Supplemental Materials

Fakih M. et al. Immune overdrive signature in colorectal tumor subset predicts poor clinical outcome

Supplemental Tables

Supplemental Table S1. Survival and associations with clinicopathologic characteristics using Cox regression.

Supplemental Table S2. Microsatellite instability status across *CD8A*/*CD274* -stratified risk groups (TCGA Stage IV).

Supplemental Table S3. Patient characteristics of public data sets.

Supplemental Figures and Legends

Supplemental Figure S1. Log-rank statistics for the determination of the optimal cut-point for patient stratification based on *CD8A* gene expression.

Supplemental Figure S2. Log-rank statistics for the determination of the optimal cut-point for patient stratification based on *CD274* gene expression.

Supplemental Figure S3. Investigation of clinical outcome for the population having high *CD8A* and intensive *CD274* expression in TCGA melanoma data set.

Supplemental Figure S4. Relapse-free survival analysis of the CRC risk subpopulation using NCBI-GEO data set.

Supplemental Figure S5. The protein expression of PDCD1 (PD-1) and CD274 (PD-L1) in the tumor microenvironment (City of Hope cohort).

Supplemental Figure S6. Expression levels of genes encoding commonly used cell type-specific markers across the CRC overall survival risk groups in TCGA and NCBI-GEO GSE39582 stage II and III samples.

Supplemental Figure S7. The gene expression of *PDCD1* (PD-1*)* and *CD274* (PD-L1) in the tumor microenvironment (public data sets).

Supplemental Figure S8. The expression of CD274 (PD-L1) regulatory genes across the CRC risk groups.

Supplemental Figure S9. Correlation among the expression of pro-inflammatory and immune regulatory genes in human melanoma and CRC in TCGA data set, along with the expression pattern of immune genes across the CRC risk groups.

Supplemental Figure S10. Estimation of the immune cell infiltration across the *CD8A* and *CD274* (PD-L1) expression-stratified CRC risk groups using two tumor deconvolution methods TIMER and CIBERSORT with TCGA gene expression data.

Supplemental Figure S11. The expression of representative immune checkpoint genes across the CRC risk groups.

Supplemental Figure S12. Expression of TGFβ-encoding and C-ECM signature genes and the distribution of consensus molecular subtypes (CMS) across the CRC risk groups in NCBI-GEO GSE39582 data set.

Supplemental Figure S13. Multispectral fluorescent IHC staining of the protein products of DNA mismatch-repair genes.

Supplemental Figure S14. Inference of microsatellite instability using microarray gene expression profiles in a NCBI-GEO data meta-analysis for stage II and III samples.

Supplemental Table S1:
Survival and associations with clinicopathologic characteristics using Cox regression.

Clinicopathologic variable	**TCGA**[A] (N=391)			**NCBI-GEO GSE39582**[A] (N=461)			**NCBI-GEO Meta-analysis**[A,B] (N=828)		
	Overall Survival			*Overall Survival*			*Relapse-free Survival*		
	HR	95%CI	p-value	HR	95%CI	p-value	HR	95%CI	p-value
Age									
Continuous	1.04	(1.02~1.06)	< 0.001	1.04	(1.02~1.05)	< 0.001	1.00	(0.99~1.01)	0.932
≥65 vs. <65	2.35	(1.33~4.17)	0.0033	1.81	(1.24~2.64)	0.0021	0.85	(0.65~1.10)	0.222
Stage									
III vs. II	1.96	(1.22~3.15)	0.0056	1.21	(0.87~1.69)	0.261	2.08	(1.59~2.71)	< 0.001
Gender									
Male vs. Female	1.00	(0.62~1.59)	0.989	1.30	(0.92~1.82)	0.136	1.25	(0.96~1.63)	0.091
CD8A expression									
Continuous	1.00	(0.87~1.16)	0.985	1.03	(0.86~1.23)	0.756	0.97	(0.83~1.14)	0.713
Categorical (High vs. Low)[C]	0.82	(0.51~1.32)	0.417	0.87	(0.63~1.22)	0.424	0.90	(0.70~1.17)	0.436
CD274 (PD-L1) expression									
Continuous	1.00	(0.85~1.17)	0.995	1.59	(0.99~2.56)	0.054	1.23	(0.89~1.70)	0.217
Categorical (High* vs. Low*)[D]	1.83	(1.04~3.20)	0.035	1.98	(1.19~3.29)	0.0085	1.53	(1.07~2.19)	0.021
Risk Group[E]									
Group IV* vs. Group III*	2.83	(1.40~5.74)	0.0038	2.37	(1.33~4.20)	0.0033	1.67	(1.09~2.57)	0.019
Group I+II vs. Group III*	1.82	(1.00~3.31)	0.048	1.31	(0.92~1.87)	0.134	1.24	(0.93~1.64)	0.141
Microsatellite instability									
MSS vs. MSI	0.82	(0.46~1.47)	0.507	0.96	(0.62~1.49)	0.845	1.03	(0.73~1.45)	0.876

[A] Analysis based on stage II and III patients.

[B] Including GSE39582, GSE14333, GSE17538, GSE31595.

[C] Dichotomized using median value of *CD8A* expression.

[D] Dichotomized using optimal *CD274* cut-point based on Log-rank test statistics (denoted with an asterisk for the cut-point determined in *CD8A* high group).

[E] Group I+II: *CD8A* low expression; Group III*: *CD8A* high / *CD274* low* expression; Group IV*: *CD8A* high / *CD274* high* expression.

Supplemental Table S2:
Microsatellite instability status across *CD8A*/*CD274*-stratified risk groups (TCGA Stage IV).

Cohort	Risk Group	MSI	MSS	Total	% MSI
TCGA[A,B] (Stage IV; N=85)	**Group I+II** (*CD8A* low)	1	55	56	1.8%
	Group III* (*CD8A* high / *CD274* low*)	0	20	20	0.0%
	Group IV* (*CD8A* high / *CD274* high*)	2	7	9 (10.6% of total)	22.2%

[A] Based on overall survival analysis.
[B] Data based on primary tumors.

Supplemental Table S3: Patient characteristics of public data sets.

TCGA

Characteristic	Number (%) of Patients[A]					
	Stage II Disease (n=217)		Stage III Disease (n=174)		All Patients[B] (n=599)	
Age at diagnosis (year)						
Median age (range)	68	(31~90)	64	(31~90)	66	(31~90)
Cancer type						
Colon adenocarcinoma	169	77.9%	126	72.4%	440	73.5%
Rectum adenocarcinoma	48	22.1%	48	27.6%	159	26.5%
Gender						
Male	119	54.8%	85	48.9%	323	53.9%
Female	98	45.2%	89	51.1%	276	46.1%
Pathologic T						
Tis	0	0	0	0	1	0.2%
T1	0	0	2	1.2%	20	3.4%
T2	1	0.5%	13	7.5%	105	17.7%
T3	201	93.9%	137	79.2%	404	68.2%
T4	12	5.6%	21	12.1%	62	10.5%
Pathologic N						
N0	214	100.0%	0	0.0%	335	56.6%
N1	0	0	108	62.4%	146	24.7%
N2	0	0	64	37.0%	108	18.2%
NX	0	0	1	0.6%	3	0.5%
Pathologic M[C]						
M0	198	100.0%	142	100.0%	440	84.3%
M1	0	0	0	0	70	13.4%
M1a	0	0	0	0	10	1.9%
M1b	0	0	0	0	2	0.4%
Microsatellite instability[D]						
MSI (MSI-H)	47	22.1%	16	9.2%	83	13.9%
MSS (MSI-L and MSS)	166	77.9%	158	90.8%	512	86.1%

NCBI-GEO GSE39582

Characteristic	Number (%) of Patients[A]					
	Stage II Disease (n=260)		Stage III Disease (n=201)		All Patients[B] (n=557)	
Age at diagnosis (year)						
Median age (range)	68	(24~94)	67	(22~97)	67	(22~97)
Gender						
Male	153	58.8%	105	52.2%	443	53.5%
Female	107	41.2%	96	47.8%	385	46.5%
Microsatellite instability[E]						
MSI	54	20.8%	30	14.9%	ND	ND
MSS	206	79.2%	171	85.1%	ND	ND

NCBI-GEO Meta-analysis (Stage II and III)

Characteristic	Number (%) of Patients[A]					
	Stage II Disease (n=444)		Stage III Disease (n=384)		All Patients (n=828)	
Age at diagnosis (year)						
Median age (range)	68	(24~94)	66	(22~97)	67	(22~97)
Gender						
Male	245	55.2%	198	51.6%	443	53.5%
Female	199	44.8%	186	48.4%	385	46.5%
Microsatellite instability[E]						
MSI	95	21.4%	66	17.2%	161	19.4%
MSS	349	78.6%	318	82.8%	667	80.6%

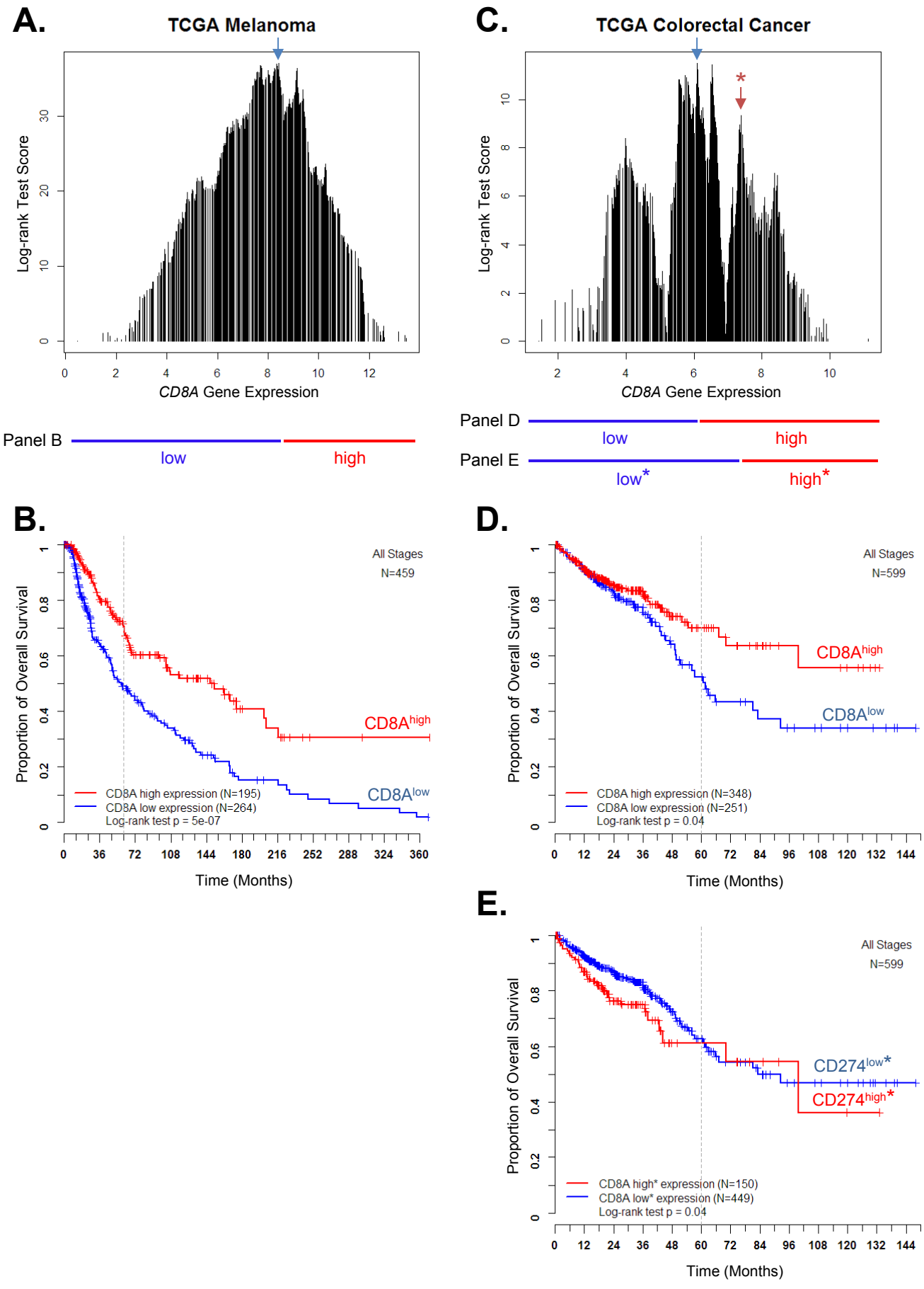
[A] All percentages were based on weighted analysis.

