OXFORD

Briefings in Bioinformatics, 2023, **24(3)**, 1–13
https://doi.org/10.1093/bib/bbad078
Advance access publication date 22 March 2023
Case Study

Prioritizing prognostic-associated subpopulations and individualized recurrence risk signatures from single-cell transcriptomes of colorectal cancer

Mengsha Tong†, Yuxiang Lin†, Wenxian Yang, Jinsheng Song, Zheyang Zhang, Jiajing Xie, Jingyi Tian, Shijie Luo, Chenyu Liang, Jialiang Huang and Rongshan Yu

Corresponding authors: Mengsha Tong. Tel.: +86-0592-2181809; E-mail: mstong@xmu.edu.cn; Jialiang Huang. Tel.: +86-0592-2181809; E-mail: jhuang@xmu.edu.cn; Rongshan Yu. Tel.: +86-0592-2580132; E-mail: rsyu@xmu.edu.cn

†Mengsha Tong and Yuxiang Lin contributed equally.

Abstract

Colorectal cancer (CRC) is one of the most common gastrointestinal malignancies. There are few recurrence risk signatures for CRC patients. Single-cell RNA-sequencing (scRNA-seq) provides a high-resolution platform for prognostic signature detection. However, scRNA-seq is not practical in large cohorts due to its high cost and most single-cell experiments lack clinical phenotype information. Few studies have been reported to use external bulk transcriptome with survival time to guide the detection of key cell subtypes in scRNA-seq data. We proposed scRank^{XMBD}, a computational framework to prioritize prognostic-associated cell subpopulations based on within-cell relative expression orderings of gene pairs from single-cell transcriptomes. scRank^{XMBD} achieves higher precision and concordance compared with five existing methods. Moreover, we developed single-cell gene pair signatures to predict recurrence risk for patients individually. Our work facilitates the application of the rank-based method in scRNA-seq data for prognostic biomarker discovery and precision oncology. scRank^{XMBD} is available at https://github.com/xmuyulab/scRank-XMBD. (XMBD:Xiamen Big Data, a biomedical open software initiative in the National Institute for Data Science in Health and Medicine, Xiamen University, China.)

Keywords: single-cell gene pair signatures, relative expression orderings, colorectal cancer, recurrence risk signatures

INTRODUCTION

Colorectal cancer (CRC) is the third most common cancer and a leading cause of morbidity and mortality worldwide [1]. The implementation of curative resection and adjuvant chemotherapy has improved the overall prognosis of CRCs. However, about 25–40% of patients would relapse after primary radical resection [2]. The tumor-node-metastasis (TNM) staging system from International American Joint Committee on Cancer/Union for International Cancer Control (AJCC/UICC) remains the most important guideline for classifying patients and making therapeutic decisions. Unfortunately, due to high heterogeneity in CRC, the clinical outcomes for patients of the same stage can be very different [2]. Therefore, there is an urgent clinical need for molecular biomarkers that predict early relapse in CRC patients for more precise patient stratification.

The interpatient heterogeneity of CRC has been revealed by genomic and epigenetic analysis, gene expression profiles and tumor microenvironment (TME). At the genetic level, several DNA biomarkers including microsatellite instability (MSI), BRAF and KRAS mutations, CpG island methylator phenotype and chromosomal instability have been reported [3]. At the transcriptome level, Guinney *et al.* [4] proposed four consensus molecular subtypes (CMSs) with different molecular and clinical features. However, the translational values of these molecular markers remain unclear. It has been reported that MSI only occurs in a small proportion of CRC patients [5]. CMS has not been

Mengsha Tong, PhD, is an Assistant Professor of School of Life Sciences, National Institute for Data Science in Health and Medicine, Xiamen University. Her research is focused on discovering biomarkers for precision oncology using bioinformatics.
Yuxiang Lin is a PhD student in National institute for Data Science in Health and Medicine, Xiamen University, His research focuses on multiomics data analysis in bioinformatics.
Wenxian Yang is the CTO of Aginome Scientific. Her research interests include signal processing, data analytics and bioinformatics.
Jinsheng Song is a undergraduate student in School of Life Sciences, Xiamen University. His research focuses on multiomics data analysis in bioinformatics.
Zheyang Zhang is a PhD student in School of Life Sciences, Xiamen University. His research focuses on single-cell multi-omics integrative analysis and cancer biology.
Jiajing Xie is a PhD student in National institute for Data Science in Health and Medicine, Xiamen University. Her research focuses on multi-omics data analysis in bioinformatics.
Jingyi Tian is a Master Degree Candidate in School of Life Sciences, Xiamen University. Her research focuses on epigenetic data analysis of developmental and cancer biology.
Chenyu Liang is a Master Degree Candidate in School of Life Sciences, Xiamen University. His research focuses on single-cell transcriptome data in bioinformatics.
Shijie Luo is a master student in School of Life Science, Xiamen University. His research focuses on single-cell multiomics analysis.
Jialiang Huang is a Professor at the School of Life Sciences, Xiamen University. His research interests include bioinformatics and epigenomics.
Rongshan Yu is currently with National Institute for Data Science in Health and Medicine, School of Informatics, Xiamen University, as a professor. His research interests include statistical signal processing and its applications in bioinformatics.
Received: October 10, 2022. **Revised:** January 11, 2023. **Accepted:** February 11, 2023

adequately validated for clinical use [6, 7]. On the other hand, it is widely recognized that the recurrence risk for early-stage CRC was associated with TME infiltration patterns [8–11], from there more reliable biomarkers for early relapse in CRC patients can be derived. The TME of CRC consists of distinctive cell subpopulations, including tumor epithelial cells, cancer-associated fibroblasts (CAFs) and immune cells [12]. Stromal cell signatures detected from bulk transcriptomes were reported to be associated with risk of relapse and predict survival time of patients [8]. To explore the prognostic values of these multi-level biological features of CRC, Dienstmann *et al.* [9] developed a multivariable Cox model for disease-free survival (DFS). The results showed that TME infiltration patterns could represent potent determinants of the recurrence risk for early-stage CRC. The diversity of the TME also could be used for classification of cancers regarding prognosis [13]. Bruni *et al.* [13] summarized the prognostic significance of the major immune components in CRC. Pagès *et al.* proposed an immunoscore system based on the density of CD3+, CD8+ or CD45RO+ lymphocytes obtained from immunohistochemistry staining. They found that the immunoscore showed higher prognostic value than the TNM stage [10, 11]. However, the C-indexes of this system for relapse-free survival and overall survival (OS) were only 0.62 and 0.58 in their benchmarking cohorts [10].

Traditional transcriptomes largely depended on the analysis of bulk tissues, which obscures the signals of distinct cell subtypes. Single-cell RNA sequencing (scRNA-seq) provides detailed characterization of the heterogeneity of cell transcriptomes, allowing the assessment of the complex TME [14, 15]. scRNA-seq studies have been carried out in CRC [16–19], and the main results of the applications of single-cell RNA technology in CRC have been summarized in Supplementary Table S1. These studies provide valuable data source and help in understanding the heterogeneity of TME of CRC. However, only a few cell types have been reported to be prognostic and few studies focused on developing TME-related prognostic models using scRNA-seq data. Application of scRNA-seq in translational research remains a challenge, partly due to the low library sizes, high noise level and a large number of dropouts in scRNA-seq data, which might induce large experimental batch effects. Tan and Cahan [20] developed a method to annotate cell types by comparing the expression of pairs of genes within each cell. The relative expression orderings (REOs) of genes within a sample are robust against batch effects of experiments [21] and are not affected by the normalization of datasets [22, 23], making them promising for developing prognostic models using bulk transcriptome data [24, 25]. Wang *et al.* [26] proposed single-cell pairwise gene expression (scPAGE) to improve bulk RNA-seq data classification in acute myeloid leukemia. The application of rank-based method in developing prognostic models from scRNA-seq data remains to be investigated. To the best of our knowledge, REO-based prognostic risk models from scRNA-seq in CRC have not yet been studied. In this study, we developed scRank^{XMBD}, a data analysis framework to detect prognostic subpopulations and identify individualized recurrence risk signatures based on within-cell REOs of gene pairs to improve the risk stratification of CRC patients with stages II–III.

MATERIAL AND METHODS

Preprocessing for scRNA-seq datasets

We collected three scRNA-seq datasets from Gene Expression Omnibus (GEO) (Supplementary Table S2). The Seurat package (https://satijalab.org/seurat/articles/get_started.html) was used to pre-process each scRNA expression profile. Briefly, we filtered out genes measured in less than three cells. We further removed low-quality cells with small number of measured genes (<200 genes) and doublets with more than 6000 genes. Cells with more than 20% mitochondrial gene expression in gene counts were also removed.

