



सत्यमेव जयते

Government Of India

Ministry of Statistics & Programme Implementation

National Sample Survey Office

Data Processing Division

Technical Coordination

**Project Report for the Summer Internship
Scheme 2024-25 of MoS&PI**

Conducted under

**Anthropological Survey of India
Ministry of Culture, Government of India
EN-79, Sector V, Salt Lake City,
Kolkata 700091**

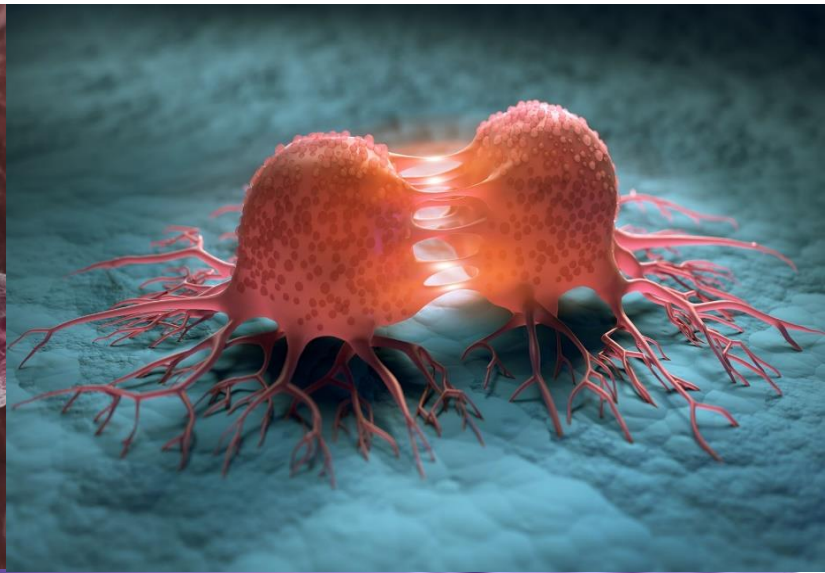
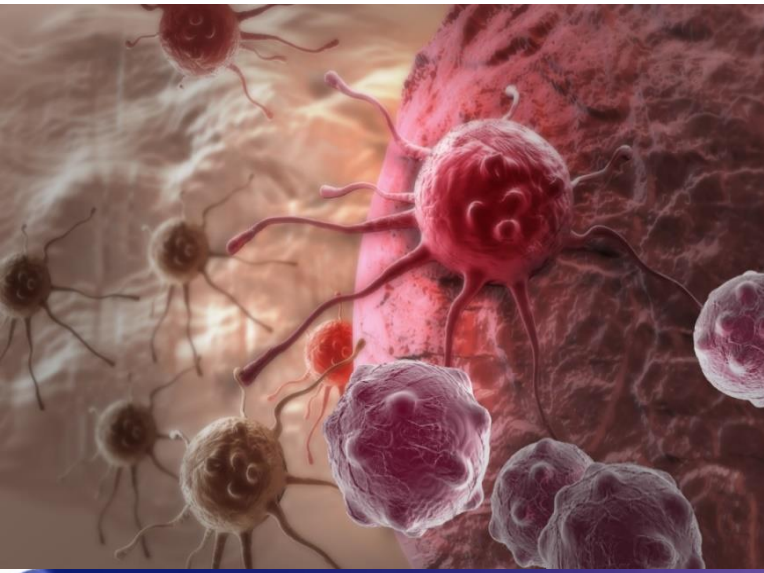
SUBMITTED BY:-

Name:- Tanmay Gayen

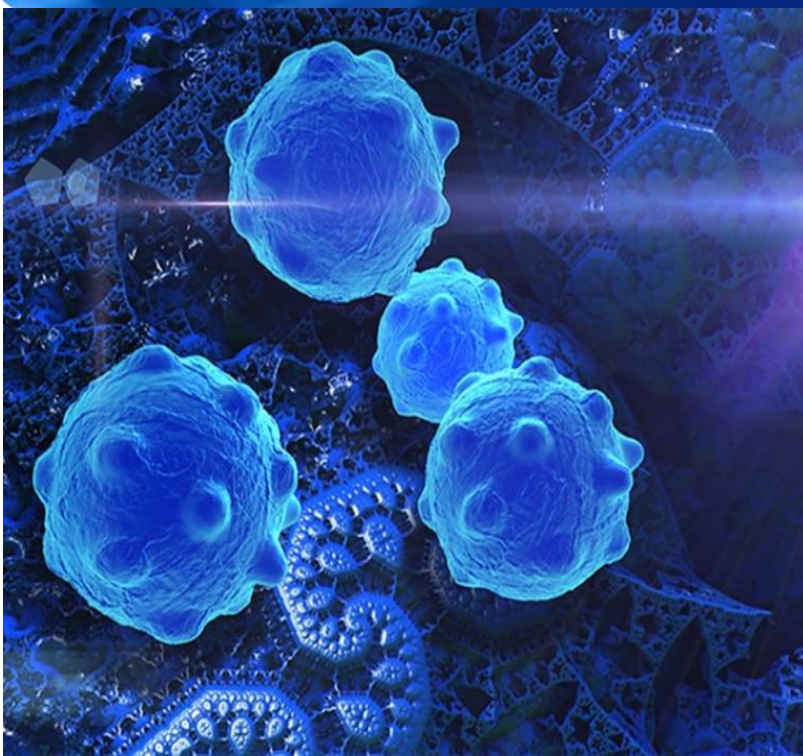
University Name:- VISVA-BHARATI

SESSION:- 2022-2024

REGISTRATION NO. :- VB-0311 of 2022-2023



**Analysis of Conserved and
Differential Gene
Expression Patterns in
Indian Cancer Patients: A
Pan-cancer Study**



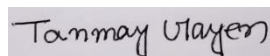
Declaration

I, TANMAY GAYEN, student of M.Sc in Statistics in Visva Bharati University(2022-24) declare that I have been working as intern under Ministry of Statistics & Programme Implementation(MOSPI) summer internship scheme from 3rd June, 2024 to 31st July, 2024 conducted under Anthropological Survey of India.

I affirm that this report is a result of my personal efforts and contributions. Any reference to existing research, direct quotations, or paraphrasing has been properly acknowledged and cited in accordance with academic standards. This report has not been previously submitted for any degree, diploma, or other qualifications at this or any other institution.

I understand the importance of this declaration and the potential consequences of any breach of academic integrity, including but not limited to disciplinary action by my institution. I hereby certify that the information presented in this report is true and accurate to the best of my knowledge and belief.

Signature:

A rectangular box containing a handwritten signature in black ink. The signature appears to read "Tanmay Gayen".

Acknowledgment

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all among the completion of my project. All that I have done is only due to such supervision and assistance and I would not regret to thank them.

I am indebted to DG, NSSO , for selecting me and giving me the opportunity to work as intern Ministry of Statistics & Programme Implementation(MOSPI) summer internship scheme.

I would like to express my gratitude to Prof. B.V.Sharma, Director, Anthropological Survey of India Kolkata for allowing me to work in this esteemed institution.

I take this opportunity to express a deep sense of gratitude towards Subhra Bhattacharyya maam for providing excellent guidance, encouragement and inspiration throughout the project work. I thank Mr. Sahid Mollick, Junior Research Fellow for guiding me in analysis of the data and Mr. P.K Mandal for his constant support throughout the period. They have constantly inspired me to work harder and perform better towards the completion the project work.

I would also like to thank my all friends for their valuable suggestions and helpful discussions. Finally, I expressed my gratitude to my parents and other family members for their constant support and motivation throughout my life.

Date:-

Reg. No: - VB-0311 of 2022-23

CONTENTS

Page no.

❖ Introduction	6
❖ Review of literature	7
❖ Objectives of study	8
❖ Material and Methods	9-11
❖ Data Analysis	11-32
❖ Conclusion	33
❖ Reference	34

1. Introduction

Cancer is a multifaceted disease characterized by uncontrolled cell growth and proliferation, resulting from genetic and epigenetic alterations. The incidence and mortality rates of cancer are significantly influenced by geographical, environmental, and genetic factors, leading to unique patterns of cancer in different populations. India, with its diverse population and varying environmental exposures, presents a distinct landscape for cancer epidemiology.

Pan-cancer analysis involves assessing frequently mutated genes and other genomic abnormalities common to many different cancers, regardless of tumor origin. Pan-cancer studies aim to uncover commonalities and differences across these cancer types by analyzing comprehensive datasets. These studies have the potential to reveal conserved gene expression patterns that are fundamental to cancer biology and to identify differential gene expression patterns that may be unique to specific cancer types or populations. Understanding these patterns can aid in the development of targeted therapies and personalized medicine approaches.

The primary objective of this study, entitled ‘Analysis of Conserved and Differential Gene Expression Patterns in Indian Cancer Patients: A Pan-cancer Study’ is to analyze conserved and differential gene expression patterns in Indian cancer patients across multiple cancer types. By leveraging high-throughput sequencing data and advanced bioinformatics tools. It is aimed to:

- I. Identify Conserved Gene Expression Patterns: Determine genes that exhibit consistent expression changes across different cancer types, providing insights into common mechanisms of oncogenesis.
- II. Identify Differential Gene Expression Patterns: Discover genes with variable expression across different cancer types, which may highlight unique biological pathways and potential therapeutic targets specific to each cancer type.
- III. Explore Population-Specific Variations: Investigate the extent to which genetic and environmental factors specific to the Indian population influence gene expression patterns in cancer.

With the development of research on cancer genomics and microenvironment, a new era of oncology focusing on the complicated gene regulation of pan-cancer research and cancer immunotherapy is emerging. This study aimed to identify the common gene expression characteristics of multiple cancers – lung cancer, liver cancer, kidney cancer, cervical cancer, and breast cancer – and the potential therapeutic targets in public databases.

Hypothesis

To identify groups of genes that behave similarly across samples and identify the distribution of samples corresponding to each cancer type. Therefore, this study focuses on applying various clustering techniques,

- First, clustering technique on all genes is applied to identify:
 - Genes whose expression values are similar across all samples of each cancer type
 - Samples of the same class (cancer type) which also correspond to the same cluster is analysed
 - Samples identified to be belonging to another cluster but also to the same cancer type.

2. Review of literature

Cancer is driven by genetic change, and the advent of massively parallel sequencing has enabled systematic documentation of this variation at the whole-genome scale. Pan-cancer analysis aims to examine the similarities and differences among the genomic and cellular alterations found across diverse tumor types. Several researches have been carried out in this area. Rashmi and Majumdar (2022) conducted a pan-cancer analysis of the THAP9 and THAP9-AS1 gene pair in various cancers and observed that although the expression levels of the two genes, THAP9 and THAP9-AS1, varied in different tumors, the expression of the gene pair was strongly correlated with patient prognosis; higher expression of the gene pair was usually linked to poor overall and disease-free survival. Thus, THAP9 and THAP9-AS1 may serve as potential clinical biomarkers of tumor prognosis.

Kumar et al. (2019) worked on pancreatic cancer and found it to be the is a leading cause of mortality in the Western world. Differential gene expression (DGE) analysis on the mRNA transcripts was performed after quantile filtration and did correlation analysis of enhancer expression with patient survival. They concluded that coupling of enhancer profiles with gene expression changes has possibly unearthed a powerful approach to treat disease, and can be expected to strengthen personalized medicine in the near future.

Research on pan-cancer chromatin analysis of the human vtRNA genes, uncovers their association with cancer biology (Fort, 2021). They uncovered new evidence linking the vtRNAs with the immune response, cell proliferation and overall survival in cancer, which guarantees further investigation.

Varn, et al., (2018) used The Cancer Genome Atlas (TCGA) for systematic pan-cancer analysis which reveals immune cell interactions in the tumor microenvironment. While the previous analyses were performed in a pan-cancer manner, we also sought to examine the distribution of immune infiltration across different cancer types. To accomplish this, we stratified the TCGA dataset by cancer type and compared the distribution of immune cell infiltration scores for our four representative cell lineages found that in nearly all cancer types, patients sharing a diagnosis did not exhibit an association between mutation load and immune infiltration. Exceptions to this were in colorectal and endometrial cancers, where MSI-based genomic instability was associated with increased immune infiltration. Going forward, it will be important to identify the immunogenic determinants specific to each tumor type, as current immunotherapeutic approaches are dependent on the presence of immune cells at the tumor microenvironment.

3. Objectives of study

Indian Council of Medical Research(ICMR) aims to investigate various types of cancers, including breast cancer, renal cancer, colon cancer, lung cancer, and prostate cancer, which have become a growing concern in recent years. They seek to identify the likely genetic causes responsible for each cancer type.

