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)?

A. Commonalities: DNA shearing and amplification, differences, the massively parallel nature

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?

A. 3rd generation generates longer reads, poorer quality in general and does not use amplification

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

A. Mismatches, it calls one base at a time

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?

If you define a cycle as any time dNTPs are introduced then you need ACGTACGTAC if bases are added in the A,C,G,T order. Otherwise, if you define a cycle as only when molecules are bound then it will occur if a sequence has no duplicated consecutive bases e.g. TAAC, A is found twice.

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

4 lines header, sequencing, plus sign and quality

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

20 means an error rate of 1/10^2.0 = 1/100. This can be considered not great but not terrible.

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?

A. Single-end 200-75=125 bases paired: 200-75*2 = 50 bases

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?

A. Single-end 0 bases, paired-end will meet for 50 bases in the middle

Q9. What does it mean to have sequenced a genome to 50X?

A. That every base **on average** will have been covered 50 times

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

A. To avoid scanning the entire genome, we seed the alignment with highly similar words between the query and genome and then extend the alignment by using a more sophisticated aligner around these seeds.

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

A. Alignments are less ambiguous. If they span a repeat, the read might be larger than the repeat regions. For de novo assembly, longer reads will result in more paths in the de Bruijn graph and create longer contigs.

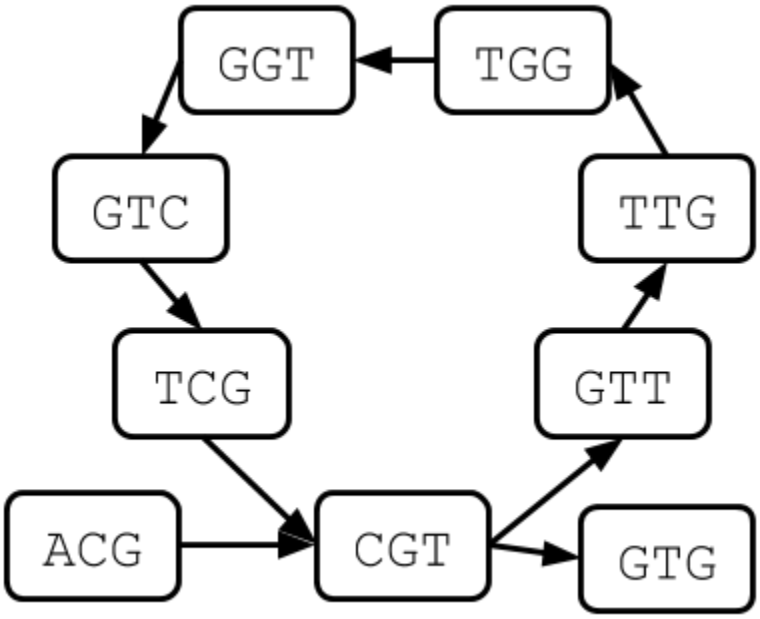
Q12. Create the Burrows-Wheeler Transformation of this sequence “TAGC”.

Shift and the $:
TAGC$
AGC$T
GC$TA
C$TAG
$TAGC

sort:
$TAGC
AGC$T
C$TAG
GC$TA
TAGC$

The BWT is CTGA$

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?

A. Find a Eulerian path through the graph, form contigs from this path, merge contigs into scaffolds using mate-pairs

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

A. Kmers from different species will have edges in common due to similar kmers

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

For 16s rRNA amplicon sequencing rarefaction curves are used to estimate the fraction of the microbiome that the sample covers. A rarefaction curve measures the number of unique Operational Taxonomic Units (OTUs) as a function of the number of reads/sequences. The shape of the rarefaction curve will reveal how many new OTUs are added to the sample, when including more reads.

For shotgun-sequenced metagenomes, nonpareil curves are used. Nonpareil uses the redundancy of the reads in metagenomic datasets to estimate the average coverage and predict the number of sequences that will be required to achieve “nearly complete coverage”.

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.

A. Contamination
B. Mismapping
C. The alignment is not the correct one (problems around indels or homopolymers)
D. Sequencing error
E. Heterozygosity
F. Copy number variation
G. Somatic mutations

Q18. Top: Illumina, Middle: PacBio, bottom: Oxford Nanopore

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?

Second-generation sequencing technology (Illumina mostly)

Why weren’t able to close the gaps then but did so in 2020? Which technology allowed us to do this?

Illumina produced more data than Sanger but the read length was shorter and therefore did not allow us to span highly repetitive regions. This was accomplished using 3rd generation sequencing technology (PacBio mostly).

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