degradation	0.51	0.0048	0.37	0.0520

* Here, IGT = impaired glucose tolerance, C = Carbohydrate Metabolism; L = Lipid Metabolism; E = Energy Metabolism; N = Nucleotide Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); SM = Biosynthesis of Secondary Metabolites; G = Glycan Biosynthesis and Metabolism; V = Metabolism of Co-factors and Vitamins; X = Xenobiotics Biodegradation and Metabolism; GI = Genetic Information Processing; T = Metabolism of Terpenoids and Polyketides.

Table 3.14: Pathways identified as significantly variable between T2D patients and healthy controls in the T2D-Qin data set using functional profiles generated by HUMAnN2. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.11.

Category	**Name**	**logFC**	**Adjusted *p*-value**
C	Citrate cycle (TCA cycle)	-0.13	0.0020
L	Synthesis and degradation of ketone bodies	-0.28	0.0419
E	Oxidative phosphorylation	-0.18	0.0145
E	Photosynthesis	-0.25	0.0172
AA	Tryptophan metabolism	0.15	0.0402
L	Glycerolipid metabolism	0.12	0.0184
L	Glycosphingolipid biosynthesis - globo and isoglobo series	0.15	0.0215
X	Toluene degradation	1.73	0.0416
X	Nitrotoluene degradation	-0.13	0.0312
V	Vitamin B6 metabolism	-0.31	0.0039
T	Terpenoid backbone biosynthesis	-0.12	0.0309
SM	Biosynthesis of secondary metabolites - unclassified	-0.32	0.0386

* ∗ Here, C = Carbohydrate Metabolism; L = Lipid Metabolism; E = Energy Metabolism; N = Nucleotide Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); SM = Biosynthesis of Secondary Metabolites; G = Glycan Biosynthesis and Metabolism; V = Metabolism of Co-factors and Vitamins; X = Xenobiotics Biodegradation and Metabolism; GI = Genetic Information Processing; T = Metabolism of Terpenoids and Polyketides.

Table 3.15: Pathways identified as significantly variable between T2D patients and normal glucose tolerance (NGT) individuals in the T2D-Karlsson data set using functional profiles generated by HUMAnN2. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.11.

Category	**Name**	**T2D-NGT**		**IGT-NGT**	
		logFC	**adj *p*-value**	**logFC**	**adj *p*-value**
C	Fructose and mannose metabolism	-0.12	0.0483	0.14	0.0180
L	Fatty acid degradation	-0.18	0.0219	-0.05	0.6458
AA	Lysine degradation	-0.30	0.0113	-0.31	0.0730
AA	Histidine metabolism	-0.59	0.0019	-0.34	0.0348
AA	Cyanoamino acid metabolism	0.32	0.0170	0.18	0.1089
V	One carbon pool by folate	0.14	0.0371	0.09	0.5909
X	Atrazine degradation	0.63	0.0122	0.65	0.0705

[*] ∗ Here, IGT = impaired glucose tolerance, C = Carbohydrate Metabolism; L = Lipid Metabolism; E = Energy Metabolism; N = Nucleotide Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); SM = Biosynthesis of Secondary Metabolites; G = Glycan Biosynthesis and Metabolism; V = Metabolism of Co-factors and Vitamins; X = Xenobiotics Biodegradation and Metabolism; GI = Genetic Information Processing; T = Metabolism of Terpenoids and Polyketides.

Table 3.16: Pathways identified as significantly variable between T2D patients and healthy controls in the T2D-Qin data set using functional profiles generated by Kraken2. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.11.

Category	**Name**	**logFC**	**Adjusted *p*-value**
L	Fatty acid biosynthesis	0.19	0.0025
E	Photosynthesis	-0.20	0.0026
SM	Aflatoxin biosynthesis	-0.27	0.0088
AA	Tryptophan metabolism	0.27	0.0021
SM	Novobiocin biosynthesis	-0.19	0.0374
AA	Glutathione metabolism	0.20	0.0034
G	Glycosaminoglycan degradation	0.21	0.0232
L	Sphingolipid metabolism	0.22	0.0086
L	Glycosphingolipid biosynthesis - globo and isoglobo series	0.20	0.0243
V	Nicotinate and nicotinamide metabolism	0.21	0.0051
L	Biosynthesis of unsaturated fatty acids	-0.27	0.0017
T	Biosynthesis of ansamycins	-0.25	0.0031

[*] Here, C = Carbohydrate Metabolism; L = Lipid Metabolism; E = Energy Metabolism; N = Nucleotide Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); SM = Biosynthesis of Secondary Metabolites; G = Glycan Biosynthesis and Metabolism; V = Metabolism of Co-factors and Vitamins; X = Xenobiotics Biodegradation and Metabolism; GI = Genetic Information Processing; T = Metabolism of Terpenoids and Polyketides.

Table 3.17: Pathways identified as significantly variable between T2D patients and normal glucose tolerance (NGT) individuals in the T2D-Karlsson data set using functional profiles generated by Kraken2. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.11.

Category	**Name**	**T2D-NGT**		**IGT-NGT**	
		logFC	**adj p-value**	**logFC**	**adj p-value**
C	Ascorbate and aldarate metabolism	-0.13	0.0314	-0.03	0.5178
L	Synthesis and degradation of ketone bodies	-0.86	0.0467	-0.86	0.0103
E	Oxidative phosphorylation	0.15	0.0108	0.07	0.4175
E	Photosynthesis	0.19	0.018	0.09	0.5748
AA	Arginine biosynthesis	0.12	0.0207	0.09	0.1038
SM	Aflatoxin biosynthesis	0.18	0.032	0.15	0.1355
SM	Monobactam biosynthesis	0.11	0.0034	0.1	0.0612
X	Fluorobenzoate degradation	-0.58	0.0236	0.24	0.681
AA	Selenocompound metabolism	0.18	0.0055	0.09	0.1107
G	Mannose type O-glycan biosynthesis	-0.27	0.0357	0.15	0.5178
L	Glycerolipid metabolism	0.15	0.0251	0.01	0.6402
X	Chloroalkane and chloroalkene degradation	0.16	0.0042	0.14	0.0657
C	C5-Branched dibasic acid metabolism	0.11	0.0125	0.08	0.0539
X	Atrazine degradation	0.5	0.0043	0.37	0.1072

* Here, IGT = impaired glucose tolerance, C = Carbohydrate Metabolism; L = Lipid Metabolism; E = Energy Metabolism; N = Nucleotide Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); SM = Biosynthesis of Secondary Metabolites; G = Glycan Biosynthesis and Metabolism; V = Metabolism of Co-factors and Vitamins; X = Xenobiotics Biodegradation and Metabolism; GI = Genetic Information Processing; T = Metabolism of Terpenoids and Polyketides.

Table 3.18: Significant differentially abundant ECs identified by Carnelian in the CD-HMP data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.58.

EC	**logFC**	**Adjusted *p*-value**
2.4.1.292	-0.67	9.55E-05
1.10.3.10	-0.60	2.42E-03
2.7.7.39	0.79	2.42E-03
4.2.1.12	-0.75	2.61E-03
1.2.1.19	-0.60	4.31E-03
1.3.3.3	-0.71	6.07E-03
4.3.1.15	-0.64	6.07E-03
1.17.5.3	-1.44	8.42E-03
3.2.1.28	-0.65	8.98E-03
1.1.1.60	-0.72	1.23E-02
3.2.2.21	-0.71	1.23E-02
3.1.1.41	-0.67	1.65E-02
3.2.2.8	-0.88	1.65E-02
4.2.1.42	-0.84	1.75E-02
2.4.2.52	-0.89	1.86E-02
2.7.7.19	-1.08	2.08E-02
1.3.1.101	-0.69	2.20E-02
2.7.1.186	-0.72	2.20E-02
5.1.3.26	-0.79	2.20E-02
3.6.1.25	-0.68	2.33E-02
4.2.1.40	-0.79	2.60E-02
5.3.3.10	-0.78	2.65E-02
2.7.7.61	-0.89	3.05E-02
3.1.3.74	-1.22	3.05E-02
3.4.23.49	-0.93	3.05E-02
3.5.1.16	-0.76	3.39E-02
4.1.1.65	-0.93	3.57E-02
2.7.1.55	-0.88	3.96E-02
3.2.1.31	-0.65	3.96E-02
2.4.1.12	-1.13	4.17E-02
3.1.4.14	-1.20	4.60E-02
5.3.1.26	0.74	4.60E-02
1.17.1.9	-0.90	4.84E-02
1.8.5.5	-0.76	4.84E-02

Table 3.19: Significant differentially abundant ECs identified by Carnelian in the CD-Swedish data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.58.

EC	**logFC**	**adj *p*-value**	**EC**	**logFC**	**adj *p*-value**
2.4.1.20	-1.46	1.00E-05	3.4.24.55	-1.23	5.46E-03
3.6.3.5	-1.19	1.23E-05	6.3.4.14	-1.25	5.84E-03
1.4.1.2	-1.35	3.53E-05	2.4.1.288	-1.32	6.12E-03
2.4.2.47	-1.20	3.87E-05	1.3.1.31	-1.26	6.20E-03
1.1.1.40	-1.41	4.64E-05	4.3.1.24	-1.35	6.21E-03
3.6.3.42	-0.93	4.64E-05	2.3.2.3	-0.92	6.26E-03
2.4.1.52	-1.08	5.08E-05	3.6.3.2	-0.88	6.99E-03
2.7.9.1	-1.04	5.55E-05	1.2.7.4	-0.90	7.38E-03
1.4.4.2	-1.39	6.62E-05	1.2.99.7	-1.14	7.56E-03
3.4.24.69	-1.31	7.22E-05	1.14.13.171	-1.71	7.57E-03
2.4.1.25	-1.00	1.81E-04	4.1.2.27	-1.74	8.32E-03
3.4.21.53	-1.00	1.81E-04	6.1.1.10	-0.61	8.67E-03
5.4.99.2	-1.27	1.81E-04	4.6.1.1	-0.89	1.02E-02
1.8.98.3	-1.26	2.29E-04	1.2.7.5	-0.80	1.07E-02
3.2.1.3	-0.91	2.34E-04	2.4.1.247	-1.35	1.12E-02
2.4.2.46	-1.40	2.48E-04	6.3.5.4	-0.69	1.19E-02
3.2.1.21	-1.16	2.48E-04	2.7.8.47	-1.50	1.36E-02
3.2.1.14	-0.97	4.21E-04	6.2.1.51	-0.84	1.36E-02
2.3.2.27	-1.16	5.25E-04	3.2.1.4	-1.10	1.38E-02
4.2.1.135	-1.15	5.41E-04	2.4.1.19	-1.35	1.56E-02
3.2.1.18	-1.37	6.51E-04	2.4.99.21	-0.77	1.60E-02
2.4.2.48	-1.47	9.38E-04	3.2.1.52	-0.80	1.60E-02
5.99.1.2	-0.72	9.88E-04	1.13.11.61	-1.15	1.94E-02
3.2.1.133	-1.23	1.53E-03	3.2.1.176	-1.51	1.97E-02
3.1.26.12	-1.13	1.79E-03	2.7.1.195	-0.66	2.51E-02
3.2.1.169	-0.98	1.95E-03	3.1.11.5	-0.73	2.57E-02
3.2.1.131	-0.64	2.11E-03	4.2.1.9	-0.62	2.57E-02
3.1.7.2	-1.01	2.18E-03	3.1.21.3	-1.11	2.80E-02
1.4.7.1	-0.67	2.45E-03	1.2.7.6	-1.19	2.97E-02
1.1.1.39	-1.54	2.50E-03	1.2.4.2	-0.59	3.05E-02
3.6.3.4	-0.86	2.77E-03	2.7.1.193	-0.79	3.05E-02
4.1.1.32	-1.55	2.80E-03	4.1.99.17	-0.74	3.21E-02
3.2.1.41	-1.31	3.05E-03	1.8.7.1	-1.31	3.47E-02
6.1.1.18	-0.76	3.13E-03	3.4.21.72	-0.68	3.53E-02
2.4.1.9	-1.26	3.15E-03	1.7.7.1	-1.10	3.58E-02
3.2.1.35	-1.08	3.48E-03	3.2.1.8	-0.77	3.82E-02
4.1.1.18	-1.10	3.68E-03	3.2.1.1	-0.87	3.84E-02
2.2.1.7	-0.81	3.97E-03	5.4.99.15	-0.78	3.92E-02
2.3.1.41	-0.89	3.97E-03	3.2.1.187	1.39	4.91E-02
3.2.1.177	-0.87	4.46E-03	4.2.1.82	-0.95	4.97E-02
1.8.5.5	-1.13	5.44E-03			

* *

Table 3.20: Pathways identified as significantly variable between CD patients and healthy controls in the CD-HMP data set using Carnelian-generated functional profiles. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value $<$ 0.05 and absolute logFC $\geq$ 0.11.

Category	**Name**	**logFC**	**adj *p*-value**
GI	Aminoacyl-tRNA biosynthesis	-0.11	4.02E-03
AA	Lysine biosynthesis	-0.17	7.90E-03
C	Pyruvate metabolism	-0.17	1.23E-02
L	Glycerolipid metabolism	0.23	1.30E-02
AA	Valine, leucine and isoleucine biosynthesis	-0.18	2.33E-02
AA	Cyanoamino acid metabolism	-0.15	2.60E-02
AA	Selenocompound metabolism	-0.18	2.89E-02
C	Propanoate metabolism	-0.18	3.05E-02
C	Starch and sucrose metabolism	-0.15	3.39E-02
X	Caprolactam degradation	0.25	1.56E-02
SM	Flavone and flavonol biosynthesis	0.65	3.96E-02
T	Polyketide sugar unit biosynthesis	-0.31	4.17E-02
SM	Streptomycin biosynthesis	-0.28	4.17E-02
L	Fatty acid elongation	0.25	4.17E-02
G	Various types of N-glycan biosynthesis	0.24	4.38E-02
T	Geraniol degradation	0.27	1.65E-02
T	Biosynthesis of ansamycins	0.21	1.75E-02
T	Biosynthesis of siderophore group nonribosomal peptides	0.33	3.57E-02
T	Insect hormone biosynthesis	0.41	1.66E-03
X	Aminobenzoate degradation	0.22	2.81E-03
T	Limonene and pinene degradation	0.26	6.49E-03
E	Carbon fixation in photosynthetic organisms	-0.11	1.08E-02
E	Photosynthesis	-0.17	1.15E-02
GI	Bacterial chemotaxis	0.15	1.38E-02
GI	Flageller assembly	-0.21	7.40E-03

[*] Here, C = Carbohydrate Metabolism; L = Lipid Metabolism; E = Energy Metabolism; N = Nucleotide Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); SM = Biosynthesis of Secondary Metabolites; G = Glycan Biosynthesis and Metabolism; V = Metabolism of Co-factors and Vitamins; X = Xenobiotics Biodegradation and Metabolism; GI = Genetic Information Processing; T = Metabolism of Terpenoids and Polyketides.

Table 3.21: Pathways identified as significantly variable between CD patients and healthy controls in the CD-Swedish data set using Carnelian-generated functional profiles. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and absolute logFC ≥ 0.11.

Category	**Name**	**logFC**	**adj *p*-value**
C	Glycolysis / Gluconeogenesis	-0.79	2.45E-03
C	Pyruvate metabolism	-0.65	2.95E-03
AA	Glycine, serine and threonine metabolism	-1.33	4.24E-05
C	Starch and sucrose metabolism	-0.83	9.04E-06
C	Amino sugar and nucleotide sugar metabolism	-0.48	9.64E-03
V	Thiamine metabolism	-0.76	1.45E-02
AA	Arginine biosynthesis	-1.35	3.53E-05
AA	Valine, leucine and isoleucine degradation	-1.27	1.81E-04
AA	Arginine and proline metabolism	-0.88	1.81E-04
AA	Cyanoamino acid metabolism	-1.16	2.48E-04
AA	Lysine degradation	-0.94	2.89E-04
AA	Valine, leucine and isoleucine biosynthesis	-0.62	2.57E-02
C	Propanoate metabolism	-0.56	3.82E-02
AA	Selenocompound metabolism	-0.46	3.98E-02
L	Sphingolipid metabolism	-0.35	4.91E-02
L	Fatty acid biosynthesis	-1.07	3.12E-04
C	Glyoxylate and dicarboxylate metabolism	-0.89	3.63E-04
C	Pentose and glucuronate interconversions	-0.88	1.57E-02
L	Glycosphingolipid biosynthesis - globo and isoglobo series	-0.8	1.60E-02
G	Various types of N-glycan biosynthesis	-0.8	1.60E-02
E	Sulfur metabolism	-0.98	9.64E-03
E	Nitrogen metabolism	-0.82	1.36E-05
AA	Phenylalanine metabolism	-0.98	1.74E-06
C	Pantothenate and CoA biosynthesis	-0.62	2.57E-02
SM	Tropane, piperidine and pyridine alkaloid biosynthesis	-0.99	8.04E-04
X	Nitrotoluene degradation	-0.91	2.03E-03
E	Carbon fixation in photosynthetic organisms	-0.89	6.62E-05
G	Lipoarabinomannan (LAM) biosynthesis	-1.2	3.87E-05
G	Arabinogalactan biosynthesis - Mycobacterium	-1.31	4.51E-08
AA	Taurine and hypotaurine metabolism	-1.32	2.76E-06
SM	Phenylpropanoid biosynthesis	-1.09	4.79E-06
E	Carbon fixation pathways in prokaryotes	-0.8	2.89E-04
V	Biotin metabolism	-0.89	3.97E-03
T	Terpenoid backbone biosynthesis	-0.81	3.97E-03
SM	Biosynthesis of secondary metabolites - unclassified	-1.71	7.57E-03

[*] Here, C = Carbohydrate Metabolism; L = Lipid Metabolism; E = Energy Metabolism; N = Nucleotide Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); SM = Biosynthesis of Secondary Metabolites; G = Glycan Biosynthesis and Metabolism; V = Metabolism of Co-factors and Vitamins; X = Xenobiotics Biodegradation and Metabolism; GI = Genetic Information Processing; T = Metabolism of Terpenoids and Polyketides.

Table 3.22: Significant differentially abundant ECs identified by mi-faser in the CD-HMP data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test *p*-value < 0.05 and abs (log fold change) > 0.58.

EC	**logFC**	**Adjusted *p*-value**	**EC**	**logFC**	**Adjusted *p*-value**
2.3.1.n3	-0.58	1.21E-03	1.18.1.3	0.78	2.65E-02
2.1.1.113	-2.64	1.02E-02	3.2.1.156	-0.80	2.89E-02
4.2.3.152	-2.52	1.23E-02	6.3.1.13	-2.52	2.92E-02
3.2.1.28	0.91	1.75E-02	3.9.1.2	-0.65	2.93E-02
4.2.1.51	-4.20	2.08E-02	3.2.1.70	-0.75	3.22E-02
3.4.19.5	1.00	2.08E-02	2.3.1.41	0.59	3.22E-02
2.7.7.2	-3.62	2.33E-02	1.17.2.1	1.83	3.25E-02
1.1.1.108	-3.59	2.33E-02	1.2.7.4	-0.78	4.62E-02
2.3.1.169	-1.25	2.37E-02	4.2.2.2	-2.03	4.67E-02
1.1.1.310	-0.85	2.65E-02	2.7.7.39	-0.81	4.80E-02

Table 3.23: Significant differentially abundant ECs identified by mi-faser in the CD-Swedish data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test *p*-value < 0.05 and abs (log fold change) > 0.58.

EC	**logFC**	**adj *p*-value**	**EC**	**logFC**	**adj *p*-value**
1.4.4.2	-1.17	3.53E-05	2.4.1.247	-2.91	1.12E-02
4.1.1.31	1.61	9.03E-05	3.2.1.135	1.49	1.24E-02
3.2.1.11	4.67	9.89E-05	1.17.4.2	7.77	1.39E-02
2.7.1.195	2.84	1.10E-04	3.1.7.2	0.9	1.46E-02
3.2.1.4	-1.93	1.21E-04	3.6.3.8	0.64	1.60E-02
1.17.4.1	1.27	1.42E-04	2.7.1.197	0.91	1.68E-02
2.4.1.20	-1.07	1.54E-04	2.1.1.13	-0.9	1.72E-02
3.2.1.18	3.21	1.83E-04	3.2.1.3	-0.98	1.82E-02
2.3.2.3	3.4	6.38E-04	4.1.1.38	3.34	1.92E-02
3.2.1.97	2.16	6.51E-04	2.4.1.279	-3.08	2.52E-02
1.1.98.6	1.29	8.04E-04	6.2.1.1	-1.64	2.67E-02
3.2.1.187	3.69	1.30E-03	2.7.1.207	2.18	2.68E-02
3.4.21.96	3.27	1.82E-03	3.2.1.41	-0.68	2.81E-02
5.4.99.2	-0.92	3.88E-03	4.2.1.162	1.03	3.18E-02
1.1.5.12	0.83	4.82E-03	3.4.24.70	-0.97	3.24E-02
3.4.14.12	-1.09	5.64E-03	1.97.1.2	-3.06	3.39E-02
3.2.1.68	1.8	5.82E-03	3.4.11.2	2.64	3.59E-02
3.2.1.8	-1.36	6.40E-03	3.6.3.4	0.93	3.60E-02
4.1.1.32	-1.89	7.20E-03	4.2.2.8	-0.58	3.99E-02
1.2.3.3	2.84	7.58E-03	4.2.1.135	-0.67	4.52E-02
1.2.4.1	1.75	7.76E-03	1.4.3.21	1.86	4.89E-02
6.2.1.36	2.69	1.09E-02	2.4.1.8	3.01	4.94E-02

Table 3.24: Pathways identified as significantly variable between CD patients and healthy controls in the CD-HMP data set using mi-faser-generated functional profiles. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and absolute logFC ≥ 0.11.

Category	**Name**	**logFC**	**adj *p*-value**
AA	D-Glutamine and D-glutamate metabolism	0.33	1.38E-02
L	Glycerolipid metabolism	0.27	1.86E-02
T	Geraniol degradation	0.43	2.89E-02
L	Fatty acid elongation	0.43	3.05E-02
X	Caprolactam degradation	0.51	3.39E-02

[*] Here, L = Lipid Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); X = Xenobiotics Biodegradation and Metabolism; T = Metabolism of Terpenoids and Polyketides.

Table 3.25: Pathways identified as significantly variable between CD patients and healthy controls in the CD-Swedish data set using mi-faser-generated functional profiles. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value $<$ 0.05 and absolute logFC ≥ 0.11.

Category	**Name**	**logFC**	**adj *p*-value**
AA	Glycine, serine and threonine metabolism	-1.12	5.08E-05
AA	Glutathione metabolism	1.23	7.22E-05
X	Drug metabolism - other enzymes	0.95	3.12E-04
V	Pantothenate and CoA biosynthesis	-0.52	2.16E-03
AA	Valine, leucine and isoleucine biosynthesis	-0.52	2.16E-03
AA	Valine, leucine and isoleucine degradation	-0.92	3.88E-03
V	Thiamine metabolism	-0.43	5.92E-03
T	Terpenoid backbone biosynthesis	-0.44	1.38E-02
C	Amino sugar and nucleotide sugar metabolism	-0.23	1.94E-02
C	Galactose metabolism	0.58	2.24E-02
	Biosynthesis of ansamycins	0.43	2.45E-02
C	Citrate cycle (TCA cycle)	0.51	3.35E-02
AA	beta-Alanine metabolism	1.86	4.89E-02
AA	Tyrosine metabolism	1.86	4.89E-02
SM	Isoquinoline alkaloid biosynthesis	1.86	4.89E-02

[*] Here, C = Carbohydrate Metabolism; L = Lipid Metabolism; E = Energy Metabolism; N = Nucleotide Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); SM = Biosynthesis of Secondary Metabolites; G = Glycan Biosynthesis and Metabolism; V = Metabolism of Co-factors and Vitamins; X = Xenobiotics Biodegradation and Metabolism; GI = Genetic Information Processing; T = Metabolism of Terpenoids and Polyketides.

Table 3.26: Significant differentially abundant ECs identified by HUMAnN2 in the CD-HMP data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.58.

EC	**logFC**	**Adjusted *p*-value**	**EC**	**logFC**	**Adjusted *p*-value**
3.5.1.104	-2.00	3.33E-03	2.6.1.57	-1.46	8.29E-03
2.4.1.329	-7.09	3.93E-03	3.2.1.70	-1.10	2.69E-02
2.7.8.36	-0.61	4.91E-02	3.2.1.151	-1.01	1.77E-02
3.2.1.37	-0.66	4.95E-02	6.3.1.12	-2.76	4.87E-02
4.2.2.26	-9.15	2.08E-02	2.4.99.16	-2.07	4.49E-02
1.18.1.3	-0.59	2.39E-02	3.6.3.40	-0.80	4.05E-02
2.3.1.35	-2.11	4.31E-02	1.3.1.n3	-0.67	1.36E-02
4.2.1.77	-3.34	3.19E-03	2.4.1.342	-2.29	2.61E-02
4.1.1.96	-1.60	1.42E-02	2.3.1.89	-1.74	3.14E-02
2.7.1.162	-1.58	2.24E-02	1.1.1.3	-1.31	1.19E-02

Table 3.27: Significant differentially abundant ECs identified by HUMAnN2 in the CD-Swedish data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.58.

EC	**logFC**	**Adjusted *p*-value**	**EC**	**logFC**	**Adjusted *p*-value**
1.17.4.1	1.19	7.92E-03	2.3.3.9	1.46	9.31E-03
4.2.1.53	0.83	4.89E-02	6.2.1.1	-0.87	3.84E-02
2.4.1.12	1.06	2.18E-02	2.4.1.8	4.62	2.23E-03
3.2.1.187	2.41	6.36E-03	2.4.1.5	1.98	2.24E-02
3.2.1.185	2.07	1.01E-02	1.4.1.2	-1.59	3.65E-04
3.2.1.20	1.63	3.78E-03	1.2.5.1	3.34	6.86E-03
3.2.1.21	-1.42	1.09E-02	6.3.5.5	0.74	1.53E-02
3.2.1.28	-1.19	4.77E-02	1.2.4.2	1.31	4.35E-02
2.2.1.9	3.44	4.59E-03	1.3.5.4	1.65	3.98E-02
6.1.1.19	1.01	3.03E-02	2.4.1.211	1.38	3.13E-03
6.1.1.5	1.25	2.86E-02	3.2.1.170	0.68	2.61E-02
3.6.3.8	1.43	3.89E-02	4.1.1.31	2.12	9.31E-03
1.8.4.13	1.41	1.98E-02			

Table 3.28: Pathways identified as significantly variable between CD patients and healthy controls in the CD-HMP data set using HUMAnN2-generated functional profiles. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value $<$ 0.05 and absolute logFC $\geq$ 0.11.

Category	**Name**	**logFC**	**adj *p*-value**
X	Biosynthesis of vancomycin group antibiotics	-0.47	1.65E-02
X	Polycyclic aromatic hydrocarbon degradation	-0.54	3.56E-02
V	One carbon pool by folate	-0.13	3.76E-02
L	Glycerolipid metabolism	0.34	4.84E-02

* Here, L = Lipid Metabolism; V = Metabolism of Co-factors and Vitamins; X = Xenobiotics Biodegradation and Metabolism;

Table 3.29: Pathways identified as significantly variable between CD patients and healthy controls in the CD-Swedish data set using HUMAnN2-generated functional profiles. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value $<$ 0.05 and absolute logFC $\geq$ 0.11.

Category	**Name**	**logFC**	**adj *p*-value**
AA	Arginine biosynthesis	-1.59	3.65E-04
AA	Taurine and hypotaurine metabolism	-1.59	3.65E-04
AA	Arginine and proline metabolism	-1.27	1.81E-03
AA	Glutathione metabolism	1.11	4.42E-03
C	Starch and sucrose metabolism	-0.98	4.55E-03
V	Ubiquinone and other terpenoid-quinone biosynthesis	3.44	4.59E-03
GI	Two-component system	-0.49	7.79E-03
AA	Cyanoamino acid metabolism	-1.42	1.09E-02
C	Pentose and glucuronate interconversions	0.96	1.56E-02
SM	Phenylpropanoid biosynthesis	-1.36	2.43E-02
X	Drug metabolism - other enzymes	0.82	3.44E-02
C	Glyoxylate and dicarboxylate metabolism	0.26	3.46E-02
L	Fatty acid biosynthesis	2.47	3.98E-02
E	Nitrogen metabolism	-0.84	4.71E-02

* Here, C = Carbohydrate Metabolism; L = Lipid Metabolism; E = Energy Metabolism; N = Nucleotide Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); SM = Biosynthesis of Secondary Metabolites; G = Glycan Biosynthesis and Metabolism; V = Metabolism of Co-factors and Vitamins; X = Xenobiotics Biodegradation and Metabolism; GI = Genetic Information Processing; T = Metabolism of Terpenoids and Polyketides.

Table 3.30: Significant differentially abundant ECs identified by Kraken2 in the CD-HMP data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.58.

EC	**logFC**	**Adjusted *p*-value**	**EC**	**logFC**	**Adjusted *p*-value**
2.4.1.161	-1.54	2.76E-02	1.14.13.92	-1.74	1.02E-02
2.1.1.44	-1.89	1.89E-02	3.1.4.3	-1.09	2.10E-02
1.2.3.3	-1.43	4.77E-02	1.14.14.5	0.98	6.07E-03
1.14.13.70	-2.36	4.49E-02	6.3.1.20	0.79	1.65E-02
1.3.99.28	-1.77	4.21E-02	2.7.7.43	1.70	2.54E-02
3.2.1.183	-1.97	1.24E-02	3.5.4.9	-1.27	1.84E-02
1.3.1.86	-3.18	4.23E-03	2.8.1.15	-2.12	1.15E-02
3.2.2.26	-1.21	1.43E-02	2.4.1.15	0.59	2.42E-03
3.2.1.28	0.87	2.46E-02	4.2.3.155	-1.43	7.85E-03
1.11.2.4	-1.14	5.23E-03	1.14.13.154	2.87	3.81E-02
2.7.4.6	0.64	2.42E-03	6.3.5.11	-1.06	1.87E-02
3.2.1.31	0.69	2.60E-02	3.2.1.156	-1.09	1.30E-02
3.2.1.37	-0.70	2.38E-02	3.1.1.74	-0.76	4.34E-02
3.2.1.35	-1.66	9.80E-03	1.1.1.374	-1.48	2.45E-02
1.5.1.43	-1.39	4.49E-02	2.7.7.62	-0.99	2.11E-02
2.4.1.109	-3.89	2.33E-02	1.3.5.4	0.71	1.75E-02
5.5.1.25	3.50	4.44E-02	4.1.1.1	-2.75	2.25E-02

Table 3.31: Significant differentially abundant ECs identified by Kraken2 in the CD-Swedish data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.58.

EC	**logFC**	**Adjusted *p*-value**	**EC**	**logFC**	**Adjusted *p*-value**
6.3.5.1	-1.05	2.01E-05	3.2.1.141	1.49	8.36E-03
4.1.1.31	1.41	3.32E-05	3.2.1.20	1.09	9.19E-03
3.2.1.97	2.52	4.43E-05	4.1.1.38	2.83	1.13E-02
2.4.1.20	-1.34	4.43E-05	3.2.1.4	-0.58	1.25E-02
2.4.1.279	-5.14	4.48E-05	4.2.2.1	2.43	1.34E-02
2.7.1.197	1.58	5.07E-04	4.2.2.24	-1.17	1.55E-02
6.3.5.3	1.69	8.29E-04	3.2.1.135	2.64	1.78E-02
3.2.1.185	2.56	1.22E-03	1.17.4.2	7.40	1.80E-02
3.2.1.187	3.64	1.46E-03	3.2.1.170	2.04	1.80E-02
1.1.1.40	-0.83	1.91E-03	3.2.1.177	-0.72	1.94E-02
2.2.1.7	-0.59	2.95E-03	3.2.1.1	-0.84	1.97E-02
2.6.1.97	2.70	3.02E-03	3.4.24.68	0.85	2.03E-02
3.2.1.11	1.04	3.14E-03	1.4.3.21	2.14	2.50E-02
1.1.5.12	-2.17	3.39E-03	3.2.1.131	-0.62	2.72E-02
1.2.3.3	2.38	4.32E-03	4.2.2.23	-1.77	2.83E-02
2.4.1.247	-1.42	4.76E-03	1.1.1.39	-0.70	2.98E-02
1.1.98.6	1.14	4.80E-03	3.2.1.21	-0.61	4.33E-02
1.8.7.1	2.19	4.82E-03	1.2.4.1	0.74	4.41E-02
3.6.3.8	0.77	5.29E-03	4.2.1.3	0.92	4.71E-02
2.2.1.1	0.73	6.26E-03	2.4.1.288	1.89	4.89E-02
4.3.1.24	1.83	6.40E-03	4.1.1.32	-2.10	4.90E-02
5.4.99.2	-0.93	6.54E-03			

Table 3.32: Pathways identified as significantly variable between CD patients and healthy controls in the CD-HMP data set using Kraken2-generated functional profiles. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and absolute logFC ≥ 0.11.

Category	**Name**	**logFC**	**adj *p*-value**
T	Biosynthesis of 12-, 14- and 16-membered macrolides	2.25	1.13E-02
AA	Phenylalanine, tyrosine and tryptophan biosynthesis	-0.13	1.15E-02
L	Ether lipid metabolism	-1.46	1.25E-02
X	Chlorocyclohexane and chlorobenzene degradation	1.34	2.20E-02
SM	Phenazine biosynthesis	-0.22	2.20E-02
SM	Flavone and flavonol biosynthesis	0.69	2.60E-02
AA	beta-Alanine metabolism	-0.11	2.60E-02
GI	Aminoacyl-tRNA biosynthesis	-0.11	3.39E-02

* Here, AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); L = Lipid Metabolism, SM = Biosynthesis of Secondary Metabolites; X = Xenobiotics Biodegradation and Metabolism; GI = Genetic Information Processing; T = Metabolism of Terpenoids and Polyketides.

Table 3.33: Pathways identified as significantly variable between CD patients and healthy controls in the CD-Swedish data set using Kraken2-generated functional profiles. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value $<$ 0.05 and absolute logFC ≥ 0.11.

Category	**Name**	**logFC**	**adj *p*-value**
T	Terpenoid backbone biosynthesis	-0.59	2.95E-03
T	Biosynthesis of ansamycins	0.73	6.26E-03
AA	Valine, leucine and isoleucine degradation	-0.93	6.54E-03
C	Phosphotransferase system (PTS)	0.5	7.79E-03
C	Pentose phosphate pathway	0.67	1.02E-02
C	Starch and sucrose metabolism	-0.27	1.52E-02
C	Amino sugar and nucleotide sugar metabolism	-0.27	1.60E-02
C	Pantothenate and CoA biosynthesis	-0.4	2.45E-02
AA	Valine, leucine and isoleucine biosynthesis	-0.4	2.45E-02
AA	beta-Alanine metabolism	2.14	2.50E-02
AA	Tyrosine metabolism	2.14	2.50E-02
SM	Isoquinoline alkaloid biosynthesis	2.14	2.50E-02
AA	Glycine, serine and threonine metabolism	2.14	2.50E-02
E	Carbon fixation in photosynthetic organisms	-0.23	2.57E-02
V	Thiamine metabolism	-0.41	2.94E-02
AA	Cyanoamino acid metabolism	-0.61	4.33E-02
C	Pyruvate metabolism	-0.19	4.33E-02

* Here, C = Carbohydrate Metabolism; E = Energy Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); SM = Biosynthesis of Secondary Metabolites; V = Metabolism of Co-factors and Vitamins; T = Metabolism of Terpenoids and Polyketides.

