Final Study Guide: DNA Sequencing and Beyond

Part 1: The Big Picture - Why This Matters

- **The Goal:** The ultimate goal of DNA sequencing is to determine the exact order of the A's, C's, G's, and T's in an organism's genome.
- **The Human Genome Project (HGP):** This was a massive, international "race" to sequence the first human genome. It was a landmark achievement that proved large-scale sequencing was possible and kickstarted a technological revolution.
  - **Key Outcome 1 (Gene Count):** The number of human genes (~20,000) was found to be much lower than the initial estimates (>100,000).
  - **Key Outcome 2 (Cost):** The cost of sequencing a genome has dropped exponentially, even faster than Moore's Law, making it a routine tool in modern biology and medicine.

Part 2: The Core Technology - How Second Generation Sequencing Works

This is the central process you need to understand. The modern method is called **Sequencing by Synthesis (SBS)**.

The Analogy: We figure out the sequence of a hidden, original DNA strand by **building a new, complementary copy** of it one piece at a time and taking a picture after adding each piece.

The Step-by-Step Process:

1. **Preparation ("Library Prep"):**
   - The long DNA from an organism is physically shattered into millions of short, random, overlapping fragments.
   - These fragments are attached to a glass slide called a **flow cell**.
   - Each fragment is copied thousands of times in its spot, creating a dense **cluster** of identical molecules. This makes the signal strong enough to be detected.
2. **The Sequencing Cycle (Repeated ~150 times):**
   This is a four-step chemical process that happens simultaneously for all clusters on the slide.
   - **Step A: Incorporation.**
     - **What happens:** The machine floods the slide with four types of special nucleotides (A, C, G, T) and the enzyme **DNA Polymerase**.
     - **Key Features of the Nucleotides:**
       1. Each type has a unique **fluorescent color** (e.g., A is blue, T is green).
       2. Each has a **reversible terminator** (a chemical "cap") that allows **only one** base to be added at a time.
     - **The Chemistry:** The DNA Polymerase adds the single, correct complementary base to the growing DNA strand in each cluster. For a template 'G', a 'C' will bind; for a template 'A', a 'T' will bind.

    - **Step B: Imaging.**
        - **What happens:** A laser illuminates the slide, and the newly added nucleotides fluoresce (glow). A high-resolution camera takes a picture of the entire slide.
        - **The Result:** The computer has an image showing the color of the *first* base for all billion clusters.
    - **Step C: Cleavage.**
        - **What happens:** A chemical wash is applied that does two things: 1) it removes the fluorescent dye, and 2) it removes the terminator "cap."
        - **The Result:** The new DNA strands are now ready for the next base to be added.
    - **Step D: Repeat.**
        - The entire cycle is repeated. In the second cycle, the camera captures the color of the *second* base for all clusters, and so on.
3. **The Result (The "Reads"):** After 150 cycles, the computer has 150 images. For each cluster, it traces its color sequence through all the images and translates it into a 150-base-pair DNA sequence. This sequence is called a **read**.

Part 3: The Data - Understanding the FASTQ File

The final output of a sequencing run is a text file called a **FASTQ file**, which contains the reads and their quality scores. Each read is represented by **four lines**.

- **Line 1: Header:** Starts with @. A unique ID for the read.
- **Line 2: Sequence:** The string of A, C, G, T's (the read itself).
- **Line 3: Separator:** Always starts with +.
- **Line 4: Quality String:** A string of characters that encodes the quality of each base in the sequence.

The Math: Phred Quality Scores (Q)

- **What it is:** A numerical score representing the confidence in a base call. **Higher Q = Better Quality.**
- **The Formula:** Q = -10 * log10(p), where p is the probability that the base call is incorrect.
    - Q = 10 means p = 1/10 (1 in 10 chance of error).
    - Q = 20 means p = 1/100 (1 in 100 chance of error).
    - Q = 30 means p = 1/1000 (1 in 1000 chance of error).
- **Encoding (Phred+33):** The Q score is converted to a single text character for the quality string using the formula: character = chr(Q + 33). To decode, you do the reverse: Q = ord(character) - 33.

Edge Cases, Key Jargon, and Potential Exam Questions

- **Q: What is the main cause of errors in Second Generation Sequencing?**
    - **A: Phasing/Dephasing.** This is when some of the DNA strands within a single cluster get out of sync with the main reaction (either falling behind or jumping ahead). This causes the camera to see a "muddy" or mixed color signal, leading to a lower quality score and potential mis-calling of the base.
- **Q: What is the purpose of the "terminator" in Sequencing by Synthesis?**
    - **A:** The terminator is a chemical cap that **allows only one nucleotide to be added per cycle**. This is the key to keeping all the billions of reactions on the slide synchronized. Without it, the polymerase would add multiple bases at once, and we couldn't read the sequence one base at a time.
- **Q: Why was the Human Genome Project initially proposed by the Department of Energy (DOE) and not the NIH?**
    - **A:** This is a classic trivia/history question. The DOE was interested in understanding the effects of radiation on DNA, which required mapping and sequencing technology. This unique historical starting point gave them the initial impetus to launch the project.
- **Q: The quality string for a read is #!J. What are the Q scores for these three bases?**
    - **A:** Q=0, Q=2, Q=41.
    - **Calculation:**
        - ord('!') - 33 = 33 - 33 = 0
        - ord('#') - 33 = 35 - 33 = 2
        - ord('J') - 33 = 74 - 33 = 41
- **Q: Why is the method called "massively parallel"?**
    - **A:** Because billions of different DNA fragments (the clusters) are all sequenced **simultaneously** on a single slide. The camera captures the state of all of them in a single picture for each cycle.

