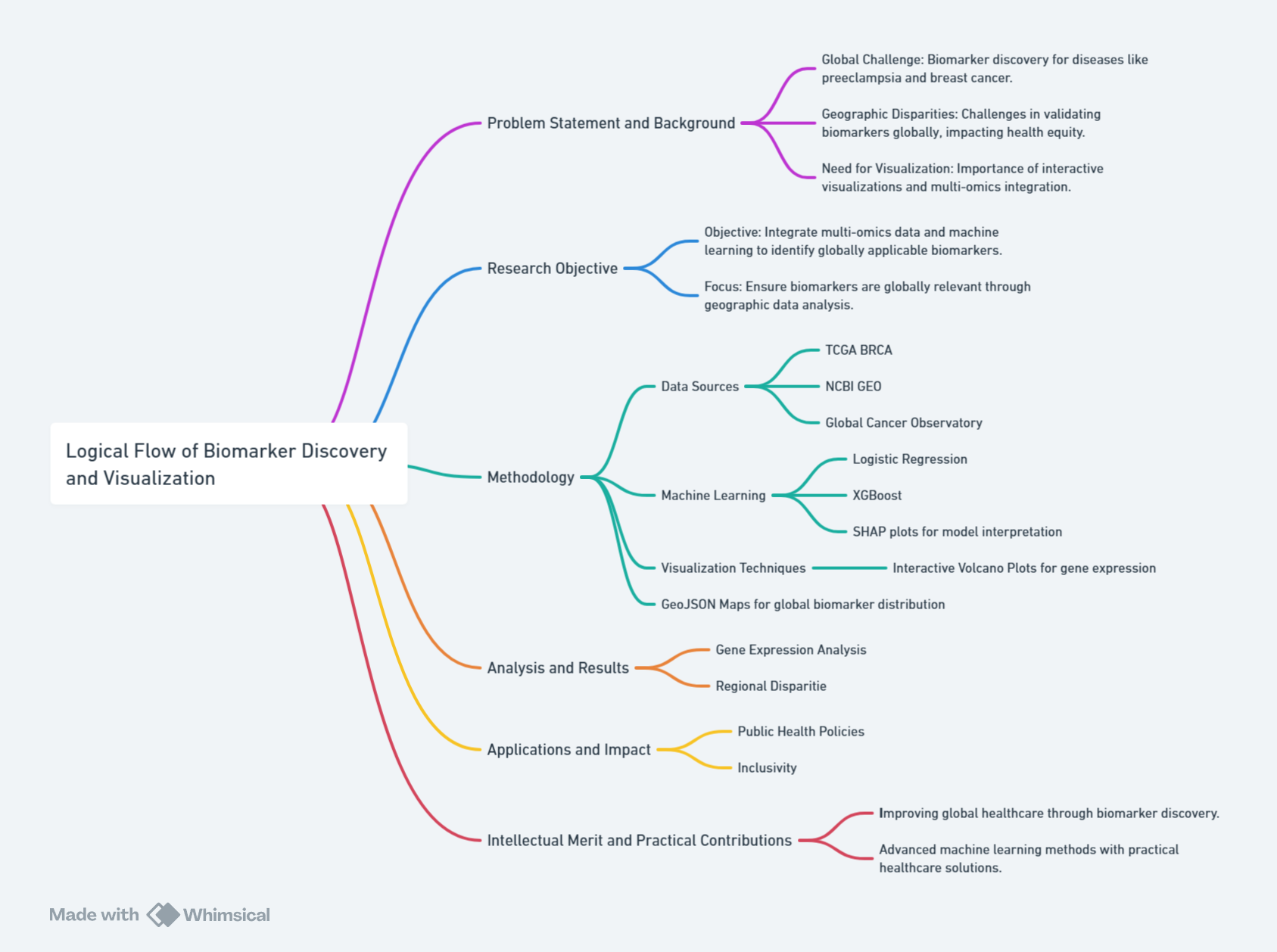
**Uncovering Geographic Variability in Breast Cancer Biomarkers: Integrating Machine Learning and Interactive Visualizations for Global Insights**

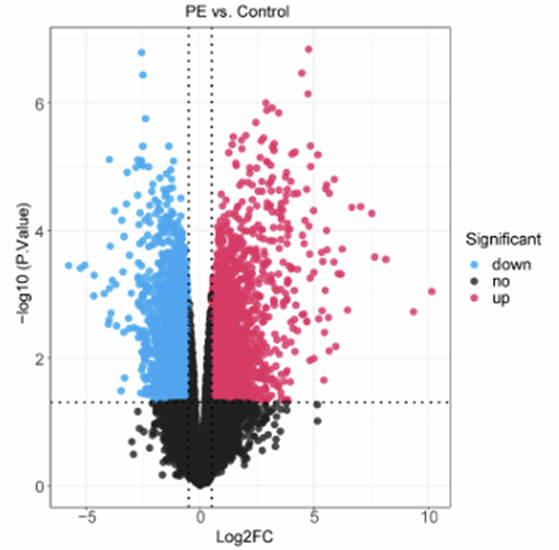
Yian Pei, Munkh-Orshikh Munkhbold, Lauris Vo



**Figure 1**: Research Methodology Flowchart (Created by Whimsical)

1. **Background and Motivation**

Breast cancer is one of the most common cancers diagnosed worldwide, making early detection and effective treatment essential for better patient outcomes. To achieve these goals, biomarkers—such as specific genes, proteins, or other molecules—play a critical role. These biological indicators are linked to cancer development and progression and provide valuable insights into the unique characteristics of each patient’s tumor. By revealing tumor characteristics, predicting therapy responses, and monitoring disease progression, it helps detect diseases, predict outcomes, and guide personalized treatments. In this context, bioinformatics plays a vital role in studying and visualizing these biomarkers, especially given the complexity and size of genomic data. Effective visualizations make complicated patterns in biomarker data easier to understand for researchers, doctors, and non-specialists. These insights are pivotal for advancing personalized medicine in breast cancer, where tailored strategies can significantly enhance survival rates. In this project, we will focus on implementing bioinformatics in breast cancer analysis, and the overall structure is shown in **Figure 1**.



