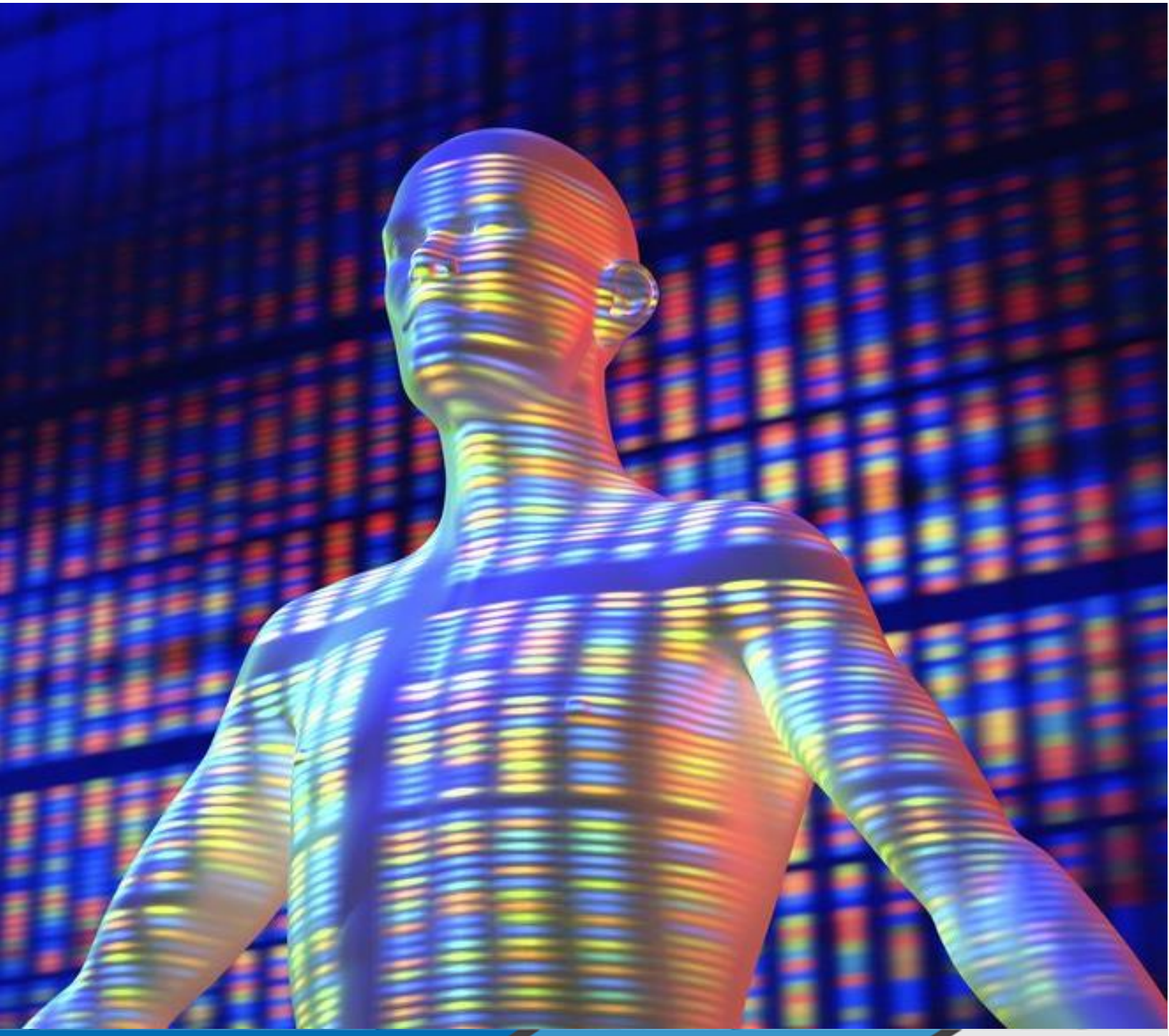


Solution Brief

大内存软件加速单细胞 RNA 测序解决方案

MemVerge

介绍

现代基因测序技术可以并行分析数百万个遗传物质片段，从而以高通量产生结果。越来越多的关注点是如何利用单细胞 RNA 测序（以下简称 scRNA-seq）技术了解单个细胞之间的差异基因活性。

基因是在 DNA 的特定区域中编码的核苷酸碱基对的独特序列。基因并非全部是激活的。当一个基因被激活（即表达）时，其编码序列被转录到信使 RNA（mRNA）上。一个标本内 mRNA 分子的完整集合称为转录组。基因越活跃（表达量越高），该基因在转录组测序过程中被检测到的次数就越多。

