Computational and Structural Biotechnology Journal 23 (2024) 752–758

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Short communication

GCclassifier: An R package for the prediction of molecular subtypes of gastric cancer

Jiang Li [a,b,c], Lingli He [a,b,c], Xianrui Zhang [a,b,c], Xiang Li [a,b], Lishi Wang [a,b], Zhongxu Zhu [a,b,d], Kai Song [a,b,e], Xin Wang [a,b,e,*]

[a] Department of Surgery, The Chinese University of Hong Kong, Shatin, Hong Kong Special Administrative Region of China
[b] Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong Special Administrative Region of China
[c] Department of Biomedical Sciences, City University of Hong Kong, Hong Kong Special Administrative Region of China
[d] HIM-BGI Omics Center, Hangzhou Institute of Medicine (HIM), Chinese Academy of Sciences, Hangzhou, China
[e] Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, Region of China

ARTICLE INFO

Keywords:
Cancer subtyping
Gastric cancer
GCclassifier
R package
Shiny application

ABSTRACT

Gastric cancer (GC) is one of the most commonly diagnosed malignancies, threatening millions of lives worldwide each year. Importantly, GC is a heterogeneous disease, posing a significant challenge to the selection of patients for more optimized therapy. Over the last decades, extensive community effort has been spent on dissecting the heterogeneity of GC, leading to the identification of distinct molecular subtypes that are clinically relevant. However, so far, no tool is publicly available for GC subtype prediction, hindering the research into GC subtype-specific biological mechanisms, the design of novel targeted agents, and potential clinical applications. To address the unmet need, we developed an R package *GCclassifier* for predicting GC molecular subtypes based on gene expression profiles. To facilitate the use by non-bioinformaticians, we also provide an interactive, user-friendly web server implementing the major functionalities of *GCclassifier*. The predictive performance of *GCclassifier* was demonstrated using case studies on multiple independent datasets.

1. Introduction

Gastric cancer (GC) is a leading cause of cancer-related death and one of the most commonly diagnosed cancers [1,2]. Surgical resection with subsequent adjuvant chemotherapy (CT), radiotherapy, immunotherapy, and targeted therapy has been proven as effective treatments for patients with GC, with substantial benefits in the reduction of mortality [3]. However, since GC is a highly heterogeneous disease entity with complex genetic features, these approaches are usually accompanied by overtreatment, which may result in unnecessary medical costs and patients' anxieties, or undertreatment, which may lead to the persistence and recurrence of GC [4–8]. An effective stratification of GC patients is imperative to avoid such adverse outcomes [9].

Multiple studies on GC heterogeneity have identified distinct molecular subtypes, with comprehensive dissection of clinical characteristics, biological properties, and prioritized potential druggable targets [10–14]. More specifically, a comprehensive study was performed by The Cancer Genome Atlas (TCGA) with four genomic subtypes identified: Epstein–Barr virus (EBV), microsatellite instability (MSI), genomic stability (GS), and chromosomal instability (CIN). The Asian Cancer Research Group (ACRG) performed molecular subtyping using gene expression profiles for 300 primary GC tumor specimens. As a result, ACRG identified four clinically relevant molecular subtypes, namely MSS (microsatellite stable)/TP53-, MSS/TP53+, MSI (microsatellite instability), and MSS/epithelial–mesenchymal transition (EMT), in which the MSS/EMT subtype was associated with the worst survival and

Abbreviations: GC, gastric cancer; TCGA, The Cancer Genome Atlas; EBV, Epstein–Barr virus; MSI, microsatellite instability, GS, genomic stability; CIN, chromosomal instability; ACRG, The Asian Cancer Research Group; MSS, microsatellite stable; MSI, microsatellite instability; EMT, epithelial–mesenchymal transition; MP, mesenchymal phenotype; EP, epithelial phenotype; OS, overall survival; RFS, recurrence-free survival; UCSC, the University of California Santa Cruz; TPM, transcripts per million; FPKM, fragments per kilobase million; GEO, Gene Expression Omnibus; ROC, receiver operating characteristic curve; AUC, area under the curve.

