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

Sarbojit Das (231080075)[1], Swapnonil Mondal (231080098)[1], Sujash Krishna Basak (231080093)[1], Sameer Verma (220949)[1], Sayanta Biswas (231080080)[1]

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

March 27, 2024

**Abstract**

Cancer remains a significant global health challenge, characterized by immense suffering and often limited treatment options. We aim to analyze genetic data related to five common cancer types, each presenting unique challenges and implications for diagnosis and treatment. Our goal is to identify the probable genetic causes underlying these cancers, providing leads for early identification and reducing fatality rates. Focusing on India, we explore the spatial and genetic aspects of cancer, utilizing statistical concepts and analyzing data from specific years. We aim to shed light on cancer prevalence, distribution and genetic underpinnings within the Indian population. Additionally, we investigate gender-specific and age-specific cancer incidence to offer detailed insights into the cancer landscape in India.

## 1 Introduction

Cancer is a formidable adversary to human health. It persists as a significant global burden, causing immeasurable suffering and presenting substantial challenges to healthcare systems worldwide. Its insidious nature lies in the uncontrolled growth and dissemination of abnormal cells, disrupting the delicate balance of cellular regulation inherent in the human body's natural processes. Despite remarkable advancements in medical science, the complexities of cancer remain a formidable challenge, often accompanied by limited treatment options and devastating outcomes for affected individuals and their families.

This project endeavors to delve into the multifaceted realm of cancer. With a diverse population and a unique set of demographic, environmental and genetic factors, India provides a rich tapestry for exploring the spatial, genetic, gender-specific and age-specific dimensions of cancer prevalence and distribution. Furthermore, our

study extends beyond genetic analyses to explore the spatial aspects of cancer incidence in India. Leveraging statistical concepts and data from specific years, we seek to illuminate patterns of cancer prevalence and distribution across different regions of the country. This spatial perspective is crucial for understanding the geographic disparities in cancer burden and for informing targeted interventions and resource allocation efforts. Such insights hold the potential to inform tailored prevention, screening and treatment strategies, thereby improving outcomes and mitigating the impact of cancer on individuals and society at large.

Through this comprehensive exploration of cancer incidence in India, our project aims to contribute to a deeper understanding of the disease landscape. By integrating these aspects, we endeavor to shed light on the complexities of cancer and to ultimately strive towards a future where the burden of cancer is significantly reduced.

## 2 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 this 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.
- KIRC (Renal Cancer): Renal cell carcinoma, or kidney cancer, occurs in the lining of small tubes in the kidney.
- 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.
- 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.
- 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.

**2. 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.

**3. 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.

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

- Cum-risk: Cumulative risk of developing cancer in the age range of 0 to 74 years. It represents the likelihood or probability of an individual developing cancer within this specified age range.

$$\boxed{\text{Cumulative Risk} = \frac{\text{Number of Cancer Cases}}{\text{Total Population}} \times 100}$$

- Crude Rate (CR): Crude rate refers to the total number of cancer cases occurring in a population divided by the total population, expressed as a rate per a specific unit of time (usually per $100,000$ population). It provides a general overview of cancer incidence within a population, without considering factors such as age distribution.

$$\boxed{\text{Crude Rate} = \frac{\text{Number of Cancer Cases}}{\text{Total Population}} \times 100,000}$$

- Age-Adjusted Rate (AAR): Age adjusted rate is a standardized rate that takes into account the age distribution of a population. It allows for fair comparisons of cancer rates between different populations or over time by removing the influence of age as a confounding factor.

$$\boxed{\text{Age-Adjusted Rate } = \sum_{i=1}^{n} \left( \frac{\text{Age-specific rate}_i \times \text{Population}_i}{\text{Total Population}} \right) \times 100,000}$$

- Malig Imn.Prol D (Malignant Immunoproliferative Diseases): This term refers to a group of disorders characterized by the abnormal proliferation of immune cells, leading to the development of malignancies. These diseases involve the uncontrolled growth of cells of the immune system, such as lymphocytes or plasma cells, and can include conditions like lymphomas, leukemias and multiple myeloma.

**4. 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.

# 3 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. Pre-processing:** 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.

