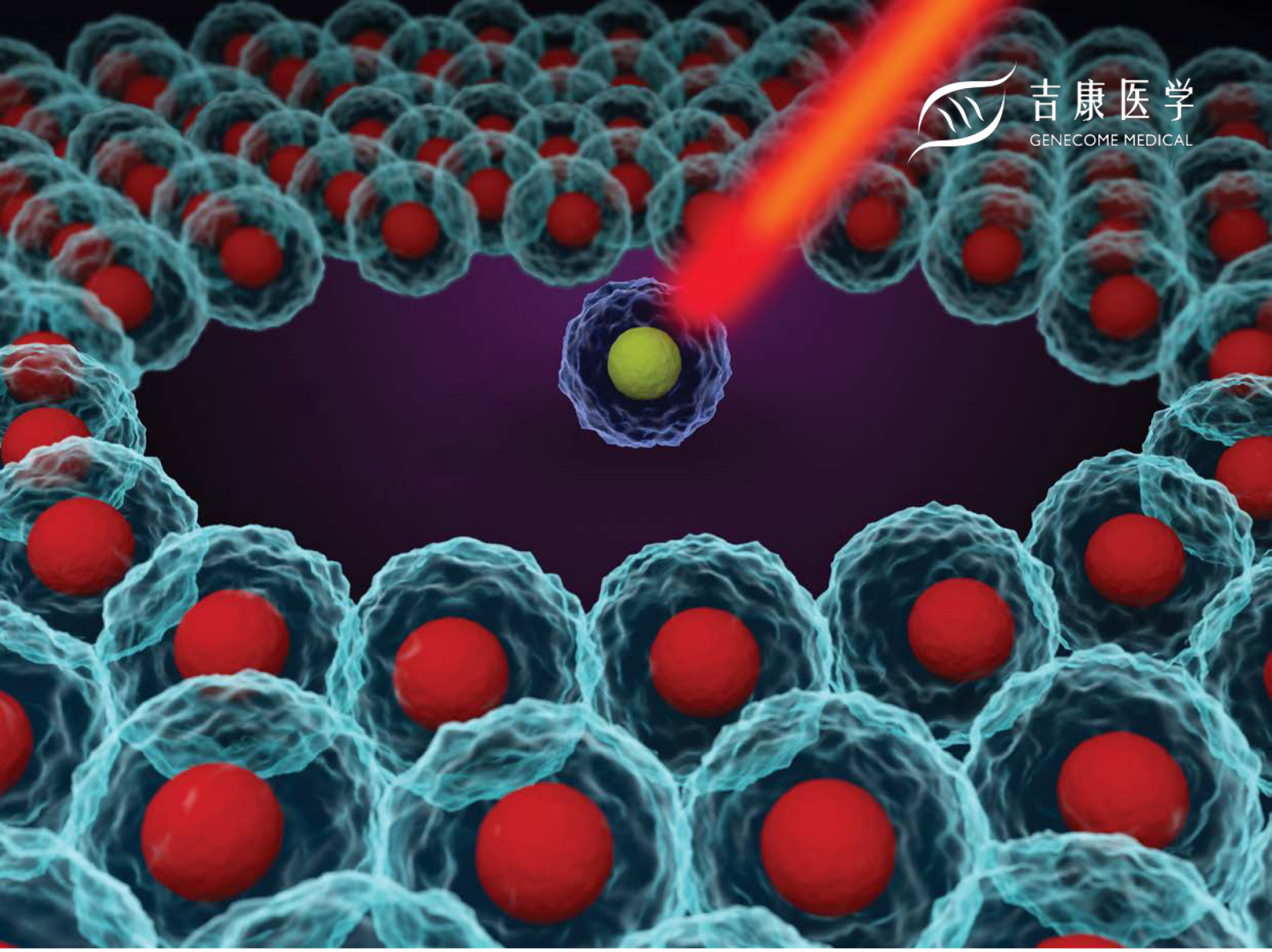


吉康医学
GENECOME MEDICAL

10X Genomics 单细胞转录组测序服务

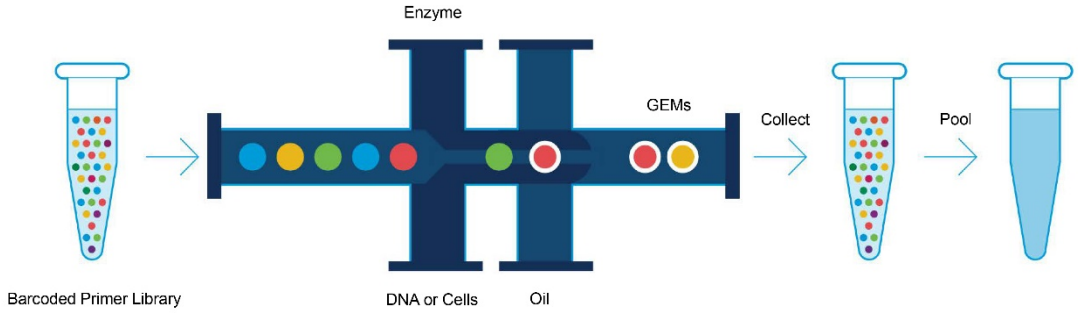
客户手册

北京吉康医学科技有限公司 | 服务热线：010-67860414 | 联系邮箱：genecomeservice@163.com
技术引领医学转化，专业创造行业口碑 | www.genecome.cn

目录

10X Genomics 平台概述

10X Genomics 平台是基于微滴的核酸标签分配系统，这个系统提供了数百万个携带唯 一标签（Barcode）的微滴，通过微流控技术，将这些标签分配给不同的 DNA 片段或单个细 胞，实现了长片段的 DNA 测序和单细胞分辨率的转录组测序。



基于 10X Genomic Chromium 系统平台，吉康医学提供以下解决方案：	
10X Genomic Chromium 系统解决方案	**应用方向**
Chromium Single Cell 3’ Solution	单细胞基因表达谱
Chromium Single Cell 5’ Solution	单细胞表达谱、免疫组库分析
Feature barcoding technology	细胞表面蛋白检测、抗原特异性 T 细胞分析
Chromium Single Cell ATAC Solution	单细胞染色质调控图谱

吉康医学技术优势

（1）**专业的技术操作：**丰富的样本处理经验、多年建库技术沉淀和测序实验操作经验；

（2）**强硬的硬件支撑：**5S 实验规范管理，多套计算集群，保证数据产出质量和分析速度；

（3）**全面的分析内容：**具有专业、成熟的生物信息分析团队，熟悉各种分析算法和软件，保证标准、个性化和售后调整分析的全面支撑，并在分析过程中与客户充分互动；

（4）**更高的服务标准：**采取严格的质量标准、多重实验质控和完备的实验记录，确保更高的成功率、更低的样品损失及实验失败风险，并且客户可以参与整个实验过程，接受客户监督，更加透明和可信；

（5）**完善的后续支持：**持续专注于转录组学定量检测，每年为上千位客户提供相关服务；

（6）**优秀的服务理念：**核心团队平均 12 年以上技术服务领域工作经验，具备良好的职业操守及服务客户理念。

1. 实验技术

新一代测序技术（Next Generation Sequencing，NGS）的发展和应用给生物学研究带来了极大的便利和突破，而传统的批量转录组测序（Bulk RNA-seq）作为其中的突出代表，在科学研究中发挥着极大的作用。尽管从一群细胞中获得全基因组范围的 RNA 表达量对理解生物学机制非常有用，但这种方法只能获得所有细胞的基因表达平均水平，无法体现细胞的异质性。

单细胞转录组测序是在单个细胞水平对细胞转录组进行高通量测序的一项新技术，其原理是将分离的单个细胞中微量的 mRNA 通过高效扩增后再进行高通量测序。单细胞转录组测序能够有效解决组织样本细胞异质性以及常规 RNA-seq 被掩盖的细胞群内的转录组异质性难题，有助于发现新的稀有细胞类型，深入了解细胞生长过程中的表达调控机制。

基于 10X Genomics 最新 Chromium 系统利用油包水的微反应体系，通过序列标签（Cell barcode 和 UMI）区别群体中的不同细胞和转录本，获得单细胞水平的基因表达谱，实现数千甚至上万个单细胞群体分析，解决常规 scRNA-seq 方法在通量或扩展性方面存在的不足，为单细胞研究开拓了新的思路。

1.1 技术优势

（1）**通量高**：一次可以同时测 8 个样本，每个样本最多可以检测到上万个细胞；

（2）**周期短**：10 分钟内完成上万个细胞封装，一天之内完成细胞悬液制备、单细胞捕获、扩增以及建库；

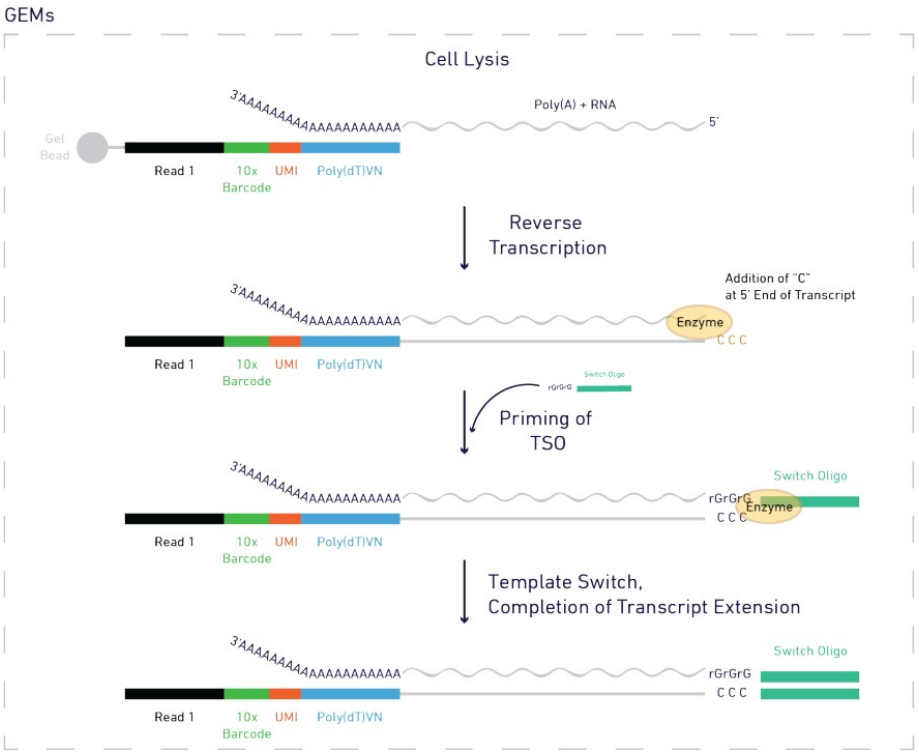
（3）**真单细胞测序**：真正意义上实现单细胞测序；

（4）**捕获效率高**：细胞捕获效率高达 65 %；

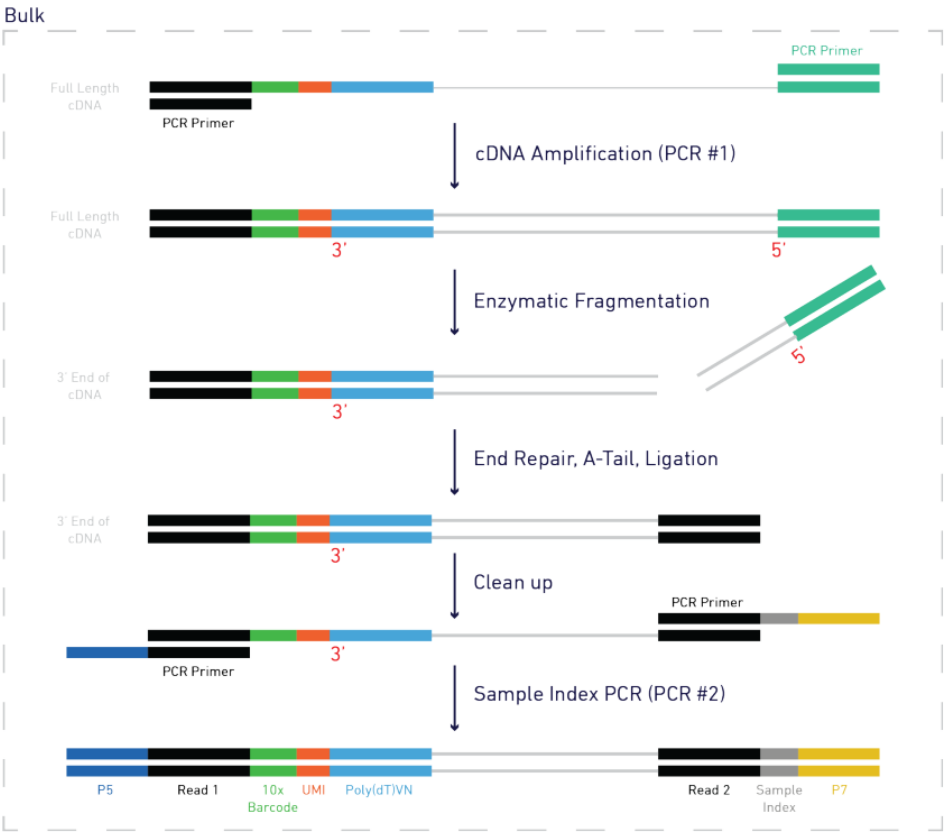
（5）**应用范围广**：动物细胞和植物细胞均可以进行单细胞测序，已广泛应用于肿瘤细胞异质性、免疫细胞群体检测、植物生长发育等研究。

1.2 技术原理

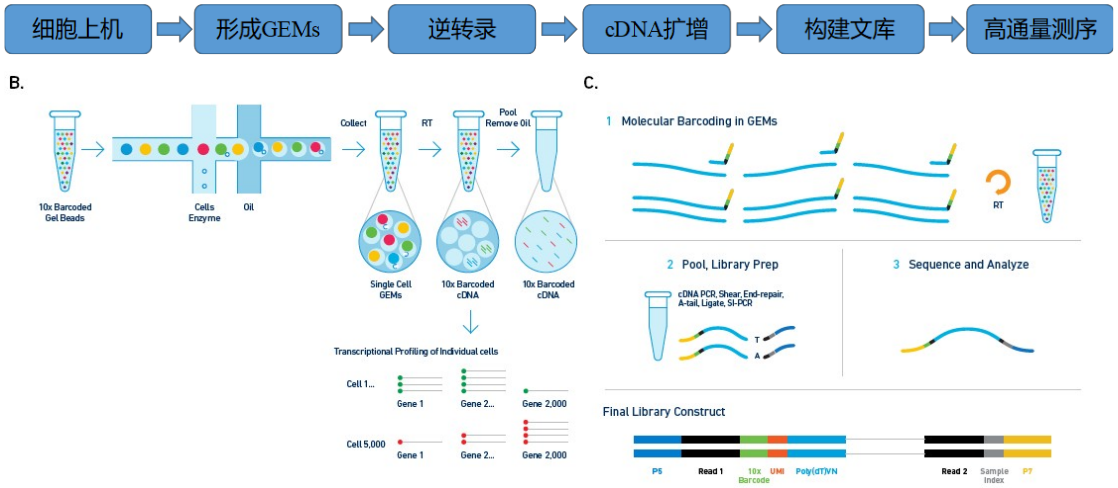
10X Genomics 的 Chromium 系统利用 8 通道的微流体双十字交叉系统，将含 barcode 的凝胶珠（Gel Beads）、细胞和酶的混合物、油三者混合，形成 GEMs（油包水的微体系）。GEMs 形成后，细胞裂解，凝胶珠自动溶解释放大量 barcode 序列，随后 mRNA 逆转录产生带有 10X barcode 和 UMI 信息的 cDNA，构建标准测序文库。下图展示 Chromium Single Cell 3’ Solution 中 mRNA 逆转录产生带有 10X barcode 和 UMI 信息的 cDNA 过程。



