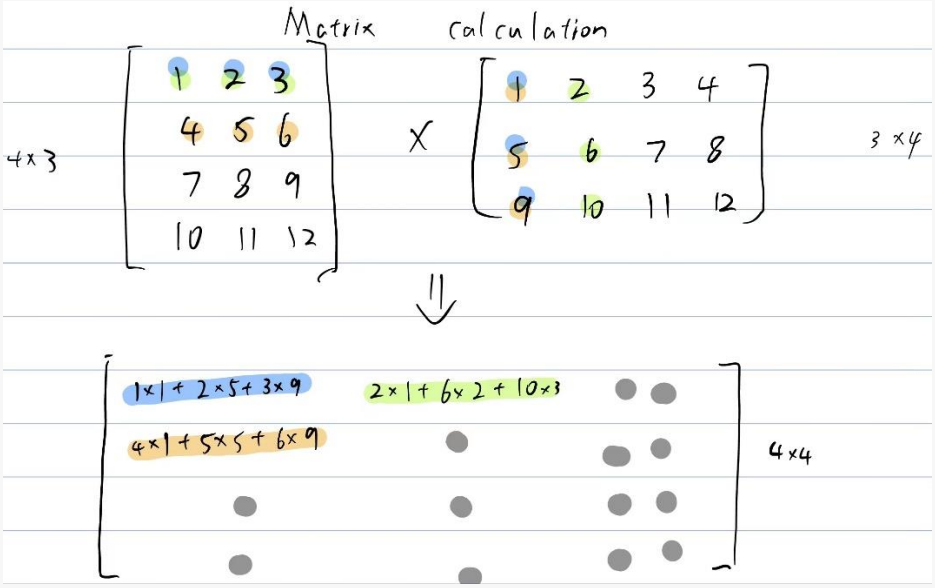
An overview of the Chinese version [7]

Matrix calculation

矩阵相乘的准则是：对于两个矩阵 A 和 B，如果 A 的列数等于 B 的行数，那么它们可以相乘，结果矩阵 C 的大小为 A 的行数乘以 B 的列数。具体地，如果 A 是一个 m 行 n 列的矩阵，B 是一个 n 行 p 列的矩阵，那么它们的乘积 C 是一个 m 行 p 列的矩阵，且 C 中的每个元素 c(i,j)等于 A 的第 i 行与 B 的第 j 列的乘积之和（长对长）。

例如，对于一个 4 行 3 列的矩阵 A 和一个 3 行 4 列的矩阵 B，它们可以相乘，结果矩阵 C 的大小为 4 行 4 列。矩阵 C 中的每个元素 c(i,j)都可以表示为 A 的第 i 行与 B 的第 j 列的乘积之和。

具体地，矩阵 A 和 B 的乘积可以表示为：

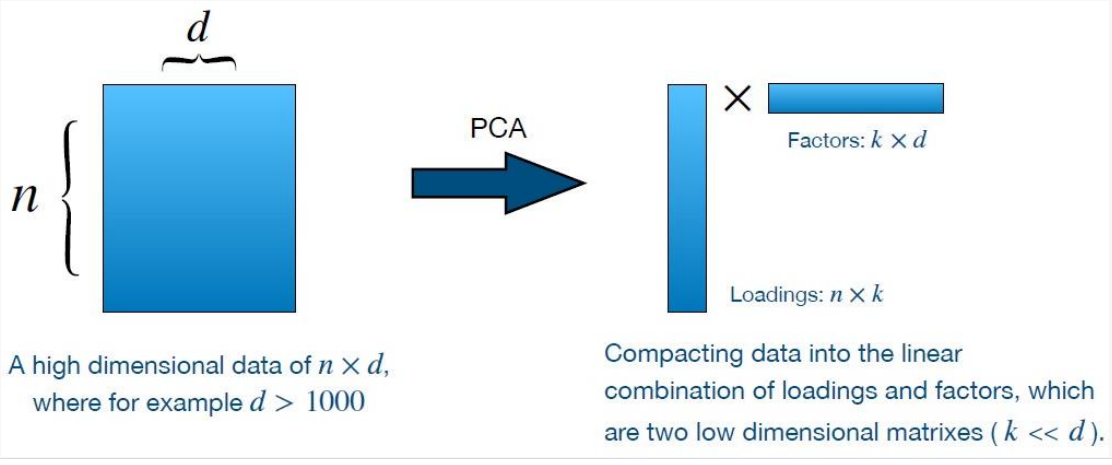


因此，矩阵 A 和 B 的乘积 C 是一个 4 行 4 列的矩阵，其中每个元素 c(i,j)都可以表示为 A 的第 i 行与 B 的第 j 列的乘积之和。

PCA in genomics

在基因表达谱分析中，PCA 可以用来寻找最具有代表性的基因表达模式，这些模式被称为主成分（principal components），每个主成分都是一组线性组合的基因表达值，它们可以描述整个数据集中的最大方差。当使用 PCA 对基因表达数据进行降维后，我们可以获得一个主成分矩阵，其中每个主成分都是一个向量，被称为 eigengene（特征基因），它代表了一组基因的联合表达模式。

