Online Methods

Tumor sample preparation

Human brain metastasis of cutaneous melanoma (CM) patients were diagnosed and treated at the Vall d'Hebron University Hospital (Barcelona, Spain). A written consent, indicating that the samples obtained as standard of care in the clinical practice could be used for research, was obtained from the parents/legal representant in agreement with the declaration of Helsinki. The study was approved by the local IRB. The uveal melanomas (UM), liver metastasis of the uveal melanoma and ovarian cancer (OC) were obtained from the Catalan Institute of Oncology (ICO). The study was approved by the local IRB: PR167/18 - Identification of immunity against GNAQ/11 mutations in uveal melanoma patients (PR070/18 - Identification of the antigenic potential of driver mutations in Metastatic Uveal Melanoma. Generating therapeutic windows of opportunity).

After tumor resection, tissues were enzymatically digested using human tumor dissociation kit (Miltenyi Biotec). To isolate PTPRC+ cells, PTPRC TIL microbeads (Miltenyi Biotec) were used following manufacturer's instructions and stored in cold 1X PBS supplemented with 0.1% BSA until partitioning for single cell RNA-sequencing. The uveal melanoma and ovarian cancer samples were transferred from the surgery room to the single-cell laboratory in DMEM medium (Gibco) on ice (<1h). Upon arrival to the laboratory, samples were washed twice with cold 1X HBSS (ThermoFisher Scientific) and minced on ice using razor blades. Next, the samples were incubated during 15 min at 37°C under constant rotation at 700 rpm with 2 ml of a pre-warmed dissociation mix (200 U/ml Collagenase IV, 1X HBSS). Every 5 min, samples were pipetted with a wide bore p1000 tip to help tissue dissociation. After 15 min, samples were passed 10 times through a 0.9 mm needle and 10 additional times through a 0.6 mm needle. Then, the enzymatic digestion was stopped by adding 10% fetal bovine serum (FBS). Samples were filtered with a 70 µM cell strainer (pluriSelect) and centrifuged during 5 min at 500 rcf at 4°C. Pellets were washed twice with cold 1X HBSS and resuspended in 1X PBS supplemented with 0.05% BSA (Miltenyi Biotec). Cell concentration and viability were verified by counting with a TC20™ Automated Cell Counter (Bio-Rad Laboratories, S.A). In the case of UM liver metastasis, PTPRC+ cells were isolated by Magnetic-activated cell sorting (MACS) using the OctoMACS™ Separator and MS columns (Miltenyi Biotec) following manufacturer's instructions. Briefly, cells were incubated with human PTPRC MicroBeads during 30 min at 4°C and PTPRC+ cells were separated by applying the cell suspension onto a pre-rinsed MS column. Cell concentration and viability were again verified with a TC20™ Automated Cell Counter (Bio-Rad Laboratories, S.A).

For the ovarian cancer, cells were subjected to a Cell Hashing protocol before proceeding to single-cell RNA sequencing. Cell hashing was performed following manufacturer's instructions (Cell hashing and Single Cell Proteogenomics Protocol Using TotalSeq™ Antibodies; BioLegend). Briefly, the sample was resuspended in Cell Staining Buffer (Bio Legend) and split into four tubes with an equal number of cells. Next, sample tubes were incubated during 10 min at 4°C with Human TruStain FcX™ Fc Blocking reagent (BioLegend) and then a specific TotalSeq-A antibody-oligo conjugate was added to each tube and incubated on ice for 1h. Cells were then washed three times with cold PBS-0.05% BSA (ThermoFisher) and centrifuged for 5 min at 500 rcf at 4°C. Finally, cells were resuspended in an appropriate volume of 1X PBS-0.05% BSA in order to obtain a final cell density > 500 cells/ul, suitable for 10x Genomics scRNA-seq. An equal volume of hashed cell suspension from each of the four tubes was mixed and filtered with a 40 µm strainer. Cell concentration was verified by counting with a TC20™ Automated Cell Counter (Bio-Rad Laboratories, S.A).

