////// 高级生信系列课程 //////

利用Python进行生信分析

原理 · 算法 · 实战

编程课

高级生信系列课程

项目二
BED文件的操作与处理

孟浩巍
menghaowei1992@126.com
2022年9月25日

课程资料下载

- 由于腾讯课堂改版，所有课程讲义、测试数据、运行代码均**需要加入课程讨论群**进行下载！

- 课程讨论QQ群号： 828030346

本节课程授课顺序

1. 讲解相关背景；
2. 讨论程序编写的逻辑；
3. 使用jupyter编写程序，并调试程序；
4. 将程序编写为命令行形式；

////// 高级生信系列课程 //////

Part II
获取BED文件的交集与并集

程序分析（2个BED文件求交集）

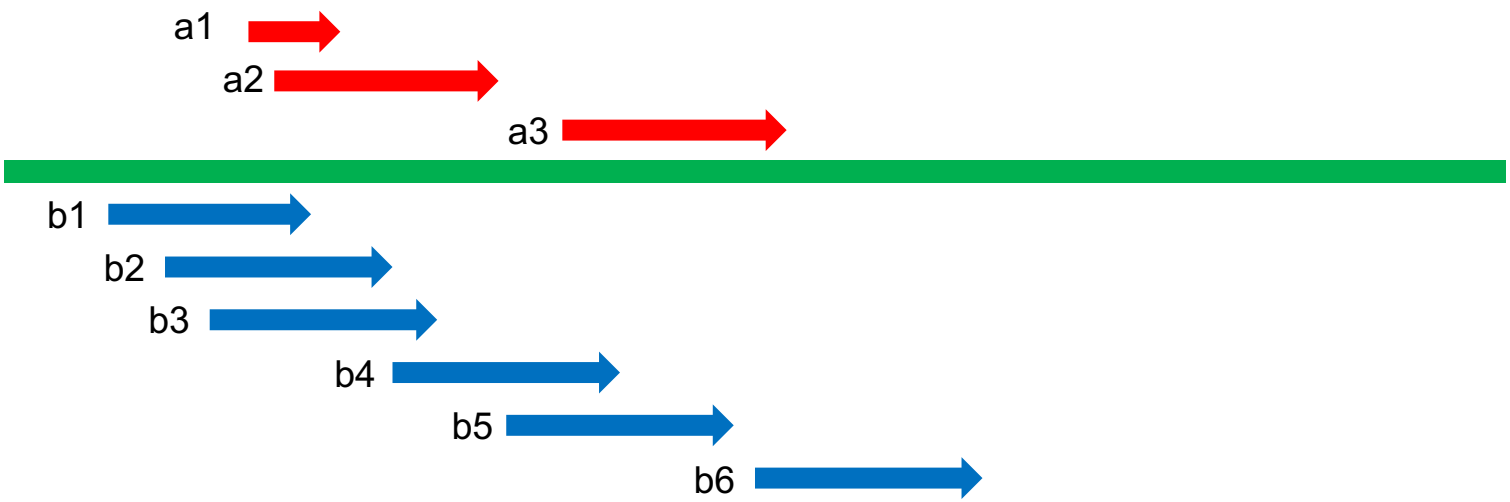
- 程序的输入
- 程序的输出
- 程序的用途
  - 求解交集在a文件中的情况；
  - 求解交集的部分；
  - 求解交集部分的占比；
  - 其它个性化需求……

复杂情况的程序编写

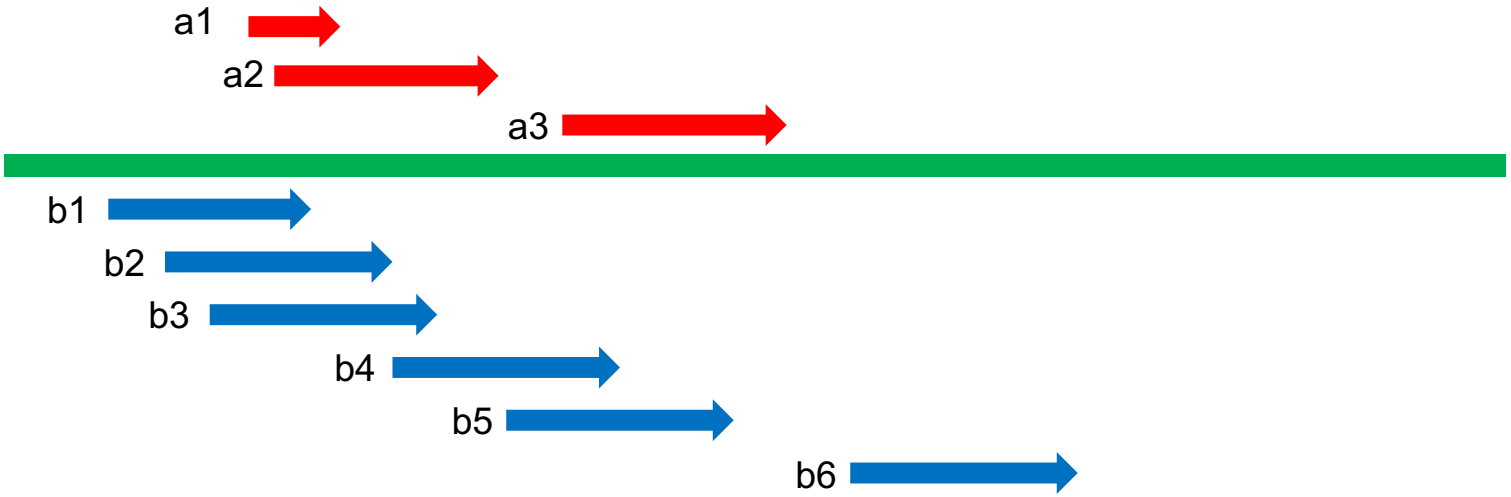
- 假设存在比较复杂的情况：
  - 文件a与文件b中的区域存在一对多，或者多对多的情况

复杂情况的程序编写

- 假设存在比较复杂的情况：
  - 文件a与文件b中的区域存在一对多，或者多对多的情况



复杂情况的程序编写



- 输出结果:
  - a1 (b1, b2, b3)
  - a2 (b1, b2, b3, b4)
  - a3 (b4, b5)

bedtools intersect结果

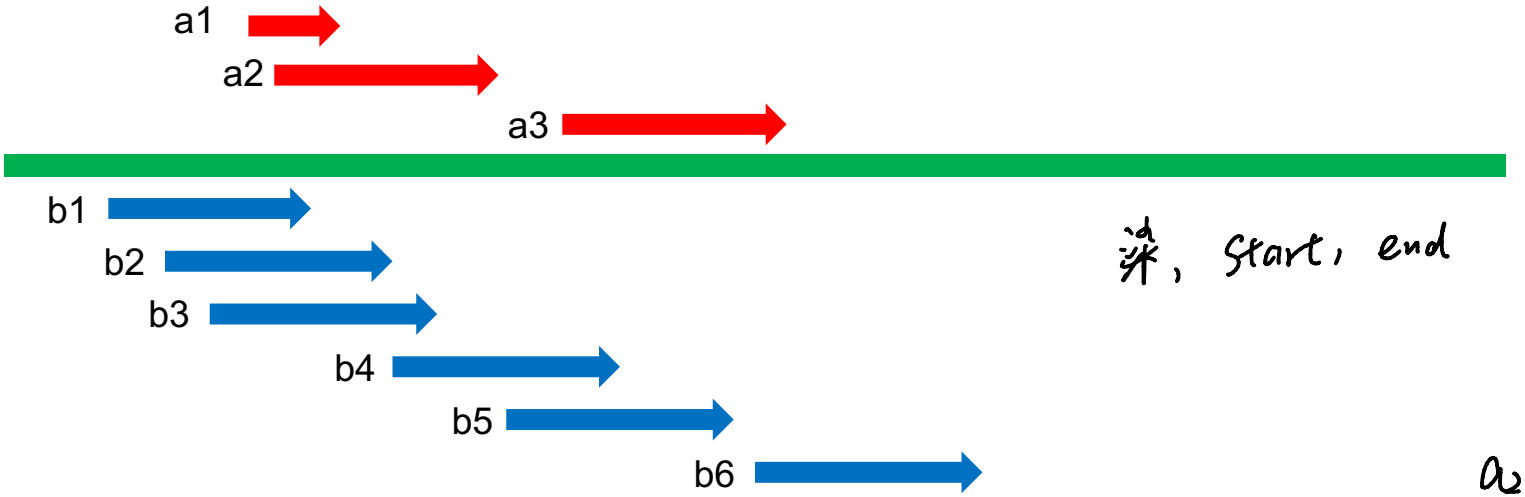
运行命令:

```
bedtools  intersect -a test_multi_region_a.sort.bed  -b
test_multi_region_b.sort.bed -wa -wb
```