* Correspondence to: Department of Surgery, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, Hong Kong Special Administrative Region of China.
E-mail address: xinwang@cuhk.edu.hk (X. Wang).

https://doi.org/10.1016/j.csbj.2024.01.010
Received 22 October 2023; Received in revised form 14 January 2024; Accepted 15 January 2024
Available online 17 January 2024

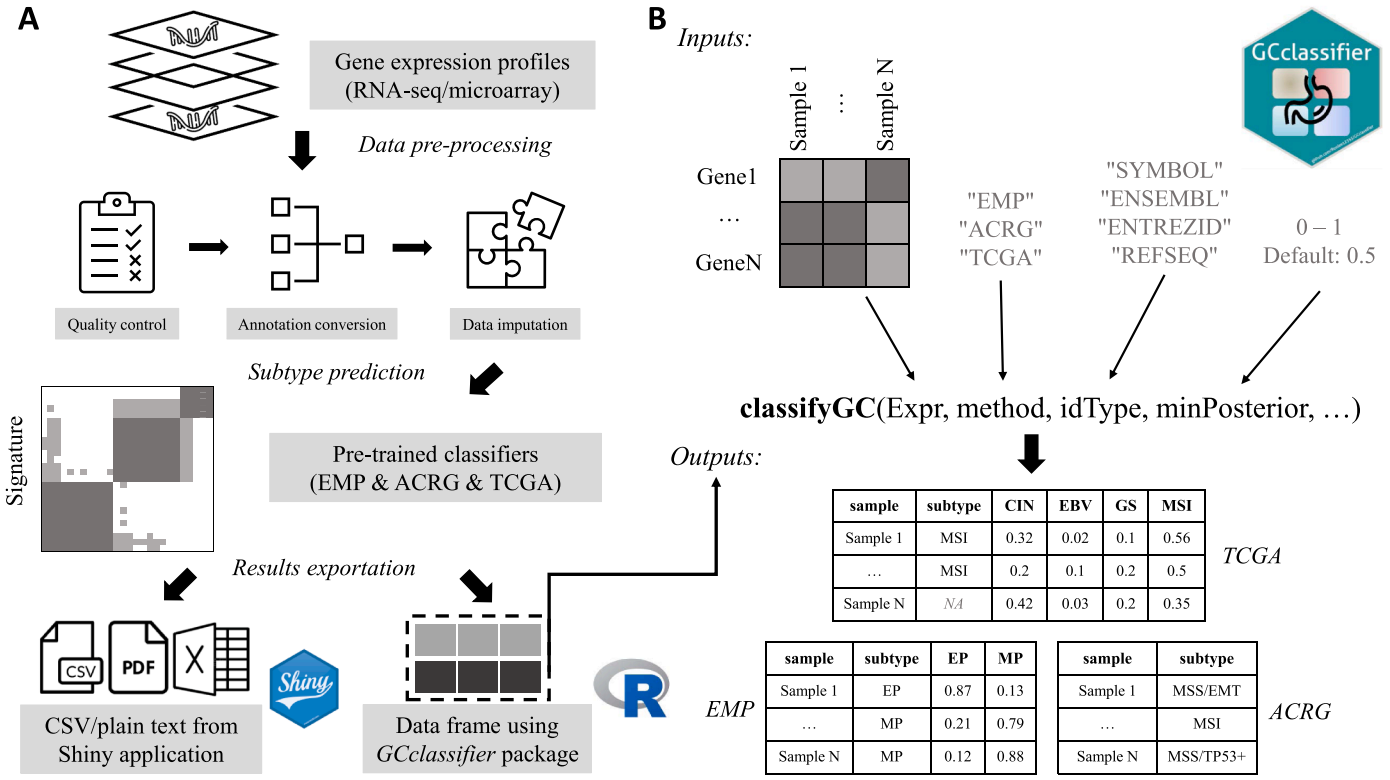


Fig. 1. A schematic workflow illustrating the design and major functionalities of *GCclassifier* package. (A) Processed gene expression profiles of GC were taken as input, and the molecular subtypes would be predicted after data quality control, gene annotation conversion, and missing data imputation. The results of subtype prediction could be exported diversely either using the package or the online Shiny application depending on the used platforms. (B) Possible choices for the four main parameters *Expr*, *method*, *idType* and *minPosterior* are shown.

the highest recurrence rate [15]. More recently, a systematic subtyping study based on integrative analysis of genomic and proteomic data identified two distinct molecular subtypes: mesenchymal phenotype (MP) and epithelial phenotype (EP) [12]. The MP subtype was characterized by worse overall survival (OS) and recurrence-free survival (RFS), lower genomic mutation, and poorer response to chemotherapy. These studies have identified molecular subtypes of GC with distinct biological features and clinical relevance, setting a strong foundation for better understanding the biology underlying GC subtypes and precision oncology.

However, the abovementioned GC subtyping frameworks are difficult to implement due to a lack of publicly available tools, and an easy-to-use computational platform for GC molecular subtype prediction is urgently called for. To address the unmet need, we developed an R package *GCclassifier*, which utilizes publicly available gene expression datasets and existing classification labels to build highly concordant classifiers implementing the subtyping systems proposed by TCGA, ACRG, and MD Anderson Cancer Center. Based on independent validations on multiple public datasets, we demonstrated the accuracy and robustness of *GCclassifier*. For general applicability, we developed an interactive Shiny application to enable subtype prediction without programming, making the package more user-friendly for clinicians and biologists in the broad areas of GC biology and precision oncology.