# 4 Data Insights

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

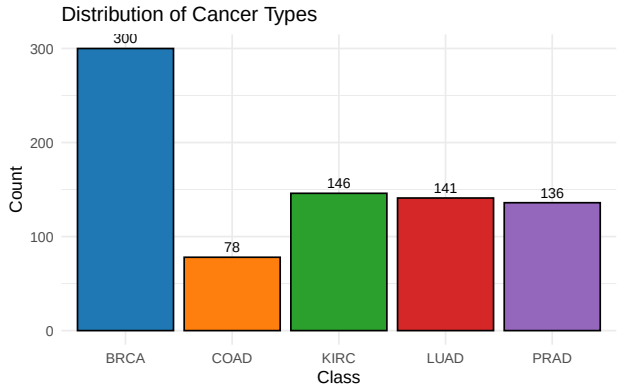



Figure 1: Frequency of Classes

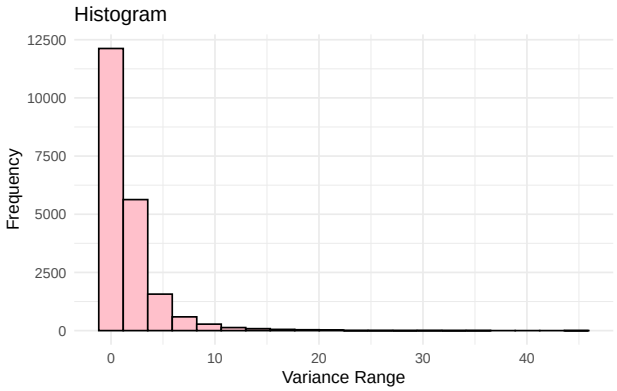
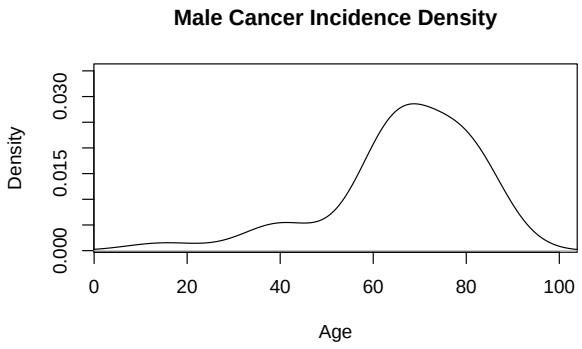
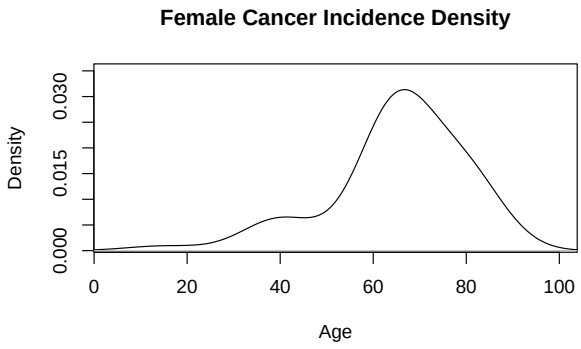



Figure 2: Histogram of Variance Range

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.

- Variance Threshold: From the graph, it's easy to see that most of the columns have low variance. Features with high variance have data points that are spread out over a wide range. These features are considered significant because they contain valuable information that can help in discriminating between different classes or categories. Similarly, features with low variance have data points that have little variation. Such features may not carry much discriminative power and can be considered less significant.
- Age specific genderwise cancer incidences: The shapes of the density curves represents the leptokurtic curve for females and mesokurtic curve for males.



# 5 F-test for comparing gene expression across cancer groups

We want to determine whether there are statistically significant differences between the mean gene informations encoded by a particular gene (*response*) among the independent cancer groups (*covariates*). In this case, with reference to the `ICMR` dataset, the independent categorical variable is the '`Class`' column which represents the 5 different cancer types. The dependent variable is the gene expression levels of a specific gene. We analyze each gene individually in this setup, making it multiple F-tests.

We have,

- 5 groups (categories of cancer types)
- $n_i$ observations in the $i^{\text{th}}$ group (where $i = 1, 2, ..., 5$)
- $N$ total observations ($N = n_1 + n_2 + ... + n_5$)
- $X_{ij}$ is the observation in the $i^{\text{th}}$ group and the $j^{\text{th}}$ gene expression level ($j = 1, 2, ..., 20532$)
- $X \approx \mathcal{N}(\mu, \sigma^2)$ for large sample size where $X$ is a random variable that denotes gene expression for a particular gene with mean $\mu$ and variance $\sigma^2$.

