HiSeq Human DNA Resequencing Data Analysis Protocols

Reads Mapping

文档作者	彦林林
提交日期	2012-03-14
更新说明	
文档备注	未完善，需验证

Input:

1. Sequencing data: FASTQ files (Pair-end)
2. Reference genome: FASTA files (Download URL)

Output:

1. Alignments: BAM files

By BWA:

1. Reference sequence index (Only run at first time):

```
a) bwa index -a bwtsw hg19.fa
```

2. Alignment:

```
a) bwa aln -t 2 hg19.fa <(zcat XXX_1.fq.gz) -f XXX_1.sai
b) bwa aln -t 2 hg19.fa <(zcat XXX_2.fq.gz) -f XXX_2.sai
```

```
3. Output SAM:
   a) bwa sampe -P -r ‘@RG\tID:XXX\tSM:XXX\tPL:Illumina\tLB:XXX’
      hg19.fa XXX_1.sai XXX_2.sai <(zcat XXX_1.fq.gz) <(zcat
      XXX_2.fq.gz) | samtools view -Sb - > XXX.bam
   b) rm XXX_?.sai
4. Sort:
   a) run_picard.sh SortSam SO=coordinate I=XXX.bam
      O=XXX.sorted.bam CREATE_INDEX=true
   b) rm XXX.bam
5. Statistics:
   a) samtools idxstats
   b) samtools flagstat
   c) (depth)
   d) (unique mapped)
Comments:
1. In alignment (bwa aln):
   a) -i 10 (default is 5, will check in future).
   b) -e 10 (default is -1, may be useful in SV calling, with -L).
```

2. In output sam (bwa sampe):
   a) -a (default is 500, change as library, exome library is 200 + 3 * sd).
   b) -c (ref. to UCSC hg19, need to be checked).
3. Change .fa to .fn
4. hg19 vs. GRCh37 (need to be checked, mappable reads & variants).