nature research

https://doi.org/10.1038/s43018-020-00159-4

Supplementary information

Pathway-based classification of glioblastoma uncovers a mitochondrial subtype with therapeutic vulnerabilities

In the format provided by the authors and unedited

Supplementary Note.

Benchmarking and computational framework of the manuscript “Pathway-based classification of glioblastoma uncovers a mitochondrial subtype with therapeutic vulnerabilities”.

1. scRNASeq data processing and quality controls

Processing and quality control filtering of the three scRNAseq datasets have been described[1-3] with some modifications. In particular, for dataset 2, to retain only high-quality cells and avoid dead cells and library construction artifacts, we filtered out cells that had less than 1,000 detected genes, resulting in a median number of detected genes of 1,392 per cell (including both tumor and non-tumor cells). We also removed those tumors whose number of cells after this high-quality cell filter was lower than 500, or the percentage of cells passing QC lower than 25% (PJ017, PJ032, and PJ035). PJ016 was removed from the analysis as this tumor harbored an IDH1 mutation. Cell doublets that result from the simultaneous capture of multiple cells in a single well or droplet were identified (and removed) by the following steps: 1) inspection of the distribution of expressed genes per cell (complexity) across each dataset, that indicates the absence of cells with highly divergent number of expressed genes (potential doublets) in comparison to the complexity of the respective dataset (Fig. SN1a); 2) application of DoubletDecon[4], a deconvolution-based strategy to remove doublets by assessing the underlying contribution of concurrent gene expression programs within a single-cell library. To provide a metrics for the MAGIC imputation of missing values, we report the percentage of zero expression genes per cell in the three datasets before application of MAGIC imputation (Fig. SN1b). Concerning quality control parameters based on mitochondrial read counts, as the 5% cutoff has been uniformly adopted for unsorted scRNAseq

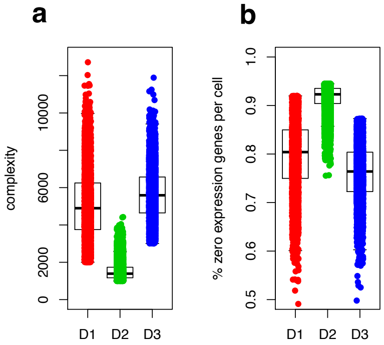

Fig. SN1. a, Boxplot of gene complexity by cell in each dataset. **b,** Boxplot of the percentage of genes with zero expression per cell in each dataset (dataset 1, *n* = 4,227; dataset 2, *n* = 10,315; dataset 3, *n* = 5,742). Boxplots span the first to third quartiles and whiskers show the 1.5× interquartile range.

data[5-9], we controlled that this threshold was independently met by each of the three processing methods for the different datasets, without imposing additional filters.

2. Single cell RNA-sequencing analysis and bulk GBM multi-omics

Definition of single-sample pathway activity

We aggregated the MSigDB c5.bp, c5.mf, c5.cc, Hallmark and KEGG collections of gene sets, retaining only pathways composed of at least 15 genes, resulting in 5,032 gene sets. Pathway enrichment in each individual cell was computed by adapting the Mann–Whitney–Wilcoxon Gene Set test (MWW-GST) originally developed for the analysis of unbalanced datasets. When used in comparative analysis, MWW-GST requires as input a gene set and a ranked list representing the gene-wise differential expression between the two groups. When adapted to single cell analysis (single sample MWW-GST, ssMWW-GST), to determine the relative expression of individual cells in each tumor, the expression of each gene is standardized for the expression in the cell cohort and used to generate a cell-specific ranked list. Ranked lists of single cells and pathway gene sets are used as input for ssMWW-GST. The resulting normalized enrichment score, NES, is an estimate of the probability that the expression of genes in the gene set is greater than the expression of genes outside the gene set:

$$NES = 1 - \frac{U}{mn}$$

where *m* is the number of genes in a gene set, *n* is the number of those outside the gene set, $U = nm + \frac{m(m+1)}{2} - T$, and *T* is the sum of the ranks of genes in the gene set. Thus, NES is a reporter of pathway activity with values near zero meaning down-regulation of the pathway and values near one indicating up-regulation of the pathway. In addition to the NES, MWW-GST generates a *p*-value for each pathway activity, a parameter considered for the selection of enriched pathways (see *Step 3-i (Binarization)* of *scBiPaD* method). MWW-GST function is available in the yaGST package (https://github.com/miccec/yaGST)[10].

Single cell Biological Pathway Deconvolution (scBiPaD)

We developed a computational approach designed as *s*ingle *c*ell *B*iological *P*athway *D*econvolution (*scBiPaD*) to identify coherent functional states in single cells across multiple tumors. Cancer phenotypes classification methods based on gene-level genome-wide expression profiles fail to capture the relationships and interactions between system components of the different cellular states within a single tumor. *scBiPaD* acquires pathway-based aggregation of gene information and incorporates gene-gene relationships. *scBiPaD* includes the three following steps (Extended Data Fig. 1): Step 1) identification of cell sub-populations in each individual tumor that share activation of similar biological functions; Step 2) determination of enriched biological pathways in each cell sub-population by defining cluster-specific ranked-lists; Step 3) identification of cell sub-populations that share coherent biological functions across multiple tumors.

- *Step 1-i (Standardization and Ranking):* for each cell, we ranked genes after standardization for the expression of each gene across cells composing each tumor.
- *Step 1-ii (Pathway Activity):* we calculated the activity of all the 5,032 biological pathways (NES) for each single-cell with MWW-GST using the ranked list of the individual cell. Thus, each cell was represented by a vector of 5,032 values of NES that were used to derive the tumor sample-specific activity matrix.
- *Step 1-iii (Euclidean Distance):* we used the activity matrix to generate the Euclidean distance matrix between every pair of cells in each tumor.
- *Step 1-iv (Consensus Clustering):* the Euclidean distance matrix was then used to inform a consensus clustering between cells of each tumor (10,000 random samplings using 70% of the cells and the Ward linkage method). For each tumor, the optimal number of clusters was determined using the Calinski and Harabasz criterion[11]. Only cells having a silhouette score > 0.5 and clusters composed of at least 10 cells were retained for further analysis. The application of this approach to each of the 36 tumors from three single cell datasets revealed 94 sub-populations, with a number of sub-populations in each tumor ranging from 2 to 5, and 91% of

cells retained after the filtering step. All retained clusters were further inspected in order to elucidate their biological significance.

- *Step 2-i (Gene Scoring):* to define the biological pathways enriched in sub-populations of individual tumors composing distinct clusters, we first derived a cluster-specific ranked-list of genes comparing the expression profiles of the cells in the cluster with all other cells in the same tumor using the Mann-Whitney-Wilcoxon test, defining a score for each gene *j* as

$$MWW - score = \frac{U_j}{nm}$$

where U_j is the MWW test statistic for the *j*-th gene, *n* is the number of cells in the cluster, and *m* is the number of cells in the other clusters.

- *Step 2-ii (Pathway Activity):* the cluster-specific ranked lists were used to identify pathways activated in each cell sub-population using MWW-GST as in *Step 1-ii*.
- *Step 3-i (Binarization):* to identify biologically coherent cell sub-populations across multiple tumors from the combined datasets, we represented each cell subpopulation with a binary vector of length 5,032, with 1 indicating the enriched biological pathways [logit(NES) > 0.58 and FDR < 0.01].
- *Step 3-ii (Jaccard Distance):* the degree of overlap of enrichment between sub-populations was then computed by using the Jaccard coefficient of similarity (index) defined as:

$$J(p_{it'}, p_{jt''}) = \frac{|p_{it'} \cap p_{jt''}|}{|p_{it'} \cup p_{jt''}|},$$

where $p_{it'}$ and $p_{jt''}$ are the enriched biological pathways of sub-population *i* of tumor *t'* and sub-population *j* in tumor *t"*. The Jaccard index is a measure of similarity between two sets, with 0 indicating no overlap and 1 indicating complete overlap. Then, we derived the Jaccard distance, defined as 1-(Jaccard index).

- *Step 3-iii (Consensus Clustering):* the Jaccard distance was used to cluster cell sub-populations using consensus clustering (Ward linkage method, 10,000 resampling steps with 70% of sub-populations). Calinski and Harabasz criterion was used to derive K=4 as the optimal number of clusters.

Characterization of the biological states of the single cell subpopulations

To characterize the four clusters of cell sub-populations, we obtained the medoids of each cluster by applying the Partitioning Around Medoids (PAM) clustering algorithm[12]. A medoid is defined as an object that minimizes the sum distance of this object to the other objects within its cluster, thus reflecting all objects in the cluster. In our datasets, the medoid is a binary vector having a value of 1 for the enriched pathways in the cell sub-population. The pathways were then used to construct Gene Ontology-guided maps using the Enrichment Map application[13] of Cytoscape (version 3.7.2)[14] after removing gene sets that included more than 250 genes (650 gene pathways), thus avoiding the preferential selection of very large uninformative pathways and applying the enrichment set cover algorithm[15]. To extract the biological functions that specifically drive each cluster, we selected pathways with significant differential activity in a specific cluster compared to the others using two-sided MWW test (effect size > 0.3 and FDR < 0.0001). Cluster assignments by *scBiPaD* were further confirmed by applying the entire workflow independently to each of the three single cell datasets. This analysis produced 4 clusters for each dataset having very similar enrichments as the combined analysis (see section 4). When the sub-populations from the individual datasets were combined, we obtained ~95% concordance of cell sub-population identity with the analysis performed on the aggregated datasets.

