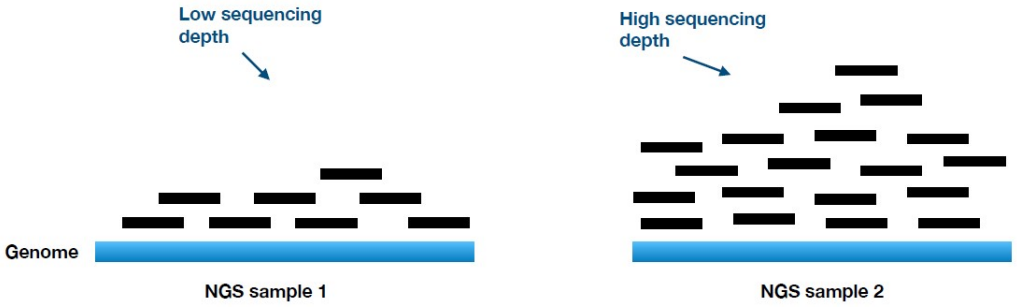
An overview of the Chinese version [5]

Sequencing depth

Sequencing depth 指的是在 DNA 或 RNA 测序实验中，每个碱基被测序的次数或测序覆盖度。通常用"X"来表示测序深度，表示平均每个碱基被测序的次数。例如，一个测序深度为 30X 的样本意味着每个碱基被测序的平均次数为 30 次。

测序深度是影响测序数据质量和可靠性的一个重要因素。深度越高，数据的信噪比越高，检测到的低频突变和 SNP 的准确性也更高。同时，测序深度也影响到检测到的变异的类型和数量。较低的测序深度可能会导致错过一些低频突变或未能检测到某些罕见的基因变异，而较高的测序深度则可能会增加检测到假阳性变异的概率。

因此，在进行测序实验时，需要根据研究的目的、样本特性和预算等因素进行测序深度的选择。例如，对于疾病的临床诊断，需要选择更高的测序深度来确保高准确性和可靠性，而对于大规模基因组学研究，则可以适当降低测序深度以节省成本。



- 样品中的初始细胞数
用不同数量的起始细胞构建 NGS 库。
- PCR 扩增的有效性
PCR 温度和周期的变化会影响片段的扩增率。
- NGS 平台
不同的测序通道和平台，其片段检测率也不同。

~estimate

可以通过从读数计算来估算测序深度。一般而言，测序深度可以通过以下公式计算：

Sequencing depth = total number of reads / size of genome (or targeted regions)

其中，"total number of reads"指的是测序获得的总 reads 数，而"size of genome (or targeted regions)"指的是参考基因组的大小或目标区域的大小（如果使用了目标测序方法）。

例如，对于一个包含 100 万个 reads 的样本，参考基因组大小为 3 亿个碱基，则该样本的测序深度为：

Sequencing depth = 1,000,000 reads / 300,000,000 bp = 0.0033X

在实际应用中，可以根据研究需要和测序目的来选择适当的测序深度。对于不同的测序方法和数据类型，也可以使用不同的计算方法来估算测序深度。例如，在 RNA-Seq 分析中，可以将 reads 数除以表达基因的数量来估算测序深度，而在 ChIP-Seq 分析中，可以将 reads 数除以测序覆盖的基因组区域的大小来估算测序深度。

Effect of feature length

在基因表达分析中，特征长度（Feature length）指的是 RNA 分子的长度，即在基因表达实验中测量的转录本或基因的长度。特征长度对基因表达分析结果有很大的影响，因为它会影响到对基因表达量的计算和比较。

一方面，特征长度越长，基因表达量就越小。这是因为在计算基因表达量时，使用的是 reads 数除以特征长度，所以分母越大，比率就越小，计算得到的基因表达量也就越小。因此，如果在分析中将较长的特征与较短的特征进行比较，可能会导致误差和不准确性。

另一方面，特征长度还会影响到基因表达的比较。如果不同样本的特征长度不同，那么就很难比较它们的基因表达水平。为了解决这个问题，常常会将基因表达量标准化为每个单位长度（通常为百万个碱基对），从而消除特征长度对基因表达量的影响，使得不同特征之间可以进行可靠的比较。

