METWARE
迈 维 代 谢

微生物宏基因组、代谢组、表型的关联分析

武汉迈特维尔生物科技有限公司

www.metware.cn

湖北省武汉市东湖技术开发区高新大道 666 号光谷生物城 C2-3 栋

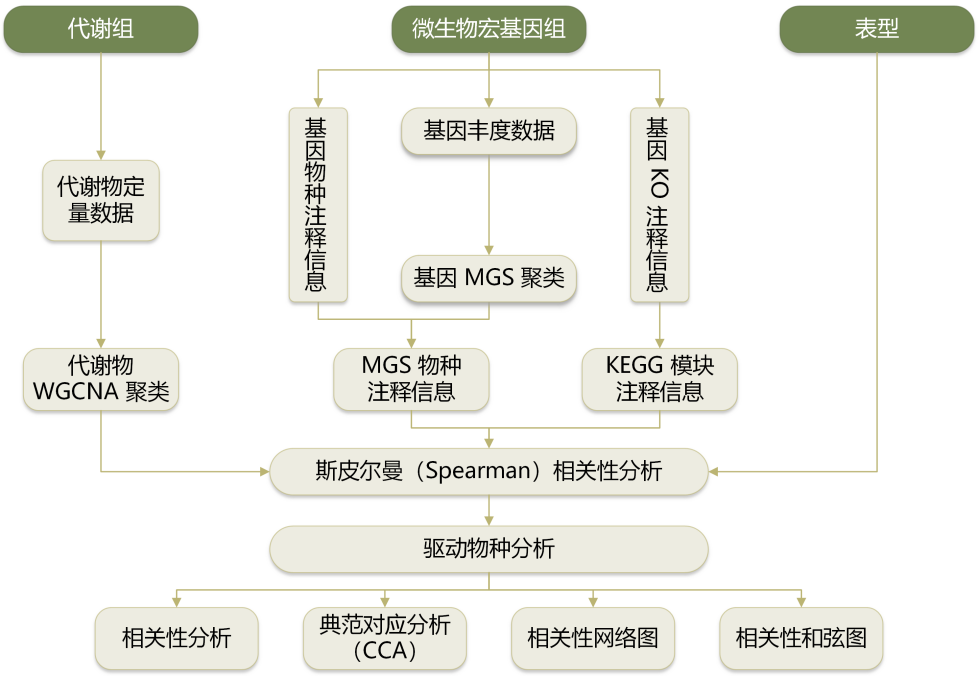
宏基因组代谢个性化联合分析报告

1 摘要

宏基因组是环境中全部微小生物遗传物质的总和，是用来研究微生物群落组成、功能基因、代谢产物的学科。采用宏基因组学研究技术分析微生物菌群，能揭示菌群与宿主的关系，进而挖掘微生物种群结构、进化关系、功能活性以及与环境之间的关系。代谢组学能测量宿主生态系统某个时间点的代谢变化。因此，为了研究宿主微生物群落结构的改变可能引起的表型变化，需要将代谢组和宏基因组进行关联分析。

由于宏基因组和代谢组数据具有高维度和复杂性等特征，直接将两组数据进行整合分析具有很大挑战。因此我们设计的关联分析方案基于降维的思想，将代谢组数据通过加权共表达网络分析（WGCNA）(Langfelder and Horvath 2008) 方法进行数据驱动降维，微生物组物种数据通过 Canopy 聚类进行数据驱动降维 (Pedersen et al. 2018)，微生物组基因数据通过 KEGG (Kanehisa et al. 2016, 2017) 模块分类法进行知识驱动降维，筛选与表型显著相关的数据特征进行跨组学关联分析，鉴定与表型显著相关的 KEGG 功能模块的驱动物种。

