

Title: Early signs of gut microbiome aging: biomarkers of inflammation, metabolism, and macromolecular damage in young adulthood

Authors:

Audrey Renson, M.P.H.

Department of Epidemiology, The Gillings School of Global Public Health, UNC-Chapel Hill, Chapel Hill, NC, USA

arenson@live.unc.edu

Telephone: 347-743-492

Kathleen Mullan Harris, Ph.D.

Department of Sociology and Carolina Population Center, UNC-Chapel Hill, Chapel Hill, NC, USA

kathie_harris@unc.edu

Jennifer B. Dowd, Ph.D.

Leverhulme Centre for Demographic Science, University of Oxford, Oxford, UK

jennifer.dowd@sociology.ox.ac.uk Lauren Gaydosh, Ph.D.

Center for Medicine, Health, and Society, Vanderbilt University, Nashville, TN, USA

lauren.m.gaydosh@vanderbilt.edu

Matthew B. McQueen, Sc.D.

Department of Integrative Physiology, University of Colorado Boulder, Boulder, CO, USA

matt.mcqueen@colorado.edu

- 1 Kenneth S. Krauter, Ph.D.
- 2 Department of Molecular, Cellular, and Developmental Biology, University of Colorado
- 3 Boulder, Boulder, CO, USA
- 4 krauter@colorado.edu
- 5 Michael Shannahan, Prof. Dr.
- 6 Institute of Sociology, University of Zurich, Zurich, Switzerland
- 7 michael.shanahan.uzh@gmail.com
- 8 Allison E. Aiello, Ph.D.
- 9 Department of Epidemiology, The Gillings School of Global Public Health, UNC-Chapel Hill,
- 10 Chapel Hill, NC, USA
- 11 aaello@email.unc.edu
- 12

Abstract:

Emerging links between gut microbiota and diseases of aging point to possible shared immune, metabolic, and cellular damage mechanisms, operating long before diseases manifest. We conducted 16S rRNA sequencing of fecal samples collected from a subsample (n=668) of Add Health Wave V, a nationally representative longitudinal study of adults aged 32-42. An overlapping subsample (n=345) included whole blood RNA-seq. We examined associations between fecal taxonomic abundances and dried blood spot-based markers of lipid and glucose homeostasis and C-reactive protein (measured in Wave IV), as well as gene expression markers of inflammation, cellular damage, immune cell composition, and transcriptomic age (measured in Wave V), using Bayesian hierarchical models adjusted for potential confounders. We additionally estimated a co-abundance network between inflammation-related genes and bacterial taxa using penalized Gaussian graphical models. Strong and consistent microbiota associations emerged for HbA1c, glucose, C-reactive protein, and principal components of genes upregulated in inflammation, DNA repair, and reactive oxygen species, with *Streptococcus infantis*, *Pseudomonas spp.*, and *Peptoniphilus* as major players for each. This pattern was largely echoed (though attenuated) for immunologic cell composition gene sets, and only *Serratia* varied meaningfully by transcriptomic age. Network co-abundance indicated relationships between *Prevotella sp.*, *Bacteroides sp.*, and *Ruminococcus sp.* and gut immune/metabolic regulatory activity, and *Ruminococcus sp.*, *Dialister*, and *Butyrivibrio crossotus* with balance between Th1 and Th2 inflammation. In conclusion, many common associations between microbiota and major physiologic aging mechanisms are evident in early-mid adulthood, and suggest avenues for early detection and prevention of accelerated aging.

1 **Key words:** gastrointestinal microbiota; cellular senescence; lipid metabolism; glucose
2 metabolism; Bayesian analysis

3

4

5

6

1 **Introduction**

2 Commensal bacteria and other microorganisms in the human gut represent a co-evolved
3 community, hypothesized to have a central role in the development and maintenance of many
4 human physiological systems across the life course, including the metabolic, neuroendocrine,
5 and immune systems.(1) Relatedly, accumulating research has linked between-person variation
6 in microbiota composition to conditions of aging, including cardiometabolic diseases,(2)
7 dementia,(3) frailty,(4) and cancers.(5) Interconnected aging-related and inflammatory
8 mechanisms proposed to underlie these conditions include accumulation of aged immune cells,
9 metabolic dysregulation, and oxidative stress-mediated macromolecular damage.(6)

10 Multiple intestinal bacterial metabolites, including toll-like receptor agonists, short chain fatty
11 acids (SCFAs), and secondary bile acids, have direct effects both on immune/inflammatory
12 processes and metabolic functions like insulin sensitivity, as well as immune-mediated metabolic
13 effects.(7) Moreover, murine experiments have illustrated that commensals are necessary for
14 regulating intestinal inflammation.(8-10) Specifically, bacteria in the murine gut aid in regulating
15 inflammatory responses by decreasing the level of responsiveness of pattern-recognition
16 receptors (PPRs) to lipopolysaccharide (LPS),(8) a membrane component of Gram-negative
17 bacteria which otherwise induces inflammation and insulin resistance.(9) Murine gut
18 commensals also participate in immune regulation by promoting differentiation of regulatory T-
19 cells (Treg) and production of IL-10,(10) a major anti-inflammatory cytokine. Conversely, age-
20 related chronic inflammation appears causally dependent on gut bacteria, as the raised circulating
21 IL-6 seen in aged mice (similarly to humans) is absent in germ free mice,(11) whereas transfer of
22 cecal contents from aged conventional mice to germ-free mice results in elevation of specific Th
23 subsets consistent with inflammaging.(12) Importantly, there is also evidence of bidirectionality

1 in relationships between immune and metabolic health and the murine gut microbiota -- for
2 example, both genetically-determined obesity(13) and immunodeficiency(14) can alter gut
3 commensal composition.

4 In humans, there is preliminary evidence to support a role of gut bacteria in immune and
5 metabolic function. A small randomized study showed that antibiotic exposure can alter response
6 to influenza vaccine -- causing decreased antibody production and raised systemic inflammation
7 -- a response echoing immunosenescence.(15) Another randomized study found that a probiotic
8 blend of *Lactobacillus*, *Streptococcus*, and *Bifidobacterium* spp. can decrease inflammatory
9 markers in overweight adults.(16) Moreover, an accumulating body of human research has
10 linked gut microbiota composition to metabolic disorders, including type 2 diabetes, obesity, and
11 associated systemic inflammation.(17)

12 Thus far, most human research on how the microbiota affect physiologic aging and age-related
13 disease has been conducted in older adults.(18, 19) However, biological aging likely begins early
14 in the life course, with immune and metabolic signals of aging emerging as early as the 30s.(20)
15 Detailing relationships between gut microbiota and markers of immune and metabolic aging
16 processes in younger populations is therefore an important step towards understanding how age-
17 related changes in disease status and microbiota interact over the life course, ultimately
18 influencing health status in older age. This study examines early adult associations between the
19 gut microbiome and hallmarks of aging and immunity, including biomarkers of inflammation,
20 immunosenescence, oxidative stress, and cardiometabolic health.

21 **Methods**

22 *Study Design, Participants, Informed consent*

1 This analysis draws on data from the National Longitudinal Study of Adolescent to Adult
2 Health (Add Health), a nationally representative longitudinal cohort of more than 20,000
3 individuals who were in grades 7-12 in 1994-95, from 80 communities and 134 middle and high
4 schools throughout the United States, with oversampling of highly educated African Americans
5 and multiple Latinx and Asian-American ethnic subgroups. Add Health has followed participants
6 through adolescence into adulthood in five waves; further design details of Add Health are
7 published.(21) In each wave, in-person interviews collected extensive information about
8 demographics, the social environment, and health, with immune/inflammatory, cardiometabolic,
9 and gene expression biomarker data added in Waves IV and V. This study draws on data from
10 Waves IV and V, when participants were ages 24-32 and 32-42, respectively.

11 Wave V embedded an ancillary study to collect tongue and stool microbiome specimens
12 on a subsample of participants. Wave V data collection was divided into four sampling groups
13 (Samples 1, 2A, 2B, and 3); the eligible pool for the microbiome subsample was nationally
14 representative and included those who accepted the microbiome pack from either the biomarker
15 home visit in Sample 1 or the in-person interview in Sample 2B (n=2,341). In sum, n=763
16 (response rate, 32.5%) returned their mail-in self-collection kits, and n=706 (92.5%) provided
17 sufficient fecal material for sequencing. Both Add Health Wave V and the microbiome ancillary
18 study were approved by the Institutional Review Board of the University of North Carolina-
19 Chapel Hill, and all participants provided written or digital informed consent.

20 *Fecal microbiome sample collection and processing*

21 Stool specimens were collected via mail-in self-collection kit. Respondents were asked to dab
22 fecal material on the “first wipe” toilet paper with dual cotton swabs (BBL® CultureSwab® EZ

Sterile Kits BD, New Jersey, USA) simultaneously before placing them in the tube containing desiccant and sealing with the supplied stopper. Specimen tubes were shipped directly to CU Boulder (Krauter lab) with transit times of <4 days and immediately stored frozen at -80°C. We used PCR to amplify the V4 region of the 16S rDNA using universal primers. Amplicons were sequenced using the Illumina MiSeq Personal Sequencer and amplicon sequence variants were determined using the Deblur pipeline and clustered into operational taxonomic units (OTUs) at 97% similarity, with closed-reference taxonomic assignment using GreenGenes. Further details of the microbial DNA processing, storage, sequencing, and data management are given in the eMethods.

Measurement of Biomarkers and medications

As part of Wave IV, dried blood spots were assayed for glucose (mg/dl), low-density lipoprotein (LDL), high-density lipoprotein (HDL), and high-sensitivity c-reactive protein (CRP), and HbA1c (%). Self-reported antidiabetic, antihyperlipidemic, and anti-inflammatory (COX-2 inhibitors, NSAIDs, inhaled corticosteroids, corticotropins/glucocorticoids, antirheumatic/antipsoriatics, or immunosuppressives) medications taken in the past four weeks were recorded at the same visit. Further details of the dried blood spot collection and assays are given in the eMethods.

Measurement of Gene Expression

A separate random subsample of 5,000 participants in Wave V was selected for whole-blood RNA-seq transcriptome analysis (currently, n=1,132 samples have been fully processed.) Details about the venous blood sampling, RNA isolation and sequencing are given in the eMethods. We created gene expression variables based on nine Molecular Signature Database gene sets: the

Hallmark sets for inflammation, reactive oxygen species (ROS), and DNA repair;(22) as well as immunological signatures for naive vs. effector CD8+; CD4+ vs. CD8+; Th1 vs. naive CD4+; Treg vs, naive CD4+; naive vs. KLRG1 high CD8+; and CD57+ vs, CD57- NK.(23) To construct these variables, we used the first principal component of the centered log-ratio transformed expression values of all genes in each expression set; a similar approach has been used previously.(24) Additionally, we computed a measure of transcriptomic age including 1,497 genes developed previously based on a large meta analysis.(25) This measure is a model-based prediction of the expected age of a person given their gene expression, which we computed using the linear regression coefficients provided in the supplementary data of the article by Peters et al. (25) The final variable used in analyses was the difference between their transcriptomic age and chronological age (hereafter, Δ age).

