Exploring the Interplay Between Fibre Intake, Exercise, and Gut Microbiota in Modulating Cardiovascular Health in a Westernization Context

Rui Xiang Yu, Houria Afshar Moghaddam, Brooke Macleod, Quinlan Torstensen

Table of contents

1 Introduction

This report contains the methods and results generated throughout the making of this project. Our research question is:

In a westernization context, how do dietary fibre intake and exercise affect cardiometabolic health, and does this associate with unique microbiome profiles?

Our aims are:

1. Investigate the optimal predictors for cardiovascular health, such as dietary fibre and exercise.
2. Compare diversity metrics of groups with adequate and inadequate dietary fibre intake, and high and low exercise levels in relation to cardiometabolic health.
3. Identify taxa that are differentially abundant and co-occur among the different fibre intake and exercise groups, as well as cardiovascular condition.

2 Methods

2.1 Data processing

The dataset was sourced from de la Cuesta-Zuluaga et al. 2018's paper, whose raw DNA FASTQ files can be found at the SRA-NCBI under BioProject PRJNA417579 (Cuesta-Zuluaga

et al. (2018)). The dataset comprises the sequences corresponding to the V4 hypervariable region of the 16S rRNA gene, collected in 2014 (Cuesta-Zuluaga et al. (2018)). The primers used were F515 and R806 and sequencing was done with Illumina MiSeq (Cuesta-Zuluaga et al. (2018)). There are 441 samples of men and women, across a wide range of ages and body mass indices, that lived in different Colombian cities (Bogotá, Medellín, Cali, Barranquilla, Bucaramanga). None of the participants were underweight (Body Mass Index <18.5 kg/m^2), pregnant, consumed antibiotics or antiparasitics three months before sample collection, with cancer, with gastrointestinal diseases, or with neurodegenerative diseases (Cuesta-Zuluaga et al. (2018)).

The accompanying metadata for each patient consists of multiple demographic parameters (age, biological sex, city of residence), anthropometric measures (body mass index, body fat percentage, waist circumference), lipid profiles in blood (HDL, LDL, adiponectin), macronutrient consumption, blood pressure, glucose metabolism profiles, and stool consistencies. A full list of available metadata with each variable's range or factors can be found in Table **??**.

Table 1: Table of the multiple metadata information available for each patient. For categorical variables, it contains all factors. For numerical variables, it contains the numerical range of available values.

Variable	Range_or_Categories
adiponectin	0 – 28.21
age_years	18 – 62
age_range	18_40, 41_62
BMI	18.6 – 47.4
BMI_class	Lean, Obese, Overweight
Body_Fat_Percentage	18.7 – 48.7
Calorie_intake	634 – 4034
Cardiometabolic_status	Abnormal, Healthy
city	Barranquilla, Bogota, Bucaramanga, Cali, Medellin
country	Colombia
diastolic_bp	50 – 126
fiber	7 – 44
glucose	64 – 335
Hemoglobin_a1c	4.6 – 10.77
CRP	0.12 – 44.3
insulin	1.95 – 57.07
latitude	3.42 – 10.96
Total_Cholesterol	67 – 302
HDL	11 – 134
LDL	30 – 219
VLDL	6 – 218
Triglycerides	28 – 1090

Table 1: Table of the multiple metadata information available for each patient. For categorical variables, it contains all factors. For numerical variables, it contains the numerical range of available values.

Variable	Range_or_Categories
medication	No, Yes
per_carbohydrates	45.89 – 64.87
per_total_protein	11.95 – 21.11
per_total_fat	21.51 – 36.83
per_animal_protein	39.41 – 74.48
per_monoinsaturated_fat	6.85 – 12.55
per_polyunsaturated_fat	3.48 – 7.86
per_saturated_fat	7.46 – 16.06
sex	female, male
smoker	No, Yes
stool_consistency	Diarrheic, Hard, Normal, soft
systolic_bp	76 – 204
MET_mins_per_week	0 – 45204.6
waist_circumference	65.2 – 131.3

The data was processed with QIIME2 version 2025.4.0 (Bolyen et al. (2019)). The script to import the data and demultiplex the sequences can be found in `bin/01-qiime2_data_processing.sh`. The maximum read length was 251 nucleotides (nts). Based on a random subsample of 10,000 reads, 98% of reads have a read length of 251 nts, while 2% have a read length of 250 nts. The maximum number of reads in a sample was 117,562, whereas the minimum number was 4,305. The mean number of reads was 40,657.89.

The script for filtering and downstream QIIME2 processes are in `bin/02-qiime2_data_filtering.sh`. The denoising tool chosen was DADA2 (Callahan et al. (n.d.)). The Phred quality score threshold chosen was 30. The base pair at position 251's median quality was 29. A truncation length of 250 was chosen. After denoising, all 441 samples were still retained. The maximum number of reads in a sample changed to 106,116, whereas the minimum number changed to 3,665.

With QIIME2's taxonomic classifier, amplicon sequence variants (ASVs) were classified with a pre-trained Naïve Bayes model built on the SILVA version 138 99% OTUs, trained for the primer pair 515F/806R, which targets the V4 region of the 16S rRNA gene. A phylogenetic tree was subsequently generated by aligning multiple sequences with MAFFT, masking hypervariable regions, constructing an approximately maximum-likelihood tree with FastTree2, and applying midpoint rooting to obtain a rooted phylogeny for downstream analyses.

Afterwards, sequences corresponding to mitochondria or chloroplasts, as well as non-bacterial sequences were filtered out. Based on the subsequently generated rarefaction curve (Figure **??**),

ASVs saturate at around ~10,000 reads for most samples. A cutoff of 24,406 reads was chosen as this retained 79.49% of the original number of features (8,705 out of 10,951) and 333 samples. Thus, 108 samples were discarded. No batch correction was applied to the data as de la Cuesta-Zuluaga et al. found no significant difference across runs and an internal control was also present (Cuesta-Zuluaga et al. (2018)).

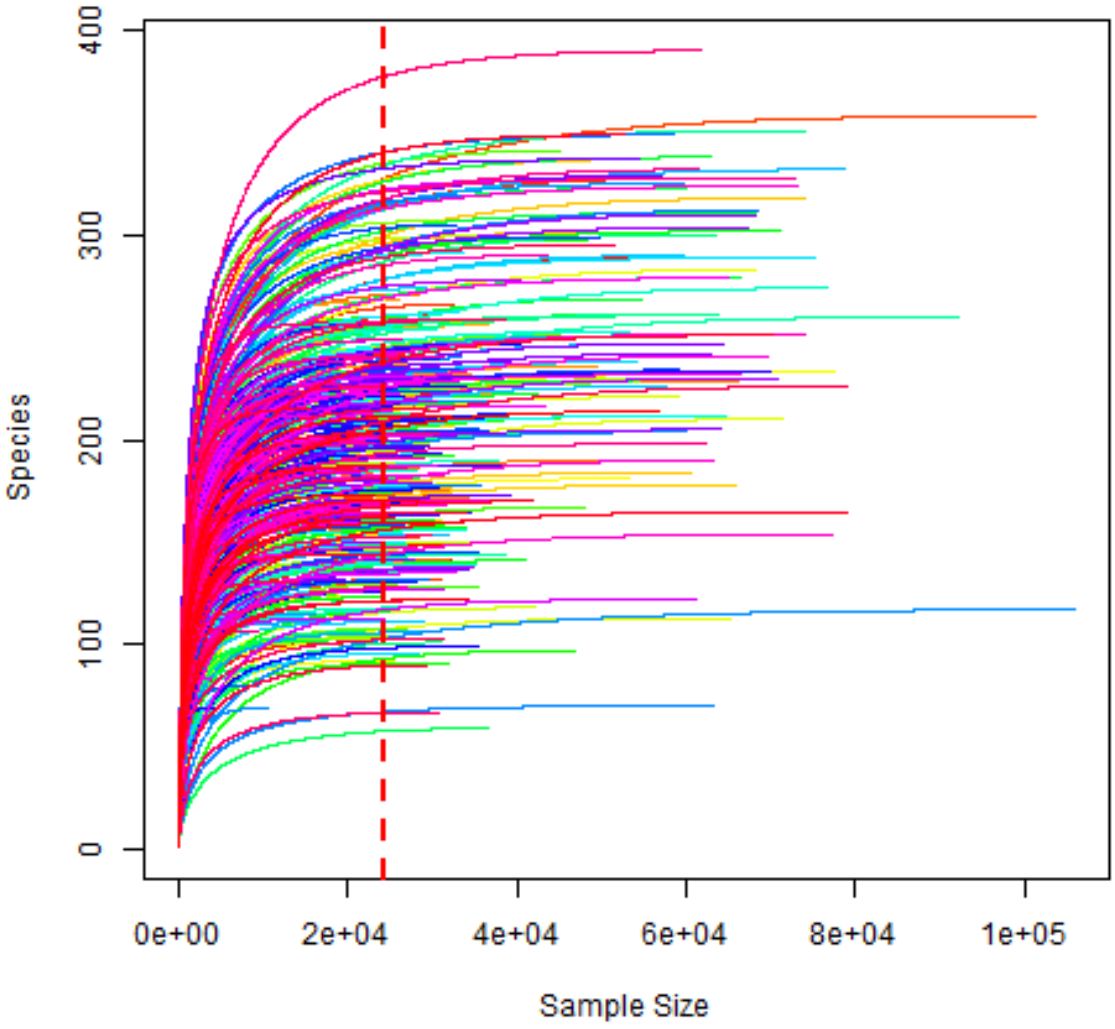

Figure 1: Alpha rarefaction plot where each line represents a sample. The dotted red line represents a sequencing depth of 24,406 reads.

2.2 Exploratory Data Analyses and Preliminary Analyses

Prior to rarefying, using all 441 samples, a redundancy analysis (RDA) was performed. The OTU table was aggregated to the phylum level and a Hellinger transformation was applied. The transformed and aggregated OTU table was used as the response variable and the entire metadata was used as the predictors. An ANOVA Permutation test was run on the model to determine significance of the covariates. Results can be found in Table **??**. We can see that fibre appears to be significantly associated to the OTU table.

Table 2: Redundancy Analysis results with hellinger-transformed phylum counts as a response matrix. Contains for each variable its degrees of freedom, variance explained by the variable, its F-statistic, and the p-value from ANOVA-like permutation tests.

variable	Df	Variance	F	Pr(>F)
adiponectin	1	0.0019950	5.0756737	0.001
age_years	1	0.0004970	1.2643629	0.265
age_range	1	0.0004762	1.2114433	0.302
BMI	1	0.0014115	3.5910102	0.007
BMI_class	2	0.0015760	2.0048381	0.035
Body_Fat_Percentage	1	0.0022959	5.8410633	0.001
Calorie_intake	1	0.0008806	2.2403008	0.061
Cardiometabolic_status	1	0.0004781	1.2162441	0.273
city	4	0.0168991	10.7484504	0.001
diastolic_bp	1	0.0010521	2.6767860	0.026
fiber	1	0.0014159	3.6022940	0.005
glucose	1	0.0003692	0.9394260	0.442
Hemoglobin_a1c	1	0.0002473	0.6290934	0.656
CRP	1	0.0000917	0.2332970	0.959
insulin	1	0.0002153	0.5478133	0.727
Total_Cholesterol	1	0.0003405	0.8663439	0.465
HDL	1	0.0001065	0.2709362	0.940
LDL	1	0.0003832	0.9748675	0.404
VLDL	1	0.0006079	1.5465834	0.183
Triglycerides	1	0.0001867	0.4751187	0.763
medication	1	0.0005710	1.4525957	0.185
per_carbohydrates	1	0.0007468	1.8999085	0.097
per_total_protein	1	0.0002285	0.5813990	0.702
per_total_fat	1	0.0001462	0.3720172	0.885
per_animal_protein	1	0.0007834	1.9930137	0.087
per_monoinsaturated_fat	1	0.0001688	0.4293863	0.829
per_polyunsaturated_fat	1	0.0002719	0.6918226	0.607
per_saturated_fat	1	0.0004138	1.0526674	0.360

Table 2: Redundancy Analysis results with hellinger-transformed phylum counts as a response matrix. Contains for each variable its degrees of freedom, variance explained by the variable, its F-statistic, and the p-value from ANOVA-like permutation tests.

variable	Df	Variance	F	Pr(>F)
sex	1	0.0005928	1.5082614	0.164
smoker	1	0.0005700	1.4501757	0.227
stool_consistency	3	0.0035411	3.0030428	0.001
systolic_bp	1	0.0006574	1.6726463	0.133
MET_mins_per_week	1	0.0001677	0.4266723	0.807
waist_circumference	1	0.0007175	1.8254798	0.096
Residual	399	0.1568304	NA	NA

Furthermore, a logistic regression model was fit onto the metadata, where cardiovascular status was the response variable. Model results are in Table **??**. We can see that MET units are significantly associated with cardiovascular status.

Table 3: Logistic regression results with cardiometabolic status as the response variable. Contains for each variable the log-odds coefficient, the standard error of the coefficient, its z-statistic, and its p-value.

Estimate	Std. Error	z value	Pr(>|z|)	Variable
-9.9122263	77.4539894	-0.1279757	0.8981682	(Intercept)
-0.0602070	0.0648303	-0.9286861	0.3530518	adiponectin
0.0656814	0.0441824	1.4865949	0.1371218	age_years
-0.3639803	0.8645281	-0.4210161	0.6737433	age_range41_62
0.0936441	0.1754121	0.5338521	0.5934439	BMI
-0.4970514	1.4419124	-0.3447168	0.7303073	BMI_classObese
-0.4844568	0.8015177	-0.6044244	0.5455615	BMI_classOverweight
-0.0052189	0.0811479	-0.0643130	0.9487210	Body_Fat_Percentage
-0.0003404	0.0007326	-0.4646448	0.6421858	Calorie_intake
-1.9390412	1.3234127	-1.4651825	0.1428711	cityBogota
-0.3654200	1.2887662	-0.2835425	0.7767610	cityBucaramanga
0.2606411	1.0634218	0.2450966	0.8063816	cityCali
-0.1763447	1.0270975	-0.1716922	0.8636795	cityMedellin
0.0339491	0.0339727	0.9993058	0.3176466	diastolic_bp
-0.0285924	0.0631854	-0.4525165	0.6508969	fiber
0.0955358	0.0324401	2.9449870	0.0032297	glucose
-0.2352485	0.5294736	-0.4443064	0.6568211	Hemoglobin_a1c
0.1768625	0.0453103	3.9033602	0.0000949	CRP
0.2162195	0.0508586	4.2513881	0.0000212	insulin

Table 3: Logistic regression results with cardiometabolic status as the response variable. Contains for each variable the log-odds coefficient, the standard error of the coefficient, its z-statistic, and its p-value.

Estimate	Std. Error	z value	Pr(>|z|)	Variable
-0.0525445	0.0539141	-0.9745964	0.3297605	Total_Cholesterol
-0.1135320	0.0632763	-1.7942248	0.0727773	HDL
0.0593667	0.0535188	1.1092685	0.2673144	LDL
-0.1594781	0.1546671	-1.0311058	0.3024912	VLDL
0.0587170	0.0305979	1.9189893	0.0549857	Triglycerides
0.8986201	0.4667142	1.9254185	0.0541770	medicationYes
-0.0338268	0.7495101	-0.0451318	0.9640022	per_carbohydrates
-0.1803468	0.7899163	-0.2283113	0.8194043	per_total_protein
0.1254492	0.8568957	0.1463997	0.8836059	per_total_fat
-0.0357073	0.0661648	-0.5396719	0.5894233	per_animal_protein
-0.4547436	0.3918152	-1.1606076	0.2458015	per_monoinsaturated_fat
-0.4089035	0.3731494	-1.0958171	0.2731588	per_polyunsaturated_fat
0.2903317	0.2742740	1.0585463	0.2898065	per_saturated_fat
-1.3558918	0.8241822	-1.6451361	0.0999418	sexmale
0.4601585	0.6921807	0.6647953	0.5061814	smokerYes
-1.4746828	1.2264204	-1.2024285	0.2291975	stool_consistencyHard
0.3962243	1.0608879	0.3734837	0.7087885	stool_consistencyNormal
-0.8717309	1.5102800	-0.5771982	0.5638056	stool_consistencysoft
0.0479254	0.0220354	2.1749271	0.0296356	systolic_bp
-0.0000771	0.0000433	-1.7807330	0.0749561	MET_mins_per_week
-0.0153338	0.0485372	-0.3159191	0.7520639	waist_circumference

The distributions of the variables of interest did not change considerably after rarefying. The distribution metabolic equivalent of task (MET) was heavily right-skewed (Figure **??**). The distribution of daily fibre intake was bell-shaped (Figure **??**). The number of female and male participants, as well as participants with a healthy or abnormal cardiometabolic status, are balanced throughout (Figure **??**). The number of participants with healthy cardiometabolic status went from 269 to 204, while those with an abnormal status went from 172 to 129. The number of female participants went from 229 to 178, while male participants went from 212 to 155. The distribution of age in years was also explored and it was roughly bell-shaped, almost uniform-like (Figure **??**). The distributions of these variables prior to filtering (all 441 initial samples portrayed) can be found in Figure **??**.

