What is Genomic Data Science?

- The study of genomes (complete sets of DNA) to understand how genetic information shapes biological functions, traits, and diseases.
- Utilizes advanced technologies like sequencing to analyze DNA, RNA, and proteins, often producing massive data sets that require powerful computing for interpretation.

Why Study Genomics?

- Human Genomics: All humans are 99.9% genetically identical, yet we have immense diversity in traits like height, longevity, and disease susceptibility. Genomics helps explain these differences.
- Development: Our genome encodes the entire developmental process from a single cell to a full organism. It determines how different cell types (like neurons vs. skin cells) function despite having identical DNA.
- Cancer: Cancer is driven by mutations in our DNA that disrupt normal cell functions, causing uncontrolled division. Understanding these mutations can help in treatment.
- Mutations: Mutations (DNA changes) can occur from replication errors or damage, and can sometimes lead to diseases like cancer if they affect key genes.

The Central Dogma of Molecular Biology

- DNA → RNA → Protein: Information flows from our DNA to RNA, which is then translated into proteins. Proteins perform most cell functions, like digestion and metabolism.
- Gene Regulation: Beyond the basic flow, proteins and other factors can interact with DNA to regulate which genes are active or inactive, helping explain cellular diversity.

Sequencing & Data Analysis

- Sequencing: The process of reading an organism's DNA sequence. It has become much faster and cheaper, making genomics research more accessible.
- The Human Genome Project: Completed in 2001, it took 12 years and millions of dollars to sequence one human genome. Today, sequencing can be done in a few days for ~$1,000 per genome.
- Data Complexity: Even though sequencing has become faster, analyzing the data generated remains a massive challenge. Powerful computers are needed to process and interpret genomic data.

Applications of Genomics

- By sequencing cancerous tissues, we identify mutations that drive abnormal cell growth, leading to better-targeted therapies.
- Repositories like the NCBI's Sequence Read Archive (SRA) store vast amounts of genomic data, allowing for widespread research and discovery across species.

The Genomic Revolution

- Sequencing costs have plummeted from ~$25 million per human genome to ~$1,000 in just 15 years.
- Public archives provide open access to vast genomic data, enabling researchers to explore and make new discoveries globally.

Key Aspects of Genomics

1. Genome Structure:
    - DNA: The genome consists of DNA molecules made up of four types of nucleotides: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T).
    - Human Genome: Composed of ~3 billion nucleotides across 23 chromosome pairs. The chromosomes are inherited in pairs, with one from the mother and one from the father, except for the sex chromosomes (X and Y).
    - Chromosomal Features: Chromosomes have centromeres (central regions) and telomeres (protective caps at the ends).
2. Genome Function:
    - The genome encodes everything an organism needs for development, from forming organs to controlling processes like respiration, metabolism, and even complex tasks like building a brain.
    - Genes within the genome carry instructions for producing proteins, which perform the functional tasks in the cell.
3. Evolution of Genomes:
    - Genomes evolve over time. While human genomes change very little across generations, large evolutionary changes can be seen when comparing human genomes

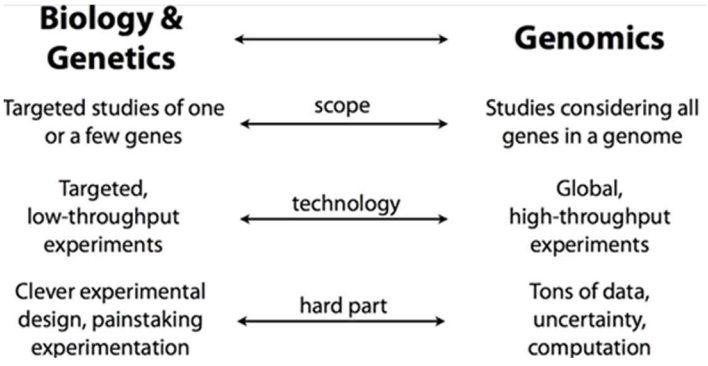
to those of closely related species like chimpanzees or even more distantly related organisms like bacteria.

- Shared DNA across species indicates fundamental biological processes that are conserved throughout evolution.

4. Genome Mapping:
    - Mapping a genome goes beyond just sequencing. It involves locating and identifying genes within the sequence.
    - Genes are often defined as sections of the genome that code for proteins or other functional elements. Our understanding of genes has evolved to include complex regulatory regions, not just protein-coding sequences.

