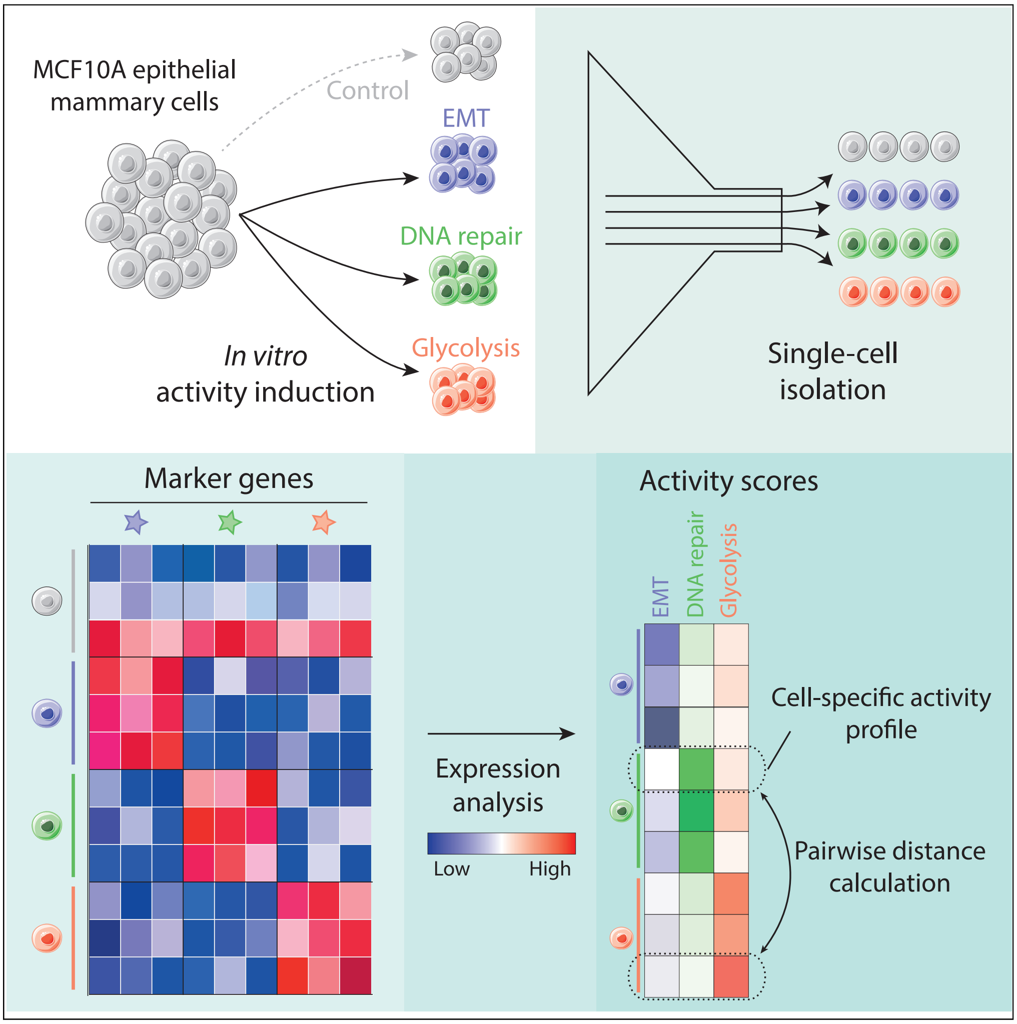
iScience

CellPress
OPEN ACCESS

Article

Assessing Cell Activities rather than Identities to Interpret Intra-Tumor Phenotypic Diversity and Its Dynamics



Laloé Monteiro, Lydie Da Silva, Boris Lipinski, Frédérique Fauvet, Arnaud Vigneron, Alain Puisieux, Pierre Martinez

pierre.martinez@lyon.unicancer.fr

HIGHLIGHTS

Cells categorized as having the same identity can perform different activities

Single-cell expression data can be used to infer the activities cells take part in

Activity profiles provide a basis to measure phenotypic cell-cell divergence

Cell activity can quantify intra-tumor heterogeneity more fully than identity

Monteiro et al., iScience *23*, 101061
May 22, 2020 © 2020 The Authors.
https://doi.org/10.1016/j.isci.2020.101061

iScience

CellPress OPEN ACCESS

Article

Assessing Cell Activities rather than Identities to Interpret Intra-Tumor Phenotypic Diversity and Its Dynamics

Laloé Monteiro,[1,2] Lydie Da Silva,[1,2] Boris Lipinski,[1,2] Frédérique Fauvet,[1] Arnaud Vigneron,[1] Alain Puisieux,[1] and Pierre Martinez[1,3,*]

SUMMARY

Despite advances in single-cell and molecular techniques, it is still unclear how to best quantify phenotypic heterogeneity in cancer cells that evolved beyond normal, known classifications. We present an approach to phenotypically characterize cells based on their activities rather than static classifications. We validated the detectability of specific activities (epithelial-mesenchymal transition, glycolysis) in single cells, using targeted RT-qPCR analyses and *in vitro* inductions. We analyzed 50 established activity signatures as a basis for phenotypic description in public data and computed cell-cell distances in 28,513 cells from 85 patients and 8 public datasets. Despite not relying on any classification, our measure correlated with standard diversity indices in populations of known structure. We identified bottlenecks as phenotypic diversity reduced upon colorectal cancer initiation. This suggests that focusing on what cancer cells do rather than what they are can quantify phenotypic diversity in universal fashion, to better understand and predict intra-tumor heterogeneity dynamics.

INTRODUCTION

Somatic evolution naturally occurs in all multicellular organisms, as cells accumulate genetic alterations upon replication and exposure to mutagenic environments (Gatenby and Brown, 2017). This can eventually select for highly adapted cells breaking free of the constraints imposed by homeostatic regulation on proliferation and motility, leading to cancer (Greaves and Maley, 2012; Trigos et al., 2018). This evolutionary nature implies that cancer cells originating from a common ancestor can display extensive diversity at both the genetic and phenotypic levels (Gerlinger et al., 2012). This diversity, known as intra-tumor heterogeneity (ITH) (McGranahan and Swanton, 2015), can foster resistance and facilitate adaptation upon the environmental changes induced by therapeutic regimens (Nowell, 1976). To limit the risk of resistant populations emerging upon treatment and predict cancer evolution, it is thus necessary to better understand the dynamics of ITH (Lässig et al., 2017; Maley et al., 2006).

Being able to follow the evolution of ITH first implies that one should be able to reliably quantify it. Although there exist multiple methods for genetic ITH thanks to alteration frequencies in the population (Nik-Zainal et al., 2012; Andor et al., 2014; Fischer et al., 2014; Martinez et al., 2017; Williams et al., 2018), phenotypic ITH is more challenging. Many studies have relied on the identification of static classifications (Frazer et al., 2007; Patel et al., 2014; Zhang et al., 2019), often based on lineage markers (Almendro et al., 2014; Nguyen et al., 2018), allowing the calculation of standard diversity metrics such as the Shannon (Bertucci et al., 2019), Simpson (Martinez et al., 2016), or GINI (Ferrall-Fairbanks et al., 2019) indices. Although these classifications make perfect sense in the context of normal tissue homeostasis, they may not be relevant in cancer cells bypassing the host's regulatory mechanisms through abnormal transcriptional programs. Cancer cells drift away from well-characterized normal phenotypes according to evolutionary trajectories specific to each tumor. They, however, display strong convergence at the phenotypic level, with key pathways and cellular activities recurrently dysregulated across both patients and tumor types (Hanahan and Weinberg, 2000, 2011). Aside from static subtype classifications, other methods have focused on expression variation among specific gene sets (Davis-Marcisak et al., 2019) and uneven repartition of expressed transcripts per gene (Hinohara et al., 2018). Yet, there is no golden standard approach to quantify phenotypic diversity in cancer.

[1] Univ Lyon, Université Claude Bernard Lyon 1, INSERM 1052, CNRS 5286, Centre Léon Bérard, Cancer Research Center of Lyon, Lyon, 69008, France, INSERM U1052, Cancer Research Center of Lyon, Lyon, F-69008, France

[2] These authors contributed equally

[3] Lead Contact

*Correspondence: pierre.martinez@lyon.unicancer.fr

https://doi.org/10.1016/j.isci.2020.101061

Here we investigated the feasibility of predicting the activities that a single cell partakes in and the relevance of considering them as traits to describe the cell's overall phenotypic profile. We performed targeted single-cell experiments on three cellular activities induced *in vitro* (epithelial-mesenchymal transition, DNA repair, glycolysis), which suggested that targeted panels can reliably identify the presence of a given activity from single cell RNA expression data. To expand on this limited data, we then analyzed 50 hallmark activity signatures from the Molecular Signature database (MSigDB) in eight publicly available single-cell tumor datasets. We used leave-one-out procedures to avoid overfitting, along with Principal Component and clustering analyses to account for the redundancy among the 50 activities. By using activity-based phenotypic profiles to quantify cell-cell divergence and sample-wise phenotypic diversity, we report that such an approach is relevant in pan-cancer fashion. It could furthermore recapitulate diversity indices based on known population structures, independently of tissue and cell types. Finally, such a method allowed a glimpse into the evolutionary dynamics of phenotypic diversity, hinting at the existence of evolutionary bottlenecks reducing phenotypic diversity upon colorectal cancer initiation. Although more work is necessary to provide specific and accurate quantitative tools and software, our results suggest that focusing on cell activities to measure phenotypic ITH can provide a more relevant angle than standard classification and marker-based methods.

RESULTS

Detecting Hallmark Signatures in Single Cells

We assessed the relevance of three MSigDB hallmark gene signatures in single cells via *in vitro* inductions: epithelial-mesenchymal transition (EMT), DNA repair, and glycolysis. We aimed to take advantage of the higher accuracy of single-cell RT-qPCR compared with whole transcriptome scRNA-seq (Mojtahedi et al., 2016) and designed reduced panels of 9–13 marker genes to detect each activity in single cells (see Methods). To do so, we first analyzed gene expression in 1,036 cell lines samples from the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012) for marker gene discovery and 10,885 pan-cancer samples from The Cancer Genome Atlas (TCGA) (Chang et al., 2013) for cross-validation. The activity-specific markers, respectively, achieved areas under the curve (AUCs) of 0.96, 086, and 0.79 in teasing out the top and bottom scoring TCGA samples for EMT, DNA repair, and glycolysis, respectively (Table S3).This suggested that these reduced gene panels satisfactorily recapitulated the signal from whole-gene set enrichment analyses, implying that analyzing the expression of few marker genes could help quantify the presence of activity-based phenotypic traits in single cells.

We analyzed the expression of 48 selected marker genes in 48 single epithelial mammary cells (MCF10A), in which each activity had been induced or not (12 EMT-induced, 12 DNA-repair-induced, 12 glycolysis-induced, 12 control cells with no induction, Figure 1A). Significantly differentially expressed genes could be identified in all experiments (Figure 1B). We inferred Beta-Poisson expression distributions for each gene in active/inactive conditions, which we used to calculate the likelihood that expression values from marker genes corresponded to cells in which the related activity was induced (Figure 1C). Differentially expressed genes, generalized linear models, and leave-one-out procedures were used to predict cells undergoing each activity induction (see Transparent Methods). We could achieve AUCs of 0.99, 0.72, and 0.86 for, respectively, the EMT, DNA repair, and glycolysis activities (Figures 1D, S2, S3, and Table S4). The absence of expression patterns clearly separating DNA repair cells from the other three types, for most DNA repair genes, impaired prediction for this activity (Figure S3). This targeted experiment, however, suggests that the expression of adequate marker genes can be used to identify whether an activity is present in a given cell with satisfying accuracy.

Whole Transcriptome Cell Activity Scores

Following these targeted *in vitro* results supporting the feasibility of predicting the activities of single cells, we investigated the relevance of an activity-centered approach to quantify phenotypic diversity in high-throughput patient datasets. In the absence of single-cell inference methods tailored to each of the 50 hallmark cell activities, we used standard tools to investigate the behavior of the related signatures in patient data. We used the AUCell (Aibar et al., 2017) software to score the enrichment of all MSigDB hallmark gene sets in all cells from eight datasets (Fan et al. 2018; Filbin et al. 2018; Li et al. 2017; Neftel et al. 2019; Patel et al. 2014; Tirosh, Izar et al. 2016; Tirosh, Venteicher et al. 2016; Venteicher et al. 2017). We normalized these data per set and merged them into a meta-dataset of 50 activity scores per cell in 28,513 cells from different cancer types (see Methods). No major batch effect could be observed as samples did not specifically cluster according to their sets of origin, whereas similar cell types appear to cluster together (Figure 2). However, the most common cell types (T cells, macrophages,

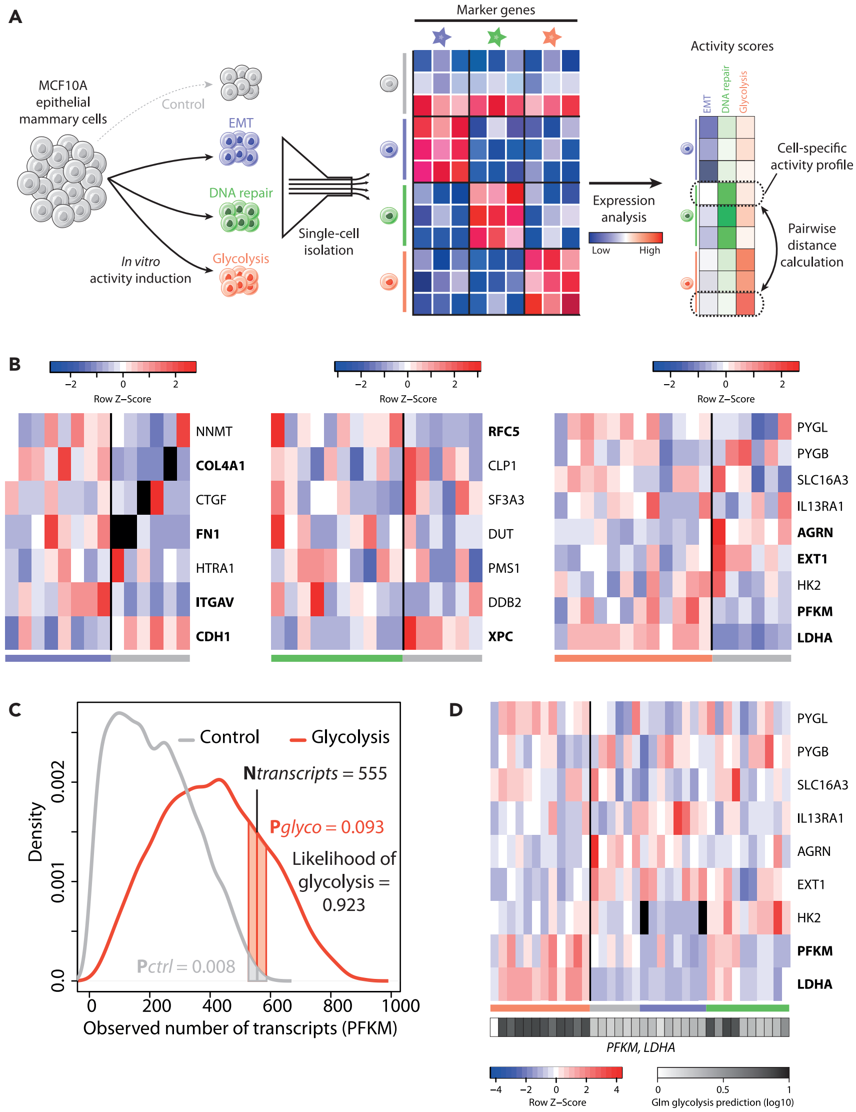


Figure 1. Detection of Selected Activities Induced *In Vitro* Using Single-Cell Expression of Targeted Genes

(A) Overall scheme. EMT (blue), DNA repair (green), and glycolysis (red) activities are induced *in vitro* in MCF10A cells, prior to single-cell analysis and RNA quantification. Targeted marker genes expression is used to assess the likelihood that an activity, considered as a phenotypic trait, is present in a cell. All quantified traits are used to create cell-specific phenotypic profiles and serve as a basis to calculate pairwise cell-cell divergence and overall phenotypic diversity.

(B) Row-normalized single-cell expression for the marker genes of EMT (left), DNA repair (center), and glycolysis (right). Blue: lower expression; red: higher expression. Cells in which the activity was induced are on the left and indicated by colored bars below. Control cells having undergone no induction are on the right and indicated by a gray bar. Significantly differentially expressed genes in bold ($p < 0.05$, BPglm function).

(C) PFKM marker gene expression in glycolysis and control conditions. Blue curve: number of transcripts in cells in which glycolysis cells was induced; gray curve: control conditions. Confidence intervals around the observed values are used to calculate the probability that a value comes from glycolytic (p_{glyco}, blue) and control (p_{ctrl}, gray) conditions. The $p_{glyco}/(p_{glyco} + p_{ctrl})$ ratio gives the likelihood that the observed value comes from a cell in which glycolysis was induced.

(D) Glycolysis prediction in single cells from all four populations: glycolysis (red), EMT (blue), and DNA repair (green) inductions and control (gray). Black and white bar underneath indicates the reported probability of each cell to be glycolytic (log10 scale). Black: missing values.

and malignant cells) segregated into more than one cluster each. This suggests that cells with similar identity tend to behave similarly across batches and tissues but that different subset of activity profiles could also be observed among cells of identical classification.

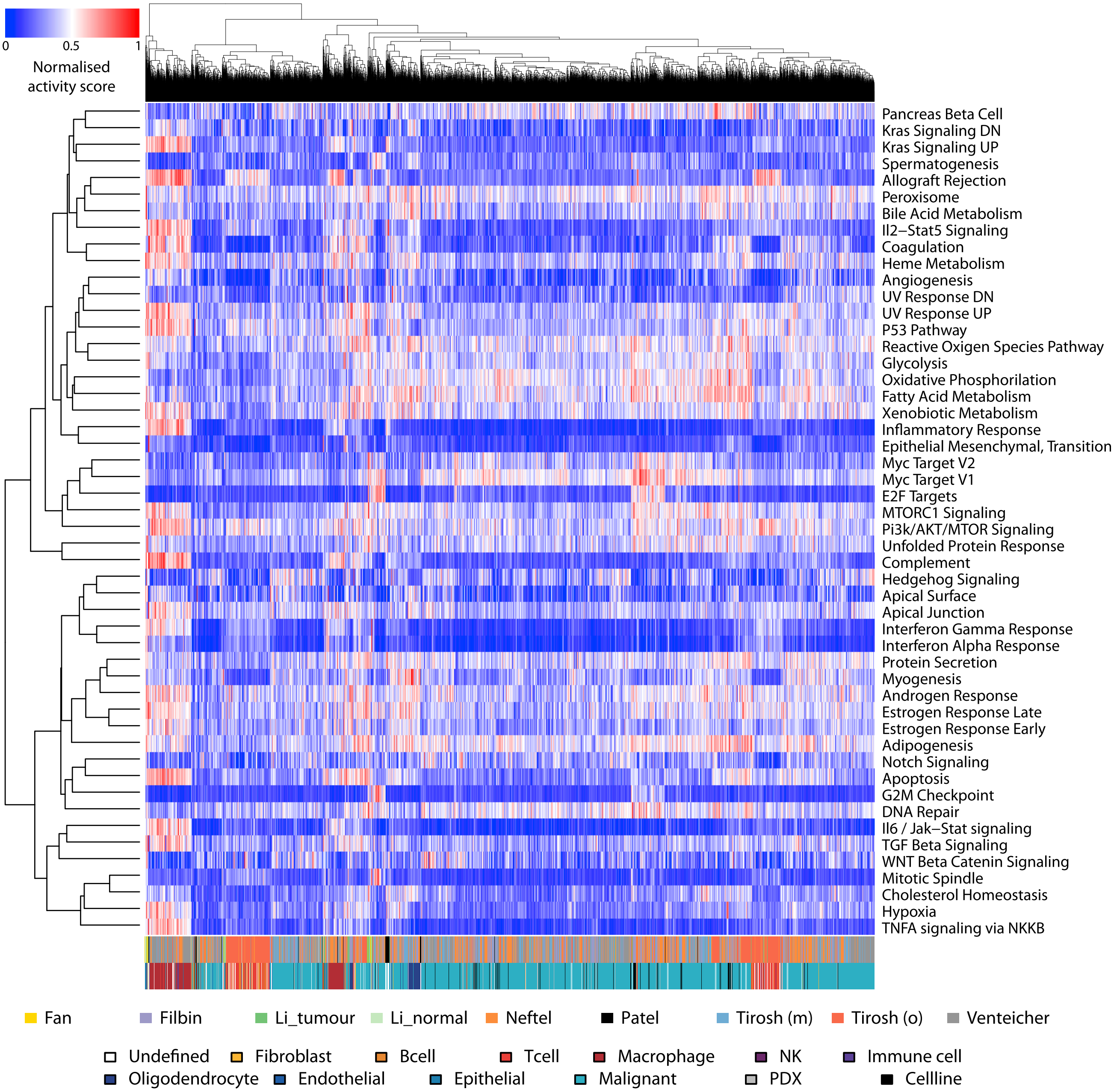


Figure 2. Normalized Activity Scores in the Meta-Dataset

Heatmap of activity scores in the meta-dataset, normalized per activity per set. Dendrograms highlight relationships between activities (left) and cells (top). The dataset of origin of each cell is reported by the bottom color bar. The top row below the score heatmap indicates the dataset of origin of each cell, whereas the bottom one indicates its reported type.

Our analysis, however, revealed extensive redundancy among the 50 activities scored (Figure 3A), suggesting that the signal from the hallmark signatures likely corresponded to fewer than 50 distinct activity-based phenotypic traits. We furthermore assigned cell-cycle phases (G1/S/G2M) to cells using the cyclone software (Scialdone et al., 2015). The cell-cycle phase in which a cell is influences its transcriptome, which can in turn bias cell-type assignment. However, because our approach is cancer oriented and based on cellular activities rather than identities, we considered this information as part of the phenotypic state of a cell and purposely did not correct for it. Cell-cycle phase assignment was found to correlate with the *G2M Checkpoint*, *E2F Targets*, and *Mitotic Spindle* signatures, highlighting that such cycle phase information was indeed taken into account in our phenotypic profiling of cells (Figure S4).

