nature computational science

RESOURCE

https://doi.org/10.1038/s43588-021-00025-y

Check for updates

Enabling single-cell trajectory network enrichment

Alexander G. B. Grønning[1,2]✉, Mhaned Oubounyt[3,4], Kristiyan Kanev[5], Jesper Lund[6], Tim Kacprowski[7], Dietmar Zehn[5], Richard Röttger[1] and Jan Baumbach[1,3,4]✉

Single-cell sequencing (scRNA-seq) technologies allow the investigation of cellular differentiation processes with unprecedented resolution. Although powerful software packages for scRNA-seq data analysis exist, systems biology-based tools for trajectory analysis are rare and typically difficult to handle. This hampers biological exploration and prevents researchers from gaining deeper insights into the molecular control of developmental processes. Here, to address this, we have developed Scellnetor; a network-constraint time-series clustering algorithm. It allows extraction of temporal differential gene expression network patterns (modules) that explain the difference in regulation of two developmental trajectories. Using well-characterized experimental model systems, we demonstrate the capacity of Scellnetor as a hypothesis generator to identify putative mechanisms driving haematopoiesis or mechanistically interpretable subnetworks driving dysfunctional CD8 T-cell development in chronic infections. Altogether, Scellnetor allows for single-cell trajectory network enrichment, which effectively lifts scRNA-seq data analysis to a systems biology level.

Single-cell RNA sequencing (scRNA-seq) allows researchers to perform cellular developmental studies with a hitherto unseen fine granularity. Single-cell transcriptomes have paved the way for novel discoveries in various biomedical fields by improving the understanding of how transcriptional profiles relate to cell phenotypes. A range of algorithms have been invented for clustering of scRNA-seq data and for inferring differentiation trajectories[1,2]. Clustering assumes that single cells can be divided into distinct groups, whereas trajectory inference aims to arrange cells such that continuous phenotypes can be traced on a low-dimensional cell map[3]. Important examples of the latter include diffusion maps[4] and pseudotemporal ordering of single cells[2,5]. Both algorithms seek to position single cells such that their coordinates reflect their developmental statuses in relation to the other cells. Additionally, several software packages have been developed for the entire analysis pipeline, from pre-processing to clustering and identification of differentially expressed genes. Scanpy[6], Seurat[7] and SINCERA[8] are examples of such software packages. Although scRNA-seq data are still challenged by noise[9], combinations of different tools and algorithms have helped to unravel hidden intercellular mechanisms and shed light on unknown cellular paths of differentiation and disease progression[10,11].

Typical computational analyses of single-cell gene expression data involve a pre-processing step, where, for example, cells with high levels of mitochondrial DNA and few expressed genes are removed. This is often followed by steps like normalization and dimensionality reduction of the data[12,13], which are typically succeeded by clustering of the single cells' transcriptional profiles and/or inference of developmental trajectories[12,14]. Normally, the clusters or trajectory segments are validated and identified using the expression of marker genes[9–12,15]. A notable tool for this type of analysis is Switchde[16], which can model differential expression over pseudotime and thereby identify relevant genes based on the pseudotemporal ordering of single cells. A way to examine development trajectories more mechanistically is by inferring gene regulatory networks from the scRNA-seq data in question[9,12,17–19]. Despite being useful in specific scenarios, such (pseudo) gene regulatory networks are limited in their power to describe the interactome beyond transcription factors. The mechanistic patterns they describe do not grant a view of the full picture of the complexity of cellular developments. In bulk RNA-seq data analysis, network enrichment technology (for example, KeyPathwayMiner[20], GiGa[21] or ActiveModules[22]) is typically applied to find such mechanistic patterns. However, these methods have not been developed to consider the noise of scRNA-seq data and to work with the computationally inferred pseudotime relationships of cells, making it difficult to obtain meaningful results when applying them to single-cell expression data. To meet some of these challenges, the tool scPPIN[23] was recently proposed. This algorithm allows for the comparison of single-cell groups by combining differentially expressed genes with a protein–protein interaction (PPI) network. By constructing a node-weighted network, the tool finds maximum-weight connected subgraphs containing genes that are differently expressed in the compared groups.

Even though the approaches to scRNA-seq analysis outlined above can help to deduce new insights from scRNA-seq, they cannot identify mechanistic patterns that explain pseudotemporal cellular developmental programs at an interactome level. Moreover, no tools exist that can use the pseudotemporal ordering of single cells to identify molecular subnetworks enriched with genes that are differentially expressed in two distinct differentiation trajectories. From a systems medicine point of view, as no tool for direct comparison of healthy differentiation trajectories and disease-associated development trajectories exists, it remains impossible to locate genes that in synergy, as a mechanism, are responsible for disease progression.

To fill this gap, we have developed Scellnetor, which stands for 'Single-cell Network Profiler for Extraction of Systems Biology

[1]Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark. [2]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [3]Chair of Experimental Bioinformatics, TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany. [4]Chair of Computational Systems Biology, University of Hamburg, Hamburg, Germany. [5]Division of Animal Physiology and Immunology, TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany. [6]Department of Biostatistics and Epidemiology, University of Southern Denmark, Odense, Denmark. [7]Division Data Science in Biomedicine, Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Brunswick, Germany. ✉e-mail: alexander.groenning@sund.ku.dk; jan.baumbach@uni-hamburg.de

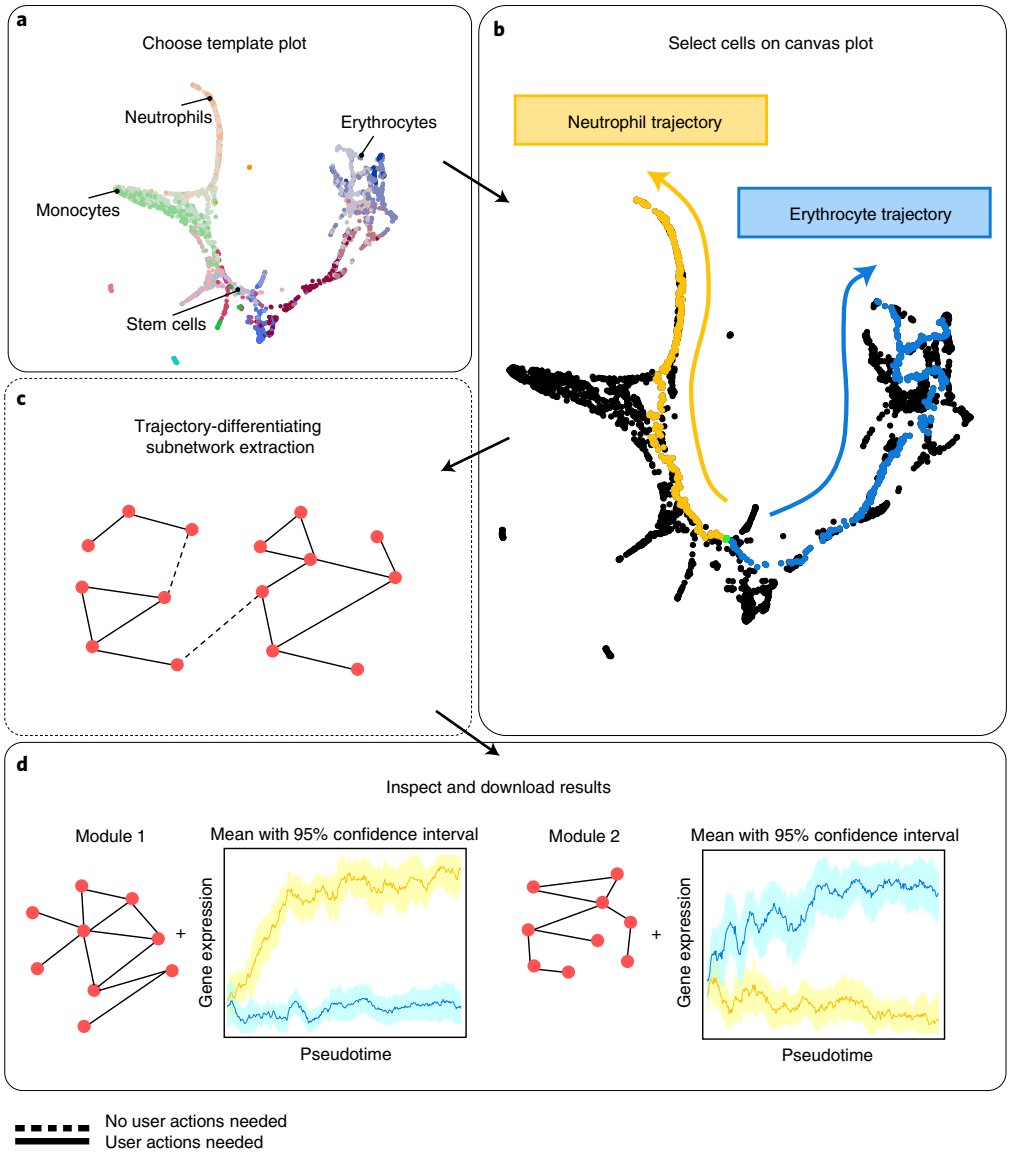


Fig. 1 | Workflow of Scellnetor. a, The user chooses the desired template plot. **b**, The user selects cells on the canvas plot that represent differentiation trajectories. **c**, Scellnetor performs a constrained agglomerative hierarchical clustering based on expression data from the selected cells. **d**, The user can inspect and download the results. Scellnetor outputs connected subnetworks of genes (modules) and mean gene expression together with the 95% confidence intervals.