反转录成 cDNA 之后，酶切打断，加上测序接头，构建二代测序文库。如下图所示。



1.3 建库流程



（1）凝胶珠进入第一个进样口与细胞悬液和酶等混合，通过第二个进样口时被油滴包裹，形成 GEMs，随后细胞裂解，凝胶珠溶解，进行逆转录；

（2）油滴破裂，收集 cDNA 产物，进行 cDNA 扩增；

（3）构建测序文库进行测序。

1.4 样本要求

1.4.1 样本准备原则

（1）样品总量：制备成细胞悬浮液，一个样本的细胞起始量需大于 1×10^5 个；

（2）样品浓度：最终上样前细胞浓度在 700 - 1200 cells / μL；

（3）细胞活性：活细胞数目在 85 %以上；

（4）细胞大小：直径小于 40 μm；

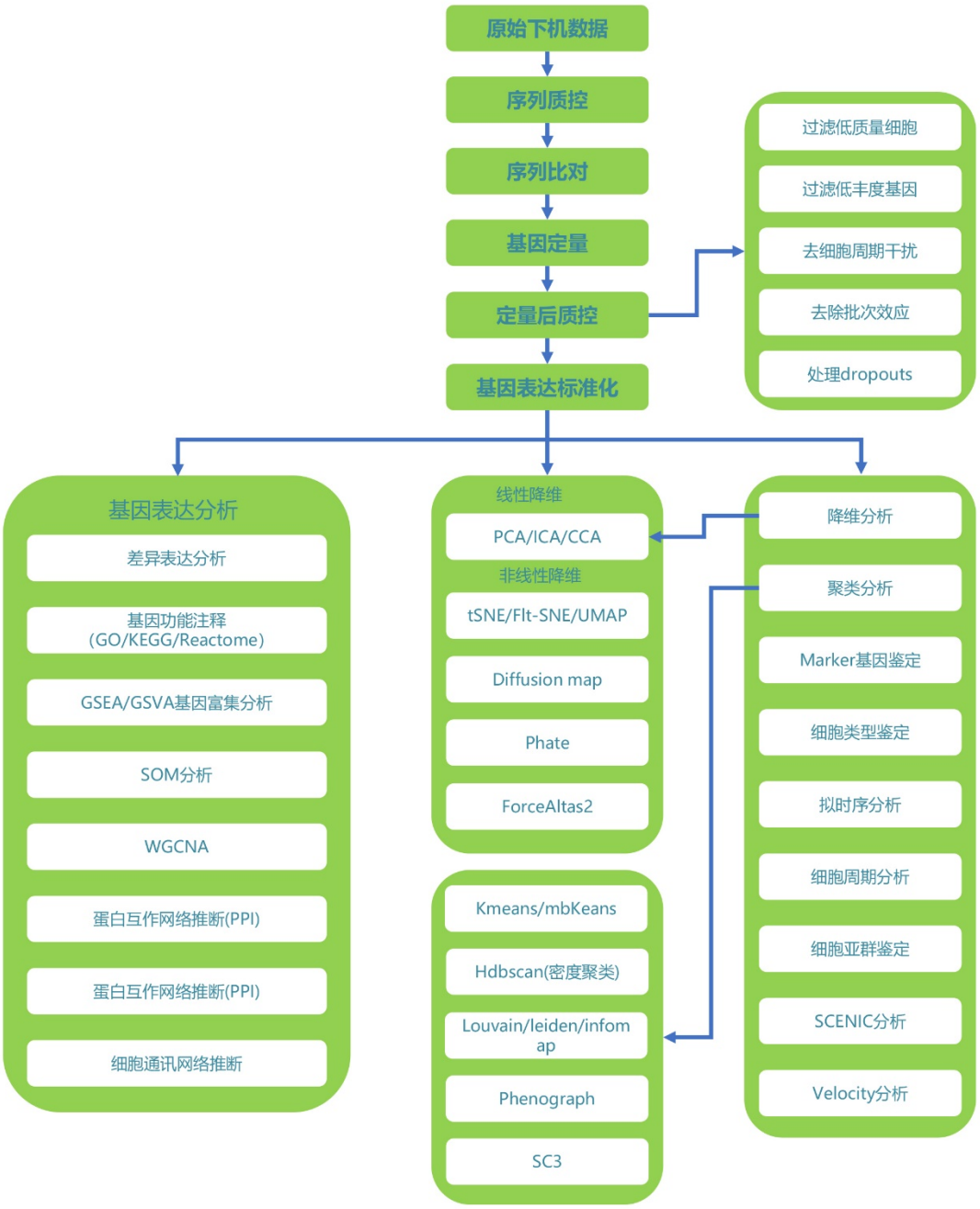
（5）细胞培养基及缓冲液不能含有 Ca^{2+} 和 Mg^{2+} 等影响酶活性的物质；

（6）组织需解离成单细胞悬液，植物细胞需制成原生质体。

注：单细胞悬液的准备方法可参考《吉康医学10X Genomics 单细胞测序样品准备指南》，请联系当地销售索取。

2. 数据分析

2.1 分析流程



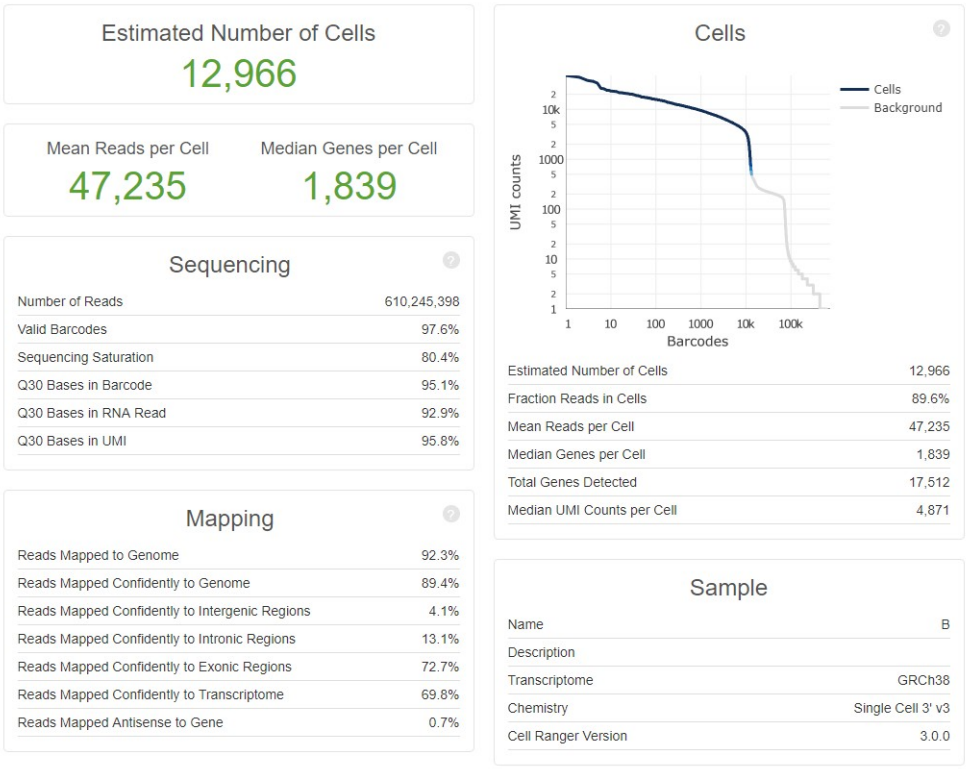
10X Genomics 3’单细胞转录组测序生物信息分析流程与内容

2.2 分析内容展示

2.2.1. 测序数据质量评估

在得到测序数据后，首先需要对样本数据进行质量评估，根据比对到参考基因组的结果计算每个样品数据的测序饱和度，测序随机性以及 reads 在不同基因元件的富集情况。我们使用 10X Genomics 官方软件 Cell Ranger（https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger）对原始数据进行质控，其内部整合了 STAR[1] 软件，可将 reads 比对到参考基因组（Ensembl 基因注释数据库）上后可获得样本数据的高质量细胞数、基因数、基因组比对率等质控结果。此外，该软件可通过识别 序列中的 Barcode 信息来区分不同的细胞和 UMI 信息对每个细胞内不同 mRNA 分子进行定量。

结果展示如下：



测序数据质量统计结果示例

样本细胞质量统计结果说明		
评估参数	**术语**	**术语说明**
代表性指标	Estimated Number of Cells	高质量细胞数，小于所有细胞的 UMI 总数的 99%分位数的 10%定义为背景噪音细胞
	Mean reads per Cell	每个细胞中的平均序列数
	Median genes per Cell	每个细胞中检测到的基因数中位值，UMI 数大于 0 被定义为检测到的基因
测序质量	Number of reads	原始下机数据的序列数
	Valid Barcodes	有效 cell barcodes 序列数百分比
	Sequencing Saturation	测序饱和度
	Q30 Bases in Barcode	测得的 Cell barcode 序列质量大于 Q30 的序列数百分比
	Q30 Bases in RNA Read	测得的 R2 reads 中质量大于 Q30 的序列百分比
	Q30 Bases in UMI	测得的 UMI 序列中序列质量大于 Q30 的序列百分比
比对质量	Reads Mapped to Genome	比对到基因组上的序列百分比
	Reads Mapped Confidently to Genome	高置信度比对到的基因组上的序列百分比
	Reads Mapped Confidently to Intergenic Regions	比对到参考基因组的基因间隔区域的序列数百分比
	Reads Mapped Confidently to Intronic Regions	比对到参考基因组的内含子区域的序列数百分比
	Reads Mapped Confidently to exonic Regions	比对到数据库的基因组外显子区域的序列数百分比
	Reads Mapped Confidently to Transcriptome	比对到参考物种的转录组序列上的序列百分比
	Reads Mapped Antisense to Gene	比对到基因的负链的序列百分比
细胞质量	Estimated Number of Cells	估计检测到的高质量细胞数
	Fraction Reads in Cells	在高质量细胞的序列数百分比
	Mean Reads per Cell	每个高质量细胞的平均序列数
	Median Genes per Cell	每个高质量细胞的基因数中值
	Total Genes Detected	所有细胞检测到的基因总数
	Median UMI Counts per Cell	每个高质量细胞的平均 UMI 数

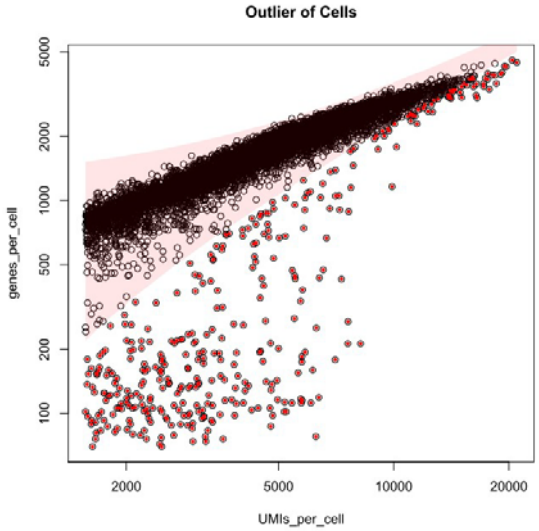
2.2.2. 基因定量质控

单细胞转录组测序采用测得的转录本序列结合 UMI 和 cell barcode ，以获知单细胞内每一个转录本分子的绝对数量。一般情况下一个凝胶珠中的细胞数目只有一个，但可能会出现没有细胞或 2 个甚至多个细胞的情况。另外，当细胞死亡或破碎时，细胞中的线粒体基因比例也会上升。因此在 Cell Ranger 初步质控的基础上进一步对实验数据进行质控。

理论上每个细胞中表达的基因数（nGene）、每个细胞的 UMI 数（nUMI）和每个细胞的线粒体基因的表达量会集中分布在某一区域内，根据它们的分布特征可以拟合分布模型，使用该模型找到其中的离域值。因此，我们将每个细胞中表达的基因数（nGene）、每个细胞的 UMI 数（nUMI）和每个细胞的线粒体 RNA 比例（percent.mito）作为单细胞数据的质控参数 [2]。我们通常通过去除这些参数过高的和过低的细胞以剔除多细胞、双细胞或者未结合上细胞的数据，再进行后续分析。

