Genetic Correspondence and Comparative Spatial Analysis with Age and Gender Specification of ICMR Data on Cancer Incidence in India

Sarbojit Das[1], Swapnonil Mondal[1], Sujash Krishna Basak[1], Sameer Verma[1], Sayanta Biswas[1]

[1]Department of Mathematics & Statistics, Indian Institute of Technology Kanpur, India

March 24, 2024

Abstract

Cancer remains a significant global health challenge, characterized by immense suffering and often limited treatment options. We focus on analyzing genetic data related to five common cancer types. Each cancer type presents unique challenges and implications for diagnosis and treatment. We would like to identify the probable causes of these cancers in terms of genes responsible for each cancer type. This would give us sufficient leads for early identification of each cancer type, thereby, reducing the fatality rate. Our project aims to explore the cancer scenario in India, focusing on spatial and genetic viewpoints. By utilizing statistical concepts and analyzing specific years' data, we aim to shed light on the prevalence, distribution and genetic underpinnings of cancer in the Indian population. Additionally, we investigate gender-specific cancer types and age-specific cancer incidence to provide detailed insights into the cancer landscape in India.

Dataset

In this project, we deal with multiple datasets that are based on cancer patients and the nature of cancer in both male and female human bodies.

1. Cancer causing genes dataset: The input dataset contains 802 samples corresponding to 802 people who have been detected with different types of cancer. Each sample contains expression values of more than 20K genes. Samples are categorized into one of the following types of tumors: BRCA, KIRC, COAD, LUAD and PRAD.

Here's an extended description of the tumors associated with the dataset:

- BRCA (Breast Cancer): Breast cancer is one of the most common cancers among women. It originates in the breast tissue and can occur in both men and women, although it is more prevalent in women. The dataset likely includes samples from individuals with breast cancer, and the analysis aims to identify the genes associated with this type of cancer.

- KIRC (Renal Cancer): Renal cell carcinoma, or kidney cancer, occurs in the lining of small tubes in the kidney. It is one of the common forms of kidney cancer. The dataset may contain samples from individuals with renal cancer, and the analysis aims to uncover the genetic factors associated with this type of cancer.

- COAD (Colon Cancer): Colorectal cancer includes colon cancer (affecting the large intestine) and rectal cancer (affecting the rectum). Colon cancer is a major cause of cancer-related deaths. The dataset could include samples from individuals with colon cancer, with a focus on identifying the genes responsible for this type of cancer.

- LUAD (Lung Cancer - *Adenocarcinoma*): Lung *adenocarcinoma* is a subtype of non-small cell lung cancer. It originates in the cells lining the airways and is one of the most common types of lung cancer. The dataset may contain samples from individuals with lung adenocarcinoma and the analysis aims to uncover the genetic factors specific to this form of lung cancer.

- PRAD (Prostate Cancer): Prostate cancer occurs in the prostate, a small gland in men that produces seminal fluid. It is one of the most common cancers in men. The dataset could include samples from individuals with prostate cancer, focusing on identifying the genes associated with this type of cancer.

2. Cancer causing labels dataset: Samples contain different types of tumors: BRCA, KIRC, COAD, LUAD and PRAD.

3. Estimated State-wise cancer incidences: The table provides estimated incidence of cancer cases in India by States and Union Territory-wise for all sites and for both sexes.

4. Gender-wise different sites of cancer: The table presents the estimated cancer incidence, number of cases, crude rate, and cumulative risk by sex and anatomical sites in India for the year 2022.

5. Age-wise cancer incidences: The table provides gender-disaggregated, estimated top five leading sites of cancer (%) in India by age group ($0-14, 15-39, 40-64$ & 65^+ age groups) for the year 2022.

Data Pre-processing and Cleaning

We conducted data pre-processing and cleaning tasks on two datasets: "`data.csv`" and "`labels.csv`". These datasets contain information related to cancer research, focusing specially on genetic data and associated labels.

1. Essential Packages: For this project, we utilized several essential packages in `R`, including `tibble`, `tidyverse` and `dplyr`.

2. Data Pre-processing and Cleaning: This step is very crucial pertaining to the complexity and volume of genetic data involved.

- Data Importing: Our code reads the CSV files "`data.csv`" and "`labels.csv`" into `R` as tibbles using the `read.csv()` function. These datasets contain genetic data and label information related to cancer research.

- Exploratory Data Analysis (EDA): We performed EDA to gain insights into the structure and content of the datasets. This involved examining the structure of both datasets using the `str()` function and displaying the first few entries of each dataset using the `head()` function.
- Categorical Variable Conversion: We converted some columns in the datasets from *character* type to *factor* type using the `as.factor()` function. This was necessary for categorical variables. We also converted the final dataframe into a more structured format, *tibble* to facilitate efficient analysis.
- Data Merging: We merged the two datasets into a single dataframe using the `cbind()` function, combining the label information with the genetic data.
- Data Quality Checks: We checked for null values (`is.null()`), NA values (`is.na()`), and duplicated entries (`duplicated()`) in the merged dataframe. The absence of such issues indicates clean data.
- Summary Statistics: We calculated summary statistics for the genetic data to understand its distribution and characteristics using the `summary()` function.

The processed dataframe contains both label information and genetic data. It is saved in the "`.Rdata`" format for further analysis and modelling.

Statistical Data Analysis:

The data we have, has $20,534$ columns with 800 rows, as mentioned in the description. And distribution of classes is given below:

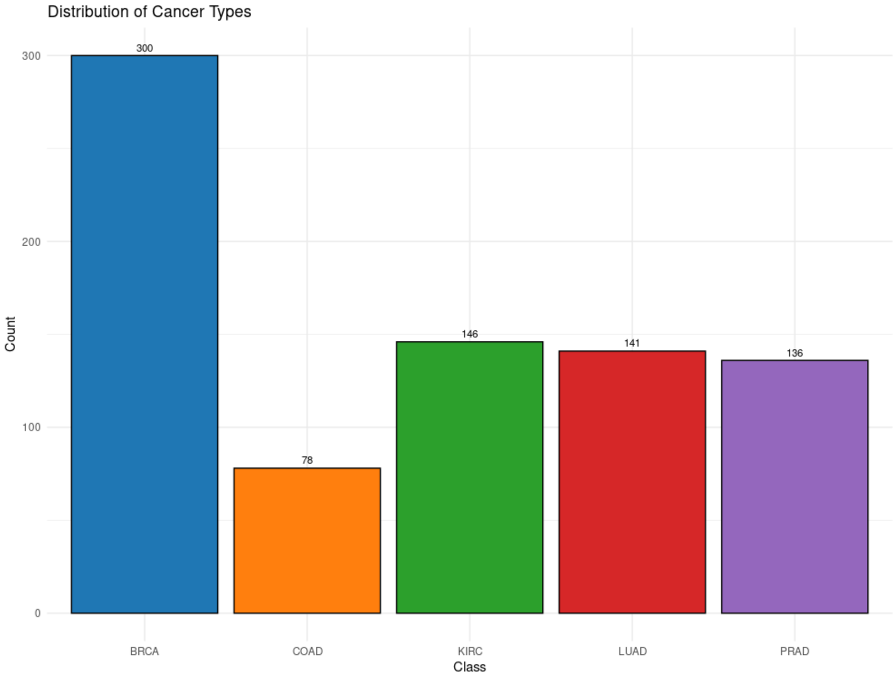

Figure 1: Frequency of Classes

From the frequency plot, we can see that instances for class BRCA are more than that of others followed by KIRC, LUAD and PRAD and class COAD has the least number of instances.