Runjelly_v1.3.1

Overview

Runjell_v1.3.1 是 pbjelly 的优化版。

- 在运行 pbjelly 之前用 minimap2 进行过滤，去除与补洞无关的 reads,比对速度可以提升 8-10 倍。
- Pbjelly 使用的 blasr1.3 会出现内存回收不完全而导致的内存飙升问题，例如比对一个 15G 的 fastq，mapping 内存峰值可以达到 80G,如果资源不够会出现任务被杀的情况；runjelly 的比对软件使用 blasr5.3，没有这个问题。Blasr5.3 的实际补洞效果也略好于 blasr1.3。
- 可以自动投递任务。

Installation

集群路径：
阿里云：*/ALBNAS12/Plant/Project/WORK/PBjelly/software/tar/runjelly_v1.3.1.tar.gz*
将软件包拷贝到流程目录下，解包：

```
$ tar -zxvf runjelly_v1.3.1.tar.gz
```

完成后，进入目录，运行 setup.sh:

```
$ sh setup.sh
```

将检查 pip list 中是否包含 networkx1.1。（若没有将尝试通过 pip 安装 networkx1.1,安装路径在 runjelly_v1.3.1/pythonlib 中）

Requirements

- Python 2.7+
- Python package: networkx 1.1+ (networkx 2.0 及以上版本会报错）
- blasr5, bedtools，minimap2,libhdf5 (这些软件在软件包中已包含）

Usage

1、准备好 reads .fastq 文件和 reference .fasta(<4G)文件。
如果是 bam 格式，用 prepare.sh 脚本（此脚本需要 samtools）转换为 fastq 格式：

```
$ path_to_bin_dir/prepare.sh input_bam_dir Out_fastq_dir Size Opts
```

Input_bam_dir: 所有 bam 文件所在目录路径
Out_fastq_dir: 保存输出 fastq 文件的目录路径
Size:每个 fastq 文件的大小，单位为 G,建议设置为 2 或者 3
Opts：集群队列信息，例如-P aliyun -q alyun.q,alpag01.q

如果已经有 fastq 文件，fastq 文件的大小最好相接近，这样比对所需时间相近，也方便分配内存大小和线程数；可以用 fastq_split.sh 脚本将一个目录下的 fastq 文件分割成大小相同的 N 份(以 vf=1g,p=1 投递即可）：

$ path_to_bin_dir/fastq_split.sh input_fastq_dir output_fastq_dir N

2、建立 pbjelly 输出文件夹，将软件包目录下的 example.cfg 文件拷贝到输出文件夹下，修改其内容：

```
#input
fastq_dir=/ALBNAS12/Plant/Project/WORK/PBjelly/workdir/fastq_2g
ref=/ALBNAS12/Plant/Project/WORK/PBjelly/workdir/pbjelly_0321/HLM.fasta

#parameter
blasr="--minMatch 8 --minPctIdentity 70 --bestn 1 --nCandidates 10 --maxScore -500 --noSplitSubreads "
dist=2000
vf=10g
p=4
opts="-P  aliyun  -q  alyun.q,alpag04.q"

#output
out_dir=/ALBNAS12/Plant/Project/WORK/PBjelly/workdir/pbjelly_0326_2k
email=libenping@novogene.com

```

fastq_dir: fastq 文件存放目录。
ref: reference 文件路径。
blasr:blasr 参数，一般不需要修改。
dist:minimap2 过滤时挑选 gap 周围多少 bp 的 reads,默认为 2000。dist 值越小，挑选的 reads 越少，比对速度相应也越快，但补洞指标可能会因此降低。
vf:每个任务投递的内存，一般设置为 10g 即可。
p:每个任务投递的线程数。
opts:其他投递参数，例如队列，没有可以不填。
out_dir:输出路径。
email:任务完成后发送一封邮件到指定邮箱，可以不填(阿里云测试无法发送邮件）。

3、在输出文件夹下运行：

$ path_to_bin_dir/runjelly.sh example.cfg

Tips

1、如果 reference>4G,由于 sawriter 的内存限制，将无法生成.sa 文件。请用 bin 目录下的 fasta-splitter.pl 将 fasta 文件切成小于 4G 的若干份，分别跑 runjelly，跑完后将 jelly.out.fasta 合并。
2、如果需要重新运行程序，在输出目录下运行:

$ sh path_to_bin_dir/cleanup.sh

即可停止运行，并且删除除了*.cfg 以外的所有文件。
如果只想停止而不删除文件，可以在输出目录下运行：

$ sh ./scripts/stop_jelly.sh

Contact

如有任何问题，请联系 zhouyiqi@novogene.com

测试结果：

山梨(640M 基因组，120G fastq):

比对软件	contig number	contig N50（kb）	scaffold number	scaffold N50(kb)	gap数	比对CPU时间（h）
原始数据	996	2882	737	4935	259	/
blasr1.3	924	3363	710	5383	214	1112
blasr5.3	919	3467	706	5383	213	1176
minimap2过滤1+blasr 5.3	919	3467	707	5383	212	223
minimap2过滤2+blasr 5.3	922	3390	709	5383	213	150

过滤 1：过滤范围 2000bp

过滤 2: 过滤范围 1500bp

黄梁木（700M 基因组，174G fastq）

比对软件	contig number	contig N50（kb）	scaffold number	scaffold N50(kb)	gap数	比对CPU时间（h）
原始数据	3584	599	1850	1135	1734	/
blasr1.3	2725	796	1823	1143	902	5356
minimap2过滤+blasr 5.3	2723	801	1823	1146	900	481