SCENIC: single-cell regulatory network inference and clustering

Sara Aibar[1,2], Carmen Bravo González-Blas[1,2], Thomas Moerman[3,4], Vân Anh Huynh-Thu[5], Hana Imrichova[1,2], Gert Hulselmans[1,2], Florian Rambow[6,7], Jean-Christophe Marine[6,7], Pierre Geurts[5], Jan Aerts[3,4], Joost van den Oord[8], Zeynep Kalender Atak[1,2], Jasper Wouters[1,2,8] & Stein Aerts[1,2]

We present SCENIC, a computational method for simultaneous gene regulatory network reconstruction and cell-state identification from single-cell RNA-seq data (http://scenic.aertslab.org). On a compendium of single-cell data from tumors and brain, we demonstrate that *cis*-regulatory analysis can be exploited to guide the identification of transcription factors and cell states. SCENIC provides critical biological insights into the mechanisms driving cellular heterogeneity.

The transcriptional state of a cell emerges from an underlying gene regulatory network (GRN) in which a limited number of transcription factors (TFs) and cofactors regulate each other and their downstream target genes. Recent advances in single-cell transcriptome profiling have provided exciting opportunities for high-resolution identification of transcriptional states and of transitions between states—for example, during differentiation[1,2]. Statistical techniques and bioinformatics methods that are optimized for single-cell RNA-seq have led to new biological insights[3], but it is still unclear whether specific and robust GRNs underlying stable cell states can be determined. This may indeed be challenging given that at the single-cell level, gene expression may be partially disconnected from the dynamics of TF inputs on account of stochastic variation of gene expression from transcriptional bursting and other sources[4]. A few methods have been developed that infer coexpression networks from single-cell RNA-seq data[5–7], but these methods do not use regulatory sequence analysis to predict interactions between TFs and target genes.

We reasoned that linking *cis*-regulatory sequences to single-cell gene expression could overcome dropouts and technical variation and thus optimize the discovery and characterization of cell states. To this end, we developed single-cell regulatory network inference and clustering (SCENIC) to map GRNs and then identify stable cell states by evaluating the activity of the GRNs in each cell. The SCENIC workflow consists of three steps (**Fig. 1a**, **Supplementary Fig. 1** and see Online Methods). In the first step, sets of genes that are coexpressed with TFs are identified using GENIE3 (ref. 8) (**Supplementary Fig. 1a**). Since the GENIE3 modules are only based on coexpression, they may include many false positives and indirect targets. To identify putative direct-binding targets, each coexpression module is subjected to *cis*-regulatory motif analysis using RcisTarget (**Supplementary Fig. 1b** and see Online Methods). Only modules with significant motif enrichment of the correct upstream regulator are retained, and they are pruned to remove indirect targets lacking motif support. We refer to these processed modules as regulons.

As part of SCENIC, we developed the AUCell algorithm to score the activity of each regulon in each cell (**Supplementary Figs. 1c** and **2**, and see Online Methods). For a given regulon, comparing AUCell scores across cells makes it possible to identify which cells have significantly higher subnetwork activity. The resulting binary activity matrix has reduced dimensionality, which can be useful for downstream analyses. For example, clustering based on this matrix identifies cell types and states based on the shared activity of a regulatory subnetwork. Since the regulon is scored as a whole, instead of using the expression of individual genes, this approach is robust against dropouts (**Supplementary Fig. 3**).

To evaluate the performance of SCENIC, we applied it to an scRNA-seq data set with well-known cell types from the adult mouse brain[9] (**Fig. 1b–e**). This analysis provided 151 regulons—out of 1,046 initial coexpression modules—with significantly enriched motifs for the corresponding TFs (7% of the initial TFs). Scoring regulon activity for each cell revealed the expected cell types (**Fig. 1d**,**e**) alongside a list of potential master regulators for each cell type (e.g., the microglia network in **Supplementary Fig. 4**). Clustering by cell type (overall sensitivity of 0.88, specificity of 0.99, and adjusted Rand index (ARI) > 0.80) is more accurate than many dedicated single-cell clustering methods[10].

To assess the robustness of SCENIC, we reanalyzed the mouse brain data: the full data set; samples of 100 randomly selected cells to simulate small data sets; or one-third of the sequencing reads to simulate low-coverage data. SCENIC identified cell types that are represented by only a few cells (e.g., two to six cells from microglia, astrocytes or interneurons; **Supplementary Fig. 5**). In addition, the predicted associations of TFs with cell type are consistent with previously established

[1]VIB Center for Brain & Disease Research, Laboratory of Computational Biology, Leuven, Belgium. [2]KU Leuven, Department of Human Genetics, Leuven, Belgium. [3]KU Leuven ESAT/STADIUS, VDA-lab, Leuven, Belgium. [4]IMEC Smart Applications and Innovation Services, Leuven, Belgium. [5]Department of Electrical Engineering and Computer Science, University of Liège, Liège, Belgium. [6]VIB Center for Cancer Biology, Laboratory for Molecular Cancer Biology, Leuven, Belgium. [7]KU Leuven, Department of Oncology, Leuven, Belgium. [8]KU Leuven, Department of Imaging and Pathology Translational Cell and Tissue Research, Leuven, Belgium. Correspondence should be addressed to S.A. (stein.aerts@kuleuven.vib.be).