Confounders

All models adjust for age, sex, and educational attainment (coded categorically as high school degree or less, some college or vocational/technical school, bachelor's degree, or master's/doctoral/professional degree); models for HDL and LDL additionally adjust for lipid medications in Wave IV; models for glucose and HbA1c additionally adjust for glucose medications in Wave IV; models for log-CRP additionally adjust for lipid and inflammation medications in Wave IV; models for all gene expression markers additionally adjust for medications for asthma, lipids, glucose, and psychiatric conditions in Wave V.

Statistical Analysis

All statistical analyses were conducted in R v. 3.5.1,(26) and code used to perform all analyses is available in a GitHub repository (27). This analysis focused on estimating associations between

1 taxonomic abundances and biomarkers of immune and cardiometabolic aging, which were
2 HbA1c, glucose, HDL, LDL, log-CRP, and the selected gene set principal component vectors,
3 described above. Each biomarker was treated in a separate model. We implemented a Bayesian
4 hierarchical log-linear modelling approach, implemented in the R packages ‘DESeq2’ v
5 1.22.2(28) and ‘ashr’ v 2.2-32.(29) The model for the i^{th} individual and j^{th} taxon can be written
6 as:

$$y_{ij} \sim \text{NegativeBinomial}(\text{mean} = \mu_{ij}, \text{variance} = \mu_{ij} + a_j \mu_{ij}^2)$$

$$\log_2(\mu_{ij}) = \log_2(\text{offset}) + \alpha_j + \beta_j X_{ij} + \sum_{k=1}^{p-2} \gamma_{jk} \mathbf{Z}_{ijk}$$

$$\beta_j \sim \sum_{l=1}^{\infty} \pi_l \text{Normal}(0, \sigma_l^2)$$

$$a_j \sim \text{LogNormal}(f(\alpha_j), \sigma_j^2)$$

7 The response variable y_{ij} is the observed count of bacterial taxon j for person i , regressed on X ,
8 the biomarker of interest, and potential confounders, \mathbf{Z} . Because we standardized all biomarkers
9 to a variance of one, the biomarker regressor coefficient β_j represents the \log_2 ratio change of a
10 given taxon’s relative abundance that would be expected for a standard deviation change in
11 biomarker value (hereafter, \log_2 -fold change or logFC). Because the number of reads varies
12 between samples, we use an offset in the model to accomplish normalization. As offset, instead
13 of DESeq2’s default, we use a measure called the geometric mean of pairwise ratios,(30) a
14 function of the number of reads that accounts for the compositional and zero-inflated nature of
15 microbiome sequencing data. Estimation of the dispersion parameter a_j is improved by sharing
16 information across samples; that is, each estimate is shrunk towards $f(\alpha_j)$, a smooth function

of the mean normalized count of taxon j . Details of the estimation procedure are given in references (28) and (29).

This is a Bayesian hierarchical model in the sense that a prior probability distribution, estimated from the data, is placed on each biomarker's coefficient. Specifically, the prior on the vector β is the best-fitting zero-mean Gaussian mixture estimated from the maximum likelihood estimates, $\hat{\beta}$.(29) Importantly, this prior accomplishes two goals: (a) it encodes our belief that few taxa are likely to have strong associations with biomarkers, and (b) it strongly regularizes noisy estimates to prevent undue confidence in discoveries. Figure 1 shows samples from priors for β for each biomarker, illustrating that these are strongly regularizing priors: the range of coefficients estimated using maximum likelihood is much wider (approximately -5 to 5) across taxonomic levels and biomarkers, with priors contributing on average 7.5 times as much information as the data.

We estimated these models for all taxa present in the dataset after collapsing separately at the species, genus, family, and phylum levels, and after filtering out low-prevalence taxa; in sum, we performed 3,645 statistical tests. We did not perform multiplicity correction. Instead, we rely on the *s-value*, the posterior probability that the sign of a coefficient is incorrect, as a continuous descriptive measure corresponding to relative confidence in a discovery. This value is estimated directly from the posterior distribution and represents the proportion of the probability mass lying on the same side of zero as the posterior mean. Because of the strong shrinkage, this value does not require multiplicity correction and is more conservative than the Benjamini–Hochberg false discovery rate in our data (not shown). In some cases, we use cutoff thresholds of the *s-value* and/or logFC, which are selected arbitrarily to select a manageable number of findings for

discussion and/or highlighting in figures. As such, these cutoffs should be understood as heuristic rather than confirmatory.

Bacteria / Blood Gene Expression Network

To estimate a co-occurrence network among bacterial taxa levels and inflammation-related gene expression levels, we used Gaussian graphical modelling (GGM), implemented in the R package ‘huge’ v 1.3.0.(31) GGMs have been used extensively to infer interaction networks in multi-omics data,(32) and in microbiome data alone,(33), but to our knowledge this is the first study to apply GGMs to combined 16S and RNA-Seq data. GGMs estimate pairwise covariances between nodes (genes and taxa), conditional on all other nodes, using regularization to shrink noisy estimates towards zero. To be as granular as possible, we considered only raw OTUs, not higher levels of taxonomic classification. Given the small sample size for this analysis (n=345), we selected a limited set of genes broadly related to inflammation pathways by searching Gene Ontology for acute or chronic inflammation, response to antigenic stimuli or wounding, leukocyte activation & migration, neuroinflammation, and mediators of cytokine, histamine, serotonin, nitric oxide, and respiratory burst pathways.(34) The search revealed 138 unique genes, of which 122 were expressed in any of our samples; these 122 were used to estimate the network (see eTable 1 (27) for the full list). Some covariate values were missing for between 3 and 12% of participants, so we first performed single imputation using fully conditional specification for age, sex, all medication variables, educational attainment, and all dried blood spot values used in the main analysis, performed using the R package ‘mice’ v. 3.1.0.(35) To further reduce the risk of chance discoveries, we randomly split the data into a discovery set of n=177 (used for initial estimation) and a replication set of n=168.

Before splitting, we transformed the data according to the centered log-ratio (log of the count, divided by the geometric mean of counts in a sample), performed separately for OTUs and genes. We then filtered out OTUs and genes present in fewer than 10 individuals in the full data set. We then applied the nonparanormal transformation (a non-parametric transformation which relaxes the multivariate Gaussian assumption in GGMs(36)) to the entire data matrix. We adjusted for potential confounders by taking residuals from a linear regression model for each nonparanormal, centered log ratio-transformed OTU and gene regressed on age, sex, educational attainment, medications for lipids, glucose, and inflammation in Wave IV, and medications for lipids, glucose, asthma, and mental health in Wave V. These residuals were the ultimate input into the model.

Model selection in GGMs requires selection of a penalty parameter denoted λ , which controls the level of sparsity in the resulting network. We performed model selection in the discovery set as follows: to select the GGM penalty λ , we first selected the λ values for which the degree distribution fit to a scale-free topology had R^2 greater than 0.6 (using the R package ‘WCGNA’ v. 1.68 (37)). This initial step restricts to the set of biologically plausible networks. Then we fit GGMs to a grid of λ values in this range using the Meinshausen-Buhlmann method, and selected the model based on the STARS criterion, which maximizes the edge stability under 100 bootstrap resamples. After model selection on the discovery set, we fit a GGM to the replication data using the same value of λ . We considered the final network to be only the replicated edges.

Results

Descriptive statistics

The complete microbiome subsample of Add Health Wave V includes fecal samples from $n=706$ participants, and $n=668$ after removing 38 samples with less than 1,000 reads. Of these, $n=345$ also overlapped with the peripheral blood gene expression subsample. Because some participants were missing covariate data, final analytic sample sizes range from $n=565$ - 632 for dried blood spot and $n=342$ for gene expression measures. Although Add Health is nationally representative, the microbiome subsample has a higher proportion female (65% vs. 50%), non-Hispanic White (69% vs 53%), and college graduates (46 vs 31%) compared to the total sample. Restricting attention to the microbiome subsample, a substantial number (30%) had $HbA1c > 5.6\%$, indicating higher than average diabetes risk (mean=5.6, standard deviation=0.8), and a large percentage (44%) had elevated (>3 mg/L) CRP (geometric mean=2.1, standard deviation=4.2). Because $\Delta\text{age}(25)$ is scaled to the age distribution in the sample, the mean of Δage was 0; 95% of the sample fell in the range (-3.8, +5.7).

Correlations between biomarkers and the general marker of Shannon diversity (adjusted for confounders) were small (all with absolute value < 0.05 or with large standard errors), and we did not observe large shifts in compositionally-robust distances for any of the aging/disease biomarkers ($R^2 < 0.35\%$, distance-based redundancy analysis on Aitchison distances, also adjusted for confounders).

Differential Abundance

After removing samples with <1000 reads, samples had mean \pm SD 11,543 \pm 13,216 reads, with an average of 173 unique taxa detected per sample. After filtering taxa to keep only those with a count greater than 3 in $n=25$ or more samples, we detected a total of 11 phyla, 59 families, 95 genera, and 72 species. We applied this same filter before calculating all the results that follow.

Figure 2 shows a manhattan plot of the negative log₁₀-transformed s-value (posterior probability of sign error), for reads collapsed at the species, genus, family, and phylum level, arranged according to the phylogenetic tree to illustrate the phylogenetic distribution of estimates. Focusing on the s-value (vs. the log₂-fold change estimate) highlights the associations for which we have the highest level of confidence in the sign (direction of association) - this does not suggest these are large associations, only that they are estimated with precision. Overall, we have high confidence in the sign for many taxa - for example, there are 47 taxon-biomarker associations for which we are more than 99% certain of the sign (ie. direction of association), given our modeling assumptions hold. The majority of confidently estimated associations are for taxa that belong to the phylum Firmicutes (followed by Proteobacteria) and the majority are at the genus level (e.g. 55% of those with s-values less than 5% are genus-level associations, but the same pattern holds regardless of the threshold). Most of the highly confident associations are for HbA1c, glucose, LDL, HDL, and CRP (all measured in Wave IV), although a few are for the ROS and CD57+ vs. - NK gene sets (measured in Wave V). Notably, confidence in the associations with most Wave 5 immunological gene expression signatures is much lower (Th1 vs. naive CD4+, Treg vs. naive CD4+, and CD4+ vs. CD8+ each had all s-values greater than 23%), which may suggest these are noisy measures in this dataset, or may reflect the sample size available for gene expression.

