Project Title - Advanced EDA for genomic data analysis: Identifying genetic variations through visualization

Team Members:

1. Name: Pallavi

   CAN ID Number: CAN_33443581

2. Name: Sandhya Manjunath Naik

   CAN ID Number: CAN_33452403

3. Name: Sneha M E

   CAN ID Number: CAN_33661118

4. Name: Thoufeeque

   CAN ID Number: CAN_33642456

Institution Name : P.A.College of Engineering Mangalore

Phase 1: Problem Definition and Data Understanding

1.1 Project Overview

The objective of this project is to develop a platform for advanced exploratory data analysis (EDA) in genomic data, focusing on identifying and visualizing genetic variations critical to understanding population diversity, disease susceptibility, and personalized medicine.

Traditional genomic data analysis often struggles with the high dimensionality and complexity of datasets, limiting meaningful insights. To address this, we leverage advanced visualization techniques and machine learning models, offering an intuitive interface for exploring genomic variations. The project empowers researchers, bioinformaticians, and healthcare professionals to

Visualize complex relationships between genetic variations, Identify patterns and clusters in genomic datasets., **support better decision-making in research and clinical settings.**

1.2 Objective of the Project

To create an interactive and user-friendly platform that enables detailed EDA of genomic data, focusing on identifying genetic variations and presenting them through effective visualization techniques.

Target Users:

Researchers in genomics and bioinformatics.

Healthcare professionals interested in personalized medicine.

Data scientists working on biological and genomic data.

Potential Applications:

- Identifying genetic markers associated with specific diseases.
- Exploring population-level genetic diversity.
- Supporting research in pharmacogenomics and personalized treatment strategies.
- Accelerating discovery in evolutionary and comparative genomics.

1.3 Dataset Overview and Data Requirements

To achieve the goal of advanced EDA and visualization for genomic variations, the dataset must include features that capture key genetic information, such as:

Single Nucleotide Polymorphisms (SNPs): Variations at a single nucleotide position in the genome.

Structural Variants: Insertions, deletions, duplications, and translocations in genomic sequences.

Genomic Annotations: Functional information like gene locations, coding regions, and regulatory elements

Overview of the data set:

Our project utilizes datasets from diverse sources, including the **Kaggle Human Genome Variation Data, the 1000 Genomes Project, and Ensembl GRCh38 Gene Annotations** These datasets contain information on genetic variations, such as SNPs, structural variants, and functional annotations, alongside metadata like population groups and gene locations. By integrating high-dimensional data in formats like CSV and VCF, these datasets enable comprehensive analysis of genomic variations, supporting visualization, clustering, and discovery of biologically meaningful patterns.

Dataset name and its format

- Dataset Name: Human Genome Variation Data
- Format: CSV/TSV

- Dataset Name: 1000 Genomes Phase 3 Variants
- Format: VCF (Variant Call Format)

- Dataset Name: GRCh38 Gene Annotations
- Format: GFF3/CSV

Explanation of all the features in the datasets with its importance:

Dataset Name: Human Genome Variation Data

Key Features:

Chromosome: The chromosome number where the variant is located.

Importance: Helps locate the exact genomic region.

Position: The base-pair position on the chromosome.

Importance: Critical for pinpointing specific variations.

Reference Allele: The original nucleotide.

Importance: Useful for comparison against the variant allele.

Variant Allele: The altered nucleotide.

Importance: Indicates the mutation/variation type.

Frequency: The frequency of this variant in the population.

Importance: Important for assessing rarity or commonality.

2. 1000 Genomes Project Dataset

Key Features:

ID: A unique identifier for each genetic variant.

Importance: Enables easy cross-referencing with other databases.

Genotype: Encodes the variation type at each position (e.g., homozygous/heterozygous).

Importance: Helps in analyzing zygosity patterns.

Population: The population group (e.g., EUR, AFR, EAS).

Importance: Useful for studying population-specific variations.

Functional Annotation: Indicates if the variation impacts coding, regulatory, or intronic regions.

Importance: Essential for understanding functional implications.

3. Ensembl: Genomic Annotations Dataset

Key Features:

Gene ID: Unique identifier for each gene.

Importance: Allows tracking of specific genes.

Gene Name: The name or symbol of the gene.

Importance: Adds interpretability to the analysis.

