### 基本分析&高级分析

环境变量设置：

/home/xingsk/workdir/pipeline/PacBio\_RNA\_Denovo/Bin/setup\_for\_PacBio\_RNA\_Denovo\_pipeline.txt

根据这个文件将所需的路径加到个人目录下的.bashrc文件中后，source ~/.bashrc

在进行主程序分析前，建议对三代下机数据先进行污染排查。

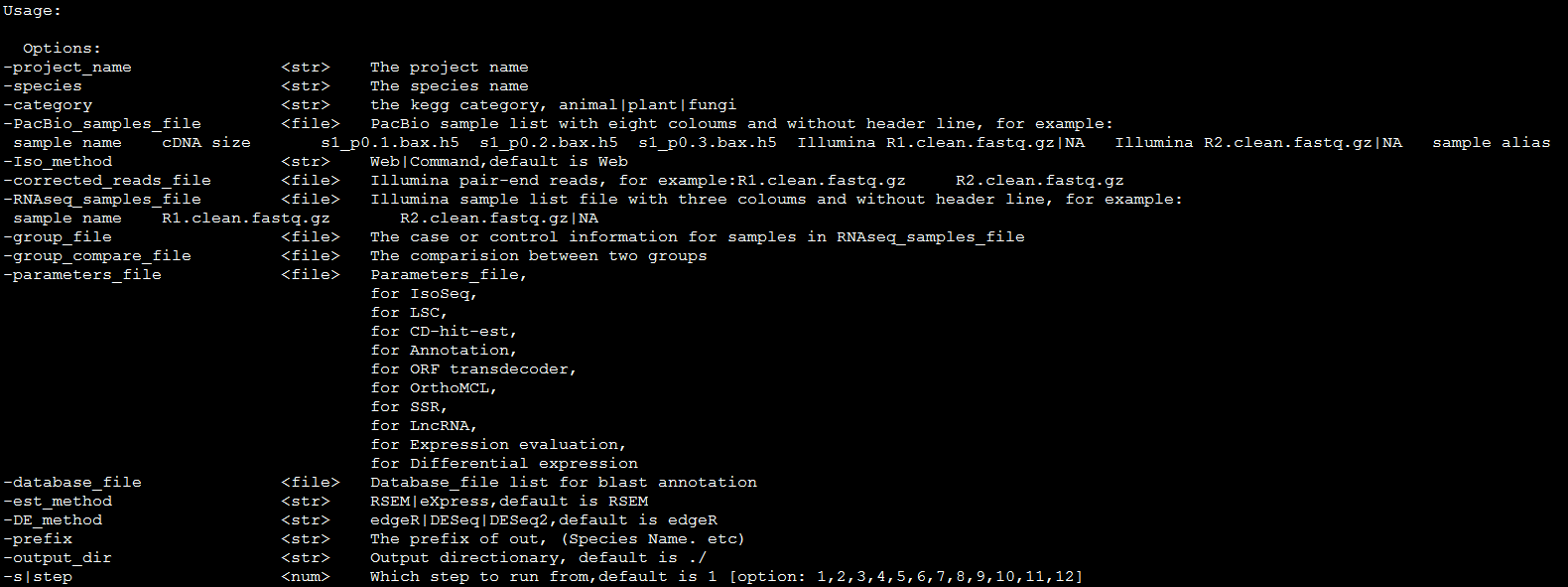
利用subread的fasta文件截取其中若干行同nt库比对。脚本可参照

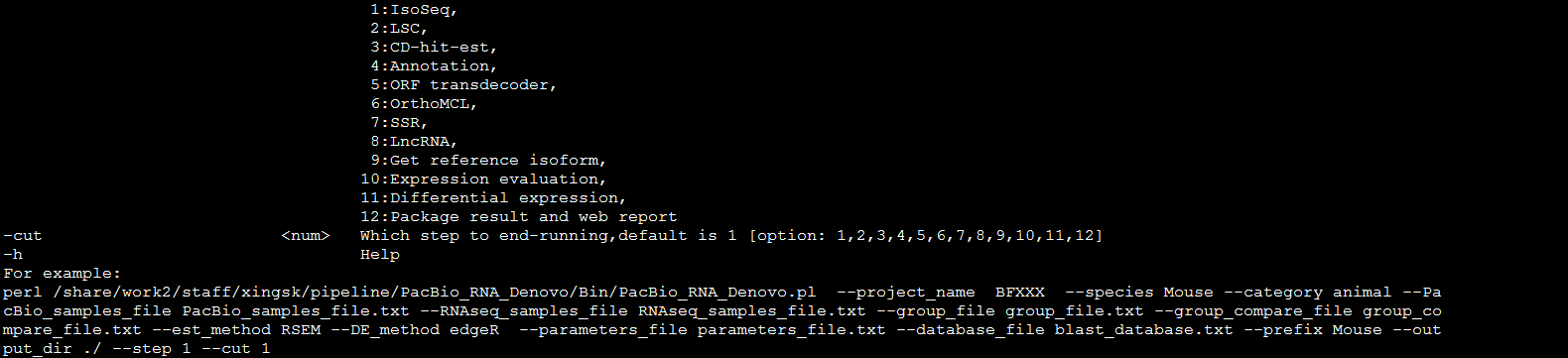
/home/xingsk/workdir/pipeline/PacBio\_RNA\_Denovo/Bin/Pollution.blast.sh

