

Original Article

Immune infiltration and prognostic and diagnostic use of LGALS4 in colon adenocarcinoma and bladder urothelial carcinoma

Animesh Acharjee^{1,2,3}, Prasoon Agarwal^{4,5}, Katrina Nash⁶, Subia Bano⁷, Taufiq Rahman⁸, Georgios V Gkoutos^{1,2,3,9,10,11}

¹College of Medical and Dental Sciences, Institute of Cancer and Genomic Sciences, University of Birmingham, B15 2TT, UK; ²Institute of Translational Medicine, University Hospitals Birmingham NHS, Foundation Trust, B15 2TT, UK; ³NIHR Surgical Reconstruction and Microbiology Research Centre, University Hospital Birmingham, Birmingham, B15 2WB, UK; ⁴KTH Royal Institute of Technology, School of Electrical Engineering and Computer Science, Stockholm, Sweden; ⁵Science for Life Laboratory, Solna, Sweden; ⁶College of Medical and Dental Sciences, University of Birmingham, Birmingham, B15 2TT, UK; ⁷Elvesys Microfluidic Innovation Centre, Paris 75011, France; ⁸Department of Pharmacology, Tennis Court Road, University of Cambridge, Cambridge, CB2 1PD; ⁹MRC Health Data Research UK (HDR UK); ¹⁰NIHR Experimental Cancer Medicine Centre, Birmingham, B15 2TT, UK; ¹¹NIHR Biomedical Research Centre, University Hospital Birmingham, Birmingham, B15 2TT, UK

Received March 7, 2021; Accepted June 3, 2021; Epub October 15, 2021; Published October 30, 2021

Abstract: Colon adenocarcinoma (COAD) is a common tumor of the gastrointestinal tract with a high mortality rate. Current research has identified many genes associated with immune infiltration that play a vital role in the development of COAD. In this study, we analysed the prognostic and diagnostic features of such immune-related genes in the context of colonic adenocarcinoma (COAD). We analysed 17 overlapping gene expression profiles of COAD and healthy samples obtained from TCGA-COAD and public single-cell sequencing resources, to identify potential therapeutic COAD targets. We evaluated the abundance of immune infiltration with those genes using the TIMER (Tumor Immune Estimation Resource) deconvolution method. Subsequently, we developed predictive and survival models to assess the prognostic value of these genes. The LGALS4 (Glectin-4) gene was found to be significantly ($P < 0.05$) downregulated in COAD and bladder urothelial carcinoma (BLCA) compared to healthy samples. We identified LGALS4 as a prognostic and diagnostic marker for multiple cancer types, including COAD and BLCA. Our analysis reveals a series of novel candidate drug targets, as well as candidate molecular markers, that may explain the pathogenesis of COAD and BLCA. LGALS4 gene is associated with multiple cancer types and is a possible prognostic, as well as diagnostic, marker of COAD and BLCA.

Keywords: Omics integration, translational research, immune infiltration, LGALS4, biomarker, BLCA

Introduction

Colon adenocarcinoma (COAD), or colon cancer, affects over 1.8 million people each year worldwide and ranks second in terms of cancer-associated mortality. In 80% of reported cases, colorectal carcinomas develop from benign colonic adenomas. Since the survival rate of COAD decreases according to cancer stage at diagnosis, early detection and removal of these premalignant adenomas is crucial.

Numerous studies have demonstrated that the pathogenesis of COAD is closely associated

with mutations and altered epigenetic expression of particular genes. It is widely accepted that specific oncogenes (e.g. Ras, EGFR (Erb-B1), Erb-B2, TGF α and TGF- β 1) and tumor suppressor genes (e.g. APC, P53, p27, MSI, LH 18q) are implicated in the development of colorectal cancer [1]. However, immune infiltration within the tumor microenvironment is also thought to participate in a complex interplay with malignant cells, potentially through bidirectional modulation of gene expression [2]. In recent studies, omics-based technologies, including RNA sequencing, microarrays, proteomics and metabolomics have been used

Table 1. Datasets employed in this study

Dataset	Number of samples	Technology/Platform	Reference
TCGA-COAD	COAD = 459; Normal = 41	RNA sequence	[7]
TCGA-BLCA	BLCA = 408; Normal = 19	RNA sequence	[7]
GSE71187	COAD = 46; Normal = 12	GPL6480, Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	[39]
GSE13507	Primary bladder cancer = 165; Normal bladder mucosa = 9	GPL6102, Illumina human-6 v2.0 expression beadchip	[40]

extensively and were extremely useful for the identification of novel diagnostic and prognostic COAD markers. Ge et al. recently identified immune cell infiltration, within the COAD micro-environment, as paramount to tumor growth and progression, in addition to immune-related gene (e.g. IL11, EREG, IL17C, and CALCA) mutations, which were significantly related to prognosis of the disease [3]. Peltekova et al. also identified the COLCA1 and COLCA2 genes as implicated in the pathogenesis of colon cancer through the involvement of immune pathways [4].

Furthermore, the tumor microenvironment may alter the response to treatment and play a role in therapeutic approaches. Wang et al. identified multiple epigenetic pathways, most notably the PPAR signalling pathway [2], that were altered in expression within an inflamed tumor microenvironment, directly affecting responsiveness to chemotherapy agents. The PPAR gene induces growth inhibition and is expressed in multiple cancer types, including colon, prostate, breast, and gastric cancer [5].

In this study, we used TCGA-COAD and public single-cell sequencing datasets and identified 17 genes common to COAD and normal samples. We further linked those genes with the abundance of immune infiltration of those genes using TIMER [6]. Subsequently, we performed an overall Cox survival analysis to identify the prognostic potential of those genes. We identified particular LGALS4 gene mutations with prognostic and diagnostic value for multiple cancer types, including COAD and BLCA. We validated those findings against Gene Expression Omnibus (GEO) independent datasets. Our analysis reveals novel drug targets, as well as candidate molecular markers, that may allow an improved understanding of the pathogenesis of COAD and BLCA.

Materials and methods

Data source

The TCGA-COAD and TCGA-BLCA datasets used in this study were downloaded from the TCGA 11354

database [7]. The gene expression matrix and clinical data were downloaded from the GEO database repository [8]. All datasets were downloaded on the 1st of November 2020. Details of the TCGA and GEO datasets used in this study are summarized in **Table 1**. **Figure 1** depicts the study workflow.

Selection of candidate genes from RNA sequencing and single-cell sequencing

We analyzed the differential expression of genes involved in colonic adenocarcinoma using the TCGA-COAD (The Cancer Genome Atlas-Colon Adenocarcinoma) RNA-Seq level 3 raw count data [7]. We used DESeq2 to investigate the RNAseq raw count so as to identify differentially expressed genes [9]. The *P*-values were found by the Wald test and were corrected for multiple testing using the Benjamini and Hochberg method. Since the unbalanced class distribution of labels (COAD: 459 vs. Normal: 41) can affect predictive performance, in particular for minority classes [10], we applied data under-sampling. We chose 41 colon adenocarcinoma samples and 41 normal samples iteratively (non-repeating random samples of COAD) and analyzed them using DESeq2. We then combined the results ($P < 0.0001$) by taking the union of the list of genes in the resultant table from each iteration. The adjusted *P*-value < 0.001 resulted in the identification of 6066 genes across all iterations. We used random forest-based Recursive Feature Elimination (RFE) to select a small subset of genes from a broad range of gene expression data [11]. Due to class unbalance, we performed RFE iteratively over the undersampled data. This resulted in the identification of a total of 345 genes across all iterations. The union of the genes from each iteration resulted in the selection of 121 genes. Out of the 345 identified genes, 76 appeared more than once among 11 iterations. The FPKM values of the significant genes from the single-cell data were obtained from the dataset provided by Li et al. [12], and Zhang et al. [13]. We then selected 17 common overlapping genes across the TCGA-COAD RNA seq