在单细胞 RNA 测序中，由于单个细胞的 RNA 数量很少，通常会使用基因表达量的 TPM（Transcripts Per Million）或 RPKM（Reads Per Kilobase Million）来计算基因表达量，以消除特征长度的影响。这种方法可以将每个特征的长度标准化为 1 kb 或 1 百万个 RNA 分子，从而使不同长度的特征之间可以进行可靠的比较。

综上所述，特征长度对基因表达分析结果具有重要影响，需要在数据处理和分析过程中进行适当的标准化和控制。

RPKM, FPKM, TPM

RPKM is viewing RNA-Seq experiment as a pool of dice rolls [repeat]

RPKM（Reads Per Kilobase Million）是 RNA-Seq 中最早使用的方法之一，它将每个基因的表达量除以该基因的长度（以千碱基为单位）和测序数据中的总 reads 数（以百万为单位），从而得到一个标准化的基因表达量。RPKM 的优点是可以通过比较不同基因在不同样本中的表达量来推断基因的相对表达水平，但是它假定了测序数据是完全随机的，并且不能处理基因转录本的多样性。

FPKM（Fragments Per Kilobase Million）是基于 RPKM 的一种改进方法，它考虑到了测序数据中的 fragment（片段），而不仅仅是 reads，从而更准确地计算了基因表达量。FPKM 还能够处理基因转录本的多样性，从而更准确地计算基因表达量。然而，与 RPKM 类似，FPKM 也无法对测序数据的偏差进行修正。

TPM（Transcripts Per Million）是一种相对较新的基因表达量计算方法，它通过将每个基因的表达量除以它在样本中所有基因的表达量之和，从而得到一个标准化的基因表达量。TPM 考虑了基因转录本的多样性和测序深度的变化，而且可以用于不同样本之间的比较。因此，TPM 是当前最常用的基因表达量计算方法之一。

应用场景方面，RPKM 和 FPKM 在一些早期的研究中经常被使用，特别是在对转录本水平的研究中。然而，由于它们不能处理多样性和测序深度偏差的问题，以及无法进行样本间比较，因此在单细胞 RNA 测序分析和基因表达谱分析中，TPM 是目前最常用的基因表达量计算方法。但是，具体使用哪种方法仍需根据研究设计和实验目的进行选择。

1. **RPKM** (reads per kilobase of transcript per million reads mapped)

$$\text{RPKM} = \frac{\text{Read Count}}{\text{Gene length} \times \sum_{\forall genes} \text{Read Count}} \times 10^9$$

2. **FPKM** (Fragments per kilobase of transcript per million reads mapped)

$$\text{FPKM} = \frac{\text{Fragment Count}}{\text{Gene length} \times \sum_{\forall genes} \text{Fragment Count}} \times 10^9$$

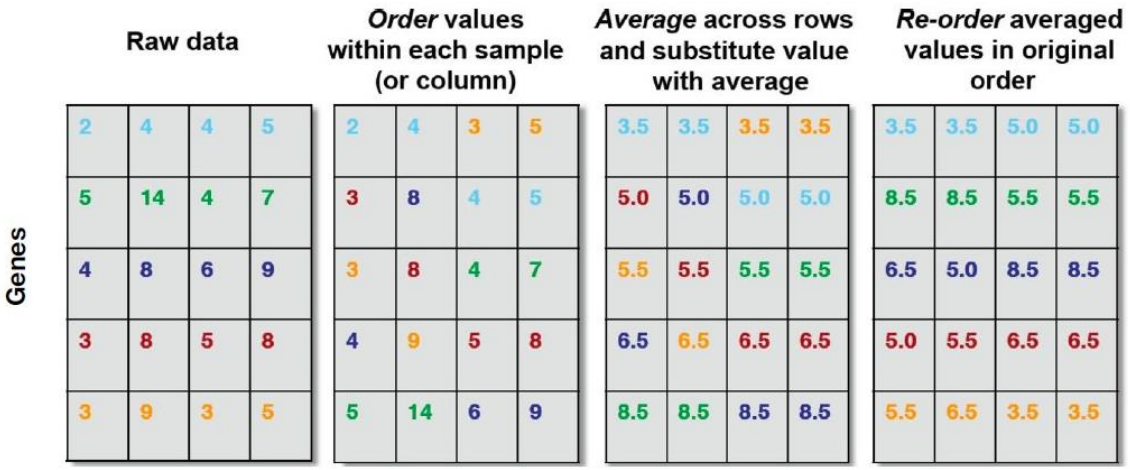
Sum over all genes within a sample

3. **TPM** (Transcripts per million)

