bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Title: Total RNA Sequencing reveals microbial communities in human blood and disease specific effects

Authors: Serghei Mangul[#1], Loes M Olde Loohuis[#2], Anil P Ori[2], Guillaume Jospin[3], David Koslicki[4], Harry Taegyun Yang[1], Timothy Wu[2], Marco P Boks[5], Catherine Lomen-Hoerth[6], Martina Wiedau-Pazos[7], Rita M Cantor[8], Willem M de Vos[9,10], René S Kahn[5], Eleazar Eskin[1], Roel A Ophoff[*,2,5,8]

Affiliations:

[1]Department of Computer Science, University of California Los Angeles, Los Angeles, USA

[2]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University California Los Angeles, Los Angeles, USA

[3]Davis Genome Center, University of California, Davis, USA

[4]Mathematics Department, Oregon State University, Corvallis, USA

[5]Department of Psychiatry, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands

[6]Department of Neurology, University of California San Francisco, San Francisco, USA

[7]Department of Neurology, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, USA

[8]Department of Human Genetics, University of California Los Angeles, Los Angeles, USA

[9]Laboratory of Microbiology, Wageningen University, Ede, The Netherlands

[10]Department of Bacteriology and Immunology, Immunobiology Research Program, University of Helsinki, Helsinki, Finland.

* Correspondence to: Roel A. Ophoff ophoff@ucla.edu

# Equal contribution

Running title: RNA-Sequencing reveals microbiome in human blood

Key words: RNA sequencing, blood microbiome, schizophrenia, unmapped reads

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract: An increasing body of evidence suggests an important role of the human microbiome in health and disease. We propose a 'lost and found' pipeline, which examines high quality unmapped sequence reads for microbial taxonomic classification. Using this pipeline, we are able to detect bacterial and archaeal phyla in blood using RNA sequencing (RNA-Seq) data. Careful analyses, including the use of positive and negative control datasets, suggest that these detected phyla represent true microbial communities in whole blood and are not due to contaminants. We applied our pipeline to study the composition of microbial communities present in blood across 192 individuals from four subject groups: schizophrenia (n=48), amyotrophic lateral sclerosis (n=47), bipolar disorder (n=48) and healthy controls (n=49). We observe a significantly increased microbial diversity in schizophrenia compared to the three other groups and replicate this finding in an independent schizophrenia case-control study. Our results demonstrate the potential use of total RNA to study microbes that inhabit the human body.

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Main text:

Introduction

Microbial communities in and on the human body represent a complex mixture of eukaryotes, bacteria, archaea and viruses. High-throughput sequencing offers a powerful culture-independent approach to study the underlying diversity of microbial communities in their natural habits across different human tissues and diseases. Increasing numbers of sequence-based studies investigate the role of the human microbiome in health (Human Microbiome Project 2012), disease (Turnbaugh et al. 2006; Turnbaugh et al. 2009; Abu-Shanab and Quigley 2010; Cho and Blaser 2012; Greenblum et al. 2012), and behavior (Hsiao et al. 2013). Advancing methods to study microbial communities is therefore important in aiding our understanding of the human microbiome.

Little is known about the human microbiome in the blood of donors in the absence of sepsis, as blood has been generally considered a sterile environment lacking proliferating microbes (Drennan 1942). However, over the last few decades, this assumption has been challenged (Nikkari et al. 2001; McLaughlin et al. 2002), and the presence of a microbiome in the blood has received increasing attention (Amar et al. 2011; Sato et al. 2014; Paisse et al. 2016).

The majority of the current studies of the microbiome use fecal samples and targeted 16S ribosomal RNA gene sequencing (de Vos and de Vos 2012). With the availability of comprehensive compendia of reference microbial genomes and phylogenetic marker genes (Darling et al. 2014), it has become feasible to use non-targeted sequencing data to identify the

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

microbial species across different human tissues and diseases in a relatively inexpensive and easy way.

Here, we use whole blood RNA sequencing (RNA-Seq) reads to detect a variety of microbial organisms. Our ‘lost and found’ pipeline utilizes high quality reads that fail to map to the human genome as candidate microbial reads. Since RNA-Seq has become a widely used technology in recent years with many large datasets available, we believe that our pipeline has great potential for application across tissues and disease types.

We applied our ‘lost and found’ pipeline to study the blood microbiome in almost two hundred individuals including patients with schizophrenia, bipolar disorder and amyotrophic lateral sclerosis. There is evidence of involvement of the microbiome in brain function and disease including schizophrenia (Foster and McVey Neufeld 2013; Hsiao et al. 2013; Castro-Nallar et al. 2015; Erny et al. 2015). These three disease groups represent complex traits affecting the central nervous system with both genetic and non-genetic components whose etiology remains largely elusive. Samples have been collected and processed using standardized lab procedures and thus allow us to explore the connection between the microbiome and diseases of the brain using our pipeline. We observed an increased diversity of microbial communities in schizophrenia patients, and replicate this finding in an independent dataset.

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

Studying blood microbiome using RNA-Seq data

To study the composition of the active microbial communities, we determined the microbial meta-transcriptome present in the blood of unaffected controls (Controls, n=49) and patients with three brain-related disorders: schizophrenia (SCZ, n=48), amyotrophic lateral sclerosis (ALS, n=47) and bipolar disorder (BPD, n=48). Peripheral blood was collected from all samples, and RNAseq libraries were prepared from total RNA after using ribo-depletion protocol (Ribo-Zero). (Figure 1A-1C, Table 1 and Table S1A).

We separated human and non-human reads, and use the latter as candidate microbial reads for taxonomic profiling of microbial communities. To identify potentially microbial reads we developed the 'lost and found' pipeline. First, we filtered read pairs and singleton reads mapped to the human genome or transcriptome (Figure 1.D). For normalization purposes, unmapped reads were then sub-sampled to 100,000 reads for each sample. Next, we filtered out low-quality and low-complexity reads using FASTX and SEQCLEAN (see urls). Finally, the remaining reads were realigned to the human references using the Megablast aligner (Camacho et al. 2009) to exclude any potentially human reads. The remaining 33,546 of 100,000 reads are high-quality, unique, non-host reads used as candidate microbial reads in subsequent analyses to determine the taxonomic composition and diversity of the microbial communities in blood (Figure 1.E).

