Deconvoluting single-cell transcriptomics reveals cellular programs regulated by cell-cell communication in colorectal cancer

Han Zhang[1], Binfeng Lu[3], Gregory F. Cooper[1], Anwaar Saeed[2], John W. Paisley[4], Xinghua Lu[1], Lujia Chen[1]*

[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, 15206, USA

[2]UPMC Hillman Cancer Center, Pittsburgh, PA 15232, USA

[3]Center for Discovery and Innovation, Hackensack Meridian Health, Nutley, NJ 07110, USA

[4]Columbia University, New York, NY 10027, USA

*Correspondence: luc17@pitt.edu

Conflicts of Interest Statement:
Anwaar Saeed reports a leadership role with Autem therapeutics, Exelixis, KAHR medical and Bristol-Myers Squibb; consulting or advisory board role with AstraZeneca, Bristol-Myers Squibb, Merck, Exelixis, Pfizer, Xilio therapeutics, Taiho, Amgen, Autem therapeutics, KAHR medical, and Daiichi Sankyo; institutional research funding from AstraZeneca, Bristol-Myers Squibb, Merck, Clovis, Exelixis, Actuate therapeutics, Incyte Corporation, Daiichi Sankyo, Five prime therapeutics, Amgen, Innovent biologics, Dragonfly therapeutics, Oxford Biotherapeutics, Arcus therapeutics, and KAHR medical; and participation as a data safety monitoring board chair for Arcus therapeutics. The remaining authors have no relevant financial interests to disclose.

Abstract

Cells within a tissue microenvironment communicate through intricate cell-cell communication (CCC) networks. In this meta-analysis of eight single-cell cohorts encompassing 153 patients and 279 samples, we advance the understanding of CCC networks in colorectal cancers through a novel analytical framework. Employing hierarchical language modeling, we identify gene expression modules (GEMs) that mirror single-cell signaling states, crucial for deciphering the complexity of intercellular interactions. By applying causal discovery methods, we systematically uncover GEMs likely regulated by ligand-receptor signaling and cross-cell-type communication. This analysis reveals cross-cell-type CCC programs, marked by highly correlated GEMs across various

cell types, shedding light on the intricate CCC networks within the tumor microenvironment. Spatial transcriptomics further validate these findings by demonstrating the co-localization of GEMs within CCC programs in distinct spatial domains, emphasizing the spatial dynamics of tumor intercellular communication. Our interactive website (http://44.192.10.166:3838/) and analytical framework equip researchers with powerful tools to explore these complex mechanisms, potentially uncovering novel drug targets and refining strategies for precision immunotherapies. This comprehensive study not only presents a detailed catalog of CCC networks driven by ligand-receptor interactions in colorectal cancer but also highlights the significance of integrating multi-sample and patient data to unravel the molecular underpinnings of cancer communication pathways.

Introduction

Colorectal cancer (CRC) is the 3rd most common cancer and the 2nd cause of cancer death[1]. The discovery of the consensus molecular subtypes (CMS) among CRC tumors reflects the heterogeneity of composition and functional states of cells in the tumor microenvironments (TME), which in turn may underlie diverse behaviors of disease progression and responses to anticancer treatments by tumors. For example, except for those with deficient mismatch repair (dMMR), most CRCs do not respond to anti-PD-1 immune therapy [2]. Such relationship between genomic pattern and clinical response is the result of cell-cell communications (CCC) within in the TME, through which the downstream signaling pathways are mediated and regulated. Therefore, the comprehensive understanding of the cell states and communication network in the TME are vital for uncovering the heterogeneous mechanisms of tumor growth, immune evasion and treatment resistance, which are essential for precision oncology.

Modeling CCC requires inferring the cellular states of individual cells, identifying synchronous changes in the cell-state among various cell types that may arise due to CCC, and investigating the molecular mechanism underlying CCC[3-6]. Single-cell RNA sequencing (scRNA-seq) is a powerful technology for profiling the transcriptomes of individual cells. As gene expressions are tightly regulated in a cell, the co-expression patterns of genes serve as proxies of the states of signaling pathways in a cell, which can be discovered by deconvoluting scRNA-seq data. Furthermore, the emergence of spatial transcriptomic technology provides additional spatial relationships between cells within their native spatial context [7,8]. Thus, integrating scRNA-seq with spatial transcriptome provides the spatial context for interactions among different cell populations [9] and enhances the accuracy of modeling CCC.

63

64 Single-cell transcriptomes can be used to investigate cellular states at different levels. At the whole
65 transcriptome level, cells sharing similar overall transcriptomic profiles can be discovered as
66 discrete cell clusters using state-of-the-art clustering techniques [10-13]. Although these approaches
67 are good at classifying different cell populations, it does not allow examination of transcriptomic
68 programs within individual cells that underpin the biological functions of these cells [14]. To discover
69 and represent transcriptomic programs within each cell, different statistical- and factorization-
70 based models, e.g., non-negative matrix factorization (NMF) and latent Dirichlet allocation (LDA),
71 have been used to deconvolute transcriptional programs by capturing the co-expression pattern of
72 certain cell subsets [15-18]. However, these models often have strong assumption of the independence
73 among gene programs. Moreover, these methodologies fail to capture the well-established
74 phenomenon that cells belonging to a lineage exhibit highly coordinated transcriptomic programs.
75 These programs are more accurately represented as a hierarchical structure, wherein transcriptomic
76 programs shared by numerous cells within a lineage are positioned close to the root of the hierarchy.
77 In contrast, transcriptomic programs underlying the differentiation of specialized cells are
78 positioned as leaves. The methods of capturing such hierarchical relationships of transcriptomic
79 programs remain relatively underdeveloped.

80

81 Transduction of signals across cells involves ligand-receptor interactions. Different methods have
82 been developed to find potential ligand-receptor-mediated intercellular communication, e.g.,
83 CellChat, CellPhoneDBv2.0, and Nichenet [4,19,20]. However, those methods are mainly based on
84 the expression and enrichment of ligand-receptor pairs in specific cell clusters but do not explicitly
85 model the impact of the signals transmitted by a ligand-receptor pair, e.g., genes regulated by PD-

86 L1/PD-1 signaling. Furthermore, few methods exist for modeling coordinated cellular states of
87 different cells within a tumor, e.g., certain subtype of myeloid cells and subtype of stromal cells
88 tend to co-exist in tumors, or inferring whether the changed state of one type of cell directly
89 influences the state of other types of cells.
90
91 To solve these challenges, we propose a novel framework for integrating bulk, single-cell, and
92 spatial RNA-seq data to depict a comprehensive landscape of CRC TMEs (**Fig. 1**). We employed
93 the nested Hierarchical Dirichlet process (nHDP) model to deconvolute a large collection of single-
94 cell transcriptomes from 8 cohorts with over 153 CRC patients and over 626,858 single-cells (**Fig.**
95 **1A, Extended Data Table 1 & Fig. 1**). nHDP organizes the discovered gene expression modules
96 (GEMs) in a hierarchy, enabling it to separate GEMs (transcriptomic programs) representing major
97 cell lineage markers from those reflecting the functional state of cells (**Fig. 1B-C**). We then
98 systematically investigated whether the expression of a GEM was regulated by ligand-receptor
99 signaling (**Fig. 1D**) using the hill function that is used in pharmacology to measure the binding
100 affinity between ligand and receptor. By studying the coordinated expression of GEMs, we
101 systematically investigated the statistically significant intercellular communications across cell
102 types (**Fig. 1E**) and discovered patterns of intercellular communication programs (GEM programs)
103 shared within subpopulations of tumors (**Fig. 1F**). Employing techniques of causal analyses, we
104 further searched for ligand-receptor pairs that contribute to mediate signal transduction of a CCC
105 and further constructed a CCC network among cells expressing coordinated GEMs (**Fig. 1G**). We
106 envision this comprehensive framework enables researchers to study complex mechanisms
107 underlying distinct TMEs, investigate different immune evasion mechanisms, discover potential
108 drug targets, and guide precision immune therapies.

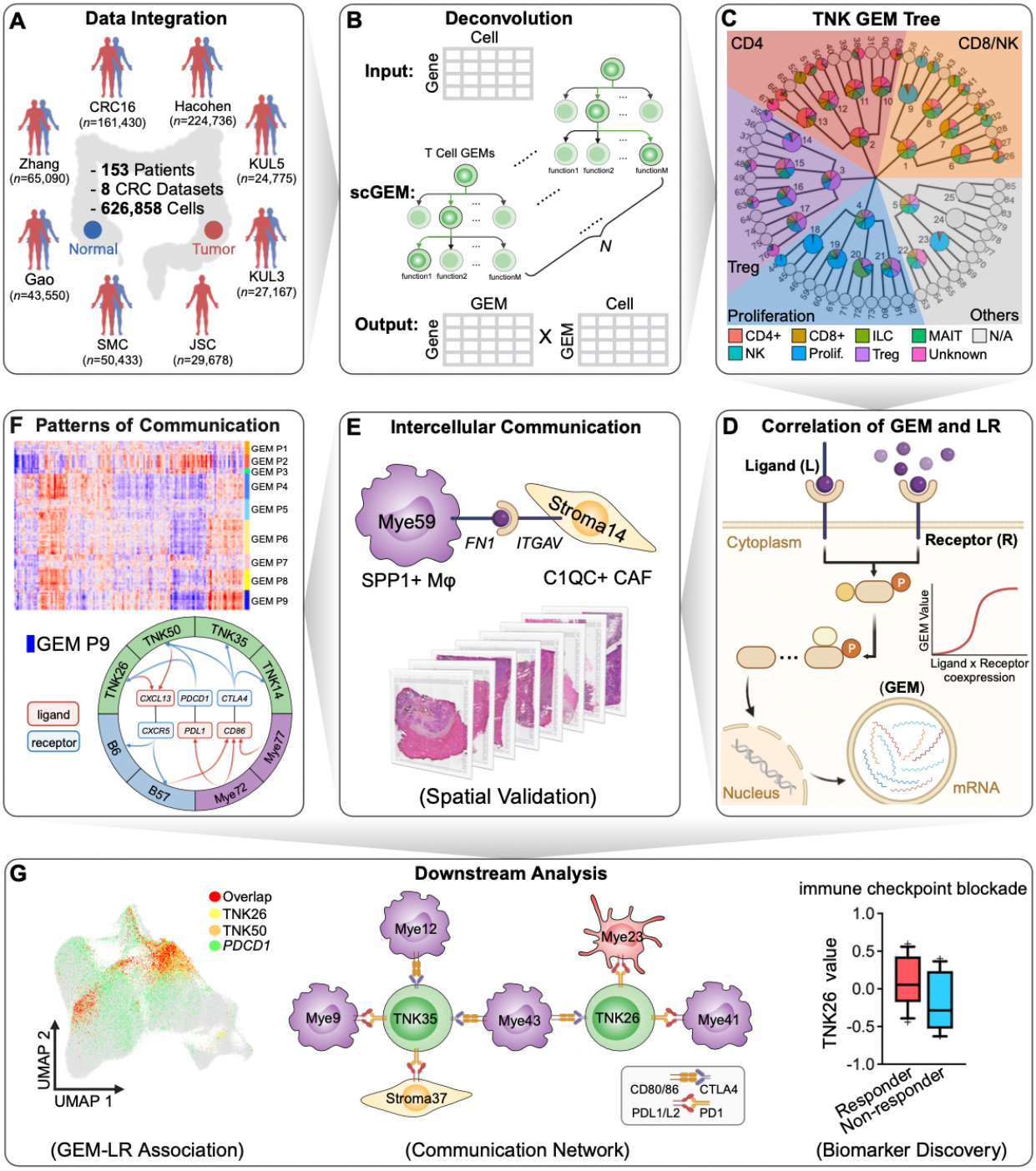


Fig. 1: General design of the study. **A.** Merged CRC single-cell RNAseq data from 8 cohorts. **B.** Diagram of the nHDP model using the expression of genes (a *cell-by-gene* matrix) in TNK cells as input. The transcriptome of one cell can be represented by a unique mixture of GEMs in the nHDP tree. The nHDP model outputs two matrices: a *gene-by-GEM* matrix reflecting the gene

114 composition of a GEM and a *cell-by-GEM* matrix representing the GEM compositions of each cell.
115 **C.** nHDP discovered the tree-structured organization of GEMs. Each node in the tree represents a
116 GEM. A pie chart represents the percentage of each cell type the GEM is expressed. **D.** Illustration
117 of modeling the relationship between ligand-receptor (LR) and its downstream GEMs. **E.**
118 Modeling cell-cell communication by combining single-cell and spatial transcriptomes. **F.**
119 Learning patterns of coordinated expression of GEMs across cell types to discover distinct GEM
120 programs and associated tumor microenvironments. **G.** Downstream discoveries enabled by
121 deconvoluting single-cell transcriptomes by nHDP: **1)** GEMs highly associated with the signal of
122 LR. **2)** Subnetwork of intercellular communication for GEMs and LRs of interest. **3)** GEM as
123 potential novel biomarker for therapies.

124 # Result

125 **A comprehensive atlas of GEMs expressed by cells in CRC tumor microenvironments.**

126 We collected single-cell RNA-seq data from 8 independent CRC cohorts. Then, we separated cells
127 into 5 major lineages: TNK: mainly including T and natural killing (NK) cells (N=246,560);
128 Myeloid: myeloid cells, including monocytes, macrophages, neutrophils, and dendritic cells
129 (N=94,663); PlasmaB: including B and plasma cells (N=96,637); Stroma: stromal cells, including
130 endothelial, fibroblast and myofibroblast cells (N=69,128); Epi: epithelial cells (N=119,870)
131 (**Extended Data Table 1**). The merged cells for each cell category were clustered and labeled
132 using Seurat pipeline and canonical marker genes. After the integration and batch removal, we
133 identified 18 clusters in TNK cells (**Fig. 2A),** 18 cell populations in both Myeloid cells and
134 PlasmaB cells and 17 cell groups in Stroma cells (**Extended Data Fig. 1**). We also showed that
135 the distributions of datasets are well maintained in all clusters in TNK cells (**Fig. 2B**), suggesting
136 the batch effects were regressed out; thus, the pooled TNK cells reflect the landscape of tumor-
137 infiltrating TNK cells in CRC.

138

139 To model the tree structure of the TNK gene expression modules (GEMs), we utilized the nested
140 hierarchical Dirichlet process (nHDP) model to identify a series of nested gene expression modules
141 (GEMs) [21], where a GEM is a probability distribution over the gene space that captures the
142 coordinated expression pattern of a set of genes in a subpopulation of cells. The nHDP model
143 organizes inferred GEMs in a hierarchical tree structure, and we set branching factors of 5, 4, and
144 3 at 1st, 2nd, and 3rd layers, respectively, leading to a total of 85 nodes to capture GEMs with
145 different levels of granularity and specificity. **Fig. 2C** shows the tree organization of the GEMs
146 learned using TNK cells as input. The nHDP model populated the nodes in the hierarchy through

147 a data-driven approach. Each node in the tree represents a GEM. The ones close to the root (TNK
148 GEMs 1-5 in **Fig. 2C**) represent GEMs that are more broadly expressed in all T, innate lymphoid
149 cells (ILC), mucosal-associated invariant T cells (MAIT) and NK cells, which can be used to
150 define major lineage of T and NK (TNK) cells. The ones close to the leaves represent GEMs
151 specifically expressed in more differentiated cells. In the TNK GEM tree, only 43 TNK GEMs
152 were identified as biologically meaningful and represented as filled nodes. The nodes identified as
153 biologically meaningless were shown as empty nodes without color.
154
155 We systematically inspected the TNK GEMs reflecting TNK transcriptomic processes. The pie
156 chart within each GEM illustrates the proportion of cells expressing that specific GEM across
157 various T cell subtypes, including $CD8^+$ T cells, $CD4^+$ T cells, ILC, MAIT, NK cells, regulatory
158 T cells (Treg) and Proliferation T cells. It is interesting to note that major branches of the nHDP
159 hierarchy tree, as defined by GEMs 1-4, represent the major TNK lineages, including CD8/NK,
160 CD4, Treg, and proliferation T cells (**Fig. 2C**). Projecting the expression of GEMs onto the UMAP
161 of TNK cells (**Fig. 2D**), we noticed that certain GEMs were concentrated in a particular cell cluster,
162 representing a characteristic transcriptomic program active in the cell cluster. For example, TNK50
163 is concentrated in cells belonging to cluster #4, which consists of T helper (Th) cells expressing
164 CXCL13; TNK35 is mainly expressed in T cell cluster #6, a subset of T_{reg}. Conversely, there are
165 also GEMs shared across multiple cell clusters, suggesting a transcriptomic program that is shared
166 across multiple subtypes of T cells. For example, TNK26 is expressed in a subset of
167 $CXCL13^+CD4^+$ T cells (cluster #4), $MKI67^+$ proliferative T cells (cluster #10), and
168 $CXCL13^+CD8^+$ T cells (cluster #11). TNK42 is expressed in MAIT cells (cluster #9) and ILC
169 (cluster #17). Finally, TNK12 is expressed broadly among Treg, Th, and proliferating T cells.

