Bioinformatics Study Guide With Notes

30 day study plan:

Week 1: Fundamentals of Biology and Bioinformatics

Day 1:

- Understand the basics of molecular biology: DNA, RNA, and proteins

  DNA

  - DNA stands for deoxyribonucleic acid.
  - It is a molecule that carries genetic information.
  - DNA is made up of four nucleotide bases: adenine (A), cytosine(C), guanine (G) and thymine (T).
  - The sequence of these bases determines the genetic code.

  RNA

  - RNA stands for ribonucleic acid.
  - It is a molecule that helps to carry out the instructions in DNA.
  - RNA is made up of four nucleotide bases: adenine (A), cytosine(C), guanine (G) and uracil (U).
  - There are several types of RNA, including messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA) and small nuclear RNA (snRNA).

  Proteins

  - Proteins are large molecules made up of amino acids.
  - They perform a wide variety of functions in the body, including catalyzing metabolic reactions, replicating DNA, responding to stimuli and transporting molecules.

- Familiarize yourself with central dogma and the genetic code

  Central Dogma

  - The central dogma of molecular biology explains the flow of genetic information within a biological system.
  - It states that DNA codes for RNA, which codes for proteins.
  - This process involves two main steps: transcription and translation.
  - In transcription, the genetic code in DNA is transcribed into RNA.
  - In translation, this RNA is then translated into proteins.

  Genetic Code

  - The genetic code is the set of rules by which information encoded in DNA is translated into proteins.
  - It is a triplet code, meaning that each amino acid is specified by a sequence of three nucleotide bases (codon).
  - There are 64 possible codons (4^3), but only 20 amino acids. This means that most amino acids are specified by more than one codon.

Day 2:

- Study gene structure and function: transcription, translation, and regulation
  - The regulated transcription of genes determines cell identity and function [1].
  - Gene expression involves multiple levels of regulation and is often influenced by transcription factors [2].
  - Transcription factors can function as activators or enhancers [2].
  - Transcription takes place in three stages: initiation, elongation, and termination [3].
- Learn about genome organization and sequencing
  - **Genome organization** refers to the linear order of DNA elements and their division into chromosomes. It can also refer to the 3D structure of chromosomes and the positioning of DNA sequences within the nucleus[1].

  - A **genome** is the complete set of DNA sequences in an organism and contains all of the instructions required for that organism to function[2].
  - The **Genome Sequencing Program** was initially created to sequence the human genome, as part of the Human Genome Project[3].
  - There are many methods for analyzing 3D genome organization, including sequencing-based techniques such as Hi-C and its derivatives, Micro-C, DamID, etc., as well as microscopy-based techniques[4].

Day 3:

- Understand protein structure and function, including primary, secondary, tertiary, and quaternary structures

  Protein Structure and Function

  - Proteins are made up of long chains of amino acids.
  - The arrangement and placement of amino acids give proteins certain characteristics.

  Levels of Protein Structure

  1. **Primary structure**: A simple linear chain of amino acids (also known as a polypeptide chain).
  2. **Secondary structure**: Determined by the angles of the covalent bonds between and within amino acids.
  3. **Tertiary structure**: Determined largely by attraction between different parts of the molecule.
  4. **Quaternary structure**: Refers to how multiple polypeptide chains come together to form a functional protein.

- Familiarize yourself with protein-protein interactions and protein-DNA interactions

  Protein-DNA Interactions

  - **Role in biological processes**: Protein-DNA interactions play a significant role in many biological processes such as regulation of gene expression, DNA replication, repair, transcription, recombination and packaging of chromosomal DNA.

  - **Types of interactions**: Proteins interact with DNA through electrostatic interactions (salt bridges), dipolar interactions (hydrogen bonding), entropic effects (hydrophobic interactions) and dispersion forces (base stacking).

Day 4:

- Learn about biological databases (e.g., NCBI, Ensembl, UniProt)

  Biological Databases:

  - **NCBI**, **UniProt**, and **Ensembl** are examples of biological databases.
  - These databases integrate a vast number of biological data from various public resources.
  - They provide a queryable interface to all the data available and can convert identifiers from one database into another .