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

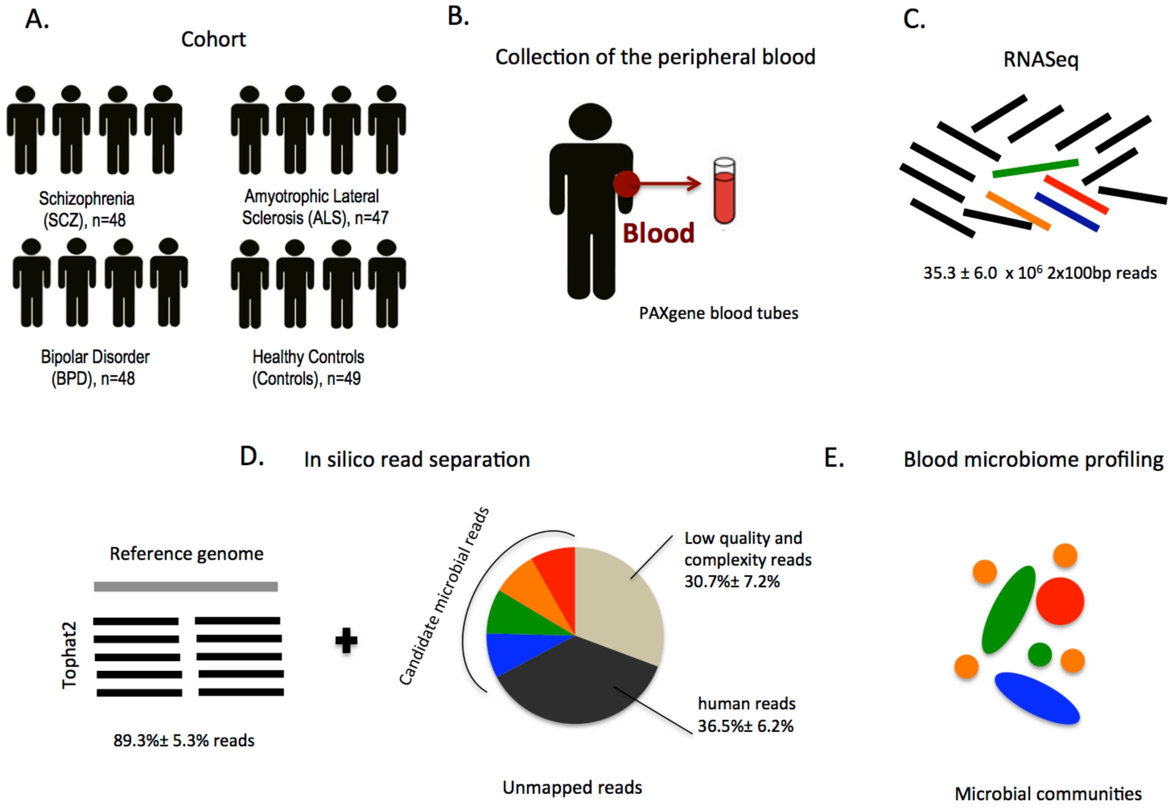
A.
Cohort
Schizophrenia (SCZ), n=48
Amyotrophic Lateral Sclerosis (ALS), n=47
Bipolar Disorder (BPD), n=48
Healthy Controls (Controls), n=49
B.
Collection of the peripheral blood
Blood
PAXgene blood tubes
C.
RNASeq
35.3 ± 6.0 x 10⁶ 2x100bp reads
D.
In silico read separation
Reference genome
Tophat2
89.3%± 5.3% reads
Candidate microbial reads
Low quality and complexity reads 30.7%± 7.2%
human reads 36.5%± 6.2%
Unmapped reads
E.
Blood microbiome profiling
Microbial communities

Figure 1. Framework of blood microbiome profiling using the ‘lost and found’ pipeline. (A) We analyzed a cohort of 192 individuals from four subject groups, i.e. Schizophrenia (SCZ, n=48), amyotrophic lateral sclerosis (ALS n=47), bipolar disorder (BPD n=48), unaffected control subjects (Controls n=49). (B) Peripheral blood was collected for RNA collection. (C) RNAseq libraries were prepared from total RNA using ribo-depletion protocol. (D) Reads that failed to map to the human reference genome and transcriptome were sub-sampled and further filtered to exclude low-quality, low complexity, and remaining potentially human reads. (E) High quality, unique, non-host reads are used to determine the taxonomic composition and diversity of the blood microbiome communities. See also Table S1.

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. Sample Description

Disease Status	Control	SCZ	BPD	ALS
N	49	48	48	47
Age Mean (SD)	41.1 (10.7)	29.9 (5.8)	46.5 (9.9)	56.4 (10.3)
Age Range	[21 – 60]	[22-46]	[26-71]	[35-76]
Male/Female	38/11	39/9	20/28	29/18

Assembly and richness of the blood microbiome

To access the assembly and richness of the blood microbiome we used phylogenetic marker genes to assign the candidate microbial reads to the bacterial and archaeal taxa. We used Phylosift (Darling et al. 2014) to perform phylogenetic and taxonomic analyses of the whole blood samples and compare across individuals. Phylosift makes use of a set of protein coding genes found to be relatively universal (in nearly all bacterial and archaeal taxa) and having low variation in copy number between taxa. Homologs of these genes in new sequence data (e.g., the transcriptomes used here) are identified and then placed into a phylogenetic and taxonomic context by comparison to references from sequenced genomes. We were able to assign 1235 reads (1.24% ± 0.41%) on average to the bacterial and archaeal gene families. A total of 1,880 taxa were assigned with Phylosift, with 23 taxa at the phylum level (Figure 2). Most of the taxa we observed derived from bacteria (relative genomic abundance 89.8% ±

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

7.4%), and a smaller portion from archea (relative genomic abundance 12.28% ±6.4%). We observed no evidence of the presence of nonhuman eukaryotes or viruses.

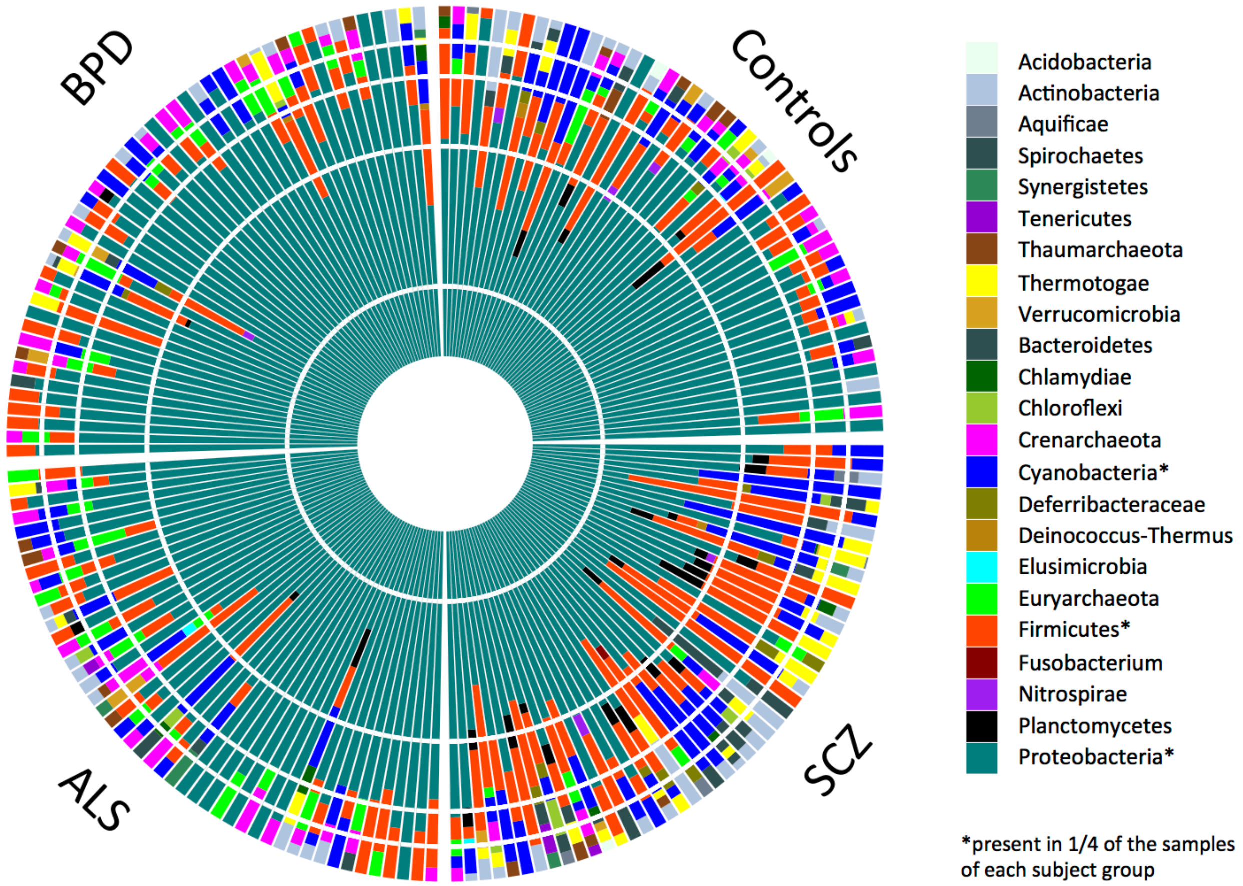

Figure 2. Genomic abundances of microbial taxa at phylum level of classification. Phylogenetic classification is performed using Phylosift able to assign the filtered candidate microbial reads to the microbial genes from 23 distinct taxa on the phylum level.

