Title – Tumor to normal single cell mRNA comparisons reveal a pan-neuroblastoma cancer cell

Gerda Kildisiute[1†], Waleed M. Kholosy[2†], Matthew D. Young[1†*], Kenny Roberts[1], Rasa Elmentaite[1], Sander R. van Hooff[2], Clarissa N. Pacyna[1], Eleonora Khabirova[1], Alice Piapi[3], Christine Thevanesan[3], Eva Bugallo Blanco[3], Christina Burke[3], Lira Mamanova[1], Kaylee M. Keller[2], Karin P.S. Langenberg-Ververgaert[2], Philip Lijnzaad[2], Thanasis Margaritis[2], Frank C.P. Holstege[2], Michelle L. Tas[2], Marc H.W.A. Wijnen[2], Max M. van Noesel[2], Ignacio del Valle[3], Giuseppe Barone[4], Reinier van der Linden[5] ,Catriona Duncan[4], John Anderson[3,4], John C. Achermann[3], Muzlifah Haniffa[1,6,7], Sarah A. Teichmann[1], Dyanne Rampling[4], Neil J. Sebire[4], Xiaoling He[8,9], Ronald R. de Krijger[2,10], Roger A. Barker[8,9], Kerstin B. Meyer[1], Omer Bayraktar[1], Karin Straathof[3,4*], Jan J. Molenaar[2*], Sam Behjati[1,11,12*]

Affiliations

[1]Wellcome Sanger Institute, Hinxton CB10 1SA, UK.

[2]Princess Máxima Center for Pediatric Oncology, Heidelberglaan 25, 3584 CS Utrecht, Netherlands.

[3]UCL Great Ormond Street Institute of Child Health, WC1N 1EH, London, UK.

[4]Great Ormond Street Hospital NHS Foundation Trust, WC1N 3JH, London, UK.

[5]Hubrecht Institute, KNAW, 3584 CT Utrecht, The Netherlands.

[6]Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne NE2 4HH, UK.

[7]Department of Dermatology and NIHR Newcastle Biomedical Research Centre, Newcastle Hospitals, NHS Foundation Trust, Newcastle upon Tyne NE2 4LP, UK.

[8]MRC-WT Cambridge Stem Cell Institute, University of Cambridge, CB2 0QQ, Cambridge, UK.

[9]Department of Clinical Neuroscience, University of Cambridge, Cambridge, CB2 0QQ, UK.

[10]Dept. of Pathology, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands.

[11]Cambridge University Hospitals NHS Foundation Trust, Cambridge CB2 0QQ, UK.

[12]Department of Paediatrics, University of Cambridge, Cambridge CB2 0QQ, UK.

[†]These authors contributed equally to this work.

*Corresponding authors.

Email: my4@sanger.ac.uk (M.D.Y..); k.straathof@ucl.ac.uk (K.S.); j.j.molenaar@prinsesmaximacentrum.nl (J.J.M.); sb31@sanger.ac.uk (S.B.)

Abstract: Neuroblastoma is a childhood cancer that resembles developmental stages of the neural crest. It is not established what developmental processes neuroblastoma cancer cells represent. Here, we sought to reveal the phenotype of neuroblastoma cancer cells by comparing cancer (n=19,723) with normal fetal adrenal single cell transcriptomes (n=57,972). Our principal finding was that the neuroblastoma cancer cell resembled fetal sympathoblasts, but no other fetal adrenal cell type. The sympathoblastic state was a universal feature of neuroblastoma cells, transcending cell cluster diversity, individual patients and clinical phenotypes. We substantiated our findings in 650 neuroblastoma bulk transcriptomes and by integrating canonical features of the neuroblastoma genome with transcriptional signals. Overall, our observations indicate that there exists a pan-neuroblastoma cancer cell state which may be an attractive target for novel therapeutic avenues.

One Sentence Summary: Comparisons of neuroblastoma cells and relevant normal adrenal cells reveal a uniform cell state that characterizes neuroblastoma across patients and disease phenotypes.

Introduction

Neuroblastoma is a childhood cancer that exhibits a diverse pattern of disease (*1*), from spontaneously resolving tumors to a highly aggressive cancer. Neuroblastoma arises from aberrant differentiation of the neural crest, mostly within the fetal adrenal medulla, via the sympathetic lineage. Different stages of adrenal medullary cell differentiation have been implicated as the cell of origin of neuroblastoma (*2–4*) and proposed to underlie the diversity of clinical phenotypes *(2, 5)*.

Advances in single cell transcriptomics have enabled the direct comparison of cancer and normal reference cells. Applied to childhood cancer, such analyses have revealed specific cell types or developmental processes that tumors adopt and adapt (*6*). Here, we sought to establish, through cancer to normal single cell mRNA comparisons, which developmental processes and cell types neuroblastoma recapitulates and correlate our findings with disease phenotypes.

Results

Building a reference of human fetal medullary cells

In the first instance, we built a normal cell reference for our analyses, by quantitatively defining transcription of cells underpinning human adrenal development. We undertook single cell mRNA sequencing (10x Genomics Chromium platform) of seven adrenal glands obtained from first and second trimester human fetuses (**Fig. 1A, table S1**). To verify technical and biological replicability we included bilateral adrenal glands in two cases. Following stringent data filtering, including removal of doublets and ambient mRNAs, we obtained count tables of gene expression from a total of 57,972 cells. These cells could broadly be divided into cortical, medullary, mesenchymal cells, and leukocytes (**Fig. 1B, fig. S1A**), using canonical human adrenal markers (**fig. S1B**, **table S2**). We focused our analyses on medullary cells (n=6,451), from which most cases of neuroblastoma arise (*7*). The absolute and relative number of medullary cells varied across medullas, likely reflecting technical and biological differences, such as the relative proportion of cortex to medulla (**table S1**).

Recent single cell analyses of murine fetal adrenals glands have provided transcriptional definitions of four principal medullary cell types (*8–10*) : Schwann cell precursor (SCP), bridge cells, chromaffin cells, and sympathoblasts. To search for these cell types within our data, we used two orthogonal unbiased methods (logistic regression; canonical correlation analysis) to quantitatively map murine reference signals onto our human data. Unlike manual annotation of cells by a selection of (biased) marker genes, these approaches avoid the risk of misclassifying cells when interspecies differences on individual markers exist. Accordingly, we identified the human correlates of the four medullary cell types (**Fig. 1C-D)**. Of particular note were the neural crest like SCP cells, identified previously in murine (*8*), but not human fetal tissues. We validated these cells by single molecule FISH of SCP defining markers (SOX10, ERBB3, MPZ, PLP1) in the fetal medulla and other human fetal organs - fetal gut, eye, skin, and bone (**fig. S2**). Comparing the four medullary cell types across different adrenals did not reveal significant differences in their overall transcriptional "identity" (**Fig. 1E**), although nuances in transcription may exist across gestational ages (**table S3**). We therefore merged cell signals from different adrenals to form the reference for cancer-to-normal cell comparisons and for additional analyses on fetal medullary cells, including trajectory analyses and comparison to murine data (**fig. S1C, fig. S3**).

Identification of cancer cells in neuroblastoma

We next sought to establish the relationship between neuroblastoma cells and human fetal medullary development. We generated single cell mRNA readouts from 21 fresh neuroblastoma specimens from two different centres, using two platforms (Chromium 10x, CEL-seq2). We analysed data from these different platforms and centres independently to ascertain technical and biological replicability of our findings. We obtained tissue from untreated patients at diagnosis (n=4), or following treatment with cytotoxic agents at resection (n=17) (**table S1**). As pre-treated neuroblastomas tend to be largely necrotic at surgical resection, we guided sampling to viable tumor areas through pre-operative metabolic cross-sectional imaging in these specimens (meta-iodobenzylguanidine scan), coupled with morphological assessment of frozen sections in 15/17 pretreated cases. Our study cohort represented the three principal prognostic categories of neuroblastoma: low (n=3), intermediate (n=5), and high risk (n=13) cases. In total we obtained 19,723 cells, with variable contribution from each tumor (**Fig. 2A-B; table S1**), which

segregated into four main cell types: leukocytes (n=10,593), mesenchymal cells (n=4,227), cells bearing markers of Schwannian stroma (n=665), and putative tumor cells exhibiting adrenal medullary-like features (n=3,396).

The first challenge was to identify *bonafide* cancer cells amongst tumor derived cells. Marker based cell typing alone was of limited value as controversy exists as to which cell types in neuroblastoma represent cancer cells. Although there is a general consensus that neuroblastoma cancer cells exhibit medullary-like features, it has also been suggested that interstitial (mesenchymal) and Schwannian stroma cells, commonly found in neuroblastoma, may be cancerous (*11–13*). We therefore used somatic copy number changes to identify cancer cells, by interrogating each cell's mRNA sequence for evidence of the somatic copy number changes underpinning each tumor. Extending a previously developed method (*6*), we integrated single cell tumor RNA information with single nucleotide polymorphisms (SNP) arrays or whole genome sequencing of the tumor DNA (**table S4**). This allowed us to assess the allelic imbalance of SNPs in each cell's mRNA sequence across patient specific copy number segments. This analysis revealed that in our cohort only adrenal medullary-like cells were cancerous, but not interstitial or Schwannian stroma cells (**Fig. 2C-D**).

Comparison of cancer to normal cells at single cell resolution

Next, we investigated which stage of adrenal medullary development cancer cells recapitulate by performing cancer to normal cell comparisons, using two independent methods. We found that neuroblastoma cancer cells did not recapitulate adrenal development but had assumed the state of sympathoblasts (**Fig. 2E-G, fig. S4A**). Importantly, although cancer cells formed discrete clusters within, and across patients (**Fig. 2A-B**), the cancer to normal cell comparison resolved this diversity into a common sympathoblast-like phenotype (**Fig. 2H-I**). The sympathoblast state was present in all patients (**Fig. 2H-I**).

