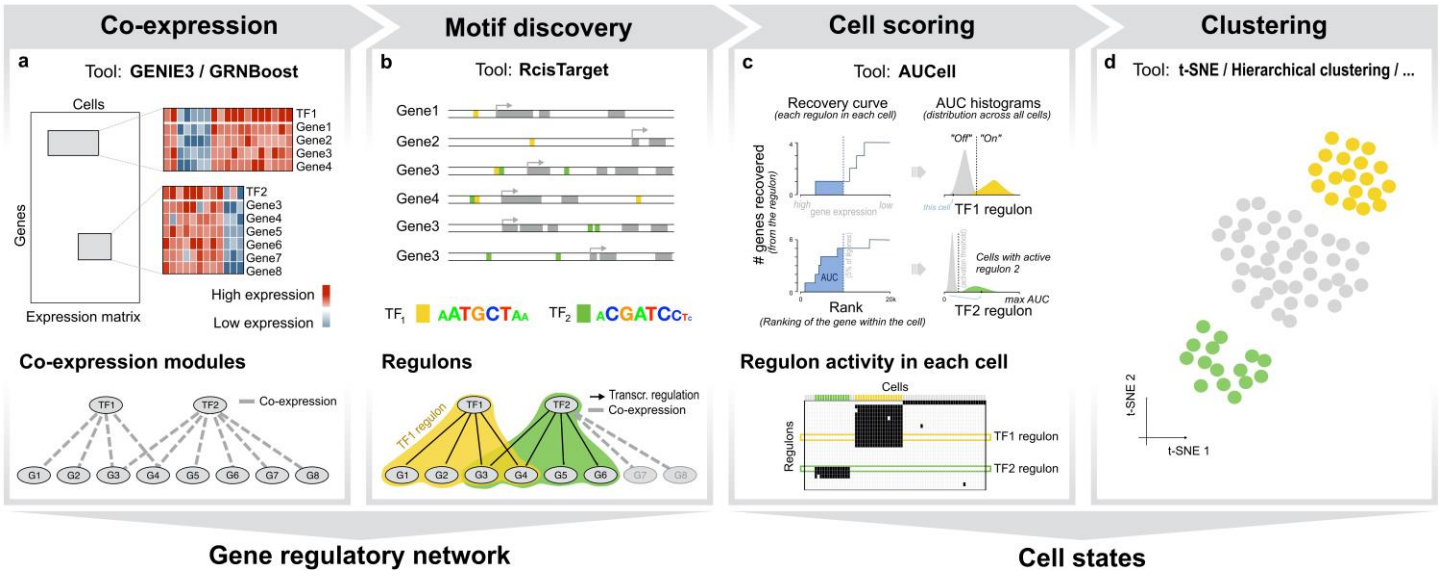
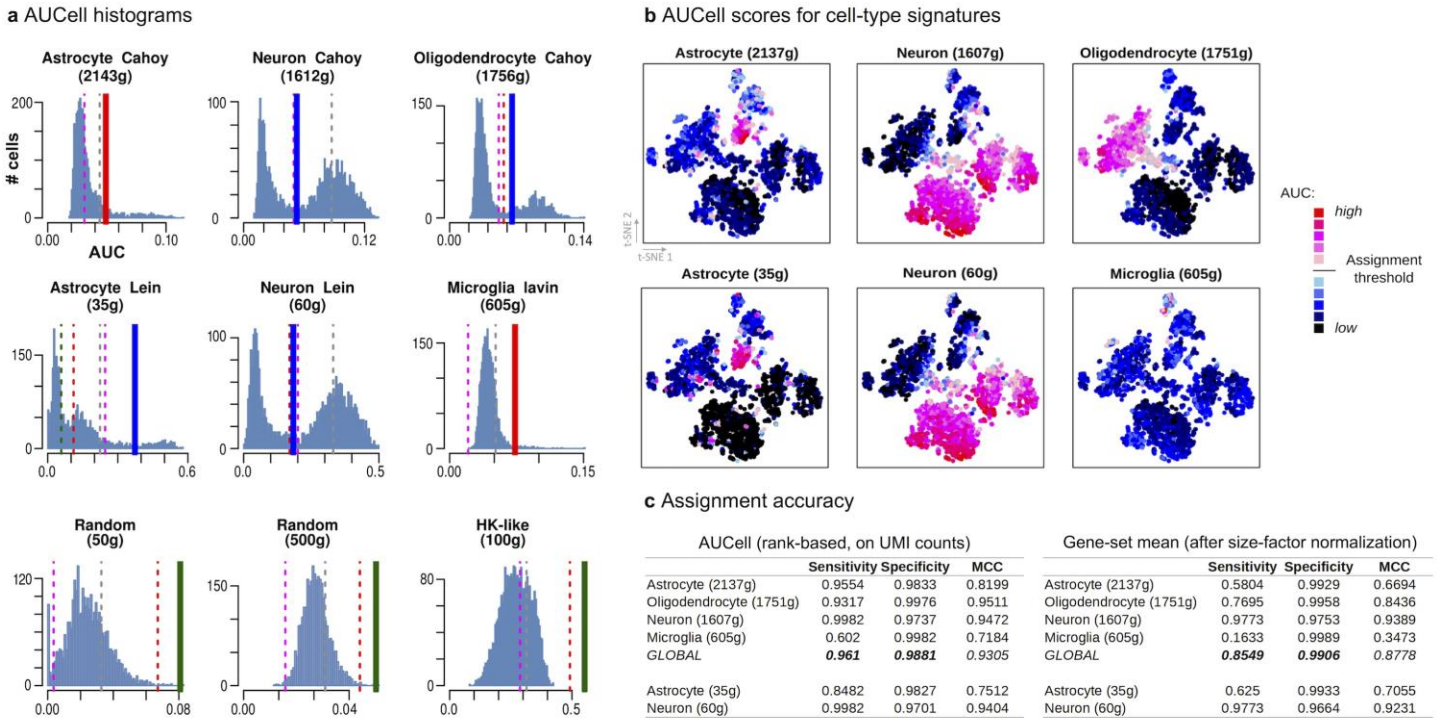


Supplementary Figure 1

The SCENIC workflow.

(**a**) In the first step, co-expression modules between transcription factors and candidate target genes are inferred with GENIE3 (Random Forest) or GRNBoost (Gradient Boosting). Each module consists of a transcription factor together with its predicted targets, purely based on co-expression. (**b**) In the second step, each co-expressed module is analyzed with RcisTarget to identify enriched motifs; only modules and targets for which the motif of the TF is enriched are retained. Each TF together with its potential direct targets is a regulon. (**c**) In the third step, the activity of each regulon in each cell is evaluated using AUCell, which calculates the Area Under the recovery Curve, integrating the expression ranks across all genes in a regulon. The AUCell scores are used to generate the Regulon Activity Matrix. This matrix can be binarized by setting an AUC threshold for each regulon, which will determine in which cells the regulon is “on”. (**d**) The Regulon Activity Matrix can be used to cluster the cells (e.g. t-SNE) and, thereby, identify cell types and states based on the shared activity of a regulatory subnetwork.



c Assignment accuracy

AUCell (rank-based, on UMI counts)

	Sensitivity	Specificity	MCC
Astrocyte (2137g)	0.9554	0.9833	0.8199
Oligodendrocyte (1751g)	0.9317	0.9976	0.9511
Neuron (1607g)	0.9982	0.9737	0.9472
Microglia (605g)	0.602	0.9982	0.7184
GLOBAL	***0.961***	***0.9881***	*0.9305*
Astrocyte (35g)	0.8482	0.9827	0.7512
Neuron (60g)	0.9982	0.9701	0.9404

Gene-set mean (after size-factor normalization)

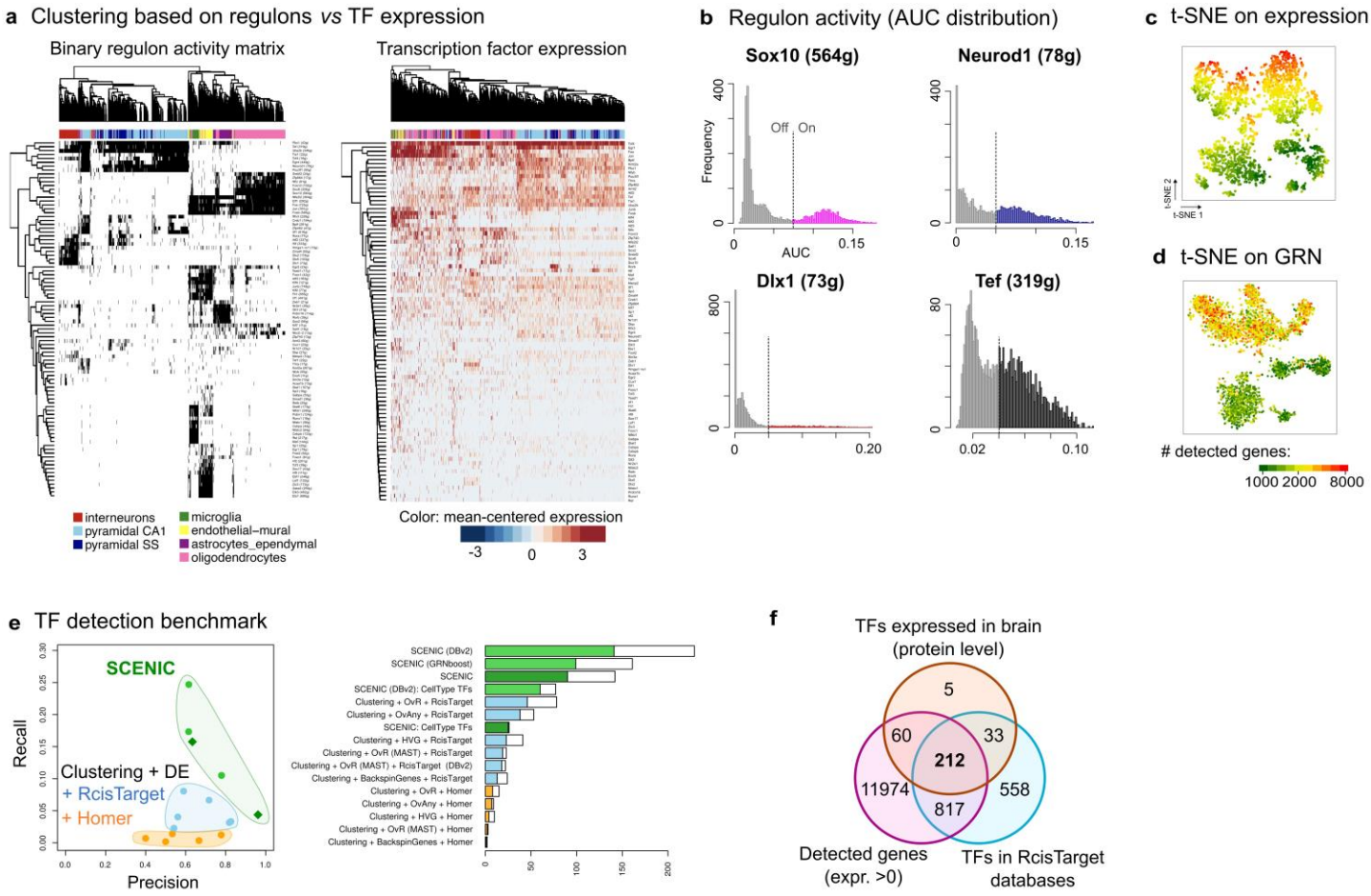
	Sensitivity	Specificity	MCC
Astrocyte (2137g)	0.5804	0.9929	0.6694
Oligodendrocyte (1751g)	0.7695	0.9958	0.8436
Neuron (1607g)	0.9773	0.9753	0.9389
Microglia (605g)	0.1633	0.9989	0.3473
GLOBAL	***0.8549***	***0.9906***	*0.8778*
Astrocyte (35g)	0.625	0.9933	0.7055
Neuron (60g)	0.9773	0.9664	0.9231

Supplementary Figure 2

AUCell applied to gene signatures of known cell types

(**a**) AUC distributions for multiple gene-sets scored on the mouse brain data set with AUCell. The AUC represents the activity of the regulon or gene signature in each of the cells. The selection of cells with the regulon "active" is based on the distribution of the AUC across all the cells in the dataset. The ideal situation of a regulon or gene signature being active in only a subset of the cells would return a bimodal distribution (e.g. neurons or oligodendrocytes), or a distribution with a long tail (e.g. microglia). On the contrary, normal-like distributions are more likely to occur from non-differentially expressed gene sets. This situation is illustrated here through random gene sets (e.g. gene names taken randomly from the dataset) and housekeeping-like gene sets (genes detected in most cells). AUCell automatically explores the distributions of AUC scores and calculates several possible thresholds for each gene-set: (1) Inflection point of the density curve, which is usually a good option for the ideal situation with bimodal distributions (blue), and (2) Outliers of the global distribution (grey/green) sub-distributions (adjusting a mixture of two or three normal distributions, red or pink). The thresholds associated to these distributions are plotted in dashed lines over the histograms; the selected threshold for each gene-set is highlighted with a thicker continuous line. Note that the threshold selection in the current version is not exhaustive, and we highly recommend checking the AUC histograms and manually adjusting the threshold if needed. We also recommend being cautious about gene-sets with few genes (10-15) and thresholds that are set extremely low. (**b**) Expression-based t-SNEs (mouse brain dataset by Zeisel et al.) colored according to the AUC of each cell for the given gene-set. Shades of pink/red are used when the cell AUC is greater than the assignment threshold, in shades of blue otherwise. (**c**) Sensitivity and specificity calculated using the cell type provided in Zeisel et al. as correct labels, and the automatic AUCell assignment thresholds. Using the AUC (left) and the mean of the gene-set expression after normalization with Scran[26] (right).

