npj | precision oncology www.nature.com/npjprecisiononcology

ARTICLE OPEN

Towards precision oncology discovery: four less known genes and their unknown interactions as highest-performed biomarkers for colorectal cancer

Yongjun Liu[1], Yuqing Xu[2], Xiaoxing Li[3]✉, Mengke Chen[3], Xueqin Wang[4], Ning Zhang[5], Heping Zhang[6] and Zhengjun Zhang[2,7,8]✉

The goal of this study was to use a new interpretable machine-learning framework based on max-logistic competing risk factor models to identify a parsimonious set of differentially expressed genes (DEGs) that play a pivotal role in the development of colorectal cancer (CRC). Transcriptome data from nine public datasets were analyzed, and a new Chinese cohort was collected to validate the findings. The study discovered a set of four critical DEGs - CXCL8, PSMC2, APP, and SLC20A1 - that exhibit the highest accuracy in detecting CRC in diverse populations and ethnicities. Notably, PSMC2 and CXCL8 appear to play a central role in CRC, and CXCL8 alone could potentially serve as an early-stage marker for CRC. This work represents a pioneering effort in applying the max-logistic competing risk factor model to identify critical genes for human malignancies, and the interpretability and reproducibility of the results across diverse populations suggests that the four DEGs identified can provide a comprehensive description of the transcriptomic features of CRC. The practical implications of this research include the potential for personalized risk assessment and precision diagnosis and tailored treatment plans for patients.

npj Precision Oncology (2024)8:13 ; https://doi.org/10.1038/s41698-024-00512-1

INTRODUCTION

Colorectal cancer (CRC) is a significant public health issue, being one of the most prevalent human malignancies worldwide and the second leading cause of cancer-related deaths[1–3]. While surgical resection, chemoradiation, and immunotherapy have advanced, they remain inadequate in many cases. Moreover, the incidence of CRC is increasing in younger individuals, particularly in the United States and other countries[4–6]. Genetic predisposition plays a crucial role in the development of CRC, with hereditary and sporadic causes accounting for a significant proportion of cases[2,6,7]. The etiology of CRC can be broadly classified into two categories: hereditary or sporadic. Hereditary CRC accounts for 10−15% of the overall incidence and is attributable to mutations in APC or DNA mismatch repair genes. Sporadic CRC is more frequent, representing >80% of CRCs, and is characterized by chromosomal instability, microsatellite instability (MSI), or CpG island methylation[6,7].

Over the past decades, many transcriptomic studies have been performed which have shed light on the molecular mechanisms underlying CRC development, with a large number of genes being identified as differentially expressed between tumor and non-tumor tissues[8–15]. At the molecular level, CRC are classified into four consensus molecular subtypes (CMS), each of which is characterized by distinct expression profiles of oncogenic/tumor suppressive genes and pathways, mutation states of particular genes, MSI, and clinical outcomes[16,17], however their clinical utility remains to be validated.

So far, most transcriptomic studies have used traditional analytical approaches which rely on fold changes of individual genes between tumor and control tissues or pathway enrichment analysis based on current knowledge of genes and biological processes[11,18–20]. As a result, the number of genes/transcripts reported is large and it is uncertain which of them plays a critical role in cancer identification and classification. Furthermore, gene-gene interactions were not well addressed in traditional analytical models. Thus, there is a need to develop novel analytical methods to identify critical DEGs with high sensitivity and specificity. Recent advances in the machine learning community have shown great promise for applying new methods to improve cancer identification/classification and have demonstrated superior performance over traditional methods[21–23].

In this study, we applied a newly proven and powerful machine-learning method to identify a parsimonious subset of critical differentially expressed genes (DEGs) for CRC. Our method is based on the max-logistic competing structure, which takes into account the competing relationships among genes in predicting the outcome variable, including gene-gene interactions, a feature not captured by traditional analytical models[21–23]. We analyzed ten transcriptome profiling datasets, including nine public datasets and one separate transcriptome dataset collected from a Chinese population. Using the max-logistic competing risk factor models, we identified four critical DEGs, namely, CXCL8 (C-X-C Motif Chemokine Ligand 8), PSMC2 (Proteasome 26S Subunit, ATPase 2), APP (Amyloid Beta Precursor Protein), and SLC20A1