To validate that neuroblastoma cancer cells principally resemble sympathoblasts, we interrogated neuroblastoma bulk transcriptomes. We curated gene expression profiles of 650 neuroblastomas from two different clinically annotated cohorts that represent the entire clinical spectrum of neuroblastoma (TARGET (n=152) (*14*) and SEQC (n=498) (*15*) cohorts). We probed these

expression data for transcripts that define the four populations of the normal human fetal medulla: SCPs, bridge cells, sympathoblasts, and chromaffin cells (**fig. S1B**). Across cohorts, we found a clear signal of sympathoblast mRNAs (**Fig. 3A**), thus verifying the sympathoblastic state of neuroblastoma. The sympathoblast signal in bulk transcriptomes was reproduced by using murine reference data (**fig. S4C**). Neuroblastomas that arose outside the adrenal gland also exhibited sympathoblast signals (**Fig. 3B**). The sympathoblast signal was present across the principal risk groups of neuroblastoma, suggesting that it transcends the clinical diversity of neuroblastoma (**Fig. 3C**).

Features of the neuroblastoma genome corroborate the sympathoblast state

To further corroborate the transcriptional evidence that neuroblastoma cells resemble sympathoblasts, we overlaid features of the neuroblastoma cancer genome with transcriptional developmental pathways. We first analysed how somatic cancer (*14*) and germline predisposition (*16*) genes of neuroblastoma are utilised by fetal medullary cells. We found that the majority of these genes were most highly expressed in sympathoblasts (**Fig. 3D; fig. S5A**), in particular the two most common neuroblastoma predisposition genes, *ALK* and *PHOX2B*, and the cancer gene, *MYCN*, somatic amplification of which is a key adverse marker used clinically for neuroblastoma risk stratification. Thus, the same genes that operate as cancer genes in neuroblastoma are utilised in normal development predominantly by sympathoblasts. Similarly, genes encoding the synthases of disialoganglioside (GD2), which is a near ubiquitous marker of neuroblastoma that is therapeutically targeted by antibody treatment (*17*), were predominantly expressed in sympathoblasts (**Fig. 3D**). We note that the expression of the enzyme catalyzing the final step in GD2 synthesis, *B4GALNT1*, was confined to sympathoblasts, whereas genes further upstream in the synthesis pathway, *ST3GAL5* and *ST8SIA1*, showed a less restricted pattern of expression within the medulla.

A variety of copy number changes have been described in neuroblastoma, the most recurrent and pertinent of which occur on chromosomes 1, 11, and 17 (*18*). The presence of these copy number changes carries fundamental prognostic significance and determines treatment intensity in most clinical contexts and treatment protocols. Given the sympathoblastic state of neuroblastoma, it seemed conceivable that somatic copy number changes may impact on the

expression of genes that define sympathoblasts. We were able to directly test this hypothesis. For example, we compared gene expression of cancer cells that harbour loss of chromosomes 1 and 11 with cancer cells that do not carry these changes (**table S5**). We found that the resulting differentially expressed genes were significantly enriched for high confidence sympathoblast marker genes (tf-idf>1, 283 out of 329 markers, $p<0.0001$, hypergeometric test) (**table S6**). Furthermore, the majority of these genes (254/329) exhibited lower expression in cells with chromosome 1 and 11 loss, even if those genes did not reside on chromosomes 1 or 11. Following on from this, we correlated genomic regions of recurrent chromosome 1 or 11 loss with the genomic localisation of gene expression in fetal medullary cells. The localisation of the boundaries of altered copy number segments is associated with gene expression in human cancer (*19*). Thus, the genomic position of these boundaries may encode information about the expression profile of the cancer cell of origin. We found regions of sympathoblast specific gene expression that statistically significantly overlapped with the most common boundaries of chromosome 11q loss ($p<0.05$; permutation test) (**Fig. 3F; fig. S5B**) and the most recurrent region of chromosome 1p loss ($p<0.05$; permutation test) (**Fig. 3F; fig. S5C**), thus further corroborating the sympathoblast state of neuroblastoma.

Differences between low and high risk tumors

The measurement of medullary cell signals in bulk transcriptomes indicated that compared to lower risk tumors, the sympathoblast signal was less pronounced in high risk tumors in both cohorts (**Fig 3C**). This weakened sympathoblast signal may be driven by differences in tissue composition, or by distinct properties of cancer cells themselves. We reanalysed the sympathoblast signal in single cancer cells, this time subdividing neuroblastoma cells by whether they originated from low or from high risk tumors. We found that some high, but not low risk cells, exhibited a significantly decreased sympathoblast signal (**Fig 3E**). This difference could not be accounted for by technical variation (**fig. S4B**). Differential gene expression between the two groups of high risk cells showed that the weakened sympathoblast signal was underpinned by, amongst other mRNAs, transcripts defining neuronal and neuroendocrine lineages (**table S7**). In aggregate, these findings indicate that a subtype of high risk cell may exist that is characterised by a fainter sympathoblast signal.

Utilisation of fetal transcripts in neuroblastoma

As neuroblastoma cells utilise transcripts of fetal medullary cells, it is conceivable that the expression of some of these genes is relatively restricted to the fetal period and may thus represent formidable therapeutic targets. The precedent for this concept is the disialoganglioside, GD2, which is widely expressed by neuroblastoma cells and some fetal lineages including sympathoblasts (**Fig. 4A**). In post-natal tissues, however, expression of key genes encoding GD2 synthases is most confined to the central and peripheral nervous systems (*20*). Therapeutic antibodies targeting GD2 are deployed in most treatment protocols of high risk neuroblastoma. Although pain, probably mediated by expression in peripheral nerves, is a common adverse effect, significant CNS toxicity is rare. To identify transcripts with a similar distribution to GD2 synthases, we searched for mRNAs utilised by single neuroblastoma and fetal medullary cells (hereinafter referred to as fetal cancer transcript). We then assessed their distribution across pan-body post-natal tissues (from GTEx data), whilst validating their expression in neuroblastoma bulk transcriptomes. These analyses revealed two patterns of fetal cancer gene expression across post-natal tissues. One pattern followed the distribution of mRNAs encoding GD2, whereby expression of fetal cancer transcripts were predominantly expressed in neuroblastoma and in CNS tissues. Some genes, including those pertaining to catecholamine synthesis, exhibited an even more restricted expression pattern. Together these analyses reveal a number of fetal cancer genes that may lend themselves as targets in neuroblastoma (**table S8**).

Differences between neuroblastoma and normal fetal medullary cells

Having established the similarities between neuroblastoma cancer cells and fetal medullary development, we now looked for differences. Previous bulk mRNA analyses have described gene expression changes that characterise neuroblastoma tissues (*3*) (*21*) . Our ability to directly compare cancer cells with their normal counterpart, fetal medullary cells, enabled us to distil the transcriptional essence of neuroblastoma cells, which comprises 91 differentially expressed genes (tf-idf > 0.85, **table S9**), significantly enriched . To verify, and to correlate, these findings with disease phenotypes, we measured mRNA levels of these differentially expressed genes in bulk neuroblastoma transcriptomes and ascertained that a significant number of them had above average gene expression (n=37/89, p<0.001, hypergeometric test; **fig. S6**). The expression of 10 of these varied significantly by clinical risk group, even after controlling for key determinants of

clinical risk, age and *MYCN* amplification (negative binomial generalised linear model) (**Fig. 4B**). In addition to known markers of clinical risk (e.g., *PRAME*), we identified novel markers that have not previously been described and could in theory be used to improve risk stratification in clinical practice in the future. Biologically, the transcription factor gene, *SIX3*, may be one of the most interesting findings amongst differentially expressed genes, representing an embryonal gene that normally regulates forebrain and eye development (*22*).

Discussion

The principal finding of our investigation was that the neuroblastoma cancer cell represented an aberrant fetal sympathoblast. Our expectation might have been that neuroblastoma exhibits a cellular hierarchy adopted from adrenal development, similar to, for example, the childhood kidney cancer, Wilms tumor, recapitulating fetal nephrogenesis. However, neither our direct analyses of single cancer cells, nor indirect insights from bulk cancer transcriptomes would corroborate such a hypothesis.

We identified cancer cells by directly assessing SNP ratios in regions of cancer defining copy number changes, rather than inferring copy number from gene expression levels. Accordingly, we did not find evidence that cells other than sympathoblast-like cells were malignant. Therefore, our findings suggest that the sympathoblastic state predominates across the entire spectrum of neuroblastoma. The transcriptional similarity between neuroblastoma cancer cells and sympathoblasts may indicate that neuroblastoma directly derives from sympathoblasts. However, it is also conceivable that neuroblastoma arises from other cell states of the neural crest lineage that assume the transcriptional state of sympathoblasts upon malignant transformation.

The neuroblastoma transcriptome has been studied extensively from bulk tissues, revealing a plethora of potential targets that have been evaluated over past decades. Nevertheless, leveraging the resolution of single cells to perform precise comparison of cell types, we identified additional potential targets in neuroblastoma including fetal genes that are utilised by neuroblastoma cells with a restricted expression profile in post-natal tissues. Furthermore, direct comparison of

cancer and fetal medullary cells revealed neuroblastoma genes that correlate with outcome in two independent validation cohorts, which may prove useful for refined risk stratification.

Overall our findings reveal a surprising homogeneity in the cellular identity of neuroblastoma. Although future, larger scale analyses may define the nuances of this signal and perhaps identify rare cancer cell types, our study would suggest that the principal normal cell correlate of neuroblastoma is the human fetal sympathoblast.

Materials and Methods

Ethics statement

Informed consent for research was obtained from participants (or their carers). Studies underlying this paper have received appropriate approval by ethics review boards as per national legislation. Dutch tumor samples were obtained through an institutionally approved research study. UK tumor samples were collected under the following studies: NHS National Research Ethics Service reference 16/EE/0394 (tumor samples) and NHS National Research Ethics Service reference 96/085 (fetal tissues). Additional fetal tissue was provided by the Joint MRC / Wellcome Trust-funded (grant # 099175/Z/12/Z) Human Developmental BiologyResource (HBDR, http://www.hdbr.org; (10)), with appropriate maternal written consent and approval from the Newcastle and North Tyneside NHS Health Authority Joint Ethics Committee. HDBR is regulated by the UK Human Tissue Authority (HTA;www.hta.gov.uk) and operates in accordance with the relevant HTA Codes of Practice.

