nature cancer

Resource

https://doi.org/10.1038/s43018-025-00957-8

The Curated Cancer Cell Atlas provides a comprehensive characterization of tumors at single-cell resolution

Received: 19 January 2024

Accepted: 24 March 2025

Published online: 08 May 2025

Check for updates

Michael Tyler[1,2] ✉, Avishai Gavish[1], Chaya Barbolin[1], Roi Tschernichovsky[1,3], Rouven Hoefflin[1,4], Michael Mints[1,5], Sidharth V. Puram[6,7] & Itay Tirosh[1] ✉

Recent years have seen a rapid proliferation of single-cell cancer studies, yet most of these studies profiled few tumors, limiting their statistical power. Combining data and results across studies holds great promise but also involves various challenges. We recently began to address these challenges by curating a large collection of cancer single-cell RNA-sequencing datasets, leveraging it for systematic analyses of tumor heterogeneity. Here we greatly extend this repository to 124 datasets for over 40 cancer types, together comprising 2,836 samples, with improved data annotations, visualizations and exploration. Using this vast cohort, we generate an updated map of recurrent expression programs in malignant cells and systematically quantify context-dependent gene expression and cell-cycle patterns across cell types and cancer types. These data, annotations and analysis results are all freely available for exploration and download through the Curated Cancer Cell Atlas, a central community resource that opens new avenues in cancer research.

A tumor is a complex ecosystem of different cell types, genetic clones and dynamic cellular states. This intratumor heterogeneity (ITH) is central to tumor development and poses a major barrier to cancer therapy, with resistant tumor subpopulations driving continued disease progression[1]. Single-cell RNA sequencing (scRNA-seq) has recently emerged as a powerful tool to study ITH, paving the way to a more complete understanding of cancer progression and treatment effects. Early studies applying scRNA-seq to tumor samples identified, for example, a neuronal progenitor cell (NPC) state in oligodendroglioma[2], a partial epithelial–mesenchymal transition (EMT) state in head and neck cancer[3] and an antigen-presenting population of cancer-associated fibroblasts in pancreatic cancer[4]. Such discoveries were possible only with the high resolution and whole-genome coverage of scRNA-seq.

The generation of tumor scRNA-seq data has accelerated dramatically, with hundreds of recent publications. Collectively, the global cancer research community has generated a vast pool of high-resolution tumor transcriptomic profiles[5,6] that have the potential to transform our understanding of cancer and drive the development of new treatment strategies. These datasets could ultimately define a foundational resource, replacing widely used bulk cohorts, such as The Cancer Genome Atlas (TCGA), with single-cell compendia that will be used routinely in cancer studies. However, because of cost and various technical constraints, individual scRNA-seq studies are only able to profile relatively few tumor samples, typically 5–20. Each dataset is, therefore, severely underpowered to identify robust and clinically important expression patterns. At the same time, the ability to compare

[1]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [2]Georg-Speyer-Haus, Institute for Tumor Biology and Experimental Therapy, Frankfurt, Germany. [3]Davidoff Cancer Center, Rabin Medical Center, Petah Tikva, Israel. [4]Department of Medicine I, Medical Center—University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany. [5]Department of Oncology–Pathology, Karolinska Institute, Stockholm, Sweden. [6]Department of Otolaryngology—Head and Neck Surgery and Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA. [7]The Robert Ebert and Greg Stubblefield Head and Neck Tumor Center at Siteman Cancer Center, St. Louis, MO, USA. ✉e-mail: m20tyler@gmail.com; itay.tirosh@weizmann.ac.il

data between studies is hindered by batch effects and inconsistencies in methods, formats and annotations.

We address these problems by curating a large number of published scRNA-seq datasets for combined analysis. We previously published a repository of 71 such datasets, in a pan-cancer scRNA-seq study characterizing recurrent programs of transcriptional ITH[7]. We now considerably expand this cohort, almost doubling its size to 124 datasets, 2,836 samples and over 5.6 million single cells, enabling an even deeper exploration of ITH. We use this extended compendium to systematically identify context-dependent gene expression profiles, characterize cell type markers and identify genes that distinguish malignant cells in various contexts. Furthermore, we present a comprehensive quantification of cell cycle patterns, revealing high variability in proliferation rates across cell types and cancer types and uncovering biases in cell-cycle phases that are associated with driver mutations, most notably *TP53*. These data, analyses and tools to explore them together constitute the Curated Cancer Cell Atlas (3CA), an updated resource that is available to the entire cancer research community through a website (https://www.weizmann.ac.il/sites/3CA/) and enables comprehensive characterization of tumors at single-cell resolution.

Results

Curating a comprehensive scRNA-seq data resource

To build 3CA, we conducted a thorough literature search to identify scRNA-seq cancer studies representing a wide range of cancer types, prioritizing those having a relatively high number of samples, as well as smaller datasets for understudied cancer types. While the majority of these datasets were generated from patient samples, they also include cancer cell lines, organoids and mouse models. Following this comprehensive search, we downloaded these datasets from their respective repositories, standardized their format and verified their cell annotations (Fig. 1a, top). To ensure flexibility and preserve biological variability, we retain the expression data in as raw a format as possible (unique molecular identifier counts or transcripts per million (TPM)) and do not apply batch correction or integration methods. An initial version of 3CA consisting of datasets from 71 studies was reported previously[7] and the updated version presented here consists of 124 datasets for over 40 cancer types, together comprising 2,836 samples and 5,658,705 cells. This new version greatly improves the statistical power of 3CA, more than doubling the number of cells of each of the most common cell types, with the highest proportional gain occurring in nonmalignant cell types (Fig. 1b). It also better captures normal samples (from both tumor-adjacent tissue and healthy individuals) and samples away from the primary tumor location, including blood and distant metastases (Fig. 1c). Lastly, the expansion includes new cancer types such as skin basal cell carcinoma and diffuse large B cell lymphoma, alongside substantially increased representation of major cancer types such as hepatocellular carcinoma and head and neck squamous cell carcinoma (Fig. 1d).

Ensuring consistent annotations is especially crucial to data curation and enables combined analysis of this large cohort. Where possible, we obtained cell type annotations from the original studies; for 12 studies, we defined cell types ourselves de novo and, for a further 9, we obtained annotations from TISCH2 (ref. 8). We then standardized the cell type annotations and validated them in two ways. Firstly, we inferred copy-number alterations (CNAs), using a method described previously[2,9] (Methods), to confirm the annotation of malignant cells. Some studies presented additional evidence for cell malignancy, such as whole-exome sequencing, which reduced our reliance on CNA inference. Secondly, we verified the identities of nonmalignant cell types by analyzing the expression of canonical marker genes (Supplementary Data 1). In a minority of cases, we amended the annotations to resolve discrepancies evident from these validation measures. Along with these cell annotations, for the new 3CA version, we invested great effort in curating additional clinical annotations, where available, including patient age and sex, tumor stage and disease extent, treatment history and response and status of relevant driver mutations.

An online portal for open access to 3CA

With its unparalleled size and rich annotations, 3CA is a valuable data resource for the entire cancer research community. To make it accessible to all researchers, we built and recently extended a website (https://www.weizmann.ac.il/sites/3CA/) through which the curated datasets are freely available to download, without the need for user registration or permissions. The home page summarizes all datasets, organizing them into 15 categories of cancer types (Fig. 1e). Separate category-specific pages contain links to download the expression matrices and/or cell and sample metadata (Fig. 1f).

We further sought to enrich 3CA with detailed data visualizations, functionality to explore the datasets and new pan-cancer analyses (Fig. 1a, bottom). The category-specific pages of the website contain summary statistics for each dataset, including disease name, sequencing technology and cell and sample numbers, along with various visualizations for each dataset, including cell type composition, expression of canonical cell type marker genes, CNA matrices and uniform manifold approximation and projection (UMAP) plots colored by cell type and sample (Fig. 1g,h and Extended Data Fig. 1a–c). Additional website features are described below. Through this online portal, 3CA serves as a central source of data and analyses for all cancer researchers, which will continue to expand as the tumor scRNA-seq literature grows further.

Exploring transcriptional tumor heterogeneity with 3CA

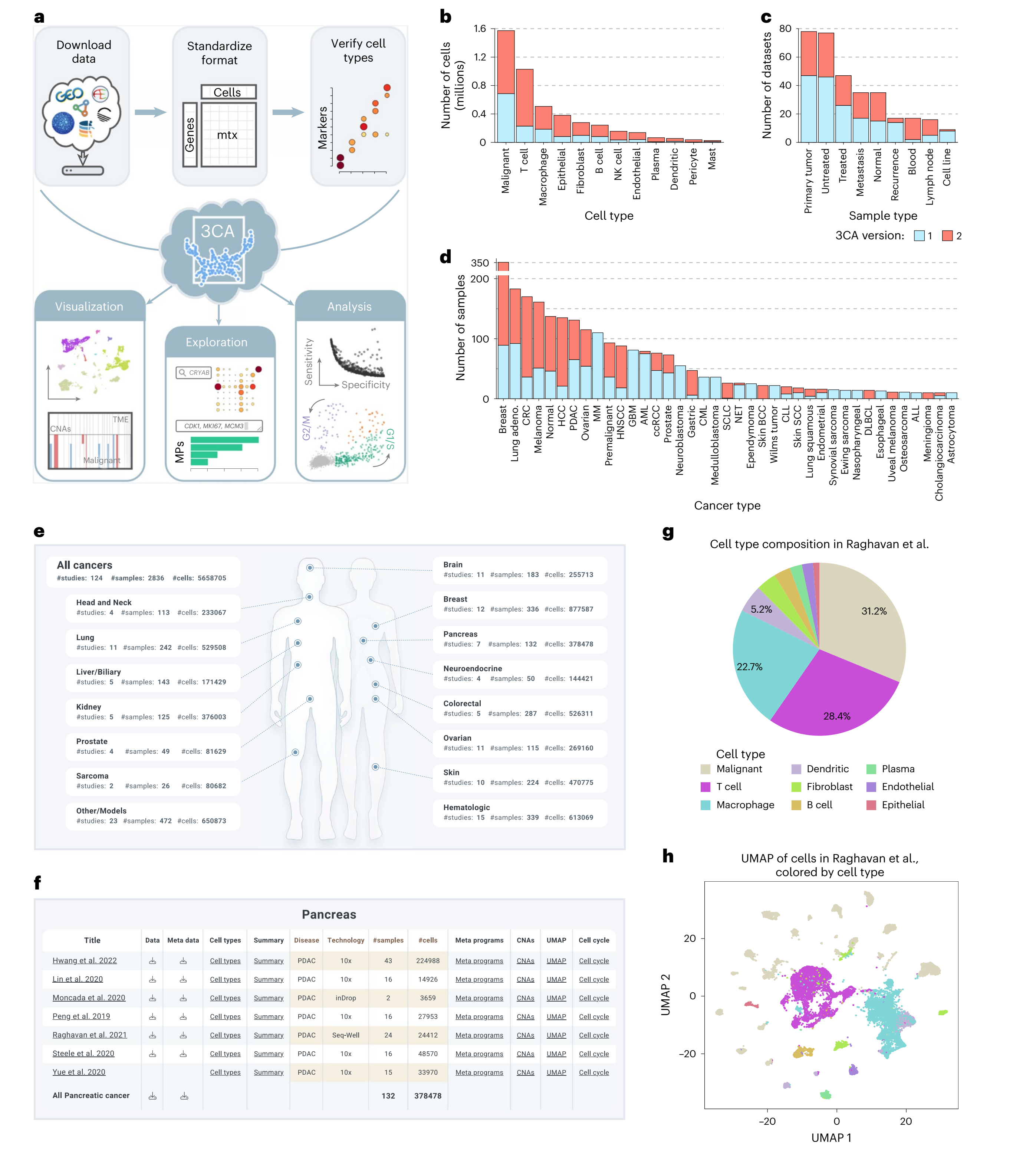
To enhance the functionality of the website beyond visualization of the cell and sample composition of individual 3CA datasets, for the latest version, we included new features that enable exploration of ITH across datasets (Fig. 1a, bottom). We previously used 71 3CA datasets in a pan-cancer characterization of ITH[7]. In particular, we identified recurrent programs of transcriptional ITH, which we term 'metaprograms' (MPs). We defined a total of 149 MPs across eight cell types, which collectively explain the majority of expression ITH. Thus, a given tumor sample may be described by quantifying the extent to which these MPs are variable across the cells in the tumor. We, therefore, added to 3CA a summary, for each dataset, of the expression of MPs across cells of each type in each sample (Fig. 2a,b) and a feature for users to enter their own gene sets and view their overlap with these MPs (Fig. 2c).

We also added functionality to address a common question in cancer research: Given a gene of interest, what is its typical expression in each cell type and cancer type? Because of the resolution and comprehensiveness of 3CA, it offers the possibility to query the expression of any gene across many contexts. We, thus, added to the 3CA website a search tool that returns, for any gene, a visualization of its average expression and the proportion of cells expressing it, per cell type and cancer type (Fig. 2d). This is further broken down per dataset to enable the examination of study-specific effects. This tool also returns plots showing the correlation of the query gene with the different MPs per cell type and cancer type (Fig. 2e).

An updated map of recurrent ITH programs

While the MPs we defined previously[7] explain the majority of expression variability in tumors, the larger data collection in the new 3CA version could enable the detection of new MPs that capture even more ITH. We, therefore, regenerated MPs for malignant cells (Methods). We defined 67 recurrent expression programs, most of which directly corresponded to MPs that we defined previously; indeed, all 41 of the earlier malignant MPs were captured in the updated list (Fig. 3a,b and Supplementary Data 2 and 3). Many other MPs reflected variations of the earlier MPs, including, for example, two new variations of EMT and one new program related to interferon response and major histocompatibility complex (MHC) class II. Interestingly, some MPs were composed