- This study will utilize publicly available datasets from Indian cancer patients, encompassing a range of cancer types.
- Comprehensive bioinformatics analyses, including differential expression analysis, pathway enrichment, and network analysis, will elucidate the underlying molecular mechanisms driving cancer in this population.
- Ultimately, this pan-cancer analysis aims to contribute to the global understanding of cancer biology, with a particular focus on the Indian context.
- This research will facilitate early detection of each cancer type, thereby reducing the fatality rate.
- This study investigates the variability in gene expression levels among Indian cancer patients with five distinct cancer types, highlighting how different genes and gene clusters distinguish tissue-specific cancers, such as adenocarcinomas from other carcinomas.
- Using techniques like PCA and clustering, the study reduces the dimensionality of the high-dimensional expression dataset to identify key genes and gene groups of interest. Modern visualization techniques illustrate these differences.
- Future research directions include exploring these comparisons in an integrative setting by combining gene expression data with copy number alterations and protein expressions.
- The findings could pave the way for the development of more effective diagnostic, prognostic, and therapeutic strategies tailored to the unique genetic and environmental landscape of Indian cancer patients.

4. Material and Methods

I. Data:

Indian Council of Medical Research provided the dataset on genes responsible for the five cancers. The dataset consists of 801 samples from 801 individuals diagnosed with various types of cancer. Each sample contains expression values for over 20,000 genes. The samples are categorized into one of the following tumor types:

BRCA (breast adenocarcinoma), COAD (colon adenocarcinoma), KIRC (kidney renal clear cell carcinoma), PRAD (prostate adenocarcinoma) and Lung adenocarcinoma (*LUAD*). The data you have provided is a table with gene expression values for different samples. Here is a brief explanation of the data:

- **Rows (samples):** Each row represents a different sample.
- **Columns (genes):** Each column represents the expression levels of different genes.
- **Values:** The values in the table indicate the expression levels of each gene in each sample. For instance, the expression level of gene_1 in sample_0 is approximately 2.017, and the expression level of gene_3 in sample_1 is approximately 7.586.

Notably, gene_0 has a constant expression level of 0 across all samples, indicating no expression or a possible control/reference gene. The other genes show varying levels of expression across the samples.

Source: The data was collect from <https://www.kaggle.com/datasets/shibumohapatra/icmr-data/code>.

Table 1: Number of different cancer patient in India up to 2022.

Cancer		Patient	
Name	Code	Number	Percentage
Breast adenocarcinoma	BRCA	300	37.45%
Colon adenocarcinoma	COAD	78	9.74%
Kidney renal clear cell carcinoma	KIRC	146	18.23%
Prostate adenocarcinoma	PRAD	136	16.98%
Lung adenocarcinoma	LUAD	141	17.60%
Total		801	

The data represents 37.45% of patients detected from breast cancer which is the most prevalent, followed by Kidney renal clear cell carcinoma (18.23%), Prostate adenocarcinoma (16.98%) and Lung adenocarcinoma (17.60%) and 9.74 5 of Colon adenocarcinoma.

II. Methodology

The methods applied to analyse the data are the following: Mean-variance summary, low-varying covariate filtering, missing data imputation, visual summaries of differential expression (scatter plot , boxplot etc), univariate multi-class logistic regression, principal component analysis, Cluster analysis etc.

Mean-Variance Summary

A mean-variance summary is a statistical method used to understand the distribution of a dataset by summarizing the mean (average) and variance (spread) of each variable. This summary helps to identify which variables have the most variability and might therefore be most informative or significant.

Low-Varying Covariate Filtering

Low-varying covariate filtering is a preprocessing step in data analysis where covariates (variables) with low variability across samples are removed. This is because low-variance covariates typically contribute little to distinguishing between different conditions or outcomes, and their removal can reduce noise and improve the performance of subsequent analyses.

Missing Data Imputation

Missing data imputation involves replacing missing values in a dataset with substituted values. Methods for imputation include:

- **Mean/Median Imputation:** Replacing missing values with the mean or median of the observed values.
- **Regression Imputation:** Using regression models to predict and fill in missing values.
- **Multiple Imputation:** Creating multiple complete datasets by filling in missing values using a statistical model, analyzing each dataset separately, and then combining the results.

Visual Summaries of Differential Expression

Visual summaries are used to depict differential expression of genes or other variables. Common visualizations include:

- **Scatter Plots:** Show the relationship between two variables, often used to compare expression levels between conditions.
- **Boxplots:** Display the distribution of expression levels across different conditions or groups, highlighting medians, quartiles, and outliers.

Univariate Multi-Class Logistic Regression

Univariate multi-class logistic regression is a statistical method used to predict the probability of a categorical dependent variable with more than two classes, based on a single predictor variable. The model estimates the odds of each class relative to a baseline class, providing insights into how changes in the predictor affect the likelihood of each outcome.

Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique used to transform a large set of variables into a smaller set of uncorrelated components. These components capture the maximum variance in the data, making it easier to visualize and analyze patterns without losing significant information.

Cluster Analysis

Cluster analysis is a technique used to group similar observations into clusters based on their characteristics. Common methods include:

- **K-Means Clustering:** Partitions data into K clusters by minimizing the within-cluster variance.

- **Hierarchical Clustering:** Builds a tree of clusters by iteratively merging or splitting clusters based on similarity measures.
- R programming was used for analysis the data for the present study.

5. DATA ANALYSIS

Table 2 : Sample data used for the study

	id	cancer	gene	expression
1	sample_0	PRAD	A1BG	9.5135381
2	sample_0	PRAD	A1CF	0.0000000
3	sample_0	PRAD	A2BP1	4.0636582
4	sample_0	PRAD	A2LD1	7.7648053
5	sample_0	PRAD	A2ML1	4.7476559
6	sample_0	PRAD	A2M	13.7143958
7	sample_0	PRAD	A4GALT	10.0344964
8	sample_0	PRAD	A4GNT	0.0000000
9	sample_0	PRAD	AAA1	0.0000000
10	sample_0	PRAD	AAAS	9.8334578
11	sample_0	PRAD	AACSL	0.0000000
12	sample_0	PRAD	AACS	10.8612653
13	sample_0	PRAD	AADACL2	0.0000000
14	sample_0	PRAD	AADACL3	0.0000000
15	sample_0	PRAD	AADACL4	0.5918709
16	sample_0	PRAD	AADAC	1.0102786
17	sample_0	PRAD	AADAT	6.9628497
18	sample_0	PRAD	AAGAB	10.9597050
19	sample_0	PRAD	AAK1	9.7695089
20	sample_0	PRAD	AAMP	12.2454042
21	sample_0	PRAD	AANAT	0.5918709
22	sample_0	PRAD	AARS2	8.9688280
23	sample_0	PRAD	AARSD1	10.1753000
24	sample_0	PRAD	AARS	12.3990643
25	sample_0	PRAD	AASDHPPT	8.4798772
26	sample_0	PRAD	AASDH	7.3557651
27	sample_0	PRAD	AASS	8.3310513
28	sample_0	PRAD	AATF	10.6434594
29	sample_0	PRAD	AATK	6.8212255
30	sample_0	PRAD	ABAT	10.6374943

This table shows the data of expression of genes in a cancer. A particular expression of a particular gene of a particular cancer patient is given. That means, for a cancer disease PRAD (prostate adenocarcinoma) of a patient sample_0 and for the gene A1BG the expression is 9.5135381.

Table 3: Data Summary 1

	▲	cancer	gene	exp_mean	exp_sd
1		BRCA	A1BG	7.12476272	1.20268284
2		BRCA	A1CF	0.10344457	0.26374707
3		BRCA	A2BP1	0.52204815	0.91279158
4		BRCA	A2LD1	6.45178696	0.80921772
5		BRCA	A2M	13.40641268	0.98708979
6		BRCA	A2ML1	3.04687295	3.18884286
7		BRCA	A4GALT	7.99071024	1.15963370
8		BRCA	A4GNT	0.60963810	0.61939396
9		BRCA	AAA1	0.03856682	0.17823596
10		BRCA	AAAS	9.44074119	0.43210006
11		BRCA	AACS	9.94684800	0.70689459
12		BRCA	AACSL	0.66085502	1.12968643
13		BRCA	AADAC	1.21385504	1.57278847
14		BRCA	AADACL2	0.16068995	0.66080276
15		BRCA	AADACL3	0.11104783	0.30017595
16		BRCA	AADACL4	0.03994810	0.18143920
17		BRCA	AADAT	5.51552378	1.48749770
18		BRCA	AAGAB	10.64318540	0.54966833
19		BRCA	AAK1	9.77924491	0.51495307
20		BRCA	AAMP	11.17445383	0.54739662
21		BRCA	AANAT	0.36444454	0.47997686
22		BRCA	AARS	11.47667256	0.66346066
23		BRCA	AARS2	8.74587163	0.52769634
24		BRCA	AARSD1	8.94313000	0.65160999
25		BRCA	AASDH	8.19487523	0.58484866
26		BRCA	AASDHPPT	9.72702367	0.60674980
27		BRCA	AASS	7.35021816	1.16168883
28		BRCA	AATF	10.26339910	0.57410463
29		BRCA	AATK	5.66829861	1.30841602
30		BRCA	ABAT	9.50299950	2.02752287

In table 3, the expression of mean and standard deviation of the data is calculated for each specific cancer and the gene responsible.

Table 3: Data summary of across cancer

	cancer	gene	exp_mean	exp_sd
1	BRCA	A1BG	7.124762716	1.20268284
2	COAD	A1BG	4.228513088	0.82983111
3	KIRC	A1BG	5.439405401	1.22077644
4	LUAD	A1BG	6.757651658	1.46334954
5	PRAD	A1BG	5.428152347	1.21141350
6	BRCA	A1CF	0.103444568	0.26374707
7	COAD	A1CF	7.015067243	1.50982096
8	KIRC	A1CF	6.501890408	2.06189516
9	LUAD	A1CF	0.603240053	1.65248765
10	PRAD	A1CF	0.743142809	1.54719151
11	BRCA	A2BP1	0.522048150	0.91279158
12	COAD	A2BP1	1.671764044	1.85393855
13	KIRC	A2BP1	0.616393095	0.94345002
14	LUAD	A2BP1	1.165228267	2.04440922
15	PRAD	A2BP1	2.902388567	1.70579454
16	BRCA	A2LD1	6.451786957	0.80921772
17	COAD	A2LD1	7.297964949	0.59702709
18	KIRC	A2LD1	7.017807751	1.08129508
19	LUAD	A2LD1	6.625939175	0.61483273
20	PRAD	A2LD1	7.169659196	0.65038800
21	BRCA	A2M	13.406412679	0.98708979
22	COAD	A2M	12.044995875	1.12913994
23	KIRC	A2M	14.920322220	0.75296231
24	LUAD	A2M	14.126835101	1.16246618
25	PRAD	A2M	13.610295935	1.05948148
26	BRCA	A2ML1	3.046872954	3.18884286
27	COAD	A2ML1	1.027825126	1.88463520
28	KIRC	A2ML1	0.393517304	0.54194209
29	LUAD	A2ML1	1.828781699	2.49721134
30	PRAD	A2ML1	3.824730188	1.96683550

Table 3, gives the summary of data across all cancer data, used in the study.

