

26/11/2025

# ENDOTARGET/ MICROBIOTA ASSOCIATED WITH SE

WP1, Deliverable 1.3

DELIVERABLE VERSION:  
D1.3, V.0.3

DISSEMINATION LEVEL:  
PU

AUTHOR(S):  
Samuel Neuenschwander  
Marco Pagni



## DOCUMENT HISTORY

PROJECT ACRONYM	ENDOTARGET	
<hr/>		
Project Title	Systemic Endotoxemia as the driver of chronic inflammation - Biomarkers and novel therapeutic targets for Arthritis	
Grant Agreement No	101095084	
Project Coordinator	HUS	
Project Duration	01/01/2023 – 31/12/2026	
Deliverable No.	1.3	
Diss. Level	Public	
Deliverable Lead	SIB	
Status		Working Verified by other WPs/Partners
	X	Final Version
Due date	30.11.2025	
Submission date	26.11.2025	
Work Package	WP1	
Work Package Lead	HUS	
Contribution	ETHZ; UTARTU; HUS	
Beneficiary(ies)		
DoA		

DATE	VERSION	AUTHOR	COMMENT
07/11/2025	0.1	Samuel Neuenschwander (SIB), Marco Pagni (SIB)	First draft
19/11/2025	0.2	Samuel Neuenschwander (SIB), Marco Pagni (SIB), Flavia Marzetta (SIB), Elin Org (UTARTU), Gonçalo Barreto (HUS)	Second draft
24/11/2025	0.3	Samuel Neuenschwander (SIB), Marco Pagni (SIB), Flavia Marzetta (SIB), Marcy Zenobi-Wong (ETHZ)	Third draft

©2023–2026 ENDOTARGET Consortium Partners. All rights reserved.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authorities can be held responsible for them.

ENDOTARGET is a Horizon Europe project supported by the European Commission under grant agreement No 101095084 and it is supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00462. All information in this deliverable may not be copied or duplicated in whole or part by any means without express prior agreement in writing by the ENDOTARGET partners. All contents are reserved by default and may not be disclosed to third parties without the written consent of the ENDOTARGET partners, except as mandated by the Grant Agreement with the European Commission, for reviewing and dissemination purposes. All trademarks and other rights on third party products mentioned in this document are acknowledged and owned by the respective holders. The ENDOTARGET consortium does not guarantee that any information contained herein is error-free, or up-to-date, nor makes warranties, express, implied, or statutory, by publishing this document. For more information on the project, its partners and contributors, please see the ENDOTARGET website (<https://endotargetproject.eu/>).

## EXECUTIVE SUMMARY

This deliverable (D1.3) reports analyses conducted within the ENDOTARGET project to investigate the relationship between the gut microbiota and the risk of rheumatic diseases—specifically osteoarthritis (OA). A first round of statistical analyses across multiple cohorts revealed no strong taxonomic or functional associations between OA and the gut microbiota. Subsequently, SIB developed a novel search algorithm, the diffusion-based network approach (SGNDB). It relies on the Unified Human Gastrointestinal Genome (UHGG) resource to construct a large-scale gene/protein network and applies diffusion statistics to identify sets of genes associated with clinical traits.

Key findings include:

- **Taxonomic analysis**, independently on four cohorts (FINRISK, EstMB, Lifelines, TwinsUK) showed no significant differences in microbiome composition between OA patients and controls.
- **Functional analysis** (TwinsUK, EstMB) using KEGG and Gene Ontology annotations of UHGG, revealed a clear association with BMI and a weaker one with OA. Because OA and BMI are known to be correlated, the observed relationship between microbiomes and OA was not deemed significant.
- **SGNDB analysis** (FINRISK, EstMB, TwinsUK) identified groups of UHGG proteins associated with OA and BMI.
  - At higher taxonomic ranks, the numbers of groups showing positive and negative associations were relatively balanced. However, at lower taxonomic ranks, such as the species level, distinct taxa appear positively associated with OA and BMI, while others show negative associations. Overall, associations were stronger for OA than for BMI.
  - Several **Bacteroidaceae species** tend to be negatively correlated with OA and positively correlated with BMI
  - A few uncultivated **Firmicutes species** showed positive correlations with OA.
  - Cross-cohort comparisons indicate that the patterns observed for OA can be reproduced across cohorts to some extent.



# TABLE OF CONTENT

Document History .....	2
Executive Summary .....	4
Table of Content .....	5
List of Tables.....	6
List of Figures .....	6
List of Abbreviations .....	6
1 Introduction .....	9
2 Methods .....	11
2.1 Taxonomic Analysis.....	11
2.2 The UHGG Resource .....	11
2.3 Functional analysis.....	13
2.4 New Approach.....	13
3 Results .....	14
3.1 Taxonomic Analysis.....	14
3.2 Functional analysis.....	15
3.3 SGNDB SIB Gene/Node Diffusion-Based Algorithm .....	17
3.3.1 Algorithm considerations .....	17
3.3.2 Network preparation.....	19
3.3.3 Metagenome screening.....	21
3.3.4 Diffusion-statistics .....	23
3.4 SGNDB Analysis of OA and BMI.....	25
3.4.1 Analyses per cohort .....	26
3.4.2 Comparing cohorts.....	31
4 Discussion.....	36
5 References .....	39





## LIST OF TABLES

Table 1. UHGP annotations..... 12  
 Table 2. Metagenome mappings..... 16  
 Table 3. Network size statistics..... 21  
 Table 4. Characteristics of the three cohorts..... 22  
 Table 5. Top nodes and subnetwork count..... 25

## LIST OF FIGURES

Figure 1. Visual representation of the UHGG resource..... 12  
 Figure 2. Microbial diversity in the gut microbiome..... 15  
 Figure 3. The heatmaps illustrate the Pearson correlation between module eigengenes and clinical traits..... 17  
 Figure 4. Workflow showing the three parts of the SGNDB approach ..... 19  
 Figure 5. Schematic representation of the protein network ..... 20  
 Figure 6. Top subnetwork example ..... 24  
 Figure 7. Distribution of OA Z-scores related to 100 randomized traits ..... 25  
 Figure 8. The figure visualizes table 5..... 26  
 Figure 9. Top operons per taxa at different ranks. .... 28  
 Figure 10. Top operons per taxa at species level..... 30  
 Figure 11. Intersections of subnetworks across cohorts ..... 31  
 Figure 12. Top six subnetworks associated with OA in all cohorts ..... 34  
 Figure 13. Matthews correlation coefficient..... 34  
 Figure 14. Cross-cohort validation test ..... 36  
 Figure 15. Graphical summary of the discussion about microbiology ..... 38

## LIST OF ABBREVIATIONS

ACRONYM	DESCRIPTION
SE	Systemic Endotoxemia
LPS	Lipopolysaccharides

RDs	Rheumatic Diseases
RA	Rheumatoid Arthritis
SpA	Spondylarthritis
OA	Osteoarthritis
ME	Metabolic Endotoxemia
MAG	Metagenome-Assembled Genomes
BMI	Body Mass Index
WGCNA	Weighted Gene Co-expression Network Analysis
HPC	High-Performance Computing
TwinsUK	UK Adult Twin Registry
EstMB	Estonian Microbiome Biobank
FINRISK	Finnish Biobank
Lifelines	A three-generation cohort study and biobank
EggNOG	Evolutionary genealogy of genes: Non-supervised Orthologous Groups
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG Orthologues
GO	Gene Ontology
MMC	Matthew Correlation Coefficient
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG Orthologues
GO	Gene Ontology
MMC	Matthew Correlation Coefficient
HGT	Horizontal Gene Transfer
CTRL	Control group

**SGNDB vocabulary**

SGNDB	The SIB new genes-network/diffusion-based approach
UHGR	Unified Human Gastrointestinal Resource
UHGG	Unified Human Gastrointestinal Genomes
UHGG-species	UHGG taxonomy annotation of clustered UHGG
UHGP	Unified Human Gastrointestinal Proteins
UHGP-90	UHGP clustered at the 90% amino acid identity
UHGP-90'	UHGP re-clustered with CD-hit by SIB
UHGP-90'-protein	In SGNDB, the proteins of the network
UHGP-90'-operon	In SGNDB, the (meta-)operons of the network
UHGP-90'-oset-A	In SGNDB, the nodes enforcing inter-species operons
UHGP-90'-oset-B	In SGNDB, the nodes connecting UHGP-90'-oset-A nodes based on the taxonomy from EggNOG
overall network	Initial complete (huge) network made of UHGP-90'-protein, UHGP-90'-operon, and additional nodes, connected by edges. It contains several disconnected components.
fragmented network	Fragmented and reduced 'Overall network' to computationally feasible dimensions.
subnetwork	Fully connected network with less than 1000 nodes, resulting from the fragmentation of the overall network.
top node/operon/subnetwork	Node/operon or subnetwork with an absolute Z-score equal or greater than 6. Positively signed nodes are positively correlated with the trait, whereas negatively signed nodes are negatively correlated.

# 1 INTRODUCTION

The overarching goal of ENDOTARGET is to explore the role of chronic systemic inflammation caused by intestinal microbiota derived immunologically active compounds, as a driver in the transition from health to disease, with a special focus on three RDs; osteoarthritis (OA), rheumatoid arthritis (RA), and spondylarthritis (SpA).

Task 1.4 (reported in Deliverable 1.3) focused on investigating the relationship between gut microbiota and systemic inflammation (SE). While the original task aimed to examine the impact of microbiota composition on intestinal permeability, SE and the risk of RA, SpA and OA by integrating multiple cohorts, the availability of datasets and the results of preliminary analyses led us to refine the scope of our analysis.

Accessing the FINRISK cohort took much longer than expected. Access to the EstMB cohort also took longer than anticipated, with limited access preventing us from running all possible analyses. SIB therefore decided to start working with the TwinsUK cohort, which is publicly available.

Pre-analyses by ETHZ to find an association between the gut microbiota and OA and BMI did not detect any significant association (Bevc et al. 2025). Further analyses by SIB, searching for a functional association between OA, BMI, and the gut microbiome, also failed to reveal a significant association between OA and the gut microbiota (Bevc et al. 2025).

The SIB therefore decided to develop a novel diffusion-based approach to identify sets of proteins associated with a clinical trait. SIB demonstrated the potential of this approach using the clinical trait OA and BMI. Extending this to other clinical traits should be simple when the data is available.

A microbiome represents a metapopulation of predominantly bacterial organisms, and identifying associations between it and a clinical trait is challenging due to the inherently multidimensional nature of the data. Broadly, two categories of association searches are used:

- **Taxonomic level:** Differences may be observed in the taxonomic composition of the microbiota, with taxa enriched or depleted in clinical cases.
- **Functional level:** Different species may perform the same function; therefore, for the host's well-being, it is not important whether one species or another is present, as long as the necessary functions are performed.

Microbial taxa are commonly analyzed in association studies to identify taxa that are enriched or depleted in clinical conditions. However, bacterial classification is continually being revised, and metagenomic data often cannot be reliably assigned down to the species level. This taxonomic instability complicates interpretation.

Concurrently, detecting associations at the functional level is even more challenging. Current functional annotation databases are often incomplete, biased, or incorrect. Most genes in metagenomes still lack any functional annotation at all (Table 1), while those that are annotated originate largely from a limited set of organisms. As a result, association testing at the functional level is restricted to the subset of genes whose functions are already known.

In other words, the scope and quality of taxonomic and functional databases are major limiting factors in interpreting microbiome research results.