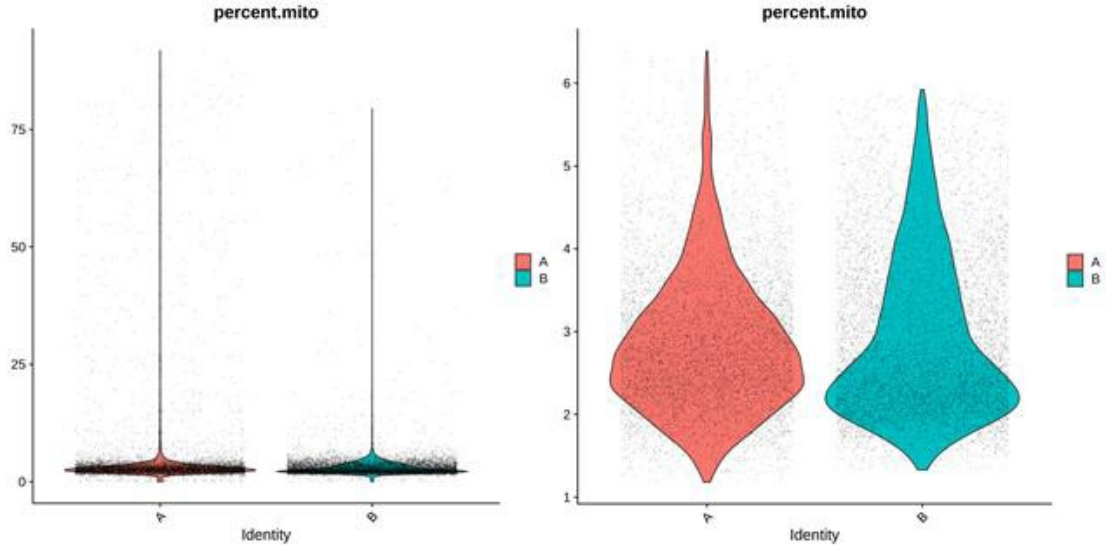
结果展示如下：

线性模型拟合曲线及质控前后每个细胞的基因数量（nGene）、UMI 数量（nUMI）和线粒体基因所占比例（percent_mito）的小提琴图展示如下：



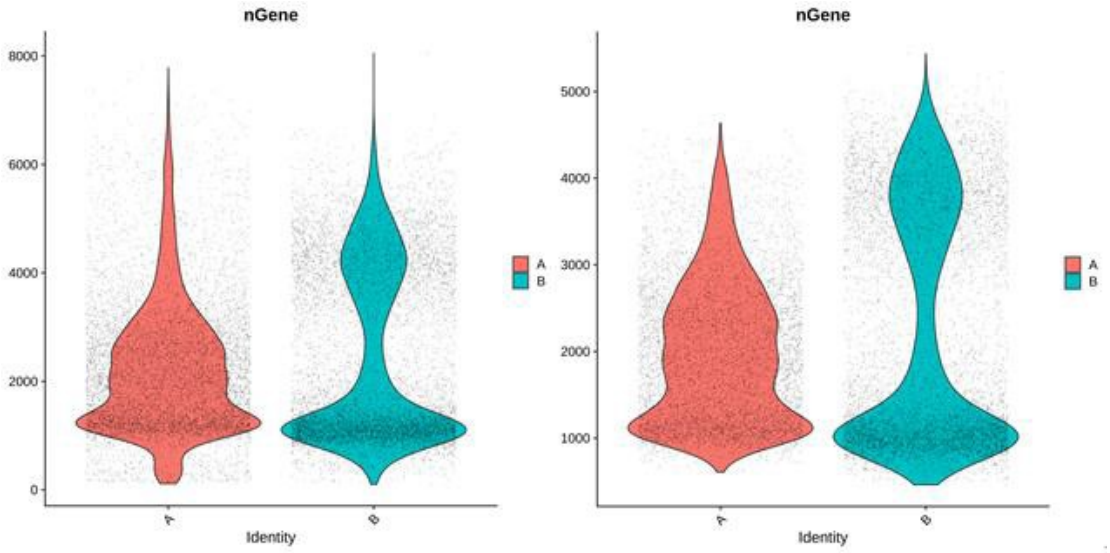
单细胞转录组每个细胞 UMI 数与基因数的线性拟合模型

图片说明：横轴为每个细胞内的 UMI 数，纵轴为每个细胞内的基因数，根据二者的线性关系拟合分布模型，着色的点表示离域细胞，在下游分析中会剔除。



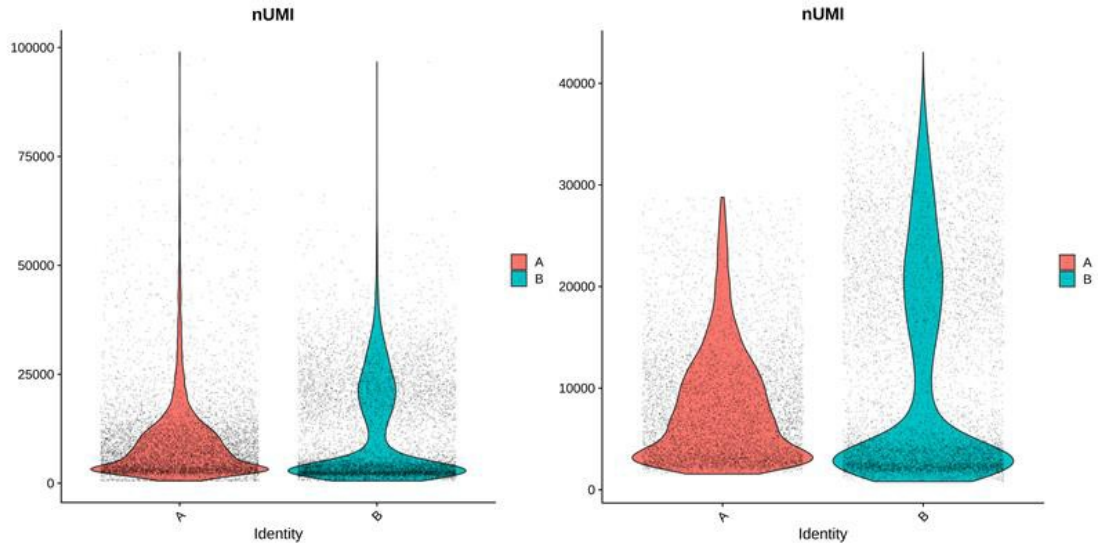
质控前后样本中每个细胞线粒体基因转录本的比例分布图

图片说明：纵轴表示线粒体转录本占单个细胞所有转录本的比例（%），图中每个点代表一个油包水微滴中细胞的线粒体转录本比例，每个小提琴图反映对应样本中所有细胞的线粒体转录本在细胞所有转录本中所占的比例，一般要求大部分细胞的线粒体转录本比例越低越好（特殊样本除外）。左图为质控前，右图为质控后。



质控前后样本中每个细胞中表达基因数目的小提琴分布图

图片说明：纵轴表示细胞中有表达的基因数目，图中每个点代表一个油包水微滴中细胞的基因数目。该图反映了样本中的每一个细胞表达基因的数目，基因数目异常过多的点很有可能是由于对应的油包水微 滴中包含多个细胞，需要根据实际需求设置合理的阈值过滤掉。左图为质控前，右图为质控后。



质控后样本中每个细胞中 UMI 数目的小提琴分布图

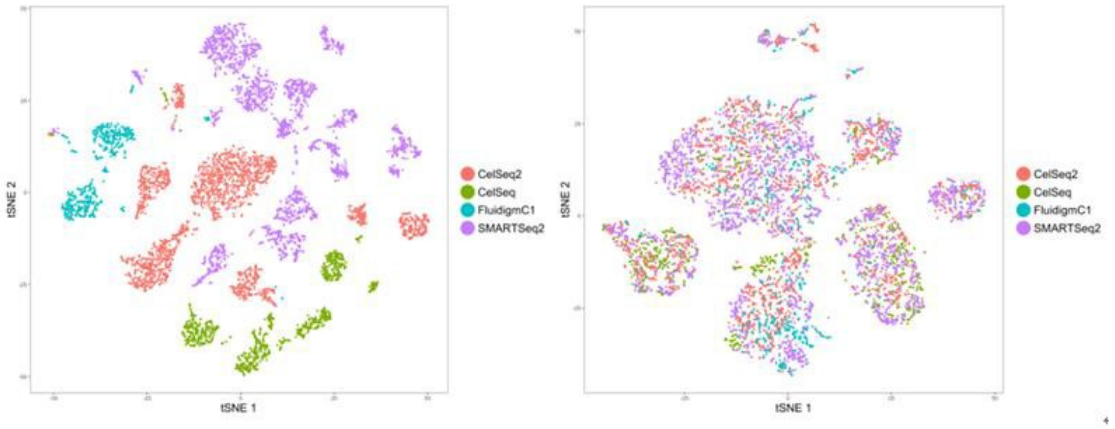
图片说明：纵轴表示 UMI 数，图中的每个点代表一个油包水微滴中细胞的 UMI 数目，即转录本的 数目，该图反映了样本中每一个细胞的转录本数目，转录本数目异常过多的细胞需要通过设置合理的阈值 将其过滤。左图为质控前，右图为质控后。

2.2.3. 去除批次效应

相同组织由于取样时间、实验操作、实验技术等造成的样本的差异属于批次效应，这种情况下我们必须将该差异去除，从而保证样本中的异质性均来自于实验设计或者细胞类型的 差异。

我们通过互享最近邻（MNN）[3]的方法找到不同批次样本中相似的细胞，并将所有这种 相似细胞之间的差异去除，最大程度的保证数据的质量。

结果展示如下：



批次效应处理前后不同批次样本细胞 tSNE 可视化

图片说明：横纵坐标分别代表降维后的主成分，图中的每个点代表一个细胞，距离相近的细胞认为是 同一类型细胞，不同批次的样本细胞以不同的颜色区分。左图为去除批次效应前，右图为去除批次效应后。

2.2.4. 降维和可视化

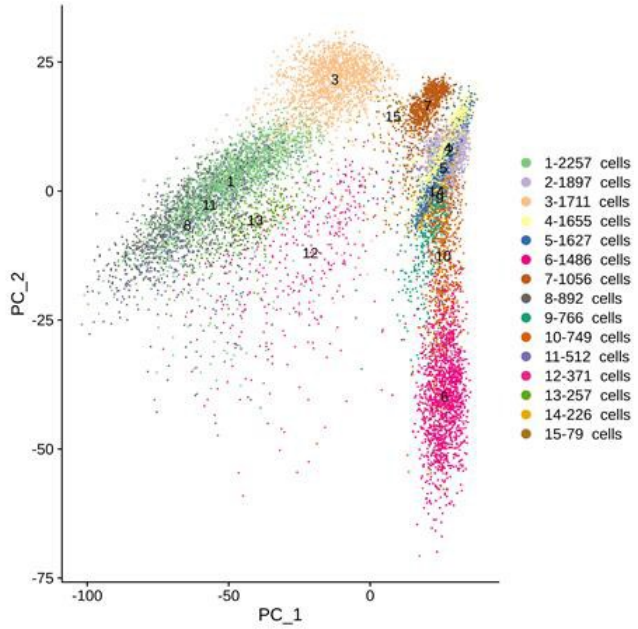
单细胞转录组定量矩阵是一个 M * N 维的矩阵，矩阵中的行一般为基因，列一般都是细胞。通常一个单细胞转录组测序样本的定量矩阵可以达到上万行*上万列的超高维度，在 这样一个超高维度下进行聚类分析，不仅运算量极大，而且难以获得较好的聚类结果。因此， 在对单细胞转录组的定量结果聚类之前，一般需要先降维。所谓降维，也就是从上万的基因 中提取新的维度，使用新维度表示的数据，一方面既能最大限度地保留样本中数据的信息，另一方面能够有效减少数据中的冗余，从而提高后续聚类运算效率。单细胞转录组目前常用的降维算法有 PCA[4]、tSNE[5]、diffusion maps[6] 及 UMAP[7] 等。

（1） PCA (principal component analysis) 降维

PCA 降维首先将数据标准化处理后建立共变异数矩阵，然后利用奇异值分解（SVD） 求得特征向量和特征值；通常特征值会由大到小排列，选取 k 个特征值与特征向量，最后将 原本的数据映射到特征向量上，得到新的特征数，从而实现降维的目的。PAC 属于线性降

维算法，它通过前 N 个主成分来表示数据集，使得降维空间中距离在该空间的所有区域都有一致的解释。通常，PCA 分析用于非线性降维方法的预处理步骤。 我们利用基因的表达量进行主成分分析（PCA 分析），考察样品分布情况，对样本间或不同细胞群关系进行探究或者对实验设计进行验证。

结果展示如下：



样本 PCA 降维聚类结果图

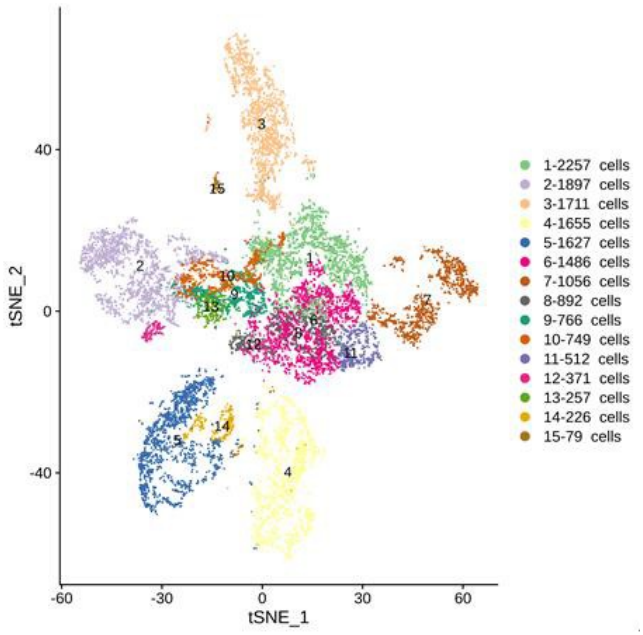
图片说明：横纵坐标分别代表降维后的成分，图中的每个点代表一个细胞，距离相近的细胞认为是同一类型细胞，不同群细胞以不同的颜色区分。

（2） tSNE (t-distributed stochastic neighbour embedding) 降维

t-分布邻域嵌入算法（tSNE）是一种非线性降维算法，它将数据点之间的相似度转换为概率。原始空间中的相似度由高斯联合概率表示，嵌入空间的相似度由“学生 t 分布”表示。tSNE 能将在高维空间中非常相似的点在降低到低维度之后依然保持这些点在空间上非常相近。

我们使用 tSNE 降维算法对获得的单细胞 UMI 定量矩阵进行降维分析。

结果展示如下：



样本 tSNE 降维聚类结果图

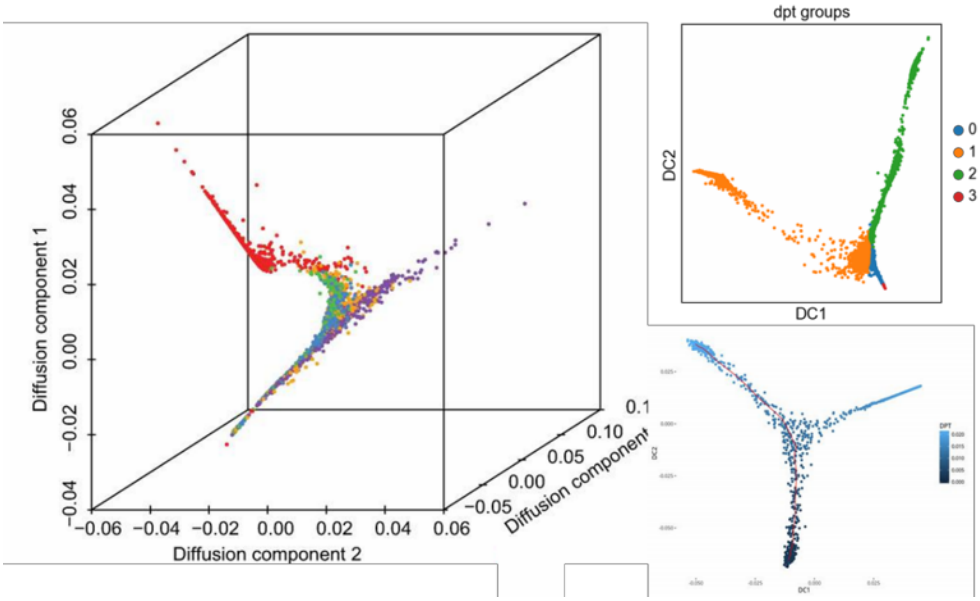
图片说明：横纵坐标分别代表降维后的成分，图中的每个点代表一个细胞，距离相近的细胞认为是同一类型细胞，不同群细胞以不同的颜色区分。

（3） diffusion maps 降维

扩散图（Diffusion maps）是 COIFMAN R 等人在 2006 年提出的一种基于流形学习的非 线性降维方法，其主要思想来自于动力系统。作为一种新的流形学习框架，扩散映射通过在 扩散过程中尽可能地保持扩散距离来进行降维，即保持样本点的局部结构不变，通过局部关 系定义全局关系，使样本点在低维空间中仍保持这种稳定的全局关系。作为一个非线性降维 工具，diffusion maps 可应用于单细胞组学数据降维分析。由于扩散成分强调数据中的转换， 因此它们主要用于分化等连续过程的分析。通常，每个扩散成分（即扩散图维度）突出了不 同细胞群的异质性。

