**Visualizing the Complexity: Multi-Variate Analysis of Genetic Disorder**

## Introduction

Genetic disorders affect millions of people and represent a critical area of medical research. Understanding relationships between genetic mutations, inheritance patterns, and patient outcomes is essential for advancing diagnosis, treatment, and genetic counseling. However, the multidimensional nature of genomic data makes identifying patterns challenging through traditional methods. This project creates an interactive visualization system enabling researchers to explore genetic disorder data dynamically, examining quantitative variables like blood cell counts alongside qualitative factors such as disorder types and inheritance patterns to make complex genomic data more accessible and actionable.

## Related Work

Glusman, G., Caballero, J., Mauldin, D. E., Hood, L., & Roach, J. C. (2011). Kaviar: an accessible system for testing SNV novelty. Bioinformatics, 27(22), 3216-3217. https://doi.org/10.1093/bioinformatics/btr540

This paper presents a system for analyzing single nucleotide variants, also known as SNVs in genetic data, which is very closely related to our project as it addresses the challenge of managing and interpreting large-scale genomic datasets. The visualization approaches used for SNV analysis can inform how we present mutation data across different genetic disorder types in our interactive system.

Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., Felix, V., Jeng, L., Bearer, C., Lichenstein, R., Bisordi, K., Campion, N., Hyman, B., Kurland, D., Oates, C. P., Kibbey, S., Sreekumar, P., Le, C., Giglio, M., & Greene, C. (2019). Human Disease Ontology 2018 update: Classification, content and workflow expansion. *Nucleic Acids Research, 47*(D1), D955-D962. https://doi.org/10.1093/nar/gky1032

This work on disease classification and ontology provides a framework for understanding how genetic disorders can be categorized and related to one another. The classification system described in this paper will inform our visualization design, particularly in how we organize and allow users to filter between different disorder types and subclasses.

## Related Visualization Examples

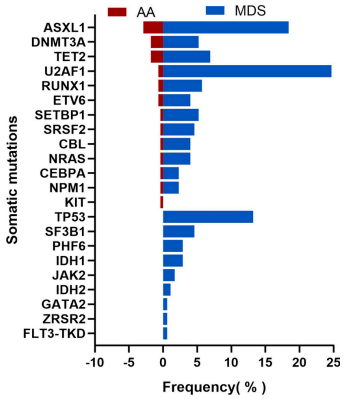


**Figure 1: Comparative bar chart of genetic mutation frequencies**

Reference: Bar chart from ResearchGate (https://www.researchgate.net/figure/Bar-chart-showing-the-frequency-of-mutated-genes-and-type-of-mutations-in-each-gene_fig2_370203557) demonstrating grouped categorical comparison methods for genetic data analysis.

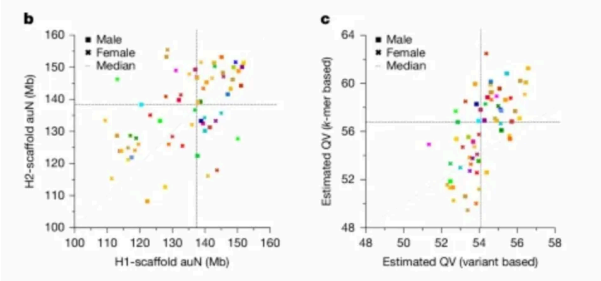


**Figure 2: Scatter plot visualization of genetic variant data**
Reference: Scatter plot from Nature (https://www.nature.com/articles/s41586-025-09140-6) showing how color-coded quantitative genetic measurements can reveal clustering patterns across patient groups.

**Data Description**
**Data Source**: The dataset we're using for this project was obtained from Kaggle (AI Buzz Genetic Disorders Dataset), specifically the training dataset file train_genetic_disorders.csv.

**Data Size:** The dataset contains approximately 22,000 patient records with 45 features that consist of demographic information, clinical measurements, genetic disorder classifications, and patient outcomes.