https://doi.org/10.1038/s43018-025-00957-8



All cancers
#studies: 124 #samples: 2836 #cells: 5658705

Head and Neck
#studies: 4 #samples: 113 #cells: 233067

Lung
#studies: 11 #samples: 242 #cells: 529508

Liver/Biliary
#studies: 5 #samples: 143 #cells: 171429

Kidney
#studies: 5 #samples: 125 #cells: 376003

Prostate
#studies: 4 #samples: 49 #cells: 81629

Sarcoma
#studies: 2 #samples: 26 #cells: 80682

Other/Models
#studies: 23 #samples: 472 #cells: 650873

Brain
#studies: 11 #samples: 183 #cells: 255713

Breast
#studies: 12 #samples: 336 #cells: 877587

Pancreas
#studies: 7 #samples: 132 #cells: 378478

Neuroendocrine
#studies: 4 #samples: 50 #cells: 144421

Colorectal
#studies: 5 #samples: 287 #cells: 526311

Ovarian
#studies: 11 #samples: 115 #cells: 269160

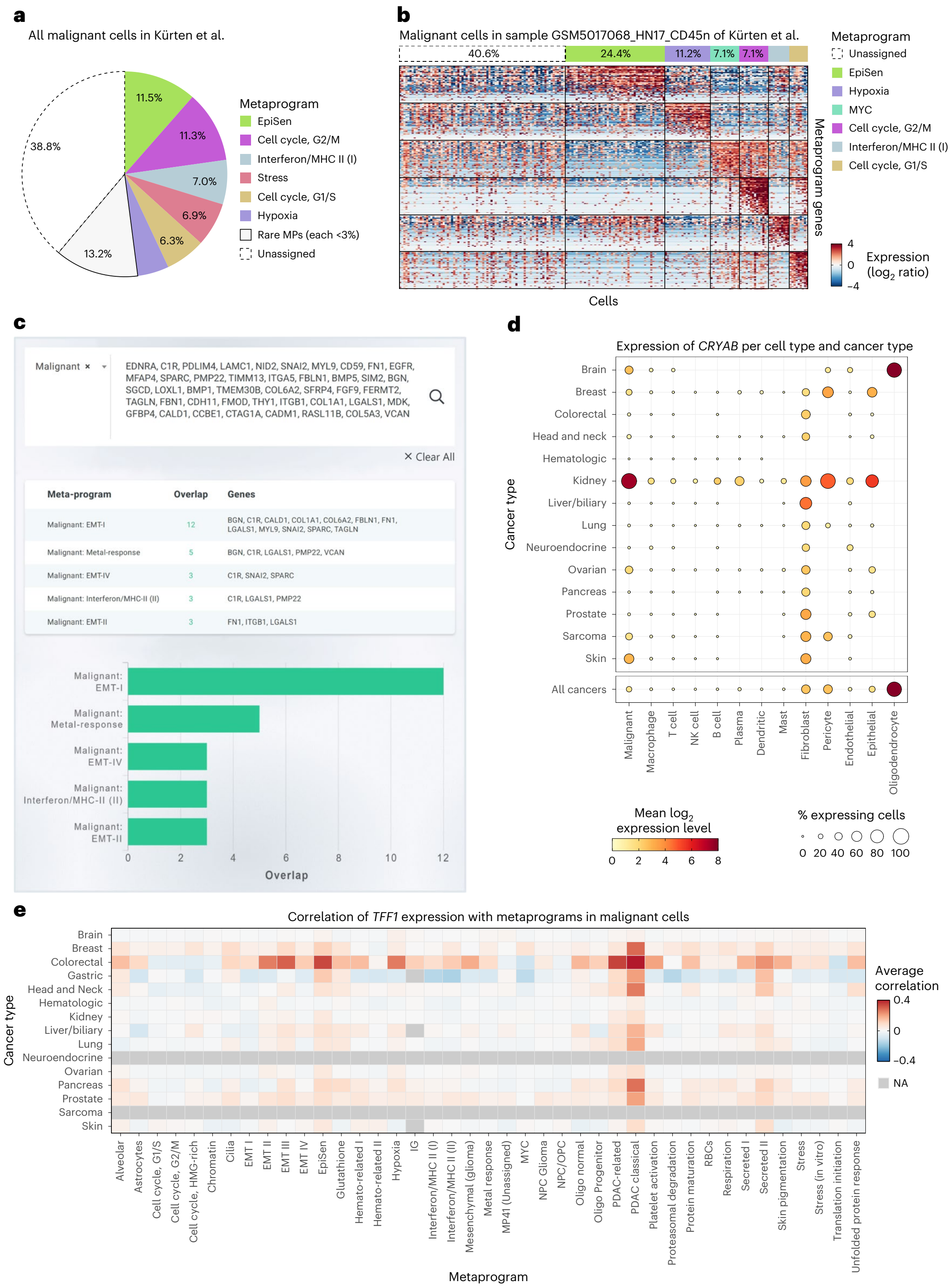
Skin
#studies: 10 #samples: 224 #cells: 470775

Hematologic
#studies: 15 #samples: 339 #cells: 613069

Pancreas

Title	Data	Meta data	Cell types	Summary	Disease	Technology	#samples	#cells	Meta programs	CNAs	UMAP	Cell cycle
Hwang et al. 2022	⤓	⤓	Cell types	Summary	PDAC	10x	43	224988	Meta programs	CNAs	UMAP	Cell cycle
Lin et al. 2020	⤓	⤓	Cell types	Summary	PDAC	10x	16	14926	Meta programs	CNAs	UMAP	Cell cycle
Moncada et al. 2020	⤓	⤓	Cell types	Summary	PDAC	inDrop	2	3659	Meta programs	CNAs	UMAP	Cell cycle
Peng et al. 2019	⤓	⤓	Cell types	Summary	PDAC	10x	16	27953	Meta programs	CNAs	UMAP	Cell cycle
Raghavan et al. 2021	⤓	⤓	Cell types	Summary	PDAC	Seq-Well	24	24412	Meta programs	CNAs	UMAP	Cell cycle
Steele et al. 2020	⤓	⤓	Cell types	Summary	PDAC	10x	16	48570	Meta programs	CNAs	UMAP	Cell cycle
Yue et al. 2020			Cell types	Summary	PDAC	10x	15	33970	Meta programs	CNAs	UMAP	Cell cycle
All Pancreatic cancer	⤓	⤓					**132**	**378478**				

Fig. 1 | Overview of 3CA and data summary features. a, Scheme outlining the construction of 3CA and the features it contains. 'Visualization' features were included in the initial version of the 3CA website[7] but all 'exploration' and 'analysis' features are new to this extended version. **b**, Stacked bar chart showing, for a selection of the most common cell types, the number of cells per cell type in each 3CA version. **c**, Stacked bar chart showing the number of datasets containing different sample types in each 3CA version. **d**, Stacked bar chart showing the number of samples per cancer type in each 3CA version. Cancer types with fewer than ten samples in total are not shown. HCC, hepatocellular carcinoma; PDAC, pancreatic ductal adenocarcinoma; MM, multiple myeloma; HNSCC, head and neck squamous cell carcinoma; GBM, glioblastoma multiforme; AML, acute myeloid leukemia; ccRCC, clear cell renal cell carcinoma; CML, chronic myeloid leukemia; SCLC, small cell lung cancer; NET, neuroendocrine tumor; BCC, basal cell carcinoma; CLL, chronic lymphocytic leukemia; DLBCL, diffuse large B cell lymphoma; ALL, acute lymphocytic leukemia. **e**, Screenshot of the 3CA website home page, summarizing the available datasets. **f**, Screenshot of the 3CA web page for pancreas, showing the available pancreatic cancer datasets with links and summary statistics. **g**, Pie chart showing the percentage of cells in the Raghavan et al. dataset[48] assigned to each cell type. **h**, UMAP plot of all cells (points) in the Raghavan et al. dataset[48], colored by cell type.



Malignant ×

EDNRA, C1R, PDLIM4, LAMC1, NID2, SNAI2, MYL9, CD59, FN1, EGFR, MFAP4, SPARC, PMP22, TIMM13, ITGA5, FBLN1, BMP5, SIM2, BGN, SGCD, LOXL1, BMP1, TMEM30B, COL6A2, SFRP4, FGF9, FERMT2, TAGLN, FBN1, CDH11, FMOD, THY1, ITGB1, COL1A1, LGALS1, MDK, GFBP4, CALD1, CCBE1, CTAG1A, CADM1, RASL11B, COL5A3, VCAN

× Clear All

Meta-program	Overlap	Genes
Malignant: EMT-I	12	BGN, C1R, CALD1, COL1A1, COL6A2, FBLN1, FN1, LGALS1, MYL9, SNAI2, SPARC, TAGLN
Malignant: Metal-response	5	BGN, C1R, LGALS1, PMP22, VCAN
Malignant: EMT-IV	3	C1R, SNAI2, SPARC
Malignant: Interferon/MHC-II (II)	3	C1R, LGALS1, PMP22
Malignant: EMT-II	3	FN1, ITGB1, LGALS1

Fig. 2 | Data query features on the 3CA website. a, Pie chart showing the MP composition of malignant cells in the Kürten et al. dataset[49]. Each cell was assigned to at most one MP, with ambiguous cells classified as unassigned. Percentages denote the proportion of cells in each category. **b**, Heat map showing relative expression levels (color, quantified as $\log_2$ ratio) of MP genes (rows) in malignant cells (columns) in sample GSM5017068_HN17_CD45n of the Kürten et al. dataset[49]. Each cell was assigned to at most one MP, with ambiguous cells classified as unassigned. Top: the color bar shows the classification of each cell and percentages denote the proportion of cells in each category. **c**, Example output from the MP gene set query tool. The table shows, for each malignant cell MP having at least two genes in common with the input gene set, the size of its overlap with the input gene set and the genes residing in this overlap. The bar plot shows the size of the overlap (*x* axis) of each MP from the table (*y* axis) with the input gene set. **d**, Dot plot showing the average $\log_2$ expression level (color) and percentage expressing cells (point size) of *CRYAB* in each cell type (columns) and cancer type (rows; top) and averaged across cancer types (bottom). **e**, Heat map showing the average correlation (color) between the expression of *TFF1* and scores for MPs (*x* axis) in malignant cells in each cancer type (*y* axis). Gray indicates insufficient data (NA, not available). OPC, oligodendrocyte progenitor cell.

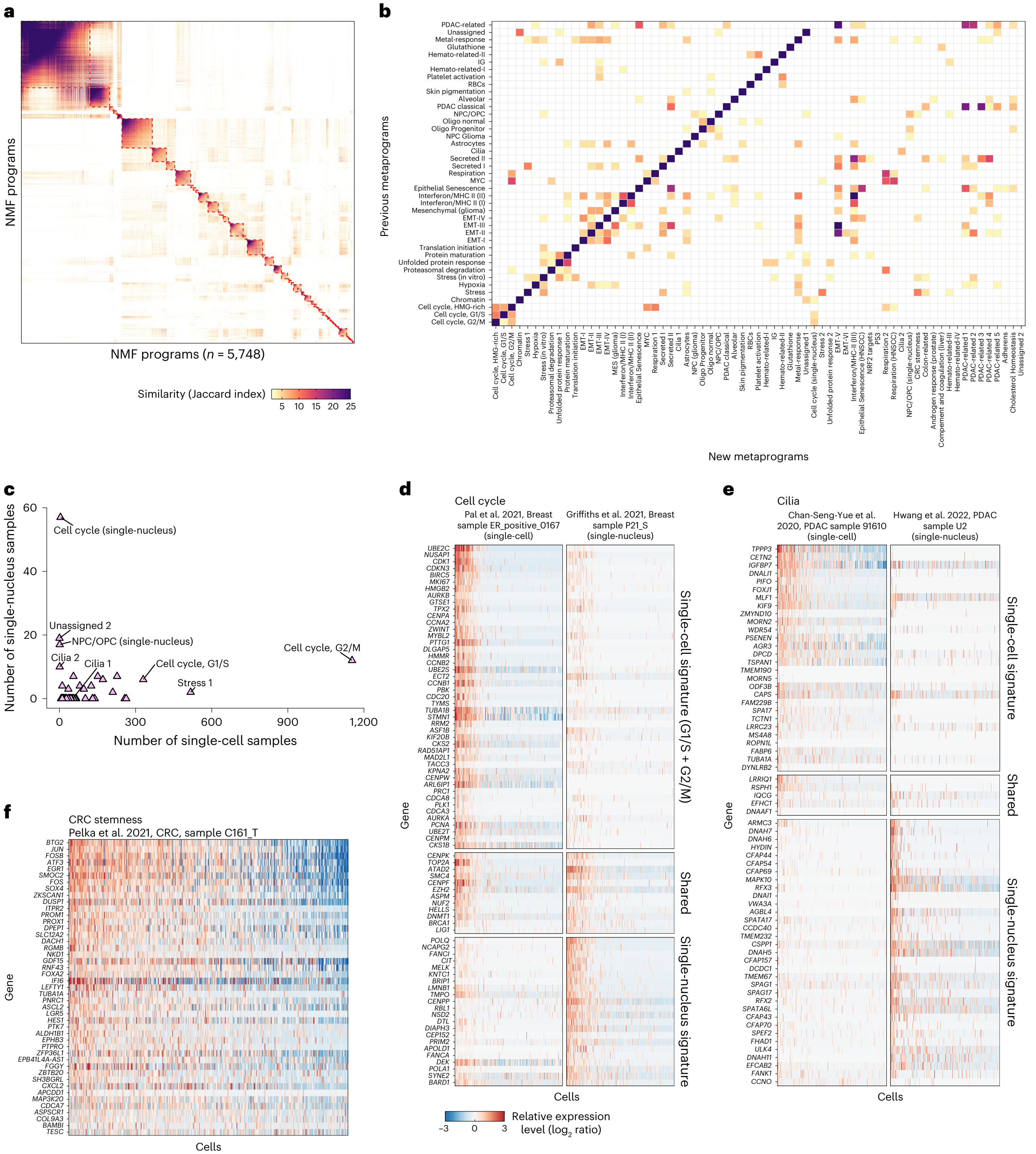


Fig. 3 | Updated MPs in malignant cells. a, Heat map showing the similarity (color, measured by Jaccard index) among 5,748 robust NMF programs based on their top 50 genes. Programs are ordered by clustering, with clusters (outlined in red) representing MPs. **b**, Heat map showing the similarity (color, as in **a**) between MPs generated from the current (*x* axis) and previous (*y* axis) 3CA versions. **c**, Scatter plot showing the number of samples profiled by snRNA-seq (*y* axis) versus scRNA-seq (*x* axis) that contributed NMF programs to each MP (points). Selected MPs are labeled. **d**, Heat map showing relative expression levels (color) of cell-cycle MP genes (rows) in individual cells (columns) from two representative breast cancer samples, one profiled by scRNA-seq and the other profiled by snRNA-seq[31,50]. Genes are partitioned according to their MPs of origin: 'single-cell signature' denotes the single-cell-derived G1/S and G2/M MPs; 'single-nucleus signature' denotes the single-nucleus-derived cell-cycle MP; 'shared' denotes the overlap. Cells in each sample are ordered by scores for the respective signatures and genes are ordered by correlation with these scores. **e**, Heat map as in **d** showing expression of genes from single-cell-derived and single-nucleus-derived cilia MPs in two representative PDAC samples[30,47]. **f**, Heat map showing relative expression levels (color, as in **d**,**e**) of genes in the CRC stemness MP (rows) in individual cells (columns) from a representative CRC sample[51]. Cells are ordered by scores for the CRC stemness gene signature and genes are ordered by correlation with these scores.

predominantly of programs from samples profiled by single-nucleus RNA-seq (snRNA-seq) (Fig. 3c). This reflects the increased representation of snRNA-seq among datasets in the new 3CA version and may also reflect biological differences in the distribution of RNA transcripts between the nucleus and cytosol[10]. For example, we detected a largely snRNA-seq-specific cell-cycle MP that correlates and partially overlaps with the G1/S and G2/M MPs derived from scRNA-seq samples (Fig. 3d). We also observed a second cilia MP mostly in snRNA-seq samples, exhibiting some correlation with the cilia program we defined previously (Fig. 3e). This new cilia variant contains multiple key cilia-associated genes that were not detected in the previous version, including *HYDIN*, *RFX3*, *CFAP44* and *DNAH7*.

A handful of entirely new MPs were detected, including cholesterol homeostasis, complement and coagulation and NRF2 targets. Notably, we also observed an MP capturing a 'stemness' phenotype in colorectal cancer (CRC), including genes such as *LGR5*, *PROM1* and *ASCL2* (Fig. 3f), possibly enabled by the large increase in CRC samples in the new 3CA version (Fig. 1d). Overall, the updated MPs are consistent with the hallmarks of ITH that we defined previously, supporting the robustness of this model. However, they indeed capture additional biological processes and future iterations on further expanded 3CA versions may warrant the introduction of new hallmarks.

Characterizing context-dependent gene expression patterns

The ability to resolve gene expression per cell type and cancer type, along with the increase in data quantity with the new 3CA version (Fig. 1b,d), enables an unbiased search across genes to identify cases of highly context-dependent gene expression. This includes the characterization of cell type markers, which we undertook for each of the most common cell types. First, for a given cell type, we identified those genes whose average expression (across cancer types) was highest in this cell type. For each of these genes, we then defined measures of their specificity and sensitivity as a marker of this cell type (Methods and Supplementary Data 4 and 5). Specificity reflects the extent to which a gene's expression is unique to a cell type, while sensitivity reflects the likelihood of detecting expression of a gene in a cell of this type. Marker genes can be identified as those with unusually high specificity and/or sensitivity (Fig. 4a). While this analysis is partially biased by the prior annotation of cell types in 3CA datasets, these annotations were generally determined at the level of clusters of cells and, thus, do not fully reflect marker performance at the level of individual cells.

