Barcode 空间组定位算法

赛题简介

空间转录组学是近些年基因分析技术中的一个重要突破，它可以获得细胞的转录异质性和空间坐标，从而进一步研究和分析细胞功能等生物特性。空间组学实现测序数据的空间定位原理是通过一段 25bp 的 barcode 序列来标记空间位置和测序 reads，然后通过 barcode 序列匹配将测序得到的 reads 定位回空间原位。但是，由于目前的测序仪测到的 DNA 序列不是百分百准确，有一定的错误概率，所以在进行 barcode 序列匹配时，需要进行容错处理。目前的容错策略是将 barcode 的序列上的每一位碱基替换成其它三种碱基（基因序列是由 A/G/C/T 四种碱基组成），然后去做匹配。这种容错方式的时间复杂度比较高，对于一位容错时间复杂度为 25*3*n（n 为需要进行容错的 barcode 数目），对于两位容错，时间复杂度为 25*3*n + 25*24*3*3*n/2，以此类推。

算法原理

程序首先会读取空间组定位芯片的 mask 文件，将编码成 Unint64 数据类型的 barcode 作为 key，将空间位置信息作为 value 存到 hashMap 中，然后根据 mismatch 参数生成用来做容错的二进制 mask 数组，然后开始处理 fastq 文件，用一个线程读取 fastq 文件并将读取的 reads 信息封装存到矩阵中，然后用多个线程从读取的数据包中分别取数据进行分析，分析过程为：首先提取 read1 中的 barcode 序列，并判断是否含有 N 碱基，没有 N 碱基的 barcode 编码成 Unint64，然后去 hashMap 中查找对应的位置信息，如果能找到位置信息，就返回位置信息，如果找不到，就通过和用来容错的 mask 数组中的数值进行二进制异或运算，将 barcode 中每个位置的碱基替换成其它三种，再去 hashMap 中查找，如果通过这种方法查找到的位置只有一个，就返回该位置信息，如果查找到的位置信息有多个，丢掉该 read，对于有 N 碱基的 barcode，如果 mismatch 参数大于 1，就处理对应有 mismatch 参数输入的数值个数的 N 碱基的 barcode，将 N 替换成 A｜T｜C｜G，然后去 hashMap 中查找相应的空间位置信息，如果查找到的位置只

有一个，就返回该位置信息，如果查找到的位置信息有多个，丢掉该 read；提取 barcode 质量值，并统计 Q30，Q20 和 Q10 碱基数目；提取 read 中的 UMI 序列信息和质量值信息，统计 UMI 中的 Q30，Q20 和 Q10 碱基数目，同时过滤掉含有 N 碱基的 UMI，以及含有两个及以上质量值小于 Q10 的碱基的 UMI。对于能够找到对应空间位置，同时 UMI 通过过滤的 reads，将空间位置信息和 UMI 信息作为标签打到 read2 的 readid 中，然后写到输出的 fastq 格式文件中。

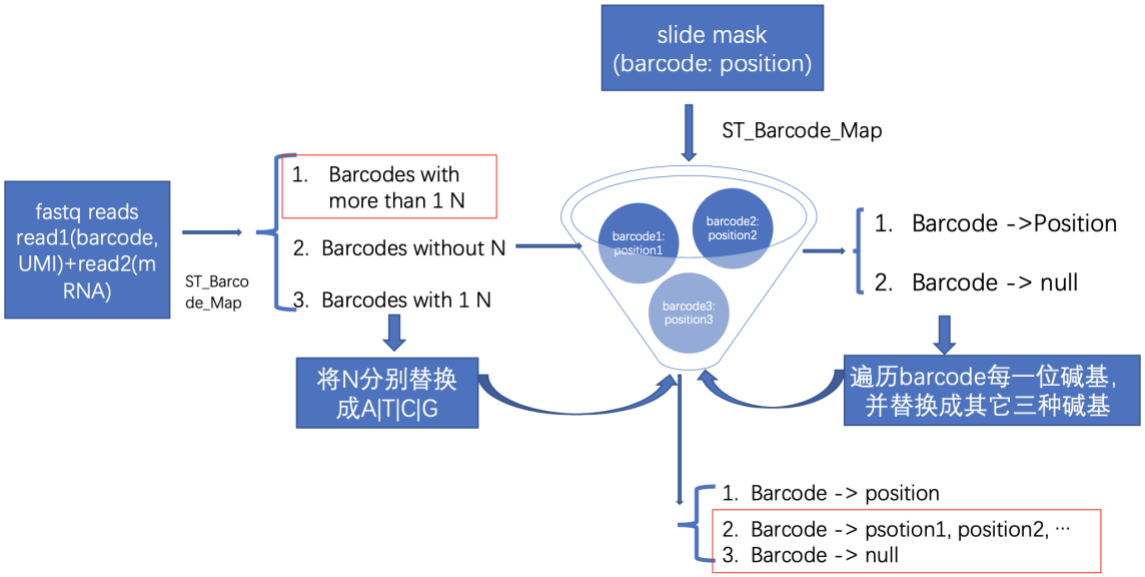

图 1. ST_Barcode_Map 软件运行过程中容错部分的数据处理逻辑示意图

优化思路：

由于当 Mask 文件较大时，直接将 barcode 作为 key，坐标位置作为值的方式会导致内存使用量巨大，加载时间较慢，并且在后续的 fastq 位置寻找时也会很慢，因此采用二级索引（类似分治法）的思路。具体过程是：

1、以 barcode 的部分信息作为分类依据，这里采用前 8 个碱基

2、将前 8 个碱基按照某种数值编码方式（比如 A 对应 0，C 对应 1，G 对应 2，T 对应 3），把碱基序列转换成 16 位的数值 n，然后对其进行一定数量 c 的分类（比如 n%c）

3、按照上述的分类方法，分别对 mask 文件和 fastq 文件进行分类，拆分成多个小文件

4、将属于同一类的 mask 小文件和 fastq 小文件进行 barcodemapping，这一步可以多进程或者分布式来进行，从而解决峰值内存大和运行效率的问题

5、根据需要，在下游分析中选择是否需要合并上述结果

注：以上优化思路仅供参考，选手可以提出更优实现方法。

题目要求：

1、决赛源码包已放在集群上，文件路径为：

/opt/PAC2021/final/ST_BarcodeMap-main.tar

解压源码包后有以下文件：

ST_BarcodeMap-main，主路径

src，源码路径

data，数据及提交脚本路径

run.sh，提交脚本

stat-check.txt，基准结果文件

Makefile，编译文件

软件依赖 hdf5 和 boost_serialization，集群已安装 GNU 环境版本，各队可按需求各自安装其他版本。

环境加载方法：

source /opt/soft/hdf5-1.10.6-gnu/env.sh

source /opt/soft/boost_1_70_0-gnu/env.sh

加载环境后，编译方法如下：

>cd ST_BarcodeMap-main

>make

编译生成 ST_BarcodeMap-0.01 可执行程序，提交程序：

>cd data

>sbatch run.sh

其他参数不需要修改。

2、比赛考察程序计时部分的时间戳的位置不可修改！

3、可以改变数据结构或者数据类型。

4、如认为有必要，可以将必要的代码修改为其它语言如 Fortran、汇编、intrinsic 等等。

5、验证结果正确性: 程序生成的统计信息文件 stat.txt 中的统计信息与标准 stat-check.txt 文件中相同。

6、有违反以上规则者，视为犯规，取消决赛成绩。

7、决赛上机成绩最终将以现场发布的数据为准！现有测试数据不计入最终上机成绩！