Dimension reduction and annotation

The classic workflow in Seurat was used to perform dimension reduction and unsupervised clustering for each of the scRNA-seq datasets. Notably, the optimal value of parameter 'resolution' of *FindClusters* was determined by clustree [27] package. To ensure that annotation results could be compared among different scRNA-seq datasets, we combined manual and supervised annotations (Figure 1, Step1). In the training dataset (GSE144735), we identified seven major cell subpopulations according to the expression levels of classic markers (Supplementary Table S3). Then, we applied the same workflow in each major cell type to further reduce dimension and clustered cells into subpopulations based on several classic markers (Supplementary Table S3). As tumor-derived epithelial cells were heterogeneous in different patients, we followed the annotation method from Lee *et al.* [17] and used *classifyCMS* in the CMSclassifier [4] package to assign a CMS label to each tumor-derived epithelial cell. The Single Sample Predictor (SSP) was chosen as the classifier. Furthermore, we performed single sample Gene Set Enrichment Analysis (ssGSEA) on the CMS-related pathway gene sets from the CMScaller package [28] to confirm the annotation results. To apply the established landscapes of cell subtypes to two independent scRNA-seq datasets (GSE132465 and GSE132257), we used SciBet [29], a supervised cell type annotation toolkit, to predict cell identities for cells from training set to query sets.

Bulk transcriptomic data and preprocessing

We downloaded the CEL format files of ten bulk microarray datasets from GEO (Supplementary Table S2). The expression values were quantile normalized and log2 transformed by using the *justRMA* function in the *affy* package [30]. We further matched the probes with gene symbols in expression profiles according to the platform probes file, respectively. Probes matched to no or more than one gene were deleted. For a gene mapped to multiple probes, the arithmetic mean of the values of the multiple probes was calculated for its expression value. Finally, we downloaded clinical information files including TNM stages, gender, age, adjuvant chemotherapy and survival information (Supplementary Table S2). In this study, we only focused on the CRC patients at stage II/III and did not receive chemotherapy after surgery.

Overview of scRank^{XMBD} construction

(i) Identification of cell type specific gene pairs (GPs) with cell subpopulation-classifying value.

Tumor-derived cells were used to identify cell-subtype-specific gene pairs (C-GPs). We filtered genes according to the following criteria: detected in at least 5% of cells; the average expression value within cell type was larger than a predefined threshold (0.01 in this study). Next, we used linear least square regression as implemented in *GetClassGenes* of SingleCellNet [20], and selected top 100 discriminative feature genes for each cell type based on their regression coefficients (Figure 1A, Step2). Then, we constructed GPs from these 100 genes, and denoted a GP as 1 if the expression of the first gene of this GP was greater than that of the second gene (gene1 > gene2), and 0 (gene1 < gene2) otherwise. We then performed linear least square regression again on GPs

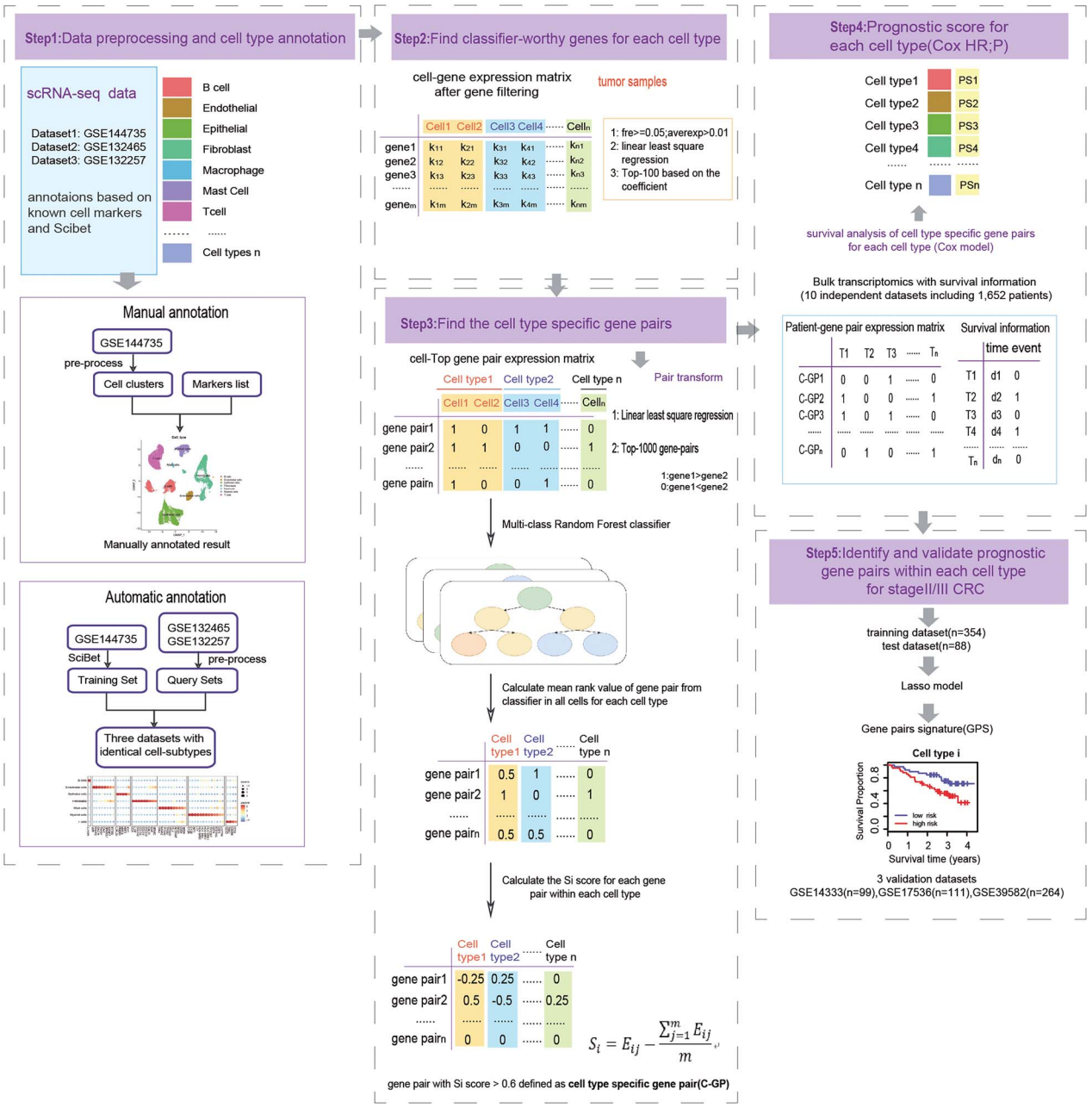

Figure 1. Workflow in this study.

using *ptGetTop* of SingleCellNet, and selected top 1000 GPs based on their regression coefficients (Figure 1A, Step3). Based on this filtered gene pair expression matrix, we trained a random forest (RF) classifier for each cell subtype. For GPs in RF classifier of cell subtypes, we calculated the specific score S_{ij} to indicate how its average REO in cell subtype j deviates from its average REO across all cell subtypes as follows:

$$S_{ij} = E_{ij} - \frac{\sum_{j=1}^{m} E_{ij}}{m}, \quad (1)$$

where E_{ij} is the average REO for GP i cell subtype j, which is calculated as

$$E_{ij} = \frac{\sum_{c \in C_j} R_i(c)}{|C_j|} \quad (2)$$

where $|C_j|$ denotes the number of cells of subtype j, and $R_i(c)$ represents the REO of i in cell c, respectively. Finally, we select GP i as a cell-subtype-specific pair (C-GP) for subtype j if$S_i \geq 0.6$.

(ii) Evaluation of prognostic value for each cell type.

Let Ga and Gb represent the expression values of gene a and gene b, respectively. We applied the univariate Cox proportional-hazards regression model to evaluate the correlation of the REO pattern (Ga > Gb or Ga < Gb) of each C-GP with the recurrence survival time of CRC patients. *P* value was adjusted by the Benjamini and Hochberg (BH) method. Furthermore, to obtain a sufficient number of C-GPs to evaluate the prognostic value for all cell types, C-GPs with adjusted *P* value less than 0.2 were defined as recurrence related. Next, we applied the C-GPs of T cells, epithelial cells, endothelial cells and fibroblasts to a bulk RNA-seq with fluorescence-activated cell sorting (GSE39396) [31] to evaluate their cross-platform stability.

(iii) Development of individualized recurrence risk signatures.