宏基因组与代谢组联合分析流程如下：



联合分析流程

2 宏基因组和代谢组数据预处理

2.1 代谢组数据预处理

加权共表达网络分析 (Langfelder and Horvath 2008)（Weighted correlation network analysis，WGCNA）是用来描述不同样品之间基因关联模式的系统生物学方法，可以用来鉴定表达模式相似的基因集。本项目使用 WGCNA 方法对代谢组数据进行降维处理，将具有相似表达模式的代谢物聚成不同的 clusters，从而减少分析的维度。每个聚类在每个样本中的定量数据定义为，该聚类包含的所有代谢物在该样本中的定量数据取中位值。聚类的定量数据后续与微生物组、表型特征进行关联分析。程序默认设置时会对数据进行 log2 标准化。

表 1 代谢物 WGCNA 聚类的定量数据

Metabolome_cluster	AA1	AA2	AA3	AA4	BB1
black	162630	178715	700510.0	1917300	206785.0
blue	103665	128655	53702.5	97066	54290.5
brown	1057500	488205	222010.0	509810	154300.0
green	46007	78865	671150.0	198090	658770.0
magenta	481290	810170	257140.0	401730	144550.0

文件路径：2.Cluster/metabolome_cluster/module.quant.xlsx

- Metabolome_cluster：代谢物聚类结果
- 其他：代谢物聚类的定量数据

2.2 宏基因组数据预处理

使用数据驱动和知识驱动两种方法对微生物数据进行降维处理。数据驱动降维方法为 Canopy(Pedersen et al. 2018) 聚类，将具有相似表达模式的基因进行聚类，生成不同的 MGS(Nielsen et al. 2014)（Metagenomics Species），然后根据基因的物种注释信息，将 MGS 注释为优势物种。每个 MGS 中有大于 50% 的基因注释到某个物种，则将 MGS 注释为该物种。知识驱动降维方法为 KEGG 模块分析。方法为，首先计算 KEGG Orthology（KO）的定量数据，定义为注释到 KO 的所有基因的定量数据的中位值。然后计算 KEGG 模块的定量数据，定义为模块内所有 KO 的定量数据的中位值。

表 2 微生物基因 MGS 降维的定量数据

Microbiome_cluster	Taxonomy	Ratio	AA1	AA2
CAG006	k__Bacteria	0.5851852	3.00e-07	1.60e-06
CAG011	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides	0.8095238	8.00e-07	3.00e-06
CAG012	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Unclassified;s__Lachnospiraceae bacterium A2	0.8571429	8.00e-07	3.00e-07
CAG015	k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Staphylococcaceae;g__Staphylococcus;s__Staphylococcus nepalensis	0.7714286	1.55e-05	1.40e-06
CAG017	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales	0.6862745	5.72e-05	2.96e-05

文件路径：2.Cluster/microbiome_cluster/cluster.taxon.profile.xlsx

- Microbiome_cluster：微生物基因的 MGS 聚类
- Taxonomy：MGS 聚类注释到的优势微生物类别
- Ratio：优势微生物类别的比例
- 其他：样本的定量数据

表 3 微生物基因 KEGG 模块降维的定量数据

KEGG_module	Level1	Level2
M00001	Carbohydrate metabolism	Central carbohydrate metabolism
M00002	Carbohydrate metabolism	Central carbohydrate metabolism
M00003	Carbohydrate metabolism	Central carbohydrate metabolism
M00307	Carbohydrate metabolism	Central carbohydrate metabolism
M00009	Carbohydrate metabolism	Central carbohydrate metabolism

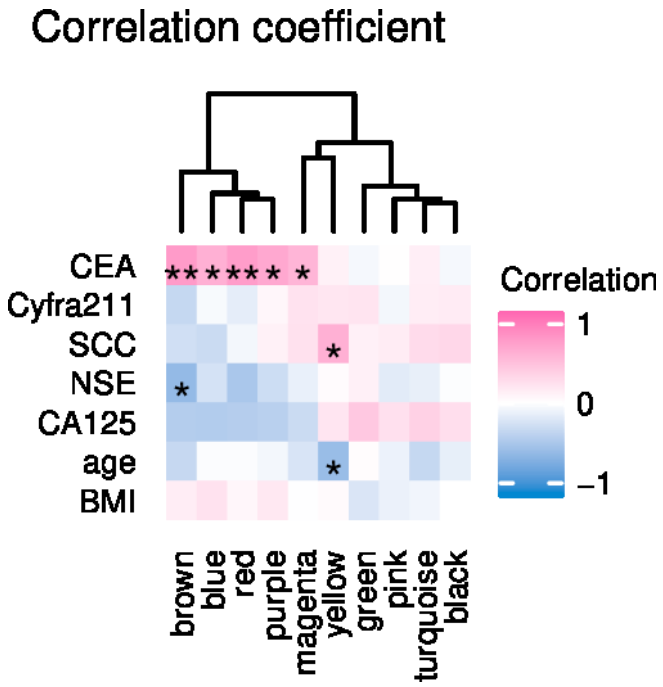
文件路径：2.Cluster/microbiome_cluster/KEGG.module.profile.xlsx