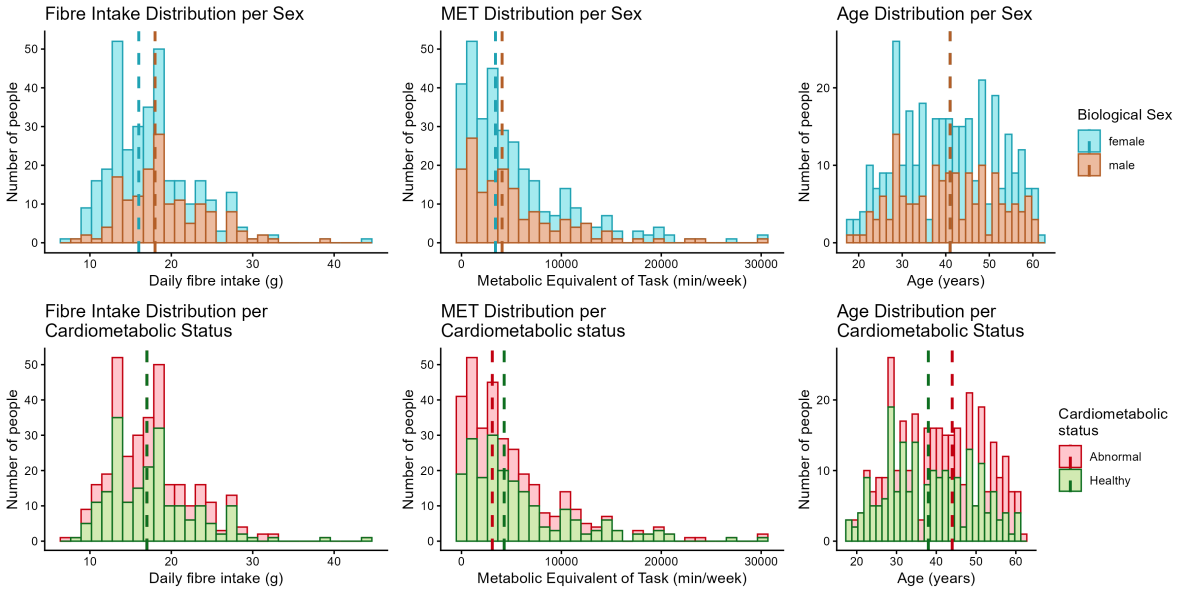

Figure 2: Histogram distributions of the 333 participants' fibre intake (left), metabolic equivalent of tasks (middle), and age (right). Distributions are colored by biological sex (top) and cardiometabolic status (bottom). Dotted vertical lines represent the median for each sub-group.

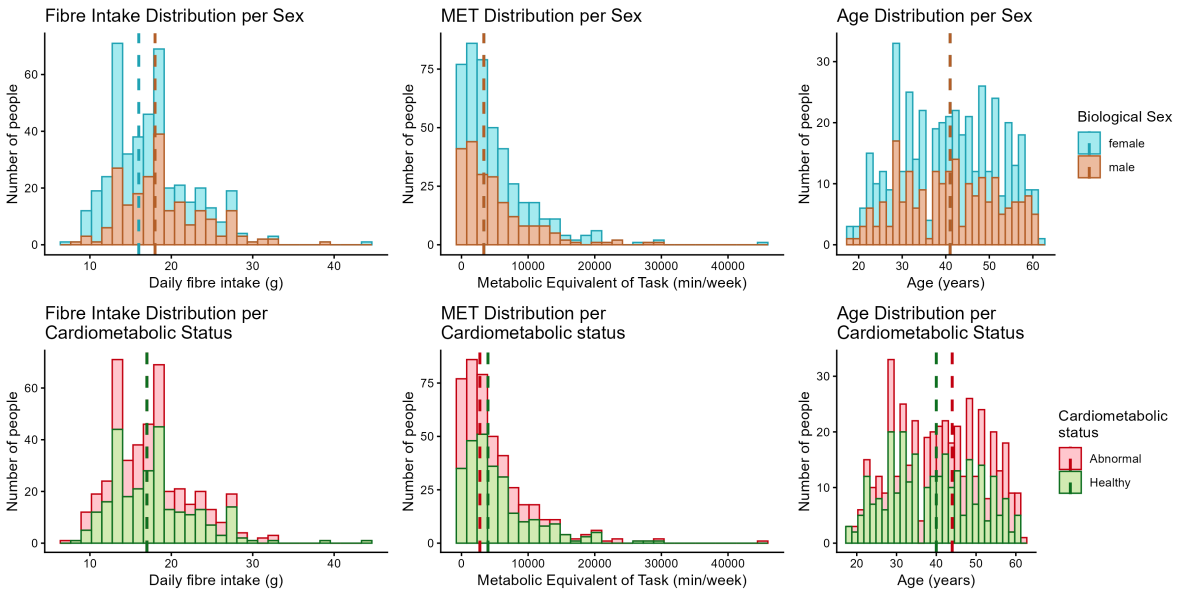

Figure 3: Histogram distributions of the 441 participants' fibre intake (left), metabolic equivalent of tasks (middle), and age (right). Distributions are colored by biological sex (top) and cardiometabolic status (bottom). Dotted vertical lines represent the median for each sub-group.

As mentioned previously, participants will be divided into two fibre intake and exercise level groups, while considering cardiovascular status as well. High fibre intake is considered as 20 grams of fibre or above. High exercise level is considered as more than 1000 metabolic equivalent of task minutes per week. These cut-offs were considered based in the distributions observed as well as guidelines on the adequate number of MET units and fibre intake. The sample sizes of each sub-group are reported in Table **??**. The smallest groups, representing participants with high fibre intake and low exercise, included 8 with a healthy and 9 with an abnormal cardiovascular status. Given these small sample sizes, all 333 samples were retained for downstream analyses, with no further filtering performed. Based on the observed distributions and lack of clear outliers, this approach was considered valid.

Table 4: Distribution of Participants by Cardiometabolic Health Status, Fibre Intake, and Exercise Level. High fibre intake is 20 grams or more of daily fibre. High exercise level is more than 1000 MET minutes per week.

Cardiometabolic_status	fibre_group	exercise_group	n
Abnormal	high	high	28
Abnormal	high	low	9
Abnormal	low	high	66
Abnormal	low	low	26

Table 4: Distribution of Participants by Cardiometabolic Health Status, Fibre Intake, and Exercise Level. High fibre intake is 20 grams or more of daily fibre. High exercise level is more than 1000 MET minutes per week.

Cardiometabolic_status	fibre_group	exercise_group	n
Healthy	high	high	51
Healthy	high	low	8
Healthy	low	high	120
Healthy	low	low	25

2.3 Aim 1: Multiple Linear Regression

The rarefied data's counts were aggregated at the phylum level. Then, relative abundances were calculated with Total Sum Scaling (TSS). Then, the data was normalied with an arcsine transformation. To see which phylum's relative abundances are significantly associated with caridometabolic status, a multiple linear regression model was fit with cardiometabolic status as the response variable, and the phylum as the predictors. The coefficients for each phylum are in Table **??**. We can see that with an alpha significance level of 0.05, none of the phylum's relative abundances are significantly associated with cardiometabolic status.

Table 5: MLR results.

Estimate	Std. Error	t value	Pr(>|t|)	variable
-0.1030853	0.5436388	-0.1896209	0.8497299	(Intercept)
0.3226526	0.2989240	1.0793801	0.2812547	p___Bacteroidota
4.3189756	3.3982094	1.2709563	0.2046940	p___Elusimicrobiota
15.3572986	43.0460063	0.3567648	0.7215098	p___Acidobacteriota
-0.2847704	0.4407976	-0.6460344	0.5187335	p___Cyanobacteriota
-27.4913393	34.5745686	-0.7951318	0.4271432	p___Deinococcota
0.1479361	0.2968721	0.4983159	0.6186136	p___Actinomycetota
5.8501334	8.2441111	0.7096136	0.4784755	p___Patescibacteria
-1.0730460	0.8334397	-1.2874908	0.1988805	p___Spirochaetota
0.5085949	0.2831205	1.7963902	0.0734028	p___Pseudomonadota
84.4610640	57.1732573	1.4772827	0.1406118	p___Myxococcota
60.8082138	44.5702238	1.3643237	0.1734521	p___Bdellovibrionota
-5.3129137	3.4197532	-1.5535957	0.1212979	p___Campylobacterota
-0.4470110	0.6634521	-0.6737652	0.5009610	p___Thermodesulfobacteriota
-7.0689198	16.1496800	-0.4377127	0.6618984	p___Deferribacterota
-12.9090115	22.5568051	-0.5722890	0.5675396	p___Planctomycetota
-15.9792680	19.8722745	-0.8040986	0.4219542	p___Chloroflexota
0.2050304	0.2572284	0.7970755	0.4260153	p___Verrucomicrobiota

Table 5: MLR results.

Estimate	Std. Error	t value	Pr(>|t|)	variable
-59.9128970	34.7258146	-1.7253129	0.0854645	p___Halanaerobiaeota
1.2556620	1.1518442	1.0901318	0.2764990	p___Synergistota
0.2620647	0.3145957	0.8330206	0.4054724	p___Bacillota
-0.9165472	1.9677564	-0.4657828	0.6416970	p___Fusobacteriota

2.4 Aim 1: Redundancy Analysis

The rarefied data's counts were aggregated at the phylum level. Then, the data was transformed with Hellinger, which results in normalized relative abundances. A redundancy analysis (RDA) model was fit, where the response variables are the phylum relative abundances, and the predictors are the metadata. In the metadata, redundant variables were removed (country, latitude, age range). The results are in Table **??**. With an alpha signficance level of 0.05, we see that these variables explain the variance in ordination space in phylum relative abundances are: city, stool consistency, adiponectin, calorie intake, BMI, and fiber.

Table 6: RDA results.

variable	Df	Variance	F	Pr(>F)
adiponectin	1	0.0018172	3.6659357	0.007
age_years	1	0.0005841	1.1783261	0.299
BMI	1	0.0016551	3.3389867	0.013
BMI_class	2	0.0011432	1.1531723	0.286
Body_Fat_Percentage	1	0.0011079	2.2349983	0.061
Calorie_intake	1	0.0016099	3.2478488	0.011
Cardiometabolic_status	1	0.0004875	0.9834606	0.412
city	4	0.0190262	9.5957642	0.001
diastolic_bp	1	0.0006976	1.4074114	0.211
fiber	1	0.0015811	3.1895962	0.020
glucose	1	0.0004913	0.9912166	0.405
Hemoglobin_a1c	1	0.0005754	1.1608817	0.324
CRP	1	0.0000466	0.0940227	0.996
insulin	1	0.0002641	0.5327352	0.739
Total_Cholesterol	1	0.0003318	0.6692897	0.583
HDL	1	0.0001224	0.2470105	0.930
LDL	1	0.0008442	1.7029788	0.143
VLDL	1	0.0003859	0.7784929	0.524
Triglycerides	1	0.0000905	0.1826502	0.940
medication	1	0.0004315	0.8704178	0.465

Table 6: RDA results.

variable	Df	Variance	F	Pr(>F)	
per_carbohydrates	1	0.0004974	1.0034106	0.389	
per_total_protein	1	0.0004163	0.8398633	0.492	
per_total_fat	1	0.0001005	0.2027760	0.955	
per_animal_protein	1	0.0009943	2.0059340	0.088	
per_monoinsaturated_fat	1	0.0001180	0.2379887	0.926	
per_polyunsaturated_fat	1	0.0001490	0.3006193	0.894	
per_saturated_fat	1	0.0004862	0.9807815	0.358	
sex	1	0.0011361	2.2919732	0.065	
smoker	1	0.0005372	1.0837177	0.360	
stool_consistency	3	0.0057403	3.8601305	0.001	
systolic_bp	1	0.0008095	1.6330857	0.172	
MET_mins_per_week	1	0.0003885	0.7837052	0.534	
waist_circumference	1	0.0006363	1.2835684	0.239	
Residual	293	0.1452382	NA	NA	

However, when we look at the four subgroups with an ordination plot, we do not see any differences (Figure **??**).

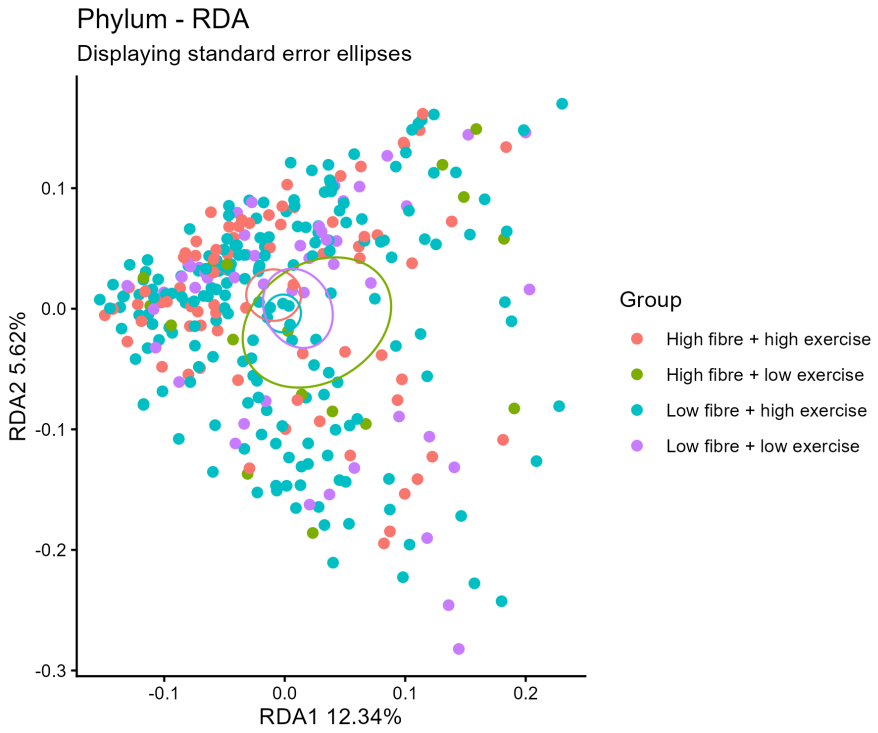

Figure 4: RDA ordination plot, of all four subgroups.

For the values that were selected as statistically significant, a new model was fit using them with PERMANOVA (adonis2 from the vegan package). The coefficients can be found in Table **??**. We see that BMI, calorie intake, and fiber do not significantly structure the multivariate distances. However, it is to be noted that the p-value for fiber was barely insignificant (p = 0.058).

Table 7: adonis2 results with the statistically significant variables from the RDA.

variable	Df	SumOfSqs	R2	F	Pr(>F)
city	4	2.3173325	0.1264281	12.390983	0.001
stool_consistency	3	0.6072894	0.0331323	4.329641	0.001
adiponectin	1	0.1322041	0.0072127	2.827629	0.022
Calorie_intake	1	0.0573931	0.0031312	1.227544	0.290
BMI	1	0.0994836	0.0054276	2.127791	0.093
fiber	1	0.1073785	0.0058583	2.296650	0.058
Residual	321	15.0081656	0.8188097	NA	NA
Total	332	18.3292469	1.0000000	NA	NA

2.5 Aim 1: Least Absolute Shrinkage and Selection Operator (metadata only)

To first explore the data, numeric variables' correlation (Spearman) were visualized in a heatmap. Only numeric variables were included as LASSO does not support categorical variables. We see that some variables are highly correlated with each other in Figure **??** (BMI with waist circumference, diastolic bp with systolic bp, VLDL with triglycerides, and cholesterol with LDL). Due to multicollinearity, we do expect that during feature selection, one of the highly correlated variables will be dropped.

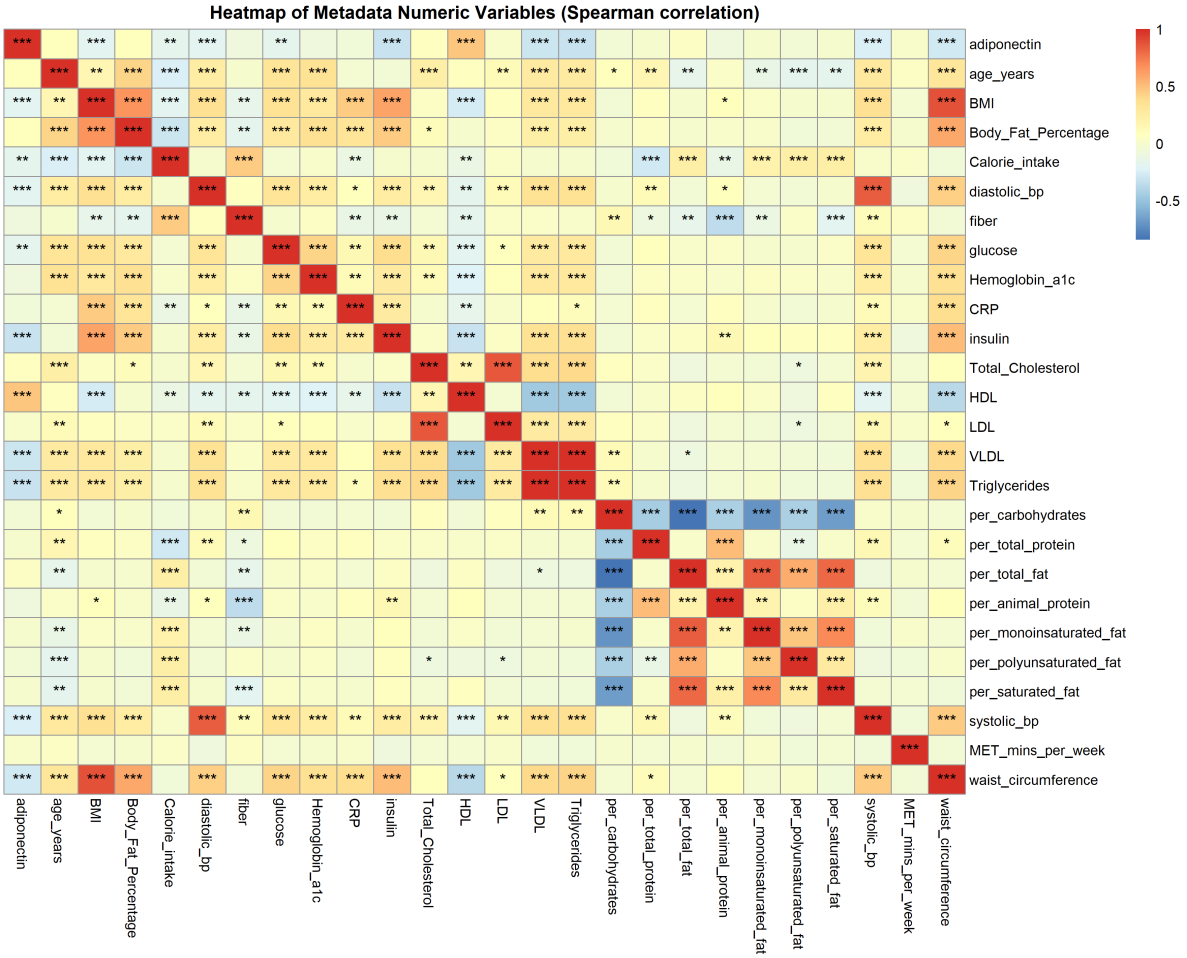

Figure 5: Heatmap of numeric variables' correlation. Presence of asterisks indicates the significance of the Spearman correlation value.

