Quantifying wellness and disease with personal, dense, dynamic data clouds

John C. Earls

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Nathan D. Price, Chair

Walter L. Ruzzo, Chair

Ed Lazowska

Program Authorized to Offer Degree:
Computer Science and Engineering

University of Washington

Abstract

Quantifying wellness and disease with personal, dense, dynamic data clouds

John C. Earls

Co-Chairs of the Supervisory Committee:
Affiliate Professor Nathan D. Price
Paul G. Allen School of Computer Science and Engineering

Professor Walter L. Ruzzo
Paul G. Allen School of Computer Science and Engineering

Precision Medicine, where medical treatment is guided by deep molecular knowledge of the individual, has gained momentum in recent years. Rapid advancement in biological measurement technologies such as genome sequencing, mass spectrometry, protein capture assays, microfluidics and quantified self devices provide an unprecedented opportunity to quantify, explain, and affect each person's health. The key challenge now is how to utilize these new capabilities to maximize wellness and prevent disease. These developments are concurrent with and aided by the increased availability of robust data analytic tools and cheap, scalable computation. In this dissertation, I present three steps taken to advance Precision Medicine. I present the first large multi-omic wellness study, where information from these -omics were integrated and used to provide personalized wellness guidance through a trained wellness coach. I present a holistic and modifiable wellness marker based on aging, generated from longitudinal multi-omic data. Finally, I apply systems approaches with dense phenotypic longitudinal data to profiling cancer, highlighting one approach to personalized 'N of 1' medicine. The research I present in this dissertation has led to the formation of two companies, so far.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I start by thanking God for bringing all of the wonderful people listed here into my life. I acknowledge the support, encouragement, and love given me by my family. I thank my parents; Denvis and Linda Earls, and my siblings and their families; Bryan Earls, Liz Henkel, and Jennifer Watson. I thank my step-children and their spouses; James and Jessica Davis, and Bradney and Annika Mullens. I thank my little guys; James "Kole" Davis, Garret "Garr" Baumgartner, Andrew "Drewno" Springman, Braxton "Brax" Springman, Linkoln "Linkie" Davis, and Luka "Carl" Davis. I thank my roses; Kloee "Kloebug" Davis, Jenna "Honey-π" Mullens, and Gracie "Grace" Baumgartner. I am thankful to the American taxpayer for their financial support, and hope their investment pays them dividends. I thank the many friends I have made along the way, especially Chris Schlossberg, Vangelis Simeonidis, and Daniel Hart. I thank the clients of Arivale and participants in the Pioneer 100. I thank my partner in crime at Arivale and good friend, Andrew Magis. I thank the many brilliant researchers I have worked with at the Institute for Systems Biology, especially James "Big G" Eddy, Shuyi Ma, Seth Ament, Jocelynn Pearl, Noa Rappaport, and Tomasz Wilmanski. I thank Daniela Witten, who on several occasions helped clarify statistical issues, and proposed fruitful paths. I thank Ed Lazowska, whose entrepreneurship course gave me my first taste of the "Start-up bug." I thank Walter "Larry" Ruzzo, whose good humor, and quiet encouragement have made a huge difference and demonstrated to me that you do not need to always be serious to be a seriously good computer scientist. I thank the inestimable Leroy Hood, whose patient mentorship, determined optimism, and indomitable will continue to humble and amaze me. Finally, I thank Nathan D. Price, "O Captain! My Captain!"

DEDICATION

To my beautiful niece, Hannah Earls, for whom I started this journey, and my amazing wife, Lita Earls, without whom I could not have survived.

Chapter 1

INTRODUCTION

1.1 Overview

A central theme of research at the Institute for Systems Biology is centered on the idea that healthcare in the 21st Century should be predictive, preventative, personalized, and participatory (P4 Medicine). A key enabling approach has been to generate thousands of multi-modal, longitudinal measurements to quantify the health status of individuals. We refer to these as personal, dense, dynamic data clouds (PD3 clouds): personal because each data cloud corresponds to the particulars of an individual; dense to reflect the large number of measurements taken per person; and dynamic to reflect the longitudinal nature of the data monitoring over time. These PD3 clouds enable the development of systems and holistic biomarkers that integrate high-dimensional, multi-modal data sources for quantifying wellness and identifying disease, which are critical aspects of P4 Medicine. The National Research Council outlined their vision for a 21st-century learning healthcare system in their report "*Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy for Disease*." Their charge was "to explore the feasibility and need for a new taxonomy of human disease based on molecular biology" [1]. This report proposed the generation of an "Information Commons," where information accumulated during the course of health care is integrated into a "Knowledge Network" that can then guide patient care by generation of a new taxonomy of disease. It was proposed that the Information Commons contain genetics, epigenetics, microbiome, exposures, and associated signs and symptoms from individuals followed over time. Such a resource could help improve patient outcomes by providing information to inform "Precision Medicine" decisions, where treatment is guided by an individual's unique biochemistry, exposures, and genetics. Such an

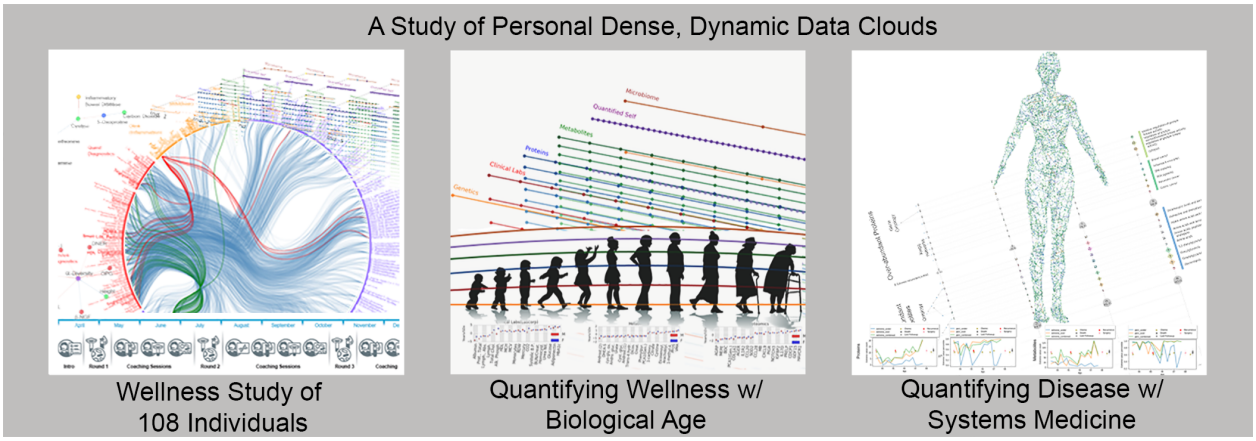
approach is in contrast to most current practices which provide treatment based solely on signs and symptoms without accounting for whether the treatment effectively addresses the uniquely perturbed biochemical mechanism in the patient. Current advances in cancer treatment are a prime example of the precision medicine envisioned in this report. Drugs like Crizotinib, an inhibitor of MET and ALK, are given to treat non-small cell lung cancer only in patients with a particular chromosomal translocation in the gene encoding ALK [2, 3].

In my graduate studies, I have had the opportunity to help bring this vision of 21st-century healthcare closer to fruition, and even extend this vision by incorporating wellness and disease prevention. I have addressed many challenges and contributed not only by writing grants and papers but also by developing infrastructure and analytical tools that have been used to power many studies at the Institute for Systems Biology. These tools allow multi-omic integrative analysis to be performed more effectively. I am primarily interested in applying computational methods to improve individuals' health and healthcare as a whole.

This dissertation is organized into three main chapters, where the first two chapters each represent a first author paper written during my graduate studies [4, 5]. The third chapter presents my work on developing systems analysis tools and methodology to understand disease better using PD3 clouds and will be published later. Several contributions of each work are highlighted in this overview.

Chapter 2 examines the results of the first large-scale multi-omic, longitudinal wellness study, termed the Pioneer 100 or P100. The P100 had four objectives which were to a) establish cost-efficient procedures for generating, storing, and analyzing multiple sources of health data obtained over time from participants and analyzed in combination with genomic data, b) develop and use analytic tools to integrate these diverse data sets and derive actionable information from their observed interrelationships, c) identify patterns within the health data that correlate with wellness or transitions between wellness and disease, and d) learn how to best work with and present longitudinal health information to individuals by studying the participants' reactions and feedback as they are presented with actionable information. I developed much of the computational architecture and analysis techniques in

Figure 1.1: Dissertation overview



Chapter 2 focuses on the Hundred Person Wellness Project. Chapter 3 shows how estimation of Biological Age from multi-omic data represents a holistic marker for wellness. Chapter 4 demonstrates the application of systems approaches to disease using multi-omic longitudinal measurement.

achieving the first three goals, in addition to performing the analysis. This chapter led to the formation of a direct-to-consumer wellness company, Arivale, that ran for four years but is now closed due to financial constraints. I took a year and a half leave from my graduate studies to help get Arivale off the ground.

Chapter 3 examines the quantification of wellness using multi-omic data. Customers of Arivale provided this data, and the analysis used data pipelines and tools developed over my time at Arivale and ISB. This study showed that computation of Biological Age (BA) from multi-omic measures was a) elevated in the presence of chronic diseases, b) modifiable through healthy lifestyle changes, c) and primarily representable as three axes; metabolic health, inflammation, and toxin accumulation. BA is shown to be a general and interpretable "metric for wellness." The results and techniques developed in this study led to my co-founding a healthy aging company, Aevum Aging, which is being acquired by Onegevity Health.

Chapter 4 examines the application of PD3 clouds to cancer. I have developed an integrative cloud-hybrid system to automate deeply-phenotyped longitudinal data analysis. I used this system and applied it to longitudinal muli-omic data on cancers to: a) develop personalized, extreme value profiles for cancer patients, b) apply standard systems analysis techniques, using a multi-omic knowledge graph, c) report on common perturbations under breast and ovarian cancer treatment, and d) identify perturbed systems on an N-of-1 basis while each patient undergoes cancer treatment. Based on these individuals' unique PD3 profiles, I propose possible personalized adjuvant and alternative therapies, identify early warning signs of known comorbidities, and demonstrate that at least one individual was likely misdiagnosed.

Each of these chapters explores a different set of challenges to data-driven Precision Medicine. This research was motivated by my desire to improve health and wellness by applying deep phenotyping and computation. This field is rapidly advancing, in both academia and industry.

I look forward to seeing these changes become part of the wider world.

Chapter 2

A WELLNESS STUDY OF 108 INDIVIDUALS USING PERSONAL, DENSE, DYNAMIC DATA CLOUDS

Note: This chapter can be read here or in the cited publication [4]. There is no extra information here that was not released with the original publication.

2.1 Abstract

Personal data for 108 individuals were collected during a 9-month period, including whole genome sequences; clinical tests, metabolomes, proteomes, and microbiomes at three time points; and daily activity tracking. Using all of these data, we generated a correlation network that revealed communities of related analytes associated with physiology and disease. Connectivity within analyte communities enabled the identification of known and candidate biomarkers (e.g., gamma-glutamyltyrosine was densely interconnected with clinical analytes for cardiometabolic disease). We calculated polygenic scores from genome-wide association studies (GWAS) for 127 traits and diseases, and used these to discover molecular correlates of polygenic risk (e. g. , genetic risk for inflammatory bowel disease was negatively correlated with plasma cystine). Finally, behavioral coaching informed by personal data helped participants to improve clinical biomarkers. Our results show that measurement of personal data clouds over time can improve our understanding of health and disease, including early transitions to disease states.

2.2 Main

In order to understand the basis of wellness and disease, we and others have pursued a global and holistic approach termed 'systems medicine'[6]. The defining feature of systems

medicine is the collection of diverse longitudinal data for each individual. These data sets can be used to unravel the complexity of human biology and disease by assessing both genetic and environmental determinants of health and their interactions. We refer to such data as personal, dense, dynamic data clouds: personal, because each data cloud is unique to an individual; dense, because of the high number of measurements; and dynamic, because we monitor longitudinally. The convergence of advances in systems medicine, big data analysis, individual measurement devices, and consumer-activated social networks has led to a vision of healthcare that is predictive, preventive, personalized, and participatory (P4)[7], also known as 'precision medicine'. Personal, dense, dynamic data clouds are indispensable to realizing this vision[8]. The US healthcare system invests 97% of its resources on disease care[9], with little attention to wellness and disease prevention. Here we investigate scientific wellness, which we define as a quantitative data-informed approach to maintaining and improving health and avoiding disease.

Several recent studies have illustrated the utility of multi-omic longitudinal data to look for signs of reversible early disease or disease risk factors in single individuals. The dynamics of human gut and salivary microbiota in response to travel abroad and enteric infection was characterized in two individuals using daily stool and saliva samples[10]. Daily multi-omic data collection from one individual over 14 months identified signatures of respiratory infection and the onset of type 2 diabetes[11]. Crohn's disease progression was tracked over many years in one individual using regular blood and stool measurements[12]. Each of these studies yielded insights into system dynamics even though they had only one or two participants.

We report the generation and analysis of personal, dense, dynamic data clouds for 108 individuals over the course of a 9-month study that we call the Pioneer 100 Wellness Project (P100). Our study included whole genome sequences; clinical tests, metabolomes, proteomes, and microbiomes at 3-month intervals; and frequent activity measurements (i.e., wearing a Fitbit). This study takes a different approach from previous studies, in that a broad set of assays were carried out less frequently in a (comparatively) large number of people.

Furthermore, we identified 'actionable possibilities' for each individual to enhance her/his health. Risk factors that we observed in participants' clinical markers and genetics were used as a starting point to identify actionable possibilities for behavioral coaching.

We report the correlations among different data types and identify population-level changes in clinical markers. This project is the pilot for the 100,000 (100K) person wellness project that we proposed in 2014 (ref. [13]). An increased scale of personal, dense, dynamic data clouds in future holds the potential to improve our understanding of scientific wellness and delineate early warning signs for human diseases.

2.3 Results

The P100 study had four objectives. First, establish cost-efficient procedures for generating, storing, and analyzing multiple sources of health data obtained over time from participants and analyzed in combination with genomic data. Second, develop and use analytic tools for integrating these diverse data sets and deriving actionable information from their observed interrelationships. Third, identify patterns within the health data that correlate with wellness, or transitions between wellness and disease. And fourth, learn how to best work with and present longitudinal health information to individuals by studying the reactions and feedback from participants as they are presented with actionable information.

2.3.1 Data collection

108 individuals (ages 21-89+ years; 59% males, 41% females; 89% of European descent; not recruited based on any specific phenotype) (Supplementary Table 2.S1) participated in this study. In month 4, one participant reported to her coach that she had become pregnant and was withdrawn per protocol and informed consent. Health history and behavioral assessments were performed at the beginning of the study to establish a baseline for health coaching, including tobacco (4 reported users) and alcohol consumption (91 reported users). Each individual had their genome sequenced in full. Blood was collected in clinics every 3 months. Additionally, participants completed at-home collections of saliva, stool, and first

morning void urine every 3 months. Stool and saliva samples were shipped directly to the vendor by the participant, while urine was given back to the study coordinators for distribution to the proper sample vendor (Figure 2.1). We called each of these three collection periods 'rounds'. For each participant in each round we carried out 218 clinical laboratory tests, measured up to 643 metabolites and 262 proteins, and measured the abundance of 4,616 operational taxonomic units in the gut microbiome using 16S rRNA sequencing. We used the whole genome sequence to calculate 127 polygenic scores for disease risks and quantitative traits based on previous studies selected from the National Human Genome Research Institute (NHGRI) GWAS catalog[14]. Three common copy number variations were also included as genomic features, bringing the total to 130. All vendor measurements are listed in Supplementary Table 2.S1. Participants were asked to record weight, blood pressure, and resting heart rate weekly, and to track activity and sleep daily using a wearable device (Fitbit), although compliance with quantified self-tracking was relatively low.

2.3.2 Community structure in the correlation networks

We built two age- and sex-adjusted correlation networks based on Spearman correlations across our cohort of individuals (Figure 2.2 and Supplementary Figure2.S1). Cross-sectional correlations were calculated from mean measurements of analytes calculated using all three rounds (mean A is correlated with mean B across all individuals). Delta correlations were calculated on the change in analytes between rounds (the change in A between time points is correlated with the change in B across all individuals). In these networks, vertices (V) correspond to analytes, and an edge (E) exists between two vertices if and only if a significant ($p_{adj} < 0.05$) correlation was observed after correction for multiple hypotheses[15]. The inter-omic cross-sectional correlation network contains 766 nodes and 3,470 edges. The majority of edges involved a metabolite (3,309) or a clinical laboratory test (3,366), with an additional 20 edges involving the 130 genetic traits tested, 46 with microbiome taxa or diversity score, and 207 with quantified proteins. The inter-omic delta correlation network contained 822 nodes and 2,406 edges. 375 of the edges in the delta correlation network

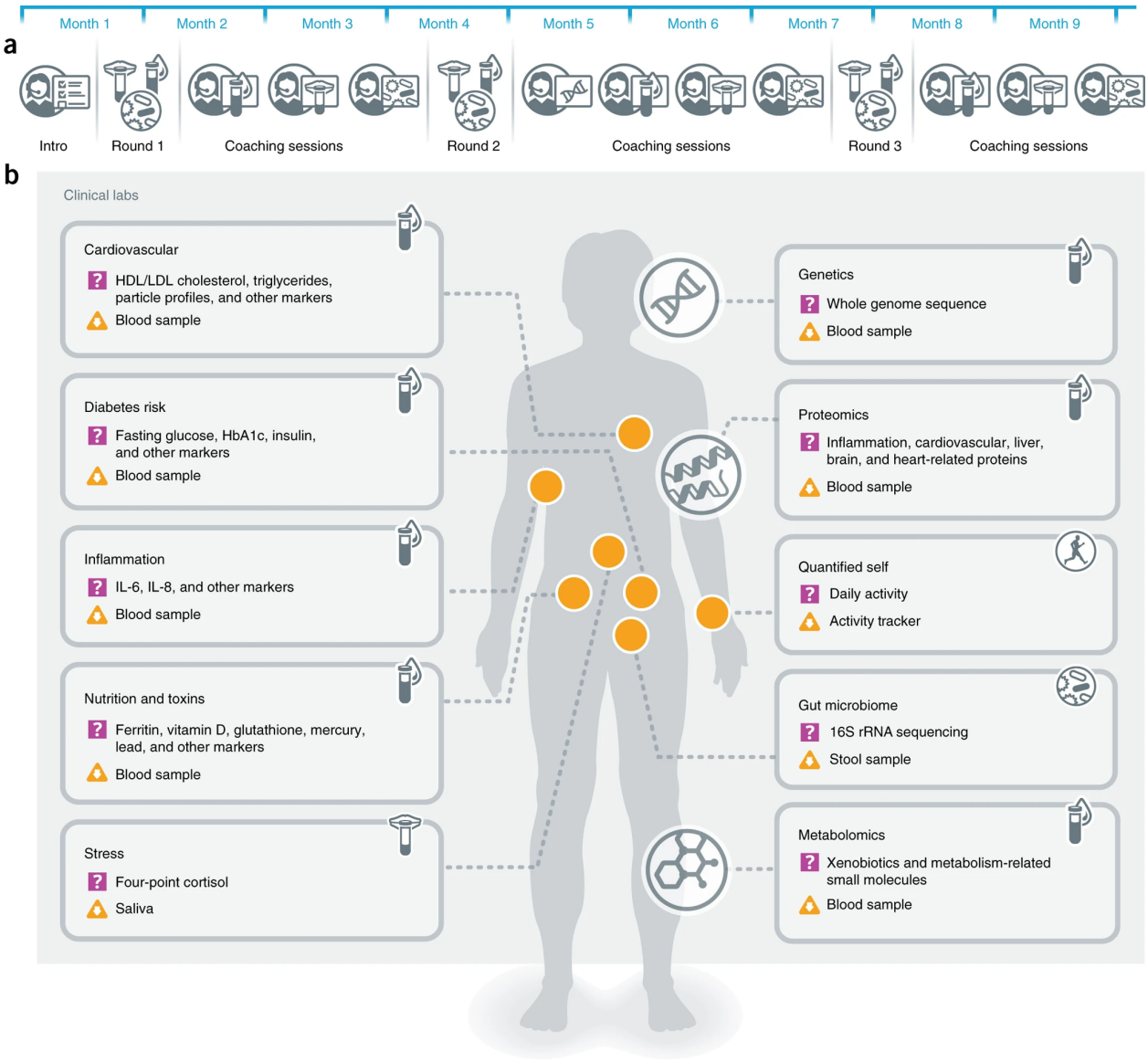


Figure 2.1: Types of longitudinal data collected.

(a) Timeline of important events in the P100. (b) Schematic of the data collected every 3 months throughout the study.

were also present in the cross-sectional network. The cross-sectional correlation network is provided in Supplementary Table 2.S2 (inter-omic only) and Supplementary Table 2.S3 (full). The delta correlation network is provided in Supplementary Table 2.S4 (inter-omic only) and Supplementary Table 2.S5 (full).

We identified clusters of related measurements from the cross-sectional inter-omic correlation network using community analysis, an unsupervised (i.e., using unlabeled data to find hidden structure) approach that iteratively prunes the network (removing the edges with the highest betweenness) to reveal densely interconnected subgraphs (communities)[16]. Seventy communities of at least two vertices (mean of 10.9 V and 34.9 E) were identified in the cross-sectional inter-omic network at the cutoff with maximum community modularity[17] (Supplementary Figure 2.S2), and are fully visualized as an interactive graph in Cytoscape[18] (Supplementary Dataset 1). 70% of the edges in the cross-sectional network remained after community edge pruning. The communities often represented a cluster of physiologically related analytes, as described below.

The largest community (246 V; 1,645 E) contains many clinical analytes associated with cardiometabolic health, such as C-peptide, triglycerides, insulin, homeostatic risk assessment-insulin resistance (HOMA-IR), fasting glucose, high-density lipid (HDL) cholesterol, and small low-density lipid (LDL) particle number (Figure 2.3). The four most-connected clinical analytes by degree (the number of edges connecting a particular analyte) were C-peptide (degree 99), insulin (88), HOMA-IR (88), and triglycerides (75). The four most-connected proteins measured using targeted (i.e., selected reaction monitoring analysis) mass spectrometry or Olink proximity extension assays by degree are leptin (18), C-reactive protein (15), fibroblast growth factor 21 (FGF21) (14), and inhibin beta C chain (INHBC) (10). Leptin and C-reactive protein are indicators for cardiovascular risk[19, 20]. FGF21 is positively correlated with the clinical analytes C-peptide (Spearman's $\rho = 0.51; p_{adj} = 3.1 \times 10^{-3}$), triglycerides ($\rho = 0.50; p_{adj} = 3.3 \times 10^{-3}$), HOMA-IR ($\rho = 0.50; padj = 3.6 \times 10^{-3}$), insulin ($\rho = 0.47; p_{adj} = 9.0 \times 10^{-3}$), and small LDL particle number ($\rho = 0.42; p_{adj} = 4.3 \times 10^{-2}$), and is a recently reported biomarker for cardiometabolic disorders[21]. INHBC, a member

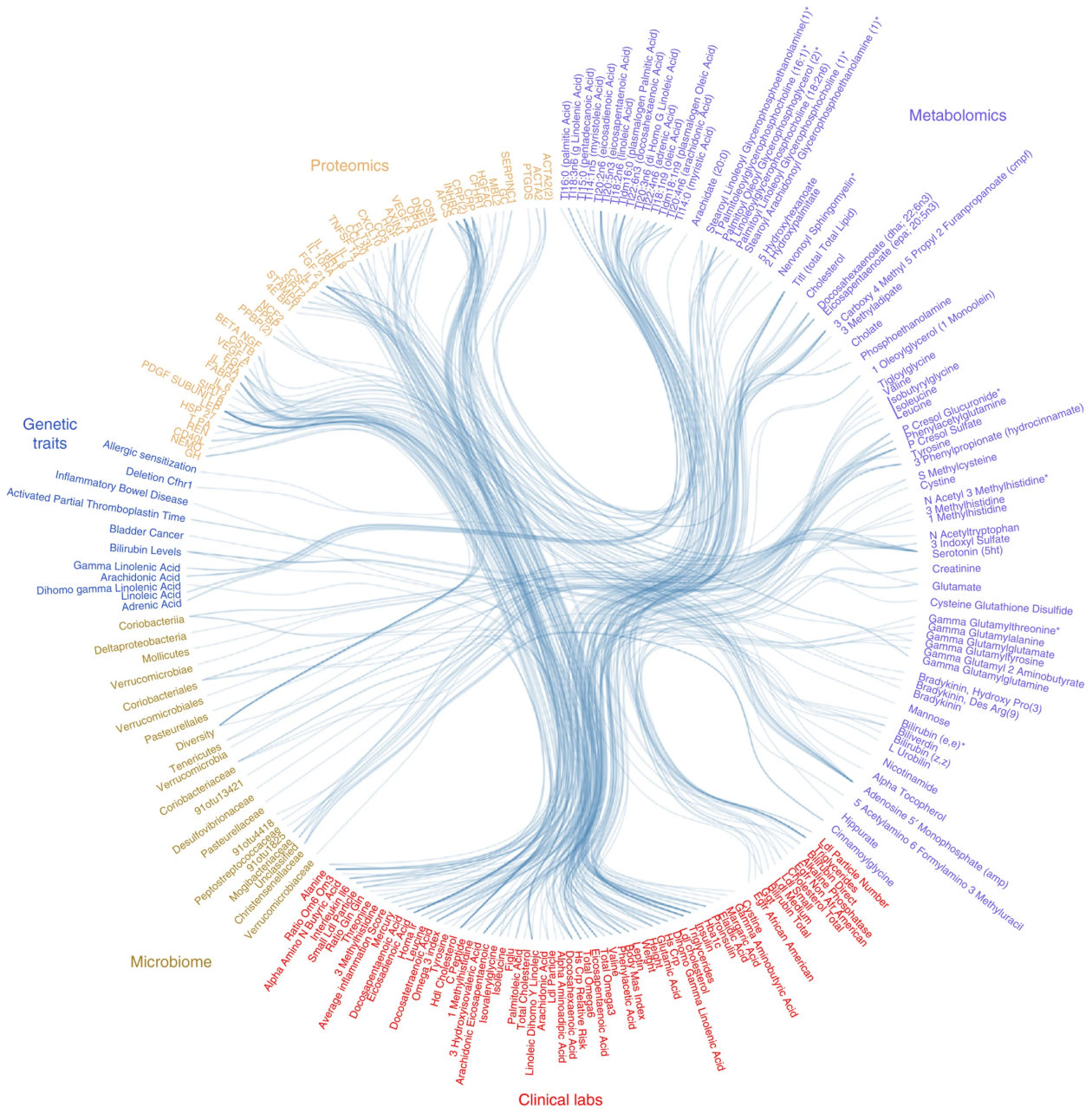


Figure 2.2: Top 100 correlations per pair of data types.

Subset of top statistically significant Spearman inter-omic cross-sectional correlations between all data sets collected in our cohort. Each line represents one correlation that was significant after adjustment for multiple hypothesis testing using the method of Benjamini and Hochberg[15] at $p_{adj} < 0.05$. The mean of all three time points was used to compute the correlations between analytes. Up to 100 correlations per pair of data types are shown in this figure. See Supplementary Figure2.S1 and Supplementary Table 2.S2 for the complete inter-omic cross-sectional network.

of the TGF-beta superfamily, is positively correlated with the clinical analytes triglycerides ($\rho = 0.45; p_{adj} = 3.0 \times 10^{-3}$), small LDL particle number ($\rho = 0.43; p_{adj} = 6.8 \times 10^{-3}$), C-peptide ($\rho = 0.40; p_{adj} = 1.8 \times 10^{-2}$), HOMA-IR ($\rho = 0.38; p_{adj} = 3.4 \times 10^{-2}$), and insulin ($\rho = 0.38; p_{adj} = 3.8 \times 10^{-2}$); it has not been reported to be a marker for cardiovascular risk and therefore represents an interesting candidate for follow-up. Serum amyloid P component (SAP) was positively correlated with LDL particle number ($\rho = 0.39; p_{adj} = 1.8 \times 10^{-2}$). SAP is a universal constituent of amyloid deposits observed in Alzheimer's disease[22], and is associated with myocardial infarction[23].

Total cholesterol and LDL cholesterol (LDL-C) segregate into a separate community from the cardiometabolic community (22 V; 48 E) with a broad array of plasma lipids (Figure 2.4a). Thyroid hormone L-thyroxine is also present and is negatively correlated with total cholesterol levels ($\rho = -0.44; p_{adj} = 5.0 \times 10^{-4}$) as well as LDL cholesterol ($\rho = -0.41; p_{adj} = 2.1 \times 10^{-3}$). Hypothyroidism has long been recognized clinically as a cause of elevated cholesterol values[24].

A community formed around plasma serotonin (18 V; 25 E) containing 12 proteins listed in Supplementary Table 2.S6, for which the most significant enrichment identified in a STRING ontology analysis[25] was platelet activation ($p_{adj} = 1.7 \times 10^{-3}$) (Figure 2.4b). Serotonin is known to induce platelet aggregation[26]; accordingly, selective serotonin reuptake inhibitors (SSRIs) may protect against myocardial infarction[27].

We identified several communities containing microbiome taxa, suggesting that there are specific microbiome-analyte relationships. Hydrocinnamate, L-urobilin, and 5-hydroxyhexanoate clustered with the bacterial class Mollicutes and family Christensenellaceae (8 V; 8 E). Another community emerged around the Verrucomicrobiaceae and Desulfovibrionaceae families and p-cresol-sulfate (7 V; 6 E). The Coriobacteriaceae and Mogibacteriaceae families were associated (12 V; 19 E) with phenylacetic acid, eicosadienoic acid, p-cresol-glucuronide, taurine, and phenylacetylglutamine. Phenylacetylglutamine, a known microbial metabolite[28], was recently identified as a risk factor for mortality and cardiovascular disease in chronic kidney disease patients[29]. Finally, the bile acid cholate clusters with the Peptostreptococ-

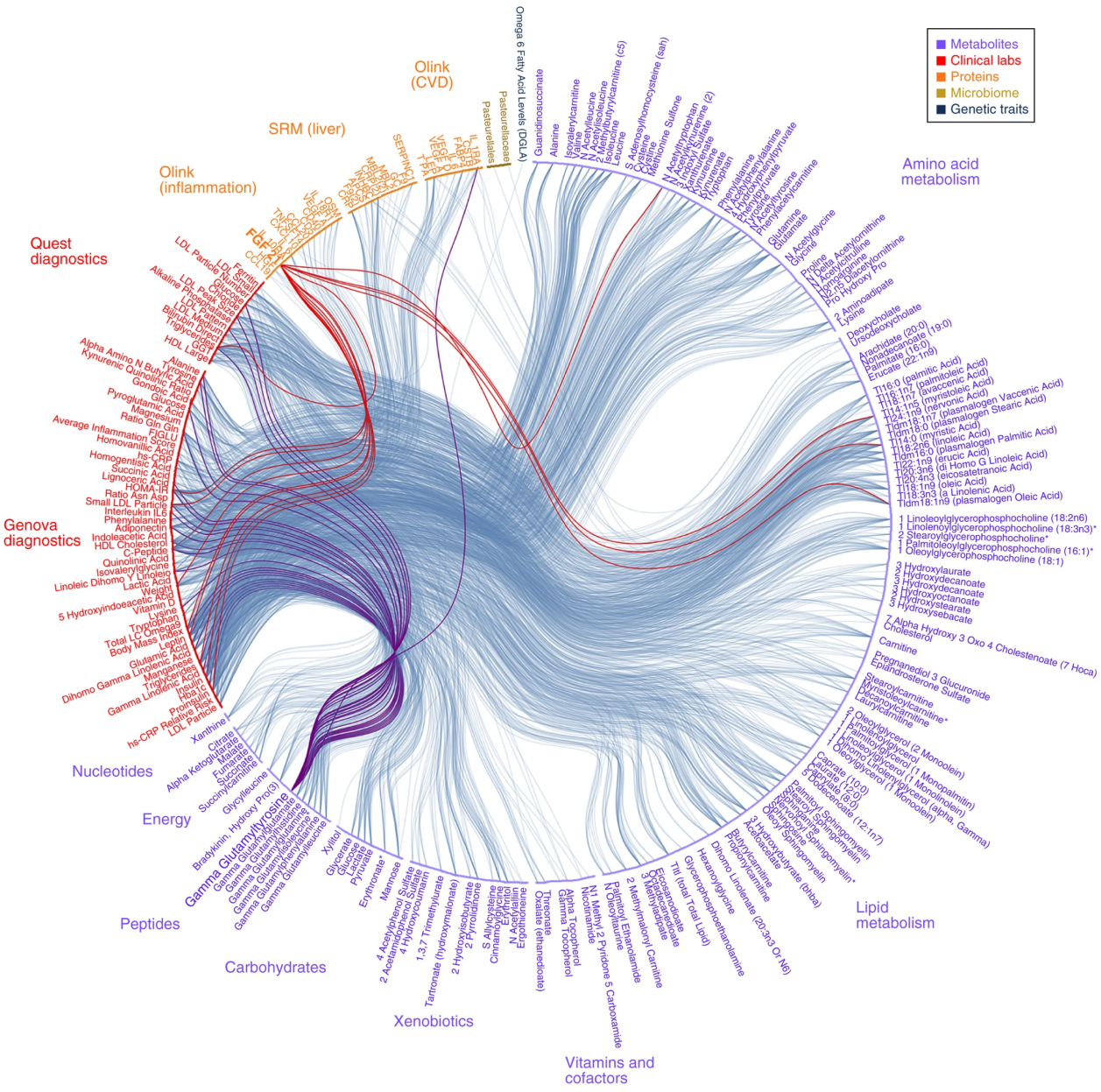


Figure 2.3: Cardiometabolic community

All vertices and edges of the cardiometabolic community, with lines indicating significant (padj < 0.05) correlations. Associations with FGF21 (red lines) and gamma-glutamyltyrosine (purple lines) are highlighted.

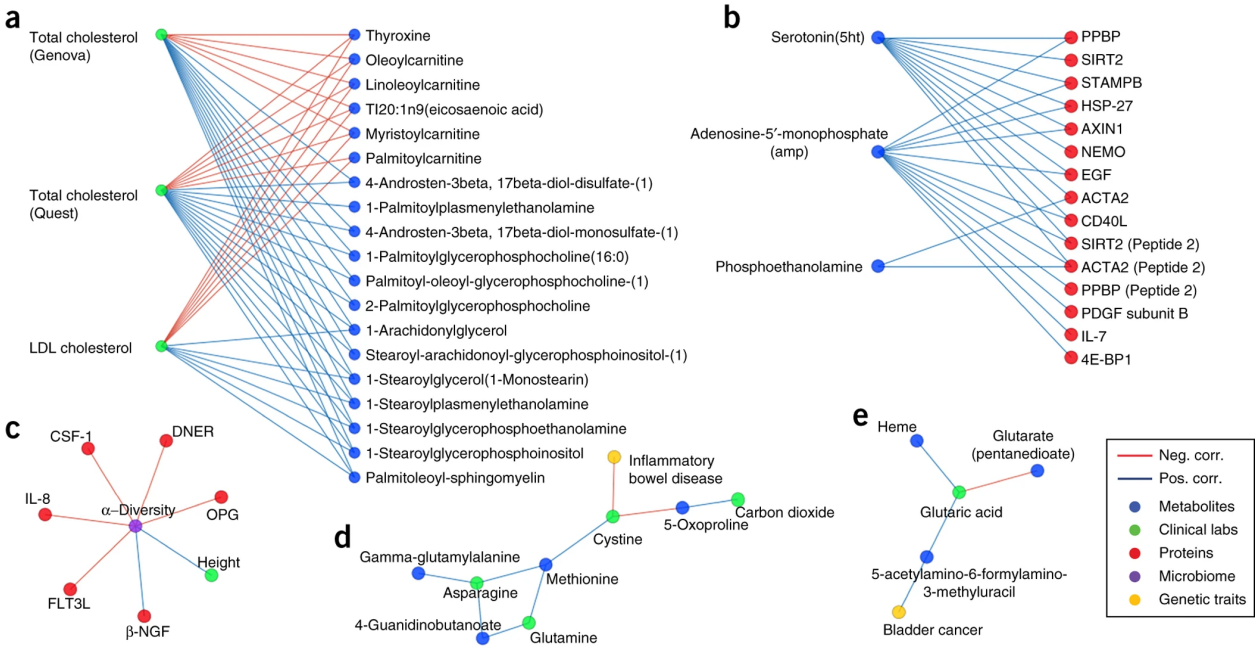


Figure 2.4: Cholesterol, serotonin, α-diversity, IBD, and bladder cancer communities.

(a) Cholesterol community. (b) Serotonin community. (c) α-diversity community. (d) The polygenic score for inflammatory bowel disease is negatively correlated with cystine. (e) The polygenic score for bladder cancer is positively correlated with 5-acetylamino-6-formylamino-3-methyluracil (AFMU).

caceae family (2 V; 1 E).

A community formed around microbiome α-diversity (8 V; 7 E), a measure of the number of operational taxonomic units observed and the evenness of their distributions; elevated diversity is generally thought to be associated with better health in part by ameliorating inflammation[30]. Microbiome α-diversity was negatively correlated with inflammatory and immune-related proteins, including interleukin-8 (IL-8), FMS-related tyrosine kinase 3 (FLT3LG), and macrophage colony-stimulating factor 1 (CSF1) (Figure 2.4c). In contrast, B-nerve growth factor (NGF) was positively correlated with microbiome α-diversity. An analysis using STRING20 on α-diversity community members revealed a significant enrichment in the KEGG pathway cytokine-cytokine receptor interaction ($p_{adj} = 1.1 \times 10^{-4}$); other pathway members have been implicated in the pathogenesis of inflammatory bowel disease[31].

2.3.3 Mining multi-omic communities for potential biomarkers

One highly interconnected metabolite in the cardiometabolic community, gamma-glutamyltyrosine (degree 27), was significantly correlated with markers of cardiometabolic disease: glucose ($\rho = 0.41; p_{adj} = 1.6 \times 10^{-3}$), HOMA-IR ($\rho = 0.38; p_{adj} = 6.0 \times 10^{-3}$), and insulin ($\rho = 0.36; p_{adj} = 9.7 \times 10^{-3}$), as well as triglycerides ($\rho = 0.41; p_{adj} = 1.5 \times 10^{-3}$), small LDL particle number ($\rho = 0.35; p_{adj} = 1.5 \times 10^{-2}$), and HDL cholesterol ($\rho = -0.35; p_{adj} = 1.6 \times 10^{-2}$). Gamma-glutamyltyrosine is produced by the enzyme gamma-glutamyl transferase (GGT), a known biomarker of diabetes risk independent of body mass index (BMI)[32, 33]. We carried out an ordinary least-squares (OLS) regression with homeostatic risk assessment (HOMA-IR, a common marker for insulin resistance), as the dependent variable and GGT, gamma-glutamyltyrosine, age, sex, and BMI as the regressors ($R^2_{adj} = 0.46$) (Supplementary Table 2.S7). In this model, gamma-glutamyltyrosine has a more significant effect on HOMA-IR ($P = 4.3 \times 10^{-6}$) than does GGT (P = 0.09). If this finding is confirmed in a larger number of unrelated samples, gamma-glutamyltyrosine could be a candidate biomarker for diabetes risk independent of BMI.

2.3.4 Delta correlation network identifies changes over time

Thirty-three communities of at least two vertices (mean of 24.9 V and 59.2 E) were identified in the inter-omic delta correlation network at the cutoff with maximum community modularity12 (Supplementary Dataset 2). 81% of the edges in the delta network remained after community edge pruning. This network contains many interesting relationships not found in the cross-sectional network. For example, changes in HDL cholesterol were positively correlated with changes in galanin ($\rho = 0.36; p_{adj} = 4.8 \times 10^{-3}$), a neuropeptide hormone with many physiological functions, including therapeutic associations with diabetes and Alzheimer's disease[34]. One of the delta communities (V = 15; E = 28) involved the omega-3 fatty acids eicosapentaenoic acid and docosahexaenoic acid (DHA), as well as the clinical analyte omega-3 index. Also present in this delta community is the furan fatty acid metabolite 3-carboxy-4-methyl-5-propyl-2-furanpropanoate (CMPF). CMPF is elevated in the plasma of type 2 diabetes patients and directly implicated in beta cell dysfunction[35], and has previously been observed to increase in response to omega-3 fatty acid supplements in diabetic patients[36].

Polygenic scores correlate with disease-risk analytes Several edges in the cross-sectional network represented correlations between genetic traits and corresponding biomarkers already identified in published studies. For example, blood levels of dihomo-γ-linolenic acid (DGLA) in our study were strongly correlated ($\rho = 0.52; p_{adj} = 1.8 \times 10^{-4}$) with a polygenic score computed from genotypes in six variants that were previously associated with DGLA levels[37] (Figure 2.5a). We observed similar results for other omega-6 fatty acids including arachidonic acid, linoleic acid, and eicosadienoic acid as well as bilirubin, a marker of liver dysfunction ($\rho = 0.52; p_{adj} = 2.3 \times 10^{-4}$)[38] (Figure 2.5b). All tested associations with quantitative traits are presented in Supplementary Table 2.S8.

Although GWAS studies that model quantitative traits are most directly applicable to the measurements made in our study, other edges in the network occurred between polygenic disease risk and specific analytes. For example, the genetic risk of inflammatory bowel disease

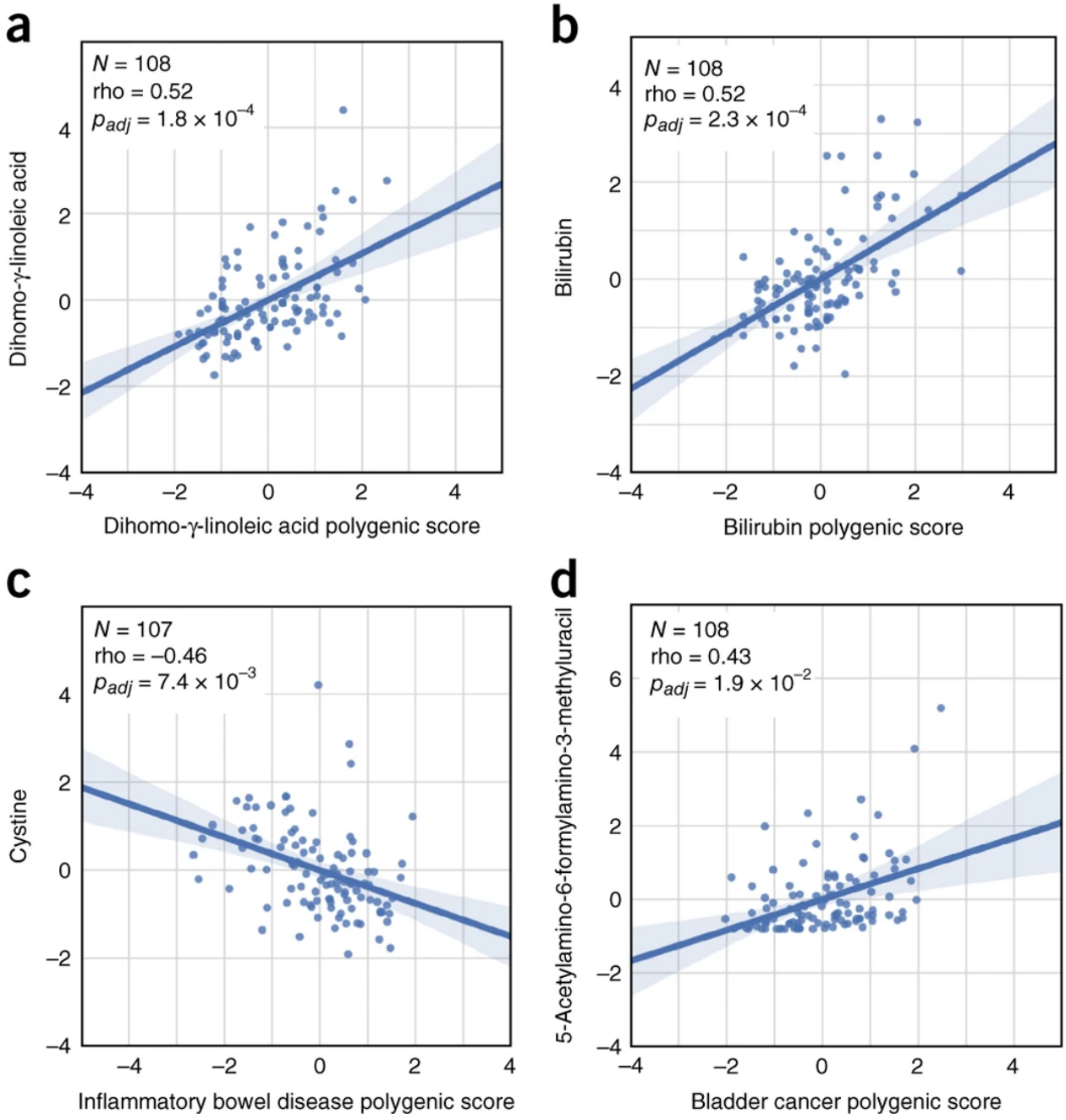


Figure 2.5: Polygenic scores correlate with blood analytes.

Spearman correlations between polygenic scores (x axis) and analyte measurements (y axis) from our correlation network. The number of measurements used for each pairwise comparison, correlation coefficients, and adjusted P-values are indicated on each figure. Values have been age and/or sex adjusted as described in Online Methods. The line shown is a y $\sim$ x regression line, and the shaded regions are 95% confidence intervals for the slope of the line.

(IBD) in Europeans has been associated with 110 single-nucleotide polymorphisms (SNPs)26. In our cohort, the polygenic score for IBD calculated from all 110 SNPs was significantly negatively correlated with plasma cystine, the disulfide form of cysteine ($\rho = -0.46; p_{adj} = 7.4 \times 10^{-3}$) (Figs. 2.4d and 52.5c).

We computed a bladder cancer polygenic score for all of our participants from nine SNPs previously associated with bladder cancer in a European cohort[39]. We identified an edge between this polygenic bladder cancer score and plasma levels of 5-acetylamino-6-formylamino-3-methyluracil (AFMU), an acetylated metabolite of caffeine, in our cohort. ($\rho = 0.43; p_{adj} = 1.9 \times 10^{-2}$). One variant is located downstream of NAT2, which encodes the enzyme N-acetyltransferase-2 responsible for acetylating carcinogenic compounds in urine. Polymorphisms in NAT2 are known to produce 'fast' and 'slow' acetylator phenotypes, of which the latter conveys higher risk for bladder cancer[40] (Figs. 2.4e and 2.5d).

2.3.5 Coaching and biomarker improvements

In order to help participants modify their behavior and potentially improve their health throughout the 9-month period of this study, a behavioral coach talked participants through actionable possibilities from their data. Each month the coach worked with the participants and made recommendations for lifestyle changes with the aim of altering markers of known clinical significance and/or compensating for genetic predispositions (Figure 2.1 and Supplementary Figure 2.S3). Specific coaching recommendations based on personal data were customized by the coach, in consultation with the study physician. Specifically, for each measurement in an individual that was outside the clinical reference range (out-of-range) recommended by the clinical laboratory, the coach would recommend lifestyle changes that have been previously demonstrated to produce improvements in that marker. In making personalized recommendations, the coach used evidence-based behavioral approaches tailored to the participant's preferences and behavioral skill level. For example, for individuals identified with elevated fasting glucose or HbA1c at baseline (pre-diabetes), the coach made recommendations based on the Diabetes Prevention Program[41], customized for each per-

son's lifestyle. These individual recommendations typically fell into one of several major categories: diet, exercise, stress management, dietary supplements, or physician referral, as relevant for each participant. Coaching focused on four primary health areas: cardiovascular, diabetes, inflammation, and nutrition. Weight loss was not a primary focus area. The clinical tests that were actively coached on are provided in Supplementary Table 2.S9. For clinical tests actively coached on, we used generalized estimating equations to calculate the average population change in each clinical analyte by round while controlling for the effects of age, sex, and self-reported ancestry. The results are shown in Table 2.1 and Supplementary Table 2.S10. The most significant improvements for those who began the study out of range were observed in vitamin D (+7.2 ng/mL/round), mercury (-0.002 mcg/g/round), and HbA1c (-0.085%/round). We observed consistent improvements in total cholesterol measured by both Quest and Genova (-6.4 mg/dL/round and -5.4 mg/dL/round, respectively). LDL cholesterol, measured only with Genova, significantly decreased (-4.8 mg/dL/round), while HDL cholesterol significantly increased (+4.5 mg/dL/round). Other significant improvements were observed in other diabetes risk factors (fasting insulin and HOMA-IR), and inflammation (IL-8 and TNF-alpha). Lipoprotein fractionation, performed by both laboratory companies, produced significant but discordant results for LDL particle number. We observed significant improvements in fasting glucose with Quest and concordant but non-significant improvements in fasting glucose with Genova.

During the introductory coaching call one participant, a 65-year-old male, reported decreased mobility during hiking trips with his family and that his primary care physician had identified cartilage damage in his ankle. The baseline blood collection revealed that he had ferritin levels of 399 ng/mL, above the clinical reference range, and subsequent genetic analysis by our clinical team revealed he was homozygous for HFE C282Y, the primary genetic risk factor for hereditary hemochromatosis. Given his reported ferritin levels and genetic risk factors, our clinical team referred him to a hematologist, who diagnosed hemochromatosis and prescribed therapeutic phlebotomy. At the next blood draw, ferritin levels had dropped to 175 ng/mL and remained normal throughout the remainder of the study (Supplementary

Table 2.1: Longitudinal analysis of clinical changes by round

Health area	Name	N	Change per round	P-value
Nutrition	Vitamin D	95	+7.2 ng/mL/round	7.1×10^{-25}
Nutrition	Mercury	81	-0.002 mcg/g/round	8.9×10^{-9}
Diabetes	HbA1c	52	-0.085%/round	9.2×10^{-6}
Cardiovascular	LDL particle number (Quest)	30	+130 nmol/L/round	9.3×10^{-5}
Nutrition	Methylmalonic acid (Genova)	3	-0.49 mmol/mol creatinine/round	2.1×10^{-4}
Cardiovascular	LDL pattern (A or B)	28	-0.16 /round	4.8×10^{-4}
Inflammation	Interleukin-8	10	-6.1 pg/mL/round	5.9×10^{-4}
Cardiovascular	Total cholesterol (Quest)	48	-6.4 mg/dL/round	7.2×10^{-4}
Cardiovascular	LDL cholesterol	57	-4.8 mg/dL/round	8.8×10^{-4}
Cardiovascular	LDL particle number (Genova)	70	-69 nmol/L/round	1.2×10^{-3}
Cardiovascular	Small LDL particle number (Genova)	73	-56 nmol/L/round	3.5×10^{-3}
Diabetes	Fasting glucose (Quest)	45	-1.9 mg/dL/round	8.2×10^{-3}
Cardiovascular	Total cholesterol (Genova)	43	-5.4 mg/dL/round	1.2×10^{-2}
Diabetes	Insulin	16	-2.3 IU/mL/round	1.5×10^{-2}
Inflammation	TNF-alpha	4	-6.6 pg/mL/round	1.8×10^{-2}
Diabetes	HOMA-IR	19	-0.56 /round	2.0×10^{-2}
Cardiovascular	HDL cholesterol	5	+4.5 mg/dL/round	2.2×10^{-2}
Nutrition	Methylmalonic acid (Quest)	7	-42 nmol/L/round	5.2×10^{-2}
Cardiovascular	Triglycerides (Genova)	14	-18 mg/dL/round	1.4×10^{-1}
Diabetes	Fasting glucose (Genova)	47	-0.98 mg/dL/round	1.5×10^{-1}
Nutrition	Arachidonic acid	35	+0.24 wt%/round	1.9×10^{-1}
Inflammation	hs-CRP	51	-0.47 mcg/mL/round	2.1×10^{-1}
Cardiovascular	Triglycerides (Quest)	17	-14 mg/dL/round	2.4×10^{-1}
Nutrition	Glutathione	6	+11 micromol/L/round	2.5×10^{-1}
Nutrition	Zinc	4	-0.82 mcg/g/round	3.0×10^{-1}
Nutrition	Ferritin	10	-14 ng/mL/round	3.1×10^{-1}
Inflammation	Interleukin-6	4	-1.1 pg/mL/round	3.8×10^{-1}
Cardiovascular	HDL large particle number	8	+210 nmol/L/round	4.9×10^{-1}
Nutrition	Copper	10	+0.006 mcg/g/round	6.0×10^{-1}
Nutrition	Selenium	6	+0.035 mcg/g/round	6.2×10^{-1}
Cardiovascular	Medium LDL particle number	20	+2.8 nmol/L/round	8.5×10^{-1}
Cardiovascular	Small LDL particle number (Quest)	14	-2.3 nmol/L/round	8.8×10^{-1}
Nutrition	Manganese	0	N/A	N/A
Nutrition	EPA	0	N/A	N/A
Nutrition	DHA	0	N/A	N/A

Generalized estimating equations (GEE) were used to calculate average changes in clinical laboratory tests over time, for those analytes that were actively coached on. The 'Change per round' column is the average change in the population for that analyte by round adjusted for age, sex, and self-reported ancestry. 'Out-of-range at baseline' indicates the average change using only those participants who were out-of-range for that analyte at the beginning of the study. Rows in boldface indicate statistically significant improvement, while the italicized row indicates statistically significant worsening. N/A values are present where no participants were out-of-range at baseline. For example, the average improvement in vitamin D for the 95 participants that began the study out-of-range was +7.2 ng/mL per round. Several analytes are measured by both Quest and Genova; with the exception of LDL particle number, the direction of effect for significantly changed analytes was concordant across the two laboratories. An independence working correlation structure was used in the GEE. See Supplementary Table 2.S10 for the complete results.