[B] Including Stage I, Stage IV and stage-unknown patients.

[C] 77 patients have either MX status or no information.

[D] 4 patients have no microsatellite instability information.

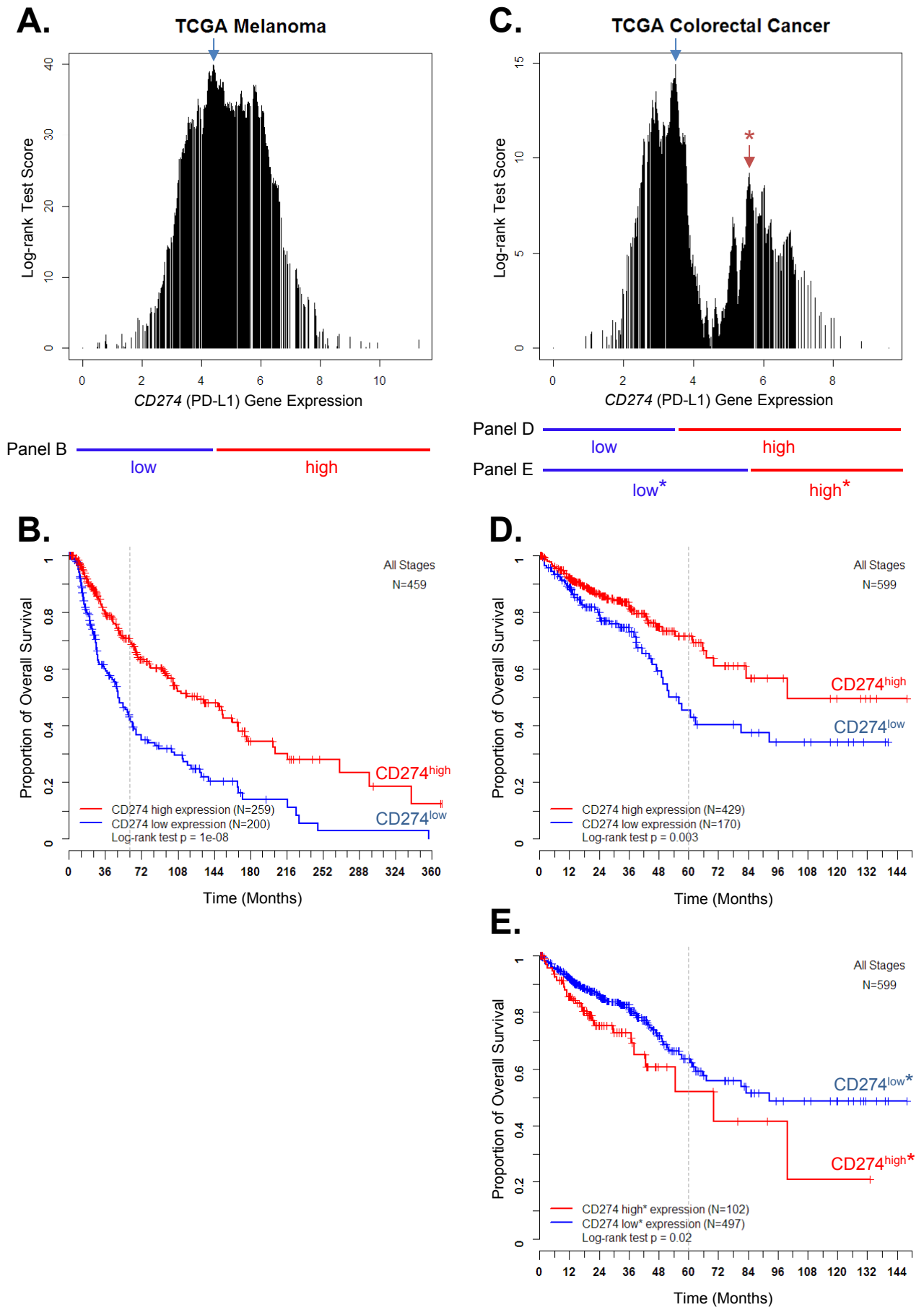
[E] Inferred by clustering analysis.



Supplemental Figure S1

Supplemental Figure S1. Log-rank statistics for the determination of the optimal cut-point for patient stratification based on *CD8A* gene expression.

Log-rank statistics were applied to identify the optimal cut-point for transforming the continuous variable of *CD8A* gene expression into categorical high and low expression groups in a survfit model. The test score at each candidate cut-point across the log-transformed gene expression values was plotted in panel (A) and (C) for TCGA melanoma and CRC data sets, respectively. For melanoma data set, the *CD8A* expression value showing the highest test score (indicated with a blue arrow) was applied for dichotomizing the patients. The Kaplan-Meier survival curves for the two patient groups were plotted using blue (low expression) and red (high expression) colors in panel (B). Grey dashed line was employed for visualizing the survival rates at 5-year mark. Log-rank test *p*-value was shown in the bottom left legend. For CRC data set, a more complicated score distribution was observed. Multiple cut-points along the distribution were tested for dichotomizing the patients for survival analysis. As shown in panel (D) and (E), Kaplan-Meier survival curves for low and high *CD8A* expression groups dichotomized by cut-points indicated with blue and red arrows showed a reverse pattern of survival trends, suggesting the existence of a novel risk group which is absent in melanoma.

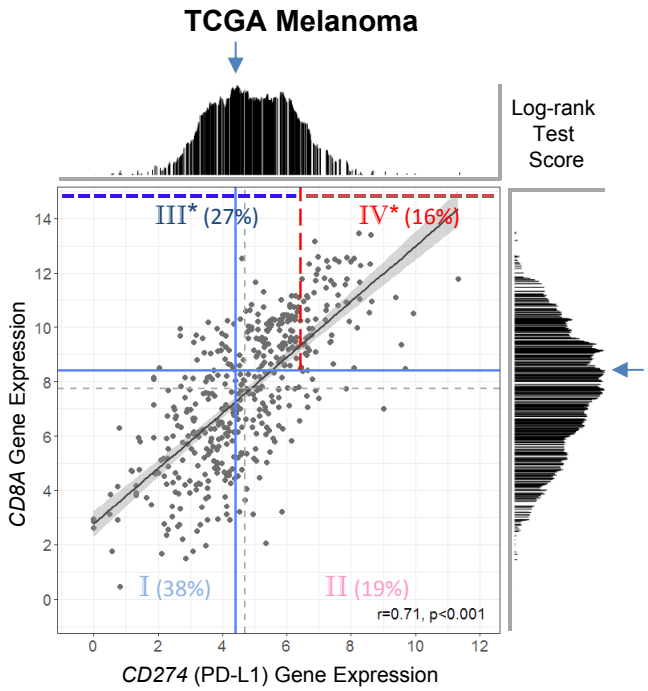


Supplemental Figure S2

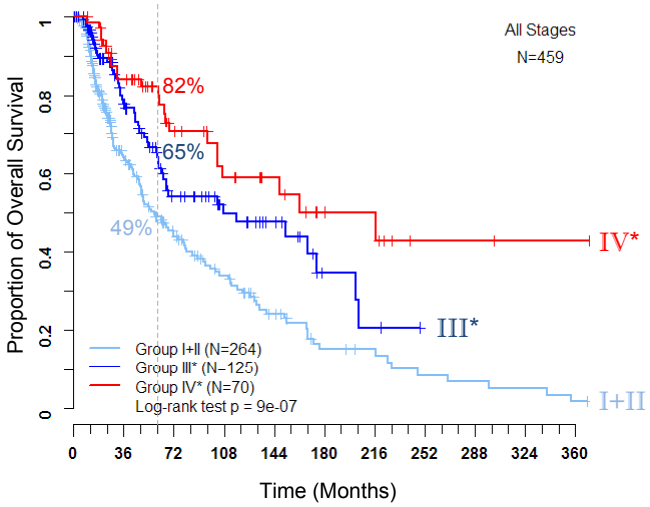
Supplemental Figure S2. Log-rank statistics for the determination of the optimal cut-point for patient stratification based on *CD274* gene expression.

Following the convention in Supplemental Figure S1 except for *CD274*, instead of *CD8A*, gene expression.

A.



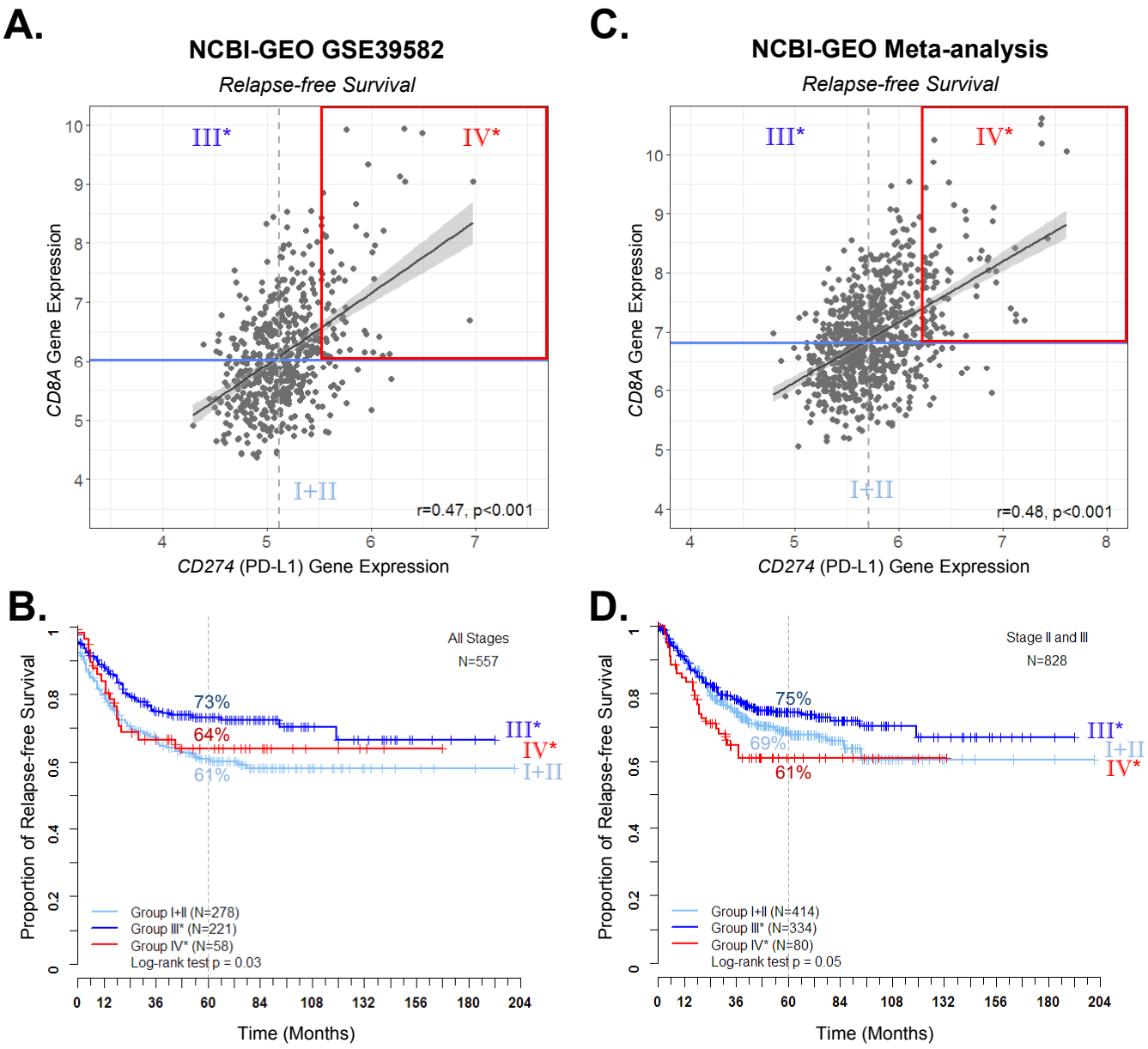
B.



