Supplementary Methods:

Ethics approval and consent to participate

Study protocols were approved by the UCSF Institutional Review Board. All clinical samples were provided by the Neurosurgery Tissue Core in a deidentified fashion. All experiments were carried out in conformity to the principles set out in the WMA Declaration of Helsinki as well as the Department of Health and Human Services Belmont Report. Informed written consent was provided by all patients.

Tissue processing and sc/snRNA-seq

Fresh tissue was dissociated mechanically in the presence of collagenase, deoxyribonuclease, trypsin inhibitor, and DPBS. The homogenate was shaken in a thermomixer at 37˚ for 15 minutes, triturated via pipette, and shaken again at 37˚ for 15 minutes. The cells were then filtered through a 70 micron strainer, pelleted, and resuspended in neural basal media with FBS. Frozen tissues were dissociated mechanically and incubated with Nuclei EZ lysis buffer (Sigma) according to the manufacturer's protocols. Nuclei were then filtered through a 70 micron strainer. Debris was further depleted via a sucrose-based density gradient. The 10X Chromium single-cell platform (v3 3'-chemistry) was used for cell/nuclei capture and library preparation was performed per 10X Genomics' protocols. Sequencing was performed on an Illumina NovaSeq using a paired-end 100bp protocol.

Public data acquisition and processing

Peripheral blood mononuclear cell (PBMC) data were obtained from 10x Genomics (https://support.10xgenomics.com/single-cell-gene-expression/datasets). Human pancreas Smart-seq2 data were obtained from the Sandberg lab (Segerstolpe *et al.*, 2016). InDrop data were obtained from the Yanai lab (Baron *et al.*, 2016). SnRNA-Seq and scRNA-Seq cortex data were obtained from the Zhang lab (Lake *et al.*, 2017), Quake lab (Darmanis *et al.*, 2015) and Kriegstein lab (Nowakowski *et al.*, 2017). Bulk RNA-Seq data were obtained from the Ivy Foundation Glioblastoma Atlas Project (Puchalski *et al.*, 2018) (http://glioblastoma.alleninstitute.org/)(Puchalski *et al.*, 2018). Single-cell data were normalized via the Seurat "LogNormalize" function (Butler *et al.*, 2018).

Selection of training data

For the classification in Figure 1 the average expression level of commonly used marker-genes for PBMCs (Table S3) was used to score cells. Cells with scores within the top 2% were used for training, the remaining data were used for testing. The classifications were compared to the cell-type labels provided by 10X Genomics. For the projections we utilized the cell-type labels provided by the atlas or publication from where the data were derived.

Random-forest algorithm for gene selection

ELSA uses the RandomForestClassifier package in Sklearn (Klikauer, 2016) to optimize feature selection for classification. ELSA bootstrap resamples the input training data, choosing samples uniformly and at random with replacement. For each bootstrap iteration and each gene, the marginal out-of-bag (OOB) error is computed from a decision tree that is fit to 2/3 of the bootstrapped data. The standardized OOB errors, taken over a forest of trees, are then used to rank genes. In particular, the algorithm is as follows:

Input: training data D and number of trees C (default C=50)

For each gene i do

For c = 1 to C:

1. Draw a bootstrap sample Z^c from D, uniformly at random with replacement. By default the size of Z^c is set to $\frac{2}{3}$ that of D, chosen based on published heuristics(Ruppert, 2004).
2. Obtain a classification tree(Breiman and Ruppert, 1984) T_c from the bootstrapped sample Z^c using Gini impurity as a metric for determining data splits.
3. Classify the remaining data ($\frac{1}{3}$ by default) using T_c and calculate the OOB error-rate e_c, i.e. the average error over all trees $T_1, \dots, T_c$, when classifying the remaining $\frac{1}{3}$ data (complement of Z^c).
4. For each gene i, permute its expression levels across cells in the training data D and recompute the OOB error (E_c). Set each gene's importance to be the increase in OOB error ($d_c = E_c - e_c$).

End for

Aggregate total OOB error rate from all trees and calculate its variance.