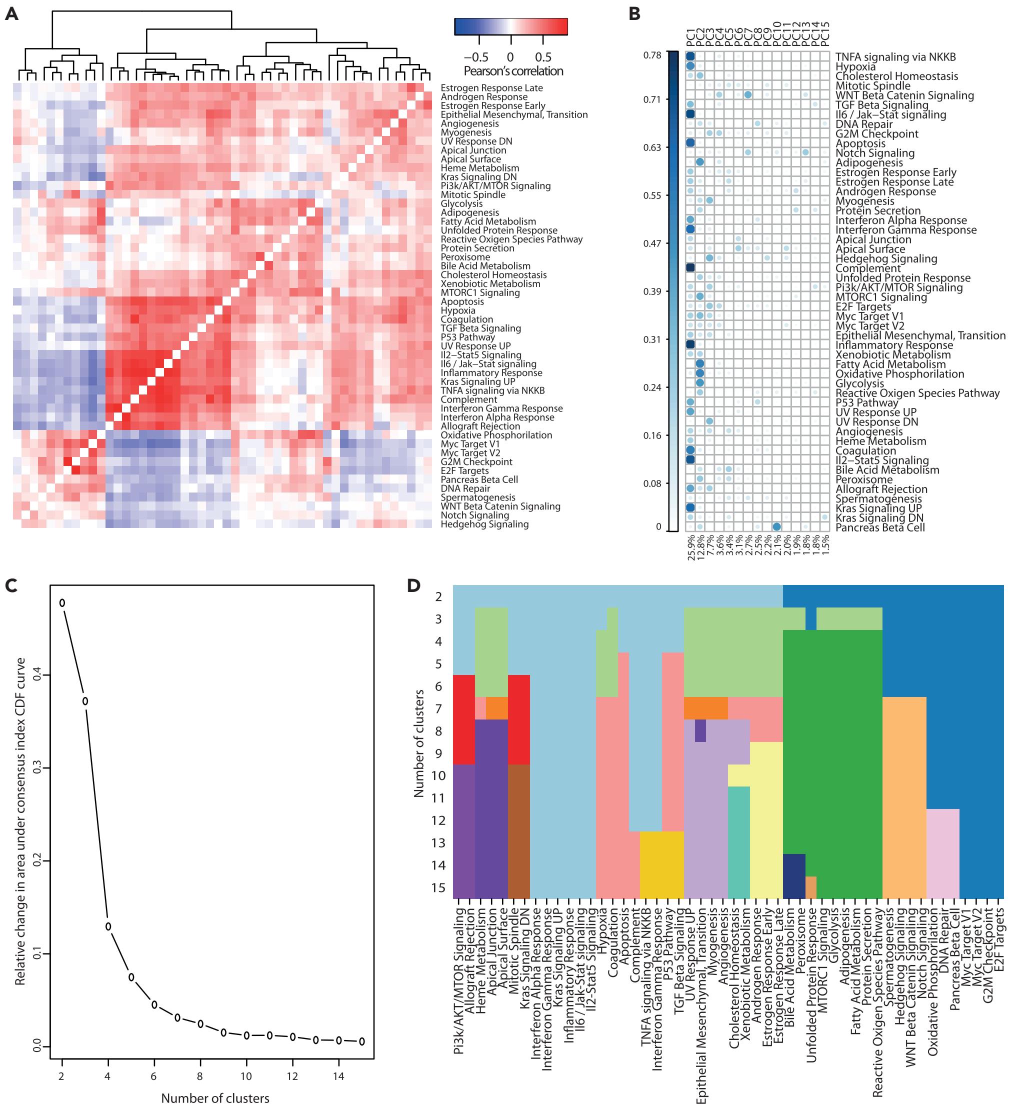


Figure 3. Principal Component and Clustering Analyses to Circumvent Hallmark Activity Redundancy

(A) Correlation heatmap between all 50 MSigDB hallmark activities on a meta-dataset comprising 28,513 cells from 8 different datasets.

(B) Importance of the 15 Principal Components (PC) for each activity (squared cosine, indicated by increasing circle size and blueness). Below, the proportion of total variance in the dataset explained by each PC.

(C) Relative increase in measure of clustering consensus as the number of clusters is increased. CDF, cumulative distribution function.

(D) Cluster assignment of all 50 activities, for a number of 2–15 clusters.

Redundancy Reduction to Obtain Phenotypic Profiles

We designed two methods to tackle redundancy, based on Principal Component (PC) and clustering analyses (see Methods). The first three PCs of the entire meta-dataset, respectively, explained 25.9%, 12.8%, and 7.7% of the variance in the data, whereas 11 PCs explained more than 2% of the variance (Figure 3B). For the clustering analyses, we investigated the relevance of splitting the data into 2–15 clusters. Using the consensus indices from bootstrapping experiments, we defined an optimal range between 6 and 10 clusters, after which increasing the number of clusters would not improve consensus (Figures 3C and 3D).

We defined phenotypic profiles for each cell based on either the PC scores or the average activity scores per cluster. We analyzed the six sets that provided metadata describing the predicted (sub)type of each cell (see Methods), using leave-one-out procedures to prevent overfitting. In line with our observations that cells clustered according to their type rather than set of origin, defining PC weights and optimal cluster compositions on all sets but the one analyzed still allowed one to identify patterns differentiating cell types (Figures S5–S10).

Cell-Cell Divergence across Tissue and Cancer Types

Pairwise Euclidean distances between phenotypic profiles then served to measure the phenotypic divergence between cells. We used different thresholds to calculate PCA- and cluster-based divergence, respectively, based on the minimum percentage of variance for a PC to be included in phenotypic profiles (0%, 1%, 2%, 3%, and 5%) and on the numbers of clusters to summarize all 50 activities (6–10 clusters). Phenotypic heterogeneity measures were highly correlated regardless of the thresholds in both methods (all Spearman's rho $\geq$ 0.72, all $p < 0.001$, Table S5), suggesting they are nearly equivalent. However, we observed less redundancy between PC scores than between cluster scores, independently of the number of clusters (Figure S10). We therefore use PCA-based phenotypic heterogeneity measures hereafter, with a 2% minimum threshold on explained variance for PC inclusion.

We investigated the pan-cancer relevance of our activity-based phenotypic divergence measure, using the six datasets for which cell type metadata were available. We report differences in cell-cell divergence distributions, according to whether two cells are of the same type or not and what that cell type is (Figures 4 and S11). In agreement with our pan-cancer observations that cells clustered by type more than dataset, the divergence between cells of different cell types was always the highest distribution (compared with same-type distributions) in all six datasets. This suggests that our metric will assign smaller divergence scores to cells from the same cell type. Using bootstrapped clustering analyses, we also investigated if different recurrent activity profiles could be observed among cancer cells only, in each set (Figures S12–S16, see Methods). Clusters related to proliferation and immune response could be observed in most analyses, whereas the most discriminant activities, and PC scores derived from them, varied between datasets. In the Venteicher astrocytoma dataset, a discernible sub-population tied to immune activities can be distinguished on the left, with marked differences in interferon alpha and gamma signatures (Figure 5). A separate sub-population with strong proliferation signaling can be observed in the center, whereas cells on the right side do not display particularly strong proliferation or immune-related signal. This suggests that activity-based distances can separate distinct subpopulations of malignant cells presenting different phenotypic characteristics.

Phenotypic Diversity Quantification

We further analyzed the relevance of activity-based approaches on two subsets with extended characterization in a large number of patients: 7 non-malignant cell types (T cell, B cell, Macrophage, Endothelial, Fibroblast, NK, Undefined) in 19 patients from the Tirosh melanoma dataset; 6 malignant subtypes (AC-like, OPC-like, MES1-like, MES2-like, NPC1-like, NPC2-like) in 28 patients from the Neftel glioma dataset. The average divergence in a group of cells was used as a surrogate for the group's phenotypic heterogeneity. We observed differences across the average profiles calculated for the distinct cell types, suggesting they are each characterized by specific activity patterns. The differences between the most divergent cells in each category, however, exemplify that individual cells can strongly deviate from these overall profiles (Figures 6A, 6B, and S17). Such variability, possibly due to the stochastic nature of gene expression, would be absent from standard classifying methods.

We proceeded to reclassify all cells according to the smallest Euclidean distance between their PCA-based profiles and the average profiles of each classification in both datasets. We observed a stronger concordance ($p = 0.022$, Wilcoxon test) when reclassifying cells from established normal cell types in melanoma samples according to their activities (Figure 6C, 82% ± 14 correctly reclassified samples), compared with subtypes of malignant glioma cells (Figure 6D, 54% ± 23). This confirmed that cells of similar type tend to partake in similar activities. However, in the glioma samples we analyzed, the differences between marker-based malignant subtypes were not as closely reflected by activity profiles as was observed in normal cell types.

We then computed the standard Simpson diversity index on a per-patient basis, according to the repartition of all cells from a patient into the relevant categories in both subsets. We found that it correlated very significantly with our divergence-based phenotypic heterogeneity score in both non-malignant cells from melanoma samples and malignant glioma cells (Figures 6E and 6F, Spearman's rho = 0.73 and rho = 0.49;

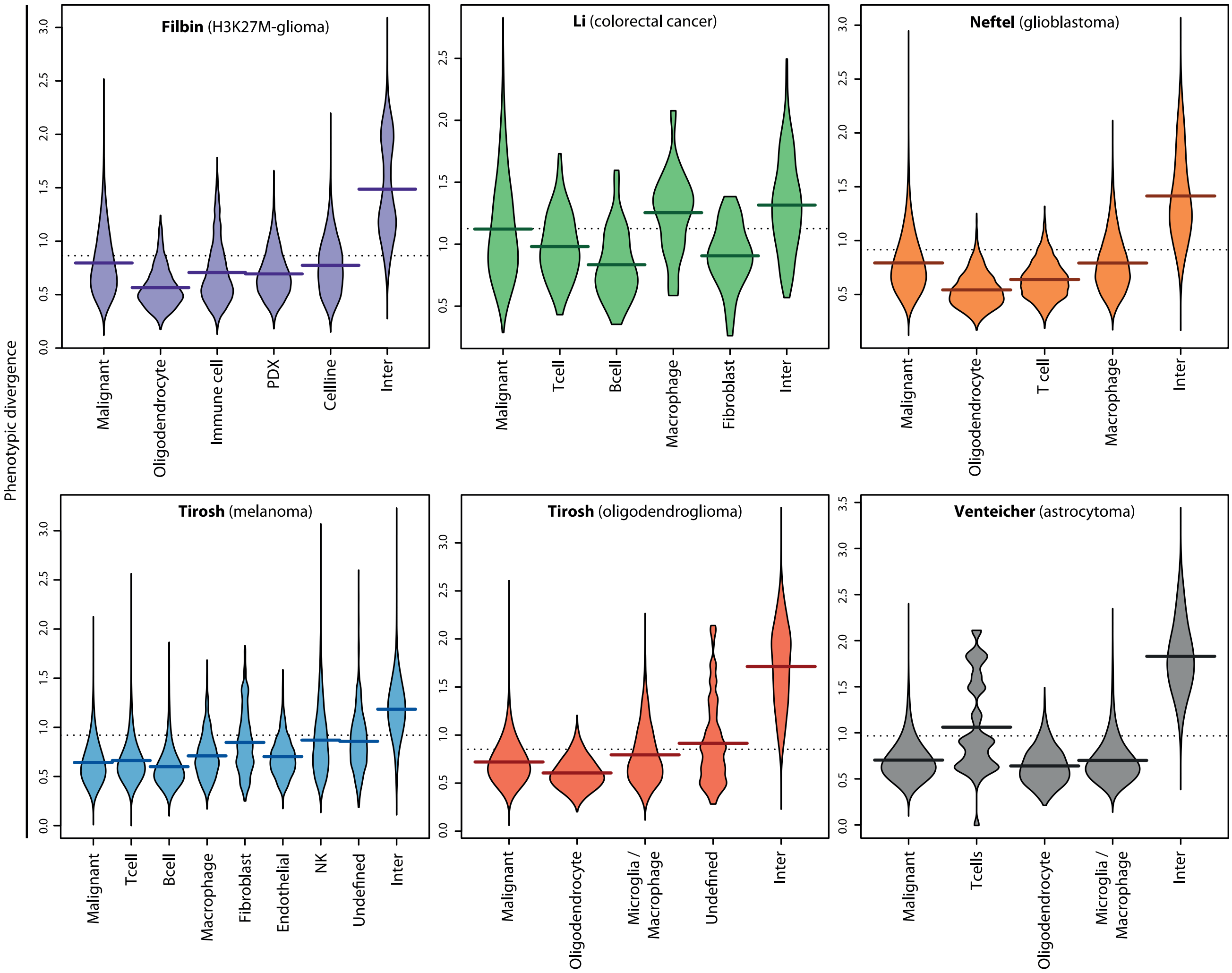


Figure 4. Pan-Cancer Phenotypic Cell-Cell Divergence

Pairwise cell-cell divergence distributions per cell type in each of the six datasets with curated metadata. Inter: inter-type divergence (between cells of different subtypes). All other distributions are between cells of the reported type. Dashed horizontal line: total average; broad horizontal lines: individual distribution averages.

p = 0.001 and p = 0.009, respectively). This suggests that this approach, although not relying on cell classification, can accurately capture the diversity of populations whose structure is known, both for malignant and normal cells from different tissues. Similar observations were reported using cluster-based distances (Figure S18).

Using the average activity-based divergence between malignant cells, we quantified intra-tumor phenotypic heterogeneity in all samples from the six datasets with metadata and compared them (Figure 7A). The mean phenotypic divergence of colorectal cancers (Li et al.) was significantly higher than other datasets, whereas melanoma heterogeneity was significantly lower (Wilcoxon test, Benjamini-Hochberg (BH) correction, $p < 0.001$ and $p = 0.004$, respectively). We furthermore report that between-samples variation in phenotypic diversity was the highest in melanoma (i.e., most heterogeneous in heterogeneity levels) and the lowest in oligodendroglioma (Figures 7B and 7C).

Phenotypic Diversity Evolution

We finally took advantage of cancer samples paired with normal tissue in the colorectal dataset to investigate the evolution of phenotypic diversity. In the five patients with colorectal cancer from Li et al. for which we could find paired tumor-normal data, diversity stayed at similar levels in three cases (CRC04, CRC06,

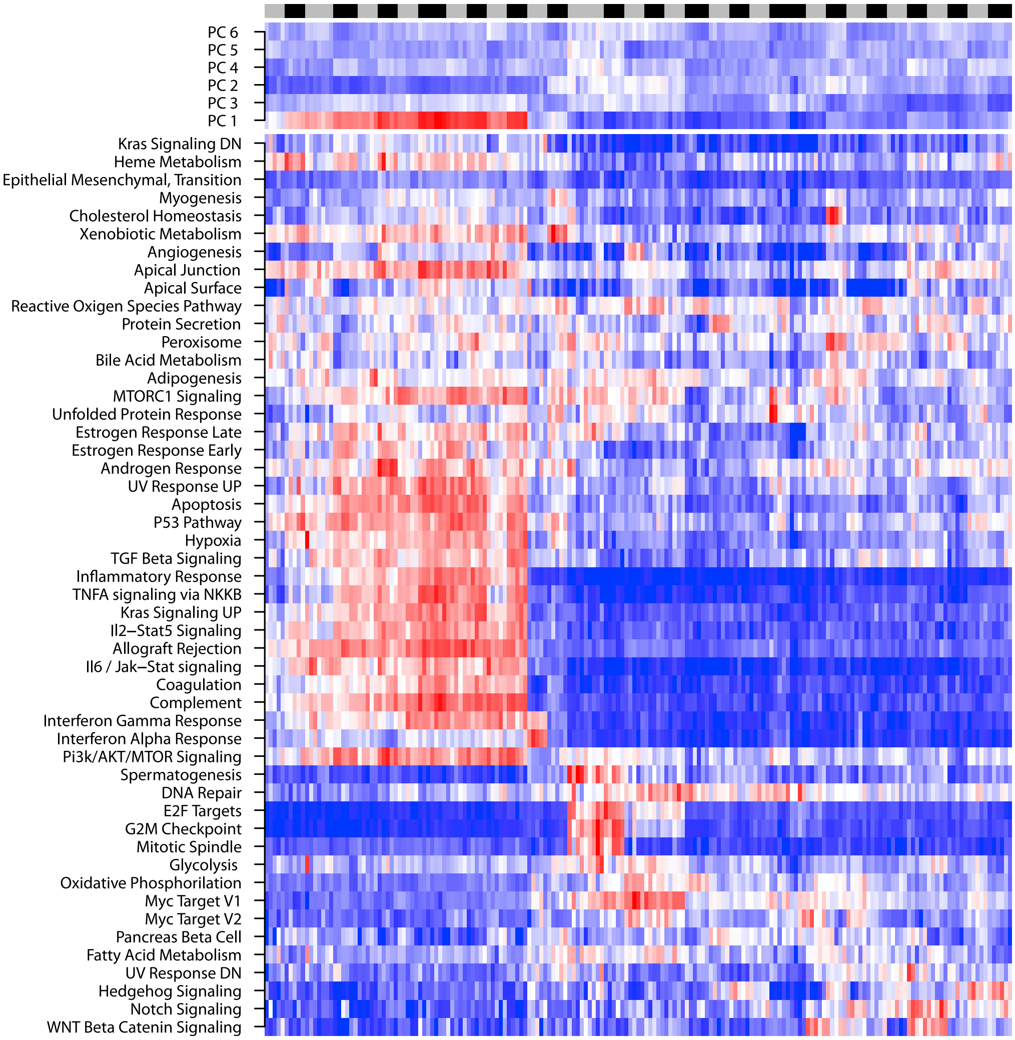


Figure 5. Isolated Activity Profiles of Significant Clusters of Malignant Cells in the Venteicher et al. Astrocytoma Dataset

Top: distinct significant clusters are identified by alternating black and gray color bars. Cells are ordered left to right according to the overall cluster data including all cells, although only significant clusters of five cells or more are displayed. Middle: Heatmap of PCA-based activity scores. All principal components were used for clustering analyses, but only those explaining >3% of total variance are displayed. PCA scores are ordered top to bottom according to complete hierarchical clustering based on Euclidean distances. Bottom: Heatmap of normalized activity scores, ordered top to bottom according to complete hierarchical clustering based on Euclidean distances.

CRC10), whereas it decreased very significantly in the tumor material in two cases (Figure 7D, CRC05, CRC08, $p < 0.001$, Wilcoxon test). Such decrease in diversity was not observed in other cell types in these patients (Figure S19). This fits a scenario in which cells go through a phenotypic bottleneck at tumor initiation, followed by the expansion of few selected clones.

DISCUSSION

Better understanding the dynamics of intra-tumor heterogeneity will help tailor better therapeutics to control and funnel cancer evolution. During malignant somatic evolution, cells drift away from their well-characterized normal ancestors by following trajectories unique to each patient (Tokutomi et al., 2019), whereas there is convergence across patients to (de)activate the necessary cellular activities (Hanahan and Weinberg, 2000, 2011). Consequently, we investigated the relevance of focusing on what cancer cells do, rather than what they are, to measure phenotypic diversity in the cancer context. We considered cellular activities

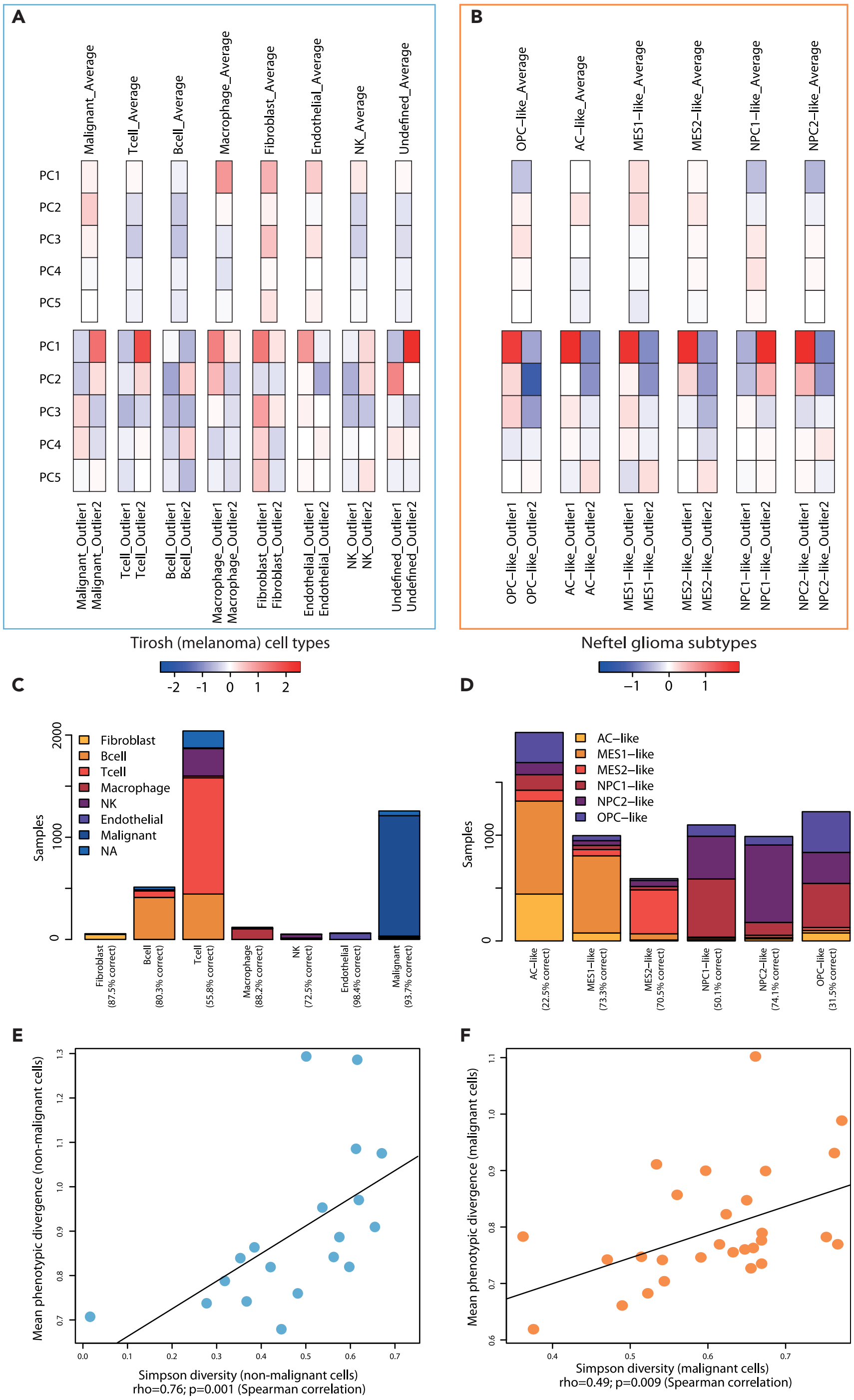


Figure 6. Phenotypic Diversity in Populations of Known Structure

(A and B) PCA-based phenotypic profiles of (A) seven non-malignant cell types from the Tirosh et al. melanoma dataset and (B) six glioma subtypes from the Neftel et al. H3K27M-glioma dataset. Average profiles on top were obtained by averaging all cells from a given subtype across all patients. The outlier profiles at the bottom were obtained from the same-type cell pairs displaying the highest activity-based divergence for each cell type. Only the first five principal components are shown.