我们使用 diffusion maps 降维算法对获得的单细胞 UMI 定量矩阵进行降维分析。

结果展示如下：



样本 diffusion maps 降维聚类结果图

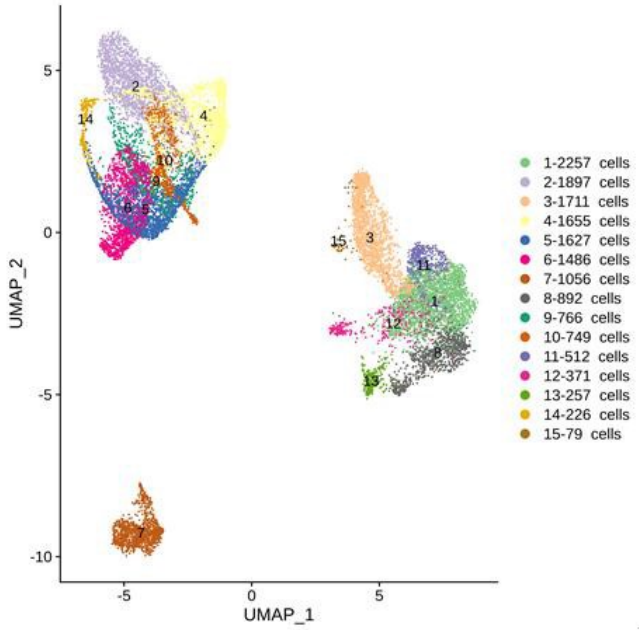
图片说明：横纵坐标分别代表降维后的成分，图中的每个点代表一个细胞，距离相近的细胞认为是同 一类型细胞，不同群细胞以不同的颜色区分。

（4）UMAP（Uniform Manifold Approximation and Projection）降维

统一流形逼近与投影（UMAP）是一种新的流形学习技术。UMAP 是建立在黎曼几何和 代数拓扑理论框架上的。UMAP 是一种非常有效的可视化和可伸缩降维算法。在可视化质 量方面，UMAP 算法与 t-SNE 具有竞争优势，但是它保留了更多全局结构、具有优越的运 行性能、更好的可扩展性。此外，UMAP 对嵌入维数没有计算限制，这使得它可以作为机 器学习的通用维数约简技术。

我们使用 UMAP 降维算法对获得的单细胞 UMI 定量矩阵进行降维分析。

结果展示如下：



样本 UMAP 降维聚类结果图

图片说明：横纵坐标分别代表降维后的成分，图中的每个点代表一个细胞，距离相近的细胞认为是同一类型细胞，不同群细胞以不同的颜色区分。

2.2.5. 聚类

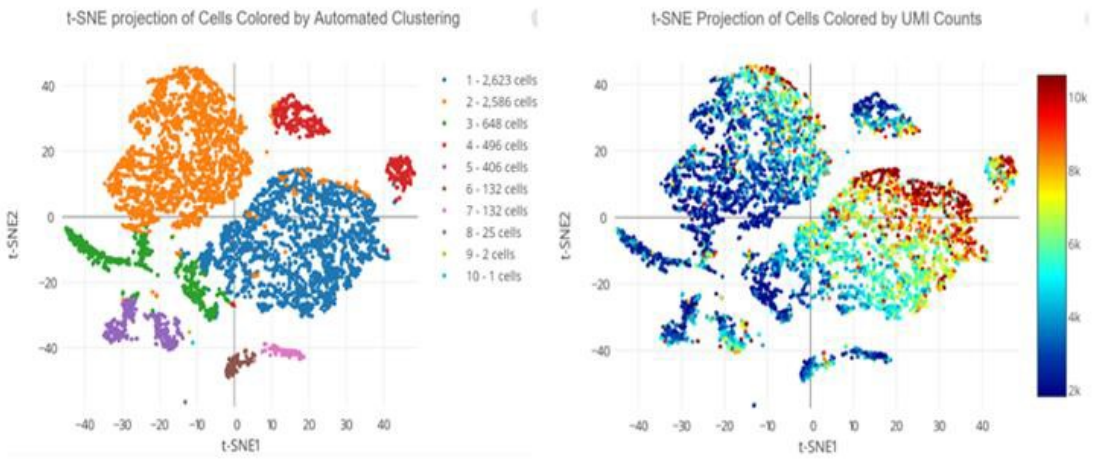
一般而言，相似的细胞具有相似的基因表达谱，因此根据每个细胞的基因表达结果可以将相同类型的细胞聚类到一起，形成一个细胞簇。单细胞转录组常用聚类算法为 K-means 聚类、SNN 聚类[8]等。

（1） K-means 聚类

K-means 算法是无监督的聚类算法，其基本思想是：对于给定的样本数据集，按照样本之间的距离大小，将样本数据集划分为 K 个簇；让簇内样本差异尽量较小，而让簇间样本之间的差异尽量的大。K-means 算法的优势在于运行速度快，简单；对于大数据集有较高的效率并且可伸缩，适合挖掘大规模的数据集。

我们使用 K-means 方法对单细胞 UMI 定量矩阵的 tSNE 降维结果进行聚类分析。

结果展示如下：



K-means 聚类结果图

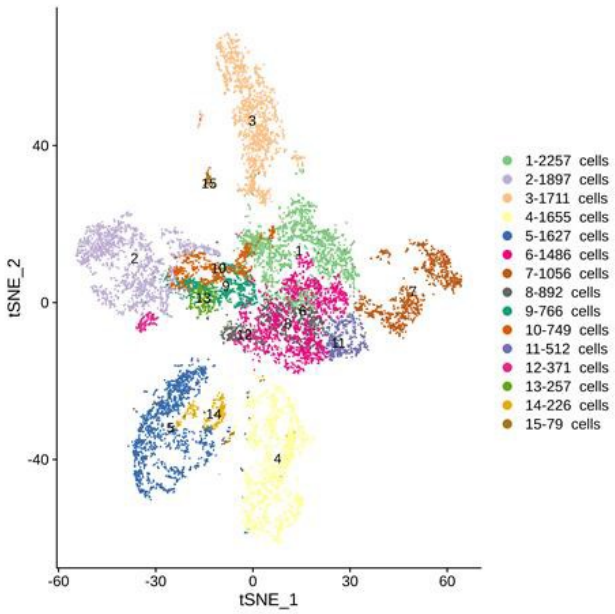
图片说明：横纵坐标分别代表降维后的成分，图中的每个点代表一个细胞，距离相近的细胞认为是同一类型细胞，不同群细胞以不同的颜色区分。

（2） SNN 聚类

共享最近邻（Shared Nearest Neighbor, SNN）算法是一种无监督聚类算法。SNN 聚类方法首先通过查看共享点的最近邻的数量来重新定义点之间的相似性。利用这种相似性度量，根据一个点的最近邻的相似性之和定义了 SNN 密度的概念。高密度的点成为代表点或核心点，而低密度的点则代表噪音或异常值并被消除。然后，通过找到与代表点非常相似的所有点的集合来对数据点进行划分。

我们使用 SNN 的方法对单细胞 UMI 定量矩阵 tSNE 降维的结果进行聚类分析。

结果展示如下：



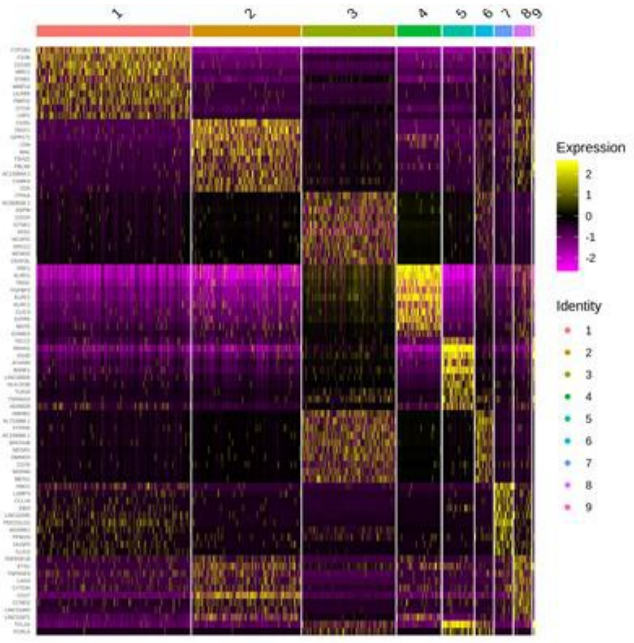
SNN 聚类结果图

图片说明：横纵坐标分别代表降维后的成分，图中的每个点代表一个细胞，距离相近的细胞认为是同一类型细胞，不同群细胞以不同的颜色区分。

2.2.6. Marker 基因鉴定与可视化

针对指定细胞群与其余细胞群进行差异检验，寻找各细胞群中的特异性 Marker 基因。Marker 基因是在指定细胞群的绝大多数细胞中有较高表达，而在其余细胞类群中只有少部分表达的基因，且该基因在此细胞群相对于其他细胞群中显著上调表达。

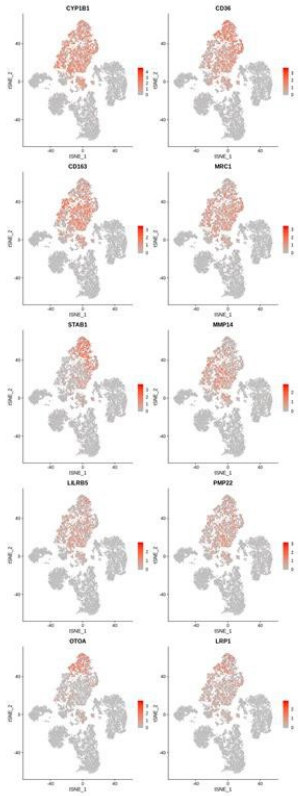
结果展示如下：



Marker 基因表达谱热图

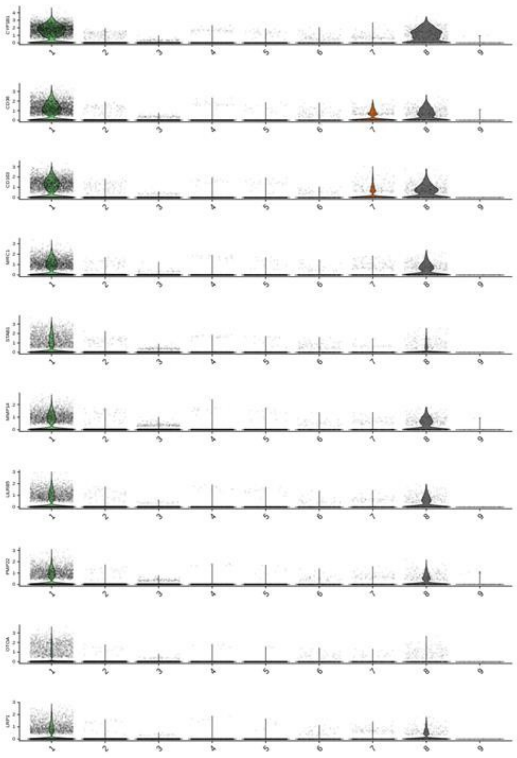
图片说明：行表示 Marker 基因，列表示细胞群，图中黄色表示高表达，紫色表示低表达。

每个细胞群中 Top10 Marker 基因的可视化展示如下：



Marker 基因在 tSNE 聚类结果中的可视化图

图片说明：Marker 基因的featureplot。红色越深表示该细胞中对应基因的表达量越高。



Marker 基因特异表达小提琴图

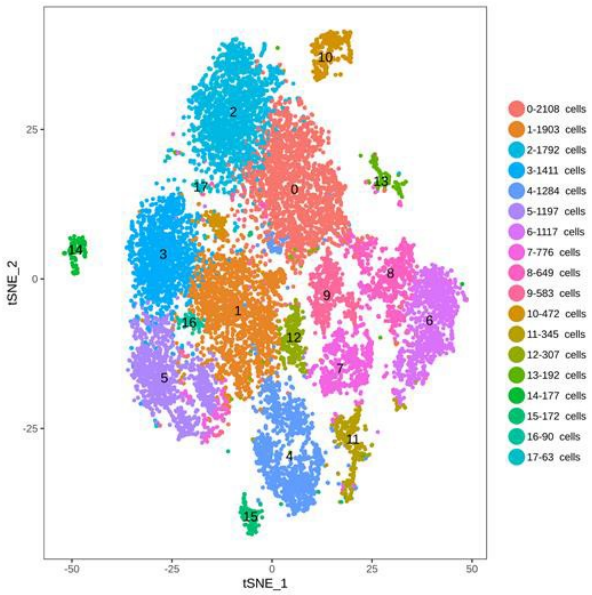
图片说明：横坐标为细胞群编号，纵坐标为标准化后的基因表达值，图中每个点表示每个细胞中 Marker 基因的表达值。

2.2.7. 细胞亚群鉴定

根据先前细胞聚类的结果，可以从中挑选感兴趣的特定细胞群进行亚群细分。

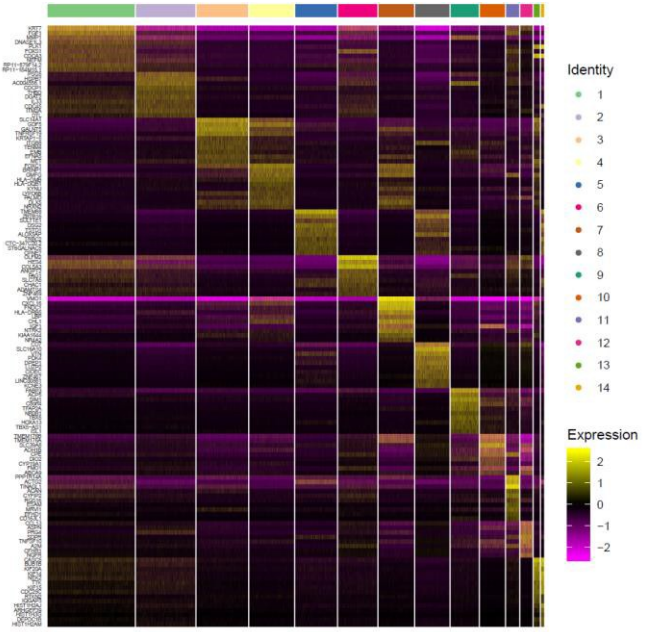
我们使用非线性降维 tSNE 方法对特定细胞群进行重新聚类， 然后对重新聚类后的每个 cluster 亚群鉴定 Marker 基因，并且可以使用参考数据集对特定的细胞亚群进行 Celltype 注释。