**Figure 2**: Volcano Plot of Differential Gene Expression in Preeclampsia vs. Control Cohorts (Lu et al., 2024)

The research is inspired by the study “Exploring the ceRNA network involving AGAP2-AS1 as a novel biomarker for preeclampsia.” **Figure 2** shows a volcano plot comparing gene expression in preeclampsia (PE) patients and control groups. Significantly upregulated genes (those with increased expression in PE) are highlighted in red, while downregulated genes (those with reduced expression in PE) are shown in blue. The study lacks good visual aids, which makes it hard for non-specialists to understand the findings. This points out the challenge of making complex bioinformatics data clear for people outside the research field.

Moreover, we found a more crucial problem that is common in many other biomarker discovery papers: people often overlook the geographic origin of the data. However, studies have shown that genes and their expression can vary significantly between populations, affecting disease susceptibility and the effectiveness of biomarkers. For instance, research has demonstrated notable genetic differences between populations of European and African ancestry in the context of breast cancer. A study by Dumitrescu and Cotarla (2005) highlighted that African American women are more likely to develop aggressive, triple-negative breast cancer subtypes compared to women of European descent, due in part to differences in genetic risk factors such as BRCA1 and BRCA2 mutations and the frequency of specific alleles associated with cancer susceptibility. These findings emphasize the importance of considering population-specific genetic variation in biomarker discovery and disease prediction models, as biomarkers validated in one population may perform poorly in others if genetic and environmental contexts are not accounted for.

In this paper (Lu et al., 2024 ) the data comes from a specific geographic area, which raises questions about whether the biomarkers can be applied globally. Second, Based on that, our project aims to explore the applicability of identified biomarkers in a global context by integrating datasets from diverse populations and analyzing regional differences in biomarker efficacy. It will develop innovative visualization techniques to make findings more accessible and engaging, bridging the gap between research and real-world applications. The project seeks to enhance existing visualizations, providing a broader perspective on biomarker data for breast cancer diagnosis and treatment, ultimately contributing to global health equity.

1. **Research Question:**

1. How do biomarkers for breast cancer vary across different geographic regions?

2. What regional patterns, trends, and anomalies in biomarkers can be identified through integrated datasets, and how can interactive data visualization combined with machine learning provide cross-disciplinary insights for global biomarker discovery?

1. **Application Scenarios:**

The findings from this research on regional variability in breast cancer biomarkers can have significant real-world applications across healthcare, public policy, and global research collaboration. By understanding how biomarkers differ across populations, healthcare providers can develop personalized treatment plans tailored to account for regional genetic variations, ultimately improving patient outcomes worldwide. For instance, healthcare providers in regions with a higher prevalence of certain genetic markers could use this information to adjust medication dosages or recommend specific treatments, ensuring more effective care for local populations. This research also has the potential to inform public health policies by enabling more targeted interventions and resource allocation for cancer care in different regions. For example, insights into regional biomarker variability could help policymakers prioritize screening programs and prevention strategies in areas where certain genetic predispositions are more common. Additionally, industries such as biotechnology, pharmaceutical, and AI-based healthcare technologies could benefit from the findings by developing more effective, globally applicable treatments and diagnostic tools. Pharmaceutical companies, for instance, could use these insights to create location-specific drugs, while AI algorithms can integrate region-specific data to improve early diagnosis and personalized treatment plans. Ultimately, this research aims to enhance the accessibility and effectiveness of breast cancer care, contributing to global health equity and advancing the field of personalized medicine.

This project can solve practical problems in both personalized medicine and global health strategies by visualizing regional differences in biomarkers. In personalized medicine, understanding region-specific gene expressions allows for the refinement of therapies, increasing their efficacy for local populations. For example, in regions with higher rates of specific genetic mutations, treatment plans could be adjusted to include therapies that target those mutations more effectively. Additionally, these insights can guide improvements in biomarker-driven diagnostics, ensuring more accurate results across different regions. On a larger scale, highlighting differences in biomarkers across populations can inform global health policies. By identifying genetic predispositions in various regions, countries and organizations such as the WHO can prioritize cancer screening programs or implement tailored prevention strategies, ensuring that health interventions are more focused and effective.

Addressing social needs, this project emphasizes health equity and public awareness. Visualizing disparities in biomarker distribution underscores the need for inclusive genetic research, advocating for better representation of underrepresented populations in genomic datasets. For example, ensuring that populations in low-income regions or developing countries are represented could lead to more equitable healthcare interventions. Moreover, interactive tools like the improved volcano plot make complex genomic data more accessible to the public, fostering a greater understanding of genetic research and empowering individuals to take charge of their health decisions. These tools can also help community health organizations and local governments in underserved areas make data-driven decisions, even without specialized knowledge, improving public health outreach and resource allocation.

The project also links to various disciplines and industries, with significant implications for medical genomics, data visualization, and healthcare policy development. In medical genomics, the identification of biomarkers is essential for precision medicine, especially in oncology, as seen in breast cancer research using TCGA datasets. Pharmaceutical companies can use these findings to develop treatments targeting genetic markers prevalent in specific regions. In data visualization and bioinformatics, industries specializing in genomic analytics can integrate interactive visualizations into their platforms, offering dynamic tools for researchers to explore multi-regional datasets. Insights into regional biomarker variability can also assist global organizations like the WHO in crafting region-specific healthcare strategies, ensuring that health interventions align with the genetic makeup of different populations. This research benefits several sectors: the research community by encouraging the use of advanced interactive visualizations, healthcare providers by enabling more accurate and effective diagnostic and treatment plans for patients from diverse genetic backgrounds, and the tech industry by demonstrating how advanced tools, such as Plotly, can improve the interpretation and application of genomic data.

In summary, this research has the potential to bridge the gap between genomic data and real-world applications, addressing practical problems in personalized medicine and global health, while contributing to social needs such as health equity and public awareness. The insights gained from understanding regional genetic differences can revolutionize healthcare strategies, benefiting a range of industries, policymakers, and individuals alike.

