Article

Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours

https://doi.org/10.1038/s41586-023-06130-4

Received: 5 May 2022

Accepted: 25 April 2023

Published online: 31 May 2023

Check for updates

Avishai Gavish[1], Michael Tyler[1], Alissa C. Greenwald[1], Rouven Hoefflin[1,2], Dor Simkin[1], Roi Tschernichovsky[1,3], Noam Galili Darnell[1], Einav Somech[1], Chaya Barbolin[1], Tomer Antman[1], Daniel Kovarsky[1], Thomas Barrett[4,5], L. Nicolas Gonzalez Castro[6,7,8,9], Debdatta Halder[1], Rony Chanoch-Myers[1], Julie Laffy[1], Michael Mints[1,10], Adi Wider[1], Rotem Tal[1], Avishay Spitzer[1], Toshiro Hara[6,7], Maria Raitses-Gurevich[11], Chani Stossel[11,12], Talia Golan[11,12], Amit Tirosh[12,13], Mario L. Suvà[6,7], Sidharth V. Puram[4,14] & Itay Tirosh[1] ✉

Each tumour contains diverse cellular states that underlie intratumour heterogeneity (ITH), a central challenge of cancer therapeutics[1]. Dozens of recent studies have begun to describe ITH by single-cell RNA sequencing, but each study typically profiled only a small number of tumours and provided a narrow view of transcriptional ITH[2]. Here we curate, annotate and integrate the data from 77 different studies to reveal the patterns of transcriptional ITH across 1,163 tumour samples covering 24 tumour types. Among the malignant cells, we identify 41 consensus meta-programs, each consisting of dozens of genes that are coordinately upregulated in subpopulations of cells within many tumours. The meta-programs cover diverse cellular processes including both generic (for example, cell cycle and stress) and lineage-specific patterns that we map into 11 hallmarks of transcriptional ITH. Most meta-programs of carcinoma cells are similar to those identified in non-malignant epithelial cells, suggesting that a large fraction of malignant ITH programs are variable even before oncogenesis, reflecting the biology of their cell of origin. We further extended the meta-program analysis to six common non-malignant cell types and utilize these to map cell–cell interactions within the tumour microenvironment. In summary, we have assembled a comprehensive pan-cancer single-cell RNA-sequencing dataset, which is available through the Curated Cancer Cell Atlas website, and leveraged this dataset to carry out a systematic characterization of transcriptional ITH.

ITH is a fundamental property of tumours that is driven by genetics, epigenetics and microenvironmental influences, and is central to treatment failure, metastasis and other cancer phenotypes[1]. Single-cell RNA sequencing (scRNA-seq) efficiently enables the characterization of ITH, and has seen a rapid expansion of its use across virtually all common cancer types[2]. One emerging concept from previous scRNA-seq studies is the existence of ITH 'expression programs', consisting of sets of dozens of genes with coordinated variability in their expression across malignant cells within a given tumour. In melanoma, a skin-pigmentation program driven by *MITF* and an epithelial–mesenchymal transition (EMT)-like program associated with *AXL* varied within individual tumours and had important functional consequences[3,4]. In glioblastoma, four expression programs were identified as a central source of transcriptional heterogeneity[5]. In head and neck squamous cell carcinoma (HNSCC), EMT-like and epithelial senescence (EpiSen) programs were identified and shown to affect the likelihood of metastasis and drug responses[6,7]. A stress-response program was found in several cancer types with important functional implications[8].

Importantly, similar ITH programs are identified across tumours of the same cancer type, and in some cases even across different cancer types[7,8]. These similarities suggest that ITH expression programs reflect fundamental aspects of tumour biology. We therefore seek to identify the consensus among related ITH programs from different tumours, which we denote as meta-programs (MPs).

High expression of any particular MP may be considered as defining a cellular state. However, it is important to note that MPs tend to be limited to dozens of genes whose expression is superimposed on the cells' baseline expression profiles and therefore reflect a relative cellular state. For instance, two subpopulations of cells from two distinct tumours may upregulate the same MP (for example, of cell-cycle genes) while retaining the extensive expression differences between these two tumours (for example, due to unique driver mutations). These two

[1]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [2]Department of Medicine I, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany. [3]Davidoff Cancer Center, Rabin Medical Center, Petah Tikva, Israel. [4]Department of Otolaryngology-Head and Neck Surgery, Washington University School of Medicine, St Louis, MO, USA. [5]Department of Genetics, Washington University School of Medicine, St Louis, MO, USA. [6]Department of Pathology and Center for Cancer Research, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [7]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [8]Center for Neuro-Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [9]Department of Neurology, Brigham and Women's Hospital, Boston, MA, USA. [10]Department of Oncology-Pathology, Karolinska Institute, Stockholm, Sweden. [11]The Oncology Institute, Chaim Sheba Medical Center, Ramat Gan, Israel. [12]The Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel. [13]Division of Endocrinology, Diabetes and Metabolism, Chaim Sheba Medical Center, Ramat Gan, Israel. [14]Siteman Cancer Center, Washington University School of Medicine, St Louis, MO, USA. ✉e-mail: itay.tirosh@weizmann.ac.il

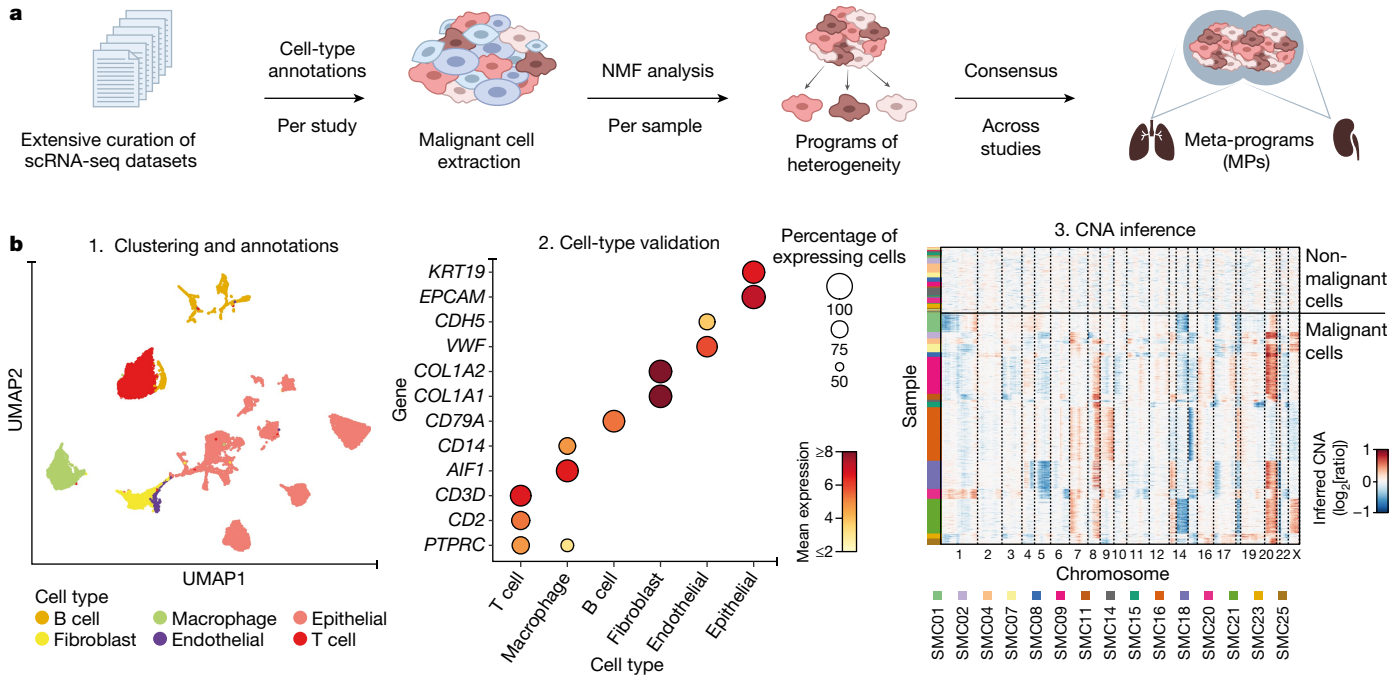


Fig. 1 | Defining MPs from a curated and annotated collection of scRNA-seq datasets. a, Workflow. **b**, Cell annotations of the Lee et al. 2020 dataset (ref. 33) by three steps, exemplifying the approach for annotation of all datasets. 1. UMAP plot of all cells coloured by their cell-type assignment as determined by the original study. 2. Validation of the cell-type assignments based on expression of canonical marker genes. 3. Inference of CNAs from gene expression profiles separates malignant cells with CNAs (bottom) from non-malignant cells without CNAs (top). Panel **a** designed by T. Bigdary.

subpopulations would be in a different global cellular state (reflecting the tumour identity) but in the same relative cellular state (reflecting the activation of a particular MP). In this work we focus primarily on relative cellular states by defining the expression programs (and MPs) of each subpopulation of cells relative to the other cells from the same tumour, hence highlighting the patterns of intratumour rather than intertumour heterogeneity.

The functional and clinical significance of MPs identified previously and the emerging view that MPs explain a large fraction of expression ITH raises the need to comprehensively define the MPs in cancer and understand their functions. We previously profiled 198 cancer cell lines by scRNA-seq and uncovered 12 in vitro MPs[7]. Another recent study analysed scRNA-seq data for 62 primary tumours from several cancer types and identified 16 MPs[9]. Here we aim to markedly expand this analysis by integrating scRNA-seq datasets across 77 studies that profiled more than a thousand patient samples from diverse cancer types. We define MPs in malignant and in non-malignant cells and investigate their functions, context specificity and interactions.

Curation of cancer scRNA-seq datasets

To systematically define cancer MPs (Fig. 1a and Methods), we searched for and prioritized all studies that reported scRNA-seq data for human tumours. We added unpublished datasets on neuroendocrine tumours, head and neck cancer and schwannoma. Finally, we incorporated selected datasets for mouse models or cell models. Altogether we obtained data from 77 studies, encompassing 1,456 samples covering 24 cancer types and 2,591,545 cells (Supplementary Table 1).

We used two complementary approaches to annotate cells from each dataset. First, we assigned cells to 38 distinct cell types, while ensuring expression of canonical cell-type markers, and excluding dubious clusters and apparent doublets (Fig. 1b, steps 1 and 2, and Supplementary Table 1). Second, we inferred copy-number alterations (CNAs) from the gene expression profiles, and assigned cells as malignant or non-malignant[3] (Fig. 1b, step 3). Notably, 67% of carcinoma samples appear to contain non-malignant epithelial cells, such that an epithelial assignment is not sufficient to define cells as malignant. Cells with borderline CNA signals were excluded from further analysis as they probably reflect doublets or low-quality data. Overall, we defined 686,690 high-confidence malignant cells and 1,199,312 non-malignant cells.

Owing to the importance of a large and consistently annotated compendium, we provide all data at the Curated Cancer Cell Atlas (3CA; https://weizmann.ac.il/sites/3CA). 3CA provides the original datasets, the cell annotations, inferred CNAs, uniform manifold approximation and projection (UMAP) plots, associated statistics and other advanced analyses described below and in upcoming publications. We aim to continuously expand 3CA with new datasets and functionalities.

Defining and annotating MPs

Next, we utilized 3CA to comprehensively characterize ITH among malignant cells (Fig. 1a). The diverse methodologies of the curated studies pose a challenge for analysing them collectively, and although computational methods can improve data integration and reduce batch effects, they cannot fully distinguish between technical and biological variability. Importantly, however, as our primary interest is in variability within individual tumours (rather than between tumours), direct integration of datasets is not required for our analyses. Thus, instead of integrating datasets, we defined expression programs that vary within each tumour and subsequently compared the resulting ITH programs (sets of correlated genes) across tumours from all studies. Although primary expression data suffer from prominent batch effects, the ITH programs defined from comparisons within the same batch are less sensitive to batch effects and thereby recover high similarities across studies (Supplementary Fig. 1).

For each tumour, we utilized non-negative matrix factorization (NMF) to characterize the ITH programs that vary among its malignant cells, each summarized by its top-scoring 50 genes. NMF was applied

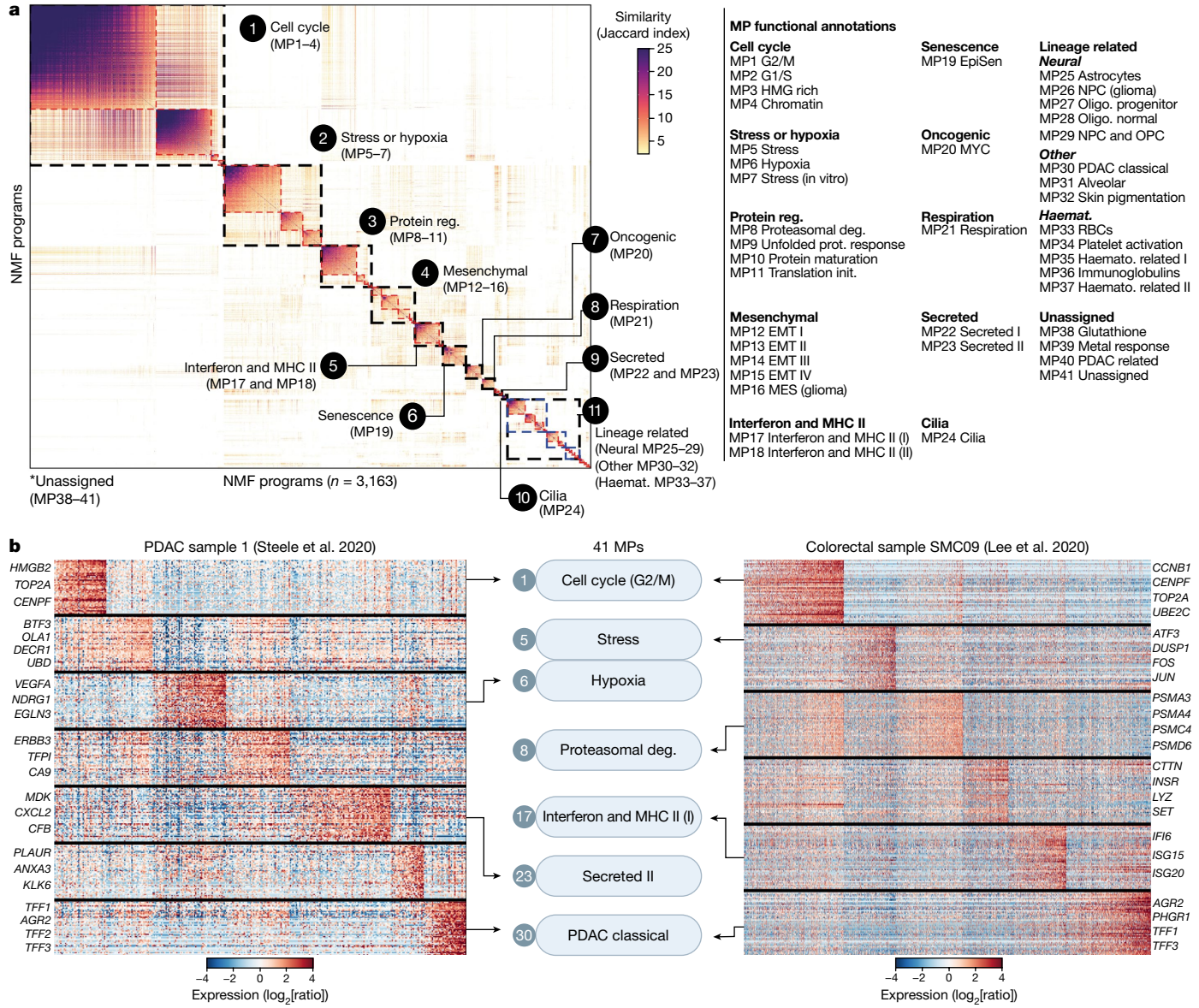


Fig. 2 | MPs and their functional annotations. a, Left: heatmap showing Jaccard similarity indices for comparisons among 3,163 robust NMF programs based on their top 50 genes. Programs are ordered by clustering and grouped into MPs (marked by red dashed lines) and families of MPs with related functions (marked by black dashed lines); MP families are numbered and labelled. The lineage related family was divided into three subgroups (marked by blue dashed lines). Right: list of all MP names, separated into MP families. **b**, Robust NMF programs in two tumours—SMC09 from Lee et al. 2020 (ref. 33) (right) and sample 1 from Steele et al. 2020 (ref. 34) (left). In each case, a heatmap shows the relative expression of the top 50 genes from each robust NMF program (rows), across all malignant cells from the tumour (columns). Genes are arranged by NMF programs and selected genes are labelled. The middle section shows the association of 7 of the 41 NMF programs with MPs and the associated MP names. Haemato. or haemat., haematological; RBCs, red blood cells; NPC, neural progenitor cells; OPCs, oligodendrocyte-progenitor cells; reg., regulation; deg., degradation; prot., protein; init., initiation; oligo., oligodendrocytes.

with multiple parameter values (K = [4–9]) and the programs that were consistently identified in a tumour were denoted as robust (Methods). Overall, we identified 5,547 robust malignant NMF programs (Supplementary Table 2).

To identify recurrent patterns of ITH, we clustered the robust NMF programs by their fractions of shared top genes (Supplementary Fig. 1), and filtered out programs that were associated with low quality or doublets. This resulted in 41 clusters of programs (Fig. 2 and Supplementary Table 2). All clusters were derived from several studies and 83% of them were derived from several cancer types. The clusters covered 66% of all robust NMF programs, indicating that most expression ITH reflects recurrent patterns that may be described by MPs. Thus, for each cluster, an MP was defined as the set of genes most commonly shared between programs from that cluster (Methods). Of all malignant cells, 54% were significantly enriched for at least one MP. Similar results were obtained with an alternative computational approach (Methods and Supplementary Fig. 2). MPs were detected comparably across scRNA-seq platforms, and included both highly expressed and non-highly expressed genes (Supplementary Fig. 3).

MPs were annotated on the basis of their functional enrichments (Extended Data Fig. 1 and Supplementary Table 3), and MPs with related functions were manually grouped into 11 MP families (Fig. 2a). The 41 MPs and their context specificity are described briefly below and in detail in Supplementary Note 1. Nine of the MPs were similar to those seen in vitro[7], and 16 were similar to those from a smaller tumour cohort[9], whereas 21 were new (Supplementary Table 4). Note that the

identification of an MP in a certain tumour does not imply overall high MP expression but rather that malignant cells with high and those with low MP expression coexist within that tumour.

The most broadly identified MP families were associated with cell cycle (MP1–4), stress or hypoxia (MP5–7), and mesenchymal (MES) or EMT-like states (MP12–16). These MPs relate to core cellular processes and to metabolism (Supplementary Table 5). Several MPs within each of these families highlight variation in the associated processes. For example, apart from the canonical G2/M and G1/S cell-cycle MPs (MP1 and MP2, respectively), two less frequent MPs consisted mostly of cell-cycle-related genes but were specifically enriched with genes encoding HMG-box proteins (MP3) or chromatin regulators (MP4). Among mesenchymal or EMT-like MPs, variants differ in their cancer type specificity, and a 'hybrid' program (MP14) includes both mesenchymal and epithelial markers.