Early preliminary analyses from ETHZ indicated that, across four cohorts, no strong associations could be detected between OA and the gut microbiome at the taxonomic level. Similarly, preliminary analyses from SIB, based on two cohorts, suggested that no strong functional-level associations between the gut microbiome and OA could be identified.

To overcome the limitation of classical statistical approaches, SIB has decided to develop a new approach focusing on feature-search and leveraging the large numbers of bacterial genomes available nowadays. We will refer to this new method as SGNDB approach (SIB new genes-network/diffusion-based), because it includes one step based on so-called diffusion statistics on a gene/protein network. This strategy uses the genetic context to construct a genetic network, which is then systematically searched for genes and operons associated with the clinical trait of interest.

In this deliverable, we apply the SGNDB approach to three cohorts to identify sets of genes associated with OA and BMI.

## 2 METHODS

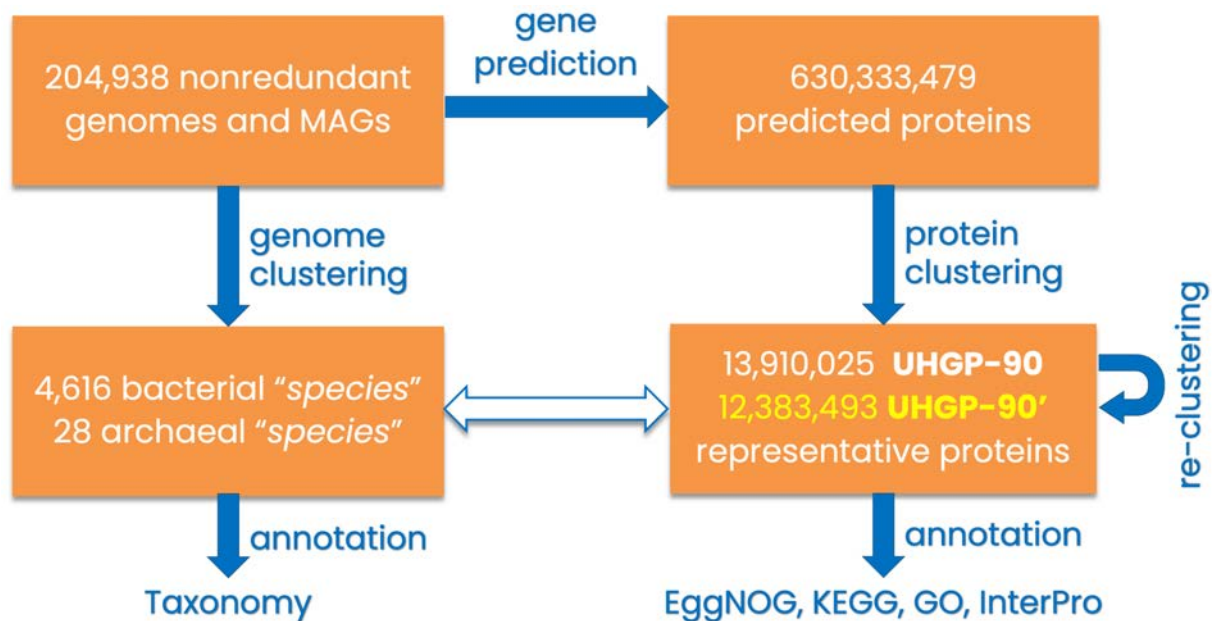
### 2.1 TAXONOMIC ANALYSIS

ETHZ analyzed the gut microbiota of a total of 1395 OA patients from the FINRISK cohort, EstMB, Lifelines and TwinsUK study, and compared them with 1395 healthy control subjects matched by age, gender, BMI, and physical activity selected from a pool of total 17'641 participants. Participants with conditions known to affect the gut microbiome (Inflammatory Bowel Syndrome, Inflammatory Bowel Disease, Coeliac disease, Pregnancy, Medicated for Depression and Diabetes) were excluded. Whole genome sequencing (WGS) raw reads were quality checked with fastqc and then trimmed accordingly, removing potential remaining adapter sequences, human reads as well as low quality bases. The resulting high-quality reads were analyzed with the software MAPseq (Matias Rodrigues et al. 2017) and mOTUs (Ruscheweyh et al. 2022) to obtain taxonomic profiles that were used for association testing. For this, we used the software package SIAMCAT (Wirbel et al. 2021) in R, based on a non-parametric Wilcoxon test and performing multiple testing correction with the Benjamini-Hochberg procedure. Additionally, we performed machine learning by training a LASSO model and checked if the microbiome composition could be predictive of the disease state.

### 2.2 THE UHGG RESOURCE

For the functional analyses and the SGNDB approach, we relied on a recently developed and highly comprehensive resource: the Unified Human Gastrointestinal Genome (UHGG) resource, version 2.0.1 (Figure 1; Almeida et al. 2021; Richardson et al. 2023). The UHGG resource is a compilation of publicly available human gut microbiome data, comprising 204,938 non-redundant genomes and metagenome-assembled genomes (MAGs). The non-redundant genomes and MAGs have been partitioned into 4,644 clusters, thereafter, referred to as *UHGG-species*. Furthermore, over 630 million genes were predicted from the genomes and MAGs. The protein-coding genes were translated into protein sequences (*UHGP*) and clustered at 90% amino acid identity, resulting in 13,910,025 representative proteins, referred to as *UHGP-90*. UHGP-90 proteins were further annotated using eggnoG (Hernandez-Plaza et al. 2023), InterPro (Mulder and Apweiler 2007), COG (Galperin et al. 2025), GO (Ashburner et al. 2000) and KEGG (Kanehisa et al. 2016)

(Table 1). Thus, the UHGG and UHGP resources provide a comprehensive representation of available gut genomes and proteins.



**Figure 1.** Visual representation of the UHGG resource. SIB re-clustered the published representative proteins for the diffusion-based analysis (in yellow).

Investigating the UHGG resource revealed that the clustering of the UHGP-90 proteins is not perfect, leaving a substantial proportion of the representative proteins exhibited an amino acid identity greater than 90%. To allow a sensitive mapping of the metagenomes (removing reads mapping to multiple proteins) SIB re-clustering UHGP-90 dataset using a more sensitive clustering approach. This resulted in the expected significant reduction in the number of representative proteins showing an amino acid identity greater than 90% and reduced the number of proteins in the UHGP-90 dataset by 11%, leaving 12,383,493 representative proteins, referred to as UHGP-90' (see Figure 1).

**Table 1.** UHGP annotations of KEGG Orthologues (KO) and Gene Orthologues (GO). 'Annotated proteins' lists the number of proteins with a given annotation. While many proteins have no annotation, some proteins have multiple annotations of the same type. 'Unique features' lists the number of unique features present. Many proteins contain the same annotation.

	annotated proteins		unique features	
<b>proteins</b>	13,910,025	100%	13,910,025	100%
<b>KO</b>	4,851,159	35%	9,885	0.07%
<b>GO</b>	521,685	4%	15,406	0.11%

## 2.3 FUNCTIONAL ANALYSIS

SIB conducted a functional analysis using KEGG and Gene Ontology annotations. The same control-matched participants included in the taxonomic analysis were selected for this step: 56 participants of the UKtwins cohort and 557 participants of the EstMB cohort. All metagenomes from these participants were aligned to the ~14 million UHGP-90 proteins using the software diamond blastx (Buchfink et al. 2015). The alignment parameters were adapted to the cohort data, optimized for specificity rather than coverage.

For each participant, alignments were counted, filtered, and summed up by KO or GO. We conducted a weighted gene co-expression network analysis (WGCNA; Langfelder and Horvath 2008) to group proteins into modules based on their similar expression patterns across samples). The expression profile of each module was summarized by an eigengenes, such as the first principal component of the expression matrix of the genes in that module. The biological relevance of the modules was then assessed by correlating eigengenes with clinical traits such as with OA and BMI traits for both cohorts and with Physical Activity and Gender (note, that the TwinsUK cohort just consist of females) only for the EstMB cohort.

## 2.4 NEW APPROACH

Such a functional analysis is however a priori impaired by the incompleteness of KEGG and Gene Ontology annotations with respect to bacterial protein diversity (see Table 1). To overcome this limitation, SIB developed a new numerical method to better exploit the UHGG resource of the gut microbiome, by less relying on pre-existing functional annotations.

Unlike in the functional analysis, the metagenomes in this step were mapped to a re-clustered version of the representative UHGP-90 proteins, referred to as UHGP-90' proteins (Figure 1). This re-clustering was necessary to help obtaining spherical clusters with a consistent radius during the alignment phase. Furthermore, reducing the number of proteins increases the proportion of reads that could ultimately be mapped.

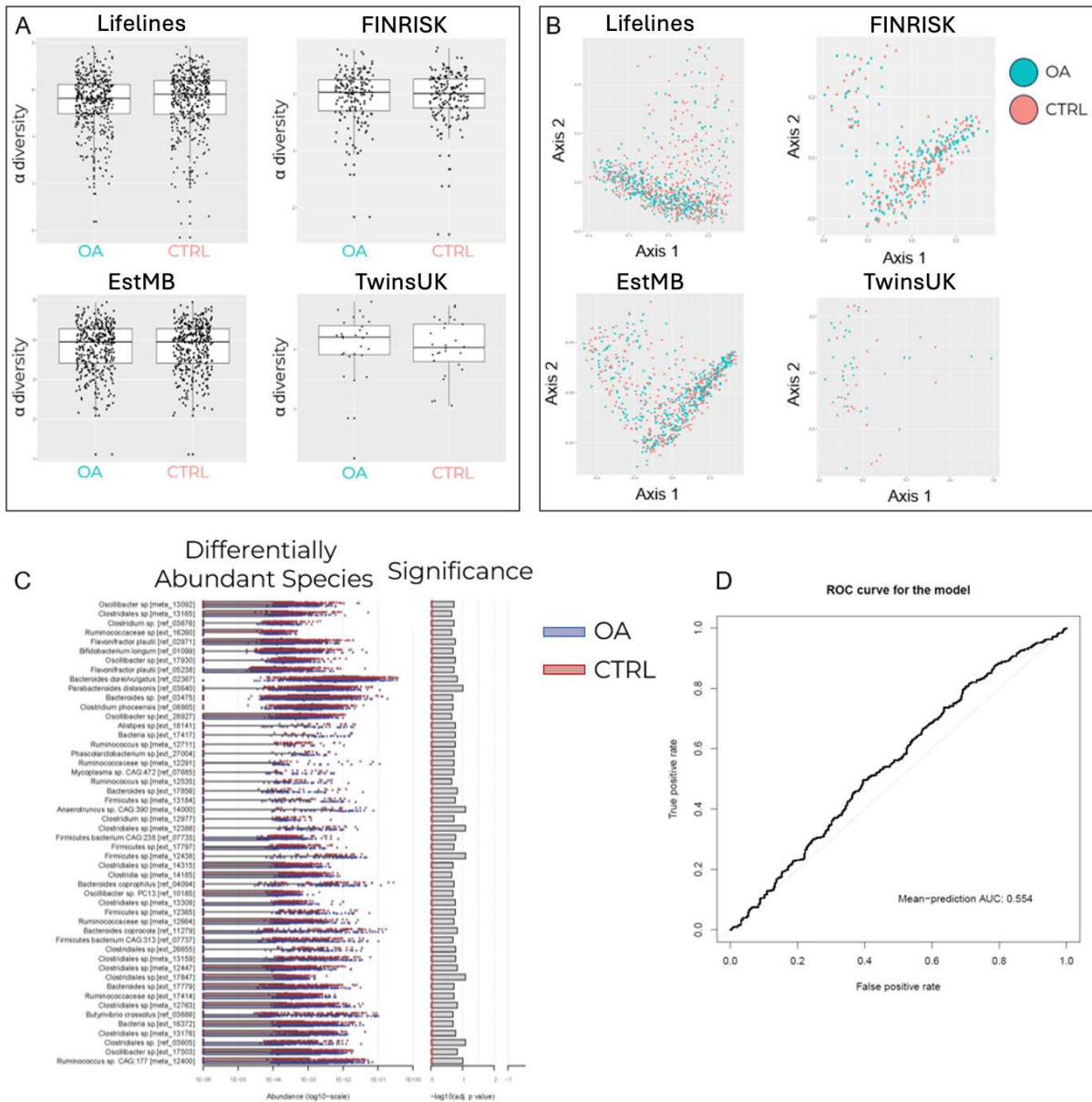
For the functional analysis, it makes no difference whether metagenomes are mapped to the original UHGP-90 representative proteins or to the re-clustered UHGP-90' set, because per-participant alignments were aggregated by KO or GO terms before filtering was applied. In contrast, under the SGNDDB approach, the post-

mapping filtering occurs directly at the protein level, making the choice of protein clustering relevant.