In total, we observed 23 distinct microbial phyla with on average 4.1 ± 2.0 phyla per individual. The large majority of taxa that were observed in our sample are not universally

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

present in all individuals, except for *Proteobacteria* that are dominating all samples with 73.4% ± 18.3% relative abundance (Figure 2 dark green color). Several bacterial phyla show a broad prevalence across individuals and disorders (present in 1/4 of the samples of each subject group). Those phyla include the *Proteobacteria, Firmicutes,* and *Cyanobacteria* with relative abundance 73.4% ± 18.3%, 14.9 ±10.9%, and 11.0% ± 8.9% (Table S2). This is in line with recent published work on the blood microbiome using 16S targeted metagenomic sequencing reporting between 80.4-87.4% and 3.0-6.4% for *Proteobacteria* and *Firmicutes* at the phylum level, respectively (Païssé et al. 2016). Although *Proteobacteria and Firmicutes, are* are commonly associated with the human microbiome (Consortium and others 2012), some members of these phyla might be associated with reagent and environmental contaminants (Salter et al. 2014) (See also **Validation and potential contamination**).

To compare the inferred blood microbial composition with that in other body sites, we used taxonomic composition of 499 metagenomic samples from Human Microbiome Project (HMP) obtained by MetaPHlAn (v 1.1.0)(Segata et al. 2012) for five major body habitats (gut, oral, airways, and skin) (Human Microbiome Project 2012) (see urls). Of the 23 phyla discovered in our sample, 15 were also found in HMP samples, of which 13 are confirmed by at least ten samples. Our data suggest that the predominant phyla of the blood microbiome are most closely related with the known oral and gut microbiome (Table S2).

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Validation and potential contamination

To investigate the possibility of DNA contamination introduced during RNA isolation, library preparation, and sequencing steps, we performed the following negative control experiment. We applied our 'lost and found' pipeline to RNA-Seq reads from six B-lymphoblast cell line (LCLs) samples that are expected to be sterile and lack any traces of microbial species. Neither Phylosift nor MetaPHlAn detected bacterial or archaeal microorganisms in the LCLs samples (See Table S1.C). This experiment also serves as a positive control, as the only virus Phylosift does detect is the Epstein-Barr virus, used for transfection and transformation of lymphocytes to lymphoblasts (Santpere et al. 2014).

We used a more direct positive control dataset to validate the feasibility of using human RNA-Seq to detect microbial organisms and applied the 'lost and found' pipeline to RNA-Seq data collected from epithelial cells infected with *Chlamydia* (Humphrys et al. 2013). The authors collected data using ribo-depletion and polyA selection protocols at 1 and 24 hours post infection. Phylosift was able to detect the Chlamydia phylum in 100,000 reads randomly subsampled from unmapped reads, confirming, as with above mentioned Epstein-Barr virus in LCLS, the validity of the bioinformatic pipeline used (Table S3).

The design of experimental procedures such as blood draw and subsequent downstream lab procedures may lead to global contamination effects. In our data, there is minimal evidence that the detected microbial communities are confounded by contamination due to experimental procedures. First, all RNA samples were subjected to the same standardized RNA isolation protocols, library preparation, and sequencing procedures. With the exception of Proteobacteria, which has been reported to be the most abundant phylum in

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

whole blood (Paisse et al. 2016), we observe no phylum present in all individuals, suggesting absence of a uniform contaminator due to experimental procedures applied across all samples.

Second, we collected two blood tubes per individual of which one is randomly chosen for subsequent RNA sequencing. If skin contamination upon first blood draw occurs, due to contact with the needle, its effect will be randomly distributed across half of individuals in our cohort and should therefore not affect downstream between-group analyses.

Third, it is vital to scrutinize the potential impact of parameters that are variable between samples, such as experimenter (i.e. lab technician who extracted RNA from blood collections) (Weiss et al. 2014). To investigate these potential effects we grouped samples by various experimental variables, including sequencing run and experimenter. We observe no evidence that the detected microbial communities are confounded by contamination, which is in agreement with previously reported low background signal introduced by such variables (Paisse et al. 2016) (See also Figure S1 and S2). In addition, we include all available technical covariates such as RNA integrity number (RIN), batch, flow cell lane and RNA concentration, in our disease specific analyses.

Finally, an independent technology was used to validate the detected microbial composition in our RNA-Seq cohort. We used available blood whole exome sequence data from two individuals from the cohort (See Table S1.B). We applied the ‘lost and found’ pipeline and compared results from both technologies. Despite the use of different technologies and reagents, microbiome profiles from both sequencing procedures were found to be in close agreement. For both individuals, we were able to detect several microbial phyla, all of which

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

were also identified using RNAseq. Conversely, RNAseq was able to detect several microbial phyla not detected using exome sequencing (Table S4). Taken together, these results confirm the validity and potential of our 'lost and found' pipeline.

Increased microbial diversity in schizophrenia samples

To evaluate potential differences in microbial profiles of individuals with the different disorders (SCZ, BPD, ALS) and unaffected controls, we explored the composition and richness of the microbial communities across the groups. We focused on alpha diversity to study microbial differences at a personal level. To compute alpha diversity, we used the inverse Simpson index which simultaneously assesses both richness (corresponding to the number of distinct taxa) and relative abundance of the microbial communities within each sample (Simpson 1949). In particular, this index allows to effectively distinguish between the microbial communities shaped by the dominant taxa and the communities with many taxa with even abundances (Whittaker 1972).

We observed increased alpha diversity in schizophrenia samples compared to all other groups (Table 2). These differences are statistically significant after adjusting for sex and age, and technical covariates (RIN value, batch, flow cell lane and RNA concentration) using normalized values of alpha (Figure 3a) (*ANCOVA* P < 0.005 for all groups Table 2 and Table S5), and survive *Bonferroni* correction for multiple testing. These differences are independent of potential confounders, such as experimenter and RNA extraction run (Figure S1 and S2) and are not the consequence of a different number of reads being detected as microbial in

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

schizophrenia samples (see Supplementary Results). No significant differences were observed between the three remaining groups (BPD, ALS, Controls). In our sample, alpha diversity was found to be a significant predictor of schizophrenia status and explained 5.0% of the variation as measured by reduction in Nagelkerke's R^2 from logistic regression. To investigate a potential relation between genetic load of schizophrenia susceptibility alleles and microbial diversity, we tested for a correlation between polygenic risk scores (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014) and alpha diversity in samples for which both RNA-Seq and genotyping data was available. No such correlation was observed in our schizophrenia sample (n= 32, Kendall's tau= 0.008, *P* = 0.96). We also did not observe differences in alpha diversity between sexes or across ages. Alpha diversity at other main taxonomic ranks yields a similar pattern of increased diversity in Schizophrenia (Figure S3).

The increased diversity observed in schizophrenia patients may be due to specific phyla characteristic to schizophrenia, or due to a more general increased microbial diversity in people affected by the disease. To investigate this, we compared diversity across individuals within the schizophrenia group to control samples. We used the Bray-Curtis beta diversity metric to measure the respective notions of internal beta diversity (within samples from the same group) and external beta diversity (across samples from different groups). Beta diversity measures the turnover of taxa between two samples in terms of gain or loss of taxa, as well as the differences in abundances between the shared taxa.

In our data, we compared beta diversity across pairs of samples with schizophrenia and controls, resulting in three subject groups: SCZ_Controls, SCZ_SCZ, and Controls_Controls. The lowest diversity was observed in the Controls_Controls group (0.43 ± 0.21), followed by

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SCZ_SCZ (0.50 ± 0.14), and the highest beta diversity values for SCZ_Controls (0.51 ± 0.17) (P< 0.05 for each comparison, by ANCOVA after correcting for three tests). Thus, the observed increased alpha diversity in schizophrenia is not caused by a particular microbial profile, but most likely represents a non-specific overall increased microbial burden (see also Figure S4 and Supplementary Results).

