Supplementary Figures and Text for "*destiny* – diffusion maps for single-cell time-course data"

Philipp Angerer [1], Laleh Haghverdi [1], Maren Büttner [1], Fabian J. Theis [1,2], Carsten Marr [1,*], and Florian Buettner [1,†,*]

[1]Helmholtz Zentrum München - German Research Center for Environmental Health, Institute of Computational Biology, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany
[2]Technische Universität München, Center for Mathematics, Chair of Mathematical Modeling of Biological Systems, Boltzmannstr. 3, 85748 Garching, Germany

SUPPLEMENTARY TEXT

S1 Preprocessing

The data properties required by destiny include a variance not spanning many orders of magnitude due to the global σ. Therefore, RNAseq count data should be adapted using a variance-stabilizing transformation like a logarithm or the square root (**stegle_computational_2015**) For single-cell qPCR data, expression values can be normalized by dividing (or subtracting in the case of logarithmic values) by housekeeper gene expression levels. However, as the expression of housekeeping genes is also stochastic, it is not clear whether such normalization is beneficial. Consequently, if such housekeeping normalization is performed, it is crucial to take the mean of several genes.

S2 Parameter selection

For the choice of the Gaussian kernel width, we pick a σ close to the intrinsic dimensionality of data. Such choice of the kernel width ensures a maximum connectivity of the data point as a graph while restricting the use of euclidean distances in the Gaussian kernel to local proximities where such distances are valid. How the intrinsic dimensionality of data can be approximated has been shown in **haghverdi_diffusion_2015**

If the nearest neighbors approximation is used, the input parameter k controls the number of nearest neighbours for each cell to be considered. Guideline for k is a small enough number to make the computation cost limited, but not too small to alter the connectivity of data as a graph, which would result in a noisy embedding. A typical k is between 200 and 1000 cells.

S3 Projection of new data

Projection of new data onto existing diffusion components is done by first computing the transition probability matrix M' from new data points to the existing data points. The same transformation that brought the existing data points' transition matrix M to the low dimensional embedding, $M \times C$, is applied to the newly built $(n_{new} \times n_{init})$ transition matrix M', $M' \times C$.

This way we provide an approximation for the diffusion distances of new points to all initial points. This approach is however inefficient if newly introduced points are too distant to any part of the initially existing population, since in a better approximation the new points will alter the position of initial points on the low dimensional map as well (**homrighausen_spectral_2011**)

SUPPLEMENTARY FIGURES AND TABLES

	guo_resolution_2010 (429 cells × 48 genes)	**guo_resolution_2010** (429 cells × 48 genes) accounting for censored & missing values	**moignard_decoding_2015** (3934 cells × 46 genes)	**trapnell_dynamics_2014** (271 cells × 20 SVs)	**zunder_continuous_2015** (256k cells × 36 marker)
MATLAB fast	0.3s	NA	4.3s	0.6s	>24h
MATLAB	21s	94s	>1h	4.7s	NA (out of memory)
R *destiny*	0.5s	1.4s	6.0s (k=500)	0.2s	1.4h (k=1000)
R diffusionMap	0.4s	NA	21s	0.2s	NA (out of memory)

Table S1. Runtime comparison of *destiny* with previously available diffusion map implementations for three different single cell data sets. *destiny* utilizes sanity checking and imputing, slightly increasing runtime with respect to a fast MATLAB implementation. The fast MATLAB version does not include the censoring and missing value models. The data from **guo_resolution_2010** **moignard_decoding_2015** and **trapnell_dynamics_2014** have been processed on a 2GHz laptop with 8GB memory. As explained in S4, the latter was created using 20 singular values (SVs) of the data. The data from **zunder_continuous_2015**

† *Current address:* European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, CB10 1SD, UK

have been processed on a 1.4GHz server with 954GB memory. The measurements for diffusionMap are always done with full distance matrices due to a lack of nearest neighbor approximation.