Figure 2.S4). Hemochromatosis leads to excessive accumulation of dietary iron in various tissues and can be associated with serious complications later in life, including cartilage damage, liver disease, diabetes, and cardiac decompensation. After diagnosis, this individual's primary care physician attributed his cartilage damage to early symptoms of hemochromatosis. Six other males had high ferritin levels but neither of the common genetic risk factors; four of the six were of Asian ancestry, out of only six male Asians in our study. It has previously been observed that Asians and Pacific Islanders have the highest mean population levels of ferritin despite a very low prevalence of risk factors for hemochromatosis[42]. These individuals were referred to their physicians for monitoring.

2.4 Discussion

We report here the main findings from the P100 Wellness Project. We computed thousands of statistically significant inter-omic correlations using personal, dense, dynamic data clouds to identify many associations that could be followed up with perturbation experiments. We partitioned the correlations into data communities, which placed biomarkers in context within biological networks. This in turn led to the identification of putative biomarkers such as gamma-glutamyltyrosine, which was highly interconnected with clinical analytes for cardiometabolic disease. We identified molecular correlates of polygenic disease risk scores computed from published GWAS data, revealing possible ways in which genetic predisposition is manifested through analyte changes. Finally, the clinical biomarkers of many participants significantly changed in a healthy direction (Table 2.1 and Supplementary Table 2.S10) during the course of the study (e.g., type 2 diabetes and cardiovascular risk factors). Together these findings show that personal, dense, dynamic data clouds embody the essence of precision medicine[8] and present possibilities for the discovery of important medical applications.

Data integration generated 3,470 significant ($p_{adj} < 0.05$) cross-sectional correlations and 2,406 significant delta (change over time) correlations after multiple hypothesis correction. Two known correlations indicate the potential existence of therapeutically valuable relation-

ships. First, our analysis identified FGF21 as a potential contributor to cardiometabolic health. Indeed, obese diabetic patients treated with an FGF21 analog have shown improvements in triglycerides and other cardiovascular markers[43]. Second, L-thyroxine, through a negative correlation, was placed in a data community with cholesterol markers; supplementation with L-thyroxine lowered total cholesterol and LDL-C levels in patients with hypothyroidism in a clinical trial[44]. These two examples were identified from our data in an unsupervised manner.

We detected gamma-glutamyltyrosine, a metabolite of the enzyme biomarker gamma-glutamyl transferase (GGT). GGT is a clinical biomarker for liver disease, diabetes, and cardiovascular disease risk[45]. GGT catalyzes the transfer of the gamma-glutamyl moiety of glutathione to a substrate, commonly another amino acid, producing gamma-glutamyl dipeptides[46]. One of these dipeptides, gamma-glutamyltyrosine, is highly interconnected within the cardiometabolic community and is a better predictor of HOMA-IR (insulin resistance) than GGT. Gamma-glutamyltyrosine might be a useful diagnostic marker for diabetes risk if our findings are replicated in an unrelated larger cohort. In clinical studies gamma-glutamyl dipeptides also discriminate different forms of liver disease[47] and predict 28-day mortality in intensive care unit patients[48].

We identified correlations between calculated polygenic scores derived from common GWAS variants and measured analytes. For several studies (Supplementary Table 2.S8) we were able to independently validate the cumulative associations of these variants with the expected quantitative trait (e.g., DGLA, LDL cholesterol, and bilirubin).

The polygenic score for IBD was significantly negatively correlated with levels of cystine in plasma samples of our cohort. A case-control study of IBD patients with either Crohn's disease or ulcerative colitis reported that plasma cystine and cysteine levels were abnormally low in affected individuals relative to controls, with the effect increasing with disease severity[49]. Decreased availability of the limiting substrate cystine suggests an impairment of glutathione synthesis in the intestine. Glutathione is an important intracellular antioxidant that is depleted in IBD inflammatory episodes, leading to excess reactive oxygen species

and subsequent colonic inflammation and oxidative damage. Although Sido et al. [49] discuss cystine deficiency as an effect rather than a cause of IBD, our preliminary data suggest that lower levels of blood cystine may be more common in individuals at higher genetic risk for IBD before the disease manifests itself.

Specific genetic variants can be used to explain metabolite profiles using targeted variant-pathway interactions[50]. Our data suggest that GWAS polygenic scores can identify analyte associations with disease risk in a non-targeted manner (e.g., AFMU versus bladder cancer) and in the absence of direct associations between GWAS loci and plausible metabolic pathways (e.g., cystine vs. IBD). It is possible that supplementation with cystine in healthy individuals with high IBD genetic risk could prevent long-term low-grade inflammation and oxidative damage and thereby stop the wellness-to-disease transition to IBD. This hypothesis requires validation in follow-up experiments.

Most (89%) of our study participants were of European ancestry, and most (87%) of the 127 GWAS used as features for the correlation network were determined using European-ancestry populations (Supplementary Table 2.S1). Principal component analysis plots of the population distribution of the P100 participants are shown in Supplementary Figures 2.S5 and 2.S6. We are now evaluating approaches to control for ancestry of individuals in the computation of polygenic scores. This study was constrained to a small population of individuals living primarily in Seattle and northern California, but as we expand to other geographic areas our population diversity will increase.

We provided activity trackers (Fitbits) to our participants with the goal of measuring activity and sleep, but observed only modest compliance. We required a minimum of 40 days of Fitbit usage to estimate the average activity for each participant; 64% of the participants met this criterion. We included mean activity calories as a feature in our correlation network, but did not observe any statistically significant correlations with this feature. We observed even lower compliance with sleep tracking.

We developed an internal pipeline and curation protocol based on the American College of Medical Genetics and Genomics (ACMG) recommendations for reporting incidental find-

ings in disease-associated genes (Online Methods). Three individuals were identified with previously annotated pathogenic or likely pathogenic variants. Our clinical team contacted each individual directly and securely provided them with a custom genetic report to take to their physician, along with a recommendation for follow-up clinical sequencing to confirm or disprove the incidental finding.

The opportunities for observing health transitions in a cohort of 108 individuals over 9 months are limited. We are now extending this pilot program with the aim of recruiting a large population[51] of more than 100,000 individuals by 2020.

Our study has the following limitations. The study design was not randomized in that none of the participants were denied wellness coaching or access to personal data.

We considered potential confounding from seasonality in the vitamin D improvements reported in Table 2.1 and Supplementary Table 2.S10. The improvements we observed in vitamin D levels (and attribute to supplementation) were considerably higher than would be expected based on seasonal changes in sun exposure[52]. Furthermore, we collected information on vitamin D dosage and report dose-dependent effects of supplementation (Supplementary Fig. 2.S7). We conclude that seasonality is a relatively minor confounding factor. Even after stringent multiple hypothesis correction, false discoveries are statistically inevitable. However, our approach could inform major efforts to translate omics-based data and build a learning healthcare system, as recently advocated by the US National Academy of Medicine[53].

We hope that analyses of personal, dense, dynamic data clouds for a much larger cohort will enable the identification of network perturbations that result in common diseases, the design of diagnostics to detect early disease transitions, and the development of drugs and other interventions to reverse disease at the earliest stages. Personal, dense, dynamic data clouds are the essence of what precision medicine should be.

2.5 Methods

2.5.1 Approval for the study.

Procedures for the P100 were run under the Western Institutional Review Board (IRB Protocol Number 20121979) at the Institute for Systems Biology (ISB). The study ran April 2014 to January 2015.

2.5.2 Recruitment of participants.

Individuals in Washington state and California were informally identified as interested in the P100 via personal communication and social networks of the authors. These individuals were then sent a formal e-mail announcement of the study from L.H. with an invitation to join. All 108 participants gave written informed consent for analysis of their data. Supplementary Figure 2.S3 is a flowchart of recruitment/dropout and other events in the P100.

2.5.3 Analysis plan.

At three time points throughout the study blood and urine samples from each participant were collected and processed using the procedures outlined by Genova Diagnostics and Quest Diagnostics and couriered to the testing facilities to maintain maximum sample stability.

Additional whole blood and plasma samples were collected from participants and shipped to BioStorage Technologies, an international CAP-accredited biorepository. Additional samples were used for metabolomics (Metabolon), SRM proteomics (ISB), Proseek Multiplex protein panels (Olink), and whole genome sequencing (Complete Genomics and the New York Genome Center). Participants collected stool samples at home for 16S rRNA sequencing (Second Genome), and were asked to provide daily activity and sleep data using personal monitoring devices (Fitbit).

Participants were asked to fast for 12 h before all blood collections. We observed a 99.3% compliance rate in fasting. Participants were asked by the phlebotomist to confirm compliance with the 12-h fast before each blood draw, and this was recorded on the requisition

document. The P100 project manager sent out reminder e-mails before each blood draw period ('round') with instructions on how long to fast. Our clinical team reviewed the clinical data from each blood draw before its use in coaching.

Data were used throughout the study for coaching. Results presented in this report were analyzed at the end of the 9-month study using uniform standardized bioinformatics pipelines. All raw data collected as part of the P100 are available from dbGaP with accession ID phs001363.v1.p1.

2.5.4 Clinical laboratory tests.

For Genova, a total of one urine tube and nine blood tubes were collected. The blood tubes consisted of two Na-Heparin Trace Element tubes, three Serum Separator Tubes (SST), three EDTA purple top tubes, and one NMR black-top LipoTube. First morning void urine was collected in the Genova-provided green-top tube by participants the morning of their blood draw. Urine was sent frozen to Genova. Both Na-Heparin tubes were spun for 15 min at 3,000 r.p.m. The plasma from one Na-Heparin tube was transferred to a blue-top preservative tube provided by Genova and shaken and spun for 5 min at 2,500 r.p.m. Supernatant was then transferred to the yellow top transfer tube provided by Genova and shipped frozen. Plasma from the second Na-Heparin was transferred to an amber top transfer tube and shipped frozen. Each SST tube was left to clot for 15 min then spun for 15 min at 3,000 r.p.m. The plasma for all three was pipetted to transfer tubes and shipped frozen. All three EDTA-lavender top tubes were refrigerated after collection and shipped refrigerated. The single NMR black-top LipoTube was clotted for 30 min then spun for 15 min at 3,000 r.p.m. The specimen was left in the tube and shipped refrigerated.

Each saliva collection consisted of four samples within a single day (four-point cortisol test). For collection of the four saliva samples, participants were instructed to abstain from eating or drinking 30 min before each collection. All participants were given the following collection times for each of their four samples. Sample 1: collect before breakfast, between 7 a.m. and 9 a.m. and 1 hour after waking up. Sample 2: collect before lunch, between 11

a.m. and 1 p.m. Sample 3: collect before dinner, between 3 p.m. and 5 p.m. Sample 4: collect before bedtime, between 10 p.m. and 12 a.m. All samples were frozen overnight after collection and shipped directly to Genova.

Two SST tubes were collected for Quest Diagnostics. After collection the two tubes were left to clot for 15 min and then spun for 15 min at 3,000 r.p.m. Samples were left in the tube and shipped at ambient temperature.

All clinical laboratory tests measured using Quest and Genova are listed in Supplementary Table 2.S1.

2.5.5 Whole genome sequencing.

Participant whole blood samples were submitted to either Complete Genomics Inc. (41 participants) or the New York Genome Center (NYGC; 67 participants) for whole genome sequencing (WGS). Complete Genomics conducted the whole genome sequencing using their standard complete sequencing platform employing high-density DNA nanoarrays populated with DNA nanoballs for 40x average coverage. The New York Genome Center used Illumina's 2 150 bp HiSeq X technology for 30x average coverage, using TruSeq kits for library prep. Both vendors aligned sequenced reads to human reference sequence GRCh37/hg19. NYGC used BWA v0.7.8-r455.

Complete Genomics provided a vcfBeta file for each sequenced sample calculated with CGAPipeline v2.5.0.20. NYGC provided a VCF4.1 file for each sequenced sample calculated with GATK HaplotypeCaller, following duplicate marking with Picard v1.83, and indel re-alignment and base quality recalibration. GATK v3.1.1-g07a4bf8 was used for BAM file post-processing and variant calling. Only variants with a FILTER value of PASS were used in downstream analyses for both CGI and Illumina data. Copy number variant status was determined using Reference Coverage Profiles49. Variant frequencies were annotated using Kaviar50. For comparison of the two technologies, we used monozygotic twins sequenced using separate technologies. We observed 99.12% concordance in variant calls across technologies in 6,601 distinct loci from the GWAS catalog, while 0.21% were fully observed and

discordant. Supplementary Table 2.S11 lists the full statistics of this comparison.

2.5.6 Gut microbiome 16S rRNA sequencing.

Gut microbiome data in the form of 16S OTU (Operational Taxonomic Unit) read counts were provided by Second Genome. 250 bp paired-end MiSeq profiling of the 16S v4 region was performed as described previously51, with 50,000-150,000 reads generated per sample. 16S sequence clustering and open reference OTU picking[54] were performed using USEARCH with a proprietary strain database. Each OTU was then represented as a fraction of an individual's total microbiome composition. These OTU proportions were placed in a vendor-provided taxonomy and aggregated at the kingdom, phylum, class, order, family, genus, and species levels (Supplementary Tables 2.S1 and 2.S12). α-diversity[55], a measure of the number of OTUs observed as well as the evenness of their distributions, was calculated as the within-sample Shannon diversity index:

$$H_j = -\sum_i p_{ij} \ln(p_{ij})$$

where p_{ij} is the relative abundance of OTU i in sample j.

Second Genome performed our microbiome OTU picking using their proprietary strain database. Many microbiome studies are performed using OTU picking against the publicly available Greengenes database, but Second Genome recommended that we use their curated database. Their database is specifically customized for microbes that exist in the human gut, whereas the Greengenes database spans a broad range of microbes, for example, soil and water microbes and those found in other organisms. The proprietary strain database used for microbiome analyses can be downloaded using the following URL: `http://secondgenome.com/solutions/resources/data-analysis-tools/strain-select/`. We used α-diversity to assess the degree to which each participant's microbiome composition resembled itself over time (Supplementary Fig. 2.S8). In nearly all cases, individuals' microbiome composition was more similar to their previous sample than to other individuals'. For these inter-individual

comparisons, representative sequences were aligned using PyNAST 1.2.2 ([56]l) via QIIME 1.9.1 (ref. [57]) with the Greengenes[58] 85% OTU representative sequences as a template. The alignment was filtered to remove high entropy positions using the Lane mask[59]. A phylogeny was reconstructed using FastTree 2.1.7. Unweighted UniFrac distances[60–62] were computed on the table using QIIME. scikit-bio 0.2.3 (`http://scikit-bio.org`) was used in a custom Jupyter Notebook61 with matplotlib[63] and seaborn to process the distance matrix. Specifically, for each sample, the distance between it and the participant's successive time point was determined (the red points in Supplementary Fig.2.S8). All the distances from that sample to all other samples at the successive time point were then retrieved (the box-whisker plots in Supplementary Fig.2.S8). Subsequent statistics were computed using SciPy 0.17.0.

2.5.7 Metabolomics.

Metabolon Inc. conducted the metabolomics assays on participant plasma samples at three time points for each participant throughout the course of the study. Metabolon Inc. generated the data using their DiscoveryHD4 platform in addition to their Fatty Acid Metabolism (FAME) panel that uses a combination of ultra-high-performance liquid chromatography with tandem mass spectrometry (MS) and gas chromatography (GC) in the identification of metabolites and fatty acids. The metabolite values were reported relative to their concentrations among all participants, except for lipids that were measured via GC-FID, which were reported as molar percentages of each participant's total fatty acids. For analysis, the metabolomics data were median-scaled, such that the median value for each metabolite was one, and values that fell beneath the range of detection were imputed to be the minimum observed value. This scaling was performed across all samples. All time points were run as a single batch. Counts of metabolites detected using each technology are listed in Supplementary Table 2.S13. See Supplementary Table 2.S1 for all metabolites detected.

2.5.8 Olink proximity extension assays.

Protein levels in plasma were determined by Proximity Extension Assays (PEA) using two Olink (Uppsala, Sweden) Proseek Multiplex 96 x 96 kits and quantified by real-time PCR using the Fluidigm (South San Francisco, California) BioMark HD system. Each kit provides a microtiter plate for measuring 92 protein biomarkers in 90 samples. Each well contains 96 pairs of DNA-labeled antibody probes. When a matched pair of probes bind to their target protein, their DNA labels are brought into close proximity and a PCR target sequence is formed by a proximity-dependent DNA polymerization. One plate contains 96 wells for processing 90 samples, 3 positive controls, and 3 negative controls to determine the lower detection limit. Each sample is also spiked with four controls to monitor variation in the three steps of the PEA process. Two non-human antigens serve as incubation controls, one DNA-labeled antibody serves as an extension control, and an oligonucleotide serves as a detection control.

The Proseek cardiovascular (CVD I) and inflammation (Inflammation I) panels target 158 different proteins with 19 overlapping measurements. Plasma samples from 80 subjects drawn at three intervals were assayed. One sample was assayed in triplicate on all plates and additional samples were replicated for a total of 270 multiplex cardiovascular and 270 multiplex inflammation assays. A total of 41,085 data points were collected. Assays were run according to the manufacturer's instructions. In short, 1 μl of each sample was incubated with the antibody probes at 4 $^\circ$ C overnight. After binding, the extension mix was added and the products were extended and amplified using 17 cycles of PCR (Applied Biosystems 9700, Life Technologies, Carlsbad, California). Next, 2.8 μl of each PCR product was added to the detection mix and loaded into the sample wells of a Fluidigm 96.96 Dynamic Array plate (Fluidigm Corporation) while kit primers were loaded into the primer wells. The Dynamic Array was primed in a Fluidigm HX IFC controller and then loaded into the Fluidigm BioMark imaging thermocycler for quantitative PCR. Quantification cycle (Cq) values for each measurement were determined using Fluidigm's Real-Time PCR Analysis software and

BiomarkDataCollection version 4.1.3. Data were normalized using the extension positive control and the negative control Cq values. The limit of detection was defined as three times the s.d. of the negative controls. See Supplementary Table 2.S1 for all proteins detected using Olink proximity extension assays.

2.5.9 Selected reaction monitoring (SRM) analysis.

SRM assay and method development. SRM assays were developed for 200 peptides representing 100 proteins. See Supplementary Table 2.S1 for all SRM peptides. For each peptide sequence the heavy-isotope-labeled analog was synthesized (PEPotec SRM library Grade 1, Thermo-Fisher Scientific, Huntsville, AL) with cysteine residues carbamidomethylated and the C-terminal arginine as R[13C6, 15N4] or lysine as K[13C6, 15N2] to allow for relative quantification. The 200 synthetic peptides were individually analyzed on a 6530 accurate-mass Q-TOF liquid chromatography mass spectrometry (LC-MS) system (Agilent Technologies, Santa Clara, CA) using a ProtID-Chip-150 (II) (Agilent Technologies, Santa Clara, CA) to verify and confirm successful peptide synthesis. The 200 peptides were pooled as internal standard. Multiplexed SRM assays were established with the human SRMAtlas63 (www.srmatlas.org) and the synthetic peptides on a 6460 QQQ MS system equipped with Jet Stream ESI technology and a 1290 Series UHPLC (Agilent Technologies, Santa Clara, CA). SRM assays were optimized with regard to sensitivity and specificity, and with the aim to target 200 peptides in a single analysis. 1,200 transitions were determined, 3 transitions to target each light endogenous peptide and 3 transitions to target each isotope-labeled heavy peptide, and peptides separated on a reversed phase column (Zorbax SB-C18, 50 mm x 2.1 mm I.D., 1.8 μm dp, Agilent Technologies, Santa Clara, CA) using a gradient from 3% to 30.5% acetonitrile/0.1% formic acid/water over 55 min at a flow rate of 0.2 mL/min. Data were acquired in dynamic MRM mode with a fixed cycle time of 2,500 ms and a minimum dwell time of 10 ms.

2.5.10 Plasma sample preparation.

Plasma samples were thawed on ice and centrifuged for 10 min at 14,000 r.p.m. to separate tissue debris or a lipid layer. 110 μL plasma were depleted from the 14 most abundant plasma proteins using the multiple affinity removal system (MARS Hu-14, 4.6 x 100 mm, Agilent Technologies, Santa Clara, CA) according to the manufacturer's protocol. The depleted fraction was collected in 1.25 mL of MARS Hu-14 Buffer A and denatured by adding 600 mg urea to 8 M final concentration. Samples were reduced with 5 mM dithiothreitol for 30 min at 55 C, alkylated with 14 mM iodoacetamide for 30 min at room temperature in darkness and desalted using a GE HiPrep 26/10 column (GE HealthCare Life Sciences, Pittsburgh, PA) and 1200 HPLC system (Agilent Technologies, Santa Clara, CA). The protein concentration of the desalted samples was determined by bicinchoninic acid assay (BCA) (Thermo-Fisher Scientific, San Jose, CA). An aliquot of the pooled 200 synthetic peptides was spiked into an aliquot of each plasma sample (equal protein amounts) before the digestions with trypsin (Promega, Madison, WI) at 1:50 enzyme:substrate ratio for 16 h at 37 C. Digests were dried under centrifugal vacuum evaporation (Savant, Thermo-Fisher Scientific, San Jose, CA) and reconstituted to 1 μg/μL protein concentration.

Plasma sample analysis. 20 μg of each plasma sample spiked with the 200 isotope-labeled peptides was subjected to SRM analysis using the method described above. SRM data were analyzed with Skyline[64]. SRM traces were integrated with default settings and manually inspected to verify correct peak assignment and co-elution of endogenous and isotope-labeled standard peptides. The relative peptide abundance level was reported as ratio of endogenous light to the heavy standard.

2.5.11 Quantified self-tracking.

Participants were asked to wear a Fitbit activity tracker throughout the 9-month study. Participants were offered either a Fitbit Flex (wrist) or a Fitbit One (clip-on). These Fitbit models measure activity using the number of steps an individual takes each day. The models

available at the time of the study did not measure heart rate, as current models do, resulting in inconsistent activity measurements for, for example, running versus cycling. Furthermore, the devices required manual entry and exit of 'sleep mode' for sleep tracking, for which compliance was too low to provide useful data. We required a minimum of 40 days of Fitbit usage in order to estimate the average activity for each participant; 64% of the participants met this criterion. The Fitbit device estimates user-specific 'activity calories' independently of basal metabolic rate (BMR). For all calculations, we used only the estimated 'activity calories', excluding BMR. We used these data only as a relative indicator of activity levels rather than an absolute measure of caloric burn.

2.5.12 Genomic traits.

The National Human Genome Research Institute's GWAS catalog lists results from more than 2,000 published studies comprising over 1,000 genetic traits [14]. We applied a strict filtering procedure to identify GWAS used for this study. First, we excluded studies which did not contain at least one SNP with $P < 1.0 \times 10^{-8}$ Studies which contain few SNPs are likely to produce a vector of cumulative genetic variation with low entropy, where almost all values are identical save a few. Such low entropy measurements are more likely to produce spurious correlations in our relatively small number of samples. We therefore excluded all traits associated with five or fewer SNPs. Furthermore, we required studies to have a sample size of at least 5,000 individuals. In the event that multiple studies examined the same trait, we kept the study with the largest sample size. Finally, we manually excluded traits with too-generic descriptions (e.g., 'common traits' or 'metabolic traits'), which did not provide a useful description of the purpose of the original study. The combination of these filters retained 127 genetic traits that we used for further analysis. Three common copy number variations (CNVs) were included as additional genetic features computed using Reference Coverage Profiles[65], bringing the total to 130. See Supplementary Table 2.S1 for all genetic traits computed for this manuscript.

Included in the GWAS catalog are the beta-coefficients/odds ratios as well as the P-values

for the predicted effect of each variant for that trait based on the association models from the original paper. We made two assumptions to simplify the calculation of the polygenic scores. First, we assumed that the beta-coefficients (or log odds ratios) combined in an additive manner based on the number of effect alleles present in each individual. Therefore, if a single effect allele was present we added the beta-coefficient or log odds ratio for that variant into the cumulative polygenic score. If two copies of the effect allele were present we added twice the value of the beta-coefficient or log odds ratio into the cumulative genetic effect for that individual. The second assumption was that the effects of each variant are independent of the effects of all other variants used in the model. These two simplifying assumptions allowed us to calculate the polygenic score for each trait across each individual in our study.

There are a number of pitfalls to this approach that served to temper our expectations. First, GWAS only identify variants that occur commonly enough in the population to be associated statistically with a trait. Unless one is able to genotype a substantial fraction of the human population at risk for a particular trait, rare variants will never rise above the level of noise in a GWAS. Furthermore, because they employ genotyping chips most GWAS ignore CNVs or structural variations (SVs) that may have a significant effect on genetic traits. We included as part of our study three common CNVs as additional genomic features. Finally, many GWAS are applied to cohorts of individuals from similar ancestries to improve their likelihood of discovering associated variants; it is therefore possible that results from these studies do not generalize to individuals from differing ancestral populations.

There are other analysis options available for WGS data that would be appropriate for subsequent studies with an 'N of 1' focus. For example, one could perform rare or de novo variant analysis, which identify genetic variants that are either very rare in the population or unique to an individual, respectively. GWAS focus on variants which are common enough in the population to find significant associations with quantitative traits or diseases. The interpretation of rare and de novo variants can be difficult, as each variant must be interpreted in the context of functional impact. Sequencing and phenotyping relatives

(e.g., family-based analysis) is a method to assist in interpreting the functional impact of de novo variants. Another possible analysis technique for WGS data is burden analysis, which calculates a cumulative burden score on each gene and attempts to associate these scores with phenotypes.

Coaching, charting, and compliance tracking. The P100 was designed as a prospective study that attempted to help participants modify their behavior to enhance their health throughout the 9-month period. Participants were assigned to a behavioral coach, who walked them through actionable possibilities from their data and made recommendations on lifestyle changes. These lifestyle changes were recommended in an attempt to alter markers of known clinical significance and/or compensate for genetic predispositions for which reliable published evidence is available. Each participant was eligible for one 30-min coaching session per month, though participants were not penalized or excluded from the study if they chose not to participate in the coaching sessions. Participants were also able to communicate privately and securely with the coach via a website portal created specifically for this project. Participants also received their data through the website portal. The P100 collected statistics on participation in the coaching calls and compliance with sample collection.

As previously stated, clients were offered specific coaching recommendations based on their genetics and clinically actionable data. These recommendations were customized before each call by the study clinician and coach, in consultation with the study physician. All clinical markers and recommendations were reviewed and approved by the study physician before their communication to each participant. While these recommendations were specific to each individual based on their data, they typically fell into one of several major categories, including diet, exercise, stress management, dietary supplements, or physician referral. Coaching focused on four primary health areas: cardiovascular, diabetes, inflammation, and nutrition. The clinical tests that were actively coached on are provided in Supplementary Table 2.S9. We used generalized estimating equations (GEE) to calculate the average change for each clinical laboratory test by round while controlling for the effects of age, sex, and self-reported ancestry. Coefficients, 95% confidence intervals, and P-values for all participants as well as

those who began the study out-of-range are listed in Supplementary Table 2.S10. See also Table 2.1.

Action items were recorded in each participant's chart by our behavioral coach during each coaching call. These charts were used to keep participants on track and follow standard clinical practice guidelines. Post-study we reviewed each de-identified chart in detail with our behavioral coach to extract compliance data for each recommendation. We learned a great deal about how to merge standard clinical practice (e.g., charting in free-text fields, as practiced by clinicians) with the need for automated database storage of pre-defined and enumerated recommendations. Subsequent studies will investigate specific effects of recommendations and compliance on clinical data as well as other omics data with far larger N.

2.5.13 Data preprocessing.

Each data set was transformed into comparable data vectors for statistical analysis. All measurements were mean-centered and scaled by the standard deviation of the observed measurements. The microbiome measurements were compared independently at the phylum, class, order, and family taxonomic levels. With the exception of the median-scaled metabolomics data, missing data were not imputed; participants that had a missing value were dropped from pairwise comparisons using that value. Each analyte was age- and/or sex-corrected if a trimmed mean robust regression identified a significant relationship ($P < 0.01$, unadjusted) between age and/or sex and the dependent variable. Age and sex corrections were performed independently, so it was possible for an analyte to be age corrected but not sex corrected, and vice versa. If an analyte was corrected the residuals of the model were used in place of the original observations. If no corrections were made, the original mean-centered and scaled measurements were used. See Supplementary Table 2.S14 for statistics on age and sex correction.

2.5.14 Correlation network and community analysis.

We created two different types of correlation networks: 'cross-sectional' and 'delta' correlations. Cross-sectional correlations were calculated from mean measurements of analytes calculated across all rounds (i.e., mean A is correlated with mean B across all individuals). Delta correlations were calculated on the change in analytes between rounds (i.e. the change in A for an individual between time points is correlated with the change in B, where the correlation is again calculated across all individuals). We used each pair of adjacent time points (r2-r1) and (r3-r2) to build the delta correlation network, where all such comparisons were used in the two vectors that were being compared. Therefore, each individual with three observations is represented twice for each calculated delta correlation. For example, while the cross-sectional correlation network was created by correlating vectors of maximum length N = 108, the delta correlation network was created by correlating vectors of maximum length N = 216. Our reasoning is that each pair of adjacent time points is an independent observation of a potential correlation in time, even though they are not drawn from a completely independent set of individuals. For each pairwise set of data (e.g., clinical tests versus proteomics, clinical tests versus metabolomics, etc.), each measurement from the first data set was correlated with every measurement from the second data set using Spearman's ρ. P-values were adjusted for multiple hypothesis testing using the method of Benjamini and Hochberg[15]; we chose an adjusted P-value (padj) cutoff of 0.05 as our significance level. Only inter-omic correlations were used for community analysis. Both inter-omic and intra-omic (e.g., metabolomics versus metabolomics) cross-sectional and delta correlations are reported in Supplementary Tables 2.S2 and 2.S5 and visualized in Figure 2.2 and Supplementary Figure2.S1. We assessed reproducibility of duplicate measurements across two clinical laboratories (Supplementary Fig. 2.S9). As correlations between repeat measurements do not represent physiologically relevant information, they are not included in our network or subsequent analysis. We note the poor correlation between LDL particle number measured with both Genova and Quest as an explanation for the discordant laboratory

results for this analyte reported in Table 2.1 and Supplementary Table 2.S10.

We performed community analysis using the method of Girvan and Newman[16]. This method involves iteratively calculating edge betweenness centrality on a network: the number of weighted shortest paths from all vertices to all other vertices that pass over that edge. After each iteration, the edge(s) with the highest betweenness centrality were removed, and the process was repeated until only individual nodes remain. All communities can also be dynamically visualized in Cytoscape[18] (Supplementary Datasets 1 and 2).

Community analysis forms a dendrogram that can be analyzed at multiple hierarchical levels. For this manuscript we analyzed our network at a cut level determined using an unbiased method, the modularity of the community structure[17]. Briefly, modularity of community structure corresponds to an arrangement of edges that is statistically improbable when compared to an equivalent network with edges placed at random. At every iteration of the community analysis, we computed the modularity, and analyzed the communities at the iteration which maximized this quantity. A visualization of community modularity vs. iteration is shown in Supplementary Figure 2.S2.

2.5.15 OLS regression tests.

To test for heteroscedasticity in our HOMA-IR regression model, we fit the model using White's heteroscedasticity-consistent estimator (HCE), and the results were consistent with those reported in the manuscript: gamma-glutamyltyrosine was still significantly more predictive ($P = 2.3 \times 10^4$) of HOMA-IR than GGT ($P = 0.02$). To test for the effects of outliers, we fit a robust regression model and again the results were consistent with those reported in the manuscript. To test for multicollinearity, we calculated the variance inflation factors (VIF) for each predictor. The maximum VIF was 1.7, indicating a low amount of correlation between the predictors of the model.

2.5.16 Statistical analyses.

All data types used in the cross-sectional and delta correlation networks were normalized as described in their respective method sections. Additionally, where we were able to identify a significant effect ($P < 0.01$, unadjusted) with age and/or sex using trimmed mean robust regression we used the residuals as the comparison value. These adjustments were performed independently for age and sex. We report the calculated P-values for age and sex with every variable, as well as whether the variable was age and/or sex adjusted in Supplementary Table 2.S14. All transformations were performed with the Python Statsmodels package (v0.6) with robust linear models using the trimmed mean norm. Unadjusted analytes were compared using the original mean-centered and scaled measurements.

We created two different types of correlation networks: 'cross-sectional' and 'delta' correlations. Cross-sectional correlations were calculated from mean measurements of analytes calculated across all rounds (i.e., mean A is correlated with mean B across all individuals). Delta correlations were calculated on the change in analytes between rounds (i.e., the change in A for an individual between time points is correlated with the change in B, where the correlation is again calculated across all individuals). We used each pair of adjacent time points (r2-r1) and (r3-r2) to build the delta correlation network, where all such comparisons were used in the two vectors that were being compared. Therefore, each individual with three observations is represented twice for each calculated delta correlation. For example, while the cross-sectional correlation network was created by correlating vectors of maximum length $N = 108$, the delta correlation network was created by correlating vectors of maximum length $N = 216$. For each pairwise set of data (e.g., clinical tests versus proteomics, clinical tests versus metabolomics, etc.), each measurement from the first data set was correlated with every measurement from the second data set using Spearman's ρ. P-values were adjusted for multiple hypothesis testing using the method of Benjamini and Hochberg[15]; we chose an adjusted P-value (padj) cutoff of 0.05 as our significance level. Only inter-omic correlations were used for community analysis.

We performed community analysis using the method of Girvan and Newman[16]. This method involves iteratively calculating edge betweenness centrality on a network: the number of weighted shortest paths from all vertices to all other vertices that pass over that edge. After each iteration, the edge(s) with the highest betweenness centrality were removed, and the process was repeated until only individual nodes remain.

For this manuscript we analyzed our network at a cut level determined using an unbiased method, the modularity of the community structure[17]. At every iteration of the community analysis, we computed the modularity, and analyzed the communities at the iteration that maximized this quantity (Supplementary Fig. 2.S2).

OLS regression on the dependent variable HOMA-IR, with regressors including sex, GGT, gamma-glutamyltyrosine, age, and body mass index was performed (Supplementary Table 2.S7). To test for heteroscedasticity in our HOMA-IR regression model, we fit the model using White's heteroscedasticity-consistent estimator (HCE), testing for the effects of outliers using a robust regression model and testing for multicollinearity by calculating the variance inflation factors (VIF) for each predictor.

Generalized estimating equations (GEE) were used to calculate the average change in each clinical analyte by round while controlling for the effects of age, sex, and self-reported ancestry (Table 2.1 and Supplementary Table 2.S10). An independence working correlation structure was used in the GEE.

2.5.17 Software packages.

NYGC used BWA v0.7.8-r455 to align sequences. Complete Genomics CGAPipeline v2.5.0.20. NYGC provided a VCF4.1 file for each sequenced sample calculated with GATK HaplotypeCaller, following duplicate marking with Picard v1.83, and indel realignment and base quality recalibration. GATK v3.1.1-g07a4bf8 was used for BAM file post-processing and variant calling. Copy number variant status was determined using Reference Coverage Profiles[65]. Variant frequencies were annotated using Kaviar[66].

OTU picking for the microbiome results used for the correlation networks was performed

using USEARCH with a proprietary strain database, which can be downloaded at: http://secondgenome.com/solutions/resources/data-analysis-tools/strain-select/. Inter-individual microbiome comparisons were performed using PyNAST 1.2.2 (ref. [54]) via QIIME 1.9.1 (ref. [57]) with the Greengenes [58] 85% OTU representative sequences as a template. The alignment was filtered to remove high entropy positions using the Lane mask[59]. A phylogeny was reconstructed using FastTree 2.1.7. Unweighted UniFrac distances[60–62] were computed on the table using QIIME. scikit-bio 0.2.3 (http://scikit-bio.org) was used in a custom Jupyter Notebook[67] with matplotlib[63] and seaborn to process the distance matrix.

Olink Cq values for each measurement were determined using Fluidigm?s Real-Time PCR Analysis software and BiomarkDataCollection version 4.1.3.

Multiplexed SRM assays were established with the human SRMAtlas[68] (http://www.srmatlas.org/) and analyzed with Skyline[64].

All pairwise statistical tests (Spearman) were performed using the Python scipy.stats package (v0.14). All regression models (OLS regressions and GEE) were fit using the Python Statsmodels package (v0.6)[69]. An independence working correlation structure was used for GEE. Variance inflation factors were calculated using Python Statsmodels (v0.6) package. Correlation network P-values were adjusted for multiple hypothesis using the Benjamini-Hochberg [15] method via the Python Statsmodels (v0.6) package for each inter-datatype comparison. Community analysis and modularity calculations were performed in Python with the NetworkX66 package, the python-louvain package (v0.3), and custom code. All custom code used in this manuscript can be downloaded from Github using the link below and is also available as Supplementary Code. The Github version used for the manuscript is 'v-release'.

https://github.com/trueprint/p100-network-code/tree/v-release

2.5.18 ACMG incidental finding reporting.

Genes listed in the American College of Medical Genetics and Genomics (ACMG) recommendations for incidental findings[70] were tabulated along with associated reporting requirements (known pathogenic and expected pathogenic) and inheritance pattern (autosomal recessive, autosomal dominant, X-linked). The ClinVar database was obtained from the NCBI (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/). Records from ClinVar were filtered to overlap with genes reported in the ACMG incidental finding list by gene symbol. The remaining records were further filtered to only include variants where at least one submitter indicated a variant was pathogenic or likely pathogenic, and no submitter indicated a variant was benign or likely benign (i.e., no conflicting interpretations of pathogenicity). P100 participants' VCF files were filtered for variant calls intersecting the chromosome and position of these filtered ClinVar records. Only calls with a FILTER value of PASS were included. If the record was a SNP, the alternate allele of each variant call was compared against the effect allele of the ClinVar record to ensure a match. Matching variant calls were further filtered to match inheritance pattern (e.g., if a disorder was reported as autosomal recessive, the reported call must be homozygous alternate). 12 variant calls were identified using this protocol. 8 of these 12 were the same variant (rs34677591), and were discarded as likely false positives. One of the remaining four was a small indel, and manual curation of the alignments determined that it was likely a false positive due to misalignment. The remaining three were manually curated and determined to be valid calls.

2.5.19 Running the custom code.

This code is meant to be run in a Jupyter[67] notebook that has the scipy stack installed. We recommend using the datascience docker image created by the Jupyter group at

`https://github.com/jupyter/docker-stacks/tree/master/datascience-notebook`

The raw data are available from dbGap in a compressed tar.gz file and should be extracted to the same directory containing the code. We recommend downloading the Jupyter docker

image using docker pull jupyter/datascience-notebook on a machine with docker installed. This is not required, but is recommended and all instructions will be based on the use of this image. An example shell script is provided with the custom code (startup-notebook.sh). The startup command is:

```
docker run -d -p <SOME LOCAL PORT>:8888 -e USE_HTTPS=yes -e GRANT_SUDO=yes -v \
   <LOCAL PATH TO p100-network-code>:<ROOT PATH OF NOTEBOOKS ON JUPYTER>/p100-network-co
   -v <LOCAL PATH TO UNZIPPED data>:<ROOT PATH OF NOTEBOOKS ON JUPYTER>/data \
   jupyter/datascience-notebook
```

Then, navigate in your browser to https://<yoururl>:<SOMELOCALPORT>. For example, if you ran this on your localhost, with SOME LOCAL PORT = 8888, then you would navigate to https://localhost:8888. Note: it will give you a warning about an invalid certificate. The default password for datascience-notebook is empty, i.e., just hit ENTER. The notebooks provided are: Generate correlation network.ipynb which generates a correlation network of all data for the P100 project, Community Generation-DELTA.ipynb which generates a correlation network for change in data measurements for the P100 project, Community Generation.ipynb which performs community analysis using the intraomic correlation network, Community Generation-DELTA.ipynb which performs community analysis using the intraomic delta(change) correlation network and GEE longitudinal clinical changes.ipynb, which uses GEE (generalized estimating equations) to calculate average change over the course of the study in clinically relevant biomarkers. We also provide the notebook Convert csvs to pickles.ipynb, which converts raw CSV files and associated meta data in JSON from the csv folder of the data into Python pickles appropriate to the currently installed pandas version.

2.5.20 Data availability.

All raw data collected as part of the P100 are available from dbGaP with accession ID phs001363.v1.p1 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_

id=phs001363.v1.p100-6).

2.6 Acknowledgments

The results presented in this chapter have been released as a paper in collaboration with my co-authors [4]: Nathan D Price, Andrew T Magis, Gustavo Glusman, Roie Levy, Christopher Lausted, Daniel T McDonald, Ulrike Kusebauch, Christopher L Moss, Yong Zhou, Shizhen Qin, Robert L Moritz, Kristin Brogaard, Gilbert S Omenn, Jennifer C Lovejoy, Leroy Hood.

2.S Supplementary Materials

2.S.1 Supplementary Figures

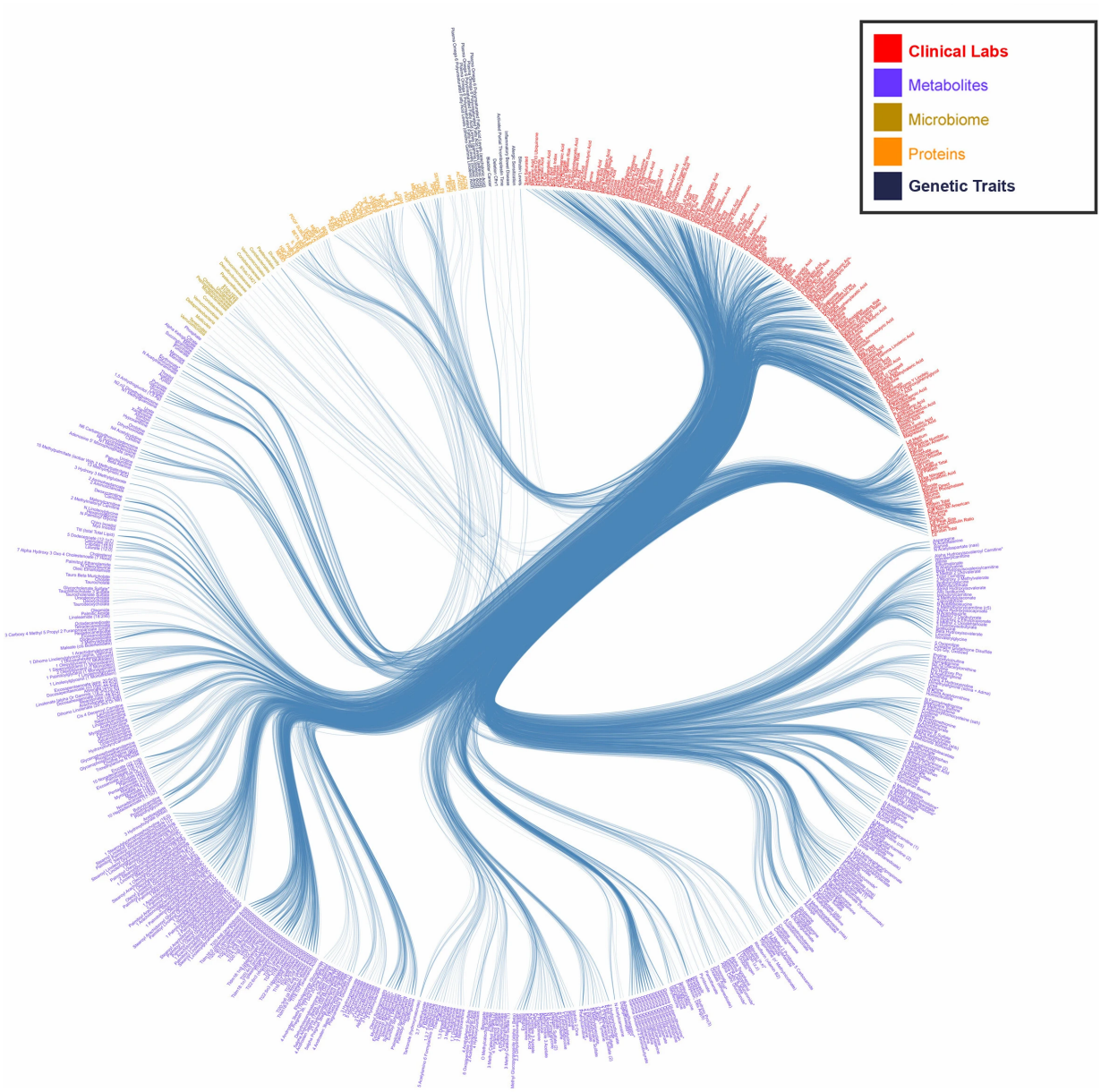


Figure 2.S1: Full inter-omic cross-sectional correlation network

Statistically-significant inter-omic cross-sectional Spearman correlations ($p_{adj} < 0.05$) between all datasets collected in our cohort.

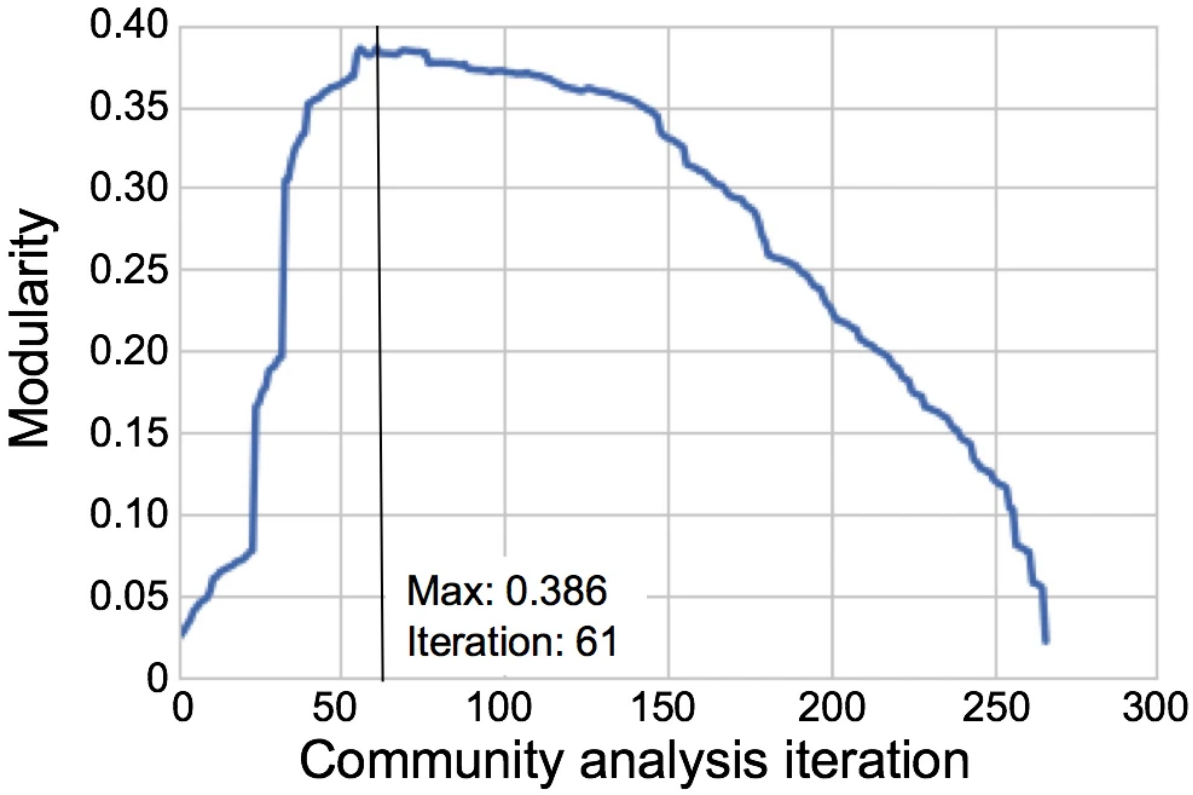


Figure 2.S2: Modularity vs. community analysis iteration

The maximum modularity observed in our inter-omic cross-sectional community analysis was 0.386 at iteration 61 of community pruning. There were 267 total iterations of community analysis.

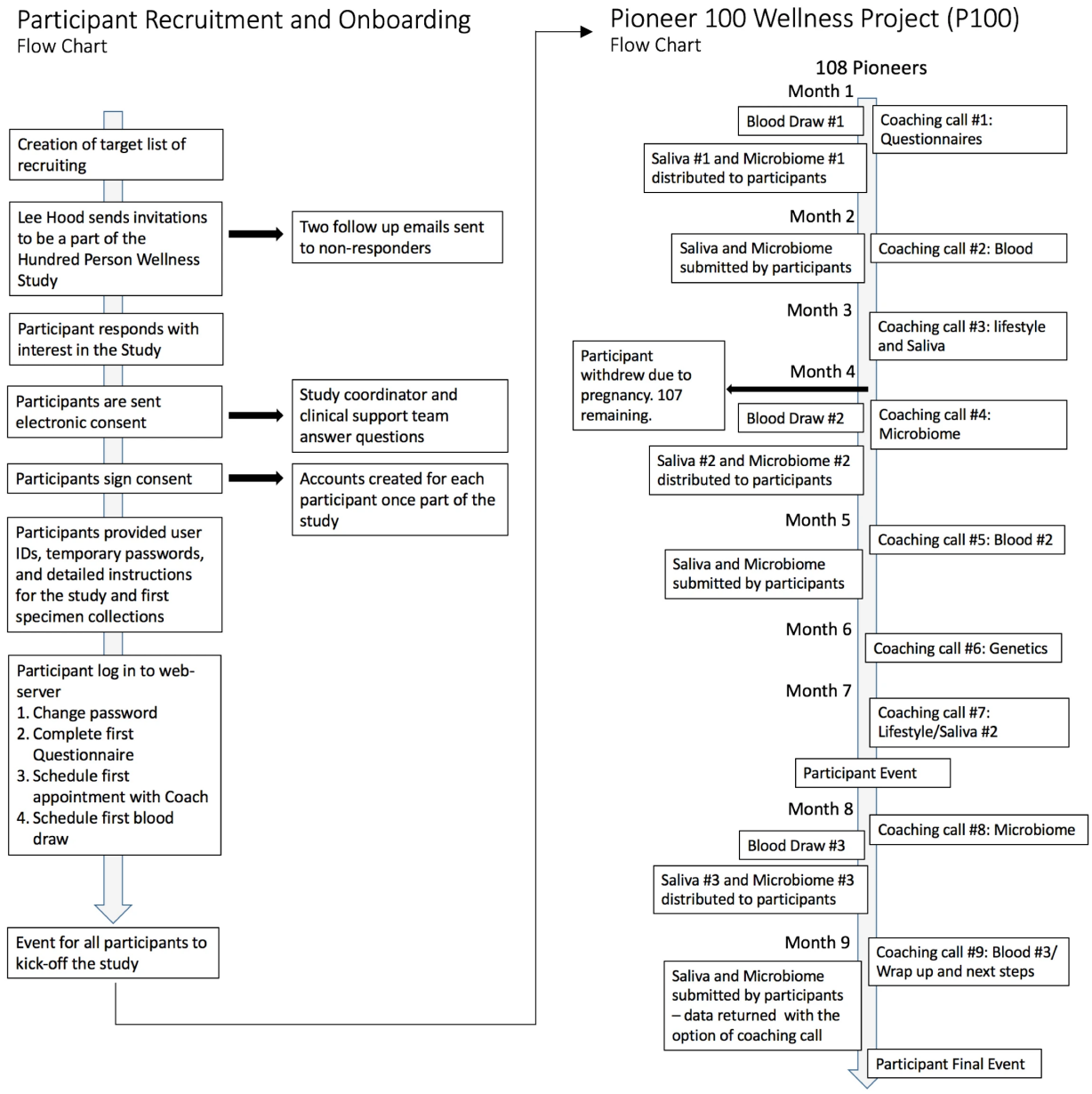


Figure 2.S3: Recruitment, onboarding, and other important events in the P100

Flowchart of important events in the P100, including recruitment, onboarding, withdrawals, data collection, coaching calls, and events.