RECEIVED 5 DECEMBER 2016; ACCEPTED 7 SEPTEMBER 2017; PUBLISHED ONLINE 9 OCTOBER 2017; DOI:10.1038/NMETH.4463

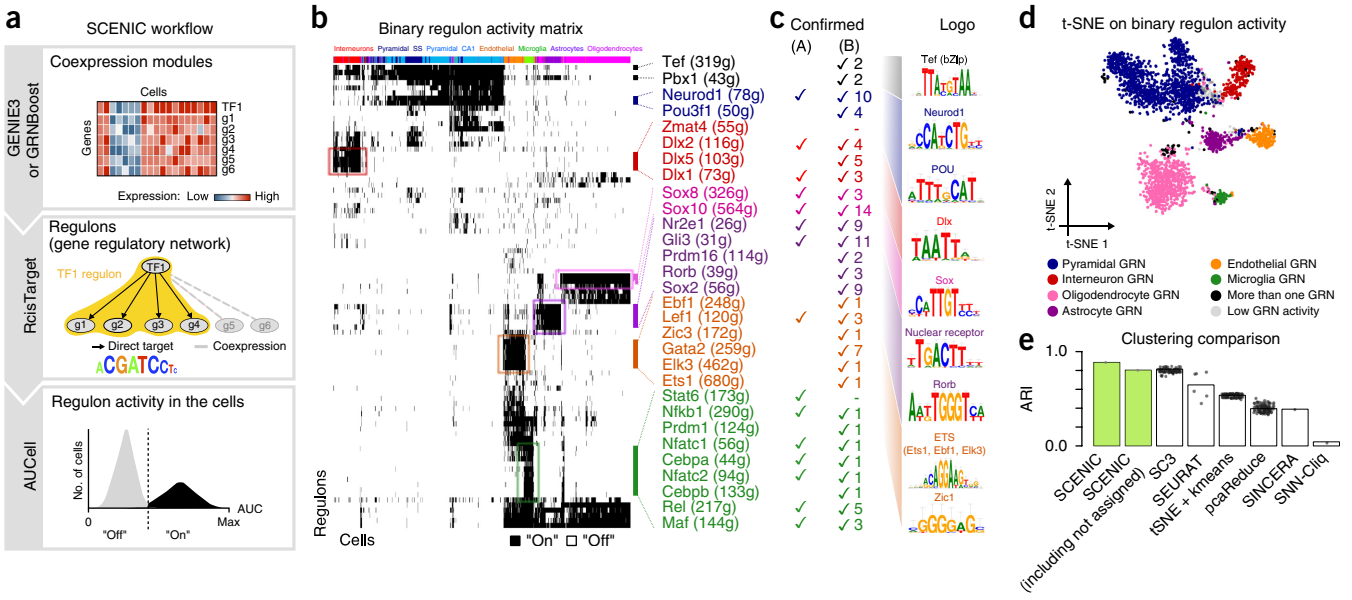

Figure 1 | The SCENIC workflow and its application to the mouse brain. (**a**) In the SCENIC workflow, coexpression modules between TFs and candidate target genes are first inferred using GENIE3 or GRNBoost. RcisTarget then identifies modules for which the regulator's binding motif is significantly enriched across the target genes and creates regulons with only direct targets. AUCell scores the activity of each regulon in each cell, thereby yielding a binarized activity matrix. The prediction of cell states is based on the shared activity of regulatory subnetworks. (**b**) SCENIC results on the mouse brain[9]. Cluster labels correspond to those used in ref. 9; master regulators are color matched with the cell types they control. (**c**) TFs confirmed by literature (A) or having brain phenotypes from Mouse Genome Informatics (B); their corresponding enriched DNA-binding motifs are shown (Logo). (**d**) t-SNE on the binary regulon activity matrix. Each cell is assigned the color of the most active GRN. (**e**) Accuracy of different clustering methods on this data set.

roles (**Fig. 1c**), and this accuracy outperforms standard analysis pipelines (**Supplementary Fig. 3e**).

To validate the Dlx1/2 network identified for mouse interneurons, we analyzed a single-nuclei RNA-seq data set of the human brain[11] (**Supplementary Fig. 6**). On the human data, SCENIC also identifies a cluster of interneurons strongly driven by DLX1/2 that has the same recognition motif as in mouse, and it identifies a set of conserved targets including DLX1 itself (**Fig. 2a**,**b**). Next, we expanded this cross-species analysis to other cell types[12]. In contrast to standard clustering based on normalized expression, which yields a strong species-driven clustering (**Supplementary Fig. 7**), the SCENIC analysis effectively grouped cells by their cell type (**Fig. 2c**). This suggests that the scoring of network activity is robust and can be exploited to overcome batch or technical effects (**Supplementary Fig. 3d**).

We also applied SCENIC to identify complex cell states in scRNA-seq data sets from oligodendroglioma[13] (4,043 cells from six tumors) and melanoma[14] (1,252 cells from 14 lesions). Because of tumor-specific mutations and complex genomic aberrations, the identification of cancer cell states is more challenging than that of normal cell states[15]. Standard clustering groups cells by their tumor of origin (**Fig. 3a**,**b**), but SCENIC reveals a different picture. For oligodendroglioma, three cancer cell states are identified across tumors (**Fig. 3c**–**e**), and each state is driven by the expected TFs—including SOX10/4/8, OLIG1/2, and ASCL1 for the oligodendrocyte-like state; SOX9, NFIB and AP-1 for the astrocyte-like state; and E2F and FOXM1 for the cycling cells.

Furthermore, applying diffusion maps to the binary SCENIC matrix (**Supplementary Fig. 8**) reconstructed a differentiation trajectory from stem-like to oligodendrocyte-like and astrocyte-like branches. Note that this path represents a different 'trajectory' compared with normal oligodendrocyte differentiation (see **Supplementary Fig. 9** for the SCENIC analysis of 5,069 oligodendrocytes). We observed a similar tumor-effect correction on the melanoma data, where SCENIC identifies groups of cells across tumors (**Supplementary Fig. 10**), including a cluster of cycling cells driven by similar TFs as in oligodendroglioma (e.g., E2F1/2/8 and MYBL2; **Fig. 3f**–**h** and **Supplementary Fig. 10**). In contrast to dedicated batch-effect removal methods such as Combat[16] and Limma[17], which require specifying the source of batch effect a priori (**Supplementary Fig. 11**), SCENIC removes the tumor effect automatically by using biologically driven features.

The melanoma cells largely fall into two groups, one corresponding to a $MITF^{high}$ state—the archetypical proliferative state—with MITF and STAT/IRF as key regulators, and one corresponding to an $MITF^{low}$ state with upregulated expression of WNT5A, LOXL2 and ZEB1—known markers of invasive states (**Supplementary Fig. 10e**,**f**). SCENIC identifies two new TFs in this $MITF^{low}$ state, NFATC2 (114 predicted target genes) and NFIB (15 predicted target genes). NFATC2, a transcriptional repressor in the JNK/MAPK pathway, is involved in melanoma dedifferentiation and immune escape[18]. NFIB, on the other hand, is linked to stem-cell behavior of hair follicle and melanocyte stem cells[19], and it plays an important role in metastatic progression of small-cell lung cancer[20].

To further explore the potential roles of NFATC2 and NFIB in the $MITF^{low}$ state, we performed immunohistochemistry on 25 melanoma specimens with varying tumor progression. We found the highest NFIB and NFATC2 expression in the sentinel lymph nodes. This colocalizes with ZEB1 expression, which suggests a relationship between the expression of these markers and the