The data was divided into a training and testing set, with a 50/50 split. As mentioned above, only numeric variables were included. Cardiometabolic status, the response variable, was converted to a binary value (0 and 1). On the training set, LASSO selection is performed with

10-fold cross-validation to find the minimum lambda. We used the minimum lambda, instead of the 1-se lambda, as we wished to create a model with the highest AUC. We can see the crossvalidation in Figure **??** that the lambda that maximizes the AUC value is associated with between 16 and 20 variables.

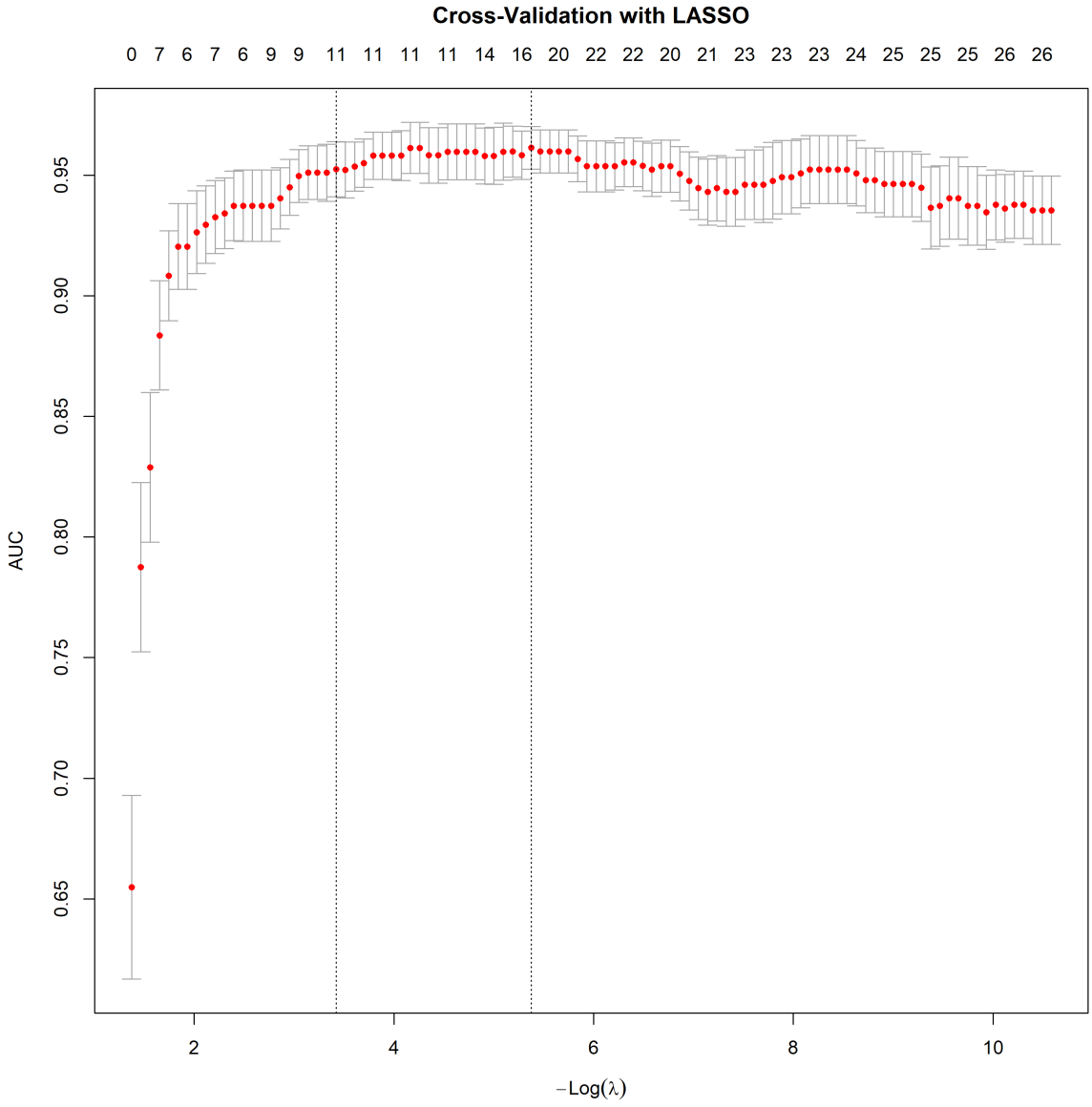

Figure 6: LASSO crossvalidation.

We find the coefficients from a logistic regression using the LASSO selected model, as these will be needed for prediction. These are: adiponectin, BMI, calorie intake, diastolic bp, fiber,

glucose, CRP, insulin, total cholesterol, HDL, triglycerides, percentage total fat, percentage animal protein, percentage monounsaturated fat, percentage polyunsaturated fat, percentage saturated fat, systolic bp, and waist circumference.

The selected model was fit onto the testing set for inference purposes. The results are in Table **??**. We see that from the selected variables, only these ones were significantly associated with cardiometabolic status outcome: glucose, HDL, insulin, CRP, and systolic bp.

Table 8: Exponentiated estimates from the Logistic Regression, on the testing set.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.000	6.544	-1.890	0.059	0.000	1.090
adiponectin	0.993	0.083	-0.089	0.929	0.829	1.158
BMI	0.975	0.151	-0.168	0.867	0.719	1.310
Calorie_intake	0.999	0.001	-0.623	0.533	0.998	1.001
diastolic_bp	1.009	0.045	0.191	0.849	0.923	1.105
fiber	0.929	0.076	-0.965	0.335	0.788	1.070
glucose	1.139	0.041	3.168	0.002	1.059	1.246
CRP	1.230	0.099	2.085	0.037	1.069	1.535
insulin	1.182	0.056	2.973	0.003	1.070	1.336
Total_Cholesterol	1.001	0.009	0.137	0.891	0.983	1.020
HDL	0.884	0.040	-3.084	0.002	0.812	0.951
Triglycerides	1.010	0.005	1.912	0.056	1.000	1.022
per_total_fat	0.649	0.408	-1.061	0.289	0.271	1.361
per_animal_protein	0.956	0.075	-0.601	0.548	0.821	1.109
per_monoinsaturated_fat	2.671	0.706	1.392	0.164	0.698	11.480
per_polyunsaturated_fat	0.684	0.571	-0.665	0.506	0.221	2.136
per_saturated_fat	1.614	0.420	1.141	0.254	0.729	3.861
systolic_bp	1.066	0.032	2.009	0.045	1.003	1.138
waist_circumference	0.977	0.046	-0.506	0.613	0.889	1.069

For prediction, we first perform the coefficients selected by LASSO on the testing set. Then, we used a confusion matrix for evaluation, with a threshold of 0.5.

The number of true positives is 91, for true negatives is 50, for false positives is 15, and for false negatives is 11. The sensitivity of the model is 0.77 whereas the specificity is 0.89. Thus, the model is better at predicting when the cardiometabolic status is healthy, but not at predicting when the status is abnormal. The accuracy is 0.84, while Cohen's Kappa value is 0.67.

We then looked at the ROC of this model on the test set in order to see prediction performance under all threshold levels. From Figure **??**, we see that the AUC is quite high (0.93).

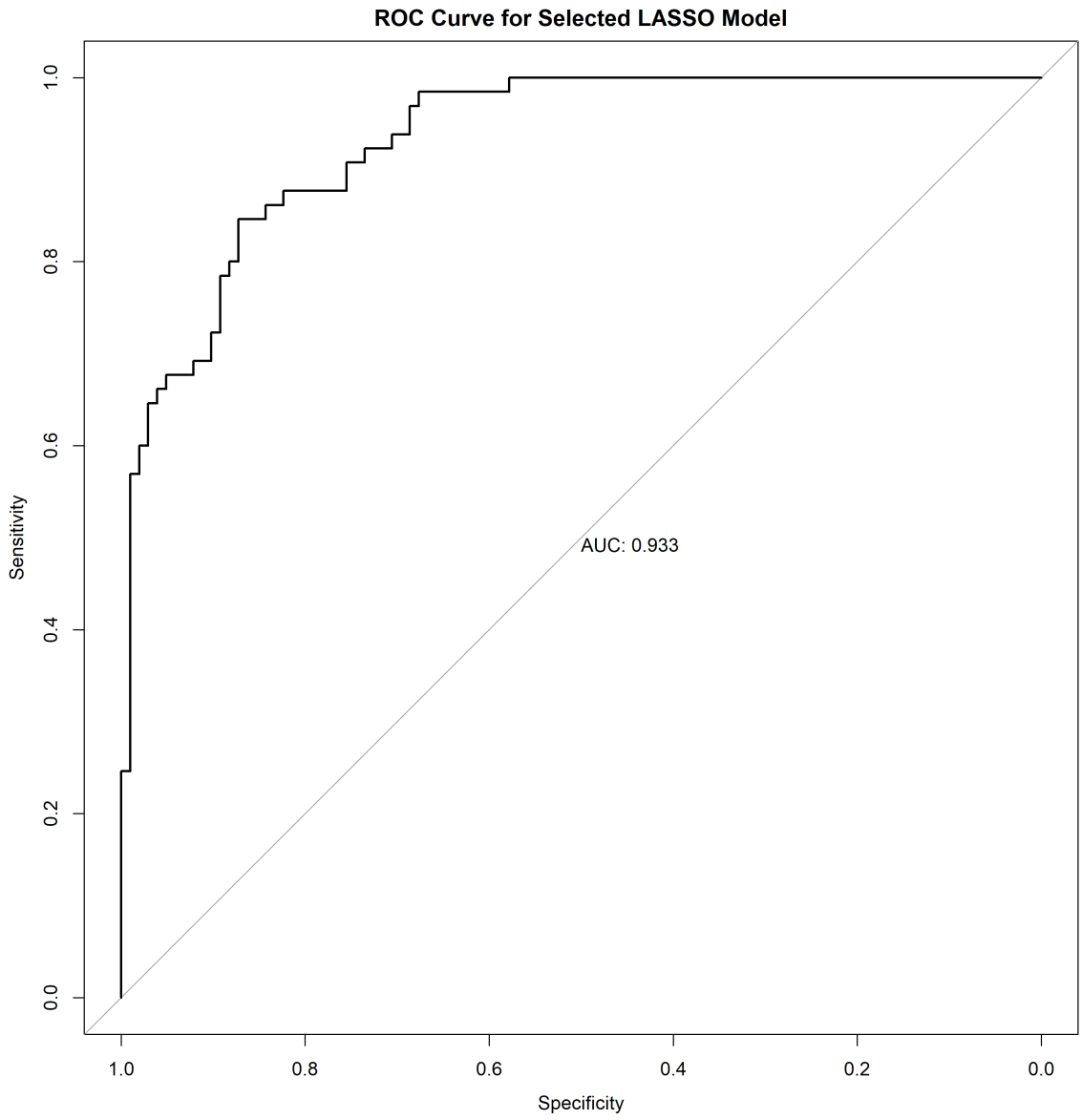

Figure 7: ROC plot of the LASSO logistic model on the testing set.

Overall, we can conclude that although fibre was selected as a predictor by LASSO, fibre on its own is not significantly associated with the outcomes of cardiometabolic status.

2.6 Aim 2: Alpha diversity

For Faith's phylogenetic diversity, the phylogenetic tree had to be resolved. Polyomies were resolved and converted into a fully dichotomous tree, branch lengths were assigned with Grafen, the tree was made ultrameric, and any zero-length edges were replaced with 1e-6. Node ages were also calculated.

For the eight subgroups (factoring in fibre, exercise, and cardiometabolic status), Faith's phylogenetic diversity was factored in. For each cardiometabolic status, Kruskal-Wallis tests were computed to test differences among the 4 groups. From Figure **??**, we see that there is no difference across groups.

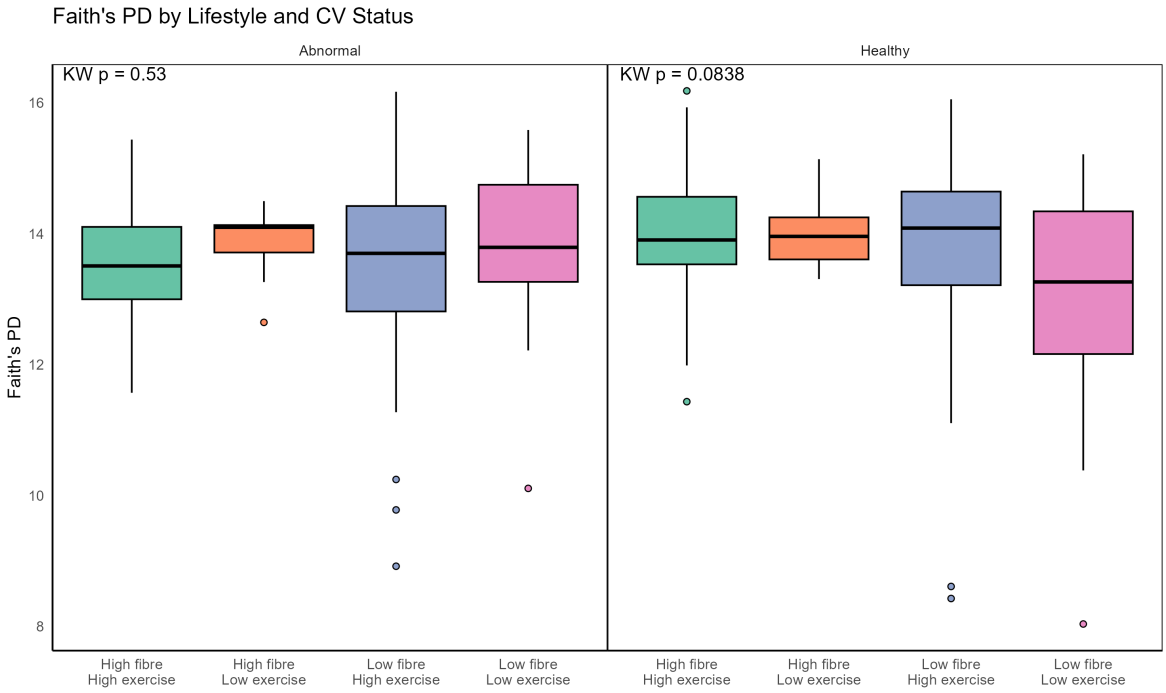

Figure 8: Faith's PD boxplots, by lifestyle and cardiovascular status.

Pairwise comparisons using Wilcoxon tests with Bonferroni-Holm correction were also computed, which are in Table **??**. We see that there are no significant differences in any pair either.

Table 9: Pairwise Wilcoxon test results with Bonferroni-Holm correction on Faith's PD metric.

Group1	adequate fibre;high exercise_Abnormal	adequate fibre;high exercise_Healthy	adequate fibre;low exercise_Abnormal	adequate fibre;low exercise_Healthy	inadequate fibre;high exercise_Abnormal	inadequate fibre;high exercise_Healthy	inadequate fibre;low exercise_Abnormal
adequate fibre;high exercise_Healthy	0.2957946	NA	NA	NA	NA	NA	NA
adequate fibre;low exercise_Abnormal	0.6507254	0.9028916	NA	NA	NA	NA	NA
adequate fibre;low exercise_Healthy	0.6507254	0.9122966	0.8373655	NA	NA	NA	NA
inadequate fibre;high exercise_Abnormal	0.9122966	0.2957946	0.7710407	0.6564888	NA	NA	NA
inadequate fibre;high exercise_Healthy	0.3938542	0.8373655	0.8211022	0.9516659	0.3072419	NA	NA
inadequate fibre;low exercise_Abnormal	0.6507254	0.8211022	0.9556802	0.9304010	0.6507254	0.9304010	NA
inadequate fibre;low exercise_Healthy	0.6665215	0.2957946	0.6507254	0.5246644	0.6564888	0.2957946	0.3794662

Shannon's diversity was also computed, which is seen in Figure **??**. Kruskwal-Wallis tests were also computed, and we also see no difference amongst grops.