Supplemental Figure S3

Supplemental Figure S3. Investigation of clinical outcome for the population having high *CD8A* and intensive *CD274* expression in TCGA melanoma data set.

(A) Adopted from Figure 1A, a virtual risk Group IV* applying the same CRC cut-point (based on *CD274* percentile) was isolated from the rest of *CD8A* high expression population in melanoma. (B) Kaplan-Meier survival curves are compared for the risk groups.

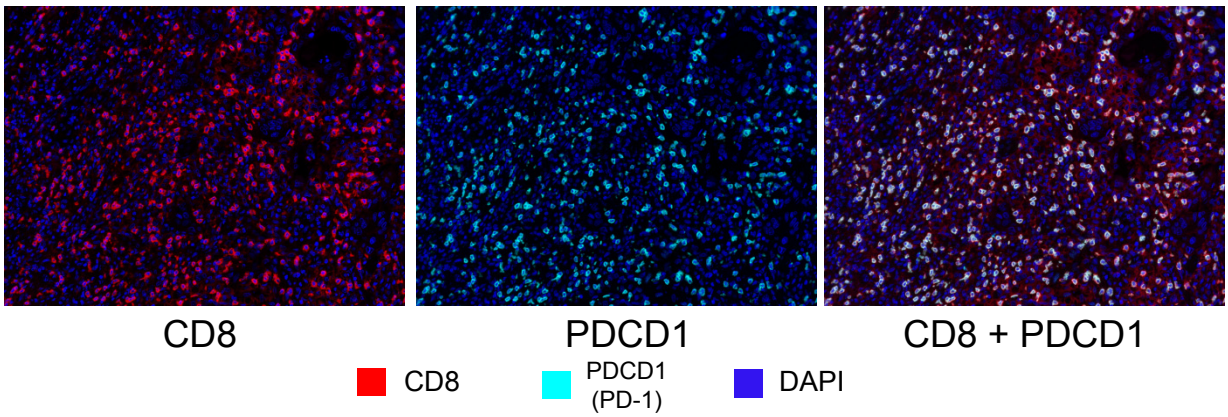


Supplemental Figure S4

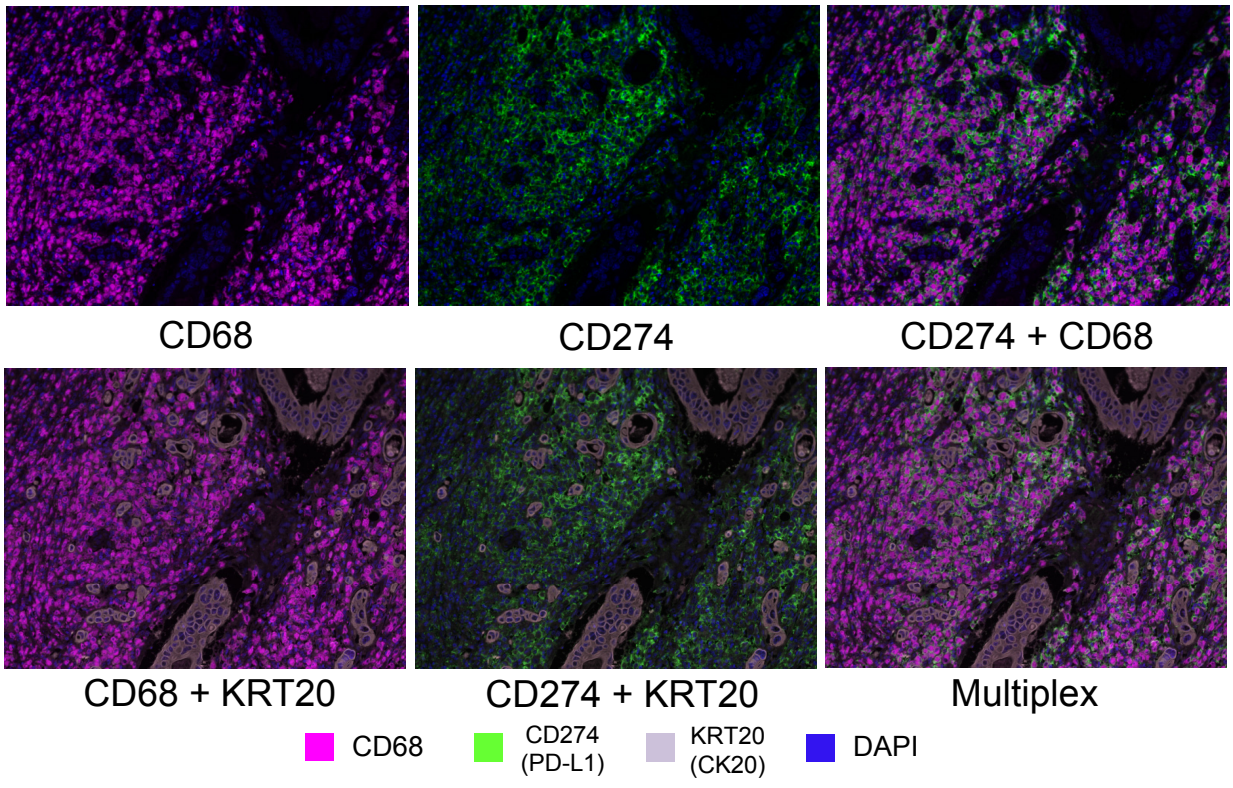
Supplemental Figure S4. Relapse-free survival analysis of the CRC risk subpopulation using NCBI-GEO data set.

Scatter plots of log_2-transformed *CD8A* and *CD274* gene expression values are shown for NCBI-GEO GSE39582 data set (A) and a NCBI-GEO meta-analysis (C), with risk groups indicated (Group I+II as $CD8A^{low}$, III* and IV* as $CD8A^{high}$ dichotomized by *CD274* expression). For relapse-free survival analysis, Kaplan-Meier survival curves for the three risk groups are compared for NCBI-GEO GSE39582 stage I to IV samples (B) (N=557) and stage II and III samples from a NCBI-GEO meta-analysis (D) (N=828). Log-rank test *p*-values are shown for each plot.

A.

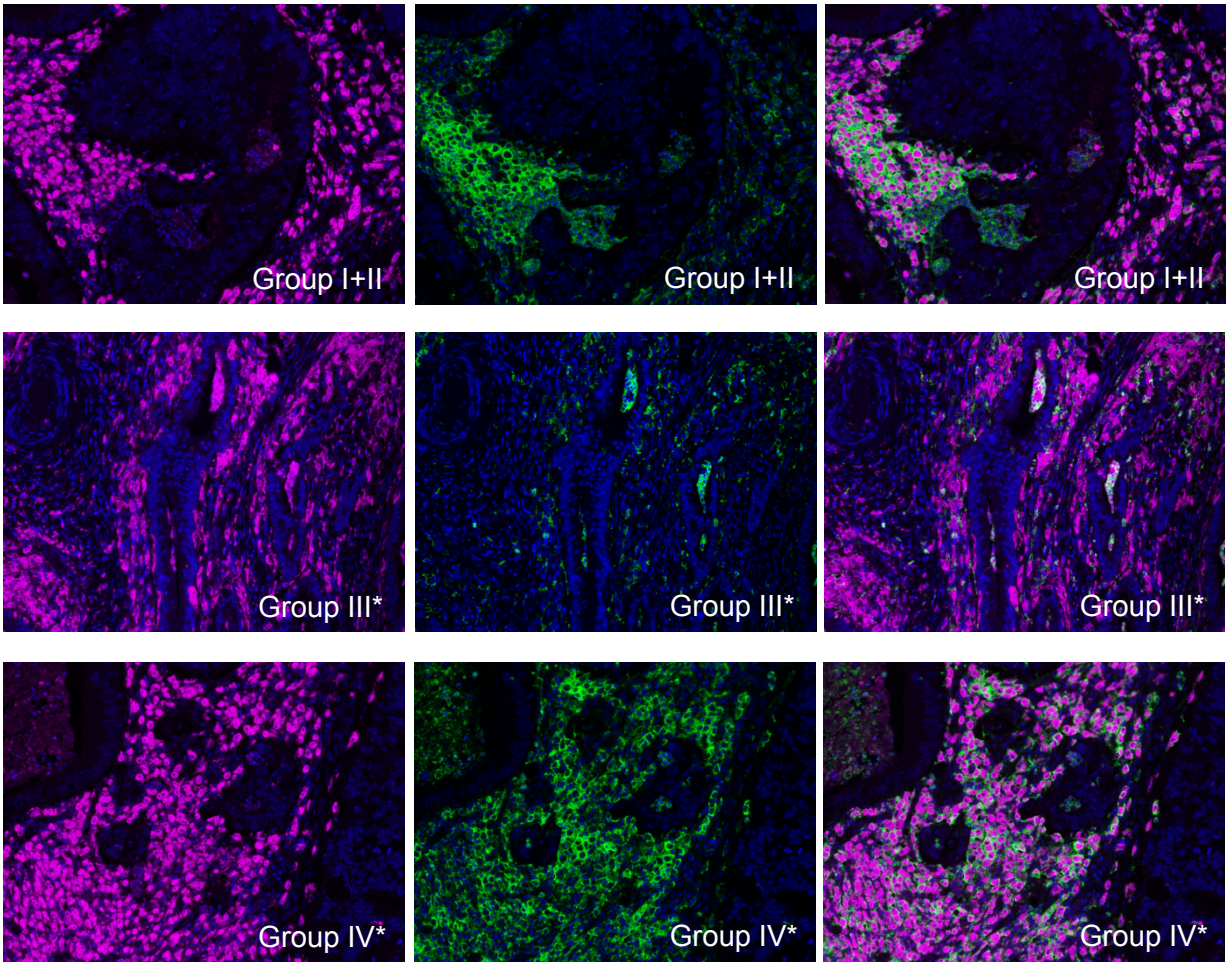


B.