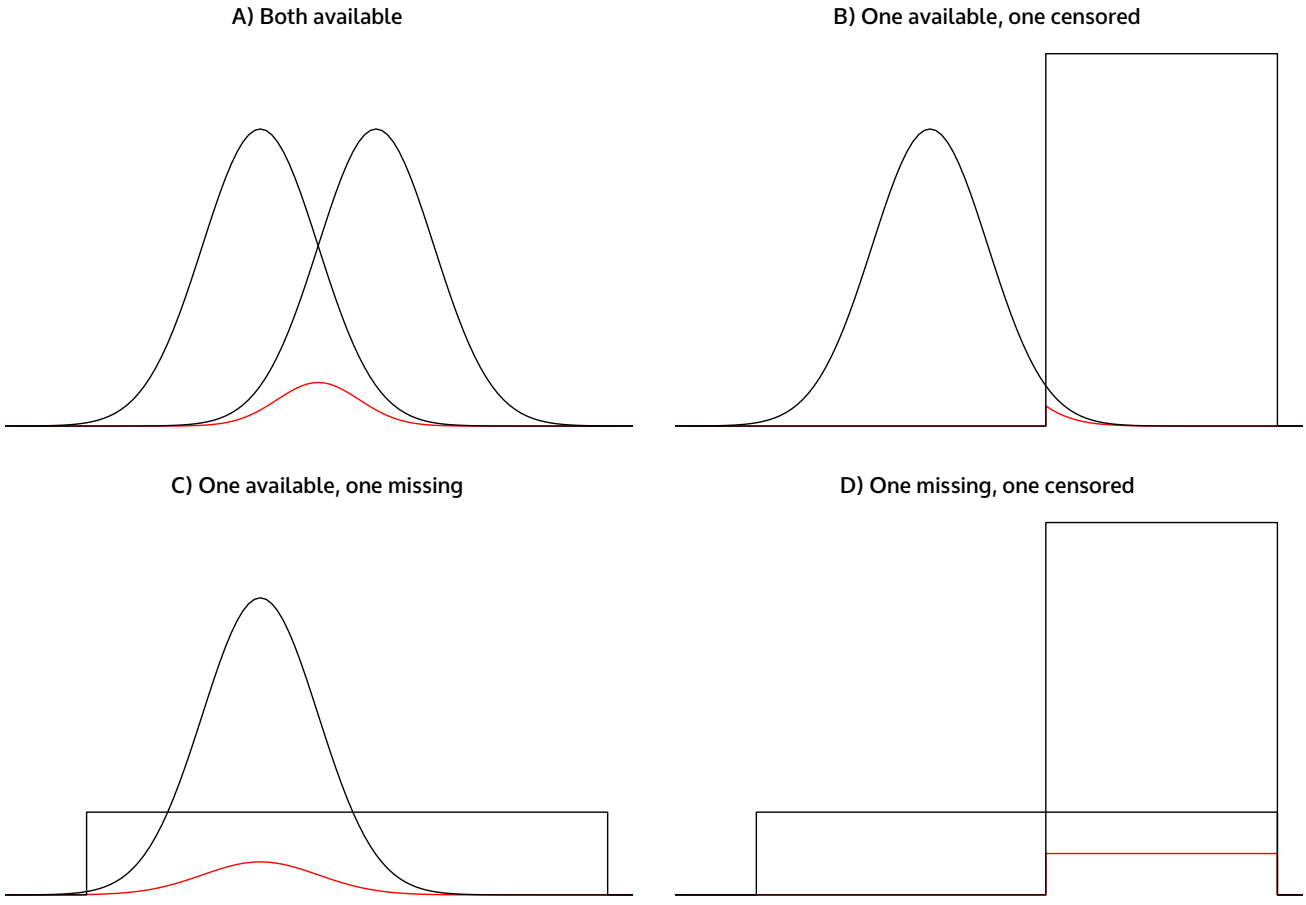

Figure S1. Transition probability models used for censored and missing values. A) The measurement of the current gene product is available for both cells. B) One measurement is missing. C) One measurement is set to the level of detection (i. e. censored) D) One measurement is missing, the other one censored.