(C) Barplots showing the breakdown of how non-malignant cells from melanoma samples would be re-categorized, based on the average activity profiles of each category in the Tirosh melanoma dataset.

Figure 6. *Continued*

(D) Barplots showing the breakdown of how malignant glioblastoma cells would be re-categorized, based on the average activity profiles of each category in the Neftel dataset.

(E) Relationship between mean phenotypic divergence between non-malignant cells in the melanoma dataset and the Simpson diversity index calculated on the repartition of cells into the seven non-malignant classes.

(F) Relationship between mean phenotypic divergence between malignant cells in the glioma dataset and the Simpson diversity index calculated on the classification of cells into the six glioma subtypes. Black lines: linear models.

as traits describing the phenotypic state of cells and used pairwise distances to quantify cell-cell divergence and overall diversity. Unlike many existing methods (Almendro et al., 2014; Ferrall-Fairbanks et al., 2019; Zhang et al., 2019), such an approach does not rely on classifying cells into putative, static identities that cancer cells drift away from in patient-specific fashion. It furthermore encompasses the temporal variability inherent to populations of cells replicating asynchronously and exhibiting stochastic differences in gene expression, which can itself foster resistance (Shaffer et al., 2017). In addition, such a method is not tissue-type specific and was relevant in all investigated datasets.

We first performed *in vitro* analyses, which revealed that it was possible to reliably predict in which cells a given activity had been induced, using targeted panels based on the MSigDB hallmark gene sets and the literature. This was done using single-cell RT-qPCR technology, which is more precise than RNA-seq on specific genes of interest (Mojtahedi et al., 2016). Our analysis, however, revealed that some of the best markers for activity detection were absent from the hallmark gene sets. Although this is likely to be attenuated when using entire gene sets rather than targeted panels, it exemplifies the need for more reliable gene signatures, particularly ones taking into account single-cell level specificities (Hwang et al., 2018; Larsson et al., 2019).

We then scored 50 hallmark activity signatures in 28,513 cells from eight publicly available datasets using the AUCell software. AUCell is based on a ranking procedure, which efficiently deals with normalization and is not affected by the dissimilarity in using either FPKM or TPM units across the datasets (Aibar et al., 2017). This was illustrated by cells not clustering according to their dataset of origin in the meta-dataset. "Dropouts" occurring when transcripts are not captured before sequencing can, however, affect ranking in low-expressed genes (Davis-Marcisak et al., 2019). Gene set enrichment analyses, in which multiple genes can contribute to the overall enrichment signal for an activity in each cell, are, however, less affected by dropouts than gene-specific differential expression analyses.

We reported high redundancy among the 50 activities scored, which we addressed by using PC and clustering analyses. We found that both methods were by and large equivalent. Importantly, hallmark activities do not focus on lineage-specific markers. Using their output, which summarizes multiple genes, is thus less likely to separate cells according to the expression of few highly discriminating lineage markers, such as can occur when focusing on the entire transcriptome. This is particularly relevant for cancer cells that broke free of homeostatic control and differentiation hierarchies, in which lineage markers inherited from ancestors may no longer correlate with phenotype and behavior.

We applied such an activity-based approach to investigate the divergence between and among cell types in six datasets with available metadata. We found that cells of the same type were less divergent than cells of different types. This can be explained by the fact that most reported cell types are non-malignant, with cells from the same type thus likely to partake in similar activities. We also observed that activity profiles recapitulated normal cell types better than malignant subtypes, although with very limited data (n = 1 in both cases). Furthermore, we could identify distinct clusters of malignant cells showing marked differences in their activity profiles in all datasets. Therefore, although same cell identity often implied similar activities, it was not always the case, especially in cancer cells for which our activity-based approach was aimed. This also indicated that this approach could reflect the divergence between cells similarly considered malignant using a blanket classification, but which appeared to engage in different activities.

Interestingly, the divergence between malignant cells was not recurrently higher or lower than that between normal subtypes, and patterns varied according to tumor type. In the two datasets with high numbers of both patients and cell type sub-classifications, the mean cell-cell divergence correlated significantly with standard diversity indices based on the repartition of individuals into subpopulations. These results suggest that avoiding the use of known lineage markers did not hamper the relevance of this approach across the investigated tissue types. Although we used a leave-one-out design to avoid

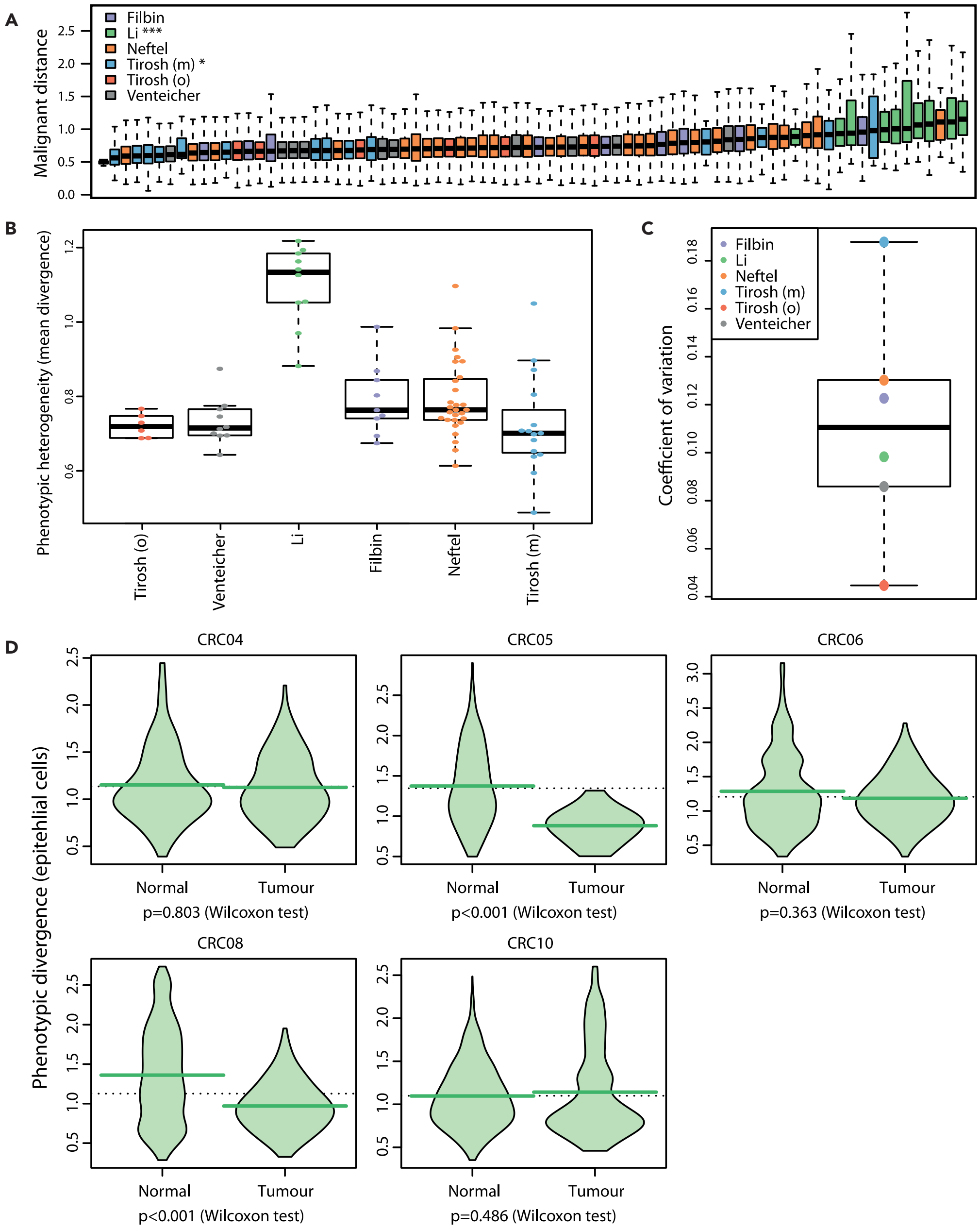


Figure 7. Differences and Dynamics of Phenotypic Diversity

(A) Distribution of phenotypic divergence between malignant cells in each sample across six datasets. Samples ordered by sample-wise phenotypic diversity (average divergence). ***: $p < 0.001$; *: $p < 0.05$ (Wilcoxon test, BH correction). Boxes represent the middle quartiles; black horizontal bars represent the median of each distribution; whiskers extend up to 1.5 times the interquartile range (box height) away from the box. Outliers (beyond the whiskers) are not displayed.

(B) Per-sample phenotypic diversity in all six sets.

(C) Coefficient of variation in phenotypic diversity across samples in each set.

(D) Phenotypic divergence distributions in normal and cancerous epithelia in five patients from the Li et al. dataset. Dashed horizontal line: total average; broad horizontal lines: individual distribution averages.

overfitting, it is, however, worth noticing that brain cell and tumor data are likely overrepresented in this study. Finally, using this approach on five patients with paired tumor-normal data suggested the existence of evolutionary bottlenecks on phenotypic diversity at tumor initiation. This would be in agreement with the genetic diversity decrease observed at this stage in orthogonal studies (Cross et al., 2020).

In this work, we focused on the quantification of phenotypic diversity according to cancer's atavistic evolutionary nature, as cells deviate from normal healthy cell types and regress toward ancestral unicellular growth (Davies and Lineweaver, 2011). We used single-cell expression analyses to quantify activity-based traits for each cell to create individual phenotypic profiles differing from static subtype classifications. This provides an alternative to marker-based methods, which can rely on markers not relevant anymore in the cancer context and that often cannot allow quantification of the differences between cells classified similarly. Not relying on markers furthermore bypasses tissue specificity and provides a universal approach applicable to all tumor types.

Limitations of the Study

In this study we used pre-defined activity signatures based on bulk data that were not specifically designed for relevance in cancer studies. More work is therefore needed to provide standardized tools to reproducibly measure phenotypic ITH from single-cell RNA data. The development of accurate single-cell-specific expression signatures for the most recurrently dysregulated pathways in cancer would provide enhanced precision to build per-cell phenotypic profiles. This will require determination of the most relevant activities that contribute to the convergence toward the ''cancer hallmarks'' (Hanahan and Weinberg, 2011) dysregulation common to most cancer types. It will also be necessary to reliably assess their predictability in single cells, taking into account the specificity of single cell expression data and design methods accounting for the redundancy among them. Finally, it will also be critical to understand how intra-tumor heterogeneity at single-cell level can be extrapolated from bulk samples, how this reflects inter-patient heterogeneity, and how it ties to genetic and clinical features.

Successful implementations will improve future similar activity-based approaches to quantify phenotypic diversity in the evolutionary context of cancer. This will in turn allow one to better monitor the evolution of phenotypic diversity over time and space and facilitate the identification of therapeutic opportunities to control intra-tumor heterogeneity. This would ultimately help thwart the emergence of resistant populations and thereby enhance clinical outcomes.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

DATA AND CODE AVAILABILITY

The R scripts and data for this project are available on github: https://github.com/pierremartinez/PhDiv.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101061.

ACKNOWLEDGMENTS

The authors wish to thank Anne-Pierre Morel and Christelle Chassot for valuable discussions and assistance on the experimental setup.

AUTHOR CONTRIBUTIONS

L.M. and F.F. performed *in vitro* experiments. F.F., A.V., A.P., and P.M. designed *in vitro* experiments. L.D.S., B.L., and P.M. performed bioinformatics analyses. A.V., A.P., and P.M. supervised the work. P.M. wrote the manuscript. All authors revised the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 27, 2019
Revised: March 2, 2020
Accepted: April 9, 2020
Published: May 22, 2020

REFERENCES

Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nat. Methods *14*, 1083–1086.

Almendro, V., Kim, H.J., Cheng, Y.K., Gönen, M., Itzkovitz, S., Argani, P., van Oudenaarden, A., Sukumar, S., Michor, F., Polyak, K., et al. (2014). Genetic and phenotypic diversity in breast tumor metastases. Cancer Res. *74*, 1338–1348.

Andor, N., Harness, J.V., Müller, S., Mewes, H.W., and Petritsch, C. (2014). EXPANDS: expanding ploidy and allele frequency on nested subpopulations. Bioinformatics *30*, 50–60.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.

Bertucci, F., Ng, C.K.Y., Patsouris, A., Droin, N., Piscuoglio, S., Carbuccia, N., Soria, J.C., Dien, A.T., Adnani, Y., Kamal, M., et al. (2019). Genomic characterization of metastatic breast cancers. Nature *569*, 560–564.

Chang, K., Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas pan-cancer analysis project. Nat. Genet. *45*, 1113–1120.

Cross, W., et al. (2020). Stabilising selection causes grossly altered but stable karyotypes in metastatic colorectal cancer. bioRxiv, 2020, https://doi.org/10.1101/2020.03.26.007138.

Davies, P.C.W., and Lineweaver, C.H. (2011). Cancer tumors as Metazoa 1.0: tapping genes of ancient ancestors. Phys. Biol. *8*, 015001.

Davis-Marcisak, E.F., et al. (2019). Differential variation analysis enables detection of tumor heterogeneity using single-cell RNA-sequencing data. Cancer Res. https://doi.org/10.1158/0008-5472.can-18-3882.

Fan, J., Lee, H.O., Lee, S., Ryu, D.E., Lee, S., Xue, C., Kim, S.J., Kim, K., Barkas, N., Park, P.J., Park, W.Y., and Kharchenko, P.V. (2018). Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. Genome Res. *28*, 1217–1227.

Ferrall-Fairbanks, M.C., Ball, M., Padron, E., and Altrock, P.M. (2019). Leveraging single-cell RNA sequencing experiments to model intratumor heterogeneity, JCO clinical cancer informatics. Am. Soc. Clin. Oncol. *3*, 1–10.

Filbin, M.G., Tirosh, I., Hovestadt, V., Shaw, M.L., Escalante, L.E., Mathewson, N.D., Neftel, C., Frank, N., Pelton, K., Hebert, C.M., et al. (2018). Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. Science *360*, 331–335.

Fischer, A., Vázquez-García, I., Illingworth, C.J.R., and Mustonen, V. (2014). High-definition reconstruction of clonal composition in cancer. Cell Rep. *7*, 1740–1752.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

Gatenby, R.A., and Brown, J. (2017). Mutations, evolution and the central role of a self-defined fitness function in the initiation and progression of cancer. Biochim. Biophys. Acta. https://doi.org/10.1016/j.bbcan.2017.03.005.

Gerlinger, M., Rowan, A.J., Horswell, S., Math, M., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N. Engl. J. Med. *366*, 883–892.

Greaves, M., and Maley, C.C. (2012). Clonal evolution in cancer. Nature *481*, 306–313.

Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57–70. http://www.ncbi.nlm.nih.gov/pubmed/10647931.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Hinohara, K., Wu, H.J., Vigneau, S., McDonald, T.O., Igarashi, K.J., Yamamoto, K.N., Madsen, T., Fassl, A., Egri, S.B., Papanastasiou, M., et al. (2018). KDM5 histone demethylase activity links cellular transcriptomic heterogeneity to therapeutic resistance. Cancer Cell *34*, 939–953.e9.

Hwang, B., Lee, J.H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp. Mol. Med. *50*, 96.

Larsson, A.J.M., Johnsson, P., Hagemann-Jensen, M., Hartmanis, L., Faridani, O.R., Reinius, B., Segerstolpe, Å., Rivera, C.M., Ren, B., Sandberg, R., et al. (2019). Genomic encoding of transcriptional burst kinetics. Nature *565*, 251–254.

Lässig, M., Mustonen, V., and Walczak, A.M. (2017). Predicting evolution. Nat. Ecol. Evol. *1*, https://doi.org/10.1038/s41559-017-0077.

Martinez, P., Timmer, M.R., Lau, C.T., Calpe, S., Sancho-Serra Mdel, C., Straub, D., Baker, A.M., Meijer, S.L., Kate, F.J., Mallant-Hent, R.C., et al. (2016). Dynamic clonal equilibrium and predetermined cancer risk in Barretts oesophagus. Nat. Commun. *7*, 12158.

Li, H., Courtois, E.T., Sengupta, D., Tan, Y., Chen, K.H., Goh, J.J.L., Kong, S.L., Chua, C., Hon, L.K., Tan, W.S., et al. (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat. Genet. *49*, 708–718.

Maley, C.C., Galipeau, P.C., Finley, J.C., Wongsurawat, V.J., Li, X., Sanchez, C.A., Paulson, T.G., Blount, P.L., Risques, R.A., Rabinovitch, P.S., et al. (2006). Genetic clonal diversity predicts progression to esophageal adenocarcinoma. Nat. Genet. *38*, 468–473.

Martinez, P., Kimberley, C., BirkBak, N.J., Marquard, A., Szallasi, Z., and Graham, T.A. (2017). Quantification of within-sample genetic heterogeneity from SNP-array data. Sci. Rep. *7*, 3248.

McGranahan, N., and Swanton, C. (2015). Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. Cancer Cell *27*, 15–26.

Mojtahedi, M., Skupin, A., Zhou, J., Castaño, I.G., Leong-Quong, R.Y., Chang, H., Trachana, K., Giuliani, A., and Huang, S. (2016). Cell Fate Decision as High-Dimensional Critical State Transition. PLoS Biol. *14*, e2000640..

Neftel, C., Laffy, J., Filbin, M.G., Hara, T., Shore, M.E., Rahme, G.J., Richman, A.R., Silverbush, D., Shaw, M.L., Hebert, C.M., et al. (2019). An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. Cell *178*, 835–849.e21.

Nguyen, Q.H., Pervolarakis, N., Blake, K., Ma, D., Davis, R.T., James, N., Phung, A.T., Willey, E., Kumar, R., Jabart, E., et al. (2018). Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. Nat. Commun. *9*, 2028.

Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., et al. (2012). The life history of 21 breast cancers. Cell *149*, 994–1007.

Nowell, P.C. (1976). The clonal evolution of tumor cell populations. Science *194*, 23–28. http://www.ncbi.nlm.nih.gov/pubmed/959840.

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science *344*, 1396–1401.

Scialdone, A., Natarajan, K.N., Saraiva, L.R., Proserpio, V., Teichmann, S.A., Stegle, O., Marioni, J.C., and Buettner, F. (2015). Computational assignment of cell-cycle stage from single-cell transcriptome data. Methods *85*, 54–61.

Shaffer, S.M., Dunagin, M.C., Torborg, S.R., Torre, E.A., Emert, B., Krepler, C., Beqiri, M., Sproesser, K., Brafford, P.A., Xiao, M., et al. (2017). Reprogramming as a mode of cancer drug resistance. Nat. Publ. Group *546*, 431–435.

Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189–196.

Tirosh, I., Venteicher, A.S., Hebert, C., Escalante, L.E., Patel, A.P., Yizhak, K., Fisher, J.M., Rodman, C., Mount, C., Filbin, M.G., et al. (2016). Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature *539*, 309–313.

Tokutomi, N., Moyret-Lalle, C., Puisieux, A., Sugano, S., and Martinez, P. (2019). Quantifying local malignant adaptation in tissue-specific evolutionary trajectories by harnessing cancers repeatability at the genetic level. Evol. Appl. *12*, 1062–1075.

Trigos, A.S., Pearson, R.B., Papenfuss, A.T., and Goode, D.L. (2018). How the evolution of multicellularity set the stage for cancer. Br. J. Cancer *118*, 145–152.

Venteicher, A.S., Tirosh, I., Hebert, C., Yizhak, K., Neftel, C., Filbin, M.G., Hovestadt, V., Escalante, L.E., Shaw, M.L., Rodman, C., et al. (2017). Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. Science *355*, eaai8478.

Williams, M.J., Werner, B., Heide, T., Curtis, C., Barnes, C.P., Sottoriva, A., and Graham, T.A. (2018). Quantification of subclonal selection in cancer from bulk sequencing data. Nat. Genet. *50*, 895–903.

Zhang, A.W., O'Flanagan, C., Chavez, E.A., Lim, J.L.P., Ceglia, N., McPherson, A., Wiens, M., Walters, P., Chan, T., Hewitson, B., et al. (2019). Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. Nat. Methods, 1–9, https://doi.org/10.1038/s41592-019-0529-1.

iScience, Volume *23*

Supplemental Information

Assessing Cell Activities rather than Identities to Interpret Intra-Tumor Phenotypic Diversity and Its Dynamics

Laloé Monteiro, Lydie Da Silva, Boris Lipinski, Frédérique Fauvet, Arnaud Vigneron, Alain Puisieux, and Pierre Martinez

Transparent Methods

In-vitro activity induction

MCF10A are mammary epithelial cells that spontaneously immortalised with no external stimulus (Qu *et al.*, 2015). We selected them as a model given our previous work demonstrating that adding TGF-β in the medium promoted a spontaneous epithelial to mesenchymal transition (EMT) (Morel *et al.*, 2017). All cell culture replicates were grown in a physiological concentration of glucose (6 mM) with daily medium renewal, so that no cell population would heavily rely on glycolysis to produce energy. Glycolytic activity was induced through increased glucose concentration (20mM), which was controlled for by monitoring pH to validate that the medium became more acid following increased glycolysis and subsequent lactate production. DNA damage was induced by irradiating cells at 2 Gy at the Centre Léon Bérard radiology facility. EMT was induced by adding TGF-β to the medium and harvesting cells after 4 days and monitoring for morphological changes through microscopy (Supplementary Figure 20).

Single-cell RT-qPCR gene panel design

We optimised panels comprising subsets of each MSigDB hallmark gene set (Liberzon *et al.*, 2015) for all 3 activities (EMT, DNA repair, glycolysis), by selecting those whose expression was most correlated to the activity's enrichment score reported. We used 1,036 samples from the Cancer Cell Line Encyclopedia (Barretina *et al.*, 2012) (CCLE) for discovery and 10,885 samples from The Cancer Genome Atlas (TCGA) for validation, with Area Under the Curve and Spearman correlation statistics to optimise the panel designs. The sample breakdown among the 32 TCGA sets is reported in Supplementary Table 6.