Mouse Colorectal cancer (CRC)

Isolation of tumor cells for generation of mouse tumor organoids (MTOs) and in vitro expansion was previously described (Tauriello et al. 2018). Briefly, for mouse liver metastasis generation, 3-day grown MTOs were trypsinized followed by mechanical dissociation to obtain a single cell suspension. 6-8-week-old C57BL/6J mice (Janvier Labs) were subjected to intrasplenic injection of 3x10^5 single cells in 70µl HBSS, followed by splenectomy to prevent splenic tumor growth. The mice were euthanized 16 days post-injection and livers were collected in HBSS.

Tumor nodules were dissected from livers and minced with scalpels. The tissue was enzymatically digested with 0.2 mg/ml Collagenase IV (Sigma, ref: C5138), 0.2 mg/ml Dispase II (Sigma, ref: D4693) and 0.04 mg/ml DNase I (Sigma, ref: 10104159001) in 10%FBS/DMEM (Life Technologies). Enzymatic digestion consisted of mechanical dissociation during 4 cycles of 20 minutes at 37ºC. The enzymatic reaction was quenched at the end of each cycle, by the addition of 30 ml of ice-cold HBSS (10% FBS supplemented). The cell suspension was filtered using 100 µm strainers (Corning). Lysis of erythrocytes was performed using red cell lysis buffer (150 mM NH4Cl, 10mM NaHCO3, 0.1 mM EDTA). Cells were incubated with FACS buffer (PBS, 2% FBS) containing blocking antibodies against CD16/CD32 at 4ºC for 20 minutes to block the Fc receptor. Then, BV605-conjugated antibodies against EpCAM were used to stain the cells for 20 minutes at 4ºC. Lastly, EpCAM-negative cells were sorted in a FACS Aria Fusion flow cytometer (BD Biosciences) from total viable cells, determined by nuclear staining with DAPI (Sigma).

Single-cell RNA sequencing (scRNA-seq)

Cells were partitioned into Gel Bead-In-Emulsions by using the Chromium Controller system (10x Genomics) aiming at Target Cell Recovery of 5000-7000 total cells per sample, in the case of CM, UM, and mouse CRC, and a Target Cell Recovery of 10,000 total cells in the case of the OC. For the CM brain metastasis and a mouse CRC sample, single-cell Gene Expression (GEX) and T cell receptor (TCR)-enriched libraries were prepared using the Chromium Single Cell 5´ Library and Gel Bead Kit (10x Genomics, Cat. N. 1000006) following manufacturer's instructions. In the case of the UM, OC samples and one mouse CRC, gene expression libraries were prepared using the Chromium Single-cell 3’ mRNA kit (V3; 10X Genomics, Cat. N. 1000075), following manufacturer's instructions. Briefly, after GEM-RT clean up, cDNAs were amplified during 16 cycles for CM and mouse CRC samples, 12 cycles for UM samples and 11 cycles for the ovarian cancer sample. cDNA QC and quantification were performed on an Agilent Bioanalyzer High Sensitivity chip (Agilent Technologies).