Supplemental Figure S5 A-B

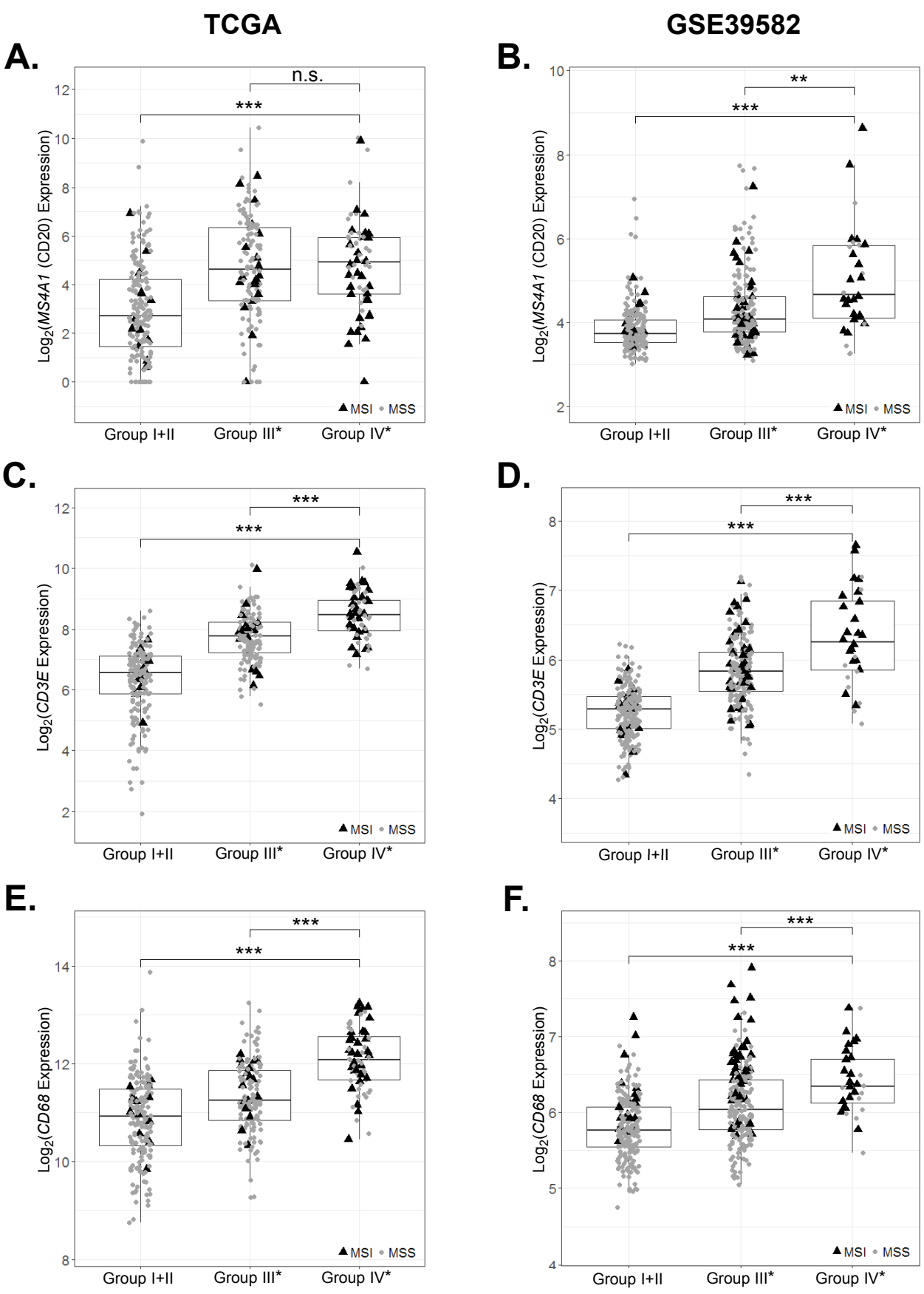
C.



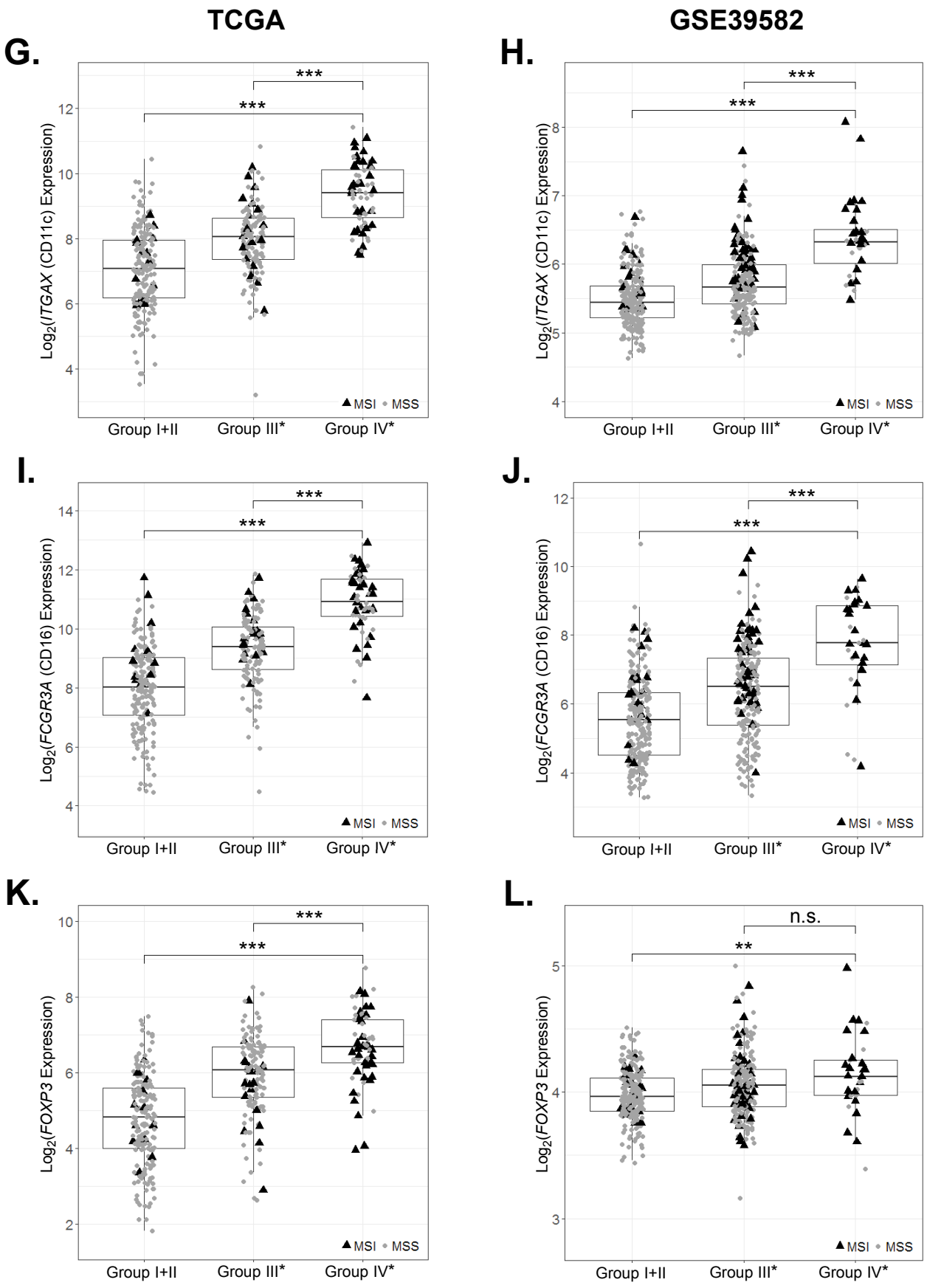
Supplemental Figure S5C

Supplemental Figure S5. The protein expression of PDCD1 (PD-1) and CD274 (PD-L1) in the tumor microenvironment (City of Hope cohort).

Two panels of multispectral fluorescent biomarkers were developed for IHC staining. The first panel, including CD8, CD274, PDCD1, KRT20, and DAPI, was employed to observe (A) the co-localization of CD8 and PDCD1 (PD-1). Representative multiplex image was shown in Figure 4A. The second panel, including CD68, CD274, CK20, and DAPI, was applied to observe (B) the CD274 (PD-L1) expression on KRT20- (CK20-) non-tumor cells and the expression of CD274 (PD-L1) on CD68+ tumor associated macrophages. Representative images from different risk groups are shown in (C). Color combination: CD8 (red), PDCD1 (PD-1) (cyan), CD274 (PD-L1) (green), KRT20 (CK20) (lavender), CD68 (magenta), DAPI (blue).