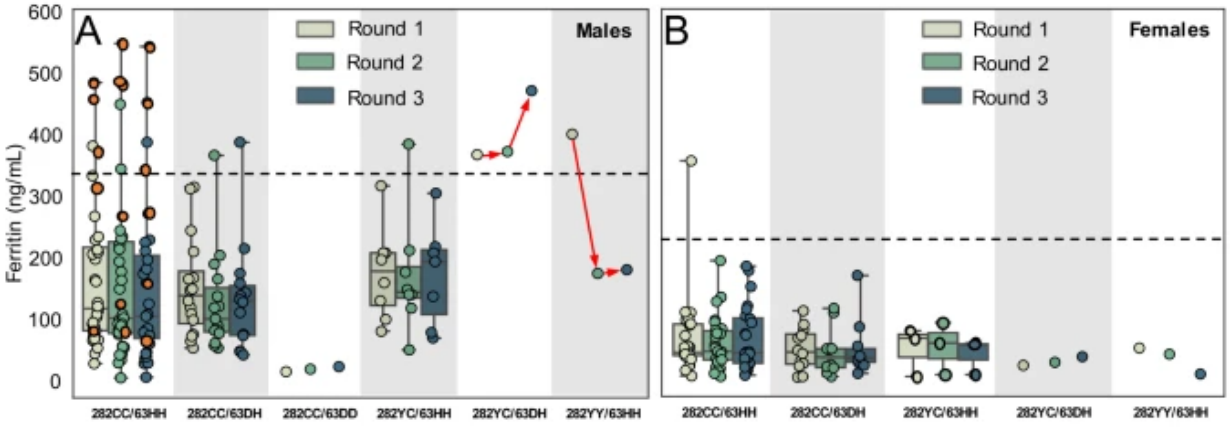


Figure 2.S4: Genetic risk factors for hemochromatosis and ferritin levels

Boxplots for ferritin levels of male (A) and female (B) participants by round in the P100. Only one male in our study was homozygous for 282YY and was diagnosed with hemochromatosis after physician referral. Changes in ferritin levels are shown by the red arrows. A second male who was heterozygous for both risk factors (282YC/63DH) did not receive therapeutic phlebotomy, and ferritin levels increased. Six other males presented at baseline (round 1) with elevated ferritin levels but neither of these genetic risk factors; they were referred to their physician for monitoring. Four of the six were of self-reported Asian ancestry (orange dots).

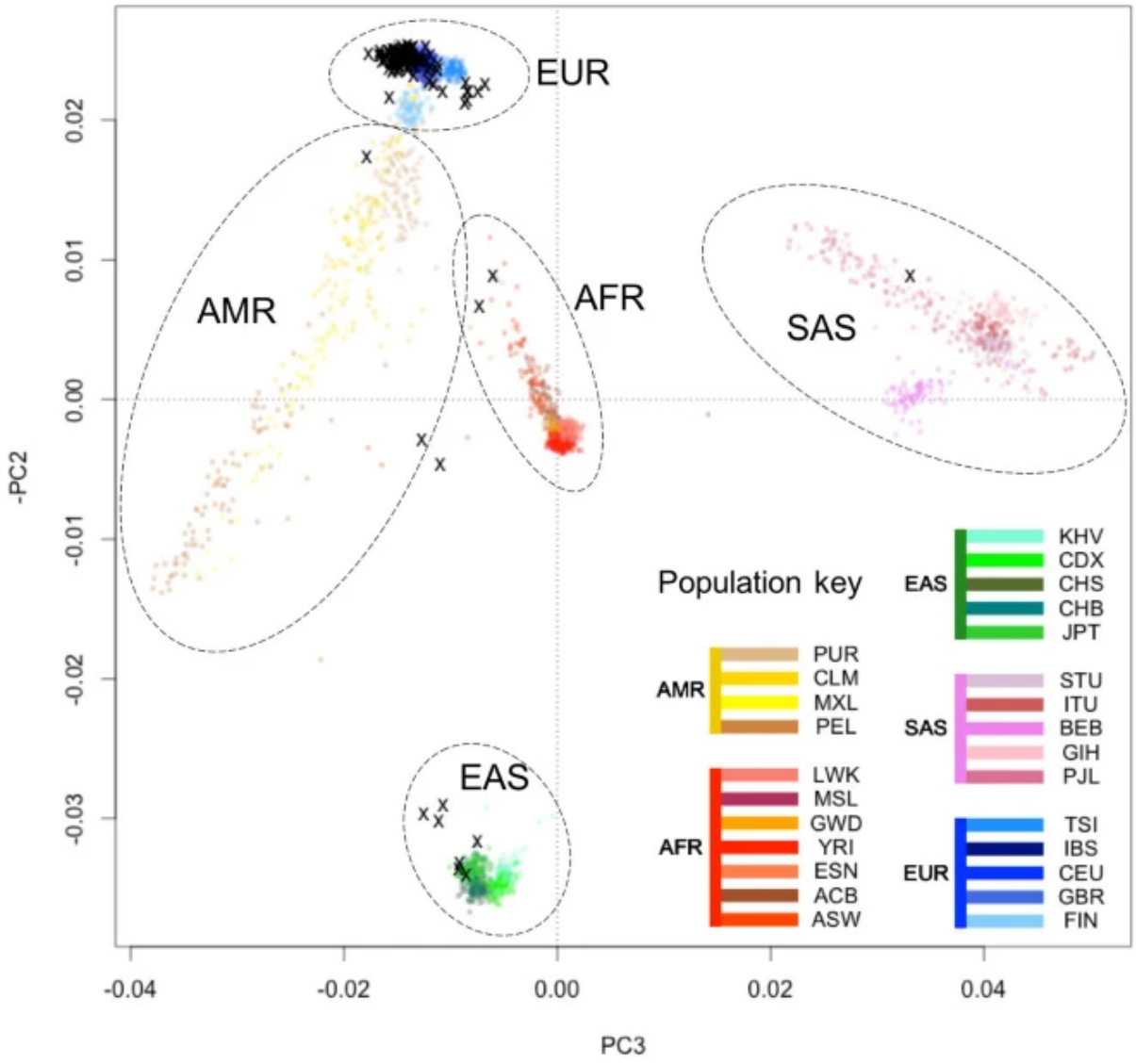


Figure 2.S5: Population distribution of the 108 Pioneers (PC2 vs PC3)

PCA using a sample of 250,000 common SNPs. Translucent colored points represent the 2504 individuals in the Thousand Genomes Project, phase 3, color-coded by population. Black points represent the 108 Pioneers. AFR, EUR, SAS, EAS, and AMR represent the continental populations: African, European, South Asian, East Asian and Admixed Americans, respectively.

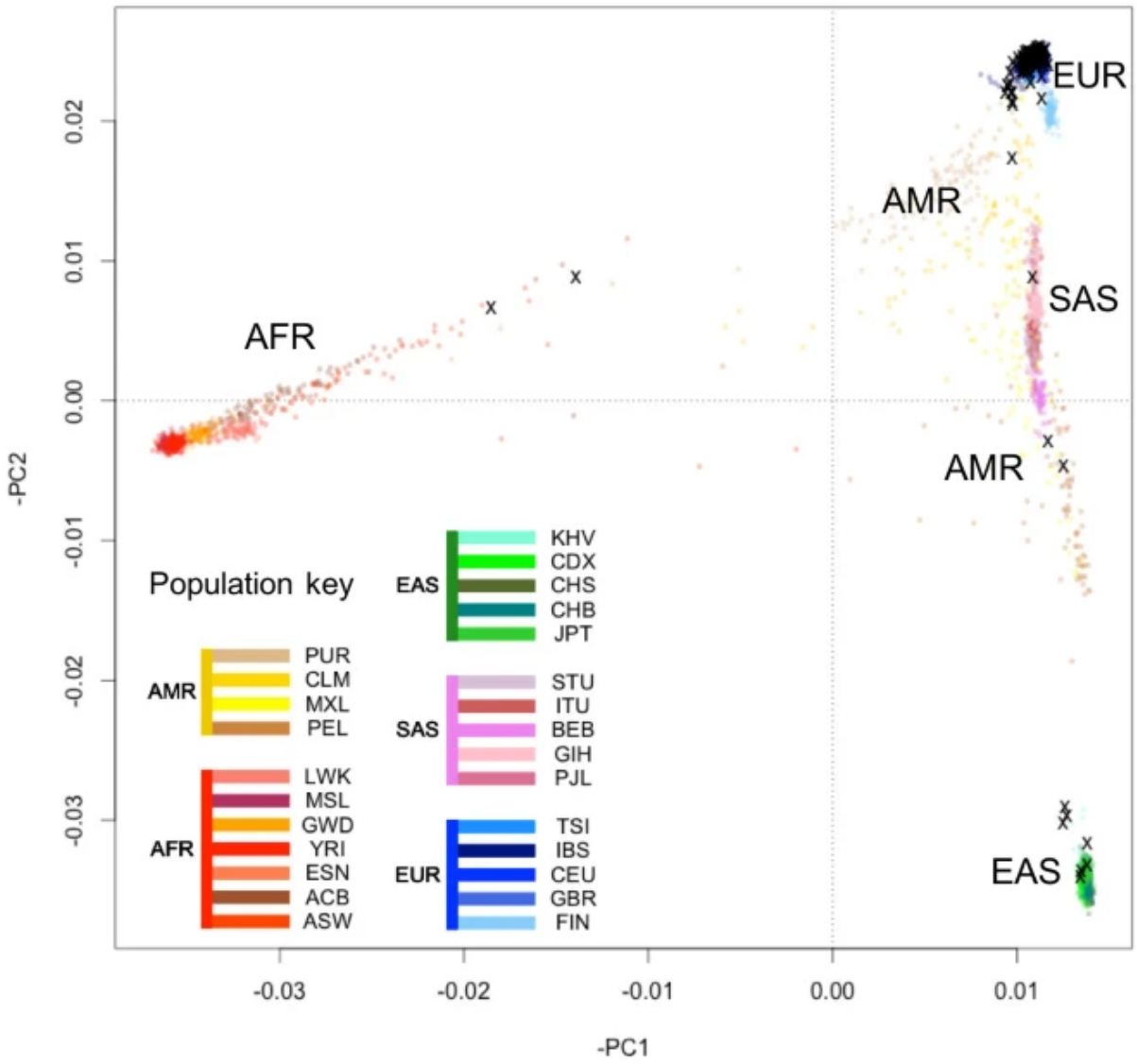


Figure 2.S6: Population distribution of the 108 Pioneers (PC1 vs PC2)

PCA using a sample of 250,000 common SNPs. Translucent colored points represent the 2504 individuals in the Thousand Genomes Project, phase 3, color-coded by population. Black points represent the 108 Pioneers. AFR, EUR, SAS, EAS, and AMR represent the continental populations: African, European, South Asian, East Asian and Admixed Americans, respectively.

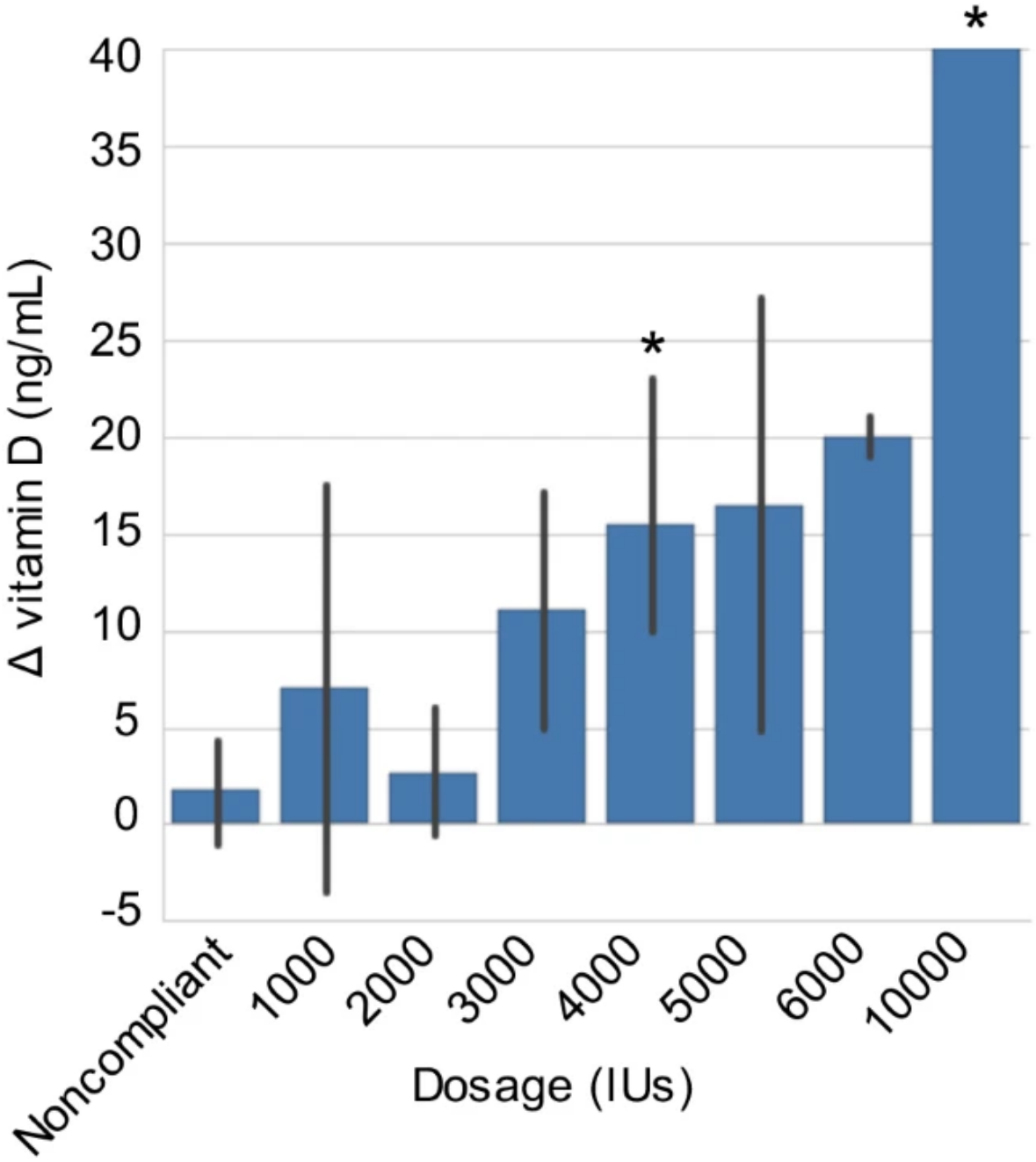


Figure 2.S7: Dose-dependent effects of vitamin D supplementation

A common intervention for our participants was vitamin D supplementation. The Institute of Medicine has recommended a minimum 25-hydroxyvitamin D level of 20 ng/mL, while the Endocrine Society recommends a minimum level of 30 ng/mL. Shown here are the dose-dependent effects of supplementation on 25-hydroxyvitamin D levels from round 1 to round 2, with individuals taking less than 3000 IUs/day exhibiting relatively little gains in 25-hydroxyvitamin D levels. Individuals that were noncompliant with the recommendations (N=13) made no gains in 25-hydroxyvitamin D levels. A one-way ANOVA yields $p=0.004$ that a significant difference exists between one of the groups. Significant differences with noncompliant participants at the $p<0.05$ level are indicated with asterisks, as determined by Tukey's range test. Individuals with low blood levels of 25-hydroxyvitamin D were recommended doses between 1000 and 5000 IU. If over time blood levels did not increase at the highest doses, individuals were referred to their physician for evaluation and, if appropriate, higher doses.

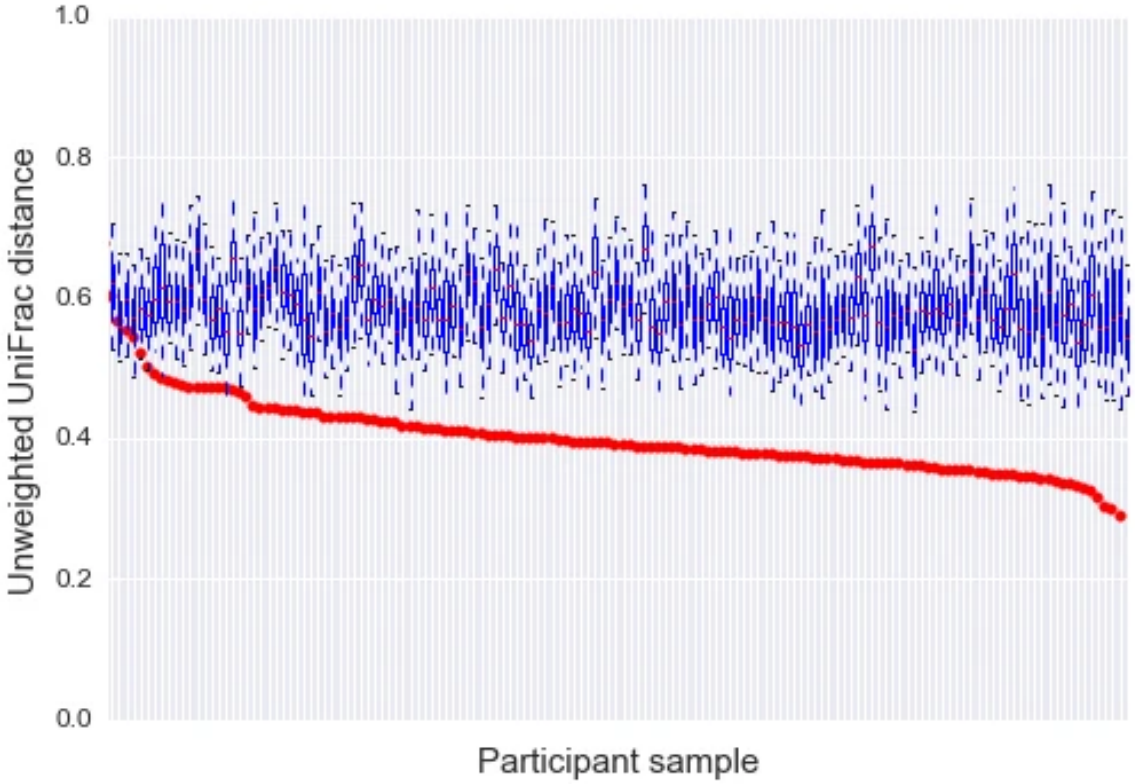


Figure 2.S8: Gut microbiome stability over nine months

Participant microbiomes tend to resemble themselves over time. Plotted in red is the unweighted UniFrac distance between consecutive microbiome samples for all participants. The blue box-and-whisker plots represent the distance distribution between each sample and all others in the same time points. In 97% of cases, an individual's cross-timepoint distance is lower than the median inter-individual distance.

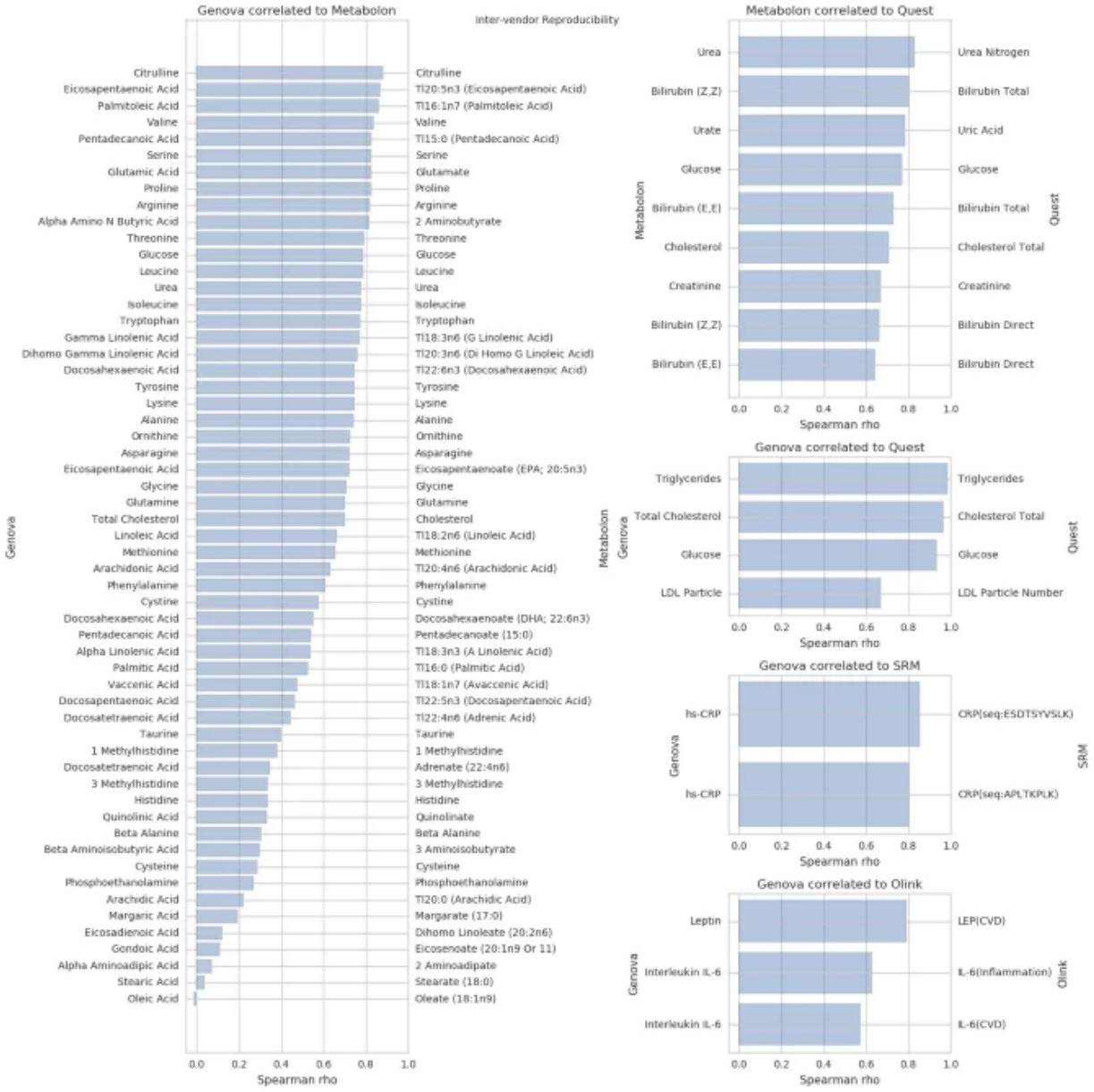


Figure 2.S9: Correlation across different vendors

Most clinical laboratory measurements were assayed by only one of the lab vendors (Quest or Genova) but certain measurements were measured by both due to overlaps in the standard analysis panels. Additionally, some analytes from the metabolomics and proteomics data were also measured by the clinical labs. This figure shows a comparison of these duplicated analytes. For example, triglycerides, total cholesterol, and fasting glucose show high correlation between clinical lab vendors, while LDL particle number is less correlated. The correlations represented in this figure have been removed from our correlation networks.

2.S.2 Supplementary Tables

Table 2.S1: All analytes measured in the P100 (XLSX 317 kb)

Available for download at https://static-content.springer.com/%3A10.1038%2Fnbt.3870/MediaObjects/41587_2017_BFnbt3870_MOESM30_ESM.xlsx

Table 2.S2: Complete inter-omic correlation network for cross-sectional correlations (XLSX 247 kb)

Available for download at https://static-content.springer.com/esm/art%3A10.1038%2Fnbt.3870/MediaObjects/41587_2017_BFnbt3870_MOESM31_ESM.xlsx

Table 2.S3: Complete intra- and inter-omic correlation network for cross-sectional correlations (XLSX 1240 kb)

Available for download at https://static-content.springer.com/esm/art%3A10.1038%2Fnbt.3870/MediaObjects/41587_2017_BFnbt3870_MOESM32_ESM.xlsx

Table 2.S4: Complete inter-omic correlation network for delta correlations (XLSX 191 kb)

Available for download at https://static-content.springer.com/esm/art%3A10.1038%2Fnbt.3870/MediaObjects/41587_2017_BFnbt3870_MOESM33_ESM.xlsx

Table 2.S5: Complete intra- and inter-omic correlation network for delta correlations (XLSX 1834 kb)

Available for download at https://static-content.springer.com/esm/art%3A10.1038%2Fnbt.3870/MediaObjects/41587_2017_BFnbt3870_MOESM34_ESM.xlsx

Table 2.S6: Proteins present in the serotonin community

symbol	name	uniprot
ACTA2	actin, alpha 2, smooth muscle, aorta	P62736
EGF	epidermal growth factor	P01133
SIRT2	sirtuin 2	Q8IXJ6
PPBP	pro-platelet basic protein	P02775
CD40LG	CD40 ligand	P29965
AXIN1	axin 1	O15169
HSPB1	heat shock 27kDa protein 1	P04792
STAMBP	STAM binding protein	O95630
EIF4EBP1	eukaryotic translation initiation factor 4E binding protein 1	Q13541
PDGFB	platelet-derived growth factor beta polypeptide	P01127
IL7	interleukin 7	P13232
IKBKG	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma	Q9Y6K9

Table 2.S7: OLS regression on the dependent variable HOMA-IR.

regressor	coefficient	std error	t	p-value	95% confidence
Body Mass Index	1.5423	0.318	4.844	4.5e-6	0.911, 2.174
gamma-glutamyltyrosine	1.2984	0.267	4.859	4.3e-6	0.768, 1.828
GGT	0.1233	0.073	1.696	0.093	-0.021, 0.267
Age	-0.2744	0.22	-1.247	0.215	-0.711, 0.162
Sex(Male)	-0.0114	0.01	-1.169	0.245	-0.031, 0.008

OLS regression on the dependent variable HOMA-IR. Regressors include sex, GGT, gamma-glutamyltyrosine, age, and body mass index. Body mass index and gamma-glutamyltyrosine are significant regressors in the model, while GGT is marginally significant. N=108 for this regression.

Table 2.S8: Polygenic score quantitative traits tested in the P100 (XLSX 51 kb)

Available for download at https://static-content.springer.com/esm/art%3A10.1038%2Fnbt.3870/MediaObjects/41587_2017_BFnbt3870_MOESM35_ESM.xlsx

Table 2.S9: Clinical laboratory tests used to analyze changes in the health areas targeted by coaching

Cardiovascular	Diabetes	Inflammation	Nutrition
Total cholesterol	Fasting glucose	Interleukin-6	Vitamin D
LDL cholesterol	HbA1c	Interleukin-8	Glutathione
HDL cholesterol	Insulin	TNF-alpha	Ferritin
Triglycerides	HOMA-IR	hs-CRP	Methylmalonic acid
LDL pattern (A or B)			Selenium
LDL particle number			Copper
LDL medium particle number			Manganese
LDL small particle number			Arachidonic acid
HDL large particle number			EPA

Table 2.S10: Longitudinal analysis of clinical changes by round

Clinical Laboratory Test			OOR Baseline				All			
Quadrant	Name	Unit	N	Coef.	95% conf.	Pvalue	N	Coef.	95% conf.	Pvalue
Nutrition	Vitamin D	ng/mL	95	7.2	[5.8, 8.5]	7.1E-25	108	6.5	[5.2, 7.9]	2.5E-22
Nutrition	Mercury	mcg/g	81	-0.002	[-0.003, -0.0015]	8.9E-09	108	-0.0015	[-0.002, -0.0008]	6.7E-05
Diabetes	HbA1c	%	52	-0.085	[-0.12, -0.047]	9.2E-06	108	-0.047	[-0.071, -0.022]	1.6E-04
Cardiovascular	LDL particle number (Quest)	nmol/L	30	130	[63, 190]	9.3E-05	108	78	[43, 110]	1.2E-05
Nutrition	Methylmalonic acid (Genova)	mmol/mol creat	3	-0.49	[-0.75, -0.23]	2.1E-04	108	0.012	[-0.029, 0.053]	5.6E-01
Cardiovascular	LDL pattern	A(0), B(1)	28	-0.16	[-0.25, -0.07]	4.8E-04	108	-0.01	[-0.047, 0.028]	6.1E-01
Inflammation	Interleukin-8	pg/mL	10	-6.1	[-9.6, -2.6]	5.9E-04	108	-0.59	[-1.4, 0.24]	1.6E-01
Cardiovascular	Total cholesterol (Quest)	mg/dL	48	-6.4	[-10.0, -2.7]	7.2E-04	108	-0.76	[-3.5, 1.9]	5.8E-01
Cardiovascular	LDL cholesterol	mg/dL	57	-4.8	[-7.6, -2.0]	8.8E-04	107	-1.4	[-3.6, 0.7]	1.9E-01
Cardiovascular	LDL particle number (Genova)	nmol/L	70	-69	[-110, -27]	1.2E-03	108	-42	[-75.0, -9.2]	1.2E-02
Cardiovascular	Small LDL particle number (Genova)	nmol/L	73	-56	[-93, -18]	3.5E-03	108	-38	[-65, -11]	6.2E-03
Diabetes Fasting glucose (Quest)	mg/dL	45	-1.9	[-3.3, -0.49]	8.2E-03	108	-1.1	[-2.1, -0.21]	1.7E-02	
Cardiovascular	Total cholesterol (Genova)	mg/dL	43	-5.4	[-9.7, -1.2]	1.2E-02	108	0.21	[-2.4, 2.8]	8.7E-01
Diabetes	Insulin	IU/ml	16	-2.3	[-4.1, -0.44]	1.5E-02	108	-0.64	[-1.0, -0.28]	5.0E-04
Inflammation	TNF-alpha	pg/mL	4	-6.6	[-12.0, -1.1]	1.8E-02	108	0.31	[-0.045, 0.66]	8.7E-02
Diabetes	HOMA-IR	(calc)	19	-0.56	[-1.0, -0.089]	2.0E-02	108	-0.15	[-0.26, -0.05]	3.5E-03
Cardiovascular	HDL cholesterol	mg/dL	5	4.5	[0.64, 8.4]	2.2E-02	108	1.9	[0.98, 2.8]	3.7E-05
Nutrition	Methylmalonic acid (Quest) n	mol/L	7	-42	[-85.0, 0.43]	5.2E-02	108	-8.6	[-13.0, -3.9]	3.8E-04
Cardiovascular	Triglycerides (Genova)	mg/dL	14	-18	[-42.0, 6.2]	1.4E-01	108	1	[-5.8, 7.8]	7.7E-01
Diabetes	Fasting glucose (Genova)	mg/dL	47	-0.98	[-2.3, 0.35]	1.5E-01	108	-0.28	[-1.1, 0.51]	4.9E-01
Nutrition	Arachidonic Acid	wt%	35	0.24	[-0.12, 0.59]	1.9E-01	108	-0.21	[-0.39, -0.028]	2.4E-02
Inflammation	hs-CRP	mcg/mL	51	-0.47	[-1.2, 0.28]	2.1E-01	108	-0.09	[-0.47, 0.29]	6.5E-01
Cardiovascular	Triglycerides (Quest)	mg/dL	17	-14	[-37.0, 9.4]	2.4E-01	108	-0.68	[-8.0, 6.6]	8.5E-01
Nutrition	Glutathione	micromol/L	6	11	[-7.6, 29.0]	2.5E-01	108	-0.49	[-20.0, 19.0]	9.6E-01
Nutrition	Zinc	mcg/g	4	-0.82	[-2.4, 0.74]	3.0E-01	108	-0.37	[-0.49, -0.24]	9.9E-09
Nutrition	Ferritin	ng/mL	10	-14	[-42, 13]	3.1E-01	108	-5.7	[-10.0, -1.1]	1.5E-02
Inflammation	Interleukin-6	pg/mL	4	-1.1	[-3.6, 1.4]	3.8E-01	108	0.12	[-0.058, 0.31]	1.8E-01
Cardiovascular	HDL large particle number	nmol/L	8	210	[-390, 800]	4.9E-01	108	110	[-62, 280]	2.2E-01
Nutrition	Copper	mcg/g	10	0.006	[-0.018, 0.031]	6.0E-01	108	0.001	[-0.005, 0.007]	7.1E-01
Nutrition	Selenium	mcg/g	6	0.035	[-0.1, 0.17]	6.2E-01	108	0.014	[-0.003, 0.032]	1.1E-01
Cardiovascular	Medium LDL particle number	nmol/L	20	2.8	[-27, 32]	8.5E-01	108	21	[13, 29]	3.3E-07
Cardiovascular	Small LDL particle number (Quest)	nmol/L	14	-2.3	[-31,27]	8.8E-01	108	13	[4.9,22]	2.0E-03
Nutrition	Manganese	mcg/g	N/A	N/A	N/A	N/A	108	-0.001	[-0.0011, -0.00025]	1.9E-03
Nutrition	EPA	wt%	N/A	N/A	N/A	N/A	108	0.034	[-0.051, 0.12]	4.3E-01
Nutrition	DHA	wt%	N/A	N/A	N/A	N/A	108	-0.07	[-0.11, -0.028]	1.0E-03

Generalized estimating equations (GEE) were used to estimate average changes in clinical labs over time. The coefficient is an estimate of the average change in the population for that analyte by round adjusted for age, sex, and self-reported ancestry. Each coefficient shown has the unit of the analyte it represents. "Out-of-range at baseline" shows the estimates using only those participants who were out-of-range for that analyte at the beginning of the study. N/A values are present where no participants were out-of-range at baseline. "All participants" shows the estimates using all participants in the study. Several analytes are measured by both Quest and Genova; with the exception of LDL particle number, the direction of effect for significantly changed analytes was concordant across the two labs. An independence working correlation structure was used in the GEE.

Table 2.S11: Concordance of 6601 loci between monozygotic twins sequenced on Illumina and CGI

Count	Percent*	Description
6543	99.12%	Fully observed and concordant between Illumina and CGI
13	0.20%	Partially observed in CGI but compatible with Illumina
29	0.44%	NOCALL in CGI
2	0.03%	NOCALL in Illumina
14	0.21%	Fully observed and discordant between Illumina and CGI

Table 2.S12: Number of unique taxa measured for each taxonomic level

Domain	Phylum	Class	Order	Family	Genus	Species	OTU
2	13	27	52	205	779	1275	4616

Table 2.S13: Number of metabolites observed by detection method

GC-FID	GC/MS	LC/MS(neg)	LC/MS(pos)	LC/MS(polar)
34	29	347	159	74

Table 2.S14: Age and sex adjustments for the correlation networks (XLSX 81 kb)

Available for download at https://static-content.springer.com/esm/arti%3A10.1038i%2Fnbt.3870/MediaObjects/41587_2017_BFnbt3870_MOESM36_ESM.xlsx

Chapter 3

MULTI-OMIC BIOLOGICAL AGE ESTIMATION AND ITS CORRELATION WITH WELLNESS AND DISEASE PHENOTYPES: A LONGITUDINAL STUDY OF 3,558 INDIVIDUALS

Note: This chapter can be read here or in the cited publication [5]. There is no extra information here that was not released with the original publication.

3.1 Abstract

Biological age (BA), derived from molecular and physiological measurements, has been proposed to better predict mortality and disease than chronological age (CA). In the present study, a computed estimate of BA was investigated longitudinally in 3,558 individuals using deep phenotyping, which encompassed a broad range of biological processes. The Klemera-Doubal algorithm was applied to longitudinal data consisting of genetic, clinical laboratory, metabolomic, and proteomic assays from individuals undergoing a wellness program. BA was elevated relative to CA in the presence of chronic diseases. We observed a significantly lower rate of change than the expected 1 year/year (to which the estimation algorithm was constrained) in BA for individuals participating in a wellness program. This observation suggests that BA is modifiable and suggests that a lower BA relative to CA may be a sign of healthy aging. Measures of metabolic health, inflammation, and toxin bioaccumulation were strong predictors of BA. BA estimation from deep phenotyping was seen to change in the direction expected for both positive and negative health conditions. We believe BA represents a general and interpretable "metric for wellness" that may aid in monitoring aging over time.

3.2 Introduction

Age is the most important risk factor for most common diseases. There is considerable interest in mitigating aging-related disease risks through lifestyle, pharmaceutical, and environmental interventions that attenuate biological aging. A hurdle in this quest is the quantification of an individual's 'wellness', which is not only the absence of disease but also their resilience to future disease, general satisfaction with one's health and wellbeing, and energy for activities that enrich a person's life. While a multitude of signals relevant to an individual's health and wellness can be captured, meaningful clinical relevance remains a challenge. The development of tools and methods for the collation, integration, analysis, and application of these signals is essential to realizing the goals of precision and personalized medicine[71]. More sensitive and precise assessments of health status and trajectory, guided by dense longitudinal phenotyping, will enable a transformation in modern health care. Such a paradigm shift can only occur by converting these sophisticated, high-dimensional measures into actionable metrics. Biological age (BA), to the extent it can be estimated, may provide one such personalized and intuitive metric of overall health status that can be communicated effectively to a general population.

Estimation of BA was first proposed in 1969 [72]. In 1988, Baker and Sprott proposed that a biomarker of aging is a biological parameter of an organism that either alone or in some multivariate composite will, in the absence of disease, predict physiologically functional capacity at some later stage better than chronological age (CA) [73]. More recently, BA has been assessed via epigenetic markers [74], proteomics [75], and Electronic Medical Records [76]. The Klemera-Doubal (KD) method has been suggested as a better predictor of all-cause mortality than either CA alone or using multiple linear regression of ten clinical biomarkers [77, 78]. Studies using small numbers of highly informative clinical variables to develop KD-computed BA measures have demonstrated these measures associated with poor balance, physical weakness, declining cognitive performance, physical appearance, cardiovascular risk, frailty indices, extrinsic epigenetic age, caloric restriction (CR), and gene expression [79–83]

.

Deep phenotyping offers the opportunity to explore multiple systems that contribute to BA in greater depth, and generate more comprehensive metrics of overall health that change over time, aimed at reflecting an individual's changing health [84]. Herein, we also explore estimated BA by distinct data types: the metabolome, proteome, and clinical labs, as well as a BA calculation that integrates all of these together. BA appears to be modifiable, and thus may be a simple metric that is useful to monitor general health.

In this work, KD was applied to over 900 disparate (principal component analysis, PCA, transformed) biomarkers, including metabolites, proteins, genomics, and clinical measures. This collection of biomarkers is herein termed personal, dense, dynamic data (PD3) clouds [4]. Data type (eg, different omics measures)-specific BA estimates were compared to each other and changes in BA over time were examined by data type, and among subgroups that were hypothesized to have different BA trajectories (including stratifications by sex, ethnicity, age group, and baseline BA). Differences between biological and chronological age (BA-CA) were utilized as a metric, noted as ΔAge (more negative indicates scoring younger than CA), and associations between ΔAge and lifetime prevalence of common health conditions were examined.

In this study, we ascertained the effects of conditions and behaviors generally thought of as being healthy or unhealthy upon the introduced BA measure. We found that "healthy" behaviors, such as participation in a scientific wellness program [85], were found to be associated with a decreasing ΔAge over time. Conversely, "unhealthy" conditions, such as self-reported diseases, were found to increase ΔAge in every condition we had data for where there was a significant effect (no significant effects in the opposite direction). The observation indicated that ΔAge was sensitive to changes in the blood associated with common disease states. Association strength and computed BA estimates varied significantly by data type (proteomics, metabolomics, and clinical labs), demonstrating that BA depends on the systems being interrogated. These results support the construction of a BA measure that integrates diverse information across different -omics, biological systems, and disease

biomarkers – and/or the use of multiple BA measures to reflect different biological systems – to help assess individual health and for the quantification and exploration of aspects of the aging process in humans.

3.3 Methods

3.3.1 Study Population

The sample studied consisted of men and women participating in a consumer data-intensive wellness program (Arivale, now closed) that varied by age and health status (demographics given below). The program involved lifestyle coaching on exercise, nutrition, stress management, and sleep all tailored to the participants' health goals, specific genetic markers, and clinical metrics as detailed in a prior publication [85]. Deidentified data from consenting participants were collected from July 2015 to July 2018. A total of 3,558 participants were observed for an average of 214 days, with an average of 2.1 longitudinal data points with a total of 7,634 observations. In total, 1,354 participants had a single time point, 1,105 had two, 711 had three, and 388 had four or more, with two participants having the maximal (8) number of time points. Average time between observations was 190 days among participants with multiple time points. The study was approved by the Western Institutional Review Board.

3.3.2 Personal, Dense, Dynamic Data Clouds (PD3 Clouds)

We previously developed and published analyses incorporating proteomic, metabolomic, microbiomic, and genetic data (the PD3 cloud) on 108 participants in the context of health and wellness [4]. This cohort ultimately expanded to 3,558 individuals at the time data were collected for this study. Participants' genetic profiles were assessed either by whole genome sequencing (2845) or by single nucleotide polymorphism (SNP) chip (713). Detailed information on the acquisition, storage, generation, and analyte-specific pre-processing of these measures is available in the Supplementary Methods 3.S.1. After pre-processing, the PD3

clouds included genomics plus longitudinal measures from blood, including 54 clinical lab tests from LabCorp or 67 clinical lab tests from Quest, 243 proteins, and 611 metabolites with CA ranging across the adult lifespan (18-89+ years).

3.3.3 Creating the BA Measure

The KD method, with a PCA transformation on the input features, was used to create the BA measure [77]. Briefly, KD is a weighted average of independent linear regressions of biomarkers to CA. Ten iterations of 10-fold cross-validation were performed to estimate BA from each data type (clinical labs, metabolites, and proteins). Male and female data were trained separately, as were observations from different laboratory vendors. Training/testing set splits were generated by randomly shuffling participants, partitioning them into ten sets, and iterating over those sets, with one set as the test set and the remaining nine being used for training. Training sets were restricted to baseline measurements, ensuring those participants had minimal wellness coaching, and only one observation of a participant was trained on. All observations of participants in the test set were predicted from the training set. Clinical labs had two vendors, so only the earliest observation among both vendors was included in the training. All samples were z-score normalized using the mean and SD estimated from the training set at each fold.

Similarly, principal components were estimated using the training, and the transformation was then applied to all samples. Principal components were used to satisfy the biomarker linear independence requirement of the KD algorithm. Slopes, SDs, and intercepts were calculated for each of the strongest components explaining up to 90% of the variance. These variables were then used to calculate BA using KD. The contribution of each analyte to BA was calculated by multiplying the weights learned for each component by the analyte contribution to each component and summing across all components. These representations are equivalent because PCA and KD are linear transformations (see Supplementary Methods3.S.1). CA was excluded as a biomarker, although KD allows its inclusion. Doing so reduces variance, but adds limited information regarding BA's relationship to health

outcomes For each data type, the 10 predictions were averaged. For each observation of a participant, all available data type predictions were averaged and presented as the overall BA prediction. A total of 2,742 observations had only one data type, 3,634 had two, and 1,258 had all three. See Supplementary Figure 3.S1 and Supplementary Table 3.S1 for details.

3.3.4 Trajectory of the BA Measure Over Time

To determine whether BA changed over time after initiation of health coaching, Generalized Estimating Equations (GEEs) with exchangeable correlation structure were utilized, which accounts for the correlation between multiple observations (time points) per participant [86]. Participants with two or more blood draw visits were included in the trajectory models. The primary model assessing linear change in BA included time in the program as the independent variable, starting at time zero (time of first blood draw), and BA as the dependent variable; baseline CA was included as a covariate. Models were stratified by baseline age group, sex, and race (white vs non-white), and interaction terms between study time and sex/race were included to assess for effect modification. Additionally, as factors that may impact BA over time are of interest, models were stratified by CA decade and BA starting point to model BA changes due to differences in initial health status at coaching initiation (model A: participants with an initial BA that was 5 or more years greater than CA; model B: participants with an initial BA that was 5 or more years less than CA. This analysis was repeated with BA $\pm$ 10 years from CA).

3.3.5 Health History, Behaviors, and Associations With ΔAge

GEE models with exchangeable correlation structure were used to examine associations between ΔAge under combined and independent data modalities and the lifetime prevalence of the 40 most common self-reported health conditions, along with lifetime and/or current smoking. In the minimally adjusted model, ΔAge was modeled as the outcome variable, and self-reported past or current condition was the predictor, with CA at each prediction included as a covariate. Each condition was modeled separately. Since obesity was hypothesized to be

strongly associated with ΔAge and many conditions, obesity (0 for body mass index [BMI] < 30, 1 for BMI *geq* 30) was included as a covariate. Association between obesity and BA itself was also calculated. A Bonferroni correction at alpha = .05/(43(conditions)*4 (modalities)) = 3E-04 threshold for statistical significance was applied. Many health outcomes are highly correlated with one another, and thus, this correction is highly conservative.

3.4 Results

3.4.1 Population Characteristics

Mean age was 47.5 years, with more females than males (58.6% female). Baseline characteristics are presented in Table 3.1. The percent of obese participants was 27.9%, lower than the Center for Disease Control reported estimate of 37.9% for all U.S. adults. This bias appears to be driven primarily by regional makeup, rather than the self-selection of lower BMI individuals. This cohort is predominantly (80%) drawn from Washington or California. Given the state/province of residence, the expected percentage of obese individuals is 27.7% [87]. Socioeconomic status of participants is presumably higher than the national average, but that information was not captured.

3.4.2 BA Estimation Through PD3 Clouds

BA estimates using the KD method are shown in Figure 3.1. The Pearson correlation between BA and CA was .78 overall, .70 for the clinical labs, .81 for the metabolomics, and .88 for the proteomics. The median absolute error, that is, the median absolute difference between BA and CA, of these predictions was 5.54 years overall, 8.04 years for clinical labs, 4.82 years for metabolomics, and 4.39 years for proteomics. Mean (SD) over repeated predictions for the same observation was 3.83 years overall, 1.05 years for clinical labs, 1.52 years for metabolomics, and 1.03 years for proteomics. ΔAge had a mean (SD) of -0.78 (9.28) years overall, -0.43 (12.18) years for clinical labs, -0.11 (7.48) years for metabolomics, and -0.73 (6.57) years for proteomics. ΔAge was largely uncorrelated with CA, at a Pearson

Table 3.1: Demographic Information

Characteristic	All[a]	Women[a]	Men[a]	p Value[b]
Chronological age, mean years (SD)	47.6 (12.2)	48.7 (11.6)	47.9(11.8)	.1
Non-white, no. (%), n = 3,452	784 (22.7)	416 (20.5)	368 (25.8)	<.001
BMI, mean (SD), n = 3,227	27.7 (6.3)	27.8 (7.1)	27.4 (4.9)	.05
Obese[c], no. (%), n = 3,227	901 (27.9)	584 (30.7)	317 (23.9)	<.001
Moderate activity $\geq 3\times$/wk, no. (%), n = 3,468	2,253 (65.0)	1,278 (62.0)	975 (69.3)	<.001
Vigorous activity $\geq 3\times$/wk, no. (%), n = 3,467	1,121 (32.3)	528 (25.6)	593 (42.1)	<.001
Sitting > 8 h/d, no. (%), n = 3,467	2,303 (66.4)	1,392 (67.6)	911 (64.7)	.09
Ever smoker, no. (%), n = 2,825	774 (27.4)	448 (25.9)	326 (29.7)	.03
Current smoker, no. (%), n = 3,469	174 (5.0)	73 (3.5)	101 (7.2)	<.0001
Cholesterol-lowering meds, no. (%), n = 2,817	337 (12.0)	163 (9.5)	171 (15.7)	<.001
Past and/or current self-report of:				
High cholesterol, no. (%), n = 3,351	788 (23.5)	408 (20.4)	380 (28.1)	<.001
Hypertension, no. (%), n = 3,361	579 (17.2)	313 (15.6)	266 (19.6)	.003
Asthma, no. (%), n = 3,389	559 (16.5)	370 (18.3)	189 (13.8)	<.001
Type 2 diabetes, no. (%), n = 3,309	125 (3.8)	78 (4.0)	47 (3.5)	.6
Breast cancer, no. (%), n = 3,235	63 (1.9)	59 (3.0)	4 (0.3)	<.001
Coronary artery disease, no. (%), n = 3,280	50 (1.5)	19 (1.0)	31 (2.3)	.003

a) Total N = 3,558; women, N = 2,087; men, N = 1,471. b) For comparisons between men and women, chi-square tests were run for categorical variables and two-sided t-tests for continuous variables. c)Obese defined as BMI $\geq$30.

r of -.06 overall, -.03 for clinical labs, -.18 for metabolomics, and -.10 for proteomics. See Supplementary Table 3.S2 for summary statistics.

Pearson correlation of ΔAge between multiple observations of the same participant, that is, ρ(ΔAget, ΔAget+1), was .66 overall, .67 for clinical labs, .67 for proteomics, and .64 for metabolomics. These correlations were stronger than the between-data type ΔAge, with clinical labs correlating with metabolomics at an r of .26, clinical labs with proteomics at .25, and metabolomics with proteomics at .27 (Supplementary Figure 3.S2).

3.4.3 BA Changes Over Time

Mean linear trajectory of BA over time, calculated using longitudinal measurements among participants with at least two visits, varied according to whether the predictions were based on clinical labs, metabolites, proteins, or a combination of all three categories (Table 3.2). In the minimal model adjusted for baseline CA, BA prediction based on all available analytes showed that BA stayed statistically stable over time. On average, BA decreased by 0.16 years for every year of participation in the wellness program (B = -0.16, 95% CI: -0.45, 0.19). This is significantly lower than the expected increase of 1 year/year. BA estimates from all data types had a B coefficient < 1, the natural rate of aging, with all data types except metabolomics being significantly < 1.