In addition to measuring individual microbial diversity (alpha), and diversity between individuals (beta), we measure the total richness of blood microbiome by the total number of distinct taxa of the microbiome community observed within an entire subject group (gamma diversity (Jost 2007)). We observed that all 23 distinct phyla are observed in schizophrenia: gamma(SCZ)=23 compared to gamma(Controls)=20, gamma(ALS)=16 and gamma(BPD)=18.

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

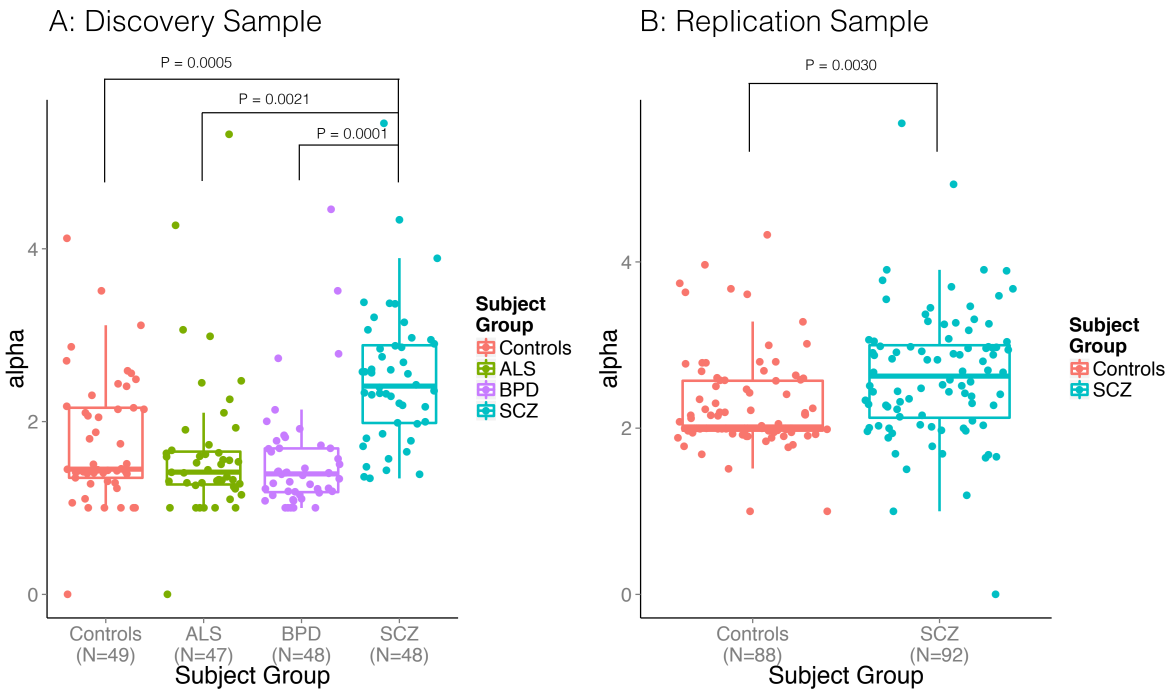

Figure. 3. Increased diversity of human blood microbiome among schizophrenia samples. (A) Alpha diversity per sample for four subject groups (Controls, ALS, BPD, SCZ), measured using the inverse Simpson index on the phylum level of classification. Schizophrenia samples show increased diversity compared to all three other groups (ANCOVA P < 0.005 for all groups, after adjustment of covariates, see also Methods, Table S5 and Figure S3). (B) Alpha diversity per sample of schizophrenia cases and controls, measured using the inverse Simpson index on the genus level of classification. Schizophrenia samples show increased within-subject diversity compared to Controls (P = 0.003 after adjustment of covariates).

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2. Microbial Diversity measures

Disease Status	Control	SCZ	BPD	ALS
N	49	48	48	47
Alpha diversity Mean (SD)	1.77 (0.74)	2.50 (0.79)	1.55 (0.66)	1.65 (0.86)
Beta diversity Mean (SD)	0.43 (0.21)	0.50 (0.14)	0.31 (0.17)	0.38 (0.22)
Gamma diversity (Per group)	20	23	18	16

Reference-free microbiome analysis

Reference-based methods (Phylosift and MetaPhlan) were complemented with the reference-independent method EMDeBruijn (see url). By this method, candidate microbial reads are condensed into a DeBruijn graph, and differences between samples are measured by quantifying how to transform one individual graph into the other. Using the resulting dissimilarities between the samples, principal coordinates are obtained by principal coordinates analysis (PCoA) (Cox and Cox 2000).

We observe that the EMDeBruijn results were in close agreement with the results obtained from Phylosift. EMDeBruin distances measured between samples correlated significantly with beta diversity (spearman rank $P < 2.2e-16$, $rho = 0.37$, including SCZ and Controls). Also, EMDeBruijn PCs significantly correlated with principal components obtained from edge PCA based on the Phylosift taxonomic classification (Correlation between

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

EMDeBruijn PC1, and Phylosift PC1 is P = 1.824e-09, rho = -0.42, Spearman rank correlation, see also Figure S5). The first three EMDeBruijn PCs are significant predictors of schizophrenia status after correcting covariates, and jointly explained 7.1% of the variance as measured by Nagelkerke R-squared (P< 0.05 for each PC).

Replication

We performed a replication study using peripheral blood from two independent subject groups: schizophrenia (SCZ n=91) and healthy controls (Controls n=88) (See Table S1.D). RNAseq libraries for the replication sample were prepared from total RNA using poly(A) enrichment of the mRNA, a more selective procedure than the total RNA that was used for the discovery sample. Microbial profiling was performed using MetaPHlAn (Segata et al. 2012).

In these samples, we replicated our main finding of increased microbial diversity in patients with schizophrenia. In particular, schizophrenia samples showed increased alpha diversity on genus level (2.73 ± 0.77 for cases, versus 2.32 ± 0.57 for controls, corrected P = 0.003 Figure 3b), and explained 2.5% of variance as measured by reduction in Nagelkerke R^2. While our original analysis was performed on the phylum level, in our discovery sample we observe a similar increase of diversity at the genus level (see Figure S3). Just as in our discovery cohort, we observed no significant correlation between alpha diversity and age or differences across gender.

For beta diversity, the pattern we observed slightly diverged from the results obtained from our discovery cohort: while Controls_Controls still has the lowest average beta diversity,

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

we observed increased beta diversity in SCZ_SCZ group versus SCZ_Controls (P < 0.0001 Figure S4). One potential explanation for this discrepancy is that beta diversity in the replication sample was computed at the genus rather than phylum level, making slight mismatches between individuals more likely, and distances between samples hard to compute based on present microbial taxa. This is expected to be more likely if both samples have a large microbial diversity. In relation to this, contrary to what we observed in the discovery sample, we did not observe a correlation between EMdeBruijn distances and Beta diversity in this sample.

However, as in our discovery sample, EMDeBruijn PCs significantly correlated with principal components obtained from edge PCA based on the MetaPHlAn taxonomic classification (Correlation between EMDeBruijn PC1, and MetaPHlAn PC1 is P = 6.091e-06, rho = -0.32 Spearman rank correlation, see also Figure S5). Finally, as in our discovery sample, the first three EMDeBruijn principal components adjusted for covariates were significant predictors of status and together explain 7.8% of the variance.

Cell type composition and diversity

We hypothesized that differences in microbial diversity may be linked to whole blood cell type composition. Since the actual cell counts were not available for these individuals, we used cell-proportion estimates derived from available DNA methylation data to test this hypothesis (Houseman et al. 2012; Aryee et al. 2014; Horvath and Levine 2015).

