An overview of the Chinese version [3]

Systematic error and Batch effect in NGS data

在 NGS（Next-Generation Sequencing）数据分析中，systematic error 和 batch effect 是两个常见的问题。它们都可能导致数据的偏差，影响后续分析的准确性和可靠性。

Systematic error 是指在实验过程中由于某些因素（如 PCR 扩增、文库制备、测序仪器等）导致数据出现常态偏差。这些因素可能会对所有样本产生相同的影响，从而导致整个数据集的偏差。Systematic error 可以通过一些校正方法（如 BQSR）来消除或降低。

Batch effect 是指在实验过程中，将样本分为多个批次进行处理，导致不同批次之间的数据出现偏差。这些偏差可能来自于不同批次的实验条件、仪器、处理过程等因素，从而导致样本之间的比较和分析出现偏差。Batch effect 可以通过一些校正方法（如 ComBat）来消除或降低。

需要注意的是，Systematic error 和 Batch effect 都可能导致数据的偏差，但它们的原因和解决方法不同。因此，在数据分析过程中需要仔细分析和区分，并选择合适的校正方法进行处理，以保证数据的准确性和可靠性。

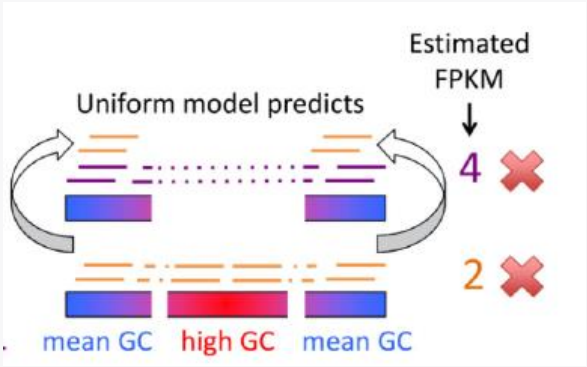
Fragment GC content bias

Fragment GC content bias 是指在测序过程中，由于 PCR 扩增和文库制备等步骤的影响，导致不同片段（Fragment）的 GC 含量分布不均衡，从而对测序结果产生偏差的现象。

PCR 扩增和文库制备过程中的反应条件和化学物质浓度等因素可能会影响到片段的 GC 含量分布。例如，如果 PCR 扩增反应的温度和时间不够准确，或者反应中的 dNTPs、酶、模板 DNA 等成分不够均匀，就可能导致片段的 GC 含量分布出现偏差。类似地，如果文库制备过程中使用的酶或试剂批次不同，也可能导致片段的 GC 含量分布出现偏差。

这种片段 GC 含量分布的不均衡，会对不同测序平台、不同测序化学体系和不同文库制备方法等产生不同的影响。例如，在 Illumina 测序中，高 GC 含量的片段可能会导致簇密度降低、偏移等问题，而低 GC 含量的片段可能会导致信噪比下降、错误率升高等问题。

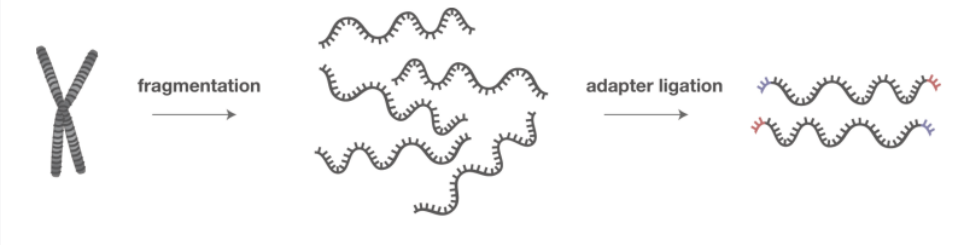
为了减少 Fragment GC content bias 的影响，可以采用一些措施来优化 PCR 扩增和文库制备过程，例如：控制反应条件、使用均匀的酶和试剂、采用多次 PCR 扩增等。此外，也可以采用一些数据处理方法，如基于 k-mer 的校正方法，对不同 GC 含量的片段进行校正，从而减少偏差的影响。



