TITLE

An unsupervised learning approach uncovers divergent mesenchymal-like gene expression programs across human neuroblastoma tumors, preclinical models, and chemotherapy-exposed tumors.

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

23 [7] Animal Resources Center, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

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

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

28 [10] Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA

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

30

31 **ABSTRACT**

32 Neuroblastoma is a common pediatric cancer, where preclinical studies have suggested
33 chemotherapy resistance is driven by a mesenchymal-like gene expression program. However,
34 the poor clinical outcomes imply we need a better understanding of the relationship between
35 patient tumors and preclinical models. Here, we generated single-cell RNA-seq maps of
36 neuroblastoma cell lines, patient-derived xenograft models (PDX), and a genetically engineered
37 mouse model (GEMM). We developed an unsupervised machine learning approach to compare
38 the gene expression programs found in preclinical models to a large cohort of human
39 neuroblastoma tumors. The dominant adrenergic programs were well preserved in preclinical
40 models, but contrary to previous reports do not unambiguously map to an obvious cell of origin.
41 The mesenchymal-like program was less clearly preserved, and primarily restricted to cancer-
42 associated fibroblasts and Schwann-like cells *in vivo*. Surprisingly however, we identified a
43 subtle, weakly expressed, mesenchymal-like program in otherwise adrenergic cancer cells in
44 some high-risk tumors. This program appears distinct from mesenchymal cell lines but was
45 maintained in PDX and a similar program could be chemotherapy-induced in our GEMM after

46 only 24 hours, suggesting an uncharacterized therapy-escape mechanism. Collectively, our
47 findings advance the understanding of neuroblastoma heterogeneity and can inform the
48 development of new treatments.

BACKGROUND

Neuroblastoma is the most common extracranial pediatric solid tumor, which arises in the developing sympathetic nervous system, usually in the adrenal gland[1]. Overall survival in high-risk neuroblastoma is typically reported as approximately 50%, with survivors sometimes suffering long-term consequences of chemotherapy[2]. This is despite new therapies showing clear promise in cell lines and mouse models[3], indicating 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 adults[4] and heterogeneity at the level of cell state[5] is thought to be a major driver of therapy resistance[6-10]. For example, it was recently shown that neuroblastoma *cell lines* express two discreet super-enhancer-driven gene expression programs, referred to as the “adrenergic” and the (comparatively chemoresistant) “mesenchymal” program[6,10]. These programs exhibited mutually exclusive expression, which was not driven by genetic differences, as cells expressing one program could readily interconvert to the other *in vitro*.

Single-cell RNA-seq represents the ideal tool to study such non-genetic tumor heterogeneity. Indeed, in the past 3 years, at least 7 different studies[11-17] have reported single-cell RNA-seq maps of neuroblastoma patient tumors. However, the primary conclusions of these studies disagreed considerably, with 6 different cell types posited as the neuroblastoma cell-of-origin (chromaffin cell[14], sympathoblast[13,15], Schwann cell precursors[11], neuroblast[17], “sympatho- and chromaffin cells”[16] and “a subtype of TRKB+ cholinergic progenitor”[16]). Additionally, arguments were made both in favor of[11-13,16] and against[14,15,17] the *in vivo* relevance of the cell-line-derived[6,10] putatively-chemoresistant mesenchymal-like gene expression program. Thus, the gene expression programs driving therapy resistance in neuroblastoma, and which preclinical models can be used to study/target these programs remains unclear.

73 Here, we generated single-cell RNA-seq data from neuroblastoma preclinical models and
74 compared the gene expression programs identified in these preclinical models against human
75 tumors. To achieve this, we developed a novel computational method called “automatic
76 consensus nonnegative matrix factorization” (acNMF; details below). Using acNMF, we
77 determined that the dominant adrenergic gene expression programs were preserved across
78 human tumors and preclinical models, but did not cleanly map to a unique cell of origin. In
79 contrast, the presumptive drug-resistant mesenchymal program demonstrated less consistent
80 behavior, and while clearly expressed in cell lines, exhibited high expression primarily in cancer-
81 associated normal cells *in vivo*. Notably, we identified a distinct weak-mesenchymal program in
82 predominantly adrenergic cells of high-risk tumors that could be elicited 24 hours after
83 administering chemotherapy *in vivo*. Overall, our findings improve the understanding of
84 neuroblastoma tumor heterogeneity in patient tumors and preclinical models, which will help
85 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 pathway activity and cell state programs—the factors believed to influence therapy resistance in 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 neuroblastoma 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." Our inspiration came from observing how the existing cNMF algorithm[18] 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 Figures 1a-d), where all ground truth programs were recovered and noise-filtered (Fig. 1d-h; Supplementary Figures 1e-r). In the simulated data, the 14 ground truth gene expression programs were recovered at a k value of

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

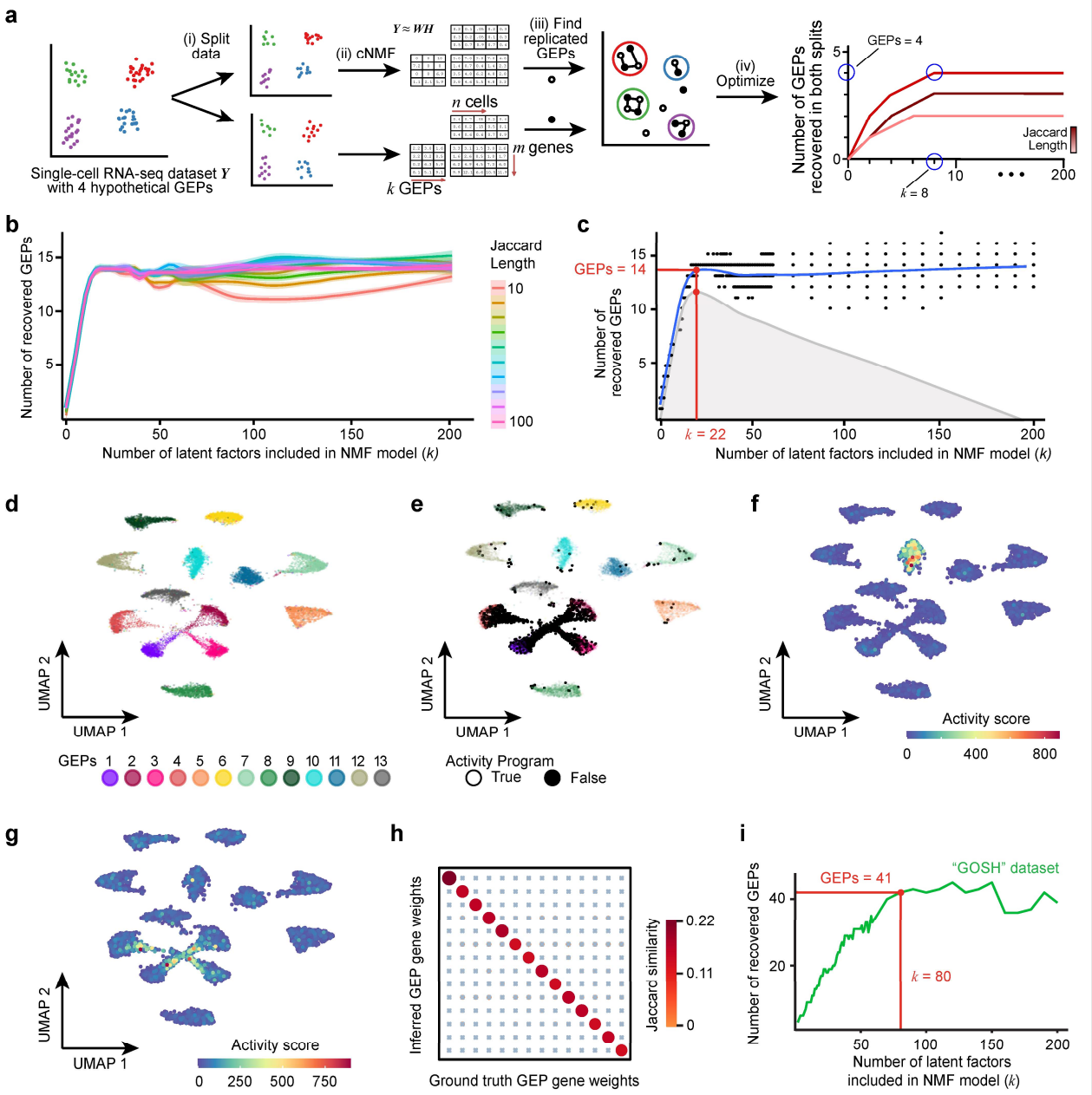


Figure 1: Overview of acNMF methodology and performance in simulated data.

(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 Figure 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 Figure 1t.

151 GEP; Gene expression program.

152

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

155 We first applied acNMF to the published "GOSH" dataset from Kildisiute *et al.*[15]. The five
156 samples therein cover low-to-high risk neuroblastoma and included one sample from normal
157 adrenal gland. Previously, the insights derived from this dataset were confined to the
158 identification of six cell types. However, the acNMF approach identified 41 gene expression
159 programs (at k = 80; Fig. 1(i)), suggesting possible new insights. We sought to annotate these
160 41 gene expression programs. Despite efforts at automation, annotation of single-cell RNA-seq
161 is subjective[19]—especially in rare/understudied tissues like neuroblastoma. Thus, two co-
162 authors (Y.K. and P.G.) blindly and independently annotated the gene expression programs.
163 This was done using a web-based platform we developed, presenting each annotator with a
164 catalogue of relevant information, including, the dominant gene weights (loadings) for each
165 program, program-program co-expression patterns, inferred copy number profiles, gene
166 ontology term enrichments, cell type annotations from SingleR[20], and 282 neuroblastoma-
167 relevant marker gene sets that we curated from the literature[10-15,17,21-28]. The web-based platform
168 presented to annotators is available at pscb.stjude.org.