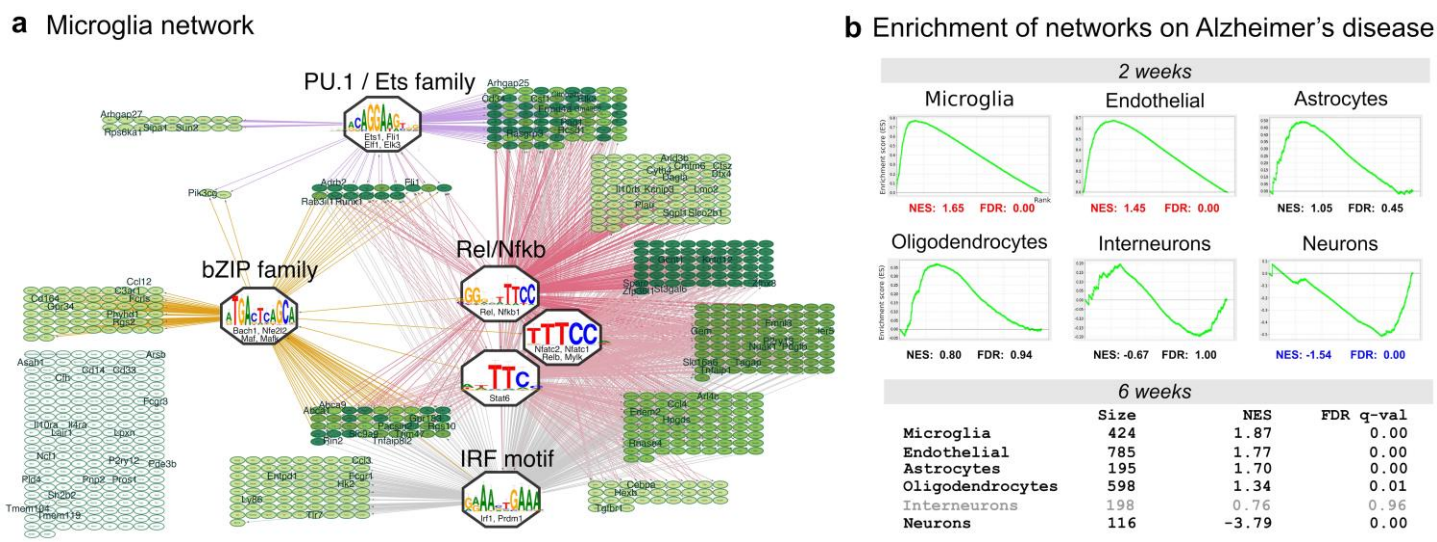
Source of the gene sets: Cahoy et al. [45] (gene signatures with more than 1000 genes, top row), Lein et al. [46] (gene signatures with less than 100 genes for astrocytes and neurons), and Lavin et al. [47] (microglia: microglia versus other tissue-resident macrophages).



Supplementary Figure 3

Validation of the regulon-centric approach.

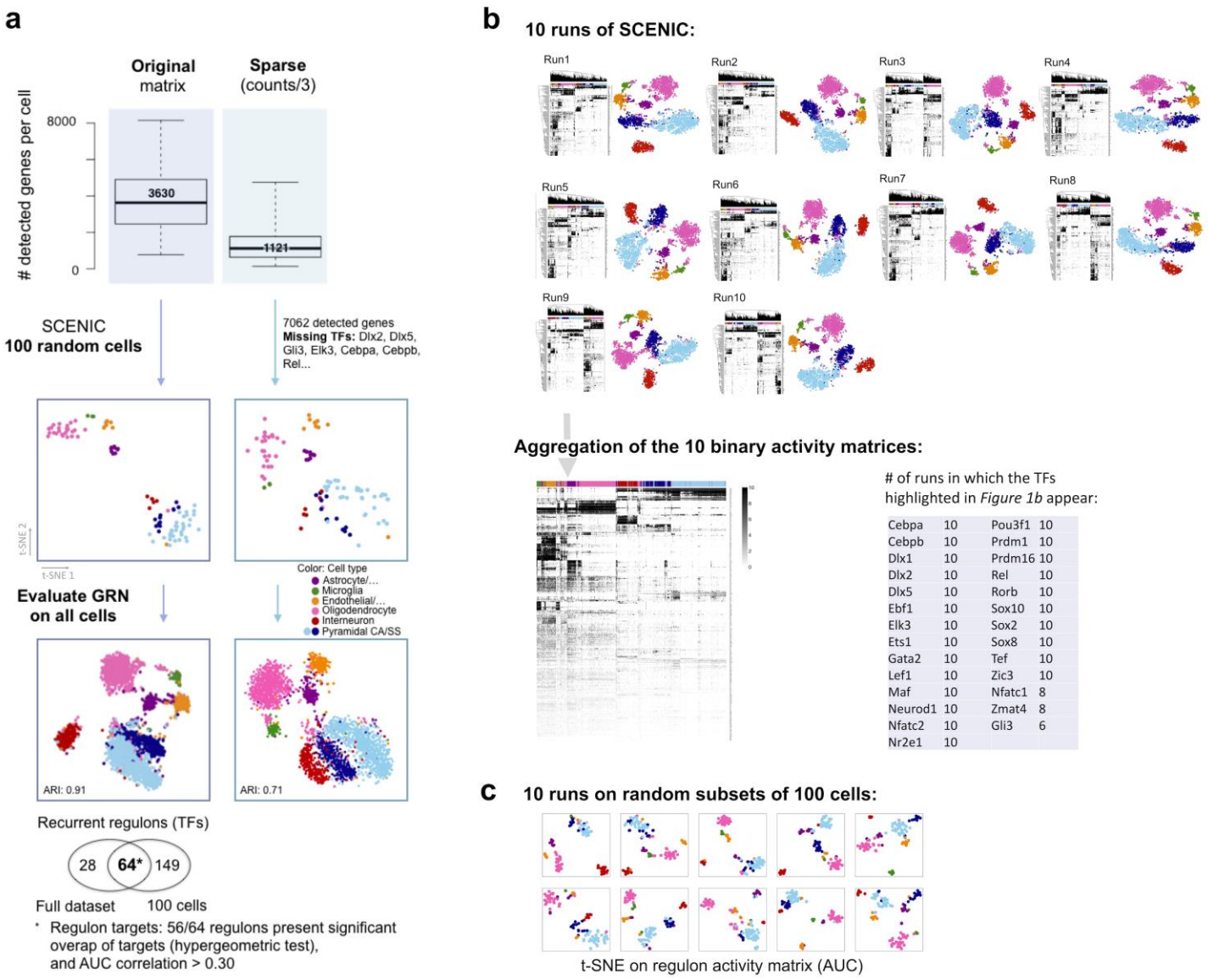
(**a**) Comparison of cell clustering resulting from the SCENIC regulon activity and from the TF expression alone**.** Left: Clustered binary regulon activity matrix. Right: Clustering based on the normalized expression of the 92 TFs (within-cluster size factor normalization with scran, heatmap color: median centered by gene). (**b**) AUC histograms for a few key regulons. The AUC allows to split the populations of cells with high versus low activity of a regulon. (**c**) t-SNE on the expression matrix (same input as to SCENIC: UMI counts with no further normalization) and (**d**) t-SNE on the binary regulon activity matrix. Both t-SNEs are PCA-based and colored according to the number of genes detected (expression over 0) in each cell. The clustering based on SCENIC effectively corrects for the intra-cluster bias, while the true biological difference between neurons (more genes expressed) and glia (less genes expressed) is unaffected [12,48]. (**e**) Comparison of SCENIC with alternative approaches for identification of cell-type associated TFs (see Methods for details). The bar plot on the right shows the number of TFs identified by each method (white) and the number of TFs in the validation set (colored). SCENIC retains more transcription factors compared to a differential expression analysis. (**f**) Proportion of TFs that can be detected by SCENIC. Venn diagram comparing the TFs detected in the mouse brain at protein level by Zhou et al. [49], in the scRNA-seq dataset by Zeisel et al, and the TFs available in RcisTarget databases (i.e. known motif).



Supplementary Figure 4

Microglia gene regulatory network and association of brain networks with Alzheimer's disease.

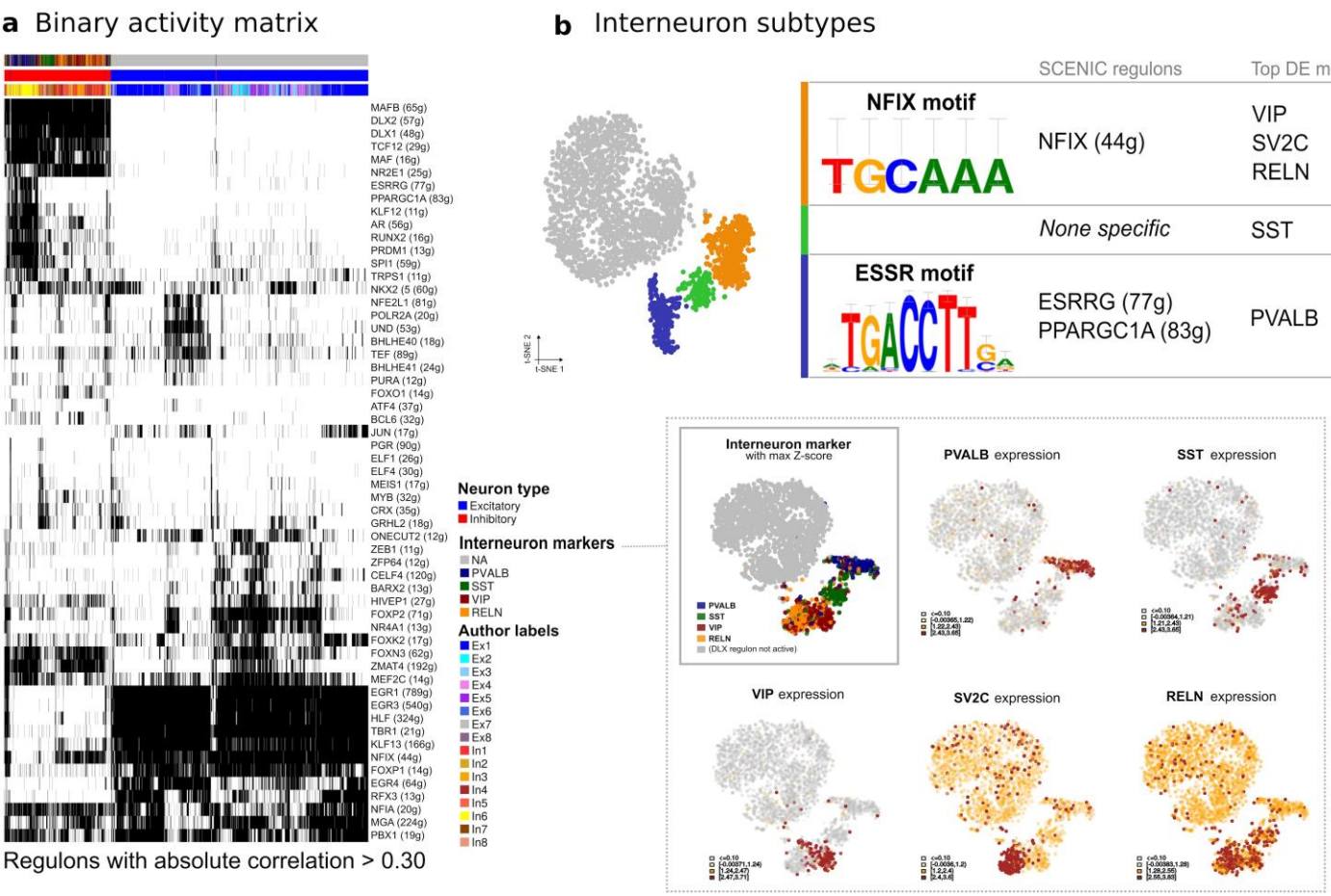
(**a**) The regulons associated to microglia, inferred on the mouse brain data, can be summarized based on the binding motif of the associated TF (network built in iRegulon). The genes that are included in a previously published microglia signature (Lavin et al. [47]) are indicated by a larger font size; the color of the node indicates the number of regulators (lighter: fewer, darker: more). The predicted network for microglia contains many well-known regulators of microglial fate and/or microglial activation, including PU.1, Nfkb, Irf, and AP-1/Maf. (**b**) When we compared the predicted microglial network to previously published gene signatures of microglial "activation" in a mouse Alzheimer's disease model, we found the microglia network to be strongly activated and the neuronal network to be down-regulated during AD progression, indicating that the microglia network captures a relevant regulatory program. The plots shown are results of GSEA analysis of the networks associated to each of the wild type cell types (union of the genes in the regulons) against the gene expression-based ranking in a mouse model for Alzheimer's disease (AD). Dataset by Gjoneska et al. [50]: transcriptional changes in hippocampus of CK-p25 mouse models of AD compared to CK littermate controls, 2 and 6 weeks after p25 induction.



Supplementary Figure 5

SCENIC is robust to down-sampling of cells and sparse expression matrices.