3.4.4 Potential Modifiers of BA Trajectory Over Time

Exploratory analyses to examine several baseline factors (sex, ethnicity, age group, and baseline health status) that were hypothesized a priori might have an impact on BA trajectories were performed. In sex-stratified models based on the "all analyte" BA predictions (clinical labs, metabolites, and proteins combined), BA decreased in women over time (coefficient: -.48, 95% CI: -0.93, -0.04), but stayed constant in men (coefficient: .19, 95% CI: -0.36, 0.74) (Table 3.2). The sex-time interaction term was weakly significant ($p < .05$), indicating a difference between men and women in their BA trajectories over time. A similar pattern in BA derived from proteins and clinical labs was observed (Supplementary Table 3.S3), with

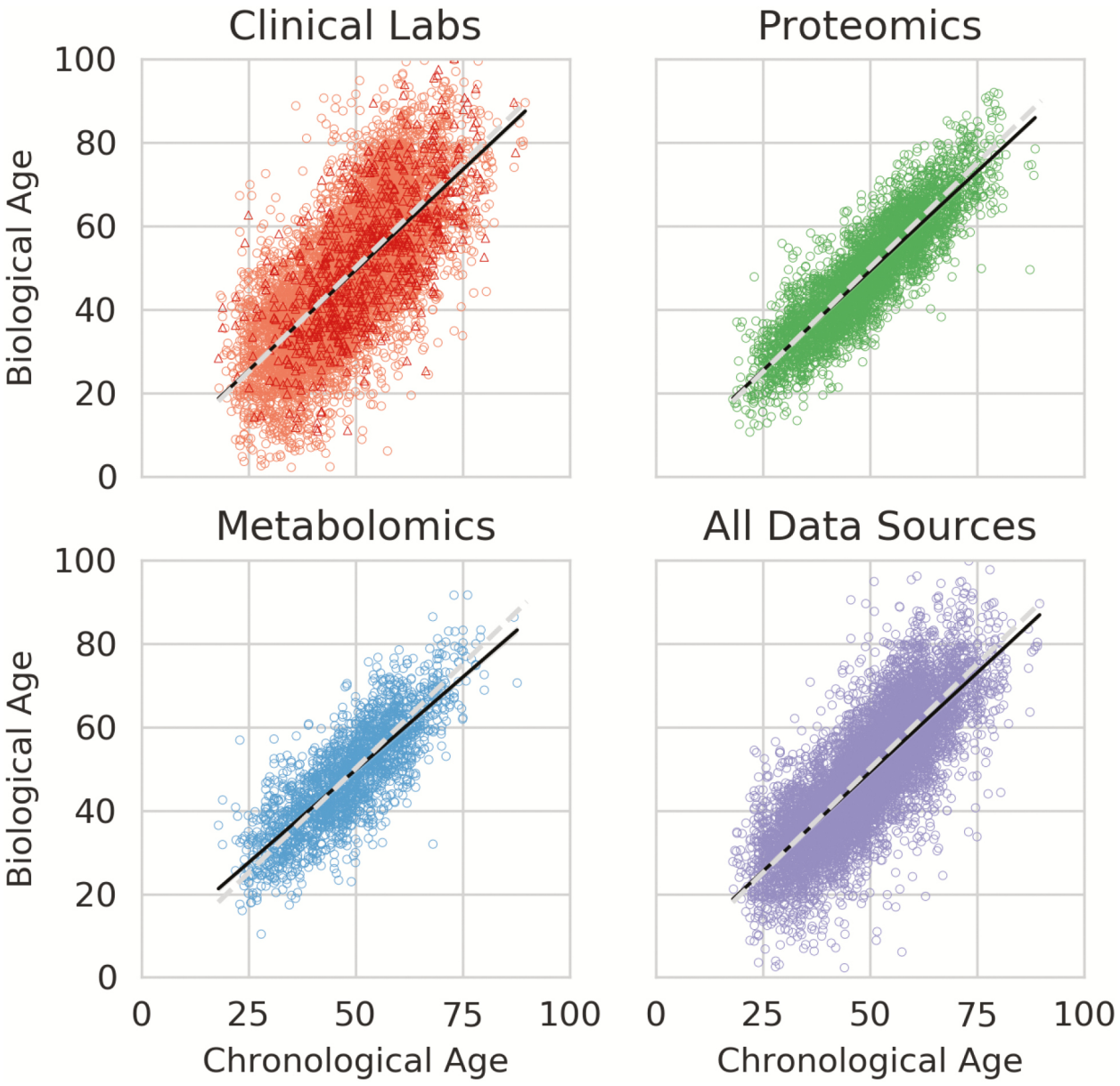


Figure 3.1: Multi-omic Biological Age Estimates.

Scatter plots of Biological Age estimates using the KlemeraDoubal algorithm for each data type individually, and in aggregate (All Data Sources). Each point is one observation of an individual. The solid line in each plot is the ordinary least squares regressed line, and the dotted line is biological age = chronological age. The Clinical Lab plot contains estimates from two vendors: Labcorp (circle) and Quest (triangle).

Table 3.2: Change in Δ Age over time.

Data set used for BA prediction[a]	*B* Coefficient	Std. error	95% CI	Interaction p[c]
All analytes	-0.160	0.181	-0.452, 0.194	NA
Proteins	0.524	0.179	0.174, 0.874	NA
Metabolites	0.274	0.471	-0.648, 1.196	NA
All clinical chemistries	-0.374	0.216	-0.798, 0.049	NA
Stratified analyses[b]				
Sex				
Males (n=908)	0.192	0.281	-0.359, 0.742	
Females (n=1297)	-0.484	0.229	-0.932, -0.036	0.043
Self-reported ethnicity				
White (n=1705)	-0.111	0.198	-0.498, 0.276	
Non-white (n=433)	-0.589	0.465	-1.500, 0.322	0.371
Age at baseline, by decade				
18-29 years (n=130)	-2.720	0.744	-4.180, -1.270	
30-39 years (n=388)	-0.055	0.419	-0.877, 0.766	
40-49 years (n=658)	-0.356	0.332	-1.010, 0.295	
50-59 years (n=638)	0.530	0.356	-0.168, 1.230	
60-69 years (n=322)	-0.285	0.393	-1.050, 0.484	
70+ years (n=69)	-0.895	0.853	-2.570, 0.776	NA
Baseline BA prediction				
BA=5 years > CA (n=481)	-0.985	0.423	-1.814, -0.157	
BA=5 years < CA (n=540)	0.024	0.377	-0.715, 0.764	
BA=10 years > CA (n=167)	-1.540	0.686	-2.88, -0.194	
BA=10 years < CA (n=187)	1.613	0.651	0.388, 2.890	NA

[a] GEE Model: BA*sim*time in Arivale + baseline CA; clustered by client ID, family=Gaussian, with an exchangeable correlation matrix; only individuals with at least two visits were included [b] GEE Models, stratified by sex, ethnicity, age group, and baseline BA prediction: ΔAge (BA-CA) ∼ time in Arivale + baseline CA; clustered by client ID, family=Gaussian, with an exchangeable correlation matrix; All models use BA predictions based on the "All analyte" data set [c] Interaction models: ΔAge (BA-CA) ∼ time in Arivale + predictor variable + baseline CA + predictor variable × time in Arivale; clustered by client ID, family=Gaussian, with an exchangeable correlation matrix.

BA from clinical labs also indicating a weakly significant difference between men and women (interaction p = .02). Since race was unevenly distributed throughout CA, self-reported race (white vs non-white) was also stratified; the interaction term was not significant, and both groups had slowed BA compared to CA. In models stratified by age (in decades), the youngest age group had the slowest rate of biological aging (using BA estimates from the all-analyte data set), though all age groups except 50-59 years indicated slowed aging (upper 95% CI < 1). This effect was not dose dependent (ie, rate of aging did not increase monotonically) and was roughly consistent across data types. This analysis for BA was repeated for each data type and observed similar patterns in BA derived from clinical labs; *B* coefficients derived from proteins and metabolites were highly variable and had wide CIs, likely due to small N per age group (Supplementary Table 3.S3).

Lastly, baseline ΔAge was stratified, with the idea that participants with higher ΔAge at study entry would be less healthy (under the assumption that ΔAge is an adequate summary metric for health and wellness), and therefore experience greater benefit from health coaching. Participants with BA 5 or more years higher than their CA at baseline experienced approximately 1-year decline in BA for every 1 year in the program (coefficient: -.99, 95% CI: -1.81, -0.16), while participants who entered the program with BA at least 5 years less than CA maintained their youthful BA over time based on the all-analyte BA estimates (coefficient: .02, 95% CI: -0.72, 0.76). These effects were more pronounced in individual data modalities, though many stratified analyses had small N, which likely inflated estimates. Regression-to-the-mean effects could not be ruled out in the absence of a control group, particularly when N was small or when baseline deviations were extreme (ie, the >10 years plus or minus for ΔAge) [88].

ΔAge Is Associated With Health and Behavior and Especially With Type 2 Diabetes Among the top 43 most common health conditions and behaviors in our cohort, after correcting for multiple comparisons, obesity, hypertension, high cholesterol, lung infection, type 2 diabetes (T2D), and breast cancer were associated with increased ΔAge in models adjusted for CA and obesity (Figure 3.2). T2D had the highest increase in ΔAge in combined models,

such that these participants had a BA that was higher than their CA by an average of 6.4 years (95% CI: 4.6, 8.2). This effect was consistent among data types for T2D. However, effects varied slightly among the different data types. For instance, the combined data type derived BA provided highest statistical significance for increased ΔAge among participants with high cholesterol, the estimates based on metabolomics and clinical labs were also associated with increased ΔAge, though below the threshold for multiple corrections. Estimates derived from proteomics did not show as pronounced an effect, with the coefficient having a p-value >.05. Since the all-data modality and the clinical labs had the largest N, these associations were well powered and most likely to show significant associations. However, several nonsignificant trends of interest were observed indicating potential disease-specific differences in sensitivity among different data modalities, with some health conditions having a trending association ($p < .05$)

with only one of the four modalities (such as concussion, endometriosis, kidney stone, gallstones, cataracts, and coronary artery disease). While these trends were not strong enough to be significant individually after a conservative Bonferroni correction for multiple hypothesis testing, collectively, every one is in the direction of increasing ΔAge with none in the opposite direction, adding confidence in their likely validity.

3.4.5 Analytes That Are Most Predictive of a High or Low BA Measure

The top mean model coefficients, representing the importance of individual analytes in the model, are shown in Figure 3.3. The value of each analyte coefficient corresponds to the contribution of that analyte to the computed BA. For instance, a coefficient of +1 indicates an increase in BA of 1 year per SD higher than the mean, while a coefficient of -1 indicates a corresponding decrease in BA of 1 year per SD above the mean. Most markers that were strongly predictive of BA were dominated by three axes of aging: metabolic health, inflammation, and bioaccumulation of toxins.

In clinical labs, glycated hemoglobin (HbA1c) was the strongest positive predictor of BA independent of sex, with other (highly correlated) metabolic health markers demonstrating

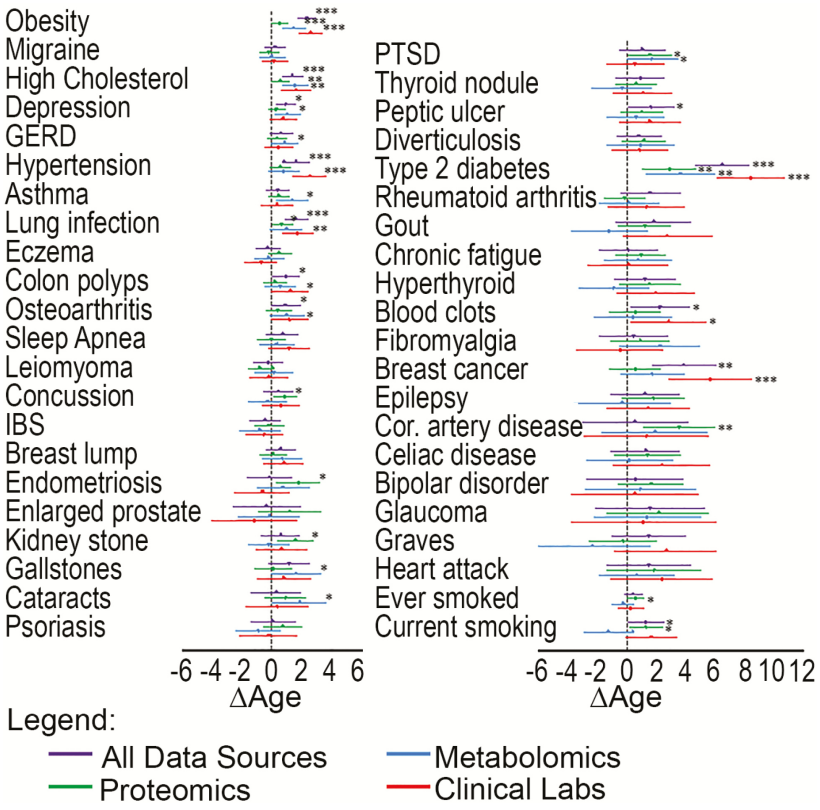


Figure 3.2: Disease associated with increased Age.

Forest plot of Age estimates and 95% confidence intervals associated with the 40 most common health conditions, plus ever smoking, current smoking, and obesity. Each condition or behavior was modeled individually, with Age as the dependent variable, the health condition/behavior as the independent variable, and further adjustment for chronological age (CA) and obesity (body mass index > 30) in Generalized Estimating Equation models clustered by client ID with an exchangeable correlation matrix to account for multiple observations from individual clients. The obesity outcome was adjusted for CA only. Biological age (BA) estimates for each data type are shown. The blue dotted line at 0 indicates no difference between BA and CA; point estimates to the right of the blue line indicate higher BA than CA associated with the health condition/behavior (eg, based on the all-data-type BA estimate, individuals with type 2 diabetes have BAs that are, on average, 6.4 years greater [95% CI: 4.6, 8.2] than their CAs, after adjustment for CA and obesity). ***p $< .0003$ (Bonferroni threshold); **p $< .005$; *p $< .05$. GERD = gastroesophageal reflux disease; IBS = irritable bowel syndrome; PTSD = Post-traumatic stress disorder

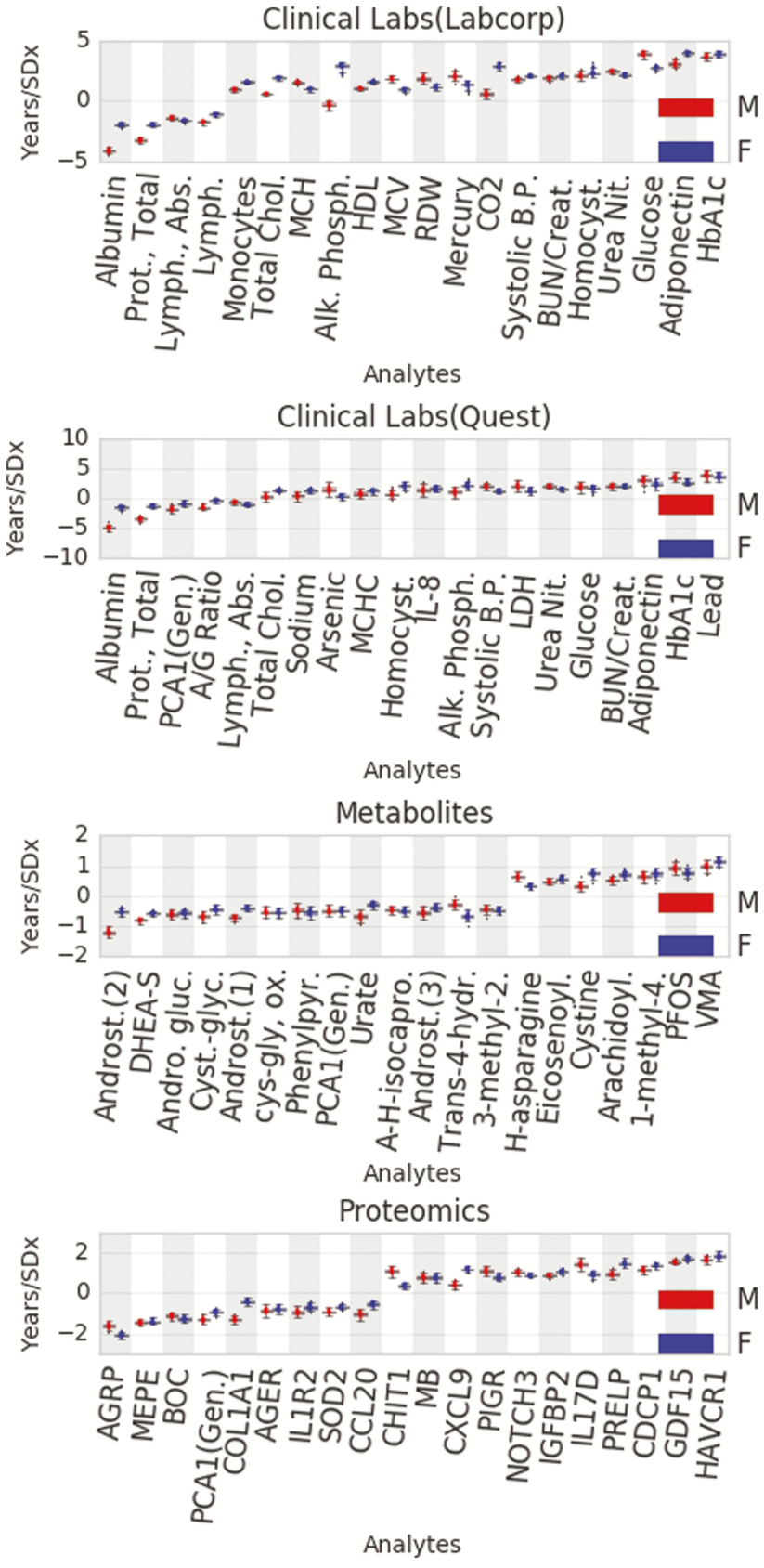


Figure 3.3: Top 20 aging predictors by data type.

similar effects, that is, adiponectin and glucose. Metabolic health was also reflected in proteomics, where agouti-related peptide (AgRP) was the strongest negative predictor of BA for both men and women. AgRP is involved in energy balance through regulation of appetite and energy expenditure (RN19). Similarly, analytes reflective of redox balance, an integral component of metabolic homeostasis, were strongly predictive of BA. The metabolite subfamily of glutathione was one of the strongest predictors of BA for both men and women (Supplementary Figure 3.S3).

Multiple proteomic markers of inflammation were associated with BA, including chemokine C-X-C motif ligand 9 (CXCL9), interleukin 17D (IL17D), and growth/differentiation factor 15 (GDF15); additionally, the lymphocyte produced lymphotoxin alpha (LTA) was a negative predictor of BA in men, but not in women. Inflammation plays a crucial role in BA prediction in the clinical labs as well, where monocyte count was a strong positive predictor of BA and lymphocytes were a strong negative predictor.

Several environmental pollutants, including the heavy metals lead and mercury, were identified as strong predictors of BA. Within the metabolomics, the bioaccumulated toxin perfluorooctanesulfonic acid (PFOS) emerged as the second strongest positive predictor. The related metabolite, perfluorooctanoic acid (PFOA) was a strong positive predictor in men but not in women (Figure 3.3).

Sex steroid hormones dominated the calculation of BA in metabolites for men and women. Dehydroepiandrosterone (DHEA-S) and its direct metabolite androstenediol monosulfate were strong negative predictors with the stress-related hormone vanillylmandelic acid (VMA) being a strong positive predictor. Similar to PFOA and LTA, several other analytes demonstrated sex-specific differences (Supplementary Table 3.S4). Alkaline phosphatase (ALP) proved to be a strong positive predictor in women but not in men (Figure 3.3). In contrast to ALP, creatine metabolites emerged as an important subfamily for calculation of BA in men, but not women (Supplementary Figure 3.S3). Within the creatine metabolite subfamily, creatinine was one of the stronger negative predictors in men. Several elements of the immune system also showed sex-specific differences in calculating BA, including Spondin 2

(SPON2), which was a negative predictor in women, but not in men. Macrophage receptor with collagenous structure (MARCO) was a positive predictor in women, but a negative predictor in men. IL-16 was a positive predictor in women, but a negative predictor in men. The intestinal mucosa secreted protein trefoil factor 3 (TFF3) was a positive predictor in men, but a negative predictor in women.

3.5 Discussion

The BA measure was generated by integrating and comparing diverse data types, including clinical labs, proteomics, metabolomics, and genetics. The key findings of this paper are as follows: 1) Higher ΔAge was shown to be associated with lifetime prevalence of common disease conditions, and BA was seen to decrease over time after joining a wellness program, supporting the hypothesis that BA is reflective of increasing or decreasing health as commonly understood, 2) The degree of plasticity in BA is dependent on several factors such as sex, current health status, and CA, 3) Blood factors corresponding to metabolic health, inflammation, and bioaccumulation of toxins were found to be the most strongly related to BA across data types, 4) Men and women showed distinct differences in the features most relevant to the determination of BA, especially those related to aspects of sex-specific physiology, such as bone density, muscle mass, immune system function, and sex-related metabolism of environmental pollutants, 5) BA was affected by the data type used in their determination, and different BAs can thus be derived from different data sources.

The complexity and variability of the aging process justify the development of system-level predictive and analytical models to describe it, with the ultimate goal of maintaining healthy aging and improving the extent and quality of healthspan through actionable lifestyle, environmental, and pharmaceutical interventions. Following an unscreened sample enabled the observation of health-related changes across the spectrum of commonly observed health conditions. The highest ΔAges, perhaps indicating poor wellness relative to CA, are in the T2D subpopulation (+6 years) which is consistent with studies observing 5-9 years shortened life expectancy with T2D[89]. Average ΔAge was higher among participants self-reporting

multiple types of current or past health conditions. This finding does not suggest that BA is a useful diagnostic for any specific disease, but instead that ΔAge maps consistently to the concept of general wellness, where every disease condition in which a statistically significant association was discovered (or even a lower-threshold trend observed) was in the direction of increasing ΔAge. Stationary or negative BA trajectories over time, after initiation of a wellness program, was consistent with the potential utility of BA as an aggregate marker (metric) of increasing wellness. In general, the all-analyte predictions of BA did not increase or decrease over time in this sample on average, despite increasing CA. Further study is required to determine the persistence of these effects or efficacy relative to other interventions.

Stratified analyses highlight differences between groups in response to their engagement in a wellness program. Both men and women experienced slowed BA on average, but the effect was stronger in women with their BA decreasing over time, while men maintained their BA. On average, the youngest participants (18-29) tended to show more ability to reduce their BA, while older participants decreased their ΔAge but maintained their initial BA. Importantly, a dose-dependent response was not observed, that is, participants over 29 have roughly similar trends. Participants tended to maintain a high degree of consistency in ΔAge over time for all data types.

Of interest is baseline health status. Participants with high ΔAge experienced a greater decline in BA over time in the program, which may be expected, given that less healthy participants had more actionable "wellness targets" to work on. Diminishing returns were also observed as those with extremely low baseline BA relative to CA had a slope of approximately the expected standard rate of BA (though confidence intervals were wide). While this seems expected from a biological perspective, the direction and magnitude of these trends are consistent with regression-to-the-mean effects, especially at the most extreme strata (ie, $>$ 10 years —ΔAge—). An independent control group, not undergoing wellness coaching, would be required to differentiate these two effects regression to the mean and improvement from the wellness program.

Metabolic health, inflammation, and bioaccumulation of toxins represent dominant themes

under our BA models across data types. The importance of metabolic health is well supported in aging literature, and a major concern in the developed world with nearly 40% of Americans expected to develop T2D in their lifetime and diagnosed diabetes patients accounting for one in four health care dollars in the United States in 2017 [90, 91]. The substantial effect of HbA1c, where 1 SD increase corresponded to a roughly 4-year increase in BA, partially explained the considerable effect on BA observed in participants that self-reported T2D. Adiponectin and AgRP are involved in the regulation of appetite and energy balance, with their levels in the blood rising in response to fasting and CR [92, 93]. Interestingly, adiponectin was a positive predictor of BA in our models, despite its aforementioned beneficial role in metabolic regulation. This is consistent with the proposed "adiponectin paradox," where despite its beneficial role throughout the life span, increased circulating adiponectin levels in elderly populations are associated with a higher risk of mortality [94]. The purported health benefits of CR are, in part, attributed to its ability to slow down metabolic decline and decrease oxidative stress. Consistently, strong beneficial effects from the anti-oxidant glutathione subfamily observed in the metabolites are consistent with these inter-relationships. Chronic inflammation is a common risk factor in many age-associated diseases, including heart disease, depression, cancer, osteoarthritis, and diabetes [95, 93]. Concordantly, changes in immune activity as people age were reflected in BA [75, 96]. CXCL9, a strong positive predictor of BA, is involved in the chemo-attraction of T cells and NK cells and has been demonstrated as a biomarker for the development of heart failure [97, 98]. CXCL9 and GDF15 were shown to explain significant variability in arterial stiffness and myocardial relaxation [99]. The negative association of LTA with BA is aligned with its broad anti-tumor active, via multiple pathways, including the recruitment of NK cells [100]. Bioaccumulation of toxins is known to be detrimental to human health, especially in Alzheimer's disease, and a growing concern as people age [101, 102]. Several environmental pollutants, including the heavy metals lead and mercury, were identified as strong positive predictors of BA.

While most of the strongest predictors of BA were shared, sex-specific analyte contribu-

tions illuminate some differences in the biological aging process. For example, ALP was a strong positive predictor of BA in women, but not in men (Figure 3.3). Circulating ALP levels are commonly used as a marker for liver or bone disease, as total ALP consists mainly of bone and liver-derived isoforms. Particularly relevant to bone, increase in total and bone-specific ALP levels has been associated with increased rates of bone turnover ([103, 104]). Given postmenopausal women experience higher bone turnover rate and accelerated bone mineral density loss with age compared to men, the difference in the effect of ALP on BA between men and women may result from sex-specific differences in bone physiology across the life span ([105]).

In contrast to ALP, creatine metabolites emerged as a notable subfamily for BA estimation in men, but not in women (Supplementary Figure 3.S3). Within the creatine metabolite subfamily, creatinine was one of the stronger negative predictors for men. While creatinine build-up can be an indicator of reduced kidney function, it is also commonly used as a surrogate marker for muscle mass [106–108]. This difference may reflect age-related muscle loss (sarcopenia) that is generally more pronounced in males than females [109]. PFOA was also a strong positive predictor of BA in men only. Kinetic studies suggest sex differences in the excretion of PFO metabolites, which may in part explain the observed effects [102, 110]. Additionally, animal studies have shown that higher testosterone levels increase the rate of elimination of PFOA [111]. Decreasing testosterone levels as men age or with obesity may partially explain the predictive capacity of PFOA levels in men but not in women.

Particularly intriguing is the fact that different data types illuminate different facets of wellness (Supplementary Figure 3.S2), even though each data type was independently effective at estimating CA (Figure 3.1 and Supplementary Table 3.S2). While each data type provides rich information about an individual's biological state, the view into that state is inextricably affected by the modality of those measures. It has previously been demonstrated that different omics profiles of the same individuals do not cluster together[112]. Data type-dependent differences among associations between ΔAge and some health conditions were observed (Table 3.2). For instance, ΔAge estimates derived from proteomics were associated

with coronary artery disease, while estimates from the other data types had CIs showing little effect. While this association was not significant after FDR correction (unadjusted p = .004), the protein panels used were heavily focused on inflammation and cardiovascular disease, and so this result is not surprising. Determining which data types are most appropriate for certain diseases may help create condition-specific calculations of BA, and lead to greater precision based on an individual's specific health concerns and history. This study argues that a fuller picture of an individual's health emerges by incorporating multiple views of aging systems. As costs decrease over the next 10-15 years, expanding the protein, clinical chemistry, and metabolites panels to the largest extent reasonable will enable each of the different analyte classes to reflect in the broadest possible manner the "integrated" aging process.

This study confirms previously identified biomarkers that also estimated BA. Eight of the 10 biomarkers identified in Levine (2012) are measured in the clinical labs, with 6 being top predictors in our clinical lab models (Figure 3.3 and Supplementary Table 3.S4) [78]. Creatinine was not directly a top predictor in the clinical labs, but the blood urea nitrogen and creatinine ratio was, and creatinine is one of the strongest predictors of BA in men in the metabolomics. Presence of a large number of inflammatory markers may explain why C-reactive protein does not emerge as a particularly strong predictor. Another study demonstrated GDF15 as a potent predictor of BA [75]. These verifications reinforce the generalizability and relevance of these biomarkers to BA.

Strengths of this study include deep phenotyping, large cohort size, a broad age distribution (18-89+), and longitudinal measurement of participants actively improving their health through lifestyle changes. Limitations of this study include the lack of many aging-specific covariates (such as grip strength, balance, and cognition), the short duration of observation relative to earlier epidemiological studies, and the lack of uniformity of measures across all people and observations. As mentioned, since a suitable control population (individuals not enrolled in a wellness program) was not available, regression-to-the-mean effects in analyses stratified by baseline BA subgroups could not be ruled out, particularly those with the largest deviation (outside of ± 10) of ΔAge away from zero. The lack of a control group

additionally raises issues for interpretation of these results. Neither causality nor the relative efficacy of this program compared to other interventions can be determined. Data type-specific stratified analyses were often underpowered, yielding large CIs and inconsistent estimates. Nevertheless, these exploratory analyses demonstrate intriguing trends for future studies. This study focuses on the applicability of BA to the whole adult life span as a general measure of wellness by assessing through hundreds of blood analytes literally 100s of biological networks. The lack of uniformity of measured variables over time presents challenges in integration and analysis, which are inevitable in the process of utilizing real-world data. Interest in repurposing incidental measures, electronic medical records, patient-contributed data, and mining of public databases is high. Thus, developing flexible methods that robustly integrate existing data is a superior strategy to ignoring essential features of human health due to partial missingness.

One question raised by application of deep phenotyping to calculate BA is whether measuring these large sets of variables is justified. They are at the level of discovery,that is, you want to survey the largest possible set of analytes to discover those which have the dominant effects on BA. Once these are discovered, far more limited feature tests can likely be assembled to calculate BA. Notably, a perfect predictor of CA would be useless as a wellness marker, giving no more information than the individual's birthday[113]. The main point is whether deviations in prediction represent deviations from wellness states and the extent to which this measure is modifiable. Longitudinal, deep phenotyping of individuals allow us to fully realize the broad dimensionality of a given population – and they allow us to stratify the population based on personal data clouds of the individuals and not on averaged data from populations. Additionally, if BA or ΔAge were used as a summary metric for wellness, a drop in BA over time may encourage participants to persist with healthful behaviors in order to maintain their "healthy" progress and allow one to carry out individual $N = 1$ studies on interesting compounds to facilitate healthy aging with lower BA as a target measure. Thus, it is proposed that ΔAge can be a useful metric to facilitate healthy aging. While the population insights herein are robust, reducing the high variance in the metric,

however, is clearly an important factor in how such a measure might be used in the future on an individual basis.

This study estimates BA measures from PD3 clouds as gross, aggregate measures of health and wellness, which are useful because they constitute the averaging of many different biological systems. Importantly, BA has the potential to serve as a metric that can be used to track progress towards healthy aging. The factors affecting BA represent acute and cumulative damage that occurs over an individual's lifetime and are mostly actionable through lifestyle, environmental, and pharmaceutical intervention. BA measures may be positive or negative wellness markers that can be used in instances where an individual lacks any specific disease conditions but is still interested in increasing wellness and preventing disease. Additionally, as CA is used to determine risk categories for many prophylactic tests such as colonoscopies, prostate exams, and mammograms, so too might BA provide personalized guidance on the relevance of those tests. As health care moves its focus from treatment to prevention, this actionable, holistic, and easily interpretable metric of wellness can be a valuable tool.

3.6 Funding

This paper was published as part of a supplement sponsored and funded by AARP. The statements and opinions expressed herein by the authors are for information, debate, and discussion, and do not necessarily represent official policies of AARP.

3.7 Acknowledgments

The results presented in this chapter have been released as a paper in collaboration with my co-authors [5]: Noa Rappaport, Laura Heath, Tomasz Wilmanski, Andrew T Magis, Nicholas J Schork, Gilbert S Omenn, Jennifer Lovejoy, Leroy Hood, Nathan D Price. The authors acknowledge important contributions to this project made by Scott Lundberg, James Yurkovich, and Matt Conomos. The authors also gratefully acknowledge the participants that consented to share their deidentified data for this study. J.C.E., N.D.P., and L. Hood

conceived and directed the study. J.C.E., N.R., L. Heath, and T.W. performed the analyses and generated figures and tables. All authors drafted and revised the manuscript and aided in the interpretation of the results.

3.8 Conflict of Interest

At the time some of this research was conducted, L. Hood and N.D.P. were co-founders of Arivale (where these data come from) and held stock in the company. N.D.P. was on the board of directors; L. Hood was chair of, and G.S.O. a member of, Arivale's scientific advisory board. A.T.M. and J.L. were employees and had stock options in the company, as did G.S.O. and J.C.E. Arivale is now closed and the authors declare no ongoing competing financial interests.

3.S Supplementary Materials

3.S.1 Supplementary Methods

Clinical labs

Blood and saliva samples were drawn from each subscriber every 189 days on average, and a battery of clinical chemistry measures was conducted using standard procedures. The clinical lab work contained many clinical analytes associated with biological health measures such as cardiometabolic health (including triglycerides, highdensity lipoprotein (HDL), small low-density lipoprotein (LDL) particle number), diabetes (such as insulin, Hemoglobin A1c and fasting glucose), inflammation (such as TNF-alpha, interleukin 6, interleukin-8) and nutrition (including vitamin D (blood 25-dihydroxyvitamin D), copper and ferritin). Two vendors were used for clinical labs (Quest and LabCorp); their measurements are analyzed independently to account for vendor-specific effects. Measurements related to supplements commonly recommended by health coaches or derived from other analytes or age were dropped from the analysis in order to minimize confounding (for instance, individuals with worse clinical health metrics may be more inclined to take supplements). These dropped measures are

detailed in Supplemental Table 3.S5.

Metabolomics

Prior to processing, plasma was stored in a bio-storage facility at -80 C. Frozen plasma samples in anticoagulant Ethylenediaminetetraacetic acid (EDTA) were sent to Metabolon, Inc. to conduct metabolomics assays. Data were generated using the Metabolon HD4 discovery platform, a combination of ultra high-performance liquid chromatography (HPLC) tandem mass spectrometry (MS) and gas chromatography (GC) for identification of metabolites and fatty acids. Relative concentration values were reported for over 700 different metabolites, while the platform itself has the potential to measure up to 2200 unique metabolites though many remain unidentified. Existing metabolomics samples were run in several batches. Between four and sixteen previously generated pooled control samples were run with each batch and used for batch correction. Roughly 1300 metabolomics samples were used in the analyses for this paper. KEGG pathway associations for each identified metabolite were provided by Metabolon. Measurements related to supplements commonly recommended by health coaches or derived from other analytes or age were identified in the Clinical Labs, and metabolites strongly correlated (r2 $>$ .05) with these measures were dropped from the analysis in order to minimize confounding (for instance, individuals with worse clinical health metrics may be more inclined to take supplements). These dropped measures are detailed in Supplemental Table 3.S5.

Proteomics

Proteomics analysis was performed on EDTA-anticoagulated plasma extracted from whole blood using Olink's proximity extension assay panels, including Cardiovascular II (http://www.olink.com/products/cvd-iipanel/), Cardiovascular III (http://www.olink.com/products/cvd-iii-panel/), and Inflammation (http://www.olink.com/products/inflammation/); 92 proteins are measured on each panel. Prior to processing, plasma was stored in a bio-storage facility at -80 C. Existing proteomics samples were run in several batches. Two

control samples were run with each batch. Batch correction was performed using median centering. 10 proteins were shared by at least 2 panels, providing another level of internal control.

Genomics

Whole genome sequencing was performed on 2806 participants. All whole genome sequencing was performed on DNA extracted from whole blood with library preparation using the Illumina TruSeq Nano Library prep kit and sequenced using Illumina technology (Illumina HiSeq X, PE-150, target 30X coverage) at a single CLIA-approved sequencing laboratory. Raw sequencing data were processed using a consistent bioinformatics pipeline, including BWA 0.7.12 for alignment to reference sequence hg19 and duplicate marking with biobam2 2.0.70. Variant calling was performed using GATK best practices for whole genome data with GATK 3.3.0, including indel local realignment followed by base quality recalibration. VCF files were produced by GATK HaplotypeCaller followed by GenotypeGVCFs. CNV calling from WGS data was performed using a bioinformatics pipeline on BAM files using CNVnator 0.3. Polygenic score computation and ancestry estimates were performed using the bioinformatics pipeline. 814 participants were genotyped using the Illumina Multi-Ethnic Global SNP Array at a single CLIA-approved lab. This array consisted of roughly 1.8 million variants. An additional roughly 38 million variants were imputed using the Haplotype Reference Consortium (HRC) panel as part of the bioinformatics pipeline. 56 individuals were missing genomic information; their genetic components were mean imputed. Genetic ancestry was represented by principal components (PCs) 1-7 from an analysis of 107,280 linkage disequilibrium pruned autosomal SNPs with minor allele frequency $> 5\%$ using the combined PC-AiR(1) and PCRelate(2) approach as described by Conomos et al. (3).

Lifestyle Information (Quantified self)

Health history and behavioral assessments were performed at baseline and then every 6 months to obtain self-reported data on health status, including (but not limited to): to-

bacco and alcohol consumption; past and/or current incidence of multiple health outcomes (including cancer, cardiovascular and metabolic diseases, infections, respiratory diseases, mental health issues such as depression and anxiety, and others), family history of health outcomes (maternal, paternal, and sibling); and self-reported use of prescription drugs and nutritional supplements.

Data processing

Proteomics, clinical labs, and metabolomics data were measured from the blood at the same blood draw for each participant. Analytes that were missing in more than 20% of the samples were removed from the analysis. Observations missing more than 10% of the remaining values were removed from the analysis. In order to minimize the effect of outliers, values greater than 3 standard deviations from the mean were iteratively shrunk to be within 3 standard deviations from the mean. Analytes that were calculated from other analytes, or were partially calculated using participant age (such as estimated Glomerular Filtration Rate), or were measures of values directly targeted by wellness coaching, were removed from the analysis (see Supplementary Table 3.S5). Mean imputation was performed on the remaining missing data values. To account for variation in populations, the first 7 principal components (calculated using the method of Conomos, et al. (3) from each participant?s genetic profile were added to each observation. All baseline ages were rounded to birth year, with age at observation being that rounded age plus the number of days in the wellness program at the time of the blood draw.

Demonstration of Equivalence of PCA KDM and reported analyte specific effect sizes

The Klemara-Doubal Algorithm estimates the m-dimensional vectors of slopes(k), intercepts (q), and standard deviations (s) of each element of the m-dimensional input vector (y), and uses these parameters to estimate biological age. In this study, the input vector (y) is a PCA transformation (the nxm matrix W) of the original data vector (x), as represented in equation 3.2. The Biological Age estimate (BAE) from the y vector is computed by

equation 3.3. Here we demonstrate that the summed effect sizes for each original analyte in data vector x are equivalent to the originally estimated BAE from the PCA transformed data equations(Equations 3.2-3.11).

Variables from PCA transform

$W -$ an mxn matrix for the linear transformation from R^n to R^mlearned via PCA

$x -$ Original data vector of size n

$y = W^T x -$ A PCA transformed vector of size m used in KD

Parameters learned from Klemera-Doubal on PCA transformed data.

$k -$ a vector of m slopes

$q -$ a vector of m intercepts

$s -$ a vector of m standard deviations

$$y_i = \sum_{j=1...n} W_{ij}^T x_j \tag{3.1}$$

$$BA_E = \frac{\sum_{i=1...m}(y_i - q_i)\frac{k_i}{s_i^2}}{\sum_{i=1...m}(\frac{k_i}{s_i})^2} \tag{3.2}$$

$$\text{let } a = \frac{\frac{k}{s^2}}{\sum_{i=1...m}(\frac{k_i}{s_i})^2}, \text{ for simplicity such that} \tag{3.3}$$

$$BA_E = \sum_{i=1...m}(y_i - q_i)a_i \tag{3.4}$$

$$= \sum_{i=1...m}(y_i a_i - q_i a_i) \tag{3.5}$$

$$= \sum_{i=1...m} y_i a_i - \sum_{i=1...m} q_i a_i \tag{3.6}$$

$$= \sum_{i=1...m}(\sum_{j=1...n} W_{ij}^T x_j)a_i - \sum_{i=1..m} q_i a_i \tag{3.7}$$

$$= \sum_{j=1...n} x_j \sum_{i=1...m} W_{ij}^T a_i - \sum_{i=1..m} q_i a_i \tag{3.8}$$

$$= \beta_0 + \sum_{j=1..n} \beta_j x_j, \text{ where} \tag{3.9}$$

$$\beta_j = \sum_{i=1...m} W_{ij}^T a_i \text{ and} \tag{3.10}$$

$$\beta_0 = -\sum_{i=1..m} q_i a_i \tag{3.11}$$

3.S.2 Supplementary Figures

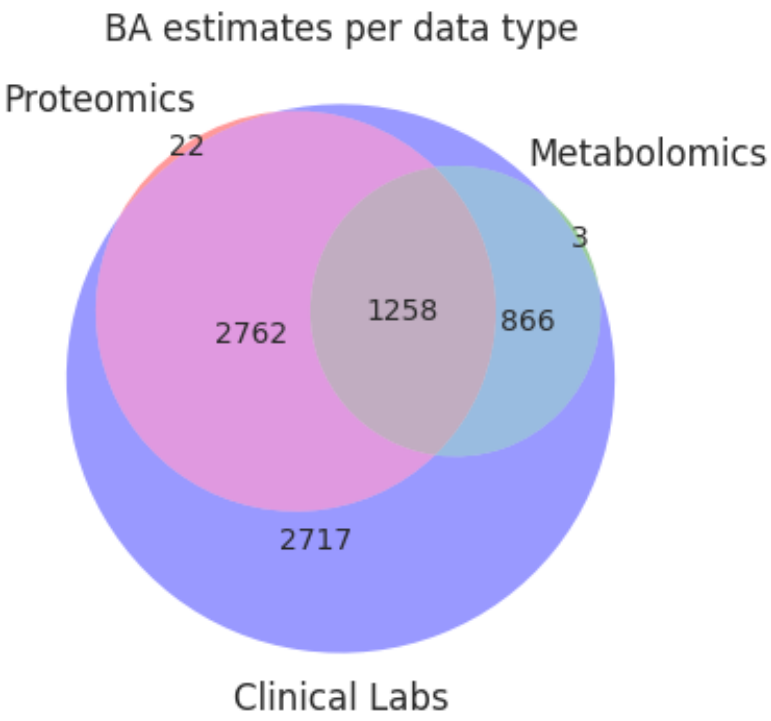


Figure 3.S1: BA estimates per data type

Venn diagram showing the number of observations per combination of data types.

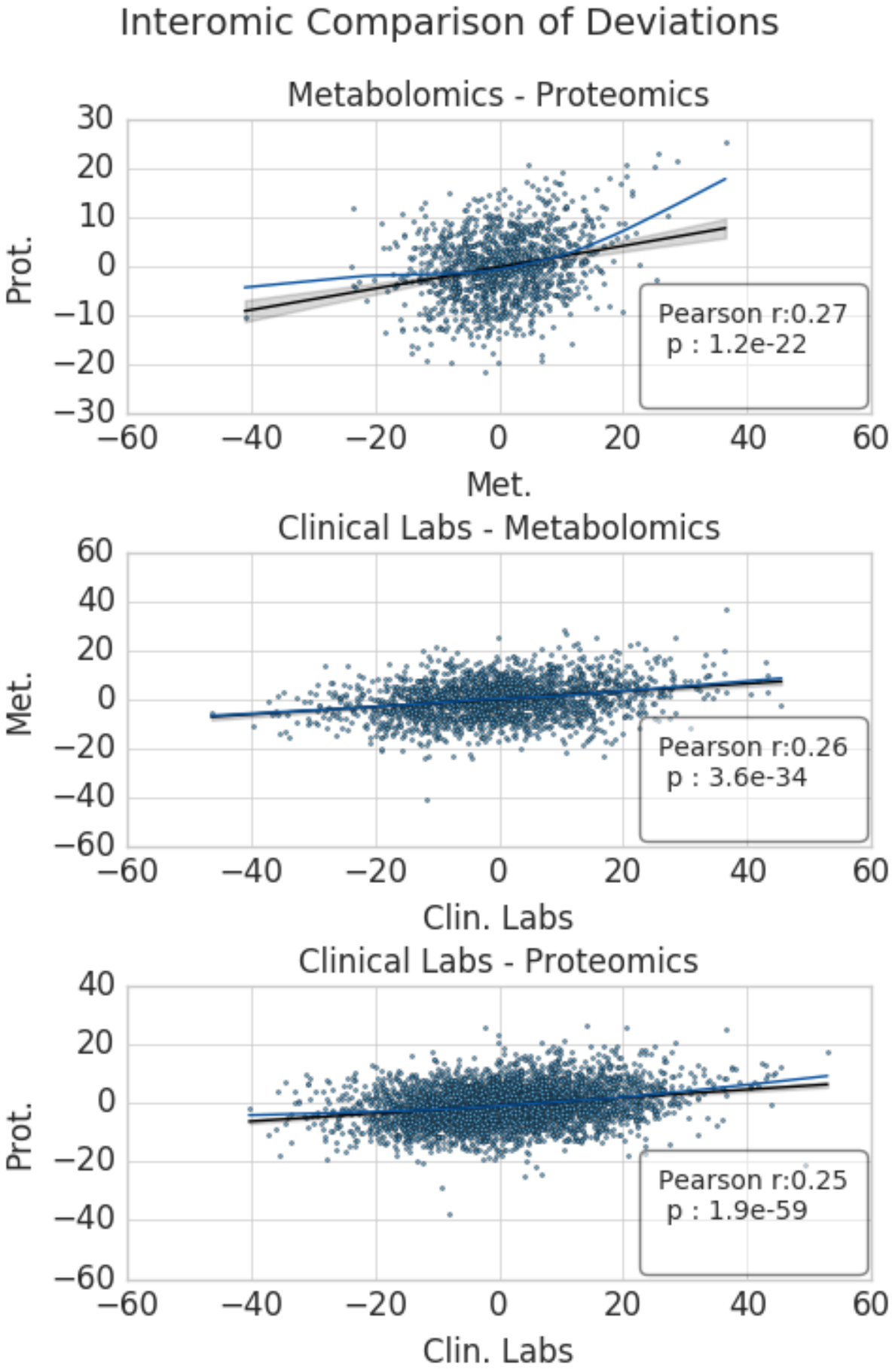


Figure 3.S2: Interomic comparison of deviations

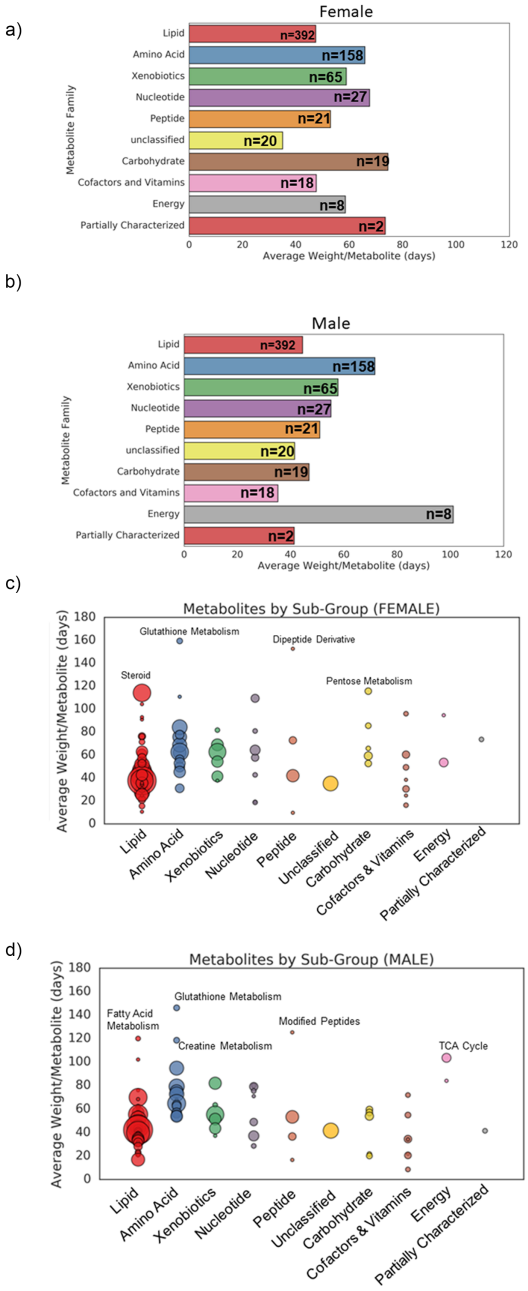


Figure 3.S3: Contribution of metabolite families and subfamilies to BA predictions for males and females, measured by average weight per metabolite in each group

Average weights are expressed in days, corresponding to the average change in BA (positive or negative) for 1 standard deviation change in metabolite concentration. a) and b) The average contribution of each metabolite family in predicting BA for males (a) and females (b). The absolute value of the mean ?-coefficient for each metabolite obtained using the KD algorithm across the cross-validation procedure was summed within each metabolite family and divided by the total number of metabolites for that family. The number of metabolites in each family is noted. c) and d) The same analysis was performed as in figures a) and b), but at the level of metabolite subfamilies. The y-axis corresponds to the mean contribution (positive or negative) of each metabolite subfamily to the overall BA prediction. Each subfamily data point is further sized by the number of metabolites measured in that subfamily in our metabolomics panel.

3.S.3 Supplementary Tables

Table 3.S1: Baseline self-reported characteristics of the wellness program sample

Characteristic	Total N=3558	Women N=2087	Men N=1471	P-value
Past and/or current self-report of:				
Migraine, no.(%), n=3273	774(23.6)	605	169	<0.001
High cholesterol, no.(%), n=3351	788(23.5)	408	380	<0.001
Depression, no.(%), n=3312	750(22.6)	550	200	<0.001
Gastroesophageal reflux disease, no.(%),	619(19.0)	381	238	0.4
Hypertension, no.(%), n=3361	579(17.2)	313	266	0.003
Asthma, no.(%), n=3389	559(16.5)	370	189	<0.001
Lung infection, no.(%), n=3265	501(15.3)	352	149	<0.001
Eczema, no.(%), n=3269	468(14.3)	337	131	<0.001
Colon polyps, no.(%), n=3285	458(13.9)	261	197	0.3
Osteoarthritis no.(%), n=3398	406(11.9)	292	114(8.3)	<0.001
Sleep apnea, no.(%), n=3295	392(11.9)	179(9.1)	213	<0.001
Leiomyoma(fibroids), no.(%), n=1960b	332(10.2)	329	3(2.3)	<0.001
Concussion, no.(%), n=3265	330(10.1)	182(9.4)	148	0.1
Irritable Bowel Syndrome, no.(%), n=3225	300(9.3)	227	73(5.7)	<0.001
Breast lump, no.(%), n=3239	299(9.2)	290	9(0.7)	<0.001
Endometriosis, no.(%), n=1945b	162(8.3)b	162(8.3)	NA	NA
Enlarged prostate, no.(%), n=1315c	100(7.6)c	NA	100(7.6)	NA
Kidney stones, no.(%), n=3239	193(6.0)	94(4.9)	99(7.6)	0.002
Gallstones, no.(%), n=3228	186(5.8)	158(8.1)	28(2.2)	<0.001
Cataracts, no.(%), n=3315	180(5.4)	113(5.7)	67(5.0)	0.4
Psoriasis, no.(%), n=3249	154(4.7)	90(4.6)	64(4.9)	0.8
Post-traumatic stress disorder, no.(%),	149(4.6)	122(6.3)	27(2.1)	<0.001
Thyroid nodules, no.(%), n=3257	143(4.4)	129(6.6)	15(1.1)	<0.001
Peptic ulcer, no.(%), n=3275	134(4.1)	87(4.5)	47(3.6)	0.2
Diverticulosis, no.(%), n=3227	129(4.0)	72(3.7)	57(4.4)	0.4
Type 2 Diabetes, no.(%), n=3309	125(3.8)	78(4.0)	47(3.5)	0.6
Rheumatoid arthritis, no.(%), n=3390	100(2.9)	61(3.0)	39(2.9)	0.9
Gout, no.(%), n=3233	90(2.8)	24(1.2)	66(5.1)	<0.001
Chronic fatigue syndrome, no.(%), n=3237	78(2.4)	65(3.4)	19(1.4)	0.001
Hyperthyroid, no.(%), n=3259	78(2.4)	63(3.2)	15(1.1)	<0.001
Blood clots, no.(%), n=3213	70(2.2)	41(2.3)	29(2.3)	0.9
Fibromyalgia, no.(%), n=3212	68(2.1)	59(3.1)	9(0.7)	<0.001
Breast cancer, no.(%), n=3235	63(1.9)	59(3.0)	4(0.3)	<0.001
Epilepsy, no.(%), n=3230	55(1.7)	34(1.8)	21(1.6)	0.9
Coronary artery disease, no.(%), n=3280	50(1.5)	19(1.0)	31(2.3)	0.003
Celiac disease, no.(%), n=3234	45(1.4)	35(1.8)	10(0.8)	0.02
Bipolar disorder, no.(%), n=3381	38(1.1)	15(0.7)	23(1.7)	0.07
Glaucoma, no.(%), n=3231	38(1.1)	19(1.0)	16(1.2)	0.6
Graves disease, no.(%), n=3204	32(1.0)	27(1.4)	5(0.4)	0.008
Heart attack, no.(%), n=3301	31(0.9)	16(0.8)	15(1.1)	0.5

a) Obese defined as BM≥30 b) Evaluated in women only c) Evaluated in men only

Table 3.S2: Detailed Prediction Statistics by data type

	C.L.(Quest)	C.L.	C.L.(Labcorp)	Prot. or Met. or C.L.[a]	Met.	Prot.
n Ind.	581	3553	3503	3558	1631	2162
n Obs.	783	7603	6820	7634	2133	4048
n Females	307	2083	2057	2087	968	1259
n Males	274	1470	1446	1471	663	903
MAE	8.15	8.04	8.03	5.54	4.81	4.39
RMSE	12.14	12.19	12.19	9.32	7.48	6.61
Pearson r of BA and CA	0.67	0.7	0.7	0.78	0.81	0.88
Pearson p of BA and CA[b]	1.32E-101	NA	NA	NA	NA	NA
Pearson r of delta BA and CA	-0.12	-0.03	-0.02	-0.06	-0.18	-0.10
Pearson p of delta BA and CA	7.27E-04	2.46E-03	4.16E-02	2.03E-08	2.63E-17	1.12E-09
Mean SD of repeated predictions(10x)	2.3	1	0.9	3.8	1.5	1

Cont. on next page

a) the "Overall predictions" b) an NA represents a p-value less than machine precision, at least p<1E-200

Detailed Prediction Statistics by data type(Cont)

	C.L.(Quest)	C.L.	C.L.(Labcorp)	Prot. or Met. or C.L.	Met.	Prot.
SD of SD of repeated predictions(10x)	0.84	0.58	0.32	3.66	0.44	0.31
Mean SD of personal longitudinal predictions	5.27	5.96	5.71	4.53	3.46	3.22
SD of SD of personal longitudinal predictions	4.17	3.87	3.87	3.19	2.70	2.17
Pearson r of delta age in personal longitudinal predictions	0.71	0.67	0.70	0.66	0.64	0.67
Mean days between longitudinal observations	116.8	190.4	199.4	190.1	197.2	166.9
Std days between longitudinal observations	20.0	65.9	65	65.6	44.1	51.3
Mean delta age	-0.59	-0.43	-0.42	-0.79	-0.12	-0.73
Std dev delta age	12.1	12.1	12.2	9.3	7.5	6.6

Table 3.S3: Beta coefficients for stratified analyses by data type.