169 The two annotators assigned qualitatively similar annotations to 37 out of 41 programs (Fig.
170 2(a); Supplementary Table S1). Upon discussion with coauthor R.C., a consensus was reached
171 on the final annotations (Fig. 2(a); Supplementary Table S1). The annotations contained 4
172 broad groups of cell types, neuroblastoma cancer cells, immune cells, stromal cells, and
173 endothelial cells (Fig. 2(a)). We also identified 7 "activity" programs that identified processes
174 such as the cell cycle and antigen presentation by MHC. A total of 32 discrete cell types or cell

states were recovered, with 7 assigned to the stromal group, 3 to endothelial, 19 to immune and 3 assigned to neuroblastoma cancer cells, aided by inferred copy number data (Supplementary Figure 2(a-f); see Methods; further details in subsequent subsections). We refer to these 3 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 signature (Fig 3(a-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]). Thus, the adrenergic program is preserved in patient tumors but can potentially be subdivided.

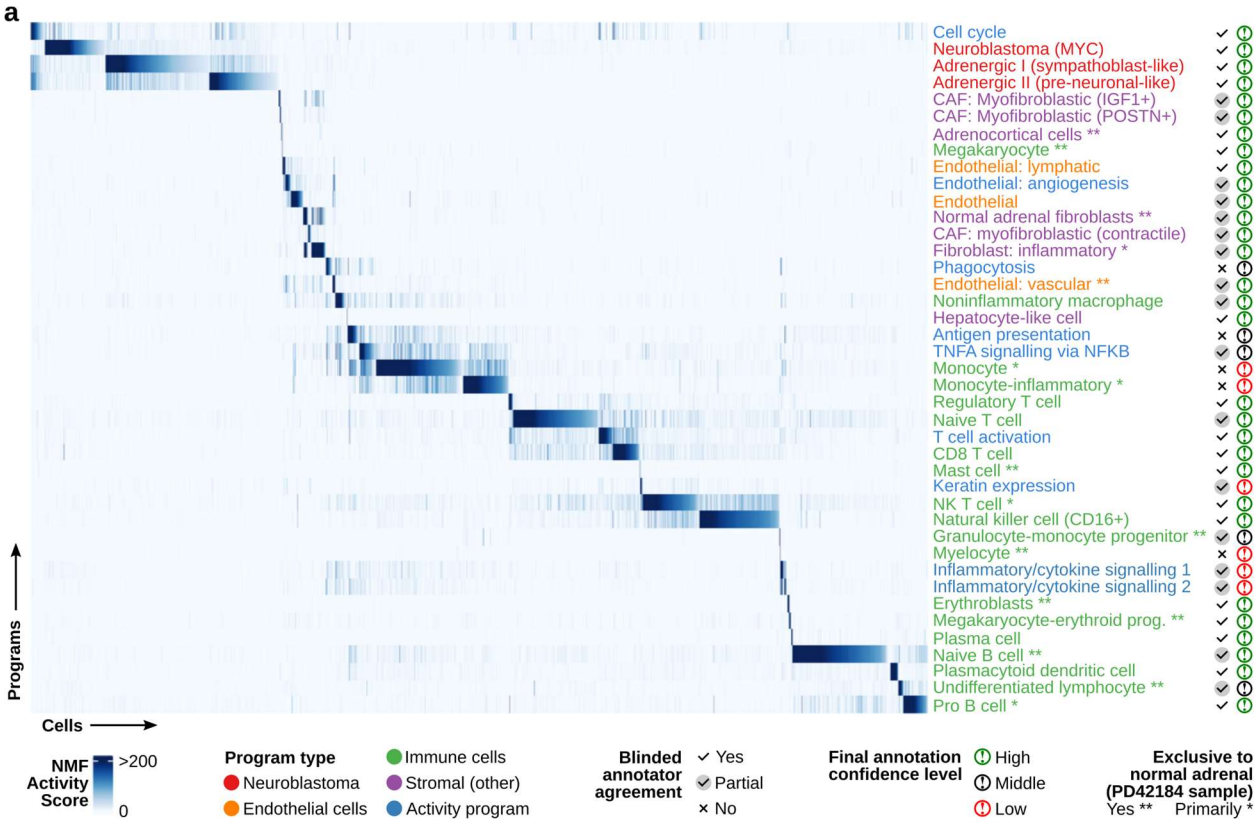


Figure 2: Gene expression programs recovered by acNMF in the GOSH neuroblastoma single-cell RNA-seq dataset.

187 (a) Heatmap showing the activity scores (color scale) for the 41 recovered programs (y axis) across
188 each single cell (x axis) in the GOSH dataset. The final consensus annotation for each gene
189 expression program is shown on the right, colored by the type of program. Tick marks indicate
190 whether the two blinded annotators agreed or not, and the far-right column indicates whether the
191 annotators reached a high, middle, or low level of confidence in the final annotation.

192

193 For all 3 adrenergic programs the top curated human cell-type signature enrichment was
194 sympathoblast (Fig. 3d, $P = 1 \times 10^{-6}$, 3.6×10^{-3}, 7.6×10^{-6} for adrenergic I, adrenergic II, NB-
195 MYC respectively, marker genes curated from Kameneva *et al.*[13]). However, there was
196 complexity beyond a clean resemblance to sympathoblasts.

197 Specifically, the Adrenergic I (sympathoblast-like) program was defined by the expression of
198 genes involved in the production of adrenaline and noradrenaline, whose expression is
199 maintained in chromaffin cells (Fig. 3(e-g)). This includes tyrosine hydroxylase (*TH*), dopamine
200 beta hydroxylase (*DBH*), dopa decarboxylase (*DDC*) ("Chromaffin and progenitors" gene list
201 curated from Jansky *et al.*[17]). Notably, markers specific to mature chromaffin cells (*PNMT* and
202 *PENK*) were not enriched, consistent with the idea neuroblastoma cells do not resemble mature
203 chromaffin cells[25,29,30].

204 The Adrenergic II (pre-neuronal-like) program over-expressed markers of neuroblasts defined
205 by Hanemaaijer *et al.*[23] and Jansky *et al*[17]. The top 5 program-defining marker genes included
206 neuronal specific genes, neurofilament light chain (*NEFL*), neuronatin (*NNAT*), brain acid
207 soluble protein 1 (*BASP1*) and neurofilament medium chain (*NEFM*). This program can be co-
208 expressed with the Adrenergic I program (Fig. 3g & 3i, Supplementary Figure 3a-c). Thus, these
209 programs were independent, but not mutually exclusive.

210 The NB-MYC program was only expressed in the high-risk patient sample (PD43255) and was
211 characterized by lower expression of marker genes of differentiating chromaffin and neuronal
212 cell types (Fig. 3d, 3f, Supplementary Figure 3a). The defining genes of this program were
213 dominated by translation initiation and elongation factors (Supplementary Figures 3d & 3e),
214 consistent with the undifferentiated pathology and hyper-aggressive growth of high-risk
215 neuroblastoma. The top marker genes of the adrenergic I & II programs were expressed 2-3-
216 fold lower in the high-risk tumor (Fig. 3(j-k)). This is consistent with low-risk tumors exhibiting
217 differentiating histology (Fig. 3l)[31].

218 Collectively, these results point to a possible source of the disagreement on neuroblastoma's
219 "cell of origin", showing neuroblastoma can co-express a mixture of independent programs,
220 each resembling different developmental cell types (Fig. 3(g, i), Supplementary Figure 3b-c).
221 Thus, marker genes of multiple cell types will always be identifiable and pursuing a clean
222 mapping between neuroblastoma cells and a specific developmental cell type may represent a
223 futile exercise for determining neuroblastoma's specific cell of origin.

224

225 **Cancer associated fibroblasts of the neuroblastoma tumor microenvironment strongly**
226 **express mesenchymal features associated with neuroblastoma cell lines.**

227 We identified 9 gene expression programs in the GOSH dataset that statistically resembled the
228 neuroblastoma cell line-derived mesenchymal gene signature, with some very strong
229 enrichments (e.g., $P = 1.78 \times 10^{-88}$ for Program 40 for enrichment of Van Groningen
230 "mesenchymal" signature, Supplementary Table S1). This seems to mean mesenchymal
231 neuroblastoma cells have been identified in these tumors. However, one of these programs was
232 exclusively expressed in fibroblasts of the reference normal adrenal sample (Fig. 3(m-n); $P =$
233 3.75×10^{-40} for mesenchymal enrichment). The shared gene expression features of fibroblasts

234 and mesenchymal-like neuroblastoma cell lines suggest a source of the disagreements about
235 whether mesenchymal-like neuroblastoma cells exist *in vivo*. Indeed, in the 4 neuroblastoma
236 tumors, our acNMF decomposition again identified several programs strongly resembling
237 mesenchymal-like neuroblastoma cell lines. However, in all cases, we annotated these as non-
238 cancerous fibroblast cell populations, not mesenchymal-like neuroblastoma cells (Fig. 3(o-q);
239 these annotations leveraged inferred copy number - details in subsequent section).

240 Additionally, acNMF revealed a more complex landscape of cancer associated fibroblasts
241 (CAFs) than previously reported in neuroblastoma[27]. Specifically, we identified IGF1+ (P = 1.92
242 $\times 10^{-7}$), POSTN+ ($P = 3 \times 10^{-15}$) and contractile myofibroblast populations ($P = 2.5 \times 10^{-3}$). We
243 also identified inflammatory CAFs ($P = 2.9 \times 10^{-9}$) expressing interleukin (*IL6*, *IL33*) and
244 complement component genes (*C7*, *C3*), using marker genes from Lavie *et al*.[27]. Overall, our
245 acNMF-based decomposition of the GOSH dataset yielded us a conceptual model to describe
246 the behavior of neuroblastoma cancer cells (Fig. 3(r)) but did not identify clear evidence for
247 mesenchymal-like neuroblastoma cells *in vivo*. We next explore whether these characteristics
248 are consistent in other human datasets and preclinical models.