170 These results indicate that our model can capture transcriptomic programs that are both shared and
171 specialized by cells belonging to different lineages. To test the performance of our model, we also
172 compared the distributions of the transcriptomic programs that were found by our model and NMF
173 [22]. Our result demonstrates that the nHDP excels in capturing more specific cellular programs
174 shared across multiple cell clusters when compared to NMF as shown in **Extended Data Fig. 2**.
175 For instance, TNK15 captures a cellular program of greater specificity, shared among cells from
176 4 TNK cell clusters, surpassing the performance of NMF5. NMF5 captures a more generalized
177 cellular program that combines programs represented by TNK15, TNK42, and TNK51. Similarly,
178 NMF58 captures the cellular programs combining programs represented by TNK17, TNK29, and
179 TNK50.
180
181 To make the GEMs more biologically explainable, we depict the distinctive gene composition
182 profile of each GEM (top 20) through a word cloud, and the top three associated transcription
183 factors identified by transcription factor analysis using ChEA3 [23] were illustrated as black dots in
184 **Fig. 2D**. It is interesting to note that TNK26 is enriched with genes characteristic of exhausted T
185 cells including PDCD1, CXCL13 and TIGIT. Based on the transcription factor analysis, NFATC2
186 regulated the top 50 genes associated with TNK26. Several studies have shown that the
187 transcription factor NFAT is involved in the exhaustion process of activated $CD8^+$ T cells [24]. In
188 addition, the top genes associated with TNK26 and TNK50 are both regulated by transcription
189 factor ZNF831, which has been shown to inhibit the functional activity of effector cells through
190 STAT pathway [25]. This indicates the potential existence of a signal leading to the regulation of T
191 cell exhaustion. We further examined the distribution of GEMs in different TNK clusters (**Fig.
192 2E**), unveiling which GEMs are broadly shared among cells assigned to multiple clusters and

193 which GEMs are specific to a particular cell cluster. **Extended Data Fig.3** displays the detailed
194 comparison between GEMs and TNK cell clusters by measuring the correlation between top 50
195 genes of a GEM and highly variable genes in each TNK cell subtype. Furthermore, we performed
196 pathway enrichment analyses in Reactome to investigate the functions of top genes associated with
197 GEMs (**Fig. 2F**). Interestingly, several TNK GEMs, including TNK6, TNK26, TNK35, and
198 TNK41, are associated with PD-1 signaling pathway. The dot plots showing the distribution of
199 GEMs in traditional cell clusters and the enriched signaling pathways for other cell types (PlasmaB,
200 Myeloid, Stromal) are listed in **Extended Data Fig. 4 & 5**. The detailed annotation of each GEM,
201 including the UMAP, cell composition, and top associated genes, is listed in our interactive website.
202 In summary, by deconvoluting the transcriptomes of a large number of single cells using nHDP,
203 our approach revealed GEMs that represent transcriptomic programs within or across traditional
204 cell subtypes, which reflect certain signaling events within these cells.

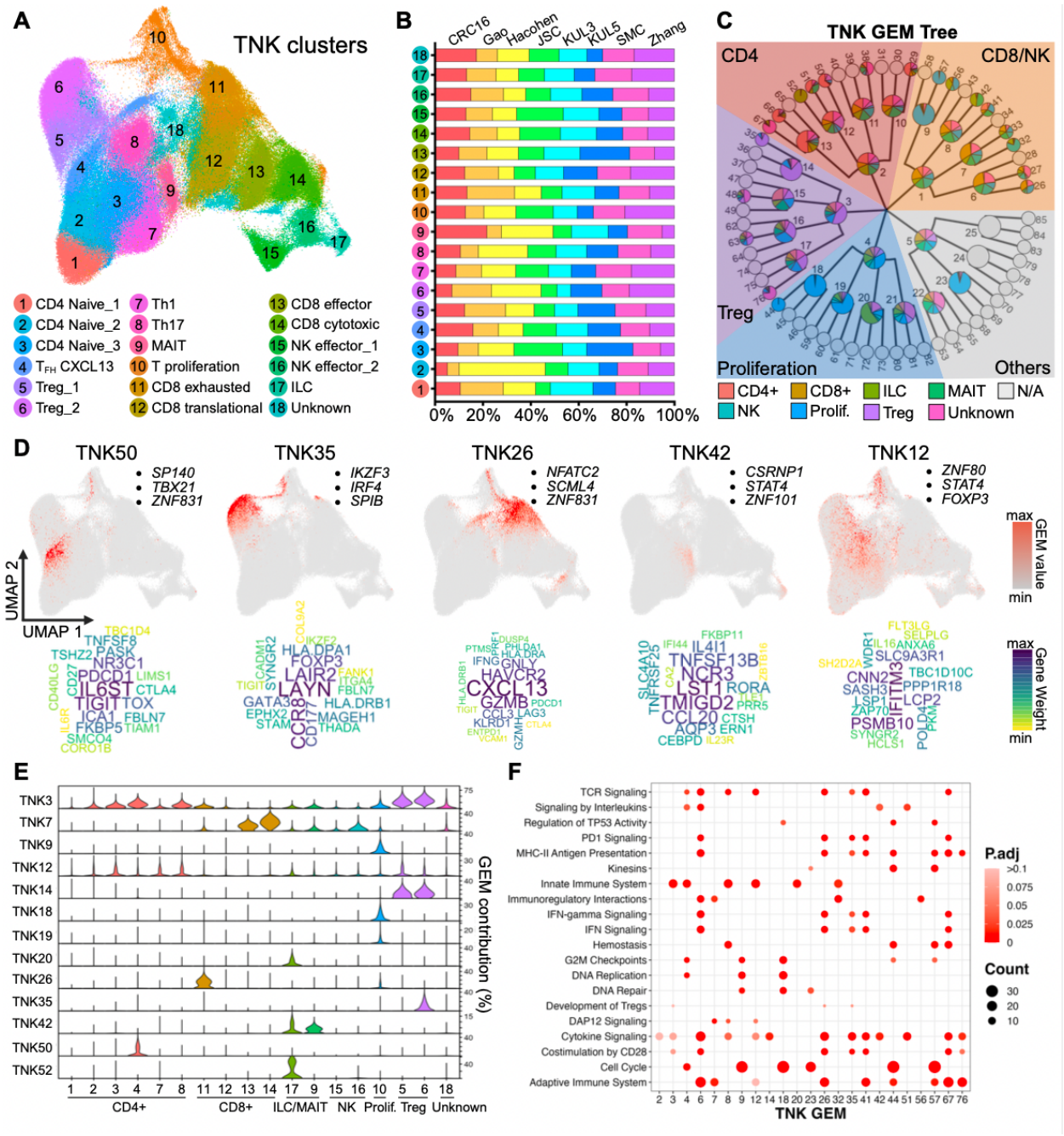


Fig. 2: Deconvoluting single-cell transcriptomics to discover GEMs reflecting the state of signaling pathways. **A.** UMAP plot of 246,560 merged TNK cells from 8 CRC cohorts. **B.** The proportion plot reflects the origins of cells in each cell cluster. **C.** The hierarchical tree structure of TNK GEMs was learned from the nHDP model, which was trained with merged TNK cells. The nHDP model has three levels. The first level contains GEMs from 1-5. The second level

211 contains GEMs from 6-25. The third level contains GEMs from 26-85. Each color represents a
212 subtype of T cell. The unused GEMs are colored grey. **D.** The UMAP plot, word cloud plot, and
213 top associated TFs of selected TNK GEMs. **E.** The functional involvement of TNK GEMs within
214 each cell cluster. **F.** The enriched signaling pathways associated with TNK GEMs.
215
216 **Discovering potential regulatory relationships between ligand-receptor signaling and GEMs**
217 We set out to examine if the expression of a GEM in cells is potentially regulated by certain ligand-
218 receptor signaling by testing the hypothesis posited in **Fig. 3A**: If a ligand-receptor pair regulates
219 the expression of the member genes of a GEM, the overall expression value of the GEM can be
220 modeled with a sigmoid dose-response model with respect to ligand-receptor pair. We modeled
221 the relationship of the expression of ligand, receptor, and GEM at the tumor sample level. We first
222 generated pseudo-bulk RNAseq (a *tumor-by-GEM* matrix) from the scRNAseq (a *cell-by-GEM*
223 matrix) for tumor samples. Then we estimated the concentration of ligand and receptor proteins
224 based on the pseudo-bulk expression (see Methods) of the ligand ($[L]$) and receptor ($[R]$) for each
225 of the 151 tumors. Additionally, we approximated the concentrate of protein complex as the
226 product of ligand and receptor: $[LR] \propto [L]*[R]^3$. Pseudo-bulk expression of GEMs was estimated
227 using the GSVA analysis. Subsequently, we conducted a systematic search against all GEMs with
228 known LR pairs from NicheNet database and retained the LR pairs that could predict the
229 expression of the GEM with statistical significance (**Extended Data Table 2**).
230
231 We investigated the GEMs that are strongly associated with immune checkpoints programmed cell
232 death protein 1 (PD-1, encoded by *PDCD1*) and cytotoxic T lymphocyte-associated protein 4
233 (*CTLA4*) signaling (**Fig. 3B**). For the PD-1 signaling, the top two TNK GEMs that could be

234 predicted by approximated [*LR*] with the strongest non-linear correlation were TNK26 and TNK50,
235 with Pearson's correlation coefficients between prediction and observation being $\rho = 0.782$, $p =$
236 7.2831e-9 and $\rho = 0.685$, $p = 6.8891e\text{-}07$, respectively. The UMAPs show that the distributions of
237 *PDCD1* and the two GEMs are highly correlated at the single-cell level (**Fig. 3B-C**). The results
238 also suggest potential PD-L1/PD-1 signaling is associated with distinct GEMs in different T cell
239 subtypes. Specifically, [*CD274*]*[*PDCD1*] is associated with TNK50 in $CXCL13^{+}CD4^{+}$ T cells,
240 whereas [*CD274*]*[*PDCD1*] is associated TNK26 mainly in exhausted $CD8^{+}$ T cells. Importantly,
241 the approximated ligand-receptor complex [*LR*] is more predictive of the expression of these two
242 GEMs when compared with only using the receptor expression [R] alone. Similarly, the results
243 show that the expression values of TNK35 can be predicted by approximated *CD86* and *CTLA4*
244 complex, and the expression of *CTLA4* is highly correlated with the expression of TNK35 in Treg
245 cells (**Fig. 3B-C**), suggesting the formation of this complex at these cells is associated with
246 expression of the GEM in a non-linear dose-response fashion. In summary, our results show that
247 identifying GEMs enabled the investigation of the functional impact of potential ligand-receptor
248 signaling. We visualized a subset of ligand-receptor pairs and TNK GEMs that exhibited
249 statistically significant nonlinear dose-response relationships (**Fig. 3D, Extended Data Table 2**).
250
251 The strong non-linear dose-response relationship between a ligand-receptor pair and a GEM (e.g.,
252 *CD274-PDCD1* and TNK26) suggests the hypothesis that the expression of some members of a
253 GEM may be regulated by the ligand-receptor interaction. Thus, PD-1 signaling pathway may be
254 more active in tumors expressing a high level of TNK26, and thereby, these tumors might be more
255 susceptive to anti-PD-1 treatment. To test the association of the expression level of TNK26 and
256 response to anti-PD-1 related treatment. We collected the clinical and transcriptomic data of CRC

257 patients from the CAMILLA trial [26]. This study involves metastatic CRC patients with proficient
258 microsatellite stable (MSS) tumors. The patients received treatment with the combination of
259 carbozantinib (a small molecule inhibitor of tyrosine kinases including VEGFR2) and durvalumab
260 (an immune checkpoint inhibitor targeting PD-L1). The transcriptome and clinical responses to
261 the combined treatment were measured. We separated the patients into responder and non-
262 responder groups, according to designation by the studies, and we compared the express level of
263 GEMs learned from nHDP model in these two groups. The analyses indicate a significant elevation
264 in TNK26 expression in the responder group compared to the non-responder group (**Fig. 3E**). This
265 association was further tested across three established melanoma studies [27] involving patients
266 treated with anti-PD-1 antibodies. Consistently, TNK26 expression is significantly higher in the
267 responder group compared to the non-responder group (**Extended Data Fig. 6C**). The results
268 support the notion that the expression level of TNK26 reflects the involvement of the PD-1
269 signaling in the tumors, and thus can potentially be used as a novel biomarker for responsiveness
270 to the anti-PD-1 therapy. Regarding TNK50, a significant difference is observed exclusively in the
271 melanoma dataset. The TNK50 expression is significantly higher in responder group than in non-
272 responder group. However, in CAMILLA trial, even though TNK50 is shown to be correlated with
273 TNK26, TNK50 did not reach statistical significance, likely due to the small number of responders
274 with one apparent outlier, as depicted in **Extended Data Fig.6B**. More examples of GEMs
275 showing significant difference between responder group and non-responder groups in CAMILLA
276 trial and melanoma cohorts are listed in **Extended Data Fig. 6A & 6C**, respectively.
277
278