Patterns from scRNA-seq Trajectories'. Scellentor is the first pseudotemporal scRNA-seq network enrichment technique that can unravel connected subnetworks of genes crucial for explaining the progression of single-cell development trajectories. The method allows users to compare single-cell trajectories selected on low-dimensional cell maps. The selected cell sets are pseudotime-sorted and clustered using a hierarchical clustering algorithm that is constrained by the PPI network from the BioGRID[24] (or a user-chosen network). Scellnetor identifies gene modules (connected subnetworks of genes) that are either differently or similarly expressed in two selected sets of cells (Fig. 1). The tool is therefore able to extract mechanisms that are fundamental cellular driver programs for differentiating between distinct development courses.

To clarify terminology, throughout this paper we will refer to subnetworks of connected genes as 'modules', which can be represented by a set of genes and a set of edges. By contrast, we refer to a set of cells as a 'cluster' or a 'group' if the cells have been picked without any ordering (for example, representing a cell type) or we refer to them as a 'trajectory' if a set of cells are ordered, for example by pseudotime. Scellnetor now identifies 'modules' (subnetworks of genes in a given interaction network) by comparing two 'clusters', two 'groups' or two 'trajectories' of user-selected cells.

Results

Overview of the Scellnetor method. Scellnetor allows for comparisons of user-chosen single-cell sets and to unravel network modules driving cell differentiation or disease progression on a system-biological level over pseudotime. Scellnetor requires Scanpy-generated ANNDATA[6] objects in H5AD file format as raw data input and Scanpy-generated plots as template plots. The template plots are cell maps contained within the uploaded ANNDATA object that the user wishes to apply for data representation in the downstream analysis pathway (Fig. 1a and Supplementary Fig. 1). The coordinates of the single cells on the template plots are converted into points on a canvas plot, on which the user can interactively select cell clusters or trajectories to be analysed (Fig. 1b and Supplementary Fig. 1). In Fig. 1b, a user has selected two sets

of single cells by creating two paths through the canvas plot. See Methods for a detailed explanation of the algorithms underlying the single-cell selection on the canvas plot. The selected cells are extracted and converted into expression matrices in which the cells are sorted in ascending order based on a user-chosen sorting key. We recommend using pseudotime as sorting key, as it provides information about the intercellular differentiation status over time[5], which further enhances Scellnetor's ability to identify interaction network modules important for the analysed development trajectories. Scellnetor computes a hyper-similarity matrix when comparing two single-cell sets (like in Fig. 1). The hyper-similarity matrix contains information on how the genes are expressed compared to one another in the individual single-cell sets and compared to the genes in the single-cell set against which they are compared (Methods and Fig. 2). Scellnetor clusters the data of the hyper-similarity matrix using a constrained agglomerative hierarchical clustering algorithm for extracting gene modules. The clustering is constrained by the interactions of biological networks, by default extracted from the BioGRID[24] database (Methods).

Scellnetor outputs (1) network modules enriched with genes that are differently expressed in the two compared cell sets, (2) plots that show mean expressions and 95% confidence intervals of the genes in the modules and (3) TSV files with statistically significant Gene Ontology (GO) terms of the modules' genes (Methods).

Differences between neutrophil and erythrocyte development trajectories. To validate the Scellnetor methodology, we used scRNA-seq data from ref. [15]. In their article, Paul et al. analysed gene expression patterns of mouse haematopoietic cells while they differentiated to progenies from a pool of progenitor cells: common myeloid progenitors (CMPs), granulocyte-macrophage progenitors (GMPs) and megakaryocyte-erythrocyte progenitors (MEPs). Using an expectation maximization-based clustering approach, Paul et al. divided the single cells into 19 different groups. Finally, they constructed a detailed map of the dynamic transcriptional states within the myeloid progenitor populations. Using the single cells from the 19 groups and their scRNA-seq gene expression dataset (GSE72857), we constructed an ANNDATA object, created cell maps and computed pseudotime. Our ANNDATA object contained 2,730 single cells that expressed 3,451 genes (Methods). Our pseudotime calculation was based on the progenitor groups defined by Paul et al. (groups 7–11 in Supplementary Fig. 2a) and coordinates from a cell map (used as the template plot) based on principles of force-directed graph drawing[25], which showed clear branching of the differentiated cells. Also, the plot showed clear co-localization of the Paul et al. groups that were highly similar. The 'start cell' for the pseudotime computation[5] was the cell closest to the average position of all relevant progenitor groups (groups 7–10 in Supplementary Fig. 2a and Fig. 2b; see Supplementary Information for the proposed rationales behind finding a 'start cell'). Group 11 was omitted, as it was dislocated from the remainder of the cells on the plot (Supplementary Fig. 2a). The connected area where Paul et al. groups 7–10 are co-localized will be referred to as the 'stem cell area' (Supplementary Fig. 2a and Fig. 3b).

We ran our Scellnetor algorithms to extract comparative systems-biology profiles between the trajectory from stem cells towards differentiated neutrophils versus the trajectory from stem cells towards differentiated erythrocytes (Fig. 3a). The drawn paths go through several of the Paul et al. defined groups. Outside the stem cell area, the neutrophil path goes through groups 15–17, where 16–17 are the neutrophil-specific groups. The erythrocyte path goes through cells in the stem cell area and groups 1–6, which all are erythrocyte-specific groups (Fig. 3a and Supplementary Fig. 2a). Using Scellnetor, we identified seven gene modules with a minimum size of five genes using pseudotime as the sorting key. The distance metric was Euclidean, the linkage type was complete and the size of the moving average was 20 (information about the hyperparameter settings is provided in the Methods).

In Fig. 3, we demonstrate three of the modules. Figure 3b,d,f presents plots of the means and 95% confidence intervals of the smoothed average expression in each of the three modules. One can see how the gene expression in the modules differs over pseudotime between the neutrophil and erythrocyte trajectories. Using the Wilcoxon signed-rank test, we show that the average expression over time of the genes of the modules from the two distinct paths are statistically significantly different (q value of 2.52×10^{-59} for all modules).

Consistency in gene expression. In Fig. 3a, the neutrophil path passes through stem cells and cells from Paul et al. groups 15–17, whereas the erythrocyte path goes through stem cells and cells from Paul et al. groups 1–6 . The paths contain subsets of these groups, so it was of interest to check if the differences in expression between the paths were consistent with the differences in expression between the corresponding groups (for example, groups 15–17 versus groups 1–6, according to ref. [15]). We extracted all genes from the Scellnetor modules (Fig. 3 and Supplementary Fig. 3) and created two cell sets, one containing all cells from Paul et al. groups 15–17 and one containing all cells from Paul et al. groups 1–6. Based on the two sets, we used our initial count matrix (Methods) to compute the average expression of the genes in the Scellnetor modules. We plotted the averaged gene expression values and compared the resulting distributions using the Wilcoxon signed-rank test (Supplementary Fig. 4a–g). The same was done for the subsets of the above Paul et al. groups that were included in the paths shown in Fig. 3a (Supplementary Fig. 4h–n). For these calculations, we used the pre-processed expression matrix from our ANNDATA object (Methods). We found that the cells in the Scellnetor paths (Fig. 3a) expressed genes in a manner that was consistent with the gene expression of the cells in the relevant Paul et al. groups.

Unravelling of marker genes. The clustered genes as gene modules are shown in Fig. 3c,e,g. The q values next to each module are based on Mann–Whitney U tests and indicate that these modules are statistically significantly different from the entire distribution of hyper-similarities (the q values for modules 1, 2 and 3 are 4.29×10^{-9}, 1.25×10^{-5} and 2.58×10^{-50}, respectively). In other words, a low q value (and P value) indicates that it is unlikely that the connections of the module are similar to connections randomly sampled from the entire distribution of hyper-similarities, and thus are similar to a randomly generated module. The genes *P4HB*, *CALR* and *HSPA5* in module 1 (Fig. 3c) produce surface markers that are expressed at high levels in neutrophils. The genes *GATA1*, *ZFPM1* and *GTF2F1* in module 3 (Fig. 3g) code for transcription factors that are upregulated in erythrocyte differentiating cells. The genes *NCL*, *HBA2*, *C1QBP* and *ATP5IF1* are all known marker genes associated with the erythrocyte lineage (Fig. 3g). Additional transcription factors upregulated in erythropoiesis are *GFI1B*, *LMO2* and *CBFA2T3*[15], which were found in Scellentor module 7 (Supplementary Fig. 3h). The genes in the Scellnetor module 3 (Fig. 3g) are more highly expressed on average in the cells located in the erythrocyte trajectory. This is corroborated by higher expression of *HBA2*, which codes for a subunit of haemoglobin[26], and *ATP5IF1*, which codes for a mitochondrial ATPase inhibitor that is involved in the synthesis of haemoglobin[27].