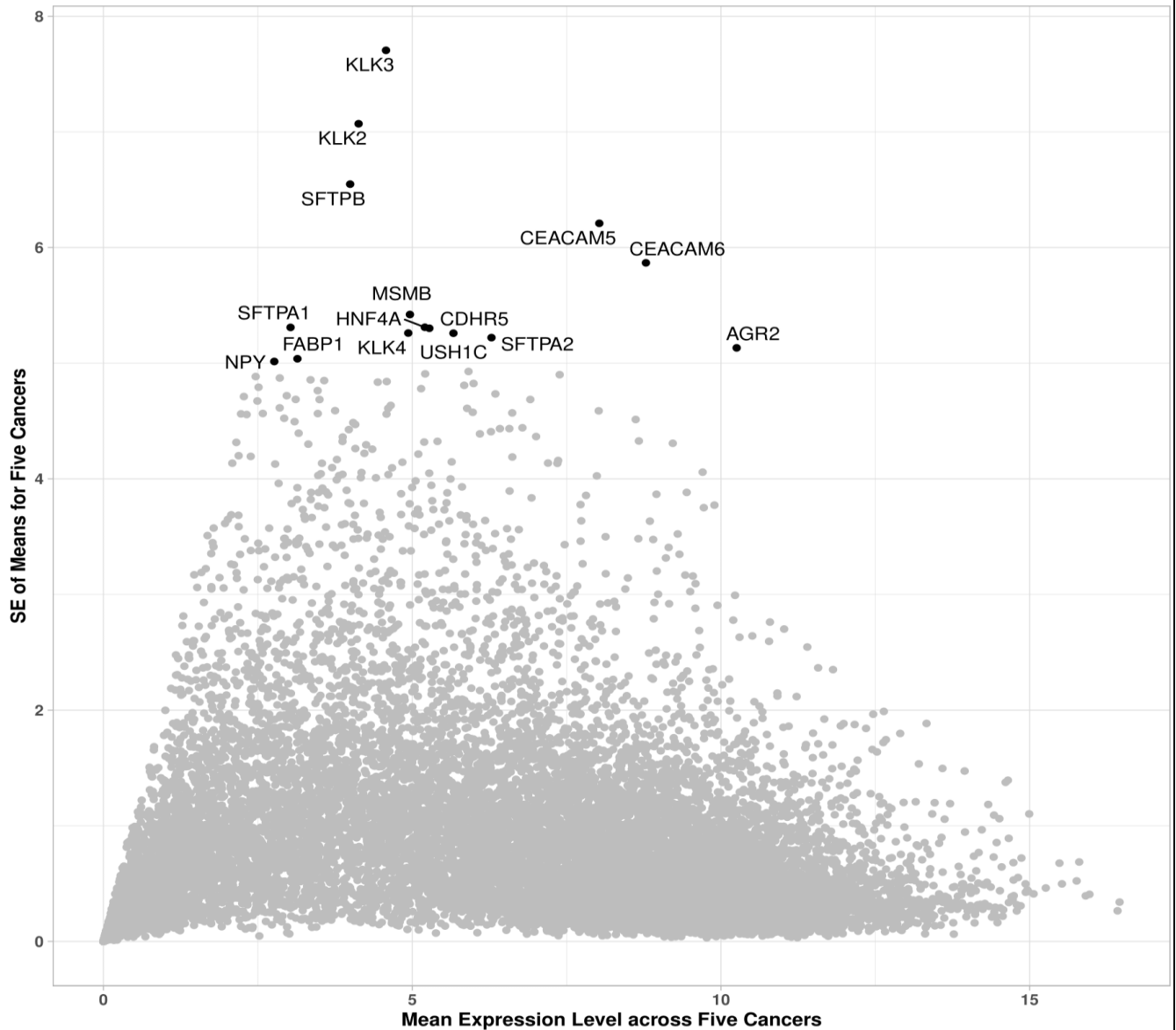
Table 4: Data Summary 2

	gene	exp_mean2	exp_se	label	color
1	A1BG	5.79569704	1.16295643		grey
2	A1CF	2.99335702	3.45006075		grey
3	A2BP1	1.37556442	0.97056667		grey
4	A2LD1	6.91263161	0.36061660		grey
5	A2M	13.62177236	1.05759819		grey
6	A2ML1	2.02434545	1.41225406		grey
7	A4GALT	8.53360359	1.14546458		grey
8	A4GNT	1.15274701	1.21240734		grey
9	AAA1	0.31562315	0.43729916		grey
10	AAAS	9.54253701	0.20603161		grey
11	AACS	9.73112643	0.42825556		grey
12	AACSL	1.40864828	1.68393457		grey
13	AADAC	1.79445556	1.65464144		grey
14	AADACL2	0.09469792	0.05878337		grey
15	AADACL3	0.07323771	0.03314445		grey
16	AADACL4	0.16897148	0.14759179		grey
17	AADAT	6.72318248	1.21322562		grey
18	AAGAB	10.03762520	0.50958484		grey
19	AAK1	9.84513415	0.56297897		grey
20	AAMP	11.34790319	0.29797412		grey
21	AANAT	0.53065288	0.20144897		grey
22	AARS	11.63055692	0.17462170		grey
23	AARS2	8.90311389	0.15887836		grey
24	AARSD1	9.15509286	0.21841730		grey
25	AASDH	7.90730855	0.27564429		grey
26	AASDHPPT	9.69411043	0.23657626		grey
27	AASS	7.99825845	1.04606848		grey
28	AATF	10.31951674	0.25216144		grey
29	AATK	6.54627152	0.62581309		grey
30	ABAT	9.17853242	1.05868127		grey

Table 4 gives the summary of the data for plotting the scatter plot of only the genes with an standard error of five or higher across the five cancers which are labeled.

❖ Scatter plot:

Only the genes with an SE of five or higher across the five cancers are labeled.

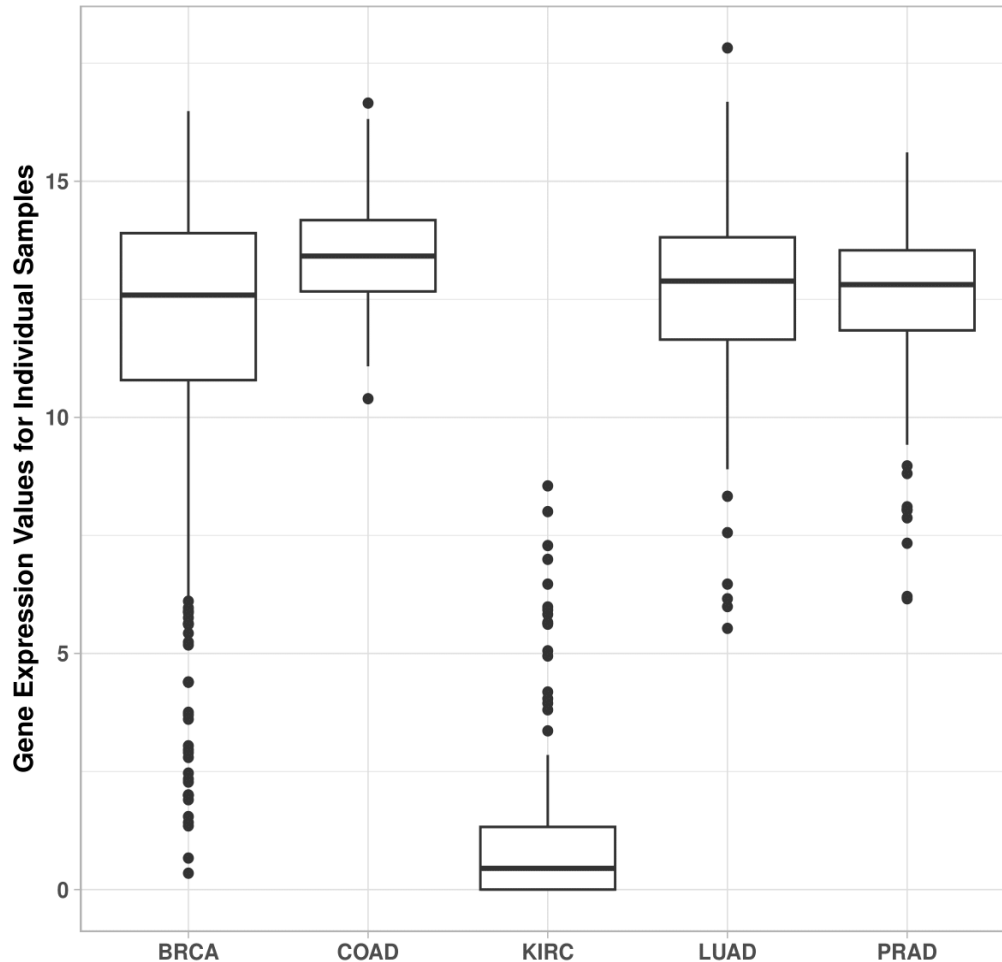


From the above plot, the gene which is high variability and low mean expression or, low variability and high mean expression or approximately same variability and mean expression is the gene which has the least expression.

The gene, which has high expression end and high variability is discussed. The most important gene are KLK3, KLK2, SFTPb, CEACAM5, CEACAM6, MSMB, AGR2, CDHR5, NPY, KLK4, USH1C etc.

❖ Box plot:

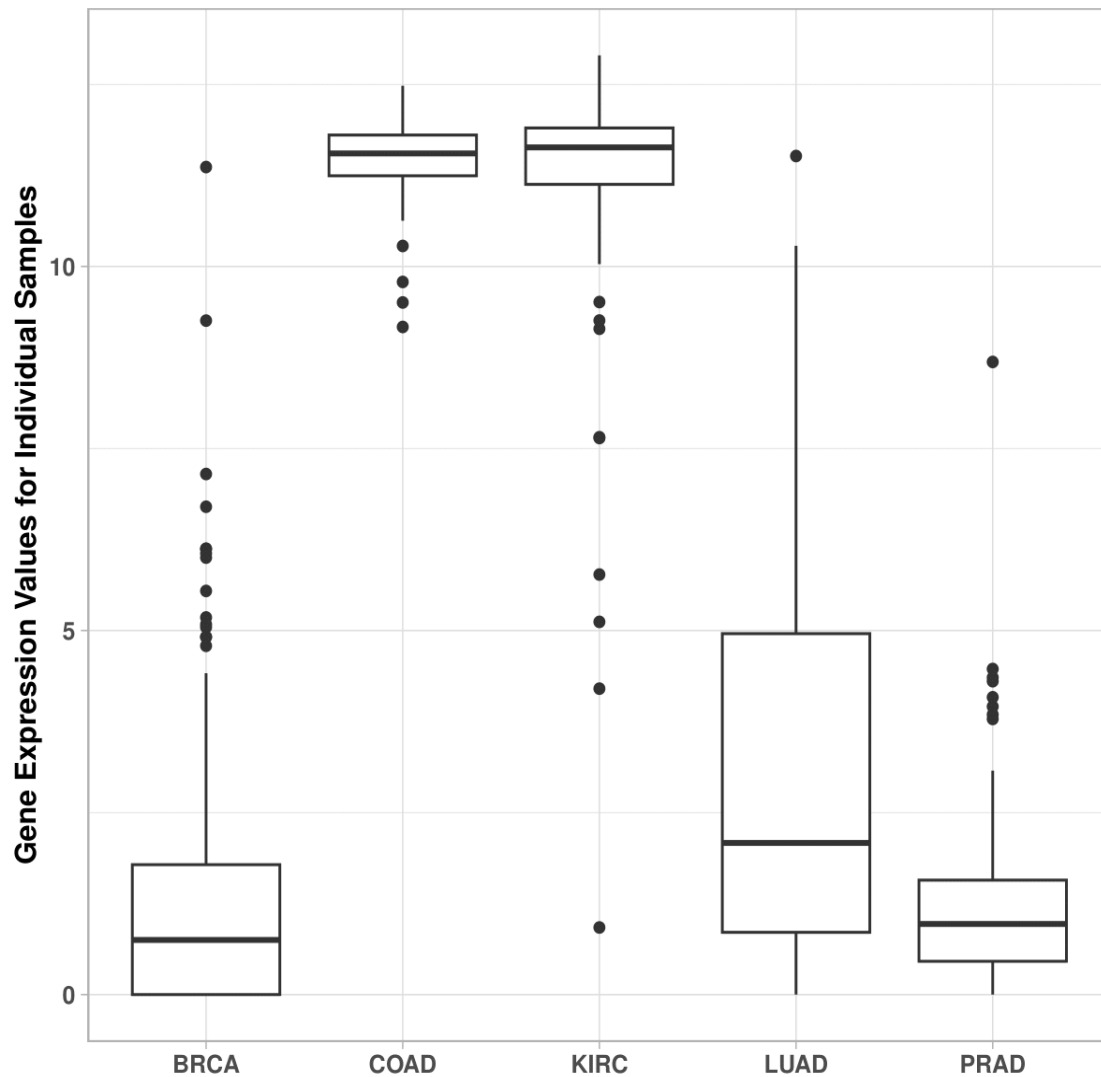
Cross-cancer summary of the expression of AGR2.