We assessed methylation data from 65 controls from our replication sample, and compared methylation-derived blood cell proportions to alpha diversity after adjusting for age, gender,

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RIN, and all technical parameters. We tested whether alpha diversity levels are associated to cell type abundance estimates. Our analysis shows one cell type, $CD8^{+}CD28^{-}CD45RA^{-}$ cells, to be significantly negatively correlated with alpha diversity after correction for all other cell-count estimates (correlation = -0.41, P=7.3e-4, Figure S6, Table S6). These cells are T cells that lack $CD8^{+}$ naïve cell markers CD28 and CD45RA and are thought to represent a subpopulation of differentiated $CD8^{+}$ T cells (Koch et al. 2008; Horvath and Levine 2015). We observed that low alpha diversity correlates with high levels of this population of T cells cell abundance.

Discussion

We used high throughput RNA sequencing from whole blood to perform microbiome profiling of active microbial communities and identified an increased diversity in schizophrenia patients. Using our 'lost and found' pipeline, we consistently detected a wide range of microbial phyla in blood. The detection of microbial RNA transcripts in blood is consistent with the possibility of microbial activity in blood and with the possible role of such microbes in health and disease.

While other studies of human microbiome using RNA-Seq have been conducted (Croucher and Thomson 2010)(McClure et al. 2013), this is the first study assessing the microbiome from whole blood by using unmapped non-human total RNA-Seq reads as microbial candidates. Despite the fact that transcripts are present at much lower fractions than human reads, we were able to detect microbial transcripts from bacteria and archaea in almost all samples. The microbes found in blood are thought to be originating from the gut as well as oral cavities (Potgieter et al. 2015; Spadoni et al. 2015), which is in line with our finding that the microbial

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

profiles found in our study most closely resemble the gut and oral microbiome as profiled by the HMP (Human Microbiome Project 2012). The taxonomic profile of the cohort samples suggests the prevalence of the several phyla, *Proteobacteria*, *Firmicutes* and *Cyanobacteria*, across individuals and different disorders included in our study. This is in line with a recent study using 16S targeted metagenomic sequencing, which reported *Proteobacteria* and *Firmicutes* among the most abundant phyla detected in blood (Païssé et al. 2016).

Our study demonstrates the value of analyzing non-human reads present in the RNA-Seq data to study the microbial composition of a tissue of interest (Kostic et al. 2011; Jorth et al. 2014). The RNA-Seq approach avoids biases introduced by primers in 16S ribosomal RNA gene profiling. In addition, compared to genome sequencing, RNA-Seq might offer a potential advantage of avoiding contamination of genomic DNA by dead cells (Ben-Amor et al. 2005). Given the many large-scale RNA-Seq datasets that are already available or currently being generated, we anticipate that high-throughput metatranscriptome-based microbiome profiling will find broader applications in studies across different tissues and disease types.

Rigorous quality control is critically important for any high-throughput sequencing project, especially in the context of studying the microbiome (Salter et al. 2014). To this end, we only considered high quality non-human reads and map them to genes that allow for differentiation between taxa. In our study we carefully evaluated possible contamination effects introduced during the experiments, and accounted for potential bias of relevant RNA-Seq technical aspects in all our analyses. In addition, we performed both negative and positive control experiments to test the feasibility and applicability of our pipeline. To address potential contamination, we performed our pipeline on sterile microbiome-free B-lymphoblastoid cell

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

lines and detected no microbiome other than the Epstein Barr virus used to transfect the cells. As a positive control, we used RNA-Seq data from cells infected with Chlamydia and were able to detect the Chlamydia phylum. By comparing results from RNA-Seq and exome sequencing of two individuals from our cohort, we also tested robustness with respect to sequencing technique. These findings validate our 'lost and found' bioinformatics pipeline in its ability to detect microbial communities using unmapped non-human reads derived from total RNA-Seq.

The most striking finding of our study relating to brain-related diseases is that schizophrenia patients have an increased microbial alpha diversity compared to controls as well as to the other two disease groups (ALS, bipolar disorder). This observation is replicated in an independent sample of schizophrenia cases and controls. The discovery sample was based on total RNA sequencing after depletion of ribosomal transcripts while the replication sample was based on polyA selected RNA. We estimate that total RNA may be better equipped to study the metagenome in whole blood. Despite these differences in methods, the replication sample provides strong evidence for a schizophrenia-specific increased alpha diversity of the blood microbiome, explaining roughly 5% of disease variation. We do not only observe an increased individual microbial diversity, but also an increased diversity between individuals (Beta diversity) with schizophrenia compared to controls, rendering it unlikely that a single phylum or microbial profile is causing the disease-specific increase in diversity.

For the study of microbiome diversity we employed reference-based methods (Phylosift and MethPhlan) as well as the EMDebruin method, a purely reference-agnostic approach. The latter showed strong correspondence to both reference-based methods, highlighting the value of this unbiased sequence-based analysis for investigating microbial differences across groups.

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We recognize, however, that the EMDebruin captures variation in sequence data that may not only depend on differences in distribution of microbial transcripts but also of transcripts of other yet unknown origin.

The increased microbial diversity observed in schizophrenia could be part of the disease etiology (i.e. causing schizophrenia) or may be a secondary effect of disease status. In our sample, we observed no correlation between increased microbial diversity and genetic risk for schizophrenia as measured by polygenic risk scores (Ripke et al. 2013a). In addition, it is remarkable that bipolar disorder, which is genetically and clinically related to schizophrenia (Bulik-Sullivan et al. 2015), does not show a similar increased diversity. We did observe however, a strong inverse correlation between increased diversity and estimated cell abundance of a population of T-cells in healthy controls. Even though this finding is based on indirect cell count measures using DNA methylation data (Horvath and Levine 2015), the significant correlation highlights a likely close connection between the immune system and the active blood microbiome, a relationship that has been documented before (Belkaid and Hand 2014). In the absence of a direct link with genetic susceptibility and the reported correlation with the immune system, we hypothesize that the observed effect in schizophrenia is secondary to disease. This may be a consequence of lifestyle differences of schizophrenia patients including smoking, drug use, or other environmental exposures. Future targeted and/or longitudinal studies with larger sample sizes, detailed clinical phenotypes, and more in-depth sequencing are needed to corroborate this hypothesis.

We hope that our finding of increased diversity in schizophrenia will ultimately lead to a better understanding of the functional mechanisms underlying the connection between immune

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

system, blood microbiome, and disease etiology. With the increasing availability of large scale RNA-Seq datasets collected from different phenotypes and tissue types, we anticipate that the application of our ‘lost and found’ pipeline will lead to the generation of a range of novel hypotheses, ultimately aiding our understanding of the role of the microbiome in health and disease.

Methods

Sample Description

Schizophrenia, bipolar patients and control subjects included in this study were recruited at the University Medical Center Utrecht, The Netherlands. Detailed medical and psychiatric histories were collected, and only patients with a DSM-IV diagnosis of schizophrenia or bipolar disorder were included as cases; controls were neurologically healthy individuals, free of any psychiatric history (Buizer-Voskamp et al. 2011; Ripke et al. 2013b; Loohuis et al. 2015). ALS patients were recruited from ALS clinics at UCLA and UCSF. Whole blood was collected in PAXgene Blood RNA tubes and total RNA was isolated using the PAXgene extraction kit (Qiagen). For DNA, whole blood was collected in EDTA tubes and the extraction performed using the Kleargene XL blood DNA extraction kit. HapMap B-lymphoblast cell lines (n=3) were cultured up to 5 days. All cell lines were grown in RPMI 1640 (Sigma-Aldrich) supplemented with 15% fetal bovine serum (Fisher Scientific) and 2mM L-glutamine (Fisher Scientific) at 37C and 5% (vol/vol) CO2 in a humidified incubator. Cell pellets were lysed with Buffer RLT and RNA was extracted using the RNeasy Mini Kit (Qiagen). All study methods were approved by the institutional review board of

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the University of California at Los Angeles, San Francisco or the Medical Research Ethics Committee of the university Medical Center Utrecht at The Netherlands. All individuals enrolled in these studies provided written informed consent.