2. Materials and methods

2.1. Data collection and processing

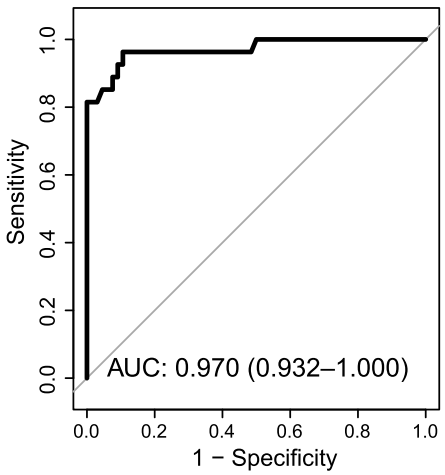
Level-3 gene expression profiles for GC in the TCGA-STAD dataset were obtained from the University of California Santa Cruz (UCSC) Xena data portal [16]. The expression profiles were converted from fragments per kilobase million (FPKM) to transcripts per million (TPM) values, followed by log2-transformation. Molecular subtype classification labels and signature genes were downloaded from the corresponding publications to train subtyping classifiers. Curated microarray-based gene expression profiles and clinical information of public datasets (GSE26899 [12], GSE26901 [12], GSE62254 [11], GSE26942 [17] and GSE13861 [12]) were downloaded from Gene Expression Omnibus (GEO). The maximal expression level was selected if multiple probes were mapped to the same gene symbol, and the missing values were imputed by the R package 'impute'. The detailed molecular subtype information of TCGA-STAD and other validation datasets were summarized in Supplementary Table S1. The gene signatures used in three classifiers were summarized in Supplementary Table S2.

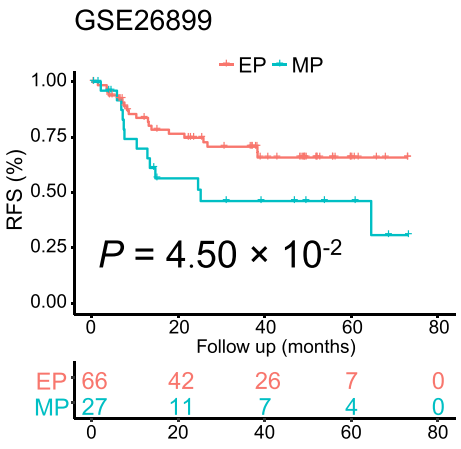
2.2. Classifier development

The development workflow was shown in Fig. 1A. This package implemented three major subtype classifiers for GC as described by ACRG [11], MD Anderson Cancer Center (epithelial and mesenchymal phenotype, EMP) [12] and TCGA [10,17]. For ACRG subtyping system, we utilized the identified signature genes described in the publication and then built a random forest classifier [18]. The classifier first distinguished MSI and MSS/EMT subtypes, and the remaining MSS/TP53- and MSS/TP53+ samples were stratified by the Youden index [19,20] of receiver operating characteristic curve (ROC) trained on the average Z-normalized expression levels of *TP53* signature genes (*MDM2* and *CDKN1A*) [11]. For training a classifier for the MD Anderson Cancer Center subtyping system, we prioritized 210 signature gene candidates commonly found in the microarray and RNA-seq data. Subsequently, a random forest classifier was constructed using the 210-gene panel, and the Youden index of the ROC trained on the TCGA-STAD dataset was utilized to stratify the EP/MP subtypes. For the TCGA taxonomy, the signature genes for each subtype were downloaded [17], and

A

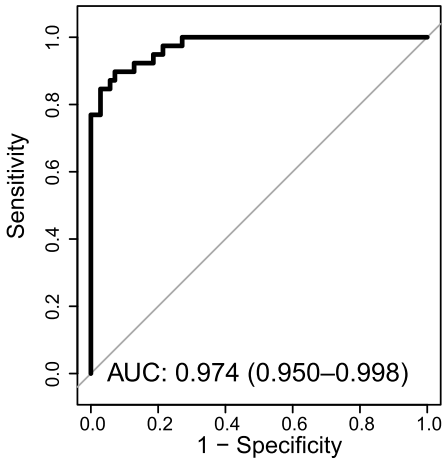
GSE26899 (n = 93)	Real MP	Real EP
Predict MP	23	4
Predict EP	4	62
Totals	27	66
Sensitivity (%)	85.2	
Specificity (%)		93.9