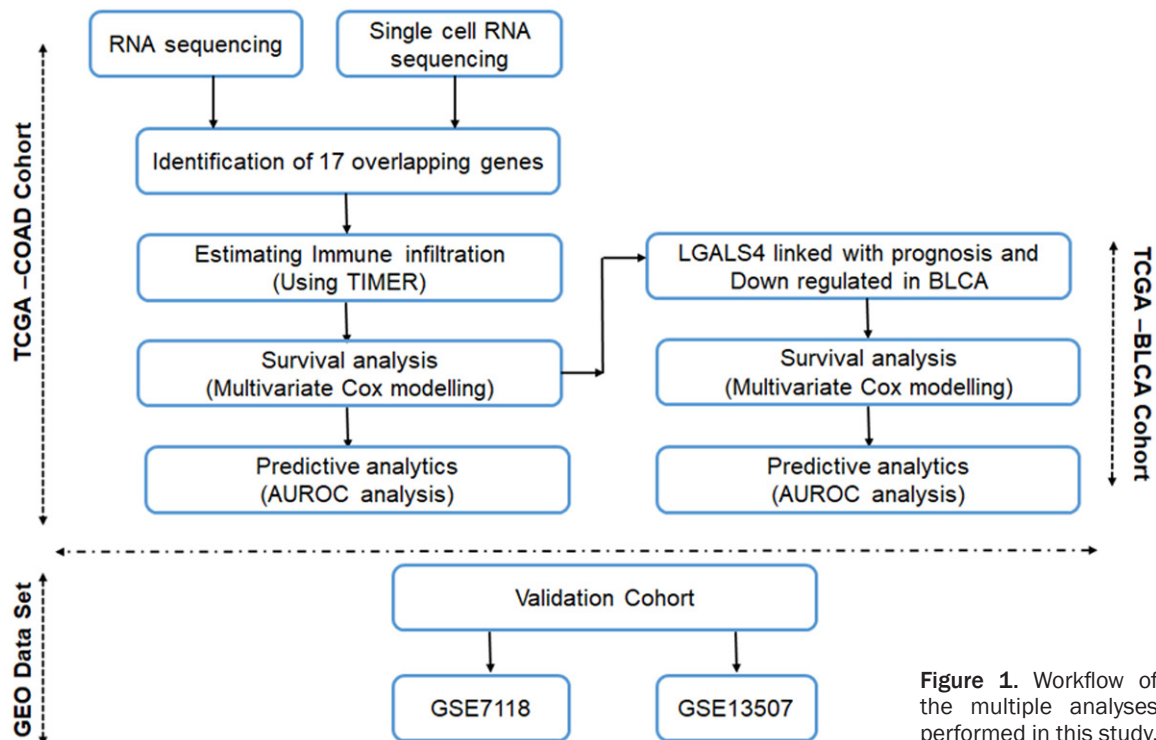


Figure 1. Workflow of the multiple analyses performed in this study.

and single cell RNA sequencing datasets, which were used for further analysis.

Evaluation and identification of immune cell infiltration

To identify and evaluate the abundance of immune infiltrates, we uploaded 17 genes to TIMER2.0 [6, 14]. TIMER is a resource for systematic and extensive analysis of immune infiltrates across diverse cancer types. The abundances of six cell types, namely: B cells, CD4+ T cells, CD8+ T cells, neutrophils, macrophages, and dendritic cells were estimated by the TIMER algorithm. The TIMER web server provides immune infiltrates' abundances estimated by multiple immune deconvolution methods and allows to generate high-quality figures dynamically and comprehensively. We used the "Survival module" [15] to explore associations between clinical outcome and abundance of immune infiltrates and gene expression.

Statistical analysis

Principal component analysis (PCA) and Area under the receiver operating characteristic (AUROC) analysis: We applied unsupervised Principal component analysis (PCA) to identify correlation structures or clusters within the normal vs. colon cancer patients across TCGA (COAD and BLCA) and GEO datasets (GSE71187

and GSE13507). We then used a logistic regression model to identify prediction performance, and AUROC curves were produced.

Survival and Kaplan-Meier analysis: We used the Kaplan-Meier method [16] to analyse the overall survival (OS). We also used the log-rank test to compare the differences between groups normal vs. colon cancer patients across TCGA (COAD and BLCA) and GEO datasets. A multivariate Cox proportional hazards model was applied, and the results were presented as hazard ratios (HRs) with 95% confidence intervals. Results with significant differences ($P < 0.05$) in the multivariate Cox proportional hazard model and log-rank test are listed in **Tables 3** and **4**.

All statistical analyses were performed using R version 4.0.3 software (<http://www.r-project.org>). The "survival" and "survminer" R packages were used for the Cox modeling and survival analysis [15]. In all the statistical tests, P -values < 0.05 were considered significant.

Results

Significant associations of the selected genes with immune cell infiltration

To identify the significant associations between the genes of interest and their expression in

Table 2. Associations between the 17 genes and the immune cell types with Spearman correlation values

Gene Symbol	B cell	CD8+ T Cell	CD4+ T Cell	Macrophage	Neutrophil	Dendritic cell
ADH1B	0.189	-0.007	0.386	0.374	0.17	0.247
KIAA1199	-0.046	0.073	0.181	0.081	0.071	0.08
CDH3	-0.031	0.000	0.094	-0.022	0	0.053
CA7	0.075	-0.105	0.066	-0.054	-0.071	-0.007
GUCA2B	0.091	-0.11	0.079	0.000	0.058	0.078
ABCC13	0.169	-0.142	0.049	-0.094	-0.158	-0.064
ABCG2	0.247	0.183	0.233	0.328	0.329	0.316
CPNE7	-0.153	-0.315	0.031	-0.18	-0.179	-0.277
HHLA2	0.221	0.014	0.168	-0.035	0.089	0.104
CEACAM7	0.183	-0.03	0.098	-0.077	0.054	0.07
AQP8	0.129	-0.178	0.168	-0.022	-0.063	0.019
GTF3A	-0.095	-0.112	-0.128	0.047	-0.114	-0.201
MMP28	0.215	0.049	0.18	0.054	0.077	0.19
LGALS4	0.141	-0.07	-0.043	-0.194	-0.188	-0.057
HSD11B2	0.035	-0.228	0.053	-0.194	-0.3	-0.202
CHP2	0.133	-0.178	0.134	0.068	-0.1	-0.059
NR3C2	0.353	0.261	0.165	0.085	0.232	0.305

Positive and significant correlations ($P < 0.05$) are highlighted in bold.

Table 3. A multivariate Cox model showing the Hazard Ratio (HR) and corresponding P values for the TCGA-COAD dataset

Parameters measured	Coefficient	Hazard Ratio (HR)	95% (Lower)	95% (Higher)	P value
Age	0.026	1.026	1.009	1.044	0.003
Gender	0.049	1.051	0.699	1.579	0.812
B cell	1.224	3.402	0.044	2.648	0.582
CD8+ T Cell	-4.688	0.009	0.000	0.410	0.016
CD4+ T Cell	-0.481	0.618	0.006	9.210	0.836
Macrophage	3.962	2.556	0.446	5.073	0.103
Neutrophil	-2.373	0.093	0.000	4.651	0.493
Dendritic	0.778	2.178	0.140	4.867	0.578
LGALS4	-0.247	0.781	0.626	0.974	0.028

Significant ($P < 0.05$) P values from the Cox proportional-hazards model are marked as bold in the last column.

COAD, we analysed the abundances of six immune cell types using the immune deconvolution framework TIMER. Significant ($P < 0.05$) Spearman correlation values are listed in **Table 2**. Detailed analysis of the 17 genes and their immune infiltration is presented in the [Supplementary Figure 1](#). ABCG2 expression was positively correlated with six immune cell types, namely B cells ($P = 5.23 \times 10^{-7}$), macrophages ($P = 8.64 \times 10^{-9}$), CD4+ T cells ($P = 2.25 \times 10^{-5}$), CD8+ T cells ($P = 1.86 \times 10^{-9}$), neutrophils ($P = 2.12 \times 10^{-8}$), and dendritic cells ($P = 2.01 \times 10^{-8}$). ADH1B gene expression was also positively associated with the levels of B cells

($P = 1.38 \times 10^{-4}$), CD4+ T cells ($P = 2.25 \times 10^{-5}$), macrophages ($P = 8.64 \times 10^{-9}$), neutrophils ($P = 2.12 \times 10^{-8}$), and dendritic cells ($P = 2.01 \times 10^{-8}$). Of the six cell types examined, B cell and CD4+ cell infiltration exhibited higher degrees of association with the majority of the genes. Those selected genes may play a vital role not only in the activation but also in the recruitment of multiple immune cell types in COAD.