Figure 3 shows volcano plots, where strength of association (log₂ fold change [logFC]) is elucidated. For figure readability, attention is restricted to associations with relatively large magnitude and precision; defined as those with an estimated logFC with an absolute value greater than 0.8 and s-value less than 0.01. Ie, these are taxa for which we are 99% sure of the direction (given assumptions), and where a standard deviation change in the blood biomarker is

1 estimated to be associated with at least $2^{0.8}=1.75$ times greater or lesser relative abundance of the
2 taxon.

3 An overall pattern in figure 3 is that all strong associations between gut bacteria and DNA repair-
4 and ROS-related gene expression signatures in Wave V are negative, suggesting many gut
5 bacteria may be specifically associated with reduced cellular aging. In contrast, there are many
6 gut bacteria positively and negatively associated with Wave IV HbA1c, LDL, glucose, and CRP,
7 suggesting some harmful and some beneficial associations of certain common gut commensals
8 with metabolic health. Another pattern evident in figure 3 is that nearly every clade that
9 demonstrates a strong association with more than one biomarker points to potential contradictory
10 effects. For example, *Pseudomonas* and Pseudomonadaceae correlate with healthier CRP and
11 CD57+ profiles, but worse lipid and glucose measures. A similar pattern is visible for *S. infantis*
12 and *Peptoniphilus*.

13 Here, we report associations as logFC estimate (posterior mode) and a 90% highest posterior
14 density interval in brackets. Several short-chain fatty acid-producing clades in Firmicutes show
15 substantial associations with several biomarkers. For example, *Streptococcus infantis* is a
16 prevalent (35%), low abundance (<1% average) acetate producer that is associated with higher
17 HbA1c (1.6 [1.2, 2.0]) and CRP (0.9 [0.6, 1.3]), but lower LDL (-1.1 [-1.5, -0.8]) and ROS (-1.1
18 [-1.4, -0.7]). *Peptoniphilus* is also a common (54%) low abundance (<1% average) butyrate-
19 producing Firmicutes member associated with lower glucose (-0.9 [-1.2, -0.6]) and higher CRP
20 (0.9 [0.6, 1.1]). Additionally, several Proteobacteria clades appear. *Pseudomonas*, its family
21 Pseudomonadaceae, or the species *P. veronii* have associations with healthier profiles of CRP
22 (*Pseudomonas*: -1.5 [-2.2, -0.9]; Pseudomonadaceae: -1.2 [-2.0, -0.6]) and CD57+ vs. - NK (*P.*
23 *veronii*: -2.6 [-3.5, -1.8]) , but worse profiles of naive vs. KLRG1 high CD8+ (*P. veronii*: -1.9 [-

3.0, -0.9]), hallmark inflammation gene expression (*P. veronii*: 1.6 [0.6, 2.6]), glucose (*Pseudomonas*: 2.1 [1.4, 2.7]), and HDL (Pseudomonadaceae: -1.4 [-2.2, -0.8], *Pseudomonas*: -1.2 [-1.8, 0.0]). *Serratia* and its family Enterobacteriaceae appear associated with higher LDL (*Serratia*: 1.0 [0.0, 2.4]; Enterobacteriaceae: 0.9 [0.4, 1.4]) and glucose (*Serratia*: 2.1 [0.7, 3.5]), but lower Δ age (-3.5 [-5.0, -2.2]). *Pseudomonas* and *Serratia* are both relatively rare (<14%) members of Proteobacteria, which is indeed the only confidently estimated phylum-level association (not labeled in figure 3) -- this phylum correlates with healthier profiles of HbA1c (-0.7 [-0.9, -0.5]), LDL (-0.5 [-0.7, -0.4]) and HDL (0.5 [0.3, 0.7]), and is also weakly related to lower Δ age (-0.2 [-0.3, 0.0]). Notably absent from this list are well-studied butyrate-producing anaerobes in Clostridium clusters IV and XIVa making up significant portions of individuals' microbiomes, such as *Faecalibacterium*, *Ruminococcus*, *Roseburia*, *Blautia*, and *Coprococcus*, all of which have logFC estimates close to zero and narrow posterior intervals for all biomarkers (eTable 2 (27)). Also notably, many immunological gene expression markers (CD4+ vs. CD8+, Th1 vs. naive CD4+, and Treg vs. naive CD4) did not demonstrate associations with sufficient confidence or size to be highlighted in either figure, although their microbiome signatures frequently correlated strongly with other markers in expected directions (eFigure 1). All taxon-biomarker associations are listed in eTable 2.(27)

Bacteria / gene expression network

The network fit to the discovery data (n=177) revealed a sparse (edge density=2.2%, average degree=7.8) network with 1364 total edges, 149 of which were between OTUs and genes. The final network (discovery \cap replication) had 575 edges and 53 OTU-gene edges, for an overall replication rate of 42% and an OTU-gene edge replication rate of 36%, with greater sparsity as a result (edge density=1.7%, average degree=4.4), and strong modularity (68%, Louvain's

community detection). Of 10 clusters identified with more than two nodes, five clusters (A1-A5) contained both OTUs and genes, four (B1-B4) contained only OTUs, and two (C1-C2) contained only genes. Note that these clusters are not truly separate; clusters A1-A5 each share between 1-13 edges with one another (A1 and A2 are especially well connected), suggesting many biological processes in common. Because primary interest is in gene-bacteria interconnectedness, we focus attention on clusters A1-A5 containing both OTUs and genes, which are illustrated in Figure 4 (the entire network is shown in eFigure 2, and all members are listed in eTables 3 and 4 (27)).

Cluster A1 highlights several important regulators of intestinal metabolism and type 1 inflammatory response. These include NLRP6 and PPARG, both of which are involved in regulating gut homeostasis, and which directly correlate with *Desulfovibrio sp.* and *Bilophila sp.*, respectively. NPPA and NPY5R both suppress inflammatory cytokines and are involved in cardiovascular regulation; these appear positively related to *Sutterella sp.* and *Bifidobacterium sp.* This cluster also includes two inducers of the acute phase response, C3 and LBP, the latter of which is the pro-inflammatory receptor of the gram-negative surface glycoprotein lipopolysaccharide, which connects directly to *Butyricimonas sp.*

Similarly, cluster A3 includes GPR17, a negative regulator of Th2 response, and SELE, a neutrophil adhesion mediator activated by IL-1 and TNF- α . These two genes are positively correlated, as expected, and also negatively correlated with a large group of bacterial taxa, predominantly members of *Bacteroides*, *Parabacteroides*, *Prevotella*, and *Ruminococcus*.

Relatedly, cluster A4 highlights several Clostridiales sp., Ruminococcaceae sp., *Ruminococcus sp.*, *Dialister*, and *Butyrivibrio crossotus*, and a member of Order ML615J-28. These taxa potentially represent a cluster of major immune activity -- the group correlates either directly or

indirectly with downregulation of IL12B and upregulation of IL-4, major upstream inducers of Th1 and Th2 cell differentiation, respectively. Cluster A4 and A5 contain the orthologous C2CD4A and C2CD4B which are induced by IL-1b and likely involved in type 2 diabetes; both clusters also contain members of Order ML615J-28, suggesting its common involvement in this pathway.

In contrast, cluster A2 includes OSMR, SLC7A2, and TNFSF11, all of which are upregulated in the type 2 inflammation associated with IgE-mediated hypersensitivities. OSMR and TNFSF11 in particular have been implicated in inflammatory bowel disease.(38) *Prevotella copri* appears to be a central player in this group, with *Megasphaera sp.*, *Desulfovibrio sp.*, and Ruminococcaceae sp. closely involved. The large number of negative covariance estimates in cluster A2 suggest that several of these taxa are in competition and may be involved in downregulating IgE-mediated proinflammatory activity. In short, clusters A3-A5 highlight potential microbial mechanisms of balance between type 1 and type 2 inflammation response, whereas cluster A2 highlights taxa potentially relevant to allergic and atopic activation.

Discussion

To our knowledge, this study is the first to explore relations between biomarkers of physiologic aging and gut microbiota composition in a population-based sample of young or middle-aged adults. We found that multiple taxa show strong and consistent associations with multiple biomarkers of interconnected aging mechanisms, suggesting that microbiota are linked to aging earlier in the life course. Alpha- or beta-diversity-based shifts in composition according to biomarker levels were not evident, which suggests that shifts in the gut microbiota influencing

(or influenced by) physiologic aging at this life stage are on the level of specific taxa rather than broad ecosystem-wide changes. Generally, the taxa showing the strongest relationships with metabolic, inflammation, and macromolecular damage markers (*Streptococcus infantis*, Pseudomonadaceae, *Veillonella*, *Serratia*) were not core members classically associated with central ecosystem services (such as Clostridium clusters VI and XVIa), but rarer groups that have been less studied in the context of intestinal ecosystems. In contrast, bacteria highlighted in our estimated bacteria-gene expression network *do* include members of these core groups, many of which have previously shown markedly lower abundance in older vs. younger adults, including *Prevotella*, *Bacteroides*, Ruminococcaceae, Lachnospiraceae,(39) as well as *Bifidobacterium*, which has been associated with extreme longevity.(40)

One contribution of our study is simultaneous examination of biomarkers related to interconnected aging mechanisms of metabolism, macromolecular damage, inflammation, and immunosenescence, as well an overall transcriptomic marker of aging. The general correspondence between the most strongly associated taxa for nearly all these systems, and between gene expression and more conventional markers, allows us to highlight taxa potentially involved in cellular aging processes, and thus age-related disease, with relative confidence. Specifically, *Veillonella* and *Streptococcus infantis* are each associated with markers of inflammation, macromolecular damage, and metabolism; *Peptoniphilus* with the latter two; and *Serratia* with Δ age and metabolic markers. Of note, some of these associations contradict one another: for example, *Streptococcus infantis* appears protective for LDL and hallmark inflammation gene expression, but potentially harmful for CRP and HbA1c; many taxa follow a similar pattern. Similarly, our finding of the phylum Proteobacteria associated with salutary parameters of LDL, HDL, and HbA1c is inconsistent with the hypothesis that this pathobiont-

1 enriched clade upregulates inflammaging,(12, 18) possibly reflecting functional diversity within
2 this phylum or different effects at varying life stages. Additionally, network results at the level of
3 specific genes and specific OTUs, intended to be as granular as possible, highlight strikingly
4 different taxa than the analysis collapsed at clade and pathway levels. This may again reflect
5 functional diversity of such core clades as *Ruminococcus*, *Prevotella*, and Lachnospiraceae,
6 which, when collapsed, may attenuate the respective associations of their specific, unclassified
7 OTUs. This may highlight a limitation of 16S taxonomic assignment, which can be addressed by
8 more granular methods such as shotgun sequencing.