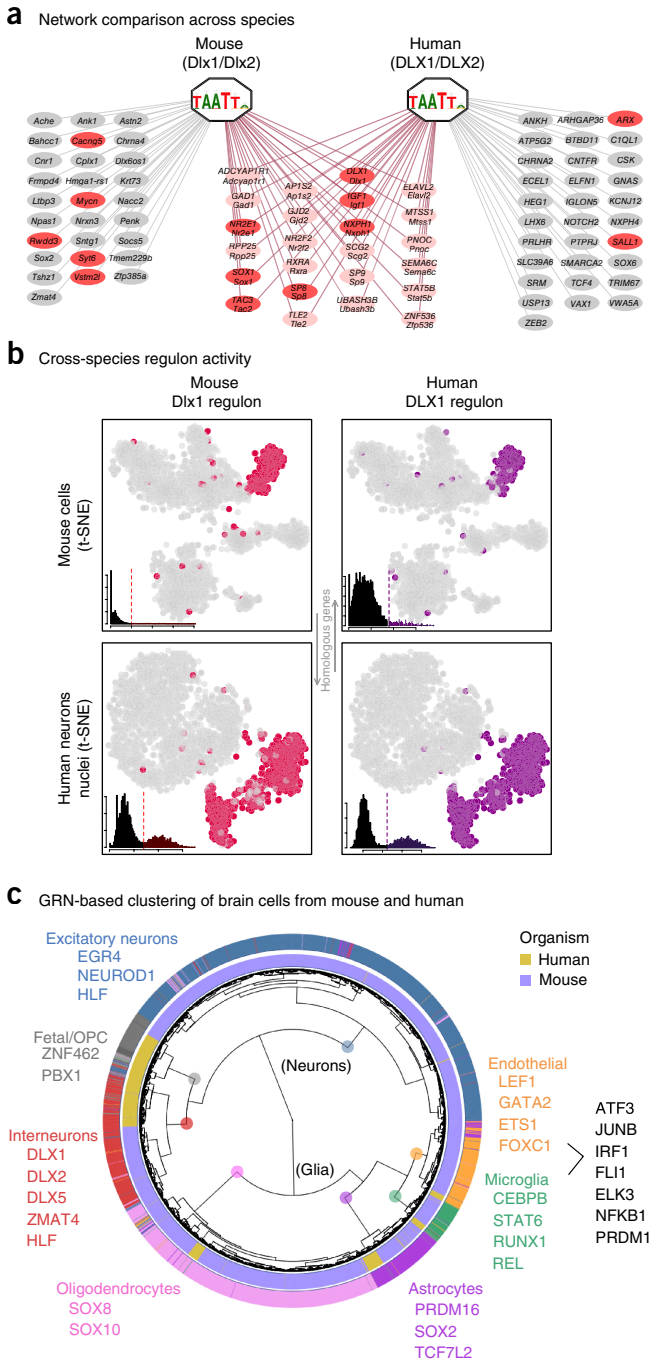

Figure 2 | Cross-species comparison of neuronal networks and cell types. (**a**) DLX1/2 regulons inferred from mouse and human brain scRNA-seq data. Genes in red have associations with Dlx1/2 in GeneMANIA. (**b**) Reciprocal activity of human and mouse Dlx1/2 regulons on mouse and human single-cell data. In each SCENIC t-SNE plot, cells are colored according to the corresponding binary regulon activity. The inset illustrates the AUCell score distribution for the regulon. (**c**) Joint clustering of human and mouse brain scRNA-seq data based on GRN activity. Colored TF names correspond to regulons identified both in the human and mouse SCENIC runs.

earliest metastatic events (**Fig. 3i** and **Supplementary Fig. 12**). When we knocked down NFATC2 using siRNA in A375, a melanoma cell line with high NFATC2 and NFIB expression (**Supplementary Fig. 13**), we found that the genes in the NFATC2 regulon were significantly upregulated (see Online Methods). This is consistent with NFATC2's previously established role as a repressor[21]. In addition, genes involved in regulation of cell adhesion and extracellular matrix and several previously published gene signatures representing the melanoma invasive state were also upregulated (**Supplementary Table 1**), which suggests that NFATC2 may indeed play an important role in disease progression. As a second validation of the melanoma regulons, we confirmed the predicted targets of MITF and STAT using ChIP-seq data (**Fig. 3j**).

As single-cell data sets increase in size, we suggest two complementary approaches to scale the network inference. The first approach is to infer the GRN from a subsampled data set and to include all cells in the scoring step with AUCell. We illustrate this approach on a data set with more than 40,000 single cells from the mouse retina (**Supplementary Fig. 14**). The second approach aims to use more efficient machine learning and big-data handling solutions. We implemented GRNBoost, a new variant of GENIE3, in Scala on Apache Spark, replacing the random-forest regression with gradient boosting. This implementation drastically reduces the time needed to infer a GRN (**Supplementary Fig. 15**) and will pave the way to network inference on very large data sets, such as the forthcoming Human Cell Atlas[22].

SCENIC is a generally applicable method for the analysis of scRNA-seq data that exploits TFs and *cis*-regulatory sequences to guide the discovery of cell states. Our results show that GRNs constitute robust guides to identify cellular states, and that scRNA-seq data are well suited to trace gene regulatory programs in which specific combinations of TFs drive cell-type-specific transcriptomes.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work is funded by The Research Foundation - Flanders (FWO; grants G.0640.13 and G.0791.14 to S. Aerts; G092916N to J.-C.M.), Special Research Fund (BOF) KU Leuven (grants PF/10/016 and OT/13/103 to S. Aerts), Foundation Against Cancer (2012-F2, 2016-070 and 2015-143 to S. Aerts) and ERC Consolidator Grant (724226_cis-CONTROL to S. Aerts). S. Aibar is supported by a PDM Postdoctoral Fellowship from the KU Leuven. Z.K.A. and J.W. are supported by postdoctoral fellowships from Kom op Tegen Kanker; V.A.H.-T. is supported by the F.R.S.-FNRS Belgium; and H.I. is supported by a PhD fellowship from the agency for Innovation by Science and Technology (IWT). Funding for T.M. and J.A. is provided by Symbiosys and IMEC HI^2 Data Science. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. T.M. would like to thank J. Simm for helpful comments and suggestions regarding gradient boosting.

AUTHOR CONTRIBUTIONS

S. Aerts and S. Aibar conceived the study; S. Aibar implemented SCENIC and related packages with help of V.A.H.-T. and P.G. for GENIE3 and G.H. for RcisTarget; S. Aibar and C.B.G.-B. analyzed the data with the help of Z.K.A. and H.I.; T.M. and J.A. implemented GRNBoost; J.W. performed the IHC and knockdown experiments; F.R., J.-C.M. and J.v.d.O. contributed reagents and helped with the interpretation of the melanoma analyses; S. Aibar, J.W. and S. Aerts wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

