**第一章 课题前景**

本次项目数据全部来自WeBirth项目。WeBirth（Westlake Precision Birth Cohort）是一项基于妊娠糖尿病（GDM）孕妇及其子代的前瞻性出生队列。WeBirth旨在研究GDM患者的精准营养，结合标准化膳食测试和连续葡萄糖监测来反应参与者对膳食摄入的个性化血糖响应。参与者将在入组时（孕24 ~ 28周）接受14天的连续血糖监测和客观体力活动测量，并在此期间按照三个标准餐（早餐、早加餐、中餐）来控制饮食。该研究收集了有关人口统计、生活方式、饮食习惯、疾病史和心理健康等信息，以及血液、唾液、尿液、粪便、胎盘和母乳等生物样本；并将进行多组学检测（如粪便宏基因组、代谢组和蛋白质组）。WeBirth通过在GDM人群中整合精准营养、可穿戴设备和多组学数据等，为前瞻性出生队列研究提供了一个新的思路和框架。[1]

肠道菌群是人体内重要的微生物群落，它可以参与食物消化、营养吸收和代谢调节等过程。因此，通过探究肠道菌群对人体内环境的影响，可以为改善妊娠期孕妇的血脂水平提供新的思路。李兰娟院士在首届西湖生物大数据与健康创新论坛上曾做过一个关于肠道微生态概述、肠道微生态与宿主免疫代谢相关性的报告。在研究中，无菌大鼠的免疫力普遍低于非无菌大鼠，印证了肠道菌群在动物正常生命活动中的重要作用。因此，对肠道菌群在人体内环境稳态的调控作用的研究是非常有必要且富有前景的。