9 We were unsurprised to find many taxonomic associations with conventional markers of lipid
10 and glucose homeostasis, as a growing body of literature suggests there are causal links
11 there.(17, 41) However, immune and inflammation pathways have been consistently observed in
12 animal models but less examined in humans, particularly in younger adults. Strong associations
13 with CRP and inflammation gene expression that echo those with metabolic markers support the
14 hypothesis that inflammation is a central mechanism of the effect of microbiota on
15 metabolism.(6) Moreover, a few taxa (primarily *Pseudomonas veronii*, whose role in the human
16 gut is largely uncharacterized, and *Succinivibrio spp.*, producers of pro-inflammatory succinate)
17 appear associated with genes upregulated in lymphoid cell markers of immunosenescence,
18 CD57+ NK and KLRG1, markers which are known to be associated with CMV activation.(42)
19 Of note, among markers of T-cell senescence examined, these two markers appeared most
20 strongly associated with the bacterial taxa assessed here. These findings are salient because
21 CD57+NK and KLRG1 are key markers of immunosenescence and associated with terminal
22 differentiation, suggesting that certain microbiota may drive cellular mechanisms of aging even
23 in middle adulthood. As animal research has shown effects of commensal metabolites on T cell

populations (43), future studies directly measuring immune cell populations are warranted to continue to examine this pathway in humans.

Strengths and Limitations

Although this is an observational study, our underlying scientific question is about causal effects of gut bacteria on physiological aging. While never perfect in this regard, observational studies can contribute information to causal questions, when certain assumptions are met.(44) One assumption here, causal consistency (a.k.a. well-defined intervention), is challenged in 16S taxonomy because most taxa are only confidently classified at the genus level, while bacteria are often extremely functionally diverse even below the species level. This assumption can be relaxed by the addition of shotgun metagenomics, metatranscriptomics, and/or metabolomics, which come closer to relevant functional specificity. Moreover, a major issue that has been largely ignored in microbiome research(45) has been the compositional nature of sequencing data, for which common statistical procedures produce associations which, at best, correspond to effects which are likely not of interest.(46) Our work improves on this by using robust denominators developed under Aitchison's framework of compositional data analysis.(30) To help meet the assumption of conditional exchangeability (ie, no unmeasured confounding), we adjusted for a number of possible confounders, but many likely remain. For example, we did not have good measures for diet, which likely influences bacterial levels as well as immune and metabolic parameters through pathways other than bacteria. We may have partially addressed this by adjusting for educational attainment, which is strongly correlated with diet in the US. Moreover, identifying causal effects of specific bacteria (e.g., for use as potential probiotics) is unrealistic in observational data because gut bacteria exhibit strong pairwise correlations whose

causal structure is typically unknown, and thus many confounders (ie, other bacteria) cannot be controlled.

Another threat to conditional exchangeability is lack of clear temporal ordering in our data: metabolic panel markers were measured roughly 10 years *before* the microbiome, and gene expression was measured nearly concurrently; therefore we cannot ascertain the direction of association for these measures. Despite this, we have some confidence that our data capture effects of the microbiota for two reasons. (a) The powerful action of commensals in training the immune system in the first several years of life is thought to lead to a durable and homeostatic community that largely persists throughout adulthood,(47) except for environmental shocks such as major dietary changes(48) -- indeed, longitudinal evidence suggests that gut microbial taxonomic composition is relatively stable within persons in adulthood.(49) (b) While findings from microbiome studies have been difficult to compare because of differences in normalization and analytic approaches, our findings are consistent with previous results indicating that butyrate and acetate producers appear to play major roles in metabolism and inflammation.(19, 41). Despite this, there are known feedback loops between gut bacteria and metabolic and immune processes;(13, 14) thus, our measures at best capture both (a) effects of gut bacteria, and (b) alteration of the community in response to human bodily processes, including immunity.

Lastly, several of our measurements have notable limitations. Although our Gene Ontology search terms likely captured a large portion of inflammation-related genes, given our limited sample size, we were unable to capture the complete construct of immunity in our network analysis, resulting in some simplifications and selectivity of immune markers we hypothesized to be most closely related to aging. Additionally, glucose, LDL, and HDL are reported in Add Health as deciles, and transcriptomic age is standardized to the sample age distribution; therefore

these measurements contain only relative (rather than absolute) information about any given individual. Dried-blood spot-based measures are highly reliable relative to whole-blood analysis for HbA1c and CRP, although perhaps less so for lipid homeostasis.(50) Finally, high-throughput 16S rRNA sequencing and RNA-seq are measurements known to contain error, which may be compounded by large variation in library size typical of these data (see eFigure 3); we make the critical assumption that this error is random and is thus less likely to produce systematic bias in our effect estimates.

Conclusion

Our study adds to the evidence that gut bacteria may influence physiologic mechanisms underlying a wide range of age-related diseases and biological phenotypes. Of note, we present some of the first data to characterize these associations far earlier in the life course than the majority of age-related disease onset and age-driven health declines. Our results therefore suggest that strategies aimed at repairing or improving gut microbiome community dynamics at or before this life stage may be useful avenues to explore for prevention of premature aging, whether through e.g. probiotic supplementation, targeted diet changes, or vaccines. Moreover, our results point to a few diverse groups of taxa potentially warranting further investigation for the physiologic aging potential of the microbiome in early to mid adulthood. Replication of these findings by future studies with richer confounder control, larger sample sizes, longitudinal follow-up, and direct immunologic measurements would build further support that gut microbiota could aid in the early detection and prevention of accelerated aging and age-related disease.

Funding

Research reported in this paper was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) of the National Institutes of Health grant numbers P01HD031921, R01HD060726, R01HD087061; National Institute of Aging grant number R01AG042794-01A1, and the National Institute of Minority Health and Health Disparities grant number R01MD013349. We are grateful to the Carolina Population Center for training support (grant number T32HD091058) and for general support (grant number P2C HD00924).

Acknowledgements

The authors would like to acknowledge Grace Noppert, PhD, for her critical feedback and expertise in developing this manuscript. This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations.

Reference List

Figure Captions

Figure 1. Adaptive prior distributions for β for each phenotype, for models estimated at the genus level, illustrating zero-centering and strong regularization such that very low prior probability is placed on absolute values greater than 0.01 for a standard deviation change in continuous variables, and 0.03 for binary variables. Although only genus priors are shown, priors at other taxonomic levels are nearly identical.

Figure 2. Phylogenetic distribution of posterior sign confidence for log-linear model coefficients for all phenotypes and at all taxonomic levels. A large $-\log_{10}$ S-value (y-axis) indicates low posterior probability of incorrectly declaring an association to be positive or negative. A non-random distribution across the phylogenetic tree is evident, with most high-confidence estimates in Firmicutes and Proteobacteria. To maximize figure readability, we labeled taxa with $-\log_{10}(\text{S-value})$ greater than 3 (less than 1 in 1,000 chance of sign error). The direction of association is indicated by + or - in the label.

Figure 3. Volcano plots, showing beta estimates (posterior mode of log2-Fold Change) on the x axis and posterior sign confidence (-log10(svalue)) on the y axis, for (A) immune and gene expression markers, and (B) metabolic markers. We required an absolute value of beta greater than 1 and a sign error probability less than 0.01 in order to be labeled. These cutoffs are selected to maximize figure readability. *HDL has been inverse coded so that higher values indicate lower HDL, to ease interpretation of taxa as related to health vs. disease.

Figure 4. Estimated conditional correlation network between OTUs and a curated set of inflammation-related genes expressed in whole blood, estimated using a penalized Gaussian graphical model and adjusting for age, sex, and medications by residualizing. Colors correspond to clusters flagged by Levain's community detection algorithm. Cluster identifiers indicated in large text are also referenced in eTables 3 and 4 (27). Size of nodes corresponds to degree - larger nodes have more connections than smaller nodes..Square nodes are genes, circular nodes are bacterial taxa.

1. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. *Nat Med.* 2018;**24**:392-400. 10.1038/nm.4517.
2. Heianza Y, Ma W, Manson JE, Rexrode KM, Qi L. Gut Microbiota Metabolites and Risk of Major Adverse Cardiovascular Disease Events and Death: A Systematic Review and Meta-Analysis of Prospective Studies. *J Am Heart Assoc.* 2017;**6**. 10.1161/JAHA.116.004947.
3. Ticinesi A, Tana C, Nouvenne A, Prati B, Lauretani F, Meschi T. Gut microbiota, cognitive frailty and dementia in older individuals: a systematic review. *Clin Interv Aging.* 2018;**13**:1497-1511. 10.2147/CIA.S139163.
4. Jackson MA, Jeffery IB, Beaumont M, Bell JT, Clark AG, Ley RE, *et al.* Signatures of early frailty in the gut microbiota. *Genome Med.* 2016;**8**:8. 10.1186/s13073-016-0262-7.
5. Gopalakrishnan V, Helmink BA, Spencer CN, Reuben A, Wargo JA. The Influence of the Gut Microbiome on Cancer, Immunity, and Cancer Immunotherapy. *Cancer Cell.* 2018;**33**:570-580. 10.1016/j.ccell.2018.03.015.
6. Franceschi C, Garagnani P, Parini P, Giuliani C, Santoro A. Inflammaging: a new immune-metabolic viewpoint for age-related diseases. *Nat Rev Endocrinol.* 2018;**14**:576-590. 10.1038/s41574-018-0059-4.
7. Zmora N, Bashardes S, Levy M, Elinav E. The Role of the Immune System in Metabolic Health and Disease. *Cell Metab.* 2017;**25**:506-521. 10.1016/j.cmet.2017.02.006.
8. Alenghat T, Osborne LC, Saenz SA, Kobuley D, Ziegler CGK, Mullican SE, *et al.* Histone deacetylase 3 coordinates commensal-bacteria-dependent intestinal homeostasis. *Nature.* 2013;**504**:153-157. 10.1038/nature12687.

