第四次作业

邹翔宇 | 2410833001

1. Give a few examples of the biological function of DNA methylation. (10’)

DNA 甲基化是表观遗传学中的一个重要机制，其生物学功能主要包括：

1. **基因调控**：DNA 甲基化能够调控基因的表达。特别是在基因启动子区域的甲基化，可以抑制基因的表达。
2. **发育过程中的关键作用**：DNA 甲基化在生物体的发育过程中扮演着关键角色，它帮助细胞在分化过程中建立身份。
3. **遗传性**：DNA 甲基化能够在细胞分裂过程中遗传给子代细胞，从而维持基因表达的模式。
4. **环境影响**：DNA 甲基化可以受到环境因素的影响，是基因与环境相互作用（GxE 互作）的一个良好候选。
5. **胚胎发育中的表观遗传重编程**：在胚胎发育过程中，DNA 甲基化模式会发生重编程，这对于正常发育和健康至关重要。
6. **帮助细胞建立身份**：在细胞和组织分化过程中，DNA 甲基化有助于细胞建立和维持其特定的身份。

2. Briefly describe why the sequence alignment of BS-seq is more complicated. (10’)

1. **亚硫酸氢盐处理的影响**：亚硫酸氢盐处理会将未甲基化的胞嘧啶（C）转化为尿嘧啶，因此在比对时，需要允许读段中的胸腺嘧啶（T）与参考基因组中的 C 匹配，这增加了比对的复杂性。
2. **C 到 T 的转换**：在“in silico”亚硫酸氢盐转化策略中，需要将参考基因组中的所有 C 转化为 T，这导致了一个三字母的基因组，增加了比对的难度。
3. **比对算法的挑战**：不同的比对工具采用不同的策略，如 BS-Seeker、Bismark 等采用“in silico”亚硫酸氢盐转化策略，而 BSMAP 采用通配符方法，这些不同的方法都需要特别设计以处理亚硫酸氢盐转化后的数据。
4. **处理大规模数据集**：随着高通量测序技术的发展，产生的数据量急剧增加，对现有的比对工具提出了更高的要求，需要及时处理大量信息。
5. **比对效率和准确性**：BS-Seeker3 等工具通过优化比对流程，如改进索引/高吞吐量参考基因组处理、超快速比对和局部比对通过 Ukkonen 算法等，来提高比对的效率和准确性。
6. **后处理步骤的优化**：比对后的数据处理步骤占据了相当大比例的运行时间，需要对候选匹配之间的不精确匹配进行检查和重新计算不匹配数量，这些复杂的后处理步骤增加了比对的复杂性。
7. **计算资源的需求**：随着生物信息学实验室可用的计算资源的增长，BS-Seeker3 等工具现在可以将大的读文件分割成小文件，并利用服务器的高内存容量同时加载多个索引并行处理每个小文件，这要求比对工具能够适应不同的计算资源。

3. Briefly describe the goal of differential methylation analysis. (10’)

1. **识别差异甲基化位点（Differential Methylation Loci, DML）**：这些是甲基化水平在不同样本组之间存在统计学显著差异的单个 CpG 位点。
2. **识别差异甲基化区域（Differential Methylation Regions, DMR）**：这些是包含多个 DML 的区域，这些区域在不同样本组之间的甲基化水平存在显著差异，可能涉及较大的基因组区域。
3. **评估甲基化状态的变化**：通过比较不同样本或条件下的甲基化水平，差异甲基化分析有助于揭示甲基化状态的变化，这些变化可能与生物学功能、疾病状态或环境因素有关。
4. **提供生物学意义的见解**：差异甲基化分析的结果可以提供关于基因表达调控、细胞身份和疾病机制等方面的见解，因为甲基化能调控基因的表达。
5. **统计测试和模型应用**：使用各种统计方法和模型（如 Fisher 精确检验、t 检验、贝叶斯层次模型等）来评估甲基化水平的差异是否具有统计学意义，并控制假阳性率。

4. Briefly describe the data from BS-seq. (10’)

BS-seq（Bisulfite sequencing，即亚硫酸氢盐测序）产生的数据包含了每个检测位置的总读段数以及甲基化的读段数。例如，在某个特定的胞嘧啶位置，会记录覆盖该位置的所有测序读段，以及其中显示为甲基化的读段数量。这样的数据格式允许对单个碱基水平上的甲基化状态进行测量。

具体来说，BS-seq 数据在对齐后会展示如下格式的信息：
- 染色体编号或序列名称
- 胞嘧啶（C）的位置
- 在该位置上的总读段数量
- 在该位置上被识别为甲基化的读段数量

5. In the differential methylation example given in the class, why Fisher’s exact test and t-test give drastically different results? (20’)

这主要是由于它们各自所基于的基本单位的不同以及对于数据不确定性的处理方式不同。

Fisher 精确检验是以每个测序读段作为基本单位来考虑的，而这些读段在每个样本内并不是完全独立的，因为来自同一个样本的读段可能会具有相似的甲基化比例。在这个例子中，由于没有考虑到样本内的相关性，并且将所有读段合并到一起进行比较，所以得到的结果表明两组之间存在显著差异（p 值为 0.0264）。