Fetal adrenal tissue processing

Tissue was minced with a scalpel and then placed in 15ml falcon tube filled with 5ml Digestion solution (5ml, dilute Liberase TH stock solution (1 vial of 5mg Liberase TH (Sigma 5401135001) powder in 2ml of 1X PBS with final concentration is 2.5 mg/ml.) in 1X PBS (100ul enzyme stock + 4.9ml 1x PBS). Tissue was incubated in 37’C for 30min (water bath) and mixed with 1ml pipette after 15 minutes to facilitate the digestion. 5ml STOP solution (40ml, 2% FBS in 1x PBS (40ml 1x PBS + 800ul FBS)) was added. The mixture was filtered through a 30um strainer and spun down, 750g, 5min at 4’C. If the pellet was red 1ml of 1X RBC (00-4300-54) lysis buffer was added and incubated for 3min at RT. 10ml of STOP solution was added and cells were gently mixed to wash, then centrifuged, 750g, 5min, 4’C. Pellets were resuspended into appropriate amount of 1x PBS or STOP solution, counted and processed on 10x Chromium platform.

Neuroblastoma tissue processing - GOSH pipeline

Surplus tumor tissue obtained at diagnostic biopsy or tumor resection was processed immediately after receipt in the histopathology laboratory (< 1 hour after interventional radiology/surgical

procedure). Tissue was minced using a scalpel, and then incubated in RPMI, supplemented with 10% FCS, 1% L-glutamine and 1% Penicillin/Streptomycin, with Collagenase IV (1.6 mg/ml; cat. No. 11410982; MP Biomedicals), for 30 minutes at 37°C, inverting the tube every 10 minutes. The digested tissue was passed through a 70 µm filter and incubated in 1x Red Blood Cell lysis buffer (cat. No. 420301; Biolegend) for 10 minutes, at room temperature. The obtained single cell suspension was used for downstream processing. Part of the single cell suspension was depleted of $CD45^+$ cells to enrich for tumor cells using a CD45 MicroBeads kit (cat. No. 130-045-801, Miltenyi Biotec), following the manufacturer's protocol. Both CD45 non-depleted and CD45-depleted single cell suspensions were depleted of dead cells using a Dead Cell Removal Kit (cat. No. 130-090-101, Miltenyi Biotec), following the manufacturer's protocol. Obtained viable single cell suspensions (CD45-depleted: 2 channels and CD45-non depleted (1 channel) were processed on the 10x Chromium platform.

Neuroblastoma tissue processing - PMC pipeline

tumor tissue was collected, after obtaining informed consent of the parents of neuroblastoma patients and after the study was approved by the ethical committee at the Princess Maxima Centrum (PMC). MIBG-positive neuroblastoma samples were included if they were diagnosed with histologically proven viable tumor cells. Most samples were debulking neuroblastomas (patients were treated with debulking type surgical procedure) and preoperative radiation and chemotherapy had been given to the patients (**table S1**). We extensively optimized the workup protocols and samples were gently dissociated to the single-cell level and sorted by flow cytometry. Single-cell suspensions were subsequently subjected to single-cell RNA-Seq using the CEL-seq2 platform. In brief, freshly resected human neuroblastoma specimens were collected and processed immediately following histological confirmation of the presence of viable tumor tissue. Tumor tissue was minced into 3-4 mm pieces and digested with collagenase type I, II and IV (2 mg/ml) in the optimized medium at 37°C, with agitation for 2 hours. The resulting suspension containing small tumor fragments was passed through a 70µm cells strainer, washed with a cold organoid medium. Half of the tumor fragments were used to generate tumor organoids, while the other half were fragments were further dissociated with NeuroCult kit until the single-cell suspension was obtained. Cells were washed with cold organoid medium and stained with 7AAD or DAPI to enrich for live cells while others were stained with FITC-GD2 or

PE-GD2 antibodies to enrich for tumor and PE-CD3 antibodies to enrich for T cells and to exclude erythrocytes. Cells were stored on ice until sorting using a FACS Jazz or FACS AriaII. Single live cells were sorted into 384-well hard-shell plates with 10 µl mineral oil, 50 nl of RT primers, dNTPs and synthetic mRNA Spike-Ins and immediately spun down, snap-frozen on dry ice, and stored in -80C to proceed with the total transcriptome amplification, library preparation and sequencing.

10X library preparation and sequencing

The concentration of single cell suspensions was manually counted using a hemocytometer and adjusted to 1000 cells/ul or counted by flow cytometry. Cells were loaded according to the standard protocol of the Chromium single cell 3' kit (v2 and v3 chemistry). All the following steps were performed according to the standard manufacturer's protocol. We used one lane of an Illumina Hiseq 4000 per 10x chip position.

CEL-seq2 library preparation and sequencing

All samples that were processed proceeded with the total transcriptome amplification, library preparation and sequencing into Illumina sequencing libraries as previously described(*23*). Paired-end 2X75 bp sequencing read length was used to sequence the prepared libraries using the Illumina NextSeq sequencer.

Bulk DNA processing of GOSH samples

DNA was extracted from fresh frozen tissue. Peripheral blood DNA was used as a matched normal. Short insert (500bp) genomic libraries were constructed, flowcells prepared and 150 base pair paired-end sequencing clusters generated on the Illumina HiSeq X platform according to Illumina no-PCR library protocols(*24*). The average sequence coverage was 43X and 42X for tumor and matched peripheral blood samples, respectively.

Bulk DNA processing of PMC samples

DNA was extracted from fresh tissue via AllPrep DNA/RNA/Protein Mini Kit (QIAGEN) according to standard protocol.

Bulk DNA alignment

DNA sequencing reads were aligned to the GRCh 37d5 reference genome using the Burrows-Wheeler transform (BWA-MEM) (*25*). Sequencing depth at each base was assessed using Bedtools coverage v2.24.0.

Single molecule fluorescent in situ hybridisation

Fresh tissue samples were embedded in OCT and frozen at -80°C on an isopentane-dry ice slurry. Cryosections were cut at a thickness of 16 μm using a Leica CM3050 S cryostat and placed onto SuperFrost Plus slides (VWR). Prior to staining, tissue sections were post-fixed in 4% paraformaldehyde in PBS for 15 minutes at 4°C, then dehydrated through a series of 50%, 70%, 100%, and 100% ethanol, for 5 minutes each. Tissue sections were then processed using a Leica BOND RX to automate staining with the RNAscope Multiplex Fluorescent Reagent Kit v2 Assay and RNAscope 4-plex Ancillary Kit for Multiplex Fluorescent Reagent Kit v2 (Advanced Cell Diagnostics, Bio-Techne), according to the manufacturers' instructions. Automated processing included pre-treatment with Protease IV for 30 minutes, but no heat treatment. Tyramide signal amplification with Opal 520, Opal 570, and Opal 650 (Akoya Biosciences) was used to develop three probe channels. The fourth was developed using TSA-biotin (TSA Plus Biotin Kit, Perkin Elmer) and streptavidin-conjugated Atto 425 (Sigma Aldrich). Stained sections were imaged with a Perkin Elmer Opera® Phenix™ High-Content Screening System, in confocal mode with 1 μm z-step size, using 20× (NA 0.16, 0.299 μm/pixel) and 40× (NA 1.1, 0.149 μm/pixel) water-immersion objectives.

Copy number detection from DNA data

The ascatNGS algorithm(*26*) (v4.0.1) was used to estimate tumor purity and ploidy and to construct copy number profiles prior to running the Battenberg algorithm (v2.2.5) (github.com/cancerit/cgpBattenberg) to allow for tumor sub-clonality in bulk DNA sequencing data. For DNA data measured with Illumina SNP arrays, copy number segments were defined by manual inspection of logR and B allele frequency data.

Mapping and quantification of scRNAseq - 10X

Single cell RNA-seq data were mapped and counts of molecules per barcode quantified using the 10X software package cellranger (versions 2.0.2 and 3.0.2) to map sequencing data to version 2.1.0 of the build of the GRCh38 reference genome supplied by 10X.

Mapping and quantification of scRNAseq - CEL-seq2

Sharq preprocessing and the QC pipeline have been applied to process the single-cell RNA-seq data as described (*27*). Read mapping has been done using STAR version 2.6.1 on the hg38 Patch 10 human genome. The featureCounts function of the subread package (version 1.5.2) was used to assign reads based on GENCODE version 26.

Quality control of fetal adrenal single cell data

Ambient mRNA contamination was removed from each channel individually using SoupX (*28*). Cleaned count data was then normalised, scaled to have a mean of 0 and standard deviation of 1 for each gene, principal component analysis was performed using highly variable genes, and grouped into clusters using a community detection finding algorithm taking the first 75 principal components as inputs and a resolution parameter of 10. This was performed using the Seurat package in R (*29*, *30*). The clustering parameter was deliberately set to this extremely high value to produce a very large number of small clusters.

Following clustering, we designated each cell as either passing or failing a series of quality control metrics. We marked as failing QC any cell with >30% expression due to mitochondrial genes, fewer than 300 genes detected, fewer than 1000 UMIs detected, or a doublet score greater than 0.2 as determined by Scrublet (*31*). In addition to excluding these individual cells, we also marked as failing QC any cell belonging to a high resolution cluster containing more than 50% of cells that fail QC due to one of the above filters. Finally, we excluded from all further analysis any cell marked as QC failed for any of the above reasons (**fig. S7A**).

Next we calculated a S and G2M phase score for each cell using Seurat (*29*, *30*) and excluded any cell where either score was greater than 0. This removed any cells that showed evidence of being in a phase of the cell cycle other than G0/G1. We excluded these cells from our reference map as cells in the S/G2M phase tend to cluster together based on their phase of cell cycle and not their underlying cell type, reducing the utility of the data as a reference map of cell types present in the developing adrenal gland.

As before, data was normalised by sequencing depth, scaled to 10,000 counts, log transformed, and genes were scaled to have mean 0 and standard deviation 1. Principal component analysis was performed on the scaled data using the most highly variable genes. We selected the number of principal components to use for clustering and visualisation to minimise the molecular cross validated error (*32*) (https://github.com/constantAmateur/MCVR). Using these principal components, we calculated a uniform manifold approximation and projection (*33*) two dimensional representation of the data for visualisation and calculated clusters using a k-nearest neighbours algorithm with resolution parameter set to 1.