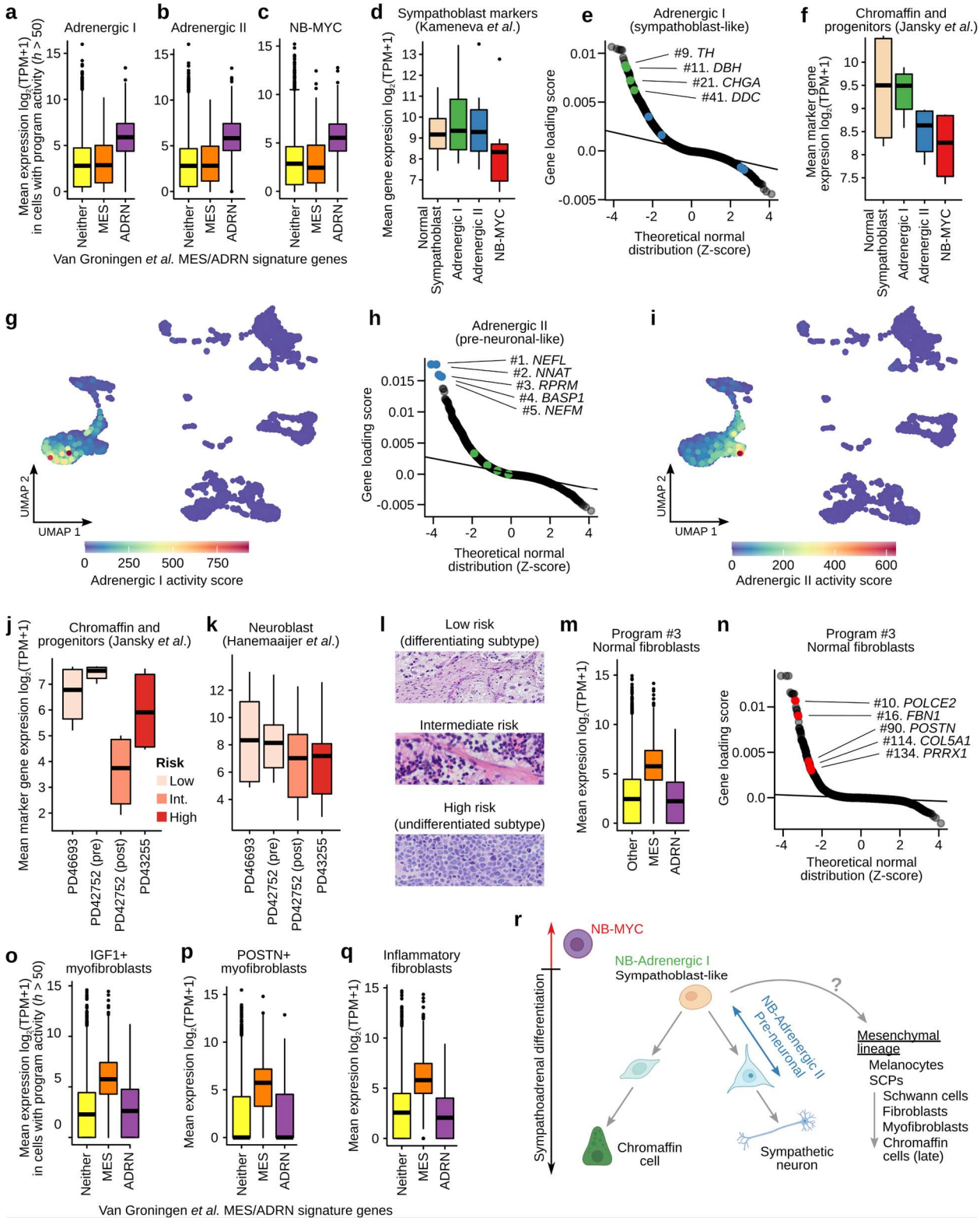


Figure 3: Cells expressing multiple adrenergic and mesenchymal-like gene expression programs are evident in neuroblastoma tumors.

252 (a) Boxplots showing the mean expression (y axis) of the genes from the classical Van Groningen
253 mesenchymal (MES; orange box) or adrenergic (ADRM; purple box) signature, in cells that are
254 highly expressing (activity $h > 50$) the acNMF-identified "Adrenergic I (sympathoblast-like)"
255 program in the GOSH dataset. The yellow box shows the mean expression of all other genes in
256 these cells.
257 (b) Like (a) but for the "Adrenergic II (pre-neuronal-like)" program.
258 (c) Like (a) but for the "Neuroblastoma-MYC" program.
259 (d) Boxplots showing the mean gene expression (y axis) of sympathoblast marker genes (curated
260 from Kameneva *et al.*) in cells highly expressing (activity $h > 50$) each of the 3 acNMF recovered
261 adrenergic programs, or in normal sympathoblasts (obtained from the Descartes Cell Atlas).
262 (e) QQ-plot showing the gene weights (loading scores) (y axis) for the "Adrenergic I (sympathoblast-
263 like)" program. Key sympathoblast marker genes have been highlighted in green, and key
264 neuronal marker genes in blue.
265 (f) Like (d) but for the "Chromaffin and progenitors" marker gene set, curated from Jansky *et al*.
266 (g) UMAP plot of the GOSH dataset; each point represents a single-cell, colored by the activity of
267 the "Adrenergic I (sympathoblast-like)" program.
268 (h) Like (e) but for the "Adrenergic II (pre-neuronal-like)" program.
269 (i) Like (g) but for the "Adrenergic II (pre-neuronal-like)" program.
270 (j) Boxplot showing the mean gene expression of "Chromaffin and progenitors" marker genes
271 (curated from Jansky *et al*.) in the cancer cells of the 4 neuroblastoma samples profiled in the
272 GOSH dataset. Boxes are colored by tumor risk classification.
273 (k) Like (j) but for the "Neuroblast" marker gene set curated from Hanemaaijer *et al*.
274 (l) Representative H&E stains of low, intermediate and high-risk human neuroblastoma tumors,
275 illustrating varying degrees of differentiated features.
276 (m) Boxplots showing the mean expression (y axis) of the genes from the classical Van Groningen
277 mesenchymal (MES; orange box) or adrenergic (ADRM; purple box) signature, in cells that are

278 highly expressing (activity $h > 50$) the acNMF-identified "Normal fibroblasts" program in the
279 GOSH dataset. The yellow box shows the mean expression of all other genes in these cells.
280 (n) QQ-plot showing the gene loading scores (y axis) for the "Normal fibroblasts" program. A subset
281 of genes classically used as markers of mesenchymal-like neuroblastoma cells are highlighted in
282 red.
283 (o) Like (m) but for the "IGF1+ myofibroblasts" program.
284 (p) Like (m) but for the "POSTN+ myofibroblasts" program.
285 (q) Like (m) but for the "Inflammatory fibroblasts" program.
286 (r) A schematic diagram overlaying the acNMF-recovered programs on normal sympathoadrenal
287 differentiation[32].

288

289 **The reproducibility of neuroblastoma gene expression programs across human tumors**
290 **and preclinical models.**

291 We next assembled 3 more of the publicly available datasets, from Dong *et al.*, Jansky *et al.*
292 and the PMC dataset from Kildisuite *et al.*[14-17], totaling 50 samples. Additionally, to assess the
293 preservation of the human programs in preclinical models, we performed single-cell RNA-seq
294 profiling across 13 neuroblastoma cell lines, 17 patient derived xenografts (PDX), and 11
295 samples from a genetic mouse model (GEMM)[33] (metadata for this (largest of its kind) cohort
296 assembled in Supplementary Table S2, and in our web platform).

297 We applied the acNMF decomposition to each of these datasets independently, yielding
298 between 17 (cell lines) and 45 (Dong *et al.*) gene expression programs per-dataset
299 (Supplementary Figure 1t, Supplementary Table S3). To pairwise cross-compare the programs
300 from each dataset we represented the programs as nodes on a graph. We drew an edge
301 between nodes if the program was replicated in another dataset (*FWER* < 0.01 by an

302 asymmetric Jaccard similarity test; Fig. 4(a); Supplementary Figure 4(a); see Methods). The
303 resulting graph's structure indicated that similar gene expression programs had often been
304 recovered in multiple datasets, including preclinical models. Consensus hubs (see Methods)
305 were evident for gene expression programs representing several immune cell types,
306 neuroblastoma-specific programs, stromal cells such as fibroblasts and activity programs such
307 as the cell cycle and hypoxia (Fig. 4(a-d)).

308 The adrenergic I (sympathoblast-like) program from GOSH was directly or indirectly connected
309 to programs in all other datasets (Fig. 4(b)). These connected programs retained the
310 characteristic hallmarks including expression of genes like *DBH*, *TH* and *CHGA* (Fig. 4 (e-h)).
311 The adrenergic II pre-neuronal-like program was also observed in multiple other datasets (Fig. 4
312 (f & h)) and the "neuroblastoma-MYC" (high risk) program formed a dense hub (Fig. 4(c)), with
313 similar programs recovered in high-risk tumors in all datasets and all preclinical models (Fig.
314 4(i)).

315 This cross-dataset analysis also revealed 3 groups of novel programs expressed in
316 neuroblastoma cancer cells. We were able to find an explanation for one, which was expressed
317 in *ALK* activated tumors (Fig. 4(a); supported by 1 patient tumor and 2 PDXs, but no other
318 preclinical model). Additionally, we found 5 programs unique to a single human neuroblastoma
319 tumor implying a possible long tail of divergent neuroblastoma programs that currently remain
320 unexplained and are absent in our profiled preclinical models.

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

327 Overall, this cross-dataset integrated map argues that, despite disagreement in the literature,
328 neuroblastoma single-cell RNA-seq datasets are reasonably consistent, and that the dominant
329 adrenergic programs are preserved in all preclinical models. However, the status of
330 mesenchymal-like tumor cells and cell lines requires further exploration. Note: this graph-based
331 representation of these datasets (Fig. 4(a)) is available as part of our interactive web-based
332 platform (accessible at pscb.stjude.org).