## 3 RESULTS

### 3.1 TAXONOMIC ANALYSIS

ETHZ observed no difference in species and community composition diversity (alpha and beta diversity) between the OA and control groups (Figure 2A, 2B). Furthermore, we were not able to identify any microbial species that were statistically significantly enriched in the microbiota of OA patients. Our validated AI model could not accurately predict whether the microbiota belonged to an OA patient or not. This means that also more complex combinations of microbial species abundances, which can be missed in the association analysis, were not enough to distinguish cases from controls (Figure 2C). Additionally, and in line with previous literature, we observed that age was the key factor significantly impacting the microbial gut community also in OA patients (Figure 2D). Our data suggests that in the current population cohorts, OA neither elicits nor drives gut microbiota changes and that the previously reported differences were likely derived from lack of exclusion criteria, small sample size and/or lack of age correction. Together, this work underlines the importance of matched controls and demonstrates that the microbiome composition alone is not correlated with OA disease state.



**Figure 2.** Microbial diversity in the gut microbiome of OA patients of all four different cohorts. A)  $\alpha$ -diversities represented by Shannon index. Each dot corresponds to a different alpha-diversity value. Boxes represent the first and third quartiles (25% and 75%), the inner horizontal line the median. B)  $\beta$ -diversity (Bray Curtis) represented with PCoA in each of the four cohorts comparing OA (red) and control (blue) cases. C) differentially abundant taxa (non-significant) enriched in OA (red) and control (blue). D) AUROC Curve of the LASSO model calculated for the FINRISK.

### 3.2 FUNCTIONAL ANALYSIS

SIB performed a functional analysis using KEGG and GO annotations. Only 35% of UHPG proteins carried a KEGG annotation and just 4% had a GO annotation.

Moreover, these annotations were highly redundant: they comprised only 9,885 unique KEGG features and 15,406 unique GO features and thus represent less than 1% of the granularity present in the UHGP protein set. As the KEGG and GO analyses produced broadly similar outcomes, we report only the KEGG results here.

Metagenomic reads from the EstMB and TwinsUK cohorts aligned to 16% and 9% of UHGP proteins, respectively, yet covered 71% and 66% of all KO (Table 2).

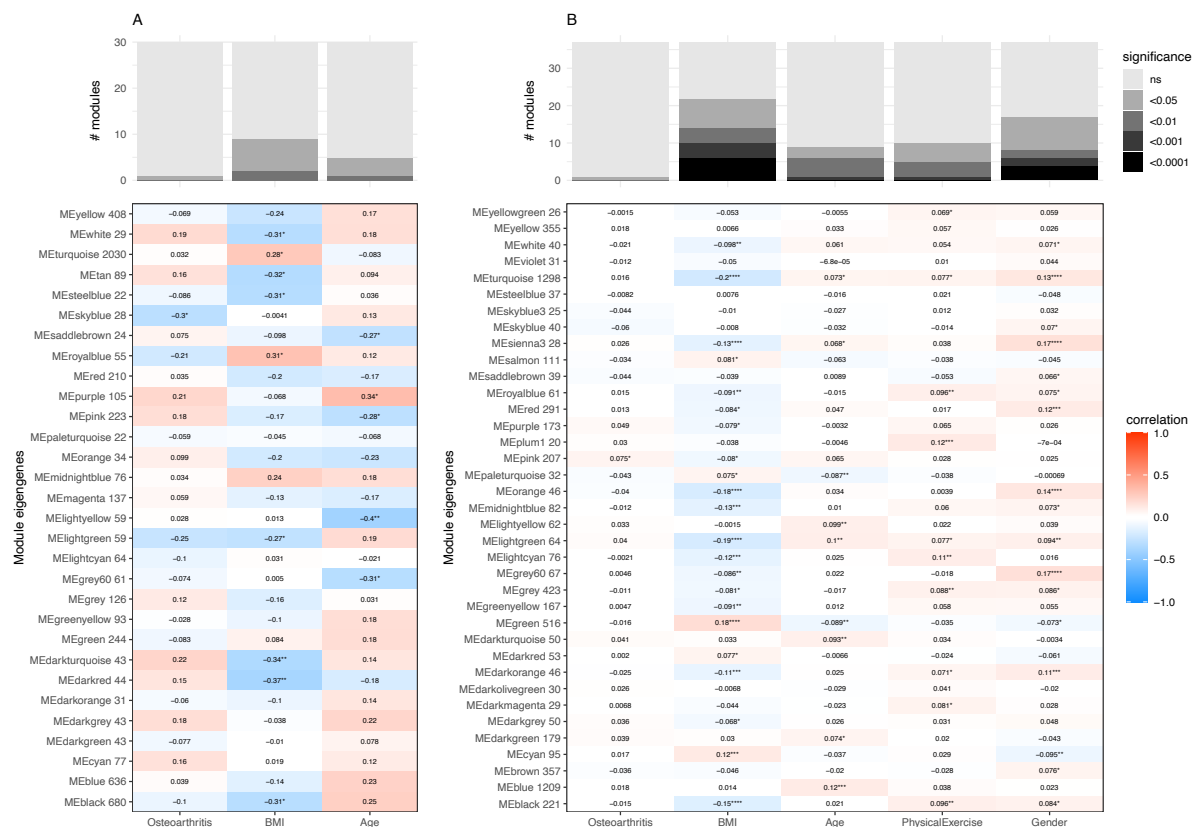
**Table 2.** Metagenome mappings to proteins, KEGG Orthologues (KO) and Gene Orthologues (GO).

	features	EstMB		TwinsUK	
proteins	12,383,493	1,962,482	16%	1,084,913	9%
KO	9,885	7,058	71%	6,570	66%
GO	15,406	10,424	68%	8,436	55%

The WGCNA revealed several KEGG Orthologues (KO) modules that were differently correlated with the investigated traits (Figure 3). More KO modules were associated with BMI than with sex. KO modules related to age and physical activity (PhysicalExercise) were also differentially distributed among participants, but to a lesser extent. KO modules related to OA were the least differentially distributed among participants. The observed patterns were similar across both studied cohorts; however, the effect was more pronounced in the EstMB cohort.

Only a single pathway was found to be significantly enriched with one of the traits. The pathway ko02040 ‘Flagellar assembly’ is significantly enriched ( $p_{adj}=1e-07$ ) for BMI in the TwinsUK cohort.

The correlation between traits and abundance profile of the KOs was, on average, significantly higher in the TwinsUK cohort than in the EstMB cohort. However, despite this, most correlation coefficients are not significant. It is noteworthy that OA shows the least association with the gut microbiome of all the traits investigated.



**Figure 3.** The heatmaps illustrate the Pearson correlation between module eigengenes and clinical traits. Panel A shows the results for the TwinsUK cohort, while panel B shows the results for the EstMB cohort. The KO modules are derived from an unsupervised weighted gene co-expression network analysis (WGCNA) of read counts on KEGG orthologues. The names of the modules are followed by the number of KEGG orthologues that are clustered together. There is no correspondence between the names of the modules identified in A and B. The values represented in the heatmap indicate the Pearson correlation coefficient, with the stars denoting the statistical significance (t-test), as follows: \*:  $p < 0.05$ ; \*\*:  $p < 0.01$ ; \*\*\*:  $p < 0.001$ ; \*\*\*\*:  $p < 0.0001$ . The bar plots displayed at the top illustrate the number of modules identified at each level of significance and for each trait.

### 3.3 SGNDB SIB GENE/NODE DIFFUSION-BASED ALGORITHM

#### 3.3.1 Algorithm considerations

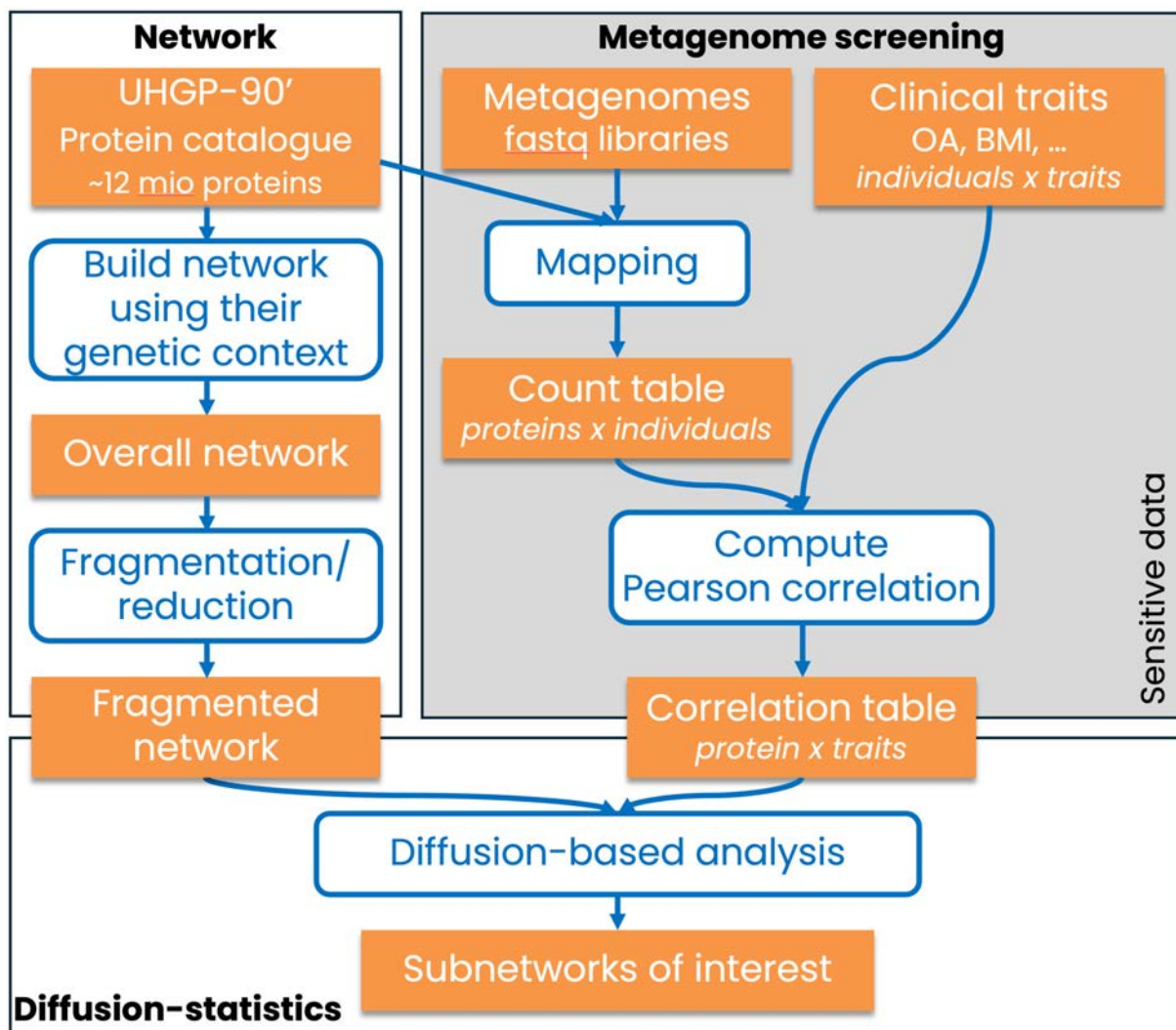
Typically, an association study between two multidimensional entities is conducted via an intermediate feature-by-sample count table. To perform a statistical analysis, the number of features must not exceed the number of samples; otherwise, there will not be sufficient statistical power. To ensure sufficient statistical power, features can be filtered out (as done for the KEGG and GO analyses) or aggregated (as done for the taxonomy, KEGG, and GO analyses). Therefore, an

