Detection of stable community structures within gut microbiota co-occurrence networks from different human populations

Matthew A Jackson[1], Marc Jan Bonder[2], Zhana Kuncheva[3], Jonas Zierer[1,4], Jingyuan Fu[2,5], Alexander Kurilshikov[2], Cisca Wijmenga[2,6], Alexandra Zhernakova[2], Jordana T Bell[1], Tim D Spector[1], and Claire J Steves[1*]

[1]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK
[2]University Medical Center Groningen, Department of Genetics, Groningen, University of Groningen, The Netherlands
[3]Department of Mathematics, Imperial College London, London, UK
[4]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Neuherberg, Germany
[5]University Medical Center Groningen, Department of Pediatrics, University of Groningen, Groningen, The Netherlands
[6]K.G. Jebsen Coeliac Disease Research Centre, Department of Immunology, University of Oslo, Norway

Corresponding author:
Claire J Steves[1]

Email address: claire.j.steves@kcl.ac.uk

ABSTRACT

Microbes in the gut microbiome form sub-communities based on shared niche specialisations and specific interactions between individual taxa. The inter-microbial relationships that define these communities can be inferred from the co-occurrence of taxa across multiple samples. Here, we present an approach to identify comparable communities within different gut microbiota co-occurrence networks, and demonstrate its use by comparing the gut microbiota community structures of three geographically diverse populations. We combine gut microbiota profiles from 2764 British, 1023 Dutch, and 639 Israeli individuals, derive co-occurrence networks between their operational taxonomic units, and detect comparable communities within them. Comparing populations we find that community structure is significantly more similar between datasets than expected by chance. Mapping communities across the datasets, we also show that communities can have similar associations to host phenotypes in different populations. This study shows that the community structure within the gut microbiota is stable across populations, and describes a novel approach that facilitates comparative community-centric microbiome analyses.

INTRODUCTION

The gut microbiome is a complex bacterial community, with its structure determined by many factors including the interactions between its members. Bacteria can interact in numerous ways - in either an actively targeted or passive manner, which can result in beneficial, neutral, or detrimental effects for the parties involved (Faust and Raes, 2012). Given the increasing evidence of the importance of the gut microbiome in human health, it is necessary to understand the inter-microbial effects underlying its composition.

Gut microbiota are frequently profiled using marker gene sequencing. Sequencing reads from amplicons of the selected marker are typically collapsed to operational taxonomic units (OTUs), analytical units used to approximate taxonomic abundances (Schloss et al., 2009; Navas-Molina et al., 2013). One

approach to infer interactions between bacteria in the gut microbiota is to quantify the co-occurrence of OTUs across multiple samples (Faust and Raes, 2012). High correlation between OTUs can reflect the interactions between their source bacteria and similarities/differences in their responses to environmental conditions (Lozupone et al., 2012; Faust and Raes, 2012). However, OTU counts are relative to the sequencing depth of a sample, which introduces inherent correlations to the data (Friedman and Alm, 2012). As a result, several specialised approaches have been developed to estimate correlations from microbiota data (Faust et al., 2012; Friedman and Alm, 2012; Deng et al., 2012; Gevers et al., 2014; Fang et al., 2015; Kurtz et al., 2015). Whilst these have seen use within the research community (Gevers et al., 2014; Goodrich et al., 2014; Tong et al., 2013; McHardy et al., 2013), correlation metrics for microbiome studies have only recently been compared systematically by Weiss and Van Treuren et al. who found that using an ensemble of metrics can improve the precision of co-occurrence detection (Weiss et al., 2016).

Given co-occurrence between OTUs it is possible to generate networks of their inferred interactions. Within this we would not expect all bacteria to interact but rather that subsets of taxa are more likely to interact within one another forming identifiable and distinct interacting groups. We will hence refer to these sub-groups of co-occurring microbes as communities. The formation of such communities can be driven by factors such as cross-species metabolism and geospatial environmental variation (Faust and Raes, 2012; Levy and Borenstein, 2014; Lozupone et al., 2012). Previous studies have identified communities within microbial co-occurrence networks, often using the WGCNA (weighted gene co-expression network analysis) method developed to identify modules in gene co-expression networks (Lozupone et al., 2012; Tong et al., 2013; Jackson et al., 2016b; Duran-Pinedo et al., 2011; Langfelder and Horvath, 2008). However, whilst they may be more biologically accurate, such approaches allow OTUs to have weighted contributions to multiple communities, which complicates comparison and mapping of equivalent communities between networks. Using a community detection algorithm that assigns OTUs to single communities simplifies such analyses enabling comparison across datasets. This was demonstrated by the use of a modularity maximisation approach to detect and compare community structures between gut microbiome networks of irritable bowel syndrome patients and healthy controls (Baldassano and Bassett, 2016).

Here, we use the ensemble approach outlined in the recent methods comparison by Weiss and Van Treuren et al. to quantify co-occurrence between gut microbiota, and apply a modularity maximisation approach to detect communities in the resultant networks. In this approach there are necessarily steps in which thresholds for edge inclusion and parameters for community detection must be selected. We describe biologically motivated and data driven approaches to inform these decisions. This method produces communities that can be compared and mapped across datasets. We apply this method to data from human cohorts from the UK, Netherlands, and Israel to establish whether gut microbiota form similar communities in different human populations. We find that OTUs form similar community structures across all three, and that these communities have similar associations with the health-related host factors of age and body mass index (BMI) in their respective populations. This study also provides a framework for future studies aiming to identify and, most importantly, replicate community level effects in microbiota studies.

MATERIALS AND METHODS

Data aggregation

Given our objective to compare network community structures across datasets (Figure 1A), we required OTUs that would be comparable between them. To this end, we carried out clustering of sequences across combined data from multiple sources. To maximise sequencing similarity, we selected two datasets with experimental approaches best matching those of the gut microbiota profiles obtained for TwinsUK, which used 16S rRNA gene sequencing of faecal samples. The LLDEEP and Israeli-PN datasets were selected as they carried out gut microbiota profiling by amplifying the V4 region of the 16S rRNA gene using the same PCR primers, and used paired-end sequencing on the Illumina MiSeq platform with read lengths sufficient to capture the whole V4 region. Notable differences between studies include faecal sampling and DNA extraction techniques. Both TwinsUK and LLDEEP utilised aliquots from faecal samples stored at -80°C, whilst the Israeli-PN study utilised a mixture of faecal swabs stored at -80°C and OMNIgene-GUT stool collection kits stored at -20°C. All three studies used both chemical and mechanical lysis in DNA extraction but employed different protocols: TwinsUK utilised the MoBio PowerSoil HTP extraction kit, the LLDEEP cohort utilised the Qiagen AllPrep kit, and the Israeli-PN DNA was extracted using the
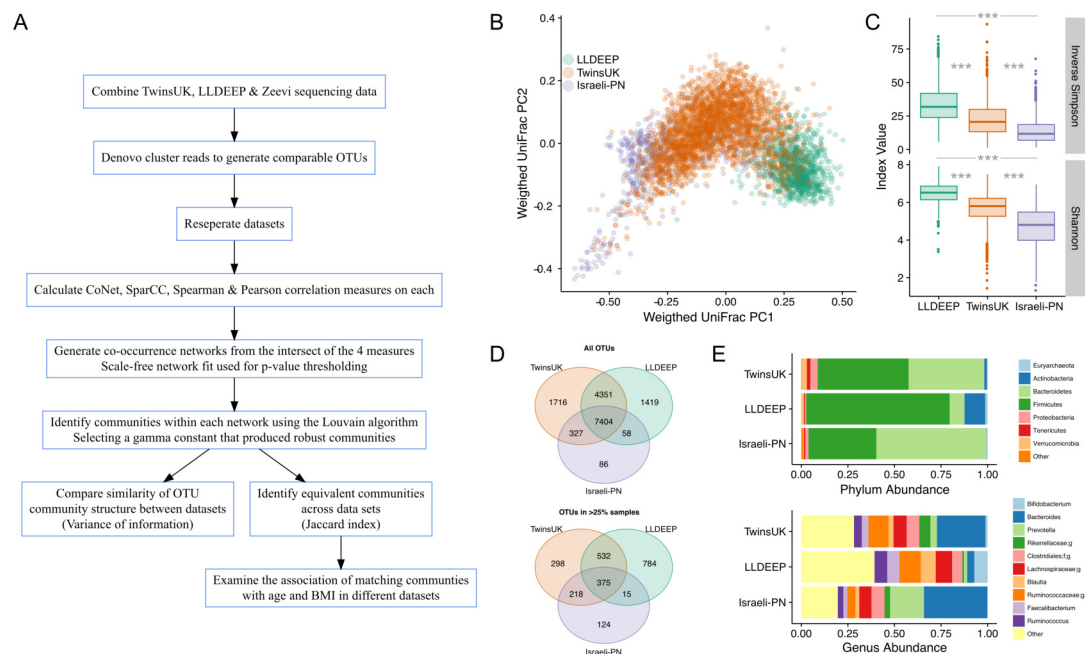
Figure 1. An overview of the study and the data used in analyses. A) An outline of the study design. B) Plot of the first two components of PCoA from weighted UniFrac distance measures between samples in the study, coloured by dataset, shows there is some separation but significant overlap in microbiota compositions by study. C) Comparison of alpha diversity measures in all three datasets. There was a significant difference in all pair-wise comparisons for both measures (Mann-Whitney U p<0.0001). D) Venn diagrams showing the number of OTUs shared across datasets. The top includes all OTUs, the lower diagram only those found in at least 25% of samples in each dataset. E) Comparison of the mean relative abundances of taxonomies at the phylum and genus level across the complete table for each dataset. Phyla at less than 1% abundance and genera at less than 5% abundance in all sets are collapsed as Other. In genera f; and g; represent unassigned family and genus names in the Greengenes reference.

99 MoBio PowerMag Soil DNA extraction kit.

100   Extraction of DNA from faecal samples, amplification and sequencing of the V4 region of the 16S
101 rRNA gene, and demultiplexing of sequencing reads has previously been described for the TwinsUK
102 cohort (Goodrich et al., 2014). The paired-end demultiplexed reads were joined using join_paired_ends
103 within QIIME with an overlap of at least 200nt to form single reads covering the full V4 region (Navas-
104 Molina et al., 2013). DNA extraction and 16S rRNA gene sequencing within the LLDEEP samples has
105 been described in detail previously (Fu et al., 2015). Data from the LLDEEP cohort was provided in
106 a similar format having used custom scripts to merge the paired end data to full length reads covering
107 the V4 region and split data by individual (Gevers et al., 2014; Fu et al., 2015). Ethics approval for
108 the TwinsUK study was given by the NRES Committee London - Westminster (REC Reference No. :
109 EC04/015), and the Lifelines-DEEP study was approved by the institutional ethics review boards of the
110 University Medical Center Groningen (ref. M12.113965).