Adaptor contamination

Adaptor contamination 指的是在文库制备过程中，由于引物（adaptor）没有被完全去除，导致在测序数据中出现未知的序列片段（contaminant reads）。这些未知的序列片段可能是引物本身，也可能是引物与文库中插入片段的连接序列（junction）。

Adaptor contamination 可能会导致测序数据中出现偏差和误解析的情况。例如，如果引物序列和待测序列存在相同的部分，那么这些部分可能会被错误地解析为待测序列，从而导致误解析。此外，未知的引物序列还可能会干扰序列比对和变异分析等后续分析过程，从而影响数据的质量和可靠性。



为了减少 Adaptor contamination 的影响，可以采取以下措施：

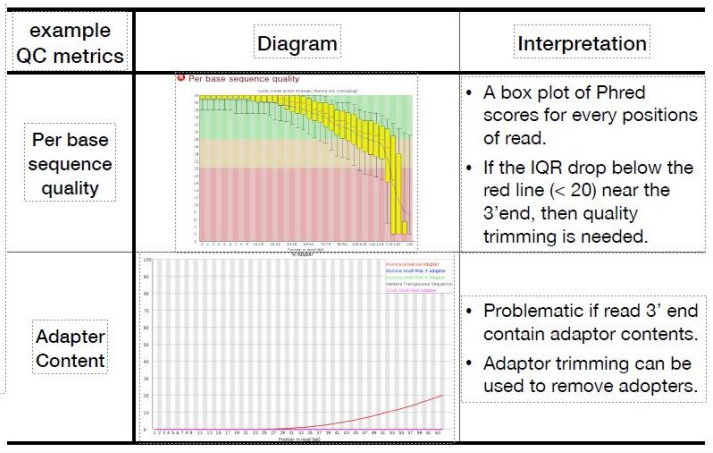
1. 使用高质量的引物和试剂，避免引物污染和批次变化的影响。
2. 对文库进行充分的纯化和质检，确保文库中不存在未知的引物和其他污染物。
3. 在数据预处理过程中，使用一些工具和软件，如 Trimmomatic、Cutadapt 等，对序列数据进行质量控制和去除引物等处理，以降低 Adaptor contamination 的影响。
4. 在序列比对和变异分析等后续分析过程中，考虑到 Adaptor contamination 的可能性，采用严格的数据过滤和质量控制策略，以保证数据的准确性和可靠性。

FastQC

FastQC 是一款用于质量控制的软件，可以用于评估测序数据的质量和检测常见的质量问题。它可以接受 FASTQ 格式的测序数据，并生成一个 HTML 报告，其中包含各种图表和统计信息，用于评估数据的质量和检测常见的质量问题。

FastQC 的主要功能包括：

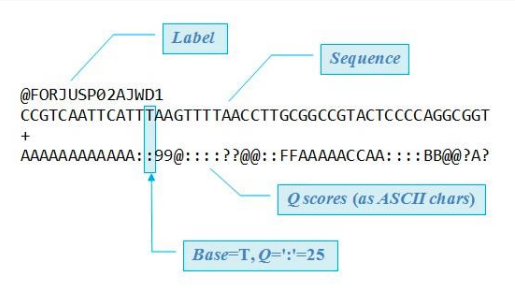
1. 测序数据质量评估：FastQC 会计算每个碱基的质量值，并生成一个质量分布图表，帮助用户评估数据的质量。它还可以检测测序数据中存在的一些问题，如碱基偏差、过度重叠和低复杂度序列等。
2. Adapter 污染检测：FastQC 可以检测到文库制备过程中引物（adaptor）残留和污染的情况，提示用户可能需要去除这些污染物。
3. 序列长度分布：FastQC 可以统计序列长度分布，帮助用户检测到存在异常的序列长度分布情况，如过度短或过长的序列。
4. 过度重复的序列：FastQC 可以检测出过度重复的序列，帮助用户识别可能存在的 PCR 扩增偏差或文库制备问题。



使用 FastQC 可以很方便地对测序数据进行质量评估和质量控制。通过对 FastQC 报告的分析和解释，可以快速识别和解决可能存在的问题，从而提高数据质量和分析结果的可靠性。

Fastq format

FASTQ 格式是常用于存储测序数据的一种文本格式，它包含两部分信息：序列（sequence）和质量（quality）。FASTQ 格式由四行组成，每个序列的信息包含在四行中，格式如下：