$\hat{d} = \frac{1}{C}\sum_{c=1}^{C} d_c$ and $S_d^2 = \frac{1}{C-1}\sum_{c=1}^{C}(d_c - \hat{d})^2$

Calculate importance of variable *i*: $v_i = \frac{\hat{d}}{\mathrm{S_d}}$ and rank variables by their importance.

Boosting algorithm for cell classification

RUSBoost combines under sampling with boosting, providing an efficient method for improving classification performance when trained on imbalanced data (Seiffert *et al.*, 2009). The training algorithm proceeds as follows:

1. Input:
    a. Training data $\mathrm{D} = (x_1, y_1), \ldots, (x_N, y_N)$. The x_i are cells, y_i are cell-type labels.
    b. M, the maximum number of estimators at which boosting is terminated (50 by default).
    c. A weak learner, by default a decision stump (a 1-level decision tree) is used.
    d. A sampling strategy, random down sampling without replacement was used for all examples, as described here.
2. Output: $H(x)$, the final strong learner.
3. Initialize weights $w_i = \frac{1}{N}\ for\ all\ i\ \in \{1, \ldots, N\}$

For $m = 1$ to M do

a. Randomly down-sample each class (x_i, y) to the size of the smallest class, $min_j|(x, y_j)|$. By default, this is done without replacement.
b. Invoke the weak learner on the down-sampled training data D_m, to obtain the classifier $h_m: X \times Y \rightarrow [0, 1]$
c. Calculate the pseudo-loss over D:
$\epsilon_m \leftarrow \sum_{(i,y):y_i \neq y} w_i(1 - h_m(x_i, y_i) + h_m(x_i, y))$
d. Calculate the weight update parameter: $\alpha_m \leftarrow \frac{\epsilon_m}{1-\epsilon_m}$
e. Update the weights w_i:

$$w_i \leftarrow w_i \alpha_m^{\frac{1}{2}(1+h_m(x_i,y_i)-h_m(x_i,y:y\neq y_i))}$$

$$w_i \leftarrow \frac{w_i}{\sum_i w_i}$$

End for

Output the final strong learner:

$$H(x) = argmax_{y \in Y} \sum_{m=1}^{M} h_m(x, y) \log \frac{1}{\alpha_m}$$

Cross-validation and testing

10-fold cross-validation was performed using the KFold cross validation routine of the Sklearn package (Klikauer, 2016). Summary metrics were computed via the PyCM package in python (Haghighi *et al.*, 2018).

Supplemental Bibliography

Baron,M. *et al.* (2016) A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.*

Breiman,L. and Ruppert,D. (1984) Classification and Regression Trees 1st ed. New York.

Butler,A. *et al.* (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.

Darmanis,S. *et al.* (2015) A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci.*, **112**, 7285–7290.

Haghighi,S. *et al.* (2018) PyCM: Multiclass confusion matrix library in Python. *J. Open Source Softw.*

Klikauer,T. (2016) Scikit-learn: Machine Learning in Python. *TripleC*.

Lake,B.B. *et al.* (2017) A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. *Sci. Rep.*

Nowakowski,T.J. *et al.* (2017) Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science (80-.).*, **358**, 1318–1323.