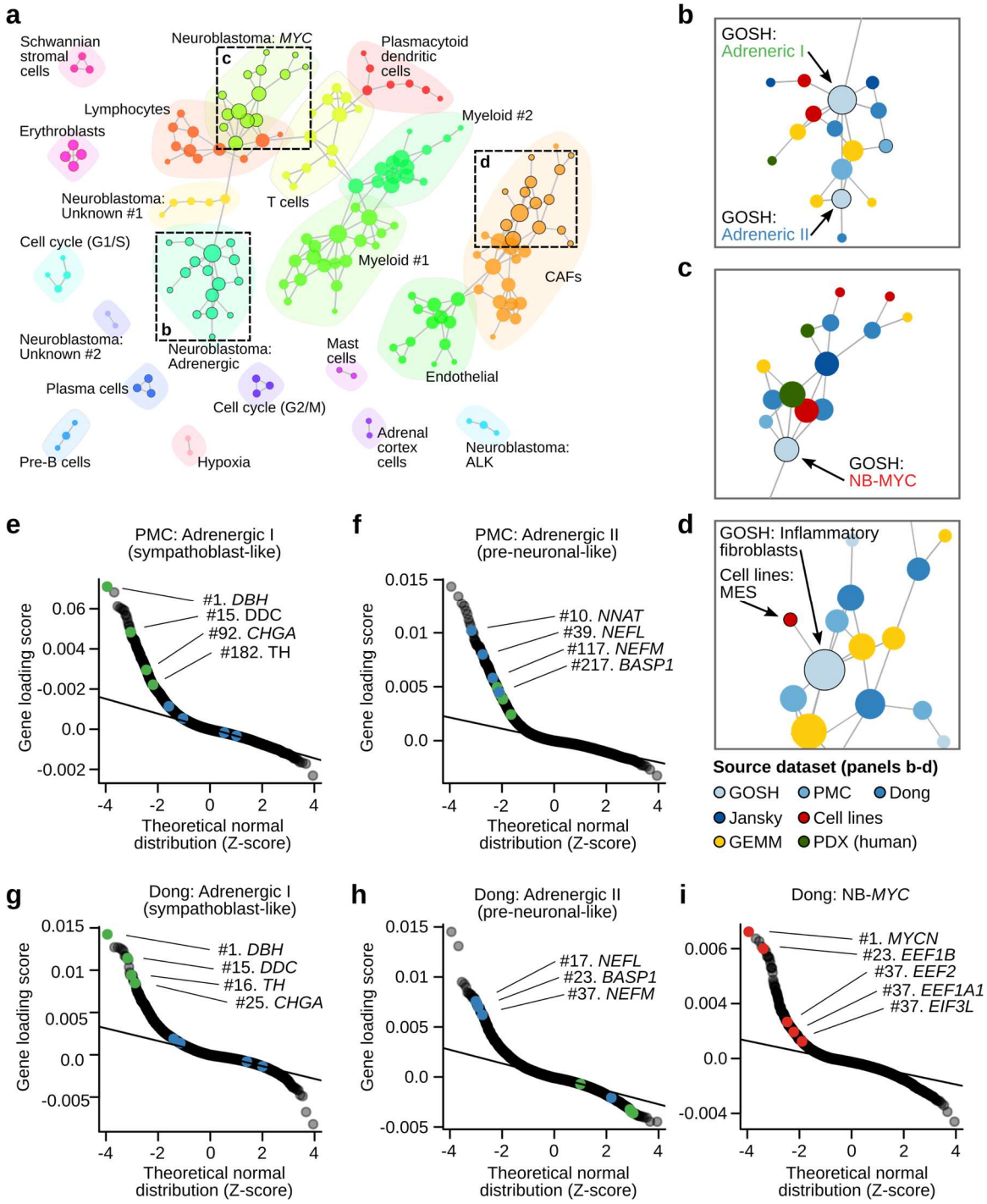


Figure 4: The reproducibility of neuroblastoma gene expression programs across human tumors and preclinical models.

(a) A graph of gene expression programs identified by acNMF across 7 neuroblastoma single-cell RNA-seq datasets, from humans, PDX, GEMM and cell lines. Nodes represent a single

program's gene weights (i.e. loading scores), and an edge connects similar nodes by a Jaccard similarity test. The colors highlight related communities of nodes. The black boxes highlight regions of interest, which are shown in detail in panels (b-d).

(b) Inset showing a detailed view of region (b) from panel (a). This region has connected Adrenergic I and Adrenergic II programs. In this inset, nodes have been colored by dataset, rather than community (as in panel (a)).

(c) Like (b) but for a dense interconnected community of "Neuroblastoma MYC" programs.

(d) Like (c) but showing a region for a community of gene expression programs expressed in cancer associated fibroblasts. A neuroblastoma program, exclusive to the GIMEN cell line, is connected to this community of nodes.

(e) QQ-plot showing the gene weights (loading scores) (y axis) for the "Adrenergic I (sympathoblast-like)" program recovered in the PMC dataset. Similar to Fig. 3(e) from the GOSH dataset key sympathoblast marker genes have been highlighted in green, and key neuronal marker genes in blue.

(f) Like (e) but for the "Adrenergic II (pre-neuronal-like)" program.

(g) Like (e) but for program 24 "Adrenergic I (sympathoblast-like)" in the Dong *et al.* dataset.

(h) Like (e) but for program 19 "Adrenergic II (pre-neuronal-like)" in the Dong *et al.* dataset.

(i) Like (e) but showing the "Neuroblastoma MYC" program from Dong *et al*. Top marker genes, similar to those identified in the GOSH dataset, have been highlighted in red.

Note: An interactive version of the graph plot (panel a), and QQ-plots (panels e-i) for every program in every dataset are available in our web platform.

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

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

369 Thus, we applied InferCNV[34] across all datasets (see Methods), a computational method for
370 inferring copy number variant (CNV) profiles from single-cell RNA-seq data. However, the CNV
371 profiles of mesenchymal-like cells were always similar to the reference normal, whereas
372 aberrations in adrenergic cells were typically obvious (Supplementary Figure 2, Supplementary
373 Figures 5a-c). The exception was cell lines, where both mesenchymal and adrenergic cells were
374 always consistent with cancer cells (Supplementary Figure 5d-f).

375 However, this basic InferCNV analysis cannot rule out a small proportion of mesenchymal-like
376 cancer cells. Thus, we constructed a machine learning model, that by leveraging the inferred
377 copy number profiles of the high-confidence-cancer and high-confidence-normal cells, could
378 *predict* the cancer/normal status of the ambiguous mesenchymal-like cells on a per-cell basis
379 (Supplementary Figure 5g; Supplementary Table 4 & 5; see Methods). These models could
380 predict the cancer/normal status of the held-out cancer/normal-endothelial cells with median
381 accuracy approaching 100% (Fig. 5a, Supplementary Figure 5h). We applied the trained models
382 to the held-out ambiguous population of mesenchymal-like cells (candidate mesenchymal
383 neuroblastoma cells, cancer associated fibroblasts, or Schwann-like cells). The model never
384 classified mesenchymal cells as cancer cells beyond the model's misclassification rate,
385 estimated from normal endothelial cells (Fig. 5b-c; Supplementary Figure 5i-n). Thus, these

386 analyses did not reveal convincing evidence for subsets of neuroblastoma cells highly
387 expressing a mesenchymal-like gene expression program *in vivo*.

388 We further probed these putative fibroblast and Schwann-like cells using a highly sensitive
389 spatial transcriptomics assay, multiplexed RNA *in situ* hybridization (RNA-ISH) (Fig 5d,
390 Supplementary Figure 6a-c). We applied this to our GEMM, which develops neuroblastoma
391 because of Cre-conditional activation of the *human* copy of *MYCN* in dopamine beta
392 hydroxylase (*Dbh*) expressing cells—thus, human *MYCN* expression should generally be
393 excluded from non-cancer cells. In all tumors, the human *MYCN* transcript was highly
394 expressed across most of the tissue, but absent from some large regions (Fig. 5e,
395 Supplementary Figure 7a-d), which resemble Schwannian stroma (Fig. 5f, Supplementary
396 Figure 7e). These *MYCN*-absent regions clearly co-expressed the myofibroblast markers *Tagln*
397 and *Acta2* (Fig. 5e), consistent with the idea these mesenchymal cells are fibroblasts, not
398 cancer cells. Regions excluding *MYCN* also expressed the inflammatory fibroblast marker *C3*
399 (Fig. 5e). Interestingly, we identified some cells seemingly co-expressing all combinations of
400 chosen fibroblast markers, consistent with the idea of intermediate subpopulations (Fig 5g).
401 Schwann cell markers *Sox10* and *Plp1* were also excluded from regions expressing human
402 *MYCN* (Fig. 5h; Supplementary Figure 7e). Collectively, despite strong transcriptional
403 resemblance to neuroblastoma mesenchymal cell lines, the data indicate these *in vivo* cells are
404 predominantly Schwann-like cells and previously uncharacterized fibroblast subpopulations (see
405 Discussion).