我们尝试进行多组学分析，来找出孕妇血脂与多维度影响因子间的关联性，并希望以此找到治疗妊娠期糖尿病的新靶点以及为相关患者提供更准确的营养学建议。

**第二章 相关准备工作**

2.1 R语言的学习

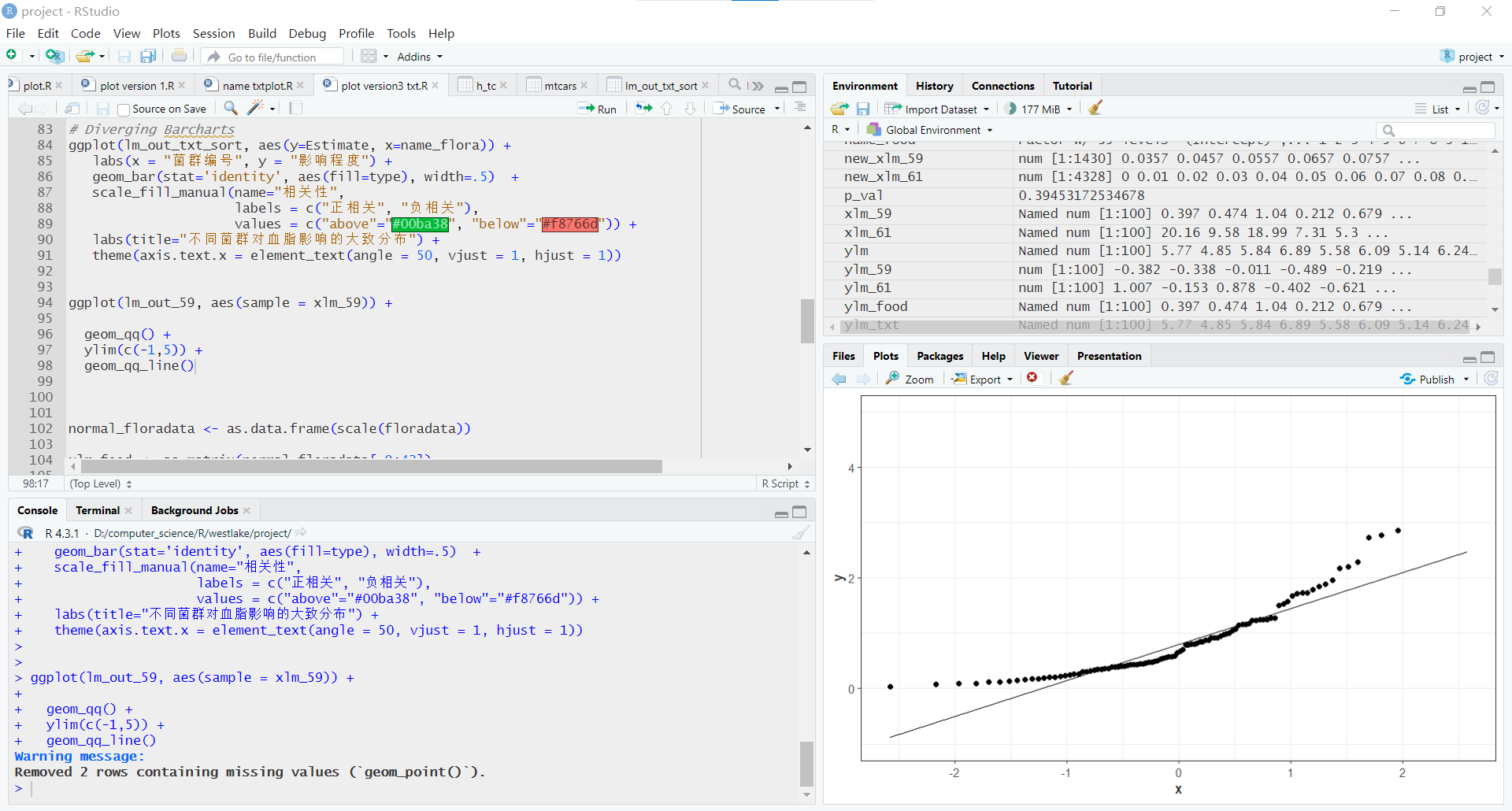
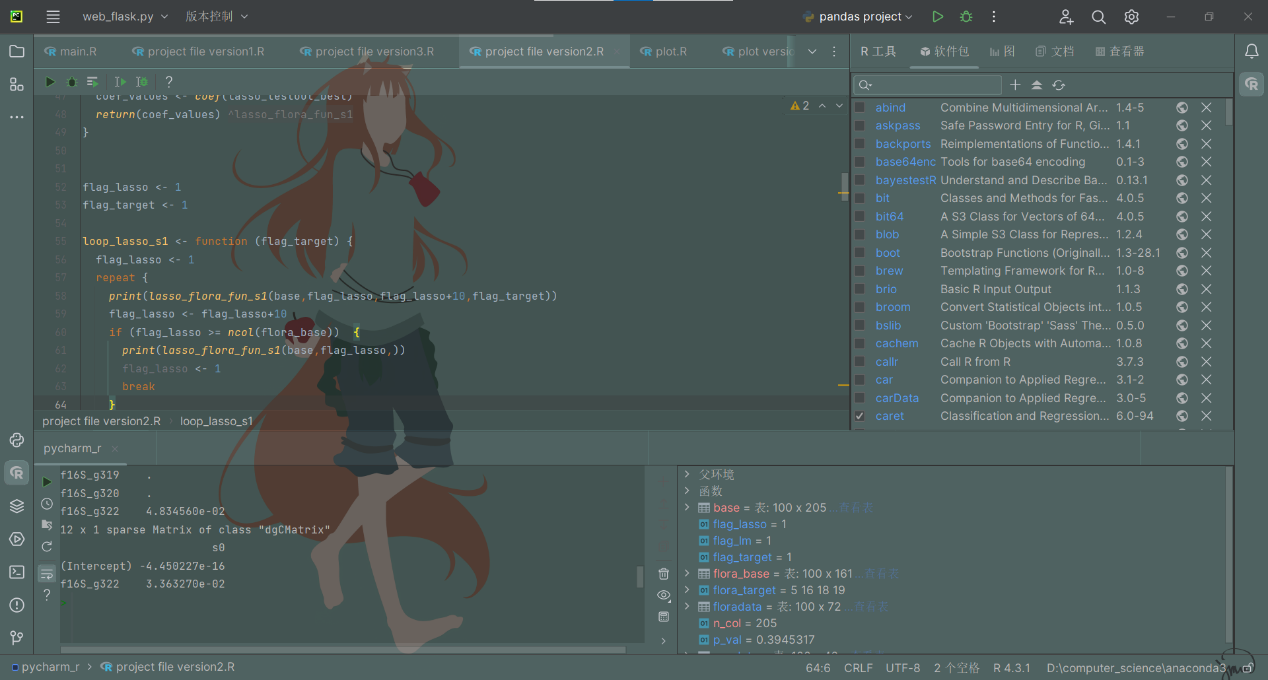
由于此次分析的工具主要为R语言，因此我在网上查找了相关资料，学习了R语言的基本操作。在北京大学教授李东风的个人主页（网址：https://www.math.pku.edu.cn/teachers/lidf），我找到了一份详细的R语言学习资料，并且主要依靠这份文件学习了大部分R语言操作，并掌握了readxl、ggplot2、cowplot等业内流行的R包的基础运用。