Quality control of 10x tumor data

We excluded nuclear mitochondrial genes, heat shock proteins and ribosomal genes from our analysis. To filter lower quality cells, we removed any cells that had greater than 20% expression originating from mitochondrial genes, expressed fewer than 300 distinct genes, or contained fewer than 1000 UMIs (**fig. S7B**). Data was normalised by sequencing depth, scaled to 10,000 counts, log transformed, and genes were scaled to have mean 0 and standard deviation 1 using the Seurat 'NormalizeData' function. Principal component analysis was performed on the scaled data using the 2,000 most variable genes. Using 75 first principal components, we calculated a uniform manifold approximation and projection (UMAP (*33*)) two dimensional representation of the data for visualisation and calculated clusters using a community detection algorithm with resolution parameter set to 1. We performed these steps using the Seurat package in R (*29*, *30*).

Quality control of CEL-seq2 tumor data

We removed RNA spike-ins and genes that were not present in the 10X gene list. We excluded nuclear mitochondrial genes, heat shock proteins and ribosomal genes from our analysis. To filter lower quality cells, we removed any cells that had greater than 20% expression originating from mitochondrial genes, expressed fewer than 200 distinct genes, and contained fewer than 500 UMIs (**fig. S7C**). We used less stringent thresholds for CEL-seq data than for 10X data due to data quality. Data was normalised by sequencing depth, scaled to 10,000 counts, log transformed, and genes were scaled to have mean 0 and standard deviation 1 using the Seurat 'NormalizeData' function. Principal component analysis was performed on the scaled data using

2,000 most variable genes. Using 50 first principal components, we calculated a uniform manifold approximation and projection (UMAP (*33*)) two dimensional representation of the data for visualisation and calculated clusters using a community finding algorithm with resolution parameter set to 1. We performed these steps using the Seurat package in R (*29*, *30*).

Cluster annotation

Using well established marker genes of different cell types curated from the literature (**table S2**), we assigned a cell type to each cluster. Where two clusters were annotated as the same cell type, we merged them together. As further confirmation of our annotation, we next identified marker genes for each population algorithmically (**table S6**). To do this we used a method which uses the tf-idf metric to identify genes specific to each population, as implemented in the "quickMarkers" function in the SoupX R package (*28*). We further filtered genes to include only those genes with a p-value less than 0.01 after multiple hypothesis correction (hypergeometric test).

Comparison of mouse and human adrenal medulla

Using an established map of the mouse medulla as our gold standard (*8*) we mapped this reference to our human data in two orthogonal ways. For this comparison, we restricted both human and mouse transcriptomes to those genes that could be unambiguously mapped between species (1-to-1 orthologs). Using data from mice taken from E12.5 and E13.5 as independent references, we calculated a similarity score using logistic regression (*6*) for each human medulla cell for each of the four murine medullary populations. Additionally, we used a canonical correlation analysis based method to transfer the murine labels to the human medulla, as implemented in the Seurat R package (*29*). We identified differentially expressed genes between matching mouse and human medulla using edgeR (*34*) as described below. To obtain a list of genes with a large effect size difference between mouse and human, we filtered this list to include only genes with FDR less than 0.01 absolute log fold change greater than 1, fraction of cells expressed greater than 80% in the species with highest expression and less than 1% in the species with lowest expression. We further removed any gene that showed a significant difference between species in all four medullary populations (**table S10)**.

Differential expression in early and late medulla

To identify genes that were differentially expressed between early (<21 weeks) and late (>21 weeks) cell types in the adrenal medulla, we performed a differential gene expression analysis between early and late cells using FindMarkers function in Seurat for each medullary cell type (SCPs, Bridge, Sympathoblats and Chromaffin cells). We pre-filtered genes that were detected at <25% frequency in either population. We identified all genes that passed the 0.01 p-value threshold after multiple hypothesis correction (**table S3**).

Differential expression in high risk cancer cells with low and high sympathoblast signal

We categorised high risk tumor cells as "low sympathoblast signal" and "high sympathoblast signal" by applying k-means clustering with k=2 to the sympathoblast similarity score (logistic regression). To identify genes that were differentially expressed between low and high sympathoblast signal in high risk cancer cells, we performed a differential gene expression analysis between low sympathoblast signal and high sympathoblast signal cells using FindMarkers function in Seurat. We pre-filtered genes that were detected at <25% frequency in either population. We identified all genes that passed the 0.01 p-value threshold after multiple hypothesis correction (**table S7**).

Pseudotime analysis

We ordered cells into a trajectory based on the similarity of their transcriptomes to create a pseudotime ordering of the medulla using monocle3 (*35–37*). We constructed a monocle3 object using data extracted from a Seurat object, rather than re-processing the data in monocle3. We used the "learn_graph" function to build a principal graph of the data. We then used "order_cells" function to place cells on a pseudotime trajectory, selecting a node (as defined in "learn_graph") as a starting point of the trajectory. We used an SCP node based on prior knowledge from murine data that medullary cells originate from SCPs (*8*). In addition, to further investigate the identity spectrum between bridge and sympathoblasts, we performed the same analysis after subsetting the Seurat object to only include bridge cells and sympathoblasts.

Differential expression of genes along pseudotime trajectory

To identify transcription factors that were differentially expressed along the graph of the medulla, we performed Moran's I test using "graph_test" function in monocle3. We limited our gene set to the 1,665 transcription factors (37), and identified all genes that passed 0.001 p-value threshold after multiple hypothesis correction (**table S11**). In addition, to further investigate the identity spectrum between bridge and sympathoblasts, we performed the same analysis, this time using all genes, on just the bridge cells and sympathoblasts, and identified all genes that passed 0.001 p-value threshold after multiple hypothesis correction (**table S11**).

RNA Velocity

RNA velocity was calculated on the foetal adrenal reference map data using velocyto (*38*) to calculate the number of reads supporting spliced/unspliced isoforms of each gene in each cell. The velocity field projection to the UMAP embedding was then calculated using the velocyto.R R package, pooling 20 k nearest neighbours and performing a gamma fit to the top/bottom 2% expression quantiles.

Cell similarity calculation

To measure the similarity of a target single cell transcriptome to a reference single cell data-set we used the methodology based on logistic regression outlined in detail in (*6*). Briefly, we trained a logistic regression model with elastic net regularization (alpha=0.99) on the reference training set. We then used this trained model to infer a similarity score for each cell in the query data set for each cell type in the reference data.

Softmax normalization was not used to allow for the possibility that some cells in the query data set do not resemble any of the cell types in the reference data set. Predicted logits were averaged within each cluster in the query dataset. This approach was implemented using the "glmnet" package in R(*39*).

Genotyping individual cells in single cell RNA-seq data

The details of the genotyping of individual cells is explained in detail in the supplementary materials with a brief description provided here. For each sample, regions of clonal copy number change and heterozygous SNPs were identified from either whole genome sequencing or copy number arrays (**table S4**). Taking copy number regions resulting in allelic imbalance,

heterozygous SNPs were phased together across the entire copy number aberration. In each single cell transcriptome, the number of counts derived from either the major or minor allele was calculated for each copy number change, by investigating the nucleotide sequence of the mRNA reads allele counter ((*40*) (https://github.com/cancerit/alleleCount). For each cell, a posterior probability was calculated for both a diploid genome and the tumor genome. **Fig. 2 C-D** show the posterior probabilities of the tumor genotype for each cell. For downstream analysis, we considered only clusters of cells that had more cells that were definitively tumor (posterior probability > 99%) than definitively normal (posterior probability < 1%).

Identification of cell-type specific genes

To identify genes that were specific to one cell type and no other in the adrenal medulla, we regenerated algorithmically defined markers as above, but with podocyte cells (*6*) included as a negative control. To use only the most specific markers, we applied a tfidf cut-off of 1 to select genes that were specific to each medullary cell type relative to all other cells in the dataset. Note that a tf-idf cut-off of t implies that the global rate at which a gene occurs in cells is less than exp(-t/rate_local), where rate_local is the rate at which the gene occurs in the cells in the target cluster.

We also removed genes that were expressed in more than 20% of cells in any other single cluster in the dataset. The resulting gene list consisted of genes specific to each cell type in developing adrenal medulla("stringent_markers category in the "dataset" column, **table S6**).

Adrenal gland cell type signal in bulk RNA-seq data

To identify which adrenal cell type bulk RNA-seq Neuroblastomas resembled most, we measured the expression levels of genes defined in **table S6** in two bulk RNA-seq Neuroblastoma datasets, TARGET (*14*) and SEQC (*15*). For both data sets, the transcripts per million reads (TPM) value was calculated for each gene in each sample. We defined a gene as being "present" in a sample when its $\log_2$(TPM) expression value exceeded that of the peak of the $\log_2$(TPM) distribution across all genes across all samples. For each gene, we then calculated the fraction of samples in which it was present, both globally and stratified by risk group.

Defining neuroblastoma samples outside the adrenal gland

To define which neuroblastoma samples in TARGET (*14*) arose outside the adrenal gland, we filtered ICD-O(*41*) description of each tumor by sites of occurrence. We excluded all samples where descriptions were absent or contained words “kidney”, “adrenal” , “abdominal”, “abdomen”, “unknown”, “retroperitoneum”, “other”, identifying 21 samples where tumors occured outside the adrenal gland.

Differential expression between fetal adrenal medulla and tumor cells

To identify the key transcriptomic differences between Neuroblastoma tumor cells and the normal fetal medulla, we performed differential expression analysis between fetal medulla single cells and tumor cells from CEL-seq2 and 10X tumor datasets. For each dataset, we extracted tumor clusters and merged them with fetal adrenal medullary cells. We labelled all tumor cells as “tumor”, and used the “quickMarkers” function from SoupX R package (*28*) to identify tumor-specific genes for each dataset. We used a cut-off of 0.01 p-value after multiple hypothesis correction (hypergeometric test), and genes that were expressed in more than 20% of cells in any other single cluster in the dataset, to define a set of tumor markers.

We merged the two marker lists and calculated an average tfi-df value for each gene present in either or both lists, and used a tf-idf cut off of 0.85 to define a single list of tumor markers. We then filtered the list to exclude genes expressed by more than 25% of leukocytes in each tumor dataset (**table S9**).