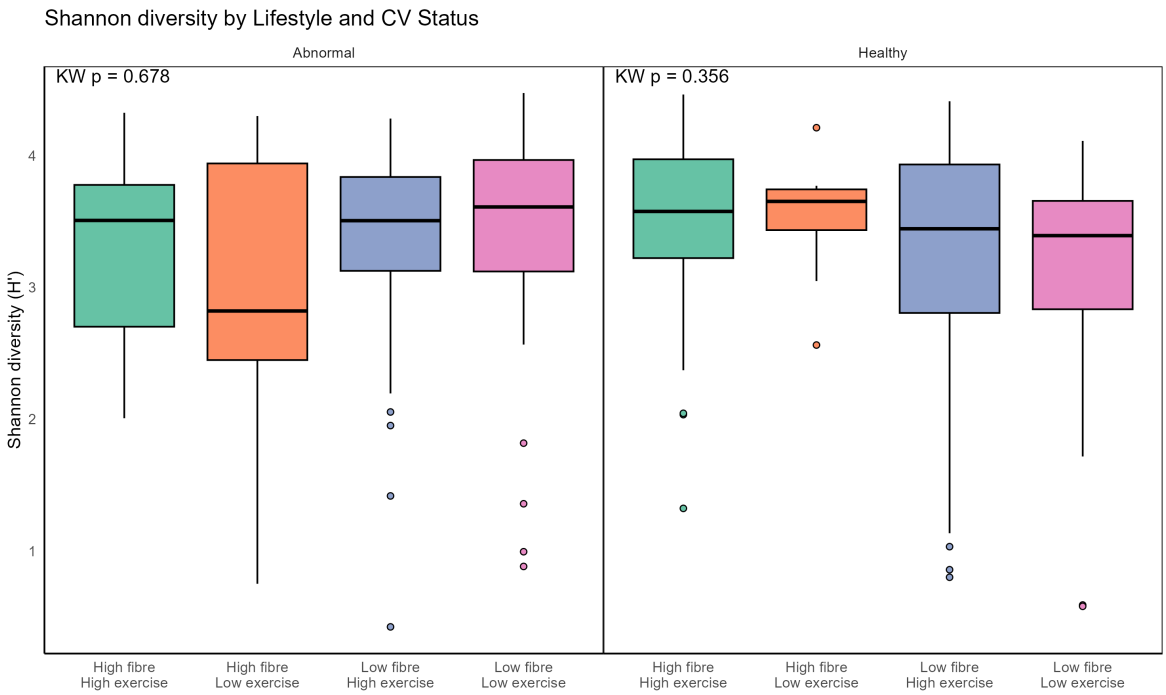

Figure 9: Shannon's diversity boxplots, by lifestyle and cardiovascular status.

Similarly, pairwise comparisons using Wilcoxon tests with Bonferroni-Holm correction were also computed, which are in Table **??**. We see that there are no significant differences in any pair either.

Table 10: Pairwise Wilcoxon test results with Bonferroni-Holm correction on Shannon's Diversity metric.

Group1	adequate fibre;high exercise_Abnormal	adequate fibre;high exercise_Healthy	adequate fibre;low exercise_Abnormal	adequate fibre;low exercise_Healthy	inadequate fibre;high exercise_Abnormal	inadequate fibre;high exercise_Healthy	inadequate fibre;low exercise_Abnormal
adequate fibre;high exercise_Healthy	0.7919832	NA	NA	NA	NA	NA	NA
adequate fibre;low exercise_Abnormal	0.7919832	0.7919832	NA	NA	NA	NA	NA

Table 10: Pairwise Wilcoxon test results with Bonferroni-Holm correction on Shannon's Diversity metric.

Group1	adequate fibre;high exercise_Abnormal	adequate fibre;high exercise_Healthy	adequate fibre;low exercise_Abnormal	adequate fibre;low exercise_Healthy	inadequate fibre;high exercise_Abnormal	inadequate fibre;high exercise_Healthy	inadequate fibre;low exercise_Abnormal
adequate fibre;low exercise_Healthy	0.7919832	0.9877015	0.7919832	NA	NA	NA	NA
inadequate fibre;high exercise_Abnormal	0.9877015	0.7919832	0.7919832	0.9345250	NA	NA	NA
inadequate fibre;high exercise_Healthy	0.9902344	0.7919832	0.7919832	0.7919832	0.9877015	NA	NA
inadequate fibre;low exercise_Abnormal	0.7919832	0.9877015	0.7919832	0.9877015	0.7919832	0.7919832	NA
inadequate fibre;low exercise_Healthy	0.8086218	0.7919832	0.8238520	0.7919832	0.7919832	0.7919832	0.7919832

2.7 Aim 2: Beta diversity

Similarly to alpha diversity, beta diversity for the subgroups was calculated with Bray-curtis. The ordination was calculated, and PCoA plots were produced for each cardiometabolic status. Within each cardiometabolic status, the PERMANOVA of the groups was calculated. From Figure **??**, we see that there are no differences.

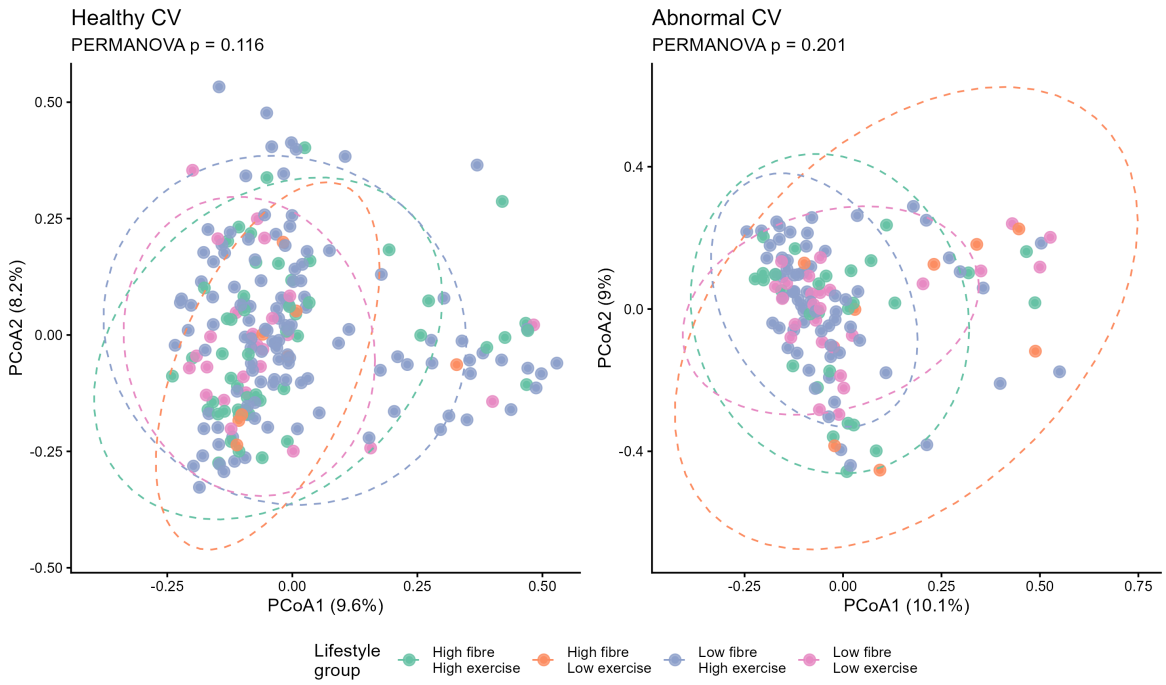

Figure 10: Bray-Curtis PCoA plot, with 95% confidence interval ellipses drawn.

2.8 Aim 3: DESeq2

On the raw, non-rarefied data, counts had 1 added to their counts to avoid non-zero data. The data was then split by cardiometabolic status. On each subset, the fibre and exercise group were added as contrasts, with the variables selected by RDA as covariates to control for them (fibre_group + exercise_group + adiponectin + BMI + Calorie_intake + city + stool_consistency).

The volcano plots are in Figure **??**. The absolute logFC cutoff was 1.5, and the FDR cutoff was 0.05. We see that there are many taxa over- or under-represented in each group.

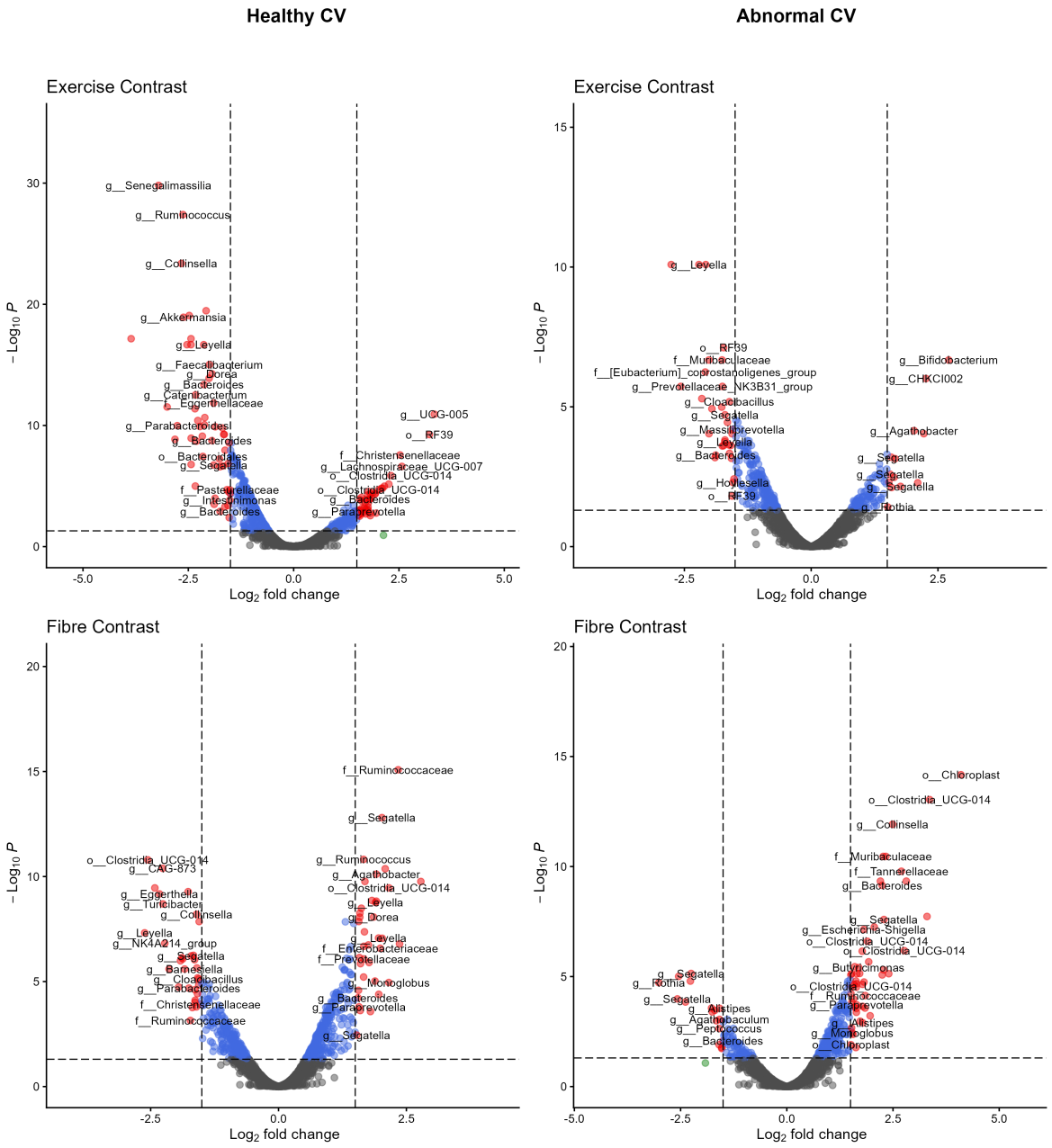

Figure 11: Volcano plots of fibre and exercise contrasts, separated by cardiometabolic status.

To investigate the taxa further, logFC plots were made, aggregated by family. For these plots, only taxa with an absolute logFC bigger than 2 and a FDR under 0.05 were considered. The plots are in Figure **??**. Within the healthy cardiometabolic groups, adequate exercise showed an increase in three taxa including Lachnospiraceae, Christensenellaceae, and Oscillospiraceae.

Several taxa were also decreased in this group, with the most abundant being Oscillospiraceae and Eggerthellaceae. Adequate fibre resulted in an increase in different taxa, with the greatest increase in Ruminococcaceae.

Among the abnormal cardiometabolic groups, adequate exercise resulted in an increase in taxa including Bifidobacteriaceae and Eggerthellaceae. A large decrease in the abundance of Lachnospiraceae was also observed, along with other taxa to a lesser extent. Adequate fibre within the abnormal cardiometabolic group resulted in an increase in a large number of taxa, with the greatest difference in Prevotellaceae, which was simultaneously the most increased and decreased.

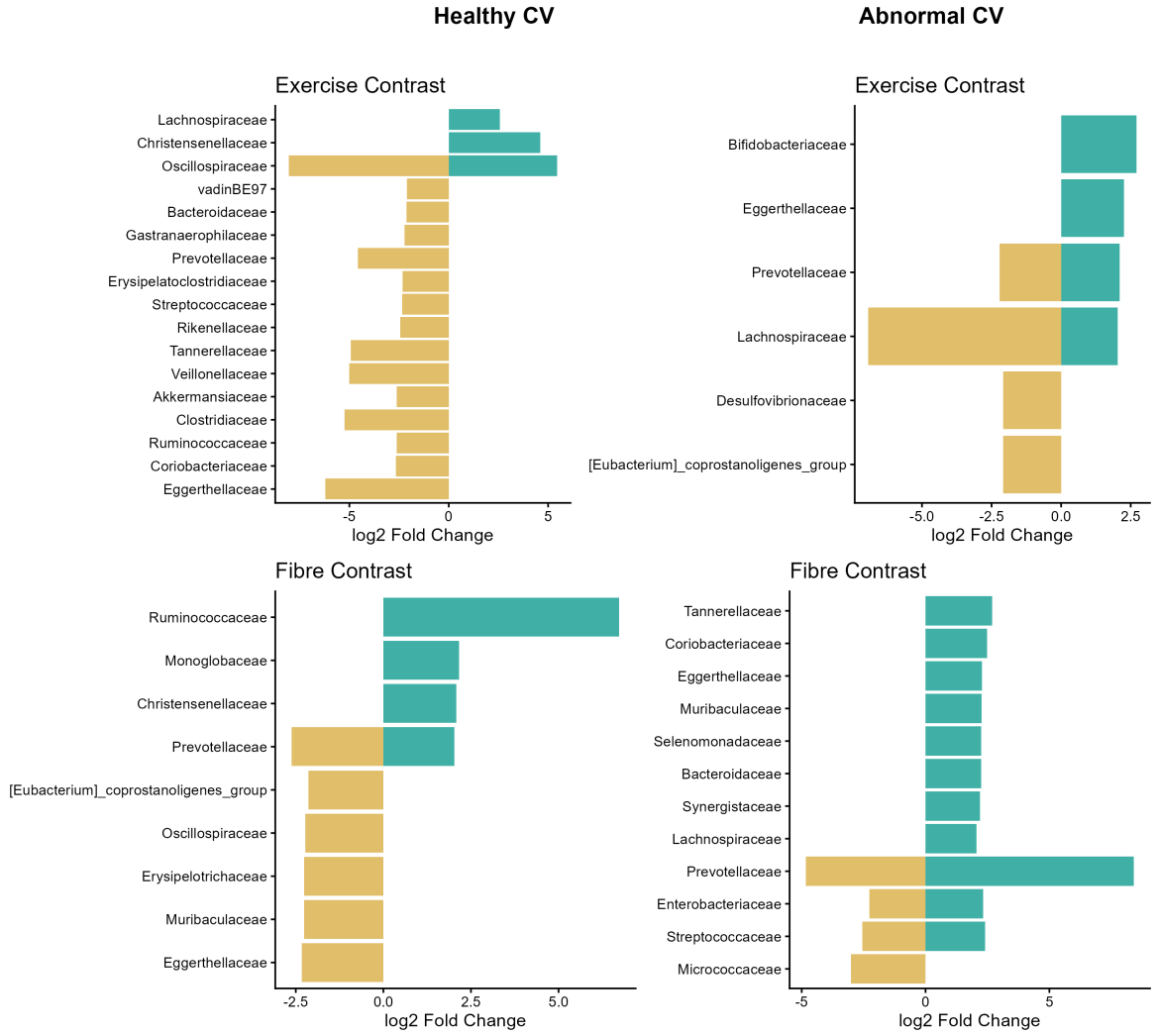

Figure 12: LogFC plots of fibre and exercise contrasts aggregated by family, separated by cardiometabolic status.

The same plot was done, but aggregated at the phylum level, which is in Figure **??**

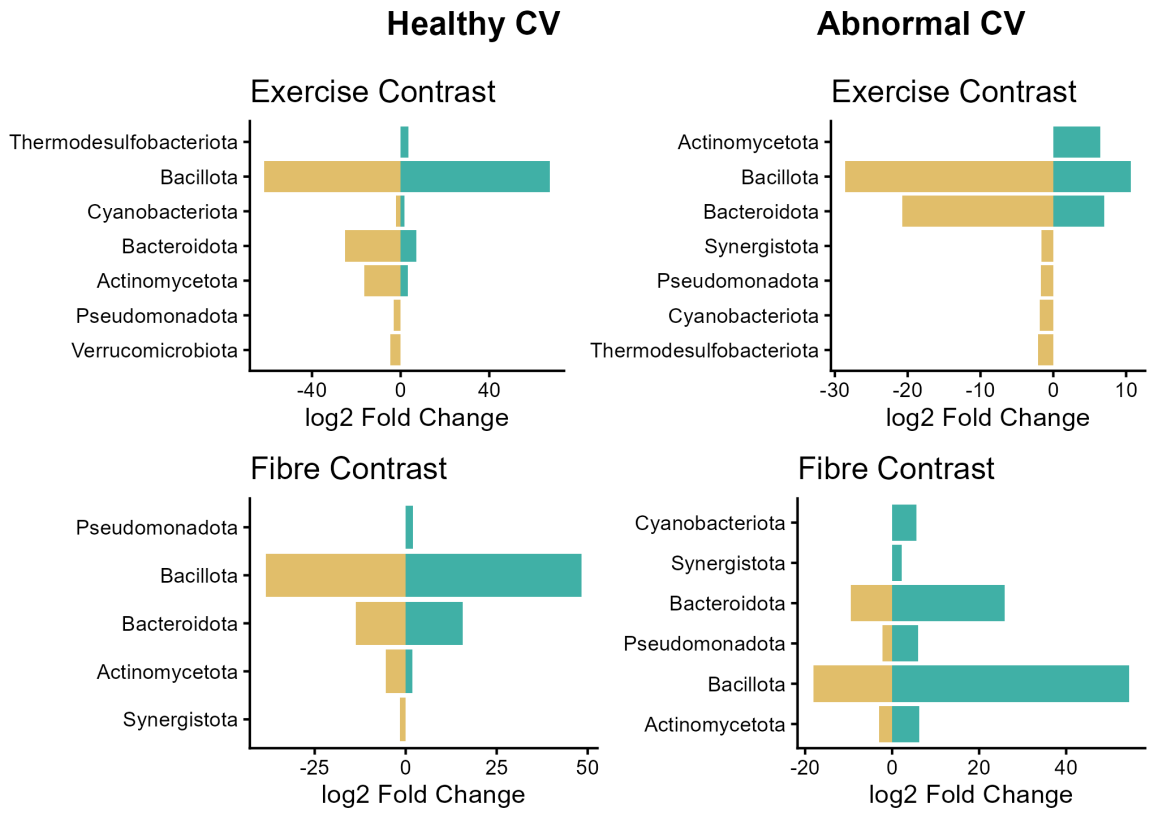

Figure 13: LogFC plots of fibre and exercise contrasts aggregated by phylum, separated by cardiometabolic status.