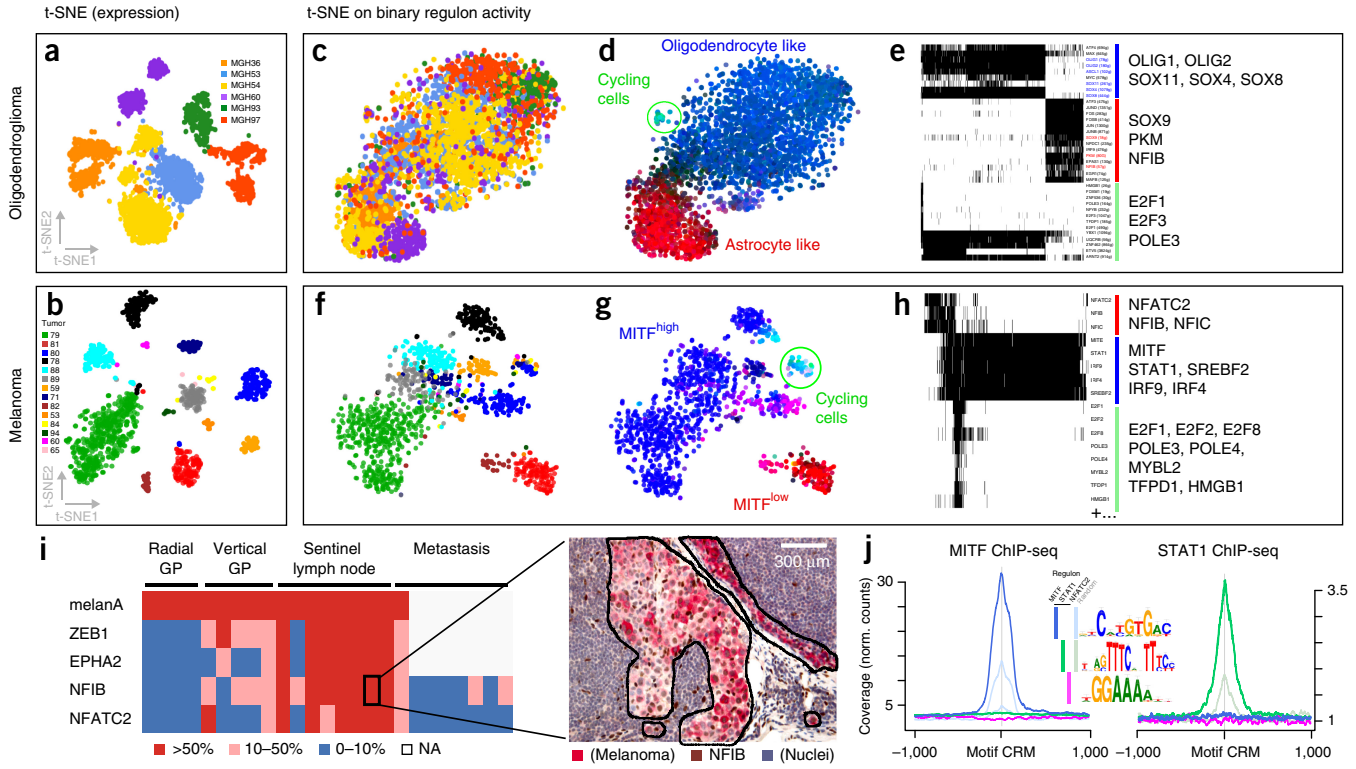

Figure 3 | SCENIC overcomes tumor effects and unravels relevant cell states and GRNs in cancer. (**a**,**b**) t-SNE plot based on the expression matrices, colored by tumor of origin. (**c**,**d** and **f**,**g**) t-SNE plots based on the binary activity matrix (**e**,**h**) after applying SCENIC. In **d** and **g**, cells are colored by GRN activity. (**i**) Immunohistochemistry (IHC) on 25 human melanomas using NFATC2, NFIB, ZEB1 and EPHA2 antibodies. The heatmap shows the percentage of cells that are positive for each marker in the given sample. Right, a representative example of IHC for NFIB on a sentinel lymph node is shown (for additional images, see **Supplementary Fig. 13**). NA, not applicable. (**j**) Aggregation plots for MITF and STAT1 ChIP-seq signal on the predicted target regions and on randomly selected genomic regions with MITF/STAT motif occurrences as a control.

1. Linnarsson, S. & Teichmann, S.A. *Genome Biol.* **17**, 97 (2016).
2. Wagner, A., Regev, A. & Yosef, N. *Nat. Biotechnol.* **34**, 1145–1160 (2016).
3. Stegle, O., Teichmann, S.A. & Marioni, J.C. *Nat. Rev. Genet.* **16**, 133–145 (2015).
4. Raj, A. & van Oudenaarden, A. *Cell* **135**, 216–226 (2008).
5. Moignard, V. *et al. Nat. Biotechnol.* **33**, 269–276 (2015).
6. Pina, C. *et al. Cell Rep.* **11**, 1503–1510 (2015).
7. Guo, M., Wang, H., Potter, S.S., Whitsett, J.A. & Xu, Y. *PLoS Comput. Biol.* **11**, e1004575 (2015).
8. Huynh-Thu, V.A., Irrthum, A., Wehenkel, L. & Geurts, P. *PLoS One* **5**, e12776 (2010).
9. Zeisel, A. *et al. Science* **347**, 1138–1142 (2015).
10. Kiselev, V.Y. *et al. Nat. Methods* **14**, 483–486 (2017).
11. Lake, B.B. *et al. Science* **352**, 1586–1590 (2016).
12. Darmanis, S. *et al. Proc. Natl. Acad. Sci. USA* **112**, 7285–7290 (2015).
13. Tirosh, I. *et al. Nature* **539**, 309–313 (2016).
14. Tirosh, I. *et al. Science* **352**, 189–196 (2016).
15. Alizadeh, A.A. *et al. Nat. Med.* **21**, 846–853 (2015).
16. Johnson, W.E., Li, C. & Rabinovic, A. *Biostatistics* **8**, 118–127 (2007).
17. Ritchie, M.E. *et al. Nucleic Acids Res.* **43**, e47 (2015).
18. Perotti, V. *et al. Oncogene* **35**, 2862–2872 (2016).
19. Chang, C.-Y. *et al. Nature* **495**, 98–102 (2013).
20. Denny, S.K. *et al. Cell* **166**, 328–342 (2016).
21. Müller, M.R. & Rao, A. *Nat. Rev. Immunol.* **10**, 645–656 (2010).
22. Regev, A. *et al. bioRxiv* Preprint at: http://www.biorxiv.org/content/early/2017/05/08/121202 (2017).

ONLINE METHODS

SCENIC workflow. SCENIC is a workflow based on three new R/bioconductor packages: (i) GENIE3, to identify potential TF targets based on coexpression; (ii) RcisTarget, to perform the TF-motif enrichment analysis and identify the direct targets (regulons); and (iii) AUCell, to score the activity of regulons (or other gene sets) on single cells. We also provide GRNBoost, implemented on Spark[23], as scalable alternative to build the coexpression network on bigger data sets (step i, replacing GENIE3).

The three R/bioconductor packages, and GRNBoost, include detailed tutorials to facilitate their use within an automated SCENIC pipeline, as well as independent tools. Links to the tools, SCENIC code and tutorials are available at http://scenic.aertslab.org.

GENIE3. GENIE3 (ref. 8) is a method for inferring gene regulatory networks from gene expression data. In brief, it trains random forest models predicting the expression of each gene in the data set and uses as input the expression of the TFs. The different models are then used to derive weights for the TFs, measuring their respective relevance for the prediction of the expression of each target gene. The highest weights can be translated into TF-target regulatory links[8]. Since GENIE3 uses random-forest regression, it has the added value of allowing complex (e.g., non-linear) coexpression relationships between a TF and its candidate targets. GENIE3 is available in Python, Matlab and R. To allow for inclusion in SCENIC workflow, we optimized the previous R implementation of GENIE3. The core of this new implementation is now written in C (which makes it orders of magnitude faster), it requires lower memory, and it supports execution in parallel. GENIE3 was the top-performing method for network inference in the DREAM4 and DREAM5 challenges[24]. The new package provides similar results in the DREAM challenge to those of previously existing implementations, but with improved speed. The comparison is available at the following website: http://www.montefiore.ulg.ac.be/~huynh-thu/GENIE3.html.