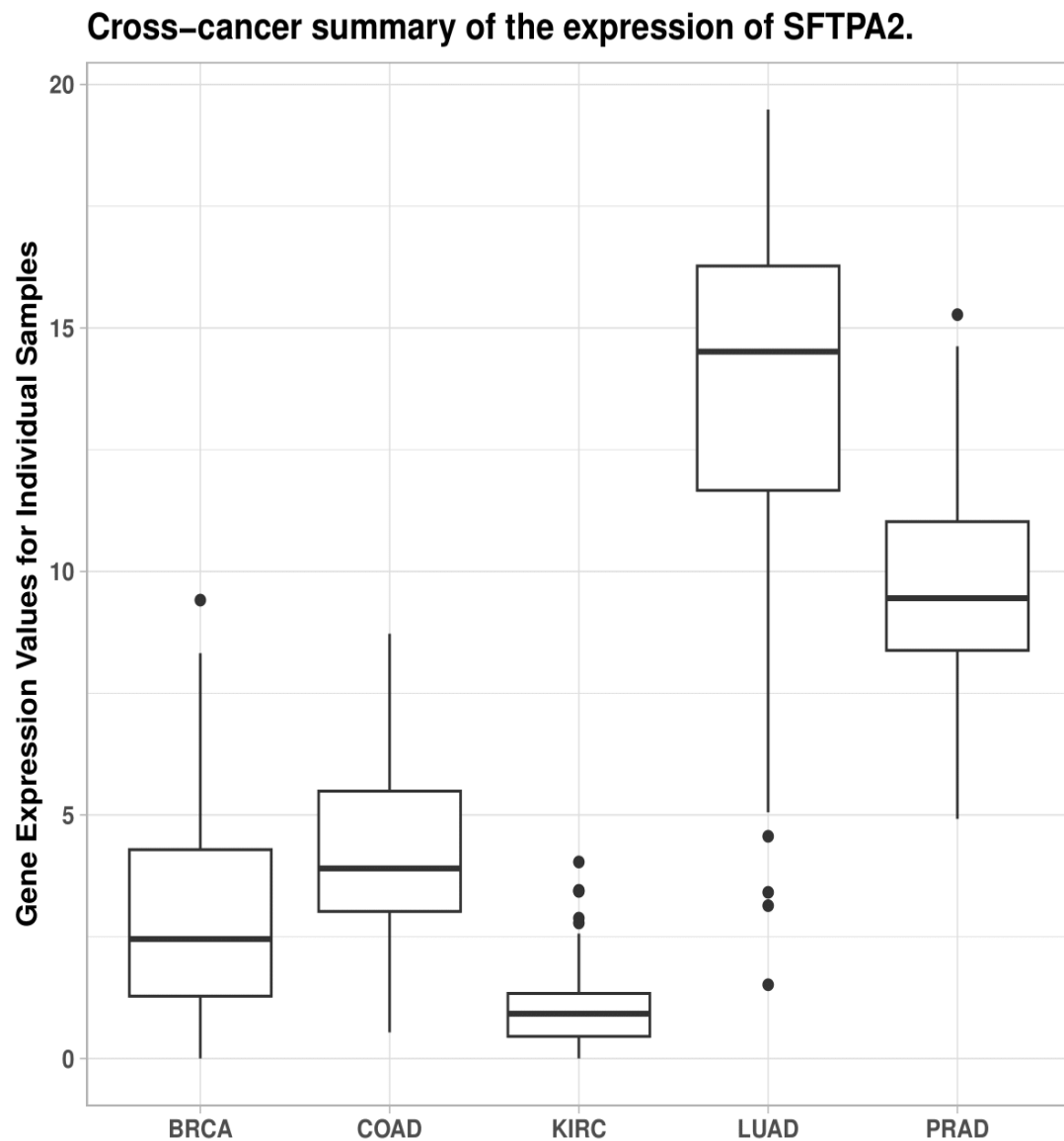
This gene encodes a member of the protein disulfide isomerase (PDI) family, which are endoplasmic reticulum (ER) proteins involved in catalyzing protein folding and thiol-disulfide interchange reactions. As an ER-localized molecular chaperone, it is crucial for the folding, trafficking, and assembly of cysteine-rich transmembrane receptors and the cysteine-rich intestinal glycoprotein mucin. This gene has been linked to inflammatory bowel disease and cancer progression in previous studies.

In the present study, a high expression of AGR2 gene is observed in all the adenocarcinomas but substantially low expression is observed in the renal cell carcinoma. AGR 2 is expressed and secreted during pancreatic cancer development and plays an important role in cancer cell growth and survival. Increased AGR2 expression is a valuable prognostic factor to predict the clinical outcome of the prostate cancer patients

Cross-cancer summary of the expression of USH1C.



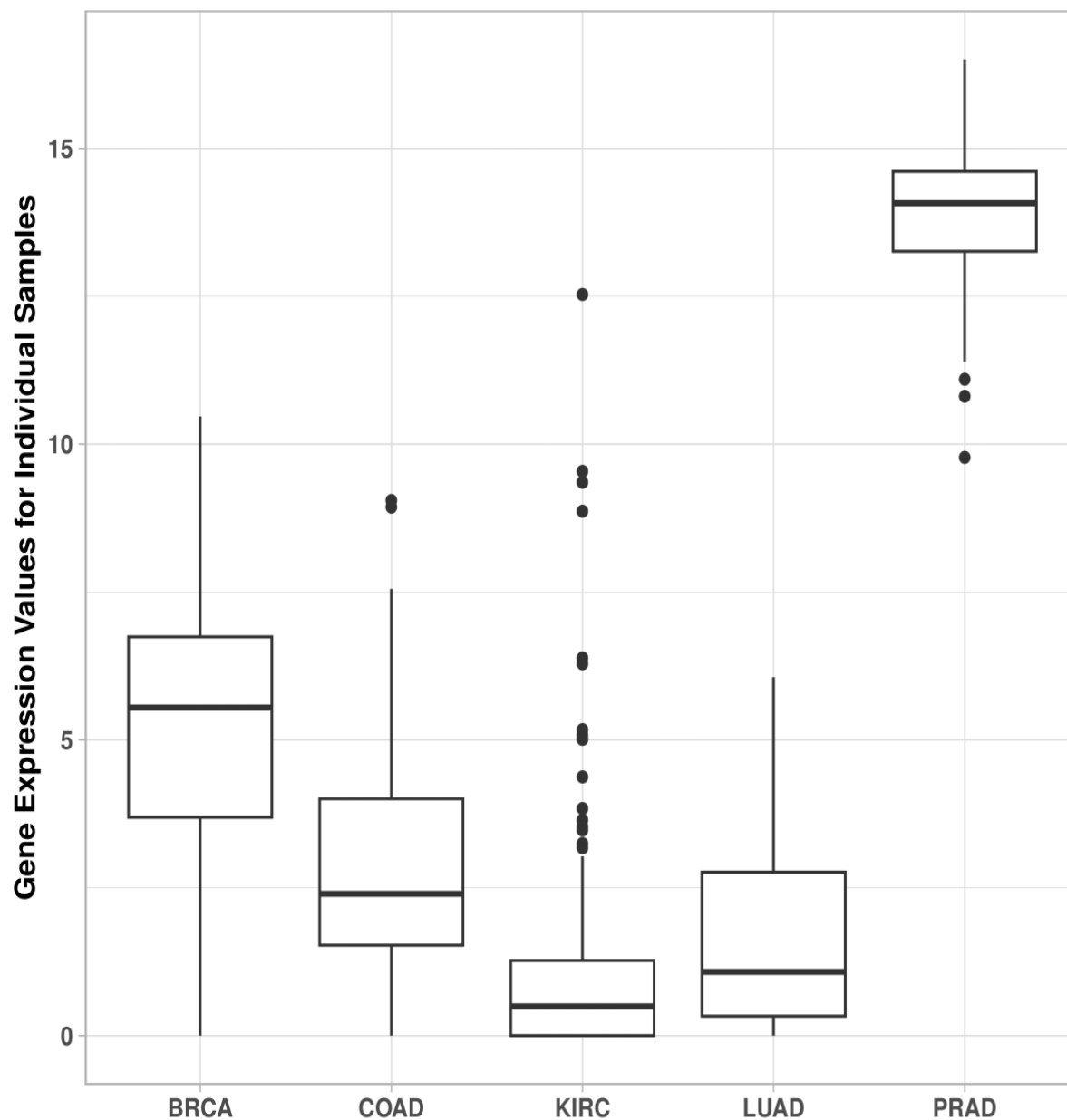
USH1C (USH1 Protein Network Component Harmonin) is a Protein Coding gene for the protein Harmonin. The USH1C gene is one of several genes encoding pulmonary-surfactant associated proteins (SFTPA) located on chromosome 11. Mutations in this gene and a highly similar gene located nearby, which affect the highly conserved carbohydrate recognition domain, are associated with idiopathic pulmonary fibrosis. This is in line with the figure above – the box plot of the gene shows high expression in the pulmonary cancer (LUAD), but much lower expression in all the other cancers.



Mutations in the genes encoding Surfactant protein (SP-A) (SFTPA2) have been associated with the phenotypes of pulmonary fibrosis and lung cancer in adults. SFTPA2 mRNA expression has also been detected in the trachea, prostate, pancreas, thymus, colon, eye, salivary gland and other tissues.

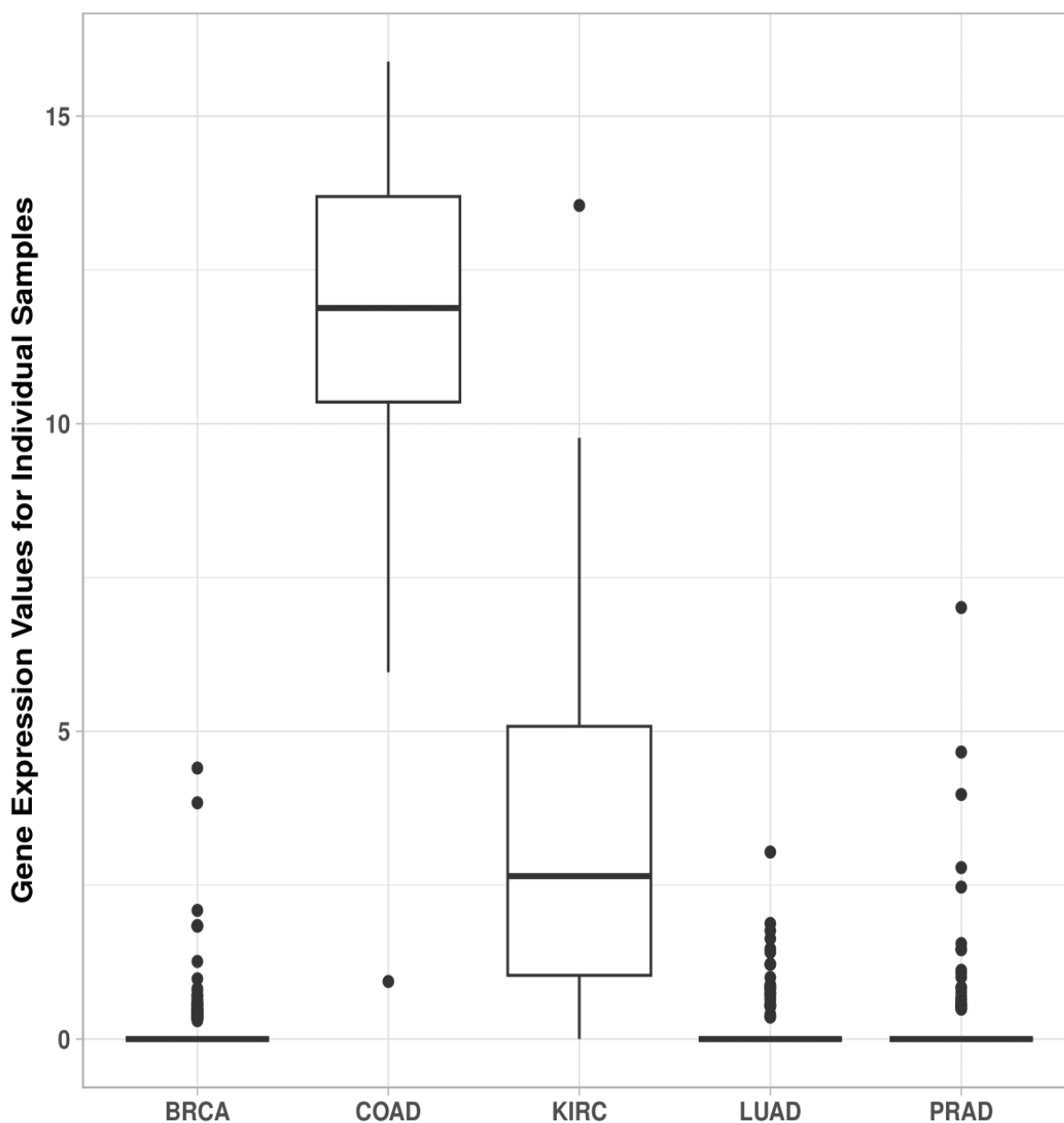
The Human Proteome Atlas shows that in the TCGA dataset, this protein was found to be highly expressed in colon and renal cancers. The gene-level findings are exactly the same, as can be seen in the above figure.

Cross-cancer summary of the expression of KLK4.