结果展示如下：



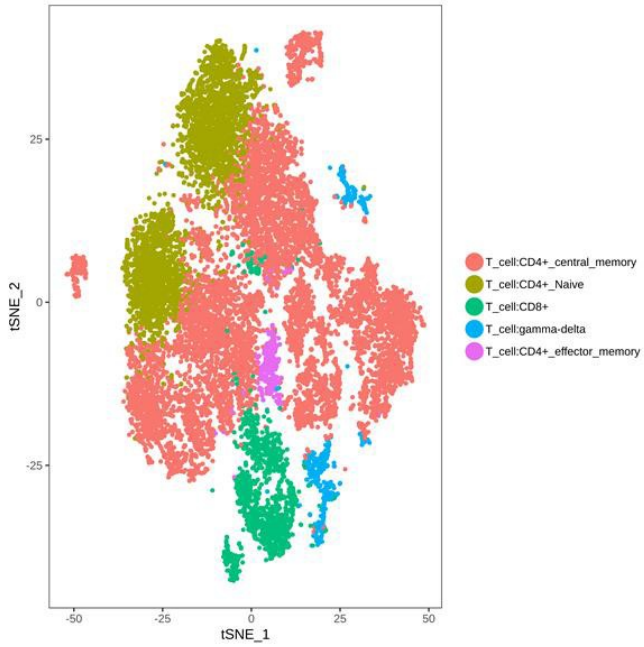
亚群细胞 tSNE 聚类结果图

图片说明：横纵坐标分别代表降维后的成分，图中的每个点代表一个细胞，距离相近的细胞认为是同一类型细胞，不同群细胞以不同的颜色区分。



每个 Cluster 亚群的 Marker 基因表达谱热图

图片说明：行表示 Marker 基因，列表示细胞亚群，图中黄色表示高表达，紫色表示低表达。



亚群细胞类型鉴定结果图

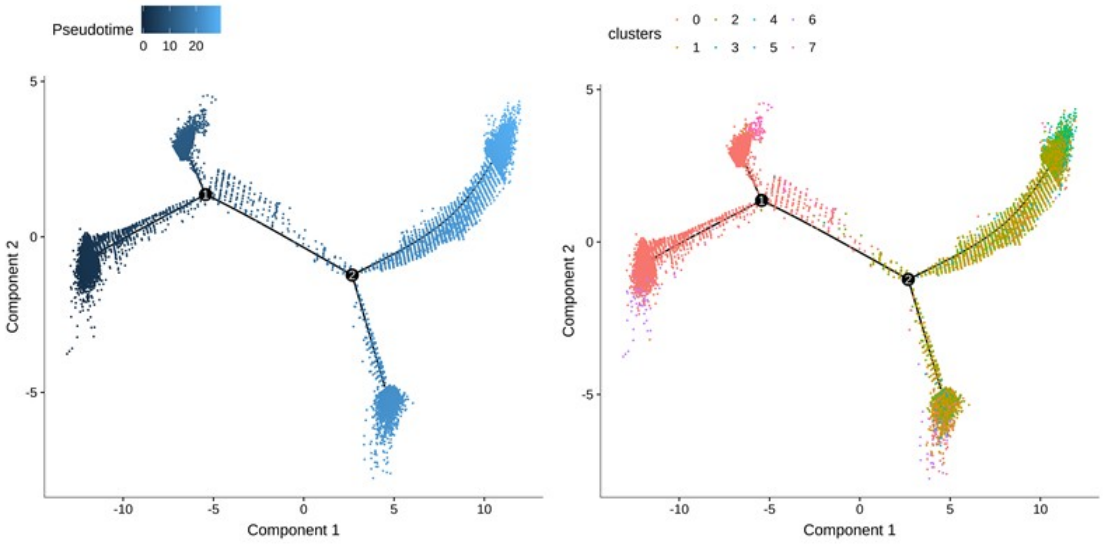
图片说明：亚群细胞类型注释结果在 tSNE 图上的展示，每种细胞类型以不同颜色区分。

2.2.8. 拟时序分析

拟时序（pseudotime）分析[9]，又称细胞轨迹（Cell trajectory）分析，通过拟时序分析可以推断出细胞发育过程的分化轨迹或细胞亚型的演化过程，通过亚群基因的表达特征，描绘细胞分化过程中基因表达的动态过程，揭示分化过程中的关键调控基因。该分析可以推测疾病发生过程中细胞亚型的演变轨迹或细胞凋亡路径，或者推断干细胞在发育过程的分化轨迹或某类细胞的分化来源，在发育相关研究中使用频率较高。

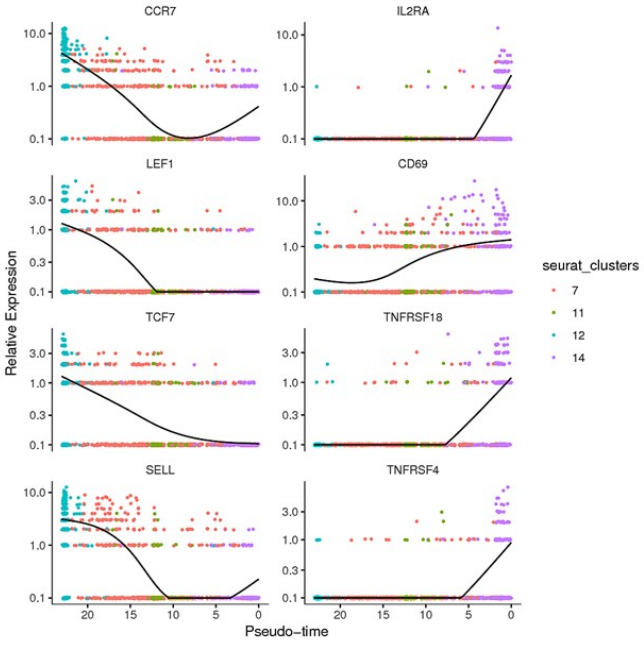
我们使用 Monocle[10] 软件包，基于关键基因的表达模式进行机器学习， 进而模拟出时间发育过程的动态变化。首先挑选出在细胞间基因表达变异程度较大的基因，依据它们的表达谱进行空间降维，再构建最小生成树（minimum spanning tree，MST），之后通过该 MST 找到最长路径代表转录特征相似细胞的分化轨迹。

结果展示如下：



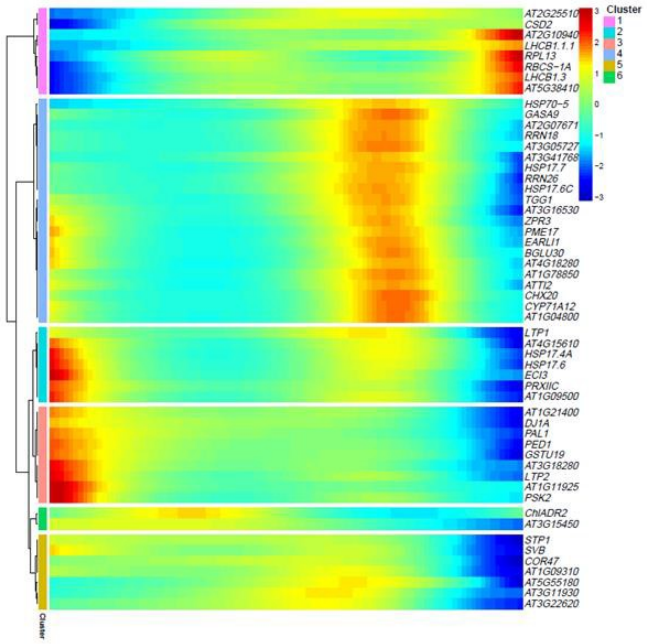
拟时间轨迹图和细胞分化拟时间轨迹图

图片说明：左图为拟时间轨迹图，颜色由浅到深，表示分化的时间由早及晚。右图为细胞分化拟时间 轨迹图，展示了对应左图的细胞群分布情况，可初步推断各细胞群之间的分化关系。



拟时序基因表达量图

图片说明：横轴由左到右表示时间从早到晚，纵轴为基因的表达量，不同颜色的点代表不同细胞群， 可以看到随着时间变化在不同细胞群中基因表达趋势的变化。



拟时序基因表达谱热图

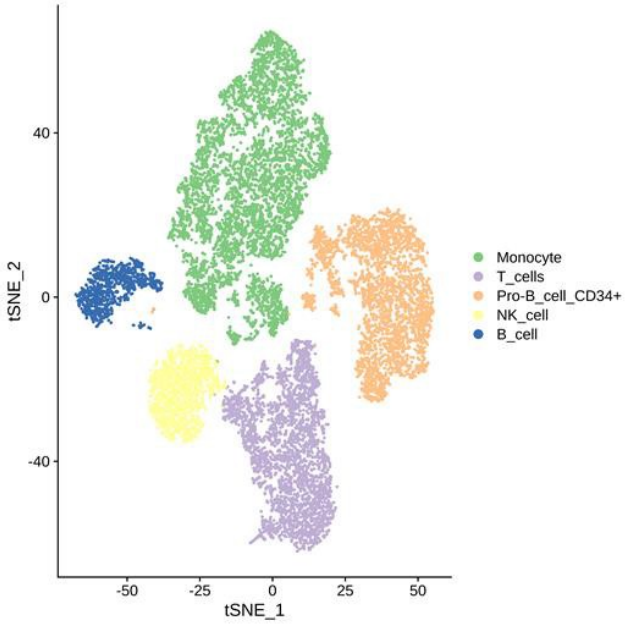
图片说明：蓝色表示低表达，红色表示高表达，并且对时间发育轨迹中具有相似表达模式的基因进行 聚类。

2.2.9. 细胞类型鉴定

在鉴定了每个细胞群的 marker 基因之后我们可以通过 marker 基因来鉴定每个细胞群的真实的细胞类型。目前主流的细胞类型鉴定方法有两种：一是基于已知细胞类型的 Marker 基因人为鉴定这些特定 Marker 所对应的细胞类型，二是基于单细胞参考表达谱数据集鉴定。以上两种方法均可获得细胞类型鉴定结果，但各有优劣，前者受制于目前已有的 Marker 基因与细胞类型的注释，也存在较大的人为主观因素干扰，而基于参考数据集的细胞类型鉴定方式则摒除了研究人员主观因素的干扰，能够有效识别细胞亚型，但同样也会受制于参考数据集的注释来源与数据质量。相对而言，随着目前可获得的单细胞表达谱数据越来越多，后者能够鉴定的细胞类型会越来越精细。就当前已有的经验来说，结合两种方法来鉴定细胞类型，结果最为可靠。

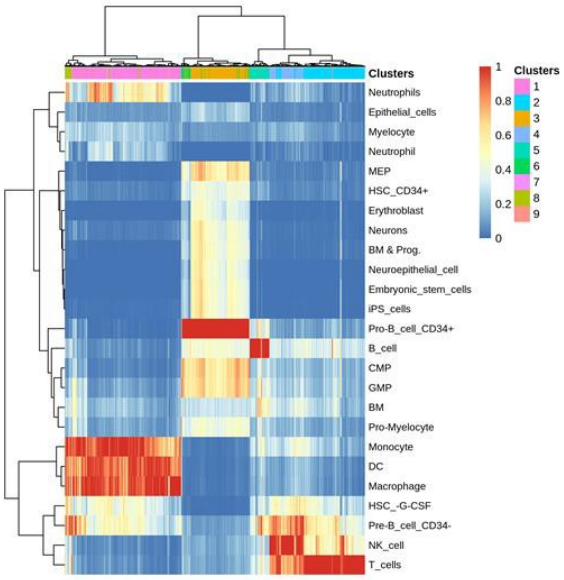
我们使用 SingleR[11] 包提供的基于参考数据集鉴定细胞类型的方法进行细胞类型注释。该方法通过计算单细胞参考表达谱数据集与待鉴定的细胞表达谱之间的相关性，将待鉴定细胞注释为与参考数据集中相关性最高的一种细胞类型。

结果展示如下：



细胞类型鉴定结果

图片说明：细胞类型注释结果在 tSNE 图上的展示，不同的细胞类型以不同颜色区分。



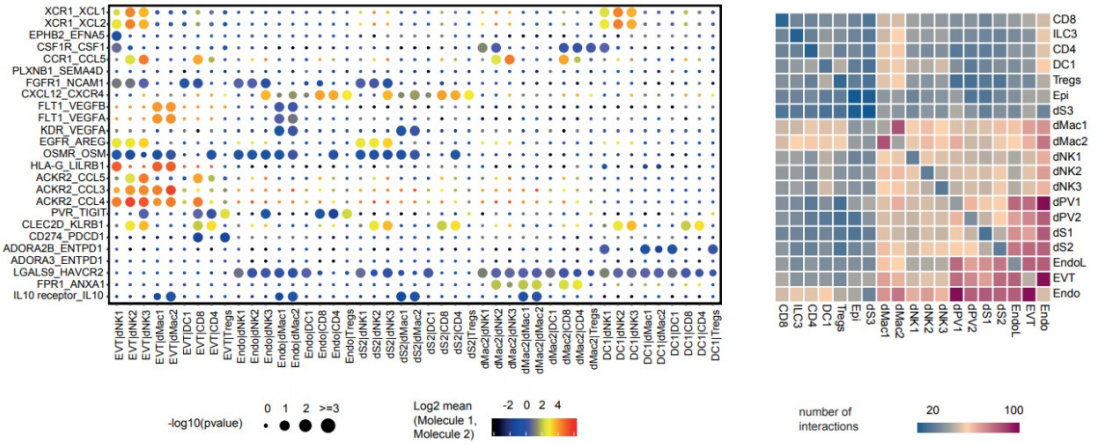
细胞类型鉴定相关性热图

图片说明：纵轴表示每一个待鉴定的细胞，横轴表示参考数据集中的细胞类型注释名称。颜色越红代 表相关性值越大，表明待鉴定的细胞类型与参考数据集中的该细胞类型注释最为相似。

2.2.10. 细胞间通讯

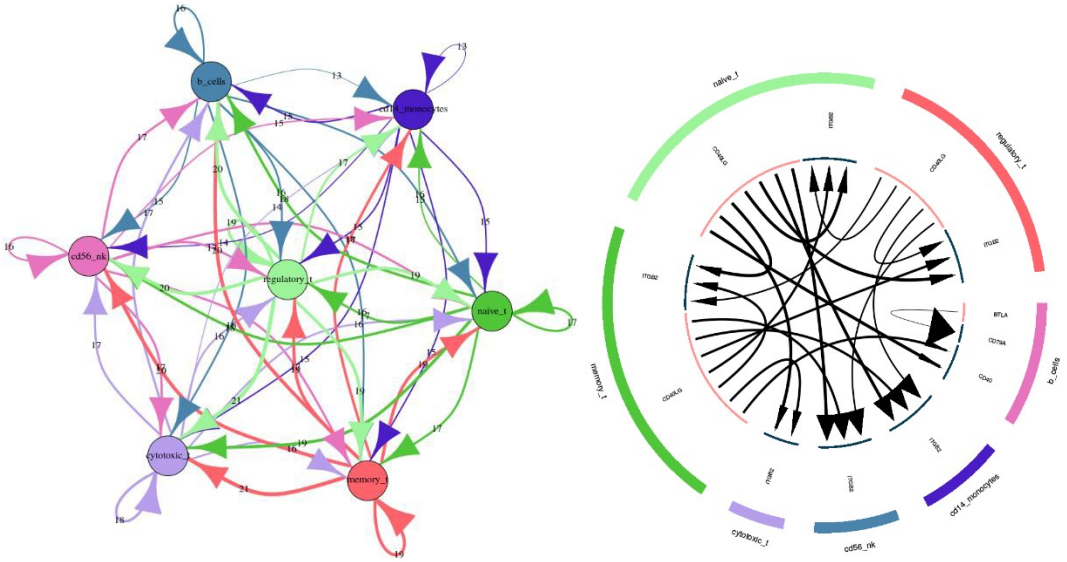
根据不同细胞膜表面和游离的蛋白之间的 ligand-receptor 关系鉴定不同细胞之间可能的相互作用，这种作用包括自分泌和旁分泌。利用单细胞 RNA 测序数据，以细胞亚群的基因表达量数据为研究对象，获取细胞中的配体及受体信息，获得细胞间的信号通讯关系，以阐明在广泛的生物学过程中细胞间通讯的复杂性、多样性和动态性。我们可以捕获高度富集的 ligand-receptor 基因（或转录本）对，通过比较分析来识别细胞的相互作用，并可跟踪样本中细胞间通讯信号的动态变化；通过借助 ligand-receptor 数据库，可自动添加配体受体基因的功能注释，并通过其高效的数据可视化工具，可以以不同的格式来展示输出结果。