3.2.3 Enabling accurate patient vs control classification using functional metagenomic markers

Patients and controls in case-control cohorts of type-2 diabetes (T2D), Crohn's disease (CD), and Parkinson's disease (PD) can be classified with much higher accuracy using the enzyme markers (EC terms) identified by Carnelian, implying that Carnelian's additional labeling of unalignable reads is meaningful. To test the power of significantly variable EC terms in discriminating patients from controls in the disease data sets, we performed N-fold cross-validation experiments (T2D: 10-fold, CD: 5-fold, and PD: 5-fold). In each trial, ECs exhibiting significant differences in terms of relative abundances between patients and controls in the training partition (Wilcoxon rank-sum test p-value < 0.05) were selected as features input to a set of random forest classifiers and accuracy was measured on the test partition.

In the Chinese T2D cohort, with Carnelian-identified differentially abundant ECs we achieved an average area under the curve (AUC) of 0.75 (95% CI: 0.69 - 0.82), whereas using the ECs identified by mi-faser, HUMAnN2 and Kraken2, average AUCs of 0.69, 0.63, and 0.63 were achieved, respectively (Figure 3-2(a)). In discriminating European T2D patients from NGT individuals, we achieved an average AUC of 0.72 (95% CI: 0.61 - 0.82) with Carnelian-identified ECs which is significantly higher than the other three methods (Figure 3-2(b)).

In the CD cohort from the US, we achieved an average AUC of 0.73 (95% CI: 0.56 - 0.89) with Carnelian-identified differentially abundant ECs, whereas using the differentially abundant ECs identified by mi-faser, HUMAnN2 and Kraken2, average AUCs of 0.61, 0.54, and 0.55 were achieved, respectively (Figure 3-2(c)). In discriminating Swedish CD patients from the healthy controls, we achieved an average AUC of 0.95 (95% CI: 0.89 - 1.00) with Carnelian-identified variable ECs which is significantly higher than the other three methods (Figure 3-2(d)). In the PD cohort, Carnelian-identified markers achieved an AUC of 0.85 (95% CI: 0.72 to 0.98) in discriminating between patients and healthy controls, whereas the differentially abundant EC terms found by other methods did not achieve more than 0.75 average AUC (Figure 3-2(e)).

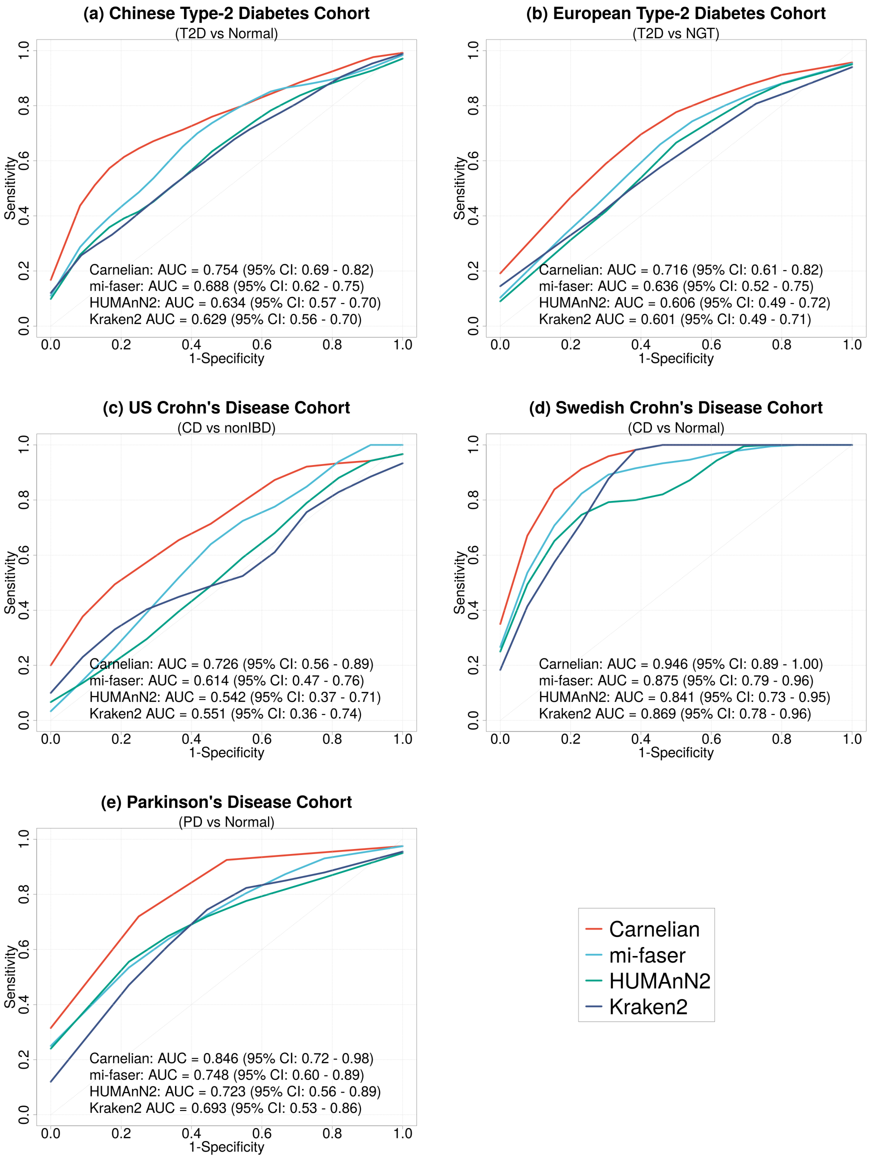


Figure 3-2: **Classification of patients vs controls using Enzyme Commission (EC) markers (N-fold cross-validation experiments).** **(a)** T2D vs controls in the T2D-Qin data set (Chinese cohort); **(b)** T2D vs Normal Glucose Tolerance (NGT) individuals in the T2D-Karlsson data set (European cohort). **(c)** CD patients vs controls in the CD-HMP data set (individuals from the US). **(d)** CD patients vs healthy individuals in the CD-Swedish data set (Swedish twin studies). **(e)** PD vs controls in the PD-Bedarf data set. In each trial, one of the N subsets was selected as the test set and the rest $N-1$ subsets were used as the training set. Differentially abundant ECs were selected from the training set as features input to a set of random forest classifiers. Performance of classification was measured on the test set. Carnelian-identified EC terms achieve a larger average area under the curve (AUC) in all the cases compared to those identified by other methods.

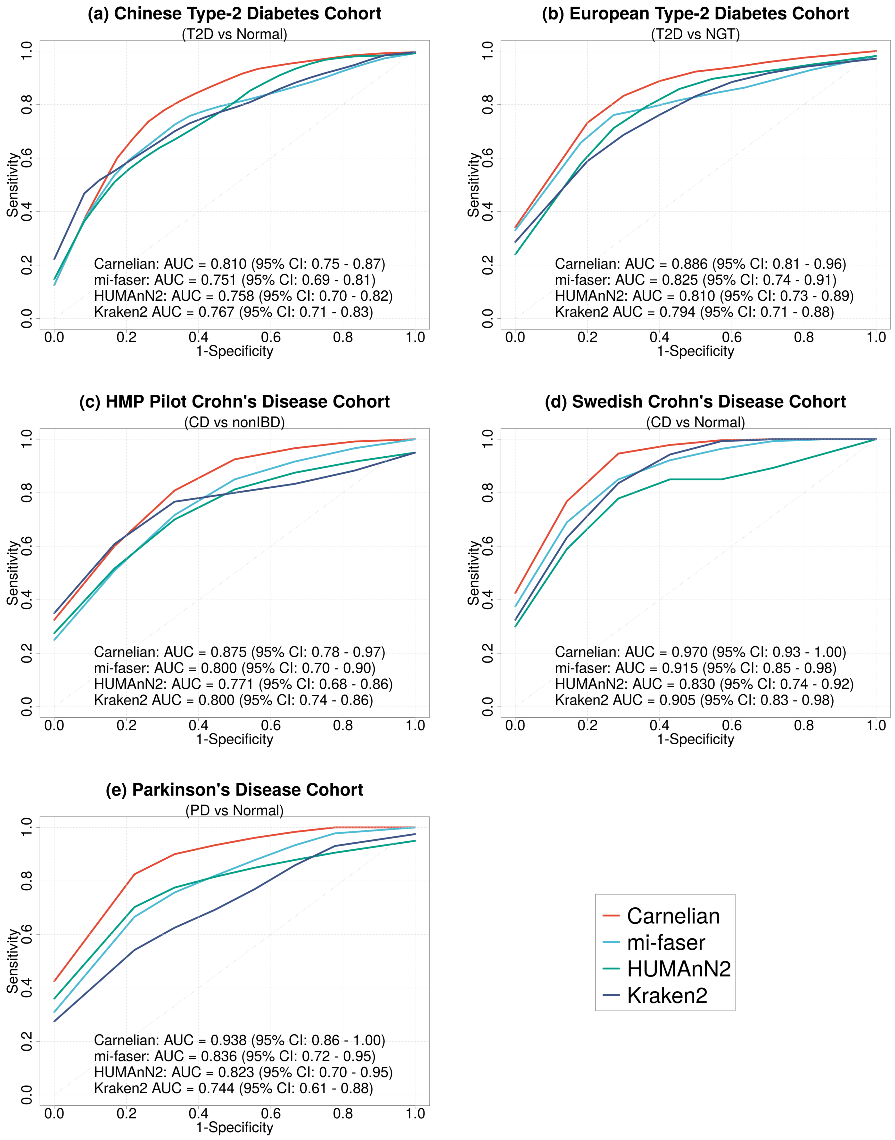


Figure 3-3: **Classification of patients vs. controls using Enzyme Commission (EC) markers selected from the entire data set (N-fold cross-validation experiments).** **(a)** T2D vs controls in the T2D-Qin data set (Chinese cohort); **(b)** T2D vs. Normal Glucose Tolerance (NGT) individuals in the T2D-Karlsson data set (European cohort). **(c)** CD patients vs. controls in the CD-HMP data set (individuals from the US). **(d)** CD patients vs. healthy individuals in the CD-Swedish data set (Swedish twin studies). **(e)** PD vs controls in the PD-Bedarf data set. Differentially abundant ECs were selected from the entire data set as features input to a set of random forest classifiers. Average area under the curve (AUC) over all cross-validation trials is reported as a measure of accuracy. Carnelian-identified EC markers achieve a larger area under the curve (AUC) in all the cases compared to those identified by mi-faser, HUMAnN2, and Kraken2.

Note that while it is common practice in the metagenomics literature to select classification features from the entire data set, even when running cross-validation experiments [182,183], in all our cross-validation analyses, we instead followed standard machine learning best-practices and avoid information leakage in feature selection by choosing EC labels only from the training sets. For completeness, we also performed the classification experiments choosing EC labels from the entire data set. Using this experimental design, Carnelian-generated ECs again achieved higher accuracy compared to the other three methods in all the study cohorts (Figure 3-3).

To test for generalizability of Carnelian-identified ECs in the CD and T2D cohorts, we combined the EC markers identified in the geographically separated cohorts and redid the classification of patients vs. controls. For CD, using the unified ECs identified by Carnelian as features, we could achieve ∼0.94 AUC on average, whereas, the combined ECs identified by other tools achieved average AUCs between 0.79 and 0.88 (Figure 3-4). For T2D, with the unified ECs from Carnelian as features, we were able to classify the functional profiles of T2D patients and controls with an average AUC of ∼0.80, whereas using the combined ECs identified by other methods in both cohorts as features, the average AUCs remained between 0.73 and 0.76 (Figure 3-5). The lists of combined EC markers for T2D and CD identified by Carnelian are provided in Tables 3.34 and 3.35.

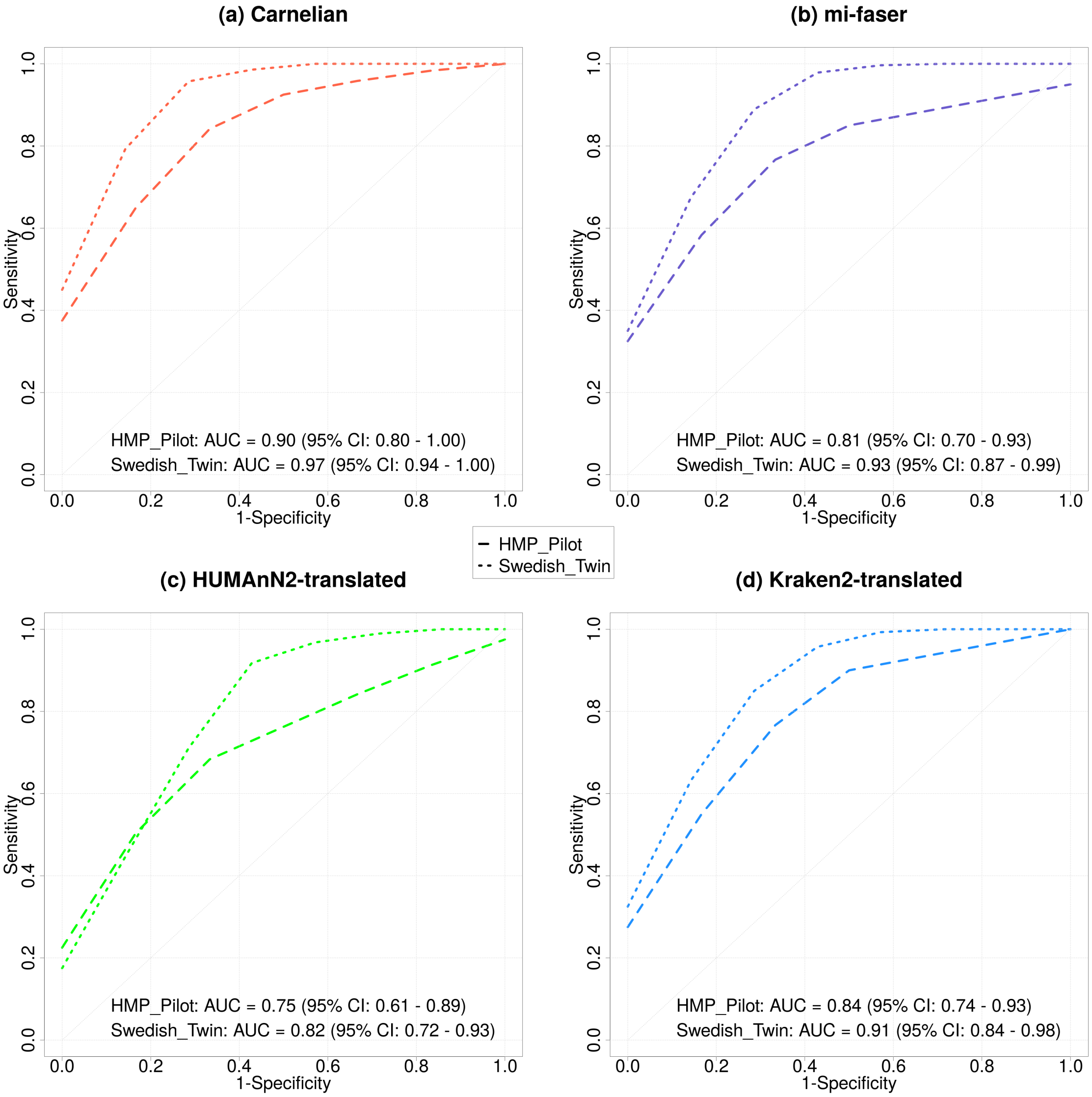


Figure 3-4: **Classification of patients vs. controls in Crohn's disease cohorts from the US and Sweden using the combined set of markers identified by Carnelian, mi-faser, HUMAnN2-translated, and Kraken2-translated.** Using Carnelian-identified significant ECs, we can consistently achieve ≥ 0.90 area under the curve (AUC) on average on both CD-HMP and CD-Swedish data sets which is higher than the other three tools.

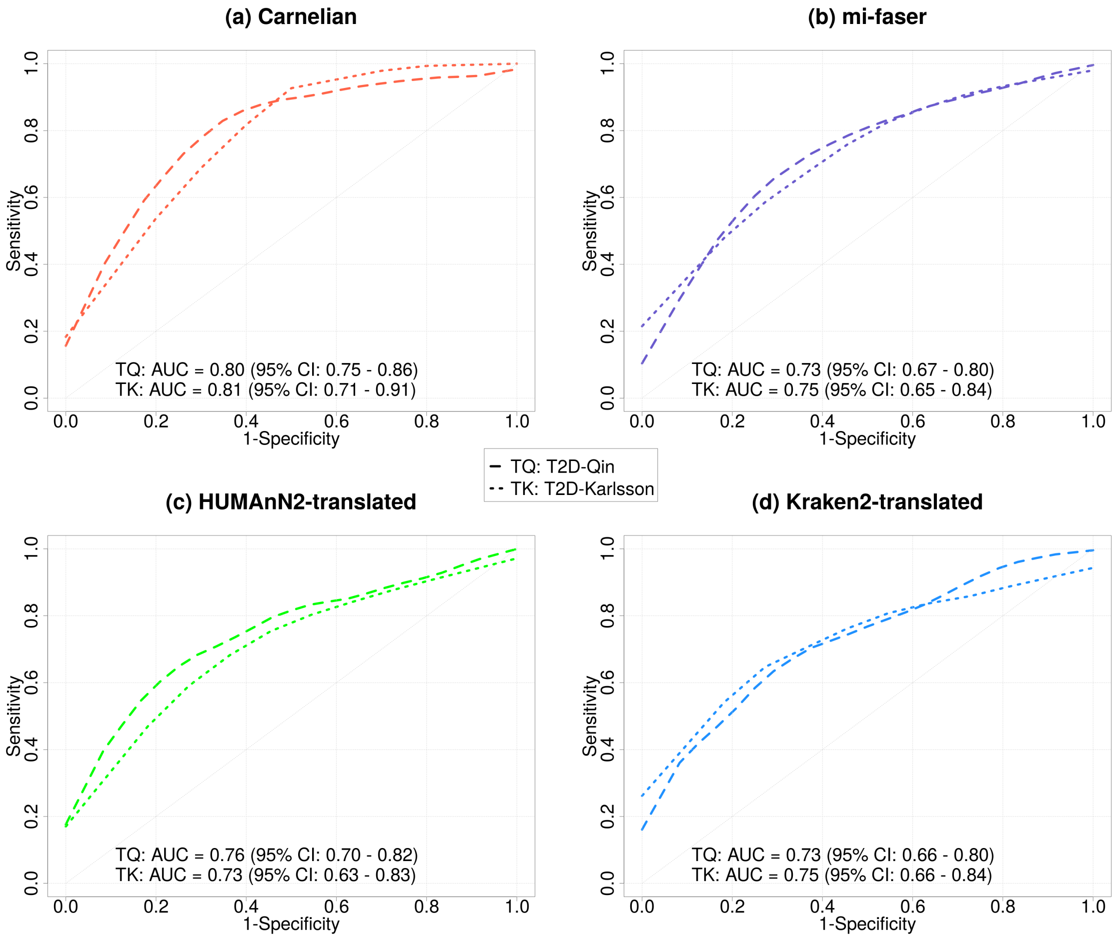


Figure 3-5: **Classification of patients vs. controls in Chinese and European T2D cohorts using the combined set of markers identified by Carnelian, mi-faser, HUMAnN2-translated, and Kraken2-translated.** Using Carnelian-identified highly variable ECs, we can consistently achieve ~ 0.80 area under the curve (AUC) on average on both T2D-Qin and T2D-Karlsson data sets, whereas using the highly variable ECs identified by other methods, an average AUC of 0.73-0.76 can be achieved on both data sets.

Table 3.34: Combined EC markers identified by Carnelian that can classify T2D patients vs. controls in both Chinese and European population with ∼80% area under the curve on average.

5.4.99.62	1.17.7.4	4.2.1.147	3.1.3.12	2.8.4.1
3.6.1.23	1.17.7.3	2.4.1.7	1.12.98.1	1.2.99.7
2.4.2.2	1.8.2.3	5.4.3.2	6.2.1.44	2.4.1.15
2.4.2.6	1.1.1.28	3.1.3.85	2.3.2.21	6.2.1.3
1.1.1.100	3.2.1.52	2.4.1.329	5.4.99.20	4.2.1.119
2.7.8.35	2.4.1.1	4.1.1.101	2.1.1.10	3.4.13.9
3.7.1.8	1.4.1.24	4.2.99.20	5.3.1.22	3.1.4.16
2.7.2.4	1.4.1.4	3.1.3.8	1.5.1.36	3.1.4.12
1.13.11.27	2.6.1.84	2.6.1.113	4.2.1.42	2.1.1.90
4.1.1.33	2.7.7.61	3.2.2.23	1.1.1.251	3.2.1.80
4.2.1.20	4.2.1.120	4.2.2.n2	4.1.2.48	2.4.1.182
2.7.1.220	5.4.2.11	4.1.1.79	6.3.2.33	1.13.11.3
1.1.1.408	4.3.1.15	1.14.13.127	1.8.98.1	6.3.2.36
1.12.2.1	1.3.1.70	3.5.1.5	2.3.1.5	1.13.11.6
1.8.4.14				

Table 3.35: Combined EC markers identified by Carnelian that can classify CD patients vs. controls in both the US and Swedish population with ∼94% area under the curve on average.

2.4.1.292	3.1.3.74	5.4.99.2	3.2.1.35	3.2.1.4
1.10.3.10	3.4.23.49	1.8.98.3	4.1.1.18	2.4.1.19
2.7.7.39	3.5.1.16	3.2.1.3	2.2.1.7	2.4.99.21
4.2.1.12	4.1.1.65	2.4.2.46	2.3.1.41	3.2.1.52
1.2.1.19	2.7.1.55	3.2.1.21	3.2.1.177	1.13.11.61
1.3.3.3	3.2.1.31	3.2.1.14	3.4.24.55	3.2.1.176
4.3.1.15	2.4.1.12	2.3.2.27	6.3.4.14	2.7.1.195
1.17.5.3	3.1.4.14	4.2.1.135	2.4.1.288	3.1.11.5
3.2.1.28	5.3.1.26	3.2.1.18	1.3.1.31	4.2.1.9
1.1.1.60	1.17.1.9	2.4.2.48	4.3.1.24	3.1.21.3
3.2.2.21	1.8.5.5	5.99.1.2	2.3.2.3	1.2.7.6
3.1.1.41	2.4.1.20	3.2.1.133	3.6.3.2	1.2.4.2
3.2.2.8	3.6.3.5	3.1.26.12	1.2.7.4	2.7.1.193
4.2.1.42	1.4.1.2	3.2.1.169	1.2.99.7	4.1.99.17
2.4.2.52	2.4.2.47	3.2.1.131	1.14.13.171	1.8.7.1
2.7.7.19	1.1.1.40	3.1.7.2	4.1.2.27	3.4.21.72
1.3.1.101	3.6.3.42	1.4.7.1	6.1.1.10	1.7.7.1
2.7.1.186	2.4.1.52	1.1.1.39	4.6.1.1	3.2.1.8
5.1.3.26	2.7.9.1	3.6.3.4	1.2.7.5	3.2.1.1
3.6.1.25	1.4.4.2	4.1.1.32	2.4.1.247	5.4.99.15
4.2.1.40	3.4.24.69	3.2.1.41	6.3.5.4	3.2.1.187
5.3.3.10	2.4.1.25	6.1.1.18	2.7.8.47	4.2.1.82
2.7.7.61	3.4.21.53	2.4.1.9	6.2.1.51	

3.2.4 Uncovering functional relatedness of diverse industrialized and non-industrialized microbiomes

In addition to finding trends in functional changes across disease cohorts, Carnelian enables us to compare the functional potential of healthy human gut microbiomes from industrialized and non-industrialized communities. We analyzed the fecal microbiomes of 84 individuals from Boston with an urban lifestyle (industrialized society; unpublished data set from Alm lab), 35 hunter-gatherer Baka individuals from Cameroon (unpublished data set from Alm lab), 50 non-industrialized individuals from Gimbichu, Ethiopia [26], and 112 individuals of Betsimisaraka and Tsimihety ethnicities from Madagascar [26]. The expectation is that healthy individuals across populations ought to share similar core metabolic pathways [25, 197]. Carnelian's analyses met this expectation, finding pathway-level similarity in core metabolic functionality of both the industrialized and non-industrialized communities.

Using our curated EC database, Carnelian detects more ECs compared to other methods (Table 3.36) and finds slightly higher diversity of identified ECs in the non-industrialized communities compared to the industrialized community indicated by the Shannon-Wiener diversity index (Figure 3-6(a)). At the pathway level, the diversity of identified functionality in both communities is comparable, as hoped (Figure 3-6(b)). At both levels, Carnelian captures significantly more diversity than the other three methods (Figure 3-6). Importantly, the fecal microbiomes of Baka individuals from Cameroon could not be characterized well, even running the full HUMAnN2 pipeline using its default databases (Appendix B). Despite incorporating taxonomic information, out-of-the-box HUMAnN2 could map on average $\sim$ 10% of the reads and detect less than 30 known species and 996 ECs per sample (Shannon diversity index for ECs: 5.58) (Appendix B: Table B.1).

Principal component analysis of the EC profiles generated by Carnelian shows a marked separation by population (Figure 3-7(a)). Mean EC profiles of industrialized and non-industrialized microbiomes show a moderate degree of correlation (Kendall's $\tau = 0.75$). Much of this separation washes away at the pathway level (Figure 3-7(b));

mean pathway profiles of industrialized and non-industrialized microbiomes show a high degree of correlation (Kendall's $\tau = 0.93$). This finding suggests a high degree of pathway-level functional similarity between industrialized and non-industrialized healthy microbiomes—which was not observed by earlier studies.

Table 3.36: Performance of Carnelian, mi-faser, HUMAnN2, and Kraken2 on population data sets from Boston, Cameroon, Ethiopia, and Madagascar. Carnelian annotates significantly more reads and identifies more ECs compared to mi-faser, HUMAnN2, and Kraken2 in all four data sets.

		Boston **84** **individuals**	**Cameroon** **35** **individuals**	**Ethiopia** **50** **individuals**	**Madagascar** **112** **individuals**	**Industrialized** **(B)**	**Non-industrialized** **(CEM)**
Carnelian	# Annotated reads per sample	1,430,026	269,720	2,182,173	2,157,741	1,430,026	1,828,507
	# ECs per sample	1981	2003	2002	2003	1981	2003
mi-faser	# Annotated reads per sample	743,877	268,684	1,181,031	1,527,111	743,877	1,215,695
	# ECs per sample	1230	1252	1309	1368	1230	1310
HUMAnN2	# Annotated reads per sample	83,131	21,383	281,640	819,424	83,131	541,147
	# ECs per sample	791	827	919	1064	791	937
Kraken2	# Annotated reads per sample	466,709	149,191	709,543	1,046,199	466,709	801,387
	# ECs per sample	1238	1219	1280	1233	1238	1244

In order to identify the ECs that characterize the separation for industrialized and non-industrialized population, we looked at the weights of the ECs in the first nine principal components, which together explain 80% of the variability among individuals (Table 3.37). The majority of ECs with high weights were involved in the carbohydrate, amino acid, nucleotide, and energy metabolism pathways. Using the highly weighted ECs, we performed Ward-linkage hierarchical clustering based on Pearson correlation coefficients of the EC profiles of the industrialized and non-industrialized individuals; we observed a clear separation of the two groups (Figure 3-8).

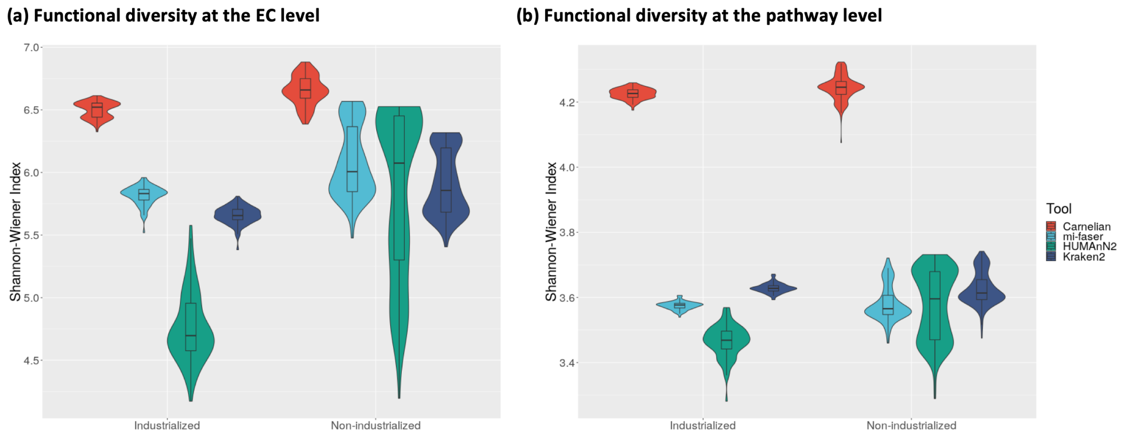


Figure 3-6: **Functional diversity found by Carnelian and other methods in the non-industrialized and industrialized microbiomes.** **(a)** Diversity at the EC-level. **(b)** Diversity at the pathway level. Carnelian identifies more diversity at both the levels than other methods as indicated by Shanon-Wiener indices.

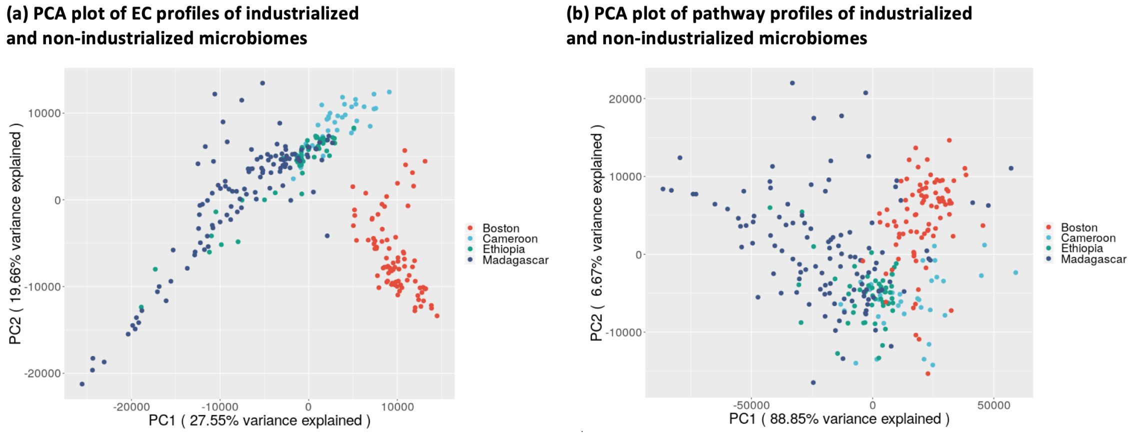


Figure 3-7: **Principal component analysis (PCA) plots depicting Carnelian-derived EC and pathway profiles of the microbiomes of non-industrialized and industrialized communities.** **(a)** PCA plot of EC profiles show separation of populations. **(b)** Pathway profiles of the microbiomes do not show much separation indicating a high degree of similarity between the microbial metabolic functionality of industrialized and non-industrialized communities.

Table 3.37: Weights of top 52 EC terms in the first nine principal components which cumulatively explain ∼80% variance in the principal component analysis of the EC profiles of the industrialized and non-industrialized microbiomes. These ECs are mostly involved in microbial carbohydrate, amino acid, nucleotide, and energy metabolism pathways.