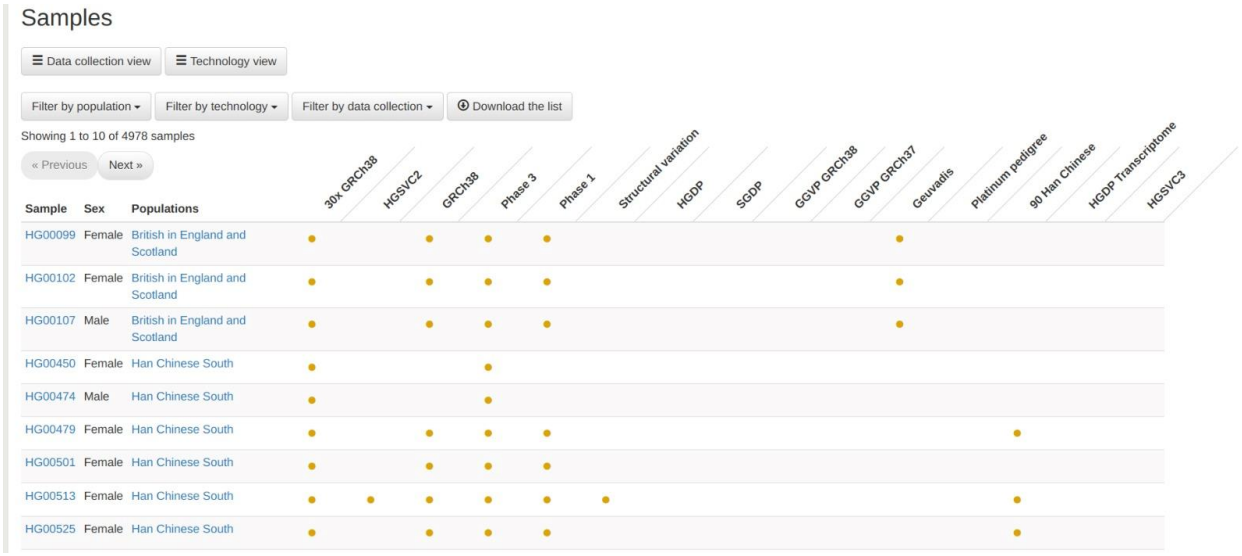
Feature Type: Indicates if the feature is an exon, intron, or regulatory region.

Importance: Provides context for the variation's location.

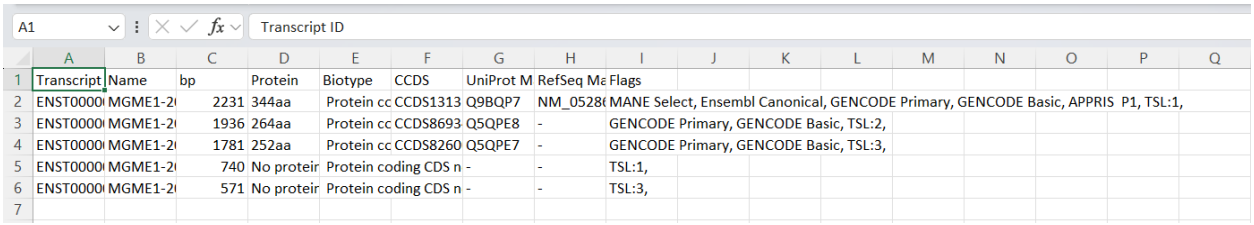
Start/End Position: The genomic coordinates of the feature.

Importance: Defines the span of the feature in the genome.