For the OC sample, the GEX library was prepared using the single-cell 3’ mRNA kit (V3; 10x Genomics) with some adaptations for cell hashing, as indicated in TotalSeq™-A Antibodies and Cell Hashing with 10x Single Cell 3' Reagent Kit v3 3.1 Protocol by BioLegend. Briefly, 1 µl of 0.2 µM HTO (Hashtag Oligonucleotides) primer (Integrated DNA Technologies, IDT) was added to the cDNA amplification reaction in order to amplify the hashtag oligos together with the full-length cDNAs. A SPRI selection clean-up was done in order to separate mRNA-derived cDNA (>300 bp) from antibody-oligo-derived cDNA (<180 bp), as described in the above-mentioned protocol from BioLegend. The cDNA sequencing library was prepared following 10x Genomics single-cell 3’ mRNA kit protocol, while HTO cDNA was indexed by PCR as follows. Briefly, 5 µl of purified hashtag oligo cDNA were mixed with 2.5 µl of 10 µM Illumina TruSeq D708_s primer (IDT), 2.5 µl of SI primer from 10 X single-cell 3’ mRNA kit, 50 µl of 2X KAPA HiFi PCR Master Mix (KAPA Biosystem) and 40 µl of nuclease-free water. The reaction was carried out using the following thermal cycling conditions: 98°C for 2 min (initial denaturation), 12 cycles of 98°C for 20 sec, 64°C for 30 sec, 72°C for 20 sec, and a final extension at 72°C for 5 min. The HTO library was purified by adding 1.2X SPRI select reagent to the PCR reaction.

All cDNA libraries were indexed by PCR using the PN-220103 Chromiumi7 Sample Index Plate. Size distribution and concentration of 3' and 5' GEX libraries, TCR-enriched and HTO libraries, were verified on an Agilent Bioanalyzer High Sensitivity chip (Agilent Technologies). Finally, sequencing of all libraries was carried out on an Illumina NovaSeq6000 sequencer to obtain approximately 40,000 reads/cell, in the case of GEX libraries, and 2000 reads/cell for the TCR-enriched and HTO libraries.

Antibodies hashtag oligo sequences.

ANTIBODY NAME HASHTAG	OLIGO SEQUENCE
anti-human Hashtag 1	GTCAACTCTTTAGCG
anti-human Hashtag 2	TGATGGCCTATTGGG
anti-human Hashtag 3	TTCCGCCTCTCTTTG

anti-human Hashtag 4	AGTAAGTTCAGCGTA

Primers used for single cell library construction.

NAME	**SEQUENCE**
10x Genomics SI-PCR primer	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGC*T*C
HTO cDNA PCR additive primer	GTGACTGGAGTTCAGACGTGTGC*T*C
Illumina TruSeq D708_s primer	CAAGCAGAAGACGGCATACGAGATGCGCATTAGTGACTGGAGTTCAGACGTGT*G*C

*Phosphorothioate bond

Benchmarking of data integration methods

To ensure that the defined cell types and states were independent of the computational tool used for data integration, we compared Seurat's correction with Scanorama (Hie et al. 2019) and Harmony (Korsunsky et al. 2019), previous identified to preserve biological variation and being scalable to hundreds of thousands of cells (Luecken et al. 2020). We ran Harmony v1.0 within the framework of the SeuratWrappers package (v0.2.0). Specifically, we used the RunHarmony function, with the top 50 principal components as latent space and the dataset of origin (also referred to as *source*) as batch label. The resulting 50 batch-corrected principal components were used as input to the RunUMAP function of Seurat v3.2.0. To run Scanorama (v1.7), we first split the atlas into a list of cancer subtype-specific *anndata* objects, and subsetted these keeping only highly variable genes. We then run the integrate_scanpy function with the aforementioned list and setting the *dimred* parameter to 50. Finally, we calculated the Local Inverse Simpson Index (LISI), with the calculate_lisi function of the lisi v1.0 package.

Validation of clusters and signatures using a random forest (RF) classifier

To assess the robustness of the clusters and their respective signatures, we trained a RF classifier as follows. First, we computed 25 cell type-specific signatures with the *AddModuleScore* function from *Seurat* using the markers in **Supplementary Table 3** that were kept in the integrated expression matrix. Additionally, we calculated a random signature per cell type by sampling without replacement as many genes as present in the real signatures. Second, we performed 5-fold cross validation to test both the bias and the variance of the classifier. We sought to ensure that all cell types were present in the five test sets. Thus, each test set of each fold was built by taking 20% of the cells in each cell type, which resulted in each cell appearing in one test set and four training sets. Third, we trained a RF classifier in each of the five training sets with the *randomForest* v4.6.14package; using the signatures as features and the cell type annotation as target variable. Finally, we tested the accuracy of the classifier in the respective test set using the *confusionMatrix* function from the *caret* v6.0.86 package.