1. **Methodology**

**4.1. Visualization Techniques**

The methodology employed in this study integrates advanced visualization techniques, diverse data sources, and foundational concepts from visualization principles to effectively analyze and present complex biomarker data. The visualizations draw heavily on concepts from Chapters 1–10 of the *Visualization Basics* repository (Munzner 2014), emphasizing marks and channels, the arrangement of spatial data, and data abstraction techniques to create meaningful representations of biomarker variability across regions.

Marks and channels were strategically chosen to ensure clarity and enhance interpretability across visualizations. For example, in the volcano plot, points are used as marks to represent individual genes, with their position encoded on quantitative axes to convey log-fold changes and statistical significance. As discussed in Chapter 2, **position** is one of the most effective channels for quantitative comparison, maximizing perceptual accuracy (Munzner 2014). Additionally, color intensity serves as an extra channel to highlight genes appearing in multiple datasets, drawing on Chapter 6’s guidance on using **color** for preattentive processing and emphasizing critical data points (Munzner 2014). This approach ensures that both primary and secondary information is effectively communicated to the user, leveraging multiple channels to enhance understanding.

Incorporating spatial data into visualizations, such as the GeoJSON map for country-specific biomarkers, required thoughtful application of **spatial data arrangement** principles. Geographic regions were encoded as areas with color intensity mapped to age-standardized rates (ASR) of breast cancer, following Chapter 8's emphasis on spatial arrangement and the use of **position and area** to convey quantitative data effectively (Munzner 2014). The design also included interactivity, enabling users to click on a country to reveal detailed information about associated biomarkers and their statistical relevance. These features align with the task abstraction framework discussed in Chapter 7, which emphasizes supporting **exploration**, **filtering**, and **comparison tasks** to provide users with actionable insights (Munzner 2014).

Data and task abstraction played a central role in the visualization design, ensuring that each tool supports specific analytical needs. The SHAP beeswarm plot was tailored to highlight feature importance in machine learning models, with SHAP values encoded on the x-axis and ranked features along the y-axis. This design reflects Chapter 5’s guidance on **arranging data in hierarchical or ranked orders** to enhance clarity and focus attention on important features (Munzner 2014). Color was employed as an auxiliary channel to encode gene-specific attributes, following Chapter 6’s principle of using **color encoding to add dimensions without cluttering visualizations** (Munzner 2014). These abstractions balance the need for high-level summarization with the capacity for detailed examination, ensuring that users can efficiently transition between global patterns and granular insights.

Diverse datasets were integrated to enrich the visualizations, ensuring that each representation captures a broad and accurate view of biomarker variability. The TCGA BRCA dataset provided robust gene expression data, while NCBI GEO contributed additional depth with varied experimental conditions and biological contexts. The Global Cancer Observatory dataset added a crucial spatial dimension, providing ASR data to contextualize biomarkers within geographic and epidemiological trends. This integration aligns with Chapter 4's recommendation on **data abstraction**, particularly the combination of heterogeneous datasets to achieve a richer representation of patterns and trends (Munzner 2014). Combining these datasets enabled insights that would not be achievable with a single source, such as identifying population-specific biomarkers and linking them to regional disease patterns.

Color theory and design choices were carefully considered to enhance the clarity and accessibility of the visualizations. As outlined in Chapter 10 of Munzner’s Visualization Analysis and Design (2014), color serves as a critical tool in both differentiating categories and emphasizing key data patterns. In the context of biomarker visualization, color was employed not only for aesthetic appeal but also to improve interpretability and facilitate user engagement. For example, in the volcano plot, color intensity was used to highlight genes that appeared in multiple datasets, making it easier to distinguish between genes of high and low relevance. The color scale was designed to ensure perceptual uniformity and avoid misleading interpretations, using a color palette that is distinguishable for individuals with color vision deficiencies. Similarly, in the GeoJSON map, color gradients were applied to represent varying levels of age-standardized breast cancer rates (ASR), with darker hues corresponding to higher rates, thus allowing users to quickly identify geographical regions with the highest incidence. This strategic use of color supports not only quantitative comparisons but also strengthens the viewer’s ability to recognize patterns and outliers in the data. By adhering to principles of effective color encoding, the visualizations ensure that color enhances, rather than distracts from, the information being communicated, aligning with Munzner's emphasis on clear and meaningful design choices.

### **4.2 Data Sources and Integration**

#### **Original Paper**

* **Source**: [Nature Scientific Reports (DOI: 10.1038/s41598-024-79224-2)](https://www.nature.com/articles/s41598-024-79224-2).
* **Description:** The study focuses on a molecule in the body called AGAP2-AS1, which is part of a system where different molecules interact and influence how cells work. This system is like a conversation between molecules, helping the body regulate important processes. Scientists think AGAP2-AS1 could be a clue (or biomarker) to identify preeclampsia—a serious pregnancy condition that can affect the mother and baby. By studying how AGAP2-AS1 interacts with other molecules, researchers hope to understand preeclampsia better and find ways to detect or treat it earlier.
* **Purpose:** To improve understanding of preeclampsia’s molecular mechanisms and advance early detection and therapeutic strategies.

#### **Original Dataset:**

* **Source:** [KEGG: Kyoto Encyclopedia of Genes and Genomes](https://www.genome.jp/kegg/)
* **Description:** The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a comprehensive database that provides detailed information on biological pathways, diseases, drugs, and genomes (Kanehisa et al. 2004). KEGG’s pathway database is particularly valuable for understanding molecular interactions and cellular processes, as it categorizes genes based on their involvement in specific biochemical pathways. This database integrates data from various organisms and offers a functional map of molecular pathways that are crucial for understanding cellular functions, disease mechanisms, and drug responses. KEGG is widely used in pathway enrichment analysis, helping researchers identify significant biological pathways associated with particular genes or phenotypes.
* **Purpose:** KEGG is a crucial step in our gene selection process, enabling the identification of genes involved in specific metabolic or signaling pathways. By integrating KEGG pathway annotations, we can explore the functional roles of genes, assess their relevance in disease mechanisms, and guide the selection of candidate genes for further investigation. This makes KEGG a powerful tool in pathway discovery and functional annotation, essential for understanding the biological context of our gene sets.