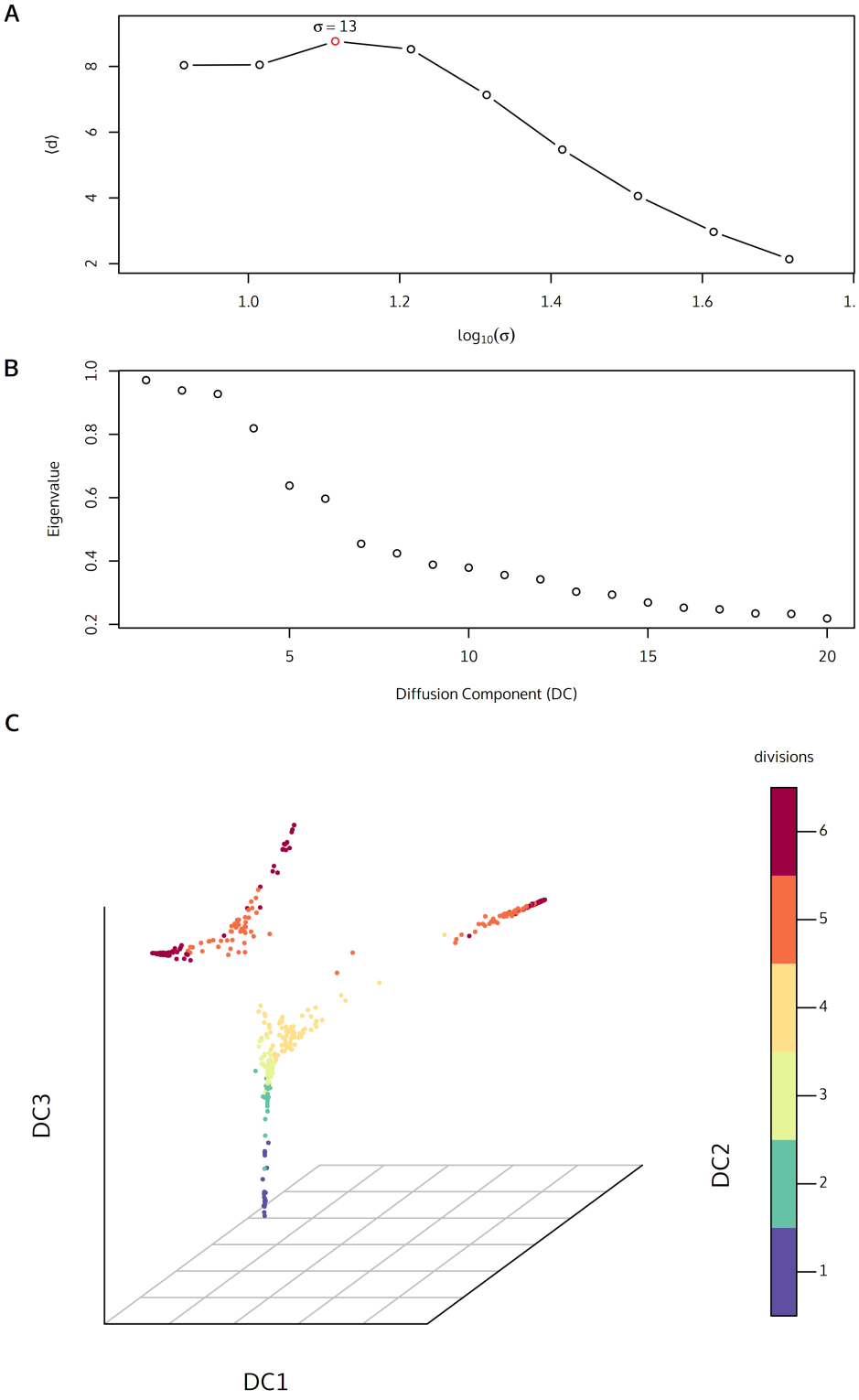

Figure S2. Diffusion map on single-cell qRT-PCR data of mouse embryonic cells from oocyte to 64-cell stage (**guo_resolution_2010**) A) Determination of the optimal Gaussian kernel width σ. B)The Eigenvalues of the first twenty diffusion components. The gap after the first three eigenvalues indicates an adequate dimension reduction on the first diffusion components. C) The *destiny* output for 48 genes from 429 cells arranges the cells in branches. Two different branching events can be identified leading to three clearly separated cell types in the 64-cell stage (6 divisions).

