主成分分析PCA

刘华，中国科学技术大学

2020/6/18

### 主成分分析 PCA

本节作者：刘华，中国科学技术大学

版本1.0.5，更新日期：2020年6月19日

#### 什么是PCA(Principal Component Analysis)

##### 相关背景

在许多领域的研究与应用中，通常需要对含有多个变量的数据进行观测，收集大量数据后进行分析寻找规律。多变量大数据集无疑会为研究和应用提供丰富的信息，但是也在一定程度上增加了数据采集的工作量。更重要的是在很多情形下，许多变量之间可能存在相关性，从而增加了问题分析的复杂性。如果分别对每个指标进行分析，分析往往是孤立的，不能完全利用数据中的信息，因此盲目减少指标会损失很多有用的信息，从而产生错误的结论。

因此需要找到一种合理的方法，在减少需要分析的指标同时，尽量减少原指标包含信息的损失，以达到对所收集数据进行全面分析的目的。由于各变量之间存在一定的相关关系，因此可以考虑将关系紧密的变量变成尽可能少的新变量，使这些新变量是两两不相关的，那么就可以用较少的综合指标分别代表存在于各个变量中的各类信息。主成分分析与因子分析就属于这类降维算法。

##### 数据降维

降维就是一种对高维度特征数据预处理方法。降维是将高维度的数据保留下最重要的一些特征，去除噪声和不重要的特征，从而实现提升数据处理速度的目的。在实际的生产和应用中，降维在一定的信息损失范围内，可以为我们节省大量的时间和成本。降维也成为应用非常广泛的数据预处理方法。

降维具有如下一些优点：

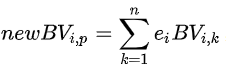
（1） 使得数据集更易使用。 （2） 降低算法的计算开销。 （3） 去除噪声。 （4） 使得结果容易理解。

降维的算法有很多，比如奇异值分解(SVD)、主成分分析(PCA)、因子分析(FA)、独立成分分析(ICA)。

##### PCA的概念

PCA(Principal Component Analysis)，即主成分分析方法，是一种使用最广泛的数据降维算法。PCA的主要思想是将n维特征映射到k维上，这k维是全新的正交特征也被称为主成分，是在原有n维特征的基础上重新构造出来的k维特征。PCA的工作就是从原始的空间中顺序地找一组相互正交的坐标轴，新的坐标轴的选择与数据本身是密切相关的。其中，第一个新坐标轴选择是原始数据中方差最大的方向，第二个新坐标轴选取是与第一个坐标轴正交的平面中使得方差最大的，第三个轴是与第1,2个轴正交的平面中方差最大的。依次类推，可以得到n个这样的坐标轴。通过这种方式获得的新的坐标轴，我们发现，大部分方差都包含在前面k个坐标轴中，后面的坐标轴所含的方差几乎为0。于是，我们可以忽略余下的坐标轴，只保留前面k个含有绝大部分方差的坐标轴。事实上，这相当于只保留包含绝大部分方差的维度特征，而忽略包含方差几乎为0的特征维度，实现对数据特征的降维处理。

变换的步骤

① 第一步计算矩阵 X 的样本的协方差矩阵 S（此为不标准PCA，标准PCA计算相关系数矩阵C） : ② 第二步计算协方差矩阵S（或C）的特征向量 e1,e2,…,eN和特征值t = 1,2,…,N； ③ 第三步投影数据到特征向量张成的空间之中。利用公式 ，其中BV值是原样本中对应维度的值。

PCA 的目标是寻找 r （ r<n ）个新变量，使它们反映事物的主要特征，压缩原有数据矩阵的规模，将特征向量的维数降低，挑选出最少的维数来概括最重要特征。每个新变量是原有变量的线性组合，体现原有变量的综合效果，具有一定的实际含义。这 r 个新变量称为“主成分”，它们可以在很大程度上反映原来 n 个变量的影响，并且这些新变量是互不相关的，也是正交的。通过主成分分析，压缩数据空间，将多元数据的特征在低维空间里直观地表示出来。

#### PCA实例

##### PCA+特征箭头

华大基因研究团队发表于Microbiome的文章(Zhong et al., 2019)，研究早期事件和生活方式对幼龄儿童肠道菌群和代谢的影响。在这里，我们选择了文章中两个PCA分析结果图为例进行讲解：

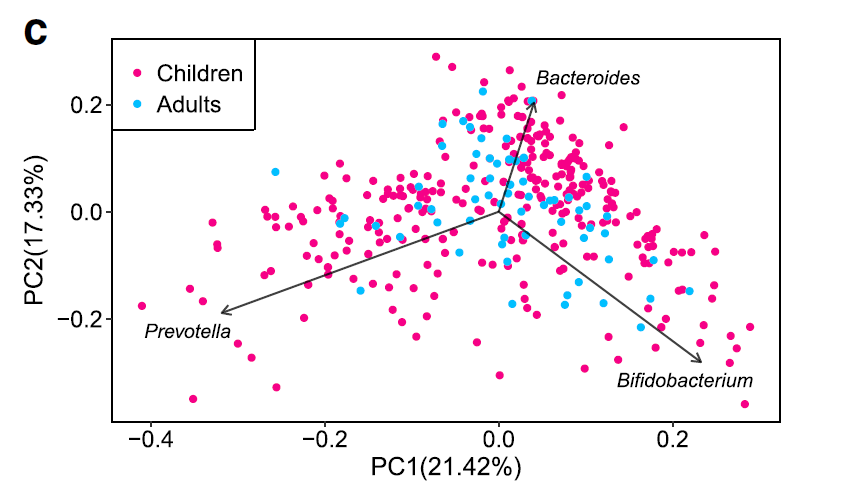


图1c. 属水平PCA主成分分析

Figure 1c. Genus-based principal component analysis (PCA) of children and adults.

结果：PCA分析揭示儿童和成年人肠道菌群没有明显区别

Principal component analysis (PCA) based on genus profiles showed no separation between Dutch children and adults.

##### PCA+特征箭头+箱线图

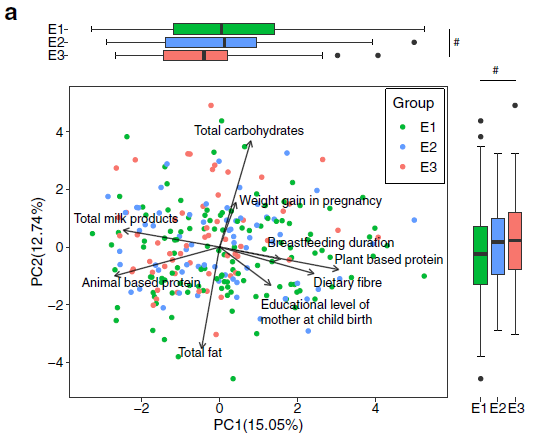


图3. 多种早期事件和学龄前生活方式与肠道菌群有关。a：PCA显示了儿童的多变量以及不同因素对PC1和PC2的主要影响。将包括早期事件和学龄前生活方式在内的18个因素进行PCA分析，其中PC1或PC2成分得分 ≥ 0.2的因素为主要影响因素。箱形图显示各肠型内PC1和PC2评分的总体分布(#P<0.05；Wilcoxon秩和检验)。

Multiple early events and pre-school lifestyle associated with the school-age gut microbiota. a PCA showing the multivariate variation of children and the major contributions of different factors to PC1 and PC2. A total of 18 factors including early events and pre-school lifestyle were subjected to PCA, and those factors with component scores for PC1 or PC2 ≥ 0.2 were shown as major contributors. Box plots showing the overall distribution of PC1 and PC2 scores within each enterotype (#P<0.05; Wilcoxon rank-sum test).

结果：我们发现母乳喂养持续时间、母亲教育水平和学前饮食模式，包括摄入的蛋白质、纤维、及奶制品在PC1中贡献最多 (15.05%)，PC2中儿童的碳水化合物和脂肪总摄入量是第二个最重要的变量 (12.74%)。E3组儿童的PC1得分低于E1组，但PC2得分高于E1组(图3a, Wilcoxon秩和检验，P < 0.05)。这种肠型间的差异是由PC1评分的主要因素决定的，如 E3组中较短的母乳喂养时间以及较少的膳食纤维和植物性蛋白摄入量 (Kruskal-Wallis检验, P < 0.05)。

We found that breastfeeding duration, educational level of mother at childbirth, and pre-school dietary patterns including intake of protein, fiber, and milk products contributed most to the variability in PC1 (15.05%, Fig. 3a), and total intake of carbohydrates and fat represented the second most important variation among children, as displayed in PC2 (12.74%, Fig. 3a). Interestingly, children in E3 exhibited lower PC1 scores but higher PC2 scores than children in E1 (Fig. 3a, Wilcoxon rank-sum test, P < 0.05). This inter-enterotype difference was governed by specific major contributors of the PC1 scores, including shorter breastfeeding duration and less intake of dietary fiber and plant-based protein in E3 as compared to the two other enterotypes (Kruskal-Wallis test, P < 0.05).

##### PCA+形状分组+门着色

例3：Robert D. Finn团队发表于Nature的文章(Almeida et al., 2019)，构建了人类肠道微生物群的基因组蓝图。

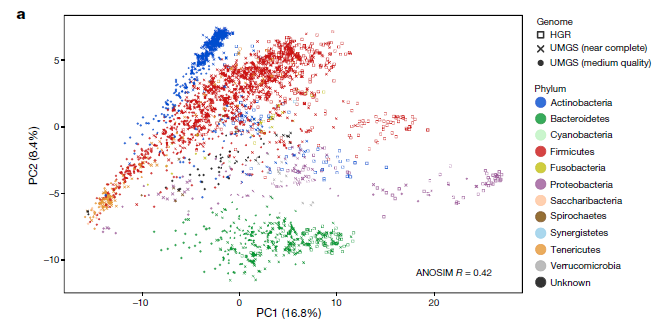


Fig 5：未培养的物种具有独特的功能。a、以已知人类参考基因组（HGR：553个基因组）和未分类物种宏基因组（UMGS：1952个基因组）的基因组特性(Genome Properties, GPs)进行PCA主成分分析，以门水平上色。

The uncultured species have a distinct functional capacity.a, Principal component analysis (PCA) based on GPs of the HGR (n = 553 genomes) and the UMGS (n = 1,952 genomes) coloured by phylum.

结果：我们使用GhostKOALA生成KEGG Orthology (KO)注释，以跟踪不同UMGS和HGR集合中特定功能类别的差异丰度。在全球范围内，通过对GPs的分类组成分析，发现按门水平分类分离良好，特别是拟杆菌门和变形菌门显示出独特的功能特征。

In parallel, we used GhostKOALA to generate KEGG Orthology (KO) annotations to track the differential abundance of specific functional categories across the UMGS and HGR sets. Globally, by analyzing the repertoire of GPs according to the taxonomic composition, we observed a good separation by phylum (ANOSIM R = 0.42, P < 0.001), with the Bacteroidetes and Proteobacteria taxa in particular displaying very distinctive functional profiles (Fig. 5a).