Applications of Genomics

- Medicine: Personalized medicine, where treatments are tailored to an individual's genetic makeup; understanding diseases at the genetic level → better diagnosis, treatment, and prevention strategies.
- Agriculture: Genomics aids in breeding crops and livestock that are more resilient, nutritious, or resistant to diseases.
- Pharmaceuticals: Identifying genetic factors that influence drug responses and disease mechanisms.
- Evolutionary biology, anthropology, and ecology (how species evolve and interact with their environments)



Genomics vs. Traditional Biology and Genetics

1. Scale of Study

- Traditional genetics focused on studying one gene at a time, often taking years to understand each gene's role.
- In contrast, genomics enables the study of all genes at once, providing a broader and more comprehensive view of the genetic landscape.

2. Technology-Drive
    - High-throughput technologies allow scientists to analyze thousands of genes simultaneously.
    - This shift has been fueled by massive computational power, allowing for large-scale data analysis that wasn't feasible before.
3. Data Challenges
    - The downside to this technology is the data explosion. With the ability to sequence entire genomes or thousands of genes at once, scientists now face the challenge of managing and interpreting vast amounts of data. The volume of data requires advanced computational tools and statistical methods to identify meaningful patterns and insights.
4. Big Data and Computation
    - Genomics requires handling complex data, statistical uncertainty, and large-scale computational processes. The challenge isn't just gathering data, but making sense of it to uncover biological insights.

The Process in Genomic Data Science

1. Sample Collection
    - Genomic research begins by collecting samples from subjects. This could be from humans, animals, or model organisms. For instance, skin cells from humans might be collected to study normal human development or disease.
2. Sequencing
    - The samples are then processed in the lab and sent for sequencing. Sequencing generates short fragments of DNA, referred to as reads, which represent parts of the genome.
3. Alignment to the Reference Genome
    - These short fragments (reads) are aligned to a reference genome to compare how they differ from the "average" genome. Currently, there is one reference genome that

represents a standard human genome, though future efforts will likely include several reference genomes.

- A human genome has two copies of each chromosome—one from the mother and one from the father. Genomic data scientists look at the differences between these two copies, which can provide insights into genetic variation and disease susceptibility.

4. Preprocessing and Normalization
    - Sequencing data is subject to various biases and errors that need correction. For example, some regions might be overrepresented, or the sequencing machine itself may make random or systematic errors. Normalization aims to reduce these biases so that the data accurately represents the biological sample.
    - This step is critical because, despite the advanced technology, sequencing isn't perfect. You need to account for these biases before moving on to biological analysis.
5. Statistical and Computational Analysis
    - Once the data is preprocessed, statistical methods and machine learning are applied to draw conclusions from the data. Scientists use these techniques to make sense of large-scale data, identifying genes, mutations, or other features that are significant.

Key Areas in Genomic Data Science

1. Experimental Design
    - Genomic experiments require careful planning. A researcher needs to decide: What scientific question are they trying to answer? How much data is needed? How many subjects are required? What specific types of data (e.g., RNA-seq, ChIP-seq, methyl-seq) will provide the answers?
    - Poor design can lead to data that can't answer the intended scientific question.
2. RNA Sequencing (RNA-Seq)
    - Widely used to study gene expression by capturing and sequencing RNA from cells. By mapping RNA sequences to the genome, you can determine which genes are turned on and how much they are expressed.
3. Software Development
    - Specialized software for processing data, from preprocessing and normalization to drawing biological conclusions.
    - Standardized to handle experiments like RNA-Seq in consistent ways and carefully engineered to work in many contexts/ensure reliability across different data sets.

4. Population Genomics
    - Rather than individual cases, population genomics looks at how genomes differ within larger groups or populations to identify genetic factors that influence disease susceptibility, resistance, and other traits within a population.
    - Example: why some individuals are more resistant to diseases like malaria or HIV within a particular population can provide insights into genetic variation.
5. Integrative Genomics / Systems Biology
    - Combines data from different types of experiments to gain a holistic understanding of biological processes.
    - Combine genomic sequencing with proteomic data or epigenomic data to provide a deeper biological understanding.
    - Crucial for systems biology, which seeks to understand how different parts of the biological system work together to produce complex functions.

Technologies and Techniques in Genomic Data Science

- **ChIP-seq**: Studies DNA-protein interactions by identifying the binding sites of DNA-binding proteins across the genome.
- **Methyl-seq**: Examines DNA methylation patterns, which can influence gene expression without changing the DNA sequence.
- Both technologies, alongside RNA-seq, provide various ways to study the genome, its regulation, and its expression in different conditions.