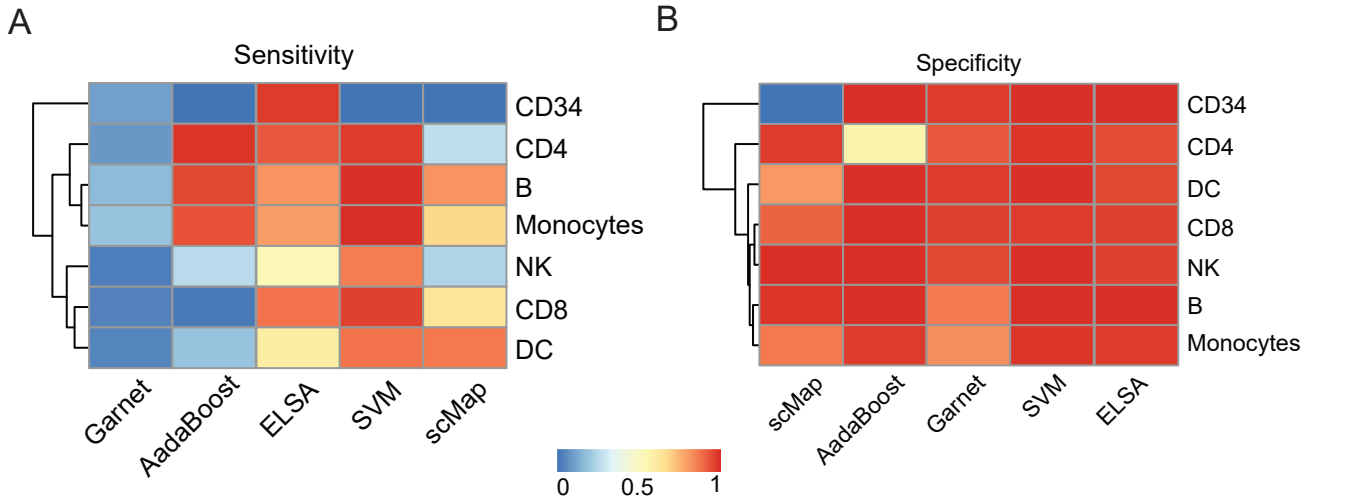
Puchalski,R.B. *et al.* (2018) An anatomic transcriptional atlas of human glioblastoma. *Science (80-.).*

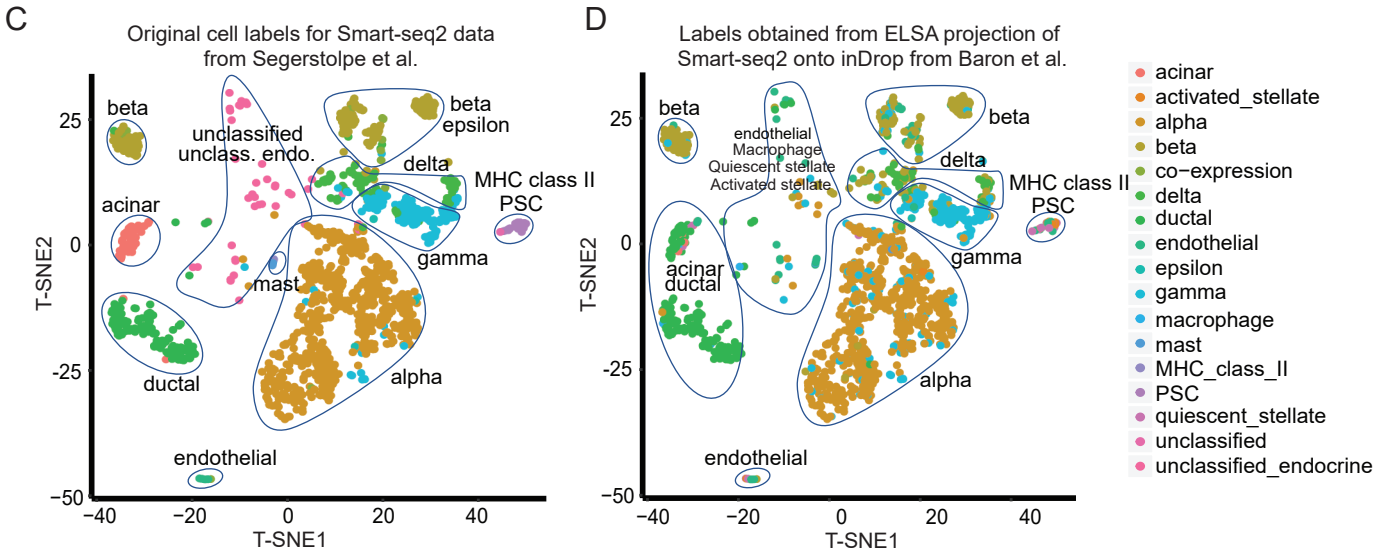
Ruppert,D. (2004) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *J. Am. Stat. Assoc.*