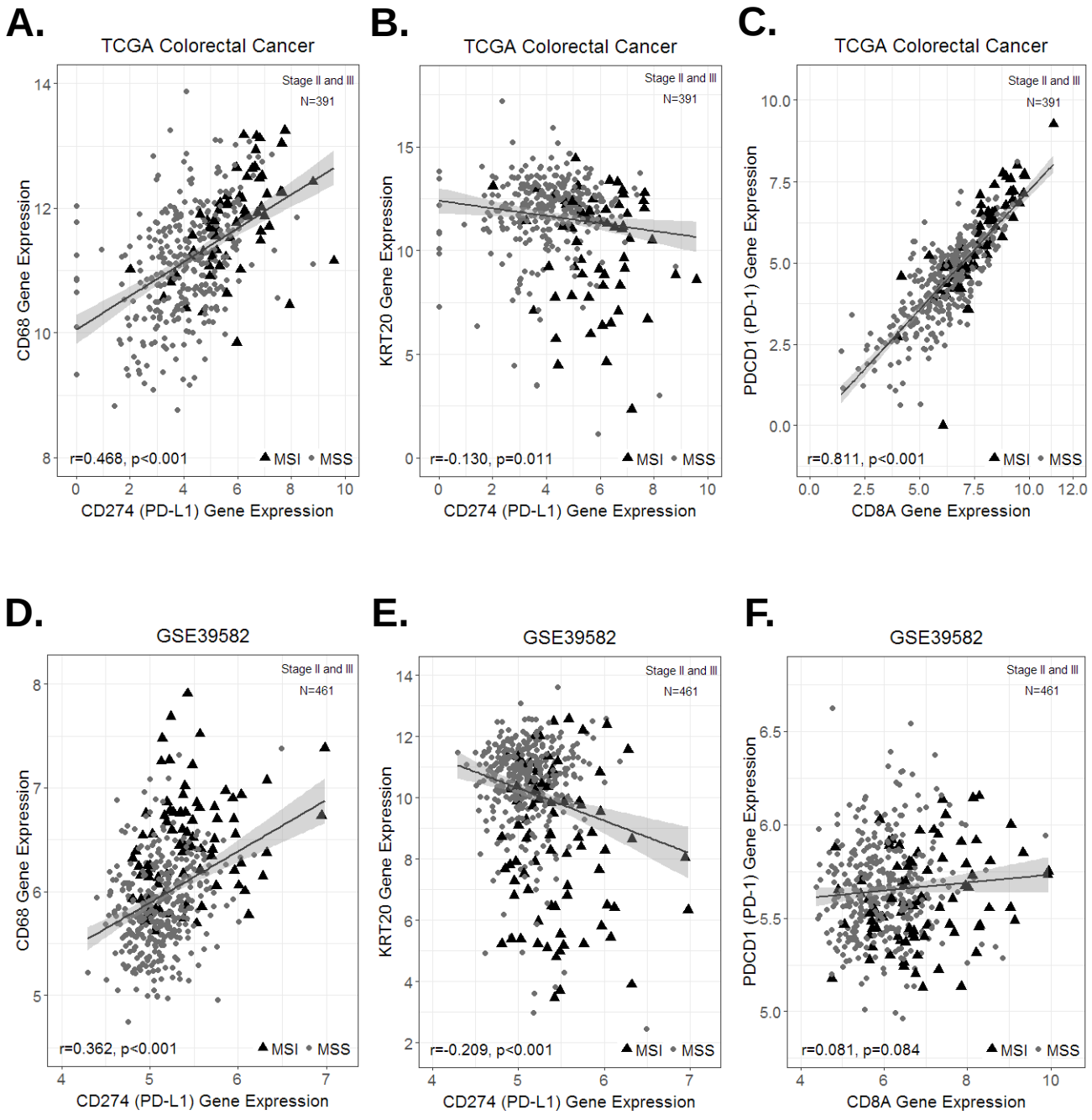


Supplemental Figure S6 A-F



Supplemental Figure S6 G-L

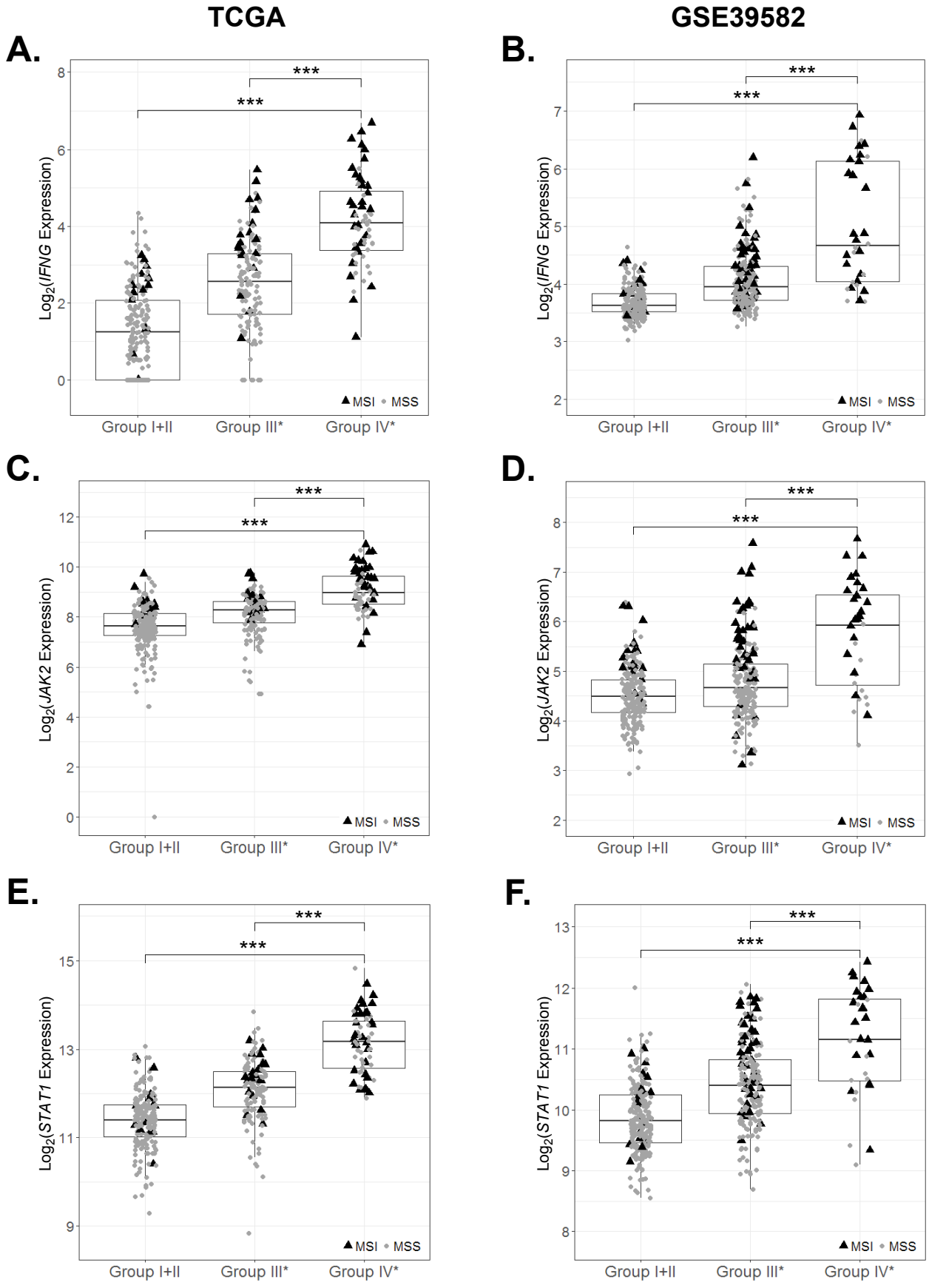
Supplemental Figure S6. Expression levels of genes encoding commonly used cell type-specific markers across the CRC overall survival risk groups in TCGA (left panels; N=391) and NCBI-GEO GSE39582 (right panels; N=461) stage II and III samples. Standard boxplots are applied to visualize gene expression levels across the three risk groups following the convention in Figure 5. Representative genes expressed in major immune cell types: *MS4A1* (CD20) for B cells (A and B), *CD3E* for T cells (C and D), *CD68* for macrophages (E and F), *ITGAX* (CD11c) for dendritic cells (G and H), *FCGR3A* (CD16) for neutrophils (I and J), *FOXP3* for regulatory T cells (K and L). Statistical *p*-values between groups were determined by Welch's t-tests after Bonferroni correction for multiple comparisons: ***$p<0.001$, **$p<0.01$, *$p<0.05$, n.s.: not significant.



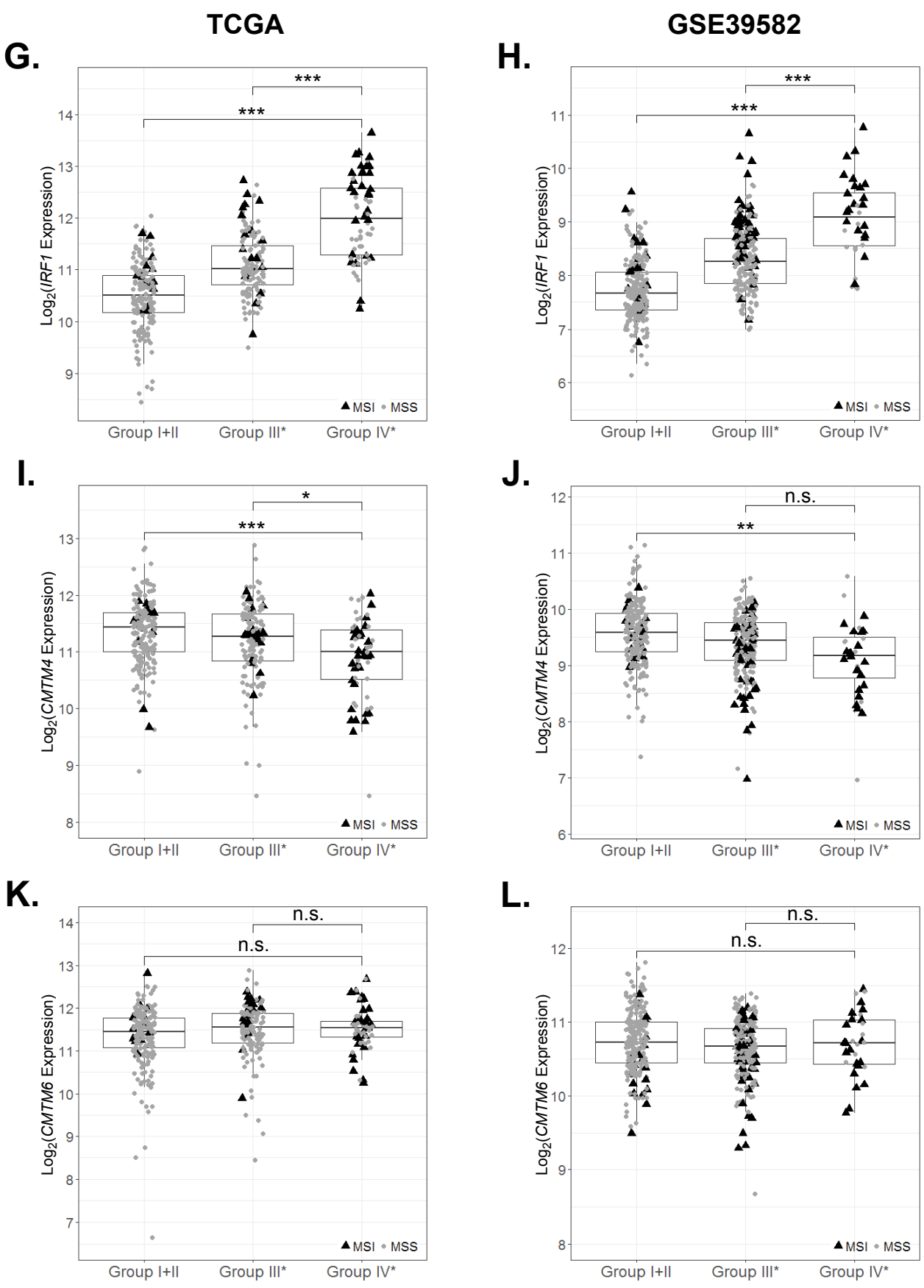
Supplemental Figure S7

Supplemental Figure S7. The gene expression of *PDCD1* (PD-1*)* and *CD274* (PD-L1) in the tumor microenvironment (public data sets).

Scatter plots for the expression of *CD274* vs. *CD68* (A and D), *CD274* vs. *KRT20* (B and E) and *PDCD1* vs. *CD8A* (C and F). Panel A to C applied normalized RNA-Seq data of stage II or III patients from TCGA data set. Panel D to F applied normalized microarray expression data from NCBI-GEO GSE39582 data set. Each dot represents the gene expression data for an individual. MSI (black triangles) and MSS (grey circles) statuses are labeled. A regression line was plotted based on the linear model with the grey shaded region showing the 95% confidence interval. The linear regression *p*-values and the Pearson correlation coefficient r were given at the bottom of the plots.



Supplemental Figure S8 A-F

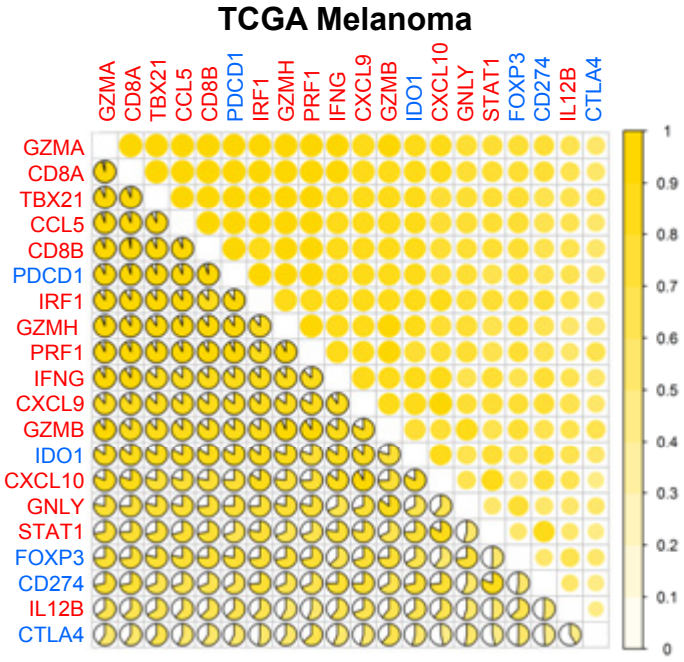


Supplemental Figure S8 G-L

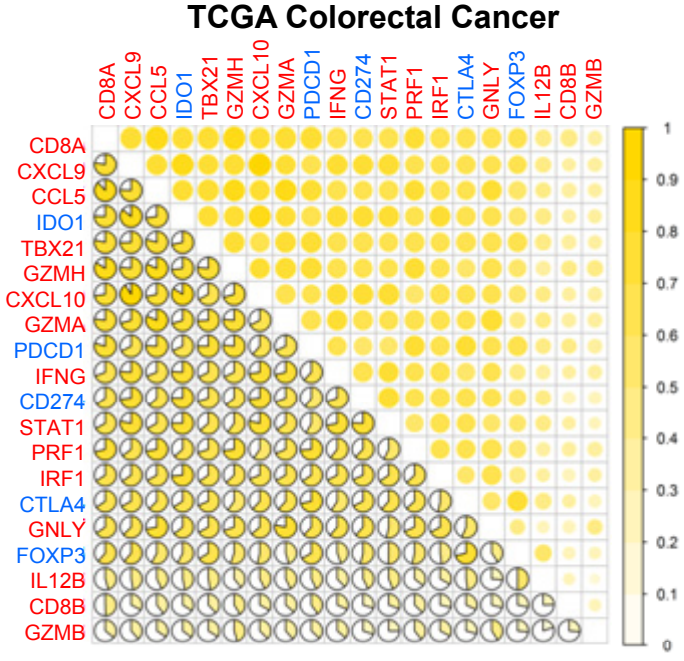
Supplemental Figure S8. The expression of CD274 (PD-L1) regulatory genes across the CRC risk groups.

Standard boxplots are applied to visualize the gene expression across the three risk groups following the convention in Figure 5, with data derived from TCGA and NCBI-GEO GSE39582 plotted to the left and right panels, respectively. The CD274 (PD-L1) regulatory genes exemplified in the presentation include *IFNG* (panel A and B), *JAK2* (panel C and D), *STAT1* (panel E and F), *IRF1* (panel G and H), *CMTM4* (panel I and J), *CMTM6* (panel K and L). Statistical *p*-values between groups were determined by Welch's t-tests after Bonferroni correction for multiple comparisons: ***$p<0.001$, **$p<0.01$, *$p<0.05$, n.s.: not significant.