Stratified analyses[b]	β Coefficient	Std. error	95% CI	Interaction p[c]
PROTEOMICS-DERIVED BA				
Gender				
Males(n=505)	0.762	0.255	0.263, 1.26	
Females(n=544)	0.271	0.249	-0.216, 0.759	0.163
Self-reported ethnicity				
White(n=839)	0.740	0.188	0.372, 1.107	
Non-white(n=145)	-0.509	0.616	-1.715, 0.697	0.059
Age at baseline, by decade				
18-29 years(n=67)	-0.239	0.865	-1.930, 1.460	
30-39 years(n=155)	-0.033	0.519	-1.05, 0.985	
40-49 years(n=295)	0.059	0.366	-0.658, 0.776	
50-59 years(n=292)	0.800	0.342	0.129, 1.47	
60-69 years(n=182)	1.12	0.362	0.414, 1.83	
70 years and over(n=58)	0.849	0.546	-0.221, 1.92	NA
Baseline BA prediction				
BA=5 years > CA(n=227)	-2.31	0.530	-3.347, -1.27	
BA=5 years < CA(n=242)	3.369	0.256	2.87, 3.871	
BA=10 years > CA(n=64)	-3.44	1.037	-5.477, -1.41	
BA=10 years < CA(n=68)	0.97	0.039	0.894, 1.05	NA

Continued on next page

* GEE Model: BA∼time in Arivale + baseline CA; clustered by client ID, family=Gaussian, with an exchangeable correlation matrix; only individuals with at least two visits were included b) GEE Models, stratified by sex, ethnicity, age group, and baseline BA prediction: ΔAge (BA-CA) ∼ time in Arivale + baseline CA; clustered by client ID, family=Gaussian, with an exchangeable correlation matrix; All models use BA predictions based on the "All analyte" data set c) Interaction models: ΔAge (BA-CA) ∼ time in Arivale + predictor variable + baseline CA + predictor variable x time in Arivale; clustered by client ID, family=Gaussian, with an exchangeable correlation matrix.

Beta coefficients for stratified analyses by data type. (cont. Table 3.S3)

Stratified analyses[b]	β Coefficient	Std. error	95% CI	Interaction p[c]
METABOLOMICS-DERIVED BA				
Gender				
Males(n=178)	0.352	0.680	-1.009, 1.658	
Females(n=273)	0.237	0.644	-1.025, 1.498	0.883
Self-reported ethnicity				
White(n=331)	0.23	0.531	-0.81, 1.269	
Non-white(n=120)	0.281	1.003	-1.684, 2.25	0.890
Age at baseline, by decade				
18-29 years(n=24)	-2.38	1.4	-5.13, 0.37	
30-39 years(n=85)	-0.669	1.399	-3.41, 2.07	
40-49 years(n=147)	1.68	0.763	0.185, 3.18	
50-59 years(n=140)	0.678	0.85	-0.987, 2.34	
60-69 years(n=47)	-0.496	1.07	-2.59, 1.59	
70 years and over(n=8)	-5.8	3.19	-12.1, 0.459	NA
Baseline BA prediction				
BA=5 years > CA(n=101)	-4.782	0.971	-6.68, -2.88	
BA=5 years < CA(n=113)	6.348	0.834	4.713, 7.984	
BA=10 years > CA(n=31)	-8	1.415	-10.776, -5.23	
BA=10 years < CA(n=46)	7.348	0.0491	0.845, 1.04	NA

Continued on next page

Beta coefficients for stratified analyses by data type. (cont. Table 3.S3)

Stratified analyses[b] CHEMISTRIES-DERIVED BA	β Coefficient	Std. error	95% CI	Interaction p[c]
Gender				
Males(n=1327)	0.25	0.332	-0.401, 0.901	
Females(n=1881)	-0.75	0.257	-1.254, -0.246	0.021
Self-reported ethnicity				
White(n=2417)	-0.245	0.231	-0.697, 0.208	
Non-white(n=692)	-0.657	0.516	-1.668, 0.355	0.617
Age at baseline, by decade				
18-29 years(n=235)	-2.710	0.798	-4.27, -1.14	
30-39 years(n=622)	0.0355	0.454	-0.854, 0.926	
40-49 years(n=950)	0.024	0.387	-0.734, 0.782	
50-59 years(n=836)	0.161	0.422	-0.655, 0.987	
60-69 years(n=447)	-0.888	0.508	-1.880, 0.107	
70 years and over(n=118)	-0.316	0.791	-1.870, 1.230	NA
Baseline BA prediction				
BA=5 years > CA(n=1007)	-3.88	0.3855	-4.633, -3.12	
BA=5 years < CA(n=1111)	3.088	0.3678	2.287, 3.73	
BA=10 years > CA(n=615)	-4.93	0.452	-5.82, -4.05	
BA=10 years < CA(n=692)	0.983	0.024	0.936, 1.03	NA

Table 3.S4: Coefficient estimates for each analyte by model

This table is available for download at https://doi.org/10.1093/gerona/glz220

Table 3.S5: Dropped values and their reason for exclusion.

This table is available for download at https://doi.org/10.1093/gerona/glz220

Chapter 4

SYSTEMS BIOMARKERS FOR DISEASE FROM PERSONAL DENSE DYNAMIC DATA CLOUDS

4.1 Abstract

This chapter presents a prototypical application of personal, dense, dynamic data clouds (PD3 clouds) to studying disease trajectories, specifically for cancers. Multi-omic quantification of changes to wellness are performed using biological age calculations, and differences in holistic wellness status between stages and -omics profiles are detailed. Breast and ovarian cancer characteristics are compared and explored, with differences in severity and progression examined and quantified. Techniques for identifying disease-relevant features in multi-omic longitudinal data are demonstrated, as are applications of Systems Biology approaches that culminate in the N-of-1 analyses of two cancer patients with possible data-informed adjuvant and alternative therapies hypothesized. This approach is then applied to an individual that was diagnosed with Chronic Lymphocytic Leukemia after a year of observation. Deep-phenotyping revealed that this individual's condition was more likely Hairy Cell Leukemia, a far more treatable condition. Limitations and next steps are explored and proposed as a guide to future attempts to leverage deep-phenotyping and computational approaches towards creating a 21st-century health care system that is personalized, predictive, preventive, and participatory.

4.2 Introduction

Systems Biology approaches to understanding cell function have been successfully applied to diverse omics and phenotypes [114–119]. A tenet of Systems Biology is that the emergent behavior of complex interactions between biological components of shared functionality is

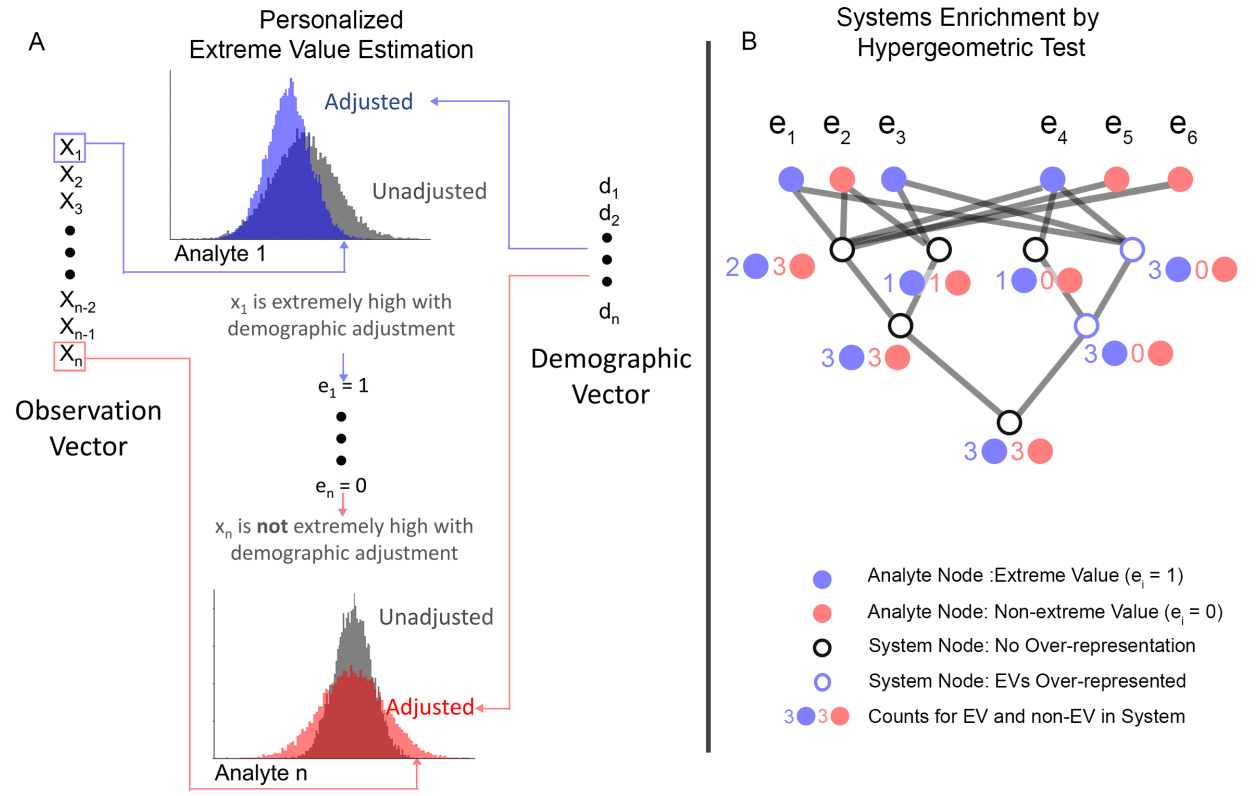
central to biology and that complex phenotypes of common interest, such as disease, are network perturbations. Variations these networks' components, whether brought about by genetics, exposures, or chance, have profound effects on individual health. Systems Biology, therefore, views it as essential that these networks are understood in the context of the individual. Systems Medicine is the extension of that paradigm, where common phenotypes are understood as multi-level perturbations of networks, to the treatment of disease and wellness [120–122].

20$^{\text{th}}$-century medicine treats extreme deviations of a few biomarkers from population average as markers and risk factors for disease, e.g., cholesterol and triglycerides for heart disease, insulin, and glucose for diabetes [123, 124]. The traditional thinking is that these biomarkers must remain few or be condensed into single measures of clinical value to be interpretable and economically scalable, leading to one-size-fits-all treatments, where an individual gets diagnosed with a disease based on the common biomarkers and is treated with the single drug that works best on the average population with that disease [125]. The human mind cannot reason effectively about more than about four variables at a time [126]. Simple explanations of complex phenotypes ignore essential facets of both the occurrence and pathology of disease, e.g., the accompanying inflammatory response in insulin resistance [127]. Systems Medicine proposes incorporating a multitude of biomarkers in the form of inter-omic biological networks, powered by the biochemical data available and the insights developed by Systems Biology into the complex and interconnected workings of the cell [128]. Interpreting this multitude of signals as systems permits deeper reasoning about the phenotype. Systems Biology, as a multi-disciplinary field, also leverages computational approaches, which are free of human limitations in multivariate reasoning. This approach has already been powerfully applied to cancer research and treatment and is poised to become the foundation of 21$^{\text{st}}$-century healthcare [129–132].

In this dissertation, personalized "extreme values" were identified as values from the proteome and metabolome that are greater or less than some percentile compared to an age, sex, and race matched cohort (Figure 4.1A). These extreme values are examined in women under-

going treatment for breast and ovarian cancer. Common and unique extreme values within and between cancer types are discussed, as is their known relevance to their phenotypes. These extreme values are then used as root nodes to select a subgraph from a knowledge graph, which provides functional and ontological relationships between them (Figure 4.1B). These relationships are explored in the context of their disease classes and an N-of-1 context where they are applied to individual patients in two case studies. The analytical system used to perform these analyzes is also presented, highlighting the challenges, opportunities, and needs involved in performing systems analysis of large-scale patient data. The software is freely available under an open-source BSD license. Finally, there is a discussion of challenges, limitations, and future directions for understanding disease with PD3 clouds.

Figure 4.1: N-of-1 analysis overview



Basic overview of analysis. A) Process for generating personalized background distributions for a given analyte from demographic information and subsequent identification of extreme variables. B) Example system network, with extreme values mapped to that system, showing over-representation or not within systems of extreme values based on Hypergeometric Test.

4.3 Methods

4.3.1 Scientific Wellness cohort

The Scientific Wellness cohort (SWC) consisted of men and women participating in a consumer data-intensive wellness program (Arivale, now closed) that varied by age and health status (client characteristics are presented in Table 4.1). The program involved lifestyle coaching on exercise, nutrition, stress management, and sleep, all tailored to the participants' health goals, specific genetic markers, and clinical metrics as detailed in a prior publication [85]. A total of 6,133 (4,881 with blood draws) individuals were observed for an average of 279 days, with an average of 2.3 longitudinal time points and a total of 11,167 observations. In total, 1,974 individuals had one time point, 1193 had two time points, 879 had three time points, and 835 having four or more time points. Four participants had the maximal (10) number of observations. The average time between observations for participants with multiple time points was 197 days. Supplementary Methods 3.S.1 provides a general description of data generation and pre-processing.

4.3.2 CARE cohort

The Tor Biorepository provided dense time-series of 188-time points that ranged from 3 to 24 time points per patient with an average of 9.4 (Supplementary Figure 4.S1). These time series followed 20 women undergoing treatment for ovarian (11) and breast (9) cancers, of which 14 experienced a cancer recurrence. The mean age at diagnosis was 53.9 years.

Stratified patient characteristics are presented in Table 4.2. In this dissertation, each observation in the CARE cohort had a metabolomic panel from Metabolon, consisting of 1296 metabolites, and two proteomic panels (Oncology II and Immuno-oncology) from Olink for a total of 184 proteins. 161 proteins were mapped to existing proteins in the SWC and are included in this dissertation. All included proteins are listed in Supplementary Table 4.S3. All metabolites were used, except where detailed below. Supplementary Table 4.S4 contains a complete list.

Table 4.1: Scientific Wellness cohort characteristics

Characteristic	All	Men	Women
Chronological age, mean years (SD)	46.4 (13.0)	48.4 (12.9)	45.3 (12.8)
Non-white, no. (%), n = 6133	1325 (21.6)	452 (21.1)	873 (21.9)
BMI, mean (SD), n = 5649	27.4 (6.0)	27.5 (4.9)	27.3 (6.5)
Obese, no. (%), n = 5649	1382 (25.0)	457 (23.4)	925 (25.9)
Mod. activity $\geq$ 3 times /wk, no. (%), n = 5095	3424 (67.2)	1374 (70.7)	2050 (65.0)
Vig. activity $\geq$ 3 times /wk, no. (%), n = 5062	1774 (35.0)	816 (42.4)	958 (30.5)
Sitting > 8 h/d, no. (%), n = 5071	3217 (63.4)	1220 (63.1)	1997 (63.6)
Smoke in past, no. (%), n = 2763	709 (25.7)	276 (26.1)	433 (25.4)
Smoke currently, no. (%), n = 5095	235 (4.6)	137 (7.1)	98 (3.1)
Past and/or current report of:			
Cholesterol-lowering meds, no. (%), n = 5149	696 (13.5)	393 (20.4)	303 (9.4)
High cholesterol, no. (%), n = 5026	1060 (21.6)	471 (24.6)	589 (19.6)
Hypertension, no. (%), n = 5026	709 (14.4)	330 (17.3)	379 (12.6)
Asthma, no. (%), n = 5026	794 (16.2)	265 (13.9)	529 (17.6)
Type 2 diabetes, no. (%), n = 5026	166 (3.4)	67 (3.5)	99 (3.3)
Breast cancer, no. (%), n = 5026	99 (2.0)	6 (0.3)	93 (3.1)
Coronary artery disease, no. (%), n = 5026	65 (1.3)	41 (2.1)	24 (0.8)

All characteristics were compared for differences between men and women, using t-tests for real variables and chi-square tests for categorical variables. In all cases, the p-value $< .001$.

Batch correction

Metabolomic assays are run in batches, with values reported relative to a pooled blood sample generated by combining samples from the SWC. Each metabolite from the CARE cohort was normalized to this reference sample as observed quantity over the reference sample quantity as detailed in Section 3.S.1 of this dissertation.

The batch correction method used to create the 'corrected' proteomic version of the data used in this dissertation is detailed in an unpublished whitepaper and summarized here [133]. Eight technical replicates from a pooled blood sample served as overlapping reference samples in each plate. Data were batch corrected by normalizing against the overlapping reference samples within the plate, as described below.

Denote the original protein value for sample i by x_i. Samples belong to exactly one run and plate. Define the set of samples from the pth plate of the rth run by M_{pr}. Define the

Table 4.2: CARE cohort characteristics

Characteristic	All Patients	Breast Cancer	Ovarian Cancer
Patients, no.	20	9	11
Age at diagnosis, mean years (SD)	53.9 (8.3)	49.0 (6.4)	57.8 (7.6)
Observations per patient, mean (SD)[a]	9.4 (5.5)	7.4 (3.7)	11.0 (6.4)
Progression free interval, mean mos. (SD)	36.5 (35.3)	59.3 (38.7)	17.9 (17.9)
Follow-up status (expired), no.	9	2	7
Stage, no.:			
I	2	2	0
IIA	2	2	0
IIB	3	2	1
IIC	1	0	1
IIIA	3	2	1
IIIC	7	1	6
IVA	2	0	2

Overview of patient statuses of the CARE cohort acquired from the Tor Biorepository. a) Total observations is 188 time-points.

set of reference samples by R. The corrected protein value for ith sample in M_{pr} is defined as follows,

$$x'_i = x_i - median_{i \in M_{pr} \cap R} + offset$$

where $offset$ is the median value of the non-reference samples from a reference run of earlier samples. Intuitively, since the original protein values are on the log2 scale, the corrected values are scale-shifted to the reference samples and the original delivered data using the reference run.

Biological age estimation of CARE cohort

The CARE cohort's change in general wellness was assessed via Biological Age Estimation in the manner described in Chapter 3. Using the SWC as the training set, Biological Age Estimates (BA_E) were calculated. Metabolomic BA_E was generated using 862 shared metabolites from 977 women who identified themselves as 'white.' Proteomic BA_E was generated from 161 proteomic measures shared with 354 clients of varying gender and race. The training set for this analysis is much smaller and less specific than the metabolomic train-

ing set due to a greater number of missing values in the overlapping proteins. Any SWC individuals with a cancer diagnosis were excluded. Any SWC individuals missing $> 10\%$ of these proteomics values were excluded from the training set. This was generally due to most individuals in the SWC not having the Oncology II panel. Any remaining missing values were median imputed. Only baseline observations were used to train the models to minimize confounding from subsequent wellness interventions. In contrast to the methods described in Chapter 3, as of the time of this analysis, genetics of the CARE cohort were not available and, therefore, not included in these calculations.

See Supplementary Methods 3.S.1 for a description of data acquisition, pre-processing, and training methods.

4.3.3 Extreme value identification

Extreme values were identified by selecting a subset of the SWC as a representative background set (S) to compare against each observation in the CARE cohort. Each analyte shared with the background set was then compared to the observed value (x_i) and was chosen as being extremely high (e_i^h) when its percentile relative to the background set $\phi_{S_i}(x_i)$ was above some threshold, or extremely low (e_i^l) when its percentile was below some threshold(l).

Precisely, an analyte i from observation x was extremely high if,

$$e_i^h = \mathbb{1}\{\phi_{S_i}(x_i) > u\}$$

or extremely low if,

$$e_i^l = \mathbb{1}\{\phi_{S_i}(x_i) < l\}$$

or simply extreme e where:

$$e_i = e_i^h \vee e_i^l.$$

for every analyte $i \in A$, where A was the set of all possible analytes.

The selection of S is critical in this process and essential to these values' personalization. In this dissertation, background set S was chosen by selecting sex and race matched individ-

uals in the SWC, and selecting the k nearest individuals in age to the patient, generating a ϕ_S unique to each individual.

4.3.4 Systems enrichment

Each analyte is represented by a vertex in a Knowledge Graph (KG) that provides information about the relationship of that measure to various biological entities and processes. This KG was drawn from numerous publicly available databases, including Kegg, HMDB, Chemical Ontology, Uniprot, Gene Ontology, LOINC, and information and relationships provided by test vendors [134–139].

While the complete KG may have cycles, queries here were restricted to predefined directed acyclic subgraphs of the KG(KG'). An example of a possible KG' would be the subgraph containing all paths of type (protein, Gene Ontology entity, Gene Ontology entity(*)), or in other words, a graph that connects proteins to their related Gene Ontology concept and follows the ontology concepts until it reaches a base concept. Supplementary Figure 4.S2 presents all available named paths and the number of entities connected by each path.

From KG', two subgraphs are generated, a background subgraph G, which contains the nodes ($N_{original}$) that mapped to the initially considered set of analytes and their descendants. The second subgraph g is a subset of G which contains the nodes that map to the extreme analytes ($N_{extreme}$) and their descendants in KG'

For each vertex $\{v \in g\}$ a set enrichment score is calculated using the Hypergeometric test[140]. The Hypergeometric distribution is a discrete distribution that measures the probability of k successes in n draws from a finite population of size N that contains K possible successes:

$$f(x; n, N, K) \sim HG(n, N, K);$$

In this instance k is the number of extreme analytes ($N_{original}$) that are ancestors of v. n is the size of $N_{extreme}$, N is the size of $N_{original}$. Finally, K is the number of ancestors of

v in G that are in $N_{original}$. The hypergeometric test for over-representation calculates the probability of collecting k or more successes from n draws given N and K,

$$p_{over} = \sum_{i=k}^{K} f(i; n, N, K)$$

Given the large number of hypotheses generated by this procedure, it is important to correct for multiple hypotheses, which is done by applying the method of Benjamini-Hochberg, and Bonferroni [15, 141].

4.3.5 Software

The software for performing the analysis is a python-based cloud-hybrid system developed by the author of this dissertation called the Scalable Retrieval Extraction Correlation and Ontologization Workflow or Scarecrow. This system is backed by several databases (AWS DynamoDB, Neo4j, Redis) and can run analyses as asynchronous tasks using the Celery distributed task queue within a Docker Swarm service [142–145]. It maintains the Knowledge Graph in Neo4j. This knowledge base is linked to the PD3 cloud analytes and provides interrelationships, ontological information, and metadata for the diverse multi-omic data present. Scarecrow provides functions for standard Extraction, Transform, and Load operations and provides distributed data management securely. The following sections detail the tools, processes, and architecture of Scarecrow, with a focus on the methods used for to manage the data securely, distribute tasks scalably, manage metadata, access and maintain the Knowledge Graph, calculate extreme values, perform systems enrichment, and tools produced to analyze and present the results of these analyses.

Secure data management

In developing a system to query and analyze sensitive personal health information, it is essential to provide secure data management[146]. The current legal definition of what constitutes health information and personally identifiable information in the context of health

in the United States is presented in the Health Insurance Portability and Accountability Act of 1996, with updates from the HITECH Act provisions of the American Recovery and Reinvestment Act of 2009. These restrictions are primarily concerned with the confidentiality of the patients' healthcare information. The confidentiality of patient records is highly regulated.

HIPAA has a well-defined set of variables regarding what it considers to be personally identifiable information ([147]). Health organizations that want to share information acquired during the course of care freely must remove these identifiers before sharing that information, creating what is known as a "Limited Data Set". All data analyzed in this chapter qualifies as "Limited Data Sets."

No restrictions on the use or disclosure of de-identified health information exist. According to HIPAA, de-identified health information neither identifies nor provides a reasonable basis to identify an individual. There are two legally acceptable ways to de-identify information; either: (1) a formal determination by a qualified statistician; or (2) the removal of specified identifiers of the individual and the individual's relatives, household members, and employers is required and is adequate only if the covered entity has no actual knowledge that the remaining information could be used to identify the individual.

The data analyzed here have been de-identified to conform with HIPPA rules on personally identifiable information[147]. To prevent data leakage of primary or secondary results, as much of the analysis done through Scarecrow is intended to occur on ephemeral virtual machines and scale across diverse clusters, all individual-related data is transferred as encrypted files.

The "CachedFileManager" maintains a database of file descriptions and a set of tools for managing the secure distribution and retrieval of source and derived data files. The primary repository for metadata about "CachedFileRecords" is a DynamoDB database. A unique name keys each CachedFileRecord that can be autogenerated from an MD5 hash of the record content. The CachedFileManager maintains queryable metadata on each CachedFileRecord, e.g., tags, date created, date modified, expiration date, description, etc.

The data file for each CachedFileRecord is maintained in an Amazon S3 bucket. Before the data file is sent to S3, it is encrypted using AES encryption via the pycrypto package [148, 149]. Each data file is encrypted in 1-megabyte chunks to allow large data files to be encrypted without requiring enormous amounts of memory and is seeded with a random IV, thereby preventing identity attacks where two files containing the same information can be identified. Each file is encrypted by a random key that is then encrypted using the AWS Key Management System (KMS) [150]. The encrypted encryption key is then stored in the CachedFileRecord. This multi-layered security approach means that to read an encrypted file an adversary must access the DynamoDB database to get the key, access the S3 bucket to get the file and access the AWS KMS system in order to decrypt the encryption key. It is possible to prevent the object's future decryption of the object by revoking the AWS KMS key. This key can also be restricted to specific users and teams via the AWS Identity Access Management (IAM) system.

It is possible to encrypt existing local files or any pickleable python object [151]. Provided the AWS IAM user account has the proper permissions, one can simply provide the CachedFileManager with the proper file name, and it will download the encrypted file to a local directory, decrypt the file in memory, and return the object as it was when it was generated. The file is never stored on disk or in the cloud as an unencrypted object.

This system is used not just to store analysis results but also maintains intermediate analysis files providing provenance management over analyses. Finally, this framework is also utilized to allow secure data transfer between distributed components.

Distributed task management

Scarecrow utilizes Celery task management workers to distribute asynchronous tasks easily [144]. Celery utilizes task queues to distribute work across threads or machines. Each unit of work is called a task. Celery workers are distributed processes that monitor the task queues for available tasks. These workers can autoscale the number of subprocesses available to do work, increasing or decreasing as the workload increases and decreases. Each Celery

worker is run as a Docker service that can be distributed across any number of machines, as long as the machine is part of a Docker Swarm and part of the shared docker overlay network. This allows cloud-hybrid processing where some of the processing can be performed on a local network, but in the event of a large processing job, it can easily be scaled to any number of servers on the cloud that only need the proper Docker image. By using multiple workers and brokers, Celery provides Scarecrow high availability and horizontal scaling. Celery requires a message broker to manage the task queues and share messages between clients and workers. Scarecrow uses the Redis in-memory key-value database as a broker[143]. This also allows Scarecrow to use Redis as a distributed in-memory cache for common data access operations, preventing frequent communications to higher latency databases, e.g., DynamoDB.

Many functions in Scarecrow provide a "celery" boolean parameter, which then packages the request, sends it to the message broker, and returns a Celery AsyncTaskResult. That AsyncTaskResult can then be queried to determine if the task has finished, and return the expected result once it has.

When a task has a large data object that needs to be shared, that data object is packaged as an auto-generated and encrypted CachedFileRecord, which is then provided to the Celery task for loading. This allows secure distributed file processing and minimizes the need to replicate data architectures among the distributed components.

Additionally, Celery maintains a task scheduler that allows the scheduling of tasks at regular intervals. This allows system maintenance, data ingestion, and common transformations to occur regularly and at off-peak hours. This task distribution system allowed a task to generate 100 extreme value systems analyses in 14 minutes using 80 processes, rather than the 10 hours required when run under one process. Caching operations are submitted in the background to a low priority queue, allowing new data to be efficiently cached without requiring primary processes to pause, thereby decreasing the apparent overhead incurred by caching operations. One final capability the distributed task management system provides Scarecrow is running multiple queries simultaneously against primary datastores such as DynamoDB and the Neo4j databases. These databases are designed to allow large numbers of

concurrent queries, and the separation of these queries into parallel tasks can dramatically increase throughput.

Metadata management

A key challenge when working with deeply-phenotyped data is organizing the metadata around each measured analyte. Scarecrow has a hierarchical and scalable metadata management system. The key to achieving this functionality is the judicious application of the object-oriented principle of inheritance [152]. A BaseMetaData class contains the shared code for communicating with the DynamoDB, where the records are stored, the Redis database, where records are cached and manages necessary operations like invalidating cached records when the base records are changed. This class also contains functions for access the Neo4j Knowledge Graph, where relationships between an instantiated piece of Metadata can be accessed. The BaseMetaData class provides a common API to accessing the information contained in each instantiated metadata object. Each metadata object is uniquely keyed by its datasource_id, which denotes the class to which it belongs, and its source_key, which uniquely identifies the object within its class. Each instantiable object also maintains a metadata property that wraps all information available to this object. All metadata objects can be converted to JSON to allow the eventual development of a RESTful interface[153].

The class hierarchy is detailed below. Note that (*)- indicates the class is instantiable. Uninstantiable classes allow polymorphic operations on similar objects and the ability to select all objects of a related type.

1. **BaseMetaData** - The base metadata class. This contains most of the functionality and provides a common interface.

    (a) **ColumnMD** - The base class for analytes that have measurements. These primarily serve as a base to link out to more semantically meaningful classes.

        i. **MetabolitesColumnMD(*)** - The source keys are columns from the Arivale metabolite data source.

ii. **ProteinColumnMD(*)** - The source keys are columns from the Arivale proteomic data source.

iii. **ChemistriesColumnMD(*)** - The source keys are columns from the Arivale clinical labs data source.

(b) **ChemistriesMD** - The base class for metadata objects related to clinical chemistries.

i. **ArivaleChemistriesMD(*)** - General information provided by Arivale on chemistries.

ii. **LabcorpChemistriesMD(*)** - Information provided by a blood vendor including units, descriptions, and links to conditions related to the lab.

iii. **LoincChemistriesMD(*)** - Related LOINC objects [139].

(c) **KeggMD** A base class for the many different KEGG data types. These metadata objects are densely linked to each other [134].

i. **KeggCompoundMD(*)**, **KeggDrugMD(*)**, **KeggReactionMD(*)**, **KeggPathwayMD(*)**, **KeggDiseaseMD(*)**, **KeggModuleMD(*)**, **KeggOrthologyMD(*)**, **KeggGenesMD(*)**, **KeggNetworkElementMD(*)**, **KeggReactionClassMD(*)**, **KeggDrugGroupMD(*)**, **KeggBriteMD(*)**, **KeggEnvironMD(*)**, **KeggEnzymeMD(*)**, **KeggGenomeMD(*)**, **KeggGlycanMD(*)**, **KeggVariantMD(*)**

(d) **ProteinMD** - Base class for metadata specific to proteins.

i. **GOMD(*)**- Gene ontology annotations [138].

ii. **UniprotGOMD(*)** - A many-to-many mapping between UniprotMDs and GOMDs.

iii. **UniprotMD(*)** - Uniprot annotations of proteins [137]

iv. **UniprotKeggGenesMD(*)** - A one-to-many mapping from uniprot to KeggGenesMD.

(e) **MetaboliteMD** A base class for metadata on metabolites.

i. **HmdbMD(*)** - Descriptions of metabolites from the Human Metabolite Database [135].

ii. **ChemOntMD(*)** - An ontology of metabolites from Classyfire [136].

The above hierarchy is intended to be extendable to other data modalities, such as microbiome, or genetics, and additional data sources, such as Reactome, the Small Molecule Pathway Database, or the Ensemble gene database [154–156].

An additional advantage of the above architecture is the ability to wrap database-specific identifiers into common functionality. For example, all classes have 'name' and 'description' properties, but the source database may denote these as 'gene_name' or 'def'. This class hierarchy's extensibility allows aliasing these identifiers to provide a consistent interface for commonly needed information, without losing the original format.

Knowledge graph

While each metadata object contains links to its immediate neighbors, recursively querying for these relationships is slow when using a NoSQL datastore such as DynamoDB. To facilitate querying deep linkages between metadata objects, Scarecrow utilizes a Neo4j graph database[142]. This provides an efficient means to select multi-layer relationships between metadata objects. The database is automatically generated from the metadata objects.

A GraphResult object is provided to simplify everyday interactions with the graph database, including the generation of NetworkX graph objects, the preferred in-memory graph analysis package for the Python programming language[157]. The Neo4j database only maintains metadata object identifiers and their relationships, so the GraphResult object provides easy access to each node's metadata objects. As of this dissertation's writing, the Neo4j database contains 300,000 metadata objects as nodes and connects those objects with over 1,500,000 relationships.

To simplify accessing these relationships, common paths such as relationships from proteins to their associated GO ontologies or metabolites to their associated Chemical Ontologies

are provided as prepackaged queries or "named paths." There are currently 30 named paths available where a seed list of metadata objects of interest can be presented, and all associated paths will be returned (Supplementary Figure 4.S2). Custom Cypher queries are also available [158].

Extreme value identification

The calculation of extreme or unusual values is central to this dissertation (Section 4.5.3). Identifying the optimal method for identifying these values for any given use case is an open question. To facilitate future exploration of these methods, a flexible interface was developed to allow the user to change the extreme value identification method. Again, a base class (ExtremeValueBase) maintains common interactions to facilitate database communications, secure intermediate and final result storage (Section 4.3.5), and distributed task processing (Section 4.3.5).

The base class provides an interface that assumes there will be some background data source that will contain values that map to a given observed vector, as well as a mapping data source of relevant metadata associated with each individual in the background data source, such as age, sex, race, etc. It also provides a function to get the extreme values that takes an observed vector, that vectors associated metadata to be used to generate the background distribution from similar individuals in the background dataset and the parameters for cut-offs, directionality, and so on.

As of this dissertation, there are two instantiable sub-classes of the base class. The simplest class uses hard selects to identify a subset of records in the background dataset from client metadata using categorical similarities such as age and sex. These categories are not known beforehand but based on the information provided at the time of the function call. The method used in the reported results provides the above functionality, plus the selection of k-nearest neighbors in age after the hard selection.

After processing, all classes return subsets of measurements from the observed vector that meet the criteria to be considered extreme. This process is made efficient by utilizing

the distributed task processing detailed earlier to allow many concurrent jobs (Section 4.3.5). While this does increase the latency of any given query, this is made up for by the enormous increase in throughput. If latency is the greater concern, then one can simply choose not to use the distributed task processing by setting the celery parameter to False, in which case the analysis will proceed synchronously.

Systems enrichment

The output of the extreme value identification is one or more sets of analytes. In this work, those sets were of analytes that were extremely high, extremely low, and the union of those two sets for a combined set. Additionally, to provide a background distribution of extreme values, the extreme value process is repeated for a user-determined number of random individuals in the background set. The systems enrichment process takes each of those sets and the set of all initial analytes as a background set, and queries against a named path (Section 4.3.5) with the given set as the root. The Knowledge Graph then returns the subgraph that is the result of that query. For each set, every node from the graph gets a count of the number of analytes from the set that are ancestors of that node. Node ancestor counts for each extreme set are then compared to the node ancestor counts for the background set using the hypergeometric test for over- and under-representation. This generates a probability for each node of the likelihood of having as many ancestors as observed or more, given the background set. Those probabilities are adjusted for the large multiple hypotheses number of multiple hypotheses using the method of Benjamini-Hochberg, after filtering sets where it is impossible to get enrichment, such as sets where all members are part of the set [15]. Each enrichment result is returned as a table and the generated subgraph for the extreme value set.

All of these processes are run as distributed tasks. Each observation's procedure is managed as a simple state machine where each step's completion leads to the following step's initiation. Errors are handled by setting the procedure back to the previous state and retrying the procedure. The most common errors are communication errors, and retries with

exponential backoffs tend to resolve these. There is an upper limit to the number of retries allowed that can be set by the user. If a process was unrecoverable, its final state would be 'error' to allow debugging.

The cumulative result of this process is then saved as a CachedFileRecord (Section 4.3.5) for downstream analysis and visualization of the result.

Analysis and interpretation

Scarecrow provides several tools for managing the results of the above analyzes. The primary tool is a class wrapper for each of the above results that provides methods for graphing, and the calculation of basic statistics. Additionally, a collection class is provided that aggregates the set enrichments by the given patient ids to allow characterization and visualization of longitudinal results.

4.4 Results

4.4.1 CARE

Cohort-wide descriptive statistics

Comparing the CARE cohort to an age- and sex-matched background population showed 190 metabolites and 18 proteins to be significantly (after Bonferroni adjustment) differentially abundant using Generalized Estimating Equation models (Figure 4.2). Using a Benjamini-Hochberg cut-off of 0.05, 47 proteins and 360 metabolites were differentially abundant. The top 5 most differentially abundant proteins and metabolites are presented in Tables 4.3 and 4.4. Supplementary Tables 4.S6 and 4.S5 provide full results.

Biological Age

Biological age estimates were made for each time point using metabolomics and proteomics. The breast cancer group showed no increase in their proteomic ΔAge (Figure 4.3b) with a mean of -0.4 years and a standard deviation of 6.8 years, approximately consistent with the

Table 4.3: Top 5 differentially abundant metabolites

Name	Super-Pathway	Sub-Pathway	p-value*
Cysteinylglycine	Amino Acid	Glutathione Metabolism	$< 10^{-16}$
Cystine	Amino Acid	Methionine, Cysteine, SAM and Taurine Metabolism	$< 10^{-16}$
Ornithine	Amino Acid	Urea cycle; Arginine and Proline Metabolism	$< 10^{-16}$
Sarcosine	Amino Acid	Glycine, Serine and Threonine Metabolism	$< 10^{-16}$
Oleoyl-ethanolamide	Lipid	Endocannabinoid	$< 10^{-16}$

The most differentially abundant identifiable metabolites between the whole CARE cohort and the Scientific Wellness cohort.

* - Each analyte was modeled individually, with the analyte as the dependent variable, cancer or no cancer as the independent variable with adjustment for age. This was run only against white females using Generalized Estimating Equation models clustered by client ID with an exchangeable correlation matrix to account for multiple observations from individual clients.

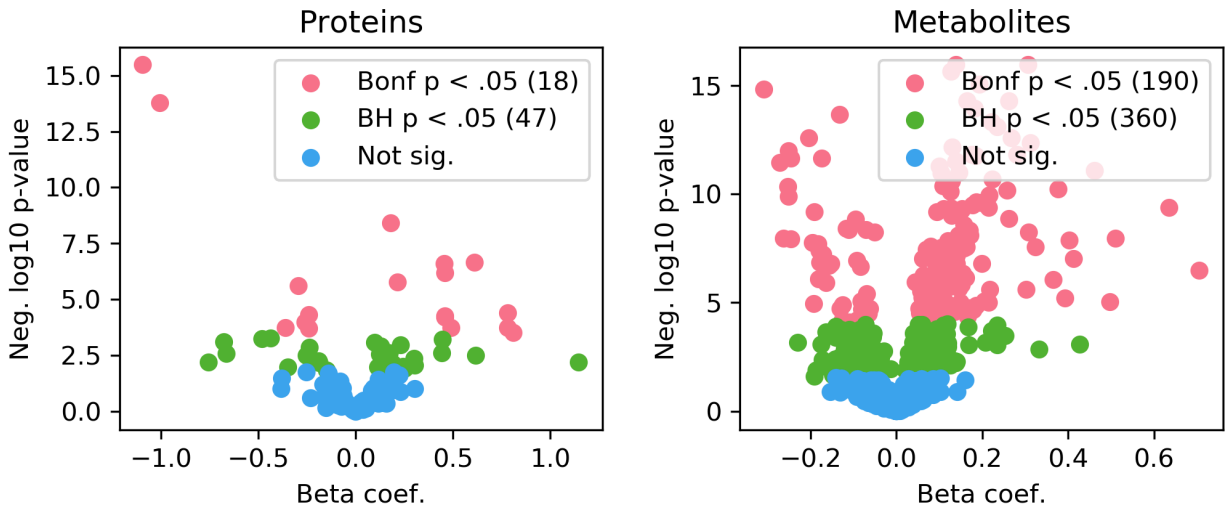
Table 4.4: Top 5 differentially abundant proteins

Gene Name	Description	Uniprot	p-value
VIM	vimentin	P08670	3.3×10^{-16}
CASP8	caspase 8, apoptosis-related cysteine peptidase	Q14790	1.7×10^{-14}
GPNMB	glycoprotein (transmembrane) nmb	Q14956	3.8×10^{-9}
KLK13	kallikrein-related peptidase 13	Q9UKR3	2.3×10^{-7}
CD40	CD40 molecule, TNF receptor superfamily member 5	P25942	2.54×10^{-7}

The most differentially abundant proteins between the CARE cohort and the Scientific Wellness cohort.

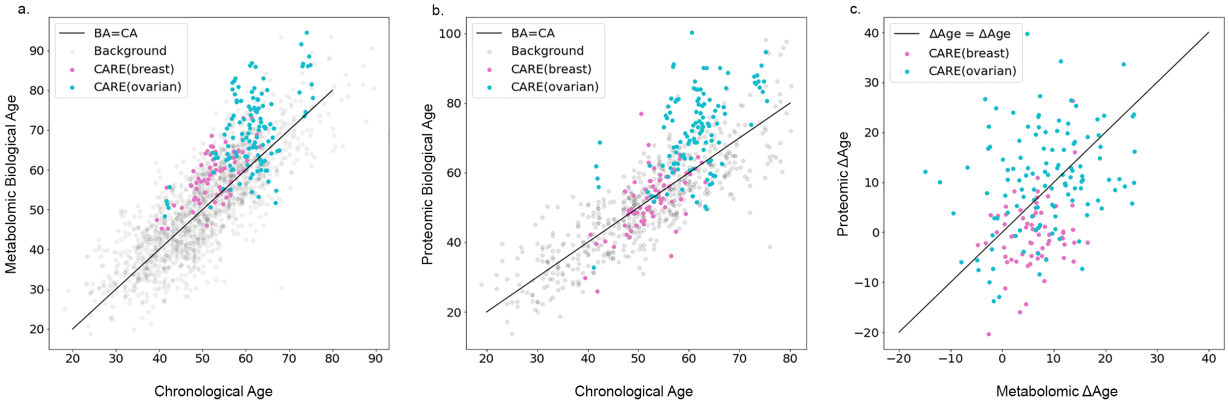
* - Each analyte was modeled individually, with the analyte as the dependent variable, cancer or no cancer as the independent variable with adjustment for age. This was run only against white females using Generalized Estimating Equation models clustered by client ID with an exchangeable correlation matrix to account for multiple observations from individual clients.

Figure 4.2: Differential expression between CARE and Scientific Wellness cohorts



Each analyte was modeled individually, with the analyte as the dependent variable, cancer or no cancer as the independent variable with adjustment for age. This was run only against white females using in Generalized Estimating Equation models clustered by client ID with an exchangeable correlation matrix to account for multiple observations from individual clients.

Figure 4.3: Biological Age estimates with CARE



Biological age estimates of individuals in the CARE cohort. Models were trained on individuals from the Scientific Wellness cohort. Hold-out samples from that cohort are shown in gray in a and b for comparison. a) BA_E using metabolomics. b) BA_E using proteomics. c) Comparison of proteomic and metabolomic ΔAges.

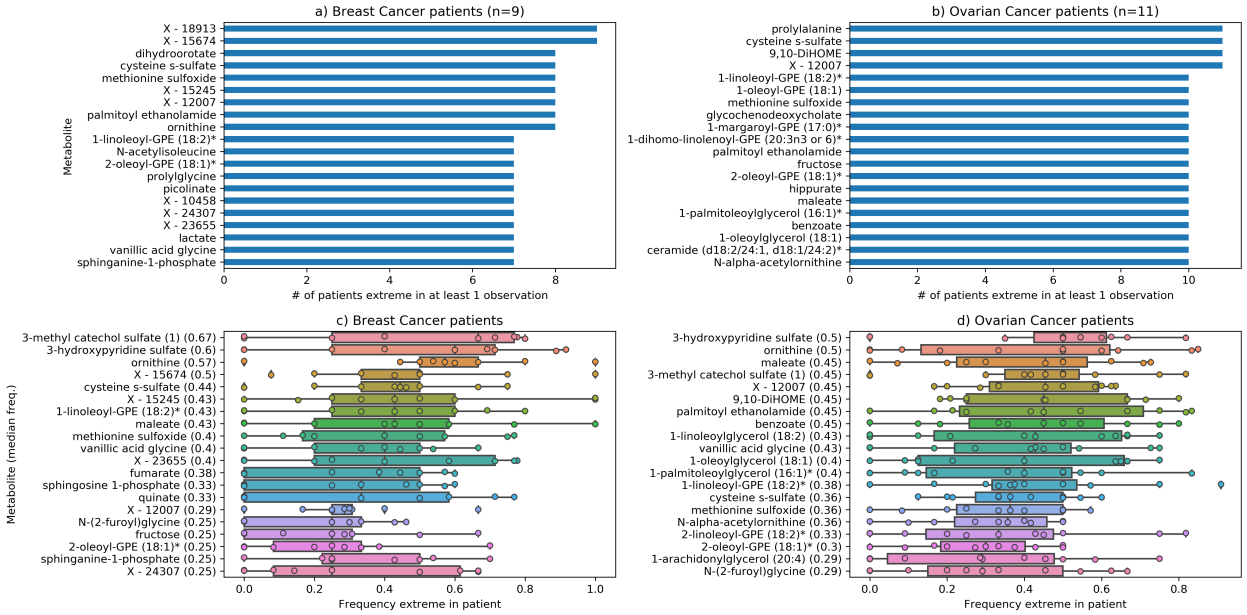
mean ΔAge of the SWC, which was -0.7 years with a standard deviation of 7.6 years. The ovarian cancer group showed a much higher mean ΔAge of +10.5 years, with a standard deviation of 10.0 years. The breast cancer group presented higher metabolomic ΔAges (Figure 4.3a) on average with a mean of +6.0 years and a standard deviation of 4.8 years. Similarly, the ovarian cancer group showed higher metabolomic ΔAges on average with a mean of +8.4 years and a standard deviation of 8.4 years. Metabolomic and proteomic ΔAges were largely uncorrelated under breast cancer with a Spearman r of 0.16 (p-value $<$ .20), weakly positively correlated among the ovarian cancer patients Spearman r of .25 (p-value $<$ .008), and, while quite significant, only marginally correlated over all cancer patients with a Spearman r of .31 (p-value $< 10^{-5}$)(Figure 4.3c). This is consistent with the inter-omic correlations observed in Chapter 3 and Figure 3.S2.

Extreme Values

The most commonly and frequently over-abundant metabolites, at an extreme cut-off of the 99th percentile, among the CARE cohort are presented by cancer type in Figure 4.4. Many similarities exist between extreme analytes in the breast and ovarian cancer cohorts. Eleven of the top 20 most frequently over-abundant metabolites were shared between ovarian and breast cancer; 2-oleoyl-GPE (18:1), X - 12007(unannotated), Vanillic acid glycine, 3-hydroxypyridine sulfate, Methionine sulfoxide, cysteine s-sulfate, 3-methyl catechol sulfate (1), N-(2-furoyl)glycine, ornithine, maleate, and 1-linoleoyl-GPE (18:2). Of the metabolites (with annotations) present in the top 20 most frequently over-abundant breast cancer, but not in the ovarian cancer top 20, Fructose, Quinate, Sphinganine-1-phosphate, Fumarate, and Sphingosine 1-phosphate had median frequencies (rank out of 791) of .25 (24), .18 (38.5), .17 (52.5), .17 (52.5), .10 (83), respectively. Only two unannotated metabolites were never present in the ovarian cancer patients (X - 24307 and X - 15245). Of the metabolites (with annotations) present in the top 20 most frequently over-abundant ovarian cancer, but not in the breast cancer top 20, Palmitoyl ethanolamide was tied for the last value at a median frequency of 0.25. 2-linoleoyl-GPE (18:2)*, 9,10-DiHOME, and N-alpha-acetylornithine were relatively common in breast cancer with median frequencies (rank) of .22 (25.5), .20 (34), and .10 (51.5), respectively. 1-palmitoleoylglycerol (16:1), 1-linoleoylglycerol (18:2), 1-oleoylglycerol (18:1), 1-arachidonylglycerol (20:4), and benzoate were never present as over-abundant in breast cancer.

The most commonly and frequently over-abundant proteins, at an extreme cut-off of the 99th percentile, among the CARE cohort are presented by cancer type in Figure 4.5. In comparison to the metabolites, relatively few proteins were consistently extreme. This is possibly due to the much smaller number of proteins used. It is notable that the breast cancer subclass had very few commonly extreme proteins between patients. Four patients (less than half) had GZMB as extremely over-abundant at some point in their treatment. On the other hand, the ovarian cancer patients had four proteins that were observed in 10

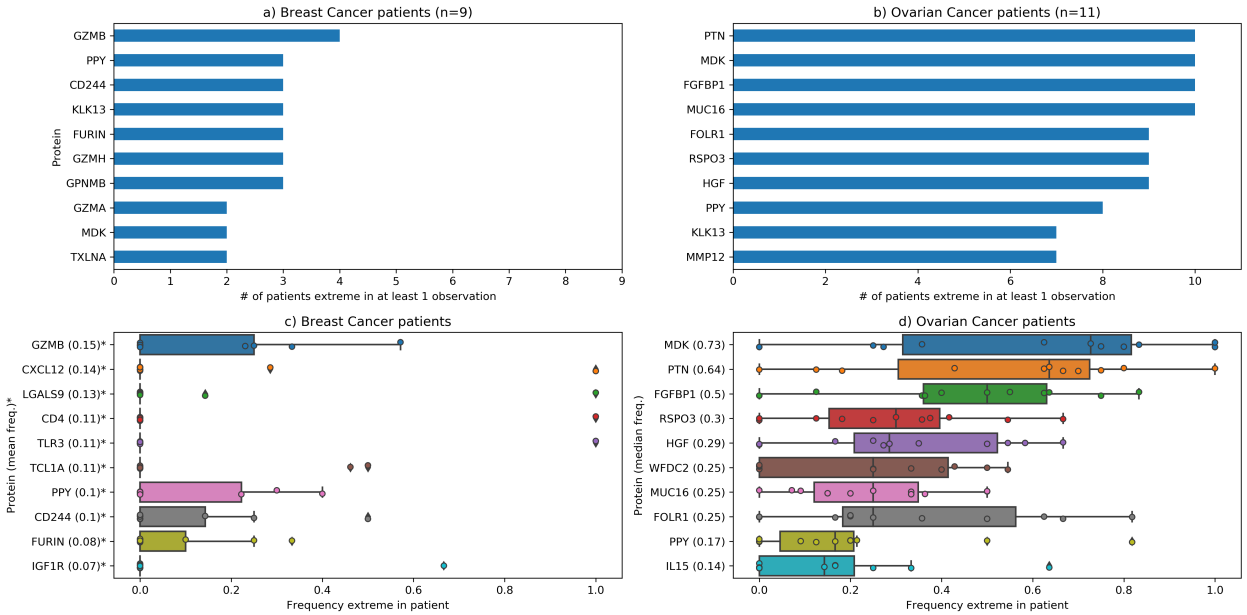
Figure 4.4: Common extremely over-abundant metabolites in CARE cohort



At an extreme cut-off of the 99th percentile, the above metabolites were commonly extreme among individuals. a) and b) present the number of individuals with this metabolite being over-abundant in at least one observation in breast and ovarian cancer patients, respectively. c) and d) present the over-abundance frequency distributions of the most frequently over-abundant metabolites (by median frequency) in breast and ovarian cancer patients, respectively.