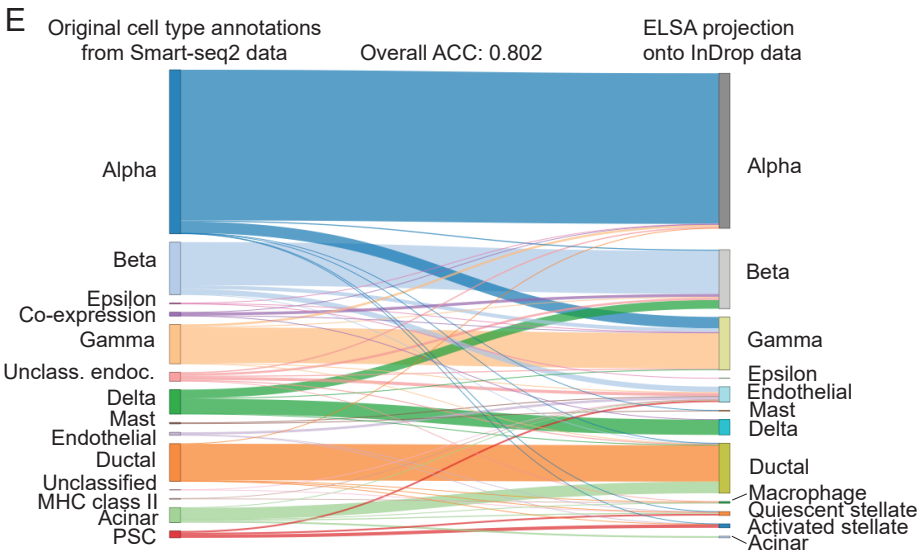
Segerstolpe,Å. *et al.* (2016) Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab.*

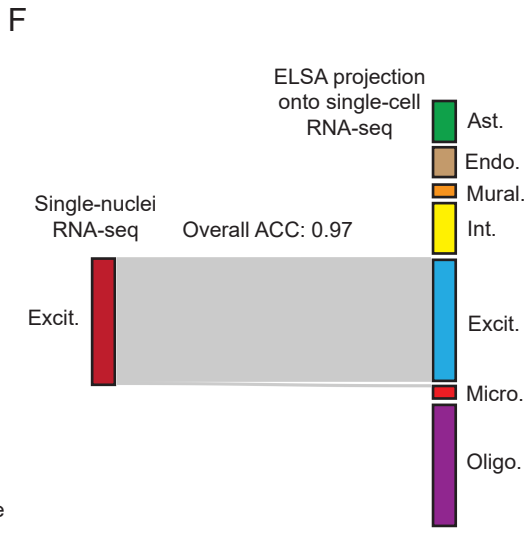
Seiffert,C. *et al.* (2009) RUSBoost: Improving classification performance when training data is skewed.

Figure S1









G

Performance on rare cell types

Reference: Baron et al. Projection: Segerstolpe et al.			
Tools	Specificity	Sensitivity	AUC
scmap-cell	0.99	0.25	0.62
ELSA	0.99	0.49	0.73
Garnett	0.96	0.06	0.49

Figure S1: A-B) Sensitivities and specificities by cell type, computed on the PBMC dataset via 10-fold cross-validation. C) T-SNE plot of Smart-seq2 scRNA-seq data from Segerstolpe et al. D) A T-SNE plot of data from Segerstolpe et al., projected onto InDrop-based scRNA-seq from Baron et al. E) A visualization of the projection of the cell-type labels from Segerstolpe et al. to the data from Baron et al. F) ELSA projection of snRNA-seq data onto scRNA-seq, both from human cortex. G) A comparison of classifier accuracy between ELSA, Garnett, and scmap, based on the Smart-seq2 and InDrop data in C).

