

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.

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 supplementary 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).

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 (version 1.4-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.
- GOrrilla (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

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 mycoplasma 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

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.