Improving predictive accuracy in multiple myeloma using a plasma cell profile derived from single-cell RNA sequencing

Lanting Liu,[1,2*] Hao Sun,[1,2*] Fangshuo Feng,[1,2] Xiyue Sun,[1,2] Jingyuan Ma,[1,2] Rui Lv,[1,2] Tengteng Yu,[1,2] Linhai Ye,[3] Xiuchun Li,[3] Zhen Yu,[1,2] Xiaoyu Zhang,[1,2] Huaqing Jing,[1,2] Yao Yao,[1,2] Fengxia Ma,[1,2] Lugui Qiu[1,2,4] and Mu Hao[1,2,4]

[1]State Key Laboratory of Experimental Hematology, National Clinical Research Center for Blood Diseases, Haihe Laboratory of Cell Ecosystem, Institute of Hematology & Blood Diseases Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Tianjin, China; [2]Tianjin Institutes of Health Science, Tianjin, China; [3]Sniper Medical Technology Co., Ltd., Suzhou,China and [4]Gobroad Healthcare Group, Beijing, China

**LL and HS contributed equally as first authors.*

Correspondence: M. Hao
haomu@ihcams.ac.cn

Received: February 17, 2025.
Accepted: May 22, 2025.
Early view: May 29, 2025.

https://doi.org/10.3324/haematol.2025.287586

Improving predictive accuracy in multiple myeloma using a plasma cell profile derived from single-cell RNA sequencing

Supplemental figures

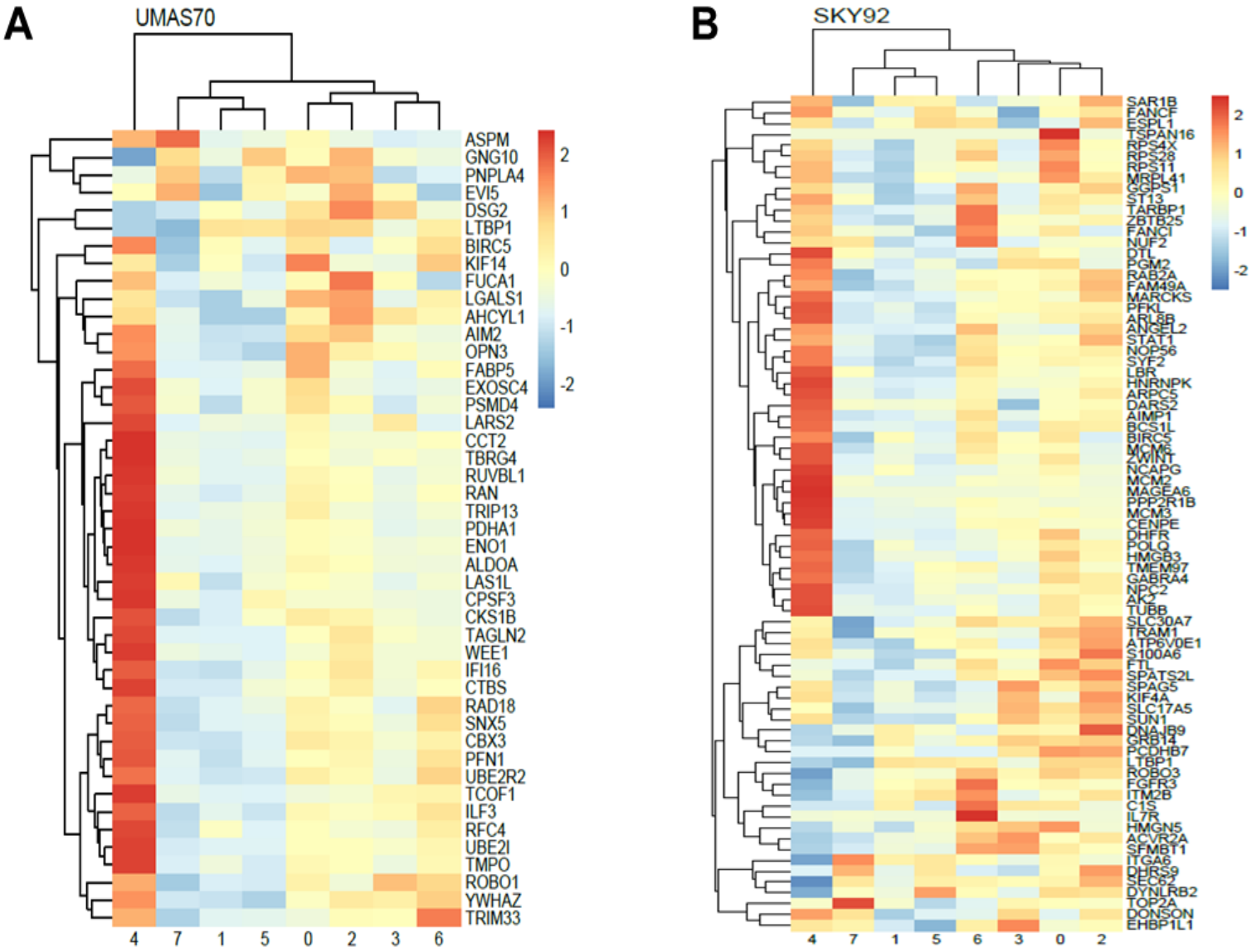
Supplemental tables

Supplemental methods

Supplemental Figures

Supplementary Figure 1. Comparative gene expression profiling of sub-cluster populations of MM cells

(A-B) Heatmaps showing the expression of 70 high-risk gene set (A) and 56 drug resistance-related gene set (B) identified by UAMS in 8 sub-clusters. The sub-cluster number is labeled on the graph bottom.