2.9 Aim 3: Least Absolute Shrinkage and Selection Operator (counts only)

To prepare the counts for LASSO, counts were aggregated at the phylum level, relative abundances calculated with TSS, and transformed with Center Log Ratio.

The data was divided into a training and testing set, with a 50/50 split. Cardiometabolic status, the response variable, was converted to a binary value (0 and 1). On the training set, LASSO selection was performed with 10-fold cross-validation to find the minimum lambda. We can see the crossvalidation in Figure **??** that the lambda that maximizes the AUC value is associated with under 4 variables. However, we also see that no matter the selected variables, the AUC values are quite low.

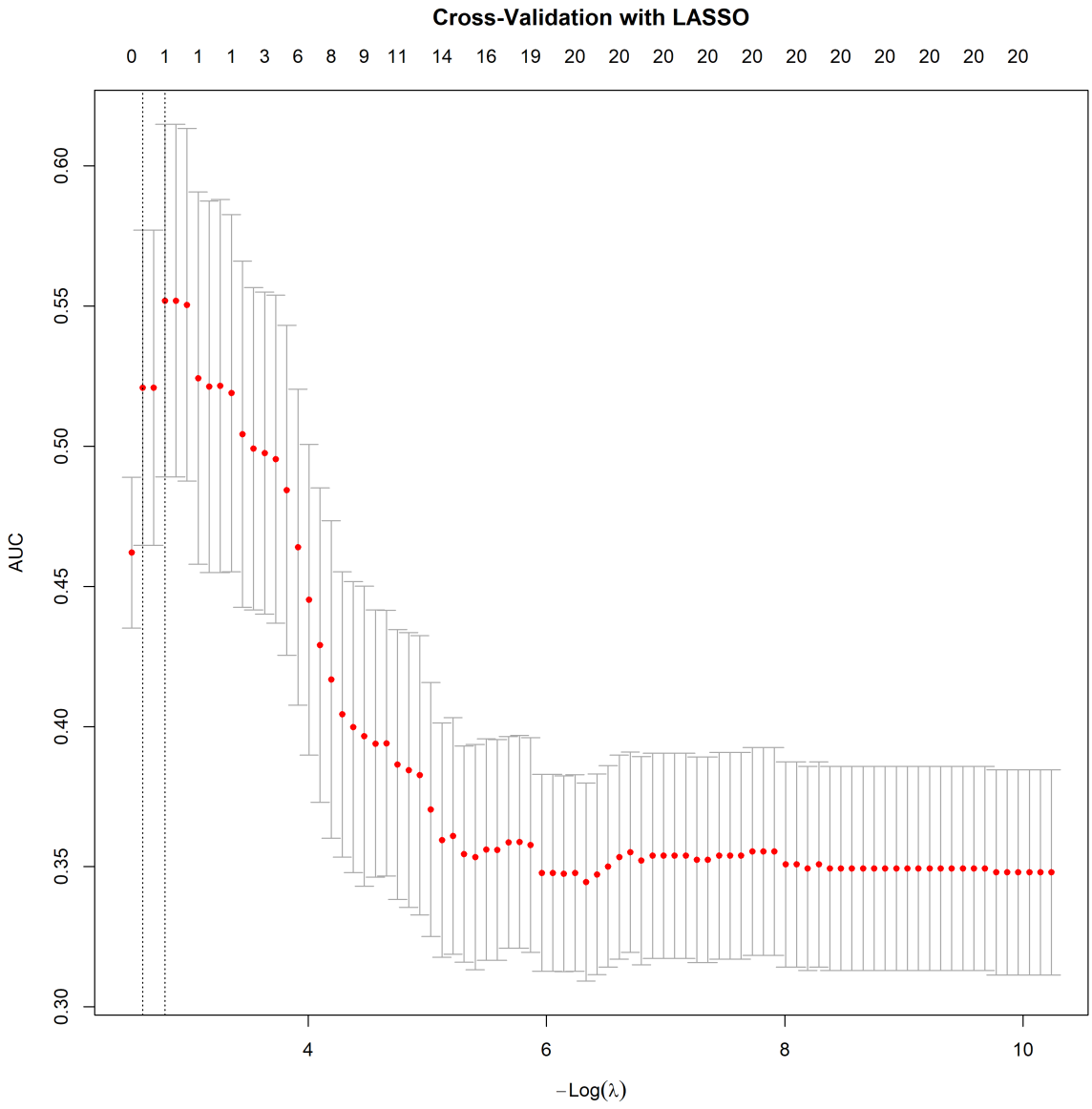

Figure 14: LASSO crossvalidation.

We find the coefficients from a logistic regression using the LASSO selected model, as these will be needed for prediction. It was only Bacteroidota. The selected model was fit onto the testing set for inference purposes. The results are in Table **??**. We see that Bacteroidota is not significantly associated with the outcomes of cardiometabolic status.

Table 11: Exponentiated estimates from the Logistic Regression, on the testing set.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.395	0.782	0.426	0.670	0.301	6.539
p__Bacteroidota	0.877	0.129	-1.020	0.308	0.678	1.128

For prediction, we first perform the coefficients selected by LASSO on the testing set. Then, we used a confusion matrix for evaluation, with a threshold of 0.5.

The number of true positives is 102, for true negatives is 0, for false positives is 0, and for false negatives is 65. The sensitivity of the model is 0 whereas the specificity is 100%. Thus, the model seems to always predict the status as healthy. The accuracy is 0.61, while Cohen's Kappa value is 0.

We then looked at the ROC of this model on the test set in order to see prediction performance under all threshold levels. From Figure **??**, we see that the AUC is 0.537, meaning that the model does not perform better than chance.

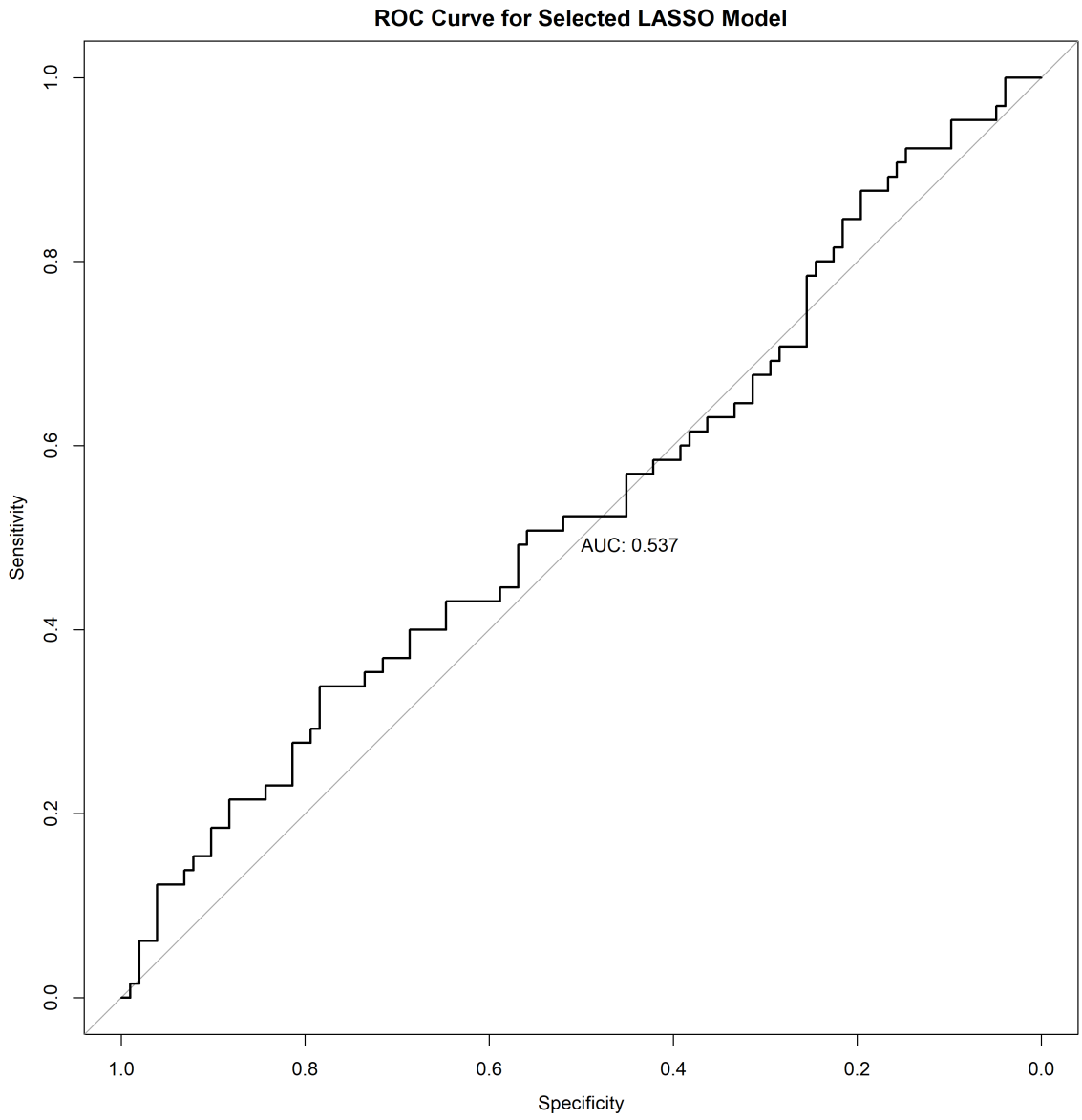

Figure 15: ROC plot of the LASSO logistic model on the testing set.

2.10 Aim 3: Networks

Microbial co-occurrence networks were constructed separately for healthy and abnormal samples using the trans_network class in the microeco R package. Genus-level abundance tables were used as input. Taxa with mean relative abundances below 0.0001 were filtered out. Pairwise associations between taxa were estimated using Spearman's rank correlation coefficient.

For each group, correlation matrices were used to infer co-occurrence networks by retaining statistically significant associations (permutation-based $p < 0.05$). The resulting adjacency matrices were then used to construct the final networks, after which network modules were identified.

10 modules were identified in the healthy cohort (Figure **??**) whereas 11 were identified in the abnormal cohort (Figure **??**).

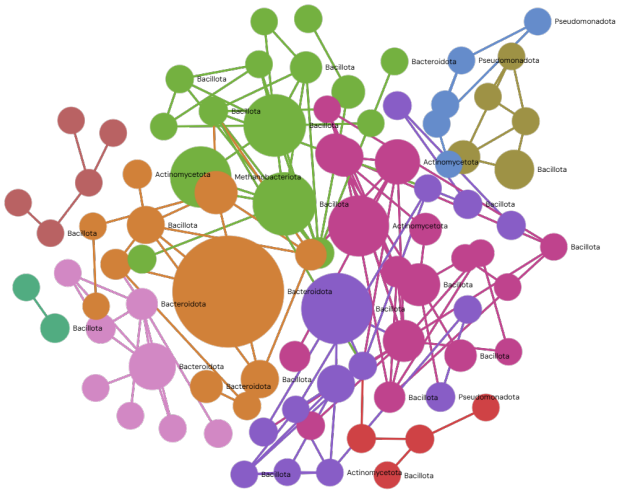

Figure 16: Networks colored by module in the healthy cohort.

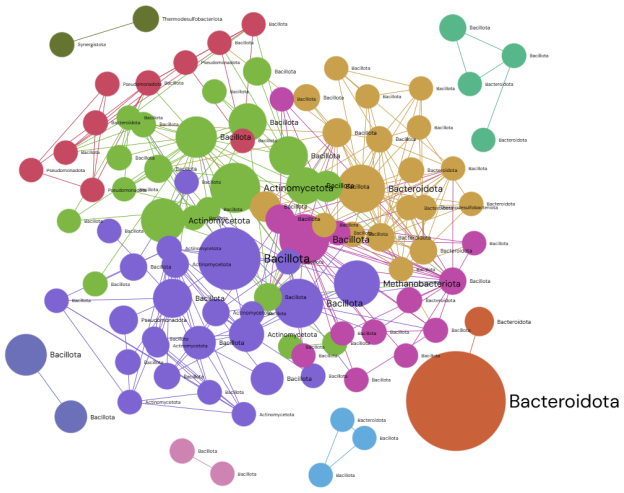

Figure 17: Networks colored by module in the abnormal cohort.

In Figure **??** and Figure **??**, we can see the same networks but colored by phylum.

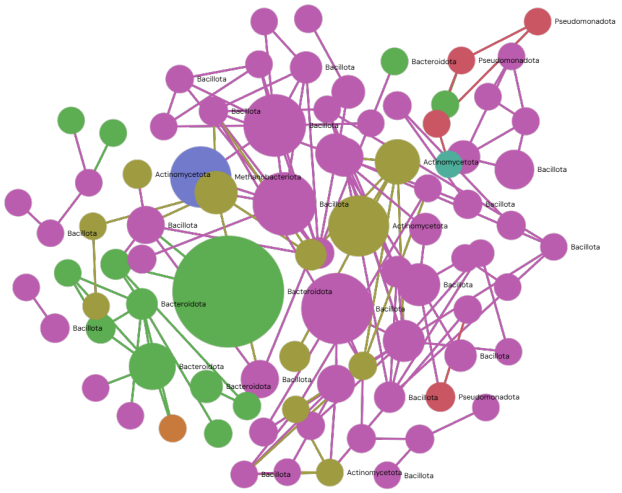

Figure 18: Networks colored by phylum in the healthy cohort.

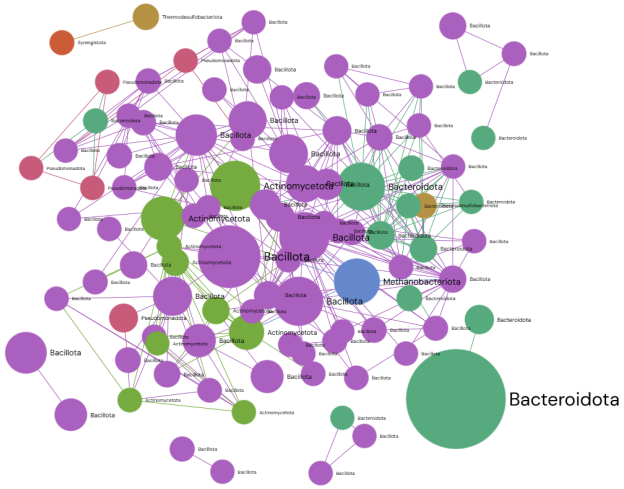

Figure 19: Networks colored by phylum in the abnormal cohort.

Next, global and node-level network attributes were computed. To characterize the ecological roles of individual taxa within each network, node-level metrics were extracted and used to classify taxa into topological roles based on within-module connectivity (Zi) and among-module connectivity (Pi). From Figure **??** and Figure **??**, we see that the abnormal cohort contains a greater number of connectors.

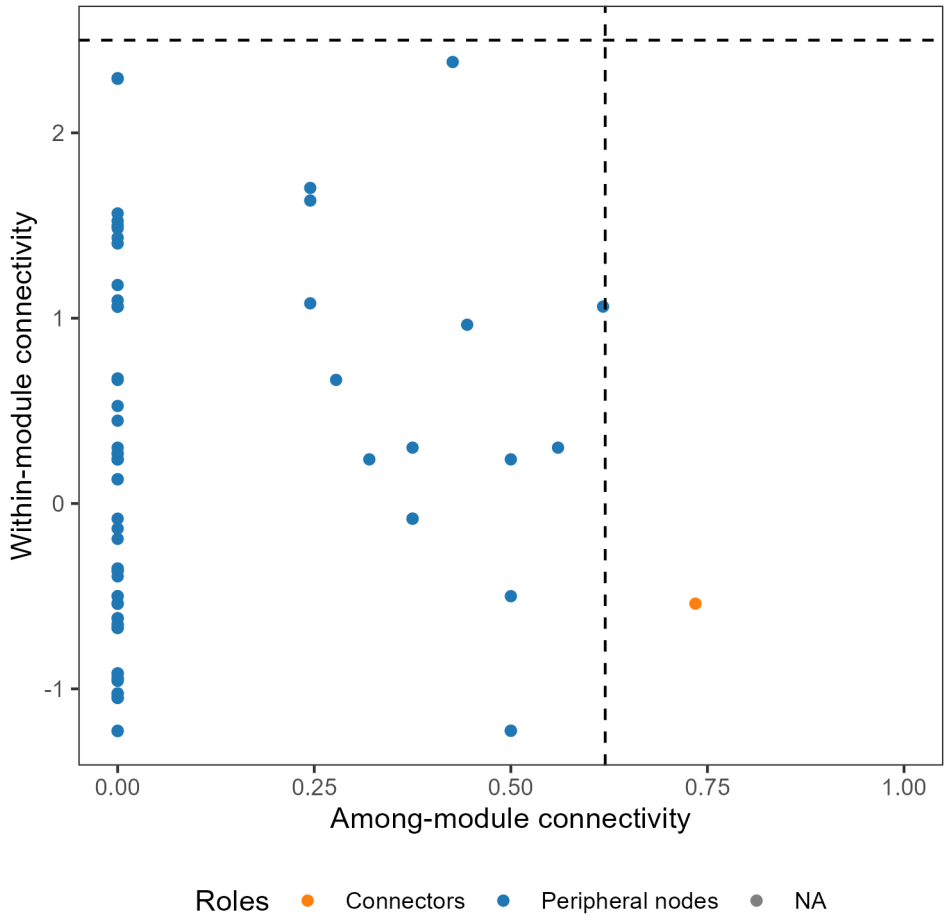

Figure 20: Taxa roles in the healthy cohort.