结果展示如下：



细胞通讯点图和热图结果

图片说明：左图为点图，行表示具有细胞间通讯关系的 ligand-receptor 对，列表示发生细胞通讯的细胞 类型，圆圈大小表示显著性差异水平，圆圈颜色越红表示细胞间通讯关系越强。右图为热图，行和列都表示为细胞，每格表示细胞间的相关作用的数量，颜色越红表示细胞间的相互作用数目越多，即细胞间通讯关系越强。



细胞通讯网络图和圈图结果

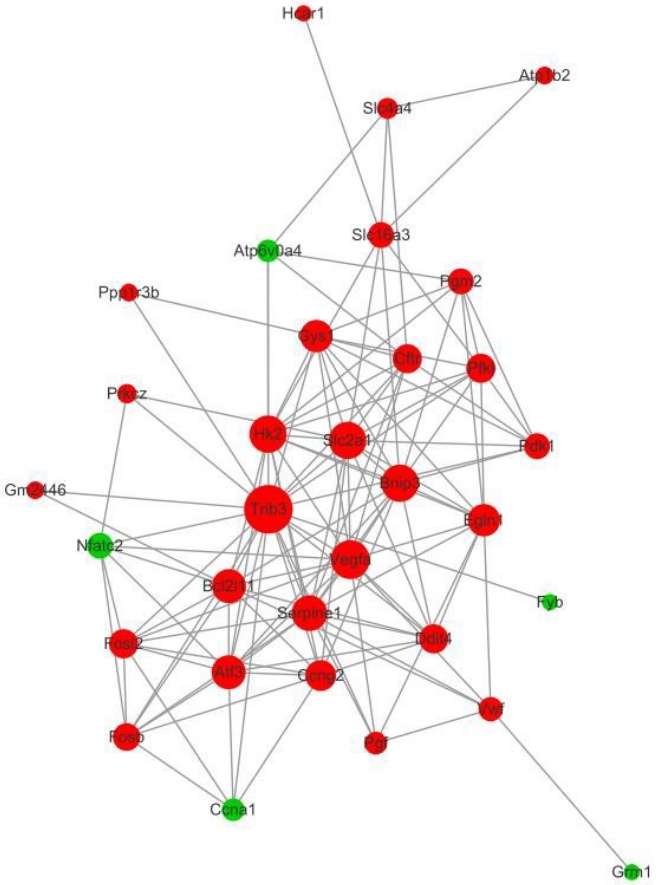
图片说明：左图为网络图，通过分别标记正向（从信号细胞到目标细胞）和反向信号，显示网络中每 种不同的细胞类型之间检测到配体-受体相互作用的数量。它还测量了每个细胞类型内的自分泌信号。网络图的节点按照细胞类型进行颜色编号，边缘按信号和目标细胞之间的相互作用的数量进行标记和缩放。右图为 circos 图，显示了每种相互作用的方向。其中外圈表示细胞类型，内圈表示每个相互作用的配体-受体基因对的详细信息。两者都用不同的颜色进行标记。circos 图中的线段和箭头分别按比例表示配体和受体的相对信号强度，用不同的颜色和线段类型来表示各种可能的变化。每个用户可以选择要在 circos 图中显示的细胞类型、基因类别和相互作用的数量。

2.2.11. 蛋白质相互作用网络预测

在同一细胞的生化过程中，所涉及的蛋白质一般是通过彼此之间的相互作用形成大分子复合物后完成其生物学功能的，例如遗传物质负值、基因表达调控、细胞信号转导、新陈代谢、细胞增殖和细胞凋亡等过程和活动都依赖于蛋白质之间的相互作用。因此，蛋白质相互作用及相互作用网络的研究和分析已成为理解生命活动中细胞组织、过程和功能的基础 [12]。

我们将基因与 STRING[13] 数据库中近缘物种进行 blast 比对，获得差异基因的单细胞关系，选取得分排名前 300，并绘制单细胞网络图。

结果展示如下：



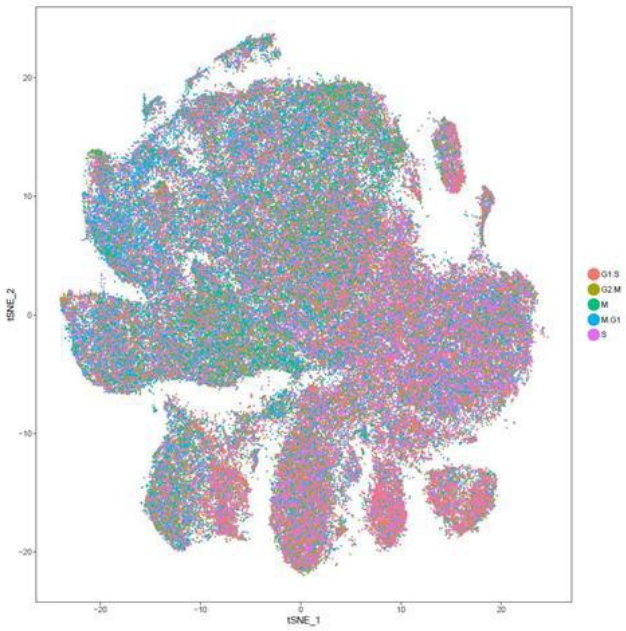
蛋白质单细胞网络预测结果图

图片说明：互作网络中节点(node)的大小与此节点的度(degree)成正比，即与此节点相连的边越多，它 的度越大， 节点也就越大。 节点的颜色与此节点的聚集系数(clustering coefficient)相关，颜色梯度由绿到红对应聚集系数的值由低到高；聚集系数表示此节点的邻接点之间的连通性好坏，聚集系数值越高表示此节点的邻接点之间的连通性越好。边(edge)的宽度表示此边连接的两个节点间的互相作用的关系强弱，互相作用的关系越强，边越宽。

2.2.12. 细胞周期鉴定

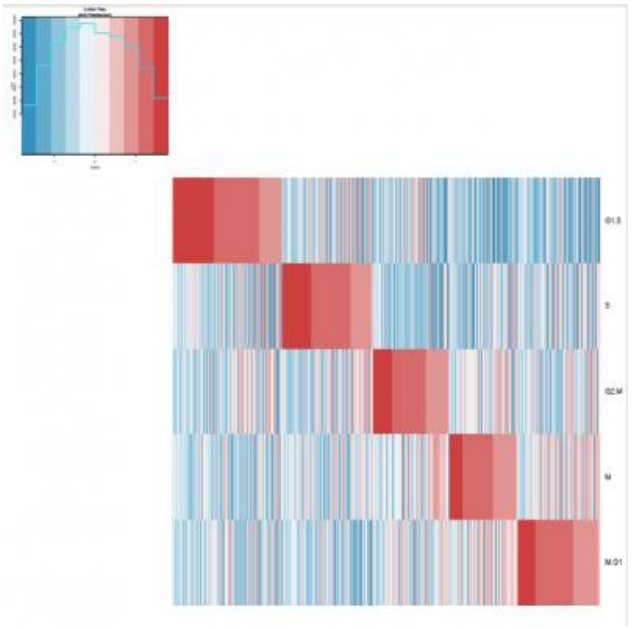
基于 Macosko 等人（2015）[14] 研究获得的细胞周期特征基因表，我们利用 R 包 Dropbead[15] 中的函数来计算细胞周期得分并将细胞分配到 5 个细胞周期阶段：G1.S、S、G2.M、M、M.G1。这种分析方法能够在不需要细胞周期同步的化学处理和高时间分辨率的情况下表征细胞周期基因的表达状况。

结果展示如下：



细胞周期鉴定 tSNE 图

图片说明：5 种颜色代表5 个分裂时期，展现细胞分裂时期在 tSNE 聚类图中的分布情况。



细胞周期鉴定热图

图片说明：列表示细胞，行表示五个细胞周期时期：G1.S、S、G2.M、M、M.G1。

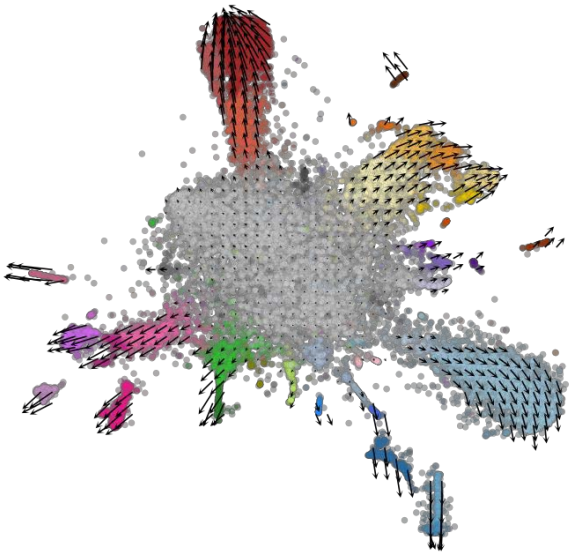
2.2.13. RNA Velocity 分析

RNA 丰度是单个细胞状态的有力指标，但不能直接揭示细胞分化等动态过程。而在生物发育过程中，关键的分化事件通常发生在数小时到数天的时间尺度上。而这些时间尺度与 mRNA 转录、剪接、核输出、翻译和降解的动力学生命周期相当，典型的半衰期为 10 小时左右，并且单个细胞中将同时存在新合成的 mRNA、成熟的 mRNA 或者完全降解的 mRNA。通过区分不同生命周期阶段的 mRNA 分子，我们可能能够观察到单个细胞的当前（mRNA 目前被翻译成蛋白质）、过去（mRNA 正在被降解）和未来（新转录的 mRNA）状态。此外，RNA 生命周期不同阶段的 mRNA 分子具有不同的特征，例如：新转录的 mRNA 分子含有内含子，剪接但不输出的 mRNA 驻留在细胞核内，当前翻译的 mRNA 与核糖体结合，降解的 mRNA 部分片段化。这些特征在原则上可用来测量转录速度（RNA velocity），即细胞中 mRNA 分子丰度的变化率[16]。

但在研究中发现，细胞内的降解信号较弱且易受干扰。而剪接则很容易被观察到，例如跨外显子/内含子的 reads 可表示未剪接的 pre-mRNA，跨外显子/外显子的 reads 可表示剪接的成熟 mRNA。研究证明，RNA 速度可以通过未剪接和剪接 mRNA 丰度来估计，它可以在数小时的时间尺度上预测单个细胞的未来状态 [16]。RNA 速度可以揭示单细胞基因表达在时间尺度上的动态变化，这种动态变化与人类和其他哺乳动物的发育、再生和反应过程相匹配。它是一种局部速度向量，可用于模型定型、命运选择和体内转录的精确动力学研究，能够详细研究复杂组织和器官的动态过程，并将极大地促进人类胚胎的谱系分析。

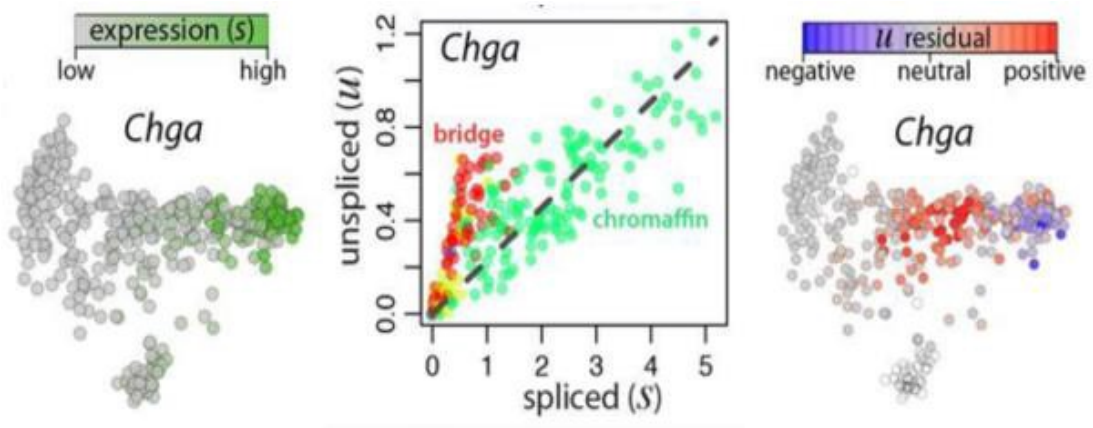
我们采用 velocyto 算法，通过计算细胞内 mRNA 剪接前后比例来估计 RNA 丰度随时间的变化，推断细胞的下一个可能的分化方向。其中，在稳态条件给定转录率 α，此时，u=s*γ，其中 u=U/N，s=S/N，N 代表总 reads counts，U 代表 unspliced reads counts ，S 代表 spliced reads counts，γ 表示退化率（对角线虚线）；稳定状态不同的转录率 α 落在对角线 γ 上。unspliced mRNA 的丰度超过 s/γ 的比例时，表明基因表达处于诱导状态，而 unspliced mRNA 低于该 比例是表明基因表达处于受抑制状态。

结果展示如下：



RNA velocity 分析结果图

图片说明：图中箭头方向代表算法预测的细胞分化方向。



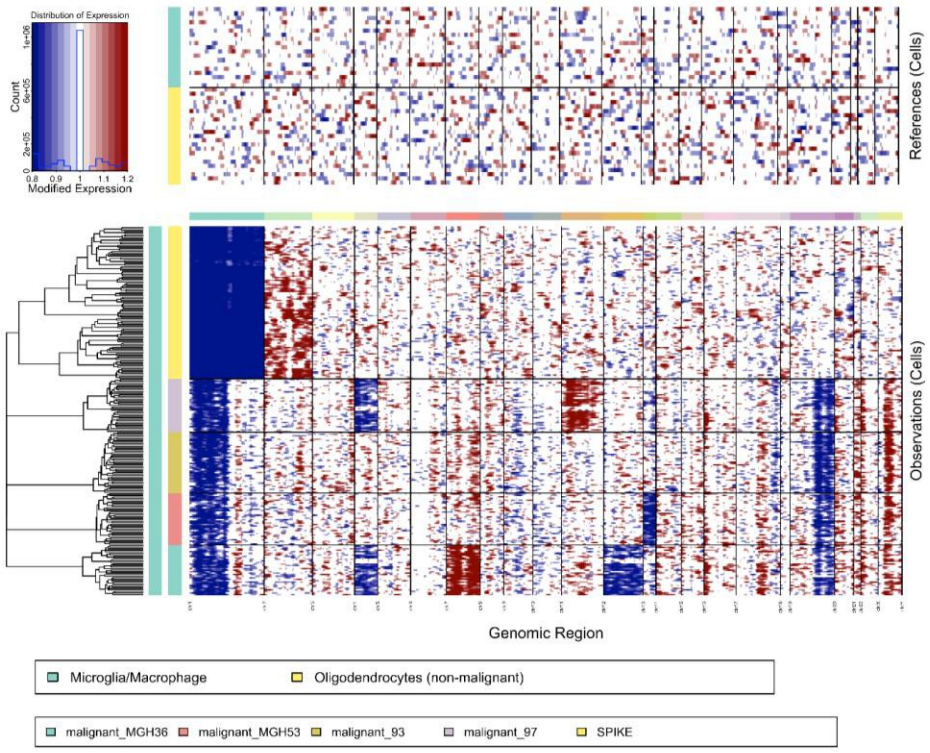
某基因的 RNA velocity 分析结果

图片说明：左图展示的是 Chga 基因的 spliced mRNA 在细胞聚类图谱上的表达模式图，颜色越绿表示 表达量越高。中图显示了相对于未剪接/剪接的 mRNA 平衡态，Chga 基因表达水平的相位图，在桥细胞（红 色和黄色）中其 unspliced mRNA 表达丰度相对于稳态时增加，说明该基因在分化过程中处于诱导状态。右图表示 u 残差在细胞聚类图谱中的分布，可以看到 Chga 的 unspliced mRNA 的在对应的细胞类群中的表达状态，与中图显示的表达状态相对应。三图中每个点都表示一个细胞。