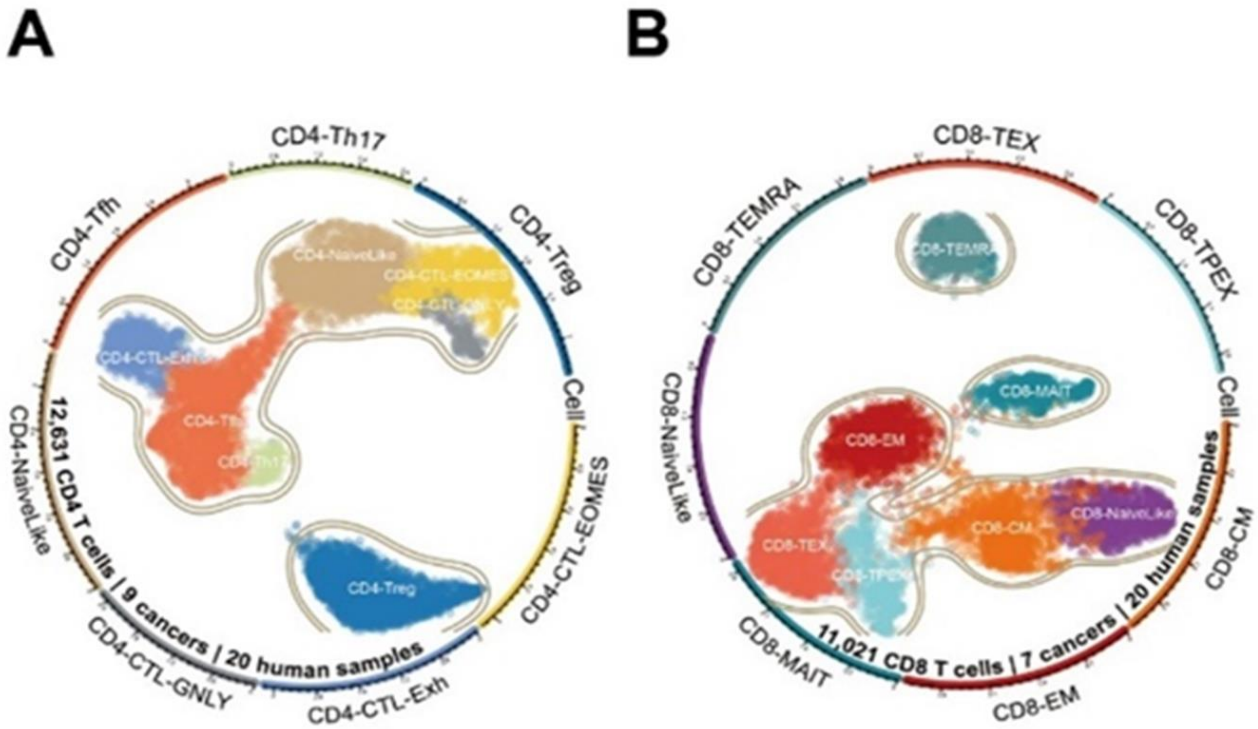
Supplementary Figure 2. Five functional CD4⁺ and seven CD8⁺ T cell clusters were identified based on canonical marker expression patterns across different conditions

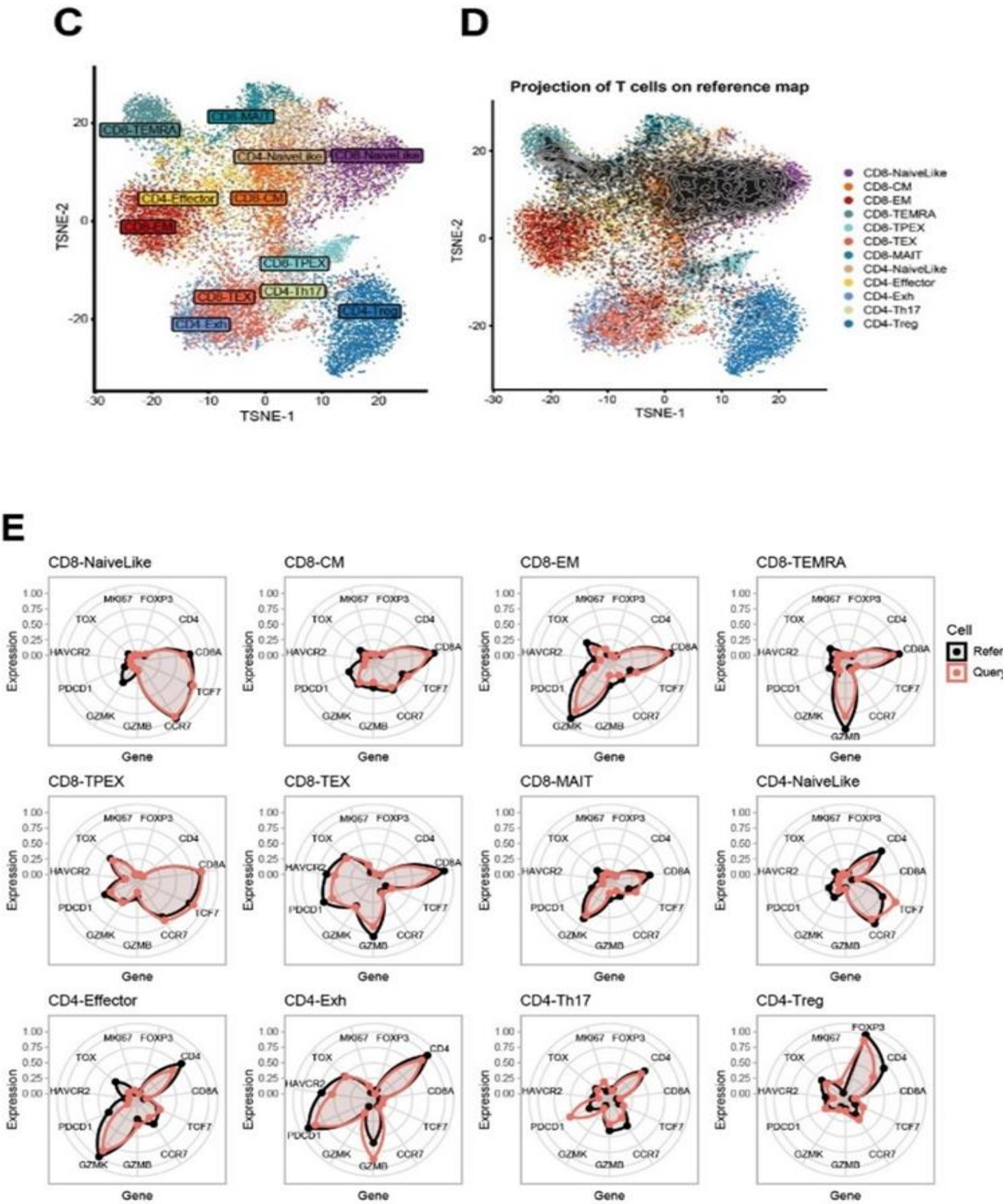
(A-B) UMAP plots displaying the referenced CD4⁺ and CD8⁺ T cells from integrated pan-cancer single-cell datasets.

