SOAPfilter--Illumina 数据过滤程序

——st_asm 史玉健

更新：

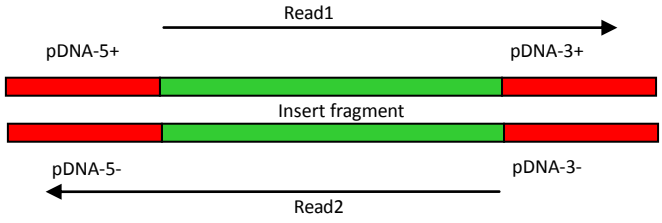
1、增加对使用 native quality 的数据的支持
2、改进 PCR duplication 的过滤算法，降低程序内存占用、减少运行时间
3、过滤 PCR duplication 由原来以 lane 为单位改为以文库为单位，使结果更准确、合理
4、将低质量数据过滤和 PCR duplication 的过滤合并，减少中间文件输出，减少磁盘占用并提高程序效率。

程序功能：

1、过滤 adapter 接头污染的 reads pair
2、过滤 insert size 过小的 reads pair
3、过滤测序质量低的数据
   a) 对于经过 EAMSS 算法对 reads 末端质量值进行 mask B 操作的数据（一般 reads 末端质量值会出现一连串的 B）：
      i. 根据用户设置参数截掉 read 首尾测序质量差的部分
      ii. 过滤掉含有过多（>=cutoff）未知碱基（N）或者为 polyA 的 reads
      iii. 过滤掉含有过多（>=cutoff）低质量值碱基的 reads
   b) 对于未进行 mask B 操作，采用 native quality 的数据：
      根据其质量值估计 reads 整体及单个碱基的错误率，根据用户设置的 reads 错误率和读长的 cutoff 进行过滤
4、过滤 PCR duplication，一个文库中多个拷贝的 reads pair 只保留一份拷贝

程序实现：

1、过滤 adapter 接头污染的 reads pair
   检测由于 insert 片段过短（小于 read 读长）导致 read 末端测到 adapter 序列的情况，示意图如下：



2、过滤 insert size 过小导致 paired-end read 末端有 overlap 的 reads pair
   检测 read1 和 read2 的末端是否有 overlap，如果 overlap>=10bp 则过滤掉。对建库设计为将 insert 片段测通的数据（例如 read 读长 100，insert size 为 170 的数据）不进行该过滤。
3、过滤测序质量低的数据

a) 对于经过 EAMSS 算法对 reads 末端质量值进行 mask B 操作的数据（一般 reads 末端质量值会出现一连串的 B）
过滤低质量碱基数达到或者超过用户设置的 cutoff 的 reads。

b) 对于未进行 mask B 操作，采用 native quality 的数据
质量值为 phred quality（Q=log10(err_rate)），可以通过质量值来估计每个碱基的测序错误率，进而采取如下过滤策略：
由用户设定过滤后数据所允许的最短 read 读长（min_len）和最大允许错误率（max_err），对于估计的整体错误率不大于 max_err 的 reads 直接保留输出；否则选取 read 中错误率不高于 max_err 的最大区域，若该区域长度大于 min_len 则保留，否则将其过滤掉。

4、过滤 PCR duplication，多个拷贝的 reads pair 只保留一份拷贝

a) 随着 reads 读长逐渐增加，靠整个 reads pair 来识别 PCR duplication 势必会导致检测假阴性增加，出于这方面考虑，程序允许用户设置分别截取 read1 和 read2 的部分片段用来识别检测其是否有 PCR duplication

b) 引入 bloom filter 技术，大大减少程序所需内存

c) 以文库为单位过滤 PCR duplication，即使同一个文库不同的 lane 之间的 duplication 依然只保留一份拷贝

使用参数：

SOAPfilter_v2.2 [options] <lane.lst> <stat_file>

-q <int> the quality shift value 64 or 33, if this value is not set, program will detect it automatically
-m <int> the reads pair number in buffer,default: 1000000
-t <int> thread number, default: 8
-i <int> library insert size, default: 500
-y filter reads with adapter
  -F <str> adapter sequence for read1,default: AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
  -R <str> adapter sequence for read2,default: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
-z filter reads with undersize insert size
-p filter PCR duplication
  -s<int> trimmed length at 5' end of read1 when distinguishing duplication, default: 0
  -l<int> the sub-length of read1 used to distinguish duplication,-1 for whole read, default: -1
  -S<int> trimmed length at 5' end of read2 when distinguishing duplication, default: 0
  -L<int> the sub-length of read2 used to distinguish duplication,-1 for whole read, default: -1