406

407

408

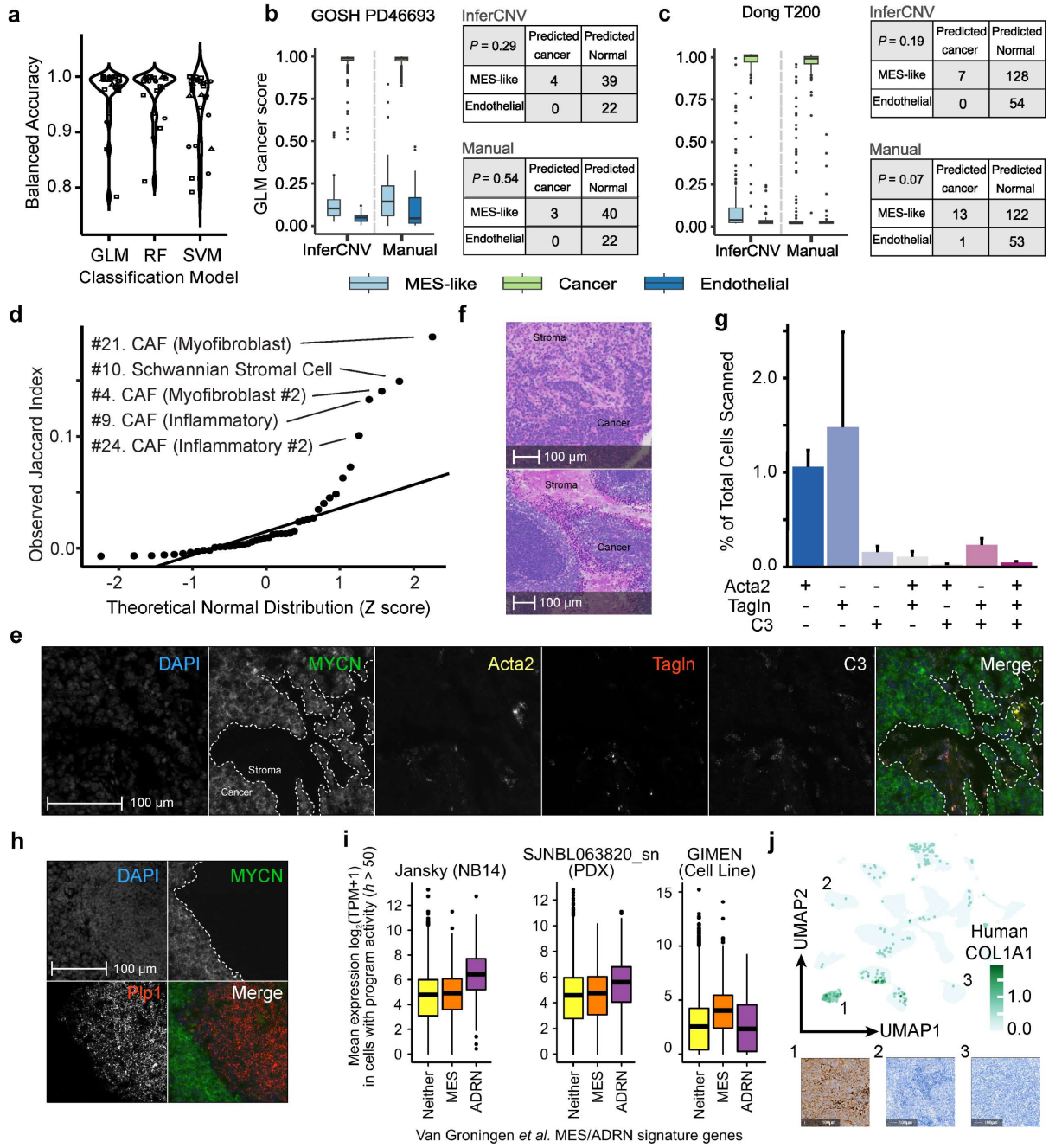
409 **Rare high-risk neuroblastoma tumors express a weakly enriched mesenchymal-like**
410 **signature *in vivo*.**

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

418 We initially suspected this was artefactual but surprisingly also identified a high-risk pretreated
419 human tumor in our PDX cohort exhibiting a similar weak enrichment of the mesenchymal
420 signature across most cancer cells of the tumor (Fig. 5i; Program 1, $P = 1.6 \times 10^{-12}$ in
421 SJNBL063820, Supplementary Figure 8c-d). We obtained this PDX tissue[35] and used
422 immunohistochemistry to stain for COL1A1 protein, a classical neuroblastoma mesenchymal
423 marker[10]. Remarkably, greater than 95% of the neoplastic cells of SJNBL063820 expressed
424 COL1A1 in the cytoplasm and membrane, but staining was not evident in two negative control
425 PDXs (Fig. 5j). Thus, this weak-mesenchymal phenotype can be recapitulated in PDX.
426 Interestingly, two treatment naïve high-risk *MYCN* amplified tumors (T200, T230) from the Dong
427 dataset, while still adrenergic (Supplementary Figures 8e-f), also showed broad overexpression
428 of the transcription factor *TWIST1*, a known driver of epithelial-to-mesenchymal transition[36]
429 (ranked #9 on Program 11 (T200-specific) and #27 on Program 33 (T230-specific)). Curiously,
430 both programs were also weakly but significantly co-expressed with the mesenchymal-fibroblast
431 Program 7 ($r_s = 0.07$ $P = 1.4 \times 10^{-33}$ and $r_s = 0.06$, $P = 7.1 \times 10^{-25}$ respectively), consistent with a
432 very weak or proto-mesenchymal cell state expressed in otherwise adrenergic cells. In general,

433 these findings uncover a complex landscape of mesenchymal-like gene expression programs *in*

434 *vivo*, where some behaviors appear different to cell line models (see Discussion).



435

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

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

463 h) Representative image of GEMM tumors stained with *MYCN* and the Schwann cell marker *Plp1* in
464 NB839. *Plp1* expression was restricted to *MYCN*-absent regions in these tumors.
465 i) Boxplot showing the mean expression (y axis) of genes in neuroblastoma cancer cells from the
466 sample NB14 from Jansky *et al*. Boxes are genes from the classical Van Groningen mesenchymal
467 (MES; orange box) or adrenergic (ADRN; purple box) signatures, with the yellow box showing
468 the expression of all other genes. The middle panel shows our PDX sample SJNBL063820 and
469 the right panels shows the predominantly mesenchymal neuroblastoma cell line GIMEN.
470 j) Immunohistochemical staining for the mesenchymal-marker gene COL1A1 in our SJNBL063820
471 PDX sample (left) and in two other PDX samples SJNBL046145 (middle) and SJNBL047443
472 (right).

473

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

476 The identification of proto-mesenchymal features (e.g. *TWIST1* expression) led us to
477 hypothesize that the weak mesenchymal-like programs in the drug pretreated tumors (NB14,
478 SJNBL063820) could have been chemotherapy induced. To test this, we treated 5 of our
479 GEMMs with a single pharmacologically relevant dose of cisplatin (a mainstay of neuroblastoma
480 treatment), along with 6 vehicle-treated control mice (see Methods). We harvested these tumors
481 24 hours after treatment and performed single-cell RNA-seq (see Methods; Supplementary
482 Figure 9a).

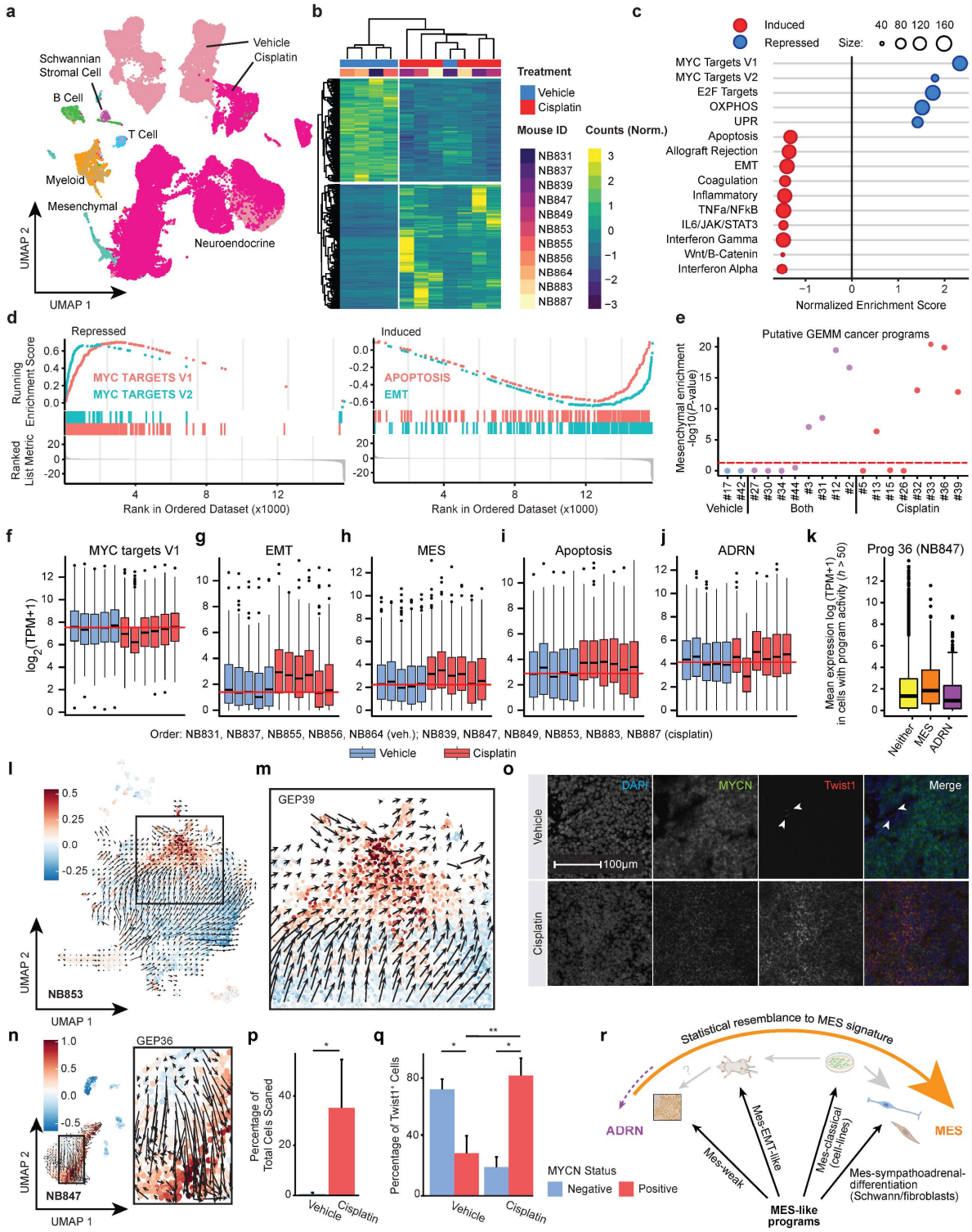
483 Encouragingly, the drug treated samples clustered separately from the vehicle treated samples
484 (Fig. 6a). To determine which genes had been induced/repressed we created “pseudo bulk”
485 expression profiles using all cells of each tumor and performed functional enrichment (drug vs
486 vehicle) using GSEA, testing the mSigDB “Hallmark” gene sets (Fig. 6b-d) (see Methods). As

487 expected, “Apoptosis” was among the top chemotherapy-induced gene sets (Fig. 6c-d and 6i, P
488 = 0.03; Supplementary Table 6). Remarkably, induced genes were also highly enriched for
489 “Epithelial to Mesenchymal Transition (EMT)” (Fig. 6c-d and 6g, $P = 7.4 \times 10^{-3}$). We probed this
490 further using acNMF, where in drug treated samples, we identified (for the first time) clear
491 mesenchymal-like programs expressed in cancer cells (Fig. 6e). Additionally, cisplatin-treated
492 samples were the first where our machine learning model classified cells highly expressing a
493 mesenchymal program as cancer cells ($P < 2.2 \times 10^{-16}$ for NB853, $P = 9.5 \times 10^{-11}$ for NB847,
494 Supplementary Figures 9b-f).