Solutions & Explanations: DNA Sequencing and Beyond

Question 1: You are given a data file with many low-quality reads (mean Q < 15). Describe how you would preprocess this data before assembly or variant calling.

This is a core task in bioinformatics called **Quality Control (QC)**. You should never trust raw sequencing data. The goal is to clean out the "noise" so that your downstream analysis (like assembly) is more accurate.

The Preprocessing Pipeline:

1. **Assess Quality (e.g., using FastQC):** The first step is to understand the quality of the data. A tool like FastQC will generate a report showing the distribution of quality scores at each position across all reads, the presence of adapter sequences, and other potential issues. This confirms that the reads are, in fact, low quality.
2. **Adapter Trimming:** The sequencing process involves ligating (attaching) short, known DNA sequences called "adapters" to the ends of your fragments. Sometimes, if the DNA fragment is short, the machine will sequence through the fragment and into the adapter on the other side. This adapter sequence is not part of the original genome and must be removed. Specialized tools are used to find and trim these adapter sequences from the ends of reads.
3. **Quality Trimming:** This is the most important step for low-quality data.
   - **Sliding Window Trim:** The program slides a small window (e.g., 4-5 bases) along the read from the 5' end to the 3' end. It calculates the average Q score within the window. Once the average score drops below a set threshold (e.g., Q15 or Q20), the program trims off that base and all subsequent bases to the end of the read.
   - **Leading/Trailing Trim:** Many tools also specifically trim low-quality bases from the absolute beginning and end of the read.
4. **Length Filtering:** After trimming, some reads may become very short (e.g., less than 30-50 bases). These short reads are often not specific enough to be useful and can cause problems in assembly. A final filtering step removes any reads that are now shorter than a specified length.

Summary: The process is to **assess** the quality, then **trim** away adapters and low-quality bases, and finally **filter** out reads that have become too short.

Question 2: What is the purpose of the Phred+33 encoding scheme in FASTQ files?

- **The Purpose:** To **compactly and unambiguously store a numerical quality score (an integer) as a single text character.**
- **The Explanation:** A FASTQ file can contain billions of reads. Storing a quality score like "30" would take two characters, while "9" would take one. This would make the files larger and more complex to parse. The solution is to map every possible Q score to a unique, single character. The universal standard for this is the ASCII character set. The "+33" is added because the first 33 characters in the ASCII table are non-printable control characters (like "tab" or "backspace"). By adding 33, we shift the entire scale so that even the lowest quality score (Q=0) is represented by a printable character (!).

Question 3: If the quality score of a base is 30, what is the estimated probability that the base is incorrect?

- **The Answer:** A 1 in 1,000 chance of error, which is a probability of **0.001** (or 0.1%).
- **The Math:** The formula is Q = -10 * log10(p), where p is the error probability.
  - 30 = -10 * log10(p)
  - -3 = log10(p)
  - p = 10^-3 = 1/1000 = 0.001

Question 4: Why do most sequencing reads have lower base quality toward the end?

- **The Answer:** This is caused by the accumulation of chemical errors in the sequencing-by-synthesis process, a phenomenon known as **phasing and dephasing**.
- **The Explanation:** A "cluster" on the flow cell contains millions of identical DNA strands being copied in sync. In each cycle, a tiny fraction of these strands can fail:
  - **Falling Behind (Phasing):** A terminator might fail to be removed, so that strand gets "stuck" for a cycle.
  - **Jumping Ahead (Dephasing):** A terminator might not be attached, allowing two bases to be added at once.

    These errors are **cumulative**. By cycle 1, nearly 100% of strands in the cluster are in sync. By cycle 150, after 149 opportunities for error, a significant fraction of the strands may be out of sync. The camera then sees a "muddy" signal with a mixture of colors instead of a pure one. The machine's software interprets this noisy signal as low confidence, assigning a lower Phred score.

Question 5: Why is it necessary to check the reverse complement of a read during genome mapping?

- **The Answer:** Because DNA is double-stranded, and we have no way of knowing which of the two original strands a given read came from.
- **The Explanation:** The reference genome we use for mapping is typically just one of the two strands (the "forward" or "Watson" strand). However, the DNA fragmentation process is random. A read in our data file could have originated from the forward strand, or it could have originated from the reverse strand. Therefore, to find the read's correct location, a mapping algorithm must try two alignments:
  1. Align the read's sequence as-is.
  2. Align the read's **reverse complement**.

     One of these two possibilities corresponds to the read’s true origin.

Question 6: A 100 bp read has high quality in the first 70 bases (Phred > 30) and drops to Phred < 15 in the last 30 bases. Would you prefer to align the full read or a trimmed read to the genome? Justify.

- **The Preference:** In virtually all standard bioinformatics pipelines, you would prefer to **align the trimmed read (the first 70 bases).**
- **Justification:** This is a classic trade-off between **signal** and **noise**.
  - The first 70 bases represent a strong, high-confidence signal.
  - The last 30 bases are noisy and have a high probability of containing errors (a Q score of 15 means a ~3% error rate, which is very high).
  - **The Risk of Aligning the Full Read:** These errors in the low-quality tail can introduce multiple mismatches into the alignment. This might cause the alignment algorithm to fail to find a match altogether, or to assign such a low alignment score that the read is discarded. Worse, it could cause the read to incorrectly map to the wrong location in the genome.
  - **The Benefit of Aligning the Trimmed Read:** By trimming off the noisy tail, you provide the aligner with a high-quality, high-confidence sequence. This sequence is much more likely to map uniquely and correctly to its true origin. While you lose some specificity by using a shorter read, this is a much smaller risk than the high probability of alignment failure or mis-mapping caused by the errors. **Accuracy is more important than length.**

–Fahad Nadim Ziad, 24341216