TITLE

An integrated single-cell RNA-seq map of human neuroblastoma tumors and preclinical models uncovers divergent mesenchymal-like gene expression programs.

AUTHORS AND AFFILIATIONS

Richard H. Chapple[1], Xueying Liu[1], Sivaraman Natarajan[1], Margaret I.M. Alexander[1], Yuna Kim[1], Anand G. Patel[2, 4], Christy W. LaFlamme[4, 5], Min Pan[1], William C. Wright[1], Hyeong-Min Lee[1], Yinwen Zhang[1], Meifen Lu[6], Selene C. Koo[6], Courtney Long[7], John Harper[7], Chandra Savage[7], Melissa D. Johnson[8], Thomas Confer[8], Walter J. Akers[9], Michael A. Dyer[2,10], Heather Sheppard[6], John Easton[1], Paul Geeleher[1, #]

[1] Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

[2] Department of Developmental Neurobiology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

[3] Department of Oncology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

[4] Department of Cell and Molecular Biology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

[5] Graduate School of Biomedical Sciences, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

[6] Department of Pathology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

[7] Animal Resources Center, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

23 [8] Center for In Vivo Imaging and Therapeutics, St. Jude Children's Research Hospital, Memphis,
24 TN 38105, USA

25 [9] Department of Biomedical Engineering, University of Texas Southwestern Medical School,
26 Dallas, TX 75390, USA

27 [10] Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA

28 [#] Correspondence: paul.geeleher@stjude.org

29

30 **ABSTRACT**

31 Neuroblastoma is a common pediatric cancer, where preclinical studies suggest that a
32 mesenchymal-like gene expression program contributes to chemotherapy resistance. However,
33 clinical outcomes remain poor, implying we need a better understanding of the relationship
34 between patient tumor heterogeneity and preclinical models. Here, we generated single-cell
35 RNA-seq maps of neuroblastoma cell lines, patient-derived xenograft models (PDX), and a
36 genetically engineered mouse model (GEMM). We developed an unsupervised machine
37 learning approach ('automatic consensus nonnegative matrix factorization' (acNMF)) to
38 compare the gene expression programs found in preclinical models to a large cohort of patient
39 tumors. We confirmed a weakly expressed, mesenchymal-like program in otherwise adrenergic
40 cancer cells in some pre-treated high-risk patient tumors, but this appears distinct from the
41 presumptive drug-resistance mesenchymal programs evident in cell lines. Surprisingly however,
42 this weak-mesenchymal-like program was maintained in PDX and could be chemotherapy-
43 induced in our GEMM after only 24 hours, suggesting an uncharacterized therapy-escape
44 mechanism. Collectively, our findings improve the understanding of how neuroblastoma patient
45 tumor heterogeneity is reflected in preclinical models, provides a comprehensive integrated

46 resource, and a generalizable set of computational methodologies for the joint analysis of

47 clinical and pre-clinical single-cell RNA-seq datasets.

BACKGROUND

Neuroblastoma is a neuroendocrine tumor, typically arising in the adrenal gland, and is the most common extracranial solid tumor in children[1]. Survival in high-risk neuroblastoma is approximately 50%, with survivors suffering long-term consequences of chemotherapy[2]. This is despite new therapies showing clear promise in cell lines and mouse models[3], indicating that we have an incomplete understanding of how neuroblastoma patient tumor heterogeneity is reflected in preclinical models.

Pediatric tumors, including neuroblastoma, exhibit very few recurrent somatic mutations compared to adult tumors[4] and heterogeneity at the level of cell state[5] is thought to contribute to therapy resistance[6-10]. For example, it was recently shown that many neuroblastoma cell lines exist in an admixed state, expressing two discrete super-enhancer-driven gene expression programs, referred to as the "*adrenergic*" (expressing sympathoadrenal features) and the "*mesenchymal*" (comparatively chemoresistant) programs[6,10]. These programs exhibited mutually exclusive expression, where cells expressing one program could interconvert to the other *in vitro*[11,12].

Single-cell RNA-seq represents the ideal tool to study this non-genetic tumor heterogeneity. Indeed, in the past three years, at least seven different studies[13-19] have reported single-cell RNA-seq maps of neuroblastoma patient tumors. However, there were disagreements among some of these studies, e.g., with multiple different cell types posited as the neuroblastoma cell-of-origin (chromaffin cell[16], sympathoblast/neuroblast[15,17,19], Schwann cell precursors[13], "sympatho- and chromaffin cells"[18] and "a subtype of TRKB+ cholinergic progenitor"[18])[15,17-19]. Additionally, arguments were made both in favor of[13-15,18] and against[16,17,19] the *in vivo* relevance of the cell-line-derived,[6,10] putatively-chemoresistant mesenchymal-like gene expression program. Thus, the gene expression programs contributing to therapy resistance in neuroblastoma—and the preclinical models that can be used to study these—remains unclear.

Here, we generated single-cell RNA-seq data from neuroblastoma cell lines, PDXs, and genetic mouse models, using these data to create the first integrated map of human neuroblastoma patient tumors and preclinical models. To achieve this, we developed a novel computational method called “automatic consensus nonnegative matrix factorization” (acNMF; details below). Using acNMF, we determined that the dominant adrenergic gene expression programs were preserved across human tumors and preclinical models. In contrast, the presumptive drug-resistant mesenchymal program demonstrated less consistent behavior, and while clearly expressed in cell lines, exhibited high expression primarily in cancer-associated normal cells *in vivo*. Notably, we confirmed a distinct weak-mesenchymal program in predominantly adrenergic cells of some high-risk patient tumors, which we showed could be elicited 24 hours after administering chemotherapy *in vivo*. Overall, our findings provide a generalizable set of computational tools that we have used to improve the understanding of neuroblastoma patient tumor heterogeneity and preclinical models, which will help inform drug development for this catastrophic disease.

RESULTS

Automatic consensus non-negative matrix factorization (acNMF) can accurately recover gene expression programs in single-cell RNA-seq data.

NMF is an unsupervised learning approach often used to estimate gene expression programs in single-cell RNA-seq datasets. Using this method, each program is represented as a vector of weights (or “loadings”) assigned to each gene (see, e.g., Fig. 3e). The activity of each program is estimated in individual cells, scoring high if genes with high weights are highly expressed in that cell (see, e.g., Fig. 3g). Like other methods, NMF can discover discrete cell types, but also, potentially, biological processes and cell state programs—factors believed to influence therapy resistance in cancers including neuroblastoma. However, there is no existing method to determine *a priori* the number of gene expression programs in a single-cell RNA-seq dataset, which can lead to inconsistent NMF-based representations of similar data. This makes it difficult to compare tumors and preclinical models across multiple datasets. To overcome this, we first aimed to create a method for identifying the number of gene expression programs (typically called "k") in our datasets. We eventually developed a generalizable method called acNMF (Fig. 1a).

Briefly (see Methods for complete details), we initially developed acNMF using a simulated single-cell RNA-seq dataset with 13 cell types and one shared "activity program" (simulating e.g. a biological process). Our inspiration came from observing how the existing cNMF algorithm[20] behaved on this dataset. We noticed that the 14 ground truth gene expression programs could be identified at most values of k, with other programs fitting noise (Fig. 1b). By plotting the value of k against the number of reproducibly identified gene expression programs in independent splits of the data, we could find an inflection point (Fig. 1c, Supplementary Figs. 1a-d) where all ground truth programs were recovered and noise-filtered (Fig. 1d-h; Supplementary Figs. 1e-r). In the simulated data, the 14 ground truth gene expression programs

112 were recovered at a k value of 22. At k=14, only 13 of the 14 programs were recovered
113 (Supplementary Fig. 1s, u). This finding implies that recovering ground truth gene expression
114 programs using NMF-based methods should be decoupled from selecting the appropriate k.
115 This is an important general consideration when applying NMF to single-cell RNA-seq data,
116 challenging the idea that finding an optimal k value necessarily recovers k biologically
117 meaningful programs. We used this method to estimate gene expression programs throughout
118 this manuscript (Fig 1i, Supplementary Fig. 1t; see subsequent sections).

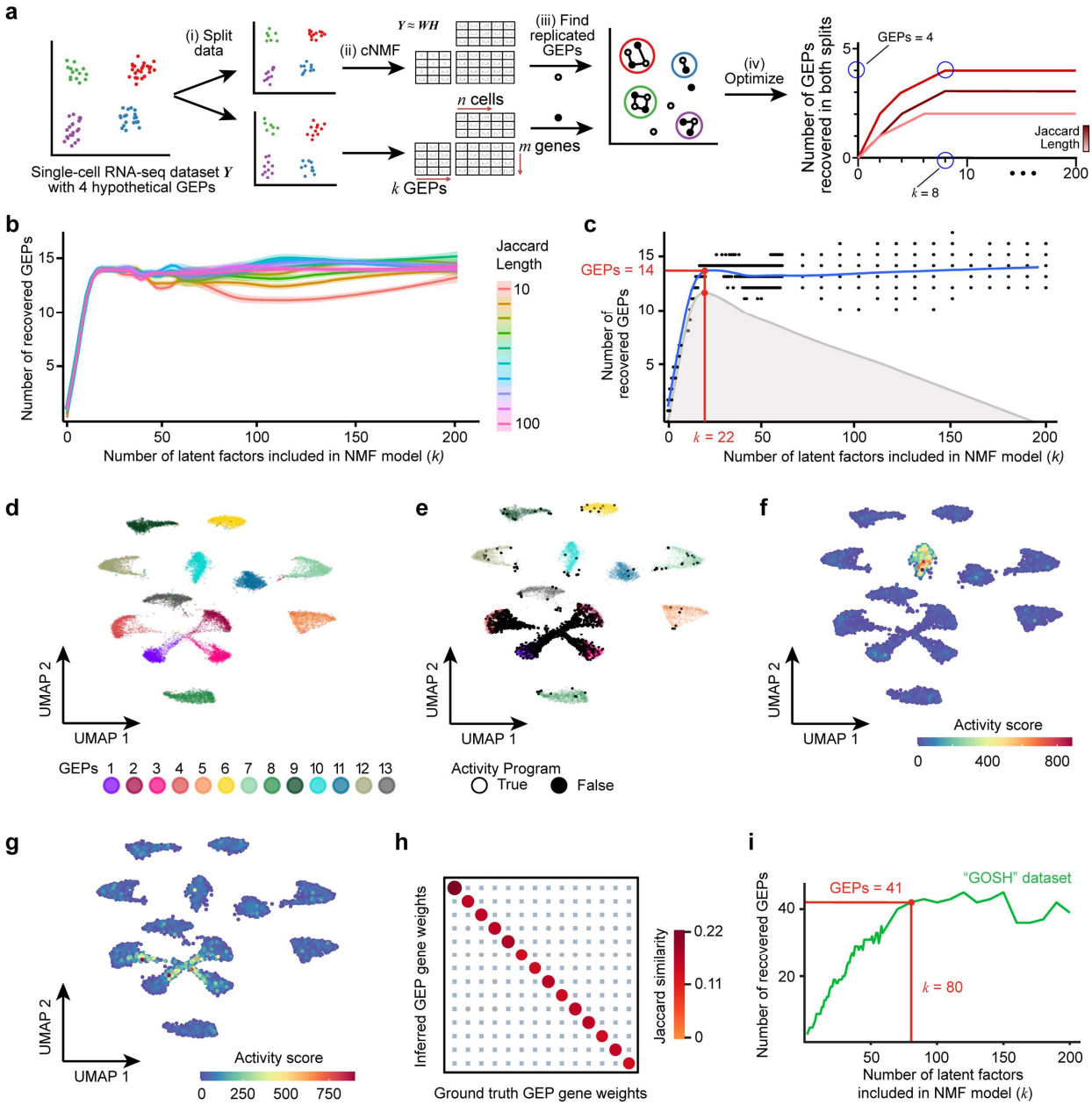


Figure 1: Overview of acNMF methodology and performance in simulated data.

(a) Schematic representation of acNMF approach for recovering gene expression programs in a single-cell RNA-seq dataset denoted ***Y***. The dataset is (i) randomly split 50/50, then (ii) each split is factorized independently into a basis matrix of gene expression programs (GEPs) ***W*** and a GEP activity matrix ***H*** using cNMF at a range of values of *k* (e.g. 5-200). At each *k* the redundant GEPs from each split are (iii) reaggregated by a community detection algorithm. This community

number is then (iv) determined for a range of values of *k* and a range of Jaccard length values, parameters which are chosen to maximize the number of GEP-communities replicated in each split. Critically, the number of ground truth GEPs is typically smaller than the value of *k* (GEPs < *k*, hypothetical example highlighted in blue).

(b) The number of GEPs reproducibly recovered in both splits of the simulated dataset (y axis) plotted against the value of *k* from NMF (x axis), over a range of values of the Jaccard length (color scale). The approach is repeated 200 times for each pair of values of *k* and Jaccard length and a loess regression curve is fit through these points, with 95% confidence intervals shown.

(c) Optimal rank *k* (x axis) is determined by calculating the inflection point (i.e., maximum residual value (gray profile)) of smoothed data in (b) (blue line).

