MethodsX 9 (2022) 101778

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/mex

Method Article

ORIGINS: A protein network-based approach to quantify cell pluripotency from scRNA-seq data

Daniela Senra, Nara Guisoni, Luis Diambra*

Centro Regional de Estudios Genómicos, Universidad Nacional de La Plata, Argentina

A B S T R A C T

Trajectory inference is a common application of scRNA-seq data. However, it is often necessary to previously determine the origin of the trajectories, the stem or progenitor cells. In this work, we propose a computational tool to quantify pluripotency from single cell transcriptomics data. This approach uses the protein-protein interaction (PPI) network associated with the differentiation process as a scaffold and the gene expression matrix to calculate a score that we call differentiation activity. This score reflects how active the differentiation network is in each cell. We benchmark the performance of our algorithm with two previously published tools, LandSCENT (Chen et al., 2019) and CytoTRACE (Gulati et al., 2020), for four healthy human data sets: breast, colon, hematopoietic and lung. We show that our algorithm is more efficient than LandSCENT and requires less RAM memory than the other programs. We also illustrate a complete workflow from the count matrix to trajectory inference using the breast data set.

- ORIGINS is a methodology to quantify pluripotency from scRNA-seq data implemented as a freely available R package.
- ORIGINS uses the protein-protein interaction network associated with differentiation and the data set expression matrix to calculate a score (differentiation activity) that quantifies pluripotency for each cell.

A R T I C L E I N F O
Method name: ORIGINS
Keywords: Stem cells, scRNA-seq, Protein-protein interaction networks, Trajectory inference
Article history: Available online 1 July 2022

* Corresponding author.
E-mail address: ldiambra@gmail.com (L. Diambra).

https://doi.org/10.1016/j.mex.2022.101778

Specifications table

Subject area:	Bioinformatics
More specific subject area:	Pluripotency quantification from single cell transcriptomics
Name of your method:	ORIGINS
Name and reference of original method:	Not applicable
Resource availability:	https://github.com/danielasenraoka/ORIGINS

Method details

Background

Recent advances in single cell RNA sequencing (scRNA-seq), that allow the transcriptional profiling of single cells, offer a promising capability to explain developmental processes. The ability to quantify pluripotency is relevant to comprehend differentiation, cell lineages and lineage hierarchy. This task may also be important for cancer research to identify cancer stem cells, which have been suggested to be responsible for metastasis, remission and resistance to therapies [3]. It may also be a critical step to perform trajectory inference, a popular application of single cell transcriptomics to unveil differentiation processes.

Several algorithms were developed to reconstruct differentiation pathways by using scRNA-seq. In 2019 Saelens et al. reported the existence of more than 70 trajectory analysis techniques [4] and in recent years many more have emerged [5–7]. Many techniques require prior information to infer the trajectory, such as a starting or root cell [8,9]. Prior biological information can help the method find the correct trajectory but on the other hand, incorrect prior knowledge can bias the trajectory. Traditionally, previously known stemness markers are used to identify the starting cell, but it is not always feasible due to the high drop-out rate of the scRNA-seq technique. Moreover, stemness markers depend on the tissue and the developmental stage and are not always available for all cases.

In this sense, in order to quantify pluripotency a systems approach methodology was proposed by members of the Teschendorf Lab. They calculated a parameter called network entropy using a protein-protein interaction (PPI) network [10]. In following works the group deepened the research on quantifying stemness, proposing different alternatives to compute entropy [1,11,12]. The final version of this method, LandSCENT, is one of the most widespread entropy-based algorithms to compute differentiation potency from scRNA-seq data [1]. The algorithm uses a highly curated a PPI network as a scaffold and the normalized expression profile to estimate the signal entropy rate (SR) for each single cell. Briefly, the gene expression profile is used to calculate the edges of the PPI network, which can be interpreted as interaction probabilities, by invoking the mass action principle. This defines a stochastic matrix that is used to compute the signaling entropy rate over the weighted network based on the Shannon entropy measure. Authors state that differentiated cells have certain specific pathways activated leading to low entropy levels whilst pluripotent cells display a broad pattern of signaling pathways activated and do not express any preference for any particular lineage. In terms of single cell transcriptomics data, this translates as more heterogeneous gene expression profiles, resulting in high entropy levels.