Definition of single cell gene meta-signature, subtype classification and state coexistence

For each cell sub-population cluster, we defined a meta-signature based on the average gene MWW-scores across cell sub-populations of the same cluster using the three single cells datasets combined. Each meta-signature consisted of the 50 highest scoring genes. Glioma cells were then assigned to each individual subtype on the basis of the highest significant score using ssMWW-GST [logit(NES) > 0 and FDR < 0.01]. ssMWW-GST was also used to classify cells according to lineage states[10,16-18] and the correlation between pathway-based functional states and lineages states was examined by χ^2 test. Analysis of the co-existence of cell states within individual tumors was performed using Spearman's correlation and classical multidimensional scaling (CMDS) of the distribution of GBM states per tumor with *k*-NN ($k = 2$) tumor clustering.

Identification of glioma subtypes in the TCGA IDH wild type glioblastoma cohort

We used the GBM dataset from The Cancer Genome Atlas (TCGA) collection profiled with Agilent chip G4502A. The matrix of the raw data was quantile normalized. The gene expression data matrix includes 534 samples and 17,814 genes. To determine the clinical relevance of the biological subtypes, we identified those pathways capable of segregating patients according to survival. Survival data are available for 527 IDH wild type GBMs from TCGA and were downloaded using TCGAbiolinks R/Bioconductor package[19]. Pathway enrichment in each individual tumor was computed by ssMWW-GST. For each pathway, we defined 3 groups of patients: (i) high activity group: patients whose tumor had activation of the pathway [logit(NES) > 0 and FDR < 0.01]; (ii) low-activity group: patients whose tumor exhibited inactivation of the pathway [logit(NES) < 0 and FDR < 0.01]; (iii) neutral activity group: patients whose tumor lacked activation or inactivation of the pathway (FDR ≥ 0.01). Survival was evaluated by the log-rank test: (i) high versus low activity group; (ii) high versus low versus neutral activity group. The positive or negative correlation with outcome was established using the Cox's proportional hazards regression model of hazard ratio (HR). For further analysis, we selected those pathways that resulted in a significant survival difference in any comparison ($p < 0.05$), totalizing 192 pathways. Given the *i*-th and *j*-th tumors and a pathway *p* from the 192 survival-associated, we defined $g_{ij}(p) = 1$ if both tumors belonged to the same pathway activity group, 0 otherwise. Finally, we constructed a dissimilarity matrix (C) as

$$C(i,j) = 1 - \frac{1}{192}\sum_{p=1}^{192} g_{ij}(p).$$

The distance induced by matrix C was then used to cluster samples using consensus clustering (Ward linkage method, 10,000 resampling steps with 70% of samples). Calinski and Harabasz criterion was then used to derive K=5 as the best number of clusters.

Characterization of GBM IDH wild type tumors in functional subtypes

We identified differential activated pathways among the 192 pathways associated with survival using two-sided MWW test (effect size > 0.3 and FDR < 0.01). This analysis produced 126 pathways. We then examined the pathways, among the entire collection of 5,032, that were differential activated between the four GBM sub-groups and identified 2,792 pathways. Finally, we used the Kruskal-Wallis H test to select genes that showed any difference in expression levels between the four subtypes (FDR < 0.01) and the post hoc Nemenyi's test for multiple comparison correction[20] to identify genes with significant differential expression in one subtype compared with the others [$\log_2$(FC) > 0.5 and FDR < 0.01].

Definition of bulk GBM gene signatures and simplicity score

We used the MWW test to derive ranked lists of genes differentially expressed in each of the subtypes compared to the others. For each subtype the final gene signature included the first 50 highest scoring genes in the ranked list. These gene signatures were used to calculate the enrichment of each functional GBM subtype (normalized enrichment score, NES) for each bulk tumor. The simplicity score for each individual tumor was then computed as the difference between the highest NES (dominant subtype) and the mean of the other subtypes (non-dominant). The simplicity score represents the subtype activation: higher scores indicate lower transcriptional complexity and lower scores multi-subtype activation.

Supervised classification of unclassified TCGA GBM

To assign state subtype memberships to the 230 tumors from the *unclassified* (black) cluster we used the 304 GBM initially classified as the training set of a *k*-NN classifier (k = 3). The classifier feature set included the expression of the 100 highest scoring genes in the ranked list of each subtype. Twenty-eight tumors with conditional probability to subtype memberships < 0.6 remained unclassified and were excluded from subsequent analyses. The samples classified by *k*-NN were integrated with 304 samples obtained from consensus clustering and used in the analysis of genetic alterations associated with GBM subgroups.

Analysis of state transition between tumor core and periphery and primary and recurrent GBM

To identify state transitions from tumor core and rim we used dataset 1, which included regional samplings, and applied STREAM (Single cell Trajectories Reconstruction, Exploration And Mapping), a recently developed tool that can reconstruct complex trajectories along with accurate pseudotime estimation from single cell transcriptomic data[21]. STREAM first identified variable genes and then, using these features, projected cells to a lower dimensional space constructing an Elastic Principal Graph (ElPiGraph algorithm). STREAM (version 0.3.8) analysis was performed using the MLLE dimensionality reduction method (*n* = 2 components) with the most variably expressed genes as features. To characterize functional state transitions from primary to recurrent tumors, we used a dataset of 61 matched primary and recurrent IDH wild type GBM[22]. Data are fpkm normalized and the transition of states from primary to recurrent tumors was assessed by the χ^2 test.

Copy Number data and processing

Copy number profiles were generated by Affymetrix SNP 6.0 array, and were available for 487 out of 506 TCGA-GBM IDH wild type tumors. Genomic regions significantly gained or lost have been inferred by GISTIC version 2.0.22[23].

Functional copy number variations (fCNV) analysis

The functional CNV (*f*CNV) genes in each sample were determined by the analysis of association of copy number with gene expression. Thresholds for CNV calls were assessed using GISTIC scores (-2 homozygous deletion, -1 heterozygous deletion, 0 no change, 1 gain, 2 high level amplification). Given an alteration (amplification or deletion) of a gene *g* in a specific sample *s* we computed the mean, μ_g, and variance, σ_g, of the expression of that gene in the set of samples without the alteration. The alteration was considered functional in sample *s* if the expression e_{gs} of the gene was such that the 95% confidence interval of $\frac{e_{gs}-\mu_g}{\sigma_g}$ was above 1 (for amplifications) or below -1 (for deletions).

Association of glioma states with copy number variations (CNVs), single nucleotide variations (SNVs) and the integrated CNVs and SNVs

We categorized as deletions CNVs with GISTIC scores of -2 and -1 and amplifications CNVs with GISTIC scores of 1 and 2. The association of CNVs with distinct GBM subtypes was determined by comparing the number of samples harboring a specific CNV in each subtype with the other groups using the two-tailed Fisher's exact test (for one subtype compared to all others combined) and a χ^2 independence test. Only genes with *f*CNV calls in at least three samples in the dataset were examined (14,948 deleted and 14,687 amplified genes). We selected CNVs that scored as statistically significant according to both tests ($p < 0.05$). These genes were used in the enrichment analysis. Genes significantly associated with multiple subtypes were excluded.

To prioritize candidate driver genes in broad CNVs, we applied the Comfocal algorithm to the Affymetrix SNP 6.0 array data from TCGA GBM[24]. Genes were ordered by their respective combined deletion score. To identify mutually exclusive homozygous deletions with the highest positive association with the MTC activity (NES), we used the UNCOVER algorithm, which identifies sets of alterations displaying complementary functional association with a quantitative profile[25]. We used the greedy algorithm implementation in Python and assessed the significance of the solution with a permutation test from 1,000 iterations. Features present in less than 3 samples were excluded from the analysis. UNCOVER also computed a weight (score) for each gene that reflects the contribution of each element to the highest significant gene module identified as the optimal solution.

For the association of cell states with SNVs, mutation data were available for 320 out of 506 GBMs and were downloaded from GDC data portal. The functional effect of missense SNVs and in-frame indels was determined using multiple prediction algorithms as we recently described[17]. Tumors with a total number of mutated genes > 300 were defined as hypermutated and excluded from the analysis. Only genes with pathogenic mutations in at least three samples in the dataset were considered in the analysis (1,270 genes). Gene mutations that were significantly associated with GBM subtypes by the Fisher's exact or χ^2 test were considered ($p < 0.05$).