(d) UMAP plot of simulated scRNA-seq data, where each dot is a cell colored by the thirteen ground-truth simulated cell types.

(e) Like (d), but with black dots highlighting cells expressing the activity program, which (akin to a biological process/pathway) is expressed across multiple simulated cell types.

(f) UMAP plot showing a representative example of a recovered ground truth cell identity GEP by acNMF in the simulated single cell RNA-seq dataset. Dots are colored by the activity score of the recovered gene expression program.

(g) Like (f) but showing the successful recovery of the activity program. Similar plots for all other recovered programs are shown in Supplementary Fig. 1e-r.

(h) Dot plot comparing Jaccard similarity index values of 14 inferred acNMF GEPs with the 14 ground truth GEPs after multiple testing correction (blue X's represent non-significant comparisons). Both the size and color of the points are scaled by Jaccard similarity.

(i) Like (b), but for the GOSH dataset. Data shown for a Jaccard length of 20, which was optimized using acNMF. 41 GEPs were reproducibly recovered in independent splits of the data, at a $k = 80$ (highlighted by the red lines). Similar plots for all datasets in this paper are shown in Supplementary Fig. 1t. GEP; Gene expression program.

152 **The acNMF approach produces an interpretable representation of a human**
153 **neuroblastoma single-cell RNA-seq dataset.**

154 Given the range of existing conclusions in the neuroblastoma literature, we decided to first
155 establish an initial baseline set of gene expression programs in patient tumors before
156 comparisons against preclinical models. To that end, we first applied acNMF to the published
157 "GOSH" dataset from Kildisiute *et al.*[17]. The five samples therein cover low-to-high risk
158 neuroblastoma and included one sample from normal adrenal gland. Previously, the insights
159 derived from this dataset were confined to the identification of six cell types. However, the
160 acNMF approach identified 41 gene expression programs (at k = 80; Fig. 1i), suggesting
161 possible new insights. Two co-authors (Y.K. and P.G.) blindly and independently annotated
162 these 41 programs. This was done using a web-based platform we developed, presenting each
163 annotator with a catalogue of relevant information, including the dominant gene weights
164 (loadings) for each program, program-program co-expression patterns, inferred copy number
165 profiles, gene ontology term enrichments, cell type annotations from SingleR[21], and 314
166 neuroblastoma-relevant marker gene sets that we curated from the literature[10,13-17,19,22-29]
167 (Supplementary Table S1). The web-based platform presented to annotators is available at
168 http://pscb.stjude.org.

169 The two annotators assigned qualitatively similar annotations to 37 out of 41 programs (Fig. 2a;
170 Supplementary Table S2). The annotations contained four broad groups of cell types,
171 neuroblastoma cancer cells, immune cells, stromal cells, and endothelial cells (Fig. 2a). We also
172 identified seven "activity" programs that captured processes such as the cell cycle and antigen
173 presentation by MHC. A total of 32 discrete cell types or cell states were recovered, with seven
174 assigned to the stromal group, three to endothelial, 19 to immune, and three to neuroblastoma
175 cancer cells, aided by inferred copy number data (Supplementary Fig. 2a-f; see Methods;
176 further details in subsequent subsections). We refer to these three neuroblastoma-specific

programs as “Adrenergic I (sympathoblast-like)”, “Adrenergic II (pre-neuronal-like)”, and “Neuroblastoma-MYC” (NB-MYC). All three neuroblastoma programs were strongly enriched for the classical cell line derived adrenergic marker genes (Fig. 3a-c; 8.1×10^{-66} for Adrenergic I, $P = 1.7 \times 10^{-72}$ Adrenergic II, 1.1×10^{-23} for Neuroblastoma-MYC from one-sided Wilcoxon rank sum test, signature defined by Van Groningen *et al.*[10]), but consistent with a subset of previous studies[16,17,19], were not strongly enriched for mesenchymal markers.

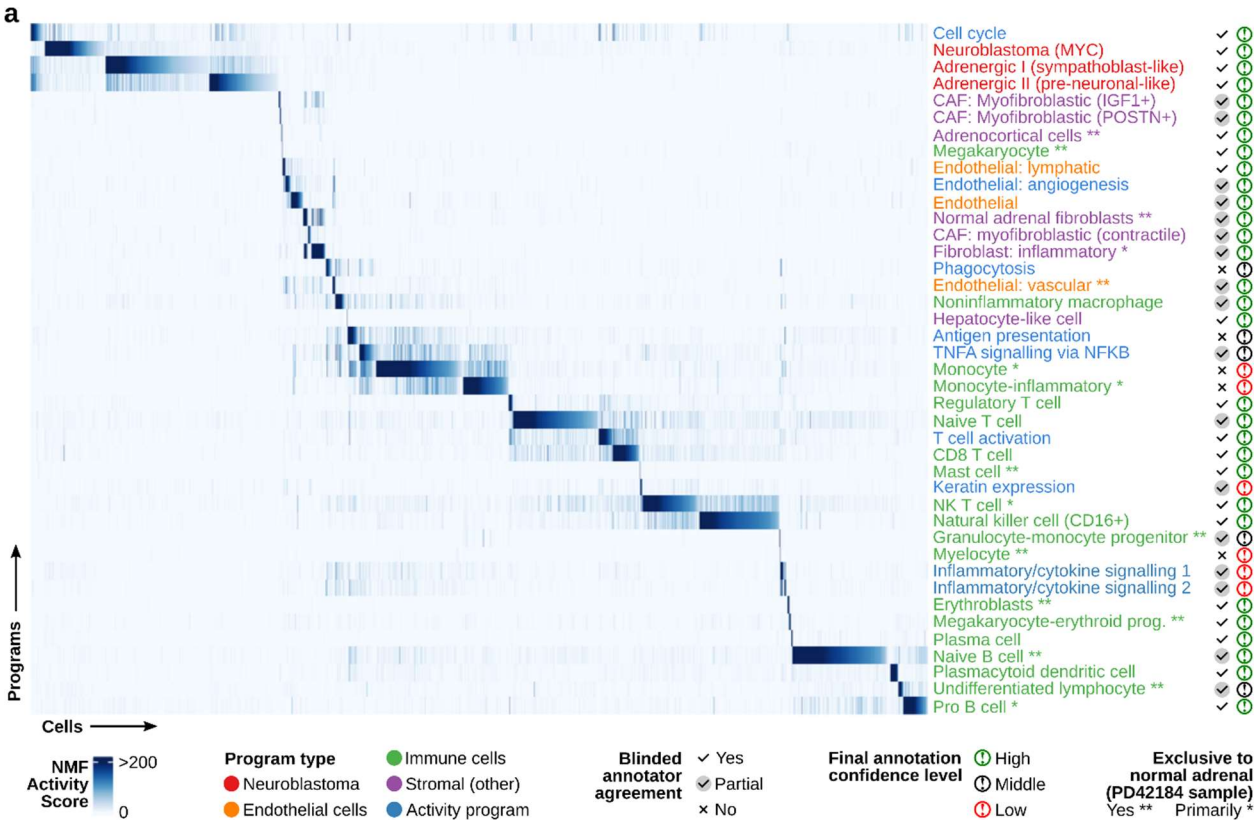


Figure 2: Gene expression programs recovered by acNMF in the GOSH neuroblastoma single-cell RNA-seq dataset.

(a) Heatmap showing the activity scores (color scale) for the 41 recovered programs (y axis) across each single cell (x axis) in the GOSH dataset. The final consensus annotation for each gene expression program is shown on the right, colored by the type of program. Tick marks indicate

189 whether the two blinded annotators agreed or not, and the far-right column indicates whether the
190 annotators reached a high, middle, or low level of confidence in the final annotation.

191

192 **The dominant neuroblastoma adrenergic programs resemble distinct aspects of**
193 **sympathoadrenal development.**

194 For all three adrenergic programs, the top ranked curated human cell-type signature enrichment
195 was sympathoblast (Fig. 3d, $P = 1 \times 10^{-6}$, 3.6×10^{-3}, 7.6×10^{-6} for adrenergic I, adrenergic II,
196 NB-MYC respectively, marker genes curated from Kameneva *et al.*[15]). Sympathoblasts are a
197 neural crest-derived precursor associated with the development of the sympathetic nervous
198 system and this result is consistent with a subset of the existing neuroblastoma single-cell RNA-
199 seq studies[15,17,19]. However, acNMF uncovered additional complexity beyond a clean
200 resemblance to sympathoblasts: Specifically, the Adrenergic I (sympathoblast-like) program was
201 defined by the expression of genes involved in the synthesis of norepinephrine in both
202 sympathoblast-precursors and differentiating chromaffin cells (Fig. 3e-g). This includes tyrosine
203 hydroxylase (*TH*), dopamine beta hydroxylase (*DBH*), dopa decarboxylase (*DDC*) ("Chromaffin
204 and progenitors" gene list curated from Jansky *et al.*[19]). Notably, markers specific to mature
205 chromaffin cells (*PNMT* and *PENK*) were not enriched, consistent with the argument that
206 neuroblastoma cells do not typically resemble differentiated chromaffin cells[26,30,31].

207 The Adrenergic II (pre-neuronal-like) program over-expressed markers of neuroblasts defined
208 by Hanemaaijer *et al.*[24] and Jansky *et al.*[19].The top five highly expressed marker genes included
209 neuronal-specific genes, neurofilament light chain (*NEFL*), neuronatin (*NNAT*), brain acid
210 soluble protein 1 (*BASP1*), and neurofilament medium chain (*NEFM*). This program can be co-
211 expressed with the Adrenergic I program (Fig. 3g & 3i, Supplementary Fig. 3a-c). Thus, these
212 programs were independent but not mutually exclusive.

213 The third recovered neuroblastoma program, NB-MYC, was only expressed in the high-risk
214 patient sample (PD43255) and was characterized by lower expression of marker genes of
215 differentiating chromaffin and neuronal cell types (Fig. 3d, 3f, Supplementary Fig. 3a),
216 consistent with the undifferentiated pathology of high-risk neuroblastoma[32]. The defining genes
217 of this program were dominated by translation initiation and elongation factors (Supplementary
218 Figs. 3d & 3e), consistent with high levels of transcription and translation supporting the hyper-
219 aggressive growth of high-risk tumors. The top marker genes of the adrenergic I & II programs
220 were expressed 2-3-fold lower in the high-risk tumor (Fig. 3j-k, $P = 6.9 \times 10^{-3}$ and $P = 5.9 \times 10^{-2}$
221 respectively). This is consistent with low/intermediate-risk tumors exhibiting differentiating
222 histology (Fig. 3l)[32]. Adrenergic marker genes were also strongly differentially expressed
223 between high-risk tumors and other tumors in 498 bulk RNA-seq samples[33] (Supplementary Fig.
224 S3f-h; $P = 1.8 \times 10^{-16}$, 1.8×10^{-5}, and 1.1×10^{-18} for Adrenergic I, II and MYC programs
225 respectively), suggesting these are generalizable patterns of differential gene expression.

226 While our study does not aim to definitively pinpoint neuroblastoma's cell-of-origin, these results
227 are consistent with the proposition that neuroblastoma tumors broadly resemble adrenergic
228 neuroblasts/sympathoblasts[15,17,19]. However, we additionally propose that these tumors can co-
229 express a mixture of independent adrenergic subprograms, resembling subtly different
230 sympathoadrenal cell types (Fig. 3g, i, Supplementary Fig. 3b-c), and that these activities differ
231 between high and low/intermediate risk disease. This suggests that the adrenergic co-
232 expression patterns in neuroblastoma patient tumors are more nuanced than previously
233 described. Evidence for these behaviors in other datasets and preclinical models is presented in
234 a later subsection.