##### 总结

PCA主成分图中坐标轴PC1/2的数值为总体差异的解释率；图中点代表样品，颜色代表分组；箭头代表原始变量，其中方向代表原始变量与主成分的相关性，长度代表原始数据对主成分的贡献度。

做PCA，首先要构建特征/变量的协方差矩阵，然后对其特征值和特征向量进行排序，根据需要取前面最重要的部分，将后面的维数省去，可以达到降维，从而达到简化模型或对数据进行压缩的效果，同时最大程度的保持了原有数据的信息。

但是PCA原理主要是为了消除变量之间的相关性，并且假设这种相关性是线性的，对于非线性的依赖关系则不能得到很好的结果。同时PCA假设变量服从高斯分布，当变量不服从高斯分布（如均匀分布）时，会发生尺度缩放与旋转。

#### PCA绘图实战

数据和代码下载：以 <https://github.com/YongxinLiu/MicrobiomeStatPlot> 上的微生物组数据进行展示。

用于计算PCA 的R软件中提供了来自不同软件包的多个函数：

prcomp()和princomp() [内置];

PCA() [ FactoMineR包];

dudi.pca() [ ade4包];

epPCA() [ ExPosition包]。

可以通过factoextraR包和ggbiplot包来轻松提取和可视化PCA的结果。

##### 安装PCA分析与可视化R包

判断每个依赖的包是否存在，没有则安装

# 安装CRAN来源R包，多个包使用循环检测和安装  
p\_list = c("FactoMineR", "dplyr", "factoextra", "ggsignif", "pca3d", "cowplot")  
for(p in p\_list){  
 if (!requireNamespace(p, quietly = TRUE))  
 install.packages(p)  
}  
  
# 安装github来源R包  
suppressWarnings(suppressMessages(library(devtools)))  
if (!requireNamespace("ggbiplot", quietly = TRUE))  
 install\_github("vqv/ggbiplot")

##### 内置数据演示PCA绘制

library("dplyr")  
# 查看鸢尾花的内置数据  
head(iris, n=3)

Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
1 5.1 3.5 1.4 0.2 setosa  
2 4.9 3.0 1.4 0.2 setosa  
3 4.7 3.2 1.3 0.2 setosa

# 获得纯数值表格，去除最后一行的分类型分组数据，  
iris <- select(iris, -Species)  
# prcomp函数进行PCA分析，需要解释参数的意义？？？cor=T???  
iris.pca <- prcomp(iris,cor = T)  
# 查看对象的名称，此处返回结果中5个列表的名称  
names(iris.pca)

[1] "sdev" "rotation" "center" "scale" "x"

# 对象摘要，需要解释参数的意义？？？loadings = T???  
summary(iris.pca,loadings = T)

Importance of components:  
 PC1 PC2 PC3 PC4  
Standard deviation 2.0563 0.49262 0.2797 0.15439  
Proportion of Variance 0.9246 0.05307 0.0171 0.00521  
Cumulative Proportion 0.9246 0.97769 0.9948 1.00000

sdev是标准偏差; center中心化，将数据减去均值; scores即降维之后的结果，当然也可以使用predict函数，结果一样。

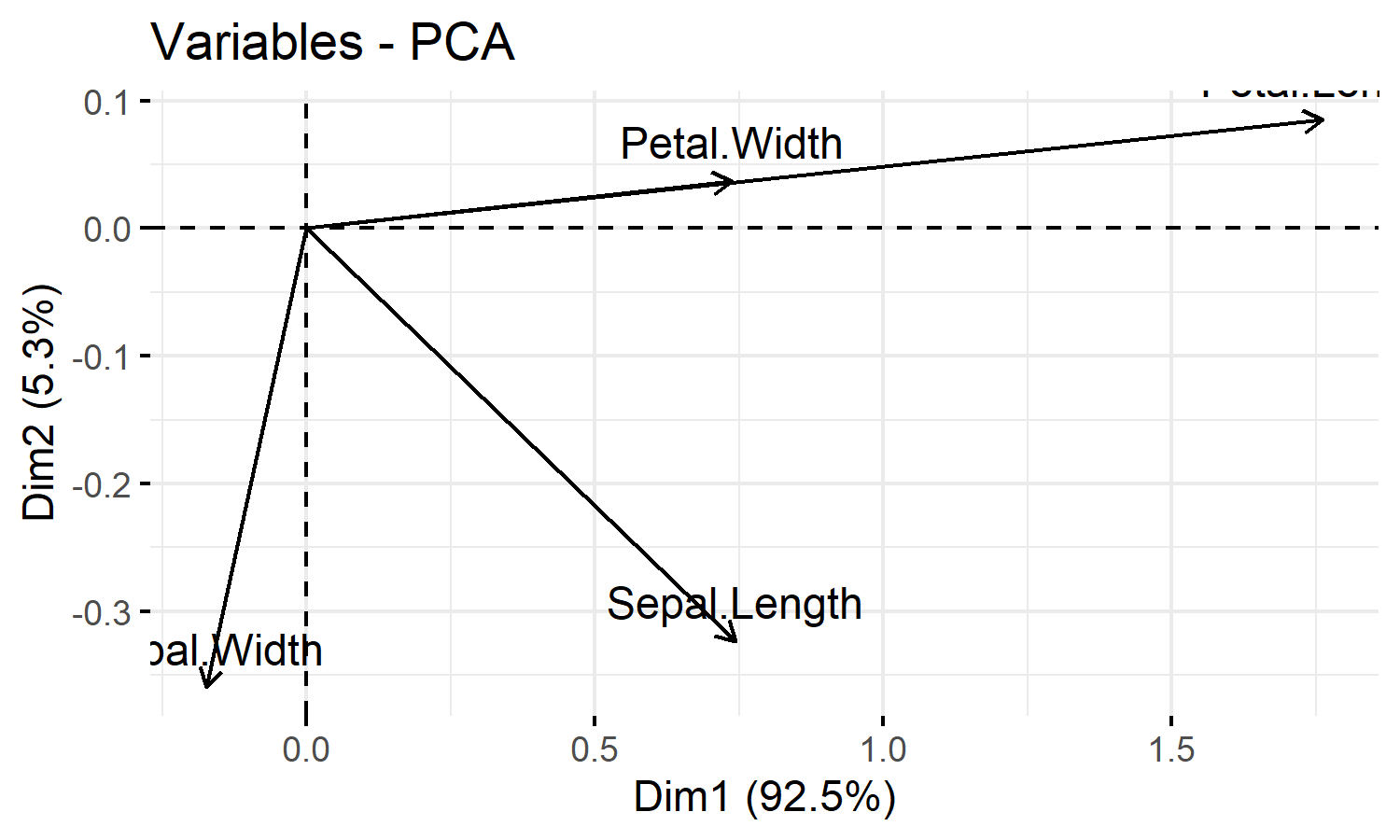
我们看Proportion of Variance，即为每一个成分方差所占比例，Cumulative Proportion代表是累计比例，为Proportion of Variance的累计值，一般达到90%左右就可以代表所有数据。

library("factoextra")  
# 提取变量的分析结果，加载包出现在函数第一次出现前，知道函数与包的关系  
get\_pca\_var(iris.pca)

Principal Component Analysis Results for variables  
 ===================================================  
 Name Description   
1 "$coord" "Coordinates for the variables"   
2 "$cor" "Correlations between variables and dimensions"  
3 "$cos2" "Cos2 for the variables"   
4 "$contrib" "contributions of the variables"

factoextra包自带了提取变量的分析结果get\_pca\_var函数，其中： coord表示用于创建散点图的变量坐标。coord实际上就是成分载荷，指观测变量与主成分的相关系数； cor表示相关系数； cos2表示因子质量，var.cos2 = var.coord \* var.coord； contrib表示包含变量对主成分的贡献(百分比)。

# 对变量作图, col.var设定线条颜色  
(p <- fviz\_pca\_var(iris.pca, col.var = "black"))



# 保存图片，指定图片为pdf格式方便后期修改，图片宽89毫米，高56毫米  
ggsave(paste0("p1.iris\_pca.pdf"), p, width=120, height=100, units="mm")  
ggsave(paste0("p1.iris\_pca.png"), p, width=120, height=100, units="mm")

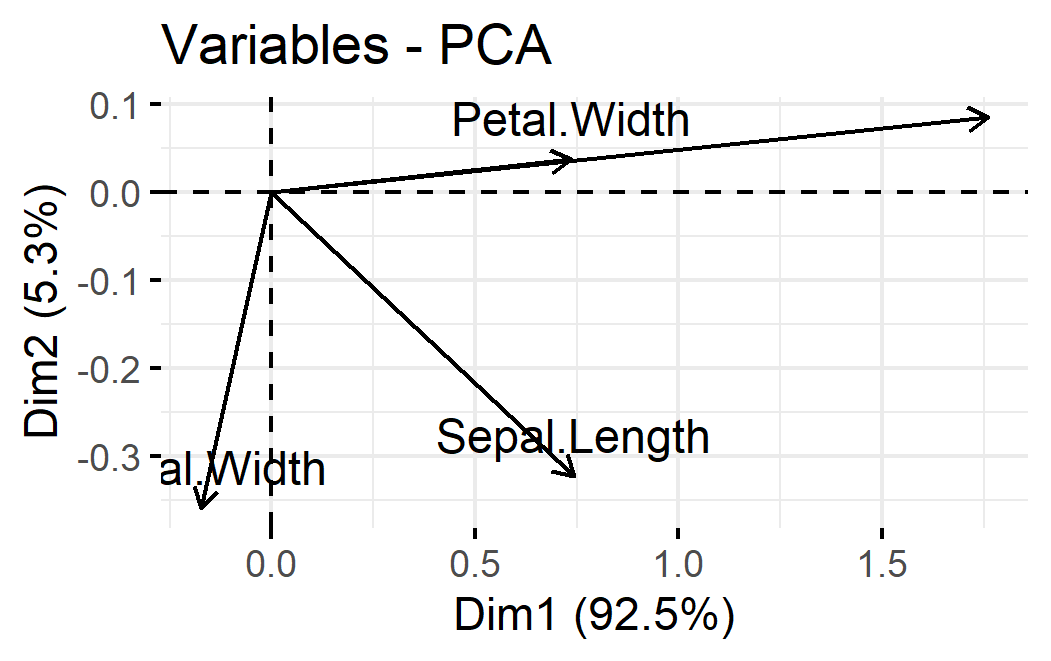


图1. PCA展示变量与主成分之间的关系，以及变量之间的关联

##### 菌群数据实战

本次测试数据来自同样来自[Science：拟南芥三萜化合物特异调控根系微生物组](https://mp.weixin.qq.com/s/KmMDEmptBKz8Fv7VSdz2Jg) 数据截取了3个实验组各6个样品的结果用于演示。数据位于Data/Science2019目录，本次需要元数据(metadata.txt)和otu表(otutab.txt)两个输入文件。

# 测试数据地址，可修改为本地下载github的目录，也可加载amplicon包获得内置数据  
dir="http://210.75.224.110/github/MicrobiomeStatPlot/Data/Science2019/"  
# 读取元数据，参数指定包括标题行(TRUE)，列名为1列，制表符分隔，无注释行，不转换为因子类型  
metadata <- read.table(paste0(dir, "metadata.txt"), header=T, row.names=1, sep="\t", comment.char="", stringsAsFactors = F)  
# 预览元数据前3行，注意分组列名  
head(metadata, n = 3)

Group Date Site CRA CRR BarcodeSequence  
KO1 KO 2017/6/30 Chaoyang CRA002352 CRR117575 ACGCTCGACA  
KO2 KO 2017/6/30 Chaoyang CRA002352 CRR117576 ATCAGACACG  
KO3 KO 2017/7/2 Changping CRA002352 CRR117577 ATATCGCGAG  
 LinkerPrimerSequence ReversePrimer Batch Species  
KO1 AACMGGATTAGATACCCKG ACGTCATCCCCACCTTCC 1 Arabidopsis thaliana  
KO2 AACMGGATTAGATACCCKG ACGTCATCCCCACCTTCC 1 Arabidopsis thaliana  
KO3 AACMGGATTAGATACCCKG ACGTCATCCCCACCTTCC 1 Arabidopsis thaliana  
 Sequencing Platform Description  
KO1 Novagene HiSeq2500 Knock-out replicate 1  
KO2 Novagene HiSeq2500 Knock-out replicate 2  
KO3 Novagene HiSeq2500 Knock-out replicate 3

# 读取otu表  
otutab <- read.table(paste0(dir, "otutab.txt"), row.names= 1, header=T, sep="\t", comment.char="", stringsAsFactors = F)  
# 过滤数据并排序  
sub\_metadata <- metadata[rownames(metadata) %in% colnames(otutab),]  
count <- otutab[, rownames(sub\_metadata)]  
# 基于OTU表PCA分析  
# 如果变量之间的数据的处于不同数量级或者变量之间的均值/方差相差很大时，建议是进行标准化，常见为：scale(count, center =T, scale. =T)对数据进行标准化，FactoMineR包的PCA()函数和基础包的prcomp()函数自带函数自带标准化参数，如下：  
# 基于OTU计算PCA，执行数据标准化  
otu.pca <- prcomp(t(count), scale. = TRUE)

##### 贡献率图(scree plot)

如果我们想判断PCA中需要多少个主成分比较好，那么可以从主成分的特征值来考虑（Kaiser-Harris准则建议保留特征值大于1的主成分）；特征值表示主成分所保留的变异量（所解释的方差）；如用get\_eigenvalue函数来提取特征值，结果中第一列是特征值，第二列是可解释变异的比例，第三列是累计可解释变异的比例。