- 1 9. Manco M, Putignani L, Bottazzo GF. Gut microbiota, lipopolysaccharides, and innate
2 immunity in the pathogenesis of obesity and cardiovascular risk. *Endocr Rev.* 2010;**31**:817-844.
3 10.1210/er.2009-0030.
- 4 10. Atarashi K, Tanoue T, Oshima K, Suda W, Nagano Y, Nishikawa H, *et al.* Treg induction
5 by a rationally selected mixture of Clostridia strains from the human microbiota. *Nature.*
6 2013;**500**:232-236. 10.1038/nature12331.
- 7 11. Thevaranjan N, Puchta A, Schulz C, Naidoo A, Szamosi JC, Verschoor CP, *et al.* Age-
8 Associated Microbial Dysbiosis Promotes Intestinal Permeability, Systemic Inflammation, and
9 Macrophage Dysfunction. *Cell Host Microbe.* 2017;**21**:455-466.e454.
10 10.1016/j.chom.2017.03.002.
- 11 12. Fransen F, van Beek AA, Borghuis T, Aidy SE, Hugenholtz F, van der Gaast-de Jongh C,
12 *et al.* Aged Gut Microbiota Contributes to Systemic Inflammation after Transfer to Germ-Free
13 Mice. *Front Immunol.* 2017;**8**:1385. 10.3389/fimmu.2017.01385.
- 14 13. Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. Obesity alters
15 gut microbial ecology. *Proc Natl Acad Sci U S A.* 2005;**102**:11070-11075.
16 10.1073/pnas.0504978102.
- 17 14. Vijay-Kumar M, Aitken JD, Carvalho FA, Cullender TC, Mwangi S, Srinivasan S, *et al.*
18 Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. *Science.*
19 2010;**328**:228-231. 10.1126/science.1179721.
- 20 15. Hagan T, Cortese M, Rouphael N, Boudreau C, Linde C, Maddur MS, *et al.* Antibiotics-
21 Driven Gut Microbiome Perturbation Alters Immunity to Vaccines in Humans. *Cell.*
22 2019;**178**:1313-1328.e1313. 10.1016/j.cell.2019.08.010.
- 23 16. Rajkumar H, Mahmood N, Kumar M, Varikuti SR, Challa HR, Myakala SP. Effect of
24 probiotic (VSL#3) and omega-3 on lipid profile, insulin sensitivity, inflammatory markers, and
25 gut colonization in overweight adults: a randomized, controlled trial. *Mediators Inflamm.*
26 2014;**2014**:348959. 10.1155/2014/348959.
- 27 17. Meijnikman AS, Gerdes VE, Nieuwdorp M, Herrema H. Evaluating Causality of Gut
28 Microbiota in Obesity and Diabetes in Humans. *Endocr Rev.* 2018;**39**:133-153. 10.1210/er.2017-
29 00192.
- 30 18. Biagi E, Nylund L, Candela M, Ostan R, Bucci L, Pini E, *et al.* Through ageing, and
31 beyond: gut microbiota and inflammatory status in seniors and centenarians. *PLoS One.*
32 2010;**5**:e10667. 10.1371/journal.pone.0010667.
- 33 19. Claesson MJ, Jeffery IB, Conde S, Power SE, O'Connor EM, Cusack S, *et al.* Gut
34 microbiota composition correlates with diet and health in the elderly. *Nature.* 2012;**488**:178-184.
35 10.1038/nature11319.

- 1 20. Parker D, Sloane R, Pieper CF, Hall KS, Kraus VB, Kraus WE, *et al.* Age-Related
2 Adverse Inflammatory and Metabolic Changes Begin Early in Adulthood. *J Gerontol A Biol Sci*
3 *Med Sci.* 2019;**74**:283-289. 10.1093/gerona/gly121.
- 4 21. Harris KM. Design features of add health. Chapel Hill, NC: University of North Carolina
5 at Chapel Hill. 2011.
6 <https://www.cpc.unc.edu/projects/addhealth/documentation/guides/DesignPaperWIIV.pdf>.
- 7 22. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The
8 Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* 2015;**1**:417-
9 425. 10.1016/j.cels.2015.12.004.
- 10 23. Godec J, Tan Y, Liberzon A, Tamayo P, Bhattacharya S, Butte AJ, *et al.* Compendium of
11 Immune Signatures Identifies Conserved and Species-Specific Biology in Response to
12 Inflammation. *Immunity.* 2016;**44**:194-206. 10.1016/j.immuni.2015.12.006.
- 13 24. Burt JB, Demirtaş M, Eckner WJ, Navejar NM, Ji JL, Martin WJ, *et al.* Hierarchy of
14 transcriptomic specialization across human cortex captured by structural neuroimaging
15 topography. *Nat Neurosci.* 2018;**21**:1251-1259. 10.1038/s41593-018-0195-0.
- 16 25. Peters MJ, Joehanes R, Pilling LC, Schurmann C, Conneely KN, Powell J, *et al.* The
17 transcriptional landscape of age in human peripheral blood. *Nat Commun.* 2015;**6**:8570.
18 10.1038/ncomms9570.
- 19 26. Team RC. R: A language and environment for statistical computing. R Foundation for
20 Statistical Computing. Austria: Vienna. 2018. <http://www.R-project.org/>.
- 21 27. Renson A. eTables and R code for 'Early signals of gut microbiome aging: immune and
22 cardiometabolic phenotypes in young adulthood'. GitHub Repository; 2020.
23 <https://github.com/audreyrenson/addhealthmb>. 10.5281/zenodo.3804462.
- 24 28. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for
25 RNA-seq data with DESeq2. *Genome biology.* 2014;**15**:550. 10.1186/s13059-014-0550-8.
- 26 29. Stephens M. False discovery rates: a new deal. *Biostatistics.* 2017;**18**:275-294.
27 10.1093/biostatistics/kxw041.
- 28 30. Chen L, Reeve J, Zhang L, Huang S, Wang X, Chen J. GMPR: A robust normalization
29 method for zero-inflated count data with application to microbiome sequencing data. *PeerJ.*
30 2018;**6**:e4600. 10.7717/peerj.4600.
- 31 31. Zhao T, Liu H, Roeder K, Lafferty J, Wasserman L. The huge package for high-
32 dimensional undirected graph estimation in R. *Journal of Machine Learning Research.*
33 2012;**13**:1059-1062.
- 34 32. Hawe JS, Theis FJ, Heinig M. Inferring interaction networks from multi-omics data-a
35 review. *Frontiers in genetics.* 2019;**10**:535. 10.3389/fgene.2019.00535

- 1 33. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and
2 compositionally robust inference of microbial ecological networks. *PLoS computational biology*.
3 2015;**11**. 10.1371/journal.pcbi.1004226
- 4 34. Newton K, Dixit VM. Signaling in innate immunity and inflammation. *Cold Spring Harb*
5 *Perspect Biol*. 2012;**4**. 10.1101/cshperspect.a006049.
- 6 35. Buuren Sv, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations
7 in R. *Journal of statistical software*. 2010:1-68. 10.18637/jss.v045.i03.
- 8 36. Liu H, Lafferty J, Wasserman L. The Nonparanormal: Semiparametric Estimation of
9 High Dimensional Undirected Graphs. *J Mach Learn Res*. 2009;**10**:2295-2328.
- 10 37. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network
11 analysis. *BMC bioinformatics*. 2008;**9**:559. 10.1186/1471-2105-9-559.
- 12 38. West NR, Hegazy AN, Owens BMJ, Bullers SJ, Linggi B, Buonocore S, *et al*. Oncostatin
13 M drives intestinal inflammation and predicts response to tumor necrosis factor-neutralizing
14 therapy in patients with inflammatory bowel disease. *Nat Med*. 2017;**23**:579-589.
15 10.1038/nm.4307.
- 16 39. Biagi E, Franceschi C, Rampelli S, Severgnini M, Ostan R, Turroni S, *et al*. Gut
17 Microbiota and Extreme Longevity. *Curr Biol*. 2016;**26**:1480-1485. 10.1016/j.cub.2016.04.016.
- 18 40. Santoro A, Ostan R, Candela M, Biagi E, Brigidi P, Capri M, *et al*. Gut microbiota
19 changes in the extreme decades of human life: a focus on centenarians. *Cell Mol Life Sci*.
20 2018;**75**:129-148. 10.1007/s00018-017-2674-y.
- 21 41. Yang Q, Lin SL, Kwok MK, Leung GM, Schooling CM. The Roles of 27 Genera of
22 Human Gut Microbiota in Ischemic Heart Disease, Type 2 Diabetes Mellitus, and Their Risk
23 Factors: A Mendelian Randomization Study. *Am J Epidemiol*. 2018;**187**:1916-1922.
24 10.1093/aje/kwy096.
- 25 42. Xu W, Larbi A. Markers of T Cell Senescence in Humans. *Int J Mol Sci*. 2017;**18**.
26 10.3390/ijms18081742.
- 27 43. Arpaia N, Campbell C, Fan X, Dikiy S, van der Veeken J, deRoos P, *et al*. Metabolites
28 produced by commensal bacteria promote peripheral regulatory T-cell generation. *Nature*.
29 2013;**504**:451-455. 10.1038/nature12726.
- 30 44. Hernan MA, Robins JM. Causal Inference. Taylor & Francis; 2019.
- 31 45. Renson A, Herd P, Dowd JB. Sick individuals and sick (microbial) populations:
32 Challenges in epidemiology and the microbiome. *Annu Rev Public Health*. 2020.
33 10.1146/annurev-publhealth-040119-094423
- 34 46. Arnold K, Berrie L, Tennant P, Gilthorpe M. A causal inference perspective on the
35 analysis of compositional data. *International journal of epidemiology*. 2020. 10.1093/ije/dyaa021

- 1 47. O'Toole PW, Claesson MJ. Gut microbiota: Changes throughout the lifespan from
2 infancy to elderly. *Int Dairy J.* 2010;**20**:281-291. 10.1016/j.idairyj.2009.11.010.
- 3 48. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, *et al.* Diet
4 rapidly and reproducibly alters the human gut microbiome. *Nature.* 2014;**505**:559-563.
5 10.1038/nature12820.
- 6 49. Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, *et al.*
7 The long-term stability of the human gut microbiota. *Science.* 2013;**341**:1237439.
8 10.1126/science.1237439.
- 9 50. Crimmins E, Kim JK, McCreath H, Faul J, Weir D, Seeman T. Validation of blood-based
10 assays using dried blood spots for use in large population studies. *Biodemography and social*
11 *biology.* 2014;**60**:38-48. 10.1080/19485565.2014.901885.