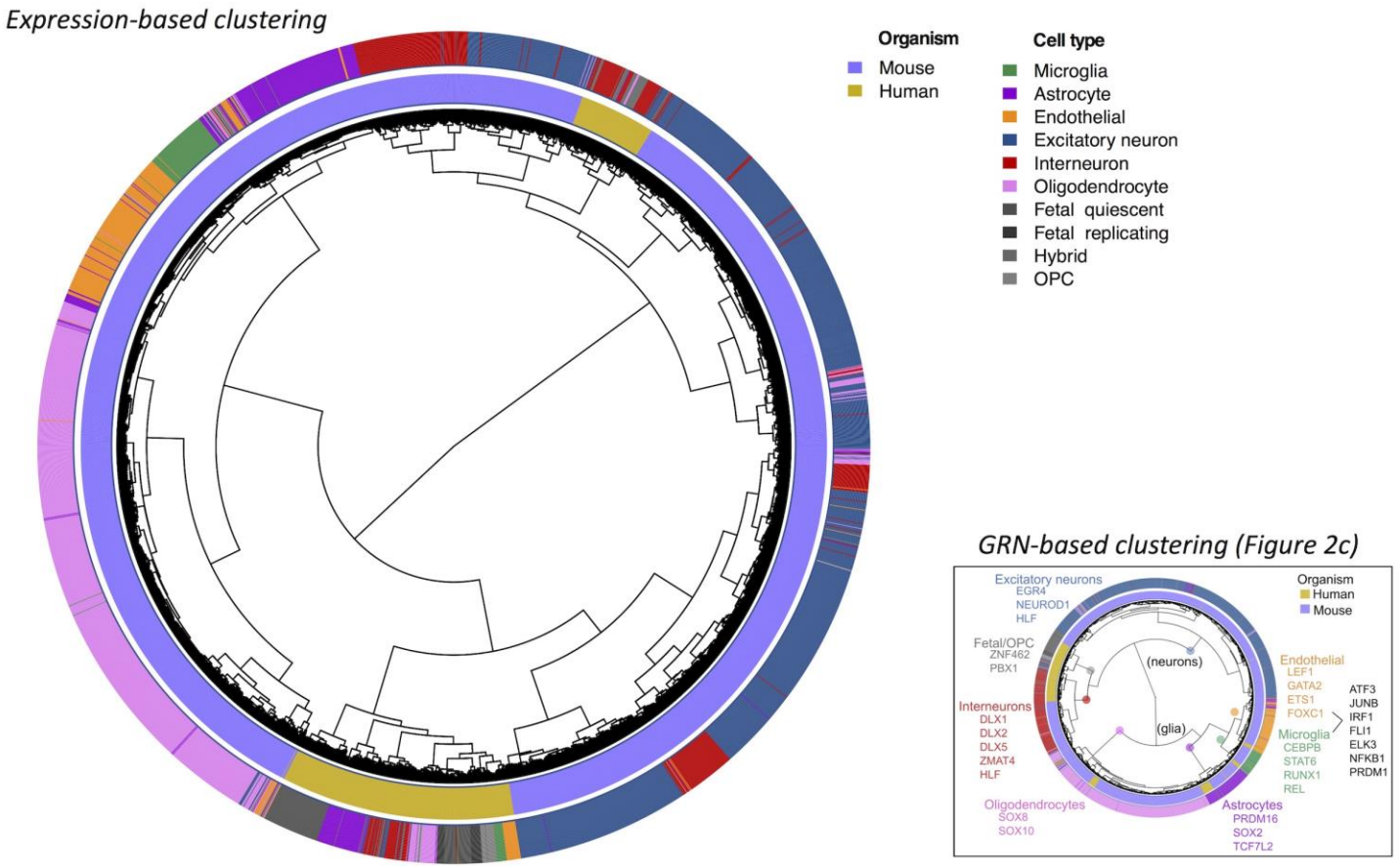
(**a**) SCENIC run on 100 random cells from the mouse brain dataset (Zeisel et al.) provides similar results to the run on the whole dataset (left column: cell states matching the cell types provided in the publication, similar relevant regulons per state, and significant overlap of targets). The GRN inferred with the 100 cells is then evaluated with AUCell on all cells to confirm that the network is generalizable to cells not included in the GRN inference. On the right column, same approach but simulating a sparse dataset (UMI count matrix divided by three and truncated, resulting on a median of 1121 detected genes). Many relevant TFs are not detected in the sparse dataset (so the associated regulons will be missed) but SCENIC is still able to identify the main cell types. (**b**) Evaluation of the stability of the results from SCENIC with 10 runs of SCENIC on the mouse brain dataset (Zeisel et al.). Top: Binary regulon activity matrices and t-SNEs (colored by the author's cell-type labels). The aggregation of the 10 binary matrices illustrates the stability of the results across runs, and the large majority of top regulators are found in 10/10 runs. (**c**) t-SNE on the AUC regulon activity resulting from running SCENIC on 10 random subsets of 100 cells.



Supplementary Figure 6

SCENIC results for the human brain single-nuclei data set (Lake et al., 2016).

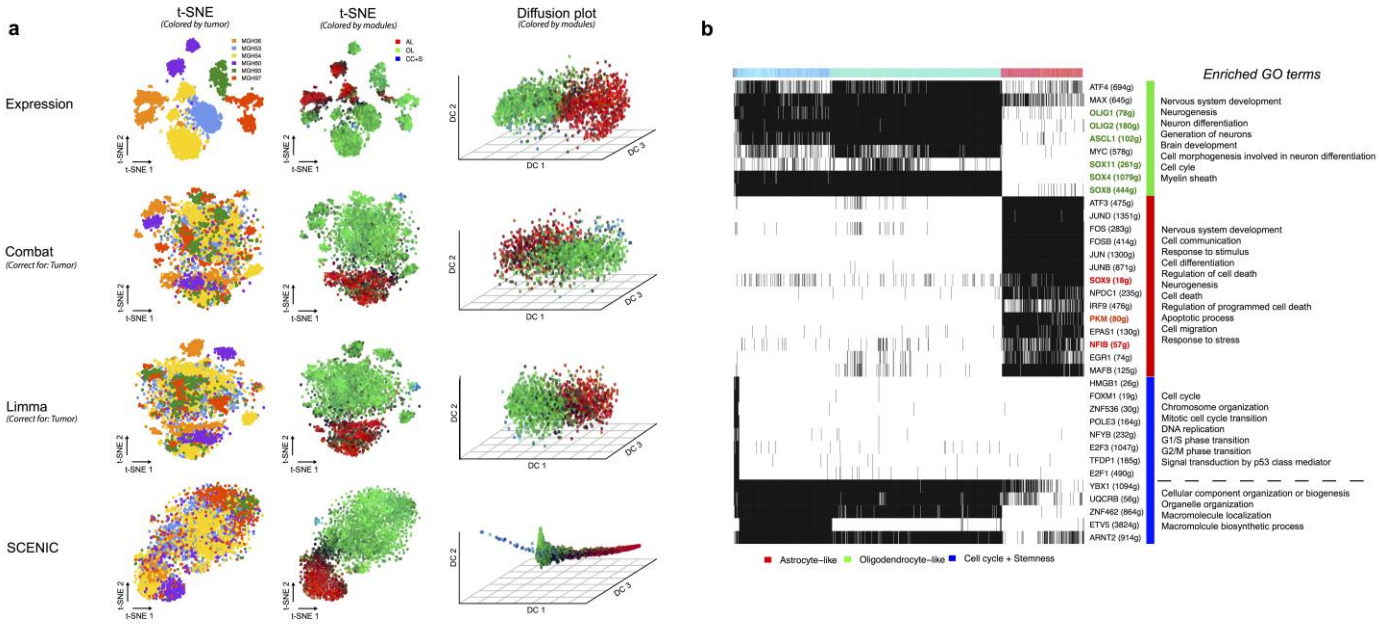
(**a**) Binary regulon activity matrix. (**b**) Three GRN-based subpopulations of human interneurons, with the main DNA motifs and transcription factors defining these groups (NFIX for the VIP interneurons, and ESRRG for the PV interneurons), and the top known markers identified in each (note that the TFs themselves are also up-regulated in the respective clusters). Bottom box: Expression of markers for interneuron subtypes [51,52]. In the first plot, interneurons are colored according to the marker with highest Z-score. The remaining plots are colored based on gene expression (grey: no expression, dark red: high expression, yellow: intermediate).



Supplementary Figure 7

Expression-based clustering of mouse and human brain cells.

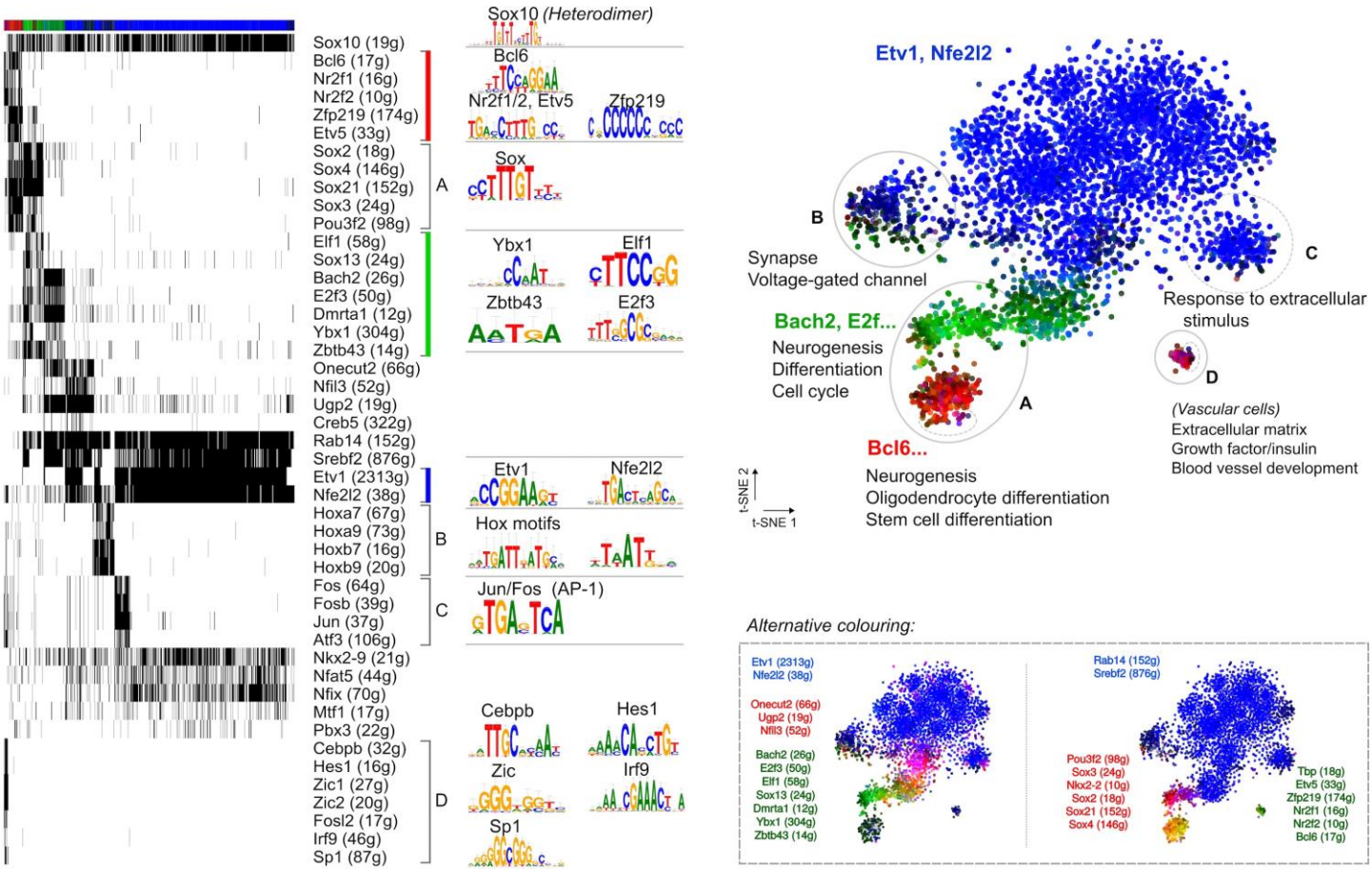
Hierarchical clustering based on the merged expression matrix, Z-score normalized, of human and mouse brain cells. Clustering groups cells by species, then by cell type. The thumbnail shows Figure 2c, for comparison, where SCENIC yields a primary clustering of the cell types.



Supplementary Figure 8

SCENIC overcomes tumor batch effect and recovers relevant cell types and GRNs in oligodendroglioma.

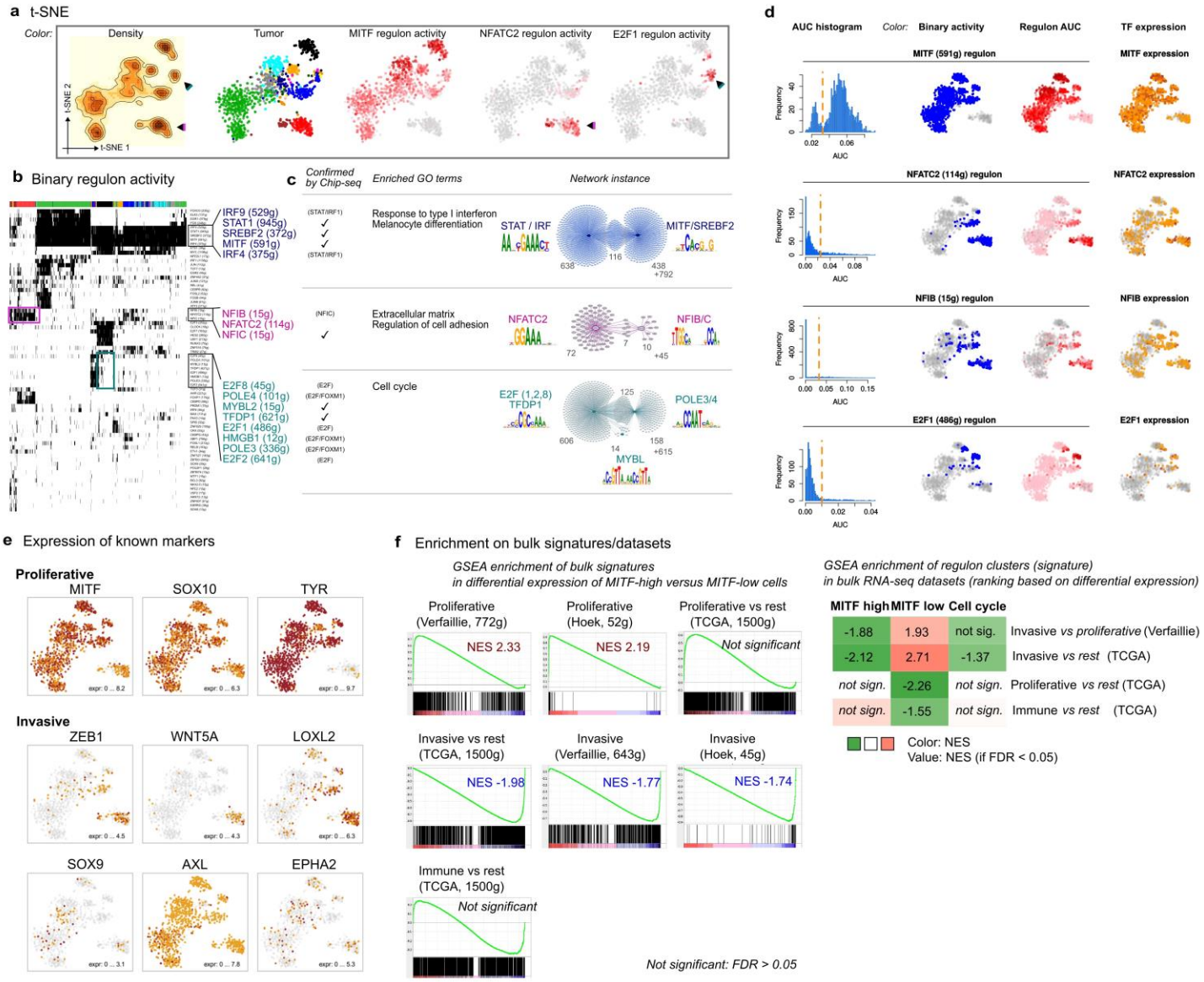
(**a**) Comparison of batch-effect removal methods on the oligodendroglioma dataset. t-SNEs and diffusion plots on the raw expression matrix (first row), after correcting by tumor of origin with Combat or Limma (rows 2-3), or on the binary activity matrix from SCENIC (row 4). The cells are colored based on the tumor of origin or GRN activity (red: astrocyte-like regulons, green: oligodendrocyte-like regulons, blue: regulons related to cell cycle or stemness). (**b**) Simplified binary regulon activity matrix (output of SCENIC) for the oligodendroglioma dataset. Highlighted regulons (colored TF names) are known to be characteristic in oligodendrocytes or astrocytes, respectively.



