

Seq-Analysis User Manual

文档作者	王萌
提交日期	2012-03-22
更新说明	2012-04-01 增加配置文件中各参数的详细说明
文档备注	

1. 简介

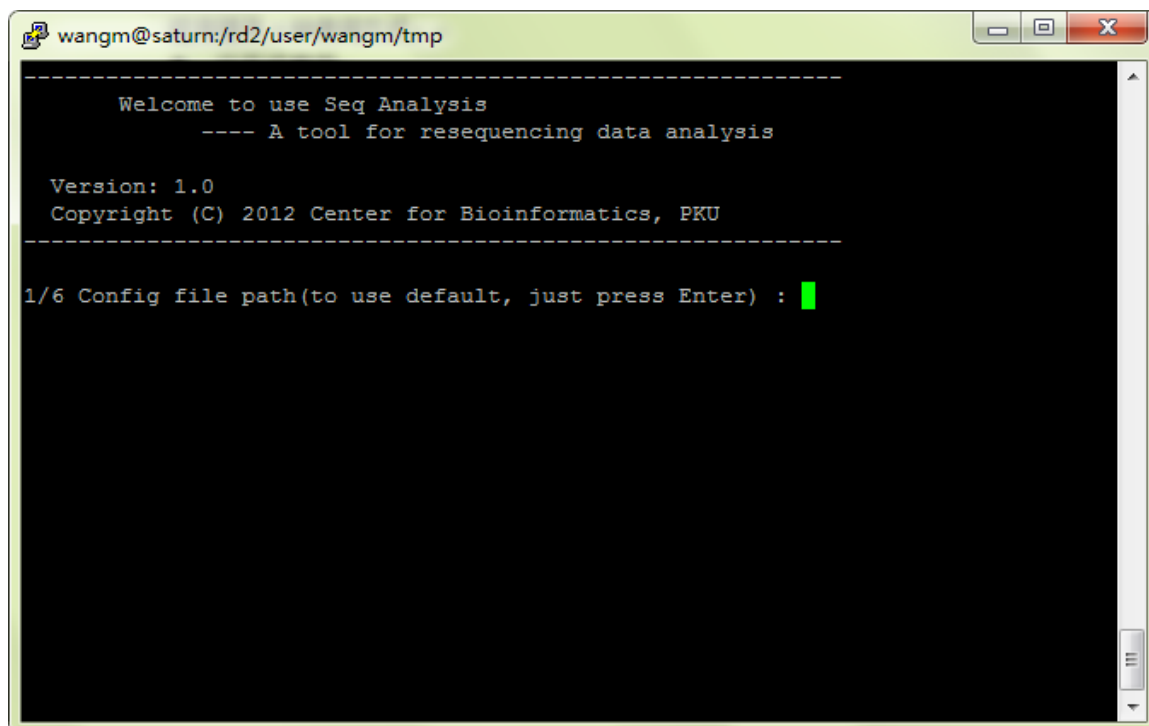
Seq-analysis 是基于 CBI 制定的高通量测序数据分析 protocols 的集成分析工具，包括 Reads mapping、Call SNP and Indel 及 Call structure variation。程序集成了数据分析中的各个工具(目前 Reads mapping 使用 BWA，Call SNP and Indel 使用 GATK，Call structure variation 使用 Pindel)，同时将各工具相关参数默认设为优化后的标准值。

程序提供 3 种使用方式：

- 交互式使用
- 非交互式使用
- 单步使用

2. 交互式使用

(1) 进入工作目录，在命令行中输入 seq_analysis



The screenshot shows a terminal window with the title bar 'wangm@saturn:/rd2/user/wangm/tmp'. The terminal content is as follows:

```
-----  
Welcome to use Seq Analysis  
---- A tool for resequencing data analysis  
  
Version: 1.0  
Copyright (C) 2012 Center for Bioinformatics, PKU  
-----  
1/6 Config file path(to use default, just press Enter) : █
```

(2) 根据提示输入

- 1/6 输入配置文件路径。如果采用默认配置文件，则直接回车。如果采用自己的配置文件，则输入文件路径。
- 2/6 测序平台。默认为 Illumina。
- 3/6 是否删除中间结果文件。默认为删除。
- 4/6 运行线程数。默认为 3 个线程。
- 5/6 库长度(Library size)。
- 6/6 Lane 数。默认为 1。

接下来按序输入各 fq 文件。

```
wangm@saturn:/rd2/user/wangm/tmp
-----
Welcome to use Seq Analysis
---- A tool for resequencing data analysis
-----
Version: 1.0
Copyright (C) 2012 Center for Bioinformatics, PKU
-----
1/6 Config file path(to use default, just press Enter) :
2/6 Platform(default is illumina) :
3/6 Remove intermediate files([y/n], default is y) :
4/6 Number of threads(default is 3) :
5/6 length between paired-end adapters : 200
6/6 Number of sample lanes(default is 1) : 2
Enter pair end reads 1 of lane 1 (.fq or .fq.gz): PE1_1.fq
Enter pair end reads 2 of lane 1 (.fq or .fq.gz): PE1_2.fq
Enter pair end reads 1 of lane 2 (.fq or .fq.gz): PE2_1.fq
Enter pair end reads 2 of lane 2 (.fq or .fq.gz): PE2_2.fq
-----
That's OK! Pipeline begin at Thu Mar 22 21:15:57 CST 2012
To see details of each program's running message, please
refer to the log directory.
This pipeline may run several hours or several days,
wish you have a good time during these days~
-----
```