Prognostics aspects of LGALS4

We used a Cox proportional hazards model to evaluate the OS of the patients with each of the

Table 4. Multivariate Cox model showing Hazard Ratios (HR) and corresponding *P* values from the TCGA-BLCA dataset

Parameters measured	Coefficient	Hazard Ratio (HR)	95% (Lower)	95% (Higher)	<i>P</i> value
Age	0.029	1.029	1.013	1.045	0.000
Gender	-0.169	0.844	0.599	1.191	0.335
B cell	-2.947	0.052	0.002	1.135	0.060
CD 8 Tcell	1.968	7.158	0.476	10.626	0.155
CD 4 Tcell	-0.352	0.703	0.017	2.119	0.853
Macrophage	3.872	4.018	3.959	8.916	0.000
Neutrophil	-2.894	0.055	0.000	6.701	0.237
Dendritic	-0.342	0.711	0.159	3.167	0.654
LGALS4	-0.096	0.908	0.836	0.987	0.023

genes separately. The model included clinical values (age and gender) as well as immune cells (B cells, CD8+ T cells, CD4+ T cells, macrophages, neutrophils, and dendritic cells). LGALS4 gene, CD8+ T cells and age (**Table 3**) were identified as significant ($P < 0.05$) values by the Cox proportional-hazards model. To understand the probability of overall survival, we used the Kaplan-Meier (KM) survival curve for both TCGA and GEO datasets, GSE71187. For the TCGA-COAD dataset (**Figure 2A**), a Hazard Ratio (HR) of 0.622 with a confidence interval (CI) 0.395-0.980 at $P = 0.041$, was reported. For the GSE71187 (**Figure 2B**), HR and CI were 0.0382 with (0.05-0.919) at $P = 0.038$.

LGALS4 was a biomarker of COAD

We performed a PCA analysis on both the TCGA cohort and GSE71187. Two separate clusters (normal and COAD) were identified for both sets (**Figure 3**). This indicates that those genes may be useful signatures (or biomarkers) for discriminating normal vs. COAD samples. Moreover, we calculated the sensitivity and specificity of the discriminative ability of the identified genes between COAD and normal patients. The results suggested that LGALS4 was highly discriminative between COAD patients and healthy controls. The AUROC results for LGALS4 for TCGA and GSE71187 were 0.973 (CI: 0.934-0.994) and 0.98 (CI: 0.93-1) respectively (**Figure 4A, 4B**).

LGALS4 expression impact on other cancer types

We performed a Wilcoxon rank-sum test and investigated the expression pattern of LGALS4

across TCGA datasets corresponding to multiple cancer types. A visual representation of all the cancer types and the normalized expression of all the genes is shown in **Figure 5A**. Interestingly, we found LGALS4 expression to be significantly downregulated both in COAD with $P = 8.3e-23$ (**Figure 5B**) and BLCA with $P = 0.024$ (**Figure 5C**). We then performed a Cox survival analysis (**Table 4**) to assess whether LGALS4 expression is associated with BLCA cancer patients' survival and identified an HR: 0.908, CI: (0.836-0.987), and P -value of 0.023. The Likelihood ratio test was significant at $P = 1.19e-07$.

Predictive aspects of the LGALS4 gene in BLCA

In order to investigate LGALS4 predictive ability between controls and BLCA patients, we employed a GEO dataset (GSE31507) and applied a logistic regression model. We measured the model's performance by evaluating the AUROC, which resulted in a value of 0.74 with CI (0.66-0.82) (**Figure 6A**). A Wilcoxon rank-sum test of the LGALS4 expression between the controls and the BLCA, was also significant ($P < 0.001$). A violin plot indicating the significant up and down-regulation of the LGALS4 gene in BLCA (GSE31507 dataset) is presented in **Figure 6B**.

Discussion

Tumor infiltration

We used genes identified from TCGA-COAD and single-cell RNA sequencing datasets, in an effort to delineate their role in the immune infiltration [17, 18]. Within the tumor microenviron-

Diagnostic use of LGALS4

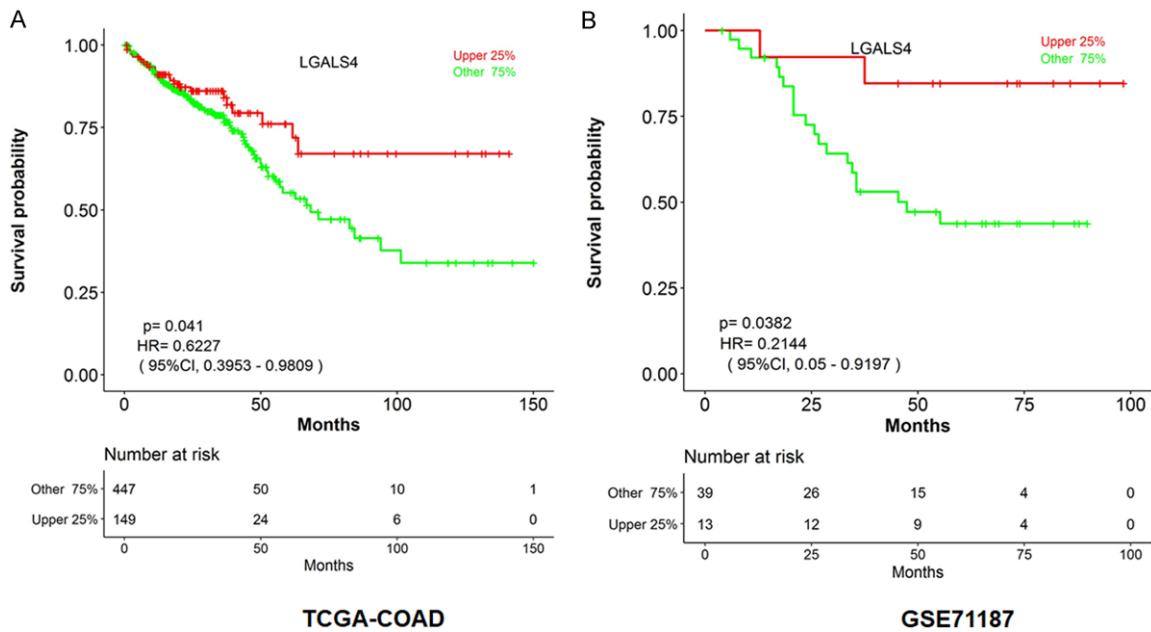


Figure 2. Kaplan-Meier (KM) survival curve for the LGALS4 gene expression. The gene expression values were divided into the upper 25% quartile as one class and the rest 75% as different classes. A risk table was also produced indicating the number of patients at risk and time in months (A) TCGA-COAD and (B) GSE71187.

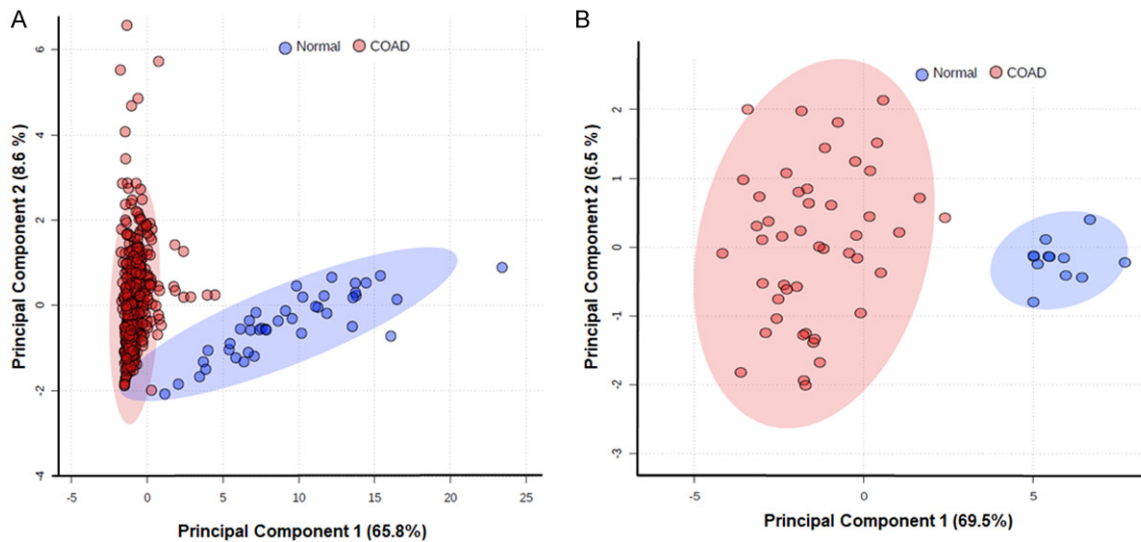


Figure 3. A PCA (principal component analysis) score plot performed on the gene expression data demonstrating clustering of the Normal vs. COAD cohorts. Small circles, representing patients, are coloured according to the patient cohort. The ellipse represents 95% confidence. The percent variation of the cohort explained by the respective x and y-axis. A. PCA on TCGA dataset. B. PCA on GSE71187.