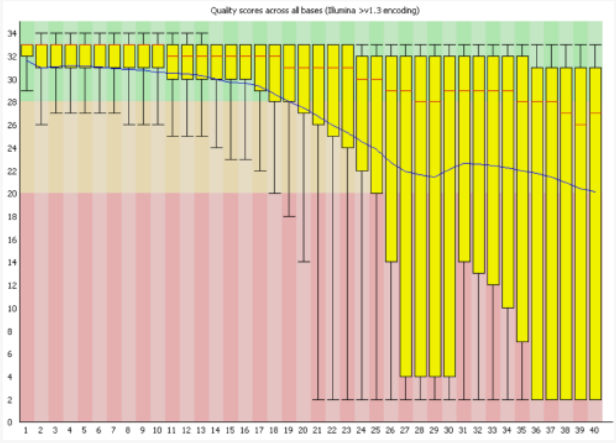
1. 第一行以“@”字符开头，后跟序列的唯一标识符，通常由仪器自动生成。
2. 第二行是序列的实际序列，通常为 ATCG 碱基。
3. 第三行以“+”字符开头，后跟与序列对应的质量值标识符，通常与第一行的标识符相同。Q = − 10 × log10(e)
4. 第四行是与序列对应的质量值，通常是 ASCII 码，用于表示测序的可靠性。

Per base sequence quality

Per base sequence quality 是指测序数据中每个碱基的质量值分布情况。对于 FASTQ 格式的测序数据，每个序列都伴随着一个对应的质量值序列，其中的每个质量值代表了对应碱基的测序可靠性。

Per base sequence quality 分析可以帮助我们评估每个碱基的可靠性，判断测序数据的质量是否符合要求。在这个分析中，通常会绘制一条曲线来表示质量随着碱基位置的变化而变化的情况。如果质量曲线下降得很快，那么就意味着在这个位置测序出现了问题，可能需要进行修剪或过滤。

Per base sequence quality 分析可以通过 FastQC 等软件进行，FastQC 可以绘制出质量曲线，并根据曲线的情况提供一些解释和建议。例如，如果在序列的末端存在较差的质量值，那么就可能需要对序列进行修剪；如果质量值过低，那么就可能需要将这些序列过滤掉，以保证后续分析的准确性。



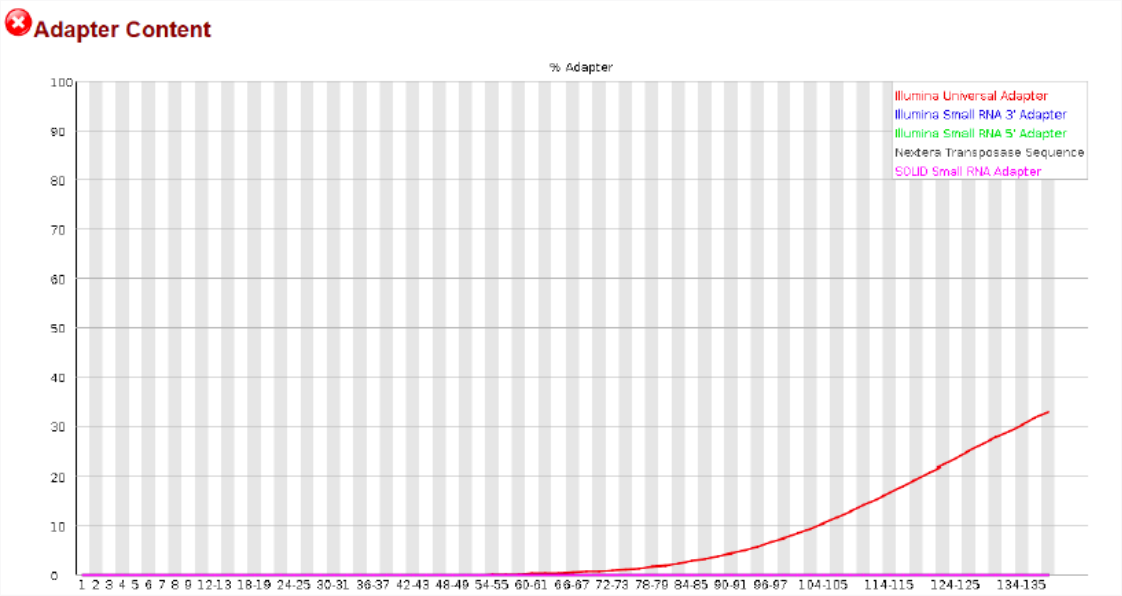
- The y-axis on the graph shows the Phred scores.
- The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red).
- Warning will be issued if the lower quartile for any bases fall below the red region.