测序后，可以通过比对碱基对序列，计算每个细胞每个基因的检测次数。这一汇总的数据随后被输入计算管线进行下游分析。与其他机器学习问题相似，单细胞 RNA 测序的计算管线步骤繁多，耗时冗长，对内存和存储容量的要求也同样高。

这些计算过程执行时间很长，对底层计算资源，特别是内存（DRAM）和存储提出了严格的要求。易失性的 DRAM 模块容量有限，需要在计算操作开始之前批量加载数据。持久内存（PMEM）是最近的一项技术突破，可在非易失性 DIMM 中实现大容量内存。Memory Machine 是 MemVerge 公司研发的软件，可与 PMEM 一起使用，提供具有数据服务功能的大容量内存池，从而消除 scRNA-seq 分析途径中的许多批量加载。其结果是过程的执行时间显著减少。

Memory Machine 通过内存虚拟化和管理层，将传统的 DRAM 与持久内存（PMEM）集成，这是非易失性内存的最新发展。基于 PMEM 的半导体技术允许供应商将 PMEM 打包安装到 DDR4 兼容模块中，这些模块的容量明显高于 DRAM，而不会增加相应的成本。虽然 与 DRAM 相比，PMEM 具有较高的延迟，但 Memory Machine 能够构建内存分层结构，从而在整个内存池中恢复 DRAM 的平均性能。内存快照可以拍摄内存状态，而不会产生任何 I/O 或消耗额外的内存，使快照过程效率很高。当将中间阶段的结果写入持久性存储，然后在随后的阶段将这些数据重新加载到内存时，内存快照的好处就显露无疑了，因为这些数据一直驻留在内存中。

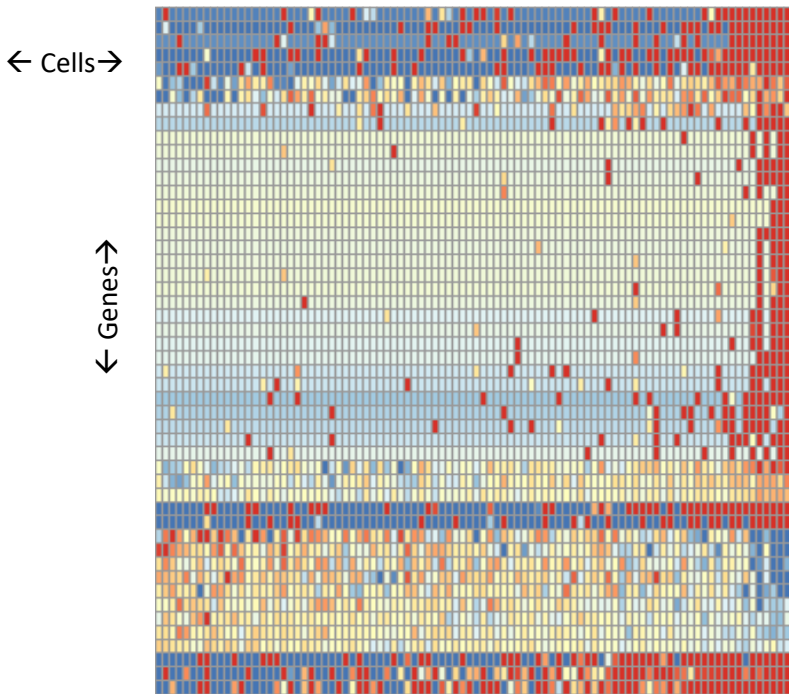
单细胞测序: 追踪单个细胞的行为

高通量测序正被越来越多地用于分析单细胞中的基因表达，例如，研究稀有细胞、异质性群体中的细胞，以及同质性群体中单个细胞基因表达的变化。具体来说， scRNA-seq 可用于追踪发育中细胞的轨迹，即可以观察单个细胞的行为，而不是整个群体的平均值。虽然确切的数据和分析手段取决于感兴趣的生物问题，但大多数分析都遵循类似的工作流程。

原始数据

在 scRNA-seq 过程中，将单细胞中的 mRNA 片段化并转换成标记的 mRNA 片段库，然后进行测序。输出通常采用二维矩阵（表达量矩阵）的形式，其中行表示研究所感兴趣的基因，列表示样本中的各个细胞。矩阵条目是每个基因在每个细胞中被鉴定的次数的计数，这是基因表达的相对量度。表达量矩阵作为热图的可视化表示如图 1 所示。

