寒假work

1.15

理解项目

- 下载G4序列和已测好的甲基化序列，寻找重叠部分
- LTR结构中具有CpG，易发生甲基化，但G4会抑制甲基化
- fasta和bed文件的转化

待做

- 理解pipeline
  - 目的：计算出染色体中G4的甲基化部分
  - Quadron_finder:
    - 输入：该条染色体的序列（fasta）
    - 结果：该条染色体中G4序列（bed）
    - 具体怎么实现的还不大明白?
  - trans_mCfile2bed_gz：
    - 输入：甲基化序列文件（tsv.gz格式）的路径，文件名
    - 结果：甲基化序列文件（bed格式）
    - 过程：

      trans2bed：
      - 输入：甲基化序列文件中的一行
      - 结果：按“/t”分开，分别将信息对应至bed文件中每列
  - calculate_coverage_mC_g4：
    - 输入：G4bed文件和甲基化序列bed文件
    - 输出：即下面函数的输出

        - 过程：
          mC_coverage_parser：
            - 输入：G4bed文件和甲基化序列bed文件合并后的文件（利用bedtools处理，还不太理解？）
            - 输出：重叠的序列，总的序列，重叠序列占比，hist（？）
    - fasta_all_5：
        - 输入：全序列文件和idfile（方便对比）
        - 输出：每个染色体自己的序列文件
        - 过程：
          按行分开，该行若有“>”则说明为新一条染色体的序列的开头，则flag+1，而从它开始到再一次读到“>”,flag均不变，即都写到对应第flag个染色体的文件中
- python小tips：
    - split（某种符号或就空着）：
        - 将序列以该符号划分，每个部分做为数组的一项
          例如：str = "Line1-abcdef \nLine2-abc \nLine4-abcd";
          print str.split( );
          #以空格为分隔符，包含 \n
          print str.split(' ', 1);
          #以空格为分隔符，分隔成两个
          ['Line1-abcdef', 'Line2-abc', 'Line4-abcd']
          ['Line1-abcdef', '\nLine2-abc \nLine4-abcd']
        - split('\n')[0]是获取第一行的信息
    - replace('符号','')就是把相应符号替换掉，无论是几个符号，只要是符号都会被替换掉
        - replace(' ',''）：把空格替换掉
    - join是字符串操作函数，操作的也是字符串，其作用结合字符串使用，常常用于字符连接操作
        - key="\t".join(('a','b','c'))（join括号里的单位只能为一个单位，所以里面那个括号去掉就会报错）
          结果：'a b c'
          result= key.split("\t")
          结果:[a,b,c]
- 爬虫批量下载文件
- 文献阅读
- tmux

记录

- bed文件
    - 3个必须的列和9个额外可选的列

- chrom（染色体名字）、目标区段起止位置
- strand： 定义链的方向，"+” 或者”-”
  thickStart： 起始位置(例如，基因起始编码位置)
  thickEnd： 终止位置（例如：基因终止编码位置)
  itemRGB： 是一个RGB值的形式, R, G, B (eg. 255, 0,0), 如果itemRgb设置为'On”, 这个RBG值将决定数据的显示的颜色。
  blockCount： BED行中的block数目，也就是外显子数目
  blockSize：用逗号分割的外显子的大小, 这个item的数目对应于BlockCount的数目
  blockStarts： 用逗号分割的列表, 所有外显子的起始位置，数目也与blockCount数目对应

1.16