To identify the genes targeted by somatic mutations and CNVs that are associated with distinct GBM subtypes, we integrated both genetic alteration data. We used 311 samples from TCGA with available mutation and copy number data. Genes altered in at least 3 samples in the dataset were considered in the analysis (6,262 mutated and/or deleted and 6,390 mutated and/or amplified genes). We determined the strength of association according to the Fisher's exact test and χ^2 test with $p < 0.05$. Then, the contribution of both CNV and SNV to the significance of the association with a specific subgroup was determined by combining Fisher's exact test p-values for CNV and SNV according to the Edgington's method[26] ($p < 0.05$).

Association between deletions of chromosomal regions and GBM subclasses was determined using Fisher's exact test and χ^2 test ($p < 0.10$). A chromosomal region was defined as deleted when at least one gene harbored functional homozygous deletion.

Integrative analysis of methylation and gene expression in primary GBM

DNA methylation data (profiled by Illumina Infinium Human Methylation 450 K platform) were downloaded using TCGAbiolinks package available on R Bioconductor. Data was normalized using functional normalization available in Minfi[27]. After removing probes targeting X and Y chromosomes and probes not associated with gene promoters[28], the final methylation data matrix included 59 GBM (15 GPM, 12 MTC, 16 NEU and 16 PPR) and 66,079 probes. Differential methylation between each GBM subtype compared with the others was performed using two-sided MWW test (absolute methylation $\log_2$ fold-change > 0.58 and $p < 0.01$). Functional methylation (DNA methylation affecting gene expression) was determined by integrating DNA methylation and gene expression and comparing each functional GBM subtype with the others (two-sided MWW test, $p < 0.05$, absolute expression $\log_2$(fold-change) > 0.4, and absolute methylation $\log_2$(fold-change) > 0.3). Methylation probes for each gene were selected according to the distance from the transcription start site (closest probes to –100-bp position, 58,059 probes for 10,337 unique genes). For data visualization in the starburst plots, if mean DNA methylation β-value or mean gene expression value was lower in one subtype compared with the other subtypes ($\log_2$(fold-change) < 0), we multiplied the $-\log_{10}$ (p-value) by -1.

Integrative analysis of miRNA and gene expression in primary GBM miRNA gene quantification data (Agilent array) was downloaded using TCGAbiolinks package available on R Bioconductor. The final expression matrix includes 534 miRNAs and 294 GBM tumors (66 GPM, 104 MTC, 52 NEU and 72 PPR). Data was quantile normalized at the probe level, followed by $\log_2$ transformation. miRNAs differentially expressed between each of the GBM subtype compared with the others were obtained using two-sided MWW test [$\log_2(FC) > 0$ and $p < 0.0005$]. To identify genes functionally regulated by specific miRNA, mRNA-miRNA gene expression data was integrated. The list of experimentally validated miRNA targets was downloaded from MiRwalk 2.0 database[29]. MiRNA targets whose expression was anti-correlated with miRNA expression in each GBM were considered (Spearman correlation, $\rho < 0$ and $p < 0.05$).

Proteomics characterization on GBM subclasses

Reverse phase protein array (RPPA) quantification data was downloaded using TCGAbiolinks package available on R Bioconductor. Data was quantile normalized on the probe level, followed by $\log_2$ transformation. The final expression matrix includes 196 protein and/or phospho-proteins from 103 GBM tumors (27 GPM, 25 MTC, 25 NEU and 26 PPR). Differential RPPA analysis was performed by comparing each subtype with the others [two-sided MWW test, $\log_2(FC) > 0.25$ and $p < 0.01$].

Analysis of genetic alterations and tumor evolution of SLC45A1-deleted GBM

A genomic dataset (WES or WGS) of 725 IDH wild type GBM was assembled by integrating TCGA (n = 581) and GLASS (n = 144) cohorts[30,31]. Somatic mutations and CNVs were retrieved from the respective studies; CNVs were further processed to infer *f*CNVs as described above. Fisher's exact test was performed to compare the frequency of alterations in GBM driver genes between *SLC45A1*-deleted ($n = 20$) and wild type ($n = 705$) GBM samples. The cancer cell fraction (CCF) of each occurring genetic alteration was computed using variant allele frequency and corresponding copy number by ABSOLUTE[32]. The CCF of mutations and *f*CNVs was analyzed to impute the genetic evolution of 8 matched pairs of primary-recurrent *SLC45A1*-deleted samples for which genomic data were available. PhyC[33] was used to construct tumor evolutionary trees that reflected the accumulation patterns of private genetic alterations and genetic alterations shared between primary and recurrent tumors. The phylogenetic structures from the 8 longitudinal GBM pairs were integrated in a three-dimensional plot representing the evolution of GBM driver genes by the relative frequencies of private alterations, exclusively occurring in primary or recurrent tumors, and truncal alterations, shared by the two tumors.

Cross-classification analysis

To assess the relationship between the GBM subtypes described in this manuscript and the two previously published glioma subgroup[34,35], we used the 304 GBM for which an unequivocal subtype assignment had been obtained (Fig. 3a) and the survival analysis had been performed (Fig. 3c). GBM subtype assignment was obtained by applying ssGSEA as described in the original description[35]. The relationship between pathway-based and gene marker-based subclasses was examined by χ^2 test. Survival for 302 out of 304 patients for which clinical data were available was determined by Kaplan Meyer curve and the log-rank test. Samples in the validation datasets were similarly typed for survival analysis.

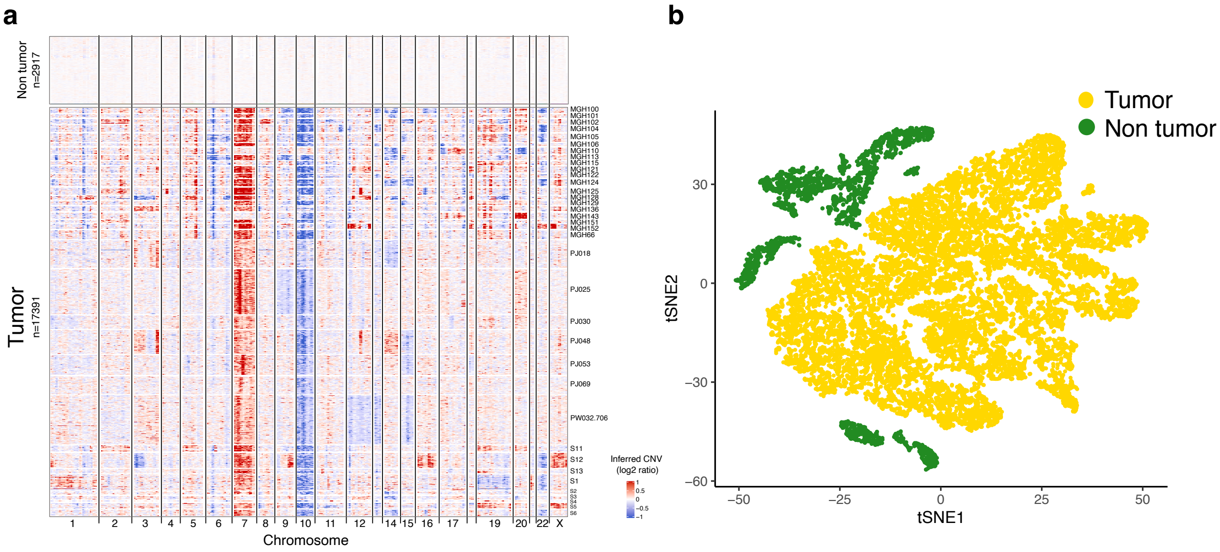

Figure SN2. a, Single cell large-scale CNV profiles predicted on the basis of the average expression in 100 gene sliding windows. Rows corresponds to individual cells with non-tumor and tumor cells grouped at the top and bottom, respectively. Tumor cells were ordered by tumor ID and showed diffuse aneuploidy compared to the non-tumor cells. **b,** Single cell gene expression *t*-SNE plot of individual cells aggregated from the three datasets under investigation; cells clustered in clearly distinct groups according to tumor and non-tumor type.

3. Single cell classification in tumor and non-tumor cells

The intra-tumor heterogeneity is characterized by multiple levels of molecular diversity including the first main layer that separates malignant cells from the non-tumor infiltrates. The preliminary analysis on the three scRNAseq datasets was the molecular annotation to resolve tumor from non-tumor cells. Towards this aim, we performed a multi-step classification that integrates CNV-based and gene expression-based approaches to distinguish malignant from non-malignant cells.

The inference of broad chromosomal aberrations from gene expression data is widely adopted in the analysis of single cell RNAseq samples and it is particularly useful in discriminating malignant from non-malignant cells in tumors with a high rate of chromosomal instability[36-41]. IDH wild type GBM is characterized by large degree of aneuploidy, whereby multiple chromosomal regions are frequently gained or lost. Among them, gain of chromosome 7 and loss of chromosome 10 are the two most frequent chromosomal-level changes[42,43]. Considering the common genetic hallmarks of GBM, we inspected the inferred copy number profiles of single cells to identify aneuploid tumor cells. Large-scale CNV were predicted within individual cells from their gene expression by deriving a moving average in sliding windows of 100 genes along each chromosome[37,40]. As shown in Fig. SN2a, all IDH wild type GBM samples we analyzed harbor chromosomal copy number alterations (mainly but not exclusively chromosome 7 amplification and chromosome 10 deletion) allowing to accurately discriminate tumor cells from the non-malignant component.