Figure S2

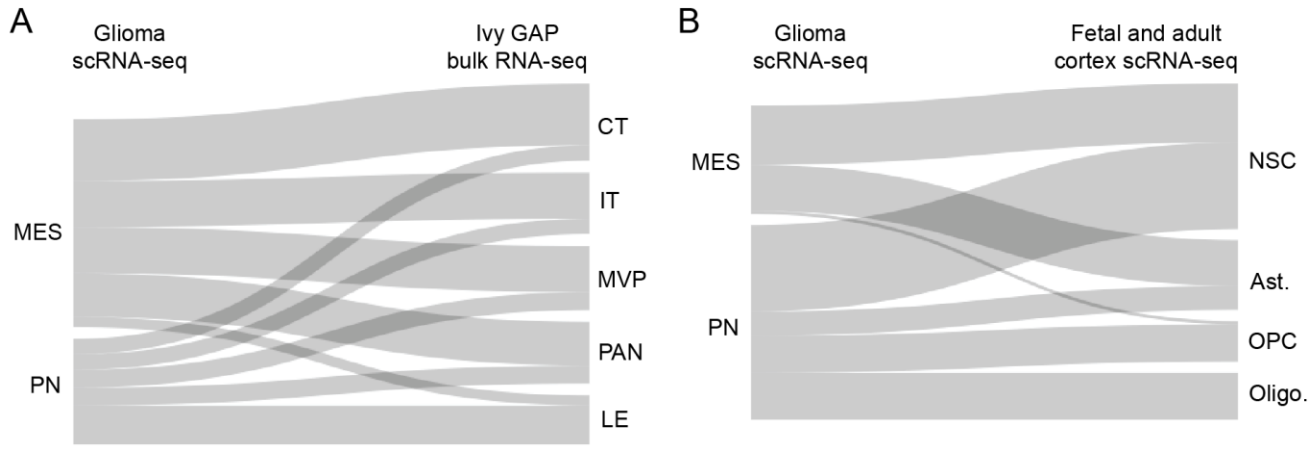


Figure S2: ELSA projections from novel scRNA-seq of human glioma to bulk RNA-seq from the Ivy Glioblastoma Atlas Project A), and from glioma scRNA-seq to scRNA-seq of human fetal and adult cortex.

Table S1. Datasets used for classifications and projections

Dataset type	**Dataset**	**Organism**	**Tissue**	**Experimental platform**
Single cell	Novel	Human	Glioblastoma	10X Genomics Chromium
Single cell	Baron	Human	Pancreas	inDrop
Single cell	Segerstolpe	Human	Pancreas	Samrt-Seq2
Single cell	10X Genomics	Human	Peripheral blood	10X Genomics Chromium
Single cell	Darmanis	Human	Brain	SMARTer
Single cell	Nowakowski	Human	Cortex	SMARTer
Single cell	Lake	Human	cortex	Samrt-Seq
Single nuc	Lake	Human	cortex	Samrt-Seq
Bulk RNA-Seq	IvyGAP	Human	Glioblastoma	Illumina HiSeq

Table S2. Specimens of novel Glioma used in the study

ID	Diagnosis	Age	Gender	IDH status	Clonal CNVs	Platform	Cell number
SF11956	Primary GBM	63	M	Wildtype	chr7p+, chr10-, chr19q-	10x	3,923
SF11977	Primary GBM	61	F	Wildtype	chr5+, chr7p+, chr9+, chr19q-	10x	705
SF11644	Primary GBM	57	M	Wildtype	chr13-	10x	1,330
SF11979	Primary GBM	76	F	Wildtype	chr7p+, chr10-, chr19q-	10x	3,331
SF12199	Primary astrocytoma	74	M	Mutant	chr7+, chr10-,1 chr15-, chr19+	10x	1834
SF111036/ SF12017	Primary astrocytoma	34/44	M/M	Mutant	chr1p-, chr7q+cd chr9-, chr17+, chr19q+	10x	5,368
S10432	Primary GBM	50	F	Wildtype	chr7p+, chr 10-	10x	1,463
SF4324	Recurrent GBM	57	F	Wildtype	chr7p+, chr 10-	10x	69,159

Table S3. The used marker-genes for PBMCs classification

Cell types	Cell markers
NK cells	NCAM1, FCGR3A
Monocytes	CD14, FCGR1A, CD68
B cells	CD19, MS4A1, CD79A
CD4 T cells	CD4, IL2RA, IL7R
CD8 T cells	CD8A, CD8B
Dendritic cells	IL3RA, CD1C, BATF3, CD209
CD34+	CD34, ENG