(C) t-SNE plot showing the integrated referenced cells of CD4⁺ and CD8⁺ T cells.

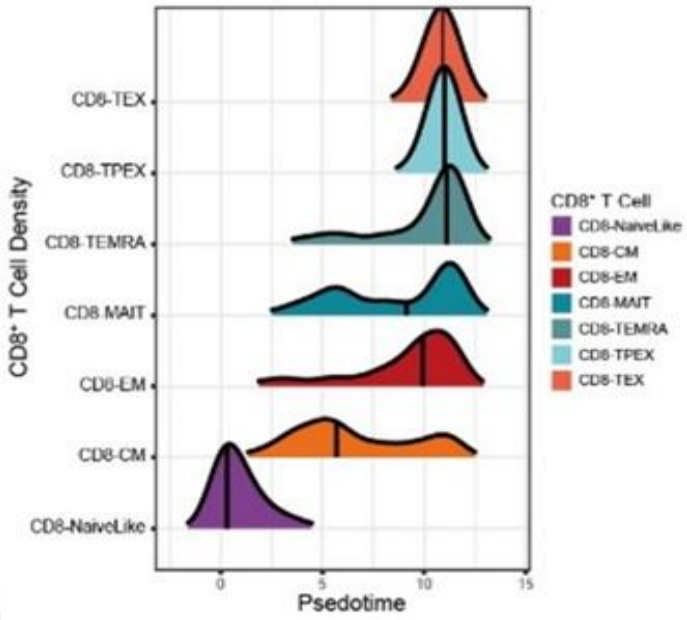
(D) t-SNE plot illustrating the projection of T cells onto the referenced cells. Contour

lines and black dots represent the projected cells.
(E) Radar chart showing the expression levels of representative genes in referenced and projected cells.
(F) Density line plot indicating the distribution of $CD8^+$ T cells along the pseudo-time.
(G) Clustered cell states of $CD8^+$ T cells identified through pseudo-time trajectory analysis.
(H) Distribution of $CD8^+$ T cell subtypes along the pseudo-time trajectory.
(I) Density line plot (upper) and PCA plot (lower) illustrating the distribution of $CD8^+$ T cells across three sample groups.
(J) Pseudo-time trajectory showing the order and distribution of $CD8^+$ T cells (left). Colored dot plot depicting the cell scores within $CD8^+$ T cells (right).
(K) Fitted line with dot plot showing the expression levels of representative transcription factors and exhaustion-related genes along pseudo-time across different T cell states.
(L) Fitted line with dot plot illustrating metabolism scores and the expression of representative genes along pseudo-time. Colors represent different T cell states.
(M) UMAP plots displaying the distribution of CD8+ T cell subtypes in two patient groups along the pseudo-time trajectory.
(N) Volcano plot highlighting differentially expressed genes in CD8+ memory T cells between the two patient groups.
(O) Bar plot showing the enriched upregulated and downregulated pathways of differentially expressed genes in memory T cells.

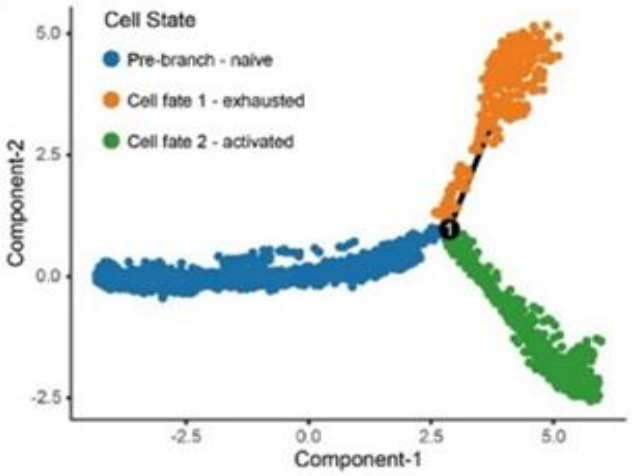



C
D
Projection of T cells on reference map
TSNE-1
TSNE-2
CD8-NaiveLike
CD8-CM
CD8-EM
CD8-TEMRA
CD8-TPEX
CD8-TEX
CD8-MAIT
CD4-NaiveLike
CD4-Effector
CD4-Exh
CD4-Th17
CD4-Treg
E
Cell
Reference
Query
Expression
Gene
MKI67
FOXP3
TOX
CD4
CD8A
HAVCR2
TCF7
PDCD1
CCR7
GZMK
GZMB

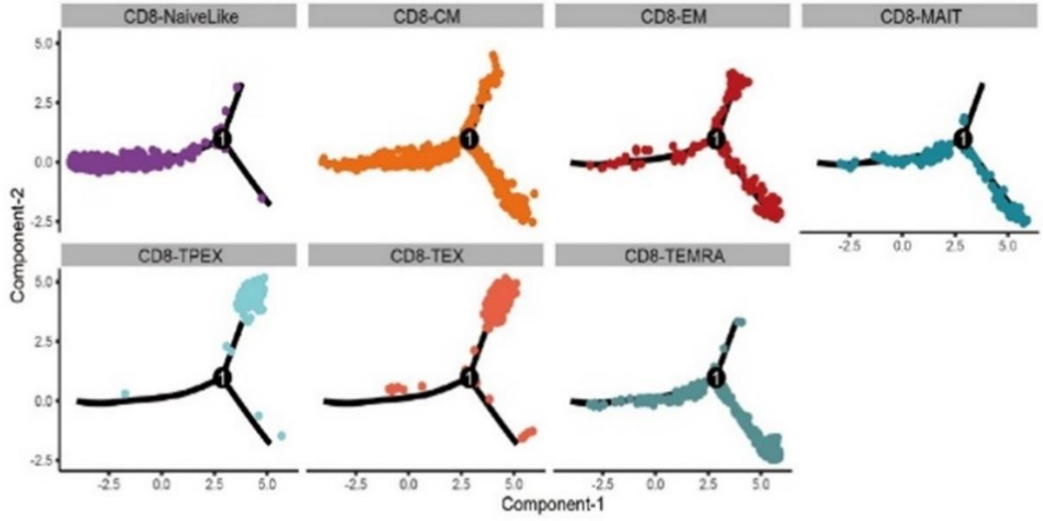
F


CD8-TEX
CD8-TPEX
CD8-TEMRA
CD8 MAIT
CD8-EM
CD8-CM
CD8-NaiveLike
CD8⁺ T Cell Density
0
5
10
15
Psedotime
CD8⁺ T Cell
CD8-NaiveLike
CD8-CM
CD8-EM
CD8-MAIT
CD8-TEMRA
CD8-TPEX
CD8-TEX