absence of statistically significant results may be due to the filtering and aggregation rather than to the absence of a biological mechanism. In order to avoid preliminary filtering or aggregation and work directly at the protein level, we aimed to design a new method to search for subsets of genes that all together correlate with the clinical trait.

Moreover, we work with sensitive individual-level data, which must be analyzed on site in a secure computer infrastructure, as individual-level data cannot be exported to another infrastructure. This requires a portable analysis workflow to accommodate the different computer infrastructures and their security-related limitations.

These considerations led us to the following approach: To avoid statistical tests, we used a diffusion-based approach involving randomization to detect sets of genes of interest. Furthermore, we divided our framework into multiple parts, minimizing the computational burden on the host computer infrastructure by quickly aggregating individual-level data into non-sensitive Pearson correlation coefficients. The analysis workflow incorporates three parts, which are described in more details below (Figure 4):

1. **Networks preparation:** Based on the UHGG resource, a network representing the proximity of genes/proteins is created. This stage relies purely on public data and only needs to be set up once for all analyses.
2. **Metagenome screening:** Metagenomes are aligned to the representative protein catalogue, proteins are counted, and a Pearson correlation coefficient is computed between the counts and the clinical trait of interest (e.g., OA). This part of the workflow involves sensitive individual-level data and must therefore be computed on the host's computer infrastructure.
3. **Diffusion statistics:** Gene/protein network and the metagenome-trait correlation tables are combined via diffusion statistics to enable the search for sets of genes of interest. As it relies on the Pearson correlation coefficients of the metagenomic screening, this step does not involve any sensitive data and can therefore be carried out at SIB. This is important, as this step is heavily memory-bound computationally, something which not every computer infrastructure is made for.



**Figure 4.** Workflow showing the three parts of the SGNDDB approach network preparation, metagenomic screening and diffusion-statistics. In gray the sensitive part is shown which must be computed on the host computer infrastructure. The other two parts can be computed on any computer infrastructure. While the network construction is cohort-independent and only needs to be computed once, the other two steps are computed for each cohort.

### 3.3.2 Network preparation

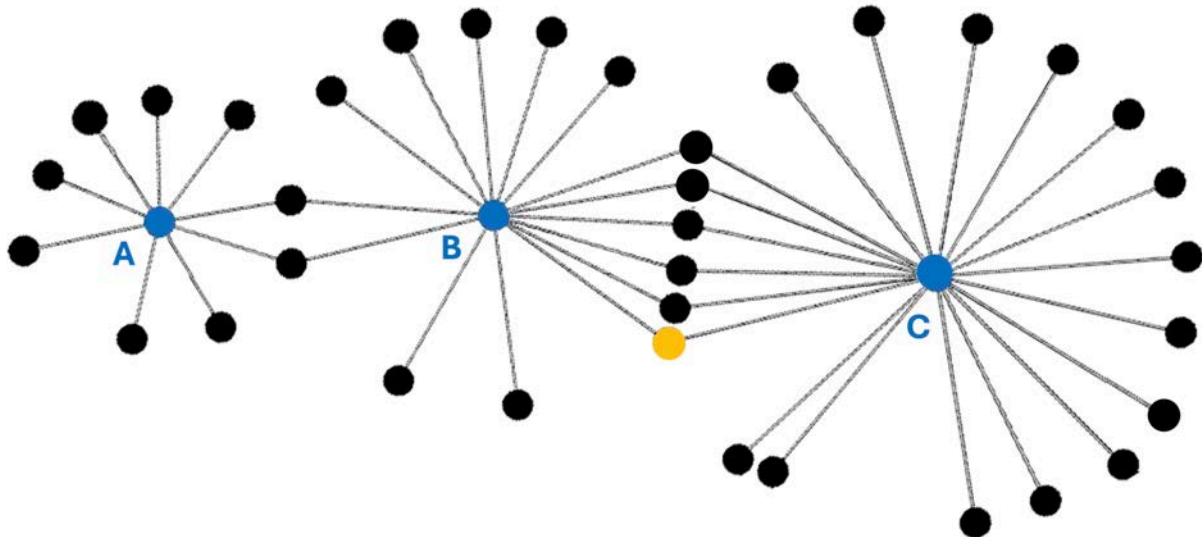
A large gene/protein network was built trying to capture the wide diversity of the pangenome, by grouping genes that are physically proximal in their source genomes, i.e., belonging to the same operon.

More precisely

1. For each UHGG-species a representative set of genomes and MAGs were selected for being as diverse as possible.

2. These genomes and MAGs were screened for adjacent UHGP-90' proteins lying on the same strand and being recovered multiple times.
3. Pairs of UHGP-90' proteins with shared proteins were grouped to operons, by connecting all UHGP-90' proteins to an UHGP-90' operon.
4. Across UHGGG-species, identical UHGP-90' proteins were combined
5. Operons with multiple shared proteins were further connected via a oset-A node to enforce the inter-species similarity.
6. Finally, the hierarchy of the taxonomy provided by EggNOG was used to further cluster the oset-A nodes via a oset-B node.

Please note that the network is purely based on the UHGG resource, and more specifically on the UHGP-90' protein dataset. Consequently, all derived nodes (proteins, operons, oset-A, and oset-B) are UHGP-90' specific. For readability, we will refer to these as proteins, operons, oset\_A, and oset\_B throughout the text. Furthermore, the network is composed of proteins; the corresponding genes were used solely to determine the physical proximity of these proteins and to group them into operons. The described network is an undirected and unweighted network representing the genomic proximity of genes (Figure 5).



**Figure 5.** Schematic representation of the protein network. The operons (blue) connect physically proximate proteins (black). Some proteins are shared between operons A and B or B and C. Operons with shared proteins originate by definition from different UHGG-species (e.g., operons B and C). The oset-A node (orange) enforces the inter-specific similarity between operons B and C.

The resulting overall network exceeded in size the technical limits required to run the diffusion-based algorithm on a HPC infrastructure. To reduce dimensionality,



uninformative nodes (i.e., highly connected nodes) were removed, and the network was recursively fragmented into subnetworks of fewer than 1,000 nodes by iteratively removing edges with the highest edge betweenness. This fragmentation strategy is expected to preserve the most densely connected parts of the original network.

The different nodes also contain different information (Table 3) which are important for the interpretation of the results.

**Table 3.** Network size statistics.

feature	fragmented network	annotation
subnetwork	746,900	
protein	6,145,426	function, often coarse if present
operon	2,062,601	Taxonomy to species level
oset-A	1,664,533	
oset-B	1,182,142	

### 3.3.3 Metagenome screening

In contrast to the functional analysis, where control groups were selected to match the positive cases, this analysis included all available participants: 2,509 metagenomes from the EstMB cohort, 249 metagenomes from the TwinsUK cohort, and, newly added, 7,072 metagenomes from the FINRISK cohort (Table 4). All were mapped to the UHGP-90' protein set.

The goal of the mapping was to use a very specificity alignment and not to maximize coverage. Not only the alignment parameters were set to maximize specificity, but also all reads aligning to more than 4 different proteins were discarded (Table 4). To compute a representative Pearson correlation for each representative protein between the counts of alignments per protein and the OA status sufficient alignments within and across participants are needed. We set the threshold to at least 10 alignments per protein in at least 10 participants. Prior to compute the Pearson correlation coefficient, the count table was normalized for library size and log transformed.

Alignment results indicate that the number of proteins with at least an alignment reflects the number of metagenomes analyses, ranging from 79% to 69% and 60%

of proteins for the FINRISK cohort, the EstMB cohort and TwinsUK cohort, respectively (Table 4). As expected, when filtering for a high enough abundant of alignments to compute the Pearson correlation coefficient, the vast majority of proteins were rejected due to too low number of mappings. The post-filtering numbers of usable proteins dropped to 16%, 9% and 6% for the EstMB, TwinsUK and FINRISK cohort, now reflecting the sequence coverage of the metagenomes.

Examining the mappings shows that the proportion of proteins with at least one alignment follows the number of metagenomes analyzed: 79% of proteins for the FINRISK cohort, 69% for the EstMB cohort, and 60% for the TwinsUK cohort (Table 4). As expected, when filtering for sufficient alignment counts and samples to calculate Pearson correlation coefficients led to the exclusion of most proteins, due to insufficient read coverage. After this filtering step, the proportion of usable proteins dropped to 16% for EstMB, 9% for TwinsUK, and 6% for FINRISK, now reflecting the effective sequence depth of the metagenomes.

**Table 4.** Characteristics of the three cohorts. The table lists the metagenomes investigated for each cohort, the mapping parameters used, and the resulting number of representative proteins retained. The mapping parameters were adapted to the characteristics of the data for each cohort in order to optimize the specificity of the mappings rather than the coverage. 'Expected usability' indicates the potential of the available data to detect an association with respect to the SGNDDB algorithm.

