Lectures on sequence analysis: Low-level sequence data processing

Hyun Min Kang,
Mark Kate Trost,
and Goncalo Abecasis

March 9th, 2011

What to start with

- 1000 Genomes pilot 3 (exon-targetted) sequence read (FASTQ)
  - Located at `/home/hyun/wed/input/*.fastq.gz`
  - Individual `NA12878`, who is also sequenced deeply in whole genome scale
  - We are using only a subset of sequence reads that mapps to a 300kb region in chr20
- Softwares to be used (located at `/home/hyun/wed/bin/`)

  bwa A rapid aligner of sequence reads

  samtools Software for manipulating sequence alignment files
  - We will `samtools-hybrid`, a modified version of `samtools` combining new and deprecated features

  SuperDeDuper A tool for mark/remove duplicated reads
  - Similar to Picard's MarkDuplicate
  - Handles overlapping paired-end fragments

  qplot A tool for assessing the quality assessment BAM files

Additional resources

Located at `/home/hyun/wed/ref`

- FASTA file (NCBI build 37, chr20 only)
  - Index files were pre-created by `bwa` and `qplot`
- dbSNP database
- GC content database

Aims for today

1. Understand sequence reads format (FASTQ)
2. Map sequence reads to reference genome
3. Merge multiple alignment files
4. View the sequence alignment format (SAM/BAM)
5. Mark duplicated reads
6. Visualize alignment to reference genome

Step 0 : Setting up environmental variables

- Visit out wiki page at *http://goo.gl/jTBqa* for easier copy-and-paste

Type or copy-paste the following commands

```
setenv BIN /home/hyun/wed/bin
setenv IN /home/hyun/wed/input
setenv REF /home/hyun/wed/ref

setenv OUT ~/seq/wednesday/output
mkdir --p ${OUT}
```

Step 1 : Understanding FASTQ format

Example commands

```
workshop:~/wed> zcat ${IN}/NA12878.exon.sample.read1.fastq.gz | head
@SRR014820.333067/1
GAGGTGTTTTGGATATTTCAGGTGGAAGGCACAGCT
+SRR014820.333067/1
>>?<@@@@@@@<;A?A@@<3@=@@=><@<;?:=>9B
@SRR014820.335184/1
GGCTCGGTCACAGGCTCAAGGGTTGGATCAAAGAGA
+SRR014820.335184/1
>4:C<85@<@;@?);D=/A56=2@8=<A@8A9:;5;
@SRR014820.337501/1
GCTCTCGGGTGCATCACAACACATGTCCCTCATTCA
```

More FASTQ input files

NA12878.exon.sample.read2.fastq.gz

Contains the other read for each read-pair, in the same order to the first-reads

```
workshop:~/wed> zcat ${IN}/NA12878.exon.sample.read2.fastq.gz | head -4
@SRR014820.333067/2
ATAGCCACTGTACCTTGACTTCTTCGGACACTTAGG
+SRR014820.333067/2
8@>?:=>;DAA@<?EAB@<EA>EA>:>@<B<DA<?;
```

NA12878.exon.sample.unpaired.fastq.gz

Single-ended reads (mostly from ealirer platforms)

```
workshop:~/wed> zcat input/NA12878.exon.sample.unpaired.fastq.gz | head -4
@SRR013620.70875
CACACACATGCATGTGCAAGCAGCTGGCCCGGACAGAGAGATCCCCCTCCG
+SRR013620.70875
3:>;B=D%6GA=CH=CC=@@BA>3@B=:A?6;9<;8%6$423%*.5),3.%
```

Step 2 : Mapping FASTQ files to the reference genome

3 step procedure for mapping

From `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/README.alignment_data`

1. Create bwa index of reference (just once) - already done
   - `bwa index -a bwtsw [ref.fa]`
2. For each fastQ file run `bwa aln` command
   - `bwa aln -q 15 [ref.fa] [fastq] > [sai]`
3. Create SAM/BAM files for single-end or paired-end reads
   - `bwa sampe [ref.fa] [sai.1] [sai1.2] [fastq.1] [fastq.2] > [bam]`
   - `bwa samse [ref.fa] [sai] [fastq] > [bam]`