Other publications address entropy without utilizing a PPI network as a scaffold. For instance, StemID aims to identify stem cells by using a score that combines the cluster median entropy and the number of inter-cluster links that define the topology of the lineage tree [13]. Another variation is SLICE, a Shannon entropy-based algorithm with some implementation modifications [14].

On the other hand, there are few techniques to quantify pluripotency that are not based on entropy. For example, Palmer et al. derived a stemness gene expression signature and used it to compute a stemness index over gene expression microarray samples [15]. They utilized the projection of the coordinates of an expression profile onto the first principal component of the gene space defined by the stemness gene signature as a relative measure of stemness.

In 2020 Gulati et al. developed a computational framework called CytoTRACE to identify stem cells using scRNA-seq data [2]. They found that the gene counts are generally correlated with the state of

differentiation. As scRNA-seq was designed to capture gene expression, they suggested determining the genes that correlate with the gene counts and creating a dataset-specific gene count signature (GCS). The authors evidenced a limitation in the case of quiescent stem cells. CytoTRACE is not useful for identifying quiescent stem cells due to their reduced metabolic activity and low RNA content, as is the case for hematopoietic stem cells [2].

Lately, increasing information about protein interactions and their role in biological processes has become available. In this work we exploit the fact that interactions between proteins underlie cell phenotypes [16]. With the widespread use of high-throughput sequencing technologies, many methodologies to integrate this type of data and network-based biological strategies have been implemented [11]. Here, we use the information of a PPI network to identify stem and progenitor cells. In this way, we define a score, which we call differentiation activity, that quantifies how active the differentiation PPI network is in each cell based on its expression profile and the set reactions involved on this PPI. This tool was implemented as an R package named ORIGINS.

ORIGINS: activity computation

Gene Ontology (GO) provides structured, controlled vocabularies and classifications associated with molecular functions, biological processes and cellular compartment [17]. Exploring GO annotations underlying a set of Differentially Expressed Genes (DEG) for insights into potential experimental meanings has become a widespread practice. Alternatively, to identify differential biological processes (BP) across a cell population we propose a strategy that exploits the set of biochemical reactions involved in the biological function of interest rather than several DEGs. In this sense, we build a protein-protein interaction (PPI) network associated with the gene products involved in Cell Differentiation BP (BP-GO: 0030154) as putative biochemical reactions.

To this end we consider 11,582 proteins from H. sapiens associated to BP-GO: 0030154 listed at QuickGo database [18], which include 87 child BP terms. These proteins constitute the nodes of the network. Further, we also consider the biochemical interactions listed in Pathways Commons (version 12), that integrate 2,424,055 interactions from 22 databases [19]. From this set of interactions we select the 191,072 interactions that involve only human proteins associated with BP-GO: 0030154, we do not take into account interactions involving chemical compounds nor other not-protein molecules. These PPI constitute the edges of the network and can be represented by an adjacency matrix A. Note that the PPI network used in our approach is an undirected graph, i.e., the links have no direction. Further, the PPI network does not distinguish if nodes are negative or positive regulators.

After building the PPI network associated with Cell Differentiation BP we define how to compute its activity level from a given expression profile. We consider that the activity level of the pathway is the accumulation of the biochemical reactions occurring in the pathway. Based on the Law of mass action for elementary reactions, we estimate the probability that a reaction will occur as the product of reactant concentrations, without considering the stoichiometric details. This approximation allows us to estimate the contribution of a PPI network edge between nodes a and b, $a \leftrightarrow b$, solely from the expression profile as: $x_a \times x_b$, where x_a and x_b are the expression levels associated with proteins A and B, respectively. Therefore, for a given expression profile $\{x_i\}$ a weighted edge matrix is defined $W_{ij} = A_{ij}x_ix_j$, where i and $j = 1, 2, ..., N_g$ and N_g is the number of genes in the pathway. A_{ij} is the adjacency matrix, a $N_g \times N_g$ square matrix, such that its element A_{ij} is one when there is an edge (when two proteins interact), and zero otherwise. Thus, the activity level associated with Cell Differentiation BP can be defined as:

