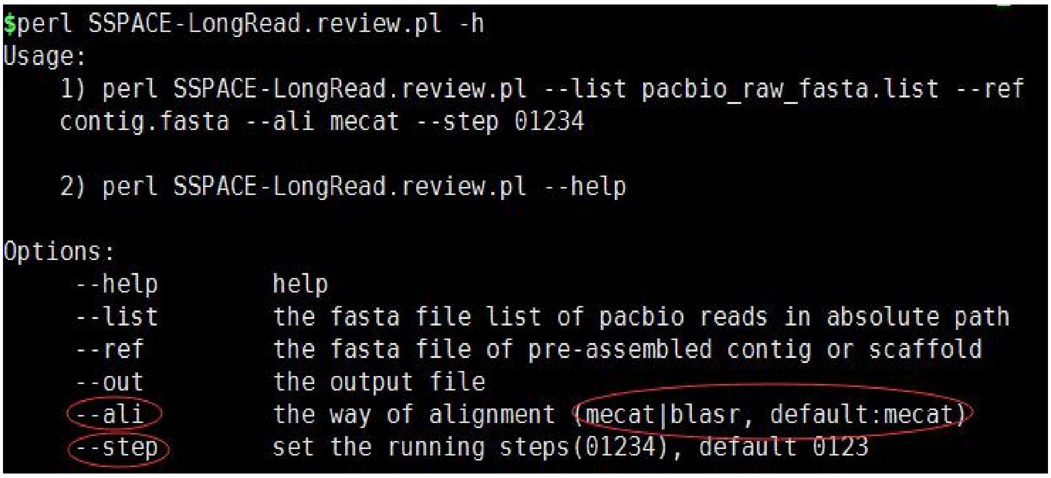
SSPACE-LongRead 优化说明文档

优化：比对软件 blasr 替换为 mecat 或 minimap2

1.流程：



```
$perl SSPACE-LongRead.review.pl -h
Usage:
    1) perl SSPACE-LongRead.review.pl --list pacbio_raw_fasta.list --ref
    contig.fasta --ali mecat --step 01234

    2) perl SSPACE-LongRead.review.pl --help

Options:
     --help         help
     --list         the fasta file list of pacbio reads in absolute path
     --ref          the fasta file of pre-assembled contig or scaffold
     --out          the output file
     --ali          the way of alignment (mecat|blasr, default:mecat)
     --step         set the running steps(01234), default 0123
```

参数：

--list　　fasta 格式的 pacbio raw read 文件路径列表（绝对路径）

--ref　　需要连接的 contig 或者 scaffold

--out　　输出路径，默认当前文件夹（./）

--ali　　比对软件（mecat|blasr），默认为 mecat

--step　　流程步骤（01234），默认为 0123

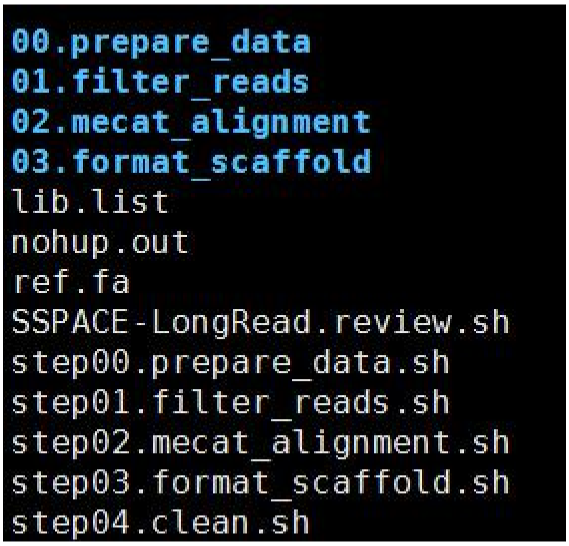
step00. 将 pacbio raw reads 整合成一个 fasta 文件；