To identify individualized recurrence risk signatures for each cell subtype, we aggregated C-GPs from three scRNA-seq datasets:

$$G_j = G_{D_1 j} \cup G_{D_2 j} \cup G_{D_3 j} \quad (3)$$

where D_x and j represent the scRNA-seq dataset x and cell subtype j, respectively. $G_{D_x j}$ represent all the C-GPs of cell subtype j in scRNA-seq dataset x. In addition, we collected six bulk transcriptome datasets ($n = 442$) and 80% of which were randomly sampled as the training dataset while the rest samples were used as the test dataset. The other three independent bulk RNA-seq datasets GSE14333 ($n = 99$), GSE17536 ($n = 111$) and GSE39582 ($n = 264$) were used as the validation datasets (Supplementary Table S2). We performed the Lasso-Cox model [32] to train a signature, which was furtherly used to score the recurrence-risk for each patient. We chose 'lambda.1se' as the final penalty for the model and *surv_cutpoint* was used to work out the optimal cutoff of the risk value. We calculated the risk score based on the REOs of C-GPs and corresponding coefficients for each patient. A patient was stratified into high-risk group if his risk score was higher than an optimal cutoff, otherwise low-risk group. After that, we performed *survdiff* function to have a log-rank test on these two groups. Finally, the multivariable Cox proportional-hazards regression model was used to evaluate whether signature performed as an independent prognostic factor after adjusting for other clinical factors including tumor stage, gender, age, primary tumor location and gene mutations (BRAF, KRAS and TP53) and mismatch repair status.

Determining cell type abundance from bulk transcriptome

Based on the high variable genes (HVG)-cell expression matrix for each scRNA-seq dataset, we inferred the abundance of each cell subtype in bulk transcriptome profile from the signature matrix established by CIBERSORTx [33]. Similarly, we applied Cox proportional hazard model to evaluate the contribution of the infiltration of each cell subtype to the recurrence survival time.

Prognostic signature enrichment analysis

Here, we hypothesized that prognostic-associated subpopulations tend to enrich more prognostic-associated genes. In each bulk RNA dataset, we applied the *coxph* function from the *survival* package on certain gene expression with recurrence time and status in clinical information. Then, we filtered out the genes with adjusted P value (BH adjusted) greater than 0.2. For those genes with hazard ratio (HR) greater than 1, we grouped them into unfavorable gene set, and those with hazard ratio smaller than 1 were denoted as favorable gene set. Single sample gene set enrichment analysis (ssGSEA) [34] was performed on average expression of different cell types from scRNAseq datasets to calculate enrichment scores with prognostic-associated gene sets of different cell types.

Identifying prognostic-associated subpopulations using Scissor

Scissor [35], a module to distinguish clinically phenotype relevant cells in scRNA-seq dataset, was used to calculate the pertinence relation between single cell in scRNA-seq profile and single sample in bulk RNA-seq profile. We used default parameters Cox mode of Scissor to detect prognostic-associated cells in each scRNA-seq profile.

Evaluation of performance for methods used for prioritizing prognostic-associated subpopulations

Besides the methods mentioned above, we used *FindMarkers* function from Seurat package to detect the genes differentially expressed among cell subtypes. We named this method as Uni-Markers. In addition, marker genes detected by SciBet, were named as SciBet-Markers. Similarly, we applied the *coxph* function to calculate the HR of Uni-Markers and SciBet-Markers in cell subtypes for each scRNA-seq dataset. To compare the performance of different methods (CIBERSORTx, ssGSEA, Scissor, Uni-Markers, SciBet-Markers and scRank^{XMBD}), we performed a literature review on the prognostic value of several cell subtypes (Supplementary Table S4). We used 'R' (risk) and 'P' (protected) to label the prognosis of each cell subtype. Under this circumstance, if the current result was consistent with the previous study, we announced it positive ('p' as a proxy) for the prognostic contribution. If not, it was labeled negative (or 'N'). Besides, 'A' (ambiguous) referred the prognostic value had conflict results in different scRNA-seq datasets or even cannot obtain the prognostic classification. 'U' (uncertain) represented the prognostic value of one cell subtype was unclear in prior research. Furthermore, we calculated three quantitative metrics to assess performance of these methods:

$$\text{Accuracy} = \frac{C_p}{C_p + C_N + C_A} \quad (4)$$

$$\text{False positive rate} = \frac{C_N}{C_p + C_N + C_A} \quad (5)$$

$$\text{Discordance} = \frac{C_A}{C_p + C_N + C_A} \quad (6)$$

where C_p represents the number of cell subtypes whose prognostic contribution was in line with previous studies and different scRNA-seq datasets, while C_N represents the number of cell subtypes with similar prognostic value among different scRNA-seq datasets but were inconsistent with previous studies. C_A and C_U represent the number of cell subtypes whose prognostic contribution were evaluated as 'A' and 'U', respectively.

RESULTS

Cell identity annotation across independent datasets

Firstly, we manually annotated the major cell types of a scRNA-seq dataset (GSE144735) (Figure 2). Based on the expression of classic markers (Figure 2B, Supplementary Figure S1B and C), four immune cell types (T cells, B cells, myeloid cells and mast cells) and three non-immune cell subtypes (endothelial cells, fibroblasts and epithelial cells) were distinguished. Then, we used SciBet [29] to find the discriminative feature genes (Supplementary Figure S1A) in the training dataset via E-test. After that, we annotated two independent scRNA-seq datasets (GSE132465 and GSE132257) separately (Supplementary Figure S1A and B). Seven major cell types similar to the training dataset were obtained and UMAP of classic markers' expression level for each cell type were also shown (Figure 2B, Supplementary Figure S1A and B). Next, we reapplied dimension reduction and clustered each major cell types into minor subtypes. For T cells, we identified 10 subpopulations: naïve T cells, CD8+ T cells, CD4+ T cells and others. Furthermore, CD8+ T cells were separated into Tex (exhausted T cells) and CTLs (cytotoxic T cells), which were named as GZMK+ CTL and KLRD1+ CTL. CD4+ T cells were separated into four

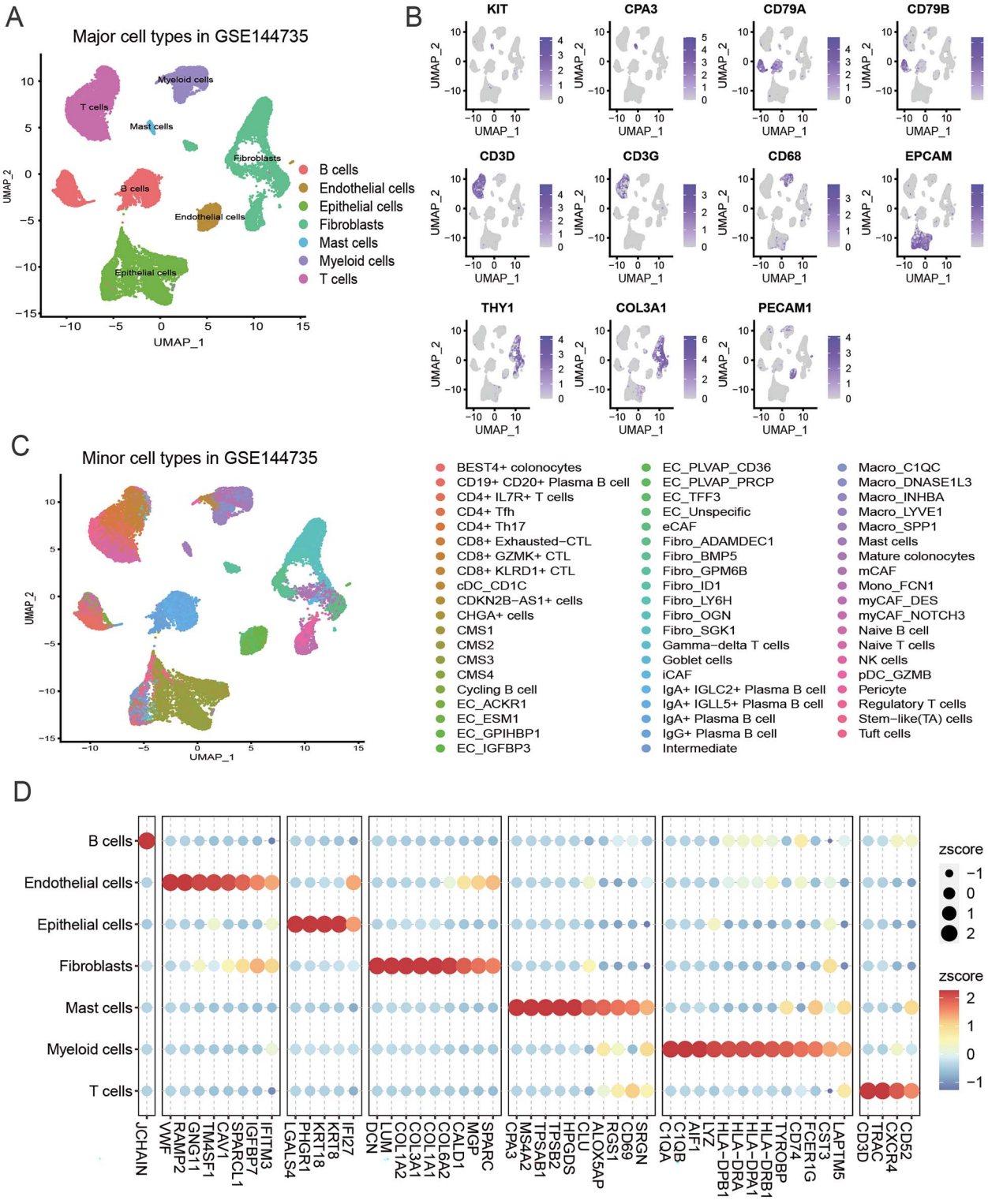

Figure 2. Cell type annotation using joint application of manual and automatic methods.

subtypes, including CD4+ Tfhs (follicular T-helper cells) and Treg (regulatory T cells). Other T cells were separated to $\gamma\delta$ T cells (gamma-delta T cells) and NK cells (natural killing cells) (Supplementary Figure S2A and B). We also identified seven sub-populations of B cells (Supplementary Figure S3A). In particular, there are two distinct IgA + plasma B cells clusters (Supplementary Figure S3B). One cluster highly expressed IGLC2 and the other highly expressed IGLL5. Note that the IGLC+ IgA+ plasma B cells cluster was also identified in CRC by Wang *et al.* [36]. For myeloid cells, we identified dendritic cells and monocyte–macrophages.