#### **New Dataset 1:**

* **Source**: [TCGA BRCA dataset (via UCSC Xena Browser)](https://xenabrowser.net/datapages/?dataset=TCGA.BRCA.sampleMap%2FHiSeqV2&host=https%3A%2F%2Ftcga.xenahubs.net&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443)
* **Description**: Global dataset representing gene expression in breast cancer, covering diverse populations.
* **Purpose**: Used as part of selection during the Machine Learning process which is described in **Section 4.3.2** (Advanced Tools - Machine Learning).

#### **New Dataset 2:**

* **Source:** [NCBI GEO (Gene Expression Omnibus)](https://www.ncbi.nlm.nih.gov/geo/)
* **Description:** NCBI GEO is a public repository that hosts a wide range of gene expression and genomics datasets, covering various diseases and biological conditions, including breast cancer. It allows researchers to explore and analyze gene expression profiles and functional genomics data.
* **Purpose:** To identify and utilize datasets specifically related to breast cancer for research, enabling the discovery of biomarkers, molecular mechanisms, and potential therapeutic targets.

#### **New Dataset 3:**

* **Source:** [Global Cancer Observatory (GCO) – IARC](https://gco.iarc.who.int/en)
* **Description:** The dataset provides age-standardized rates (ASR) for breast cancer through countries, which is a way to compare cancer incidence across countries by accounting for differences in age distribution. It allows researchers to understand how breast cancer rates vary globally.
* **Purpose:** To create a heat map of the world, visualizing breast cancer incidence rates across different countries to identify geographic patterns and disparities.

All datasets are reliable, peer-reviewed, and widely used in genomic research.

Integrating datasets from diverse sources, including the **TCGA BRCA dataset**, **NCBI GEO**, and the **Global Cancer Observatory (GCO)**, offers a comprehensive approach to uncovering insights into breast cancer biomarkers that cannot be achieved with a single dataset. Each dataset contributes unique and complementary information, enabling a cross-disciplinary analysis that combines molecular and epidemiological perspectives.

The **TCGA BRCA dataset** provides high-resolution gene expression data, capturing molecular changes in breast cancer across diverse populations. This dataset is crucial for identifying candidate biomarkers and understanding the genetic underpinnings of the disease. Complementing this, the **NCBI GEO** dataset adds breadth by including a wide range of experimental conditions and biological contexts. By integrating these datasets, duplicated genes or overlapping findings can be identified, allowing the focus to shift to high-confidence biomarkers through machine learning-based feature selection, as described in Section 4.3.2.

The inclusion of the **GCO dataset**, which provides age-standardized rates (ASR) of breast cancer across countries, adds a critical epidemiological dimension to the analysis. By mapping ASR data to the refined biomarker candidates, the geographic distribution of breast cancer incidence can be visualized, revealing regional disparities and population-specific trends. This step situates molecular findings within a global context, enabling the identification of biomarkers that may have distinct relevance across different populations.

Combining these datasets creates a pipeline that integrates molecular precision with global patterns, addressing key questions about biomarker diversity across regions. For example, it allows for the investigation of whether genetic markers align with regional trends in breast cancer incidence or whether population-specific biomarkers correlate with aggressive disease subtypes. Without the molecular insights provided by TCGA BRCA and NCBI GEO, the analysis would lack depth; without the geographic context of the GCO dataset, the findings would miss critical public health implications.

This integrated approach bridges genomics and public health, revealing patterns and trends that inform both scientific discovery and clinical application. By connecting molecular data to global health challenges, the analysis not only strengthens the validity of biomarker selection but also provides actionable insights for targeted therapies and regional cancer prevention strategies.

### **4.3. Advanced Tools**

**4.3.1. Interactive Visualization Tools**:

The volcano plot in our project addresses several key limitations of traditional static plots by integrating interactivity and enhanced data representation. Standard volcano plots often suffer from low data density and lack interactivity, providing limited information about individual genes. To overcome these limitations, we have redesigned the volcano plot to serve as a dynamic tool for gene selection, making it more informative and user-friendly.

One of the key improvements is the interactivity of the enhanced volcano plot. Users can click on individual genes to access detailed information, such as its biological function, associated pathways, and its frequency across datasets. This feature eliminates the need for researchers to manually search through external databases, streamlining the analysis process. Unlike traditional volcano plots, which may only display data from a single dataset, our plot consolidates information from multiple datasets, ensuring a more robust representation of gene significance and fold-change. This allows researchers to assess candidate genes more comprehensively.

Additionally, we have incorporated a novel color scheme that highlights genes based on their frequency across datasets. Genes that appear in multiple datasets are displayed in brighter colors, making it easier for users to visually identify consistently observed biomarkers. This visual cue adds another layer of information to the plot, helping researchers focus on more reliable gene candidates. By integrating more data and providing interactive functionality, the enhanced volcano plot enables researchers to not only identify potential biomarkers but also gain a deeper understanding of the underlying significance of the highlighted genes. This approach addresses the static nature of traditional plots, which often lack sufficient context for interpretation.