We first identified the top 50% (top hereafter) and bottom 50% samples (bottom hereafter) with respectively the lowest and highest enrichment score per activity, using single-sample Gene Set Enrichment Analysis on all 50 entire signatures (Subramanian *et al.*, 2005). We investigated the genes in each signature, only considering genes that were present in a single hallmark gene set and had a 1-to-1 orthologue in mouse, in order to focus on essential genes. Samples were bootstrapped 100 times by randomly splitting samples into equal-sized validation and discovery sets, each containing half of both the top and bottom samples. At every iteration, the 15 most differentially expressed genes were defined using the limma R library and were recorded along with the sign of the fold change. After all iterations were performed, we selected the 15 genes present most times in the top 15 differentially expressed genes for each activity.

Every combination of 1 to 15 of these genes was then assessed for its ability to correlate with whole-set signatures using a similar bootstrapping approach. At each bootstrapping iteration, we build a matrix containing the expression value of each gene combination in 100 of the top 50% scoring samples and 100 of the bottom 50% ones. The expression values are then normalised into Z-scores in this bootstrapped matrix. We then attribute a per-sample score to the combination as follows by summing the Z-scores multiplied by the sign of the expected fold change of each gene. Based on these per-sample scores, we compute the AUC in predicting if a sample belongs to the top or bottom sample set using the pROC R library. The AUCs and Spearman correlations are recorded for each bootstrapping iteration for each combination (Supplementary Figure 21).

Finally, a cross-validation analysis was performed on the TCGA metadataset for each activity. We merged the best 5 genes based on AUC and the best 5 genes based on Spearman's correlation in the

CCLE data. We assessed their performance by calculating correlations and AUCs for the top/bottom 50% scoring samples using whole-set signature in the TCGA dataset.

We selected 7 EMT genes (*NNMT*, *COL4A1*, *SNAI2*, *FBN1*, *CTGF*, *FSTL1*, *FN1*), 7 DNA repair genes (*POLR2F*, *RFC5*, *POLR2E*, *CLP1*, *SF3A3*, *POLD1*, *DUT*) and 6 Glycolysis genes (*EXT1*, *AGRN*, *SLC16A3*, *PYGB*, *PYGL*, *IL13RA1*). We furthermore included additional genes according to the literature: 6 additional EMT genes (*CDH1*, *COL5A1*, *HTRA1*, *ITGAV*, *ZEB1*, *ZEB2*), 6 additional DNA repair genes (*DDB2*, *FEN1*, *LIG1*, *XPC*, *PMS1*, *POLQ*) and 3 additional Glycolysis genes (*HK2*, *PFKM*, *LDHA*). 4 genes from the Myc targets hallmark (*BYSL*, *DCTPP1*, *GNL3*, *TCOF1*) and 7 stably expressed "housekeeping" genes (*CIAO1*, *CNOT4*, *HNRNPK*, *RAB1A*, *TIAL1*, *UBE2D3*, *YTHDC1*, Supplementary Figure 22) added as internal controls, as well as two RNA spikes from the manufacturer (spikes 1 and 4). Primers are reported in Supplementary Table 7.

Inferring cellular activities

4-parameter Beta-Poisson (BP) distributions and differential expression were estimated using the BPSC R package (Vu *et al.*, 2016). Significantly differentially expressed genes were identified comparing the activity-specific cells to all other ones (including cells with another activity induction). BP distributions were used to calculate the probability that an expression value (in number of transcripts) observed for a marker gene came from a cell in which the related activity was induced. For each gene in each cell, the likelihood that the observed value originated from a population in which the related activity had been induced was calculated. Two BP distributions were first estimated in leave-one-out fashion (removing the cell under scrutiny): one using cells in which the related activity had been induced, and another one in cells in which this activity had not been induced. We then defined a confidence interval surrounding the observed expression value ± sd_s, where sd_s corresponds to 1 standard deviation in the distribution of spike *s* expression values. The probability *P(x,d)* that a given expression value *x* was from a given distribution *d* was given by the percentage of values falling into this interval out of 10,000 draws using the rBP function of the BPSC package, given the 4 parameters of the BP distributions. The likelihood that the observed value *x* came from a cell where the activity had been induced was then given by the formula:

$$L(x) = \frac{\mathrm{P(x,I)}}{\mathrm{P(x,I)+P(x,N)}}$$

Where *P(x,I)* is the probability of observing value x in the induced cells BP distribution, while *P(x,N)* is the probability of observing value *x* in the non-induced cells BP distribution. Finally, we used generalised linear models (glm) based on these likelihoods, to assess the power of the significantly differentially expressed marker genes of each activity, as well as their combinations in predicting whether the activity was induced in a cell or not. Only genes reported as significantly differentially expressed were included in multi-gene glm analyses.

Single-cell isolation and RT-qPCR data

96 cells were isolated from 4 populations (control, EMT, DNA repair, Glycolysis) using the Fluidigm C1 technology. 12 cells from each population were then analysed using a 48.48 chip on the BiomarkHD hardware with our designed 48-genes panel, using 18 pre-amplification cycles and maximum 30 amplification cycles for quantification. Single-cell isolation and RT-qPCR experiments were performed by the ProfilExpert platform (Université Claude Bernard Lyon 1, France). We filtered out genes and cells for which >30% of the 48 wells involved did not meet the BiomarkHD PASS criterion. For each spike (1 and 4), cells corresponding to the most distant outliers Ct values were filtered out (either side),

until the Ct distribution was deemed normal (p>0.05, Shapiro test). After this pre-filtering, 36 genes per 36 cells were left. We normalised Ct values by subtracting, for each cell and each spike, the difference between the spike value for the cell with the mean obtained across all cells for this spike (two subtractions per cell, one for each spike). We then obtained transcript abundance *nt* for gene *i* in cell *j* as follows:

$$(1)\ \ nt_{ij} = 48 \times 45 \times 2^{30-18-Ct_{ij}}$$

In order to account for possible differences in total RNA per cell, we further normalised this number by the cell-specific expression of housekeeping genes. Expression for each housekeeping gene was first linearly normalised across all cells, by dividing by the maximum number of transcripts observed for the gene across all cells (maximum expression observed across all cells: 1; no expression: 0). Each cell *j* was then attributed a weight H_j corresponding to the mean normalised expression of housekeeping genes in cell *j*. Transcripts for all genes *i* in cell *j* were finally normalised as follows:

$$(2)\ \ Nt_{ij} = \frac{nt_{ij}}{H_j}$$

Datasets

We downloaded 8 cancer-related single-cell RNA (sc-RNA) datasets, through the Single Cell Portal (https://portals.broadinstitute.org/single_cell) and publications: Fan *et al.* multiple myeloma (Fan *et al.*, 2018) ; Filbin *et al.* H3 lysine27-to-methionine mutated H3K27M-glioma (Filbin *et al.*, 2018); Li *et al.* colorectal cancer (Li *et al.*, 2017); Neftel *et al.* glioblastoma (Neftel *et al.*, 2019); Patel *et al.* glioblastoma (Patel *et al.*, 2014); Tirosh *et al.* melanoma (Tirosh, Izar, *et al.*, 2016) and oligodendroglioma (Tirosh, Venteicher, *et al.*, 2016); Venteicher *et al.* astrocytoma (Venteicher *et al.*, 2017). The Li *et al.* dataset included both a tumour set and a normal set of sc-RNA data. All 8 sets were used to create a meta-dataset on which to investigate hallmark activity signatures, although only the 6 for which we could find cell type information were used for detailed analyses (Filbin, Li, Neftel, Tirosh x2 and Venteicher). We solely focused on Smart-Seq2 expression data quantified either using fragments per kilobase per million (FPKM, Li *et al.*) or transcripts per million (TPM, all other sets) metrics for consistency. This represented 28,513 cells from 85 patients (Supplementary Table 8).

Activity scores

The AUCell software (Aibar *et al.*, 2017) was used to score the enrichment of each MSigDB hallmark signatures, considered as activities, in all cells. In addition, the cyclone software (Scialdone *et al.*, 2015) was used to predict the cell cycle phase status of all cells (G1/S/G2M). The meta-dataset including all cells from the 8 sets was constructed by normalising the scores for each of the 50 activities on a per-set and per-activity basis. This was achieved by first subtracting the minimal score observed for the activity and then dividing by the maximum score for this activity. Thus, the distribution of scores for each activity had a minimum of 0 and a maximum of 1 in each of the 8 sets prior to merging.

Redundancy reduction and phenotypic distances

We developed two approaches to reduce redundancy among the 50 activities, respectively based on Principal Component Analysis (PCA) and clustering. The first approach relied on determining the PCA coefficients of each activity so as to transform the 50 activity scores of each cell into PCA coordinates, using linear combinations of the scores and coefficients thanks to the FactoMineR R package(Lê, Josse and Husson, 2008). We used the ConsensusClusterPlus R package (Wilkerson and Hayes, 2010) to identify the optimal number of clusters and their composition, with 100 bootstrapping replicates. The scores of all activities in a given cluster were then averaged to create a score for each cluster in each cell. As a result, phenotypic profiles were represented by vectors of PCA component scores or cluster scores per cell. To ensure reproducibility, PCA coefficients and cluster composition were calculated using a leave-one-out procedure: the cells from each of the 8 sets were removed from the meta-dataset to calculate coefficients and assign activities to clusters, before using this information to calculate the PCA and cluster scores for the cells of the absent set. Phenotypic distances were then calculated as the Euclidean distance between the profiles of two cells, with the average distance accounting for the phenotypic diversity for a given sample / group of cells. For comparison against standard indices, we used the Simpson index rather than the Shannon one, as it is normalised and thus more stable when the number of individuals (*i.e.* cells) vary between populations (*i.e.* patient samples).

Recurrent malignant cell clusters

We used the pvclust (Suzuki and Shimodaira, 2006) software on the PCA-based profiles of malignant cells only in the 6 datasets with available metadata, using Euclidean distances, Ward clustering and 500 bootstrap replicates. The pvpick function was used to identify groups of cells that were significantly associated to each other across bootstrap replicates. For display, only these significant clusters were selected but their display order in the cluster including all malignant cells was kept.

Cell reclassification

We determined the average profiles of non-malignant cell types from the Tirosh (melanoma) dataset and malignant subtypes from the Neftel glioblastoma dataset, using the mean of each PC score for all cells from each (sub)type. All 50 principal components were included in the calculation, yielding a vector of 50 items per (sub)type. Each cell was then re-classified as belonging to the (sub)type whose average profile it was less distant to, using Euclidean distances on all PC scores.

References

Aibar, S. *et al.* (2017) 'SCENIC: Single-cell regulatory network inference and clustering', *Nature Methods*. Nature Publishing Group, 14(11), pp. 1083–1086. doi: 10.1038/nmeth.4463.

Barretina, J. *et al.* (2012) 'The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity', *Nature*. Nature Research, 483(7391), pp. 603–307. doi: 10.1038/nature11003.

Fan, J. *et al.* (2018) 'Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data', *Genome Research*. Cold Spring Harbor Laboratory Press, 28(8), pp. 1217–1227. doi: 10.1101/gr.228080.117.

Filbin, M. G. *et al.* (2018) 'Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq', *Science*. American Association for the Advancement of Science, 360(6386), pp. 331–335. doi: 10.1126/science.aao4750.

Lê, S., Josse, J. and Husson, F. (2008) 'FactoMineR : An R Package for Multivariate Analysis', *Journal of Statistical Software*, 25(1), pp. 1–18. doi: 10.18637/jss.v025.i01.

Li, H. *et al.* (2017) 'Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors', *Nature Genetics*. Nature Research, 49(5), pp. 708–718. doi: 10.1038/ng.3818.

Liberzon, A. *et al.* (2015) 'The Molecular Signatures Database Hallmark Gene Set Collection', *Cell Systems*. Elsevier, 1(6), pp. 417–425. doi: 10.1016/j.cels.2015.12.004.

Morel, A.-P. *et al.* (2017) 'A stemness-related ZEB1–MSRB3 axis governs cellular pliancy and breast cancer genome stability', *Nature Medicine*, 23(5), pp. 568–578. doi: 10.1038/nm.4323.

Neftel, C. *et al.* (2019) ‘An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma.’, *Cell*. Elsevier, 0(0). doi: 10.1016/j.cell.2019.06.024.

Patel, A. P. *et al.* (2014) ‘Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma’, *Science*. American Association for the Advancement of Science, 344(6190), pp. 1396–1401. doi: 10.1126/science.1254257.

Qu, Y. *et al.* (2015) ‘Evaluation of MCF10A as a Reliable Model for Normal Human Mammary Epithelial Cells’, *PLOS ONE*. Edited by X. Liu. Public Library of Science, 10(7), p. e0131285. doi: 10.1371/journal.pone.0131285.

Scialdone, A. *et al.* (2015) ‘Computational assignment of cell-cycle stage from single-cell transcriptome data’, *Methods*. Academic Press, 85, pp. 54–61. doi: 10.1016/j.ymeth.2015.06.021.

Subramanian, A. *et al.* (2005) ‘Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.’, *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), pp. 15545–50. doi: 10.1073/pnas.0506580102.

Suzuki, R. and Shimodaira, H. (2006) ‘Pvclust: an R package for assessing the uncertainty in hierarchical clustering’, *Bioinformatics*, 22(12), pp. 1540–1542. doi: 10.1093/bioinformatics/btl117.

Tirosh, I., Izar, B., *et al.* (2016) ‘Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq’, *Science*, 352(6282), pp. 189–196. doi: 10.1126/science.aad0501.

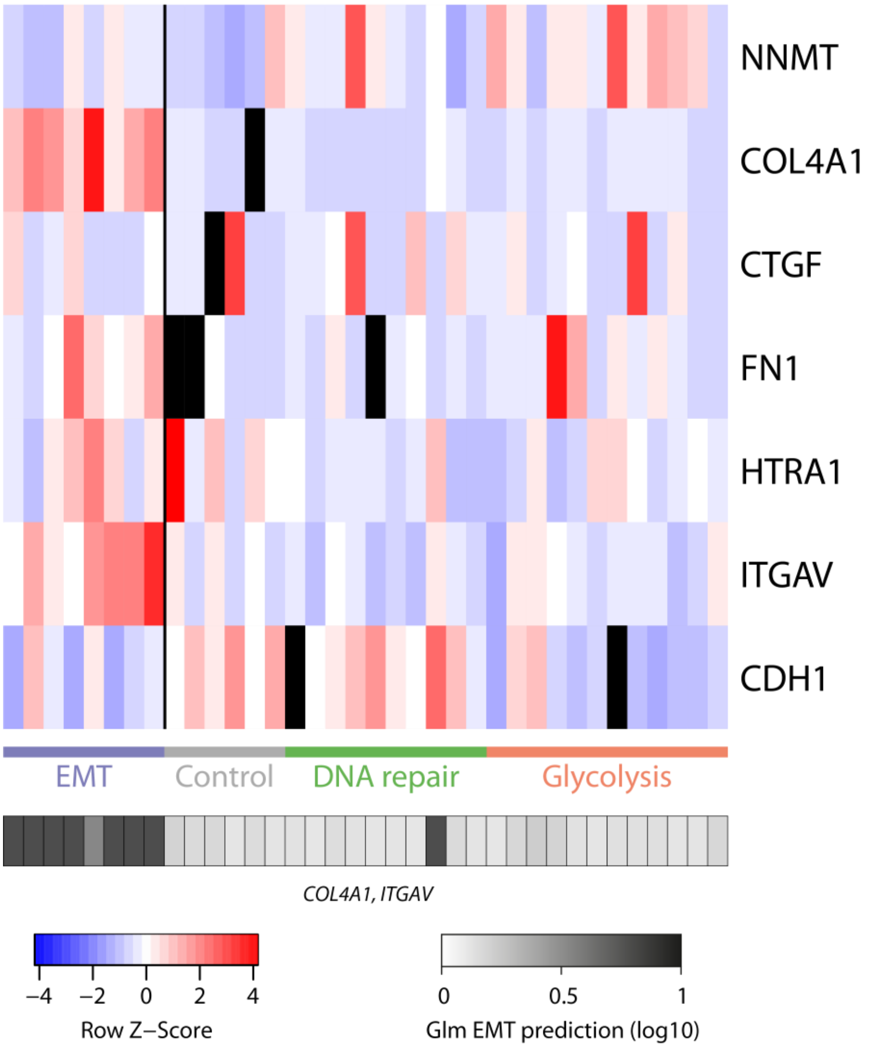
Tirosh, I., Venteicher, A. S., *et al.* (2016) ‘Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma’, *Nature*. Nature Publishing Group, 539(7628), pp. 309–313. doi: 10.1038/nature20123.

Venteicher, A. S. *et al.* (2017) ‘Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq’, *Science*, 355(6332), p. eaai8478. doi: 10.1126/science.aai8478.

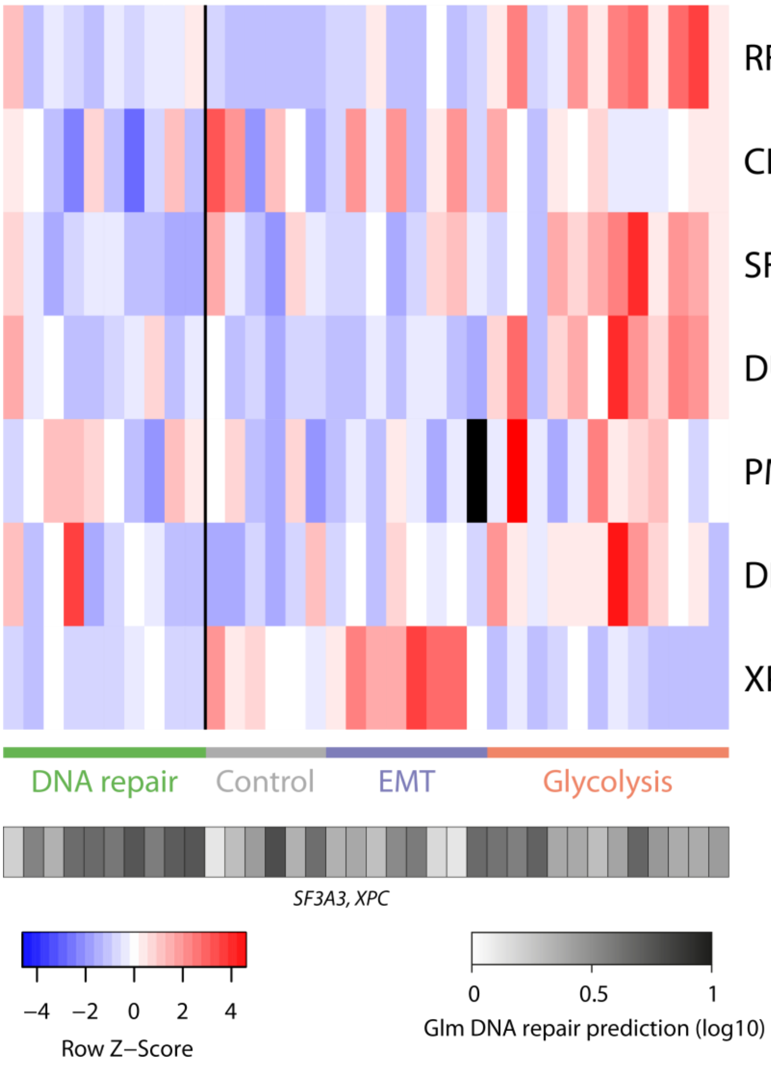
Vu, T. N. *et al.* (2016) ‘Beta-Poisson model for single-cell RNA-seq data analyses’, *Bioinformatics*. Oxford University Press, Oxford, 32(14), pp. 2128–2135. doi: 10.1093/bioinformatics/btw202.

Wilkerson, M. D. and Hayes, D. N. (2010) ‘ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking’, *Bioinformatics*. Narnia, 26(12), pp. 1572–1573. doi: 10.1093/bioinformatics/btq170.