111   Raw data accompanying the Israeli-PN publication was downloaded from the European Nucleotide
112 Archive (ENA) (Accession:PRJEB11532)(Zeevi et al., 2015). This was processed similarly to the Twin-
113 sUK data and demultiplexed using published barcode mappings. Samples not listed in the accompanying
114 metadata or with ambiguous read identifiers were removed.

OTU clustering, table filtering, and diversity analyses

116 The cleaned, joined, and demultiplexed data was concatenated to produce one sequencing file covering
117 all three studies. After quality control, 4426 samples (2764 TwinsUK, 1023 LLDEEP, 639 Israeli-PN)
118 were included in the analysis. The complete sequencing from all three sets contained 381,767,528 reads.
119 These were dereplicated, removing reads occurring only once, resulting in 5,728,288 unique reads. We

3/21

chose to use a *de novo* approach to cluster these to OTUs, as *de novo* OTU clustering is not influenced by the reference database used, can capture novel diversity, and can produce more accurate clustering of OTUs than reference-based approaches (Westcott and Schloss, 2015; Jackson et al., 2016a). *De novo* clustering was carried out using the cluster_fast command in the VSEARCH package with a 97% similarity threshold (Rognes et al., 2016). The resultant 94,070 representative sequences were filtered to remove chimeric reads using the uchime_denovo command (within VSEARCH), producing a final set of 17,123 representative sequences. These were used to generate OTU counts across all samples using the VSEARCH usearch_global command (Rognes et al., 2016). Post-processing, average read counts were 82,695±745 for TwinsUK, 49,962±964 for LLDEEP, and 130,378±5,534 for Israeli-PN (mean±SEM).

A phylogenetic tree was generated from the representative sequences using the default parameters of the make_phylogeny command in QIIME (Navas-Molina et al., 2013). Taxonomy of OTUs was assigned by matching representative sequences against the Greengenes v13_8 database using the default parameters of the assign_taxonomy command in QIIME (Navas-Molina et al., 2013). OTUs occurring in only one sample, and samples with less than 10,000 OTU counts were removed. Weighted and unweighted UniFrac beta diversity measures and subsequent principle co-ordinates analysis of them was carried out using the beta_diversity_through_plots script in QIIME (Navas-Molina et al., 2013). The combined OTU table was then split by dataset. For the purposes of alpha diversity calculations, the raw counts tables were rarefied to a depth of 10,000 reads. For each sample, Shannon and inverse Simpson diversity indices were calculated as the mean across ten rarefactions. Significant differences in alpha diversity between datasets were assessed using Mann-Whitney U-tests.

Calculating co-occurrence measures

Co-occurrence calculation was carried out on each dataset independently. Sub-tables were generated from raw (unrarefied) OTU tables that only contained OTUs found in at least 25% of the samples (Jackson et al., 2016b). All co-occurrence measures were calculated within these subsets as they are less sparse and hence more amenable to correlation measures (Weiss et al., 2016). The mean diversity was assessed on rarefied versions of the subset tables using the inverse Simpson index as for the full tables. OTU sparsity was assessed from the unrarefied table using the biom summarize_table command. Following the recommendations of Weiss et al. (2016), we then used these estimates of the mean inverse Simpson index (TwinsUK=20.2, LLDEEP=29.0, and Israeli-PN=13.1) and OTU table sparsity (0.49 in all three) to select an ensemble approach to co-occurrence detection that combines four different correlation measures: CoNet, SparCC, Pearson's, and Spearman's.

CoNet

CoNet is itself an ensemble approach. The package allows a range of co-occurrence measures to be combined with several options for how to combine the weighting of edges and their p-values. CoNet addresses compositionality within the data using the ReBoot procedure, which involves permutation followed by renormalisation of data. Calculating co-occurrence within these renormalized data allows assessment of the levels of correlation expected simply as a result of the compositionality within the data (Faust and Raes, 2012). For this study we used four measures of co-occurrence within CoNet: Spearman's and Pearson's correlations and Kullback-Leibler and Bray-Curtis distance measures (Faust and Raes, 2012). Initial correlation thresholds were selected for each of these measures that produced 2000 positive and 2000 negative edges concordant across the four metrics (Weiss et al., 2016); 1000 permutations were then used for renormalisation to account for compositionality, and bootstrapping to identify edge p-values for each metric. The Simes method was selected to merge p-values across edges by keeping the minimum. Final p-values were adjusted for multiple testing using the Benjamini-Hochberg FDR approach.

SparCC

SparCC was developed to calculate correlations between OTU abundances in microbiome data whilst accounting for their inherent sparsity and compositionality (Friedman and Alm, 2012). It uses the centered log-ratio transformation to address data compositionality. SparCC was used with default parameters to calculate correlations from the raw count OTU tables (Friedman and Alm, 2012). The MakeBoostraps command was used to generate 100 bootstrapped tables, which were in turn used to calculate SparCC correlations. The bootstrapped correlations were then used with the PseudoPvals command to generate two-tailed p-values for the SparCC correlations from the true table.

¹⁷² *Pearson's and Spearman's Correlation Coefficients*

¹⁷³ Pearson's and Spearman's correlation metrics do not take data compositionality into account, but were
¹⁷⁴ included to follow the approach outline by Weiss and Van Treuren et al.. Both measures were calculated
¹⁷⁵ using the relative abundance tables. The rcorr function from the Hmisc R package was used to calculate
¹⁷⁶ correlations and generate two-tailed p-values pairwise between all OTUs for both Pearson's and Spear-
¹⁷⁷ man's measures (Harrell Jr and Dupont, 2008). P-values were adjusted for multiple testing using the
¹⁷⁸ Bonferroni method in R, again following the approach of Weiss et al. (2016).
¹⁷⁹
¹⁸⁰ The outputs of all four co-occurrence approaches were converted into simplified unweighted edge
¹⁸¹ tables detailing the direction of association (1 or -1) and bootstrapped/adjusted p-values.

P-value thresholding to generate co-occurrence networks

¹⁸³ Intersected networks were generated by combining the edge tables from the CoNet, SparCC, Pearson's,
¹⁸⁴ and Spearman's methods. This was done independently for each dataset. Edges were retained in the final
¹⁸⁵ network if the direction of co-occurrence matched, and the edge p-values were below a given threshold
¹⁸⁶ in all four methods. This was carried out at multiple different p-value thresholds (0.05, then ranging in
¹⁸⁷ powers of ten from 0.01 to 10^{-8}), to generate multiple intersect networks for each dataset with a gradient
¹⁸⁸ of stringency for edge inclusion. We then aimed to determine which was the most appropriate threshold to
¹⁸⁹ use to generate the final networks. Rather than make an entirely arbitrary selection, we chose to use fit to
¹⁹⁰ a scale-free network as a biologically motivated method to identify the optimum p-value threshold to use.
¹⁹¹ A scale-free network has a node degree distribution that follows a power law, i.e. there are few highly
¹⁹² connected nodes and many more less connected nodes. This distribution has been observed in several
¹⁹³ biological networks, including microbiota co-occurrence networks (Jeong et al., 2000; Albert, 2005;
¹⁹⁴ Zhang and Horvath, 2005; Faust et al., 2012; Tong et al., 2013). It is also frequently used as a threshold
¹⁹⁵ for edge inclusion in the widely used WGCNA package (Langfelder and Horvath, 2008). Fit to a scale
¹⁹⁶ free network was calculated by first extracting the degree distribution of a network using igraph (Csardi
¹⁹⁷ and Nepusz, 2006). The fit of this distribution was then assessed using the scaleFreeFitIndex function
¹⁹⁸ from the WGCNA package in R (Langfelder and Horvath, 2008). This provides the R^2 of the fit to a
¹⁹⁹ scale-free model, which the creators of WGCNA suggest should be >0.8, and the slope of the fit, which
²⁰⁰ they suggest should be close to -2 to indicate a good fit. The optimum p-value threshold was selected
²⁰¹ based on these criteria across all three datasets (see Results) and the resultant intersect network at this
²⁰² threshold was used in all further analyses. Visualisation and generation of descriptive statistics from the
²⁰³ final networks was carried out using Gephi (Bastian et al., 2009).

Detecting communities within co-occurrence networks

²⁰⁵ Between OTU adjacency matrices were generated that represented the final intersect networks for each
²⁰⁶ dataset. Negative correlations were removed (considered zeroes in the adjacency matrix) to generate
²⁰⁷ unsigned networks as they represented a considerably small proportion of edges (<1% in all datasets).
²⁰⁸ Community detection was carried out on each dataset's network independently using the genLouvain 2.0
²⁰⁹ package within MATLAB (Mucha et al., 2010), which implements the Louvain approach to modularity
²¹⁰ maximisation (Blondel et al., 2008). This defines community partitions by assigning nodes to unique
²¹¹ communities, then iteratively combining neighbouring nodes into communities if it results in an increase in
²¹² modularity across the whole network. Modularity is a measure of the number of edges within communities
²¹³ relative to between communities and a higher value represents better community definition (Newman and
²¹⁴ Girvan, 2004).
²¹⁵ The genLouvain algorithm includes a γ parameter, which controls the size and number of detected
²¹⁶ communities (Mucha et al., 2010). A smaller gamma value promotes the detection of a small number
²¹⁷ of larger communities, while larger gamma values promote the detection of a high number of smaller
²¹⁸ communities. To find an optimal value for the γ parameter, we carried out community detection using a
²¹⁹ range of γ values for each dataset (0.1-1, increments of 0.01), and assessed the stability and statistical
²²⁰ significance of the resultant communities.

²²¹ *Stability*

²²² To assess the stability of community definitions at each γ, we carried out community detection 25 times
²²³ on the real network followed by pairwise comparisons of the similarity of community clustering between
²²⁴ the runs. To assess similarity between two community groupings we used the normalised variation of

information (variation of information divided by ln(number of nodes in the network)) as a measure of similarity between assignments (Ronhovde and Nussinov, 2009). A high value for variation of information means OTUs are grouped more differently in the two partitions compared, whereas a value of zero means the two partitions are identical. In figures and text, we report 1-normalised variation of information so a higher value represents more similar community structure.