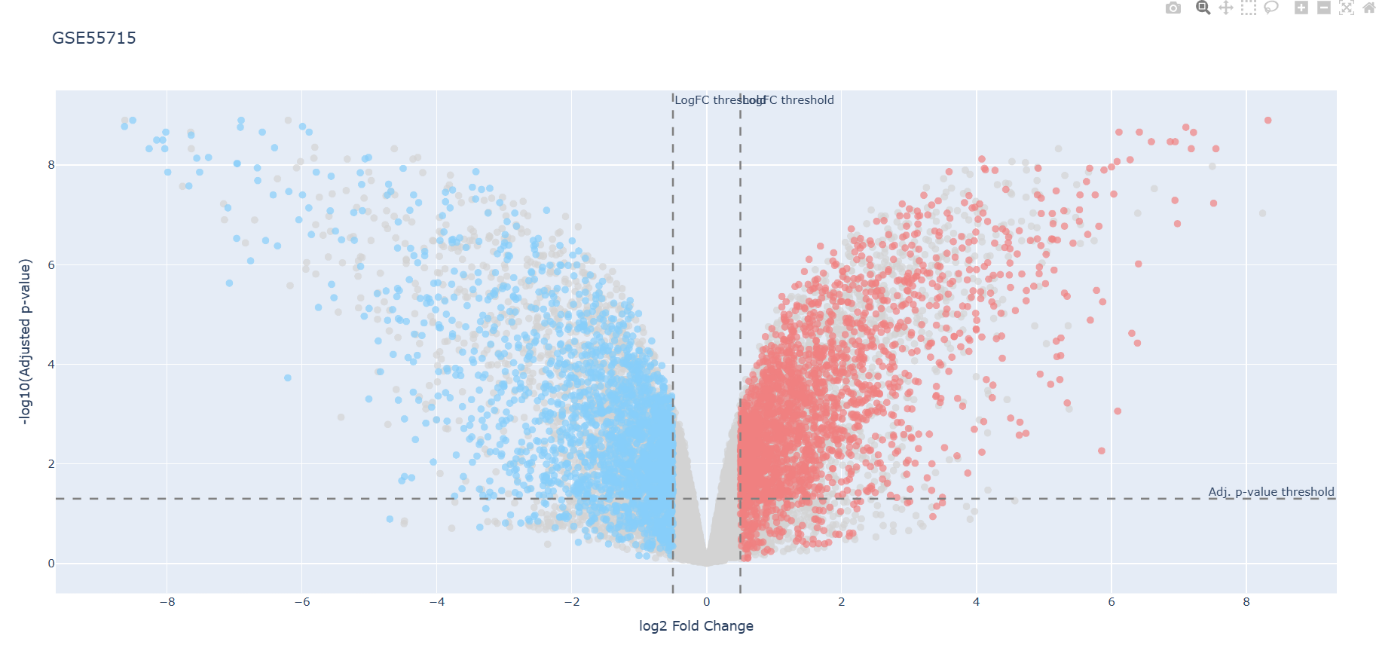


Figure 3: **Interactive Volcano Plot for GSE55715**

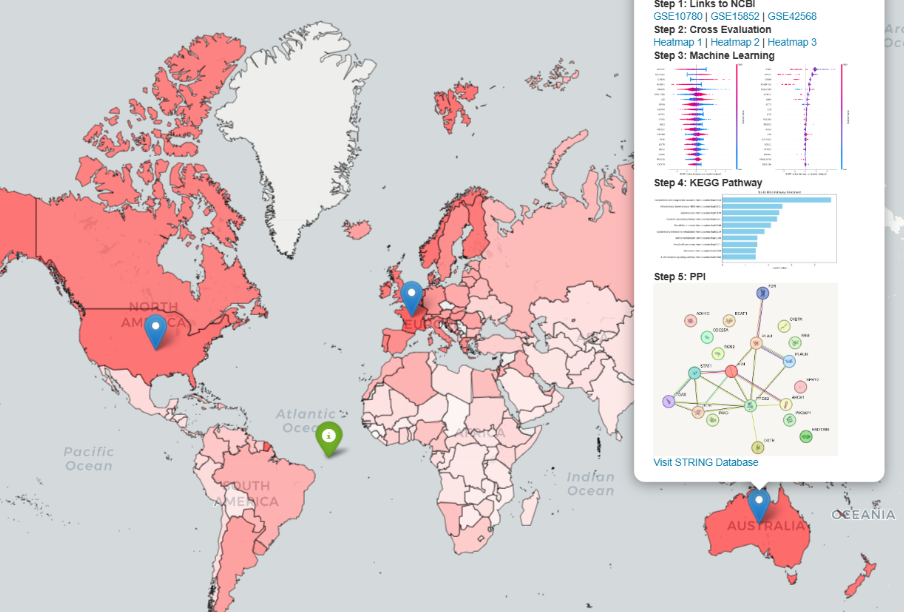
**Figure 3** shows gene expression differences in the GSE55715 dataset. Each point represents a gene, with **red** for upregulated (genes with increased expression) and **blue** for downregulated genes (genes with decreased expression). The x-axis shows log2 fold change, and the y-axis shows -log10 adjusted p-value, with hover-over details for each gene, which would be displayed by clicking on a specific point.

To complement the enhanced volcano plots, we have developed an interactive GeoJSON-based map to visualize country-specific biomarkers, offering a global perspective on the distribution of key genes associated with breast cancer. This map integrates geographic and genomic data to highlight regional trends and disparities. The GeoJSON format provides a flexible and dynamic representation of spatial data, making it ideal for this application.

The map combines biomarkers identified through machine learning and gene expression analysis with global data on breast cancer incidence. Each country is color-coded based on its age-standardized rate (ASR) for breast cancer, allowing users to directly compare disease prevalence with the occurrence of key biomarkers. Users can click on a country to access detailed information about the top biomarkers found in that region. For instance, clicking on “Australia” reveals significant genes and also the flow of our data selecting process.