Differential expression between low risk and high risk Neuroblastomas in TARGET

To test if any of the genes defined in **table S9** displayed risk group dependence, we compared their expression levels between 4S and high risk bulk Neuroblastomas in the TARGET data(*14*). To do this we used a negative binomial model, where we set the square root of the over-dispersion value to 0.4 (as genuine biological replicates were not available). We constructed a generalised linear model, with age, MYCN status, and risk group as covariates then used a quasi-likelihood F-test to calculate genes for which the risk group coefficient was significantly non-zero (0.05 FDR, edgeR (*34*) package) (**table S12**). We did not perform equivalent analysis in SEQC data as the raw counts were not available for this dataset.

Identification of neuroblastoma transcripts not expressed in post-natal tissues

To identify transcripts that were shared between cancer cells and the fetal medulla, we used the "quickMarkers" function in the SoupX R package to find the markers of cancer cells in each tumor dataset (CEL-seq and 10x). We removed genes that were expressed in fewer than 50% of cancer cells and genes that were expressed in more than 20% of cells in any other single cluster each dataset. We overlapped the gene lists for the two tumor datasets, and filtered the list to exclude genes that were expressed in fewer than 10% of adrenal medullary cells in our reference. We then looked at the average expression (log2(TPM)) of those genes (**table S8**) in low and high risk neuroblastoma bulk samples and GTEx (*42*) bulk samples. We sorted the gene list by increasing average expression outside the brain in GTEX data. The top 25 genes (i.e. genes with lowest expression outside the brain) are shown in **Fig. 4A**.

Association between recurrent CN changes and sympathoblastic expression

To compare regions of recurrent CN changes in neuroblastoma to the genomic pattern of expression in reference cell populations we first defined break-point regions and recurrently changed regions. We defined a recurrently changed region as one which had a gain (or a loss) in 200 or more samples (out of 556) (*18*). For regions with a well defined transition region (e.g. chromosome 11 change between recurrent gain and loss), we defined the break-point region to be 2Mb around the transition point.

For sympathoblasts, erythrocytes and chromaffin cells, we then calculated the average expression as a function of genomic position. To do this, we calculated gaussian kernel smoothed density with bandwidth of 200,000 bp and weights equal to the library size normalised expression in each cell type placed at the transcription start site of each gene. This is roughly equivalent to a running mean of normalised expression of each cell type calculated across the genome.

This normalised, smoothed expression is what is plotted in **Fig. 3F**, **fig. S8**. To calculate the significance of the relative abundance of the sympathoblastic and chromaffin we calculate the average difference in the smoothed expression between the two cell types in the regions of interest. To calculate a p-value for the significance of the observed difference in smoothed expression, we compared the observed value to the value in 1000 randomly selected regions of

the genome of the same size. We then defined the p-value to be the fraction of the randomly selected regions with a difference greater than or equal to the observed one (**fig. S5B-C**).

Differential expression of cells carrying prognostic CN change

In our 10X data, we identified one sample (GOSH021) that had loss of both chromosome 1p and 11q, which was not present in any of the other tumor samples. To determine the transcriptomic consequences of these prognostically important copy number changes we calculated the differentially expressed genes between those cells with and without the copy number changes. We limited this analysis to only those cells in the G1 phase of the cell cycle to exclude confounding with differences in proliferation. We then performed a quasi-likelihood F-test using the differential expression testing package edgeR(*34*) comparing the cells with/without the CN change. This test was performed with default parameters except for estimating the over-dispersion parameter of the negative binomial distribution where we set prior.df=0 to prevent information sharing between genes. This was done as there were sufficiently many cells in each group to robustly estimate genewise over-dispersion without sharing information.

To test if this list of differentially expressed genes related to sympathoblastic cells we identified all markers of sympathoblastic genes with a tf-idf cut-off of 1 (*28*). We then performed a hypergeometric test to measure if this list of sympathoblastic marker genes was over-represented in the genes differentially expressed between tumor cells with and without the prognostic CN changes.

Risk group stratification for SEQC data

As COG risk group information was not available for the SEQC dataset, we stratified samples into risk groups based on *MYCN* amplification status, age at diagnosis, and disease stage. We defined samples as low risk if age at diagnosis was <18 months and *MYCN* amplification status was negative, excluding 4s stage samples. We defined samples as high risk if age at diagnosis was >18 months and *MYCN* amplification status was positive, excluding 4s stage samples.

Code Availability

We have included the source code used to generate the Figures and Tables presented in this analysis as **Data S1**. The purpose of this code is to provide additional explanation of the analyses described in Methods, such as the precise function call and parameter values used.

Bibliography

1. K. K. Matthay *et al.*, Neuroblastoma. *Nat. Rev. Dis. Primers*. **2**, 16078 (2016).

2. A. Furlan, I. Adameyko, Schwann cell precursor: a neural crest cell in disguise? *Dev. Biol.* **444 Suppl 1**, S25–S35 (2018).

3. K. De Preter *et al.*, Human fetal neuroblast and neuroblastoma transcriptome analysis confirms neuroblast origin and highlights neuroblastoma candidate genes. *Genome Biol.* **7**, R84 (2006).

4. S. B. Turkel, H. H. Itabashi, The natural history of neuroblastic cells in the fetal adrenal gland. *Am. J. Pathol.* **76**, 225–244 (1974).

5. M. M. van Noesel, Neuroblastoma stage 4S: a multifocal stem-cell disease of the developing neural crest. *Lancet Oncol.* **13**, 229–230 (2012).

6. M. D. Young *et al.*, Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science*. **361**, 594–599 (2018).

7. J. E. Heck, B. Ritz, R. J. Hung, M. Hashibe, P. Boffetta, The epidemiology of neuroblastoma: a review. *Paediatr. Perinat. Epidemiol.* **23**, 125–143 (2009).

8. A. Furlan *et al.*, Multipotent peripheral glial cells generate neuroendocrine cells of the adrenal medulla. *Science*. **357** (2017), doi:10.1126/science.aal3753.

9. R. Soldatov *et al.*, Spatiotemporal structure of cell fate decisions in murine neural crest. *Science*. **364** (2019), doi:10.1126/science.aas9536.

10. W. H. Chan *et al.*, Differences in CART expression and cell cycle behavior discriminate sympathetic neuroblast from chromaffin cell lineages in mouse sympathoadrenal cells. *Dev. Neurobiol.* **76**, 137–149 (2016).

11. T. van Groningen *et al.*, A NOTCH feed-forward loop drives reprogramming from adrenergic to mesenchymal state in neuroblastoma. *Nat. Commun.* **10**, 1530 (2019).

12. T. van Groningen *et al.*, Neuroblastoma is composed of two super-enhancer-associated differentiation states. *Nat. Genet.* **49**, 1261–1266 (2017).

13. J. Mora *et al.*, Neuroblastic and Schwannian stromal cells of neuroblastoma are derived from a tumoral progenitor cell. *Cancer Res.* **61**, 6892–6898 (2001).

14. T. J. Pugh *et al.*, The genetic landscape of high-risk neuroblastoma. *Nat. Genet.* **45**, 279–284 (2013).

15. C. Wang *et al.*, The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat. Biotechnol.* **32**, 926–932 (2014).

16. E. K. Barr, M. A. Applebaum, Genetic predisposition to neuroblastoma. *Children (Basel)*. **5** (2018), doi:10.3390/children5090119.

17. T. Terzic *et al.*, Expression of Disialoganglioside (GD2) in Neuroblastic Tumors: A Prognostic Value for Patients Treated With Anti-GD2 Immunotherapy. *Pediatr. Dev. Pathol.* **21**, 1093526617723972 (2017).

18. P. Depuydt *et al.*, Meta-mining of copy number profiles of high-risk neuroblastoma tumors. *Sci. Data*. **5**, 180240 (2018).

19. Y. Li *et al.*, Patterns of somatic structural variation in human cancer genomes. *Nature*. **578**, 112–121 (2020).

20. B. Nazha, C. Inal, T. K. Owonikoko, Disialoganglioside GD2 expression in solid tumors and role as a target for cancer therapy. *Front. Oncol.* **10**, 1000 (2020).

21. A. Schramm *et al.*, Prediction of clinical outcome and biological characterization of neuroblastoma by expression profiling. *Oncogene*. **24**, 7902–7912 (2005).

22. G. Oliver *et al.*, Six3, a murine homologue of the sine oculis gene, demarcates the most anterior border of the developing neural plate and is expressed during eye development. *Development*. **121**, 4045–4055 (1995).

23. M. J. Muraro *et al.*, A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst.* **3**, 385-394.e3 (2016).

24. I. Kozarewa *et al.*, Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods*. **6**, 291–295 (2009).

25. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. **26**, 589–595 (2010).

26. P. Van Loo *et al.*, Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 16910–16915 (2010).

27. T. Candelli *et al.*, Sharq, a versatile preprocessing and QC pipeline for Single Cell RNA-seq. *BioRxiv* (2018), doi:10.1101/250811.

28. M. D. Young, S. Behjati, SoupX removes ambient RNA contamination from droplet based single cell RNA sequencing data. *BioRxiv* (2018), doi:10.1101/303727.

29. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

30. T. Stuart *et al.*, Comprehensive Integration of Single-Cell Data. *Cell*. **177**, 1888-1902.e21 (2019).

31. S. L. Wolock, R. Lopez, A. M. Klein, Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst.* **8**, 281-291.e9 (2019).

32. J. Batson, L. A. Royer, J. T. Webber, Molecular Cross-Validation for Single-Cell

RNA-seq. *BioRxiv* (2019), doi:10.1101/786269.

33. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* (2018).

34. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. **26**, 139–140 (2010).

35. C. Trapnell *et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol*. **32**, 381–386 (2014).

36. X. Qiu *et al.*, Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods*. **14**, 309–315 (2017).

37. X. Qiu *et al.*, Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods*. **14**, 979–982 (2017).

38. G. La Manno *et al.*, RNA velocity of single cells. *Nature*. **560**, 494–498 (2018).

39. J. Friedman, T. Hastie, R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw*. **33**, 1–22 (2010).

40. K. M. Raine *et al.*, ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Curr. Protoc. Bioinformatics*. **56**, 15.9.1-15.9.17 (2016), doi:10.1002/cpbi.17.