EC	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**
4.1.1.11	0.19	0.17	0.04	0.31	-0.25	0.03	-0.08	0.24	-0.09
5.1.3.32	-0.18	-0.14	-0.08	0.15	0.02	-0.15	0.18	0.05	-0.1
2.7.7.24	0.16	0.1	-0.12	0.24	0.06	-0.02	0.12	-0.19	0.01
6.4.1.3	0.15	-0.11	0.05	0.01	0.01	0	-0.02	-0.03	-0.02
1.11.1.15	0.15	-0.28	0.1	-0.05	-0.03	-0.01	-0.02	-0.04	-0.17
1.16.3.2	0.15	-0.17	0.08	-0.03	-0.04	-0.04	-0.06	0.07	-0.07
4.6.1.12	0.15	0.1	-0.04	-0.03	-0.07	-0.06	0.04	0.19	0.01
2.7.1.11	0.14	-0.08	0.01	-0.03	0	-0.03	-0.03	0.06	0.02
1.1.1.100	0.13	0.02	-0.03	-0.02	0.02	-0.01	-0.02	0.09	0.03
2.5.1.47	0.13	0	-0.03	-0.02	-0.01	-0.02	0.04	0.09	0.02
1.7.1.15	-0.13	-0.08	-0.04	0.07	-0.01	-0.07	0.1	0.01	-0.08
2.6.1.83	0.13	0.02	0	0.07	0.08	-0.02	0	0.05	0.04
4.2.1.11	0.12	0.02	-0.02	-0.02	-0.01	-0.04	0.03	-0.01	-0.09
5.1.3.13	0.12	0.03	-0.04	0.09	0.04	0	0	-0.04	0.05
3.2.1.3	0.12	-0.04	0.01	0.12	0.01	0	0	-0.06	0.02
5.2.1.8	-0.11	-0.02	-0.03	0.06	-0.03	-0.03	-0.02	0	0.09
6.3.5.2	0.1	0	-0.02	-0.05	-0.01	-0.01	0.02	0	0.01
5.4.99.18	0.1	0	-0.01	0	0.05	0.01	0.02	-0.03	0
3.2.1.21	0.1	-0.07	0.03	0.02	0.03	0.02	-0.07	-0.04	0.05
3.1.4.52	-0.1	-0.07	-0.04	0.05	0	-0.06	0.07	0.01	-0.02
4.2.1.47	0.1	-0.06	0.02	0.04	0.01	-0.02	-0.01	0	0
3.5.99.10	0.1	0.01	-0.06	0	0.06	-0.11	0.05	0.25	-0.04
1.4.1.16	0.09	-0.03	0.01	0.06	-0.03	0.03	-0.01	0.01	0.01
4.1.1.49	0.09	0.02	-0.03	0.06	-0.01	0	0.03	0.02	0.04
1.14.99.48	-0.09	0.28	0.12	0.23	0.17	0.17	-0.52	0	-0.48
4.2.1.46	0.09	0.02	-0.03	0.06	0.01	0.03	0.11	-0.17	-0.12
3.6.1.23	-0.09	0.15	0.38	0.02	-0.42	-0.41	-0.24	-0.32	0.25
1.6.5.11	-0.09	-0.08	-0.01	0.05	0.05	-0.07	0.06	-0.01	0.01
3.6.3.14	0.09	0.09	-0.07	-0.11	0.07	0.02	0.01	0.03	0.16
3.4.21.92	0.09	0.2	0.11	-0.12	-0.16	-0.32	0.17	-0.22	-0.17
3.2.2.27	0.09	0.05	0	0	-0.03	-0.05	0.03	-0.05	-0.06
2.1.3.9	0.09	-0.04	0.01	0.05	0.01	0	0	-0.02	0.01
1.2.1.12	0.09	0.11	-0.06	-0.07	0.02	-0.15	0.09	0.03	-0.14
3.6.4.13	0.09	-0.07	-0.01	0.05	-0.02	0	0.01	-0.01	0.03
2.7.7.65	-0.08	-0.05	-0.03	0.04	0.01	-0.05	0.05	0.01	-0.01
3.5.99.6	0.08	-0.01	-0.01	0.09	0.06	-0.02	0.04	-0.08	0.01
1.1.1.22	0.08	-0.01	-0.01	0.06	0	-0.03	0.01	-0.01	-0.04
2.4.1.281	0.08	-0.02	0	0.05	0.02	0	0	-0.02	0.03
5.3.1.5	0.08	-0.08	0.04	-0.01	-0.02	0	-0.02	0.01	-0.02
2.7.1.90	0.08	-0.02	0.01	0.04	0.02	-0.02	0	-0.01	0.02
2.1.1.45	0.07	-0.07	0.02	-0.06	-0.12	0.07	0.06	-0.03	-0.12
2.7.4.22	0.07	0.04	-0.03	-0.05	-0.02	-0.03	0.04	0.02	-0.06
5.4.2.11	0.07	-0.12	0.05	-0.06	0.02	0.02	0	-0.11	-0.13
6.1.1.20	0.07	-0.06	0.01	-0.01	-0.01	0.01	-0.02	-0.02	0.01
1.11.1.1	0.07	0.01	0.02	-0.02	0.02	-0.07	-0.04	0.04	-0.01
2.7.7.6	0.07	0.12	-0.09	-0.19	-0.07	0	0.07	-0.08	-0.01
1.1.1.205	0.07	0.02	-0.02	0.02	0.01	0.01	0.03	-0.06	0
2.3.1.31	0.07	-0.02	0.03	-0.01	0.04	0.03	0.01	-0.05	-0.02
6.3.1.2	0.07	-0.04	0.01	0.01	0	0.09	-0.01	-0.08	0.04
6.3.2.1	0.07	0	-0.01	0.07	-0.04	0.03	0.03	0	0.03
1.15.1.2	0.06	0.32	-0.13	0.35	0.07	-0.06	0.17	-0.04	0.07
3.5.1.5	-0.06	0.04	0.03	-0.06	-0.19	0.26	0.07	-0.06	-0.03

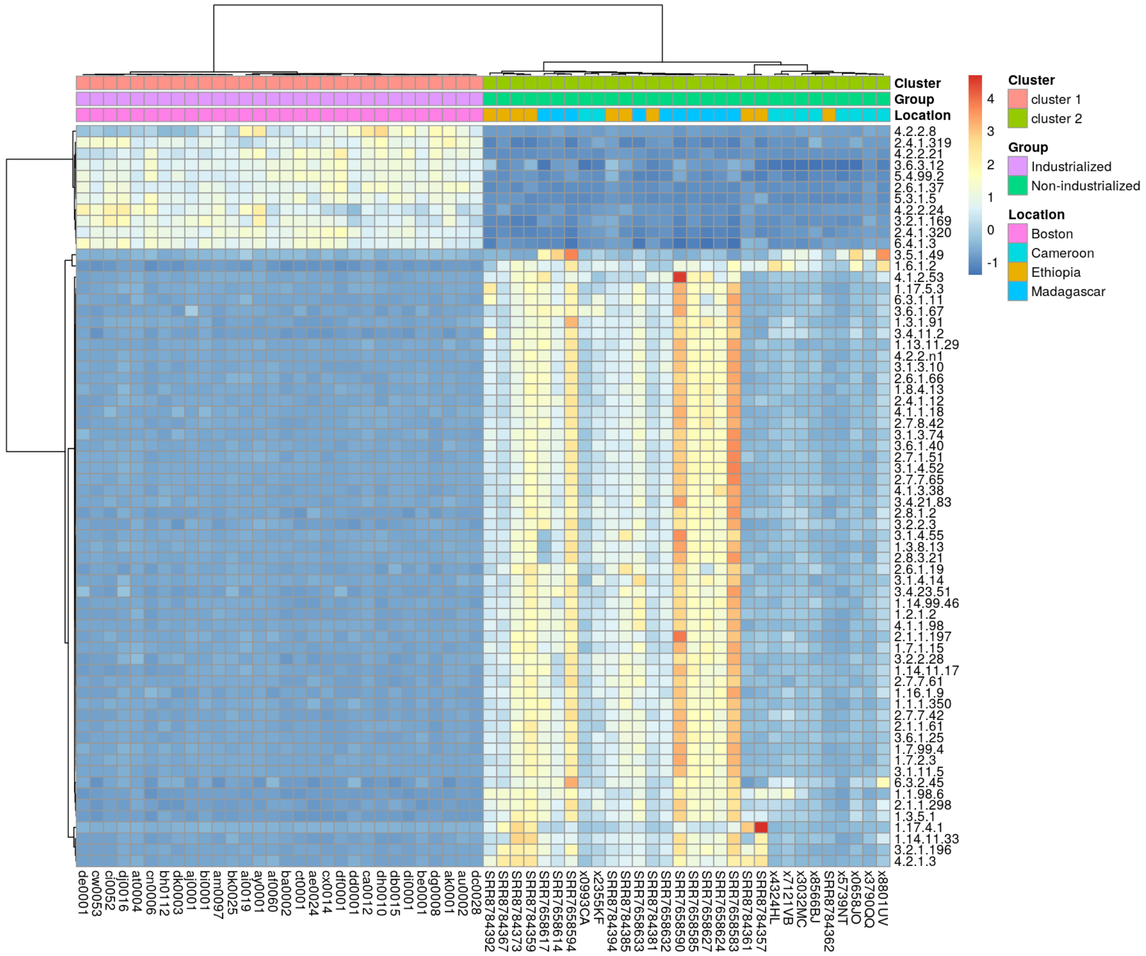


Figure 3-8: **Enzyme-level differences between the microbiomes of non-industrialized and industrialized communities (selected individuals).** The heatmap was generated using the z-scores of read abundances of the ECs with high weights in the top principal components of the EC profiles of industrialized and non-industrialized microbiomes. Ward-linkage hierarchical clustering of the EC profiles was performed using the Pearson correlation. The two top-level clusters found by hierarchical clustering perfectly capture the separation of non-industrialized and industrialized microbiomes. For display purposes, we show only individuals with read abundances falling outside one standard deviation of the mean in at least nine of the highly variable ECs. See Figure 3-9 for the corresponding heatmap and clustering of all individuals.

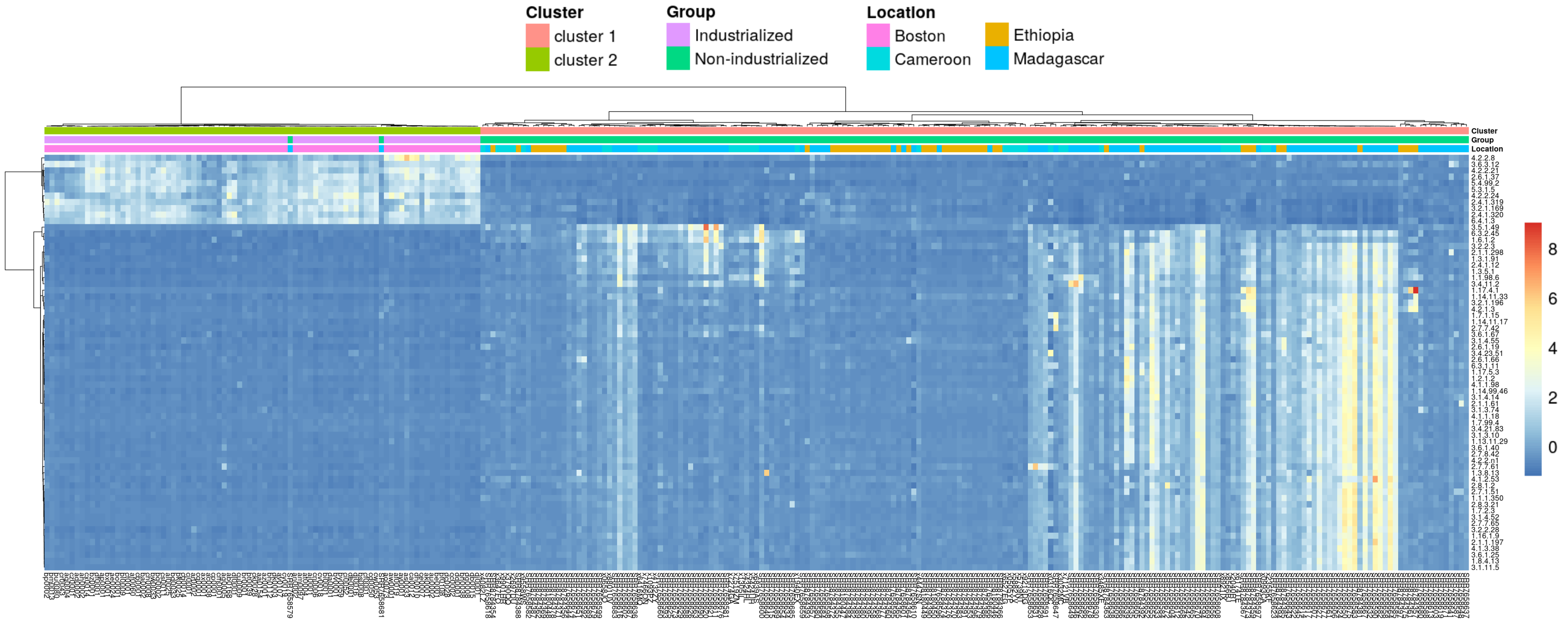


Figure 3-9: **Enzyme-level separation of non-industrialized and industrialized microbiomes (all individuals).** The figure shows a heatmap of the z-scores of relative abundances of ECs with high weights in the top principal components. Ward-linkage hierarchical clustering of the EC profiles (with highly weighted ECs in the PCA) of industrialized and non-industrialized microbiomes was performed using Pearson correlation. The two clusters found by hierarchical clustering capture the separation of non-industrialized and industrialized microbiomes perfectly except for the two outlier samples from Madagascar: SRR7658579 and SRR7658681.

Table 3.38: **Top 100 significant ECs identified by Carnelian in the industrialized (Boston) vs non-industrialized (Cameroon, Ethiopia, and Madagascar) communities as differentially abundant.** Significance thresholds used: BH corrected Wilcoxon ranksum test p-value < 0.05 and abs (log fold change) > 1. Fold change was calculated as the ratio of the mean EC abundance in non-industrialized communities to the EC abundance in the industrialized community. Carnelian identifies 454 differentially abundant ECs, whereas mi-faser, HUMAnN2, and Kraken2 identify 1009, 976, and 785 differentially abundant ECs respectively which covers > 80% of the ECs they identified in the population data sets; this indicates the presence of false positive hits among the reported ECs by other methods. Out of the 454 differentially abundant ECs identified by Carnelian, 284 were also reported significant by mi-faser, HUMAnN2, and Kraken2.

EC	**logFC**	**adjusted *p*-value**	**EC**	**logFC**	**adjusted *p*-value**	**EC**	**logFC**	**adjusted *p*-value**
4.2.2.21	-3.44	6.98E-37	2.7.7.61	2.06	8.20E-29	3.6.1.40	2.19	2.40E-28
5.3.1.5	-3.43	4.72E-37	2.6.1.19	2.06	3.81E-34	3.6.1.67	2.18	3.24E-29
2.6.1.37	-3.2	8.40E-37	1.14.11.33	2.06	8.65E-32	2.7.1.51	2.17	1.72E-26
1.7.1.15	3.16	1.41E-26	1.16.1.9	2.05	1.61E-27	3.6.1.25	2.16	3.21E-31
2.4.1.320	-2.82	2.61E-37	6.3.2.45	2.05	4.05E-29	4.1.1.18	2.16	3.50E-19
3.1.4.55	2.61	1.04E-29	4.2.1.3	2.01	1.13E-26	1.7.2.3	2.15	1.37E-26
5.4.99.2	-2.57	7.87E-37	2.8.1.2	2	1.85E-28	1.8.4.13	2.14	5.57E-24
4.2.2.24	-2.57	6.98E-37	3.5.4.1	2	5.47E-31	3.5.1.49	2.14	3.30E-23
3.1.4.52	2.55	3.56E-34	2.3.1.29	-1.99	6.98E-37	3.1.1.5	1.93	1.70E-26
2.8.3.21	2.53	2.14E-32	3.6.1.63	1.98	8.76E-32	1.1.1.346	1.91	4.24E-30
3.4.23.51	2.5	1.80E-24	2.7.4.23	1.98	2.86E-21	1.5.1.42	1.91	2.46E-21
3.1.3.10	2.49	8.73E-31	2.7.7.12	1.97	1.94E-35	3.4.16.4	1.91	1.96E-34
4.2.2.n1	2.48	1.81E-31	6.3.2.2	1.95	6.66E-34	2.3.1.15	1.91	1.94E-35
4.2.2.8	-2.46	4.60E-34	1.14.11.47	1.95	1.13E-30	1.2.7.3	-1.91	4.92E-37
1.3.8.13	2.45	2.22E-26	2.3.1.242	1.94	3.92E-28	1.3.5.3	1.91	1.72E-22
3.2.1.196	2.41	3.78E-30	6.2.1.30	-1.94	2.18E-36	3.1.21.1	1.9	3.41E-26
1.13.11.29	2.41	1.20E-29	2.4.1.319	-2.3	4.92E-37	2.1.1.265	1.89	2.29E-28
3.1.11.5	2.39	7.09E-36	2.1.1.298	2.3	1.03E-31	3.1.3.23	1.89	9.08E-25
2.1.1.197	2.38	3.46E-25	4.1.3.38	2.3	1.72E-30	2.7.1.73	1.87	1.54E-28
1.17.4.1	2.36	5.26E-32	3.1.3.74	2.29	1.94E-27	1.1.1.373	1.86	8.54E-33
2.1.1.61	2.36	4.96E-33	1.14.11.17	2.29	2.59E-31	4.2.1.80	1.86	1.79E-25
2.4.1.12	2.35	1.56E-33	3.4.21.83	2.29	2.43E-19	1.14.12.19	1.86	3.83E-22
4.1.1.98	2.34	6.68E-28	1.7.99.4	2.26	9.29E-19	1.2.1.10	1.85	1.30E-24
1.6.1.2	2.33	2.43E-35	6.3.1.11	2.25	1.60E-30	1.16.3.2	-1.85	4.95E-36
1.14.99.46	2.31	4.66E-33	1.1.98.6	2.25	2.68E-33	2.6.1.62	-1.85	1.86E-36
3.2.2.28	2.11	1.13E-30	3.2.2.3	2.25	5.53E-29	1.13.11.39	-1.85	2.74E-32
1.17.5.3	2.11	6.13E-21	1.3.1.91	2.24	1.15E-28	5.3.1.22	1.84	4.42E-21
2.7.7.42	2.11	1.12E-33	3.1.4.14	2.23	1.41E-23	1.8.5.3	1.84	1.11E-16
6.4.1.3	-2.1	2.61E-37	4.1.2.53	2.22	2.05E-33	2.3.1.193	1.83	2.74E-32
1.2.1.2	2.1	4.65E-33	3.4.11.2	2.21	6.99E-33	2.7.7.19	1.82	3.00E-27
1.1.1.350	2.09	2.25E-32	2.7.8.42	2.2	7.91E-34	3.2.1.170	1.82	3.14E-25
1.3.5.1	2.09	1.63E-35	2.7.7.65	2.2	7.95E-33	3.5.3.26	1.81	5.15E-19
3.6.3.12	-2.09	2.81E-36	2.6.1.66	2.2	4.66E-25	6.3.5.3	1.8	3.85E-33
3.2.1.169	-2.07	1.52E-36						

We also identified the significantly variable ECs between the two groups using a cutoff of BH-corrected Wilcox on rank-sum test p-value < 0.05 and absolute log fold change > 1 (Table 3.38). The differentially abundant ECs identified by Carnelian recapitulate the findings of earlier studies; those ECs match the microbial enzymatic functions related to differences in diet and lifestyle [26, 80, 198]. For example, fecal microbiota from the non-industrialized communities showed over-representation of several enzymes (exclusively identified by Carnelian) involved in the metabolism of fructose, mannose, starch, and sucrose. Examples include mannosyl-3-phosphoglycerate synthase (2.4.1.217), sucrose phosphorylase (2.4.1.7), phosphoglycerate mutase (5.4.2.11), phosphate propanoyltransferase (2.3.1.222), etc. On the other hand, fecal microbiota of industrialized individuals showed over-representation of simple sugar metabolizing enzymes, including ornithine aminotransferase (2.6.1.13), lysine 2,3-aminomutase (5.4.3.2), glycogenase (3.2.1.1), NADP-glucose-6-phosphate dehydrogenase (1.1.1.49), and phosphohexokinase (2.7.1.11). Urease enzyme (3.5.1.5) which potentially plays a role in synthesizing essential and non-essential amino acids by releasing ammonia as well as a number of amino acid metabolizing enzymes, including ornithine carbamoyltransferase (2.1.3.3; metabolizes arginine), lysine decarboxylase (4.1.1.18; metabolizes lysine), lysine racemase (5.1.1.5; metabolizes lysine), showed higher read abundance in non-industrialized communities (not found by other methods). In addition, Carnelian exclusively found read enrichment in phospholipase D (3.1.4.4; involved in lipid metabolism) and phosphoadenylate 3'-nucleotidase (3.1.3.7; involved in sulfur metabolism), and depletion of phenylacetyl-CoA ligase (6.2.1.30; involved in phenylalanine metabolism), pyrrolysyl-tRNA synthetase (6.1.1.26; involved in aminoacyl-tRNA synthesis), and potassium-importing ATPase (3.6.3.12; involved in microbial potassium import) in the non-industrialized communities compared to the industrialized one.

We then explored the co-abundance associations between the core metabolic pathways involved in energy production and the metabolism of carbohydrate, protein, lipid, vitamins, and co-factors. Although hierarchical clustering can be used to identify clusters of co-abundance pathways between the non-industrialized vs industrialized communities, the clusters were not significantly different from each other with

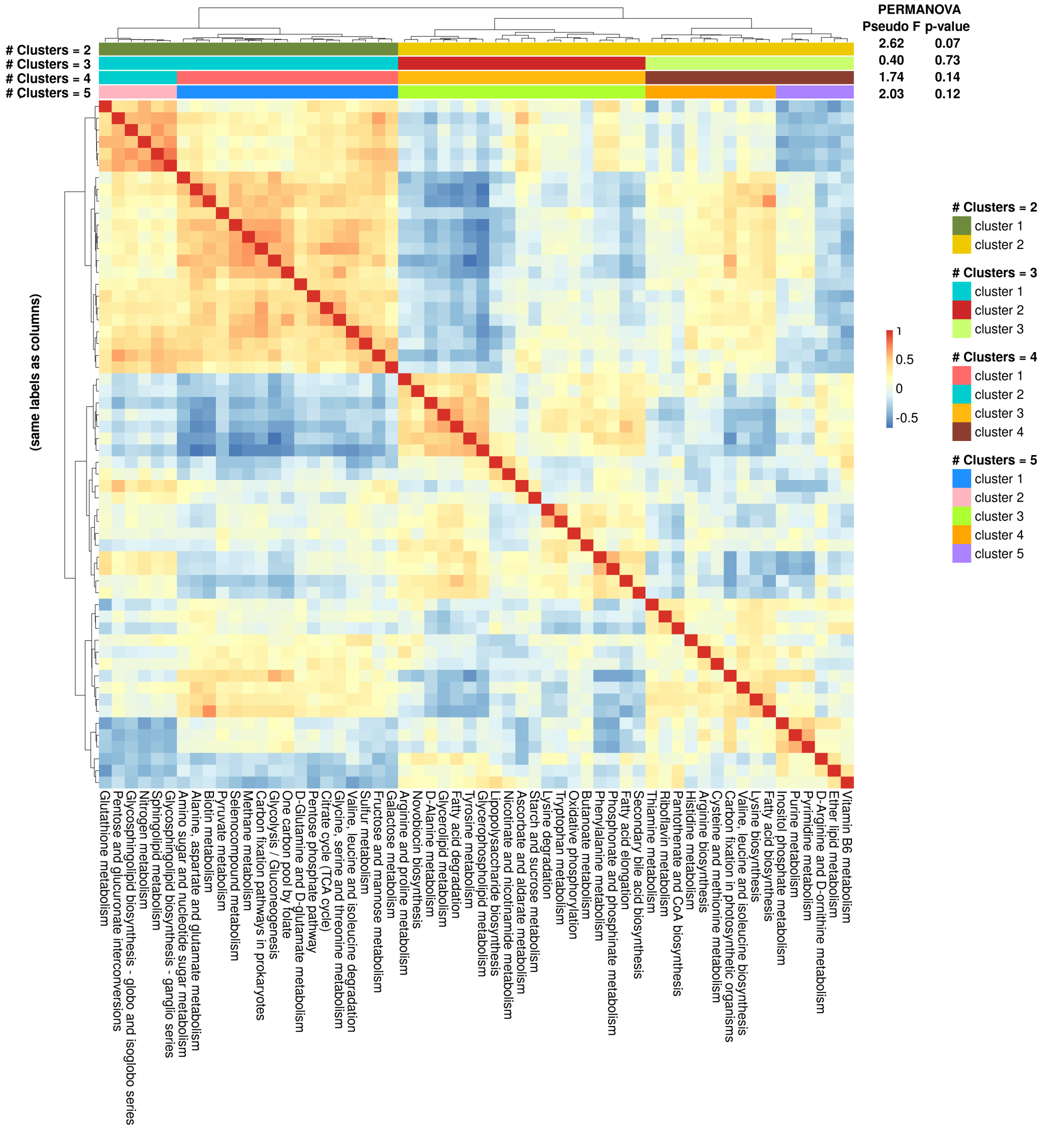


Figure 3-10: **Co-abundance association of core metabolic pathways across industrialized and non-industrialized microbiomes.** Co-abundance associations between pathways were calculated as the pairwise Kendall rank correlations between the pathway abundance profiles (obtained using Carnelian-generated EC profiles) of microbiomes from both communities considered together. Ward-linkage hierarchical clustering was used to partition the pathways using Euclidean distance, generating either 2, 3, 4, or 5 clusters. Although hierarchical clustering can be used to identify clusters of co-abundance pathways between the non-industrialized vs. industrialized communities, the clusters were not significantly different from each other concerning the industrialized/non-industrialized label (PERMANOVA test p-value > 0.05). Thus, in contrast to the top-level EC label clustering from part (a), the partitions are not merely recapitulating the industrialized/non-industrialized labels.

respect to the industrialized/non-industrialized label (PERMANOVA test p-value $>$ 0.05) (Figure 3-10). This result confirms the existence of pathway-level similarity in the core metabolic functionality (i.e., carbohydrate, amino acid, lipid, energy, vitamin, and co-factors metabolism) between the healthy gut microbiomes of non-industrialized and industrialized population. Differences at the pathway level between the two groups were mainly observed in secondary metabolism and xenobiotics degradation pathways (Table 3.39).

Table 3.39: Pathways identified as significantly variable between the microbiomes of the industrialized (Boston) vs. non-industrialized (Cameroon, Ethiopia, and Madagascar) communities using Carnelian-generated functional profiles. Significance thresholds used: BH-corrected Wilcoxon ranksum test p-value $<$ 0.05 and abs (log fold change) $>$ 1. Coverage is calculated as the ratio of the number of significant ECs mapped to a pathway to the total number of gold-standard ECs present in the pathway. When we take EC coverage of pathways into account, only six pathways remain significant.

Category	**Pathway**	**logFC**	**adjusted p-value**	**Coverage**
SM	Phenylpropanoid biosynthesis	-1.49	8.69E-38	0.16
X	Drug metabolism - cytochrome P450	1.33	5.31E-36	0.44
X	Metabolism of xenobiotics by cytochrome P450	1.02	1.73E-29	0.45
X	Beta-Lactam resistance	1.01	3.31E-22	1.00
V	Lipoic acid metabolism	-1.02	3.79E-31	0.75
SM	Caffeine metabolism	1.02	9.95E-31	0.31
X	Xylene degradation	1.26	1.04E-29	0.18
X	Atrazine degradation	1.11	1.35E-21	0.17
X	Ethylbenzene degradation	1.19	6.69E-35	0.67
X	Naphthalene degradation	1.10	8.06E-32	0.25

* Here, SM = Biosynthesis of Secondary Metabolites; V = Metabolism of Co-factors and Vitamins; X = Xenobiotics Biodegradation and Metabolism.

3.2.5 Finding novel functional dysbiosis in the gut microbiome of Parkinson's patients

Not only does Carnelian find consistent functional patterns in healthy and disease microbiomes across different geographies, but it also helps us uncover novel biology when applied to metagenomic data from a disease with poorly understood links to the gut microbiome. For example, although two-thirds of the patients with Parkinson's disease (a neurodegenerative disease of complex etiology) suffer from gastrointestinal (GI) abnormalities [199], it is not well understood how the gut microbiome is associated with the disease process. We applied Carnelian on whole metagenome sequencing data from the gut microbiomes of early-stage L-DOPA- naive Parkinson's disease (PD) patients and controls [200] to investigate the differences between the functional capacity of healthy and Parkinson's gut.

Our results reveal a hitherto unobserved functional shift in the gut microbiome of early-stage Parkinson's disease patients from the microbiome of healthy controls through performing differential abundance analyses of ECs and pathways. At the EC level, Carnelian exclusively identifies significant variation in read abundance (Benjamini-Hochberg (BH)-corrected Wilcoxon rank-sum test p-value < 0.05 and absolute log fold change > 0.58) in ribonucleoside-diphosphate reductase (1.17.4.1; implicated in glutathione metabolism), alpha-galactosidase (3.2.1.22; implicated in lipid metabolism), kynureninase (3.7.1.3; implicated in tryptophan metabolism), etc. (ECs identified by all four methods are provided in Tables 3.40–3.43).

At the pathway level, we found the PD gut to have lower read abundance in several carbohydrate metabolism pathways (BH-corrected Wilcoxon rank-sum test p-value < 0.05 and absolute log fold change > 0.11) (Table 3.44). Differential read abundances in different carbohydrate metabolism pathways were also found by HUMAnN2, mi-faser, and Kraken2 (Tables 3.45–3.47). Carnelian also identified significantly lower read abundances (BH-corrected Wilcoxon rank-sum test p-value < 0.05 and absolute log fold change > 0.11) in phenylalanine, tyrosine, and tryptophan biosynthesis (missed by both mi-faser and HUMAnN2), alanine, aspartate, and glutamate metabolism (missed by HUMAnN2), sphingolipid metabolism (missed by HUMAnN2 and Kraken2), glycosphingolipid biosynthesis, and D-alanine metabolism, notably missed by the other three methods (Tables 3.44–3.47). Note that the original study—which employed an assembly-based functional annotation approach using gene catalogs—found differences in gene abundances in only the D-Glucuronate and tryptophan metabolism pathways [200]. We also analyzed the same data set with an out-of-the-box HUMAnN2 pipeline with its default databases (Pathway results are provided in Table 3.48). Despite using additional taxonomic information, HUMAnN2 was not able to detect any significant shifts in pathways related to tryptophan metabolism (a clinically established hallmark of PD). The differentially abundant pathways identified by HUMAnN2 were primarily related to the broad category of purine and pyrimidine metabolism, which is non-specific to Parkinsonism. It detected a downward shift in some vitamin B and phospholipid metabolism pathways which might be associated with Parkinson's disease.

Table 3.40: Significant differentially abundant ECs between Parkinson's disease (PD) patients and controls identified by Carnelian in the PD-Bedarf data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.58.

EC	**logFC**	**adjusted *p*-value**	**EC**	**logFC**	**adjusted *p*-value**	**EC**	**logFC**	**adjusted *p*-value**
1.17.4.1	-0.70	0.0424	2.1.1.74	0.75	0.0024	1.7.1.13	-0.62	0.0005
1.4.1.16	-0.73	0.0042	1.1.1.271	-0.74	0.0039	3.4.13.9	-0.75	0.0452
3.4.14.12	-1.00	0.0006	2.7.1.209	0.76	0.0136	3.1.1.72	-0.58	0.0397
2.1.1.219	-0.82	0.0482	4.2.1.47	-0.69	0.0126	3.1.1.73	-0.98	0.0264
1.14.15.1	0.61	0.0424	6.3.2.2	-1.71	0.0482	3.1.4.16	0.87	0.0001
2.4.1.11	0.63	0.0283	1.12.1.2	0.70	0.0008	1.1.1.376	0.73	0.0264
2.4.1.321	0.60	0.0264	1.16.3.2	-0.68	0.0077	3.2.1.86	-0.73	0.0000
2.4.1.320	-0.78	0.0065	2.1.3.9	-0.92	0.0007	3.2.1.80	-0.59	0.0116
1.7.2.2	-0.60	0.0283	2.1.3.6	0.59	0.0246	3.2.1.177	-0.59	0.0042
3.11.1.1	-0.66	0.0264	3.5.3.18	0.68	0.0002	4.2.1.70	0.65	0.0046
5.3.1.17	-0.71	0.0020	3.7.1.3	0.63	0.0158	2.1.1.181	-0.70	0.0424
3.5.4.32	0.85	0.0011	3.8.1.2	0.66	0.0032	4.1.1.32	0.89	0.0015
2.7.14.1	0.67	0.0013	1.11.1.22	-0.72	0.0229	3.2.1.169	-0.81	0.0229
3.2.1.22	-0.58	0.0116	3.2.1.3	-1.06	0.0055	3.1.3.85	0.58	0.0482
2.1.1.72	-0.76	0.0065						

Table 3.41: Significant differentially abundant ECs identified by mi-faser in the PD-Bedarf data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.58.

EC	logFC	adjusted p-value	EC	logFC	adjusted p-value	EC	logFC	adjusted p-value
1.1.5.3	-0.79	0.0309	3.2.1.122	-0.87	0.0006	3.4.21.102	-0.65	0.0218
3.1.4.46	-0.77	0.0351	3.3.2.1	-1.07	0.0206	2.8.1.12	-0.76	0.0494
4.4.1.24	1.1	0.0033	3.4.25.2	0.61	0.0414	2.7.1.193	-0.64	0.0309
4.4.1.21	-1.01	0.0112	4.2.1.82	-0.73	0.0284	3.5.3.23	-1.17	0.0307
2.7.1.59	-1	0.0403	1.11.1.22	-0.78	0.0044	2.4.1.282	1.85	0.0006
1.7.1.7	-0.88	0.0448	2.1.1.265	-0.98	0.0373	2.1.1.35	-0.86	0.0189
3.4.14.12	-0.84	0.0025	3.5.4.40	-1.37	0.0316	2.1.1.193	-0.58	0.0151
1.5.8.2	0.8	0.0231	3.5.4.41	0.58	0.0249	5.1.3.3	-0.95	0.0412
2.1.1.217	-1.08	0.0379	3.1.3.45	-1.1	0.0336	1.12.2.1	1.17	0.0311
3.6.1.54	-1.18	0.0317	1.1.98.6	-0.78	0.0431	1.1.1.38	-0.67	0.0467
1.3.1.91	-0.81	0.0087	1.1.1.298	-1.01	0.0001	3.4.17.11	-0.74	0.0231
6.1.2.1	0.67	0.0182	2.5.1.30	0.73	0.0174	1.2.1.92	-1.12	0.0483
2.5.1.86	0.75	0.001	3.1.26.8	-1.17	0.0096	5.4.2.8	-0.67	0.0351
3.6.1.22	-0.67	0.0017	1.4.1.24	1.43	0.0249	5.1.3.20	-0.67	0.0032
1.6.99.3	-0.62	0.0209	1.4.1.3	0.71	0.0414	2.3.1.12	-0.89	0.0249
3.11.1.1	-0.83	0.0137	3.4.11.7	-0.99	0.0314	3.3.1.1	0.58	0.0076
1.1.1.103	0.63	0.0448	1.1.1.69	-0.65	0.0012	1.3.1.39	1.21	0.0058
2.5.1.90	-0.77	0.0038	5.4.3.5	0.62	0.0063	1.1.1.350	-0.95	0.0121
2.3.1.n4	-0.67	0.0376	4.1.1.32	0.95	0.0014	1.1.1.24	2.3	0.0416
2.3.1.101	1.19	0.0151	3.2.1.165	-3.64	0.0096	3.5.4.3	-0.64	0.0024
4.2.1.39	-0.64	0.0068	3.5.5.1	-0.59	0.0375	3.1.4.16	0.93	0.0004
3.5.4.39	1.77	0.0124	2.4.1.279	0.81	0.0238	3.5.2.17	-0.7	0.0318
2.7.1.130	-0.99	0.0041	1.5.1.43	-0.58	0.0144	3.2.1.85	-0.96	0.0131
2.7.14.1	0.82	0.0098	4.2.2.26	1.71	0.0186	3.6.1.67	-0.86	0.0496
3.4.15.5	-0.95	0.0001	3.1.3.16	-0.58	0.0242	1.1.1.383	1.04	0.003
5.3.99.11	-1.23	0.0048	3.1.1.11	-2.19	0.0395	1.1.1.385	-1.28	0.0315
4.1.1.75	-0.77	0.033	2.7.1.207	-0.98	0.0272	3.5.1.108	-0.81	0.0137
2.1.1.74	0.63	0.003	5.4.99.27	-0.85	0.0302	5.3.2.8	-1.09	0.0217
2.1.1.242	-1.11	0.0273	2.8.2.22	-0.75	0.0343	3.1.13.5	-0.77	0.0449
2.8.1.6	-0.73	0.0159	4.2.1.42	-1.04	0.0068	2.7.1.5	-0.67	0.0034
2.7.8.8	-0.99	0.0464	4.2.1.40	-0.91	0.0058	1.9.3.1	-2.14	0.0228
2.3.1.263	0.68	0.0388	2.4.1.54	-1.02	0.0068	3.2.1.99	-1.13	0.0358
3.1.3.27	-0.94	0.0408	2.2.1.10	1.33	0.0233	4.1.99.19	-0.72	0.0489
3.1.3.25	-0.71	0.0199	2.7.1.144	-1.16	0.0336	4.1.3.4	1.38	0.0219
1.3.4.1	1.2	0.0491	2.8.3.5	0.61	0.0035	4.1.3.3	-0.99	0.028
1.1.1.310	1.25	0.0209	5.5.1.27	1.68	0.0167	3.4.21.72	-1.89	0.0196
2.6.1.14	-0.61	0.0108	1.12.1.2	0.98	0.0017	1.14.13.2	-3.11	0.0355
1.1.1.308	0.97	0.0039	3.4.24.70	-1.05	0.0001	4.3.1.24	-1.59	0.0327
1.1.1.304	-0.98	0.0486	3.4.24.78	0.67	0.0098	3.4.11.9	-0.84	0.0343
2.7.1.219	-2.22	0.0466	1.8.98.1	0.8	0.008			

Table 3.42: Significant differentially abundant ECs identified by HUMAnN2 (translated) in the PD-Bedarf data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.58.