Statistical Significance

To assess the statistical significance of the community definitions, we also carried out community detection at each γ on 100 randomised networks with nodes following the same degree distribution as the real network (generated using the randomGraphFromDegreeSequence command in the Octave networks toolbox package). We then compared the mean modularity of the 25 runs on the real network to the 100 randomised networks. A process that has previously been proposed to assess the significance of community structures across different resolutions (Lambiotte, 2010; Traag et al., 2013).

From observation of both the variation of information (stability) and modularity (statistical significance) results we then selected a suitable value for γ that produced both stable and significant community groupings (see Results). Once a suitable value for γ was identified, community detection was repeated 100 times at this value, and the community definitions from the run producing the highest modularity were retained as the final community definitions.

Community properties and host associations in TwinsUK

The relative abundance of each community in a sample was found by summing the read counts of its constituent OTUs and dividing by the total number of reads observed in the sample. Association between the mean abundance of an OTU in a dataset and the number of OTUs in its parent community was assessed using Spearman's correlation. Taxonomy within communities was investigated by counting the number of OTUs assigned to different taxa at each taxonomic level. The mean identity between the representative sequences of the OTUs in each community was assessed for each dataset. All pairwise comparisons between the OTUs within a community were carried out using global alignments performed using BLAST within the pairwise2 command of Biopython, with a score of 1 for a match, -1 for a mismatch, -2 for opening a gap, and -1 for extending a gap (the defaults used by the Mothur align.seqs command) (Cock et al., 2009; Schloss et al., 2009).

The heritability of TwinsUK communities was estimated using the mets package in R. Log transformed relative abundances were used for each community. For each community ACE, CE, E, and AE models were fit using data from complete twin pairs. Co-variates included age, BMI, sample collection method, sex, and sequencing depth. For each community, the best fitting model was determined as that with the lowest Akaike information criterion.

Association analyses between community abundances and BMI and age was carried out using log transformed relative abundances of the communities (log10(relative abundance + 10^{-6})). These phenotypes were selected as they were also available in both the LLDEEP and Israeli-PN datasets. To investigate associations linear models were fitted for each community using the lm function in R. Community abundance was the dependent variable with BMI, age, gender, and sequencing depth as independent variables, as these were the maximal set available across all datasets. The coefficient and significance of associations were extracted for BMI and age. P-values were then adjusted for multiple testing using the FDR method in R.

Comparison of community structure between populations

Overall community structure comparisons

We carried out pairwise comparisons between datasets to assess the similarity of overall community structures between the networks. In each comparison, the two networks were subset to just the OTUs shared by both, and the normalised variation of information between their community structures calculated. To find the variation that would be expected by chance we generated randomised community sets for each network by shuffling OTU labels. Each randomised comparison therefore shared the same number OTUs between the two real networks with the same community sizes, but without the biological basis for the linkage of OTUs. We then carried out pairwise comparisons of the variation of information between the randomised communities. Shuffling and comparisons were repeated 1,000 times. The highest score

Table 1. Participant summary statistics for the datasets considered in this study.

dataset	BMI (Mean±SD)	Age (Mean±SD)	Sex (M/F/Unknown)	No. Samples
TwinsUK	26.1±4.8	59.5±12.3	308 / 2456 / 0	2764
LLDEEP	25.3±4.2	45.3±13.7	445 / 578 / 0	1023
Israeli-PN	26.4±5	43.5±12.9	376 / 251 / 12	639

277 observed (1- normalised variation of information) in any pairwise comparison, in any permutation was
278 0.41 (mean=0.34, SD=0.014).
279   To determine the robustness of the community detection approach to differences in the selected
280 parameters, overall community structure was also compared between networks created using both higher
281 and lower p-value thresholds and γ values. In each case, networks and community definitions were
282 generated as previously. Normalised variation of information was then used to compare the networks'
283 community structures as described above. Comparisons between the three datasets were carried out within
284 each p-value and γ pairing. A wider comparison was also carried out between all possible dataset, p-value,
285 and γ value combinations. In the case of the complete pairwise comparison, permutations to assess null
286 variation of information were only generated 100 times (in place of 1000) due to the high total number of
287 comparisons involved.

288 ***Mapping of individual communities between datasets***
289 To map communities between networks we carried out pairwise comparisons between all the communities
290 in all three networks. We used the Jaccard index to quantify the number of OTUs shared between the two
291 communities relative to the number not shared in each comparison. Matches were considered positive
292 with scores >0.25 (range 0-1, no shared OTUs - complete overlap). This was selected as above this
293 threshold there were no instances of multiple mapping of communities between datasets.
294   Community-types were defined where matches could be found linking the communities in all three
295 datasets. These were labelled with colour names using the standardColors function from the WGCNA
296 package in R. The log transformed abundances for the 14 community-types that could be mapped across
297 all three datasets were generated for the LLDEEP and Israeli-PN datasets as for TwinsUK. These were
298 analysed for associations with age and BMI using linear regressions as for TwinsUK.

RESULTS

Integrating microbiota data from different populations

301 The aim of this study was to compare gut microbiota community structures across populations (study
302 overview in Figure 1A). To this end comparable OTUs were generated by combining gut microbiota
303 sequencing data from three geographically diverse populations. These were the British TwinsUK and
304 Dutch Lifelines-DEEP (LLDEEP) cohorts and data from an Israeli personalised nutrition study (Israeli-
305 PN) (Table 1) (Fu et al., 2015; Zeevi et al., 2015). Principle coordinates analysis of overall microbiota
306 composition between samples found that whilst there was some grouping by dataset, there was also
307 significant overlap between them (Figure 1B). Comparison of median alpha diversities between the
308 populations found significant differences between all three (Figure 1C).
309   Across the 15361 OTUs detected in the complete data 48% were shared across all three datasets,
310 with 79% being found in at least two datasets (Figure 1D). There was also considerable overlap between
311 TwinsUK and LLDEEP of OTUs not found in the Israeli-PN data. Similar patterns were observed when
312 considering more abundant OTUs (found in >25% of the respective populations). Although LLDEEP and
313 TwinsUK shared more OTUs, examining the mean taxonomic distributions at the phyla level the TwinsUK
314 and Israeli-PN cohorts were most alike (Figure 1E). The LLDEEP cohort contained relatively lower levels
315 of the phylum Bacteroidetes, and a higher abundance of Firmicutes bacteria. Further differences were
316 observed at the genus level, where the Israeli-PN study had a higher average abundance of Prevotella.

Scale-free p-value thresholding of edges in co-occurrence networks

318 We split the OTU data and generated co-occurrence networks for each dataset using an ensemble approach
319 - taking the intersect of four different correlation measures (TwinsUK example in Figure 2A). This required
320 selection of a p-value threshold at which correlations were considered significant and retained as edges
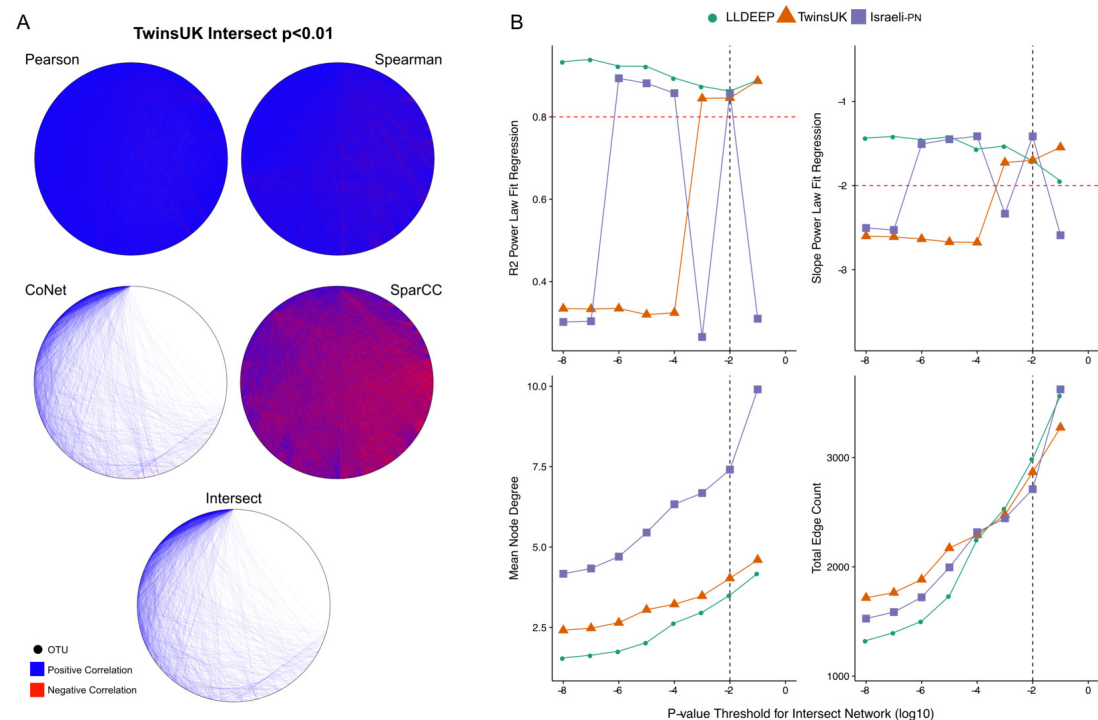
Figure 2. P-value thresholding of co-occurrence networks. A) A visualisation of the ensemble process at the p<0.01 threshold for TwinsUK, the top four networks combine to make the intersect network below. B) Selection of final p-value threshold for generating ensemble networks. Networks for each dataset were made by intersecting four correlation measures with different p-value thresholds. The R^2 and slope of the resultant intersect networks regression fit to a scale-free distribution was assessed using WGCNA, highlighted in red are the developers recommend values to consider a good fit. Also shown are summary measures of the resultant networks at each value showing the mean node degree of OTUs in the networks and the total number of edges in each. The black dashed line indicates the final p-value threshold chosen to generate networks for community detection.