ment [17], immune cells play a role in modulating and changing tumor behavior. We used the deconvolution algorithm TIMER [6] to investigate associations between identified genes and associate them with immune-related cell types in COAD and BLCA. We also found a sig-

nificant positive correlation between the expression levels of many of the genes, for example, ADH1B, ABCG2, and NR3C2. Lu *et al.* have also reported NR3C2 involvement in the immune infiltration of breast cancer [19]. They collected data from 24 immune cell types

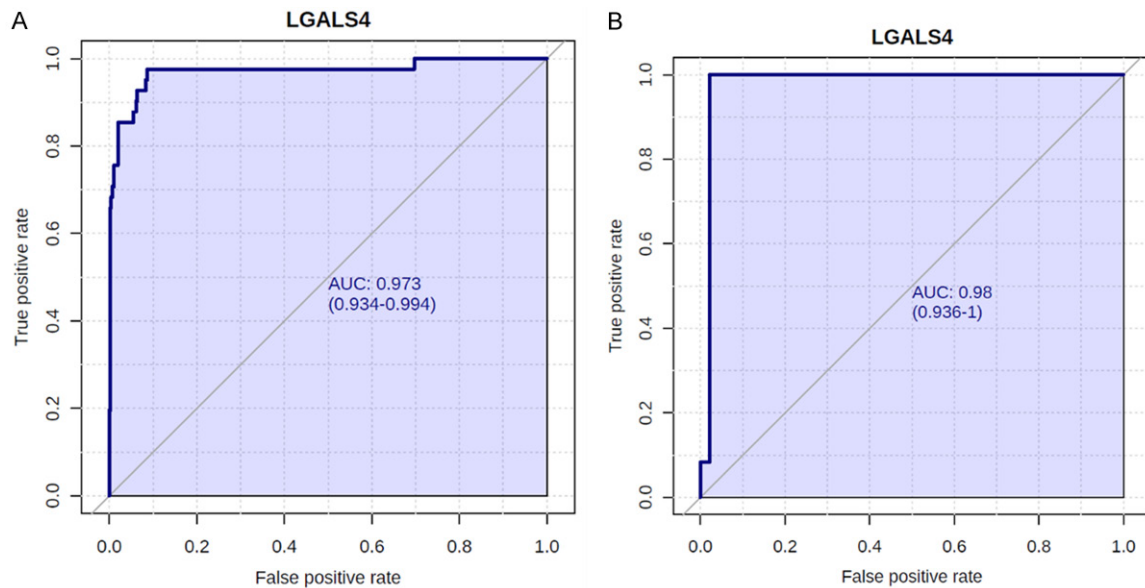


Figure 4. A. An AUROC plot with CI of LGALS4 gene on TCGA dataset discriminating normal vs. colon cancer patients. B. An AUROC plot with CI of LGALS4 gene on GSE71187 discriminating normal vs. colon cancer patients.

including macrophages, eosinophils, neutrophils, natural killer cells, and dendritic cells. The correlation between NR3C2 gene expression and immune infiltration was explored using the TIMER2.0 database and the NR3C2 expression was correlated with immune cells and some other T cell subsets [19]. Their report concluded that NR3C2 genes could be a prognostic biomarker for the treatment of breast carcinoma. Jin *et al.* reported on the role of ABCG2 in the maturation of dendritic cells (DCs) and showed that its inhibition suppresses DC maturation and promotes the generation of immune tolerogenic effect [20]. They concluded that the up- or down-regulation of the ABCG2 activity can affect the balance of immune cell activation versus immune tolerance mechanisms. This finding suggests the need for a better understanding of the function of the ABCG2 gene and its effect on immune tolerance in order to develop better treatment methods for multi-drug resistant cancer. Gharpure *et al.* reported the involvement of ADH1B gene in ovarian cancer cell invasion and metastasis [21]. Their results showed that over-expression of ADH1B gene facilitates a more infiltrative cancer cell phenotype, and increases the adhesion of cancer cells to mesothelial cells, extracellular matrix degradation, and hypoxia, all of which promote cancer progression and metastasis [21].

The LGALS4 gene has a significant positive association with B cell infiltration in COAD. Among six immune cell types, B cells and CD4+ cells were more associated with almost all the genes listed in the **Table 2**. These correlations suggest that genes may play a vital role in the activation and recruitment of immune cells in COAD. In the case of the BLCA, LGALS4 is significant positively associated with B cell infiltration. More recently, Na *et al.* identified nine immune-related genes, namely MMP9, PDGFRA, AHNK, OLR1, RAC3, IGF1, PGF, OAS1, and SH3BP2, and used them to develop a risk stratification model by multivariate Cox proportional hazards regression analysis [22].

Prognostic and diagnostic use of LGALS4 in COAD

Kim *et al.* investigated the role of LGALS4 in COAD development, both *in vitro* and *in vivo* [23]. They reported that LGALS4 silencing increases cell proliferation and activates NF- κ B and STAT3 signaling along with IL-6 up-regulation. LGALS4 is a microvillar lipid raft involved in cell adhesion [23], and is secreted to mediate cell responses [24]. Consistent with our findings, LGALS4 has been reported as strongly down-regulated in COAD [25]. Previous LGALS4 gene expression studies, for example, by Lin *et al.* [26], have identified 51 differentially