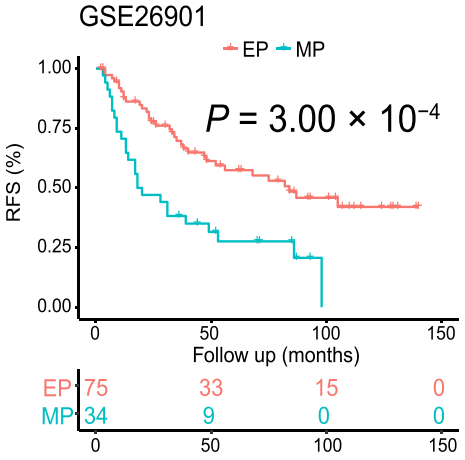




B

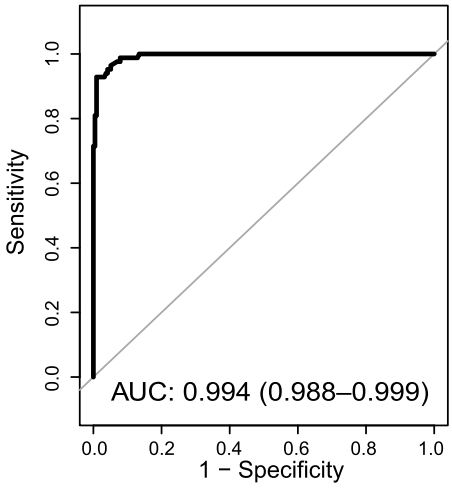
GSE26901 (n = 109)	Real MP	Real EP
Predict MP	32	2
Predict EP	7	68
Totals	39	70
Sensitivity (%)	82.1	
Specificity (%)		97.1





C

GSE62254 (n = 300)	Real MP	Real EP
Predict MP	72	2
Predict EP	12	214
Totals	84	216
Sensitivity (%)	85.7	
Specificity (%)		99.1



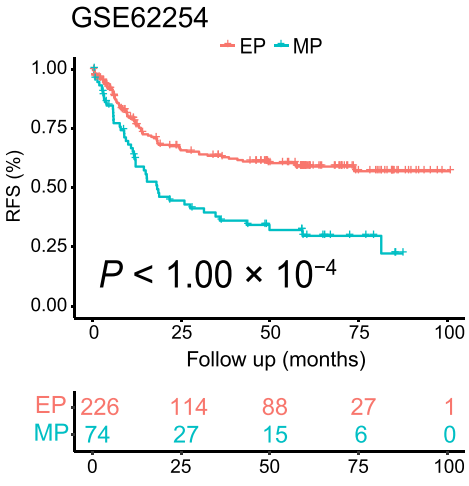
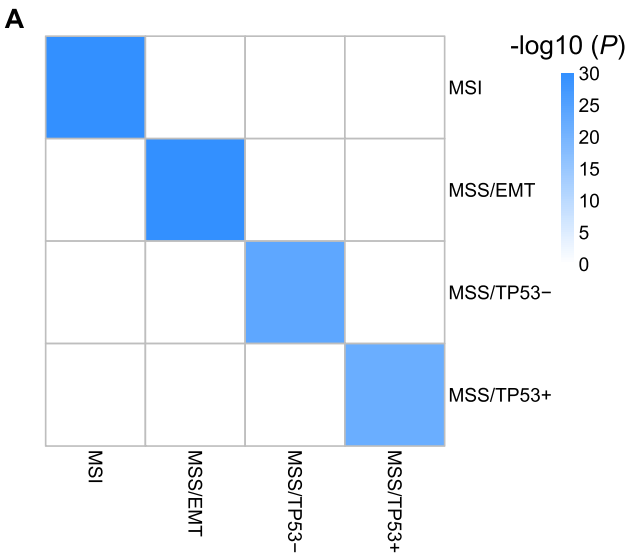


Fig. 2. Three case studies of EMP classification. The EMP subtypes were predicted by employing *GCclassifier* in GSE26899 (A), GSE26901 (B) and GSE62254 (C). The confusion matrices and ROC curves demonstrated high prediction accuracies and the provided evidence showed that the MP subtype was significantly associated with the worst RFS. In confusion matrix, true positives (TP) represent the number of correctly identified MP samples, false negatives (FN) represent the number of MP incorrectly classified as EP, true negatives (TN) represent the number of correctly identified EP, and false positives (FP) represent the number of EP incorrectly classified as MP. The definition of sensitivity is TP / (TP + FN) and specificity is TN / (TN + FP).