321 in the networks. We generated ensemble networks for each dataset at a range of p-value thresholds and
322 selected a cut-off where the resulting networks' degree distributions best fit a scale-free distribution
323 (Figure 2B). We observed no consistent trend across the p-values, but this might be expected given the
324 differences in the number of OTUs and samples in the datasets and varying levels of p-value precision
325 between the co-occurrence measures used. Indeed, investigating network properties at each threshold for
326 each method individually (Supplementary Figure 1), we observed that the p-value accuracy of SparCC
327 meant that at at all thresholds below 10^{-2} there was no change in the number of edges in the SparCC
328 networks (only edges with p=0 remained). We also observed that the drop in fit to a scale-free distribution
329 at lower p-value cut-offs in the TwinsUK and Israeli-PN networks was driven by trends in the CoNet
330 networks, which had the lowest number of edges and hence were the principal determinant in the intersects
331 at these lower thresholds. However, overall, the trends of the ensemble networks did not simply reflect
332 those of any one of their constituent methods, displaying properties only emergent upon intersection of
333 all four. From examination of the ensemble networks (Figure 2), we found that including edges with a
334 p-value <0.01 produced intersect networks with a good fit to the scale-free model in all three datasets. As
335 such, this threshold was selected to generate the networks used for community detection.

336 From the 6761 unique edges observed across in all three networks, 166 were observed in all three, 882
337 in the TwinsUK and LLDEEP, 677 in TwinsUK and Israeli-PN, and 229 in the LLDEEP and Israeli-PN
338 datasets. Summary statistics for the resultant networks are shown in Table 2. The TwinsUK and LLDEEP
339 network structures were most alike. The Israeli-PN data, which contained fewer nodes (OTUs), produced
340 a smaller and more connected network.

Table 2. Summary statistics for the final intersect co-occurrence networks (p<0.01) for each dataset. Graph density is the percentage of all possible edges represented, mean path length is the minimum calculated pairwise between all nodes, mean clustering coefficient is across all nodes in the network and provides a comparative indicator of overall clustering in the networks.

dataset	Nodes	Edges	Mean Node Degree	Graph Density	Mean Path Length	Mean Clustering Coefficient
TwinsUK	844	2843	3.3	0.008	5.9	0.29
LLDEEP	922	2967	3.2	0.007	5.9	0.28
Israeli-PN	406	2573	6.7	0.033	3.7	0.31

Detection of communities in microbial co-occurrence networks

Selecting a γ parameter for community detection

After establishing co-occurrence networks we used the Louvain modularity maximisation algorithm to detect communities within them (Mucha et al., 2010; Blondel et al., 2008; Newman and Girvan, 2004). The version used includes a constant (γ) that can be used to manipulate the number and size of resultant communities (see Materials and Methods). To determine an appropriate value for γ, we carried out repeated community detection on the three co-occurrence networks at various γ values and assessed the stability (as variation of information between runs) (Figure 3A) and statistical significance (mean modularity of real compared to randomised networks) (Figure 3B) of the community definitions at each γ.

Community definitions were stable (low variation of information between runs) at γ values approaching zero (Figure 3A). This would be expected as in this instance most OTUs fall into a few large communities (Figure 3C). Similarly, at these low γ values, a networks edges are most likely to be within communities, hence we observed high modularity values for both the true and randomised networks. As γ increased we found modularity estimates for the random networks dropped significantly lower than those of the true networks (Figure 3B). This shows that the real co-occurrence networks have significantly more community structure than would be expected by chance, and provides confidence that the communities identified at these γ values reflect the biological relationships of their member OTUs. The stability of definitions (mean variation of information) fluctuated across the γ range, however it should be noted that the lowest estimates (1 – variation of information) were above 0.9 (range 0-1, least to most similar), showing that even the least stable γ produced very similar communities between runs.

Final community definitions

From Figure 3, we selected a γ value of 0.4 for community detection. This γ provided both high modularity estimates (that were significantly higher than the random networks) and good stability in all three datasets. We then used the Louvain algorithm with γ=0.4 to detect the final OTU community definitions in the three co-occurrence networks (Figure 4, Supplementary Table 1). Communities in each network are hence referred to using arbitrary numbers ranging from 1 to the number of communities in the network.

The number of communities detected within the networks (TwinsUK=96, LLDEEP=105, Israeli-PN=31) reflected the number of OTUs within them, but was more similar when only considering communities containing at least 5 OTUs (TwinsUK=35, LLDEEP=36, Israeli-PN=11). There was a positive correlation between the mean relative abundance of an OTU and the total number of OTUs in the community it was assigned to in the TwinsUK and LLDEEP networks (rs=0.1, p=0.002 and rs=0.27, p<0.0001 respectively). However, overall each community in all three networks contained a range of both high and low abundance OTUs (Supplementary Figure 2).

Taxonomy, heritability, and host associations of communities in TwinsUK

Taxonomic distribution within communities

Investigating taxonomic assignments of OTUs within the TwinsUK communities, we found that at finer taxonomic levels (genus and species) several communities contained a mixture of OTUs assigned to different taxa (Supplementary Table 2). Whilst some communities retained a mixture of taxa at the family level, the majority were dominated by one taxon (Figure 4B), and at higher levels nearly all contained a single taxon. A similar pattern was observed in the LLDEEP and Israeli-PN networks (Supplementary Figure 3). These results also reflected the mean sequence similarity between OTUs within communities.
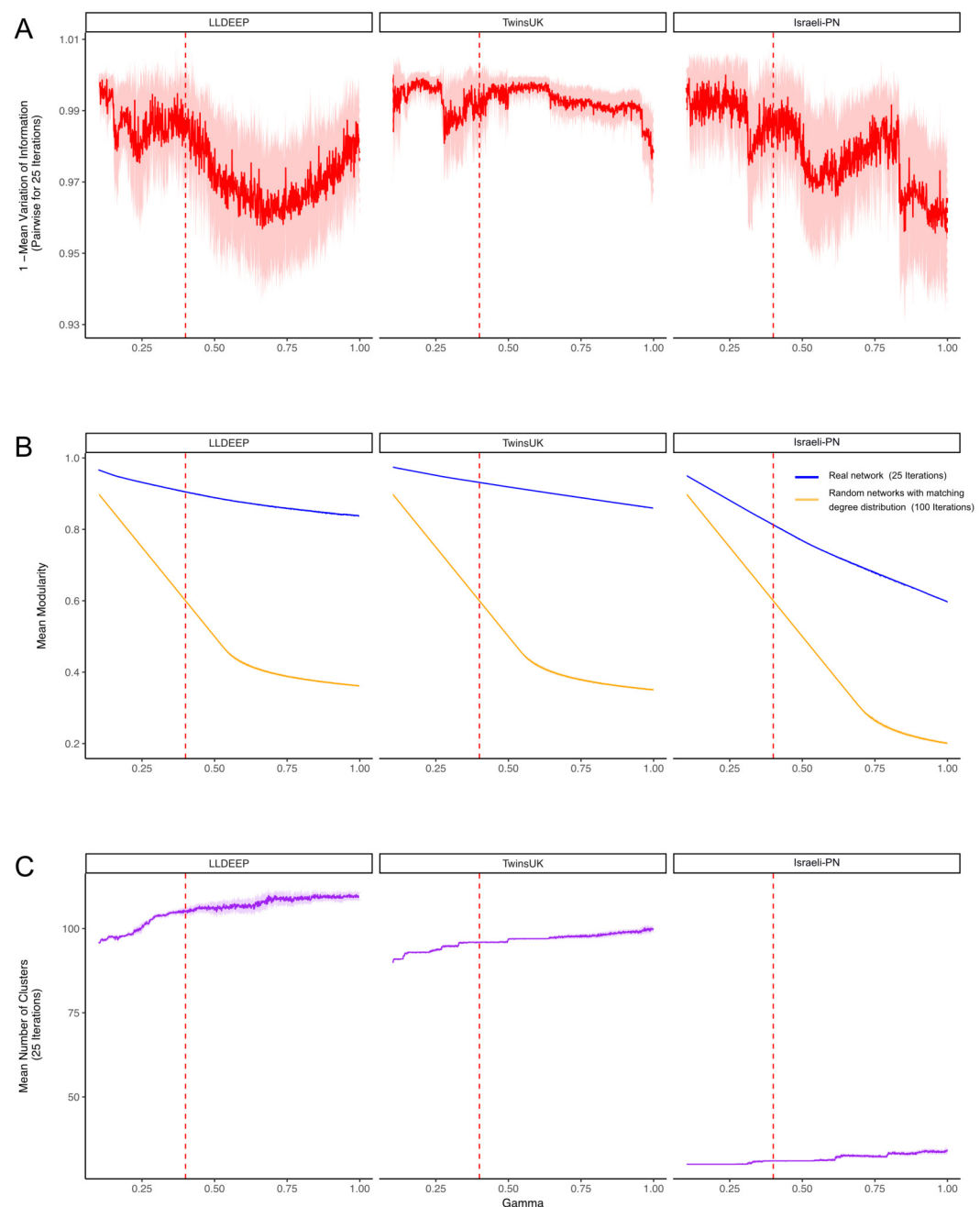
Figure 3. Selecting a γ parameter for Louvain thresholding. Louvain community detection was carried out multiple times on each network on a range of γ values. A) The mean variation of information between the 25 runs at each gamma value. Plotted is 1-variation of information so a higher value means more stability in community definitions between runs. B) The mean modularity of the 25 iterations compared to the mean modularity of 100 randomised networks with the same degree distribution. C) The mean number of communities generated at each value of γ. In all plots, the mean line is surrounded by the interval of standard deviation. This is also true for the plots in B, but it should be noted that modularity estimates had extremely low variance between runs, hence even a small difference between the random and true means is significant. The dashed red line indicates the final value selected for γ.
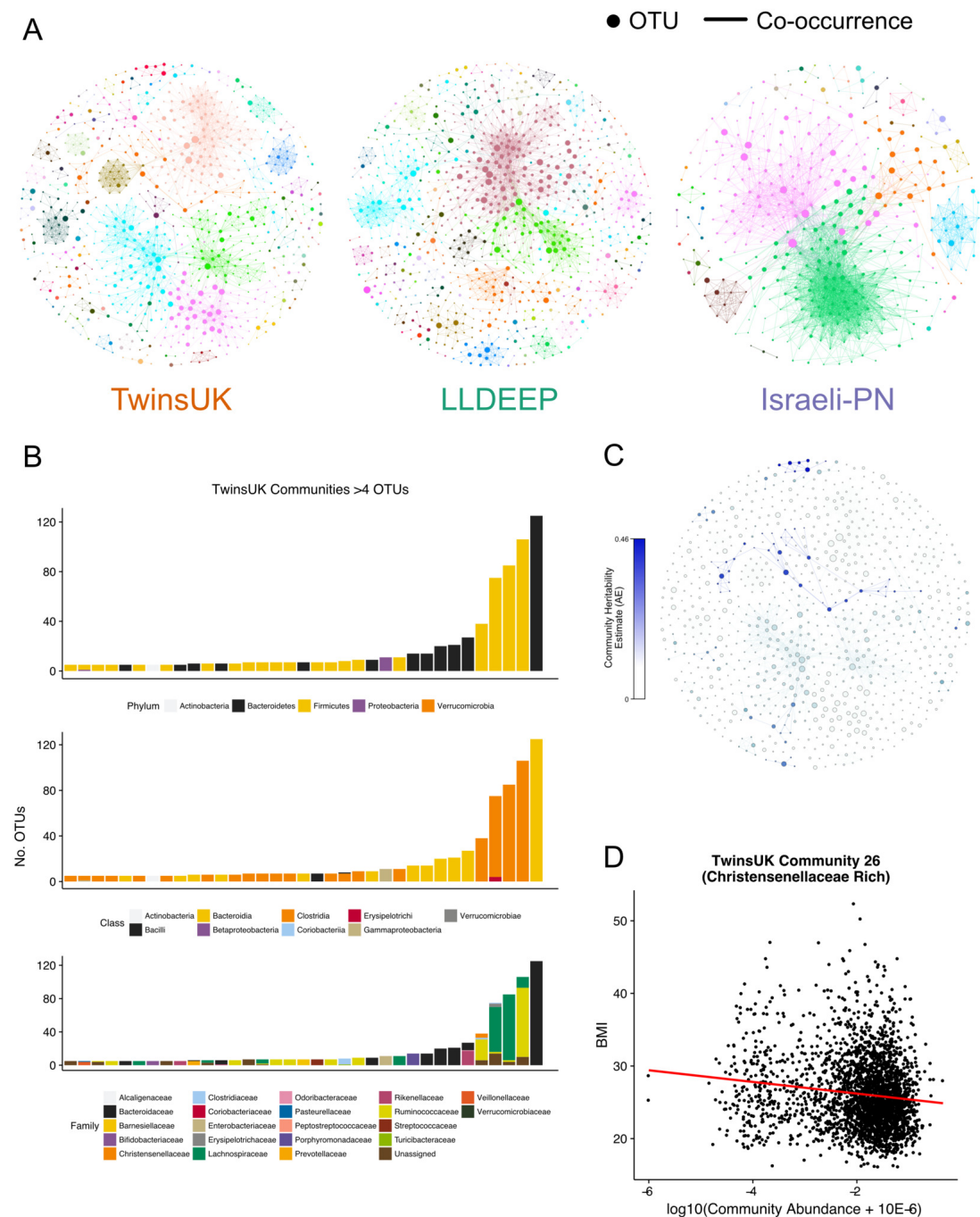
Figure 4. Communities detected within the co-occurrence networks. A) A visualisation of the communities detected by the Louvain algorithm. Communities are arbitrarily coloured. Node size represents an OTUs relative abundance within the dataset. B) Summary of the taxonomic distributions within the TwinsUK communities. Only communities with at least 5 OTUs are shown. Stacked bars represent the number of OTUs in the community assigned to each taxon. C) Community plot for TwinsUK as in A, but coloured by the heritability estimate of the community. D) Highlighting the significant negative association between BMI and the abundance of a Christensenellaceae rich community (linear regression with age, gender, and seq. depth β=-0.13, FDR q=0.001).

