PCR

- Generate billions of identical DNA copies from a single molecule
- Essential for: DNA sequencing, RNA sequencing (via cDNA), Genomics and molecular biology research
- DNA structure: double-stranded, complementary base-pairing (A-T, C-G), Directional: 5' → 3'
- DNA polymerase: enzyme that replicates DNA starting from a primer
- Primers: short, single-stranded DNA sequences (~15–20 bases) that are complementary to target sequence ends and bind (anneal) to specific locations to initiate replication
- PCR Ingredients:
    - Template DNA: The sequence to be copied
    - Primers: Two (forward and reverse) to flank the target region
    - DNA polymerase: Heat-stable (e.g., Taq polymerase)
    - dNTPs: Free A, T, C, G nucleotides
    - Buffer: Maintains optimal conditions
- PCR Process: 3 Main Steps per Cycle
    - Denaturation (94°C): Heat breaks hydrogen bonds between strands → single-stranded DNA
    - Annealing (54°C): Primers bind (anneal) to their complementary sequences
    - Extension (72°C): DNA polymerase extends from the primers, synthesizing new DNA strands
- Amplification
    - Each cycle doubles the DNA quantity:
    - Cycle 1: 1 → 2 copies
    - Cycle 2: 2 → 4 copies
    - After 30 cycles: ~2^{30} = ~1–2 billion copies
- All ingredients are mixed in one tube
- Thermocycler machine automates temperature changes
- Fast, reliable, and requires minimal manual intervention

Next Generation Sequencing (NGS, second-generation sequencing):

- Development
    - Introduced around 2007; still the dominant DNA sequencing method today.
    - Successor to first-generation Sanger sequencing (developed by Fred Sanger).
    - Sanger sequencing: slow, manual → later automated in the 1980s.
    - DNA microarrays (1990s) enabled hybridization-based analysis but were not sequencing.
- Basic Principle

    - Sequencing involves copying DNA using DNA polymerase and tracking the added bases.
    - Uses complementary base pairing: A-T, C-G.
    - Observing the copying process allows us to “read” the sequence.
- Process
    - DNA is fragmented into short pieces (hundreds to ~1,000 bases).
    - Fragments are chemically attached to a solid slide surface.
    - PCR is used on the slide to amplify fragments into clusters of identical copies.
    - Each cluster = millions of identical DNA fragments at a fixed position on the slide.
- Sequencing by Synthesis
    - Special fluorescently labeled nucleotides (A, T, C, G) are added
    - Each base fluoresces in a unique color when hit by a laser.
    - Each base also has a terminator modification (only one base added at a time).
    - After imaging, terminators are chemically removed to allow the next base addition.
    - A camera captures images at each cycle to record which base was added at each spot.
    - The process is repeated over many cycles to build the full sequence.
- Parallel Sequencing
    - Millions of DNA fragments are sequenced in parallel on the same slide.
    - Each position on the slide generates a read (a short DNA sequence).
- Data & Output
    - Final output = reads (sequences of A, C, G, T).
    - Each base has an associated quality score indicating the confidence in the base call.
    - Quality scores are based on signal clarity (color purity) in the images.
- Sources of Error
    - Errors increase in later cycles due to synchronization issues:
        - Some DNA fragments may lag or jump ahead in base addition.
        - Causes mixed signals (impure colors) in clusters.
    - Result: higher error rates in later positions of a read → lower quality scores.