-g <int> output type, 0 for text, 1 for gz compressed format, default: 1
-o <str> suffix of output file name, default: clean

-M <int> how to filter low quality reads? 0 for no filtering, 1 for native quality reads mode, 2 for read-end quality masking by EAMSS algorithm(mask B); default: 0

For quality masked by EAMSS algorithm at the end of the read(-M 2):
-f <int> trim flag:-1 for no trimming, 0 for unify trimming, 1 for trimming min(maskB length,-b/-d);default:

0

-Q <int> the maximum low quality value, default: 7

-h output help information

Note:

"lane.lst" include the input read files from same library and their low-quality filtering parameters. The format of it is as below:

For quality masked by EAMSS algorithm at the end of the read(-M 2):

lane1_seq_file1 a b B

lane1_seq_file2 c d W

...

a,b mean the trimmed length at 5' and 3' end of reads in lane1_seq_file1, default: 0

c,d mean the trimmed length at 5' and 3' end of reads in lane1_seq_file2, default: 0

B means the low quality cutoff, reads with >=B% low quality bases(set by -Q) will be filtered, default: 40

W means the N base cutoff, reads with >=W% N bases will be filtered, default: 10

For native quality(-M 1):

lane1_seq_file1 e

lane1_seq_file2 n

...

e means the maximum allowed error ratio(estimated by phred quality) of result reads, default: 0.02

n means the minimum required length of result reads to output, -1 for the raw read length, default: -1

输出结果：

1、过滤后的数据的 fastq 结果文件，根据用户是否设置“-g”参数为文本文件或者 gz 压缩文件

2、一个过滤信息的统计结果（9 列），第一行为表头，每列依次为：

raw_read_id	原始数据的文件名，用于区分同一文库的不同 lane
raw_read_pair_num	原始数据的总 reads pair 数
raw_read_length	原始数据的 reads 读长
raw_base_num	原始数据的总碱基数
low_qual_filter(%)	因测序质量低过滤掉的数据百分比
adapter_filter(%)	因 adapter 接头污染过滤掉的数据百分比
undersize_ins_filter(%)	因 insert size 过小过滤掉的数据的百分比
duplication_filter(%)	因 PCR duplication 过滤掉的数据的百分比
clean_read_pair_num	过滤后数据的总 reads pair 数
clean_read_length	过滤后数据的 reads 读长
clean_base_num	过滤后数据的总碱基数

执行范例：

./test_data/test.sh:

./SOAPfilter_v2.2 -y -z -p -i 500 -M 2 -f 0 -g 0 -o clean lane.lst stat.txt >fil.out 2>fil.err
其中 lane.lst 内容是：
110114_I481_FC81C7HABXX_L5_HUMiqvDBTDIAAPE_1.fq.gz 0 0 40
110114_I481_FC81C7HABXX_L5_HUMiqvDBTDIAAPE_2.fq.gz 0 0 10

注意事项：

1、程序各个功能可以独立实现，根据参数说明设置相应参数即可。
2、Inser size 大小的参数为 -i ，与 Filter_data_5 不同。
3、-M 参数很重要，根据原始数据的质量值模式来设置，通过查看数据的质量值分布图可很明显区分。
4、lane.lst 用来设置原始数据输入文件和相关低质量数据过滤 cutoff，且不同质量值模式对应不同的格式，需格外注意。
5、-s -l -S -L 参数只有当需要过滤 duplication（设置了-p 参数）才有效，且是独立于对于测序质量低数据过滤的 trim 参数的。即以上四个参数是作用在输入的原始数据上面的。
6、对于 native quality 的数据过滤参数（lane.lst 中的“e”）：是允许的单条 read 的估计错误率，即过滤后数据中单条 read 最大的错误率，所以过滤后的 clean data 的整体错误率会小于该值。
7、对于 mask B 的数据，lane.lst 中的 B/W 为低质量碱基或者 N 的百分比，不是个数，且 cutoff 均为闭区间，所以与 Filter_data_5 设置相同的参数结果可能会略有出入。
8、关于内存占用，程序默认读取 1M reads pair 存入内存进行操作，另外过滤 duplication 的话需要额外的内存，且 duplication 越严重所需内存越大。经测试，对于 PE100，100M reads pair（20G 数据）的小片段数据，duplication 率（因 duplication 被过滤掉的 reads 占原始 reads 的比例）0.13%，需要内存 2G 内存；对于 PE90，100M reads pair（18G 数据）的大片段数据，duplication 率 22.5%，需 3G 内存。其余情况可按比例进行估算。