$$P = \sum_{i,j=1}^{N_g} W_{ij}$$

For sc-RNAseq data each cell k has an expression profile associated and the corresponding weighted edge matrix W^k from which the activity level P^k associated with Cell Differentiation BP of the kth cell can be computed. Finally, activity levels are scaled so that activity takes values between 0 and 1 as follows: $P^k_{\text{scaled}} = \frac{P^k - \min(P^k)}{\max(P^k) - \min(P^k)}$, where $\min(P^k)$ and $\max(P^k)$ are the minimum and maximum activity levels among all cells.

Validation

We tested the performance of our proposed methodology to quantify pluripotency using four human data sets: breast epithelium [20], colon epithelium [21], bone marrow (hematopoietic cells) [22] and lungs [23]. Cell types were already annotated and provided as metadata for all datasets except the breast sample, which was annotated by ourselves according to the original publication [20]. For more details on breast cell annotation go to section Application to human mammary epithelium. Below there is a brief description of the data sets.

- Breast epithelium. Data from human breast epithelial cells that is publicly available in the GEO database (GSE113197) [20] was used. We utilized a healthy adult sample (Ind4).
- Colon epithelium. Data was downloaded from the GEO database (GSE125970). We used an adult human colon sample (Colon-2) to benchmark our algorithm.
- Hematopoietic. Publicly available scRNA-seq data of hematopoietic progenitors from human bone marrow was used. Raw data is available in the GEO database under the accession code GSE117498 [22]. We performed the pluripotency quantification using data from Donor A.
- Lungs. The scRNA-seq data included 19 lung samples, we focused on one of the five adult donors (D122), a 32 years old healthy male. Data was downloaded from the cellxgene Data Portal . Raw data is also available in the GEO database (GSE161383) [23].

We compared the performance of ORIGINS with two previously existing methodologies specifically developed for scRNA-seq data: the Signaling Entropy Rate (SR) from LandSCENT [1] and CytoTRACE [2]. Both algorithms are publicly available as R packages and were downloaded and installed from their official repositories. In addition, we propose a quicker approximation to estimate the differentiation activity by using the 2000 top highly variable features (HVF) of the expression matrices. These genes were determined using the Seurat function FindVariableFeatures() with selection.method = vst. Thus, we obtained a reduced gene expression matrix for each sample and we computed the activity on this normalized reduced matrix but with the same PPI network. This quantity is referred as activity HVF (ORIGINS).

All the code was implemented in R version 4.1.2 and the main packages used were LandSCENT version 0.99.5, CytoTRACE version 0.3.3 and Seurat version 4.1.0. The computer specifications were Kernel Version 5.13.0-30-generic, processor 12 $\times$ Intel® Core™ i7-8700 CPU @ 3.20GHz and 16 GiB of RAM.

The different pluripotency scores calculated for the breast sample are presented in the UMAP space in Figs. 1A–D. We investigated how these parameters vary according to cell types. Basal cells presented the highest average levels of SR (LandSCENT), activity (ORIGINS) and its approximation (Fig. 1E, G and H). This was expected since in the article where data was published the authors suggested the presence of breast stem cells within the basal population [20]. In the work where LandSCENT is presented, the breast sample was also used and the authors found that the majority of the multipotent cells are basal [1]. In the same way, it would be expected that Luminal 1 cells (L1) are in second place because they are more immature cells than Luminal 2 (L2), however this was only evidenced for activity (ORIGINS) and its approximation. The CytoTRACE score did not coincide with this order, it was found that on average the L1 cluster has the highest values followed by L2 and basal (Fig.1B and F).

In the case of the colon sample, where cells were already annotated (Fig. S1A), the four stemness scores are exhibited in Fig. S1B–E in the UMAP space. All the methods were limited in finding the stem cells, since the transit amplifying cells (TA) showed the highest score values on average and not the stem cells (Fig. S1F–I). In the intestine, stem cells divide asymmetrically, giving rise to another stem cell and a daughter cell called transit amplifying progenitor cell. Transit-amplifying cells are highly proliferative, they undergo a limited number of cell divisions and eventually differentiate into absorptive (enterocytes) or secretory (mucosal, enteroendocrine, Paneth cell) lineages [24–26].