然后程序会自动运行，直到这个 pipeline 结束或出错。在屏幕上会显示主要事件的时间、状态等信息。程序运行的具体信息可以在 log 文件夹下查看。

Mapping 输出结果文件为 map_result.sorted.bam

Call SNP and Indel 输出文件为 raw.snp.vcf raw.indel.vcf flt.snp.vcf flt.snp.vcf flt.indel.vcf

Call structure variation 的输出结果在 sv 文件夹下。

3. 非交互式使用

程序可以接收命令行参数以非交互式方式运行。通过执行 seq_analysis -h 可以查看各参数及其功能。使用方式及参数如下：

Usage: seq_analysis [option] -l <int> -n <int> <PE1_1.fq> <PE1_2.fq> [PE2_1.fq] [PE2_2.fq]...

-c STRING	config file path[default]
-p STRING	platform[illumina]
-r	keep intermediate files
-t INT	number of threads[3]
-n INT	number of lanes[1]
-l INT	library size(Mandatory)
-v	program version
-h	help

其中-l INT 为必选项。例：

```
seq_analysis -l 200 PE1_1.fq PE1_2.fq
```

```
seq_analysis -c ./myconfig -r -t 5 -l 500 -n 2 PE1_1.fq PE1_2.fq PE2_1.fq PE2_2.fq
```

程序输出同交互式使用。

4. 单步使用

可以单独使用 Reads mapping、Call SNP and Indel、Call structure variation. 这些脚本位于程序所在目录的相应子目录下。不带任何参数直接执行相应脚本会给出各命令行参数的相关信息。

5. 日志

在程序运行过程中会向屏幕输出主要的日志信息。详细的日志及程序运行信息输出在用户工作目录的 log 文件夹下。包括 4 部分：

主要事件日志： journal
Mapping 过程输出日志： reads_mapping.log
Call SNP and Indel 过程输出日志： call_variants.log
Call structure variation 输出日志： call_sv.log
Map 结果统计信息： map_result.sta

主要事件日志包含用户配置信息、各工具版本信息及事件信息，格式如下：

日期时间	事件信息	运行时间	状态
------	------	------	----

统计结果示例：

	6724	+	0		in total (QC-passed reads + QC-failed reads)	
	0	+	0		duplicates	
	6678	+	0		mapped (99.32%:-nan%)	
	6724	+	0		paired in sequencing	
	3362	+	0		read1	
	3362	+	0		read2	
	3492	+	0		properly paired (51.93%:-nan%)	
	6636	+	0		with itself and mate mapped	
	42	+	0		singletons (0.62%:-nan%)	
	1896	+	0		with mate mapped to a different chr	
	94	+	0		with mate mapped to a different chr (mapQ>=5)	

6. 配置文件

默认配置文件位于程序目录下。内容如下：

```

1  # reference files path
2  REF_PATH=/rd2/user/wangm/reference          #path of the reference genome
3  REF_NAME=hg19                             #name of the reference genome
4  DBSNP_PATH=/rd2/user/wangm/reference/dbsnp_132.hg19.vcf
5  #
6  GATK_HOME=/rd2/user/wangm/tools/GATK        #directory of GATK
7  PICARD_HOME=/rd2/user/wangm/tools/picard     #directory of Picard
8  #
9  # BWA related parameters
10 FQ_VERSION=1.5    #when fq version is 1.5, -I option should be set when bwa aln
11 END_IND=10        #bwa aln -i option value
12 GAP_EXT=-1        #bwa aln -e option value
13 #
14 # GATK related parameters
15 STAND_CALL_CONF=30      #gatk UnifiedGenotyper -stand_call_conf
16 STAND_EMIT_CONF=30      #gatk UnifiedGenotyper -stand_emit_conf
17
18

```

用户如果想改变个别参数，可以将该默认配置文件拷贝到自己目录下，修改相应参数。在运行程序时指定该配置文件即可。

各参数详细说明：

REF_PATH	参考序列的存放路径
REF_NAME	参考序列名称，例如参考序列为 hg19.fa，则参考序列名称为 hg19。 (注：参考序列的后缀名必须为.fa，不能是.fasta)
DBSNP_PATH	dbsnp.vcf 的存放路径，在 GATK 中会用到。
GATK_HOME	GATK jar 包的存放路径
PICARD_HOME	Picard jar 包的存放路径
FQ_VERSION	fastq 文件版本。目前主要有 1.3+、1.5+、1.8 三种版本。其中 1.3+和

	1.8 中的 base quality 是加 33 的，而 1.5 是加 64 的。对于 1.5 版本的 fastq 文件，在做 BWA 时会加上 -I 选项。
END_IND	序列末尾最大 Indel 长度，对应 bwa aln 中的 -i 选项。
GAP_EXT	Maximum number of gap extensions，对应 bwa aln 中的 -e 选项。
STAND_CALL_CONF	The minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be called.
STAND_EMIT_CONF	The minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be emitted (and filtered if less than the calling threshold).