- Understand how to access and analyze data from these databases using APIs or web interfaces

  - To access and analyze data from databases using APIs or web interfaces, you can use API to extract data and save it in a dataframe
    . APIs allow communication between an application and a database management system
    . You can also build programs that run searches on the data the server is hosting and transform that information into a different, usable format.

Day 5:

- Get started with sequence alignment: global, local, and multiple alignments

  Sequence Alignment:

  - Used to find regions of similarity between two or more sequences.

  Types of Sequence Alignment:

  1. **Global Alignment:**
     - Tries to find the best overall alignment between sequences.
     - A common global alignment algorithm is Needleman-Wunsch algorithm.
  2. **Local Alignment:**

        - Focuses on finding the best alignment between two segments of the sequences.

    Multiple Sequence Alignments:

    - Aligns more than one sequence at a time.
    - Can use either global or local alignments.

- Learn about popular alignment tools, such as BLAST, Clustal Omega, and MUSCLE
    - **Clustal Omega** is a fast and accurate aligner suitable for alignments of any size. It uses mBed guide trees and pair HMM-based algorithm which improves sensitivity and alignment quality [1].
    - **MUSCLE** is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options [2].
    - The **Multiple Sequence Alignment Viewer application (MSA)** is a web application that visualizes alignments created by programs such as MUSCLE or CLUSTAL, including alignments from NCBI BLAST results [3].

Day 6:

- Introduction to phylogenetics and its applications in bioinformatics

    Phylogenetics is the study of evolutionary relationships among biological entities – often species, individuals or genes (which may be referred to as taxa)[1]. Phylogenetic analysis refers to the evolutionary study of organisms forming a tree called a phylogenetic tree[2]. A phylogenetic tree is a visual representation of the relationship between different organisms, showing the path through evolutionary time from a common ancestor to different descendants[2].

    Phylogenetics based on sequence data provides us with more accurate descriptions of patterns of relatedness than was available before molecular sequencing. Phylogenetics now informs the Linnaean classification of new species[3].

- Learn about tree-building methods and phylogenetic software (e.g., MEGA, RAxML)
    - RAxML (Randomized Axelerated Maximum Likelihood) is a program for sequential and parallel Maximum Likelihood based inference of large phylogenetic trees[1].

- MEGA is an easy-to-use software with multiple features including aligning sequences, estimating evolutionary distances, building trees using several methods, testing tree reliability, marking genes/domains, testing for selection and computing sequence statistics**2**.

Day 7:

- Review Week 1 material and practice using online resources or real-life datasets

Week 2: Programming for Bioinformatics and Data Analysis

Day 8:

- Learn Python for bioinformatics: essential libraries (NumPy, Pandas, Biopython, and Matplotlib)

1. **NumPy** is an array manipulation library that is behind libraries such as Pandas, Matplotlib, Biopython, and scikit-learn**1**. It provides efficient array and matrix support**2**.
2. **Pandas** (Python data analysis) is a must in the data science life cycle. It is heavily used for data analysis and cleaning**3**.
3. **Biopython** is a set of freely available tools for biological computation written in Python by an international team of developers. It addresses the needs of current and future work in bioinformatics**4**.
4. **Matplotlib** is a plotting library that works hand-in-hand with NumPy.

Day 9:

- Get started with Python for bioinformatics: essential libraries (Bioconductor, ggplot2, and dplyr)
  - **Bioconductor** is a collection of R packages that can be used directly from Python via rpy2 or similar**1**. It provides tools for the analysis and comprehension of high-throughput genomic data**2**.
  - **ggplot2** is a system for declaratively creating graphics based on The Grammar of Graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details**3**.
  - **dplyr** is a fast and consistent tool for working with data frame-like objects both in memory and out of memory**4**.

- Practice data manipulation and visualization with Python
  - Data visualization is a crucial aspect of data analysis as it helps in understanding data better. Python provides several libraries for data visualization such as **pandas**, **matplotlib**, and **seaborn**
    . You can learn how to import, clean, and visualize data using these powerful libraries

Day 10:

- Understand the basics of high-throughput sequencing data (RNA-Seq, ChIP-Seq, and Whole-Exome Sequencing)

  High-throughput sequencing data refers to data generated by high-throughput sequencing technologies such as RNA-Seq, ChIP-Seq, and Whole-Exome Sequencing.