The analyzed bone marrow cells were classified as hematopoietic stem cells (HSC), multipotent progenitors (MPP), multilymphoid progenitors (MLP), pre-B lymphocytes / Natural Killer cells (PREB/NK), megakaryocyte-erythroid progenitors (MEP), common myeloid progenitors (CMP) and granulocyte-monocyte progenitors (GMP) [22] as shown in Fig. S2A. The classical hematopoietic model

Table 1
Computational time comparison between SR (LandSCENT), CytoTRACE, activity (ORIGINS) and activity HVF (ORIGINS).

	SR (LandSCENT)	CytoTRACE	activity (ORIGINS)	activity HVF (ORIGINS)
Breast epithelium	3.43 h	28.27 s	2.95 h	3.00 min
Colon	4.11 h	44.19 s	3.11 h	3.11 min
Hematopoietic	8.12 h	59.95 s	6.24 h	4.09 min
Lungs	5.37 h	35.00 s	4.71 h	3.19 min

states that hematopoietic stem cells (HSC) give rise to all blood-cells types [27,28]. HSC cells are predominantly in a quiescent state [29] and can be activated as a response to the organism demand [30]. Self-renewing HSC occupy the apex of the hierarchy and originate different progenitors. Fig. S2B–E depict the calculated scores in the UMAP representation for several progenitor cell types. It should be noted that all the methods agreed that the highest average potency corresponded to the GMP as seen in Fig. S2F–I. HSCs would have been expected to be the most pluripotent, followed by MPPs, and then MLPs and CMPs. However, none of the methods seem adequate to order cells according to pluripotency based on the hematopoietic model, this could be a consequence of the different degree of quiescence and cell commitment of the hematopoietic stem/progenitor cells [2,31,32].

The lung is a complex organ that includes several distinct cell types, as seen in the lung sample we used (Fig. S3A). Regarding the lungs hierarchical organization, alveolar type II (AT2) cells are the best described stem cells and give rise to alveolar type I (AT1) cells [33]. In addition, club cells are stem cells that differentiate into ciliated cells [34,35]. Furthermore, basal [36–38], club-like and pulmonary neuroendocrine cells were identified as progenitor cells [39]. Fig. S3B–E exhibit the pluripotency scores in the UMAP space calculated for the lung sample. Among the approximately 30 cell types present in the dataset analyzed, AT2/club-like had the highest average SR (LandSCENT) and activity (ORIGINS) scores. The highest activity (ORIGINS) values were observed for AT2 cells, although not the highest mean activity score. All cells of interest, including basal and club cells, ranked in the top 8 mean scores for SR (LandSCENT), CytoTRACE, and activity (ORIGINS) as shown in Fig. S3F,G and H, respectively. The activity approximation, HVF activity (ORIGINS), failed to classify cell types based on the expected pluripotency. This may be because the use of the top 2000 HVF is not enough to provide a good activity approximation, probably due to the great diversity of cells analyzed (Fig. S3I).

Correlation with other methods

The Pearson correlation coefficient between the methodologies was computed for all data sets as shown in Fig. 2. All quantities were positively correlated. Taking into account the four data sets analysed, the average correlation coefficient between activity (ORIGINS) and SR (LandSCENT) was around 0.77, between activity (ORIGINS) and CytoTRACE 0.44, between SR (LandSCENT) and CytoTRACE 0.63 and between activity (ORIGINS) and its approximation activity HVF (ORIGINS) 0.67.

Efficiency

The elapsed real time for all algorithms and samples are reported in Table 1. On average, SR (LandSCENT) took approximately 25% longer than activity (ORIGINS) and CytoTRACE took less than 1%. As expected, activity HVF (ORIGINS) took less than 2% than activity (ORIGINS).

