22126 NGS analysis recap test

Q1. What are the commonalities and differences between how Sanger sequencing operates and the second generation of next-generation sequencing machines (Illumina, Ion Torrent)?

Q2. What are the primary differences between how the second generation of next-generation sequencing machines (Illumina, Ion Torrent) operate and third-generation NGS platforms?

Q3. What is the main type of sequencing error seen with Illumina data? Why?

Q4. After 10 cycles (a cycle of 1 type of dNTP are bound, pH measure, unattached dNTP molecules are washed out) of Ion torrent, for which kind of sequence are you guaranteed to have exactly resolved 10 bases for a read?

Q5. How many lines is one read in fastq format? What are the lines?

Q6. What does it mean that a base in a read has a base quality of Q20?

Q7. A sequence has a length of 200 bases. An Illumina sequencer is used with 75 cycles. How many bases will have been **unsequenced** if used in single-end mode? In paired-end mode?

Q8. A sequence has a length of 100 bases. An Illumina sequencer is used with 75 cycles. How many bases will have been **sequenced twice i**f used in single-end mode? In paired-end mode?

Q9. We have sequenced a genome to 50X. Does this mean:

a) We ran the sequencer 50 times
b) Every base in the genome is covered by at least 50 reads
c) Every base in the genome is covered by at most 50 reads
d) Every base in the genome is covered on average 50 reads
e) None of the above

Q10. Briefly describe the principle of the Seed and Extend algorithm.

Q11. Why are longer reads better for aligning or assembly?

Q12. Create the Burrows-Wheeler Transformation of this sequence “TAGC”.

Q13. Create the de Bruijn graph of this sequence using k=3: ACGTTGGTCGTG

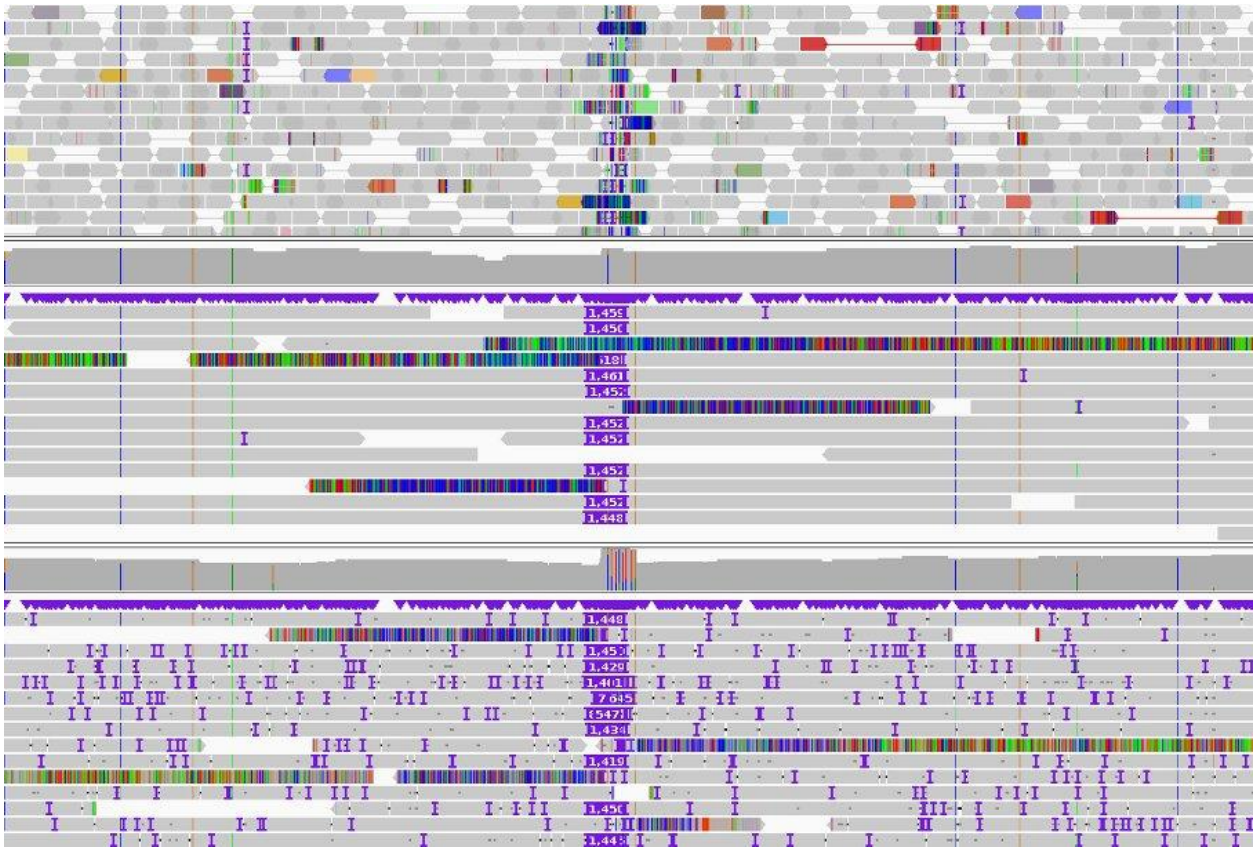
Q14. How do we create contigs and scaffolds from a de Bruijn graph?

Q15. Why is *de novo* assembly much harder for metagenomic data compared to single genome data?

Q16. How would you analyze if your metagenomic sequencing has sampled enough? Both for 16s rRNA amplicon data and for shotgun-metagenomic sequencing.

Q17. Someone says: “We have observed a read aligning at that position and there was an ‘A’, therefore, the sample is homozygous A”. I can think of half a dozen reasons why this is not true. List 3.

Q18. Someone sequenced data from Illumina, Oxford Nanopore and PacBio. This is a screenshot of the alignments in IGV. Guess which is which:



(source: Alex Rubinsteyn, UNC Chapel Hill
https://twitter.com/iskander/status/1603094882708377601)

Q19. The initial human genome was released around 2000 and was for a single genome and had several gaps. Which technology allowed us to sequence 1000 Genomes around 2010 but did not close the gaps in the reference? Why weren’t able to close the gaps then but did so in 2020? Which technology allowed us to do this?

Q20. Associate the following quality control measures to what they are applied and what computes them:

quality score	individual read	software that finds unique regions in the genome
mappability	reference genome	mapping software
mapping quality	individual DNA base	basecaller