另一方面，t 检验则是以样本为基础来进行的，它假定每个样本提供了等量的信息，并且忽略了由于测序深度不同所带来的比例估计的不确定性。在给定的例子中，两个肿瘤样本分别有 32/44 和 4/10 的甲基化读段，两个正常样本分别有 8/12 和 12/34 的甲基化读段。t 检验忽略了这些样本间不同的测序深度，并赋予每个样本相同的重要性，最终导致得出两组之间没有显著差异（p 值为 0.834）的结果。 因此，这两种方法的主要区别在于它们如何定义观察单元以及是否考虑到了样本内变异性和测序深度对甲基化比例估计的影响。

为了获得更准确的结果，可能需要采用其他方法来同时考虑样本内和样本间的变异性，比如通过跨 CpG 位点借用信息的平滑化方法或者贝叶斯分层模型。

6. Why dispersion is a better measurement than variance for the biological variability?(20’)

在生物学变异性的测量中，散度（dispersion）比方差（variance）是一个更好的度量，原因如下：

1. **方差与均值的依赖性**：
   - 在许多生物数据中，方差与均值不是独立的。例如，在二项分布中，方差与均值的平方成正比。这意味着均值的变化会直接影响方差，使得方差成为一个不可靠的离散程度度量。
2. **散度的稳定性**：
   - 散度是衡量数据分布的离散程度，它不依赖于分布的均值。这种稳定性使得散度成为衡量数据离散程度的更好指标，特别是在均值可能变化的情况下。
3. **对异质性的敏感性**：
   - 在生物学数据中，样本间的异质性（heterogeneity）很常见。散度能够更好地捕捉这种异质性，因为它考虑了数据点相对于均值的分布范围，而不是仅仅依赖于均值和方差。
4. **模型拟合和预测**：
   - 在统计建模中，使用散度而不是方差可以提高模型的拟合度和预测能力。这是因为散度能够更准确地反映数据的离散程度，从而允许更精确的模型参数估计。
5. **对异常值的鲁棒性**：
   - 散度对异常值的敏感性较低，这使得它在处理包含异常值的生物学数据时更为鲁棒。相比之下，方差对异常值非常敏感，因为异常值可以极大地影响均值和方差的计算。
6. **贝叶斯层次模型中的应用**：
   - 在贝叶斯层次模型中，散度用于模拟不同来源的变异性，如生物学变异和技术变异。通过使用散度，模型可以更准确地估计这些变异性，从而提高对差异甲基化区域（DMRs）的识别能力。
7. **数据的过度离散**：
   - 在许多生物学数据集中，观察到的变异性超过了二项分布的预期，这种现象称为过度离散（overdispersion）。散度能够更好地处理过度离散，因为它允许方差与均值的比率变化，从而更准确地描述数据。

综上所述，散度是衡量生物学变异性的一个更好的度量，因为它提供了一个更稳定、更敏感、更鲁棒的离散程度度量，特别是在处理异质性和过度离散的数据时。

7. Briefly describe the beta-binomial model used for BS-seq data. You need to provide the model, and explain how they capture different types of variations in the data. (20’)

Beta-Binomial Model（贝塔二项模型）是用于处理 BS-seq（亚硫酸氢盐测序）数据的一种统计模型，它特别适用于处理甲基化数据中的生物学和技术变异。以下是该模型的简要描述以及如何捕捉数据中的不同变异类型：

模型描述

Beta-Binomial Model 是一个层次模型，它结合了二项分布和 Beta 分布来描述和分析甲基化数据。在这个模型中：

- **二项分布（Binomial Distribution）**：用于描述在给定的甲基化水平 π 下，一个 CpG 位点被甲基化的次数 M。如果一个 CpG 位点被覆盖了 N 次，那么甲基化次数 M 服从参数为 N 和 π 的二项分布，即$M \sim Binomial(N, \pi)$。
- **Beta 分布（Beta Distribution）**：用于描述不同样本或重复实验中 π（甲基化比例）的变异性。π 的先验分布假定为 Beta 分布，这允许模型捕捉不同样本间的生物学变异。

捕捉不同类型变异

1. **Biological Variation**:
   - 这是指同一条件下不同样本间的自然甲基化水平变化。Beta-Binomial Model 通过 Beta 分布来建模 π 的分布，从而捕捉这种生物学变异。Beta 分布的参数可以根据数据调整，以反映不同样本间的甲基化水平差异。
2. **Technical Variation**:
   - 这是指由于测序过程中的随机抽样导致的甲基化水平估计的变异。在二项分布中，给定 π 和 N，甲基化读数 M 会因随机抽样而变化，这就是技术变异的体现。
3. **Overdispersion**:
   - 在实际应用中，甲基化数据常常表现出比二项分布预期更高的离散度，即过度离散现象。Beta-Binomial Model 可以通过调整 Beta 分布的参数来适应这种过度离散，使得模型更加灵活，能够更好地拟合实际数据。
4. **Borrowing Information**:
   - 当某些 CpG 位点的样本数量较少时，模型估计可能会不稳定。Beta-Binomial Model 可以通过层次结构在整个基因组的 CpG 位点间共享信息，从而提高估计的准确性，尤其是在样本数量有限的情况下。

综上所述，Beta-Binomial Model 通过结合二项分布和 Beta 分布，有效地捕捉了甲基化数据中的生物学和技术变异，使得差异甲基化分析更加准确和可靠。这种模型在处理 BS-seq 数据时，尤其适用于考虑和解释数据中的复杂变异性。