The input to GENIE3 is an expression matrix. The preferred expression values are gene-summarized counts (which might or might not use unique molecular identifiers, UMIs[25]). Other measurements, such as counts or transcripts per million (TPM) and FPKM/RPKM are also accepted as input. However, note that the first network-inference step is based on coexpression, and some authors recommend avoiding within-sample normalizations (i.e., TPM) for this task because they may induce artificial covariation[26]. To evaluate to what extent the normalization of the input matrix affects the output of SCENIC, we also ran SCENIC on the Zeisel *et al.*[9] data set after library-size normalization (using the standard pipeline from scran[27], which performs within-cluster size-factor normalization). The results are highly comparable, both in regards to resulting clusters or cell types (ARI between the cell types obtained from raw UMI counts or normalized counts: 0.90, ARI from normalized counts compared to the author's cell types: 0.87) and to the TFs identifying the groups (26 out of the 30 regulons highlighted in **Fig. 1b**). Furthermore, during the course of this project we have applied GENIE3 to multiple data sets, some of them having UMI counts (e.g., mouse brain and oligodendrocytes) and others TPM (e.g., human brain and melanoma), and both units provided reliable results.

The output of GENIE3 is a table with the genes, the potential regulators, and their 'importance measure' (IM), which represents the weight that the TF (input gene) has in the prediction of the target. We explored several ways to determine the threshold (e.g., looking at the rankings, distributions and outputs after pruning with RcisTarget) and finally opted for building multiple gene sets of potential targets for each TF: (i) setting several IM thresholds (IM > 0.001 and IM > 0.005), (ii) taking the 50 targets with highest IM for each TF and (iii) keeping only the top 5, 10 and 50 TFs for each target gene (then, split by TF). In all these cases, only the links with IM > 0.001 were taken into account. Furthermore, each gene set was then split into positive- and negative-correlated targets (i.e. Spearman correlation between the TF and the potential target) to separate likely activated and repressed targets. Finally, only the gene sets (TF coexpression modules) with at least 20 genes were kept for the following step.

GRNBoost. GRNBoost is based on the same concept as GENIE3: inferring regulators for each target gene purely from the gene expression matrix. However, GRNBoost does so using the gradient-boosting machines (GBM)[28] implementation from the XGBoost library[29]. A GBM is an ensemble learning algorithm that uses boosting[30] as a strategy to combine multiple weak learners, like shallow trees, into a strong one. This contrasts with random forest, the method used by GENIE3, which uses bagging (bootstrap aggregation) for model averaging to improve regression accuracy. GRNBoost uses gradient-boosted stumps (regression trees of depth 1)[31] as the base learner. GRNBoost's main contribution is casting this multiple regression approach into a Map/Reduce[32] framework based on Apache Spark[23]. In GRNBoost, the core data entry is a tuple of a gene name and a vector of gene expression values. Using a Spark RDD, GRNBoost first partitions the gene expression vectors over the nodes available in the compute cluster. Subsequently, it constructs a predictor matrix that contains the expression values for all candidate regulator genes. Using a Spark broadcast variable, the predictor matrix is broadcasted to the different compute partitions. In the map phase of the framework, GRNBoost iterates over the gene tuples (expression vector) and uses the predictor matrix to train the XGBoost regression models with the expression vectors as respective training labels. From the trained models, the strengths of the regulator–target relationships are extracted and emitted as a set of network edges. In the reduce phase, all sets of edges are combined into the final regulatory network.

The performance of GRNBoost and GENIE3 was compared on a workstation with 2 Intel Xeon E2696 V4 CPUs with, in total, 44 physical cores or 88 threads and 128 GB of 2133Ghz ECC memory. Large data sets and hence large predictor matrices cause the network inference to become memory bound rather than CPU bound. In order to comfortably fit the amount of memory required into the available 128 GB of memory, we decreased the number of partitions to 11, therefore having a maximum of only 11 predictor matrices in flight simultaneously. However, we increased the number of threads available to each individual XGBoost regression to 8, effectively using all available (88) threads in the workstation. GRNBoost is written in the Scala programming language and can be used as a software library or be submitted as a Spark job from the command line.

RcisTarget. RcisTarget is a new R/Bioconductor implementation of the motif enrichment framework of i-cisTarget and iRegulon.

RcisTarget identifies enriched TF-binding motifs and candidate transcription factors for a gene list. In brief, RcisTarget is based on two steps. First, it selects DNA motifs that are significantly over-represented in the surroundings of the transcription start site (TSS) of the genes in the gene set. This is achieved by applying a recovery-based method on a database that contains genome-wide cross-species rankings for each motif. The motifs that are annotated to the corresponding TF and obtain a normalized enrichment score (NES) > 3.0 are retained. Next, for each motif and gene set, RcisTarget predicts candidate target genes (i.e., genes in the gene set that are ranked above the leading edge). This method is based on the approach described by Aerts *et al.*[33], which is also implemented in i-cisTarget (web interface)[34] and iRegulon (Cytoscape plugin)[35]. Therefore, when using the same parameters and databases, RcisTarget provides the same results as i-cisTarget or iRegulon, benchmarked against other TFBS-enrichment tools in Janky *et al.*[35]. More details about the method and its implementation in R are given in the package documentation.

To build the final regulons, we merge the predicted target genes of each TF module that show enrichment of any motif of the given TF. To detect repression, it is theoretically possible to follow the same approach with the negative-correlated TF modules. However, in the data sets we analyzed, these modules were less numerous and showed very low motif enrichment. For this reason, we finally decided to exclude the detection of direct repression from the workflow and continue only with the positive-correlated targets. The databases used for the analyses presented in this paper are the "18k motif collection" from iRegulon (gene-based motif rankings) for human and mouse. For each species, we used two gene-motif rankings (10 kb around the TSS or 500 bp upstream the TSS), which determine the search space around the transcTSS.

AUCell. AUCell is a new method that allows researchers to identify cells with active gene regulatory networks in single-cell RNA-seq data. The input to AUCell is a gene set, and the output is the gene set 'activity' in each cell. In SCENIC, these gene sets are the regulons, which consist of the TFs and their putative targets. AUCell calculates the enrichment of the regulon as an area under the recovery curve (AUC) across the ranking of all genes in a particular cell, whereby genes are ranked by their expression value. This method is therefore independent of the gene expression units and the normalization procedure. In addition, since the cells are evaluated individually, it can easily be applied to bigger data sets (e.g., subsetting the expression matrix if needed). In brief, the scoring method is based on a recovery analysis where the *x*-axis (**Supplementary Fig. 1c**) is the ranking of all genes based on expression level (genes with the same expression value, e.g., '0', are randomly sorted); and the *y*-axis is the number of genes recovered from the input set. AUCell then uses the AUC to calculate whether a critical subset of the input gene set is enriched at the top of the ranking for each cell. In this way, the AUC represents the proportion of expressed genes in the signature and their relative expression values compared to the other genes within the cell. The output of this step is a matrix with the AUC score for each gene set in each cell. We use either the AUC scores (across regulons) directly as continuous values to cluster single cells, or we generate a binary matrix using a cutoff of the AUC score for each regulon. These cutoffs are either determined automatically, or they are manually adjusted by inspecting the distribution of the AUC scores. Some examples of AUC distributions are provided in **Supplementary Figure 2a**. **Supplementary Figure 2b,c** shows the validation of AUCell using previously published neuronal and glial gene signatures. The tutorial included in the package also includes practical explanations and implications of each of the steps of the method.