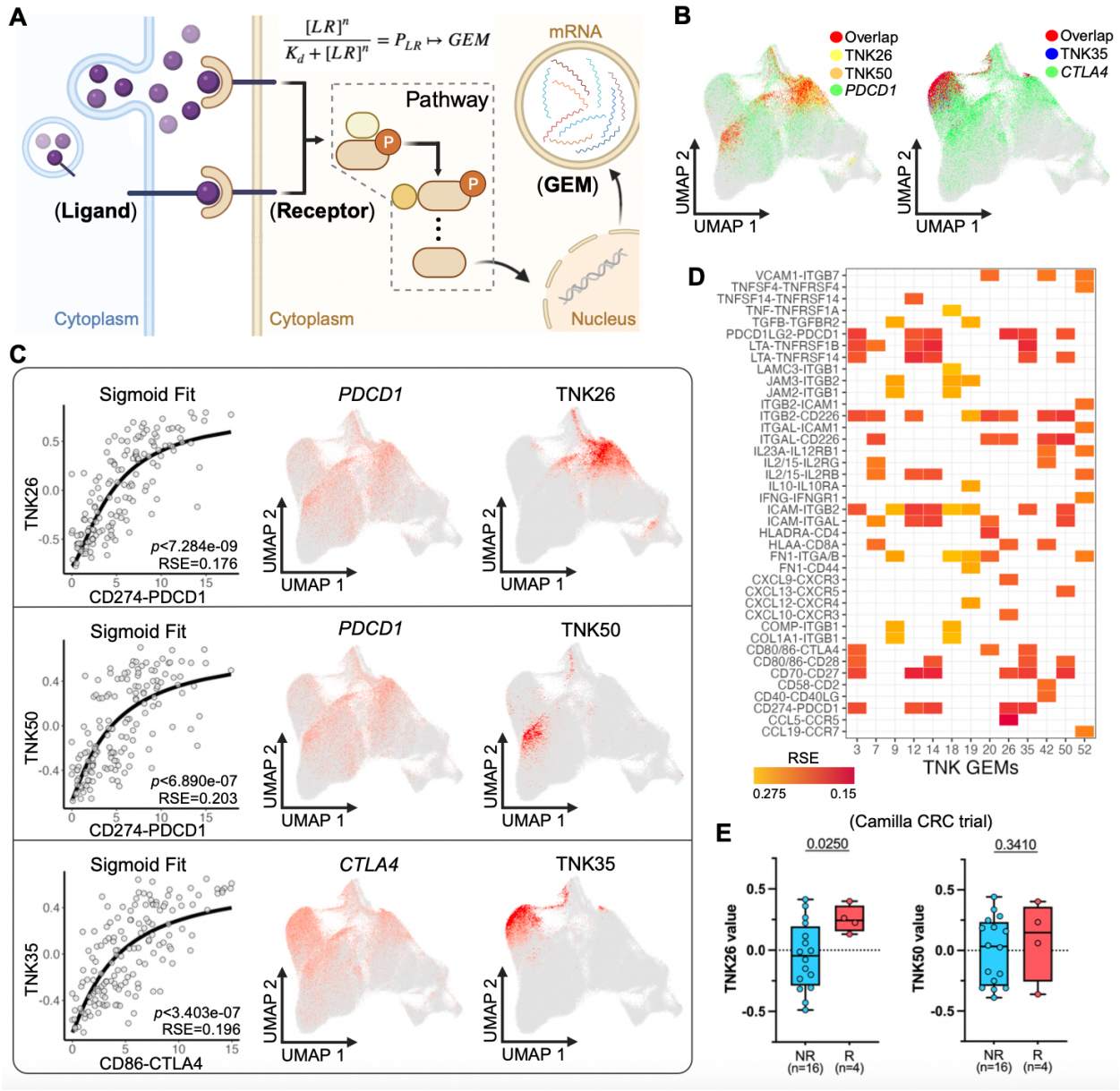


Fig. 3: Discovering ligand-receptors strongly associated with specific pathways regulating the expression of GEMs. A. Illustration of GEMs reflective of downstream transcriptome mediated by biochemical signaling through ligand-receptor binding. The significant association between a GEM and a ligand-receptor needs to satisfy the sigmoid fitting. **B.** The overlap of cells expressing receptor and GEMs of interest. Left: the overlap of *PDCD1*, TNK26 and TNK50. Right: the overlap of *CTLA4* and TNK35. **C.** From left to right, the plot of sigmoid fitting between GEM

286 and ligand-receptor pair with *p*-value and residual standard error (RSE), UMAP plot of the receptor
287 gene, and UMAP plot of the GEM. **D.** The summarization of top ligand-receptor pairs significantly
288 associated with TNK GEMs. **E.** Comparison of the expression level of TNK26 and TNK50
289 between the responder group and non-responder group treated with Cabozantinib plus durvalumab.
290 Carbozantinib is a small molecule inhibitor of tyrosine kinases including VEGFR2 and
291 durvalumab is an immune checkpoint inhibitor of PD-L1. The p-value was calculated using one-
292 sided Wilcoxon test.
293

294 **CCC between cancer-associated fibroblast and SPP1-macrophage through ligand-receptor**
295 **pairs *FN1-ITGAV* and *SPP1-ITGAV***

296 The significance of crosstalk between macrophages and cancer-associated fibroblasts (CAFs) has
297 garnered attention in recent studies [28-30]. Motivated by this, we embarked on an investigation into
298 this crucial aspect. After searching for highly correlated GEMs that are expressed in stromal cells
299 and myeloid cells, our analysis led to a pair of GEMs, Mye59 and Stroma14, which exhibited the
300 strongest correlation (Pearson ρ=0.78) at the pseudo-bulk RNA level. A further search of myeloid
301 GEMs correlated to Stroma14 led to the additional discovery of Mye43 (Pearson ρ=0.58) (**Fig. 4A**
302 top row). The results led to a hypothesis that myeloid cells expressing Mye59 or Mye43 may
303 communicate with stromal cells expressing Stroma14. Based on the annotation of GEM
304 distribution, Mye59 and Mye43 are expressed in a separate subpopulation of M2 macrophage, and
305 Stroma14 is expressed in a subpopulation of $FAP^{+}POSTN^{+}$ fibroblast cells defined as matrix CAF
306 (mCAF)[31]. The search for the *LR* potentially mediating such communication led to *FN1-ITGAV*
307 and *SPP1-ITGAV* pairs, which predict the expression of Stroma14 (**Fig. 4A** bottom row) in a
308 nonlinear dose-response fashion. Examining the distribution of the ligands and receptor genes and

309 the above GEMs in UMAPs showed that *FN1* and Mye59 were co-expressed in a subpopulation
310 of macrophage cells, and *SPP1* and Mye43 were co-expressed in another distinct macrophage
311 subpopulation (**Fig. 4B**). While *ITGAV* (encoding a unit of integrin) was widely expressed in
312 stromal cells, Stroma14 was expressed in a subpopulation of mCAF, suggesting that integrin
313 signaling may have different downstream effects in distinct subpopulations of stromal cells (**Fig.**
314 **4B**). The above results led to the hypothesis: a subpopulation of *ITGAV*-expressing mCAF in CRC
315 may interact with distinct subpopulations of macrophage cells that express cells *FN1* and *SPP1*,
316 respectively, and such interactions induce the expression of the Stroma14 in CAFs.
317
318 We tested the above hypothesis using spatial transcriptomic data of metastatic CRC patients from
319 Pelka et al [22], by studying the co-localization of the above ligands, receptors, and GEMs in the
320 spatial domain and modeling their interactions in spatial domains. In a CRC liver metastatic tumor
321 section, the spatial expression of ligand (*FN1*, *SPP1*) and receptors (*ITGAV*) and GEMs (Mye59,
322 Mye43, and Stroma14) were highly correlated (**Fig. 4C**), indicating that cells expressing these
323 genes and GEMs likely interact in spatial domains. There is a notable difference in the spatial
324 distributions of Mye59 and Mye43. Particularly, myeloid cells expressing Mye59 concentrate at
325 tumor-associated fibroblast-rich tissue and the tumor leading edge, interfacing with normal liver
326 tissue, whereas mey43 is mainly expressed in the tumor core. Examining the distribution of *ITGAV*
327 and Stroma14 showed that while *ITGAV* was expressed in both tumor and adjacent normal liver
328 tissues (indicating the presence of stromal cells), Stroma14 is mainly expressed in mCAF in the
329 tumor area. This indicates that the two CCC channels (Mye59-Stroma14 and Mye43-Stroma14)
330 may play different signaling roles.
331

332 The spatial distribution of the approximated *FN1- ITGAV* complex agrees well with the spatial
333 distribution of Mye59 and Stroma14, and their distribution is distinguishable from that of *SPP1-*
334 *ITGAV* and Mye43 (**Fig. 4C**). We further modeled the relationship between approximated [*LR*]
335 and expression of Stroma14 in the spatial domains. The approximated *FN1-ITGAV* complex
336 concentration showed a clear sigmoid dose-response curve ($\rho = 0.6$, p = 1.4429e-14), and the
337 association between *SPP1-ITGAV* complex and Stroma14 is also statistically significant ($\rho =$
338 $0.315, p = 0.0007$). The sigmoid fittings between Stroma14 and two ligand-receptor pairs in the
339 spatial domain are shown in **Fig. 4D**. The above results led to a hypothetic model of the TME
340 involving the above ligands, receptors, and GEMs (**Fig. 4E**). We hypothesize that myeloid cells
341 expressing Mye59 (macrophage I) may align with the tumor leading edge as the tumor invades,
342 whereas myeloid cells expressing Mye43 (macrophage II) are more concentrated within the tumor
343 core. The interaction between these macrophage cells and mCAF cells leads to the expression of
344 Stroma14, which is characteristic of mCAF but absent from stromal cells in normal liver tissue.
345 By deconvoluting cellular signaling activities as GEMs, our approach uncovers the signaling
346 mechanisms through which various SPP1+ macrophage subpopulations interact with different
347 CAF subsets in distinct spatial domains. The example discussed above demonstrates our ability to
348 identify potential CCC candidates transmitted by the ligand-receptor signal, supported by both
349 scRNA-seq and spatial transcriptomic data.

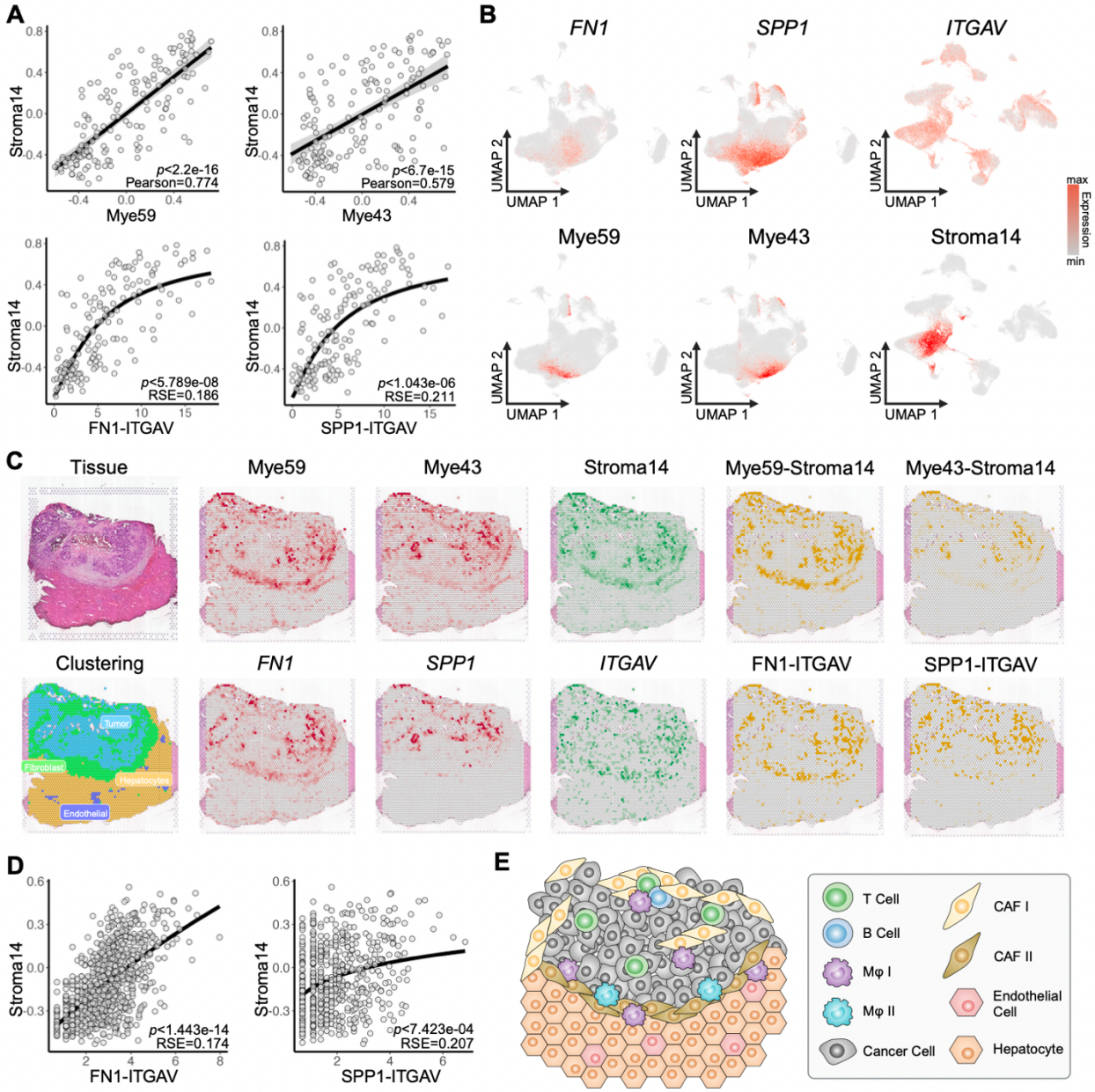


Fig. 4: Modeling intercellular communication. A. Up panels: The scatter plot of Stroma14-vs-Mye43 and Stroma14-vs-Mye59 using the pseudo-bulk transcriptome data. Lower panels: the sigmoid fitting between Stroma14 and two ligand-receptor pairs (*FN1-ITGAV* and *SPP1-ITGAV*) using the pseudo-bulk transcriptome data. **B.** The UMAP plot of cells expressing *FN1, SPP1, ITGAV*, Mye59, Mye43, and Stroma14, respectively. **C.** The spatial distribution plot of HE, tissue clustering, GEMs, ligands, receptors, and ligand-receptor pairs. **D.** The sigmoid fitting between

357 Stroma14 and two ligand-receptor pairs (*FN1-ITGAV* and *SPP1-ITGAV*) using spatial
358 transcriptome data. **E.** Illustration of a hypothetic spatial model for the tumor microenvironment
359 of the spatial sample.
360
361 **Discovering coordinated CCC programs**
362 We further investigated patterns of coordinated expression of GEMs across different cell types
363 (TNK, B, myeloid, stromal, and epithelial) in CRCs. We represented each tumor in the GEM space
364 (pseudo-bulk expression: *cell-by-GEM* matrix) and performed double clustering of GEMs and
365 tumors. The results (**Fig. 5A**) showed that the expression of many GEMs was highly coordinated.
366 We hypothesized that the coordinated expression of GEMs across multiple types of cells in a TME
367 is likely due to two possible mechanisms: environmental factors that orchestrate the expression of
368 GEMs (confounding) or a CCC network that allows cells to communicate and reach homeostasis.
369 We refer to a set of GEMs across multiple cell categories that displays a highly coordinated
370 expression pattern as a CCC program, and we have identified 9 CCC programs (rows). Based on
371 their GEM composition, clustering tumors revealed 6 sample clusters (columns). A complete list
372 of CCC programs is in **Extended Data Table 4**.
373
374 We further investigate the potential molecular basis of CCC programs. CCC Program 2 consists
375 of 16 GEMs from different cell categories (**Fig. 5B**). The pathway analysis indicates that GEMs
376 in CCC Program 2 are significantly enriched of genes involved in cell cycle pathways, such as the
377 C-MYC signaling pathway, a hallmark pathway of cell proliferation in different cell types [32]. The
378 GEMs belonging to this CCC program include GEMs from TNK, B, myeloid, epithelial and
379 stromal cells, respectively (**Fig. 5B**), suggesting they are subpopulations of cells in a proliferation