	EstMB	TwinsUK	FINRISK
<b>Cohort data</b>			
# participants	2,509	250	7,072
Public data	No	Yes	No
# OA (pos / neg / NA)	431 / 2,078 / 0	57 / 164 / 28	1,275 / 5,797 / 0
mean age per OA (pos / neg)	54.2 / 49.2	65.0 / 60.5	57.3 / 47.7
mean BMI per OA (pos / neg)	27.4 / 26.3	25.9 / 25.7	29.2 / 26.5
% female per OA (pos / neg)	69.1 / 70.6	100 / 100	60.4 / 53.5
# BMI (with value / NA)	2,509 / 0	250 / 1	7,072 / 2
mean BMI	26.5	25.9	27.0
<b>Sequence data generation</b>			
DNA extraction	stool-kit	unknown	gentle
sequencing	2x150bp	2x100bp	1x150bp
# reads per sample	15x10 <sup>6</sup>	36x10 <sup>6</sup>	1x10 <sup>6</sup>

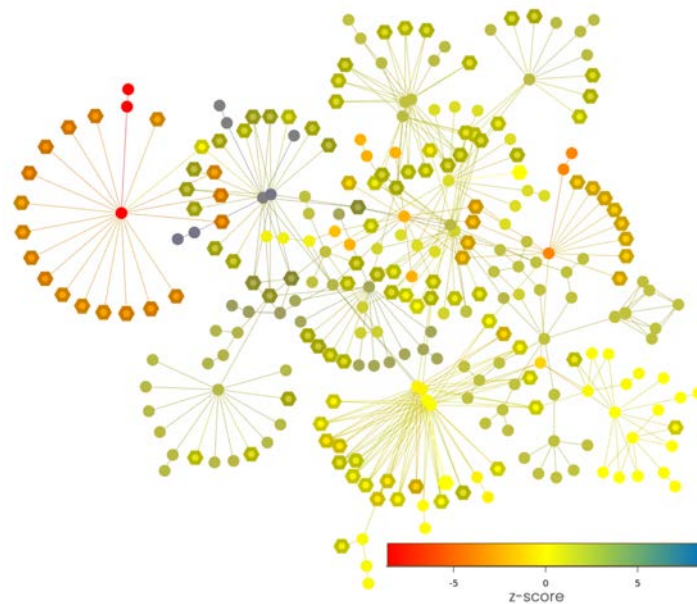


DIAMOND parameters						
min-orf	45		30		45	
min-score	75		50		75	
top	5		5		5	
Post-mapping filtering						
pair_read_constrained	Yes		Yes		No	
max-protein	4		4		4	
Protein hits (12,383,493 proteins in total)						
With at least 1 hit	8,519,744	69%	7,406,596	60%	9,782,050	79%
After filtering 10_10	1,962,482	16%	1,084,913	9%	739,713	6%
<b>Expected usability</b>	High specificity; good coverage		Good specificity; high coverage		Lower specificity; lower coverage	

### 3.3.4 Diffusion-statistics

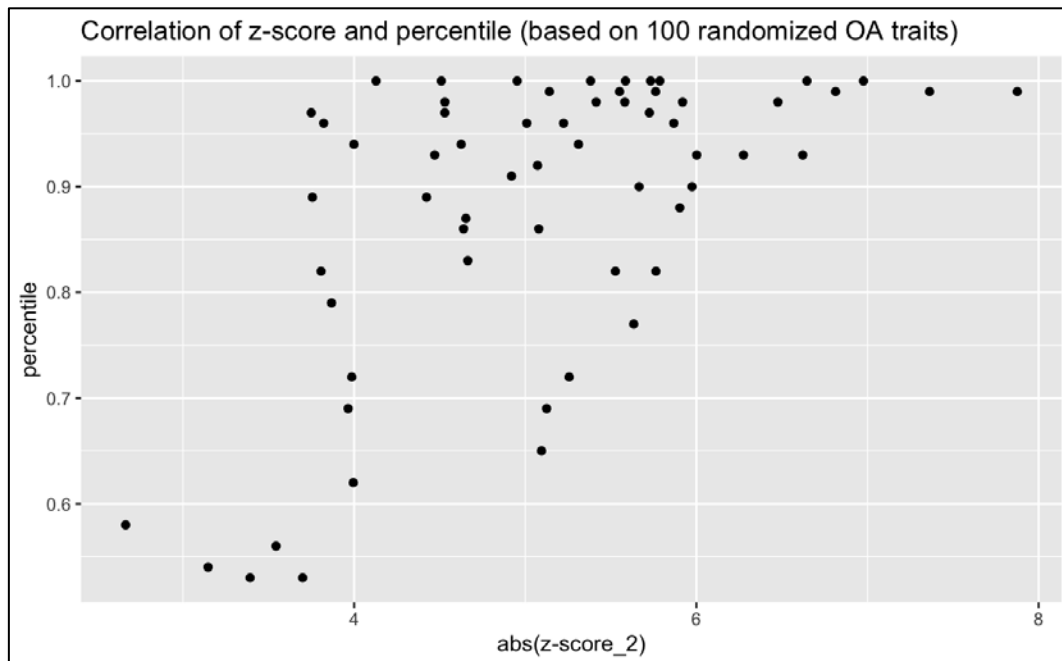
The correlation coefficients were treated as 'temperatures' within a diffusion-based model of the protein network. In this framework, the observed nodes were assigned fixed temperatures, which then diffused throughout the network according to the connectivity structure, gradually losing temperature in relation to 0. At equilibrium, each node thus attained a steady-state temperature. The diffusion model was applied only to subnetworks containing at least five observed proteins, substantially reducing the number of subnetworks requiring analysis.

To evaluate whether the observed temperature pattern could have emerged by chance, we used *diffustat* (Picart-Armada et al. 2018) to randomly permute the observed nodes and generate a null temperature distribution for every node. This enabled the calculation of a *Z-score* not only for the observed nodes but for all nodes in the network, quantifying how strongly each node deviated from the null expectation. Both the diffusion procedure and the *Z-score* computation are computationally intensive.



**Figure 6.** Top subnetwork example. This subnetwork contains multiple UHGP-90' operons, shown as the centers of the flower-like structures, with the UHGP-90' proteins depicted as petals. The hexagons represent the UHGP-90' proteins that were actually observed. In this real example, two adjacent operon structures exhibit top Z-scores—one strongly positive (blue) and one strongly negative (red). Because the positive operon has the larger absolute Z-score, we classify this subnetwork as a top positive subnetwork.

The Z-scores obtained from permuting the observations may be overly optimistic, as genes within genomes are not truly independent of one another. We therefore considered any node with an absolute Z-score greater than 6 to be a candidate associated with the clinical trait. It is important to note that the Z-score is not a *p*-value or a formal statistical test; rather it serves as an index of unexpectedness. Simulations using randomized observations showed that an absolute Z-score threshold of 6 identifies unexpected nodes with over 90% accuracy (Figure 7). Nodes with Z-scores above 6 are positively correlated with the trait and are referred to as “top positive” nodes, while those below  $-6$  are negatively correlated with the trait and referred to as “top negative” nodes. Subnetworks, i.e. fully connected subsets of nodes, with at least one top node are referred to as ‘top subnetwork’.



**Figure 7.** Distribution of OA Z-scores related to 100 randomized traits. For each randomized trait, participant OA status was shuffled without replacement. A percentile of 0.5 indicates that 50% of the randomized traits produced a larger Z-score than the observed OA trait, whereas a percentile of 1 indicates that the observed Z-score was not reached by any randomized trait. The plot demonstrates that applying a Z-score threshold of 6 enables the identification of unexpectedly high absolute Z-scores, expected to occur by chance in less than 10% of cases.

### 3.4 SGNDB ANALYSIS OF OA AND BMI

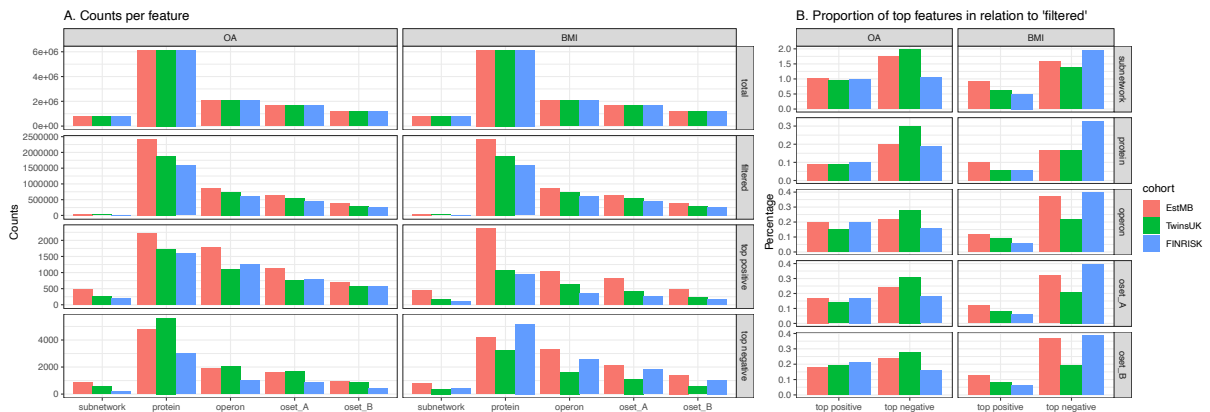
The SIB applied the SGNDB approach to OA and BMI. The two measurements are present for all three cohorts and thus allows a comparison across the cohorts and trait. The SGNDB approach enabled SIB to identify sets of proteins that seems to be associated with OA and BMI.

The diffusion statistics revealed that only a small proportion of subnetworks (0.5% to 1.99%) and even fewer nodes (0.06% to 0.4%) exhibited high positive or negative Z-scores (Table 5 / Figure 8).

**Table 5.** Top nodes and subnetwork count for OA and BMI. The table summarizes the number of subnetworks (network) and nodes (proteins, operons, oset-A, and oset-B). The “fragment network” column reports the overall counts identical to all cohorts. The “filtered” column shows the corresponding counts after restricting the fragmented network to at least 5 proteins observed per subnetwork and used for the diffusion-based analysis. The “top positive” and “top negative” columns indicate the numbers of top features with positive

and negative Z-scores, respectively. The percentages of top nodes are in relation to the features entering the diffusion-based approach (i.e. column 'filtered').

cohort	type	fragmented network		OA				BMI				
		filtered	top positive	top negative	filtered	top positive	top negative	filtered	top positive	top negative		
EstMB	subnetwork	746,900	49,118	494	1.01%	864	1.76%	49,118	447	0.91%	775	1.58%
	protein	6,145,426	2,416,195	2,226	0.09%	4,783	0.20%	2,416,195	2,364	0.10%	4,172	0.17%
	operon	2,062,601	873,777	1,775	0.20%	1,892	0.22%	873,777	1,032	0.12%	3,275	0.37%
	oset-A	1,664,533	658,093	1,132	0.17%	1,588	0.24%	658,093	819	0.12%	2,125	0.32%
	oset-B	1,182,142	381,549	702	0.18%	933	0.24%	381,549	483	0.13%	1,408	0.37%
TwinsUK	subnetwork	746,900	27,501	258	0.94%	548	1.99%	27,501	169	0.61%	380	1.38%
	protein	6,145,426	1,882,178	1,718	0.09%	5,630	0.30%	1,882,178	1,081	0.06%	3,206	0.17%
	operon	2,062,601	731,381	1,101	0.15%	2,036	0.28%	731,381	633	0.09%	1,631	0.22%
	oset-A	1,664,533	539,731	764	0.14%	1,666	0.31%	539,731	426	0.08%	1,121	0.21%
	oset-B	1,182,142	311,149	589	0.19%	867	0.28%	311,149	238	0.08%	587	0.19%
FINRISK	subnetwork	746,900	20,958	208	0.99%	217	1.04%	20,958	104	0.50%	409	1.95%
	protein	6,145,426	1,582,338	1,597	0.10%	3,002	0.19%	1,582,338	937	0.06%	5,182	0.33%
	operon	2,062,601	628,350	1,261	0.20%	1,023	0.16%	628,350	372	0.06%	2,538	0.40%
	oset-A	1,664,533	462,194	782	0.17%	833	0.18%	462,194	265	0.06%	1,845	0.40%
	oset-B	1,182,142	265,798	563	0.21%	423	0.16%	265,798	166	0.06%	1,046	0.39%



**Figure 8.** The figure visualizes table 5. A shows the counts for each type of feature, while B shows the proportion of the positive and negative top features.