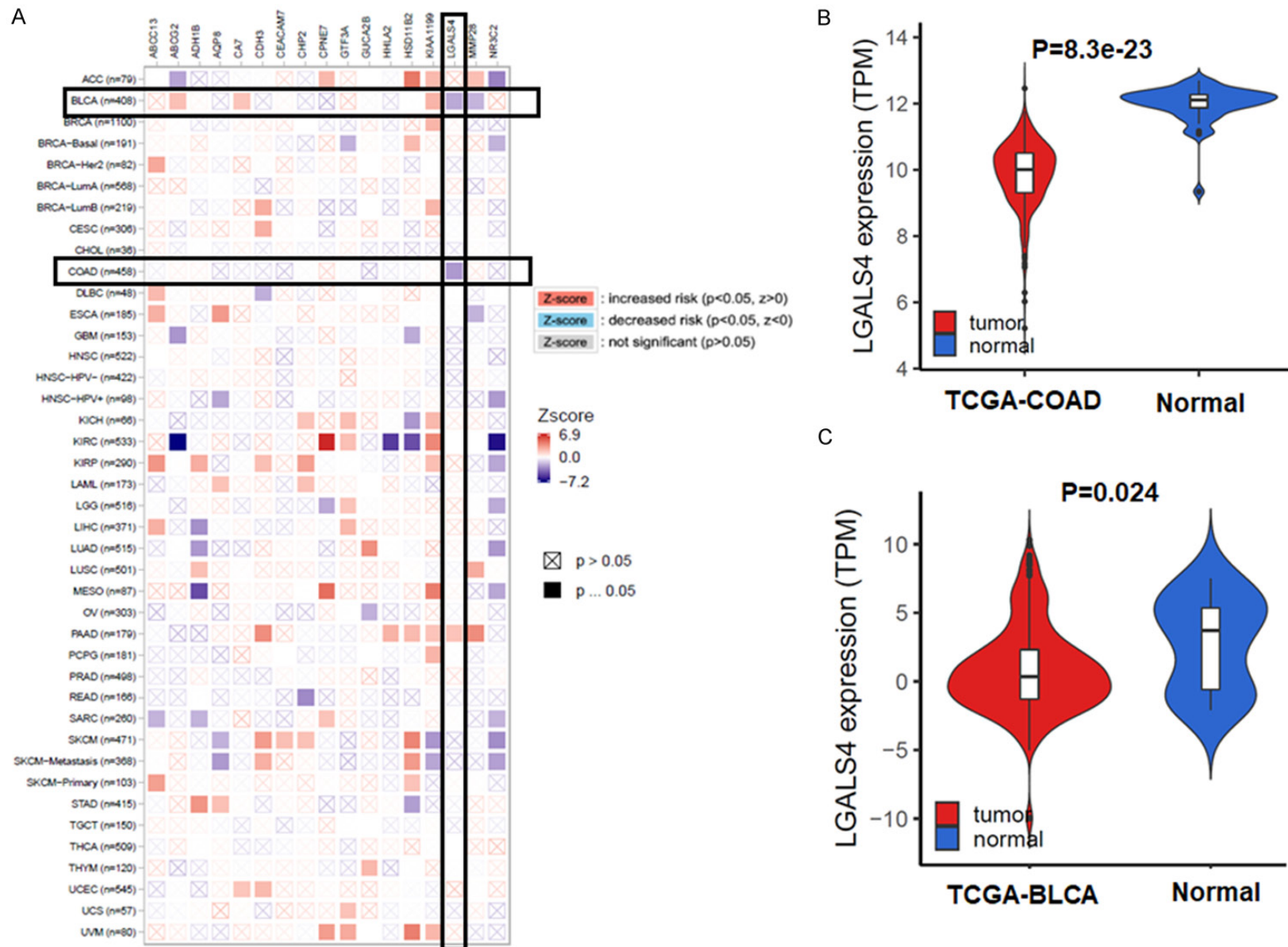


Figure 5. A. Normalised (Z score) expression values of the genes within various TCGA dataset across multiple cancer types. Significant ($P < 0.05$) up-regulation of the genes is represented in solid red and down-regulation in solid blue. B. A violin plot representing the significant up and down-regulation of the LGALS4 gene in COAD. C. A violin plot representing the significant up and down-regulation of the LGALS4 gene in TCGA-BLCA.

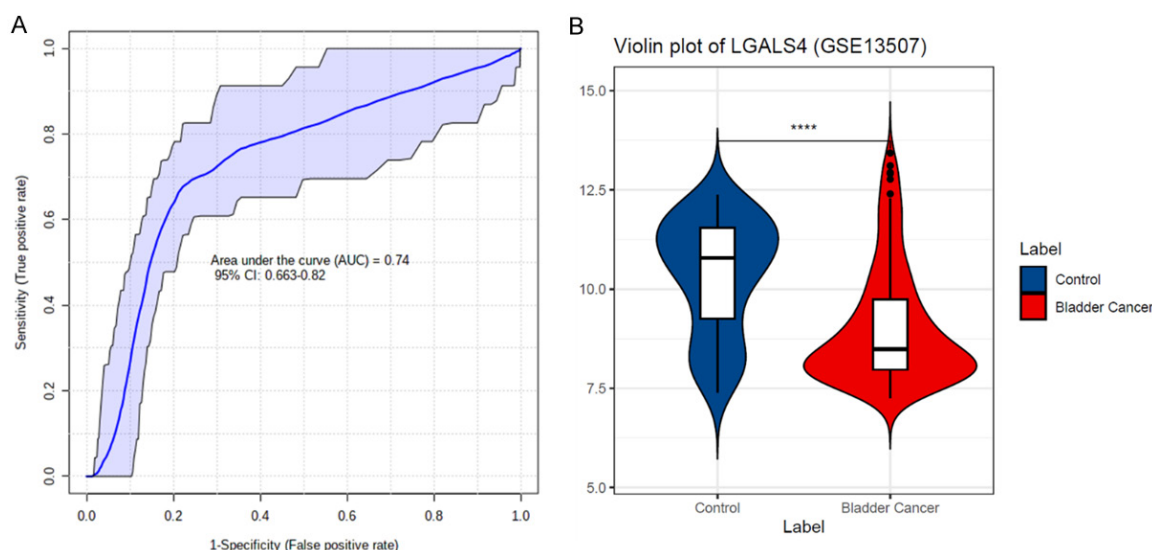


Figure 6. A. An AUROC with a 0.74 value and a CI (0.66-0.82) related to the performance of the model using the LGALS4 gene in the GEO dataset GSE31507. B. A violin plot representing the significant up and down-regulation of the LGALS4 gene in BLCA (GSE31507 dataset).

expressed genes between COAD and normal samples and the LGALS4 gene was also down-regulated in the tumor samples. Based on these findings, Lin et al. 2002 proposed and developed a scoring system to separate adenomas from carcinomas [26]. Zhang et al., 2019 integrated six GEO gene expression datasets (GSE25070, GSE44076, GSE44861, GSE21510, GSE9348, and GSE21815) to identify key colon cancer regulators [27]. Rodia et al. 2018 reported that the LGALS4 interaction with other genes, namely EACAM6, TSPAN8, and COL1A2, was able to discriminate between low risk vs. high-risk colon cancer patients, leading to the hypothesis of the LGALS4 gene's role in the development and progression of colorectal cancer [28, 29]. They also reported the combined effect of these genes, as a part of a diagnostic panel, resulting in a logistic regression model with a 0.91 AUROC value.

LGALS4 gene is involved in BLCA and other cancer types

Interestingly, the function of LGALS4 seems multifactorial with a key role in the progression of BLCA as well as other similar cancer types. Ding et al. 2019 [30], performed a Weighted Gene Co-expression Network Analysis (WGCNA) over BLCA related TCGA datasets, reporting similar results to our own, identifying LGALS4 as a hub gene that plays a critical role in the

progression of urothelial carcinoma, as well as depicting its role as a possible diagnostic biomarker and target for bladder cancer therapy.

LGALS4 produces Galectin-4 which is involved in various biological processes, including intestinal wound healing, promotion of axon growth, and myelination in a neuron [31]. LGALS4 has also been implicated in autoimmune disease, for example inflammatory bowel disease, and has been associated with the development and progression of other cancer types, including pancreatic carcinoma [32], hepatocellular carcinoma [33], COAD [25], gastric cancer [34], and lung cancer [35]. In the case of COAD, the LGALS4 expression is strikingly reduced compared to normal colonic tissues, leading to tumor progression and metastasis [25, 36, 37]. In the case of urothelial bladder cancer, LGALS4 has been demonstrated as a prognostic marker, affecting cell function and inhibiting cancer cell growth and invasiveness [30]. Furthermore, hypermethylation in the promoter of this gene has been positively related to reduced patient survival [38].

Our study was limited since the datasets employed were unbalanced (different numbers of normal vs. cancer types). Additionally, our approach remains computational in nature and, as such, as our findings require experimental validation, for example by LGALS4 func-

tional validation studies using qPCR (quantitative polymerase chain reaction) to confirm our results.

Conclusions

LGALS4 gene expression is downregulated in multiple cancer types, including COAD and BLCA. It may be a prognostic and diagnostic biomarker for the presence and progression of COAD and BLCA.

Availability of data and materials

The TCGA-COAD and BLCA dataset used can be obtained from the TCGA database (<https://cancergenome.nih.gov/>). The GEO datasets used in this paper can be obtained from GEO online database (<https://www.ncbi.nlm.nih.gov/geo/>).

Disclosure of conflict of interest

None.

Abbreviations

AUROC, Area under the receiver operator characteristic curve; BLCA, Bladder urothelial carcinoma; COAD, Colon adenocarcinoma; FPKM, Fragments Per Kilobase of transcript per Million; GEO, Gene Expression Omnibus; GO, Gene ontology; HR, Hazard ratio; IBD, Inflammatory bowel disease; LGALS4, Galactin-4; PCA, Principal component analysis; qPCR, Quantitative polymerase chain reaction; ROC, Receiver operating characteristic; TCGA, The Cancer Genome Atlas; TIMER, Tumor immune estimation resource; WGCNA, Weighted Gene Co-expression Network Analysis.