Applying this approach to all cell types illuminated the distribution of markers across contexts (Fig. 4b,c). Markers were strongest for mast cells, with multiple genes scoring exceptionally highly for both specificity and sensitivity (*TPSB2*/*AB1*, *CPA3* and *MS4A2*). In other cell types, however, the choice of markers represents a compromise between these two measures. For example, *NKG7* is a highly sensitive marker for natural killer (NK) cells but is not highly specific because of its expression in T cells. Conversely, *KLRF1* is highly specific to NK cells but is detected in only 60% of NK cells. Meanwhile, markers were unreliable for nonmalignant epithelial and plasma cells and all but absent for malignant and dendritic cells. The paucity of markers that are simultaneously sensitive and specific may be because of high context specificity or high heterogeneity within a cell type. Malignant and epithelial cells are expected to have highly context-specific gene expression such that there is no universal epithelial or malignant marker. Dendritic cells divide into conventional and plasmacytoid, precluding the existence of universal dendritic cell markers. Moreover, because of high expression similarities between certain cell types, such as between T cells and NK cells, some classical cell type markers have low specificity in scRNA-seq data (Fig. 4b,c).

As no pan-cancer markers were found for malignant cells, we instead focused on identifying cancer-type-specific malignant cell markers. We found substantial heterogeneity in the strength and abundance of malignant markers between cancer types (Fig. 4d and Extended Data Fig. 2a). Some markers were consistent with prior knowledge, including *PMEL* in melanoma, *CDKN2A* (encoding p16) in human-papillomavirus-positive (HPV$^+$) head and neck cancer and *ESR1* in breast cancer (reflecting the high representation of estrogen-receptor-positive breast tumors in our cohort).

In addition to malignant cell markers, we further characterized those genes whose expression in malignant cells was most variable between cancer types (regardless of their cell type specificity). This analysis identified many highly context-dependent genes, with one of the most prominent being *KLK3*, encoding prostate-specific antigen, which is highly specific to prostate cancer (Fig. 4e and Extended Data Fig. 2b). Other examples of cancer-type-specific genes included *APOA2* in liver cancer and *ANKRD30A* in breast cancer. We added an interactive feature to the 3CA website that allows the exploration of gene specificity and sensitivity in different contexts (Extended Data Fig. 3; https://www.weizmann.ac.il/sites/3CA/marker-genes).

Cancer type specificity is highest in malignant cells

While we could find many genes with robust cancer-type-specific expression among malignant cells (Fig. 4e), we found far fewer for tumor microenvironment (TME) cell types (Extended Data Fig. 4a–c). To better characterize the context specificity of gene expression in different cell types, we aggregated the cells of each cell type from each tumor into a pseudobulk profile and then compared these pseudobulk profiles across all tumors. This analysis revealed extensive diversity for each cell type, reflecting the combined effects of cancer type, subtype, genetics, TME composition, spatial location, technical batch effects and possibly other variables. This complexity may be explored further in future work; our focus here was only on quantifying the effect of cancer type while controlling for technical batch effects. We defined an overall expression similarity between every pair of pseudobulk samples and then quantified the effect of cancer type by the average similarity of pairs from the same cancer type versus from different cancer types. To control for batch effects, both measures were calculated only for pairs from different studies (Extended Data Fig. 5a and Methods).

Surprisingly, for most cell types, we observed comparable similarity of pairs of pseudobulk samples from the same or from different cancer types, indicating a minimal effect of cancer type (Fig. 4f). Malignant cells exhibited by far the highest cancer type specificity, followed by nonmalignant epithelial cells. Malignant cells from different cancer types are associated with distinct sets of common genetic events and, thus, are expected to be highly distinct. Nonmalignant epithelial cells reflect parenchymal cells that are also expected to vary considerably between tissues. Yet all the other TME immune and stromal cell types appeared to have very limited cancer type specificity. A similar albeit weaker effect was observed for the patient specificity of cell types, although this analysis did not control for batch effects (Extended Data Fig. 5a,b). Thus, while immune and stromal cell types exhibit substantial diversity within tumors, their average expression profiles are only minimally dependent on the cancer type. This observation supports the pan-cancer approach previously used for the discovery of cellular states in immune and stromal cell types[11–14].

Pan-cancer comparison of proliferation rates

Proliferation is a defining feature of cancer but the cell-cycle behavior of different tumor types remains poorly characterized. Dozens of canonical cell-cycle genes are highly upregulated during the cell cycle in a phase-dependent manner[2,15]. Hence, scRNA-seq provides an efficient means to detect cycling cells and assign them to specific phases. Accordingly, 3CA offers an unprecedented opportunity to systematically compare cell-cycle patterns across contexts. Moreover, because of the substantial increase in data quantity with the new 3CA version (Fig. 1b,d), we now have the power and resolution needed for robust estimates of cell cycle, distinguishing G1/S and G2/M phases, within each cell type and cancer type, which was not possible previously.

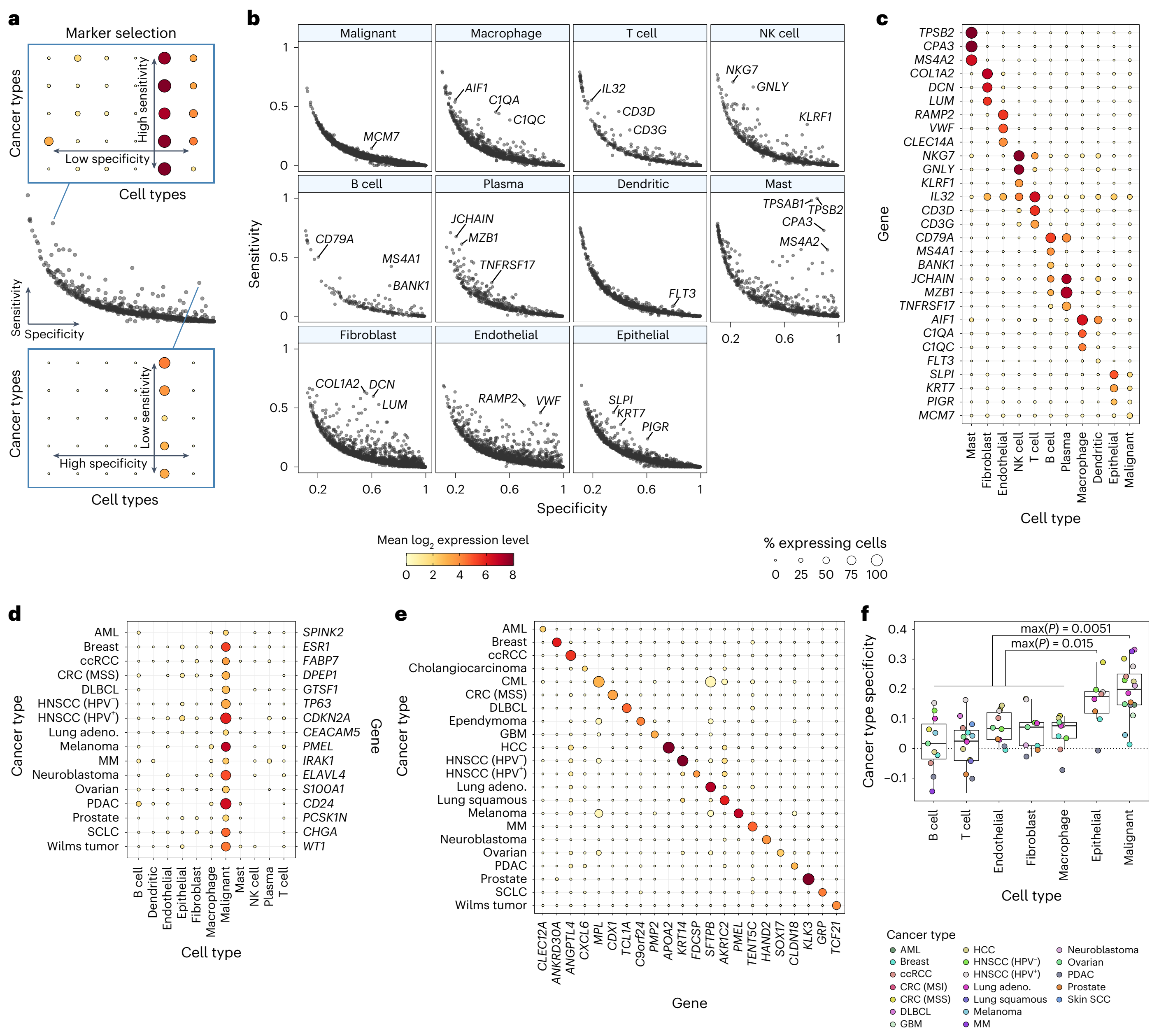


Fig. 4 | Context dependency of gene expression. a, Scheme illustrating the selection of cell type marker genes. Middle: scatter plot illustrating the sensitivity (*y* axis) and specificity (*x* axis) of genes (points) for a given cell type. Top and bottom: dot plots illustrating the expression levels (color) and percentage expressing cells (point size) of example marker genes across cell types (columns) and cancer types (rows). **b**, Scatter plots per cell type showing the sensitivity (*y* axis) and specificity (*x* axis) of genes (points) for each cell type. Selected genes with unusually high sensitivity or specificity are labeled. **c**, Dot plot showing the average $\log_2$ expression level (color) and percentage expressing cells (point size) in each cell type (columns) of marker genes labeled in **b** (rows). **d**, Dot plot showing the average $\log_2$ expression level (color) and percentage expressing cells (point size) in each cell type (columns) of selected malignant cell markers in each cancer type (rows). **e**, Dot plot showing the average $\log_2$ expression level (color) and percentage expressing cells (point size) of selected cancer-type-specific genes (*x* axis) in malignant cells in each cancer type (rows). **f**, Box plot showing the cancer type specificity (*y* axis) of each cell type (*x* axis) in each cancer type (points and color). *P* values for the differences between cell types were computed by pairwise paired, two-sided *t*-tests and adjusted to FDR < 0.05. Brackets are labeled with the maximal *P* values among pairwise comparisons between malignant cells, respectively epithelial cells, and all nonepithelial TME cell types (max(*P*) = 0.0051 and max(*P*) = 0.015, respectively). Pairwise differences between nonepithelial TME cell types and between malignant and epithelial cells were not significant. Boxes indicate the median and first and third quartiles; upper and lower whiskers extend to the maximal and minimal values no further than 1.5 × the IQR from the third and first quartiles. Groups (cell types) consist of *n* = 11, 15, 10, 9, 9, 8 and 16 data points (from left to right), corresponding to averages across pairwise correlations between biologically distinct samples.

We quantified cell cycle patterns across datasets by scoring cells for G1/S and G2/M gene signatures and learning appropriate thresholds from the distributions of these scores (Fig. 5a, Methods and Supplementary Data 6). Summary plots of these cell-cycle measurements are available for each published dataset on the 3CA website. The gene signatures were adapted for each dataset but highly similar results were obtained when using the same consensus signatures for all datasets (Extended Data Fig. 6a). The sequencing technology, number of detected genes and number of captured cells also had a negligible effect on the results (Extended Data Fig. 6b–i).

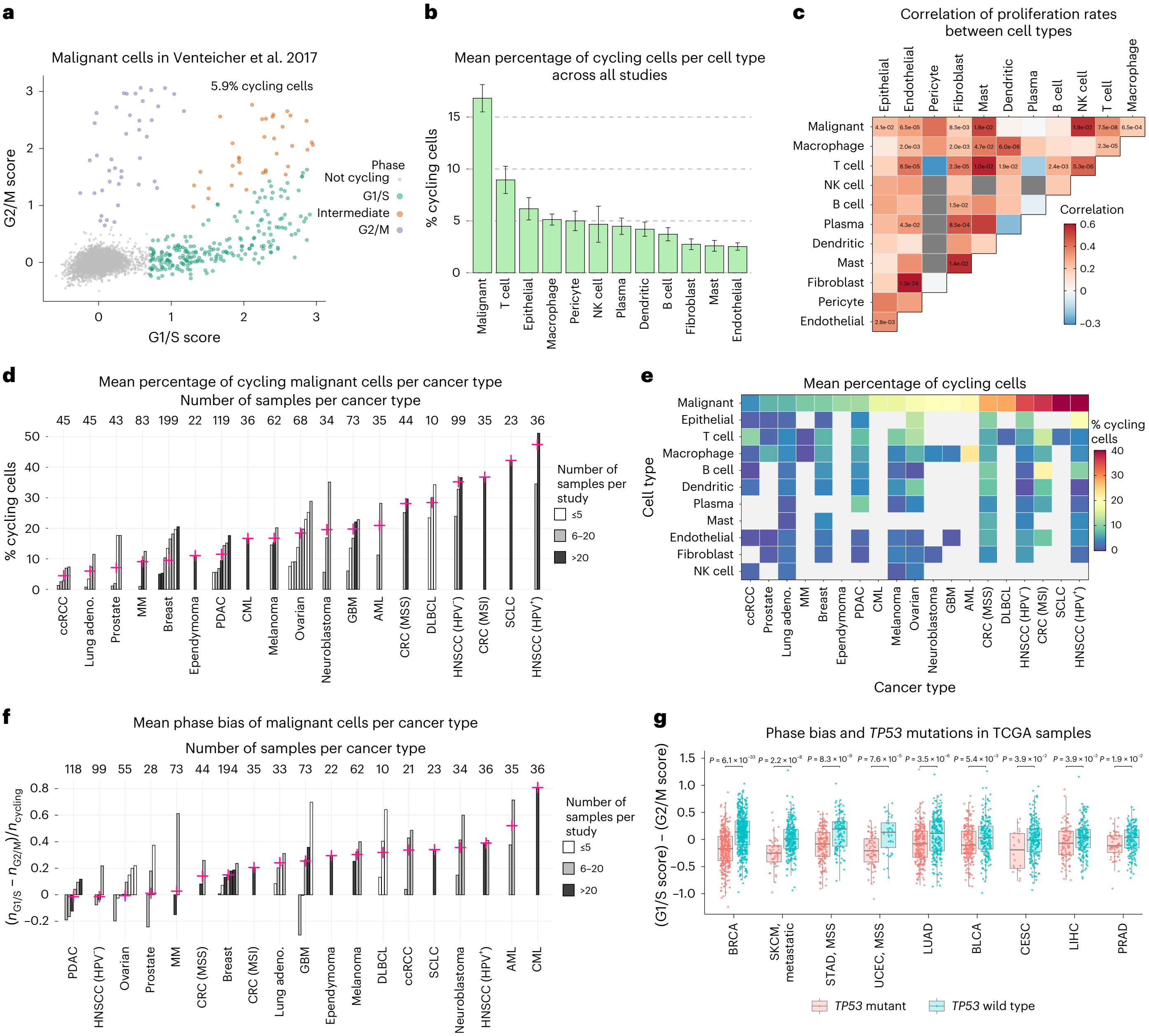


Fig. 5 | Quantification and comparison of cell-cycle patterns. a, Scatter plot showing scores for gene signatures of G1/S (*x* axis) and G2/M (*y* axis) for malignant cells (points) in the Venteicher et al. dataset[52], colored by cell-cycle phase. **b**, Bar plot showing the average percentage of cycling cells (*y* axis) per cell type (*x* axis; *n* = 75, 70, 38, 77, 10, 30, 24, 36, 52, 55, 23 and 54, from left to right). Error bars denote the standard error. Full distributions are shown in **d** and Extended Data Fig. 8. **c**, Heat map showing the Spearman correlation between cell types (color) of percentages of cycling cells. Significant correlations are labeled with *P* values (4.1×10^{-2}, 6.5×10^{-5}, 8.5×10^{-3}, 1.8×10^{-2}, 1.9×10^{-2}, 7.5×10^{-8}, 6.5×10^{-4}, 2.0×10^{-3}, 2.0×10^{-3}, 4.7×10^{-2}, 6.0×10^{-6}, 2.3×10^{-5}, 8.5×10^{-5}, 2.3×10^{-5}, 1.0×10^{-2}, 1.9×10^{-2}, 2.4×10^{-3}, 5.3×10^{-6}, 1.5×10^{-2}, 4.3×10^{-2}, 8.5×10^{-4}, 1.4×10^{-2}, 1.3×10^{-24} and 2.8×10^{-3}, from left to right and top to bottom), which were computed using a two-tailed test of zero correlation with algorithm AS 89 (ref. [53]) and adjusted to FDR < 0.05. **d**, Bar plot showing the percentage of cycling malignant cells (*y* axis) in each study (bars), grouped by cancer type (*x* axis). Crosses denote the average for each cancer type, weighted by the number of samples containing at least ten malignant cells. The bar color categorizes studies by the number of such samples and values above the plot denote the total number of such samples per cancer type. **e**, Heat map showing the average percentage of cycling cells (color, defined as for the crosses in **d**) per cancer type (*x* axis) and cell type (*y* axis). Gray squares indicate insufficient data. Cell types are ordered by average, while cancer types are ordered by the values for malignant cells. **f**, Bar plot showing the phase bias (*y* axis) of malignant cells in each study (bars), grouped by cancer type (*x* axis). Crosses, bar color and the number of samples per cancer type are defined as in **d**. **g**, Box plot showing the phase bias (*y* axis) of TCGA tumor samples (points; *n* = 334, 530, 52, 284, 161, 127, 73, 37, 241, 238, 191, 197, 20, 220, 101, 189, 56 and 177, from left to right), colored by *TP53* mutation status, for cancer types with a significant difference between mutant and wild type. Brackets are labeled with *P* values (6.1×10^{-33}, 2.2×10^{-8}, 8.3×10^{-9}, 7.6×10^{-5}, 3.5×10^{-6}, 0.0054, 0.039, 0.019 and 0.019, from left to right), which were computed using a two-sided *t*-test and adjusted to FDR < 0.05. Boxes indicate the median and first and third quartiles; upper and lower whiskers extend to the maximal and minimal values no further than 1.5 × the IQR from the third and first quartiles. All points represent distinct samples. BRCA, breast cancer; SKCM, skin cutaneous melanoma; STAD, stomach adenocarcinoma; MSS, microsatellite stable; UCEC, uterine corpus endometrial carcinoma; LUAD, lung adenocarcinoma; BLCA, bladder urothelial carcinoma; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; LIHC, liver hepatocellular carcinoma; PRAD, prostate adenocarcinoma.