Sample sequencing

Discovery sample

RNAseq libraries were prepared using Illumina's TruSeq RNA v2 protocol, including ribo-depletion protocol (Ribo-Zero Gold). Sequencing was performed by UCLA Sequencing Core using the Illumina Hiseq 2000 platform. In total we obtained 6.8 billions 2x100bp paired-end reads (1355 Gbp) of paired-end reads for the primary study(35.3 ± 6.0 million paired-end reads per sample).

Replication sample

For the replication sample, RNAseq libraries were prepared using Illumina's TruSeq RNA v2 protocol, with poly(A) enrichment. A total of 3.8 billion reads (760 Gbp) were obtained (26.3 ± 12.0 million paired-end reads per sample).

RNASeq of B-lymphoblast cell lines

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Just as our discovery samples, we used the TruSeq RNA v2 library preparation, including Ribo-Zero Gold rRNA depletion. Samples were collected from a trio (father, mother, offspring) in duplicate. We obtained 144.6 million 2x69bp paired-end reads (Rapid run).

Whole Blood Exome Sequencing

DNAseq libraries were prepared using Illumina's TruSeq protocol, using the TruSeq Exome enrichment kit.

For all samples, RIN values were obtained using Agilent's RNA 6000 Nano kit and 2100 Bioanalyzer and measures for RNA concentration were obtained using the Quant-iT RiboGreen RNA Assay Kit.

Sequence Analysis: "Lost and found pipeline"

Candidate microbial reads were obtained as follows. We filtered reads mapped to the human reference genome and transcriptome (tophat v. 2.0.12 with default parameters, ENSEMBL GRCh37 transcriptome and ENSEMBL hg19 build). Tophat2 was supplied with a set of known transcripts (as a GTF formatted file, Ensembl GRCh37) using –G option. . Unmapped reads were sub-sampled to 0.1 million reads file, Ensembl GRCh37) using –G option. We used a multi-stage procedure to filter out non-microbial reads. First, to reduce bias of coverage, unmapped reads were sub-sampled to 0.1 million reads for the taxonomic survey of the microbial communities. Then, we filtered out low-quality and low-complexity reads, that is reads with at least 75% of their base pairs with quality lower then 30 (FASTX,

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

http://hannonlab.cshl.edu/fastx_toolkit/) and reads with sequences of consecutive repetitive nucleotides (SEQCLEAN, http://sourceforge.net/projects/seqclean/), respectively. Next, the remaining reads were realigned to the reference human genome and transcriptome (ENSEMBL GRCh37 transcripome and ENSEMBL hg19 build) using the Megablast aligner (BLAST+ 2.2.30, edit distance 6) (Camacho et al. 2009) to filter out any remaining potentially human reads. We prepared the index from each reference sequence using makembindex from BLAST+. The following parameters were used for makembindex: iformat = blastdb. The following options were used to map the reads using Megablast: for each reference: task = megablast, use_index = true, perc_identity = 94, outfmt =6, max_target_seqs =1. We consider only entirely mapped reads. Reads mapped to the human reference genome and transcriptome were identified as 'unmapped human reads' and filtered out. The remaining unmapped reads were used in subsequent analyses.

Taxonomic profiling

We used Phylosift to perform taxonomic profiling of the whole blood samples **(v 1.0.1, https://phylosift.wordpress.com/)**. Phylosift makes use of a set of protein coding genes found to be relatively universal (in nearly all bacterial and archaeal taxa) and having low variation in copy number between taxa. Homologs of these genes in new sequence data (e.g., the transcriptomes used here) are identified and then placed into a phylogenetic and taxonomic context by comparison to references from sequenced genomes. Phylosift was run as follows with default parameters: $*phylosift all --output=results input.fastq*

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For our replication study, we used MetaPhlAn for microbial profiling (Metagenomic Phylogenetic Analysis, v 1.7.7, http://huttenhower.sph.harvard.edu/metaphlan). The database of the microbial marker genes is provided with the tool. MetaPhlAn was run in 2 stages as follows, the first stage identifies the candidate microbial reads (i.e. reads hitting a marker) and the second stage profiles a metagenomes in terms of relative abundances.

```
1.    $metaphlan.py <fastq> <map> --input_type multifastq --bowtie2db bowtie2db/mpa -t
      reads_map --nproc 8 --bowtie2out
2.    $metaphlan.py --input_type blastout <bowtie2out.txt> -t rel_ab <tsv>
```

The reason for using MetaPHlAn rather that Phylosift was that due to differences in library preparation and sequence procedure, there were not sufficiently many reads matching the database of the marker genes curated by Phylosift for adequate microbial profiling.

Estimating Microbial Diversity

Microbial diversity within a sample was determined using the richness and alpha diversity indices. Richness was defined as the total number of distinct taxa in a sample. We use Inverse Simpson's formula incorporating richness and evenness components to compute alpha diversity $\frac{\mathbf{1}}{\boldsymbol{\lambda}} = \frac{\mathbf{1}}{\boldsymbol{\Sigma p_i^2}}$ (R package asbio, http://www.inside-r.org/packages/cran/asbio). To measure sample-to-sample dissimilarities between microbial communities we use Bray-Curtis beta diversity index accounting for both changes in the abundances of the shared taxa and account for taxa

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

uniquely present in one of the samples. Higher beta diversity indicates higher level of dissimilarity between microbial communities, providing a link between diversity at local scales (alpha diversity) and the diversity corresponding to total microbial richness of the subject group (gamma diversity (Koleff et al. 2003)). We calculate beta diversity per each combination of the samples resulting in a matrix of all pair-wise sample dissimilarities. Bray-Curtis beta diversity index is measured taxonomically as $\mathbf{1}-\frac{\mathbf{2}J}{A+B}$, where J is the sum of the lesser values for the shared taxa, A and B are the sum of the total values for all taxa for each sample respectively. Beta diversity was computed using 'vegan' R package (https://cran.r-project.org/web/packages/vegan/index.html). Total diversity of the groups is estimated as a function of the total number of taxa (gamma diversity). We use gamma diversity to estimate diversity of the group as well as total diversity of the study.

Statistical analysis of microbiome diversity

Alpha diversity

To test for differences in alpha diversity between disease groups, we fit the following analysis of covariance (ancova) model

$$alpha\_norm \sim Sex\ +\ Age\ +\ Technical\ covariates\ +\ Disease\ status$$

Where *Alpha_norm* = alpha values after inverse normal transformation, and *Age* = Individual's age at blood draw. Technical covariates include: *RIN, Batch (Plate_number)*, *Concentration*, and

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Flow cell lane, where *RIN* = RNA integrity value, a measure for RNA quality and *Concentration* = RNA concentration prior to normalization at the genotyping core.

The effect of disease status was estimated by first regressing out the effects of the included covariates. Adjustment for pairwise comparisons for all possible disease status pairs (6 comparisons) is performed using Bonferroni correction for multiple testing.

Because of the large age differences within the included groups, we tested for differences in alpha by sex or age directly within each group separately by correlating normalized alpha values with sex/age, using spearman rank correlation. We also repeated the above ancova analysis using only younger samples (with Age<47, the maximum age in the schizophrenia cohort, resulting in n=107 samples), and obtained similar results (i.e. *ANCOVA* P < 0.007 between schizophrenia and all other and no significant differences observed between BPD, ALS and Controls). To determine the relative effect size of alpha diversity on schizophrenia status, we fit the following logistic regression model:

$$SCZ \sim Sex + Age + Technical\ covariates + alpha\_norm$$

Where *SCZ* is a binary variable, which is coded as true if the sample belongs to the SCZ cohort.

Variation explained is by *alpha_norm* is measured by the reduction in R^2 comparing the full logistic regression model versus a reduced model with alpha_norm removed.

Beta diversity

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To assess difference in Beta diversity we fit a similar model as above, now correcting for Sex, Age and technical covariates for each individual:

$$beta\_norm \sim Sex1 + Sex2 + Age1 + Age2 + Technical\ covariates1 \\ + Technical\ covariates2 + Group$$

where *beta_norm* = beta values for each pair of individuals after inverse normal transformation, and *Group* contains set SCZ_SCZ (both individuals from SCZ), SCZ_Control (one SCZ, one control), Control_Control (both controls).