  RNA-Seq (RNA-sequencing) is a technique that can examine the quantity and sequences of RNA in a sample using next-generation sequencing (NGS). It analyzes the transcriptome, indicating which of the genes encoded in our DNA are turned on or off and to what extent.

  ChIP-Seq (Chromatin Immunoprecipitation Sequencing) is a method used to analyze protein interactions with DNA. ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. It can be used to map global binding sites precisely for any protein of interest.

  Whole Exome Sequencing (WES) is a genomic technique for sequencing all of the protein-coding regions of genes in a genome (known as the exome). It consists of two steps: selecting only the subset of DNA that encodes proteins and then sequencing it.

- Learn about the file formats used in bioinformatics (FASTA, FASTQ, SAM/BAM, and VCF)
  - **FASTA**: A text-based format for representing nucleotide or peptide sequences. The sequence is represented in a single line preceded by a description line starting with '>'.
  - **FASTQ**: A text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Both the sequence

letter and quality score are encoded with a single ASCII character for brevity.

  - **SAM/BAM**: SAM (Sequence Alignment Map) is a tab-delimited text format for storing sequence alignment data. BAM (Binary Alignment/Map) is the compressed binary version of SAM, which stores the same data in a compressed, indexed, binary form.
  - **VCF (Variant Calling Format/File)**: A text file format that contains information about variations found at specific positions in a reference genome.

Day 11:

- Analyze RNA-Seq data: quality control, read mapping, and quantification
  1. **Quality control**: It's important to check the quality of your data before proceeding with analysis. This can be done using tools such as FastQC.
  2. **Read mapping**: After quality control, reads need to be mapped (i.e., aligned) to a reference genome using programs such as TopHat or STAR.
  3. **Quantification**: The simplest approach to quantifying gene expression by RNA-seq is to count the number of reads that map (i.e., align) to each gene (read count) using programs such as HTSeq-count.
- Learn about differential gene expression analysis using tools such as DESeq2 and edgeR
  - Differential gene expression analysis is an important aspect of bulk RNA sequencing (RNAseq). A lot of tools are available for this purpose, and among them **DESeq2** and **edgeR** are widely used.
    DESeq2 provides methods to test for differential expression by use of negative binomial generalized linear models; the estimates of dispersion and logarithmic fold changes incorporate data-driven prior distributions.

Day 12:

- Analyze ChIP-Seq data: quality control, peak calling, and motif analysis
  - ChIP-Seq analysis pipeline consists of several steps including raw data processing, quality control analysis, alignment to the reference genome, quality check of the aligned reads, peak calling, annotation and visualization.
    Peak calling programs help define sites of Protein:DNA binding by identifying

regions where sequence reads are enriched in the genome after mapping. . Motif analysis can be performed using tools such as RSAT peak-motifs.

- Learn about tools such as MACS2, HOMER, and MEME Suite
    - **MACS2** is a tool for analyzing ChIP-seq data. It can be installed from its page on PyPI.
    - **HOMER** is another tool for analyzing ChIP-seq data. It uses Poisson distributions to generate background models prior to statistical testing for peak significance between ChIP and control samples.
    - The **MEME Suite** is an integrated collection of tools for discovering and characterizing sequence motifs in collections of DNA or protein sequences. Its flagship program is MEME, which finds motifs in unaligned collections of DNA and sequence motifs.

Day 13:

- Analyze Whole-Exome Sequencing data: quality control, read mapping, and variant calling
    1. **Quality Control (QC)**: The first step of a variant calling pipeline involves evaluating the quality of raw sequencing data.
    2. **Raw reads preprocessing**: This step is necessary for high-quality results.
    3. **Short reads mapping**: Mapping short reads is one of several steps in whole exome sequencing data analysis.
    4. **Post-alignment processing**: This step is also necessary for high-quality results.
    5. **Variant calling and annotation**: Variant calling is a crucial procedure for whole exome, targeted panels, and whole genome sequencing.
    6. **Variant prioritization**: This step helps prioritize variants based on their relevance.
- Learn about tools such as ANNOVAR
    - • **ANNOVAR** is a variant annotator. Given a vcf file from an unknown sample and existing data about genes, other known SNPs, gene variants, etc.,

ANNOVAR will place the discovered variants in context[1]. It is one of the most powerful yet simple to run variant annotators available.