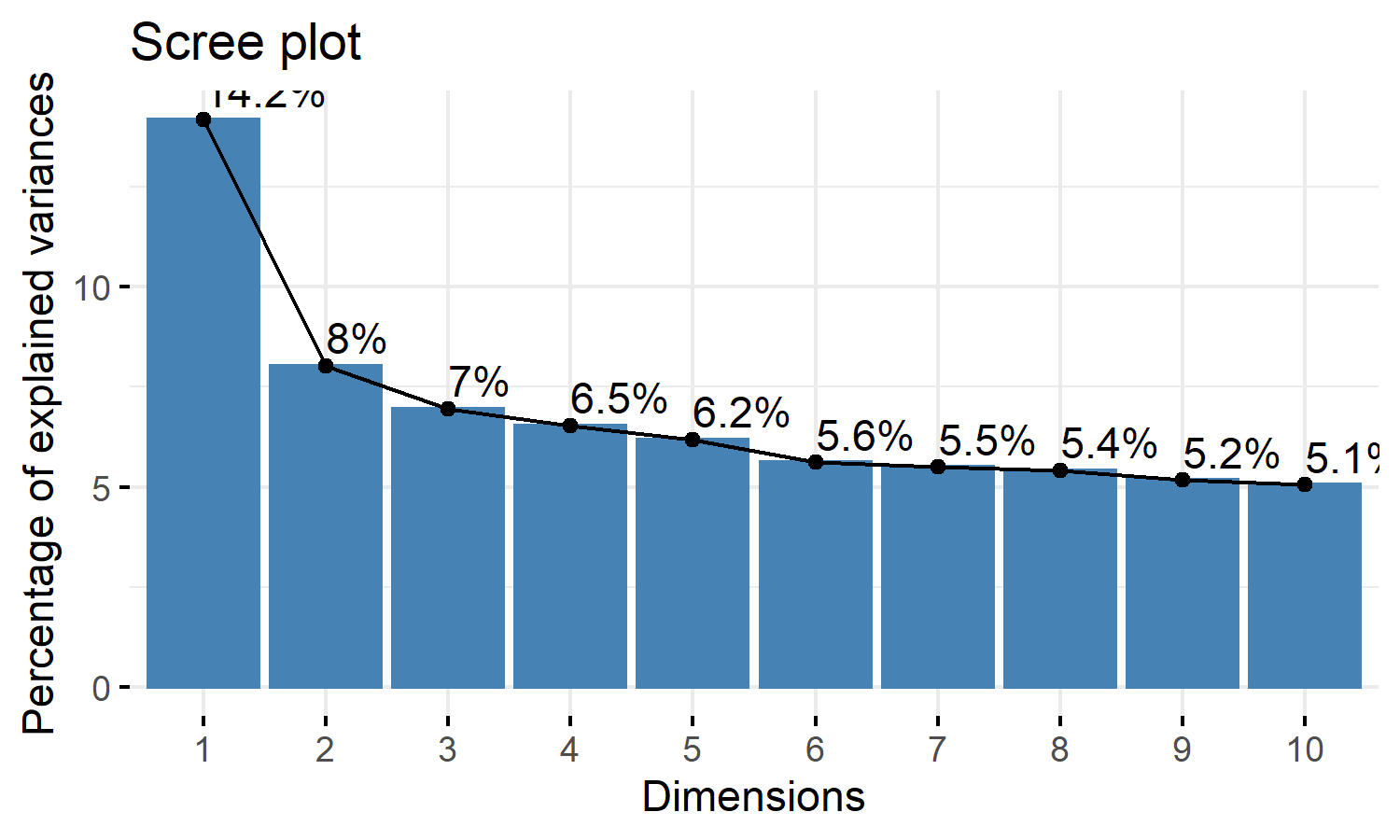
get\_eigenvalue(otu.pca)[1:3,]

eigenvalue variance.percent cumulative.variance.percent  
Dim.1 373.2903 14.182764 14.18276  
Dim.2 211.3096 8.028481 22.21124  
Dim.3 183.0639 6.955314 29.16656

除了特征值大于1作为主成分个数的阈值外，还可以设置总变异的阈值（累计）作为判断指标

除了看表格来判断，还可从图形上直观的感受下：

# 绘制崖低碎石图(scree plot)即贡献率图，外层()可对保存的图形同时预览  
(p <- fviz\_eig(otu.pca, addlabels = TRUE))

ggsave(paste0("p2.pca\_screen.pdf"), p, width=120, height=100, units="mm")  
ggsave(paste0("p2.pca\_screen.png"), p, width=120, height=100, units="mm")

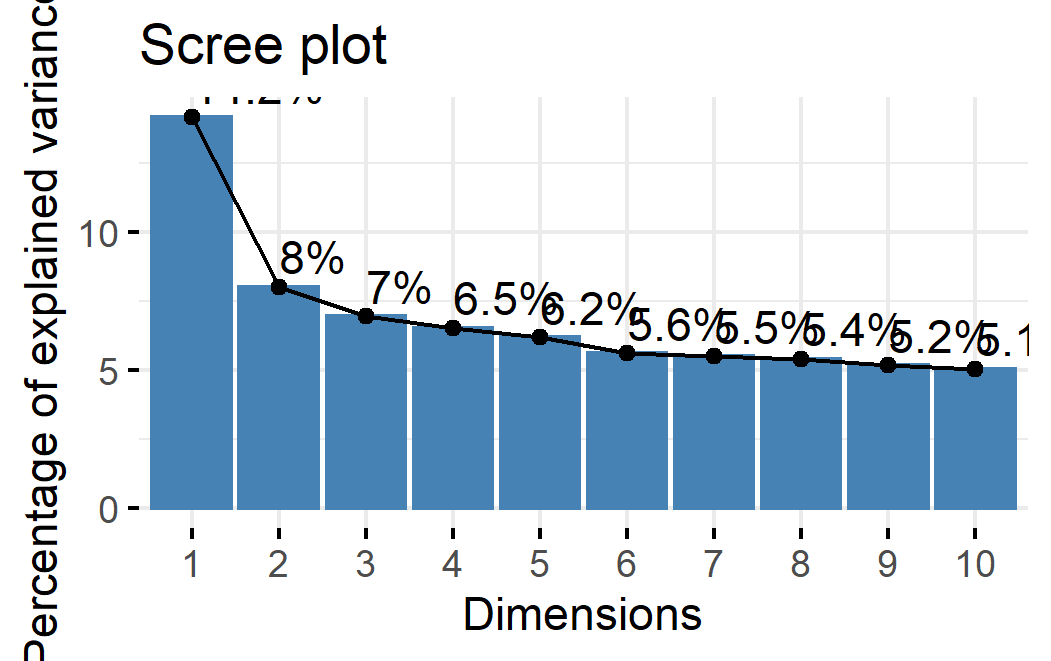


图2. 贡献率图表示各个主成分贡献率，进而决定选择多少主成分。可以将主成分数量限制为占总方差的比例。

如果我们想提取PCA结果中变量的信息，则可用get\_pca\_var()：

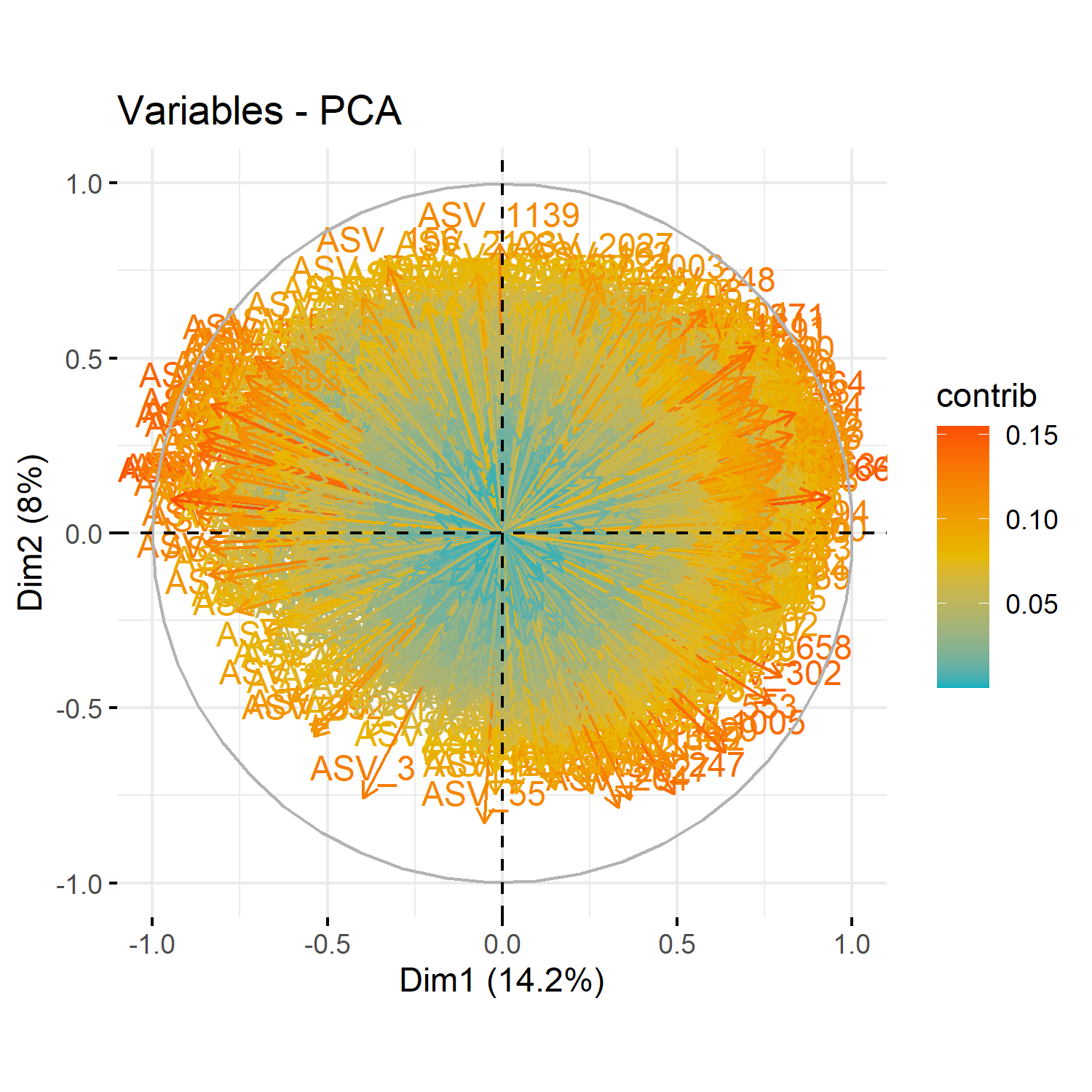
(var <- get\_pca\_var(otu.pca))

Principal Component Analysis Results for variables  
 ===================================================  
 Name Description   
1 "$coord" "Coordinates for the variables"   
2 "$cor" "Correlations between variables and dimensions"  
3 "$cos2" "Cos2 for the variables"   
4 "$contrib" "contributions of the variables"

Quality of representation（对应var$cos2），用于展示每个变量在各个主成分中的代表性（高cos2值说明该变量在主成分中有good representation，对应在Correlation circle图上则是接近圆周边上；低cos2值说明该变量不能很好的代表该主成分，对应Correlation circle图的圆心位置）。简单的说，如果一个变量在PC1和PC2的Contributions（cos2值在各个主成分中的比例）很高的话，则说明该变量可有效解释数据的变异，我们可以用图形展示各个变量在PC1和PC2上的Contributions。

##### 特征PCA图

# 绘制变量PCA主成分分析图  
(p <- fviz\_pca\_var(otu.pca, col.var = "contrib", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07")))



ggsave(paste0("p3.pca\_cos.pdf"), p, width=120, height=100, units="mm")  
ggsave(paste0("p3.pca\_cos.png"), p, width=120, height=100, units="mm")