Supplementary Figure 9

Oligodendrocyte differentiation is driven by discrete changes in gene regulatory networks.

Binary activity matrix highlighting transcription factors groups and their motifs. In the resulting t-SNE, cells are colored based on the average binary activity of three selected groups of regulons (red, green, blue), which correspond to the three main states in the differentiation trajectory: OPC network, driven by Bcl6 and co-regulatory factors; intermediate network driven by Bach2 and other factors; and mature oligodendrocyte network with many transcription factors including Etv1 and Nfel2l2. Sox10 is found as regulator of all subtypes of oligodendrocytes. Within the set of mature oligodendrocytes (blue cells), two outgroups are detected that fall slightly outside the differentiation trajectory: oligodendrocytes with neuronal properties (B) and oligodendrocytes with AP-1 activation signatures (C). Next to each cluster in the t-SNE, enriched GO terms are shown. In the box the t-SNEs are colored in an alternative scheme, showing other oligodendrocyte networks and states. Note that in spite to being cells in differentiation, the dominant networks are rather discrete. This suggests that transitions between these main states must occur rapidly, since only few cells were found in transition.



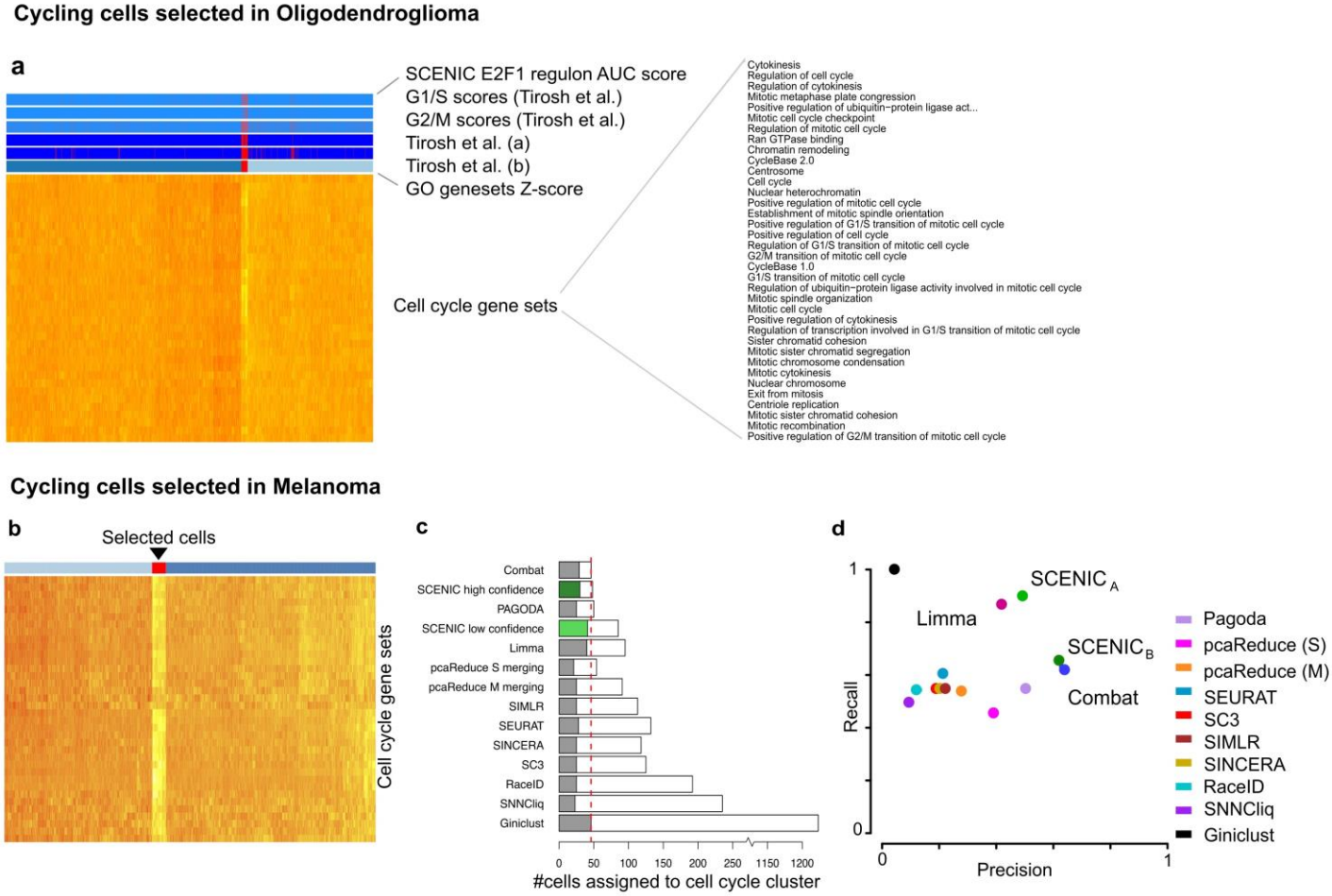
Supplementary Figure 10

SCENIC reveals melanoma heterogeneity.

(**a**) t-SNE on the binary activity matrix after applying SCENIC. (**b**) Binary regulon activity matrix for the melanoma dataset. The color bar above the heatmap indicates the tumor of origin; regulons associated to the cell cycle (green), MITF^{low}, invasive (pink) and MITF^{high}, usually known as proliferative (blue) states are zoomed in. (**c**) Details for the three most dominant networks. "Confirmed by ChIP-seq": a tick indicates that the regulon presents enrichment of targets in a ChIP-seq dataset for the same transcription factor. (**d**) Comparison of TF expression and regulon activity. For four transcription factors: histogram of AUC values, together with the chosen cutoff (orange dashed line). In the second column, the cells with AUC value over the cutoff are shown in blue. These are the cells where the regulon is considered active (i.e. "1" in the binary activity matrix). In the third column, the actual AUC values are used to color the cells. In the fourth column, the expression of the transcription factor itself is shown. The discriminative power of the TFs is much lower than that of the regulons. (**e**) Expression of known melanoma markers. Note that the MITF^{low} cluster shows up-regulated WNT5A, LOXL2 and ZEB1 expression (both known markers of the invasive state [53,54]), and correlates significantly with previously published invasive gene signatures (Figure S19). However, unlike the 'classical' invasive cell state, this MITF^{low} state retains SOX10 expression. **(f)** Comparison of melanoma bulk signatures with single-cell states. Left: The bulk signatures derived from invasive and proliferative melanoma states (e.g., Hoek and Verfaillie) are significantly enriched in the respective up- or down-regulated side of the gene ranking based on the single-cell states. In the

GSEA, the ranking (x-axis) is based on the contrast between MITF-high versus MITF-low states. Right: Similar GSEA analysis, but now only the NES scores are shown. This analysis is the reciprocal to the previous one, whereby the ranking is based on the contrast in bulk samples, and the signatures tested are derived from differential expression between the single-cell MITF-high and MITF-low states.

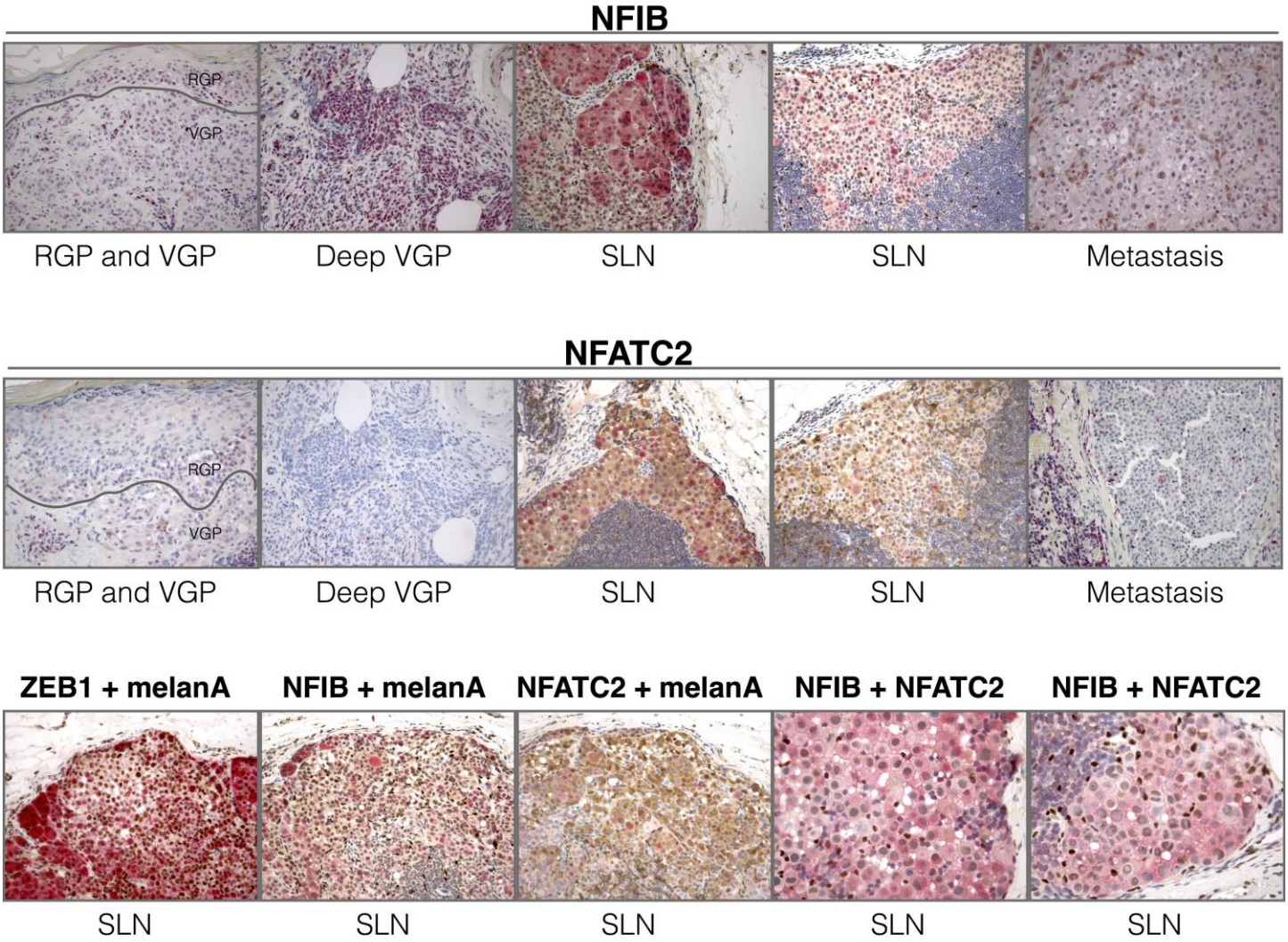
Note that cells in the $MITF^{high}$ state also have high activity of STAT and IRF downstream targets. This is difficult to detect in bulk samples because of the complex mixture of malignant cells with tumor infiltrating lymphocytes (TIL) where STAT and IRF also play an important role [55]. Here, we find that the $MITF^{high}$ cells themselves have higher STAT activity than the $MITF^{low}$ cells (we excluded all benign cells from the analysis, including immune cells). This has important consequences for the interpretation and prediction of resistance to immune therapy, because these cancer cells with high STAT and IRF activity are likely most sensitive to immunotherapy. Indeed, a recent study identified the JAK-STAT-IRF axis as driver for the expression of two major targets in immune therapy: PD-L1 and PD-L2; which results in an inhibition of the anti-tumor immune response on the one hand, but an increased response to anti-PD(L)1 immune therapies on the other [55]. Note also that the MITF-low “invasive” state largely shared by two of the 14 tumor biopsies, were both resected from auxiliary lymph nodes. This state, unlike the in vitro invasive state, which is driven by AP-1 and TEAD factors, features distinct transcription factors, including NFATC2 and NFIB, which we confirmed to be expressed in early metastatic melanoma cells (i.e. in the initial, small tumors in the sentinel lymph node, by immunohistochemistry). Using gene expression analysis after NFATC2 knock down (Supplementary Fig.13), we identified NFATC2 as a transcriptional repressor of the AP-1 and TEAD target genes. Thus, these observations suggest that NFATC2 may act as a transcriptional break that cells need to overcome to switch to a full-blown invasive cell state. NFATC2 is itself a JUN target [56], and may constitute a negative feedback mechanism. A similar repressor function of NFATC2 has been previously observed in breast cancer [57]. We believe that this is the (biological) reason why AP-1 and TEAD are not detected as regulons in this data set. Note that for TEAD there is an additional reason that it cannot be detected as regulon, because our SCENIC run selects TFs co-expressed with their targets, while TEADs are regulated at the protein level.



Supplementary Figure 11

Validation of cycling cells and comparison with other methods.

(**a**) Identification of cell cycle cells based on the Z-score of gene sets related to cell cycle. (**b**) Heatmap showing the Z-score of cell cycle gene-sets on the Oligodendroglioma dataset. The blue/red bars on the top of the heatmap highlight the cells selected as cycling (red) by three approaches: (1) the AUC scores of SCENIC's E2F1 regulon; (2) the G1/S scores according to Tirosh et al. and the G2/M scores according to Tirosh et al; Tirosh et al. approach with a permissive cut-off (cells are classified as cycling if their G1/S and G2/M scores are above twice the mean within the cell population); Tirosh et al. approach with a more restrictive cut-off (cells are classified as cycling if their G1/S and G2/M scores are above four times the mean within the cell population); and (3) the GO genesets based Z-score. (**c,d**) Comparison of the capacity of different methods to identify the cycling cells: (**c**) Number of cells recovered, sorted by True Positive Rate (TPR, also known as sensitivity or recall). Colored bars represent the number of cell cycle cells identified by the method, in white the number of non-cell cycle cells included in the selected cluster (the cluster with the highest number of cell cycle cells). (**d**) Recall vs Precision ($SCENIC_A$: High-confidence cells, $SCENIC_B$: Lower confidence/regulon activity).