Deconvolution of TCGA bulk RNA sequencing data

To deconvolute bulk RNA sequencing tumor data using the single-cell reference atlas, we applied SPOTlight (Elosua-Bayes et al. 2021) on samples from four different TCGA cancer types: 1,222 breast invasive carcinoma (TCGA-BRCA), 594 lung adenocarcinoma (TCGA-LUAD), 472 skin cutaneous melanoma (TCGA-SKCM) and 521 colon adenocarcinoma (TGCA-COAD). HTSeq counts from all tumor types were downloaded from https://portal.gdc.cancer.gov using the TCGAbiolinks package (Colaprico et al. 2016). To minimize dataset variability, we used cell types contributing >1% to the first or second principal components (Figure 2C). The Seurat's function FindAllMarkers was used to obtain the marker genes of each cell subtype, subsequently used to initialize the model. After deconvolution, we applied our Random Forest classifier to predict the corresponding single-cell derived patient immune clusters.

Human oropharyngeal cancer samples

All patients provided informed consent for the collection of human specimens and data. This was approved by the St Vincent's Hospital Research Office (2019/PID04335) in accordance with the National Health and Medical Research Council's National Statement of Ethical Conduct in Human Research. Patients undergoing surgical resection for a locally advanced oropharyngeal cancer were recruited to the study. After surgical removal, the anatomical pathologist dissected a sample of both the primary and nodal metastasis. Samples were tumour banked in accordance with our ethically approved protocol. Within 30 min of collection, tumour samples were tumour banked. Samples were cut into 1mm x 1mm chunks with a scalpel blade. For Visium, a tissue chunk was snap frozen in OCT. After freezing, samples were moved to liquid nitrogen for long term storage.

Visium Spatial Gene Expression

Frozen tissue samples were processed using the Visium Spatial Gene Expression slide and reagent kit (10x Genomics, US) following the manufacturer's instruction. Briefly, 10 μm sections were placed into the capture areas of the Visium slide. Tissue morphology was assessed with H&E staining and imaging using a Leica DM6000 microscope equipped with a 20x lens (Leica, DE). The imaged sections were then permeabilized for 12 minutes using the supplied reagents. The permeabilization condition was previously optimised using the Visium Spatial Tissue Optimisation slide and reagent kit (10x Genomics, US). After permeabilization, cDNA libraries were prepared, checked for quality and sequenced on a NovaSeq6000 platform (Illumina, US).

Around 300 million pair-ended reads were obtained for each tissue section. Read 1, i7 index and Read 2 were sequenced with 28, 8 and 98 cycles respectively.

Visium data quality control

Quality control was carried looking at the number of UMIs, genes and mitochondrial percentage. Spots with <1.000 UMIs were removed from the analysis due to poor quality. Three tissue slices were found on the same Visium slide area, two were mostly on the spots while the third was mainly off the capture area. The latter was discarded while the other two where separated and treated as different datasets. Data was scaled and normalized using *SCTransform* with default parameters. Principal component analysis was carried out to reduce the dimensionality, the top 30 principal component's along with a 0.25 resolution were used to cluster and obtain the UMAP embedding of the spots.

Human breast cancer

We used human breast cancer Visium ST data publicly available through the 10x Genomics website (https://support.10xgenomics.com/spatial-gene-expression/datasets/). Two replicates were available and used to confirm the *SPOTlight* predictions. Spatial transcriptomics data run through *spaceranger* 1.0.0 were used (10x Genomics). Further specifications on how the samples were obtained and processed can be found at the 10x Genomics website.