12

Prior density

150
100
50
0

-0.02

0.00

0.02

Prior β values

KLRG1 high vs. Naive CD8+

LDL

Glucose

CRP

Th1 vs. Naive CD4+

Hallmark ROS

HDL

Naive vs. Effector CD8+

CD57+ vs. - NK

Hallmark Inflammation

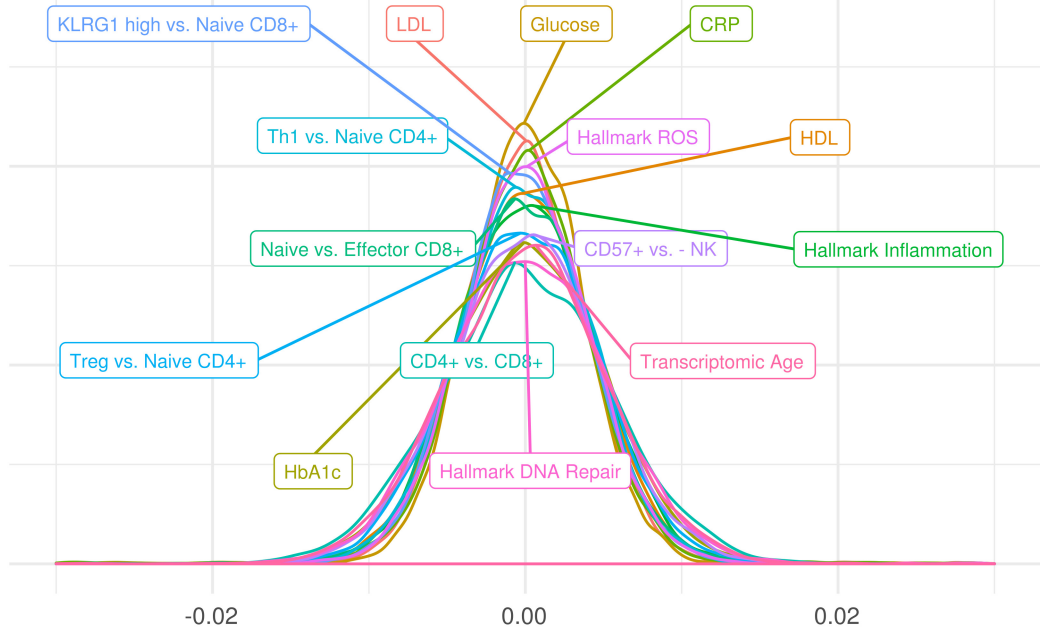
Treg vs. Naive CD4+

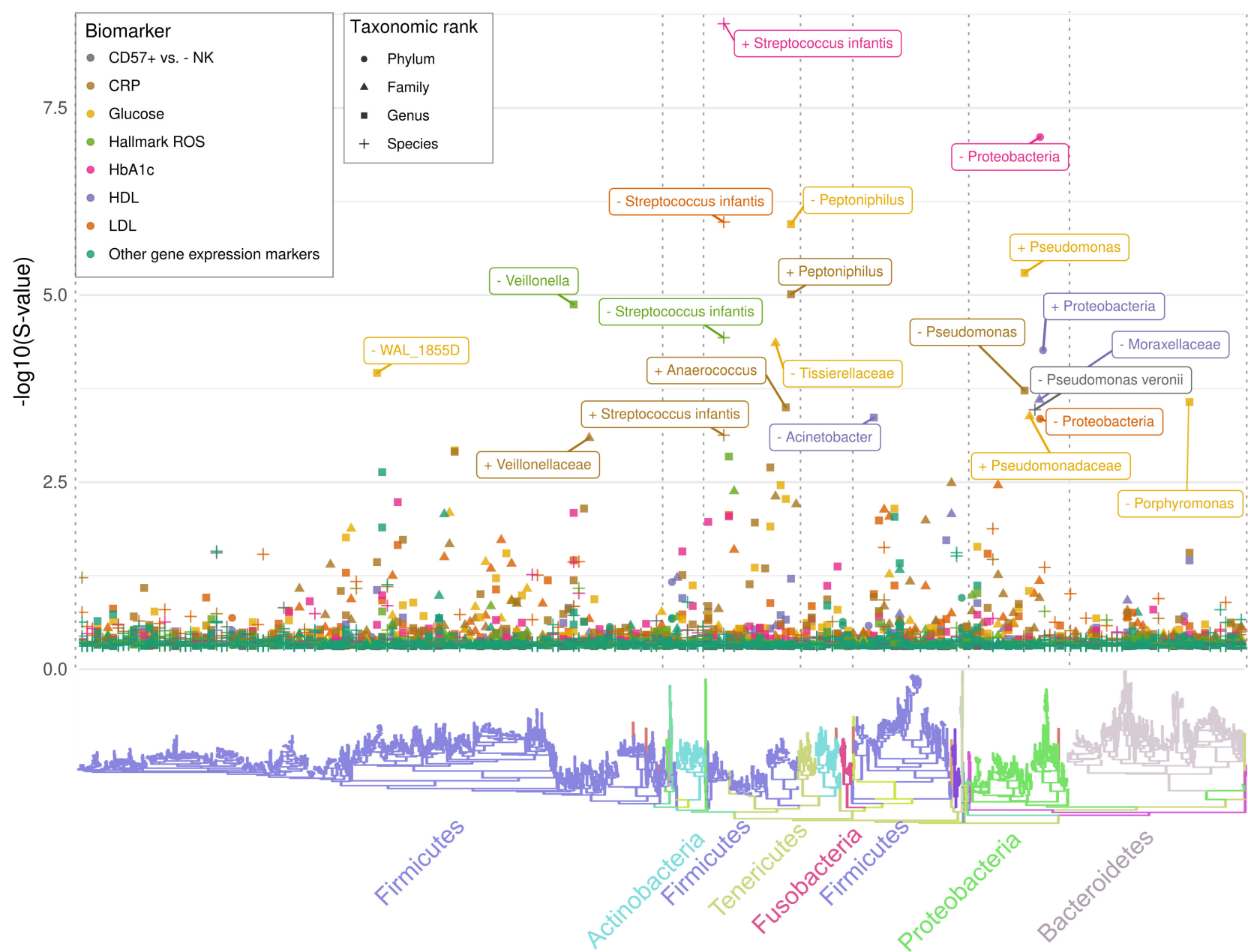
CD4+ vs. CD8+

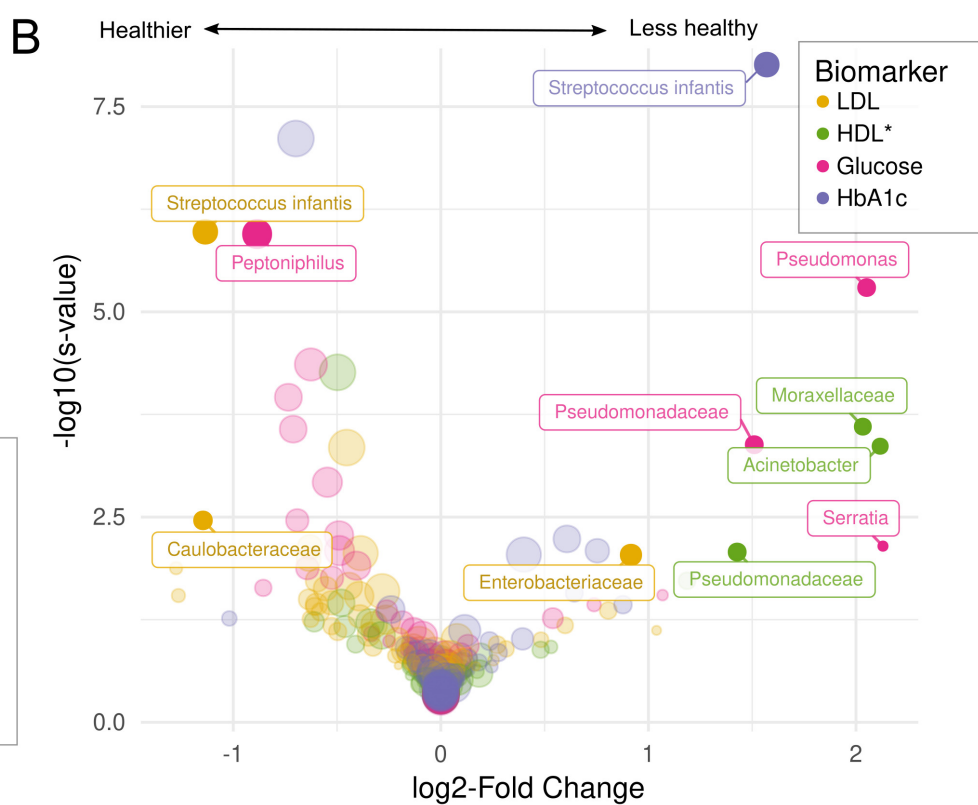
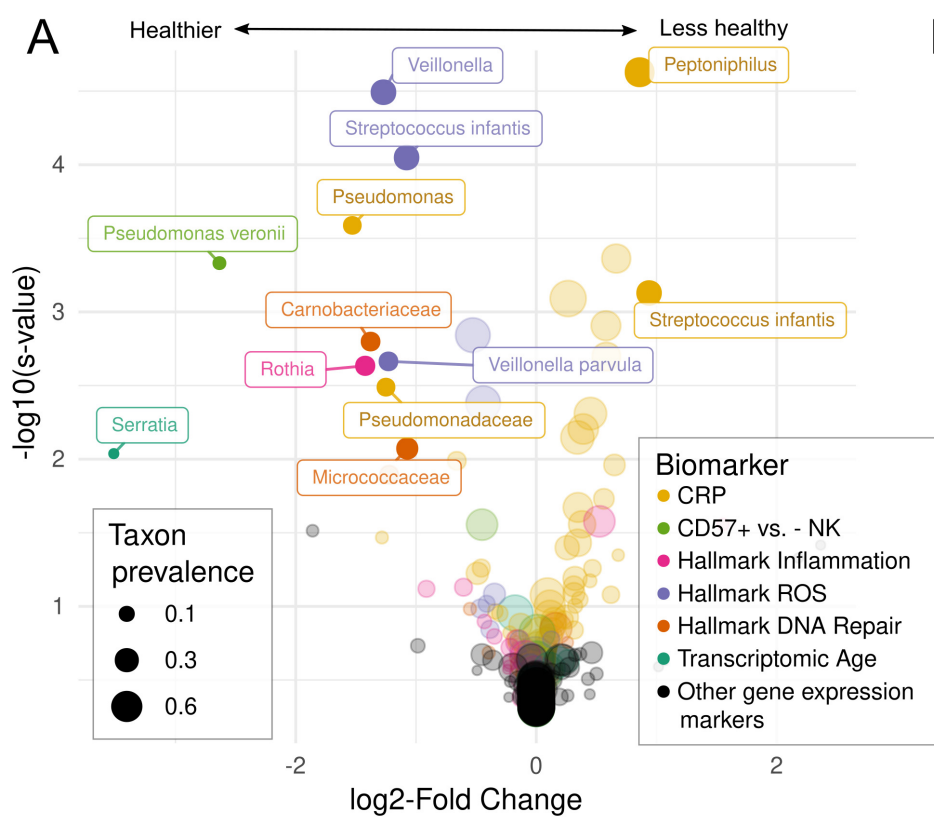
Transcriptomic Age