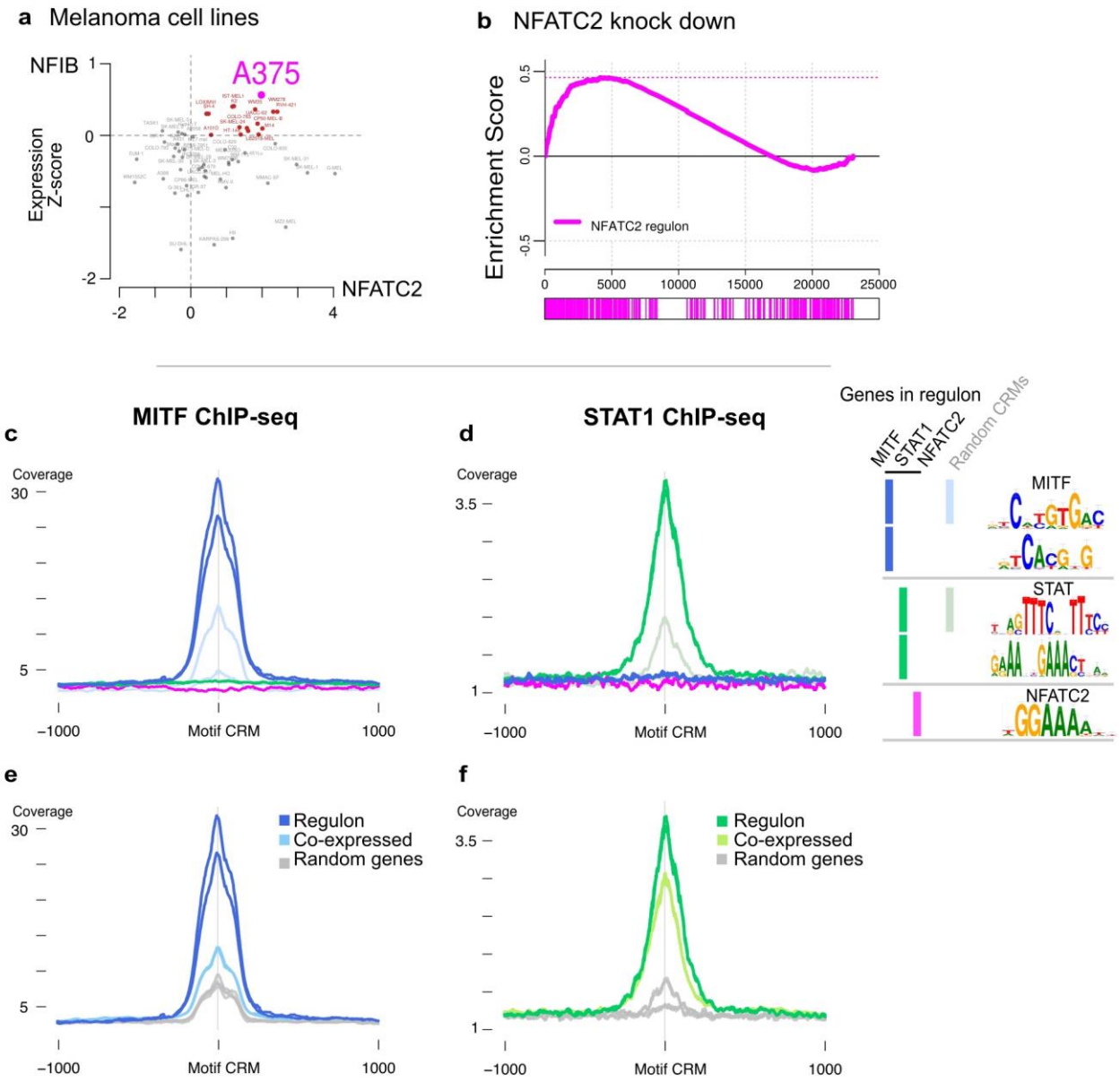


Supplementary Figure 12

Immunohistochemistry of NFATC2, NFIB and ZEB1 on human melanomas.

Complementary to Figure 3i. Here, we show additional biopsies, in different stages of melanoma progression (RGP: radial growth phase, VGP: vertical growth phase, SLN: sentinel lymp node (with small metastases), Metastasis: (full-blown) metastases).

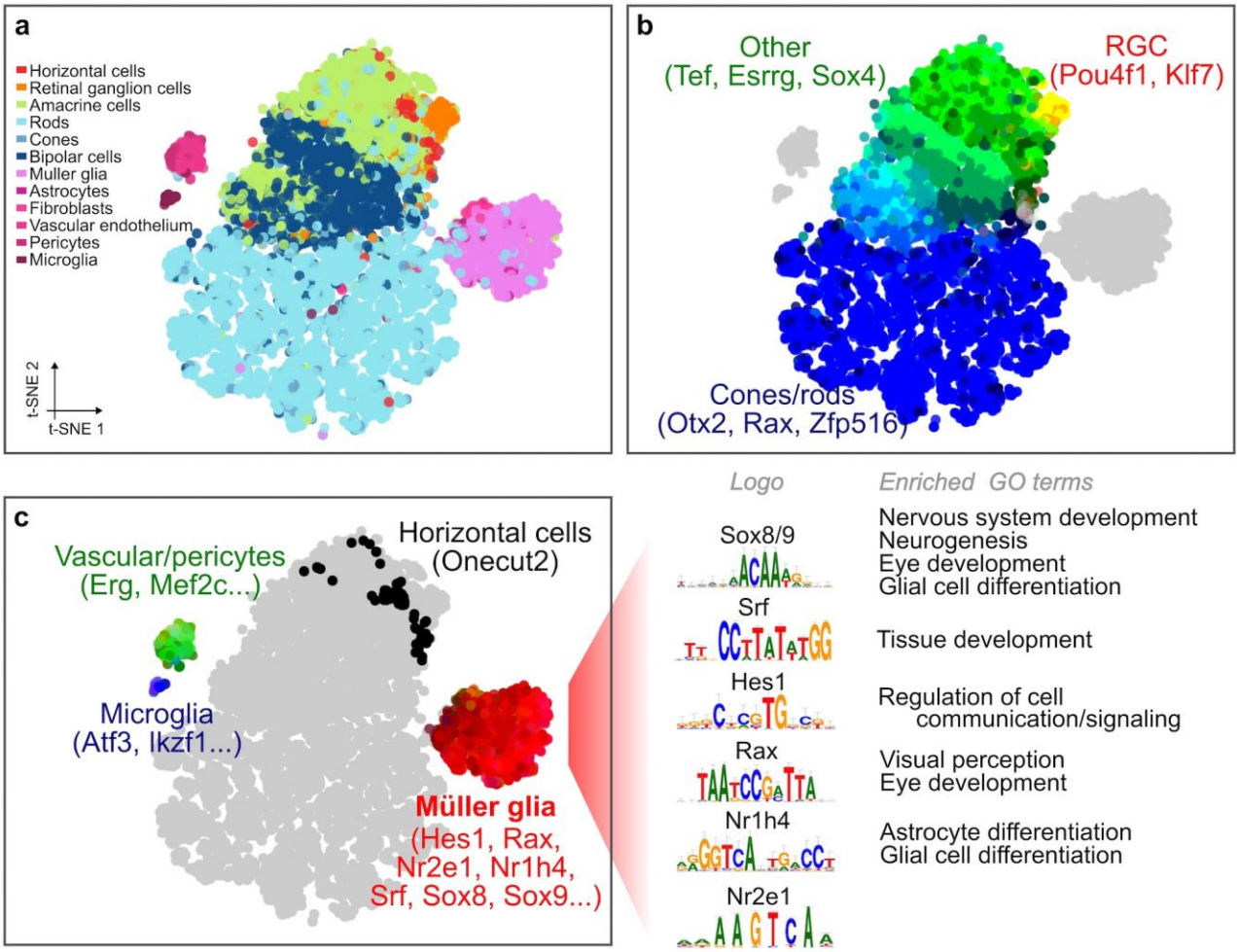
The strongest positive signals for both NFATC2 and NFIB can be seen in the sentinel lymph nodes.



Supplementary Figure 13

Validation of target gene predictions included in the regulons.

(**a**) Z-score normalized expression of NFIB and NFATC2 across melanoma cell lines from COSMIC. A375 was selected for the knock-down based on the expression of key markers which resemble the MITF-low state. (**b**) Knock down of NFATC2 in the A375 melanoma cell line. GSEA plot for genes differentially expressed after NFATC2 knock-down: the predicted NFATC2 targets are significantly up-regulated in the NFATC2 knock-down. **(c-f)** Enrichment of ChIP-seq signal in selected regulons. c-d: Aggregation plots for MITF and STAT1 ChIP-seq signal on the predicted target CRMs of MITF, STAT, and NFATC2 (i.e. regulatory regions for genes in their respective regulons). e-f: Comparison of the ChIP-seq signal on the regulons (predicted CRMs in a window of 10kb around the TSS), with TF motif occurrences in promoter-proximal regions (of randomly selected genes, and of co-expressed genes outside the regulons). The enrichment for the genes in the regulon compared to the control confirms that SCENIC increases the specificity for finding direct targets compared to individual/alternative approaches (e.g. only co-expression or motif analysis).

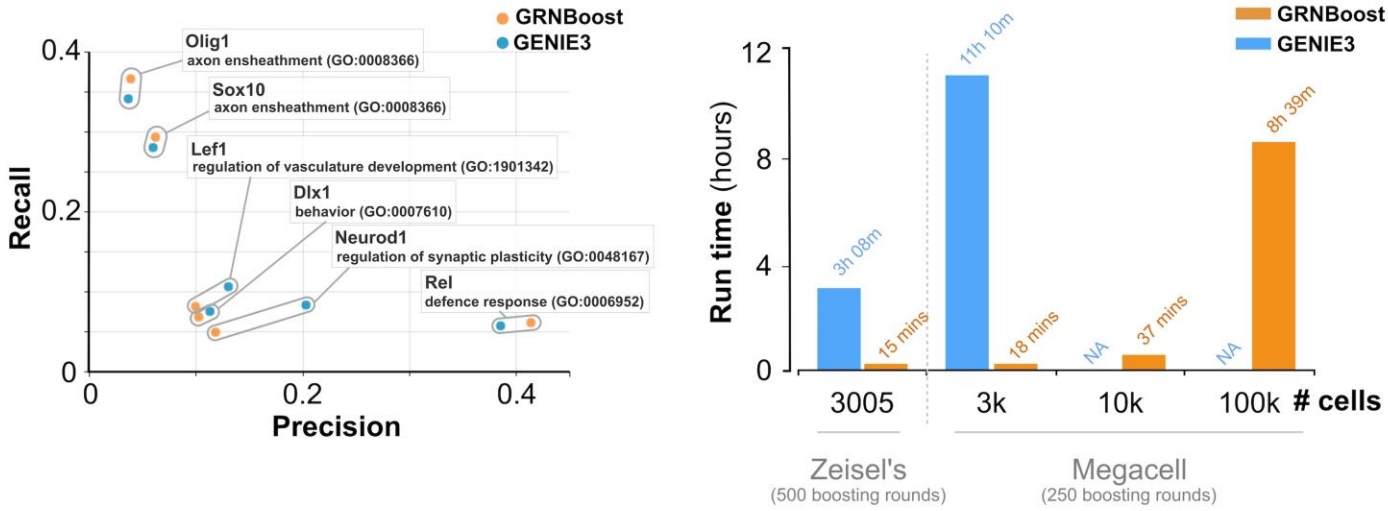


Supplementary Figure 14

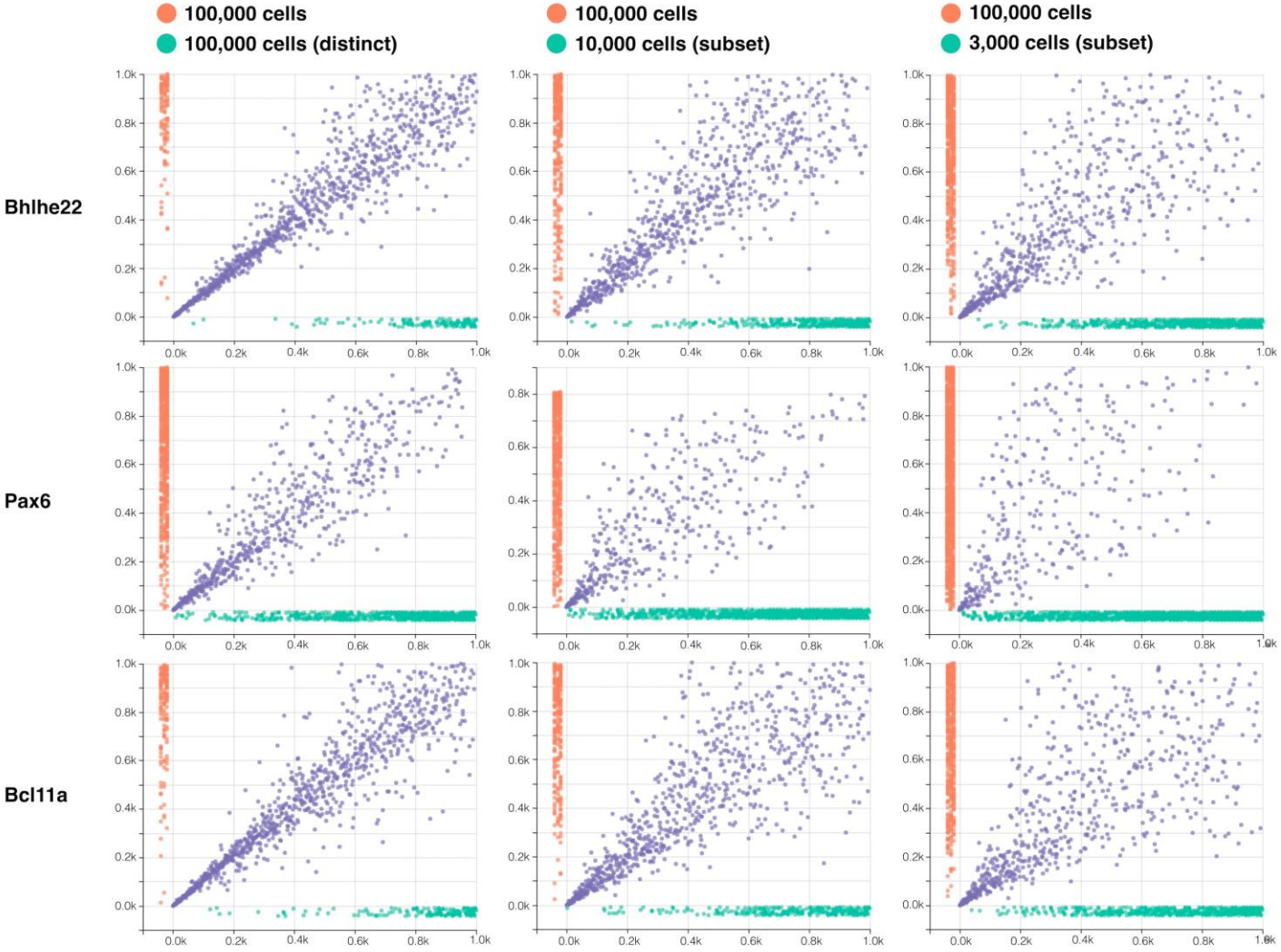
SCENIC analysis of >40000 single cells from the mouse retina.

