第七次作业

邹翔宇 | 2410833001

1. What are the two main types of ST technologies? What are their advantages and weaknesses? (10’)

- **基于测序的 ST 技术（Sequencing-based ST）**
  - **优点**
    - 能够在空间位置上测量基因表达，同时保持空间上下文
    - 可以对每个基因在每个细胞中的空间表达水平进行量化
    - 通过组织固定、HE 染色和成像，以及 cDNA 合成和 NGS 测序等步骤，能够获得基因表达数据和空间坐标
  - **弱点**
    - 基本的测序单元（spot）不能准确地匹配细胞，存在分辨率限制。
    - 例如 Visium 技术，其 spot 直径为 55 微米，中心到中心的距离为 100 微米，这限制了其在更精细尺度上解析细胞的能力。
- **基于成像的 ST 技术（Imaging-based ST）**
  - **优点**

    理论上，通过 4 色探针进行 8 轮杂交可以测量所有人类和小鼠基因的表达（4^8=65536）。能够直接在组织切片上进行分析，不需要复杂的样本制备步骤。
  - 弱点:
    - 分析过多的转录本可能导致光学拥挤问题。
    - 分辨率可能受到限制，尤其是在测量较小组织结构或单个细胞时。

2. A graph refers to a structure consisting of nodes and edges. How can a graph be constructed from ST data? Specifically, what are the nodes in this graph, and how can we determine whether there are edges between nodes? (10’)

在空间转录组学（ST）数据中，图（Graph）是一种用来表示细胞之间关系的结构，由节点（Nodes）和边（Edges）组成。

- **节点（Nodes）:**

在 ST 数据中，图中的每个节点通常代表一个细胞。这些节点包含了细胞的基因表达信息和空间坐标信息。

- **边（Edges）:**

边用来表示节点（细胞）之间的连接关系。 在 ST 中，边的确定通常基于细胞之间的空间邻近性。如果两个细胞在空间上彼此接近，它们之间就可能有一条边。边的权重可以基于细胞之间的空间距离来定义，距离越近，边的权重越大。另外，也可以通过分析细胞的基因表达模式来确定边的存在，如果两个细胞的基因表达模式相似，它们之间也可能有一条边。

通过这种方式构建的图可以帮助研究者理解细胞之间的空间关系和相互作用，进而揭示组织结构和细胞功能。

3. Briefly explain what spatially variable genes (SVGs) are and describe the difference between SVGs and highly variable genes (HVGs). (20’)

空间可变基因（Spatially Variable Genes）是指在不同空间位置上表达水平有显著变化的基因。这些基因的表达水平与它们在组织中的位置相关，可能受到细胞类型定位、细胞类型内部变异或与细胞类型无关的变异的影响。SVGs 的识别有助于研究者理解基因表达如何在空间上变化，以及这种变化如何与组织结构和功能相关联。

与 SVGs 相比，高变异基因（Highly Variable Genes）是指在整个数据集中或在不同样本中表达变异性很高的基因。HVGs 的识别通常用于单细胞 RNA 测序（scRNA-seq）数据分析中，以筛选出在细胞间表达差异大的基因，这些基因可能与细胞状态和功能密切相关。

SVGs 与 HVGs 的主要区别在于:

1. 空间依赖性: SVGs 强调的是基因表达在空间位置上的变化，而 HVGs 强调的是基因表达在不同样本或细胞中的总体变异性。
2. 生物学意义: SVGs 可能揭示了与空间位置相关的生物学过程，如细胞间的相互作用和微环境的影响，而 HVGs 可能揭示了细胞状态和功能的变化。
3. 分析方法: 识别 SVGs 通常需要考虑基因表达数据和空间坐标信息, 可能涉及到空间自相关分析等方法; 而识别 HVGs 则主要依赖于统计方法来评估基因表达的变异性，不一定需要空间信息。

简而言之，SVGs 关注的是基因表达的空间异质性，而 HVGs 关注的是基因表达的整体变异性。两者都是转录组数据分析中识别重要基因的有用方法，但它们关注的焦点和应用场景有所不同。