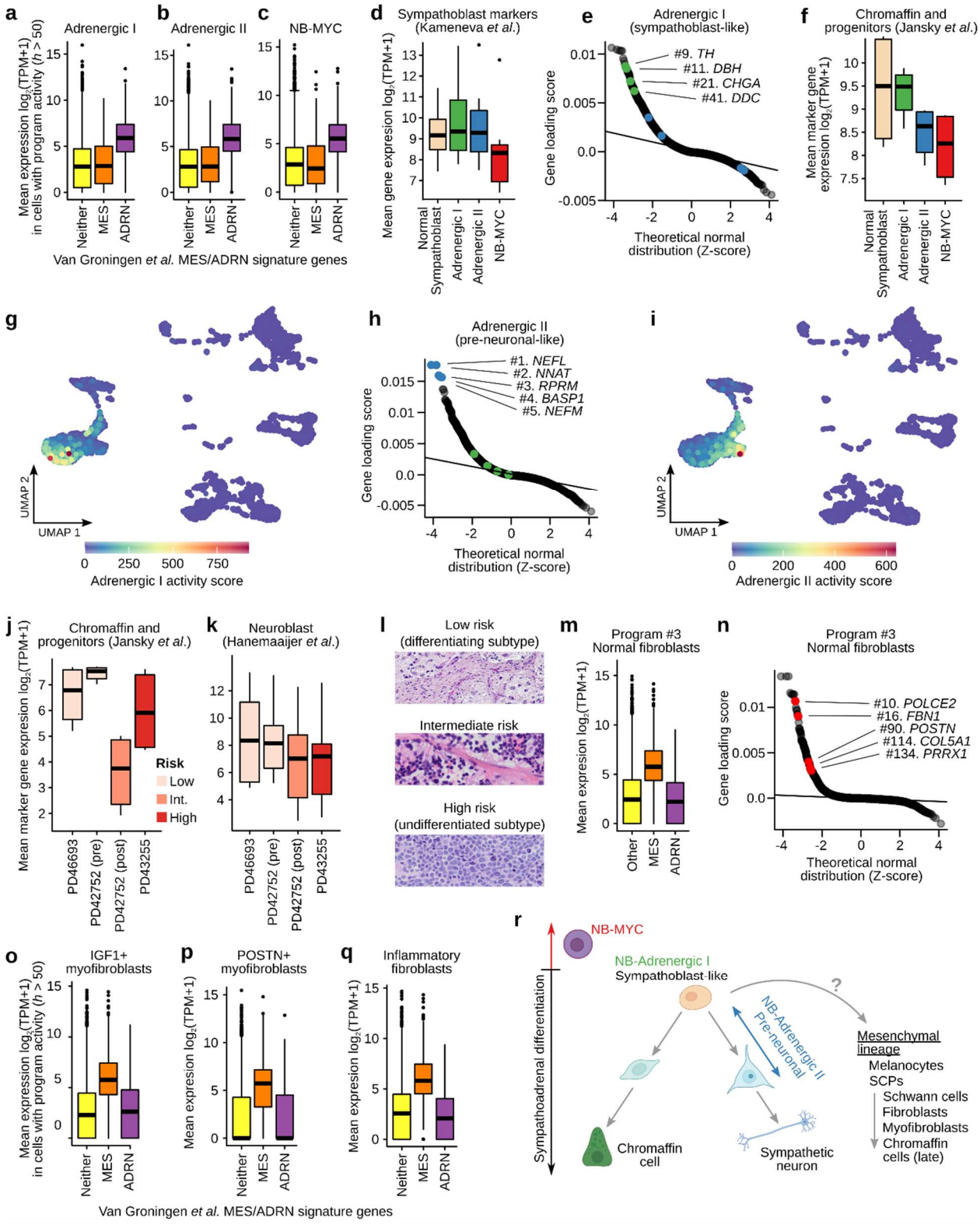
235

236 **Cancer associated fibroblasts subpopulations in the neuroblastoma tumor**
237 **microenvironment strongly express mesenchymal features, heavily resembling**
238 **mesenchymal-neuroblastoma cell lines.**

239 We identified nine gene expression programs in the GOSH dataset that statistically resembled
240 the neuroblastoma cell line-derived presumptive-drug-resistant mesenchymal gene expression
241 program. This includes some very strong enrichments (e.g., $P = 1.78 \times 10^{-88}$ for Program 40 for
242 enrichment of Van Groningen "mesenchymal" signature, Supplementary Table S2). This seems
243 to imply that mesenchymal neuroblastoma cells have been identified in these tumors. However,
244 one of these programs was expressed exclusively in fibroblasts of the reference normal adrenal
245 sample (Fig. 3m-n; $P = 3.75 \times 10^{-40}$ for mesenchymal enrichment) and thus this specific program
246 cannot be expressed in cancer cells. Hence, these shared gene expression features of
247 fibroblasts and mesenchymal-like neuroblastoma cell lines suggest a source of disagreement
248 about whether mesenchymal-like neuroblastoma cells exist *in vivo*. Indeed, in the four
249 neuroblastoma tumors, our acNMF decomposition again identified several programs strongly
250 resembling mesenchymal-like neuroblastoma cell lines. However, in all cases, we annotated
251 these as non-cancerous fibroblast cell populations rather than as mesenchymal-like
252 neuroblastoma cells (Fig. 3o-q; these annotations leveraged inferred copy number - details in
253 subsequent section). Notably, however, acNMF revealed a much more complex landscape of
254 cancer-associated fibroblasts (CAFs) than previously reported in neuroblastoma[28]. Specifically,
255 we identified IGF1+ ($P = 1.92 \times 10^{-7}$), POSTN+ ($P = 3 \times 10^{-15}$) and contractile myofibroblast
256 populations ($P = 2.5 \times 10^{-3}$). We also identified inflammatory CAFs ($P = 2.9 \times 10^{-9}$) expressing
257 interleukin (*IL6*, *IL33*) and complement component genes (*C7*, *C3*) using marker genes from
258 Lavie *et al.*[28]. Overall, our acNMF-based decomposition of the GOSH dataset yielded a
259 conceptual model with which to describe the behavior of neuroblastoma cancer cells in patient
260 tumors (Fig. 3r) but did not identify clear evidence for mesenchymal-like neuroblastoma cancer

261 cells. We next explore whether these initial characteristics are recapitulated in other human

262 datasets and the range of preclinical models.



263

264 **Figure 3: Cells expressing multiple adrenergic and mesenchymal-like gene expression programs are**
265 **evident in neuroblastoma tumors.**

266 (a) Boxplots showing the mean expression (y axis) of the genes from the classical Van Groningen
267 mesenchymal (MES; orange box) or adrenergic (ADRM; purple box) signature, in cells that are
268 highly expressing (activity $h > 50$) the acNMF-identified "Adrenergic I (sympathoblast-like)"
269 program in the GOSH dataset. The yellow box shows the mean expression of all other genes in
270 these cells.
271 (b) Like (a) but for the "Adrenergic II (pre-neuronal-like)" program.
272 (c) Like (a) but for the "Neuroblastoma-MYC" program.
273 (d) Boxplots showing the mean gene expression (y axis) of sympathoblast marker genes (curated
274 from Kameneva *et al.*) in cells highly expressing (activity $h > 50$) each of the 3 acNMF recovered
275 adrenergic programs, or in normal sympathoblasts (obtained from the Descartes Cell Atlas).
276 (e) QQ-plot showing the gene weights (loading scores) (y axis) for the "Adrenergic I (sympathoblast-
277 like)" program. Key sympathoblast marker genes have been highlighted in green, and key
278 neuronal marker genes in blue.
279 (f) Like (d) but for the "Chromaffin and progenitors" marker gene set, curated from Jansky *et al*.
280 (g) UMAP plot of the GOSH dataset; each point represents a single-cell, colored by the activity of
281 the "Adrenergic I (sympathoblast-like)" program.
282 (h) Like (e) but for the "Adrenergic II (pre-neuronal-like)" program.
283 (i) Like (g) but for the "Adrenergic II (pre-neuronal-like)" program.
284 (j) Boxplot showing the mean gene expression of "Chromaffin and progenitors" marker genes
285 (curated from Jansky *et al*.) in the cancer cells of the 4 neuroblastoma samples profiled in the
286 GOSH dataset. Boxes are colored by tumor risk classification.
287 (k) Like (j) but for the "Neuroblast" marker gene set curated from Hanemaaijer *et al*.

288 (l) Representative H&E stains of low, intermediate and high-risk human neuroblastoma tumors,
289 illustrating varying degrees of differentiated features.
290 (m) Boxplots showing the mean expression (y axis) of the genes from the classical Van Groningen
291 mesenchymal (MES; orange box) or adrenergic (ADRM; purple box) signature, in cells that are
292 highly expressing (activity $h > 50$) the acNMF-identified "Normal fibroblasts" program in the
293 GOSH dataset. The yellow box shows the mean expression of all other genes in these cells.
294 (n) QQ-plot showing the gene loading scores (y axis) for the "Normal fibroblasts" program. A subset
295 of genes classically used as markers of mesenchymal-like neuroblastoma cells are highlighted in
296 red.
297 (o) Like (m) but for the "IGF1+ myofibroblasts" program.
298 (p) Like (m) but for the "POSTN+ myofibroblasts" program.
299 (q) Like (m) but for the "Inflammatory fibroblasts" program.
300 (r) A schematic diagram overlaying the acNMF-recovered programs on normal sympathoadrenal
301 differentiation[34].

302

303 **The fidelity of neuroblastoma gene expression programs across human tumors and**
304 **preclinical models.**

305 Next, to assess the generality of the gene expression programs found in the GOSH dataset and
306 assess their fidelity in neuroblastoma preclinical models, we assembled five more of the publicly
307 available patient tumor datasets: from Dong *et al.*, Jansky *et al.*, Kameneva *et al.*, Verhoeven *et*
308 *al.*, and the PMC dataset from Kildisuite *et al.*[15-19,35], totaling 69 samples. Additionally, we
309 performed single-cell RNA-seq profiling across 13 neuroblastoma cell lines, 17 patient derived
310 xenografts (PDX), and 11 samples from a genetic mouse model (GEMM)[36] (metadata for this
311 (largest-of-its-kind) cohort is assembled in Supplementary Table S3, and in our web platform).

312 We applied the acNMF decomposition to each of these datasets independently, yielding
313 between 17 (cell lines) and 45 (Dong *et al.*) gene expression programs per dataset
314 (Supplementary Fig. 1t, Supplementary Table S4). To pairwise cross-compare the programs
315 from each dataset, we represented the programs as nodes on a graph. We drew an edge
316 between nodes if the program was replicated in another dataset (*FWER* < 0.01 by an
317 asymmetric Jaccard similarity test; Fig. 4a; Supplementary Fig. 4a; see Methods). A force-
318 directed algorithm facilitated the layout of the graph whereby interconnected nodes are found in
319 closer proximity than sparsely connected nodes. The resulting graph's structure indicated that
320 similar gene expression programs had often been recovered in multiple datasets, including
321 preclinical models. Consensus hubs (see Methods) were evident for gene expression programs
322 representing several immune cell types, neuroblastoma-specific programs, stromal cells such as
323 fibroblasts, and activity programs such as the cell cycle and hypoxia (Fig. 4a-d).

324 The adrenergic I (sympathoblast-like) program from GOSH was directly or indirectly connected
325 to programs in all other datasets, including all preclinical models (Fig. 4b). These connected
326 programs retained the characteristic hallmarks, including overexpression of genes like *DBH*, *TH*
327 and *CHGA* (Fig. 4e-h). The adrenergic II pre-neuronal-like program was also observed in
328 multiple other datasets (Fig. 4f & h) and the "neuroblastoma-MYC" (high risk) program formed a
329 dense hub (Fig. 4c), with similar programs recovered in high-risk (both *MYC* and *MYCN*
330 overexpressing) tumors in all datasets (Fig. 4i).

331 This cross-dataset analysis also revealed three groups of novel programs expressed in
332 neuroblastoma cancer cells. We were able to find a potential explanation for one, which was
333 expressed in *ALK* activated tumors (Fig. 4a; supported by one patient tumor and two PDXs but
334 no other preclinical model). Additionally, we found five programs unique to a single human
335 neuroblastoma tumor, implying a possible long tail of divergent neuroblastoma programs that
336 currently remain unexplained and are absent in our profiled preclinical models.

337 The mesenchymal-like programs observed in the GOSH dataset formed a large
338 mesenchymal/fibroblast subgraph. Interestingly, programs from cancer cell lines were also
339 connected to this fibroblast subgraph. For example, a program expressed exclusively in the
340 GIMEN cell line[10] was directly connected to the inflammatory fibroblasts program in the GOSH
341 dataset (Fig. 4d), again consistent with similar gene expression features in mesenchymal-like
342 neuroblastoma cells in culture and untransformed fibroblasts *in vivo* (Fig. 3o-q).

343 Overall, this cross-dataset-integrated map argues that, despite some disagreements in the
344 literature, neuroblastoma single-cell RNA-seq datasets are reasonably consistent and the
345 dominant adrenergic programs are preserved in all preclinical models. However, the status of
346 mesenchymal-like tumor cells and cell lines requires further exploration. Note: this graph-based
347 representation of these datasets (Fig. 4a) is available as part of our interactive, web-based
348 platform (accessible at http://pscb.stjude.org).