### 3.4.1 Analyses per cohort

We next focused on the UHGP 90' operons. Taking advantage of their taxonomic annotations, we were able to investigate taxonomic composition at different taxonomic ranks and between different sets. It is important to note that the composition of taxa does not change a lot when moving from the fragmented network to the filtered network of each cohort (results not shown).

We examined which UHGG species were represented among the top positive and top negative operons (Figures 9 and 10). Among the 3,895 UHGG species present in the fragmented network, only 118 species had at least one analysis with more than 20 top operons. Figure 9 summarizes the top operons aggregated at various taxonomic ranks. At higher taxonomic levels, the proportions of top negative and top positive operons within a given trait and cohort are often similar. In contrast, as the taxonomic resolution becomes more specific, the imbalance becomes more pronounced, with clearer enrichment of either top positive or top negative operons. This pattern is evident at the genus level and is even more distinct at the species level (Figure 10).

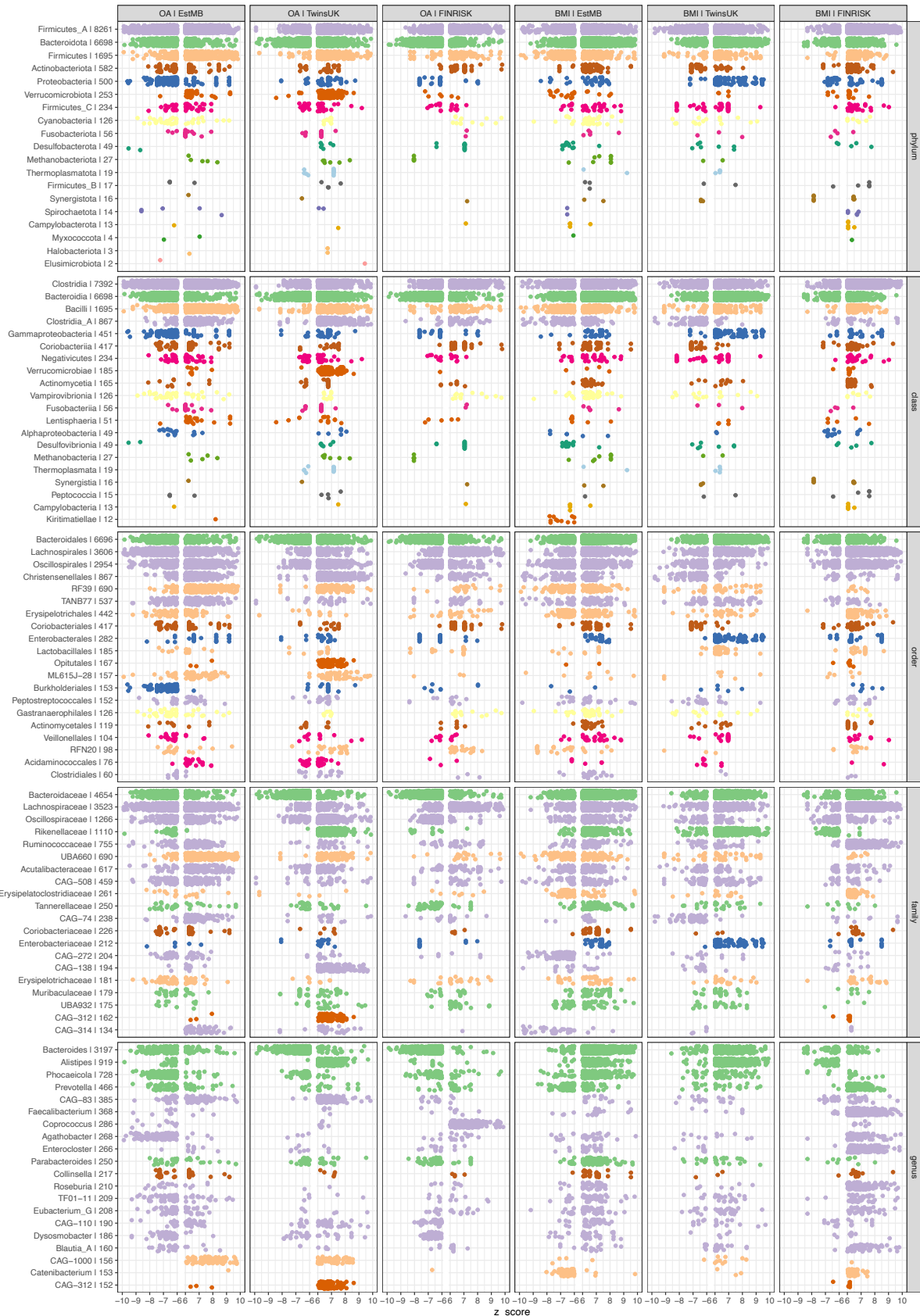


Figure 9. Top operons per taxa at different ranks.

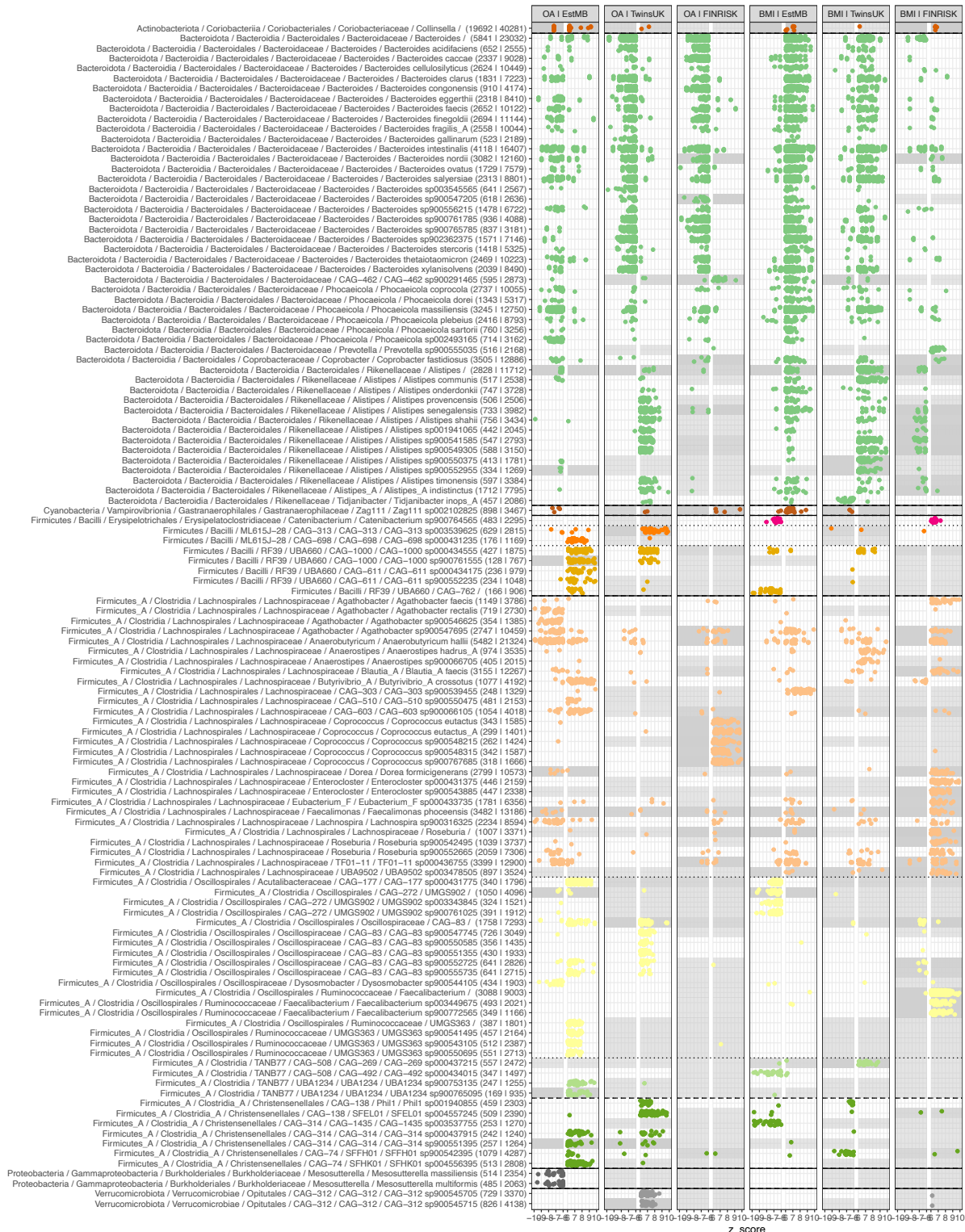
For each taxonomic rank the taxa with the most top operons (listed behind the label). Each panel shows the Z-score of the top operons of the most abundant taxa associated with OA or BMI in EstMB, TwinsUK and FINRISK cohorts. The color refers to the phylum rank.

A big advantage of the SGNDB approach is that it allows us to distinguish whether a species fails to appear because no biological association exists, or because there is insufficient statistical power, due to low sequencing coverage, to detect a potential association. As a proxy for power, we use the proportion of observed UHGP-90'-proteins: the higher the proportion of proteins with adequate read coverage to compute a Pearson correlation (i.e., observed UHGP-90'-protein), the higher the probability of detecting an association.

In Figure 10, this power is visualized through a grayscale background. Darker shading indicates lower power (i.e., lower proportion of observed proteins), while lighter shading indicates higher power. When at least 50% of the proteins associated with a species are observed, the background is shown in white.

In Figure 10 we see that for the top 118 UHGG-species that the power to detect a potential association is lowest for the FINRISK cohort (largest gray parts, followed by TwinsUK and EstMB where most UHGG-species have at least half of the associated UHGP-90'-proteins observed).

For some species, the OA and BMI traits show correlated patterns, whereas for others they are inversely related. For instance, several *Bacteroidaceae* species rank among the top UHGG-species and typically exhibit negative Z-scores, and thus negative associations, with OA, while showing positive Z-scores, and thus positive associations, with BMI. Notably, in the EstMB cohort, the number of top operons identified for *Bacteroidaceae* in relation to OA is lower than in the other two cohorts. In contrast, for BMI, EstMB displays a large number of top positive operons, more than those detected in the other two cohorts.



**Figure 10.** Top operons per taxa at species level. Same as Figure 9, but at the species level. All UHGG-species are listed with at least 20 top operons—either with positive or negative Z-scores—in any cohort or trait are included. The gray background indicates the detection “power” for a potential top operon: the darker the shading, the fewer proteins were observed

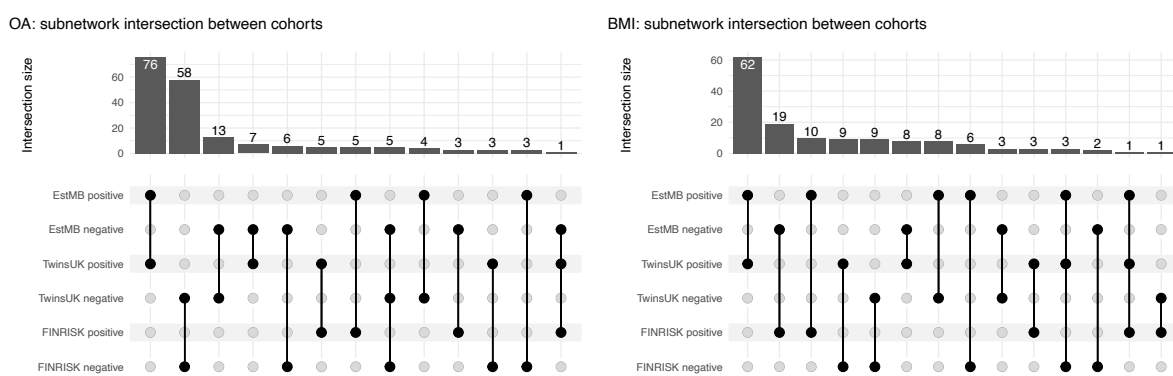


for that species in the cohort, making it more difficult to detect an association. A white background indicates high detection potential, meaning at least 50% of the associated proteins were observed. Taxonomic labels on the left show the full taxonomy (starting at the phylum), followed in brackets by the number of operons and proteins associated with that UHGG-species. Colors indicate the taxonomic rank 'order'. Solid lines separate phyla, dashed lines classes, and pointed lines families.