Gene set enrichment analysis results. Scellnetor automatically conducts gene set enrichment analysis when gene modules have been identified. To demonstrate this functionality, the tables in Fig. 3b,d,f display four biologically meaningful statistically significant biological-process GO terms associated with each module. The GO terms associated with the modules are all provided in Supplementary

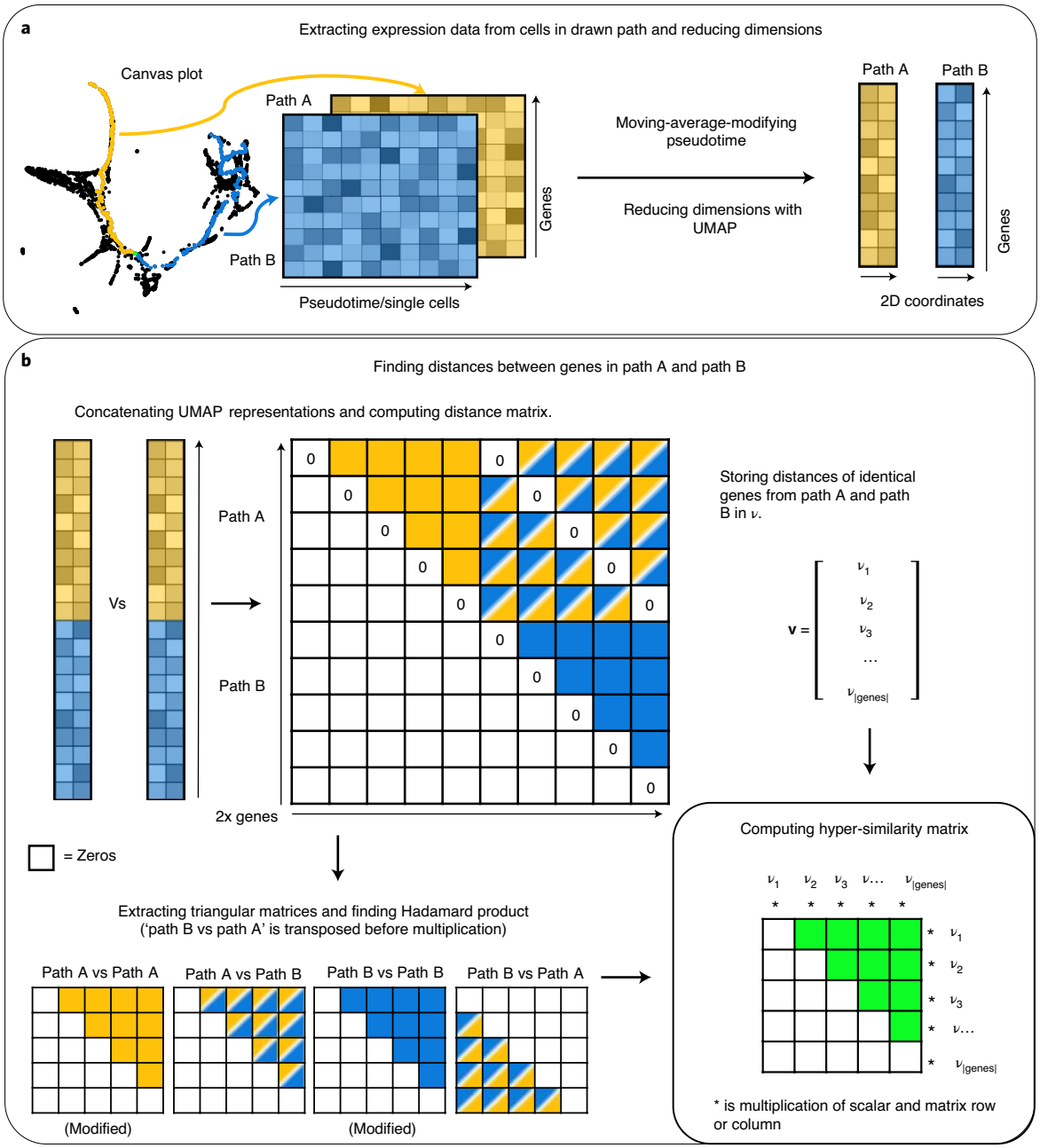
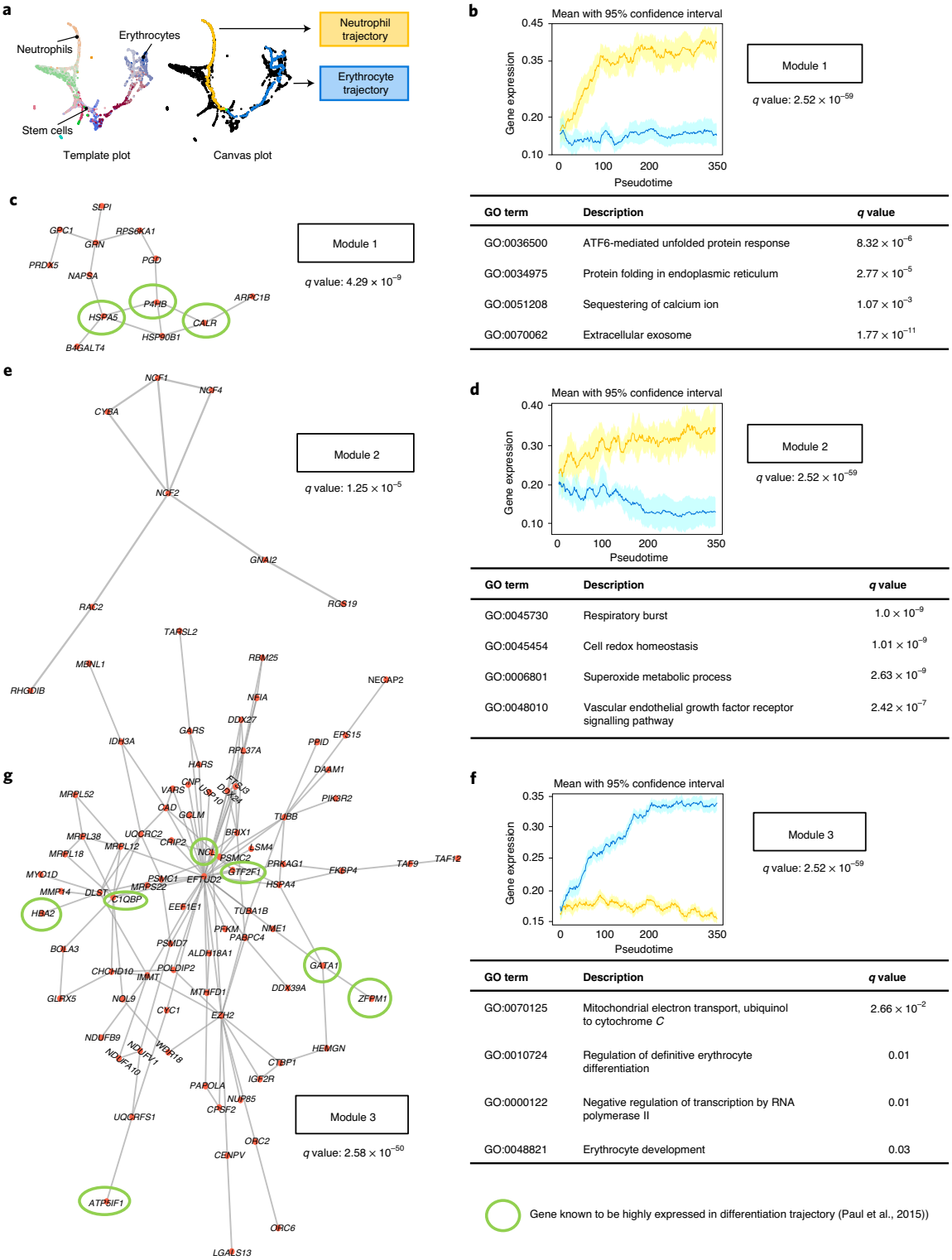


Fig. 2 | Computation of hyper-similarity matrix for gene module discovery. a, Two paths are drawn on the canvas plot: path A and path B. Expression matrices (one for each path) are created based on data from selected cells. All data that are derived from path A or path B will be labelled as path A or path B in the figure. In the matrices, rows are genes and columns are cells (pseudo-timepoints). Rows of the matrices are smoothed using a moving average function. The smoothed expression matrices are dimensionality-reduced with UMAP. **b**, The resulting two-dimensional (2D) coordinate sets from path A and path B, respectively, are concatenated along the *x* axes. The concatenated coordinate sets are used to produce a distance matrix. The diagonal, the lower triangular matrix and the diagonal of the upper right quadrant of the distance matrix are zeroed out. The distances between genes with identical IDs, but from different sets, are computed and stored in the vector **v** (top, right). Triangular matrices of the distance matrix quadrants I, II and IV (as in the cartesian coordinate system) are extracted. Values of the matrices 'path A versus path A' and 'path B versus path B' are converted into similarities. The matrix 'path B versus path A set' is transposed and the Hadamard product of all triangular matrices is computed (bottom, left). This produces the hyper-similarity matrix. The rows and columns of the hyper-similarity matrix are weighted by the values of **v** (bottom, right).