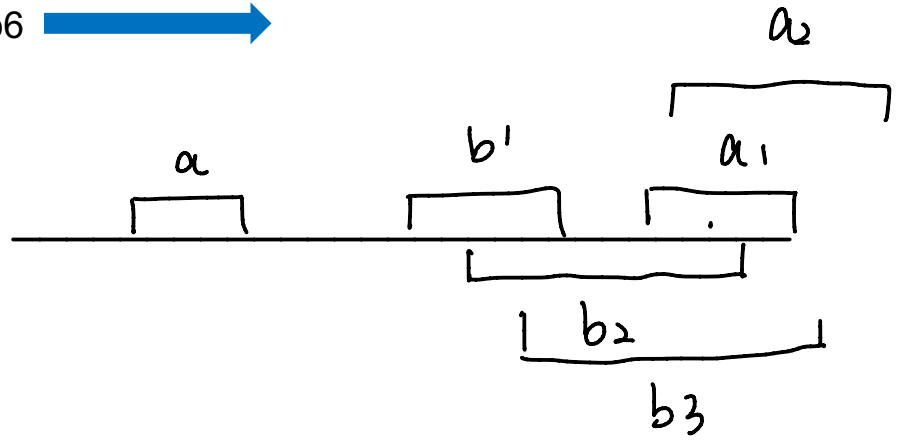
运行结果:

chr1	140	170	a1	chr1	100	150	b1
chr1	140	170	a1	chr1	120	180	b2
chr1	140	170	a1	chr1	130	200	b3
chr1	145	280	a2	chr1	100	150	b1
chr1	145	280	a2	chr1	120	180	b2
chr1	145	280	a2	chr1	130	200	b3
chr1	145	280	a2	chr1	175	300	b4
chr1	145	280	a2	chr1	260	400	b5
chr1	290	350	a3	chr1	175	300	b4
chr1	290	350	a3	chr1	260	400	b5

程序设计思路

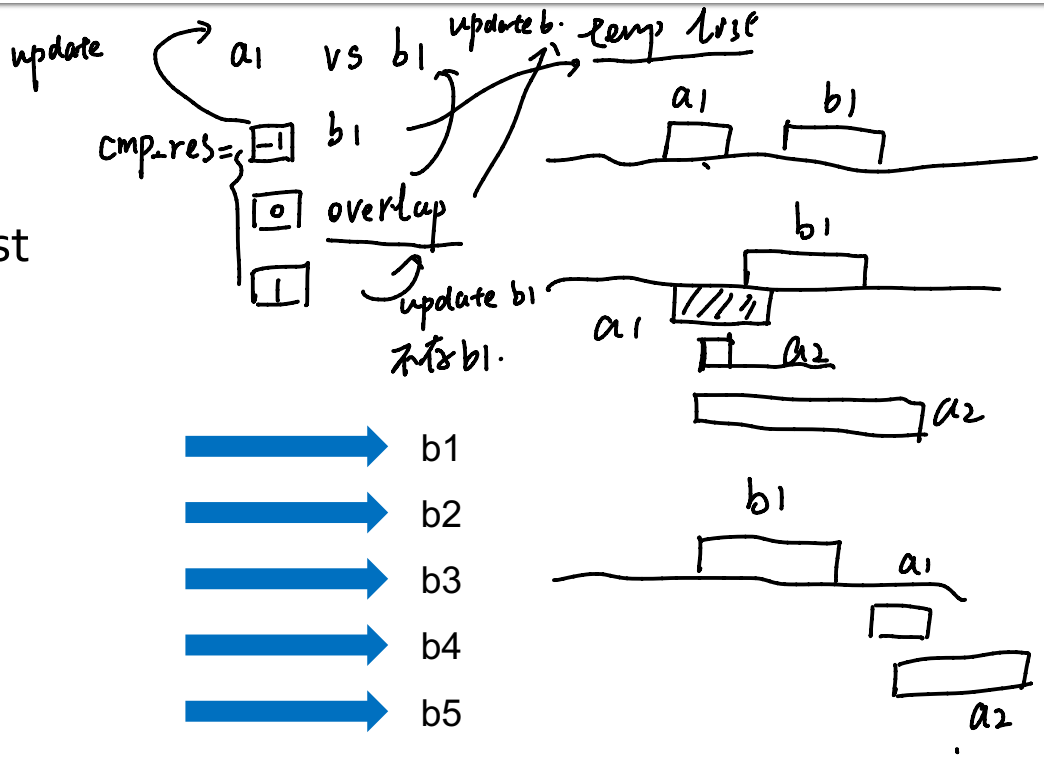


- 用函数简化代码；模块化
- 重新确定边界条件；
- 引入保留行信息的temp_list



程序设计思路

- 用函数简化代码；
- 重新确定边界条件；
- 引入保留行信息的temp_list



把程序写成命令行版本

- 程序的输入
  - --input_a
  - --input_b
  - --reference_index

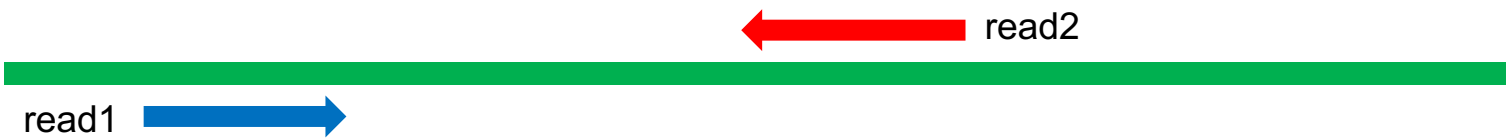
高级生信系列课程

项目四
BAM文件的操作与处理

孟浩巍
menghaowei1992@126.com
2022年10月03日

Part I
SAM文件与BAM文件的介绍与基本操作

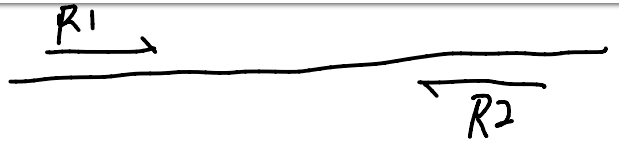
比对文件需要记录哪些信息?



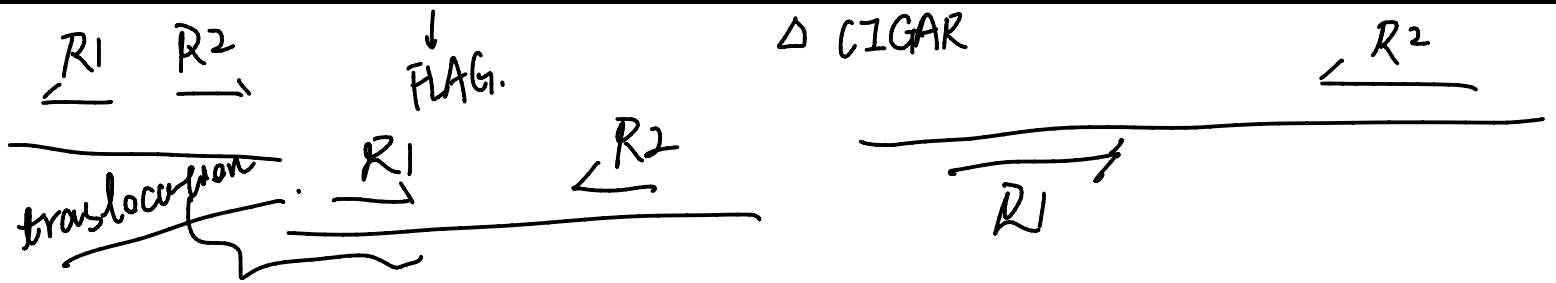
1. 比对位置
2. 比对质量
3. +- 链
4. inser del mutation
5. 原始序列信息、

SAM文件与BAM文件是用来储存比对结果的

- 一个比对结果的例子




```
@SQ     SN:chr17        LN:83257441
@SQ     SN:chr18        LN:80373285
@SQ     SN:chr19        LN:58617616
@SQ     SN:chr20        LN:64444167
@SQ     SN:chr21        LN:46709983
@SQ     SN:chr22        LN:50818468
@SQ     SN:chrX LN:156040895
@SQ     SN:chrY LN:57227415
@SQ     SN:chrM LN:16569
@RG     ID:293T_Mock-Input-covaris_rep1 SM:293T_Mock-Input-covaris_rep1 PL:ILLUMINA
@PG     ID:bwa  PN:bwa  VN:0.7.17-r1188 CL:/gpfs/build/bin/bwa mem /home/menghaowei/menghw_HD/reference/bwa_hg38/hg38_only_chromosome.fa fix.fastq/293T-Mock-Input-
@PG     ID:MarkDuplicates       VN:2.18.17-SNAPSHOT     CL:MarkDuplicates INPUT=[bam/293T-Mock-Input-covaris_bwa_hg38_rep1_sort.bam] OUTPUT=bam/293T-Mock-Input-cov
@PG     ID:samtools     PN:samtools     PP:bwa  VN:1.9  CL:samtools reheader -i rep1.header 293T-Mock-Input-covaris_bwa_hg38_rep1_sort_rmdup.bam
@PG     ID:samtools.1   PN:samtools     PP:MarkDuplicates       VN:1.9  CL:samtools reheader -i rep1.header 293T-Mock-Input-covaris_bwa_hg38_rep1_sort_rmdup.bam
E00495:573:H375LCCX2:8:1111:21115:31142 129     chr1    10002   43      104M1I41M       chr3    44587195        0       AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
E00495:573:H375LCCX2:8:1206:22079:2856  99      chr1    10005   24      45M1D100M5S     =       10094   138     CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTACCCTA
E00495:573:H375LCCX2:8:2103:22150:30017 163     chr1    10054   27      7M1D89M1D17M    =       10124   218     CTAACCCAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAA
E00495:573:H375LCCX2:8:1206:22079:2856  147     chr1    10094   24      101S49M =       10005   -138    TCATGTCCTTCCCTGTTACCTCATCCCTTTCCTGCACCCTACCCCCATCCATTCCCCGT
E00495:573:H375LCCX2:8:2103:22150:30017 83      chr1    10124   27      20M1I13M1I6M1I13M4D12M2I52M1I28M        =       10054   -218    CCTAACCCTAACCCTAACCCTTAACCC
E00495:573:H375LCCX2:8:2207:28392:22370 163     chr1    10150   40      7M1I74M1D59M    =       10329   303     CTAACCCTTAACCTTAACCTTAACCTTAACCTAACCCTAACCCTAACCCTA
E00495:573:H375LCCX2:8:2207:28392:22370 83      chr1    10329   43      27S115M1D8M     =       10150   -303    CTTAACCCTAACCCTACCCCTACCCCTACCCCTACCCCTAACCCTAACCCTA
E00495:573:H375LCCX2:8:2217:7984:60888  147     chr1    10354   20      57S93M  =       10034   -413    CCCCAACCCCCCCCCTCCCCCCCCCCCTCACCCTCCCCCCCCCACCCCTACCCCTACCC
E00495:573:H375LCCX2:8:2102:13738:25710 99      chr1    10446   47      7M1I142M        =       10466   170     ACCCTAACCCCTAACCCTAACCCTCGCGGTACCCTCAGCCGGCCCGCCCGC
E00495:573:H375LCCX2:8:2102:13738:25710 147     chr1    10466   48      150M    =       10446   -170    CCTCGCGGTACCCTCAGCCGGCCCGCCCGCCCGGGTCTGACCTGAGGAGAACTGTGCTC
```