- NGS has made sequencing fast and cheap, enabling creative and powerful new experiments
- Scientific questions that were once too expensive or slow to explore can now be answered using sequencing
- Basic concept: convert molecule (DNA or RNA) into DNA, then apply second-gen sequencing to measure

NGS Applications

1. **Exome Sequencing**
    - Exome: all exons in the genome
    - Exons: protein-coding regions; remain after RNA splicing; translated into proteins
    - Used for: finding mutations that affect protein function
    - Exons only ~1.5% of the genome, so sequencing just them is efficient
    - Process:
        - Fragment whole genome DNA
        - Use capture kits with magnetic beads containing exon-specific DNA
        - Heat DNA to single strands
        - Exonic DNA hybridizes with complementary strands on beads
        - Pull down beads, elute exonic DNA, sequence only the exons
    - Kits today capture ~50–60 million base pairs
2. **RNA-Seq** (RNA Sequencing)
    - Goal: identify which genes are being expressed (turned on)
    - Process:
        - Genes transcribed into RNA, then translated into proteins
        - RNA with poly-A tails (added post-transcription) is targeted
        - Use complementary T sequences to bind poly-A tails and isolate mature mRNA
        - Reverse transcription:
            - RNA → DNA using reverse transcriptase (evolution-derived enzyme)
        - Resulting DNA is sequenced

- Used to identify gene expression patterns in specific cells/tissues
- Computationally intensive to interpret, but essential to understand data origin

3. **ChIP-Seq** (Chromatin Immunoprecipitation Sequencing)
    - Goal: find where specific proteins (e.g., transcription factors) bind to DNA
    - Importance: protein binding affects gene expression regulation
    - Process:
        - Cross-link proteins to DNA in cells to "freeze" their binding positions
        - Fragment DNA; some pieces have proteins attached
        - Use antibodies to isolate specific proteins
        - Pull down protein-DNA complexes
        - Remove protein, sequence attached DNA
    - Result: identifies DNA regions where proteins bind, based on pulled-down fragments

4. **Bisulfite Sequencing / Methyl-Seq**
    - Goal: determine methylation sites on DNA (epigenetic regulation)
    - Methylation: affects gene expression; can be inherited during cell division
    - Methyl groups attach to cytosine (C) bases
    - Process:
        - Split DNA into two identical samples
        - Treat one with bisulfite, which converts unmethylated C → U
        - Methylated Cs remain unchanged
        - Sequence both samples
        - Compare sequences using special aligners that handle C→U conversion
    - Used to measure DNA methylation in cells or tissues

Quiz

1. Genome assembly refers to

   A method for capturing gene sequence

   A computational method to identify the genes being expressed in a cell or tissue

   A computational method for reconstructing chromosomes from short reads

   The process whereby a cell copies its DNA

2. Which of the following is not true about DNA?

   It doesn't matter which direction you write the sequence in

   Each strand has a direction

   It is a double-stranded molecule

   One strand is complementary to the other

3. RNA molecules are translated into

   DNA molecules

   Proteins

   Modified RNA molecules

   Introns

4. Messenger RNA is

   A copy of DNA in which thymine (T) is changed to uracil (U)

   A special signal that helps a cell communicate with other cells

   A reverse copy of DNA

   A shortened version of DNA

5. DNA is copied into DNA in order to

    Create species diversity

    Encourage evolutionary changes

    Replicate a cell

    Respond to an infection

6. Evolutionary biology involves the study of

    Changes in DNA that occur from generation to generation and that sometimes lead to the appearance of new species

    The purpose of life on earth

    The origin of the very first living organisms

    How the cell membrane is formed

7. Which of the following can we measure with next generation sequencing?

    DNA variants

    Protein transport rates

    Nucleus size

    RNA secondary structure

8. What is the first step in ChIP--sequencing to measure protein--DNA binding?

    Sequencing the bound DNA fragments

    Sequencing the DNA molecules

    Cross--linking proteins to the DNA

    Antibody pulldown of the linked protein-DNA fragments

9. Which of the following can be measured using bisulfite conversion and then sequencing?

    DNA methylation

    DNA secondary structure

    DNA--protein binding

    DNA variants

10. What is the primary measurement technology used in most modern genomics experiments?

    Western blotting

    Sanger sequencing

    Nanopore sequencing

    Next generation sequencing

    Oligonucleotide arrays

    Polymerase chain reaction