41. World Health Organisation, *International classification of diseases for oncology (ICD-O) – 3rd edition, 1st revision* (2013).

42. GTEx Consortium *et al.*, Genetic effects on gene expression across human tissues. *Nature*. **550**, 204–213 (2017).

43. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. **27**, 2987–2993 (2011).

44. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*. **57**, 289–300 (1995).

45. 1000 Genomes Project Consortium *et al.*, A global reference for human genetic variation. *Nature*. **526**, 68–74 (2015).

46. W. C. Aird, Phenotypic heterogeneity of the endothelium: II. Representative vascular beds. *Circ. Res.* **100**, 174–190 (2007).

47. R. Nawroth *et al.*, VE-PTP and VE-cadherin ectodomains interact to facilitate regulation of phosphorylation and cell contacts. *EMBO J.* **21**, 4885–4895 (2002).

48. V. G. Sankaran, S. H. Orkin, The switch from fetal to adult hemoglobin. *Cold Spring Harb. Perspect. Med.* **3**, a011643 (2013).

49. J. A. Donovan, G. A. Koretzky, CD45 and the immune response. *J. Am. Soc. Nephrol.* **4**, 976–985 (1993).

50. D. M. Stocco, StAR protein and the regulation of steroid hormone biosynthesis. *Annu. Rev. Physiol.* **63**, 193–213 (2001).

51. M. A. Wood *et al.*, Fetal adrenal capsular cells serve as progenitor cells for steroidogenic and stromal adrenocortical cell lineages in M. musculus. *Development*. **140**, 4522–4532 (2013).

52. A. Tokunaga *et al.*, PDGF receptor beta is a potent regulator of mesenchymal stromal cell function. *J. Bone Miner. Res.* **23**, 1519–1528 (2008).

53. H.-S. Kim *et al.*, Schwann Cell Precursors from Human Pluripotent Stem Cells as a Potential Therapeutic Target for Myelin Repair. *Stem Cell Reports*. **8**, 1714–1726 (2017).

54. M. E. Kastriti *et al.*, Schwann cell precursors generate the majority of chromaffin cells in zuckerkandl organ and some sympathetic neurons in paraganglia. *Front. Mol. Neurosci.* **12**, 6 (2019).

55. E. Di Zanni *et al.*, Targeting of PHOX2B expression allows the identification of drugs effective in counteracting neuroblastoma cell growth. *Oncotarget*. **8**, 72133–72146 (2017).

56. L. Longo *et al.*, PHOX2A and PHOX2B genes are highly co-expressed in human neuroblastoma. *Int. J. Oncol.* **33**, 985–991 (2008).

57. M. Schwab *et al.*, Enhanced expression of the human gene N-myc consequent to amplification of DNA may contribute to malignant progression of neuroblastoma. *Proc Natl Acad Sci USA*. **81**, 4940–4944 (1984).

58. N. E. Kohl, C. E. Gee, F. W. Alt, Activated expression of the N-myc gene in human neuroblastomas and related tumors. *Science*. **226**, 1335–1337 (1984).

Acknowledgements

General: The human embryonic and fetal material was provided by the Joint MRC / Wellcome (MR/R006237/1) Human Developmental Biology Resource (www.hdbr.org). All research at Great Ormond Street Hospital NHS Foundation Trust and UCL Great Ormond Street Institute of Child Health is made possible by the NIHR Great Ormond Street Hospital Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. We are indebted to the children and families who participated in our research. We gratefully acknowledge Liz Tuck and the Sanger Cellular Generation and Phenotyping (CGaP) Core Facility, for their assistance in tissue preparation, staining, and imaging.

Funding: This study was funded by the Wellcome Trust (references 110104/Z/15/Z, 206194 and 211276/Z/18/Z, 211276/C/18/Z, 209328/Z/17/Z). Additional funding was received from the St Baldrick's Foundation (Robert J Arceci Award to S.B.), NIHR (Great Ormond Street Biomedical Research Centre), ERC-START grant PREDICT-716079, NWO-Vidi grant 91716482, H2020-iPC-826121 grant, and National Institute for Health Research (NIHR 146281) Cambridge Biomedical Research Centre.

Author contributions: S.B., J.J.M., and K.S. conceived of the experiment. G.K., M.D.Y. and S.B. analyzed data, aided by R.E., S.R.v.H., E.K., I.d.V.T., J.C.A., C.N.P. Statistical expertise was provided by M.D.Y. K.R. and O.B. performed smFISH experiments. S.A.T. contributed fetal single cell data, together with M.H., K.B.M., R.A.B., X.H., L.M. Tumor samples were curated and/or experiments were performed by W.M.K., A.P., C.T., E.B.B., C.B., G.B., C.D., J.A., P.L., T.M., F.C.P.H., M.L.T., M.H.W.A.W., M.M.v.N., R.v.d.L, K.M.K., K.P.S.L.V.. Pathological expertise was provided by D.R., N.J.S., and R.R.d.K. G.K., M.D.Y., and S.B. wrote the manuscript. M.D.Y. directed analytical method development and statistical analyses. K.S., J.J.M., and S.B. co-directed the study.

Competing Interests: The authors declare no competing interests.

Data and materials availability: Raw sequencing data have been deposited in EGA. Bulk RNA-seq data were obtained from (*14*) and (*15*).

Figures

Figure 1

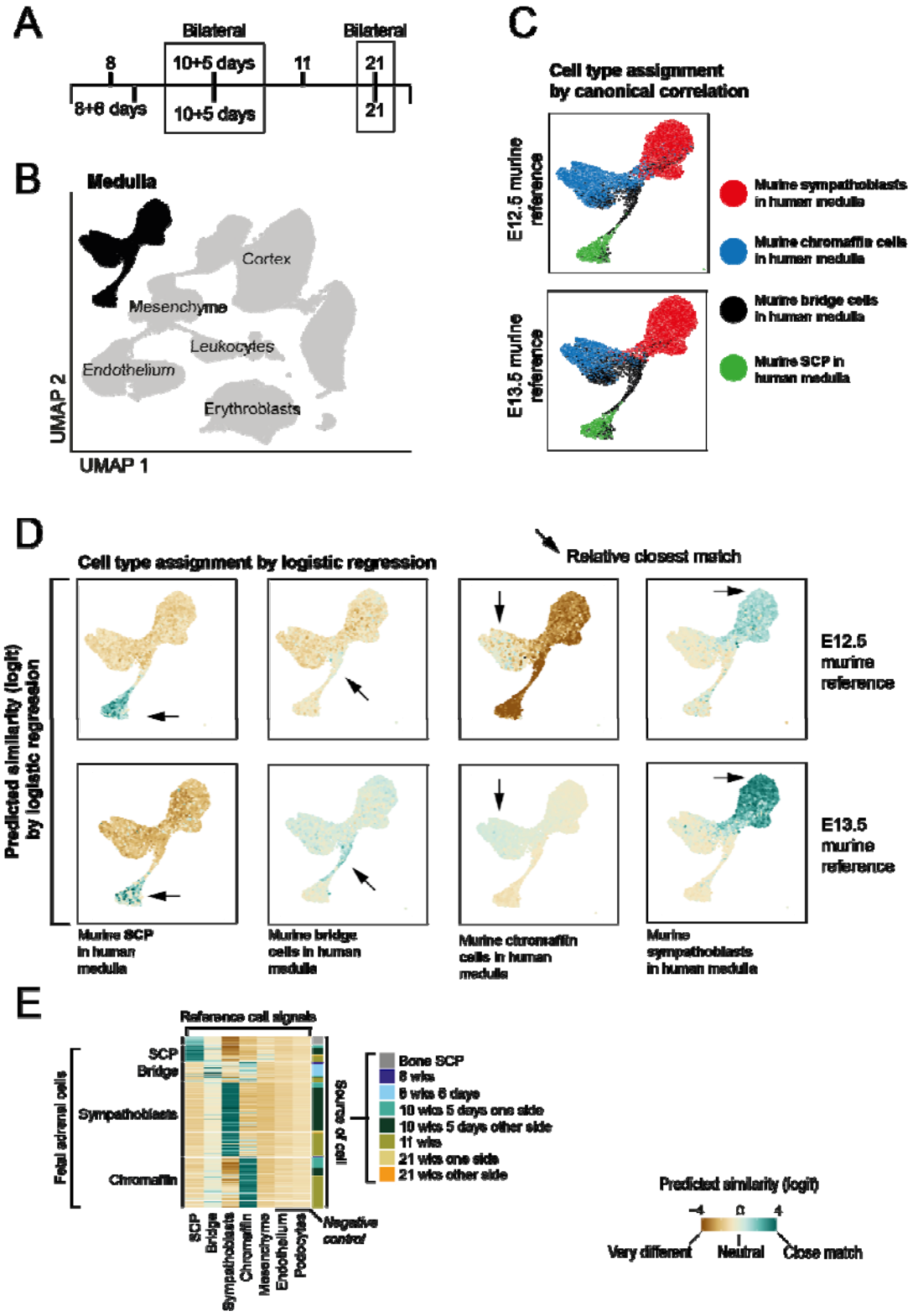
A
Bilateral
Bilateral
8
10+5 days
11
21
8+6 days
10+5 days
21
B
Medulla
Cortex
Mesenchyme
Leukocytes
Endothelium
Erythroblasts
UMAP 2
UMAP 1
C
Cell type assignment
by canonical correlation
E12.5 murine reference
E13.5 murine reference
Murine sympathoblasts in human medulla
Murine chromaffin cells in human medulla
Murine bridge cells in human medulla
Murine SCP in human medulla
D
Cell type assignment by logistic regression
Relative closest match
Predicted similarity (logit) by logistic regression
E12.5 murine reference
E13.5 murine reference
Murine SCP in human medulla
Murine bridge cells in human medulla
Murine chromaffin cells in human medulla
Murine sympathoblasts in human medulla
E
Reference cell signals
Fetal adrenal cells
SCP
Bridge
Sympathoblasts
Chromaffin
Source of cell
Bone SCP
8 wks
8 wks 6 days
10 wks 5 days one side
10 wks 5 days other side
11 wks
21 wks one side
21 wks other side
SCP
Bridge
Sympathoblasts
Chromaffin
Mesenchyme
Endothelium
Podocytes
Negative control
Predicted similarity (logit)
-4
0
4
Very different
Neutral
Close match

Fig. 1. Single cell reference data from seven human fetal adrenal glands.