SAM文件的标准

Sequence Alignment/Map Format Specification

The SAM/BAM Format Specification Working Group

3 Jun 2021

The master version of this document can be found at `https://github.com/samtools/hts-specs`.
This printing is version 53752fa from that repository, last modified on the date shown above.

1 The SAM Format Specification

SAM stands for Sequence Alignment/Map format. It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. If present, the header must be prior to the alignments. Header lines start with '`@`', while alignment lines do not. Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information.

This specification is for version 1.6 of the SAM and BAM formats. Each SAM and BAM file may optionally specify the version being used via the `@HD VN` tag. For full version history see Appendix B.

Unless explicitly specified elsewhere, all fields are encoded using 7-bit US-ASCII [1] in using the POSIX / C locale. Regular expressions listed use the POSIX / IEEE Std 1003.1 extended syntax.

1.1 An example

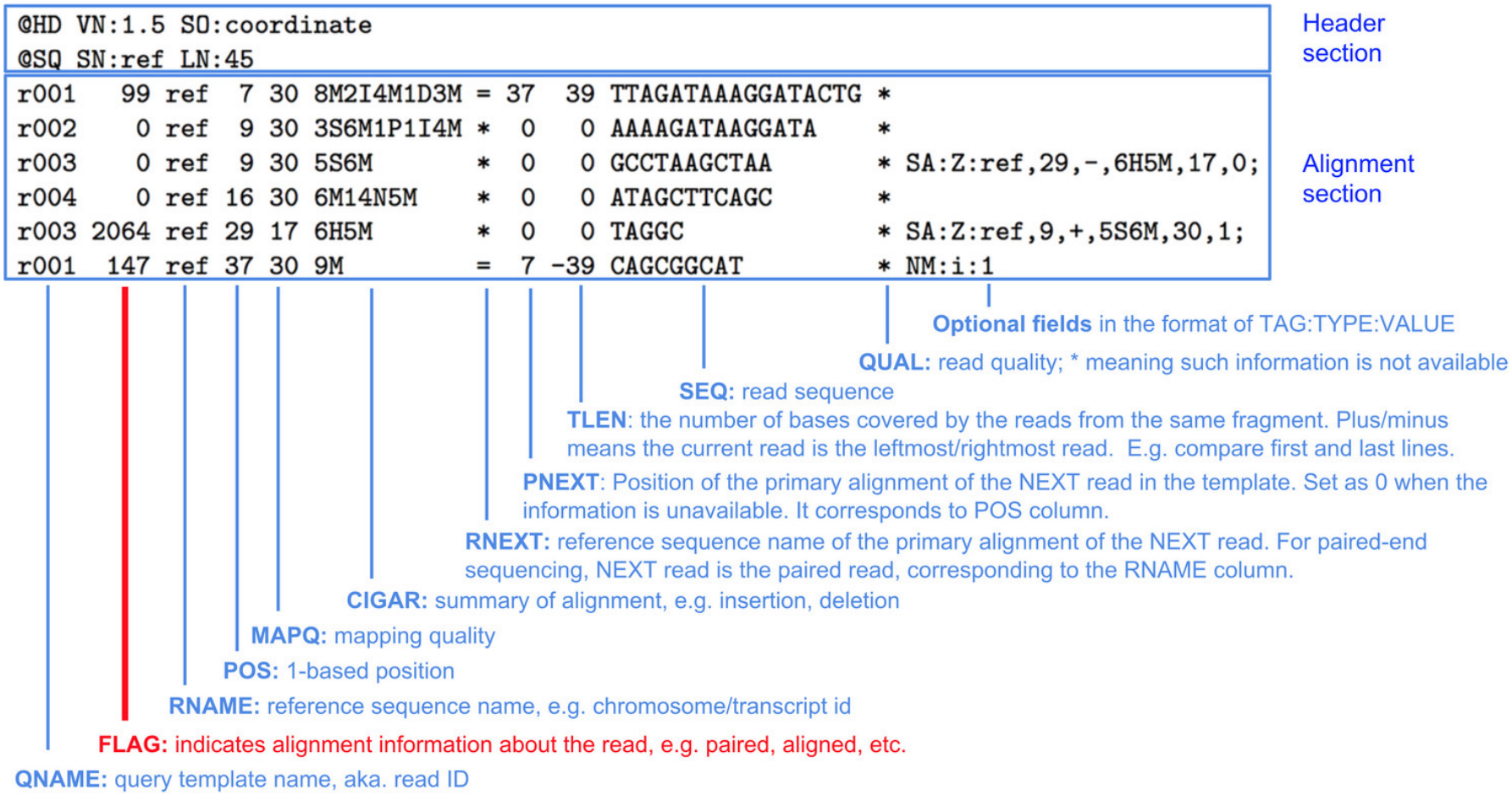
Suppose we have the following alignment with bases in lowercase clipped from the alignment. Read `r001/1` and `r001/2` constitute a read pair; `r003` is a chimeric read; `r004` represents a split alignment.