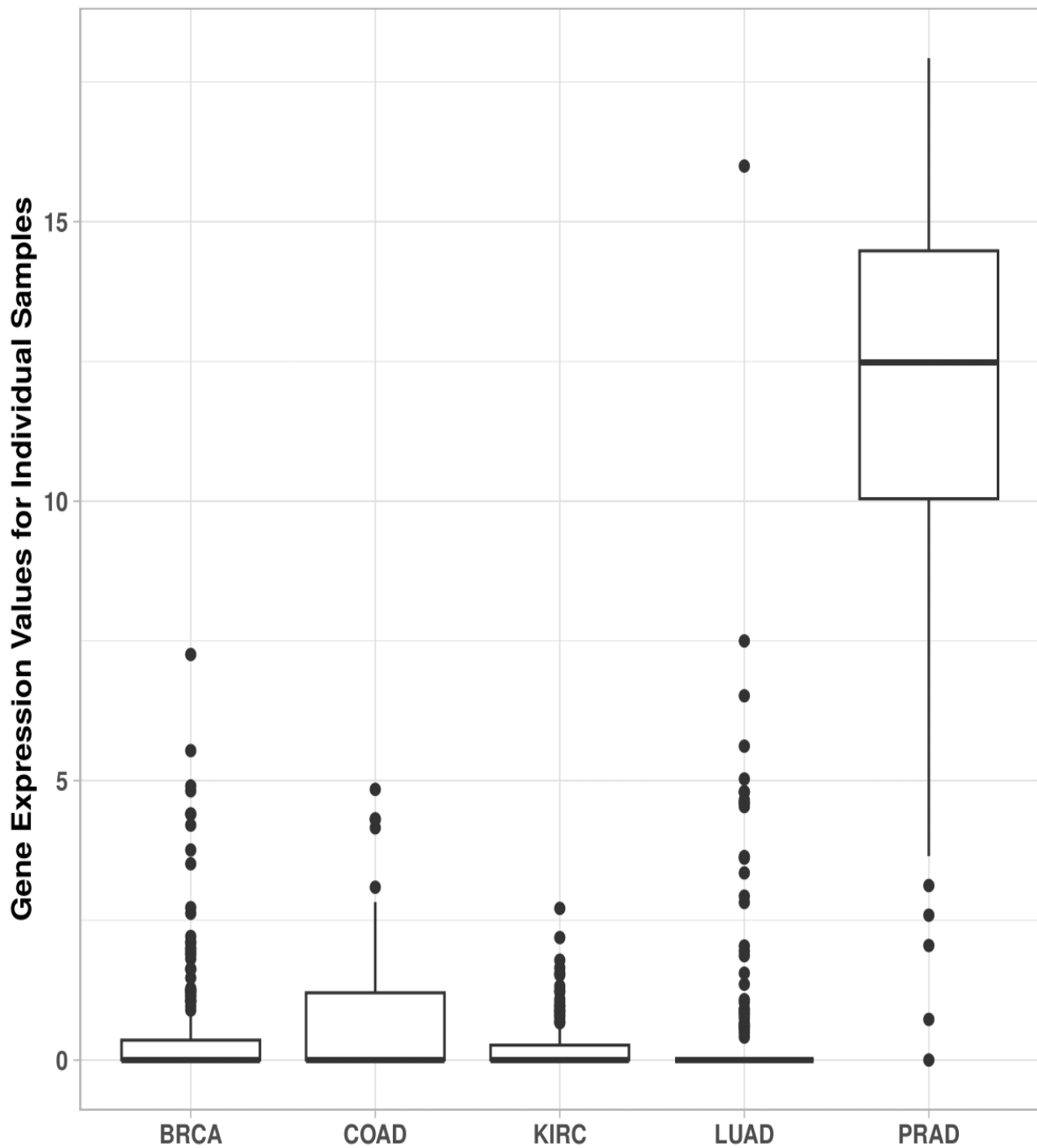
Kallikrein-related peptidase 4 (KLK4), also known as prostase/KLK-L1, is a serine protease gene implicated in the development and progression of certain cancers. KLK4 is highly expressed in the prostate epithelial cells of both premalignant and malignant lesions, suggesting a role in prostate cancer biology. It can activate ERK1/2 signaling in prostate cancer cell lines and induce cancer-associated fibroblast characteristics in prostate-derived stromal cells. In our cohort, it is clearly showing highly outlying expression in the prostate samples compared to any other cancer.

Cross-cancer summary of the expression of FABP1.



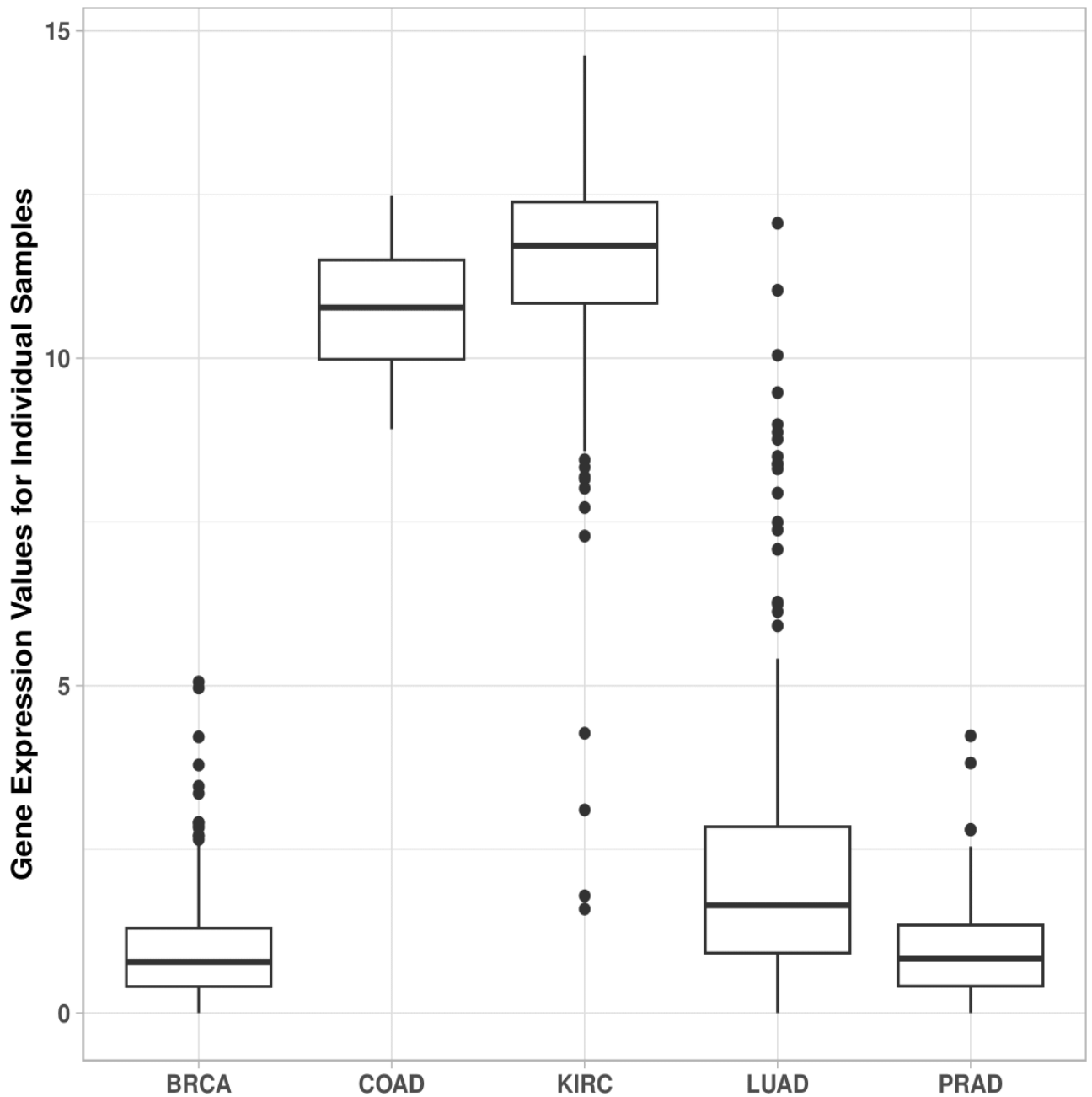
Previous studies have identified links between bile acids (BAs) and colorectal carcinogenesis in animal models. Exposure of the colonic epithelium to BAs can promote carcinogenesis through various mechanisms, including oxidative stress with DNA damage/genomic instability, apoptosis, epigenetic changes, and alterations in the gut microbiota. These findings, along with the observation that BAs significantly increase FABP expression, prompted the investigation of genes associated with BA homeostasis in the development of colorectal cancer. Expression of FABP mRNA was found to be significantly elevated in subjects with colorectal tumour. In the dataset too, the same pattern is observed where colorectal samples exhibit extremely high expression of the gene.

Cross-cancer summary of the expression of NPY.



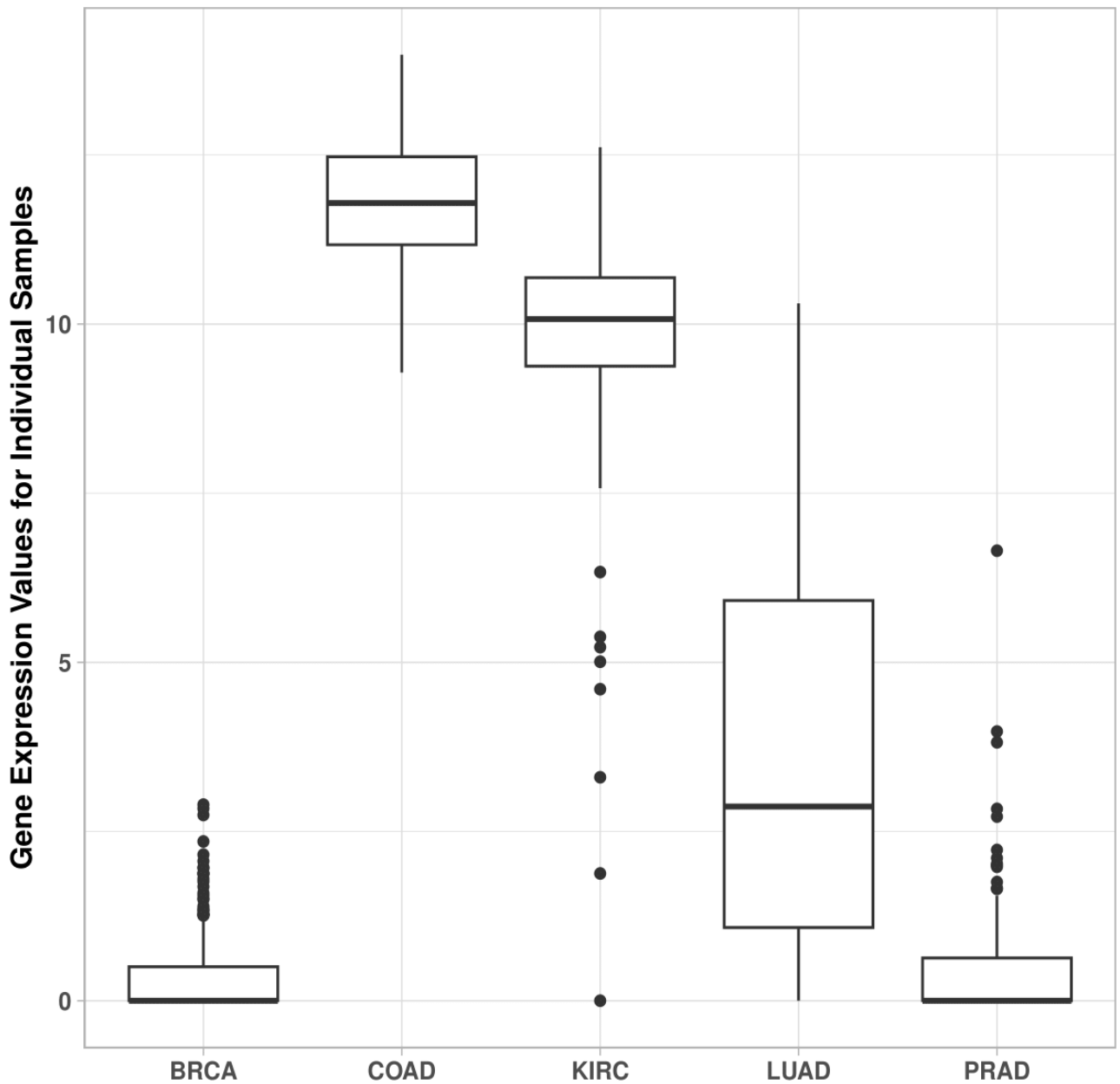
Neuropeptide Y (NPY) is a pleiotropic gene implicated in stress resilience and is associated with higher levels of conscientiousness. Along with environmental factors such as stressful life events, this gene may be a factor in the neurobiology of human personality. The expression of the NPY gene varies significantly across different cancer types, with PRAD showing the highest median expression and variability, while KIRC exhibits the lowest median expression with fewer outliers.

Cross-cancer summary of the expression of CDHR5.