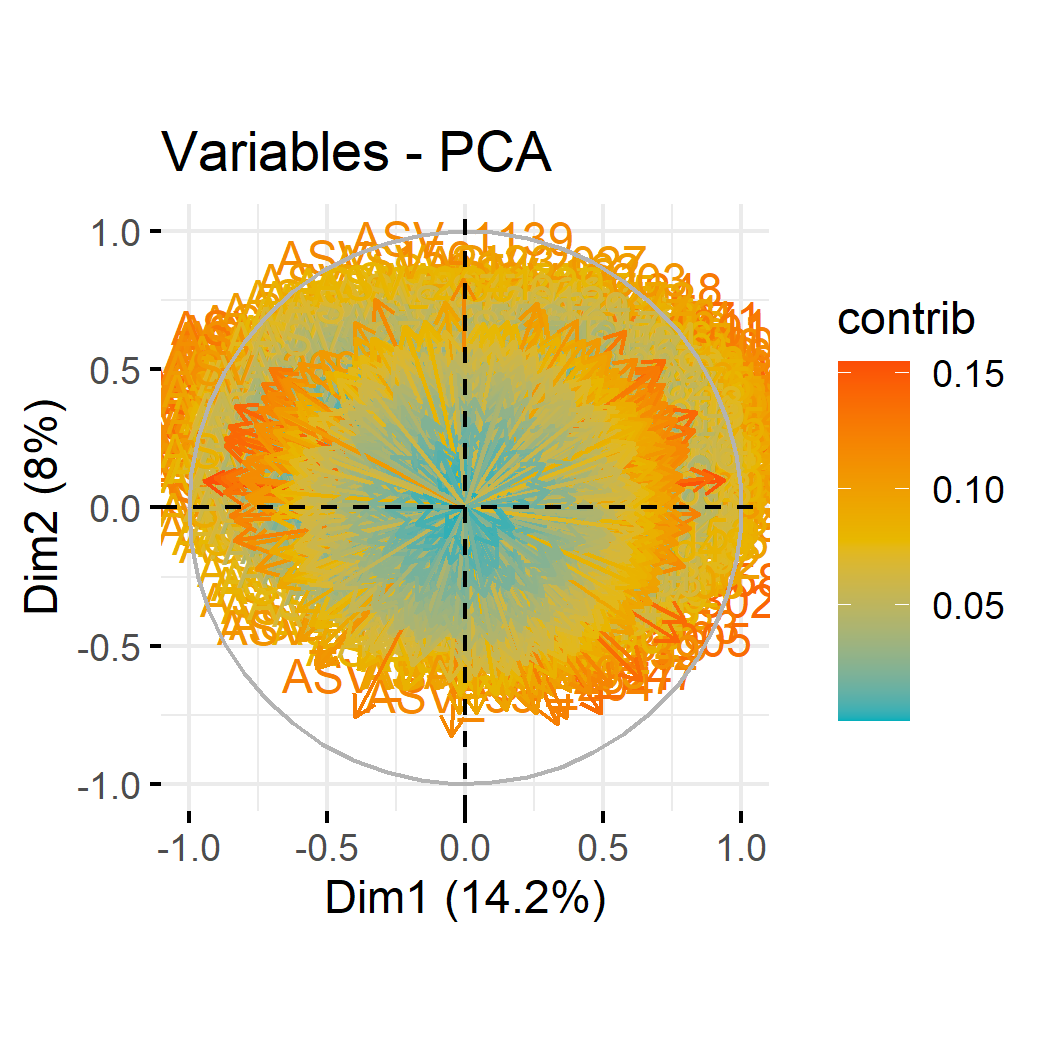


图3. PCA图展示变量

可以看到，biplot图只能用于展示变量较少的情况，当变量较多时需要进行筛选。

接下来分析观测值，先提取出individuals信息。

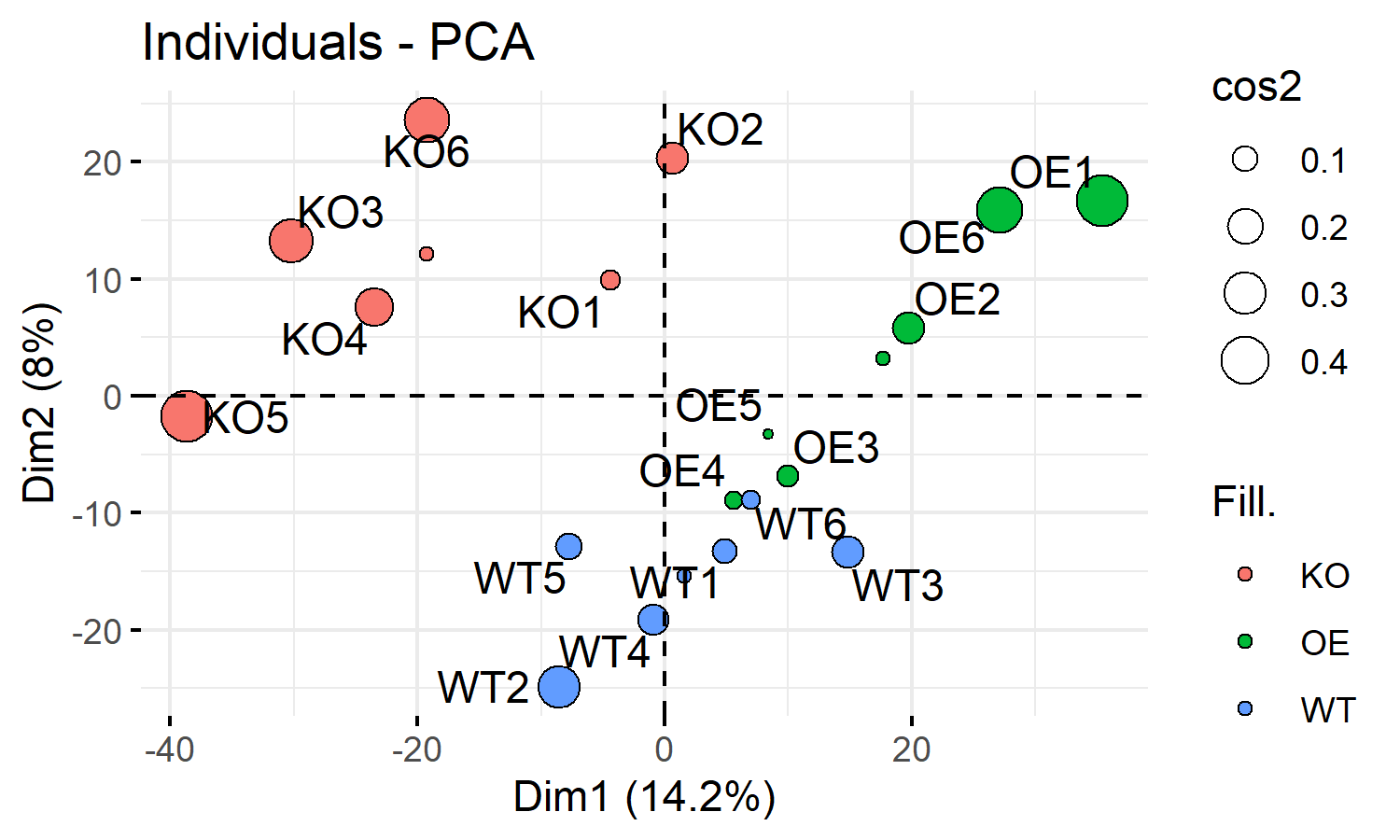
(ind <- get\_pca\_ind(otu.pca))

Principal Component Analysis Results for individuals  
 ===================================================  
 Name Description   
1 "$coord" "Coordinates for the individuals"   
2 "$cos2" "Cos2 for the individuals"   
3 "$contrib" "contributions of the individuals"

##### 样本PCA图

然后按照上面的模式来展示下individuals的点图，比如以cos2值来代表各个individuals点的圆圈大小

# 样本PCA图，点大小为cos2，形状为21圆形，按组填充，repel避免标签重叠  
(p <- fviz\_pca\_ind(otu.pca, pointsize = "cos2", pointshape = 21, fill = metadata$Group, repel = TRUE))



ggsave(paste0("p4.pca\_individuals.pdf"), p, width=120, height=100, units="mm")  
ggsave(paste0("p4.pca\_individuals.png"), p, width=120, height=100, units="mm")

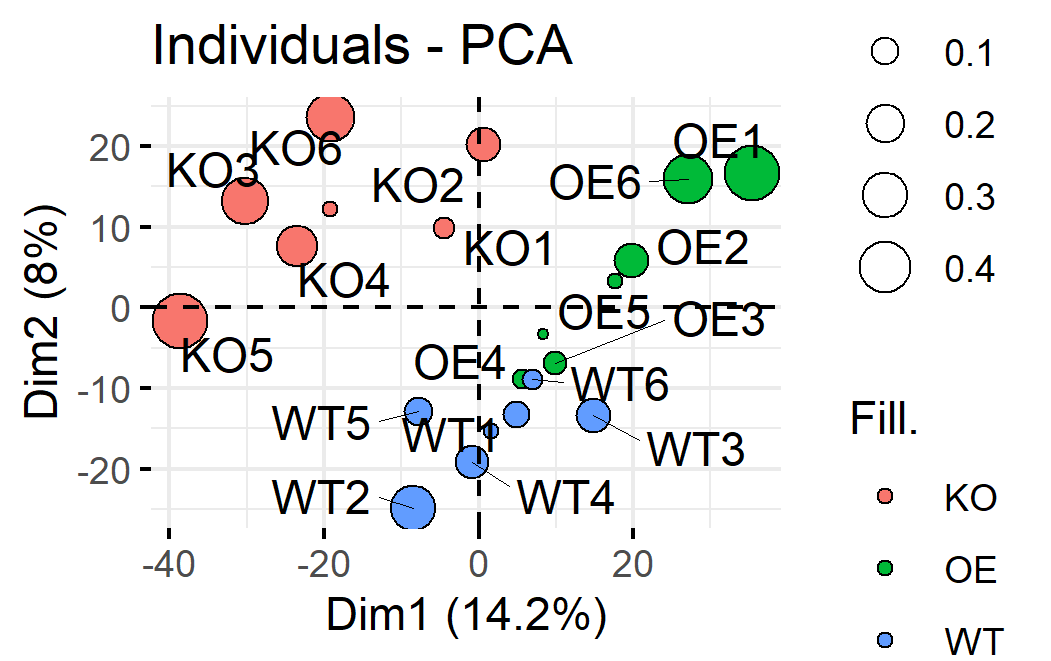
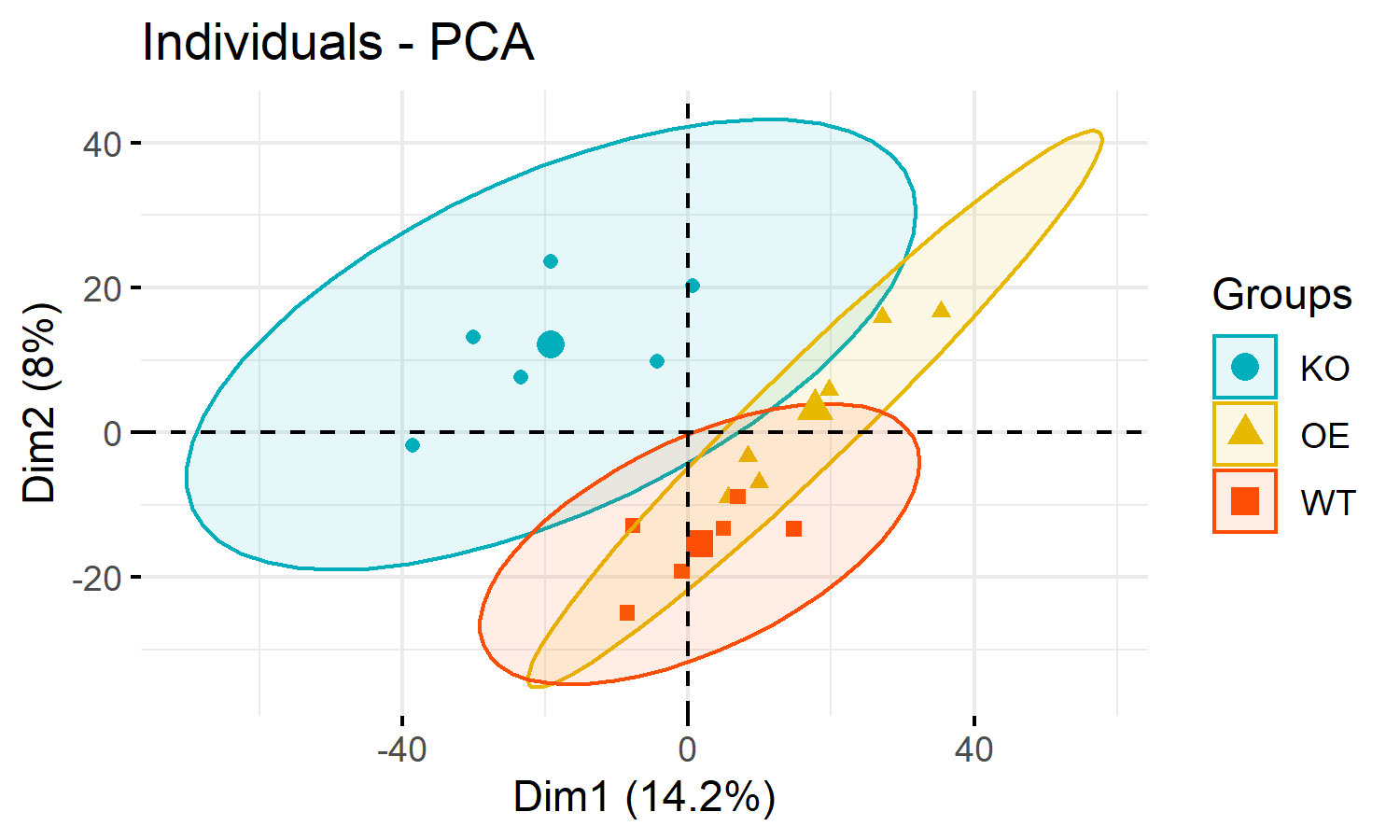


图4. PCA图展示观测值，以cos2值来代表各个individuals点的圆圈大小。

如果有分组信息，则可以将同一组的individuals圈在一起，如：

# 样本PCA图，只显示点，分组着色并手动分配颜色，添加置信椭圆和图例  
(p <- fviz\_pca\_ind(otu.pca,  
 geom.ind = "point", # show points only ( not "text")  
 col.ind = metadata$Group, # color by groups  
 palette = c("#00AFBB", "#E7B800", "#FC4E07"),  
 addEllipses = TRUE, # Concentration ellipses  
 legend.title = "Groups"))