B

GSE62254 (n = 300)	Real MSI	Real MSS/EMT	Real MSS/TP53-	Real MSS/TP53+
Predict MSI	54	0	1	1
Predict MSS/EMT	7	46	18	9
Predict MSS/TP53-	2	0	67	13
Predict MSS/TP53+	5	0	21	56

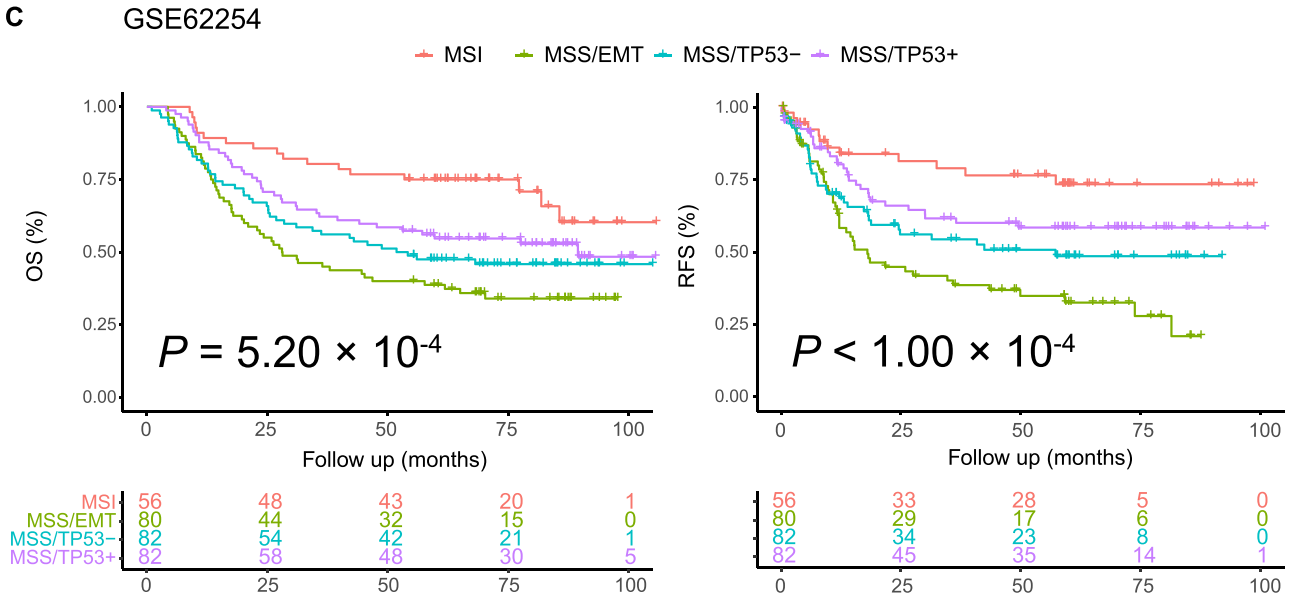


Fig. 3. A case study of ACRG classification. The hypergeometric test (A) and the confusion matrix (B) of reference and predicted labels showed high prediction consistency in GSE62254. (C) Further survival analysis using the KM method showed that the MSS/EMT subtype was consistently associated with the worst OS and RFS.

the top 200 significant genes in each subtype were selected and intersected with genes in the microarray datasets to ensure cross-platform applicability. A random forest classifier was then trained for subtype prediction. The training (TCGA-STAD) and validation datasets were Z-normalized by R function "scale" before classifier construction and subtype prediction. All models were trained with default parameters in R package "randomForest". The formula for Z-normalization was indicated as follows:

$$x^* = \frac{x - \mu}{\sigma}$$

Where x^* denoted as new value, x as original value, μ as the mean value of data, and σ as the standard deviation of data.

2.3. Data and code availability

The *GCclassifier* package was written in R language and released under GPL-3 License. The source code and documents are publicly available in the GitHub repository: https://github.com/Ronlee12355/GCclassifier. The web tool *GCclassifier* was implemented as a Shiny application and can be freely accessed online (https://compbio-cityuhk.shinyapps.io/GCclassifier_online/). All data used in the analysis are publicly available in TCGA, GEO and corresponding publications.

3. Results

3.1. Package implementations

The *GCclassifier* package was written and implemented in R language (version >= 4.1.0 required), consisting of four major functions:

- *classifyGC*: Perform the molecular subtype prediction for the customized input gene expression dataset (Fig. 1**B**). It could automatically detect whether log2-transformation should be performed, and missing gene expression data should be imputed.
- *classifyGC_interface*: Launch the internal Shiny interactive application in the default browser.