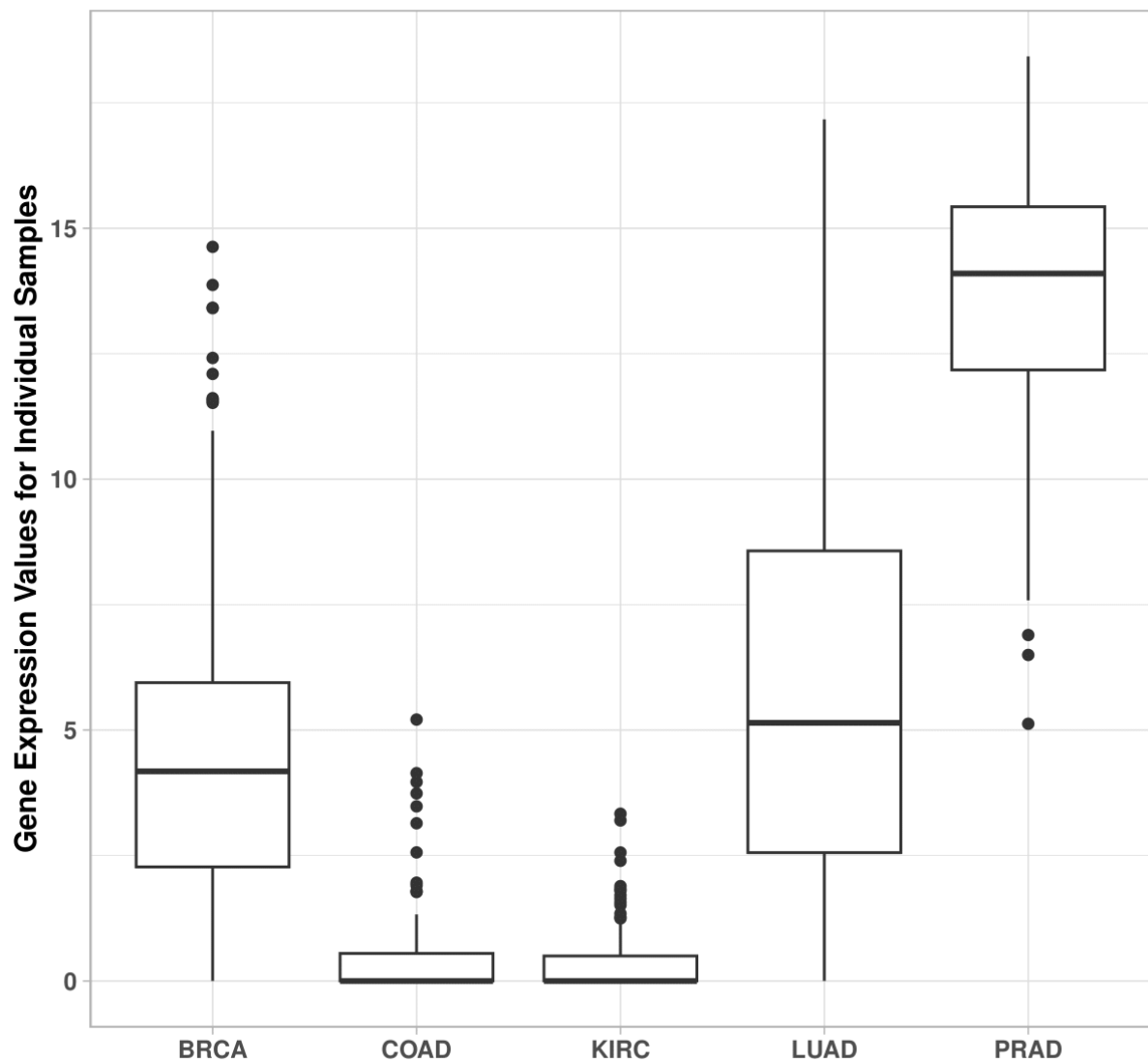
The expression of the CDHR5 gene varies across different cancer types. COAD and KIRC show relatively high median expression levels with broad variability. In contrast, BRCA and PRAD have low median expression levels with more consistency among samples, except for some outliers. LUAD has a moderate median expression level with a narrower range of variability within the middle 50% of samples but more outliers compared to PRAD.

Cross-cancer summary of the expression of HNF4A.



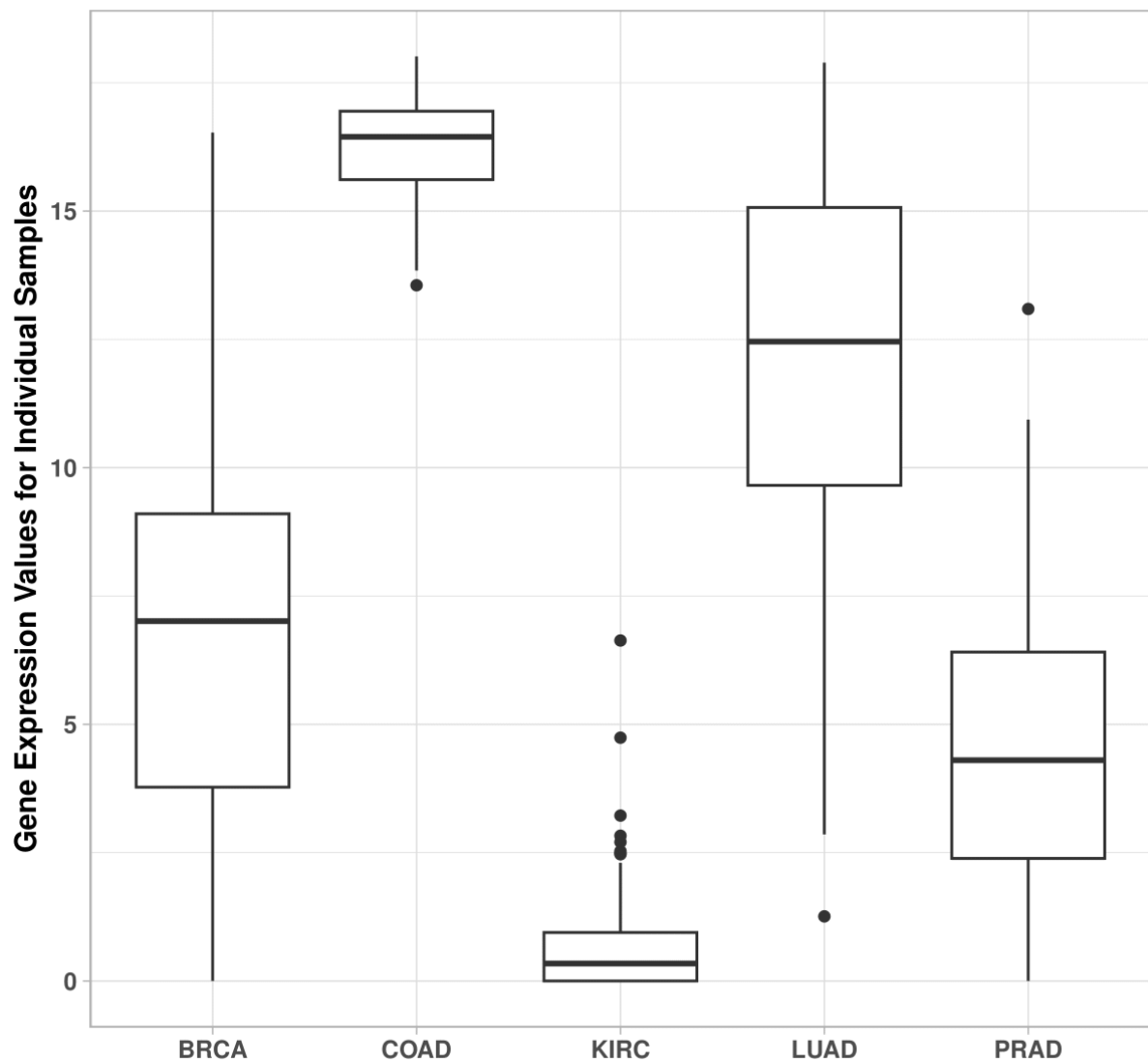
Hepatocyte Nuclear Factor 4alpha (HNF4alpha) Plays a Controlling Role in Expression of the Retinoic Acid Receptor beta (RARbeta) Gene in Hepatocytes. The expression of the HNF4A gene varies significantly across different cancer types. COAD and KIRC show relatively high median expression levels with broad variability, while BRCA and PRAD have low median expression levels with more consistency among samples, except for some outliers. LUAD has a moderate median expression level with a wide range of variability within the middle 50% of samples and several outliers.

Cross-cancer summary of the expression of MSMB.



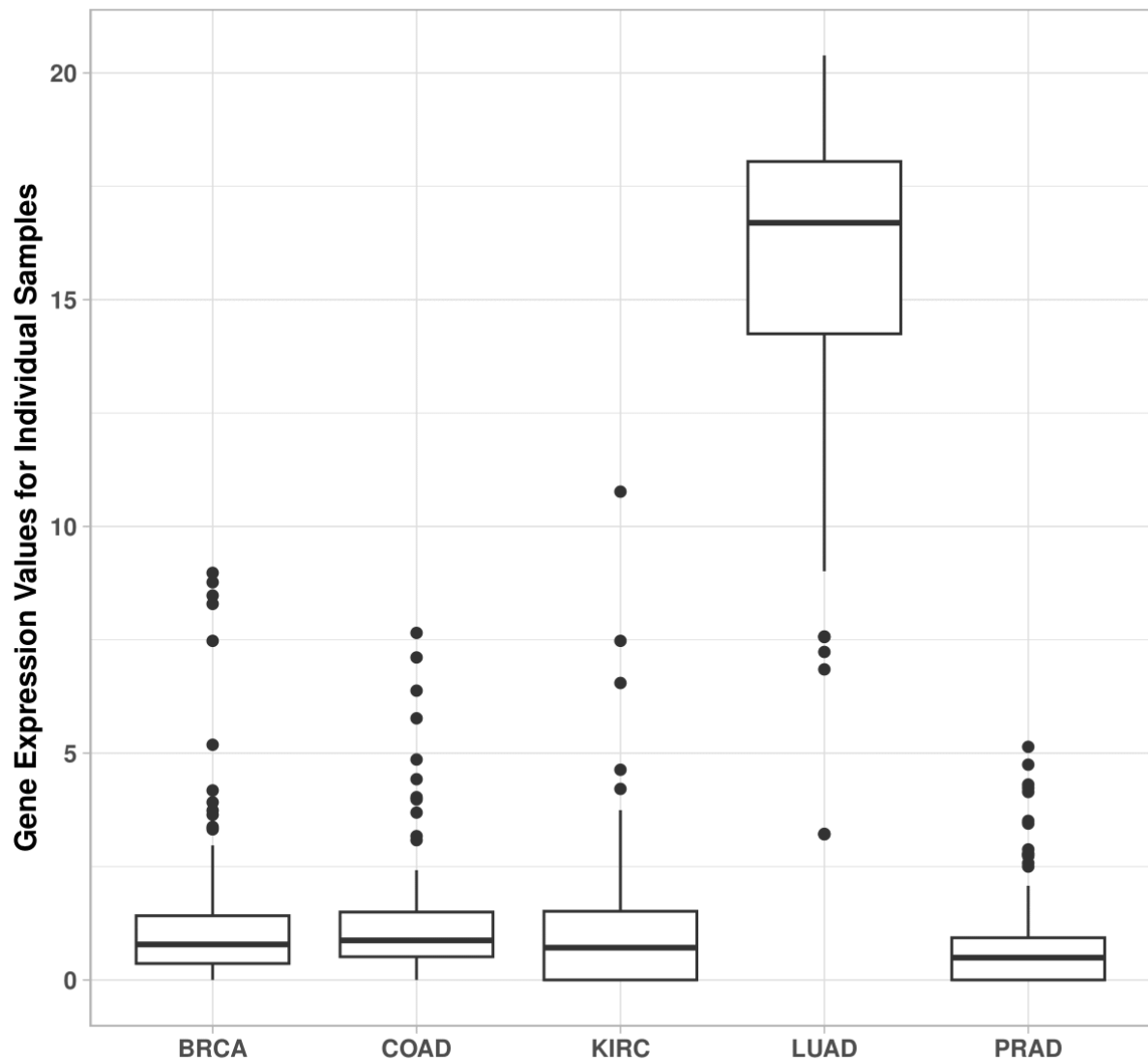
Microseminoprotein-beta (MSMB) is an abundant secretory protein contributed by the prostate, and is implicated as a prostate cancer (PC) biomarker based on observations of its lower expression in cancerous cells compared with benign prostate epithelium. The expression of MSMB varies significantly across different cancer types. BRCA, LUAD, and PRAD show higher and more variable expression levels, while COAD and KIRC exhibit lower and less variable expression levels.

Cross-cancer summary of the expression of CEACAM5.