总之，Per base sequence quality 分析可以帮助我们评估测序数据的质量，为后续的数据处理和分析提供重要的参考。

Adaptor content

Adaptor content 分析可以帮助我们评估测序数据中适配器的含量，并确定是否需要对数据进行适当的处理。例如，在 RNA-seq 分析中，高含量的适配器序列可能会影响基因表达水平的估计，因此需要对测序数据进行去除或修剪。同样，在 DNA 测序中，高含量的适配器序列也可能会影响变异位点的检测，因此需要对测序数据进行适当的处理。

Adaptor content 分析可以通过 FastQC 等软件进行。FastQC 会检测测序数据中是否存在适配器序列，并计算适配器序列的含量。如果适配器序列含量过高，那么就需要对测序数据进行相应的处理，例如去除或修剪适配器序列，以保证后续分析的准确性。



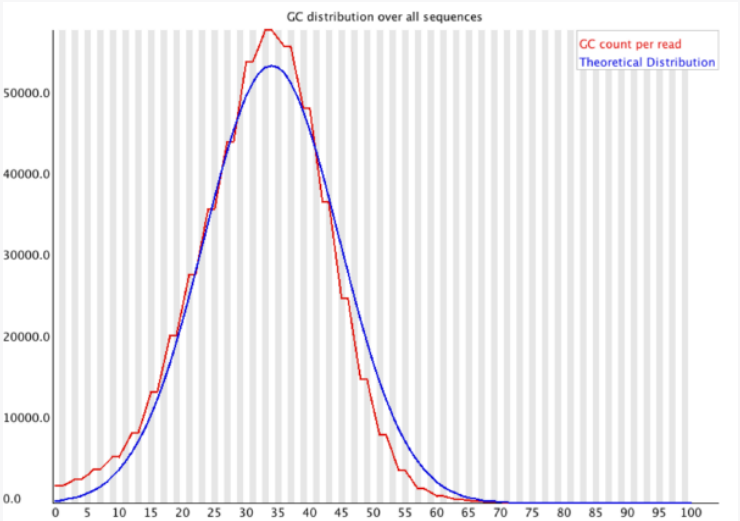
- The plot shows a cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position.

- This module will issue a warning if any sequence is presented in more than 5% of all reads.

GC content distribution

GC content distribution 是指测序样品中所有 DNA 片段的 GC 含量的分布情况。在 DNA 测序中，GC 含量对测序数据的质量和可靠性有很大影响。因此，对 GC content distribution 的分析可以帮助我们评估测序样品的质量和偏差情况。

GC content distribution 的分析通常是通过绘制直方图来进行的，其中 X 轴表示 GC 含量，Y 轴表示每个 GC 含量区间中包含的 DNA 片段数。如果 GC content distribution 存在明显的峰值或不均匀的分布情况，那么就可能会影响后续数据分析的准确性。例如，在 RNA-seq 分析中，如果测序样品中存在明显的 GC 偏差，那么就需要进行相应的纠正，以保证基因表达水平的准确性。GC content distribution 的分析可以通过 FastQC 等软件进行。FastQC 可以计算出样品中所有 DNA 片段的 GC 含量，并绘制出 GC content distribution 的直方图。如果 GC content distribution 存在异常情况，那么 FastQC 会提供相应的解释和建议，帮助我们对测序数据进行相应的处理。



- The graph displayed a histogram of GC content over all reads.