380 stage from each cell category. Of interest, we noted that TNK9 is enriched in the T-cell cluster
381 #10, which is labeled as proliferative T cells. Not surprisingly, an epithelial GEM belonging to
382 this program, Epi9, is broadly active in epithelial cells, reflecting a shared transcriptomic program
383 involved in cell proliferation by cancer cells. The UMAPs in **Fig. 5C** indicate that various cell
384 categories expressing these GEMs also express the proliferation marker gene *MKI67*. We
385 hypothesized that the expression of GEMs of proliferation nature in different cell types may be
386 associated with signals that promote proliferation of cells, e.g., receptor tyrosine kinase (RTK)
387 pathways, Wnt/β-catenin pathway, and PI3K-AKT-mTOR pathway. Our pathway enrichment
388 analysis shows that the top genes associated with most of GEMs in CCC Program 2 are enriched
389 in the MYC pathway, which is downstream of RTKs. To test the hypothesis, we investigated the
390 spatial distribution of the GEMs in tumors. The results (**Fig. 5D**) show the co-localization of GEMs
391 in the CCC Program 2. We calculated the spatial correlations between MKI67 with respect to B21,
392 Stroma68, and Epi23, using their expression values in spatial spots. The correlation coefficients
393 are 0.708, 0.762, and 0.804, respectively, providing evidence for the strong coordinated expression
394 of these GEMs in spatial domains. We further checked the cell composition of spatial spots
395 expressing these GEMs. The inference of cell composition in tumor spots (**Extended Data Fig.
396 7A**) show that those spots contain a mixture of cells potentially expressing these GEMs. The
397 compositions of B21, Stroma68 and Epi23, represented by their top associated genes, are not
398 overlapping (**Extended Data Fig. 7B**). Each of the three GEMs captures cell-type-specific
399 proliferation signals. The indicates that the diverse cells expressing GEMs in GEM program 2 may
400 be stimulated by proliferation-related factors existing in tissue microenvironments.
401 Moreover, we performed deconvolution analysis on the GEMs in CCC Program 2 utilizing the
402 single-cell RNA-seq transcriptome data obtained from the peripheral blood sample prior to the

403 treatment of the regorafenib and nivolumab combination therapy[33]. Our analysis revealed that the
404 expression levels of GEMs, including TNK9, B18, and Mye21, in the proliferation program
405 (CCC program 2), exhibited significant differences between responders and non-responders, as
406 illustrated in Extended Data Fig.8. This observation aligns with the findings reported in the
407 original paper and highlights the potential of cellular signals related to MKI67 proliferation as
408 indicative markers for predicting clinical outcomes in patients treated with a combination of
409 multi-kinase inhibitors and immune checkpoint inhibitors.
410 We also searched for CCC programs in which coordinated expression of GEMs is likely due to
411 CCC between the GEMs. As an example, CCC Program 9 includes TNK26, TNK35, TNK50,
412 Mye72, Mye77, B57, and B6. To investigate whether cells expressing these GEMs communicate
413 with each other, we searched for LR pairs that were co-expressed with these GEMs, e.g., (*GEM X*
414 and *L*) and (*R* and *GEM Y*), as a candidate pair of communicating cells, using the pseudo-bulk
415 RNA-seq data. We found three main LR pairs through which GEMs in the program may
416 communicate, including *CD274-PDCD1*, *CD86-CTLA4*, and *CXCL13-CXCR5*, which are closely
417 associated with T-cell immune checkpoint molecules in the tumor microenvironment. We illustrate
418 the possible communication channels among the cells expressing those GEMs using directed edges
419 based on the co-expression of *GEM X-L* and *R-GEM Y* (**Fig. 5E**). The expression level of the
420 ligands and receptors in the cells expressing GEMs in the CCC Program 9 is shown in **Fig. 5F**.
421
422 The cells expressing TNK GMEs in CCC Program 9, such as TNK26 and TNK35, mainly express
423 receptor genes *PDCD1* and *CTLA4*. GEM B57 is concentrated in IgA/G cells expressing the
424 CXCR5, which is the receptor of CXCL13. Two myeloid GEMs, including Mye72 and Mye77,
425 concentrated in dendritic cells, express ligand genes *CD274*, *CD80*, and *CD86*. We then

426 investigated whether these ligands, receptors, and GEMs are colocalized in spatial domains as
427 evidence of possible CCC among the cells expressing those GEMs. Indeed, the spatial
428 transcriptomic data show that *GEM X-LR-GEM Y* triplets of inferred CCC channels, e.g., Mye72—
429 *CD274/PDCD1*—TNK50 (Pearson correlations of Mye72-*CD274* and *PDCD1*-TNK50 are 0.059
430 and 0.545 respectively) and TNK41—*CXCL13/CXCR5*—B6 (Pearson correlations of TNK41-
431 *CXCL13* and CXCR5-B6 are 0.574 and 0.628 respectively), tend to be colocalized in spatial
432 domains (**Fig. 5G**). The low correlation between *CD274* and Mye72 is due to the sparse
433 distribution of *CD274* with diffusion in the spatial domain. These results support the hypothesis
434 that CCC between cells in the spatial domain underlies the observed coordinated expression of
435 GEMs, i.e., CCC programs.
436
437 In summary, the findings presented in this section collectively illustrate the feasibility of
438 uncovering a unique CCC mechanism within a micro-TME. This is achieved by modeling the
439 synchronized expression of GEMs, identifying potential LRs responsible for mediating
440 communication, and providing evidence of spatial colocalization for the components involved in
441 the identified CCC channel.

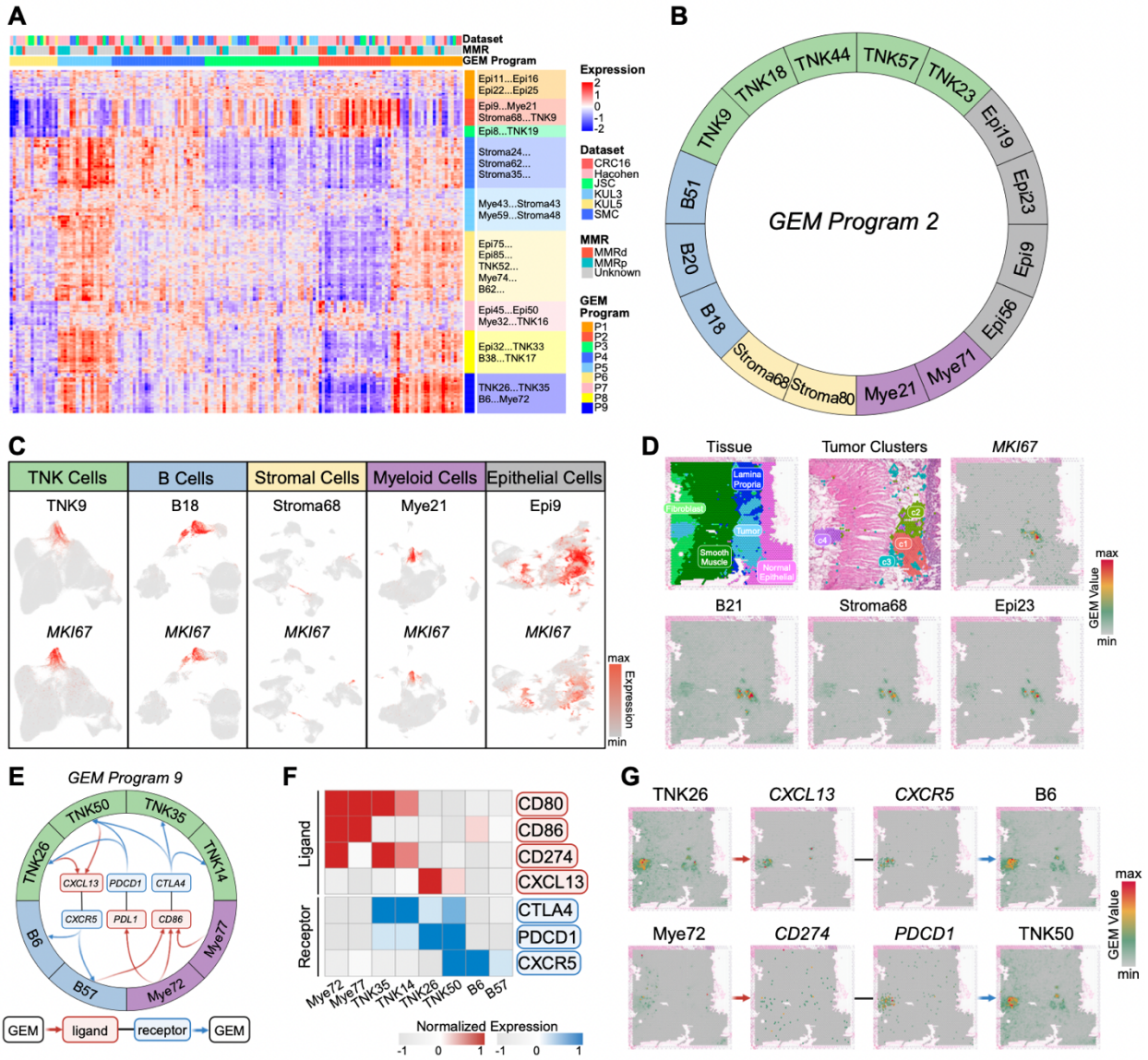


Fig. 5: Discovering patterns of intercellular communication channels that define transcriptome programs across different cell types. A. Clustering of GEMs and CRC tumors. Pseudo-bulk profile for each tumor was constructed from scRNA-seq data. Expression values of a GEM were estimated based on the GSVA using the top 50 weighted genes in a GEM. The GEMs (rows) are grouped into 9 CCC programs by clustering analysis, representing patterns of intercellular communication channels. The tumors (columns) are grouped into 6 clusters. **B.** The GEM members of the CCC Program 2. The majority of genes associated with these GEMs are

450 MYC pathway targets, as indicated by the HALLMARK_MYC_TARGETS in GSEA. **C.** UMAP
451 of MKI67 and GEMs in CCC program 2 in different cell-type categories. **D.** Spatial colocalization
452 of proliferation marker MKI67 and GEMs in CCC Program 2. **E.** The potential CCC network and
453 the ligand-receptors involved in GEMs in CCC Program 9. The red and blue rectangle in the center
454 represents ligands and receptors respectively. The edge between a ligand/receptor and a GEM is
455 added if the ligand/receptor is expressed in the cells expressing a GEM. **F.** The expression level of
456 receptors (blue) and its associated ligands (red) in cells expressing the GEMs in CCC program 9.
457 TNK and B GEMs in CCC Program 9 mainly express receptors (*PDCD1, CTLA4, and CXCR5*).
458 Myeloid GEMs mainly express ligands (*CD80, CD86 and CD274*). **G.** The spatial colocalization
459 of ligand, receptor and GEMs using Mye73 -> *CD274-PDCD1*-> TNK50 and TNK26 -> *CXCL13-*
460 *CXCR5* -> B6 as examples.
461
462 **Construction of cell-cell communication subnetworks connected by ligand-receptor signaling**
463 Discovering a pair of highly correlated GEMs across cell types suggests possible CCC between
464 the cells expressing the GEMs. Subsequently, we conducted a systematic search and examination
465 of potential CCC channels and networks. Using CRC pseudo-bulk RNA-seq data, we first searched
466 for highly correlated GEMs (*GEM X* and *GEM Y*) that are expressed in different cell types (or
467 clusters) to discover potential CCCs. In addition, we performed a causal analysis to find potential
468 ligands and receptors that mediate CCCs. To that end, we then search for *LR* pairs such that *L* is
469 co-expressed with *GEM X* and *R* is co-expressed with *GEM Y.* Finally, we assess whether
470 conditioning on LR, the two marginally correlated GEMs become independent, which is denoted
471 as $GEM\ X \perp\!\!\!\perp GEM\ Y \mid LR$. Conditional independence is an important concept in causal analysis,
472 and in our setting, if a pair of highly correlated GEMs becomes independent when

473 conditioning on a LR pair, this LR pair is potentially involved in transmitting signals between
474 cells expressing *GEM X* and *GEM Y*. Finding a *GEM X-LR-GEM Y* triplet leads to the hypothesis
475 depicted in **Fig. 6A**. More specifically, a change in the state of the signaling pathway (*X*) in one
476 cell promotes the release of the ligand and, at the same time, leads to transcriptomic changes
477 represented by *GEM X*, which serves as a proxy of the pathway (*X*) state. The binding of the ligand
478 with its receptor in another cell results in a cellular effect represented by the signaling
479 pathway (*Y),* which leads to the changed gene transcription represented by *GEM Y*. Further
480 extending such CCC analysis to multiple GEMs and cell types leads to a CCC network (CCCN),
481 which may be used to understand how different cells reach homeostatic states in a TME through
482 CCCs. If there are GEMs/LR pairs of interest, a CCC sub-network can be constructed based on
483 the *GEM X-LR-GEM Y* triplets. We identified and illustrated a CCC sub-network related to
484 immune checkpoint signaling in **Fig. 6B**. The network consists of T cells expressing TNK GEMs
485 (TNK26, TNK35 and TNK50) and non-TNK cells expressing GEMs (Stroma37, B44, Mye21,
486 Mye43, and Mye46), likely transmitted through PD-1, CTLA4 and CXCL13 signaling. In this
487 network, the communication between cells expressing Mye9 (a combination of dendritic and
488 macrophage cells) and TNK26 (PDCD1^{+}CXCL13^{+} CD4^{+} and CD8^{+} T cells) can be mediated
489 through *CD274-PDCD1*; the communication between the cells expressing Mye43 (TAM: tumor-
490 associated macrophage) and TNK35 (LAYN^{+} Effector Treg)/TNK26 is transmitted through
491 *CD80/86-CTLA4*; the communication between Mye26-expressing NLRP3^{+}INHBA^{+} macrophage
492 cells and TNK50-expressing CXCL13 producing CD4^{+} T cells is facilitated via *CD274-PDCD1*;
493 the communication between TNK50 and b44-expressing follicular B helper cell is transmitted
494 through *CXCL13-CXCR5* which is consistent with the current literature [34]. Moreover, we also
495 discovered the communication between a specific subset of mCAF and effector Treg. The

496 communication between Stroma37 expressing MMP1^{+}CXCL5^{+}mCAF cells and TNK35
497 expressing LAYN1^{+}Treg cells is transmitted through *CD274-PDCD1*. *MMP1* ranks the top 3rd
498 gene associated with Stroma37 and has been shown to play an important role in CAF-induced
499 protection and lead to therapy resistance [35]. The communication between stromal cells and Tregs
500 has been validated in several studies [36]. Besides, TNK51 expressing Th1 cells communicate with
501 myeloid GEMs (Mye8 and Mye42) expressing suppressor-like macrophage (M2) through *TNF-*
502 *TNFRSF1B*[37,38]. The above results demonstrate that re-constructing CCC networks provides
503 insights into potential molecular mechanisms of CCCN underlying immune evasion mechanisms
504 of tumors.
505
506 NK cells are a promising alternative platform for cellular immunotherapy, and clinical trials of NK
507 cell immunotherapy have produced many encouraging clinical results [39]. In **Fig. 6C**, a CCC sub-
508 network illustrates the primary communications between TNK32-expressing NK cells and specific
509 cellular functions of stromal, myeloid cells, and B cells through *ICAM1/3-LFA-1*, *IL10-IL10RA*,
510 and *IL15-IL15RB* interactions. Besides, we discovered a highly ranked potential CCC channel
511 involving Stroma22 expressing CAF, TNK14 expressing Treg and TNK50 expressing
512 CXCL13^{+}CD4^{+} T cells through *OX40L-OX40,* which is a costimulatory receptor as shown in **Fig.**
513 **6D**. Recent studies have shown the benefit of combining OX40 agonist and PD-1 inhibitor to
514 improve the immunotherapy outcome [40-42]. In both the CAMILLA trial dataset, where combined
515 therapy involved a VEGFR2 inhibitor and anti-PD-L1, and the dataset from melanoma patients
516 treated solely with anti-PD1 immunotherapy, Stroma22 exhibits a notable difference between the
517 responder and non-responder groups. (**Extended Data Fig. 6**). In addition to the aforementioned
518 findings, we have identified additional *GEM X-LR-GEM Y* triplets as potential CCCNs in

colorectal cancer (CRC). These discoveries offer valuable insights into exploring intercellular communication within the TME and understanding the key cellular transcriptomic processes associated with ligand-receptor signaling. The details are listed in **Extended Data Table 3**. To enhance the user-friendliness of our framework, we've created an interactive website where users can explore the annotation of each GEM, identify the top-ranked GEMs associated with each ligand-receptor pair, and investigate the *GEM X-LR-GEM Y* triplets of interest (http://44.192.10.166:3838/).