For non-immune cell types, we identified CAFs, eCAF (extracellular CAF) and myCAF (myofibroblast), which were named based on differential expressed markers [17, 37, 38] (Supplementary Figure S5). Using the markers from Sharma *et al.* [39], we identified eight subpopulations of endothelial cells (Supplementary Figure S6). We found that the distribution of epithelial cells was heterogeneous among different patients

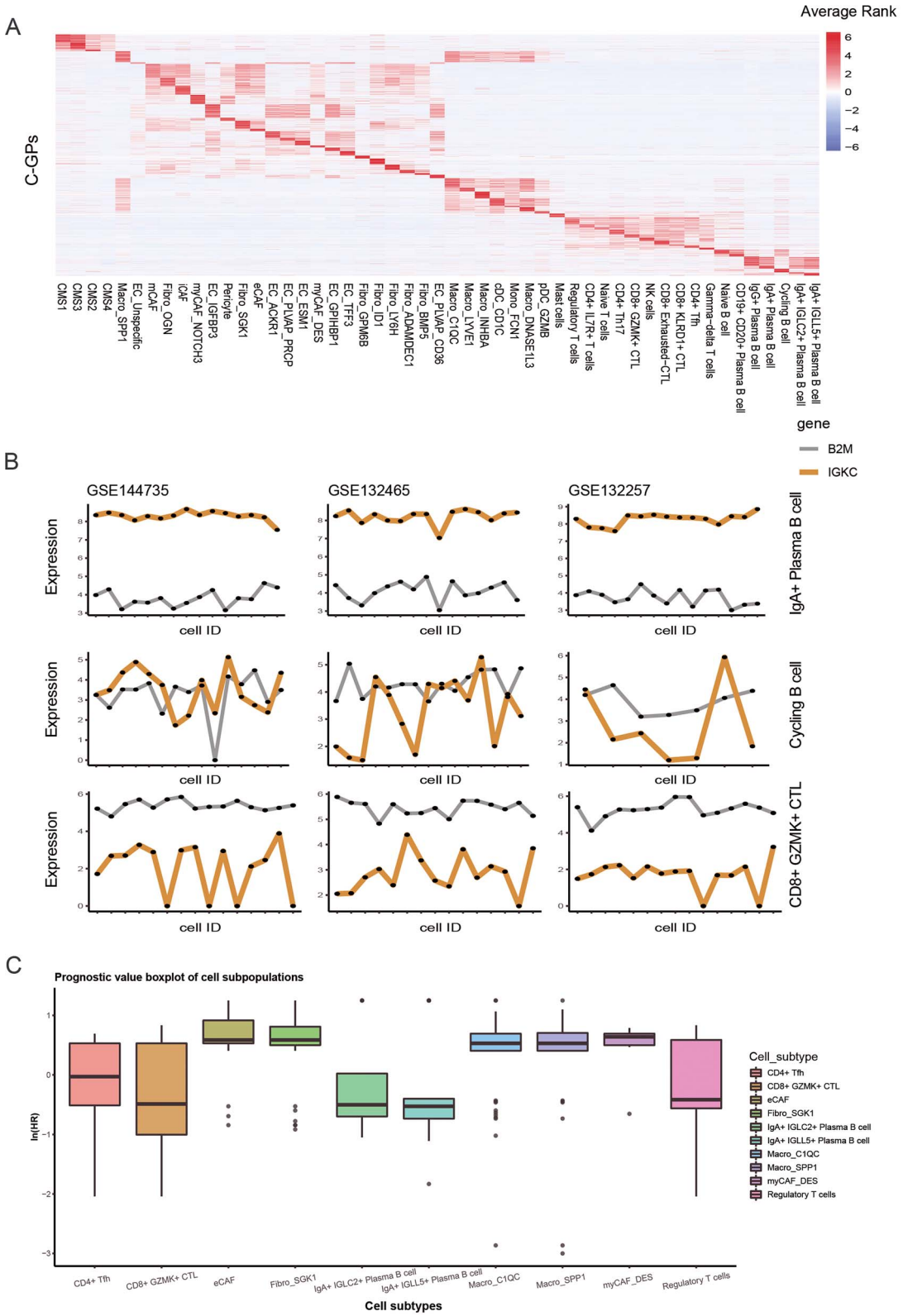

Figure 3. Cell subtype specific gene pairs in three scRNA-seq datasets.

(Supplementary Figure S7C); therefore, we separated epithelial cells from tissues. For epithelial cells from normal tissues, we applied classic markers [17] to identify subpopulations (Supplementary Figure S7A). We assigned CMS labels to epithelial cells from tumors and nearby tissues by calculating the enrichment score of the CMS-related signaling pathway gene sets. It turns out that CMS4-like cells enriched genes related to EMT and TGF-β signaling pathways. CMS2-like cells mainly enriched genes related to WNT and MYC signaling pathways. CMS3-like and CMS1-like cells enriched genes related to differentiation and MSI, respectively (Supplementary Figure S7B). These results were consistent with Lee *et al* [17] and in accordance with

Table 1. Comparison of scRank^{XMBD} with existing methods used for prioritizing prognostic-associated subpopulations based on single-cell transcriptomes

Methods	Input for prognostic analysis	Accuracy | False positive rate | Discordance					Mean
		T cells	Myeloids	B cells	Endothelial cells	Fibroblasts	
CIBERSORTx	Cell type abundance	0.14 | 0.29 | 0.57	0.88 | 0.13 | 0.00	0.17 | 0.50 | 0.33	0.33 | 0.67 | 0.00	0.45 | 0.36 | 0.18	0.39 | 0.39 | 0.22
Scissor	The similarity between single-cell data and bulk data	0.00 | 0.00 | 0.00	0.50 | 0.00 | 0.50	0.00 | 0.00 | 0.00	0.00 | 0.00 | 0.00	0.64 | 0.27 | 0.09	0.57 | 0.13 | 0.30
ssGSEA	Enrichment scores of prognostic associated gene sets	0.71 | 0.14 | 0.14	0.88 | 0.00 | 0.13	0.33 | 0.33 | 0.33	0.50 | 0.17 | 0.33	0.73 | 0.27 | 0.00	0.63 | 0.18 | 0.19
Uni-Markers	Genes differentially expressed among cell subtypes	0.86 | 0.14 | 0.00	0.38 | 0.63 | 0.00	0.33 | 0.67 | 0.00	0.17 | 0.83 | 0.00	0.82 | 0.00 | 0.18	0.51 | 0.45 | 0.04
SciBet-Markers	Genes for cell subtype annotation	0.71 | 0.14 | 0.14	0.50 | 0.38 | 0.13	0.17 | 0.50 | 0.33	0.17 | 0.67 | 0.17	0.64 | 0.09 | 0.27	0.44 | 0.36 | 0.20
scRank^{XMBD}	Cell-type-specific gene pairs	0.71 | 0.00 | 0.29	0.88 | 0.00 | 0.13	0.50 | 0.17 | 0.33	0.83 | 0.00 | 0.17	0.73 | 0.27 | 0.00	0.73 | 0.09 | 0.18

the conventional understanding of CMS molecular subtypes [4, 7, 40].

Cell subpopulations were also identified in GSE132465 and GSE132257 datasets using the same workflow (Supplementary Figure S8). The proportions of cell subpopulations were shown in Supplementary Figure S9. As for mast cells with low cell counts, no further subtypes were annotated.

Prioritizing prognostic-associated subpopulations based on C-GPs

Firstly, we identified C-GPs (Figure 1, Steps 2–3). SingleCellNet [20] was used to train RF classifiers for each cell subtype based on GPs in the three independent scRNA-seq datasets, respectively. Then, we plotted the precision–recall (PR) curve and calculated area under the precision-recall curve (AUPR) to evaluate the performance of the classifiers. In general, the classifiers performed well for most cell subtypes (Supplementary Figure S10), suggesting that cell subtypes could be annotated accurately based on REOs. We selected C-GPs for each cell type and observed that cell subpopulations derived from the same major cell type shared some specific GPs as expected (Figure 3A, Supplementary Figure S11A). In addition, these cell subpopulations still had their own specific GPs (Figure 3A, Supplementary Figure S11A). For example, we randomly selected 20 IgA+ plasma B cells in all scRNA-seq datasets and plotted the REOs of B2M-IGKC, respectively. It was found that this GP maintained cross-dataset stability (Figure 3B). Another example was the specific GP CCL5-GADD45B of CD8+ KLRD1+ T cells. Notably, we observed that the REOs of this GP were consistent (Supplementary Figure S11C).