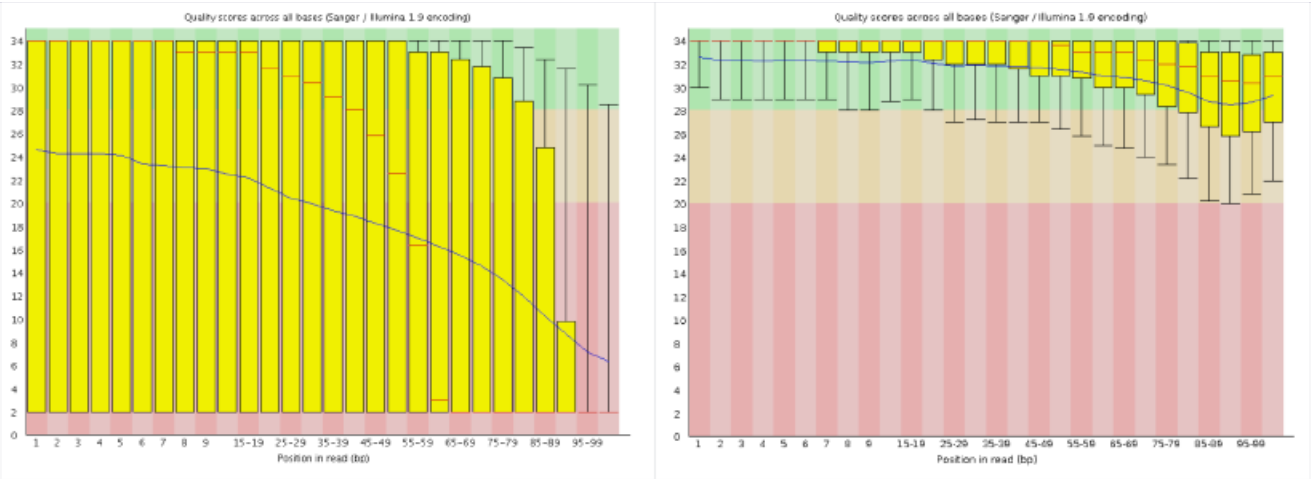
- Warning is issued when observed read GC content distribution (red) is significantly deviant from the expected normal distribution (blue).

Trim Galore

Trim Galore 是一个用于高通量测序数据的质量控制和处理的软件工具，可以帮助用户去除测序数据中的低质量序列和适配器序列等，从而提高测序数据的质量和可靠性。

Trim Galore 的主要功能包括以下几个方面：

1. 去除低质量序列：Trim Galore 可以根据用户设定的阈值去除测序数据中质量较低的序列，从而提高数据的质量。
2. 去除适配器序列：Trim Galore 可以识别测序数据中的适配器序列，并将其去除，以避免适配器序列的干扰。
3. 质量控制统计：Trim Galore 可以生成有关测序数据的各种统计信息，例如序列长度分布、GC 含量分布、N 含量分布等，以帮助用户评估测序数据的质量和偏差情况。
4. 并行处理：Trim Galore 支持多线程处理，可以快速处理大量的测序数据。



Before & After

Trim Galore 支持多种测序数据格式，包括 FASTQ、BAM 和 CRAM 等。同时，Trim Galore 还支持多种测序平台，包括 Illumina、SOLiD、454 和 Ion Torrent 等。

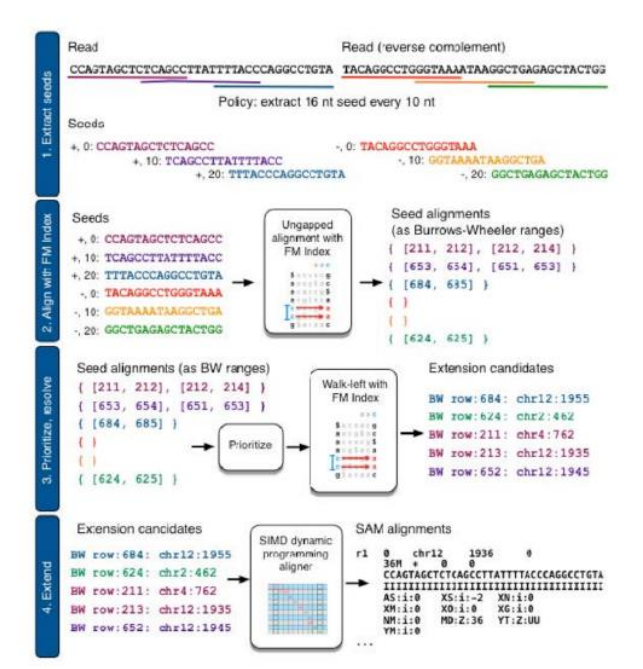
Bowtie2

Bowtie2 是一种快速的序列比对工具，它是 Bowtie 的改进版，可以用于比对高通量测序数据。Bowtie2 支持比对单端和双端测序数据，同时还支持多个测序数据的同时比对，可以有效地提高比对效率和准确性。