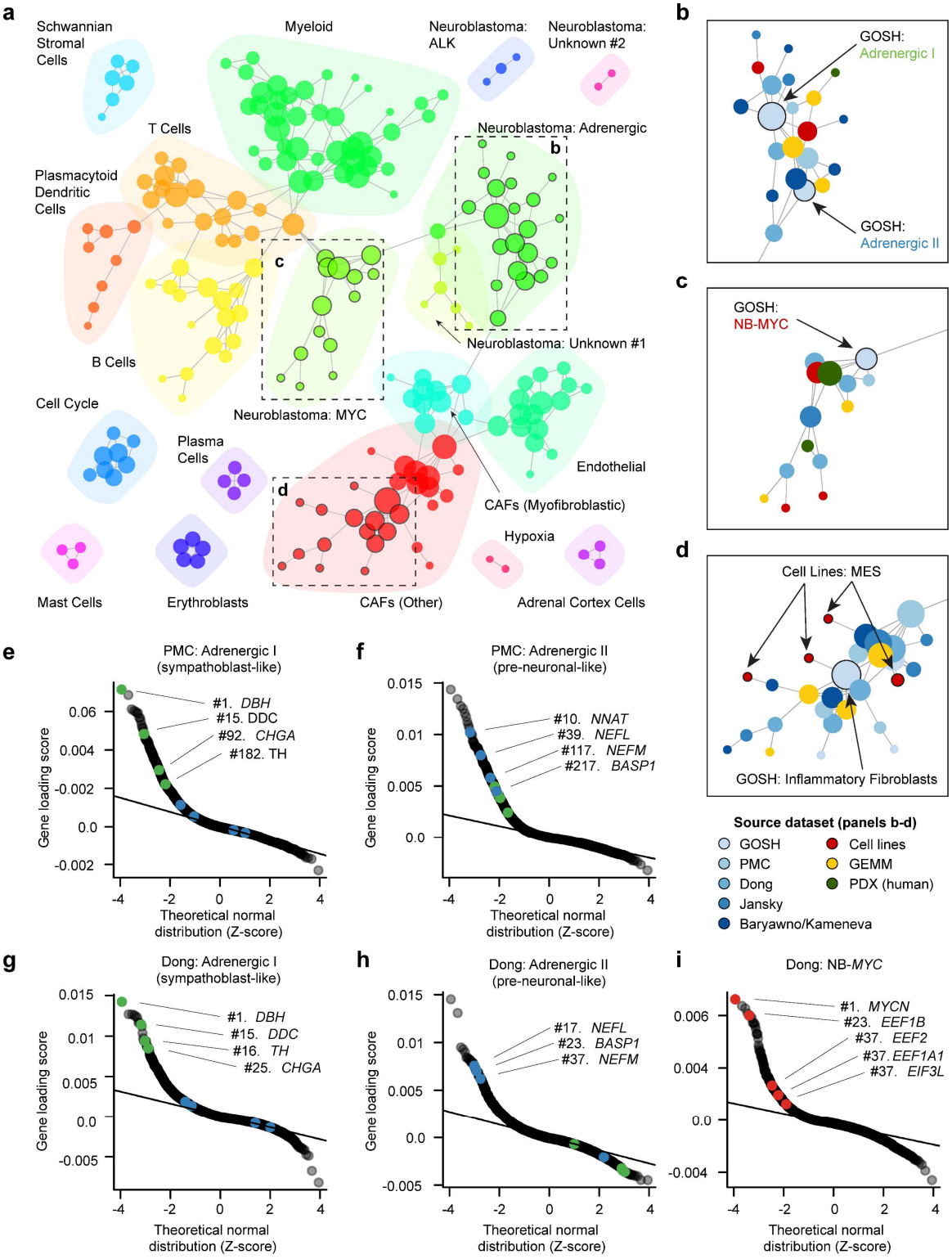


Figure 4: The fidelity of neuroblastoma gene expression programs across human tumors and preclinical models.

352 (a) A graph of gene expression programs identified by acNMF across eight neuroblastoma single-cell
353 RNA-seq datasets, from humans, PDX, GEMM and cell lines. Nodes represent a single
354 program's gene weights (i.e. loading scores), and node size is determined by the degree (i.e. # of
355 connections). Edges connect similar nodes by a Jaccard similarity test. The colors highlight
356 related communities of nodes. The black boxes highlight regions of interest, which are shown in
357 detail in panels (b-d).
358 (b) Inset showing a detailed view of region (b) from panel (a). This region has connected Adrenergic
359 I and Adrenergic II programs. In this inset, nodes have been colored by dataset, rather than
360 community (as in panel (a)).
361 (c) Like (b) but for a dense interconnected community of "Neuroblastoma MYC" programs.
362 (d) Like (b) but showing a region for a community of gene expression programs expressed in cancer
363 associated fibroblasts. A neuroblastoma program, exclusive to the GIMEN cell line, is connected
364 to this community of nodes.
365 (e) QQ-plot showing the gene weights (loading scores) (y axis) for the "Adrenergic I (sympathoblast-
366 like)" program recovered in the PMC dataset. Similar to Fig. 3e from the GOSH dataset key
367 sympathoblast marker genes have been highlighted in green, and key neuronal marker genes in
368 blue.
369 (f) Like (e) but for the "Adrenergic II (pre-neuronal-like)" program.
370 (g) Like (e) but for program 24 "Adrenergic I (sympathoblast-like)" in the Dong *et al.* dataset.
371 (h) Like (e) but for program 19 "Adrenergic II (pre-neuronal-like)" in the Dong *et al.* dataset.
372 (i) Like (e) but showing the "Neuroblastoma MYC" program from Dong *et al*. Top marker genes,
373 similar to those identified in the GOSH dataset, have been highlighted in red.
374 Note: An interactive version of the graph plot (panel a), and QQ-plots (panels e-i) for every program
375 in every dataset are available in our web platform.

376

377 **Cells highly expressing a mesenchymal-like program are consistent with non-cancerous**
378 **cells in all datasets except cell lines.**

379 Our cross-dataset analysis above also revealed other candidate mesenchymal-like
380 neuroblastoma cells, including Schwann-like cells (e.g., Dong dataset, Program 27, $P = 1 \times 10^{-9}$
381 for enrichment of Descartes Schwann cell markers, $P = 5.3 \times 10^{-43}$ for Van Groningen
382 mesenchymal enrichment; Supplementary Fig. 5a). We were interested in whether any of these
383 could plausibly represent *in vivo* mesenchymal neuroblastoma cancer cells.

384 We therefore applied InferCNV[37], a computational method for inferring copy number variant
385 (CNV) profiles from single-cell RNA-seq data, across the entire integrated largest-of-its-kind
386 cohort (see Methods). We found that the CNV profiles of mesenchymal-like cells were always
387 similar to the reference normal, whereas CNV aberrations in adrenergic cells were typically
388 obvious (Supplementary Fig. 2, Supplementary Figs. 5a-c). The exception was cell lines, where
389 both mesenchymal and adrenergic cells always exhibited massive CNV aberrations consistent
390 with cancer cells (Supplementary Fig. 5d-f).

391 However, because CNV-based inference tools act at the level of clustered groups of cells,
392 InferCNV cannot rule out a small proportion of mesenchymal-like cancer cells. We therefore
393 constructed a machine learning model that, by leveraging the inferred copy number profiles of
394 the high-confidence-cancer and high-confidence-normal cells, could *predict* the cancer/normal
395 status of the ambiguous mesenchymal-like cells on a per-cell basis (Supplementary Fig. 5g;
396 Supplementary Tables 5 & 6; see Methods). These models could predict the cancer/normal
397 status of the held-out cancer/normal-endothelial cells with median accuracy approaching 100%
398 (Fig. 5a, Supplementary Fig. 5h). We applied the trained models to the held-out ambiguous
399 population of mesenchymal-like cells (candidate mesenchymal neuroblastoma cells, cancer-
400 associated fibroblasts, or Schwann-like cells). The model never classified mesenchymal cells as
401 cancer cells beyond the model's misclassification rate estimated from normal endothelial cells

402 (Fig. 5b-c; Supplementary Fig. 5i-n). Thus, these analyses, given the scale of the data currently
403 available, did not provide convincing evidence for subsets of neuroblastoma cells that highly
404 express a mesenchymal-like gene expression program *in vivo*.

405 We validated these putative fibroblast and Schwann-like cells using an orthogonal assay—a
406 highly sensitive spatial transcriptomics assay, multiplexed RNA *in situ* hybridization (RNA ISH)
407 (Fig 5d, Supplementary Fig. 6a-c). We applied this in our GEMM tumors, a model that develops
408 neuroblastoma because of Cre-conditional activation of the *human* copy of *MYCN* in dopamine
409 beta hydroxylase (*Dbh*) expressing cells—thus, human *MYCN* expression should be excluded
410 from non-cancer cells. In all tumors, the human *MYCN* transcript was highly expressed across
411 most of the tissue, but absent from some large regions (Fig. 5e, Supplementary Fig. 7a-d) that
412 resemble Schwannian stroma (Fig. 5f, Supplementary Fig. 7e). These *MYCN*-absent regions
413 clearly co-expressed the myofibroblast markers *Tagln* and *Acta2* (Fig. 5e), consistent with the
414 idea that these mesenchymal cells are fibroblasts and not cancer cells. Regions excluding
415 *MYCN* also expressed the inflammatory fibroblast marker *C3* (Fig. 5e). Interestingly, we
416 identified some cells that seemed to co-express all combinations of chosen fibroblast markers,
417 consistent with the notion of intermediate subpopulations (Fig 5g). Schwann cell markers
418 *Sox10* and *Plp1* were also excluded from regions expressing human *MYCN* (Fig. 5h;
419 Supplementary Fig. 7e). Collectively, the data indicate that, despite strong transcriptional
420 resemblance to neuroblastoma mesenchymal cell lines, these *in vivo* cells are predominantly
421 Schwann-like cells and previously uncharacterized fibroblast subpopulations (see Discussion).

422

423

424

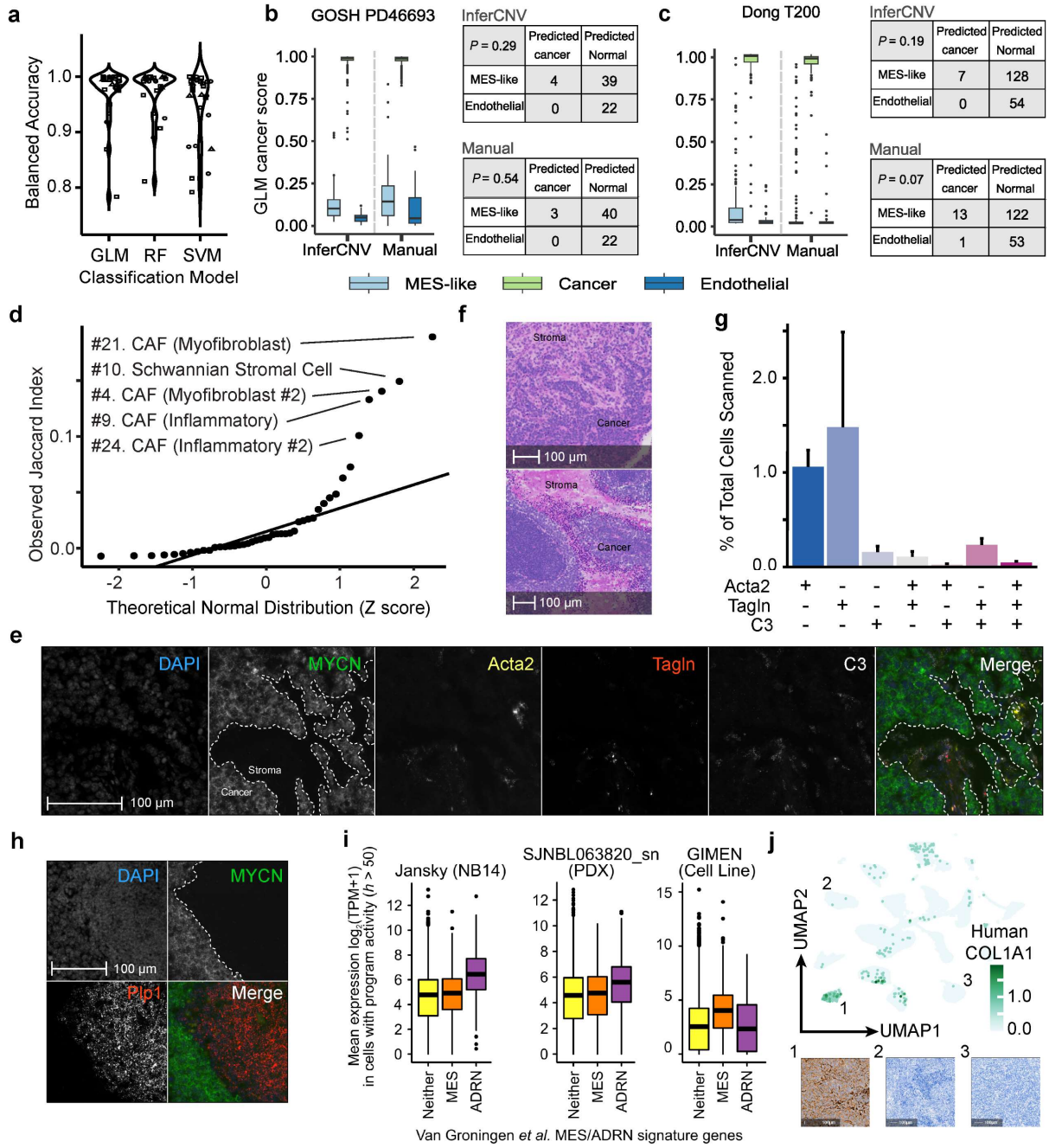
425 **Rare high-risk neuroblastoma tumors express a weakly enriched mesenchymal-like**
426 **signature *in vivo* that can be recapitulated in PDX.**

427 Interestingly, we identified some drug pretreated tumors that seem to represent a partial
428 exception to the above. In high-risk sample NB14 from the Jansky *et al.* dataset, the tumor's
429 cancer cells were broadly (but weakly) statistically enriched for the Van Groningen cell-line-
430 derived mesenchymal gene expression signature (Fig 5i; Program 12 $P = 1.7 \times 10^{-6}$;
431 Supplementary Fig. 8a-b). Notably, these cells were still predominantly adrenergic (Fig. 5i,
432 purple boxes). This is different from the behavior of cell lines, where the expression of
433 mesenchymal-like genes corresponds to a loss of expression of adrenergic genes[6,10] (Fig. 5i).