MPs with intermediate frequency mostly resembled known ITH patterns (protein regulation, interferon response, EpiSen or cilia) but also included MPs that to our knowledge were not described previously as heterogeneous in tumour scRNA-seq datasets, including MYC targets (MP20). MPs with low frequencies (<1% of NMF programs) were primarily not described previously, to our knowledge, highlighting the increased sensitivity in detecting recurrent ITH programs with a large and diverse compendium. Many low-frequency MPs were enriched with functional annotations linked to a specific tissue or lineage (for example, brain or blood related), which we grouped together in an MP family denoted as lineage-related.

Regulation of MPs

To uncover MP regulators, we applied SCENIC to all tumour samples. SCENIC integrates scRNA-seq data with known protein–DNA interactions to infer the regulons of transcription regulators[10]. We then searched for inferred regulons that correlate with MP expression across many samples (Extended Data Fig. 2 and Supplementary Table 6). This approach identified expected regulators (for example, E2F transcription factors as regulators of cell-cycle MPs), and many putative regulatory interactions, such as GRHL1 regulating EpiSen (MP19).

Next, we explored the genetic regulation of MPs both within and across tumours. Within tumours, we investigated how often MP expression varies between genetically distinct subclones. We identified genetic subclones from inferred CNA profiles within 16% of the tumours, typically with 2–3 subclones per tumour. Only 24% of the subclones were associated with significantly high or low expression of at least one MP, suggesting that MPs do not primarily reflect genetic subclones and often reflect non-genetic plasticity (Extended Data Fig. 3).

Across tumours, we investigated whether MP expression is associated with particular mutations or CNAs. As only few tumours were profiled by both scRNA-seq and whole-exome or whole-genome sequencing, we turned to bulk data from The Cancer Genome Atlas (TCGA), thereby extending the analysis to thousands of genetically annotated tumours, but limiting the analysis to a tumour's average MP expression. We found many associations between MP expression and specific mutations or CNAs (Extended Data Fig. 4 and Supplementary Table 7). The strongest associations include high expression of cell-cycle MPs in tumours with *RB1* and *TP53* mutations. Another strong association was found in kidney clear cell carcinoma (KIRC) between the glutathione MP and several genetic features including lack of *VHL* and *PBRM1* mutations.

Several genetic associations were observed for the interferon and major histocompatibility complex (MHC) II MP, including *STK11* and *KEAP1* mutations in lung adenocarcinoma (LUAD) and *CASP8* mutations in HNSCC (Extended Data Fig. 4). *CASP8* is involved in apoptosis, and its expression is induced by interferon[11]. Thus, inactivating *CASP8* mutations may enable cancer cells to survive in an interferon-high environment by inducing an interferon response program but without activating apoptosis.

MP clinical associations

To explore the significance of MPs, we first examined their association with proliferation. In each tumour, we calculated the correlation between proliferation scores of cells (maximal expression of cell-cycle MPs) and expression of other MPs. The average correlations across tumours define the overall propensity of cells in each state to proliferate (Supplementary Fig. 4). Most MPs are slightly negatively correlated with proliferation, suggesting that cycling cells may repress other programs and divert resources to proliferation. The most significant negative correlation ($P = 9.8 \times 10^{-32}$) was found for MP19, consistent with its annotation as EpiSen.

Conversely, MYC targets (MP20), proteasomal degradation (MP8) and respiration (MP21) were positively correlated with proliferation ($P = 1.8 \times 10^{-23}$, 8.8×10^{-14} and 7.8×10^{-11}, respectively). The association of respiration with proliferation suggests that despite the tendency of many cancers to rely primarily on glycolysis[12], subsets of cells with increased respiration tend to have increased proliferation capacity.

Next, we examined the association of MPs with clinical features. Given the limited clinical annotation of scRNA-seq datasets, we again turned to analysis of (average) MP expression in bulk TCGA samples, and identified associations with overall survival, grade and stage, lymph node metastasis and therapy resistance (Extended Data Fig. 5a–c and Supplementary Table 8). Some associations are context-specific, such as the associations of the alveolar MP with higher overall survival and with decreased therapy resistance in LUAD, and the associations of the glutathione MP with higher overall survival and lower proliferation in KIRC (Extended Data Fig. 4k,l). Other associations are more consistent across cancer types, such as cell cycle, hypoxia and proteasomal degradation correlating with worse outcomes (Extended Data Fig. 5d).

MP context specificity

We classified the frequency of each MP in each cancer type as absent, low, medium, high or high and significantly enriched (Fig. 3a and Methods). Seven MPs were found at medium or high frequency in most cancer types and are denoted as general (Fig. 3b). These include MPs of cell cycle, stress, hypoxia, interferon responses, EMT III and MYC targets. By contrast, 13 MPs found only in one or two cancer types were denoted as context-specific. The remaining MPs (21 out of 41) were detected in 3 to 12 cancer types and denoted as shared.

To validate MPs and their context specificity, we examined spatial transcriptomics (Visium) data for ovarian cancer, skin squamous cell carcinoma and glioblastoma[13–15]. We detected 28 of the 29 MPs defined in these cancer types by scRNA-seq, and their context specificity was highly consistent between Visium and scRNA-seq (Supplementary Fig. 5).

Several MPs had unexpected context specificity. For example, MP38 and MP39, enriched with glutathione and metal-response genes, respectively, were identified as variable only within KIRC. Expression of these MPs is not unique to KIRC and is even higher in other cancer types, but their variability is seen only in KIRC, highlighting the distinction between specificity of an expression program and the specificity of its intratumour variability (see Supplementary Fig. 6 for a systematic analysis).

Other MPs were identified as variable more broadly than expected. EpiSen was previously described in HNSCC[6,7], but here we find it in subpopulations of cells from 131 tumours across 9 cancer types and validated it in 4 cancer types (Fig. 3c, Extended Data Fig. 6a and Supplementary Fig. 7). As a second example, MP30 resembles signatures of the 'classical' subtype of pancreatic ductal adenocarcinoma (PDAC)[16,17], but apart from PDAC tumours, it was identified as variable within lung, colorectal, liver, and head and neck cancers (Figs. 2b and 3a).

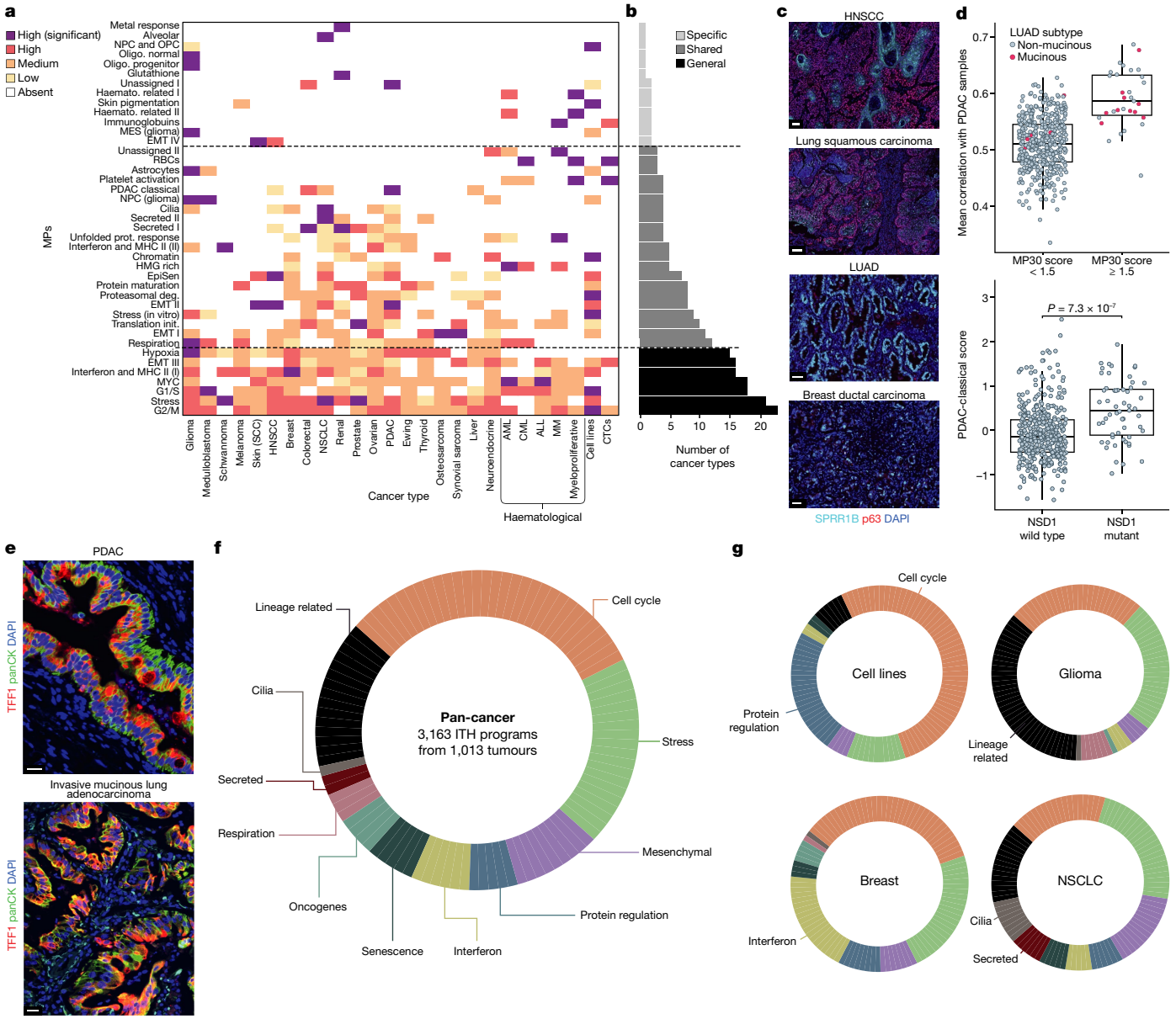


Fig. 3 | MP context specificity. a, Abundance of each MP (rows) in each cancer type (columns), defined as absent, low, medium, high or high and significant (Methods). **b**, Bar plot showing, for each MP, the number of cancer types with a medium or higher abundance. Rows correspond to the MPs as labelled in **a**. MPs are further divided into three abundance categories by the dashed black lines. **c**, Validation of EpiSen marker expression by IHC in four cancer types (representative image from three independent experiments per cancer type). Scale bars, 50 µm (first, third and fourth images) and 100 µm (second image). **d**, Top: average correlation of each LUAD TCGA sample with all PDAC TCGA samples. LUAD samples are divided into those with or without high expression of the PDAC-classical MP (score > 1.5) and are coloured on the basis of their histological classification as mucinous or as non-mucinous LUADs. Bottom: average PDAC-classical scores in NSD1-mutant and NSD1-wild-type TCGA HNSCC samples. **e**, Validation of PDAC-classical marker expression by IHC in lung invasive mucinous tissues (representative image from four independent experiments per cancer type). Scale bars, 50 µm. **f**, The circle reflects the 11 hallmarks of transcriptional ITH, with each corresponding to one family of MPs. The size of each section is proportional to the abundance of the MP family across all tumour samples (Supplementary Table 9). **g**, Same as **f**, with each panel corresponding to hallmark frequencies in a particular type of samples—cell lines, glioma, breast cancers and lung cancers. Hallmarks with high relative frequency in each type of samples are labelled. CTCs, circulating tumour cells; SCC, squamous cell carcinoma; NSCLC, non-small-cell lung cancer; AML, acute myeloid leukaemia; CML, chronic myeloid leukaemia; ALL, acute lymphoblastic leukaemia; MM, multiple myeloma. Panels **f**,**g** designed by T. Bigdary.

A PDAC-classical MP linked to mucins

To understand the variability of MP30 in non-PDAC tumours, we examined the TCGA tumours with highest MP30 expression. In HNSCC, tumours with high MP30 expression were enriched with *NSD1* mutations (Fig. 3d). In LUAD, tumours with high MP30 expression had high expression of other PDAC-enriched genes (Fig. 3d and Supplementary Fig. 8) and an enrichment of *KRAS* mutations common in PDAC ($P = 0.0032$, hypergeometric test). This subset of LUADs is enriched ($P = 9.7 \times 10^{-10}$) with histological classification as invasive mucinous adenocarcinoma (Fig. 3d), suggesting a link between this LUAD histology and PDAC features[18]. We validated the expression of the PDAC-classical marker TFF1 (encoded by the top gene in MP30) in lung invasive mucinous tissues (Fig. 3e and Extended Data Fig. 6b). Thus, MP30 is common in PDAC but is also observed at a lower frequency in other contexts, in which it is associated with specific genetics (*NSD1*

mutations in HNSCC) and a mucinous histology (in LUAD; Extended Data Fig. 4g).

Two observations further support a link between MP30 and mucin production. First, MP30 and PDAC-classical signatures[16,17] contain mucin genes (*MUC13*, *MUC5AC* and *MUC5B*) and genes associated with mucin production such as *GCNT3* and *TFF1-3*. Second, the MP30 regulators inferred by SCENIC (CREB3L1 and FOXA3; Extended Data Fig. 2) are associated with differentiation of goblet cells[19,20], which are specialized for mucin production. Thus, MP30 might reflect the aberrant induction of a mucin production program that could occur in a variety of cancer cells but is particularly common in PDAC.

Hallmarks of transcriptional ITH

Owing to the large breadth of samples used to derive MPs, the 11 families of MPs may be considered as hallmarks of transcriptional ITH (Fig. 3f). Two abundant hallmarks (cell cycle and stress) are common in nearly all cancer types and together cover more than half of all programs of transcriptional ITH. The remaining nine hallmarks have lower frequencies and unique context specificities (Fig. 3f,g). For example, interferon response covers 6% of all ITH programs in solid tumours, but is almost absent in sarcoma (1%) and particularly abundant in breast cancer (19%).

The most common hallmark after cell cycle and stress is of 'lineage-related' MPs, including programs that resemble developmental and differentiated cell types such as skin pigmentation, alveolar cells, and neuronal and oligodendrocyte progenitors. These MPs are consistent with the notion that ITH recapitulates developmental trajectories. Abundance of this hallmark varied markedly between cancer types, covering 36% of all ITH programs in glioma[5] but not being detected in other cancer types such as ovarian and breast cancer (Supplementary Table 9). A more lenient method for MP detection identified few additional lineage-related MPs (for example, androgen receptor MP in some breast and prostate tumours; Supplementary Fig. 9 and Supplementary Table 2). Yet, lineage-related MPs remained rare in most cancer types, whereas the main patterns of ITH reflect core cellular processes that are shared across cancer types, such as cell cycle, stress, anti-viral responses and respiration.

Cancer cell lines recapitulate most hallmarks, except for hypoxia (probably reflecting higher oxygen concentration in cell cultures), MYC targets (invariably high in cell lines, possibly reflecting selection for MYC activity during establishment of cell lines), respiration, cilia and secreted (Fig. 3g). Future analysis of mouse models and organoids may shed light on the tumour microenvironment (TME) components required for these ITH patterns. By contrast, the cell cycle and protein regulation hallmarks were particularly abundant in cell lines.

MPs of non-malignant TME cell types

The 3CA compendium contained 1,199,312 non-malignant cells from 38 TME cell types. The 6 most common cell types (fibroblasts, macrophages, T cells, B cells, epithelial cells and endothelial cells) were each represented by more than 50,000 cells from more than 200 tumours (Fig. 4a and Supplementary Table 1). We applied the analysis described above to define MPs in these cell types and annotate them by functional enrichments (Extended Data Fig. 7 and Supplementary Table 10).

Most malignant MPs resemble the non-malignant MPs of epithelial cells but not the MPs of other cell types (Fig. 4b,c and Extended Data Fig. 8), suggesting that a large fraction of the heterogeneity seen in malignant cells already exists in the cells of origin. For example, we identified an EpiSen MP in malignant and in non-malignant epithelial cells, both of which were validated by immunohistochemistry (IHC) (Extended Data Fig. 6). Analysis of pre-malignant tissues (adenomas)[21] suggested an intermediate stage with higher similarity in MP expression to malignant than to non-malignant epithelial cells (Supplementary Fig. 10). Similarly, for each of the six common TME cell types, comparison of MPs between tumours and healthy tissues[22–25] demonstrated extensive similarities (Supplementary Fig. 11 and Supplementary Table 11).

Despite the similarity of malignant MPs to non-malignant epithelial MPs, the remaining differences highlight cancer-specific features (Supplementary Fig. 12 and Supplementary Table 12). In some cases, these differences reflect context-specific coupling between related pathways. For example, the EpiSen malignant MP harbours both epithelial differentiation genes and secreted factors, whereas these are observed as two distinct MPs of non-malignant epithelial cells. This suggests that differentiation and secretion are coupled only in the malignant cells, possibly reflecting an aberrant senescence response.

Coupling of MHC II with interferon MPs

Context-specific coupling was especially common for MHC II genes. The alveolar MP of non-malignant cells includes several MHC II genes, consistent with the antigen presentation of normal lung AT2 cells[26]. Yet, the highly similar malignant alveolar MP lacks MHC II genes, suggesting that specifically in malignant cells, alveolar differentiation is decoupled from antigen presentation (Fig. 4d and Supplementary Fig. 12a).

Conversely, MPs of interferon response tend to be coupled to MHC II expression in malignant, but not in non-malignant, cells. In non-malignant immune and stromal cells, we identified separate MPs for interferon response genes and for MHC II genes, whereas the two sets of genes were highly correlated and part of the same MPs in malignant cells (Fig. 4c,d and Extended Data Fig. 9e–g). This coupling between interferon response and MHC II genes was also observed in non-malignant epithelial cells (unlike other non-malignant cells), albeit more weakly. We validated by IHC the coupling in malignant cells, which was typically seen in proximity to T cells, and the decoupling in non-malignant (for example, endothelial) cells (Fig. 4e and Extended Data Fig. 10). These results suggest that interferon-γ signalling from T cells efficiently activates the expression of MHC II genes in epithelial cells (malignant and non-malignant), whereas additional mechanisms induce MHC II in other cell types, thereby decoupling it from interferon responses.

Among malignant cells, interferon and MHC II are coupled in most tumours (75%) but not in others (Fig. 4d and Extended Data Fig. 9e–g). IHC demonstrated a gradient from coupled to decoupled malignant cells, highlighting the coexistence of these states in the same tumour (Fig. 4e, right panels, and Extended Data Fig. 10). Taken together, our findings show that MHC II, interferon response and alveolar differentiation reflect coherent expression programs that are observed across several cell types and cancer types; yet, the coupling among those programs varies among cell types, among tumours and even spatially within the same tumour, indicating its complex regulation. Expression of MHC II and interferon responses have important effects on T cell functions, raising the possibility that coupling (or decoupling) may be under selection in tumours.

