Mapping DNA Reads with BWA

Author: Brian Nadel (brian.nadel@gmail.com)

July 7th, 2016

BWA (Burrows Wheeler Aligner) is memory-efficient mapping algorithm. The full man page containing all commands and parameter choices can be found here: http://bio-bwa.sourceforge.net/bwa.shtml. Below is an outline of a typical pipeline. If using hoffman2 importing bwa is simple:

```
module load bwa
```

BWA requires a reference genome, which must be indexed prior to alignment. This is done with the command:

```
bwa index ref.fa
```

This creates several index files in the same directory. The hg19 assembly of the human genome is usually appropriate when working with human data, though there are other choices as well. The chromosome files for hg19 can be found at this URL: http://hgdownload.cse.ucsc.edu/goldenpath/hg19/chromosomes/.

After indexing, use the "bwa mem" command to align the reads. A typical command for non-paired reads might be:

```
bwa mem -M -R "TAG"  hg19.fa   reads.fa > mappedReads.sam
```

The -R "TAG" refers to the read group tag, which can be important for downstream steps. The sam file specifications describes the fields for the read group tag (https://samtools.github.io/hts-specs/SAMv1.pdf).

The -M argument is necessary if you plan to use Picard and/or GATK tools downstream, otherwise it can be omitted.

If mapping paired end reads, there are two possible formats. The first is if the reads are in two separate files (e.g. the first read in File1 is the pair of the first read in File2, and so on). In this case, the command looks like this:

```
bwa mem -M -R "TAG"  hg19.fa   reads1.fa reads2.fa > mappedReads.sam
```

If the paired reads are interleaved in the same file, use the -P command:

```
bwa mem -M -P -R "TAG"  hg19.fa   reads.fa > mappedReads.sam
```

After mapping, you will likely want to convert to bam format. This can be done with samtools:

```
module load samtools (if on hoffman2)
samtools view -Sb MappedReads.sam > MappedReads.bam
```

The -S denotes that the input is sam format, and the -b denotes that the output is bam format.

The bam file can also be sorted, which further shrinks the file size:

samtools sort MappedReads.bam MappedReads.sorted.bam

and indexed, as required for many downstream steps

samtools index MappedReads.sorted.bam

Bwa mem with default parameters will output at least one line for each of the input reads. Some reads will be mapped more than once if they have two or more distinct regions that map to different places, i.e. due to introns or chimeric reads. Note, this is not multi-mappers - reads that map equally well to more than once reference genome position. Multi-mappers are mapped only once, randomly at one of the positions and are given a mapping quality of zero. Bam files can be converted back to fastq files, so once you've confirmed all reads are in the output, you can delete the original fastq files.

The sam flag field gives information about each reads alignment (or lack thereof). See bwa documentation for a basic description. A summary of the sam flag entries is a good standared QC step for your alignments.

samtools flagstat MappedReads.sorted.bam

Example samtools flagstat output. QC-failed reads refers to the Illumina pass/fail field. In this example failing reads and duplicated reads have already been removed.

406826 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
399433 + 0 mapped (98.18%:nan%)
406826 + 0 paired in sequencing
203958 + 0 read1
202868 + 0 read2
349205 + 0 properly paired (85.84%:nan%)
390689 + 0 with itself and mate mapped
8744 + 0 singletons (2.15%:nan%)
42009 + 0 with mate mapped to a different chr
15292 + 0 with mate mapped to a different chr (mapQ>=5)