File 1. Genes involved in the 'ATF6-mediated unfolded protein response', 'protein folding in endoplasmic reticulum' and 'sequestering of calcium ion' have been found highly expressed in module 1 (Fig. 3b). The first two GO terms describe events that are related to protein folding and endoplasmic reticulum (ER) stress. It has been shown that ER stress is decreased during neutrophil differentiation[28]. Scellnetor showed that *CALR* is highly expressed in the neutrophil trajectory compared to the erythrocyte trajectory. *CALR* codes for a multi-functional protein called calreticulin, which counteracts ER stress and is involved in the correct maintenance of calcium ions in the ER[29]. The gene *P4HB* (Fig. 3e) has also been found to downregulate ER stress[30]. Eleven of 13 genes in module 1 (Fig. 3b; except for *B4GALT4* and *RPS6KA1*) are associated with the GO term 'extracellular exosome'. Releasing exosomes is an important



GO term	Description	q value
GO:0036500	ATF6-mediated unfolded protein response	8.32×10^{-6}
GO:0034975	Protein folding in endoplasmic reticulum	2.77×10^{-5}
GO:0051208	Sequestering of calcium ion	1.07×10^{-3}
GO:0070062	Extracellular exosome	1.77×10^{-11}

GO term	Description	q value
GO:0045730	Respiratory burst	1.0×10^{-9}
GO:0045454	Cell redox homeostasis	1.01×10^{-9}
GO:0006801	Superoxide metabolic process	2.63×10^{-9}
GO:0048010	Vascular endothelial growth factor receptor signalling pathway	2.42×10^{-7}

GO term	Description	q value
GO:0070125	Mitochondrial electron transport, ubiquinol to cytochrome *C*	2.66×10^{-2}
GO:0010724	Regulation of definitive erythrocyte differentiation	0.01
GO:0000122	Negative regulation of transcription by RNA polymerase II	0.01
GO:0048821	Erythrocyte development	0.03

Fig. 3 | Comparison of neutrophil and erythrocyte differentiation trajectories. a, The chosen template plot (left) and selected cells on the canvas plot (right). The selected cells form paths representing trajectories from stem cells towards differentiated neutrophils and from stem cells towards differentiated erythrocytes, respectively. **b**,**d**,**f**, Plots of mean expression values of the modules over pseudotimelines and the respective 95% confidence intervals of modules 1 (**b**), 2 (**d**) and 3 (**f**). The mean pseudotimelines from both trajectories of each module were compared using Wilcoxon signed-rank tests. The resulting *q* values are provided below each plot (*q* values were determined using the Benjamini–Hochberg procedure). Below each plot are tables with the four statistically most significant GO terms linked to the modules. **c**,**e**,**g**, Modules identified by Scellnetor. The encircled genes in **c**, **e** and **g** are known to be highly expressed in the differentiation trajectory that on average has the highest expression in **b**, **d** and **f**, respectively. The associated *q* values are based on Mann–Whitney *U* tests and indicate that these modules are statistically significantly different from randomly generated sets of genes of the same size (*q* values were determined using the Benjamini–Hochberg procedure).

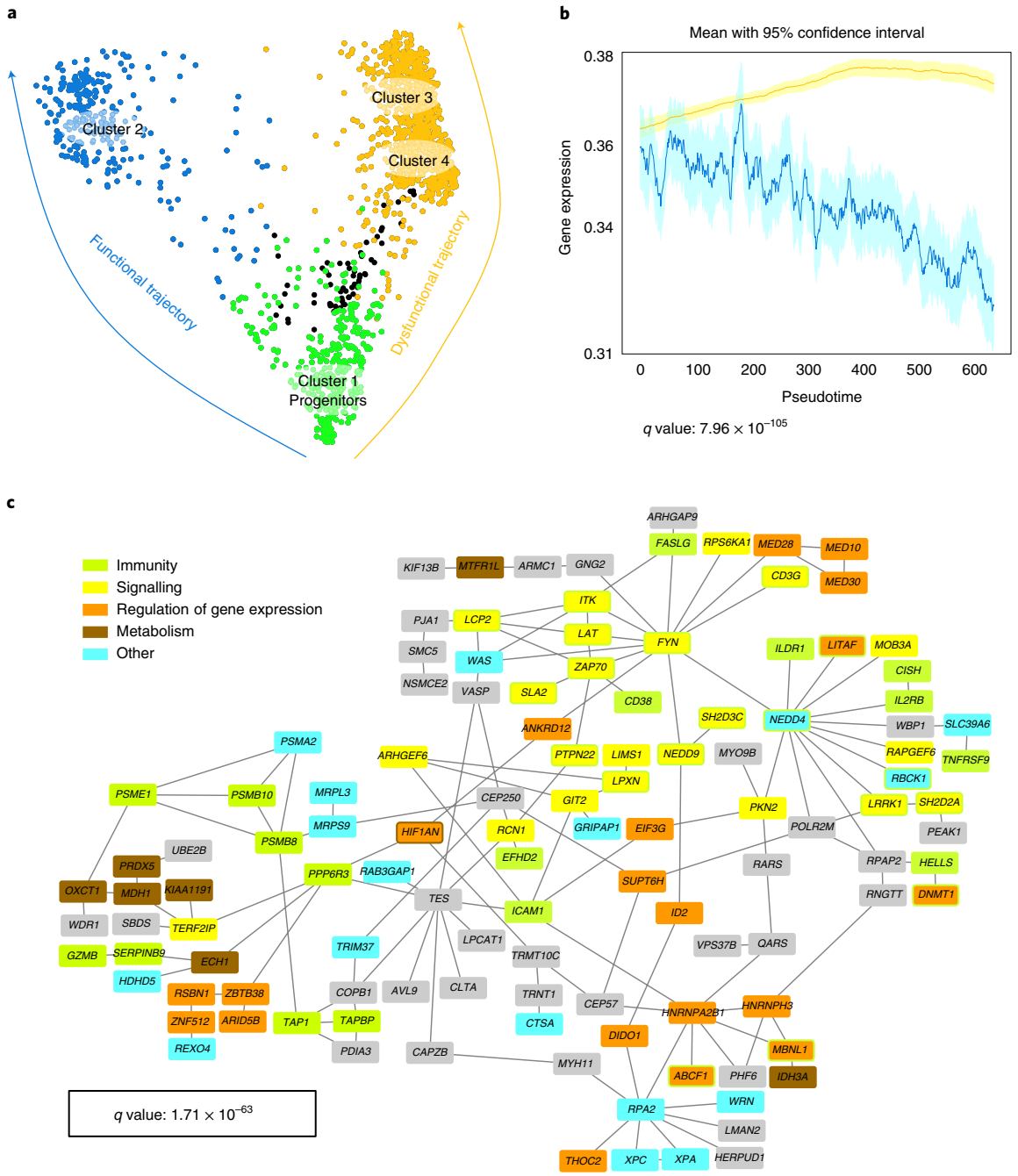


Fig. 4 | Differentiation of progenitor cells to dysfunctional CD8 T cells. a, Diffusion map of the single cells that differentiate into functional and dysfunctional CD8 T cells. On the left, in navy blue, is the development trajectory of cells from progenitor cells (cluster 1 in ref. [10]) to functional CD8 T cells (cluster 2 in ref. [10]). On the right, in amber, is the development trajectory of cells from progenitor cells to dysfunctional CD8 T cells (clusters 3 and 4 in ref. [10]). **b**, Mean gene expression and 95% confidence intervals of the selected cluster. The genes in the cluster shown in **c** are on average expressed at a higher level in the cells developing via the functional trajectory. The *q* value below the plot is based on a Wilcoxon signed-rank test comparing the average gene expressions. **c**, A Scellnetor-identified module. The genes are colour-coded as follows: green genes are involved in immune responses, yellow genes are involved in signalling, orange genes are involved in regulation of gene expression, brown genes are involved in metabolism, blue genes are involved in other processes and grey genes do not belong to any of the mentioned groups. The colour coding is based on our subjective assessments of the genes' general functions. The associated *q* value is based on the Mann–Whitney *U* test. This indicates that the cluster is statistically significantly different from randomly generated clusters. All *q* values in this figure were found using the Benjamini–Hochberg procedure.

part of neutrophil signalling[31]. The GO terms 'respiratory burst', 'cell redox homeostasis' and 'superoxide metabolic process' from module 2 (Fig. 3d) all relate to well-known neutrophilic cellular processes. A distinctive feature of the inflammatory actions of neutrophils is the respiratory or oxidative burst, where large amounts of oxygen are consumed to produce superoxide[32]. Recent studies revealed that neutrophils might play a key role in angiogenesis, as they can store and synthesize molecules with known angiogenic activity (fourth GO term, Fig. 3d)[33–35]. Module 2 does not contain any genes that were highlighted by ref. [15] as noteworthy for the neutrophilic differentiation course. However, it has been described that neutrophil cytosolic factors 1, 2 and 4 (*NCF1*, *NCF2* and *NCF4*) interact with

CYBA and *RAC2* at the membrane as subunits of the NOX2 complex. Interestingly, the inclusion of *RAC2* in this molecular assembly is specific for neutrophils[36–38].

The GO terms 'regulation of definitive erythrocyte differentiation' and 'erythrocyte development' from module 3 (Fig. 3f) indicate that the cells from the erythrocyte trajectory in fact do express genes associated with erythropoiesis. The GO term 'mitochondrial electron transport, ubiquinol to cytochrome *c*' (Fig. 3f) is interesting, because the proteins resulting from the genes linked to this term (*UQCRC2*, *UQCRFS1* and *CYC1*; Fig. 3g) have been found in high abundance in erythroid progenitor cells compared to haematopoietic stem cells[39]. Apparently, these genes are more highly expressed in erythrocyte differentiating cells than in neutrophil differentiating cells. The GO term 'negative regulation of RNA polymerase II' fits well with genes involved in erythropoiesis, because mature erythrocytes lose their nuclei, which is where RNA polymerase II catalyses transcription of genes to pre-mRNA[40].