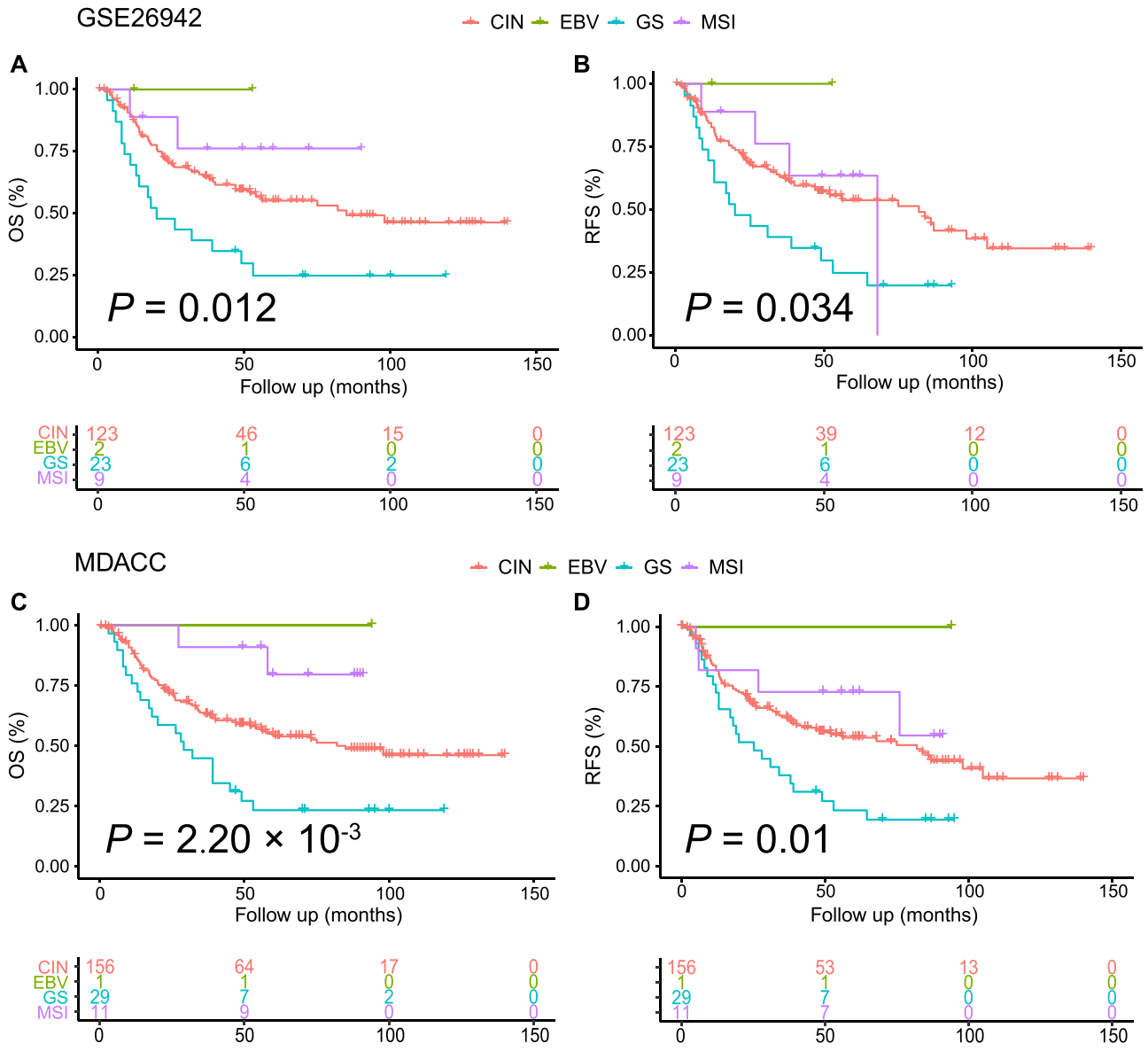


Fig. 4. Two case studies of TCGA classification. Patients were stratified by TCGA classification labels. Survival analysis showed that the GS subtype was significantly associated with the worst OS and RFS in GSE26942 (A-B), and MDACC (C-D).

- *probe_to_symbol*: Aggregate gene expression profile(s) from probeset IDs to gene identifiers for microarray-based datasets.
- *get_signature*: Extract signature genes from a specific GC molecular subtyping system.

The package includes example datasets (GSE62254 and GSE26901), with gene expression profiles and clinical information attached, to demonstrate the usage of exported functions. We also provide a complete vignette documentation (Supplementary File S1), including the explanation of function parameters, a step-by-step guide, and package dependencies.