# 保存图片，指定图片为pdf格式方便后期修改，图片宽89毫米，高56毫米  
ggsave(paste0("p5.sample\_group\_ellipse.pdf"), p, width=120, height=100, units="mm")  
ggsave(paste0("p5.sample\_group\_ellipse.png"), p, width=120, height=100, units="mm")

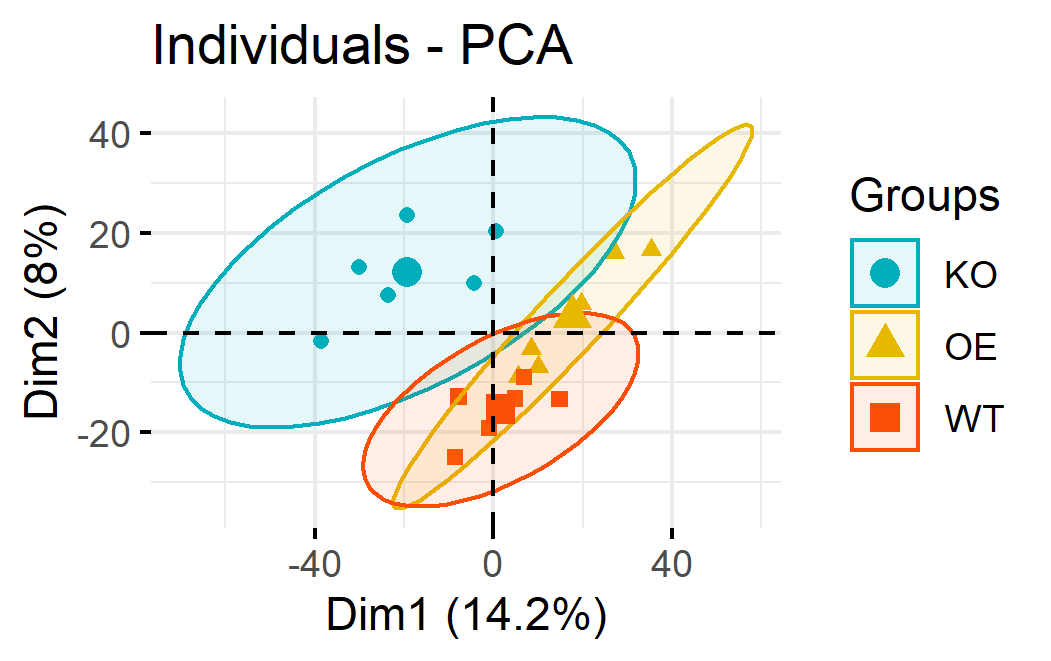


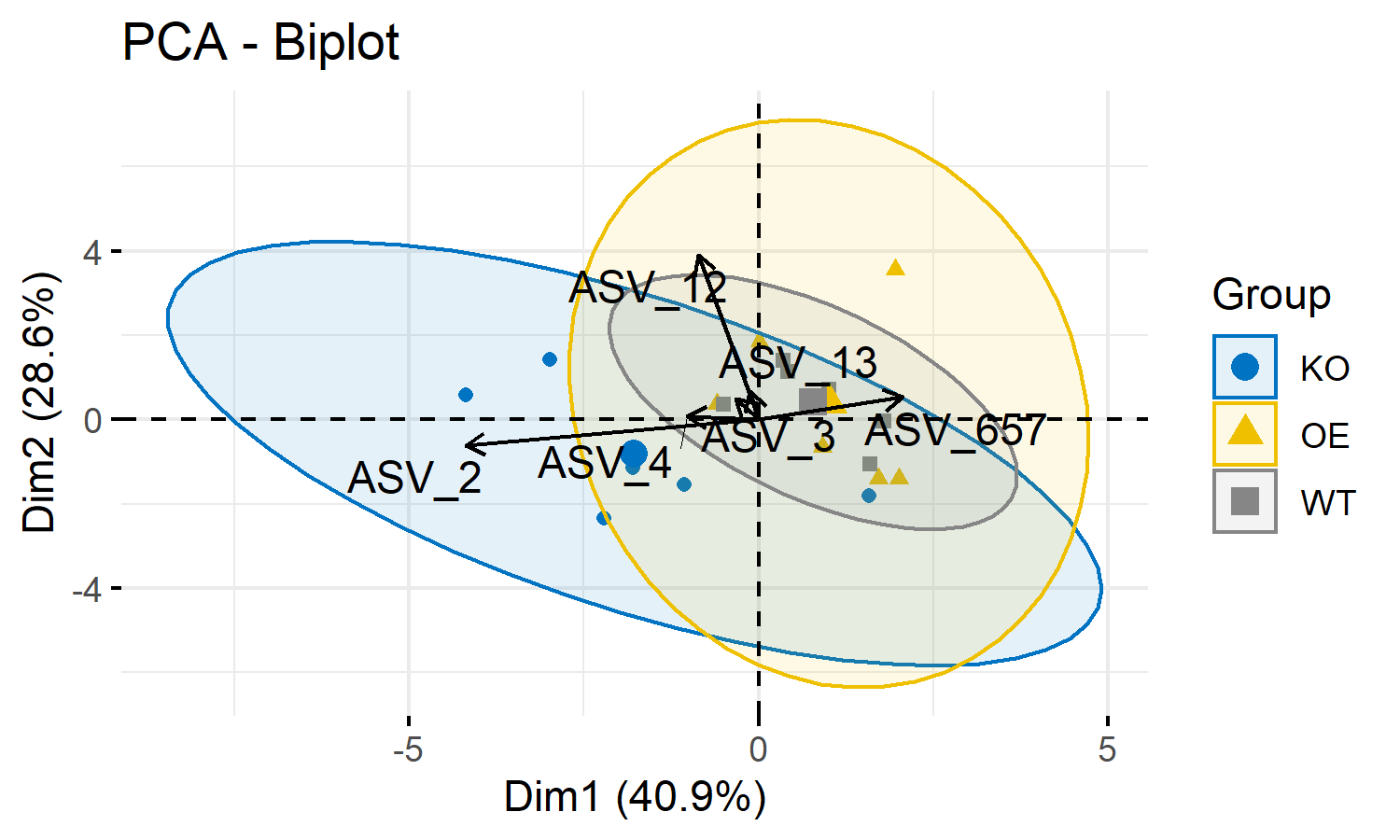
图5. 样本PCA图按分组着色并添加置信椭圆

##### biplot样本和特征图

如果我们想将特征变量(vars)和样本(individuals)同时在一张biplot图中展示，那么就要对主要的OTUs/ASVs/分类进行排序挑选。

展示主要差异ASV与主成分的关系

# 转换原始数据为百分比  
norm <- t(t(count)/colSums(count,na=T)) \* 100  
# 筛选mad(median absolute deviation,中位数偏差的绝对值的中位数,衡量特异波动的方法)值大于0.5的ASV,  
mad.5 <- norm[apply(norm,1,mad)>0.5,]  
# 另一种方法：按mad值排序取前N个，如6个波动最大的ASVs  
# mad.5 <- head(norm[order(apply(norm,1,mad), decreasing=T),],n=6)  
# 计算PCA和菌与菌轴的相关性  
otu.pca <- prcomp(t(mad.5))  
# 绘制观测值PCA主成分分析图，外层()可对保存的图形同时预览  
(p <- fviz\_pca\_biplot(otu.pca, col.ind = metadata$Group, palette = "jco", addEllipses = TRUE, label = "var",  
 col.var = "black", repel = TRUE, legend.title = "Group"))



# 保存图片，指定图片为pdf格式方便后期修改，图片宽89毫米，高56毫米  
ggsave(paste0("p6.sample\_group\_ellipse2.pdf"), p, width=120, height=100, units="mm")  
ggsave(paste0("p6.sample\_group\_ellipse2.png"), p, width=120, height=100, units="mm")

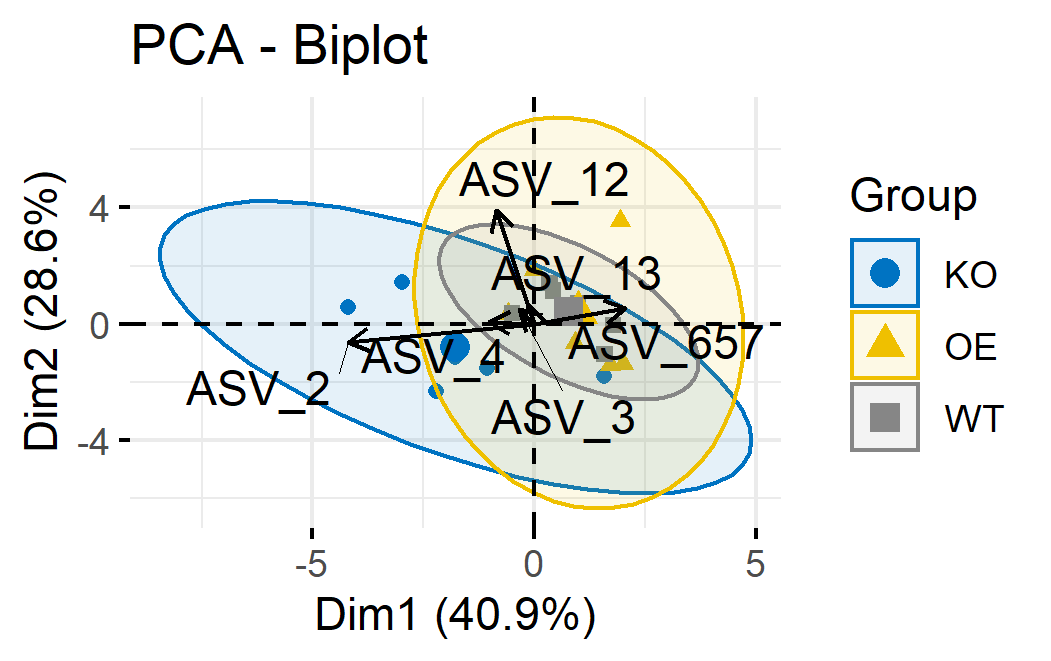


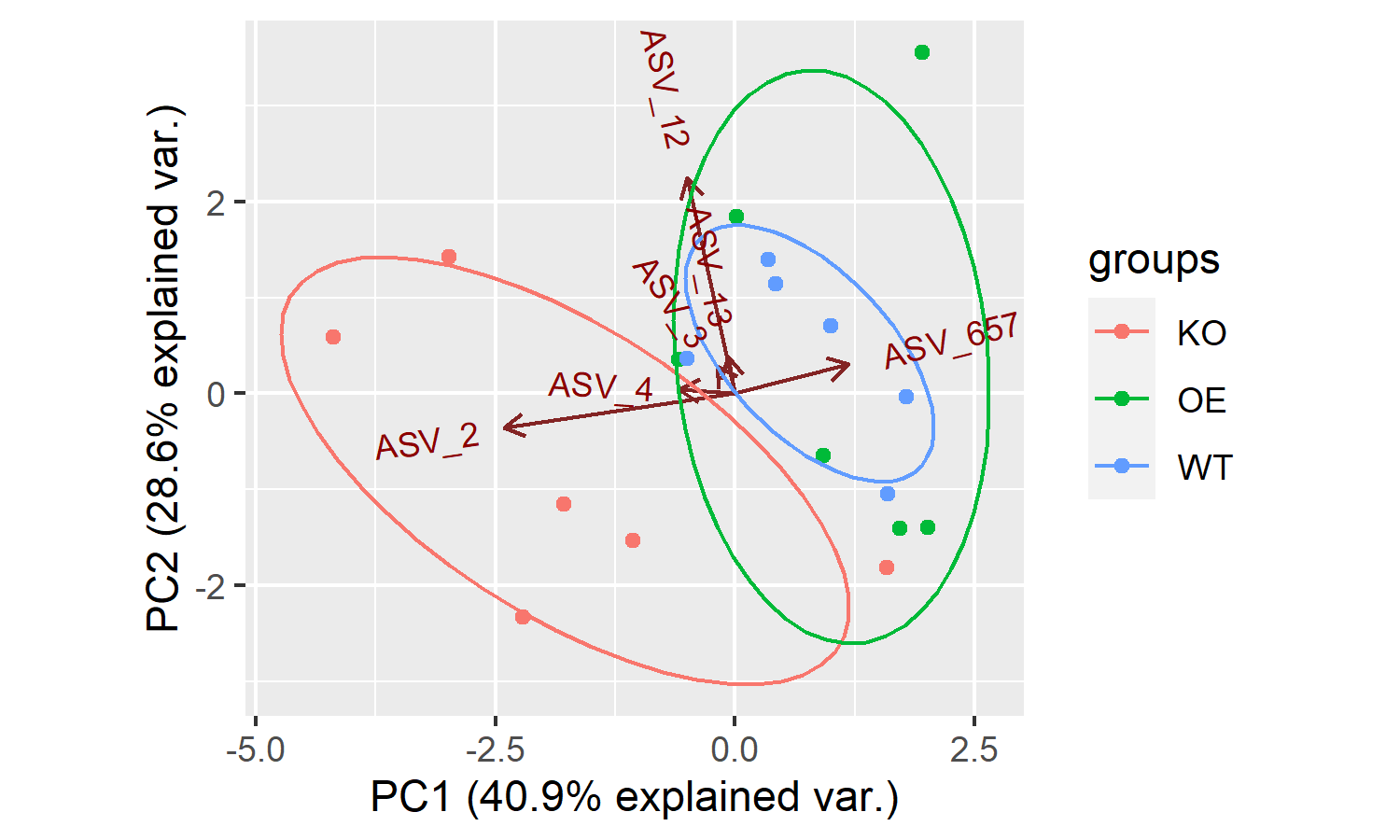
图6. PCA图展示主要ASVs与主成分的关系。

我们仅用中值绝对偏差(mad)大于0.5的6个OTUs进行主成分分析，即可将三组样品明显分开。图中向量长短代表差异贡献，方向为与主成分的相关性。可以看到最长的向量ASV\_2与X轴近平行，表示PC1的差异主要由此菌贡献。其它菌与其方向相反代表OTUs间可能负相关；夹角小于90%的代表两个OTUs有正相关。

