An overview of the Chinese version [6]

Batch effect

Batch effect 是指在实验设计和数据采集过程中，由于技术和环境因素的影响，导致不同批次（或实验）的数据之间存在系统性偏差，而这些偏差与我们要研究的生物学差异无关。因此，它是一种技术差异，不应该被错误地解释为生物学上的差异。

Batch effect 可能会掩盖或误报生物学上的信号，降低实验的可重复性和可靠性，因此需要进行识别和纠正。常见的 Batch effect 包括实验日期、操作员、批次、处理时长、试剂批次等。Batch effect 可以通过各种方法进行识别和纠正，包括但不限于以下几种：

1. 手动调整：通过手动调整各批次数据的表达量，使它们在整个数据集上保持一致。然而，这种方法需要较高的主观性和经验，并且可能会导致其他偏差的引入。
2. 基于 PCA 的批次效应校正：通过主成分分析（PCA）等方法来确定批次效应所占的主要变异，并使用线性回归或正交回归等方法来消除批次效应。这种方法通常适用于只有一个或少数批次的情况。
3. 基于统计模型的批次效应校正：使用批次作为固定效应或随机效应，将批次因素纳入到线性模型中，从而消除批次效应。这种方法适用于有多个批次或复杂设计的情况。
4. ComBat 批次效应校正：这是一种特定的统计模型，使用贝叶斯方法来估计批次效应，并将其纳入到线性模型中进行校正。它适用于多个批次和复杂的设计，并且可以处理各种数据类型，例如基因表达数据、蛋白质组学数据、代谢组学数据等。

纠正批次效应是基因表达分析中一个重要的步骤，它可以提高数据的可重复性和可靠性，从而更准确地发现生物学差异和识别生物标记物

Calculation of the adjustment batch effect

Dividing by more feature specific size factors

	Sample 1	Sample 2
Gene A	16/(s_j*l_i*gc_ij*M_i*…)	5/(s_j*l_i*gc_ij*M_i*…)
Gene B	13/(s_j*l_i*gc_ij*M_i*…)	3/(s_j*l_i*gc_ij*M_i*…)
Gene C	7/(s_j*l_i*gc_ij*M_i*…)	0/(s_j*l_i*gc_ij*M_i*…)
Gene D	28/(s_j*l_i*gc_ij*M_i*…)	12/(s_j*l_i*gc_ij*M_i*…)

Multiplicative model behind, K_{ij} is the read count for the i th gene and j th sample.

$$K_{ij} = \theta_{ij} \times s_j \times l_i \times f_j(\mathbf{gc}_i) \times M_i \times \ldots$$

- θ_{ij} : the true gene expression level (target of estimation).
- s_j : sequencing depth.
- l_i : gene length.
- $f_j(gc_i)$: **GC content bias.**
- M_i: **read mappability.**

这个公式是一个典型的 RNA-seq 表达量计算公式，其中各个变量的含义如下：

- Kij 表示基因 i 在样本 j 中的表达量；
- θij 表示 RNA-seq 测序深度归一化系数；
- sj 表示样本 j 的测序深度（即所得的 read 数）；
- li 表示基因 i 的长度（以 bp 计算）；
- fj(gci) 表示基因 i 的剪接形式对应的比例因子，gci 表示第 i 个基因的第 c 个剪接形式；
- Mi 表示基因 i 的匹配到的 reads 数；
- … 表示其它可选参数。

其中，θij 是 RNA-seq 中的归一化因子，它的目的是将不同样本之间的测序深度进行调整。sj 是样本 j 的测序深度，它代表每个样本的测序覆盖度，是 RNA-seq 分析中一个重要的质控指标。li 表示基因 i 的长度，由于计算的是基因的表达量，因此基因的长度也是需要考虑的因素。fj(gci)表示基因 i 的各个剪接形式的比例因子，考虑到同一个基因的不同剪接形式会对表达量计算产生影响。Mi 是基因 i 匹配到的 reads 数，也是表达量计算中的重要参数之一。综合考虑这些因素，就可以得到基因 i 在样本 j 中的表达量 Kij。

Read genome mappability

Read genome mappability 是指基因组上可以被 reads 准确比对的区域或位置的比例。由于基因组中存在一些高度相似或反向互补的序列，这些区域可能会导致 reads 比对到多个位置，从而影响后续的数据分析和解释。因此，了解基因组的 mappability 是非常重要的。

通常，mappability 可以通过在参考基因组上模拟一定数量的 reads，然后计算成功比对到基因组上的 reads 数量与模拟的 reads 数量之比来估计。另外，一些公共数据库也提供了基因组的 mappability 信息，例如 UCSC Genome Browser 的 "mappability track"。

在基因组数据分析中，mappability 的考虑可以用于过滤低质量的 reads 和区域，以及确定表达量计算的有效基因组区域。同时，在一些结构变异检测、基因注释和突变检测等分析中，mappability 也被用于帮助识别高可信度的变异和基因组区域。

Sequencing artifacts

Sequencing artifacts 是指在测序过程中产生的偏差或误差，这些误差不是由样品本身引起的，而是由测序技术或实验过程中的其他因素引起的。这些误差可以影响测序数据的质量和可靠性，从而影响后续的数据分析和解释。