- KEGG_module：KEGG 模块
- Level1：KEGG 模块第一级分类描述
- Level2：KEGG 模块第二级分类描述
- Module_description：KEGG 模块功能详细描述
- 其他：样本的定量数据

3 微生物、代谢物和表型的关联分析

3.1 代谢物聚类与表型关联

将代谢物通过数据驱动聚类（WGCNA）进行降维处理后生成不同的聚类，每个聚类中代谢物的中位值与表型数据进行 Spearman 关联分析，寻找显著关联的聚类进行后续的跨组学关联分析。结果如下图所示。



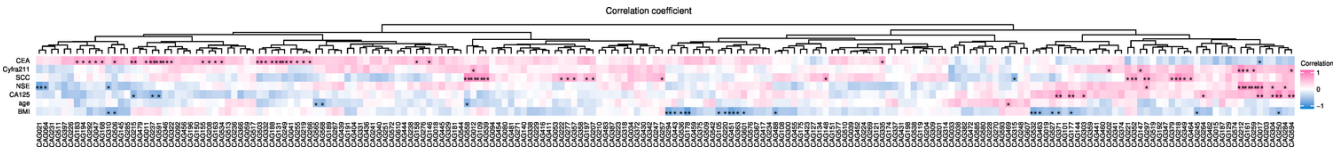
代谢物聚类与表型的相关性热图

横坐标表示代谢物聚类，纵坐标表示表型。红色表示正相关，蓝色表示负相关，其中 * 表示相关系数显著性检验的 P 值 < 0.05，** 表示 P 值 < 0.01。

文件路径：3.Correlation_analysis/1.metabolome_phenotype_correlation

3.2 微生物 MGS 与表型关联

将微生物基因通过数据驱动（MGS）法进行降维处理后生成不同的 MGS，微生物 MGS 中所有基因丰度取中位值，然后与表型数据进行 Spearman 关联分析，寻找显著关联的 MGS 进行后续的跨组学关联分析。结果如下图所示。



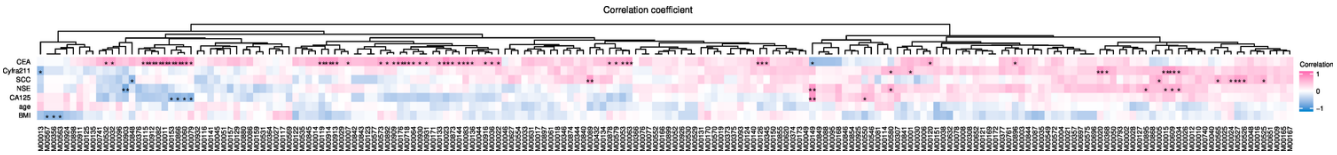
微生物聚类与表型的相关性热图

横坐标表示微生物 MGS，纵坐标表示表型。红色表示正相关，蓝色表示负相关，其中 * 表示相关系数显著性检验的 P 值 < 0.05，** 表示 P 值 < 0.01。

文件路径：3.Correlation_analysis/2.microbiome_phenotype_correlation

3.3 微生物 KEGG 功能模块与表型关联

将微生物基因通过 KEGG 功能模块进行降维处理后，生成不同的 KEGG 功能模块，每个 module 取所有基因丰度的中位值，然后与表型数据进行 Spearman 关联分析，寻找显著关联的模块进行后续的跨组学关联分析。结果如下图所示。



