Assignment 1 Sample Solution

Task 1
See the submission by Sequn Ma on the LMS.

Task 2
The complexity of the aligner is O(n*l^2*g + g). The index has to be read, which is in order of the genome size (+g). Each read (n) has to be aligned. Each k-mer in the read (l) needs to be looked up in the index. In theory, a k-mer could point to every position in the genome (g)!! Each of those positions needs to be compared to the read (l).
This makes the theoretical complexity worse than that of the workshop method (O(ngl)).
Measuring the runtime for each for the references (Figure 1), we observe the opposite: The indexed aligner is several orders of magnitude faster. Further, the runtime is constant with the genome size, instead of scaling linearly (there is a slight trend, which could just be due to the +g component or variance).
The reason for this is that the g in the first term is only theoretical – in practice the number of positions to be visited for each k-mer is very limited (mostly unique). Also the l^2 is only theoretical as many of the position lists of the O(l) k-mers in the read lead to already explored alignments in the reference and don't have to be checked again.
The independence of g might change with smaller values of k.

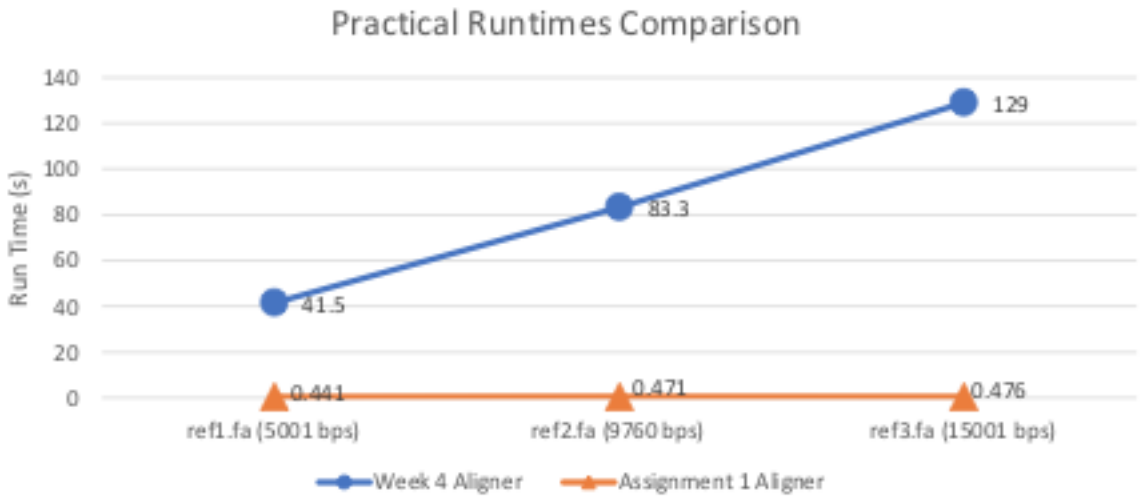

Figure 1: Runtime comparison of two alignment programs. The (rounded) runtimes in second is shown next to each data point.

Task 3
Observe the summary of Phred base qualities of reads in Figure 3. Quality scores (QS) go down towards the ends of reads for both distributions (phasing becoming an issue).
Mismatches' QSs are overall lower than the average base.
Mismatches are also wider in range (box size in Figure 2b), showing greater variance of QS values. This could be because they are comprised of errors and SNPs. Sequencing errors should

be generally low in QS if the QS is indeed correlated with sequencing quality. SNPs however, result in mismatched bases with perfectly normal quality.
Tendency of errors to have lower mismatched quality could be factored in during alignment (maybe having Hamming distance scaled by QS – a low quality mismatch not counting as much as a high quality one?), but the differences are subtle, so this may not do much.

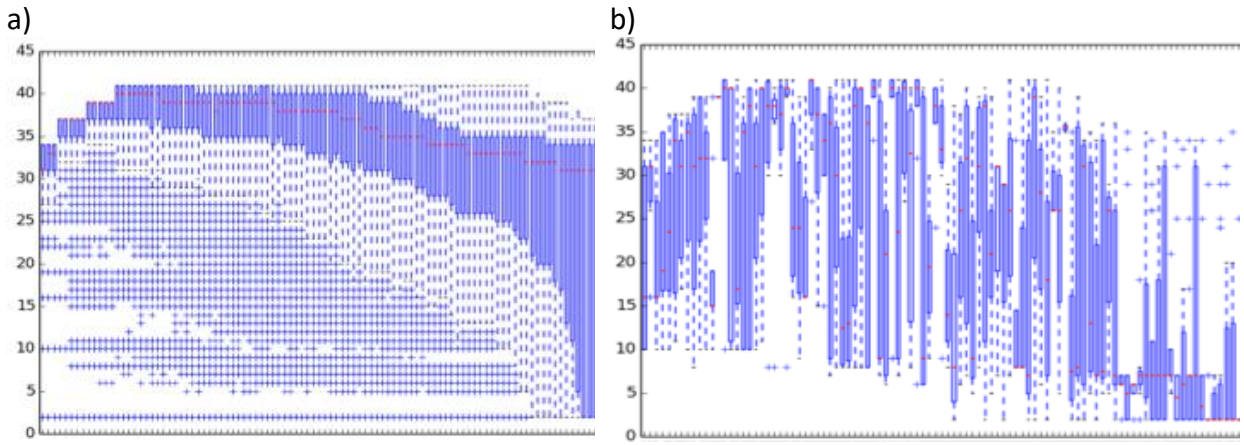

Figure 2: Boxplot of quality scores of (a) all read bases, and (b) mismatched bases.