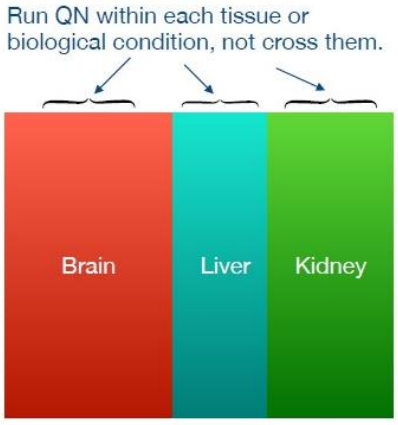
$$\text{TPM} = \frac{\text{Read Count}}{\text{Gene length} \times \sum_{\forall genes} (\text{Read Count}/\text{Gene length})} \times 10^6$$

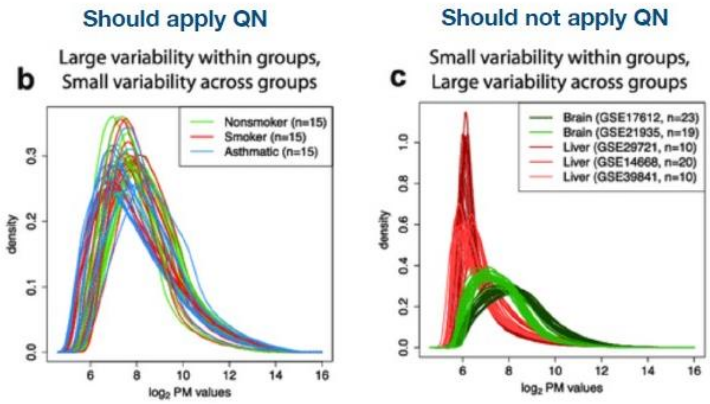
Sequencing depth estimated on the length normalized count, ensuring sample wised sum of TPM = constant

Quantile normalization



- 量子归一化（QN）可以在任何测序样本中强制执行相同的分布。
- QN 的步骤。1.对列（样本）值进行排序。2.用行（基因）平均数代替数值。3. 返回到原来的顺序。
- 该程序可以有效地消除基因组数据的批次效应。





在不同的生物群体中进行 QN 可能会扭曲有意义的生物信号。因此，最好在主要的生物条件下（如组织和细胞类型）进行 QN。

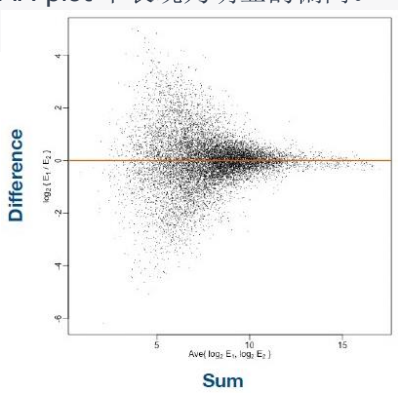
MA-plot

MA-plot 是一种常用的数据可视化方法，用于比较两个不同样本的基因表达水平差异。它是由基因表达水平的平均值（M）和差异值（A）所构成的散点图。

在 MA-plot 中，每个点代表一个基因在两个不同样本中的表达水平的差异。其中，M 表示两个样本中该基因的表达水平之间的对数比值，即 M=log2(sample1/sample2)。A 表示两个样本中该基因的表达水平的平均值的对数，即 A=(log2(sample1) + log2(sample2))/2。因此，MA-plot 的横坐标是平均表达水平的对数值，纵坐标是表达水平的对数比值，散点图上每个点代表一个基因。

MA-plot 可以用于检查样本之间的重现性，即不同样本之间的差异是否与总表达水平有关。如果样本之间的重现性良好，MA-plot 中的散点应该分布在水平线上，并且没有明显的趋势。反之，如果样本之间的重现性较差，则 MA-plot 中的散点会出现明显的倾斜或曲线趋势。此外，如果样本中存在一些异常值或离群点，它们可能会在 MA-plot 中表现为明显的偏离。