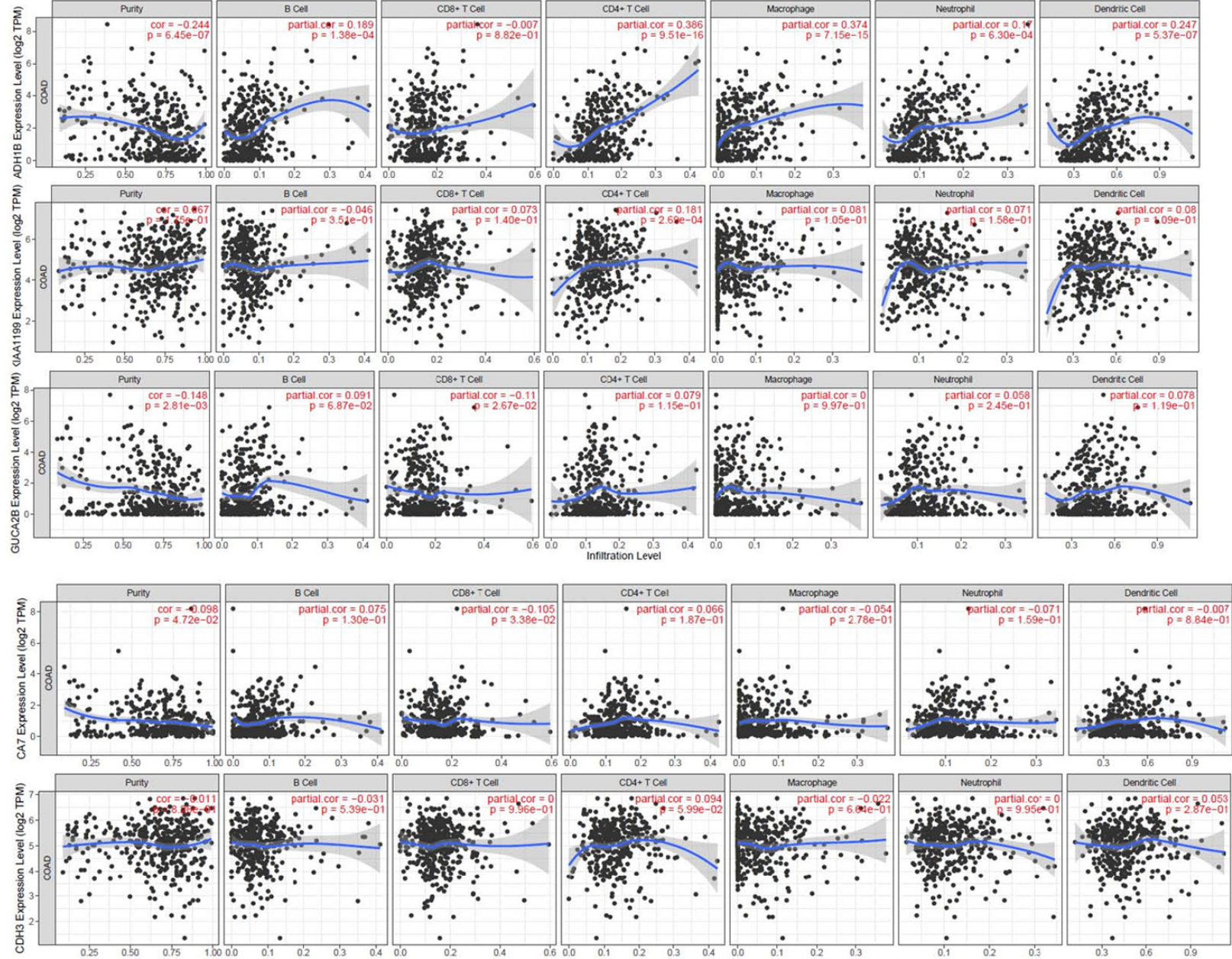
Address correspondence to: Dr. Animesh Acharjee, College of Medical and Dental Sciences, Institute of Cancer and Genomic Sciences, University of Birmingham, B15 2TT, UK. Tel: +4407403642022; E-mail: a.acharjee@bham.ac.uk

References

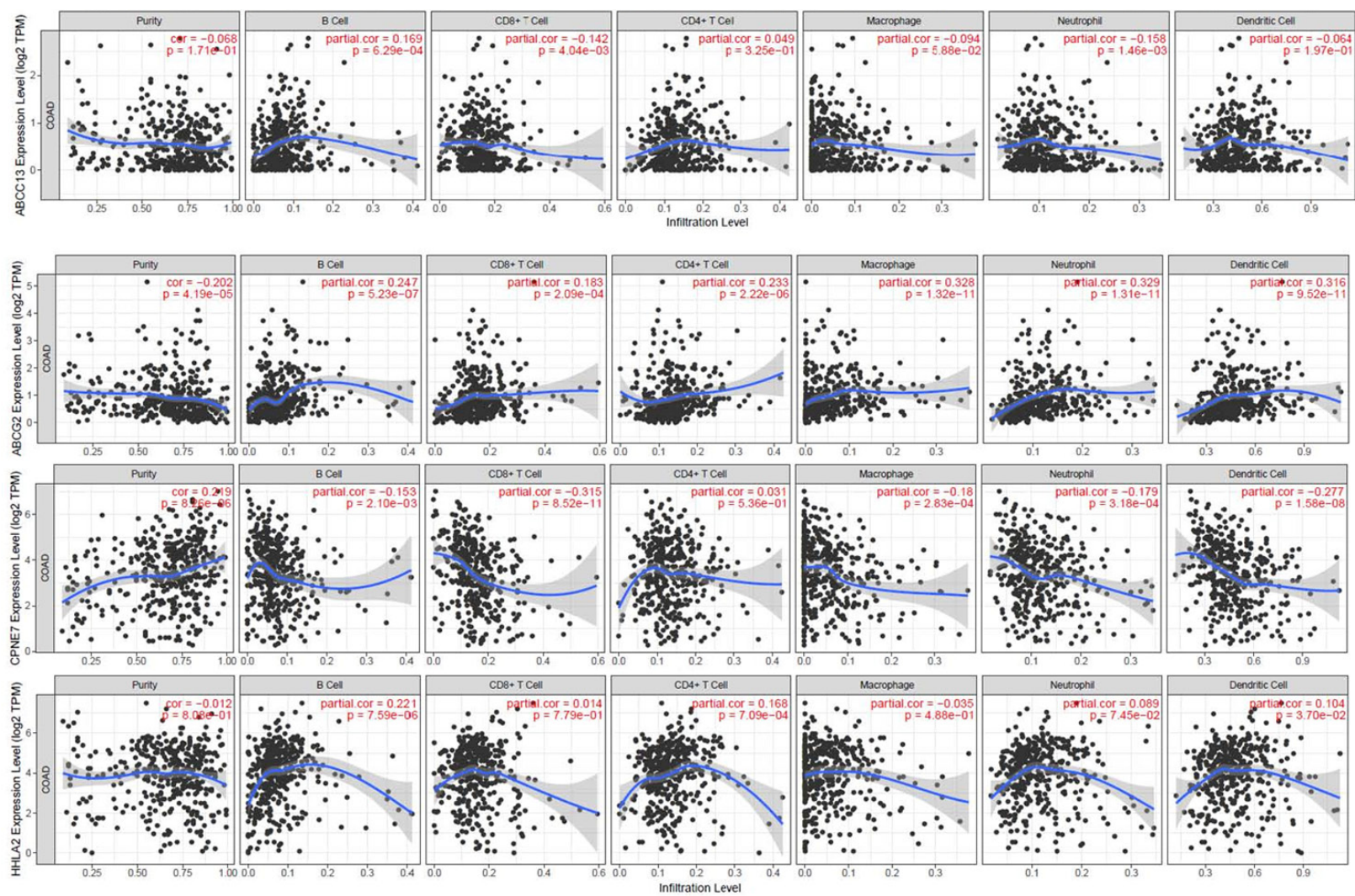
- [1] Munteanu I and Mastalier B. Genetics of colorectal cancer. *J Med Life* 2014; 7: 507-11.
- [2] Wang B, Liu M, Ran Z, Li X, Li J and Ou Y. Analysis of gene signatures of tumor microenvironment yields insight into mechanisms of resistance to immunotherapy. *Front Bioeng Biotechnol* 2020; 8: 348.
- [3] Ge P, Wang W, Li L, Zhang G, Gao Z, Tang Z, Dang X and Wu Y. Profiles of immune cell infiltration and immune-related genes in the tumor microenvironment of colorectal cancer. *Biomed Pharmacother* 2019; 118: 109228.
- [4] Peltekova VD, Lemire M, Qazi AM, Zaidi SH, Trinh QM, Bielecki R, Rogers M, Hodgson L, Wang M, D'Souza DJ, Zandi S, Chong T, Kwan JY, Kozak K, De Borja R, Timms L, Rangrej J, Volar M, Chan-Seng-Yue M, Beck T, Ash C, Lee S, Wang J, Boutros PC, Stein LD, Dick JE, Gryfe R, McPherson JD, Zanke BW, Pollett A, Gallinger S and Hudson TJ. Identification of genes expressed by immune cells of the colon that are regulated by colorectal cancer-associated variants. *Int J Cancer* 2014; 134: 2330-41.
- [5] Sato H, Ishihara S, Kawashima K, Moriyama N, Suetsugu H, Kazumori H, Okuyama T, Rumi MA, Fukuda R, Nagasue N and Kinoshita Y. Expression of peroxisome proliferator-activated receptor (PPAR)gamma in gastric cancer and inhibitory effects of PPARgamma agonists. *Br J Cancer* 2000; 83: 1394-400.
- [6] Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, Jiang P, Shen H, Aster JC, Rodig S, Signoretti S, Liu JS and Liu XS. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol* 2016; 17: 174.
- [7] The cancer genome atlas (TCGA) RNA-seq meta-analysis. 2018.
- [8] <https://www.ncbi.nlm.nih.gov/gds>.
- [9] Love MI, Huber W and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014; 15: 550.
- [10] Lin WJ and Chen JJ. Class-imbalanced classifiers for high-dimensional data. *Brief Bioinform* 2013; 14: 13-26.
- [11] Darst BF, Malecki KC and Engelman CD. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genet* 2018; 19: 65.
- [12] Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JLL, Kong SL, Chua C, Hon LK, Tan WS, Wong M, Choi PJ, Wee LJK, Hillmer AM, Tan IB, Robson P and Prabhakar S. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet* 2017; 49: 708-718.
- [13] Zhang GL, Pan LL, Huang T and Wang JH. The transcriptome difference between colorectal tumor and normal tissues revealed by single-cell sequencing. *J Cancer* 2019; 10: 5883-5890.
- [14] <http://timer.cistrome.org/>.
- [15] <https://rviews.rstudio.com/2017/09/25/survival-analysis-with-r/>.
- [16] Kaplan EL and Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958; 53: 457-481.