Drop-seq data on cells from the mouse retina was analyzed with SCENIC by running GENIE3 on a subset of ~11K cells, and evaluating the resulting GRN on all cells. (**a**) tSNE colored according to the expected cell types as annotated by Macosko et al. (**b**, **c**). Main networks according to the activity of regulons, shown based in the red-green-blue coloring scheme. In these tSNEs, the cells shown in grey are not taking into account for the coloring. Logos corresponding to the significant motifs found in iRegulon for the regulons identified in Müller glia and the corresponding GO terms are included. The identified master regulators (such as Sox8/9, Hes1, Rax, Nr1h4, Srf, and Nr2e1 for the Muller glia, which are confirmed in literature [58–61], illustrate that correct networks can be inferred even on sparse data. The sub-sampling approach is especially interesting on sparse datasets, such as the ones resulting from Drop-seq or 10x, because the best quality-cells can be used to infer the GRN, and then this high-quality GRN can be scored on all cells.

a Comparison with GENIE3



b Stability across multiple runs



Nature Methods: doi:10.1038/nmeth.4463

Supplementary Figure 15

GRNBoost benchmark.

(**a**) Comparison of the performance of GRNBoost and GENIE3. **Left**: As a biological validation for GRNBoost, we devised a gene-set-based precision and recall benchmark. We asked wh1ether GRNBoost and GENIE3 predict similar sets of target genes for a selection of transcription factors (Olig1, Sox10, Rel, Lef1, Neurod1, Dlx1) from the Zeisel et al. mouse brain expression data. Using the network inferred by GENIE3, we constructed for each of the 6 TFs its ranked list of target genes. For each ranked list, we used GOrilla [62] to query the top enriched GO term for each TF target list. For each of the 6 top GO terms, we consulted QuickGO [63] to obtain the set of association protein annotation gene symbols (filter by mouse taxon), keeping only the genes present in the expression matrix, to obtain six "master lists" of gene symbols. These master lists were finally used to calculate the precision and recall scores for the sets of target genes predicted by both GENIE3 and GRNBoost for each of the aforementioned TFs. **Right**: To benchmark the time performance of GRNBoost, and as proof of concept, we used the Chromium Megacell demonstration dataset, which contains over 1 million cells from embryonic mouse brain. We took random subsets of 3000, 10000 and 100,000 cells to infer the GRNs. (**b**) Evaluation of stability across multiple runs. Each scatter plot compares the ranking for the targets of each TF across two independent runs (with equal or different numbers of cells).

Supplementary Note 1: Supplementary methods

SCENIC runs on the different datasets

Dataset: Mouse cortex and hippocampus (Zeisel et al.)

The mouse brain dataset, published by Zeisel et al. [9], includes single-cell RNA-seq of 3005 cells from somatosensory cortex and hippocampus (CA1 region) of juvenile mice (21-31 days old). This data set has been used extensively for benchmarking purposes [10,64–70] and contains the main cell types in hippocampus and somatosensory cortex, namely neurons (pyramidal excitatory neurons, and interneurons), glia (astrocytes, oligodendrocytes, microglia), and endothelial cells. Most of the cells were sequenced after dissociation with no specific selection by markers or cell type (wild type CD-1 mice). In addition, the dataset also includes 116 cells selected by FACS from 5HT3a-BACEGFP transgenic mice (likely Htr3a interneurons). The expression matrix was downloaded from GEO (GSE60361). This matrix contains the UMI counts for 19972 genes across the 3005 cells that passed their quality controls (e.g. low quality cells and potential doublets). To run GENIE3, this matrix was filtered to keep the 13063 genes with more than 90 counts (which corresponds to 3 counts in 1% of cells) and detected in more than 30 cells (1% of cells). The rest of the SCENIC workflow was run as described in the previous section, leading to an activity matrix including 151 regulons. For the purpose of visualization, very sparse regulons can be filtered-out. For example, in **Fig. 1b**, we have plotted only regulons active in at least 1% of the cells and correlated with other regulons in the matrix (absolute correlation > 0.30). However, the downstream analyses include all the regulons.

Dataset: Human neurons (Lake et al.)

The human neurons dataset, published by Lake et al. [11], includes single-cell RNA-seq data of 3083 neuronal cells from a normal brain (retrieved postmortem from a 51-year old female, from six different Brodmann areas: BA8, BA10, BA17, BA21, BA22, BA41/42). The expression matrix (available at the host laboratory webpage: http://genome-tech.ucsd.edu/ZhangLab/index.php/data/epigenomics-and-transcriptomics/sns/) contains expression values (in TPM) for 25122 genes in 4039 cells. Of these, only 3083 cells are retained after filtering out low mapping outliers and potential doublets. Repeated genes, mitochondrial genes and non protein coding genes were removed and the matrix was renormalized as log2(TPM+1). To run GENIE3, this matrix was filtered to keep the 14941 genes with more than 154 normalized counts (which corresponds to 5 normalized counts in 1% of cells) and detected in more than 31 cells (1% of

Nature Methods: doi:10.1038/nmeth.4463

cells). The rest of the SCENIC workflow was run as described in the previous section, except that the selected AUCell threshold was 0.20 instead of 0.03, leading to an activity matrix including 130 regulons.

Dataset: Human brain (Darmanis et al.)

The human brain data set from Darmanis et al.[12] provides scRNA-seq data from 466 cells from adult and fetal human brains. The fetal samples were taken from four different individuals at 16 to 18 weeks post-gestation. The adult brain samples were taken from healthy temporal lobe tissue (according to the test) from 8 different patients (21, 22, 37, 47, 50, 63 years old) during temporal lobectomy surgery for refractory epilepsy and hippocampal sclerosis. The expression profiles for 22085 genes in each cell (expressed as raw reads) were downloaded from GEO (GSE67835), merged into an expression matrix, and converted to logged CPM [$\log_{10}$(Reads per gene in a cell/Total reads in a cell)*1000000+1)]. Genes expressed in overall with less than 9.32 logged CPM counts (corresponding to at least 2 counts in 1% of the population) and expressed in less than 5 cells (1% of the population) were removed, resulting in an expression matrix with 14703 genes. The rest of the SCENIC workflow was run as described previously, resulting in a Regulon Activity Matrix with 259 regulons.

Dataset: Mouse oligodendrocytes (Marques et al.)

The oligodendrocytes data set from Marques et al. [37] contains scRNA-seq data of 5069 cells from the oligodendrocyte lineage. Cells were obtained from several different mouse strains and isolated from ten different regions of the anterior-posterior and dorsal-ventral axis of the mouse juvenile and adult CNS; including white and grey matter. The expression matrix, downloaded from GEO (GSE75330), provides the expression values in UMI counts for 23556 genes in those 5069 cells. Genes expressed in overall with less than 100 counts (corresponding to 2 counts in 1% of the population) or expressed in less than ~51 cells were filtered out for GENIE3 analysis, resulting in 11985 genes. The rest of the SCENIC workflow was run as described in the previous section, leading to an activity matrix including 128 regulons.

Dataset: Oligodendroglioma (Tirosh et al.)

The oligodendroglioma data set from Tirosh et al. [13] includes scRNA-seq expression profiles for 4347 cells from 6 untreated grade II oligodendroglioma tumors with either IDH1 or IDH2 mutation, and 1p/19q co-deletion. The expression data, given as $\log_2$(TPM+1), was downloaded from GEO (GSE70630). We only used the tumoral cells for the analysis. Most of the non-tumoral cells were removed from the data set by the authors based on CNV profile analysis. However, a total of 303 non-tumoral cells that lacked detectable CNVs were still included in the data set. We removed these non-tumoral cells from the expression matrix using hierarchical clustering based on the markers cited in the article (mature oligodendrocytes and microglia, respectively). Out of the 23686 genes in the expression matrix, we run GENIE3 on the genes expressed with more than 202 logged TPM counts (at least 5 logged TPM counts in 1% of the population) and detected in more than 40 cells (1% of the total data set), resulting in an expression matrix with 14728 genes and 4043 cells. The SCENIC pipeline was executed as previously described, resulting in a Regulon Activity Matrix with 159 regulons.

Dataset: Melanoma (Tirosh et al.)

The melanoma dataset from Tirosh et al. [14] provides expression profiles of 23689 genes in 4645 cells from 19 melanoma tumors. These cells include both, malignant (melanoma cells), and non-malignant cells (e.g. immune cells). Here we analyze the 1252 melanoma cells (from 14 different tumors) that are labeled as malignant by the authors based on their CNV profiles. The expression matrix, as downloaded from GEO (GSE72056, on Apr 2016), is provided as logged TPM [log2(TPM/10+1)]. Therefore, for running GENIE3 we included the 14566 genes with more than 62.6 normalized counts per row (5 x 12.52 cells), that were detected (expression>0) in more than 12 cells (1%). In this way, the application of SCENIC on this dataset, lead to an activity matrix including 185 regulons.

Nature Methods: doi:10.1038/nmeth.4463

Dataset: Mouse retina (Macosko et al.)

The dataset from Macosko et al. [38] contains scRNA-seq data of 44808 cells (after pruning singletons) obtained through Drop-seq from mouse retina (14 days post-natal). The expression matrix was obtained from GEO (GSE63472), while the cluster information was obtained from the host laboratory webpage (http://mccarrolllab.com/dropseq/). We used the normalized expression matrix [given as *log((UMI counts per gene in a cell/Total UMI counts in cell)*10000)+1)*].

In order to reduce the computational cost of the analysis, the dataset was down-sampled into a smaller set in which all the given cell types are represented. In mouse retina, the majority of the cells are rods [71], which according to the authors, in this data set correspond to more than 29000 cells. Since rods are the smallest cell type in mouse retina [72] and express fewer genes, they also contain higher levels of noise. In order to take a representative sample not overtaken by the rods content, Macosko et al. selected cells which express more than 900 genes. We used this same down-sampling approach, which resulted in a selection of 11020 cells, to build the gene regulatory network. Running GENIE3 (and RcisTarget) on the 12953 genes with more than 55.1 normalized counts (0.5 in 1% of the population) and detected in more than ~55 cells (0.5% of the population). This network was then evaluated on all the cells in the dataset, which led to an activity matrix including 123 regulons.

Dataset: Embryonic mouse brain (10X Genomics)

The Chronium Megacell demonstration dataset contains 1,306,127 cells from cortex, hippocampus and subventricular zone of two E18 mice (strain: C57BL/6). We downloaded the expression matrix from the authors website (https://support.10xgenomics.com/single-cell-gene-expression/datasets/1M_neurons), which contains the expression data as 3 arrays in CSC (compressed sparse columnar) format compressed into a HDF5 file. Several subsets of this matrix were used to benchmark GRNboost (See GRNboost section).

Gene Ontology

To identify enriched GO terms or pathways associated to the states identified, we performed functional enrichment analysis on the union of regulons associated to each state (i.e. for the synaptic oligodendrocyte group, the union of targets of Hoxa7, Hoxa9, Hoxb7 and Hoxb9). The analyses were performed mainly through Mouse Mine [73] (although we also checked DAVID [74,75], Enrichr [76,77], which provided similar results).

Differential expression analysis

Differential expression between clusters of single cells was performed using MAST [78]. Genes expressed with less than 9 counts in the population were excluded. Differentially expressed genes were evaluated according to their log fold change and adjusted p-values.

Nature Methods: doi:10.1038/nmeth.4463

Method comparisons

Method comparison for cell clustering

To determine whether the clustering based on gene regulatory network activity matches real cell types, we calculated sensitivity and specificity for the GRN assigned to the cells of the mouse brain (**Fig. 1d**) according to the cell types assigned by Zeisel et al. For the other datasets, we compared the clustering (mainly t-SNE) based on the regulon activity matrices to the cell labels provided in the corresponding publications. For the comparison with clustering methods, we build on the benchmark presented in the SC3 publication [10], which also uses Zeisel et al., and provides the adjusted Rand index (ARI) on this dataset for 6 clustering methods commonly used on single-cell RNA-seq data. We extended the comparison with SEURAT by adding the results obtained with different resolution values.

Method comparison for TF-motif discovery

The validation of the TFs identified by SCENIC was mainly done by confirming their role in the given cell type in literature (e.g. **Fig. 1c**). However, we also compared SCENIC to an alternative approach to identify TFs potentially regulating cell states: Applying transcription factor motif enrichment analysis on genes differentially expressed between clusters (i.e. gene signature, or markers for a cell type). To do so, we started from the 'gold standard' for the mouse brain dataset: the cell type labels assigned by the authors, which are based on their own biclustering algorithm (Backspin) plus annotation based on markers to find the correspondence of each cluster to a given cell type. The 'signatures' for each cell type or cluster are defined based on four alternative approaches: (1) The genes assigned by Backspin to each cluster, (2) differentially expressed genes of each cluster versus all other cluster (One versus rest, OvR), (3) differentially expressed genes in each cluster versus any of the other clusters (One versus any other, OvAny), (4) highly variable genes across clusters (HVG). All the differential expression analyses, and the identification of HVG were run using EdgeR. We also run the best performing method, OvR, with an alternative differential expression tool, MAST, to confirm that the differential expression tool doesn't have a major impact on the results. Homer (version 4.9) and RcisTarget were then run on each of the gene-sets resulting from these contrasts: Backspin (pyramidal: 1960g, interneurons: 1126g, oligodendrocyte: 579g, microglia: 392g, astrocytes (only): 206g), OvR edger (interneurons: 574g, pyramidal: 1031g, oligodendrocytes: 859g, microglia: 1421g, endothelial_mural: 1541g, astrocytes_ependymal: 1121g), OvR MAST (pyramidal: 818g, interneurons: 688g, oligodendrocytes: 1057g, astrocytes_ependymal: 487g, microglia: 607g, endothelial_mural: 549g), OvAny (astrocytes_ependymal: 871g, endothelial_mural: 1019g, interneurons: 916g, microglia: 1060g, oligodendrocytes: 1013g, pyramidal: 816g), and 2013 highly variable genes (FDR<0.01, logFC>1).