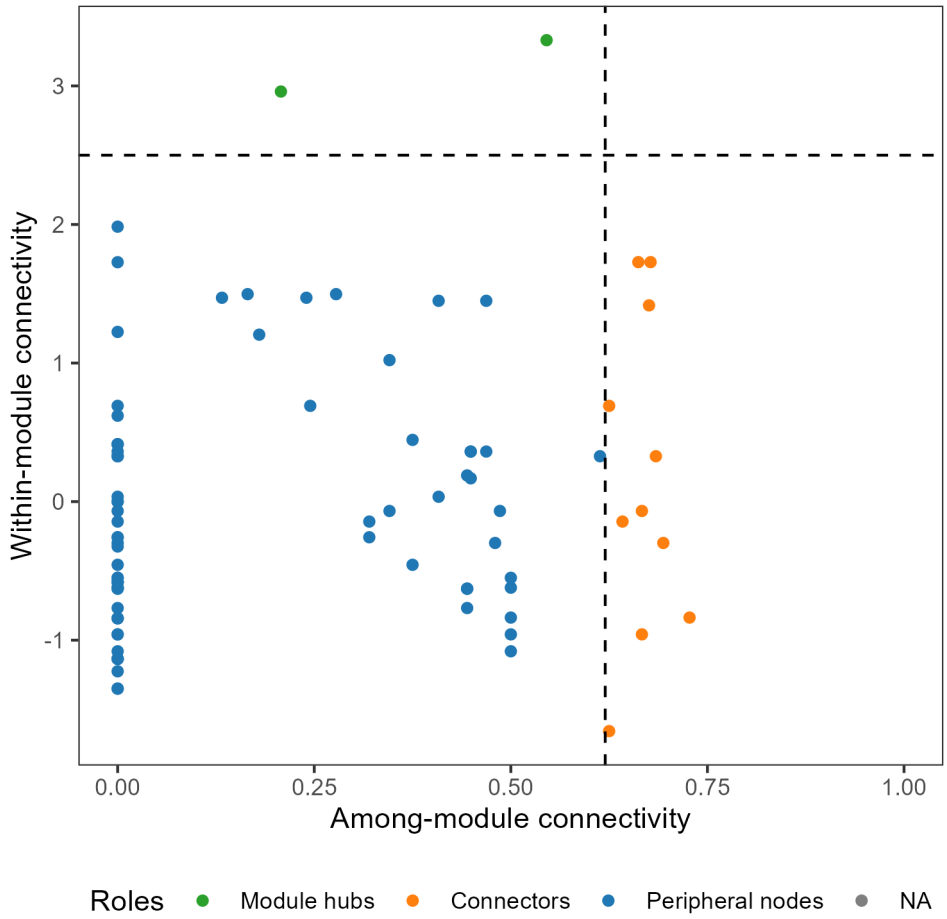

Figure 21: Taxa roles in the abnormal cohort.

In Figure **??** and Figure **??**, we can see which phylum were categorized into which roles. We see that the connectors in the abnormal cohort were from the Bacillota phylum.

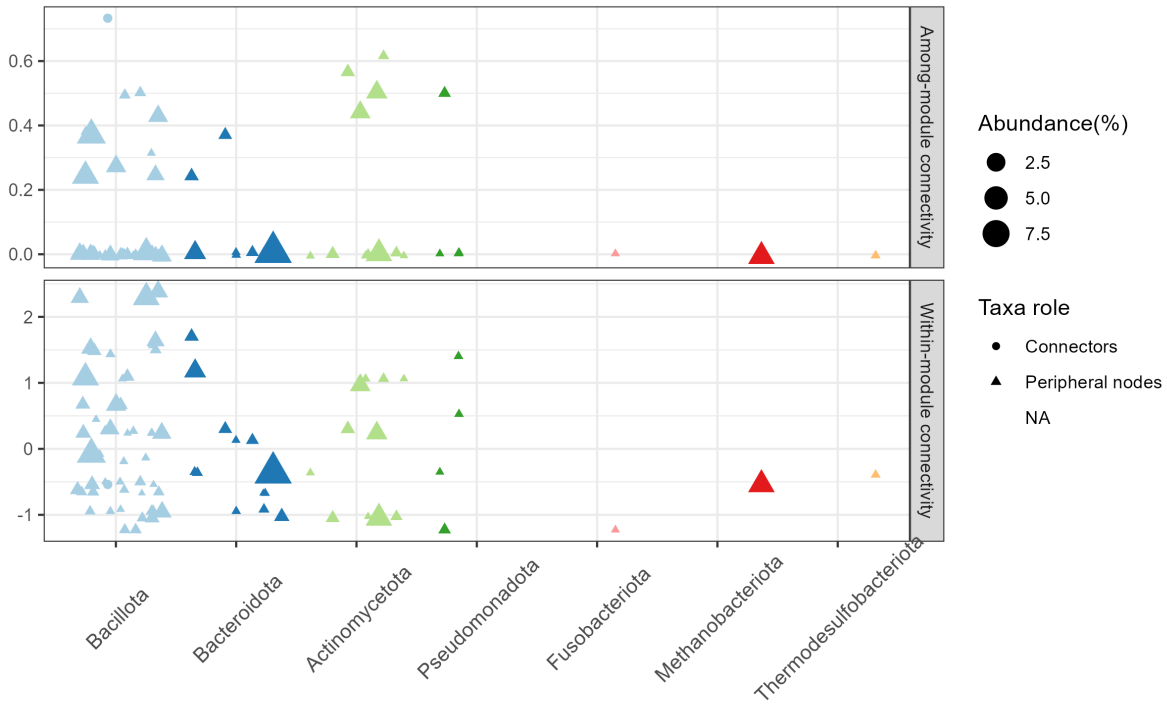

Figure 22: Taxa roles in the healthy cohort, by phylum.

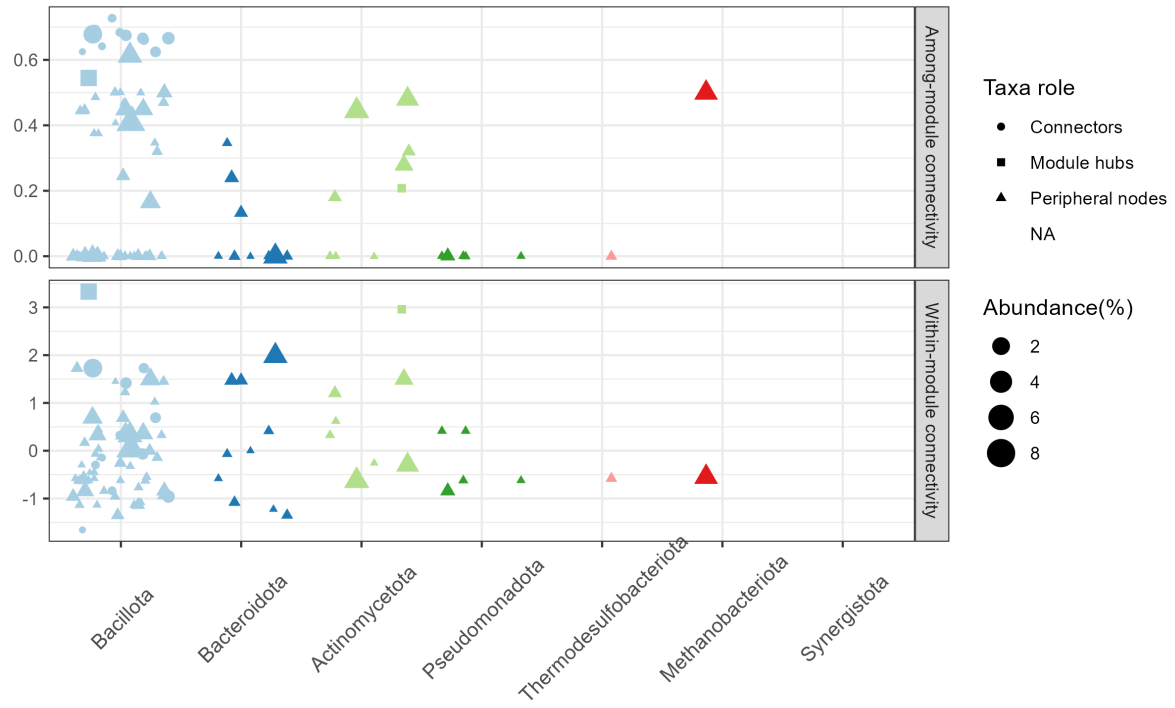

Figure 23: Taxa roles in the abnormal cohort, by phylum.

To summarize the abundance patterns of network modules, module eigengenes were calculated separately for the healthy and abnormal co-occurrence networks. Associations between module eigengenes and host or environmental variables were evaluated using the trans_env class. For each dataset, environmental variables including dietary fiber intake, physical activity (MET minutes per week), age, total caloric intake, body mass index (BMI), and circulating adiponectin levels were selected. Spearman rank correlations were calculated between module eigengenes and environmental variables, with module eigengene tables supplied as the response matrix. We see in Figure **??** that no variables were significantly associated with any of the modules, whereas in Figure **??** module 6 was significantly associated with exercise (METs).

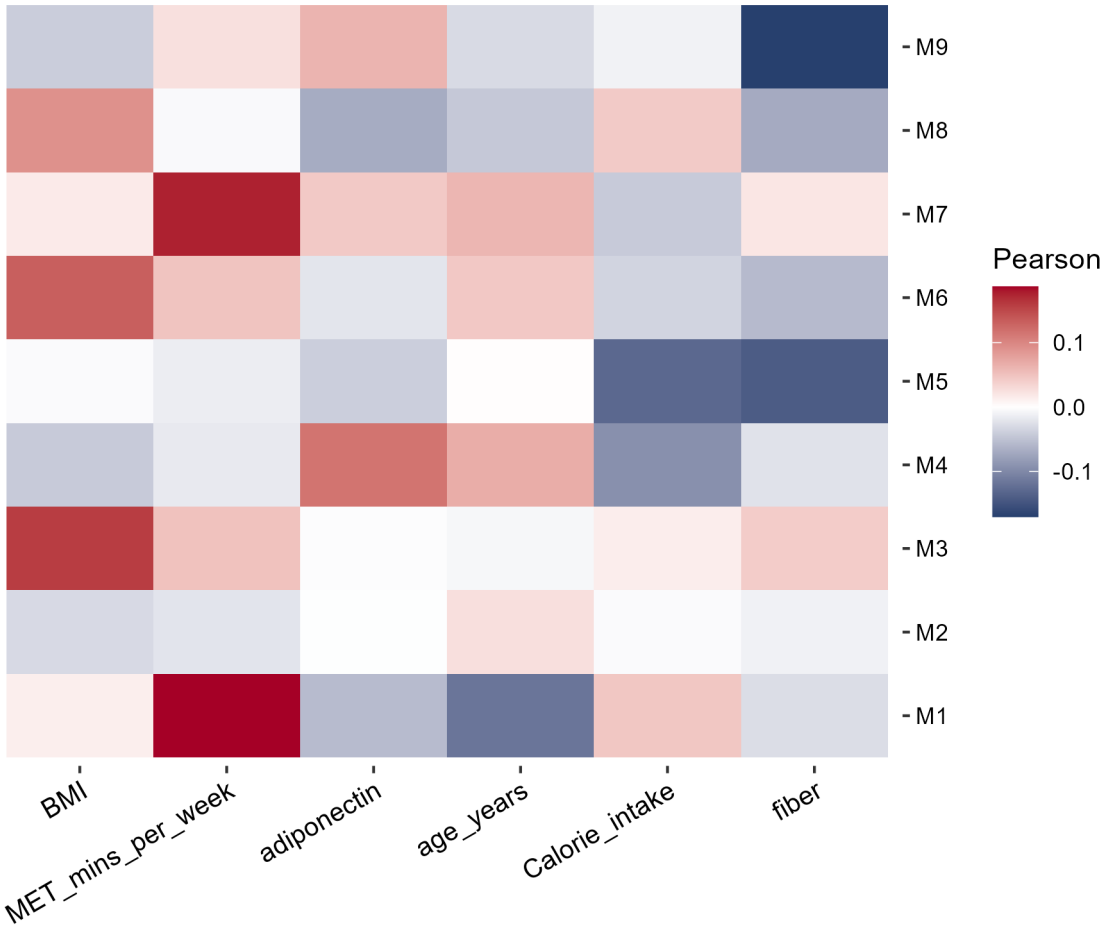

Figure 24: Heatmap of Spearman correlation of module eigengenes with metadata attributes in the healthy cohort.

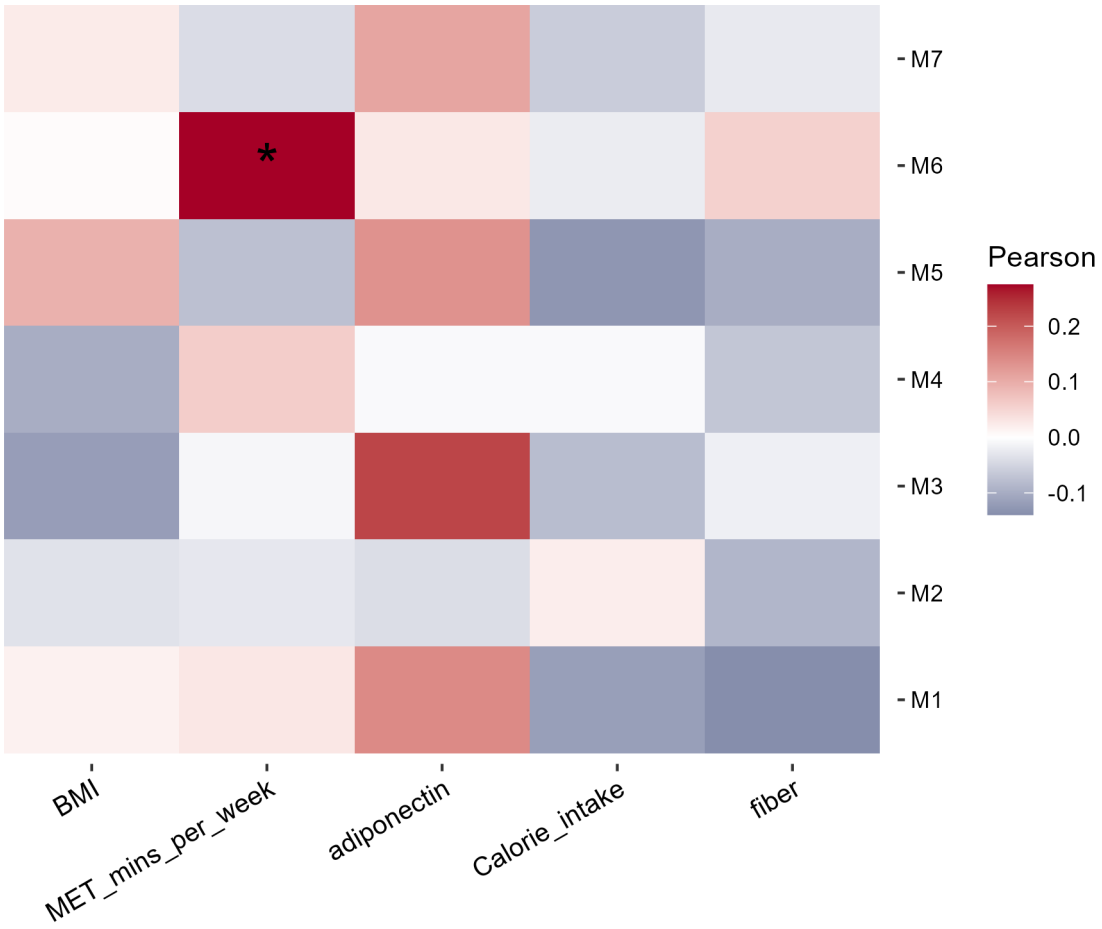

Figure 25: Heatmap of Spearman correlation of module eigengenes with metadata attributes in the abnormal cohort.

2.11 Aim 4: Functional analysis

Functional potential of the microbial communities was inferred using PICRUSt2, and predicted KEGG Ortholog (KO) abundances were obtained from the unstratified metagenome output. Differential abundance analysis (DAA) of predicted KOs was performed separately for healthy and abnormal cardiometabolic status groups, with fibre and exercise as contrasts. Statistical testing was performed using the edgeR framework, which models count data using negative binomial distributions and accounts for differences in library size. Significantly differentially abundant KOs were subsequently annotated, specifying KEGG Ortholog pathways and enabling mapping of KO identifiers to KEGG pathway annotations.

A threshold of an absolute log2FC of 4 and a FDR under 0.01 were selected. In Figure **??**, these pathways were visualized.

Compared to the abnormal cardiometabolic groups, we observed a greater number of increased pathways overall. Specifically, in the adequate exercise group, several metabolic and signalling pathways were increased. The adequate fibre group showed an increase in protein metabolism and cellular signalling, as well as a decrease in some signal transduction and energy metabolism. In contrast, we observed a larger number of decreased pathways in the abnormal cardiometabolic groups. Adequate exercise promoted few pathways including the metabolism of cofactors and vitamins, as well as cellular processes. However, there was a large decrease in infectious disease pathways and xenobiotic biodegradation. Within the adequate fibre group, the greatest decrease was observed in the infectious disease pathway.

