**HCA施工大队 · 2018年2月28日 会议记录**

**及每日总结**

|  |  |
| --- | --- |
| 记录人 | 花奎、陈斯杰 |
| 发布时间 | 2018年2月26日 |
| 文件版本 | V1 |

**工作例会 (早)**

**Drop-seq组（冯晨、刘昌毅、王震毅）：**

对于10xGenomics的数据，刘昌毅对手中有100 cells（700MBytes）的数据集进行了从reads到matrix的前端计算，耗时1h12min；冯晨根据之前对~1000cells的数据集进行前端的经验，断定时间消耗同数据量是成正比的。王震毅对于>4000cells(4.9GBytes)的数据进行mapping。

在alpha集群上mapping遇到了报内存错误的问题。这是由于没有使用slurm网格计算平台。

**Smart-seq组（宋绍铭、陈盛泉、鄞启进）：**

数据下完了，开始分析

**ATAC-seq组（张威、魏征）：**

Mapping结束

陈斯杰和魏征就如何实现workflow的思路未达成共识。

总结昨天遇到的问题

1. 关于服务器内存的问题

2. 关于技术实现路线目前的两套思路，目前还未达成共识

重申今日任务

1. 各小组拿真实数据完成对分析流程的测试

2. 魏征、斯杰确定技术实现方案

**下午需求确认会**

定义了Pipeline和Step：

* Step是计算的一个环节，它接受数据及其元数据，输出结果及其元数据。



* Pipeline是一个有向无环图，图的节点是Step，边由各Step在分析流程中的先后顺序决定。



为我们的开发产品定义了初级阶段、高级阶段；

初级阶段为命令行界面，能够动态地选择分析流程中的step，生成pipeline，分析结果用可交互的html展示；

高级阶段提供基于web的图形交互界面，在服务端调用初级阶段的程序。

接下来一段时间里的任务为：

1. 完成初级阶段的交付
2. 提供简便易用的Software Development Kit，使得研究部同学可以参与到pipeline中各Step的开发中来。
   1. 尽量让同学们的工作变成填空工作；
   2. 如果一个非常易用的SDK比较难开发，至少也要有一份详细的文档教大家怎么来开发各个Step；
   3. Keras的Sequence Model和Functional Model是个很好的参考对象；
3. 完成高级阶段的交付

未来的展望：

* 研读各实验室paper，把他们的computational methods映射到我们的pipeline上。
* 支持单核测序、microwell测序。

**工作例会(晚)**

**各组报告今日工作情况**

**Drop-seq组（冯晨、刘昌毅、王震毅）：**

10x数据已拿100个细胞跑通，与标准结果有差异，但是差异不大Mapping率>90%，猜想可能是参数问题，或者是reference版本选择的问题。

其他Drop-seq数据还没开始跑。

10x的数据是所有细胞在一个大fastq里面，不像其他的都在不同的fastq里面。

**Smart-seq组（宋绍铭、鄞启进、陈盛泉）：**

重复现有文章中的工作，前端后端的步骤全部安装成功，在10个细胞上做了前端分析，但是算出来的matrix值要做细胞间normalization，细胞数量少了没法做。

**ATAC-seq组（张威）:**

部分真实数据比对完毕，其他数据正在下载。

明天看看call peak的结果，算overlap率。如果全下完了就在全部数据上做，如果没下完，就取一个subset在上面做。

**魏征、斯杰选出了几种备选的程序实现方案**

看airflow是否能够总结成keras风格。

**下一步任务：**

1. 将现有对各流程的摸索写成文档，明天结束对现有工具的摸索；

2. 确定程序实现方案；

3. 开始做进度计划，制定milestone和甘特图。