Next, for each scRNA-seq dataset, we collected the C-GPs and correlated them with the recurrence survival time of the patients in bulk transcriptome datasets using a univariable Cox model (Supplementary Table S2). The prognostic values of several cell types evaluated here were in line with previous studies (Supplementary Table S5). T cells' subpopulations, such as CD4+ Tfh and CD8+ GZMK+ CTL, correlated to good prognosis (Figure 3C). CD4+ Tfh was shown to be an independent prognostic predictor in breast cancer and correlated to improved prognosis [41]. CD8+ GZMK+ CTL, a significant component of cell-mediated immunity, plays a central role in tumor cytotoxicity [42]. Besides, the results of other cell subpopulations were consistent with previous studies (Figure 3C). For example, in B cells' subpopulations, low-expressed IGLC2 was considered as a factor correlated to worse prognosis in triple negative breast cancer patients [43] and the expression level of IGLL5 was positively correlated to tumor size in clear cell renal cell carcinoma [44]. When it comes to the subpopulations of myeloid cells, it has been reported previously that Macro_SPP1 correlated to worse prognosis for CRC patients [18]. As for subpopulations of fibroblasts, eCAF that highly expressed CST1 was reported to correlate to worse prognosis and tumor generating in CRC [45]. myCAF was found to promote cancer development and progression [46].

scRank ^{XMBD} achieves higher precision and concordance compared with existing methods

The performance of scRank^{XMBD} was compared with five methods in common practice. Firstly, we used CIBERSORTx to evaluate the relevance between cell subtype infiltration and prognosis of CRC patients from bulk transcriptome, which was performed in most existing studies [47–49]. We found that the infiltration of a few cell subtypes, such as eCAF and EC_GPIHBP1, correlated with worse prognosis in CRC patients (Supplementary Figure S12). Fibro_ADAMDEC1 and EC_IGFBP3 were related to good prognosis (Supplementary Figure S12). Secondly, the results of ssGSEA revealed that prognosis-protected genes mainly enriched on the subpopulations of B cells and T cells (Supplementary Figure S13). However, the recurrence-related genes mainly enriched on fibroblasts, myeloid cells and endothelial cells, especially SPP1+ and C1QC+ macrophages (Supplementary Figure S14). Thirdly, Scissor was used to integrate the phenotype of patients in bulk transcriptome data and cells in scRNA-seq data. With Scissor, we observed that Macro_SPP1, Macro_INHBA, Fibro_OGN, Fibro_SGK1, CAFs, CMS1-like cells and CMS3-like epithelial cells were related to CRC recurrence. CMS2-like epithelial cells were related to good prognosis. Interestingly, results of subtypes of B cells and T cells from Scissor were not associated to clinical outcome, which may indicate potential limitations of this method (Supplementary Figure S15). Finally, applying Uni-Markers and SciBet-Markers, many cell subtypes including CD8+ GZMK+ CTL and cDC_CD1C were recognized as related to prolong patient's DFS time. However, it could not evaluate exactly for CAFs and some macrophages, which were considered as risk factors in CRC [8, 50–53].

To have a more systematic evaluation on these methods, we used three metrics (accuracy, false positive rate and discordance, Method section) to compare the prognostic value of these cell subtypes with consensus results from existing literature

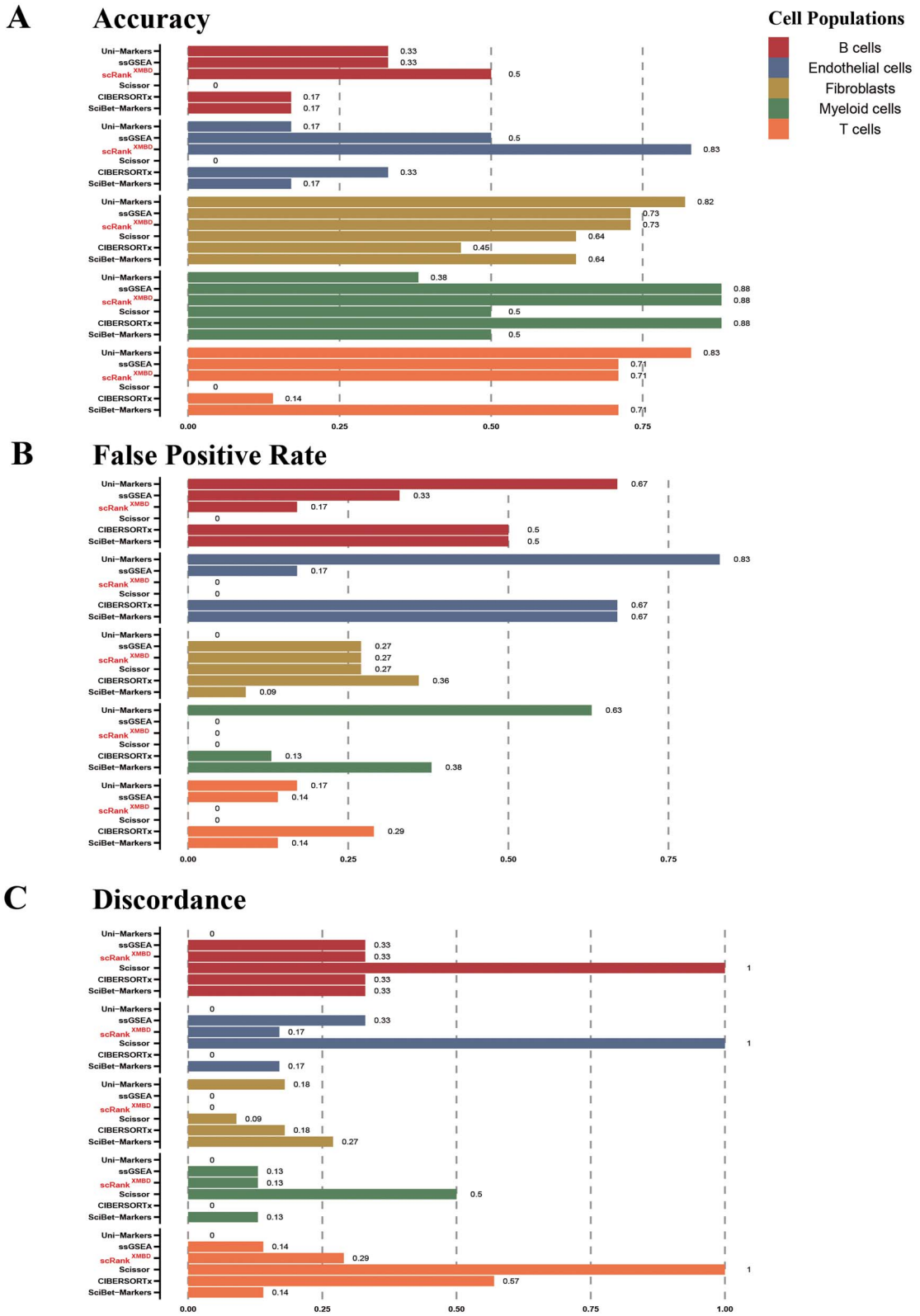

Figure 4. Comparison with the existing methods used for evaluation of prognostic value for cell types.

(Supplementary Table S5). The results showed that scRank^{XMBD} achieved higher precision and concordance. (Table 1, Figure 4). Moreover, scRank^{XMBD} ensured the stability cross different datasets and REO was a reliable signature for specific cell phenotypes in bulk RNA (Supplementary Figure S16).