(A) Gestational age of fetal adrenal glands.

(B) UMAP (Uniform Manifold Approximation and Projection) representation of 57,972 fetal adrenal gland cells. Medullary cells (n=6,451) are highlighted in black.

(C) Cell type assignment by canonical correlation analysis as implemented in the TransferData function in Seurat. Murine medullary cell types, as defined by Furlan *et al. (8)*, from two developmental ages (E12.5 and E13.5) were transferred to human medullary cells (black cell population in (A)) to identify human correlates (as per colour legend).

(D) Cell type assignment by logistic regression comparing murine medullary cell types, as defined by Furlan *et al. (8)* from two developmental ages (E12.5 and E13.5) with our human data (black cell population in A). Colours represent the similarity score (logit scale): 0 default, positive (green) indicating strong similarity and negative (brown) strong dissimilarity.

(E) Comparison of medullary cells from different gestational ages. Similarity scores are shown by color, comparing the reference week 10, day 5 adrenal gland and extrinsic control (x-axis) to medullary cell types across all timepoints (y-axis). The annotation track to the right of the heatmap shows the sample each cell belongs to.

Figure 2

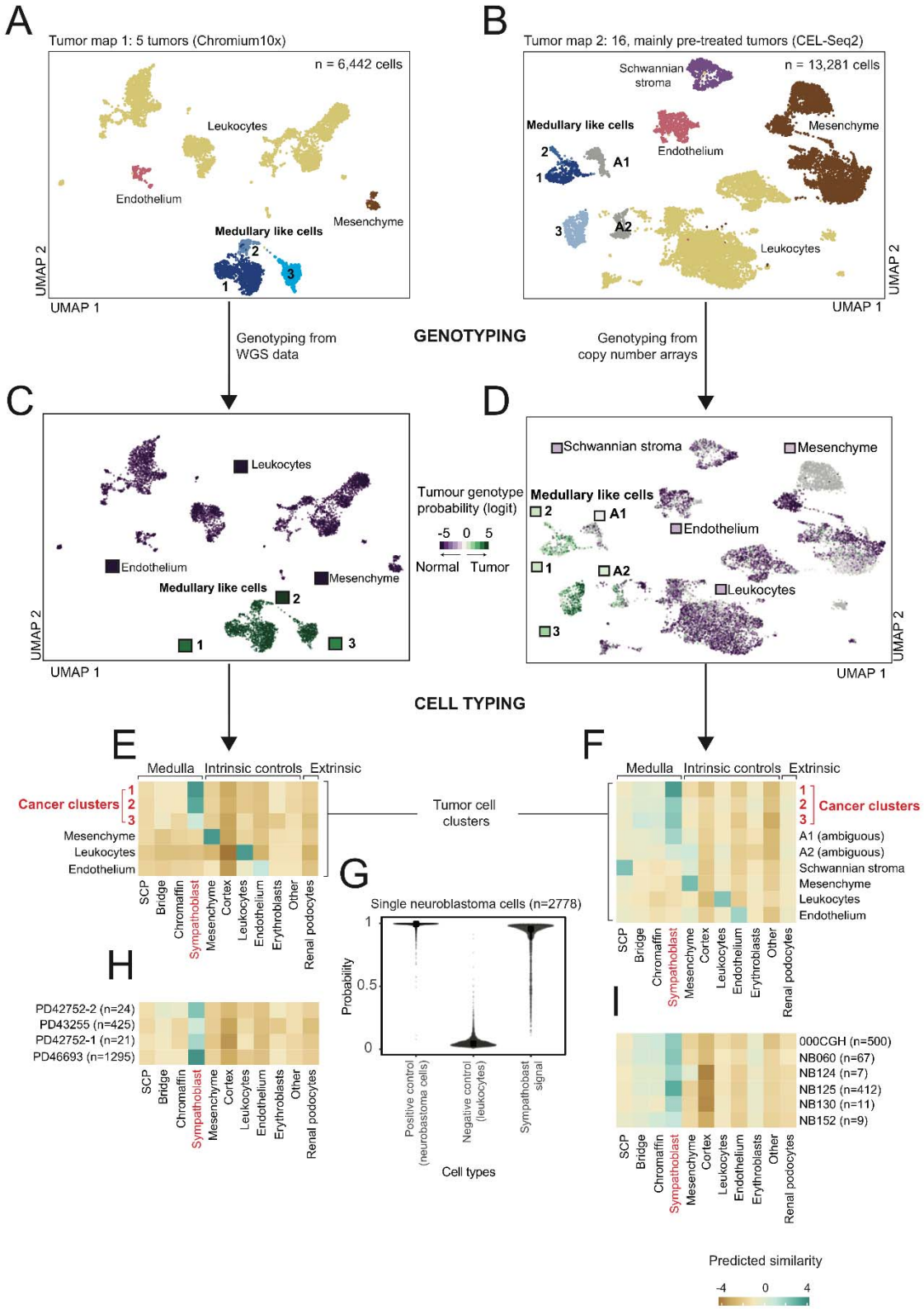
A
Tumor map 1: 5 tumors (Chromium10x)
n = 6,442 cells
Leukocytes
Endothelium
Mesenchyme
Medullary like cells
UMAP 2
UMAP 1
B
Tumor map 2: 16, mainly pre-treated tumors (CEL-Seq2)
n = 13,281 cells
Schwannian stroma
Medullary like cells
Endothelium
Mesenchyme
A1
A2
Leukocytes
GENOTYPING
Genotyping from WGS data
Genotyping from copy number arrays
C
D
Tumour genotype probability (logit)
-5 0 5
Normal Tumor
CELL TYPING
E
F
Medulla
Intrinsic controls
Extrinsic
Cancer clusters
Tumor cell clusters
A1 (ambiguous)
A2 (ambiguous)
Schwannian stroma
SCP
Bridge
Chromaffin
Sympatholast
Cortex
Erythroblasts
Other
Renal podocytes
G
Single neuroblastoma cells (n=2778)
Probability
Positive control (neuroblastoma cells)
Negative control (leukocytes)
Sympathobast signal
Cell types
H
PD42752-2 (n=24)
PD43255 (n=425)
PD42752-1 (n=21)
PD46693 (n=1295)
I
000CGH (n=500)
NB060 (n=67)
NB124 (n=7)
NB125 (n=412)
NB130 (n=11)
NB152 (n=9)
Predicted similarity
-4 0 4

Fig. 2. Neuroblastoma single cell transcriptomes.

(A,B) UMAP representation of 6,442 **(A)** and 13,281 **(B)** neuroblastoma cells. Colours and labels indicate different cell types defined using marker gene expression.

(C,D) The same UMAP representation as (A,B) but here colour represents posterior probability of the cancer genotype calculated for each cell. The square next to each cluster title shows the average posterior probability for that cluster. Scores are given on a logit scale, with 0 indicating no information, positive (green) values evidence for the tumor genotype, and negative (purple) evidence for the normal genotype.

(E,F) Similarity scores (logistic regression, logit scale) of the reference fetal adrenal gland (x-axis) to neuroblastoma cells (y-axis), containing intrinsic (non-cancer cells) and an extrinsic (kidney podocytes) control cell population.

(G) tumor to normal similarity score, represented as posterior probability, comparing neuroblastoma cells (positive control), leukocytes (negative control), and sympathoblasts to neuroblastoma cells. The similarity of tumor cells to sympathoblasts is comparable to the similarity of tumor cells to themselves.

(H,I) Similarity of cells from the cancer clusters in (E,F) grouped by patient. Numbers in parentheses indicate the total number of cells in cancer clusters in (E,F). Only patients with at least 10 tumor cells, or at least one tumor cell with validated genotype were included.