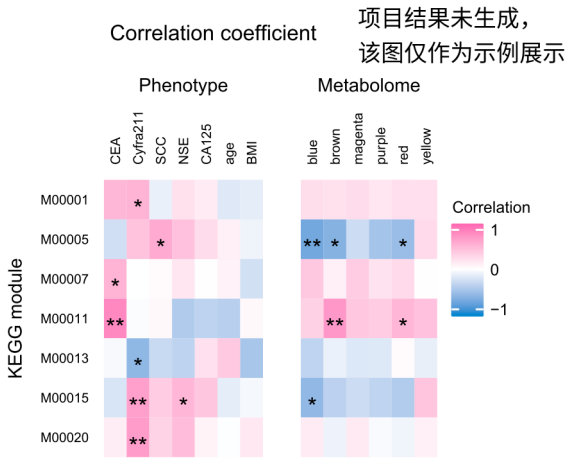
微生物 KEGG 功能模块与表型的相关性热图

横坐标表示微生物 KEGG 模块，纵坐标表示表型。红色表示正相关，蓝色表示负相关，其中 * 表示相关系数显著性检验的 P 值 < 0.05，** 表示 P 值 < 0.01。

文件路径：3.Correlation_analysis/3.KEGG_module_phenotype_correlation

3.4 微生物 KEGG 功能模块与代谢物聚类、表型关联

将与表型显著关联的代谢物聚类和与表型显著关联的 KEGG 功能模块进行跨组学关联，揭示微生物与代谢物之间的相互作用关系。结果如下图所示。



微生物 KEGG 功能模块与代谢物聚类、表型的相关性热图组合图

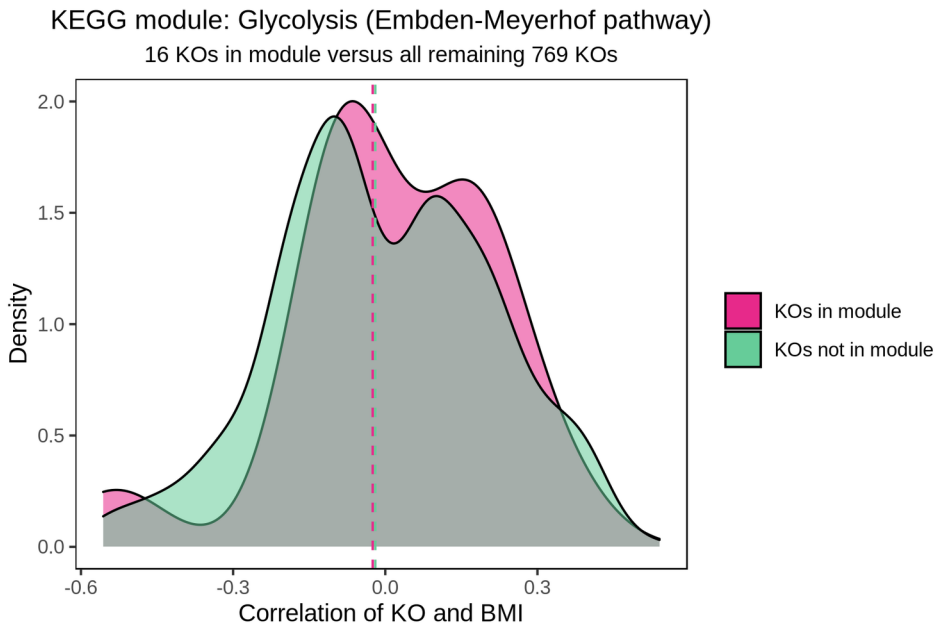
图形纵坐标表示不同的微生物 KEGG 功能模块。左侧为微生物 KEGG 模块与表型的相关性热图，横坐标表示表型；右侧为微生物 KEGG 模块与代谢物聚类的相关性热图，横坐标表示不同的代谢物聚类。红色表示正相关，蓝色表示负相关，其中 * 表示相关系数显著性检验的 P 值 < 0.05，** 表示 P 值 < 0.01。