495 While the repressed genes were quite consistent across all drug-treated samples, e.g., the loss
496 of the MYC target genes (Fig. 6b-d and 6f), the patterns of induced genes showed greater
497 variation (Fig. 6b and 6f-j). The EMT signal was strongest in NB839, NB847, NB849 and NB853
498 (Fig. 6g) where the classical Van Groningen mesenchymal signature was also enriched (Fig.
499 6h). These tumors also had the highest induction of apoptosis genes (Fig. 6i), with a loss of the
500 adrenergic genes in NB847 (Fig. 6j)—reminiscent of trends we had only previously observed in
501 cell lines (Fig. 6k; Fig 5i). We performed RNA-velocity analysis in NB847 and NB853, which was
502 consistent with the mesenchymal programs resulting from an acute cell state transition (Fig. 6l-
503 n, Supplementary Figure 10a). Overall, these recently drug-treated tumors demonstrate that a
504 mesenchymal-like gene expression program can be chemotherapy induced in neuroblastoma.

505 Our GEMM, which expresses human *MYCN* in cancer cells, again allowed us to validate these
506 conclusions by an orthogonal RNA-ISH assay. This time, we assessed the co-localization of
507 human *MYCN* and the classical EMT marker *Twist1* (Supplementary Figure 10b-c) across 4
508 drug and 4 vehicle treated samples. While *Twist1* expression was evident to some degree in
509 both drug and vehicle treated samples, the co-localization of *MYCN* and *Twist1* was much
510 higher in drug treated samples, again consistent with the induction of a mesenchymal program
511 (Fig. 6o and 6q).

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



518

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

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

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

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

565

566

567 **DISCUSSION**

568 Several single-cell RNA-seq maps of neuroblastoma patient tumors have now been
569 published[6,10]. However, their conclusions were in marked disagreement, with 6 different cell
570 types proposed as the neuroblastoma cell of origin and major disagreements on evidence for a
571 cell line-derived drug resistant mesenchymal-like cell state *in vivo*. To address these issues, we
572 have performed the first integrative analysis of single-cell RNA-seq data from human
573 neuroblastoma tumors and preclinical models. To achieve this, we developed a novel NMF-
574 based computational method and a front-end user interface for interpreting the resulting gene
575 expression programs, which provides a new public resource for interrogating neuroblastoma
576 biology (pscb.stjude.org). Using these tools, we discovered that classical neuroblastoma
577 adrenergic and mesenchymal programs can be subdivided, with implications for drug resistance
578 and cell of origin.

579 Specifically, we identified novel neuroblastoma adrenergic programs, mimicking the
580 sympathoadrenal developmental lineage, and showed that neuroblastoma tumors are co-
581 expressing marker genes of multiple developmental cell types. The co-expression pattern of
582 these programs potentially explains why so many conflicting cell types have been proposed as
583 the neuroblastoma cell of origin, because neuroblastoma cells concurrently express features of
584 e.g. sympathoblasts and developing chromaffin/neuronal cells. These behaviors were preserved
585 between human tumors and preclinical models. We propose a conceptual mapping of
586 neuroblastoma gene expression programs to sympathoadrenal development in Fig. 3r, naming
587 these adrenergic programs “Adrenergic I (sympathoblast-like)”, “Adrenergic II (pre-neuronal-
588 like)” and “Neuroblastoma-MYC”.

589 We also found that the classical cell-line-derived mesenchymal-like program has differences
590 between human tumors and preclinical models. First, while the neuroblastoma mesenchymal
591 signature was clearly identifiable in neuroblastoma cell lines, we could not identify strong

592 evidence that a similar program is expressed in cancer cells in human tumors—neither by
593 inferred copy number nor by RNA ISH. Regardless, it is logically impossible to conclude the
594 non-existence of such cells. Indeed, they could be identifiable in other tumors, by profiling larger
595 numbers of cells, or by using more sensitive technologies (e.g., our RNA ISH experiments had
596 difficulty distinguishing cancer vs fibroblast/Schwann-like cells at the boundaries of stromal
597 regions, it is conceivable some cell types could have dropped out of the single-cell profiling, and
598 tools for inferring copy number are imperfect[37]). Despite this, we can conclude that such cells
599 are likely less abundant than fibroblasts and Schwann-like cells, which strongly express a
600 similar mesenchymal-like signature and thus, on the basis of marker gene expression, are
601 easily confused for mesenchymal-like neuroblastoma cells. These analyses also identified novel
602 subpopulations of cancer associated fibroblasts in most datasets/tumors, which we could
603 validate using RNA ISH in our GEMM, revealing at least inflammatory, myofibroblast and
604 potentially intermediate subpopulations. Similar cell types are critical in regulating
605 immunosuppression, metastasis and progression in other cancers[27] and determining the
606 biological/therapeutic relevance of these new cell types in neuroblastoma will form the basis for
607 future research.

608 Lastly, we identified rare, high-risk, drug pretreated tumors, which weakly express
609 mesenchymal-like features in cancer cells that remain predominantly adrenergic. This contrasts
610 with the behavior in cell lines, where a mesenchymal state coincides with loss of adrenergic
611 features. This may be related to the concept of “high risk tumors with mesenchymal-like
612 features”, which was recently proposed by Jansky *et al*[17], but in contrast to Jansky we showed
613 these cells at least partially resemble the classical Van Groningen mesenchymal signature. We
614 also identified a potential source of this phenotype, showing that a mesenchymal-like program
615 could be strongly induced *in vivo* in an acutely chemotherapy treated GEMM. This is consistent
616 with observations across several other cancers that EMT can represent an apoptosis escape

617 mechanism that leads to drug resistance[38,39]. It appears high-risk neuroblastoma cells may be
618 primed to undergo such a transition. It is tempting to speculate that targeting this process could
619 circumvent chemoresistance and it is plausible that some mesenchymal-like cell lines may
620 represent a good model for this behavior. Because of the very short 24-hour treatment window,
621 this finding also implies that the expression of mesenchymal-like genes need not necessarily
622 emerge in relapsed neuroblastoma tumors by purifying selection, as previously speculated[6,10,40].
623 However, the clinical significance of these drug-induced and weakly enriched mesenchymal
624 programs, and their potential role in facilitating escape from chemotherapy induced apoptosis,
625 remains to be determined. In general, our results argue that mesenchymal-like programs in
626 neuroblastoma could be subdivided into at least (1) a weak mesenchymal-like signature co-
627 expressed in still-predominantly-adrenergic cells of some high-risk tumors, which (2) could
628 potentially be chemotherapy induced; (3) the classical mesenchymal-like signature in cell lines,
629 which also strongly express features similar to (4) cancer associated fibroblasts and Schwann-
630 like cells found in tumors. As these represent (at least partially) qualitatively different entities, we
631 propose naming these (1) Mes-weak, (2) Mes-EMT-like, (3) Mes-classical, and (4) Mes-
632 sympathoadrenal-differentiation (Fig. 6r).

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

639 **METHODS**

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

641 LSL-MYCN;Dbh-iCre mice[33] (129X1/SvJ) were obtained from Michael A. Dyer's lab (St. Jude).
642 This study was carried out in strict accordance with the recommendations in the Guide to Care
643 and Use of Laboratory Animals of the National Institute of Health. The protocol was approved by
644 the Institutional Animal Care and Use Committee at St. Jude Children's Research Hospital. All
645 mice were housed per approved IACUC protocols, on a 12−12 light cycle (light on 6 am off
646 6 pm) and provided food and water ad libitum. Neuroblastoma-like tumors comparable to
647 human neuroblastoma arose in the abdomen from either the adrenal gland or the celiac and
648 superior cervical ganglia with a penetrance of approximately 50%, at a median of about 6
649 months old. The tumor volumes were monitored by ultrasound using the Vevo 3100 system
650 (Fujifilm-Visualsonics, Toronto, Canada), and once an adrenal tumor had reached 500mm^3 in
651 diameter, the mice were treated with either a single pharmacologically relevant[41] 2 mg/kg dose
652 of cisplatin (n = 6), or vehicle control (n = 5), by intraperitoneal injection. The mice were
653 euthanized at 24 hours, and tumors were immediately harvested and subjected to single-cell
654 RNA-seq, using the protocols described below.

655

656 **Orthotopic patient-derived xenografts**

657 Flash-frozen fragments of orthotopic patient-derived xenografts were obtained through the St.
658 Jude Childhood Solid Tumor Network (https://cstn.stjude.cloud/)[35,42]. Approximately 50-100 mg
659 of frozen tissue underwent nuclei extraction using manual disruption within a chilled Tween
660 containing buffer[43]. To generate fresh xenograft tissue, cells were injected into the para-adrenal
661 space using ultrasound guidance[44] into immunocompromised nude mice. Tumors were

662 harvested from xenograft-bearing mice, and approximately 250-500 mg of tissue were
663 dissociated using the Papain Dissociation System (Worthington Biochemical)[43].

664

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

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

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

674 *Library Preparation and Sequencing:* The dissociated/suspended single cells were filtered and
675 strained through a 40µm strainer and cells were counted using the Luna cell counter. Next, they
676 were loaded into Chromium Chips from Single cell 3' Gene expression V3.1 kit (10X Genomics,
677 Pleasanton, CA) with target capture of 6000 cells. The cDNA generation, library preparation,
678 and QC was performed following 10X Genomics protocol. The library was sequenced in
679 Novaseq-V1 reagents, and the sequenced data were processed using the Cell Ranger Software
680 (10X Genomics) mapped to mm10 (mouse) or GRCh38 (cell line). Poor quality and outlier cells
681 were identified and removed. Analysis downstream of acNMF, was conducted on normalized
682 read counts (SCTransform, Seurat R package) of the resulting cells.