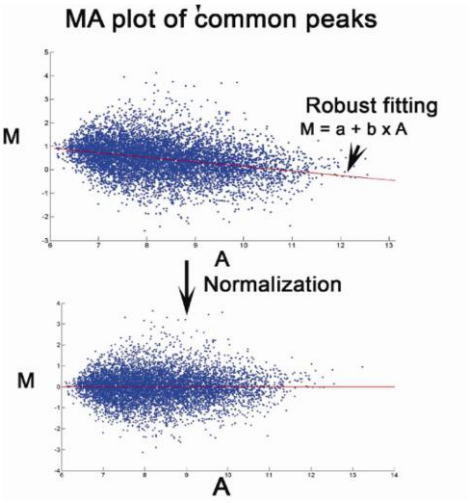
为了检查样本之间的重现性，可以绘制两个样本之间的 MA-plot，并比较它们的分布是否相似。如果两个样本之间的 MA-plot 分布相似，则说明它们之间的差异不是由总表达水平所引起的，这种差异可能是生物学上的真实差异。反之，如果两个样本之间的 MA-plot 分布不同，则可能是由于技术变异或批次效应等因素引起的，而不是生物学上的真实差异。因此，在研究分析中使用 MA-plot 是一种有用的方法，可以帮助鉴别生物学上的差异和技术上的差异。



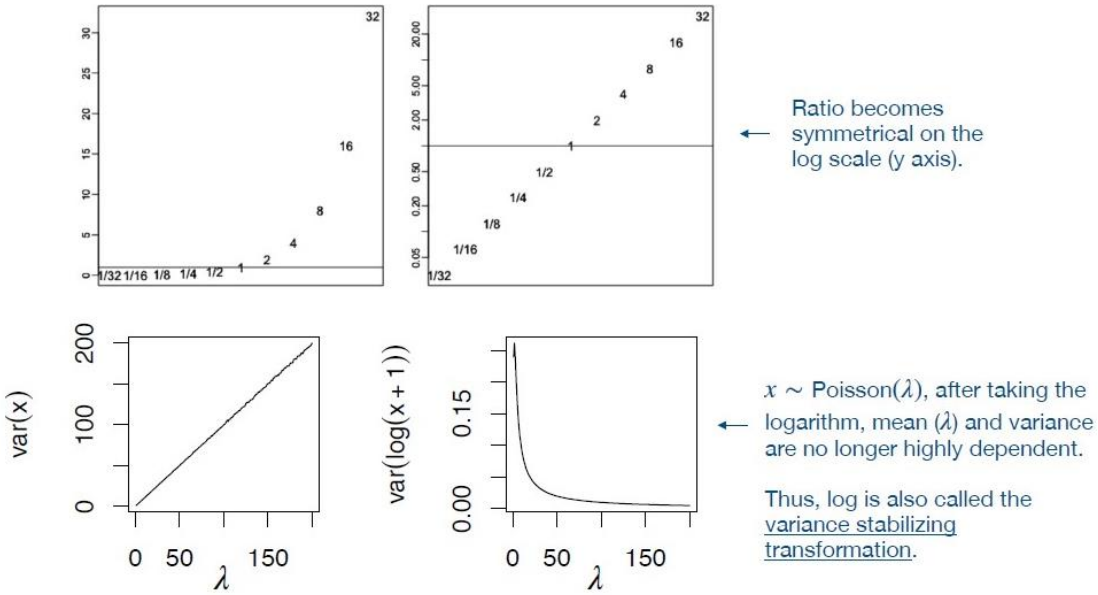
MA-normalization

人们可以通过 MA 归一化来校正基因组学数据。

1. 选择一个参考样本，通常通过基因的平均数来计算。
2. 通过与参考样本的比较为每个样本生成一个 MA 图，并对每个图进行线性回归。
3. 通过减去拟合值使每个样本正常化，以考虑到从预期的水平线通过原点的偏差。



Log transformation



- **Count** and **ratio** data types are often beneficial from log transformation.
- log(count + 1) and log fold changes are commonly used in genomic data visualization and data analysis.
- log is also a mathematically natural transformation for ratio and count.