434 We initially suspected that this was artefactual, but, surprisingly, we also identified a high-risk
435 pretreated human tumor in our PDX cohort that exhibited a similar weak enrichment of the
436 mesenchymal signature across most cancer cells of the tumor (Fig. 5i; Program 1, $P = 1.6 \times 10^{-}$
437 ^{12} in SJNBL063820, Supplementary Fig. 8c-d). We obtained this PDX tissue[38] and used
438 immunohistochemistry to stain for COL1A1 protein, a classical neuroblastoma mesenchymal
439 marker[10]. Remarkably, greater than 95% of the neoplastic cells of SJNBL063820 expressed
440 COL1A1 in the cytoplasm and membrane, but staining was not evident in two negative control
441 PDXs (Fig. 5j, Supplementary Fig 8g). Thus, this weak-mesenchymal phenotype can be
442 recapitulated in PDX. Interestingly, two treatment-naïve high-risk *MYCN* amplified tumors (T200,
443 T230) from the Dong dataset, while still adrenergic (Supplementary Figs. 8e-f), also showed
444 broad overexpression of the transcription factor *TWIST1*, a known driver of epithelial-to-
445 mesenchymal transition[39] (ranked #9 on Program 11 (T200-specific) and #27 on Program 33
446 (T230-specific)). Curiously, though, both programs were also weakly but significantly co-
447 expressed with the mesenchymal-fibroblast Program 7 ($r_s = 0.07$ $P = 1.4 \times 10^{-33}$ and $r_s = 0.06$, P
448 $= 7.1 \times 10^{-25}$ respectively), also consistent with a very weak or proto-mesenchymal cell state
449 expressed in otherwise adrenergic cells. In general, these findings uncover a complex

450 landscape of mesenchymal-like gene expression programs *in vivo*, where the behaviors appear

451 different to cell line models (see Discussion) but can be maintained in PDX.



452

453 **Figure 5: The cancer/normal status of ambiguous mesenchymal-like cells in neuroblastoma tumors.**

454 a) Violin plots comparing the balanced accuracy (y axis) for prediction of cancer/normal status of
455 known cancer or immune cells for the 3 different initial models (x axis) across 13 different tumor
456 samples (shown for both "InferCNV" and "manual" features (see methods)). GLM, generalized
457 linear model; RF, random forest; SVM, support vector machine.
458 b) Boxplots showing the predicted probability of each cell being a cancer cell from the GLM model
459 (y axis) for the ambiguous-mesenchymal cells (light blue), for high confidence cancer cells held
460 out from the initial model fitting (green) and for normal endothelial cells held out from the initial
461 model fitting (dark blue). The predictions are shown for models using features defined based on
462 InferCNV (left 3 boxes) or from manual bins of 50 adjacent genes (right 3 boxes). The
463 contingency tables (right panels) estimate whether the number of mesenchymal-like cells
464 classified as cancer cells differs statistically from the misclassification rate of normal endothelial
465 cells, calculated by a Fisher's exact test. This panel shows a representative sample PD46693 from
466 the GOSH dataset.
467 c) Like (b) but for the Dong sample T200, all other samples are in Supplementary Fig. 5 and 8.
468 d) QQ-plot showing the Jaccard Similarity (y axis) of other programs with Program #37
469 (Myofibroblastic CAF) in our GEMM dataset.
470 e) Representative RNA ISH image of CAF markers in GEMM tumors. Neuroblastoma markers
471 (human *MYCN,* and *Ddc*) were used to define cancer cells. DAPI counterstaining confirmed the
472 presence of cells located in regions that lack *MYCN* expression. The co-expression of CAF
473 markers was identified within the *MYCN*-absent stromal regions (NB831 shown).
474 f) H&E staining of mouse tumors identifies regions of densely packed dark stained neuroblastoma
475 cells, and pink-staining stromal cells in our GEMM, which resemble the *MYCN*-absent regions of
476 mesenchymal marker expression in our RNA ISH experiments.
477 g) Bar plot showing the co-expression of CAF markers in GEMM tumors (y axis). All
478 combinations of CAF marker expression were quantified in stromal cells (x axis), and they are
479 represented as proportions of the total cell numbers identified in each sample.

480 h) Representative image of GEMM tumors stained with *MYCN* and the Schwann cell marker *Plp1* in
481 NB839. *Plp1* expression was restricted to *MYCN*-absent regions in these tumors.
482 i) Boxplot showing the mean expression (y axis) of genes in neuroblastoma cancer cells from the
483 sample NB14 from Jansky *et al*. Boxes are genes from the classical Van Groningen mesenchymal
484 (MES; orange box) or adrenergic (ADRN; purple box) signatures, with the yellow box showing
485 the expression of all other genes. The middle panel shows our PDX sample SJNBL063820 and
486 the right panels shows the predominantly mesenchymal neuroblastoma cell line GIMEN.
487 j) Immunohistochemical staining for the mesenchymal-marker gene COL1A1 in our SJNBL063820
488 PDX sample (#1, left) and in two other PDX samples SJNBL046145 (#2, middle) and
489 SJNBL047443 (#3, right). The upper panel shows a UMAP plot colored by the expression of
490 *COL1A1* in our PDX cohort, with the numbers indicating the UMAP clusters corresponding to
491 the 3 stained PDX samples.

492

493 **A mesenchymal-like gene expression program can be chemotherapy induced in**
494 **neuroblastoma cells *in vivo*.**

495 The confirmation of proto-mesenchymal features in some drug-pretreated tumors in multiple
496 datasets (e.g., NB14, SJNBL063820)—but not in baseline tumors—led us to hypothesize that
497 the weak mesenchymal-like programs could have been chemotherapy induced. To test this, we
498 treated five of our GEMMs with a single pharmacologically relevant dose of cisplatin (a mainstay
499 of neuroblastoma treatment) and treated six GEMMs with vehicle-control (see Methods). We
500 harvested these tumors 24 hours after treatment and performed single-cell RNA-seq (see
501 Methods; Supplementary Fig. 9a).

502 Consistent with a treatment effect, the drug-treated samples clustered separately from the
503 vehicle-treated samples (Fig. 6a). To determine which genes had been induced/repressed, we

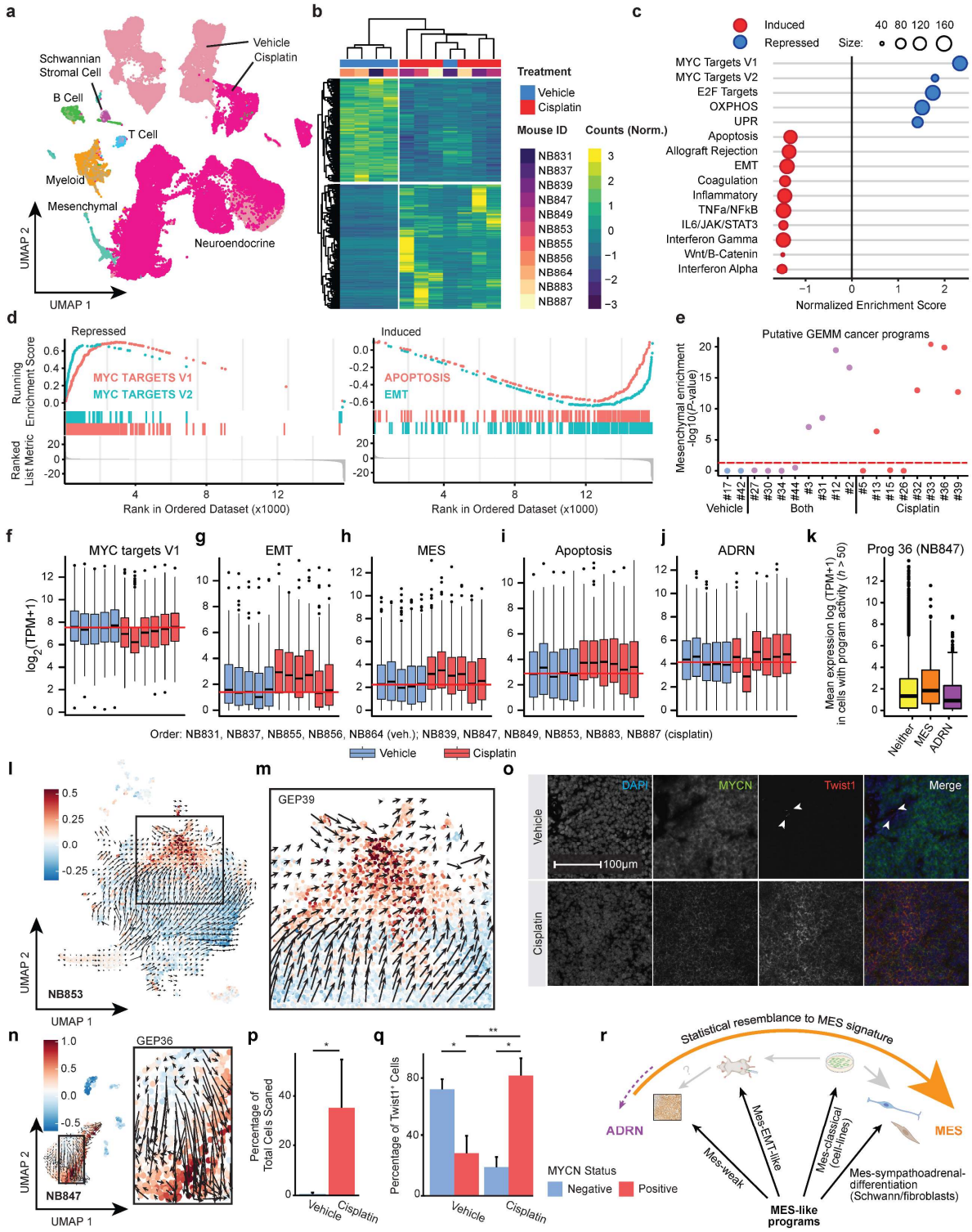
504 created “pseudo bulk” expression profiles using all of the cells of each tumor and then
505 performed functional enrichment (drug vs. vehicle) using GSEA, testing the mSigDB “Hallmark”
506 gene sets (Fig. 6b-d) (see Methods). As expected, “Apoptosis” was among the top
507 chemotherapy-induced gene sets (Fig. 6c-d and 6i, $P = 0.03$; Supplementary Table 7).
508 Remarkably, induced genes were also highly enriched for “Epithelial to Mesenchymal Transition
509 (EMT)” (Fig. 6c-d and 6g, $P = 7.4 \times 10^{-3}$). We probed this further using acNMF, where in the
510 drug treated samples, we identified (for the first time) clear mesenchymal-like programs
511 expressed in cancer cells (Fig. 6e). Additionally, cisplatin-treated samples were the first where
512 our machine learning model classified cells highly expressing a mesenchymal program as
513 cancer cells ($P < 2.2 \times 10^{-16}$ for NB853, $P = 9.5 \times 10^{-11}$ for NB847, Supplementary Figs. 9b-f).

514 While the repressed genes were quite consistent across all drug-treated samples, e.g., the loss
515 of the MYC target genes (Fig. 6b-d and 6f), the patterns of induced genes showed greater
516 variation (Fig. 6b and 6f-j, Supplementary Table 8 for *P*-values). The EMT signal was strongest
517 in NB839, NB847, NB849, and NB853 (Fig. 6g), where the classical Van Groningen
518 mesenchymal signature was also enriched (Fig. 6h). These tumors also had the highest
519 induction of apoptosis genes (Fig. 6i), with a loss of the adrenergic genes in NB847 (Fig. 6j)—
520 reminiscent of trends we had previously observed only in cell lines (Fig. 6k; Fig. 5i). We
521 performed RNA-velocity analysis in NB847 and NB853, the results of which were consistent
522 with the mesenchymal programs resulting from an acute cell state transition (Fig. 6l-n,
523 Supplementary Fig. 10a-e, Supplementary Figure 11a-c). Overall, these recently drug-treated
524 tumors demonstrate that a mesenchymal-like gene expression program can be chemotherapy-
525 induced in neuroblastoma.

526 Our GEMM, which expresses human *MYCN* in cancer cells, again allowed us to validate these
527 conclusions by an orthogonal RNA ISH assay. This time we assessed the co-localization of
528 human *MYCN* and the classical EMT marker *Twist1* across four drug- and four vehicle-treated

529 samples. While *Twist1* expression was evident to some degree in both the drug- and vehicle-
530 treated samples, the co-localization of *MYCN* and *Twist1* was much higher in drug-treated
531 samples, again consistent with the induction of a mesenchymal program (Fig. 6o and 6q).

532 Overall, the landscape of mesenchymal programs in neuroblastoma differs between cell lines
533 and other tumor models. In the *in vivo* setting, this involves weakly enriched expression in still
534 predominantly adrenergic cells (Fig. 6r), which our GEMM data showed could relate to a
535 mesenchymal transition during chemotherapy. Further assessment of the relevance of
536 mesenchymal programs in cell lines to the behaviors of tumors *in vivo* should provide a clear
537 basis for future research (see Discussion).