683

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

685 For each dataset, raw read counts were merged into a single matrix containing all samples.
686 50,000 cells were subsampled per dataset, and then randomly split into two subsets containing
687 roughly equal numbers of cells. cNMF[18] (https://github.com/dylkot/cNMF/) was applied to each
688 split independently by performing 200 NMF iterations per rank *(k),* spanning a range of k = 2 to k
689 = 200. Outlier replicates, defined by Euclidean distance, were filtered at a rate of 10%. The
690 consistency of gene expression programs between the data splits was assessed by calculating
691 the Jaccard similarity index of *Z*-score normalized (see Kotlier *et al.*[18] for details) gene weight
692 coefficients at each value of *k*. *P*-values for all pairwise comparisons between the columns of
693 the component matrices were calculated using the jaccard.test.mca function from the
694 'jaccardtest' R package[45] followed by Bonferroni multiple testing correction. This was carried out
695 over a range of Jaccard lengths. All significant pairs ($P < 0.05$) of components between data
696 splits were then represented as node pairs in an undirected network graph using the R package
697 *igraph*[46]. Community detection was performed on the underlying graph using edge
698 betweenness centrality, and the number of distinct communities were recorded. These
699 communities represent replicated gene expression programs between data splits, and these
700 were used as the basis of comparison to determine optimal parameters for rank (k) and Jaccard
701 length on each dataset. For each Jaccard length the number of discrete communities was
702 plotted against all values of k, and a loess regression was performed to facilitate curve
703 smoothing. The inflection point of these concave curves was calculated using the 'elbow' R
704 package (https://nicolascasajus.fr/elbow/). Cost values, representing residual values after the
705 inflection point, were used to evaluate the stability of the solution. A linear regression was
706 performed on cost values of each Jaccard length curve, and the slope was recorded. The
707 optimal Jaccard length was established by determining which slope was closest to -1.0, which
708 represents a solution whereby the number of reproducible programs remains approximately
709 constant for all values of k greater than ground truth (Supplementary Figure 1d). Using the
710 established parameters for rank (k) and Jaccard length, final reproducible programs for each

711 dataset were calculated by averaging over all nodes (i.e., cNMF gene loading scores (i.e.
712 weights)) per community in the network graph.

713

714 **Identifying gene expression programs replicated across different datasets using a**
715 **network graph**

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

724

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

727 For each gene expression program, an interactive web-based report was created, which was
728 used to support annotation and interpretation. To estimate variability between subjective
729 annotators, the GOSH dataset was independently annotated by co-authors P.G. and Y.K, and a
730 consensus annotation was subsequently reached by discussing and merging these annotations,
731 including subjective assignment of confidence levels. The web-based reports were created
732 using the R package *knitr*[47], which for each program summarized a rich catalogue of biologically
733 relevant information/enrichments, including 282 marker gene lists we curated from the literature.
734 The interactive plots were created using *plotly*[48]. The broader network-based representation

735 comparing gene expression programs across datasets, as well as the tools for organizing the
736 reports and cataloguing/displaying relevant metadata was created using *Shiny*[49]. The complete
737 web platform is accessible at pscb.stjude.org.

738

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

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

756

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

758 We used InferCNV[34] to identify regions of copy number variation from single-cell RNA-seq,
759 which we used as evidence to differentiate between cancer and non-cancerous "normal" cells.
760 InferCNV was run on each patient/sample independently. Reference normal cells were selected
761 from clear immune cells, identified by acNMF (selecting cells expressing an immune program at
762 $h > 100$). In the case of the GOSH dataset, we used the author's original immune cell
763 annotations, which were essentially identical to the acNMF annotations. In cell lines, we defined
764 the reference normal using immune cells from the Dong dataset. InferCNV's six-state Hidden
765 Markov Model (i6) predicted CNV regions and states, with posterior probabilities of CNV regions
766 assigned to a given state using a Bayesian network at a value of P(Normal) = 0.3 (all other
767 parameters were used at default settings).

768

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

771 To predict the cancer/normal status of mesenchymal-like cells on a per-cell basis, we tested 3
772 classification models, a regularized logistic generalized linear model (GLM), a random forest
773 (RF), and a support vector machine (SVM), which all achieved comparable performance. These
774 were implemented in the R packages *glmnet*, *randomForest,* and *SVM* (linear kernel)
775 respectively. We identified cancer cells, endothelial cells, or ambiguous mesenchymal-like cells
776 if an appropriate program was expressed in that cell at $h > 100$. We randomly split clear cancer
777 cells into 75% training set and 25% test set. Cell-specific copy number estimates, which were
778 treated as the model's input features, were obtained from inferCNV's predicted regions of copy
779 number variation (inferCNV-features) or were obtained manually, by taking the average
780 standardized TPM expression level in groups of 50 collocated genes along the genome
781 (manual-features). For each sample, a classifier was then trained using sample-specific cancer
782 cells and immune cells (representing normal cells). For each sample, these trained models were

783 then applied to the held-out reference normal endothelial cells, or the held-out ambiguous
784 mesenchymal-like cells, outputting a predicted class of 0 (normal) or 1 (cancer) for each cell.
785 These results were then assembled into a 2 × 2 contingency table, where we applied a Fisher's
786 exact test to compare the frequency with which the model classified mesenchymal-like cells as
787 cancer cells, against the frequency it misclassified reference normal endothelial cells as cancer
788 cells. Note that for model training/testing in the GEMM data, because the mice were genetically
789 identical, we pooled the immune cells and endothelial cells in the drug-treated and vehicle-
790 treated groups, which increased the size of the reference normal pool of cells and should
791 improve power.

792

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

795 We performed differential expression analysis of our cisplatin treated vs vehicle treated GEMM
796 using the R package DESeq2[50] by 'pseudobulking' cells from each mouse using the
797 aggregate.Matrix function in the R package Matrix.utils. All genes were ranked by logFC
798 between vehicle and cisplatin treated mice, and then were subjected to GSEA[51] for functional
799 enrichments of the mSigDB Hallmarks gene set, using default parameter settings. Trajectory
800 inference and pseudotime analysis were performed with the R package Monocle3[52]. RNA
801 velocity was assessed by applying the dynamical module of scVelo[52] (v0.2.5) under default
802 settings.

803

804 **Spatial pathology analyses and immunohistochemistry.**

805 For H&E staining, the harvested mouse tumor tissue was embedded in Tissue-Tek O.C.T.
806 embedding medium (Sakura Finetek) and flash frozen on a metal block in liquid nitrogen then

807 preserved at –80°C until further processing. Each tissue block was equilibrated to -20°C in a
808 cryostat chamber (CM3050S, Leica BioSciences) for 30 minutes before making 10-µm sections
809 collected onto Superfrost slides. The slides were incubated at 37°C for 1 minute and transferred
810 to a container containing pre-chilled methanol for 30 minutes at -20C in the upright position.
811 After methanol fixing, the excess methanol was wiped out from the bottom of the slides without
812 disrupting the tissue section. After methanol fixation, the slides were immersed into water 10
813 times gently and incubated in hematoxylin solution for 1 minute followed by immersing slides 10
814 times into 800 ml water. The slides were then submerged 10 times into Bluing buffer and
815 immersed into water to remove the excess Bluing buffer. Next, slides were dipped into 96%
816 ethanol 10 times followed by incubation with Eosin for 30 seconds. The excess Eosin was
817 removed by immersing slides 10 times into 3 different containers containing 100% ethanol. After
818 that, tissue slides were mounted, and images were acquired using Axioscan slide scanner
819 (Zeiss, Oberkochen, DE) up to a 20X scalable magnification. For our immunohistochemistry
820 analysis of fresh frozen slides used for the RNA ISH experiments serial unstained sections of
821 murine NB-like tumors were fixed in acetone for 20 minutes. Fixed slides were then wet loaded
822 on the Ventana Discovery Ultra autostainer (Roche, Basel, CH) using the Ventana Discovery
823 Ultra antigen retrieval CC1 (catalogue #950-500) for 64 minutes followed by the application of
824 anti-PHOX2B antibody (Abcam, ab183741) at a 1:100 dilution for 60 minutes. OmniMap anti-
825 rabbit HRP (catalogue #760-4311) for 16 minutes, followed by the ChromoMap DAB kit
826 (catalogue #760-159) for 8 minutes, followed by counterstained with hematoxylin, and the post-
827 counterstain Bluing Reagent (catalogue #760-2037) for 8 minutes was used for the detection
828 and visualization of PHOX2B in tissue sections. Slides immunolabeled for COL1A1 (Cell
829 Signaling Technology, Danvers, MA; catalogue #66948) were deparaffinized and underwent
830 HIER Staining performed on Leica Bond RX autostainer (Leica Biosystems, Bannockburn, IL)
831 using Epitope Retrieval solution 2 (ER2, AR9640, Lecia Biosystems) for 30 minutes followed by
832 incubation with the primary antibody at 1:500 for 15 minutes and visualization with a secondary

833 rabbit anti-mouse antibody (Abcam, ab133469) and the Bond Polymer Refine Detection kit
834 (DS9800, Leica Biosystems). Positive and negative tissue controls and isotype controls were
835 used for examining the specificity of staining and species cross-reactivity for both PHOX2B and
836 COL1A1 in comparison with the experimental tissue slides. Slides were then cleaned with
837 warm water and detergent, dehydrated, and coverslipped for pathology review and imaging.