Within-cell REOs of GPs predict recurrence risk in CRC

We used the C-GPs as features of each cell subpopulation and used Lasso-Cox model to train a signature to predict the recurrence risk (Figure 1, Step5). Results show that the trained

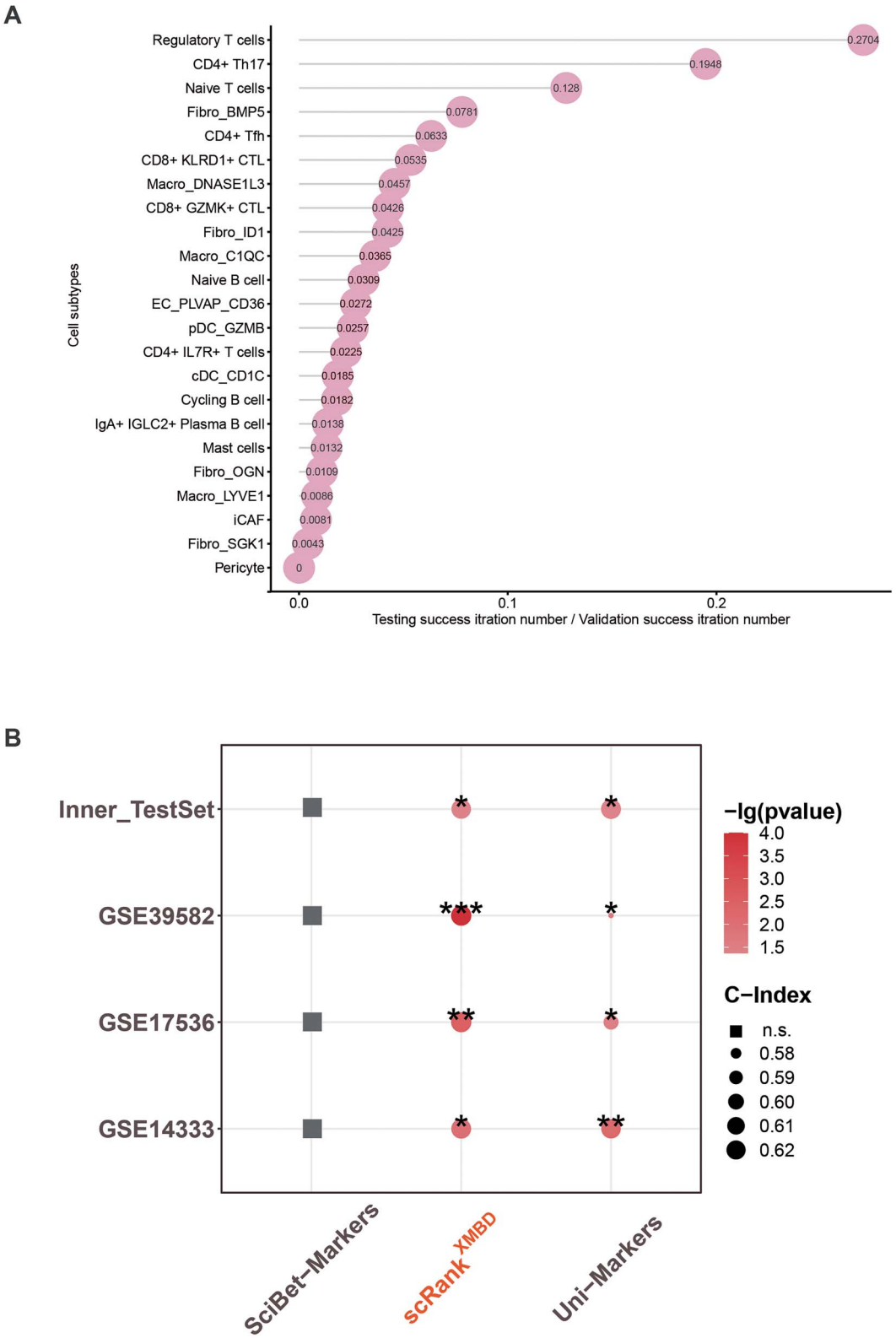

Figure 5. Prioritizing recurrence risk -associated subpopulations for CRC patients.

signature for 28 cell subtypes, including CD4+ Tfh and CD4+ IL7R+ T cell, was able to predict the recurrence risk for CRC patients with early stage (II/III) (Figure 5A). The infiltration of CD4+ Tfhs was considered as a potent prognosis predictor in breast cancer [41]. The expression level of IL7R was relevant to prolonged DFS and OS in lung adenocarcinoma [54]. Besides, for non-immune cell subpopulations, the low-expression level of BMP5 was considered as a predictor to worse prognosis in CRC [55].To further validate the signature derived by scRank^{XMBD}, we selected Macro_DNASE1L3 as an example. The signature of

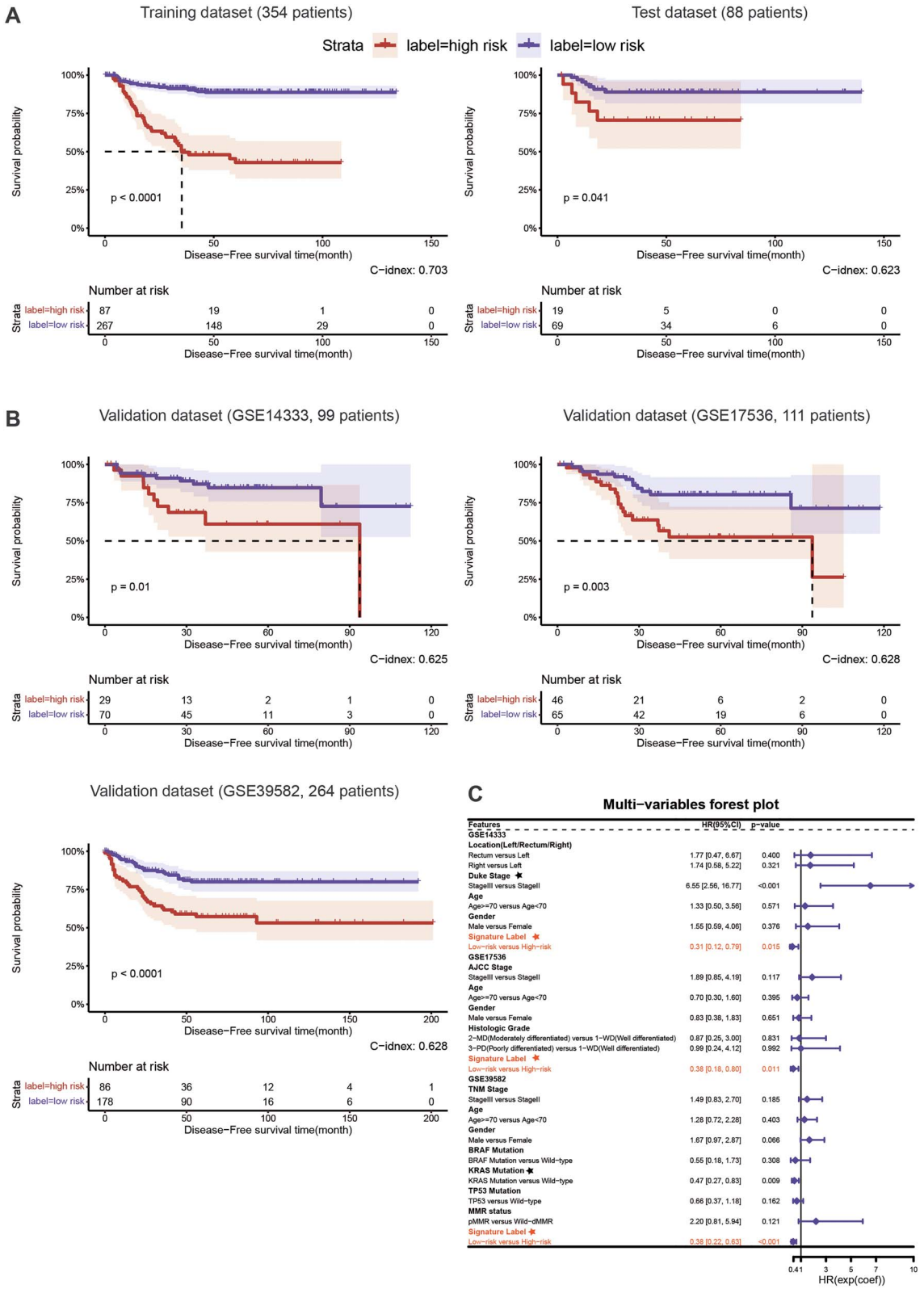

Figure 6. The performance of the individualized prognostic signature from Macro_DNASE1L3 for predicting the DFS of CRC patients.

Macro_DNASE1L3 consisted of 18 GPs (Supplementary Table S5) and it performed well in training set ($n = 354$, P value <0.001, log rank test; C-index = 0.703) and inner test set ($n = 88$, P value = 0.041, log rank test). Moreover, it also could separate patients into high-risk and low-risk groups exactly in three independent validation sets GSE14333 ($n = 99$, P value <0.01, log rank test), GSE17536 ($n = 111$, P value = 0.003, log rank test) and GSE39582 ($n = 264$, P value <0.001, log rank test) (Figure 6A and B). The key parameters of the Lasso-Cox model were shown (Figure S17). Furthermore, multivariable Cox analysis on our signature and other clinical factors including age, gender, clinical stage and genomic biomarkers showed that our signature was still an

independent recurrence-risk predictor (Figure 6C). Features and their weights were shown (Supplementary Table S5). Similar results were observed in CD8+ GZMK+ CTLs and IgA+ IGLC2+ B cells (Supplementary Figures S18 and S19). The lower level of DNASE1L3 correlated with poorer prognosis in various cancers including breast invasive carcinoma, hepatocellular carcinoma, kidney cancer, stomach cancer, lung adenocarcinoma and sarcoma [56, 57]. CD8+ GZMK+ CTLs were related to the prolonged DFS [58]. IgA+ IGLC2+ plasma B cells were related to worse prognosis in CRC [36]. In general, CD8 CTLs have been extensively studied in the literature, while IgA+ IGLC2+ B cell and Macro_DNASE1L3 are novel cell subtypes, which could potently predict the recurrence risk in CRC.