2.2.14. 拷贝数变异（CNV）分析

InferCNV（https://github.com/broadinstitute/inferCNV）[17] 主要用来从肿瘤单细胞转录组数据中研究体细胞染色体大片段变异，包括整条染色体的重复与缺失以及染色体大片段的插入和缺失。原理是通过比较基因在肿瘤细胞基因组和一系列“正常”细胞参考基因组上的基因表达密度来确定，然后通过热图对每条染色体上基因的相对表达密度进行可视化，在以正常细胞作为参考的情况下可以直观地判断肿瘤基因组上的哪些区域是过表达和低表达。为了提高推断结果的可靠性，inferCNV 采用几种 residual expression 过滤方法减少数据中噪声信号，从而发现真正的 CNV 的信号。另外，inferCNV 还可以预测 CNV 区域，并根据基因表达的异质性聚类细胞。

结果展示如下：



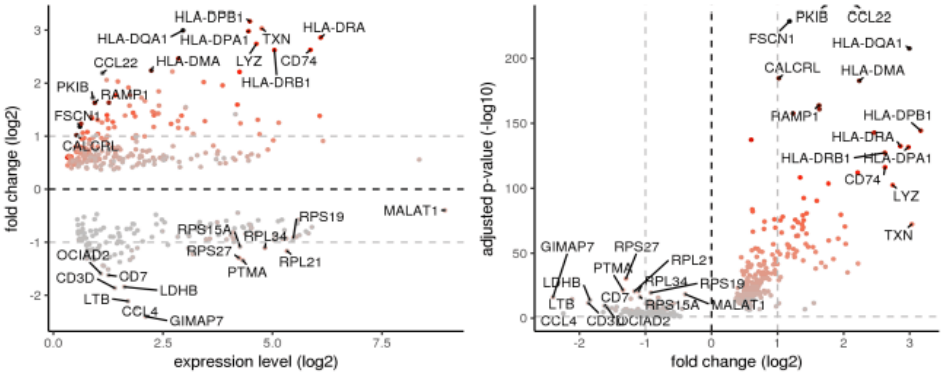
拷贝数变异（CNV）分析结果图

图片说明：底部的热图显示参考细胞中基因表达水平变化，底部的热图显示了包含层次聚类排序的实 际分析样本细胞的基因表达变化，染色体上的基因从左向右排列。将正常细胞表达的有效数据从样本测序的表达数据中减去产生差异，其中红色区块表示染色体区域扩增，蓝色区块表示染色体区域缺失。两热图中的行表示细胞，列表示按染色体位置排列的基因。两热图中间的颜色条表示不同的染色体区域。

2.2.15. 差异基因表达分析

关于表达数据的一个常见问题是在两个实验条件之间基因表达是否存在差异。相对于 bulk 水平的基因差异分析，我们可以通过在细胞类群内进行差异基因表达分析来解释在单细胞水平的细胞异质性问题，发现在特定实验条件下，单个细胞是如何发生转录反应的。

我们使用 MAST 差异检验方法，根据差异倍数（FoldChange）及差异显著性检验（pvalue）结果筛选差异表达基因。



差异基因的 MA 图和火山图

图片说明：左图为基因差异表达的 MA 图。纵坐标表示差异倍数，横坐标表示表达水平。以差异倍数 为 0 时作为水平线，在水平线之上的为上调表达的基因，水平线之下为下调表达的基因。颜色越深表达越 显著。右图为基因差异表达的火山图。纵坐标为 p-value 值，横坐标表示差异倍数。以差异倍数为 0 时作为标准，左侧为下调表达的基因，右侧为上调表达的基因。颜色越深表达越显著。

2.2.16. 差异基因功能分析

得到差异表达基因之后，对差异表达基因进行 GO 富集分析和 KEGG 富集分析，对其功能进行描述。

1） GO 富集分析

首先，统计每个 GO 条目中所包括的差异基因个数，并用超几何分布检验方法计算每个 GO 条目中差异基因富集的显著性。计算的结果会返回一个富集显著性的 p 值，小的 p 值表示差异基因在该 GO 条目中出现了富集。

$$p = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \qquad Enrichment\ score = \frac{\frac{m}{n}}{\frac{M}{N}}$$

超几何分布检验计算 p 值的公式和 Enrichment score 计算公式

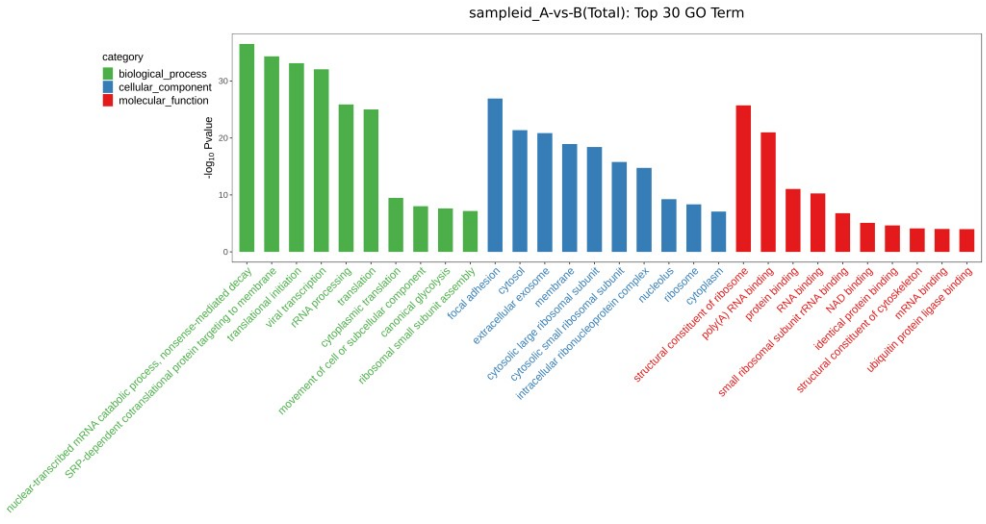
图片说明：其中，N 为所有基因中具有 GO 注释的基因数目；n 为 N 中差异表达基因中具有 GO 注释的基因数目；M 为所有基因中注释为某特定 GO Term 的基因数目；m 为注释为某特定 GO Term 的差 异表达基因数目。可以根据 GO 分析的结果结合生物学意义从而挑选用于后续研究的基因。

然后，使用 Fisher 算法分别对样品间差异基因进行 CC、BP、MF 富集分析，并使用 topGO （http://bioconductor.org/packages/release/bioc/html/topGO.html）对富集到的 Term 绘制有向无环图。topGO 有向无环图能直观展示差异表达基因富集的 GO 节点 (Term) 及其层级关系，是差异表达基因 GO 富集分析的结果图形化展示，分支代表的包含关系，从上至下所定义的功能描述范围越来越具体。

根据功能分级，一般将 GO 分为三个层级，level1 包含三个条目：biological process、cellular component 和 molecular function，level2 包含 biological adhesion、cell 和 binding 等 64 个条目，level3 即为常规富集使用的数万个条目。从 level1 到 level3 功能更具体，反之，更概括。我们一般展示差异基因在 GO Level2 水平的分布比较图。

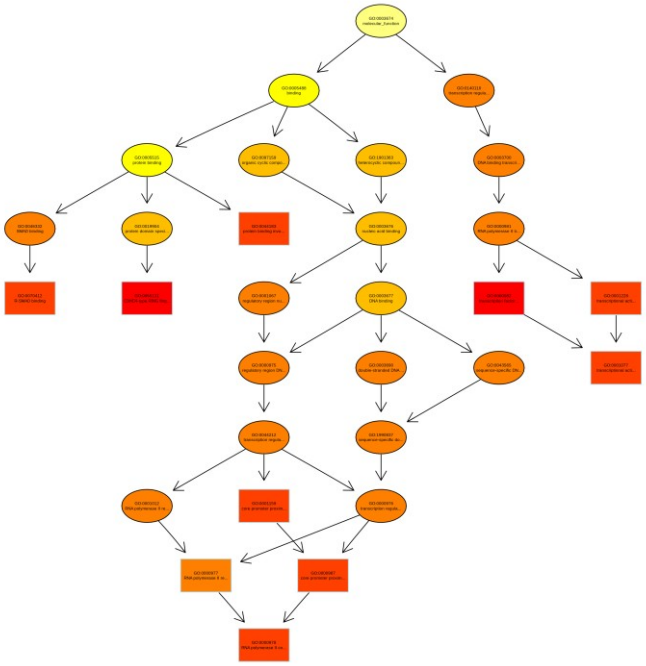
结果展示如下：

GO 富集分析 top30 （筛选三种分类中对应差异基因数目大于 2 的 GO 条目，按照 每个条目对应的 -log10pValue 由大到小排序的各 10 条）条形图展示如下：



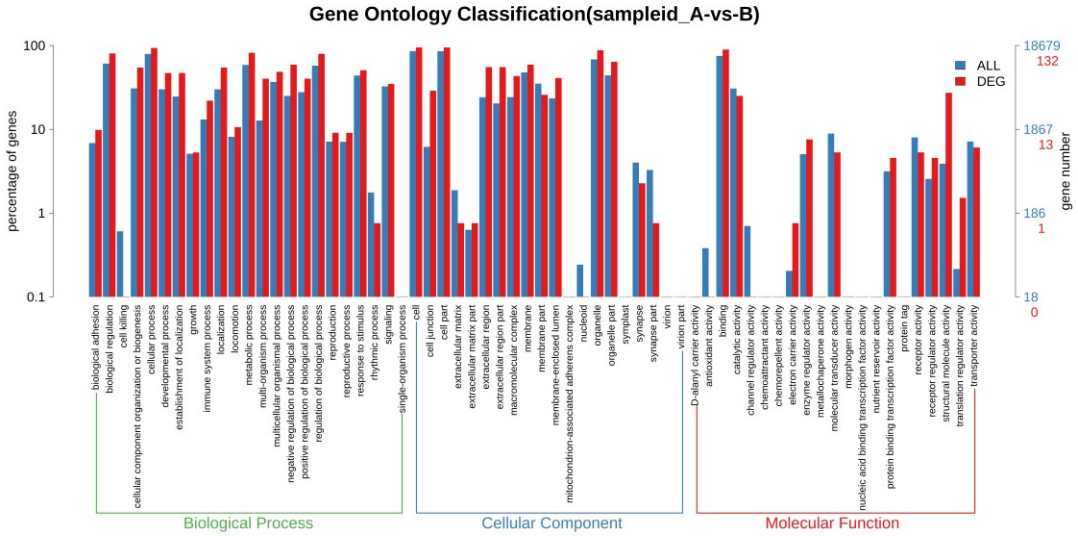
GO 富集分析结果展示

图片说明：图中横轴为 GO 条目名称，纵轴为 -log10pValue。



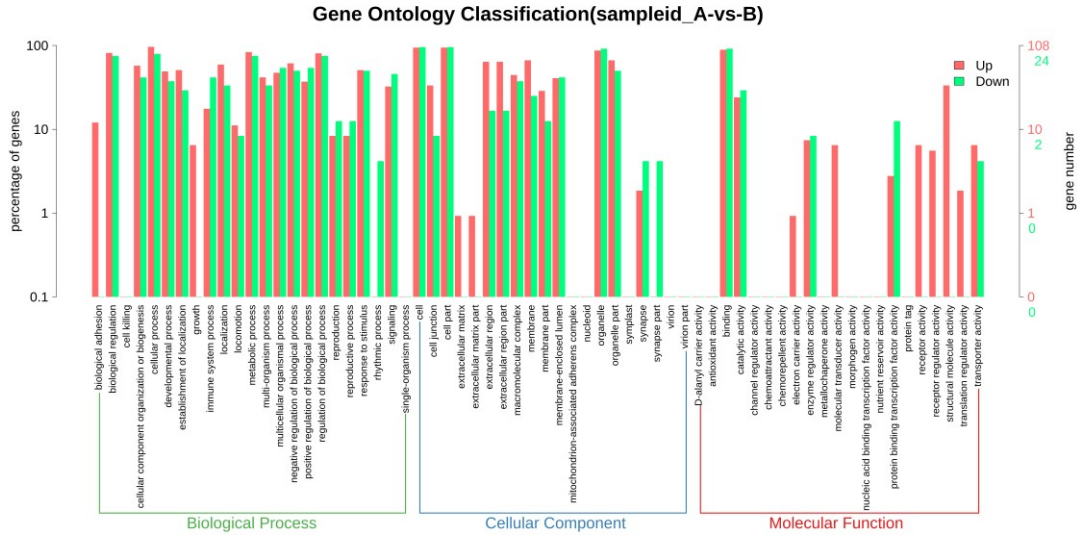
差异基因 topGO 有向无环图展示

图片说明：对每个 GO Term 进行富集，最显著的 10 个节点用矩形表示。矩形的颜色代表富集显著 性，从黄色到红色显著性越来越高。每个节点的基本信息显示在相应的图形中，为 GO ID 和 GO Term。



差异表达基因及所有基因在 GO Level2 水平分布比较图

图片说明：蓝色表示所有基因富集的 GO Level2 条目，红色表示差异基因富集的 GO Level2 条目，横轴为条目名称，纵轴表示对应条目的基因数量和其百分比。



上调差异基因和下调差异基因在 GO Level2 水平分布比较图

图片说明：红色表示上调差异表达基因富集的 GO Level2 条目，绿色表示下调差异表达基因富集的 GO Level2 条目，横轴为条目名称，纵轴表示对应条目的基因数量和其百分比。