[1]Department of Laboratory Medicine and Pathology, University of Washington Medical Center, Seattle, WA, USA. [2]Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA. [3]State Key Laboratory of Oncology in South China, Sun Yat-sen University Cancer Center, Guangzhou, Guangdong, China. [4]Department of Statistics and Finance, University of Science and Technology of China, Hefei, China. [5]Department of Gastroenterology, First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. [6]Yale School of Public Health, Yale University, New Haven, CT, USA. [7]Department of Biostatistics and Medical Informatics, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI, USA. [8]Present address: School of Economics and Management, and MOE Social Science Laboratory of Digital Economic Forecasts and Policy Simulation, University of Chinese Academy of Sciences, Center for Forecasting Sciences, Chinese Academy of Sciences, Beijing, China. ✉email: Lixiaox23@mail.sysu.edu.cn; zjz@stat.wisc.edu

performed in the TCGA Colon Cancer (COAD) cohort using the Illumina HiSeq 2000 RNA Sequencing platform[46]. Gene expression values were $\log_2$(norm count+1) transformed. The "normal" samples were adjacent nontumor colorectal tissues. The data link is https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-COAD.htseq_counts.tsv.gz. We noted that this dataset was expanded from 349 samples to 512 samples since we first downloaded from the website. We used 349 samples in our initial data analysis. In our analysis of the second dataset, we used 512 samples which had different measurements from the first dataset (see below).

The second public dataset was also obtained from the NCI's GDC and included 471 CRC samples and 41 normal controls (a total of 512 samples). This RNA-seq study was performed in the TCGA COAD cohort using the Illumina HiSeq 2000 RNA Sequencing platform[46]. The expression values were normalized with $\log_2$(Fragments Per Kilobase of transcript per Million mapped reads (FPKM) + 1). In our computation, the expression values were further transformed using a natural logarithm operator. The "normal" samples were adjacent normal colon/rectal tissues. The data link is: https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-COAD.htseq_fpkm.tsv.gz.

The first dataset and the second dataset had measurement heterogeneity which may cause batch effects when applying classical statistical models and classifications. Our newly proposed max-logistic competing regression can overcome the batch effects by logarithm transformation of the second dataset (see below for details).

The third public dataset (GEO Accession: GSE39582) was obtained from a study performed in Europe using the Affymetrix Human Genome U133 Plus 2.0 Array platform[11]. This dataset included 566 CRC samples and 19 normal controls with 54,675 genes/transcripts. The expression values were derived from $\log_2$(normalized intensity signal). This dataset included frozen tissue of primary colorectal adenocarcinoma and its "normal" samples were frozen tissue of non-tumoral colorectal mucosa.

The fourth public dataset (GEO Accession: GSE9348) was obtained from a study performed in a Han Chinese CRC cohort including 82 age-, ethnicity- and tissue-matched healthy controls using the Affymetrix U133 Plus 2 array[47]. The patients were classified as early-stage CRC (Stage 1 or 2). Gene expression values were calculated using the MAS5 algorithm. This dataset included tumor tissue collected and archived within 30 minutes after surgery, and its "normal" was colonic mucosa collected and archived within 30 minutes after biopsy.

The fifth public dataset (GEO Accession: GSE18105) was obtained from a study performed in a Japanese cohort containing 77 CRC samples and 17 paired samples from adjacent nontumor tissues[48]. The patients were classified as stages 2 or 3 CRC. Gene expression values were derived from RMA signal intensities.

The sixth public dataset (GEO Accession: GSE41258) was obtained from a study performed in an Israel population containing 299 samples, including 180 CRC, 46 polyps, 43 normal colon, 21 liver metastases, and 9 lung metastases[49]. Data were normalized using the PLIER algorithm and batch corrected, then Lowess normalized signals.

To validate the results of the discovery analysis, we collected a sample at Sun Yat-sen University Cancer Center in Guangzhou, China, which included 45 CRC samples and 47 normal controls collected from adjacent nontumor colonic tissues (seventh dataset). The genes identified in the discovery analysis were validated using real-time quantitative RT-PCR with the TaqMan Gene Expression assays (Applied Biosystems, Inc.). In addition, the patients' age, sex, TNM tumor stage, and histologic grade of the tumor were included in the analyses. Regarding this new data collection, the following procedure had been utilized. • Complying with the 'Guidance of the Ministry of Science and Technology (MOST) for the Review and Approval of Human Genetic Resources', this project obtained approval from the institutional ethics committee (IEC) of Sun Yat-sen University Cancer Center. Experimental procedures and data collection involving Chinese patients were conducted in China with the participation of Chinese co-authors. • This study obtained approval from the institutional ethics committee (IEC) of Sun Yat-sen University Cancer Center, adhering to the principles of the Declaration of Helsinki. All experimental procedures were conducted in compliance with the guidelines and regulations for the protection of human subjects. Informed consent was obtained from all patients prior to their participation in the study. • No sequencing experiment was performed using the clinical samples collected from Sun Yat-sen University Cancer Center in Guangzhou, China. • Participants provided written informed consent to take part in all studies conducted by Sun Yat-sen University Cancer Center. • The four gene expression values are made available for download from the datalink specified in Code availability.