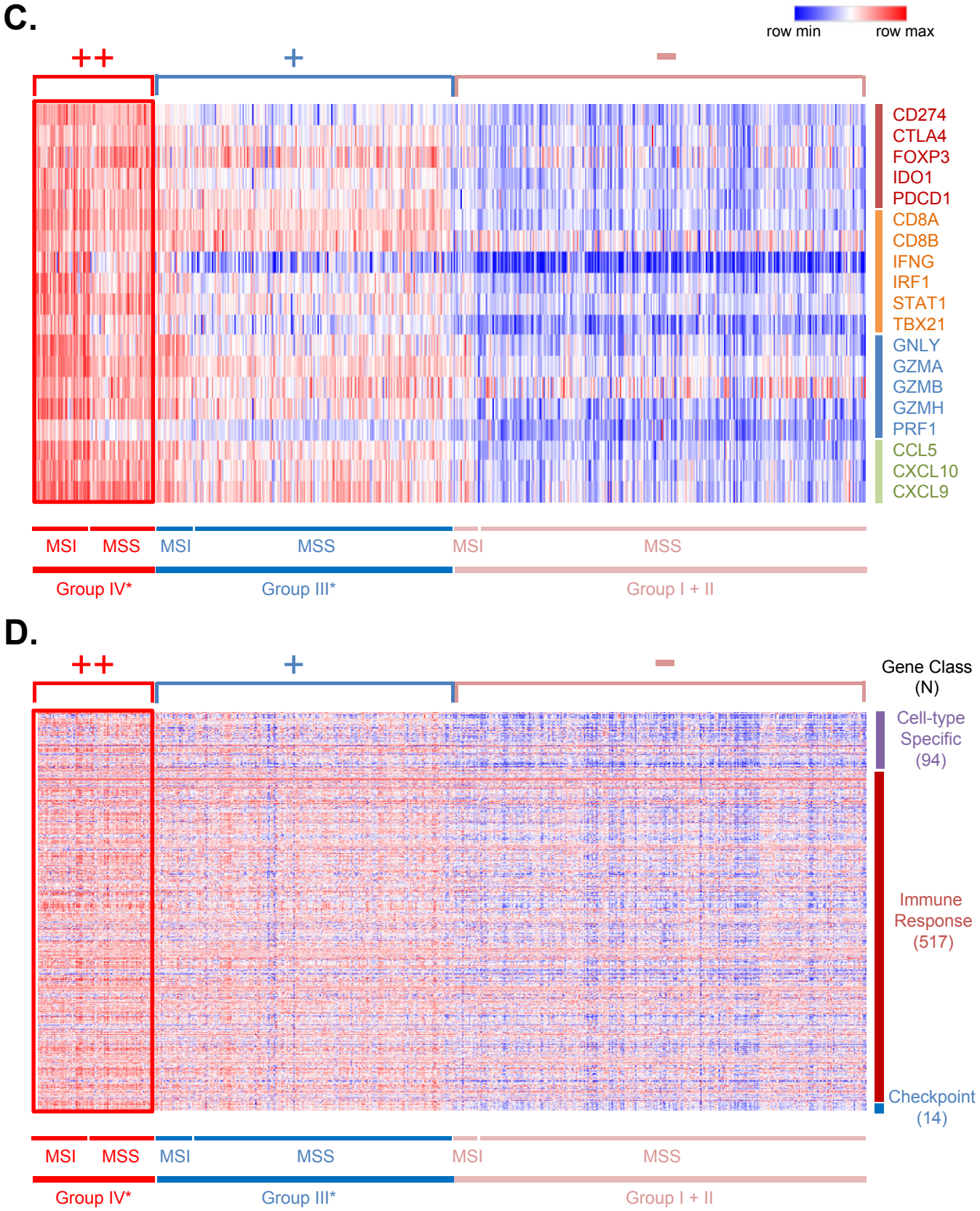
A.



B.



Supplemental Figure S9 A-B

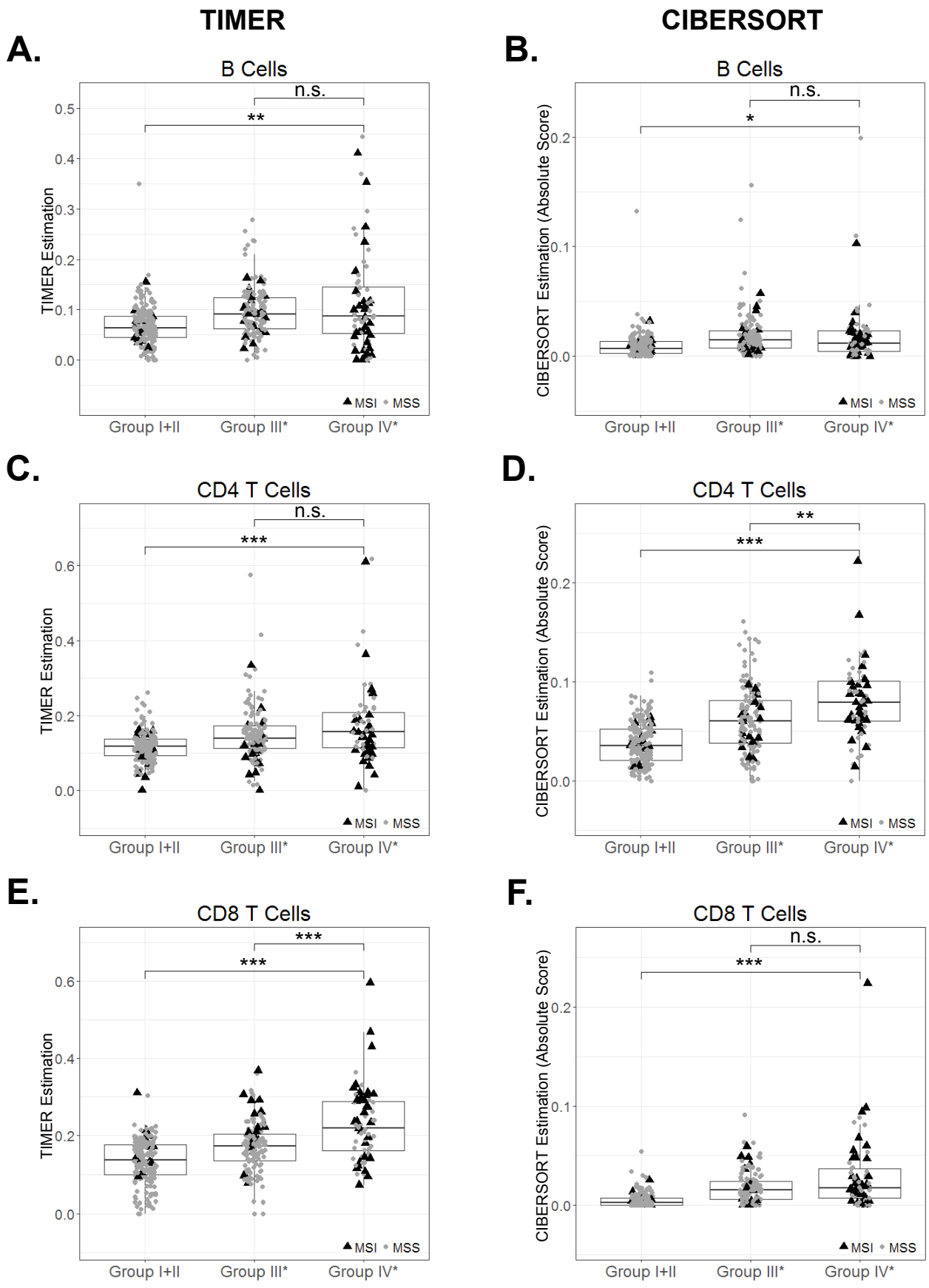


Supplemental Figure S9 C-D

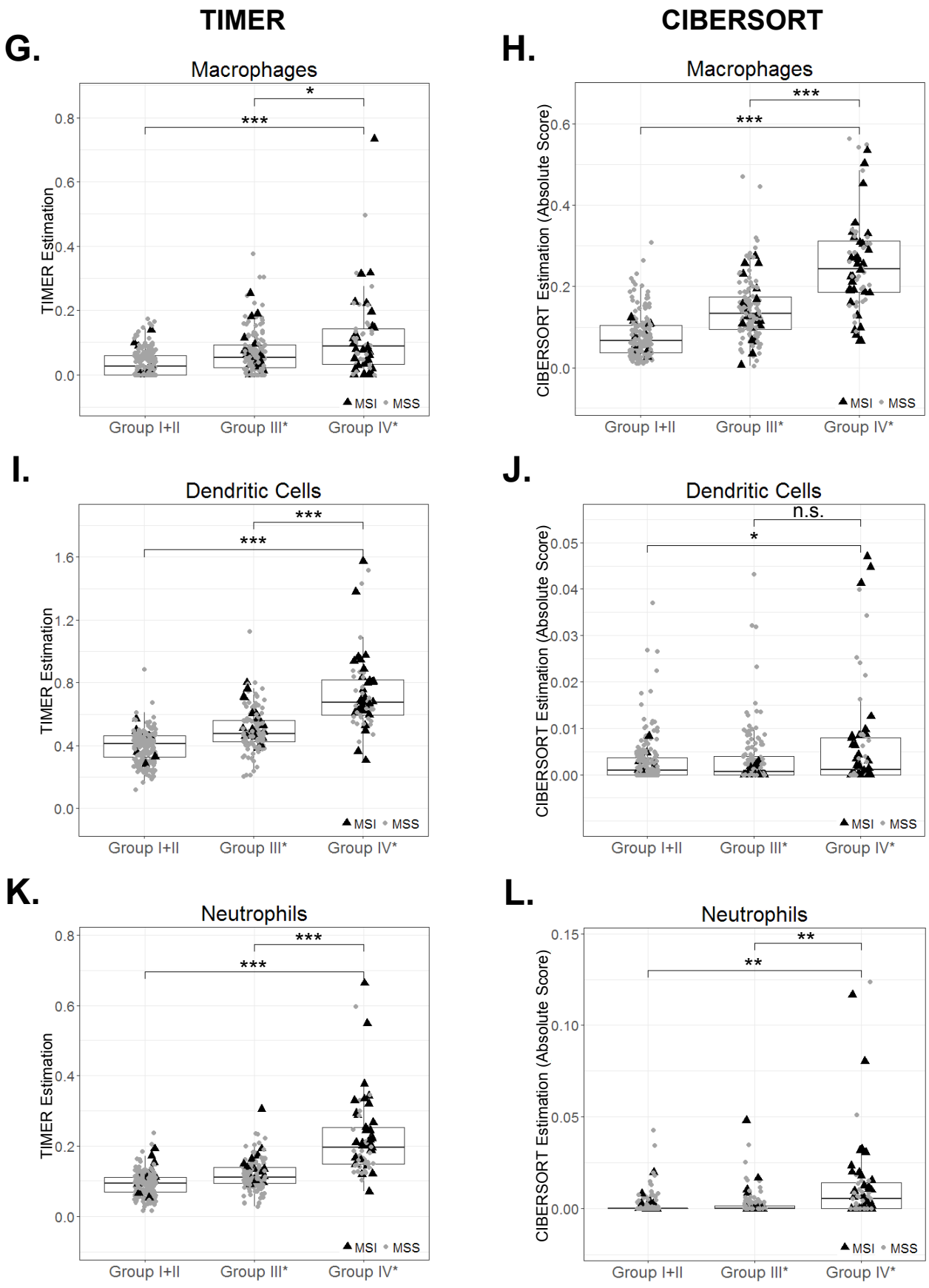
Supplemental Figure S9. Correlation among the expression of pro-inflammatory and immune regulatory genes in human melanoma and CRC in TCGA data set, along with the expression pattern of immune genes across the CRC risk groups.