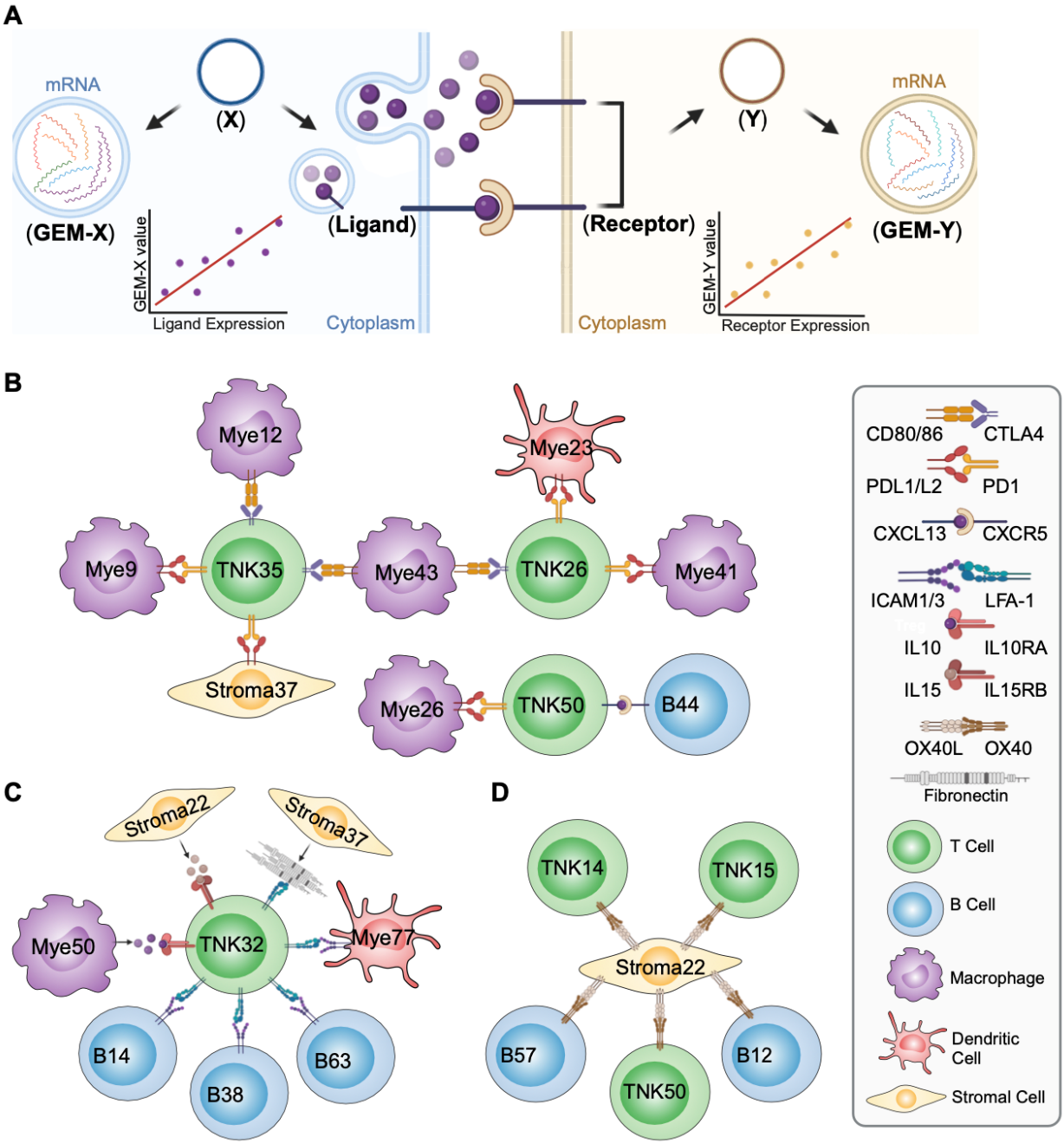


Fig. 6: The newly discovered cellular networks associated with ligand-receptor interactions involved in the immune checkpoint inhibitor therapy. A. The framework of the conditional independence analysis to find the causal relationship between GEM-expressing cells transmitted by ligand-receptor signals. **B.** Subnetwork of intercellular communication related to PDL1/2-PD1, CD80/CD86-CTLA4, and CXCL13-CXCR5 based on the conditional independence analysis. The

532 detailed annotation of each GEM is listed in the interactive website. **C.** Subnetwork of intercellular
533 communication related to TNK32 which represents the cellular function of NK cells. **D.**
534 Subnetwork of intercellular communication between Stroma22 and GEMs from other cell types
535 through the OX40L-OX40 signaling.
536

537 # Discussion

538 In this study, we showed that by deconvoluting the hierarchically organized transcriptomic
539 programs and closely representing cell differential programs, the nHDP model can identify GEMs
540 that are potentially regulated through CCC, discover possible CCC channels, and generate the
541 hypothesis of the molecular mechanisms of CCC underlying heterogeneous tumor
542 microenvironments of CRC. Detecting and representing the states of cellular signaling systems in
543 individual cells is a fundamental task in single-cell biology. While there are different approaches
544 for deconvoluting single-cell transcriptomic components [16,43], our results indicate that by modeling
545 the hierarchical and compositional relationships among transcriptomic programs of cells, the
546 nHDP model captures fine-grained transcriptomic signals better than other non-hierarchical
547 models.
548

549 Although scRNAseq assays offer unparalleled cell-type-specific insights, their practical use is
550 impeded by several obstacles (e.g. high cost of library preparation and sequencing) that hinder
551 their broad application in clinical contexts. The evolution of bulk RNA-seq data over the past
552 decades has yielded a substantial volume of samples conducive to the execution of pattern and
553 machine learning investigations. The study of CCC, especially the TME associated with drug
554 treatments, could be further promoted by integrating insights leaned from scRNA-seq data.

555 Therefore, effective methods for bulk RNA-seq deconvolution are still urgently needed. The
556 GEMs learned from nHDP could be used as novel signatures for deconvoluting the specific
557 transcriptomic processes in both bulk RNA-seq data and spatial transcriptomic data. Compared
558 with two state-of-the-art statistical models, non-negative matrix factorization (NMF) and latent
559 Dirichlet allocation (LDA), for identifying transcriptomic programs, nHDP model has the
560 advantage of capturing the hierarchical structure of signaling pathways. GEMs derived from the
561 nHDP model reflect the cellular states across multiple cell subtypes or specific to a particular cell
562 subtype. Our recent study [44] showed that GEMs offered a more precise estimation of biological
563 signals, such as immune cell infiltration, in comparison to the state-of-the-art deconvolution
564 methods like Cibersort and SCNA [45,46]. This improved accuracy enables the identification of
565 different lymphocyte infiltration levels.
566
567 Identifying highly specialized GEMs enabled us to use the expression level of GEMs as proxies to
568 represent the state of signaling pathways (unobserved) that regulate their expression. This, in turn,
569 enabled us to search ligand-receptors that likely transmit extracellular signals to the pathways
570 inside single cells, generating the hypothesis of how across-cell signaling influences intracellular
571 signaling through CCC. Our results indicate that the *CD274-PDCD1* ligand-receptor pair may
572 influence the state of the cellular signaling pathways that regulate the expression of certain
573 members of TNK26 in exhausted T cells. The hypothesis is supported by multiple lines of evidence:
574 The non-linear dose-response relationship between the approximated ligand-receptor complex and
575 the GEM provides support from a biochemical point of view. The significant difference in TNK26
576 expression between responding and non-responding tumors treated with anti-PD-1 agents (both
577 newly developed clinical trials in CRC and melanoma datasets) indicates that tumors with high

578 TNK26 expression display active PD-L1/PD-1 signaling, rendering them more susceptible to the
579 treatment. Conversely, tumors with low TNK26 expression exhibit relatively inactive PD-L1-PD-
580 1 signaling, leading to their reduced responsiveness to the treatment. Based on these results, we
581 hypothesize that the PD-L1/PD-1 signaling likely regulates some members of the TNK26. Further
582 investigation of which members are regulated by this signal, e.g., using principled causal discovery
583 methods, may enable researchers to delineate mechanisms of PD-1 signaling. Similar studies of
584 the impact of other commonly co-expressed immune checkpoint proteins (receptors) in tumor-
585 experienced T cells may reveal how different checkpoint signaling pathways cooperate and lead
586 to immune suppression.
587
588 The discovery of a strong correlation of GEMs across cell types indicates that the expression of
589 GEMs by different cell types is coordinated, and such coordination is likely due to CCC. We
590 identified coordinated GEM expression patterns, referred to as CCC programs, and their spatial
591 colocalization provides additional support for intercellular communication networks. We further
592 noticed the differences in the activities of CCC programs may underlie the heterogeneity of TMEs,
593 leading to tumor clusters that recapitulate well-appreciated molecular subtyping of CRC. Further
594 investigation identifies ligand-receptors that likely transmit signals across GEMs in CCC channels.
595
596 We anticipate further investigating CCC channels active in each tumor/subcluster may provide
597 mechanistic insights into the TME heterogeneity. In our population-wide exploration of CCC, we
598 observed instances where multiple ligands could communicate with a single receptor, as illustrated
599 by the TNK32 in **Fig. 6C**. In the causal theory, those ligands are multiple parents of the receptor.
600 If we examine whether instantiated values of variables in a local network lead to context-specific

601 independence, only certain parents may contribute to the state of children from the aspect of an
602 individual tumor sample. For instance, in an individual tumor sample, only one type of cell among
603 stromal, myeloid, or B cells may be involved in the interaction with TNK32-expressing cells. Our
604 forthcoming efforts will involve employing the individualized Causal Bayesian Network (iCBN)
605 model [47] to reveal the individualized CCCN within each tumor sample. This effort will contribute
606 to a deeper understanding of the mechanisms underlying heterogeneity.
607
608 We envision that the framework presented in this study for deconvoluting single-cell
609 transcriptomic programs to model CCC can be applied to various cancer types. Enhancing the
610 nHDP model involves refining the initialization process and filtering variable genes, leading to a
611 more effective capture of relationships within transcriptomic programs. This improvement will
612 boost the model's ability to detect GEMs regulated by specific transcriptomic programs. Future
613 methodological developments for detecting fine-grained transcriptomic programs may involve
614 leveraging recently developed language models to capture the compositionally co-expressed genes.
615 Additionally, integrating scRNAseq with single-cell chromatin accessibility sequencing (scATAC)
616 could aid in discovering active motifs associated with CCC networks. In conclusion, this study
617 advances the understanding of CCCs mechanisms underlying heterogeneous TME and immune
618 evasion mechanisms in CRC. These insights may uncover novel target for immunotherapy and
619 guide precision immunotherapy [21].
620

621 # Online Method
622
623 ## Data Collection and Preprocessing

624 ***Single-cell transcriptomics***

625 We collected single-cell RNA sequencing data from 8 publicly available colorectal cancer data
626 sources (**Extended Data Table 1**) [22,48-50]. The total number of patients is 153, of which there are
627 179 and 100 scRNA tumor and normal samples respectively. We retained cells that met the
628 following criteria: nUMI > 500, nGene > 200, log10GenesPerUMI > 0.8, and mitoRatio < 0.2. To
629 perform gene-level filtering, we retained only genes expressed in at least 0.1% of the cells. After
630 the quality control (QC) analysis, we applied the total-count correction (library-size correction)
631 and normalized the data matrix to 10,000 reads per cell. The annotations of cell major type were
632 obtained from original publications and used for further analysis. We then divided cells into five
633 major categories: 246,560 T and natural killer (TNK) cells, 94,663 myeloid cells, 96,637 plasma
634 and B cells, 69,128 stromal cells, and 119,870 epithelial cells. When performing clustering
635 analysis within each cell category, the Harmony algorithm [51] was applied to remove the batch
636 effect, cell clustering was carried out using the Leiden algorithm [52], and the UMAP (Uniform
637 Manifold Approximation and Projection) algorithm was used to visualize clusters.
638 Seurat pipeline was used to find subclusters and its top variable genes within TNK, myeloid,
639 plasmaB, stromal, and epithelial cells, respectively [13,53]. The cell types were annotated with well-
640 known marker genes: *CD3D, CD8A, FOXP3, IL7R, NKG7, GNLY* for T and NK cells; *CD79A,*
641 *MS4A1, MZB1* for B and plasma cells; *CD14, CD68, CD1C, S100A8, C1QA, LYZ* for myeloid
642 cells. The detailed markers we used were listed in **Extended Data Table 5**.
643 ***Bulk RNA-seq***
644 The TCGA bulk RNA-seq data of colorectal cancer (COADREAD) was downloaded from UCSC
645 Xena [54]. The total number of tumor samples is 290. Bulk RNA-seq data of 20 patients with
646 metastatic CRC, treated with the combination of Cabozantinib plus durvalumab in the CAMILLA
647 clinical trial and characterized by proficient microsatellite stable (MSS) tumors, were collected

upon request [26]. The number of responders and non-responders is 4 and 16, respectively. The bulk RNA-seq data with PD1 response of melanoma was downloaded from GSE96061, PRJEB23709 and dbGap:phs000452.v3.p1 for validation analysis [55-57]. To facilitate the clustering of tumors and ligand receptor search using the CRC single-cell data atlas, we generated pseudo-bulk RNAseq results for tumor samples (a *tumor-by-GEM* matrix). For the tumor samples (n=151) that include both $CD45^-$ and $CD45^+$ cells, UMI counts of the same gene in each sample were aggregated, resulting in the expression data in raw count format. We then normalized the raw count expression data to $\log_2(TPM+1)$.

Spatial transcriptomics

Spatial transcriptomic data of colorectal cancer and its corresponding liver metastasis using 10x Visium technology were collected from the study by Wu et al [49]. We first used sctransform method[58] to normalize the raw count in each spot of the slide (a *spot-by-gene* matrix). The GSVA method was then used to estimate the GEM values from normalized expression data. The deconvoluted spatial transcriptomic data is represented as a *spot-by-GEM* matrix. STdeconvolve was used to infer the cell-type composition in tumor region [59].