Homer was run using the default parameters (promoter region: -300 to +50bp around TSS, motifs of length 8,10,12 and masking repeats). For the rest of the analysis we only took into account known motifs (ignored de-novo motifs), using the TF on the motif name as annotation to transcription factors. Note that Homer was selected as alternative TF-discovery method, because it was the second-best performer in the iRegulon benchmark [35]. The equivalent analysis was also run with RcisTarget (the tool used by SCENIC), also using the default parameters with the two available databases (0 to - 500b, and -10kbp to +10kbp around TSS) and only 'direct annotation' (TFs annotated to the motif in the original motif source). With both tools, we only took into account those TFs that are differentially expressed themselves, as this is also a standard approach to prioritize TFs and reduce false positives. For comparison with SCENIC, we used the results on the mouse brain presented in this paper ("diamond" shape in **Supplementary Fig. 3e**). Two versions of statistics are presented: One including all the TFs returned by SCENIC (including the sparse regulons), and one including only the "cell type" regulons, the regulons that are mostly specific to one of the final clusters. During the progress of this project, we updated the database of RcisTarget to contain about 2k more motifs, and developed GRNboost. Therefore, the figure includes the results for SCENIC using these new features (labelled as "DBv2" and "GRNboost") as "true" TFs for the validation. This *List of TFs* include gene sets from mouse genome informatics (MGI) mammalial phenotype database (sample terms: abnormal [cell type] morphology/physiology, increased/decreased [cell type] number, abnormal brain morphology, abnormal blood-brain barrier function) and using the cell type as keyword (oligodendrocyte, astrocyte,

Nature Methods: doi:10.1038/nmeth.4463

interneuron, pyramidal, neuron, microglia, brain endothelial), and from the gene ontology (e.g. GO:0014013 Regulation of gliogenesis, GO:0022008 Neurogenesis, …). Finally, the precision and recall for each method were calculated according to the TFs identified across all the cell types (e.g. joining all cell types, not cell-type specific), since multiple cell-type specific TFs known from literature were only available in generic terms (e.g. "brain", "gliogenesis"), not in the cell specific annotations.

List of TFs *used for validation (in brackets, number of gene-sets in which it appears): Emx1* (18), Nf1 (16), Foxg1 (14), Rb1 (14), Sox10 (14), Ascl1 (13), Olig1 (13), Olig2 (13), Ctnnb1 (12), Emx2 (12), En1 (12), Trp53 (12), Gli2 (11), Gli3 (11), Hes1 (11), Hes5 (11), Pax6 (11), Phox2b (11), Pitx3 (11), Rora (11), Sox11 (11), Zbtb18 (11), Abl1 (10), Eomes (10), Hoxa2 (10), Id2 (10), Isl1 (10), Lhx5 (10), Lhx6 (10), Meox2 (10), Mycn (10), Neurod1 (10), Neurog2 (10), Nr4a3 (10), Otx2 (10), Prnp (10), Setdb1 (10), Trp73 (10), Atoh1 (9), Bax (9), Cux2 (9), Id1 (9), Kmt2a (9), Mecp2 (9), Mnx1 (9), Nkx2-2 (9), Nkx6-2 (9), Nr2e1 (9), Pou3f2 (9), Rest (9), Sox2 (9), Xrcc1 (9), Arx (8), Bhlhe22 (8), Dgcr8 (8), Dmrta2 (8), Esr2 (8), Ferd3l (8), Fezf1 (8), Gli1 (8), Gsx2 (8), Hdac2 (8), Hif1a (8), Hspa5 (8), Id4 (8), Insm1 (8), Lmx1a (8), Mbd1 (8), Nr3c1 (8), Olig3 (8), Rbfox2 (8), Rbm8a (8), Six1 (8), Smarcc2 (8), Sod1 (8), Sox1 (8), Trib2 (8), Egr2 (7), En2 (7), Foxc1 (7), Gata2 (7), Kdm2a (7), Lbx1 (7), Lmx1b (7), Magoh (7), Mapk1 (7), Mef2c (7), Nkx6-1 (7), Nr1h2 (7), Nr1h3 (7), Nr4a2 (7), Parp1 (7), Pax3 (7), Pitx2 (7), Pou1f1 (7), Pou4f2 (7), Pparg (7), Rbpj (7), Rela (7), Smarca1 (7), Tbr1 (7), Tfap2a (7), Bcl11a (6), Canx (6), Creb1 (6), Fev (6), Fezf2 (6), Foxa2 (6), Gata3 (6), Hey1 (6), Hoxa3 (6), Hoxb1 (6), Jun (6), Klf7 (6), Lhx8 (6), Mef2d (6), Ncor2 (6), Nfe2l2 (6), Nr2f1 (6), Prox1 (6), Ptf1a (6), Pura (6), Runx1 (6), Sim1 (6), Sp2 (6), Thrb (6), Tlx3 (6), Zfp521 (6), Adarb1 (5), Atf1 (5), Cers2 (5), Dbx1 (5), Ddit3 (5), Dlx5 (5), Ebf2 (5), Etv1 (5), Foxj1 (5), Gbx2 (5), Hoxa1 (5), Hoxb3 (5), Hoxd3 (5), Isl2 (5), Lhx1 (5), Mdm2 (5), Mef2a (5), Neurog1 (5), Otp (5), Otx1 (5), Pax2 (5), Pbx3 (5), Phox2a (5), Pik3c3 (5), Ppargc1a (5), Rel (5), Rps6ka5 (5), Smad2 (5), Smad4 (5), Sox4 (5), Sp8 (5), Spi1 (5), Stat3 (5), Tal1 (5), Tcf4 (5), Tet1 (5), Thra (5), Vax1 (5), Vsx2 (5), Ywhae (5), Zfp24 (5), Zic2 (5), Atf2 (4), Atf5 (4), Barhl2 (4), Ckmt1 (4), Cux1 (4), Dlx2 (4), Dmrt3 (4), Egr1 (4), Foxo6 (4), Foxp2 (4), Gbx1 (4), Gcm2 (4), Gfi1 (4), Hey2 (4), Heyl (4), Hhat (4), Hltf (4), Hmga2 (4), Hoxb2 (4), Hsf1 (4), Kdm2b (4), Lhx2 (4), Lhx3 (4), Lin28a (4), Msx1 (4), Myc (4), Myrf (4), Neurod2 (4), Nkx2-9 (4), Npas4 (4), Nr2c2 (4), Pax8 (4), Plg (4), Pou3f1 (4), Pou4f1 (4), Pou4f3 (4), Rarb (4), Sall4 (4), Shox2 (4), Slc18a1 (4), Tcf12 (4), Tcf7l2 (4), Tfam (4), Tgif1 (4), Tgif2 (4), Vax2 (4), Zeb1 (4), Adnp (3), Anxa1 (3), Arnt (3), Arntl (3), Atoh7 (3), Barhl1 (3), Bcl11b (3), Bcl6 (3), Cbfb (3), Clock (3), Crem (3), Crx (3), Dlx1 (3), E2f1 (3), E2f3 (3), Egr3 (3), Ep300 (3), Epas1 (3), Esr1 (3), Etv4 (3), Etv5 (3), Fez1 (3), Fos (3), Foxn4 (3), Foxo3 (3), Gata4 (3), Gcm1 (3), Grhl3 (3), Gsx1 (3), Gtf2ird1 (3), Hand2 (3), Hdac1 (3), Helt (3), Hmx1 (3), Hoxc10 (3), Hoxd10 (3), Hoxd9 (3), Kdm4a (3), Klf4 (3), Lef1 (3), Lhx4 (3), Maf (3), Meis1 (3), Myf5 (3), Ncor1 (3), Nfatc4 (3), Nfia (3), Nhlh1 (3), Nhlh2 (3), Nkx2-5 (3), Nr1d1 (3), Nr2f2 (3), Nr3c2 (3), Nr5a1 (3), Nrl (3), Pdlim5 (3), Plag1 (3), Pou3f3 (3), Prdm13 (3), Rfx4 (3), Rorb (3), Runx3 (3), Rxra (3), Sall1 (3), Satb1 (3), Sema4a (3), Six3 (3), Sox8 (3), Sox9 (3), Spr (3), Stat1 (3), Tbx1 (3), Tlx1 (3), Tlx2 (3), Tulp1 (3), Ube2k (3), Zeb2 (3), Zfp335 (3), Zfp536 (3), Adnp2 (2), Aire (2), Alx4 (2), Ar (2), Arnt2 (2), Ascl2 (2), Atf3 (2), Atf7 (2), Bad (2), Bhlhe23 (2), Bhlhe40 (2), Brca1 (2), Cbfa2t2 (2), Cd59a (2), Cdx2 (2), Celf4 (2), Cers6 (2), Csnk2b (2), Cycs (2), Dab2 (2), Dbp (2), Dnmt3a (2), E2f2 (2), E2f5 (2), Ebf3 (2), Erg (2), Etv6 (2), Evx1 (2), Fli1 (2), Fosb (2), Foxa1 (2), Foxb1 (2), Foxc2 (2), Foxd1 (2), Foxq1 (2), Gabpa (2), Gadd45a (2), Hdac8 (2), Hhex (2), Hlf (2), Hmg20a (2), Hmg20b (2), Hmgb2 (2), Hoxa13 (2), Hoxb13 (2), Hoxb8 (2), Hoxc8 (2), Hoxd1 (2), Hsf2 (2), Irx3 (2), Irx5 (2), Irx6 (2), Kdm4c (2), Luzp1 (2), Mafb (2), Mafg (2), Mafk (2), Mesp1 (2), Mitf (2), Msi1 (2), Msx2 (2), Mycl (2), Neurog3 (2), Nfib (2), Nfix (2), Nkx2-1 (2), Nme1 (2), Nr2e3 (2), Nr2f6 (2), Nup133 (2), Ovol2 (2), Pax5 (2), Pbx1 (2), Pick1 (2), Pou3f4 (2), Ppard (2), Pqbp1 (2), Prdm16 (2), Prkaa2 (2), Prop1 (2), Prrx1 (2), Prrxl1 (2), Rab18 (2), Rara (2), Rufy3 (2), Runx2 (2), Rxrb (2), Sall2 (2), Scrt1 (2), Scrt2 (2), Sim2 (2), Smad5 (2), Smad9 (2), Sox14 (2), Sox21 (2), Sox5 (2), Sp4 (2), Srf (2), Stau2 (2), Stub1 (2), Tal2 (2), Tbp (2), Tbx21 (2), Tbx4 (2), Tbx6 (2), Tcf3 (2), Tef (2), Thoc2 (2), Twist1 (2), Vsx1 (2), Ybx1 (2), Zbtb20 (2), Zhx2 (2), Zic5 (2), Abcf2 (1), Acaa1a (1), Aff4 (1), Ahctf1 (1), Ahrr (1), Akr1a1 (1), Alx1 (1), Alx3 (1), Arg1 (1), Arid3a (1), Aspscr1 (1), Atf4 (1), Atoh8 (1), Aven (1), Bach1 (1), Bach2 (1), Banp (1), Barx2 (1), BC005561 (1), Bcl6b (1), Bclaf1 (1), Bhlhe41 (1), Bmyc (1), Bnc2 (1), Bptf (1), Carf (1), Cat (1), Cbx7 (1), Ccnt2 (1), Cdx1 (1), Cdx4 (1), Cebpa (1), Cebpb (1), Cebpd (1), Cebpe (1), Cebpz (1), Celf6 (1), Cic (1), Clk1 (1), Cnot3 (1), Cnot6 (1), Cphx1 (1), Crebzf (1), Deaf1 (1), E2f6 (1), Ebf1 (1), Edn1 (1), Elk3 (1), Erf (1), Esx1 (1), Ets1 (1), Ets2 (1), Etv2 (1), Foxd3 (1), Foxe3 (1), Foxh1 (1), Foxi1 (1), Foxn1 (1), Foxp1 (1), Foxp4 (1), Foxs1 (1), Fubp1 (1), Gata5 (1), Gata6 (1), Glis1 (1), Grhl2 (1), Gsc (1), Gtf2i (1), Hand1 (1), Hesx1 (1), Hic1 (1), Hif3a (1), Hlx (1), Hnf4a (1), Hoxa11 (1), Hoxd11 (1), Ikzf1 (1), Il21 (1), Irf8 (1), Kcnip1 (1), Klf15 (1), Klf2 (1), Klf3 (1), Lbx2 (1), Lhx9 (1), Lin28b (1), Mecom (1), Msrb3 (1), Mtf1 (1), Mylk (1), Nanos1 (1), Nfatc1 (1), Nfatc2 (1), Nfatc3 (1), Nfkb1 (1), Nkx2-3 (1), Nkx3-2 (1), Nobox (1), Noto (1), Npas2 (1), Nr6a1 (1), Onecut2 (1), Patz1 (1), Pax7 (1), Pds5a (1), Peg3 (1), Phf2 (1), Pitx1 (1), Pou2f1 (1), Ppara (1), Prdm1 (1), Prdm12 (1), Prdm6 (1), Prox2 (1), Rax (1), Rxrg (1), Sall3 (1), Six4 (1), Six6 (1), Smad1 (1), Smad3 (1), Smad6 (1), Smarcc1 (1), Snai1 (1), Sox17 (1), Sox18 (1), Sox3 (1), Sox6 (1), Sp1 (1), Sp3 (1), Srebf2 (1), Ssbp3 (1), Tbx19 (1), Tbx20 (1), Tcf7 (1), Tcf7l1 (1), Tead1 (1), Tead2 (1), Tead3 (1), Tfap2c (1), Tfap2d (1), Tfap2e (1), Tfdp1 (1), Tppp (1), Trip10 (1), Twist2 (1), Ubp1 (1), Uncx (1), Usf1 (1), Usf2 (1), Vamp3 (1), Vdr (1), Vezf1 (1), Xbp1 (1), Yy1 (1), Zbtb14 (1), Zfp385a (1), Zic1 (1), Zic3 (1), Zic4 (1), Zscan10 (1)

Method comparison for batch effect correction

The comparison of batch effect removal methods was performed using the Oligodendroglioma dataset by Tirosh et al. [13]. SCENIC was run in the standard way (see Methods: Oligodendroglioma dataset), the t-SNE and diffusion plots were run on the full binary regulon activity matrix (the heatmap in **Supplementary Fig. 8** illustrates selected stage-specific regulons). Combat [16,41] and Limma [17,42] were run to correct for "patient of origin" as source of batch effect (input matrix: 14728 genes and 4043 cells, same as GENIE3/SCENIC). Diffusion plots were done using the R/Bioconductor package destiny [79].