EC	**logFC**	**adjusted *p*-value**	**EC**	**logFC**	**adjusted *p*-value**	**EC**	**logFC**	**adjusted *p*-value**
1.1.5.2	-0.99	0.0327	4.4.1.8	-0.81	0.0366	2.7.1.193	-1.13	0.0345
2.7.7.19	-0.85	0.0201	4.4.1.5	-1.18	0.0485	1.16.3.1	-1.12	0.026
4.4.1.24	3.31	0.0216	3.1.3.45	-1.38	0.0049	3.5.3.7	-2.17	0.0427
4.4.1.21	-0.67	0.0441	2.7.1.17	-0.98	0.0397	2.7.1.25	-0.87	0.009
2.7.1.59	-0.86	0.0332	1.3.1.108	0.77	0.0372	2.1.1.193	-1.07	0.0284
3.5.1.104	0.86	0.0323	5.1.1.7	-0.65	0.0188	2.1.1.198	-1.12	0.045
1.1.1.58	-0.62	0.0166	6.5.1.2	-0.79	0.0167	5.1.3.9	-0.66	0.0229
1.2.7.3	1.46	0.0197	5.4.99.19	-1.13	0.007	2.1.1.171	-1.17	0.0158
3.6.1.55	-0.90	0.0178	1.1.1.283	2.08	0.0446	2.1.1.173	2.40	0.0316
1.3.1.91	-0.70	0.0366	3.5.2.17	-0.91	0.0236	1.1.1.346	-1.22	0.0325
2.5.1.86	2.59	0.0188	3.2.1.85	-1.14	0.0452	1.1.1.38	-0.81	0.0331
4.1.1.17	-0.64	0.0226	6.3.5.7	0.98	0.0032	2.3.3.13	-0.70	0.049
2.4.2.3	-0.89	0.0382	1.2.4.2	-0.79	0.0248	2.1.1.189	-1.01	0.0342
2.4.2.2	0.69	0.0119	4.1.2.25	-1.50	0.0291	6.4.1.3	1.02	0.0125
3.5.4.27	2.79	0.0474	4.1.2.21	-0.66	0.045	2.3.1.180	-0.62	0.0324
2.3.1.241	-0.85	0.0361	4.3.1.3	0.76	0.0425	3.3.1.1	1.05	0.0045
2.3.1.247	1.20	0.0076	1.1.1.383	0.73	0.0397	2.6.1.48	-1.02	0.0086
2.7.7.39	1.27	0.0158	2.3.1.174	1.26	0.0086	4.1.2.14	-0.98	0.0204
1.2.1.11	-0.87	0.0373	2.7.8.14	5.65	0.0271	2.7.7.56	-0.59	0.0403
2.5.1.90	-1.01	0.0463	5.4.99.2	0.69	0.0174	4.2.1.82	-2.13	0.0167
2.3.1.n4	-1.75	0.0029	1.3.3.11	-2.32	0.0399	1.1.1.385	2.72	0.0483
5.3.3.14	-1.37	0.0046	5.4.99.27	-1.00	0.0395	2.5.1.55	-0.58	0.0446
3.5.4.32	2.70	0.0361	5.4.99.24	-0.71	0.0415	4.2.1.1	-0.97	0.0397
2.1.2.9	-0.78	0.0384	5.3.1.28	-1.14	0.01	2.7.7.75	-0.96	0.0191
1.2.7.1	0.82	0.003	3.6.3.4	0.84	0.0173	3.6.3.42	2.39	0.0016
2.3.2.6	-1.23	0.0332	3.4.21.116	1.66	0.0021	1.14.11.47	-0.93	0.0318
6.3.4.20	-0.59	0.0042	2.7.1.148	-0.82	0.0468	1.5.1.3	-0.94	0.0403
2.7.8.8	-1.04	0.027	6.4.1.1	1.52	0.0005	1.1.1.60	-0.75	0.019
2.3.1.263	1.99	0.0056	3.6.5.n1	0.66	0.003	2.7.7.6	0.62	0.0027
2.7.9.1	0.81	0.0024	3.4.21.92	0.66	0.0047	2.1.1.181	-0.67	0.0497
3.1.3.27	-0.85	0.0297	6.3.2.2	-1.00	0.0445	5.4.3.5	1.91	0.0022
3.1.3.25	-0.73	0.0345	4.2.1.2	0.63	0.0358	5.4.3.4	0.97	0.0023
5.1.1.13	1.83	0.0007	6.3.2.8	-0.75	0.0244	5.4.3.3	0.92	0.0028
2.7.6.1	-0.66	0.0491	6.3.4.3	0.78	0.0021	5.4.3.2	1.00	0.0158
2.6.1.11	0.84	0.0442	2.7.10.2	1.14	0.0146	4.1.1.32	2.09	0.0022
4.2.2.24	2.05	0.007	2.3.1.35	5.63	0.0253	4.1.1.37	-0.67	0.0483
3.6.3.29	1.36	0.0311	1.10.3.14	-0.63	0.0337	6.3.4.18	-0.65	0.0417
2.3.1.54	0.71	0.0182						

Table 3.43: Significant differentially abundant ECs identified by Kraken2 in the PD-Bedarf data set. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.58.

EC	**logFC**	**adjusted *p*-value**	**EC**	**logFC**	**adjusted *p*-value**	**EC**	**logFC**	**adjusted *p*-value**
4.4.1.21	-0.65	0.0152	4.2.2.26	1.15	0.0039	2.3.1.228	1.13	0.021
2.5.1.72	-0.65	0.0052	3.2.2.8	-0.92	0.0241	1.14.14.5	-0.91	0.0231
3.5.1.104	0.83	0.0448	1.12.98.4	1.01	0.0079	4.2.3.170	3.25	0.0187
1.1.1.53	1.87	0.0007	1.12.98.2	1.01	0.0385	1.1.1.35	0.66	0.0025
1.2.7.8	0.63	0.0174	4.2.1.40	-0.93	0.0093	1.1.2.8	1.07	0.0082
3.4.14.12	-0.95	0.0018	5.3.1.29	1.09	0.0355	1.2.99.8	1.34	0.0145
1.13.11.75	-1.59	0.0296	2.4.2.36	1.89	0.0413	2.4.99.21	1.32	0.0108
1.1.99.6	-1.08	0.0111	2.4.2.31	-1.00	0.0211	2.6.1.48	-0.63	0.0209
1.1.1.302	-0.95	0.0018	3.1.4.16	0.91	0.0024	4.1.2.14	-0.68	0.0361
2.5.1.86	0.69	0.0006	3.2.1.85	-1.11	0.0058	2.7.7.1	2.04	0.0039
4.2.2.6	-3.24	0.0148	2.1.1.315	0.80	0.0206	1.1.1.350	-1.10	0.0133
2.4.1.11	1.45	0.0295	3.5.1.108	-0.72	0.0454	4.2.1.109	3.97	0.0253
2.4.1.12	-0.94	0.0096	3.4.21.72	1.06	0.0012	3.5.4.3	-0.58	0.0103
6.3.2.44	0.67	0.0472	3.5.1.25	-0.65	0.0058	2.7.1.219	-0.63	0.0454
4.1.1.104	-0.98	0.0115	1.3.5.1	-0.60	0.0166	3.2.1.122	-0.83	0.0028
3.11.1.1	-0.92	0.0199	4.1.1.87	1.52	0.0068	3.8.1.7	-1.59	0.0436
1.1.1.107	0.72	0.031	4.1.1.81	0.64	0.0208	3.4.23.42	0.99	0.0304
2.4.1.250	0.76	0.0079	5.4.99.17	1.34	0.0295	4.2.3.154	-0.98	0.0459
1.14.15.12	-1.51	0.0164	6.3.2.5	-0.62	0.0249	1.1.1.298	-0.88	0.0317
2.7.14.1	0.83	0.0103	2.1.1.298	-0.59	0.0296	1.7.1.15	-0.82	0.044
2.1.1.74	0.75	0.0007	1.8.98.1	0.65	0.0467	1.1.1.14	0.64	0.0293
5.5.1.16	3.62	0.0022	3.1.3.78	2.18	0.0062	5.4.99.19	-0.71	0.0358
3.1.3.25	-0.73	0.0065	3.4.21.105	-0.84	0.0468	1.1.1.286	0.89	0.0431
1.3.4.1	1.15	0.0485	3.1.1.31	-0.87	0.0268	2.5.1.113	-0.59	0.0166
2.7.4.2	-0.83	0.0258	3.5.3.23	-0.95	0.0194	1.14.13.2	-2.39	0.0216
2.3.1.129	-1.03	0.0454	2.1.1.289	-2.33	0.0007	4.3.1.23	1.51	0.0204
2.4.1.279	0.93	0.0065	6.6.1.1	0.71	0.0043	3.4.11.7	-1.07	0.0431
2.3.1.94	1.19	0.0303	2.7.4.29	-0.84	0.0463	5.4.3.5	0.76	0.0013
3.6.4.9	1.47	0.005	2.1.1.196	0.96	0.0068	4.1.1.32	0.94	0.0034

Table 3.44: Pathways identified as significantly variable between PD patients and healthy controls in the PD-Bedarf data set using Carnelian-generated functional profiles. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.11.

Category	**Name**	**logFC**	**adjusted p-value**
C	Pentose and glucuronate interconversions	-0.31	0.0009
C	Fructose and mannose metabolism	-0.22	0.0039
C	Galactose metabolism	-0.18	0.0013
C	Ascorbate and aldarate metabolism	-0.28	0.0171
AA	Alanine, aspartate and glutamate metabolism	-0.18	0.0213
X	Benzoate degradation	0.15	0.0015
AA	Phenylalanine, tyrosine and tryptophan biosynthesis	-0.19	0.0190
AA	D-Alanine metabolism	0.25	0.0029
AA	Glutathione metabolism	-0.49	0.0016
C	Starch and sucrose metabolism	-0.19	0.0020
SM	Streptomycin biosynthesis	-0.24	0.0397
T	Polyketide sugar unit biosynthesis	-0.32	0.0452
G	Glycosaminoglycan degradation	-0.32	0.0372
G	Peptidoglycan biosynthesis	0.20	0.0022
G	Lipoarabinomannan (LAM) biosynthesis	0.34	0.0060
L	Sphingolipid metabolism	-0.35	0.0107
L	Glycosphingolipid biosynthesis - globo and isoglobo series	-0.43	0.0325
L	Glycosphingolipid biosynthesis - ganglio series	-0.37	0.0372
V	Folate biosynthesis	-0.13	0.0303
V	Porphyrin and chlorophyll metabolism	0.12	0.0482
T	Zeatin biosynthesis	0.26	0.0099
X	Drug metabolism - other enzymes	-0.19	0.0066
SM	Biosynthesis of secondary metabolites - unclassified	0.15	0.0264
T	Biosynthesis of vancomycin group antibiotics	-0.36	0.0424

[*] Here, C = Carbohydrate Metabolism; L = Lipid Metabolism; E = Energy Metabolism; N = Nucleotide Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); SM = Biosynthesis of Secondary Metabolites; G = Glycan Biosynthesis and Metabolism; V = Metabolism of Co-factors and Vitamins; X = Xenobiotics Biodegradation and Metabolism; GI = Genetic Information Processing; T = Metabolism of Terpenoids and Polyketides.

Table 3.45: Pathways identified as significantly variable between PD patients and healthy controls in the PD-Bedarf data set using mi-faser-generated functional profiles. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.11.

Category	**Name**	**logFC**	**adjusted *p*-value**
C	Pentose and glucuronate interconversions	-0.31	0.0022
C	Fructose and mannose metabolism	-0.21	0.0099
C	Galactose metabolism	-0.18	0.0018
C	Ascorbate and aldarate metabolism	-0.31	0.0099
L	Fatty acid elongation	-0.45	0.0482
T	Ubiquinone and other terpenoid-quinone biosynthesis	-0.60	0.0184
N	Purine metabolism	0.12	0.0071
N	Pyrimidine metabolism	0.19	0.0010
SM	Phenazine biosynthesis	-0.25	0.0283
AA	Phosphonate and phosphinate metabolism	-0.79	0.0229
AA	Selenocompound metabolism	-0.13	0.0184
AA	D-Arginine and D-ornithine metabolism	0.69	0.0116
AA	Glutathione metabolism	-0.43	0.0158
C	Starch and sucrose metabolism	-0.16	0.0264
G	Other glycan degradation	-0.60	0.0007
G	Glycosaminoglycan degradation	-0.59	0.0005
C	Inositol phosphate metabolism	0.26	0.0065
L	Sphingolipid metabolism	-0.53	0.0016
L	Glycosphingolipid biosynthesis - ganglio series	-0.60	0.0009
X	Nitrotoluene degradation	0.18	0.0325
V	One carbon pool by folate	0.12	0.0424
V	Biotin metabolism	-0.20	0.0042
V	Folate biosynthesis	-0.31	0.0013
T	Limonene and pinene degradation	-0.48	0.0371
T	Zeatin biosynthesis	0.44	0.0020
X	Caprolactam degradation	-0.54	0.0424
GI	Aminoacyl-tRNA biosynthesis	0.15	0.0055
X	Drug metabolism - other enzymes	-0.12	0.0012
X	Steroid degradation	-0.57	0.0303
SM	Biosynthesis of secondary metabolites - unclassified	0.33	0.0032

* Here, C = Carbohydrate Metabolism; L = Lipid Metabolism; E = Energy Metabolism; N = Nucleotide Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); SM = Biosynthesis of Secondary Metabolites; G = Glycan Biosynthesis and Metabolism; V = Metabolism of Co-factors and Vitamins; X = Xenobiotics Biodegradation and Metabolism; GI = Genetic Information Processing; T = Metabolism of Terpenoids and Polyketides.

Table 3.46: Pathways identified as significantly variable between PD patients and healthy controls in the PD-Bedarf data set functional profiles generated by HUMAnN2 (translated). Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.11.

Category	**Name**	**logFC**	**adjusted *p*-value**
C	Citrate cycle (TCA cycle)	0.28	0.005511
C	Pentose phosphate pathway	-0.29	0.015847
C	Pentose and glucuronate interconversions	-0.39	0.030336
C	Galactose metabolism	-0.19	0.008417
L	Synthesis and degradation of ketone bodies	0.41	0.026405
N	Purine metabolism	0.41	0.002425
N	Pyrimidine metabolism	0.52	0.007124
SM	Monobactam biosynthesis	-0.31	0.03248
AA	Valine, leucine and isoleucine degradation	0.45	0.012579
AA	Lysine degradation	0.77	0.001641
AA	beta-Alanine metabolism	-0.69	0.030336
AA	D-Arginine and D-ornithine metabolism	2.18	0.001629
C	Starch and sucrose metabolism	-0.47	0.028313
C	Pyruvate metabolism	0.34	0.000714
X	Nitrotoluene degradation	0.74	0.000325
C	Propanoate metabolism	0.42	0.000032
C	Butanoate metabolism	0.28	0.022911
V	One carbon pool by folate	0.3	0.00991
E	Methane metabolism	0.37	0.003526
E	Carbon fixation in photosynthetic organisms	0.37	0.019818
E	Carbon fixation pathways in prokaryotes	0.6	0.000021
C	Pantothenate and CoA biosynthesis	-0.5	0.015847

[*] Here, C = Carbohydrate Metabolism; L = Lipid Metabolism; E = Energy Metabolism; N = Nucleotide Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); SM = Biosynthesis of Secondary Metabolites; G = Glycan Biosynthesis and Metabolism; V = Metabolism of Co-factors and Vitamins; X = Xenobiotics Biodegradation and Metabolism; GI = Genetic Information Processing; T = Metabolism of Terpenoids and Polyketides.

Table 3.47: Pathways identified as significantly variable between PD patients and healthy controls in the PD-Bedarf data set functional profiles generated by Kraken2. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05 and abs (log fold change) > 0.11.

Category	**Name**	**logFC**	**adjusted *p*-value**
C	Pentose and glucuronate interconversions	-0.35	0.0016
C	Fructose and mannose metabolism	-0.17	0.0264
C	Galactose metabolism	-0.23	0.0002
C	Ascorbate and aldarate metabolism	-0.32	0.0084
L	Primary bile acid biosynthesis	0.81	0.0013
L	Steroid hormone biosynthesis	2.27	0.0014
N	Purine metabolism	0.13	0.0126
N	Pyrimidine metabolism	0.22	0.0006
AA	Alanine, aspartate and glutamate metabolism	-0.14	0.0126
AA	Arginine and proline metabolism	-0.15	0.0482
AA	Histidine metabolism	-0.18	0.0171
AA	Phenylalanine, tyrosine and tryptophan biosynthesis	-0.15	0.0325
AA	Selenocompound metabolism	-0.23	0.0046
AA	D-Arginine and D-ornithine metabolism	0.59	0.0091
AA	Glutathione metabolism	-0.41	0.0147
C	Starch and sucrose metabolism	-0.13	0.0371
G	Other glycan degradation	-0.42	0.0099
C	Amino sugar and nucleotide sugar metabolism	-0.13	0.0452
G	Glycosaminoglycan degradation	-0.42	0.0055
G	Lipopolysaccharide biosynthesis	-0.47	0.0397
G	Peptidoglycan biosynthesis	0.18	0.0424
L	Glycosphingolipid biosynthesis - ganglio series	-0.42	0.0099
X	Toluene degradation	0.87	0.0006
V	Riboflavin metabolism	-0.48	0.0032
V	Folate biosynthesis	-0.37	0.0029
T	Zeatin biosynthesis	0.38	0.0229
E	Nitrogen metabolism	-0.15	0.0482
GI	Aminoacyl-tRNA biosynthesis	0.2	0.0032
SM	Biosynthesis of secondary metabolites - unclassified	0.32	0.0136

[*] Here, C = Carbohydrate Metabolism; L = Lipid Metabolism; E = Energy Metabolism; N = Nucleotide Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); SM = Biosynthesis of Secondary Metabolites; G = Glycan Biosynthesis and Metabolism; V = Metabolism of Co-factors and Vitamins; X = Xenobiotics Biodegradation and Metabolism; GI = Genetic Information Processing; T = Metabolism of Terpenoids and Polyketides.

Table 3.48: Pathways identified as significantly variable between PD patients and healthy controls in the PD-Bedarf data set functional profiles generated by out-of-the-box HUMAnN2 using ChocoPhlAn, Uniref, and MetaCyc databases. Significance thresholds used: BH corrected Wilcoxon rank-sum test p-value < 0.05, abs (log fold change) > 0.11 and coverage > 0.10.

Category	**MetaCyc ID**	**Name**	**logFC**	**adjusted p-value**	**coverage**
V	1CMET2-PWY	N10-formyl-tetrahydrofolate biosynthesis	-0.60	0.03	0.55
AA	ASPASN-PWY	Superpathway of L-aspartate and L-asparagine biosynthesis	-0.76	0.01	0.17
T	NONMEVIPP-PWY	methylerythritol phosphate pathway I	-0.59	0.04	0.95
V	PANTO-PWY	phosphopantothenate biosynthesis I	-0.59	0.03	0.86
AA	PWY-2942	L-lysine biosynthesis III	-0.59	0.03	0.95
V	PWY-3841	folate transformations II	-0.55	0.04	0.64
AA	PWY-5097	L-lysine biosynthesis VI	-0.55	0.02	0.94
L	PWY-5667	CDP-diacylglycerol biosynthesis I	-0.65	0.01	0.88
N	PWY-5686	UMP biosynthesis I	-0.53	0.04	1.00
N	PWY-5695	inosine 5'-phosphate degradation	-0.74	0.02	0.77
L	PWY-5973	cis-vaccenate biosynthesis	-0.70	0.01	0.70
N	PWY-6126	superpathway of adenosine nucleotides de novo biosynthesis II	-0.69	0.04	0.18
AA	PWY-6151	S-adenosyl-L-methionine cycle I	-0.82	0.01	0.95
N	PWY-6609	adenine and adenosine salvage III	-0.49	0.02	0.15
N	PWY-6700	queuosine biosynthesis	-0.72	0.03	0.80
V	PWY-6897	thiamine salvage II	-0.63	0.04	0.19
N	PWY-7219	adenosine ribonucleotides de novo biosynthesis	-0.58	0.03	1.00
N	PWY-7221	guanosine ribonucleotides de novo biosynthesis	-0.58	0.04	1.00
N	PWY-7229	superpathway of adenosine nucleotides de novo biosynthesis I	-0.66	0.03	0.19
N	PWY0-1296	purine ribonucleosides degradation	-0.70	0.04	0.27
L	PWY0-1319	CDP-diacylglycerol biosynthesis II	-0.65	0.01	0.88
L	PWY4FS-7	phosphatidylglycerol biosynthesis I (plastidic)	-0.99	0.05	0.12
L	PWY4FS-8	phosphatidylglycerol biosynthesis II (non-plastidic)	-0.99	0.05	0.12

[*] Here, L = Lipid Metabolism; N = Nucleotide Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); V = Metabolism of Co-factors and Vitamins; T = Metabolism of Terpenoids and Polyketides.

3.2.6 Robustness to sequencing technology biases

To test for Carnelian's robustness against the biases introduced by different sequencing technologies, we analyzed the sequencing reads generated by Roche 454 FLX Titanium and the Illumina Genome Analyzer (GA) II on the same DNA sample obtained from a complex planktonic community from a temperate freshwater lake (Lake Lanier, Atlanta, GA) from the Luo et al. study [201]. Raw sequencing reads were downloaded from the JGI Genomic Portal (`https://genome.jgi.doe.gov/portal/`) with a free account. Carnelian could capture similar functional diversity at both EC and pathway levels (Spearman correlation coefficients 0.87 and 0.89 respectively) despite the differences in sequencing technologies.

3.2.7 Applicability to environmental metagenomes

Since many of the species found in the human microbiome are well annotated, and many of the proteins in the reference databases come from human commensal bacteria, Carnelian, as well as other functional annotation methods, is expected to provide high-quality annotations for metagenomic reads from human body sites. Despite the existence of such bias in the reference data set, Carnelian can find meaningful biological insights from environmental metagenomic samples which we demonstrated using six aquatic metagenomes from an asbestos mine pit pond in Vermont (VAG-pond data set [185] and six beach sand metagenomes from the Deepwater Horizon oil spill site (DWH-spill data set [202]).

In the VAG-pond data set, we found the functional profiles of the samples from all three layers of the pond to be different from the freshwater samples; they showed high intra-layer correlations and relatively low inter-layer correlation as expected (Table 3.49). We identified several highly variable ECs which were abundant in the surface layer (epilimnion) where the sunlight, temperature, and amount of dissolved oxygen is higher and less abundant at the middle (hypolimnion) and bottom layers (hypolimnion). For example, EC terms 1.3.15.15, 4.99.1.4, and 2.7.1.177, vital in porphyrin and chlorophyll metabolism, were depleted both in the metalimnion

Table 3.49: Kendall rank correlation between the functional profiles of VAG-pond samples generated by Carnelian.

		Epilimnion			**Metalimnion**			**Hypolimnion**		**Freshwater**	
		1-S	**2-S**	**3-S**	**1-M**	**2-M**	**3-M**	**2-B**	**3-B**	**1-F**	**2-F**
	1-S	1	-	-	-	-	-	-	-	-	-
Epilimnion	2-S	0.81	1	-	-	-	-	-	-	-	-
(Surface)	3-S	0.83	0.86	1	-	-	-	-	-	-	-
	1-M	0.79	0.76	0.79	1	-	-	-	-	-	-
Metalimnion	2-M	0.73	0.77	0.78	0.74	1	-	-	-	-	-
(Middle)	3-M	0.78	0.8	0.79	0.82	0.78	1	-	-	-	-
Hypolimnion	2-B	0.72	0.77	0.75	0.74	0.74	0.78	1	-	-	-
(Bottom)	3-B	0.7	0.74	0.73	0.7	0.75	0.74	0.85	1	-	-
Freshwater	1-F	0.57	0.61	0.59	0.59	0.6	0.75	0.65	0.66	1	-
(Control)	2-F	0.53	0.57	0.56	0.54	0.67	0.6	0.6	0.63	0.6	1

and hypolimnion layers compared to the epilimnion layer (Table 3.50). Conversely, EC terms 2.7.4.31, 3.5.4.27, and 4.2.1.147 (implicated in methane metabolism), and 3.1.3.87, 4.1.1.50, and 5.3.1.23 (involved in sulfur-containing amino-acid metabolism) were found abundant in the hypolimnion layer and depleted in the epilimnion layer (Table 3.50). Interestingly, several ECs, such as 1.14.11.7 (involved in sulfur metabolism) were found enriched in the metalimnion layer compared to both epilimnion and hypolimnion layers (Table 3.50).

We also observed high variability in several pathways, including the synthesis and degradation of ketone bodies, monobactam biosynthesis, Geraniol degradation, and D-arginine, D-ornithine metabolism between all three layers (Table 3.51). Reduced rate of oxidative phosphorylation was observed in hypolimnion compared to epilimnion, which was expected in the presence of less dissolved oxygen and sunlight in the bottom layer. Overall, the functional profiles of the samples from the bottom layer showed slightly more functional variability compared to the top two layers as indicated by Shannon-Wiener index (Figure 3-11) which agrees with the taxonomic-level findings of the original study [185].

Table 3.50: Highly variable ECs between the epilimnion, metalimnion, and hypolimnion layers found by Carnelian in the VAG-pond data set.

EC	EM_FC	EH_FC	MH_FC
2.7.1.202	6.19	13.82	2.23
4.1.1.47	1.02	2.91	2.85
2.7.1.177	1.38	2.86	2.07
1.3.7.15	1.33	2.86	2.15
3.4.17.n1	0.57	2.41	4.24
1.1.1.108	0.54	2.09	3.9
2.7.7.76	0.35	1.99	5.63
5.4.1.4	0.64	1.83	2.85
1.14.11.17	0.36	1.06	2.94
3.5.4.3	0.25	0.33	0.08
4.99.1.4	2.5	3.45	8.33
5.1.3.29	0.4	0.31	0.13
5.1.99.1	0.53	0.24	0.13
1.13.11.48	0.53	0.25	0.13
2.1.1.156	0.63	0.23	0.14
1.20.4.3	0.56	0.29	0.16
2.7.14.1	0.61	0.27	0.17
4.1.1.98	0.49	0.34	0.17
4.2.1.171	1.13	0.15	0.17
1.11.1.6	0.51	0.34	0.18
1.4.3.23	0.75	0.25	0.18
1.14.99.50	1.21	0.15	0.18
1.1.1.412	0.97	0.19	0.19
2.7.8.47	0.67	0.28	0.19
4.1.1.50	1	0.19	0.19
2.7.4.31	0.82	0.24	0.19
1.1.1.286	0.64	0.31	0.2
3.6.1.7	0.75	0.26	0.2
3.1.1.17	1.05	0.19	0.2
1.21.98.1	1.08	0.19	0.2
2.7.1.162	1.09	0.19	0.2
4.2.1.83	0.66	0.31	0.2
2.4.1.1	0.74	0.29	0.21
2.2.1.10	0.67	0.32	0.21
1.1.1.374	0.71	0.3	0.21
3.6.3.4	0.63	0.34	0.21
3.5.1.102	0.69	0.32	0.22
4.3.1.7	0.96	0.23	0.22
2.7.1.164	0.9	0.25	0.22
5.5.1.16	0.86	0.26	0.22
2.5.1.97	1.21	0.2	0.24
1.1.1.14	0.78	0.31	0.24
1.2.1.80	0.79	0.31	0.24
1.4.1.24	0.71	0.35	0.24
2.5.1.105	0.71	0.35	0.25
3.6.3.12	1.11	0.23	0.25
5.3.1.23	1.04	0.24	0.25
1.1.1.382	1.52	0.17	0.25
3.2.1.37	0.83	0.31	0.26
3.6.1.17	0.91	0.28	0.26
1.6.1.1	0.76	0.34	0.26
4.2.1.82	1.21	0.21	0.26
4.2.1.147	1.14	0.23	0.26
2.6.1.83	0.89	0.3	0.27
2.6.1.59	0.94	0.29	0.27
3.5.1.44	1.43	0.19	0.27
5.1.3.30	1	0.28	0.28
4.1.2.27	0.91	0.31	0.28
1.12.98.4	0.84	0.34	0.28
1.13.11.49	1.1	0.26	0.28
4.2.2.22	0.95	0.3	0.28
3.4.17.19	0.82	0.35	0.29
3.1.3.87	1.04	0.28	0.29
3.4.17.14	0.91	0.33	0.3
1.1.1.390	0.9	0.33	0.3
1.3.1.104	0.86	0.35	0.3
3.1.1.61	1.2	0.25	0.3
3.2.1.67	0.89	0.34	0.3
1.1.1.343	1.25	0.25	0.31
4.2.1.5	0.91	0.35	0.32
3.5.4.27	1.06	0.31	0.32
1.5.3.1	0.96	0.35	0.33
4.2.3.156	1.14	0.3	0.34
2.4.1.320	1.95	0.18	0.35

* Here, EM = Epilimnion vs. Metalimnion, EH = Epilimnion vs. Hypolimnion, MH = Metalimnion vs. Hypolimnion, and FC = Fold Change.

Table 3.51: Highly variable pathways between the epilimnion, metalimnion, and hypolimnion layers found by Carnelian in the VAG-pond data set.

Category	**Pathway**	**Coverage**	**EM logFC**	**EH logFC**	**MH logFC**
L	Synthesis and degradation of ketone bodies	1	0.13	0.37	0.23
SM	Monobactam biosynthesis	0.5	0.17	0.29	0.12
T	Geraniol degradation	0.4	0.09	0.26	0.18
X	Benzoate degradation	0.39	0.07	0.24	0.17
G	Lipoarabinomannan (LAM) biosynthesis	0.5	0.15	0.24	0.1
SM	Carbapenem biosynthesis	0.4	0.09	0.24	0.15
L	Biosynthesis of unsaturated fatty acids	0.32	0.03	0.23	0.2
G	Arabinogalactan biosynthesis - Mycobacterium	0.86	0.13	0.22	0.09
AA	Valine, leucine and isoleucine degradation	0.58	0.06	0.2	0.14
AA	Valine, leucine and isoleucine biosynthesis	0.71	0.01	0.18	0.17
L	alpha-Linolenic acid metabolism	0.33	0.06	0.18	0.12
AA	Lysine degradation	0.39	0.03	0.17	0.14
C	C5-Branched dibasic acid metabolism	0.43	0.01	0.17	0.16
C	Butanoate metabolism	0.58	0.04	0.17	0.13
AA	Phosphonate and phosphinate metabolism	0.38	0.14	0.17	0.03
T	Terpenoid backbone biosynthesis	0.56	0.04	0.17	0.13
L	Fatty acid degradation	0.38	0.03	0.16	0.13
AA	Lysine biosynthesis	0.78	0.08	0.16	0.08
X	Dioxin degradation	0.62	0.07	0.15	0.08
T	Ubiquinone and other terpenoid-quinone biosynthesis	0.49	0.02	0.15	0.13
X	Ethylbenzene degradation	0.67	0.03	0.14	0.11
X	Steroid degradation	0.54	0.11	0.13	0.01
AA	Glycine, serine and threonine metabolism	0.66	0.04	0.12	0.08
V	Lipoic acid metabolism	0.75	-0.03	0.11	0.15
C	Galactose metabolism	0.61	-0.06	-0.11	-0.05
SM	Penicillin and cephalosporin biosynthesis	0.43	-0.01	-0.12	-0.1
E	Oxidative phosphorylation	0.62	0.04	0.13	0.1
C	Starch and sucrose metabolism	0.68	0	-0.14	-0.13
G	Other glycan degradation	0.44	-0.01	-0.14	-0.14
G	Mannose type O-glycan biosynthesis	0.38	-0.12	-0.14	-0.02
AA	D-Arginine and D-ornithine metabolism	0.64	-0.07	-0.24	-0.16

* Thresholds used: coverage ≥ 0.30 and absolute log fold-change (logFC) > 0.11. Coverage is calculated as the ratio of the number of Carnelian-identified ECs mapped to a pathway to the total number of gold-standard ECs in the pathway. Here, EM = Epilimnion vs. Metalimnion, EH = Epilimnion vs. Hypolimnion, MH = Metalimnion vs. Hypolimnion, C = Carbohydrate Metabolism; L = Lipid Metabolism; E = Energy Metabolism; N = Nucleotide Metabolism; AA = Amino Acid Metabolism (includes metabolism of other amino acids as well); SM = Biosynthesis of Secondary Metabolites; G = Glycan Biosynthesis and Metabolism; V = Metabolism of Cofactors and Vitamins; X = Xenobiotics Biodegradation and Metabolism; GI = Genetic Information Processing; T = Metabolism of Terpenoids and Polyketides.

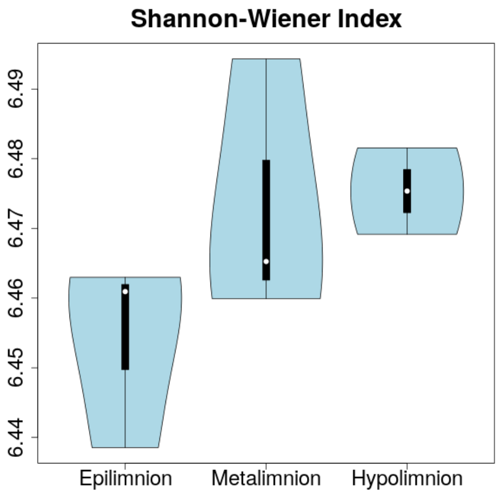


Figure 3-11: **Violin plot showing the functional diversity observed at different layers of the Vermont mine pit pond (VAG-Pond data set).** Samples from the hypolimnion (bottom) layer show higher functional diversity than both epilimnion (surface) and metalimnion (middle) layers, as indicated by the Shannon-Wiener indices.

In the DWH-spill data set, we observed a high intra-phase correlation between the samples (Figure 3-12). We also observed much higher functional diversity in the oil and post-oil phases compared to pre-oil phase (Shannon-Wiener Index: pre-oil: 3.43, oil: 5.82 post-oil: 5.78) which suggests a shift in the microbial functionality in the area due to the disastrous event of an oil spill. Carnelian-generated functional profiles showed a greater abundance of a number of ECs involved in the BTEX (Benzene, Toluene, Ethylbenzene, and Xylenes) degradation pathways in the oil phase compared to the pre-oil phase (Table 3.52). Notably, we observed an enrichment of catechol 1,2-dioxygenase (EC 1.13.11.1), catechol-2,3-dioxygenase (1.13.11.2), protocatechuate 3,4-dioxygenase beta chain (1.13.11.3), and muconolactone delta-isomerase (EC 5.3.3.4), key players in the aerobic degradation of aromatic hydrocarbons [203], in the oil phase samples. Many of the oil-degrading functions were also enriched in the post-oil phase which might suggest that the recovery process may not have finished at the time of sample collection—a finding that agrees with other independent studies of the same data set [72, 204]. Notably, Carnelian found significant enrichment in all BTEX metabolism pathways in the oil phase (Table 3.53).