The few tumors cells that may not harbor gain of chromosome 7 and/or loss of chromosome 10 are identified by other, less recurrent but clearly identifiable chromosomal aberrations.

From the analysis of scRNAseq data, individual cell clustering based on gene expression profiles separates the malignant component from non-tumor cells independently of sample of origin[36,39,41]. Therefore, we corroborated the distinction of malignant from non-malignant cells that emerged from InferCNV using gene expression profile clustering with non-linear dimensionality reduction (*t*-Distributed Stochastic Neighbor Embedding, *t*-SNE). In this analysis, tumor cells classified by InferCNV were distinct from non-tumor sub-populations that instead grouped in spatially separated clusters (Fig.

SN2b). The InferCNV annotation was combined with the *t*SNE-based clustering to establish concordant assignment of the cell types considered in subsequent analyses.

4. Determination of patient-specific batch effects from different scRNAseq datasets and platforms

The use of gene signatures-based approaches from scRNAseq data generated from human tumors has been ridden with major limitations, among which the two most prominent are (*i*) the patient-based clustering of tumor cells that masks the identity of biologically meaningful clusters across patients[44]; and (*ii*) the inability to compare results generated from different platforms (i.e. droplet-based scRNAseq approaches such as 10X genomics vs full-length Smartseq2 sequencing). The *s*ingle *c*ell *Bi*ological *Pa*thway *D*econvolution (*scBiPaD*) approach was generated to identify coherent functional states in scRNAseq data across multiple tumors, regardless of the particular platform used for single cell RNA sequencing. *scBiPaD* uses pathway activity as identifier of individual cell types. Pathway-based classifications of transcriptomic cancer data have shown higher stability of functional relationships of biological activities and exhibited better performance than gene-based classifiers[45].

As described in detail in the section 2 of the Supplementary Note, step 1 of *scBiPaD* (identification of sub-populations of cells in each individual tumor that share activation of similar biological functions) and step 2 (determination of enriched biological pathways in each sub-population by defining cluster-specific ranked-lists) are completely batch-independent as they are applied to each individual tumor of each dataset. In the third and final step of *scBiPaD* (identification of cell sub-populations that share coherent biological functions across multiple tumors), we combine cell sub-populations from each of the tumors analyzed. To avoid potential contaminations from batch effects and introduce rigorous quality control measures of pathway activities, rather than clustering directly the activity values of cell sub-populations from the different datasets, we represented each sub-population with a binary vector of length 5,032, with 1 indicating the enriched biological pathway. The final cluster assignment was derived from the

Jaccard coefficient of similarity evaluating the degree of enrichment overlap between cell populations. Thus, the resolution of tumor and dataset-specific batch effects is intrinsic to the criteria that have inspired the design of the *scBiPaD* computational framework.

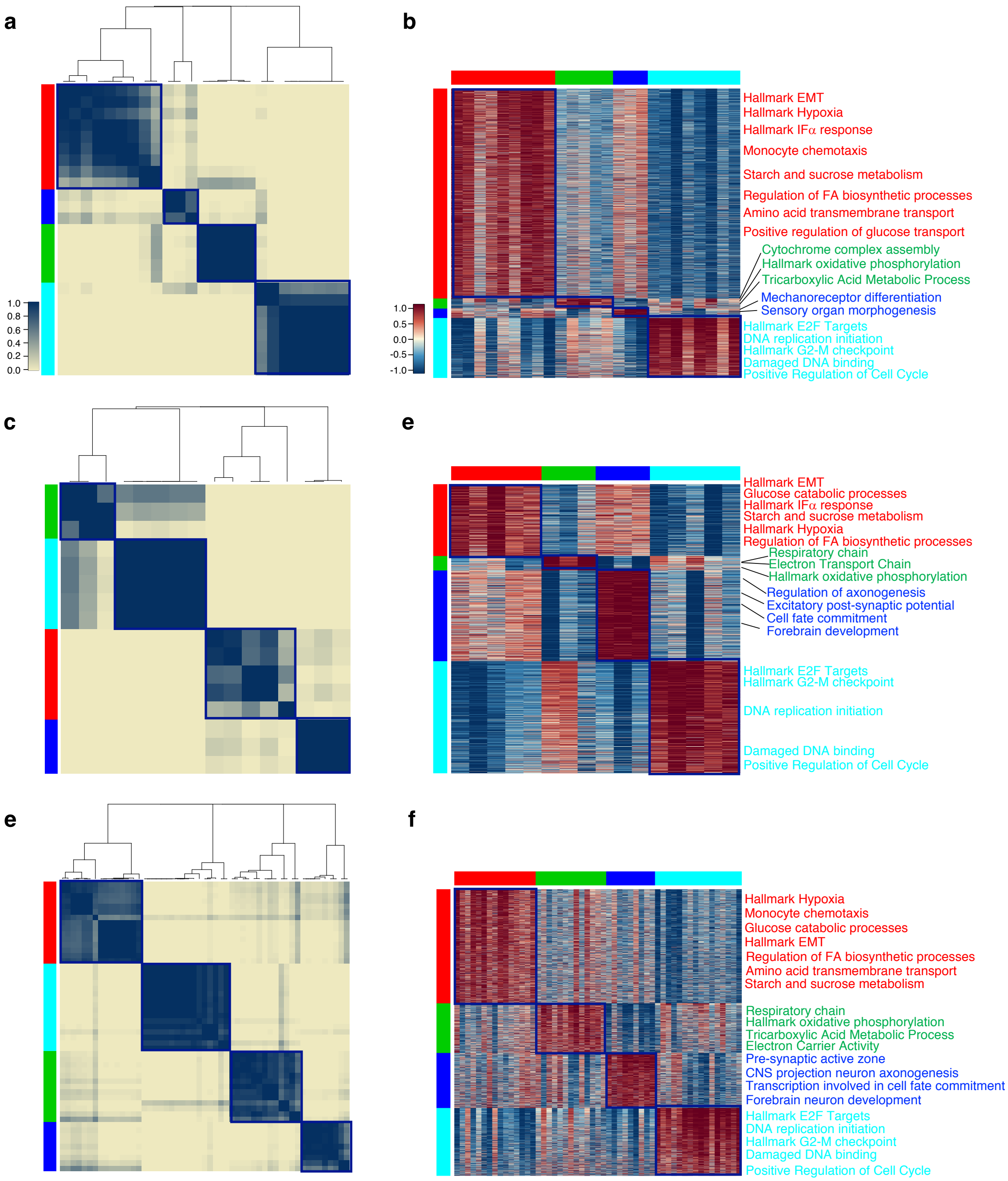

Figure SN3. Consensus clustering and heatmap of biological activities of cell sub-populations from dataset 1 (**a, b**), dataset 2 (**c, d**) and dataset 3 (**e, f**). Columns and rows of the consensus clustering and columns of the heatmaps are cell sub-populations, while rows of the heatmaps are biological activities. Colors in the left and upper tracks define the four clusters. Representative pathways specifically activated in each group are indicated according to the effect size. Left and upper tracks: red, GPM; green, MTC; blue, NEU; cyan, PPR.

To provide additional evidence that the analysis/clustering across multiple datasets was not affected by technical factors introduced during single cell processing, we also applied *scBiPaD* to each individual dataset. The three independent analyses identified four clusters defined by the same biological core activities that had been obtained by analyzing the combined datasets shown in Fig. 1 (Fig. SN3a-f). Consistent with this finding, the concordance of sub-population identity from the examination of individual and aggregated datasets was ~95%.