##### ggbiplot包可视化PCA图

我们也可以选择ggbiplot包可视化PCA图

# 加载ggbiplot并绘制观测值PCA主成分分析图  
suppressWarnings(suppressMessages(library("ggbiplot")))  
# 绘制观测值PCA图  
(p <- ggbiplot(otu.pca, obs.scale = 1, var.scale = 1, groups = metadata$Group, ellipse = TRUE,var.axes = T))



# 保存图片，指定图片为pdf格式方便后期修改，图片宽89毫米，高56毫米  
ggsave(paste0("p7.sample\_group\_ellipse3.pdf"), p, width=120, height=100, units="mm")  
ggsave(paste0("p7.sample\_group\_ellipse3.png"), p, width=120, height=100, units="mm")

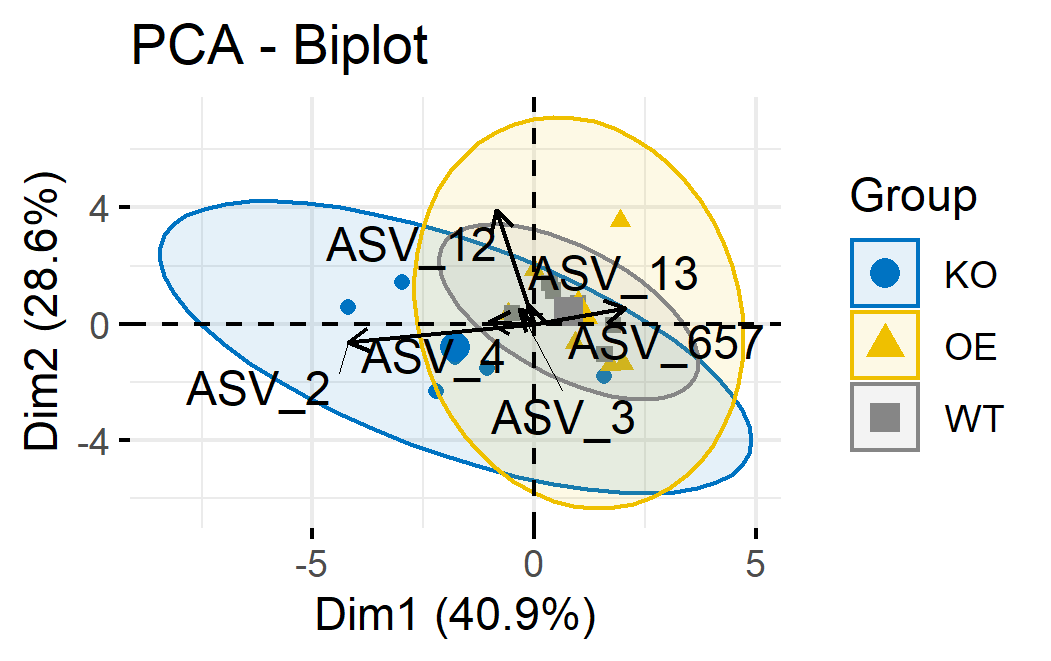
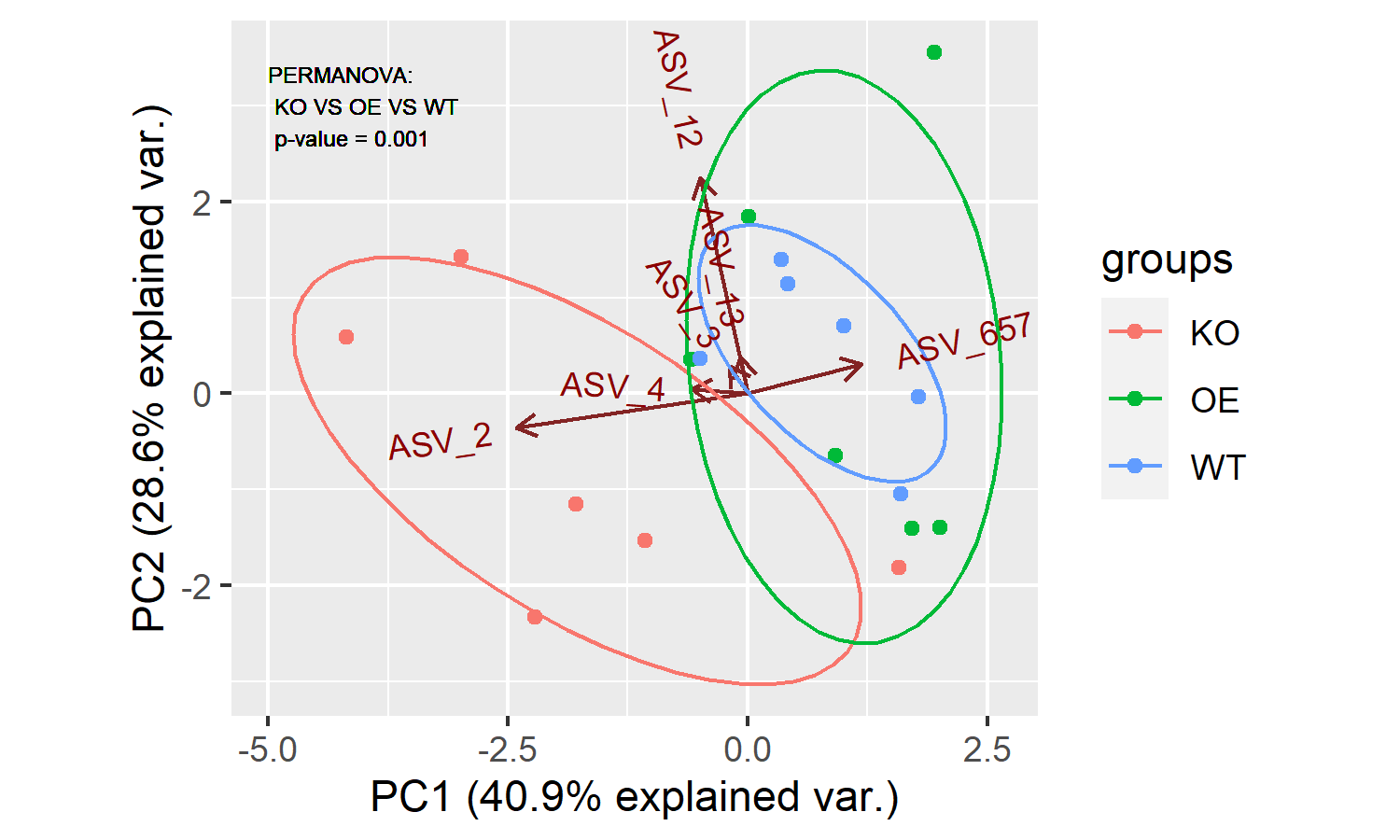


图7. ggbiplot可视化PCA结果

通常情况下，需要使用PERMANOVA来检验不同组样本间的微生物群落是否具有显著差异

# 使用vegan包中的adonis函数进行PERMANOVA分析  
library("vegan")  
otu.adonis <- adonis(t(count) ~ Group, data = metadata, permutations = 999)  
# 之后在绘图代码中将PERMANVOA结果在PCA图中进行展示,同时修改图例  
(p1 <- p + geom\_text(aes(x = - 5 , y = 3, label = paste("PERMANOVA:\n KO VS OE VS WT \n p-value = ",   
 otu.adonis$aov.tab$`Pr(>F)`[1], sep = "")), size = 2, hjust = 0))



ggsave(paste0("p8.sample\_group\_ellipse4.pdf"), p1, width=120, height=100, units="mm")  
ggsave(paste0("p8.sample\_group\_ellipse4.png"), p1, width=120, height=100, units="mm")