step01. 如果 mecat 比对，过滤长度大于 100k 的序列；如果 blasr 比对，将序列以 1G 为单位拆分；

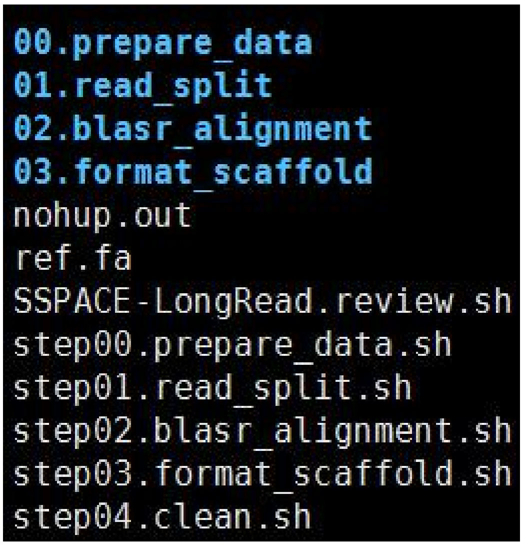
step02. 采用 mecat、minimap2 或 blasr 进行比对；

step03. 将 contig 根据比对结果连成 scaffold；

step04. 删除大的中间文件。



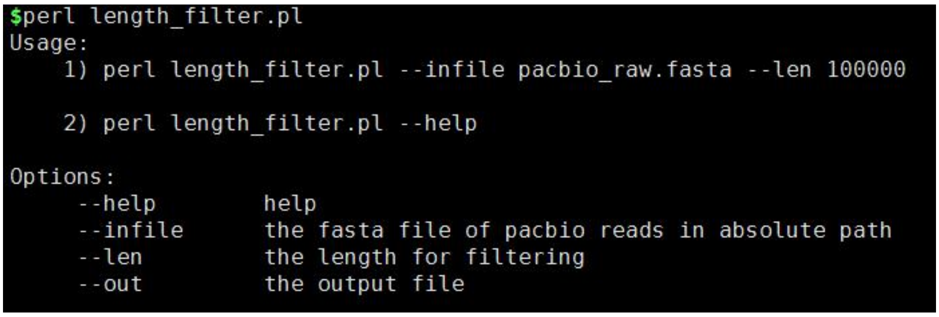
```
00.prepare_data
01.filter_reads
02.mecat_alignment
03.format_scaffold
lib.list
nohup.out
ref.fa
SSPACE-LongRead.review.sh
step00.prepare_data.sh
step01.filter_reads.sh
step02.mecat_alignment.sh
step03.format_scaffold.sh
step04.clean.sh
```



```
00.prepare_data
01.read_split
02.blasr_alignment
03.format_scaffold
nohup.out
ref.fa
SSPACE-LongRead.review.sh
step00.prepare_data.sh
step01.read_split.sh
step02.blasr_alignment.sh
step03.format_scaffold.sh
step04.clean.sh
```

step00: step00.prepare_data.sh

cat *.subreads.fasta >pacbio.merge.subreads.bam.fasta step01:　step01.filter_reads.sh



```
$perl length_filter.pl
Usage:
    1) perl length_filter.pl --infile pacbio_raw.fasta --len 100000

    2) perl length_filter.pl --help

Options:
     --help        help
     --infile      the fasta file of pacbio reads in absolute path
     --len         the length for filtering
     --out         the output file
```

过滤掉长度大于 100K 的序列

```
perl length_filter.pl --infile *.subreads.fasta --len 100000 --out mecat_test/01.filter_reads
```

Step02:　step02mecat_alignment.sh

mecat 比对

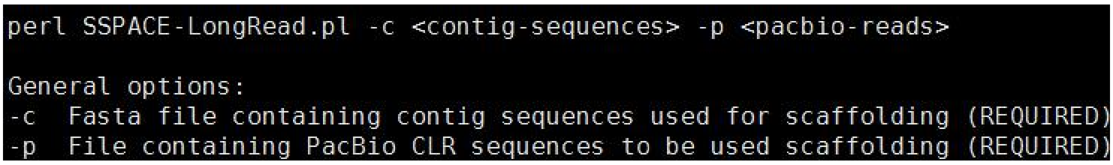
```
export
LD_LIBRARY_PATH=$LD_LIBRARY_PATH:$Bin/../software/hdf5/lib;export
PATH=$Bin/../software/MECAT/Linux-amd64/bin:$PATH;export
PATH=$Bin/../software/DEXTRACTOR:$PATH;
mecat2ref -d *.subreads.fasta.less100000.fa -r ref.fa -w wrk_dir.N -t 16 -o
mecat2ref.N.out -m 1 -x 0
```

替换比对 ID 并输出与 blasr 相同的 m1 格式

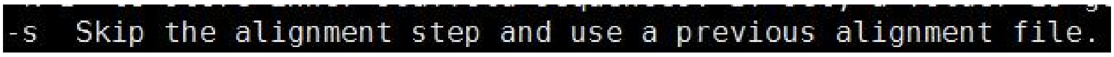
```
perl change_alignment_mecat_to_blasr.ID.pl *.subreads.fasta.less100000.fa
mecat2ref.N.out mecat2blasr.N.out
```

Step03:　step03.format_scaffold.sh

```
perl SSPACE-LongRead.pl -c ref.fa -p pacbio.merge.subreads.bam.fasta -s
1 -b PacBio_scaffolder_results
```



```
perl SSPACE-LongRead.pl -c <contig-sequences> -p <pacbio-reads>

General options:
-c  Fasta file containing contig sequences used for scaffolding (REQUIRED)
-p  File containing PacBio CLR sequences to be used scaffolding (REQUIRED)
```



```
-s  Skip the alignment step and use a previous alignment file.
```

step04:　step04.clean.sh

```
rm -rf 00.prepare_data/* 01.filter_reads/* 02.mecat_alignment/wrk_dir.*
02.mecat_alignment/mecat2blasr.*.out 02.mecat_alignment/mecat2ref.*
```