538

539 **Figure 6: An acutely chemotherapy treated neuroblastoma genetic mouse model strongly**

540 **overexpresses a mesenchymal-like program in cancer cells.**

541 a) UMAP plot of single-cell RNA-seq data obtained from the tumors of all 11 mice. Each point
542 represents a single cell, colored by whether it was obtained from a cisplatin-treated or a vehicle-
543 treated mouse. Brighter colored chroma indicates cisplatin treated animals.
544 b) Heatmap of normalized gene expression values highlighting differentially expressed genes
545 identified using DESeq2.
546 c) Dot plot showing GSEA enrichment scores for the top mSigDB "Hallmark" gene sets. Expression
547 in pseudobulk cancer cells was compared between (n=6) cisplatin treated mice and (n=5) vehicle
548 treated mice. A negative enrichment score implies the gene set was more highly expressed in drug
549 treated mice.
550 d) GSEA plots summarizing key induced and repressed gene sets. The colored tick marks show the
551 rank of the genes in the gene set of interest and the colored curves show the running enrichment
552 score.
553 e) Statistical enrichment of Van Groningen mesenchymal (MES) signature (y axis) in gene
554 expression programs (x axis) expressed in high-confidence cancer cells in our GEMM. P-values
555 were calculated using a Wilcoxon rank sum test (rank of MES genes vs all other genes).
556 f) Boxplot showing the pseudobulk expression in cancer cells from all 11 mice for genes in the
557 Hallmark gene set "MYC targets V1". Order of samples is NB831, NB837, NB855, NB856,
558 NB864 (vehicle treated, blue); NB839, NB847, NB849, NB853, NB883, NB887 (cisplatin
559 treated, red). The red line indicates the median of vehicle treated samples.
560 g) Like (f) but for Hallmark gene set "Epithelial to mesenchymal transition".
561 h) Like (f) but for the Van Groningen mesenchymal signature (MES) genes.
562 i) Like (f) but for the Hallmark gene set "Apoptosis".
563 j) Like (f) but for the Van Groningen adrenergic (ADRN) signature.
564 k) Boxplots showing the mean expression (y axis) of the genes from the classical Van Groningen
565 mesenchymal (MES; orange box) or adrenergic (ADRM; purple box) signature, in cells that are
566 highly expressing (activity $h > 50$) the acNMF-identified program 36, which has characteristics

567 consistent with cancer cells and is expressed primarily in cisplatin treated mouse NB847. The
568 yellow box shows the mean expression of all other genes in these cells.
569 l) UMAP plot showing RNA-velocity analysis of mouse NB853. The cellular trajectory, inferred
570 from splicing dynamics, shown by the arrows, is consistent with cancer cells actively
571 transitioning to a mesenchymal state, which is marked by the expression of the mesenchymal-
572 cancer Program 39 (color scale).
573 m) Detailed view of the box in (l).
574 n) Like (l) but for the sample NB847 and mesenchymal Program 36.
575 o) Representative RNA ISH images for co-localization of *Twist1* with human *MYCN* in vehicle
576 treated (top panels, NB855) and drug treated samples (bottom panels, NB887).
577 p) Bar plot showing the percentage of *Twist1*$^{+}$ cells in each of our sampled mouse slides (y axis),
578 quantified using the HALO software. (*, $P < 0.05$ by t-test, error bars represent SEM)
579 q) Bar plot showing the percentage of *Twist1*$^{+}$ cells (y axis) in either vehicle treated, or cisplatin
580 treated tumors that do or do not co express *MYCN* (colors). (*, $P < 0.05$; **, $P < 0.01$ by t-test)
581 r) Schematic representation of the distinct mesenchymal-like gene expression programs that can be
582 identified across neuroblastoma datasets and preclinical models. These programs have a varying
583 statistical resemblance to the classical cell line derived Van Groningen mesenchymal signature
584 (MES), but despite this, may capture qualitatively different aspects of neuroblastoma cell biology.

585

586

587 **DISCUSSION**

588 Multiple preclinical cancer models are commonly applied in drug discovery and mechanistic
589 experimental work. However, few formal methods have been developed to determine how these
590 various models represent/resemble primary patient tumors. Here, we developed a novel NMF-
591 based computational method and a web-based front-end user interface for integrating and
592 interpreting gene expression programs obtained from human tumors and preclinical models. We
593 used these tools to perform the first integrative analysis of single-cell RNA-seq data from human
594 neuroblastoma tumors and preclinical models, and we provide a new public resource for
595 interrogating neuroblastoma biology (http://pscb.stjude.org).

596 Using our tools, we identified novel neuroblastoma adrenergic programs mimicking the
597 sympathoadrenal developmental lineage and showed that neuroblastoma tumors co-express
598 marker genes of developmental cell types. These behaviors were preserved between human
599 tumors and preclinical models. We propose a conceptual mapping of neuroblastoma gene
600 expression programs to sympathoadrenal development in Fig. 3r, naming these adrenergic
601 programs “Adrenergic I (sympathoblast-like)”, “Adrenergic II (pre-neuronal-like)” and
602 “Neuroblastoma-MYC”.

603 We also found that the classical cell-line-derived mesenchymal-like program has differences
604 between human tumors and preclinical models. First, while the neuroblastoma mesenchymal
605 signature was clearly identifiable in neuroblastoma cell lines, we could not identify strong
606 evidence that a similar program is expressed in cancer cells in human tumors—neither by
607 inferred copy number nor by RNA ISH. Regardless, it is logically impossible to conclude the
608 non-existence of such cells. Indeed, they could be identifiable in other tumors, by profiling larger
609 numbers of cells, or by using more sensitive technologies. For instance, our RNA ISH
610 experiments have difficulty distinguishing cancer vs. fibroblast/Schwann-like cells at the
611 boundaries of stromal regions, it is conceivable some cell types could have dropped out of the

612 single-cell profiling, and tools for inferring copy number are imperfect[40]. Accordingly, a recent
613 large spatial atlas reported evidence of malignant Schwann-like programs in some
614 neuroblastoma tumors[41], thus, as the available data continues to grow, it is likely that further
615 nuances will continue to be discovered. However, despite the complexity, we can conclude that
616 mesenchymal-like neuroblastoma cells are likely less abundant than fibroblasts and Schwann-
617 like cells *in vivo*, which also strongly express a similar mesenchymal-like signature and thus, on
618 the basis of marker gene expression, are easily confused with neuroblastoma cells. These
619 analyses also identified novel subpopulations of cancer-associated fibroblasts in most
620 datasets/tumors, which we could validate using RNA ISH in our GEMM, revealing at least
621 inflammatory, myofibroblast, and, potentially, intermediate subpopulations. Similar cell types are
622 critical in regulating immunosuppression, metastasis, and progression in other cancers[28], and
623 determining the biological/therapeutic relevance of these new cell types in neuroblastoma will
624 form the basis for future research.

625 Lastly, in multiple datasets we identified rare, high-risk, drug pretreated tumors that weakly
626 express mesenchymal-like features in cancer cells that remain predominantly adrenergic. This
627 contrasts with the behavior in cell lines, where cells expressing a mesenchymal state lose
628 adrenergic features. This may be related to the concept of "high risk tumors with mesenchymal-
629 like features", which was recently proposed by Jansky *et al*[19], but in contrast to Jansky we
630 showed these cells at least partially resemble the classical Van Groningen mesenchymal
631 signature and that this could be maintained in PDX. We also identified a potential source of this
632 phenotype, showing that a mesenchymal-like program could be strongly induced *in vivo* in an
633 acutely chemotherapy-treated GEMM. This is consistent with observations across several other
634 cancers that EMT can represent an apoptosis escape mechanism that leads to drug
635 resistance[42,43]. Notably, our observation of the loss of expression of MYC target genes and gain
636 of EMT/mesenchymal-like features in cancer cells also resembles a previously reported

637 “embryonic diapause-like” drug/chemo persistent state change in breast cancer[44]. Our data
638 suggest that high-risk neuroblastoma cells may be primed to undergo a similar transition. It is
639 tempting to speculate that targeting this process could circumvent chemoresistance, and it is
640 plausible that some mesenchymal-like cell lines may represent a good model for this behavior.
641 Because of the very short 24-hour treatment window, this finding also implies that the
642 expression of mesenchymal-like genes need not necessarily emerge in relapsed neuroblastoma
643 tumors by purifying selection, as was previously speculated based on patient data[6,10,45] and
644 tumors obtained at later time points post-chemotherapy and -relapse in mice[45]. However, the
645 clinical significance of these drug-induced and weakly enriched mesenchymal programs and
646 their potential role in facilitating escape from chemotherapy-induced apoptosis remains to be
647 determined. In general, our results argue that mesenchymal-like programs in neuroblastoma
648 can be subdivided into at least (1) a weak mesenchymal-like signature co-expressed in still-
649 predominantly-adrenergic cells of some high-risk tumors, some of which (2) could potentially be
650 chemotherapy induced; (3) the classical mesenchymal-like signature in cell lines, which also
651 strongly express features similar to (4) cancer-associated fibroblasts and Schwann-like cells
652 found in tumors. As these represent (at least partially) qualitatively different entities, we propose
653 naming these (1) Mes-weak, (2) Mes-EMT-like, (3) Mes-classical, and (4) Mes-
654 sympathoadrenal-differentiation (Fig. 6r).

655 Finally, our findings only touch on the biological insights that could be gained from an
656 interpretable integrated representation of single-cell RNA-seq data from human neuroblastoma
657 and preclinical models. For example, our NMF-based models recovered a rich representation of
658 the immune component of neuroblastoma tumors, largely recapitulated in our GEMM. We have
659 made all the data and models available in our interactive web-based platform
660 (http://pscb.stjude.org), which we encourage readers to review, discuss, and use as a map for
661 future discoveries.

662 **CONCLUSIONS**

663 We have developed a set of computational methods for integrative analysis of human single-cell
664 RNA-seq datasets and pre-clinical models. We have used these tools to identify the divergent
665 behavior of gene expression programs in neuroblastoma tumors and preclinical models,
666 providing insight into the behavior of putative drug resistance mesenchymal-like programs. We
667 have provided this dataset as an integrated community resource and our computational
668 approach is applicable in other diseases, where the fidelity of pre-clinical models remains
669 largely unknown.

670 **METHODS**

671 **Establishing a neuroblastoma genetic mouse model.**

672 LSL-MYCN;Dbh-iCre mice[36] (129X1/SvJ) were obtained from Michael A. Dyer's lab (St. Jude).
673 This study was carried out in strict accordance with the recommendations of the National
674 Institute of Health's Guide to Care and Use of Laboratory Animals. The protocol was approved
675 by the Institutional Animal Care and Use Committee at St. Jude Children's Research Hospital.
676 All mice were housed, per approved IACUC protocols, on a 12−12 light cycle (light on 6 am, off
677 6 pm) and provided food and water ad libitum. Neuroblastoma-like tumors comparable to
678 human neuroblastoma arose in the abdomen from either the adrenal gland or the celiac and
679 superior cervical ganglia with a penetrance of approximately 50%, at a median age of about 6
680 months old. The tumor volumes were monitored by ultrasound using the Vevo 3100 system
681 (Fujifilm-Visualsonics, Toronto, Canada), and once an adrenal tumor had reached 500mm^3 in
682 diameter, the mice were treated with either a single, pharmacologically relevant[46] 2 mg/kg dose
683 of cisplatin (n = 6) or vehicle control (n = 5) by intraperitoneal injection. The mice were
684 euthanized at 24 hours, and tumors were immediately harvested and subjected to single-cell
685 RNA-seq using the protocols described below.

686

687 **Orthotopic patient-derived xenografts**

688 Flash-frozen fragments of orthotopic patient-derived xenografts were obtained through the St.
689 Jude Childhood Solid Tumor Network (https://cstn.stjude.cloud/)[38,47]. Approximately 50-100 mg
690 of frozen tissue underwent nuclei extraction using manual disruption within a chilled Tween
691 containing buffer[48]. To generate fresh xenograft tissue, cells were injected into the para-adrenal
692 space of immunocompromised nude mice using ultrasound. Tumors were harvested from

693 xenograft-bearing mice, and approximately 250-500 mg of tissue were dissociated using the
694 Papain Dissociation System (Worthington Biochemical)[48].

695

696 **Single-cell/single-nuclear RNA-seq data generation and preprocessing**

697 *Mouse:* Single cells from mouse tumors were obtained through dissociation using the Papain
698 Enzymatic dissociation method following the Worthington kit (cat. No. LK003150), and final
699 dissociated cells were suspended in the DMEM medium.

700 *Cell Lines:* The following NB cell lines were maintained in the following culture conditions (5%
701 CO_2, 37° C, RPMI medium, 10% FBS, 2mM L-glutamine, 100 IU/mL penicillin, 100 μg/mL
702 streptomycin); BE2C, BE2M17, CHP134, CHP212, GIMEN, IMR5, KELLY, MHHNB11, NGP,
703 SKNAS, SKNFI, SKNSH, TGW. Cells were harvested, and viabilty was measured using Lunc
704 Cell Counter AO/PI dye. We processed samples that showed greater than 90% viability.