此外，通过在知乎、bilibili、CSDN、Runebook等网站上查询相关代码与语法，我解决了学习过程中的绝大部分问题。其余的疑难问题也在与指导教授郑矩圣教授与其助手苟望龙老师的两次线下交流与两次线上会议中得到了完美解决。

在开始工作前，除了与指导老师交流以外，我还与一名现居美国的统计行业从业者探讨了WeBirth项目，并且获得了一定的指导。

2.2 R语言环境的搭建

一个好的IDE可以大大降低编写程序时的工作量。我准备了两套平行的开发环境：R 4.3.1 + Rstudio；R 4.3.1 + Anaconda + PyCharm Community Edition。大部分R包从国内镜像站下载。相关的R语言程序也已上传至个人GitHub主页（https://github.com/Ives-Natsume/Gestational-diabetes.git）。由于PyCharm Community并不支持4.3.1版本的R作图，因此绘图工作主要在Rstudio中完成。PyCharm高效的代码检查、代码查询等功能大大提高了我编写程序的速度和正确率。



2.3 统计学的学习

在《数学 选择性必修 第三册（A版）》中，我学习了关于统计的基本知识，包括期望、方差、正态分布、最小二乘法、线性回归、χ2检验等基本的统计学知识。在处理数据时，发现相关知识的欠缺后，我在知乎等平台查找了相关资料，并向指导老师、指导教授、网友学习了相关的统计学知识。但是由于数学知识储备的欠缺，无法完全了解所有的数学原理。

**第三章 数据分析与结果呈现**

3.1建立血脂预测模型的尝试

起初，我希望基于现有的100组数据得到一个较为准确的血脂预测模型，因此尝试建立了一个基础的多变量线性回归模型，但是出现了以下意料之外的情况：

Call:

lm(formula = Y ~ X, data = data)

Residuals:

ALL 99 residuals are 0: no residual degrees of freedom!

Coefficients: (105 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.214e-14 NaN NaN NaN