Here, we report the results of the analysis we performed to compare the output of *scBiPaD* with state-of-the-art gene expression single cell clustering methods. Fig. SN4 shows the comparison of the *t*-SNE plots of the tumor cell sub-populations that emerged from the pathway enrichment vectors of *scBiPaD* and used to generate the clustering in Fig. 1a of the manuscript (Fig. SN4a-c) with the *t*-SNE plots of gene expression clusters from the same scRNAseq data (Fig. SN4d- f). The results are presented to highlight either the origin of the tumor (Fig. SN4a, d), the dataset (Fig. SN4b, e), or the GBM subtype

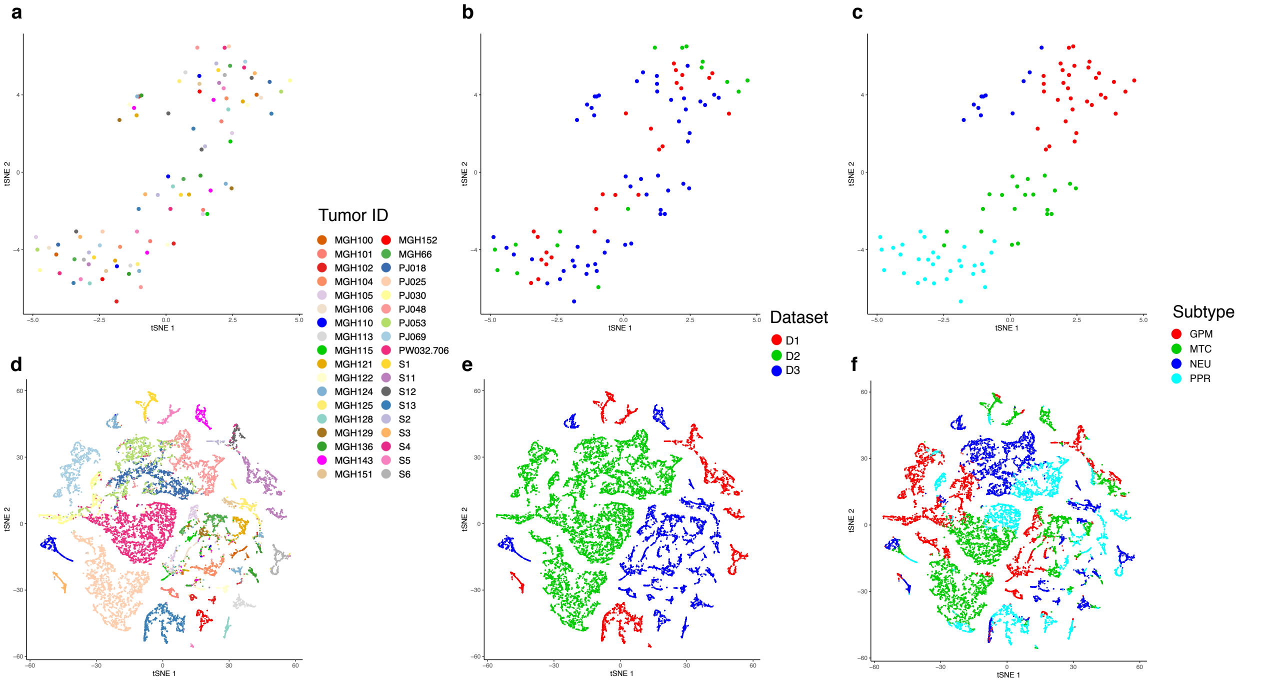

Figure SN4. a, b, c, *t*-SNE plot derived from the analysis using pathway activity (NES) for each subpopulation classified by *scBiPaD* and colored by tumor (**a**), dataset (**b**) or GBM subtype (**c**). **d, e, f,** *t*-SNE plot derived from the analysis using gene expression of each cell and colored by tumor (**d**), dataset (**e**) or GBM subtype (**f**).

(Fig. SN4c, f). Using the pathway features returned by *scBiPaD*, cell sub-populations are uniformly distributed in the *t*-SNE space without enrichment and/or bias towards specific tumors or datasets, thus indicating that *scBiPaD* outputs are not influenced by batch effects (Fig. SN4a-c).

Conversely, when gene expression is adopted for clustering, tumor origin and dataset features dominate the clustering, a finding consistent with previous knowledge (Fig. SN4d-f).

To provide additional and independent evidence for the resolution of tumor-specific batch effects by *scBiPaD*, we benchmarked the ability of *scBiPaD* to resolve batch effects in comparison to a recently reported state-of-the-art approach for real world datasets. We considered the clustering obtained by applying to the scRNAseq datasets the integration and clustering approaches recently reported and available in the Seurat v3 package[46]. In the absence of a truth table, we measured how frequently cells from each individual tumor are homogeneously located in a single cluster (indication of tumor-specific batch effect) or distributed within multiple clusters using the *entropy* of the clustering. First, we applied both *scBiPaD* and the Seurat v3 data integration pipeline and then clustered the integrated matrices considering five different number of clusters (*nCL* = 3, 4, 5, 6, 7). The clustering entropy, for each tumor sample *t*, was computed as:

$$H_t^{nCL} = -\sum_{cl=1}^{nCL} p\left(x_{t,cl}\right) \cdot \log p\left(x_{t,cl}\right),$$

where $p\left(x_{t,cl}\right)$ is the frequency of cells from tumor *t* falling in cluster *cl*. The entropy is a basic quantity in information theory, which can be interpreted as the average level of “information” or “uncertainty” inherent in the variable’s possible outcomes: higher the entropy, higher the uncertainty of the outcome of the random variable. In our case, higher entropy values indicate higher variation, that is cells of a particular tumor are randomly distributed within the various clusters and they become more varied. Entropy reaches the maximum value in case of uniform distribution at $\log(nCL)$. To compare settings with different number of clusters, we normalized the entropy to the interval of 0-1 by dividing it by its maximum value depending on the number of clusters:

$$\eta_t = \frac{H_t^{nCL}}{\log{(nCL)}}.$$

In Fig. SN5, we report the difference of η_t between the *scBiPaD* platform and Seurat ($\Delta\eta_t$) for each tumor *t* and number of clusters *nCL*.

Thus, $\Delta\eta_t > 0$ when scBiPaD outperforms the Seurat-based method; $\Delta\eta_t < 0$ in the opposite scenario. In Fig. SN5a, we ordered the tumors by the value $\Delta\eta_t$ in the case of four clusters (the final number of clusters presented in the manuscript). For all five clustering, the large majority of tumors had a positive $\Delta\eta_t$, thus indicating that our method outperforms or has a performance similar to one of the state-of-the-art approaches designed to resolve the batch effect. The same conclusion can be drawn from Fig. SN5b, where we report the same data aggregating in boxplots the $\Delta\eta_t$ of all tumors by number of clusters. The horizontal line at 0 clearly shows that most of the points are above that threshold (89% for 3 clusters, 78% for 4, 76% for 5 and 6, 73% for 7).

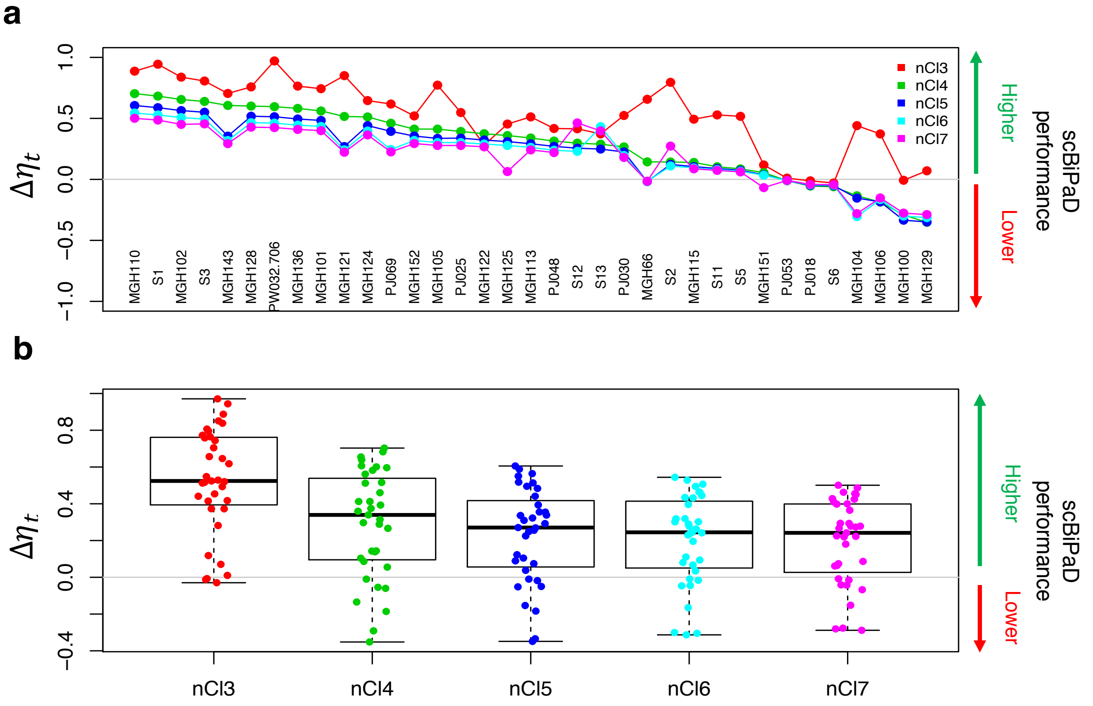

Figure SN5. Comparison of clustering entropies after scBiPaD and Seurat v3 integration pipelines. **a,** Each line represents the difference of η_t between the scBiPaD platform and Seurat ($\Delta\eta_t$) for each tumor *t* (*n* = 36) with a defined number of clusters ($nCL = 3, 4, 5, 6, 7$). Tumors are ordered by the *value* $\Delta\eta_t$ obtained for four clusters (the number of clusters presented in the manuscript). **b,** Boxplots of the $\Delta\eta_t$ for each tumor *t* (*n* = 36) and number of clusters. The horizontal gray line at $\Delta\eta_t = 0$, represents the threshold above which scBiPaD outperforms Seurat. Value below 0 indicate that Seurat v3-based method outperforms scBiPaD. Boxplots span the first to third quartiles and whiskers show the 1.5× interquartile range.