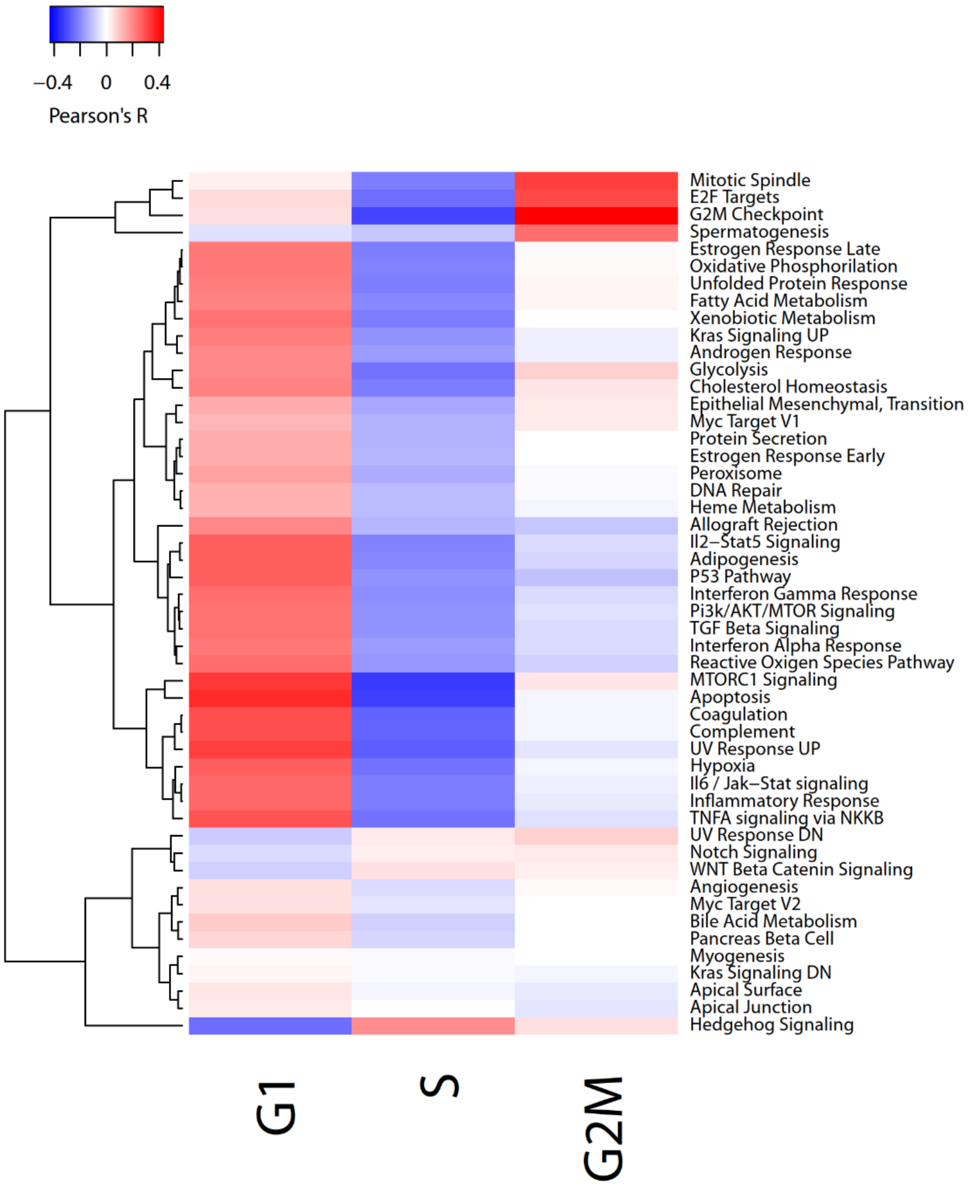
Supplementary Figures



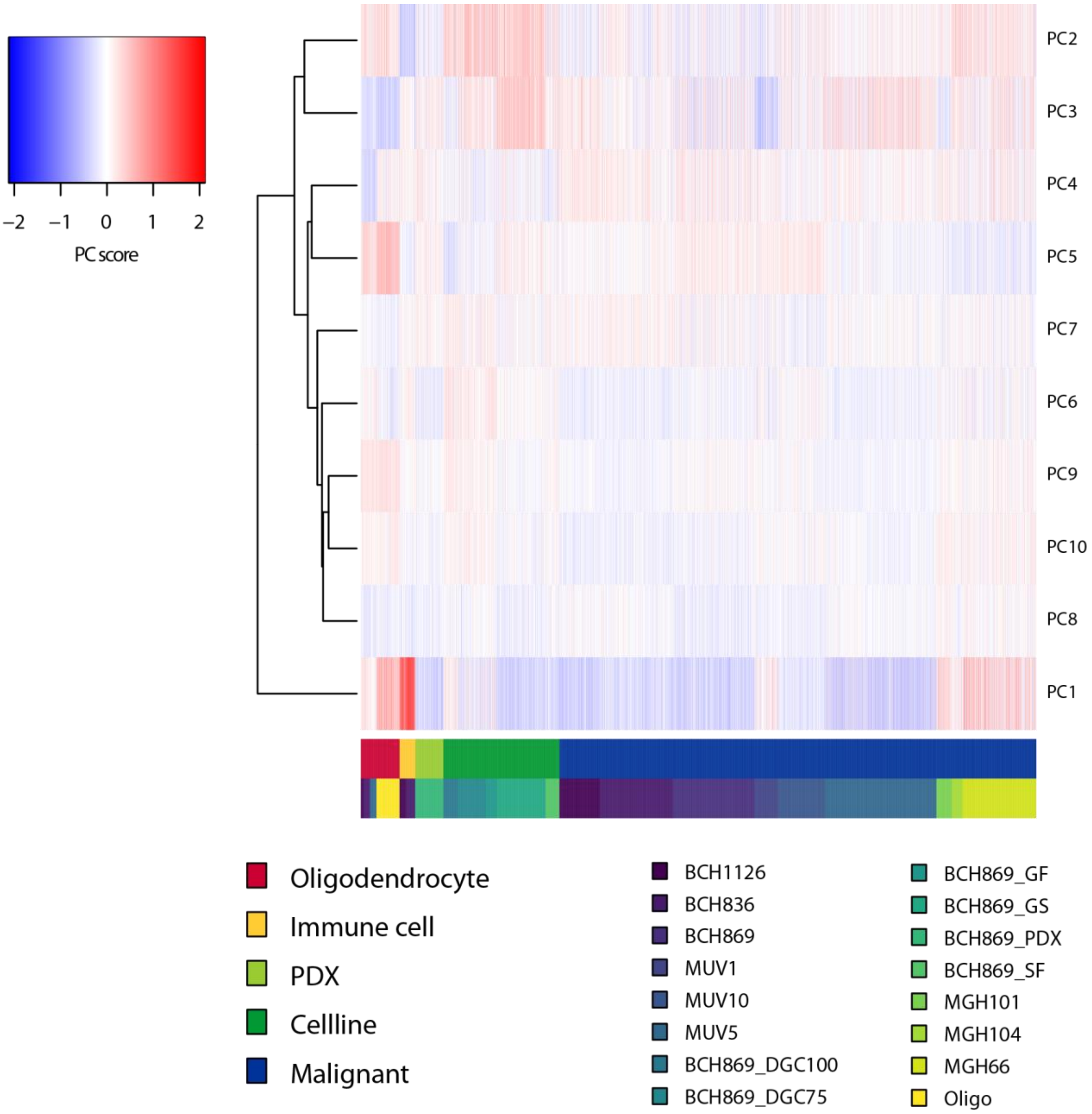
Supplementary Figure 1 - EMT prediction in single-cells from all 4 populations, Related to Figure 1. EMT (blue), DNA repair (green) and glycolysis (red) inductions and control (grey). Black and white bar underneath indicates the reported probability of each cell to be glycolytic (log10 scale). Black: missing values.



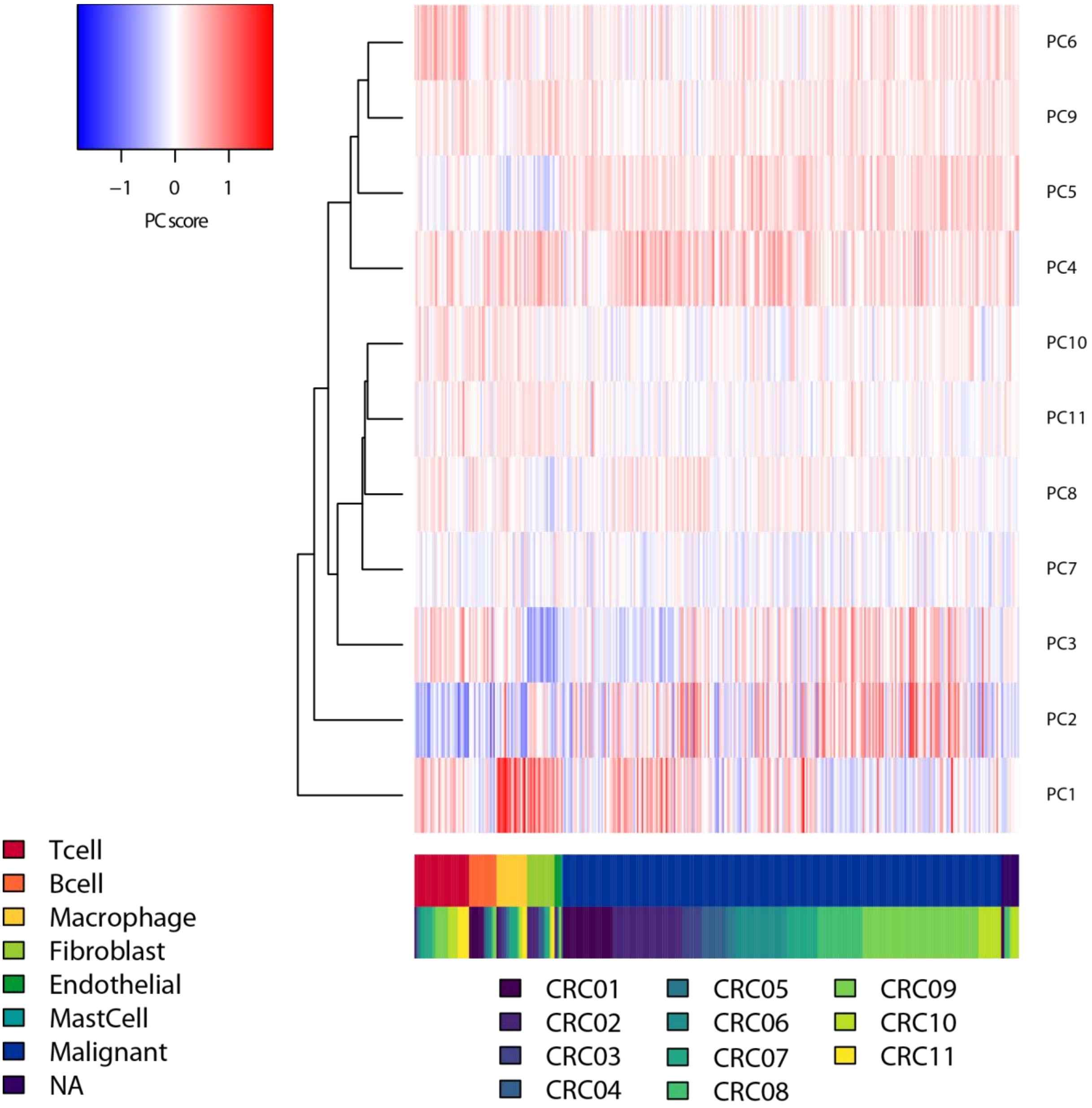
Supplementary Figure 2 – DNA repair prediction in single-cells from all 4 populations, Related to Figure 1. DNA repair (green), EMT (blue) and glycolysis (red) inductions and control (grey). Black and white bar underneath indicates the reported probability of each cell to be glycolytic (log10 scale). Black: missing values.



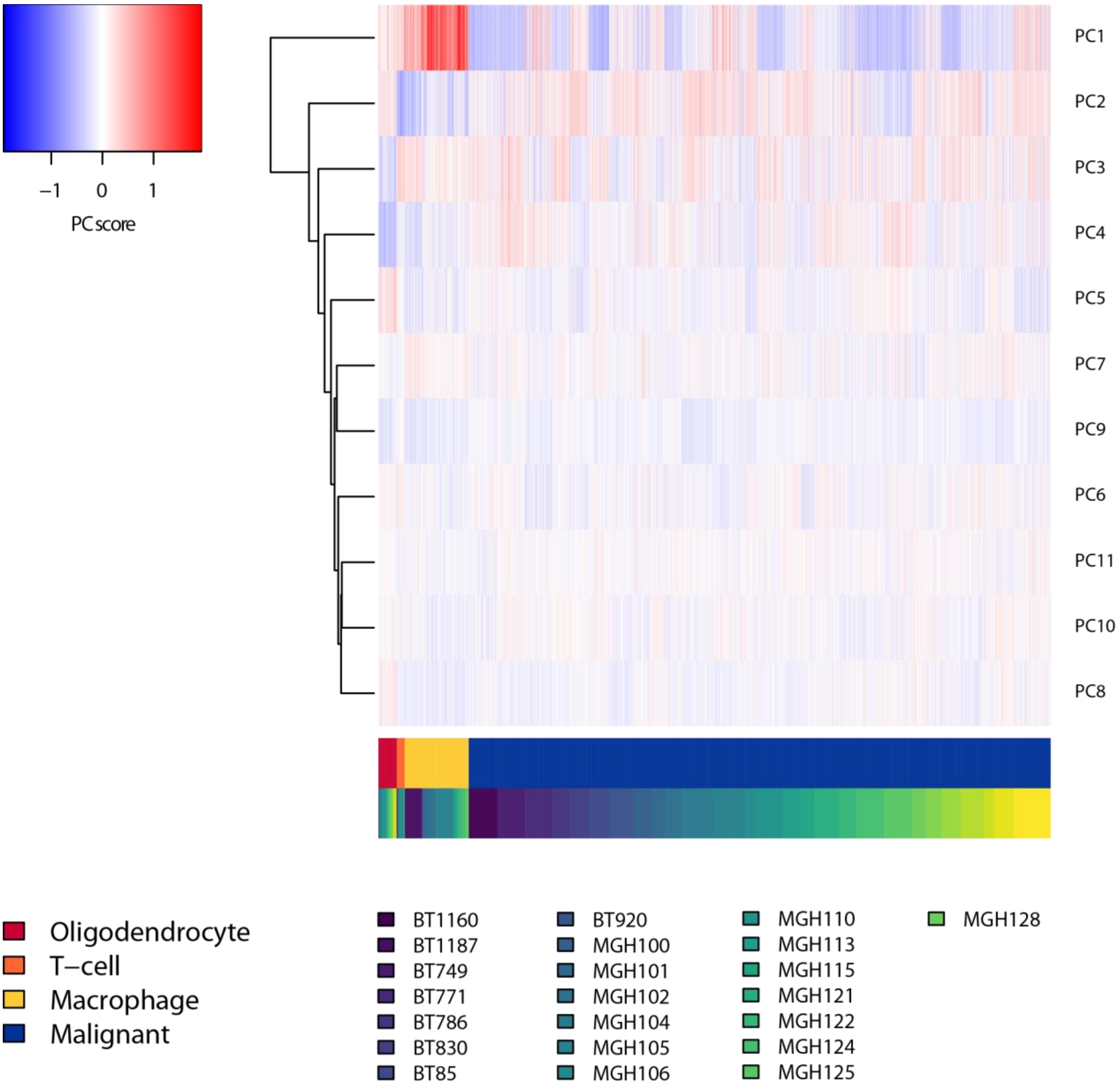
Supplementary Figure 3 – Correlation between scores for cell cycle phases and hallmark activities, Related to Figure 2. The heatmap displays the positive (red) and negative correlation between all cyclone cell cycle phase scores and the normalised activity scores across 20,583 cells from the 6 datasets with metadata.



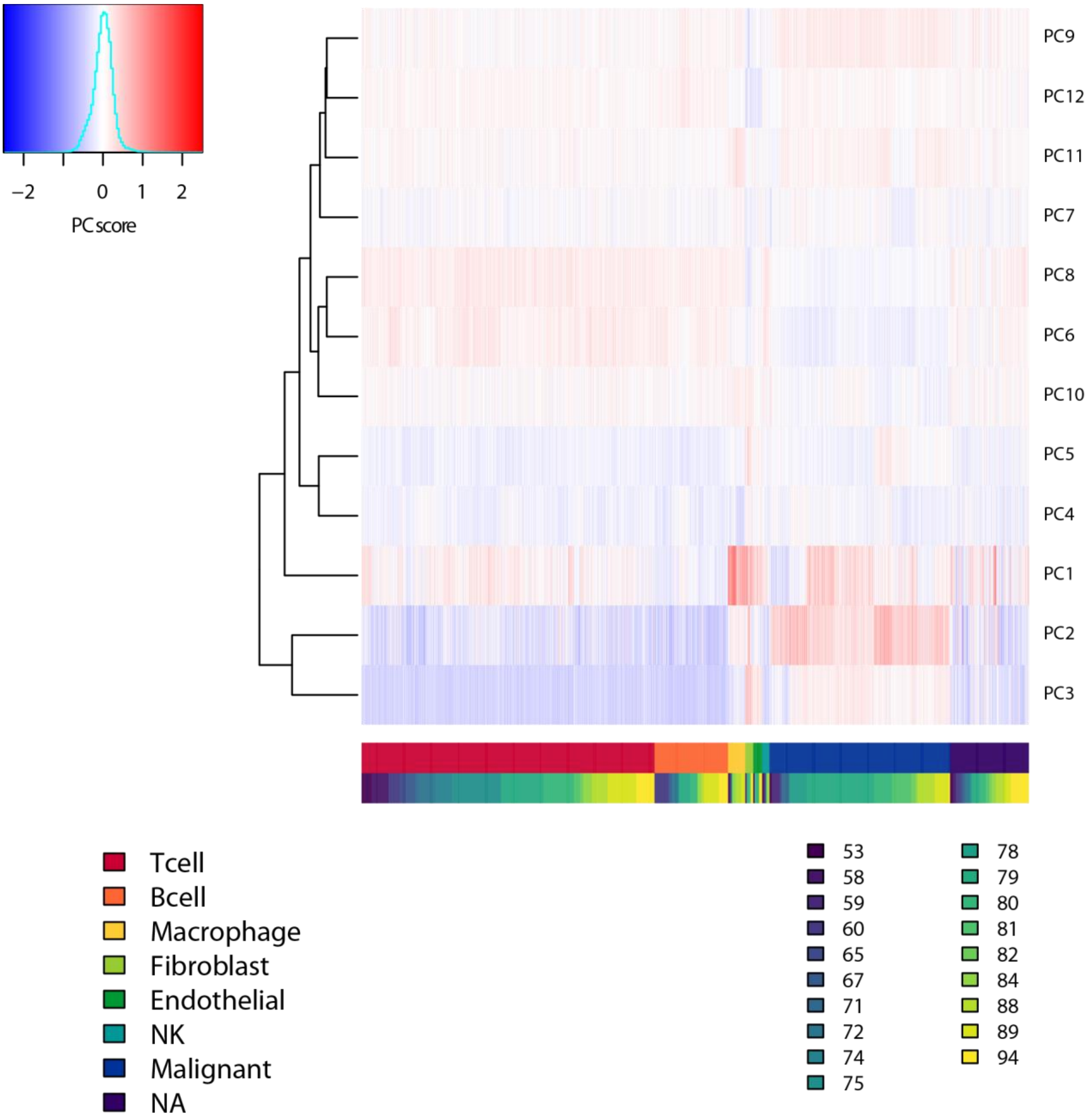
Supplementary Figure 4 – Filbin *et al.* (H3K27M-glioma) activity scores according to patients and tumour types, Related to Figure 2. Heatmap of PCA-based activity scores in the metadataset, normalised per activity per set. Dendrogram highlights relationships between activities (left). Bottom colour bars report the cell type (top) and patient of origin (bottom) of each cell. Top left colour scale corresponds to the normalised activity score for each activity in each cell. Only principal components explaining >2% of total variance were included.



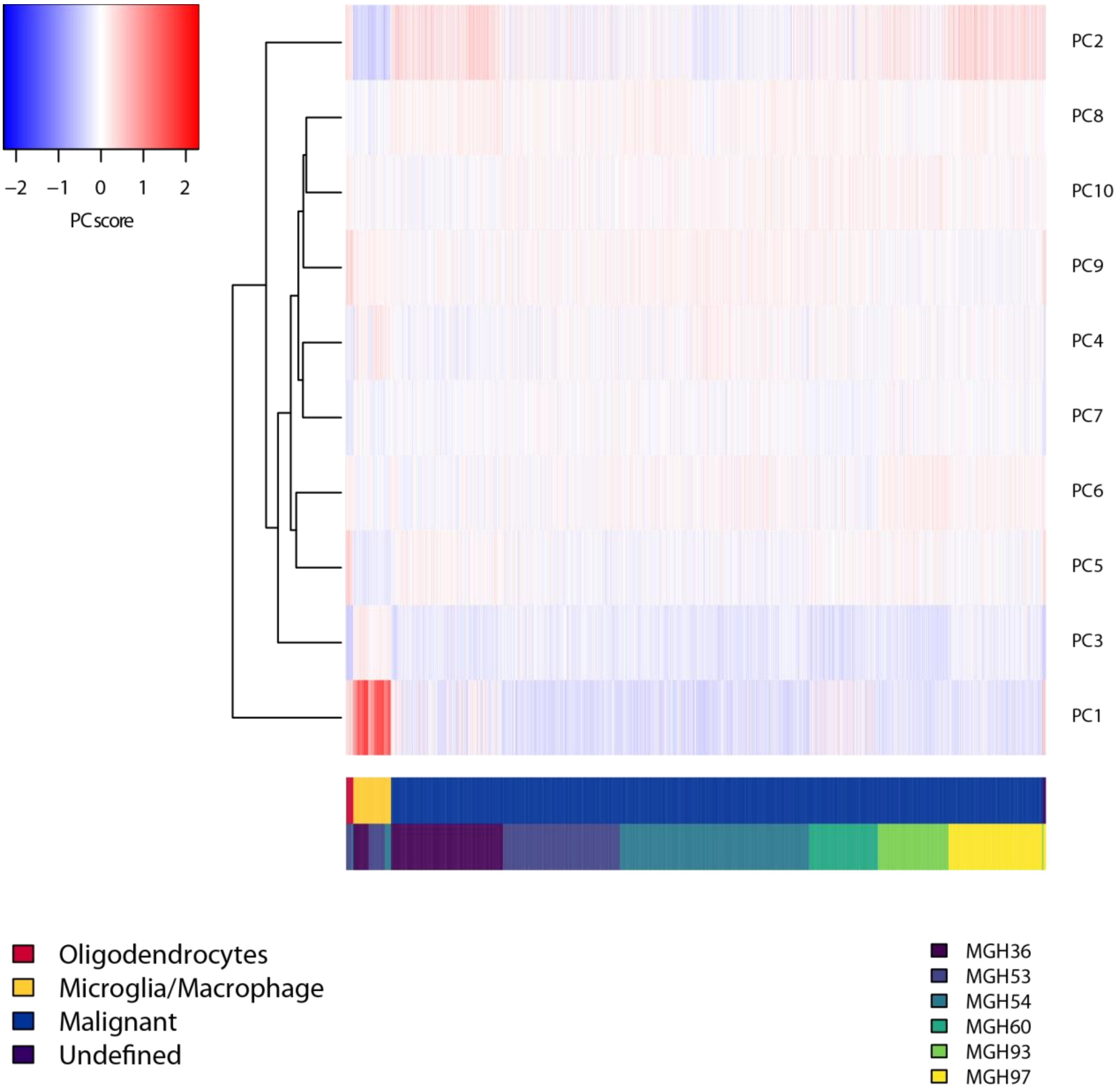
Supplementary Figure 5 – Li *et al.* (colorectal cancer) activity scores according to patients and tumour types, Related to Figure 2. Heatmap of PCA-based activity scores in the metadataset, normalised per activity per set. Dendrogram highlights relationships between activities (left). Bottom colour bars report the cell type (top) and patient of origin (bottom) of each cell. Top left colour scale corresponds to the normalised activity score for each activity in each cell. Only principal components explaining >2% of total variance were included.



