

# **BISC-577: Homework # 1**

Due on Tuesday, March 31, 2015

**Saket Choudhary**  
**2170058637**

## Contents

Question # 1 . . . . .	3
Question # 2 . . . . .	3
Question # 3 . . . . .	3
Question # 4 . . . . .	4
Question # 5 . . . . .	4
Question # 6 . . . . .	4
Question # 7 . . . . .	4

**Question # 1**

**Uniquely mappable reads:** Reads that map to a unique position in the reference genome. These reads cannot come from repeated regions of DNA.

**Ambiguously mapping reads:** A read may often map to more than one positions in the reference genome. This is often true for reads coming from the region of say short tandem repeats. It is also possible to get more than one mapping positions for a read if mismatches are allowed. With increasing number of allowed mismatches, the number of positions that read gets mapped to also increases.

**PCR duplicate reads:** Upon ligation with adapters, the fragments are PCR amplified so that they are enough to be detected by the flow channel. Multiple PCR copies of the same fragment if sequenced in two different wells.

**Concordantly mapped paired-end reads:** In a paired end/ mate pair experiment, a fragment is sequenced from both the ends. Thus while mapping, there is an 'expectation' that these 'mates' 1 and 2 will have certain orientation. These mates are expected to be separated by an 'insert size'. However it is possible that the sequenced read comes from say a structural variation, in which the sequence is likely to map in a flipped manner, resulting in discordant mapping.

**Sequenced fragment length:** The sequenced fragment length refers to the 'piece' of chunked sequence that is sequenced at single or both ends post ligation and PCR amplification

**Uniquely mappable part of a genome:** Genome sans the repeats(Tandem repeats, interspersed repeats)

**Question # 2**

Single End: <http://www.ncbi.nlm.nih.gov/sra/SRX175791>[accn]

SRA size: 2.7G

FastQ size : 20G

Paired End: <http://www.ncbi.nlm.nih.gov/sra/SRX109558>[accn]

SRA size: 9.7G

FastQ size : 25G + 25G = 50G

**Question # 3**

Organism: Homo Sapiens

The reference was downloaded as a 2 *bit* file from UCSC

Build: hg19

Reference came as a single 2 *bit* file and was converted to FASTA using 'twoBitToFa' utility available on UCSC.

Besides the 22 autosomes and the two sex chromosomes, the FASTA contains few scaffolds for some chromosomes and the mitochondrial sequence.

Given that these datasets come from a WGS study, it would make sense to include all sequences(including scaffolds, mitochondrial) for mapping. The overhead of having extra sequences in the reference is going to result in increased time required for searching.

**Question # 4**

```
time bowtie2-build -f hg19.fa hg19 1&&bowtie2build.log 2&&bowtie2build.err
real 96m38.714s user 95m56.664s sys 0m28.580s
time bwa index hg19.fa 1&&bwa.log 2&&bwa.err
real 60m55.489s user 59m17.654s sys 0m20.558s
```

**Question # 5**

Mapping results are presented in SAM format. SAM stands for Sequence Alignment/Map and is a generic format for storing alignments.

SAM contains reference sequence name, the leftmost positions where the alignment starts, the query sequence(read sequence), it's quality sequence. Match, mismatch information is encoded in CIGAR format. CIGAR is a space efficient way to store matches, mismatches.

*bwa* does not print out the number of concordant/discordant reads that were mapped explicitly, *bowtie2* does. *bwa* does not explicitly print out number of reads that are ambiguously mapped. Both print the total number of reads.

**Paired End:**

**bwa:** 107m48s for 79367217 reads

**bowtie2:** 69m19s for 79367217 reads

**Single End**

**bwa:** 37m54s for 86574968 reads

**bowtie2:** 16m37s for 86574968 reads

Memory requirement was bounded by 16GB for both *bwa* and *bowtie2*.

**Question # 6****Question # 7**