Figure 3

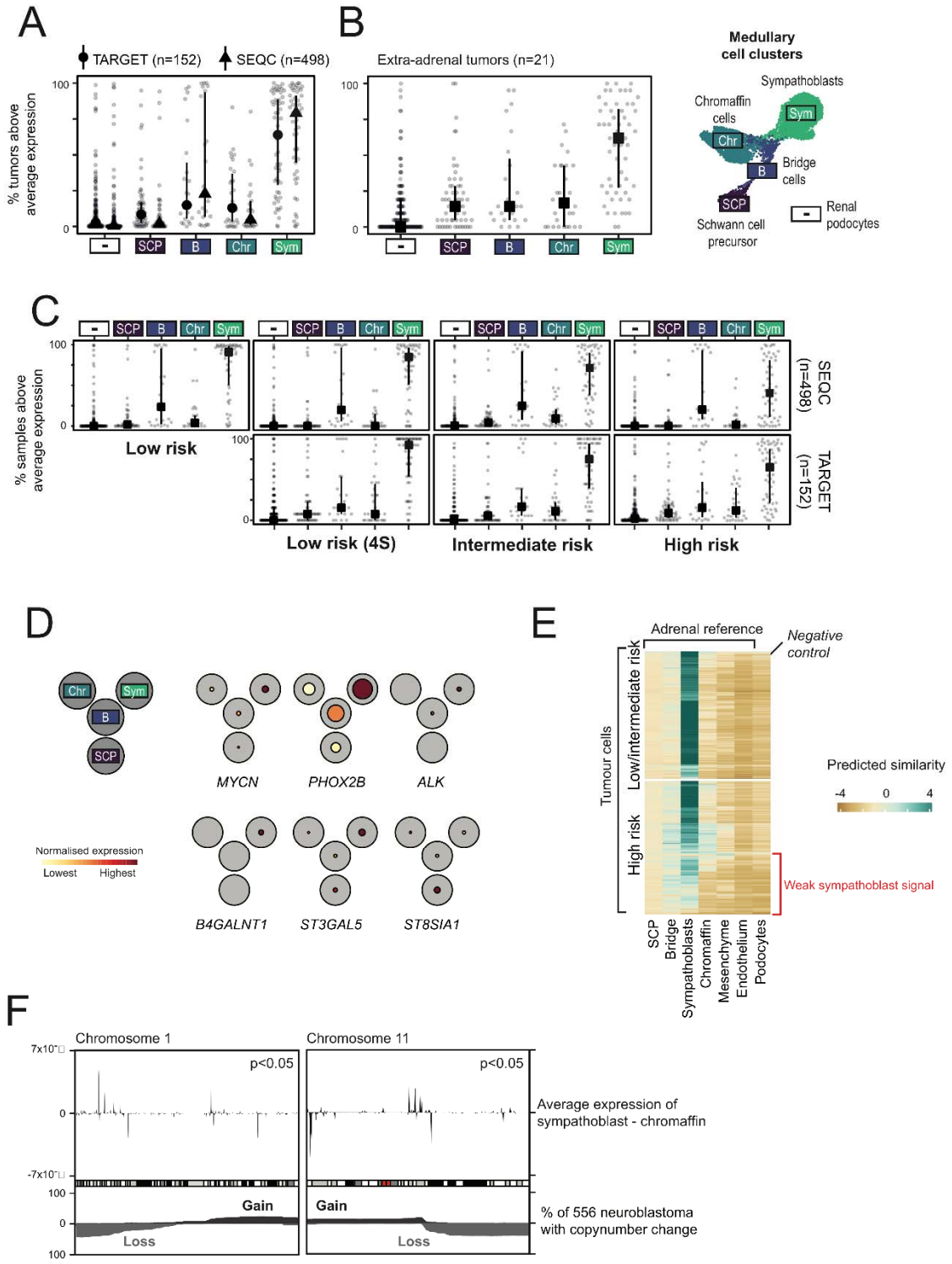
A
TARGET (n=152)
SEQC (n=498)
% tumors above average expression
100
0
SCP
B
Chr
Sym
B
Extra-adrenal tumors (n=21)
Medullary cell clusters
Sympathoblasts
Chromaffin cells
Bridge cells
Schwann cell precursor
Renal podocytes
C
% samples above average expression
SEQC (n=498)
TARGET (n=152)
Low risk
Low risk (4S)
Intermediate risk
High risk
D
Normalised expression
Lowest
Highest
MYCN
PHOX2B
ALK
B4GALNT1
ST3GAL5
ST8SIA1
E
Adrenal reference
Negative control
Tumour cells
Low/intermediate risk
High risk
Predicted similarity
-4
0
4
Weak sympathoblast signal
SCP
Bridge
Sympathoblasts
Chromaffin
Mesenchyme
Endothelium
Podocytes
F
Chromosome 1
Chromosome 11
p<0.05
Average expression of sympathoblast - chromaffin
Gain
Loss
% of 556 neuroblastoma with copynumber change

Fig. 3. Validation in clinically annotated neuroblastoma cohorts.

(A) Fraction of bulk neuroblastoma samples in TARGET (Therapeutically Applicable Research To Generate Effective Treatments) cohort and SEQC (Sequencing Quality Control) with above average marker expression. Dataset/cell-type indicate by points/x-axis labels (see UMAP on right), lines indicate interquartile ranges, symbols the median.

(B) As in (A) but limited to tumors sampled outside the adrenal gland.

(C) As in (A) but with samples split by clinical risk group (horizontal facets) and datasets (vertical facets).

(D) Fetal medulla expression of GD2 synthesis genes, and principal somatic, and germline predisposition genes. Black circles represent cell populations in (B), with average expression (normalised for each gene)/fraction of expressing cells indicated by color/colored circle radius. Further genes shown in **fig. S5**.

(E) Similarity scores (logistic regression, logit scale) comparing fetal adrenal (x-axis) to neuroblastoma cells (y-axis) and extrinsic control (kidney podocytes). In high risk tumor cells, a subset of cells exhibit a weakened sympathoblast signal, indicated by the red label.

(F) Average expression difference between sympathoblastic and chromaffin cells versus of genomic position on chromosomes 1 and 11 (top panel). Positive (negative) values indicate higher expression in sympathoblasts (chromaffin) cells. The bottom panel shows fraction of 556 neuroblastomas with copy number gain (black) or loss (grey).

Figure 4

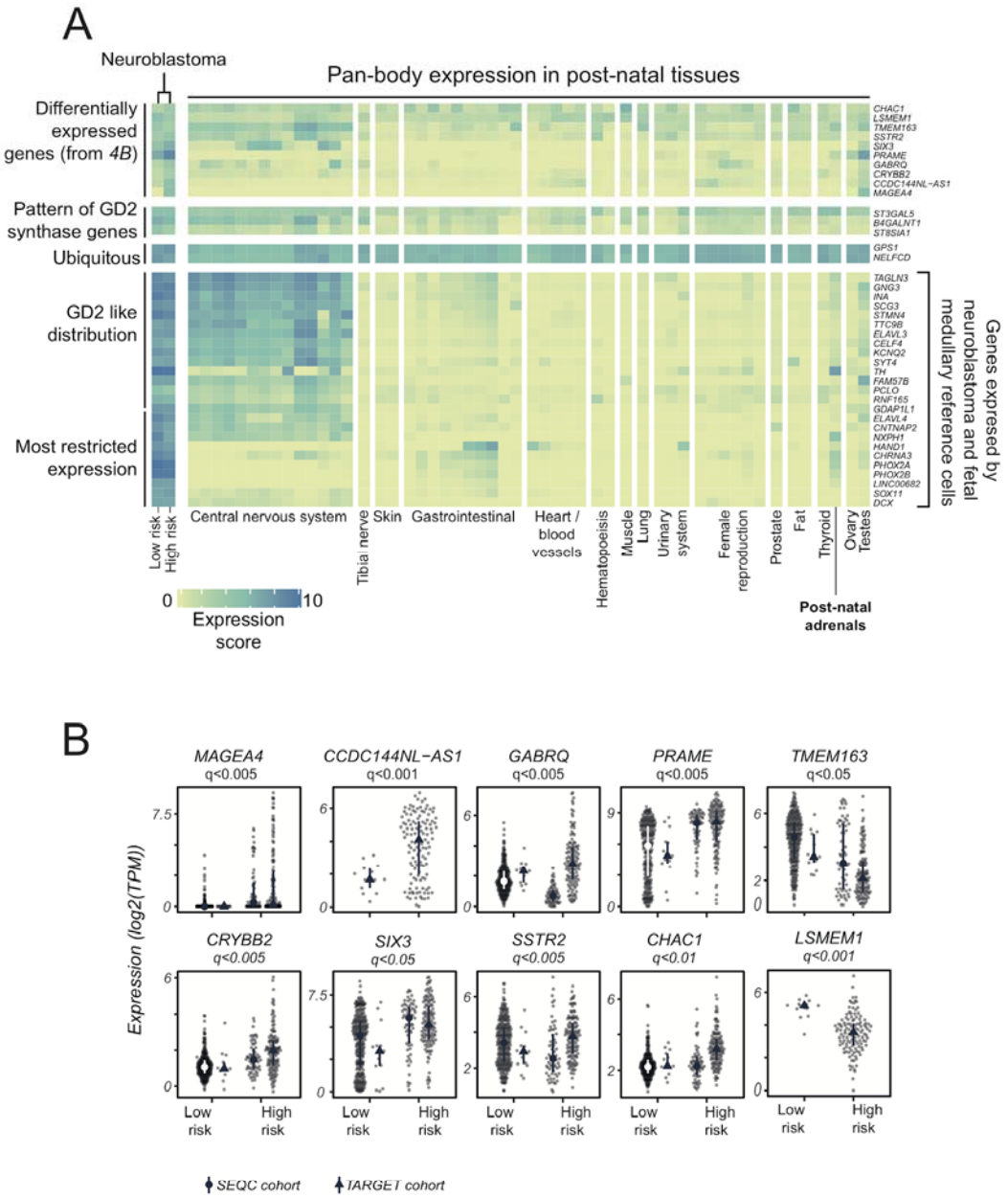
A
Neuroblastoma
Pan-body expression in post-natal tissues
Differentially expressed genes (from 4B)
Pattern of GD2 synthase genes
Ubiquitous
GD2 like distribution
Most restricted expression
CHAC1
LSMEM1
TMEM163
SSTR2
SIX3
PRAME
GABRQ
CRYBB2
CCDC144NL-AS1
MAGEA4
ST3GAL5
B4GALNT1
ST8SIA1
GPS1
NELFCD
TAGLN3
GNG3
INA
SCG3
STMN4
TTC9B
ELAVL3
CELF4
KCNQ2
SYT4
TH
FAM57B
PCLO
RNF165
GDAP1L1
ELAVL4
CNTNAP2
NXPH1
HAND1
CHRNA3
PHOX2A
PHOX2B
LINC00682
SOX11
DCX
Genes expresed by neuroblastoma and fetal medullary reference cells
Low risk
High risk
Central nervous system
Tibia nerve
Skin
Gastrointestinal
Heart / blood vessels
Hematopoeisis
Muscle
Lung
Urinary system
Female reproduction
Prostate
Fat
Thyroid
Ovary
Testes
Post-natal adrenals
0
10
Expression score
B
MAGEA4
q<0.005
CCDC144NL-AS1
q<0.001
GABRQ
q<0.005
PRAME
q<0.005
TMEM163
q<0.05
CRYBB2
q<0.005
SIX3
q<0.05
SSTR2
q<0.005
CHAC1
q<0.01
LSMEM1
q<0.001
Expression (log2(TPM))
Low risk
High risk
SEQC cohort
TARGET cohort

Fig. 4. Fetal neuroblastoma genes and prognostically significant neuroblastoma mRNAs

(A) A heatmap showing the expression of GD2 synthesis genes, genes from (B), and genes utilized by both neuroblastoma cells and the fetal medulla (y-axis) in bulk transcriptomes of normal tissues (GTEX) and high/low risk neuroblastomas from TARGET. The color indicates log2 of the average gene expression (TPM) for the corresponding gene and sample set.

(B) Genes differentially expressed between neuroblastoma tumor cells and foetal medulla that exhibit significant risk group dependence after controlling for MYCN status and age (negative-binomial test, $\log_2 FC>1$, $FDR<0.05$) in TARGET dataset. For the gene listed above each panel, dots indicate expression ($\log_2(TPM)$) in individual samples, lines the interquartile range, and symbols the median for SEQC and TARGET datasets. Two genes (*LSMEM1*, *CCDC144NL-AS1*), are not present in annotation used by SEQC.