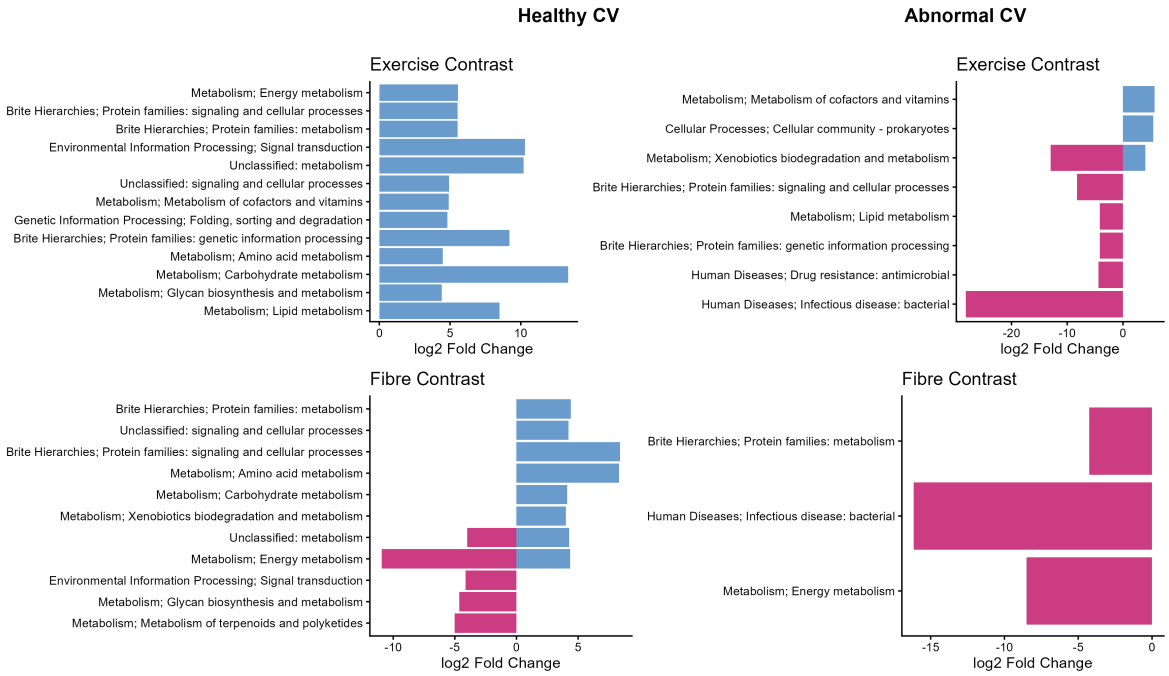

Figure 26: Log2FC plots of pathways.

To see the correlation of pathways, KO abundance tables were first transformed to relative abundances by normalizing each sample to its total KO count. For each cardiometabolic status group, KOs identified as significantly differentially abundant in fibre and exercise contrasts were selected based on an adjusted p-value under 0.01 and an absolute log2FC > 4. Spearman rank correlation matrices were calculated among the selected differentially abundant KOs using sample-wise relative abundance profiles. Correlation analyses were conducted separately for fiber- and exercise-associated KOs within each cardiometabolic status group. These correlations can be found in Figure **??**, Figure **??**, Figure **??**, and Figure **??**.

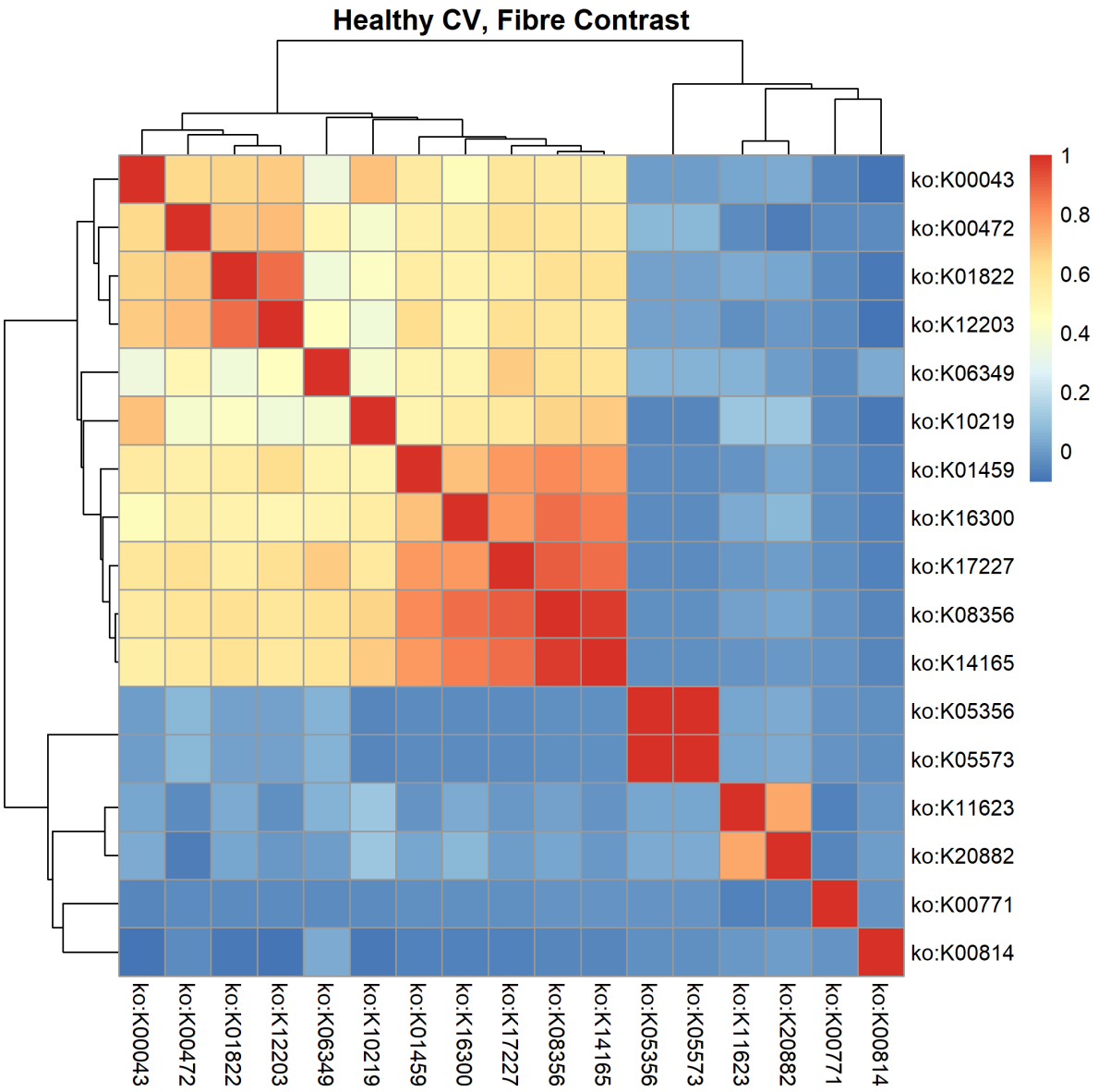

Figure 27: Heatmap of KO terms with fibre contrast, in the healthy cohort.

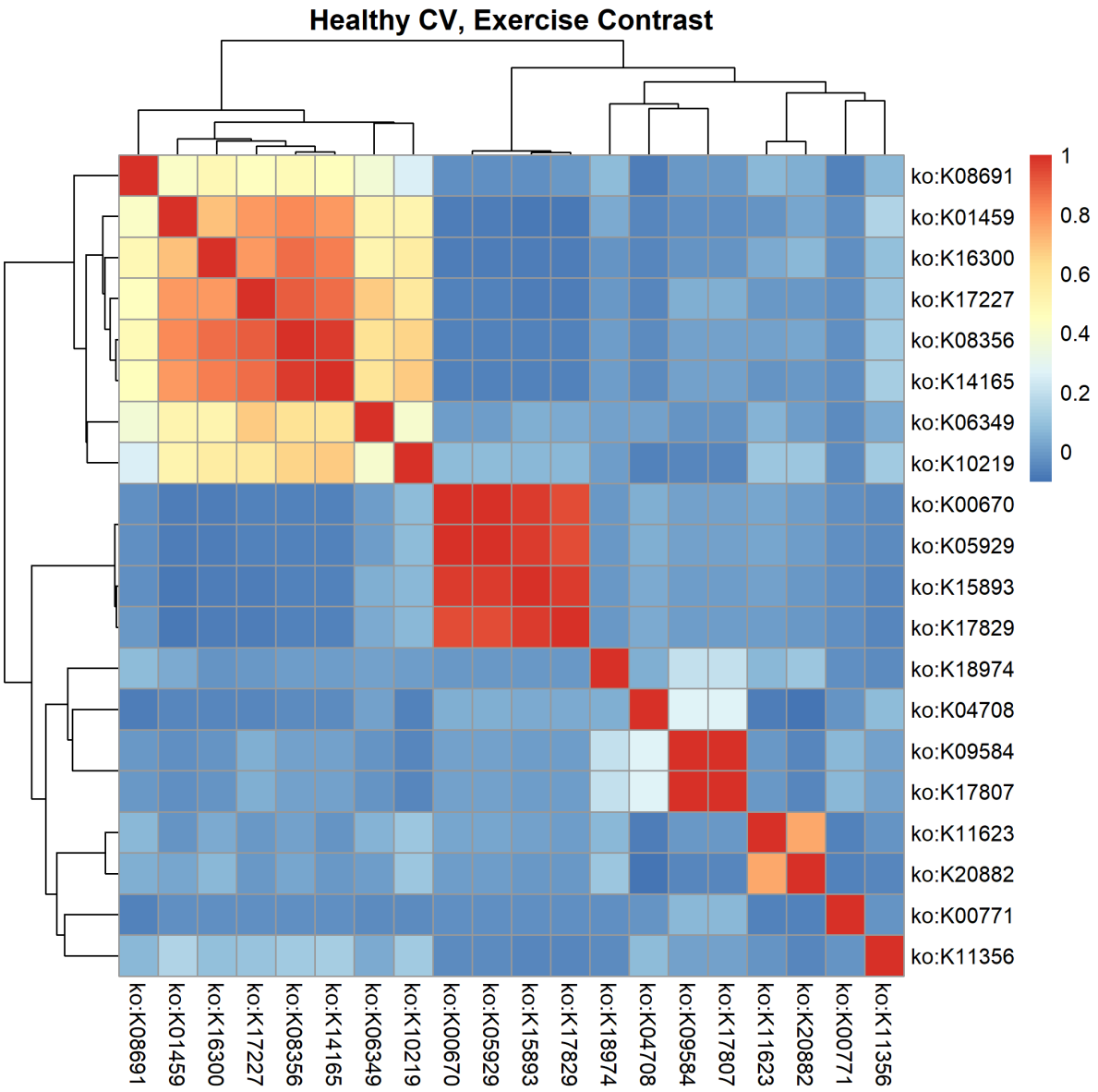

Figure 28: Heatmap of KO terms with exercise contrast, in the healthy cohort.

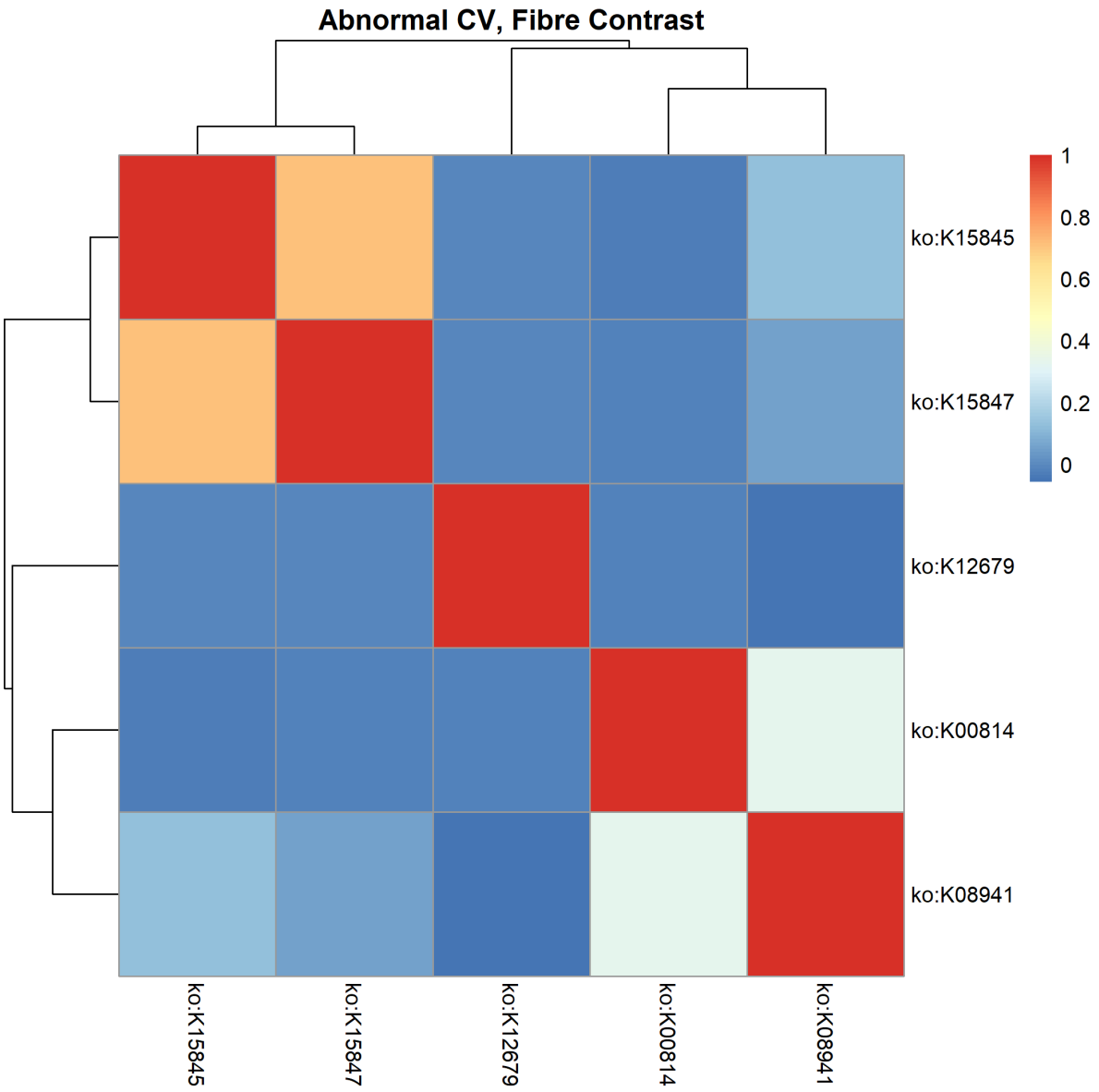

Figure 29: Heatmap of KO terms with fibre contrast, in the abnormal cohort.

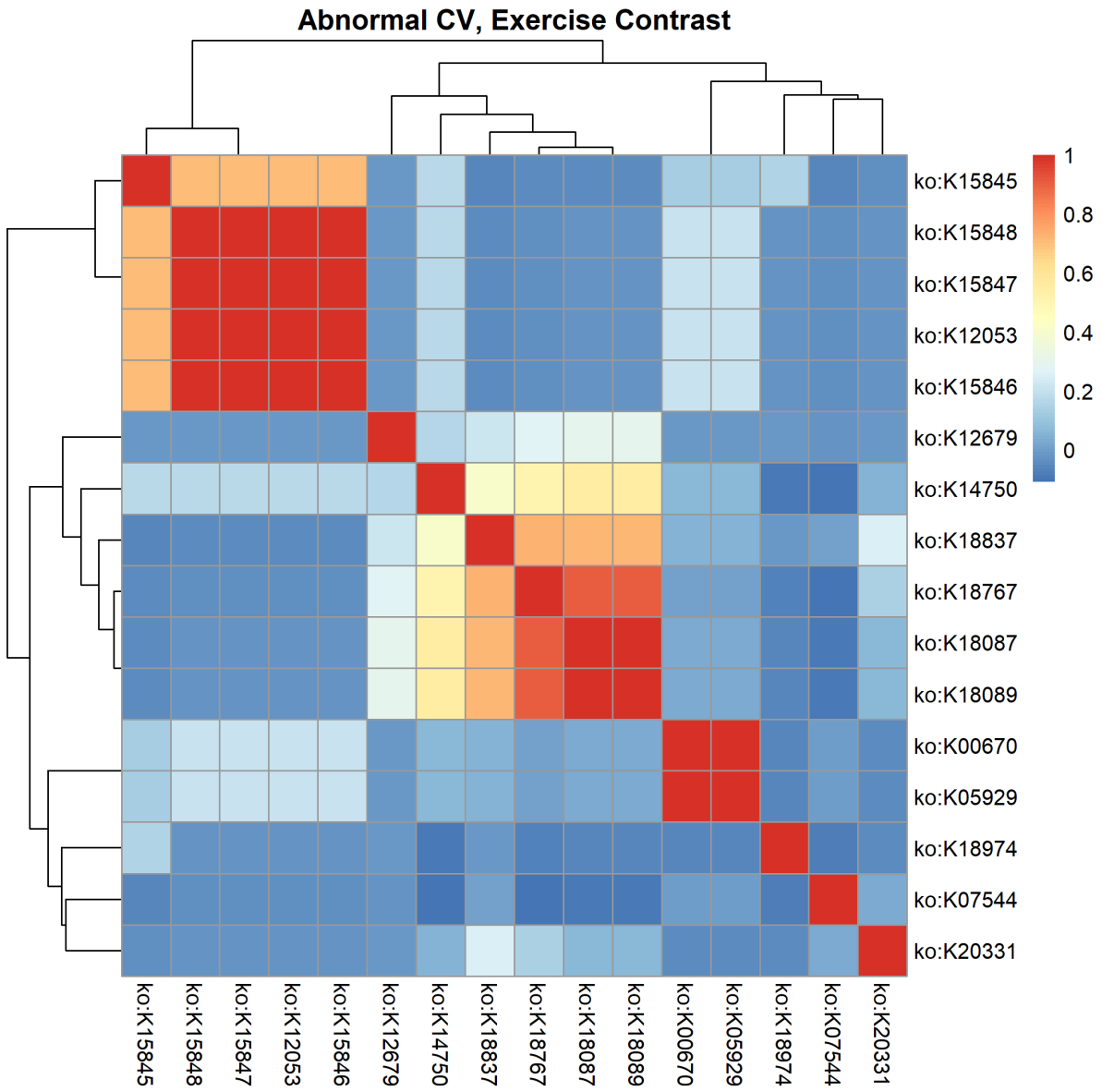

Figure 30: Heatmap of KO terms with exercise contrast, in the abnormal cohort.