A comparison of proliferation rates across cell types confirmed malignant cells as the most proliferative on average (Fig. 5b), with more than 15% of malignant cells typically observed cycling. However, all cell types showed some cycling activity, with surprisingly high proliferation in nonmalignant cells, especially T cells and normal epithelial cells. Moreover, we observed an overall positive correlation of proliferation rates between cell types across tumors (Fig. 5c and Extended Data Fig. 7), consistent with recent findings[7]. This suggests that the cell cycle may be stimulated in multiple cell types at once by TME factors and intercellular communication. An especially high correlation was observed between the proliferation of fibroblasts and the proliferation of endothelial cells.

Proliferation rates were highly variable across cancer types, especially for malignant cells (Fig. 5d,e and Extended Data Fig. 8a–j). The proliferation of malignant cells was lowest in clear cell kidney cancer (~5% cycling cells), consistent with the slow growth of kidney tumors and their resistance to chemotherapy[16]. At the opposite extreme, HPV$^+$ head and neck cancer was the most highly proliferative cancer type (>45% cycling cells). This may be explained by the mechanism of action of HPV, which silences p53 and pRb activity to promote progression through the cell cycle[17]. However, HPV$^-$ head and neck cancer was also among the most highly proliferative cancer types (~35% cycling cells), suggesting that high proliferation is a general feature of head and neck cancer.

Variability in phase bias explained by genomic alterations

As our method for measuring cell-cycle state distinguishes G1/S and G2/M phases, we were able to explore patterns of phase bias, that is, the relative proportion of cycling cells detected in G1/S versus G2/M. The expanded 3CA version is crucial for these comparisons because the proportion of cycling cells may be small (Fig. 5b,d,e); accordingly, many cells are required to accurately estimate this relative fraction within the population of cycling cells. Technical factors, including the choice of gene signatures, sequencing technology, number of detected genes and number of captured cells, had a negligible effect on our estimates of phase bias and downstream results (Extended Data Fig. 6).

As with overall proliferation, we observed high variability in phase bias across cancer types, especially in malignant cells, with acute and chronic myeloid leukemias being the most strongly biased toward G1/S and pancreatic and HPV$^-$ head and neck cancers having the strongest relative bias toward G2/M (Fig. 5f and Extended Data Fig. 9a–c). Interestingly, while HPV$^+$ and HPV$^-$ head and neck cancers were both among the most proliferative cancer types overall (Fig. 5d,e), they had opposite patterns of phase bias, with HPV$^+$ exhibiting a strong bias toward G1/S.

To test whether genomic alterations may explain the variation in phase bias across cancer types, we also computed phase bias scores in bulk RNA-seq profiles from TCGA for a variety of cancer types (Methods and Supplementary Data 7). Using these scores, we observed a strong association between G2/M bias and *TP53* mutations in multiple cancer types, consistent with the role of p53 as a gatekeeper of the G1/S transition (Fig. 5g and Extended Data Fig. 10a). Moreover, an unbiased analysis of many genes commonly mutated in cancer[18] identified *TP53* mutations as the most consistently associated with G2/M bias across cancer types (Extended Data Fig. 10b). Interestingly, this analysis suggested *RB1* mutations as the most consistently associated with G1/S bias (Extended Data Fig. 10b,c). As HPV acts in part through the degradation of pRb[19], which is not mutated in HPV$^-$ head and neck tumors, this association could explain the opposite phase bias patterns observed in HPV$^+$ and HPV$^-$ head and neck tumors. Various other driver mutations had strong context-specific associations with phase bias, such as *SMARCA4* and *EGFR* in lung adenocarcinoma, *PIK3CA* and *CDH1* in breast and stomach cancers and *CTNNB1* in endometrial cancer. Together, this analysis indicates that phase bias in cycling malignant cells is influenced by various cancer driver alterations, including those affecting *TP53* and *RB1*.

Discussion

3CA brings together many individual scRNA-seq efforts from across the cancer research community to unlock their combined potential. The increased volume and variety of curated data in this updated 3CA version, along with the enhanced website, confers heightened accessibility and statistical power to cancer scRNA-seq. This data resource will be immensely valuable to many research groups for various tasks, such as (1) to search for and download individual datasets most suited to a particular question; (2) to examine expression of genes or gene signatures of interest across cell types and cancer types; (3) to conduct pan-cancer analyses and uncover relationships between diseases; and (4) to fine-tune statistical models and algorithms. 3CA fills the role of a central source of scRNA-seq data for all cancer researchers, who may in turn contribute new datasets to further enrich this resource.

While other studies have presented repositories for processed scRNA-seq data[8,20–24], these have limited size, lack a cancer focus or are centered around TME factors such as immune cells. 3CA prioritizes malignant cells and contains extensive data visualizations and analyses detailing their diversity within and between cancer types. Our efforts also included careful curation of tumor clinical annotations, which are typically sparse in individual scRNA-seq datasets. As 3CA continues to grow, these clinical annotations will enrich future research efforts by uncovering single-cell expression patterns that correlate with clinical outcomes.

In our processing and analysis of 3CA datasets, we avoided using data integration methods, such as those offered by scANVI[25], Harmony[26] and Seurat[27]. These methods apply certain assumptions to identify and remove technical effects in scRNA-seq data but there is currently no widely agreed standard and they likely remove some biological signal[28]. This is especially true in cancer, where much of the transcriptional variation between tumors arises from their unique genetic and epigenetic profiles rather than from batch effects. Importantly, our analysis either focused on ITH or reported only average values across many samples. The high sample size in 3CA allows confidence in these averages and their accuracy will improve further with the inclusion of more datasets. However, there would be clear advantages to a fully integrated scRNA-seq data resource of this size in which expression levels may be directly compared between any two samples. Further research is needed to establish optimal methods to integrate 3CA data while preserving biological signal.

We anticipate the further extension of 3CA in at least three ways. Firstly, sequencing efforts are becoming larger, with individual studies sequencing upward of 50 samples. Including more such studies will greatly boost 3CA's statistical power. Secondly, we expect new studies to include cancer types currently underrepresented in 3CA, as well as rare sample types such as post-treatment tumors, metastatic lesions and circulating tumor cells. Thirdly, snRNA-seq from frozen tumor samples is quickly becoming common[29–34] and methods also exist for profiling fixed tissue[35–39]. These technologies open up many new possibilities by lifting the restriction to sequencing fresh tissue; hence, we expect 3CA to grow and diversify substantially in the coming years.

Methods

All analyses in this study were performed using R version 4.1.1. Analyses for Figs. 3–5 incorporated additional unpublished datasets on glioma, Schwannoma and head and neck cancer.

Data preprocessing

Datasets were filtered to remove cells with few detected genes. The minimum threshold was usually 1,000 and never lower than the cutoff used in the original studies. Following this, expression levels were converted to $\log_2(\mathrm{TPM}/10 + 1)$, where the factor 1/10 reflects an estimated upper bound of 100,000 for the number of transcripts in single-cell libraries. An exception to this was Fig. 3d–f where, to ensure visually similar color scales, expression levels were divided by the median of the cell totals instead of 100,000.

Data visualizations on the 3CA website

To generate UMAP plots for each dataset, we removed samples having fewer than ten cells and then restricted the data to genes for which $\log_2(\text{mean(TPM)} + 1) \geq 4$. A partial singular value decomposition was computed using the IRLBA algorithm[40] and a UMAP was computed using the first 50 of the resulting principal components. Visualizations of cell type marker expression in each dataset were made using a manually curated list of canonical cell type markers and restricting the data to cell types having at least ten cells in the given dataset.

CNAs were inferred from the scRNA-seq data using a method described previously[2,9]. Briefly, in each dataset, we ordered the genes by chromosomal position and then computed the running average of their centered expression levels. After median centering per cell, we adjusted these values relative to those of a selection of reference cells. These reference cells were chosen to be confidently nonmalignant and to have expression profiles as similar as possible to those of the presumed malignant cells. The exact choice of reference cells was specific to each dataset, depending on the cell types detected therein.

MP exploration features of the 3CA website

To measure the distribution of MP expression among cells of a given type in each sample or dataset, we computed MP scores using a method described previously[2], which measures the expression of signature genes relative to a set of control genes. These control genes are chosen to have comparable expression levels to the signature genes but no coherent association with any biological process. Briefly, genes were ranked by average expression and divided into a number of discrete bins. Then, for each signature gene *g*, we sampled a set of control genes from the corresponding expression bin and computed the difference between the expression of *g* and the average expression of these control genes. The signature score was then defined as the average of these relative expression levels across all genes in the signature.

Having scored the cells for each of the MPs for the corresponding cell type, we assigned each cell to an MP as follows. If none of a cell's MP scores was greater than 1, this cell was classified as 'unassigned', whereas, if at least one score was greater than 1, the cell was assigned to the MP with maximal score. Among the cells given an assignment, rare MPs (with 'rarity' depending on whether this analysis was per sample or per dataset) were considered spurious and cells assigned to them were reclassified as unassigned.

Defining MPs

Definition of MPs in malignant cells closely followed the approach used previously[7]. Expression levels were centered per gene and negative values were set to zero. We then applied non-negative matrix factorization (NMF), with $K = 4, \ldots, 9$, to the malignant cells in each sample separately. Summarizing each of the resulting NMF programs by its top 50 genes (according to the NMF coefficients), we filtered these according to the following scheme. First, we identified NMF programs that were (1) robust within a tumor, defined as having an overlap of at least 35 genes with other NMF programs derived from the same sample with different *K* values and (2) robust across tumors, defined as having an overlap of at least ten genes with an NMF program from a different sample. We then filtered these robust NMF programs to remove redundancy by ranking programs in each tumor in decreasing order of similarity with programs in different tumors and, iterating along this list, removing each program with an overlap of more than ten genes with the one preceding it.

We then applied a custom algorithm to cluster these NMF programs, in which the program with the highest number of considerable overlaps (at least ten genes) with other programs was selected as the founder of a cluster and the programs with strongest considerable overlap with this cluster were successively added and a consensus gene list updated until no considerably overlapping programs remained and a new cluster was initiated. Each cluster then defined an MP. Genes in each MP were ranked by the number of datasets whose NMF programs contributed each gene to this MP. For downstream analysis, we used the top 50 genes in each MP.

This process yielded 111 initial MPs. We further filtered these to remove (1) MPs deriving from a single study; (2) MPs deriving from two studies, one of which contributed only a single NMF program; (3) MPs defined by fewer than five NMF programs; (4) MPs enriched in genes that likely reflected poor data quality or technical artifacts (for example, ribosomal or mitochondrial genes); and (5) MPs that were suspected to reflect doublets or errors in cell type annotation because of a high similarity with signatures of nonmalignant cell types. This filtering yielded a final list of 67 MPs. When counting the number of snRNA-seq versus scRNA-seq samples that contributed NMF programs to each MP (Fig. 3c), studies that contained both single-cell and single-nucleus samples were excluded[41–43].

Average gene expression per cell type, study and cancer type

For a given gene, and for each of expression level and percentage expressing cells, the average for a given cell type and study or cancer type was computed as follows. First, we computed the average across all cells of that type from a given study that were profiled with the same sequencing technology. If a study used more than one sequencing technology, these values were further averaged across technologies within the study. We then calculated the weighted average of these values across datasets for each cancer type, with weights equal to the number of samples in each dataset having at least ten cells of that type.

Sensitivity and specificity of gene expression

We included in this analysis only those cancer types having at least two studies and 10 samples in total or one study with at least 20 samples. We included genes that were detected in datasets for at least 20 cancer types. Then, for each gene, we conducted three comparisons of its average expression levels (average per cell type and cancer type, defined above) to define sensitivity and specificity to three different contexts:

(i) For the first comparison, we first computed the median of the given gene's averages for each cell type (median across cancer types) and then compared these medians between cell types. The values of interest are, therefore, the cell type medians and the contexts for comparison are the different cell types, with the context of interest being a single chosen cell type.

(ii) For the second comparison, we first fixed a cancer type and then compared the given gene's averages between malignant cells and all other cell types within this cancer type. The values of interest are the gene averages and the contexts for comparison are the cell types within this cancer type, with the context of interest being the malignant cells. This comparison was repeated with each cancer type, excluding carcinomas with insufficient data for nonmalignant epithelial cells.

(iii) For the third comparison, we first fixed a cell type (namely malignant cells) and then compared the given gene's averages between malignant cells from different cancer types. The values of interest are the gene averages and the contexts for comparison are malignant cells in different cancer types, with the context of interest being malignant cells in a single chosen cancer type.

In each case, if the relevant value was highest in the context of interest, its sensitivity to this context was defined as this value (the highest value across contexts) and its specificity was defined as $1/(x + 1)$, where x is the second highest value across contexts. The sensitivity measures were then scaled to the interval [0, 1] by dividing by the maximum across genes.

Number of cancer-type-specific genes per cell type

For each cell type, we quantified the number of cancer-type-specific genes for this cell type as follows. For each gene whose average expression was highest in this cell type, we defined a value y to be this average value (corresponding to the sensitivity described above) and a value x to be its highest average expression level among all other cell types (corresponding to the value x described above). Then, for each value c from a manually selected set of thresholds, we defined the number of cancer-type-specific genes for this cell type to be the number of genes with $y > x + c$. As expression levels were defined in log space, $c = 1$, 2 and 3 denote genes with ≥2-fold, ≥4-fold and ≥8-fold higher expression in this cell type than in other cell types.

Cancer type and patient specificity of cell types

Pseudobulk profiles were computed for each cell type in each distinct tumor sample by averaging the expression levels (transcripts per 100,000) for each gene across cells of this cell type in this sample. Then, after excluding ribosomal genes, we restricted the data to the top 5,000 genes with highest variance across pseudobulk profiles. We then log-transformed these profiles and computed their pairwise Pearson correlations. Cancer type specificity was then defined, for each cell type and cancer type, by the average correlation between pseudobulks of this cell type and cancer type from different studies minus the average correlation between pseudobulks of this cell type from different cancer types. Patient specificity was defined as 1 minus the average correlation between pseudobulks of this cell type from the same study. Differences between cell types were assessed using pairwise paired t-tests. P values were adjusted to a false discovery rate (FDR) < 0.05.

