pilon][https://github.com/broadinstitute/pilon/wiki]

输入fasta 组装后文件和比对到fasta 组装后文件的一个或多个bam文件, 使用read 比对分析识别输入fasta组装文件和bam文件中的不一致情况, 进而提高fasta 组装文件质量:

- 单碱基差异
- 小的indels
- 大的indel或重组(block substitution events)
- 补洞
- 识别局部组装错误

输出改善后的fasta 组装序列文件和包含read和fasta 组装文件差异的VCF文件

Methods of Operation

输入文件为fasta格式基因组和一个或多个包含序列比对输入基因组的BAM文件. BAM文件需根据位置顺序排序且索引. 对于Illumina数据, 使用BWA或BOWTIE2生成BAM文件, pilon使用三种类型BAM文件:

- Fragments: 短插入序列的双端read数据, 一般小于1kp
- Jumps: 长插入序列的双端read数据, 一般大于1kp
- Unpaired: 单端测序read数据

强烈推荐使用双端测序read, 建议read长度至少75bp, 且覆盖度最小50X, 100X的深度更佳. 同时pilon还可以使用更长的reads, 例如一代测序序列, PacBio 校正后的reads, circular-consensus序列. 当时, pilon当前不适用于原始PacBio reads数据.

Output files

pilon生成一修正后的fasta基因组文件, 校正了snp, small indel, gap filling, including all single-base, small indel, gap filling, misassembly and large-event corrections from the input genome. 针对组装而言, 该过程提高了组装的一致性, 对于变异检测, 输出的基因组经过修正后更接近给定样本.

除了输出VCF文件记录基因组的snp和indel信息外, 也可以输出一个"changes"文件, 记录输入和输出文件改变信息及位置. pilon还可输出一系列共IGV查看的可视化路径文件(visualization track, bed/wig).

[Output File Descriptions] [https://github.com/broadinstitute/pilon/wiki/Output-File-Descriptions]

默认输出文件前缀为pilon.*: `--output` `--outdir`

输出改善后fasta文件序列名称添加_pilon后缀以区别 `--fix`

使用`--change`, 输出pilon.change文件, 记录scaffold修改信息

使用`--vcf`, 记录差异信息

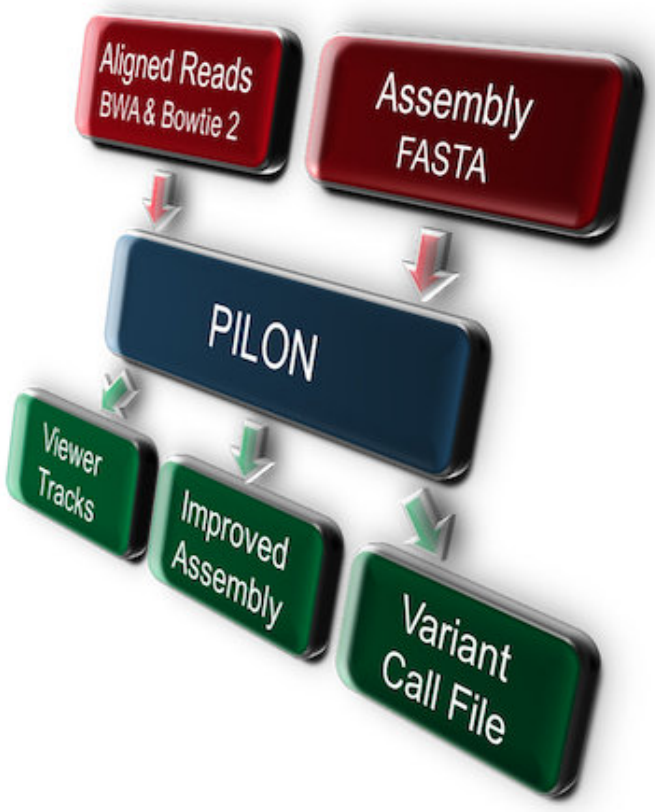
VCF Pileups

Calls are classified by small number VCF FILTER tags:

Filter Tag	Description
PASS	A passing call, either reference confirmation or difference
Amb	Ambiguous; significant evidence for more than one allele at this position. Meant for haploid genomes, this filter tag is suppressed by the `--diploid` argument, as it will result in a heterozygous call for a diploid genome
LowCov	Valid read coverage less than the threshold controlled by the `--mindepth` argument
Del	Provides pileup information for loci which were removed by a variation in another line; this gives a sense of the alignment evidence at that locus had the larger variation not been called

使用`--tracks`, 输出.bed/.wig文件用于可视化查询

Requirements & Usage




```
java -Xmx16G -jar <pilon.jar>
```

```
pilon --genome genome.fasta [--frags frags.bam] [--jumps jumps.bam] [--unpaired unpaired.bam]
```

`--genome`: genome.fasta 输入用于改善的基因组, 同时也是read比对生成bam文件的参考序列

`--frags` frags.bam, bwa或bowtie2比对双端reads得到的bam文件, 插入序列小于1kp

`--jumps` jumps.ba, 同上, 插入序列大于1kp

`--unpaired` unpaired.bam, 同上, 单端reads

`---bam` any.bam 任意, pilon自动识别区分

`--output` 输出前缀

`--outdir` 输出文件路径

`--changes` 输出改变信息文件

`--vcf` 输出vcf差异文件/ `--vcfqe` 包含QE(quality-weighted evidence)的改变信息文件

`--tracks` 输出.bed/.wig文件, 用于查看

`--variant` 设置heuristics用于变异检出, 相对于改善组装过程, 等同于`--vcf --fix all,breaks`

`--chunksize` 拆分fasta文件分析, 默认10M

`--diploid` 样本来自二倍体物种

`--fix` 逗号分隔的fix分类信息: snps, 修正snp; indels, 修正小indels; gaps, 填补gaps; local, 检测并修正局部错误组装; all, 以上所有. 另外处于实验阶段: breaks, 允许局部重组开启新的gaps; circles, 使用长修正后reads时实现环状基因组闭合; novel, 使用为比对的non-jump reads组装新的序列

`--dumpreads` dump reads for local-assemblies

`--dumplicates` 使用输入bam中标记了的duplicates的reads, 默认取消该参数

`--targets` 逗号分隔用于修正的fasta序列, 例如： scaffold00001, scaffold00002:10000-20000

`--debug` 输出debugging信息

HEURISTICS:略

Analysis flow

```
##https://github.com/broadinstitute/pilon/wiki

##组装后fasta和以fasta为参考基因组比对的bam文件,bwa/bowtie2
ref=icu43_contigs.fasta
index=index/icu43_50x
r1=ICU43_galore_R1_trimmed_1P_val_1.fq.gz
r2=ICU43_galore_R2_trimmed_2P_val_2.fq.gz
out=icu43_50x

mkdir index
bwa index -p $index $ref

bwa mem $index $r1 $r2 | samtools sort -O bam -o ${out}_align.bam

java -jar /home/huizhen/bin/picard-tools-1.119/MarkDuplicates.jar \
I=${out}_align.bam \
O=${out}_align.dedup.bam \
REMOVE_DUPLICATES=true \
METRICS_FILE=${out}.dedup.metric

samtools view -b -q 30 ${out}_align.dedup.bam > ${out}_align.dedup.filtered.bam

samtools index ${out}_align.dedup.filtered.bam

###pilon polish
pilon --genome ${ref} --frags ${out}_align.dedup.filtered.bam --outdir ${out}_pilon --output ${o
ut}_polish --vcf --tracks --changes --fix all --debug
```