Cell clustering based on gene regulatory networks. The cell regulon activity is summarized in a matrix in which the columns represent the cells and the rows the regulons. In the binary regulon activity matrix, the coordinates of the matrix that correspond to active regulons in a given cell will contain a "1," and "0" otherwise. The equivalent matrix, which contains the continuous AUC values for each cell regulon, is normally referred to as the AUC activity matrix. Clustering of either of the regulon activity matrices reveals groups of regulons (jointly, a network) that are recurrently active across a subset of cells. The binary activity matrix tends to highlight higher order similarities across cells (and therefore highly reduces batch effects and technical biases); on the other hand, the AUC matrix allows researchers to observe more subtle changes. For visualization, we have mostly used t-SNEs (Rtsne package[36], we always tested consistency across several perplexity values and distance metrics/number of PCs), and heatmaps with hierarchical clustering (although the heatmap figures feature selected regulons, the t-SNEs are always run on the whole matrices). In the tutorials, we have also included several options to explore the results. For example, how to detect most likely stable states (higher density areas in the t-SNE), and to help identify key regulators, known cell properties (based on the data set annotation) and GO terms (GO enrichment analysis of the genes in the cluster of regulons) that might be associated to the detected states.

SCENIC runs on the different data sets. SCENIC was run on all the data sets using the expression matrices provided by the authors (downloaded from GEO or the authors' website), including only the cells that passed their quality control, and the default gene filtering for GENIE3 (which in all these data sets resulted in 12,000–15,000 genes). The standard SCENIC workflow was run on all data sets (the package versions at the time of publication are available as **Supplementary Software**, updated versions will be posted at http://scenic.aertslab.org). A more detailed description of the data sets and the any peculiarities for each analysis are available in **Supplementary Note 1**. Here we provide a brief description of the data sets:

Mouse cortex and hippocampus. Single-cell RNA-seq of 3005 brain cells of juvenile mice (21–31 d old). It contains the main cell types in hippocampus and somatosensory cortex, namely neurons (pyramidal excitatory neurons, and interneurons), glia (astrocytes, oligodendrocytes, microglia), and endothelial cells. Expression matrix units: UMI counts. (Zeisel *et al.*[9], GSE60361)

Human neurons. Single-nuclei RNA-seq of 3,083 neuronal cells from a normal human brain (retrieved postmortem from a 51 year old female, from six different Brodmann areas). Expression matrix units: TPM. (Lake *et al.*[11])

Human brain. scRNA-seq from 466 cells from adult and fetal human brains. The fetal samples were taken from four different individuals at 16 to 18 weeks postgestation. The adult brain

doi:10.1038/nmeth.4463

samples were taken from healthy temporal lobe tissue from eight different patients (21–63 years old) during temporal lobectomy surgery for refractory epilepsy and hippocampal sclerosis. Expression matrix units: logged CPM. (Darmanis *et al.*[12], GSE67835)

Mouse oligodendrocytes. scRNA-seq data of 5,069 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 axes of the mouse juvenile and adult CNS, including white and gray matter. Expression matrix units: UMI counts. (Marques *et al.*[37], GSE75330)

Oligodendroglioma. scRNA-seq expression profiles for 4,347 cells from six untreated grade II oligodendroglioma tumors with either IDH1 or IDH2 mutation, and 1p/19q codeletion. Only the tumoral cells were used for the analysis (selected by the authors based on CNV profile). Expression matrix units, $\log_2(TPM + 1)$. (Tirosh *et al.*[13], GSE70630)

Melanoma. scRNA-seq of 1,252 melanoma cells from 14 different tumors. These include only the cells that are labeled as malignant by the authors, based on their CNV profiles. Expression matrix units: $\log_2(TPM/10 + 1)$. (Tirosh *et al.*[14], GSE72056)

Mouse retina. scRNA-seq data of 44,808 cells obtained through Drop-seq from mouse retina (14 d postnatal). Expression matrix units: log((UMI counts per gene in a cell/total UMI counts in cell) × 10,000) + 1). (Macosko *et al.*[38], GSE63472)

Embryonic mouse brain. Chronium Megacell demonstration data set containing 1,306,127 cells from cortex, hippocampus and subventricular zone of two E18 mice (strain: C57BL/6). (10X Genomics)

Gene filtering. For gene filtering to run GENIE3, we applied a soft filter based on the total number of counts of the gene and the number of cells in which it is detected. The first filter, the total number of reads per gene, is meant to remove genes that are most likely unreliable and provide only noise. The specific value depends on the data set; for the ones used in this paper we set the thresholds at, for example, 3 UMI counts (slightly over the median of the nonzero values) multiplied by 1% of the number of cells in the data set (e.g., in mouse brain: 3 UMI counts × 30 (1% of cells) = minimum 90 counts per gene). The second filter, the number of cells in which the gene is detected (e.g., >0 UMI, or >1 $\log_2(TPM)$), is to remove genes that are only expressed in one or very few cells (they would gain a lot of weight if they happen to coincide in a given cell). In the workflow, we recommend to set the second filtering lower than the smallest population of cells to be detected. For example, since microglia cells represent approximately 3% of the total cells in the data set, we used a detection threshold of at least 1% of the cells.

Cross-species network comparisons. SCENIC was run independently for each of the three data sets used for the GRN comparison: Zeisel *et al.*[9] (mouse brain cells), Lake *et al.*[11] (human neurons nuclei) and Darmanis *et al.*[12] (human brain cells). To compare the networks across species, the genes in the human regulons were converted into the homologous mouse genes using Biomart (through biomaRt R package[39]) and vice versa (the mouse regulons into human genes). In **Figure 2a**, the genes highlighted in red also have associations with Dlx1/2 in GeneMANIA[40] (protein–protein interactions, genetic interactions, coexpression, or literature comentioning).

For the cross-species cell clustering (**Fig. 2c**), the genes in the mouse expression matrix were converted into the homologous human genes and merged with the Darmanis *et al.*[12] expression matrix by row (only genes available in both matrices were kept). The 259 human regulons from the Darmanis *et al.*[12] data set and the human homologs of the mouse regulons were evaluated on this merged matrix to obtain the binary regulon activity containing 410 regulons. The cells were clustered based on the binary activity matrix using Ward's hierarchical clustering with Spearman's distance. Similar results were obtained for the reverse approach (converting the expression matrix into mouse genes to evaluate the mouse regulons). In order to provide an alternative approach based only on expression (**Supplementary Fig. 7**), we also generated a merged expression matrix. Since the merged data sets use different measurement units (CPM in human and UMI in mouse), each matrix was Z-score normalized by gene before merging.

Method comparisons. We performed different evaluations and benchmark comparisons, each assessing a different aspect of SCENIC (e.g., cell type identification, TF identification, cofounding effect correction). The detailed description of how these comparisons were performed is available in **Supplementary Note 1**. Here we provide a brief summary:

Cell clustering. To determine whether the clustering based on gene regulatory network activity matches real cell types, we compared the clustering based on the regulon activity matrices to the cell labels provided in the corresponding publications. To compare SCENIC performance to other methods, we reused the benchmark presented in the SC3 publication[10], which provides the adjusted Rand index (ARI) for six clustering methods on the mouse brain data set.

