454 Amplicon sequencing 数据过滤流程说明文档

1. 实验原理：

1.1 4-Primer Amplicon Tagging

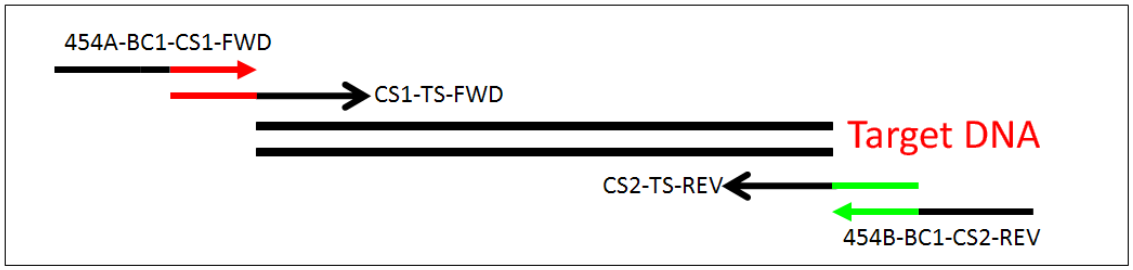
本实验使用 2 对引物对样品进行扩增。

第一对引物：CS (Consensus Sequence) + TS (target-specific) 序列
Forward: 5'-ACACTGACGACATGGTTCTACATGCCCTAAACGTTCCGAAAAA-3'
Reverse: 5'-TACGGTAGCAGAGACTTGGTCTCCCTGTTTCCAGCCAGAATCC-3'

第二对引物：454 Adapter sequence (key), barcode + CS
Forward: 5'- CGTATCGCCTCCCTCGCGCCATCAGACGAGTGCGTACACTGACGACATGGTTCTACA-3'
Reverse: 5'- CTATGCGCCTTGCCAGCCCGCTCAGACGAGTGCGTTACGGTAGCAGAGACTTGGTCT-3'

1.2 过程

在对样品核苷酸进行扩增时，依次加入第一对引物和第二对引物，如下图所示：



TS 是我们要测序部分的引物序列，CS 的作用是为了第二对引物的扩增。当使用了第一对引物后产物是：CS1->TS1->Target->RC_TS2->RC_CS2 和 CS2->TS2->Target->RC_TS1->RC_CS1 （其中 RC 表示反向互补）。

当使用了第二对引物后产物是 Adapter (key)->Barcode-> CS1->TS1->Target->RC_TS2->RC_CS2->RC_Barcode->RC_Adapter(key) 和 Adapter (key)->Barcode-> CS2->TS2->Target->RC_TS1->RC_CS1->RC_Barcode->RC_Adapter(key)。其中 Barcode 的作用是区分不同的样品，这样做的好处是我们可以在同一个 Run 中同时测很多的样品，这样可以减少成本。

2. 问题描述：

我们上面已经知道了，完整的 PCR 产物是 Adapter (key)->Barcode-> CS1->TS1->Target->RC_TS2->RC_CS2->RC_Barcode->RC_Adapter(key) 和 Adapter (key)->Barcode-> CS2->TS2->Target->RC_TS1-

>RC_CS1->RC_Barcode->RC_Adapter(key)。如果 Target 序列较短，那么由于 454 读长相对较长（平均 400bp-500bp），得到的序列 3’端就可能含有 RC_CS2->RC_Barcode->RC_Adapter(key)和 RC_CS1->RC_Barcode->RC_Adapter(key)的一部分或者全部。

下机数据拆分组目前的流程是根据 5’端 Barcode 来拆分不同的样品，对于 3’端的 Barcode 以及 Adapter 未作处理。本流程便是通过寻找 5’和 3’端的 CS 序列来去除 CS 序列及 CS 之外的 Barcode 和 Adapter 等污染序列的。实际上最优的办法是获得 TS 引物序列，需找 5’和 3’端的 TS 引物序列，然后去除 TS 之外的 CS，Barcode 和 Adapter 等污染序列。但是由于这些引物往往是客户自己设计的，除非客户提供，我们无法知道，所以采取的是次优的方法，寻找 CS 并去除 CS 序列及 CS 之外的 Barcode 和 Adapter 等污染序列。

3.用法和参数

3.1 用法

perl SSH.pl [Opriont s] [<sample_name> <sff_file>]