Inferring cell-type associations

TME cell types influence one another through secreted factors, physical interactions or competition over nutrients and oxygen. To uncover such effects, we examined the co-occurrence between MPs of different cell types. The fractions of cells scoring highly for each MP were defined in each tumour, and then centred within each study and combined to an integrated dataset that was used to define MP correlations (Methods). We constructed a graph from the positive correlations (Fig. 5a), as well as from the negative correlations, which were fewer and weaker (Supplementary Fig. 13).

The graph of positive correlations highlighted five clusters (shaded areas in Fig. 5a). One cluster consists of five MPs enriched in LUAD and linked to angiogenesis (grey shaded area in Fig. 5a): two endothelial

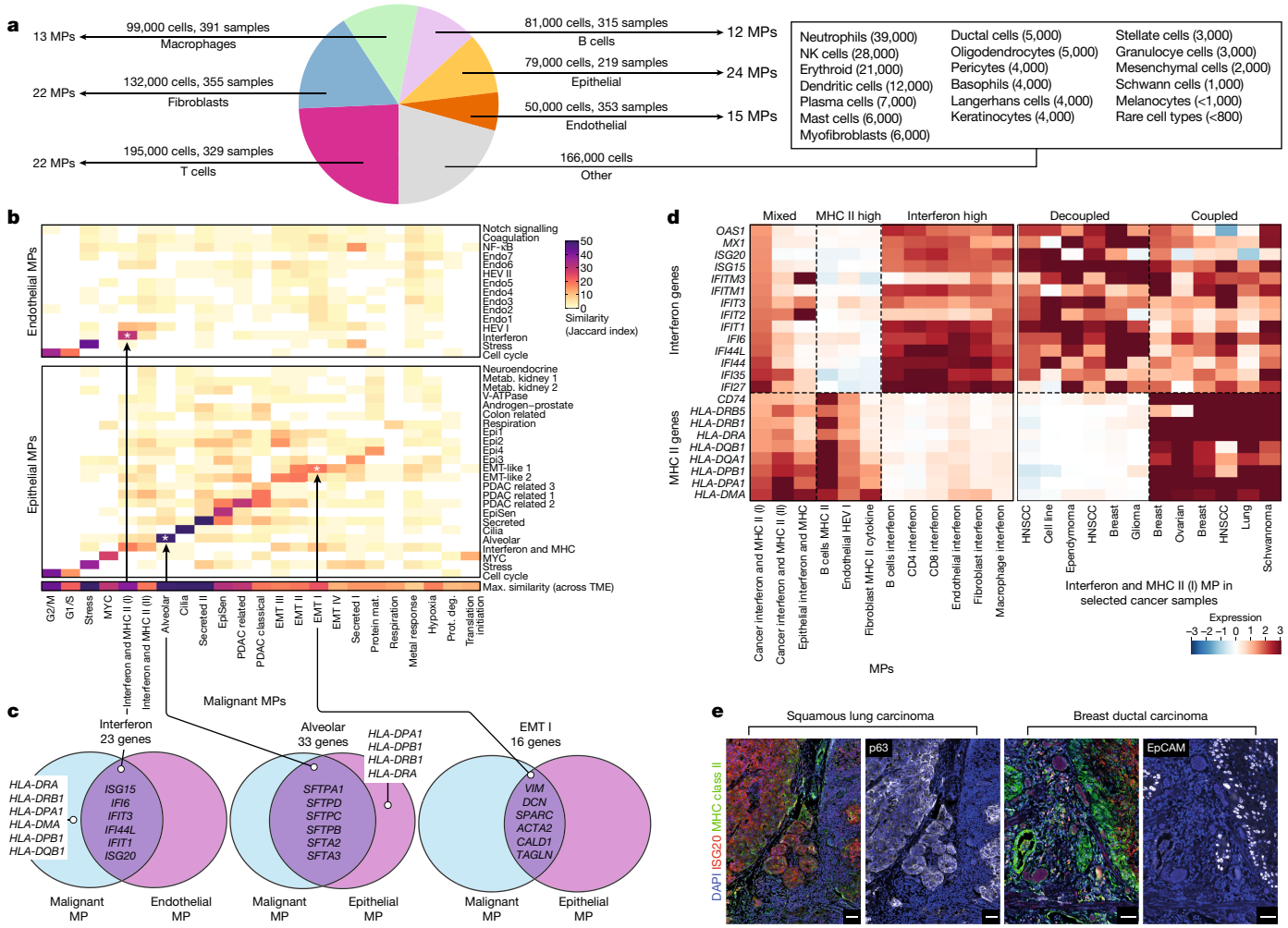


Fig. 4 | Non-malignant MPs and their similarity to malignant MPs. a, Pie chart depicting the annotation of all non-malignant cells in the compendium. Each colour corresponds to a common cell type, and an arrow indicates the corresponding number of cells, tumour samples and MPs; 'Other' (in grey) corresponds to all non-common TME cell types. **b**, Similarity between malignant MPs (*x* axis) and non-malignant epithelial and endothelial MPs (*y* axis), as defined by overlapping genes (Jaccard index). The bottom bar shows the maximal similarity of each malignant MP compared to all MPs from non-malignant (TME) MPs. **c**, Venn diagram depicting the similarities and differences between malignant and non-malignant MPs for selected cases; the number of shared genes is indicated at the top and selected genes are labelled. **d**, Left: mean expression levels of selected MHC II and interferon genes in samples with the related MPs in different cell types. MPs are classified as those with high expression of MHC II, interferon or both (mixed). Right: mean marker expression in selected samples of the cancer interferon and MHC II (I) MP, separated into those for which interferon and MHC II are coupled or decoupled. **e**, Validation of coupling and decoupling of MHC II and interferon marker expression by IHC (representative image from three independent experiments per cancer type). Scale bars, 50 µm. Androgen–prostate, androgen receptor program in the context of prostate cancer; Endo (1 or 2), endothelial; Epi (1 or 2), epithelial; metab., metabolism; mat., maturation.

MPs with genes encoding angiogenesis factors[27] (for example, *EGFL7*, *PLVAP* and *EPAS1*), and three non-endothelial MPs with genes encoding secreted pro-angiogenic factors: *VEGF-A*, *GAS6*, *CLIC3* and *ANXA3* (in Epi1 MP); *VEGF-D* and *SLIT2* (in CAF3 MP); and *CCL18*, *FABP4* and *IGFBP2* (in MAC2 MP). These MPs also included genes encoding anti-angiogenic factors (*NOTCH4*, *TIMP3* and *IGFBP7*), potentially reflecting a negative feedback loop during angiogenesis. These correlated MPs suggest a multicellular TME network linked to angiogenesis in LUAD.

The other four MP clusters were each made up of MPs that reflect the same process in distinct cell types (blue shade in Fig. 5a). These include a cluster of interferon responses (in seven cell types), stress responses (in five cell types), heat-shock responses (in four cell types) and cell cycle (in six cell types). The correlations within these MP clusters remained significant when we excluded shared genes between the cell types, indicating that they cannot be explained by ambient RNA (Supplementary Fig. 13). These clusters may reflect a concomitant response of several adjacent cell types to the same TME features (for example, interferon and other molecules). The cell-cycle cluster seemed unexpected given that proliferation of malignant cells is typically thought to be intrinsically driven by oncogenic mutations. Using spatial proteomics datasets[28–30], we validated the prediction that cycling cells of several cell types are co-localized in 'high-proliferation' niches (Fig. 5b, Supplementary Fig. 5 and Supplementary Fig. 14). Correlated proliferation of diverse cell types may be driven by TME mitogens or by interactions between cell types, such as transfer of proteins through exosomes or other mechanisms. Such transfer could potentially contribute also to the other MP clusters described above, as well as to the correlation that we observe between the MYC-related MPs of malignant cells and B cells[31] (Fig. 5a).

Immune-related TME associations

Next, we shifted our attention from MP clusters to pairwise correlations among MPs. Such correlations could reflect ligand–receptor

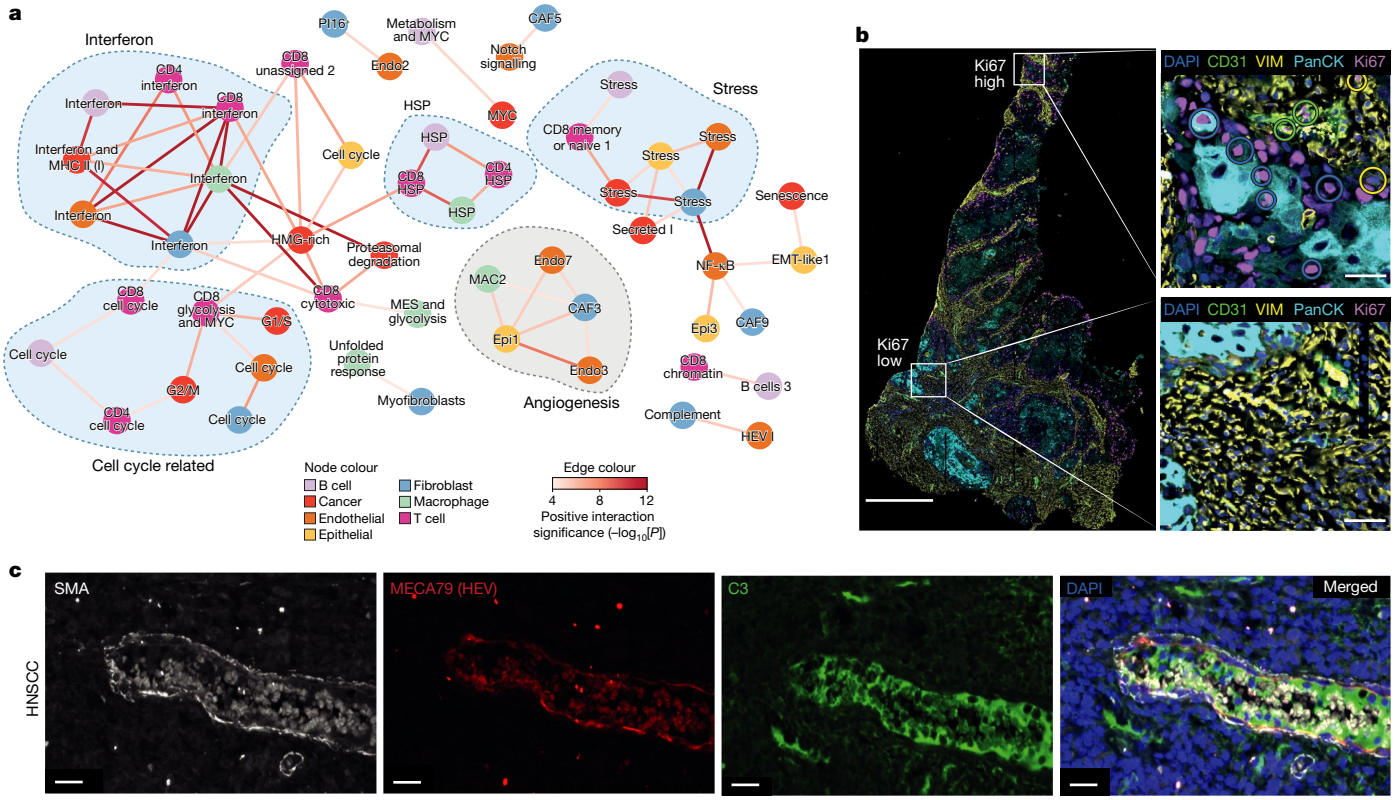


Fig. 5 | Correlations between MPs of different cell types. a, Cytoscape graph of positive correlations (edges) between MPs (nodes), as calculated in the integrated dataset. Node colour indicates cell type and edge colour indicates the correlations' significance. **b**, Validation of cell-cycle coordination using published CODEX data[28–30]. Top inset depicts cycling cells of different cell types that are in spatial proximity (Ki67-high niche), compared to non-cycling cells in the bottom inset (Ki67-low niche). PanCK (blue), CD31 (green) and VIM (yellow) mark malignant, endothelial and fibroblast cells, respectively. Scale bars, 200 µm (main image, left) and 20 µm (insets, right). **c**, IHC validation in a HNSCC sample (representative image from 3 independent experiments) of the co-localization between HEVs (expressing MECA79, red) and complement-expressing fibroblasts (expressing C3 and SMA in green and white, respectively). HSP, heat shock proteins; CAF, cancer associated fibroblast; PI16, peptidase inhibitor 16.

pairs expressed by the respective cell types (Supplementary Table 13). Owing to the large number of potential interactions, we focus here on those that involve the cytotoxicity or recruitment of T cells and hence might influence the response to immunotherapies.

$CD8^+$ T cell cytotoxicity was most strongly correlated with the macrophage interferon response (Supplementary Fig. 15), probably reflecting the secretion of interferon-γ by cytotoxic T cells that then induces an interferon response of macrophages. The second strongest correlation of $CD8^+$ T cell cytotoxicity was with the proteasomal degradation MP of malignant cells. This may reflect the role of proteasomal degradation in antigen presentation through MHC I, which could facilitate the recognition of malignant cells by T cells and their subsequent activation. This association was observed in the combined dataset as well as in datasets of several specific studies (Supplementary Fig. 15 and Supplementary Table 14). Thus, upregulation of the proteasomal MP might be beneficial in the context of immunotherapies.

Although negative associations were weaker, we noticed that the alveolar MP of malignant cells is significantly ($P < 0.05$) negatively correlated both with the cell cycle and with the cytotoxicity MPs of $CD8^+$ T cells (Supplementary Fig. 15 and Supplementary Table 14). This may suggest a negative effect of malignant cell alveolar differentiation on immune activation, possibly linked to the decoupling of this program from antigen presentation, as noted above.

Recruitment of T cells to tumours depends on their migration through specialized vessels, such as high endothelial venules (HEVs)[32]. The frequently observed endothelial MP2 included HEV markers (for example, *ACKR1* as the number 1 gene), MHC II genes and many other immune-related genes such as those encoding chemokines (*CCL2*, *CCL14* and *CXCL14*) and selectins (*SELP* and *SELE*) that mediate leukocyte recruitment and interactions, respectively (Supplementary Table 10). This HEV-like MP is positively correlated with the complement MP of fibroblasts (Fig. 5a). We further validated by IHC the co-localization of HEVs with complement-expressing fibroblasts in HNSCC (Fig. 5c and Supplementary Fig. 16). These results provide evidence of a multicellular stromal organization that may facilitate leukocyte migration, consistent with the previously described correlation between fibroblast-derived complement and T-cell infiltration[3].

Discussion

The transcriptome of a cancer cell may be considered as a proxy to its 'global' state. Comparison of global cell states across tumours underscores a high degree of cellular heterogeneity that remains difficult to interpret. We argue that dividing the differences in global cell states into intertumour and intratumour heterogeneity helps to make sense of such high heterogeneity. Intertumour differences reflect the cumulative effects of genetic and epigenetic aberrations that have been acquired during the lengthy oncogenic process of each tumour, creating distinctive tumour-specific profiles, whereas intratumour differences primarily reflect recent events that have shaped the state of cells, including their phase along the cell cycle, their short-term responses to surrounding cells, cytokines and nutrients (or lack thereof), and their stochastic fluctuations. Whereas intertumour differences have been widely studied during the past two decades through bulk RNA-seq profiles, the ability to analyse intratumour differences emerged only recently and has been carried out at a limited scale. Here we carry out

a systematic analysis of intratumour differences through curation of a large number of scRNA-seq datasets.

We find that relative cell states tend to be shared across tumours. Thus, certain aspects of ITH are predictable, such as subpopulations of EMT-like or senescent cells in HNSCC. Predictable subpopulations may warrant new therapeutic strategies. For example, combination therapies may target coexisting cellular states, and differentiation therapies may shift cells from an aggressive state (for example, EMT-like) to a more benign or responsive state (for example, senescent cells).

Some patterns of ITH may have been overlooked in our analysis. First, rare patterns may not be sampled sufficiently in this cohort. Second, our approach efficiently detects large-scale expression programs but could miss programs of only a handful of genes or those primarily reflecting proteins or metabolites rather than mRNAs. Third, our approach for inferring TME interactions, by MP co-occurrence across the entire compendium, highlights generic interactions over context-specific ones. For example, potential interactions of $CD4^+$ T cells in pancreatic cancer seem to be absent in other cancers (Supplementary Fig. 15). At present, we are underpowered to evaluate each interaction in each cancer context and will revisit these analyses as 3CA is further expanded.

In summary, we curated a large pan-cancer atlas that enabled us to define a comprehensive map of transcriptional ITH. To support the use of this framework by future studies, we provide software to quantify and visualize the MPs reported here in any new scRNA-seq dataset (Extended Data Fig. 11). This framework may be refined by future studies and will guide our understanding of ITH towards new therapeutic strategies.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-06130-4.

1. Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat. Rev. Cancer* **12**, 323–334 (2012).
2. Suva, M. L. & Tirosh, I. Single-cell RNA sequencing in cancer: lessons learned and emerging challenges. *Mol. Cell* **75**, 7–12 (2019).
3. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
4. Rambow, F. et al. Toward minimal residual disease-directed therapy in melanoma. *Cell* **174**, 843–855 (2018).
5. Neftel, C. et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell* **178**, 835–849 (2019).
6. Puram, S. et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171**, 1611–1624 (2017).
7. Kinker, G. S. et al. Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nat. Genet.* **52**, 1208–1218 (2020).
8. Baron, M. et al. The stress-like cancer cell state is a consistent component of tumorigenesis. *Cell Syst.* **11**, 536–546 (2020).
9. Barkley, D. et al. Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. *Nat. Genet.* **54**, 1192–1201 (2022).
10. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
11. Meister, N. et al. Interferon-gamma mediated up-regulation of caspase-8 sensitizes medulloblastoma cells to radio- and chemotherapy. *Eur. J. Cancer* **43**, 1833–1841 (2007).
12. Vander Heiden, M. G., Cantley, L. C. & Thompson, C. B. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science* **324**, 1029–1033 (2009).
13. Denisenko, E. et al. Spatial transcriptomics reveals ovarian cancer subclones with distinct tumour microenvironments. Preprint at https://doi.org/10.1101/2022.08.29.505206 (2022).
14. Ji, A. L. et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell* **182**, 1661–1662 (2020).
15. Ravi, V. M. et al. Spatially resolved multi-omics deciphers bidirectional tumor-host interdependence in glioblastoma. *Cancer Cell* **40**, 639–655 e613 (2022).
16. Moffitt, R. A. et al. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat. Genet.* **47**, 1168–1178 (2015).
17. Raghavan, S. et al. Microenvironment drives cell state, plasticity, and drug response in pancreatic cancer. *Cell* **184**, 6119–6137 (2021).
18. Shim, H. S. et al. Unique genetic and survival characteristics of invasive mucinous adenocarcinoma of the lung. *J. Thorac. Oncol.* **10**, 1156–1162 (2015).
19. Asada, R. et al. The endoplasmic reticulum stress transducer OASIS is involved in the terminal differentiation of goblet cells in the large intestine. *J. Biol. Chem.* **287**, 8144–8153 (2012).
20. Chen, G. et al. Foxa3 induces goblet cell metaplasia and inhibits innate antiviral immunity. *Am. J. Respir. Crit. Care Med.* **189**, 301–313 (2014).
21. Chen, B. et al. Differential pre-malignant programs and microenvironment chart distinct paths to malignancy in human colorectal polyps. *Cell* **184**, 6262–6280 (2021).
22. Guilliams, M. et al. Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. *Cell* **185**, 379–396 (2022).
23. Deprez, M. et al. A single-cell atlas of the human healthy airways. *Am. J. Respir. Crit. Care Med.* **202**, 1636–1645 (2020).
24. Burclaff, J. et al. A proximal-to-distal survey of healthy adult human small intestine and colon epithelium by single-cell transcriptomics. *Cell Mol. Gastroenterol. Hepatol.* **13**, 1554–1589 (2022).
25. The Tabula Sapiens Consortium. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
26. Toulmin, S. A. et al. Type II alveolar cell MHCII improves respiratory viral disease outcomes while exhibiting limited antigen presentation. *Nat. Commun.* **12**, 3993 (2021).
27. Lugano, R., Ramachandran, M. & Dimberg, A. Tumor angiogenesis: causes, consequences, challenges and opportunities. *Cell. Mol. Life Sci.* **77**, 1745–1770 (2020).
28. Schurch, C. M. et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell* **182**, 1341–1359 (2020).
29. Zhang, W. et al. Identification of cell types in multiplexed in situ images by combining protein expression and spatial information using CELESTA. *Nat. Methods* **19**, 759–769 (2022).
30. Blise, K. E., Sivagnanam, S., Banik, G. L., Coussens, L. M. & Goecks, J. Single-cell spatial architectures associated with clinical outcome in head and neck squamous cell carcinoma. *NPJ Precis. Oncol.* **6**, 10 (2022).
31. Borzi, C. et al. c-Myc shuttled by tumour-derived extracellular vesicles promotes lung bronchial cell proliferation through miR-19b and miR-92a. *Cell Death Dis.* **10**, 759 (2019).
32. Hua, Y. et al. Cancer immunotherapies transition endothelial cells into HEVs that generate $TCF1^+$ T lymphocyte niches through a feed-forward loop. *Cancer cell* **40**, 1600–1618 (2022).
33. Lee, H.-O. et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat. Genet.*, https://doi.org/10.1038/s41588-020-0636-z (2020).
34. Steele, N. G. et al. Multimodal mapping of the tumor and peripheral blood immune landscape in human pancreatic cancer. *Nat. Cancer* **1**, 1097–1112 (2020).

Methods

Data curation

A total of 77 single-cell RNA-seq (scRNA-seq) datasets, which include 1,456 samples (that is, tumours), were curated. Most of these datasets are available at https://www.weizmann.ac.il/sites/3CA aside from unpublished datasets and datasets for which we did not obtain sharing permission from the authors of the original studies. For dataset selection, we first constructed a list of potentially relevant studies through PubMed search, and continuously updated it through literature review. Each study was then examined for the type and amount of scRNA-seq data generated in it, prioritizing studies with data for several patient tumours and including a reasonable fraction of malignant cells. When seeking to add each new dataset to our cohort, we initially checked whether the data are publicly available for downloading (for example, through the Gene Expression Omnibus database repository). Most publications freely provided the data in the form of an expression matrix, together with a list of associated genes and cells (barcodes). Several datasets were only available on author consent, in which case we contacted the authors for permission. The cohort also includes unpublished and published datasets that were previously analysed in our laboratory, and that were already sequenced, aligned and processed. Most of the external datasets we downloaded, even when freely available, did not include the cell annotations presented in the published manuscript. In these cases, we contacted the leading authors and requested that they provide us with the annotations they used. In some cases, annotations of the original study were either not provided by the authors or were limited (for example, not distinguishing between malignant and non-malignant epithelial cells), in which case we inferred the annotations ourselves.

Verification of cell annotation

For each dataset—both for author-based annotations and for the annotations that we defined—we carried out the following analyses to ensure the validity of the annotations. First, we carried out dimensionality reduction using UMAP and examined whether the cells annotated as different cell types clustered separately. Second, we validated the annotations by verifying that the top differentially expressed genes of each cell type match known marker genes. Finally, we inferred CNAs (using the package available at https://github.com/jlaffy/infercna), to verify the annotation of cells as malignant. Some samples in which we could not resolve the annotations, possibly owing to low data quality, were excluded during this process and were not used for further analysis.

Data preprocessing

The following preprocessing steps were carried out before conducting downstream analysis.

Cell filtering. We excluded cells with a low number of detected genes (n_{genes}). For 10x data, our cutoff was typically $n_{genes} > 1{,}000$. For smart-seq2 data, we typically used a higher cutoff of $n_{genes} > 2{,}000$ genes. For other single-cell platforms, we adapted the cutoff and in some cases used a threshold lower than 1,000 genes, but never lower than the cutoff used by the original study.

Sample filtering. After cell filtering, we excluded samples having fewer than 10 malignant cells. When appropriate, we also excluded samples with unresolved CNA patterns. In total, 1,163 samples were retained for downstream analysis.

Gene filtering. Given an expression matrix A with n genes (rows) and m cells (columns), the mean expression of gene i across cells is given by $E_{i\{1..n\}} = \sum_{j=1}^{m} \frac{A_{ij}}{m}$. For most analyses, we kept the 7,000 genes with the highest E_i value in each sample.

Normalization. UMI counts were converted to counts per million (CPM). For most analyses, each entry in the matrix was then normalized according to $E = \log_2\left(\frac{\mathrm{CPM}}{10} + 1\right)$. The same normalization was used for transcripts per million (TPM) values. The values were divided by 10 as the actual complexity is assumed to be in the realm of about 100,000 and not 1 million as implied by the CPM and TPM measures.

Centring. For most analyses, the data were centred (each gene was centred across all cells). Centring was carried out separately for each study.

Defining a non-redundant set of robust NMF programs

We carried out NMF for each sample separately, to generate programs that capture the heterogeneity within each sample. Negative values in each centred expression matrix were set to zero. As application of NMF requires a 'K' parameter that influences the results, we ran NMF using different values (K = 4, 5, 6, 7, 8 and 9), thereby generating 39 programs for each tumour. Each NMF program was summarized by the top 50 genes based on NMF coefficients. We reasoned that the most meaningful NMF programs are those that would recur across different values of K as well as across tumours; such programs (denoted here as robust NMF programs) were defined by the following three criteria: robust within the tumour (a program that is represented by several similar NMF programs, as defined for the same tumour when analysed by multiple K values; two NMF programs were considered as similar if they had at least 70% gene overlap (35 out of 50 genes)); robust across tumours (NMF programs that had at least 20% similarity (by top 50 genes) with any NMF program in any of the other tumours analysed); non-redundant within the tumour (within each tumour, NMF programs were ranked by their similarity (gene overlap) with NMFs from other tumours and selected in decreasing order; once an NMF was selected, any other NMF within the tumour that had 20% overlap (or more) with the selected NMF was removed, to avoid redundancy). This approach yielded a total of 5,547 robust NMF programs

Defining MPs

We next clustered the robust NMF programs according to Jaccard similarity. The clustering was carried out using a custom approach (Supplementary Fig. 1) that defined clusters of NMF programs and a list of 50 genes that constitute the MP. In brief, each robust NMF program was compared to all other robust NMF programs to assess the degree of gene overlap between programs. Considering overlap instances of at least 10 genes, the programs with the maximal number of considerable overlaps was selected as a potential founder of a new cluster. If the number of overlapping NMF programs (>10 genes) exceeded 5 cases, the NMF program with the highest gene overlap to the founder NMF program was added, and thus a cluster was formed. The MP for the cluster was initially defined by the genes that appeared in both programs. To complete the list to 50 genes, the genes with the top NMF scores (in either program) were selected. The process was then repeated by searching for the program with maximal overlap with the MP, adding it to the cluster as long as the overlap was at least 10 genes, and updating the MP (selecting genes that appeared in the highest number of programs, and completing the MP list to 50 genes according to NMF scores in the set of programs). In this way, the MP was updated after addition of each NMF, reflecting the genes common to the programs constituting the cluster. The cluster was completed when no NMF could be added, and an attempt to form a new cluster was made as described above.

This approach yielded 67 initial MPs. We further removed MPs that: were suspected to reflect low-quality data or other technical confounders, with a strong enrichment of either ribosomal protein genes or mitochondrial-encoded genes; included NMF programs from only a single study, even though that study reflected a cancer type that is also examined by additional studies; or were suspected to reflect

doublet cells on the basis of high similarity to the expression profile of a non-malignant cell type (for example, T cells or macrophages). We retained 41 MPs, and assessed their enrichment in functionally annotated gene sets.

Gene sets for functional enrichment analyses

We primarily used signatures from MsigDB, including the following collections of gene sets: Gene Ontology (C5.GOBP, C5.GOCC and C5.GOMF), Hallmark (H) and Cell Types (C8). We also added selected additional signatures (not taken from MsigDB) related to brain or glioma, alveolar and PDAC signatures. Signatures with a false discovery rate (FDR)-adjusted $P < 0.05$ (hypergeometric test) were considered significantly enriched (Supplementary Table 3). The 41 MPs were further grouped by functional similarities into 11 hallmarks (Fig. 2). Metabolism-related signatures were defined as all gene lists that contained in their title the word 'metabolism', or words associated with metabolism, that were derived from MSigDB (including Gene Ontology, Hallmark, Cell Types and curated sets; Supplementary Table 5).

Inferring MP regulators

We applied the SCENIC method[10] to malignant cells in each of the samples in which we identified cancer MPs. The SCENIC workflow includes: identifying gene sets that are co-expressed with transcription factors; retaining modules with significant cis-regulatory motif enrichment of the correct upstream regulator (termed regulons); and deriving regulon expression scores in each cell (using the AUCell algorithm).

In each sample that participated in a malignant MP, we calculated the correlation between the cells' malignant MP scores and their regulon scores. We then selected the regulons with a statistically significant correlation ($P < 0.05$ after FDR correction) that is above 0.5 or below −0.5. We retained the top three positively or negatively correlated regulons as potential MP regulators. We also retained regulons that were consistently selected in many samples (at least 10 different samples or at least 30% of all samples with the MP), whereas those selected only in one or very few samples were excluded. Correlation values of selected regulons were averaged across samples in each MP and then clustered (Extended Data Fig. 2). We also searched for regulons that had significantly higher or lower correlations in samples of one or more cancer types compared to the rest based on $P < 0.05$ by a hypergeometric test with FDR correction (Extended Data Fig. 2 and Supplementary Table 6). Notably, some of the regulons that came up according to SCENIC are not known transcription factors according to our knowledge or the literature. We removed the questionable regulons from Extended Data Fig. 2b, but retained all SCENIC output in Extended Data Fig. 2a and in the full regulon list in Supplementary Table 6. The top correlated transcription factors in Supplementary Table 6 were defined as regulons with a mean correlation of above 0.5 (positive regulons) or below −0.5 (negative regulons). For MPs with fewer than three such positive regulons, we added the subsequent top positive ones so that the list was made up of at least three positive regulons when possible.

Subclone analysis

We systematically inferred CNA patterns throughout the scRNA-seq compendium to identify the malignant cells, as described above. Each CNA matrix was then first filtered, retaining the top 67% genes with highest CNA absolute signal. We next derived several clusters using Louvain clustering with $k = 15$. A chromosomal arm was considered to be deleted or amplified in a cluster when the cluster cells' average CNA values across genes in a chromosome were smaller than −0.15 or larger than 0.15. To ensure that clusters correspond to distinct genetic subclones, we then iteratively merged all pairs of clusters for which we could not find any chromosome arm with evidence for a distinct copy number according to the following criteria: clusters had the same assignment as deleted, amplified or neither across all chromosomes; clusters had a maximal difference across all chromosomes that was less than 0.15.

After the merging process was completed, the remaining clusters were considered as distinct subclones and were examined for differences in MP expression. Finally, in each subclone we calculated: the CNA signal, defined as the mean of the absolute CNA values across the genes with the top 67% values; the CNA correlation, defined as correlation between CNA values (per cell) and the mean CNA values of the top 25% cells as defined by CNA signal.

We determined the thresholds for CNA signal and correlation as the values observed in the lower percentile (1%) of cells by each measure. Cells passing both thresholds were considered to be malignant cells, cells that passed only one of the thresholds were unresolved, and cells that passed neither threshold were considered non-malignant (Extended Data Fig. 3).

Bulk RNA-seq gene set scores

Bulk tumour RNA-seq profiles were downloaded from TCGA via the Broad GDAC Firehose (https://gdac.broadinstitute.org). Malignant-cell-specific profiles, obtained through the deconvolution algorithm CIBERSORTx, were provided by the authors of Luca et al. (ref. 35) Expression levels per tumour were defined as $\log_2$[TPM]. For a given gene set (for example, an MP), a score was defined for each tumour through a method described previously (ref. 36). In brief, for each gene in the given gene set, the expression levels of this gene were centred relative to the average across a set of control genes, which are chosen to have similar average expression levels (across all tumours) to the given gene. The tumour score for this gene set is then defined as the average of these centred expression levels across all genes in the set. Scores for a given MP were computed only for those cancer types in which this MP was detected through scRNA-seq.

Association of bulk RNA-seq MP scores with genomic alterations

Mutation annotations and discretized CNA values for TCGA samples were obtained from the Broad GDAC Firehose. To limit our analysis to the most relevant events, we considered only functional mutations, focal CNAs and whole-arm CNAs. Mutations were considered only for genes in the Cancer5000-S set from Lawrence et al.[37], and were considered functional if they were nonsense, non-stop, frameshift indels or occurring at a splice site or translation start site, or missense or in-frame indels occurring in at least two patients. Focal CNAs were defined as genes having discretized values of ±2, whereas whole-arm CNAs were defined for chromosome arms that had at least 100 genes and which had an average discretized value greater than 0.9 or less than −0.9. Of all these genomic alterations, in each cancer type we considered only those that were detected in at least 10 samples and in at least 10% of samples, and which were not detected in at least 10 samples. For cancer types having highly distinct subtypes that might confound observed associations (for example, genetic or expression subtypes), we considered these subtypes separately. For a given genomic alteration in a given cancer type or subtype, effect sizes were defined as the difference in average MP scores between samples with and without this alteration, and P values were computed by t-test. Only associations with effect size greater than 0.4 were retained for further analysis. Among these, P values were adjusted by Benjamini–Hochberg correction, and cases were deemed significant if their adjusted P values were less than 0.05. As there were many more significant whole-arm CNAs than mutations or focal CNAs, a stricter effect-size threshold of 0.6 was used for Extended Data Fig. 4c,d.

Association of bulk RNA-seq MP scores with clinical features

Clinical annotations for TCGA samples were obtained from the Broad GDAC Firehose and from Liu et al.[38], and data from these two sources were amalgamated to achieve the highest possible number of annotated samples. For overall survival and progression-free interval, effect size was defined as the hazard ratio, computed through Cox regression, with P value obtained by overall likelihood test. For all other clinical

variables, the samples were divided into two groups and effect size was defined as the difference in average MP scores between these two groups, and *P* values were obtained by *t*-test. The criteria for dividing samples into groups varied between clinical features and cancer types, depending on the data available and the distribution of values across samples (see source code in https://github.com/tiroshlab/3ca/tree/main/ITH_hallmarks). For cancer types having highly distinct subtypes that might confound observed associations (for example, genetic or expression subtypes), we considered these subtypes separately. After carrying out these computations using both non-deconvolved and deconvolved profiles, *P* values were adjusted across all of these computations using the Benjamini–Hochberg method (Extended Data Fig. 5).

MP abundance assessment

We first calculated, for each MP, the observed number of MP-related NMF programs in each cancer type. The expected abundance was then defined by multiplying the number of MP-related NMF programs (that is, the MP size) by the total number of NMF programs identified in the cancer type, across all MPs, and dividing that by the total number of robust NMF programs. Finally, for each combination of cancer type and MP, we calculated $A = \log_2\left[\frac{\text{observed}+1}{\text{expected}+1}\right]$ and the Bonferroni-adjusted *P* value using a hypergeometric test. The abundance classification in Fig. 3a was defined as follows—absent: 0 MP-related NMF programs in that cancer type; low: 1 MP-related NMF program in that cancer type and $-1.5 < A \leq 0$; medium: between 2 to 10 MP-related NMF programs in that cancer type or $0 < A \leq 1$; high: >10 MP-related NMF programs in that cancer type or $A > 1$; high (significant): same as high and adjusted *P* value < 0.05.

IHC staining protocol

Following deparaffinization and rehydration, antigen retrieval of FFPE sections was carried out using citrate buffer (pH 6; Sigma catalogue number C9999) in a pressure cooker. Sections were blocked in CAS block (Thermo Fisher catalogue number 8120) and incubated with primary antibody diluted in CAS block overnight at 4 °C. Primary antibodies included: mouse MHC class II (1:100, ab55152; Abcam), rabbit ISG20 (1:100, ab154393; Abcam), goat SMA (1:200, LS-b3933-50; LifeSpan Biosciences), goat p63 (1:200, AF1916; R&D Systems), rat CD3 (1:100, LS-B8765-50; LifeSpan Biosciences), goat EPCAM (1:100, AF960; Millipore), rabbit TFF1 (1:100, ab92377; Abcam), mouse pan-cytokeratin (1:100, ab86743; Abcam), rat MECA79 (1:100, NB100-77673; Novus Biologicals), rabbit C3 (1:100, ab7462; Abcam), rabbit SLPI (1:100, PA582990; Invitrogen), rabbit SPRR1B (1:100, LS-C161464-400; LifeSpan Biosciences) and mouse LAMC2 (1:100, NBP2-42388; Novus Biologicals). Primary antibodies were initially validated with positive controls using HNSCC (SPRR1B, SLPI, EPCAM, panCK, p63 and LAMC2), PDAC (TFF1 and EPCAM) and a tissue microarray containing tonsil, lymph node spleen, placenta and Hodgkin's lymphoma (CD3, C3, SMA, MECA79, ISG20 and MHC class II). Sections were stained with fluorescent secondary antibodies (all used at 1:200) for 2 h at room temperature and included: donkey anti-rabbit Cy3 (711-165-152, Jackson), donkey anti-goat AlexaFlour 647 (705-605-003, Jackson), donkey anti-rat AlexaFluor 647 (712-605-150, Jackson), donkey anti-goat Cy3 (705-165-003, Jackson), donkey anti-rabbit FITC (711-095-152, Jackson) and donkey anti-mouse FITC (711-095-152, Jackson). Sections were mounted with Fluoroshield mounting medium containing DAPI (Sigma, catalogue number F6057). Whole-slide image scanning was carried out using a PhenoImager Fusion (Akoya Biosciences) and images were analysed with QuPath[39] version 0.3.2. The antibodies and samples that were used for staining are listed in Supplementary Table 15.