We quantified the average pairwise distance between the representative sequences of the OTUs within each community within each dataset (Supplementary Figure 4 and Supplementary Table 3). We found that most communities had a mean sequence identity of 0.94-0.97 between their OTUs. This further suggests communities largely consist of OTUs from taxonomically similar sources. However, we also observed communities with higher divergence between the representative sequences of their OTUs. These included communities 1 (0.84) and 3 (0.82) from the Israeli-PN network, communities 4 (0.82) and 19 (0.84) from the LLDEEP network, and communities 5 (0.88) and 26 (0.88) from the TwinsUK network amongst others (mean identity between OTU representative sequences).

The influence of host genetics on communities

To determine if host genetics influenced the abundance of the microbial communities in TwinsUK, we used an ACE model (which estimates the variance attributable to additive genetics - A, to the environment common within twin pairs - C, and to the environment unique to individuals - E) to estimate the heritability of community abundances within 654 monozygotic and 495 dizygotic pairs. From the 96 communities in the TwinsUK network 52 had some variance attributed to additive genetic effects (Supplementary Table 2). Within these, the variance due to genetic effects ranged from 0.07-0.46 (Figure 4C). The estimates for the most heritable communities are of a similar magnitude to those of the most heritable taxa previously reported within TwinsUK (Goodrich et al., 2014). Indeed, examination of taxonomies within these communities found that they contained highly heritable taxa (Supplementary Table 2). For instance, the most heritable community (Community 39) contained a *Turicibacter* OTU, and the second most (Community 26) contained numerous Christensenellaceae and Ruminococcaceae OTUs (Goodrich et al., 2014).

Community associations with age and BMI

To further explore host effects, we investigated if communities were associated with properties relating to health. We carried out linear regression analysis between community abundances and age and BMI in TwinsUK (Supplementary Table 4). These phenotypes were selected as they were available for all three datasets enabling replication. In TwinsUK, we found that of the 96 communities 47 were significantly associated with BMI (FDR adjusted p<0.05). The highly heritable Community 26 that contained multiple Christensenellaceae OTUs had a significant negative correlation with BMI (FDR adjusted $p<10^{-10}$, β=-0.13)(Figure 4D), reflecting our previous observation that Christensenellaceae in conjunction with correlated taxa are protective against weight gain in mice (Goodrich et al., 2014). There were several communities containing exclusively OTUs belonging to the order Clostridiales that were negatively associated with BMI. The strongest positive association was with Community 5 (FDR adjusted $p<10^{-7}$, β=0.11), which was a large community also dominated by Clostridiales OTUs. However, these were also assigned lower level taxonomies with the majority being various genera from the Lachnospiraceae family.

There were 48 communities significantly associated with age (FDR adjusted p<0.05). The three most significant associations were negative, two of these (Community 54 and 24), were dominated by *Bifidobacterium* OTUs. There was also a significant negative association with Community 27 (FDR adjusted p=0.007, β=-0.05), containing *Faecalibacterium prausnitzii* OTUs. The strongest positive association with age was observed with Community 17 (FDR adjusted $p<10^{-7}$, β=0.11), which contained exclusively Enterobacteriaceae OTUs. There were also several small communities positively associated with age that contained exclusively OTUs assigned to the Ruminococcaceae family.

Comparison of communities across populations

Similarities in overall community structure

Having identified communities within each dataset we aimed to determine how similar the segmentation of OTUs was between them. We again used the normalised variation of information measure to compare groupings, and found that the similarity of community assignments was significantly higher in all three pair-wise comparisons than would be expected by chance (Figure 5A). This shows that OTUs are forming similar communities across the three geographically diverse datasets.

We also repeated community structure comparisons between networks using both higher and lower p-values to threshold edges, and higher and lower values for γ in community detection (Supplementary Figure 5 and Supplementary Table 5). In all instances the community structure was higher than expected by chance, and similar communities were observed both within and between datasets when using alternate parameters. This shows that the approach is robust to differences in parameterisation, and that the

12/21
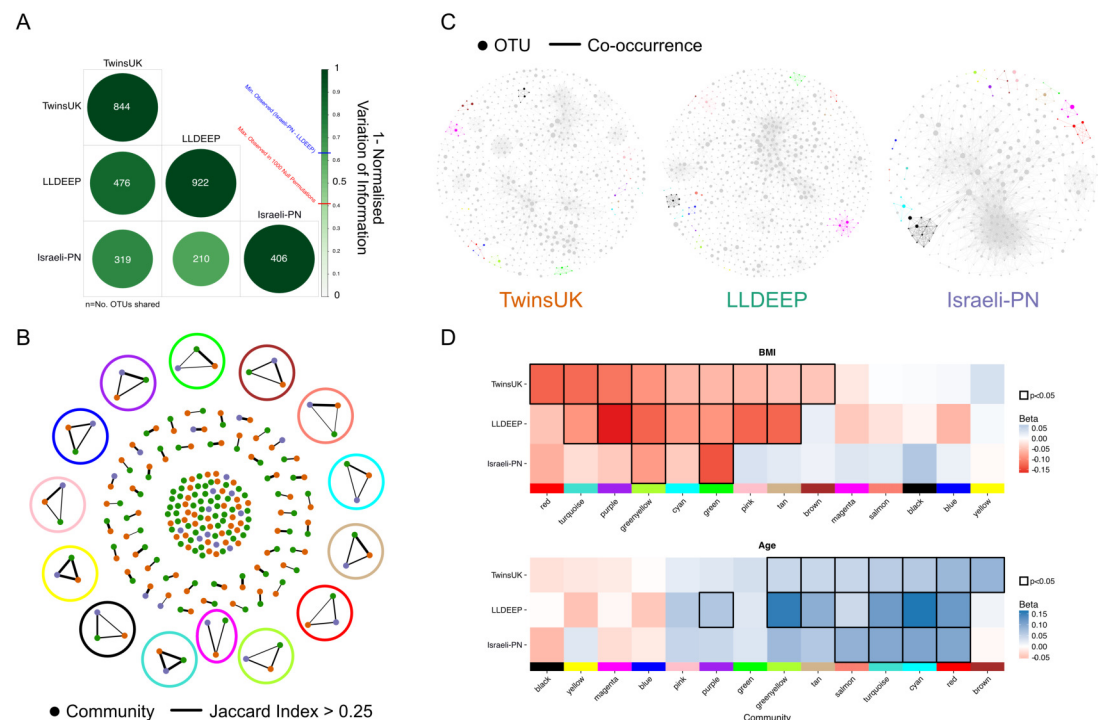
Figure 5. Comparison of communities across populations. A) Pair-wise comparison of variation of information between community definitions for OTUs shared between networks. 1-variation of information is shown where 1 represents identical segmentation of OTUs to communities and 0 indicates no similarity, with the highest score observed in randomised permutations being 0.41. Numbers represent the number of OTUs shared between the sets (OTUs with at least one edge). B) A network showing the matching of communities based on Jaccard index. Edges are considered where the index >0.25 and are weighted by the index 0.25-1. Highlighted are the community-types selected for later analysis, using the colour matching their name. C) Visualisation of networks as in 4A but coloured based on community-types as in B. Communities not in these 14 types are coloured grey. D) Replication of linear regression analysis for each of the 14 mapped communities in the LLDEEP and Israeli-PN for associations with both age and BMI. Squares are coloured by the Beta estimate for the association and nominally significant associations (p<0.05) are highlighted by a black border. Community-types are ordered by their associations within TwinsUK.