Bowtie2 的主要特点包括以下几个方面：

1. 高效率：Bowtie2 使用一种叫做“比对索引”的数据结构来组织参考基因组，可以快速地比对大规模的测序数据。
2. 高准确性：Bowtie2 使用一种先进的比对算法，可以对含有多个错配或缺失的序列进行准确比对。
3. 灵活性：Bowtie2 支持多种比对模式，包括全局比对、局部比对和端对端比对，可以根据不同的数据类型和比对任务进行选择。
4. 可视化：Bowtie2 可以生成详细的比对报告，包括比对结果的统计信息和可视化图表等。

Bowtie2 广泛应用于基因组、转录组和芯片测序等领域，可以用于各种测序平台和数据格式的比对任务。同时，Bowtie2 还有许多辅助工具和插件，可以进一步优化比对结果和性能。



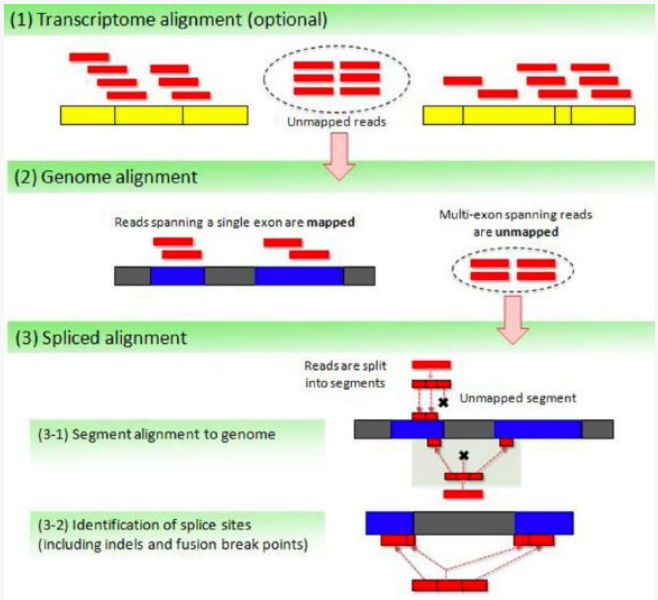
Tophat2 pipeline

Tophat2 是一个用于分析 RNA 测序数据的软件工具，可以将 RNA 测序数据与参考基因组进行比对，并生成基因转录本的注释信息。Tophat2 一般被用作 RNA 测序数据预处理的第一步，它会将原始测序数据进行清洗、修剪、去除低质量序列等处理，并将清洗后的数据比对到参考基因组上，最终生成比对结果。

Tophat2 的主要特点包括以下几个方面：

1. 支持多种比对模式：Tophat2 可以进行全局比对、局部比对和跨越内含子的比对，可以根据不同的比对任务和数据类型进行选择。
2. 高效率和准确性：Tophat2 使用了一种称为“splice junctions”的数据结构，可以快速地和准确地比对含有内含子的转录本。
3. 集成 Bowtie2：Tophat2 使用 Bowtie2 作为其比对引擎，可以有效地提高比对效率和准确性。
4. 可视化和注释：Tophat2 可以生成详细的比对报告，包括比对结果的可视化图表和基因转录本的注释信息等。

Tophat2 的工作流程一般包括以下几个步骤：



1. 读入原始测序数据，进行质量控制和过滤，得到高质量的序列数据。

2. 利用 Bowtie2 将序列比对到参考基因组上，得到比对结果。
3. 利用 Tophat2 将比对结果进行进一步处理，去除错误比对和重复比对等。
4. 利用 Cufflinks 对比对结果进行拼接，生成基因转录本的注释信息。
5. 可选地，可以使用其他工具对基因转录本进行定量和差异表达分析等。