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                    ATAGCT..............TCAGC
-r003                          ttagctTAGGC
-r001/2                                     CAGCGGCAT
```

https://github.com/samtools/hts-specs

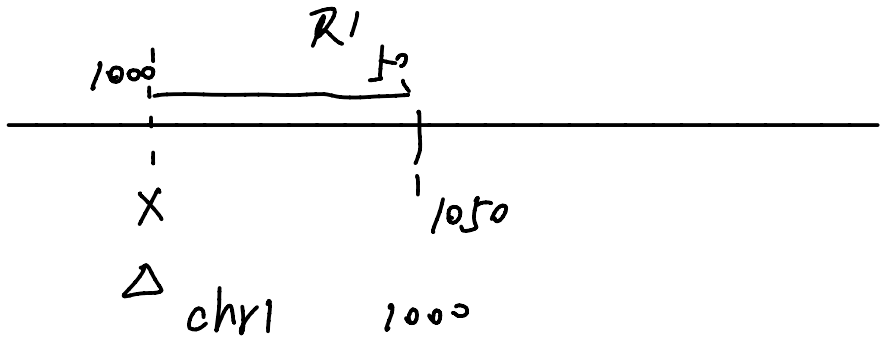
SAM文件的标准

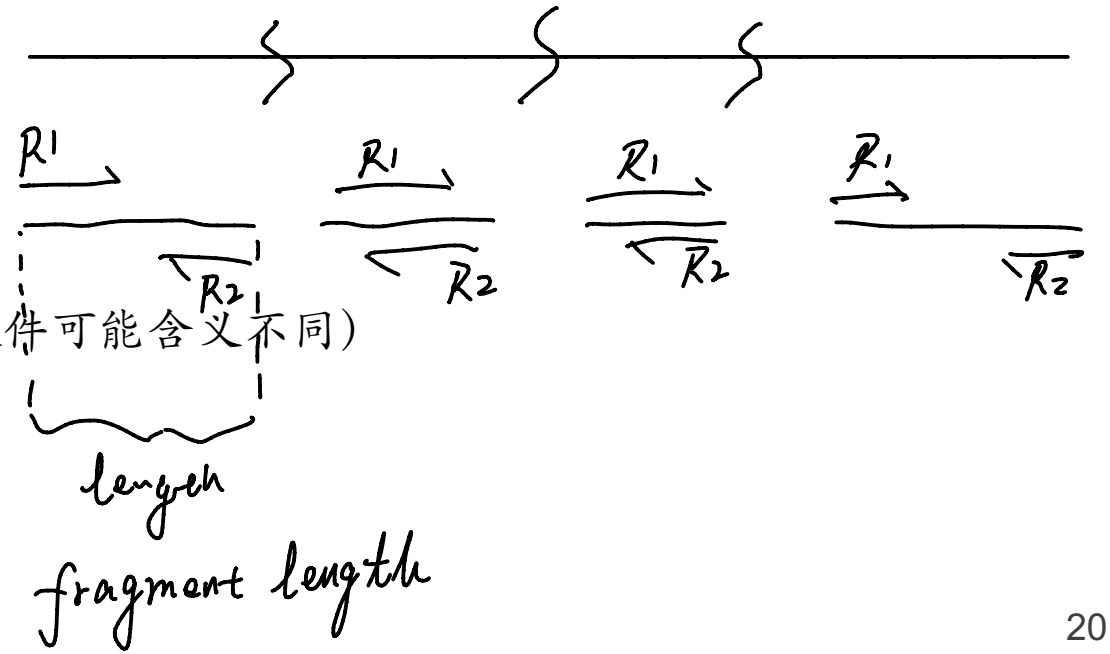


https://www.samformat.info/sam-format-flag

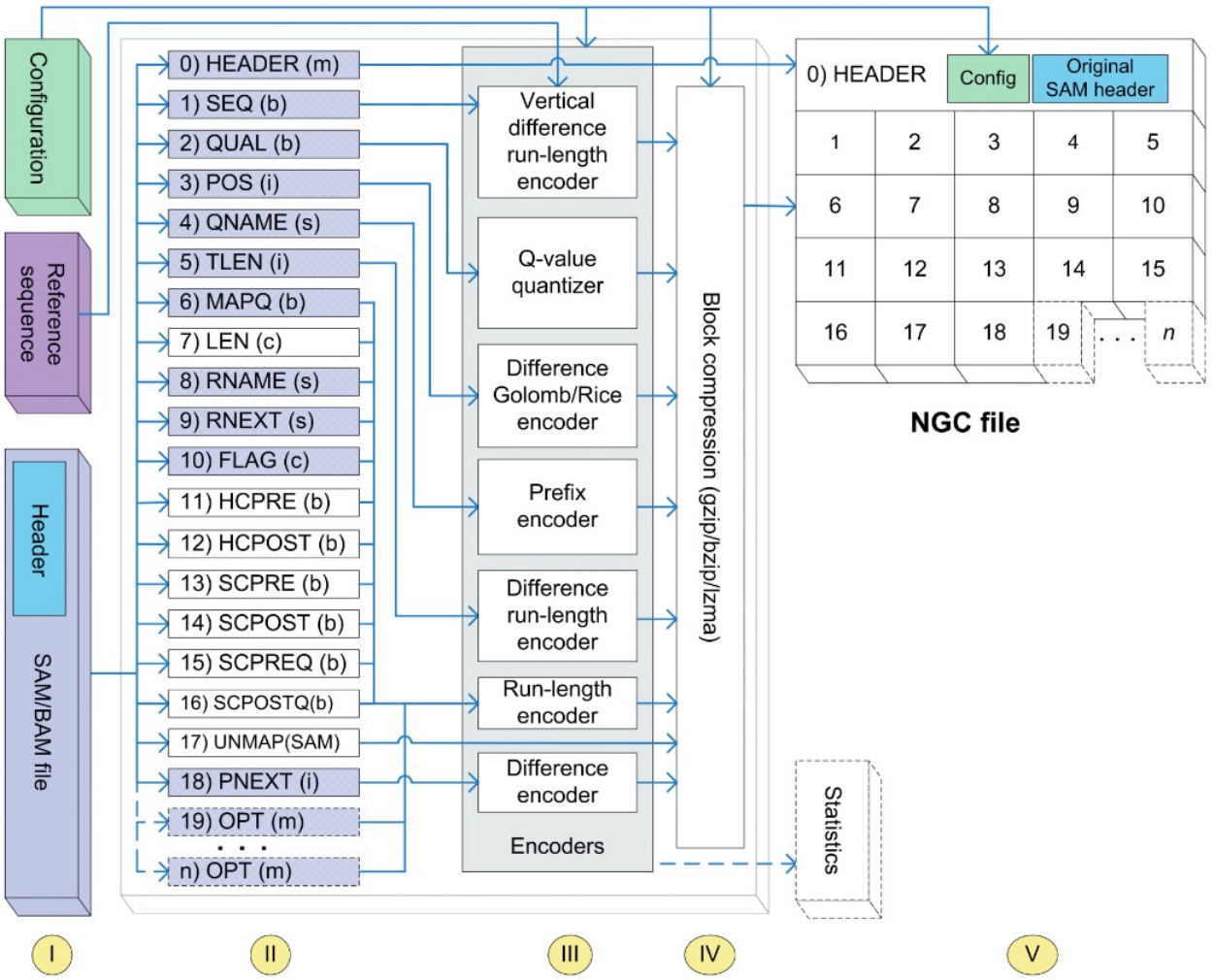
SAM/BAM文件中的重要信息

- CIGAR值 用来储存比对的简略情况
- 双端测序的配对Reads name
- Fragment length
- 比对坐标
- 其它Tag信息（不同比对软件可能含义不同）





SAM文件的压缩与重编码



Niko Popitsch and Arndt von Haeseler, NAR, 2013

SAM文件的压缩与重编码

一个简单的类比

SAM文件是无序的文本文件

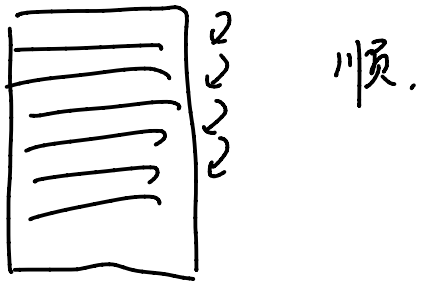
只能进行顺序读取

经过sort和index后的BAM文件
是有序的二进制文件

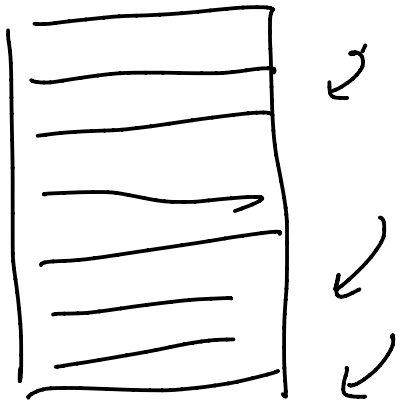
可以进行随机读取

顺序读取与随机读取

- 顺序读取



- 随机读取

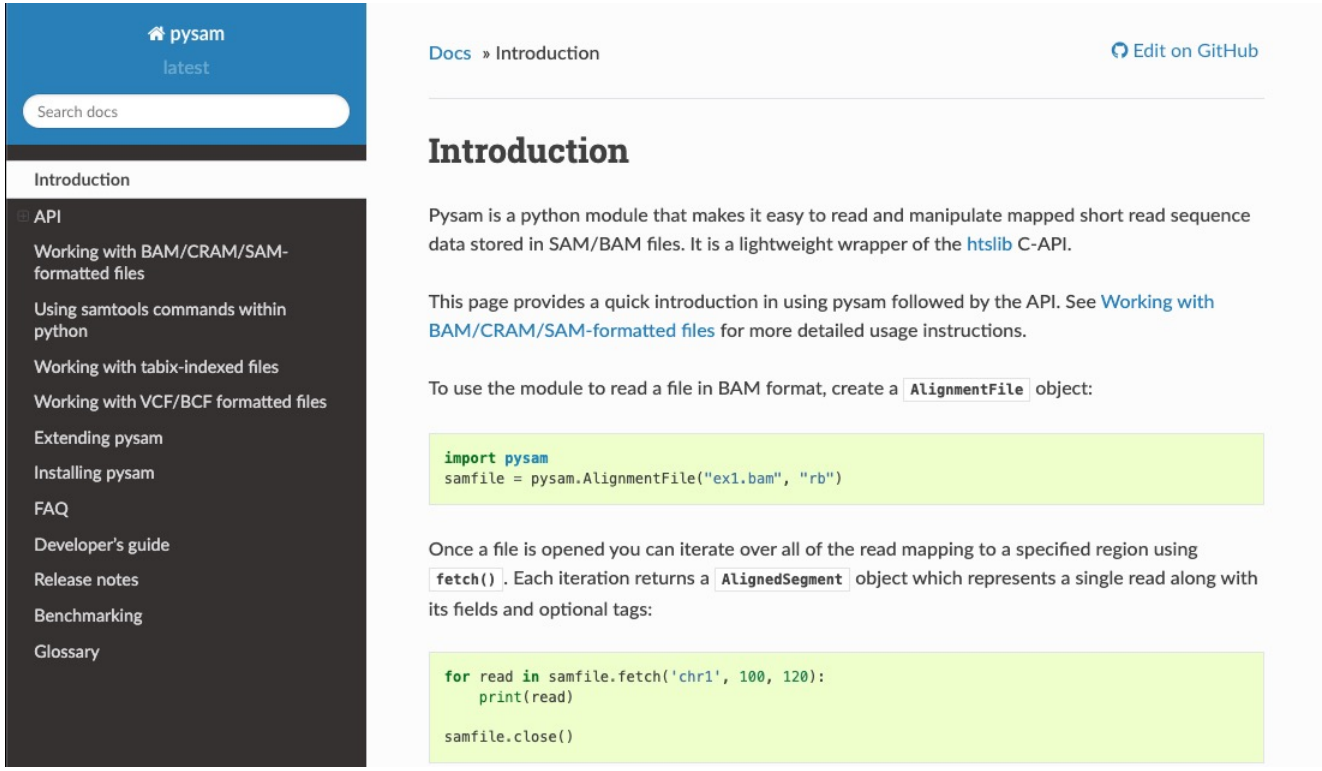


BAM文件的一些实用操作

- 选取任一区域，提取BAM文件
  - `samtools view`
- 计算基因组每一个位置的覆盖度
  - `samtools depth`
- 计算基因组每一个位置的详细覆盖情况（包括突变信息等）
  - `samtools mpileup`
- 快速获取各条染色体的比对情况
  - `samtools idxstats`

Part II
使用Pysam读取BAM文件中的比对信息

为什么实用Pysam包?



- 重点在解决问题，而不是去解决细枝末节的接口问题；
- 要造车，在有轮子的情况下，尽量不要造轮子！

https://pysam.readthedocs.io/en/latest/api.html

Pysam的一些基础操作

- 打开、关闭SAM/BAM文件;
- 对SAM/BAM文件进行顺序读取;