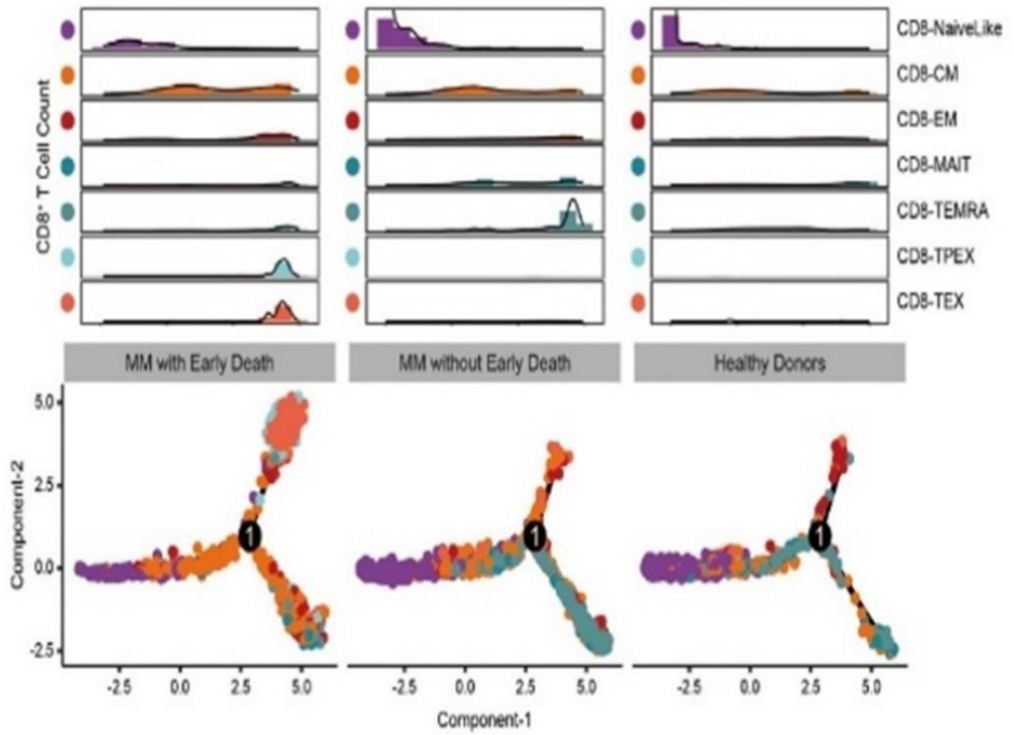
G


Cell State
Pre-branch - naive
Cell fate 1 - exhausted
Cell fate 2 - activated
5.0
2.5
0.0
-2.5
Component-2
-2.5
0.0
2.5
5.0
Component-1

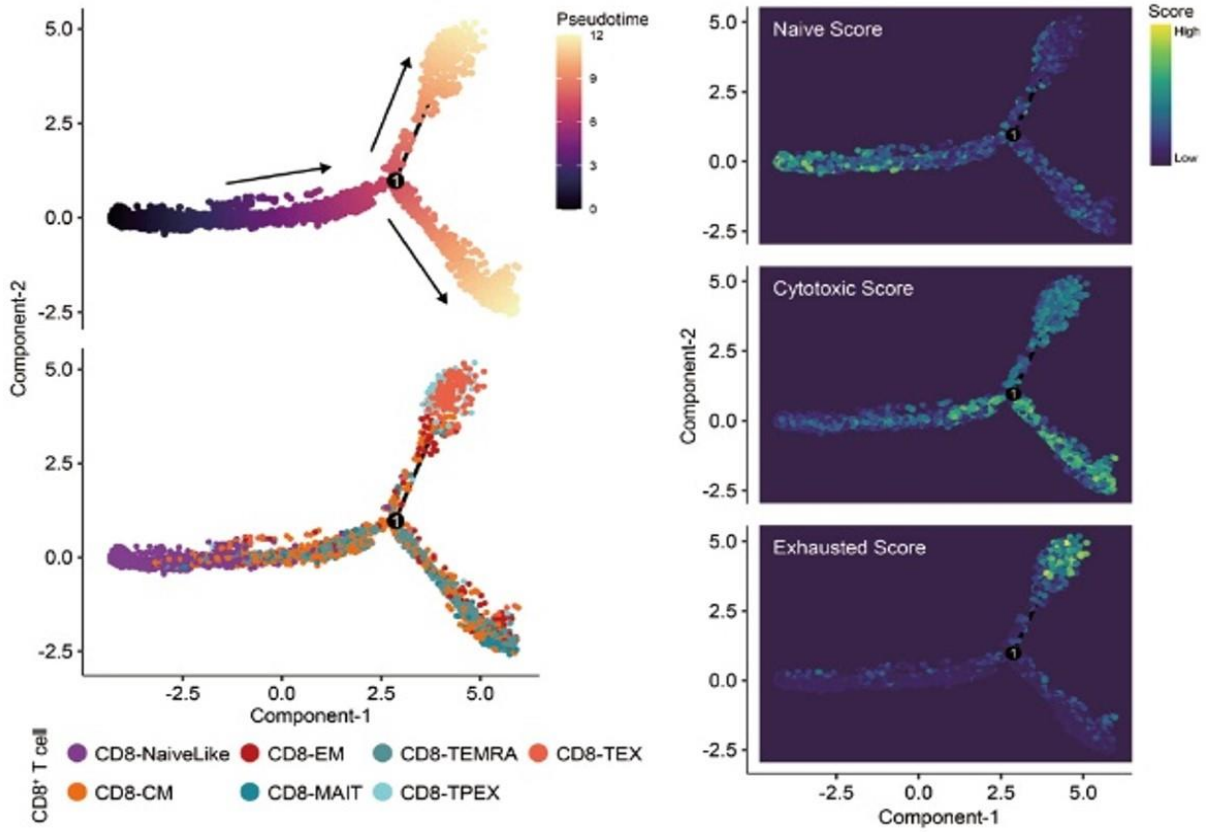
H


CD8-NaiveLike
CD8-CM
CD8-EM
CD8-MAIT
CD8-TPEX
CD8-TEX
CD8-TEMRA
Component-2
Component-1

