# FANSe3使用手册

2016-1-20

## FANSe3目前的限制

* 内存耗用很大。146M的人转录组RefSeq-RNA库需峰值6G内存运行，稳定运行需要3G内存。所以需要预留出至少50倍的内存空间来。
* 即便有无限大的内存，目前的版本也不能超过2G参考序列
* 当存在multi mapping时，mapping到正链可以堆reads到该参考序列的第一个位置。但mapping到负链时会堆reads到该参考序列最后的位置。
* 不能考虑masked genome

## FANSe3参数说明

FANSe3的命令行参数分为参数与开关两种。参数为-X形式，X为大写字母，后面不加空格直接接值，若有空格须用双引号引起整个参数。开关为两个减号开头，有之则为该开关为true，无则为该开关为false。

FANSe3的参数：（注意是一个减号开头，大小写无所谓）

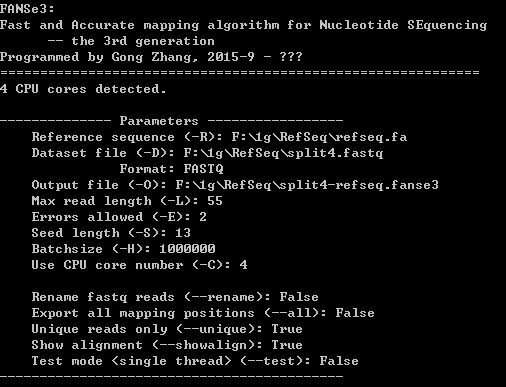
|  |  |  |
| --- | --- | --- |
| -R | 必选 | 参考序列FASTA文件名。可以支持网络路径（即\\开头的UNC路径） |
| -D | 必选 | 测序数据集文件名。目前支持FASTQ与FQC格式（自动判断）。可以支持网络路径 |
| -O | 可选 | 指定输出文件名，可以支持网络路径。若不指定，默认为”dataset-reference.fanse3”，默认路径与测序数据集文件名相同。 |
| -L | 可选 | 测序数据集中的最大读长。若有read长度大于该设定值则将出错退出或者导致mapping结果错误。默认值为2000. |
| -E | 可选 | 错误数量，指mismatch+indel的总和。默认为3（即3个）。  可设置为整数或百分比。设置为整数时，无论reads多长，都允许这么多个错误。例如-E4，表明允许4个碱基的错误。设为百分比时（必须以%结尾），对不同长度的reads以百分比来设置允许错配值，四舍五入到整数。例如-E12%，即允许read长度的12%错误数量。这个设计对于可变长度reads的测序平台（除illumina以外的所有测序平台）非常有用。 |
| -S | 可选 | Seed长度，默认为13. 可设为8-14中的任意整数值，但不建议设为14（有内存溢出风险）。Seed越小则精度越高，但速度会越慢。通常建议设置13。对reads很短的情况（如miRNA等应用，reads小于25nt）建议设为10-11。  Reads的长度不得小于Seed长度，否则会自动扔掉这个read（算作不可匹配）。 |
| -I | 可选 | 是否开启indel检测。0为不开启，1为开启。默认为0.  保留这个参数是为了和FANSe2的参数体系相兼容。FANSe3里面还有一个开关--indel是干同样的事情的，两者选其一即可。 |
| -C | 可选 | 并行核数，可设为任何整数。默认为自动检测系统的CPU核数（逻辑核心数，即超线程会被识别为2个物理核心）。不要设置超过系统的逻辑核心数，否则性能将严重下降。 |
| -H | 可选 | 每批次读取的reads数量，单位为百万，即-H1表示每批次读取100万reads。经过测试，不建议超过2. |
| -T | 可选 | 截取Reads的部分。形式为-Ta,b 例如-T5,75  两个数值a和b，中间以逗号分隔。a为扔掉reads开头的a个碱基；b为保留多少个碱基。例如-T5,75为扔掉开头的5个碱基，然后将6-80位（共75nt）保留下来，其后的也扔掉。 |

FANSe3的开关：（注意是两个减号开头，大小写无所谓）

|  |  |
| --- | --- |
| --rename | 强制重命名reads名称为1,2,3,…。若无此参数，则输出的结果中reads名称完全按照fastq文件中的来。  FQC文件因不储存reads名称，因此此开关无效（有和没有都一样，都会重命名为1,2,3…）。 |
| --all | 输出所有最佳的mapping位点。 |
| --unique | 将unique mapped reads和multi mapped reads分别输出为两个文件 |
| --showalign | 在fanse结果文件中直接输出比对结果。”.”为能比对上，”x”为错配。目前暂不支持indel。开启此开关大约会降低一半的速度，并显著增大结果文件的体积。在开indel时必须同时开启--showalign，否则当负链有indel时报告的mapping位置会有少量误差。 |
| --indel | 开启indel检测。作用等同于-I1。 |
| --test | 以单线程模式运行，用于排错。 |