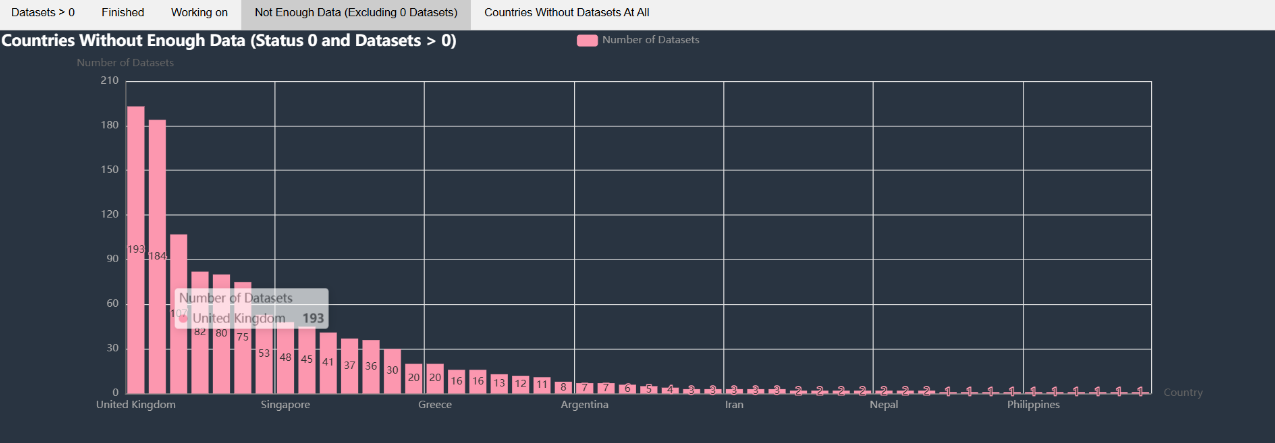
The ASR for each country is represented using a heatmap color scale, with darker shades of red indicating higher rates. This dynamic color encoding helps users easily identify regions with a higher burden of breast cancer and explore the biomarkers that may contribute to these trends. Additionally, the brightness of marker annotations reflects the frequency of a gene across multiple datasets, ensuring that the visualization emphasizes biomarkers with broader relevance while also focusing on regional specificity.

By integrating genomic and epidemiological data, the map bridges the fields of public health and molecular biology. This cross-disciplinary approach offers a more comprehensive understanding of how biomarkers and disease burden vary globally.



**Figure 4:** Geographic Visualization of Dataset Origins

**Figure 4** displays the origins of the datasets. Markers highlight regions like China, Japan, USA. The map reveals geographic gaps in dataset representation, emphasizing the need for diverse genomic studies. When a user clicks on a specific region, an annotation appears, offering detailed information about the genes identified from datasets originating in that area. These annotations include concise summaries of key gene markers, their roles in biological pathways, and potential implications in breast cancer research. Additionally, each region shows our data filtering progress, and is linked to results of every step in our work. We also attach the STRING interactive platform, a powerful tool for exploring protein-protein interaction networks. This integration allows users to delve deeper into the functional relationships of genes within the dataset.

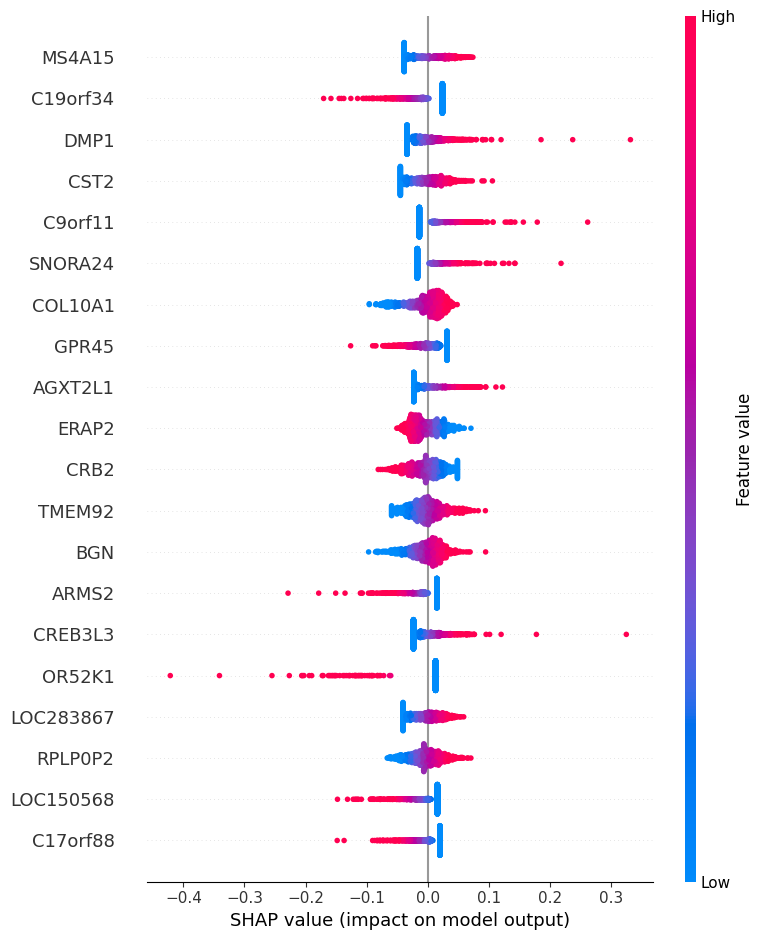


**Figure 5:** Number of Dataset Found per Country and their Current Status

Additionally, we conducted a statistical analysis of the number of datasets available in the NCBI database using the keywords "breast cancer," "GEO2R," and "country name" (e.g., "China"), specifically focusing on datasets for *Homo sapiens*. We assessed the usability of these datasets and found that only 5 countries had datasets that could be used, while 4 countries had datasets that could potentially be used. Many countries, however, did not have any datasets available at all. **Figure 5** highlights the countries with datasets that are not usable, underscoring the challenges researchers face when attempting to analyze data by country. This emphasizes the need for more research and greater data sharing to improve global analysis and understanding.

**4.3.2. Machine Learning:**

In this study, we employed multiple machine learning models—Logistic Regression and XGBoost—to address the challenges of gene selection in high-dimensional datasets, leveraging the complementary strengths of these methods. Logistic Regression, a regression-based model, assumes a linear relationship between gene expression levels and the target variable, assigning coefficients to genes that reflect their influence on outcomes. Its simplicity and interpretability make it particularly valuable for identifying direct, independent gene effects, offering insights that are both biologically meaningful and accessible to clinicians. XGBoost, a modern tree-based model, enhances our methodology by capturing complex, non-linear relationships and gene interactions through an iterative process of building decision trees. Unlike traditional methods, XGBoost incorporates regularization techniques to mitigate overfitting, ensuring robust performance in noisy, high-dimensional datasets. Additionally, its computational efficiency allows it to handle large datasets effectively while identifying subtle patterns that simpler models might overlook. By combining these models, we capitalize on their unique capabilities—Logistic Regression for linear effects and XGBoost for complex patterns—ensuring a comprehensive and reliable approach to gene selection that facilitates accurate and biologically relevant discoveries.  
  