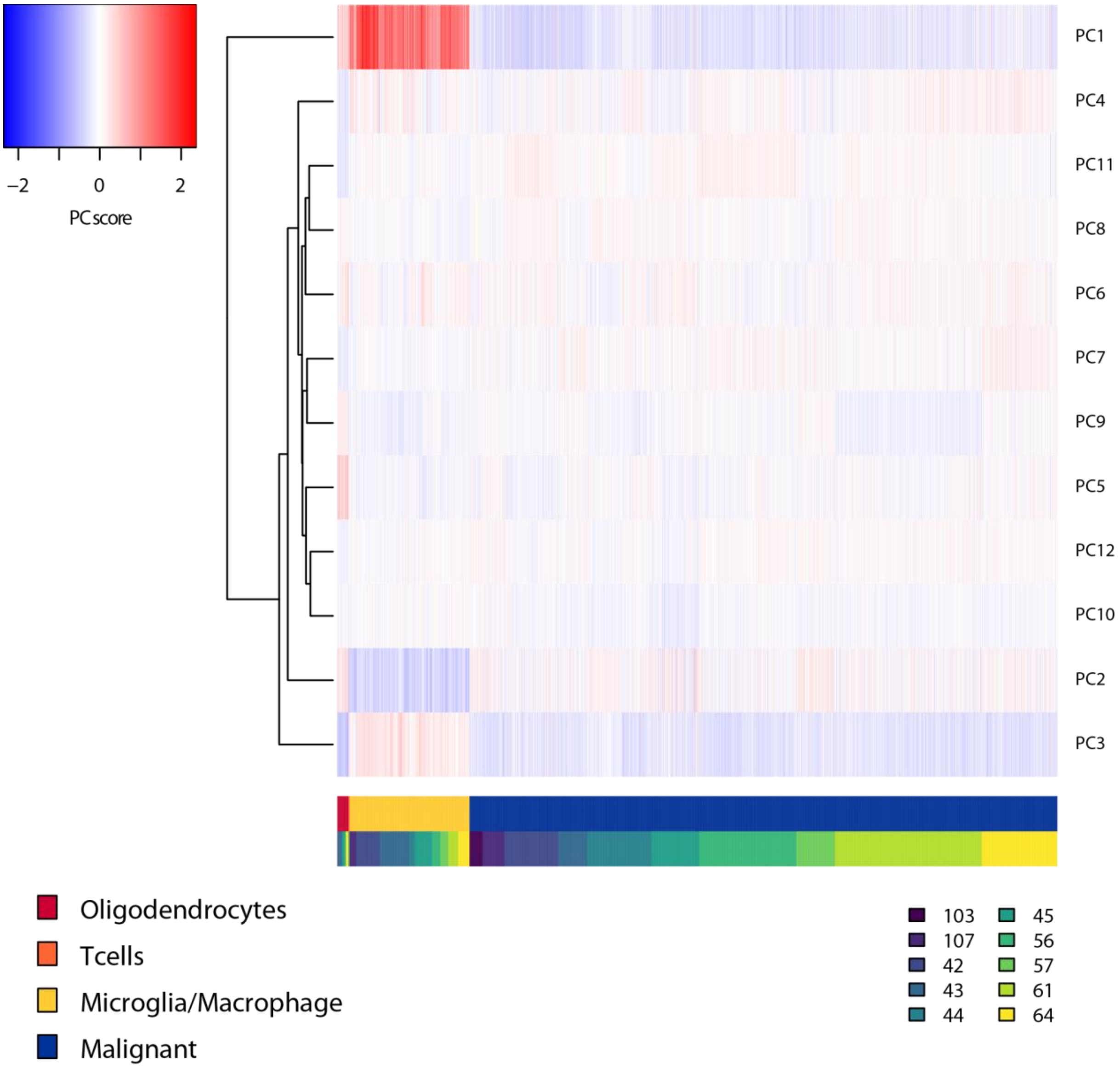
Supplementary Figure 6 – Neftel *et al.* (glioblastoma) activity scores according to patients and tumour types, Related to Figure 2. Heatmap of PCA-based activity scores in the metadataset, normalised per activity per set. Dendrogram highlights relationships between activities (left). Bottom colour bars report the cell type (top) and patient of origin (bottom) of each cell. Top left colour scale corresponds to the normalised activity score for each activity in each cell. Only principal components explaining >2% of total variance were included.



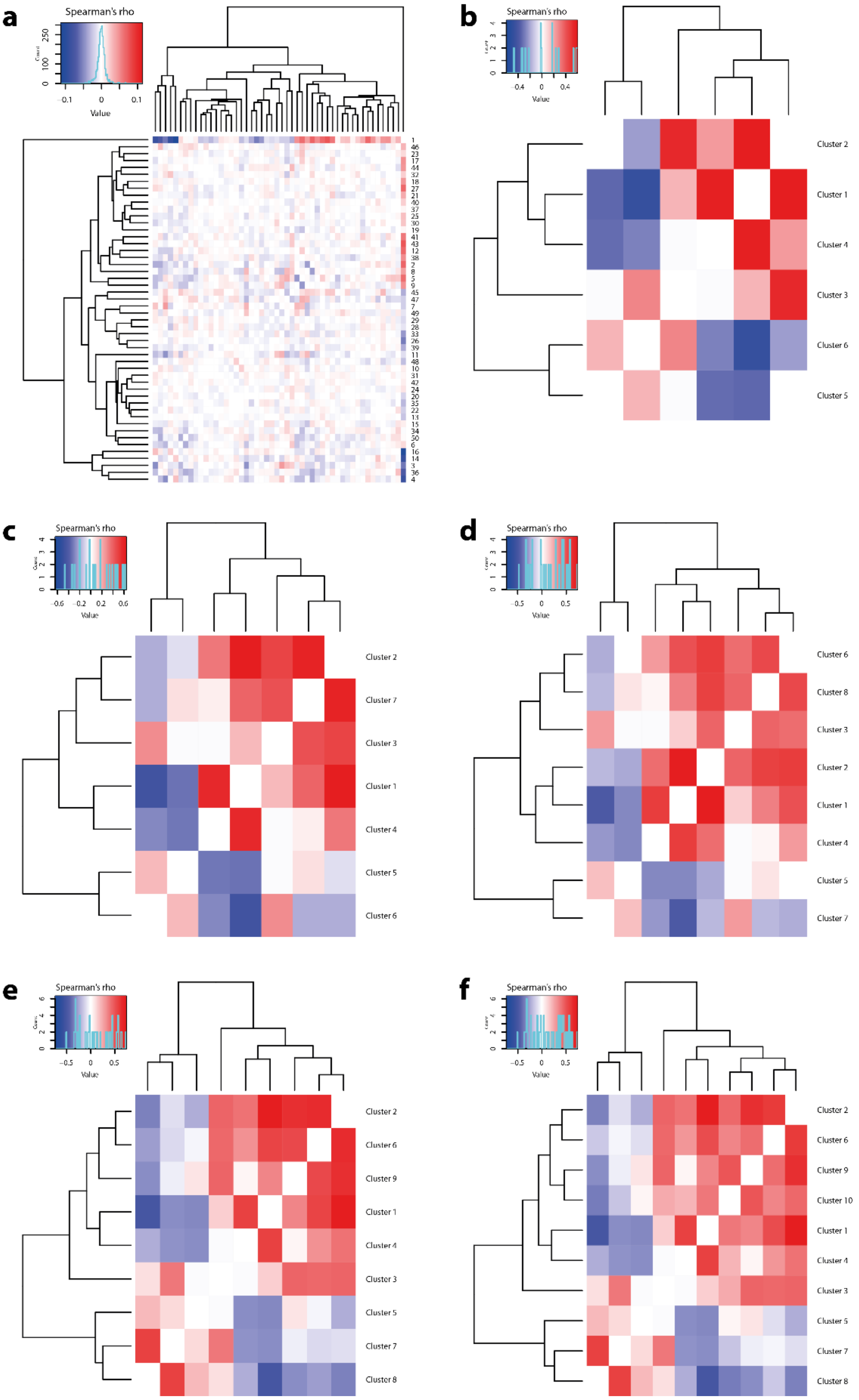
Supplementary Figure 7 – Tirosh *et al.* (melanoma) activity scores according to patients and tumour types, Related to Figure 2. Heatmap of PCA-based activity scores in the metadataset, normalised per activity per set. Dendrogram highlights relationships between activities (left). Bottom colour bars report the cell type (top) and patient of origin (bottom) of each cell. Top left colour scale corresponds to the normalised activity score for each activity in each cell. Only principal components explaining >2% of total variance were included.



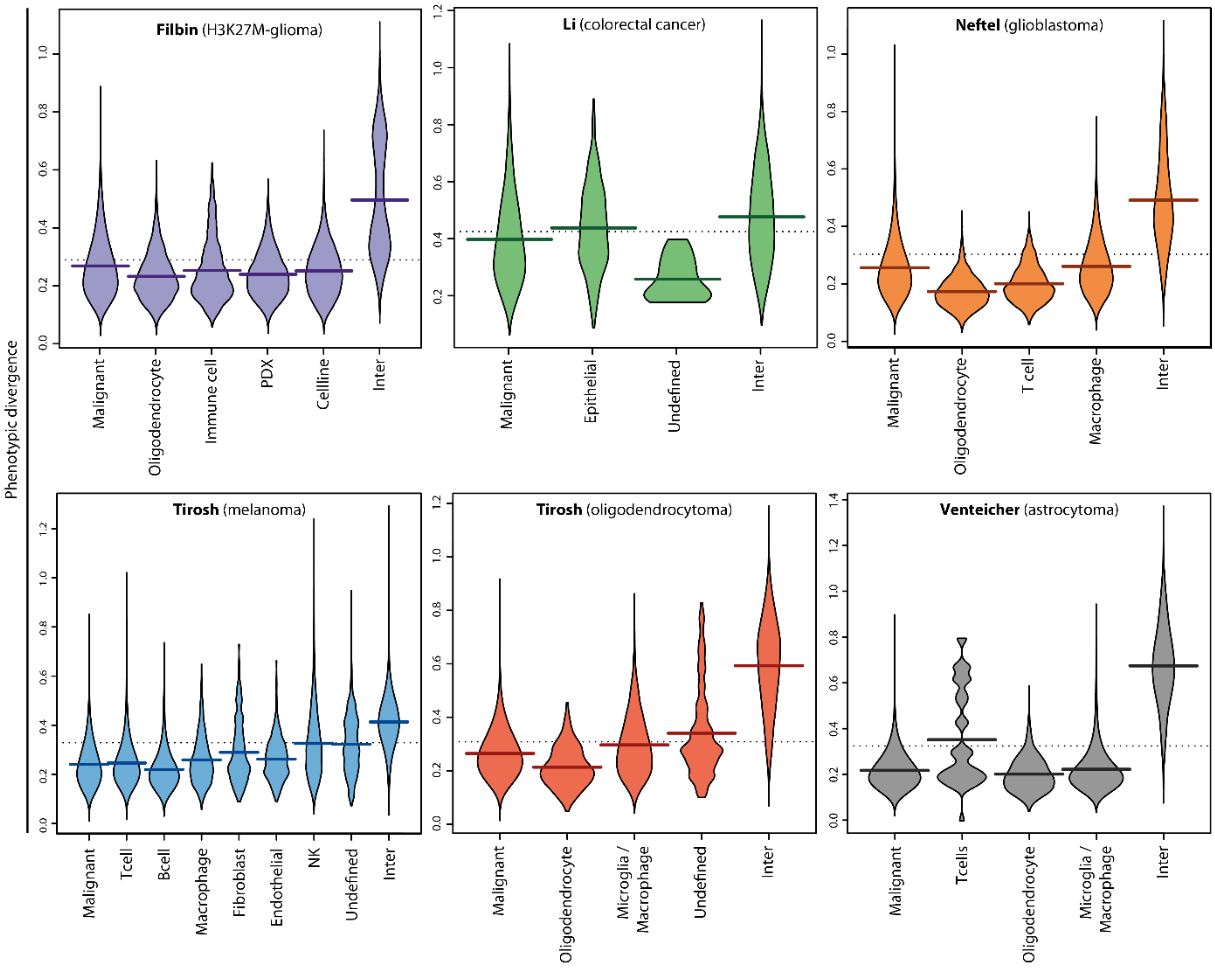
Supplementary Figure 8 – Tirosh *et al.* (oligodendroglioma) activity scores according to patients and tumour types, Related to Figure 2. Heatmap of PCA-based activity scores in the metadataset, normalised per activity per set. Dendrogram highlights relationships between activities (left). Bottom colour bars report the cell type (top) and patient of origin (bottom) of each cell. Top left colour scale corresponds to the normalised activity score for each activity in each cell. Only principal components explaining >2% of total variance were included.



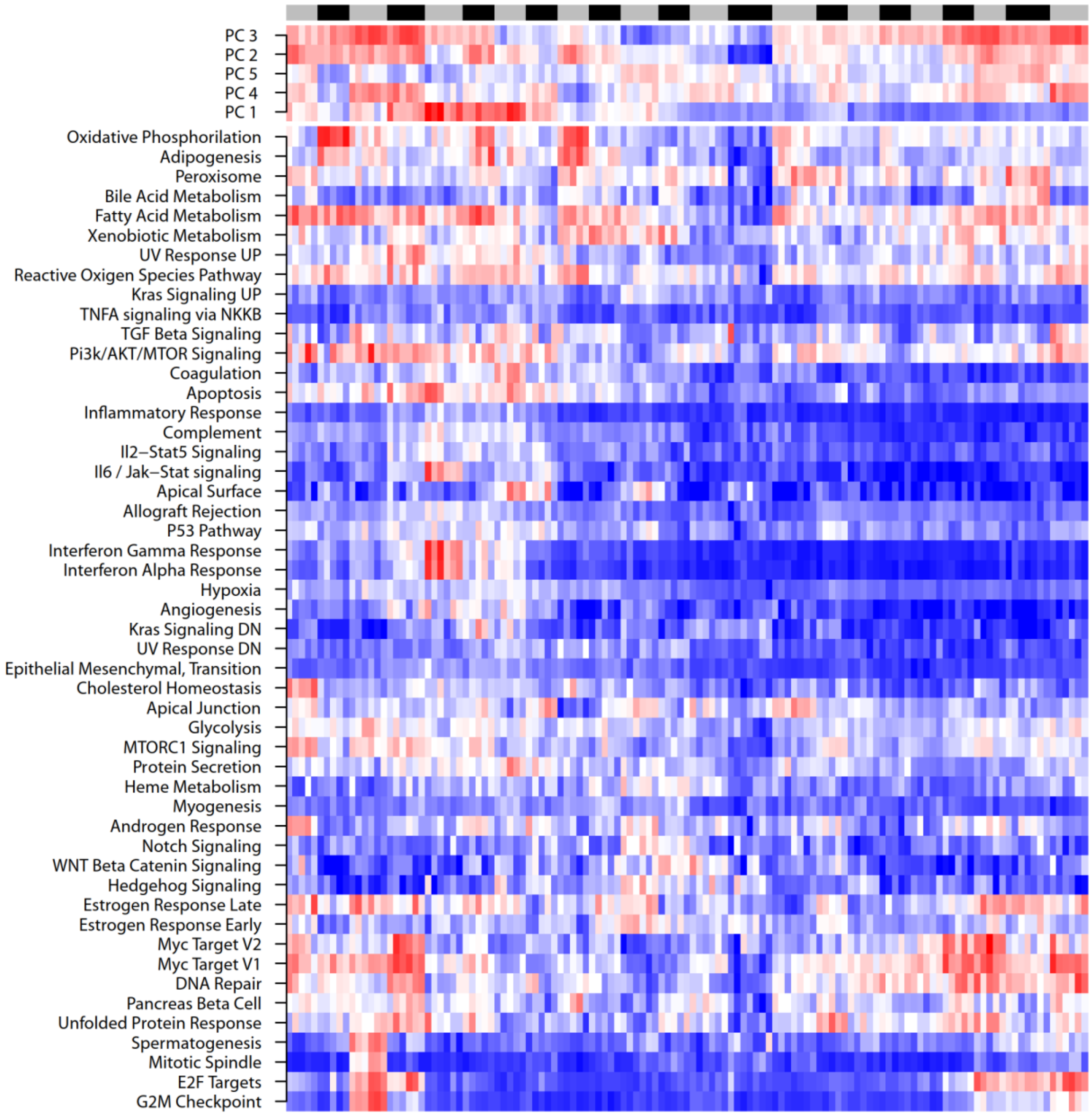
Supplementary Figure 9 – Venteicher *et al.* (astrocytoma) activity scores according to patients and tumour types, Related to Figure 2. Heatmap of PCA-based activity scores in the metadataset, normalised per activity per set. Dendrogram highlights relationships between activities (left). Bottom colour bars report the cell type (top) and patient of origin (bottom) of each cell. Top left colour scale corresponds to the normalised activity score for each activity in each cell. Only principal components explaining >2% of total variance were included.



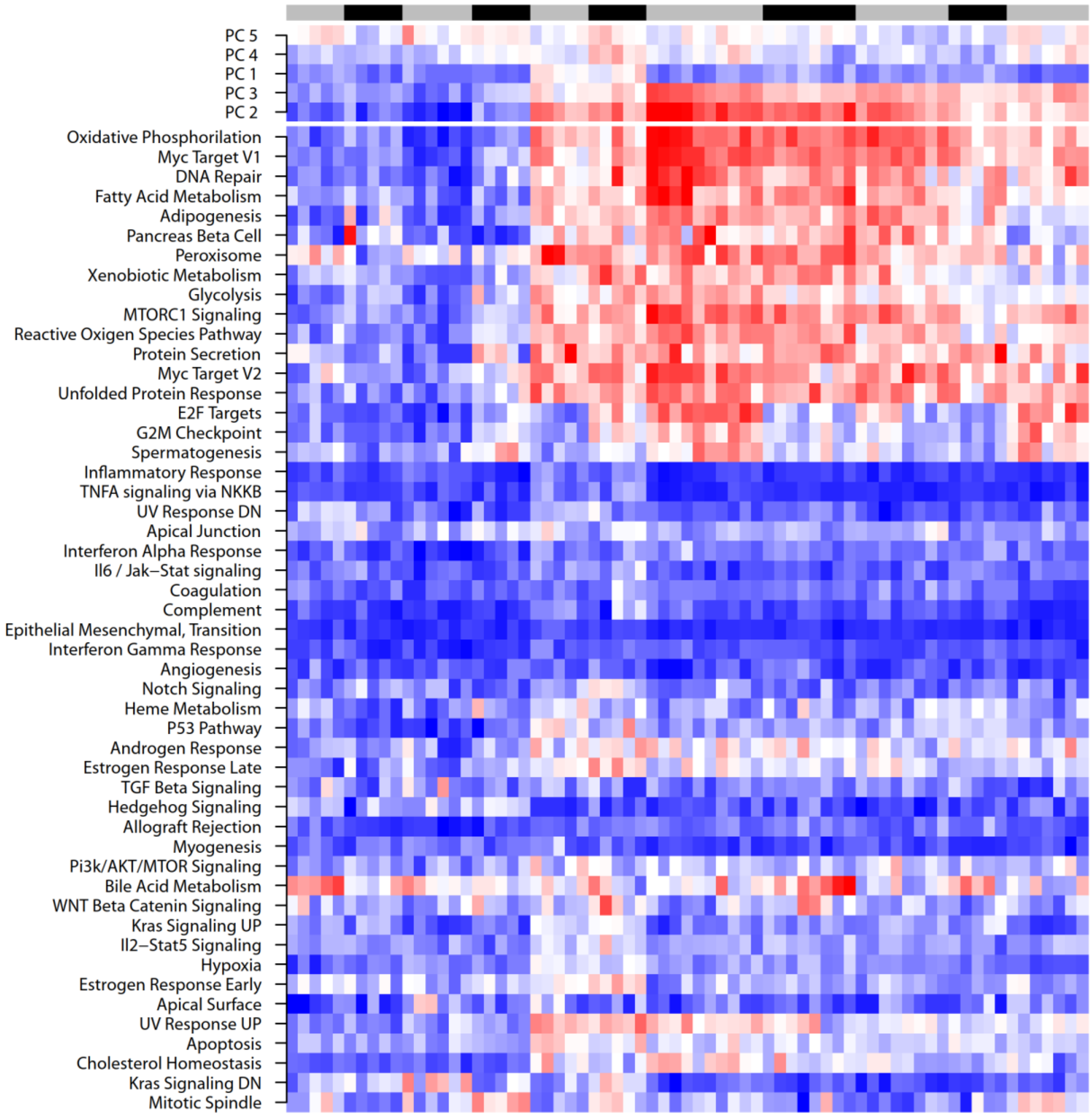
Supplementary Figure 10 – Redundancy among principal components and cluster scores, Related to Figure 3. Spearman correlations between a) all 50 principal components scores and b-f) cluster scores, for 6 to 10 clusters. Scale ranges differ between panels.



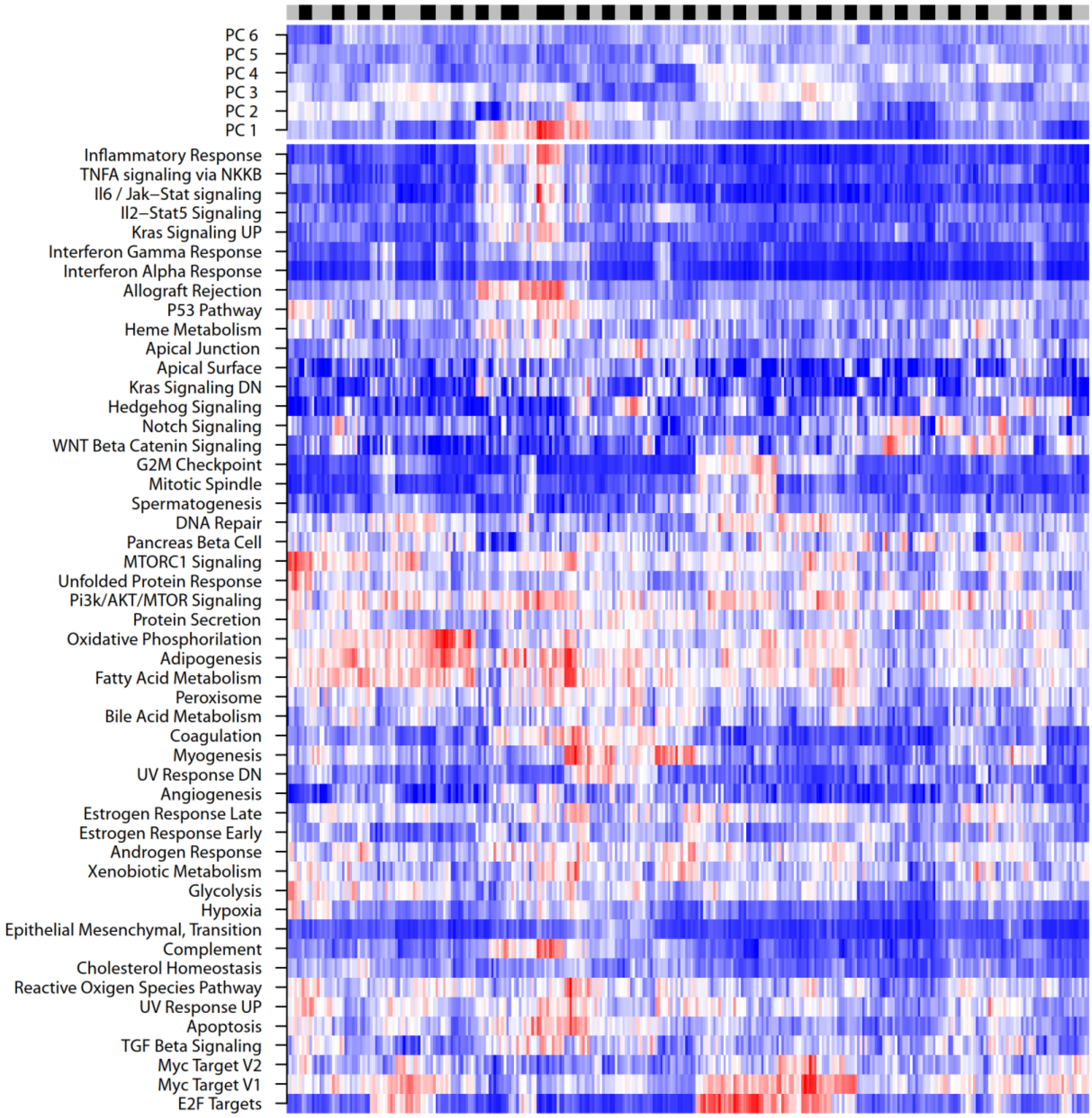
Supplementary Figure 11 – Cluster-based pan-cancer phenotypic cell-cell divergence, Related to Figure 4. Pairwise cell-cell divergence distributions per cell type in each of the 6 datasets with curated metadata. Activities were split into 8 clusters. Inter: inter-type divergence, between cells of different subtypes. All other distributions are between cells of the reported type. Dashed horizontal line: total average; broad horizontal lines: individual distribution averages.