X总胆固醇mmolL\_v1 1.000e+00 NaN NaN NaN

X甘油三酯mmolL\_v1 -1.292e-16 NaN NaN NaN

X高密度脂蛋白胆固醇mmolL\_v1 2.081e-15 NaN NaN NaN

X低密度脂蛋白胆固醇mmolL\_v1 1.265e-15 NaN NaN NaN

Xage\_v1 3.378e-17 NaN NaN NaN

Xpre\_BMI -4.615e-17 NaN NaN NaN

Xincome -1.415e-16 NaN NaN NaN

……

Xf16S\_g352 NA NA NA NA

Xf16S\_g356 NA NA NA NA

Xf16S\_g358 NA NA NA NA

Residual standard error: NaN on 0 degrees of freedom

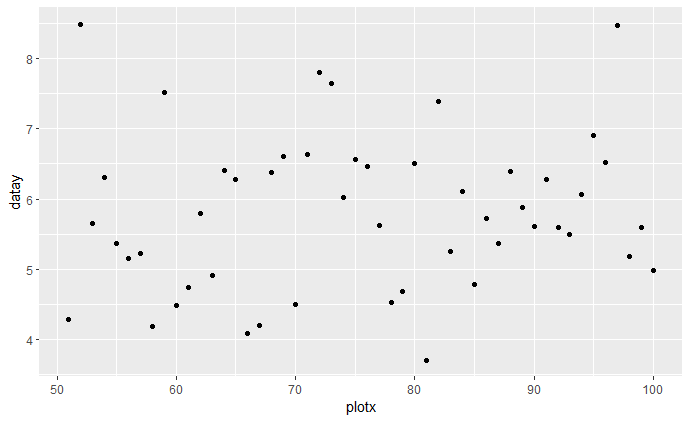
(因为不存在，1个观察量被删除了)

Multiple R-squared: 1, Adjusted R-squared: NaN

F-statistic: NaN on 98 and 0 DF, p-value: NA

反复检查程序无错误后，我们最终发现在建立模型时，我们忽视了最基本的数学原理：方程的组数必须大于等于自变量个数，从而导致了问题的发生。

为了解决相关问题，我起初选用了Ridge回归、N-Fold交叉验证等方法尝试建立相关模型，但是训练后的模型预测效果并不好。以下是一张模型预测值与实际值匹配情况的可视化图片：



可以看到，预测值与观测值的符合度很低。在与教授线下交流后，我放弃了建立预测模型的想法，转而关注不同影响因子与血脂间的联系。理由如下：

1. 数据量过少导致难以训练模型，即使训练成功也可能过拟合
2. 在我们的项目中，首要目标应为寻找血脂的影响因素，而不是建立模型
3. 调整模型的参数来提高预测能力是一项复杂的工作，已经超出了我们的能力范围

3.2 影响因子的相关性判断

Lasso回归时，对数据进行了L1正则化处理。L1正则化在逐渐加强的过程中，相对不重要的特征的参数会比相对重要的特征的参数更快地变成0。所以L1正则化可以用于特征选择。通过选出少量但重要的特征，以防止过拟合问题。这一思想特别适合我们的数据集，因此被我们采用，并借此完成了对肠道菌群的初步筛选。

以下是初筛相关R语言程序（完整程序可以在我的GitHub下载）：

lasso\_flora\_fun\_s1 <- function(dataset, col\_1, col\_2, target) {

require(ggplot2)

require(caret)

require(glmnet)

require(corrplot)

require(Metrics)

require(cowplot)

# 数据标准化

normal\_dataset <- scale(dataset)

normal\_flora\_dataset <- normal\_dataset[,44:ncol(dataset)]

# lasso回归

xlasso <- normal\_flora\_dataset[, col\_1:col\_2]

ylasso <- normal\_dataset[, target+1]

set.seed(1245)

lasso\_testout <- cv.glmnet(xlasso, ylasso, alpha = 1, nfolds = 10)

# 绘制交叉验证结果图

if (requireNamespace("ggplot2", quietly = TRUE)) {

plot(lasso\_testout)

plot(lasso\_testout$glmnet.fit, "lambda", label = TRUE)

}

lasso\_testout\_min <- lasso\_testout$lambda.min

lasso\_testout\_best <- glmnet(xlasso, ylasso, alpha = 1, lambda = lasso\_testout\_min)

# 返回系数的名称和值

#coef\_names <- rownames(coef(lasso\_testout\_best))

coef\_values <- coef(lasso\_testout\_best)

return(coef\_values)