838

839 **Data and code accessibility**

840 Simulated scRNA-seq datasets were obtained through CodeOcean
841 (https://doi.org/10.24433/CO.9044782e-cb96-4733-8a4f-bf42c21399e6) using the
842 Step1_Simulate.ipynb Jupyter Notebook. Default parameters, defined by the implementation of
843 the R package 'splatter' on a published scRNA-seq dataset, were used to generate five replicate
844 simulations across three conditions of varying signal:noise levels. The raw readcounts from
845 each of these simulations were subjected to the acNMF approach described below
846 (https://rchapple2.github.io/acNMF/). The gene and cell parameters generated from this code
847 were used to evaluate the recovery of programs following the acNMF analysis. Human
848 datasets were downloaded via the following accession numbers. Dong: GSE137804. Jansky:
849 EGAS00001004388. Kildisiute: EGAD00001008345. Cell line and mouse data deposited in
850 GSE229224 and GSE229226, respectively. PDX deposited in GSE228957.

851

852 **ACKNOWLEDGEMENTS**

853 P.G. is supported by an NIGMS R35 award [R35GM138293] an R01 grant from NCI
854 [R01CA260060]; K99/R00 [R00HG009679] from NHGRI; P.G also receive support from ALSAC.
855 CWL received funding from the St. Jude Graduate School of Biomedical Sciences. The content

is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors have declared that no conflict of interest exists.

858 **REFERENCES**

859 1 Johnsen, J. I., Dyberg, C. & Wickström, M. Neuroblastoma—a neural crest derived embryonal
860 malignancy. *Frontiers in molecular neuroscience* **12**, 9 (2019).
861 2 Armstrong, G. T. *et al.* Reduction in late mortality among 5-year survivors of childhood cancer.
862 *New England Journal of Medicine* **374**, 833-842 (2016).
863 3 Moreno, L. *et al.* Accelerating drug development for neuroblastoma: summary of the second
864 neuroblastoma drug development strategy forum from innovative therapies for children with
865 cancer and International Society of Paediatric Oncology Europe Neuroblastoma. *European*
866 *Journal of Cancer* **136**, 52-68 (2020).
867 4 Ma, X. *et al.* Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and
868 solid tumours. *Nature* **555**, 371-376 (2018).
869 5 Oren, Y. *et al.* Cycling cancer persister cells arise from lineages with distinct programs. *Nature*
870 **596**, 576-582 (2021).
871 6 Boeva, V. *et al.* Heterogeneity of neuroblastoma cell identity defined by transcriptional
872 circuitries. *Nature genetics* **49**, 1408-1413 (2017).
873 7 Chen, X., Pappo, A. & Dyer, M. A. Pediatric solid tumor genomics and developmental pliancy.
874 *Oncogene* **34**, 5207-5215 (2015).
875 8 Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer.
876 *Science* **357**, eaal2380 (2017).
877 9 Shaffer, S. M. *et al.* Rare cell variability and drug-induced reprogramming as a mode of cancer
878 drug resistance. *Nature* **546**, 431-435 (2017).
879 10 Van Groningen, T. *et al.* Neuroblastoma is composed of two super-enhancer-associated
880 differentiation states. *Nature genetics* **49**, 1261-1266 (2017).
881 11 Olsen, T. K. *et al.* Malignant Schwann cell precursors mediate intratumoral plasticity in human
882 neuroblastoma. *bioRxiv* (2020).
883 12 Yuan, X. *et al.* Single-cell RNA-sequencing of peripheral neuroblastic tumors reveals an
884 aggressive transitional cell state at the junction of an adrenergic-mesenchymal
885 transdifferentiation trajectory. *bioRxiv* (2020).
886 13 Kameneva, P. *et al.* Single-cell transcriptomics of human embryos identifies multiple
887 sympathoblast lineages with potential implications for neuroblastoma origin. *Nature genetics*
888 **53**, 694-706 (2021).
889 14 Dong, R. *et al.* Single-cell characterization of malignant phenotypes and developmental
890 trajectories of adrenal neuroblastoma. *Cancer Cell* **38**, 716-733. e716 (2020).
891 15 Kildisiute, G. *et al.* Tumor to normal single-cell mRNA comparisons reveal a pan-neuroblastoma
892 cancer cell. *Science advances* **7**, eabd3311 (2021).
893 16 Bedoya-Reina, O. *et al.* Single-nuclei transcriptomes from human adrenal gland reveal distinct
894 cellular identities of low and high-risk neuroblastoma tumors. *Nature communications* **12**, 1-15
895 (2021).
896 17 Jansky, S. *et al.* Single-cell transcriptomic analyses provide insights into the developmental
897 origins of neuroblastoma. *Nature genetics* **53**, 683-693 (2021).
898 18 Kotliar, D. *et al.* Identifying gene expression programs of cell-type identity and cellular activity
899 with single-cell RNA-Seq. *Elife* **8**, e43803 (2019).
900 19 Pasquini, G., Arias, J. E. R., Schäfer, P. & Busskamp, V. Automated methods for cell type
901 annotation on scRNA-seq data. *Computational and Structural Biotechnology Journal* **19**, 961-969
902 (2021).

903 20 Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional
904 profibrotic macrophage. *Nature immunology* **20**, 163-172 (2019).
905 21 Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, eaba7721 (2020).
906 22 Domínguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in
907 humans. *Science* **376**, eabl5197 (2022).
908 23 Hanemaaijer, E. S. *et al.* Single-cell atlas of developing murine adrenal gland reveals relation of
909 Schwann cell precursor signature to neuroblastoma phenotype. *Proceedings of the National*
910 *Academy of Sciences* **118**, e2022350118 (2021).
911 24 Hsiao, C. J. *et al.* Characterizing and inferring quantitative cell cycle phase in single-cell RNA-seq
912 data analysis. *Genome research* **30**, 611-621 (2020).
913 25 Kildisiute, G., Young, M. D. & Behjati, S. Pitfalls of applying mouse markers to human adrenal
914 medullary cells. *Cancer Cell* **39**, 132-133 (2021).
915 26 Kinker, G. S. *et al.* Pan-cancer single-cell RNA-seq identifies recurring programs of cellular
916 heterogeneity. *Nature genetics* **52**, 1208-1218 (2020).
917 27 Lavie, D., Ben-Shmuel, A., Erez, N. & Scherz-Shouval, R. Cancer-associated fibroblasts in the
918 single-cell era. *Nature cancer* **3**, 793-807 (2022).
919 28 Veglia, F., Sanseviero, E. & Gabrilovich, D. I. Myeloid-derived suppressor cells in the era of
920 increasing myeloid cell diversity. *Nature Reviews Immunology* **21**, 485-498 (2021).
921 29 Bedoya-Reina, O. C. & Schlisio, S. Chromaffin cells with sympathoblast signature: too similar to
922 keep apart? *Cancer Cell* **39**, 134-135 (2021).
923 30 Kameneva, P. *et al.* Evolutionary switch in expression of key markers between mouse and
924 human leads to mis-assignment of cell types in developing adrenal medulla. *Cancer Cell* **39**, 590-
925 591 (2021).
926 31 Klöppel, G. Neuroendocrine neoplasms: dichotomy, origin and classifications. *Visceral medicine*
927 **33**, 324-330 (2017).
928 32 Zeineldin, M., Patel, A. G. & Dyer, M. A. Neuroblastoma: When differentiation goes awry.
929 *Neuron* (2022).
930 33 Althoff, K. *et al.* A Cre-conditional MYCN-driven neuroblastoma mouse model as an improved
931 tool for preclinical studies. *Oncogene* **34**, 3357-3368 (2015).
932 34 Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary
933 glioblastoma. *Science* **344**, 1396-1401 (2014).
934 35 Stewart, E. *et al.* Orthotopic patient-derived xenografts of paediatric solid tumours. *Nature* **549**,
935 96-100 (2017).
936 36 Zhu, Q.-Q., Ma, C., Wang, Q., Song, Y. & Lv, T. The role of TWIST1 in epithelial-mesenchymal
937 transition and cancers. *Tumor Biology* **37**, 185-197 (2016).
938 37 Boen, J., Wagner, J. P. & Di Nanni, N. Inferring copy number variation from gene expression
939 data: methods, comparisons, and applications to oncology. *bioRxiv*, 2021.2010. 2018.463991
940 (2021).
941 38 Du, B. & Shim, J. S. Targeting epithelial–mesenchymal transition (EMT) to overcome drug
942 resistance in cancer. *Molecules* **21**, 965 (2016).
943 39 Shibue, T. & Weinberg, R. A. EMT, CSCs, and drug resistance: the mechanistic link and clinical
944 implications. *Nature reviews Clinical oncology* **14**, 611-629 (2017).
945 40 Mañas, A. *et al.* Clinically relevant treatment of PDX models reveals patterns of neuroblastoma
946 chemoresistance. *Science Advances* **8**, eabq4617 (2022).
947 41 Teitz, T. *et al.* Preclinical models for neuroblastoma: establishing a baseline for treatment. *PloS*
948 *one* **6**, e19133 (2011).
949 42 Stewart, E. *et al.* The Childhood Solid Tumor Network: A new resource for the developmental
950 biology and oncology research communities. *Developmental biology* **411**, 287-293 (2016).

951 43 Slyper, M. *et al.* A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human
952 tumors. *Nature medicine* **26**, 792-802 (2020).
953 44 Stewart, E. *et al.* Development and characterization of a human orthotopic neuroblastoma
954 xenograft. *Developmental biology* **407**, 344-355 (2015).
955 45 Chung, N. C., Miasojedow, B., Startek, M. & Gambin, A. Jaccard/Tanimoto similarity test and
956 estimation methods for biological presence-absence data. *BMC bioinformatics* **20**, 1-11 (2019).
957 46 Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal,*
958 *complex systems* **1695**, 1-9 (2006).
959 47 Xie, Y. in *Implementing reproducible research* 3-31 (Chapman and Hall/CRC, 2018).
960 48 Sievert, C. *Interactive web-based data visualization with R, plotly, and shiny*. (CRC Press, 2020).
961 49 Chang, W., Cheng, J., Allaire, J., Xie, Y. & McPherson, J. Shiny: web application framework for R.
962 *R package version* **1**, 2017 (2017).
963 50 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for
964 RNA-seq data with DESeq2. *Genome biology* **15**, 1-21 (2014).
965 51 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for
966 interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*
967 **102**, 15545-15550 (2005).
968 52 Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by
969 pseudotemporal ordering of single cells. *Nature biotechnology* **32**, 381-386 (2014).

970