HbA1c

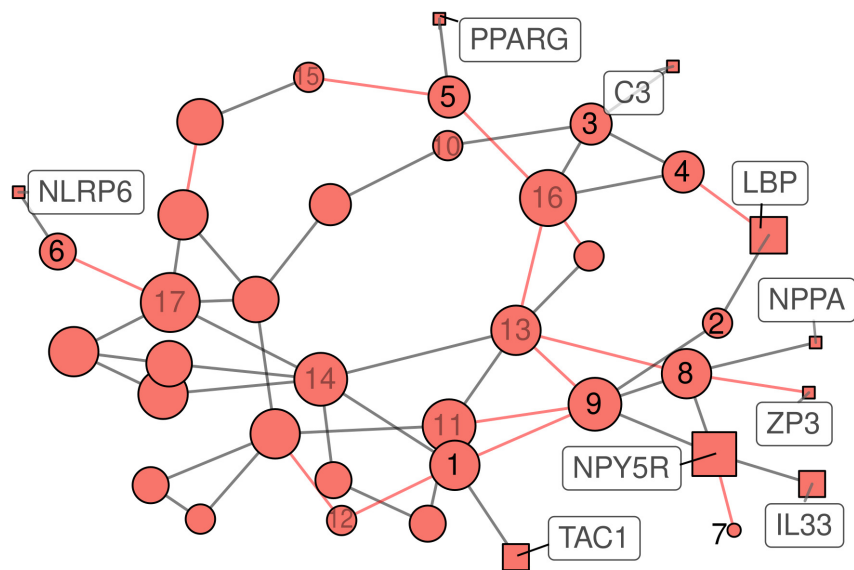
Hallmark DNA Repair





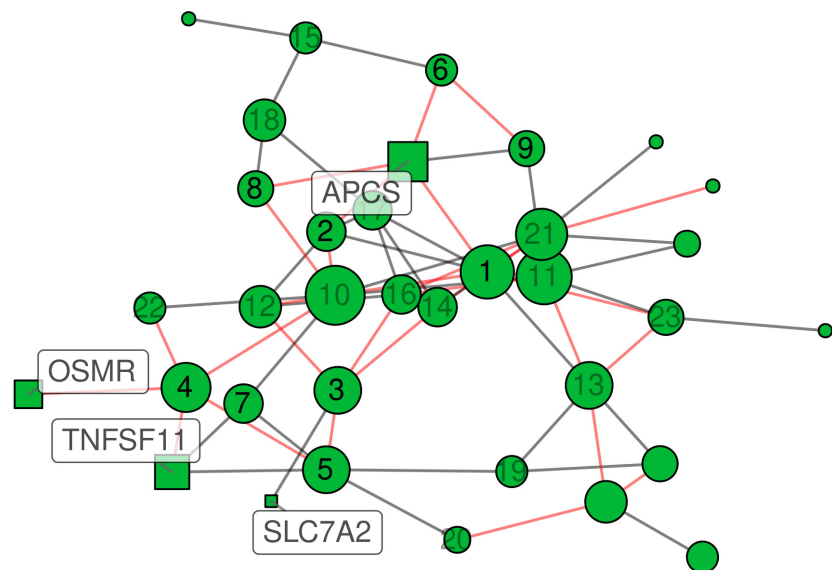


A1



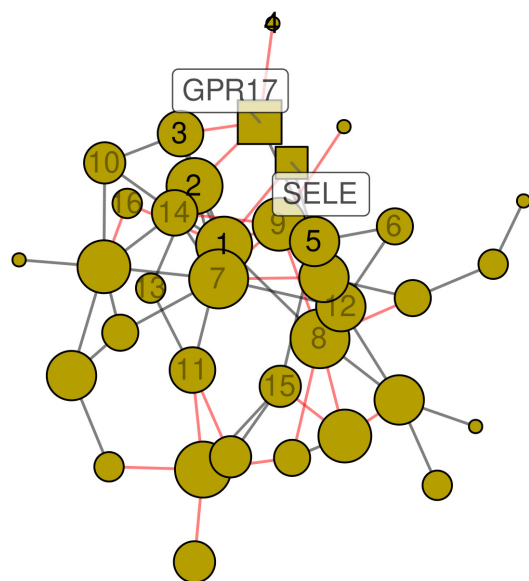
1: *Paraprevotella* sp. (iii); 2: *Butyricimonas* sp. (iii); 3: *Odoribacter* sp. (iii); 4: Order Bacteroidales sp. (i); 5: *Bilophila* sp. (ii); 6: *Desulfovibrio* sp. (i); 7: *Achromobacter* sp.; 8: *Sutterella* sp. (iii); 9: *Bifidobacterium* sp. (i); 10: *Akkermansia muciniphila*; 11: *Prevotella stercorea*; 12: Family [Barnesiellaceae] sp. (ii); 13: *Bacteroides* sp. (v); 14: Order RF32 sp. (ii); 15: *Serratia marcescens*; 16: Order RF39 sp. (iii); 17: Order RF39 sp. (iv)

A2



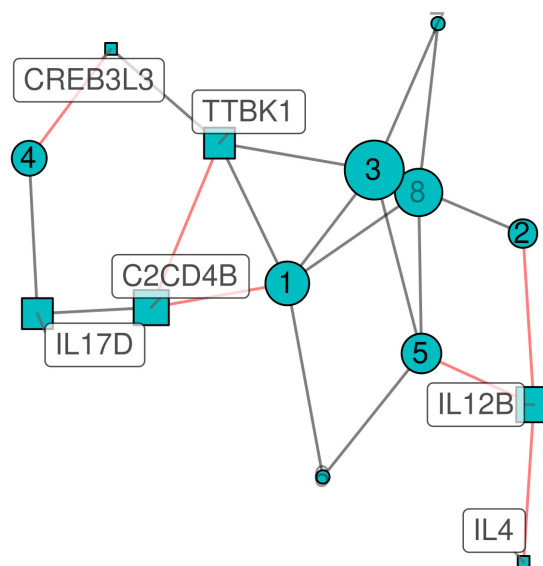
1: *Prevotella copri* (iv); 2: *Desulfovibrio* D168; 3: *Acinetobacter* sp.; 4: *Megasphaera* sp.; 5: Family Planococcaceae sp.; 6: Order Clostridiales sp. (vii); 7: Family Ruminococcaceae sp. (ii); 8: *Coprococcus eutactus*; 9: Family Lachnospiraceae sp. (iv); 10: *Butyricimonas* sp. (ii); 11: Family [Barnesiellaceae] sp. (i); 12: Family [Barnesiellaceae] sp. (vii); 13: *Chryseobacterium* sp.; 14: Family Rikenellaceae sp. (i); 15: Family Rikenellaceae sp. (iii); 16: *Bacteroides* sp. (ii); 17: *Bacteroides* sp. (vii); 18: *Desulfovibrio* sp. (ii); 19: *Sphingomonas* sp.; 20: Order RF32 sp. (iii); 21: Order Clostridiales sp. (iii); 22: *Ruminococcus* sp. (ii); 23: *Ruminococcus* sp. (iii)

A3



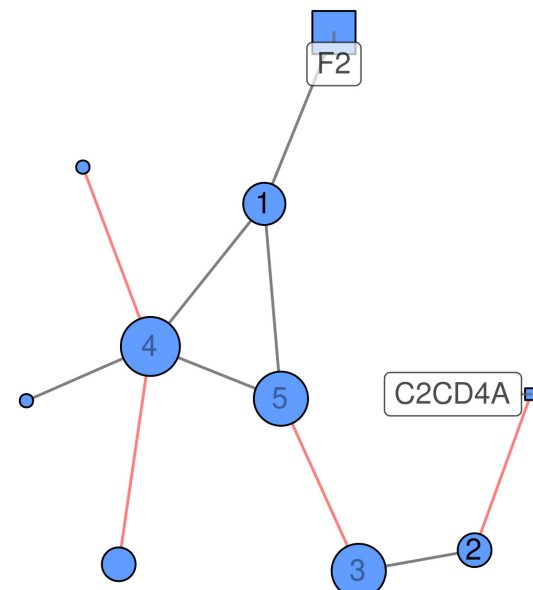
1: [*Prevotella*] sp.; 2: *Parabacteroides* sp. (i); 3: *Bacteroides* sp. (iv); 4: *Bacteroides* sp. (xii); 5: *Ruminococcus* sp. (viii); 6: *Paraprevotella* sp. (ii); 7: *Paraprevotella* sp. (v); 8: *Butyricimonas* sp. (i); 9: *Odoribacter* sp. (ii); 10: *Parabacteroides* sp. (ii); 11: Family [Barnesiellaceae] sp. (v); 12: Order Bacteroidales sp. (ii); 13: *Desulfovibrio* sp. (iii); 14: Order RF32 sp. (i); 15: *Clostridium septicum*; 16: *Roseburia* sp. (i)