Quality control analysis was carried out looking at the number of unique molecular identifiers (UMIs), genes, and mitochondrial percentage in each spot. No spots were removed from the analysis after inspection of the aforementioned parameters. Data was scaled and normalized using *SCTransform* with default parameters. Principal component analysis was carried out to reduce the dimensionality, the top 30 principal component's along with a 0.1 resolution were used to cluster and obtain the UMAP embedding of the spots. Clusters were annotated using a pathologist's annotation to separate tumor from fibrotic. Within tumor, clusters were annotated according to the expression of *ESR1*, *PGR* and *ERBB2*. We used *SPOTlight* to map the tumor immune cell atlas to the spatial spots. Therefore, a subset was used to train the model, we selected up to 100 cells coming from melanoma cancers for each cell type/state. Selecting cells from one of the cancers types allowed us to reduce dataset-specific noise which could confound the model. The gene set used to train the model was the union between the marker genes of the cell types along with the top 3000 variable genes. Marker genes for each cell type were obatined with Seurat's function *FindAllMarkers* considering only positive markers and setting the logFC and min.pct to 0 to include all genes. All markers were used to initialize the model basis and unit variance normalization was carried out. Nonsmooth nonnegative matrix factorization was the method used to carry out the factorization. Cell types contributing <1% to the spot's predicted composition were considered fitting noise and were set as 0. We then used *SPOTlight* to map the atlas cell types to the spatial spots.

Code versions and availability

All analyses were carried out using R3.6.0 and the data was analyzed using *Seurat* v3.2 (Stuart et al. 2019). Furthermore, *SPOTlight* is developed to run with R versions ≥3.5; docker images with the appropriate environment are available at Docker hub: marcelosua/spotlight_env_rstudio and marcelosua/spotlight_env_r.

References (Online Methods)

Alashwal H, El Halaby M, Crouse JJ, Abdalla A, Moustafa AA. 2019. The Application of Unsupervised Clustering Methods to Alzheimer's Disease. *Front Comput Neurosci* **13**. https://www.frontiersin.org/articles/10.3389/fncom.2019.00031/full (Accessed September 20, 2020).

Chen B, Harrison R, Pan Y, Tai PC. 2005. Novel Hybrid Hierarchical-K-means Clustering Method (H-K-means) for Microarray Analysis. In *Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference - Workshops*, *CSBW '05*, pp. 105–108, IEEE Computer Society, USA https://doi.org/10.1109/CSBW.2005.98 (Accessed September 20, 2020).

Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, et al. 2016. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* **44**: e71.

Elosua-Bayes M, Nieto P, Mereu E, Gut I, Heyn H. 2021. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res*. https://doi.org/10.1093/nar/gkab043 (Accessed April 23, 2021).

Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, et al. 2015. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16**. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4676162/.

Hie B, Bryson B, Berger B. 2019. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat Biotechnol* **37**: 685–691.

Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh P, Raychaudhuri S. 2019. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**: 1289–1296.

Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L, Dugas M, Colomé-Tatché M, et al. 2020. Benchmarking atlas-level data integration in single-cell genomics. *bioRxiv* 2020.05.22.111161.

Mereu E, Lafzi A, Moutinho C, Ziegenhain C, McCarthy DJ, Álvarez-Varela A, Batlle E, Sagar, Grün D, Lau JK, et al. 2020. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat Biotechnol* 1–9.

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, Satija R. 2019. Comprehensive Integration of Single-Cell Data. *Cell* **177**: 1888-1902.e21.

Tauriello DVF, Palomo-Ponce S, Stork D, Berenguer-Llergo A, Badia-Ramentol J, Iglesias M, Sevillano M, Ibiza S, Cañellas A, Hernando-Momblona X, et al. 2018. TGFβ drives immune evasion in genetically reconstituted colon cancer metastasis. *Nature* **554**: 538–543.

van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, Burdziak C, Moon KR, Chaffer CL, Pattabiraman D, et al. 2018. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**: 716-729.e27.