Lastly, multivariate patterns in predicted functional composition were explored using principal component analysis (PCA) of KEGG Ortholog (KO) abundance profiles. PCA was performed separately for healthy and abnormal cardiometabolic status groups. KO abundance tables were used as input, and sample metadata were incorporated to define grouping variables. Ordinations were generated for contrasts based on dietary fiber intake group and physical activity (exercise) group. These PCA plots can be found in Figure **??**, Figure **??**, Figure **??**,

and Figure **??**.

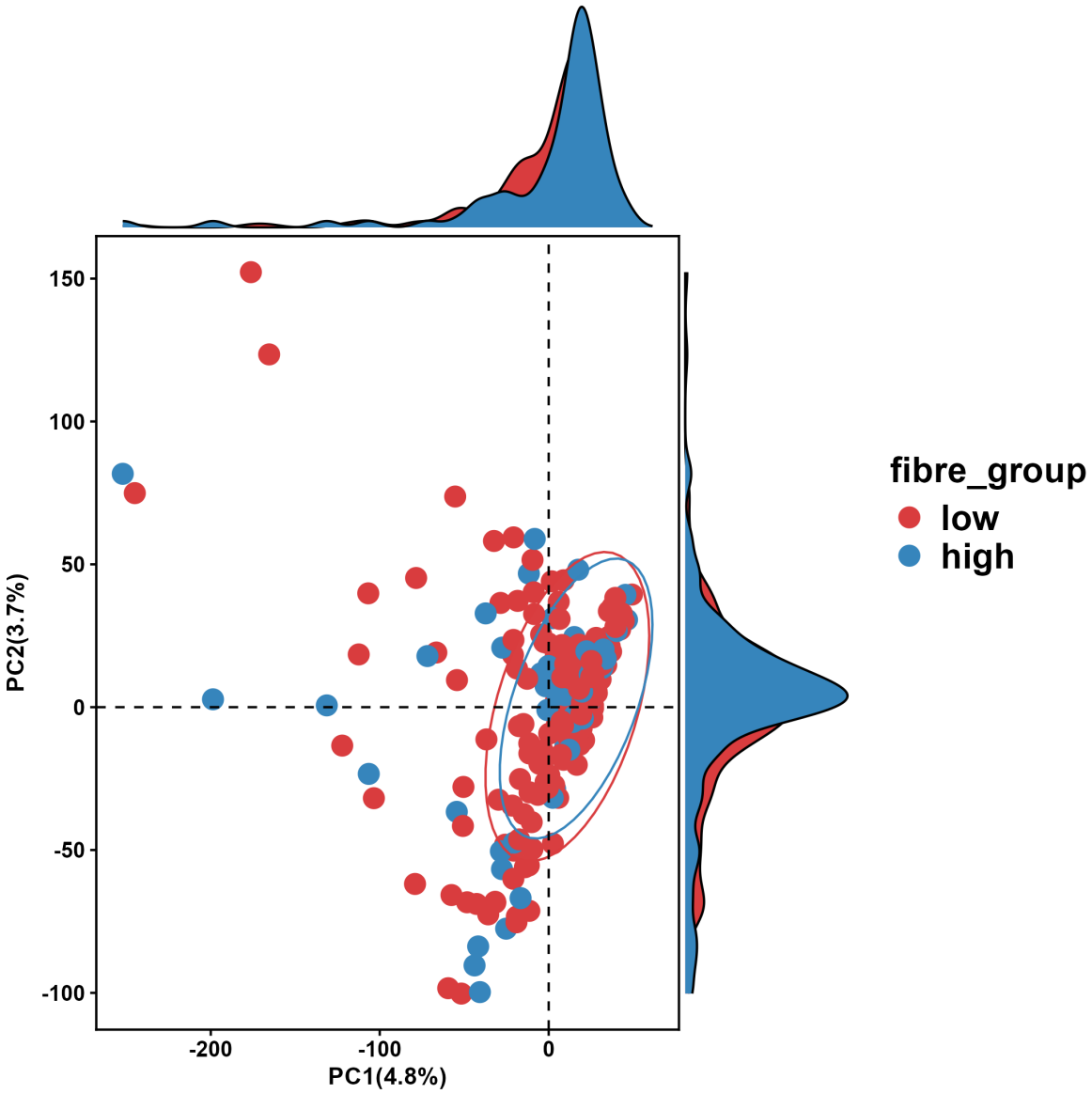

Figure 31: PCA of KO profiles with fibre contrast, in the healthy cohort.

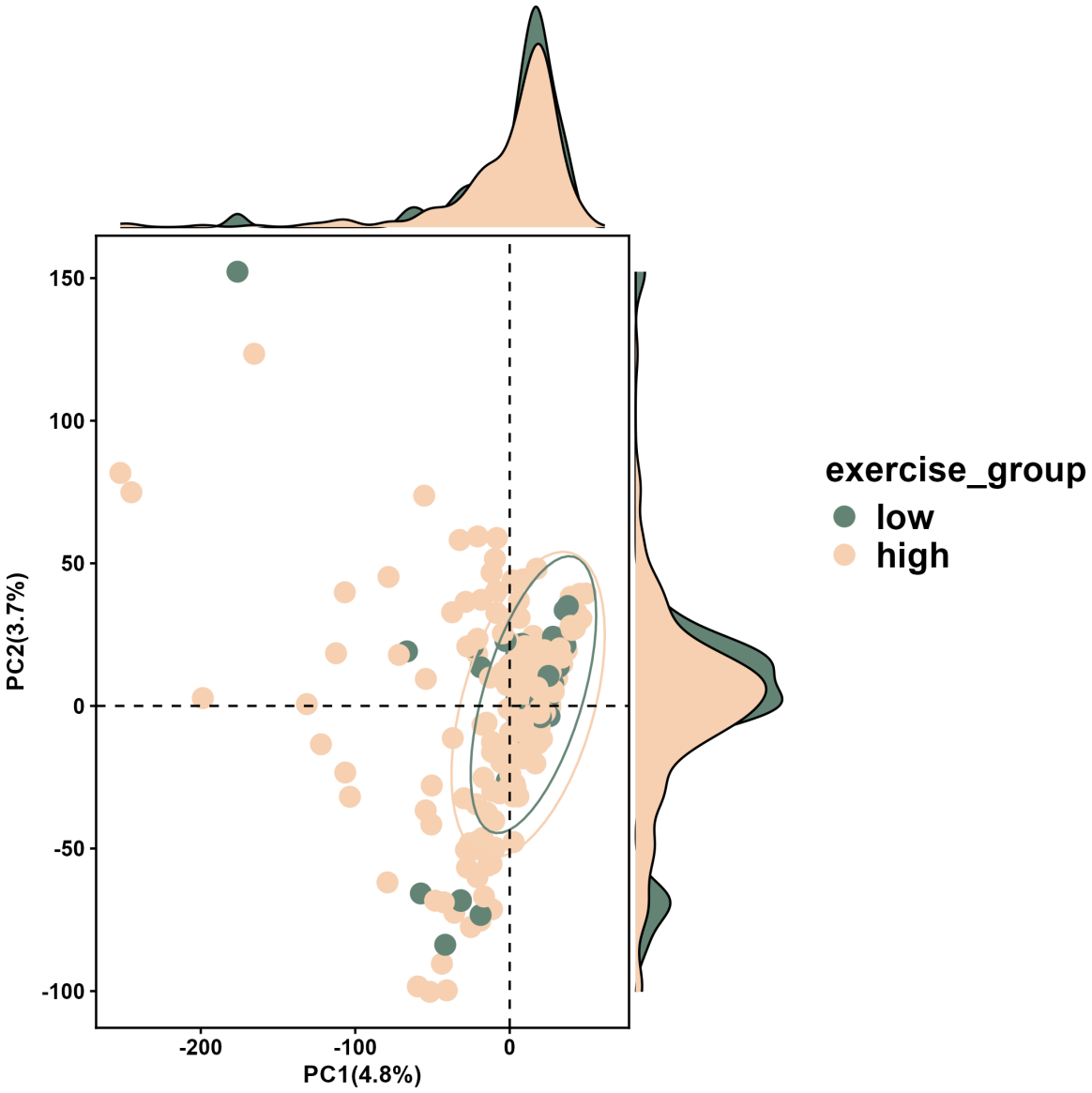

Figure 32: PCA of KO profiles with exercise contrast, in the healthy cohort.

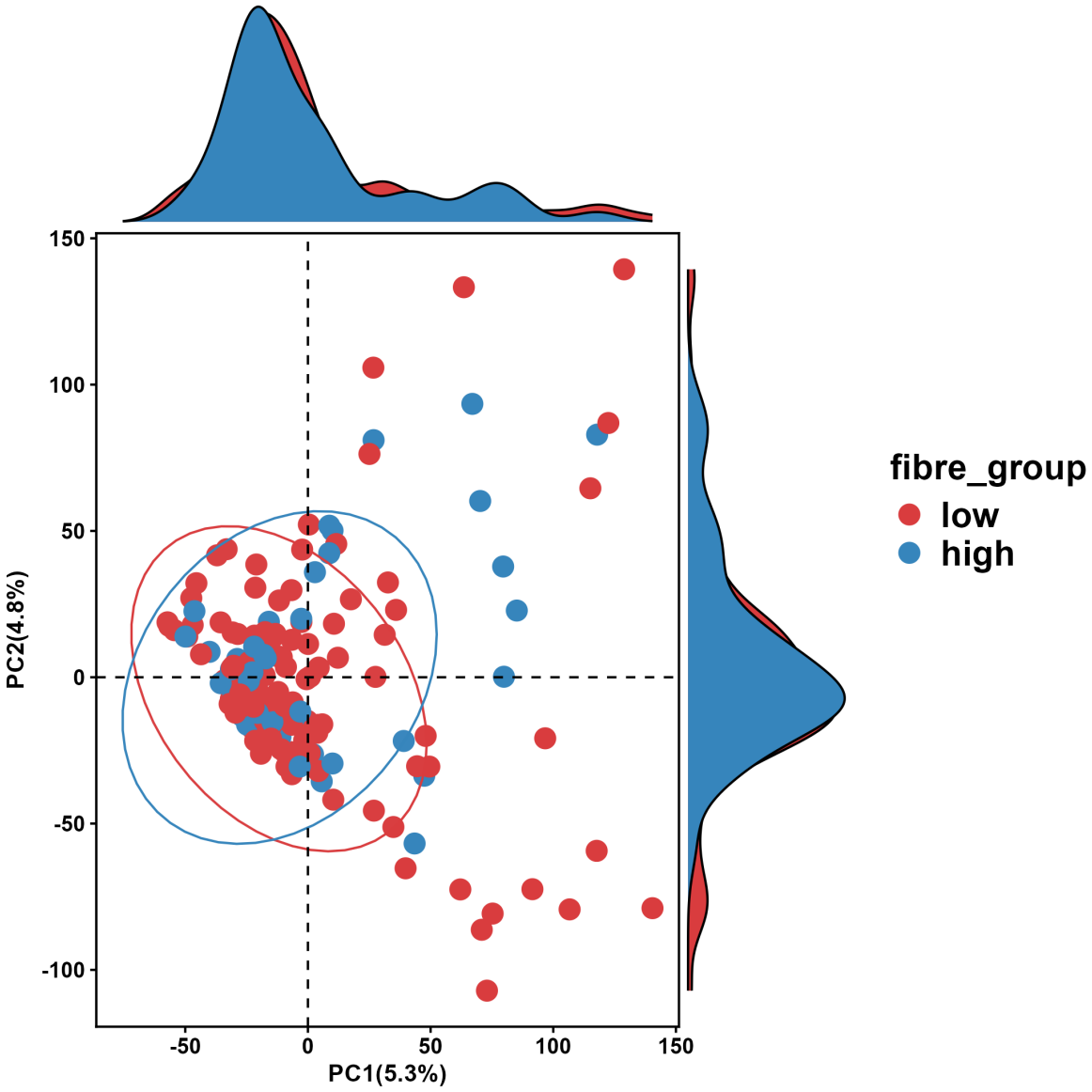

Figure 33: PCA of KO profiles with fibre contrast, in the abnormal cohort.

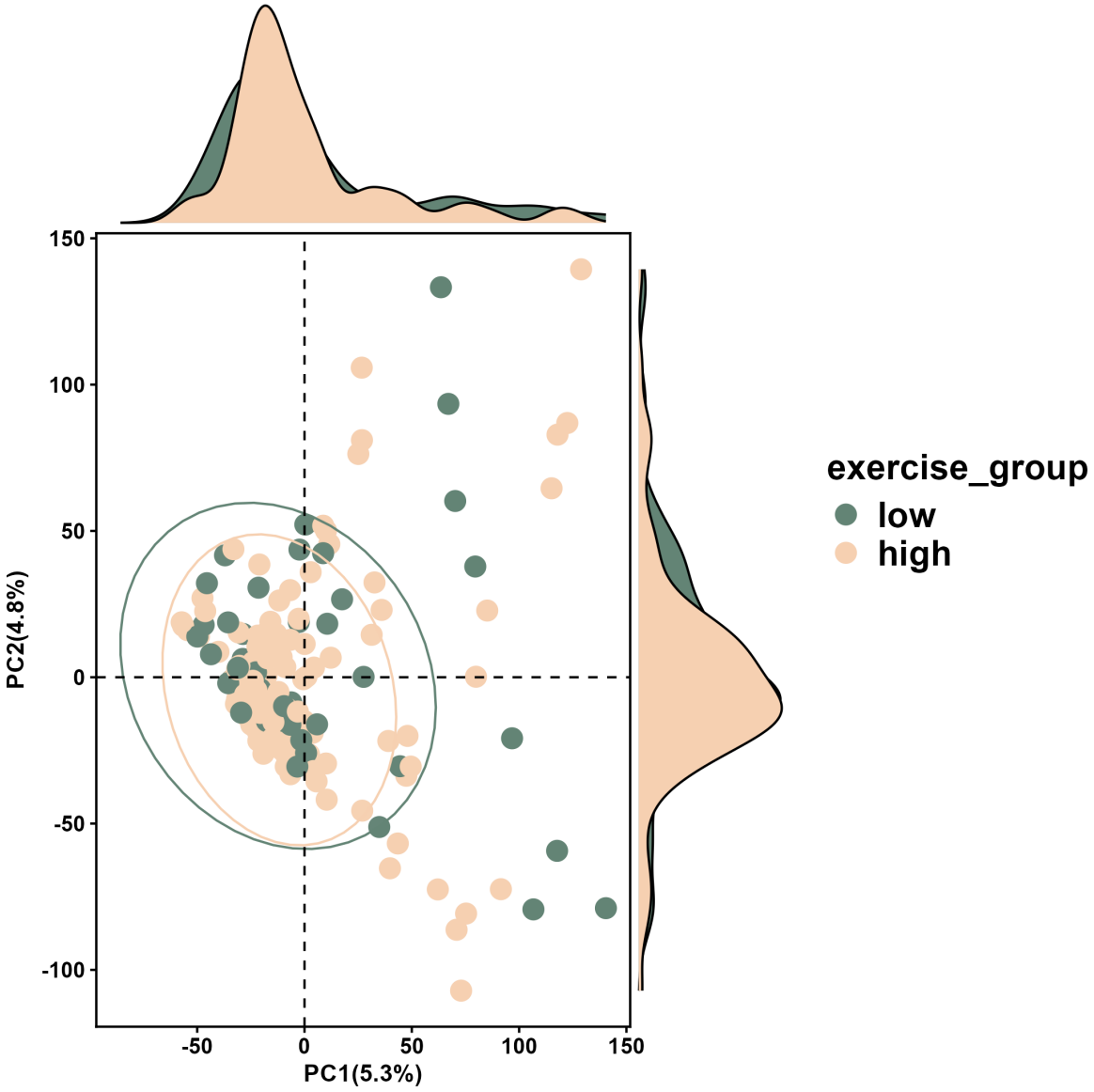

Figure 34: PCA of KO profiles with exercise contrast, in the abnormal cohort.

3 Conclusion

In a Colombian population newly undergoing westernization, we found that fibre and exercise promote several SCFA-producing bacteria as well as reduce pathogenic pathways. Overall, this may have a synergistic effect and support cardiometabolic health. Despite several studies

reporting that fibre and exercise promote greater microbial diversity, it appears that these changes may need more time to accumulate.

References

Bolyen, Evan, Jai Ram Rideout, Matthew R Dillon, Nicholas A Bokulich, Christian C Abnet, Gabriel A Al-Ghalith, Harriet Alexander, et al. 2019. "Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2." *Nature Biotechnology* 37 (8): 852–57.

Callahan, BJ, PJ McMurdie, MJ Rosen, AW Han, AJA Johnson, and SP Holmes Dada. n.d. "High-Resolution Sample Inference from Illumina Amplicon Data., 2016, 13." *DOI: Https://Doi. Org/10.1038/Nmeth* 3869: 581–83.

Cuesta-Zuluaga, Jacobo de la, Vanessa Corrales-Agudelo, Eliana P Velásquez-Mejía, Jenny A Carmona, José M Abad, and Juan S Escobar. 2018. "Gut Microbiota Is Associated with Obesity and Cardiometabolic Disease in a Population in the Midst of Westernization." *Scientific Reports* 8 (1): 11356.