705 *Library Preparation and Sequencing:* The dissociated/suspended single cells were filtered and
706 strained through a 40μm strainer and cells were counted using the Luna cell counter. Next, they
707 were loaded into Chromium Chips from Single cell 3' Gene expression V3.1 kit (10X Genomics,
708 Pleasanton, CA) with target capture of 6000 cells. The cDNA generation, library preparation,
709 and QC was performed following 10X Genomics protocol. The library was sequenced in
710 Novaseq-V1 reagents, and the sequenced data were processed using the Cell Ranger Software
711 (10X Genomics) mapped to mm10 (mouse) or GRCh38 (cell line). Poor quality and outlier cells
712 were identified and removed. Analysis downstream of acNMF, was conducted on normalized
713 read counts (SCTransform, Seurat R package) of the resulting cells.

714

715

716 **Automatic Consensus Non-negative Matrix Factorization (acNMF)**

717 For each dataset, raw read counts were merged into a single matrix containing all samples.
718 50,000 cells were subsampled per dataset, and then randomly split into two subsets containing
719 roughly equal numbers of cells. cNMF[20] (https://github.com/dylkot/cNMF/) was applied to each
720 split independently by performing 200 NMF iterations per rank *(k),* spanning a range of k = 2 to k
721 = 200. Outlier replicates, defined by Euclidean distance, were filtered at a rate of 10%. The
722 consistency of gene expression programs between the data splits was assessed by calculating
723 the Jaccard similarity index of *Z*-score normalized (see Kotlier *et al.*[20] for details) gene weight
724 coefficients at each value of *k*. *P*-values for all pairwise comparisons between the columns of
725 the component matrices were calculated using the jaccard.test.mca function from the
726 'jaccardtest' R package[49] followed by Bonferroni multiple testing correction. This was carried out
727 over a range of Jaccard lengths. All significant pairs ($P < 0.05$) of components between data
728 splits were then represented as node pairs in an undirected network graph using the R package
729 *igraph*[50]. Community detection was performed on the underlying graph using edge
730 betweenness centrality, and the number of distinct communities were recorded. These
731 communities represent replicated gene expression programs between data splits, and these
732 were used as the basis of comparison to determine optimal parameters for rank (k) and Jaccard
733 length on each dataset. For each Jaccard length the number of discrete communities was
734 plotted against all values of k, and a loess regression was performed to facilitate curve
735 smoothing. The inflection point of these concave curves was calculated using the 'elbow' R
736 package (https://nicolascasajus.fr/elbow/). Cost values, representing residual values after the
737 inflection point, were used to evaluate the stability of the solution. A linear regression was
738 performed on cost values of each Jaccard length curve, and the slope was recorded. The
739 optimal Jaccard length was established by determining which slope was closest to -1.0, which
740 represents a solution whereby the number of reproducible programs remains approximately

741 constant for all values of k greater than ground truth (Supplementary Fig. 1d). Using the
742 established parameters for rank (k) and Jaccard length, final reproducible programs for each
743 dataset were calculated by averaging over all nodes (i.e., cNMF gene loading scores (i.e.
744 weights)) per community in the network graph.

745

746 **Benchmarking acNMF against other methods**

747 We benchmarked acNMF against 3 different published latent factor models, iNMF, ScVI, and
748 SignatureAnalyzer, using our simulated dataset of 15,000 cells, encompassing 13 cell types
749 plus 1 activity program. We benchmarked all models using the same 2000 highly variable
750 genes. iNMF was run with miniBatch_size = 5000, max.epochs = 5, where we first ran the
751 suggestK function at lambda = 5, thresh = 1e-4 and max.iters = 100 with values of k ranging
752 from 5 to 100 (from 5:60 in intervals of 5, from 70:100 in intervals of 10). We then determined
753 the optimal value of k by LOESS (same approach as in acNMF) on the resulting median KL
754 divergence data. ScVI was run with max_epochs = 200, gene_likelihood = “nb” and
755 latent_distribution = “ln”. k was chosen based on the marginal log-likelihood obtained with the
756 dimensionality of the bottleneck layer ranging from 2 to 100 (from 3:60 in intervals of 1, from
757 70:100 in intervals of 10), where the encoder and decoder were both structured with 2 layers
758 (note that the marginal log-likelihood fluctuated with increasing dimensionality and we picked
759 the smallest dimensionality with a comparable marginal log-likelihood). SA-GPU was run with
760 K_0 = 100, max_iter = 100000. For ARI (adjusted rand index) calculation, we labeled each cell
761 with the program associated with the highest program expression score (h) and ARI was
762 calculated by ARI function in R library aricode.

763

764 **Identifying gene expression programs replicated across different datasets using a**
765 **network graph**

766 To determine which gene expression programs were replicated across datasets, or between
767 human tumors and preclinical models, all pairwise comparisons between programs from all
768 datasets were performed using an asymmetric Jaccard similarity test, fixing the Jaccard length
769 to that tuned for each dataset above, followed by Bonferroni correction (keeping connections at
770 *FWER* < 0.05). We visually represented connected nodes in a network graph (e.g., Fig. 4a) –
771 the layout of which was established using the force-directed Fruchterman-Reingold algorithm to
772 minimize overlapping of edges. The network structure facilitated the interpretation of these
773 programs, with communities detected by modularity scores calculated from random walks of the
774 graph.

775

776 **Web-based platform for annotation, exploration, and cross-dataset integration of acNMF**
777 **output.**

778 For each gene expression program, an interactive web-based report was created, which was
779 used to support annotation and interpretation. To estimate variability between subjective
780 annotators, the GOSH dataset was independently annotated by co-authors P.G. and Y.K, and a
781 consensus annotation was subsequently reached by discussing and merging these annotations,
782 including subjective assignment of confidence levels. The web-based reports were created
783 using the R package *knitr*[51], which for each program summarized a rich catalogue of biologically
784 relevant information/enrichments, including 282 marker gene lists we curated from the literature.
785 The interactive plots were created using *plotly*[52]. The broader network-based representation
786 comparing gene expression programs across datasets, as well as the tools for organizing the

787 reports and cataloguing/displaying relevant metadata was created using *Shiny*[53]. The complete
788 web platform is accessible at http://pscb.stjude.org.

789

790 **RNA *in situ* hybridization (RNA ISH) and HALO analysis**

791 RNA ISH was performed in fresh frozen mouse tissue samples (NB837, NB839, NB853, NB855,
792 NB856, NB864, NB883, NB887) using the RNAscope HiPlex V2 assay Reagent Kit. These data
793 were generated by Advanced Cell Diagnostics (7707 Gateway Blvd., Newark, CA 94560, USA).
794 Briefly, fresh frozen tissue sections were pretreated with 10% NBF and protease prior to
795 hybridization with the target oligo probes (Mm-Tagln, Mm-Acta2, Mm-C3, Mm-Sox10, Mm-Plp1,
796 Hs-MYCN, Mm-Twist1, and Mm-Ddc). Immediately following hybridization, the RNAscope
797 Fluorescent V2 Assay was performed with the preamplifier, amplifier, and RNAscope HiPlex
798 FFPE Reagent to reduce autofluorescence. Fluorophores T1-T4 were then hybridized
799 sequentially, followed by RNAscope DAPI counterstain. After imaging, the first group of
800 multiplexed fluorophores were then cleaved, and the process was repeated for each additional
801 group of fluorophores. Fluorescence images were acquired using a Polaris 40X full-slide digital
802 slide scanner. The resulting multispectral images were then uploaded into HALO Image
803 Analysis platform v3.5.3577.140 (Indica Labs), and the High-Plex FL v2.0 module was used to
804 construct the image analysis algorithms. These included nuclei detection, cytoplasmic
805 segmentation, and FISH detection criteria based on intensity thresholds for each probe in each
806 sample.

807

808 **CNV Inference in scRNA-seq data**

809 We used InferCNV[37] to identify regions of copy number variation from single-cell RNA-seq,
810 which we used as evidence to differentiate between cancer and non-cancerous “normal” cells.

811 InferCNV was run on each patient/sample independently. Reference normal cells were selected
812 from clear immune cells, identified by acNMF (selecting cells expressing an immune program at
813 $h > 100$). In the case of the GOSH dataset, we used the author's original immune cell
814 annotations, which were essentially identical to the acNMF annotations. In cell lines, we defined
815 the reference normal using immune cells from the Dong dataset. InferCNV's six-state Hidden
816 Markov Model (i6) predicted CNV regions and states, with posterior probabilities of CNV regions
817 assigned to a given state using a Bayesian network at a value of P(Normal) = 0.3 (all other
818 parameters were used at default settings).

819

820 **Machine learning classifiers for predicting cancer/normal status of mesenchymal-like**
821 **cells using inferred copy number profiles.**

822 To predict the cancer/normal status of mesenchymal-like cells on a per-cell basis, we tested 3
823 classification models, a regularized logistic generalized linear model (GLM), a random forest
824 (RF), and a support vector machine (SVM), which all achieved comparable performance. These
825 were implemented in the R packages *glmnet*, *randomForest,* and *SVM* (linear kernel)
826 respectively. We identified cancer cells, endothelial cells, or ambiguous mesenchymal-like cells
827 if an appropriate program was expressed in that cell at $h > 100$. We randomly split clear cancer
828 cells into 75% training set and 25% test set. Cell-specific copy number estimates, which were
829 treated as the model's input features, were obtained from inferCNV's predicted regions of copy
830 number variation (inferCNV-features) or were obtained manually, by taking the average
831 standardized TPM expression level in groups of 50 collocated genes along the genome
832 (manual-features). For each sample, a classifier was then trained using sample-specific cancer
833 cells and immune cells (representing normal cells). For each sample, these trained models were
834 then applied to the held-out reference normal endothelial cells, or the held-out ambiguous
835 mesenchymal-like cells, outputting a predicted class of 0 (normal) or 1 (cancer) for each cell.

836 These results were then assembled into a 2 × 2 contingency table, where we applied a Fisher's
837 exact test to compare the frequency with which the model classified mesenchymal-like cells as
838 cancer cells, against the frequency it misclassified reference normal endothelial cells as cancer
839 cells. Note that for model training/testing in the GEMM data, because the mice were genetically
840 identical, we pooled the immune cells and endothelial cells in the drug-treated and vehicle-
841 treated groups, which increased the size of the reference normal pool of cells and should
842 improve power.

843

844 **Differential expression, gene set enrichment analysis, and RNA-velocity analysis of**
845 **cisplatin treated GEMM.**

846 We performed differential expression analysis of our cisplatin treated vs vehicle treated GEMM
847 using the R package DESeq2[54] by 'pseudobulking' cells from each mouse using the
848 aggregate.Matrix function in the R package Matrix.utils. All genes were ranked by logFC
849 between vehicle and cisplatin treated mice, and then were subjected to GSEA[55] for functional
850 enrichments of the mSigDB Hallmarks gene set, using default parameter settings. Trajectory
851 inference and pseudotime analysis were performed with the R package Monocle3[56]. RNA
852 velocity was assessed by applying the dynamical module of scVelo[56] (v0.2.5) under default
853 settings.

854

855 **Spatial pathology analyses and immunohistochemistry.**

856 For H&E staining, the harvested mouse tumor tissue was embedded in Tissue-Tek O.C.T.
857 embedding medium (Sakura Finetek) and flash frozen on a metal block in liquid nitrogen then
858 preserved at –80°C until further processing. Each tissue block was equilibrated to -20°C in a
859 cryostat chamber (CM3050S, Leica BioSciences) for 30 minutes before making 10-μm sections

860 collected onto Superfrost slides. The slides were incubated at 37°C for 1 minute and transferred
861 to a container containing pre-chilled methanol for 30 minutes at -20C in the upright position.
862 After methanol fixing, the excess methanol was wiped out from the bottom of the slides without
863 disrupting the tissue section. After methanol fixation, the slides were immersed into water 10
864 times gently and incubated in hematoxylin solution for 1 minute followed by immersing slides 10
865 times into 800 ml water. The slides were then submerged 10 times into Bluing buffer and
866 immersed into water to remove the excess Bluing buffer. Next, slides were dipped into 96%
867 ethanol 10 times followed by incubation with Eosin for 30 seconds. The excess Eosin was
868 removed by immersing slides 10 times into 3 different containers containing 100% ethanol. After
869 that, tissue slides were mounted, and images were acquired using Axioscan slide scanner
870 (Zeiss, Oberkochen, DE) up to a 20X scalable magnification. For our immunohistochemistry
871 analysis of fresh frozen slides used for the RNA ISH experiments serial unstained sections of
872 murine NB-like tumors were fixed in acetone for 20 minutes. Fixed slides were then wet loaded
873 on the Ventana Discovery Ultra autostainer (Roche, Basel, CH) using the Ventana Discovery
874 Ultra antigen retrieval CC1 (catalogue #950-500) for 64 minutes followed by the application of
875 anti-PHOX2B antibody (Abcam, ab183741) at a 1:100 dilution for 60 minutes. OmniMap anti-
876 rabbit HRP (catalogue #760-4311) for 16 minutes, followed by the ChromoMap DAB kit
877 (catalogue #760-159) for 8 minutes, followed by counterstained with hematoxylin, and the post-
878 counterstain Bluing Reagent (catalogue #760-2037) for 8 minutes was used for the detection
879 and visualization of PHOX2B in tissue sections.