图 1. 用热图表示的表达量矩阵



虽然确切的数据和分析手段因感兴趣的生物问题不同而有差异，但大多数分析都遵循类似的工作流程。

scRNA-seq 工作流

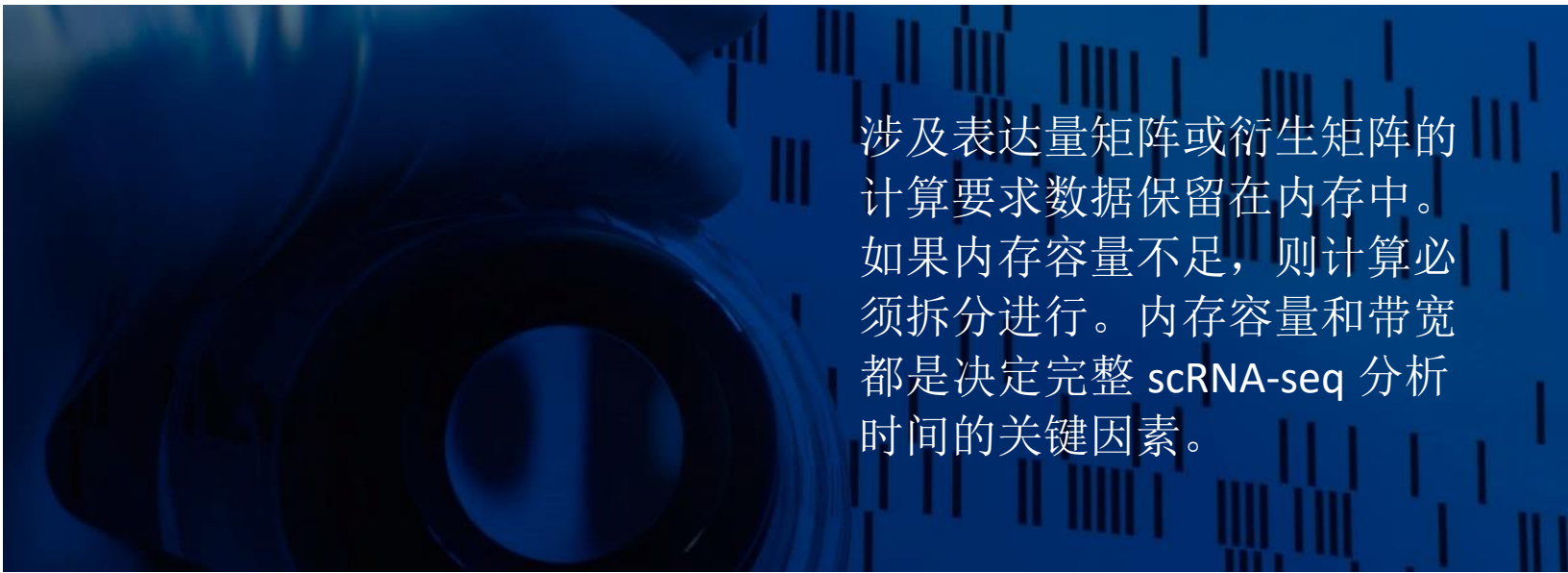
用于分析 scRNA-seq 数据的工具数据库（目前有超过 800 种工具）已可公开获得。超过 80% 的工具使用 R 或 Python 作为基础平台。表达量矩阵生成完毕后，后续工作流程通常有以下几个步骤。

1. **质量控制 (QC).** scRNA-seq 的数据本质上是有噪声的，这是由样品的准备和细胞间随机变异造成的。QC 的工作就是剔除可能已损坏的细胞数据，或去除不完整的测序读数。
2. **标准化.** 将单个细胞表达量标准化，以消除细胞特异性偏差，以便可以在后续步骤中进行跨细胞的直接比较。
3. **特征选择.** 其目的是通过只保留可能对细胞异质性有影响的基因来减少其他无关基因的噪声。
4. **降维.** 通过将特征转换为仍能捕获大多数数据关系的较少数量的因子，可以压缩数据并进一步降低噪声。主成分分析（PCA）通常用作初始步骤，然后进行更积极的修剪。
5. **聚类.** 细胞根据其（标准化）基因表达向量的相似性分群。这些分群反映了不同的生物学状态。

具体的计算任务数量和步骤划分取决于所使用的 scRNA-seq 分析工具：一些工具仅处理该过程中的单个步骤；一些工具能进行多个步骤的处理。根据所使用的工具包，上述工作流程可以进一步细分（例如，在下面描述的研究中，使用了 11 个步骤）。

当数据量大于内存容量

随着单细胞分析技术的改进，单次实验所解析的细胞数量不断增加，表达量矩阵的大小也随之增加。与机器学习的惯例一致，分析中的几个步骤需要反复迭代，以便调整模型参数。迭代要求将前一步骤保存的数据重新加载到内存中，以便成为下一个周期的输入。此外，涉及表达量矩阵或衍生矩阵的计算要求数据保留在内存中。如果内存容量不足，则计算必须拆分进行。内存容量和带宽都是决定完整 scRNA-seq 分析耗时的关键因素。