Differences between functional and dysfunctional exhausted CD8 T cells in chronic infection. To demonstrate that Scellnetor can be used to identify mechanisms underlying disease progression, we re-analysed a dataset of pathways of CD8 T-cell differentiation in a well-defined standard model of chronic infections[10]. In their original work, the authors of ref. [10] investigated the role of CD4 help for the maintenance and function of the stem-like progenitors and their terminally differentiated CD8 T-cell progeny using scRNA-seq. Using previously established signature genes, they identified five clusters. Cluster 1 represented the critical stem-like progenitors and cluster 2 the functional and clusters 3–5 the dysfunctional differentiated effector cell subpopulations. Absence of CD4 help did not affect the maintenance of the progenitors, but caused a massive decline in the effector compartment, the most affected of which was the functional cluster 2. In our study, we used the cells from these five clusters (GSE137007) to generate an ANNDATA object, calculate pseudotime and compute a diffusion map (Methods). We used the cluster annotations of the above presented clustering from ref. [10] to colour-code the cells (Fig. 4a).

We defined two trajectories of differentiation starting from the progenitor subpopulation (cluster 1), one towards the functional effector cells (cluster 2) and one towards the dysfunctional effector cells (clusters 3 and 4). We excluded cluster 5 (Fig. 4a, cells in black) from the analysis as it is condition-specific and was only observed in the absence of CD4 help. Using Scellnetor, we unravelled a gene module with increased expression in cells progressing in the dysfunctional trajectory and decreased expression in those progressing in the functional trajectory (Fig. 4b,c). The complexity of the gene module is highlighted by the diversity of the genes and their functions, including processes like regulation of gene expression (epigenetic, transcriptional and translational), signalling, immune defence, cell migration, cytoskeletal reorganization and genes involved in the T-cell receptor (TCR) signalling pathway (*CD3G*, *FYN*, *ZAP70*, *LAT*, *ITK*, *PTPN22*, *LCP2*, *SLA2*, *NEDD4*, *NEDD9*, *LRRK1*, *SH2D2A* and *SH2D3C*), whose excessive stimulation is known to be one of the main drivers of T-cell exhaustion[41]. The two most statistically significant GO terms associated with this gene module are 'T-cell receptor signalling pathway' and 'T-cell activation' (Supplementary File 2). TCR signalling is connected to two therapeutically interesting receptors involved in the regulation of the T-cell response, ILDR1 and TNFRSF9, which could potentially be used for its modulation. ILDR1 function as a regulator of T-cell response in chronic infection and cancer is intriguing, especially taking into account that ILDR2 was recently described as a negative regulator of T-cell function[42]. *TNFRSF9* (coding 4–1BB) is a positive regulator of T-cell effector function and survival, and its stimulation has already been reported to ameliorate T-cell exhaustion[43]. The mechanistic gene module also includes diverse regulators of nuclear factor-κB signalling (*PPP6R3*, *TERF2IP* and *EFHD2*), which is critical for T-cell survival and cytokine production[44]. The negative regulator EFHD2 is of particular interest, as it is necessary for PD-1-mediated inhibition of proliferation and cytokine secretion in dysfunctional CD8 T cells[45]. As previously reported, the transition from functional to dysfunctional CD8 T cells is associated with metabolic reprogramming marked by a switch from glycolysis to oxidative phosphorylation (OXPHOS) as a main pathway for the generation of adenosine triphosphate (ATP). In line with this, our dysfunctional gene module includes multiple genes involved in fatty-acid beta oxidation and OXPHOS (*MDH1*, *KIAA1191*, *IDH3A*, *ECH1*, *OXCT1* and *PRDX5*) and putative regulators of this metabolic adaptation (*HIF1AN*, *ARID5B* and *MTRF1L*). Interestingly, HIF1AN is an inhibitor of HIF1α (HIF1, hypoxia inducible factor 1), which is known to trigger the expression of genes promoting the use of glycolysis over mitochondrial oxidative phosphorylation as the main energy generating pathway[46–48]. Moreover, HIF activity has been shown to enhance the effector CD8 T-cell response and influence the expression of pivotal transcription, effector and co-stimulatory molecules in chronic infection[49]. Altogether, this demonstrates that the gene modules extracted by Scellnetor are functionally related mechanisms that precisely reflect the opposing nature of progenitor differentiation towards functional or dysfunctional CD8 T cells. Thus, Scellnetor is a promising systems medicine hypothesis generator for identifying the molecular subnetworks mechanistically driving dynamic differentiation of complex cell populations.

Discussion

To investigate Scellnetor's robustness and sensitivity, we conducted different analyses of Scellnetor in different settings and compared it to scPPIN, Switchde and standard single-cell differential expression analyses (Supplementary Information, Supplementary Tables 1–9 and Supplementary Fig. 5). Our experiments demonstrate that Scellnetor finds relevant module-specific and mechanism-specific genes and associated GO terms that could not be detected by any of the compared approaches.

Note that we have implemented a moving average function to remove noise, and we hypothesize that using a pseudotemporal ordering of the data points (single cells) before smoothing is more informative than using an arbitrary sorting key. To investigate how different sorting keys help define the resulting modules, we compared two Scellnetor-generated module sets that were based on the same cell sets and generated using two different sorting keys (Supplementary Information and Supplementary Table 10). The comparison shows that the final modules of the two module searches are different, which suggests that choosing a sorting key is an important step in the Scellnetor pipeline. In general, we recommend using pseudotime as the sorting key for single-cell ordering before clustering.

Scellnetor utilizes per default networks from the BioGRID to constrain the agglomerative hierarchical clustering. As mentioned, any user-uploaded network should use human Entrez IDs as node identifiers, because Scellnetor converts human gene symbols, human Ensemble stable IDs or mouse gene symbols to human Entrez IDs. However, because of overlapping gene symbol usage and the identical human–mouse gene orthology of certain genes, some human and mouse gene symbols are mapped to the same human Entrez IDs. In these cases, Scellnetor will automatically represent these Entrez IDs by the first instance of the genes in question that it meets in its search. As this might introduce minor gene imprecisions in the resulting modules, we recommend that users integrate a list with either human Entrez IDs or Ensemble stable IDs in the applied ANNDATA object. This will preserve the highest number of genes that can be used for clustering. Although Scellnetor can utilize a user-uploaded network, it only allows undirected edges in the networks and thus cannot use all information stored in, for example,

gene regulatory networks. Finally, Scellnetor can only compare two single-cell sets, which hampers the analysis of biological phenomena where, for example, one progenitor cell develops into more progenies. Future versions of Scellnetor will account for these limitations by allowing usage of mouse interaction networks, integrate directionality in the clustering process and enable comparisons of more than two single-cell sets. One could also imagine having the networks not provided by the user but inferred directly from the single-cell data. Such a methodology could be developed and integrated in the future.

Methods

Selecting cells on the canvas plot. *Generating paths.* When drawing paths with Scellnetor, users select cells on the canvas plot, which will be connected automatically with a path. The order in which the cells are selected directs the construction of the path through the plot. Let Z be the set of all cells and $X=\{x_1,x_2,x_3\ldots x_n\}$ the set of user-selected cells. The path is created by forming a path between x_1 and x_2 followed by a path between x_2 and x_3, and so on. A path between cell x_i and cell x_{i+1} is generated by finding the cell x_{ia} that is closest to x_i and closer to x_{i+1} than x_i is. This is followed by finding the cell x_{ib} that is closest to x_{ia} and closer to x_{i+1} than x_{ia} is. This is repeated until a sub-path between x_i and x_{i+1} has been drawn. This is done for all possible pairs of x_i and x_{i+1} in X. The result is a minimal greedy path, which is the path connecting the cells in X with the fewest number of cells possible when following the rules of the greedy path algorithm. The cells in the resulting path are stored in the set, X_{path}.

The paths can be expanded (that is, made thicker to include more cells) by including a user-defined percentage of cells that neighbour cells in the minimal greedy path. This is achieved by greedily adding the closest cells to cells that are already in X_{path} until the desired number of cells is reached. Euclidean distance is used as the distance metric and $X_{\text{expand}}=X_{\text{path}}$ when $i=0$ for $X_{\text{path}}=\{x_1,x_2,x_3\ldots x_N\}$ where N is the number of cells in the minimal greedy path. Additionally, the user can even out the sizes of the generated paths so they contain the same number cells.

Finding similar-sized paths. Scellnetor has integrated functionalities that allow users to even out the sizes of their drawn paths. If two paths have been drawn and one has been expanded by inclusion of the *pct* percent closest points, and the other is in its minimal greedy path form, then let X_{expanded} be the set containing the cells of the expanded path and Y_{path} be the set containing the cells of the path, which is in its minimal greedy path form. To even out the sizes of X_{expanded} and Y_{path}, first expanded by the inclusion of the *pct* percent closest cells for every $y_i \in Y_{\text{path}}$. The result is the set Y_{expanded}. If $|Y_{\text{expanded}}|=|X_{\text{expanded}}|$, then the objective has been reached. If $|Y_{\text{expanded}}|<|X_{\text{expanded}}|$, a cell from Y_{expanded} is chosen at random and its closest neighbour that is not already in Y_{expanded} is added to the set. This is repeated until $|Y_{\text{expanded}}|=|X_{\text{expanded}}|$. Adding cells like this will promote a uniform expansion of the path. If $|Y_{\text{expanded}}|>|X_{\text{expanded}}|$, distances from the cells in Y_{path} to cells in $Y_{\text{expanded}}-Y_{\text{path}}$ are calculated. The cells in $Y_{\text{expanded}}-Y_{\text{path}}$ that have the average longest distance to all cells in Y_{path} are removed until $|Y_{\text{expanded}}|=|X_{\text{expanded}}|$. Users can even out their paths even if both paths have been expanded or none of them has. However, it is not possible to get a path smaller than the path's minimal greedy path form.