### 3.4.2 Comparing cohorts

Despite substantial differences in participant characteristics and sequencing strategies between EstMB, TwinsUK and FINRISK cohorts (Table 4), we were interested in identifying features that consistently appeared across multiple cohorts. To this end, we searched for nodes, operons, and subnetworks that displayed high Z-scores in more than one cohort.

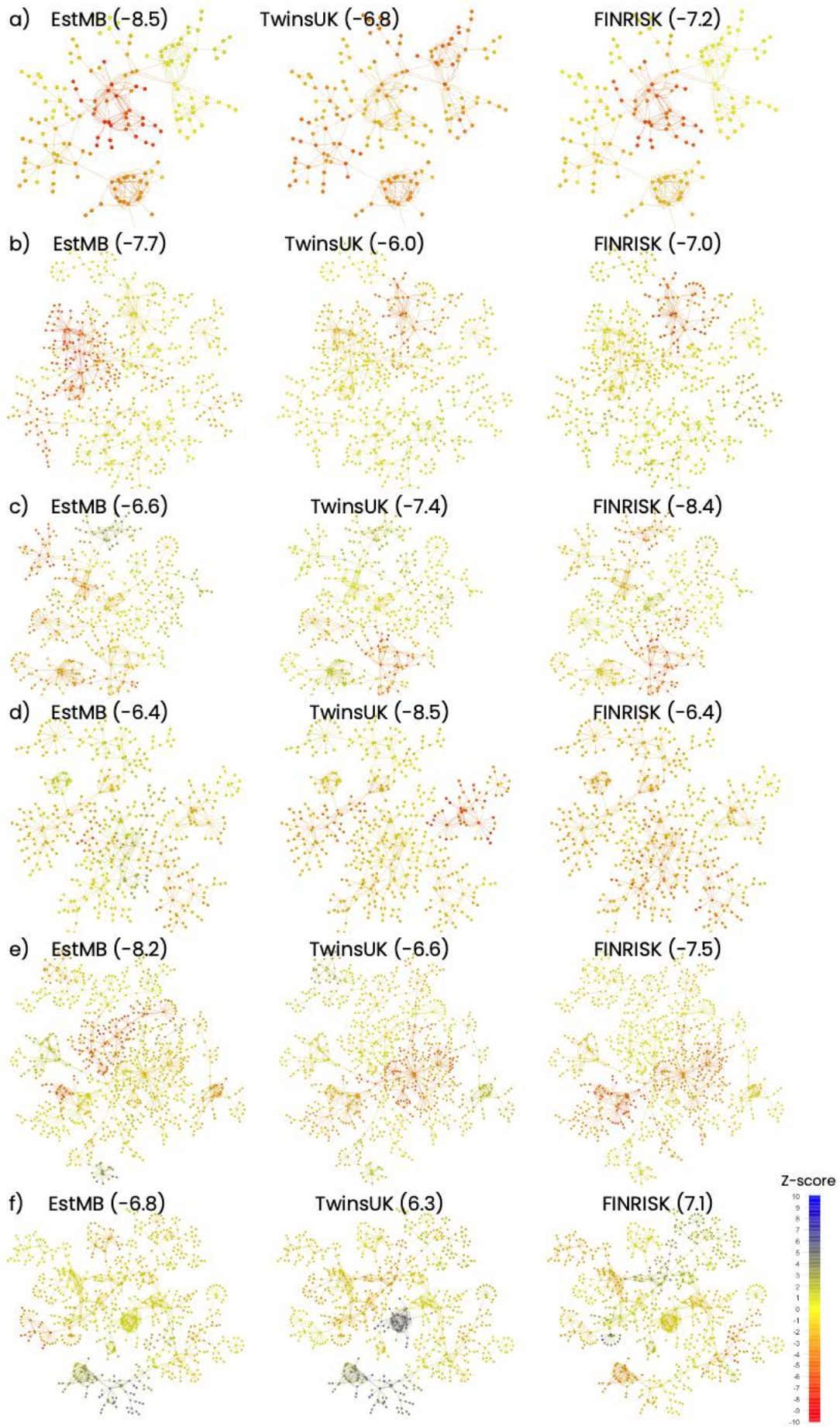
We first assessed the overlap of top subnetworks between cohorts (Figure 11). For OA, six subnetworks were identified as top features. Five of these showed negative Z-scores in all three cohorts, while one exhibited a positive Z-score in the TwinsUK and FINRISK cohorts but a negative Z-score in the EstMB cohort. For BMI, four subnetworks were identified as top features: one demonstrated consistent Z-scores across all three cohorts, whereas the remaining three showed positive Z-scores in the EstMB and TwinsUK cohorts but negative Z-scores in the FINRISK cohort (Figure 11).



**Figure 11.** Intersections of subnetworks across cohorts for OA (left panel) and BMI (right panel) for top positive and top negative subnetworks.

When examining the six top subnetworks associated with OA (see Figure 11) more in detail, we observed that their sizes vary widely, ranging from 168 to 829 nodes

(Figure 12). Most operon structures display low *Z-scores*, with only one or two operon structures in each subnetwork standing out—either positively or negatively. These subnetworks rank among the top in all three cohorts. Although the same operon structure occasionally have the same top hits across cohorts (*correlated Z-scores across cohorts*), more commonly different operon structures emerge as top hits in each cohort, and in some cases these topo operons differ not only in identity but also in the direction of their association (not shown).

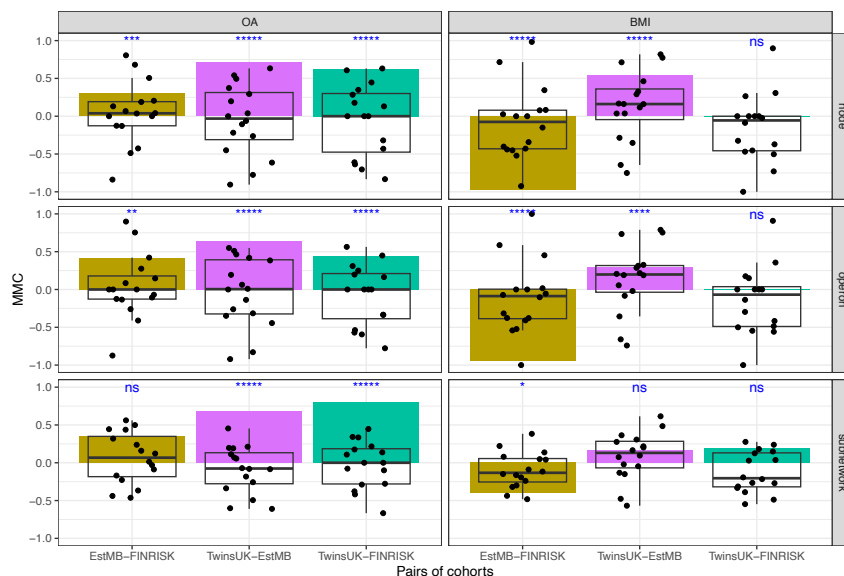




**Figure 12.** Top six subnetworks associated with OA in all cohorts. Nodes color represents the Z-score. The Z-score of the subnetwork is written in brackets behind the cohort's name. The subnetworks are: a) net\_22813 with 168 nodes; b) net\_10018 with 564 nodes; c) net\_10830 with 448 nodes; d) net\_22003 with 461 nodes; e) net\_22903 with 829 nodes; f) net\_4398 with 743 nodes. Subnetworks a) to e) all appear in all cohorts with a top positive Z-score; subnetwork f) has a top positive Z-score in TwinsUK and FINRISK, and a top negative Z-score in EstMB.

### 3.4.2.1 Matthews Correlation Coefficient

To collect evidence of results reproducibility across cohorts, we conducted a pairwise cohort comparison. For each pair of cohorts, we counted how many top nodes, operons, and subnetworks shared the same sign (positive-positive or negative-negative) versus the opposite sign (positive-negative or negative-positive). These counts were summarized in a contingency table and concordance of signs was evaluated using the Matthews Correlation Coefficient (MCC; Chicco and Jurman 2023).



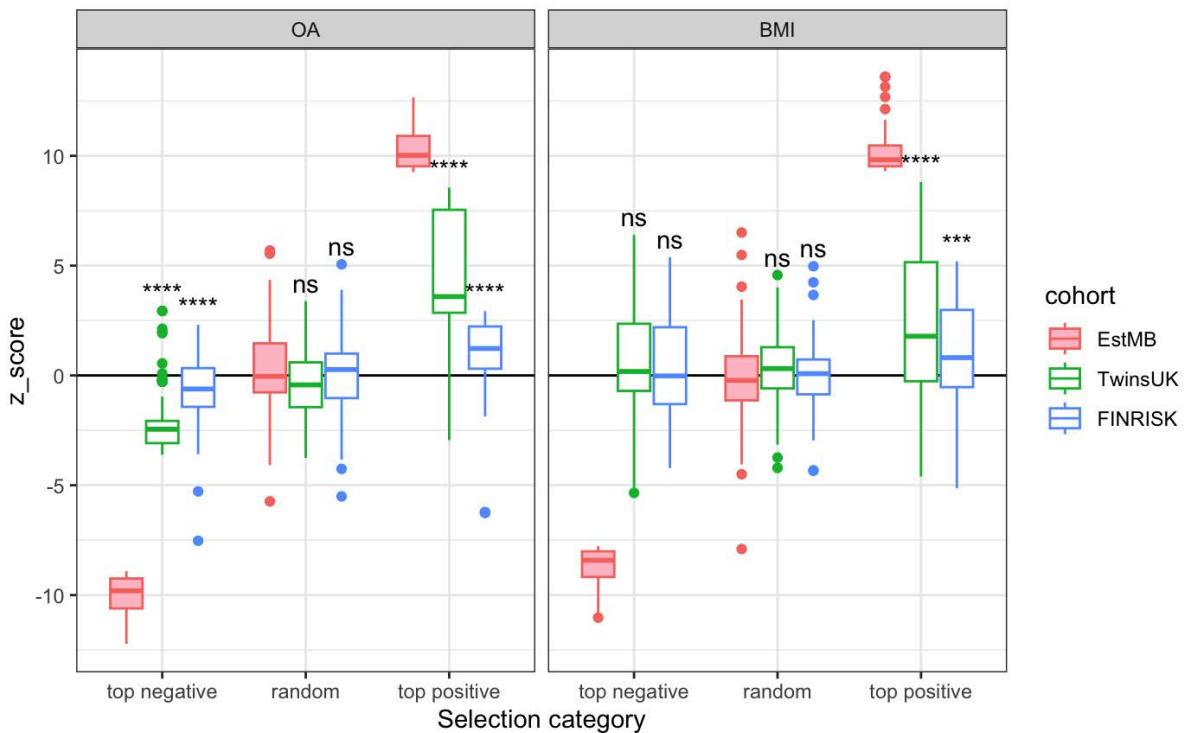
**Figure 13.** Matthews correlation coefficient (MCC) between pairs of cohorts. A value of 1 indicates that all top features share the same sign, -1 indicates that all have opposite signs, and 0 reflects a random distribution of positive and negative signs. The color represents the pair of cohorts. The boxplots and black points represent the distribution of MCC values obtained from randomized traits. The stars and 'ns' represent the p-value of the observed MCC (ns: not significant, \*: <0.05; \*\*: <0.01; \*\*\*: <0.001; \*\*\*\*<0.0001; \*\*\*\*\*<0.00001)

The analysis revealed that, for OA, top nodes, operons, and subnetworks were reproducible across all cohort pairs, notably for the TwinsUK-EstMB and TwinsUK-FINRISK comparisons (Figure 13). In contrast, the EstMB-FINRISK pair showed weaker

or even non-significant agreement at the subnetwork level. The pattern for BMI differed substantially: EstMB–FINRISK shows a strong and highly significant inverse correlation, TwinsUK–FINRISK shows no correlation, and TwinsUK–EstMB shows a positive correlation at the node and operon levels. In other words, while OA appears to be driven by similar underlying forces across all cohorts, BMI seems to be influenced by different forces in each cohort.