3.2. Case studies

To assess the predictive performance of the package, we utilized four publicly available datasets with corresponding subtype classification labels (GSE26899, GSE26901, GSE62254, GSE26942 and GSE13861). Using gene expression profiles of these datasets, we predicted the GC subtypes of matched corresponding molecular subtyping systems using *GCclassifier* and evaluated the prediction performance accordingly. For EMP classification, we generated ROC curves and confusion matrices to demonstrate the prediction performance of the assigned labels and predicted probabilities. The accuracy for subtype prediction was promising in both GSE26899 (91.40% accuracy and area under the curve [AUC] = 0.97) (Fig. 2A), GSE26901 (91.74% accuracy and AUC = 0.97) (Fig. 2B) and GSE62254 (95.33% accuracy and AUC = 0.99) (Fig. 2C) datasets. Kaplan-Meier analysis of predicted labels revealed consistent survival differences in RFS, with the MP subtype samples exhibiting worse RFS compared to EP samples (all $P < 0.05$, log-rank tests). For ACRG classification, a hypergeometric test on the GSE62254 dataset demonstrated a significant association (all $P < 0.05$) and high subtype prediction accuracies in overall samples (~75%) and each subtype (sensitivity = 100% in MSS/EMT subtype and 79.41% in MSI subtype, respectively) were achieved compared to reference labels (Fig. 3A-B). Moreover, survival analysis confirmed the consistency in survival outcomes, with the MSS/EMT subtype patients showing the worst OS ($P = 5.20 \times 10^{-4}$, log-rank test) and RFS ($P < 1.00 \times 10^{-4}$, log-rank test) (Fig. 3C). For TCGA classification, the survival outcomes of predicted labels on GSE26942 and MDACC (GSE26942 + GSE13861) datasets were compared with a previous study [17], revealing consistent

findings that the GS subtype associated with the worst OS and RFS (all $P < 0.05$, log-rank tests) in GSE26942 (Fig. 4A-B) and MDACC (Fig. 4C-D). In conclusion, our *GCclassifier* package demonstrated robust performance in predicting molecular subtypes across independent public datasets, underscoring its significance as a robust tool for GC subtype prediction.

4. Discussion

the *GCclassifier* package, of its first kind, presents a highly robust and user-friendly approach for predicting GC molecular subtypes. With its streamlined input requirements and flexible operational modes, this tool offers a wide range of potential applications for precision oncology and the development of subtype-specific therapeutics. The ability to operate within the R software environment or as an interactive webpage enhances its accessibility to researchers and clinicians alike.

It is known that certain molecular classes of GC exhibit preferential associations with distinct Lauren histology classification subtypes, namely diffuse and intestinal. To further enhance the clinical credibility of our package, we have analyzed our test sets to examine the reproducibility of the association. Consistent with the reference publications [11,12,17], our results showed the MP subtype, the MSS/EMT subtype and GS subtype patients showed a significantly higher association with the diffuse subtype ($P < 0.05$, Chi-square test) (Supplementary Fig. S1). These patterns corroborate with well-established findings in the field, thereby lending additional credence to the robustness and clinical relevance of our model.

Recently, deep learning has been introduced to molecular subtyping, and we have explored the potential of deep learning by training a feed-forward multilayer artificial neural network and compared it with our random forest model. In the benchmark study, our random forest model outperformed in terms of sensitivity and accuracy (Supplementary Fig. S2), highlighting the robust generalization capabilities of our model. Deep learning models typically require a large volume of data to effectively learn and generalize. Considering the relatively small number of gene expression features, using a random forest model for classification suffices, as it consumes fewer computing resources compared to a deep learning model. Additionally, the limited size of the training dataset cannot guarantee the training effectiveness of a deep learning model. Together, our study contributes to the growing understanding that deep learning is not a one-size-fits-all solution and underscores the importance of matching the model to the dataset characteristics to achieve the best possible outcomes.

Moreover, we would like to acknowledge some limitations of our study. First, for ACRG classification, the MSS/TP53- and MSS/TP53+ subtypes were classified by the Youden index value of a two-gene *TP53* signature, and such strategy was relatively small in terms of feature size compared to the other subtypes, thus hindering its capacity to discriminate the two *TP53* related subtypes. Second, it is noteworthy to mention that the lack of reference subtype labels in the TCGA classification, as observed in independent studies, posed a challenge in assessing the performance of our model. Consequently, the ability to further substantiate the prediction reliability of our model was hindered.

In summary, our findings suggest that the *GCclassifier* package holds promise in enhancing molecular subtype prediction for GC and serves as a valuable tool in precision oncology investigations of GC.