Analysis of MP30 (PDAC-classical) in LUAD

Expression data for LUAD and PDAC tumours were obtained from TCGA and normalized as described above, and patient samples were removed if they had no accompanying mutations data or if several tumour samples existed for the same patient. LUAD tumours were scored for the 50-gene PDAC-classical signature using the method described above. By manual inspection of the distribution of these scores, a long tail was observed above 1.5; hence, this threshold was chosen to distinguish PDAC-classical-high LUAD tumours. Hypergeometric tests were used to quantify the enrichment of mucinous samples, and likewise the enrichment of KRAS mutations, among the PDAC-classical-high tumours. The mean correlation of each LUAD sample with PDAC samples was calculated as follows. First, for each of LUAD and PDAC, we computed the variability of all genes by median absolute deviation from the median and selected the 2,500 genes with the highest such variability. Restricting to the intersection of these two gene sets (about 1,500 genes), we computed the pairwise correlations between LUAD and PDAC expression profiles. For each LUAD sample, we then computed the mean such correlation across PDAC samples (Fig. 3d). We also showed that LUAD samples that contributed an NMF program to the PDAC-classical MP had higher resemblance to PDAC samples by averaging the top 100 differentially expressed PDAC genes (obtained by comparing all PDAC samples to all LUAD samples) in LUAD samples (Supplementary Fig. 8, $P < 0.05$ by two-sample *t*-test).

Spatial transcriptomics Visium data analysis

We used published Visium datasets, spatially profiling tumour samples from the ovary (8 samples), skin (4 samples) and GBM (9 samples)[13–15]. The Visium platform averages the transcriptional profile of cells in each spatial spot, resulting in a higher depth per spot but lower sensitivity per transcript. This effect, along with other technical factors such as tissue dissociation and lateral diffusion between spots, results in platform batch effects that can distort the results scoring spots to programs. To this end, we first tested whether the MP genes are truly captured as part of the program in the spatial data, by correlating the gene expression levels to the MP score per spot. Genes showing a Pearson correlation coefficient equal to or greater than 0.2 were selected, and only MPs with at least 10 genes were kept for further scoring. We developed a semi-automated approach that enabled us to iterate through each sample and assign a per-spot malignant score (using CNA inference), and cell-type annotation (by scoring for canonical cell-type marker genes). To assign cellular states, spots that were confidently assigned as malignant (>95 percentile for both CNA score and CNA correlation) were then scored to the scRNA-seq-derived cancer MPs (Supplementary Table 2). Non-malignant cell types were each scored to their corresponding specific cell-type MPs (Supplementary Table 10). The proportion of MP abundance per cancer type in the spatial data (that is, the proportion of spots assigned to an MP per sample) was compared to the observed NMF abundance in the scRNA-seq (computed for Fig. 3a) after centring the values across cancer types and using Pearson correlation (Supplementary Fig. 5).

Global expression versus variability

To calculate the mean (global) expression of an MP and its variability within individual samples (Supplementary Fig. 6), we first averaged the (non-centred) expression of MP genes in each cell and $\log_2$-normalized each value. To compare the global MP expression and variability within each cancer type, the global MP expression was averaged across samples from the same caner type. The variability of the MP expression in a given cancer type was determined as the fraction of tumours (out of the total number of tumours in that cancer type) that contributed at least one NMF program to the MP. For each MP, we then calculated the Pearson correlation between global expression and variability across all cancer types.

Estimating the fraction of NMF programs and Louvain clusters accounted for by MPs

Each individual cell was considered to score positively to an MP on the basis of a one-sample *t*-test across the 50 genes of the MP, with null = 0

(because the input expression data have been centred per gene). We used a threshold of $P < 0.05$ after adjusting for multiple comparisons. To determine whether an NMF scored significantly to an MP, we first centred all of the NMF programs in each sample (centring each rank separately), and then scored against MPs as we did with cells. We then carried out Louvain clustering for each sample using $k = 10$, which resulted in an average of about 4 ± 1.8 clusters per tumour (33 tumours that had only 1 cluster were removed from this analysis). Clusters with more than 50% cells that scored positive to an MP were considered as being accounted for by an MP and their frequency in each sample was compared to the frequency of robust NMF programs in the sample that scored significantly positive to an MP (Supplementary Fig. 2).

MP association with proliferation

In each sample, we scored every cell to the 41 MPs and calculated the correlations between the scores (across cells) of the four cell-cycle-related MPs and the rest of the MPs that were identified as variable within that sample. We kept the maximal correlation out of these four, to represent correlation of an MP with cell cycle. We averaged the correlations of each MP across samples, retaining only the 29 MPs that were represented in at least 7 samples, and using one-sample *t*-test (with null = 0) to define its significance (Supplementary Fig. 4).

Defining TME MPs

We adopted an identical approach to that described above for the cancer cells to define MPs for the six main TME components: macrophages, fibroblasts, endothelial cells, B cells, T cells and non-malignant epithelial cells. The T cells were split into CD4 or CD8 cells as follows: if {CD4,CD8A,CD8B} > 0, a cell was assigned as $CD4^+$ when CD4 > mean{CD8A,CD8B} and as $CD8^+$ otherwise; if CD8A = 0, a cell was assigned as $CD4^+$ when CD4 > CD8B and as $CD8^+$ otherwise (similarly, if CD8B = 0, a cell was assigned as $CD4^+$ when CD4 > CD8A and as $CD8^+$ otherwise); when a cell had values of 0 for CD4, CD8A and CD8B it was unassigned.

After removing programs that we suspected represented low-quality cells or mis-annotations, we were able to define 13 macrophage MPs, 22 fibroblast MPs, 15 endothelial MPs, 12 B cell MPs, 10 CD4 MPs, 12 CD8 MPs and 24 epithelial MPs (Supplementary Table 10).

Defining MPs in non-malignant tissue

We curated published normal scRNA-seq data from 4 large studies that included 184 samples and 741,309 cells[22–25]. We then followed a similar approach to how we defined MPs in malignant tissues, in which we verified the published cell annotations, derived robust NMF programs and finally generated MPs for epithelial cells, T cells, B cells, macrophages, fibroblasts and endothelial cells (see sections above). In the last step of generating MPs, we integrated the robust malignant NMFs with the robust normal NMFs of each cell type before the clustering was carried out to test which clusters were homogeneously mixed (that is, not enriched with malignant or non-malignant NMFs), and which clusters were enriched or depleted of a given NMF type (Supplementary Fig. 11). In each MP, we calculated the relative enrichment score, which is the log ratio between the observed and expected number of non-malignant NMF programs per cluster (see the above section entitled MP abundance assessment for details of how the observed and expected values were defined). MP annotations were obtained on the basis of functional enrichment as defined in earlier sections. We also derived MPs from normal NMFs without prior integration with malignant NMFs (Supplementary Table 11).

Cancer-type- and cell-type-specific associations within MPs

To search for differentiating features between different cancer types within the same MP, we focused on cancer types that contributed at least ten samples to an MP, given that there were at least ten remaining other samples in the MP. To this end, we applied two approaches—correlation based: we calculated the correlation of each gene with the scores of the MP within each sample, then averaged the correlations per cancer type and compared these average correlations, retaining genes with an absolute difference of 0.2 or above between samples of a given cancer type and the other samples; differential expression (DE)-based: we calculated the DE between cells with MP scores above 1 and cells with MP scores below 0 within each sample, then averaged the differential expression per cancer type and compared the average DE values, retaining genes with an absolute difference of 1 or above between samples of a given cancer type and the other samples.

The two methods had high overall agreement and in subsequent analysis, we included only genes that were upregulated or downregulated according to both approaches (Extended Data Fig. 9a–d and Supplementary Table 12).

To compare the gene expression patterns between cancer and epithelial MPs of a similar function we grouped the cancer and epithelial NMF programs that constituted the respective MPs, and calculated the fraction of NMFs in each group with an NMF score that is larger than the median of the top 5 percentile of all cancer and epithelial NMF scores. After normalizing the fraction of each gene by its maximal value, we assigned each gene according to Euclidian proximity to one of the following vectors: {[0.5,0.5];[0.5,1];[1,0.5];[1,1]}. Genes belonging to [1,0.5] were considered as differentially expressed in the cancer cells, whereas genes belonging to [0.5,1] were considered as differentially expressed in the epithelial cells. Genes belonging to [1,1] were high in both groups, whereas genes belonging to [0.5,0.5] were low in both groups. Supplementary Fig. 12 shows the non-normalized fractions for three MPs (alveolar, EMT and EpiSen) after dividing the genes into groups by the approach described above.

To test the degree of MHC II and interferon coupling in malignant and TME cells, we curated lists of genes that represented MHC I, MHC II and interferon signalling, along with several other gene groups listed in Supplementary Table 12. We compared their individual expression levels, NMF scores and Jaccard similarity to the MPs that were annotated as related to MHC or interferon signalling (Fig. 4d and Extended Data Fig. 9e,f). To test the degree of MHC II and interferon coupling per cancer type within MPs, we focused on cancer types that contributed at least six samples to the corresponding MP, and averaged the NMF scores in cells with a score > 1 (Extended Data Fig. 9f). A similar approach was used for comparing average gene expression values within samples that contributed to the cancer interferon and MHC II (I) MP—after centring across genes in each sample, we averaged the expression values in cells with a score > 1 to the MP (Fig. 4d and Extended Data Fig. 9g).

Correlations between MPs of different cell types

To test the significance of the co-occurrence of cancer and TME MPs across tumours, we selected studies that contained at least ten samples with both cancer cells and TME cells (of at least one type). We next scored cancer cells for all cancer MPs, and similarly scored TME cells for their respective MPs (using the sigScores function from https://github.com/jlaffy/scalop) and calculated the adjusted score of each MP—defined as the percentage of cells of a given type in each sample with a score > 1 for their respective MPs. Next we calculated the Pearson and Spearman correlations across all tumours between all possible adjusted scores, and their significance ($-\log_{10}$[*P* value]). We removed cases for which the Spearman correlation was not significant (*P* value ≥ 0.05), and also ignored interactions between pairs of TME fractions that we reasoned might be technical. Using Cytoscape software (version 3.9.1), we generated graphs depicting the positive and negative Pearson correlations and their significance. For the positive correlations, we considered interactions with a minimal significance of 4 (Fig. 5a) whereas for the negative correlations in which the significance values tended to be lower all interactions with a significant *P* value ($P_{value} < 0.05$) were shown (Supplementary Fig. 13).

Validating MP associations with CODEX data

We downloaded annotated CODEX datasets from three independent recent studies of HNSCC and CRC tumours[28–30], resulting in a total of 243 annotated regions of interest coming from 59 tumours and 50 patients. We assessed Ki67 indices reflecting the proportion of cycling cells for every annotated cell type and carried out correlation analysis within all regions of interest (n = 243, area = 2,300–25,002 μm^2). For a given pair of cell types, spots or regions of interest were excluded if fewer than 20 cells of these cell types were present in that area. In the Schuerch et al. dataset (ref. 28), mean Ki67 intensity per TMA spot was assessed and outlier spots with suspected artificial high Ki67 expression were removed (mean Ki67 intensity > 100, n = 33). Ki67 intensity was $\log_2$-transformed and a cutoff for Ki67-positivity calling was calculated for each spot by the formula: mean +1.5 standard deviations. In the Zhang et al. dataset (ref. 29), to increase comparability to the other studies that used regions of interest or TMA spots rather than large whole-slide samples, each of the eight samples was randomly segmented into regions of interest with 20,002 μm^2 (n = 100). Cutoffs for Ki67-positivity calling were manually chosen for each sample. In the Blise et al. dataset (ref. 30), regions of interest (n = 46, area = 25,002 μm^2) and Ki67-positivity calling were used as defined by the authors.

Ligand–receptor interactions

To test for ligand–receptor interactions in the associations shown in Fig. 5a and Supplementary Fig. 13, we grouped the cells with a score > 1 in each node (that is, the cells that contributed to the association) and ranked the genes in these cells according to mean expression. We checked for existence of known ligands or receptors in the top 4,000 ranked genes in each node with paired receptors or ligands in MP genes of connected nodes (Supplementary Table 13). The list of putative ligand–receptor pairs was taken from Ramilowski et al.[40].

Reporting summary

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

Data availability

This work relied on curation and integrative analysis of external studies and did not involve generation of new primary data. The curated data from 71 studies are available at https://www.weizmann.ac.il/sites/3CA, including the primary datasets and results of multiple downstream analyses. Datasets from one curated study (PDAC study by Chan-Seng-Yue et al. 2020 (ref. 41)) are available only through EGA with accession code EGAS00001002543 (permissions for sharing through 3CA were denied). Additional datasets from unpublished studies will be added when possible.

Code availability

The code for generating MPs and inferring MP distribution across and within samples in a study (Extended Data Fig. 11) is provided in https://github.com/tiroshlab/3ca/tree/main/ITH_hallmarks. Additional code for downstream analysis will be provided in the future in the 3CA website and through GitHub (https://github.com/tiroshlab/3ca)[42].

35. Luca, B. A. et al. Atlas of clinically distinct cell states and ecosystems across human solid tumors. *Cell,* https://doi.org/10.1016/j.cell.2021.09.014 (2021).
36. Tirosh, I. et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).
37. Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
38. Liu, J. et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416 (2018).
39. Bankhead, P. et al. QuPath: open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878 (2017).
40. Ramilowski, J. A. et al. A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat. Commun.* **6**, 7866 (2015).
41. 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).
42. Tyler, M. & Gavish, A. tiroshlab/3ca: First release (v1.0.0). *Zenodo,* https://doi.org/10.5281/zenodo.7688626 (2023).

Acknowledgements This work was supported by grants from the Israel Science Foundation, the Zuckerman STEM Leadership Program, the Mexican Friends New Generation, the Neuroendocrine Tumor Research Foundation, the Israel Cancer Research Fund, the Benoziyo Endowment Fund and E. Harari. I.T. is the incumbent of the Dr. Celia Zwillenberg-Fridman and Dr. Lutz Zwillenberg Career Development Chair.

Author contributions A.G. and I.T. conceived and designed the study, interpreted the data and wrote the manuscript. Data curation for the 3CA website was carried out by A.G., M.T., A.C.G., R.H., D.S., R. Tschernichovsky, C.B., D.K., D.H., R.C.-M., J.L., M.M., A.W., R. Tal. and A.S. Computational analysis relating to scRNA-seq data was mainly carried out by A.G., including defining the MPs and their abundance in malignant and TME cells, inferring MP regulators and correlations with one another. M.T. carried out the computational analysis relating to TCGA bulk RNA data, determined the associations between MPs and various clinical outcomes and interpreted the data. R.H., D.S. and N.G.D. carried out computational analysis relating to spatial transcriptomics and proteomics data. E.S. carried out the subclone analysis. C.B. tested whether MP abundance is influenced by the sequencing platform. T.A. determined MPs in non-cancer scRNA-seq data. A.C.G. and R.H. carried out the IHC experiments and interpreted the data. T.B., L.N.G.C., T.H., M.R.-G., C.S., T.G., A.T., M.L.S. and S.V.P. contributed reagents and samples. M.T., A.C.G., M.L.S. and S.V.P. reviewed the manuscript and provided feedback. The study was supervised by I.T.

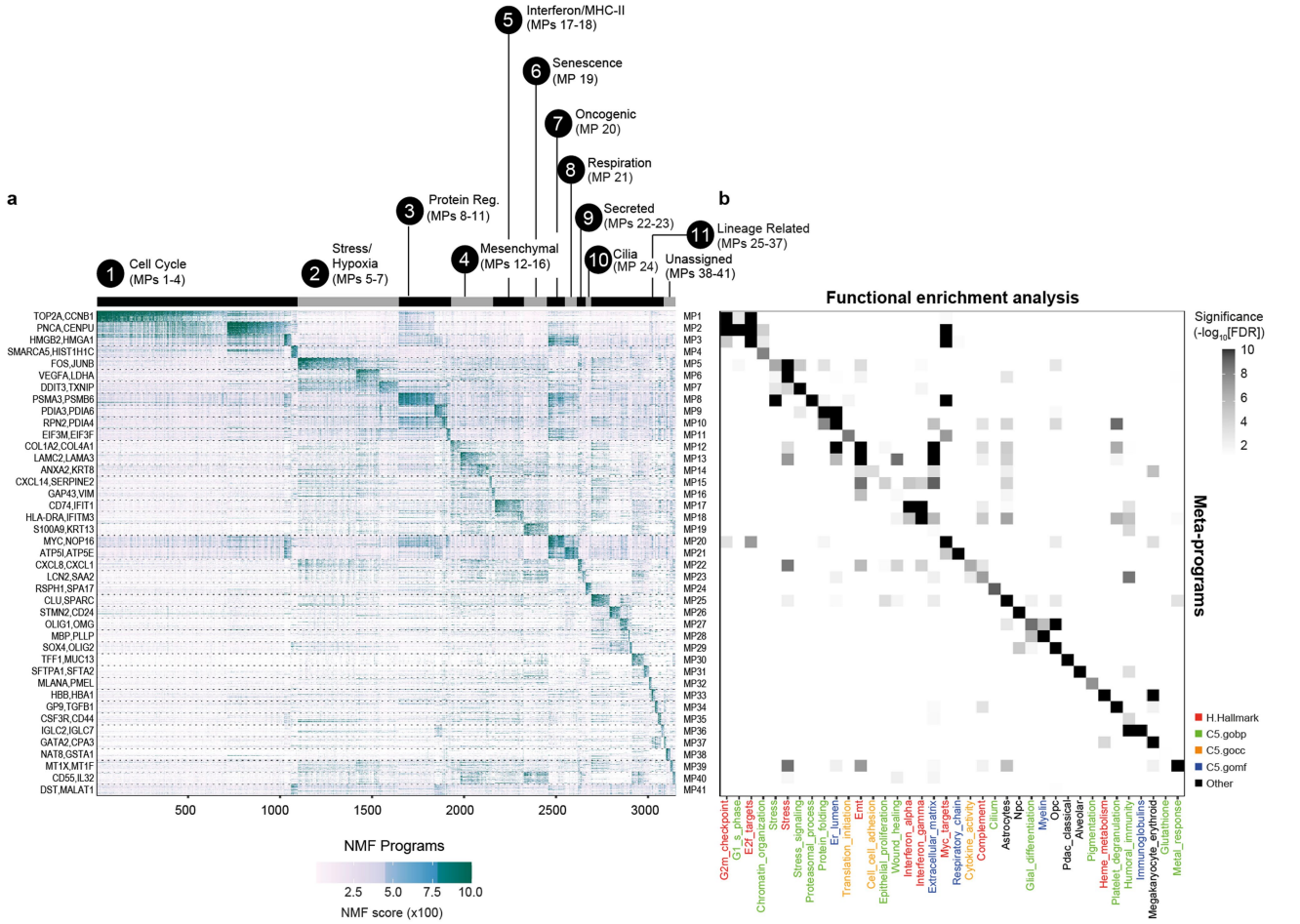
Competing interests I.T. is an advisory board member of Immunitas Therapeutics. M.L.S. is an equity holder, scientific co-founder and advisory board member of Immunitas Therapeutics.