Adjustment for pairwise comparisons for all possible disease status pairs (3 comparisons) is performed using Bonferroni correction for multiple testing. We also determined a possible effect of alpha diversity on the above model by adding normalized values of alpha as a covariate to the model.

Reference-free microbiome analysis

We complement the reference-based taxonomic analysis with a reference independent analysis. We use EMDeBruijn (https://github.com/dkoslicki/EMDeBruijn) a reference-free approach able to quantify differences in microbiome composition between the samples. EMDeBruijn compresses the k-mer counts of two given samples onto de Bruijn graphs and then measures the minimal cost of transforming one of these graphs into the other (in terms of how many k-mers moved how far). This direct comparison of samples allows one to circumvent the

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

many issues involved with selecting a phylogenetic classification algorithm, choosing which training database to use, and deciding how to compare two classifications.

Other reference-free comparison metrics have been used before (such as treating k-mer frequencies as vectors in $\mathbb{R}^n$ and then using the Euclidean distance, Jensen-Shannon divergence, Kullback-Liebler divergence, cosine similarity, etc.). However, treating k-mer frequencies as vectors in $\mathbb{R}^n$ ignores the dependencies induced by the amount of overlap between two given k-mers. Instead of Euclidean space, EMDeBruijn considers k-mer frequencies as existing on an underlying de Bruijn graph, a structures that naturally takes into consideration such overlap-induced dependencies.

Fixing a k-mer size, we first form the undirected de Bruijn graph, with vertices given by k-mers, and an edge between two k-mers if the first (or last) k-1 nucleotides of one k-mer overlaps with the last (or first) k-1 nucleotides of the other k-mer. Let $d(\cdot,\cdot)$ represent the resulting graph distance. Then given two metagenomic samples, $S_1$and S_2, let the frequencies of k-mer be given by $freq_k(S_1)$ and $freq_k(S_2)$ respectively. These frequencies are thought of as weights on the vertices of the de Bruijn graph. Now to represent the transformation of one set of weights into the other, we use the term flow (or coupling) which is any real-valued matrix γ with rows and columns indexed by k-mers, such that the row sums equals $fre⍰_k(S_1)$ and the column sums equal $freq_k(S_2)$. A flow represents how much weight was moved where. There are infinitely many flows possible, but we choose the most efficient flow which is defined to be the one that minimizes the total cost (in terms of weight times distance). This leads to the

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

definition of the EMDeBruijn metric $EMD_k(S_1, S_1)$:

$$EMD_k(S_1, S_1) \coloneqq \min_{\gamma} \sum_{x,y \text{ kmers}} \gamma(x, y) * d(x, y)$$

Hence, the EMDeBruijn metric measures the minimal cost of transforming one sample's k-mer frequency vector into the other sample's k-mer frequency vector when allowable transformations are restricted to moves along edges of the de Bruijn graph.

To compute this quantity, we used the FastEMD implementation of the Earth Mover's Distance since the graph metric $d(\cdot,\cdot)$ is naturally thresholded. We found that a good trade-off between algorithmic run-time and effectiveness of the resulting metric was to use the k-mer size of k=6. To determine the variation explained by EMdeBruin principal components, we adopted a similar approach as described above and fit the following logistic regression model:

$$SCZ \sim Sex\ +\ Age\ +\ Technical\ covariates\ +\ PC1 + PC2 + PC3$$

where *PCi* denotes the ith EMdeBruin principal component.

To determine overlap between the results from Phylosift and EMdeBruin, we correlated principal components of EMdeBruin PC1 and Phylosift PC1 by spearman rank correlation, including all samples.

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Correlation of microbial diversity with genetic risk for schizophrenia

To determine a correlation between genetic risk for schizophrenia and alpha diversity, we compared alpha diversity to the polygenic risk score for schizophrenia. The polygenic risk score represents the cumulative genetic load of disease risk alleles, and is defined as the sum of trait-associated alleles across many genetic loci, weighted by effect sizes estimated from a genome-wide association study. We based our scores on the most recent genome wide association study (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014) with our samples removed (Ripke et al. 2013a), and used a P-value cut-off of $P < 0.05$. For a total of 32 Schizophrenia cases, we had both polygenic risk score and alpha diversity measures available and we performed a spearman rank correlation. Similar results are obtained if we use different P-value cutoffs to determine the polygenic risk score.

Analyses of the replication sample were performed in an analogous fashion to the methods described above. Statistical analysis was performed in R. We represent data as mean ± standard deviation. Boxplots are represented with the first and third quantiles.

Estimation of DNA methylation-derived cell proportions in whole blood

DNA methylation profiles of heterogeneous tissue types reflect variability in underlying cellular composition (29, 30). Recent studies, using flow-sorted cell populations, identified CpG sites discriminatory for distinct cell populations and developed sophisticated methods to estimate blood cell proportions from DNA methylation data derived from whole blood (28-30). We use

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

these methods to investigate a potential link between microbial diversity and the immune system.

In a control cohort of 220 individuals blood-based genome-wide methylation data was collected using the Infinium HumanMethylation450 BeadChip. We used the epigenetic clock software (Horvath 2013) with normalization to estimate cell abundance measures. Briefly, this software uses Houseman's estimation method (Houseman et al. 2012; Aryee et al. 2014) to estimate monocytes, granulocytes, CD8 T, CD4 T, natural killer, and B cells. In addition, it predicts abundance measures for plasmablasts (i.e. immature plasma cells), CD8.naive, CD4.naive, and CD8pCD28nCD45RAn cells (i.e. differentiated CD8 T cells) based on a penalized elastic net regression model (Horvath 2013; Horvath and Levine 2015).

Quality control of the DNA methylation data was performed as follows**.** CpG sites with bead counts less than 5 or a detection p-value greater than 0.01 in more than 5% of samples were removed using the pfilter function in the wateRmelon package in R. In addition, sample having more than 5% of CpG sites with a detection p-value greater than 0.01 or having gender discrepancies were excluded from further analyses. Next, we removed CpG sites with probes containing known SNPs (EUR, MAF > 0.01) and probes that are cross-reactive, i.e. non-specific (Chen et al. 2013; Price et al. 2013). Data was background corrected using the danen function in R (wateRmelon package) and beta values were extracted for further analyses.

We investigated the relationship between microbiome diversity and the immune system as follows. From a cohort of n=220 controls for which methylation-derived cell proportions were

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

available, we first obtained residuals for each cell-type using the following model:

$$Proportion\_cell\_type \sim Sex\ +\ Age\ +\ Beadchip\ +\ Beadchip\ position\ +\ dataset.$$

In addition, we used residuals from the above-described regression on alpha diversity using our full replication cohort. Using all samples with both alpha levels and methylation-based cell abundance measures available (a total of n=65), we next fitted a linear regression model with alpha diversity residuals as response variable and all blood cell proportion residuals as independent variables. Each independent variable was analyzed as it was put in the model last to account for correlations among cell proportions. We thus model the relationship between alpha diversity and individual cell types while adjusting for all other cell types.

Data access:

The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus (Edgar et al. 2002) and are accessible through GEO Series accession number GSE80974 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE80974).

Acknowledgments:

This work is supported by NIH/NIMH R01 5R01MH090553 (BP samples), 5R01NS058980 (ALS samples) and R01MH078075 (SCZ, Control samples), R21MH098035 (replication sample) awarded to R.A.O. L.M.O.L was financially supported by the National Institute of Neurological Disorders And Stroke of the National Institutes of Health under Award Number T32NS048004.

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The authors thank Dr. Jonathan Eisen for helpful discussions and insights throughout the course of this project. We thank the study subjects for their participation.

Disclosure declaration:

The authors have no conflicts of interest to declare.