Funding

This work was supported by a grant from Guangdong Basic and Applied Basic Research Foundation (Project No. 2019B030302012), a grant from Shenzhen Science, Technology and Innovation Commission (Project No. 基2020N368), a startup fund (Project No. 4937084), and direct grant (2021.077) from the Chinese University of Hong Kong, grants from the Research Grants Council (Project No. 11103619, 11103921, 14111522, 14104223, C4024-22GF, R4007-23) of the Hong Kong Special Administrative Region, China, awarded to Xin Wang. This work was also partially sponsored by Shenzhen Bay Scholars Program awarded to Xin Wang.

CRediT authorship contribution statement

Jiang Li: Conceptualization, Software, Methodology, Investigation, Formal analysis, Visualization, Data curation, Writing – original draft. **Lingli He:** Visualization, Writing - review & editing. **Xianrui Zhang:** Writing - review & editing. **Xiang Li:** Writing – review & editing. **Lishi Wang**: Writing - review & editing. **Zhongxu Zhu:** Writing - review & editing. **Kai Song:** Writing - review & editing. **Xin Wang:** Conceptualization, Writing - review & editing, Funding acquisition, Project administration, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors gratefully acknowledge the valuable contributions of all members of Dr. Xin Wang's laboratory, who provided insightful feedback and shared their expertise in implementing the package and designing the Shiny webpage.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.01.010.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71:209–49.
[2] Ilic M, Ilic I. Epidemiology of stomach cancer. World J Gastroenterol 2022;28:1187–203.
[3] Joshi SS, Badgwell BD. Current treatment and recent progress in gastric cancer. CA Cancer J Clin 2021;71:264–79.
[4] Bang Y-J, Kim Y-W, Yang H-K, Chung HC, Park Y-K, Lee KH, et al. Adjuvant capecitabine and oxaliplatin for gastric cancer after D2 gastrectomy (CLASSIC): a phase 3 open-label, randomised controlled trial. Lancet 2012;379:315–21.
[5] Aoyama T, Yoshikawa T, Watanabe T, Hayashi T, Ogata T, Cho H, et al. Survival and prognosticators of gastric cancer that recurs after adjuvant chemotherapy with S-1. Gastric Cancer 2011;14:150–4.
[6] Smyth EC, Nilsson M, Grabsch HI, van Grieken NCT, Lordick F. Gastric cancer. Lancet 2020;396:635–48.
[7] Jiang Y, Xie J, Huang W, Xi S, Li T, Chen C, et al. Overtreatment of younger adults with gastric cancer: More chemotherapy use with unmatched survival gains. SSRN Electron J 2019. https://doi.org/10.2139/ssrn.3393684.
[8] Liu N, Molena D, Stem M, Blackford AL, Sewell DB, Lidor AO. Underutilization of treatment for regional gastric cancer among the elderly in the USA. J Gastrointest Surg 2018;22:955–63.
[9] Ho SWT, Tan P. Dissection of gastric cancer heterogeneity for precision oncology. Cancer Sci 2019;110:3405–14.
[10] Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature 2014;513:202–9.
[11] Cristescu R, Lee J, Nebozhyn M, Kim K-M, Ting JC, Wong SS, et al. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nat Med 2015;21:449–56.
[12] Oh SC, Sohn BH, Cheong J-H, Kim S-B, Lee JE, Park KC, et al. Clinical and genomic landscape of gastric cancer with a mesenchymal phenotype. Nat Commun 2018;9. https://doi.org/10.1038/s41467-018-04179-8.
[13] Tan IB, Ivanova T, Lim KH, Ong CW, Deng N, Lee J, et al. Intrinsic subtypes of gastric cancer, based on gene expression pattern, predict survival and respond differently to chemotherapy. Gastroenterology 2011;141:476–85. , 485.e1-11.
[14] Lei Z, Tan IB, Das K, Deng N, Zouridis H, Pattison S, et al. Identification of molecular subtypes of gastric cancer with different responses to PI3-kinase inhibitors and 5-fluorouracil. Gastroenterology 2013;145:554–65.
[15] Liao P, Jia F, Teer JK, Knepper TC, Zhou H-H, He Y-J, et al. Geographic variation in molecular subtype for gastric adenocarcinoma. Gut 2019;68:1340–1.

[16] Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol 2020; 38:675–8.
[17] Sohn BH, Hwang J-E, Jang H-J, Lee H-S, Oh SC, Shim J-J, et al. Clinical significance of four molecular subtypes of gastric cancer identified by the Cancer Genome Atlas project. Clin Cancer Res 2017;23:4441–9.
[18] Kaihara M, Kikuchi S. Discriminant analysis of countries growing wakame seaweeds: a preliminary comparison of visible-near infrared spectra using soft independent modelling, randomforests and classification and regression trees. J Infrared Spectrosc 2007;15:371–7.
[19] Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. Biom J 2005;47:458–72.
[20] Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. Biom J 2008;50:419–30.