Additional information

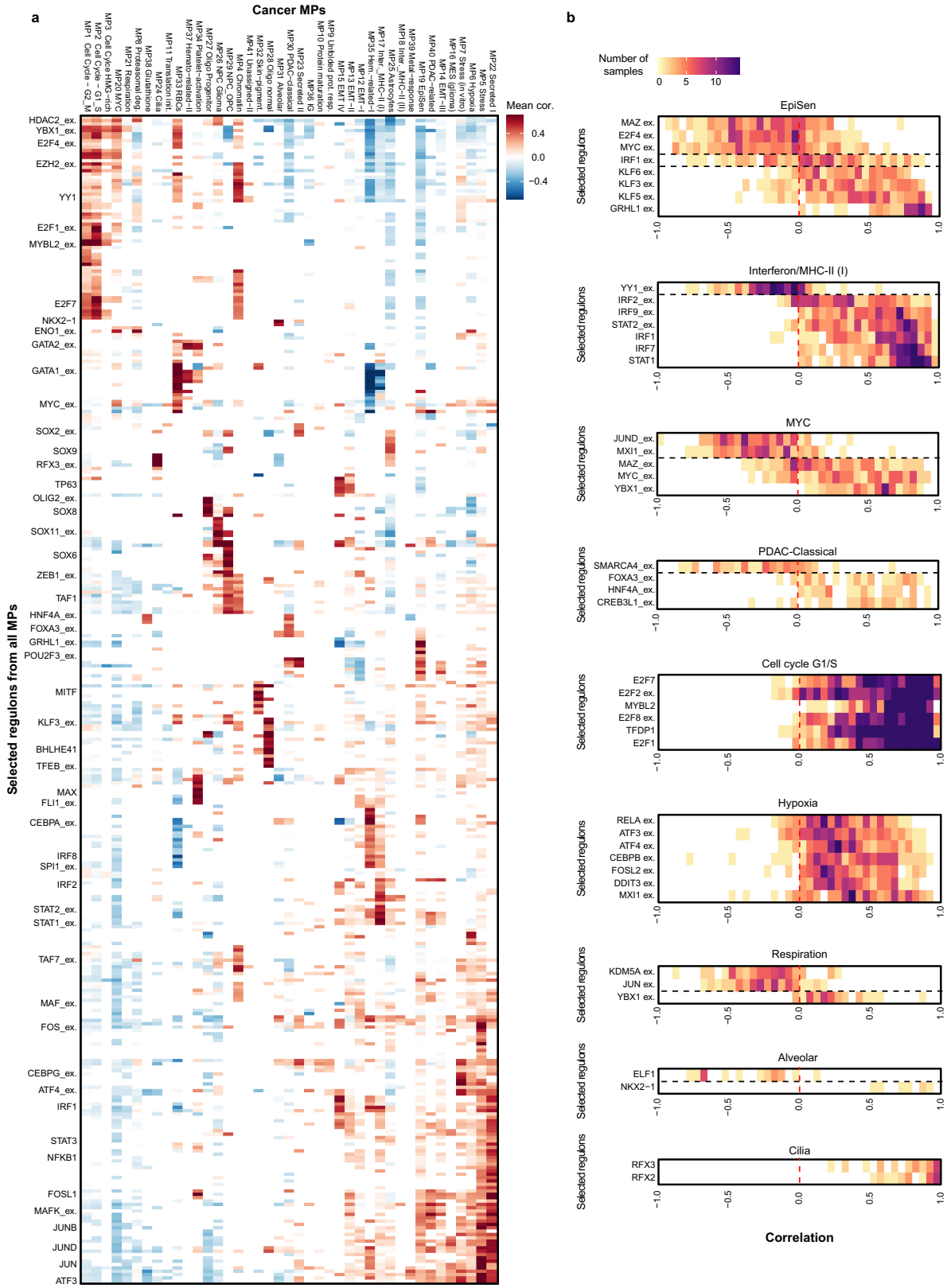
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-06130-4.

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

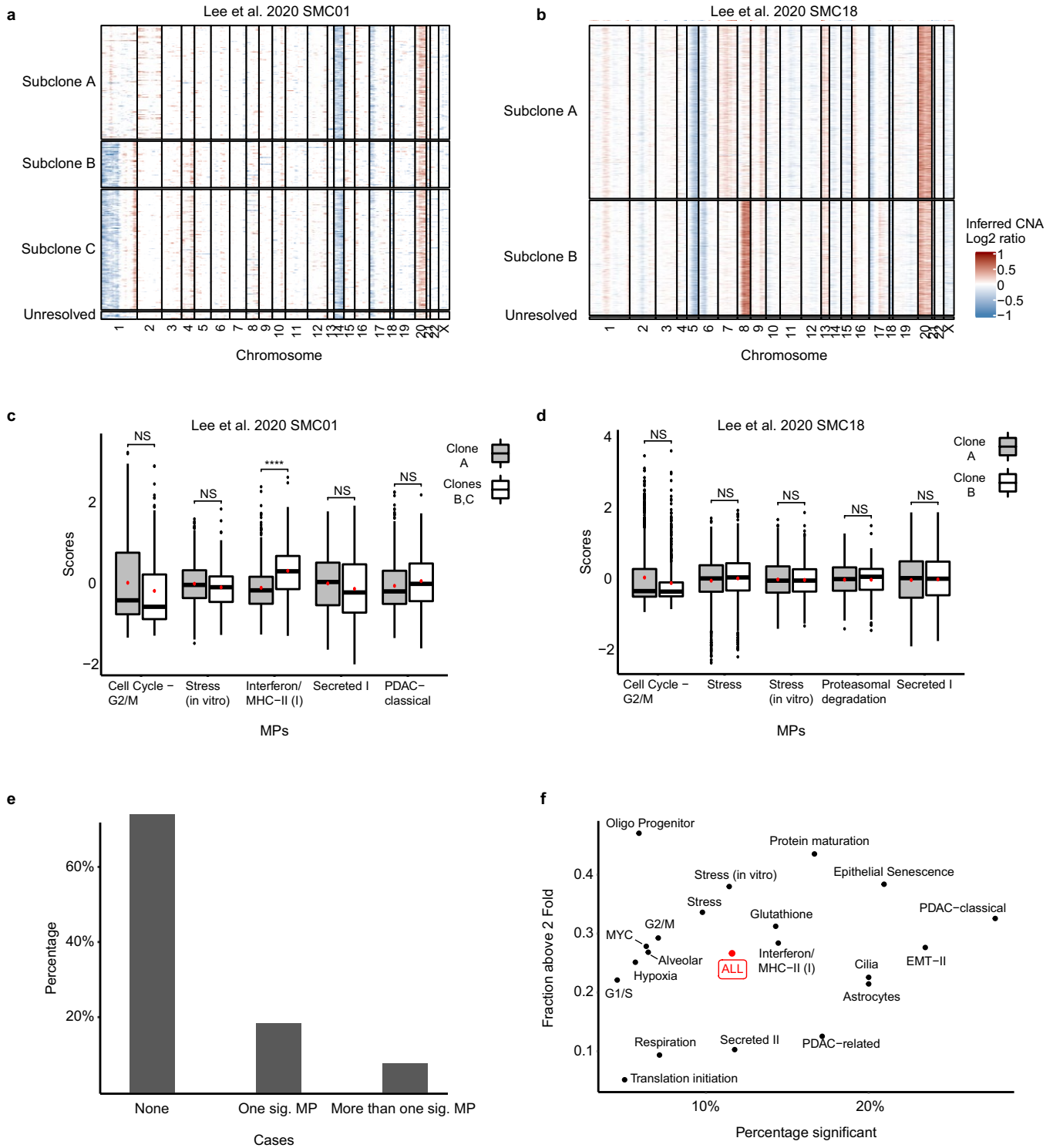
Peer review information *Nature* thanks Dominic Gruen, Sydney Shaffer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



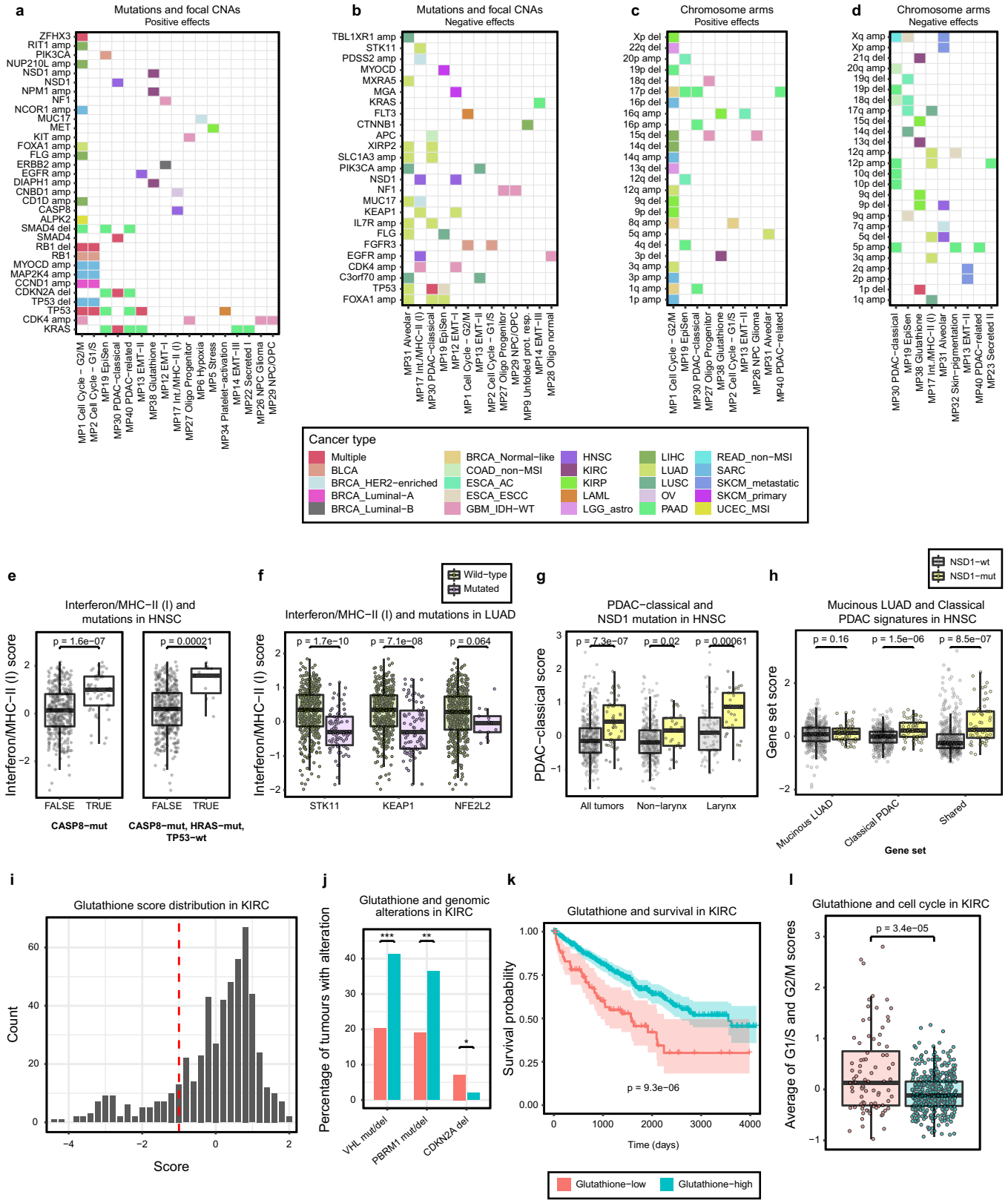
Extended Data Fig. 1 | Genes and functional annotations of malignant MPs. (**a**) Heatmap showing NMF scores for all MP genes (rows) across all robust NMF programs (columns, arranged as in Fig. 2a). Dashed lines separate the genes of different MPs; MP numbers (right) and selected genes (left) are indicated. (**b**) Significance (-log10(FDR)) of the enrichment of selected functional annotations (columns) for all of the MPs (rows, ordered as in **a**). Annotations are colored by their database origin, as indicated (right legend). See Table S3 for additional functional annotations.



Extended Data Fig. 2 | MP regulators inferred by SCENIC. (**a**) Heatmap showing the mean correlations between regulon scores and MP scores (correlations were calculated separately within each sample, and then averaged across samples). Regulons were included if they had a consistent association with at least one MP, defined as correlation above 0.5 (or below −0.5) in at least 10 samples or in at least 30% of the samples with that MP (see Table S6 for a list of such regulons). (**b**) Distributions of the correlations whose means are shown in **a** for selected regulons associated with nine MPs (each shown in a separate panel). ex., extended.

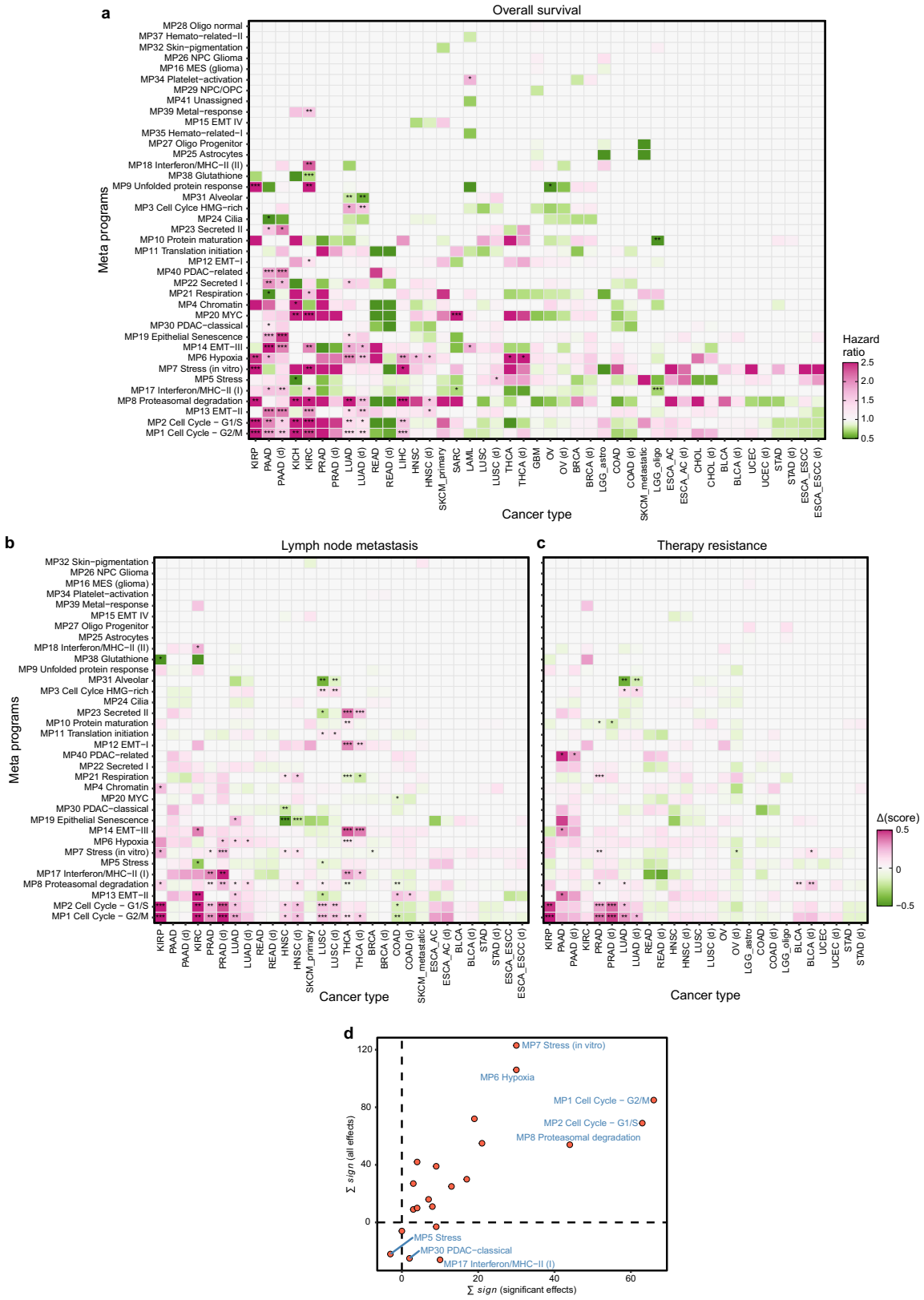


Extended Data Fig. 3 | Genetic subclones explain limited variability in MP expression. (**a**-**b**) CNA patterns for two samples from the Lee et al. 2020 colorectal study, in which 3 and 2 subclones were detected, respectively. Few cells were unresolved with respect to subclone assignment. (**c**-**d**) Boxplots depicting MP scores across cells in subclones detected in **a** and **b**, for all MPs detected in the respective samples. Also shown is the significance of the difference between subclones. In sample SMC01, only the Interferon/MHC-II (I) was significantly different between subclone A and the other two subclones; in sample SMC18, none of the MPs was significantly different between subclones. Boxes indicate the median and 1st and 3rd quartiles, while the upper and lower whiskers respectively extend to maximal and minimal values which are no further than 1.5 times the interquartile range from the 3rd and 1st quartiles. **e** Percentage of all 481 detected subclones in which zero, one or >1 MPs were found as significantly differentially expressed in comparison to the other subclones from the same tumor (as exemplified in panel **c**). **f** For each MP, the X-axis shows the percentage of subclones in which that MP was significantly different from other subclones in the same tumor (calculated only among subclones in which the MP is detected); for that subset of significant subclones, the Y-axis shows the mean fraction of cells with at least a 2-fold difference from the average of the tumor: cells with score above 1 for subclones with increased MP expression, or cells with score below −1 for subclones with decreased MP expression. Thus, even in subclones with a significant upregulation/ downregulation, only a minority of cells have larger than 2-fold effect. Red dot ('ALL') indicates the average across all MPs.