Day 14:

- Review Week 2 material and practice with real-life datasets

Week 3: Machine Learning and AI in Bioinformatics

Day 15:

- Understand the basics of machine learning and AI, including supervised and unsupervised learning
  - **Artificial intelligence (AI)** refers to the general ability of computers to emulate human thought and perform tasks in real-world environments.
  - **Machine learning** is a subset of AI that enables systems to identify patterns, make decisions, and improve themselves through experience and data.
  - Machine learning algorithms allow AI to not only process data but also use it to learn and get smarter without needing any additional programming.
- Learn about popular algorithms, such as linear regression, k-means clustering, and decision trees
  - **Linear Regression**
    . It is a machine learning algorithm based on supervised learning that performs a regression task[1]
    . The algorithm uses independent variables to model a target prediction value and is mainly used for finding out the relationship between variables and forecasting.
  - **k-means clustering**. It is one of the simplest and most popular unsupervised machine learning algorithms for data scientists[1]. The algorithm is deployed to discover groups that haven’t been explicitly labeled within the data[1].
  - **Decision trees**, on the other hand, are a non-parametric supervised learning algorithm utilized for both classification and regression tasks[2]. The algorithm has a hierarchical tree structure consisting of a root node, branches, internal nodes and leaf nodes[2].

Day 16:

- Dive into deep learning: neural networks, convolutional neural networks (CNNs), and recurrent neural networks (RNNs)
    - Deep learning is a subset of machine learning that includes deep neural networks, deep belief networks, and recurrent neural networks.
    - There are three fundamental architectures of neural network that perform well on different types of data: feedforward neural network (FFNN), recurrent neural network (RNN), and convolutional neural network (CNN).
    - CNN is the most commonly used architecture for deep learning. RNNs are a type of neural network that have feedback loops in the recurrent layer which allows them to maintain information in memory over time and process data sequentially.
- Understand their applications in bioinformatics (e.g., protein structure prediction and gene expression analysis)

    Machine learning has many applications in bioinformatics. For example, it can be used for protein structure prediction[1][2], drug discovery[3][4], and gene expression analysis[5][6].

    In protein structure prediction, machine learning methods are widely used to predict the folded structure of a protein from its sequence information[1]. The most recent successes of machine learning for protein structure prediction arise with the application of deep learning to evolutionary information[2].

    In drug discovery, machine learning approaches provide a set of tools that can improve discovery and decision making for well-specified questions with abundant, high-quality data[4]. Machine learning approaches provide tools and algorithms to improve drug discovery by effectively modeling many physicochemical properties of drugs like toxicity, absorption, drug-drug interaction, carcinogenesis, and distribution using QSAR techniques[7].

    In gene expression analysis, a machine learning algorithm can be used for differential expression analysis by selecting genes that are more relevant for discriminating between different health outcomes (e.g., sick versus healthy patients)[5]. Machine learning-based feature selection approaches help to select required genes from large datasets while preserving informative attributes[6].

Day 17:

- Learn about natural language processing (NLP) techniques and their applications in biomedical text mining
    - Natural language processing (NLP) and text mining are branches of biomedical informatics that deal with processing prose for purposes such as extracting information and cohort retrieval.
      Biomedical text mining incorporates ideas from NLP, bioinformatics, medical informatics and computational linguistics.
      Text mining systems are powerful tools that automatically extract and integrate information in large textual collections.
- Practice using NLP libraries, such as NLTK, spaCy, and gensim
    - **NLTK** (Natural Language Toolkit) is one of the premier libraries for developing Natural Language Processing (NLP) models.
    - **spaCy** is a new NLP library that's designed to be fast, streamlined, and production-ready. It's not as widely adopted but it's an excellent choice for building a new application.
    - **gensim** is most commonly used for topic modeling and similarity detection. It also provides methods to remove stopwords.