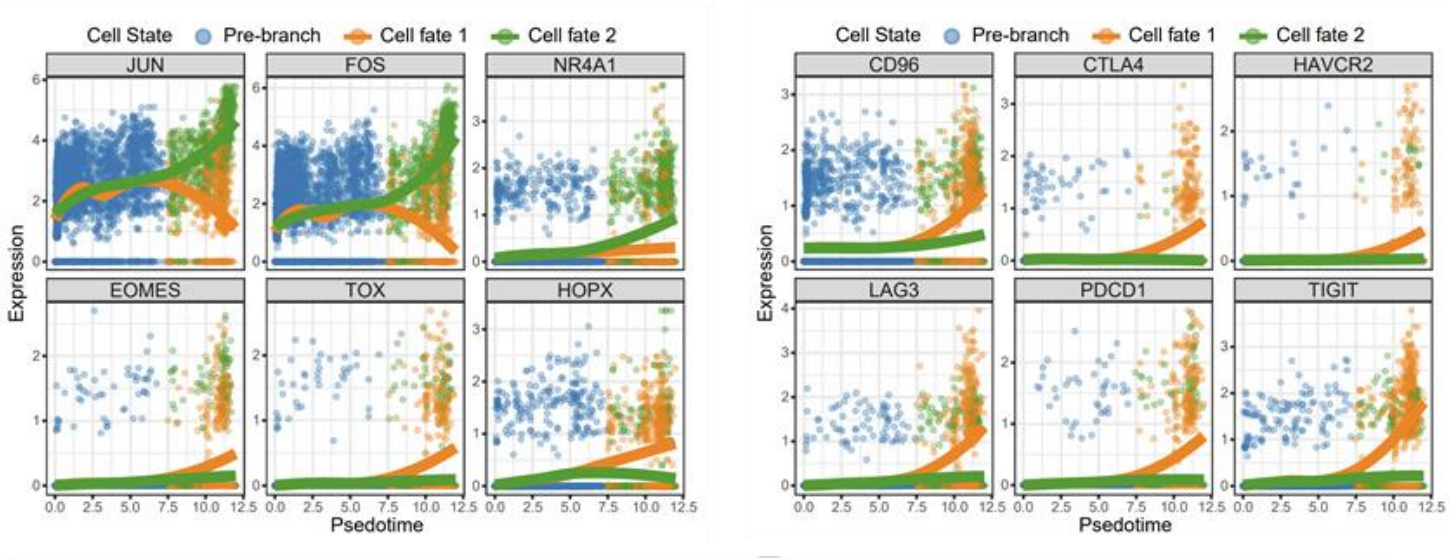
I


CD8+ T Cell Count
CD8-NaiveLike
CD8-CM
CD8-EM
CD8-MAIT
CD8-TEMRA
CD8-TPEX
CD8-TEX
MM with Early Death
MM without Early Death
Healthy Donors
Component-2
Component-1