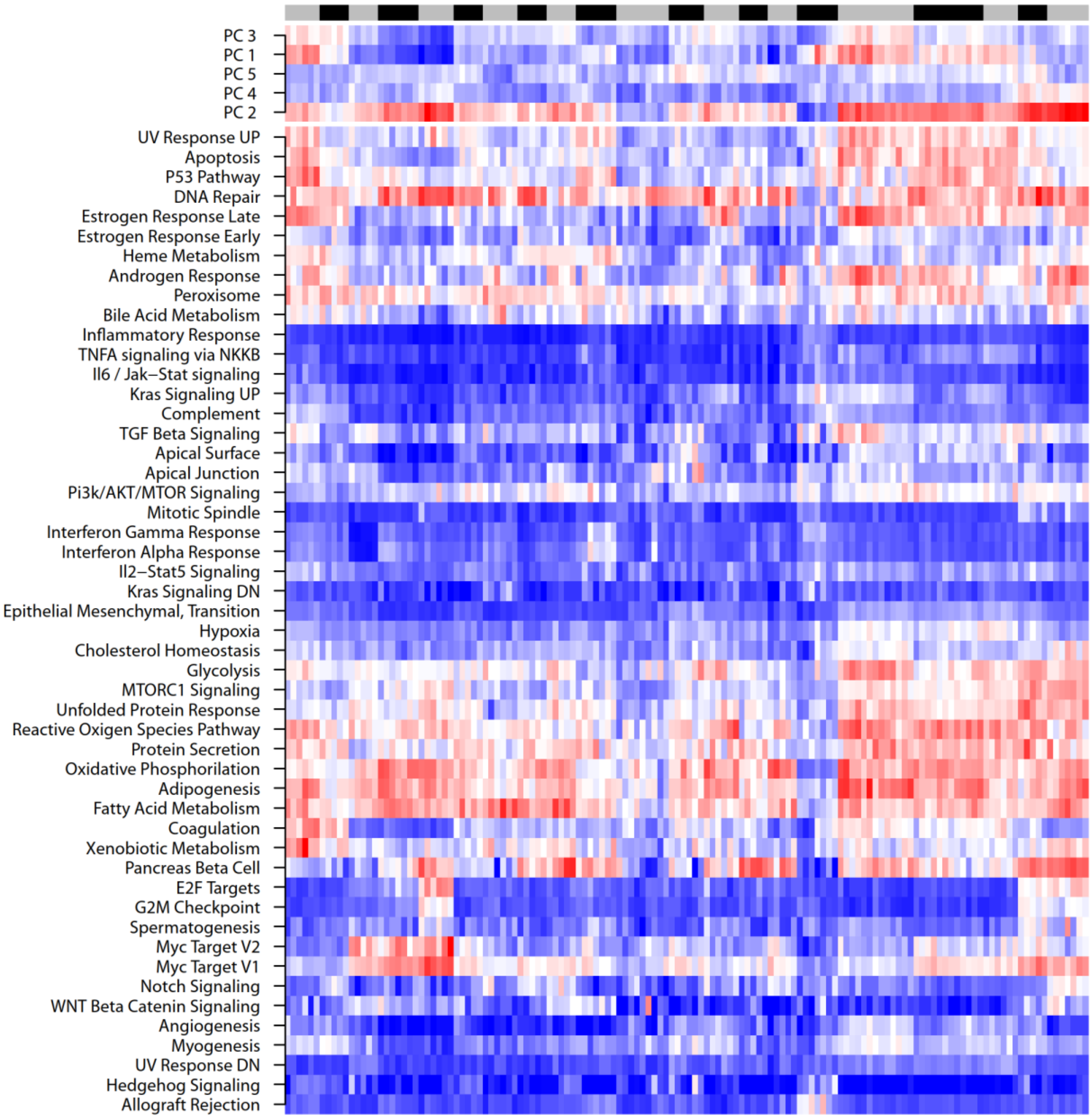
Supplementary Figure 12 – Isolated activity profiles of significant clusters of malignant cells in the Filbin *et al.* pediatric glioma dataset, Related to Figure 5. Top: distinct significant clusters are identified by alternating black and grey colour bars. Cells are ordered left to right according to the overall cluster data including all cells, although only significant clusters of 5 cells or more are displayed. Middle: Heatmap of PCA-based activity scores. Only principal components explaining >3% of total variance were included. PCA scores are ordered top to bottom according to complete hierarchical clustering based on Euclidean distances. Bottom: Heatmap of normalised activity scores, ordered top to bottom according to complete hierarchical clustering based on Euclidean distances.



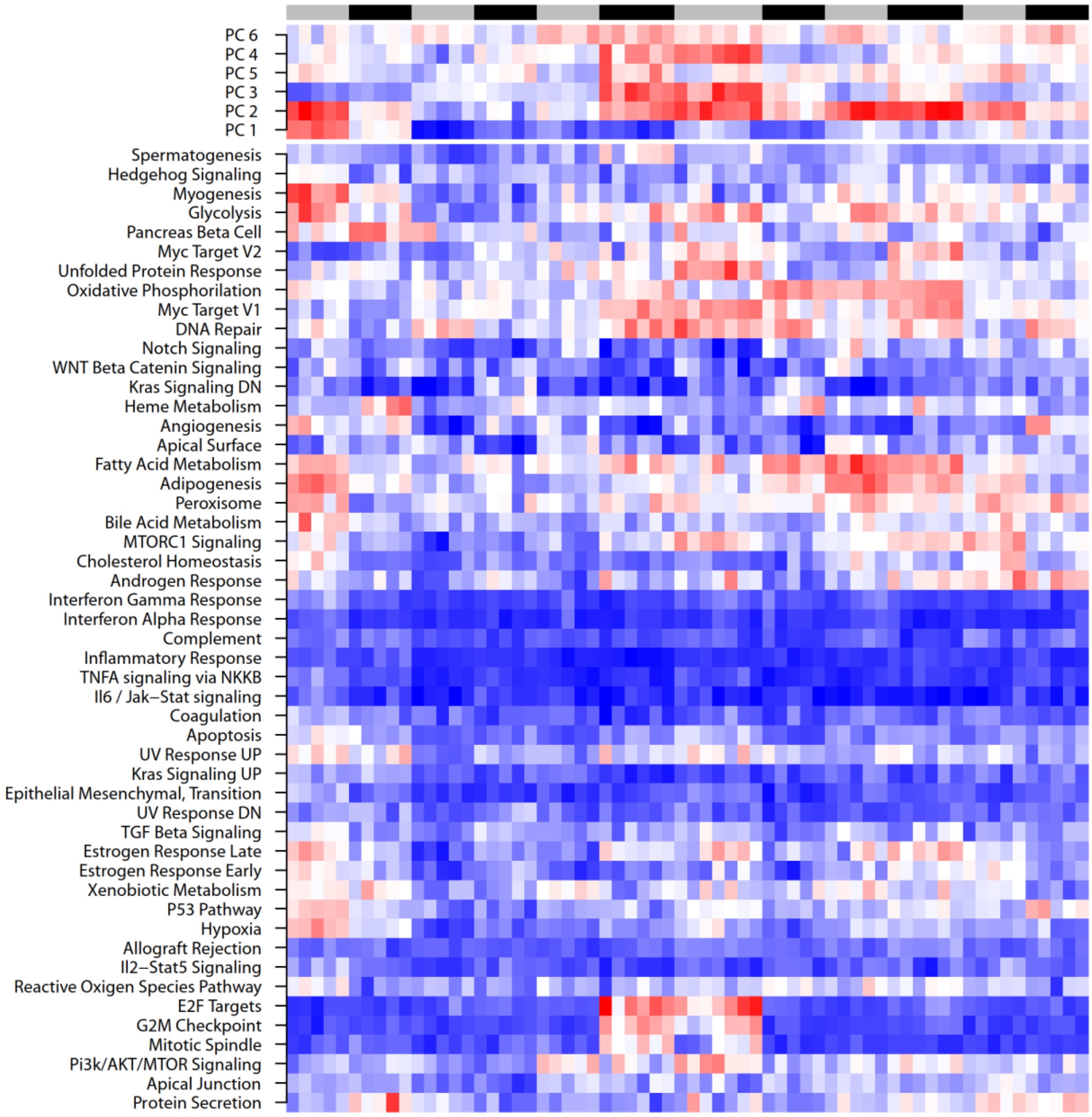
Supplementary Figure 13 – Isolated activity profiles of significant clusters of malignant cells in the Li *et al.* colorectal dataset, Related to Figure 5. Top: distinct significant clusters are identified by alternating black and grey colour bars. Cells are ordered left to right according to the overall cluster data including all cells, although only significant clusters of 5 cells or more are displayed. Middle: Heatmap of PCA-based activity scores. All principal components were used for clustering analyses, but only those explaining >3% of total variance are displayed. PCA scores are ordered top to bottom according to complete hierarchical clustering based on Euclidean distances. Bottom: Heatmap of normalised activity scores, ordered top to bottom according to complete hierarchical clustering based on Euclidean distances.



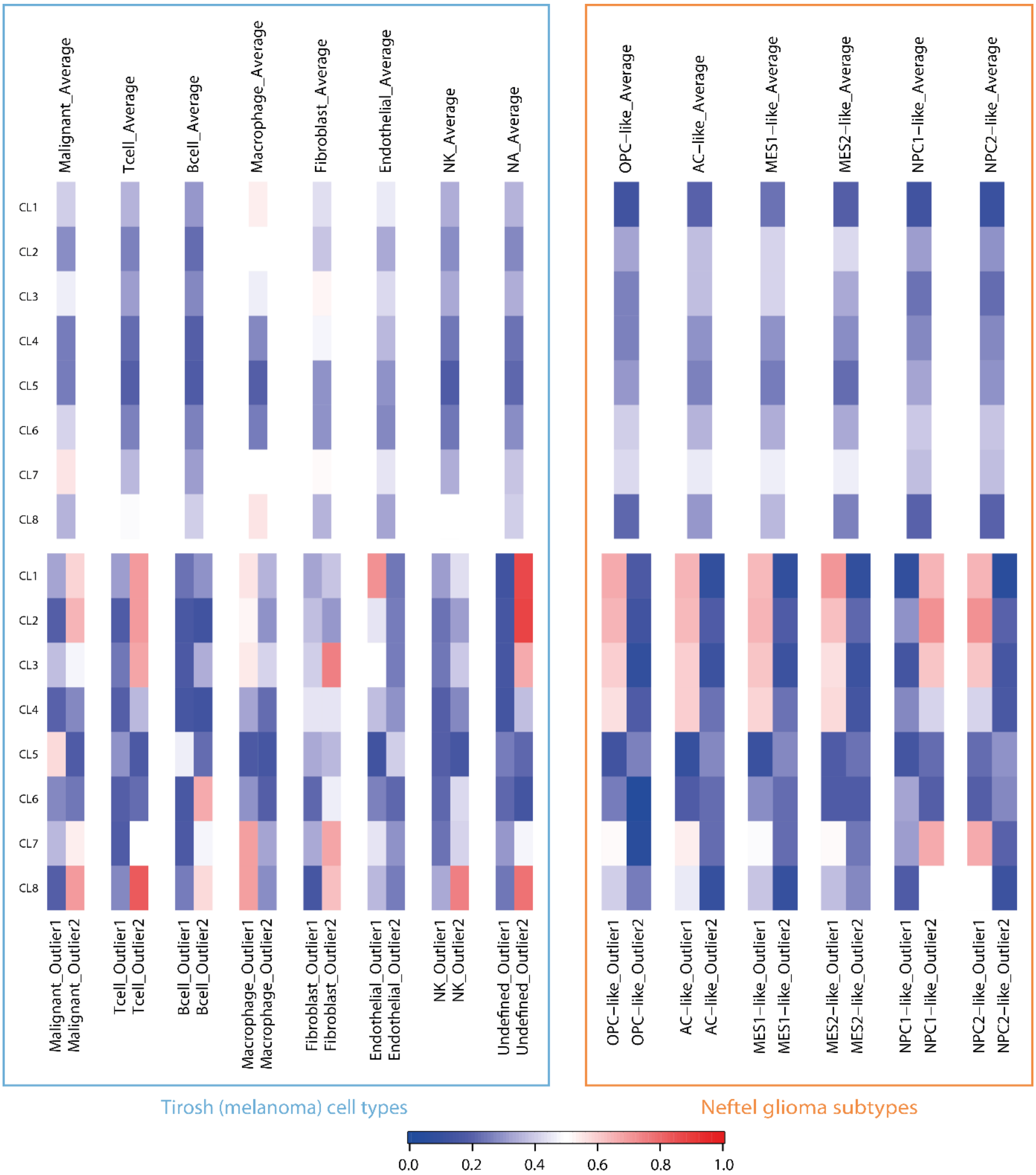
Supplementary Figure 14 – Isolated activity profiles of significant clusters of malignant cells in the Neftel *et al.* glioma dataset, Related to Figure 5. Top: distinct significant clusters are identified by alternating black and grey colour bars. Cells are ordered left to right according to the overall cluster data including all cells, although only significant clusters of 5 cells or more are displayed. Middle: Heatmap of PCA-based activity scores. All principal components were used for clustering analyses, but only those explaining >3% of total variance are displayed. PCA scores are ordered top to bottom according to complete hierarchical clustering based on Euclidean distances. Bottom: Heatmap of normalised activity scores, ordered top to bottom according to complete hierarchical clustering based on Euclidean distances.



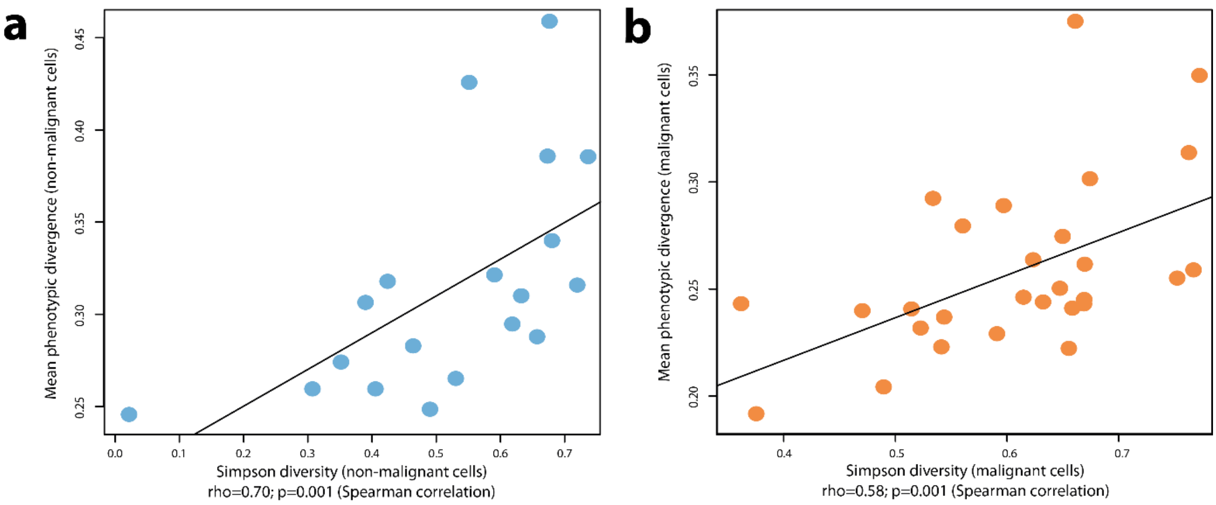
Supplementary Figure 15 – Isolated activity profiles of significant clusters of malignant cells in the Tirosh *et al.* melanoma dataset, Related to Figure 5. Top: distinct significant clusters are identified by alternating black and grey colour bars. Cells are ordered left to right according to the overall cluster data including all cells, although only significant clusters of 5 cells or more are displayed. Middle: Heatmap of PCA-based activity scores. All principal components were used for clustering analyses, but only those explaining >3% of total variance are displayed. PCA scores are ordered top to bottom according to complete hierarchical clustering based on Euclidean distances. Bottom: Heatmap of normalised activity scores, ordered top to bottom according to complete hierarchical clustering based on Euclidean distances.



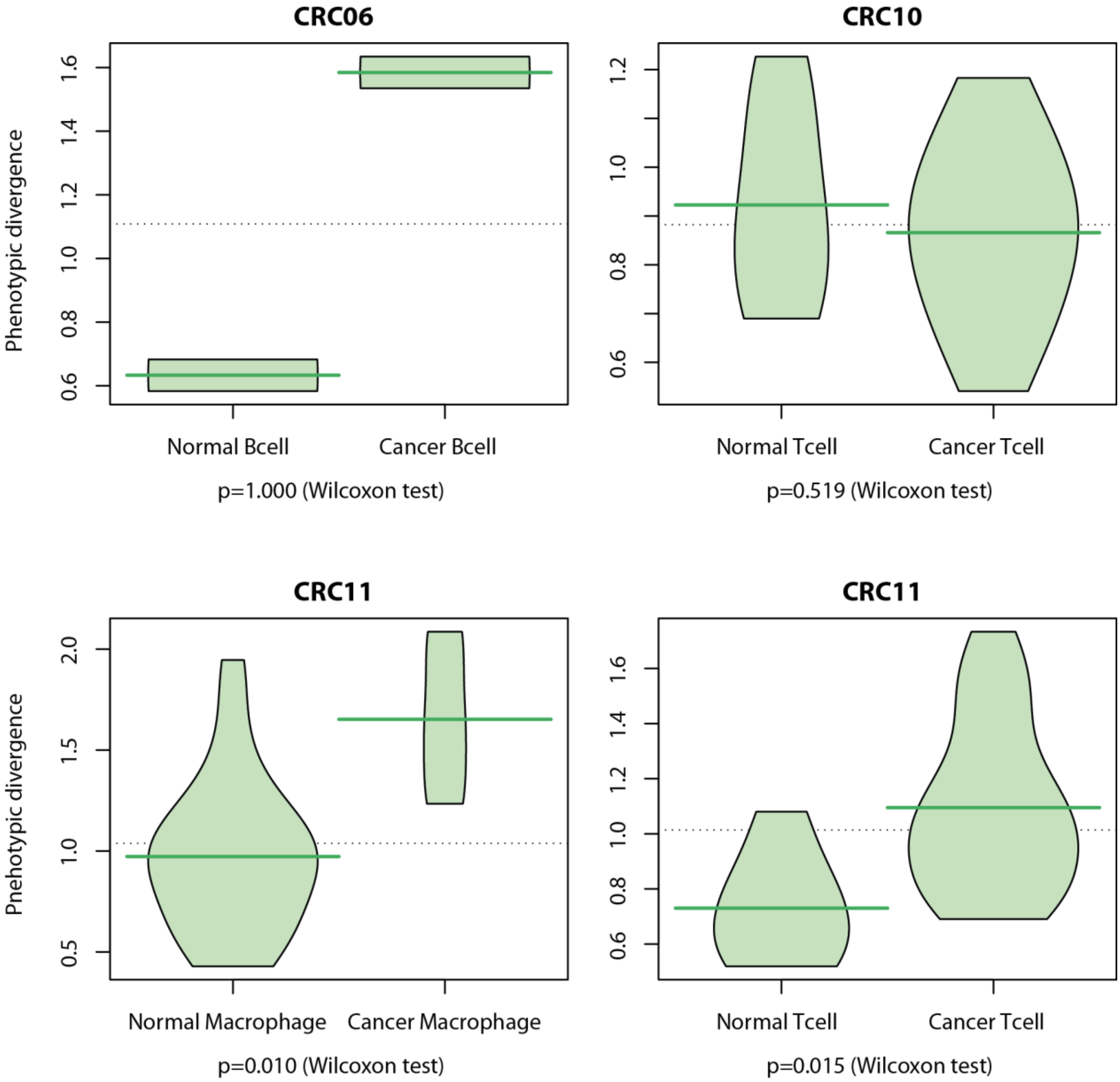
Supplementary Figure 16 – Isolated activity profiles of significant clusters of malignant cells in the Tirosh *et al.* oligodendroglioma dataset, Related to Figure 5. Top: distinct significant clusters are identified by alternating black and grey colour bars. Cells are ordered left to right according to the overall cluster data including all cells, although only significant clusters of 5 cells or more are displayed. Middle: Heatmap of PCA-based activity scores. All principal components were used for clustering analyses, but only those explaining >3% of total variance are displayed. PCA scores are ordered top to bottom according to complete hierarchical clustering based on Euclidean distances. Bottom: Heatmap of normalised activity scores, ordered top to bottom according to complete hierarchical clustering based on Euclidean distances.