- [17] Seton-Rogers S. A multiplexed view of immune infiltration. *Nat Rev Cancer* 2018; 18: 666-667.
- [18] Jochems C and Schlom J. Tumor-infiltrating immune cells and prognosis: the potential link between conventional cancer therapy and immunity. *Exp Biol Med (Maywood)* 2011; 236: 567-79.
- [19] Lu J, Hu F and Zhou Y. NR3C2-related transcriptome profile and clinical outcome in invasive breast carcinoma. *Biomed Res Int* 2021; 2021: 9025481.
- [20] Jin JO, Zhang W, Wong KW, Kwak M, van Driel IR and Yu Q. Inhibition of breast cancer resistance protein (ABCG2) in human myeloid dendritic cells induces potent tolerogenic functions during LPS stimulation. *PLoS One* 2014; 9: e104753.
- [21] Gharpure KM, Lara OD, Wen Y, Pradeep S, LaFargue C, Ivan C, Rupaimoole R, Hu W, Mangala LS, Wu SY, Nagaraja AS, Baggerly K and Sood AK. ADH1B promotes mesothelial clearance and ovarian cancer infiltration. *Oncotarget* 2018; 9: 25115-25126.
- [22] Na L, Bai Y, Sun Y, Wang Z, Wang W, Yuan L and Zhao C. Identification of 9-core immune-related genes in bladder urothelial carcinoma prognosis. *Front Oncol* 2020; 10: 1142.
- [23] Danielsen EM and van Deurs B. Galectin-4 and small intestinal brush border enzymes form clusters. *Mol Biol Cell* 1997; 8: 2241-51.
- [24] Ideo H, Seko A and Yamashita K. Recognition mechanism of galectin-4 for cholesterol 3-sulfate. *J Biol Chem* 2007; 282: 21081-9.
- [25] Rechreche H, Mallo GV, Montalto G, Dagorn JC and Iovanna JL. Cloning and expression of the mRNA of human galectin-4, an S-type lectin down-regulated in colorectal cancer. *Eur J Biochem* 1997; 248: 225-30.
- [26] Lin YM, Furukawa Y, Tsunoda T, Yue CT, Yang KC and Nakamura Y. Molecular diagnosis of colorectal tumors by expression profiles of 50 genes expressed differentially in adenomas and carcinomas. *Oncogene* 2002; 21: 4120-8.
- [27] Zhang H, Du Y, Wang Z, Lou R, Wu J and Feng J. Integrated analysis of oncogenic networks in colorectal cancer identifies GUCA2A as a molecular marker. *Biochem Res Int* 2019; 2019: 6469420.
- [28] Rodia MT, Solmi R, Pasini F, Nardi E, Mattei G, Ugolini G, Ricciardiello L, Strippoli P, Miglio R and Lauriola M. LGALS4, CEACAM6, TSPAN8, and COL1A2: blood markers for colorectal cancer-validation in a cohort of subjects with positive fecal immunochemical test result. *Clin Colorectal Cancer* 2018; 17: e217-e228.
- [29] Ferlizza E, Solmi R, Miglio R, Nardi E, Mattei G, Sgarzi M and Lauriola M. Colorectal cancer screening: assessment of CEACAM6, LGALS4, TSPAN8 and COL1A2 as blood markers in faecal immunochemical test negative subjects. *J Adv Res* 2020; 24: 99-107.
- [30] Ding Y, Cao Q, Wang C, Duan H and Shen H. LGALS4 as a prognostic factor in urothelial carcinoma of bladder affects cell functions. *Technol Cancer Res Treat* 2019; 18: 1533033819876601.
- [31] Cao ZQ and Guo XL. The role of galectin-4 in physiology and diseases. *Protein Cell* 2016; 7: 314-24.
- [32] Belo AI, van der Sar AM, Tefsen B and van Die I. Galectin-4 reduces migration and metastasis formation of pancreatic cancer cells. *PLoS One* 2013; 8: e65957.
- [33] Cai Z, Zeng Y, Xu B, Gao Y, Wang S, Zeng J, Chen L, Huang A, Liu X and Liu J. Galectin-4 serves as a prognostic biomarker for the early recurrence/metastasis of hepatocellular carcinoma. *Cancer Sci* 2014; 105: 1510-7.
- [34] Hippo Y, Yashiro M, Ishii M, Taniguchi H, Tsutsumi S, Hirakawa K and Kodama T, Aburatani H. Differential gene expression profiles of scirrhous gastric cancer cells with high metastatic potential to peritoneum or lymph nodes. *Cancer Res* 2001; 61: 889-95.
- [35] Hayashi T, Saito T, Fujimura T, Hara K, Takamochi K, Mitani K, Mineki R, Kazuno S, Oh S, Ueno T, Suzuki K and Yao T. Galectin-4, a novel predictor for lymph node metastasis in lung adenocarcinoma. *PLoS One* 2013; 8: e81883.
- [36] Satelli A, Rao PS, Thirumala S and Rao US. Galectin-4 functions as a tumor suppressor of human colorectal cancer. *Int J Cancer* 2011; 129: 799-809.
- [37] Kim SW, Park KC, Jeon SM, Ohn TB, Kim TI, Kim WH and Cheon JH. Abrogation of galectin-4 expression promotes tumorigenesis in colorectal cancer. *Cell Oncol (Dordr)* 2013; 36: 169-78.
- [38] Wu MM, Li CF, Lin LF, Wang AS, Pu YS, Wang HH, Mar AC, Chen CJ and Lee TC. Promoter hypermethylation of LGALS4 correlates with poor prognosis in patients with urothelial carcinoma. *Oncotarget* 2017; 8: 23787-23802.
- [39] An N, Shi X, Zhang Y, Lv N, Feng L, Di X, Han N, Wang G, Cheng S and Zhang K. Discovery of a novel immune gene signature with profound prognostic value in colorectal cancer: a model of cooperativity disorientation created in the process from development to cancer. *PLoS One* 2015; 10: e0137171.
- [40] Lee JS, Leem SH, Lee SY, Kim SC, Park ES, Kim SB, Kim SK, Kim YJ, Kim WJ and Chu IS. Expression signature of E2F1 and its associated genes predict superficial to invasive progression of bladder tumors. *J Clin Oncol* 2010; 28: 2660-7.