}

flag\_lasso <- 1

flag\_target <- 1

loop\_lasso\_s1 <- function (flag\_target) {

flag\_lasso <- 1

repeat {

print(lasso\_flora\_fun\_s1(base,flag\_lasso,flag\_lasso+10,flag\_target))

flag\_lasso <- flag\_lasso+10

if (flag\_lasso >= ncol(flora\_base)) {

print(lasso\_flora\_fun\_s1(base,flag\_lasso,))

flag\_lasso <- 1

break

}

}

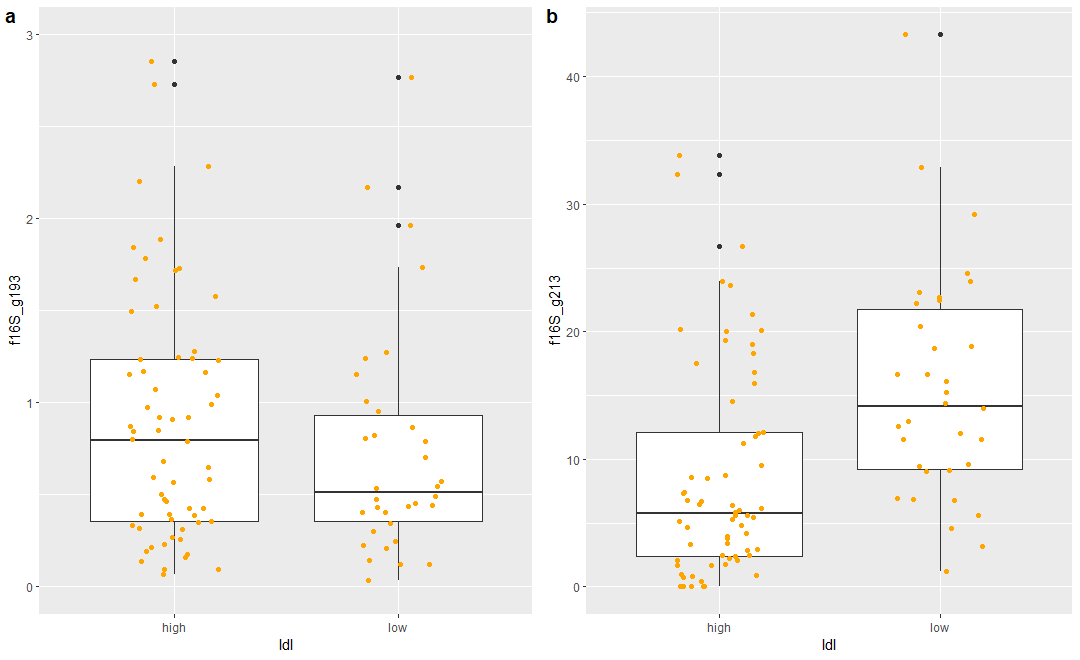
} #该函数可对所有血脂筛选菌群

在初筛的基础上，我们以剩余的肠道菌群为自变量，并且纳入了饮食、体力活动等因素，共同建立了有关血脂的线性回归模型。如果希望获得更好的拟合效果，可以通过观察Q-Q图等参数进一步调整参数，但是由于时间等因素，我们目前未能进一步探究。最后我们得出了两类对血脂影响最显著的菌群：