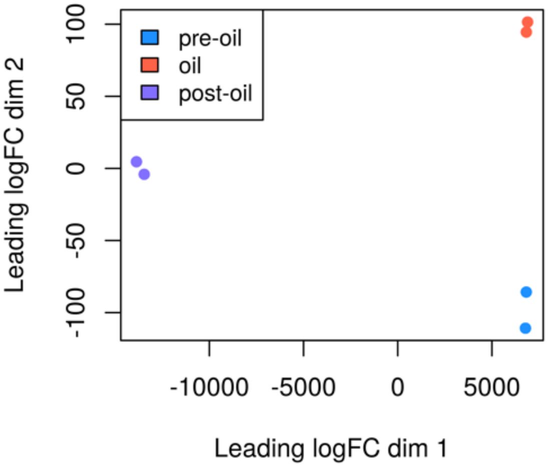


Figure 3-12: **Non-metric multidimensional scaling (NMDS) plot depicting the Carnelian-derived functional profiles of beach sand metagenomes from the DWH-spill data set.** Samples in each phase are functionally similar to each other in both the leading log fold-change dimensions. Samples in the oil-phase are functionally more similar to the samples in the post-oil phase.

Table 3.52: Hydrocarbon-degrading ECs involved in BTEX metabolism pathways found enriched in the oil phase compared to the pre-oil phase in the DWH-spill data set by Carnelian.

EC	**Name**	**Oil/Pre FC**	**Oil/Post FC**
1.1.1.35	3-hydroxyacyl-CoA dehydrogenase	2.23	0.97
1.12.98.4	Sulfhydrogenase 1 subunit beta	2.25	0.86
1.13.11.1	Catechol 1,2-dioxygenase	1.97	0.89
1.13.11.2	Catechol-2,3-dioxygenase	2.33	0.97
1.13.11.3	Protocatechuate 3,4-dioxygenase beta chain	1.89	0.9
1.13.11.39	Manganese-dependent 2,3-dihydroxybiphenyl 1,2-dioxygenase	2.58	1.18
1.13.11.57	Gallate dioxygenase	1.99	0.87
1.13.12.16	Nitronate monooxygenase	2.55	1.89
1.14.11.17	Alpha-ketoglutarate-dependent taurine dioxygenase	2.74	0.99
1.14.12.1	Anthranilate 1,2-dioxygenase large subunit	3.85	1.49
1.14.13.2	p-hydroxybenzoate hydroxylase (PHBH)	1.78	1.09
1.14.13.24	3-hydroxybenzoate 6-hydroxylase 1	2.26	1
1.14.13.7	Phenol 2-monooxygenase	2.13	1.04
1.14.14.1	Cytochrome P450 3A56	2.61	0.94
1.14.14.5	Alkanesulfonate monooxygenase	2.43	0.94
1.14.99.15	Cytochrome p450 CYP199A2	2.29	1
1.14.99.39	Ammonia monooxygenase alpha subunit (AMO)	1.74	0.83

(continued on next page)

Table 3.52: *cont.* Hydrocarbon-degrading ECs involved in BTEX metabolism pathways found enriched in the oil phase compared to the pre-oil phase in the DWH-spill data set by Carnelian.

EC	**Name**	**Oil/Pre FC**	**Oil/Post FC**
1.2.1.10	Acetaldehyde dehydrogenase	2.53	1.5
1.2.1.39	Phenylacetaldehyde dehydrogenase (PAD)	2.19	1.01
1.3.1.32	Maleylacetate reductase 2	1.95	0.85
1.3.8.10	Cyclohex-1-ene-1-carbonyl-CoA dehydrogenase (Ch1CoA)	3.31	1.58
1.3.8.11	Cyclohexane-1-carbonyl-CoA dehydrogenase (ChCoA)	2.6	1.47
1.97.1.2	Pyrogallol hydroxytransferase large subunit	1.88	0.88
2.3.1.16	3-ketoacyl-CoA thiolase	2.05	1.18
2.3.1.9	Acetyl-CoA acetyltransferase A	1.68	1.26
2.8.3.12	Glutaconate CoA-transferase subunit A	2.42	0.91
2.8.3.6	3-oxoadipate CoA-transferase subunit A	1.8	0.96
2.8.3.8	Acetate CoA-transferase subunit alpha	2.55	0.95
3.1.1.24	3-oxoadipate enol-lactonase 2	1.89	1.03
3.1.1.45	Putative carboxymethylenebutenolidase	3.37	1.21
3.1.8.1	Aryldialkylphosphatase	2.44	0.84
3.5.1.4	Acetamidase	2.33	0.99
3.5.5.1	Nitrilase 3	2.18	0.95
3.8.1.2	(S)-2-haloacid dehalogenase	3.58	0.94
3.8.1.3	Haloacetate dehalogenase H-1	1.52	0.92
3.8.1.5	Haloalkane dehalogenase	1.52	0.99
3.8.1.7	4-chlorobenzoyl coenzyme A dehalogenase-1	2.19	0.95
4.1.1.61	4-hydroxybenzoate decarboxylase subunit C	2.64	0.93
4.1.1.7	Benzoylformate decarboxylase (BFD)	2.39	0.92
4.1.1.70	Glutaconyl-CoA decarboxylase subunit gamma	1.6	1.03
4.1.3.17	4-carboxy-4-hydroxy-2-oxoadipic acid aldolase	2.25	0.96
4.1.3.39	4-hydroxy-2-oxovalerate aldolase (HOA)	2.44	1.27
4.2.1.17	enoyl-CoA hydratase	2.23	1.22
4.2.1.80	2-keto-4-pentenoate hydratase	2.46	1.19
4.2.1.83	4-oxalmesaconate hydratase (OMA hydratase)	2.38	0.93
5.1.2.2	Mandelate racemase (MR)	2.25	0.93
5.3.2.8	4-oxalomesaconate tautomerase	2.69	1.12
5.3.3.4	Muconolactone Delta-isomerase (MIase)	3.91	0.92
5.4.4.3	3-hydroxylaminophenol mutase (3HAP mutase)	1.88	1.37
5.5.1.2	3-carboxy-cis,cis-muconate cycloisomerase	2.19	0.94
5.5.1.7	Chloromuconate cycloisomerase	2.04	0.84
6.2.1.32	Anthranilate–CoA ligase	2.47	1.06

Table 3.53: Highly variable hydrocarbon metabolism pathways found by Carnelian in the DWH-spill data set.

Pathway	**Oil-Pre Fold Change**	**Oil-Post Fold Change**	**# ECs Mapped**	**Coverage**
Benzoate degradation	15.32	1.23	24	0.85
Aminobenzoate degradation	18.03	1.07	15	0.94
Chloroalkane and chloroalkene degradation	17.15	1.1	11	1
Chlorocyclohexane and chlorobenzene degradation	13.48	1.04	9	1
Dioxin degradation	21.31	1.31	7	0.88
Toluene degradation	17.5	1.01	6	1
Fluorobenzoate degradation	17.49	0.97	5	1
Styrene degradation	16.89	1.03	5	0.71
Xylene degradation	19.18	1.29	4	0.8
Ethylbenzene degradation	18.52	1.13	1	0.33
Polycyclic aromatic hydrocarbon degradation	18.73	1.07	1	1

Here, coverage is calculated as the ratio of the number of Carnelian-identified ECs mapped to a pathway to the total number of gold-standard ECs in the pathway. Pathways having coverage > 0.30 are reported.

3.2.8 Benchmarking Results

We benchmarked Carnelian against state-of-the-art alignment-based tools: mi-faser, and HUMAnN2, as well as a state-of-the-art alignment-free tool: Kraken2 using our gold-standard database, EC-2010-DB. Off-the-shelf HUMAnN2 and Kraken2 use taxonomic information in addition to translated searches; to ensure fair comparison hence we used only their "translated-search" modes. All comparisons were based on the estimation of EC terms identified by each method using the same gold-standard reference database. The reference databases used by mi-faser and HUMAnN2 and the Kraken2 reference index were created with Carnelian's gold-standard reference database for unbiased comparison.

Benchmarks for performance in functional inference

Testing a tool's capability to infer the functional capacity of a metagenome is very difficult in the absence of true functional labels. To achieve this, we simulated a synthetic human gut metagenome containing 5 million single-ended, 250-nucleotide DNA reads drawn from ChocoPhlAn pangenomes of the 20 most abundant bacterial species in Human Microbiome Project (HMP) stool samples [197] by following an approach similar to the HUMAnN2 paper by Franzosa and colleagues [73]. Species abundances were geometrically staggered from 0.1x to 70x (Table 3.54). To ensure that the synthetic metagenome has the desired relative abundance distribution of the 20 species, we drew fragments from the ChocoPhlAn pangenomes of each species with probability proportional to the product of the genome's size and corresponding species' target relative abundance. To create a gold-standard EC profile of the synthetic metagenome, we grouped the UniRef90 gene families present in the synthetic metagenome under ECs using the annotations from UniProt and cross-referenced the EC labels with our gold-standard database. The synthetic metagenome contained 9% read with 605 ECs from our reference database. These ECs were mapped to 117 KEGG metabolic pathways, which we consider as the pathway gold-standard.

Since the generation of synthetic metagenome does not account for the random

Table 3.54: Composition of the synthetic gut metagenome used for the task of functional capacity inference.

Species	**# Reads**	**Proportion**
Alistipes onderdonkii	1455410	29.11%
Alistipes putredinis	1031759	20.64%
Alistipes shahii	731372	14.63%
Bacteroides caccae	519394	10.39%
Bacteroides cellulosilyticus	368759	7.38%
Bacteroides dorei	261169	5.22%
Bacteroides massiliensis	185225	3.70%
Bacteroides ovatus	131266	2.63%
Bacteroides stercoris	93525	1.87%
Bacteroides thetaiotaomicron	65776	1.32%
Bacteroides uniformis	46771	0.94%
Bacteroides vulgatus	33147	0.66%
Barnesiella intestinihominis	23559	0.47%
Dialister invisus	16766	0.34%
Eubacterium rectale	11940	0.24%
Faecalibacterium prausnitzii	8770	0.18%
Parabacteroides distasonis	6115	0.12%
Parabacteroides merdae	4226	0.08%
Prevotella copri	2971	0.06%
Ruminococcus bromii	2080	0.04%
Total	5000000	100.00%
UniRef_annotated	4071989	81%
UniRef_unknown	928011	19%
EC_annotated	456107	9%

Random isolates of the listed species were selected from the ChocoPhlAn database, and reads were drawn from their annotated coding sequences, ensuring the target coverage in the synthetic metagenome.

variations introduced by the fragment sampling and may not give the true magnitude of the EC abundances present in the metagenome, we do not compare against the magnitude of the EC abundances. We calculated sensitivity as the ratio of the number of correct functional terms (EC or pathway) identified by a method to the total number of functional terms (EC or pathway) present in the synthetic metagenome (determined by mapping the UniRef90 gene families present in the sample to gold-standard ECs and KEGG pathways). Precision was calculated as the ratio of the number of correct functional terms identified by a method to the total number of functional terms identified by the method. F1-score was calculated as the geometric mean of precision and sensitivity. Carnelian achieved 82.15% sensitivity at the EC level and 90.55% sensitivity at the pathway level, which were higher than the other three tools (Table 3.55).

Table 3.55: Performance of Carnelian, mi-faser, HUMAnN2, and Kraken2 in functional capacity inference from the synthetic gut metagenome data. Carnelian achieves higher sensitivity and F1-score at both the EC and pathway levels as compared to mi-faser, HUMAnN2, and Kraken2 searches on a synthetic gut metagenome comprising of the 20 most abundant bacterial species in Human Microbiome Project (HMP) stool samples. Here, sensitivity is calculated as the ratio of the number of correct functional terms (EC or pathway) identified by a method to the total number of gold-standard functional terms (EC or pathway) present in the synthetic metagenome. Precision is calculated as the ratio of the number of correct functional terms identified by a method to the total number of functional terms identified by the method. F1-score is calculated as the geometric mean of precision and sensitivity. Best performances are shown in boldface.

	Tool	**Sensitivity (%)**	**Precision (%)**	**F1-score (%)**
EC level	Carnelian	**82.15**	64.21	**72.08**
	mi-faser	74.88	63.45	68.69
	HUMAnN2	30.41	**70.23**	42.45
	Kraken2	59.83	52.24	56.78
Pathway level	Carnelian	**90.55**	86.47	**88.46**
	mi-faser	84.25	91.45	87.7
	HUMAnN2	70.87	**91.84**	80
	Kraken2	85.04	84.38	84.71

Like functional capacity inference, performance in functional difference inference is equally difficult to measure. One way to do such benchmark was described by Lindgreen et al. [78] which we replicated. The data set consisted of different proportions of cyanobacteria (more abundant in set A), Bradyrhizobium and Rhizobium (more abundant in set A), and known pathogens (more abundant in set B). The shifts in taxa were used as a proxy for the expected pathway shifts between the two sets in the original study. Since the magnitude of the pathway abundances might differ from the differences observed at the taxonomic level, we tested for the direction of the change as suggested by Lindgreen and colleagues [78]. Carnelian identifies the expected direction of change in each of the three categories, where the rest of the methods don't (Figure 3-13).

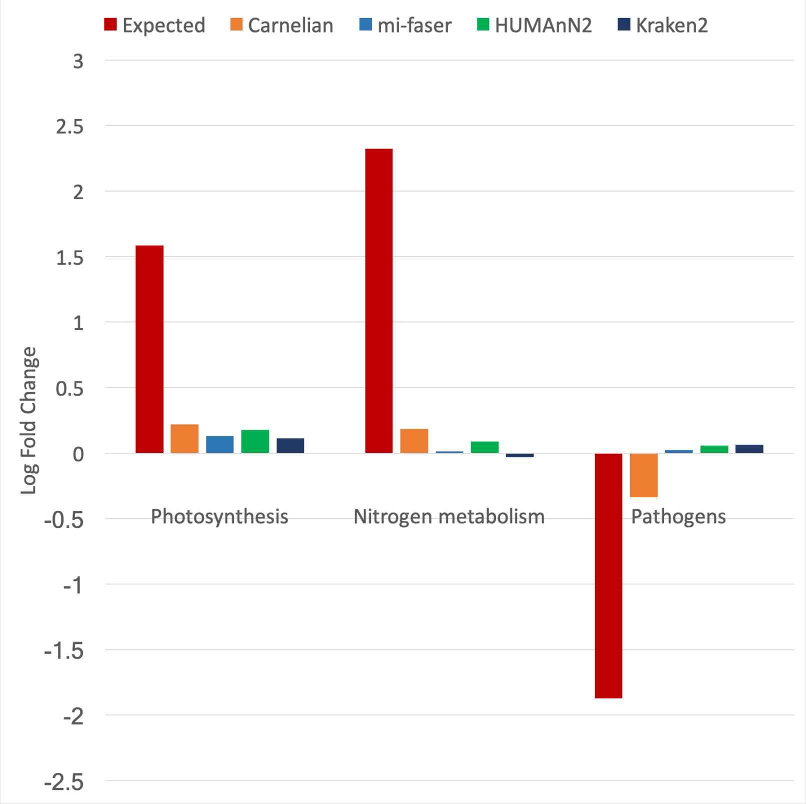


Figure 3-13: **Functional shifts predicted by Carnelian, mi-faser, HUMAnN2, and Kraken2 on the Lindgreen et al. data set.** A positive log fold change means an increase in set A relative to set B and vice versa. The expected fold change amounts were given by the original paper based on the taxonomic differences of the two sets. The test data sets were created with differences in the relative abundance of cyanobacteria (photosynthesis; more abundant in set A), Bradyrhizobium and Rhizobium (nitrogen fixation; more abundant in set A), and known pathogens (more abundant in set B). We profiled the metagenomes from the two sets with all four methods and calculated pathway abundances by mapping the identified ECs to the carbon-fixation, photosynthesis, nitrogen-fixation, two-component systems, bacterial chemotaxis, and cell motility pathways and summing the abundances. Carnelian identifies the expected direction of change in each of the three categories, where the rest of the methods don't.

Inspired by the above test, we simulated two sets of six complex metagenomes with varying proportions of coding sequences from random isolates of 20 different species of proteobacteria, cyanobacteria, photosynthetic bacteria, nitrogen-fixing bacteria and known pathogens from the ChocoPhlAn database (Table 3.56). The gold-standard

Table 3.56: Composition of the data set used for the task of functional change inference.

Species	**Set A**			**Set B**			**Average**	
	A1	**A2**	**A3**	**B1**	**B2**	**B3**	**Set A**	**Set B**
Anabaena sp 90	125384	100254	149865	751064	875154	650519	3%	15%
Bradyrhizobium diazoefficiens	249101	224954	274745	49967	50035	40149	5%	1%
Bradyrhizobium elkanii	250393	264246	214801	50136	40075	40236	5%	1%
Campylobacter coli	499756	449839	495120	75506	90185	100080	10%	2%
Chlorobium chlorochromatii	499910	400318	474744	74501	75118	100118	9%	2%
Chloroflexus aurantiacus	499942	600412	525030	99725	99630	75098	11%	2%
Erythrobacter litoralis	500580	549474	505666	74980	109706	75165	10%	2%
Escherichia albertii	50223	74874	24907	250429	274864	299179	1%	5%
Escherichia coli	50138	74624	49633	249811	275877	249701	1%	5%
Helicobacter canadensis	49893	49770	24998	250529	250215	224633	1%	5%
Microcystis aeruginosa	125138	150456	100086	748434	625162	624557	3%	13%
Nodularia spumigena	124423	150483	100215	749842	650422	848959	3%	15%
Nostoc punctiforme	124725	99389	150116	751316	849131	876120	2%	17%
Rhizobium freirei	250230	224940	200277	50064	59833	49784	5%	1%
Rhizobium grahamii	250005	274160	298599	49916	65349	75238	5%	1%
Rhodomicrobium vannielii	500826	525809	400816	99805	49724	60128	10%	1%
Rhodospirillum centenum	499860	475189	600069	75496	75315	90070	11%	2%
Salmonella enterica	50019	24909	74949	249450	224496	199586	1%	4%
Trichodesmium erythraeum	249610	261117	260705	49908	35508	45189	5%	1%
Vibrio campbellii	49844	24783	74659	249121	224201	275491	1%	5%
Total	5000000	5000000	5000000	5000000	5000000	5000000	100%	100%
Uniref annotated	4419647	4421728	4420501	4253105	4240995	4251735	88%	85%
Uniref unknown	580353	578272	579499	746895	759005	748265	12%	15%
EC annotated	456097	401600	446782	421600	456079	465900	9%	9%

Two sets of six complex metagenomes were created with varying proportions of coding sequences from 20 different species of proteobacteria, cyanobacteria, photosynthetic bacteria, nitrogen-fixing bacteria, and known pathogens.

EC and pathway profiles of the metagenomes were created in a similar way as the synthetic gut metagenome described above. Each of the metagenomes contained 5 million single-ended, 250-nucleotide DNA reads, 9% of which had EC annotations. We tested all four methods with the task of detecting the ECs and pathways with the correct direction of the change, as demonstrated by the reference profiles.

The reference values for abundances of ECs were calculated by grouping the Uniref90 gene families present in the simulated metagenomes by UniProt annotations and normalizing the summed counts by fragment length and average gene length per EC label. Similarly, reference values for pathway abundances were determined by mapping the ECs to pathways and summing their abundances. The gold-standard directions of the functional changes between two groups were determined by taking the fold-change of the average abundances of each functional term in the two sets. Carnelian achieves slightly higher sensitivity at both EC and pathway level compared to mi-faser and significantly higher sensitivity than HUMAnN2 and Kraken2. Sensitivity is measured as the proportion of functional terms identified by a method with the correct direction of shift between two sets (Table 3.57).

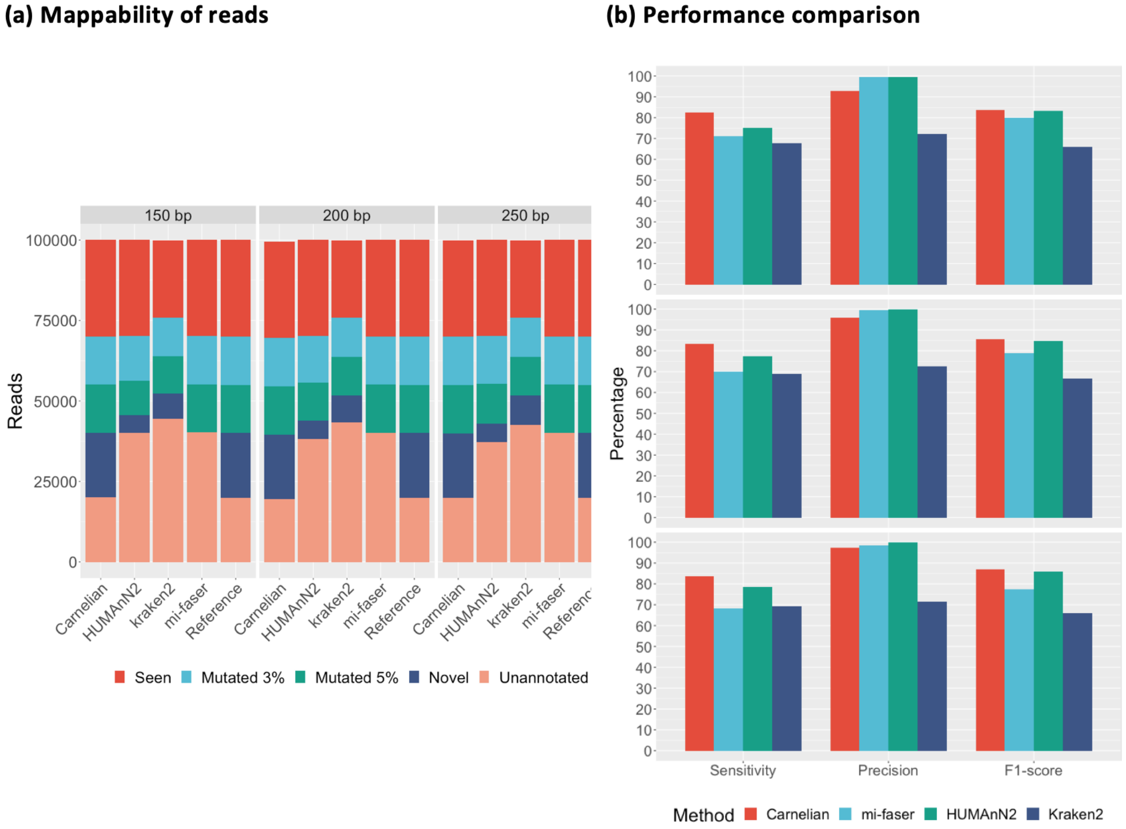
Table 3.57: Performance of Carnelian, mi-faser, HUMAnN2, and Kraken2 in functional change inference from the two-set complex metagenomes. The ECs with expected log fold change > 1 are termed as highly variable ECs. Similarly, the pathways with expected log fold change > 0.58 are termed as highly variable pathways. Sensitivity is measured as the proportion of functional terms identified by a method with the correct direction of shift between two sets. Carnelian achieves slightly higher sensitivity at both EC and pathway level compared to mi-faser and significantly higher sensitivity than HUMAnN2 and Kraken2.

Method	**Sensitivity at EC level (%)**		**Sensitivity at pathway level (%)**	
	All ECs	**Highly variable ECs**	**All pathways**	**Highly variable Pathways**
Carnelian	65.75	74.38	64.96	56.25
mi-faser	65.13	73.6	63.5	54.17
HUMAnN2	52.77	62.27	60.58	52.08
Kraken2	52	59.63	48.18	54.17

Accuracy benchmarks on our curated EC database

To test for accuracy, we benchmarked the tools on synthetic data sets generated from Carnelian's gold-standard reference proteins because actual functional labels for reads in the real-world data sets are not available. Three synthetic read data sets with read lengths 150 bp, 200 bp, and 250 bp were constructed, each consisting of 80% coding and 20% shuffled (non-coding) reads. The coding reads comprised of 30% seen sequences, 15% sequences drawn with 3% mutation rate, 15% sequences drawn with 5% mutation rate from the reference database, and 20% novel reads drawn from prokaryotic proteins with homology-based complete EC annotation from the Uniprot/Swissprot database that are not present in our gold-standard database.

Carnelian bins more reads than all other methods; especially in the case of novel reads, Carnelian's mappability is significantly better than other methods (Figure 3-14(a)). We achieve significantly higher sensitivity compared to the other methods at comparable accuracy (Figure 3-14(b)).


(a) Mappability of reads

(b) Performance comparison

Figure 3-14: **Comparison of Carnelian's performance against mi-faser, HUMAnN2 and Kraken2 on our in-house synthetic data set.** **(a)** Mappability of reads in in-house benchmarking data sets. Three synthetic read data sets were created with read lengths 150, 200, and 250 base pair lengths respectively. Each data set contained 80% coding reads (Seen: 30%; Mutated 3%: 15%; Mutated 5%: 15%; Novel: 20%) and 20% shuffled reads. Carnelian maps more coding reads compared to the other three methods. **(b)** Performance comparison on in-house benchmarking data sets. Carnelian achieves higher sensitivity and F1-score at comparable precision on the benchmarking data sets described in (b) compared to the other three methods.

We also performed a set of cross-validation experiments: we first drew amino acid (AA) fragments of length, $l = 50$ AA, 68AA, 84AA from the EC-2010-DB sequences ensuring every position of the reference proteins are covered at least five times by the fragments. We removed duplicate fragments from the sets. For each fragment length group, we then divided the fragments into the training and test sets as required for a five-fold cross-validation experiment. The fragments in the test sets were back-translated using standard codon table to mimic nucleotide reads of lengths 150-bp, 200-bp, and 250-bp. All four methods (Carnelian, mi-faser, HUMAnN2, and Kraken2) used the protein fragments contained in the training sets as the reference database and were tested on the nucleotide fragments in the test sets.

Additionally, we simulated the effect of the presence of a novel protein in metagenomic read data sets by performing two sets of experiments. First, we back-translated reference proteins to nucleotide sequences using the standard codon table. We then simulated reads of lengths 150-bp, 200-bp, and 250-bp from those sequences introducing 3% and 5% mutations within the fragments using *wgsim*. Carnelian's gold-standard reference database was used as a reference by all four methods, and they were tested on the *wgsim*-generated nucleotide reads with random mutations. To mimic the presence of functionally similar proteins with relatively less sequence similarity, we held out different proportions of proteins from the multi-protein EC bins in our gold-standard database and created two test sets by drawing 100-bp and 150-bp fragments from the back-translated held-out proteins. All four methods were trained on the remaining proteins in the multi-protein EC bins and tested on the nucleotide fragments from the held-out proteins.

For longer reads, Carnelian achievers higher sensitivity and accuracy compared to all other methods in cross-validation and mutation experiments. Especially in the case of novel proteins, Carnelian demonstrates significant improvement in sensitivity (Figure 3-15 and Table 3.58).

Performance evaluation metric

For evaluating the performance in cross-validation, mutation, and hold-out experiments, we used the macro-averaged sensitivity (ρ), precision (π) and F1-score as evaluation metrics. For each gold-standard functional label (functional bin), i, we

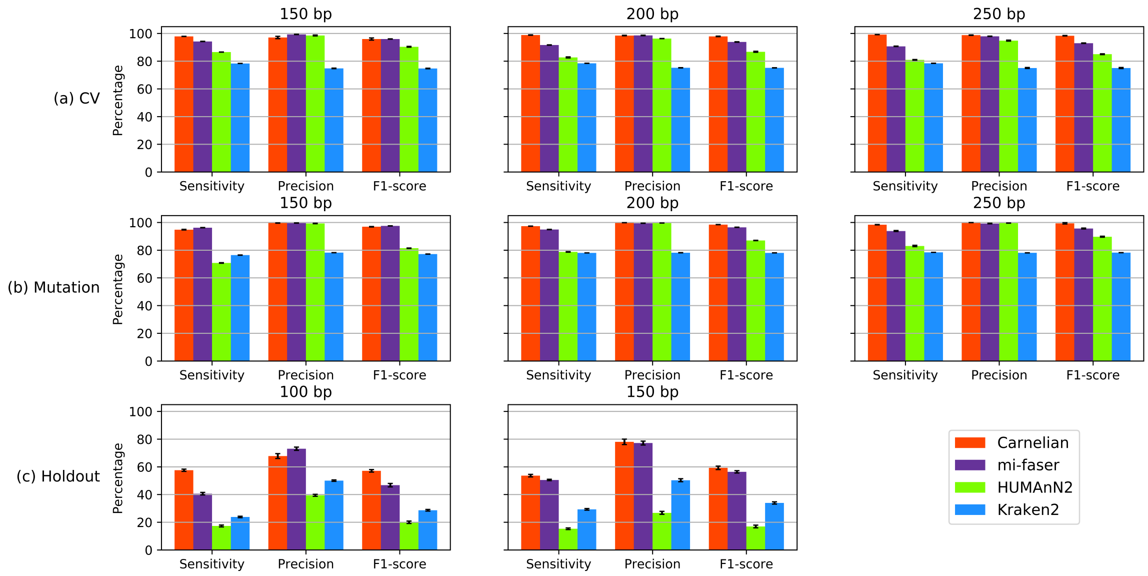


Figure 3-15: **Comparison of Carnelian's performance against mi-faser, HUMAnN2 and Kraken2 on Carnelian's curated EC database. (a) 5-fold Cross-validation experiments (testing across "seen" proteins).** While Carnelian and mi-faser achieve comparable sensitivity and F1-score for 150-bp reads, for longer reads (200-bp, 250-bp), where the field is heading, Carnelian achieves significantly higher sensitivity and F1-score compared to other methods. **(b) Mutation (5%) experiments (novel sequences with high similarity).** Similar to the cross-validation experiment, Carnelian achieves significantly higher sensitivity and F1-score compared to other methods for longer reads. **(c) Hold-out experiments (novel functionally similar proteins with moderate sequence similarity).** On short sequences (100-bp and 150-bp – the longest we could test with available data) from held-out proteins, Carnelian achieves significantly higher sensitivity and F1-score. In all experiments, all the methods demonstrate comparable precision, even though Carnelian alone doesn't perform exact alignment or exact k-mer matching. The error bars indicate standard deviation from the mean.

calculated ρ, π, and F1-score as follows:

$$\pi_i = \frac{TP_i}{TP_i + FP_i}, \ \rho_i = \frac{TP_i}{TP_i + FN_i}, \text{ and } F_i = \frac{2\pi_i\rho_i}{\pi_i + \rho_i}$$

Here, TP_i (True Positive) denotes the number of fragments binned correctly under label i; FP_i (False Positive) denotes the number of fragments that do not have label i but are binned under label i by the classifier model; and FN_i (False Negative) is the number of fragments that belong to the bin of label i but were incorrectly assigned to some other bin. The overall F1-score of the entire binning problem can be computed

Table 3.58: Comparison of Carnelian's performance against mi-faser, HUMAnN2, and Kraken2 on fragments of different lengths with 3% mutations using Carnelian's curated EC database.

Read Length	**Method**	**Mean**			**Stddev**		
		Sensitivity	**Precision**	**F1-score**	**Sensitivity**	**Precision**	**F1-score**
150-bp	Carnelian	96.19	99.71	97.73	0.08	0.02	0.08
	mi-faser	95.44	99.48	96.93	0.05	0.06	0.06
	Humann2	92.82	99.9	95.91	0.05	0.03	0.03
	Kraken2	78.15	78.16	78.12	0.04	0.01	0.03
200-bp	Carnelian	98.1	99.84	98.85	0.11	0.05	0.07
	mi-faser	94.32	99.33	96.05	0.14	0.06	0.14
	Humann2	96.46	99.91	97.99	0.26	0.01	0.16
	Kraken2	78.45	78.06	78.21	0.02	0.02	0.02
250-bp	Carnelian	98.87	99.9	99.3	0.04	0	0.02
	mi-faser	92.86	98.85	94.78	0.05	0.18	0.08
	Humann2	98.14	99.91	98.74	0.54	0	0.03
	Kraken2	78.47	78	78.19	0	0.02	0.02

by macro averaging, where F1-score for each bin, F_i, is calculated first and then averaged over all bins as:

$$\pi_m = \frac{\sum_{i=1}^{M} \pi_i}{M},\ \rho_m = \frac{\sum_{i=1}^{M} \rho_i}{M},\ \text{and}\ F_m = \frac{\sum_{i=1}^{M} F_i}{M}$$

where M is the total number of unique functional labels. Macro-averaged measures have advantages over micro-averaged ones because they give equal weight to each functional bin, regardless of how many examples of each label the classifier model has seen in the training set. Thus, the performance of the classifier model is not dominated by common bins; relatively rare categories also get equal importance.

Performance on examples of functionally similar proteins which are not-so-sequence-similar

We compared Carnelian's performance with mi-faser, HUMAnN2, and Kraken2 on three sets of examples of functionally similar proteins which are not-so-sequence-similar. Each experiment includes three proteins—A, B, and C, where A and B do not have much similarity at the protein sequence level but have the same enzymatic function (EC label). C has a different enzymatic function and serves as a control. Carnelian, mi-faser, HUMAnN2, and Kraken2 only see fragments of protein A in the reference as a positive example. Carnelian's training data also includes shuffled human sequences as negative examples. The test set contains randomly drawn nucleotide reads from back-translated protein sequences of A, B, and C. Reads of length 100 base pairs were drawn from back-translated protein sequences in such a way that every position was covered at least 10 times. Ideally, methods should be able to annotate reads from B with the same function as A and leave the reads from C unannotated. We measured performance in terms of sensitivity, precision, and F1-score.

Example 1:

A: Beta-galactosidase, gene lacZ from Escherichia coli (strain K12). EC label: 3.2.1.23

B: Evolved Beta-galactosidase, gene ebgA from Escherichia coli (strain UTI89 / UPEC). EC label: 3.2.1.23

C: 6-phospho-alpha-glucosidase, gene BET80_00230 from Escherichia coli. EC label: unverified 3.2.1.122 (used as negative here)

Similarity between proteins:

A-B: 34.18% (blastp e-value $= 5e-177$), A-C: 28.57% (blastp e-value $= 0.018$), and B-C: 29.03% (blastp e-value $= 2.3$)

Performance:

Method	**Sensitivity**	**Precision**	**F1-score**
Carnelian	0.9857	0.8257	0.8988
mi-faser	0.4857	1	0.6538
HUMAnN2 (translated)	0.4857	1	0.6538
Kraken2	0.4857	1	0.6538

HUMAnN2 (translated), Kraken2 and mi-faser couldn't annotate any reads from B.