Challenges in Genomic Data Science

1. Big Data
    - The sheer scale of genomic data is a major challenge. With entire genomes being sequenced and datasets containing millions of reads, data management and analysis can become overwhelming.
    - Statistical uncertainty and computational resources are key challenges that require advanced algorithms and powerful computing infrastructure.
2. Biases and Errors
    - Data quality is always an issue, as sequencing machines can introduce biases, and different tissues or conditions might affect gene expression levels. Biases in data

collection (e.g., underrepresentation of certain genes) must be accounted for to ensure accurate results.

The Three Domains of Life

- At the most basic level, life is divided into three domains: Eukaryotes, Archaea, and Bacteria (prokaryotes = archaea and bacteria)
- Eukaryotes are more complex organisms, including humans, plants, animals, and yeast (a single-celled organism with a nucleus).
- Evolutionary Perspective: Eukaryotes, Archaea, and Bacteria diverged from a common ancestor long ago. Archaea and bacteria share certain similarities, but eukaryotes are more complex with a nucleus and other organelles.

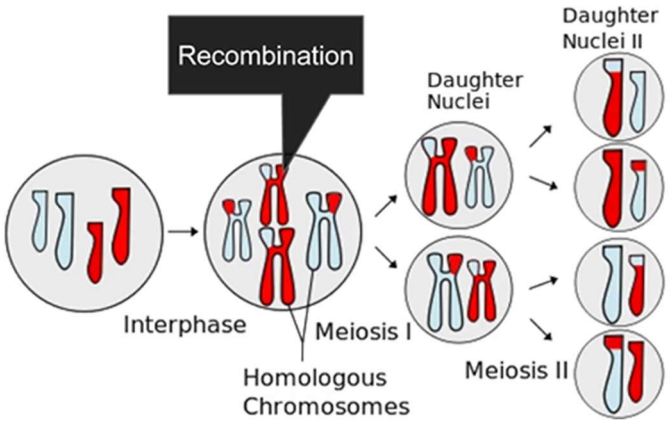
Cell Structure and Organization

1. Eukaryotic Cells
    - Nucleus: DNA is sequestered in a membrane-bound nucleus, which organizes the genetic material into chromosomes.
    - Organelles: In addition to the nucleus, eukaryotic cells contain other organelles (mitochondria, endoplasmic reticulum, golgi apparatus, etc.).
    - Mitochondria: mitochondria have their own DNA, inherited from the mother. Their genome is small (less than 1% of the total human genome), but crucial for energy metabolism.
2. Prokaryotic Cells (Archaea and Bacteria)
    - Prokaryotes organize their DNA in a nucleoid instead of a nucleus; their DNA isn't enclosed within a membrane.

The Cell Cycle

1. Mitosis
    - DNA is replicated, so each chromosome is duplicated. This results in two identical daughter cells, each with a full set of chromosomes (**diploidy**).
    - Human cells are diploid, meaning they contain two copies of each chromosome (one from mother and one from father).
2. Meiosis:

- In the production of egg and sperm cells (gametes), meiosis occurs. This process reduces the chromosome number by half, so each gamete has only one copy of each chromosome.
- **Recombination**: sections of chromosomes from the mother and father "cross over," leading to genetic diversity in offspring.
- Crossing over occurs at random locations, making each child genetically unique, even though they inherit half of their DNA from each parent.



Stem Cells and Differentiation

- Stem cells are undifferentiated cells that can develop into many different types of cells in the body. The process of differentiation allows a stem cell to become specialized for particular functions, such as becoming a blood cell or a nerve cell.
- Example: all blood cells come from a hematopoietic stem cell, which can differentiate into red blood cells, white blood cells, and platelets.

Genetic Diversity and Recombination

- Genetic Diversity: The combination of mutation and recombination contributes to the genetic diversity observed in offspring. While mutation creates new variations in DNA, recombination reshuffles the genetic material during meiosis, ensuring that each child is genetically distinct.
- Family Diversity: Even though children inherit DNA from both parents, the way chromosomes recombine during meiosis results in unique combinations in each individual, which is why siblings are not genetically identical, even though they share the same parents.