* - Unnannoted metabolites are labeled as X - number. These metabolites have not yet been identified.

Figure 4.5: Common extremely over-abundant proteins in CARE cohort



At an extreme cut-off of the 99th percentile, the above proteins (labeled by gene symbol) were commonly extreme among individuals. a) and b) present the number of individuals with this protein being over-abundant in at least one observation in breast and ovarian cancer patients, respectively. c) and d) present the over-abundance frequency distributions of the most frequently over-abundant proteins in breast (by mean frequency) and ovarian (by median frequency) cancer patients, respectively.
* - The frequency of breast cancer patients is ordered by mean rather than median due to no protein having a median frequency greater than zero.

out of the 11 patients as being extreme at some point. Out of the top 10 proteins, three (PPY, KLK13, MDK) were present in both ovarian and breast cancer. Descriptive statistics for extreme value counts by cohort and cut-off are available in Supplementary Table 4.S1.

4.4.2 N-of-1 analyzes

The results of examining three individuals in an N-of-1 fashion are discussed below. Note, these patients were explicitly chosen due to having interesting profiles that exemplified the

goals of performing this analysis. They do not represent the results as a whole, which is addressed in limitations and future work. Full results from all patients are available in Supplementary Data 4.S.1.

CARE patient: 7

CARE patient 7 (CP7) was a white female diagnosed with stage IIIC invasive lobular breast cancer at age 55. She had unknown chemosensitivity status, with a progression-free interval of 35 months and overall survival of 87 months. Her final follow-up status was "alive with breast cancer." The first blood draw available

to this dissertation was at surgery, which has a de-identified age of 56.5, with subsequent blood draws at 57.5, 58.4, and 59.3 years of age for a total of four observations. Her first round of chemotherapy was at 56.6, with hormone therapy at 56.9 years and radiation therapy immediately following at 57.0 years of age. She was diagnosed with recurrence at 59.4 years of age, closely following our final observation.

CP7's initial BA_E was below her Chronological Age (CA) in both metabolites and proteins (Supplementary Figure 4.S6). She experienced a large jump in BA_E in both metabolomics and proteomics following initial treatment, with gradual decrease in BA_E up to her recurrence event. Metabolomic BA_E was much higher than proteomic BA_E at all time points.

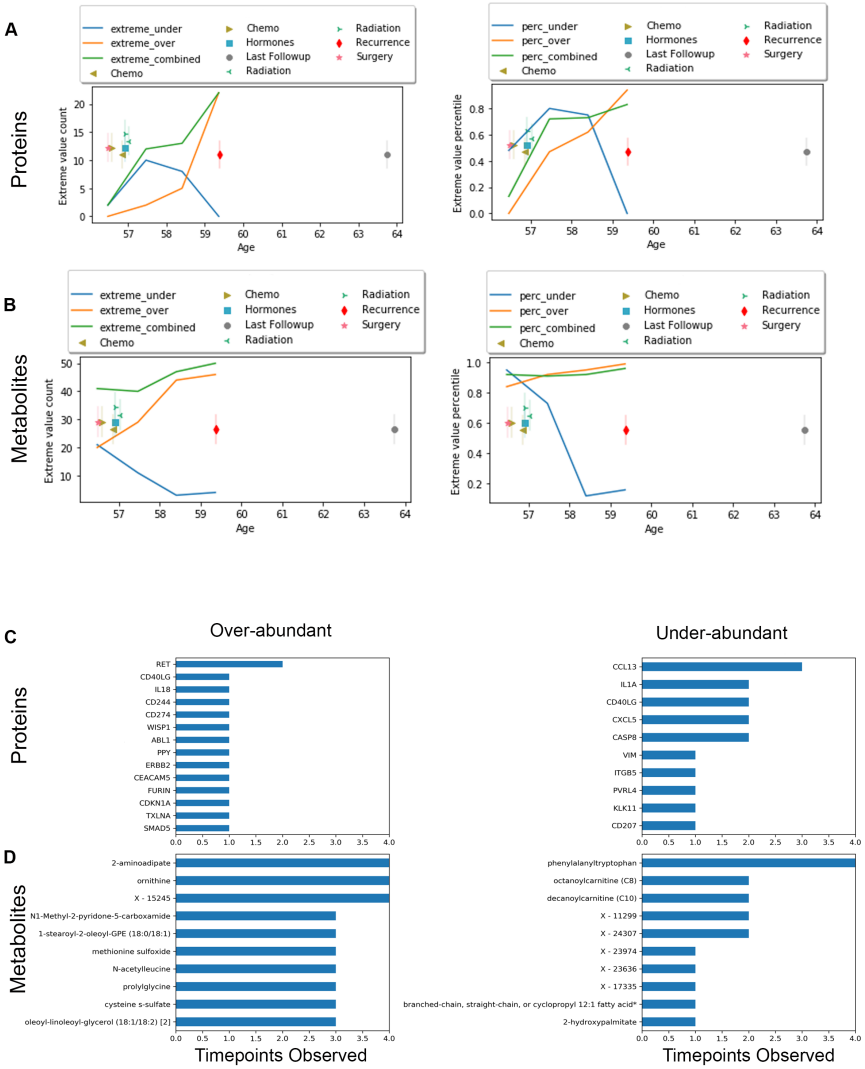
The proteomic signal was generally weaker in the breast cancer patients, with only 13 proteins being extreme at a cut-off of greater than 99th percentile in this individual and none of those being extreme at more than one time point. Therefore, our proteomic analysis of CP7 was restricted to extreme values greater than the 97.5th percentile and less than the 2.5th percentile. Metabolites were examined at the 99th and 1st percentiles.

CP7's initial proteomic extreme value signature contained no extremely over-abundant proteins (Figure 4.6A). This increased mildly a year later, mildly again a year after that until finally there was an extreme jump from 5 to 22 extremely over-abundant proteins at the final observation where she received a recurrence diagnosis. She was initially high in over-

abundant metabolites, and again monotonically increased with the greatest change between the second and third time points (Figure 4.6B). The single protein extreme at more than one time point was RET (Ret Proto-Oncogene) (Figure 4.6C). At all four observations in the metabolites, 2-aminoadipate and ornithine were over-abundant. At three out of 4 time points, CP7 was over-abundant in methionine sulfoxide and cysteine-s-sulfate (Figure 4.6D).

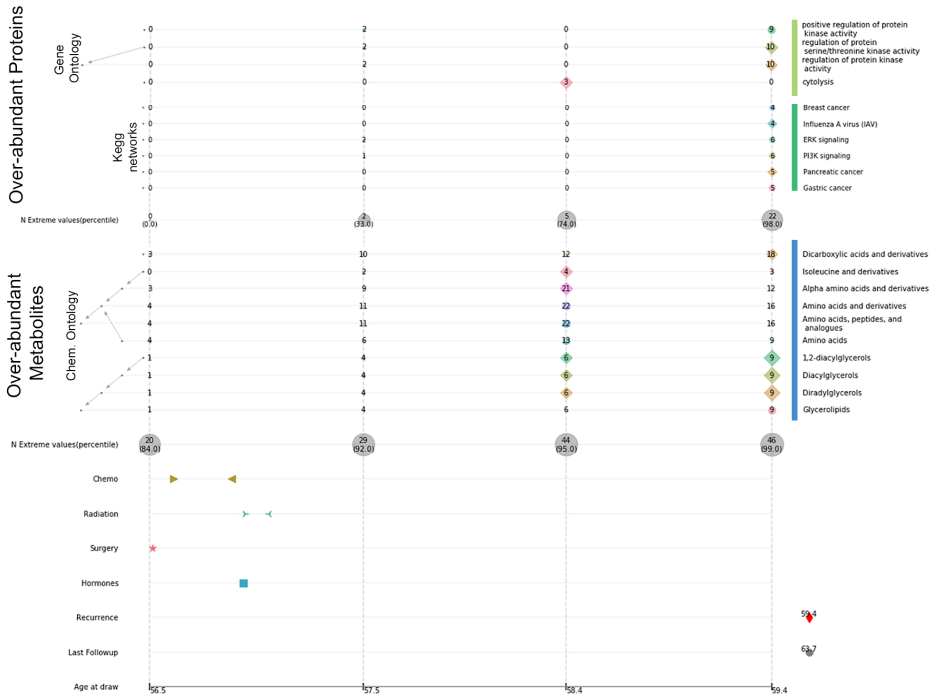
In the systems enrichments for over-abundant proteins via the Gene Ontology, an increasing amount of enrichment was seen as CP7's recurrence event approached. There was an increase in and eventual significant enrichment of the serine/threanine protein kinase activity network. In the Kegg Networks, immediately preceding recurrence, ERK and PI3K signaling is present, as is enrichment for several cancers, including breast cancer. In metabolites, there is a gradual increase in enrichment for amino acids and a monotonic increase in diacylglycerols based on the Chemical Ontology (Figure 4.7).

Figure 4.6: CARE Patient 7: Longitudinal extreme value profile



Extreme value counts and percentiles of those counts relative to the SWC, and the most common extreme values among proteins and metabolites. The extreme value cut-offs presented for metabolites are less than or equal to the 1st percentile for under-abundance and greater than or equal to the 99th percentile for over-abundance.
The extreme value cut-offs presented for proteins are less than or equal to the first 1st percentile for under-abundance and greater than or equal to the 99th percentile for over-abundance.
A and B present the trajectories of extreme values in proteins and metabolites, respectively. C and D show the number of time points a protein or metabolite value was over- or under-abundant in Patient 7 of the CARE cohort.

Figure 4.7: CARE Patient 7: Longitudinal systems



The significantly enriched sets of over-abundant metabolites and proteins for Patient 7 using the Gene Ontology, Kegg Network annotations, and the Chemical Ontology. The number of extreme values belonging to each set are presented. They are marked with a circle if the probability of over- or under-abundance was Benjamini-Hochberg adjusted p-value $< .1$, and are marked with a diamond if they had a Bonferroni adjusted p-value $< .1$. Ontology structure is presented on the left of each graph. Important events throughout treatment are presented on the bottom.
Only over-representation of metabolites and proteins was tested.

Care patient: 10

Care patient 10 (CP10) was a white female, diagnosed with stage IIIC serous ovarian cancer at age 63. She was classified as chemosensitive, with a progression-free interval of 21 months and overall survival of 63 months. The first blood draw available to this dissertation was at surgery, which has a de-identified age of 63.41 years old, with subsequent blood draws at 63.48, 63.8, 64.0, 64.3, 64.5, 64.8, 65.1, 66.1, 66.4, 66.8, 67.0, 67.3, and 67.7 years of age for a total of 14 observations. The first round of chemotherapy was initiated at 63.48 years of age, with subsequent chemotherapy treatments at 65.23, 66, 66.7, 67.0, and 68.5 years. Recurrence was diagnosed at 65.2 years, and she finally succumbed to the disease at 68.6 years.

CP10's initial metabolomic BA_E was slightly below her CA, and her proteomic BA_E was very slightly above her CA. (Supplementary Figure 4.S7). She experienced a massive jump in BA_E in both metabolites and proteins immediately following initial treatment, followed by a gradual decrease in metabolic BA_E up to her recurrence event, where it increased dramatically for two time points, then reached its lowest point and recovered to a level slightly under CA for two time points. After the initial jump in proteomic BA_E, she stayed consistently high until her recurrence event, where she jumped dramatically to a high of almost 120 years. While there was considerable variance, she stayed extremely high in proteomic BA_E, with four observations being over 100 years throughout the rest of the observations. Proteomic BA_E was much higher than metabolomic BA_E at all time points.

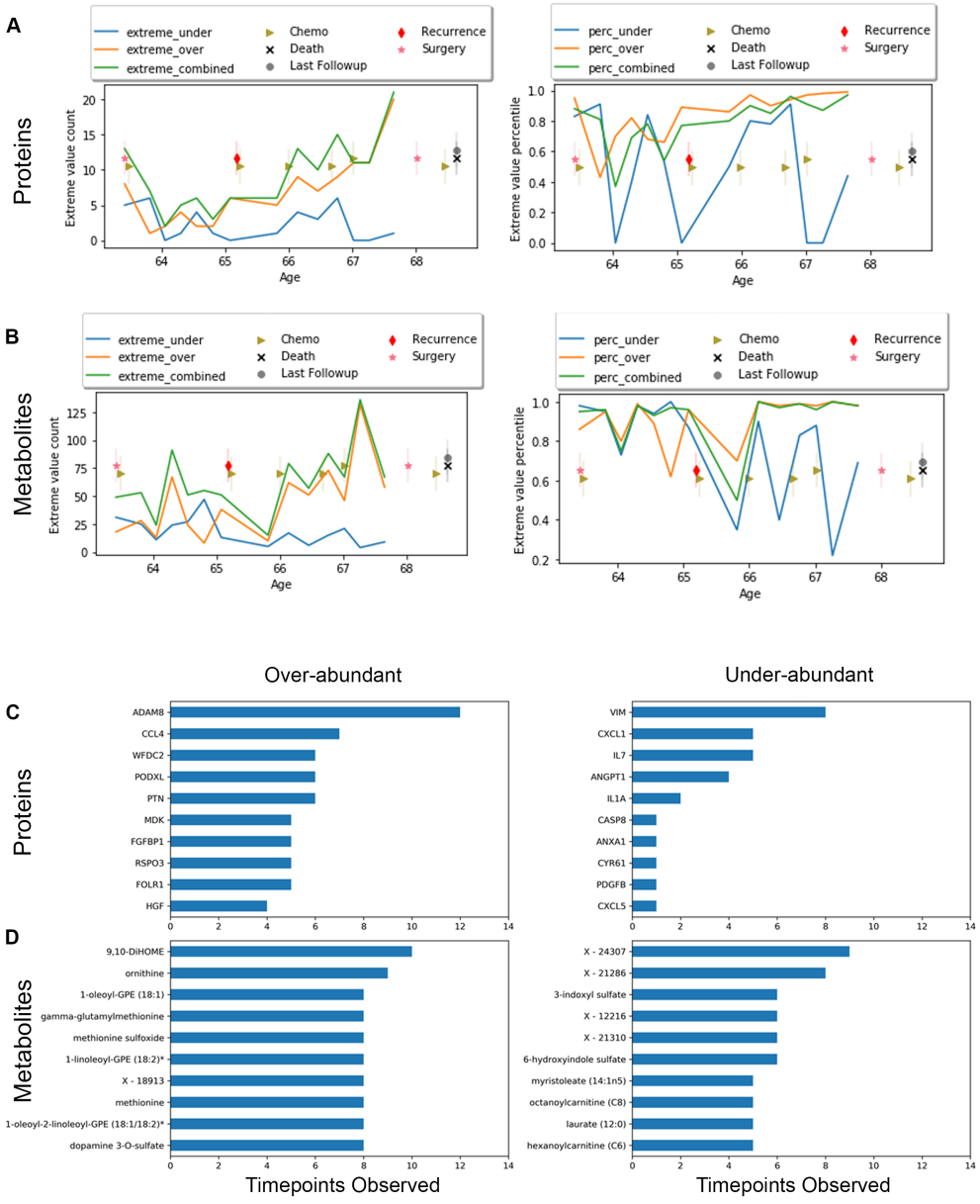
Initially, CP10 presented with many extreme values (above 80th percentile) in metabolomics and proteomics, both in under-abundance and over-abundance. Six months following initial surgery, extreme protein levels dropped to a relatively normal percentile, with a similar, although less precipitous, drop in metabolite levels. After this drop, a slow and steady increase in extremely high proteins was observed, with (other than one time point immediately following recurrence) over-abundant metabolites quickly returning to the 99th percentile, while under-abundant metabolites and proteins fluctuating between the 20th and 80th percentiles

(Figure 4.8). The most commonly over-abundant protein was ADAM8 (A Disintegrin and metalloproteinase domain-containing protein 8), which was more abundant than in 99% of the matched SWC cohort in 12 out of 14 observations. The most commonly over-abundant metabolite was 9,10-DiHome (9,10-dihydroxyoctadecenoic acid), with several methionine-related compounds also being regularly over-abundant.

Systems analysis of the metabolites provides insight into the disease's progression in this patient (Figure 4.9). Throughout treatment, a strong over-abundance of Amines/Primary Amines/Monalkylamines is observed, as well as amino acids of different types. Post-recurrence, there is an increase in extremely over-abundant Monoacylglycerols. Sporadic increases in Diacylglycerols was also present. A marked under-abundance of Long, Straight, and Unsaturated fatty acids preceding tumor recurrence was also seen.

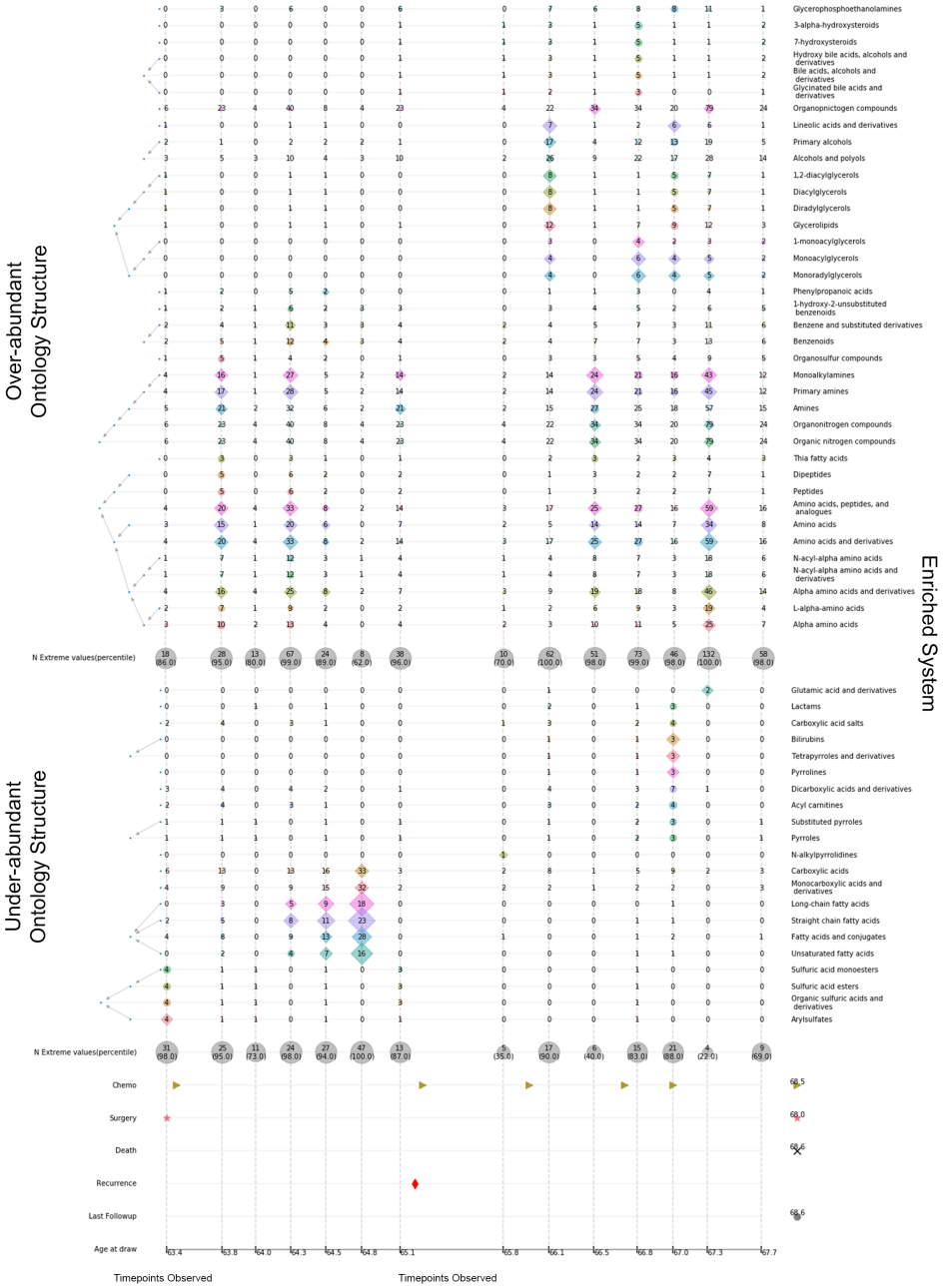
Systems analysis of proteins yielded nothing of significance. While some top Kegg disease networks were of various cancer types, none were significantly enriched, and the top networks also contained many equally weighted networks that were unrelated to ovarian cancer. Among Uniprot tissue annotations, the only weakly significant tissues were immune-related (Supplementary Figure 4.S8). While cancer was in the top 10 tissues, it was not significant. Possible alternative strategies for the future are discussed in the limitations section.

Figure 4.8: CARE Patient 10: Longitudinal extreme value profile



Extreme value counts and percentiles of those counts relative to the SWC, and the most common extreme values among proteins and metabolites. The extreme value cut-offs presented are less than or equal to the 1st
percentile for under-abundance and greater than or equal to the 99th percentile for over-abundance.
A and B present the trajectories of extreme values in proteins and metabolites, respectively. C and D show the number of time points a protein or metabolite value was over- or under-abundant in Patient 10 of the CARE cohort.

Figure 4.9: CARE Patient 10: Longitudinal systems



The significantly enriched sets of metabolites for Patient 10 using the Chemical Ontology. The number of extreme values belonging to each set are presented. They are marked with a circle if the probability of over- or under-abundance was Benjamini-Hochberg adjusted p-value $< .1$, and are marked with a diamond if they had a Bonferroni adjusted p-value $< .1$. Ontology structure is presented on the left of each graph. Important events throughout treatment are presented on the bottom.
Only over-representation of metabolites was tested.

SWC Transitions: Chronic Lymphocytic Leukemia

One client from the SWC was diagnosed with Chronic Lymphocytic Leukemia (CLL) after three observations. He was the first male in the SWC to be diagnosed with cancer following joining the program, and was therefore given the pseudonym "Adam." Adam was a white male with a de-identified age of 68.1 at his first blood draw and subsequent blood draws at the age of 68.4, 68.8 and 69.3. He received his diagnosis at the age of 69.0, which he received after being referred to his physician by Arivale based on abnormal clinical measures. At last contact, Adam indicated to his health coach that he did not intend to undergo treatment for his condition.

For the first three time points (pre-diagnosis), 13 Olink protein panels (Organ Damage, Cardiometabolic, Cardiovascular II and III, Metabolism, Immune Response, Neuro-Exploratory, Inflammation, Oncology II and III, Cell Regulation, and Neurological I) were run for a total of 1193 Proteins. For the final time point (post-diagnosis), 3 Olink panels (Cardiovascular II and III, and Inflammation) were run, for a total of 274 proteins.

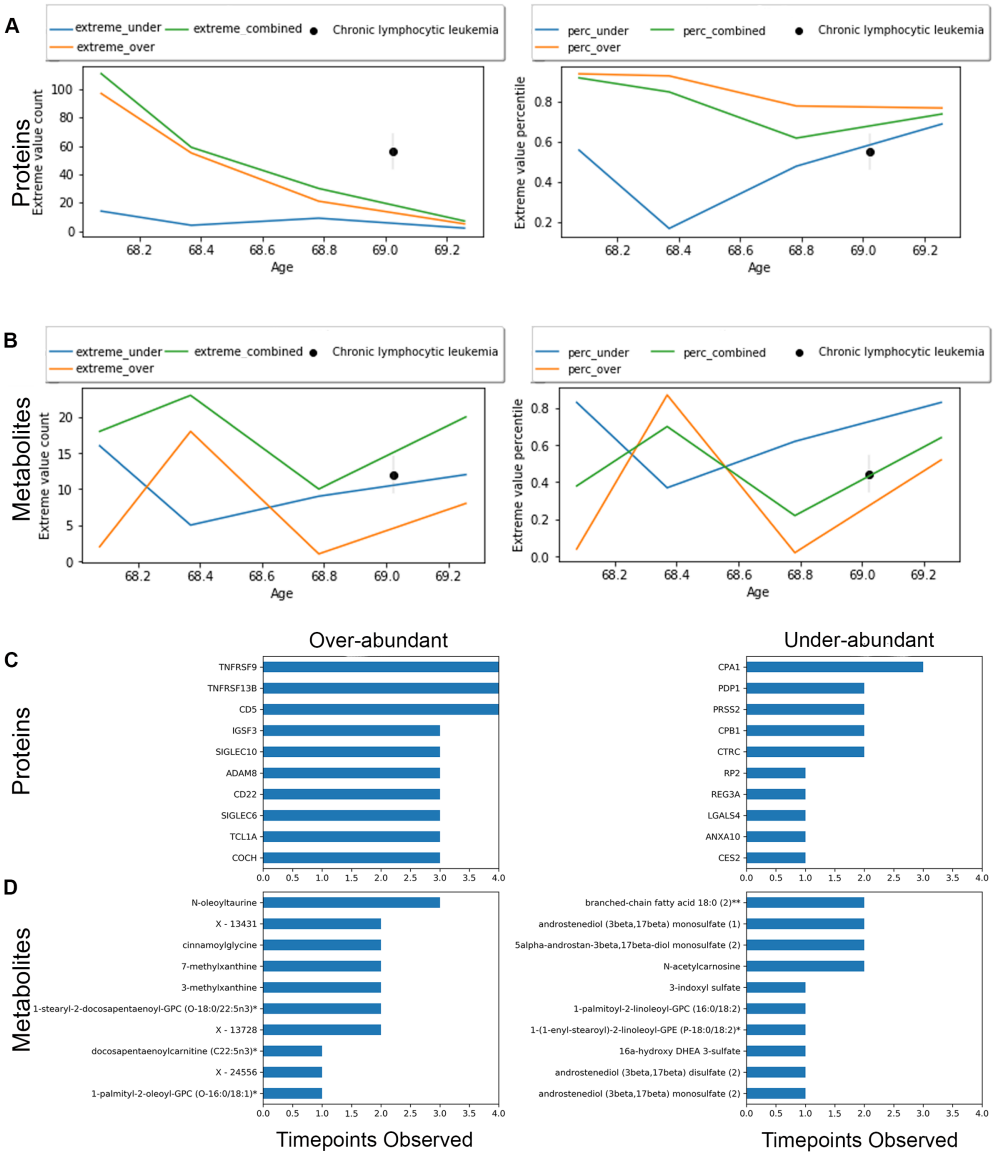
Adam had extreme over-abundance of the proteins TNFSRF9, TNFRSF138, and CD5 at all four time points with TCL1A, SIGLEC6, CD22, IGSF3, ADAM8, SIGLEC10, COCH, CD200, CGA, FCER2, LAG3, GALNT7, and FCRL1 extreme at all time points where they were measured (Figure 4.10C). The most commonly over-abundant metabolite was N-oleoyltaurine, with several androgens, such as androstenediol and DHEA, being under-abundant (Figure 4.10D). In aggregate, at the first two time points, the number of extreme values in over-abundant extreme proteins was above the 90th percentile, with the following two time points having percentiles in the high 70s and low 80s, relative to randomly sampled members of the SWC (Figure 4.10A). Under-abundant proteins were initially in the 50th percentile, dropped to the 20th percentile, and subsequently returned to the 50th percentile. Few over-abundant metabolites were present at the first observation, with a sharp jump at the second time point, a subsequent drop, and a final jump. Under-abundant metabolites were generally high, although not enormously so, with a drop at the second time

point(Figure 4.10B) .

The metabolites' systems enrichment showed an increase of pyrimidines and purines in the second observation with an over-enrichment of under-abundant androgens at the third and fourth time points. The second observation of over-abundant proteins showed an over-representation of nucleotide-related processes from the Gene Ontology. Additionally, a moderate amount of enrichment was present in cell adhesion processes at the first three time points, although only significant at the first observation. Uniprot tissue annotation showed strong over-enrichment in immune-related proteins, especially from the B-cells.

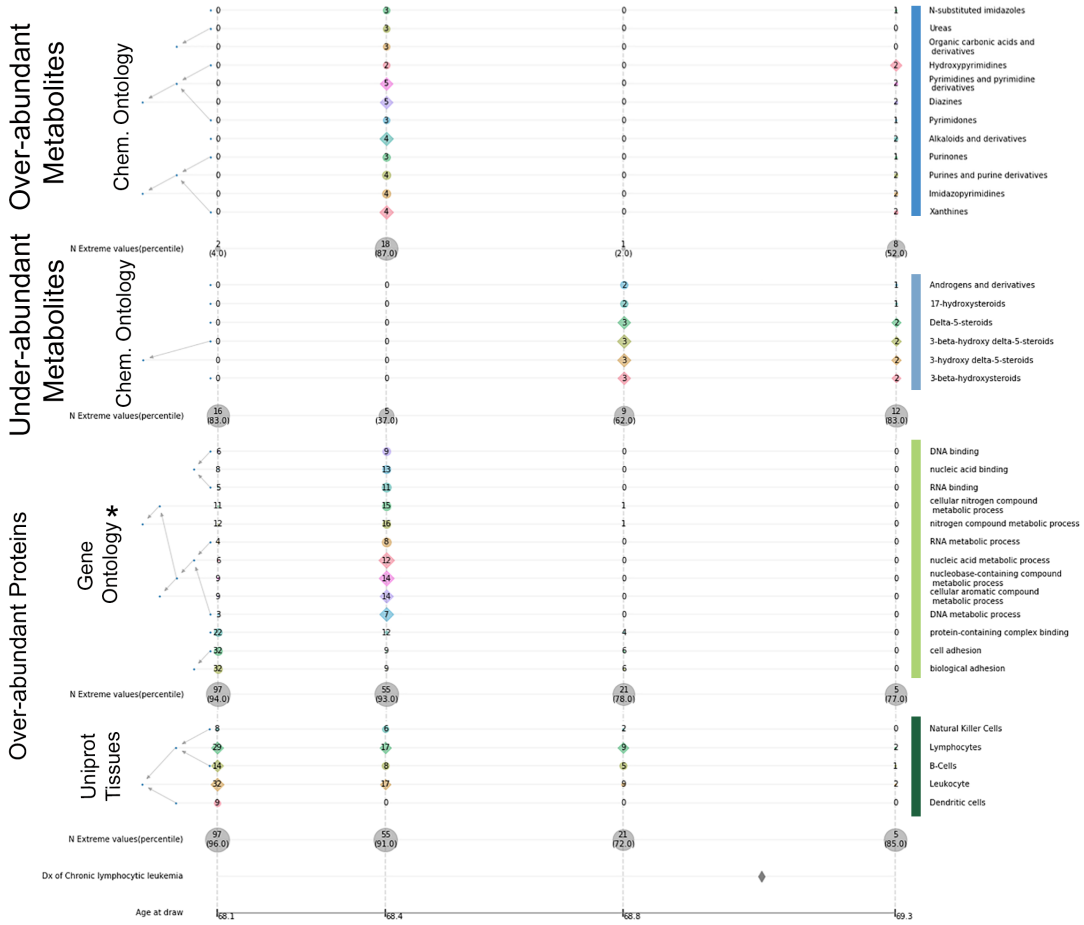
Subsequent investigation of this client's genome revealed that he was homozygous for the B-RAF V600E effect SNP, rsid:rs113488022(A;A).

Figure 4.10: SWC Transitions (CLL): Longitudinal extreme value profile



Extreme value counts and percentiles of those counts relative to the SWC, and the most common extreme values among proteins and metabolites. The extreme value cut-offs presented are less than or equal to the 1st percentile for under-abundance and greater than or equal to the 99th percentile for over-abundance.
A and B present the trajectories of extreme values in proteins and metabolites, respectively. C and D show the number of time points a protein or metabolite value was over- or under-abundant in the CLL transition from the SWC.
Note: The first three time points use 1193 proteins, while the final time point has 274.

Figure 4.11: SWC Transition (CLL): Longitudinal systems



The significantly enriched sets from extreme values in proteins and metabolites at a cut-off of less than the 1st percentile or greater than the 99th percentile. The number of extreme values belonging to each set are presented. They are marked with a circle if the probability of over- or under-abundance was Benjamini-Hochberg adjusted p-value $< .1$, and are marked with a diamond if they had a Bonferroni adjusted p-value $< .1$. Ontology structure is presented on the left of each graph. Age at diagnosis is presented at the bottom.

* - Gene ontology sets annotated as "cellular components" were excluded for brevity from the graph, but were included in multiple hypothesis adjustment.

Note: The first three time points use 1193 proteins, while the final time point has 274.

4.5 Discussion

4.5.1 Cohort-level observations

Measures that are commonly extreme, that is are extreme in at least one observation in many individuals, and measures that are frequently extreme, that is are extreme in multiple observations of the same individual, are described below. Several of the most frequent and common metabolites observed in the CARE cohort (Figure 4.4) are well-known products of cancer metabolism[159]. Fumarate is an onco-metabolite which accumulates to very high levels in tumors[160]. Lactate is a product of the Warburg effect [161]. Sphingosine-1-phosphate is a known marker of ovarian cancer invasiveness and breast cancer progression, with the closely related sphinganine-1-phosphate being a product of the same kinase (Sphingosine kinase 2) [162, 163]. While less studied, sphinganine-1-phosphate has been shown to have opposite modulatory effects to sphingosine-1-phosphate through the PTEN/PPM1A-dependent pathway [164]. 9,10-DiHOME (9,10-dihydroxy-12Z-octadecenoic acid) is a cytotoxic Linoleic Acid derivative that has been observed as being over-abundant in several conditions where oxidative stress and inflammation are present, including cancer [165, 166].

Of particular interest is the commonly and frequently over-abundant metabolite cysteine-s-sulfate (CSS). CSS is a very potent NMDA receptor agonist, second only to glutamate [167]. Accumulation of this metabolite contributes to severe neurological impairment in the genetic disease, Molybdenum cofactor deficiency[168]. A metabolomic study of the Breast Cancer Family Registry, where participants with a family history of cancer were tracked prospectively, showed a 125% increase in CSS in women diagnosed with breast cancer relative to controls [169]. This dissertation shows that this trend maintained in women undergoing treatment for breast cancer, and it exists for women undergoing treatment for ovarian cancer [169]. The additional relationship of CSS with neurological disorders may point to a causal factor in the well-known "Chemobrain" effect observed where patients that survive cancer treatment experience diminished mental capacity [170]. This may suggest possible treatments based on current therapies for sulfite oxidase deficiencies [171]. This observa-

tion is being followed up in an ongoing study of wellness in Breast Cancer Survivors via a partnership between the Institute for Systems Biology and Swedish Medical Center.

The uniquely frequent metabolites in the ovarian cancer patients (1-palmitoleoylglycerol (16:1), 1-linoleoylglycerol (18:2), 1-oleoylglycerol (18:1), 1-arachidonylglycerol (20:4)) are all 1-monoacylglycerols. Monoacylglycerols are formed biochemically via the release of a fatty acid from diacylglycerol by Diacylglycerol lipase or Hormone-sensitive lipase, which is generated by the break down of triacylglycerides (triglycerides), the primary long-term energy storage molecule of the body, generally stored in adipocytes. The protein MAGL (Monoacylglycerol lipase) controls free fatty acid levels in cancer cells by pairing lipogenesis and high lipolytic activity to promote protumorigenic signals, especially in ovarian cancer [172]. In intra-abdominal tumors such as ovarian cancer, where metastatic cells migrate and induce lipolysis, adipocytes provide fatty acids for rapid tumor growth [173]. The high levels of monoacylglycerols demonstrated here may be a consequence of the increased release of free fatty acids from diacylglycerol and the death of the exhausted adipocytes, dumping their waste into the bloodstream.

Many of the remaining metabolites are possibly byproducts of the many drugs and therapies these patients were undergoing. Maleic acid and benzoate are components of common pharmaceuticals, and often metabolites that are unable to be annotated are xenobiotic and byproducts of the metabolizing of various less common drugs.

Notably, breast cancer patients appear to have far weaker and less consistent signal in extreme proteins than ovarian cancer patients. This may be due to the breast cancer patients being predominantly at an earlier stage than the ovarian cancer patients (Table 4.2), or it may be a result of bias in the protein panels used in this dissertation. Interestingly, this is consistent with the observations from the biological age calculations where the breast cancer patients had similarly high metabolomic BAs to the ovarian cancer patients, while the proteomic BAs of the ovarian cancer patients was far higher than that of the breast cancer patients(Figure 4.3). This may indicate that proteomic analysis of the blood from cancer patients is most useful when examining higher grades of cancer, where metastasis and

cell-death provide sufficient proteins to the blood for identification. A larger, less restricted set of protein markers would be required to test this.

The ovarian extreme values show ample evidence of cancer progression, which is to be expected as the protein panels were chosen to be oncogenic. Here is discussed, the top 4 most commonly over-abundant proteins. MUC16 (Mucin-16), also commonly known as CA-125, is a large integral membrane glycoprotein that can be detected in over 80% of ovarian carcinomas, and recently was proposed as a target for T-cell therapies, although this is complicated by glycosylation [174, 175]. While it was extreme in 10 of 11 ovarian cancer patients, it had the lowest median frequency of the four most commonly extreme proteins, possibly indicating an optimal time window of secretion. It appears to be elevated early in treatment with a resurgence immediately before recurrence (Supplementary Figure 4.S3). FGFBP1 (Fibroblast growth factor-binding protein 1) acts as a carrier protein that releases fibroblast-binding factors (FGFs) from storage in the extracellular matrix and thus enhances the mitogenic activity of FGFs [176]. MDK (Midkine) is expressed in various tumor cell lines and regulates processes such as inflammatory response, cell proliferation, cell adhesion, cell growth, cell survival, tissue regeneration, cell differentiation, and cell migration [177–179]. This protein also promotes epithelial to mesenchymal transition through interaction with NOTCH2 making it a major contributor to late-stage tumorigenesis [180]. PTN (Pleiotrophin) is commonly expressed in several tumor types and shares receptors with MDK [181].

4.5.2 N-of-1 Analysis

CARE Patient 7

CP7 shows characteristic metabolomic over-abundance of ornithine, cysteine-s-sulfate, and methionine metabolism byproducts, which is discussed elsewhere. Of particular interest to CP7 is her consistent overabundance of 2-aminoadipate, a well-known risk factor for type 2 diabetes [182]. This may be reflected in her high metabolomic BA_E, relative to her proteomic BA_E. Breast cancer is more common in women with Type-2-diabetes (T2D), although this

risk is mediated by obesity with a 27% risk increase with T2D in general, and 16% after adjusting for BMI [183, 184]. It has been shown that T2D is associated with a 40% increase in mortality within the first five years following breast cancer, although that may be due to diabetes rather than cancer [185]. It is unknown if CP7 had pre-existing T2D, but it has been shown that breast cancer survivors have a 1.22 OR for subsequently developing T2D [186]. The impact of surviving breast cancer on general health and wellness is an important area of future research, with poor general health outcomes expected among cancer survivors. As treatments improve and more survivors pass the critical stage of surviving their cancers, focus on post-treatment wellness will become crucial to returning these individuals to their prior states. In other words, healthcare that focuses on not only curing cancer, but healing the patient. This is the essence of Systems Medicine.

From the systems analysis of metabolites, once again, alterations in fatty and amino acid metabolism leading up to cancer recurrence were detected, presenting more evidence of changes in catabolic and anabolic processes consistent with rapid cancer growth. Additionally, the systems analysis of proteins provides a window into the eventual recurrence of CP7s breast cancer. ERK and PI3K pathways, from Kegg network annotations, are classic cancer pathways, and perturbations are commonly associated with breast cancer [187]. Intriguingly, some evidence has shown that Simvastatin may promote cancer cell death in breast cancer, specifically through the deactivation of these pathways [188]. Similarly, the upregulation of serine/threonine protein kinases from the Gene Ontology is another classic marker of cancer, with several being explored for their prognostic value in breast cancer [189, 190]. Attempts have been made to inhibit these kinases as part of cancer treatment, but definitive trials are ongoing, and results are mixed [191].

CARE Patient 10

While most of the recurrent ovarian cancer patients showed a distinct pattern of resurgence in MUC16 abundance before recurrence, CP10 is a notable exception (Supplementary Figure 4.S3). CP10's extremely high proteomic BA_E, especially post-recurrence, seems in-

formative given her eventual outcome. That said, there were obvious patterns present in her commonly extreme values. The most commonly observed extreme metabolite was the leukotoxin 9,10-DiHOME, which was also commonly over-abundant in ovarian cancer in general, as discussed above. A persistent signal in the most commonly over-abundant metabolites is from methionine and methionine-cysteine related metabolites. This is a common phenomenon that has frequently been observed with these metabolites directly promoting cancer cells' proliferation and protecting the cancer cells from chemotherapy, specifically in ovarian cancer [192, 193]. Intriguingly, methionine-deprivation can be an effective adjuvant therapy in many cancers [194–196]. ADAM8 has been identified as a significant player in aggressive malignancies, including breast, pancreatic, and brain cancer, with high expression levels associated with invasiveness and poor outcomes [197]. In pancreatic cancer cells, ADAM8 induces migration and invasion and contributes to chemoresistance [198]. The ubiquity of the over-abundance of this protein in CP10 leads to the question of whether this patient may have benefitted from therapies that directly target this protein, which, as a well-known metastatic marker in other cancers, is a well-studied drug target [193, 199, 200]. While ADAM8-inhibitors and methionine-deprivation may not be a standard treatment for ovarian cancer, one must wonder if this is a path that could have made a difference in the outcome for CP10, were she viewed on an N-of-1 basis with deep phenotyping in real time with information going to her physician rather than in a retrospective analysis.

Systems analysis of CP10 provides further insight into the progression of her cancer during treatment. Monoacylglycerols and their relationship to metastasis in ovarian cancer are discussed above in Section 4.5.1. Excess amino acids provide several advantages to cancer metabolism. It has been shown in pancreatic adenocarcinoma, that elevated amino acids are present and the result of increased whole-body protein breakdown in the early stages[201, 202]. The Alkylamines are downstream metabolites of ornithine and ODC (ornithine decarboxylase), which is controlled by the potent oncogenic transcription factor cMyc[203]. They are derived from amino acids and a common feature of many cancers[204].

SWC Transitions: Chronic Lymphocytic Leukemia

CLL is characterized by the proliferation of CD5+ B-cells, so the over-abundance of CD5 is expected [205]. CLL is a B-cell leukemia that progresses slowly initially and where treatment focuses on controlling the symptoms rather than attempting to cure, as chemotherapy and radiation treatments do not appear to improve overall survival significantly [206]. The Tumor Necrosis Factor receptor superfamily proteins, TNFRSF9 and TNFRSF13B, are secreted by T- and B-cells, respectively. Increased expression of TCL1A (T-cell leukemia/lymphoma 1A) has been associated with shorter time to treatment in Chronic Lymphocytic Leukemia, where "watchful waiting" is the standard first line treatment [207]. The primary transgenic mouse model for CLL is a model where TCL1A expression is overstimulated, which leads to the over-abundance of CD5+ B-cells [208]. In the same mouse model, over-expression of TNFRSF13B, commonly also known as TACI, accelerated the progression of CLL [209]. In short, there is ample evidence of CLL in the extremely over-abundant proteins, with many other commonly over-abundant proteins playing important roles in carcinogenesis, especially the already discussed ADAM8, which was commonly over-abundant in the CARE ovarian cancer patients. Tissue enrichment demonstrates that there was also an enrichment specifically in B-cell related proteins.

The metabolite, n-oleoyltaurine, is an endogenously produced N-acyl-taurine conjugate of oleic acid and the amino acid taurines. It has been shown to have anti-proliferative effects in cancer as an apoptosis-inducer and an anti-neoplastic agent and may be an early warning sign of the body fighting cancer [210].

The enrichment of over-abundant purines and pyrimidines among the metabolites, with concordant enrichment of RNA- and DNA-related metabolic processes, was not found to be exceptionally informative in the case of common CLL. Neither was the dramatic drop in androgens/steroids. Interestingly, these are concordant with a specific subtype of CLL, Hairy Cell Leukemia (HCL) where androgen therapy was an early treatment [211–213]. HCL is rare (less than 2% of all leukemias) and generally presents as typical B-cell lymphoid

leukemia [214]. HCL is commonly misdiagnosed as CLL, but is very easily treated when properly diagnosed via a bone marrow biopsy. Treatment with purine analogs, which have very mild side effects compared to traditional chemotherapy regimens, result in median follow-ups of 12.5 years [215, 216]. The BRAF v600E variant, while present in other cancers, was shown to be highly specific to HCL, such that 47 patients with HCL carried the variant and 195 patients with other peripheral B-cell lymphomas or leukemias did not [217]. It seems likely that Adam does not have classical CLL, but instead has HCL. Hopefully, Adam decided to undergo treatment, and the appropriate bone marrow biopsy was performed. As of this time, this student is not legally allowed to contact or re-identify this client.

4.5.3 Limitations and Future work

Extreme Value Analysis

In this dissertation, I used a simple non-parametric method to identify a given value as extreme if above or below certain arbitrary percentile thresholds relative to a sample of similar individuals based on age, sex, and race. This is not the only, and not proposed as the optimal, method to identify a value as extreme. The ideal background distribution for identifying a deviation from an individual's prior good health would be previous measurements of that individual, where an individual is used as their own control. While the data here is longitudinal, there are not enough observations per individual (especially observations of good health before transition into disease) to develop an empirical distribution. It is hoped that as multi-omic data becomes ubiquitous, incorporation of prior measurements will be a pillar of 21st Century Systems Medicine. As this is not feasible currently, this section will discuss possible alternative implementations.

A background distribution could be generated empirically as done in this dissertation, or parametrically. When generating this distribution empirically, several options are still possible and remain to be investigated. The selection of the subset through a combination of categorical and real variables, as done here, is a simple method for selecting the distribution,

given the background sample is large, and the categorical restrictions are not too numerous. The size of the distribution must be sufficient to get a reasonable estimate of the likelihood of the observed value, and every added categorical variable exponentially diminishes the size of your sample. A weighted k-Nearest Neighbors approach is recommended for any selection composed of a large number of variables [218]. It is proposed that each demographic variable have its own weight based on their effect on the analyte of interest. In that way, a variable that changes a great deal with age, but not sex, would weigh age higher than sex, while a variable with the opposite relationship would weigh the age lower than sex. These weightings could then contain very rich demographic information, including known disease conditions and medications, without diminishing the sample population, creating a background distribution where the analyte will only be extreme if there is an unknown factor at play, e.g., a diabetic would not be repeatedly informed their blood glucose is high. Alternatively, one could use a training set to develop a posterior distribution for each variable in any manner of ways, such as Bayesian sampling, linear modeling, or assumptions about the shape of the underlying distribution [219–221].

Additionally, the choice of the metric for extremity relative to the expected distribution is another variable. In this dissertation, percentile cut-offs were used, but several other possibilities exist based on standard deviations, median absolute deviations, or one of many methods for estimation of a posterior probability for a given analyte value [222]. An approach similar to the above was successfully applied to SWC individuals that transitioned to cancer, using a learned median absolute deviation, that has generated putative early-warning biomarkers for metastatic cancer years prior to diagnosis [223].

Future work must include an exploration and comparison of these options. No matter the method used to identify these extreme values, the goal remains the same; given a large vector of variables, identify the variables that fall outside the expected range, conditioned on what is known about the individual.

Systems Enrichment of Extreme Analytes

One key question this dissertation attempts to address is, given that a set of variables that are identified as extreme or unusual by some criteria in an individual, how is that translated into actionable information. With a large number of possible variables, incidental or fluke extreme values are a critical concern. Even in studies of disease with enormous statistical power, false positives and unreproducible results are commonplace [224, 225]. The challenge of controlling these false positives is enormously amplified when observing many variables in a single individual, and essential if one is interested in performing medicine on an N-of-1 basis. The difficulty of managing this complexity is a common criticism of the practice of deeply phenotyping individuals. Several methods of controlling for this problem with PD3 clouds are suggested. One is the use of longitudinal measures. Consistency of extremity can be a clear indicator that signal and not noise is being observed. Another technique is the use of deep phenotyping that consists of many different -omics. Each -omic measure will possess batch effects, selection biases, and noise that make any given observation questionable, but by incorporating many different -omics and analytes, much (although certainly not all) of the noise will be orthogonal, which will allow the useful signal to reinforce itself [112]. In essence, one gets multiple witnesses to the same event. A key challenge to this approach is identifying when you see the same condition or issue multiple times. This issue is addressed here through the use of Knowledge Graphs and systems enrichment. The key insight is that while having a single metabolite in the TCA cycle extreme may not tell you much, having several metabolites and proteins from the TCA cycle does.

Unfortunately, as observed in the CARE proteomic systems enrichment, one must be careful with the metrics one applies to these systems and the set of measures one applies to them. One hypothesis for the poor enrichment identified in this dissertation, which is so at odds with the obviously cancer-relevant extreme proteins, is the bias of the background set of proteins used. The proteins used in the CARE cohort are well-known cancer markers. This makes set enrichment a poor choice for identifying perturbed systems. Set enrichment

works best on an unbiased background set, as seen in the metabolomic markers. This does not mean that one cannot focus one's systems analysis on measures that are known to be relevant to the phenotype of interest, such as cancer markers in this case. Systems Biology has developed several techniques for identifying perturbed networks that do not rely on the background set's diversity, such as DIRAC, WGCNA, and others [226–230]. These should be explored.

This dissertation's knowledge graph was limited and should be significantly expanded with other knowledge bases, as noted in Section 4.3.5. In addition to the noted publically available databases, additional multi-omic subgraphs should be examined based on data-driven approaches, such as correlation, disease-perturbed, and tissue-specific networks.

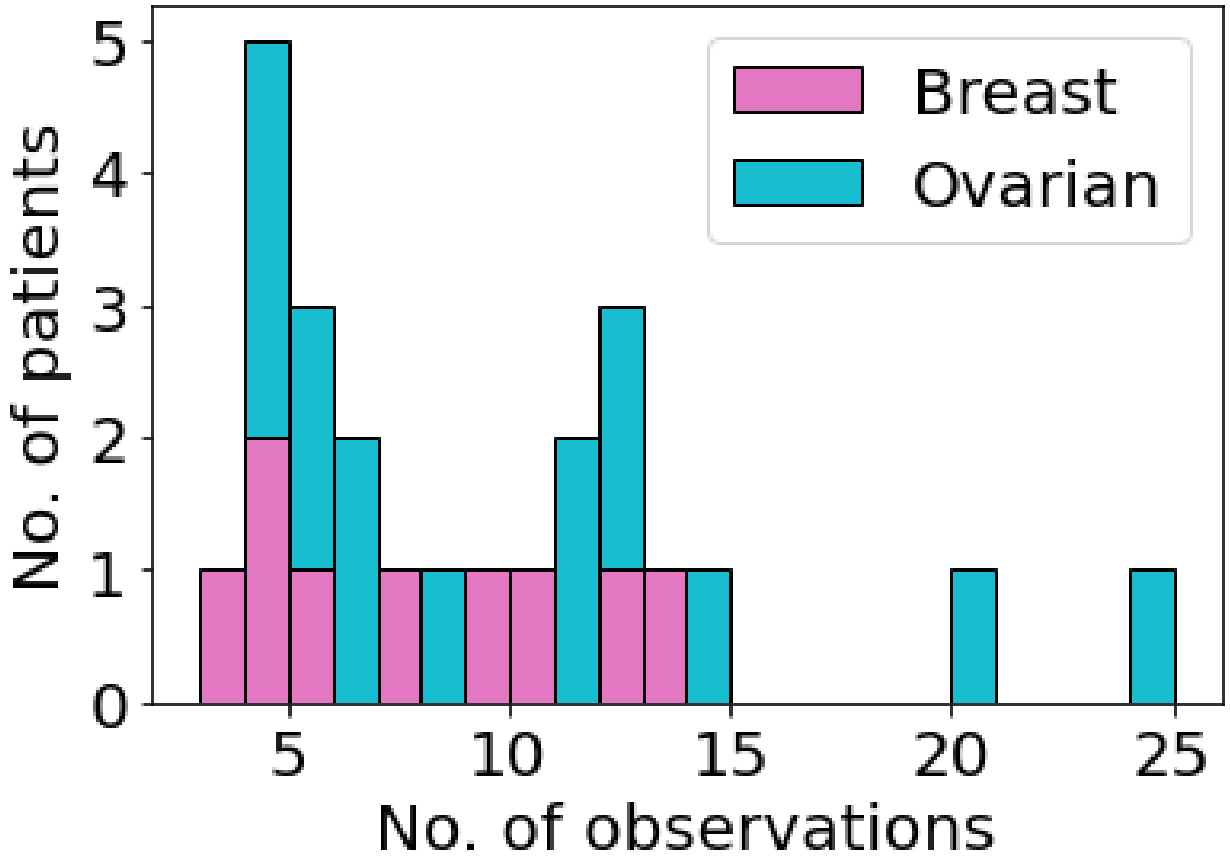
If I only had a brain

Finally, the insights generated by the above analyses tend to be too broad to be of serious translational use at present. It is necessary to develop machine learning methods for surfacing the most relevant pieces of information and tuning the parameters in so that the appropriate level of information is returned. Cancer is a very dramatic disease with large swings in protein and metabolite concentrations. Tuning the parameters to fit less dramatic, changes from diseases such as heart disease, diabetes, liver disease, etc. remains a tremendous challenge well-suited to machine learning approaches.