- 对BAM文件进行随机读取;

- 获得reads的比对信息
  - FLAG值的解析
  - CIGAR值的解析
  - 突变信息的解析
  - Tag信息的解析

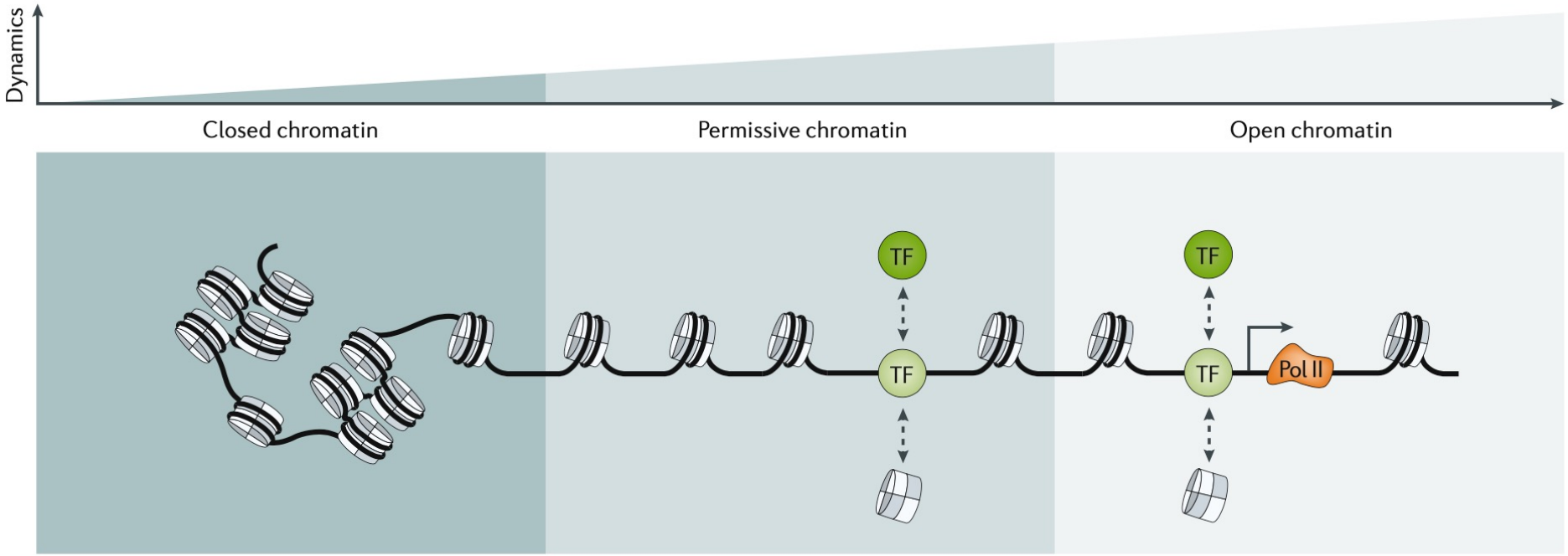
Part III
对BAM文件进行过滤

为什么要对BAM文件进行过滤

- samtools等工具已经提供了很多过滤操作
  - 筛选MAPQ > 20的比对结果 samtools view -q 20
  - 筛选比对到基因组正链的结果 -F 16.
  - 筛选某一个区域的所有比对结果 samtools view chr21:xx-xx.

- samtools不能够完成的
  - R1和R2中有1条MAPQ < 30就过滤这一对reads;
  - 通过fragment筛选

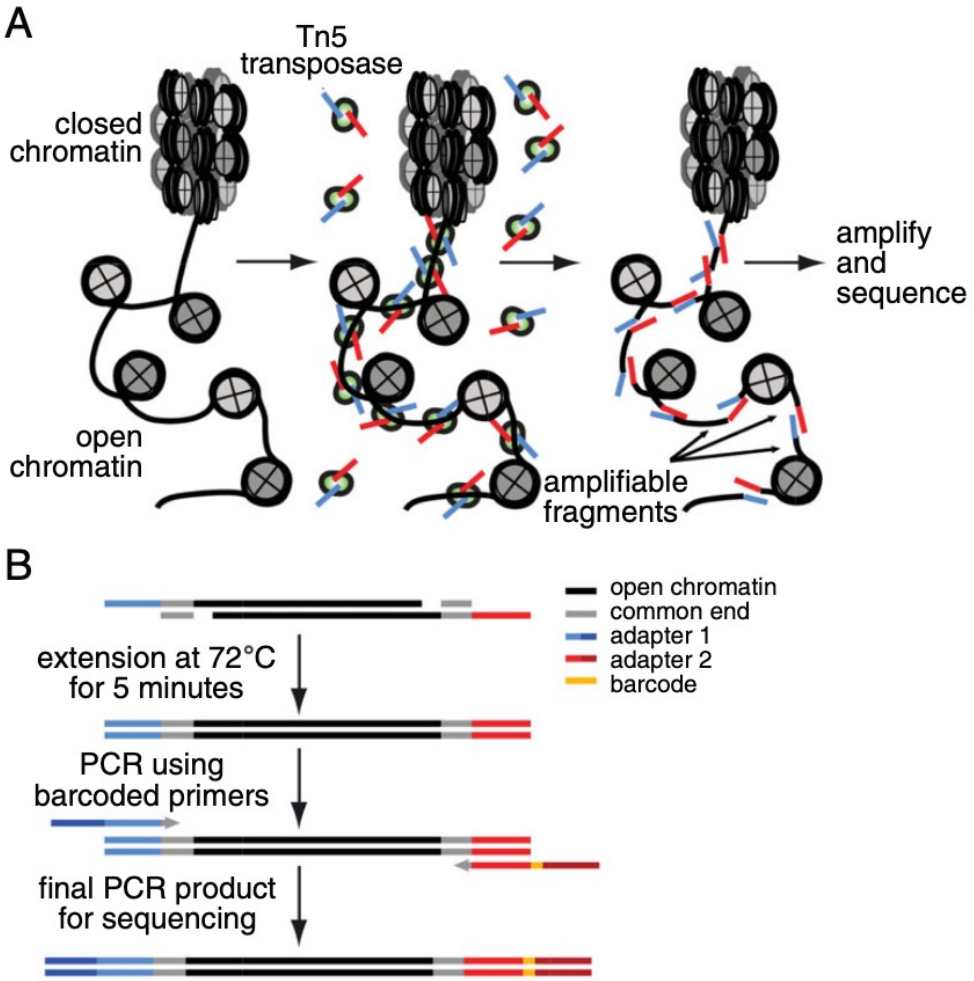
Chromatin accessibility



- 如何定量染色质可及性?

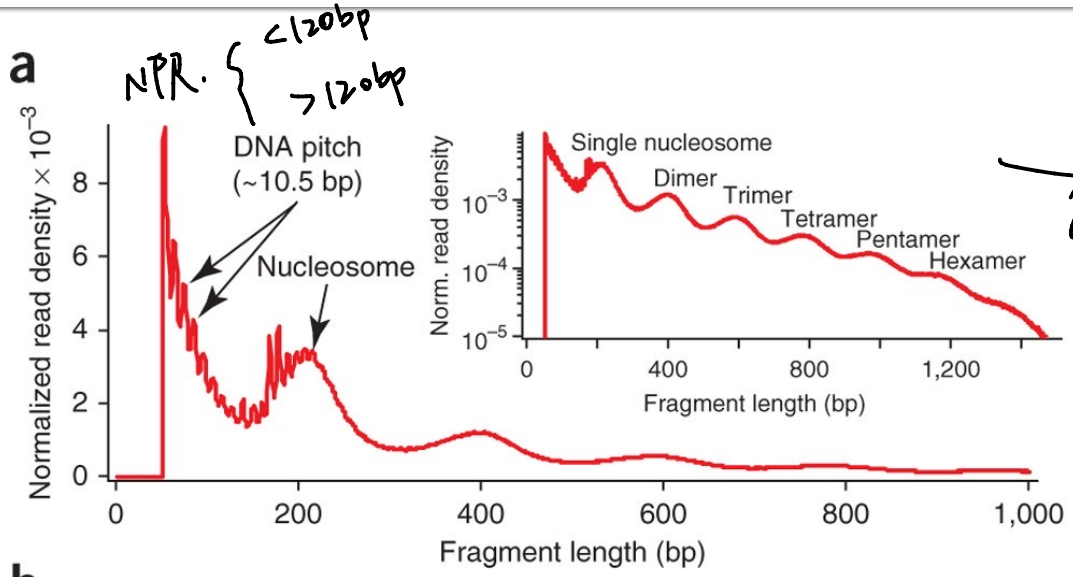
Sandy L. Klemm et al., Nat Rev Genetics, 2019

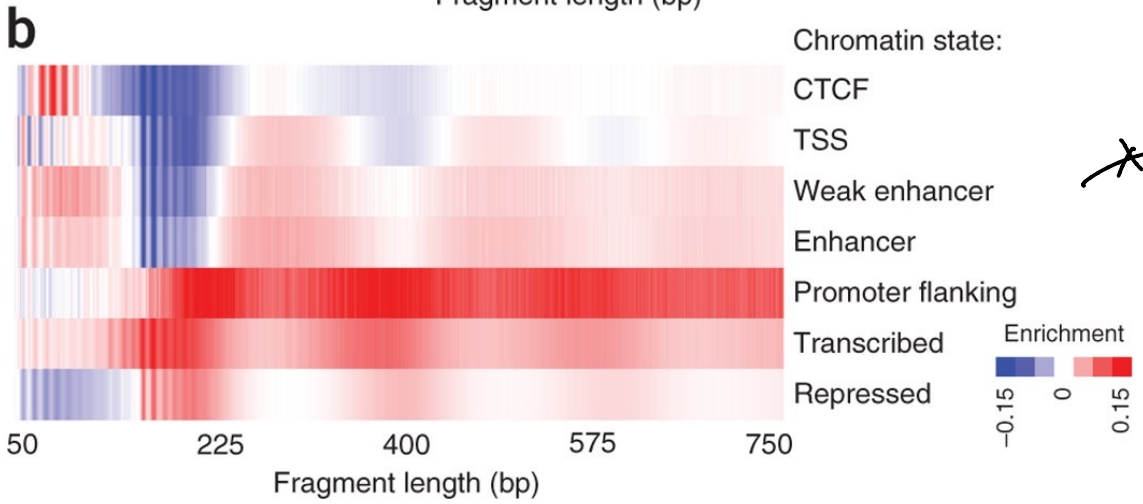
ATAC-seq技术原理

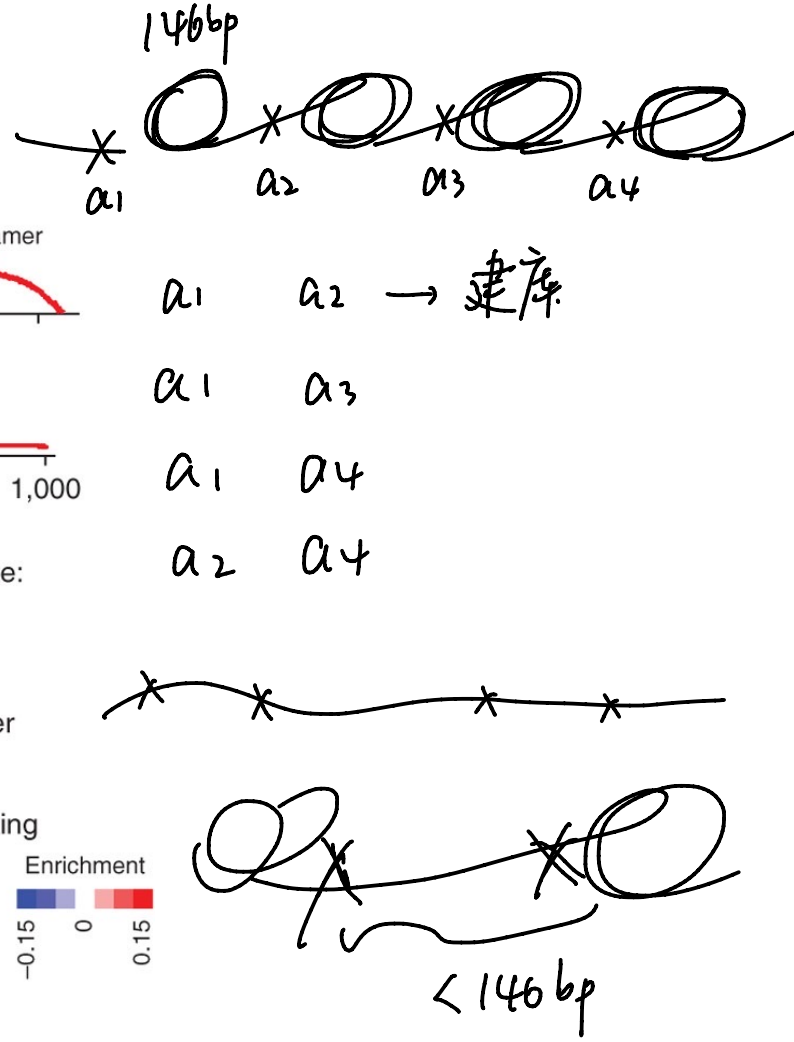


Jason D. Buenrostro et al., Curr. Protoc. Mol. Biol., 2015

ATAC-seq对基因组开放区域进行捕获

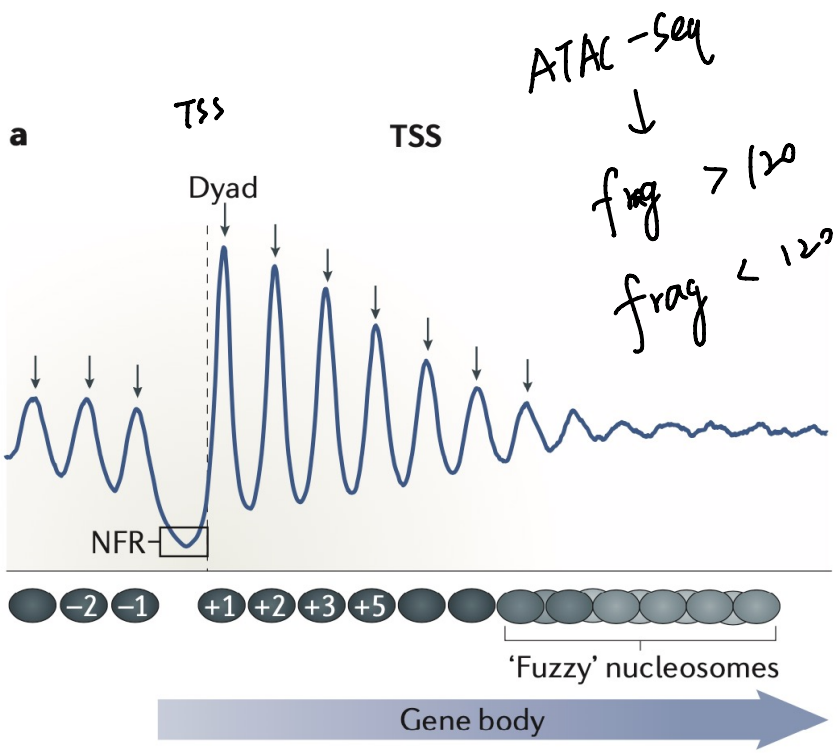


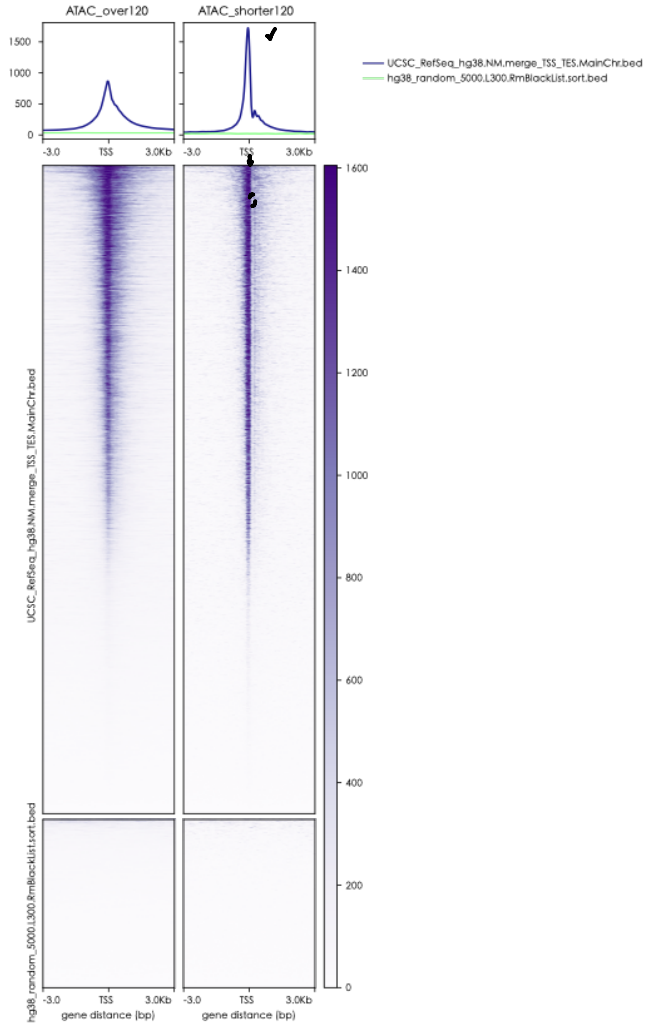




Jason D. Buenrostro et al., Nat Methods, 2013

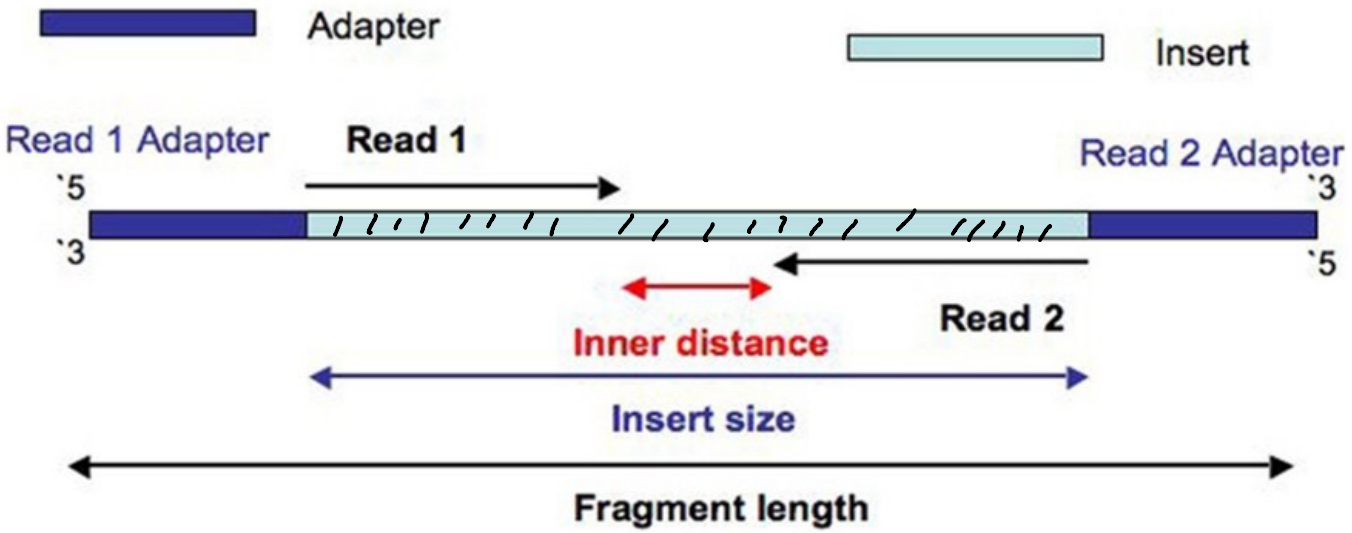
不同fragment length的ATAC-seq信号





William K. M. Lai and B. Franklin Pugh, Nat Rev Genetics, 2017

Fragment length的解释



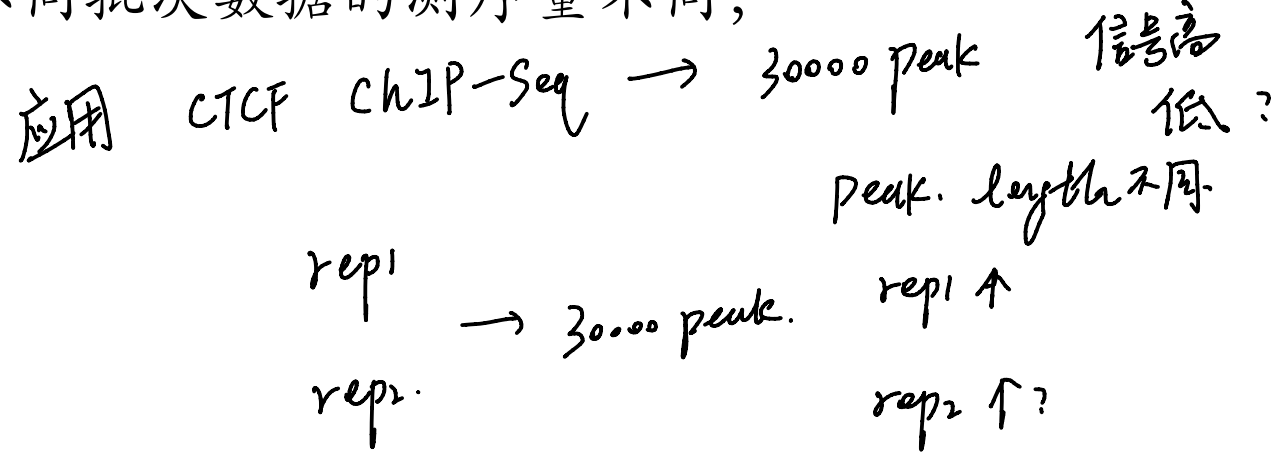
- 在衡量文库构建片段长度时，fragment length一般指包含adapter序列的长度；
- 在进行mapping以后的片段长度计算时，fragment一般指不包含adapter序列的长度，即上图的insert size

Part IV
计算给定区域的信号值

FPKM RPKM CPM

为什么要对数据进行标准化（normalization）？