2） KEGG 富集分析

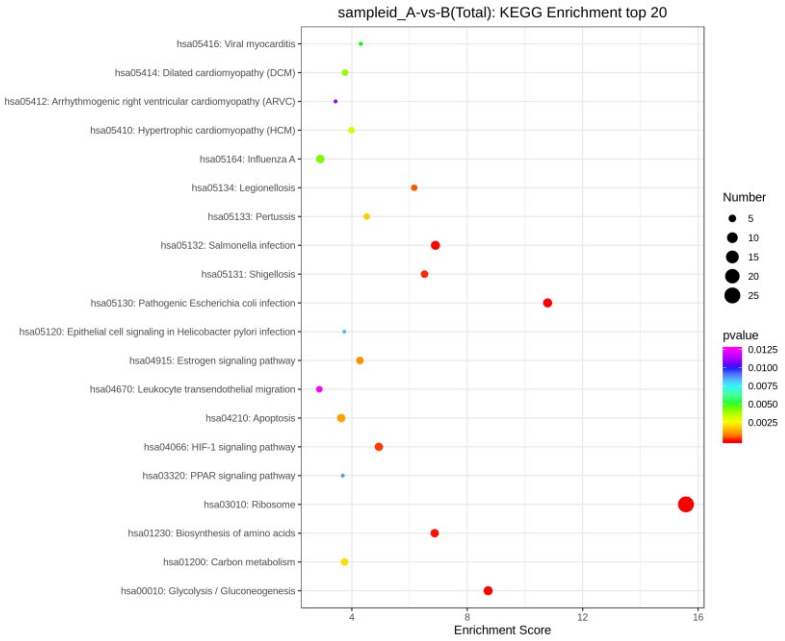
KEGG[18] 是有关 Pathway 的主要公共数据库，利用 KEGG 数据库对差异蛋白编码基因进行 Pathway 分析（结合 KEGG 注释结果），并用 超几何分布检验的方法计算每个 Pathway 条目中差异基因富集的显著性。计算的结果会返回一个富集显著性的 p 值，小的 p 值表示差异基因在该 Pathway 中出现了富集。相应的计算公式参见 GO 富集分析。Pathway 分析对实验结果有提示的作用，通过差异基因的 Pathway 分析，可以找到显著富集差异基因的 Pathway 条目，寻找不同样品的差异基因可能和哪些细胞通路的改变有关。

根据功能分级，通常将 KEGG 分为三个层级，level1 包含六个分类：Metabolism、Genetic Information Processing、Environmental Information Processing、Cellular Processes、Organismal Systems 和 Human Diseases（具体物种注释可能有删减）。level2 包含 Cell

growth and death、Transcription 和 Development 等 44 个分类(具体物种注释可能有删减)，level3 即为常规富集使用的数百个 Pathway，从 level1 到 level3 功能更具体，反之，更概括。

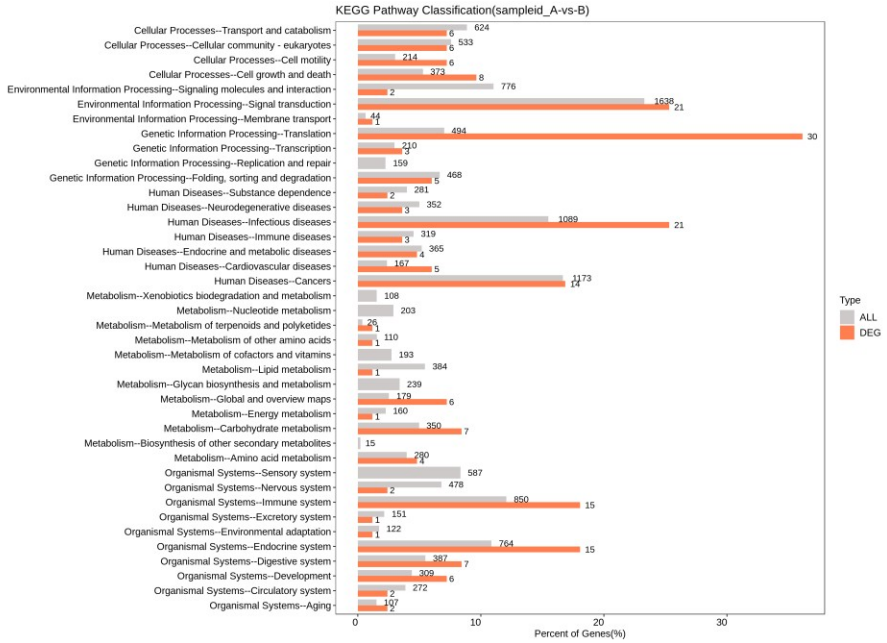
结果展示如下：

KEGG 富集分析 top20（筛选对应差异基因数目大于 2 的 Pathway 条目，按照每个条目对应的 -log10Pvalue 由大到小排序）气泡图如下：



KEGG 富集 top20 气泡图

图片说明：图中横轴 Enrichment Score 为富集分值，气泡越大的条目包含的差异蛋白编码基因数目越多，气泡颜色由紫-蓝-绿-红变化，其富集 pValue 值越小，显著程度越大。



差异表达基因及所有基因在 KEGG Level2 水平分布比较图

图片说明：横轴是注释到各 Level2 通路的基因（差异表达基因）和所有注释到 KEGG 通路的基因（差异表达基因）总数的比值（%），纵轴表示 Level2 Pathway 的名称，柱子右边数字代表注释到该 Level2 Pathway 下的差异表达基因数量。

上调差异表达基因及下调差异表达基因在 KEGG Level2 水平分布图

图片说明：横轴是注释到各 Level2 通路的上调（下调）差异表达基因和所有注释到 KEGG 通路的上调（下调）差异表达基因总数的比值（%），纵轴表示 Level2 Pathway 的名称，柱子右边数字代表注释到该 Level2 Pathway 的上调（下调）差异表达基因数量。

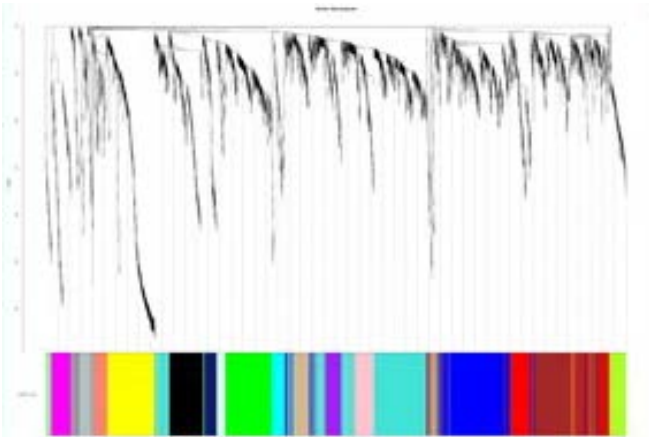
2.2.17. 加权基因共表达网络分析（WGCNA）

基因共表达网络(gene co-expression network)分析致力于寻找协同表达的基因模块(module)，并探索基因网络与研究者关注的表型之间的关联关系。它基于高通量的微阵列技术(microarray)或者 RNA-Seq 技术，应用检测得到的实验数据，从转录组(基因)水平探索基因网络与疾病或者研究者关注的性状之间的关联关系，因此应用于复杂疾病的易感基因鉴定、新药开发等生物医学研究领域。

加权基因共表达网络构建(weighted gene co-expression network，WGCNA)[19]，是一种从高通量的表达数据中挖掘模块（module）信息的算法。在该算法中 module 被定义为一组具有类似表达谱的基因，如果某些基因在一个生理过程或不同组织中总是具有相类似的表达变化，那么我们有理由认为这些基因在功能上是相关的，可以把他们定义为一个模块(module)。该算法作为一种高效、准确的生物信息学生物数据挖掘方法，理论不断完善，应用日渐广泛，主要包括不同器官或组织类型发育调控、同一组织不同发育调控、非生物胁迫不同时间点应答、病原菌侵染后不同时间点应答等研究方向。

我们采用 R 语言中的 WGCNA 包完成数据分析。从方法上讲，WGCNA 分析分为表达量聚类分析和表型关联两部分，主要包括基因质检的相关系数计算、加权、模块确定、模块与性状关联、模块核心基因挖掘等步骤。

结果展示如下：



模块识别分析结果图

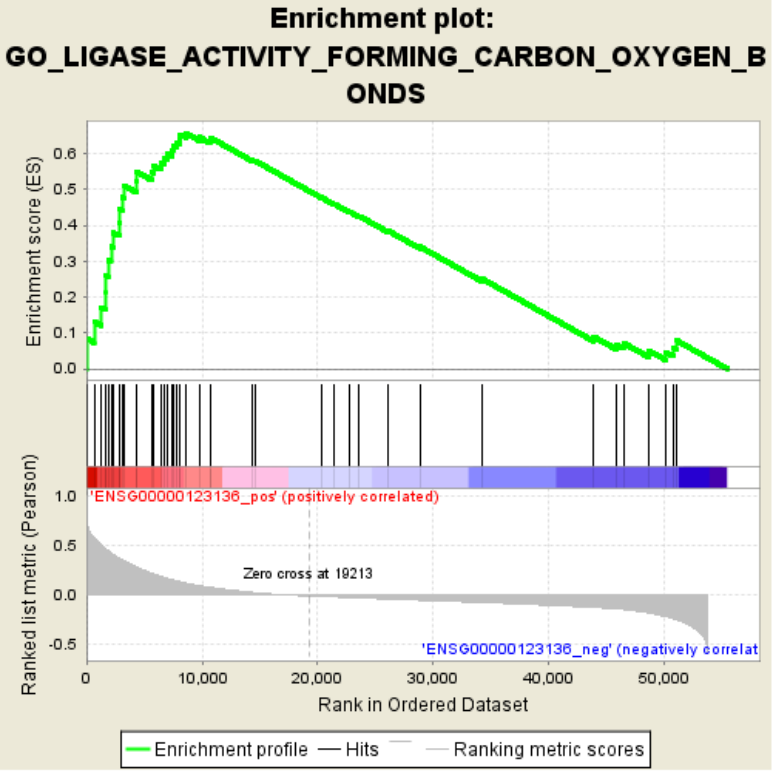
图片说明：图中上部为对加权后的相关系数构建的 dissTOM 矩阵构建的基因聚类树，图中下部为每个 模块基因的分布情况，同一种颜色表示同一个模块，Dynamic Tree Cut 颜色为利用 dynamicTreeCut 方法识别得到的模块，由于某些模块之间存在一定的相关，则将对应的模块合并为同一个模块，即下方的 Merged Dynamic 为最终得到的模块。

2.2.18. 基因集富集分析

（1）GSEA 分析

基因集富集分析（Gene Set Enrichment Analusis，GSEA）[20]，是一种基于基因集的富集分析方法。其基本思想是使用预定义的基因集（通常来自功能注释或先前实验的结果），将基因按照在两类样本中的差异表达程度排序，然后检验预先设定的基因集合是否在这个排序表的上部或下部富集，从而判断此基因集内基因的协同变化对表型变化的影响。

结果展示如下：

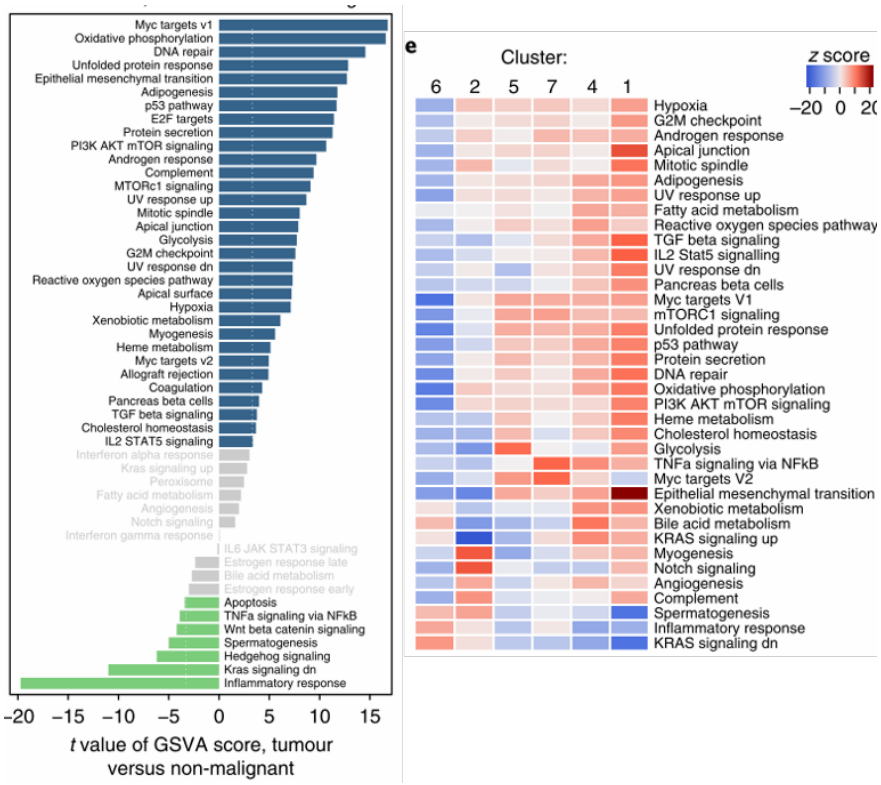


GSEA 分析结果图

图片说明：GSEA 结果图分 3 个部分，图片最上部为基因富集分数（enrichment score， ES）折线图， 横轴为该基因下的每个基因，纵轴为对应的ES 值，在折线图中有个峰值，该峰值就是这个基因集的ES 值，峰值之前的基因就是该基因集下的核心基因。中间部分为用线条标记位于该基因集下的基因hit，每一条线代表基因集中的一个基因，及其在基因列表中的排序位置。最下部为所有基因的rank 值分布图， 其中上部为基因与表型关联的矩阵，红色表示与第一个表型（class A）正相关，在class A 中高表达，蓝色表示与第二个表型（class B）正相关，在class B 中高表达。

（2）GSVA 分析

基因集变异分析（Gene Set Variation Analysis, GSVA）[21]，是一种非参数的无监督分析方法。该分析主要通过将基因在不同样品/不同细胞群间的表达量矩阵转化成基因集在样品/细胞群间的表达量矩阵，从而来评估不同的代谢通路在不同样品间/不同细胞群间是否富集。结果展示如下：



GSVA 分析结果图

图片说明：通过 GSVA 分析显示每个细胞信号通路活性评分差异。

参考文献

1. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

2. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, (2019).

3. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA sequencing data are corrected by matching mutual nearest neighbours. *Nat. Biotechnol.* **36**, 421–427 (2018).

4. Pearson, K. LIII. *On lines and planes of closest fit to systems of points in space*. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **2**, 559–572 (1901).

5. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

6. Coifman, R. R. *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 7426–7431 (2005).

7. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* (2018).

8. Aguilar, J. S., Ruiz, R., Riquelme, J. C. & Giráldez, R. SNN: A Supervised Clustering Algorithm. in *Engineering of Intelligent Systems* (eds. Monostori, L., Váncza, J. & Ali, M.) 207–216 (Springer Berlin Heidelberg, 2001).

9. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).

10. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

11. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).

12. Zhang, A. Protein Interaction Networks: Computational Analysis. *Protein Interact. Netw. Comput. Anal.* (2009). doi:10.1017/CBO9780511626593

13. Snel, B., Lehmann, G., Bork, P. & Huynen, M. A. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* **28**, 3442–3444 (2000).

14. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).

15. Alles, J. *et al.* Cell fixation and preservation for droplet-based single-cell transcriptomics.

a) *BMC Biol.* **15**, 44 (2017).

16. Manno, G. L. *et al.* RNA velocity in single cells. *bioRxiv* 206052 (2017). doi:10.1101/206052

17. Puram, S. V. *et al.* Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171**, 1611-1624.e24 (2017).

18. Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, D480–D484 (2008).

19. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

20. Subramanian, A. *et al.* From the Cover: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545 (2005).

21. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* **14**, 7 (2013).

特别申明

本项目报告由北京吉康医学科技有限公司提供给项目相关客户。本公司承诺：未经客户同意，不向第三方泄露数据及数据分析内容，不将客户数据用于任何商业行为（遵循合同保密协议）。客户未经本公司同意，不得以任何目的向第三方出示项目报告。本报告的最终解释权归北京吉康医学科技有限公司。