Three additional public datasets (i.e., eighth, ninth and tenth) were included in this study to further validate the findings and to fit our models into preclassified CMS.

The eighth public dataset was also obtained from the NCI's GDC and included 167 rectal cancer samples and 10 normal controls (adjacent normal rectal tissues). The data link is: https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-READ.htseq_fpkm.tsv.gz. Data from the same sample but from different vials/portions/analytes/aliquotes was averaged; data from different samples was combined into genomicMatrix. This RNA-seq study was performed in the TCGA COAD cohort using the Illumina HiSeq 2000 RNA Sequencing platform[46]. The expression values were normalized with $\log_2$(Fragments Per Kilobase of transcript per Million mapped reads (FPKM) + 1).

The ninth public dataset (GEO Accession: GSE103512) contained formalin-fixed and paraffin-embedded normal and tumor tissues of four cancer types, in which colon cancer was included. The platform used was Affymetrix HT-U133plus-2-PM microarrays[50].

In this study, 57 CRC samples and 12 matched normal colon samples were analyzed[50]. The tenth public dataset (GEO Accession: GSE156451) contained tumor tissues from 72 CRC and adjacent nontumor colorectal tissues from a Chinese population. The platform was Illumina NovaSeq 6000 (Homo sapiens)[17].

CODE AVAILABILITY

A MATLAB® R13 demo code for solving Eq. (4) ($\lambda 2 = 0$) and readme files are also available. The final dataset organized from the original datasets and our computer program generated datasets and formulas are also available at https://pages.stat.wisc.edu/~zjz/POsubmissions.zip for review. Readers can also make requests by sending emails to the corresponding authors if the link is not working.

Received: 2 June 2023; Accepted: 5 January 2024;
Published online: 20 January 2024

REFERENCES

1. Jemal, A. et al. Global cancer statistics. *CA: Cancer J. Clin.* **61**, 69–90 (2011).
2. Keum, N. & Giovannucci, E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nat. Rev. Gastroenterol. Hepatol.* **16**, 713–732 (2019).
3. Edwards, B. K. et al. Annual report to the nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates. *Cancer.: Interdiscip. Int. J. Am. Cancer. Soc.* **116**, 544–573 (2010).
4. Patel, P. & De, P. Trends in colorectal cancer incidence and related lifestyle risk factors in 15-49-year-olds in Canada, 1969-2010. *Cancer Epidemiol.* **42**, 90–100 (2016).
5. Siegel, R. L. et al. Colorectal cancer incidence patterns in the United States, 1974-2013. *J. Natl. Cancer Inst.* **109**, djw322 (2017).
6. Young, J. P. et al. Rising incidence of early-onset colorectal cancer in Australia over two decades: report and review. *J. Gastroenterol. Hepatol.* **30**, 6–13 (2015).
7. Fearon, E. R. Molecular genetics of colorectal cancer. *Annu. Rev. Pathol.: Mech. Dis.* **6**, 479–507 (2011).
8. Dienstmann, R. et al. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nat. Rev. cancer* **17**, 79–92 (2017).
9. Schlicker, A. et al. Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC Med. Genom.* **5**, 1–15 (2012).
10. Roepman, P. et al. Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *Int. J. Cancer* **134**, 552–562 (2014).
11. Marisa, L. et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med.* **10**, e1001453 (2013).
12. Sadanandam, A. et al. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat. Med.* **19**, 619–625 (2013).
13. De Sousa, E. M. F. et al. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nat. Med.* **19**, 614–618 (2013).
14. Wang, H., Wang, X., Xu, L., Zhang, J. & Cao, H. Analysis of the transcriptomic features of microsatellite instability subtype colon cancer. *BMC Cancer* **19**, 1–16 (2019).
15. Budinska, E. et al. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *J. Pathol.* **231**, 63–76 (2013).
16. Guinney, J. et al. The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).
17. Li, Q. L. et al. Genome-wide profiling in colorectal cancer identifies PHF19 and TBC1D16 as oncogenic super enhancers. *Nat. Commun.* **12**, 6407 (2021).
18. Alon, U. et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci.* **96**, 6745–6750 (1999).
19. Kim, E. et al. Upregulation of SLC2A3 gene and prognosis in colorectal carcinoma: analysis of TCGA data. *BMC Cancer* **19**, 1–10 (2019).
20. Salazar, R. et al. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. *J. Clin. Oncol.* **29**, 17–24 (2011).