- 不同区间的长度不同；
- 不同批次数据的测序量不同；



RPM (Reads per million mapped reads)
CPM (Counts per million mapped reads)

- 不同批次数据的测序量不同；

CPM.

peak region.

CPM = 10 ?

$$\text{RPM or CPM} = \frac{\text{Number of reads mapped to gene} \times 10^6}{\text{Total number of mapped reads}}$$

region

$$\frac{20}{100M} \cdot = 0.2$$

CPM = 0.2 ? 1M, 0.2

RPKM (*Reads per kilo base per million mapped reads*)
FPKM (*Fragments per kilo base per million mapped reads*)

- 不同基因的长度不同；
- 不同批次数据的测序量不同；

$$\mathrm{RPKM} = \frac{\text{Number of reads mapped to gene} \times 10^3 \times 10^6}{\text{Total number of mapped reads} \times \text{gene length in bp}}$$

分子：测到的 Reads 数 / 总测序 × 长度

20个 / 100bp

$RPKM = \frac{20}{100M \cdot 0.1k} = 2$

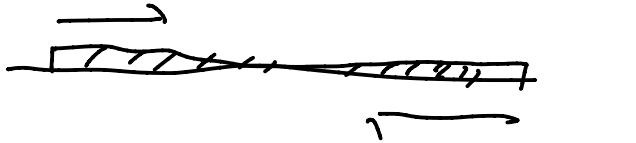
RPKM = 10

1M Reads，平均 1kbp，2条；

1M Reads 平均 1kbp，10个。

RPKM (*Reads per kilo base per million mapped reads*)
FPKM (*Fragments per kilo base per million mapped reads*)

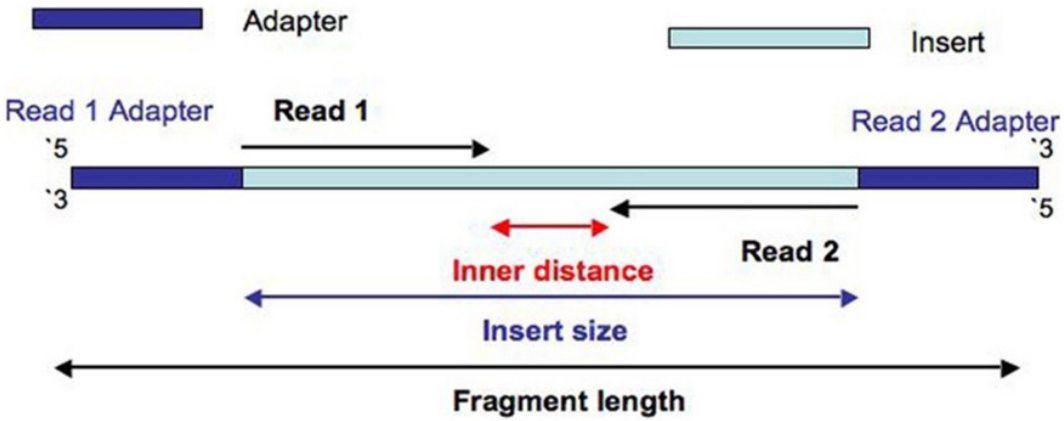
- 不同基因的长度不同；
- 不同批次数据的测序量不同；



Reads count = 2 * fragment

FPKM = RPKM / 2

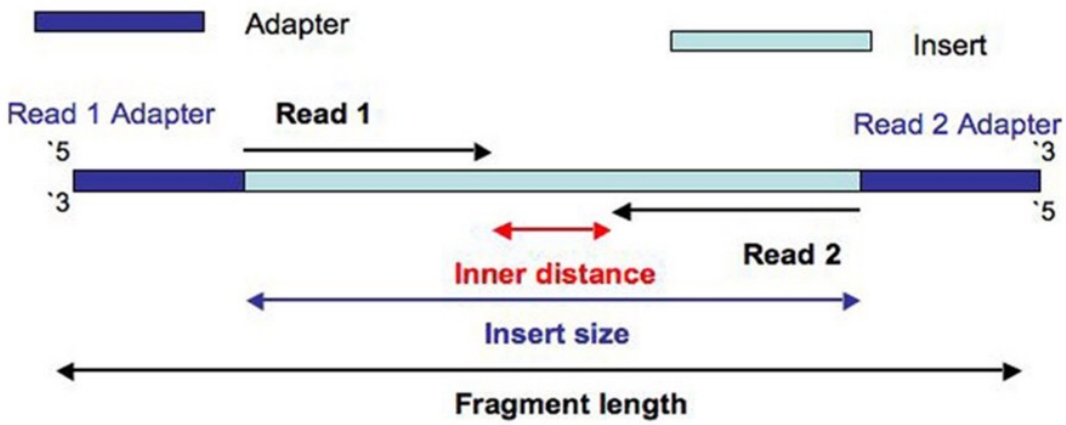
$$\mathrm{RPKM} = \frac{\text{Number of reads mapped to gene} \times 10^3 \times 10^6}{\text{Total number of mapped reads} \times \text{gene length in bp}}$$