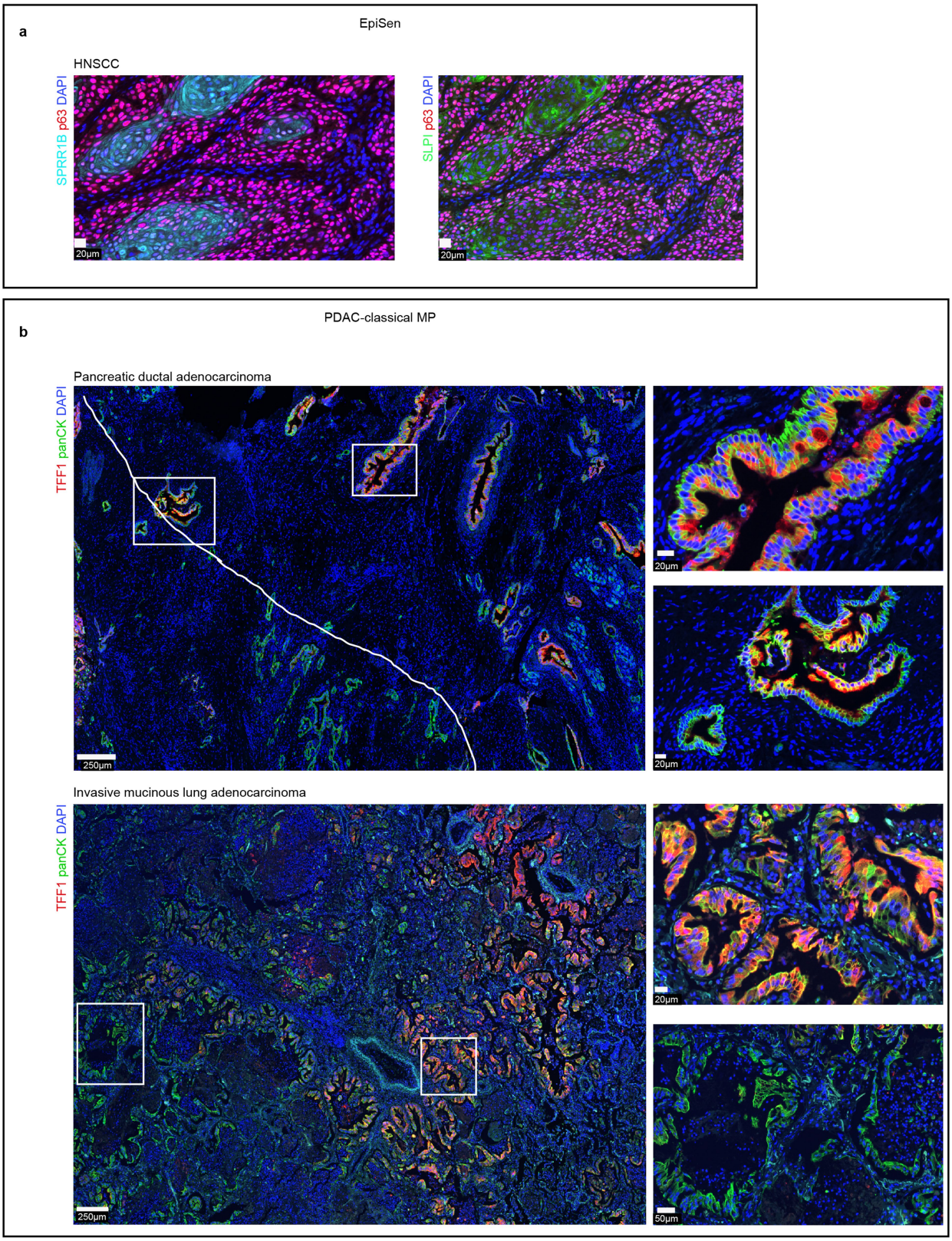
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Associations of MPs with genomic alterations in TCGA data. (**a**-**b**) Heatmaps showing significant associations between MP scores (columns) and mutations or focal amplifications/deletions (rows), colored by the cancer types in which they were observed, or in red if observed in multiple cancer types. Panels separate positive and negative effects, that is, associations where the difference in score between tumors with and without a given genomic alteration is positive or negative, respectively. Significance levels were computed by two-sided t-test and adjusted by Benjamini-Hochberg correction. Exact p values are shown in Table S7. (**c**-**d**) Same as **a**-**b** for chromosome arm gains and losses. (**e**) Boxplot showing scores (Y-axis) for the Interferon/MHC-II (I) meta-program in HNSCC tumors (points, n = 480 biologically independent samples), comparing CASP8-mut tumors, or CASP8-mut/HRAS-mut/TP53-wt tumors, with all others (X-axis). Significance levels were computed by two-sided t-test, without adjustment. Boxes indicate the median and 1st and 3rd quartiles, while the upper and lower whiskers respectively extend to maximal and minimal values which are no further than 1.5 times the interquartile range from the 3rd and 1st quartiles. (**f**) Boxplot showing scores (Y-axis) for the Interferon/MHC-II (I) meta-program in LUAD tumors (points, n = 492 biologically independent samples), comparing tumors with and without STK11 and KEAP1 mutations, respectively. Significance levels were computed by two-sided t-test, without adjustment. Boxes and whiskers are defined as in **e**. (**g**) Boxplot showing scores (Y-axis) for the PDAC-classical meta-program in NSD1-mut and NSD1-wt HNSCC tumors (points), first amongst all HNSCC tumors (n = 480 biologically independent samples), then separately for non-laryngeal and laryngeal tumors (n = 373 and n = 107 respectively). Significance levels were computed by two-sided t-test, without adjustment. Boxes and whiskers are defined as in **e**. (**h**) Boxplot showing scores (Y-axis) in NSD1-mut and NSD1-wt HNSC tumors (points, n = 480 biologically independent samples) for three MP30-related gene-sets: those specific to the Mucinous LUAD tumors, those specific to Classical PDAC tumors, and those shared between Mucinous LUAD and Classical PDAC tumors. Significance levels were computed by two-sided t test, without adjustment. Boxes and whiskers are defined as in **e**. (**i**) Histogram of scores for the Glutathione meta-program in KIRC tumors (n = 532 biologically independent samples). The dashed red line indicates the chosen threshold of −1, used to define Glutathione-low and -high populations. (**j**) Bar plot showing the percentage of KIRC tumors (Y-axis) in Glutathione-low and -high categories (color) having alterations in VHL, PBRM1 and CDKN2A genes (X-axis). Significance levels were computed by two-sided Fisher test, without adjustment, and the notation '*', '**' and '***' indicates $p < 0.05$, $p < 0.01$ and $p < 0.001$, respectively (exact p-values are 0.00064, 0.0025 and 0.020 for the VHL, PBRM1 and CDKN2A comparisons, respectively). (**k**) Kaplan-Meier plot comparing survival probability (Y-axis) across time (X-axis) between KIRC tumors in Glutathione-low and -high categories (color), with p-value computed by log-rank test. Error bands represent 95% confidence intervals. (**l**) Boxplot comparing the average scores for the G1/S and G2/M MPs (Y-axis) between KIRC tumors (points, n = 532 biologically independent samples) in the Glutathione-low and -high categories (color), with P-value computed by two-sided t-test. Boxes and whiskers are defined as in **e**.

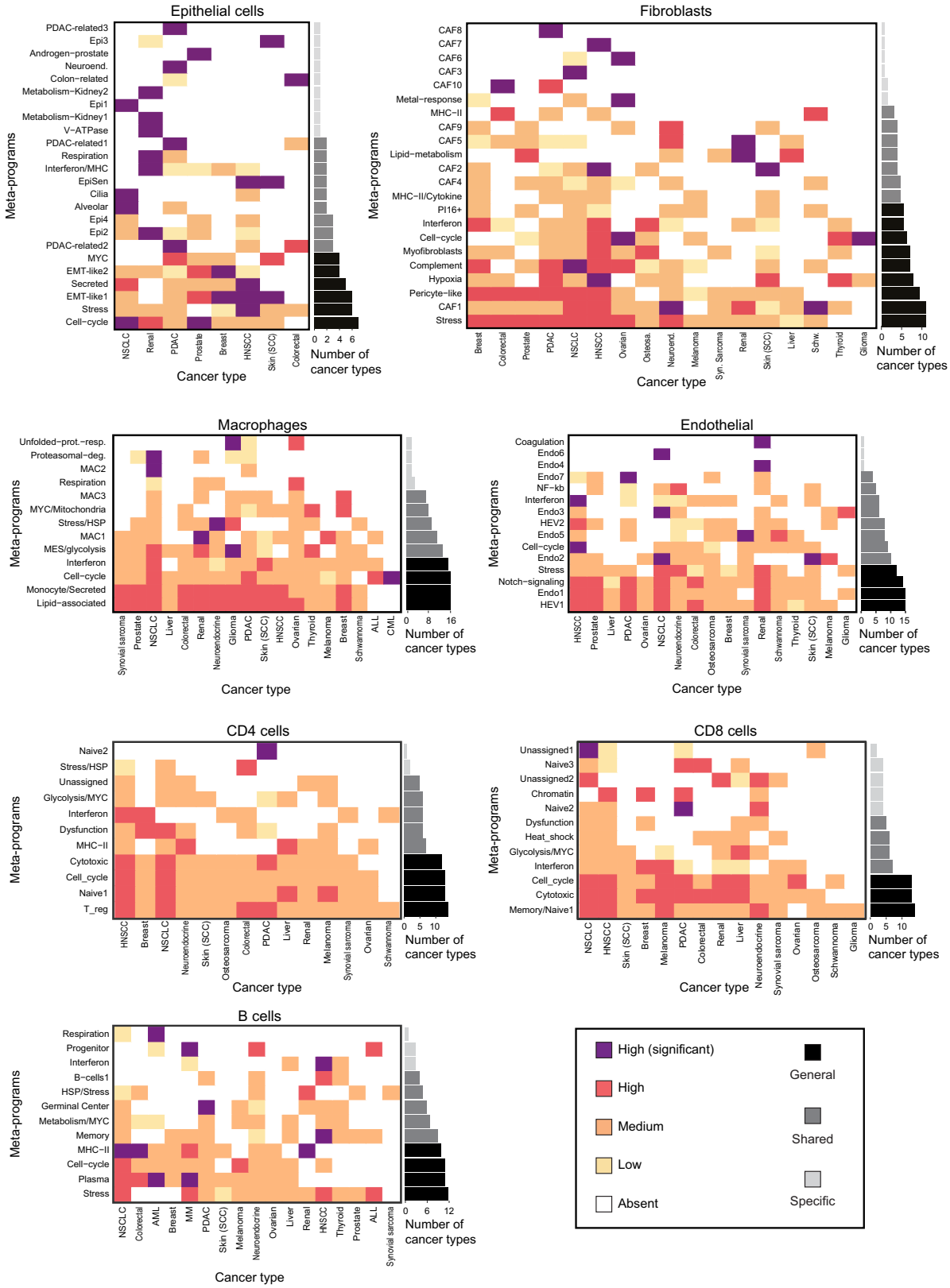


Extended Data Fig. 5 | See next page for caption.

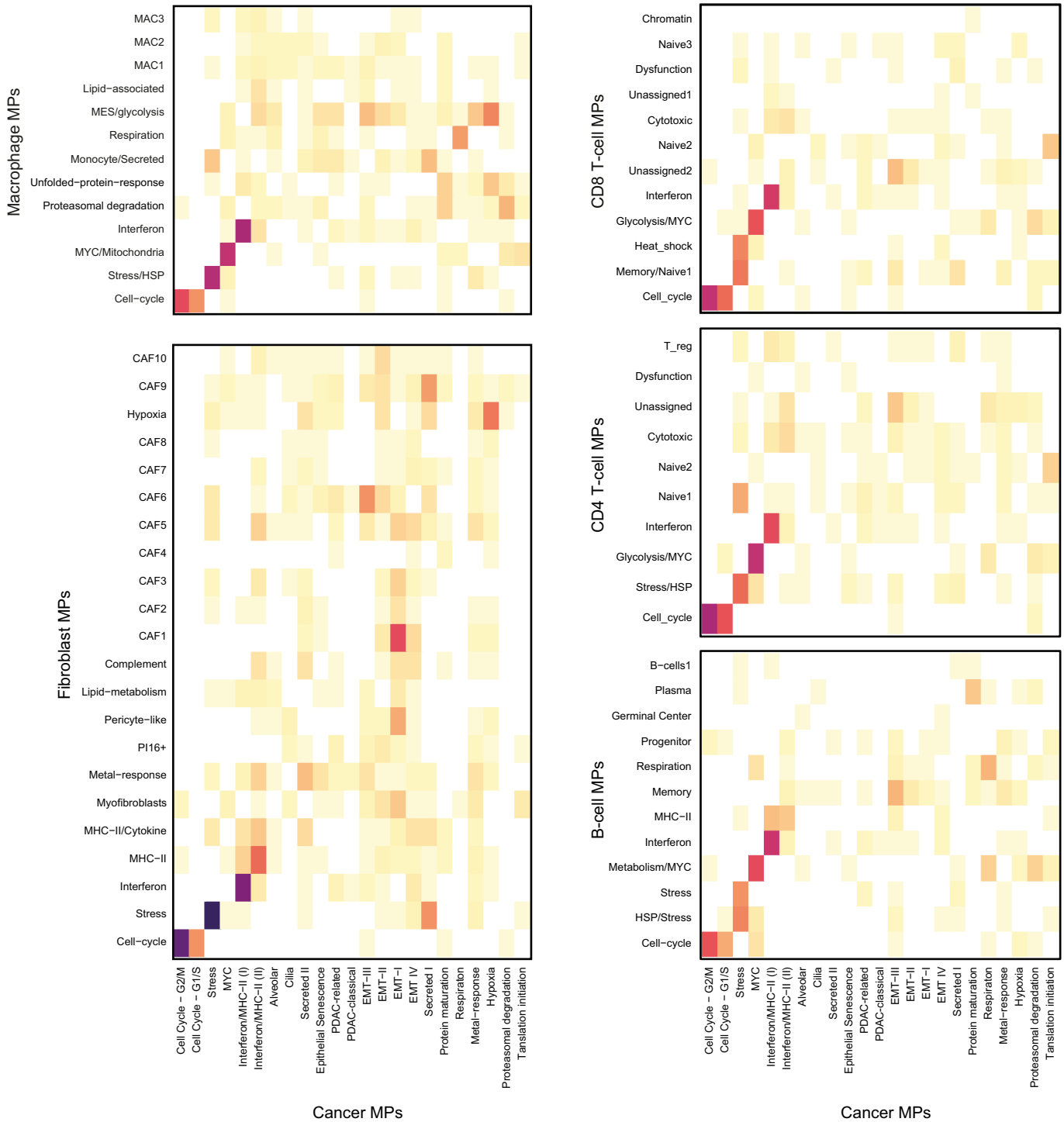
Extended Data Fig. 5 | Associations of MPs with clinical features in TCGA data. (**a**) Heatmap showing the hazard ratio of scores for MPs (rows) with respect to overall survival in each cancer type (columns). Purple and green represent association with worse and better survival, respectively. For relevant carcinoma types, the columns of results from data deconvolved by CIBERSORTx are presented with the suffix "(d)", adjacent to the corresponding column for non-deconvolved data. Significant associations are labelled with '*' for $p < 0.05$, '**' for $p < 0.01$ and '***' for $p < 0.001$. Significance levels were computed via Cox regression and likelihood ratio test, and adjusted by Benjamini-Hochberg correction. Exact p values are shown in Table S8. (**b**-**c**) Heatmaps as in **a** showing the difference in average scores between tumors stratified by lymph node metastasis and therapy resistance, respectively. Significance levels were computed by two-sided t-test and adjusted by Benjamini-Hochberg correction. Exact p values are shown in Table S8. (**d**) Scatterplot showing consistency of associations of MPs with clinical features in TCGA data. Each point represents an MP, its X-value being the sum of the signs (1 for positive effect, −1 for negative effect) of all significant effects (adjusted p value < 0.05, with adjusted p values as in **a**-**c**) for that MP across all cancer types and clinical features, and its Y-value being the sum of the signs of all effects for that MP. Selected points are labelled with their corresponding MPs.



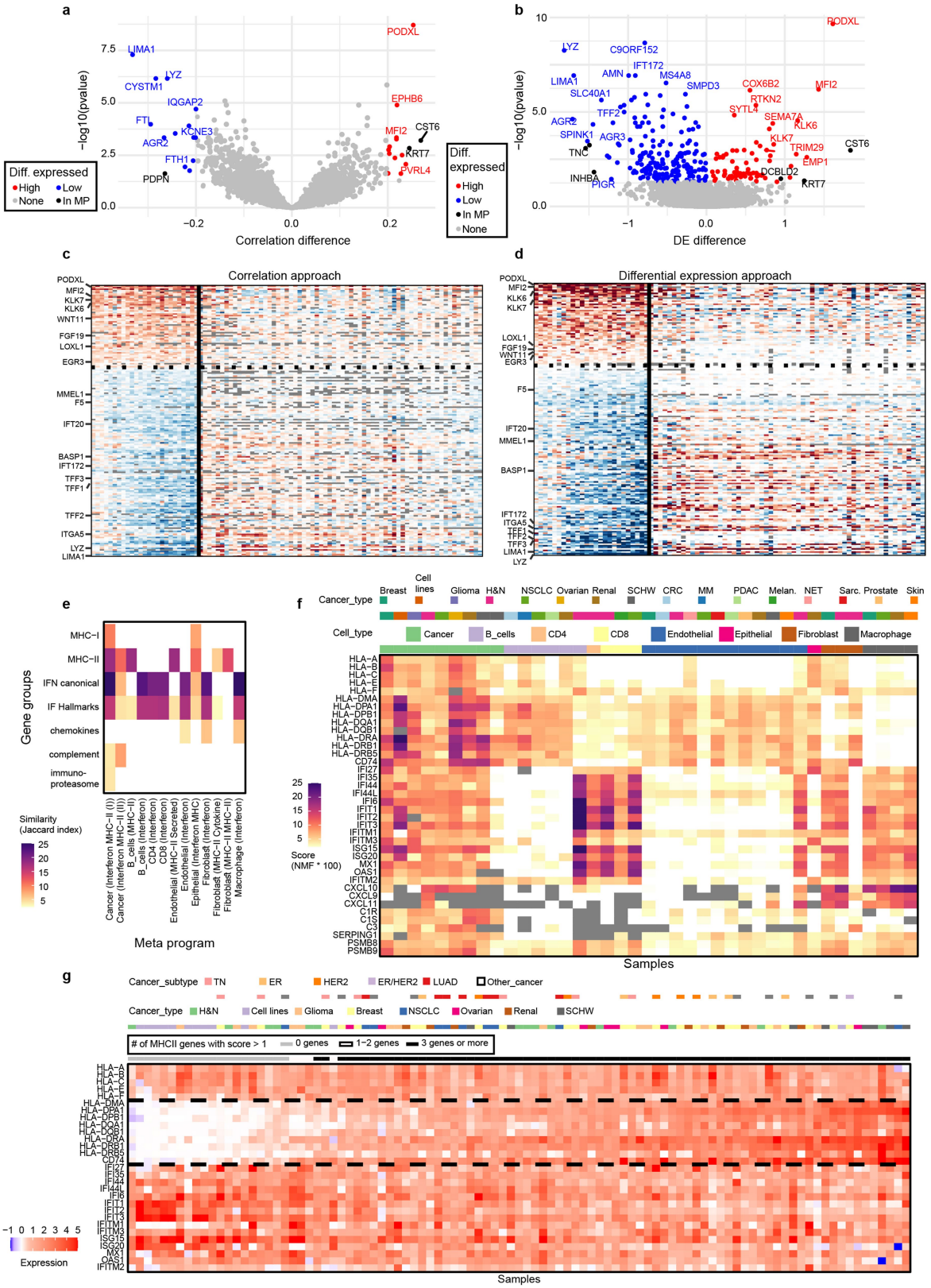
Extended Data Fig. 6 | Experimental validation related to EpiSen and PDAC-classical MPs. (**a**) Example of EpiSen in HNSCC (SPRR1B+) with the senescence-associated secretory-like phenotype (SLPI+) (representative image from 3 independent experiments). (**b**) The PDAC-classical MP (panCK+TFF1+) is also observed in other cancer types such as invasive mucinous lung adenocarcinoma (also shown in Fig. 3e). In both PDAC and invasive mucinous lung adenocarcinoma, subsets of cancer cells expressing TFF1 are spatially zonated (representative images from 4 independent experiments per cancer type).