5. Impact of gene dropout from scRNAseq on pathway enrichment estimation by *scBiPaD*

The frequent high dropout rate in scRNAseq data introduces significant challenges for differential expression analysis. To begin the evaluation of the impact of gene dropout when analyzing the three scRNAseq datasets in the manuscript, we first assessed whether the frequency of missing values for genes included in each annotated biological pathway was different in each of the three datasets. Dataset 2 had the highest frequency of missing values/genes by pathway (median, dataset 1: 6%; dataset 2: 12%; dataset 3: 4%, two-sided MWW test p = 2.2e-16 in any comparison, Fig. SN6a). Accordingly, dataset 2 also exhibited the highest absolute number of missing data across all the aggregated biological pathways (Fig. SN6b).

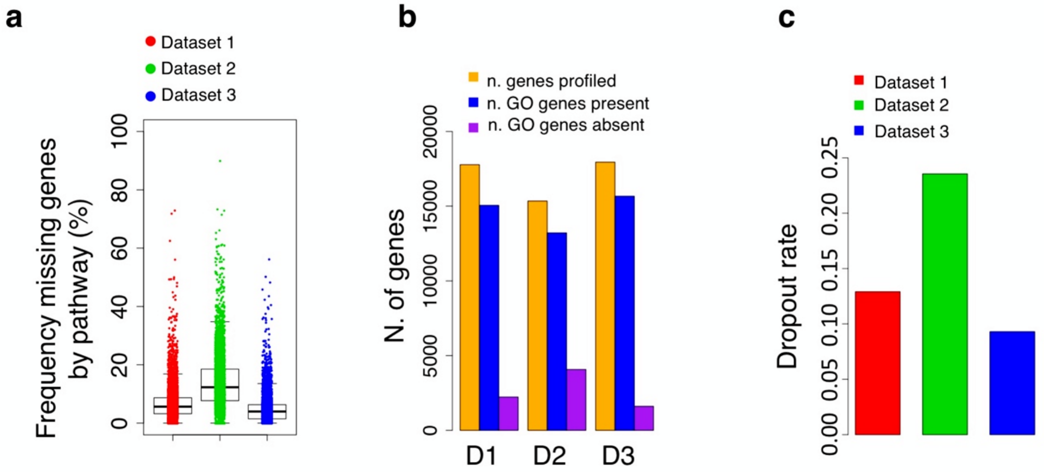

Figure SN6. a, Box plot showing the frequency (in percentage) of genes in each biological pathway that were not profiled by expression, in each single-cell dataset (n = 5,032 pathways). Statistical significance of the difference has been evaluated using two-sided MWW test (p = 2.2e-16 in any comparison). Boxplots span the first to third quartiles and whiskers show the 1.5× interquartile range. **b,** Bar plots of the number of genes profiled by expression (orange), the number of genes across all biological pathways that were included in the expression matrix (blue), total number of genes across all biological pathways that were not included in the expression matrix (purple). **c,** Bar plot of the dropout rate from each dataset across all biological pathways.

If n is the number of gene expression values in any scRNA-Seq profile, and m is the number of missing values due to sequencing factors, then the ratio $d = \frac{m}{n}$ is the percentage of zero-value genes in each cell. We defined the ratio of gene set dropout d_p by considering n as the total number of genes in the whole collection of 5,032 biological pathways, and m as the overall number of genes with missing values. In the individual datasets, d_p was 0.13 for dataset 1, 0.24 for dataset 2 and 0.09 for dataset 3

(Fig. SN6c). With this modest difference in the frequency of missing values among the three datasets, the independent *scBiPaD* analysis of each dataset produced an output essentially identical to the result of the merged dataset (compare Fig. 1 in the manuscript with Fig. SN3).

Having established the fraction of missing values for the three datasets, we asked whether and to what extent the performance of the *scBiPaD* approach is affected by different frequency of missing values. In order to interrogate how the pathway-based features degrade as the dropout increases, we performed a set of simulations in a controlled setting where we can have a ground truth (gold standard) that evaluates the stability of the pathway-based features of *scBiPaD* as the dropout rate increases. Simulated data were adjusted to explore different levels of dropout and used to compare the accuracy and sensitivity of discovery of enriched biological pathways using a ranked list of genes. The controlled dataset used unprocessed RNAseq data and RNAseq data that had been pre-processed using the MAGIC imputation algorithm to model the enrichment of pre-established biological classes. The Area Under the Curve (AUC) was computed as measure of accuracy.

Synthetic data generation. We generated a standard RNA-Seq profile by averaging 146 samples from TCGA-GBM database. The averaged profile defined an empirical distribution function without dropout that we assumed to be the theoretical model. We then derived from this profile the significant enriched pathways using MWW-GST (logit(NES) > 0.58)[10], and we refer to these gene sets as "true positive" gene sets. We then generated a data matrix X of s samples by first assigning to each sample the same standard expression profile. To simulate dropout of a gene i in a single cell j, we implemented a model built with similar parameters as the Splat model that exhibited the most accurate and reproducible simulation of scRNA-seq data[47]. Briefly, we used the relationship between the mean expression of a gene and the proportion of zero counts in that gene to model this process and use a logistic function to produce a probability that a count should be zero. The rationale of this approach was motivated by the notion that the probability of dropout in scRNAseq experiments is dependent on the level of expression of the gene, with the lowest level of gene expression corresponding to the highest probability of missing values of those genes. The logistic function is defined as:

$$\pi_{ij}^{D} = \frac{L}{1 + e^{-k(x_{ij} - x_0)}}$$

where L is the curve's maximum value and as we model a probability, the maximum value of L is set to 1; k is the logistic growth rate or steepness of the curve; x_0 is the value of the sigmoid midpoint, here set as 0; x_{ij} is the expression level of the *i*-th gene in the *j*-th sample. In this model, k should be interpreted as the desired level of dropout (d), by leaving all other parameter fixed (Fig. SN7a). As an example, the logistic function was modeled for two different levels of k (or dropout rate) with k= -200 (dropout 1%) and k= -0.8 (dropout 20%) and calibrated within the range of expression of the averaged profile x used as theoretical model (red boxes, Fig. SN7b).

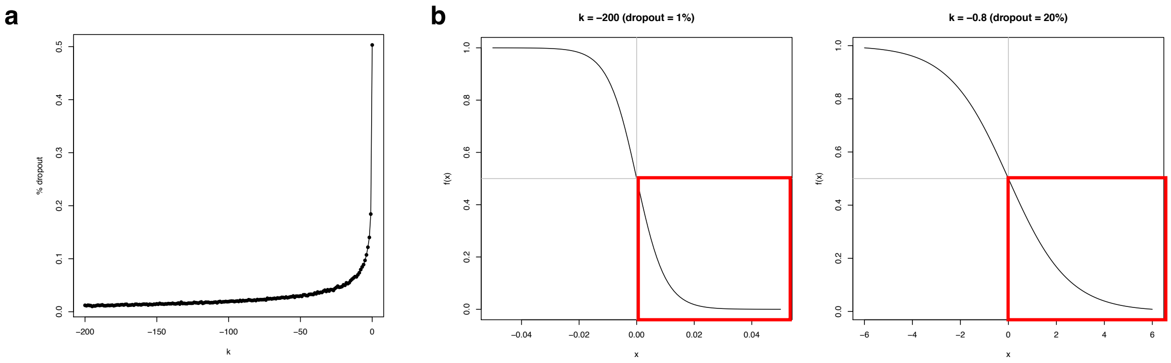

Figure SN7. a, Change in gene dropout on the basis of the parameter *k* of the logistic function. **b,** Logistic function for two different levels of *k* (or dropout rate). The red boxes indicate the range of expression of the averaged profile used as theoretical model.

The probability of a zero for each gene was then used to randomly replace some of the simulated counts with zeros using a Bernoulli distribution defined as:

$$1_{ij}^{D} \sim Ber\left(\pi_{ij}^{D}\right).$$

The final count matrix is then $X_{ij} = 1_{ij}^{D} \cdot X_{ij}$. Lastly, we derived X_{ij}^{M} by applying MAGIC imputation on the final matrix X_{ij}.

Results. Having fixed $s = 100$ simulated samples, we evaluated different levels of k and imputation, by performing seven experiments corresponding to $d = 0.01, 0.02, 0.05, 0.10,\ 0.15, 0.20, 0.50$. Using the levels of $k = -200, -104, -20, -4.5, -1.6, -0.8, 0$ and from $s = 100$ random samples, we determined the simulated level of dropout, which corresponds to the desired level of percentage of zero-value genes d in each cell (Fig. SN8a). To extract the predicted gene sets from the model, we established the significantly enriched pathways from each random sample and each level of k using MWW-GST (logit(NES) > 0.58).