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References and Notes

Abu-Shanab A, Quigley EM. 2010. The role of the gut microbiota in nonalcoholic fatty liver disease. *Nature reviews Gastroenterology & hepatology* **7**: 691-701.

Amar J, Serino M, Lange C, Chabo C, Iacovoni J, Mondot S, Lepage P, Klopp C, Mariette J, Bouchez O et al. 2011. Involvement of tissue bacteria in the onset of diabetes in humans: evidence for a concept. *Diabetologia* **54**: 3055-3061.

Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. 2014. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**: 1363-1369.

Belkaid Y, Hand TW. 2014. Role of the microbiota in immunity and inflammation. *Cell* **157**: 121-141.

Ben-Amor K, Heilig H, Smidt H, Vaughan EE, Abee T, de Vos WM. 2005. Genetic diversity of viable, injured, and dead fecal bacteria assessed by fluorescence-activated cell sorting and 16S rRNA gene analysis. *Applied and environmental microbiology* **71**: 4679-4689.

Buizer-Voskamp JE, Muntjewerff JW, Genetic R, Outcome in Psychosis Consortium M, Strengman E, Sabatti C, Stefansson H, Vorstman JA, Ophoff RA. 2011. Genome-wide analysis shows increased frequency of copy number variation deletions in Dutch schizophrenia patients. *Biol Psychiatry* **70**: 655-662.

Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, ReproGen C, Psychiatric Genomics C, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control C, Duncan L et al. 2015. An atlas of genetic correlations across human diseases and traits. *Nature genetics* **47**: 1236-1241.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC bioinformatics* **10**: 421.

Castro-Nallar E, Bendall ML, Perez-Losada M, Sabuncyan S, Severance EG, Dickerson FB, Schroeder JR, Yolken RH, Crandall KA. 2015. Composition, taxonomy and functional diversity of the oropharynx microbiome in individuals with schizophrenia and controls. *PeerJ* **3**: e1140.

Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R. 2013. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**: 203-209.

Cho I, Blaser MJ. 2012. The human microbiome: at the interface of health and disease. *Nature reviews Genetics* **13**: 260-270.

Cox TF, Cox MA. 2000. *Multidimensional scaling*. CRC Press.

Darling AE, Jospin G, Lowe E, Matsen FAt, Bik HM, Eisen JA. 2014. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**: e243.

de Vos WM, de Vos EA. 2012. Role of the intestinal microbiome in health and disease: from correlation to causation. *Nutrition reviews* **70 Suppl 1**: S45-56.

Drennan MR. 1942. What is “Sterile Blood”? *British Medical Journal* **2**: 526-526.

Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**: 207-210.

Erny D, Hrabe de Angelis AL, Jaitin D, Wieghofer P, Staszewski O, David E, Keren-Shaul H, Mahlakoiv T, Jakobshagen K, Buch T et al. 2015. Host microbiota constantly control maturation and function of microglia in the CNS. *Nat Neurosci* **18**: 965-977.

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Foster JA, McVey Neufeld KA. 2013. Gut-brain axis: how the microbiome influences anxiety and depression. *Trends Neurosci* **36**: 305-312.

Greenblum S, Turnbaugh PJ, Borenstein E. 2012. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc Natl Acad Sci U S A* **109**: 594-599.

Horvath S. 2013. DNA methylation age of human tissues and cell types. *Genome Biol* **14**: R115.

Horvath S, Levine AJ. 2015. HIV-1 Infection Accelerates Age According to the Epigenetic Clock. *J Infect Dis* doi:10.1093/infdis/jiv277.

Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. 2012. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* **13**: 86.

Hsiao EY, McBride SW, Hsien S, Sharon G, Hyde ER, McCue T, Codelli JA, Chow J, Reisman SE, Petrosino JF et al. 2013. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell* **155**: 1451-1463.

Human Microbiome Project C. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* **486**: 207-214.

Jorth P, Turner KH, Gumus P, Nizam N, Buduneli N, Whiteley M. 2014. Metatranscriptomics of the human oral microbiome during health and disease. *MBio* **5**: e01012-01014.

Jost L. 2007. Partitioning diversity into independent alpha and beta components. *Ecology* **88**: 2427-2439.

Koch S, Larbi A, Derhovanessian E, Ozcelik D, Naumova E, Pawelec G. 2008. Multiparameter flow cytometric analysis of CD4 and CD8 T cell subsets in young and old people. *Immun Ageing* **5**: 6.

Koleff P, Gaston KJ, Lennon JJ. 2003. Measuring beta diversity for presence–absence data. *Journal of Animal Ecology* **72**: 367-382.

Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RG, Getz G, Meyerson M. 2011. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat Biotechnol* **29**: 393-396.

Loohuis LM, Vorstman JA, Ori AP, Staats KA, Wang T, Richards AL, Leonenko G, Walters JT, DeYoung J, Consortium G et al. 2015. Genome-wide burden of deleterious coding variants increased in schizophrenia. *Nat Commun* **6**: 7501.

McLaughlin RW, Vali H, Lau PC, Palfree RG, De Ciccio A, Sirois M, Ahmad D, Villemur R, Desrosiers M, Chan EC. 2002. Are there naturally occurring pleomorphic bacteria in the blood of healthy humans? *J Clin Microbiol* **40**: 4771-4775.

Nikkari S, McLaughlin IJ, Bi W, Dodge DE, Relman DA. 2001. Does blood of healthy subjects contain bacterial ribosomal DNA? *J Clin Microbiol* **39**: 1956-1959.

Paisse S, Valle C, Servant F, Courtney M, Burcelin R, Amar J, Lelouvier B. 2016. Comprehensive description of blood microbiome from healthy donors assessed by 16S targeted metagenomic sequencing. *Transfusion* doi:10.1111/trf.13477.

Potgieter M, Bester J, Kell DB, Pretorius E. 2015. The dormant blood microbiome in chronic, inflammatory diseases. *FEMS Microbiol Rev* **39**: 567-591.

Price ME, Cotton AM, Lam LL, Farre P, Emberly E, Brown CJ, Robinson WP, Kobor MS. 2013. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin* **6**: 4.

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ripke S O'Dushlaine C Chambert K Moran JL Kahler AK Akterin S Bergen SE Collins AL Crowley JJ Fromer M et al. 2013a. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* **45**: 1150-1159.

Ripke S O'Dushlaine C Chambert K Moran JL Kahler AK Akterin S Bergen SE Collins AL Crowley JJ Fromer M et al. 2013b. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* **45**: 1150-1159.

Santpere G, Darre F, Blanco S, Alcami A, Villoslada P, Mar Alba M, Navarro A. 2014. Genome-wide analysis of wild-type Epstein-Barr virus genomes derived from healthy individuals of the 1,000 Genomes Project. *Genome biology and evolution* **6**: 846-860.

Sato J, Kanazawa A, Ikeda F, Yoshihara T, Goto H, Abe H, Komiya K, Kawaguchi M, Shimizu T, Ogihara T et al. 2014. Gut dysbiosis and detection of "live gut bacteria" in blood of Japanese patients with type 2 diabetes. *Diabetes Care* **37**: 2343-2350.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**: 421--427

.

Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods* **9**: 811-814.

Simpson EH. 1949. Measurement of diversity. *Nature*.

Spadoni I, Zagato E, Bertocchi A, Paolinelli R, Hot E, Di Sabatino A, Caprioli F, Bottiglieri L, Oldani A, Viale G et al. 2015. A gut-vascular barrier controls the systemic dissemination of bacteria. *Science* **350**: 830-834.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP et al. 2009. A core gut microbiome in obese and lean twins. *Nature* **457**: 480-484.

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**: 1027-1031.

Weiss S, Amir A, Hyde ER, Metcalf JL, Song SJ, Knight R. 2014. Tracking down the sources of experimental contamination in microbiome studies. *Genome biology* **15**: 564.

Whittaker RH. 1972. Evolution and measurement of species diversity. *Taxon*: 213-251.

bioRxiv preprint doi: https://doi.org/10.1101/057570; this version posted June 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.