文件路径：3.Correlation_analysis/5.KEGG_module_metabolome_phenotype

4 驱动物种与差异显著代谢物的关联

4.1 驱动物种的鉴定

为了探究与表型显著相关的 KEGG 功能模块中是哪个或哪些物种起主要的驱动作用，本项目采用 leave-one-MGS-out 方法进行驱动物种的鉴定。该方法先计算每个 KEGG 模块中每个 KO 与表型的 Spearman 相关系数，然后用 KO 相关系数的中位值表示某个 KEGG 模块与表型的相关性。为了鉴定某个 MGS 是否为驱动物种，在每个模块中通过去掉某个 MGS 的所有基因，然后重新计算每个 KO 与表型的 Spearman 相关系数，用相关系数的中位值表示去掉某个 MGS 后 KEGG 模块与表型的相关性。如果去掉某个 MGS 后，SCC 值显著变化，则认为该 MGS 对应的物种是驱动物种。

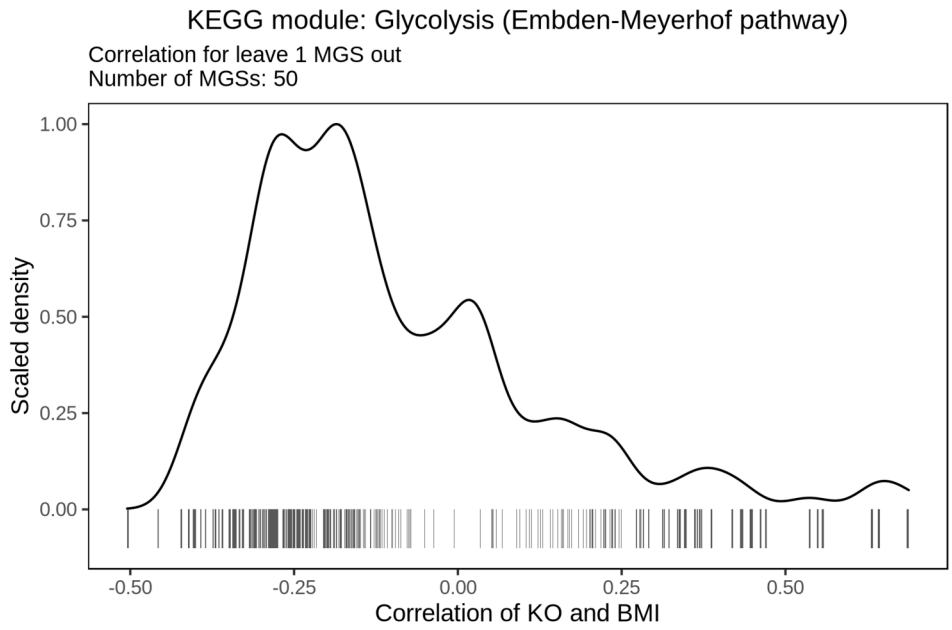


KEGG 模块的 KO 和其他所有 KO 与表型的 Spearman 相关系数密度分布图

红色区域表示模块包含的 KO 与表型的相关系数分布情况，红色虚线表示这些相关系数的中位值。绿色区域表示去掉该模块 KO 后剩余 KO 与表型的相关系数分布情况，绿色虚线表示这些相关系数的中位值。