 We also implemented SHAP (SHapley Additive exPlanations) summary plots generated from a machine learning model used for binary classification of cancer and normal tissue in TCGA breast cancer data, where 0 represents normal tissue and 1 represents cancer. The plot visually explains how each gene (feature) contributes to the model’s predictions. The genes are listed in order of importance on the vertical axis, with the most influential genes (e.g., "MS4A15") at the top. Each dot on the horizontal axis represents a SHAP value, which quantifies the impact of a gene on moving the prediction toward cancer or normal. Positive SHAP values push the prediction toward cancer (1), while negative SHAP values push it toward normal (0). The colors of the dots indicate the gene expression level: red represents high expression, and blue represents low expression. For example, if a gene's high expression (red) is associated with positive SHAP values, it means that the high expression of that gene strongly influences the model to classify a sample as cancer. Conversely, if blue (low expression) is associated with negative SHAP values, it indicates that low expression contributes to a normal classification. This plot provides a clear and interpretable way to identify which genes are most critical for distinguishing between cancerous and normal tissues, allowing researchers to explore potential biomarkers or drivers of breast cancer.



**Figure 6:** SHAP Summary Plot for Logistic Regression Model

**Figure 6** shows the impact of gene features on the logistic regression model’s predictions. Each row represents a gene, and the x-axis indicates SHAP values, reflecting the magnitude and direction of each gene's contribution. Red dots represent high feature values, and blue dots indicate low values. Key genes like **MS4A15** and **COL10A1** are the most influential.

### 

### **Figure 7: SHAP Summary Plot for XGBoost Model**

### **Figure 7** shows the impact of gene features on the predictions of the XGBoost model. Each row represents a gene, and the x-axis indicates SHAP values, reflecting the strength and direction of each gene's contribution. **Red dots** indicate high feature values and **blue dots** indicate low values. Key genes such as **SPRY2** and **COL10A1** exhibit significant influence on the model’s predictions.

### 5. **Results:**

The visualizations generated in this study provide a range of insights into the data. The volcano plot (**Figure 3**) highlights significant genes that are either upregulated (red) or downregulated (blue) in preeclampsia compared to controls, helping to identify key candidate genes for further analysis. These genes stand out due to their high statistical significance and large changes in expression, making them prime candidates for further biomarker studies. The SHAP beeswarm plots (**Figures 6 and 7**) reveal the distribution and importance of individual genes as features in machine learning models. Interestingly, the two machine learning models differ in the genes they prioritize, reflecting each model’s unique approach. This divergence underscores the importance of combining insights from both models, as it leads to a more comprehensive understanding of key genetic factors. Meanwhile, the pilot world map visualization (**Figure 4**) presents geographic disparities in breast cancer incidence, showing variations in age-standardized rates (ASR) across countries. This spatial mapping provides valuable insights into regional differences, which can inform further investigation into the environmental or genetic factors contributing to the disease.

In terms of initial results and observations, the volcano plot provides clear distinctions between upregulated and downregulated genes in preeclampsia, with those exhibiting extreme fold changes and strong significance being ideal candidates for further investigation. These findings are consistent with previous biomarker research, reinforcing the robustness of our visualization approach. The SHAP analysis of the machine learning models further underscores this, highlighting how different models prioritize different genes. For example, one model may focus on immune response genes, while the other may highlight metabolic pathways. This difference reflects the distinct architectures and training objectives of the models. By combining their results, we aim to leverage the complementary strengths of both models, offering a more nuanced interpretation of key biomarkers. The geographic disparities shown in the world map pilot indicate significant variability in breast cancer ASR across regions, with higher rates observed in developed countries, likely due to better detection rates, and lower rates in some regions, which may point to underdiagnosis or limited healthcare access. These findings underscore the global inequities in healthcare and the need for biomarkers that are applicable across diverse populations.

This project addresses critical gaps in biomarker research by integrating diverse datasets and advanced visualization techniques. Our findings highlight the importance of considering geographic, genetic, and methodological variability when identifying biomarkers for diseases such as preeclampsia and breast cancer. By combining the outputs of multiple machine learning models and incorporating global data, we ensure that our results are both comprehensive and globally relevant. Ultimately, this work demonstrates the potential for more inclusive biomarker research to improve diagnostic and therapeutic approaches, reduce health disparities, and enhance outcomes worldwide.

**6. Intellectual Merit and Practical Impacts:**

This research contributes to the academic community by merging interactive data visualization with biomarker analysis, addressing a key challenge in genomic research: making complex datasets more accessible and interpretable. Traditional bioinformatics tools often present data in static formats, which can hinder researchers—especially those outside the field—from extracting meaningful insights. By utilizing interactive visualizations such as volcano plots and geojson maps, our work makes these datasets more accessible and easier to understand. This is especially important in biomarker research, where multi-dimensional data must be effectively communicated to reveal patterns and trends.

A unique aspect of our study is the integration of both statistical and geographic dimensions in biomarker analysis. While many biomarker studies focus solely on statistical significance, incorporating spatial data introduces a layer of understanding about population-level patterns and potential biases. This cross-disciplinary approach bridges bioinformatics, data visualization, and global health research, offering a framework that can be adapted to other fields. The interactive tools we have developed—such as dynamic volcano plots and map-based datasets—present an innovative methodology that other researchers can replicate and build upon, facilitating advancements in precision medicine and genomic equity.

Our research on regional variability in breast cancer biomarkers is aligned with recent breakthroughs recognized by the 2024 Nobel Prize in Chemistry, which celebrated advancements in biomolecular structures and computational tools for modeling biological processes. Similarly, our study uses advanced data integration and visualization techniques to explore how gene expression varies across populations, addressing a critical gap in global biomarker applicability. Furthermore, our work contributes to the development of AI-driven visualization tools that make complex biomarker data more accessible and interpretable. Although platforms like AlphaFold3 have increased the open-source availability of biological data, there remains a significant need for intuitive, interactive visualizations. Through our geojson maps and volcano plots, we aim to democratize biomarker knowledge, making it accessible to interdisciplinary scholars and the general public, bridging the gap between bioinformatics and real-world applications in global cancer research and personalized medicine.