In addition, compared with SciBet-Markers and Uni-Markers, scRank^{XMBD} showed better predictive results in the recurrence risk classification task (Figure 5B, Supplementary Figure S20).

DISCUSSION

While TME has been reported to be important for tumor treatment response and survival of patients [13, 59], bulk tumor-based transcriptome only provides averaged data and could not accurately characterize the gene expression of subpopulations of the TME. The use of scRNA-seq in cancer research has improved our understanding of the TME [60, 61]. Identifying cell subpopulations that associate with clinical outcome could facilitate the discovery of cell type targeted therapies as well as prognostic biomarkers. Most scRNA-seq datasets include fewer than 20 samples, which could not be used to identify the cell subpopulations associated with survival time for the lack of statistical power. Therefore, it is necessary to take full use of valuable clinical information to prioritize cell subpopulations from single-cell transcriptomes. Here, we developed scRank^{XMBD}, a novel computational method to identify prognostic-associated cell subpopulations based on within cell REOs. Most of the previous studies prioritized prognostic-associated subpopulations based on the proportion of cell types in bulk datasets predicted by computational methods such as CIBERSORT [62] and MCP-counter [63] . These methods usually analyzed about 10–20 cell subtypes due to limited cell-type-specific genes. In contrast, scRNA-seq enables unbiased transcriptional profiling of thousands of individual cells from a single-cell suspension, which allows for more accurate identification of prognostic-associated cell subtypes. Moreover, compared with exiting methods based on gene expression levels or cellular abundance, within cell REOs could be more robust to transfer knowledge from single-cell to bulk transcriptome. Collectively, these features enable scRank^{XMBD} to achieve higher precision and concordance in identifying prognostic-associated cell subtypes from bulk RNA datasets. To extend the application of scRank^{XMBD}, we will perform analysis to identify clinically relevant cell subpopulations by associating with other clinical phenotypes such as chemotherapy and immunotherapy response in the future.

Recent advances in high-throughput technologies facilitate the application of molecular biomarkers for prognosis prediction of CRC. However, most of the reported bulk transcriptional biomarkers were based on the expression levels of the signature genes [64–66]. Due to experimental batch effects [67], risk classification methods depend on data normalization, which could not be diagnosed at the individualized level [24]. In contrast, the REOs of genes within a sample are robust against experimental batch effects and normalization methods [21, 22], which renders them promising for building robust diagnostic and prognostic models in bulk transcriptome data across different platforms including RNA-Seq and mircoarrays [68–71]. Moreover, a sample could be individually classified without data normalization based on REOs, which is more in line with current clinical practice. In this study, we developed single-cell GP signatures to predict recurrence risk for CRC patients individually. Our evaluation reveals that compared with existing methods, the prognostic cell sub-subpopulations identified by scRank^{XMBD} were highly concordant with more published results in CRC (Supplementary Table S5). For example, CD4+ Tfh, CD8+ GZMK+ CTL and IgA+ IGLC2+ plasma B cells were related to good prognosis of CRC. Macro_SPP1 and Macro_DNASE1L3 were considered as risk factors in CRC. Their roles in CRC certainly deserve further investigation.

Several limitations should be noted. The REOs relied on accurate annotations of cell types and could be affected by the dropout events in scRNA-seq. Interestingly, in our work, we demonstrated the stability of C-GPs across different technical platforms and cell capture strategies, e.g. on microarray data from purified major cell types. Further study is warranted to validate this result on other sequencing platforms, e.g. long reads sequencing platforms and other minor cell types when such data are available.

Key Points

- We developed scRank^{XMBD} (https://github.com/xmuyulab/scRank-XMBD), a novel method to prioritize prognostic-associated subpopulations based on within-cell REOs of gene pairs.
- scRank^{XMBD} achieves higher precision and concordance compared with existing methods.
- Single-cell gene pair signatures were developed to predict recurrence risk for CRC patients individually.
- Our work facilitates the application of the rank-based method in scRNA-seq data for prognostic biomarker discovery and precision oncology.

SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/bib.

AUTHORS' CONTRIBUTIONS

R.Y, J.H. and M.T. supervised the study. M.T., Y.L. and J.S. developed the framework and performed the analysis. M.T., Y.L., W.Y, J.S., Z.Z, J.X, J.T., S.L and C.L wrote and revised the manuscript.

FUNDING

National Natural Science Foundation of China (grant nos 82002529, 31871317, 32070635), Fundamental Research Funds for the Central Universities (grant nos 20720210095) and Natural Science Foundation of Fujian Province (grant nos 2020 J05012, 2020 J01028).

DATA AVAILABILITY STATEMENT

All scRNA-seq and bulk transcriptomes analyzed in this work were available in GEO database and summarized in Supplementary Table S2.

References

1. Siegel RL, Miller KD, Fuchs HE, *et al.* Cancer statistics. *CA Cancer J Clin* 2022;**72**:7–33.
2. Siegel RL, Miller KD, Goding Sauer A, *et al.* Colorectal cancer statistics. *CA Cancer J Clin* 2020;**70**:145–64.
3. Sveen A, Kopetz S, Lothe RA. Biomarker-guided therapy for colorectal cancer: strength in complexity. *Nat Rev Clin Oncol* 2020;**17**:11–32.
4. Guinney J, Dienstmann R, Wang X, *et al.* The consensus molecular subtypes of colorectal cancer. *Nat Med* 2015;**21**:1350–6.
5. Gelsomino F, Barbolini M, Spallanzani A, *et al.* The evolving role of microsatellite instability in colorectal cancer: a review. *Cancer Treat Rev* 2016;**51**:19–26.
6. Roseweir AK, McMillan DC, Horgan PG, *et al.* Colorectal cancer subtypes: translation to routine clinical pathology. *Cancer Treat Rev* 2017;**57**:1–7.
7. Wang W, Kandimalla R, Huang H, *et al.* Molecular subtyping of colorectal cancer: recent progress, new challenges and emerging opportunities. *Semin Cancer Biol* 2019;**55**:37–52.
8. Calon A, Lonardo E, Berenguer-Llergo A, *et al.* Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat Genet* 2015;**47**:320–9.
9. Dienstmann R, Villacampa G, Sveen A, *et al.* Relative contribution of clinicopathological variables, genomic markers, transcriptomic subtyping and microenvironment features for outcome prediction in stage II/III colorectal cancer. *Ann Oncol* 2019;**30**:1622–9.
10. Pagès F, Mlecnik B, Marliot F, *et al.* International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. *Lancet* 2018;**391**:2128–39.
11. Wang Y, Lin HC, Huang MY, *et al.* The Immunoscore system predicts prognosis after liver metastasectomy in colorectal cancer liver metastases. *Cancer Immunol Immunother* 2018;**67**:435–44.
12. Maman S, Witz IP. A history of exploring cancer in context. *Nat Rev Cancer* 2018;**18**:359–76.
13. Bruni D, Angell HK, Galon J. The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy. *Nat Rev Cancer* 2020;**20**:662–80.
14. Lim B, Lin Y, Navin N. Advancing cancer research and medicine with single-cell genomics. *Cancer Cell* 2020;**37**:456–70.
15. Lei Y, Tang R, Xu J, *et al.* Applications of single-cell sequencing in cancer research: progress and perspectives. *J Hematol Oncol* 2021;**14**:91.
16. Li H, Courtois ET, Sengupta D, *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet* 2017;**49**:708–18.
17. Lee HO, Hong Y, Etlioglu HE, *et al.* Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat Genet* 2020;**52**:594–603.
18. Zhang L, Li Z, Skrzypczynska KM, *et al.* Single-cell analyses inform mechanisms of myeloid-targeted therapies in colon cancer. *Cell* 2020;**181**:442–459 e429.
19. Mei Y, Xiao W, Hu H, *et al.* Single-cell analyses reveal suppressive tumor microenvironment of human colorectal cancer. *Clin Transl Med* 2021;**11**:e422.
20. Tan Y, Cahan P. SingleCellNet: a computational tool to classify single cell RNA-Seq data across platforms and across species. *Cell Syst* 2019;**9**:207–213 e202.
21. Geman D, d'Avignon C, Naiman DQ, *et al.* Classifying gene expression profiles from pairwise mRNA comparisons. *Stat Appl Genet Mol Biol* 2004;**3**:1–19.
22. Patil P, Bachant-Winner PO, Haibe-Kains B, *et al.* Test set bias affects reproducibility of gene signatures. *Bioinformatics* 2015;**31**:2318–23.
23. Wang H, Sun Q, Zhao W, *et al.* Individual-level analysis of differential expression of genes and pathways for personalized medicine. *Bioinformatics* 2015;**31**:62–8.
24. Qi L, Chen L, Li Y, *et al.* Critical limitations of prognostic signatures based on risk scores summarized from gene expression levels: a case study for resected stage I non-small-cell lung cancer. *Brief Bioinform* 2016;**17**:233–42.
25. Tong M, Zheng W, Li H, *et al.* Multi-omics landscapes of colorectal cancer subtypes discriminated by an individualized prognostic signature for 5-fluorouracil-based chemotherapy. *Oncogenesis* 2016;**5**:e242.
26. Wang R, Zheng X, Wang J, *et al.* Improving bulk RNA-seq classification by transferring gene signature from single cells in acute myeloid leukemia. *Brief Bioinform* 2022;**23**:23.
27. Zappia L, Oshlack A. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *Gigascience* 2018;**7**:7.
28. Eide PW, Bruun J, Lothe RA, *et al.* CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. *Sci Rep* 2017;**7**:16618.
29. Li C, Liu B, Kang B, *et al.* SciBet as a portable and fast single cell type identifier. *Nat Commun* 2020;**11**:1818.
30. Gautier L, Cope L, Bolstad BM, *et al.* Affy—analysis of Affymetrix GeneChipdata at the probe level. *Bioinformatics* 2004;**20**:307–15.
31. Berdiel-Acer M, Sanz-Pamplona R, Calon A, *et al.* Differences between CAFs and their paired NCF from adjacent colonic mucosa reveal functional heterogeneity of CAFs, providing prognostic information. *Mol Oncol* 2014;**8**:1290–305.
32. Tibshirani R. The Lasso method for variable selection in the Cox model. *Stat Med* 1997;**16**:385–95.
33. Newman AM, Steen CB, Liu CL, *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* 2019;**37**:773–82.
34. Subramanian A, Tamayo P, Mootha VK, *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;**102**:15545–50.
35. Sun D, Guan X, Moran AE, *et al.* Identifying phenotype-associated subpopulations by integrating bulk and single-cell sequencing data. *Nat Biotechnol* 2021;**40**:527–38.
36. Wang W, Zhong Y, Zhuang Z, *et al.* Multiregion single-cell sequencing reveals the transcriptional landscape of the immune microenvironment of colorectal cancer. *Clin Transl Med* 2021;**11**:e253.
37. Han C, Liu T, Yin R. Biomarkers for cancer-associated fibroblasts. *Biomark Res* 2020;**8**:64.
38. Sahai E, Astsaturov I, Cukierman E, *et al.* A framework for advancing our understanding of cancer-associated fibroblasts. *Nat Rev Cancer* 2020;**20**:174–86.
39. Sharma A, Seow JJW, Dutertre CA, *et al.* Onco-fetal reprogramming of endothelial cells drives immunosuppressive macrophages in hepatocellular carcinoma. *Cell* 2020;**183**:377–394 e321.
40. Muller MF, Ibrahim AE, Arends MJ. Molecular pathological classification of colorectal cancer. *Virchows Arch* 2016;**469**:125–34.