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. 1e**). However, we also compared SCENIC to an alternative approach to identify TFs potentially regulating cell states—applying TF motif enrichment analysis on genes differentially expressed between clusters (i.e., gene signature or markers for a cell type).

Batch effect correction. The results of SCENIC on the oligodendroglioma data set (clustering of the full binary regulon activity matrix) were compared to Combat[16,41] and Limma[17,42], correcting for 'patient of origin' as source of batch effect.

Cycling cells. The cycling cells were predicted based on consistent upregulation of 46 gene sets related to the mitotic cell cycle from amiGO and cycleBase 1.0 and 2.0. We then compared the ability of different clustering methods to identify these cells (sensitivity and specificity). Since most of the methods provide multiple clusters as output, to compare their results, for each method we selected the cluster with the largest amount of CC cells.

Immunohistochemistry of melanoma biopsies. Immunohistochemistry with antibodies for melanA, EPHA2, ZEB1, NFATC2 and NFIB was performed on formalin-fixed, paraffin-embedded melanoma samples. The samples include biopsies of nine primary melanomas (four in radial growth phase and five in vertical growth phase), eight melanoma-containing sentinel lymph nodes, and eight melanoma metastases. A detailed description of how the immunohistochemistry was performed, as well as the antibodies used, is available in **Supplementary Note 1**.

Knockdown of *NFATC2* in melanoma cell culture. The A375 cell line was selected as representative of the MITF^{low} state based on expression of *NFATC2*, *NFIB* (**Supplementary Fig. 13**) and *SOX10* after comparing 59 melanoma cell lines from the COSMIC Cancer Cell lines Project[43]. Knockdown of *NFATC2* was performed in A375 using *NFATC2* siRNA, and total RNA was extracted 72 h after the knockdown. The final libraries were pooled and sequenced on a combination of NextSeq 500 and HiSeq 4000 (Illumina). RNA-seq reads were mapped to the genome (hg19) for upstream analysis. A detailed description of the methods, including cell line source, knockdown of *NFATC2*, RNA-seq protocol and bioinformatics analysis, are available in **Supplementary Note 1**.

Code availability. Updated links to the packages and tutorials related to SCENIC are available at http://scenic.aertslab.org; the package versions at the time of publication are provided as **Supplementary Software**.

Data availability statement. The *NFATC2* knockdown RNA-seq data have been deposited in NCBI's Gene Expression Omnibus[44] and are accessible through GEO accession number GSE99466. Source data files for **Figures 1–3** are available online.

A **Life Sciences Reporting Summary** is available.

23. Zaharia, M. *et al.* In *Proc. of the 9th USENIX Conference on Networked Systems Design and Implementation* 2–2 (USENIX Association, 2012).
24. Marbach, D. *et al. Nat. Methods* **9**, 796–804 (2012).
25. Islam, S. *et al. Nat. Methods* **11**, 163–166 (2014).
26. Crow, M., Paul, A., Ballouz, S., Huang, Z.J. & Gillis, J. *Genome Biol.* **17**, 101 (2016).
27. Lun, A.T.L., McCarthy, D.J. & Marioni, J.C. *F1000Res.* **5**, 2122 (2016).
28. Friedman, J.H. *Ann. Stat.* **29**, 1189–1232 (2001).
29. Chen, T. & Guestrin, C. In *Proc.of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016).
30. Freund, Y. & Schapire, R.E. *Jinko Chino Gakkaishi* **14**, 771–780 (1999).
31. Sławek, J. & Arodź, T. *BMC Syst. Biol.* **7**, 106 (2013).
32. Dean, J. & Ghemawat, S. *Commun. ACM* **51**, 107–113 (2008).
33. Aerts, S. *et al. PLoS Biol.* **8**, e1000435 (2010).
34. Herrmann, C., Van de Sande, B., Potier, D. & Aerts, S. *Nucleic Acids Res.* **40**, e114 (2012).
35. Janky, R. *et al. PLoS Comput. Biol.* **10**, e1003731 (2014).
36. Krijthe, J. Rtsne: t-distributed stochastic neighbor embedding using Barnes-Hut implementation https://github.com/jkrijthe/Rtsne (2015).
37. Marques, S. *et al. Science* **352**, 1326–1329 (2016).
38. Macosko, E.Z. *et al. Cell* **161**, 1202–1214 (2015).
39. Durinck, S. *et al. Bioinformatics* **21**, 3439–3440 (2005).
40. Warde-Farley, D. *et al. Nucleic Acids Res.* **38**, W214–W220 (2010).
41. Leek, J. sva: Surrogate Variable Analysis. R package version 3.24.4 (2017).
42. Smyth, G. limma: Linear models for microarray data. (2015).
43. Forbes, S.A. *et al. Nucleic Acids Res.* **45**, D777–D783 (2017).
44. Edgar, R., Domrachev, M. & Lash, A.E. *Nucleic Acids Res.* **30**, 207–210 (2002).

doi:10.1038/nmeth.4463

nature research

Corresponding author(s): Stein Aerts

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

SCENIC analyses: We show the application of SCENIC to 8 datasets. These datasets were selected to cover different case studies: clearly defined "static" cell types (mouse brain), developmental process (mouse oligodendrocytes, this dataset was selected among the multiple developmental datasets for comparison with the previous analysis), cross-species comparison (2 x human brain), and cancer (melanoma and oligodendroglioma). All these datasets range between 1k-5k cells. In addition, we included an sparse dataset (49k mouse retina cells with Drop-seq), and a larger dataset (Megacell demonstration).

IHC: We selected ~5 melanoma samples per category based on availability of human specimens and performed immunohistochemical stainings on four radial growth phase primary melanomas, five vertical growth phase primary melanomas, eight sentinel lymph node metastases and eight full-blown metastases (see manuscript for results).

RNA-seq on NFATC2 knock-down: We performed NFATC2 knock down on A375 cells with one replicate for each category, as we used a whole-genome, ranked list of differentially expressed genes (see below). Two replicates of NFATC2 knock down, which were sequenced at lower coverage (~two million high quality reads), reproduced the original findings reliably (data not shown in the manuscript).

2. Data exclusions

Describe any data exclusions.

SCENIC analyses: No data was excluded from the analyses. As the analysis was performed on public datasets, we used the cells selected by the authors. Any further selection is described in the methods (e.g. oligodendroglioma: CNV, mouse retina sub-sampling, and mouse brain sub-sampling).

IHC, and RNA-seq on NFATC2 knock-down: No data were excluded from the analyses.

3. Replication

Describe whether the experimental findings were reliably reproduced.

SCENIC analyses: SCENIC reliably identified the expected cell types (plus some novel cell types) in all analysed datasets. The computational replications (sub-sampling) also provided reproducible results (Supplementary Figure 5 and 15).