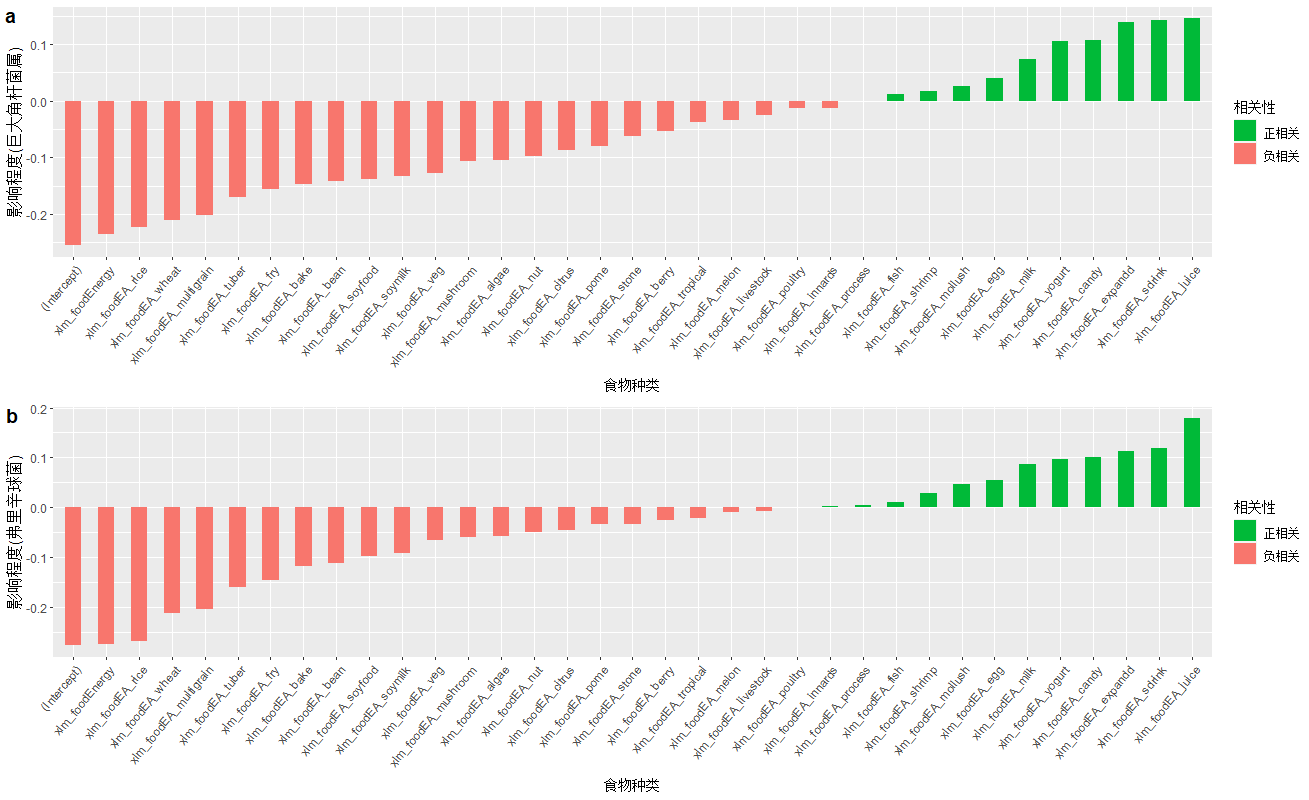
d\_\_Bacteria;p\_\_Actinobacteriota;c\_\_Coriobacteriia;o\_\_Coriobacteriales;f\_\_Coriobacteriaceae;g\_\_Enorma（中文名：巨大角杆菌属，P-value = 0.03563885，与TC、LDL-C水平呈正相关）；

d\_\_Bacteria;p\_\_Firmicutes;c\_\_Clostridia;o\_\_Lachnospirales;f\_\_Lachnospiraceae;g\_\_Frisingicoccus（中文名：弗里辛球菌，P-value = 0.003784949，与TC、LDL-C水平呈负相关）。

为了验证分析结果的正确性，我们根据患者的LDL-C水平是否高于危险阈值将数据集进行了分组。结果我们发现分组结果中的相关菌群丰度确实与分析结果相吻合，证明了以上两类菌群确实对血脂水平有较强的相关性。



同时，我们也进一步分析了菌群丰度受膳食因素影响的情况，并希望以此为妊娠期糖尿病患者制定更合理的营养学建议。结果如图：



**第四章 总结与展望**

通过以上工作，我们成功发现了对血脂水平有调控作用的菌群。未来我们可以以此作为妊娠期糖尿病的治疗靶点来开发新药。此外，我们的工作也可以为妊娠期糖尿病患者提供更好的营养学建议。

原计划里除了以上内容外，我们也希望通过贝叶斯公式、神经网络等方式进一步探究菌群与菌群，食物与菌群等因素间的相互作用。但是限于时间和精力有限，目前的工作只能暂时止步于此。以后我们将继续工作，继续挖掘数据中的真相。

**参考文献：**

[1]Liang, Xinxiu, Miao, Zelei, Lu, Sha, Ye, Meng, Wang, Jiali, Zhao, Hui, et al. 2023. Integration of multiomics with precision nutrition for gestational diabetes: Study protocol for the Westlake Precision Birth Cohort. *iMeta* 2, e96. <https://doi.org/10.1002/imt2.96>