41. Gu-Trantien C, Loi S, Garaud S, *et al.* CD4(+) follicular helper T cell infiltration predicts breast cancer survival. *J Clin Invest* 2013;**123**:2873–92.
42. Masuda K, Kornberg A, Miller J, *et al.* Multiplexed single-cell analysis reveals prognostic and nonprognostic T cell types in human colorectal cancer. *JCI Insight* 2022;**7**(7):e154646.
43. Chang YT, Tsai WC, Lin WZ, *et al.* A novel IGLC2 gene linked with prognosis of triple-negative breast cancer. *Front Oncol* 2021;**11**:759952.
44. Xia ZN, Wang XY, Cai LC, *et al.* IGLL5 is correlated with tumor-infiltrating immune cells in clear cell renal cell carcinoma. *FEBS Open Bio* 2021;**11**:898–910.
45. Li T, Xiong Q, Zou Z, *et al.* Prognostic significance of cystatin SN associated nomograms in patients with colorectal cancer. *Oncotarget* 2017;**8**:115153–63.
46. Dinh HQ, Pan F, Wang G, *et al.* Integrated single-cell transcriptome analysis reveals heterogeneity of esophageal squamous cell carcinoma microenvironment. *Nat Commun* 2021;**12**:7335.
47. Su H, Cai T, Zhang S, *et al.* Identification of hub genes associated with neutrophils infiltration in colorectal cancer. *J Cell Mol Med* 2021;**25**:3371–80.
48. Ye L, Zhang T, Kang Z, *et al.* Tumor-infiltrating immune cells act as a marker for prognosis in colorectal cancer. *Front Immunol* 2019;**10**:2368.
49. Zhu X, Tian X, Ji L, *et al.* A tumor microenvironment-specific gene expression signature predicts chemotherapy resistance in colorectal cancer patients. *NPJ Precis Oncol* 2021;**5**:7.
50. Isella C, Terrasi A, Bellomo SE, *et al.* Stromal contribution to the colorectal cancer transcriptome. *Nat Genet* 2015;**47**:312–9.
51. Zhang R, Qi F, Zhao F, *et al.* Cancer-associated fibroblasts enhance tumor-associated macrophages enrichment and suppress NK cells function in colorectal cancer. *Cell Death Dis* 2019;**10**:273.
52. Komohara Y, Takeya M. CAFs and TAMs: maestros of the tumour microenvironment. *J Pathol* 2017;**241**:313–5.
53. Kalluri R. The biology and function of fibroblasts in cancer. *Nat Rev Cancer* 2016;**16**:582–98.
54. Wang X, Chang S, Wang T, *et al.* IL7R is correlated with immune cell infiltration in the tumor microenvironment of lung adenocarcinoma. *Front Pharmacol* 2022;**13**:857289.
55. Chen E, Yang F, He H, *et al.* Alteration of tumor suppressor BMP5 in sporadic colorectal cancer: a genomic and transcriptomic profiling based study. *Mol Cancer* 2018;**17**:176.
56. Deng Z, Xiao M, Du D, *et al.* DNASE1L3 as a prognostic biomarker associated with immune cell infiltration in cancer. *Onco Targets Ther* 2021;**14**:2003–17.
57. Li B, Ge YZ, Yan WW, *et al.* DNASE1L3 inhibits proliferation, invasion and metastasis of hepatocellular carcinoma by interacting with beta-catenin to promote its ubiquitin degradation pathway. *Cell Prolif* 2022;**55**:e13273.
58. Fakih M, Ouyang C, Wang C, *et al.* Immune overdrive signature in colorectal tumor subset predicts poor clinical outcome. *J Clin Invest* 2019;**129**:4464–76.
59. Becht E, de Reynies A, Giraldo NA, *et al.* Immune and stromal classification of colorectal cancer is associated with molecular subtypes and relevant for precision immunotherapy. *Clin Cancer Res* 2016;**22**:4057–66.
60. Baslan T, Hicks J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing. *Nat Rev Cancer* 2017;**17**:557–69.
61. Zhang Y, Wang D, Peng M, *et al.* Single-cell RNA sequencing in cancer research. *J Exp Clin Cancer Res* 2021;**40**:81.
62. Newman AM, Liu CL, Green MR, *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 2015;**12**:453–7.
63. Becht E, Giraldo NA, Lacroix L, *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol* 2016;**17**:218.
64. Dienstmann R, Mason MJ, Sinicrope FA, *et al.* Prediction of overall survival in stage II and III colon cancer beyond TNM system: a retrospective, pooled biomarker study. *Ann Oncol* 2017;**28**:1023–31.
65. Dai W, Li Y, Mo S, *et al.* A robust gene signature for the prediction of early relapse in stage I-III colon cancer. *Mol Oncol* 2018;**12**:463–75.
66. Zhou R, Zeng D, Zhang J, *et al.* A robust panel based on tumour microenvironment genes for prognostic prediction and tailoring therapies in stage I-III colon cancer. *EBioMedicine* 2019;**42**:420–30.
67. Leek JT, Scharpf RB, Bravo HC, *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 2010;**11**:733–9.
68. Guan Q, Zeng Q, Yan H, *et al.* A qualitative transcriptional signature for the early diagnosis of colorectal cancer. *Cancer Sci* 2019;**110**:3225–34.
69. Wu J, Lin Z, Ji D, *et al.* Metabolism-related gene pairs to predict the clinical outcome and molecular characteristics of early hepatocellular carcinoma. *Cancers (Basel)* 2022;**14**:3957.
70. Li Y, Zhang H, Guo Y, *et al.* A qualitative transcriptional signature for predicting recurrence risk of stage I-III bladder cancer patients after surgical resection. *Front Oncol* 2019;**9**:629.
71. Guan Q, Yan H, Chen Y, *et al.* Quantitative or qualitative transcriptional diagnostic signatures? *A case study for colorectal cancer. BMC Genomics* 2018;**19**:99.