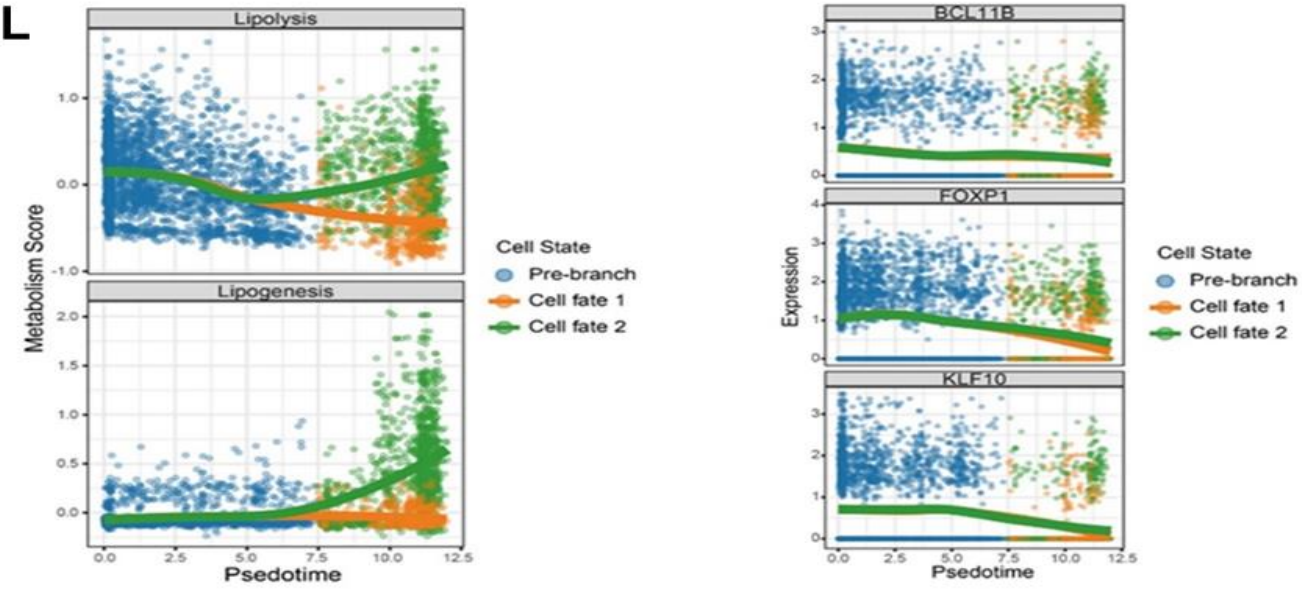
J

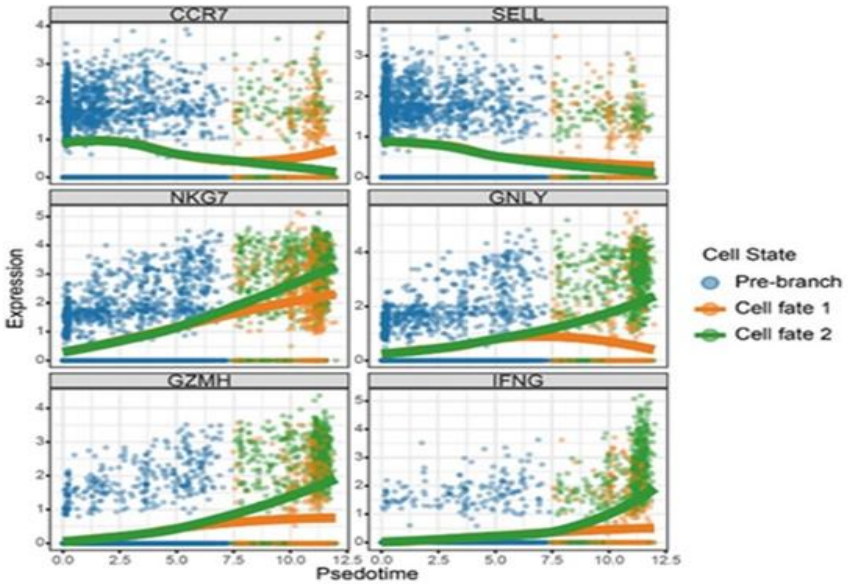

Pseudotime
Component-2
Component-1
CD8+ T cell
CD8-NaiveLike
CD8-CM
CD8-EM
CD8-MAIT
CD8-TEMRA
CD8-TPEX
CD8-TEX
Naive Score
Cytotoxic Score
Exhausted Score
Score
High
Low

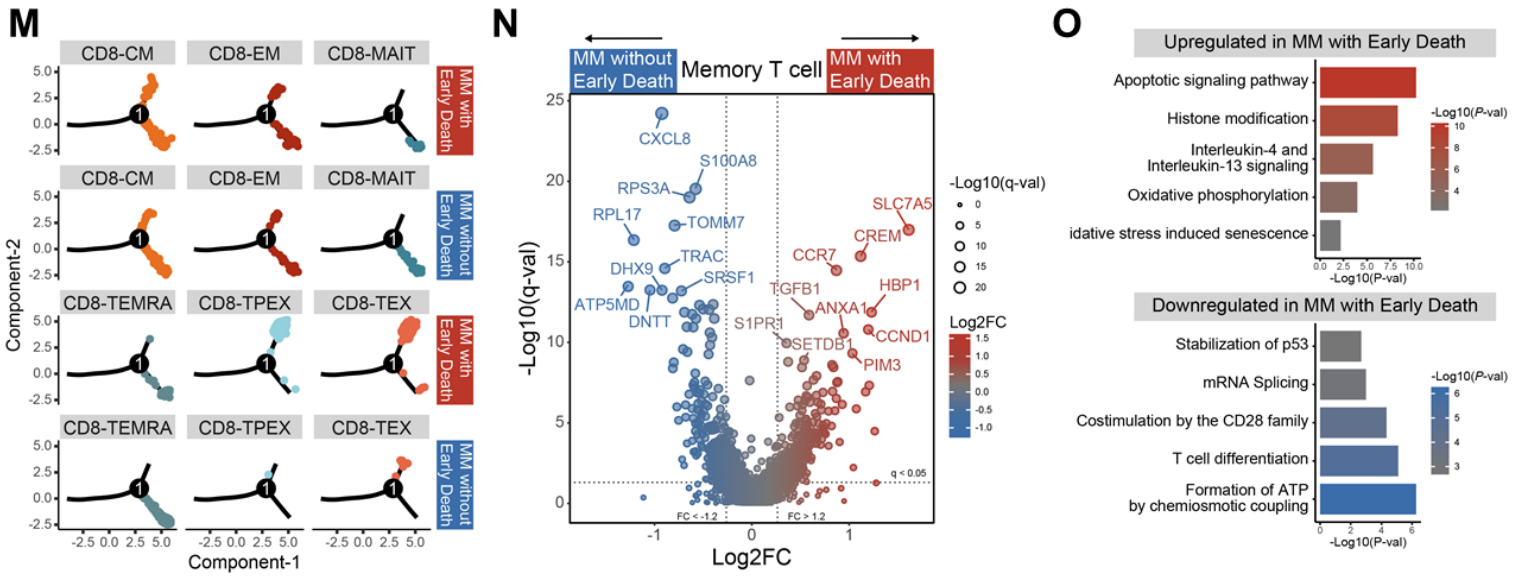
K


Cell State
Pre-branch
Cell fate 1
Cell fate 2
JUN
FOS
NR4A1
EOMES
TOX
HOPX
CD96
CTLA4
HAVCR2
LAG3
PDCD1
TIGIT
Expression
Psedotime

L

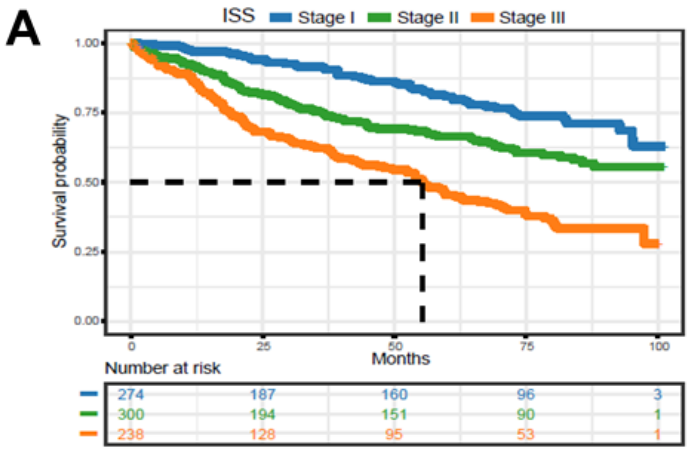

Lipolysis
Lipogenesis
Metabolism Score
BCL11B
FOXP1
KLF10
Expression
Cell State
Pre-branch
Cell fate 1
Cell fate 2
Psedotime


CCR7
SELL
NKG7
GNLY
GZMH
IFNG
Expression
Cell State
Pre-branch
Cell fate 1
Cell fate 2
Psedotime



Supplementary Figure 3. Combining GEP-Based Biomarkers With R-ISS

(A) Kaplan-Meier curve showing the overall survival of MM patients of MMRF-CoMMpass dataset by ISS.