Spearman's correlations among twenty genes displaying an active Th-1 phenotype are graphically visualized in (A) and (B) for melanoma and CRC. Pro-inflammatory genes and immune regulatory genes are typed red and blue. The Spearman's correlations are represented by the combination of color gradient, size of the circles (upper triangle) and area of yellow wedges (lower triangle).

The gene expression of the Th-1 response genes is also visualized using a heat map across the CRC risk groups in (C). Genes functionally representing the counter-activation of immune suppression, Th-1 signaling, effector functions and CXCR3/CCR5 chemokines are typed red, orange, blue and green to the right. The risk groups stratified by *CD8A* and *PD-L1* expression and the MSI status are delineated at the bottom. The overall expression levels are qualitatively shown with ++, + and – for the three risk groups. In addition, the gene expression of an expanded immune gene panel is shown in (D). Genes grouped by functional annotation as cell-type specific, immune response and immune checkpoint are typed purple, red and blue with the number of annotated genes shown to the right.

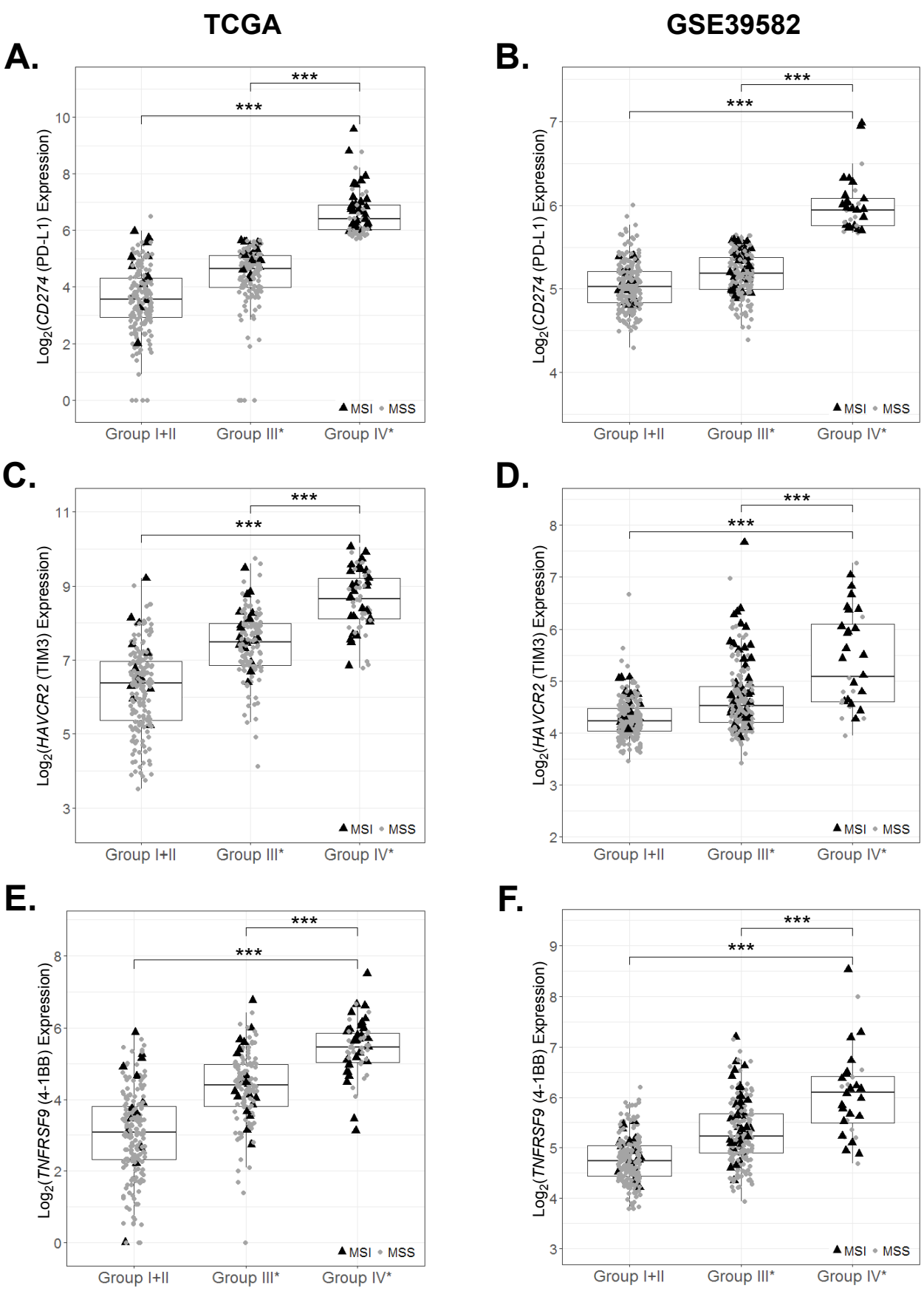


Supplemental Figure S10 A-F

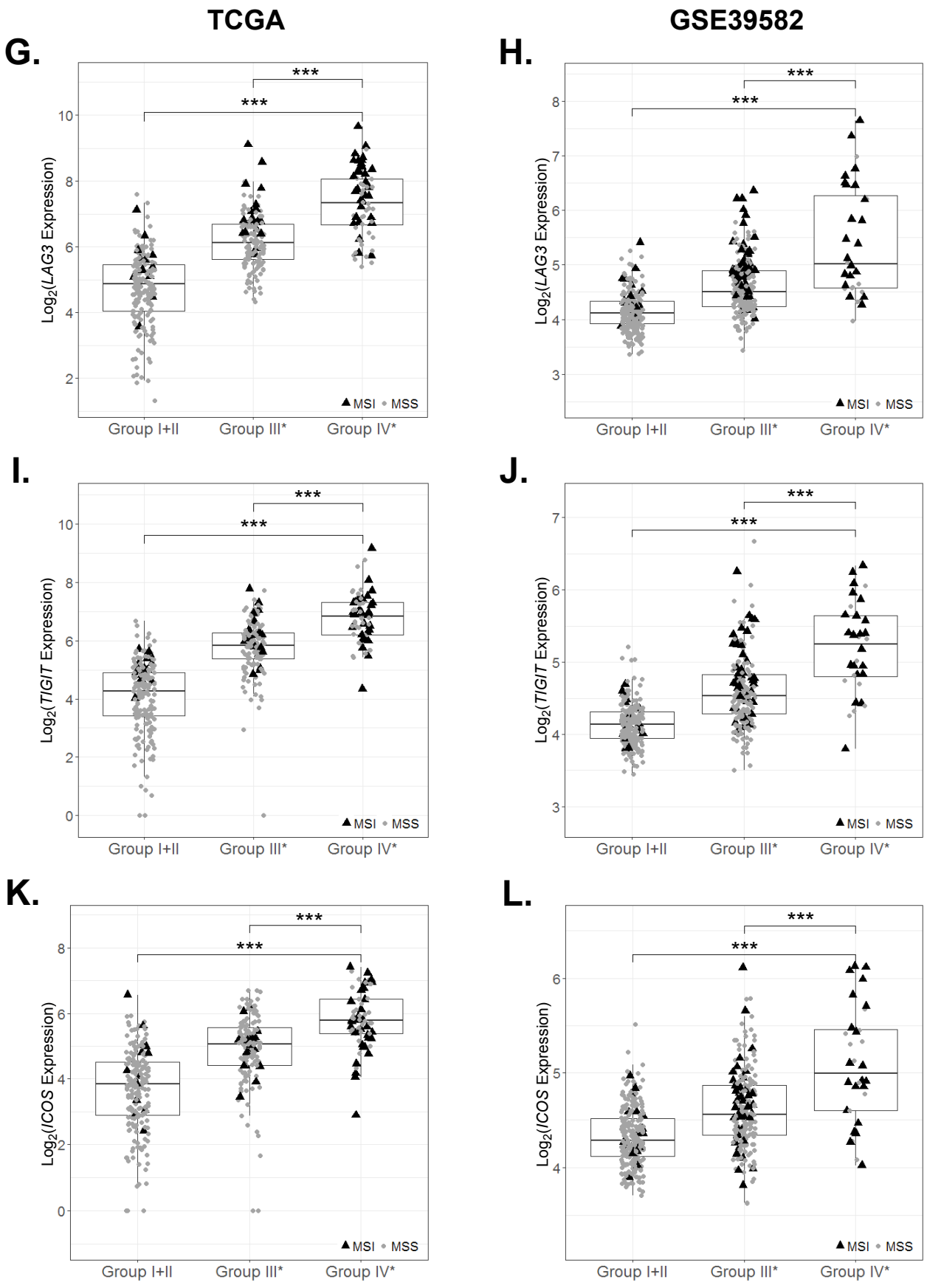


Supplemental Figure S10 G-L

Supplemental Figure S10. Estimation of the immune cell infiltration across the *CD8A* and *CD274* (PD-L1) expression-stratified CRC risk groups using two tumor deconvolution methods TIMER and CIBERSORT with TCGA gene expression data. Standard boxplots are applied to visualize the immune cell infiltration across the three risk groups. The estimates by TIMER and CIBERSORT are plotted to the left and right panels, respectively. Major immune cell types, including B cells (panel A and B), CD4+ T cells (panel C and D), CD8+ T cells (panel E and F), macrophages (panel G and H), dendritic cells (panel I and J) and neutrophils (panel K and L), are presented. For CIBERSORT, which provided estimates for 22 immune cell types), related sub-types were grouped for the presentation. Statistical *p*-values between groups were determined by Welch's t-tests after Bonferroni correction for multiple comparisons: ***$p<0.001$, **$p<0.01$, *$p<0.05$, n.s.: not significant.



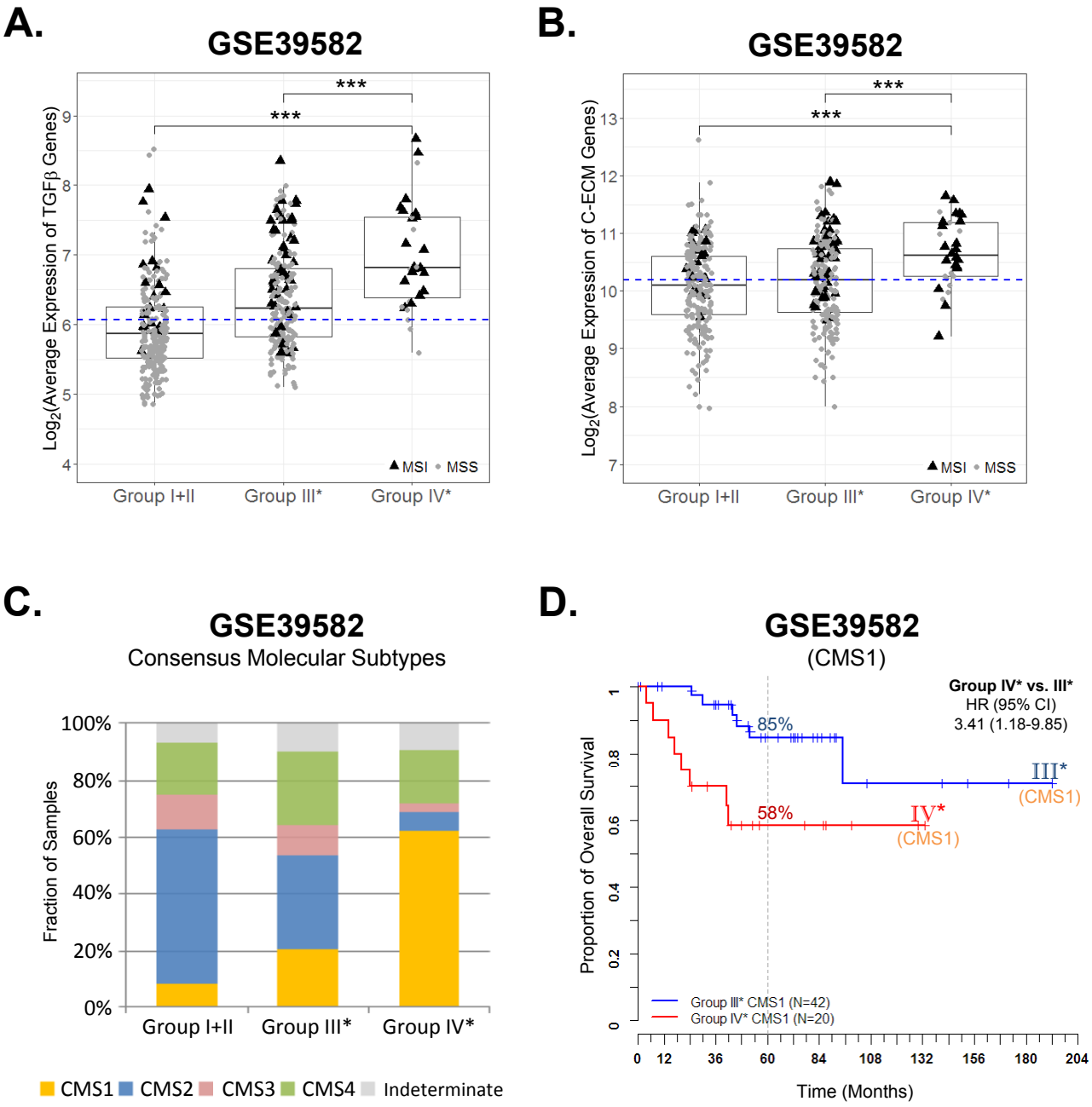
Supplemental Figure S11 A-F



Supplemental Figure S11 G-L

Supplemental Figure S11. The expression of representative immune checkpoint genes across the CRC risk groups.