主程序路径：

/home/xingsk/workdir/pipeline/PacBio\_RNA\_Denovo/Bin/PacBio\_RNA\_Denovo.pl

主程序运行方式：





----project\_name，后接项目名称

--species，后接物种名称

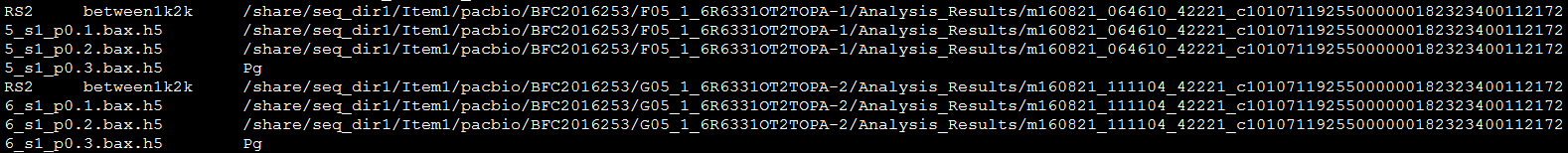
--category, 后接animal|plant|fungi中一种，用于KEGG注释

--Iso\_method，后接IsoSeq的方法，有Web和Command两种，默认前者，数据量大是选用后者

--corrected\_reads\_file，后接文件名称，每一行可以有若干tab键分割的字段，第一个字段是样矫正的样品名称，同PacBio\_samples\_file中一致；其它字段是二代reads文件完整路径