The null hypothesis ($H_0$) for each F-test is that there are no differences in mean gene expression levels among the different cancer types.

$$H_0 : \mu_1 = \mu_2 = ... = \mu_5$$

where: $\mu_i$ is the population mean of the $i^{\text{th}}$ group.

The alternative hypothesis ($H_1$) is that at least one pair of mean gene expression levels is different.

$$H_1 : \text{At least one } \mu_i \text{ is different from the others.}$$

To test these hypotheses, we calculate the F-statistic: $F = \frac{\text{Between-group variance}}{\text{Within-group variance}}$

Rejection Criteria: If the p-value associated with the F-statistic is below a certain threshold (typically 0.05), we reject the null hypothesis.

Table 1: Top 3 high variance genes

| Gene | F_statistic | p_value | variance |
|---|---|---|---|
| gene_9176 | 3463.550 | 0 | 44.76385 |
| gene_9175 | 4194.489 | 0 | 36.36194 |
| gene_15898 | 1905.191 | 0 | 34.50391 |

Table 2: Bottom 3 high variance genes

| Gene | F_statistic | p_value | variance |
|---|---|---|---|
| gene_12668 | 3.163906 | 0.0137332 | 0.0014541 |
| gene_12670 | 3.250183 | 0.0118692 | 0.0013394 |
| gene_4834 | 2.446886 | 0.0450807 | 0.0005937 |

# 6 One vs. All t-test

In a one-vs-all t-test scenario, we compare the means of one cancer group (the "*one*" group) against the means of all other cancer groups excluding that group (the "*all*" group). It's commonly used in situations where one wants to compare a specific group against the rest of the data.

- Null Hypothesis ($H_0$): The null hypothesis assumes that there is no difference between the means of the "*one*" group and the means of the "*all*" group.

$$H_0 : \mu_{\text{one}} = \mu_{\text{all}}$$

- Alternative Hypothesis ($H_1$): The alternative hypothesis states that the means of the "*one*" group and the "*all*" group are different.

$$H_1 : \mu_{\text{one}} \neq \mu_{\text{all}}$$

- Test Statistic: $t = \dfrac{\bar{x_1} - \bar{x_2}}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ where:
  - $\bar{x_1}$ and $\bar{x_2}$ are the means of the "*one*" group and the "*all*" group respectively.
  - $s_1$ and $s_2$ are the standard deviations of the "*one*" group and the "*all*" group respectively.
  - $n_1$ and $n_2$ are the sample sizes of the "*one*" group and the "*all*" group respectively.
- Rejection Criteria: If the p-value is smaller than a predetermined significance level (commonly 0.05), we reject the null hypothesis.
- Results: We perform this test for the top 3 high variance genes.

```
Performing 'one vs. all' t-tests for gene_9176 :
PRAD vs. All for gene_9176 : p-value = 0
LUAD vs. All for gene_9176 : p-value = 2.3192579e-50
BRCA vs. All for gene_9176 : p-value = 3.9070211e-28
KIRC vs. All for gene_9176 : p-value = 2.5060458e-42
COAD vs. All for gene_9176 : p-value = 0.00025240039
Performing 'one vs. all' t-tests for gene_9175 :
PRAD vs. All for gene_9175 : p-value = 2.0867204e-256
LUAD vs. All for gene_9175 : p-value = 4.2859671e-35
BRCA vs. All for gene_9175 : p-value = 8.2966265e-23
KIRC vs. All for gene_9175 : p-value = 4.8127605e-50
COAD vs. All for gene_9175 : p-value = 8.2634542e-27
Performing 'one vs. all' t-tests for gene_15898 :
PRAD vs. All for gene_15898 : p-value = 6.2853312e-37
LUAD vs. All for gene_15898 : p-value = 2.4837977e-99
BRCA vs. All for gene_15898 : p-value = 3.5424646e-33
KIRC vs. All for gene_15898 : p-value = 8.7197042e-28
COAD vs. All for gene_15898 : p-value = 7.6822615e-16
```