Identification of GEMs using nHDP

We used the nested hierarchical Dirichlet processes (nHDP) model [44,60] to identify patterns of GEMs from scRNA data. The flow of the training process is shown in **Fig.1B-C**. We binarized the UMI count of each gene within a cell as the input of the nHDP model. A trained nHDP model returns two matrices: a cell-by-GEM count matrix, which collects the counts of UMIs assigned to a GEMs in each cell; a GEM-by-gene matrix, which represents the probability profile of genes in each GEM (**Fig.1B**). We constructed a three-layer tree structure with the following branching

671 factors: 5, 4, and 3 for 1st, 2nd, and 3rd layers, respectively, as shown in **Fig.1C**. The expression of
672 GEMs (cell-by-GEM matrix) is added to the Seurat object as new features for plotting the UMAP
673 of GEMs. The dominant function represented by a GEM is based on the rank of cell subtype
674 percentage ("Distribution" in **Extended Data Fig. 4**). We used word clouds R package
675 "wordcloud2" to visualize the probability profile of the top 20 associated genes of a GEM (**Fig.1D**).
676 We trained a nHDP model for each cell category (TNK, B, Mye, Stroma, and Epi) and retained
677 GEMs expressed in over 1% of cells in a category for further analysis.
678

679 **Functional analysis of GEMs**

680 Transcription factor and enrichment analysis are applied to conduct functional analysis on GEMs.
681 CHEA3 is used to conduct the transcription factor analysis [23]. Reactome, KEGG and GO databases
682 are used to conduct more comprehensive analysis using R package "msigdbr"[61].
683

684 **Estimation of GEM level in bulk and spatial RNA-seq data**

685 To estimate the expression level of a GEM in bulk/pseudo-bulk/spatial RNA-seq data, we used
686 two methods: 1) gene set variation analysis (GSVA) of the top 50 weighted genes in each GEM [62].
687 2) mean expression of the top 50 weighted genes in each GEM.
688

689 **Fitting non-linear relationship between ligand-receptor pairs and GEMs**

690 The ligand-receptor database was downloaded from Nichenet [20]. To fit the dose-response curve
691 between the expression of ligand-receptor and GEM value, we model the quantity of LR complex
692 [LR] as the product of expression of L and R based on the fact that $[LR] \propto [L][R]$ using the

693 pseudo-bulk RNA-seq data. We first scaled the GEM value to [0, 1] and then created a 2-parameter
694 log-logistic function where the lower limit fixes at 0 and upper limit fixes at 1 using *drc/v3.0.1* [63].
695 The coefficients and its associated *p*-values were estimated and adjusted[20]. Sigmoid fitting with a
696 significant adjusted *p*-value (<0.05) demonstrates a good fit between the expression of LR complex
697 and the scaled GEM values.
698

699 **Assessing conditional independence of *GEM X-LR-GEM Y* triplets**

700 To find out the ligand-receptor pairs that potentially mediate CCC between cells across cell
701 categories expressing two correlated GEMs, conditional independence tests of $GEM\ X \perp\!\!\!\perp GEM\ Y \mid$
702 LR were performed using the *bnlearn/v4.7.1* package in R. We estimated expression status of each
703 GEM using the GSVA scores (*GSVA/v1.42.0*) based on its top 50 genes. We calculated all pairwise
704 correlations between GEMs expressed in different cell categories. If a pair of cross-cell-category
705 GEMs exhibits significant correlation, we further search for LR pairs that rendered conditional
706 independence ($GEM\ X \perp\!\!\!\perp GEM\ Y \mid LR$). *P*-values of conditional independence tests were adjusted
707 to control the false discovery rate using Benjamini-Hochberg method. We set the threshold of the
708 conditional correlation between *GEM1* and *GEM2* at an adjusted *p*-value > 0.05 as conditional
709 independent.
710

711 We further performed post-hoc proportional tests to remove false positive findings. For each
712 conditional independent *GEM X—LR—GEM Y* triplet, we investigated the co-expression of *GEM*
713 *X—L* in potential signal-originating cells and the co-expression of *R—GEM Y* in potential signal-
714 receiving cells. To conduct the proportional tests, we first construct a 2 by 2 contingency table to
715 measure the frequency of co-expression GEM and the tested gene. GEM and gene were considered

716 presenting if the corresponding values are beyond 0 and not presenting otherwise. Then the
717 statistics of proportional tests $z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ were estimated to test if the proportion of co-
718 expressing cells is significantly higher than the proportion of not co-expressing cells, where $\hat{p}$ is
719 the proportion of cells that co-express GEM and tested gene and p_0 is the proportion of cells that
720 not co-express GEM and tested gene, *n* is the total number of cells that express such GEM. *p*-value
721 of each post-hoc test was also adjusted using the Benjamini-Hochberg method for false discovery
722 rate control. Ligand-receptor pair that passes both conditional independence test and post-hoc
723 proportional test was considered valid for downstream analysis.
724
725 To prevent the false-positive rate, we set multiple criteria to filter the GEM pairs, including 1) the
726 correlation between *GEM X* and *L* is over 0.3 based on pseudo-bulk RNA-seq data 2) the
727 correlation between *LR* and *GEM Y* is over 0.3 based on pseudo-bulk RNA-seq data 3) more than
728 10% of cells expressing *GEM X* also express ligand gene *L* and more than 10% of cells expressing
729 *GEM Y* also express receptor gene *R*. We incorporated the examination of expression at the single-
730 cell level to confirm the presence of a specific molecule (ligand gene/receptor gene) in the GEM
731 expressed cells, thereby reducing false positives.
732

733 **Identifying programs of GEMs**

734 Hierarchical clustering was performed on a Euclidean distance matrix to identify GEM programs
735 using the pseudo-bulk expression of GEMs.
736

737 **Evaluating GEMs as potential biomarkers for anti-PD1 therapies**

738 We performed GSVA analysis to determine the expression levels of GEMs for each bulk tumor
739 sample. Samples were divided to response group if they showed complete response or partial
740 response to anti-PD-1 whereas non-response group if progressive disease or stable disease were
741 seen. We employed one-sided Wilcoxon rank test [64] to identify GEMs that exhibited a significant
742 difference in expression levels between these two groups.
743

744 Availability

745 nHDP model designed for analyzing single-cell transcriptome (scGEM):
746 https://github.com/hansolo-bioinfo/scGEM
747 CRC Atlas: https://github.com/hansolo-bioinfo/CRCAtlas
748 Interactive website: http://44.192.10.166:3838/
749

750 Acknowledgments

751 This research was supported in part by The National Library of Medicine; National Cancer
752 Institute at the National Institutes of Health grants R00LM013089, R01HG014023,
753 R01LM012011 and R01CA254274. The Department specifically disclaims responsibility for any
754 analysis, interpretations or conclusions.
755

756 Reference

757 1 Xi, Y. & Xu, P. Global colorectal cancer burden in 2020 and projections to 2040. *Transl*
758 *Oncol* **14**, 101174, doi:10.1016/j.tranon.2021.101174 (2021).
759 2 Sahin, I. H. *et al.* Immune checkpoint inhibitors for the treatment of MSI-H/MMR-D
760 colorectal cancer and a perspective on resistance mechanisms. *Br J Cancer* **121**, 809-818,
761 doi:10.1038/s41416-019-0599-y (2019).

762 3 Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell-cell
763 interactions and communication from gene expression. *Nat Rev Genet* **22**, 71-88,
764 doi:10.1038/s41576-020-00292-x (2021).
765 4 Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nat*
766 *Commun* **12**, 1088, doi:10.1038/s41467-021-21246-9 (2021).
767 5 Giladi, A. *et al.* Dissecting cellular crosstalk by sequencing physically interacting cells.
768 *Nat Biotechnol* **38**, 629-637, doi:10.1038/s41587-020-0442-2 (2020).
769 6 Shao, X., Lu, X., Liao, J., Chen, H. & Fan, X. New avenues for systematically inferring
770 cell-cell communication: through single-cell transcriptomics data. *Protein Cell* **11**, 866-
771 880, doi:10.1007/s13238-020-00727-5 (2020).
772 7 Marx, V. Method of the Year: spatially resolved transcriptomics. *Nat Methods* **18**, 9-14,
773 doi:10.1038/s41592-020-01033-y (2021).
774 8 Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nat Methods* **19**, 534-546,
775 doi:10.1038/s41592-022-01409-2 (2022).
776 9 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell
777 transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*
778 **36**, 411-420, doi:10.1038/nbt.4096 (2018).
779 10 Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a
780 tutorial. *Mol Syst Biol* **15**, e8746, doi:10.15252/msb.20188746 (2019).
781 11 Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression
782 data analysis. *Genome biology* **19**, 15, doi:10.1186/s13059-017-1382-0 (2018).
783 12 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902
784 e1821, doi:10.1016/j.cell.2019.05.031 (2019).
785 13 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587
786 e3529, doi:10.1016/j.cell.2021.04.048 (2021).
787 14 Kotliar, D. *et al.* Identifying gene expression programs of cell-type identity and cellular
788 activity with single-cell RNA-Seq. *Elife* **8**, doi:10.7554/eLife.43803 (2019).
789 15 Zhu, X., Ching, T., Pan, X., Weissman, S. M. & Garmire, L. Detecting heterogeneity in
790 single-cell RNA-Seq data by non-negative matrix factorization. *PeerJ* **5**, e2888,
791 doi:10.7717/peerj.2888 (2017).
792 16 Liu, L. *et al.* Deconvolution of single-cell multi-omics layers reveals regulatory
793 heterogeneity. *Nat Commun* **10**, 470, doi:10.1038/s41467-018-08205-7 (2019).
794 17 Venkatasubramanian, M., Chetal, K., Schnell, D. J., Atluri, G. & Salomonis, N.
795 Resolving single-cell heterogeneity from hundreds of thousands of cells through
796 sequential hybrid clustering and NMF. *Bioinformatics* **36**, 3773-3780,
797 doi:10.1093/bioinformatics/btaa201 (2020).
798 18 Shao, C. & Hofer, T. Robust classification of single-cell transcriptome data by
799 nonnegative matrix factorization. *Bioinformatics* **33**, 235-242,
800 doi:10.1093/bioinformatics/btw607 (2017).
801 19 Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB:
802 inferring cell-cell communication from combined expression of multi-subunit ligand-
803 receptor complexes. *Nat Protoc* **15**, 1484-1506, doi:10.1038/s41596-020-0292-x (2020).
804 20 Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication
805 by linking ligands to target genes. *Nat Methods* **17**, 159-162, doi:10.1038/s41592-019-
806 0667-5 (2020).

807 21 Chen, X. *et al.* An individualized causal framework for learning intercellular
808 communication networks that define microenvironments of individual tumors. *PLoS*
809 *computational biology* **18**, e1010761, doi:10.1371/journal.pcbi.1010761 (2022).
810 22 Pelka, K. *et al.* Spatially organized multicellular immune hubs in human colorectal
811 cancer. *Cell* **184**, 4734-4752 e4720, doi:10.1016/j.cell.2021.08.003 (2021).
812 23 Keenan, A. B. *et al.* ChEA3: transcription factor enrichment analysis by orthogonal
813 omics integration. *Nucleic Acids Res* **47**, W212-W224, doi:10.1093/nar/gkz446 (2019).
814 24 Martinez, G. J. *et al.* The transcription factor NFAT promotes exhaustion of activated
815 CD8(+) T cells. *Immunity* **42**, 265-278, doi:10.1016/j.immuni.2015.01.006 (2015).
816 25 Fan, J. *et al.* Zinc finger protein 831 promotes apoptosis and enhances chemosensitivity
817 in breast cancer by acting as a novel transcriptional repressor targeting the STAT3/Bcl2
818 signaling pathway. *Genes Dis* **11**, 430-448, doi:10.1016/j.gendis.2022.11.023 (2024).
819 26 Saeed, A. *et al.* Cabozantinib plus durvalumab in advanced gastroesophageal cancer and
820 other gastrointestinal malignancies: Phase Ib CAMILLA trial results. *Cell Rep Med* **4**,
821 100916, doi:10.1016/j.xcrm.2023.100916 (2023).
822 27 . (!!! INVALID CITATION !!! 23-25).
823 28 Hoeft, K. *et al.* Platelet-instructed SPP1(+) macrophages drive myofibroblast activation
824 in fibrosis in a CXCL4-dependent manner. *Cell Rep* **42**, 112131,
825 doi:10.1016/j.celrep.2023.112131 (2023).
826 29 Qi, J. *et al.* Single-cell and spatial analysis reveal interaction of FAP(+) fibroblasts and
827 SPP1(+) macrophages in colorectal cancer. *Nat Commun* **13**, 1742, doi:10.1038/s41467-
828 022-29366-6 (2022).
829 30 Mao, X. *et al.* Crosstalk between cancer-associated fibroblasts and immune cells in the
830 tumor microenvironment: new findings and future perspectives. *Mol Cancer* **20**, 131,
831 doi:10.1186/s12943-021-01428-1 (2021).
832 31 Chen, Z. *et al.* Single-cell RNA sequencing highlights the role of inflammatory cancer-
833 associated fibroblasts in bladder urothelial carcinoma. *Nat Commun* **11**, 5077,
834 doi:10.1038/s41467-020-18916-5 (2020).
835 32 . (!!! INVALID CITATION !!! 30).
836 33 Kim, H. D. *et al.* Regorafenib plus nivolumab in unresectable hepatocellular carcinoma:
837 the phase 2 RENOBATE trial. *Nat Med* **30**, 699-707, doi:10.1038/s41591-024-02824-y
838 (2024).
839 34 Rao, D. A. T Cells That Help B Cells in Chronically Inflamed Tissues. *Front Immunol* **9**,
840 1924, doi:10.3389/fimmu.2018.01924 (2018).
841 35 Cui, Q. *et al.* Upregulating MMP-1 in carcinoma-associated fibroblasts reduces the
842 efficacy of Taxotere on breast cancer synergized by Collagen IV. *Oncol Lett* **16**, 3537-
843 3544, doi:10.3892/ol.2018.9092 (2018).
844 36 Sun, R. *et al.* Amphiregulin couples IL1RL1(+) regulatory T cells and cancer-associated
845 fibroblasts to impede antitumor immunity. *Sci Adv* **9**, eadd7399,
846 doi:10.1126/sciadv.add7399 (2023).
847 37 Wilk, A. J., Shalek, A. K., Holmes, S. & Blish, C. A. Comparative analysis of cell-cell
848 communication at single-cell resolution. *Nat Biotechnol*, doi:10.1038/s41587-023-01782-
849 z (2023).
850 38 Nonaka, K. *et al.* Th1 polarization in the tumor microenvironment upregulates the
851 myeloid-derived suppressor-like function of macrophages. *Cell Immunol* **369**, 104437,
852 doi:10.1016/j.cellimm.2021.104437 (2021).

39 Chu, J. *et al.* Natural killer cells: a promising immunotherapy for cancer. *J Transl Med* **20**, 240, doi:10.1186/s12967-022-03437-0 (2022).
40 van der Sluis, T. C. *et al.* OX40 agonism enhances PD-L1 checkpoint blockade by shifting the cytotoxic T cell differentiation spectrum. *Cell Rep Med* **4**, 100939, doi:10.1016/j.xcrm.2023.100939 (2023).
41 Ma, Y. *et al.* Combination of PD-1 Inhibitor and OX40 Agonist Induces Tumor Rejection and Immune Memory in Mouse Models of Pancreatic Cancer. *Gastroenterology* **159**, 306-319 e312, doi:10.1053/j.gastro.2020.03.018 (2020).
42 Lao, J. *et al.* OX40 enhances T cell immune response to PD-1 blockade therapy in non-small cell lung cancer. *Int Immunopharmacol* **108**, 108813, doi:10.1016/j.intimp.2022.108813 (2022).
43 Im, Y. & Kim, Y. A Comprehensive Overview of RNA Deconvolution Methods and Their Application. *Mol Cells* **46**, 99-105, doi:10.14348/molcells.2023.2178 (2023).
44 Zhang, H., Lu, X., Lu, B. & Chen, L. scGEM: Unveiling the Nested Tree-Structured Gene Co-Expressing Modules in Single Cell Transcriptome Data. *Cancers (Basel)* **15**, doi:10.3390/cancers15174277 (2023).
45 Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* **355**, doi:10.1126/science.aaf8399 (2017).
46 Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453-457, doi:10.1038/nmeth.3337 (2015).
47 Liu, Z. *et al.* A Novel Bayesian Framework Infers Driver Activation States and Reveals Pathway-Oriented Molecular Subtypes in Head and Neck Cancer. *Cancers (Basel)* **14**, doi:10.3390/cancers14194825 (2022).
48 Joanito, I. *et al.* Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. *Nat Genet* **54**, 963-975, doi:10.1038/s41588-022-01100-4 (2022).
49 Wu, Y. *et al.* Spatiotemporal Immune Landscape of Colorectal Cancer Liver Metastasis at Single-Cell Level. *Cancer Discov* **12**, 134-153, doi:10.1158/2159-8290.CD-21-0316 (2022).
50 Liu, Y. *et al.* Immune phenotypic linkage between colorectal cancer and liver metastasis. *Cancer Cell* **40**, 424-437 e425, doi:10.1016/j.ccell.2022.02.013 (2022).
51 Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289-1296, doi:10.1038/s41592-019-0619-0 (2019).
52 Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**, 5233, doi:10.1038/s41598-019-41695-z (2019).
53 Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502, doi:10.1038/nbt.3192 (2015).
54 Goldman, M. J. *et al.* Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol* **38**, 675-678, doi:10.1038/s41587-020-0546-8 (2020).
55 Riaz, N. *et al.* Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell* **171**, 934-949 e916, doi:10.1016/j.cell.2017.09.028 (2017).
56 Gide, T. N. *et al.* Distinct Immune Cell Populations Define Response to Anti-PD-1 Monotherapy and Anti-PD-1/Anti-CTLA-4 Combined Therapy. *Cancer Cell* **35**, 238-255 e236, doi:10.1016/j.ccell.2019.01.003 (2019).