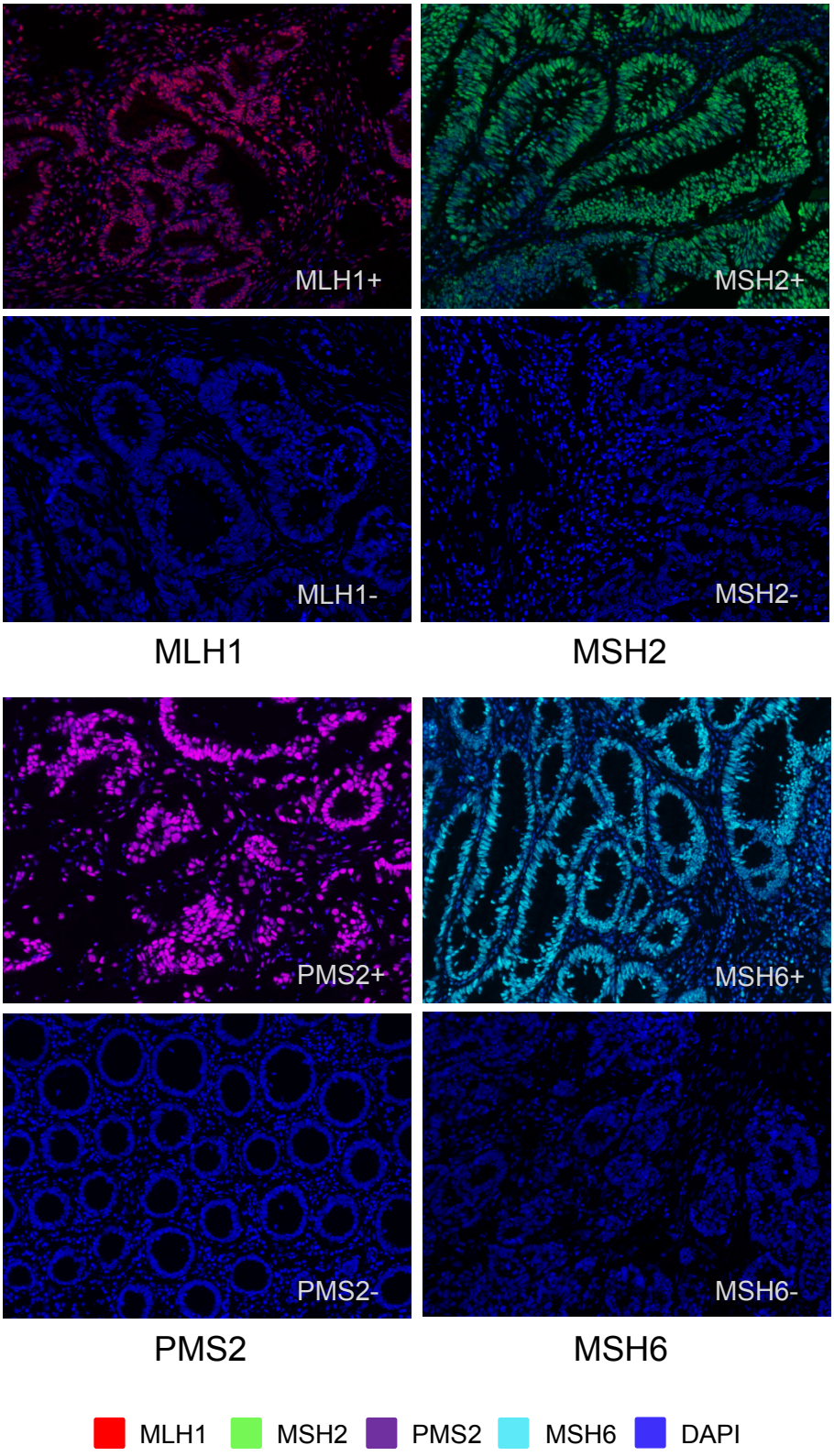
Standard boxplots are applied to visualize the gene expression across the three risk groups following the convention in Figure 5, with data derived from TCGA and NCBI-GEO GSE39582 plotted to the left and right panels, respectively. The immune checkpoint genes exemplified in the presentation include *CD274* (PD-L1; panel A and B), *HAVCR2* (TIM-3; panel C and D), *TNFRSF9* (4-1BB or CD137; panel E and F), *LAG3* (panel G and H), *TIGIT* (panel I and J), *ICOS* (panel K and L). Statistical *p*-values between groups were determined by Welch's t-tests after Bonferroni correction for multiple comparisons: ***$p<0.001$, **$p<0.01$, *$p<0.05$, n.s.: not significant.



Supplemental Figure S12

Supplemental Figure S12. Expression of TGFβ-encoding and C-ECM signature genes and the distribution of consensus molecular subtypes (CMS) across the CRC risk groups in NCBI-GEO GSE39582 stage II and III samples.

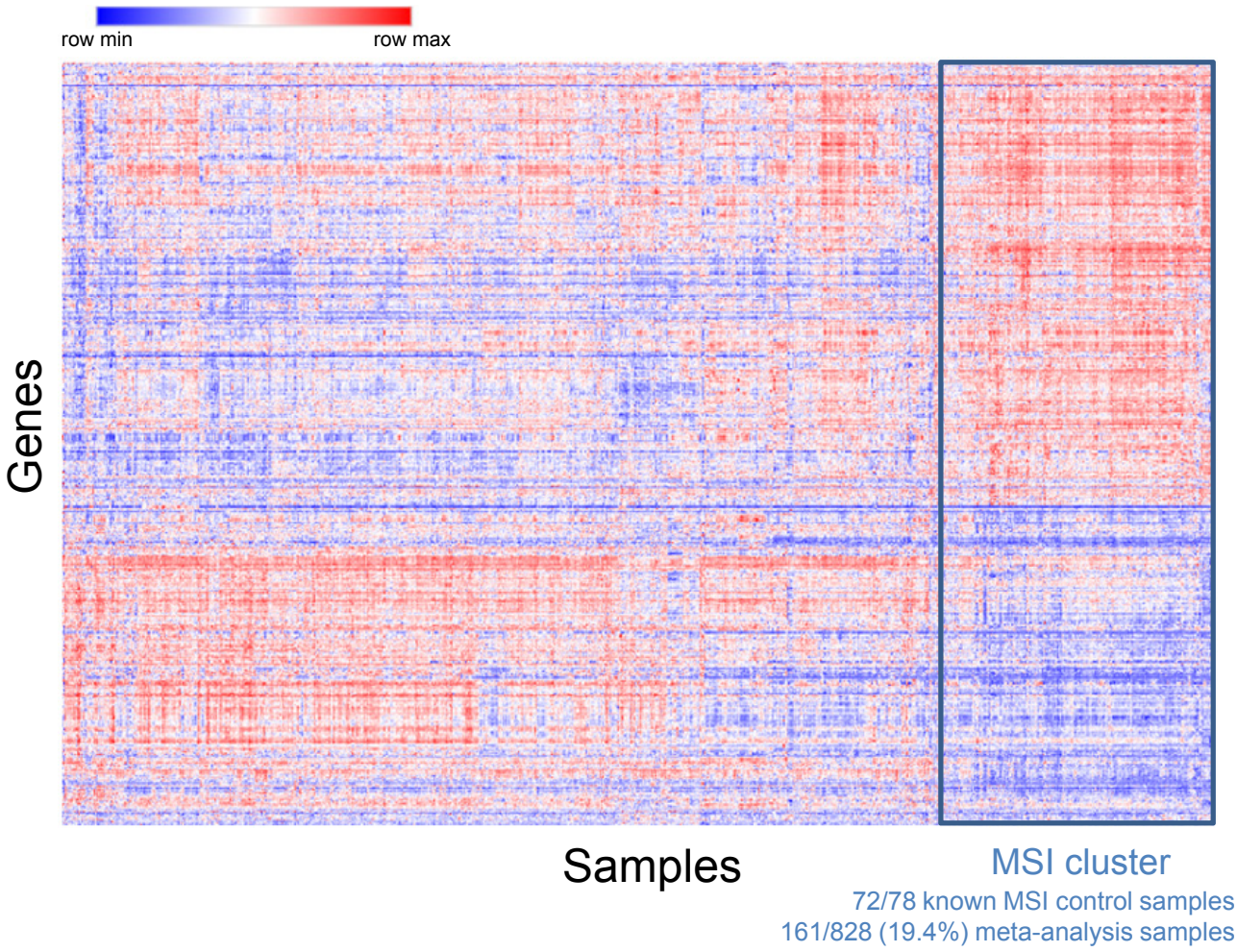
Following the convention in Figure 6 except using NCBI-GEO GSE39582 data set.



Supplemental Figure S13

Supplemental Figure S13. Multispectral fluorescent IHC staining of the protein products of DNA mismatch-repair genes (City of Hope cohort).

Seventy-one colorectal cancer tissue specimens from City of Hope Comprehensive Cancer Center were analyzed for their status of MMRD or MMRP based on the protein expression of MLH1 (top left), MSH2 (top right), PMS2 (bottom left) and MSH6 (bottom right). MMRD is based on a complete absence of nuclear staining of at least one MMR protein (exemplified on bottom panel for each protein). MMRP is defined as the presence of nuclear staining of all four MMR proteins in tumor nuclei (exemplified on top panel for each gene). Color combination: MLH1 (Red), MSH6 (Cyan), MSH2 (Green), PMS2 (Magenta), DAPI (blue).



Supplemental Figure S14

Supplemental Figure S14. Inference of microsatellite instability using microarray gene expression profiles in a NCBI-GEO data meta-analysis for stage II and III samples.

For the 828 meta-analysis samples in NCBI-GEO meta-analysis data set, a panel of 543 signature genes, known for their deferential expression between MSI and MSS samples, was applied for the determination of MSI status. Hierarchical clustering, including additional 155 samples having pre-determined MSI/MSS status, was conducted and the result was visualized using a heat map. Each column represents a sample and each row represents a gene. The expression level of a given gene for a sample was indicated by a color scheme (top left). Red color indicates higher expression and blue color indicates lower expression relative to the mean expression value of a given gene across all samples. Samples showing similar gene expression profiles are being clustered together. The resulting MSI cluster covered 72 of 78 known MSI samples and 161 of 828 meta-analysis samples.