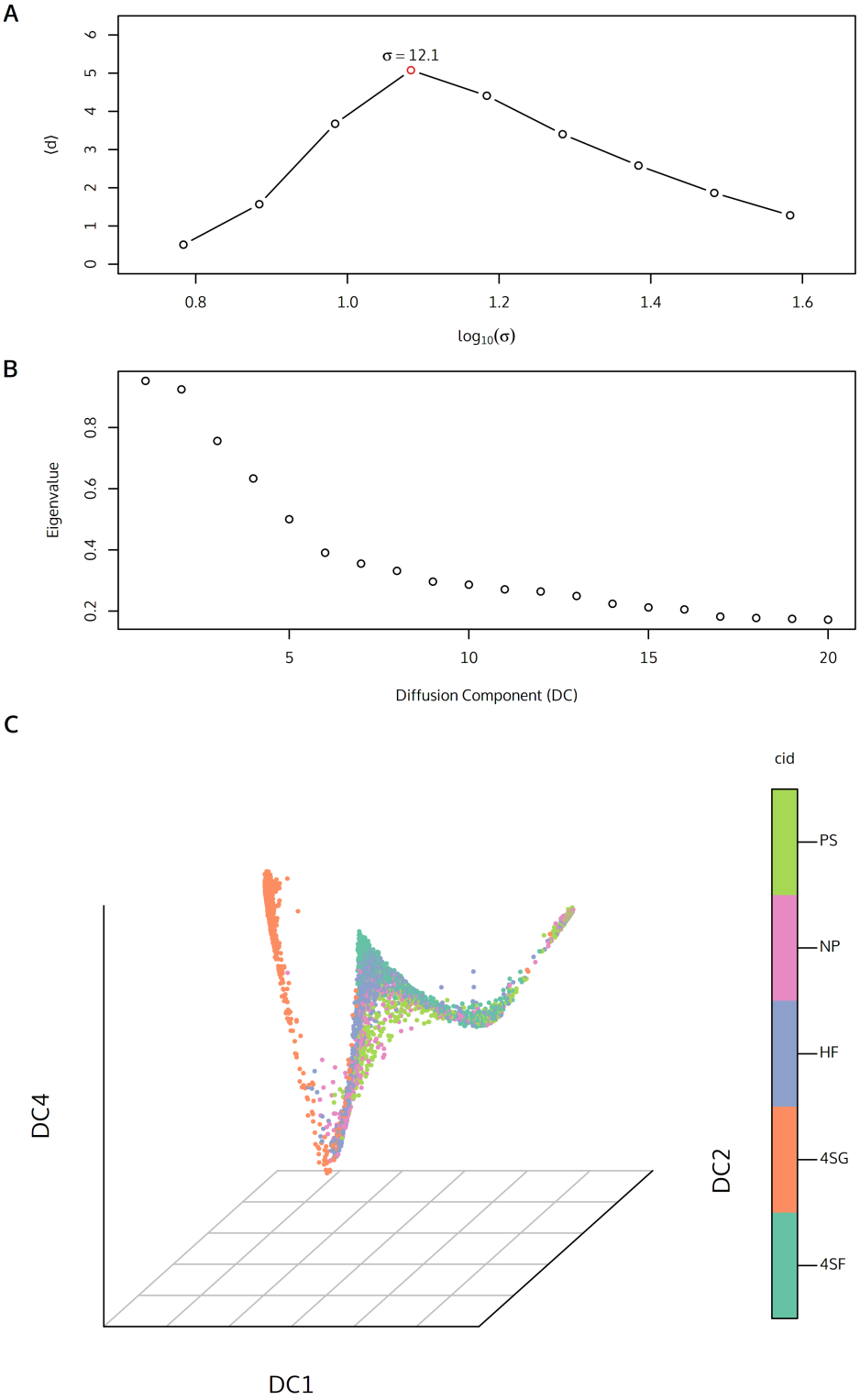

Figure S3. Diffusion map on single-cell qRT-PCR data of mouse embryonic cells with blood-forming potential (**moignard_decoding_2015**) A) Determination of the optimal Gaussian kernel width σ. B) The Eigenvalues of the first twenty diffusion components. The decrease of the Eigenvalues after the third component indicates C) The *destiny* output for 46 genes from 3934 cells. The diffusion trajectories *destiny* created correspond to the developmental stages of the progenitor cells (see **moignard_decoding_2015**).