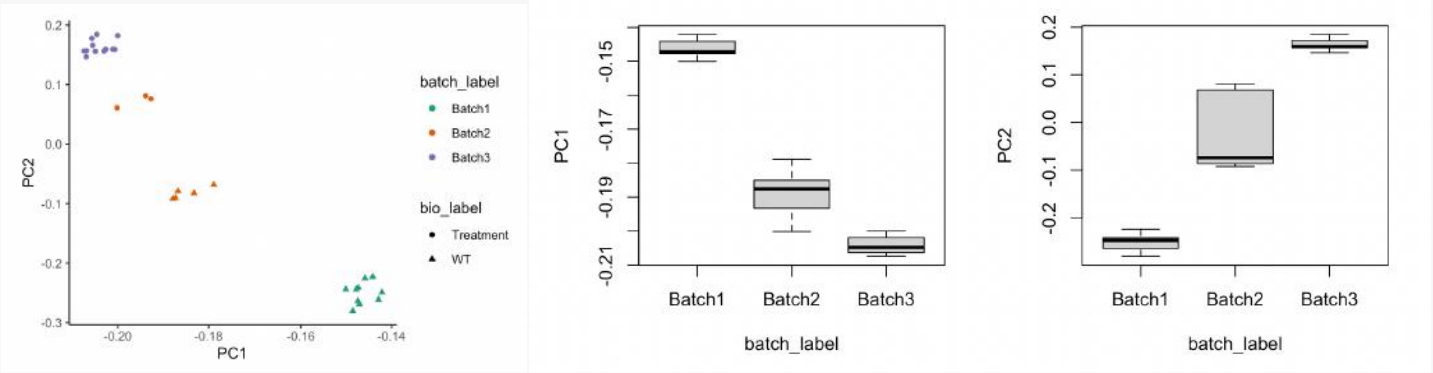
在基因表达谱分析中，eigengene 可以被用来表示一组基因的表达模式，通过计算不同条件下不同样本的 eigengene 表达值，可以揭示基因表达模式在不同生理和病理状态下的变化。eigengene 还可以被用来鉴定不同基因表达模式之间的相似性和差异性，帮助研究人员发现基因之间的相互作用和调控关系，以及揭示不同基因之间的生物学功能。



Estimation and correction for the batch effect

批效应（batch effect）是指实验过程中由于多种因素引起的样本间差异，而不是真实的生物学变异所导致的数据偏差。批效应对于生物学实验结果的解释和应用可能会带来严重的影响，因此需要对其进行估计和修正。使用特征基因（eigengene）进行批效应的估计与修正是一种常用的方法，下面是具体的步骤：

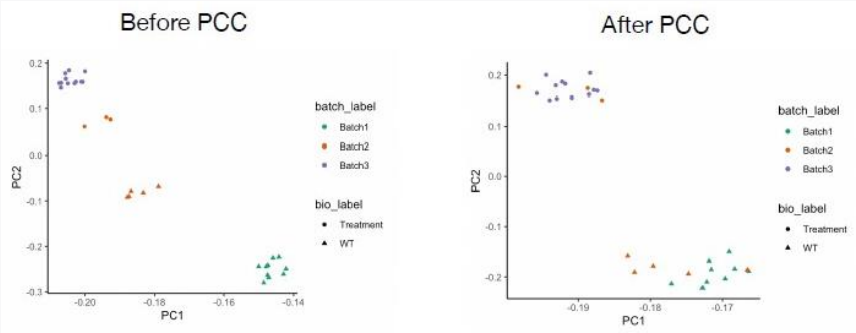
1. 通过 PCA 等方法，计算数据集的主成分矩阵，获得每个主成分对应的特征基因（eigengene）。
2. 将数据集中的所有样本按照批次（batch）进行分类，将不同批次的样本分别进行分组。
3. 对于每个批次，计算该批次中所有样本的 eigengene 的均值和方差。
4. 将每个批次的 eigengene 的均值和方差作为参数，建立线性回归模型，对每个样本进行批效应的估计和修正。
5. 通过比较估计和修正前后数据的分布和差异，评估批效应的影响，并确定是否需要进一步的数据修正。