Cell-cycle quantification and phase assignment

To analyze the expression of cell-cycle genes in each dataset, we constructed initial G1/S and G2/M gene signatures by taking the unions of published G1/S and G2/M signatures from the Scandal R package (https://github.com/dravishays/scandal) and four other sources[2,7,44,45]. We then removed genes in the overlap of the resulting G1/S and G2/M gene sets. G1/S and G2/M scores were defined similarly to the MP scores described above, with slight differences. For a given signature gene g and corresponding set of control genes, we assigned two values to g. First, we computed the difference between the expression level of g and the average expression level of the sampled control genes, capping these differences at ±3 to lessen the influence of extreme values. Second, we assigned 1 if the expression level of g was greater than the bin average and 0 otherwise. We then defined two scores for each gene signature by averaging each of these sets of values across all genes in the signature. We refer to these two scores as 'mean-based' and 'count-based', respectively.

In each dataset, we then filtered the G1/S and G2/M signatures to maximize orthogonality. To do this, we computed the correlation of each gene with the mean-based G1/S and G2/M scores and excluded genes for which the difference between these correlations was 0.1 or less. If, after this step, there were fewer than 20 genes remaining in either signature, this dataset was excluded from further analysis. If there were more than 50 genes remaining in a given signature, we ranked them by their correlation values for the same signature and by the differences between their G1/S and G2/M correlations. We retained the top 50 genes according to the average of these ranks. Note that this procedure meant that we obtained different filtered signatures for each dataset. We further defined consensus signatures by ranking all G1/S and G2/M genes by their occurrence in the dataset-specific signatures, retaining the top 50. We compared the cell-cycle estimates yielded by the consensus and dataset-specific signatures (Extended Data Fig. 6a) but all downstream scRNA-seq analysis used the dataset-specific signatures.

After recalculating the G1/S and G2/M scores using these filtered signatures, we defined thresholds to distinguish cycling from noncycling cells using a bootstrapping approach. For each cell type, we generated 1,000 null distributions consisting of mean-based and count-based scores for 100 'pseudocells'. These scores were computed as above except that, for each signature gene, expression levels were taken from a different set of 100 cells that were randomly sampled (with replacement) from the expression matrix. This approach removed the correlation between signature genes while preserving each gene's distribution of expression levels. Then, for each cell c and each score type, a P value was obtained by one-sided binomial test on the number of null distributions containing the score for c (that is, in which the highest-scoring pseudocell scored higher than c). After adjustment to FDR < 0.05, cells were classified as cycling if, for both score types, the P value for either G1/S or G2/M score was less than 0.05. To correct for biases in the thresholds because of a high proportion of cycling cells, for cell types of which more than 10% were classified as cycling, we reassigned each signature gene to a control gene set on the basis of its average expression in the noncycling cells. Scores, null distributions, P values and cycling assignments were then recomputed for these cell types as before.

For each of the G1/S and G2/M signatures, we next defined a consensus significance variable, taking the value 1 if adjusted P values for both mean-based and count-based scores were less than 0.05 and 0 otherwise. We then defined new mean-based score thresholds in each cell type by fitting a binomial regression model of the mean-based score against this consensus significance. Cells were reclassified as cycling if their mean-based scores for either G1/S or G2/M passed the corresponding regression threshold. Scores were then recentered per cell type relative to the average mean-based scores of the noncycling cells. Final mean-based score thresholds for each of G1/S and G2/M signatures were defined by manually examining the distribution of the regression thresholds across all cell types and datasets and choosing an appropriately conservative consensus value. Cells were reclassified as cycling according to these consensus thresholds. Lastly, cells were assigned to G1/S and G2/M phases using a fold change threshold of 2 on recentered mean-based scores. Cells assigned to neither phase were classified as intermediates.

Estimates of cell-cycle proportion and phase bias

Estimates of proportion of cycling cells and phase bias were computed per study, sequencing technology, cancer type and cell type only for those combinations having at least 100 cells. For each such combination, to avoid bias arising from samples with exceptionally many or few cycling cells, we defined outlier samples as those with the number of cells outside the bounds [25th percentile − 1.5 × the IQR (interquartile range), 75th percentile + 1.5 × the IQR]. Samples below the lower bound were excluded from further analysis, while those above the upper bound were downsampled to this bound. The proportion of cycling cells across these samples was then defined as the number of cells classified as cycling (of any phase) divided by the total number of cells. If there were at least 100 cycling cells across these samples, the phase bias was defined as the difference between the number of cells assigned to G1/S and the number assigned to G2/M, divided by the total number of cycling cells. If a study contained data generated by multiple sequencing technologies, the estimates for this study were averaged across sequencing technologies, weighted by the number of samples containing at least ten cells, to obtain one estimate per study, cancer type and cell type for each of cycling proportion and phase bias. To obtain pan-cancer cell-type-level estimates, these study-level estimates were averaged across all studies. To obtain cancer-type-level estimates for each cell type, the study-level estimates were averaged across studies for that cancer type, weighted by the number of samples with ≥10 cells. Cycling proportions were also calculated per sample only in those samples having at least 100 cells.

These computations were also performed after restricting to 10x datasets. To measure the correlation of cycling proportion and phase bias with number of detected genes and number of captured cells for each cell type and sequencing platform, we identified outliers manually and computed regression lines and Pearson correlation both with and without outliers. This analysis included only those cancer types represented by both sequencing platforms and cell types represented in at least five datasets of each sequencing platform.

Correlation of cell cycle between cell types

To compute the correlation of cell cycle between cell types across all cancer types, we first centered the per-sample estimates of cycling proportion within each study and cell type. Then, for each pair of cell types having such estimates in at least ten of the same samples, Spearman correlation was computed between cycling proportions across these common samples. We computed the correlation of cell cycle between cell types within each cancer type and the significance of these correlations, using the same procedure per cancer type, before filtering out cell types with fewer than 20 such correlation values across all cancer types and cell type pairs.

TCGA mutations analysis

TCGA expression and mutation data were obtained from the Broad Genome Data Analysis Center Firehose (gdac.broadinstitute.org). The expression data were downloaded in the form of RSEM[46] 'scaled estimates', multiplied by 10^6 (giving a measure similar to TPM) and log-transformed. For CRC and stomach cancer, tumors with microsatellite instability (MSI) were defined as those with more than 500 mutations in total; similarly, MSI endometrial tumors were defined as those with more than 400 mutations. Following this, we restricted our attention to a previously defined list of genes commonly mutated in cancer[18] and, among these, we focused on nonsense, nonstop, frame shift and splice or translation start site mutations, along with missense mutations and in-frame insertions and deletions occurring in at least two tumors.

Cell-cycle scores were computed per cancer type using the consensus cell-cycle gene signatures, via the method described above, with bulk profiles in place of individual cells. We then defined the phase bias as the difference between G1/S and G2/M scores. In each cancer type having at least 100 tumors and for each gene from the above list that was mutated in at least ten tumors and at least 5% of tumors and wild type in at least ten tumors, we compared phase bias scores between mutant and wild-type tumors using two-sample *t*-tests. *P* values were adjusted to FDR < 0.05 separately for *TP53* and *RB1*.

Statistics and reproducibility

This study used only external datasets. Our analysis used all available samples; we did not perform any analysis to predetermine sample size and we did not choose samples, patients or groups ourselves. In particular, no randomization or blinding was performed. We excluded data from specific analyses based on insufficient number of samples or cells. Two datasets were excluded from several analyses because of high technical noise. Where *t*-tests were used to calculate statistical significance, data distribution was assumed to be normal but this was not formally tested. All analyses may be reproduced using the code available on GitHub (https://github.com/tiroshlab/3ca).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

This study used only external datasets and did not involve the generation of new data. All published single-cell datasets are available on the 3CA website (https://www.weizmann.ac.il/sites/3CA/), with the exception of one dataset[47], for which permission for sharing through 3CA was not granted; this dataset is available through the European Genome–Phenome Archive under accession number EGAS00001002543. Additional unpublished datasets used will be added to the 3CA website when possible. TCGA data were obtained online (http://gdac.broadinstitute.org/). Data for reproducing the figures in this article are provided in the Supplementary Information. Source data are provided with this paper.

Code availability

Source code for all analyses in this study and for generating the figures available on the 3CA website is available on GitHub (https://github.com/tiroshlab/3ca).

References

1. Marusyk, A., Janiszewska, M. & Polyak, K. Intratumor heterogeneity: the Rosetta stone of therapy resistance. *Cancer Cell* **37**, 471–484 (2020).
2. Tirosh, I. et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).
3. Puram, S. V. et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171**, 1611–1624 (2017).
4. Elyada, E. et al. Cross-species single-cell analysis of pancreatic ductal adenocarcinoma reveals antigen-presenting cancer-associated fibroblasts. *Cancer Discov.* **9**, 1102–1123 (2019).
5. Vegliante, R., Pastushenko, I. & Blanpain, C. Deciphering functional tumor states at single-cell resolution. *EMBO J.* **41**, e109221 (2022).
6. Aran, D. Single-cell RNA sequencing for studying human cancers. *Annu. Rev. Biomed. Data Sci.* **6**, 1–22 (2023).
7. Gavish, A. et al. Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. *Nature* **618**, 598–606 (2023).
8. Han, Y. et al. TISCH2: expanded datasets and new tools for single-cell transcriptome analyses of the tumor microenvironment. *Nucleic Acids Res.* **51**, D1425–D1431 (2023).
9. Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
10. Zaghlool, A. et al. Characterization of the nuclear and cytosolic transcriptomes in human brain tissue reveals new insights into the subcellular distribution of RNA transcripts. *Sci. Rep.* **11**, 4076 (2021).
11. Zheng, L. et al. Pan-cancer single-cell landscape of tumor-infiltrating T cells. *Science* **374**, abe6474 (2021).
12. Tang, F. et al. A pan-cancer single-cell panorama of human natural killer cells. *Cell* **186**, 4235–4251 (2023).
13. Cheng, S. et al. A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid cells. *Cell* **184**, 792–809 (2021).
14. Buechler, M. B. et al. Cross-tissue organization of the fibroblast lineage. *Nature* **593**, 575–579 (2021).
15. Whitfield, M. L. et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13**, 1977–2000 (2002).
16. Cohen, H. T. & McGovern, F. J. Renal-cell carcinoma. *N. Engl. J. Med.* **353**, 2477–2490 (2005).
17. Puram, S. V. et al. Cellular states are coupled to genomic and viral heterogeneity in HPV-related oropharyngeal carcinoma. *Nat. Genet.* **55**, 640–650 (2023).
18. Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
19. Graham, S. V. The human papillomavirus replication cycle, and its links to cancer progression: a comprehensive review. *Clin. Sci.* **131**, 2201–2221 (2017).
20. CZI Cell Science Program et al. CZ CELL×GENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res.* **53**, D886–D900 (2025).

21. Camps, J. et al. Meta-analysis of human cancer single-cell RNA-seq datasets using the IMMUcan database. *Cancer Res.* **83**, 363–373 (2023).
22. Zeng, J. et al. CancerSCEM: a database of single-cell expression map across various human cancers. *Nucleic Acids Res.* **50**, D1147–D1155 (2022).
23. Yuan, H. et al. CancerSEA: a cancer single-cell state atlas. *Nucleic Acids Res.* **47**, D900–D908 (2019).
24. Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019**, baz046 (2019).
25. Xu, C. et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).
26. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
27. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
28. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
29. Wang, L. et al. A single-cell atlas of glioblastoma evolution under therapy reveals cell-intrinsic and cell-extrinsic therapeutic targets. *Nat. Cancer* **3**, 1534–1552 (2022).
30. Hwang, W. L. et al. Single-nucleus and spatial transcriptome profiling of pancreatic cancer identifies multicellular dynamics associated with neoadjuvant treatment. *Nat. Genet.* **54**, 1178–1191 (2022).
31. Griffiths, J. I. et al. Serial single-cell genomics reveals convergent subclonal evolution of resistance as patients with early-stage breast cancer progress on endocrine plus CDK4/6 therapy. *Nat. Cancer* **2**, 658–671 (2021).
32. Nassiri, F. et al. A clinically applicable integrative molecular classification of meningiomas. *Nature* **597**, 119–125 (2021).
33. Jansky, S. et al. Single-cell transcriptomic analyses provide insights into the developmental origins of neuroblastoma. *Nat. Genet.* **53**, 683–693 (2021).
34. Kim, C. et al. Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. *Cell* **173**, 879–893 (2018).
35. 10x Genomics. *Chromium Fixed RNA Profiling Reagent Kits for Multiplexed Samples*. Report No. CG000527 (10x Genomics, 2023).
36. Chung, H. et al. SnFFPE-seq: towards scalable single nucleus RNA-seq of formalin-fixed paraffin-embedded (FFPE) tissue. Preprint at *bioRxiv* https://doi.org/10.1101/2022.08.25.505257 (2022).
37. Wang, T. et al. snPATHO-seq, a versatile FFPE single-nucleus RNA sequencing method to unlock pathology archives. *Commun. Biol.* **7**, 1340 (2024).
38. Xu, Z. et al. High-throughput single nucleus total RNA sequencing of formalin-fixed paraffin-embedded tissues by snRandom-seq. *Nat. Commun.* **14**, 2734 (2023).
39. Phan, H. V. et al. High-throughput RNA sequencing of paraformaldehyde-fixed single cells. *Nat. Commun.* **12**, 5636 (2021).
40. Baglama, J. & Reichel, L. Augmented implicitly restarted Lanczos bidiagonalization methods. *SIAM J. Sci. Comput.* **27**, 19–42 (2005).
41. Biermann, J. et al. Dissecting the treatment-naive ecosystem of human melanoma brain metastasis. *Cell* **185**, 2591–2608 (2022).
42. Nath, A. et al. Evolution of core archetypal phenotypes in progressive high grade serous ovarian cancer. *Nat. Commun.* **12**, 3039 (2021).
43. Wang, L. et al. The phenotypes of proliferating glioblastoma cells reside on a single axis of variation. *Cancer Discov.* **9**, 1708–1719 (2019).
44. Neftel, C. et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell* **178**, 835–849 (2019).
45. Kinker, G. S. et al. Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nat. Genet.* **52**, 1208–1218 (2020).
46. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
47. Chan-Seng-Yue, M. et al. Transcription phenotypes of pancreatic cancer are driven by genomic events during tumor evolution. *Nat. Genet.* **52**, 231–240 (2020).
48. Raghavan, S. et al. Microenvironment drives cell state, plasticity, and drug response in pancreatic cancer. *Cell* **184**, 6119–6137 (2021).
49. Kürten, C. H. L. et al. Investigating immune and non-immune cell interactions in head and neck tumors by single-cell RNA sequencing. *Nat. Commun.* **12**, 7338 (2021).
50. Pal, B. et al. A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. *EMBO J.* **40**, e107333 (2021).
51. Pelka, K. et al. Spatially organized multicellular immune hubs in human colorectal cancer. *Cell* **184**, P4734–4752.E20 (2021).
52. Venteicher, A. S. et al. Decoupling genetics, lineages, and microenvironment in *IDH*-mutant gliomas by single-cell RNA-seq. *Science* **355**, eaai8478 (2017).
53. Best, D. J. & Roberts, D. E. Algorithm AS 89: the upper tail probabilities of Spearman's rho. *J. R. Stat. Soc. C Appl. Stat.* **24**, 377–379 (1975).

Acknowledgements

We thank all members of the I.T. lab for helpful discussions and comments. We thank the Weizmann Information Technology Internet and Mobile Development Team for their work on the 3CA website. This work was supported by the Israeli Council for Higher Education through the Weizmann Data Science Research Center and by the Israel Science Foundation. I.T. is the incumbent of the Dr. Celia Zwillenberg-Fridman and Dr. Lutz Zwillenberg Career Development Chair and is supported by the Zuckerman Science, Technology, Engineering and Mathematics Leadership Program, the Mexican Friends New Generation and the Benoziyo Endowment Fund. R.H. is funded by the Walter Benjamin Program from the German Research Foundation.

Author contributions

M.T. and I.T. conceptualized and designed the study, interpreted the results and wrote the paper. M.T. performed the majority of computational analyses, generated the data visualizations for the 3CA website and managed the hosting of data on the website for public download. A.G. performed the analysis defining MPs in malignant cells and quantifying the relative contributions of scRNA-seq and snRNA-seq profiles. C.B. and M.T. performed the analysis of cancer type specificity of cell types. R.T., R.H. and M.M. curated datasets for 3CA. S.V.P. contributed scRNA-seq data for head and neck tumors. A.G., R.T., R.H., M.M. and S.V.P. reviewed the paper and provided feedback. The study was cosupervised by M.T. and I.T.