文件路径：

4.Driver_species/1.total_driver_species/correlation.KO_phenotype.density_plot



去掉每个 MGS 后模块与表型的 Spearman 相关系数密度分布图

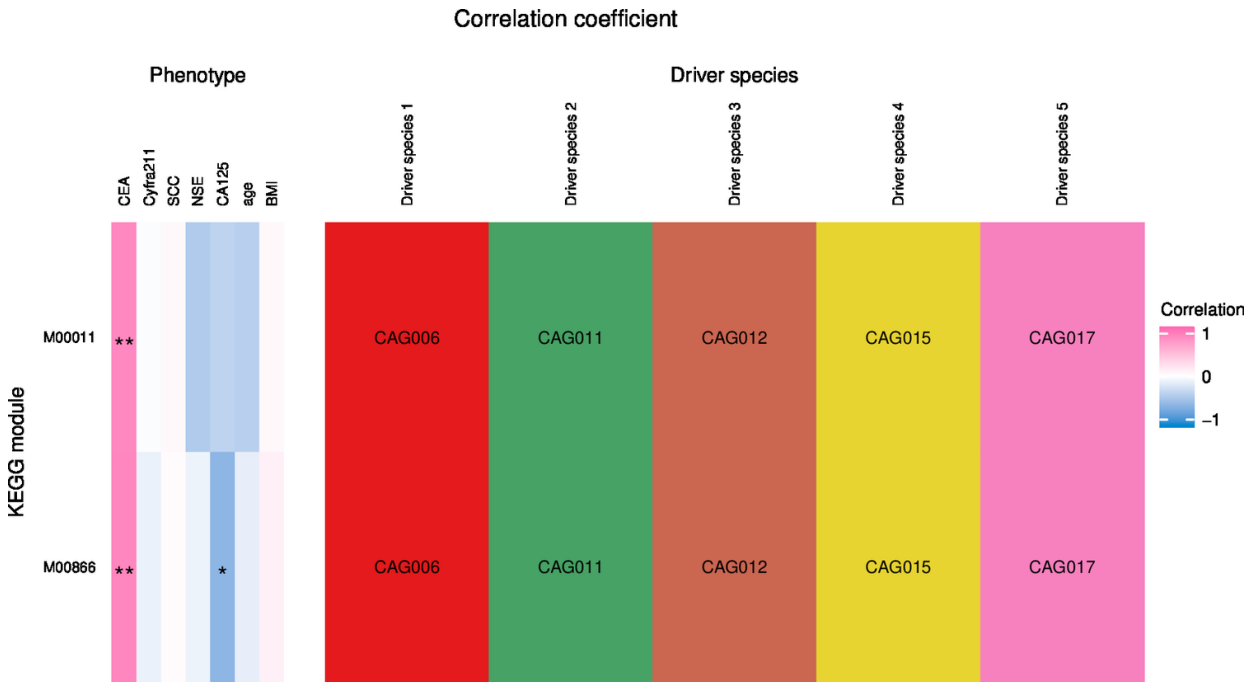
图中底部每根线表示去掉每个 MGS 后模块与表型相关系数的中位值，其中模块与表型的相关系数为模块内所有 KO 与表型相关系数的中位值。

文件路径：

4.Driver_species/1.total_driver_species/correlation.module_phenotype.leave_1_MGS_out.density_plot

4.2 驱动物种和显著代谢物的相关性

为了进一步研究驱动物种与表型显著相关代谢物之间的相互作用关系，每个 KEGG 模块挑选 top5 驱动物种，如下图所示。然后计算 top5 驱动物种与代谢物的相关性，筛选出相关系数绝对值大于 0.8，P-value < 0.05 的结果用于后续的分析。