Example 2:

A: Acetyl-coenzyme A carboxylase carboxyl transferase subunit alpha, gene accA from Escherichia coli (strain K12). EC label: 6.4.1.2

B: Acetyl-coenzyme A carboxylase carboxyl transferase subunit beta, gene accD from

Staphylococcus aureus (strain COL). EC label: 6.4.1.2

C: Pyruvate carboxylase subunit B, gene pycB from Methanocaldococcus jannaschii (strain ATCC 43067 / DSM 2661 / JAL-1 / JCM 10045 / NBRC 100440). EC label: 6.4.1.1 (used as negative here)

Similarity between proteins:

A-B: 58.33% (blastp e-value = 0.13), A-C: 41.18% (blastp e-value = 8.9), and B-C: 30.00% (blastp e-value = 1.7)

Performance:

Method	**Sensitivity**	**Precision**	**F1-score**
Carnelian	0.769	0.8162	0.7628
mi-faser	0.5273	1	0.6905
HUMAnN2 (translated)	0.5273	1	0.6905
Kraken2	0.4552	0.8632	0.5961

HUMAnN2 (translated), Kraken2 and mi-faser couldn't annotate any reads from B.

Example 3:

A: Urease subunit beta, gene ureB from Helicobacter felis (strain ATCC 49179 / NCTC 12436 / CS1). EC label: 3.5.1.5

B: Urease subunit alpha, gene ureC from Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv). EC label: 3.5.1.5

C: 4-hydroxyproline 2-epimerase, gene Arad_8151 from Agrobacterium radiobacter (strain K84 / ATCC BAA-868). EC label: 5.1.1.8 (used as negative here)

Similarity between proteins:

A-B: 55.00% (blastp e-value = 0.0), A-C: no significant similarity, and B-C: 40.00% (blastp e-value = 0.11)

Performance:

Method	Sensitivity	Precision	F1-score
Carnelian	0.9554	0.8198	0.8824
mi-faser	0.6658	1	0.7994
HUMAnN2 (translated)	0.4961	1	0.6632
Kraken2	0.5879	1	0.7405

Kraken2 and mi-faser were able to annotate some reads from B with correct function, but HUMAnN2 (translated) was not using the default similarity cut-off.

Benchmarks for runtime and memory requirement

Carnelian is practical. In terms of running time, our performance closely matches that of HUMAnN2 and is better than the standalone binary of mi-faser (Tables 3.59 and 3.60). We benchmarked the runtime and memory requirement of Carnelian against mi-faser, HUMAnN2, and Kraken2 using the synthetic gut metagenome created with the 20 most abundant species from HMP data set (Section 3.2.8). All the methods were run on a 40-core machine with 320 GB RAM; each core was Intel Xeon CPU E5-2695 v2 @ 2.40GHz.

To test how the runtime and memory requirement of Carnelian vary when the overall size of the input data remains fixed but read length varies, five million single-ended reads were generated for each data set from the synthetic gut metagenome with 20 species. Carnelian's runtime closely matched with HUMAnN2 and is better than mi-faser's standalone binary (Tables 3.59). Kraken2 is the fastest among all but significantly limited in terms of performance (Section 3.2.8).

To test how the runtime and memory requirement of Carnelian vary when read length remains fixed, but the size of the input data varies, one million, five million, and 10 million single-ended 250 base pair reads were generated from the synthetic gut metagenome created using the 20 most abundant species from HMP project. Carnelian's runtime closely matched with HUMAnN2 and is better than mi-faser's standalone binary (Table 3.60). Kraken2 is the fastest among all but significantly limited in terms of performance (Section 3.2.8). As read length increases, the running times of all the methods tend to increase.

In both cases, the memory requirement of the other three methods is smaller than Carnelian because they all use reduced size amino acid alphabets trading off sensitivity. We can get similar memory gains by using reduced size amino acid alphabets without significant loss of precision (Table 3.61).

Table 3.59: Runtime and memory requirement of Carnelian compared to mi-faser, HUMAnN2, and Kraken2 when the size of input data remains fixed but read length is varied.

Read length	**Method**	**Elapsed clock time (min)**	**Maximum resident set size (GB)**
100 bp	Carnelian	12.43	13.4
	mi-faser	22.01	0.17
	Humann2	12.08	3.33
	Kraken2	1.34	0.47
150 bp	Carnelian	15.67	13.46
	mi-faser	29.38	0.2
	Humann2	17.65	4.94
	Kraken2	2.15	0.47
200 bp	Carnelian	20.77	13.7
	mi-faser	37.63	0.28
	Humann2	24.35	6.49
	Kraken2	2.96	0.46
250 bp	Carnelian	33.17	13.7
	mi-faser	44.65	0.34
	Humann2	30.15	6.53
	Kraken2	3.75	0.46

Table 3.60: Runtime and memory requirement of Carnelian compared to mi-faser, HUMAnN2, and Kraken2 when read length remains fixed but the size of input data varies.

# Reads	**Method**	**Elapsed clock time (min)**	**Maximum resident set size (GB)**
1 Million	Carnelian	9.30	13.7
	mi-faser	9.01	0.32
	Humann2	6.51	2.15
	Kraken2	0.77	0.43
5 Million	Carnelian	33.17	13.7
	mi-faser	44.65	0.34
	Humann2	30.15	6.53
	Kraken2	3.75	0.45
10 Million	Carnelian	54.85	13.7
	mi-faser	92.48	0.35
	Humann2	58.9	6.54
	Kraken2	7.43	0.46

Table 3.61: Performance of Carnelian with reduced-size amino acid alphabets on our cross-validation test set containing ~3M 100 bp fragments.

Model	**Alphabet Size (AA)**	**Sensitivity (%)**	**Precision (%)**	**F1-score (%)**	**Peak (Memory) (GB)**
Full Alphabet	20	98.86	97.86	98.26	7.57
MWL2000[1]	15	86.77	99.52	92.12	0.76
MWL2000[1]	10	86.27	99.59	91.83	0.76
MWL2000[1]	8	78.37	99.18	86.49	0.76
Physico-Chemical[2]	5	86.65	99.53	92.03	0.76
HP Model[3]	2	75.22	98.98	84.22	0.76
mi-faser, HUMAnN2	11[4]	96.78	99.95	98.16	1.31

[1] [1] MWL2000: Murphy, L. R., Wallqvist, A., and Levy, R. M. (2000). Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Engineering, 13(3), 149-152.

[2] [2] Physio-Chemical: Amino acids grouped according to 5 physico-chemical properties; A (Aliphatic): IVL, R (aRomatic): FYWH, C (Charged): KRDE, T (Tiny): GACS, D (Diverse): TMQNP

[3] [3] HP Model: Groups amino acids as polar (hydrophilic) or hydrophobic; P: AGTSNQDEHRKP, H: CMFILVWY

[4] [4] DIAMOND aligner, used by mi-faser and HUMAnN2 (translated), inherently represents the proteins in its database with a reduced amino acid alphabet of size 11.

3.3 Methods

3.3.1 Main components of the Carnelian pipeline

We present here the full pipeline of Carnelian for whole metagenome comparative studies. Our pipeline combines more tailored database curation, probabilistic gene finding, alignment-free gapped k-mer-based functional metagenomic binning, abundance estimation, and appropriate statistical tools for performing comparative functional metagenomics. Figure 3-1 depicts the main components of our pipeline. The heart of our pipeline is a novel compositional (gapped k-mer-based) tool for functional metagenomic binning. It incorporates probabilistic ORF finding with a compositional gapped k-mer classifier ensemble to bin reads into different Enzyme Commission (EC) groups according to their gene content (if any).

Carnelian represents gold standard proteins with complete EC labels in a low-dimensional compact feature space by leveraging Opal-Gallager hashes [85, 86], a class of even-coverage, low-density, and locality-sensitive hashes [205]. These hashes guarantee that there is a high probability of collision for input sequences which are similar to each other in the k-mer space and a low collision probability for dissimilar sequences. These features are then used to train an ensemble of one-against-all classifiers (support vector machines). We implemented the classifier ensemble using the Vowpal Wabbit (v8.1.1) framework [206, 207]. Negative examples were generated using the "shuffle" program from the HMMER package [192]. The classifiers are trained in an online fashion (one example in memory at a time) using stochastic gradient descent (SGD). The online training capability makes incremental training of Carnelian easy as new verified EC annotations for proteins become available. For more details of the parameters of the classifier ensemble, see Appendix B.

To functionally profile WMS reads, Carnelian first uses FragGeneScan [191] to detect the best possible ORFs from them. FragGeneScan is a unified hidden Markov model framework that incorporates codon usage bias and sequencing error models to probabilistically detect the coding part(s) of the reads. As part of our pipeline, FragGeneScan is run with 'short reads' option, because our input is short WMS reads. Since the average substitution error rate for Illumina sequencing is $\sim 0.1\%$, we used the 'complete' option with FragGeneScan, which assumes 0% error rate. The ORFs predicted by FragGeneScan are encoded into the same compact feature space as in training using Opal-Gallager hashing. Carnelian employs the trained classifier ensemble to bin the feature vectors of the ORFs by EC labels.

All else being equal, the more abundant proteins from an EC bin in the microbial sample is, the more reads from them are likely to be sequenced. Therefore, read counts can be used as a proxy for EC abundance in the sample — used by standard functional annotation tools (e.g., mi-faser). However, in practice "all else" are never equal. These counts need to be made comparable across proteins, samples, and experiments to enable meaningful comparative analysis downstream. Hence, we borrow intuition from transcriptomics and have Carnelian construct a functional vector by normalizing the read counts as follows:

$$\text{Effective protein length in EC bin } b,\ e_b = p_b - \frac{rl}{3} + 1$$

$$\text{Abundance of EC bin } b,\ \rho_b = \frac{\frac{r_b}{e_b} \times 10^6}{\sum_b \frac{r_b}{e_b}}$$

Here, p_b is the effective protein length (in amino acids) of EC bin b, and rl is the average read length (in base pairs). This equation takes into account the effect of effective protein length in an EC bin as well as the lengths of the proteins in other EC bins while calculating the relative abundance of an EC label in a sample. This normalization further ensures that the relative abundances of the ECs sum up to the same amount in every microbial sample making the proportions directly comparable across samples.

To understand why this normalization is important let's look at the following scenarios in which using raw read counts as a proxy for EC abundance provides inaccurate estimates.

Scenario 1: Suppose, a microbial sample has only two proteins (from two different ECs) in equal proportion. These protein sequences have different lengths. If we sequence the sample, with high probability, we see more reads from the longer protein (thus more reads from the corresponding EC). If we take raw read counts as a proxy for relative EC abundance, we mistakenly assume that the EC with the longer protein is more abundant than the EC with the shorter one. For this reason, we need to normalize the read counts in an EC bin by the effective protein length (the positions in the protein sequence to which a read can map) of that bin. This value is often known as RPK when the length is measured in kilobases (used by different methods such as HUMAnN2).

Scenario 2: Suppose, we have reads from two experiments with different sequencing depths—one experiment has 10x more reads than the other. If we want to compare the relative abundance of the same EC across experiments, just normalizing by effective protein length in the corresponding EC bin does not change anything. The higher the total number of reads, the higher read count and normalized read count we see for any given EC. For relative abundances to be comparable across experiments, they need to be on the same scale.

Scenario 3: Suppose we have two microbial samples, each with two types of proteins (from two different ECs). Sample 1 has red and yellow proteins, and sample 2 has red and green proteins. The lengths of red, yellow, and green proteins are 10, 50, and 250 units respectively. Let's say, we observe 300 reads from both samples, and we want to compare the abundance of red proteins across samples. If we observe 50 reads from the red protein in both the sample, the RPK values for red protein will be the same across samples. We observe 250 reads from the yellow protein in sample 1 which means the relative abundance of red protein is much less compared to yellow protein here (RPK for red protein = 1×10^3 vs. RPK for yellow protein = 25×10^3). In sample 2, we observe 250 reads from the green protein, which means both red and protein have the same relative abundance (RPK for both proteins is 1×10^3). This means sample 2 has a higher abundance of red protein, which we will not be able to tell if we only compare the RPK values. The RPK values of other proteins in the sample affect the relative abundance of a protein in question. If we normalize by the sum of all the RPK values in the sample, then we can see the desired difference (normalized RPK values of red protein in sample 1 and 2 are $\frac{1}{26}$ and $\frac{1}{2}$ respectively).

Carnelian's effective count normalization takes the above scenarios into account and normalizes read counts by effective protein length in the EC bin and a per million scaling factor which incorporates the sum of all RPK values in the sample. This normalization ensures that the relative abundances of the EC bins in every sample effectively sums up to the same number making them directly comparable across samples and experiments.

Supporting Experiment: To show how well Carnelian's effective counts normalization works in practice, we conducted the following experiment. We randomly selected an individual from our Bostonian cohort. The original read data set contained $\sim$9M paired-end reads of length 150 base pairs. We created another read data set by performing 20x down-sampling such that the new subsampled data set has 450k reads. Ideally, the relative abundance of all ECs should be the same in these two samples, and we should observe a log fold-change (logFC) of zero (0) for all of them. We used raw read counts (used by mi-faser), RPK measure (used by HUMAnN2), and our effective read counts (TPM measure) as proxies for relative abundance and measured the logFC value for all the ECs in each case. While nearly every EC appears variable between the original and the subsampled data set in terms of raw read counts and RPKs, only Carnelian's effective counts show the expected behavior.

	Raw count	**RPK**	**Effective Count (Carnelian)**
Mean logFC	-1.1251	-1.1251	0.0094
Stddev logFC	0.4659	0.4659	0.1918

Some examples:

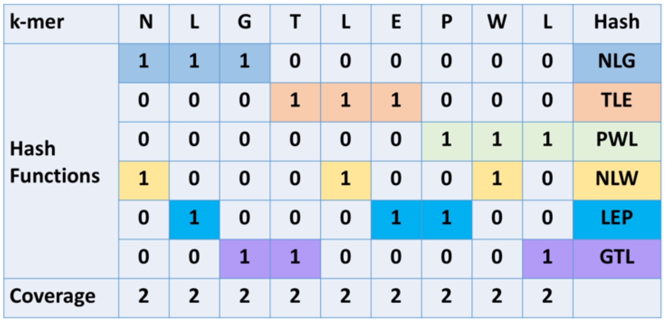
EC	Raw Count		RPK		Effective Count (Carnelian)	
	Original	**Subsampled**	**Original**	**Subsampled**	**Original**	**Subsampled**
3.6.3.19	1090	56	2939.99	151.05	2072.43	2120.91
1.2.1.11	876	45	2442.38	125.46	1721.66	1761.72
5.2.1.8	935	48	3231.71	165.91	2278.07	2329.58
3.6.3.25	2689	138	8280.22	424.94	5836.82	5966.87
4.1.1.87	78	4	187.95	9.64	132.49	135.34
2.7.7.24	3096	153	10927.06	540	7702.6	7582.46
4.2.1.24	466	23	1401.5	69.17	987.95	971.3
6.1.1.19	223	11	389.37	19.21	274.47	269.69
4.2.1.8	981	48	2468.74	120.79	1740.24	1696.15
2.5.1.47	3148	154	10092.68	4937.34	7114.44	6932.81

3.3.2 Database curation

We built our gold standard reference data set by first collecting reviewed prokaryotic proteins from UniProtKB/Swiss-Prot (Feb. 2018) [208, 209] that had both experimental evidence of existence at either the protein or the transcriptomic level and complete EC Numbers associated—EC numbers act as the primary identifiers for metabolic pathway members. We excluded any protein that had computationally inferred functional labels (e.g., by homology), an incomplete EC label, or multiple EC annotations. Indeed, some proteins can have multiple functions. However, these proteins primarily act as enzymes, and the secondary functions are mainly non-enzymatic. Therefore, we can safely assume that a protein will have a unique EC label in the reference database. We also collected prokaryotic catalytic residues with complete EC numbers for which a literature reference existed from the Catalytic Site Atlas. We combined these two sets and removed any redundant sequences, which gave us a reference data set, EC-2010-DB, consisting of 7,884 proteins with 2,010 unique EC numbers (both the dataset and a pre-trained model to bin reads into EC labels are available on the Carnelian's website). Amino acid sequences for these proteins were downloaded from UniProt [208]. This database is designed for profiling the metabolic functional capacity of the microbiome and more suited for cross-comparing healthy and disease microbiomes. Additionally, we provide a database of 1,785,722 proteins from 3,285 COG categories and a pre-trained model to classify reads into COG categories on our website (`http://carnelian.csail.mit.edu`), which can be used for microbial functional profiling beyond metabolism.

3.3.3 Constructing feature vectors using Opal-Gallager hashes

Let us consider a sequence fragment of l amino acids, $s \in \Sigma^l$, where $\Sigma =$ standard amino acid alphabet ($|\Sigma| = 20$). A k-mer, with $k < l$, is a short word of k contiguous amino acids. Similar to the bag-of-words representation of a document, we define a k-mer profile of a sequence s as a vector $f_k(s) \in \mathbb{R}^{20^k}$. We index each k-mer with an integer i, where $0 \leq i \leq 20^k$ which can be represented by a binary string of length $5k$. Each entry $f_k(s, i) \in f_k(s)$ stores the frequency of the i-th k-mer. Thus, an amino acid fragment of length l can be represented using k-mers in $O(20^k)$ space instead of a vector of $O(20^l)$. Using random locality-sensitive hash (LSH) functions, we can create k-mer profiles that specify spaced subsequences, rather than contiguous subsequences of fragment s. More specifically, we define a random hash function, $h : \Sigma^k \to \Sigma^r$ to generate a spaced (k, r)-mer such that a hashed k-mer can be represented by a binary vector of $O(20^r)$ dimensions with corresponding positions set to 1. Here r denotes the number of positions selected within a k-mer window. With this family of LSH functions, we can randomly sample a set of m LSH functions and concatenate them together to represent a k-mer profile of a sequence by only $O(m20^r) \ll O(20^k)$ space. However, k-mer profiles built with uniformly random LSH functions often have uneven coverage of positions in a sequence unless a large number of such functions are used. To evenly cover positions using a small number (m) of LSH functions, we build upon Opal's modified Gallager design algorithm [86]. Figure 3-16 depicts an example of how even coverage LSH functions are generated for an amino acid k-mer. We used a $k = 8$ and $r = 4$ for the purpose of this study.



k-mer	N	L	G	T	L	E	P	W	L	Hash
Hash Functions	1	1	1	0	0	0	0	0	0	NLG
	0	0	0	1	1	1	0	0	0	TLE
	0	0	0	0	0	0	1	1	1	PWL
	1	0	0	0	1	0	0	1	0	NLW
	0	1	0	0	0	1	1	0	0	LEP
	0	0	1	1	0	0	0	0	1	GTL
Coverage	2	2	2	2	2	2	2	2	2	

Figure 3-16: **Example of low-density even coverage hashing representation of an amino acid k-mer.** A (k, r)-hash function can be thought of as a binary vector of length k with r 1's where each 1 indicates a marked position in the k-mer. Leveraging the hashing technique from Opal, Carnelian starts with a set of hash functions, represented as a hash matrix where the first row has 1's in first r positions, the second row has 1's in second r positions, and so on. Carnelian then permutes the columns of this matrix repeatedly to generate even coverage LSH functions. The rows then give the corresponding hashes of a k-mer.

Choice of fragment length and k-mer length

In the training phase, Carnelian trains an ensemble of classifiers in multiple batches. Since our current gold-standard is small in size, to ensure that the classifier ensemble sees enough examples per batch, we draw random fragments from the gold-standard proteins by making sure that all the reference proteins have sufficient representation in the training batches. While choosing the fragment length (l), we needed to ensure that the fragments we drew were smaller than the smallest protein sequence in our data sets. Lengths of 7,884 protein sequences in EC-2010-DB ranged from 34 to 7,073 residues with a median length of 342 residues. For this reason, we used $l = 30$ in the training phase.

Our choice of the value of k needs to be such that the chance of k-mers being shared by any two protein sequences in our gold-standard data sets is minimized. In a study of 1,121 bacterial genomes, Greenfield et al. [210] showed that for a k-mer length of > 20 nucleotides (≥ 7 amino acid residues), over 96% of the nucleotide k-mers within an organism are unique and only less than 0.2% of the k-mers of length 25 nucleotides (≥ 8 amino acid residues) are shared by any two organisms; the 25-mers have the same gene annotation in both genomes. Inspired by these results, we chose $k = 8$ for our experiments. For $k = 8$, we can calculate the probability, p of a random k-mer match within a data set as follows [211]:

$$p = \frac{1}{\frac{|\Sigma|^k}{g} + 1}$$

where, $|\Sigma|$ is the alphabet size and g is the size of the data set in total number of amino acid fragments. For a gold-standard database containing 32,111,182 randomly sampled amino acid fragments from the EC-2010-DB data set, this probability is 0.12%, which is sufficiently small. For flexibility, Carnelian takes fragment length as input from the user.

3.3.4 Setup for benchmarking experiments

We benchmarked our compositional functional profiler, Carnelian against state-of-the-art alignment-based tools, mi-faser and HUMAnN2, and a state-of-the-art k-mer-based tool, Kraken2, using our gold standard database, EC-2010-DB on a number of synthetic metagenomes. Off-the-shelf HUMAnN2 and Kraken2 use taxonomic information in addition to translated searches; to ensure fair comparison we used only their "translated-search" or "protein-search" mode. All comparisons were based on the EC terms identified by each method using the same gold standard reference database. That is to say, the reference databases we used for the mi-faser and HUMAnN2 and the Kraken2 reference indexes were created with Carnelian's gold standard reference database for unbiased comparison. Detailed performance benchmarks for Carnelian against mi-faser, HUMAnN2, and Kraken2 are available in Section 3.2.8. The exact commands used for running mi-faser, HUMAnN2, and Kraken2 are given in Appendix B and scripts are available on our website.

3.3.5 Functional profiling of real data sets

We explored two large-scale type-2 diabetes (T2D) studies, two Crohn's disease (CD) studies, and a Parkinson's disease (PD) study for investigating functional dysbiosis in disease vs. healthy microbiomes. We analyzed whole metagenome sequencing data from fecal samples of 347 individuals from a Chinese T2D study cohort [182]. Raw paired-end Illumina reads were downloaded from the NCBI short read archive (SRA) (Study accession: SRP008047). We labeled this data set T2D-Qin. Additionally, we analyzed fecal metagenome sequencing data from a T2D study performed on a European cohort of 145 women with either T2D or impaired glucose tolerance (IGT) or normal glucose tolerance (NGT) [183]. Since we aimed at finding the differences in microbial metabolic function between T2D patients and healthy individuals, we did not include the IGT individuals in our analysis. We downloaded publicly available raw Illumina HiSeq 2000 paired-end reads from NCBI SRA (Study accession: ERP002469); each individual metagenome contained ~ 3 Gb on average. We labeled this data set T2D-Karlsson. We further analyzed two Crohn's disease case-control data sets: 53 US individuals from HMP pilot phase and 62 Swedish individuals from a Swedish cohort [178]. We downloaded publicly available raw Illumina HiSeq 2000 paired-end reads for the US cohort (CD-HMP data set) from the IBDMDB website [196]. Raw reads for the Swedish cohort (CD-Swedish) were downloaded from NCBI SRA (Study accession: SRP002423). We also analyzed whole metagenome sequencing reads from the fecal samples of 20 patients and 21 healthy individuals in an early stage L-DOPA naïve PD case-control study [200]. All the participants in the study were male and age-matched. We downloaded publicly available raw Illumina HiSeq 2500 paired-end reads from NCBI SRA (Study accession: ERP019674). We labeled this data set PD-Bedarf.

For investigating the functional relatedness of the healthy microbiomes in industrialized and non-industrialized communities, we analyzed gut microbiomes of four cohorts (84 individuals from Boston, 35 Baka individuals from Cameroon, 50 individuals from the Gimbichu region in Ethiopia, and 112 individuals from Madagascar of Betsimisaraka and Tsimihety ethnicity). The first two data sets were unpublished data sets from the Alm lab. The latter two data sets were contributed by a recent study [26] and are publicly available at NCBI SRA with study accessions SRP168387 and SRP156699.

We also explored the functional diversity of two environmental data sets — (i)

VAG-pond data set: eight aquatic metagenomes from a study of microbial diversity in a pond created by chrysotile asbestos mining activity at the Vermont Asbestos Group (VAG) Mine in northern Vermont, USA [185] and (ii) DWH-spill data set: six beach sand metagenomes from a study of the Deepwater Horizon oil spill [202]. For the aquatic metagenomes in the VAG-pond data set, we obtained Illumina HiSeq 2000 paired-end reads of length 101 bp from NCBI SRA (Study accession: SRP056095). The data set includes three samples from the epilimnion (surface layer), three samples from the metalimnion (middle layer), and two samples from the hypolimnion (bottom layer). In addition, as in the original study [185], we included two positive controls—(i) 1-F: single-end Illumina reads from a synthetic microbial sample simulating organisms found in the Delaware River (downloaded from BaseSpace with a free account: `https://basespace.illumina.com/projects/20039022/samples`); (ii) 2-F: a set of single-end Illumina reads from the Human Microbiome Project (HMP) mock community (downloaded from NCBI SRA with accession SRR172902). For the DWH-spill data set, we obtained Illumina HiSeq 2000 paired-end reads of length 151 bp for the beach sand metagenomes (two from pre-oil phase, two from oil phase, and two from post-oil phase) from NCBI SRA (Study accession: SRP046227).

Metadata of the samples from each study are available at `http://carnelian.csail.mit.edu/data/metadata.pdf`.

Preprocessing steps for raw reads

We used Trimmomatic v0.36 [212] for adapter trimming and quality filtering with a quality threshold of 30 and a minimum length of 60 bp (paired-end mode for Illumina reads and single-end mode for Roche 454 reads). DeconSeq v0.4.3 [213] was used to remove contaminating human sequences with the human reference genome GRCh38 as the database. For paired-end reads, we kept only the read-pairs for which both sequences survived quality control. These steps were applied to all the data sets. In the T2D-Qin data set, 241 of the samples survived the preprocessing step and were used for subsequent analyses.

3.3.6 Quantifying microbial functional variation in real data sets

Carnelian outputs the effective read counts per EC label (i.e., normalized read counts against effective protein length per EC bin and a per million scaling factor) as abundance estimates. For the other three methods, we applied the same normalization on the raw read counts produced by them to ensure an unbiased comparison. Pathway abundances were calculated by grouping the ECs into KEGG metabolic pathways and summing the effective read counts. Pathway coverage was calculated as the ratio of the number of mapped ECs identified by a method to the total number of reference ECs present in the pathway.

For the studies with two groups of microbiomes (case vs. control, industrialized vs non-industrialized), we created an effective counts matrix using Carnelian generated functional profiles and performed pairwise Wilcoxon rank-sum test (Mann-Whitney U test). A Benjamini-Hochberg (BH) false discovery rate (FDR) corrected p-value threshold of 0.05 was used as a test of significance. Additional log-fold-change thresholds have been selected for each data set (mentioned in the main text).

To determine the significance of the common pathways between geographically separated disease cohorts, we combined the individual p-values per pathway from different studies of the same disease using Fisher's combined probability test (Figure 3-1: Green). To investigate the co-abundance of microbial metabolic pathways between healthy microbiomes of industrialized and non-industrialized communities, we computed Kendall's rank correlation of the pathway abundance profiles of the two groups. Next, we performed Ward-linkage hierarchical clustering using Euclidean distance on the pathway co-abundance matrix (correlation matrix). To determine whether the centroids and dispersion of the pathway clusters are significantly different between the non-industrialized and industrialized microbiomes, permutational multivariate analysis of variance (PERMANOVA) test was performed using "adonis" function available through the "vegan" package in R (Figure 3-1: Blue). For measuring functional diversity in a sample, we calculated the Shannon-Wiener diversity indices of the EC and pathway profiles of the samples using the “vegan” package available in R.

3.4 Discussion

While the rapid advancement in sequencing technologies has helped researchers resolve the taxonomic diversity of microbial “dark matter” to a great extent, much of its functional diversity remains uncharacterized [24–26]. Even for the minimal bacterial genome designed by Hutchison et al. [28], the function of one-third of the genes could not be determined. Thus, functional annotation remains a challenging task even for well-studied genomes, and, unsurprisingly, the sensitivity of all relevant methods is low across the board. Potential reasons why reads often cannot be mapped to functional labels include unknown functionality, non-metabolic functionality, lack of coverage in reference databases, or a non-prokaryotic origin. It is possible to use a much more extensive off-the-shelf protein database containing computationally predicted functional labels, but doing so is not always advisable because incorporating such databases can increase the chance of erroneous transfer of spurious annotations [70, 214].

More than merely providing an alternative functional profiling tool, Carnelian is able to capture hidden microbial metabolic functional diversity from whole metagenome sequencing reads through its use of a gapped k-mer classifier. Being able to label additional ECs accurately manifests partially as an increase in Carnelian’s sensitivity. Additional sensitivity alone is suspect, due to the possibility of spurious labels, but we believe that our stricter criteria for database inclusion, combined with training negative examples to reduce false positives, contributed significantly to Carnelian being able to assign a functional label to unknown proteins while minimizing false positives. Indeed, we believe that this ability makes Carnelian a potential tool for annotating novel microbial proteins that are increasingly becoming available [29]. Also, we believe that it is partially due to this ability that, unlike existing methods, Carnelian is able to create functional profiles that are comparable across populations. In multiple large-scale comparative experiments, Carnelian uncovers shared and novel functional similarities and differences across diverse populations and environmental conditions that would go unseen when using existing tools, which are often implicitly designed around taxonomic profiling.

Carnelian detected a high degree of similarity in core metabolic pathways between healthy guts in industrialized and non-industrialized communities, despite significant taxonomic differences [26,80,198]. This result is notable given the differences in external pressures (e.g., diet, lifestyle, exposure to toxins) and may indicate the adaptive nature of the gut microbiome. Indeed, many of the enzyme-level variations we found did suggest an adaptive response to industrialized vs. non-industrialized dietary differences in carbohydrates (simple sugars vs. complex monosaccharides) and proteins (protein-rich vs. protein-deficient); this finding agrees with earlier studies [26, 198]. By using different enzymes involved in core metabolic pathways, the healthy guts in these communities can better maintain the overall balance in core metabolic functionality.

We did observe differential read abundance in several xenobiotics metabolism pathways between industrialized and non-industrialized microbiomes (Table 3.39). For example, non-industrialized microbiomes showed enrichment of reads in antibiotic resistance ECs and pathways (e.g., beta-lactamase, drug metabolism by cytochrome P450). On the other hand, we observed higher read abundance in lipoic acid metabolism, xenobiotics metabolism by cytochrome P450, and phenylpropanoid biosynthesis pathways in the industrialized gut. These findings agree with earlier studies [26, 80, 198]. A potential line of future inquiry would be to investigate these similarities and differences with much larger sample sizes, but such is beyond the scope of this study.

Our results with Carnelian indicate concordant dysbiosis in several microbial carbohydrate metabolism pathways in both Chinese and European cohorts for type-2 diabetes. Though existing methods identified variable read abundances in several carbohydrate metabolism pathways, they did not find any common pathways which were statistically significant in both the cohorts. T2D patient guts were found to have higher read abundance in the oxidative phosphorylation pathway, suggesting a higher degree of bacterial defense against oxidative stress and a more significant energy imbalance in the patient gut [182,183]. While the shared dysbiosis in vitamin B metabolism pathways might not be directly related to the disease process, it could be a side-effect of prolonged metformin use by T2D patients in both cohorts [182,215].

In Crohn's disease case-control cohorts from the US and Sweden, Carnelian uncovered reduced functional potential of several specific carbohydrate metabolism pathways and amino acid biosynthesis pathways; other tools did not find any concordant dysbiosis. Our results make sense given that microbial carbohydrate metabolism, amino acid synthesis, and selenocompound metabolism pathways were already known to be associated with Crohn's disease [216,217]. Valine, leucine, and isoleucine have anti-inflammatory roles and are required for intestinal growth and maintenance of mucosal integrity and barrier function; dietary amino acids have been found to be beneficial for inflammatory bowel disease (IBD) animal models [218]. Additionally, dysbiosis in the microbial biosynthesis of N-glycan can affect the intestinal health of CD patients [219].

For Parkinson's disease (PD), Carnelian's results indicate a downward shift in the gut microbial capacity to synthesize tryptophan, which was not found by mi-faser or HUMAnN2 (both the translated search and the full out-of-the-box pipeline). Microbial tryptophan metabolism has been associated with a number of diseases [220], and in particular, for Parkinson's, this might affect serotonin production in the host as tryptophan is a known precursor of serotonin. We also found microbial carbohydrate metabolism to be altered in Parkinson's disease which might be a contributor to the insulin impairment observed commonly in Parkinson's patients [221]; Glucagon-like peptide-1 receptor agonists, which act in the gut-brain axis pathway and regulate blood glucose, have shown therapeutic potential in clinical studies of PD [222].

Of course, though we find a significant alteration in the functional capacity of these microbial metabolic pathways, these diseases cannot be characterized by these shifts alone. Integrative approaches involving metabolomics, metagenomics, and metatranscriptomics will likely be required to establish causal relationships between microbial pathways and disease processes in the host. Since disease-associated shifts can often be confounded by antibiotics and other drug usages by participants in a case-control study, the results must be interpreted carefully. Despite these challenges, we were able to show that it is possible to find concordant functional trends across geographically separated case-control cohorts. Our study opens the door to a future where bioprospecting efforts using natural microbes, genetically engineered bacteria, or microbial products targeting specific metabolic pathways in a broad therapeutic context may become possible.

3.5 Conclusions

In this chapter, we have presented a full pipeline for whole metagenome comparative studies. By integrating together more tailored database curation, probabilistic gene finding, alignment-free functional metagenomic binning, abundance estimation, and the appropriate statistical tools, we show that on a variety of data sets, our tool provides a more comprehensive picture of the functional relatedness of healthy and disease microbiomes than cannot be achieved using existing tools, which implicitly rely on taxonomic binning. Carnelian's modular design enables flexibly running each step of the pipeline independently. For instance, it can be run on either raw sequencing reads (default) or transcriptomic sequences (by bypassing the ORF detection phase). Alternately, should a user prefer to employ other functional profiling tools instead of Carnelian, other components of our pipeline, such as the database curation and statistical tests, may still be of use.

To demonstrate the usefulness of our pipeline, we also analyze a variety of data sets, some publicly available and some newly collected. For type-2-diabetes and Crohn's disease, earlier studies showed only a moderate degree of taxonomic dysbiosis, which did not generalize across different geographic cohorts. With Carnelian, we newly identify concordant changes in the functional capacity of 13 metabolic pathways in European and Chinese type-2 diabetes cohorts and eight metabolic pathways in US and Swedish Crohn's disease cohorts. Moreover, Carnelian was able to identify several clinically established hallmarks of Parkinson's disease that were not found by other state-of-the-art functional annotation tools. Carnelian-identified EC terms can be used to classify patients and controls with high accuracy. In healthy microbiomes from industrialized and non-industrialized communities, Carnelian identified more functional diversity at both the EC and pathway levels compared to other methods and revealed a high degree of pathway-level similarity in core metabolic functionality.