Day 18:

- Study network biology and graph-based methods in bioinformatics
    - Network biology is useful for modeling complex biological phenomena and has attracted attention with the advent of novel graph-based machine learning methods. Graph neural networks (GNNs), as a branch of deep learning in non-Euclidean space, perform particularly well in various tasks that process graph structure data. With the rapid accumulation of biological network data, GNNs have also become an important tool in bioinformatics.
- Learn about graph databases (e.g., Neo4j) and graph algorithms (e.g., PageRank, community detection)

  Graph Databases

    - A graph database is a type of database that uses graph structures for semantic queries with nodes, edges, and properties to represent and store data.

    - **Neo4j** is a popular open-source graph database project that was first released in February 2010 .
    - Neo4j is a native graph database which means it implements a true graph model all the way down to the storage level .
    - Neo4j provides ACID transactions, cluster support, runtime failover, robust transactional guarantee and unmatched reliability .

  Graph Algorithms

    - Graph algorithms are used to evaluate how groups of nodes are clustered or partitioned.
    - **PageRank** algorithm measures the importance of each node within the graph based on incoming relationships and importance of corresponding source nodes **4**.
    - **Community detection** algorithms evaluate how groups of nodes are clustered or partitioned as well as their tendency to strengthen or break apart **5**.

Day 19:

- Understand the basics of molecular docking and virtual screening for drug discovery
    - Molecular docking and virtual screening are computational techniques used in drug discovery. Structure-based virtual screening (SBVS), also known as molecular docking, has been increasingly applied to discover small-molecule ligands based on protein structures.
      . Docking and simulation techniques have also been applied to analyze features of the active site.
    - Molecular docking is a computational technique used to predict the intermolecular framework formed between a protein and a small molecule or a protein and protein.
    - Structure-based virtual screening (VS) has been a staple for more than a decade now in drug discovery with its underlying computational technique, docking, extensively studied.
    - The idea behind VS is that a library of small compounds are docked into the binding pocket of a protein (e.g., receptor, enzyme), and among the top-ranked

    solutions per molecule, a choice is made on the fraction of compounds to be moved forward for testing toward hit identification.

  - The underlying principle of VS is that it differentiates between active and inactive compounds, thus reducing the number of molecules moving forward and possibly offering a complementary tool to high-throughput screening (HTS).

- Familiarize yourself with popular software, such as AutoDock Vina and Schrödinger Suite

  - AutoDock Vina is an open-source program for doing molecular docking. It was originally designed and implemented by Dr. Oleg Trott in the Molecular Graphics Lab at The Scripps Research Institute. Schrödinger Suite is a powerful software package for computational chemistry and materials science. It includes a range of tools for quantum chemistry, molecular dynamics, and structural biology.

  AutoDock Vina:

  - Open-source program for molecular docking
  - Designed and implemented by Dr. Oleg Trott in the Molecular Graphics Lab at The Scripps Research Institute
  - One of the fastest and most widely used open-source docking engines
  - Based on a simple scoring function and rapid gradient-optimization conformational search

  Schrödinger Suite:

  - Powerful software package for computational chemistry and materials science
  - Includes a range of tools for quantum chemistry, molecular dynamics, and structural biology
  - User-friendly interface and availability on multiple platforms
  - Used by industry leaders worldwide for drug discovery as well as materials science in fields such as aerospace and energy

Day 20:

- Explore reinforcement learning and its applications in drug discovery and protein folding

    - Reinforcement learning is a type of machine learning that can be used to improve drug discovery and protein folding. For example, deep generative networks can substantially reduce the laborious, long and costly process of drug discovery for a protein target. Another approach is using an autonomous molecule generation model that requires only a target protein structure and directly modifies ligand structures to obtain higher predicted binding affinity for the target protein without any other training data. There's also ReLeaSE which integrates two deep neural networks—generative and predictive—that are trained separately but are used jointly to generate novel targeted chemical libraries.
- Learn about the AlphaFold algorithm by DeepMind
    - AlphaFold is a powerful AI algorithm developed by DeepMind that uses deep learning to predict a protein's three-dimensional (3D) shape down to the width of an atom. It has predicted the structure of nearly all 20,000 proteins expressed by humans. In an independent benchmark test that compared predictions to known structures, AlphaFold was able to predict the shape of a protein to a good standard 95% of time.