4. Why is it important to integrate sequencing-based ST data, imaging-based ST data, and scRNA-seq data? After integration, what additional information can each type of data obtain? (30’)

整合基于测序的空间转录组学（Sequencing-based ST）数据、基于成像的空间转录组学（Imaging-based ST）数据和单细胞 RNA 测序（scRNA-seq）数据非常重要，因为每种技术都有其独特的优势和局限性。通过整合这些数据，可以互补各自的不足，获得更全面和精确的生物学信息。

1. **整合的重要性：**
   - **基于测序的 ST 数据**：提供了在预定义区域内的基因表达信息，但空间分辨率无法达到单细胞水平。
   - **基于成像的 ST 数据**：可以在单细胞分辨率下测量基因表达，但受限于可测量的基因数量。
   - **scRNA-seq 数据**：提供了丰富的单细胞基因表达信息，但缺乏空间信息。
2. **整合后获得的额外信息：**
   - **对于成像基 ST 数据**：可以通过基因填充（Gene imputation）技术，利用 scRNA-seq 数据中的基因表达模式和关系，预测成像基 ST 数据中未测量的基因表达，从而扩展基因覆盖范围到全基因组水平。
   - **对于测序基 ST 数据**：可以通过空间解混（Spatial deconvolution）技术，推断每个位置上的细胞类型组成，从而获得更精细的细胞类型信息。此外，通过空间重建（Spatial reconstruction）技术，可以将 scRNA-seq 数据中的单细胞映射回 ST 数据的物理空间，恢复细胞的空间位置信息。
   - **对于 scRNA-seq 数据**：整合 ST 数据后，可以为原本失去空间信息的单细胞数据恢复空间上下文，从而更好地理解细胞如何在组织中相互作用和影响彼此的基因表达模式。

综上所述，整合这些数据类型可以克服单一技术的限制，实现对基因表达的空间异质性、细胞类型的精确识别以及细胞间相互作用的深入理解，为研究复杂的生物学问题提供了更全面的工具。

5. Briefly describe how Gaussian processes are utilized in SpatialDE to identify SVGs. Include the model and explain how they capture different types of variations in the data. (30’)

高斯过程（Gaussian processes）是一种统计模型，用于模拟具有连续输入空间的随机变量。在空间转录组学（ST）中，高斯过程被用于识别空间可变基因（SVGs），如 SpatialDE 工具中所实现的。

1. **SpatialDE 模型**：
   - SpatialDE 将基因表达建模为一个高斯过程。对于给定的基因，其在不同空间坐标上的表达值被假设遵循一个多变量正态分布。
   - 这个多变量正态分布的均值是基因的平均表达水平，而协方差矩阵包含空间和非空间组成部分。
2. **捕获不同类型的变异**：
   - **空间组成部分**：协方差矩阵中的空间组成部分使用高斯核（Gaussian kernel）来描述。高斯核根据细胞之间的空间距离来计算空间权重，距离越近的细胞，其表达值的相关性越高。
   - **非空间组成部分**：协方差矩阵中的非空间组成部分代表基因表达中的随机变异，这可能与技术噪声或其他非空间因素有关。
3. **模型比较**：
   - SpatialDE 通过比较包含空间组成部分的模型（备择模型）与不包含空间组成部分的模型（零假设模型）来评估空间方差分量的显著性。
   - 使用对数似然比（Log-likelihood ratio, LLR）测试来比较两个模型。如果 LLR 统计量在零假设下遵循卡方分布，并且 p 值小于显著性水平（例如 0.05），则认为空间方差分量是显著的。
4. **识别 SVGs**：
   - 通过上述模型比较，SpatialDE 可以识别出那些表达模式在空间上显著变化的基因，即 SVGs。
   - 这种方法允许研究者识别出那些在空间上具有特定模式的基因，这些基因可能与细胞间的相互作用或微环境的影响有关。

总之，SpatialDE 利用高斯过程通过建模基因表达的空间和非空间变异来识别 SVGs，提供了一种强大的工具来探索空间转录组数据中的复杂空间模式。