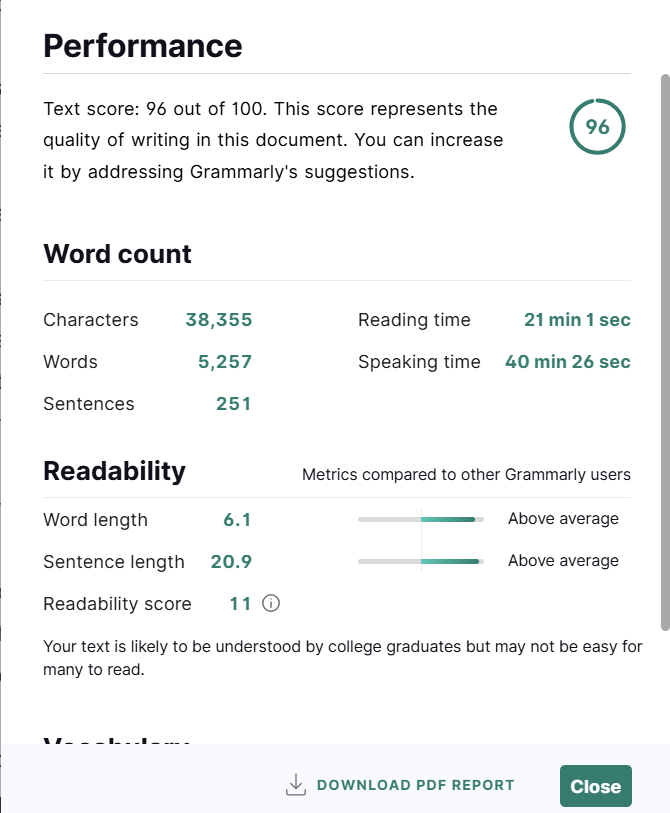
The practical implications of this research are wide-ranging, particularly in personalized medicine and global health equity. By identifying region-specific biomarkers and visualizing gaps in global genomic research, this study provides actionable insights for the development of diagnostic tools and therapies tailored to diverse populations. For example, biomarkers found to be significant in regional datasets but not in global ones could guide the creation of localized healthcare interventions. On the other hand, biomarkers shared across populations could serve as universal targets for diagnostics and treatments, helping to reduce the risk of population bias in clinical applications.

On a larger societal scale, our project promotes inclusivity in genomic research by highlighting the importance of data from underrepresented regions. The spatial visualization of datasets on a global map tells a compelling story about geographic disparities in genomic studies, encouraging policymakers and funding agencies to prioritize research in neglected regions. Additionally, the interactive nature of our tools makes complex genomic data more accessible, enabling clinicians, educators, and even policymakers to engage with the findings without needing extensive technical expertise. This democratization of data interpretation not only accelerates research but also ensures that its benefits are shared more equitably across global populations. Ultimately, our project helps bridge the gap between academic research and societal impact, contributing to a future where biomarker-based diagnostics and treatments are more inclusive and effective.

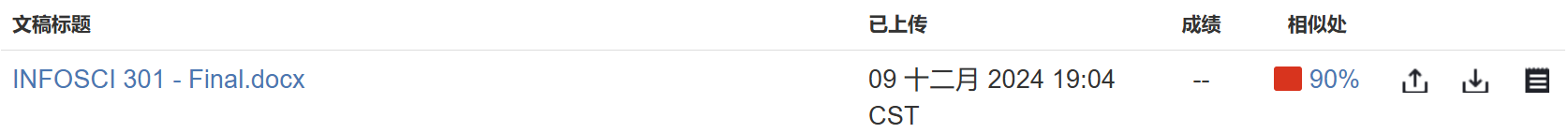
**7. Supplementary Materials:**

GitHub repository: <https://github.com/YP-118/Info301_Final/>

**Grammarly Report:**



**Turnitin Report:**



Note: This is due to being similar to our groups’ own submission. Similarity compared with other groups/paper should be less than 5%.

**References**

Dumitrescu, Raluca G., and Ioana Cotarla. "Understanding Breast Cancer Disparities in African-American Women: A Review of the Literature." *Cancer Epidemiology, Biomarkers & Prevention* 14, no. 4 (2005): 768–78. https://doi.org/10.1158/1055-9965.EPI-04-0452.

Gehlenborg, Nils, Sean I. O'Donoghue, Nitin Baliga, et al. "Visualization of Omics Data for Systems Biology." *Nature Methods* 7, Suppl 3 (2010): S56–S68.<https://doi.org/10.1038/nmeth.1436>.

Lu, F., N. Zeng, X. Xiao, et al. "Exploring the ceRNA Network Involving AGAP2-AS1 as a Novel Biomarker for Preeclampsia." *Scientific Reports* 14 (2024): 27330.<https://doi.org/10.1038/s41598-024-79224-2>.

Munzner, Tamara. *Visualization Analysis and Design.* Boca Raton, FL: CRC Press, 2014.

Kanehisa, Minoru, Satoru Goto, Shigehiko Sato, Kazuaki Furumichi, and Masayuki Tanabe. 2004. "KEGG for Linking Genomes to Life and the Environment." *Nucleic Acids Research* 32 (Database issue): D277-D280. https://doi.org/10.1093/nar/gkh012.

STRING Database. "STRING Network for Protein-Protein Interactions in Preeclampsia-Related Genes." Accessed 2024.<https://cn.string-db.org/cgi/network?taskId=bu6SgERUFBNZ&sessionId=bB6GRQlT9eTz>.

UCSC Xena Browser. "TCGA BRCA Dataset." Accessed 2024.<https://xenabrowser.net/datapages/>.

Global Cancer Observatory. "Age-Standardized Rates for Breast Cancer Incidence across Countries." International Agency for Research on Cancer (IARC), 2024.<https://gco.iarc.who.int/en>.