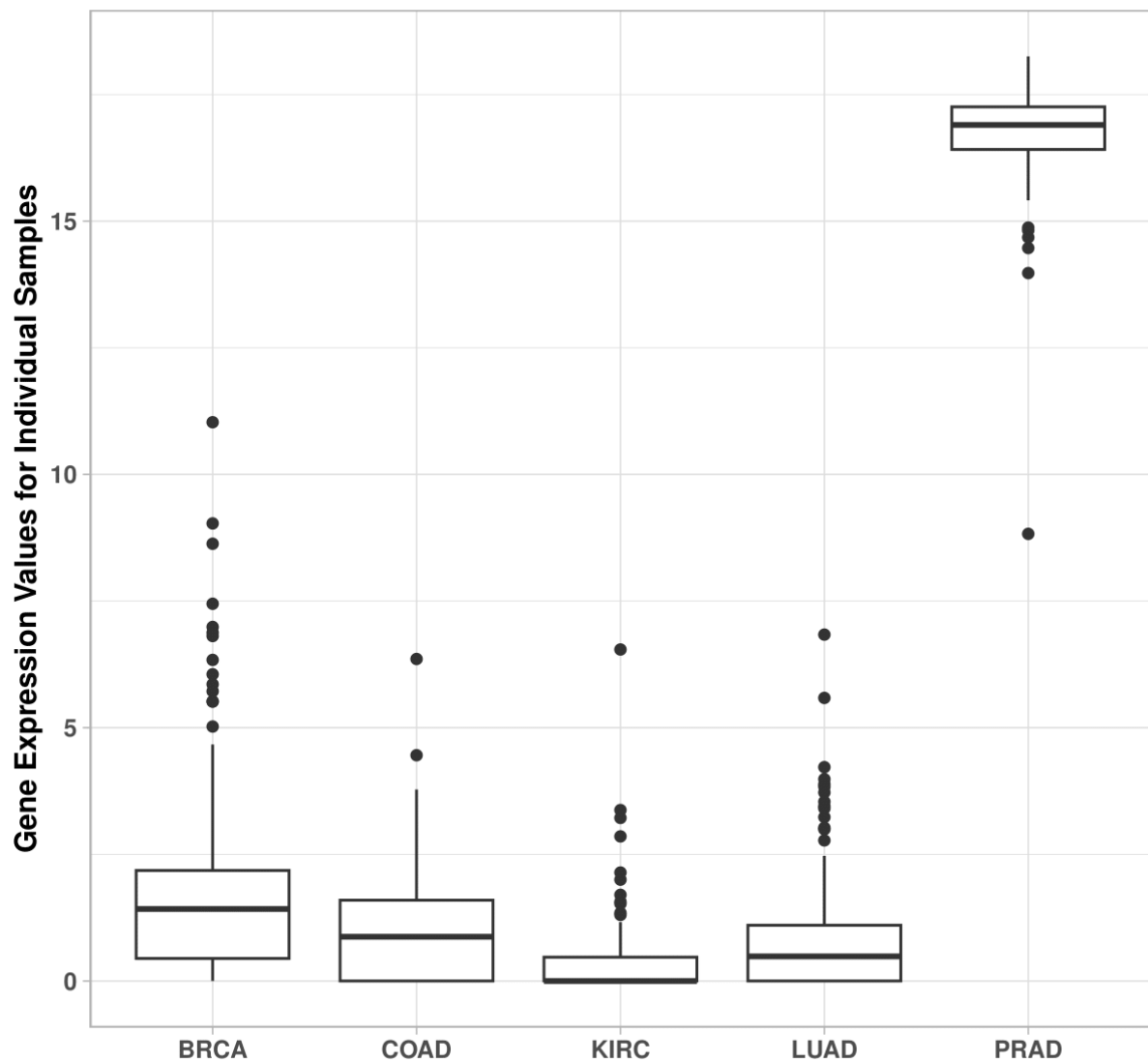
CEACAM5 is often used as a synonym for cancer embryonic antigen (CEA), a well-known biomarker of many types of malignancies, such as colorectal cancer and non-small-cell lung cancer. Its primary function in the embryonic intestine and colon tumors is adhesion between epithelial cells.

Cross-cancer summary of the expression of SFTPB.



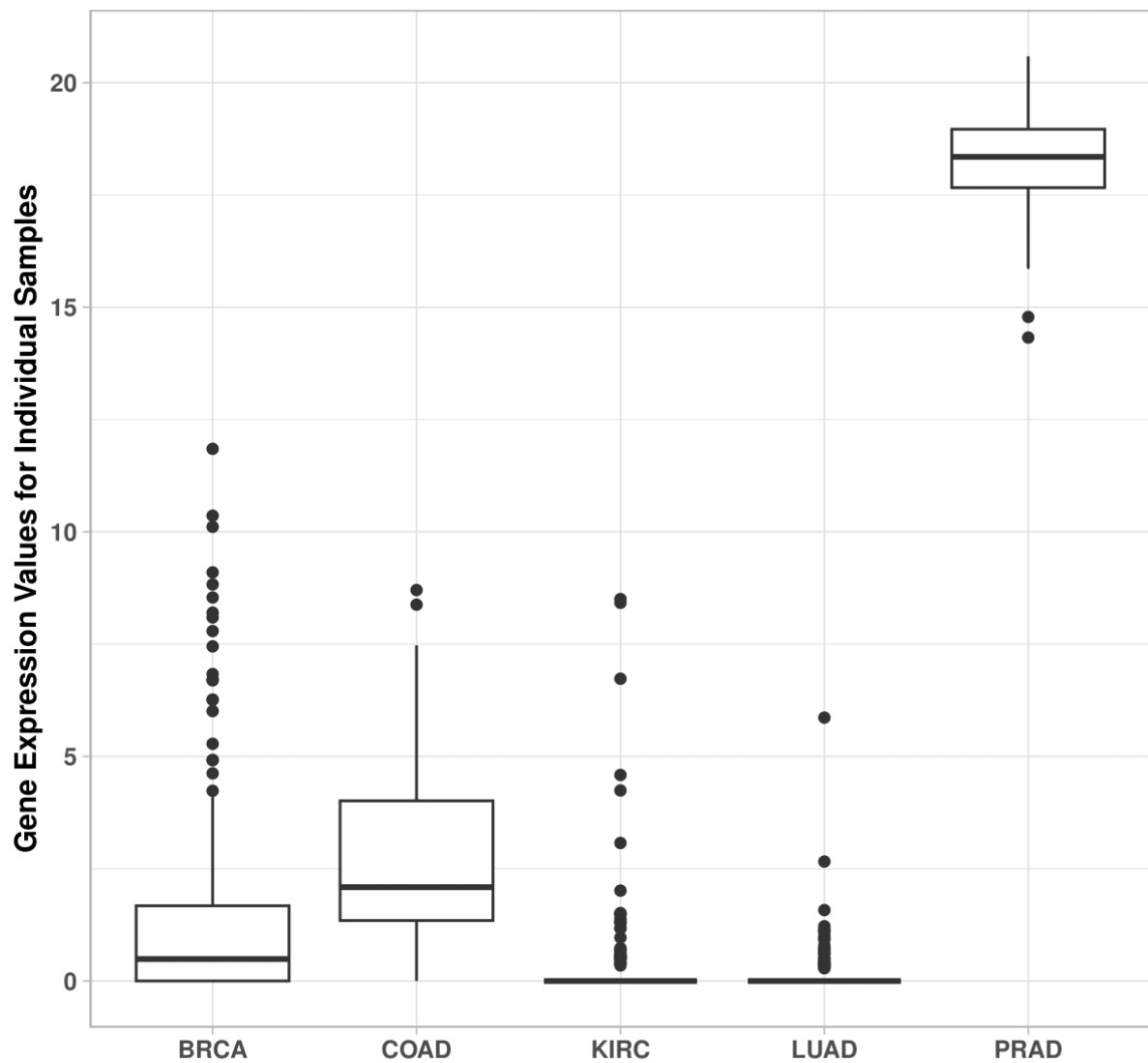
The SFTPB gene expression is significantly higher in lung adenocarcinoma (LUAD) compared to other types of cancer, suggesting its potential diagnostic or therapeutic relevance specifically for lung cancer.

Cross-cancer summary of the expression of KLK2.



The kallikrein related peptidase 2 (KLK2) gene expression is significantly higher in prostate cancer (PRAD) compared to other types of cancer, similar to KLK3, suggesting its potential diagnostic or therapeutic relevance specifically for prostate cancer.

Cross-cancer summary of the expression of KLK3.



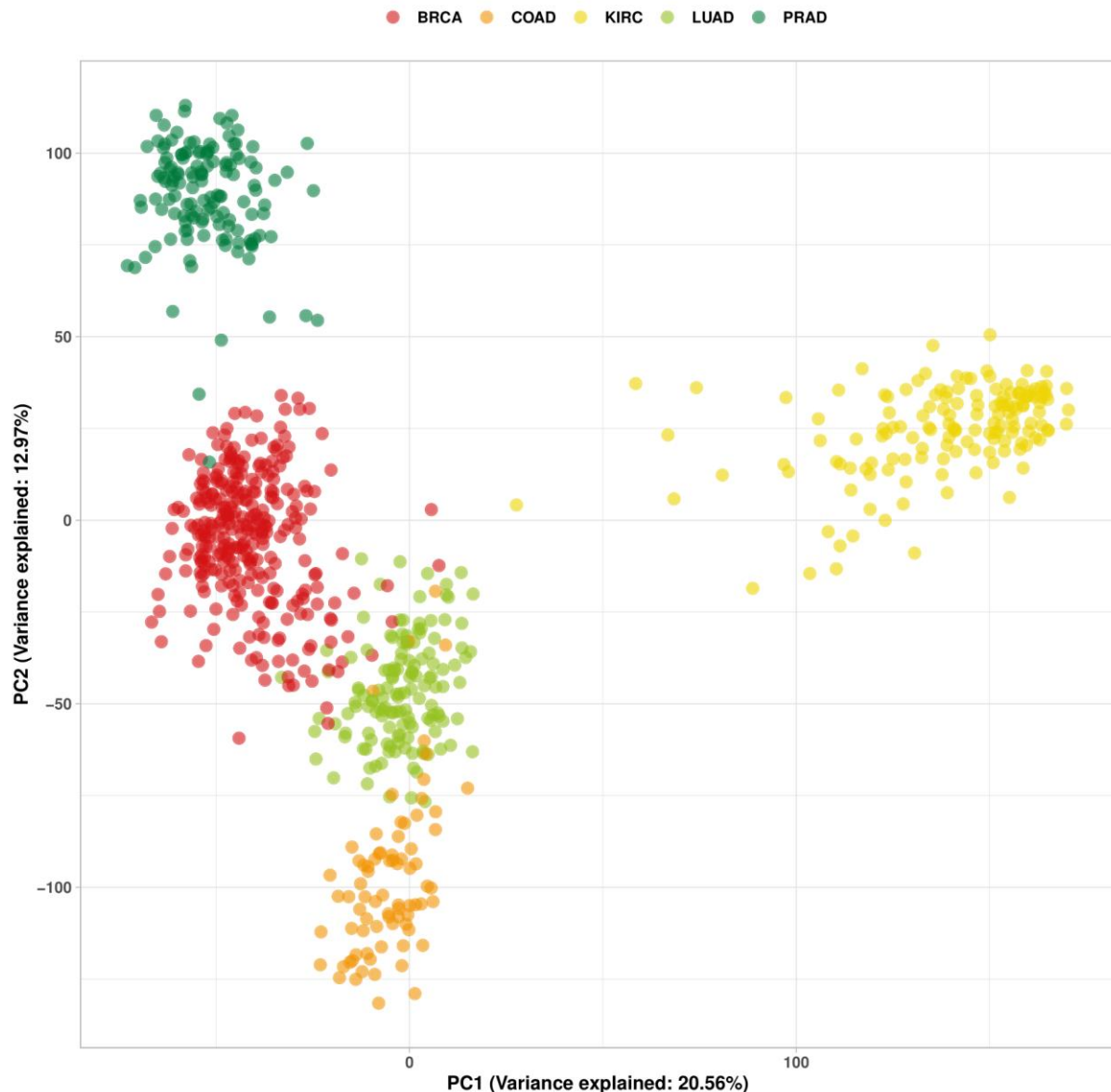
The Kallikrein Related Peptidase 3 (KLK3) gene expression is significantly higher in prostate cancer (PRAD) compared to other types of cancer, suggesting a potential diagnostic or therapeutic relevance specifically for prostate cancer.