57 Liu, D. *et al.* Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma. *Nat Med* **25**, 1916-1927, doi:10.1038/s41591-019-0654-5 (2019).

58 Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* **20**, 296, doi:10.1186/s13059-019-1874-1 (2019).

59 Miller, B. F., Huang, F., Atta, L., Sahoo, A. & Fan, J. Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. *Nature communications* **13**, 2339 (2022).

60 Paisley, J., Wang, C., Blei, D. M. & Jordan, M. I. Nested Hierarchical Dirichlet Processes. *IEEE Trans Pattern Anal Mach Intell* **37**, 256-270, doi:10.1109/TPAMI.2014.2318728 (2015).

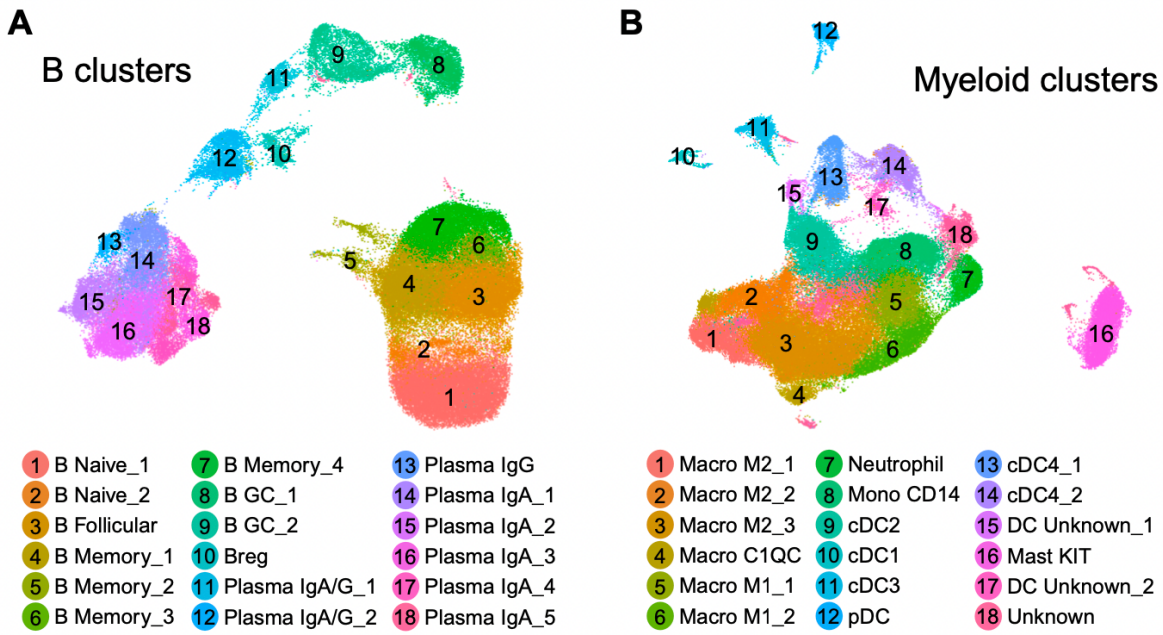
61 Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739-1740, doi:10.1093/bioinformatics/btr260 (2011).

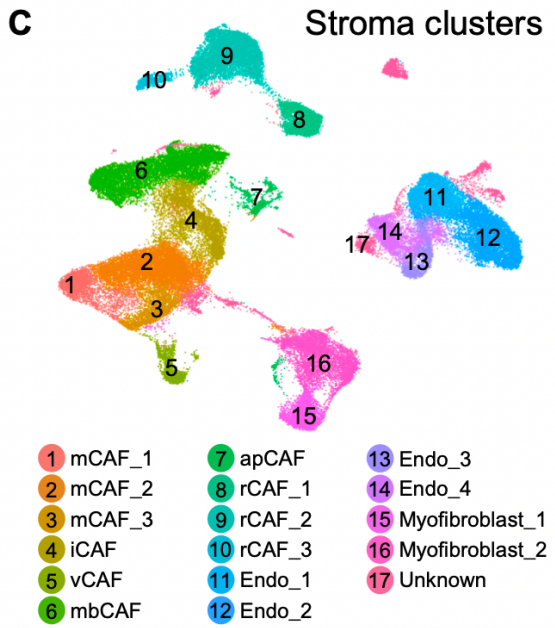
62 Hanzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7, doi:10.1186/1471-2105-14-7 (2013).

63 Ritz, C., Baty, F., Streibig, J. C. & Gerhard, D. Dose-Response Analysis Using R. *PLoS One* **10**, e0146021, doi:10.1371/journal.pone.0146021 (2015).

64 Forrester, J. C. & Ury, H. K. The Signed-Rank (Wilcoxon) test in the rapid analysis of biological data. *Lancet* **1**, 239-241, doi:10.1016/s0140-6736(69)91245-8 (1969).

Supplementary Figures

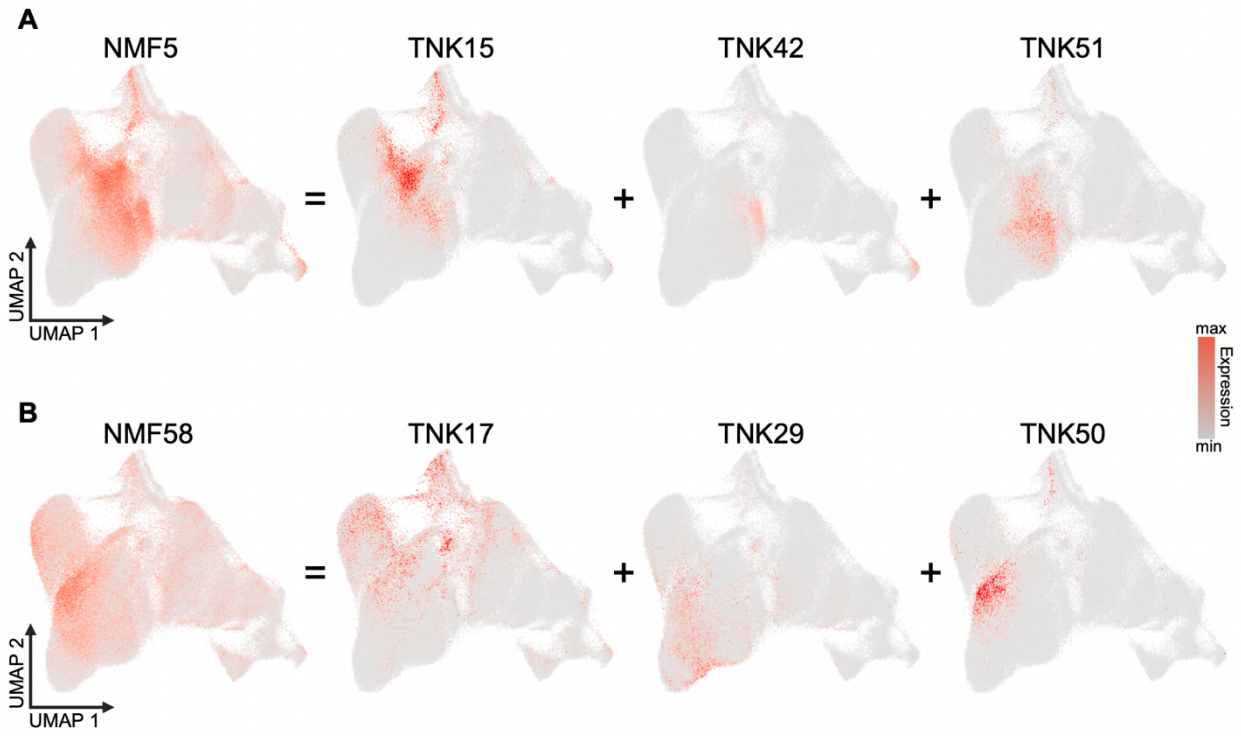




934 **Supp Fig. 1: UMAP of B, Myeloid and Stromal cells. A.** The cell clusters for B and Plasma cells.

935 **B.** The cell clusters for myeloid cells. **C.** The cell clusters for stromal cells.

936



937

938 **Supp Fig. 2: Comparison between GEMs learned from nHDP and GEMs learned from NMF.**

939 **A.** The cells expressing NMF5 are the combination of cells expressing TNK15, TNK42 and

940 TNK51. **B.** The cells expressing NMF58 are the combination of cells expressing TNK17, TNK29,

941 and TNK50.