Extended Data Fig. 7 | Observed vs. expected distribution of MPs across cancer types. Similar plots as in Fig. 3a for the TME cell types. See Methods for exact definitions.

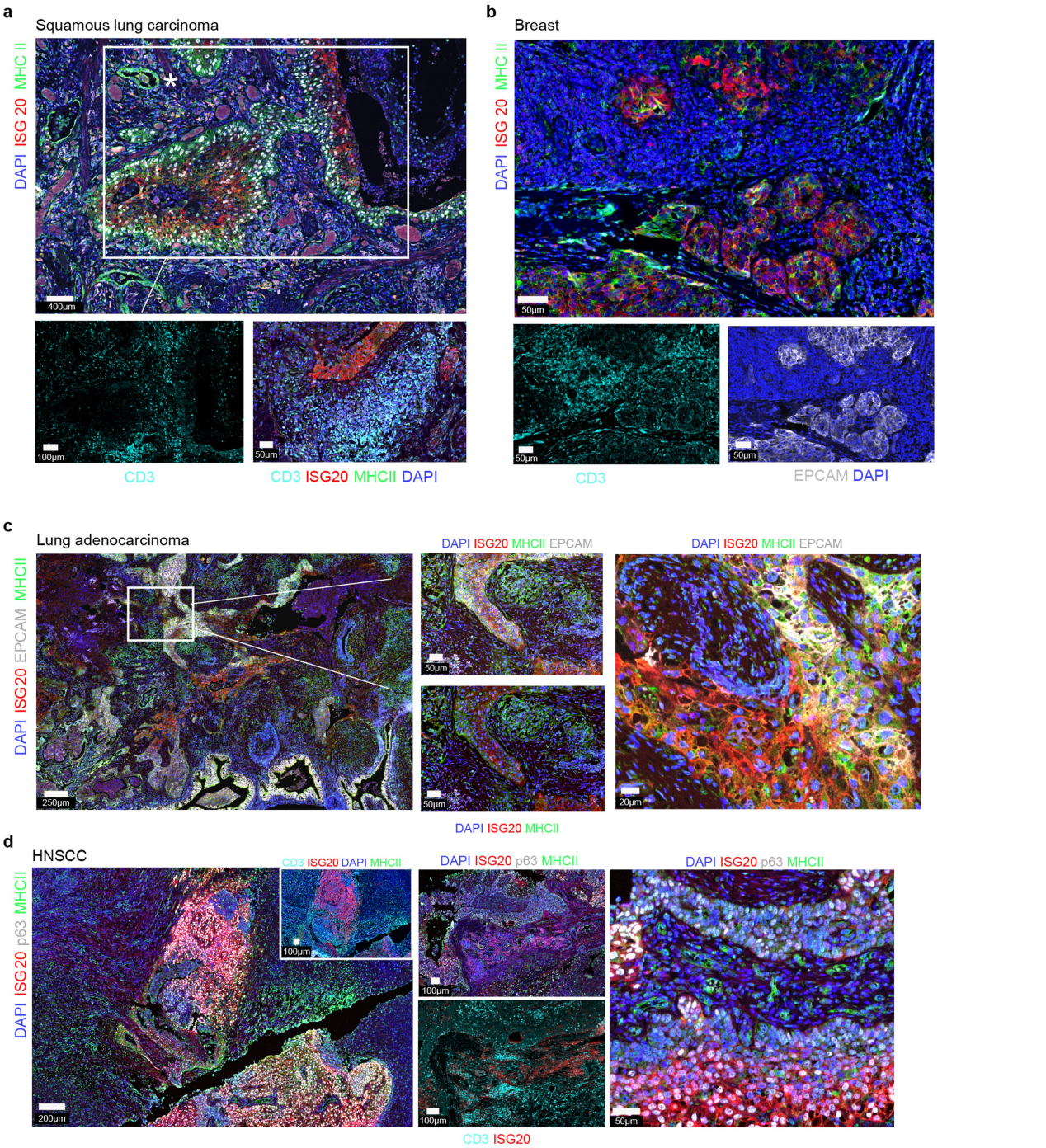


Extended Data Fig. 8 | Jaccard similarity between MPs of malignant cells and TME cells. Similar to Fig. 4b for the remaining non-malignant cell types. The most significant overlap was observed for MPs of the cell-cycle, stress, interferon and MYC.

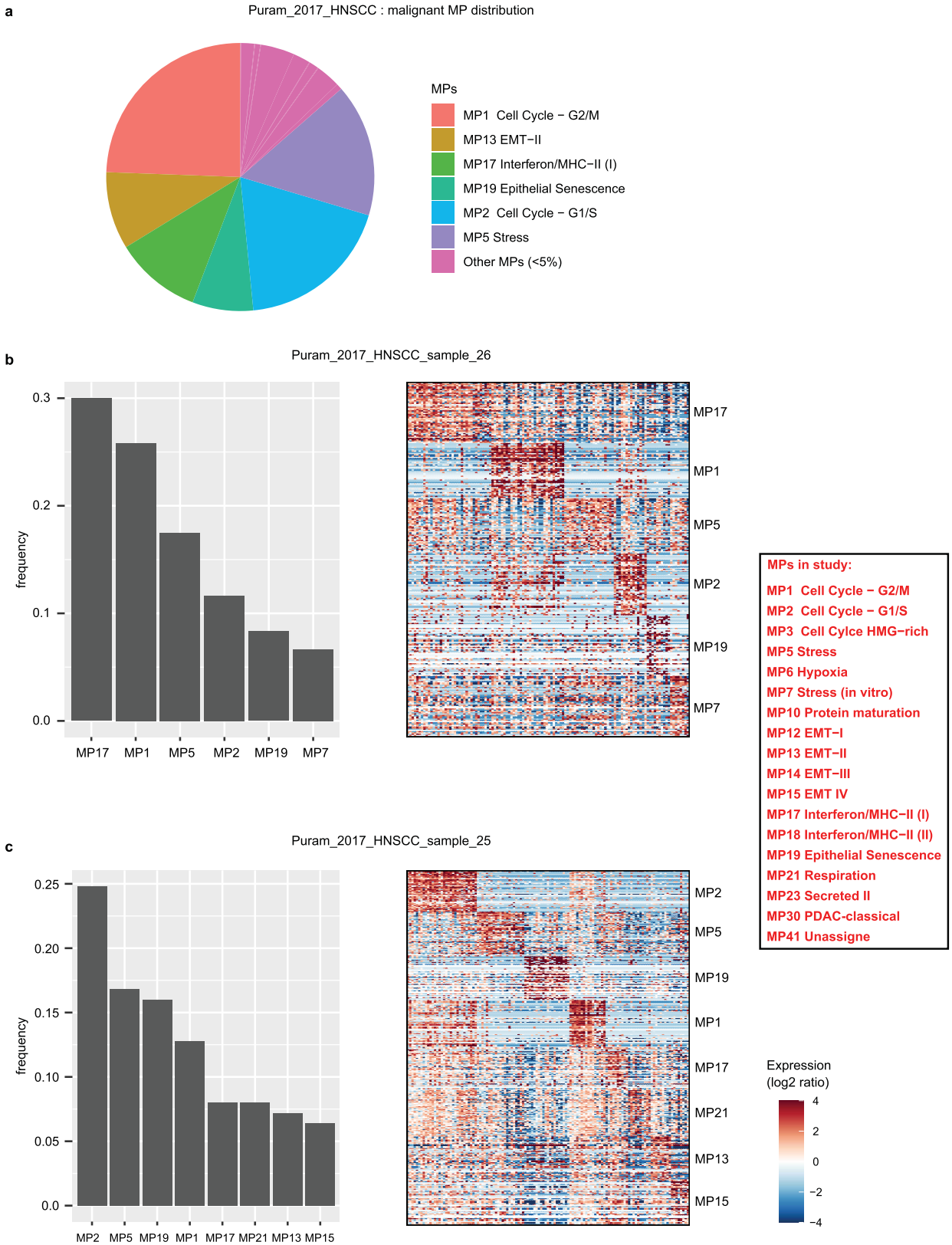


Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Dissimilarities within MPs across different cancer types and cell types. (**a**) Comparing the EMT-II MP between PDAC samples and samples from other cancer types. For each gene, the correlation with the MP scores were calculated in each sample with the MP, and those correlations are compared between PDAC and non-PDAC samples, showing their mean (X-axis) and significance (Y-axis). Red and blue reflect genes with significantly higher or lower MP correlations in PDAC, respectively, calculated by a two-sided t-test adjusted by Benjamini-Hochberg correction. (**b**) Same as **a**, but instead of correlation with MP, the measure for each gene in each sample is the differential expression (log-ratio) between cells with MP score above 1 and those with score below 1. (**c**,**d**) heatmaps corresponding to **a** and **b** that show the actual correlation or differential-expression values across the PDAC samples (to the left of the black vertical line) vs. the other samples, showing genes that were high in both approaches (above the dashed line) vs. genes that were low in both approaches. The genes in both heatmaps are the same but are sorted differently (in decreasing order according to their average in PDAC samples). (**e**) The overlap between MHC-II or Interferon related MPs from different cell types and several curated gene-sets. The only MP with considerable overlap to both MHC-II and Interferon response came from cancer. (**f**) NMF scores of samples that participated in **e** averaged according to cancer type. Grey indicates missing values. (**g**) Average expression of MHC and interferon response genes (rows) across samples (column) in which the Interferon/MHC-II(I) MP was detected. In each sample, the expression was averaged only among cells that scored above 1 to the MP. Samples are sorted by the average difference between MHC-II and Interferon genes. Genes above the top broken line reflect MHC-I, genes between the broken lines reflect MHC-II, and genes below the bottom broken line reflect Interferon response.



Extended Data Fig. 10 | Coupling and decoupling of MHC-II and Interferon response. In malignant cells, MHC-II and interferon response are often co-expressed, unlike in non-malignant cells where they are uncoupled (e.g. asterisks in **a**, right panel in **d**). Cancer cells co-expressing MHC-II and interferon response frequently form an expression gradient in which the cancer cells co-express both markers on a continuum (upper panels in **a** and **b**; panel **c**; left panel in **d**). T cells often cluster around cancer cells with high expression of interferon response and MHC class II (left lower panels in **a**; left lower panel in **b**; left and middle lower panels in **d**) (representative images from 3 independent experiments per cancer type).



Extended Data Fig. 11 | MP distribution across the Puram 2017 HNSCC study. (**a**) Pie chart depicting the proportions of malignant cell assignment to malignant MPs combining all samples in the study. Each cell was assigned to a single MP to which it scored the highest (given that the maximal score was larger than 1). MPs with total proportions of less than 5% (across the whole study) are not shown. (**b**,**c**) Two examples for the MP distribution across samples (samples 26 and 25, bar plots). Heatmaps show the expression of the MPs per sample, with the MP genes in the rows and cells in the columns. Legend lists all MPs that had a maximal score above 1 in at least 5% of the cells in at least one sample (including samples not shown). See Methods (and code availability) for exact definition and code.

nature portfolio

Corresponding author(s): Itay Tirosh

Last updated by author(s): Apr 3, 2023

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	Data was downloaded and further curated using custom code written in R (version 4.1.1). We used a custom approach for initial data filtering and preprocessing as described in more detail in the Methods. Custom code is available on Github in https://github.com/tiroshlab/3ca/tree/main/ITH_hallmarks
Data analysis	We used open source code and software for some of the analysis: we used SCENIC ersion 1.3.1 for inferring gene regulators, inferCNA for CNA analysis (https://github.com/jlaffy/infercna), Cytoscape version 3.9.1 for graph visualization, the 'NMF' R package (https://cran.r-project.org/web/packages/NMF/index.html) and QuPath version 0.3.2 for analyzing IHC slide images. We used custom written code using R version 4.1.1 for multiple analyses, including standard QC steps, cell annotations, clustering, visualization, MP-generation. Relevant code can be found on github in https://github.com/tiroshlab/3ca/tree/main/ITH_hallmarks

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

All curated data is available at (https://www.weizmann.ac.il/sites/3CA), aside from samples from unpublished studies that will be added when possible or cases in which sharing permission from the authors was denied.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	We used mainly published data, as described above, which included samples from males and females. To our knowledge, the data was not deliberately enriched for a particular sex in none of the studies, and other than specific cancer types which are sex specific (e.g. breast cancer or prostate cancer), the data was evenly distributed. We included any available metadata, including sex, for each sample, which is provided in the 3CA website.
Population characteristics	Similar to sex, we made major efforts to include other sample information, including age, treatment status and more. All collected information can be found on the 3CA website.
Recruitment	We prioritized high quality and available data for our analysis. Although we put major efforts to obtain data from a variety of tumor types, there are some types that are under-represented in our compendium (e.g. endometrial cancer or gastric cancer). Hence, we will be able to generalize our findings to these types after accessing the relevant data (work in progress).
Ethics oversight	Ethics protocols were typically obtained by the original authors. For unpublished data that was sequenced by our lab or collaborators we used....For the sample sections used for the IHC staining experiments we used....

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	We curated 1163 samples from 77 studies. Following QC, we typically used all available samples within each study. Depending on the question we tried to address, analyses were performed within each study, across samples from the same cancer type, or across samples that participated in each MP.
Data exclusions	We excluded several samples in which the malignant cell annotations were questionable, based on pre-established criteria for CNA signal and correlation, as described in the Methods. After generating MPs, we also excluded a few programs that were enriched with low quality genes (ribosomal or mitochondrial).
Replication	We believe that our findings are highly reproducible. First, our definitions for MPs were derived using a very large dataset that was sequenced by dozens of different groups using different methods. We used standard IHC staining in multiple samples of different cancer types to validate some of our findings, and also further validated our findings using published CODEX and VISIUM data, and TCGA data.
Randomization	Randomization was not relevant to our study since we aimed to infer ITH in an unsupervised manner based on a large compendium of available published data. We prioritized studies which utilized high quality data according to our judgment and that sequenced many samples. We put special effort to include also rarer cancer types, which resulted in 24 cancer types altogether.
Blinding	Blinding was not possible or relevant in this study since we did not perform a clinical trial or any other experiment/analysis that might require blinding, but rather curated all scRNAseq data that was published as part of other studies and available to us.

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

Methods

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

Antibodies

Antibodies used

We used antibodies for our IHC validation experiments (Table S15):
- mouse anti-human MHC class II; Abcam; ab55152; donkey anti-rabbit Cy3;Jackson; 711-165-152
- rabbit anti-human ISG20; Abcam; ab154393; donkey anti-goat AlexaFluor 647; Jackson; 705-605-003
- goat anti-human SMA; LSBio; LS-b3933-50; donkey anti-rat AlexaFluor 647; Jackson; 712-605-150
- goat anti-human p63; R&D; AF1916; donkey anti-goat Cy3; Jackson; 705-165-003
- rat anti-human CD3; LSBio; LS-B8765-50; donkey anti-rabbit FITC; Jackson; 711-095-152
- goat anti-human EPCAM; Millipore; AF960; donkey anti-mouse FITC; Jackson; 715-095-150
- rabbit anti-human TFF1; Abcam; ab92377
- mouse anti-human pan-cytokeratin; Abcam; ab86734
- rat anti-human MECA79 ; NovusBiologicals; NB100-77673
- rabbit anti-human C3; Abcam; ab97462
- rabbit anti-human SLPI; Invitrogen; PA582990
- rabbit anti-human SPRR1B; LSBio; LS-C161474-400
- mouse anti-human LAMC2; Novus Biologicals; NBP2-42388

Validation

Below we include respective information on antibody, cat. number or clone, manufacturer, species, expected expression and validation information:
1) Pan -Cytokeratin; ab234297; Abcam; rabbit; epithelial, cancer; tested in human skin tissue by Abcam, tested in tumor-adjacent normal epithelium in HNSCC by us
2) EPCAM; AF960; Millipore; goat; epithelial, cancer; tested in PDAC, HNSCC, and tumor-adjacent normal epithelium by us
3) p63 ;AF1916; R&D; goat; squamous, cancer; tested in human breast by R&D, tested in HNSCC by us
4) SLPI; PA582990; Invitrogen; rabbit; epithelial, cancer; tested in human cervix, uterine, and skeletal muscle by Invitrogen, tested in HNSCC by us
5) SPRR1B; LS-C161474-400; LSBio; rabbit; epithelial, cancer; tested in human breast carcinoma by LSBio, tested in HNSCC - by us
6) ISG20; ab154393; Abcam; rabbit; immune; tested in non-Hodgkin's lymphoma and tumor xenograft by Abcam,tested by us in human lymph node
7) CD31; ab9498; Abcam; mouse; endothelial; tested in human tonsil and lung by Abcam, tested in human placenta by us
8) C3; ab97462; Abcam; rabbit; immune tested in mouse brain by Abcam, tested in human lymph node and tonsil by us
9) CD3; LS-B8765-50; LSBio; rat; T cell; tested in human thymus by LSBio, tested in human lymph node, tonsil, and spleen by us
10) aSMA; LS-b3933-50; LSBio; goat; myofibroblast/CAF; tested by us in human spleen and placenta
11) TFF1/Anti-Estrogen Inducible Protein pS2; ab92377; Abcam; rabbit; mucosal glands/goblet cells, cancer; tested in human breast and ovarian carcinoma by Abcam, tested in PDAC by us
12) MHC class II; ab55152; Abcam; mouse; immune; tested in human bowel tissue by Abcam, tested in human lymph node by us
13) MECA79; NB100-77673; Novus Biologicals; rat; high endothelial venules (HEVs); tested in lymph node and tonsil by Novus, tested in human tonsil by us
14) LAMC2; NBP2-42388; Novus Biologicals; mouse; pEMT (cancer), ECM; tested in fallopian tube and liver by Novus, HNSCC in Puram et al. 2017 https://doi.org/10.1016/j.cell.2017.10.044

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)

State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.

Authentication

Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.

Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	We used a few published mouse and PDX models, as described in Table S1: Ebinger et al 2016 (ALL PDX models), Ireland et al 2020 (SCLC mouse model), Yao et al 2020 (ESCC mouse model), and unpublished GBM mouse model
Wild animals	The study did not involve wild animals.
Reporting on sex	We used mainly published datasetes, as described above, which included samples from males and females. To our knowledge, the data was not deliberately enriched for any sex in none of the studies, and other than specific cancer types which are sex specific (e.g. breast cancer or prostate cancer), the data was evenly distributed. We included any available metadata, including gender, for each sample, which is provided in the 3CA website.
Field-collected samples	This study did not involve samples collected from the field.
Ethics oversight	In most cases we used published data that did not require further ethical approval. For unpublished data that was sequenced by our lab or collaborators we used... For samples used for IHC staining we used....

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