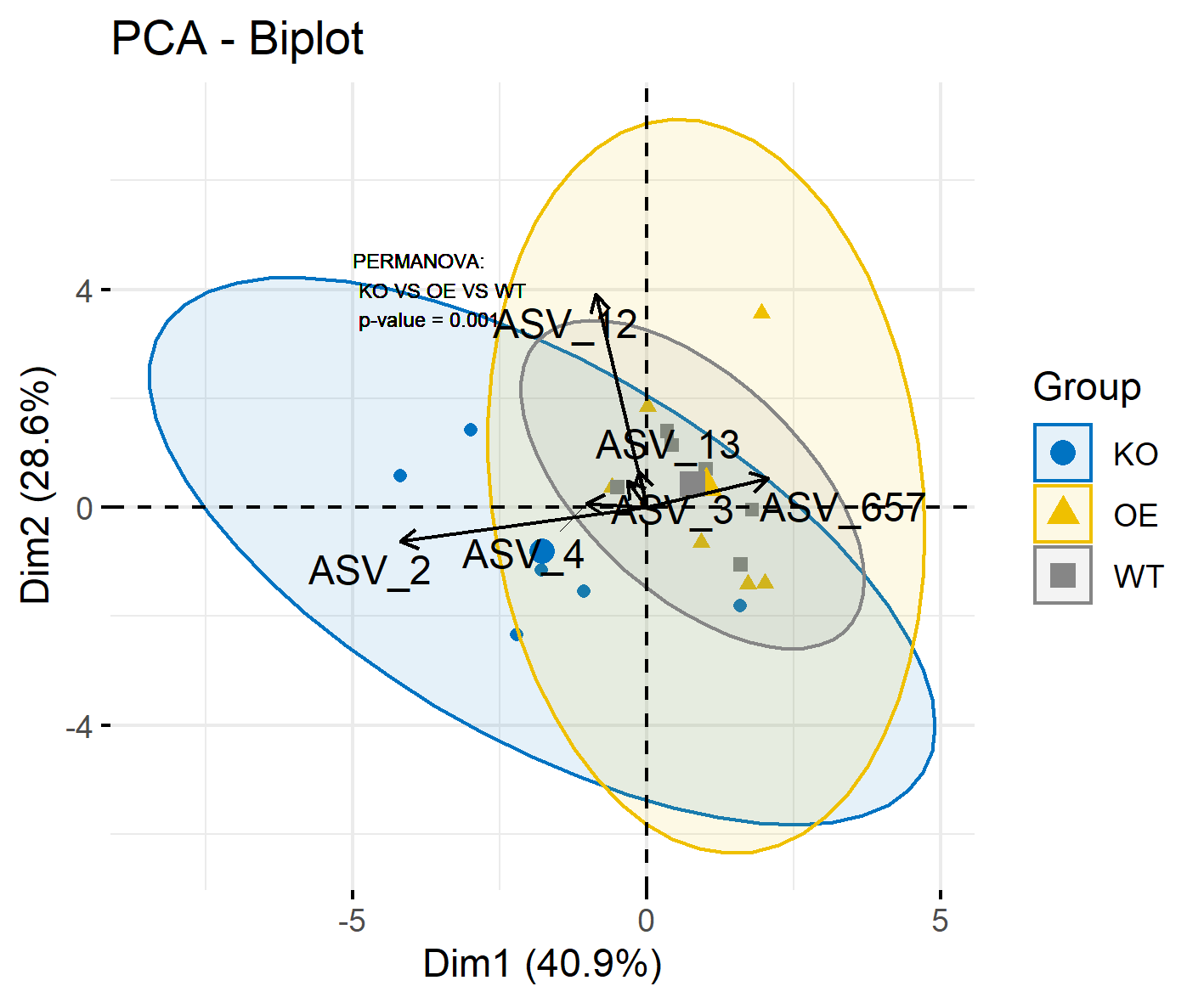
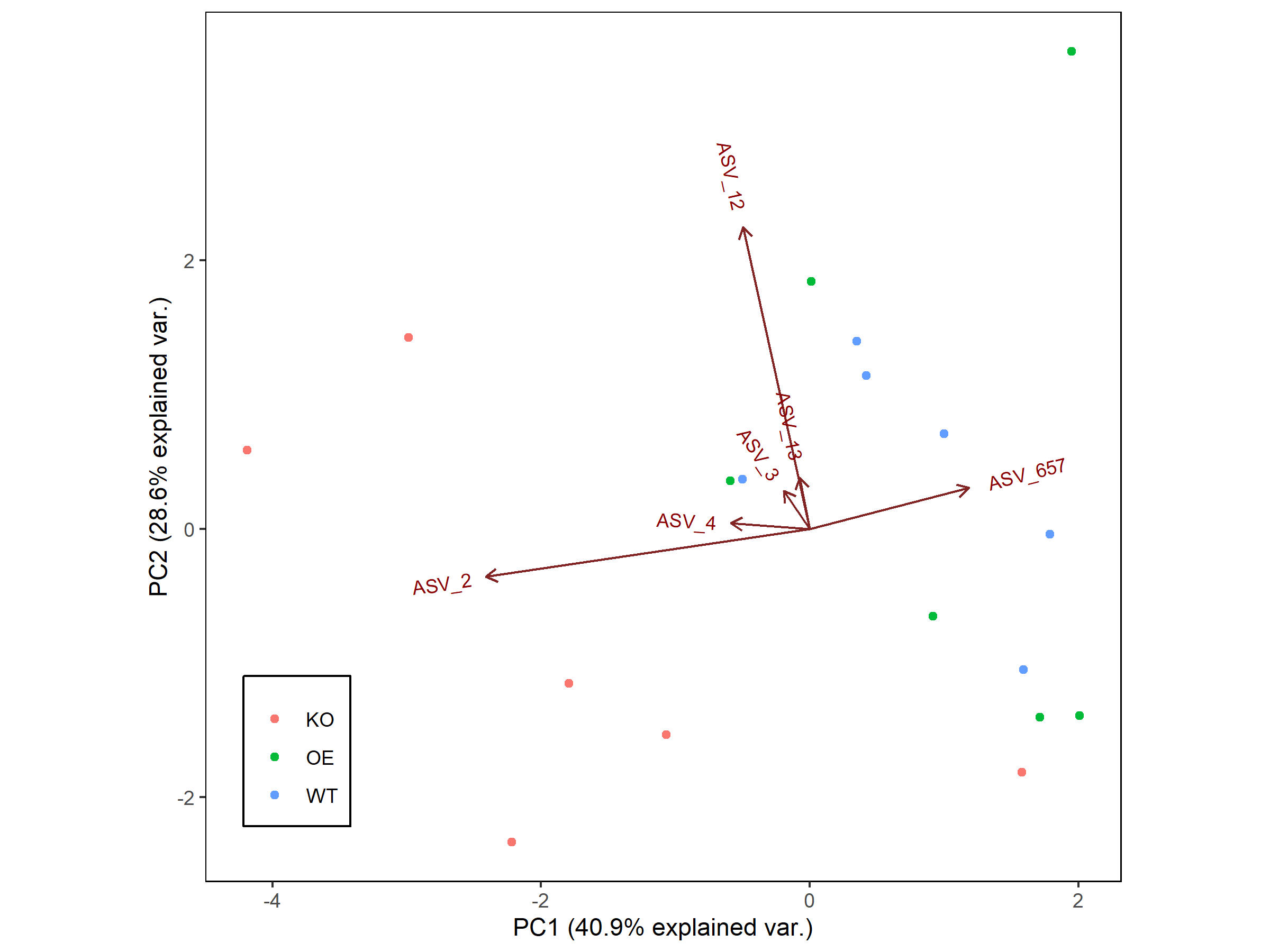


图8. PCA结果图添加PERMANOVA检验

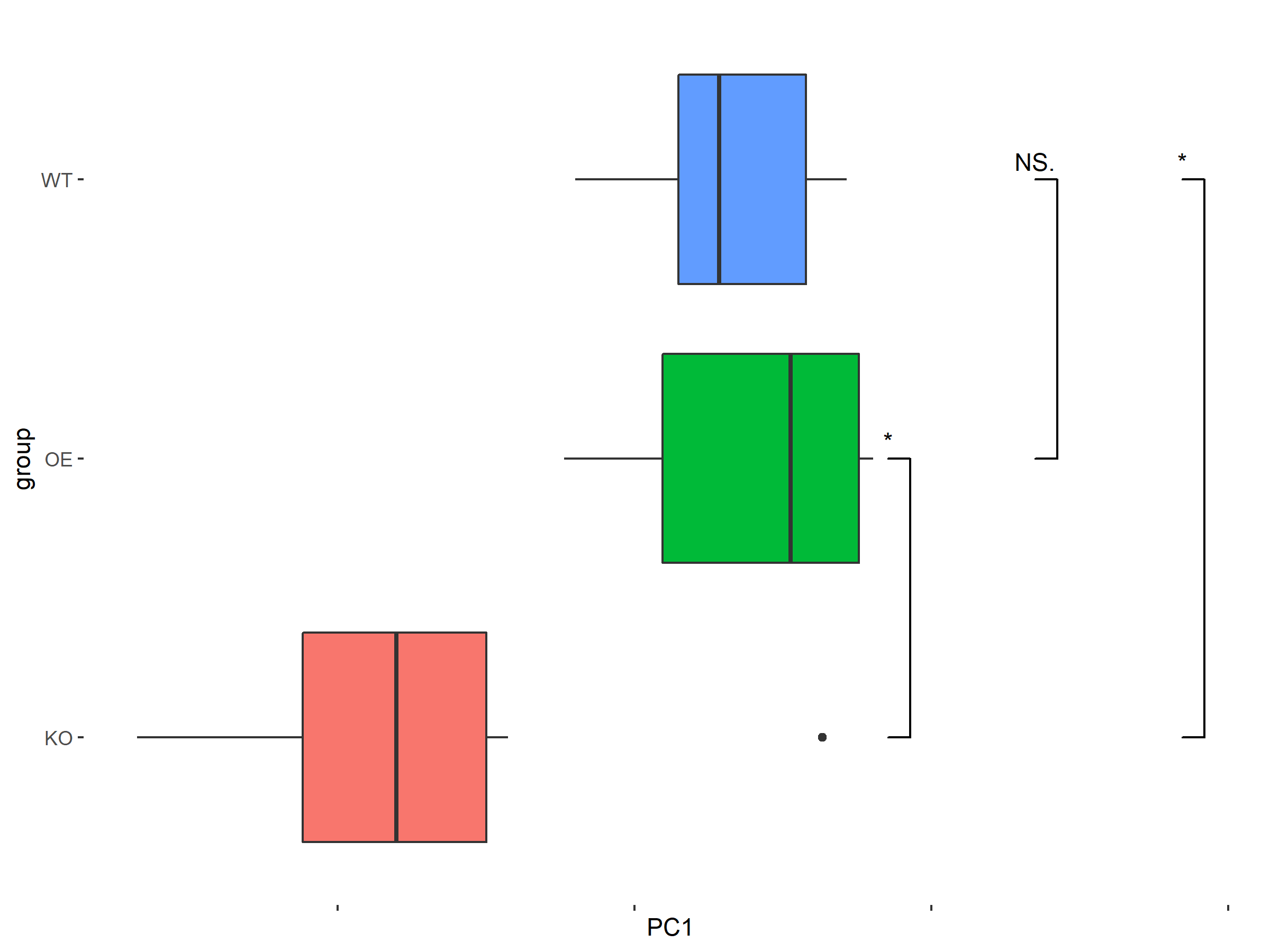
##### PCA图x/y轴添加箱线图

在PCA图的x和y轴添加箱线图，可以实现进一步展示组间差异

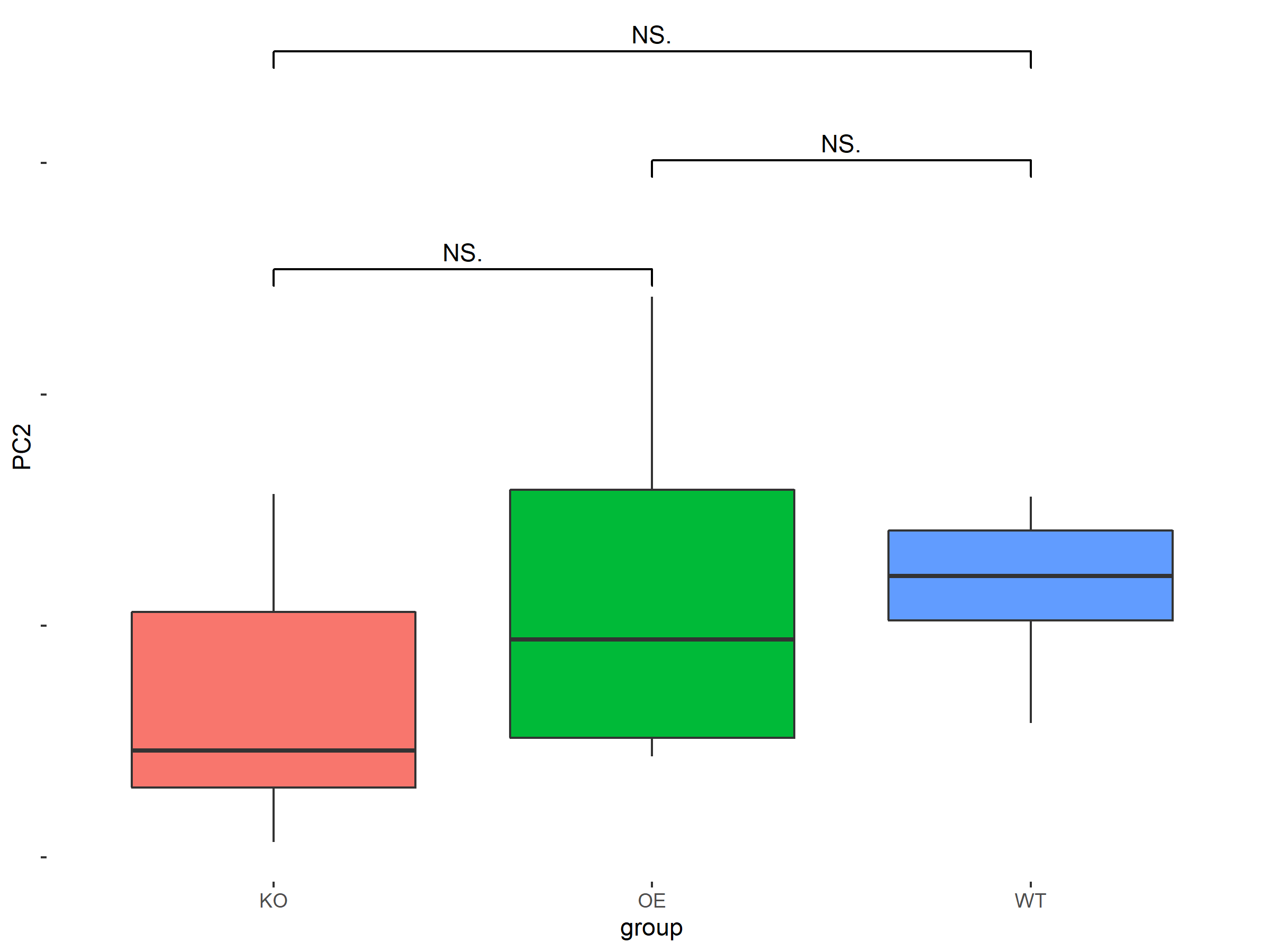
# 需要制作PCA结果可视化的绘图数据文件  
PC1 <- otu.pca$x[,1]  
PC2 <- otu.pca$x[,2]  
otu.pca.data <- data.frame(rownames(otu.pca$x),PC1,PC2,metadata$Group)  
colnames(otu.pca.data) <- c("sample", "PC1", "PC2", "group")  
# 绘制PCA主图,theme()函数调整主题样式与图例位置  
p1 <- ggbiplot(otu.pca, obs.scale = 1, var.scale = 1, groups = metadata$Group, var.axes = T) +   
 theme(panel.background = element\_rect(fill = "white",colour = "black"),  
 panel.grid = element\_blank(), legend.title = element\_blank(),  
 legend.key = element\_blank(), legend.position = c(0.1,0.15),  
 legend.background = element\_rect(size=0.5, linetype="solid",colour ="black"))  
p1