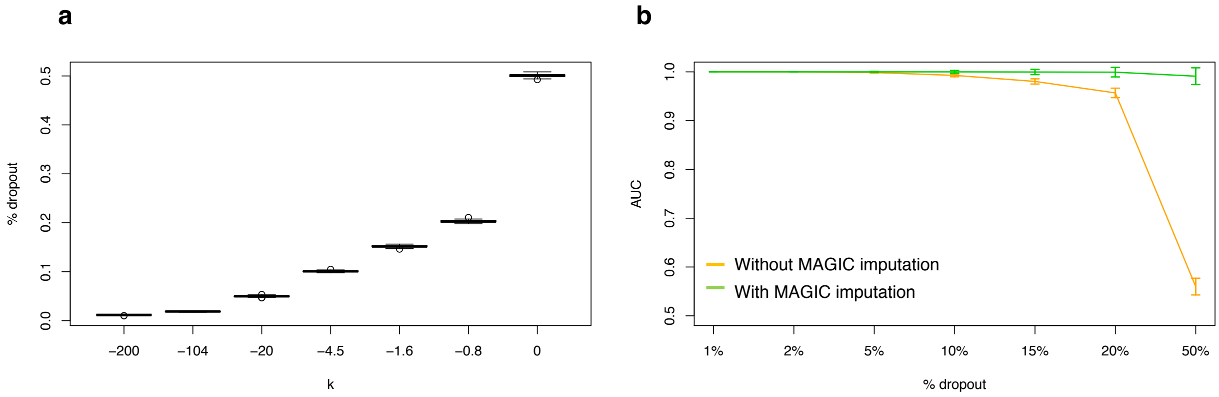

Figure SN8. a, Box plot of the *n* = 100 simulated samples with different levels of *k* (or dropout rate). Boxplots span the first to third quartiles and whiskers show the 1.5× interquartile range. **b,** Trend of the AUC value for the 100 simulated samples for different levels of *k* (or dropout rate) with or without MAGIC imputation (green and orange line, respectively). Curves indicate the average AUC and errors bars indicate the confidence interval from *n* = 100 simulations.

Finally, we compared the predicted gene sets to the true positive gene sets and derived the Area Under the Curve (AUC) from the Receiver Operating Characteristics (ROC) curve, a performance measurement for classification experiments. AUC values range from 0 to 1, with 0.5 indicating uninformative classifiers, and 1 representing perfect classification. In Fig. SN8b, we report the average and confidence interval of the AUC values among $s = 100$ random samples for each percentage of zero-value genes d in each cell. Without MAGIC imputation, AUC values were very stable as the dropout rate reaches 10%, and exhibited the first detectable loss at dropout rate of 15%. At dropout rate of 50%, the AUC markedly degraded to a level that was barely higher than random guess. The application of the MAGIC imputation stabilized the AUC above 99% even at the dropout rate of 50%, which is more than two-fold higher than the maximum level of dropout in the scRNAseq datasets used in the manuscript (Fig. SN6). Thus, the pathway-based *scBiPaD* analysis, which relies on meta-genes (the pathways) rather than the expression of individual genes, is only minimally affected by the range of gene dropout in the scRNAseq datasets (Fig. SN8b). The application of MAGIC imputation stabilized the performance of the classification to levels of dropout that are well above those detected in scRNAseq datasets.

6. Evaluation and resolution of pathway redundancy

Feature redundancy is a common issue in any feature selection experiment. To address the effect of pathway redundancy in the *scBiPaD* computational platform, we first used the set cover algorithm[15], a method that reduces pathway redundancy without merging pathways and maximizing gene coverage, to build enrichment map gene ontology networks. Next, we selected the association with patient survival to reduce the 5,032 GO pathways to a subset of 192 for the classification of bulk tumors (shown in Extended Data Fig. 5a, b of the manuscript). To assess whether the robustness of the pathway enrichment analysis was different when using the 192 survival-associated pathways or the 5,032 gene sets, we independently applied the *scBiPaD* pipeline to the single cell datasets using 192 survival-associated pathways and compared the class assignment of the two analyses (192 versus 5,032 GO terms). The experiment, which is presented in Extended Data Fig. 4, showed a concordance in biological state classification of individual cells of 92% between *scBiPaD* pipeline applied to 192 survival-associated and unselected 5,032 biological pathways. Therefore, the collection of 192 survival-associated pathways retains the ability to capture the core biological states of GBM cells.

Here, we measured pathway redundancy reduction by survival-based feature selection (192 pathways) in comparison with the full GO terms collection (5,032). It is important to note that in the entire collection there are biologically different subsets of GO categories composed of few redundant functionally related pathways that share common genes (for example, cell cycle, immune, mitochondrial categories). We used the Jaccard index to assess the similarity between pathways and set four different thresholds (0.05, 0.10, 0.20, 0.50) above which we computed the number of pathways exhibiting similar features.

The rationale for testing different thresholds was to determine whether the redundancy reduction is stable and unaffected by cutoff selection of the Jaccard index. As expected, the median number of redundant pathways decreased as the Jaccard threshold increased. Depending on the Jaccard index threshold, variable degrees of redundancy emerged from the 5,032 gene sets but redundancy was consistently very low in the 192 pathways setting, regardless of the particular index of gene overlap included in the analysis (Fig. SN9a).

To provide additional evidence for the non-random resolution of redundancy in the set of 192 survival-associated pathways, we derived a null distribution (1,000 random samples) of the average level of overlap between subsets of 192 pathways (of similar gene set size to the survival-associated pathway collection), randomly selected from the original universe of 5,032 candidates.

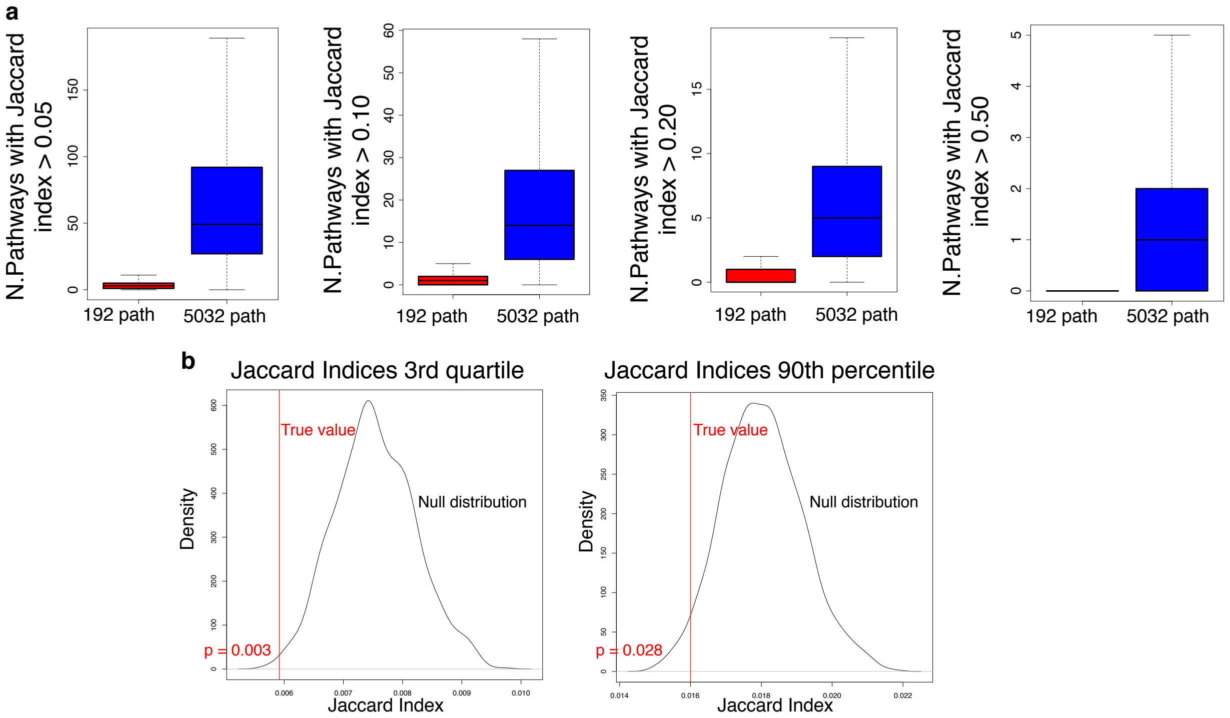

Figure SN9. a, Box plots showing the distribution, for each gene set, of biological pathway pairs (survival-based feature selection, *n* = 192; GO terms, *n* = 5,032) with Jaccard index higher than 0.05 (first panel from left), 0.10 (second panel), 0.20 (third panel) and 0.50 (fourth panel). Boxplots span the first to third quartiles and whiskers show the 1.5× interquartile range. **b,** Distribution of the third quartiles (left panel) and 90th percentile (right panel) of the Jaccard index of 192 pathways derived from of 1,000 random sampling of 192 pathways across the whole 5,032 collection from 20 bins by gene size of equal size. The vertical red line represents the true value of Jaccard Index from the 192 pathways in the two combinations.