在异构数据集中，顶部特征基因通常是批因子。需要注意的是，批效应的估计和修正可能会影响数据的统计分析和结果解释，因此在使用特征基因进行批效应的估计和修正时，需要根据具体的实验设计和研究问题，选择合适的方法和参数，并进行适当的验证和比较。

principal component correction (PCC)

Principal Component Correction（PCC）是一种常用的批次效应校正方法，它基于主成分分析（PCA）的思想，通过寻找样本数据中的主要方差方向，并利用这些方向来控制批次效应。



PCC 的基本步骤如下：

1. 将数据集按批次进行分类，每个批次内部作为一个子集，每个子集可以由不同的样本组成，但同一子集内的样本应该具有相似的生物特性。
2. 对于每个子集，利用 PCA 方法对数据进行降维处理，提取出前 k 个主成分（Principal Components）。
3. 对于每个子集，计算该子集的主成分平均值（Principal Component Mean，PCM），即该子集内所有样本在主成分空间中的平均向量。
4. 对于每个子集内的样本，将其表示为该子集的主成分和 PCM 的线性组合，即该样本在每个子集内的主成分得分（Principal Component Scores）。
5. 对于每个主成分，利用线性回归模型拟合出该主成分和批次之间的关系，即主成分得分和批次的线性回归系数（Regression Coefficient）。
6. 对于新的样本，首先利用 PCA 将其投影到主成分空间中，并计算其在每个子集内的主成分得分，然后利用对应的线性回归系数进行校正，得到纠正后的主成分得分。

7. 将纠正后的主成分得分重新转换到原始数据空间中，得到经过批次效应校正后的数据。

PCC 的优点在于可以在不需要其他信息的情况下，自适应地调整批次效应，并且可以同时处理多个批次。缺点在于，当样本之间的差异较大时，PCC 可能会引入不必要的噪声，因此需要谨慎使用。

PCA and tSNE/UMAP

PCA（Principal Component Analysis）、tSNE（t-distributed Stochastic Neighbor Embedding）和 UMAP（Uniform Manifold Approximation and Projection）都是降维方法。

PCA 是一种线性降维方法，通过线性变换将高维数据映射到低维空间中，使得新的特征空间中的数据方差最大化，从而保留原始数据的最大信息。PCA 在降维过程中比较简单，计算速度快，但是无法处理非线性数据和复杂的数据结构。

tSNE 和 UMAP 都是非线性降维方法，可以处理非线性数据和复杂的数据结构。它们的主要区别在于算法的原理和结果可视化方式。

tSNE 是一种基于概率模型的降维方法，它通过对高维空间中的数据点之间的相似性关系建模，使用 t 分布来近似高维数据的相似度，将高维数据映射到二维或三维空间中，保留数据的局部结构和相似性。tSNE 的主要优点是保留了原始数据的局部结构，适用于可视化高维数据集的局部结构。

UMAP 是一种基于图论的降维方法，它通过构建高维空间中的局部邻域图，将高维数据映射到低维空间中，并保留数据的全局和局部结构。UMAP 的主要优点是具有更好的可扩展性，适用于大规模数据集的降维和可视化。

对于优缺点，PCA 的优点是简单易实现、计算速度快，缺点是无法处理非线性数据和复杂的数据结构。tSNE 的优点是保留了原始数据的局部结构，适用于可视化高维数据集的局部结构，缺点是计算时间较长、对参数敏感，可解释性差。UMAP 的优点是具有更好的可扩展性，适用于大规模数据集的降维和可视化，缺点是在某些情况下可能会失去一些局部结构。

Method	Principle	Advantage	Disadvantage
PCA	Finding low dimensional projections that spread data as much as possible.	High interpretability as factor analysis	Work less well for non-linear patterns
tSNE / UMAP	Non-linear embedding that keep close-by points close using a probabilistic objective.	Can learn complex non-linear relationships	Axes have no meanings

总的来说，选择哪种降维方法取决于具体的问题和数据集。如果数据集是线性的，则可以使用 PCA；如果数据集是非线性的，则可以使用 tSNE 或 UMAP。如果需要保留原始数据的全局结构，则可以选择 UMAP；如果需要保留原始数据的局部结构，则可以选择 tSNE。

Performance of dimensional reduction methods in scRNA-seq