Selecting pre-defined clusters. Users select cells on a canvas plot, and the cluster to which the cell belongs is coloured in accordance with the chosen set colour. Switching between colours and selecting cells in different clusters makes it possible to create two distinct sets of clusters that can be compared.

Hyper-similarity matrix. *Extraction of cells and smoothing of timelines.* A hyper-similarity matrix is only calculated when two sets are compared. For the following elucidation of the hyper-similarity matrix computation, it will be assumed that a user has made two paths on a canvas plot, path A and path B. The cells in path A and path B are extracted and their expression patterns are compiled into matrices A and B, respectively (Fig. 2a). Matrix A has size $g\times m$ and B has size $g\times l$, where g is the number of genes, and m and l are the numbers of cells in paths A and B, respectively. The orders of the cells in the matrices are sorted using pseudotime as the sorting key. To reduce noise, a moving average function is per default applied on the rows of A and B, which results in new matrices called A_{mean} and B_{mean} with sizes $g\times m'$ and $g\times l'$, respectively. The elements in row i and column j of A_{mean} and B_{mean} are found using the following moving average formula:

$$A_{\text{mean}_{i,j}} = \frac{1}{a}\sum_{z=0}^{a-1} A_{i,j+z} \tag{1}$$

where a is the user-defined size of the moving average window and $A_{i,j}$ is the element in row i and column j of matrix A. Equation (1) is applied on every row of A and B and every jth column in the range $\{j\in\mathbb{N}|1\leq j\leq m-a+1\}$ for A and every jth column in the range $\{j\in\mathbb{N}|1\leq j\leq l-a+1\}$ for B. In cases where the two paths are of different lengths, the moving average window of the longer path is adjusted such that both paths contain the same number of smoothed values.

Uniform manifold approximation and projection of single-cell gene expression. Before the gene expression patterns in the two sets are compared, the smoothed data in A_{mean} and B_{mean} are further dimensionality-reduced utilizing uniform manifold approximation and projection (UMAP)[50]. However, it is possible to uncouple UMAP from the Scellnetor pipeline, such that only A_{mean} and B_{mean} are used for the hyper-similarity matrix computation (see section 'Computing the hyper-similarity matrix without UMAP'). Researchers can choose one of the four distance metrics—Euclidean distance, Manhattan distance, Minkowski distance or Correlation—when computing relationships between cellular expression patterns. Scellnetor sets the UMAP parameters *n_neighbors*=30, *min_dist*=0.0 and *random_state*=42 and allow users to choose the number of dimensions to which the input data should be reduced. The remaining UMAP parameters are set to their default values. A_{mean} and B_{mean} are concatenated before being transformed by UMAP, as the gene expression patterns from the two matrices will thereby be high-dimensional coordinates on the same manifold structure. This will provide more meaningful inter-coordinate Euclidean-based distances of the dimensionality-reduced data. UMAP outputs a different coordinate landscape depending on the order of the concatenation. For example, the Euclidean distances between coordinates resulting from a UMAP-embedding of the concatenation of A_{mean} and B_{mean} are similar, but slightly different from the Euclidean distances between coordinates resulting from a UMAP-embedding of the concatenation of B_{mean} and A_{mean}.

Hence, to retain determinism of the Scellnetor clustering, a dimensionality reduction is conducted on both the concatenation of A_{mean} and B_{mean} and the concatenation of B_{mean} and A_{mean}. The resulting dimensionality-reduced matrices, $C_{\text{small}_{\text{AB}}}$ and $C_{\text{small}_{\text{BA}}}$, both have size $2g\times 2$, where g is the number of genes in A_{mean} and B_{mean}. $C_{\text{small}_{\text{BA}}}$ is redefined as the concatenation of $C_{\text{small}_{\text{BA}}}[g:]$ on top of $C_{\text{small}_{\text{BA}}}[:g]$, as the order of the genes in the two coordinate sets should be identical for the further processing. $C_{\text{small}_{\text{BA}}}[g:]$ is the slice of $C_{\text{small}_{\text{BA}}}$ that contains the last g rows and $C_{\text{small}_{\text{BA}}}[:g]$ is the slice of $C_{\text{small}_{\text{BA}}}$ that contains the first g rows.

Hadamard product of distance matrix quadrants. When comparing the two paths—path A and path B—a hyper-similarity matrix is computed. $C_{\text{small}_{\text{AB}}}[:g]$ contains the moving-average-modified and dimensionality-reduced gene expression patterns from path A, and $C_{\text{small}_{\text{AB}}}[g:]$ contains the moving-average-modified and dimensionality-reduced gene expression patterns from path B. The same holds true for $C_{\text{small}_{\text{BA}}}[:g]$ and $C_{\text{small}_{\text{BA}}}[g:]$, respectively. Again, g is the number of genes in A_{mean} and B_{mean}. A $C_{\text{small}_{\text{AB}}}$ versus $C_{\text{small}_{\text{AB}}}$ distance matrix, $D_{\text{small}_{\text{AB}}}$, and a $C_{\text{small}_{\text{BA}}}$ versus $C_{\text{small}_{\text{BA}}}$ distance matrix, $D_{\text{small}_{\text{BA}}}$, are calculated using Euclidean distance as the metric. Again, to retain determinism of the Scellnetor clustering approach, a distance matrix, D_{small}, is found by

$$D_{\text{small}} = \frac{D_{\text{small}_{\text{AB}}} \oplus D_{\text{small}_{\text{BA}}}}{2} \tag{2}$$

where the values of D_{small} are modified by

$$D'_{\text{small}} = \frac{D_{\text{small}}}{\max(D_{\text{small}}) + 10^{-6}} \tag{3}$$

In the matrix D_{small} in our example (Fig. 2), the upper left quadrant corresponds to a path A versus path A Euclidean distance matrix and the lower right quadrant corresponds to a path B versus path B Euclidean distance matrix. The upper triangle of the upper right quadrant corresponds to the upper triangle of a path A versus path B Euclidean distance matrix and the lower triangle of the upper right quadrant corresponds to the lower triangle of a path B versus path A Euclidean distance matrix.

The diagonal of D_{small} is zeroed out. The diagonal of the upper right quadrant of D_{small} is zeroed out as well, because similar values will be used to weigh the hyper-similarity matrix at a later step. The upper triangular matrix of the upper left quadrant is defined as D_{AA}, the upper triangular matrix of the lower right quadrant is defined as D_{BB}, the upper triangular matrix of the upper right quadrant is defined as D_{AB}, and the lower triangular matrix of the upper right quadrant is defined as D_{BA}. The values in D_{AA} and D_{BB} are reversed by

$$D'_{XX} = |(D_{XX} - 1)| \tag{4}$$

where D_{XX} is the matrix and 1 is subtracted from it to set the values of the matrix in the range]0;1]. In this way, no values are zeroed out after reversion of the distance values and the relative distance differences remain unchanged. If the user wants modules that have similar expression patterns in path A and path B, respectively, and similar expression patterns when comparing path A and path B, then D_{AB} and D_{BA} are updated using equation (4) (module type 1).

The matrices D_{AB} and D_{BA} remain as initially defined if the resulting modules should have similar expression patterns in path A and path B, respectively, and dissimilar expression patterns when comparing path A against path B (module type 2). To compress matrices D_{AA}, D_{BB}, D_{AB} and D_{BA} into a single 'pre-hyper-similarity matrix', D, the four matrices are element-wise multiplied as follows:

$$D = D_{\text{AA}} \odot D_{\text{BB}} \odot D_{\text{AB}} \odot D_{\text{BA}}^{\top} \tag{5}$$

where $\odot$ is the Hadamard product. The diagonal and the lower triangular matrix of D are zeroed out. Now, the matrix D only needs a few processing steps before it is ready for the clustering as a hyper-similarity matrix.

Weighting values of the hyper-similarity matrix. In D, the values depend on the Euclidean distances between the gene expression patterns from the cells in path A and on the Euclidean distances between the gene expression patterns from the cells in path B. The matrices A_{mean} and B_{mean}, derived from the two paths, contain the moving-average-modified expression values of the same genes in the same order, but measured as the cells in the two sets follow different differentiation trajectories. The matrix element at $D_{i,j}$ depends on the expression values of gene i in A_{mean} and B_{mean}, respectively, and on the expression values of gene j in A_{mean} and B_{mean}, respectively. The matrix element at $D_{i,j}$ should also depend on the Euclidean distance of gene i in A_{mean} versus gene i in B_{mean} and gene j in A_{mean} versus gene j in B_{mean}. Until now, these values have been zeroed out and ignored, but they will be used to weigh the matrix D in the following steps (Fig. 2).