Step 2.1 - bwa aln

```
${BIN}/bwa aln -q 15 ${REF}/human_g1k_v37_chr20.fa \
   ${IN}/NA12878.exon.sample.read1.fastq.gz \
  > ${OUT}/NA12878.exon.sample.read1.fastq.gz.sai

${BIN}/bwa aln -q 15 ${REF}/human_g1k_v37_chr20.fa \
   ${IN}/NA12878.exon.sample.read2.fastq.gz \
   > ${OUT}/NA12878.exon.sample.read2.fastq.gz.sai

${BIN}/bwa aln -q 15 ${REF}/human_g1k_v37_chr20.fa \
   ${IN}/NA12878.exon.sample.unpaired.fastq.gz \
   > ${OUT}/NA12878.exon.sample.unpaired.fastq.gz.sai
```

Step 2.2 - bwa samse or bwa sampe

```
${BIN}/bwa samse ${REF}/human_g1k_v37_chr20.fa \
  ${OUT}/NA12878.exon.sample.unpaired.fastq.gz.sai \
  ${IN}/NA12878.exon.sample.unpaired.fastq.gz \
  | ${BIN}/samtools-hybrid view -uhS - \
  | ${BIN}/samtools-hybrid sort -m 10000000 \
  - ${OUT}/NA12878.exon.sample.unpaired.bwa.sorted

${BIN}/bwa sampe ${REF}/human_g1k_v37_chr20.fa \
  ${OUT}//NA12878.exon.sample.read1.fastq.gz.sai \
  ${OUT}/NA12878.exon.sample.read2.fastq.gz.sai \
  ${IN}/NA12878.exon.sample.read1.fastq.gz \
  ${IN}/NA12878.exon.sample.read2.fastq.gz \
  | ${BIN}/samtools-hybrid view -uhS - \
  | ${BIN}/samtools-hybrid sort -m 10000000 \
  - ${OUT}/NA12878.exon.sample.paired.bwa.sorted
```

Step 3 - Merge mulltiple BAMs

```
${BIN}/samtools-hybrid merge ${OUT}/NA12878.exon.sample.merged.bam \
   ${OUT}/NA12878.exon.sample.paired.bwa.sorted.bam  \
   ${OUT}/NA12878.exon.sample.unpaired.bwa.sorted.bam
```

Step 4 - View the SAM/BAM format

```
${BIN}/samtools-hybrid view -h ${OUT}/NA12878.exon.sample.merged.bam | less
@SQ SN:20 LN:63025520
SRR014820.15718550 117 20 2654979 0 * = 2654979 0
  AAAGTAGAATCCTCTGAGTGCCTAGCAATATGGAAA
  @%C@;C>50*,%+@;?B-:*<@7?:@@A6.@=8:?7
SRR014820.15718550 153 20 2654979 0 11M1D25M = 2654979 0
  AGCACCTTGGGGGCCGAGGCAGGTGGTTCACCTGAG
  %;>962707<$A=>-=8;<=6<<?@>%.@%;>>;B8
  XT:A:R NM:i:2 SM:i:0 AM:i:0 X0:i:2 X1:i:0 XM:i:1 XO:i:1 XG:i:1 MD:Z:5T5^A25
```

Step 5 - Mark Duplicated Reads

```
${BIN}/superDeDuper -i ${OUT}/NA12878.exon.sample.merged.bam \
  -o ${OUT}/NA12878.exon.sample.deduped.bam -v
```

Step 6 - Visualize alignment to reference genome

Run samtools tview

```
${BIN}/samtools-hybrid index ${OUT}/NA12878.exon.sample.deduped.bam
${BIN}/samtools-hybrid tview ${OUT}/NA12878.exon.sample.deduped.bam \
   ${REF}/human_g1k_v37_chr20.fa
```

- Type 'g', and 20:19989392
- TYPE 'g', and 20:20032998

What we covered today

1. Understand sequence reads format (FASTQ)
2. Map sequence reads to reference genome
3. Merge multiple alignment files
4. View the sequence alignment format (SAM/BAM)
5. Mark duplicated reads
6. Visualize alignment to reference genome