Kallisto

Kallisto 是一种用于快速和准确进行 RNA 测序定量的软件工具。与传统的 RNA-seq 定量方法不同，Kallisto 使用了一种名为“pseudomapping”的技术，将 RNA 序列比对到参考转录本组上，而非参考基因组。这种方法不仅大大降低了计算成本，同时也保证了定量结果的高精度。

Kallisto 的主要特点包括以下几个方面：

1. 快速和准确：Kallisto 的计算速度非常快，可以在几分钟内处理数百万条 reads。同时，其定量结果的精度也非常高，与传统的比对方法相当甚至更好。
2. 采用 pseudomapping 技术：Kallisto 的 pseudomapping 技术可以在不进行实际比对的情况下，将 RNA 序列映射到参考转录本组上，大大减少了计算成本。

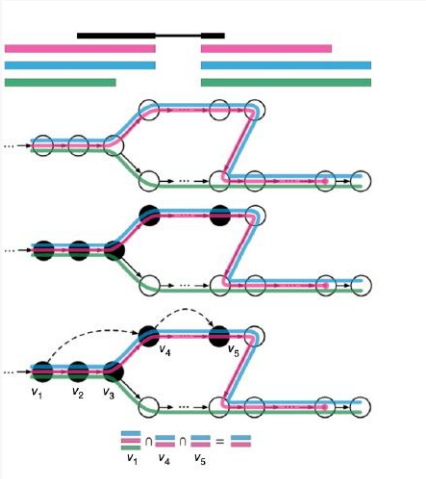
Alignment-free:

	Transcript 1	Transcript 2	Transcript 3
Read 1	1	0	0
Read 2	0	1	0
Read 3	0	0	1
Read 4	1	0	1

1: compatible
0: incompatible

3. 适用于多种测序技术：Kallisto 可以处理多种 RNA 测序数据，包括 Illumina、Ion Torrent 和 PacBio 等。
4. 可视化和交互式分析：Kallisto 的定量结果可以通过交互式图表和可视化工具进行展示和分析。

Kallisto 的工作流程一般包括以下几个步骤：



1. 读入 RNA 测序数据，进行质量控制和过滤，得到高质量的序列数据。
2. 利用 Kallisto 的 pseudomapping 技术将序列映射到参考转录本组上，得到每个转录本的表达量。
3. 可选地，可以使用其他工具对基因转录本进行定量和差异表达分析等。

Pseudo mapping

“Pseudomapping”是一种新兴的 RNA 测序（RNA-seq）分析技术，旨在通过将 RNA 测序数据与参考转录组进行快速匹配，以提高数据分析的速度和效率。以下是“pseudomapping”技术的工作原理和流程：

1. 创建索引：首先，对参考转录组进行索引构建，以便后续的快速比对。索引可以是基于 k-mer 的，也可以是基于 FM 索引的。

2. 生成伪映射（pseudomapping）：将 RNA 测序数据通过“pseudomapping”引擎与参考转录组进行匹配。这里的“pseudomapping”指的是一种快速匹配算法，它可以在不进行完全比对的情况下，根据 RNA 测序数据的特征将其分配到不同的基因或外显子中。这种方法的优点是速度快，可以处理更大的数据集。
3. 计算表达：根据伪映射的结果，计算每个基因或转录本的表达水平。这一步通常使用一些常见的表达计算方法，如 TPM（transcripts per million）或 FPKM（fragments per kilobase million）。

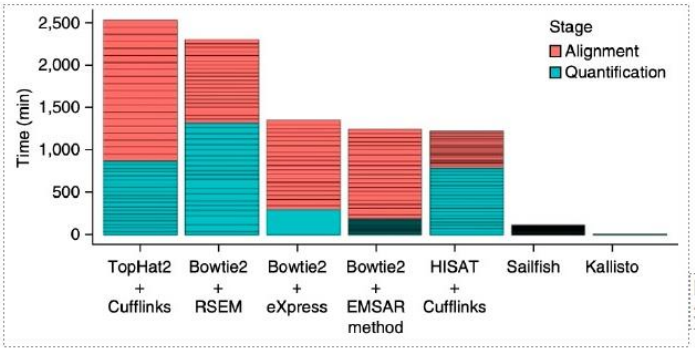
总的来说，“pseudomapping”技术通过将 RNA 测序数据快速匹配到参考转录组，避免了耗时的比对过程，并可以快速计算基因或转录本的表达水平。它比传统的基于比对的 RNA-seq 分析方法具有更高的速度和效率，尤其适用于大型 RNA-seq 数据集的处理。