Day 21:

- Review Week 3 material and practice with real-life datasets or case studies

Week 4: Applying Bioinformatics to Disease Research

Day 22:

- Understand the basics of genomics and personalized medicine
    - Genomics-based personalized medicine uses information about a person's genes to prevent, diagnose, and treat disease. It's the customization of lifestyle (diet, exercise, medication, sleep, stress) based on the study of a person's genome . Like fingerprints, everyone's genetic make-up is unique .
- Learn about Genome-Wide Association Studies (GWAS) and their role in identifying disease-associated genes
    - Genome-Wide Association Studies (GWAS) screen the genome for associations between millions of genetic variants and a disease or trait without any a priori hypothesis. As such, GWASs may reveal new genes and pathways not previously implicated in the disease pathology. The goal of GWAS is to

screen the entire genome of large numbers of individuals to look for associations between millions of genetic variants within those individuals and their disease outcomes or sometimes for associations between the variants and non-disease traits such as height.

Day 23:

- Study epigenetics and its role in diseases (e.g., cancer, neurological disorders)

  Epigenetics is the regulation of gene expression through alterations in DNA or associated factors (other than the DNA sequence). These factors control the diverse manifestations of diseases. Insights into epigenetic modification may lead to new therapies for common diseases.

  Epigenetic changes are responsible for human diseases, including Fragile X syndrome, Angelman's syndrome, Prader-Willi syndrome, and various cancers. Host epigenetic factors have been shown to play an important role throughout the viral life cycle and have been implicated in the pathogenesis of several viral infections.

- Learn about epigenetic modifications and their impact on gene expression
  - Epigenetic modifications are heritable changes in gene expression patterns that do not involve changes to the underlying DNA sequence. These modifications can affect gene expression by turning genes "on" and "off". One type of epigenetic change is DNA methylation, which works by adding a chemical group to DNA that blocks proteins from attaching to DNA and "reading" the gene.

Day 24:

- Understand the role of the microbiome in human health and disease

  The microbiome is the collection of all microbes, such as bacteria, fungi, viruses, and their genes that naturally live on our bodies and inside us. These microbiomes support and maintain your health but also have been linked to hundreds of ailments such as cancers, autoimmune diseases and cardiovascular diseases when disturbed.

  Environmental exposures can also disrupt a person's microbiome in ways that could increase the likelihood of developing conditions such as diabetes, obesity, cardiovascular diseases and neurological diseases.

- Learn about metagenomics and its applications in disease research
    - Metagenomics is a powerful tool that has been applied to identify unknown pathogens in outbreaks of disease. It can also be used for discovery and detection of pathogens in clinical samples. Applications of clinical metagenomics include infectious disease diagnostics for a variety of syndromes and sample types.

Day 25:

- Dive into cancer genomics and the role of bioinformatics in understanding tumor biology
    - Cancer genomics is the study of cancer genomes to reveal abnormalities in genes that drive the development and growth of many types of cancer. Bioinformatics plays a critical role in managing, storing, annotating, and reporting on data produced by sequencing platforms. Computational methods are used to distill biologically meaningful information from cancer genome sequencing data.
    - Cancer genomics is the study of cancer genomes to reveal abnormalities in genes that drive the development and growth of many types of cancer.
    - Bioinformatics plays a critical role in managing, storing, annotating, and reporting on data produced by sequencing platforms.
    - Computational methods are used to distill biologically meaningful information from cancer genome sequencing data.
    - Popular tools include those that identify point mutations, copy number alterations, structural variations and mutational signatures in cancer genomes.
- Learn about The Cancer Genome Atlas (TCGA) and other cancer genomics resources
    - The Cancer Genome Atlas (TCGA) is a landmark cancer genomics program that molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. This joint effort between NCI and the National Human Genome Research Institute began in 2006. The project aimed to catalogue genetic mutations responsible for cancer using genome sequencing and bioinformatics.