RPKM (*Reads per kilo base per million mapped reads*)
FPKM (*Fragments per kilo base per million mapped reads*)

- 不同基因的长度不同；
- 不同批次数据的测序量不同；

$$\mathrm{RPKM} = \frac{\text{Number of reads mapped to gene} \times 10^3 \times 10^6}{\text{Total number of mapped reads} \times \text{gene length in bp}}$$



$$\mathrm{FPKM} = \mathrm{RPKM} / 2$$

计算给定区域的FPKM/RPKM/CPM

- 程序的输入
  - BAM文件，给定的区域（BED文件）
- 程序的输出
  - BED文件包含信号值

CPM, RPKM, FPKM, → 分母. 总数.

- 程序关键步骤
  - 获得BAM文件中的总reads数
  - 读取区间，并获得区间中的reads数
  - 计算对应信号值

Reads / Fragment

Part V
读取指定坐标位置的突变信息并进行统计检验

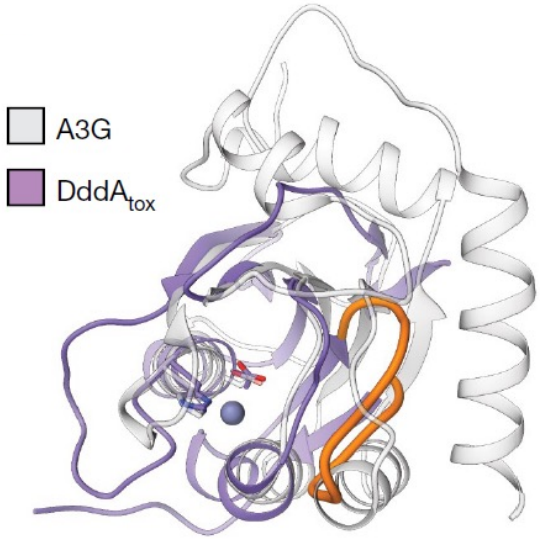
线粒体的基因编辑工具 DdCBE

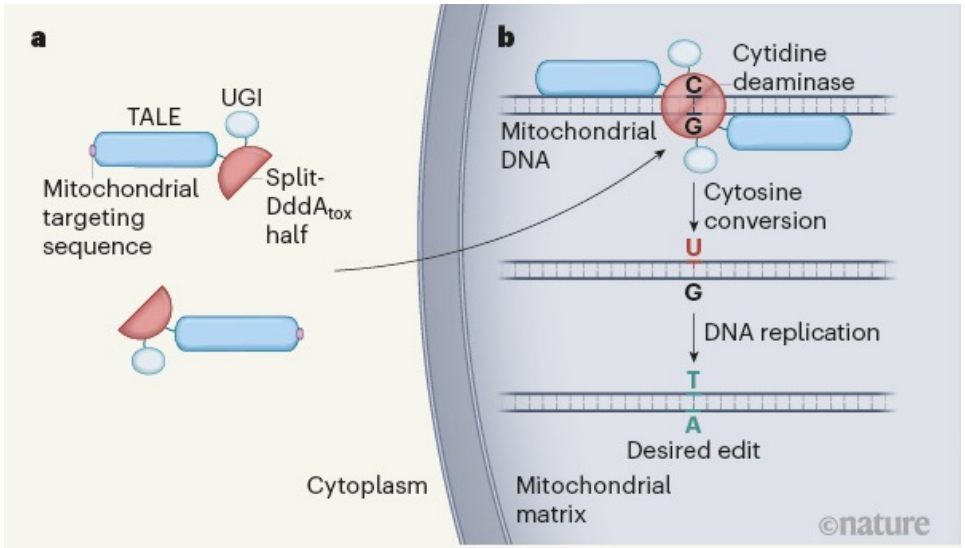
nature

Article | Published: 08 July 2020

A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing

Beverly Y. Mok, Marcos H. de Moraes, Jun Zeng, Dustin E. Bosch, Anna V. Kotrys, Aditya Raguram, FoSheng Hsu, Matthew C. Radey, S. Brook Peterson, Vamsi K. Mootha, Joseph D. Mougous ✉ & David R. Liu ✉





DddA： Double-stranded DNA deaminase toxin A

DdCBE: DddA-derived cytosine base editor

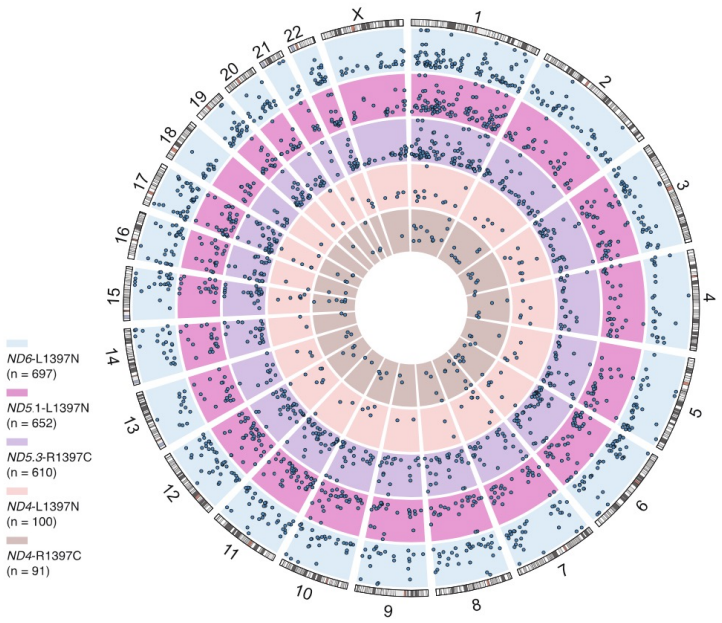
DdCBE可以引起全基因组的脱靶

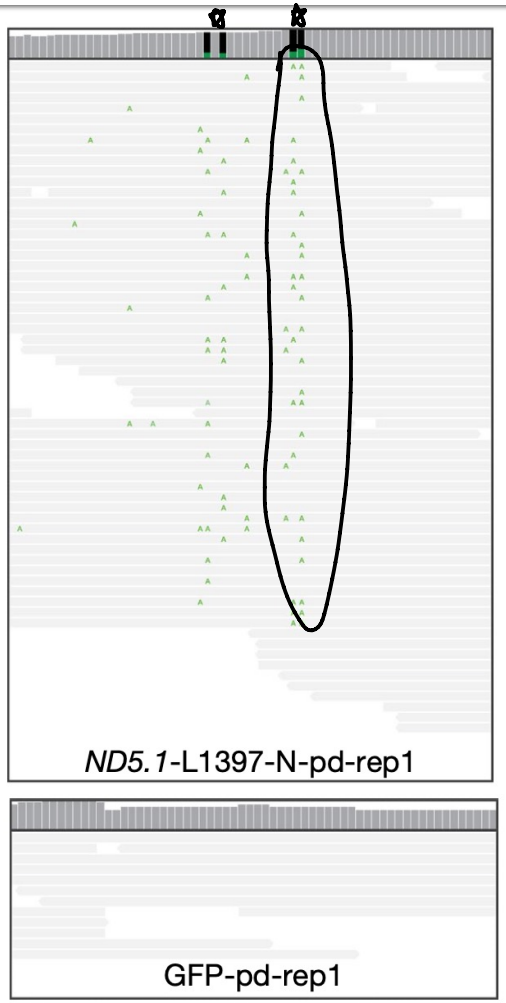
Article

Mitochondrial base editor induces substantial nuclear off-target mutations