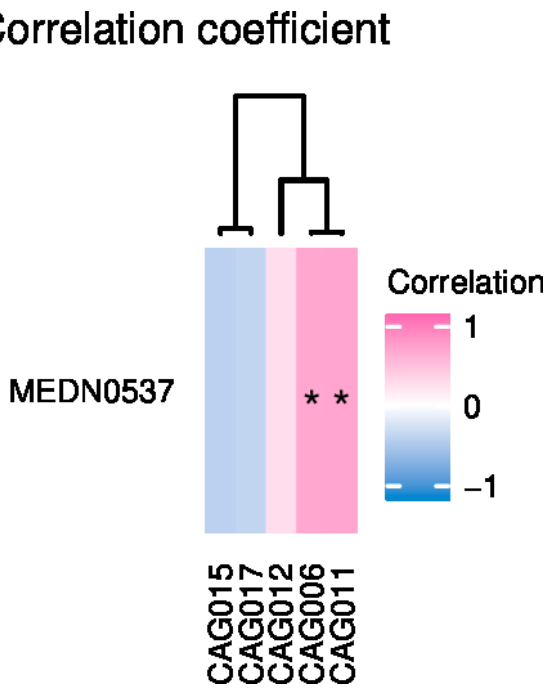


与表型显著相关的 KEGG 模块的最重要的 5 个驱动物种

左侧为微生物 KEGG 模块与表型的相关性热图，右侧为微生物 KEGG 模块的驱动物种，该图只展示与表型显著相关的微生物。红色表示正相关，蓝色表示负相关，其中 * 表示相关系数显著性检验的 P 值 < 0.05，** 表示 P 值 < 0.01。驱动物种的热图中，每个物种用不用颜色表示。

文件路径：4.Driver_species/2.top_driver_species

驱动物种与显著代谢物的相关性热图如下图所示。



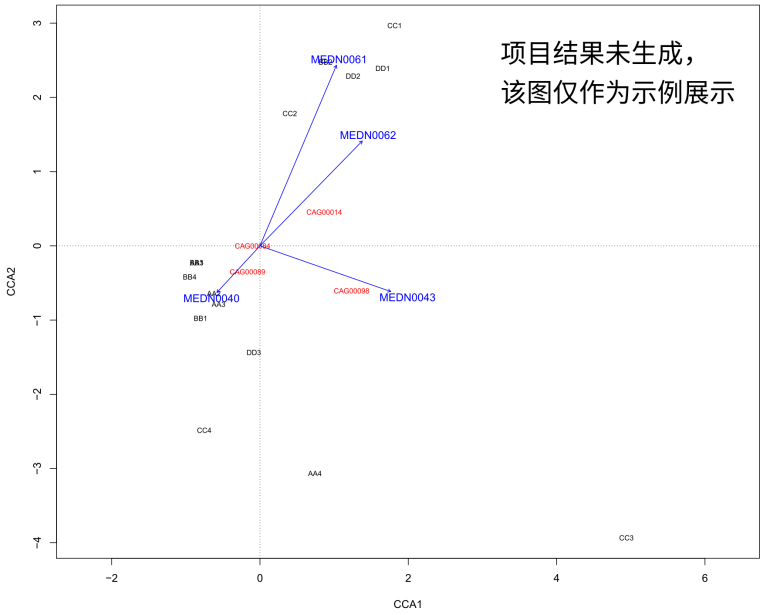
驱动物种与显著代谢物 Spearman 相关性热图

横坐标表示微生物 KEGG 模块的驱动物种，纵坐标表示与表型显著相关的代谢物。红色表示正相关，蓝色表示负相关，其中 * 表示相关系数显著性检验的 P 值 < 0.05，** 表示 P 值 < 0.01。

文件路径：4.Driver_species/4.top_driver_species_metabolome_correlation

4.3 驱动物种和显著代谢物的 CCA 分析

筛选出的显著相关结果进行典范对应分析 CCA（Canonical Correspondence Analysis）分析，结果如下图所示。



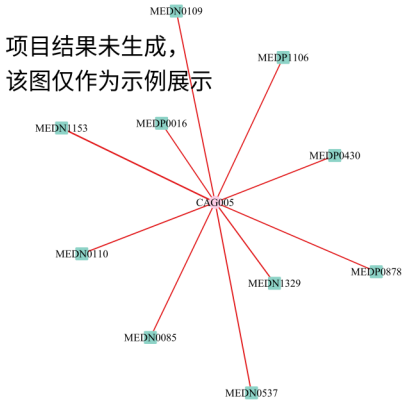
驱动物种与显著代谢物显著相关结果的 CCA 分析

红色标签表示驱动物种，蓝色标签和箭头表示代谢物，黑色标签表示样本。箭头长度表示代谢物对微生物变化影响的强度，箭头越长，表示代谢物对微生物变化的影响越大。样本点到代谢物线段及其延长线的垂直距离表示微生物对样本的影响强度，距离越近，表示代谢物对样本的影响越大。以中心原点为起点，微生物与箭头同方向，表示代谢物与微生物的变化正相关，反之表示负相关。