435 similarity observed between datasets is not solely manifest at the selected values. However, it should be
436 noted that in all comparisons the variation of information can only be used to compare matching sets, so
437 these results only show communities are similar within the OTUs shared by networks.

Mapping equivalent communities between datasets
439 We mapped the communities (those defined using γ=0.4 and a network edge threshold of p<0.01) between
440 the three networks using the Jaccard index to identify those which were equivalent to one another. Across
441 the 96 TwinsUK, 105 LLDEEP, and 31 Israeli-PN communities, we found 14 instances where communities
442 could be matched in all three datasets (Figure 5B), although there were many more that could be mapped
443 across only two datasets. For distinction, we refer to these 14 matched groups as community-types and
444 label each using arbitrary colour names (Figure 5B and Figure 5C, Supplementary Table 6). For example,
445 the Green community type was community 45 in the LLDEEP network, community 31 in the Israeli-PN
446 network, and community 36 in the TwinsUK network.

Communities have similar associations with age and BMI across populations
448 To determine if each of the 14 community-types had similar associations with age and BMI in their
449 respective populations, linear regressions were carried out as for TwinsUK. Comparing the results
450 (Figure 5D), we found that seven significant BMI associations within TwinsUK were replicated in

the LLDEEP data, with two also significant in the Israeli-PN data. There were six community-types significantly associated with age in both the TwinsUK and LLDEEP data, four of which were also significant in the Israeli-PN data. BMI had more negative associations within the community-types, whereas age had more positive associations. For both BMI and age, the significance of community associations was most similar between TwinsUK and LLDEEP, this is may be due to their higher sample sizes relative to the Israeli-PN dataset.

The Greenyellow and Green community-types had significant negative associations with BMI in all three datasets. Greenyellow communities were dominated by Ruminococcus OTUs in all three, and similarly all the Green communities consisted solely of Ruminococcaceae OTUs. More widely, most community-types with at least one significant negative association with BMI consisted of Clostridiales OTUs. The Pink community type for instance had a significant negative association with BMI in TwinsUK and LLDEEP, and contained multiple Lachnospiraceae OTUs assigned to the species *Coprococcus eutactus* in all three datasets.

The Salmon, Turquoise, Cyan, and Red community-types were significantly positively associated with age across all three datasets. All three contained exclusively Clostridiales OTUs. The strongest positive association with age in TwinsUK was with the Brown module which was not associated with age in the other sets, this contained *Veillonella* and *Haemophilus parainfluenzae* OTUs. Several associations for both age and BMI were significant in TwinsUK and LLDEEP but not in the smaller Israeli-PN dataset, but tended to share the same direction of effect. Overall, these results showed that communities associated with age and BMI in the same manner across the thee populations.

DISCUSSION

Here, we have presented a rationalised and data-driven approach to generate co-occurrence networks from 16S rRNA gene sequencing data, and to identify comparable community structures within them. Applying this to gut microbiota profiles from three different populations we have found that both co-occurrence networks and community structures are stable across all three. Furthermore, we have shown that genetics, age, and BMI are associated with the relative abundances of the identified communities, with age and BMI having similar associations with communities in all three populations.

Differences between datasets

In combining these datasets we observed significant differences in the diversity and taxonomic profiles between datasets. These likely reflect a mixture of both the differences in the study populations and the experimental approaches used. Environmental differences between cohorts, in terms of genetics and other lifestyle factors are known to influence the gut microbiome (Goodrich et al., 2014; Falony et al., 2016; David et al., 2014). For instance, low diversity Prevotella dominant microbiomes have been previously observed in cohort studies at differential levels and are potentially linked to dietary intake (Arumugam et al., 2011; Wu et al., 2011). The larger proportion of Prevotella dominance, and associated reduced diversity, in the Israeli-PN gut microbiota might therefore reflect geographical and dietary differences. However, technical disparities such as faecal sampling and extraction method, which were known to be different between the present studies, are also known to influence taxonomic composition and cannot be discounted (Walker et al., 2015; Kennedy et al., 2014; Sinha et al., 2017). The increased overlap in the TwinsUK and LLDEEP data in comparison to the Israeli-PN dataset in both the OTU and network structures might also result from the higher sample sizes in these studies. However, it is notable that even with the intrinsic population and technical differences between the three studies we still observed consistent elements across them. Most OTUs were found in at least two studies and there was overlap of samples from all three in beta diversity PCoA plots. It is within this common ground that similarities in community structure were observed.

Generating co-occurrence networks

To generate the co-occurrence networks in this study we used an ensemble approach as suggested by Weiss and Van Treuren et al. (Weiss et al., 2016). However, we found that CoNet was the main edge filter when intersecting networks and also produced the best fit to a scale-free distribution when considered alone, and thus might be sufficient. This might be expected given that it is itself an ensemble approach. However, metrics such as SparCC are not implemented within CoNet, and different correlation methods can produce better results depending on the format and properties of the input dataset (Weiss et al.,

2016). These observations may therefore only apply to the current study. Furthermore, there are several alternative approaches that were not considered in the comparisons of Weiss and Van Treuren et al. such as SPIEC-EASI and CCLasso (Fang et al., 2015; Kurtz et al., 2015). Further exploration of community detection using these methods is warranted. An approach to generate quantitative microbiota profiles using a sample's total microbial cell count has also recently been described (Vandeputte et al., 2017). This negates issues of data compostionality and was shown to reduce the number of spurious correlations observed between microbiota. Using such approaches in future studies could enable the use of more traditional correlation metrics and simplify downstream community detection.

We selected edges for inclusion in co-occurrence networks based on their fit to a scale-free topology. Comparing community structures using different p-value thresholds we observed similar community structures between different datasets when using different thresholds. This suggests different p-value cut-offs could be used for each dataset if a single threshold does not produce a reasonable scale-free fit across all datasets as in this study. However, we also observed similar community structures within datasets when using thresholds producing networks with a poorer fit to a scale-free distribution. This is likely due to the use of a more stringent ensemble approach to co-occurrence estimation, which limits edges in the network to the strongest and most consistent associations (arguably most likely to reflect biological phenomena) prior to further thresholding. Use of a scale-free approach to p-value parameterisation might therefore be unnecessary where similar stringent co-occurrence methods are used. It is also not possible to determine if the true network has a scale-free distribution, and indeed the true network could be defined in a number of different ways depending on what relationships are considered to constitute an interaction or edge.

However, it has previously been shown that microbial interaction network structures can approximate a scale-free distribution (Chaffron et al., 2010; Faust et al., 2012; Tong et al., 2013). Scale-free networks are also well conserved in other aspects of biology across different domains of life (Wolf et al., 2002; Albert, 2005), and this approach is used by existing methods such as WGCNA (Langfelder and Horvath, 2008). Using typical co-occurrence approaches a p-value cut-off will be required to determine edges for inclusion when generating a network. Using a scale-free fit, or similar data driven approach, avoids using entirely arbitrary selections and provides a uniform rationale that can be used to generate comparable network structures across different datasets.

Community detection

Following network creation, we utilised a modularity maximisation algorithm to detect communities within the network. Comparing modularity to randomised networks ensured we were detecting structure manifest as a result of non-stochastic processes. Similarly, our application of variation of information enabled identification of the most stable communities across the range of γ values tested. Such parametrization steps are often neglected in network analyses but should be considered in future microbiota studies, particularly where the aim is to identify the optimal community configuration. However, overall the community definitions were stable and similar across the range of γ values considered here, suggesting that the resultant OTU communities were robust to small changes in the γ parameter.

Modularity maximisation assigned OTUs to single communities. This may not represent the true nature of microbiota interactions but provides some analytical benefits. Unambiguous definition of communities facilitates comparison between network structures. This has applications in comparison between disease and control groups (Baldassano and Bassett, 2016) or, as we have shown, replication of community associations across datasets. Unambiguous assignments of OTUs to individual communities might also be desirable for studies aiming to design synthetic communities to replicate the benefits observed with treatments such as faecal microbiota transplants (De Roy et al., 2014). For example, these methods could be applied to identify common microbial communities in successful donor samples, and generate more controlled, rationally designed synthetic communities for use in place of current less specific approaches. There is also evidence that a community-centric approach to studying the gut microbiome in relation to human health can be more relevant than investigating individual taxa. Several studies have shown that associations between individual gut microbiota and host health can depend on their wider community context (Gevers et al., 2014; Goodrich et al., 2014; Baldassano and Bassett, 2016; Ridaura et al., 2013).

Beyond identifying communities, understanding their formation would require extension of the described approaches to quantification of taxa using metagenomic sequencing. This, especially used in

15/21

557 tandem with metatranscriptomics and metabolomics, might provide indications to the mechanisms of
558 interaction. Metabolic modeling from metagenomic data can also be used to improve the inference of
559 interactions from co-occurrence observations by predicting interspecies metabolic dependencies and/or
560 shared niche specialization (Levy and Borenstein, 2013). However, even with such improvements, cross-
561 sectional approaches are inherently limited to inference of interactions from co-occurrence across samples.
562 Time-series data and *in vitro* studies will also be required to delineate directional effects and validate
563 individual interactions (Faust et al., 2015). For instance, there are existing methods that have been used to
564 infer interactions from covariation between microbiota across time series (Steele et al., 2011), and it has
565 been shown that interactions observed in pair-wise species co-cultures can be used to predict outcomes
566 in more complex multi-species cultures (Friedman et al., 2017). Combining *in vitro* observations with
567 community interactions observed in the host environment could be a powerful tool for the design of
568 synthetic bacterial communities for use as gut microbiome targeting medicines (Lindemann et al., 2016).