Key Molecules in Molecular Biology

1. DNA (Deoxyribonucleic Acid)

- Structure: DNA is the molecule that stores genetic information. It consists of four nucleotides: Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). These nucleotides bind in specific pairs: A always pairs with T, and G always pairs with C. This pairing rule is crucial for DNA replication.
- Double Helix: DNA is structured as a double helix, with two strands winding around each other. Each strand has a direction (5' to 3' end), and the complementary strand is read in the opposite direction.
- Chromosomes: DNA is organized into chromosomes in the nucleus. Humans have 23 pairs of chromosomes, each containing millions of base pairs.
- Base Pairing: The consistent pairing of bases (A-T, G-C) means that if one strand is known, the sequence of the complementary strand is automatically determined. This property is vital for DNA replication.

2. RNA (Ribonucleic Acid)

- Difference from DNA: RNA is similar to DNA but has uracil (U) instead of thymine (T). So, A pairs with U, and G pairs with C in RNA.
- Single-Stranded: Unlike DNA, RNA is single-stranded and is used as a template for protein synthesis.
- Function: RNA carries the genetic information from DNA to the ribosome, where proteins are synthesized. It acts as a transcription of the genetic code in DNA.

3. Proteins

- Amino Acids: Proteins are made of long chains of amino acids. There are 20 standard amino acids that form proteins, although some organisms use additional ones.
- Protein Synthesis: The process of building proteins from RNA is called translation. RNA is read in triplets (codons), with each codon specifying an amino acid. The sequence of amino acids forms the protein's structure.
- Codons: There are 64 possible codons. 61 of them encode amino acids, and 3 are stop codons that signal the end of protein synthesis.

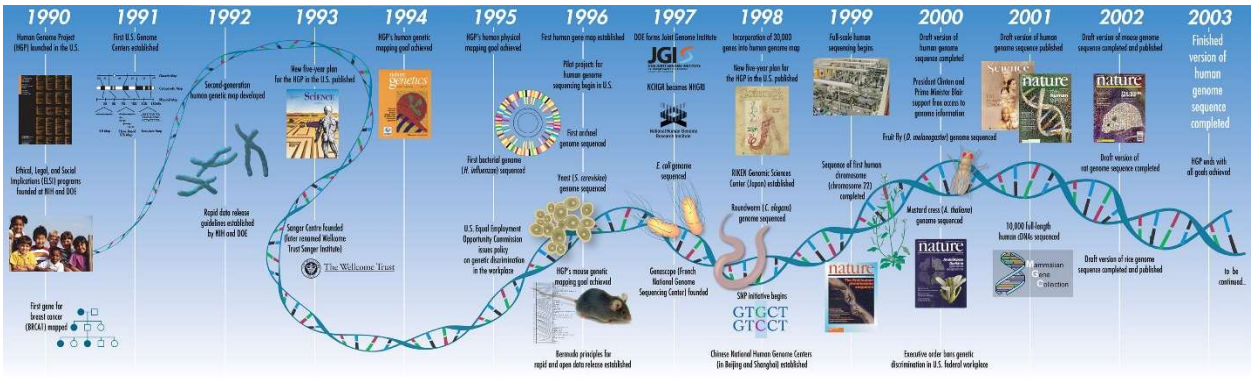
4. DNA Replication and Inheritance

- Genetic Inheritance: DNA is inherited through cell division. DNA replication allows cells to make copies of the genetic material to pass on to daughter cells.
- RNA's Role: RNA, particularly messenger RNA (mRNA), acts as an intermediary between DNA and proteins. It is transcribed from DNA and used in translation to synthesize proteins.
- Protein Structure: Proteins are made from amino acids that are strung together based on the RNA template, and they fold into specific shapes to carry out their biological functions.

Key Molecular Biology Processes

- Transcription: The process where DNA is used as a template to synthesize RNA.
- Translation: The process where mRNA is decoded to build proteins.
- DNA Replication: The process that copies DNA so that it can be passed on during cell division.

Human Genome Project



- Overview
    - Goal: Sequence the entire human genome (3 billion base pairs), identify all human genes, and understand genome function.
    - Proposed by U.S. Department of Energy (DOE) in the late 1980s, later joined by NIH and international partners. Officially began in 1989.
    - Estimated cost and duration: $3 billion (~$1 per base), 15 years (finish 2005).
- Early Challenges
    - Sequencing cost in 1980s: ~$10 per base.