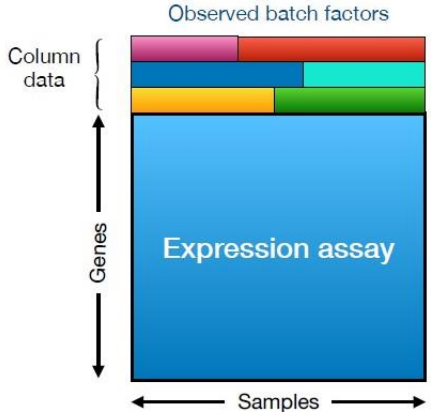
常见的测序偏差和误差包括：

1. 测序错误：由于测序仪器的误差或者试剂的质量问题，会导致碱基的读取错误，从而产生序列的错配或者错位。
2. GC 偏差：由于不同的 DNA 序列含有不同比例的 GC 碱基，如果在不同的 DNA 片段中 GC 含量的分布不均匀，会导致在测序中的碱基比例失衡，从而影响测序数据的可靠性。
3. 长度偏差：由于 PCR 扩增、DNA 片段化等实验操作，不同 DNA 片段的长度分布不同，从而影响测序深度和表达量分析的准确性。
4. 交叉污染：在实验操作过程中，不同样品之间可能会发生交叉污染，从而导致样品间的序列混淆，影响数据的准确性。

为了避免和纠正测序偏差和误差，研究人员可以采取一系列措施，如在实验设计阶段考虑样品分组、重复实验等，选择合适的测序平台和质量好的试剂，对测序数据进行过滤和校正等。同时，也可以利用一些工具和软件对数据进行检测和校正，例如 FastQC、Trim Galore 等。

Combat (supervised batch effect modeling)

Combat 是一种用于纠正批量效应的方法，当我们知道导致批量效应的关键混杂因素时。它的工作原理是将多元线性回归模型拟合到基因表达数据中，其中已知的混杂因素和实验设计因素都用作模型中的协变量。然后，该模型估计每个协变量对基因表达数据的影响，并去除由于已知混杂因素造成的不需要的变化。



Combat like approach (Regression only):

$$y = \mu + \text{Treatment } \alpha + \text{Batch } \beta + e$$

$$\hat{y}^{\text{correct}} = y - \text{Batch } \hat{\beta}$$

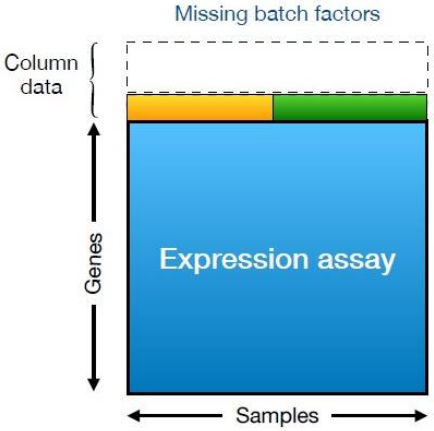
Fit linear regression first to estimate $\hat{\beta}$

Subtract the batch effect term from the gene expression vector to get the corrected gene expression

ComBat 是一种有监督的批次效应校正方法，它使用经验贝叶斯框架来估计和去除批次效应。ComBat 的基本思想是对数据拟合一个线性模型，其中批次指示变量被包括作为协变量。然后使用这个模型的残差来估计批次效应的参数，如每个批次的均值和方差。这些估计值然后被用来调整数据以去除批次效应。

SVA (unsupervised batch effect modeling)

当批处理因素未知且不能直接计算时，使用无监督方法。这些方法使用诸如 PCA 或其他因子分析算法等技术来估计"潜在"批因子。替代变量分析(SVA)是一种复杂的 PCA 形式，可以估计批量影响因素，同时也隔离实验设计因素的影响。



SVA like approach (1. PCA, 2. Regression):

Step1:
Use PCA / SVA to estimate Batch from the entire gene expression matrix

Step2:

$$y = \mu + \text{Treatment } \alpha + \hat{\text{Batch}}\ \beta + e$$
$$\hat{y}^{\text{correct}} = y - \hat{\text{Batch}}\ \hat{\beta}$$

而 SVA 是一种无监督的方法，它使用称为因子分析的统计技术来识别和去除批次效应。SVA 通过识别"代理变量"来捕捉不需要的变异源，如批次效应，然后在下游分析中将这些代理变量作为协变量。

NGS situations

大多数的 NGS 实验并没有严格的对照组。因此，在数据分析过程中，我们需要识别和校正实验中可能出现的各种偏差和误差。在实践中，这通常需要使用各种统计技术和计算工具进行调整和矫正，以确保实验结果的准确性和可靠性。

其中一种常用的方法是使用 trial and error 策略，这通常包括尝试多种校正方法，并通过比较其效果来选择最佳的方法。例如，我们可以使用 PCA 或 t-SNE 等技术对样本间的方差或相似性进行可视化，以确定是否存在批次效应或其他技术偏差。接下来，我们可以尝试不同的校正方法，例如使用 ComBat 或 SVA 等算法进行批次效应校正，或者使用 RUVSeq 等算法来消除其他来源的技术噪声。

在尝试不同的校正方法时，我们通常需要根据实验数据的特点和样本属性进行一定的调整和优化。例如，我们可能需要考虑样本大小、基因丰度分布、批次大小和数目、协变量的选择和调整等因素。通过 trial and error 策略，我们可以找到最适合特定实验数据和分析流程的校正方法，并最大程度地减少偏差和误差的影响，从而提高实验结果的可靠性和可重复性。