Visualizing the Human Genome

W209 - Spring 2022 - Final Project Report

Project Team:

- Whitney Fee
- Angel Ortiz Nuñez
- Eric Ellestad

Project Deliverables:

- Project Website Link
- Code Repository Link
- Youtube Tutorial Link

Project Responsibilities

Whitney Fee:

- Exploratory Data Analysis
- Tableau Visualization - Genome Chromosome
- Tableau Visualization - Chromosome Overview
- CSS and Website Formatting
- Definitions
- User Study Script

Angel Ortiz Nuñez:

- Data Pre-Processing and Data Wrangling
- Altair Charts - Gene Chromosome
- Altair Charts - Protein Coding Composition by Chromosome & Genome
- Altair - Gene Function Charts
- Altair - Interactivity Across Charts
- Interactive Chart Instructions

Eric Ellestad:

- Data Collection and Dataset Integration

- Protein Viewer API Integration
- Altair Charts - Gene Expression Charts
- Bootstrap and Website Formatting
- Javascript Interactivity
- Flask Backend and Website Integration
- Heroku Platform Integration and Hosting

Project Recap

Changes we would have made if we had more time

If we had more time on the project we would have focused more on the design and UX of the webpage and continued to iterate with users to continuously improve our visualizations and optimally arrive at the tasks detailed. We would have loved to add interactive hovers with definitions vs the definitions page, but within the time allotted we weren't able to extend this Javascript functionality to the Altair charts. Additionally there are so many design and UX features of a webpage we would have loved to optimize, but building out our very first webpage proved a large undertaking. We were able to do Bootstrap and start working down the path of optimizing the webpage design, but not able to create something as visually appealing as a lot of the websites we see on the internet that are optimized for web user experiences.

Project Concept

The human genome consists of 23 chromosomes that include 20,000 protein-coding genes and over 6 billion base pairs in total DNA length. The relative location of genes is driven by their evolutionary origin and plays an important role in determining their function as well as their regulatory environment. Modern high-throughput genetic sequencing technologies and worldwide scientific collaboration have resulted in large scale publicly available databases of human genomic information.

Our project concept is to create an intuitive and interactive visualization tool of the human genome that shows the chromosomal location and subcomponent organization of each gene. Protein-coding genes will include additional visualizations regarding their RNA transcripts and information regarding the resulting post-translation protein such as its functional role in the body and a 3D visualization of the folded structure of the protein. The fully-zoomed out overview will reveal the organization and scale of the human genome while the zoomed-in details will reveal information about specific genes and their biological function.

Users

The primary category of users is biology students who have been introduced to genetics and the central dogma of molecular biology and are looking for an integrated visualization to fully synthesize the relationships between the processes of transcribing DNA into RNA, translating RNA into proteins, and the role of proteins in the human body.

Tasks

1. Identify the number of chromosomes in the human genome, their name, and relative size
2. Scroll across a specific chromosome to learn the location and geographic density of genes
3. Identify the amount of each chromosome that is a gene vs not a gene
4. Look up a specific gene and identify which chromosome it is located on and what other genes are nearest to it
5. Identify the subcomponents of a given gene such as number and location of exons, untranslated regions, coding DNA sequences, start codons, and stop codons
6. Identify the gene transcription components of a given gene and the relevant components of the gene through each step in gene expression
7. Identify the Amino Acid Sequence and resulting Protein that a given gene codes for and explore its function and 3D structure

Data

Primary Dataset: The Human Genome

- Link to dataset: GENCODE Human Genome version GRCh38.p13
- Source Organization: The GENCODE Project is run by the European Bioinformatics Institute (EBI) which is part of the European Molecular Biology Lab (EMBL)
- Size of Dataset: 3,241,002 rows × 25 columns

Secondary Dataset: The Human Proteome

- Link to downloadable dataset: UniProtKB Human Proteome
- Sample protein information including 3D structure: UniProt - Zinc Finger Protein 558
- Source Organization: Universal Protein Resource (UniProt) which is a collaboration between the European Bioinformatics Institute (EMBL-EBI), the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR).
- Size of Dataset: 20,213 rows × 8 columns

Tertiary Dataset: 3D Protein Structures

- Link to protein structure database: AlphaFold Protein Structure
- Source Organization: AlphaFold 2.0 by Deepmind

Additional Sources

Protein Viewer API from Molstar 3D Protein Visualizer
Definitions and Images from the National Human Genome Research Institute