测试结果：

资源消耗测试对比：

数据：103M PE100 的 reads，总碱基数 20.58G
1、Filter_data_5
先过滤低质量数据，参数： -t 8 -m 1000000 -y -z -l 500 -a 0 -b 0 -c 0 -d 25 -B 40 -w 10；即过滤 adapter 污染、undersize insert size、read2 末端统一 trim25bp，过滤单条 read 低质量碱基超过 40 个，或者未知碱基（N）含量超过 10%的 reads pair。然后过滤 PCR duplication.
2、SOAPfilter_v2.2
参数：-t 8 -m 1000000 -g 0 -y -z -p -i 500 -M 2 -f 0；lane.lst：
XXX_1.fq.gz 0 0 40
XXX_2.fq.gz 0 25 10
即过滤 adapter 污染，undersize insert size，PCR duplication，read2 末端统一 trim25bp，过滤单条 read 低质量碱基含量超过 40%，或者未知碱基（N）含量超过 10%的 reads pair。

程序消耗资源信息如下：

	程序	时间（秒）	平均内存占用（MB）	最大内存占用（MB）
Filter_data_5	filter_data_parallel	3029.69	1215.46	1221.16
	duplication	6998.41	5129.09	10110.72
	TOTAL	10028.10	3946.70	10110.72
SOAPfilter_v2.2		3994.10	1711.44	1985.43

测试表明：SOAPfilter_v2.2 无论从时间还是空间上较之前的程序版本（Filter_data_5）都有较明显的提升，过滤 20G 的测序数据只需要 67 分钟，空间占用也只需 2G。

PCR duplication 过滤策略测试对比：

数据：同一个大片段文库的两个 lane 的数据，每个有 lane 44.3M PE90 的 reads，总碱基数 15.95G

1、 按照原策略，过滤 PCR duplication 以 lane 为单位分别过滤，参数：

a) -t 8 -m 1000000 -g 0 -y -z -p –i 2580 -M 2 -f 0；lane1.lst：
XXX_L1_XXX_1.fq.gz 2 3 40
XXX_L1_XXX_2.fq.gz 2 3 10

b) -t 8 -m 1000000 -g 0 -y -z -p –i 2580 -M 2 -f 0；lane2.lst：
XXX_L2_XXX_1.fq.gz 2 3 40
XXX_L2_XXX_2.fq.gz 2 3 10

即过滤 adapter 污染，undersize insert size，PCR duplication，read1 和 read2 均统一 5’端 trim 2bp，3’端 trim 3bp，过滤单条 read 低质量碱基含量超过 40%，或者未知碱基（N）含量超过 10%的 reads pair。注：两个 lane 分别单独过滤，即以 lane 为单位过滤 PCR duplication。

2、 按照新策略，以文库为单位过滤 PCR duplication，参数：

-t 8 -m 1000000 -g 0 -y -z -p –i 2580 -M 2 -f 0；lane.lst：
XXX_L1_XXX_1.fq.gz 2 3 40
XXX_L1_XXX_2.fq.gz 2 3 10
XXX_L2_XXX_1.fq.gz 2 3 40
XXX_L2_XXX_2.fq.gz 2 3 10

即过滤 adapter 污染，undersize insert size，PCR duplication，read1 和 read2 均统一 5’端 trim 2bp，3’端 trim 3bp，过滤单条 read 低质量碱基含量超过 40%，或者未知碱基（N）含量超过 10%的 reads pair。注：两个 lane 组织到一起过滤，即以文库为单位过滤 PCR duplication。

两个测试用例唯一不同在于 PCR duplication 的过滤，其结果对比见下表：

	filter within whole lib		filter within the lane	
	duplicated filter(%)	clean_base_num	duplicated filter(%)	clean_base_num
Lane1	3.41387	5,106,582,400	3.41387	5,106,582,400
Lane2	7.31925	4,818,266,650	3.53619	5,103,178,320

Total	5.36656	9,924,849,050	3.47503	10,209,760,720

测试表明：以文库为单位能识别出更多的 duplication 并将其过滤掉。结果更加准确，进一步降低 duplication 对后期分析的影响。

有任何问题或者建议，请联系：
shiyujian@genomics.org.cn 或 st_asm@genomics.org.cn