A4

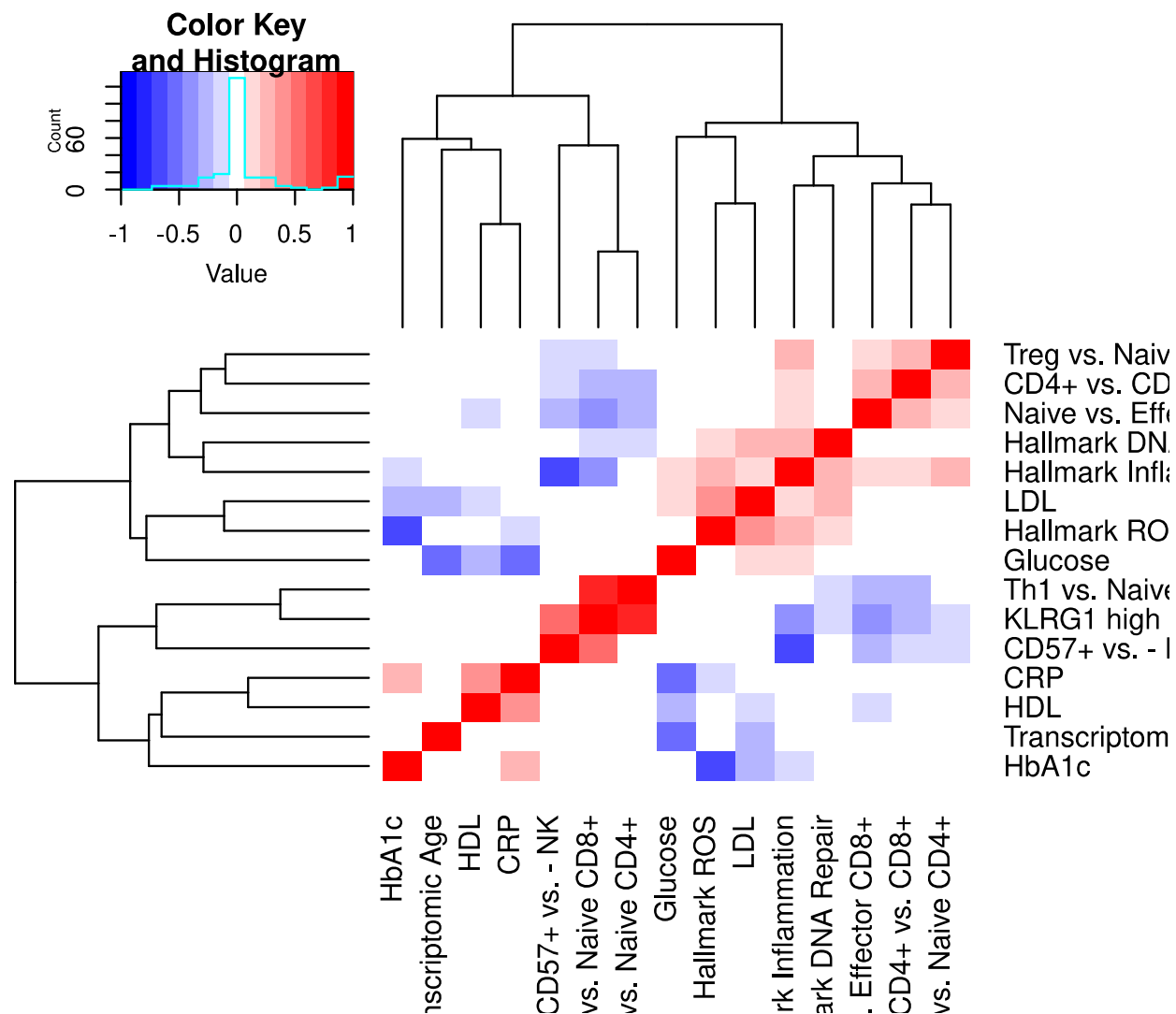


1: Order ML615J-28 sp. (ii); 2: Order Clostridiales sp. (x); 3: *Ruminococcus* sp. (i); 4: Family Ruminococcaceae sp. (iv); 5: *Butyrivibrio crossotus*; 6: *Dialister* sp. (ii); 7: Order Clostridiales sp. (ix); 8: Order Clostridiales sp. (xi)

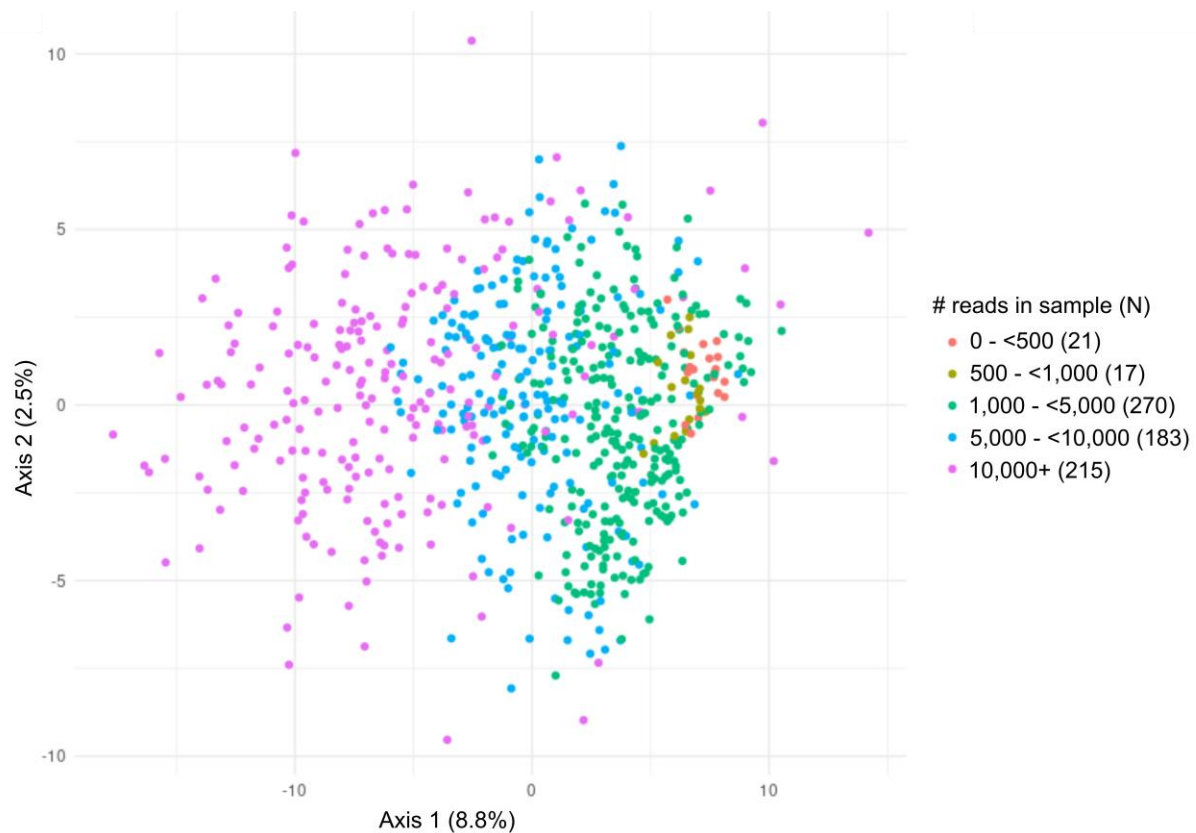
A5



1: Family [Barnesiellaceae] sp. (vi); 2: *Bacteroides eggerthii* (i); 3: *Bacteroides eggerthii* (ii); 4: Order RF39 sp. (ii); 5: Order ML615J-28 sp. (i)



eFigure 1. Pairwise correlations between microbiome signals for each biomarker. Each square represents the correlation between the log₂-fold change estimates between biomarker A and every taxon, and the log₂-fold change estimates between biomarker B and every taxon, where biomarker A and B are on the X and Y axis, respectively.



eFigure 3. Principal component analysis using centered log-ratio transformed microbiome count data, according to number of reads in sample, for all (n=706) samples with sufficient material for sequencing.

eMethods

Measurement of Biomarkers and medications

As part of Wave IV, trained and certified field examiners collected dried blood spots from voluntarily fasting (≥ 8 hours) and non-fasting participants via finger prick on seven-spot, Whatman 903® Protein Saver cards. Cards were dried on location, then shipped to the University of Washington Department of Laboratory Medicine (Seattle, WA) for assay of glucose (mg/dl), low-density lipoprotein (LDL), high-density lipoprotein (HDL), and high-sensitivity c-reactive protein (CRP); and to FlexSite Diagnostics, Inc. (Palm City, FL) for assay of HbA1c (%). Because LDL, HDL, and glucose are provided in deciles, and because gene expression biomarkers other than transcriptomic age are derived from a standardized principal

component scale, these markers are not directly interpretable as clinically relevant changes. Further information about the dried blood spot collection, assays, and medications are available at (1).

Measurement of Gene Expression

After participation in the Wave V survey, respondents were visited by a field examiner/phlebotomist to collect biological specimens including a sample of venous blood. Using RNA isolated from Pax gene, 200 ng of total RNA were converted to cDNA (Lexogen QuantSeq) and sequenced using an Illumina HiSeq 4000 instrument in the University of California Los Angeles, Neuroscience Genomics Core Laboratory. Each sample yielded >10 million 65nt single-strand sequencing reads, which were mapped to the RefSeq human genome sequence reference (ENSEMBL hg38) and quantified as transcript counts per million mapped reads using STAR aligner v 2.6.(2)

Fecal microbiome sample collection and processing

All microbial DNA was prepared, stored, and sequenced at CU Boulder. Microbial DNA was prepared with a QIAGEN high-throughput system (MagAttract Power Soil Kit EP). Approximately 50ng of DNA from each sample was subjected to PCR with barcoded sequencing primers specific for V4 of the 16S rDNA in triplicate. The following primers were used: 515F (Parada)–806R (Apprill), forward-barcoded: FWD:GTGYCAGCMGCCGCGGTAA; REV:GGACTACNVGGGTWTCTAAT. The triplicates were then pooled, quantitated using picogreen, mixed in equimolar amounts and subjected to paired-end DNA sequence analysis on an Illumina MiSeq Personal Sequencer. Illumina base quality scores of greater than 30 were required. We used QIIME2 (3) v. 2019.4 for all the following data processing steps. Filtered

reads were demultiplexed to assign reads to individual samples, using the Deblur pipeline (4) to determine operational taxonomic units (OTUs) at single-nucleotide resolution. Clustering of OTUs was performed at the 97% similarity level and taxonomy was assigned using closed-reference picking from GreenGenes v. 13-8, which has been shown to be highly effective at identifying taxa from constructed complex samples.(5) Previous work by the American Gut Project has characterized the taxonomic changes expected due to specimen tubes being shipped at room temperature with transit times up to four days, which consist of “blooms” of particular species.(6) As an additional quality control step, we used an algorithm developed by the Knight lab to remove reads identified as reflecting microbial blooms occurring during room temperature shipping using Deblur.(7) Finally, we removed samples with less than 1,000 reads prior to analysis. We selected this cutoff as a compromise between including as many samples as possible and preserving as much variability as possible in the multivariate distribution of the data in compositional PCA (shown in eFigure 3).

1. Whitsel EA, Tabor JW, Nguyen QC, Cuthbertson CC, Wener MH, Potter AJ, *et al.* Add Health Wave IV: Documentation report measures of glucose homeostasis. Available at:< http://www.cpc.unc.edu/projects/addhealth/data/guides/Glucose_HbA1c.pdf. 2012.
2. Dobin A, Gingeras TR. Mapping RNA-seq reads with STAR. *Current protocols in bioinformatics*. 2015;**51**:11.14. 11-11.14. 19.
3. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature biotechnology*. 2019;**37**:852-857.

4. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Xu ZZ, *et al.* Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems*. 2017;**2**.
5. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, *et al.* Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*. 2018;**6**:90.
6. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, *et al.* American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems*. 2018;**3**:e00031-00018.
7. Amir A, McDonald D, Navas-Molina JA, Debelius J, Morton JT, Hyde E, *et al.* Correcting for Microbial Blooms in Fecal Samples during Room-Temperature Shipping. *mSystems*. 2017;**2**.