[<sample_name> <sff_file>] 在定义了-l 参数后是可选的

3.2 参数

-sff2xxx <str>　　parameters for HOME/bin/Sff2xxx.pl, default -sff2xxx="-s -q";

-remove_CS <str>　　parameters for HOME/bin/remove_CS.pl, default -remove_CS="-cat yes -remove yes";

-fas <str>　　parameters for HOME/bin/FAS.pl, default -fas="-Q 20 -B 40 -w 20 -s 0";

-fa2fq <str>　　parameters for HOME/bin/Fa2Fq.pl,default -fa2fq="-manner Sanger";

-l or -lst <str>　　*.sff files list;

-p or -project_id <str>　Project ID, such as ASPxfbD, defualt SSH

-h or -help　　Print this information

3.3 参数说明

-l or -lst <str>　　为 SFF 文件列表文件，每一行一个样品名称和一个 SFF 文件；格式为第一列为样品名称，第二列为对应的 sff 文件；如果一个样品有多个文件，写成多行。

-p or -project_id <str> 子项目编码，此处只用作输出文件夹和打包文件名，不会获取项目信息。

其余参数说明见第 4 节

4. 过程

4.1 第一步：转化 SFF 格式为 FA 格式（*.fa）和对应的质量文件（*.qual）

4.2.1 使用脚本：HOME/bin/Sff2xxx.pl

4.1.2 允许在 SSH.pl 中定义的参数及说明（除非定义了-a，否则必须包含-q 和-s）：

-s or -seq Output just the sequences, default yes # 输出序列的 FA 格式

-q or -qual Output just the quality scores, default default yes # 输出上述序列对应的 qual 文件

-f or -flow Output just the flowgrams, default no

-t or -tab Output the seq/qual/flow as tab-delimited lines, default no

-n or -notrim Output the untrimmed sequence or quality scores, default no

-m or -mft Output the manifest text, default no

-p or -plain Output the plain text, default no # 转变 SFF 文件所有内容到一个文本文件

-a or -all Output the all above files, default no # 输出以上所有文件

4.1.3 在 SSH.pl 中对应的参数-sff2xxx <str>，定义方式-sff2xxx="-s -q"

4.2 第二步：去除 CS 及之外的污染序列

4.2.1 使用脚本：HOME/bin/remove_CS.pl

4.2.2 允许在 SSH.pl 中定义的参数及说明：

-max_mismatch <int> default 2 nucletides # CS 与序列匹配时的最大错配数

-max_gap <int> default 2 nucletides # CS 与序列匹配时的最大 gap 数

-cat <yes|no> cat the *.A/B.*.* to *.deCS, default no # 是否把不同的分类结果 cat 起来

-remove <yes|no> remove the sequnce without CS, default no # 是否去除不包含 CS 的序列

4.2.3 在 SSH.pl 中对应的参数-remove_CS <str>，定义方式-remove_CS="-cat yes -remove yes"

4.2.4 比对算法：

使用的是 SmallRNA 流程的 DPalign.pm 模块，调用的是其 local_affine 函数；

打分方式：Match 2; Mismatch 1; Gap -3; Gap_Extension -1

4.3 第三步：过滤

4.3.1 使用脚本：HOME/bin/FAS.pl

4.3.2 允许在 SSH.pl 中定义的参数及说明：

-Q <int> Base with less value than this is defined as low quality bases, default 20 #质量值低于此值为低质量碱基

-B <int> filter reads with >X percent base are low quality bases, set a cutoff, default 40 # 低质量碱基百分比超过此值被过滤

-w <int> filter reads with >X percent base are Ns, set a cutoff, default 20 # 含 N 的百分比低于此值被过滤

-s <int> filter reads with length shorter than this threshold value, default 0 nucleotides. # 读长低于此值被过滤

4.3.3 在 SSH.pl 中对应的参数-fas <str>，定义方式-fas="-Q 20 -B 40 -w 20 -s 0"

4.4 第四步：格式转换

4.4.1 使用脚本：HOME/bin/Fa2Fq.pl

4.4.2 允许在 SSH.pl 中定义的参数及说明：

-manner <Sanger|Solexa> Choose the transcoding manner between QA and ASCII, default Solexa，if you choose Sanger, ASCII = QA + 33; if you choose Solexa, ASCII = QA + 64; # 质量文件*.qual 转换成 fq 文件的 ASCII 编码方式，Sanger 方式为：ASCII 码=质量值+33；Solexa 方式为：ASCII 码=质量值+64