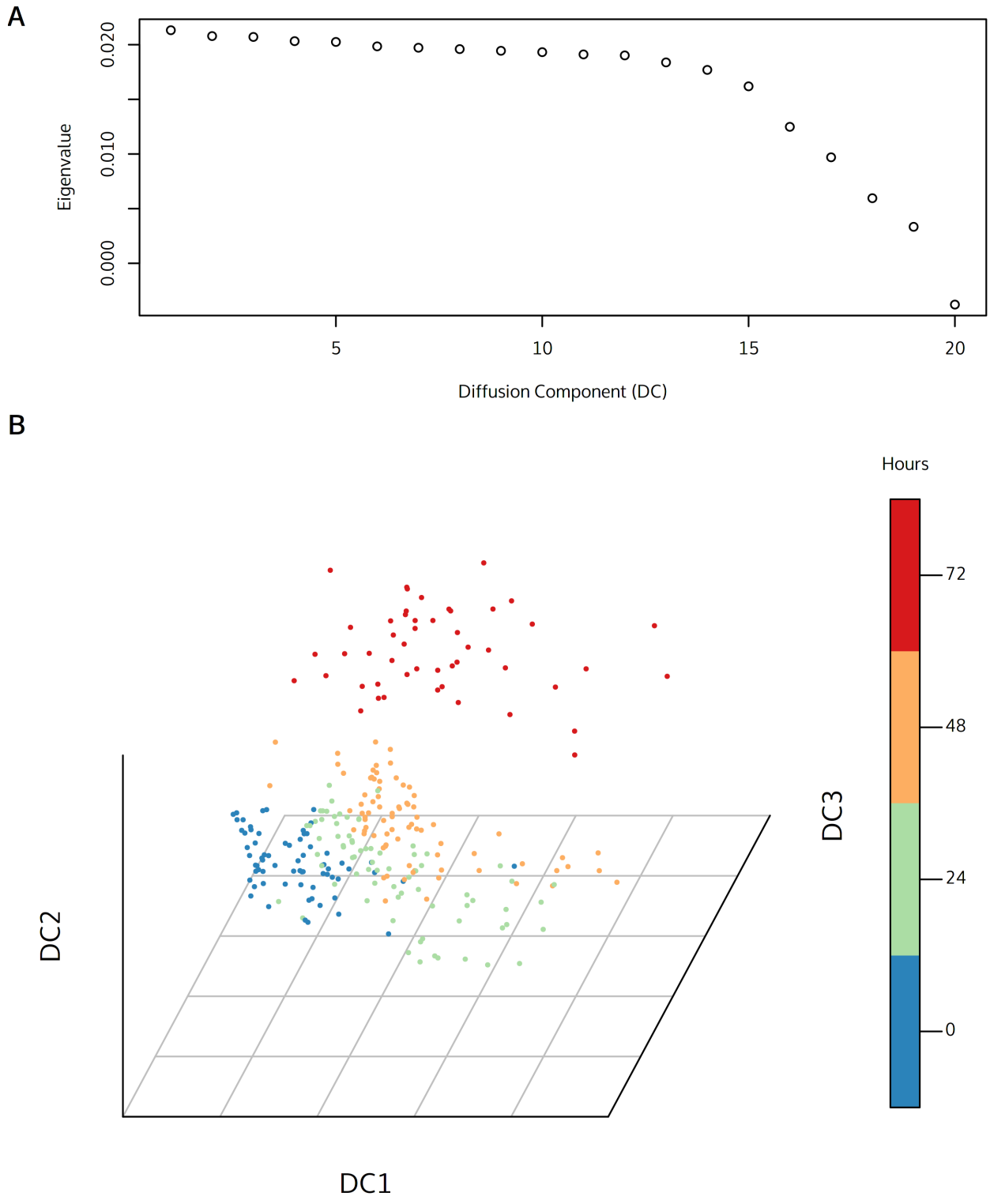

Figure S4. Diffusion map of the human skeletal muscle myoblasts RNA-Seq dataset from **trapnell_dynamics_2014** A) The Eigenvalues of the first twenty diffusion components B) The diffusion map computed from the first 20 singular values (SVs) of the RNA-Seq counts for 271 cells and 47192 genes. The sigma parameter was manually specified to 0.39.