As pathway redundancy typically involves clusters of functionally redundant gene sets, deriving Jaccard index from all possible pairs of pathways generates sparse matrices with many zero values, resulting in poor representation of the event under study by any measure of centrality. To mitigate this effect, the comparisons of the similarity values interrogating redundancy of the 192 survival-associated pathways and the 1,000 randomly selected 192 subsets of pathways were independently done using two measures of central tendency - 3rd quartile and 90th percentile – to represent average redundancy among clusters of pathways. The model showed that, with either of the two measures of central tendency of the Jaccard index (3rd quartile and 90th percentile), redundancy among pathways from the survival-associated feature set was significantly lower than expected by chance (p = 0.003 and p = 0.028, respectively, Fig. SN9b). Therefore, the 192 survival-associated pathways exhibit significantly lower redundancy than the 5,032 collection.

7. Analysis of PDCs

Whole exome sequencing, mRNA sequencing and CNV analysis of PDCs

Sequencing methods and analyses were performed as recently described[17]. Somatic copy number profiles were estimated from WES data by CNVkit[48] using the integrated profile from paired blood samples as normal reference. CNV were classified according to segmentation ratio as homozygous deletions for $\log_2$(ratio) < -2 (CNVkit score = -2), heterozygous deletions for $\log_2$(ratio) < -0.75 (CNVkit score = -1), gain for $\log_2$(ratio) > 0.5 (CNVkit score = 1) and high-level amplification for $\log_2$(ratio) >1 (CNVkit score = 2).

Classification of PDCs

To classify PDCs, we applied the random forest machine learning algorithm[49] implemented in the R package *random Forest*. The feature set of the classifier included the z-score of the 100 highest scoring genes in the ranked-list of each GBM subtype, 1,576 amplified, 1,443 deleted and 88 mutated genes representing the genetic alterations most significantly associated with the GBM subtypes in the TCGA cohort (two-sided Fisher exact test and χ^2 $p < 0.01$), which were used as the training set. To optimize algorithm parameters, we considered different parameter configurations to build decision trees and number of genes selected at each split (from 1 to 500 *ntree* and *mtry*), using a 5-fold cross validation by randomly selecting a set of 80% TCGA samples for training and evaluating the performance on the remaining 20%. We selected the best performing configuration of parameters (*mtry* = 117, *ntree* = 33), for which we obtained the lowest predictive error rate (14%). Feature importance was computed using the Gini index. A score representative of the enrichment of each individual GBM subtype was calculated for each PDC as the average of the NES of TCGA gene signatures and the meta-signatures from the combination of all three single cell datasets. Pathway enrichment in each PDC was computed by ssMWW-GST.

Association between GBM subtype and CNVs in PDCs

Significance of the associations between CNVs and the GBM subtypes was evaluated using the two-sided Fisher's exact test. CNVkit scores of -2 and -1 were considered deletions and 1 and 2 were considered amplifications. Only genes with *f*CNV calls present in at least two PDCs were tested (4,734 deleted and 3,194 amplified genes) retaining genes with a $p < 0.07$ for the specific association.

References

1. Neftel, C., *et al.* An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell* **178**, 835-849 e821 (2019).

2. Yu, K., et al. Surveying brain tumor heterogeneity by single-cell RNA-sequencing of multi-sector biopsies. *National Science Review* **7**, 1306-1318 (2020).

3. Yuan, J., *et al.* Single-cell transcriptome analysis of lineage diversity in high-grade glioma. *Genome Med* **10**, 57 (2018).

4. DePasquale, E.A.K., *et al.* DoubletDecon: Deconvoluting Doublets from Single-Cell RNA-Sequencing Data. *Cell Rep* **29**, 1718-1727 e1718 (2019).

5. Ilicic, T., *et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome Biol* **17**, 1-15 (2016).

6. Lukassen, S., Bosch, E., Ekici, A.B. & Winterpacht, A. Single-cell RNA sequencing of adult mouse testes. *Sci Data* **5**, 180192 (2018).

7. Lun, A.T., McCarthy, D.J. & Marioni, J.C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* **5**, 2122 (2016).

8. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502 (2015).

9. Wang, L., *et al.* The Phenotypes of Proliferating Glioblastoma Cells Reside on a Single Axis of Variation. *Cancer Discov* **9**, 1708-1719 (2019).

10. Frattini, V., *et al.* A metabolic function of FGFR3-TACC3 gene fusions in cancer. *Nature* **553**, 222-227 (2018).

11. Calinski, T. A Dendrite Method for Cluster Analysis. *Biometrics* **24**, 207-& (1968).

12. Van der Laan, M.J., Pollard, K.S. & Bryan, J. A new partitioning around medoids algorithm. *J Stat Comput Sim* **73**, 575-584 (2003).

13. Isserlin, R., Merico, D., Voisin, V. & Bader, G.D. Enrichment Map - a Cytoscape app to visualize and explore OMICs pathway enrichment results. *F1000Res* **3**, 141 (2014).

14. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431-432 (2011).

15. Stoney, R.A., Schwartz, J.M., Robertson, D.L. & Nenadic, G. Using set theory to reduce redundancy in pathway sets. *BMC Bioinformatics* **19**, 386 (2018).

16. Caruso, F.P.*, et al.* A MAP of tumor-host interactions in glioma at single cell resolution. GigaScience **9**, (2020).

17. D'Angelo, F.*, et al.* The molecular landscape of glioma in patients with Neurofibromatosis 1. *Nat Med* **25**, 176-187 (2019).

18. Zhang, J.*, et al.* The combination of neoantigen quality and T lymphocyte infiltrates identifies glioblastomas with the longest survival. *Commun Biol* **2**, 135 (2019).

19. Colaprico, A.*, et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* **44**, e71 (2016).

20. Hollander, M., Wolfe, D.A. & Chicken, E. *Nonparametric statistical methods*, (John Wiley & Sons, Inc., Hoboken, New Jersey, 2014).

21. Chen, H.*, et al.* Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat Commun* **10**, 1903 (2019).

22. Wang, J.*, et al.* Clonal evolution of glioblastoma under therapy. *Nat Genet* **48**, 768-776 (2016).

23. Mermel, C.H.*, et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).

24. Trifonov, V., Pasqualucci, L., Dalla Favera, R. & Rabadan, R. MutComFocal: an integrative approach to identifying recurrent and focal genomic alterations in tumor samples. *BMC Syst Biol* **7**, 25 (2013).

25. Sarto Basso, R., Hochbaum, D.S. & Vandin, F. Efficient algorithms to discover alterations with complementary functional association in cancer. *PLoS Comput Biol* **15**, e1006802 (2019).

26. Edgington, E.S. Additive Method for Combining Probability Values from Independent Experiments. *J Psychol* **80**, 351-363 (1972).

27. Aryee, M.J.*, et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363-1369 (2014).

28. Saghafinia, S., Mina, M., Riggi, N., Hanahan, D. & Ciriello, G. Pan-Cancer Landscape of Aberrant DNA Methylation across Human Tumors. *Cell Rep* **25**, 1066-1080 e1068 (2018).

29. Dweep, H. & Gretz, N. miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nat Methods* **12**, 697 (2015).

30. Barthel, F.P.*, et al.* Longitudinal molecular trajectories of diffuse glioma in adults. *Nature* **576**, 112-120 (2019).

31. Ceccarelli, M.*, et al.* Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell* **164**, 550-563 (2016).

32. Carter, S.L.*, et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* **30**, 413-421 (2012).

33. Matsui, Y.*, et al.* phyC: Clustering cancer evolutionary trees. *PLoS Comput Biol* **13**, e1005509 (2017).

34. Phillips, H.S.*, et al.* Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* **9**, 157-173 (2006).

35. Wang, Q.*, et al.* Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. *Cancer Cell* **32**, 42-56 e46 (2017).

36. Filbin, M.G.*, et al.* Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. *Science* **360**, 331-335 (2018).

37. Patel, A.P.*, et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401 (2014).

38. Suva, M.L. & Tirosh, I. Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. *Mol Cell* **75**, 7-12 (2019).

39. Tirosh, I.*, et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-196 (2016).

40. Tirosh, I.*, et al.* Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309-313 (2016).

41. Venteicher, A.S.*, et al.* Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science* **355**(2017).

42. Boisselier, B.*, et al.* Whole genome duplication is an early event leading to aneuploidy in IDH-wild type glioblastoma. *Oncotarget* **9**, 36017-36028 (2018).

43. Louis, D.N.*, et al.* The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol* **131**, 803-820 (2016).

44. Xiao, Z., Dai, Z. & Locasale, J.W. Metabolic landscape of the tumor microenvironment at single cell resolution. *Nat Commun* **10**, 1-12 (2019).

45. Kim, S., Kon, M. & DeLisi, C. Pathway-based classification of cancer subtypes. *Biol Direct* **7**, 21 (2012).

46. Stuart, T.*, et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821 (2019).

47. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol* **18**, 1-15 (2017).

48. Talevich, E., Shain, A.H., Botton, T. & Bastian, B.C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol* **12**, e1004873 (2016).

49. Breiman, L. Random forests. *Mach Learn* **45**, 5-32 (2001).