Nature Methods: doi:10.1038/nmeth.4463

Method comparison for cycling cells

Gene sets used to identify cycling cells are 46 sets related to the mitotic cell cycle, with at least 10 genes, retrieved from amiGO and cycleBase 1.0 and 2.0; Cycling cells are those that show consistent up-regulation of these genes, and are selected by hierarchical clustering on the matrix containing the z-scores for each gene-set. Since most of the methods provide multiple clusters as output, to compare their results, for each method we selected the cluster with the biggest amount of CC cells. This cluster was then used to calculate the precision and recall.

Combat [16,41] and **Limma** [17,42] were run using the same logged and filtered matrix as in SCENIC. The tumor of origin was specified as the source of batch effect and hierarchical clustering (Ward's method) was performed in the batch corrected matrix. The number of clusters was determined by the cutreeDynamic function of the dynamicTreeCut package [87]. **GiniClust** [70] was run on the unlogged TPM matrix with the default parameters, which resulted in a matrix with 17843 genes and one single cluster. **PAGODA** [64] was run using the unlogged matrix (filter: 13976 genes with more than 100 TPM counts per row –at least 3 TPM counts in 1% of the population– and detected in more than 12 cells). The error models were generated using a k-nearest neighbor model fitting (minimum of 2 reads for the gene to be initially classified as a non-failed measurement, and at least 5 non-failed measurements per gene). The models were fitted based on 1/2 of the most similar cells (assuming that there will be two main subpopulations). These thresholds were established to avoid the impact of the TPM normalization of the counts. The variance normalization –which aims to normalize out technical bias and biological noise– was performed by trimming the 3 most extreme cells and limiting the maximum adjusted variance to 5. The rest of the PAGODA steps were run using default parameters, except: The evaluation of overdispersion of 'de novo' gene sets was performed with a trimming value of 7.1 extreme cells, and 50 as the number of clusters to be determined; threshold in the determination of top aspects a p-value: 0.01; distance threshold for the reduction of redundancy: 0.9. The optimal number of clusters in the data set was set to 10. **pcaReduce** [81] was run using the unlogged TPM matrix, with a similar filter as in SC3. The number of times to repeat the pcaReduce framework was set to 100 and the number of dimensions to start with was set to 30. Both merging methods (sampling based, S, and probability based, M) were tested. The number of clusters in the data set selected was 17 in both cases. **RaceID** [82] was run using an unlogged expression matrix filtered with the package's function filterdata. The genes expressed with less than 8 TPM counts in at least 13 cells were discarded, resulting in 10371 genes. This matrix was normalized according to the RaceID procedure. The rest of the RaceID pipeline was run using default parameters. **SC3** [10] was run using the unlogged expression matrix. Default filters were applied, resulting in a matrix with 11645 genes. The number of estimated clusters was set to 21 (based on the output of the *sc3_estimate_k* function). **SEURAT** [38] was run using the unlogged TPM matrix and default parameters, correcting for tumor of origin as batch effect. The FindClusters function provided 8 clusters. **SIMLR** [69] was run with default paramaters using the unlogged expression matrix, with a similar filter to SC3. The number of clusters to be specified was set to 15. **SINCERA** [7] was run using the unlogged TPM matrix. A gene-by-gene per-cell z-score transformation was performed during the analysis, and hierarchical clustering was applied using correlation distances and average clustering. The number of clusters identified in the data set defined by the tool was 12. This threshold is set as the highest number of possible clusters that generates less than 1 singleton cluster. **SNNCliq** [83] was run using the scripts provided at http://hemberg-lab.github.io. The number of clusters found in the data set was 15. All these methods were run in R, except for SNNCliq, which is implemented in Python. Note that there are two tools related to cell cycle and single-cell RNA-seq datasets which have not been included in this analysis: (1) scLVM allows to correct for cofounding factors. These can be given in the form of a gene-set, for example GO gene-sets related to cell cycle, to correct for the cell cycle effect. However, to our knowledge, it does not provide an explicit score of the gene set on the cells. (2) Cyclone is a method to split cells according to their cell cycle stage. However, it assigns a cell cycle state to all the cells in the dataset, and thus, it is not useful for our purpose of identifying the cycling cells.

Nature Methods: doi:10.1038/nmeth.4463

Immunohistochemistry of melanoma biopsies

Immunohistochemistry was performed on formalin-fixed, paraffin-embedded melanoma samples on the Leica BOND-MAX™ automatic immunostainer (Leica Microsystems). The samples include biopsies of 9 primary melanomas (4 in radial growth phase and 5 in vertical growth phase), 8 melanoma-containing sentinel lymph nodes, and 8 melanoma metastases. Antigen retrieval was performed onboard using a citrate-based (Bond Epitope Retrieval Solution 1, pH 6.0; Leica) or a EDTA-based buffer (Bond Epitope Retrieval Solution 2, pH 9.0; Leica) according to the manufacturer's instructions. The antibodies were used for melanA (IR633 from DAKO, initially diluted at RTU, but further diluted 1:2 for better contrast; antigen retrieval: pH9.0), EPHA2 (#6997 from Cell Signaling Technology diluted at 1/50; pH9.0), ZEB1 (sc-25388 from Santa Cruz Biotechnology diluted at 1/200; pH9.0), NFATC2 (#5861 from Cell Signaling Technology diluted at 1/5000; pH6.0) and NFIB (HPA003956 from Sigma Aldrich diluted at 1/250; pH6.0). Alkaline phosphatase activity was detected with Bond Polymer Refine Red Detection (Leica) as substrate, resulting in a pink/red immunoreactivity. To help identification of melanoma cells in sentinel lymph nodes, double immunohistochemical staining with melanA were performed with sequential development of peroxidase and alkaline phosphatase with Bond Polymer Refine Detection and Bond Polymer Refine Red Detection (Leica Microsystems, Wetzlar, Germany), respectively, resulting in contrasting dark brown (marker) and pink/red immunoreactivities (melanA).

Knock-down of NFATC2 in melanoma cell culture

A375 cell line was selected based on expression of NFATC2, NFIB (**Supplementary Fig. 13a-b**), and SOX10 across 59 melanoma cell lines from the COSMIC Cancer Cell lines Project [43]. A375 cells were obtained from the ATCC and were cultured in Dulbecco's Modified Eagle's Medium with high glucose and glutamax (ThermoFisher Scientific), supplemented with 10% fetal bovine serum (Lonza) and penicillin-streptomycin (ThermoFisher Scientific). Knockdown of NFATC2 was performed using the ON-TARGETplus NFATC2 siRNA SMARTpool (Dharmacon) at a final concentration of 40nM in opti-MEM medium (ThermoFisher Scientific). Total RNA was extracted 72 hours after knockdown, using the innuPREP RNA mini kit (Analytik Jena), according to the manufacturer's instructions. Quality checks were performed using the Bioanalyzer 1,000 DNA chip (Agilent) after which libraries were constructed: After total RNA purification, mRNA was enriched using the Dynabeads mRNA purification kit (Invitrogen). To make cDNA, 1 µl of oligo(dT) primers (500ng/µl; Ambion) and 1 µl of 10 mM dNTP (Promega) was added to 10 µl of polyA-selected mRNA; incubated at 65°C for 5 min and placed on ice. First-strand cDNA synthesis was performed by adding 4 µl of first strand buffer (Invitrogen), 2 µl of 100 mM DTT (Invitrogen) and 1 µl of Superscript II (Invitrogen) and incubating the mix at 42°C for 50 min, then 70°C for 15 min. The second strand of cDNA was filled in by adding 35 µl of water, 15 µl of 5x second strand buffer (Invitrogen), 1.5 µl of 10 mM dNTP, 0.5 µl of 10 U/µl E Coli DNA ligase (Bioke), 2 µl of 10 U/µl E Coli DNA polymerase I (Bioke), 1 µl of 2 U/µl E Coli RNaseH and then incubating at 16°C for 2 hours. The cDNA was purified on a MinElute column (Qiagen) and eluted in 15 µl EB buffer. To incorporate sequencing adapters, we combined the purified cDNA with 4 µl of Nextera TD buffer (Illumina) and 1 µl of Nextera Tn5 enzyme (Illumina) on ice and incubated at 55°C for 5 min. The tagmented cDNA was purified again on a MinElute column and eluted in 20 µl EB buffer. To PCR amplify the fragments, we added 25 µl of NEBnext PCR master mix (Bioke), 5 µl of Nextera primer mix and incubated at 72°C for 5 min, then at 98°C for 30 sec, followed by 15 cycles of 98°C for 10 sec, 63°C for 30 sec and 72°C for 3 min. We purified the PCR amplicons with 55 µl AMPure beads (Analis).

Final libraries were pooled and sequenced on a NextSeq 500 and HiSeq 4000 (Illumina). Raw Fastq files of the same sample were merged and adapter sequences were removed using fastq-mcf. RNA-seq reads were mapped to the genome (hg19) using STAR (v2.5.1b) and reads with high mapping quality (Q4) were selected using SAMtools (v1.4). Read counts per gene were obtained from the aligned reads using htseq-count. The Bioconductor/R package DESeq2 was used for normalization and differential gene expression analysis. Log2FoldChange values were used for ranking the genes, and downstream GOrilla and GSEA analysis, as previously described.

Nature Methods: doi:10.1038/nmeth.4463

References

45. Cahoy, J. D. *et al.* A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function. *J. Neurosci.* 28, 264–278 (2008).

46. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445, 168–176 (2007).

47. Lavin, Y. *et al.* Tissue-Resident Macrophage Enhancer Landscapes Are Shaped by the Local Microenvironment. *Cell* 159, 1312–1326 (2014).

48. Tasic, B. *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* 19, 335–346 (2016).

49. Zhou, Q. *et al.* A mouse tissue transcription factor atlas. *Nat. Commun.* 8, 15089 (2017).

50. Gjoneska, E. *et al.* Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer/'s disease. *Nature* 518, 365–369 (2015).

51. Kelsom, C. & Lu, W. Development and specification of GABAergic cortical interneurons. *Cell Biosci.* 3, 19 (2013).

52. Rudy, B., Fishell, G., Lee, S. & Hjerling-Leffler, J. Three Groups of Interneurons Account for Nearly 100% of Neocortical GABAergic Neurons. *Dev. Neurobiol.* 71, 45–61 (2011).

53. Gjerdrum, C. *et al.* Axl is an essential epithelial-to-mesenchymal transition-induced regulator of breast cancer metastasis and patient survival. *PNAS* 107, 1124–1129 (2010).

54. Richard, G. *et al.* ZEB1-mediated melanoma cell plasticity enhances resistance to MAPK inhibitors. *EMBO Mol. Med.* 8, 1143–1161 (2016).

55. Garcia-Diaz, A. *et al.* Interferon Receptor Signaling Pathways Regulating PD-L1 and PD-L2 Expression. *Cell Rep.* 19, 1189–1201 (2017).

56. Schummer, P., Kuphal, S., Vardimon, L., Bosserhoff, A. K. & Kappelmann, M. Specific c-Jun target genes in malignant melanoma. *Cancer Biol. Ther.* 17, 486–497 (2016).

57. Oskay Halacli, S. FOXP1 enhances tumor cell migration by repression of NFAT1 transcriptional activity in MDA-MB-231 cells. *Cell Biol. Int.* 41, 102–110 (2017).

58. Furukawa, T., Mukherjee, S., Bao, Z.-Z., Morrow, E. M. & Cepko, C. L. rax, Hes1, and notch1 Promote the Formation of Müller Glia by Postnatal Retinal Progenitor Cells. *Neuron* 26, 383–394 (2000).

59. Muto, A., Iida, A., Satoh, S. & Watanabe, S. The group E Sox genes Sox8 and Sox9 are regulated by Notch signaling and are required for Müller glial cell development in mouse retina. *Exp. Eye Res.* 89, 549–558 (2009).

60. Corso-Díaz, X. & Simpson, E. M. Nr2e1 regulates retinal lamination and the development of Müller glia, S-cones, and glycineric amacrine cells during retinogenesis. *Mol. Brain* 8, 37 (2015).

61. Li, M. *et al.* Comprehensive Analysis of Gene Expression in Human Retina and Supporting Tissues. *Hum. Mol. Genet.* ddu114 (2014). doi:10.1093/hmg/ddu114

62. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10, 48 (2009).

63. Binns, D. *et al.* QuickGO: a web-based tool for Gene Ontology searching. *Bioinforma. Oxf. Engl.* 25, 3045–3046 (2009).

64. Fan, J. *et al.* Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* 13, 241–244 (2016).

65. Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* 14, 309–315 (2017).

66. Vallejos, C. A., Marioni, J. C. & Richardson, S. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLOS Comput. Biol.* 11, e1004333 (2015).

67. Lin, P., Troup, M. & Ho, J. W. K. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* 18, 59 (2017).

68. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* 33, 1179–1186 (2017).

69. Wang, B. *et al.* SIMLR: a tool for large-scale single-cell analysis by multi-kernel learning. *ArXiv170307844 Cs Q-Bio* (2017).

70. Jiang, L., Chen, H., Pinello, L. & Yuan, G.-C. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biol.* 17, 144 (2016).

71. Jeon, C. J., Strettoi, E. & Masland, R. H. The major cell populations of the mouse retina. *J. Neurosci.* 18, 8936–8946 (1998).

Nature Methods: doi:10.1038/nmeth.4463

72. Carter-Dawson, L. D. & LaVail, M. M. Rods and cones in the mouse retina. I. Structural analysis using light and electron microscopy. *J. Comp. Neurol.* 188, 245–262 (1979).

73. Motenko, H., Neuhauser, S. B., O'Keefe, M. & Richardson, J. E. MouseMine: a new data warehouse for MGI. *Mamm. Genome* 26, 325–330 (2015).

74. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13 (2009).

75. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57 (2009).

76. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 14, 128 (2013).

77. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90-97 (2016).

78. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16, 278 (2015).

79. Angerer, P. *et al.* destiny : diffusion maps for large-scale single-cell data in R. *Bioinformatics* 32, 1241–1243 (2016).

80. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 24, 719–720 (2008).

81. žurauskienė, J. & Yau, C. pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics* 17, 140 (2016).

82. Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 525, 251–255 (2015).

83. Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* 31, 1974–1980 (2015).

Nature Methods: doi:10.1038/nmeth.4463