https://doi.org/10.1038/s41586-022-04836-5
Received: 10 May 2021
Accepted: 5 May 2022

Zhixin Lei[1,2,9], Haowei Meng[3,9], Lulu Liu[3,9], Huanan Zhao[4,5,9], Xichen Rao[3], Yongchang Yan[1,2], Hao Wu[1,2], Min Liu[1,6], Aibin He[1,6] & Chengqi Yi[1,3,7,8]✉





简化问题

- 已知在Detect-seq测序中，全基因组平均突变率为0.03，某个位点中测到带有C的reads 10条，带有C-to-T突变的reads 20条，那么该位点是否为一个Detect-seq突变位点?

简化问题

- 已知在Detect-seq测序中，全基因组平均突变率为0.03，某个位点中测到带有C的reads 10条，带有C-to-T突变的reads 20条，那么该位点是否为一个Detect-seq突变位点?

- 可以使用二项分布计算pvalue

- 其中n=30，p=~~0.003~~ 0.03

$$b(k;n,p) = P(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \cdots, n,$$

称X服从**二项分布**(binomial distribution)，简记$X \sim B(n,p)$。

获得突变信息并做统计检验

- 程序的输入
  - BAM文件，给定的区域（BED文件）
- 程序的输出
  - 区域内每一个C的突变情况及pvalue

- 程序关键步骤
  - 获得BAM文件中的总突变数，计算突变背景(pileup)；
  - 读取区间，获得区间内每一个位点的突变情况；
  - 计算对应二项分布的pvalue
  - 对得到的pvalue进行校正

BAM文件的pileup操作

- 通过*?bam_file.pileup*获得说明文档
- 几个关键参数设置：
  - contig, start, stop
  - reference
  - max_depth
  - ignore_overlaps (目前版本pysam存在bug)
  - min_base_quality
  - min_mapping_quality

进行统计学检验需要的Python包

- 二项分布检验 scipy
- FDR矫正 statsmodels

Thanks for your attention!
And see you next time!

Haowei MENG
menghaowei@pku.edu.cn
2022-10-03