For example, if a user applies Euclidean distance as metric and wants to find modules of module type 1, then the value in row i and column i of D should be 'penalized' if $A_{\text{mean}i,:}$ and $B_{\text{mean}i,:}$ are far away from each other in the Euclidean space and/or if $A_{\text{mean}j,:}$ and $B_{\text{mean}j,:}$ are far away from each other in the Euclidean space. $A_{\text{mean}i,:}$ indicate the entire row i of A_{mean}. To obtain the Euclidean distances of genes with identical IDs in A_{mean} and B_{mean}, new coordinate sets are defined: $C_{AB_A} = C_{\text{small}_{AB}}[:g]$, $C_{AB_B} = C_{\text{small}_{AB}}[g:]$, $C_{BA_A} = C_{\text{small}_{BA}}[:g]$ and $C_{BA_B} = C_{\text{small}_{BA}}[g:]$, where g is the number of genes in A_{mean} and B_{mean}. Two vectors, $\mathbf{v}_{AB}$ and $\mathbf{v}_{BA}$, are calculated by finding the Euclidean distances between every identically indexed row of C_{AB_A} and C_{AB_B} and every identically indexed row of C_{BA_A} and C_{BA_B}, respectively, such that $|v_{AB}| = |v_{BA}| = g$. A vector, $\mathbf{v}$, is defined by

$$\mathbf{v} = \frac{\mathbf{v}_{AB} + \mathbf{v}_{BA}}{2} \tag{6}$$

If the aim is to find modules of module type 1, then v is normalized as follows

$$\mathbf{v}' = \mathbf{v} - \min(\mathbf{v}) \tag{7}$$

$$\mathbf{v}'' = \left| \left(\frac{\mathbf{v}'}{(\max(\mathbf{v}') + 10^{-6})} - 1 \right) \right| \tag{8}$$

If modules of module type 2 are the objective, then v is normalized by

$$\mathbf{v}' = \mathbf{v} - \left(\min(\mathbf{v}) - 10^{-6}\right) \tag{9}$$

$$\mathbf{v}'' = \frac{\mathbf{v}'}{\max(\mathbf{v}')} \tag{10}$$

In both scenarios, $\mathbf{v}$ will be in the range]0;1], which implies that nothing will be zeroed out by the weighting of $\mathbf{v}$. The hyper-similarity matrix is found by weighting row $D_{i,:}$ by element v_i and weighting column $D_{:,j}$ by element v_j, where i is in the range $\{i \in \mathbb{N} | 1 \leq i \leq g\}$ and j is in the range $\{j \in \mathbb{N} | 1 \leq j \leq g\}$. As a final step, D is normalized by

$$D' = D - \left(\min(D) - 10^{-6}\right) \tag{11}$$

$$D'' = \frac{D'}{\max(D')} \tag{12}$$

such that all possible gene–gene hyper-similarities are in the range]0;1].

Computing the hyper-similarity matrix without UMAP. Scellnetor allows users to choose whether they want to use UMAP on top of the moving average function for further dimensionality-reduction of the data. When UMAP is uncoupled from the Scellnetor pipeline, the hyper-similarity matrix computation includes the following steps presented:

1. The lengths of the genes' pseudotemporal expression patterns are reduced using the moving average function.
2. The resulting dimensionality-reduced expression data are used for calculation of a distance matrix, which is modified using equation (3).
3. Depending on whether the user wishes to find modules of module type 1 or module type 2, the relevant quadrants of the distance matrix are converted to similarities by equation (4).
4. The element-wise product of the quadrants is found using equation (5).
5. The vector containing the distances of identical genes but from the different sets are modified by equations (7) and (8) or by equations (9) and (10), depending on whether modules of module type 1 or module type 2, respectively, are desired.
6. Every row and column of the matrix found in step 4 are multiplied by the elements of the vector found in step 5.
7. The final matrix is normalized by equations (11) and (12).

Interpreting the hyper-similarity values. When comparing two sets of cells, Scellnetor computes a hyper-similarity matrix. The hyper-similarity matrix contains information on how the genes are expressed relative to each other within a set of cells as well as between the compared sets of cells. The main goal of our similarity function is to find genes whose expression patterns are highly similar and conserved within each cell set, but dissimilar between the cell sets (module type 2). As an example, a high hyper-similarity value (close to 1) between two genes (*gene*1 and *gene*2) that are connected in a network implies the following if, for example, Euclidean distance is used as the metric:

1. Both genes express pseudotimelines in two user-drawn paths, path A and path B: $gene1_A$ and $gene2_A$ are the pseudotimelines from path A and $gene1_B$ and $gene2_B$ are the pseudotimelines from path B.
2. The genes $gene1_A$ and $gene2_A$ are in close proximity to each other and $gene1_B$ and $gene2_B$ are in close proximity to each other in Euclidean space.
3. The distance between $gene1_A$ and $gene2_B$ is large and the distance between $gene1_B$ and $gene2_A$ is large in Euclidean space.
4. The distance between $gene1_A$ and $gene1_B$ is large and the distance between $gene2_A$ and $gene2_B$ is large in Euclidean space (weighting by values in vector $\mathbf{v}$).

Generating results. Users can inspect the main results online and download all data that were produced in the Scellnetor pipeline. The main results are as follows:

1. Modules of genes as PDF files and two edge lists for every module in CSV file format where nodes are denoted as both human Entrez IDs and human gene symbols.
2. One plot per module of mean expression of the genes together with the 95% confidence interval in PDF file format. On the x axis is 'Moving-average-modified number of single cells'. The cells are arranged after the variable selected as the sorting key. The y axis shows 'Normalized moving-average-modified gene expression'. It has been normalized such that the highest value of the concatenation of A_{mean} and B_{mean} is 1 and the smallest is 0. This normalization only serves visualization purposes and is done after completion of the hyper-similarity calculation.
3. TSV files containing statistically significant GO terms associated with the modules. They are generated using GOATOOLS[51], which uses Fisher's exact test to calculate P values and the Benjamini–Hochberg procedure to adjust P values.

Constrained agglomerative hierarchical clustering. Scellnetor uses a constrained agglomerative hierarchical clustering algorithm. This means that it iteratively clusters genes together pairwise in order of descending hyper-similarity until all items have been assigned to a cluster or a threshold has been reached. The threshold is defined via user-chosen parameters that defines the minimum module size and the minimum number of modules. As the clustering is constrained by the interactions of a network, Scellnetor will always output connected subnetworks (modules) of genes as clusters. Per default, the clustering is constrained by biological networks, for example from BioGRID or uploaded by the user. For the clustering, the constraint implies that (1) two single genes can only be fused into a cluster if they are connected by an edge (are neighbours) in the network, (2) a single gene needs to neighbour a gene in a cluster before it can be added to the cluster and (3) when merging two clusters, they need to have at least one gene each that are neighbours in the graph. The possible connections of a cluster are equal to the sum of all connections of genes in that cluster.

When Scellnetor encounters a tie (two or more genes with the same short distance or high similarity), it will choose the first instance that appears in the data matrix. However, when utilizing the hyper-similarity matrix for the clustering, tie encounters are unlikely, as the hyper-similarities are based on multiplications of many inter-gene distances or similarities. Scellnetor is limited to only using genes that are present in the used network and in the uploaded data. This means that genes that cannot be mapped to the applied network will be ignored, and genes that are not in the used ANNDATA object will not be included in the clustering nor in the results.

Extraction and conversion of genes from an uploaded file. Per default, Scellnetor constrains the clustering by the interaction networks from BioGRID. However, the user can also utilize any other network provided in a tab separated edge list. The only restriction is the required usage of human Entrez IDs as node identifiers. The same restriction applies to the genes of uploaded H5AD files (the file format of the ANNDATA object), which are also required to be given as Entrez IDs (or as identifiers that can be converted to Entrez IDs) and are additionally available as nodes in the network. All IDs without a corresponding network node will be discarded from further analysis. Scellnetor automatically recognizes human Entrez IDs, human gene symbols, human Ensemble stable IDs and mouse gene symbols.

Pre-processing of data for Scellnetor haematopoiesis study. Using the scRNA-seq data from the 19 cell groups defined by ref. [15] (GSE72857), we created an initial count matrix and made sure that we could reproduce the measured group-wise average gene expressions from ref. [15] before we constructed an ANNDATA object. We then generated cell maps and computed pseudotime, which was stored in the ANNDATA object. Note that we could not identify 10 genes from their list with group-wise average gene expressions, because they were not annotated as any known gene. These anonymous genes were omitted from our analysis. This gave us a total of 2,730 single cells that expressed 3,451 genes. We pre-processed the count matrix in the same way as in the study in ref. [52], with the exception of using the top 1,726 (3,451/2 ≈ 1,726) most variable genes instead of the top 1,000.

Pre-processing of data for the Scellnetor exhausted CD8 T cell study. Using the scRNA-seq data from the five clusters the authors identified in ref. [10] (GSE137007) we generated an ANNDATA object. The data were already pre-processed using the Seurat[7] package. Additionally, we computed a diffusion map and calculated pseudotime. The 'start cell' for the pseudotime inference was the cell in the progenitor cluster (cluster 1, Fig. 4a) that was furthest away from all progenies. The Supplementary Information provides a discussion of the rationale for this.

Data availability

The scRNA-seq data used for the Scellnetor haematopoiesis analysis is from GEO (GSE72857). The scRNA-seq used for the clustering of exhausted CD8 T cells in chronic infections is also from GEO (GSE137007). Scellnetor results can be downloaded from ref. [53] and from GitLab.

Code availability

Scellnetor is freely available as an online tool at https://exbio.wzw.tum.de/scellnetor/ and can be downloaded as a standalone program from ref. [53] and from GitLab.