    - Technology: Slow and expensive; sequencing required mapping large DNA chunks (BACs) of ~150,000 base pairs.
    - Skepticism: Many scientists thought most DNA was "junk" and the project wouldn't be worth the cost.
- Mapping Approach (Public Project)
    - Created BAC libraries to organize and map the genome before sequencing.
    - Sequenced mapped chunks systematically, piece by piece.
- Major Disruption: Whole Genome Shotgun Sequencing
    - 1995: TIGR (Craig Venter & Hamilton Smith) sequenced first bacterial genome (*H. influenzae*) using shotgun sequencing.
    - Method skipped mapping — broke DNA into many small fragments and used computers to assemble the genome.
- Formation of Celera Genomics (1998)
    - Craig Venter founded Celera to sequence the human genome faster and cheaper using shotgun sequencing.
    - This initiated the public vs. private "race" to complete the genome first.
- The Race Intensifies
    - Public project ramped up efforts (merging centers, increasing sequencing speed).
    - 1999-2000: Celera and public project raced toward a 2001 draft completion.
    - 2000: Clinton and Blair announced the "completion" of the human genome (draft form).
- Draft Genome (2001)
    - Two papers published:
        - Nature (public consortium): estimated 30,000–40,000 genes.
        - Science (Celera): estimated ~26,000 confirmed + ~12,000 likely genes.
    - Genome was ~92% complete at the time.
    - Represented a mosaic genome from ~12 anonymous individuals (all of Northern European descent).
- Gene Count Evolution
    - 1960s estimate: ~6.7 million genes (based on hemoglobin gene weight).
    - 1990s estimate: ~100,000 genes.

    - Post-HGP: estimate dropped to ~20,000–23,000 protein-coding genes.
    - Also: tens of thousands of non-coding RNA genes — gene count remains imprecise.
- Project Outcomes
    - Finished ahead of schedule (draft by 2001, target was 2005); costs fell below $1 per base — by 2001, ~$1 per 700 bases; ~$1 per 3 million bases (thanks to next-gen sequencing).
    - Sparked a revolution in genomics, personalized medicine, disease research, and biotechnology.

DNA Structure and Packaging

- DNA: Very long (~2 meters long per cell), contained within 23 chromosome pairs in humans.
- DNA Packaging: Wrapped around histones to form a "beads-on-a-string" structure. Further coiled into chromosomes.
- Coiling Process: DNA coils into larger structures to fit inside cells, and these structures are essential for processes like transcription.

Repeats in DNA

- Tandem Repeats: Identical sequence repeated consecutively (e.g., ATTCG repeated 3 times).
    - Can be very long (e.g., centromeres consist of repeats of 180 base pairs repeated hundreds of thousands of times).
- Interspersed Repeats: Identical or near-identical sequences scattered across chromosomes.
- Issues in Analysis: Repeats can complicate DNA sequencing, as repeated sequences may be hard to place in the genome.

RNA Structures

- Messenger RNA (mRNA): Carries genetic information from DNA to ribosomes for protein synthesis.
    - Untranslated Regions (UTRs): Portions of mRNA (5' UTR and 3' UTR) that don't code for proteins but regulate mRNA stability and translation.
    - Coding Sequence: The part of mRNA that gets translated into proteins.

- Poly-A Tail: A series of adenine (A) nucleotides added to the 3' end of mRNA after transcription; important for mRNA stability and export.

Gene Structure (Introns & Exons)

- Exons: Protein-coding regions of genes.
- Introns: Non-coding regions that are spliced out during mRNA processing.
- Alternative Splicing: Different combinations of exons are joined together to form multiple mRNA variants, allowing different proteins to be produced from the same gene. Over 90% of human genes undergo alternative splicing, increasing protein diversity.

Protein Structure

- Amino Acids: Proteins are made of long chains of amino acids (hundreds – thousands).
- Secondary Protein Structures
    - Alpha Helices: Coiled structures.
    - Beta Sheets: Flat, sheet-like structures.
- Protein Function: The 3D structure of a protein determines its function, and understanding this is essential for understanding biological processes.

Transcription Factors

- Proteins that regulate gene expression by binding to DNA and controlling the transcription of genes.
- Can activate or inhibit gene expression by binding to regions upstream or downstream of a gene.

Epigenetics

- Mechanisms that regulate gene expression without altering the DNA sequence itself.
- DNA Methylation: The addition of methyl groups to DNA, which can silence gene expression.
    - Inheritance of Methylation: Methylation marks are passed on during cell division (not between generations).

- Epigenetic changes are important in regulating cell function and gene expression in response to environmental factors.