--PacBio\_samples\_file，后接文件名称，包含样品三代转录组测序数据相关信息，共有6列。



第一列：样品名称，只能由字母和数字组成

第二列：cDNA文库大小，具体可以选择under1k,between1k2k,between2k3k,above3k 任一个

第三列：第一个bax.h5文件

第四列：第二个bax.h5文件

第五列：第三个bax.h5文件

第六列：样品的简写名，一般填属名和种名的首字母

--RNAseq\_samples\_file，后接文件名称，包含高级分析中样品二代转录组测序数据相关信息，共有3列。



第一列：样品名称，只能由字母和数字组成

第二列：样本的二代转录组数据为双端测序时，填R1端fastq.gz文件；为单端时，填fastq.gz文件

第三列：样本的二代转录组数据为双端测序时，填R2端fastq.gz文件；为单端时，填NA

--group\_file，后接文件名称，文件中包含高级分析中二代转录组样品分组的相关信息



第一列：分组名称，只能由字母和数字组成，开头不为数字

第二列：样品名称，只能由字母和数字组成，开头不为数字

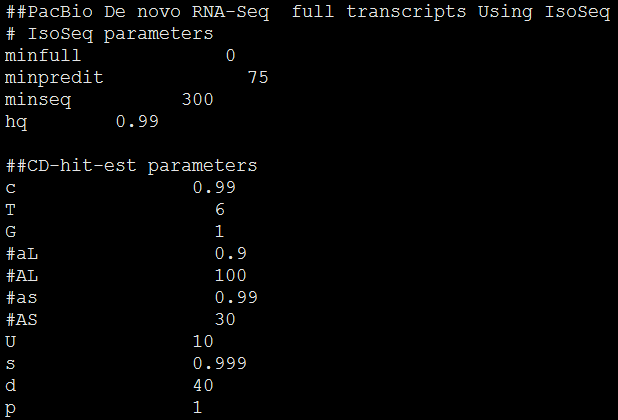
注：当样品没有组别时，第一列可以和第二列相同

--group\_compare\_file，后接文件名称，文件中每行对应高级分析中要进行差异表达分析的两个分组

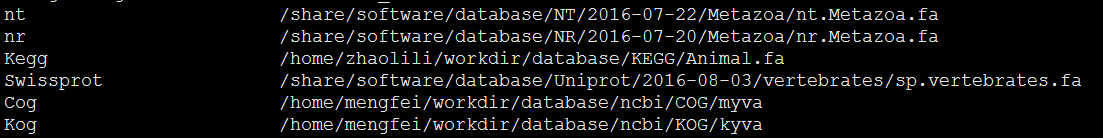


--DE\_method 可选方法有edgeR|DESeq|DESeq2，是高级分析中二代样本基因差异表达的方法

--parameters\_file ，后接文件名称，文件中对应IsoSeq、CD-hit-est、Annotation、differential expression的参数设置



--database\_file， 后接文件名称，文件中给出Annotation所需要的数据库的路径，根据具体项目需要检查更新



--prefix, 后接前缀名，用于基因家族分析（OrthoMCL）、SSR分析、高级分析中表达定量及差异表达、生成结果报告等过程；可以和物种名一样。

实例路径：

/home/xingsk/workdir/project/PacBio\_RNA\_Denovo/BFC2016253/PacBio\_RNA\_Denovo\_run.sh



可以根据以下依赖关系，在PacBio\_RNA\_Denovo\_run.sh中设置step和cut，即起始步骤和结束步骤。直接命令行跑sh PacBio\_RNA\_Denovo\_run.sh，大概2分钟跑完 (不可以nohup、也不可以qsub)。运行该脚本后会在输出目录下生成相应模块的路径。

|  |  |
| --- | --- |
| 模块 | 依赖关系 |
| 01 IsoSeq | 无 |
| 02 LSC | 01 IsoSeq、有对应的二代数据 |
| 03 CD-hit-est | 01 IsoSeq（无二代矫正数据时）；  02 LSC （有二代矫正数据时） |
| 04 Annotation | 03 CD-hit-est |
| 05 Transdecoder | 03 CD-hit-est |
| 06 OrthoMCL | 05 Transdecoder |
| 07 SSR | 03 CD-hit-est |
| 08 LncRNA | 03 CD-hit-est |
| 09 reference | 03 CD-hit-est |
| 10 Expression evaluation | 09 reference、若干二代样本数据 |
| 11 Differential expression | 10 Expression evaluation |
| 12 Package result and web report | 上面选做的分析都完成，在当前工作路径下worksh文件夹中.sh文件数和.mark文件数相等时，执行该步骤 |

注：

1）第10步中，如果一些样本的表达量跑完，另一些由于集群问题中断的，重跑这步时，原先跑完的样本会跳过并在worksh文件夹下生成带failed的文件，只需删掉带failed的文件就可以了；

2）在跑第12步生成web报告之前，要PacBio\_RNA\_denovo\_report\result\base\_info.json文件中加入例如{"info":"共2个样本各组织均匀混合，1~2K文库共测3个cell；2~3K文库共测3个cell，3~6K共测2个cell。"} 对样品测序cell的描述信息；要在\home\xingsk\workdir\project\PacBio\_RNA\_Denovo\BFC2016253\PacBio\_sample.xls中整理如下表格：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **样本名** | **cell编号** | **文库类型** | **Polymerase  Read Bases** | **Polymerase  Reads** | **Polymerase  Read N50** | **Polymerase  Read Length** | **Polymerase  Read Quality** |
| **Sample1** | **A-1** | 1-2K | 1292766343；  1232578974 | 150292；  67621 | 32966；33314 | 8601；  18227 | 0.387；  0.82 |
|  |  |  |  |  |  |  |  |

这里含两个数值的格分别是过滤前后的。

3）具体项目的流程图可能需要做一定修改更换图片

/home/xingsk/workdir/pipeline/PacBio\_RNA\_Denovo/Bin/jade2html/public/img/pipeline.png

### 个性化分析

个性化分析没有写在流程里，是直接用BFC2014375项目中已有的脚本。

项目路径：

/home/jiangdezhi/workdir/Program/RNA\_seq\_ref/BFC2014375\_Rice/BFC2014375/Specific\_Analysis