Competing interests

I.T. is an advisory board member of Immunitas Therapeutics. The other authors declare no competing interests.

Additional information

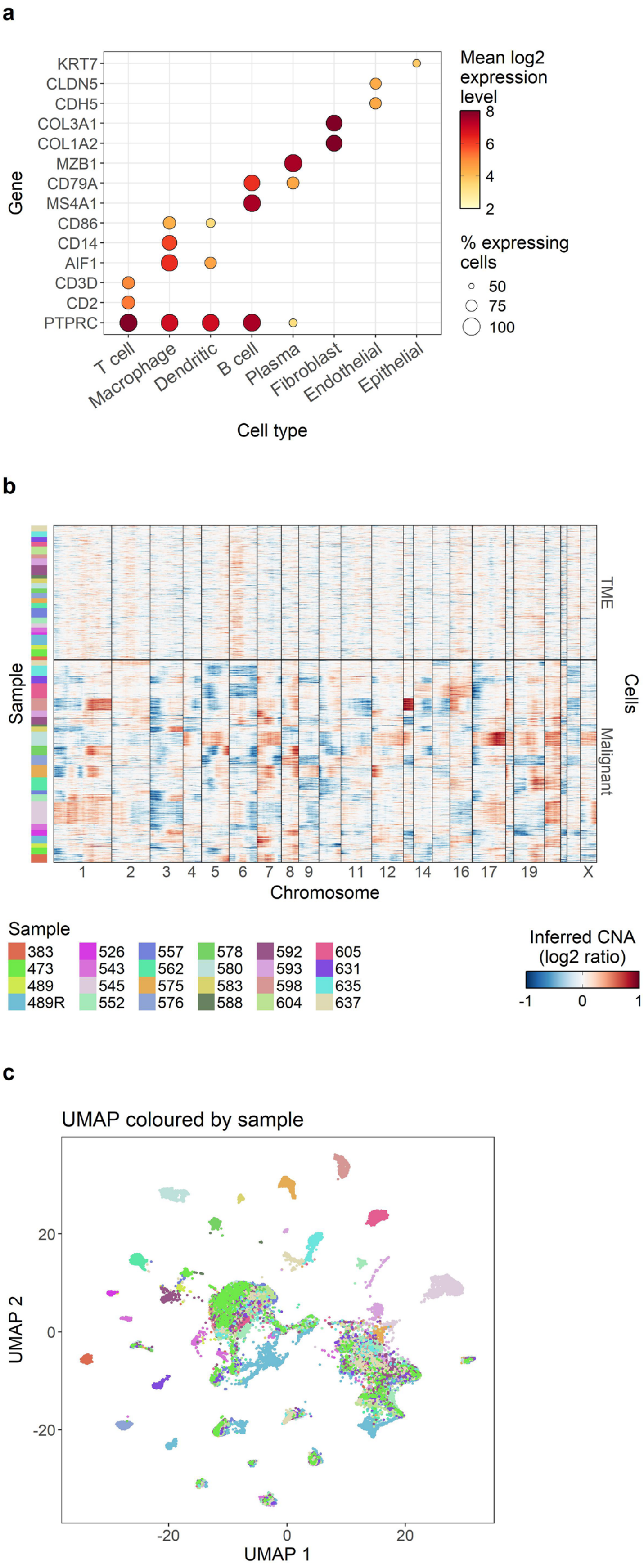
Extended data is available for this paper at https://doi.org/10.1038/s43018-025-00957-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43018-025-00957-8.

Correspondence and requests for materials should be addressed to Michael Tyler or Itay Tirosh.

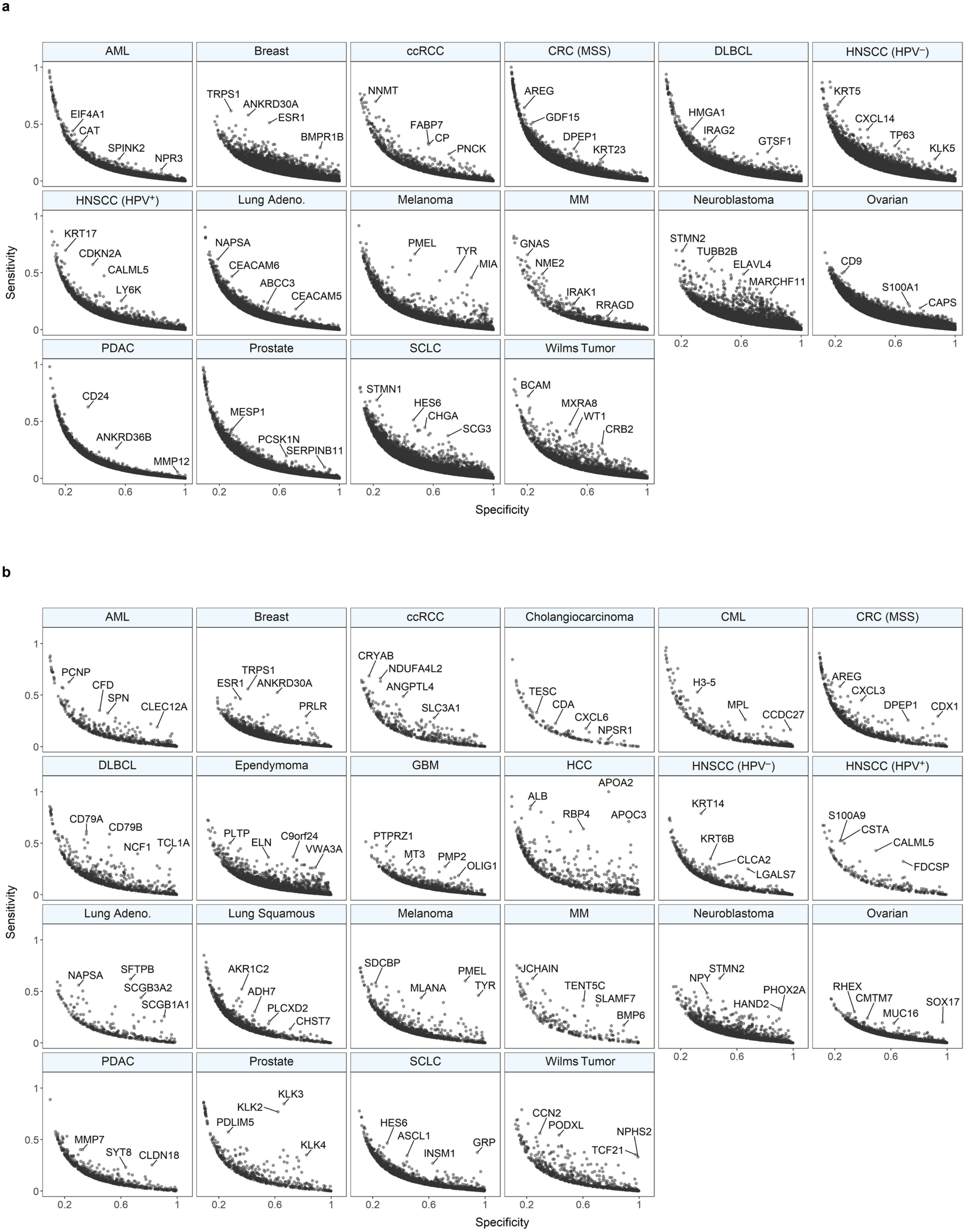
Peer review information *Nature Cancer* thanks Francisco Martínez Jiménez and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

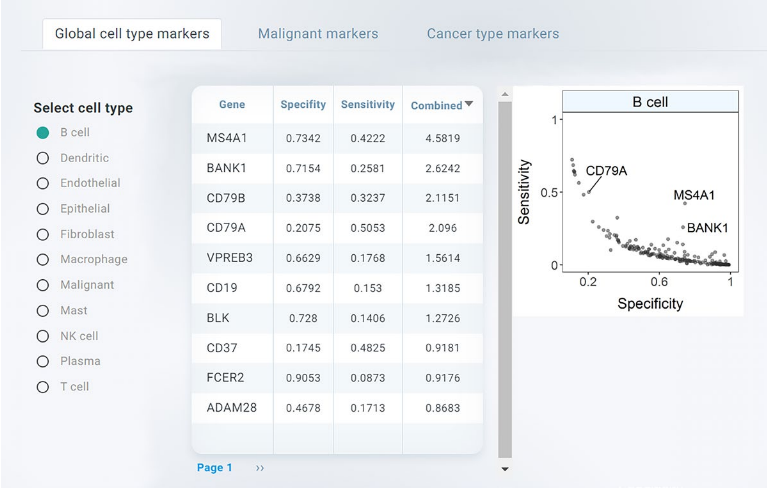


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Overview of 3CA cell annotation and data summary features. a. Dot plot showing the average expression level (colour) and percentage expressing cells (point size) of a selection of cell type marker genes (rows) in each cell type (columns) in the Raghavan et al. dataset[48]. **b**. Heatmap showing inferred copy number alteration (CNA) values (colour, quantified as $\log_2$ ratio, with blue indicating depletion and red amplification) at each chromosomal position (columns) for a representative subset of cells (rows) in the Raghavan et al. dataset[48], with colour bar (left) showing the sample each cell belongs to. **c**. UMAP plot of all cells (points) in the Raghavan et al. dataset[48], coloured by sample, with the same colours as in **b**.



Extended Data Fig. 2 | Cancer-type-dependent gene expression patterns. **a**. Scatter plots per cancer type showing sensitivity (y axis) and specificity (x axis) of genes (points) to malignant cells, relative to other cell types, within each cancer type. Selected genes with unusually high sensitivity or specificity are labelled. **b**. Scatter plots per cancer type showing sensitivity (y axis) and specificity (x axis) of genes (points) to malignant cells in each cancer type, relative to malignant cells in other cancer types. Selected genes with unusually high sensitivity or specificity are labelled.

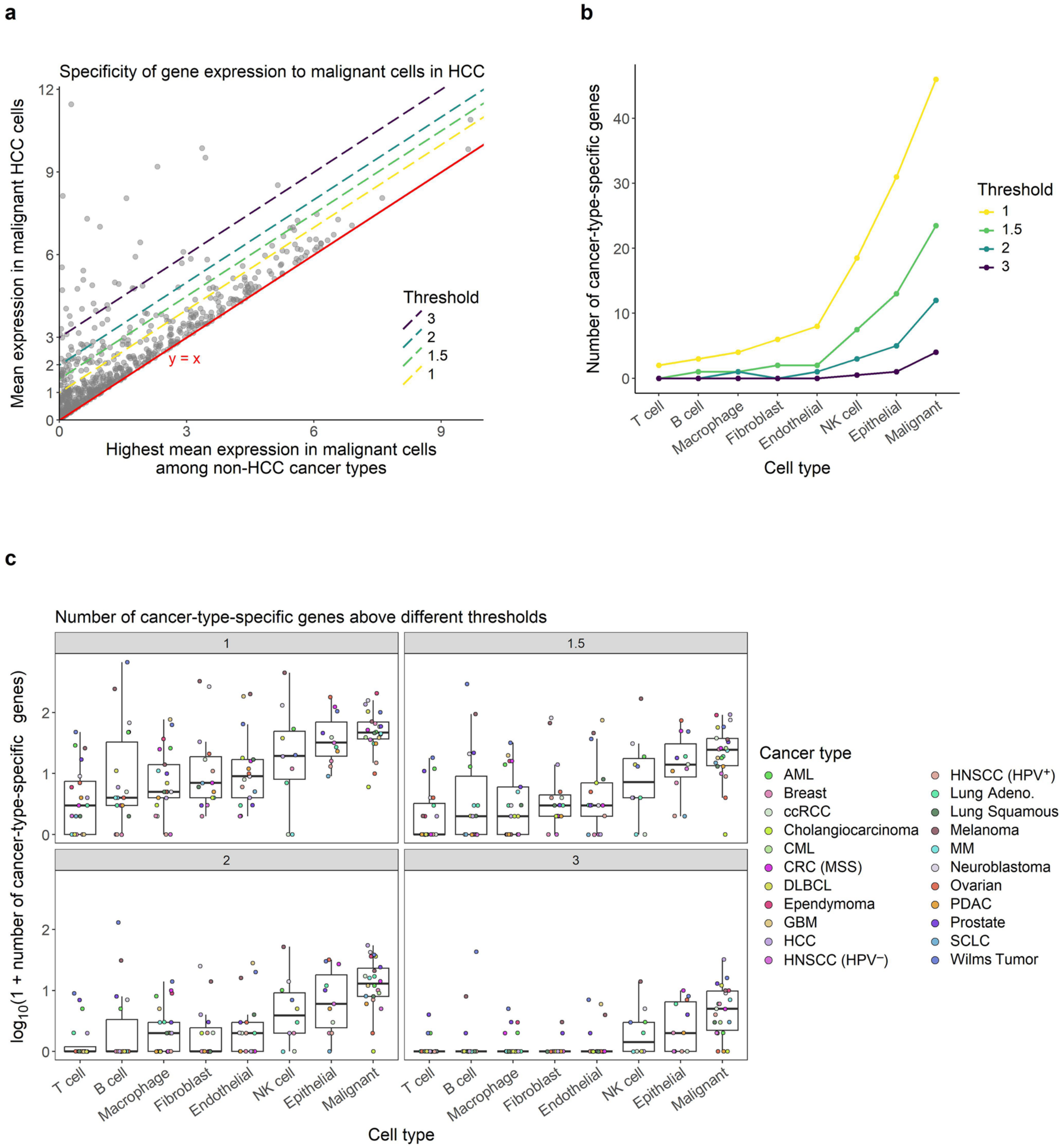


Select cell type

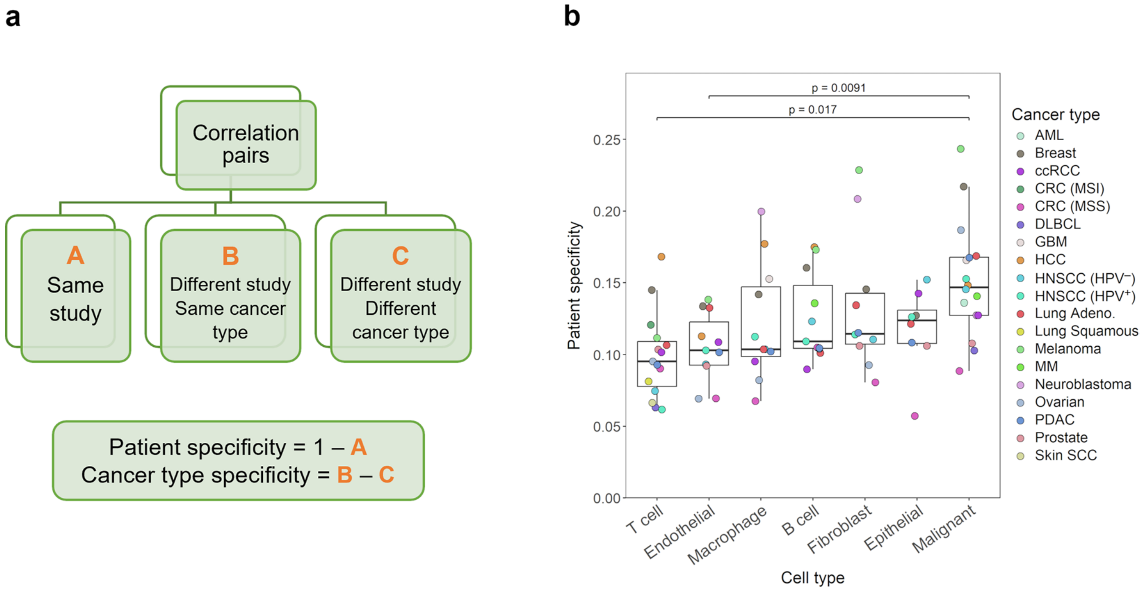
● B cell
○ Dendritic
○ Endothelial
○ Epithelial
○ Fibroblast
○ Macrophage
○ Malignant
○ Mast
○ NK cell
○ Plasma
○ T cell