The stability of communities across populations

570 Although 16S rRNA gene data cannot elucidate interaction mechanisms, we were able to determine
571 several biological phenomena from the microbial approximations provided by OTUs. Most notably, that
572 OTUs formed similar communities with similar host associations in the co-occurrence networks of three
573 different populations. This shows that the variation that exists between the populations (e.g. the OTUs
574 unique to each dataset) does not significantly alter the interactions between the OTUs shared across all
575 three.
576   Several community associations with age and BMI in TwinsUK replicated in the LLDEEP and Israeli-
577 PN datasets. These associations with age and BMI, and the heritability results within TwinsUK, broadly
578 reflected previous observations from studies investigating taxa in isolation. For instance, in TwinsUK
579 communities negatively associated with BMI were enriched with butyrate producers Ruminococcaceae
580 and *Coprococcus eutactus* (Louis and Flint, 2009). Members of the Ruminococcaceae family have
581 previously been associated with visceral fat mass in members of the TwinsUK cohort (Beaumont et al.,
582 2016). Negative associations with butyrate have also been previously observed with metabolic deficits
583 such as type 2 diabetes that are also associated with obesity (Gao et al., 2009; Qin et al., 2012). However,
584 short chain fatty acids have also been observed at higher levels in obese mice (Turnbaugh et al., 2006),
585 and specific taxa associations with obesity were not found in a recent meta-analysis of human data using
586 BMI (Sze and Schloss, 2016). We observed two *Bifidobacterium* rich communities that were negatively
587 associated with age. This is in line with two previous studies (Yatsunenko et al., 2012; Odamaki et al.,
588 2016). Although it should be noted that these considered a wider range of ages and the principal loss of
589 *Bifidobacterium* with age was observed in infants. We also observed *F.prausnitzii* communities that were
590 negatively associated and a community of Enterobacteriaceae that was positively associated with age,
591 similar to previous observations within TwinsUK with frailty (Jackson et al., 2016b). Furthermore, the
592 most heritable communities contained OTUs belonging to highly heritable taxa (Goodrich et al., 2014).

Co-occurrence patterns between taxa

594 It would be expected that community level associations reflected those of taxa based studies as each
595 community constituted of taxonomically similar OTUs. This is in agreement with a previous study that
596 found that co-occurrence to be higher between genetically similar taxa (Chaffron et al., 2010). This
597 indicates either that interactions might evolve within closely related taxa; or that co-occurrence mainly
598 detects genetically related taxa (more likely to have similar functionality) responding to environmental
599 stimuli in the same manner. Reflecting observations that niche differentiation can be a major driver of co-
600 occurrence patterns (Levy and Borenstein, 2013). A further possibility is that co-occurrence communities
601 are grouping reads from source taxa that are improperly captured by the heuristic sequence clustering
602 used to generate OTUs. Whilst this is unlikely the main driver of these communities (most had mean
603 sequence identities more divergent than the 97% threshold), further exploration is warranted to determine
604 the influence of OTU clustering threshold on co-occurrence patterns and communities.
605   We also observed several more taxonomically diverse communities with higher levels of sequence
606 divergence between OTUs. From those highlighted in the results: The Israeli-PN community 3 contained a
607 diverse number of taxa that have been associated with short chain fatty acid (SCFA) production including
608 the genera *Coprococcus* and *Blautia* and the Rikenellaceae family (Duncan et al., 2002; Vital et al., 2015;
609 Louis and Flint, 2017). Similarly, numerous taxa in Community 1 from the Israeli-PN network have been
610 associated with SCFA production, including *F.prausnitzii*, *Roseburia*, *Blautia*, and potentially *Oscillospira*

16/21

(Sokol et al., 2008; Louis and Flint, 2017; Konikoff and Gophna, 2016). Community 4 from the LLDEEP network also contained a similar combination of OTUs assigned to these same SCFA associated taxa. The LLDEEP community 19 contained 29 OTUs all assigned to the order Clostridiales but with a mixture of finer taxonomic assignments including *Oscillospira* and the family Christensenellaceae, which, as well as being a highly heritable taxon, has been experimentally shown to influence *Oscillospira* abundance in mice (Goodrich et al., 2014). OTUs within Community 5 from the TwinsUK network also contained taxa related to SCFA production including *Blautia* and *Coprococcus*, and TwinsUK community 26 contained Christensenellaceae and *Oscillospira* OTUs, similar to LLDEEP Community 19.

These results suggest that the communities containing a more diverse range of taxa are largely driven by a shared involvement in short chain fatty acid production in the gut. This could be due to each taxa having similar responses to host environment and substrate supply, or due to metabolic inter-dependencies between them. Further work using metabolic and metagenomic data will be required to determine which is the case, and is warranted given the beneficial health effects associated with gut microbial SCFA production (Maslowski and Mackay, 2011). The conserved communities containing Christensenellaceae and *Oscillospira* are also of interest given their established interaction and the heritability of the former family (Goodrich et al., 2014). Stable observation of this community across human cohorts might reflect a direct host action on the gut that selectively promotes these communities.

We are unable to determine if the co-occurrence communities observed here reflect closely related taxa responding to environmental niches in a similar manner, or direct interactions between microbes (such as metabolic dependencies) driving inter-dependencies between them. Most likely it is a combination of these factors. Nevertheless, co-occurrence communities provide a useful method to reduce the dimensionality of 16S rRNA gene sequencing data to units reflecting biological phenomena. Further analyses using the described approaches should investigate the influence of host factors such as genetics, diet and health, on these communities and their reciprocal influences on the host.

CONCLUSION

We have described a method to generate robust and comparable community definitions from microbiota co-occurrence networks. We have also described data-driven parameterisation steps and methods to map communities and compare their associations across datasets. This enabled us to demonstrate that the gut microbiome contains stable communities of bacteria that are similarly associated with host factors across geographically diverse populations. Future use of this approach will facilitate community-centric microbiota studies, in particular by aiding replication of findings across datasets.

DATA AVAILABILITY

TwinsUK 16S rRNA gene sequencing data is available from the ENA (Accession:ERP015317). LLDEEP 16S rRNA gene sequencing data is available from the EGA upon request of an account from the Lifelines-DEEP group (Accession: EGAD00001001991). The data from the Israeli-PN study was obtained from the ENA (Accession:PRJEB11532). Code used for the main stages of OTU and co-occurrence network creation and the identification of communities within the networks can be found in Supplementary Materials.

17/21

REFERENCES

Albert, R. (2005). Scale-free networks in cell biology. *Journal of Cell Science*, 118(21):4947.

Arumugam, M., Raes, J., Pelletier, E., Paslier, D. L., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H. B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E. G., Wang, J., Guarner, F., Pedersen, O., de Vos, W. M., Brunak, S., Doré, J., Antolín, M., Artiguenave, F., Blottiere, H. M., Almeida, M., Brechot, C., Cara, C., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Foerstner, K. U., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Huber, W., van Hylckama-Vlieg, J., Jamet, A., Juste, C., Kaci, G., Knol, J., Lakhdari, O., Layec, S., Roux, K. L., Maguin, E., Mérieux, A., Minardi, R. M., M'rini, C., Muller, J., Oozeer, R., Parkhill, J., Renault, P., Rescigno, M., Sanchez, N., Sunagawa, S., Torrejon, A., Turner, K., Vandemeulebrouck, G., Varela, E., Winogradsky, Y., Zeller, G., Weissenbach, J., Ehrlich, S. D., and Bork, P. (2011). Enterotypes of the human gut microbiome. *Nature*, 473:174.

Baldassano, S. N. and Bassett, D. S. (2016). Topological distortion and reorganized modular structure of gut microbial co-occurrence networks in inflammatory bowel disease. *Scientific Reports*, 6:26087.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Web and Social Media*, 8:361.

Beaumont, M., Goodrich, J. K., Jackson, M. A., Yet, I., Davenport, E. R., Vieira-Silva, S., Debelius, J., Pallister, T., Mangino, M., Raes, J., Knight, R., Clark, A. G., Ley, R. E., Spector, T. D., and Bell, J. T. (2016). Heritable components of the human fecal microbiome are associated with visceral fat. *Genome biology*, 17.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):10008.

Chaffron, S., Rehrauer, H., Pernthaler, J., and von Mering, C. (2010). A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research*, 20(7):947.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1.

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, A. S., Varma, Y., Fischbach, M. A., Biddinger, S. B., Dutton, R. J., and Turnbaugh, P. J. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505:559.

De Roy, K., Marzorati, M., Van den Abbeele, P., Van de Wiele, T., and Boon, N. (2014). Synthetic microbial ecosystems: an exciting tool to understand and apply microbial communities. *Environmental Microbiology*, 16(6):1472.

Deng, Y., Jiang, Y.-H., Yang, Y., He, Z., Luo, F., and Zhou, J. (2012). Molecular ecological network analyses. *BMC Bioinformatics*, 13(1):113.

Duncan, S. H., Barcenilla, A., Stewart, C. S., Pryde, S. E., and Flint, H. J. (2002). Acetate utilization and butyryl coenzyme a (coa): acetate-coa transferase in butyrate-producing bacteria from the human large intestine. *Applied and environmental microbiology*, 68(10):5186–5190.

Duran-Pinedo, A. E., Paster, B., Teles, R., and Frias-Lopez, J. (2011). Correlation network analysis applied to complex biofilm communities. *PloS One*, 6(12):e28438.

Falony, G., Joossens, M., Vieira-Silva, S., Wang, J., Darzi, Y., Faust, K., Kurilshikov, A., Bonder, M. J., Valles-Colomer, M., Vandeputte, D., Tito, R. Y., Chaffron, S., Rymenans, L., Verspecht, C., Sutter, L. D., Lima-Mendez, G., Dhoe, K., Jonckheere, K., Homola, D., Garcia, R., Tigchelaar, E. F., Eeckhaudt, L., Fu, J., Henckaerts, L., Zhernakova, A., Wijmenga, C., and Raes, J. (2016). Population-level analysis of gut microbiome variation. *Science*, 352(6285):560.

Fang, H., Huang, C., Zhao, H., and Deng, M. (2015). Cclasso: correlation inference for compositional data through lasso. *Bioinformatics*, 31(19):3172.

Faust, K., Lahti, L., Gonze, D., de Vos, W. M., and Raes, J. (2015). Metagenomics meets time series analysis: unraveling microbial community dynamics. *Current opinion in microbiology*, 25:56.

Faust, K. and Raes, J. (2012). Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538.

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., and Huttenhower, C. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS Computational Biology*, 8(7):e1002606.

Friedman, J. and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Computational Biology*, 8(9):e1002687.

Friedman, J., Higgins, L. M., and Gore, J. (2017). Community structure follows simple assembly rules in microbial microcosms. *Nature Ecology & Evolution*, 1(5).

Fu, J., Bonder, M. J., Cenit, M. C., Tigchelaar, E. F., Maatman, A., Dekens, J. A., Brandsma, E., Marczynska, J., Imhann, F., Weersma, R. K., Franke, L., Poon, T. W., Xavier, R. J., Gevers, D., Hofker, M. H., Wijmenga, C., and Zhernakova, A. (2015). The gut microbiome contributes to a substantial proportion of the variation in blood lipids. *Circulation Research*, page 115.

