组学统计 web平台

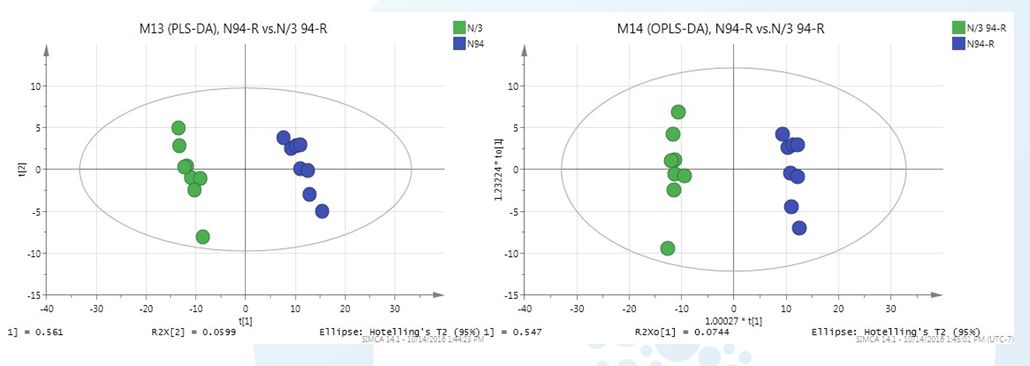
https://www.omicsolution.org/wkomics/main/

原理与流程

https://wenku.baidu.com/view/c0d69285d4d8d15abe234edc.html

PLS-DA和OPLS-DA模型有什么区别

OPLS-DA比PLS-DA多了一个正交换算，把与模型分类不相干的信号过滤掉，因此OPLS-DA解释能力更强。比如组间差异比较小，组内差异比较大时，用PLS-DA的VIP筛选出的可能是组内差异变量，容易误导，而OPLS-DA则优于PLS-DA，可更准确地筛选出组间差异。



为何代谢组学结果不像蛋白质组学结果提供所有鉴定的化合物

一般文献不体现鉴定到总的代谢物，而是描述实验总共检测到多少离子峰，进行差异分析后鉴定到多少个代谢物。 因代谢物存在多种同分异构现象较多，且存在加和形式的多样性，如果定性结果没有经过核验，会存在很多冗余甚至假阳性的结果，所以一般不建议直接使用。代谢组学一般先定量后定性，即先通过多维统计和单变量统计学方法寻找有差异的离子，后经过查库比对，人工校验得到准确的代谢物名称及结构，这样是常用的思路和分析方法。所以不提供所有鉴定的化合物。

差异代谢物的筛选标准

一般同时采用多元变量统计模型的VIP值和单变量统计T检测的P值来筛选差异代谢物。单变量统计分析方法如t检验、方差分析等更加注重代谢物水平的独立变化，多元变量统计分析更加注重代谢物之间的关系以及它们在生物过程中的促进/拮抗关系。同时考量两类统计分析方法的结果，有助于我们从不同角度观察数据，得出结论，也可以帮助我们避免只使用一类统计分析方法带来的假阳性错误或模型过拟合。筛选的阈值一般是VIP＞1和P＜0.05。如果得到的差异代谢物很多，可以再加上差异倍数这个筛选条件。

筛选不到差异代谢物怎么办

如果利用常用的筛选标准（VIP＞1和P＜0.05）没有找到差异代谢物，首先可以筛选阈值设得宽放些，如VIP＞1和P＜0. 1。如果还是没有筛选出差异代谢物，那么还可以对检测到的物质进行单维统计学分析，如FC>1.5和P＜0.05，通过每单个代谢物组间样本的表达值比较分析差异代谢物，并用火山图展现。

找到差异代谢物之后，如何验证这个代谢物的功能

首先要进行定性定量验证。主要是采用三重四级杆进行靶向验证，获得绝对定量信息。然后通过前期的代谢通路分析，我们知道这个差异代谢物属于哪条调控通路，那么可以对该通路上的分子进行功能验证。另外，还可以增加其他组学的数据，如转录组、蛋白组，进行相互补充。

统计分析：

第一步为非监督多元统计分析，通常使用PCA（主成分分析）。使用非监督分析有以下几个目的：

直观的观察被分析样本有无天然的分组

检查异常样本（在置信区间之外的点）

揭示研究中存在的隐藏的偏向性

展示样本分类的细节信息

这一步分析可以看作是一个数据质量控制的过程，如果样本点在score plot（得分图）中根据样本的分组展现出一定程度聚集，则证明数据的质量可信度。此外也可以在QC样本点被移除之前，通过观察QC样本点的空间分布来判断数据的质量，如果QC样本点紧密聚集则证明数据质量高。