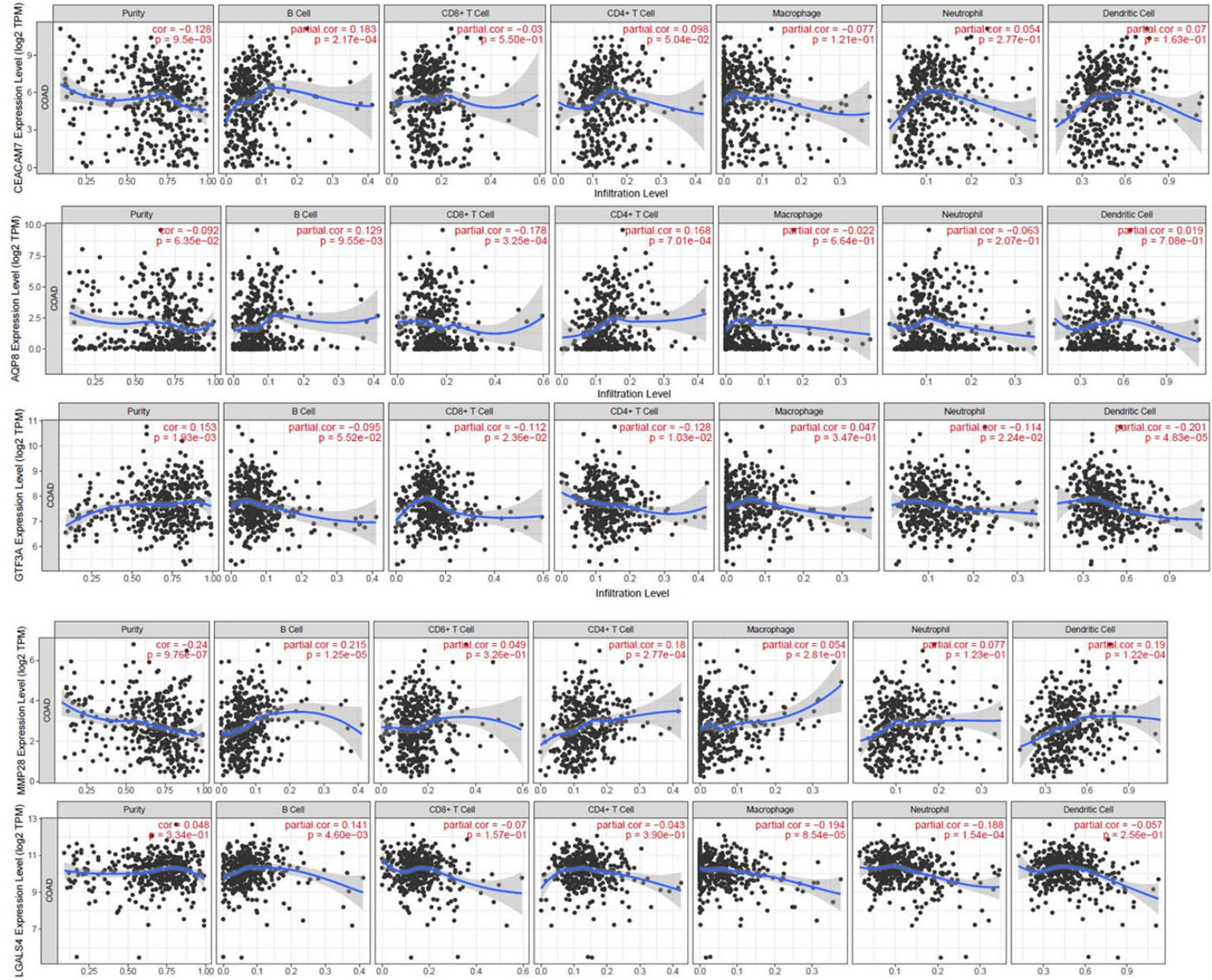
Diagnostic use of LGALS4



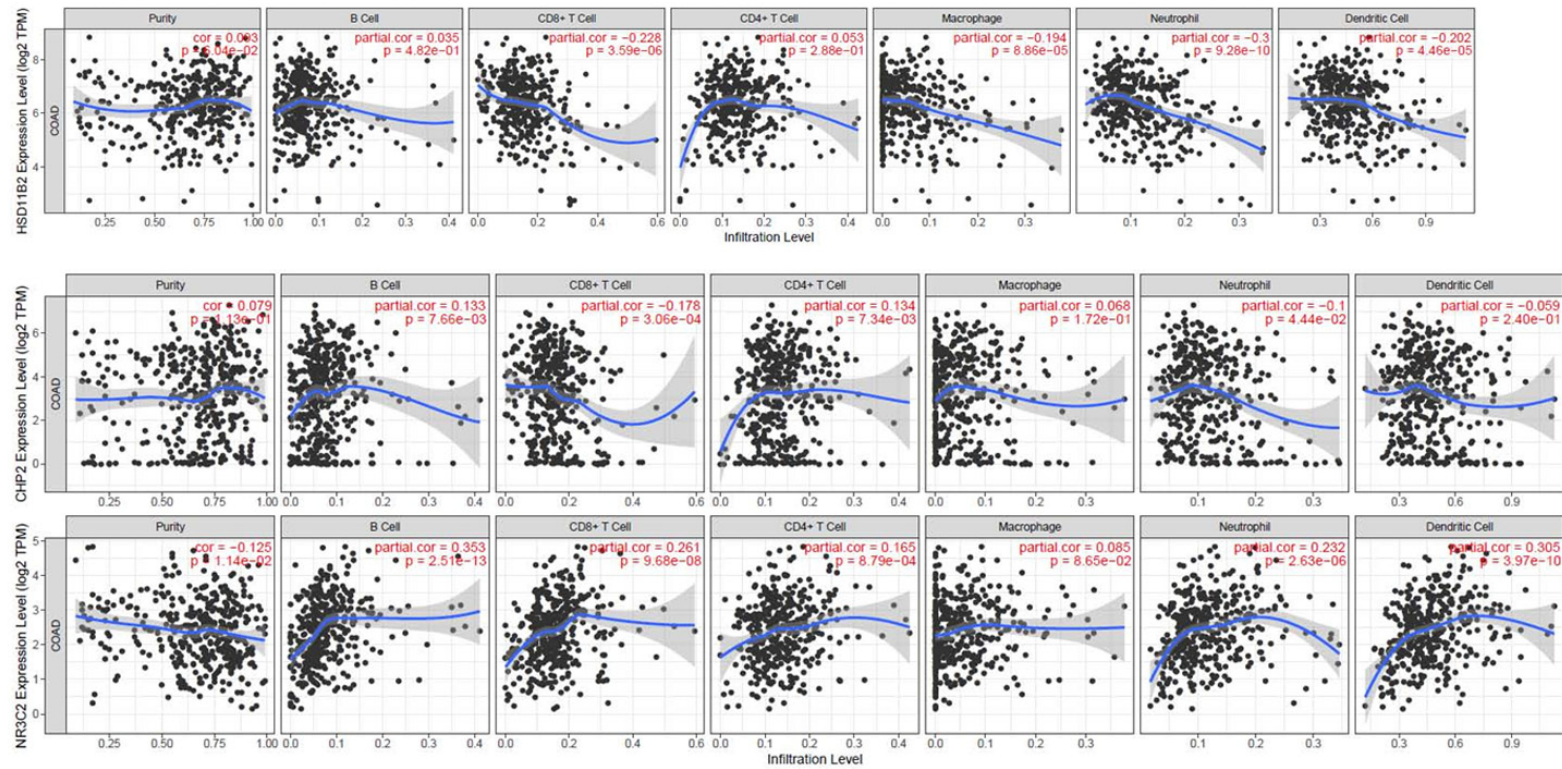
Diagnostic use of LGALS4



Diagnostic use of LGALS4



Diagnostic use of LGALS4



Supplementary Figure 1. Spearman correlation analysis with *P* values linking the 17 genes and their immune infiltration status is listed.