Gene	Specifity	Sensitivity	Combined ▼
MS4A1	0.7342	0.4222	4.5819
BANK1	0.7154	0.2581	2.6242
CD79B	0.3738	0.3237	2.1151
CD79A	0.2075	0.5053	2.096
VPREB3	0.6629	0.1768	1.5614
CD19	0.6792	0.153	1.3185
BLK	0.728	0.1406	1.2726
CD37	0.1745	0.4825	0.9181
FCER2	0.9053	0.0873	0.9176
ADAM28	0.4678	0.1713	0.8683

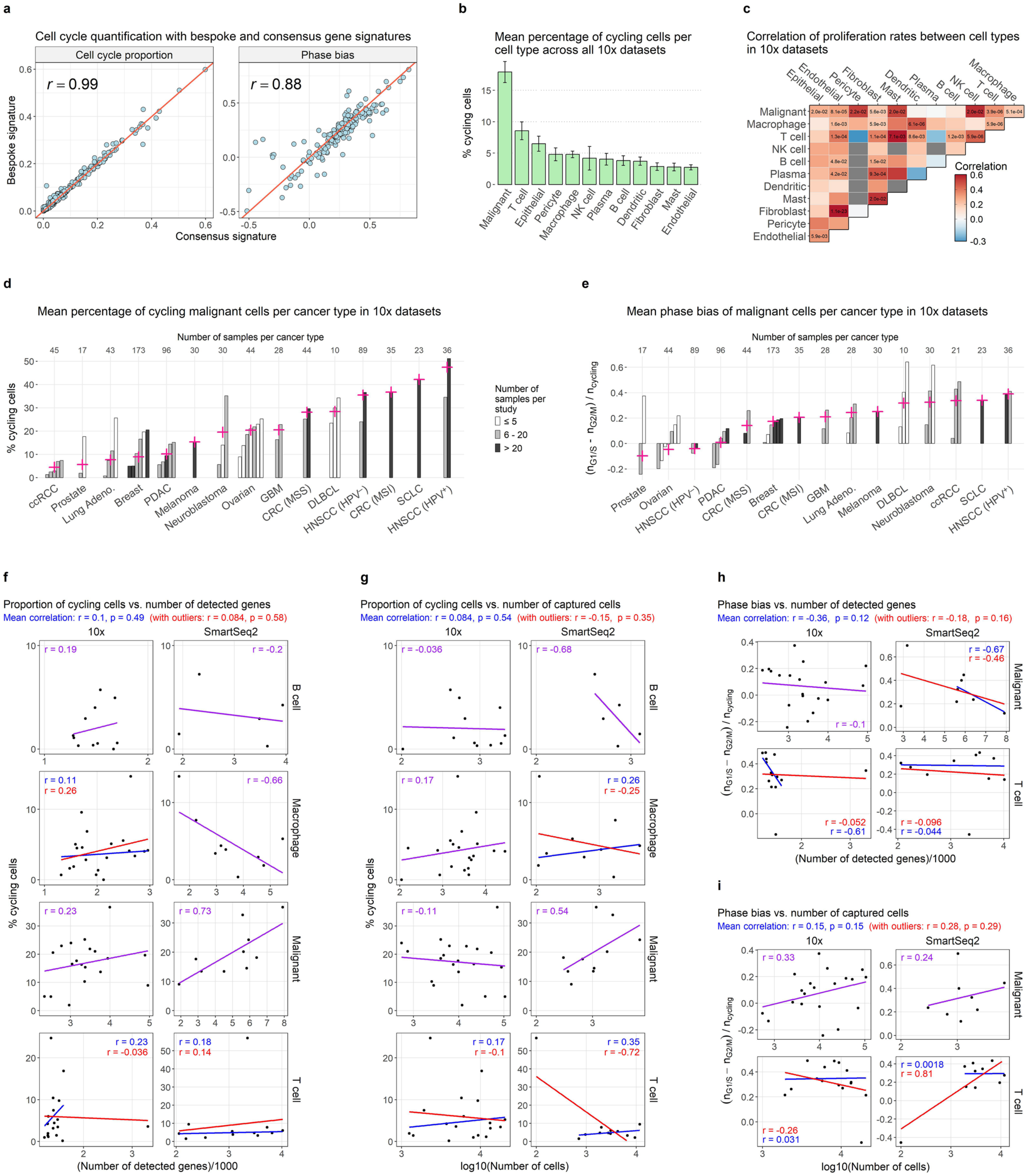
Extended Data Fig. 3 | Interactive web tool for exploring gene specificity and sensitivity in different contexts. Screenshot of 3CA web page enabling exploration of gene specificity and sensitivity values in sortable table format. Tabs distinguish analyses of global cell type markers, cancer-type-specific malignant cell markers and genes with highly variable expression between cancer types in malignant cells. In each tab, the user can select a cell type or cancer type and view the specificity and sensitivity values in a table, with the option to sort by either value or by a 'Combined' score reflecting overall marker performance. The corresponding summary scatter plot from Fig. 4b or Extended Data Fig. 2 is shown alongside.



Extended Data Fig. 4 | Number of cancer-type-specific genes per cell type. **a**. Scatter plot to illustrate the definition of cancer-type-specific gene expression at different thresholds. Points correspond to genes whose mean expression in malignant cells is highest in HCC than in all other cancer types. A point's y axis value denotes the average expression of this gene in malignant cells in HCC, while its x axis value corresponds to the maximum of its mean expression levels in malignant cells across all non-HCC cancer types. Each dashed line denotes a choice of threshold, whereby the number of genes whose expression in malignant cells is specific to HCC is defined as the number of points above this dashed line. **b**. Line plot showing the median number of cancer-type-specific genes (y axis, median across cancer types) for each cell type (x axis) for different choices of threshold (colour). Cell types are ordered by their average y values. **c**. Boxplots showing, for each choice of threshold (panels), the log-transformed number of cancer-type-specific genes (y axis) per cancer type (points/colour) for each cell type (x axis). Cell types are ordered as in **b**. Each point in **b** corresponds to the median of points for the corresponding box in **c**, after reversing the log transformation. Boxes indicate the median and 1st and 3rd quartiles, while the upper, respectively the lower whiskers extend to the maximal, resp. minimal values no further than 1.5 times the interquartile range from the 3rd, resp. 1st quartiles. Groups (cell types) consist of n = 20, 18, 21, 15, 17, 10, 11, 22 data points (in order from left to right), corresponding to differences in average expression levels across biologically distinct samples.

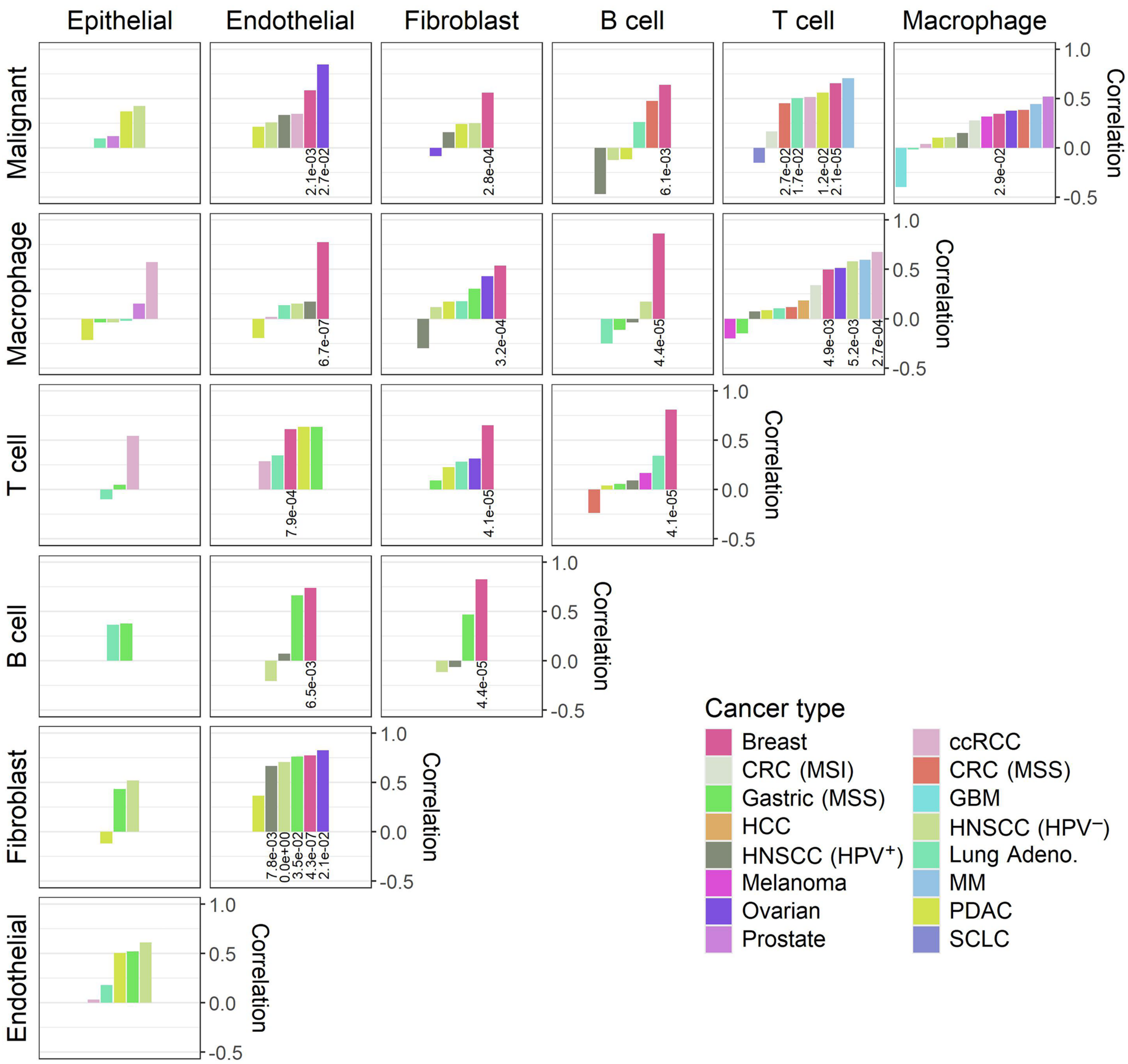


Extended Data Fig. 5 | Cancer type and patient specificity of cell type expression profiles. a. Scheme illustrating the definition of cancer type and patient specificity in terms of pairwise correlations of pseudobulk profiles. **b**. Box plot showing the patient specificity (y axis) of each cell type (x axis) in each cancer type (points/colour). Brackets indicate significant differences and are labelled with p values (0.0091 and 0.017), which were computed by pairwise paired, two-sided t tests and adjusted to FDR < 0.05 (all p values are provided in the **Source Data**). Unmarked pairwise differences are not significant. Boxes indicate the median and 1st and 3rd quartiles, while the upper, respectively the lower whiskers extend to the maximal, resp. minimal values no further than 1.5 times the interquartile range from the 3rd, resp. 1st quartiles. Groups (cell types) consist of n = 15, 11, 11, 11, 10, 8, 16 data points (in order from left to right), corresponding to averages across pairwise correlations between biologically distinct samples.

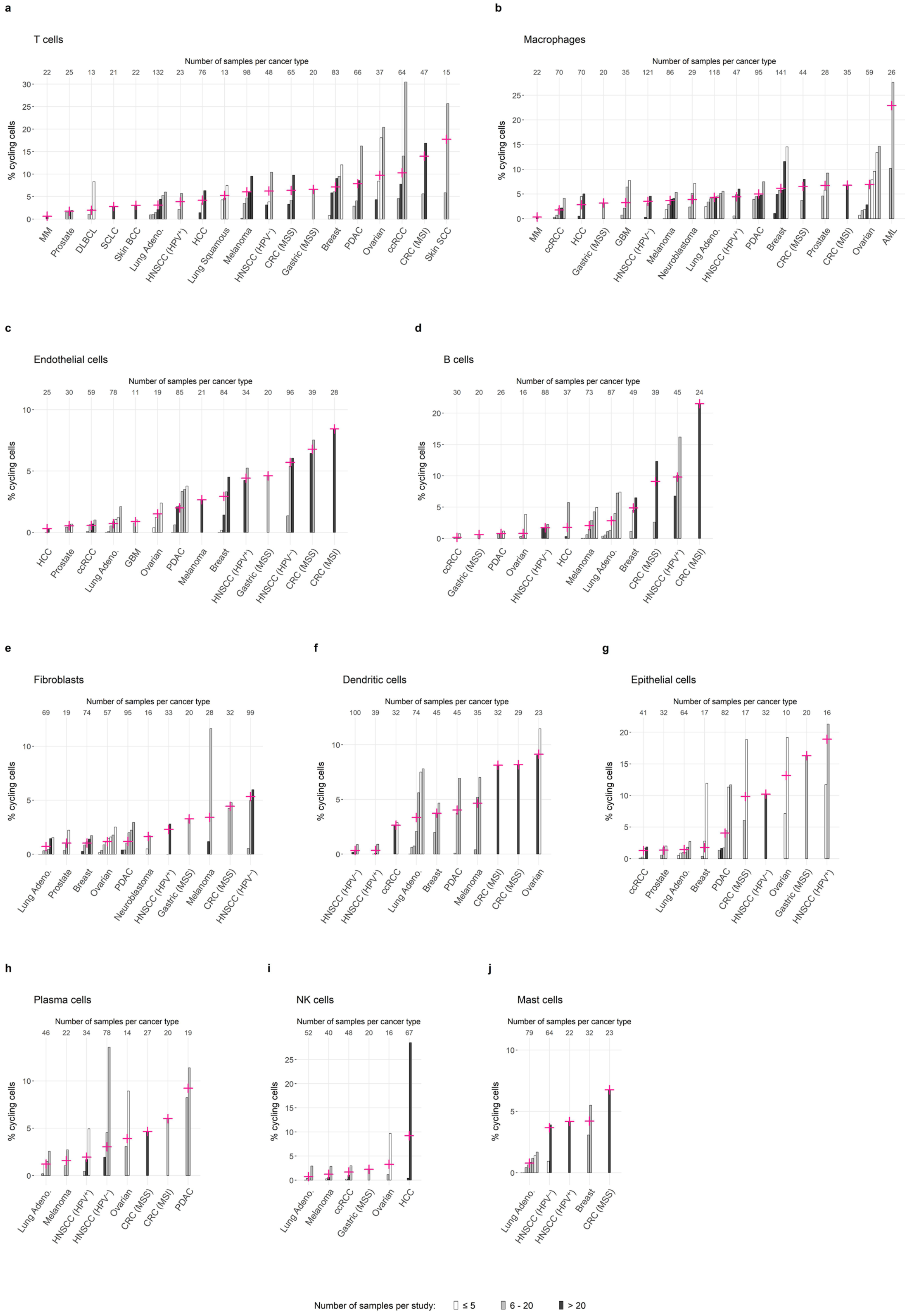


Extended Data Fig. 6 | See next page for caption.