Supplemental tables

Suppl. Table 1. Clinical characteristic of MM patients

MM Patient	Ig type	Light chain type	ISS	R-ISS	Elevated LDH	1q+	del(17p)	t(4;14)	Treatment	Best response	OS
S1	Light chain	Lambda	III	III	-	+	-	-	VRD	sCR	>24 months
S2	IgG	Kappa	I	I	-	-	-	-	VRD	VGPR	>24 months
S3	IgA	Lambda	III	II	-	-	-	-	VRD	CR	>24 months
S4	IgA	Kappa	I	II	-	+	-	+	VRD	VGPR	>24 months
S5	IgG	Kappa	II	II	-	+	-	+	VRD	PR	>24 months
S8	IgD	Lambda	III	II	-	+	-	-	VRD	MR	<24 months
S12	IgA	Lambda	III	II	-	+	-	-	VRD	PR	<24 months
S15	Light chain	Kappa	III	III	+	+	+	-	VRD	NR	<24 months
S16	IgG	Kappa	III	II	-	-	-	-	VRD	PR	>24 months
S24	IgG	Kappa	III	III	+	+	-	+	VRD	VGPR	<24 months
S25	IgAλ	Lambda	III	III	-	+	-	+	VRD	CR	>24 months
S27	IgG	Kappa	III	III	-	-	-	-	VRD	CR	>24 months

Suppl. Table 2. Top 10 up-regulated gene in sC4 compared with other sClusters

Gene	p_val	Location	avg_logFC
LILRB4	0	19q13.4	1.540618984
STMN1	0	1p36.11	1.640841816
GAPDH	0	12p13.31	1.610208026
CRIP1	5.83E-298	14q32.33	1.46268255
TUBA1B	1.40E-288	12q.13.12	1.713650663
TUBB	1.94E-288	6p21.33	1.497821746
HIST1H4C	3.68E-254	6p22.2	1.849714723
CD74	5.54E-191	5q33.1	1.494992301
ITGB7	1.57E-190	12q13.13	1.471848687
CCND2	4.34E-123	12p13.32	1.415058859

Suppl.Table 3. Diagram showing baseline characteristics of MM patients

	Training Set (n=834)	Validation Set 1 (n=426)	Validation Set 2 (n=133)
Age, years			
Median (IQR)	63 (56-70)	60 (52-65)	56 (49-63)
Sex, No. (%)			
Female	332 (40)	165 (39)	57 (43)
Male	502 (60)	261 (61)	76 (57)
ISS, No. (%)			
I	274 (33)	168 (39)	12 (9)
II	300 (36)	135 (32)	43 (32)
III	238 (29)	121 (28)	76 (57)
LDH, No. (%)			
High	117 (14)	19 (4)	32 (24)
Normal	590 (70)	397 (93)	101 (76)
del(17p), No. (%)			
Yes	80 (10)	18 (4)	17 (13)
No	661 (79)	408 (96)	110 (83)
t(4;14), No. (%)			
Yes	95 (11)	14 (3)	31 (23)
No	670 (80)	412 (97)	96 (72)
t(14;16), No. (%)			
Yes	31 (4)	1 (0)	5 (4)
No	734 (88)	425 (100)	122 (92)
1q+, No. (%)			
Yes	263 (28)	6 (1)	95 (71)
No	406 (49)	420 (99)	33 (25)

Training set (MMRF-CoMMpass cohort), Validation set 1 (GSE136337 cohort), Validation set 2 (in-house data).

Suppl.Table 4. Primer sequence of ddPCR

	Upstream primer sequence	Downstream primer sequence	probe sequence	GENE
1	CCAAGTGCAACAAG GAGGTGTA	GCCAGTCCTTGCCCA GAGA	TCGCCGAGAGGGTGA	CRIP1
2	TCGCCTTCGCCTCCT AATC	GCCAACGTGGATGGA GATG	CTAGCCACTATGCGTGAG T	TUBA1B
3	AAACTTAAAGTCGGT GGCACACA	CTGTCTCTGATCCGCA AGCA	AGCAATGAAGGTCTGAGC	CCND2
4	GGTGCTAAGCGCCAT CGT	CGGTTTTGTAATGCCC TGGAT	AGGTGCTCCGGGATA	HIST1H4C
5	GGGTTCATTCAGGCC CAAGT	CATAGCACTGGAGTG GCAGATAGT	CGACGAGAACGGC	CD74
6	CAAGGCTGTGGGCA AGGT	GGAAGGCCATGCCAG TGA	ATCCCTGAGCTGAACG	GAPDH
7	CCCAAGAACAAGGC CAGATTC	CTGCGATAGTAACAGC GGTATCTC	CCATCCCATCCATGACAG AGGACTATGC	LILRB4
8	CTGCCAGTCACCATT CAGCTT	CCCGCTCGAAGGCTT GT	CACCATGTGCTGTCCCTG ACGGG	ITGB7

Supplemental methods

Single-cell data analysis and T cell projection

We used Cell Ranger v3.0.2 to generate gene expression matrices for each cell. Quality control was performed using scCancer v2.1.0, which automatically detected outliers based on the distributions of various metrics [1]. The quality-controlled cell matrices were then processed using Seurat v4 for data normalization, scaling, dimensionality reduction, and clustering. For the integration of the cells from different samples, we corrected batch effect using the harmony v0.1.1 [2]. We integrated the GSE156728 and utility datasets (https://spica.unil.ch/refs/CD8T_human) using the canonical correlation analysis (CCA) algorithm implemented in Seurat v4.0, thereby obtaining a reference dataset of human $CD4^+$ and $CD8^+$ T cells [3]. Finally, we projected our T cells onto this reference dataset using ProjecTILs v3.0 to assign annotations to each T cell [4]. The metabolism signatures were obtained from PathCards (https://pathcards.genecards.org/) and calculated by AddModuleScore function in Seurat v4. Pseudotime-ordered analysis of CD8+ T cells was performed using Monocle2 [5].

Establishment of 7-gene score and MR-ISS

The 7-gene score for each patient in the datasets mentioned above was quantified by the following formula: 7-gene score = $\Sigma i=1,2\ldots n\ \beta(i) \times Exp(i)$, where β represents theregression coefficient for each gene derived from multivariate Cox regression and Exp indicates the expression value of each gene. The RNAseq data was represented as Transcripts Per Million (TPM) and sourced from the MMRF-CoMMpass cohort IA20

version. The specific weights and gene expressions that contribute to the score are as follows: score = 0.678×Exp(LILRB4) + 0.809×Exp(ITGB7) + 0.674×Exp(CRIP1) + 0.800×Exp(TUBA1B) + 0.120×Exp(CCND2) + 0.421×Exp(CD74) – 1.246×Exp(HIST1H4C). The optimal cut point for the 7-gene score was determined using the maximally selected rank statistics from the maxstat v0.7. Subsequently, we constructed MR-ISS model with both univariate and multivariate Cox regression models using clinical risk features and 7-gene group. The Hazard ratio of multivariate Cox regression was utilized to assign scores to these risk features. Based on the cumulative scores of individual patients, we categorized them into four distinct intervals, thereby developing an MR-ISS classification. The Kaplan–Meier method was employed to plot survival curves for log-rank survival analyses.