什么是参考转录组？

参考转录组是一个已知物种或组织的转录组序列集合，用于进行 RNA 测序分析。通常情况下，参考转录组是由已知的基因组 DNA 序列转录而来，并且在转录组水平进行注释。这种注释包括识别外显子、内含子、UTR 和剪切变异等，因此参考转录组可以用于对 RNA 测序数据进行比对和分析，从而得到基因表达和转录本水平的定量信息。

在“pseudomapping”技术中，也需要使用参考转录组进行伪比对。因为“pseudomapping”是一种无需实际比对的方法，所以它并不需要比对到参考转录组的每一个碱基，只需要伪比对到转录本的一部分即可。因此，参考转录组的选择并不是特别严格，只需要保证包含了足够的转录本序列即可。通常情况下，参考转录组是由公共数据库如 NCBI、Ensembl 等提供的。

Running time of different methods



Nature biotechnology 34.5 (2016): 525-527.

Aligner	Splice-awared	Pesudo-alignment	Speed	Memory demand
bowtie2	No	No	Fast	Small
STAR	Yes	No	Fast	Large
Tophat2	Yes	No	Slow	Large
Hisat2	Yes	No	Fast	Small
Kallisto	Yes	Yes	Altra fast	Very small
Salmon	Yes	Yes	Altra fast	Very small

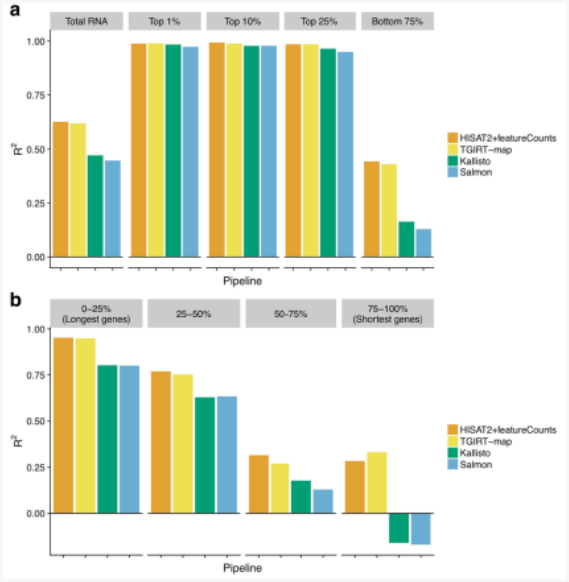
- For DNA-Seq based assays, *bowtie2* is recommended.
- For RNA-Seq based assays, *Hisat2* or *Tophat2* is recommended.

Why not only use alignment-free methods?

相较于传统的基于比对的方法，无比对的方法有多个优势，如速度更快、计算成本更低、对测序数据中的误差和变异性更不敏感等。此外，无比对的方法不需要参考基因组或已知的序列同源信息，因此更加灵活，适用范围更广。

但是，无比对的方法也有一些局限性。其中主要的一个是它们通常比基于比对的方法的准确性低。这是因为无比对的方法依赖于比较序列特征或模式，而不是直接比对序列，这可能导致准确性和灵敏度降低。此外，一些无比对的方法可能对某些序列特征或模式有偏差，导致分析的不准确性或假阳性结果。

因此，通常会将无比对的方法与基于比对的方法结合使用，以提供互补信息，提高序列分析的总体准确性和可靠性。总的来说，基于比对的方法仍被认为是许多类型序列分析的金标准，但无比对的方法已被证明是某些应用的有价值工具，并且对于分析大型和复杂的序列数据集特别有用。



* 无对齐方法和传统的基于对齐的量化方法对于常见的基因目标（如蛋白质编码基因）具有相似的性能。

* 然而，无对齐方法在分析和量化低表达的基因和小 RNA 方面有局限性，特别是当这些小 RNA 有生物变异时。

* 因此，由于其特征（bin）大小较小，在峰值调用中的滑动窗口不能使用无对齐方法进行可靠的量化。