Carnelian's unique ability to find functional relatedness in diverse metagenomic data sets at the scale of hundreds of samples opens the door to more comprehensive comparative functional metagenomic studies across different geographies, environmental conditions, and time points. We expect Carnelian to be an essential component of the metagenomic analysis toolkit, especially when cross-population comparisons are performed.

3.6 Software Availability

Carnelian is open source and freely licensed (MIT License). Source code of Carnelian is available at `http://carnelian.csail.mit.edu` and `https://github.com/snz20/carnelian` (DOI:10.5281/zenodo.3371731).

Chapter 4

Conclusion

The central theme of this thesis has been developing computational tools that can perform robust integration of heterogeneous omics data and reveal meaningful functional insights in large-scale comparisons. We have focused on two pressing challenges in life science and provided comprehensive solutions for each of them. Firstly, we addressed the challenge of finding a shared molecular connection between autism spectrum disorder (ASD) and its seemingly unrelated multi-system comorbidities. In Chapter 2, we presented a novel three-tiered integrative omics analysis pipeline that can integrate transcriptomic data from disparate sources at the gene, pathway, and disease levels in a statistically principled fashion. By for the first time integrating data across 53 transcriptomic studies of twelve disease conditions, our pipeline revealed a novel innate immunity connection between ASD and its highly prevalent comorbidities. Secondly, we addressed the challenge of functionally profiling whole metagenome sequencing reads to enable large-scale comparisons across samples, experiments, populations, and environmental conditions. In Chapter 3, we introduced Carnelian, a comprehensive framework for functional comparisons of whole metagenome sequencing data from diverse study cohorts. The heart of our framework is a new compositional (gapped k-mer) classifier model for alignment-free functional metagenomic binning that accurately classifies microbial proteins, especially from non-annotated species. By newly integrating more tailored database curation, probabilistic gene finding, alignment-free functional metagenomic binning, abundance estimation, and the appropriate statistical tools, Carnelian provides a more comprehensive picture of the functional relatedness of healthy and disease microbiomes than can not be achieved using existing tools. Carnelian uniquely enables finding concordant patterns of microbial metabolic function that can generalize across geographical borders and complement taxonomic studies.

Due to continued technological advances, we now have access to unprecedented amounts of omics data. UK Biobank currently provides access to different types of omics data as well as demographic and clinical data from 500,000 individuals in the United Kingdom, and by 2022 we are going to have access to data on at least 1 million individuals through the European infrastructure [223]. Moreover, access to explosive amounts of omics data at single-cell resolution through consortia like the Human Cell Atlas [224] is opening up new avenues for a more in-depth understanding of the functioning of cells, thereby establishing the mechanistic links between cell states and diseases. As we move into the era of omics-based precision healthcare, the challenge we face is how to integrate and interpret all these data to obtain meaningful insights into health and disease. The intuitive way to address this challenge is through mechanistic approaches, but there is still a lack of such efforts. We need computational tools that can process large-scale data sets rapidly and robustly, and identify accurate and meaningful functional patterns in large-scale comparisons. This thesis is one step in that direction.

Appendix A

Supplementary information for multi-level integrative omics analysis for ASD and its comorbidities

Microarray expression data sets of ASD and its comorbidities from the GEO

The list of selected microarray studies for ASD and its comorbid diseases is provided in Table A.1. For each selected GEO series, the table lists the accession identifier as well as abridged study details including the organism, tissue type, platform, and the number of samples.

Table A.1: Selected GEO series for ASD and its comorbidities.

Disease	**Accession**	**Platform**	**Organism**	**Tissue**	**# Samples**
ASD	GSE25507	Affymetrix	Homo sapiens	Peripheral blood lymphocytes	146
	GSE7329	Agilent	Homo sapiens	Lymphoblastoid cells	30
	GSE28521	Illumina	Homo sapiens	Brain tissue	79
	GSE26415	Agilent	Homo sapiens	Peripheral blood leucocyte	84
	GSE6575	Affymetrix	Homo sapiens	Blood tissue	47[1]
	GSE18123	Affymetrix	Homo sapiens	Peripheral blood	285
Asthma	GSE19187	Affymetrix	Homo sapiens	Nasal epithelial cells	241
	GSE27011	Affymetrix	Homo sapiens	White blood cells	54
	GSE45251	Agilent	Homo sapiens	Airway smooth muscle cells	16
	GSE470	Affymetrix	Homo sapiens	Epithelial tissue	12
	GSE8190	Agilent	Homo sapiens	Airway epithelial cells	250
Bacterial & Viral Infection	GSE40396	Illumina	Homo sapiens	Whole blood	65
	GSE42026	Illumina	Homo sapiens	Whole blood	92
	GSE47172	Affymetrix	Homo sapiens	Whole blood	15
	GSE34205	Affymetrix	Homo sapiens	PBMC	101
Chronic Kidney Disease	GSE43484	Affymetrix	Homo sapiens	Monocytes	6
	GSE41030	Agilent	Homo sapiens	Fibroblasts	6
	GSE38117	Agilent	Mus musculus	Renal tissue	6
	GSE48041	Agilent	Mus musculus	Kidney tissue	24
	GSE15072	Affymetrix	Homo sapiens	Peripheral blood mononuclear cells	29

(continued on next page)

Table A.1: *cont.* Selected GEO series for ASD and its comorbidities.

Disease	**Accession**	**Platform**	**Organism**	**Tissue**	# **Samples**
Cerebral Palsy	GSE16447	Affymetrix	Homo sapiens	Fibroblasts	9
	GSE31243	Affymetrix	Homo sapiens	Skeletal muscle biopsies	40
Dilated Cardiomyopathy	GSE29819	Affymetrix	Homo sapiens	Heart tissue	38
	GSE42955	Affymetrix	Homo sapiens	Heart tissue	29
Ear Infection	GSE49128	Affymetrix	Mus musculus	Middle & inner ear tissue	17
Epilepsy	GSE32534	Affymetrix	Homo sapiens	FFPE peritumoral neurocortex tissue	10
	GSE6834	Ion channel splice array	Homo sapiens	Temporal cortex brain tissue	201
	GSE6614	Affymetrix	Mus musculus	Brain tissue	28
	GSE47516	Affymetrix	Mus musculus	Cerebellum & granule neurons	21
	GSE20977	Illumina	Homo sapiens	CNV	15
	GSE22225	Affymetrix	Homo sapiens	Lymphocytes	15
	GSE16969	Affymetrix	Homo sapiens	Cortical tubers	10
Inflammatory Bowel Disease	GSE11223	Agilent	Homo sapiens	Colon epithelial biopsies	202
	GSE3365	Affymetrix	Homo sapiens	PBMC	127
	GSE38713	Affymetrix	Homo sapiens	Intestinal mucosa	43
	GSE9452	Affymetrix	Homo sapiens	Colonic mucosa	26
Muscular Dystrophy	GSE42806	Affymetrix	Homo sapiens	Muscle tissue	12
	GSE36398	Affymetrix	Homo sapiens	Muscle tissue	50
	GSE9397	Affymetrix	Homo sapiens	Muscle biopsies	20
	GSE6011	Affymetrix	Homo sapiens	Muscle biopsies	37

(continued on next page)

Table A.1: *cont.* Selected GEO series for ASD and its comorbidities.

Disease	**Accession**	**Platform**	**Organism**	**Tissue**	# **Samples**
Schizophrenia	GSE53987	Affymetrix	Homo sapiens	Prefrontal cortex, Striatum, Hippocampus	1031
	GSE27383	Affymetrix	Homo sapiens	PBMC	72
	GSE48072	Illumina	Homo sapiens	Blood	66
	GSE46509	Affymetrix	Homo sapiens	Parvalbumin-immunoreactive neurons	16
	GSE37981	Affymetrix	Homo sapiens	Pyramidal cells in superior temporal cortex	18
	GSE21935	Affymetrix	Homo sapiens	Post-mortem brain tissue	42
	GSE25673	Affymetrix	Homo sapiens	hiPSC-derived neurons	24
	GSE21138	Affymetrix	Homo sapiens	Prefrontal cortex	59
	GSE17612	Affymetrix	Homo sapiens	Post-mortem brain tissue	51
	GSE12654	Affymetrix	Homo sapiens	Prefrontal cortex	28[1]
Upper Respiratory Infection	GSE24132	Affymetrix	Homo sapiens	Peripheral and cord blood	12
	GSE35940	Agilent	Mus musculus	Lung tissue	43

[1] Samples that are not relevant have been excluded.

Different classification methods for microarray gene expression data

Different classification methods from the area of statistics and machine learning can be applied to microarray gene expression data of a disease, but some issues make the task non-trivial. Gene expression data is very different from what is expected by these methods. First, it has very high dimensionality, usually contains tens of thousands of genes. Second, the number of participating individuals in publicly available disease datasets is very small, often below 100. Third, most genes are irrelevant to disease case-control classification. Many researchers propose to perform gene selection before classification, which reduces the dimensionality as well as the number of irrelevant genes. For disease datasets, one common practice is to consider the differentially expressed genes between cases and controls as predictors of the classifier.

There is no single classification method that is superior over the rest in terms of classifying disease gene expression data [225]. For such data, we want a binary classification method that gives maximal classification accuracy in distinguishing disease cases from controls. First, we define the classification problem formally.

Problem A.0.1. *Given a training set $T = \{(t_1, case), (t_2, control), \ldots, (t_n, case)\}$, where n is the number of individuals in the training set, t_is are independent m-dimensional random data tuples of gene expression values, m is the total number of predictor genes, $t_i = (t_iX_1, t_iX_2, , t_iX_m)$, $m >> n$ and "case" and "control" are the class labels. Given a test set $S = s_1, s_2, \ldots, s_l$. Each s_i is a gene expression data tuple of length m, and l is the number of individuals in the test set. Each s_i is in the form of $s_i = (s_iX_1, s_iX_2, \ldots, s_iX_m)$, where X_j is the expression value of gene j. Find a classification function C, that gives maximal classification accuracy on S.*

For completeness, we consider four classification methods namely, Fisher's Linear Discriminant Analysis (FLDA), K-Nearest Neighbor (KNN), NaÃŕve Bayes Method (NB), and Support Vector Machine (SVM). For our purpose, we found that SVM performed better than other classifiers in most of the cases in terms of accuracy.

Fisher's Linear Discriminant Analysis (FLDA)

FLDA tries to find a linear combination of the predictors that maximizes the separation between the centers of the data points from different classes while at the same time minimizing the variation within each class. This approach is often is preferred in practice due to its dimension-reduction property.

More formally, given a training set T and test set S as described above, FLDA tries to find the linear combination Ma of the columns of matrix M that maximizes the Rayleigh quotient given by $a^T Ba/a^T Wa$, where B is the between-class sum of squares, W is the within-class sum of squares, and a is the transformation matrix.

Let, μ_k be the vector of average gene expression values of m predictor genes for the training tuples in class k, where $k \in \{case, control\}$. The correlation between any test sample s_i and each class is measured using the squared Euclidean distance of s_i and μ_k, denoted by $d_k(s_i)$, where,

$$d_k(s_i) = \sum_{p=1}^{h} ((s_i - \mu_k)v_p)^2$$

Here, v_ps are linearly independent eigenvectors of the matrix $W^{-1}B$ and h is the number of its non-zero eigenvalues. Class k is assigned to s_i if the distance between s_i and μ_k is minimum. Thus, for training set T, and a test sample s_i, FLDA classifies s_i using the following classification function:

$$C(T, s_i) = \arg\min_k d_k(s_i)$$

FLDA was first proposed and implemented by R.A. Fisher in 1936 [226]. Since then, it has been implemented numerous times for classifying gene expression data. We use the R implementation given by the *lda* function of the MASS package for our purpose.

K-Nearest Neighbor (KNN)

KNN is a distance metric based classifier. The main idea of this method is for each test sample s_i, find k training tuples from the training set, T with most similar expression value according to a distance measure. The class label of s_i is assigned using majority vote from the selected k training tuples while breaking ties at random. The commonly used measure of similarities includes Pearson correlation, Euclidean distance, etc. Thus, the classification function is given by,

KNN was first proposed and implemented by Fix and Hodges [227] and was applied to gene expression data by Dudoit et al. for tumor classification [228]. We use the R implementation of *knn* function provided in the class package for our purpose. This implementation uses Euclidean distance as a measure of similarity. The value of k was chosen by iterating over values from 1 to 5 and picking the k, which gives the most accuracy in classification. In most cases, the chosen value was either 2 or 3.

Naïve Bayes Method (NB)

NB method uses probabilistic induction to assign class labels to test samples, assuming independence among the predictor genes. The method models each class as a set of Gaussian distributions: one for each gene, by looking at the gene expression values of the training samples. Let, Gk denote the class variable representing the setoff Gaussian distributions where $k \in \{case, control\}$. Then G_k is given by $G_k = \{G_k^1, G_k^2, , G_k^m\}$ where, G_k^i is the Gaussian distribution of class k for gene i.

For training set T and any test sample s_i of test set S, the class label of s_i is obtained by the classification function:

$$C(T, s_i) = \arg \max_k^m \Big(\sum_{g=1}^{m} \log P(s_i^g | G_k^g) \Big)$$

$P(s_i^g | G_k^g)$ is given by Bayes rule and can be approximated from the mean and standard deviation of the Gaussian distribution for gene g from class kâĂŹs distribution set, G_k^g. NB method was first used to classify gene expression data in 2000 [229, 230].

We use the R implementation of NB method given by *naiveBayes* function of the e1071 package for our purpose.

Support Vector Machine (SVM)

SVM is a max-margin classifier that tries to find a hyperplane with maximum margin to separate the training tuples into different groups according to their classes. The margin of the hyperplane is defined as the distance from the hyperplane to the sets of points that are closest to it. The points that lie closest to the max-margin hyperplane are called support vectors. Since gene expression data can be viewed as very sparse points in a very high dimensional space, it is easy to find several hyperplanes that can separate the training tuples. However, this method is often prone to overfitting.

Formally, let, T be the training set, as defined before, with n training samples of the form $x_i = (t_i, c_i)$ where, each t_i is a expression vector of the form $(t_iX_1, t_iX_2, \dots, t_iX_m)$ and $ci \in \{1, -1\}$ is the class label with "1" representing cases and "-1" representing controls. Let, the max-margin hyperplane be denoted by the vector w and scalar b. Given a test sample s_i, SVM assigns class label to s_i based on the distance of s_i from the hyperplane in feature space. Thus the classification function is given by,

$$C(T, s_i) = \begin{cases} 1, & \text{if } sign(\langle \mathbf{w}, \phi(s_i) \rangle - b) > 0. \\ -1, & \text{otherwise.} \end{cases}$$

Here $\phi(s_i)$ denotes the mapping of test sample s_i into the feature space and denotes the dot product of two vectors x and y. SVM determines the max-margin hyperplane by applying various dot product functions as kernels depending on the separability of the training data points.

SVM was first introduced by Boser et al. and used in many data mining applications [231, 232]. We use the R implementation of SVM given by the *ksvm* function of the "kernlab" package for our purpose. We have applied both *vanilladot* (linear) and *rbfdot* (Gaussian) kernels and got similar levels of accuracy. Aggregated classifiers are often useful for improving the accuracy of classification. However, since the main focus of our study is not getting a higher accuracy in classifying cases and controls of a disease dataset but to select the appropriate multiple hypothesis correction tests which can give more informative genes, we limited our discussion to the basic classifiers.

Appendix B

Supplementary information for robust comparative functional metagenomics across diverse study populations

Performance of off-the-shelf HUMAnN2 on the gut microbiome of Baka individuals

We ran the full pipeline of out-of-the-box HUMAnN2 with ChocoPhlAn, UniRef, and MetaCyc databases on the microbiomes of all the non-industrialized Baka individuals (35 unpublished samples from Alm lab). Each sample has $\sim$ 7M paired-end reads of 150 bp length. It took HUMAnN2 $\sim$ 3 days 6 hours (4676.85 minutes) to annotate the samples using 16 threads on a server with Intel Xeon E5-2695 v2 x86_64 2.40 GHz processor and 320 GB RAM. On the same machine, it would take Carnelian a little over a day (1617 minutes) to bin the reads from this data set using 16 cpus on the same machine when we analyze the samples sequentially.

On average HUMAnN2 could annotate only 10% reads per Baka sample despite using the entire ChocoPhlAn and UniRef database. On average, HUMAnN2 detected less than 30 species and 996 Enzyme Commission (EC) terms per sample, and the average Shannon diversity index per sample was 5.58. For comparison, we also ran

Table B.1: Performance comparison of off-the-shelf HUMAnN2 with HUMAnN2 (translated) and Carnelian on the microbiomes of Baka individuals.

	Out-of-the-box HUMAnN2 (Default database)		**HUMAnN2 (translated) (EC database)**		**Carnelian (EC database)**	
	Baka	**Bostonian**	**Baka**	**Bostonian**	**Baka**	**Bostonian**
# reads/sample	722,354	24,527,448	21,383	83,131	269,720	1,430,026
# Species/sample	29.6	53.6	N/A	N/A	N/A	N/A
# ECs/sample	996	1061	827	791	2003	1981
Shannon Index	5.58	5.95	5.79	4.76	6.5	6.49

out-of-the-box HUMAnN2 on the microbiomes of 20 industrialized Bostonian individuals (unpublished data set from Alm lab). Each sample had roughly 36M reads on average. Since industrialized microbiomes are well characterized, here, HUMAnN2 can annotate more reads (40-50% per sample). On average, it detects 1061 ECs per sample (Shannon diversity index 5.95). UniRef IDs were mapped to level 4 EC numbers using the mapping provided by HUMAnN2. Table B.1 summarizes the results of full HUMAnN2 pipeline on Baka individuals. For comparison, we include the results from the Boston data set as well as the results of HUMAnN2 (translated) and Carnelian; both were run with our curated EC database.

Note that, full HUMAnN2 pipeline finds less diversity in the microbiomes of non-industrialized Baka individuals compared to the industrialized Bostonian individuals which is counter-intuitive [233]. Using our curated Enzyme Commission database, both HUMAnN2 (translated) and Carnelian can detect more enzymatic diversity in the Baka population.

Vowpal Wabbit implementation of Carnelian's one-against-all classifier model

Carnelian's gapped k-mer ensemble classifier was implemented using Vowpal Wabbit (v8.1.1). Vowpal Wabbit implementations perform better than conventional classifiers for large-scale sequence classification tasks [87, 207]. Some of its advantages are as follows: (i) it provides a dedicated of stochastic gradient descent (SGD) which makes the task of learning faster and more scalable compared to standard gradient descent; (ii) the learning can be done in an online fashion which makes retraining the model easier as new annotations become available; and (iii) the keys of the feature hash table can be stored as an integer using MurmurHash3 which saves space.

In "default" mode in Carnelian uses Vowpal Wabbit's one-against-all SVM classifiers. If users want probability scores for the predicted labels, they can use the "precise" mode of Carnelian in which we use one-against all logistic regression models from Vowpal Wabbit. The default parameters used to run Vowpal Wabbit are as follows:

- oaa: to select the one-against-all classifiers
- passes (Number of Training Passes): 1
- cache: Use a cache
- save_resume: save extra state so learning can be resumed later with new data
- bit precision: 31
- regularization parameters: l1=0, l2=0 for faster training
- To enable "precise" mode:
  - Loss function: logistic
  - probabilities: to get the probabilities for the predictions

Note that, Carnelian gives the user a choice to play with the Vowpal Wabbit training parameters, such as the number of passes, regularization parameters l1 and l2, bit precision, etc. The values that worked best for our analyses are included as defaults in the pipeline. A comparison of Carnelian's performance in "default" versus "precise" mode is available at `http://carnelian.csail.mit.edu`.

Commands used to run mi-faser, HUMAnN2, and Kraken2

mi-faser:

# Database construction:

```
DIAMOND makedb --in <gold_standard_protein_fasta> -d <database>
```

# Running annotation:

# Single-ended reads:

```
python3.6 mifaser.py -f <input_fasta> -d <database_path>
-o <path_to_output> -t 1 -c 1
```

# Paired-end reads:

```
python3.6 mifaser.py -l <forward_fq> <reverse_fq> -d <database_path>
-o <path_to_output> -t 1 -c 1
```

HUMAnN2 (translated search):

# Database construction and configuration:

```
DIAMOND makedb --in <gold_standard_protein_fasta> -d <database>
humann2_config --update database_folders protein <database>
```

# Running translated search:

```
humann2 --input <input_file> --output <out_dir>
--id-mapping <ec_mapping_file> --protein-database <database>
--bypass-nucleotide-search
```

Paired-end reads were put in a single file before running HUMAnN2 (translated) on it as instructed on their website.

Kraken2-translated:

#Index construction:

```
./kraken2-build --download-taxonomy --db $DBNAME --skip-maps
./kraken2-build --add-to-library $FASTAFILE --db $DBNAME --protein
./kraken2-build --build --db $DBNAME --protein
```

#Running translation and annotation:

```
./kraken2 --db $DBNAME --threads <num_threads> --output <out_file>
--use-names <input_file>
```

Bibliography

[1] Berger, B., Peng, J., Singh, M.: Computational solutions for omics data. Nature Reviews Genetics **14**(5), 333 (2013)

[2] Karczewski, K.J., Snyder, M.P.: Integrative omics for health and disease. Nature Reviews Genetics **19**(5), 299 (2018)

[3] Nazeen, S., Palmer, N.P., Berger, B., Kohane, I.S.: Integrative analysis of genetic data sets reveals a shared innate immune component in autism spectrum disorder and its co-morbidities. Genome Biology **17**(1), 228 (2016)

[4] Nazeen, S., Berger, B.: Carnelian: alignment-free functional binning and abundance estimation of metagenomic reads. 17th European Conference on Computational Biology (ECCB 2018), 8-12 September, 2018, Athens, Greece. Presented under Applications track. bioRxiv: `https://doi.org/10.1101/375121`. `https://doi.org/10.1101/375121`

[5] Nazeen, S., Yu, Y.W., Berger, B.: Carnelian uncovers hidden functional patterns across diverse study populations from whole metagenome sequencing reads. Genome Biology **under review** (2019)

[6] Adams, J.U.: Transcriptome: connecting the genome to gene function. Nature Education **1**(1), 195 (2008)

[7] Hasin, Y., Seldin, M., Lusis, A.: Multi-omics approaches to disease. Genome Biology **18**(1), 83 (2017)

[8] Wang, W.Y., Barratt, B.J., Clayton, D.G., Todd, J.A.: Genome-wide association studies: theoretical and practical concerns. Nature Reviews Genetics **6**(2), 109 (2005)

[9] Ragoussis, J.: Genotyping technologies for genetic research. Annual Review of Genomics and Human Genetics **10**, 117–133 (2009)

[10] Koboldt, D.C., Steinberg, K.M., Larson, D.E., Wilson, R.K., Mardis, E.R.: The next-generation sequencing revolution and its impact on genomics. Cell **155**(1), 27–38 (2013)

[11] Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., *et al.*: Targeted capture and massively parallel sequencing of 12 human exomes. Nature **461**(7261), 272 (2009)

[12] Aebersold, R., Mann, M.: Mass spectrometry-based proteomics. Nature **422**(6928), 198 (2003)

[13] Mann, M.: Origins of mass spectrometry-based proteomics. Nature Reviews Molecular Cell Biology **17**(11), 678 (2016)

[14] Schulze, A., Downward, J.: Navigating gene expression using microarrays—a technology review. Nature Cell Biology **3**(8), 190 (2001)

[15] Ozsolak, F., Milos, P.M.: RNA sequencing: advances, challenges and opportunities. Nature Reviews Genetics **12**(2), 87 (2011)

[16] Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., *et al.*: The NIH roadmap epigenomics mapping consortium. Nature Biotechnology **28**(10), 1045 (2010)

[17] Dettmer, K., Aronov, P.A., Hammock, B.D.: Mass spectrometry-based metabolomics. Mass Spectrometry Reviews **26**(1), 51–78 (2007)

[18] Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., Glöckner, F.O.: Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Research **41**(1), 1–1 (2013)

[19] Ranjan, R., Rani, A., Metwally, A., McGee, H.S., Perkins, D.L.: Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochemical and Biophysical Research Communications **469**(4), 967–977 (2016)

[20] Nazeen, S.: Integrative analysis of heterogeneous genomic datasets to discover genetic etiology of autism spectrum disorders. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, USA (2014)

[21] Cho, H., Berger, B., Peng, J.: Compact integration of multi-network topology for functional analysis of genes. Cell Systems **3**(6), 540–548 (2016)

[22] Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., Segata, N.: Shotgun metagenomics, from sampling to analysis. Nature Biotechnology **35**(9), 833 (2017)

[23] Nayfach, S., Pollard, K.S.: Toward accurate and quantitative comparative metagenomics. Cell **166**(5), 1103–1116 (2016)

[24] Joice, R., Yasuda, K., Shafquat, A., Morgan, X.C., Huttenhower, C.: Determining microbial products and identifying molecular targets in the human microbiome. Cell Metabolism **20**(5), 731–741 (2014)

[25] Lloyd-Price, J., Abu-Ali, G., Huttenhower, C.: The healthy human microbiome. Genome Medicine **8**(1), 1–11 (2016)

[26] Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., *et al.*: Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell **176**(3), 649–662 (2019)

[27] Malmstrom, R.R., Rodrigue, S., Huang, K.H., Kelly, L., Kern, S.E., Thompson, A., Roggensack, S., Berube, P.M., Henn, M.R., Chisholm, S.W.: Ecology of uncultured Prochlorococcus clades revealed through single-cell genomics and biogeographic analysis. The ISME Journal **7**(1), 184 (2013)

[28] Hutchison, C.A., Chuang, R.-Y., Noskov, V.N., Assad-Garcia, N., Deerinck, T.J., Ellisman, M.H., Gill, J., Kannan, K., Karas, B.J., Ma, L., *et al.*: Design and synthesis of a minimal bacterial genome. Science **351**(6280), 6253 (2016)

[29] Sberro, H., Fremin, B.J., Zlitni, S., Edfors, F., Greenfield, N., Snyder, M.P., Pavlopoulos, G.A., Kyrpides, N.C., Bhatt, A.S.: Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes. Cell **178**, 1–15 (2019)

[30] Loscalzo, J., Kohane, I., Barabási, A.-L.: Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. Molecular Systems Biology **3**(1), 124 (2007)

[31] Barabási, Albert-László; Gulbahce, Natali and Loscalzo, Joseph: Network medicine: a network-based approach to human disease. Nature Reviews Genetics **12**(1), 56–68 (2011)

[32] Cuthbert, B.N., Insel, T.R.: Toward the future of psychiatric diagnosis: the seven pillars of RDoC. BMC Medicine **11**(1), 126 (2013)

[33] Insel, T.R.: Mental disorders in childhood: shifting the focus from behavioral symptoms to neurodevelopmental trajectories. JAMA **311**(17), 1727–1728 (2014)

[34] Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.-h., Narzisi, G., Leotta, A., *et al.*: De novo gene disruptions in children on the autistic spectrum. Neuron **74**(2), 285–299 (2012)

[35] Neale, B.M., Kou, Y., Liu, L., MaÂŠAyan, A., Samocha, K.E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V., *et al.*: Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature **485**(7397), 242–245 (2012)

[36] Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., *et al.*: De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature **485**(7397), 237–241 (2012)

[37] O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., *et al.*: Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature **485**(7397), 246–250 (2012)

[38] Malhotra, D., Sebat, J.: CNVs: harbingers of a rare variant revolution in psychiatric genetics. Cell **148**(6), 1223–1241 (2012)

[39] Yu, T.W., Chahrour, M.H., Coulter, M.E., Jiralerspong, S., Okamura-Ikeda, K., Ataman, B., Schmitz-Abe, K., Harmin, D.A., Adli, M., Malik, A.N., *et al.*: Using whole-exome sequencing to identify inherited causes of autism. Neuron **77**(2), 259–273 (2013)

[40] Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., Thiruvahindrapuram, B., Xu, X., Ziman, R., Wang, Z., *et al.*: Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. The American Journal of Human Genetics **94**(5), 677–694 (2014)

[41] Lin, G.N., Corominas, R., Lemmens, I., Yang, X., Tavernier, J., Hill, D.E., Vidal, M., Sebat, J., Iakoucheva, L.M.: Spatiotemporal 16p11. 2 protein network implicates cortical late mid-fetal brain development and KCTD13-Cul3-RhoA pathway in psychiatric diseases. Neuron **85**(4), 742–754 (2015)

[42] Smoller, J., Craddock, N., Kendler, K., Lee, P., Neale, B., Nurnberger, J., Ripke, S., Santangelo, S., Sullivan, P.: Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. The Lancet **381**(9875), 1371–1379 (2013)

[43] Zhernakova, A., van Diemen, C.C., Wijmenga, C.: Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nature Reviews Genetics **10**(1), 43–55 (2009)

[44] Kohane, I.S., McMurry, A., Weber, G., MacFadden, D., Rappaport, L., Kunkel, L., Bickel, J., Wattanasin, N., Spence, S., Murphy, S., *et al.*: The co-morbidity burden of children and young adults with autism spectrum disorders. PLoS One **7**(4), 33224–33224 (2012)

[45] Doshi-Velez, F., Ge, Y., Kohane, I.S.: Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. Pediatrics **133**(1), 54–63 (2014)

[46] Prados-Torres, A., Calderón-Larrañaga, A., Hancco-Saavedra, J., Poblador-Plou, B., van den Akker, M.: Multimorbidity patterns: a systematic review. Journal of Clinical Epidemiology **67**(3), 254–266 (2014)

[47] Poblador-Plou, B., Calderón-Larrañaga, A., Marta-Moreno, J., Hancco-Saavedra, J., Sicras-Mainar, A., Soljak, M., Prados-Torres, A.: Comorbidity of dementia: a cross-sectional study of primary care older patients. BMC Psychiatry **14**(1), 84 (2014)

[48] Garin, N., Koyanagi, A., Chatterji, S., Tyrovolas, S., Olaya, B., Leonardi, M., Lara, E., Koskinen, S., Tobiasz-Adamczyk, B., Ayuso-Mateos, J.L., *et al.*: Global multimorbidity patterns: a cross-sectional, population-based, multi-country study. Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences **71**(2), 205–214 (2015)

[49] Mouridsen, S.E., Rich, B., Isager, T.: Epilepsy in disintegrative psychosis and infantile autism: a long-term validation study. Developmental Medicine & Child Neurology **41**(02), 110–114 (1999)

[50] Tuchman, R., Rapin, I.: Epilepsy in autism. The Lancet Neurology **1**(6), 352–358 (2002)

[51] Horvath, K., Papadimitriou, J.C., Rabsztyn, A., Drachenberg, C., Tildon, J.T.: Gastrointestinal abnormalities in children with autistic disorder. The Journal of Pediatrics **135**(5), 559–563 (1999)

[52] Horvath, K., Perman, J.A.: Autistic disorder and gastrointestinal disease. Current Opinion in Pediatrics **14**(5), 583–587 (2002)

[53] Richdale, A.L., Schreck, K.A.: Sleep problems in autism spectrum disorders: prevalence, nature, & possible biopsychosocial aetiologies. Sleep Medicine Reviews **13**(6), 403–411 (2009)

[54] Wu, J.Y., Kuban, K.C., Allred, E., Shapiro, F., Darras, B.T.: Association of Duchenne muscular dystrophy with autism spectrum disorder. Journal of Child Neurology **20**(10), 790–795 (2005)

[55] Hendriksen, J., Vles, J.: Neuropsychiatric disorders in males with duchenne muscular dystrophy: frequency rate of attention-deficit hyperactivity disorder (ADHD), autism spectrum disorder, and obsessive–compulsive disorder. Journal of Child Neurology **23**(5), 477–481 (2008)

[56] Hinton, V.J., Cyrulnik, S.E., Fee, R.J., Batchelder, A., Kiefel, J.M., Goldstein, E.M., Kaufmann, P., Darryl, C.: Association of autistic spectrum disorders with dystrophinopathies. Pediatric Neurology **41**(5), 339–346 (2009)

[57] Morgan, C.N., Roy, M., Chance, P.: Psychiatric comorbidity and medication use in autism: A community survey. The Psychiatrist **27**(10), 378–381 (2003)

[58] Meyer, U., Feldon, J., Dammann, O.: Schizophrenia and autism: both shared and disorder-specific pathogenesis via perinatal inflammation? Pediatric Research **69**, 26–33 (2011)

[59] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology **215**(3), 403–410 (1990)

[60] Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., *et al.*: The RAST Server: rapid annotations using subsystems technology. BMC Genomics **9**(1), 75 (2008)

[61] Huson, D.H., Mitra, S., Ruscheweyh, H.-J., Weber, N., Schuster, S.C.: Integrative analysis of environmental sequences using MEGAN4. Genome Research **21**(9), 1552–1560 (2011)

[62] Karlsson, F.H., Nookaew, I., Nielsen, J.: Metagenomic data utilization and analysis (MEDUSA) and construction of a global gut microbial gene catalogue. PLoS Computational Biology **10**(7), 1003706 (2014)

[63] Boulund, F., Sjögren, A., Kristiansson, E.: Tentacle: distributed quantification of genes in metagenomes. GigaScience **4**(1), 40 (2015)

[64] Kultima, J.R., Coelho, L.P., Forslund, K., Huerta-Cepas, J., Li, S.S., Driessen, M., Voigt, A.Y., Zeller, G., Sunagawa, S., Bork, P.: MOCAT2: a metagenomic assembly, annotation and profiling framework. Bioinformatics **32**(16), 2520–2523 (2016)

[65] Markowitz, V.M., Chen, I.-M.A., Palaniappan, K., Chu, K., Szeto, E., Pillay, M., Ratner, A., Huang, J., Woyke, T., Huntemann, M., *et al.*: IMG 4 version of the integrated microbial genomes comparative analysis system. Nucleic Acids Research **42**(D1), 560–567 (2013)

[66] Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., *et al.*: A human gut microbial gene catalogue established by metagenomic sequencing. Nature **464**(7285), 59 (2010)

[67] Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J.R., Prifti, E., Nielsen, T., *et al.*: An integrated catalog of reference genes in the human gut microbiome. Nature Biotechnology **32**(8), 834 (2014)

[68] Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., *et al.*: The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics **9**(1), 386 (2008)

[69] Abubucker, S., Segata, N., Goll, J., Schubert, A.M., Izard, J., Cantarel, B.L., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M., Henrissat, B., *et al.*: Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Computational Biology **8**(6), 1002358 (2012)

[70] Nayfach, S., Bradley, P.H., Wyman, S.K., Laurent, T.J., Williams, A., Eisen, J.A., Pollard, K.S., Sharpton, T.J.: Automated and accurate estimation of gene family abundance from shotgun metagenomes. PLoS Computational Biology **11**(11), 1004573 (2015)

[71] Sharifi, F., Ye, Y.: From gene annotation to function prediction for metagenomics. Protein Function Prediction: Methods and Protocols, 27–34 (2017)

[72] Zhu, C., Miller, M., Marpaka, S., Vaysberg, P., Rühlemann, M.C., Wu, G., Heinsen, F.-A., Tempel, M., Zhao, L., Lieb, W., Franke, A., Bromberg, Y.: Functional sequencing read annotation for high precision microbiome analysis. Nucleic Acids Research **46**(4), 23 (2018)