Screenshots of the datasets



	A	B	C	D	E	F	G	H	I	J	K
1	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	Ancestry
2	-10.90117144	0.798743345	-1.143300964	-1.070960009	11.85639581	-2.265965399	4.536404701	1.519959126	-2.214294195	-0.671273925	African
3	-9.990054297	1.416821349	-0.729626106	-0.443620754	10.41859419	0.44351371	2.640658887	-4.637745825	3.351628779	-0.671273925	African
4	-9.345388445	2.913053602	-0.921420784	0.029172626	10.67261467	-2.052552285	5.140475582	-1.451096013	0.444182864	-0.671273925	African
5	-11.2215074	1.733021061	-2.339816764	0.045786048	13.1950875	-3.068897297	2.863433874	-2.259194382	2.374566476	-0.671273925	African
6	-10.17515829	2.066307063	-0.785492616	-0.632400445	7.461272082	-1.643509092	0.715258231	-3.982751555	0.098680781	-0.671273925	African
7	-13.31992242	2.737273362	-1.532659031	0.058068544	-2.660751875	-0.550663024	2.141832367	-2.007177304	1.670643986	-0.671273925	African
8	-10.44169213	0.872937537	-0.100213953	0.519791326	0.094731956	0.764297776	-4.658872392	3.438304327	1.630509664	2.442466571	African
9	-9.218514509	1.758892098	-0.578552353	-0.937131202	1.373940452	-1.62012629	0.919812475	-1.998084061	1.592814334	-0.671273925	African
10	-12.673631	3.015143961	-2.100390149	-0.517233089	1.116705887	1.096178869	-3.240629907	2.39449918	0.251877547	0.885596323	African
11	-8.342017433	3.451908119	-1.668932156	-0.158435218	-1.710343582	0.813721929	-4.333145837	0.234331624	-0.549882983	0.885596323	African
12	-10.00354586	5.123690809	-0.941167859	-0.107404688	0.954424411	-3.093075065	-0.783159977	3.908338761	-0.586799913	0.885596323	African
13	-9.927064519	1.051759512	-3.225099329	-1.296031613	-3.038653364	4.477170342	0.789175627	-0.965553189	3.592118429	-0.671273925	African
14	-10.25617439	3.971784407	-2.110359934	-0.180311866	-0.622544214	1.981484299	1.543862414	1.963172052	-0.953001214	-0.671273925	African
15	-8.608074752	2.431557033	-1.526449966	0.361928942	1.653533923	-3.75090431	-0.400825647	-1.933118406	0.512476879	-0.671273925	African
16	-11.95794075	3.764007528	-1.823577195	-1.519007607	-0.68268394	2.596575651	-4.346790467	1.480340307	1.94792034	-0.671273925	African
17	-11.98361742	3.173192919	-0.842081561	-1.174821788	-1.287207348	1.556872884	-4.524589905	1.570644171	-1.819484073	-0.671273925	African
18	-9.421678436	1.922828125	-1.542939093	1.356596546	-1.738172369	-0.503768923	-2.896540581	0.12631027	-3.695712468	-0.671273925	African
19	-12.21129711	3.956428483	-2.88478883	-0.114107625	-0.872094538	6.004056581	-1.745879504	1.371193077	1.308170489	-0.671273925	African
20	-10.41113399	2.100081474	-0.757817407	0.740436579	0.048281438	-2.188839431	-2.490498267	0.159456526	-0.990683496	0.885596323	African
21	-11.28466346	0.043077027	1.021742955	-1.442500212	-3.031798226	-2.870877848	1.330861779	-1.639410679	1.865175872	-0.671273925	African



	A	B	C	D	E	F	G	H	I
1	Transcript	Name	bp	Protein	Biotype	CCDS	UniProt M	RefSeq Ma	Flags
2	ENST0000	MGME1-2	2231	344aa	Protein co	CCDS1313	Q9BQP7	NM_0528	MANE Select, Ensembl Canonical, GENCODE Primary, GENCODE Basic, APPRIS P1, TSL:1,
3	ENST0000	MGME1-2	1936	264aa	Protein co	CCDS8693	Q5QPE8	-	GENCODE Primary, GENCODE Basic, TSL:2,
4	ENST0000	MGME1-2	1781	252aa	Protein co	CCDS8260	Q5QPE7	-	GENCODE Primary, GENCODE Basic, TSL:3,
5	ENST0000	MGME1-2	740	No protein	Protein coding	CDS n	-	-	TSL:1,
6	ENST0000	MGME1-2	571	No protein	Protein coding	CDS n	-	-	TSL:3,
7									

1.4 Conclusion of Phase 1

In conclusion, our project provides a robust platform for advanced exploratory data analysis (EDA) of genomic variations, addressing the challenges posed by high-dimensional and complex genomic datasets. By integrating advanced visualization techniques and machine learning models, the platform enables researchers, bioinformaticians, and healthcare professionals to uncover meaningful patterns and insights from genomic data. This facilitates a deeper understanding of genetic variations, supporting applications in population diversity studies, disease research, and personalized medicine. With its intuitive interface and analytical capabilities, the platform empowers users to make informed decisions in research and clinical settings, bridging the gap between raw data and actionable insights.