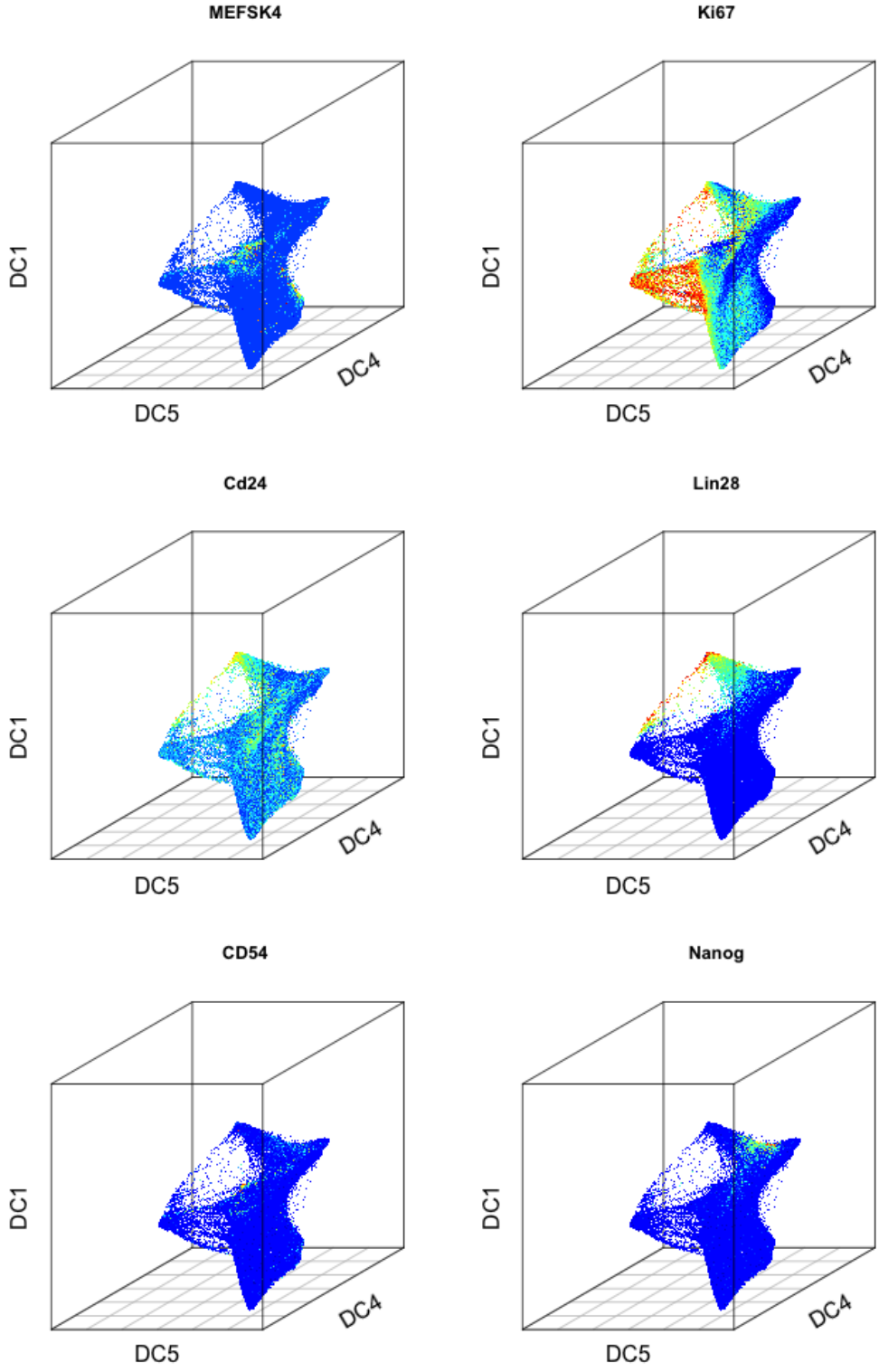

Figure S5. Expression of 6 intracellular and surface markers mapped on the diffusion map of the reprogramming dataset from **zunder_continuous_2015** (see Figure 1). Coloring normalized on the log-scale (low - dark blue, medium - green, high - red).