Gao, Z., Yin, J., Zhang, J., Ward, R. E., Martin, R. J., Lefevre, M., Cefalu, W. T., and Ye, J. (2009). Butyrate improves insulin sensitivity and increases energy expenditure in mice. *Diabetes*, 58(7):1509.

Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Treuren, W. V., Ren, B., Schwager, E., Knights, D., Song, S. J., Yassour, M., Morgan, X. C., Kostic, A. D., Luo, C., González, A., McDonald, D., Haberman, Y., Walters, T., Baker, S., Rosh, J., Stephens, M., Heyman, M., Markowitz, J., Baldassano, R., Griffiths, A., Sylvester, F., Mack, D., Kim, S., Crandall, W., Hyams, J., Huttenhower, C., Knight, R., and Xavier, R. J. (2014). The treatment-naive microbiome in new-onset crohn's disease. *Cell Host & Microbe*, 15(3):382.

Goodrich, J. K., Waters, J. L., Poole, A. C., Sutter, J. L., Koren, O., Blekhman, R., Beaumont, M., Treuren, W. V., Knight, R., Bell, J. T., Spector, T. D., Clark, A. G., and Ley, R. E. (2014). Human genetics shape the gut microbiome. *Cell*, 159(4):789.

Harrell Jr, F. E. and Dupont, C. (2008). Hmisc: harrell miscellaneous. *R package version*, 3(2).

Jackson, M. A., Bell, J. T., Spector, T. D., and Steves, C. J. (2016a). A heritability-based comparison of methods used to cluster 16s rrna gene sequences into operational taxonomic units. *PeerJ*, 4:e2341.

Jackson, M. A., Jeffery, I. B., Beaumont, M., Bell, J. T., Clark, A. G., Ley, R. E., O'Toole, P. W., Spector, T. D., and Steves, C. J. (2016b). Signatures of early frailty in the gut microbiota. *Genome medicine*, 8(1):8.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z., and Barabasi, A. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804):651.

Kennedy, N. A., Walker, A. W., Berry, S. H., Duncan, S. H., Farquarson, F. M., Louis, P., and Thomson, J. M. (2014). The impact of different dna extraction kits and laboratories upon the assessment of human gut microbiota composition by 16s rrna gene sequencing. *PloS One*, 9(2):e88982.

Konikoff, T. and Gophna, U. (2016). Oscillospira: a central, enigmatic component of the human gut microbiota. *Trends in microbiology*, 24(7):523–524.

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Computational Biology*, 11(5):e1004226.

Lambiotte, R. (2010). Multi-scale modularity in complex networks. In *Modeling and optimization in mobile, ad hoc and wireless networks (WiOpt), 2010 Proceedings of the 8th International Symposium on*, pages 546–553. IEEE.

Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559.

Levy, R. and Borenstein, E. (2013). Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proceedings of the National Academy of Sciences*, 110(31):12804.

Levy, R. and Borenstein, E. (2014). Metagenomic systems biology and metabolic modeling of the human microbiome: From species composition to community assembly rules. *Gut Microbes*, 5(2):265.

Lindemann, S. R., Bernstein, H. C., Song, H.-S., Fredrickson, J. K., Fields, M. W., Shou, W., Johnson, D. R., and Beliaev, A. S. (2016). Engineering microbial consortia for controllable outputs. *The ISME journal*, 10(9):2077.

Louis, P. and Flint, H. J. (2009). Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine. *FEMS Microbiology Letters*, 294(1):1.

Louis, P. and Flint, H. J. (2017). Formation of propionate and butyrate by the human colonic microbiota. *Environmental microbiology*, 19(1):29–41.

Lozupone, C., Faust, K., Raes, J., Faith, J. J., Frank, D. N., Zaneveld, J., Gordon, J. I., and Knight, R. (2012). Identifying genomic and metabolic features that can underlie early successional and opportunistic lifestyles of human gut symbionts. *Genome Research*, 22(10):1974.

Maslowski, K. M. and Mackay, C. R. (2011). Diet, gut microbiota and immune responses. *Nature immunology*, 12(1):5–9.

McHardy, I. H., Goudarzi, M., Tong, M., Ruegger, P. M., Schwager, E., Weger, J. R., Graeber, T. G., Sonnenburg, J. L., Horvath, S., Huttenhower, C., McGovern, D. P., Fornace, A. J., Borneman, J., and Braun, J. (2013). Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. *Microbiome*, 1(1):17.

Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., and Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876.

Navas-Molina, J. A., Peralta-Sánchez, J. M., González, A., McMurdie, P. J., Vázquez-Baeza, Y., Xu, Z., Ursell, L. K., Lauber, C., Zhou, H., Song, S. J., Huntley, J., Ackermann, G. L., Berg-Lyons, D., Holmes, S., Caporaso, J. G., and Knight, R. (2013). Advancing our understanding of the human microbiome using qiime. *Methods in Enzymology*, 531:371.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113.

Odamaki, T., Kato, K., Sugahara, H., Hashikura, N., Takahashi, S., Xiao, J.-z., Abe, F., and Osawa, R. (2016). Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. *BMC microbiology*, 16(1):90.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., Wu, P., Dai, Y., Sun, X., Li, Z., Tang, A., Zhong, S., Li, X., Chen, W., Xu, R., Wang, M., Feng, Q., Gong, M., Yu, J., Zhang, Y., Zhang, M., Hansen, T., Sanchez, G., Raes, J., Falony, G., Okuda, S., Almeida, M., LeChatelier, E., Renault, P., Pons, N., Batto, J.-M., Zhang, Z., Chen, H., Yang, R., Zheng, W., Li, S., Yang, H., Wang, J., Ehrlich, S. D., Nielsen, R., Pedersen, O., Kristiansen, K., and Wang, J. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490:55.

Ridaura, V. K., Faith, J. J., Rey, F. E., Cheng, J., Duncan, A. E., Kau, A. L., Griffin, N. W., Lombard, V., Henrissat, B., Bain, J. R., Muehlbauer, M. J., Ilkayeva, O., Semenkovich, C. F., Funai, K., Hayashi, D. K., Lyle, B. J., Martini, M. C., Ursell, L. K., Clemente, J. C., Treuren, W. V., Walters, W. A., Knight, R., Newgard, C. B., Heath, A. C., and Gordon, J. I. (2013). Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science*, 341(6150):1241214.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). Vsearch: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584.

Ronhovde, P. and Nussinov, Z. (2009). Multiresolution community detection for megascale networks by information-based replica correlations. *Physical Review E*, 80(1):016109.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Horn, D. J. V., and Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23).

Sinha, R., Abu-Ali, G., Vogtmann, E., Fodor, A. A., Ren, B., Amir, A., Schwager, E., Crabtree, J., Ma, S., Abnet, C. C., Knight, R., White, O., and Huttenhower, C. (2017). Assessment of variation in microbial community amplicon sequencing by the microbiome quality control (mbqc) project consortium. *Nature Biotechnology*.

Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermudez-Humaran, L. G., Gratadoux, J.-J., Blugeon, S., Bridonneau, C., Furet, J.-P., Corthier, G., Grangette, C., Vasquez, N., Pochart, P., Trugnan, G., Thomas, G., Blottiere, H. M., Dore, J., Marteau, P., Seksik, P., and Langella, P. (2008). Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of crohn disease patients. *Proceedings of the National Academy of Sciences*, 105(43):16731.

Steele, J. A., Countway, P. D., Xia, L., Vigil, P. D., Beman, J. M., Kim, D. Y., Chow, C.-E. T., Sachdeva, R., Jones, A. C., Schwalbach, M. S., Rose, J. M., Hewson, I., Patel, A., Sun, F., Caron, D. A., and Fuhrman, J. A. (2011). Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *The ISME journal*, 5(9):1414–1425.

Sze, M. A. and Schloss, P. D. (2016). Looking for a signal in the noise: Revisiting obesity and the

microbiome. *MBio*, 7(4):e01018.

Tong, M., Li, X., Parfrey, L. W., Roth, B., Ippoliti, A., Wei, B., Borneman, J., McGovern, D. P. B., Frank, D. N., Li, E., Horvath, S., Knight, R., and Braun, J. (2013). A modular organization of the human intestinal mucosal microbiota and its association with inflammatory bowel disease. *PloS One*, 8(11):e80702.

Traag, V. A., Krings, G., and Van Dooren, P. (2013). Significant scales in community structure. *Scientific reports*, 3.

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444:1027.

Vandeputte, D., Kathagen, G., D'hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., Wang, J., Tito, R. Y., Commer, L. D., Darzi, Y., Vermeire, S., Falony, G., and Raes, J. (2017). Quantitative microbiome profiling links gut community variation to microbial load. *Nature*, 551(7681):507.

Vital, M., Gao, J., Rizzo, M., Harrison, T., and Tiedje, J. M. (2015). Diet is a major factor governing the fecal butyrate-producing community structure across mammalia, aves and reptilia. *The ISME journal*, 9(4):832–843.

Walker, A. W., Martin, J. C., Scott, P., Parkhill, J., Flint, H. J., and Scott, K. P. (2015). 16s rrna gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and pcr primer choice. *Microbiome*, 3(1):26.

Weiss, S., Treuren, W. V., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L. C., Xu, Z. Z., Ursell, L., Alm, E. J., Birmingham, A., Cram, J. A., Fuhrman, J. A., Raes, J., Sun, F., Zhou, J., and Knight, R. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME Journal*, 10(7):1669.

Westcott, S. L. and Schloss, P. D. (2015). De novo clustering methods outperform reference-based methods for assigning 16s rrna gene sequences to operational taxonomic units. *PeerJ*, 3:e1487.

Wolf, Y. I., Karev, G., and Koonin, E. V. (2002). Scale-free networks in biology: new insights into the fundamentals of evolution? *BioEssays*, 24(2):105.

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F. D., and Lewis, J. D. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J. G., Lozupone, C. A., Lauber, C., Clemente, J. C., Knights, D., Knight, R., and Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature*, 486:222.

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., Suez, J., Mahdi, J. A., Matot, E., Malka, G., Kosower, N., Rein, M., Zilberman-Schapira, G., Dohnalová, L., Pevsner-Fischer, M., Bikovsky, R., Halpern, Z., Elinav, E., and Segal, E. (2015). Personalized nutrition by prediction of glycemic responses. *Cell*, 163(5):1079.

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1):1–45.