### 3.4.2.2 Cross-cohort validation

Another way to assess whether findings from one cohort replicate in another is to identify the top operons based on their *Z-scores* in a reference cohort, then examine the corresponding *Z-scores* of those operons in the other cohorts. Using the most reliable dataset, the EstMB cohort, as the predictor, we selected the top 100 positive or top 100 negative operons. As a control, 100 additional operons were chosen at random. For OA, the corresponding *Z-scores* in the TwinsUK and FINRISK cohorts were significantly different from zero and aligned in direction with those from EstMB (Figure 14). While the effect sizes were smaller in TwinsUK and substantially attenuated in FINRISK, the overall pattern indicates that the EstMB findings are reproducible in the other cohorts, with stronger replication in TwinsUK than in FINRISK. For BMI, the same pattern can be observed for the top positive selection, although at a lower level. For the top negative selection, no such an association exists.



**Figure 14.** Cross-cohort validation test. Three sets of 100 operons were selected based on their Z-score in the EstMB cohort (red filled boxplots): top negative, top positive and random. Corresponding operon Z-scores are showed for TwinsUK (green) and FINRISK (blue) cohorts. The left panel shows the cross-validation for OA and the right panel for the BMI. A one-sample t-test was used to test whether the z-scores in the TwinsUK and FINRISK cohorts deviate significantly from zero (ns: not significant; \*\*\*: p-value<0.001; \*\*\*\*: p-value<0.0001).

Both cross-cohort comparisons indicate that OA-related patterns are more reproducible across cohorts than BMI-related ones.

## 4 DISCUSSION

Both EstMB and TwinsUK cohorts were sequenced using paired end reads with similar sequencing depth. FINRISK was sequenced using single reads, shallow depth and its DNA was prepared using a protocol less efficient to recover DNA from “hard” bacteria like Firmicutes. We had *a priori* considered FINRISK as of lower usability, i.e. result obtained for FINRISK should be considered with great caution. The points discussed below primarily refer to EstMB and TwinsUK cohorts, although they are often supported by FINRISK as well.

The initial working hypothesis of ENDOTARGET was that endotoxins, AKA lipopolysaccharides (LPS), produced by Proteobacteria in the gut, cross the intestinal barrier and cause inflammation in joints. No strong experimental evidence for this specific claim was recently published. The analysis of metagenomes performed by the ETHZ and the SIB don't support that specific claim either. But this does not mean that the so-called gut-joint axis hypothesis is untrue, as other classes of molecules and other bacterial taxa could be responsible for it. Indeed, there is some evidence (especially from animal models) that the Bacteroidetes / Firmicutes (B/F) ratio is altered in osteoarthritis (OA), but the picture is nuanced, and human data are still limited. Our results permit to greatly refine the role of Bacteroidetes and Firmicutes with respect to the gut-joint axis.

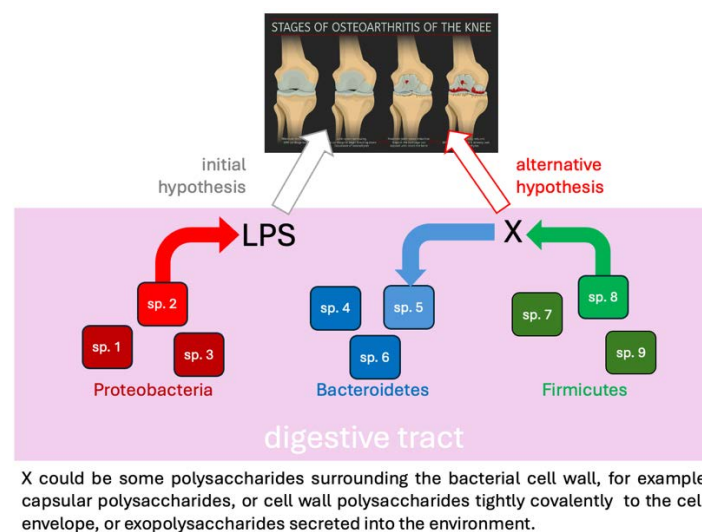
A first cycle of investigations by ETHZ, later complemented by SIB was carried on four cohorts, using classical, "aggregative" statistics, i.e. statistics that regroup (sum) multiples genes to document a limited number of high-level taxa or broad functions. Despite a lot of precaution to remove confounding variables prior to statistical analysis, the stronger detected signal was BMI, to which OA is known to be correlated. Identifying bacterial taxa specifically correlated with OA could not be achieved in this setting. These preliminary observations have motivated SIB to design a new method, named SGNDB to search for low-level taxa that could be positively or negatively correlated with OA, specifically. The key idea was to avoid the early aggregation of the experimental data. In this perspective SGNDB appears more as a "search" method, than as a traditional statistical method.

SGNDB reveals that some species are **negatively** correlated with OA, and simultaneously positively correlated with BMI in the three investigated cohorts. They are all members of the genera *Bacteroides* and *Phocaeicola*, both in family Bacteroidaceae. The species overlap is however only partial among cohorts. The three cohorts are distinct with respect to age, sex, and BMI distributions. Important biases linked to the different lifestyle in the different countries, and to the different genetic backgrounds are to be expected. These factors are likely to influence the most abundant Bacteroidaceae species within a cohort, which will be revealed by sequencing and later detected by SGNDB. A protective effect by some Bacteroidaceae seems supported by a recent paper reporting on a mouse model of OA (Liu et al. 2025)

Some Firmicute species were observed to be **positively** correlated with OA in both EstMB and TwinsUK. They were not observed in FINRISK, possibly due to the "gentle" DNA extraction method used and shallow sequencing. These Firmicute species belongs to the RF39 order (Bacilli), Oscillospirales (Clostridia), and

Christensenellales (Clostridia). They are all uncultivated species which biology is poorly known, but they were previously reported from many metagenomes.

Figure 15 presents our current working hypothesis centered around a class of compound X, distinct from lipopolysaccharides (LPS) and not yet characterized. The key idea is that its production by some Firmicutes species is positively correlated with OA, and its degradation, transformation by some Bacteroidetes species exert a protective effect for OA. Although X is unknown, we report candidates for the involved species and this result is supported by three different cohorts.



**Figure 15.** Graphical summary of the discussion about microbiology

Bacterial taxonomy traditionally relies on the 16S rRNA gene and a small set of conserved phylogenetic markers to infer evolutionary relationships. However, bacterial genomes are highly dynamic, evolving rapidly even within what is considered a single species through genome rearrangements and horizontal gene transfer (HGT). As a result, a species cannot be represented by a single static genome but rather by its *pan-genome*—the collective set of all genes found across its strains, encompassing both core and accessory genetic elements. Metabolic features such as antibiotic resistance, specific substrate utilization, or pathogenicity are often restricted to a few strains within a species rather than being shared by all its member strains. Consequently, the species designation alone is a poor predictor of functional capabilities, as key metabolic traits can vary widely across strains depending on the presence or absence of accessory genes acquired through horizontal gene transfer or other genomic rearrangements. Moreover, different bacterial species may have strains that carry the same metabolic function. Such functional convergence, often driven by horizontal gene transfer, blurs the link between taxonomic boundaries and metabolic capabilities.

The SGNDB method was developed with these ideas in mind. Operon structures contain a lot of functional significance, despite most bacterial genes are poorly annotated. The procedure to fragment and reduce the overall network into subnetworks was devised to meet technical constraints of the diffusion algorithm, but also to regroup similar operons from different species (see examples in Figure 15). It seems to be a successful attempt. There are a lot of possible optimizations and improvements to the SGNDB method, that should be readily applicable to other clinical traits.

## 5 REFERENCES

- Almeida, A., S. Nayfach, M. Boland, F. Strozzi, M. Beracochea *et al.*, 2021 A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology* 39 (1):105–114.
- Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler *et al.*, 2000 Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25 (1):25–29.
- Bevc, K., L. Malfertheiner, S. Neuenschwander, V.D. Tran, M. Pagni *et al.*, 2025 Microbiome variations in osteoarthritis reflect aging and metabolic factors, not the disease. *bioRxiv*:2025.2006.2024.661261.
- Buchfink, B., C. Xie, and D.H. Huson, 2015 Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12 (1):59–60.
- Chicco, D., and G. Jurman, 2023 The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Min* 16 (1):4.
- Galperin, M.Y., R. Vera Alvarez, S. Karamycheva, K.S. Makarova, Y.I. Wolf *et al.*, 2025 COG database update 2024. *Nucleic Acids Res* 53 (D1):D356–D363.
- Hernandez-Plaza, A., D. Szklarczyk, J. Botas, C.P. Cantalapiedra, J. Giner-Lamia *et al.*, 2023 eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Res* 51 (D1):D389–D394.
- Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, 2016 KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 44 (D1):D457–462.
- Langfelder, P., and S. Horvath, 2008 WGCNA: an R package for weighted correlation network analysis. *Bmc Bioinformatics* 9:559.

- Liu, S., H. Xu, L. Liu, W. Ma, H. Fan *et al.*, 2025 Gut microbiome dysbiosis accelerates osteoarthritis progression by inducing IFP-SM inflammation in "double-hit" mice. *Arthritis Res Ther* 27 (1):137.
- Matias Rodrigues, J.F., T.S.B. Schmidt, J. Tackmann, and C. von Mering, 2017 MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. *Bioinformatics* 33 (23):3808–3810.
- Mulder, N., and R. Apweiler, 2007 InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol Biol* 396:59–70.
- Picart-Armada, S., W.K. Thompson, A. Buil, and A. Perera-Lluna, 2018 diffuStats: an R package to compute diffusion-based scores on biological networks. *Bioinformatics* 34 (3):533–534.
- Richardson, L., B. Allen, G. Baldi, M. Beracochea, M.L. Bileschi *et al.*, 2023 MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res* 51 (D1):D753–D759.
- Ruscheweyh, H.J., A. Milanese, L. Paoli, N. Karcher, Q. Clayssen *et al.*, 2022 Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments. *Microbiome* 10 (1):212.
- Wirbel, J., K. Zych, M. Essex, N. Karcher, E. Kartal *et al.*, 2021 Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biol* 22 (1):93.