BISC-577: Project # 2

Due on Tuesday, April 7, 2015

Saket Choudhary
2170058637

Contents

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 sequnced 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 camse as a single 2 *bit* file and was converted to FASTA using 'twoBitToFa' utility available on UCSC.
Besides the 22 automosomes and the two sex chromosomes, the FASTA contains few scaffolds for some chromosomes and the mitochondrial sequence. In total there are 93 chromosomes with total base size 3,137,161,264.
Given that these datasets come from a WGS study, it would make sense to include all sequences(including scaffolds, mitrochondrial) for mapping. The overhead of having extra sequences in the reference is going to result in incresased time required for searching.

continued on next page. . .

Question # 4

Building index:
bowtie2: 96m38.714s

bwa 60m55.489s

bwa's suffix-array and bwt files are 1.5G and 3G respectively. bowtie2 createx indexes[for hg19 4 forward, 2 reverse] in mutiple files totaling 3.5G as well.

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 aligment 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, *bowtie*2 does. *bwa* does not explicitly print out number of reads that are ambigulosuly 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, not measured explicitly.

continued on next page. . .

Question # 6

SAM size
Single End:
bwa: 18095631253
bowtie2: 18234575912

Paired End
bwa: 53215198684
bowtie2: 52379271708

To determine the distribution of length of sequenced fragments, one can extract $ISIZE$ from the samfile. This will give the insert size. $FragmentSize = ISIZE + 2 * ReadLength$ for paired end. Alternatiely using the location of mapping of mate pairs. The sequence fragment would be $Mate2LeftMostPosition - Mate1RightMostPosition + 2 * ReadLength$ Sequence fragment length is determined by the library preparation strategy(restriction enzymes used for example).

Single End
Bowtie:
Reads mapped with 0 mismatches('XM:0'): 74834880
Number of mismatches(Extracted from 'XM' tag of mapped reads in SAM file): 16218708
Total Mapped Reads: 85083473
Average number of mismatches: $\frac{16218708}{85083473} = 0.19$ per mapped read = 19 mismatched bases per 100 bases mapped

Fraction of reads uniquely mapped: $\frac{83670496}{85083473} = 98.3\%$

BWA:
Reads mapped with 0 mismatches('XM:0'): 74751241
Number of mismatches(Extracted from 'XM' tag of mapped reads in SAM file): 12686684
Total number of reads that map: 84561944
Average number of mismatches: $\frac{12686684}{84561944} = 0.15$ per mapped read = 15 mismatched bases per 100 bases mapped

Fraction of reads uniquely mapped: $\frac{74588324}{84561944} = 88.2\%$

Paired End
Bowtie:

Reads mapped with 0 mismatches('XM:0'): 104436920
Number of mismatches(Extracted from 'XM' tag of mapped reads in SAM file): 11814602
Average number of mismatches: $\frac{11814602}{72991407} = 0.16 = 16$ mismatches per 100 bases mapped
Total Mapped reads: 153358953
Fraction of reads uniquely mapped: $\frac{151486001}{153358953} = 98.7\%$
Fraction of reads mapped concordantly: $\frac{151210022}{153358953} = 98.5\%$

[Question # 6] continued on next page...

Illumina's sequencing error rate = 1 in 1000
Thus comparing the average number of mismatched base pairs (around 15 per 100 or 150 per 1000) they are much higher than the sequencing error rate. Thus, a higher error rate at the experiemnt end is well adjusted for by controlling the allowed number of maximum mismatches.(This was left to default)

Question # 7

BAM size:

Single End
bwa: 4836918016 v/s 18095631253[SAM]
bowtie2: 4866929520 v/s 18234575912[SAM]

Paired End
bwa: 16107276016 v/s 53215198684[SAM]
bowtie2: 16250076788 v/s 52379271708[SAM]

samtools recognizes multiple reads that map to the same coordinates as potential PCR duplicates. However if the mate pairs are mapped discordantly(for paired end experiments) and say mate pair 1 shares the same coordinate with some other read, they are NOT regarded as duplicates. single-end duplication is treated as defined earlier. It is easier to mark reads as duplicate for paired end experiments since the evidence is higher.
For paired end:

bwa
Total mapped reads before removing duplicates: 152113718
Total mapped reads after removing duplicates: 149227033
Mapped reads removed:2886685 = 1.89%

For single end:
Bowtie
Total mapped reads before removing duplicates: 85083473
Total mapped reads after removing duplicates: 79941660
Mapped reads removed: 5141813 = 6%

Bwa
Total mapped reads before removing duplicates: 84561944
Total mapped reads after removing duplicates: 63365044
Mapped reads removed: 21196900 = 25%

Around 6% mapped reads are removed for the single end experiment when mapped with bowtie and 25% for bwa, which is strange. One possible reason is that bowtie allows lmore mismatches than bwa(19 v/s 14 per 100 bp mapped as in Q6 above) thus allowing reads to map at different locations while bwa is more stringent(using default paramters). The experiment did not inolve significant level of duplication though.

continued on next page...