Genotype vs. Phenotype

- Genotype: Genetic makeup of an organism, including all the sequences and mutations in its genes.
    - Determines how the body and cells function and influences traits or diseases.
- Phenotype: Observable traits or characteristics, such as:
    - Physical traits: hair color, eye color, height, weight.
    - Health traits: genetic diseases, health conditions.
    - Personality or behavior can also be considered part of the phenotype.
- The relationship between genotype (genes) and phenotype (traits) is studied to understand how genes affect observable characteristics.
- The relationship between genetic variation and observable traits is complex. While genetics play a significant role in traits and diseases, other factors (environment, lifestyle) also contribute to the phenotype.

Mendelian Genetics: Example of Pea Color

- Dominant trait: A trait that appears if at least one allele (gene copy) carries the mutation.
- Recessive trait: A trait that only appears if both alleles carry the mutation.
- Pea Color Example
    - Green peas (yy) have two recessive alleles (lowercase y).
    - Yellow peas (Yy or YY) have at least one dominant allele (uppercase Y).
    - Cross of yellow pea (Yy) × green pea (yy) results in 4 possible offspring genotypes (Yy and yy).
    - Yy offspring are yellow (dominant trait), yy offspring are green (recessive trait).

Genetic Variation Across Populations

- Global Variation: Different regions of the world have characteristic genetic mutations.

    - Mutations can include single nucleotide polymorphisms (SNPs) or larger DNA changes (insertions/deletions).
- Geographical Clustering: Using principal component analysis (PCA) of genome data, genetic variation tends to cluster by geographic regions (e.g., people from Spain cluster genetically).

Example of Genotype-Phenotype Association (HERC2 Gene and Eye Color)

- AA Genotype: 85% chance of brown eyes, 14% chance of green eyes, 1% chance of blue eyes.
- GG Genotype: 72% chance of blue eyes, 27% chance of green eyes.
- AG Genotype: Likely brown eyes (since A is dominant).
- Phenotype Behaves Like a Recessive Trait: If you inherit two Gs, you'll likely have blue eyes; one G results in green/brown eyes.

Genome-Wide Association Studies (GWAS)

- Studies large populations to find associations between genetic variations (e.g., SNPs) and specific traits or diseases.
    - Example: SNP1 (mutation at a specific location) was significantly more common in individuals with a certain disease.
    - SNP1 vs. SNP2: SNP1 showed a higher frequency of a specific mutation in diseased individuals compared to healthy controls, suggesting an association.
    - Important Note: Even if an SNP is associated with a disease, having that SNP doesn't guarantee you will develop the disease; it just increases the likelihood.
- Understand how specific genetic variants contribute to diseases and traits.

Quiz

1. The central dogma of molecular biology tells us that information is passed from
   DNA to epigenetics to protein
   DNA to methylation to RNA to protein
   DNA to RNA to protein
   RNA to DNA to protein

2. Which of the following is one of the major drivers of the 2008 sequencing revolution?
   Improved Sanger sequencing
   Decreased computational analysis time
   Decreased cost of sequencing
   Increased sample collection

3. Which of the following is an exclusive characteristic of genomics compared to traditional biology?
   Massive amounts of data
   Targeted studies of one or a few genes
   Measurements of molecules in the Central Dogma
   Sequencing

4. Genomic data science involves techniques from which of these disciplines?
   Molecular Biology
   Statistics
   All of the these options
   Computer Science

5. Which of the following is an activity that genomic data scientists do not perform?
   Pipetting
   Integrative genomics
   Statistics and machine learning
   Population genomics

6. Which of these is not one of the DNA nucleic acids?
   Alanine
   Adenine
   Cytosine
   Thymine

7. Transcription is a process that converts DNA to
   RNA
   polymerases
   Any other molecule
   genes

8. The cost to sequence a human genome today, in U.S. dollars, is approximately
   $3 billion
   $30 million
   None of these options
   $1000

9. DNA encodes instructions for
   Producing all the proteins that a person requires for life
   Helping us digest food
   Enveloping viruses that infect a cell
   Regulating body temperature

10. One major difference between humans and bacteria is
    Human cells contain separate organelles called mitochondria, and bacterial cells do not.
    The human genome is made of DNA, while bacteria are made of RNA.
    Human genes are first transcribed to RNA, while bacterial genes are not.
    Human proteins are made of combinations of 20 amino acids, while bacterial proteins use a smaller set of 12 amino acids.