Received: 31 May 2020; Accepted: 15 January 2021;
Published online: 22 February 2021

References

1. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* https://doi.org/10.1186/s13059-019-1663-x (2019).
2. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* https://doi.org/10.1038/nmeth.3971 (2016).
3. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet*. https://doi.org/10.1038/s41576-018-0088-9 (2019).
4. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btv325 (2015).
5. Tritschler, S. et al. Concepts and limitations for learning developmental trajectories from single cell genomics. *Development* https://doi.org/10.1242/dev.170506 (2019).
6. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* https://doi.org/10.1186/s13059-017-1382-0 (2018).
7. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies and species. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.4096 (2018).
8. Guo, M., Wang, H., Potter, S. S., Whitsett, J. A. & Xu, Y. SINCERA: a pipeline for single-cell RNA-Seq profiling analysis. *PLoS Comput. Biol.* https://doi.org/10.1371/journal.pcbi.1004575 (2015).
9. Chen, G., Ning, B. & Shi, T. Single-cell RNA-seq technologies and related computational data analysis. *Front. Genet*. https://doi.org/10.3389/fgene.2019.00317 (2019).
10. Kanev, K. et al. Proliferation-competent Tcf1^{+} CD8 T cells in dysfunctional populations are CD4 T cell help independent. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.1902701116 (2019).
11. Guo, X. et al. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat. Med*. https://doi.org/10.1038/s41591-018-0045-3 (2018).
12. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* https://doi.org/10.15252/msb.20188746 (2019).
13. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med*. https://doi.org/10.1038/s12276-018-0071-8 (2018).
14. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet*. https://doi.org/10.1038/nrg3833 (2015).
15. Paul, F. et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* https://doi.org/10.1016/j.cell.2015.11.013 (2015).
16. Campbell, K. R. & Yau, C. Switchde: inference of switch-like differential expression along single-cell trajectories. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btw798 (2017).
17. Matsumoto, H. et al. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-seq during differentiation. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btx194 (2017).
18. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* https://doi.org/10.1038/nmeth.4463 (2017).
19. Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst*. https://doi.org/10.1016/j.cels.2017.08.014 (2017).
20. Alcaraz, N. et al. De novo pathway-based biomarker identification. *Nucleic Acids Res*. https://doi.org/10.1093/nar/gkx642 (2017).
21. Breitling, R., Amtmann, A. & Herzyk, P. Graph-based iterative group analysis enhances microarray interpretation. *BMC Bioinformatics* https://doi.org/10.1186/1471-2105-5-100 (2004).
22. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A. F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* https://doi.org/10.1093/bioinformatics/18.suppl_1.S233 (2002).
23. Klimm, F. et al. Functional module detection through integration of single-cell RNA sequencing data with protein–protein interaction networks. *BMC Genomics* https://doi.org/10.1186/s12864-020-07144-2 (2020).
24. Oughtred, R. et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Res*. https://doi.org/10.1093/nar/gky1079 (2019).
25. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS ONE https://doi.org/10.1371/journal.pone.0098679 (2014).
26. Ribeiro, D. M. & Sonati, M. F. Regulation of human α-globin gene expression and α-thalassemia. *Genet. Mol. Res*. https://doi.org/10.4238/vol7-4gmr472 (2008).
27. Shah, D. I. et al. Mitochondrial Atpif1 regulates haem synthesis in developing erythroblasts. *Nature* https://doi.org/10.1038/nature11536 (2012).
28. Tanimura, A. et al. Mitochondrial activity and unfolded protein response are required for neutrophil differentiation. *Cell. Physiol. Biochem*. https://doi.org/10.1159/000491464 (2018).
29. Michalak, M., Groenendyk, J., Szabo, E., Gold, L. I. & Opas, M. Calreticulin, a multi-process calcium-buffering chaperone of the endoplasmic reticulum. *Biochem. J.* https://doi.org/10.1042/BJ20081847 (2009).
30. Sun, S. et al. Inhibition of prolyl 4-hydroxylase, beta polypeptide (P4HB) attenuates temozolomide resistance in malignant glioma via the endoplasmic reticulum stress response (ERSR) pathways. *Neuro. Oncol*. https://doi.org/10.1093/neuonc/not005 (2013).
31. Vargas, A., Roux-Dalvai, F., Droit, A. & Lavoie, J. P. Neutrophil-derived exosomes: a new mechanism contributing to airway smooth muscle remodeling. *Am. J. Resp. Cell Mol. Biol*. https://doi.org/10.1165/rcmb.2016-0033OC (2016).
32. Winterbourn, C. C., Kettle, A. J. & Hampton, M. B. Reactive oxygen species and neutrophil function. *Annu. Rev. Biochem*. https://doi.org/10.1146/annurev-biochem-060815-014442 (2016).
33. Scapini, P. et al. CXCL1/macrophage inflammatory protein-2-induced angiogenesis in vivo is mediated by neutrophil-derived vascular endothelial growth factor-A. *J. Immunol*. https://doi.org/10.4049/jimmunol.172.8.5034 (2004).
34. Gaudry, M. et al. Intracellular pool of vascular endothelial growth factor in human neutrophils. *Blood* https://doi.org/10.1182/blood.v90.10.4153 (1997).
35. Scapini, P., Calzetti, F. & Cassatella, M. A. On the detection of neutrophil-derived vascular endothelial growth factor (VEGF). *J. Immunol. Methods* https://doi.org/10.1016/S0022-1759(99)00170-2 (1999).
36. Jacob, C. O. et al. Lupus-associated causal mutation in neutrophil cytosolic factor 2 (NCF2) brings unique insights to the structure and function of NADPH oxidase. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.1113251108 (2012).
37. Nauseef, W. M. Assembly of the phagocyte NADPH oxidase. *Histochem. Cell Biol*. https://doi.org/10.1007/s00418-004-0679-8 (2004).
38. Groemping, Y. & Rittinger, K. Activation and assembly of the NADPH oxidase: a structural perspective. *Biochem. J*. https://doi.org/10.1042/BJ20041835 (2005).
39. Liu, X. et al. Regulation of mitochondrial biogenesis in erythropoiesis by mTORC1-mediated protein translation. *Nat. Cell Biol*. https://doi.org/10.1038/ncb3527 (2017).
40. Szentirmay, M. N. Survey and summary: spatial organization of RNA polymerase II transcription in the nucleus. *Nucleic Acids Res*. https://doi.org/10.1093/nar/28.10.2019 (2000).
41. Wherry, E. J. T-cell exhaustion. *Nat. Immunol*. https://doi.org/10.1038/ni.2035 (2011).
42. Hecht, I. et al. ILDR2 is a novel B7-like protein that negatively regulates T-cell responses. *J. Immunol*. https://doi.org/10.4049/jimmunol.1700325 (2018).
43. Long, A. H. et al. 4-1BB costimulation ameliorates T-cell exhaustion induced by tonic signaling of chimeric antigen receptors. *Nat. Med*. https://doi.org/10.1038/nm.3838 (2015).
44. Krishna, S. et al. Chronic activation of the kinase IKKβ impairs T-cell function and survival. *J. Immunol*. https://doi.org/10.4049/jimmunol.1102429 (2012).
45. Peled, M. et al. EF hand domain family member D2 is required for T-cell cytotoxicity. *J. Immunol*. https://doi.org/10.4049/jimmunol.1800839 (2018).
46. Lando, D.et al. FIH-1 is an asparaginyl hydroxylase enzyme that regulates the transcriptional activity of hypoxia-inducible factor. *Genes Dev*. https://doi.org/10.1101/gad.991402 (2002).
47. Kim, J. W., Tchernyshyov, I., Semenza, G. L. & Dang, C. V. HIF-1-mediated expression of pyruvate dehydrogenase kinase: a metabolic switch required for cellular adaptation to hypoxia. *Cell Metab*. https://doi.org/10.1016/j.cmet.2006.02.002 (2006).

48. Papandreou, I., Cairns, R. A., Fontana, L., Lim, A. L. & Denko, N. C. HIF-1 mediates adaptation to hypoxia by actively downregulating mitochondrial oxygen consumption. *Cell Metab*. https://doi.org/10.1016/j.cmet.2006.01.012 (2006).
49. Doedens, A. L. et al. Hypoxia-inducible factors enhance the effector responses of CD8$^+$ T cells to persistent antigen. *Nat. Immunol*. https://doi.org/10.1038/ni.2714 (2013).
50. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: uniform manifold approximation and projection. *J. Open Source Softw*. https://doi.org/10.21105/joss.00861 (2018).
51. Klopfenstein, D. V. et al. GOATOOLS: a Python library for gene ontology analyses. *Sci. Rep*. https://doi.org/10.1038/s41598-018-28948-z (2018).
52. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun*. https://doi.org/10.1038/ncomms14049 (2017).
53. Grønning, A. G. B. Scellnetor_standalone_scripts_data (2021); https://doi.org/10.5281/ZENODO.4419550

Acknowledgements

J.B. and A.G.B.G. received funding from J.B.'s VILLUM Young Investigator grant no. 13154. The work of J.B. and T.K. was further funded by H2020 project RepoTrial (no. 777111). The work of R.R. and J.B. was partially funded by H2020 project FeatureCloud (no. 826078). J.B. and T.K. are grateful for financial support from BMBF project Sys_Care. M.O. is grateful for financial support from the Collaborative Research Center SFB924.

Author contributions

A.G.B.G. developed and implemented the clustering algorithm of the Scellnetor tool. A.G.B.G. developed and implemented all basic backend functionalities of the webtool. A.G.B.G., J.L. and M.O. further developed the webtool. K.K., T.K. and D.Z. tested the webtool, provided critical feedback and, together with A.G.B.G., used Scellnetor to generate the biomedical results presented in the manuscript. All authors contributed equally to writing and improving the paper. A.G.B.G., J.B. and R.R. conceived the idea of the Scellnetor pipeline.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-021-00025-y.

Correspondence and requests for materials should be addressed to A.G.B.G. or J.B.

Peer review information *Nature Computational Science* thanks Florian Klimm and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Yann Sweeney was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.