从FANSe2迁移到FANSe3时，可以直接用FANSe2的命令行参数，以保持兼容性。那些FANSe3被取消掉的参数将自动忽略。当然，我们仍然强烈推荐

各种参数会在FANSe3运行一开始列出来，包括取默认值的参数，以供确认：



目前版本的FANSe3最大支持的参考序列为略小于1G。处理>150M的参考序列时需要将FANSe3.exe.config文件放在FANSe3.exe同一目录下。

## 输出结果说明

**最普通的形式是：**

Fanse3 -RF:\1g\RefSeq\refseq.fa -DF:\1g\RefSeq\split4.fastq -L55 -E2 -S13

这将产生两个文件：



其中，split4-refseq.fanse3存放已mapping的结果，split4-refseq.unmapped存放未mapping上的reads。

未mapping上的reads按如下方式存放：

11 CAGATGAGGAAGAACCAAAAGATGATATCGTATGCC

13 CTGCTACAAGAACCCACTGCAGATGCAGC

14 GCTTAATTTGACCCAACACGG

18 GAACACAGGCAGCAGGTGCTGAGGTC

21 ATGCCACTCCTTTCCCGTGCTCCAG

25 GCTAACTTTCAAGAGGTGCAG

26 GATAGCACTATTGCACTCCAGCCT

29 CAAAGTTTGGAATAGTCAG

39 AGAAATTCAATGAAGCGCGGGTAAATTCGTATGCCG

40 TCTGGCATGTGGAACAATGT

49 GGAGGACGAGGACGAAACCTGGAG

即 <名称> tab <序列> 的方式。若没有--rename开关，则<名称>完全依照原FASTQ文件中的reads名称。

已mapping上的reads，基本格式如下，两行为一个read的mapping结果：

2 TCTGGCACGGTGAAGAGACATGAGAG

F NR\_003287 0 3929 2

第一行两个字段：<名称> tab <read序列>

第二行五个字段：

<正负链>：F代表正链，R代表负链

<mapping参考序列的名称>

<错配数>

<mapping到参考序列上的起始位置(0-based)>

<multi mapping次数>：若该read有多个最佳匹配位点（错配数相同且最少，称为multi mapped read），则该字段表示有多少个最佳匹配位点。若只有一个最佳匹配位点（称为unique mapped read），此字段为1.

最普通的情况下，若出现multi mapping，会将read mapping到所有参考序列中第一次出现的位置上。若mapping到负链上，会mapping到负链转正链的第一个位置（即最后一个位置上）。

下面来加点开关：

首先，如果加入--showalign开关，则结果的第一行会有三个字段。第三个字段即为比对的结果，直观地给出哪里有错配。例如：

42628 AGCAAGGACTAACCCCTATACC .................x....

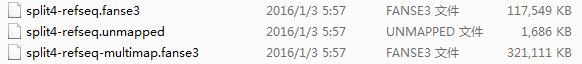
F NM\_001190470 1 142 1

第一行后面直接给出了序列比对的结果，倒数第5个碱基为错配。这样不用再行与参考序列比对即可知道哪里有错配。

这样做确实很方便，但代价是速度减慢约一倍，并且输出的文件量也会大一些。

然后来看看加了--unique开关的效果：

加了--unique开关之后，会输出三个文件：



此时，split4-refseq.fanse3里面只有unique mapped reads（即只有一个最佳匹配位点的reads），而有multi mapping的reads都被放到了split4-refseq-multimap.fanse3里面去了。这样仍然一次完成了所有reads的mapping任务，若应用严格需要unique mapped reads，则直接取第一个文件即可。若还需要multi mapped reads，则直接取-multimap.fanse3文件即可，无需重新mapping。

如果加了--all开关，则会将multi mapping的所有最佳匹配位点全部输出出来。若没有--showalign开关，则性能几乎没有损失。若与--showalign开关联用，则会拖慢3倍以上的速度。

输出的multi mapping结果示例：

369061 AGCTGGTACAGAAAGCCAAATTCGCTG ....................x......,....................x......

F,F NM\_003404,NM\_139323 1 405,310 2

可以注意到，比对结果、正负链、参考序列、位置号 都以逗号分隔多个mapping结果。

这个read可以被mapping到两个地方，都是1个错配，都是mapping在正链上，分别mapping到NM\_003404,NM\_139323两个参考序列上，位置号分别为405,310。