文件路径：4.Driver_species/4.top_driver_species_metabolome_correlation

4.4 驱动物种和显著代谢物的网络图分析

筛选出的显著相关结果绘制网络图，结果如下图所示。



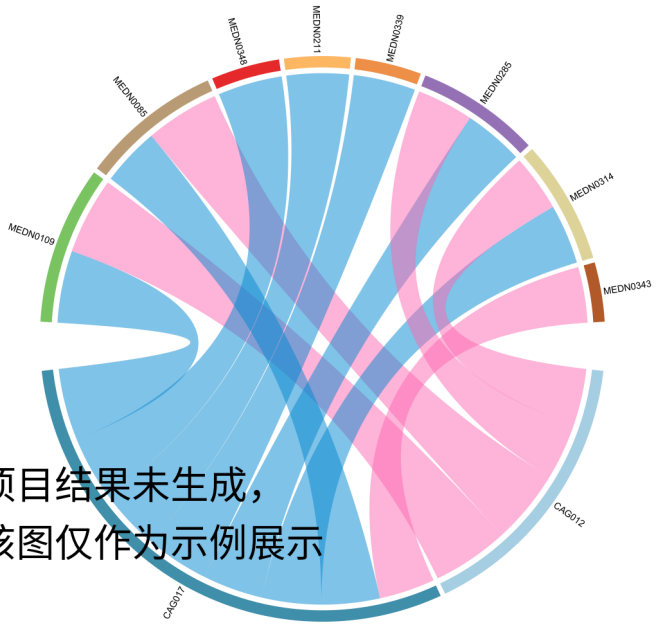
驱动物种与显著代谢物显著相关结果的网络图

微生物以粉红色表示，代谢物以浅绿色表示。微生物与代谢物之间的连接表示相关性，红色表示正相关，蓝色表示负相关，线条越粗表示相关性越大。动态网络图中节点大小表示度的大小，即连接的边越多，节点越大。

文件路径：4.Driver_species/4.top_driver_species_metabolome_correlation

4.5 驱动物种和显著代谢物的和弦图分析

筛选出的显著相关结果绘制和弦图分析，结果如下图所示。



驱动物种与显著代谢物显著相关结果的和弦图

弦连接（link）的宽度表示所连接的两个对象的相关性大小，link 越宽，相关性绝对值越大。粉红色表示正相关性，蓝色表示负相关性。

文件路径：4.Driver_species/4.top_driver_species_metabolome_correlation

参考文献

Kanehisa, Minoru, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. 2017. "KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs." *Nucleic Acids Research* 45 (D1): D353–D361. https://doi.org/10.1093/nar/gkw1092.

Kanehisa, Minoru, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2016. "KEGG as a Reference Resource for Gene and Protein Annotation." *Nucleic Acids Research* 44 (D1): D457–462. https://doi.org/10.1093/nar/gkv1070.

Langfelder, Peter, and Steve Horvath. 2008. "WGCNA: An R Package for Weighted Correlation Network Analysis." *BMC Bioinformatics* 9 (December): 559. https://doi.org/10.1186/1471-2105-9-559.

Nielsen, H Bjørn, Mathieu Almeida, Agnieszka Sierakowska Juncker, Simon Rasmussen, Junhua Li, Shinichi Sunagawa, Damian R Plichta, et al. 2014. "Identification and Assembly of Genomes and Genetic

Elements in Complex Metagenomic Samples Without Using Reference Genomes." *Nature Biotechnology* 32 (8): 822–28. https://doi.org/10.1038/nbt.2939.

Pedersen, Helle Krogh, Sofia K. Forslund, Valborg Gudmundsdottir, Anders Østergaard Petersen, Falk Hildebrand, Tuulia Hyötyläinen, Trine Nielsen, et al. 2018. "A Computational Framework to Integrate High-Throughput '-Omics' Datasets for the Identification of Potential Mechanistic Links." *Nature Protocols* 13 (12): 2781–2800. https://doi.org/10.1038/s41596-018-0064-z.