[73] 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., *et al.*: Species-level functional profiling of metagenomes and metatranscriptomes. Nature Methods **15**(11), 962 (2018)

[74] Ounit, R., Wanamaker, S., Close, T.J., Lonardi, S.: CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics **16**(1), 236 (2015)

[75] Yu, Y.W., Daniels, N.M., Danko, D.C., Berger, B.: Entropy-scaling search of massive biological data. Cell Systems **1**(2), 130–140 (2015)

[76] Buchfink, B., Xie, C., Huson, D.H.: Fast and sensitive protein alignment using DIAMOND. Nature Methods **12**(1), 59 (2015)

[77] Sasson, O., Kaplan, N., Linial, M.: Functional annotation prediction: all for one and one for all. Protein Science **15**(6), 1557–1562 (2006)

[78] Lindgreen, S., Adair, K.L., Gardner, P.P.: An evaluation of the accuracy and speed of metagenome analysis tools. Scientific Reports **6**, 19233 (2016)

[79] Finucane, M.M., Sharpton, T.J., Laurent, T.J., Pollard, K.S.: A taxonomic signature of obesity in the microbiome? Getting to the guts of the matter. PloS One **9**(1), 84689 (2014)

[80] Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., *et al.*: Human gut microbiome viewed across age and geography. Nature **486**(7402), 222 (2012)

[81] Lozupone, C.A., Stombaugh, J.I., Gordon, J.I., Jansson, J.K., Knight, R.: Diversity, stability and resilience of the human gut microbiota. Nature **489**(7415), 220 (2012)

[82] Sze, M.A., Schloss, P.D.: Looking for a signal in the noise: revisiting obesity and the microbiome. mBio **7**(4), 01018–16 (2016)

[83] Duvallet, C., Gibbons, S.M., Gurry, T., Irizarry, R.A., Alm, E.J.: Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nature Communications **8**(1), 1784 (2017)

[84] Gibbons, S.M., Duvallet, C., Alm, E.J.: Correcting for batch effects in case-control microbiome studies. PLoS Computational Biology **14**(4), 1006102 (2018)

[85] Gallager, R.: Low-density parity-check codes. IRE Transactions on Information Theory **8**(1), 21–28 (1962)

[86] Luo, Y., Yu, Y.W., Zeng, J., Berger, B., Peng, J.: Metagenomic binning through low-density hashing. Bioinformatics **35**(2), 219–226 (2018)

[87] Vervier, K., Mahé, P., Tournoud, M., Veyrieras, J.-B., Vert, J.-P.: Large-scale machine learning for metagenomics sequence classification. Bioinformatics **32**(7), 1023–1032 (2015)

[88] Sirota, M., Dudley, J.T., Kim, J., Chiang, A.P., Morgan, A.A., Sweet-Cordero, A., Sage, J., Butte, A.J.: Discovery and preclinical validation of drug indications using compendia of public gene expression data. Science Translational Medicine **3**(96), 96–779677 (2011)

[89] Levy, D., Ronemus, M., Yamrom, B., Lee, Y.-h., Leotta, A., Kendall, J., Marks, S., Lakshmi, B., Pai, D., Ye, K., *et al.*: Rare de novo and transmitted copy-number variation in autistic spectrum disorders. Neuron **70**(5), 886–897 (2011)

[90] Marshall, C.R., Noor, A., Vincent, J.B., Lionel, A.C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y., *et al.*: Structural variation of chromosomes in autism spectrum disorder. The American Journal of Human Genetics **82**(2), 477–488 (2008)

[91] Bijlsma, E., Gijsbers, A., Schuurs-Hoeijmakers, J., Van Haeringen, A., Van De Putte, D.F., Anderlid, B.-M., Lundin, J., Lapunzina, P., Jurado, L.P., Delle Chiaie, B., *et al.*: Extending the phenotype of recurrent rearrangements of 16p11. 2: deletions in mentally retarded patients without autism and in normal individuals. European Journal of Medical Genetics **52**(2), 77–87 (2009)

[92] McCarthy, S.E., Makarov, V., Kirov, G., Addington, A.M., McClellan, J., Yoon, S., Perkins, D.O., Dickel, D.E., Kusenda, M., Krastoshevsky, O., *et al.*: Microduplications of 16p11. 2 are associated with schizophrenia. Nature Genetics **41**(11), 1223–1227 (2009)

[93] Weiss, L.A., Shen, Y., Korn, J.M., Arking, D.E., Miller, D.T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M.A., Green, T., *et al.*: Association between microdeletion and microduplication at 16p11. 2 and autism. New England Journal of Medicine **358**(7), 667–675 (2008)

[94] Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., *et al.*: Functional impact of global rare copy number variation in autism spectrum disorders. Nature **466**(7304), 368–372 (2010)

[95] Béna, F., Bruno, D.L., Eriksson, M., van Ravenswaaij-Arts, C., Stark, Z., Dijkhuizen, T., Gerkes, E., Gimelli, S., Ganesamoorthy, D., Thuresson, A.C., *et al.*: Molecular and clinical characterization of 25 individuals with exonic deletions of NRXN1 and comprehensive review of the literature. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics **162**(4), 388–403 (2013)

[96] Moreno-De-Luca, D., Sanders, S., Willsey, A., Mulle, J., Lowe, J., Geschwind, D., State, M., Martin, C., Ledbetter, D., *et al.*: Using large clinical data sets to infer pathogenicity for rare copy number variants in autism cohorts. Molecular Psychiatry **18**(10), 1090–1095 (2013)

[97] Carter, M., Scherer, S.: Autism spectrum disorder in the genetics clinic: a review. Clinical Genetics **83**(5), 399–407 (2013)

[98] Robinson, W.H., Fontoura, P., Lee, B.J., de Vegvar, H.E.N., Tom, J., Pedotti, R., DiGennaro, C.D., Mitchell, D.J., Fong, D., Ho, P.P., *et al.*: Protein microarrays guide tolerizing DNA vaccine treatment of autoimmune encephalomyelitis. Nature Biotechnology **21**(9), 1033–1039 (2003)

[99] Lutterotti, A., Yousef, S., Sputtek, A., Stürner, K.H., Stellmann, J.-P., Breiden, P., Reinhardt, S., Schulze, C., Bester, M., Heesen, C., *et al.*: Antigen-specific tolerance by autologous myelin peptide–coupled cells: a phase 1 trial in multiple sclerosis. Science Translational Medicine **5**(188), 188–7518875 (2013)

[100] Becker, K.G.: Autism, asthma, inflammation, and the hygiene hypothesis. Medical Hypotheses **69**(4), 731–740 (2007)

[101] Atladóttir, H.O., Thorsen, P., Østergaard, L., Schendel, D.E., Lemcke, S., Abdallah, M., Parner, E.T.: Maternal infection requiring hospitalization during pregnancy and autism spectrum disorders. Journal of Autism and Developmental Disorders **40**(12), 1423–1430 (2010)

[102] Atladóttir, H.Ó., Henriksen, T.B., Schendel, D.E., Parner, E.T.: Autism after infection, febrile episodes, and antibiotic use during pregnancy: an exploratory study. Pediatrics **130**(6), 1447–1454 (2012)

[103] Garbett, K.A., Hsiao, E.Y., Kálmán, S., Patterson, P.H., Mirnics, K.: Effects of maternal immune activation on gene expression patterns in the fetal brain. Translational Psychiatry **2**(4), 98 (2012)

[104] Hagberg, H., Gressens, P., Mallard, C.: Inflammation during fetal and neonatal life: implications for neurologic and neuropsychiatric disease in children and adults. Annals of Neurology **71**(4), 444–457 (2012)

[105] Curatolo, P., Porfirio, M.C., Manzi, B., Seri, S.: Autism in tuberous sclerosis. European Journal of Paediatric Neurology **8**(6), 327–332 (2004)

[106] Loirat, C., Bellanné-Chantelot, C., Husson, I., Deschênes, G., Guigonis, V., Chabane, N.: Autism in three patients with cystic or hyperechogenic kidneys and chromosome 17q12 deletion. Nephrology Dialysis Transplantation **25**(10), 3430–3433 (2010)

[107] Surén, P., Bakken, I.J., Aase, H., Chin, R., Gunnes, N., Lie, K.K., Magnus, P., Reichborn-Kjennerud, T., Schjølberg, S., Øyen, A.-S., *et al.*: Autism spectrum disorder, ADHD, epilepsy, and cerebral palsy in Norwegian children. Pediatrics **130**(1), 152–158 (2012)

[108] Witchel, H.J., Hancox, J.C., Nutt, D.J.: Psychotropic drugs, cardiac arrhythmia, and sudden death. Journal of Clinical Psychopharmacology **23**(1), 58–77 (2003)

[109] Bilder, D., Botts, E.L., Smith, K.R., Pimentel, R., Farley, M., Viskochil, J., McMahon, W.M., Block, H., Ritvo, E., Ritvo, R.-A., *et al.*: Excess mortality and causes of death in autism spectrum disorders: a follow up of the 1980s Utah/UCLA autism epidemiologic study. Journal of Autism and Developmental Disorders **43**(5), 1196–1204 (2013)

[110] Konstantareas, M.M., Homatidis, S.: Brief report: ear infections in autistic and normal children. Journal of Autism and Developmental Disorders **17**(4), 585–594 (1987)

[111] Rosenhall, U., Nordin, V., Sandström, M., Ahlsen, G., Gillberg, C.: Autism and hearing loss. Journal of Autism and Developmental Disorders **29**(5), 349–357 (1999)

[112] Porges, S.W., Macellaio, M., Stanfill, S.D., McCue, K., Lewis, G.F., Harden, E.R., Handelman, M., Denver, J., Bazhenova, O.V., Heilman, K.J.: Respiratory sinus arrhythmia and auditory processing in autism: modifiable deficits of an integrated social engagement system? International Journal of Psychophysiology **88**(3), 261–270 (2013)

[113] Walker, S.J., Fortunato, J., Gonzalez, L.G., Krigsman, A.: Identification of unique gene expression profile in children with regressive autism spectrum disorder (ASD) and ileocolitis. PLoS One **8**(3), 58058 (2013)

[114] Shavelle, R.M., Strauss, D.J., Pickett, J.: Causes of death in autism. Journal of Autism and Developmental Disorders **31**(6), 569–576 (2001)

[115] Tabares-Seisdedos, R., Rubenstein, J.: Chromosome 8p as a potential hub for developmental neuropsychiatric disorders: implications for schizophrenia, autism and cancer. Molecular Psychiatry **14**(6), 563–589 (2009)

[116] Ingason, A., Rujescu, D., Cichon, S., Sigurdsson, E., Sigmundsson, T., Pietiläinen, O., Buizer-Voskamp, J., Strengman, E., Francks, C., Muglia, P., *et al.*: Copy number variations of chromosome 16p13. 1 region associated with schizophrenia. Molecular Psychiatry **16**(1), 17–25 (2011)

[117] Murdoch, J.D., State, M.W.: Recent developments in the genetics of autism spectrum disorders. Current Opinion in Genetics & Development **23**(3), 310–315 (2013)

[118] Smyth, G.K.: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology **3**(1), 1–25 (2004)

[119] Fisher, R.A.: Statistical Methods for Research Workers, 5th edn. Biological Monographs and Manuals. Oliver and Boyd Ltd., Edinburgh (1934)

[120] Mogensen, T.H.: Pathogen recognition and inflammatory signaling in innate immune defenses. Clinical Microbiology Reviews **22**(2), 240–273 (2009)

[121] Ashwood, P., Krakowiak, P., Hertz-Picciotto, I., Hansen, R., Pessah, I., Van de Water, J.: Elevated plasma cytokines in autism spectrum disorders provide evidence of immune dysfunction and are associated with impaired behavioral outcome. Brain, Behavior, and Immunity **25**(1), 40–45 (2011)

[122] Enstrom, A.M., Onore, C.E., Van de Water, J.A., Ashwood, P.: Differential monocyte responses to TLR ligands in children with autism spectrum disorders. Brain, Behavior, and Immunity **24**(1), 64–71 (2010)

[123] Verkhratsky, A., Rodríguez, J.J., Parpura, V.: Neuroglia in ageing and disease. Cell and Tissue Research **357**(2), 493–503 (2014)

[124] Won, H., Mah, W., Kim, E.: Autism spectrum disorder causes, mechanisms, and treatments: focus on neuronal synapses. Frontiers in Molecular Neuroscience **6**, 19 (2013)

[125] Zihni, C., Mills, C., Matter, K., Balda, M.S.: Tight junctions: from simple barriers to multifunctional molecular gates. Nature Reviews Molecular Cell Biology **17**(9), 564 (2016)

[126] Malkova, N.V., Collin, Z.Y., Hsiao, E.Y., Moore, M.J., Patterson, P.H.: Maternal immune activation yields offspring displaying mouse versions of the three core symptoms of autism. Brain, Behavior, and Immunity **26**(4), 607–616 (2012)

[127] Enstrom, A.M., Lit, L., Onore, C.E., Gregg, J.P., Hansen, R.L., Pessah, I.N., Hertz-Picciotto, I., Van de Water, J.A., Sharp, F.R., Ashwood, P.: Altered gene expression and function of peripheral blood natural killer cells in children with autism. Brain, Behavior, and Immunity **23**(1), 124–133 (2009)

[128] Filiano, A.J., Xu, Y., Tustison, N.J., Marsh, R.L., Baker, W., Smirnov, I., Overall, C.C., Gadani, S.P., Turner, S.D., Weng, Z., *et al.*: Unexpected role of interferon-γ in regulating neuronal connectivity and social behaviour. Nature **535**(7612), 425 (2016)

[129] Suh, H.-S., Kim, M.-O., Lee, S.C.: Inhibition of granulocyte-macrophage colony-stimulating factor signaling and microglial proliferation by anti-CD45RO: role of Hck tyrosine kinase and phosphatidylinositol 3-kinase/Akt. The Journal of Immunology **174**(5), 2712–2719 (2005)

[130] Fatemi, S.H.: Multiple pathways in prevention of immune-mediated brain disorders: Implications for the prevention of autism. Journal of Neuroimmunology **217**(1-2), 8 (2009)

[131] Parker-Athill, E., Luo, D., Bailey, A., Giunta, B., Tian, J., Shytle, R.D., Murphy, T., Legradi, G., Tan, J.: Flavonoids, a prenatal prophylaxis via targeting JAK2/STAT3 signaling to oppose IL-6/MIA associated autism. Journal of Neuroimmunology **217**(1), 20–27 (2009)

[132] Polan, M.B., Pastore, M.T., Steingass, K., Hashimoto, S., Thrush, D.L., Pyatt, R., Reshmi, S., Gastier-Foster, J.M., Astbury, C., McBride, K.L.: Neurodevelopmental disorders among individuals with duplication of 4p13 to 4p12 containing a GABAA receptor subunit gene cluster. European Journal of Human Genetics **22**(1), 105–109 (2014)

[133] Kanehisa, M., Goto, S.: KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research **28**(1), 27–30 (2000)

[134] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M.: KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Research **40**(D1), 109–114 (2011)

[135] Pramparo, T., Pierce, K., Lombardo, M.V., Barnes, C.C., Marinero, S., Ahrens-Barbeau, C., Murray, S.S., Lopez, L., Xu, R., Courchesne, E.: Prediction of autism by translation and immune/inflammation coexpressed genes in toddlers from pediatric community practices. JAMA Psychiatry **72**(4), 386–394 (2015)

[136] Davis, S., Meltzer, P.S.: GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics **23**(14), 1846–1847 (2007)

[137] Goodman, S.N.: Toward evidence-based medical statistics. 2: The Bayes factor. Annals of Internal Medicine **130**(12), 1005–1013 (1999)

[138] Sellke, T., Bayarri, M., Berger, J.O.: Calibration of ρ values for testing precise null hypotheses. The American Statistician **55**(1), 62–71 (2001)

[139] Johnson, V.E.: Bayes factors based on test statistics. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**(5), 689–701 (2005)

[140] Edgar, R., Domrachev, M., Lash, A.E.: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research **30**(1), 207–210 (2002)

[141] Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., *et al.*: NCBI GEO: archive for functional genomics data setsÂŮupdate. Nucleic Acids Research **41**(D1), 991–995 (2013)

[142] Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., *et al.*: The Reactome pathway knowledgebase. Nucleic Acids Research **42**(D1), 472–477 (2014)

[143] Monaco, M.K., Stein, J., Naithani, S., Wei, S., Dharmawardhana, P., Kumari, S., Amarasinghe, V., Youens-Clark, K., Thomason, J., Preece, J., *et al.*: Gramene 2013: comparative plant genomics resources. Nucleic Acids Research **42**(D1), 1193–1199 (2014)

[144] Nishimura, D.: BioCarta. Biotech Software & Internet Report: The Computer Software Journal for Scient **2**(3), 117–120 (2001)

[145] Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., Buetow, K.H.: PID: the pathway interaction database. Nucleic Acids Research **37**(suppl 1), 674–679 (2009)

[146] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., *et al.*: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America **102**(43), 15545–15550 (2005)

[147] Murie, C., Woody, O., Lee, A.Y., Nadon, R.: Comparison of small n statistical tests of differential expression applied to microarrays. BMC Bioinformatics **10**(1), 1 (2009)

[148] Tesse, R., Pandey, R., Kabesch, M.: Genetic variations in toll-like receptor pathway genes influence asthma and atopy. Allergy **66**(3), 307–316 (2011)

[149] Zuany-Amorim, C., Hastewell, J., Walker, C.: Toll-like receptors as potential therapeutic targets for multiple diseases. Nature Reviews Drug Discovery **1**(10), 797–807 (2002)

[150] Lin, J., Caye-Thomasen, P., Tono, T., Zhang, Q., Nakamura, Y., Feng, L., Huang, J., Ye, S., Hu, X., Kerschner, J.: Mucin production and mucous cell metaplasia in otitis media. International Journal of Otolaryngology **2012**, 745325–745325 (2011)

[151] Kimura, H., Yoshizumi, M., Ishii, H., Oishi, K., Ryo, A.: Cytokine production and signaling pathways in respiratory virus infection. Frontiers in Microbiology **4**(276), 2 (2013)

[152] Hennessy, E.J., Parker, A.E., O'Neill, L.A.: Targeting Toll-like receptors: emerging therapeutics? Nature Reviews Drug Discovery **9**(4), 293–307 (2010)

[153] Neurath, M.F.: Cytokines in inflammatory bowel disease. Nature Reviews Immunology **14**(5), 329–342 (2014)

[154] Gijsbers, K., Van Assche, G., Joossens, S., Struyf, S., Proost, P., Rutgeerts, P., Geboes, K., Van Damme, J.: CXCR1-binding chemokines in inflammatory bowel diseases: down-regulated IL-8/CXCL8 production by leukocytes in Crohn's disease and selective GCP-2/CXCL6 expression in inflamed intestinal tissue. European Journal of Immunology **34**(7), 1992–2000 (2004)

[155] Ramos, P.S., Sajuthi, S., Langefeld, C.D., Walker, S.J.: Immune function genes CD99L2, JARID2 and TPO show association with autism spectrum disorder. Molecular Autism **3**(1), 1–5 (2012)

[156] Saxena, V., Ramdas, S., Ochoa, C.R., Wallace, D., Bhide, P., Kohane, I.: Structural, genetic, and functional signatures of disordered neuro-immunological development in autism spectrum disorder. PLoS One **7**(12), 48835 (2012)

[157] Moscavitch, S.-D., Szyper-Kravitz, M., Shoenfeld, Y.: Autoimmune pathology accounts for common manifestations in a wide range of neuro-psychiatric disorders: the olfactory and immune system interrelationship. Clinical Immunology **130**(3), 235–243 (2009)

[158] Li, X., Chauhan, A., Sheikh, A.M., Patil, S., Chauhan, V., Li, X.-M., Ji, L., Brown, T., Malik, M.: Elevated immune response in the brain of autistic patients. Journal of Neuroimmunology **207**(1), 111–116 (2009)

[159] Smith, S.E., Li, J., Garbett, K., Mirnics, K., Patterson, P.H.: Maternal immune activation alters fetal brain development through interleukin-6. The Journal of Neuroscience **27**(40), 10695–10702 (2007)

[160] Kong, S., Shimizu-Motohashi, Y., Campbell, M., Lee, I., Collins, C., Brewster, S., Holm, I., Rappaport, L., Kohane, I., Kunkel, L.: Peripheral blood gene expression signature differentiates children with autism from unaffected siblings. Neurogenetics **14**(2), 143–152 (2013)

[161] Voineagu, I., Eapen, V.: Converging pathways in autism spectrum disorders: interplay between synaptic dysfunction and immune responses. Frontiers in Human Neuroscience **7**, 738 (2013)

[162] Estes, M.L., McAllister, A.K.: Immune mediators in the brain and peripheral tissues in autism spectrum disorder. Nature Reviews Neuroscience **16**(8), 469–486 (2015)

[163] Suzuki, K., Sugihara, G., Ouchi, Y., Nakamura, K., Futatsubashi, M., Takebayashi, K., Yoshihara, Y., Omata, K., Matsumoto, K., Tsuchiya, K.J., *et al.*: Microglial activation in young adults with autism spectrum disorder. JAMA Psychiatry **70**(1), 49–58 (2013)

[164] Gupta, S., Ellis, S.E., Ashar, F.N., Moes, A., Bader, J.S., Zhan, J., West, A.B., Arking, D.E.: Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. Nature Communications **5** (2014)

[165] Kim, H., Cho, M., Shim, W., Kim, J., Jeon, E., Kim, D., Yoon, S.: Deficient autophagy in microglia impairs synaptic pruning and causes social behavioral defects. Molecular Psychiatry **22**(11), 1576 (2017)

[166] Campbell, M.G., Kohane, I.S., Kong, S.W.: Pathway-based outlier method reveals heterogeneous genomic structure of autism in blood transcriptome. BMC Medical Genomics **6**(1), 34 (2013)

[167] Jyonouchi, H., Geng, L., Davidow, A.L.: Cytokine profiles by peripheral blood monocytes are associated with changes in behavioral symptoms following immune insults in a subset of ASD subjects: an inflammatory subtype? Journal of Neuroinflammation **11**(1), 187 (2014)

[168] West, P.R., Amaral, D.G., Bais, P., Smith, A.M., Egnash, L.A., Ross, M.E., Palmer, J.A., Fontaine, B.R., Conard, K.R., Corbett, B.A., *et al.*: Metabolomics as a tool for discovery of biomarkers of autism spectrum disorder in the blood plasma of children. PLoS One **9**(11), 112445 (2014)

[169] Atladóttir, H.Ó., Pedersen, M.G., Thorsen, P., Mortensen, P.B., Deleuran, B., Eaton, W.W., Parner, E.T.: Association of family history of autoimmune diseases and autism spectrum disorders. Pediatrics **124**(2), 687–694 (2009)

[170] Brown, A.S., Sourander, A., Hinkka-Yli-Salomäki, S., McKeague, I., Sundvall, J., Surcel, H.: Elevated maternal C-reactive protein and autism in a national birth cohort. Molecular Psychiatry **19**(2), 259–264 (2014)

[171] Yap, I.K., Angley, M., Veselkov, K.A., Holmes, E., Lindon, J.C., Nicholson, J.K.: Urinary metabolic phenotyping differentiates children with autism from their unaffected siblings and age-matched controls. Journal of Proteome Research **9**(6), 2996–3004 (2010)

[172] Kang, D.-W., Park, J.G., Ilhan, Z.E., Wallstrom, G., LaBaer, J., Adams, J.B., Krajmalnik-Brown, R.: Reduced incidence of Prevotella and other fermenters in intestinal microflora of autistic children. PLoS One **8**(7), 68322 (2013)

[173] Wang, L., Christophersen, C.T., Sorich, M.J., Gerber, J.P., Angley, M.T., Conlon, M.A., *et al.*: Increased abundance of Sutterella spp. and Ruminococcus torques in feces of children with autism spectrum disorder. Molecular Autism **4**(1), 42 (2013)

[174] De Angelis, M., Piccolo, M., Vannini, L., Siragusa, S., De Giacomo, A., Serrazzanetti, D.I., Cristofori, F., Guerzoni, M.E., Gobbetti, M., Francavilla, R.: Fecal microbiota and metabolome of children with autism and pervasive developmental disorder not otherwise specified. PLoS One **8**(10) (2013)

[175] Mayer, E.A., Padua, D., Tillisch, K.: Altered brain-gut axis in autism: Comorbidity or causative mechanisms? Bioessays **36**(10), 933–939 (2014)

[176] Gene Expression Omnibus. `https://www.ncbi.nlm.nih.gov/geo/`. Accessed: 2016-09-29

[177] Molecular Signatures Database. `http://www.broadinstitute.org/gsea/msigdb/collections.jsp`. Accessed: 2016-09-29

[178] Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., Gordon, J.I.: The human microbiome project. Nature **449**(7164), 804 (2007)

[179] Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., *et al.*: Critical assessment of metagenome interpretation—a benchmark of metagenomics software. Nature Methods **14**(11), 1063 (2017)

[180] Erickson, A.R., Cantarel, B.L., Lamendella, R., Darzi, Y., Mongodin, E.F., Pan, C., Shah, M., Halfvarson, J., Tysk, C., Henrissat, B., *et al.*: Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. PloS One **7**(11), 49138 (2012)

[181] Turnbaugh, P.J., Gordon, J.I.: The core gut microbiome, energy balance and obesity. The Journal of Physiology **587**(17), 4153–4158 (2009)

[182] Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., *et al.*: A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature **490**(7418), 55 (2012)

[183] Karlsson, F.H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C.J., Fagerberg, B., Nielsen, J., Bäckhed, F.: Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature **498**(7452), 99 (2013)

[184] Wang, T., Cai, G., Qiu, Y., Fei, N., Zhang, M., Pang, X., Jia, W., Cai, S., Zhao, L.: Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. The ISME Journal **6**(2), 320 (2012)

[185] Driscoll, H.E., Vincent, J.J., English, E.L., Dolci, E.D.: Metagenomic investigation of the microbial diversity in a chrysotile asbestos mine pit pond, lowell, vermont, usa. Genomics Data **10**, 158–164 (2016)

[186] Strati, F., Cavalieri, D., Albanese, D., De Felice, C., Donati, C., Hayek, J., Jousson, O., Leoncini, S., Renzi, D., Calabrò, A., *et al.*: New evidences on the altered gut microbiota in autism spectrum disorders. Microbiome **5**(1), 24 (2017)

[187] Hawksworth, G., Drasar, B., Hili, M.: Intestinal bacteria and the hydrolysis of glycosidic bonds. Journal of Medical Microbiology **4**(4), 451–459 (1971)

[188] Wood, D.E., Salzberg, S.L.: Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology **15**(3), 46 (2014)

[189] Schaeffer, L., Pimentel, H., Bray, N., Melsted, P., Pachter, L.: Pseudoalignment for metagenomic read assignment. Bioinformatics **33**(14), 2082–2088 (2017)

[190] Rath, S., Rud, T., Karch, A., Pieper, D.H., Vital, M.: Pathogenic functions of host microbiota. Microbiome **6**(1), 174 (2018)

[191] Rho, M., Tang, H., Ye, Y.: FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Research **38**(20), 191–191 (2010)

[192] Eddy, S.R.: Accelerated profile HMM searches. PLoS Computational Biology **7**(10), 1002195 (2011)

[193] Wilson, M.R., Naccache, S.N., Samayoa, E., Biagtan, M., Bashir, H., Yu, G., Salamat, S.M., Somasekar, S., Federman, S., Miller, S., *et al.*: Actionable diagnosis of neuroleptospirosis by next-generation sequencing. New England Journal of Medicine **370**(25), 2408–2417 (2014)

[194] Wang, B., Yao, M., Lv, L., Ling, Z., Li, L.: The human microbiota in health and disease. Engineering **3**(1), 71–82 (2017)

[195] Walters, W.A., Xu, Z., Knight, R.: Meta-analyses of human gut microbes associated with obesity and IBD. FEBS Letters **588**(22), 4223–4233 (2014)

[196] The IBDMDB team: The Inflammatory Bowel Disease Multi'omics Database. `https://ibdmdb.org/`. last accessed: March 20, 2019 (2018)

[197] The Human Microbiome Project Consortium: Structure, function and diversity of the healthy human microbiome. Nature **486**(7402), 207 (2012)

[198] Rampelli, S., Schnorr, S.L., Consolandi, C., Turroni, S., Severgnini, M., Peano, C., Brigidi, P., Crittenden, A.N., Henry, A.G., Candela, M.: Metagenome sequencing of the Hadza hunter-gatherer gut microbiota. Current Biology **25**(13), 1682–1693 (2015)

[199] Fasano, A., Visanji, N.P., Liu, L.W., Lang, A.E., Pfeiffer, R.F.: Gastrointestinal dysfunction in Parkinson's disease. The Lancet Neurology **14**(6), 625–639 (2015)

[200] Bedarf, J.R., Hildebrand, F., Coelho, L.P., Sunagawa, S., Bahram, M., Goeser, F., Bork, P., Wüllner, U.: Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naïve Parkinson's disease patients. Genome Medicine **9**(1), 39 (2017)

[201] Luo, C., Tsementzi, D., Kyrpides, N., Read, T., Konstantinidis, K.T.: Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PloS One **7**(2), 30087 (2012)

[202] Rodriguez-r, L.M., Overholt, W.A., Hagan, C., Huettel, M., Kostka, J.E., Konstantinidis, K.T.: Microbial community successional patterns in beach sands impacted by the Deepwater Horizon oil spill. The ISME Journal (2015)

[203] Cao, B., Nagarajan, K., Loh, K.-C.: Biodegradation of aromatic compounds: current status and opportunities for biomolecular approaches. Applied Microbiology and Biotechnology **85**(2), 207–228 (2009)

[204] Dombrowski, N., Donaho, J.A., Gutierrez, T., Seitz, K.W., Teske, A.P., Baker, B.J.: Reconstructing metabolic pathways of hydrocarbon-degrading bacteria from the Deepwater Horizon oil spill. Nature Microbiology **1**(7), 16057 (2016)

[205] Gionis, A., Indyk, P., Motwani, R., *et al.*: Similarity search in high dimensions via hashing. In: VLDB, vol. 99, pp. 518–529 (1999)

[206] Langford, J., Li, L., Strehl, A.: Vowpal Wabbit (Fast Online Learning). `https://github.com/JohnLangford/vowpal_wabbit/`. last accessed: July 12, 2018 (2007)

[207] Agarwal, A., Chapelle, O., Dudík, M., Langford, J.: A reliable effective terascale linear learning system. The Journal of Machine Learning Research **15**(1), 1111–1133 (2014)

[208] UniProt Consortium: UniProt: the universal protein knowledgebase. Nucleic Acids Research **45**(D1), 158–169 (2016)

[209] Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A.J., Poux, S., Bougueleret, L., Xenarios, I.: UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. In: Edwards, D. (ed.) Plant Bioinformatics. Methods in Molecular Biology, pp. 23–54. Humana Press, New York, NY (2016)

[210] Greenfield, P., Roehm, U.: Answering biological questions by querying k-mer databases. Concurrency and Computation: Practice and Experience **25**(4), 497–509 (2013)

[211] Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., Phillippy, A.M.: Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology **17**(1), 132 (2016)

[212] Bolger, A.M., Lohse, M., Usadel, B.: Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**(15), 2114–2120 (2014)

[213] Schmieder, R., Edwards, R.: Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PloS One **6**(3), 17288 (2011)

[214] Mahlich, Y., Steinegger, M., Rost, B., Bromberg, Y.: HFSP: high speed homology-driven function annotation of proteins. Bioinformatics **34**(13), 304–312 (2018)

[215] Valdés-Ramos, R., Ana Laura, G.-L., Beatriz Elina, M.-C., Alejandra Donaji, B.-A.: Vitamins and type 2 diabetes mellitus. Endocrine, Metabolic & Immune Disorders-Drug Targets (Formerly Current Drug Targets-Immune, Endocrine & Metabolic Disorders) **15**(1), 54–63 (2015)

[216] 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., *et al.*: Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biology **13**(9), 79 (2012)

[217] Lewis, J.D., Chen, E.Z., Baldassano, R.N., Otley, A.R., Griffiths, A.M., Lee, D., Bittinger, K., Bailey, A., Friedman, E.S., Hoffmann, C., *et al.*: Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn's disease. Cell Host & Microbe **18**(4), 489–500 (2015)

[218] Liu, Y., Wang, X., Hu, C.-A.: Therapeutic potential of amino acids in inflammatory bowel disease. Nutrients **9**(9), 920 (2017)

[219] Sicard, J.-F., Le Bihan, G., Vogeleer, P., Jacques, M., Harel, J.: Interactions of intestinal bacteria with components of the intestinal mucus. Frontiers in Cellular and Infection Microbiology **7**, 387 (2017)

[220] Agus, A., Planchais, J., Sokol, H.: Gut microbiota regulation of tryptophan metabolism in health and disease. Cell Host & Microbe **23**(6), 716–724 (2018)

[221] Santiago, J.A., Potashkin, J.A.: Shared dysregulated pathways lead to Parkinson's disease and diabetes. Trends in Molecular Medicine **19**(3), 176–186 (2013)

[222] Kim, D.S., Choi, H.-I., Wang, Y., Luo, Y., Hoffer, B.J., Greig, N.H.: A new treatment strategy for Parkinson's disease through the gut–brain axis: the glucagon-like peptide-1 receptor pathway. Cell Transplantation **26**(9), 1560–1571 (2017)

[223] Saunders, G., Baudis, M., Becker, R., Beltran, S., Béroud, C., Birney, E., Brooksbank, C., Brunak, S., Van den Bulcke, M., Drysdale, R., et al.: Leveraging European infrastructures to access 1 million human genomes by 2022. Nature Reviews Genetics, 1–9 (2019)

[224] Svensson, V., Vento-Tormo, R., Teichmann, S.A.: Exponential scaling of single-cell RNA-seq in the past decade. Nature Protocols **13**(4), 599 (2018)

[225] Lu, Y., Han, J.: Cancer classification using gene expression data. Information Systems **28**(4), 243–268 (2003)

[226] Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of Eugenics **7**(2), 179–188 (1936)

[227] Fix, E., Hodges, J.L.: Discriminatory analysis. Nonparametric discrimination: consistency properties. International Statistical Review/Revue Internationale de Statistique **57**(3), 238–247 (1989)

[228] Dudoit, S., Fridlyand, J., Speed, T.P.: Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association **97**(457), 77–87 (2002)

[229] Keller, A.D., Schummer, M., Hood, L., Ruzzo, W.L.: Bayesian classification of DNA array expression data. Technical Report UW-CSE-2000-08-01 (2000)

[230] Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian networks to analyze expression data. Journal of Computational Biology **7**(3-4), 601–620 (2000)

[231] Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery **2**(2), 121–167 (1998)

[232] Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience, New York, NY (1998)

[233] Segata, N.: Gut microbiome: westernization and the disappearance of intestinal diversity. Current Biology **25**(14), 611–613 (2015)