在PCA分析之后，我们需要去除异常值（样本及变量），因此数据集的大小将会有所改变。通常来源于分析时程中，由于操作偏差引起的异常值需要从数据集中删除；但是，有些时候这些异常值可能并不是由于操作误差引起，可能代表了数据中一些新的发现，则这些数值需要保留用作进一步分析。

第二步可能进行单变量统计分析，来筛选在不同组别中差异有统计学意义的变量。单变量统计分析在分析数据时，数据之间相互独立；多变量分析则考虑数据之间的相互作用和相关性，因此二者可以提供不同的关于数据的信息。

使用单变量分析为多元统计分析进行数据的预先筛选是一个有争议的操作，一些研究者不建议这种筛选方式，另一些则推荐使用此方法，将筛选之后的化合物进行后续有监督分析。

单变量分析多也被用在有监督分析之后，来检测通过有监督分析选择出的标记物在不同组别之间的差异有无统计学意义。（目前多数都是这么使用）

第三步是有监督的多元统计分析如PLS-DA（偏最小二乘判别分析），用以选择对样本分类贡献较大的变量即筛选标记物。这一步可以作为数据分析的最后一步，或者在这一步之后接着做单变量统计分析来检测所筛选的化合物的差异有无统计学意义。

需要注意的是，有监督模型建立之后需要进行模型的验证，如置换检验（permutation test，PLS-DA），交叉验证（cross-validation，OPLS-DA）等。

以上就是数据分析的一个常用流程，其中第二步通常有争议，目前多数都放在标记物通过有监督分析之后进行检查所选标记物的差异有无统计学意义。

关于标记物的筛选，还可以有很多其他的方法，参见上一篇公众号文章。

代谢物的鉴定和标记可以在数据分析结束之后进行，也可以在数据分析之前进行，在分析之前进行可以帮助我们选择唯一变量（即去除加合离子，碎片离子等）进行分析。

流程：

http://www.360doc.com/content/18/0929/13/54670480\_790664303.shtml

详细流程：

https://www.cnblogs.com/jessepeng/p/12116920.html

https://www.jianshu.com/p/ce3f27635019

知识点集合：

https://blog.csdn.net/fjsd155/article/details/82918974

有监督分析和无监督分析

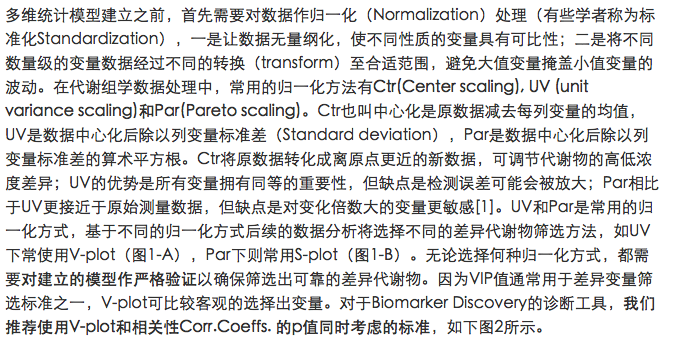
https://mp.weixin.qq.com/s?\_\_biz=MzU2MzMzOTk1Mg==&mid=2247484793&idx=1&sn=dc30ac709cf16d7c2149dc624c01b054&chksm=fc5af2a6cb2d7bb012f6f432d5a848e710a416d5d35b7830b554dde57bc988c5cb786772fc76&mpshare=1&scene=1&srcid=

OPLS-DA原理及图表

https://mp.weixin.qq.com/s?\_\_biz=MzU2MzMzOTk1Mg==&mid=2247484881&idx=1&sn=a8b079763b17d620d11a96a62fa06c9d&chksm=fc5af20ecb2d7b18905a560e6f138902a7b1d0ae996d8444d34506e3af9267393fc89fed32e1&mpshare=1&scene=1&srcid=&sharer\_sharetime=1564982593306&sharer\_shareid=57f69e562db94e4e682fc0073cfaa3cd

其它缺失值处理：

http://www.360doc.com/content/18/0920/00/54670480\_788098803.shtml



UV标准化=zscore标准化=

一、scale函数

scale函数默认的是对制定数据做均值为0，标准差为1的标准化。它的两个参数center和scale：

1）center和scale默认为真,即T

2）center为真表示数据中心化

3）scale为真表示数据标准化

中心化：所谓数据的中心化是指数据集中的各项数据减去数据集的均值。

标准化：标准化就是数据在中心化之后再除以标准差。变换后值域为[0,1]。

Standard2<-scale(iris[1:4],center=F,scale=T)

https://www.jianshu.com/p/1e7eb92247c5

新R包 muma

#先muma导入，scaling，不填充缺失值

NormalizeMets包，k邻近填补，scaling，

PLSDA 需要mixomics包

http://blog.sina.com.cn/s/blog\_942438cf0102wth0.html

ropls包

http://blog.sciencenet.cn/home.php?mod=space&uid=3406804&do=blog&id=1213011