2. 优化结果比较

黄梁木 SSPACE-LongRead 结果，全部 pacbio raw reads

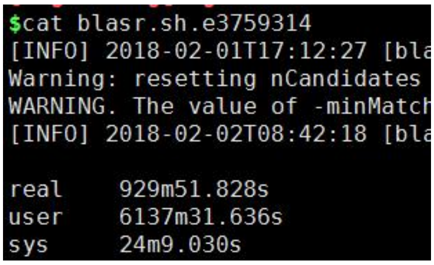
#Title	Total_len	Total_num	Average_len	Max_len	N50_len	N50_num
Contig	715797514	3587	199553	5359931	598594	299
blasr	722166616	1851	390149	11791812	1249035	148
mecat	722581833	1671	432424	9044902	1042479	183

山梨 SSPACE-LongRead 结果，全部 pacbio raw reads

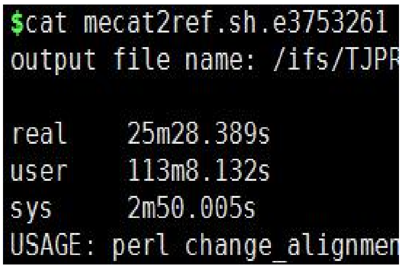
#Title	Total_len	Total_num	Average_len	Max_len	N50_len	N50_num
origin	639825632	999	640466	16181927	2882448	69
blasr	642116045	587	1093894	33856258	5548795	32
mecat	642884676	414	1552861	24217045	7946167	24
minimap2	641969645	629	1020619	23023348	5494316	33
mecat_v2	642456360	377	1704128	33365141	9881326	19
minimap2_v2	641923953	627	1023802	21763267	5470150	34

3.运行速度比较

5G 的 fasta 文件，mecat（25min）比 blasr 在 8cpu 比对时，速度快 30~60 倍，比切分成 1G 文件投递，快 6~12 倍

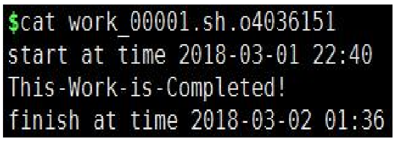


```
$cat blasr.sh.e3759314
[INFO] 2018-02-01T17:12:27 [bla
Warning: resetting nCandidates
WARNING. The value of -minMatch
[INFO] 2018-02-02T08:42:18 [bla

real    929m51.828s
user    6137m31.636s
sys     24m9.030s
```

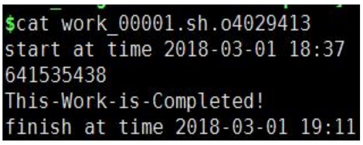


```
$cat mecat2ref.sh.e3753261
output file name: /ifs/TJPR

real    25m28.389s
user    113m8.132s
sys     2m50.005s
USAGE: perl change_alignmen
```

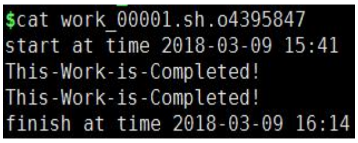
8G 的 fasta 文件，blasr 比对（2h56min*8）mecat 比对(34min)，minimap2 比对（33min）



```
$cat work_00001.sh.o4036151
start at time 2018-03-01 22:40
This-Work-is-Completed!
finish at time 2018-03-02 01:36
```



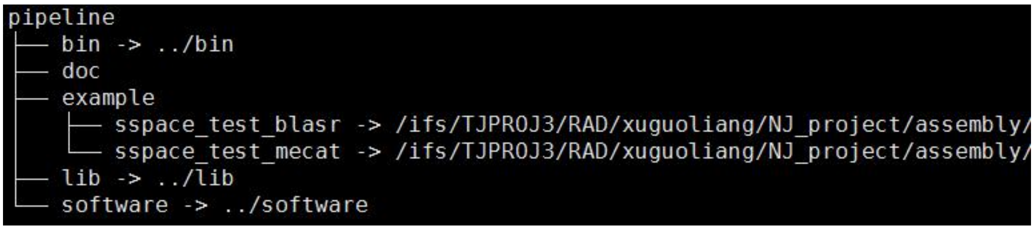
```
$cat work_00001.sh.o4029413
start at time 2018-03-01 18:37
641535438
This-Work-is-Completed!
finish at time 2018-03-01 19:11
```



```
$cat work_00001.sh.o4395847
start at time 2018-03-09 15:41
This-Work-is-Completed!
This-Work-is-Completed!
finish at time 2018-03-09 16:14
```

4.流程路径及目录结构

/ifs/TJPROJ3/RAD/xuguoliang/NJ_project/assembly/sspace-longread/pipeline



```
pipeline
├── bin -> ../bin
├── doc
├── example
│   ├── sspace_test_blasr -> /ifs/TJPROJ3/RAD/xuguoliang/NJ_project/assembly/
│   └── sspace_test_mecat -> /ifs/TJPROJ3/RAD/xuguoliang/NJ_project/assembly/
├── lib -> ../lib
└── software -> ../software
```

Copy 到其他位置请赋予所有文件执行的权限

chmod 755 -R */pipeline