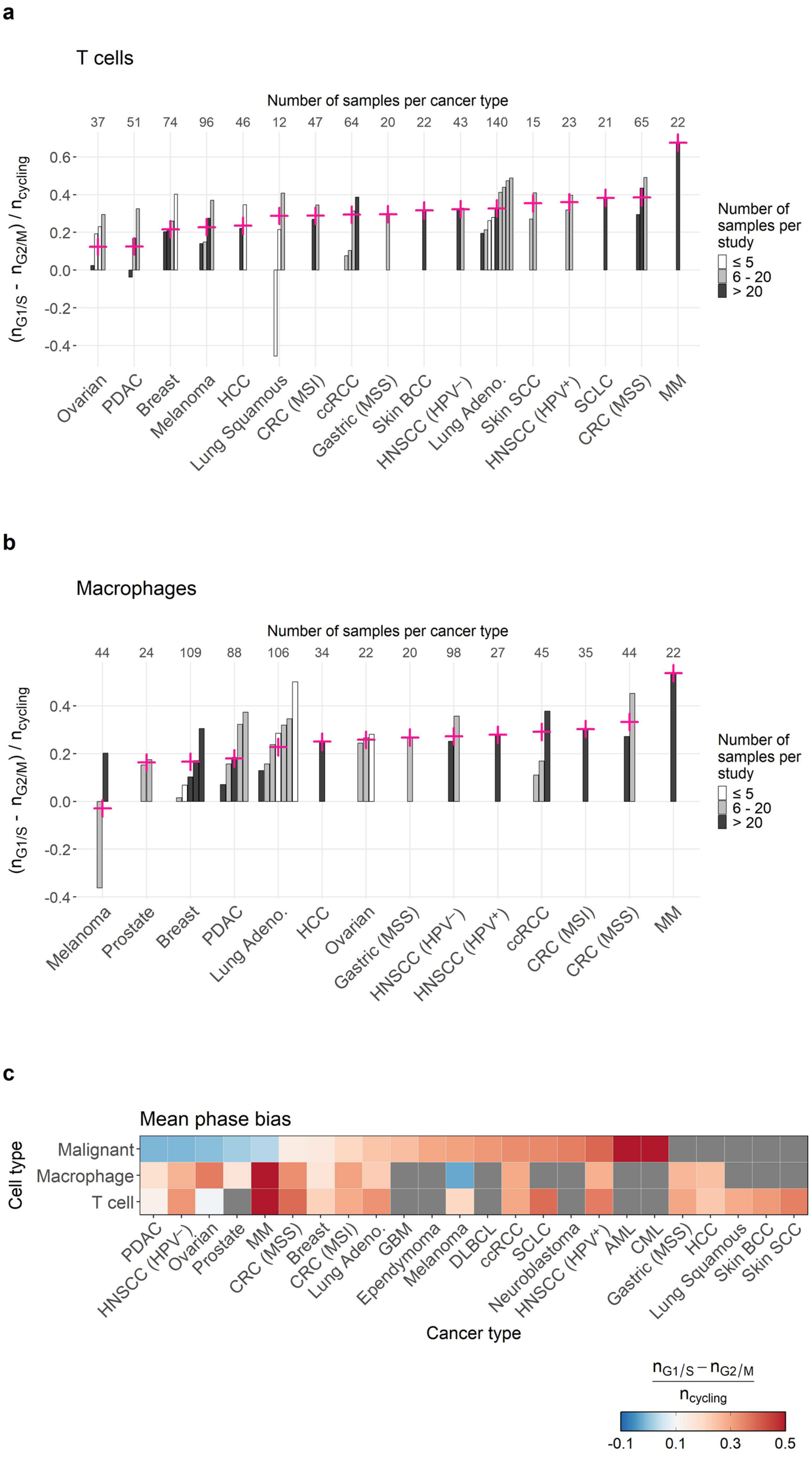
Extended Data Fig. 6 | Potential technical confounders of cell cycle estimates. **a**. Scatter plots showing cell cycle proportion and phase bias from bespoke (y axis) and consensus (x axis) G1/S and G2/M gene signatures. Each point corresponds to one cell type in one dataset. Red lines denote y = x, and r denotes Pearson correlation. **b**. Bar plot showing average percentage of cycling cells in 10x datasets (y axis) in each cell type (x axis; n = 53, 51, 33, 9, 58, 26, 17, 42, 29, 44, 19, 45, in order from left to right). Error bars denote standard error. **c**. Heatmap showing Spearman correlation between cell types (colour) of percentages of cycling cells in 10x datasets. Significant correlations are labelled with p values (2.0×10^{-2}, 8.1×10^{-5}, 2.2×10^{-2}, 5.6×10^{-3}, 2.0×10^{-2}, 2.0×10^{-2}, 3.9×10^{-6}, 5.1×10^{-4}, 1.6×10^{-3}, 5.9×10^{-3}, 6.1×10^{-6}, 5.9×10^{-6}, 1.3×10^{-4}, 1.1×10^{-4}, 7.1×10^{-3}, 8.6×10^{-3}, 1.2×10^{-3}, 5.9×10^{-6}, 4.8×10^{-2}, 1.5×10^{-2}, 4.2×10^{-2}, 9.3×10^{-4}, 2.0×10^{-2}, 1.1×10^{-23}, 5.9×10^{-3}, in order from left to right and top to bottom), which were computed by two-tailed test of zero correlation via algorithm AS 89[53], and adjusted to FDR < 0.05 (all p values are provided in the **Source Data**). **d**. Bar plot showing percentage of cycling malignant cells (y axis) in each 10x dataset (bars), grouped by cancer type (x axis). Crosses denote the average for each cancer type, weighted by number of samples containing at least 10 malignant cells. Bar colour categorises studies by number of such samples, and values above the plot denote the total number of such samples per cancer type. **e**. Bar plot showing phase bias (y axis) of malignant cells in each 10x dataset (bars), grouped by cancer type (x axis). Crosses, bar colour and number of samples per cancer type are as in **d**. **f**. Scatter plots showing, for each cell type and sequencing platform, percentage of cycling cells (y axis) against number of detected genes (x axis) in each dataset (points). Regression lines and Pearson correlation were computed with and without outliers (red and blue respectively, or purple in cases with no outliers). Average correlations and p values (computed by two-sided t test, without adjustment) are shown at the top. **g-i**. Scatter plots as in **a** showing: percentage of cycling cells against number of captured cells; phase bias against number of detected genes; phase bias against number of captured cells.



Extended Data Fig. 7 | Correlation of cell cycle between cell types, per cancer type. Bar plots, for each pair of cell types, showing the Spearman correlation (across samples) of proportion of cycling cells between those cell types in each cancer type. Significant correlations are labelled with p values (2.1×10^{-3}, 2.7×10^{-2}, 2.8×10^{-4}, 6.1×10^{-3}, 2.7×10^{-2}, 1.7×10^{-2}, 1.2×10^{-2}, 2.1×10^{-5}, 2.9×10^{-2}, 6.9×10^{-7}, 3.2×10^{-4}, 4.4×10^{-5}, 4.9×10^{-3}, 5.2×10^{-3}, 2.7×10^{-4}, 7.9×10^{-4}, 4.1×10^{-5}, 4.1×10^{-5}, 6.5×10^{-3}, 4.4×10^{-5}, 7.8×10^{-3}, 0, 3.5×10^{-2}, 4.3×10^{-7}, 2.1×10^{-2}, in order from left to right and top to bottom), which were computed by two-tailed test of zero correlation via algorithm AS 89[53] and adjusted to FDR < 0.05 (all p values are provided in the **Source Data**).



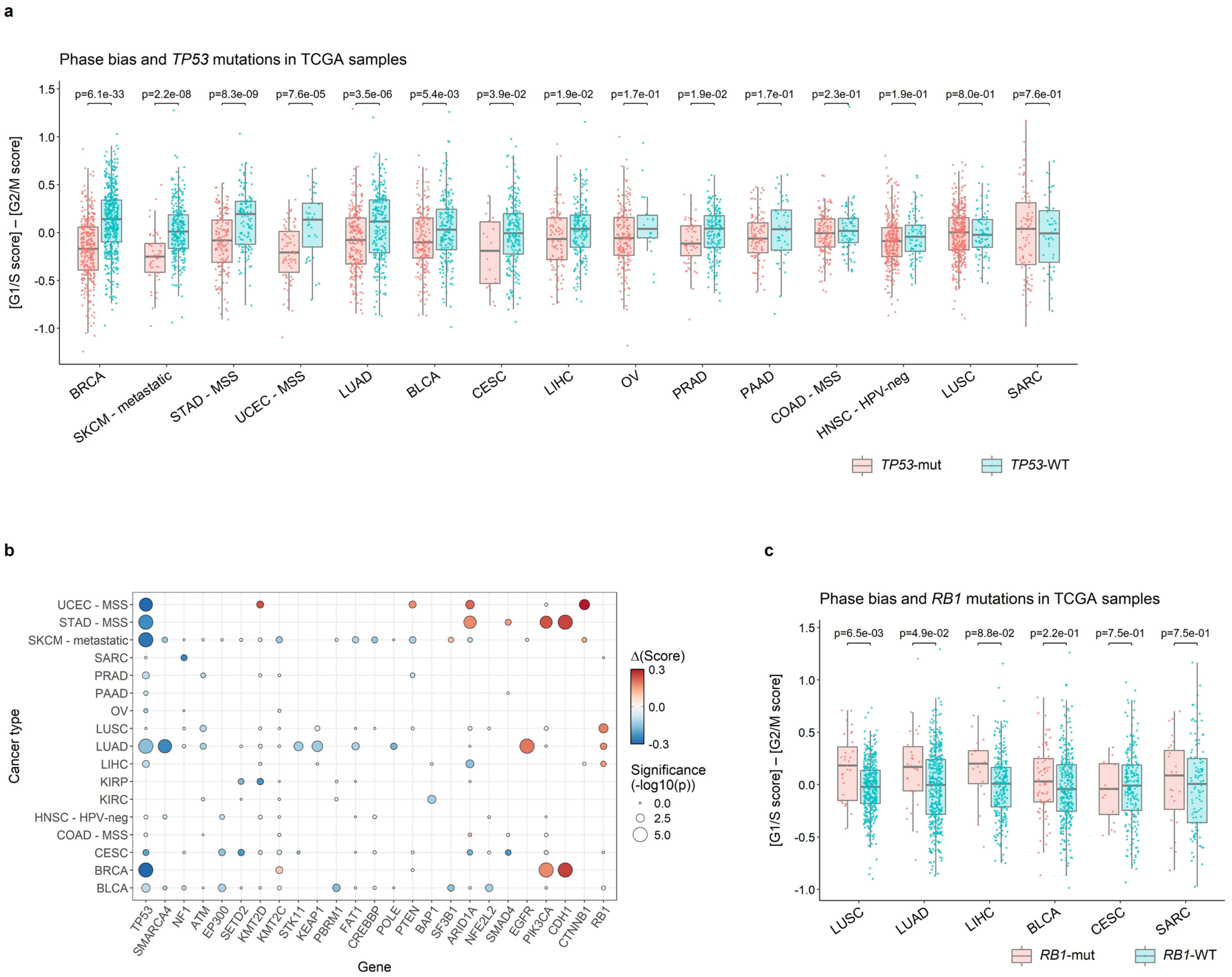
Extended Data Fig. 8 | Proliferation rates across cancer types for non-malignant cell types. a-j. Bar plots, for each of the most common non-malignant cell types, showing the percentage of cycling cells of that type (y axis) in each study (bars), grouped by cancer type (x axis), with crosses denoting the average y value for each cancer type, weighted by the number of samples in each study which contain at least 10 cells of that type. Bar colour categorises studies by number of such samples, and values above the plot denote the total number of such samples in each cancer type.



Extended Data Fig. 9 | See next page for caption.

https://doi.org/10.1038/s43018-025-00957-8

Extended Data Fig. 9 | Phase bias patterns across cancer types. a. Bar plot showing the phase bias (y axis, quantified by the relative fraction of cycling cells in G1/S versus G2/M) of T cells in each study (bars), grouped by cancer type (x axis), with crosses denoting the average y value for each cancer type, weighted by the number of samples in each study which contain at least 10 T cells. Bar colour categorises studies by number of such samples, and values above the plot denote the total number of such samples in each cancer type. Low and high y values indicate bias toward G2/M and G1/S, respectively. **b**. Bar plot as in **a** for macrophages. **c**. Heatmap showing the weighted average of the phase bias (colour, defined as for the crosses in **a**) per cancer type (x axis) and cell type (y axis). Grey squares indicate insufficient data.



Extended Data Fig. 10 | Genomic associations of phase bias. a. Box plot showing the phase bias scores (y axis, defined as the difference between scores for G1/S and G2/M gene signatures) of TCGA tumour samples (points; n = 334, 530, 52, 284, 161, 127, 73, 37, 241, 238, 191, 197, 20, 220, 101, 189, 189, 18, 56, 177, 106, 45, 139, 80, 309, 85, 385, 87, 86, 47, in order from left to right), grouped by cancer type (x axis) and coloured by TP53 mutation status. Brackets are labelled with p values (6.1×10^{-33}, 2.2×10^{-8}, 8.3×10^{-9}, 7.6×10^{-5}, 3.5×10^{-6}, 5.4×10^{-3}, 0.039, 0.019, 0.17, 0.019, 0.17, 0.23, 0.19, 0.80, 0.76, in order from left to right), which were computed by two-sided t test and adjusted to FDR < 0.05. Low and high y values indicate bias toward G2/M and G1/S, respectively. Boxes indicate the median and 1st and 3rd quartiles, while the upper, respectively the lower whiskers extend to the maximal, resp. minimal values no further than 1.5 times the interquartile range from the 3rd, resp. 1st quartiles. **b.** Dot plot showing, for a selection of genes commonly mutated in cancer (x axis), the difference in average phase bias score between mutant and wild-type tumours (point colour; phase bias score defined as in **a**) and the statistical significance of this difference (point size, computed as in **a**, before adjustment) in each cancer type (y axis). **c.** Box plot as in **a** for RB1 mutations (n = 33, 439, 30, 449, 19, 271, 73, 315, 16, 224, 27, 106; p = 0.0065, 0.049, 0.088, 0.22, 0.75, 0.75; both in order from left to right).

nature portfolio

Corresponding author(s): Michael Tyler, Itay Tirosh

Last updated by author(s): Jan 17, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐ ☒ A description of all covariates tested

☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect the data used in this study.
Data analysis	All analysis and data processing in this study used custom code written with R version 4.1.1. This code is available at https://github.com/tiroshlab/3ca. Individual R package versions are as follows: caTools: 1.18.2 colorRamps: 2.3.1 cowplot: 1.1.1 data.table: 1.14.2 dplyr: 1.0.8 easyPubMed: 2.13 ggplot2: 3.4.4 ggpubr: 0.4.0 ggrepel: 0.9.1 ggtext: 0.1.1 grid: 4.1.1 gridExtra: 2.3 gtable: 0.3.0 irlba: 2.3.5 knitr: 1.38 magrittr: 2.0.3 matkot: 0.0.0.9000

Matrix: 1.4.1
NMF: 0.24.0
plyr: 1.8.7
randomcoloR: 1.1.0.1
RColorBrewer: 1.1.3
readxl: 1.4.0
reshape2: 1.4.4
rmarkdown: 2.13
scales: 1.2.1
seriation: 1.3.5
stringr: 1.4.0
uwot: 0.1.11
viridis: 0.6.4

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

This study used only external datasets and did not involve the generation of new data. All published single-cell datasets are available on the 3CA website (https://www.weizmann.ac.il/sites/3CA/), with the exception of one dataset, for which permission for sharing through 3CA was not granted, and which is available through EGA with accession number EGAS00001002543. Additional unpublished datasets used will be added to the 3CA website when possible. TCGA data was downloaded from http://gdac.broadinstitute.org/.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	This study used only external datasets. When available, patient sex annotations were collected from these datasets and are included with the data available for download on the 3CA website. No analysis involving sex or gender was performed in this study, and these variables were not considered in study design.
Reporting on race, ethnicity, or other socially relevant groupings	This study used only external datasets. When available, patient race or ethnicity annotations were collected from these datasets and are included with the data available for download on the 3CA website. No analysis involving race or ethnicity was performed in this study, and these variables were not considered in study design.
Population characteristics	Not applicable, as this study used only external datasets.
Recruitment	Not applicable, as this study used only external datasets.
Ethics oversight	Not applicable, as this study used only external datasets.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This study used only external datasets. Our analysis used all available samples, and we did not perform any analysis to predetermine sample size.
Data exclusions	We excluded data from specific analyses based on insufficient number of samples or cells. Two datasets were excluded from several analyses due to high tehcnical noise.

Replication

All analyses may be replicated using the published code.

Randomization

This study used only external datasets. Our analysis used all available samples, and we did not choose samples/patients or groups ourselves.

Blinding

This study used only external datasets. Our analysis used all available samples, and we did not choose samples/patients or groups ourselves.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern
☒	□ Plants

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.