UNCOVERING GEOGRAPHIC VARIABILITY IN BREAST CANCER BIOMARKERS: INTEGRATING MACHINE LEARNING AND INTERACTIVE VISUALIZATIONS FOR GLOBAL INSIGHTS

Yian Pei, Munkh-Orshikh Munkhbold, Lauris Vo

INTRODUCTION

- Breast cancer is one of the most common cancers worldwide, necessitating early detection and effective treatment.
- Biomarkers are critical for early detection, therapy prediction, and personalized treatment.
- Bioinformatics enables analysis and visualization of complex genomic data.
- **Challenges:** Despite genetic variation across populations, the geographic origin of datasets is often ignored.
- **Project goals:**

1. Analyze biomarkers globally using diverse datasets.
2. Create visualizations to improve accessibility and understanding.
3. Promote equity in breast cancer diagnosis and treatment.

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?

METHODOLOGY

Visualization methods:

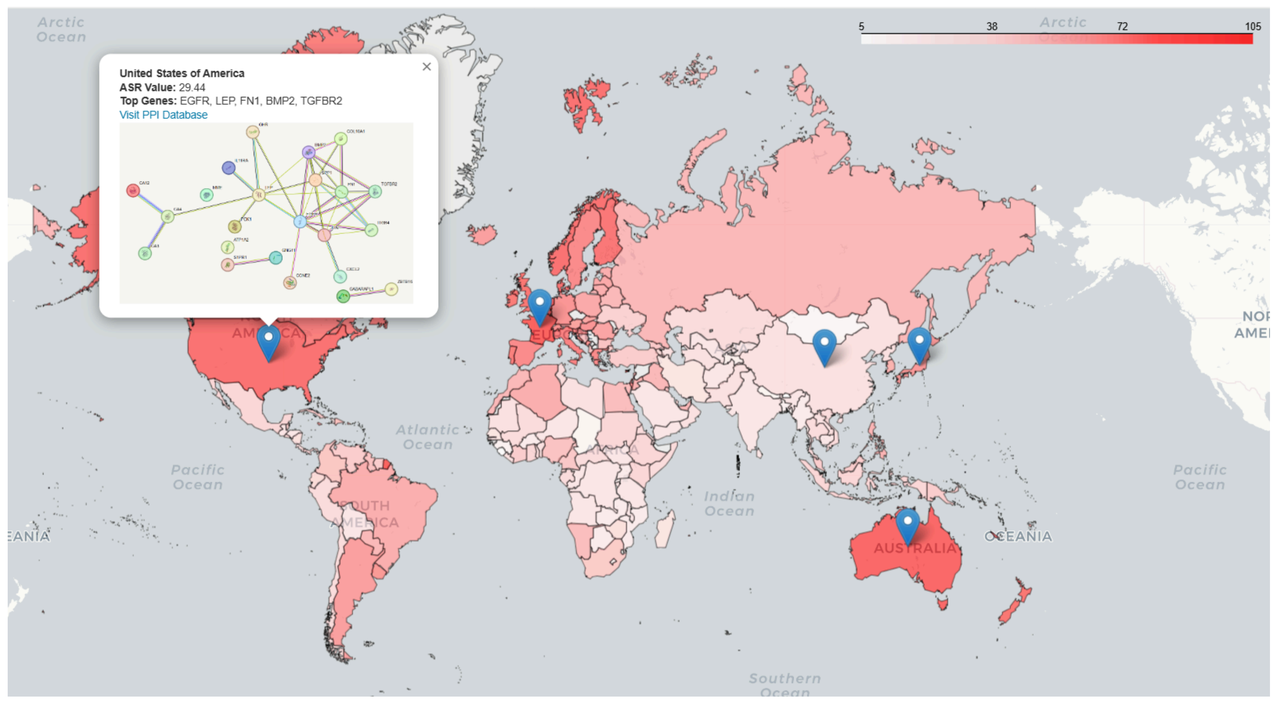
- Ploty for interactive (Volcano plot and number of datasets per country)
- GeoJSON for visualization of world map
- Folium for specific info within a country.

Visualized pipeline for data filtering for Bio Marker:

- Searched for 2-3 datasets per country (NCBI keywords: “breast cancer” + “country name”)
- Cross-analysis to identify genes duplicated across multiple datasets
- ML+XAI: We used Logistic regression and XGBoost for ML. SHAP for XAI, selected top 50 genes for each model.
- KEGG: Kyoto Encyclopedia of Genes and Genomes
- PPI (STRING)

RESULTS

- **Identification of Potential Biomarkers for Breast Cancer:** Identify several potential biomarkers associated with breast cancer that exhibit strong relevance across different geographic regions.
- **Geographic Variation in Breast Cancer Biomarkers:** breast cancer biomarkers do vary across different geographic regions
- **Challenges Due to Lack of Comprehensive Datasets:** There is the lack of comprehensive datasets, even in countries with high age-standardized rates (ASR) of breast cancer.



FUTURE RESEARCH

- **Biomarker Validation:** Wet lab testing and collaboration with biology/oncology experts are crucial for finalizing promising biomarkers.
- **Focus on Developing Countries:** There is a critical need for cancer research in countries with high disease rates to improve diagnostics and treatment.
- **Continental Analysis:** Due to data limitations in third-world countries, future studies could analyze cancer data on a regional or continental scale.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to **Professor Luyao Zhang** for her guidance and support throughout the INFOSCI-301 data visualization course. We also extend our appreciation to our fellow students for their insightful suggestions and constructive feedback, which significantly enhanced the quality of our visualizations and project presentations.

REFERENCE

- **Github link:** https://github.com/YP-118/Info301_Final