880 Xenograft tissue was fixed in 10% neutral-buffered formalin and processed as paraffin
881 embedded samples. Tissues were sectioned at a 4-µm thickness and mounted onto positively
882 charged glass slides (Superfrost Plus; ThermoFisher Scientific, Waltham, MA). Unstained
883 slides were baked for 60 minutes at 60ºC. After deparaffinization, slides underwent
884 immunohistochemical staining for anti-COL1A1 (clone E3E1X, Cell Signaling, #66948) was

885 performed on the Leica BOND-MAX automated stainer (Leica Biosytems, Buffalo Grove, IL)
886 with IHC Protocol F. Heat-induced epitope retrieval was carried out with Bond Epitope Retrieval
887 Solution 2 (ER2) for 20 minutes followed by incubation of the primary antibody at 1:200 for 30
888 minutes, and visualization with Bond Polymer Refine Detection kit (DS9800). Negative and
889 positive tissue controls included xenografts defined as having either negative or high expression
890 of COL1A1 as determined by sequencing data. Species cross-reactivity was not observed in
891 wild-type B6 murine haired skin, which served as another negative tissue control. An isotype
892 control (Mouse BALB/c IgG2a, BD Pharmingen, 349050) was used to confirm the specificity of
893 immunoreactivity for the monoclonal antibody.

894

895 **DECLARATIONS**

896 **Ethics approval and consent to participate**

897 Not applicable.

898

899 **Consent for publication**

900 Not applicable.

901

902 **Availability of data and materials**

903 Simulated scRNA-seq datasets were obtained through CodeOcean
904 (https://doi.org/10.24433/CO.9044782e-cb96-4733-8a4f-bf42c21399e6) using the
905 Step1_Simulate.ipynb Jupyter Notebook. Default parameters, defined by the implementation of
906 the R package 'splatter' on a published scRNA-seq dataset, were used to generate five replicate

907 simulations across three conditions of varying signal:noise levels. The raw readcounts from
908 each of these simulations were subjected to the acNMF approach described below
909 (https://rchapple2.github.io/acNMF/). The gene and cell parameters generated from this code
910 were used to evaluate the recovery of programs following the acNMF analysis. Human
911 datasets were downloaded via the following accession numbers. Dong: GSE137804. Jansky:
912 EGAS00001004388. Kildisiute: EGAD00001008345. Cell line and mouse data deposited in
913 GSE229224 and GSE229226, respectively. PDX deposited in GSE228957.

914

915 **Competing interests**

916 The authors declare no competing interests.

917

918 **Funding**

919 PG is supported by an NIGMS R35 award [R35GM138293] an R01 grant from NCI
920 [R01CA260060]; K99/R00 [R00HG009679] from NHGRI; PG also receives support from
921 ALSAC. CWL received funding from the St. Jude Graduate School of Biomedical Sciences. MA
922 was supported by R25CA023944 from the National Cancer Institute. The content is solely the
923 responsibility of the authors and does not necessarily represent the official views of the National
924 Institutes of Health. The authors have declared that no conflict of interest exists.

925

926 **Authors' contributions**

927 RHC and PG developed the methods, analyzed the data and drafted the paper. XL developed
928 the CNV calling method and acNMF benchmarking. SN and JE generated the single-cell data.
929 MIMA, CWL, WCW, YZ performed additional analytical work. YK, MP, HML, ML, SCK, CL, JH,

CS, MDJ, TC, WJA performed additional experimental and mouse work. HS performed the pathology work. AGP and MD generate the PDX data and provided analytical support.

Acknowledgements

Not applicable

936 **REFERENCES**

937 1 Johnsen, J. I., Dyberg, C. & Wickström, M. Neuroblastoma—a neural crest derived embryonal
938 malignancy. *Frontiers in molecular neuroscience* **12**, 9 (2019).
939 2 Armstrong, G. T. *et al.* Reduction in late mortality among 5-year survivors of childhood cancer.
940 *New England Journal of Medicine* **374**, 833-842 (2016).
941 3 Moreno, L. *et al.* Accelerating drug development for neuroblastoma: summary of the second
942 neuroblastoma drug development strategy forum from innovative therapies for children with
943 cancer and International Society of Paediatric Oncology Europe Neuroblastoma. *European*
944 *Journal of Cancer* **136**, 52-68 (2020).
945 4 Ma, X. *et al.* Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and
946 solid tumours. *Nature* **555**, 371-376 (2018).
947 5 Oren, Y. *et al.* Cycling cancer persister cells arise from lineages with distinct programs. *Nature*
948 **596**, 576-582 (2021).
949 6 Boeva, V. *et al.* Heterogeneity of neuroblastoma cell identity defined by transcriptional
950 circuitries. *Nature genetics* **49**, 1408-1413 (2017).
951 7 Chen, X., Pappo, A. & Dyer, M. A. Pediatric solid tumor genomics and developmental pliancy.
952 *Oncogene* **34**, 5207-5215 (2015).
953 8 Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer.
954 *Science* **357**, eaal2380 (2017).
955 9 Shaffer, S. M. *et al.* Rare cell variability and drug-induced reprogramming as a mode of cancer
956 drug resistance. *Nature* **546**, 431-435 (2017).
957 10 Van Groningen, T. *et al.* Neuroblastoma is composed of two super-enhancer-associated
958 differentiation states. *Nature genetics* **49**, 1261-1266 (2017).
959 11 Gartlgruber, M. *et al.* Super enhancers define regulatory subtypes and cell identity in
960 neuroblastoma. *Nature Cancer* **2**, 114-128 (2021).
961 12 Schubert, N. A. *et al.* Loss of p16INK4a in neuroblastoma cells induces shift to an immature state
962 with mesenchymal characteristics and increases sensitivity to EGFR inhibitors. *bioRxiv*,
963 2021.2010. 2027.465137 (2021).
964 13 Olsen, T. K. *et al.* Malignant Schwann cell precursors mediate intratumoral plasticity in human
965 neuroblastoma. *bioRxiv* (2020).
966 14 Yuan, X. *et al.* Single-cell RNA-sequencing of peripheral neuroblastic tumors reveals an
967 aggressive transitional cell state at the junction of an adrenergic-mesenchymal
968 transdifferentiation trajectory. *bioRxiv* (2020).
969 15 Kameneva, P. *et al.* Single-cell transcriptomics of human embryos identifies multiple
970 sympathoblast lineages with potential implications for neuroblastoma origin. *Nature genetics*
971 **53**, 694-706 (2021).
972 16 Dong, R. *et al.* Single-cell characterization of malignant phenotypes and developmental
973 trajectories of adrenal neuroblastoma. *Cancer Cell* **38**, 716-733. e716 (2020).
974 17 Kildisiute, G. *et al.* Tumor to normal single-cell mRNA comparisons reveal a pan-neuroblastoma
975 cancer cell. *Science advances* **7**, eabd3311 (2021).
976 18 Bedoya-Reina, O. *et al.* Single-nuclei transcriptomes from human adrenal gland reveal distinct
977 cellular identities of low and high-risk neuroblastoma tumors. *Nature communications* **12**, 1-15
978 (2021).
979 19 Jansky, S. *et al.* Single-cell transcriptomic analyses provide insights into the developmental
980 origins of neuroblastoma. *Nature genetics* **53**, 683-693 (2021).

20 Kotliar, D. *et al.* Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *Elife* **8**, e43803 (2019).
21 Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology* **20**, 163-172 (2019).
22 Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, eaba7721 (2020).
23 Domínguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
24 Hanemaaijer, E. S. *et al.* Single-cell atlas of developing murine adrenal gland reveals relation of Schwann cell precursor signature to neuroblastoma phenotype. *Proceedings of the National Academy of Sciences* **118**, e2022350118 (2021).
25 Hsiao, C. J. *et al.* Characterizing and inferring quantitative cell cycle phase in single-cell RNA-seq data analysis. *Genome research* **30**, 611-621 (2020).
26 Kildisiute, G., Young, M. D. & Behjati, S. Pitfalls of applying mouse markers to human adrenal medullary cells. *Cancer Cell* **39**, 132-133 (2021).
27 Kinker, G. S. *et al.* Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nature genetics* **52**, 1208-1218 (2020).
28 Lavie, D., Ben-Shmuel, A., Erez, N. & Scherz-Shouval, R. Cancer-associated fibroblasts in the single-cell era. *Nature cancer* **3**, 793-807 (2022).
29 Veglia, F., Sanseviero, E. & Gabrilovich, D. I. Myeloid-derived suppressor cells in the era of increasing myeloid cell diversity. *Nature Reviews Immunology* **21**, 485-498 (2021).
30 Bedoya-Reina, O. C. & Schlisio, S. Chromaffin cells with sympathoblast signature: too similar to keep apart? *Cancer Cell* **39**, 134-135 (2021).
31 Kameneva, P. *et al.* Evolutionary switch in expression of key markers between mouse and human leads to mis-assignment of cell types in developing adrenal medulla. *Cancer Cell* **39**, 590-591 (2021).
32 Klöppel, G. Neuroendocrine neoplasms: dichotomy, origin and classifications. *Visceral medicine* **33**, 324-330 (2017).
33 Wang, C. *et al.* The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nature biotechnology* **32**, 926-932 (2014).
34 Zeineldin, M., Patel, A. G. & Dyer, M. A. Neuroblastoma: When differentiation goes awry. *Neuron* (2022).
35 Verhoeven, B. M. *et al.* The immune cell atlas of human neuroblastoma. *Cell Reports Medicine* **3** (2022).
36 Althoff, K. *et al.* A Cre-conditional MYCN-driven neuroblastoma mouse model as an improved tool for preclinical studies. *Oncogene* **34**, 3357-3368 (2015).
37 Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401 (2014).
38 Stewart, E. *et al.* Orthotopic patient-derived xenografts of paediatric solid tumours. *Nature* **549**, 96-100 (2017).
39 Zhu, Q.-Q., Ma, C., Wang, Q., Song, Y. & Lv, T. The role of TWIST1 in epithelial-mesenchymal transition and cancers. *Tumor Biology* **37**, 185-197 (2016).
40 Boen, J., Wagner, J. P. & Di Nanni, N. Inferring copy number variation from gene expression data: methods, comparisons, and applications to oncology. *bioRxiv*, 2021.2010. 2018.463991 (2021).
41 Patel, A. G. *et al.* A spatial cell atlas of neuroblastoma reveals developmental, epigenetic and spatial axis of tumor heterogeneity. *bioRxiv*, 2024.2001.2007.574538, doi:10.1101/2024.01.07.574538 (2024).

1028 42 Du, B. & Shim, J. S. Targeting epithelial–mesenchymal transition (EMT) to overcome drug
1029 resistance in cancer. *Molecules* **21**, 965 (2016).
1030 43 Shibue, T. & Weinberg, R. A. EMT, CSCs, and drug resistance: the mechanistic link and clinical
1031 implications. *Nature reviews Clinical oncology* **14**, 611-629 (2017).
1032 44 Dhimolea, E. *et al.* An embryonic diapause-like adaptation with suppressed Myc activity enables
1033 tumor treatment persistence. *Cancer cell* **39**, 240-256. e211 (2021).
1034 45 Mañas, A. *et al.* Clinically relevant treatment of PDX models reveals patterns of neuroblastoma
1035 chemoresistance. *Science Advances* **8**, eabq4617 (2022).
1036 46 Teitz, T. *et al.* Preclinical models for neuroblastoma: establishing a baseline for treatment. *PloS*
1037 *one* **6**, e19133 (2011).
1038 47 Stewart, E. *et al.* The Childhood Solid Tumor Network: A new resource for the developmental
1039 biology and oncology research communities. *Developmental biology* **411**, 287-293 (2016).
1040 48 Slyper, M. *et al.* A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human
1041 tumors. *Nature medicine* **26**, 792-802 (2020).
1042 49 Chung, N. C., Miasojedow, B., Startek, M. & Gambin, A. Jaccard/Tanimoto similarity test and
1043 estimation methods for biological presence-absence data. *BMC bioinformatics* **20**, 1-11 (2019).
1044 50 Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal,*
1045 *complex systems* **1695**, 1-9 (2006).
1046 51 Xie, Y. in *Implementing reproducible research* 3-31 (Chapman and Hall/CRC, 2018).
1047 52 Sievert, C. *Interactive web-based data visualization with R, plotly, and shiny*. (CRC Press, 2020).
1048 53 Chang, W., Cheng, J., Allaire, J., Xie, Y. & McPherson, J. Shiny: web application framework for R.
1049 *R package version* **1**, 2017 (2017).
1050 54 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for
1051 RNA-seq data with DESeq2. *Genome biology* **15**, 1-21 (2014).
1052 55 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for
1053 interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*
1054 **102**, 15545-15550 (2005).
1055 56 Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by
1056 pseudotemporal ordering of single cells. *Nature biotechnology* **32**, 381-386 (2014).

1057