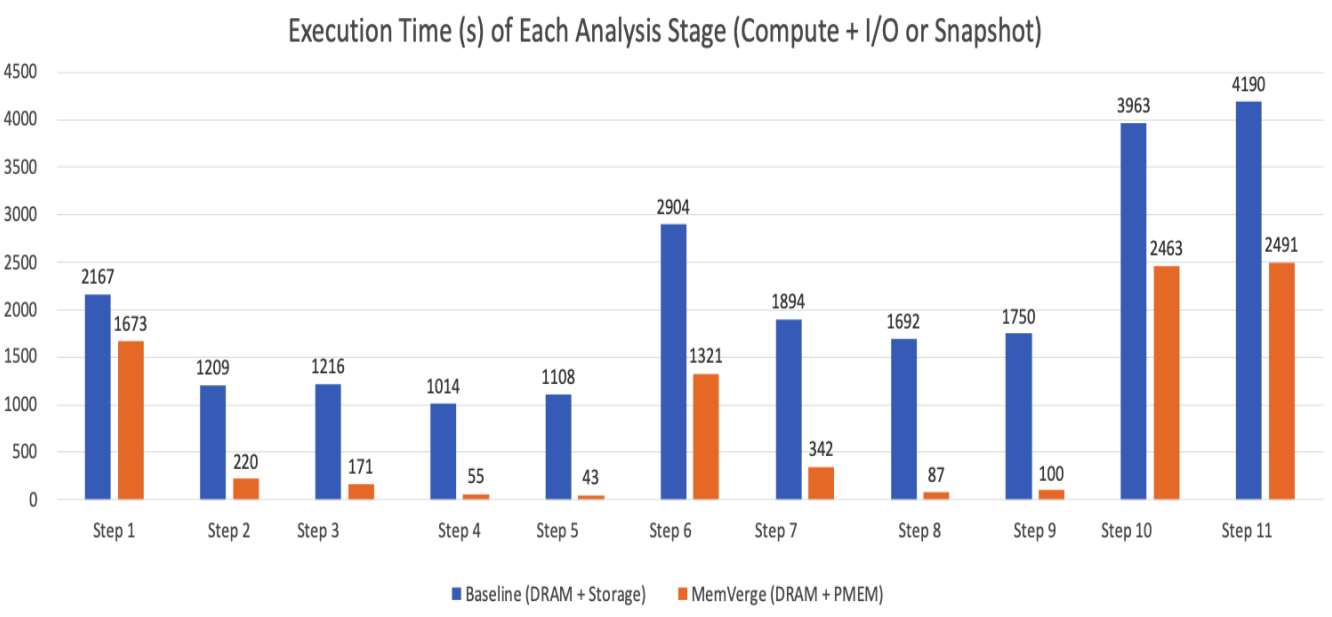


scRNA-seq 数据加载速度提高 1000 倍，执行速度提高 25 倍

为了研究 Memory Machine 对 scRNA-seq 分析（小鼠细胞图谱，表达量矩阵大小 31787 x 813348）性能的影响，我们使用了 Centos 的高端服务器（2 x Intel Gold 18 core CPU@2.60GHz），安装了 12 x 16 GB 的 DDR4 内存模块和 Intel Optane DC PMEM 的 12 x 128 GB 模块。作为基准，使用基于 R 的工具的完整分析在纯 DRAM 上完成。然后再使用 Memroy Machine 软件和组合 DRAM 加 PMEM 组合的内存池重复进行相同分析。

结果显示使用 DRAM 加 PMEM 组合时性能显著提高。对于每个步骤，通过使用内存快照，数据加载时间从~1000s 减少到~1s。图 2 显示了在迭代步骤中将数据重新加载到内存中的时间最多提高了 25 倍。

Figure 2 – Improvement in Execution Time



实时测序可以帮助控制流行病

自从 DRAM 于 1969 年发明以来，服务器内存模型变化不大，因为 DRAM 一直价格昂贵，是易失性的，且只有通过持续地与用作内存扩展的速度更慢的存储进行 IO 才能实现更大的容量。

有了大内存，包括英特尔持久内存加上 Memory Machine 软件，科学家现在就可以追踪病毒的传播，基因测序可以重构病毒过去在全球传播的过程，并确定它首次在人类中出现的时间。

单细胞 RNA 测序分析是一个系统工程，包括机器学习中通常出现的任务。该过程对内存资源（计算使用需要适配内存的非常大的矩阵），存储（中间结果必须保存并重新加载到其他阶段）提出了严苛的要求，并且该过程运行时间很长（计算密集型，许多阶段需要重复进行参数调整）。通过使用 Memory Machine 软件及其快照功能，可以显著减少执行整个 scRNA-seq 分析的总时间。因此，对于其他使用源自下一代测序技术的大矩阵的生物信息学分析，Memory Machine 软件同样可以进行加速。