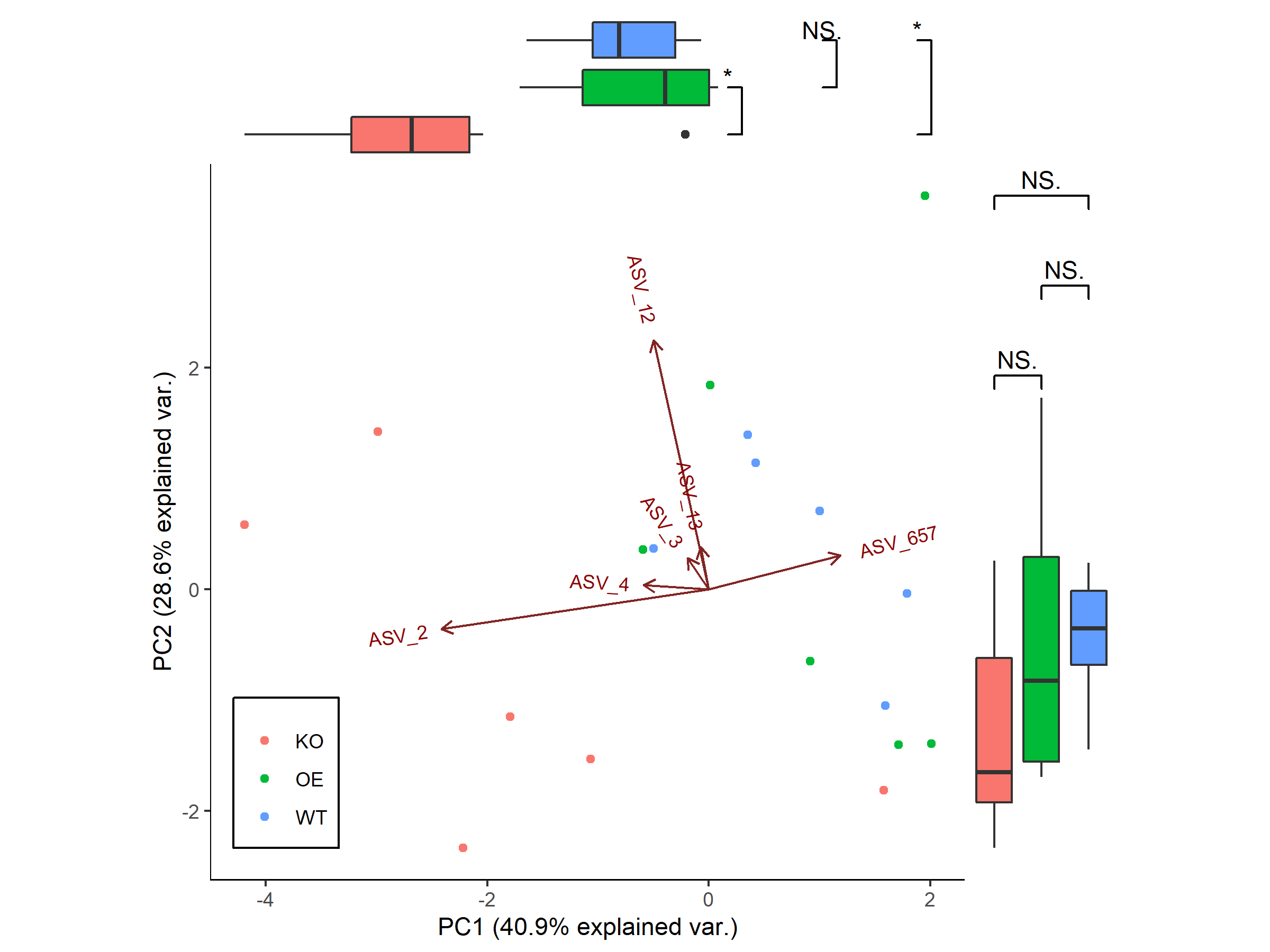
# 加载ggsignif包对组间进行统计检验（ggsignif包主要函数为：geom\_signif()和stat\_signif()）  
library("ggsignif")  
# 设置比较组  
my\_comparisons = list(c('KO','OE'),c('OE','WT'),c('KO','WT'))  
# 绘制y轴为PC1值的分组箱线图,组间比较使用ggsignif()  
p2 <- ggplot(otu.pca.data, aes(x = group, y= PC1, fill = group)) +  
 geom\_boxplot() + coord\_flip() +  
 theme(panel.background = element\_rect(fill = "white",colour = "white"),  
 panel.grid = element\_blank(), axis.text.x = element\_blank(),  
 legend.position="none") +   
 geom\_signif(comparisons = my\_comparisons, step\_increase = 0.2,   
 map\_signif\_level = T, test = wilcox.test)# 添加显著性检验,  
p2



# 绘制y轴为PC2值的分组箱线图  
p3 <- ggplot(otu.pca.data, aes(x = group, y= PC2, fill = group)) +  
 geom\_boxplot() +theme(panel.background = element\_rect(fill = "white",colour = "white"),  
 panel.grid = element\_blank(),  
 axis.text.y = element\_blank(),  
 legend.position="none") +   
 geom\_signif(comparisons = my\_comparisons, step\_increase = 0.2,   
 map\_signif\_level = T, test = wilcox.test)  
p3



# 加载cowplot()包添加副图  
library("cowplot")  
p4 <- insert\_xaxis\_grob(p1, p2, grid::unit(.2, "null"), position = "top")  
p5 <- insert\_yaxis\_grob(p4, p3, grid::unit(.2, "null"), position = "right")  
(p <- ggdraw(p5))



ggsave(paste0("p9.sample\_group\_ellipse\_boxplot.pdf"), p, width=180, height=150, units="mm")  
ggsave(paste0("p9.sample\_group\_ellipse\_boxplot.png"), p, width=180, height=150, units="mm")

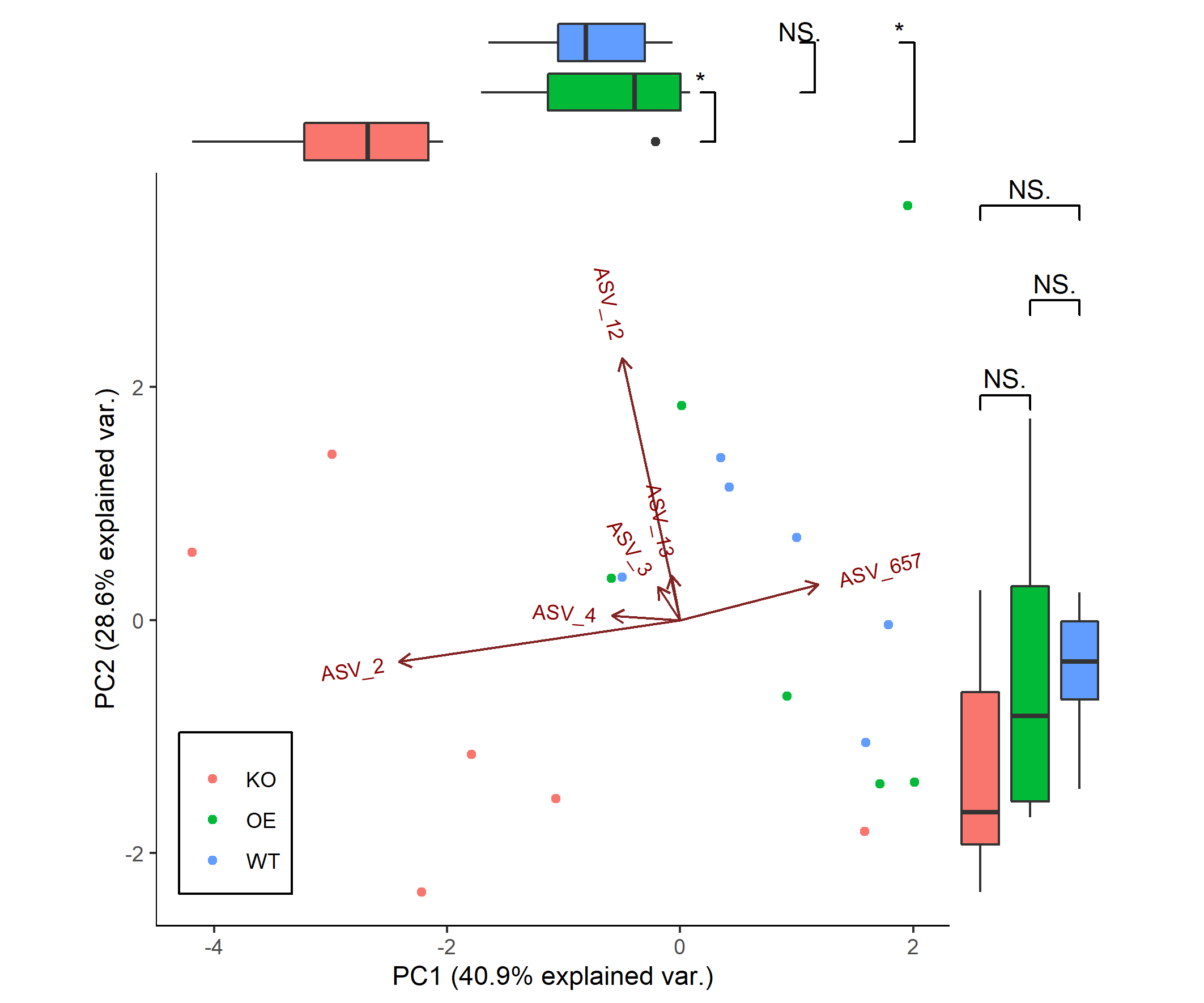


图9. PCA图添加箱线图副图

##### pca3d包可视化PCA图

我们还可以使用pca3d包对数据进行三维展示

suppressWarnings(suppressMessages(library("pca3d")))  
# 绘制样本三维PCA图，分组着色，68%置信椭圆  
(p <- pca3d(otu.pca, group = metadata$Group, show.ellipses=TRUE, ellipse.ci=0.68, show.plane=FALSE))

[1] 0.08376608 0.07110685 0.06971886  
Creating new device

groups colors shapes pch  
KO KO #E69F00 sphere 16  
OE OE #56B4E9 tetrahedron 17  
WT WT #009E73 cube 15

会打开新窗口，展示三维PCA图，而且可用鼠标托动旋转变换观察角度，变量p中保存了各组名、颜色 、形状名称和编号。

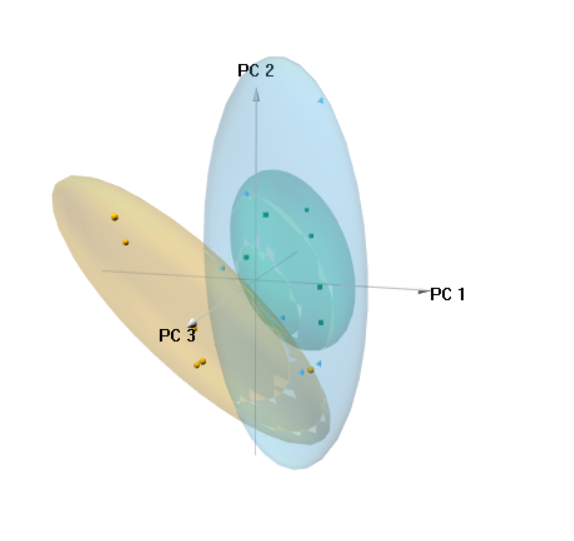


图10. 三维PCA图快照

注：三维PCA图适合交互探索数据，但不能导出矢量图，在发表文章中使用较少。

#### 参考文献

Huanzi Zhong, John Penders, Zhun Shi, Huahui Ren, Kaiye Cai, Chao Fang, Qiuxia Ding, Carel Thijs, Ellen E. Blaak, Coen D. A. Stehouwer, Xun Xu, Huanming Yang, Jian Wang, Jun Wang, Daisy M. A. E. Jonkers, Ad A. M. Masclee, Susanne Brix, Junhua Li, Ilja C. W. Arts & Karsten Kristiansen. (2019). Impact of early events and lifestyle on the gut microbiota and metabolic phenotypes in young school-age children. Microbiome 7, 2, doi: <https://doi.org/10.1186/s40168-018-0608-z>

Alexandre Almeida, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley & Robert D. Finn. (2019). A new genomic blueprint of the human gut microbiota. Nature 568, 499-504, doi: <https://doi.org/10.1038/s41586-019-0965-1>

责编：刘永鑫，中科院遗传发育所

版本1.0.5，PCA图x/y轴添加箱线图和统计及组合的方法由ggpubr替换为ggsignif+cowplot (旧版ggpubr绘制箱线图统计+组合图代码见bak目录)