Multi-color ddPCR

RNA was extracted from bead-enriched $CD138^+$ cells isolated from MM patients using the Qiagen RNeasy Micro Kit. A two-tube ddPCR protocol was employed, whereby each cDNA sample was split into two aliquots, each receiving a distinct ddPCR reaction mix containing primer/probe sets for the target gene and the internal control GAPDH, thereby enabling the simultaneous detection of seven genes. The ddPCR assays were performed using the DQ24 All-in-one Automation Digital PCR System (Suzhou Sniper Medical Technologies Co., Ltd.) based on multicolor fluorescence detection. For each sample, the absolute expression copy number was calculated using the Poisson statistical formula: absolute expression copy number = $-\ln(1 - P/N) \times M$, where P represents the number of positive droplets, N is the total number of droplets, and M

denotes the dilution factor of the reaction system. Normalization was performed by dividing the absolute copy number of the target gene by that of the internal control GAPDH.

Dimensionality reduction, clustering of cells, and visualization

The R package Seurat v.3 was used for data scaling, transformation, clustering, dimensionality reduction, differential expression analysis, and most visualizations [6, 7]. PCA was performed using the top 2,000 variable genes. Graph-based clustering was performed on the PCA-reduced data for clustering analysis. The resolution was set to 0.5 to obtain a more refined result. Briefly, the first 50 PCs of the integrated gene -cell matrix were used to construct a shared nearest-neighbor graph (SNN; 'FindNeighbors'), and this SNN was used to cluster the dataset ('FindClusters') u sing a graph -based modularity -optimization algorithm of the Louvain method for community detection. Then UMAP was performed on the top 30 principal components for visualizing the cells.

Cell cluster annotation with specific maker genes expression

The marker genes, identified by the 'FindAllMarkers' function—with the setting 'only.pos' as 'True', 'min.pct' as 0.25, 'logfc.threshold' as 0.25 and 'test.use' as 'wilcox'—were used to annotate cell clusters. Cluster annotation was confirmed using the celltypist, an automated cell type annotation tool for scRNA-seq datasets based on logistic regression classifiers optimised by the stochastic gradient descent algorithm [8].

Characterizing the cell distribution in sample groups

To characterize the group distribution of cells, odds ratios (OR) were calculated and

used to indicate preferences as reported [3]. Specifically, for each combination of cells i and group j, a 2 by 2 contingency table was constructed, which contained the number of cells i in group j, the number of cells of cell i in other groups, the number of non -i cells in group j, the number of non -i cells in other groups. Then Fisher's exact test was applied on this contingency table, thus OR and corresponding p -value could be obtained. We labeled groups with a p-value < 0.05 using an asterisk (*). A higher OR with an asterisk indicated that cells i was more preferred to distribute in group j, while a lower OR with an asterisk indicated that cells i was preferred to less distribute in group j.

InferCNV analysis

We used InferCNV to infer copy number variations (CNVs) from single-cell RNA sequencing data. The gene expression matrix was first preprocessed and normalized to reduce technical noise and batch effects. CNVs were inferred by comparing the expression profiles of individual cells to a reference population of normal cells. Specific chromosomal regions with significant deviation in expression patterns were identified as CNV regions. The results were visualized through heatmaps, which displayed the CNV landscape across the cell population. For scoring CNVs, we calculated a CNV score for each cell, representing the degree of CNV severity based on its expression profile. The CNV score was derived by assessing the deviation of each cell's gene expression from the normal reference. Higher scores indicated more pronounced copy number alterations, while lower scores suggested minimal or no CNV. These CNV scores were then used for further statistical analysis to correlate with biological or

disease-related features.

Construction of transcriptional regulatory network and gene co -expression network

We selected specific genes from cluster 4 to infer the transcriptional regulatory network and establish a co-expression network. Candidate transcription factors and transcription factor -target pairs were obtained from TRRUST v2 [9], which includes 8,444 regulatory interactions (annotated as 'activation', 'repression', and 'unknown') for 800 transcription factors in humans. We retained only the edges where both the transcription factor and target were present in the specific genes of cluster 4. Moreover, we kept only the transcription factors with at least two edges annotated as 'activation' or 'repression' in the final network. To construct the co-expression network, we calculated the Pearson correlation between the specific genes of cluster 4 in cluster 4 MM cells. Genes with a P-value < 0.05 and an absolute correlation coefficient > 0.4 were retained. Clustering of the co-expression network was performed using MCODE in Cytoscape v3.9.0, and the top 4 clusters with a degree cutoff of 5 were kept. The networks were visualized using Gephi v0.10.1 with the 'Fruchterman Reingold' layout.

References

1. Guo W, Wang D, Wang S, Shan Y, Liu C, Gu J. scCancer: a package for automated processing of single-cell RNA-seq data in cancer. Briefings in bioinformatics. 2021;22(3):
2. Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nature methods. 2019;16(12):1289-1296.
3. Zheng L, Qin S, Si W, et al. Pan-cancer single-cell landscape of tumor-infiltrating T cells. Science (New York, NY). 2021;374(6574):abe6474.
4. Andreatta M, Corria-Osorio J, Müller S, Cubas R, Coukos G, Carmona SJ. Interpretation of T cell states from single-cell transcriptomics data using reference atlases. Nature communications. 2021;12(1):2965.

5. Qiu X, Mao Q, Tang Y, et al. Reversed graph embedding resolves complex single-cell trajectories. Nature methods. 2017;14(10):979-982.
6. Ríos-Tamayo R, Sáinz J, Martínez-López J, et al. Early mortality in multiple myeloma: the time-dependent impact of comorbidity: A population-based study in 621 real-life patients. American journal of hematology. 2016;91(7):700-704.
7. Sonneveld P, Avet-Loiseau H, Lonial S, et al. Treatment of multiple myeloma with high-risk cytogenetics: a consensus of the International Myeloma Working Group. Blood. 2016;127(24):2955-2962.
8. Domínguez Conde C, Xu C, Jarvis LB, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science (New York, NY). 2022;376(6594):eabl5197.
9. Han H, Cho JW, Lee S, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic acids research. 2018;46(D1):D380-d386.