4.S Supplementary Materials

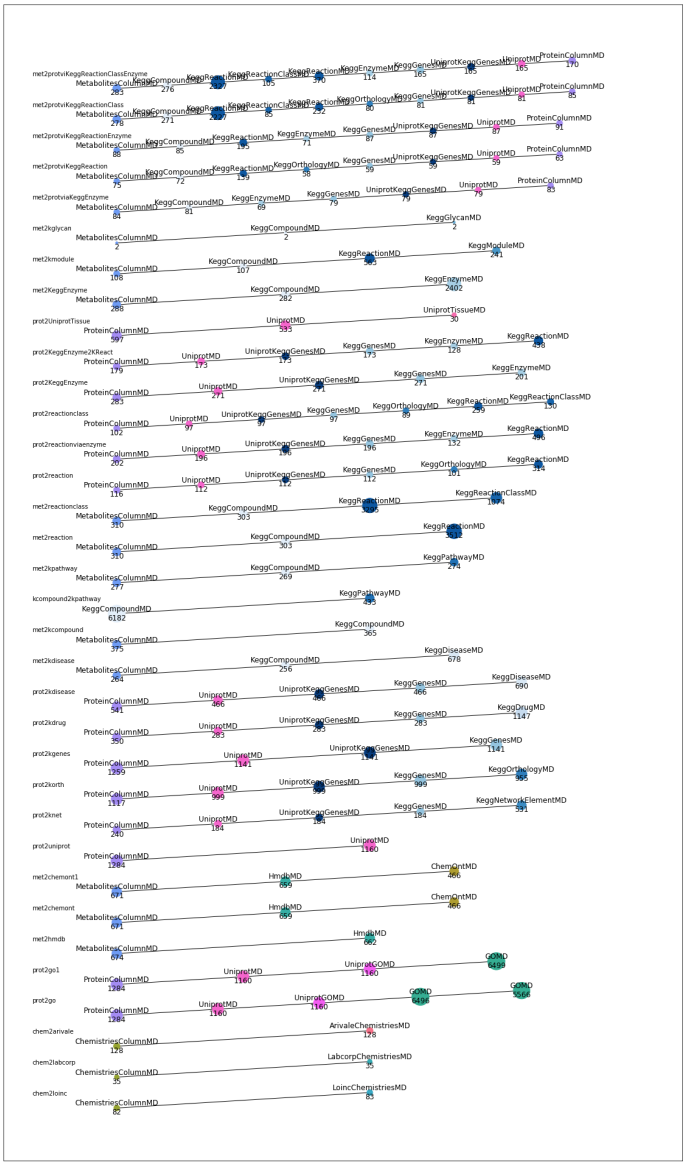
4.S.1 Supplementary Figures

Figure 4.S1: CARE cohort: Number of observations per patient



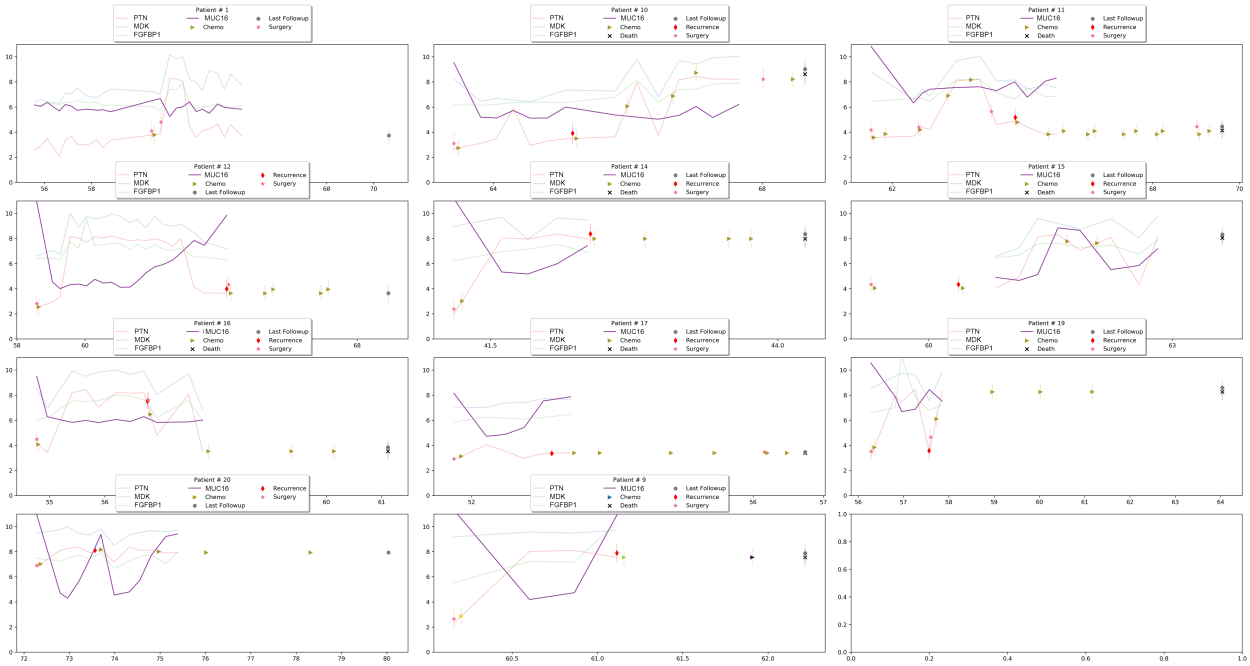
This stacked bar plot shows the number of time points per individual in the CARE cohort, stratified by cancer type.

Figure 4.S2: Knowledge Graph: Named paths



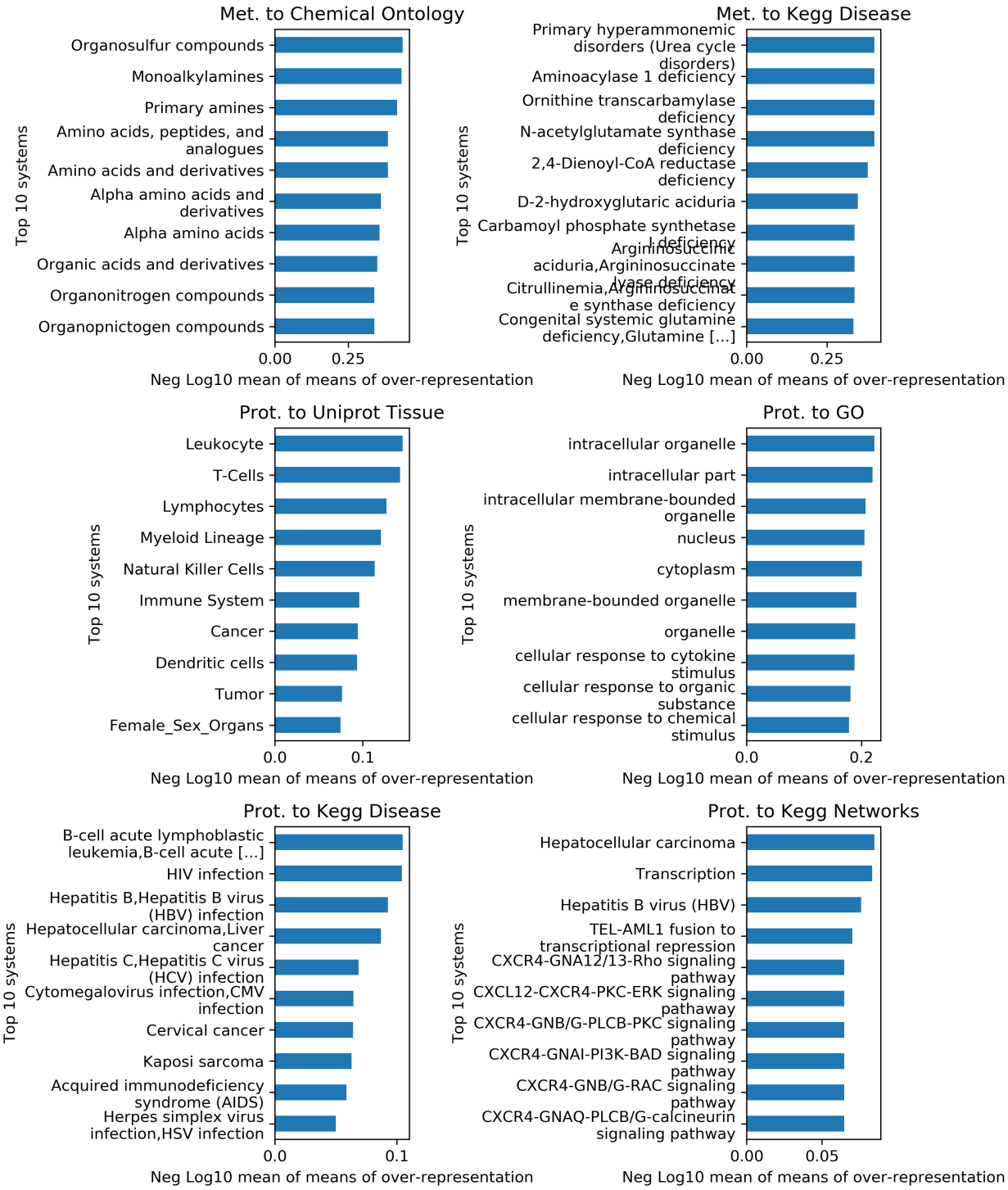
The number of connected entities by relationship in the Scarecrow Knowledge Graph. Note: this is against **all** proteins and metabolites in the Scientific Wellness cohot, not just the ones available in the CARE cohort.

Figure 4.S3: CARE Expression levels of top 4 common proteins



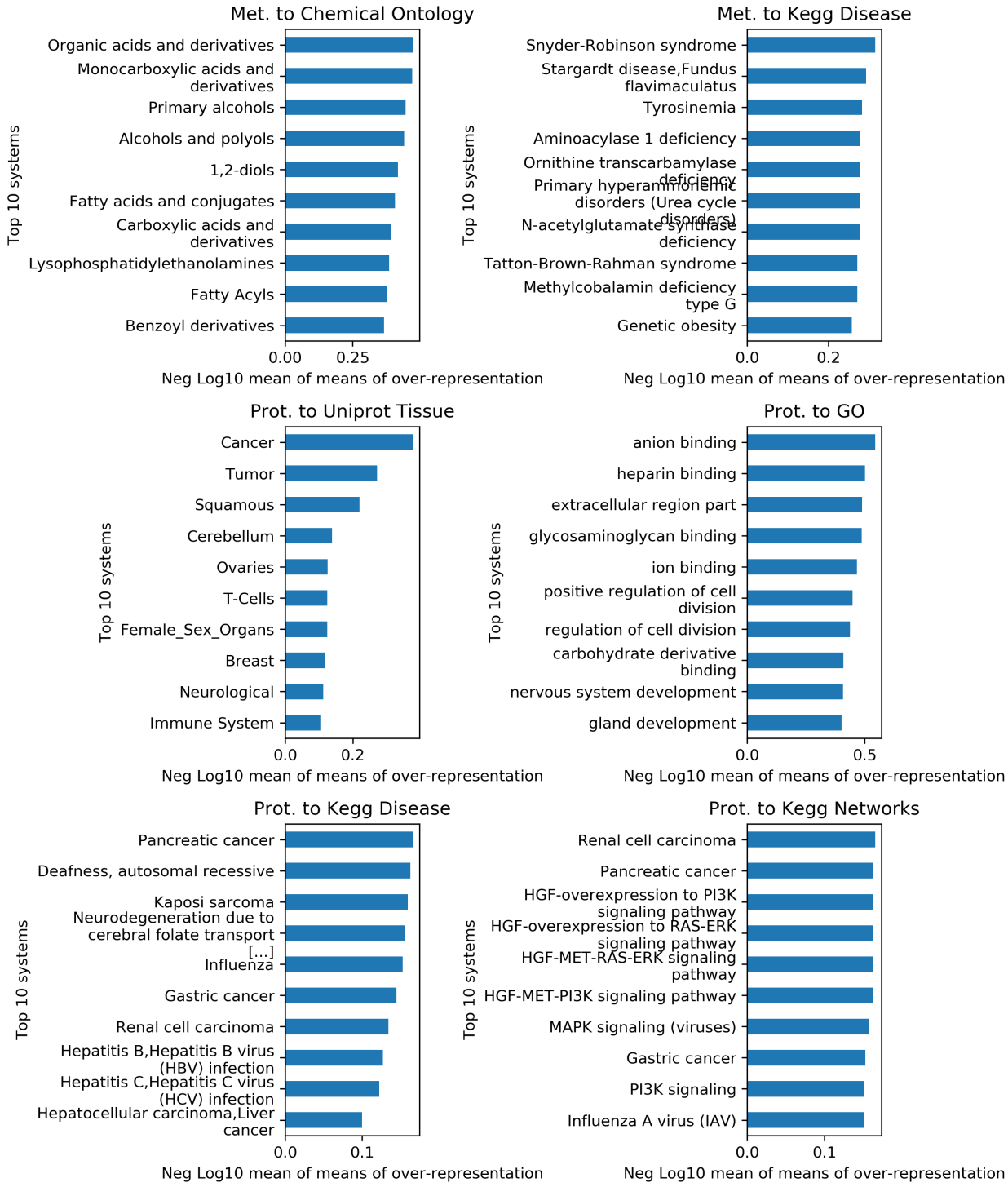
Expression profiles of the four most commonly over-abundant proteins in ovarian cancer patients. Mucin-16 (in purple) is highlighted as a possible marker of recurrence.

Figure 4.S4: CARE Breast Cancer: Top systems enrichments



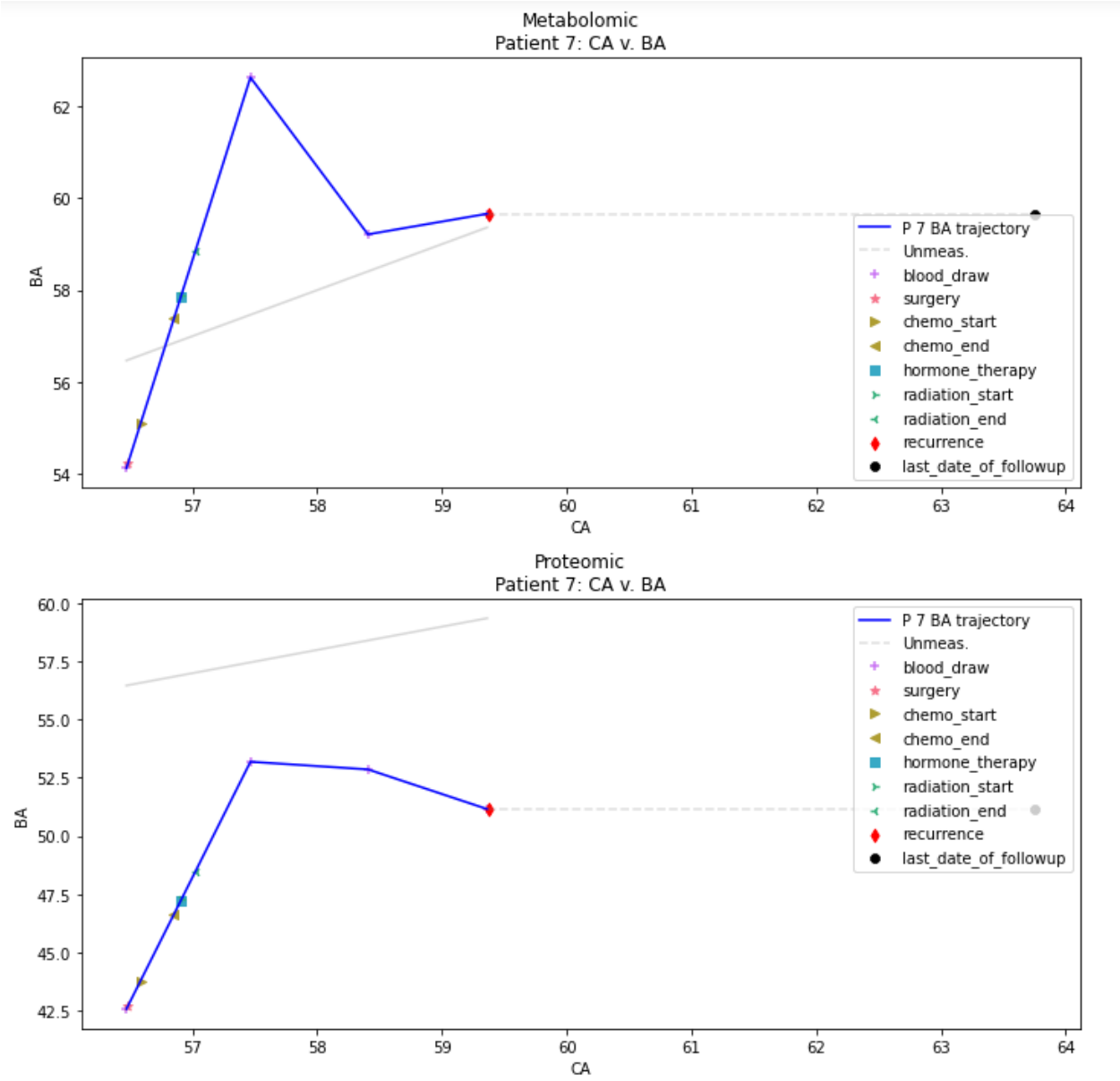
The top 10 systems enriched in breast cancer at the 1st and 99th percentile cut-offs. Each system was ranked by the mean of the mean probability of over-representation for each individual.

Figure 4.S5: CARE Ovarian Cancer: Top systems enrichments



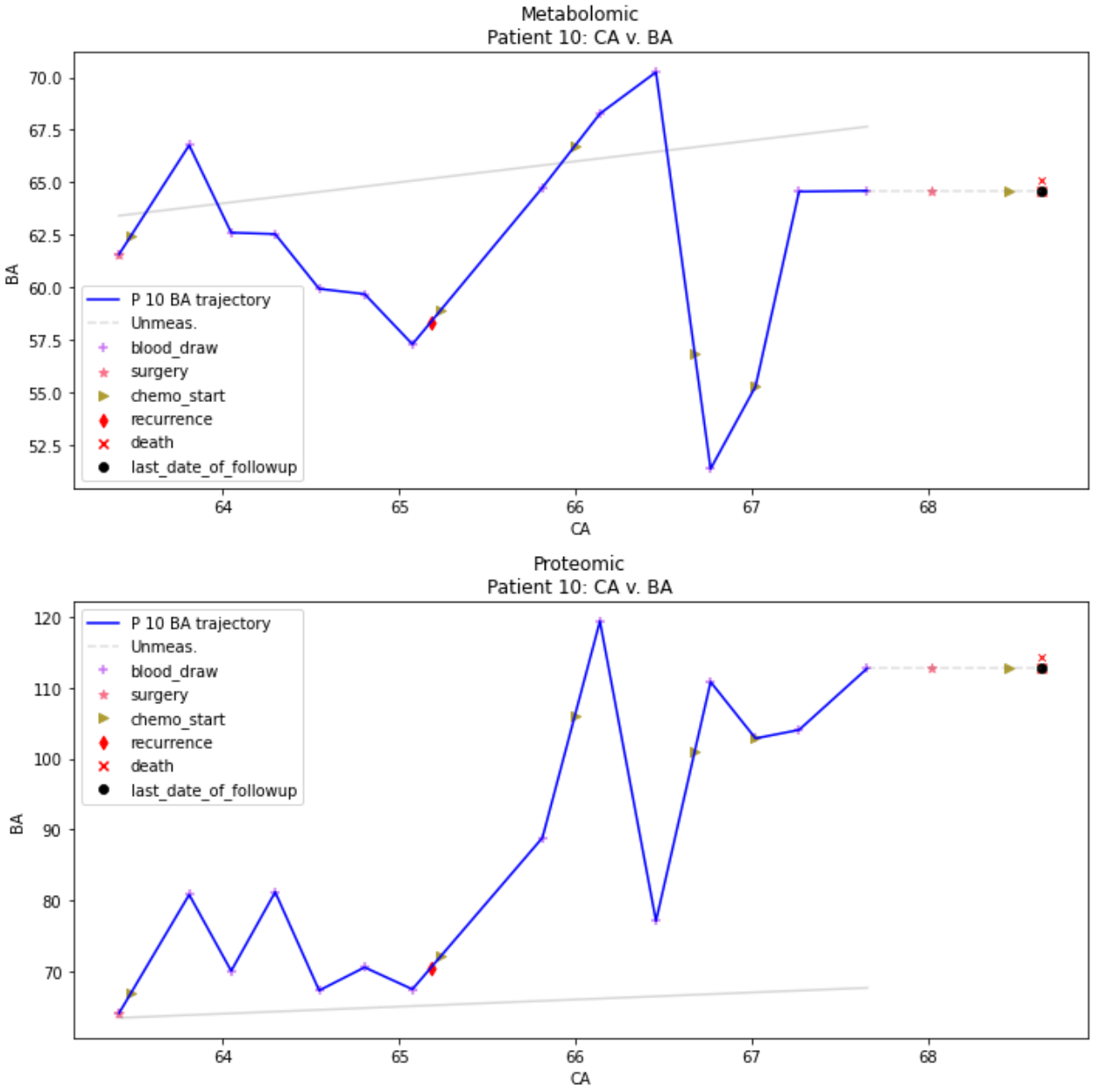
The top 10 systems enriched in ovarian cancer at the 1st and 99th percentile cut-offs. Each system was ranked by the mean of the mean probability of over-representation for each individual.

Figure 4.S6: CARE Patient 7: Biological Age trajectories



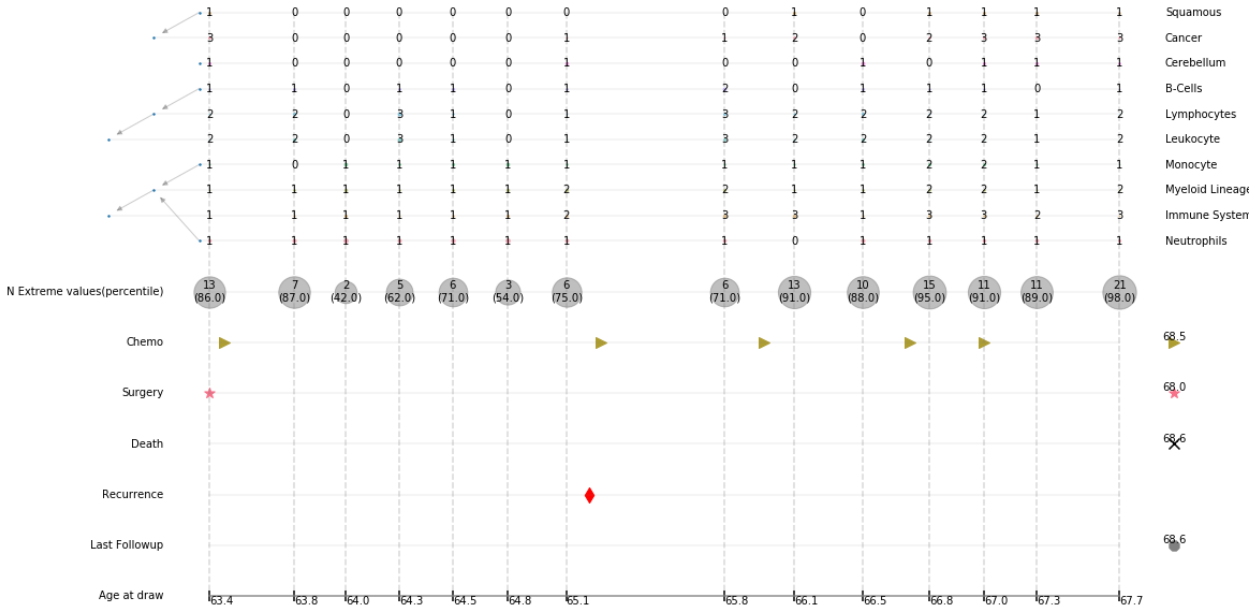
The metabolomic and proteomic biological age trajectories for CARE Patient 7, plotted with important events during treatment. The gray line shows the Chronological Age trajectory.

Figure 4.S7: CARE Patient 10: Biological Age trajectories



The metabolomic and proteomic biological age trajectories for CARE patient 10, plotted with important events during treatment. The gray line shows the Chronological Age trajectory.

Figure 4.S8: CARE Patient 10: Protein tissue enrichment



The top ten tissues from CP10's protein set enrichment. Nothing is significant and most are immune-related.

Downloadable supplementary file figures

The supplementary data file `supplement.zip` (6.1 G) contains many figures detailing the systems results for individuals that were not chosen for N-of-1 discussion. Additionally, there are figures for individuals that were discussed above, but whose systems were deemed less informative. File structure is detailed here.

The figures are contained in the `supplementary_figures` folder.
The `supplementary_figures/set_agg` folder contains the top systems by the group mean of individual mean probabilities of over-enrichment. for each system subgraph, grouped by cancer type and cutoff values. They are named `set_agg_'cancer type'_'lower percentile'_'upper percentile'.png`.

The supplementary systems figures for each patient are contained in the `supplementary_figures/systems_figures_by_patient/` folder. The subfolders are `N_of_1` which contains the individuals discussed above, `CARE_cohort` which contains members of the CARE cohort, and `SWC_cancer_transitions` which contains the set of individuals that transitioned to cancer while clients of Arivale. Each individuals root folder contains a `patient_info.txt` file that contains demographic and condition information for that individual. Under that folder is a set of folders named `'named path'_'cohort name'_'lower percentile'_'upper percentile'/`. This contains a text file with the parameters of this analysis `cfr_description.txt`, and a `figures` folder. The figures folder contains 3 folders, `enrichment_tree` which contains the graph of significant relationships for each time point individually, `frequency_individual` which contains the most common extreme values for each data type, and `longitudinal_enrichment_graph` which contains a set of figures for each `longitudinal_enrichment`. The significant values, as well as the top 10 and top 20 systems plots are present. Each significant system's metadata is presented in the `summaries` subfolder.

4.S.2 Supplementary Tables

Table 4.S1: Extreme value counts

Cohort	Data Type	Cutoff	Mean	Median	Std. Dev	Max.	Min
Care(Breast)	Metabolites	< 1 perc.	13.87	9.00	12.51	52	0
		< 2.5 perc.	26.37	20.00	17.94	83	6
		< 25 perc.	189.85	178.00	45.12	323	118
		> 75 perc.	371.99	365.00	72.66	539	227
		> 97.5 perc.	75.63	69.00	33.81	167	16
		> 99 perc.	41.42	36.00	20.96	101	4
	Proteins	< 1 perc.	3.24	1.00	5.69	39	0
		< 2.5 perc.	6.12	3.00	8.39	48	0
		< 25 perc.	43.34	44.00	29.91	118	3
		> 75 perc.	44.36	36.00	31.89	127	4
		> 97.5 perc.	5.13	4.00	5.64	27	0
		> 99 perc.	2.54	2.00	2.63	10	0
Care(Ovarian)	Metabolites	< 1 perc.	16.62	11.00	15.34	83	1
		< 2.5 perc.	32.44	26.00	23.89	139	5
		< 25 perc.	206.88	188.00	68.64	494	114
		> 75 perc.	384.54	392.00	102.31	589	93
		> 97.5 perc.	96.16	93.00	47.22	226	11
		> 99 perc.	55.58	51.00	31.59	158	7
	Proteins	< 1 perc.	2.53	2.00	3.13	18	0
		< 2.5 perc.	4.56	3.00	4.98	28	0
		< 25 perc.	35.67	34.00	20.10	90	1
		> 75 perc.	51.60	47.00	25.77	134	12
		> 97.5 perc.	9.67	7.00	9.60	65	0
		> 99 perc.	6.37	5.00	6.86	52	0
SWC(transitions)	Metabolites	< 1 perc.	7.17	6.00	4.60	16	0
		< 2.5 perc.	19.26	14.00	12.21	45	3
		< 25 perc.	235.35	221.00	85.63	396	102
		> 75 perc.	285.96	263.00	109.29	516	142
		> 97.5 perc.	29.48	24.00	18.92	69	2
		> 99 perc.	11.35	11.00	7.66	23	1
	Proteins	< 1 perc.	13.52	4.50	24.94	122	0
		< 2.5 perc.	25.83	12.00	39.12	209	0
		< 25 perc.	217.73	179.50	159.27	708	11
		> 75 perc.	268.25	239.00	196.71	668	16
		> 97.5 perc.	34.67	20.00	46.54	214	0
		> 99 perc.	18.65	7.50	27.47	128	0

Descriptive statistics of extreme value counts by cohort and cut-off.

Table 4.S2: Extreme value percentiles

Cohort	Data Type	Cutoff	Mean	Median	Std. Dev	Max.	Min
Care(Breast)	Metabolites	< 1 perc.	0.61	0.64	0.32	1.00	0.00
		< 2.5 perc.	0.58	0.59	0.30	1.00	0.04
		< 25 perc.	0.35	0.31	0.23	0.97	0.04
		> 75 perc.	0.90	0.93	0.11	1.00	0.53
		> 97.5 perc.	0.93	0.96	0.10	1.00	0.42
		> 99 perc.	0.93	0.95	0.12	1.00	0.16
	Proteins	< 1 perc.	0.46	0.52	0.38	1.00	0.00
		< 2.5 perc.	0.52	0.66	0.34	0.99	0.00
		< 25 perc.	0.50	0.61	0.34	0.97	0.00
		> 75 perc.	0.52	0.50	0.32	1.00	0.00
		> 97.5 perc.	0.53	0.62	0.32	0.98	0.00
		> 99 perc.	0.50	0.68	0.36	0.97	0.00
Care(Ovarian)	Metabolites	< 1 perc.	0.70	0.79	0.29	1.00	0.02
		< 2.5 perc.	0.68	0.77	0.26	1.00	0.00
		< 25 perc.	0.43	0.37	0.27	1.00	0.02
		> 75 perc.	0.88	0.97	0.20	1.00	0.00
		> 97.5 perc.	0.94	0.98	0.12	1.00	0.27
		> 99 perc.	0.95	0.98	0.09	1.00	0.54
	Proteins	< 1 perc.	0.52	0.61	0.33	0.96	0.00
		< 2.5 perc.	0.53	0.60	0.29	1.00	0.00
		< 25 perc.	0.45	0.49	0.27	0.95	0.00
		> 75 perc.	0.63	0.64	0.24	1.00	0.09
		> 97.5 perc.	0.75	0.82	0.21	1.00	0.00
		> 99 perc.	0.77	0.87	0.26	1.00	0.00
SWC(transitions)	Metabolites	< 1 perc.	0.43	0.44	0.29	0.89	0.00
		< 2.5 perc.	0.42	0.28	0.32	0.83	0.00
		< 25 perc.	0.53	0.57	0.34	0.97	0.01
		> 75 perc.	0.61	0.62	0.33	0.98	0.03
		> 97.5 perc.	0.55	0.55	0.31	0.92	0.00
		> 99 perc.	0.54	0.68	0.33	0.93	0.02
	Proteins	< 1 perc.	0.47	0.46	0.28	0.99	0.00
		< 2.5 perc.	0.50	0.51	0.27	0.99	0.00
		< 25 perc.	0.45	0.46	0.29	0.97	0.00
		> 75 perc.	0.54	0.58	0.31	1.00	0.03
		> 97.5 perc.	0.53	0.59	0.32	0.99	0.00
		> 99 perc.	0.52	0.61	0.34	1.00	0.00

Descriptive statistics of extreme value percentiles by cohort and cut-off.

Table 4.S3: Proteins used to calculation BA_E and CARE extreme value calculation.

Available in the downloadable supplementary data package, in the supplement/supplementary_data/ba_proteins.csv file (31K).

Table 4.S4: Metabolites used to calculation BA_E and CARE extreme calculation.

Available in the downloadable supplementary data package, in the supplement/supplementary_data/ba_metabolites.csv file (72K).

Table 4.S5: Results of comparison of metabolomics between CARE and SWC using GEE models

Available in the downloadable supplementary data package, in the supplement/supplementary_data/gee_results_metabolites.csv file (91K).

Table 4.S6: Full results of comparison of proteomics between CARE and SWC using GEE models

Available in the downloadable supplementary data package, in the supplement/supplementary_data/gee_results_proteins.csv file (91K).

Table 4.S7: Full table of systems results

Available in the downloadable supplementary data package, in the supplement/supplementary_data/all_systems_scores.csv file (634M).

Chapter 5

CONCLUSION

I walked into my new General Practitioner's office for a checkup about two years ago. He was a good doctor who seemed to enjoy getting to know his patients and asked me about myself and what I do. When I mentioned Arivale, he got quite excited and mentioned that one of his patients was a client. That patient had walked into his office with a printout of all of the data that was returned to them and asked him what he should be doing to get healthier. This physician, who was well-versed in functional medicine and in general an inquisitive sort, said he had no idea what to do with all that information. I told him that figuring that out was my job.

The data that the patient brought with them was a small fraction of everything we collected at Arivale, and yet it was still overwhelming. Our ability to affordably capture the genome, microbiome, metabolome, proteome, and all manner of health-related signals continues to expand at an tremendous rate. Transformation of this vast information into actionable signals, specific to the individual to whom they belong, in ways that improve their health and life is the basis of this dissertation.

The conversion of health data to health intelligence has been the goal of my graduate studies. While the company that grew from the work described in Chapter 2 is closed; the data generated, the systems produced, and the expertise developed continue to make significant contributions to our understanding of health and disease via published papers and currently ongoing studies of cancer, Alzheimer's disease, Scientific Wellness, Longevity, diseases of childhood and COVID [5, 223, 231–238]. Additionally, Arivale enriched and possibly extended the lives of thousands of customers, to whom this graduate student is enormously indebted and profoundly grateful.

This dissertation describes many paths taken and points to many more that need to be explored. I believe, someday soon, these paths will lead to a new era of medicine that is personalized, predictive, preventative, and participatory.

BIBLIOGRAPHY

[1] National Research Council et al. *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease*. National Academies Press, 2011.

[2] Eunice L Kwak, Yung-Jue Bang, D Ross Camidge, Alice T Shaw, Benjamin Solomon, Robert G Maki, Sai-Hong I Ou, Bruce J Dezube, Pasi A Jänne, Daniel B Costa, et al. Anaplastic lymphoma kinase inhibition in non–small-cell lung cancer. *New England Journal of Medicine*, 363(18):1693–1703, 2010.

[3] Manabu Soda, Young Lim Choi, Munehiro Enomoto, Shuji Takada, Yoshihiro Yamashita, Shunpei Ishikawa, Shin-ichiro Fujiwara, Hideki Watanabe, Kentaro Kurashina, Hisashi Hatanaka, et al. Identification of the transforming eml4–alk fusion gene in non-small-cell lung cancer. *Nature*, 448(7153):561–566, 2007.

[4] Nathan D. Price, Andrew T. Magis, John C. Earls, Gustavo Glusman, Roie Levy, Christopher Lausted, Daniel T. McDonald, Ulrike Kusebauch, Christopher L. Moss, Yong Zhou, Shizhen Qin, Robert L. Moritz, Kristin Brogaard, Gilbert S. Omenn, Jennifer C. Lovejoy, and Leroy Hood. A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nature Biotechnology*, 35:747, 2017. doi: 10.1038/nbt.3870https://www.nature.com/articles/nbt.3870#supplementary-information. URL http://dx.doi.org/10.1038/nbt.3870.

[5] John C Earls, Noa Rappaport, Laura Heath, Tomasz Wilmanski, Andrew T Magis, Nicholas J Schork, Gilbert S Omenn, Jennifer Lovejoy, Leroy Hood, and Nathan D Price. Multi-omic biological age estimation and its correlation with wellness and disease

phenotypes: a longitudinal study of 3,558 individuals. *The Journals of Gerontology: Series A*, 74(Supplement_1):S52–S60, 2019.

[6] Leroy Hood and Mauricio Flores. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *New biotechnology*, 29(6):613–624, September 2012.

[7] Leroy Hood and Stephen H Friend. Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nature reviews. Clinical oncology*, 8(3):184–187, March 2011.

[8] Francis S Collins and Harold Varmus. A new initiative on precision medicine. *The New England journal of medicine*, 372(9):793–795, February 2015.

[9] Pierre L Yong, Robert S Saunders, and LeighAnne Olsen. *The Healthcare Imperative: Lowering Costs and Improving Outcomes: Workshop Series Summary*. The National Academies Collection: Reports funded by National Institutes of Health. National Academies Press (US), Washington (DC), 2010.

[10] Lawrence A David, Arne C Materna, Jonathan Friedman, Maria I Campos-Baptista, Matthew C Blackburn, Allison Perrotta, Susan E Erdman, and Eric J Alm. Host lifestyle affects human microbiota on daily timescales. *Genome biology*, 15(7):R89, 2014.

[11] Rui Chen, George I Mias, Jennifer Li-Pook-Than, Lihua Jiang, Hugo Y K Lam, Rong Chen, Elana Miriami, Konrad J Karczewski, Manoj Hariharan, Frederick E Dewey, Yong Cheng, Michael J Clark, Hogune Im, Lukas Habegger, Suganthi Balasubramanian, Maeve O'Huallachain, Joel T Dudley, Sara Hillenmeyer, Rajini Haraksingh, Donald Sharon, Ghia Euskirchen, Phil Lacroute, Keith Bettinger, Alan P Boyle, Maya Kasowski, Fabian Grubert, Scott Seki, Marco Garcia, Michelle Whirl-Carrillo, Mercedes Gallardo, Maria A Blasco, Peter L Greenberg, Phyllis Snyder, Teri E Klein, Russ B Altman, Atul J Butte, Euan A Ashley, Mark Gerstein, Kari C Nadeau, Hua Tang,

and Michael Snyder. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 148(6):1293–1307, March 2012.

[12] Larry Smarr. Quantifying your body: a how-to guide from a systems biology perspective. *Biotechnology journal*, 7(8):980–991, August 2012.

[13] Leroy Hood and Nathan D Price. Promoting wellness and demystifying disease: The 100k project. *Clinical Omics*, 1(3):20–23, May 2014.

[14] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and Helen Parkinson. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*, 42(Database issue):D1001–6, January 2014.

[15] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(1):289–300, 1995.

[16] M Girvan and M E J Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99 (12):7821–7826, June 2002.

[17] M E J Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, June 2006.

[18] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, November 2003.

[19] Kwang Kon Koh, Sang Min Park, and Michael J Quon. Leptin and cardiovascular disease: response to therapeutic interventions. *Circulation*, 117(25):3238–3249, June 2008.

[20] Paul M Ridker. Clinical application of C-reactive protein for cardiovascular disease detection and prevention. *Circulation*, 107(3):363–369, January 2003.

[21] Y C Woo, Aimin Xu, Yu Wang, and Karen S L Lam. Fibroblast growth factor 21 as an emerging metabolic regulator: clinical perspectives. *Clinical endocrinology*, 78(4): 489–496, April 2013.

[22] T Duong, E C Pommier, and A B Scheibel. Immunodetection of the amyloid P component in Alzheimer's disease. *Acta neuropathologica*, 78(4):429–437, 1989.

[23] Nancy Swords Jenny, Alice M Arnold, Lewis H Kuller, Russell P Tracy, and Bruce M Psaty. Serum amyloid P and cardiovascular disease in older men and women: results from the Cardiovascular Health Study. *Arteriosclerosis, thrombosis, and vascular biology*, 27(2):352–358, February 2007.

[24] B U Althaus, J J Staub, A Ryff-De Lèche, A Oberhänsli, and H B Stähelin. LDL/HDL-changes in subclinical hypothyroidism: possible risk factors for coronary heart disease. *Clinical endocrinology*, 28(2):157–163, February 1988.

[25] Lars J Jensen, Michael Kuhn, Manuel Stark, Samuel Chaffron, Chris Creevey, Jean Muller, Tobias Doerks, Philippe Julien, Alexander Roth, Milan Simonovic, Peer Bork, and Christian von Mering. STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research*, 37(Database issue):D412–6, January 2009.

[26] N Li, N H Wallén, M Ladjevardi, and P Hjemdahl. Effects of serotonin on platelet activation in whole blood. *Blood coagulation & fibrinolysis : an international journal in haemostasis and thrombosis*, 8(8):517–523, November 1997.

[27] W H Sauer, J A Berlin, and S E Kimmel. Selective serotonin reuptake inhibitors and myocardial infarction. *Circulation*, 104(16):1894–1898, October 2001.

[28] Min Li, Baohong Wang, Menghui Zhang, Mattias Rantalainen, Shengyue Wang, Haokui Zhou, Yan Zhang, Jian Shen, Xiaoyan Pang, Meiling Zhang, Hua Wei, Yu Chen, Haifeng Lu, Jian Zuo, Mingming Su, Yunping Qiu, Wei Jia, Chaoni Xiao, Leon M Smith, Shengli Yang, Elaine Holmes, Huiru Tang, Guoping Zhao, Jeremy K Nicholson, Lanjuan Li, and Liping Zhao. Symbiotic gut microbes modulate human metabolic phenotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 105(6):2117–2122, February 2008.

[29] Ruben Poesen, Kathleen Claes, Pieter Evenepoel, Henriette de Loor, Patrick Augustijns, Dirk Kuypers, and Björn Meijers. Microbiota-Derived Phenylacetylglutamine Associates with Overall Mortality and Cardiovascular Disease in Patients with CKD. *Journal of the American Society of Nephrology*, 27(11):3479–3487, May 2016.

[30] C Manichanh, L Rigottier-Gois, E Bonnaud, K Gloux, E Pelletier, L Frangeul, R Nalin, C Jarrin, P Chardon, P Marteau, J Roca, and J Dore. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut*, 55(2):205–211, February 2006.

[31] Luke Jostins, Stephan Ripke, Rinse K Weersma, Richard H Duerr, Dermot P McGovern, Ken Y Hui, James C Lee, L Philip Schumm, Yashoda Sharma, Carl A Anderson, Jonah Essers, Mitja Mitrovic, Kaida Ning, Isabelle Cleynen, Emilie Theatre, Sarah L Spain, Soumya Raychaudhuri, Philippe Goyette, Zhi Wei, Clara Abraham, Jean-Paul Achkar, Tariq Ahmad, Leila Amininejad, Ashwin N Ananthakrishnan, Vibeke Andersen, Jane M Andrews, Leonard Baidoo, Tobias Balschun, Peter A Bampton, Alain Bitton, Gabrielle Boucher, Stephan Brand, Carsten Büning, Ariella Cohain, Sven Cichon, Mauro D'Amato, Dirk De Jong, Kathy L Devaney, Marla Dubinsky, Cathryn Edwards, David Ellinghaus, Lynnette R Ferguson, Denis Franchimont, Karin Fransen, Richard Gearry, Michel Georges, Christian Gieger, Jürgen Glas, Talin Haritunians, Ailsa Hart, Chris Hawkey, Matija Hedl, Xinli Hu, Tom H Karlsen, Limas Kupcinskas, Subra Kugathasan, Anna Latiano, Debby Laukens, Ian C Lawrance, Charlie W Lees, Edouard

Louis, Gillian Mahy, John Mansfield, Angharad R Morgan, Craig Mowat, William Newman, Orazio Palmieri, Cyriel Y Ponsioen, Uros Potocnik, Natalie J Prescott, Miguel Regueiro, Jerome I Rotter, Richard K Russell, Jeremy D Sanderson, Miquel Sans, Jack Satsangi, Stefan Schreiber, Lisa A Simms, Jurgita Sventoraityte, Stephan R Targan, Kent D Taylor, Mark Tremelling, Hein W Verspaget, Martine De Vos, Cisca Wijmenga, David C Wilson, Juliane Winkelmann, Ramnik J Xavier, Sebastian Zeissig, Bin Zhang, Clarence K Zhang, Hongyu Zhao, International IBD Genetics Consortium (IIBDGC), Mark S Silverberg, Vito Annese, Hakon Hakonarson, Steven R Brant, Graham Radford-Smith, Christopher G Mathew, John D Rioux, Eric E Schadt, Mark J Daly, Andre Franke, Miles Parkes, Severine Vermeire, Jeffrey C Barrett, and Judy H Cho. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124, November 2012.

[32] Ryan Bradley, Annette L Fitzpatrick, Nancy S Jenny, Duk-Hee Lee, and David R Jacobs. Associations between total serum GGT activity and metabolic risk: MESA. *Biomarkers in medicine*, 7(5):709–721, October 2013.

[33] Ji-Sun Lim, Duk-Hee Lee, Joo-Yun Park, Soo-Hee Jin, and David R Jacobs. A strong interaction between serum gamma-glutamyltransferase and obesity on the risk of prevalent type 2 diabetes: results from the Third National Health and Nutrition Examination Survey. *Clinical chemistry*, 53(6):1092–1098, June 2007.

[34] Roland Lang, Andrew L Gundlach, and Barbara Kofler. The galanin peptide family: receptor pharmacology, pleiotropic biological actions, and implications in health and disease. *Pharmacology & therapeutics*, 115(2):177–207, August 2007.

[35] Kacey J Prentice, Lemieux Luu, Emma M Allister, Ying Liu, Lucy S Jun, Kyle W Sloop, Alexandre B Hardy, Li Wei, Weiping Jia, I George Fantus, Douglas H Sweet, Gary Sweeney, Ravi Retnakaran, Feihan F Dai, and Michael B Wheeler. The furan

fatty acid metabolite CMPF is elevated in diabetes and induces β cell dysfunction. *Cell metabolism*, 19(4):653–666, April 2014.

[36] Ju-Sheng Zheng, Mei Lin, Fumiaki Imamura, Wenwen Cai, Ling Wang, Jue-Ping Feng, Yue Ruan, Jun Tang, Fenglei Wang, Hong Yang, and Duo Li. Serum metabolomics profiles in response to n-3 fatty acids in Chinese patients with type 2 diabetes: a double-blind randomised controlled trial. *Scientific Reports*, 6:29522, July 2016.

[37] Weihua Guan, Brian T Steffen, Rozenn N Lemaitre, Jason H Y Wu, Toshiko Tanaka, Ani Manichaikul, Millennia Foy, Stephen S Rich, Lu Wang, Jennifer A Nettleton, Weihong Tang, Xiangjun Gu, Stafania Bandinelli, Irena B King, Barbara McKnight, Bruce M Psaty, David Siscovick, Luc Djousse, Yii-Der Ida Chen, Luigi Ferrucci, Myriam Fornage, Dariush Mozafarrian, Michael Y Tsai, and Lyn M Steffen. Genome-wide association study of plasma N6 polyunsaturated fatty acids within the cohorts for heart and aging research in genomic epidemiology consortium. *Circulation. Cardiovascular genetics*, 7(3):321–331, June 2014.

[38] Tae-Wook Kang, Hee-Jin Kim, Hyoungseok Ju, Jeong-Hwan Kim, Yeo-Jin Jeon, Han-Chul Lee, Ka-Kyung Kim, Jong-Won Kim, Siwoo Lee, Jong Yeol Kim, Seon-Young Kim, and Yong Sung Kim. Genome-wide association of serum bilirubin levels in Korean population. *Human molecular genetics*, 19(18):3672–3678, September 2010.

[39] Nathaniel Rothman, Montserrat Garcia-Closas, Nilanjan Chatterjee, Nuria Malats, Xifeng Wu, Jonine D Figueroa, Francisco X Real, David Van Den Berg, Giuseppe Matullo, Dalsu Baris, Michael Thun, Lambertus A Kiemeney, Paolo Vineis, Immaculata De Vivo, Demetrius Albanes, Mark P Purdue, Thorunn Rafnar, Michelle A T Hildebrandt, Anne E Kiltie, Olivier Cussenot, Klaus Golka, Rajiv Kumar, Jack A Taylor, Jose I Mayordomo, Kevin B Jacobs, Manolis Kogevinas, Amy Hutchinson, Zhaoming Wang, Yi-Ping Fu, Ludmila Prokunina-Olsson, Laurie Burdett, Meredith Yeager, William Wheeler, Adonina Tardón, Consol Serra, Alfredo Carrato, Reina

García-Closas, Josep Lloreta, Alison Johnson, Molly Schwenn, Margaret R Karagas, Alan Schned, Gerald Andriole, Robert Grubb, Amanda Black, Eric J Jacobs, W Ryan Diver, Susan M Gapstur, Stephanie J Weinstein, Jarmo Virtamo, Victoria K Cortessis, Manuela Gago-Dominguez, Malcolm C Pike, Mariana C Stern, Jian-Min Yuan, David J Hunter, Monica McGrath, Colin P Dinney, Bogdan Czerniak, Meng Chen, Hushan Yang, Sita H Vermeulen, Katja K Aben, J Alfred Witjes, Remco R Makkinje, Patrick Sulem, Soren Besenbacher, Kari Stefansson, Elio Riboli, Paul Brennan, Salvatore Panico, Carmen Navarro, Naomi E Allen, H Bas Bueno-de Mesquita, Dimitrios Trichopoulos, Neil Caporaso, Maria Teresa Landi, Federico Canzian, Borje Ljungberg, Anne Tjonneland, Francoise Clavel-Chapelon, David T Bishop, Mark T W Teo, Margaret A Knowles, Simonetta Guarrera, Silvia Polidoro, Fulvio Ricceri, Carlotta Sacerdote, Alessandra Allione, Geraldine Cancel-Tassin, Silvia Selinski, Jan G Hengstler, Holger Dietrich, Tony Fletcher, Peter Rudnai, Eugen Gurzau, Kvetoslava Koppova, Sophia C E Bolick, Ashley Godfrey, Zongli Xu, José I Sanz-Velez, María D García-Prats, Manuel Sanchez, Gabriel Valdivia, Stefano Porru, Simone Benhamou, Robert N Hoover, Joseph F Fraumeni, Debra T Silverman, and Stephen J Chanock. A multistage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nature genetics*, 42(11):978–984, November 2010.

[40] H Okkels, T Sigsgaard, H Wolf, and H Autrup. Arylamine N-acetyltransferase 1 (NAT1) and 2 (NAT2) polymorphisms in susceptibility to bladder cancer: the influence of smoking. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 6(4):225–231, April 1997.

[41] Diabetes Prevention Program Research Group. 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. *The Lancet*, 374 (9702):1677–1686, November 2009.

[42] Paul C Adams, David M Reboussin, James C Barton, Christine E McLaren, John H

Eckfeldt, Gordon D McLaren, Fitzroy W Dawkins, Ronald T Acton, Emily L Harris, Victor R Gordeuk, Catherine Leiendecker-Foster, Mark Speechley, Beverly M Snively, Joan L Holup, Elizabeth Thomson, Phyliss Sholinsky, and Hemochromatosis and Iron Overload Screening (HEIRS) Study Research Investigators. Hemochromatosis and iron-overload screening in a racially diverse population. *The New England journal of medicine*, 352(17):1769–1778, April 2005.

[43] Gregory Gaich, Jenny Y Chien, Haoda Fu, Leonard C Glass, Mark A Deeg, William L Holland, Alexei Kharitonenkov, Thomas Bumol, Holger K Schilske, and David E Moller. The effects of LY2405319, an FGF21 analog, in obese human subjects with type 2 diabetes. *Cell metabolism*, 18(3):333–340, September 2013.

[44] C Meier, J J Staub, C B Roth, M Guglielmetti, M Kunz, A R Miserez, J Drewe, P Huber, R Herzog, and B Müller. TSH-controlled L-thyroxine therapy reduces cholesterol levels and clinical symptoms in subclinical hypothyroidism: a double blind, placebo-controlled trial (Basel Thyroid Study). *The Journal of clinical endocrinology and metabolism*, 86(10):4860–4866, October 2001.