4.4.3 在 SSH.pl 中对应的参数-fa2fq <str>，定义方式-fa2fq="-manner Sanger"

5. 结果文件

5.1 简单说明

DATA/filter.SAMPLE.fa # 结果 FA 文件
DATA/filter.SAMPLE.fq # 结果 FQ 文件
DATA/filter.SAMPLE.qual # 结果 QAUL 文件
DATA/raw.SAMPLE.fa.CsOnly #只包含 CS 的序列
DATA/raw.SAMPLE.fa.CsWrong #包含 CS 位置关系错误的序列
DATA/raw.SAMPLE.fa.deCS #去取 CS 及 CS 之外的污染序列，FA 格式
DATA/raw.SAMPLE.fa.deCS.fq #去取 CS 及 CS 之外的污染序列，FQ 格式
DATA/raw.SAMPLE.fa.gff #记录每条 read 去除 CS 的信息
DATA/raw.SAMPLE.qual.deCS #去取 CS 及 CS 之外的污染序列，QUAL 文件
filter.len #过滤后 read 长度值
filter.length_distribution.svg #过滤后长度分布图
filter.length_distribution.txt #过滤后长度分布
filter.quality_distribution.svg #过滤后质量分布图

filter.quality_distribution.txt　　#过滤后质量分布
filter.statistic.xls　　#过滤后统计文件
raw.len #原始数据 read 长度值
raw.length_distribution.svg　　#原始数据长度分布图
raw.length_distribution.txt　　#原始数据长度分布
raw.quality_distribution.svg　　#原始数据质量分布图
raw.quality_distribution.txt　　#原始数据质量分布
raw.statistic.xls #原始数据统计文件
SSH.log #日志文件

5.2 *.fa.gff 说明

该文件是包含去除 CS 的信息，跑完后可以检查该文件是否正常，共 8 列：

第一列：Read Id

第二列：Raw Read 总长

第三列：clean read 起始

第四列：clean read 终止

第五列：使用的搜索 CS1 序列（或 CS2）

第六列：Read 中的 CS1 序列（或 CS2）

第七列：使用的搜索 RC_CS2 序列（或 RC_CS1）

第八列：Read 中的 RC_CS2 序列（或 RC_CS1）

注：（1）RC 表示反向互补；

（2）-表示 Read 中没有搜索到改 CS 序列

（3）搜索算法在搜索 RC_CS 时采用的是先搜索完整的 RC_CS 序列，如果不存在再搜索前面的 11bp，如果仍没有结果，不再搜索。

6. 注意事项

1. 目前的流程提高运行速度的方式是采用 fork 并行，可以通过修改-remove_CS="-cat yes -remove yes -para 500"中的-para 参数修改并行数目，但请采用 qsub 方式，否则任务多可能会使节点死掉。如果样品较多建议分开投任务，这样可以节约时间。后续会优化仍在。

2.为了增加可扩展性，采用的是在主脚本中使用一个参数定义调用的脚本的参数，可在主脚本中定义的个调用脚本的参数已经在上述文档中说明，或者通过运行 perl SSH.pl (-h)查得，而且都已经有了默认值。

7.工具

HOME/tools/Fa2Fq.pl #fa 和 fq 格式相互转换

HOME/tools/FAS.pl # 根据 fq 和 quality 文件过滤

HOME/tools/remove_CS.pl # 去除 CS 及之外的污染序列

HOME /tools/Sff2xxx.pl # sff 格式转化为 fa 和 quality 文件

8. 流程目录结构

```
.
|-- bin
|   |-- SSH.pl -> SSH_v2.0.pl
|-- doc
|   `-- README.pdf
|-- example
|-- lib
|   |-- DPalign.pm
|   |-- ForkManager.pm
|   `-- QASCII.pm
|-- opt
|   |-- fastaDeal.pl
|   |-- fishInWinter.pl
|   |-- line_diagram.pl
|   `-- sffinfo
|-- V1.0
|-- V1.5
|-- V2.0
`-- tools
    |-- FAS.pl -> ../bin/FAS.pl
    |-- Fa2Fq.pl -> ../bin/Fa2Fq.pl
    |-- Sff2xxx.pl -> ../bin/Sff2xxx.pl
    `-- remove_CS.pl -> ../bin/remove_CS.pl
```

DNA 组 邓操

华大科技生物农业部

2012-10-09