IHC: Four radial growth phase primary melanomas, five vertical growth phase primary melanomas, eight sentinel lymph node metastases and eight full-blown metastases were stained (see manuscript for results).

RNA-seq on NFATC2 knock-down: Two replicates of NFATC2 knock down, which were sequenced at lower coverage (~two million high quality reads), reproduced the original findings reliably (data not shown in the manuscript).

4. Randomization

Describe how samples/organisms/participants were

SCENIC analyses: Not relevant. Each analysis was independent.

Nature Methods: doi:10.1038/nmeth.4463

allocated into experimental groups.

IHC: not relevant. Groups are determined by clinical diagnosis.

RNA-seq on NFATC2 knock-down: not relevant. The melanoma cell line was selected based on high levels of NFATC2 in the COSMIC panel of cell lines.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

SCENIC analyses: The analyses were performed on the expression matrix alone, without taking into account the cell types or any other phenotypic information provided by the authors of the dataset. Only at the end of the analyses, for validation, the cell-type/phenotypic data was compared with the clusters provided by SCENIC.

IHC: The stainings were blinded to the pathologist scoring them.

RNA-seq on NFATC2 knock-down: Not relevant.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☒	☐ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☒	☐ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

SCENIC algorithm: This paper presents a new algorithm SCENIC, three new R-packages (GENIE3, RcisTarget and AUCell) which were required for its implementation, and GRNboost as scalable alternative for GENIE3. All of them are described in the methods, and their implementation in R is available in Github. The R packages are also provided as supplementary code, and links to new versions will be kept at the authors website (http://scenic.aertslab.org).

SCENIC analyses: The analyses presented in the paper were run using the development version of the packages. These versions are provided as supplmentary code (only the interface has changed across different versions): SCENIC 0.1.5 (17 jul 2017), AUCell 0.99.5 (7 jun 2017), RcisTarget 0.99.0 (7 jun 2017), and GENIE3 0.99.3. The analyses were run using the 18k-motif databases for RcisTarget (Human: RcisTarget.hg19.motifDatabases_0.99.0, and Mouse: RcisTarget.mm9.motifDatabases_0.99.0).

Complementary analyses of public datasets:
- R version 3.3.2 and packages corresponding to Bioconductor version 3.4.
- Benchmarks: Homer (version 4.9), Seurat (version 1).
- Gene-set enrichment analysis: GSEA (version 2.0) GeneMANIA (accessed: oct. 2016), amigo, cycleBase (1.0 and 2.0).

Nature Methods: doi:10.1038/nmeth.4463

RNAseq on NFATC2 knock-down:
- fastq-mcf (as part of ea utils; version 1.1.2-686): default parameters using a list containing the common Illumina adapters; to trim adapter sequences from the raw reads.
- FastQC from Babraham Bioinformatics: for quality control of trimmed reads.
- STAR (version 2.5.1b-foss-2014a): to map the reads to the human refseq hg19 genome.
- SAMtools (version1.4-foss-2014a): to filter reads for -q4 quality only
- HTSEQ-count (version 0.6.1p1): to count the number of reads for each gene
- DESeq2 (version 1.14.1) from Bioconductor used in R-studio: to obtain a list of differentially expressed genes, ranked based on the Log2FC of up or down regulation.
- GOrilla (cbl-gorilla.cs.technion.ac.il/): an online tool to identify of enriched Gene Ontology categories, based on a whole-genome, ranked list of differentially expressed genes.
- GSEA (version 2.0), using the GSEAPreranked function: a Java Desktop Application to assess potential enrichment of gene sets in a whole-genome, ranked list of differentially expressed genes.

IHC: not relevant.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	N/A

Nature Methods: doi:10.1038/nmeth.4463

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

IHC was performed on a Leica BOND-MAX automatic immunostainer (Leica Microsystems). Antigen retrieval was performed on-board using a citrate-based (Bond Epitope Retrieval Solution 1, pH 6.0; Leica) or an EDTA-based (Bond Epitope Retrieval Solution 2, pH 9.0; Leica) buffer according to the manufacturer's instructions (see below).

Primary antibodies for stainings:
- rabbit polyclonal anti-NFIB (Sigma-Aldrich; HPA003956; pH6.0): antibody was validated extensively by the Human Protein Atlas for selectivity/specificity and for its use for immunohistochemical stainings. In-house the staining conditions were validated on human pancreas sections.
- rabbit monoclonal anti-NFATC2 (Cell Signaling Technology; #5861; pH6.0): antibody was validated by the supplier for selectivity/specificity and for its use for immunohistochemical stainings. In-house the staining conditions were validated on human lymph node sections.
- rabbit polyclonal anti-ZEB1 (Santa Cruz; sc-25388; pH9.0): antibody was validated by the supplier for selectivity/specificity, and its use for immunohistochemical stainings on human melanomas was validated by Caramel and colleagues (caramel et al., Cancer Cell, 2013). In-house the staining conditions were validated on human melanoma sections.
- rabbit monoclonal anti-EPHA2 (Cell Signaling Technology; #6997; pH9.0): antibody was validated by the supplier for selectivity/specificity and for its use for immunohistochemical stainings. In-house the staining conditions were validated on human melanoma sections.
- mouse monoclonal anti-melanA (DAKO; IR633; pH9.0): antibody was validated by the supplier for selectivity/specificity and for its use for immunohistochemical stainings. In-house the staining conditions were validated on human melanoma sections.

Secondary antibodies for stainings:
- for brown visualization: Bond Polymer Refine Detection kit (Leica)
- for red/pink visualization: Bond Polymer Refine Red Detection (Leica)

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

The A375 melanoma cell line was kindly provided by a collaborator (Professor Lionel Larue, Institut Curie, Paris)

b. Describe the method of cell line authentication used.

A375 were authenticated by verifying the presence of the three mutations in A375 cells according the ATCC, namely BRAF homozygous c.1799T>A (p.V600E), CDKN2A homozygous c.181G>T (p.E61*) and CDKN2A homozygous c.205G>T (p.E69*).

c. Report whether the cell lines were tested for mycoplasma contamination.

Cell line A375 was tested regularly for mycoplasm contamination. Results were negative.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

N/A

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

N/A

Nature Methods: doi:10.1038/nmeth.4463

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

For IHC, we selected melanoma samples (~5) for each clinicopathologic category without knowledge of the age or gender of the patient, but based on availability of human specimens and performed immunohistochemical stainings on four radial growth phase primary melanomas, five vertical growth phase primary melanomas, eight sentinel lymph node metastases and eight full-blown metastases (see manuscript for results).
The IHC experiments have been approved by the Medical Ethical Committee and Institutional Review Board (OG032) of the University Hospitals of KU Leuven (BioMel; Belgian reference number B322201524395), and by the UZ Leuven Biobank (reference number S57760).
The RNAseq experiments have also been approved by the Medical Ethical Committee and Institutional Review Board (OG032) of the University Hospitals of KU Leuven (ML10660; Belgian reference number B322201421305), and by the UZ Leuven Biobank (reference number S56777).
The entire study conformed to the World Medical Association Declaration of Helsinki.

Nature Methods: doi:10.1038/nmeth.4463