Table 5: Distribution of the principal component analysis of the data

	id	cancer	PC1	PC2
1	sample_0	PRAD	-57.5167564	8.606018e+01
2	sample_1	LUAD	-5.1133229	-7.536531e+01
3	sample_10	BRCA	-53.8303246	1.657849e+01
4	sample_100	BRCA	-43.9995608	-1.814553e+00
5	sample_101	KIRC	155.4817799	3.475200e+01
6	sample_102	BRCA	-48.9900642	5.773638e+00
7	sample_103	KIRC	160.6893834	2.645022e+01
8	sample_104	LUAD	-8.6236991	-5.221154e+01
9	sample_105	KIRC	156.1044447	3.088473e+01
10	sample_106	LUAD	8.0663249	-6.233632e+01
11	sample_107	COAD	-7.7889418	-9.079837e+01
12	sample_108	LUAD	-14.0485133	-4.961879e+01
13	sample_109	LUAD	-4.4205389	-4.301958e+01
14	sample_11	KIRC	160.1256238	2.423920e+01
15	sample_110	PRAD	-60.7131401	8.356430e+01
16	sample_111	BRCA	-44.1051858	1.614986e+01
17	sample_112	LUAD	-21.0289423	-4.126293e+01
18	sample_113	PRAD	-55.2541340	8.418767e+01
19	sample_114	BRCA	-53.1682347	-2.788534e+00
20	sample_115	KIRC	169.9164358	3.590354e+01
21	sample_116	KIRC	158.6779979	1.421874e+01
22	sample_117	KIRC	150.0736475	3.920479e+01
23	sample_118	BRCA	-34.7355793	-1.176789e+01
24	sample_119	BRCA	-48.4600161	-7.818352e+00
25	sample_12	PRAD	-40.7642620	8.336346e+01
26	sample_120	LUAD	10.6173729	-6.130608e+01
27	sample_121	KIRC	155.1396158	6.222798e+00
28	sample_122	LUAD	-15.9312046	-5.267828e+01
29	sample_123	BRCA	-54.5117679	-4.031309e+00
30	sample_124	PRAD	-53.9318331	9.487088e+01

Table 5 shows the principal component analysis of the data that is calculated for all the cancers.

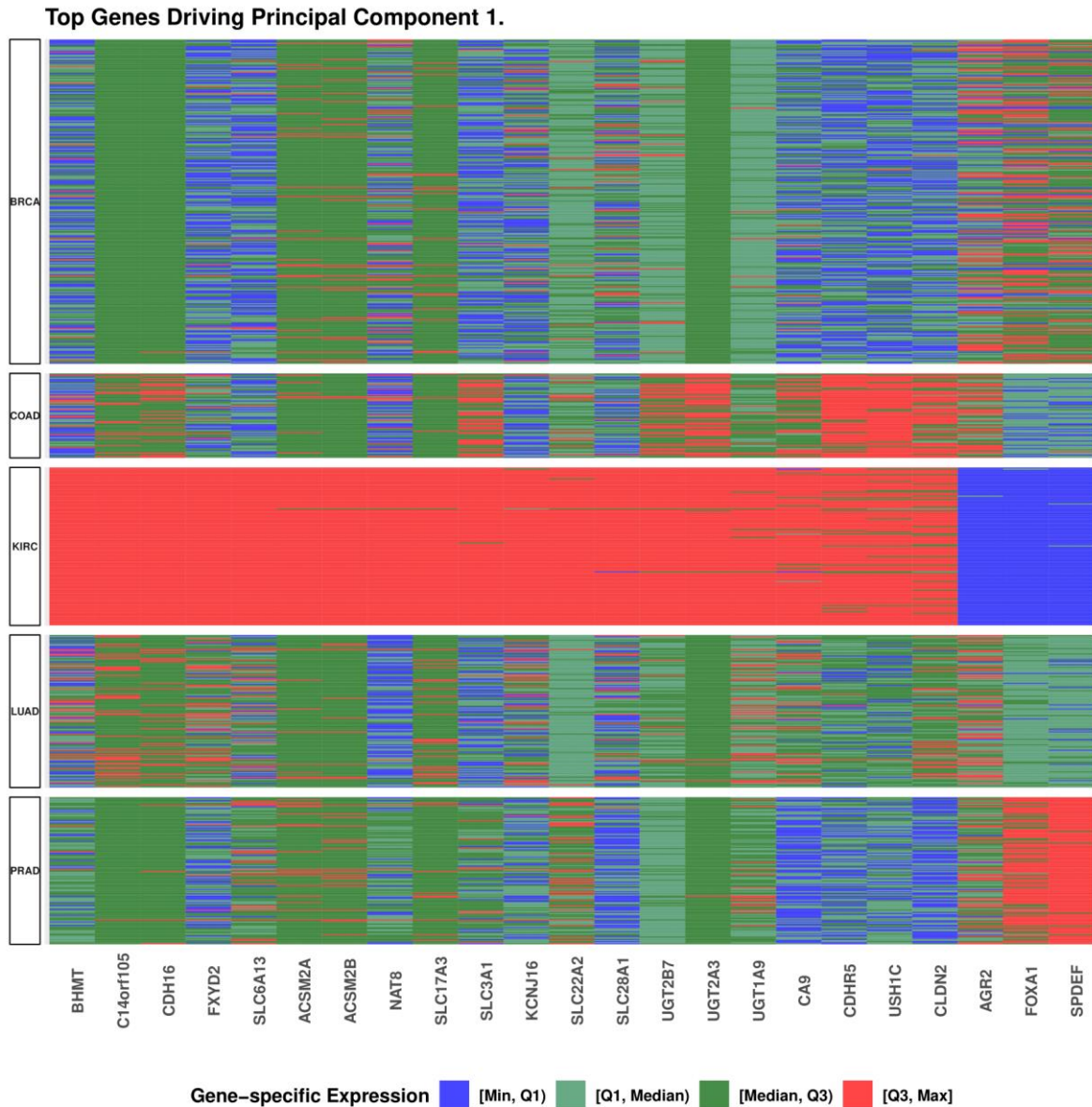
❖ Scatter plot:



From the above scatter plot it is observed, five cancer represented by the five colour respectively. X axis represents the principal component 1 which variance 20.56% explained. Y axis represents the principal component 2 which is variance 12.9% explained. Every data points is a sample.

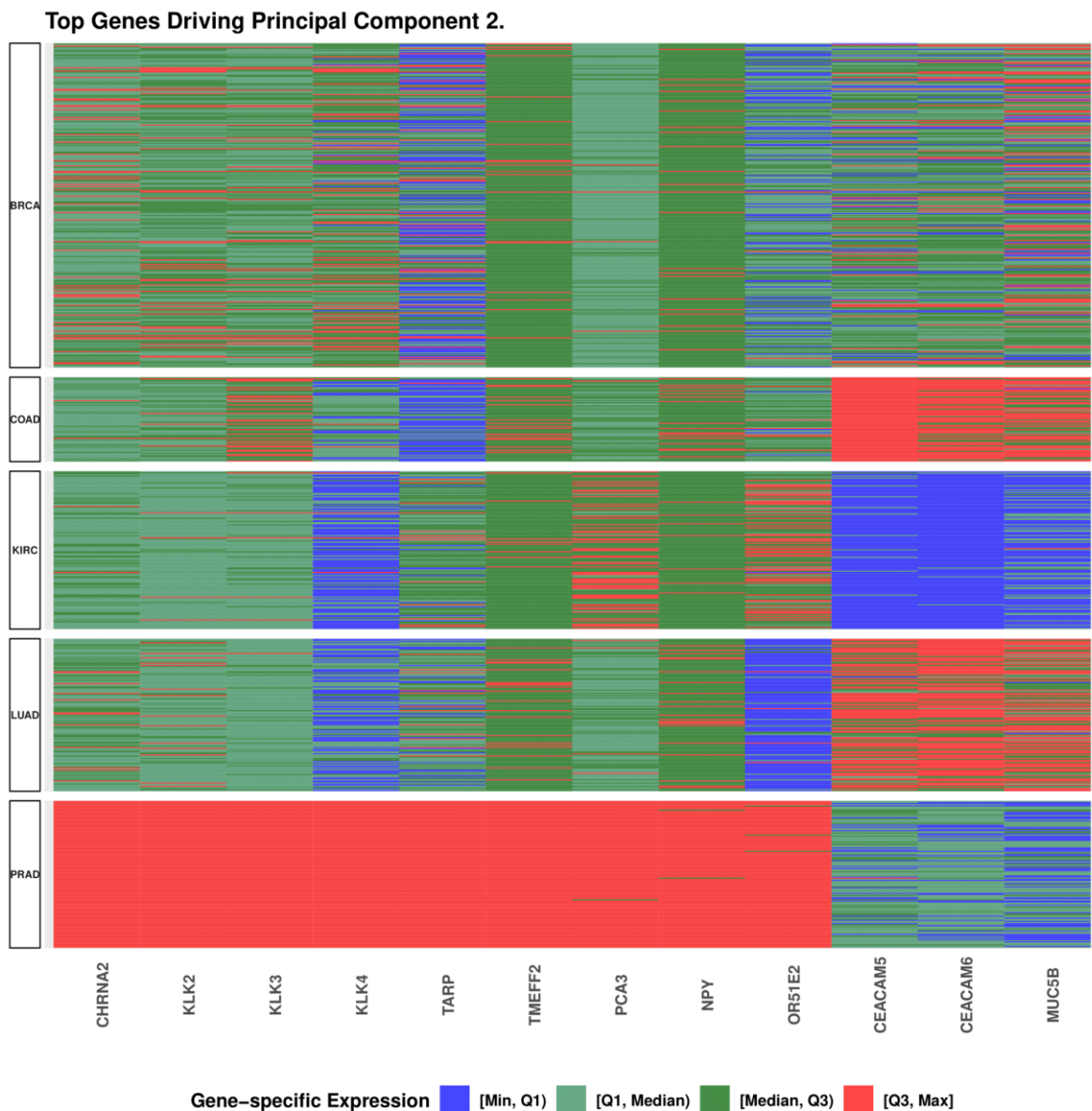
Also, we can see that the two cancer KIRC (kidney renal clear cell carcinoma), PRAD (prostate adenocarcinoma) are separate from another cancer. The cancer BRCA (breast adenocarcinoma), COAD (colon adenocarcinoma), and Lung adenocarcinoma (LUAD) are similar type biologically. The principal component 1 separate the cancer KIRC from other four cancer and principal component 2 separate the cancer PRAD from other four cancer.

❖ Heat map of principal Component 1



Kidney renal clear cell carcinoma (KIRC) is distinguished from the rest of the cancers by a cluster of 20 genes which are consistently strongly expressed in this cancer but not as strong in the others. On the other hand, the genes AGR2, FOXA1, and SPDEF are exhibiting low to no expression in the KIRC samples but relatively higher expression levels in the other cancers – specifically PRAD. These 23 genes have a profound effect on the first principal component, which results in the separation of KIRC from the other samples in the PCA scatterplot.

❖ Heat map of principal Component 2



PRAD is differentiated from the rest of the cancers by a cluster of 9 genes that have strong positive expression, but have average to low expression in the other cancers. These genes notably include the Human kallikrein-related peptidases (KLK) class of genes, are known to be functional agents of relevance in prostate cancer. The analysis of KLK panels in large sets of samples from diverse stages of the disease, including premalignant phases, will probably help to reveal how the expression profile evolves during the course of the disease.

6. Conclusion

World Health Organization (WHO) estimates cancer, as a global burden in India with India expecting more than 14.1 lakh new cancer cases in 2024. Cancer is a genetic disease, as initiation, progression, and metastasis are governed by several genetic and epigenetic changes within the genome. Different types of cancer are caused by different onco-agents, with varying causal interactions.

This study underscores such variability across cancers by assessing gene expression levels in Indian cancer patients from five distinct cancer types. It is clearly observed that differential patterns in specific genes and cluster of genes make certain tissue-specific cancers different from the others (such as adenocarcinomas being different from other types of carcinomas). Pan-cancer analysis (PCA) allows the inclusion of markers from various cancer types in a single assay, positively impacting workflows and turn around times.

Using statistical techniques like PCA and clustering, dimension reduction was performed in high-dimensional expression dataset to identify key genes and gene groups of interest. Modern visualization techniques was used to illustrate such differences. As a future direction, it would be interesting to explore the same comparisons in an integrative setting – combining data from other sources such as copy number alterations and protein expressions alongside gene expressions.

The results from the present study provide a comprehensive transcriptional architecture of the cancer matrisome and suggest the need for development of specific matrisome-targeting approaches for future therapies. Specific patterns of gene expression that is analysed, may define tumor subtypes and support a more targeted approach to therapy hence saving lives of cancer patients.

7. References

Data procurement. <https://www.kaggle.com/datasets/shibumohapatra/icmr-data/code>.

Rashmi, Richa and Majumdar, Sharmistha. 2022. Pan-Cancer Analysis Reveals the Prognostic Potential of the THAP9/THAP9-AS1 Sense--Antisense Gene Pair in Human Cancers. *Non-coding RNA* 8(4): 51

Kumar, Rajesh and Patiyal, Sumeet and Kumar, Vinod and Nagpal, Gandharva and Raghava. 2019. In silico analysis of gene expression change associated with copy number of enhancers in pancreatic adenocarcinoma. *International journal of molecular sciences* 20 (14):3582.

Fort, Rafael Sebasti, and Duhagon, Ana. 2021. Pan-cancer chromatin analysis of the human vtRNA genes uncovers their association with cancer biology. *F1000Research* 10:

Varn, F. S., Wang, Y., Mullins, D. W., Fiering, S., & Cheng, C. (2018). Systematic pan-cancer analysis reveals immune cell interactions in the tumor microenvironment. *Cancer Research*, 77(6), 1271. <https://doi.org/10.1158/0008-5472.CAN-16-2490>