样品分组及差异算法选择

您在做高通量测序分析前，可提前与我们的项目人员沟通您的实验设计思路。这样在我们的技术人员对您的样品及分组有更多了解的情况下，可以更有效率的帮助您进行项目分析，并可能为您的实验提供出更为合理的建议。

关于样品的命名与分组，我们提供如下建议供您参考：

一、 样品命名规范

1.1样本命名方法

1. **样本命名以英文字母开头**（很多软件不支持数字开头），可以是英文字母、数字和_的组合，不包含中文或者其他特殊字符（如# @ *? ( ) [ ] / \ = + <> : ; “ ‘, *^|& . —空格等）。此外 1-1,12-3 这种字样，很容易由于 excel 默认格式中变成 42370，42707 导致后期分析样品名对应错误
2. **长度建议最好不要超过 6 个字符；**
   A:名字太长，在得到的分析结果的表格中不能完整显示
   B:因为管壁和管盖空间有限，名字太长写不下，也容易出错。
3. **尽量避免使用容易混淆的字母和数字，如“0”，“o”，“1”，“l”等**

1.2 样本命名示例

1. **根据样本生物学属性意义 选取首字母数字组合。**
   样本命名最好能体现出其生物学属性和重复情况。比如：“W2R3”，“W”表示野生型，“2”表示第 2 个时期，“R”表示根（root 首字母），“3”表示第 3 个生物学重复等；
2. **根据样本的项目分组编号命名**
   如您的样品为来自于临床病人的样品，为保护病人的隐私，一定不能用病人的名字进行命名，需对样品加以编号命名，并且编号能够体现病人关于科研目的分类特征。比如：癌症病人组可用编号 C01；C02；C03；C04……；正常病人组可用编号 N01；N02；N03；N04……。

如果您在测序或质检时没来得及命名，可在项目管理与您确认项目启动时重新对样品命名。

二、 样品分组建议

2.1 设置生物学重复

目前多数差异算法都不再支持无生物学重复样品间的比较，建议您在进行实验设计的时候尽量设置生物学重复。一般建议至少设置 3 个生物学重复，对于临床样本，由于个体差异较大，建议设置 5 个以上的生物学重复。对于无生物学重复的比较组，我们也可以通过一定算法计算差异（如 DEGseq 方法），但是由于个体的差异可能较大，算法无法完全消除由个体差异导致的假阳性/假阴性结果。

2.2 唯一明确分组

对于同批下机的样品，尽量对样品进行唯一、明确的分组，分组不清晰会影响数据校正及差异分析结果。例如样品 A 同时属于 groupA（A、B、C）、GroupB（A、D、E），数据在整体校正时会出现混乱，我们后续分析时可能采用分组校正的方法进行分析，这会导致不同比较组中 A 样品的校正值不同。

2.3 分组方法一致

另外，如果您同批下机的数据有多个比较组，而多比较组中如果同时包含有多样品--多样品、单样品—单样品、多样品-单样品等混合比较的情况时，数据校正方法及适用差异算法会比较混乱，最终可能影响差异结果的计算。

2.4 合理命名比较组

建议对比较组进行合理命名，避免后续出现比较组名称过长、图表展示结果不符合文章发表规范等情况。

2.5 差异比较组示例

基于以上几点，请您在反馈比较时先为本次所有测序样品分组，再依据样品分组情况设定比较组。例如：

样品	组别
A	Tumor
B	Tumor
C	Normal
D	Normal

比较组信息

Control	Treatment
Normal	Tumor

三、 差异算法简介

目前锐博生物支持 DESeq2、DESeq、edgeR 及 DEGseq 四种算法，这四类算法也是目前高通量测序相关文献中常引用的差异算法。各种算法基于的模型及输入数据不同，适用的情况也略有不同，详情请参考下表。

方法	特点
DESeq2	DESeq2 是 DESeq 的继承者，计算方法基于负二项广义线性模型（Negative Binomial Generalize Linear Model）。DESeq2 结合了方法论的新特点--使用收缩估计计算分布和差异倍数，提高分析结果的稳定性和可解释性，使 RNA-seq 数据的差异分析更准确。 输入为**原始 count 值**，**适用于有生物学重复样品的组间差异分析。** DESeq2 为目前高分文献中常引用方法，也是**默认差异算法**。
DESeq	DESeq 基于负二项分布模型，输入为**原始 count 值**，软件自动校正表达值。**适用于有生物学重复样品的组间差异分析**
DEGseq	DEGseq 基于二项分布或泊松分布模型，输入文件为**校正后的表达值**（如 RPKM、FPKM、TPM 等）。**适用于无生物学重复的样品间差异分析**

edgeR	edgeR 同样基于负二项分布模型，但与 DEseq 基于不同的数学假设。输入文件为原始 count 值。长 RNA 测序中较少选用，小 RNA 测序中较为常用。**可用于有生物学重复样品的组间差异分析或样品间比较**

参考文献:

1. Love MI, Huber W, Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology*, **15**, 550.
2. Anders S, Huber W (2010). "Differential expression analysis for sequence count data." *Genome Biology*, **11**, R106.
3. Wang, L., Feng, Z., Wang, X., Wang, X., & Zhang, X. (2010). "DEGseq: an R package for identifying differentially expressed genes from RNA-seq data". *Bioinformatics*, *26*(1), 136-138.
4. Robinson MD, McCarthy DJ, Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics*, **26**(1), 139-140.