942

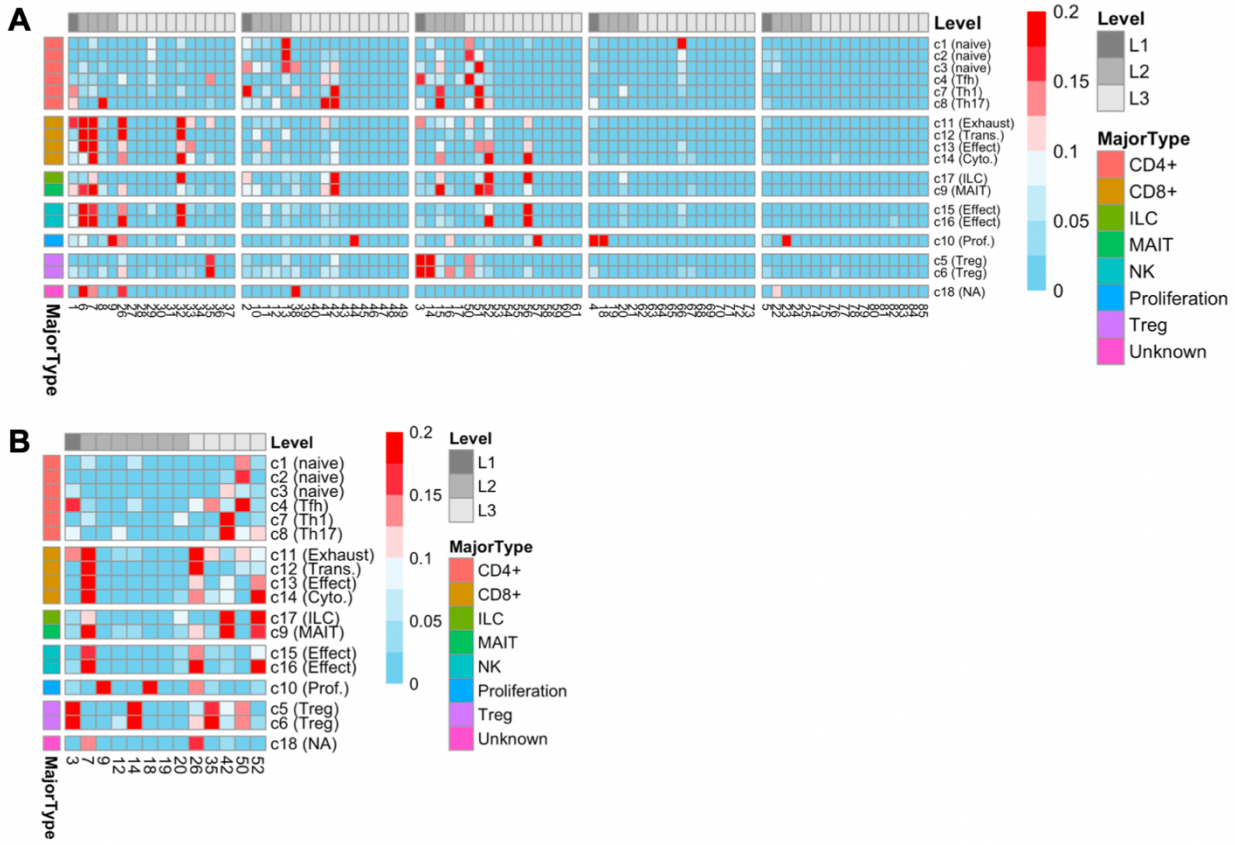
943

944

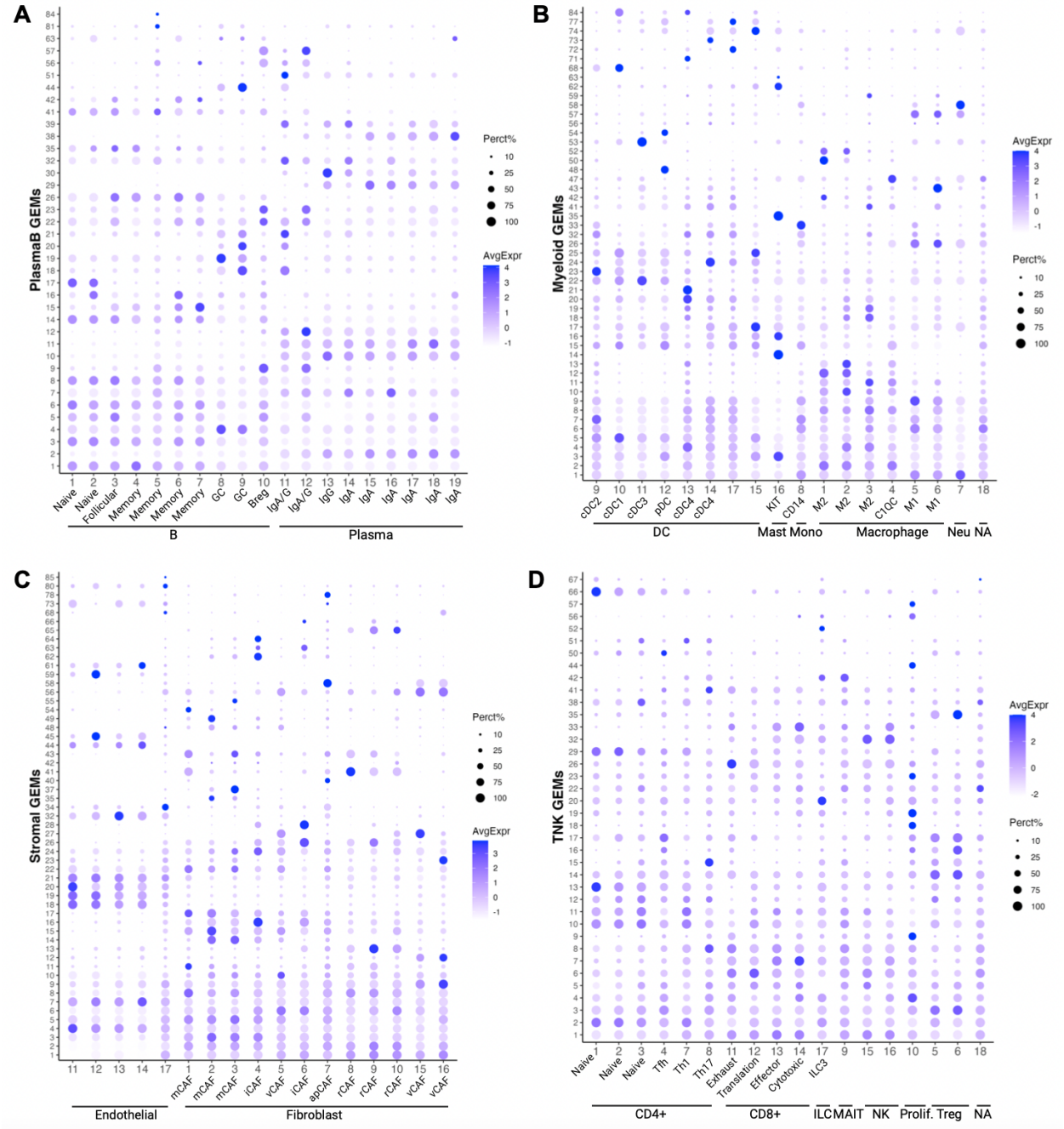
945

946

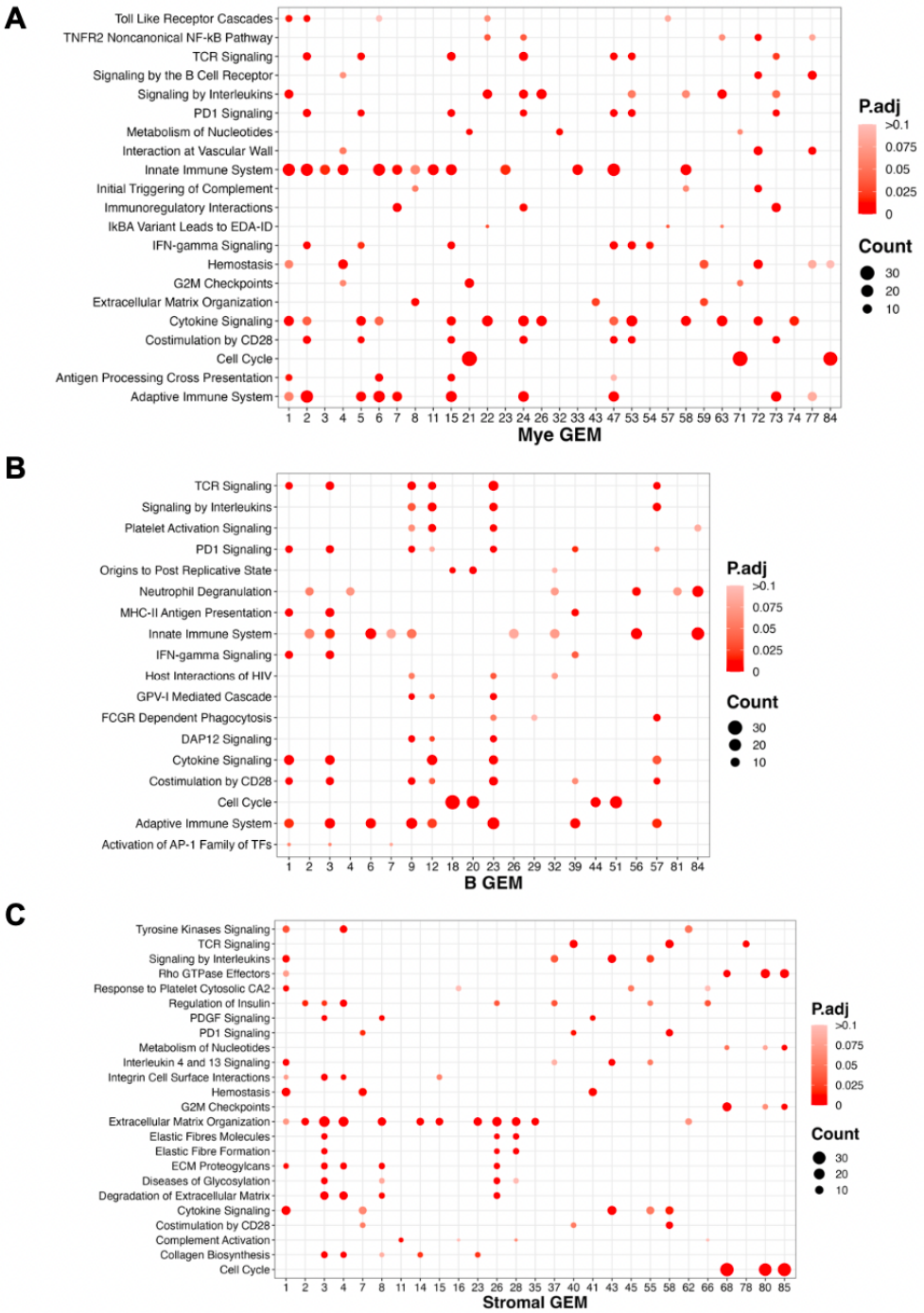
947



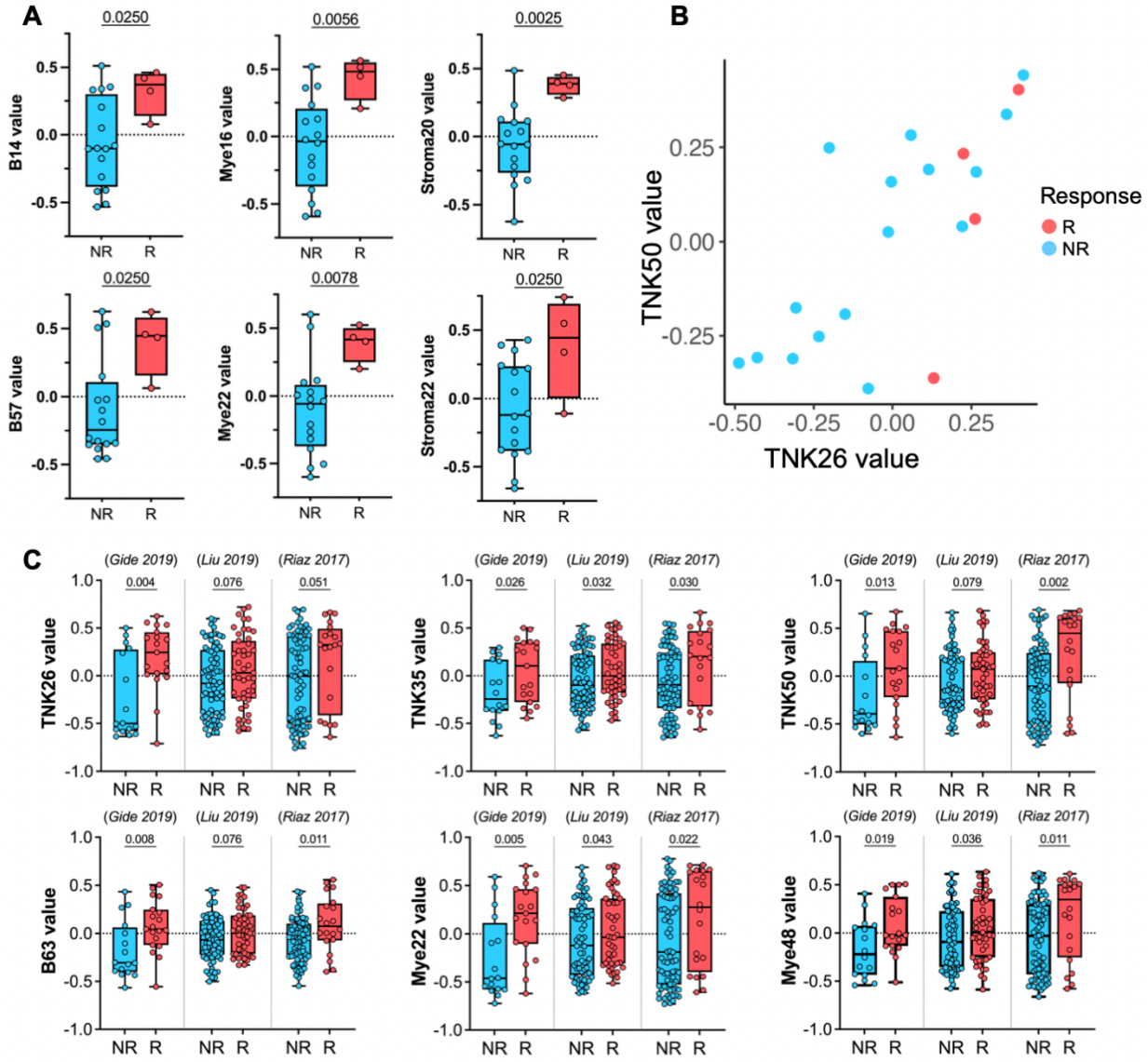
Supp Fig. 3: Cosine similarity between top50 GEM associated genes and highly expressed marker genes in T cell subtypes. A. Heatmap of cosine similarities between differentially expressed genes of T cell subtype and top 50 genes of all 85 T cell GEMs learned from nHDP model. GEMs are labeled with respect to their tree levels (L1, L2, and L3) and major T cell subtypes including $CD4^+$, $CD8^+$, ILC, MAIT, NK, Treg and proliferation T cells. **B.** The heatmap of cosine similarity for significant T cell GEMs summarized in Fig. 2E are shown.



Supp Fig. 4: The dot plot comparing the similarity between GEMs and traditional cell clusters. A. The dot plot for B and Plasma cells. **B.** The dot plot for myeloid cells. **C.** The dot plot for stromal cells. **D.** The dot plot for TNK cells. The color and size of the dot reflect the average expression and percentage of overlapped cells, respectively.

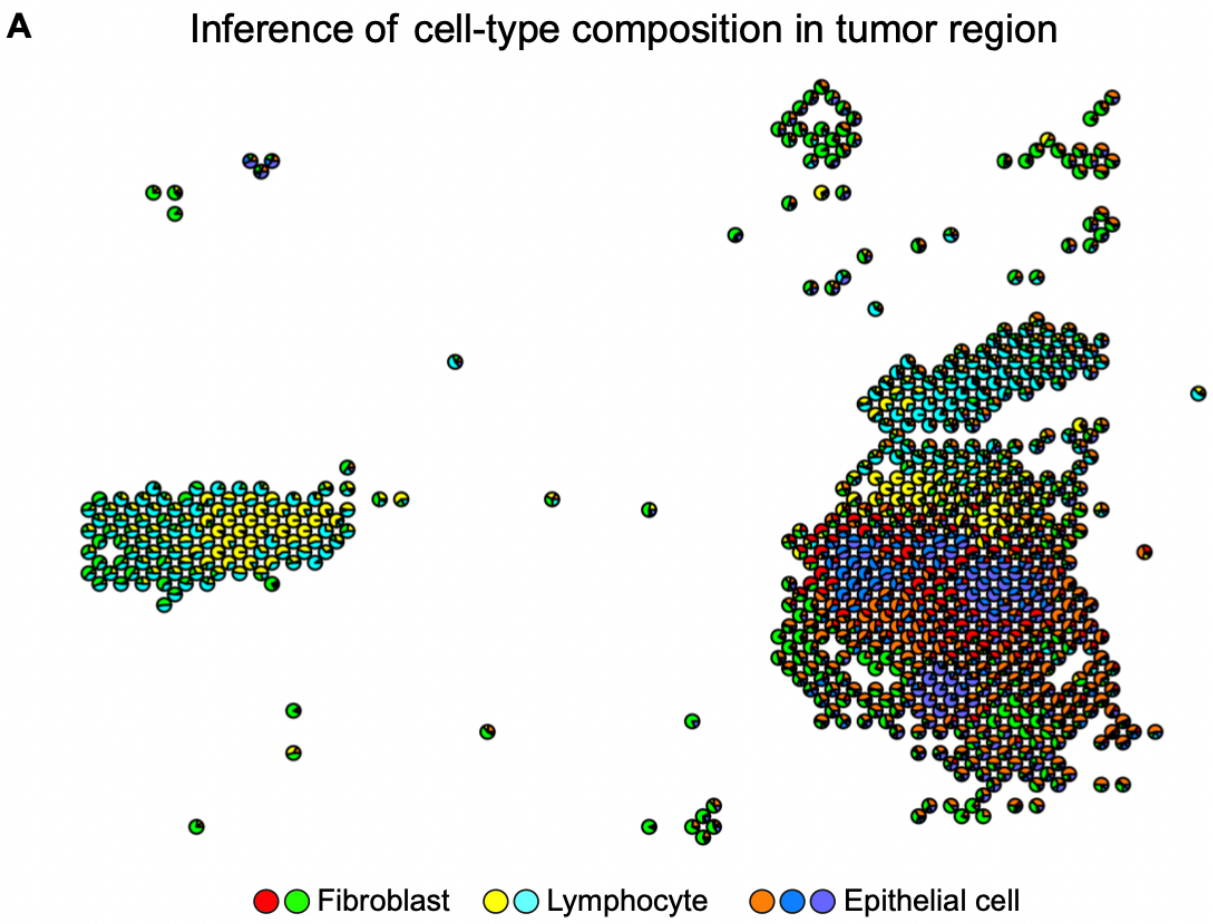


Supp Fig. 5: The enriched signaling pathways. A. The enriched signaling pathway for myeloid cells. **B.** The enriched signaling pathway for B and plasma cells. **C.** The enriched signaling pathway for stromal cells.



Supp Fig. 6: Example of GEMs significantly differentially expressed in responder group and non-responder group. A. Example of GEMs significantly differentially expressed in responder group and non-responder group treated with Cabozantinib plus durvalumab (CAMILLA trial). **B.** The correlation between TNK26 and TNK50 in CAMILLA cohort. **C.** Example of GEMs significantly differentially expressed in responder group and non-responder group treated with solo anti-PD-1 immunotherapy (three independent melanoma datasets).

A



B

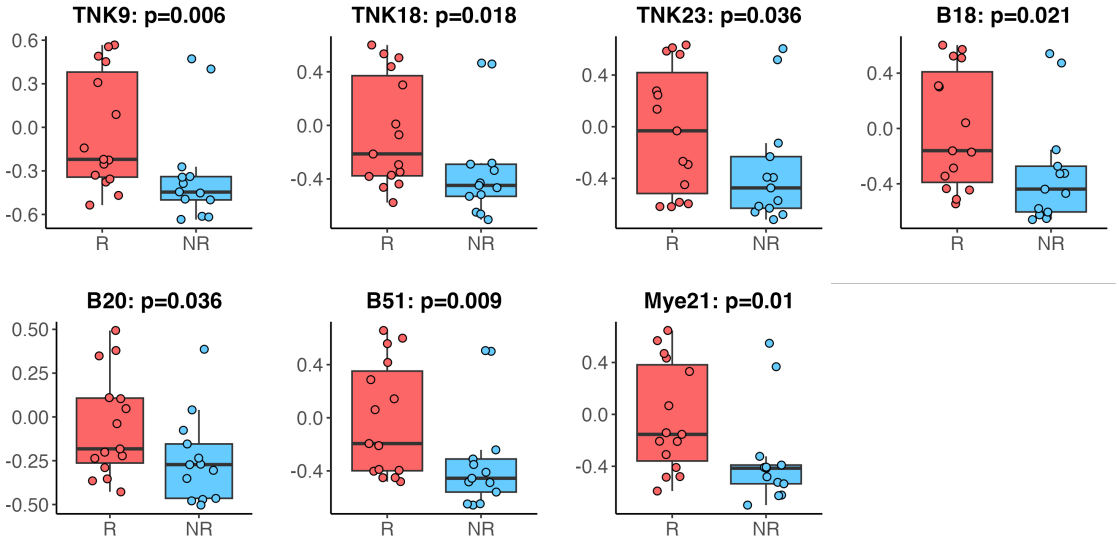
	B21	Stroma68	Epi23
B21	1	0	0
Stroma68	0	1	0.02
Epi23	0	0.02	1

Supp Fig. 7: Exploring the uniqueness of three GEMs in CCC program 2. A. The inference

982 of cell-type composition in tumor region of **Fig. 5D**. **B.** The cosine similarity between B21,

983 Stroma68 and Epi23 using the expression level of the top 50 associated genes.

984



985

986 **Suppl Fig 8: Example of GEMs in the proliferation program (CCC program2), which are**

987 **significantly differentially expressed in responders and non-responders treated with**

988 **regorafenib and nivolumab combination therapy.**

989