    - **What is TCGA?** TCGA is a landmark cancer genomics program that molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. This joint effort between NCI and the National Human Genome Research Institute began in 2006.
    - **What was the aim of TCGA?** The project aimed to catalogue genetic mutations responsible for cancer using genome sequencing and bioinformatics.
    - **How has TCGA transformed cancer research?** TCGA catalyzed considerable growth and advancement in the computational biology field by supporting the development of high-throughput genomic characterization technologies, generating a massive quantity of data, and fielding teams of researchers to analyze the data.

Day 26:

- Study the role of bioinformatics in infectious disease research (e.g., COVID-19, HIV)
    - Bioinformatics plays an integral role in combating infectious diseases by providing valuable insights into disease prevention, diagnosis, treatment and management. It can be used to identify genes associated with a particular disease and has contributed immensely to our understanding of infectious diseases .
    - Bioinformatics uses advanced computing, mathematics, and different technological platforms to physically store, manage, analyze and understand biological data [1]. This data can provide insights into the biology of disease and expedite progress towards precision medicine [2]. Precision medicine tailors prevention, diagnosis and treatment based on the molecular characteristics of a patient's disease [2].

  For example, bioinformatics tools can help biologists explore the aetiology of neurodegenerative diseases [3] and some biomarkers can be utilized to promote drug repurposing as well as de novo drug design [3]. Pharmaceutical companies are also deploying bioinformatics in cancer treatment and diagnosis [4].

- Learn about pathogen genomics and the use of bioinformatics for vaccine development

  In the field of infectious diseases, genomics can be a useful tool for guiding vaccine development. Given the inevitability and increasing prevalence of antibiotic resistance, vaccines against pathogenic microbes can be even more valuable than antibiotics as a strategy to prevent serious or deadly infectious diseases.

Genomics has catalyzed a shift in vaccine development towards sequence-based ‘Reverse Vaccinology’ approaches, which use high-throughput in silico screening of the entire genome of a pathogen to identify genes that encode proteins with the attributes of good vaccine targets. In conjunction with adjunct technologies such as bioinformatics, random mutagenesis, microarrays and proteomics, a systematic and comprehensive approach to identifying vaccine discovery can be undertaken.

Day 27:

- Explore the role of bioinformatics in rare disease research and diagnosis
    - Bioinformatics can play a crucial role in rare disease research and diagnosis. Rare diseases pose unique problems that bioinformatics can help address. Novel computational methodologies and approaches that use and combine exome and genome sequencing, combined with well-described phenotypic profiles from patients, can help understand the causes and mechanisms of rare diseases.
    -

  key points about the role of bioinformatics in rare disease research and diagnosis:

    1. **Understanding the causes and mechanisms of rare diseases**: Bioinformatics can help address unique problems posed by rare diseases. Novel computational methodologies and approaches that use and combine exome and genome sequencing, combined with well-described phenotypic profiles from patients, can help understand the causes and mechanisms of rare diseases .
    2. **Streamlining diagnosis**: Whole-genome sequencing and phenotype data sharing can be introduced in a national health system to streamline diagnosis and to discover coding and non-coding variants that cause rare diseases.
    3. **Integration of several data types**: A broad range of bioinformatic approaches applied to rare diseases can include novel methods to databases, servers, pipelines, integration of several data types, modeling and systems biology.
- Learn about tools and resources for studying rare diseases, such as OMIM and Orphanet

- OMIM (Online Mendelian Inheritance in Man) and Orphanet are two of the largest public resources for rare diseases. OMIM is a continuously updated catalog of human genes and genetic phenotypes of all known Mendelian disorders . It is freely available and updated daily . Orphanet organizes diseases based on ORDO (Orphanet Rare Disease Ontology) .

OMIM (Online Mendelian Inheritance in Man)

- A comprehensive and authoritative compendium of human genes and genetic phenotypes .
- Freely available and updated daily .
- Contains information on all known Mendelian disorders and over 15,000 genes .
- Content is generated using published peer-reviewed biomedical literature .

Orphanet

- One of the largest public resources for rare diseases .
- Organizes diseases based on ORDO (Orphanet Rare Disease Ontology) .

Day 28:

- Review Week 4 material and practice with real-life datasets or case studies