**Key Features**

| Feature Category | Variables | Purpose |
|---|---|---|
| Quantitative | Patient age, parental ages, blood cell counts, white blood cell counts, previous abortions | Histograms, scatter plots, distribution analysis |
| Categorical | Genetic disorder type (3 types), disorder subclass (9 types), gender, status, inheritance patterns, birth defects, blood test results, folic acid details | Grouping, filtering, comparative analysis |

**Analysis Plan and Visualizations**
**Data Processing**
The first step in processing is to remove missing values. After removing missing values, the dataset has 5000 rows which is within project requirements. Also, some columns have been anonymized because of privacy concerns to the patients, for example, columns (test 1 - 5, symptoms 1-5) where each of these columns has a dummy variable 1 or 0 indicating if they tested positive or negative. Without the names of the tests and symptoms, it is hard to make sense of the data in those columns. However, we propose to mitigate this shortfall by creating a new variable called severity of symptoms, calculated by counting the number of symptoms they are positive for. This variable can provide new insight for our webpage.

**Analysis Tasks**
**Task 1: How do clinical measurements vary across genetic disorder types?**
We want to understand whether quantitative indicators like blood cell count and white blood cell count show different distributions depending on the disorder type. This can help identify whether certain biomarkers are associated with specific categories of genetic disorders.

**Task 2: How do inheritance patterns connect to disorder outcomes?**
The dataset includes variables describing whether a disorder was inherited from the maternal or paternal side. We want to trace the flow from these inheritance patterns to the resulting disorder type, revealing whether certain pathways are more strongly associated with specific disorders.

**Task 3: How do demographic and environmental risk factors relate to disorder type and severity?**
Risk factors such as history of substance abuse and serious maternal illness may influence both the type and severity of genetic disorders. We want to explore how these factors interact with disorder classifications and clinical measurements like blood cell count.

**Planned Visualizations**

1. Scatter Plot (Static & Interactive) - Blood cell count vs White blood cell count
    - Color-coded by disorder type to address Task 1, revealing whether disorders cluster in specific biomarker ranges and helping identify diagnostic patterns and potential biomarkers for specific genetic disorders.
2. Grouped Bar Chart (Static & Interactive) - Gender, Inheritance Source, and Disorder Type
    - Compares disorder counts across gender and inheritance patterns (maternal vs paternal) to address Task 2, identifying gender-specific or inheritance-specific disorder patterns.
3. Sankey Diagram (Interactive) - Inheritance Flow to Disorder Type
    - Traces pathways from inheritance source to disorder type, with flow width representing patient counts. Addresses Task 2 by visualizing inheritance-to-outcome connections. The dropdown menu allows exploration of different categorical flows to identify dominant pathways.
4. Grouped Box Plot (Interactive) - Risk Factors vs Blood Cell Count
    - Compares blood count distributions across parental substance abuse and maternal illness categories to address Task 3. Dropdown toggles between blood cell count and white blood cell count, revealing how risk factors relate to clinical measurements and potential disorder severity.
5. Tree Diagram (Interactive) - Hierarchical Inheritance and Disorder Classification
    - Hierarchical breakdown from inheritance side to source to disorder type, with node sizes showing patient counts. Addresses Tasks 2 and 3 by displaying the complete inheritance-to-disorder landscape and identifying dominant versus rare pathways.

**Team Member Responsibilities**

**Dalia:**

- Data collection and data exploration
- Planned Visualizations: Plot 1 (Scatter plot) and Plot 2 (Grouped Bar Chart)
- Website design and layout for plots 1 and 2
- Introduction and related work sections of final report

**Yuqi:**

- Data Processing and feature engineering
- Planned visualizations: Plot 3 (Sankey Diagram) and Plot 4 (Grouped Box Plot)
- Website implementation for plots 3 and 4
- Data description and analysis plan sections of final report

**Jeff:**

- Data validation and quality assurance
- Planned visualization: Plot 5 (Tree diagram)
- Website integration and overall interactivity coordination
- Team member responsibilities coordination and final presentation preparation

**All Members:**

- Collaborative design decisions, user testing, debugging, and documentation