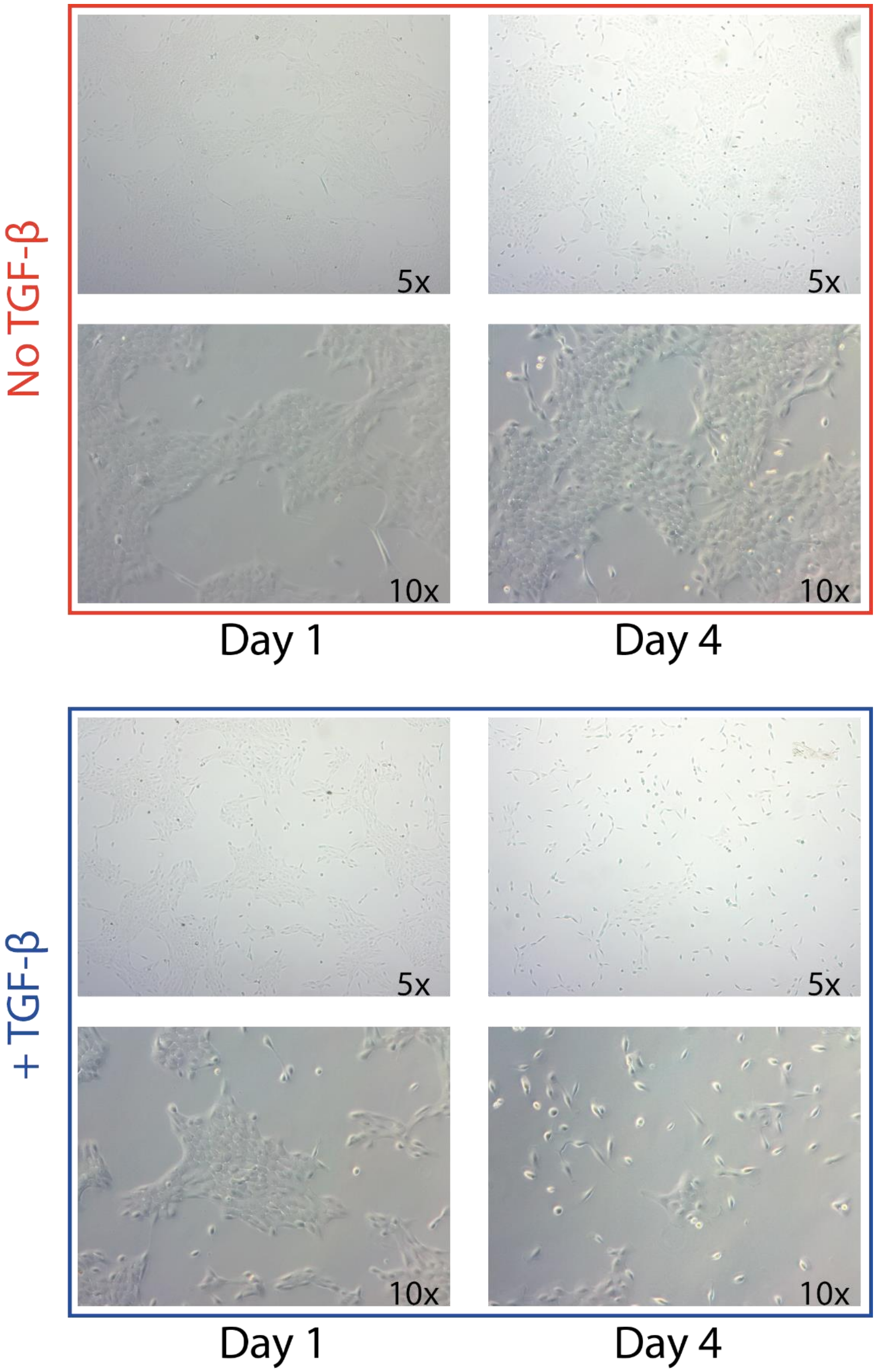
Supplementary Figure 17 – Cluster-based activity-centred phenotypic profiles in populations of known structure, Related to Figure 6. Description of 7 non-malignant cell types from the Tirosh *et al.* melanoma dataset and (blue, left) 6 glioma subtypes from the Neftel *et al.* H3K27M-glioma dataset (orange, right). Average profiles on top were obtained by averaging all cells from a given subtype across all patients. The outlier profiles at the bottom were obtained from the same-type cell pairs displaying the highest activity-based divergence for each cell type. Activities were split into 8 clusters.



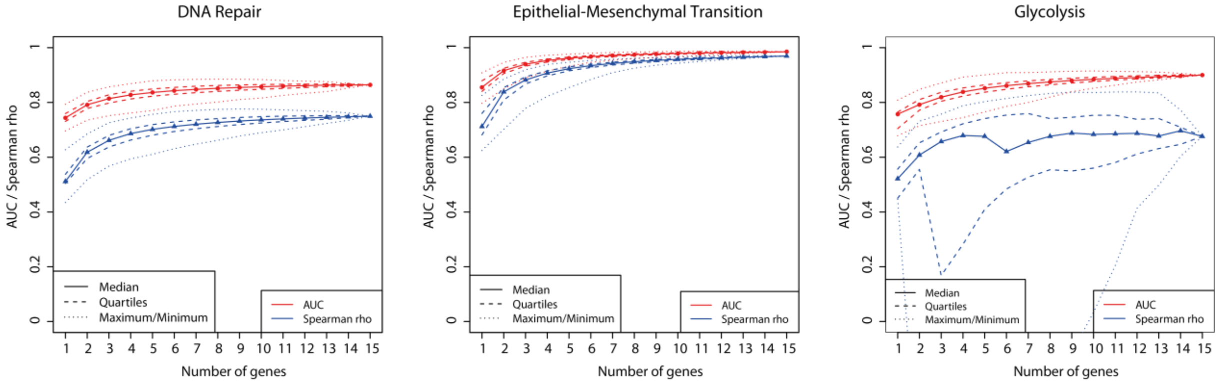
Supplementary Figure 18 – Correlation between cluster-based, activity-centred phenotypic diversity and Simpson index in populations of known structure, Related to Figure 6. a) Relationship between mean phenotypic divergence between non-malignant cells in the Tirosh *et al.* melanoma dataset and the Simpson diversity index calculated on the repartition of cells into 7 non-malignant classes. b) Relationship between mean phenotypic divergence between malignant cells in the Neftel *et al.* H3K27M-glioma dataset and the Simpson diversity index calculated on the classification of cells into 6 malignant subtypes. Black lines: linear models.



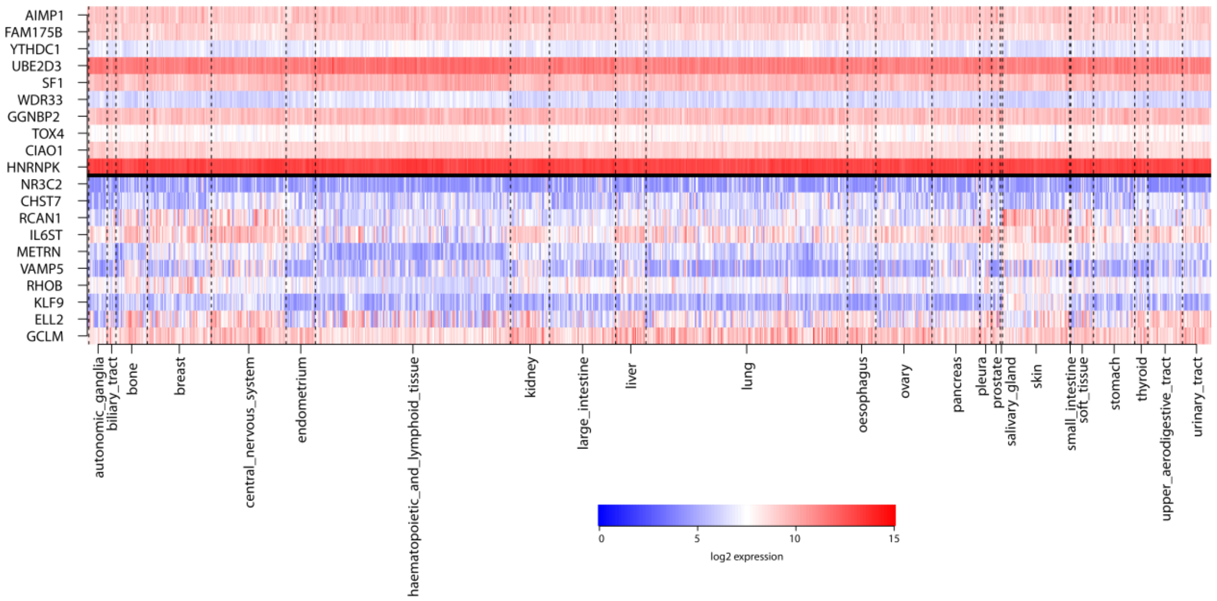
Supplementary Figure 19 – Divergence in specific cell types between normal and tumour tissue in Li *et al.* colorectal cancer data, Related to Figure 7. Dashed horizontal line: total average; broad horizontal lines: individual distribution averages. Patient of occurrence reported atop each plot, cell type below.



Supplementary Figure 20 – EMT induction by TGF-β addition to the media in MCF10A cells, Related to Figure 1. Cell morphology and organisation at days 1 and 4, at 5x and 10x zooms. No TGF-β was added to the negative control (red, top). In the positive control (blue, bottom), TGF-β was added continuously to the medium from the morning of day 1.



Supplementary Figure 21 – AUC and Spearman rho upon reduced gene panel usage, Related to Figure 1. AUC in predicting the top 50% from the bottom 50% samples from the TCGA using reduced gene panels based on CCLE data (red), and correlation between reduced panel scores and whole-signature ssGSEA scores (Spearman's rho, blue), according to the number of genes included to designed reduced panels.



Supplementary Figure 22 – Most and least variable genes across different tissues, Related to Figure 1. Least (top) and most (bottom) variable 10 genes across multiple tumour types in TCGA data. Based on publically available RSEM data, displayed using log2 transformation.

Supplementary Tables

Supplementary Table 1 – Best combinations per activity, Related to Figure 1. Predictive power and correlation between the three reduced gene panels and signal from whole signature Gene Set Enrichment Analysis. Predictive power was assessed by measuring the ability to tease the bottom 50% whole signature GSEA scoring samples from the top 50% ones, using the reduced panels in the Cancer Cell Line Encyclopedia (CCLE, discovery) and (TCGA, validation). AUC: Area Under the Curve. rho: Spearman's rho correlation.

			CCLE		**TCGA**	
Hallmark	**n Genes**	**Genes**	**AUC**	**rho**	**AUC**	**rho**
EMT	7	NNMT, COL4A1, SNAI2, FBN1, CTGF, FSTL1, FN1	0.98	0.95	0.96	0.92
DNA repair	7	POLR2F, RFC5, POLR2E, CLP1, SF3A3, POLD1, DUT	0.88	0.78	0.86	0.75
Glycolysis	6	EXT1, AGRN, SLC16A3, PYGB, PYGL, IL13RA1	0.91	0.83	0.79	0.61

Supplementary Table 2 – Activity prediction, Related to Figure 1. * Sensitivity and specificity calculated using a 0.1 cut-off on the predicted glm score. Coloured rows indicate the best AUC for each activity.

	Genes	AUC	Sensitivity*	Specificity*
EMT	COL4A1	0.84	0.88	0.96
	ITGAV	0.90	0.88	0.71
	COL4A1, ITGAV	0.99	0.88	0.96
DNA repair	SF3A3	0.68	0.90	0.35
	XPC	0.56	0.90	0.35
	SF3A3, XPC	0.72	0.80	0.50
Glycolysis	PYGL	0.63	0.83	0.00
	PYGB	0.48	1.00	0.00
	AGRN	0.66	0.92	0.00
	EXT1	0.74	0.92	0.46
	PFKM	0.67	1.00	0.00
	LDHA	0.80	0.92	0.79
	PYGL, PYGB	0.59	0.83	0.00
	PYGL, AGRN	0.58	0.92	0.04
	PYGL, EXT1	0.76	0.83	0.50
	PYGL, PFKM	0.66	0.83	0.04
	PYGL, LDHA	0.84	0.92	0.79
	PYGB, AGRN	0.53	1.00	0.08
	PYGB, EXT1	0.72	0.92	0.38
	PYGB, PFKM	0.70	0.92	0.17
	PYGB, LDHA	0.82	0.92	0.79
	AGRN, EXT1	0.74	0.92	0.42
	AGRN, PFKM	0.64	0.92	0.04
	AGRN, LDHA	0.81	0.92	0.79
	EXT1, PFKM	0.77	0.92	0.50
	EXT1, LDHA	0.79	0.83	0.71
	PFKM, LDHA	0.86	0.92	0.79
	PYGL, PYGB, AGRN	0.56	0.83	0.04
	PYGL, PYGB, EXT1	0.73	0.83	0.46
	PYGL, PYGB, PFKM	0.67	0.92	0.13
	PYGL, PYGB, LDHA	0.82	0.92	0.79
	PYGL, AGRN, EXT1	0.77	0.83	0.46
	PYGL, AGRN, PFKM	0.61	0.75	0.08
	PYGL, AGRN, LDHA	0.84	0.92	0.75
	PYGL, EXT1, PFKM	0.77	0.92	0.58
	PYGL, EXT1, LDHA	0.80	0.83	0.71
	PYGL, PFKM, LDHA	0.84	0.92	0.79
	PYGB, AGRN, EXT1	0.75	0.92	0.33
	PYGB, AGRN, PFKM	0.69	0.83	0.13
	PYGB, AGRN, LDHA	0.82	0.92	0.75
	PYGB, EXT1, PFKM	0.74	0.92	0.46
	PYGB, EXT1, LDHA	0.82	0.83	0.71
	PYGB, PFKM, LDHA	0.86	0.92	0.67
	AGRN, EXT1, PFKM	0.75	0.83	0.50
	AGRN, EXT1, LDHA	0.78	0.83	0.71
	AGRN, PFKM, LDHA	0.85	0.92	0.79
	EXT1, PFKM, LDHA	0.78	0.83	0.71

	PYGL, PYGB, AGRN, EXT1	0.73	0.83	0.54
	PYGL, PYGB, AGRN, PFKM	0.63	0.83	0.13
	PYGL, PYGB, AGRN, LDHA	0.82	0.92	0.83
	PYGL, PYGB, EXT1, PFKM	0.73	0.75	0.54
	PYGL, PYGB, EXT1, LDHA	0.81	0.83	0.75
	PYGL, PYGB, PFKM, LDHA	0.84	0.92	0.71
	PYGL, AGRN, EXT1, PFKM	0.77	0.83	0.50
	PYGL, AGRN, EXT1, LDHA	0.78	0.83	0.71
	PYGL, AGRN, PFKM, LDHA	0.83	0.92	0.79
	PYGL, EXT1, PFKM, LDHA	0.79	0.83	0.71
	PYGB, AGRN, EXT1, PFKM	0.76	0.83	0.46
	PYGB, AGRN, EXT1, LDHA	0.81	0.83	0.63
	PYGB, AGRN, PFKM, LDHA	0.85	0.92	0.67
	PYGB, EXT1, PFKM, LDHA	0.82	0.83	0.71
	AGRN, EXT1, PFKM, LDHA	0.78	0.83	0.67
	PYGL, PYGB, AGRN, EXT1, PFKM	0.72	0.83	0.50
	PYGL, PYGB, AGRN, EXT1, LDHA	0.80	0.83	0.71
	PYGL, PYGB, AGRN, PFKM, LDHA	0.82	0.92	0.71
	PYGL, PYGB, EXT1, PFKM, LDHA	0.82	0.83	0.71
	PYGL, AGRN, EXT1, PFKM, LDHA	0.78	0.83	0.67
	PYGB, AGRN, EXT1, PFKM, LDHA	0.81	0.83	0.71
	PYGL, PYGB, AGRN, EXT1, PFKM, LDHA	0.80	0.83	0.71

Supplementary Table 3 – Correlation between PCA-based divergences, Related to Figure 3. Correlation between PCA-based phenotypic divergences calculated using different thresholds on the minimal percentage of total variance required for inclusion.

PCA threshold 1	PCA threshold 2	rho	p
0	1	0.97	<0.001
0	2	0.87	<0.001
0	3	0.81	<0.001
0	5	0.81	<0.001
1	2	0.91	<0.001
1	3	0.84	<0.001
1	5	0.84	<0.001
2	3	0.91	<0.001
2	5	0.91	<0.001
3	5	1.00	<0.001

Supplementary Table 4 – Correlation between cluster-based divergences, Related to Figure 3. Correlation between cluster-based phenotypic divergences calculated using different numbers of clusters.

n clusters 1	n clusters 2	rho	p
6	7	0.9	<0.001
6	8	0.85	<0.001
6	9	0.83	<0.001
6	10	0.83	<0.001
7	8	0.94	<0.001
7	9	0.92	<0.001
7	10	0.92	<0.001
8	9	0.95	<0.001
8	10	0.95	<0.001
9	10	1	<0.001

Supplementary Table 5 – Correlations between PCA-based and cluster based divergences, Related to Figure 3. Correlation between phenotypic divergences calculated using different thresholds on the minimal percentage of total variance required for inclusion (PCA-based) and different number of clusters (cluster-based).

PCA threshold	n clusters	rho	p
0	6	0.72	<0.001
0	7	0.75	<0.001
0	8	0.78	<0.001
0	9	0.79	<0.001
0	10	0.79	<0.001
1	6	0.75	<0.001
1	7	0.76	<0.001
1	8	0.78	<0.001
1	9	0.79	<0.001
1	10	0.79	<0.001
2	6	0.79	<0.001
2	7	0.76	<0.001
2	8	0.78	<0.001
2	9	0.78	<0.001
2	10	0.78	<0.001
3	6	0.82	<0.001
3	7	0.79	<0.001
3	8	0.81	<0.001
3	9	0.80	<0.001
3	10	0.80	<0.001
5	6	0.82	<0.001
5	7	0.79	<0.001
5	8	0.81	<0.001
5	9	0.80	<0.001
5	10	0.80	<0.001

Supplementary Table 6 – TCGA samples, Related to Figure 1. Number of samples per TCGA set that were used for Hallmark activity gene reduction validation.

Set	Samples
ACC	78
BLCA	408
BRCA	1093
CESC	304
CHOL	36
COADREAD	624
DLBC	48
ESCA	184
GBMLGG	681
HNSC	521
KICH	66
KIRC	533
KIRP	290
LAML	173
LIHC	371
LUAD	516
LUSC	501
MESO	87
OV	304
PAAD	178
PCPG	179
PRAD	497
SARC	259
SKCM	469
STAD	418
STES	602
TGCT	150
THCA	501
THYM	120
UCEC	557
UCS	57
UVM	80

Supplementary Table 7 – Genes and primers for each activity, Related to Figure 1.

Activity	Gene	Forward Primer	Reverse Primer
EMT	CDH1	cccgggacaacgtttattac	gctggctcaagtcaaagtcc
EMT	COL4A1	ggcatgcctggtattggt	aggccccatatcacccttag
EMT	COL5A1	gcccggatgtcgcttacag	aaatgcagacgcagggtacag
EMT	CTGF	acattagtacacagcaccagaatgt	gctatctgatgatactaacctttctgc
EMT	FBN1	gcggaaatcagtgtattgtccc	cagtgttgtatggatctggagc
EMT	FN1	gacgcatcacttgcacttct	gcaggtttcctcgattatcct
EMT	FSTL1	gccatcaatattacaacgtatcca	tcaatgagagcatcaacacaga
EMT	HTRA1	tgatctcaggagcgtatataattga	tgacgtcgttttccttgaga
EMT	ITGAV	catgtcctccttatacaattttactgg	gcagctacagaaaatccgaaa
EMT	NNMT	ggcttcacctccaaggacacc	cccttcacaccgtctaggca
EMT	SNAI2	tggttgcttcaaggacacat	gttgcagtgagggcaagaa
EMT	ZEB1	aactgctgggaggatgacac	tcctgcttcatctgcctga
EMT	ZEB2	aagccagggacagatcagc	gccacactctgtgcatttga
DNA repair	CLP1	cccccactttgtacgcact	gatacgctcatccctacattcc
DNA repair	DDB2	gagacaacgtggggaacg	tgcattctgagattccaaagc
DNA repair	DUT	ccttctgggtgttatggaagagt	gctgtgcaattcgatcaccttt
DNA repair	FEN1	agaagggagagcgagcttag	gggccacatcagcaattagt
DNA repair	LIG1	cttcctgctggcctcctac	cactgaagccagttccaagc
DNA repair	PMS1	gaatgtagacctcgcaaagtgat	atgggtaattgtctggatagacg
DNA repair	POLD1	gcctacatgaagtcggagga	tccaggtagtactgcgtgtcaa
DNA repair	POLQ	gattgagccagagtctgttgg	tccataaatgatcccatagcaa
DNA repair	POLR2E	agctagtccctgagcacgtc	gctggttctctcggagctta
DNA repair	POLR2F	tgtcagacaacgaggacaattt	tccaagtcatctagcccttca
DNA repair	RFC5	gcgtagggctctgaacattt	tggcaatgtctgacttgagc
DNA repair	SF3A3	tggtcgttatctcgatctccat	agaggcttcactctatctgtgt
DNA repair	XPC	ggaacgagtttgggaatgtg	gtgtagattgggcaggttcag
Glycolysis	AGRN	cacacgtactcctgcaaggtt	cgctgatcaccaccttgtt
Glycolysis	EXT1	aggcttgggtccttcagatt	catccattgctgagcatcac
Glycolysis	HK2	ctcgccggtagccttctt	gtccgactgctttgtgctg
Glycolysis	IL13RA1	tgcacagtaatatggacatggaa	cgagtttccggagctattttc
Glycolysis	LDHA	gcagatttggcagagagtataatg	gacatcatcctttattccgtaaaga
Glycolysis	PFKM	gggtgtggaagcagtgatggc	gttcatgaagctccggcctct
Glycolysis	PYGB	gatcgtgaaacagtcggtctt	gccattggtcttattctggaac
Glycolysis	PYGL	acaccaaccacacagtgctc	aggaaacaaggccacaattc
Glycolysis	SLC16A3	catctcctagggcatggtg	aggagtttgcctcccgaag
Other	BYSL	tgagagccaacttgagatacca	tgacctgacacaatagccataga
Other	DCTPP1	cgcctccatgctgagtttg	ccaggttccccatcggttttc
Other	GNL3	cagtggcttcagttcacacg	ggtcatgcgtttacttgcttt
Other	TCOF1	ggtctccatccaaggtgaagc	tccccacagatggcacagat
Housekeeping	CIAO1	ccatgaaaatgaggtcaagtca	gctcttatctcggctgcaag
Housekeeping	CNOT4	acctatatccggtcagaagacg	ttctgccatctactaccacattg
Housekeeping	HNRNPK	gaaaatcatccctaccttggaa	tccacagcatcagattcgag
Housekeeping	RAB1A	gggaaaacaatcaagcttcaaa	ctggaggtgattgttcgaaat
Housekeeping	TIAL1	gggtggatttggtgctca	catatccggcttggttagga
Housekeeping	UBE2D3	cggacctttgagcatacacc	cgccatagtgtgtgcttgtc
Housekeeping	YTHDC1	aagccactgagctcatctgtta	cgcttgtttctttcagatctttg

Supplementary Table 8 – Dataset information, Related to Figure 2.

Dataset	**Number of patients**	**Number of cells**
Fan (multiple myeloma)	2	172
Filbin (H3K27M-glioma)	9	4,058
Li (colorectal cancer)	10	590 (of which 215 normal)
Neftel (glioblastoma)	28	7,930
Patel (glioblastoma)	5	430
Tirosh (melanoma)	15	4,347
Tirosh (oligodendroglioma)	6	4,645
Venteicher (astrocytoma)	10	6,341
Total	*85*	*28,513*