[45] Elfriede Ruttmann, Larry J Brant, Hans Concin, Günter Diem, Kilian Rapp, Hanno Ulmer, and Vorarlberg Health Monitoring and Promotion Program Study Group. Gamma-glutamyltransferase as a risk factor for cardiovascular disease mortality: an epidemiological investigation in a cohort of 163,944 Austrian adults. *Circulation*, 112 (14):2130–2137, October 2005.

[46] G A Thompson and A Meister. Interrelationships between the binding sites for amino acids, dipeptides, and gamma-glutamyl donors in gamma-glutamyl transpeptidase. *The Journal of biological chemistry*, 252(19):6792–6798, October 1977.

[47] Tomoyoshi Soga, Masahiro Sugimoto, Masashi Honma, Masayo Mori, Kaori Igarashi, Kasumi Kashikura, Satsuki Ikeda, Akiyoshi Hirayama, Takehito Yamamoto, Haruhiko Yoshida, Motoyuki Otsuka, Shoji Tsuji, Yutaka Yatomi, Tadayuki Sakuragawa,

Hisayoshi Watanabe, Kouei Nihei, Takafumi Saito, Sumio Kawata, Hiroshi Suzuki, Masaru Tomita, and Makoto Suematsu. Serum metabolomics reveals γ-glutamyl dipeptides as biomarkers for discrimination among different forms of liver disease. *Journal of Hepatology*, 55(4):896–905, October 2011.

[48] Angela J Rogers, Michael McGeachie, Rebecca M Baron, Lee Gazourian, Jeffrey A Haspel, Kiichi Nakahira, Laura E Fredenburgh, Gary M Hunninghake, Benjamin A Raby, Michael A Matthay, Ronny M Otero, Vance G Fowler, Emanuel P Rivers, Christopher W Woods, Stephen Kingsmore, Ray J Langley, and Augustine M K Choi. Metabolomic derangements are associated with mortality in critically ill adult patients. *PloS one*, 9(1):e87538, 2014.

[49] B Sido, V Hack, A Hochlehnert, H Lipps, C Herfarth, and W Dröge. Impairment of intestinal glutathione synthesis in patients with inflammatory bowel disease. *Gut*, 42 (4):485–492, April 1998.

[50] Lining Guo, Michael V Milburn, John A Ryals, Shaun C Lonergan, Matthew W Mitchell, Jacob E Wulff, Danny C Alexander, Anne M Evans, Brandi Bridgewater, Luke Miller, Manuel L Gonzalez-Garay, and C Thomas Caskey. Plasma metabolomic profiles enhance precision medicine for volunteers of normal health. *Proceedings of the National Academy of Sciences of the United States of America*, 112(35):E4901–10, September 2015.

[51] Leroy Hood and Nathan D Price. Demystifying disease, democratizing health care. *Science translational medicine*, 6(225):225ed5–225ed5, February 2014.

[52] Vin Tangpricha, Elizabeth N Pearce, Tai C Chen, and Michael F Holick. Vitamin D insufficiency among free-living healthy young adults. *The American journal of medicine*, 112(8):659–662, June 2002.

[53] Christine M Micheel, Sharly J Nass, and Gilbert S Omenn. *Evolution of Translational*

Omics: Lessons Learned and the Path Forward. National Academies Press (US), Washington (DC), March 2012.

[54] Jai Ram Rideout, Yan He, Jose A Navas-Molina, William A Walters, Luke K Ursell, Sean M Gibbons, John Chase, Daniel McDonald, Antonio Gonzalez, Adam Robbins-Pianka, Jose C Clemente, Jack A Gilbert, Susan M Huse, Hong-Wei Zhou, Rob Knight, and J Gregory Caporaso. Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ*, 2(5):e545, 2014.

[55] R H Whittaker. Evolution and Measurement of Species Diversity. *Taxon*, 21(2/3):213, May 1972.

[56] J Gregory Caporaso, Kyle Bittinger, Frederic D Bushman, Todd Z DeSantis, Gary L Andersen, and Rob Knight. PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics (Oxford, England)*, 26(2):266–267, January 2010.

[57] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttley, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel McDonald, Brian D Muegge, Meg Pirrung, Jens Reeder, Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld, and Rob Knight. QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335–336, May 2010.

[58] Daniel McDonald, Morgan N Price, Julia Goodrich, Eric P Nawrocki, Todd Z DeSantis, Alexander Probst, Gary L Andersen, Rob Knight, and Philip Hugenholtz. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME journal*, 6(3):610–618, March 2012.

[59] D J Lane. *16S/23S rRNA sequencing. In 'Nucleic acid techniques in bacterial systematics'*. 1991.

[60] Catherine Lozupone, Manuel E Lladser, Dan Knights, Jesse Stombaugh, and Rob Knight. Unifrac: an effective distance metric for microbial community comparison. *The ISME journal*, 5(2):169–172, 2011.

[61] Micah Hamady, Catherine Lozupone, and Rob Knight. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *The ISME journal*, 4(1):17–27, January 2010.

[62] Catherine Lozupone and Rob Knight. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, 71(12):8228–8235, December 2005.

[63] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90–95, 2007.

[64] Brendan MacLean, Daniela M Tomazela, Nicholas Shulman, Matthew Chambers, Gregory L Finney, Barbara Frewen, Randall Kern, David L Tabb, Daniel C Liebler, and Michael J MacCoss. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics (Oxford, England)*, 26(7): 966–968, April 2010.

[65] Gustavo Glusman, Alissa Severson, Varsha Dhankani, Max Robinson, Terry Farrah, Denise E Mauldin, Anna B Stittrich, Seth A Ament, Jared C Roach, Mary E Brunkow, Dale L Bodian, Joseph G Vockley, Ilya Shmulevich, John E Niederhuber, and Leroy Hood. Identification of copy number variants in whole-genome data using Reference Coverage Profiles. *Frontiers in genetics*, 6:45, 2015.

[66] Gustavo Glusman, Juan Caballero, Denise E Mauldin, Leroy Hood, and Jared C Roach.

Kaviar: an accessible system for testing SNV novelty. *Bioinformatics (Oxford, England)*, 27(22):3216–3217, November 2011.

[67] Fernando Pérez and Brian E Granger. IPython: A System for Interactive Scientific Computing. *Computing in Science & Engineering*, 9(3):21–29, May 2007.

[68] Ulrike Kusebauch, David S Campbell, Eric W Deutsch, Caroline S Chu, Douglas A Spicer, Mi-Youn Brusniak, Joseph Slagel, Zhi Sun, Jeffrey Stevens, Barbara Grimes, David Shteynberg, Michael R Hoopmann, Peter Blattmann, Alexander V Ratushny, Oliver Rinner, Paola Picotti, Christine Carapito, Chung-Ying Huang, Meghan Kapousouz, Henry Lam, Tommy Tran, Emek Demir, John D Aitchison, Chris Sander, Leroy Hood, Ruedi Aebersold, and Robert L Moritz. Human SRMAtlas: A Resource of Targeted Assays to Quantify the Complete Human Proteome. *Cell*, 166(3):766–778, July 2016.

[69] S Seabold and J Perktold. Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science ...*, pages 57–61, 2010.

[70] Robert C Green, Jonathan S Berg, Wayne W Grody, Sarah S Kalia, Bruce R Korf, Christa L Martin, Amy L McGuire, Robert L Nussbaum, Julianne M O'Daniel, Kelly E Ormond, Heidi L Rehm, Michael S Watson, Marc S Williams, Leslie G Biesecker, and American College of Medical Genetics and Genomics. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Springer Nature, July 2013.

[71] National Research Council and Others. *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease*. National Academies Press, 2011. doi: citeulike-article-id:14615917.

[72] Alex Comfort. Test-battery to measure ageing-rate in man. *The Lancet*, 294(7635): 1411–1415, 1969. doi: citeulike-article-id:14615968.

[73] George Baker and Richard Sprott. Biomarkers of aging. *Experimental gerontology*, 23 (4):223–239, 1988. doi: citeulike-article-id:14615969.

[74] Steve Horvath. Dna methylation age of human tissues and cell types. *Genome Biology*, 14(10):3156, 2013. doi: citeulike-article-id:14615965doi:10.1186/gb-2013-14-10-r115. URL http://dx.doi.org/10.1186/gb-2013-14-10-r115.

[75] Toshiko Tanaka, Angelique Biancotto, Ruin Moaddel, Ann Zenobia Moore, Marta Gonzalez-Freire, Miguel A Aon, Julián Candia, Pingbo Zhang, Foo Cheung, Giovanna Fantoni, et al. Plasma proteomic signature of age in healthy humans. *Aging Cell*, 17 (5):e12799, 2018.

[76] Zichen Wang, Li Li, Benjamin S Glicksberg, Ariel Israel, Joel T Dudley, and Avi Ma'ayan. Predicting age by mining electronic medical records with deep learning characterizes differences between chronological and physiological age. *Journal of biomedical informatics*, 76:59–68, 2017.

[77] Petr Klemera and Stanislav Doubal. A new approach to the concept and computation of biological age. *Mechanisms of ageing and development*, 127(3):240–248, 2006. doi: citeulike-article-id:14615934.

[78] Morgan Levine. Modeling the rate of senescence: can estimated biological age predict mortality more accurately than chronological age? *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 68(6):667–674, 2012. doi: citeulike-article-id:14615932.

[79] Daniel Belsky, Avshalom Caspi, Renate Houts, Harvey Cohen, David Corcoran, Andrea Danese, HonaLee Harrington, Salomon Israel, Morgan Levine, Jonathan Schaefer, and Others. Quantification of biological aging in young adults. *Proceedings of the National Academy of Sciences*, 112(30):E4104–E4110, 2015. doi: citeulike-article-id:14615933.

[80] A. Mitnitski, S. E. Howlett, and K. Rockwood. Heterogeneity of human aging and its assessment. *J Gerontol A Biol Sci Med Sci*, 72(7):877–884, 2017. ISSN 1758-535X (Electronic) 1079-5006 (Linking). doi: 10.1093/gerona/glw089. URL https://www.ncbi.nlm.nih.gov/pubmed/27216811.

[81] J. M. Murabito, Q. Zhao, M. G. Larson, J. Rong, H. Lin, E. J. Benjamin, D. Levy, and K. L. Lunetta. Measures of biologic age in a community sample predict mortality and age-related disease: The framingham offspring study. *J Gerontol A Biol Sci Med Sci*, 73(6):757–762, 2018. ISSN 1758-535X (Electronic) 1079-5006 (Linking). doi: 10.1093/gerona/glx144. URL https://www.ncbi.nlm.nih.gov/pubmed/28977464.

[82] H. Lin, K. L. Lunetta, Q. Zhao, P. R. Mandaviya, J. Rong, E. J. Benjamin, R. Joehanes, D. Levy, J. B. J. van Meurs, M. G. Larson, and J. M. Murabito. Whole blood gene expression associated with clinical biological age. *J Gerontol A Biol Sci Med Sci*, 2018. ISSN 1758-535X (Electronic) 1079-5006 (Linking). doi: 10.1093/gerona/gly164. URL https://www.ncbi.nlm.nih.gov/pubmed/30010802.

[83] Daniel W Belsky, Kim M Huffman, Carl F Pieper, Idan Shalev, and William E Kraus. Change in the rate of biological aging in response to caloric restriction: Calerie biobank analysis. *The Journals of Gerontology: Series A*, 73(1):4–10, 2018.

[84] R. Chen, G. I. Mias, J. Li-Pook-Than, L. Jiang, H. Y. K. Lam, R. Chen, E. Miriami, K. J. Karczewski, M. Hariharan, F. E. Dewey, Y. Cheng, M. J. Clark, H. Im, L. Habegger, S. Balasubramanian, M. O'Huallachain, J. T. Dudley, S. Hillenmeyer, R. Haraksingh, D. Sharon, G. Euskirchen, P. Lacroute, K. Bettinger, A. P. Boyle, M. Kasowski, F. Grubert, S. Seki, M. Garcia, M. Whirl-Carrillo, M. Gallardo, M. A. Blasco, P. L. Greenberg, P. Snyder, T. E. Klein, R. B. Altman, A. Butte, E. A. Ashley, K. C. Nadeau, M. Gerstein, H. Tang, and M. Snyder. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 148(6):

1293–307, 2012. doi: citeulike-article-id:14615920doi:10.1016/j.cell.2012.02.009. URL http://dx.doi.org/10.1016/j.cell.2012.02.009.

[85] Niha Zubair, Matthew P Conomos, Leroy Hood, Gilbert S Omenn, Nathan D Price, Bonnie J Spring, Andrew T Magis, and Jennifer C Lovejoy. Genetic predisposition impacts clinical changes in a lifestyle coaching program. *Scientific reports*, 9(1):1–11, 2019.

[86] Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986. ISSN 0006-3444.

[87] Cassandra M Pickens, Carol Pierannunzi, William Garvin, and Machell Town. Surveillance for certain health behaviors and conditions among states and selected local areasbehavioral risk factor surveillance system, united states, 2015. *MMWR Surveillance Summaries - Centers for Disease Control and Prevention*, 67(9):1, 2018.

[88] Adrian G Barnett, Jolieke C Van Der Pols, and Annette J Dobson. Regression to the mean: what it is and how to deal with it. *International journal of epidemiology*, 34(1): 215–220, 2005.

[89] J. Leal, A. M. Gray, and P. M. Clarke. Development of life-expectancy tables for people with type 2 diabetes. *Eur Heart J*, 30(7):834–9, 2009. ISSN 1522-9645 (Electronic) 0195-668X (Linking). doi: 10.1093/eurheartj/ehn567. URL https://www.ncbi.nlm.nih.gov/pubmed/19109355.

[90] Wenya Yang, Timothy M. Dall, and Kaleigh Beronjia. Economic costs of diabetes in the u.s. in 2017. *Diabetes Care*, 41(5):917–928, 2018. ISSN 0149-5992. doi: 10.2337/dci18-0007. URL https://care.diabetesjournals.org/content/41/5/917.

[91] Edward W Gregg, Xiaohui Zhuo, Yiling J Cheng, Ann L Albright, KM Venkat Narayan, and Theodore J Thompson. Trends in lifetime risk and years of life lost due to diabetes

in the usa, 19852011: a modelling study. *The lancet Diabetes & endocrinology*, 2(11): 867–874, 2014. ISSN 2213-8587.

[92] I. Shimokawa and Y. Higami. Leptin signaling and aging: insight from caloric restriction. *Mech Ageing Dev*, 122(14):1511–9, 2001. ISSN 0047-6374 (Print) 0047-6374 (Linking). URL https://www.ncbi.nlm.nih.gov/pubmed/11511393.

[93] CP Shen, KK Wu, LP Shearman, R Camacho, MR Tota, TM Fong, and LHT Van der Ploeg. Plasma agoutirelated protein level: a possible correlation with fasted and fed states in humans and rats. *Journal of neuroendocrinology*, 14(8):607–610, 2002. ISSN 0953-8194.

[94] Jennifer Poehls, CL Wassel, Tamara B Harris, Peter J Havel, Michael M Swarbrick, Steven R Cummings, Anne B Newman, Suzanne Satterfield, Alka M Kanaya, et al. Association of adiponectin with mortality in older adults: the health, aging, and body composition study. *Diabetologia*, 52(4):591–595, 2009.

[95] A. Dregan, J. Charlton, P. Chowienczyk, and M. C. Gulliford. Chronic inflammatory disorders and risk of type 2 diabetes mellitus, coronary heart disease, and stroke: a population-based cohort study. *Circulation*, 130(10):837–44, 2014. ISSN 1524-4539 (Electronic) 0009-7322 (Linking). doi: 10.1161/CIRCULATIONAHA.114.009990. URL https://www.ncbi.nlm.nih.gov/pubmed/24970784.

[96] D. Weiskopf, B. Weinberger, and B. Grubeck-Loebenstein. The aging of the immune system. *Transpl Int*, 22(11):1041–50, 2009. ISSN 1432-2277 (Electronic) 0934-0874 (Linking). doi: 10.1111/j.1432-2277.2009.00927.x. URL https://www.ncbi.nlm.nih.gov/pubmed/19624493.

[97] R. Altara, M. Manca, M. H. Hessel, Y. Gu, L. C. van Vark, K. M. Akkerhuis, J. A. Staessen, H. A. Struijker-Boudier, G. W. Booz, and W. M. Blankesteijn. Cxcl10 is a circulating inflammatory marker in patients with advanced heart failure: a pilot study.

J Cardiovasc Transl Res, 9(4):302–14, 2016. ISSN 1937-5395 (Electronic) 1937-5387 (Linking). doi: 10.1007/s12265-016-9703-3. URL https://www.ncbi.nlm.nih.gov/pubmed/27271043.

[98] R. Altara, Y. M. Gu, H. A. Struijker-Boudier, L. Thijs, J. A. Staessen, and W. M. Blankesteijn. Left ventricular dysfunction and cxcr3 ligands in hypertension: From animal experiments to a population-based pilot study. *PLoS One*, 10(10):e0141394, 2015. ISSN 1932-6203 (Electronic) 1932-6203 (Linking). doi: 10.1371/journal.pone.0141394. URL https://www.ncbi.nlm.nih.gov/pubmed/26506526.

[99] Sara Bouajila, Kegan Moneghetti, Yukari Kobayashi, Fatemeh A Gomari, Fahim Abbasi, Mark M Davis, Joseph C Wu, Tatiana Kuznetsova, Nazish Sayed, and Francois Haddad. Immune profile of healthy cardiovascular aging: Insights from a population-based study and network modeling. *Journal of the American College of Cardiology*, 71 (11):A1657, 2018. ISSN 0735-1097.

[100] K. Takei, S. Ikeda, T. Arai, N. Tanaka, M. Muramatsu, and M. Sawabe. Lymphotoxin-alpha polymorphisms and presence of cancer in 1,536 consecutive autopsy cases. *BMC Cancer*, 8:235, 2008. ISSN 1471-2407 (Electronic) 1471-2407 (Linking). doi: 10.1186/1471-2407-8-235. URL https://www.ncbi.nlm.nih.gov/pubmed/18700950.

[101] M. C. Houston. Role of mercury toxicity in hypertension, cardiovascular disease, and stroke. *J Clin Hypertens (Greenwich)*, 13(8):621–7, 2011. ISSN 1751-7176 (Electronic) 1524-6175 (Linking). doi: 10.1111/j.1751-7176.2011.00489.x. URL https://www.ncbi.nlm.nih.gov/pubmed/21806773.

[102] J. M. Donohue, T. M. Duke, and J. Wambaugh. Health effects support document for perfluorooctanoic acid. *Environmental Protection Agency, USA*, May 2016.

[103] K. Mukaiyama, M. Kamimura, S. Uchiyama, S. Ikegami, Y. Nakamura, and H. Kato. Elevation of serum alkaline phosphatase (alp) level in postmenopausal women is

caused by high bone turnover. *Aging Clin Exp Res*, 27(4):413–8, 2015. ISSN 1720-8319 (Electronic) 1594-0667 (Linking). doi: 10.1007/s40520-014-0296-x. URL https://www.ncbi.nlm.nih.gov/pubmed/25534961.

[104] P. Urena, M. Hruby, A. Ferreira, K. S. Ang, and M. C. de Vernejoul. Plasma total versus bone alkaline phosphatase as markers of bone turnover in hemodialysis patients. *J Am Soc Nephrol*, 7(3):506–12, 1996. ISSN 1046-6673 (Print) 1046-6673 (Linking). URL https://www.ncbi.nlm.nih.gov/pubmed/8704118.

[105] B. L. Riggs, 3rd Melton Iii, L. J., R. A. Robb, J. J. Camp, E. J. Atkinson, J. M. Peterson, P. A. Rouleau, C. H. McCollough, M. L. Bouxsein, and S. Khosla. Population-based study of age and sex differences in bone volumetric density, size, geometry, and structure at different skeletal sites. *J Bone Miner Res*, 19(12):1945–54, 2004. ISSN 0884-0431 (Print) 0884-0431 (Linking). doi: 10.1359/JBMR.040916. URL https://www.ncbi.nlm.nih.gov/pubmed/15537436.

[106] Sapna S Patel, Miklos Z Molnar, John A Tayek, Joachim H Ix, Nazanin Noori, Deborah Benner, Steven Heymsfield, Joel D Kopple, Csaba P Kovesdy, and Kamyar Kalantar-Zadeh. Serum creatinine as a marker of muscle mass in chronic kidney disease: results of a crosssectional study and review of literature. *Journal of cachexia, sarcopenia and muscle*, 4(1):19–29, 2013. ISSN 2190-5991.

[107] Andrew D Rule, Kent R Bailey, Gary L Schwartz, Sundeep Khosla, John C Lieske, and L Joseph Melton III. For estimating creatinine clearance measuring muscle mass gives better results than those based on demographics. *Kidney international*, 75(10): 1071–1078, 2009. ISSN 0085-2538.

[108] J. E. Schutte, J. C. Longhurst, F. A. Gaffney, B. C. Bastian, and C. G. Blomqvist. Total plasma creatinine: an accurate measure of total striated muscle mass. *J Appl Physiol Respir Environ Exerc Physiol*, 51(3):762–6, 1981. ISSN 0161-7567 (Print) 0161-

7567 (Linking). doi: 10.1152/jappl.1981.51.3.762. URL https://www.ncbi.nlm.nih.gov/pubmed/7327978.

[109] M. Iannuzzi-Sucich, K. M. Prestwood, and A. M. Kenny. Prevalence of sarcopenia and predictors of skeletal muscle mass in healthy, older men and women. *J Gerontol A Biol Sci Med Sci*, 57(12):M772–7, 2002. ISSN 1079-5006 (Print) 1079-5006 (Linking). URL https://www.ncbi.nlm.nih.gov/pubmed/12456735.

[110] J. J. Lee and I. R. Schultz. Sex differences in the uptake and disposition of perfluorooctanoic acid in fathead minnows after oral dosing. *Environ Sci Technol*, 44(1): 491–6, 2010. ISSN 0013-936X (Print) 0013-936X (Linking). doi: 10.1021/es901838y. URL https://www.ncbi.nlm.nih.gov/pubmed/19961202.

[111] Naomi Kudo and Yoichi Kawashima. Toxicity and toxicokinetics of perfluorooctanoic acid in humans and animals. *The Journal of toxicological sciences*, 28(2):49–57, 2003. ISSN 0388-1350.

[112] Lucy L Gao, Jacob Bien, and Daniela Witten. Are clusterings of multiple data views independent? *Biostatistics*, 2019.

[113] Johannes Hertel, Stefan Frenzel, Johanna Koenig, Katharina Wittfeld, Georg Fuellen, Birte Holtfreter, Maik Pietzner, Nele Friedrich, Matthias Nauck, Henry Voelzke, et al. The informative error: a framework for the construction of individualized phenotypes. *Statistical methods in medical research*, 28(5):1427–1438, 2019.

[114] Younhee Ko, Seth A Ament, James A Eddy, Juan Caballero, John C Earls, Leroy Hood, and Nathan D Price. Cell type-specific genes show striking and distinct patterns of spatial expression in the mouse brain. *Proceedings of the National Academy of Sciences*, 110(8):3095–3100, 2013.

[115] Seth A Ament, Jocelynn R Pearl, Andrea Grindeland, Jason St. Claire, John C Earls, Marina Kovalenko, Tammy Gillis, Jayalakshmi Mysore, James F Gusella, Jong-Min

Lee, et al. High resolution time-course mapping of early transcriptomic, molecular and cellular phenotypes in huntingtons disease cag knock-in mice across multiple genetic backgrounds. *Human Molecular Genetics*, 26(5):913–922, 2017.

[116] Dhimankrishna Ghosh, Cory C Funk, Juan Caballero, Nameeta Shah, Katherine Rouleau, John C Earls, Liliana Soroceanu, Greg Foltz, Charles S Cobbs, Nathan D Price, et al. A cell-surface membrane protein signature for glioblastoma. *Cell Systems*, 4(5):516–529, 2017.

[117] Yang Yang, Leng Han, Yuan Yuan, Jun Li, Nainan Hei, and Han Liang. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature communications*, 5:3231, 2014.

[118] Neelroop N Parikshak, Michael J Gandal, and Daniel H Geschwind. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nature Reviews Genetics*, 16(8):441–458, 2015.

[119] Jennifer L Gardy, David J Lynn, Fiona SL Brinkman, and Robert EW Hancock. Enabling a systems biology approach to immunology: focus on innate immunity. *Trends in immunology*, 30(6):249–262, 2009.

[120] Trey Ideker, Timothy Galitski, and Leroy Hood. A new approach to decoding life: systems biology. *Annual review of genomics and human genetics*, 2(1):343–372, 2001.

[121] Leroy Hood, James R Heath, Michael E Phelps, and Biaoyang Lin. Systems biology and new technologies enable predictive and preventative medicine. *Science*, 306(5696): 640–643, 2004.

[122] Nathan D Price, Lucas B Edelman, Inyoul Lee, Hyuntae Yoo, Daehee Hwang, George Carlson, David J Galas, James R Heath, and Leroy Hood. Systems biology and systems medicine. *Genomic and Personalized Medicine*, 1:60, 2012.

[123] Antonio Gaddi, AFG Cicero, FO Odoo, et al. Practical guidelines for familial combined hyperlipidemia diagnosis: an up-date. *Vascular health and risk management*, 3(6):877, 2007.

[124] Earl S Ford, Wayne H Giles, and William H Dietz. Prevalence of the metabolic syndrome among us adults: findings from the third national health and nutrition examination survey. *Jama*, 287(3):356–359, 2002.

[125] Steven M Haffner, Clicerio Gonzalez, Heikki Miettinen, Esmarie Kennedy, and Michael P Stern. A prospective analysis of the homa model: the mexico city diabetes study. *Diabetes care*, 19(10):1138–1141, 1996.

[126] Graeme S Halford, Rosemary Baker, Julie E McCredden, and John D Bain. How many variables can humans process? *Psychological science*, 16(1):70–76, 2005.

[127] Jamal Ahmad and Abdur R Khan. Insulin resistance, inflammation and atherosclerosis. *Anti-Inflammatory & Anti-Allergy Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Anti-Inflammatory and Anti-Allergy Agents)*, 7(3):172–186, 2008.

[128] Charles Auffray, Zhu Chen, and Leroy Hood. Systems medicine: the future of medical genomics and healthcare. *Genome medicine*, 1(1):2, 2009.

[129] Hiroaki Kitano. Cancer as a robust system: implications for anticancer therapy. *Nature Reviews Cancer*, 4(3):227–235, 2004.

[130] Jorrit J Hornberg, Frank J Bruggeman, Hans V Westerhoff, and Jan Lankelma. Cancer: a systems biology disease. *Biosystems*, 83(2):81–90, 2006.

[131] Qinghua Cui, Yun Ma, Maria Jaramillo, Hamza Bari, Arif Awan, Song Yang, Simo Zhang, Lixue Liu, Meng Lu, Maureen O'Connor-McCourt, et al. A map of human cancer signaling. *Molecular systems biology*, 3(1):152, 2007.

[132] Edison T Liu, Vladimir A Kuznetsov, and Lance D Miller. In the pursuit of complexity: systems medicine in cancer biology. *Cancer cell*, 9(4):245–247, 2006.

[133] Elisa Sheng. Arivale proteomics data, May 2019.

[134] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[135] David S Wishart, Yannick Djoumbou Feunang, Ana Marcu, An Chi Guo, Kevin Liang, Rosa Vázquez-Fresno, Tanvir Sajed, Daniel Johnson, Carin Li, Naama Karu, et al. Hmdb 4.0: the human metabolome database for 2018. *Nucleic acids research*, 46(D1): D608–D617, 2018.

[136] Yannick Djoumbou Feunang, Roman Eisner, Craig Knox, Leonid Chepelev, Janna Hastings, Gareth Owen, Eoin Fahy, Christoph Steinbeck, Shankar Subramanian, Evan Bolton, et al. Classyfire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of cheminformatics*, 8(1):61, 2016.

[137] Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 45(D1):D158–D169, 2017.

[138] Gene Ontology Consortium. The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl_1):D258–D261, 2004.

[139] Clement J McDonald, Stanley M Huff, Jeffrey G Suico, Gilbert Hill, Dennis Leavelle, Raymond Aller, Arden Forrey, Kathy Mercer, Georges DeMoor, John Hook, et al. Loinc, a universal standard for identifying laboratory observations: a 5-year update. *Clinical chemistry*, 49(4):624–633, 2003.

[140] S Falcon and R Gentleman. Hypergeometric testing used for gene set enrichment analysis. In *Bioconductor case studies*, pages 207–220. Springer, 2008.

[141] Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

[142] Jim Webber. A programmatic introduction to neo4j. In *Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity*, pages 217–218, 2012.

[143] Maxwell Dayvson Da Silva and Hugo Lopes Tavares. *Redis Essentials*. Packt Publishing Ltd, 2015.

[144] Ask Solem et al. Celery: Distributed task queue. *Dostupné z: http://www. celeryproject. org*, 32:33, 2013.

[145] Dirk Merkel. Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, 2014(239):2, 2014.

[146] Wiktoria Wilkowska and Martina Ziefle. Privacy and data security in e-health: Requirements from the user?s perspective. *Health informatics journal*, 18(3):191–201, 2012.

[147] Summary of the hipaa privacy rule. URL http://bit.ly/2oHeWGs.

[148] Joan Daemen and Vincent Rijmen. Aes proposal: Rijndael. 1999.

[149] Kyle Isom. A working introduction to crypto with pycrypto. 2011.

[150] Matthew Campagna. Aws key management service cryptographic details. *Aws key management service cryptographic details*, 2015.

[151] Guido van Rossum and Fred L Drake. pickle–python object serialization. *Available from World Wide Web: http://docs. python. org/lib/module-pickle. html*, 2009.

[152] Ralph E Johnson and Brian Foote. Designing reusable classes. *Journal of object-oriented programming*, 1(2):22–35, 1988.

[153] Alex Rodriguez. Restful web services: The basics. *IBM developerWorks*, 33:18, 2008.

[154] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 42(D1):D472–D477, 2014.

[155] Alex Frolkis, Craig Knox, Emilia Lim, Timothy Jewison, Vivian Law, David D Hau, Phillip Liu, Bijaya Gautam, Son Ly, An Chi Guo, et al. Smpdb: the small molecule pathway database. *Nucleic acids research*, 38(suppl_1):D480–D487, 2010.

[156] Daniel R Zerbino, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, et al. Ensembl 2018. *Nucleic acids research*, 46(D1):D754–D761, 2018.

[157] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[158] Nadime Francis, Alastair Green, Paolo Guagliardo, Leonid Libkin, Tobias Lindaaker, Victor Marsault, Stefan Plantikow, Mats Rydberg, Petra Selmer, and Andrés Taylor. Cypher: An evolving query language for property graphs. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1433–1445, 2018.

[159] Natalya N Pavlova and Craig B Thompson. The emerging hallmarks of cancer metabolism. *Cell metabolism*, 23(1):27–47, 2016.

[160] Ming Yang, Tomoyoshi Soga, Patrick Pollard, and Julie Adam. The emerging role of fumarate as an oncometabolite. *Frontiers in Oncology*, 2:85, 2012. ISSN 2234-

943X. doi: 10.3389/fonc.2012.00085. URL https://www.frontiersin.org/article/10.3389/fonc.2012.00085.

[161] Otto Warburg, K Posener, and EVIII Negelein. The metabolism of cancer cells. *Biochem Z*, 152:319–44, 1924.

[162] Yoel Smicun, Scott Reierstad, Feng-Qiang Wang, Cathy Lee, and David A Fishman. S1p regulation of ovarian carcinoma invasiveness. *Gynecologic oncology*, 103(3):952–959, 2006.

[163] Masayuki Nagahashi, Subramaniam Ramachandran, Eugene Y Kim, Jeremy C Allegood, Omar M Rashid, Akimitsu Yamada, Renping Zhao, Sheldon Milstien, Huiping Zhou, Sarah Spiegel, et al. Sphingosine-1-phosphate produced by sphingosine kinase 1 promotes breast cancer progression by stimulating angiogenesis and lymphangiogenesis. *Cancer research*, 72(3):726–735, 2012.

[164] Shizhong Bu, Bagrat Kapanadze, Tien Hsu, and Maria Trojanowska. Opposite effects of dihydrosphingosine 1-phosphate and sphingosine 1-phosphate on transforming growth factor-β/smad signaling are mediated through the pten/ppm1a-dependent pathway. *Journal of Biological Chemistry*, 283(28):19593–19602, 2008.

[165] Kelsey Hildreth, Sean D Kodani, Bruce D Hammock, and Ling Zhao. Cytochrome p450-derived linoleic acid metabolites epomes and dihomes: A review of recent studies. *The Journal of Nutritional Biochemistry*, page 108484, 2020.

[166] Yonghai Lu, Jinling Fang, Choon Nam Ong, Shengsen Chen, Ning Li, Liang Cui, Chong Huang, Qinxia Ling, Sin Eng Chia, and Mingquan Chen. Targeted analysis of omega-6-derived oxylipins and parent polyunsaturated fatty acids in serum of hepatitis b virus-related hepatocellular carcinoma patients. *Metabolomics*, 13(1):6, 2017.

[167] DK Patneau and ML Mayer. Structure-activity relationships for amino acid transmitter candidates acting at n-methyl-d-aspartate and quisqualate receptors. *Journal*

of Neuroscience, 10(7):2385–2399, 1990. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.10-07-02385.1990. URL https://www.jneurosci.org/content/10/7/2385.

[168] Paldeep S. Atwal and Fernando Scaglia. Molybdenum cofactor deficiency. *Molecular Genetics and Metabolism*, 117(1):1 – 4, 2016. ISSN 1096-7192. doi: https://doi.org/10.1016/j.ymgme.2015.11.010. URL http://www.sciencedirect.com/science/article/pii/S1096719215300809.

[169] Marcelle M Dougan, Yuqing Li, Lisa W Chu, Robert W Haile, Alice S Whittemore, Summer S Han, Steven C Moore, Joshua N Sampson, Irene L Andrulis, Esther M John, et al. Metabolomic profiles in breast cancer: a pilot case-control study in the breast cancer family registry. *BMC cancer*, 18(1):532, 2018.

[170] Steven Castellon and Patricia A Ganz. Neuropsychological studies in breast cancer: in search of chemobrain. *Breast cancer research and treatment*, 116(1):125–127, 2009.

[171] José Angel Santamaria-Araujo, Berthold Fischer, Tanja Otte, Manfred Nimtz, Ralf R Mendel, Victor Wray, and Günter Schwarz. The tetrahydropyranopterin structure of the sulfur-free and metal-free molybdenum cofactor precursor. *Journal of Biological Chemistry*, 279(16):15994–15999, 2004.

[172] Daniel K Nomura, Jonathan Z Long, Sherry Niessen, Heather S Hoover, Shu-Wing Ng, and Benjamin F Cravatt. Monoacylglycerol lipase regulates a fatty acid network that promotes cancer pathogenesis. *Cell*, 140(1):49–61, 2010.

[173] Kristin M Nieman, Hilary A Kenny, Carla V Penicka, Andras Ladanyi, Rebecca Buell-Gutbrod, Marion R Zillhardt, Iris L Romero, Mark S Carey, Gordon B Mills, Gökhan S Hotamisligil, et al. Adipocytes promote ovarian cancer metastasis and provide energy for rapid tumor growth. *Nature medicine*, 17(11):1498, 2011.

[174] Alison Crawford, Lauric Haber, Marcus P Kelly, Kristin Vazzana, Lauren Canova, Priyanka Ram, Arpita Pawashe, Jennifer Finney, Sumreen Jalal, Danica Chiu, et al.

A mucin 16 bispecific t cell–engaging antibody for the treatment of ovarian cancer. *Science Translational Medicine*, 11(497):eaau7534, 2019.

[175] Subhash C Chauhan, Ajay P Singh, Fernanda Ruiz, Sonny L Johansson, Maneesh Jain, Lynette M Smith, Nicolas Moniaux, and Surinder K Batra. Aberrant expression of muc4 in ovarian carcinoma: diagnostic significance alone and in combination with muc1 and muc16 (ca125). *Modern Pathology*, 19(10):1386–1394, 2006.

[176] Shaker Abuharbeid, Frank Czubayko, and Achim Aigner. The fibroblast growth factor-binding protein fgf-bp. *The international journal of biochemistry & cell biology*, 38(9): 1463–1468, 2006.

[177] Nahoko Sakaguchi, Hisako Muramatsu, Keiko Ichihara-Tanaka, Nobuaki Maeda, Masaharu Noda, Tokuo Yamamoto, Makoto Michikawa, Shinya Ikematsu, Sadatoshi Sakuma, and Takashi Muramatsu. Receptor-type protein tyrosine phosphatase ζ as a component of the signaling receptor complex for midkine-dependent survival of embryonic neurons. *Neuroscience research*, 45(2):219–224, 2003.

[178] Hisako Muramatsu, Kun Zou, Nahoko Sakaguchi, Shinya Ikematsu, Sadatoshi Sakuma, and Takashi Muramatsu. Ldl receptor-related protein as a component of the midkine receptor. *Biochemical and biophysical research communications*, 270(3):936–941, 2000.

[179] Kenji Kadomatsu, Satoshi Kishida, and Shoma Tsubota. The heparin-binding growth factor midkine: the biological activities and candidate receptors. *The Journal of Biochemistry*, 153(6):511–521, 2013.

[180] Yiping Huang, Mohammad Obaidul Hoque, Feng Wu, Barry Trink, David Sidransky, and Edward A Ratovitski. Midkine induces epithelial-mesenchymal transition through notch2/jak2-stat3 signaling in human keratinocytes. *Cell cycle*, 7(11):1613–1622, 2008.

[181] Kenji Kadomatsu and Takashi Muramatsu. Midkine and pleiotrophin in neural development and cancer. *Cancer letters*, 204(2):127–143, 2004.

[182] Thomas J Wang, Debby Ngo, Nikolaos Psychogios, Andre Dejam, Martin G Larson, Ramachandran S Vasan, Anahita Ghorbani, John O?Sullivan, Susan Cheng, Eugene P Rhee, et al. 2-aminoadipic acid is a biomarker for diabetes risk. *The Journal of clinical investigation*, 123(10):4309–4317, 2013.

[183] Karin B Michels, Caren G Solomon, Frank B Hu, Bernard A Rosner, Susan E Hankinson, Graham A Colditz, and JoAnn E Manson. Type 2 diabetes and subsequent incidence of breast cancer in the nurses? health study. *Diabetes care*, 26(6):1752–1758, 2003.

[184] Peter Boyle, Mathieu Boniol, A Koechlin, Chris Robertson, Faustine Valentini, Kim Coppens, Laura-Louise Fairley, T Zheng, Y Zhang, M Pasterk, et al. Diabetes and breast cancer risk: a meta-analysis. *British journal of cancer*, 107(9):1608–1617, 2012.

[185] Lorraine L Lipscombe, Pamela J Goodwin, Bernard Zinman, John R McLaughlin, and Janet E Hux. The impact of diabetes on survival following breast cancer. *Breast cancer research and treatment*, 109(2):389–395, 2008.

[186] Lorraine L Lipscombe, Pamela J Goodwin, Bernard Zinman, John R McLaughlin, and Janet E Hux. Increased prevalence of prior breast cancer in women with newly diagnosed diabetes. *Breast cancer research and treatment*, 98(3):303–309, 2006.

[187] Kamal S Saini, Sherene Loi, Evandro de Azambuja, Otto Metzger-Filho, Monika Lamba Saini, Michail Ignatiadis, Janet E Dancey, and Martine J Piccart-Gebhart. Targeting the pi3k/akt/mtor and raf/mek/erk pathways in the treatment of breast cancer. *Cancer treatment reviews*, 39(8):935–946, 2013.

[188] Tingting Wang, Serena Seah, Xinyi Loh, Ching-Wan Chan, Mikael Hartman, Boon-Cher Goh, and Soo-Chin Lee. Simvastatin-induced breast cancer cell death and deactivation of pi3k/akt and mapk/erk signalling are reversed by metabolic products of the mevalonate pathway. *Oncotarget*, 7(3):2532, 2016.

[189] Maria Capra, Paolo Giovanni Nuciforo, Stefano Confalonieri, Micaela Quarto, Marco Bianchi, Manuela Nebuloni, Renzo Boldorini, Francesco Pallotti, Giuseppe Viale, Mikhail L Gishizky, et al. Frequent alterations in the expression of serine/threonine kinases in human cancers. *Cancer research*, 66(16):8147–8154, 2006.

[190] Heinz Mueller, Nathalie Flury, Serenella Eppenberger-Castori, Willy Kueng, Françoise David, and Urs Eppenberger. Potential prognostic value of mitogen-activated protein kinase activity for disease-free survival of primary breast cancer patients. *International journal of cancer*, 89(4):384–388, 2000.

[191] Robert Roskoski Jr. Cyclin-dependent protein serine/threonine kinase inhibitors as anticancer drugs. *Pharmacological Research*, 139:471–488, 2019.

[192] Laila M Poisson, Adnan Munkarah, Hala Madi, Indrani Datta, Sharon Hensley-Alford, Calvin Tebbe, Thomas Buekers, Shailendra Giri, and Ramandeep Rattan. A metabolomic approach to identifying platinum resistance in ovarian cancer. *Journal of ovarian research*, 8(1):13, 2015.

[193] Weifeng Zhang, Minghui Wan, Lunchao Ma, Xiang Liu, and Jianxing He. Protective effects of adam8 against cisplatin-mediated apoptosis in non-small-cell lung cancer. *Cell biology international*, 37(1):47–53, 2013.

[194] James O Mecham, David Rowitch, C Douglas Wallace, Peter H Stern, and Robert M Hoffman. The metabolic defect of methionine dependence occurs frequently in human tumor cell lines. *Biochemical and biophysical research communications*, 117(2):429–434, 1983.

[195] Raghu Sinha, Timothy K Cooper, Connie J Rogers, Indu Sinha, William J Turbitt, Ana Calcagnotto, Carmen E Perrone, and John P Richie Jr. Dietary methionine restriction inhibits prostatic intraepithelial neoplasia in tramp mice. *The Prostate*, 74 (16):1663–1673, 2014.

[196] Elena Strekalova, Dmitry Malin, David M Good, and Vincent L Cryns. Methionine deprivation induces a targetable vulnerability in triple-negative breast cancer cells by enhancing trail receptor-2 expression. *Clinical cancer research*, 21(12):2780–2791, 2015.

[197] Catharina Conrad, Julia Benzel, Kristina Dorzweiler, Lena Cook, Uwe Schlomann, Alexander Zarbock, Emily P Slater, Christopher Nimsky, and Jörg W Bartsch. Adam8 in invasive cancers: links to tumor progression, metastasis, and chemoresistance. *Clinical Science*, 133(1):83–99, 2019.

[198] Uwe Schlomann, Garrit Koller, Catharina Conrad, Taheera Ferdous, Panagiota Golfi, Adolfo Molejon Garcia, Sabrina Höfling, Maddy Parsons, Patricia Costa, Robin Soper, et al. Adam8 as a drug target in pancreatic cancer. *Nature communications*, 6(1):1–16, 2015.

[199] Xiangdi Yu, Jinshan Shi, Xin Wang, and Fangxiang Zhang. Propofol affects the growth and metastasis of pancreatic cancer via adam8. *Pharmacological Reports*, pages 1–9, 2019.

[200] Victor Yim, Anaïs FM Noisier, Kuo-yuan Hung, Jörg W Bartsch, Uwe Schlomann, and Margaret A Brimble. Synthesis and biological evaluation of analogues of the potent adam8 inhibitor cyclo (rlskdk) for the treatment of inflammatory diseases and cancer metastasis. *Bioorganic & medicinal chemistry*, 24(18):4032–4037, 2016.

[201] Jared R Mayers, Chen Wu, Clary B Clish, Peter Kraft, Margaret E Torrence, Brian P Fiske, Chen Yuan, Ying Bao, Mary K Townsend, Shelley S Tworoger, et al. Elevation of circulating branched-chain amino acids is an early event in human pancreatic adenocarcinoma development. *Nature medicine*, 20(10):1193–1198, 2014.

[202] Elizabeth L Lieu, Tu Nguyen, Shawn Rhyne, and Jiyeon Kim. Amino acids in cancer. *Experimental & molecular medicine*, pages 1–16, 2020.

[203] Concha Bello-Fernandez, Graham Packham, and John L Cleveland. The ornithine decarboxylase gene is a transcriptional target of c-myc. *Proceedings of the National Academy of Sciences*, 90(16):7804–7808, 1993.

[204] Eugene W Gerner and Frank L Meyskens. Polyamines and cancer: old molecules, new understanding. *Nature Reviews Cancer*, 4(10):781–792, 2004.

[205] Mariella Dono, Giannamaria Cerruti, and Simona Zupo. The cd5+ b-cell. *The international journal of biochemistry & cell biology*, 36(11):2105–2111, 2004.

[206] PDQ Adult Treatment Editorial Board. Chronic lymphocytic leukemia treatment (pdq®). In *PDQ Cancer Information Summaries [Internet]*. National Cancer Institute (US), 2020.

[207] Mohit Aggarwal, Raquel Villuendas, Gonzalo Gomez, Socorro M Rodriguez-Pinilla, Margarita Sanchez-Beato, David Alvarez, Nerea Martinez, Antonia Rodriguez, Maria E Castillo, Francisca I Camacho, et al. Tcl1a expression delineates biological and clinical variability in b-cell lymphoma. *Modern Pathology*, 22(2):206–215, 2009.

[208] Roberta Bichi, Susan A Shinton, Eric S Martin, Anatoliy Koval, George A Calin, Rossano Cesari, Giandomenico Russo, Richard R Hardy, and Carlo M Croce. Human chronic lymphocytic leukemia modeled in mouse by targeted tcl1 expression. *Proceedings of the National Academy of Sciences*, 99(10):6955–6960, 2002.

[209] Valeria Lascano, Marco Guadagnoli, Jan G Schot, Dieuwertje M Luijks, Jeroen EJ Guikema, Katherine Cameron, Michael Hahne, Steven Pals, Erik Slinger, Thomas J Kipps, et al. Chronic lymphocytic leukemia disease progression is accelerated by april-taci interaction in the tcl1 transgenic mouse model. *Blood, The Journal of the American Society of Hematology*, 122(24):3960–3963, 2013.

[210] Vicky Chatzakos, Katharina Slätis, Tatjana Djureinovic, Thomas Helleday, and

Mary C Hunt. N-acyl taurines are anti-proliferative in prostate cancer cells. *Lipids*, 47 (4):355–361, 2012.

[211] J Demeter and T Fehér. Serum dehydroepiandrosterone sulphate (dheas) and dehydroepiandrosterone (dhea) levels in hairy-cell leukaemia. *European journal of haematology*, 47(4):313–315, 1991.

[212] Alan Saven and Lawrence Piro. Newer purine analogues for the treatment of hairy-cell leukemia. *New England Journal of Medicine*, 330(10):691–697, 1994.

[213] Peter H Wiernik. Androgen therapy for acute myeloid and hairy cell leukemia. *Current treatment options in oncology*, 19(1):4, 2018.

[214] J Larry Jameson. *Harrison's principles of internal medicine.* McGraw-Hill Education,, 2018.

[215] Sam O Wanko and Carlos de Castro. Hairy cell leukemia: an elusive but treatable disease. *The oncologist*, 11(7):780–789, 2006.

[216] Monica Else, Rosa Ruchlemer, Nnenna Osuji, Ilaria Del Giudice, Estella Matutes, Anthony Woodman, Andrew Wotherspoon, John Swansbury, Claire Dearden, and Daniel Catovsky. Long remissions in hairy cell leukemia with purine analogs: a report of 219 patients with a median follow-up of 12.5 years. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 104(11):2442–2448, 2005.

[217] Enrico Tiacci, Vladimir Trifonov, Gianluca Schiavoni, Antony Holmes, Wolfgang Kern, Maria Paola Martelli, Alessandra Pucciarini, Barbara Bigerna, Roberta Pacini, Victoria A Wells, et al. Braf mutations in hairy-cell leukemia. *New England Journal of Medicine*, 364(24):2305–2315, 2011.

[218] Inderjeet Mani and I Zhang. knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, volume 126, 2003.

[219] A. Mishra, V. K. Verma, M. S. K. Reddy, A. S., P. Rai, and A. Mittal. A generative approach to zero-shot and few-shot action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 372–380, 2018.

[220] George EP Box. Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A (General)*, 143(4):383–404, 1980.

[221] Matthias W Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9(Apr):759–813, 2008.

[222] Amir F Atiya. Estimating the posterior probabilities using the k-nearest neighbor rule. *Neural computation*, 17(3):731–740, 2005.

[223] Andrew T Magis, Noa Rappaport, Matthew P Conomos, Gilbert S Omenn, Jennifer C Lovejoy, Leroy Hood, and Nathan D Price. Untargeted longitudinal analysis of a wellness cohort identifies markers of metastatic cancer years prior to diagnosis. *Scientific Reports*, 10, 2020.

[224] John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2 (8):e124, 2005.

[225] Jeffrey T Leek and Roger D Peng. Opinion: Reproducible research can still be wrong: Adopting a prevention approach. *Proceedings of the National Academy of Sciences*, 112(6):1645–1646, 2015.

[226] James A Eddy, Leroy Hood, Nathan D Price, and Donald Geman. Identifying tightly regulated and variably expressed networks by differential rank conservation (dirac). *PLoS Comput Biol*, 6(5):e1000792, 2010.

[227] John C Earls, James A Eddy, Cory C Funk, Younhee Ko, Andrew T Magis, and Nathan D Price. Aurea: an open-source software system for accurate and user-friendly identification of relative expression molecular signatures. *BMC bioinformatics*, 14(1): 1–7, 2013.

[228] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008.

[229] G Pei, L Chen, and W Zhang. Wgcna application to proteomic and metabolomic data analysis. In *Methods in enzymology*, volume 585, pages 135–158. Elsevier, 2017.

[230] Bahman Afsari, Donald German, and Elana J Fertig. Learning dysregulated pathways in cancers from differential variability analysis. *Cancer informatics*, 13:CIN–S14066, 2014.

[231] Sergey A Kornilov, Isabelle Lucas, Kathleen Jade, Chengzhen L Dai, Jennifer C Lovejoy, and Andrew T Magis. Plasma levels of soluble ace2 are associated with sex, metabolic syndrome, and its biomarkers in a large cohort, pointing to a possible mechanism for increased severity in covid-19. *medRxiv*, 2020.

[232] Michael Wainberg, Andrew T Magis, John C Earls, Jennifer C Lovejoy, Nasa Sinnott-Armstrong, Gilbert S Omenn, Leroy Hood, and Nathan D Price. Multiomic blood correlates of genetic risk identify presymptomatic disease alterations. *Proceedings of the National Academy of Sciences*, 117(35):21813–21820, 2020.

[233] Roie Levy, Andrew T Magis, John C Earls, Ohad Manor, Tomasz Wilmanski, Jennifer Lovejoy, Sean M Gibbons, Gilbert S Omenn, Leroy Hood, and Nathan D Price. Longitudinal analysis reveals transition barriers between dominant ecological states in the gut microbiome. *Proceedings of the National Academy of Sciences*, 2020.

[234] Tomasz Wilmanski, Noa Rappaport, John C Earls, Andrew T Magis, Ohad Manor, Jennifer Lovejoy, Gilbert S Omenn, Leroy Hood, Sean M Gibbons, and Nathan D Price. Blood metabolome predicts gut microbiome α-diversity in humans. *Nature biotechnology*, 37(10):1217–1228, 2019.

[235] Niha Zubair, Matthew P Conomos, Leroy Hood, Gilbert S Omenn, Nathan D Price, Bonnie J Spring, Andrew T Magis, and Jennifer C Lovejoy. Genetic predisposition

impacts clinical changes in a lifestyle coaching program. *Scientific reports*, 9(1):1–11, 2019.

[236] Ohad Manor, Niha Zubair, Matthew P Conomos, Xiaojing Xu, Jesse E Rohwer, Cynthia E Krafft, Jennifer C Lovejoy, and Andrew T Magis. A multi-omic association study of trimethylamine n-oxide. *Cell reports*, 24(4):935–946, 2018.

[237] X Xu, MP Conomos, O Manor, JE Rohwer, AT Magis, and JC Lovejoy. Habitual sleep duration and sleep duration variation are independently associated with body mass index. *International Journal of Obesity*, 42(4):794–800, 2018.

[238] Ohad Manor, Chengzhen Dai, Sergey A. Kornilov, Brett Smith, Nathan D. Price, Jennifer C. Lovejoy, Sean M. Gibbons, and Andrew T. Magis. Health and disease markers correlate with gut microbiome across thousands of people. *Nature Communications*, 2020.