RAM usage

LandSCENT was the most memory demanding program. For example, an extra 6.4 Gb of RAM was required for the breast data set, while CytoTRACE needed an additional 5.8 Gb. This makes it difficult to quantify pluripotency for typical size samples (a few thousand cells) using standard personal computers. In this aspect, ORIGINS significantly outperforms the other programs as it does

not require additional RAM other than the vector where activity is stored, for the same data set the memory allocated by this vector was 26.6 Kb.

Simplicity and biological foundation

The main concept underlying the activity of the differentiation PPI network is relatively simple. Briefly, the differentiation activity of a cell is proportional to the sum of all the weights of the differentiation PPI network. The weights (edges) associated with two transcripts (nodes) are approximated as the multiplication of the expression levels of the associated proteins according to the law of mass action. Thus, an edge weight is greater if both nodes are highly expressed and vice versa. By adding all the edges of this differentiation network, we can quantify how active this network is.

ORIGINS can handle zero-containing expression matrices

Normalized expression matrices often have null elements (zeros), such as the obtained by using the R Seurat package normalization. Unlike the Signaling Entropy Rate (SR) [1], ORIGINS is not limited by the presence of zeros in the expression matrix. The user can provide any normalized non-negative expression matrix.

User friendly

The algorithm is easy to use. By typing four lines of code in R, the differentiation activity can be calculated:

```
install.packages("remotes") #if remotes package not installed
remotes::install_github("danielasenraoka/ORIGINS") library(ORIGINS)
diff_activity <-activity(expression_matrix, differentiation_edges)
```

Application to human mammary epithelium

We applied our methodology, ORIGINS, to identify stem cells in the human mammary gland. Below we describe all the steps performed, from the raw data to the inference of the differentiation trajectory. We used a scRNA-seq data set of human breast epithelial cells that is publicly available in the GEO database (GSE113197) [20]. This data set was acquired using the 10 $\times$ Genomics Chromium platform. It included approximately 25000 cells from four nulliparous women between 17 and 36 years old denoted as Individuals 4 to 7 (Ind4-7). In a previous work the Individual 4 sample (Ind4) was used to quantify pluripotency using LandSCENT [1], so for comparative purposes we will describe our analysis in detail for this donor.

Data analysis workflow

We performed the Seurat pipeline for scRNA-seq data analysis. We downloaded the UMI count matrix, cells and features were filtered to reduce noise and eliminate redundancies. Data was normalized and dimensionality reduction was performed (Fig. 3). Clustering and differential expression analysis allowed us to annotate the cells (Figs. S4A and B). We identified three main clusters and concluded that these corresponded to a basal myoepithelial cell type, a luminal immature secretory and immune-related cell type, and a luminal mature hormone-responsive cell type. The identified cell types are in accordance with the original work where the data was published [20] and other subsequent works [1]. In line with the original notation, we will refer to them as Basal, Luminal1 (L1) and Luminal2 (L2). The workflow used is described in detail below.

Filtering. We trimmed cells based on the total number of mapped reads, the number of unique features detected and proportion of mitochondrial content. In this sense, we filtered cells that have

unique feature counts over 3000 or less than 200 and total number of mapped reads less than 12000. Upper and lower bounds were applied to filter potential doublets or multiplets and debris or empty droplets, respectively. Besides, we removed cells that had a high mitochondrial content to eliminate potentially low-quality cells, an upper bound of 5% was chosen as it was done by Chen et al. [1]. Thresholds were selected so that after filtering, the main body of cells was kept while removing outliers. We also filtered out features expressed in few cells, the cutoff value of 3 cells per feature was used. Additionally, we filtered out three small non-epithelial clusters (stromal, endothelial and outliers) as was done in the work where data was published [20].

Normalization. In order to compare cells among them, it is necessary to carry out a normalization step. Gene counts are divided by the total counts for each individual cell, scaled (multiplication by a scale factor) and natural-log transformed. We used the NormalizeData() function from the R package Seurat. We later performed a step to identify features that exhibit high cell-to-cell variation.

Dimensionality reduction. We scaled the data prior to reducing the dimensionality of the data set, so that the mean expression is 0 and the variance is 1 across cells. We then performed PCA and UMAP and visualized the data (Fig. 3). Three main big clusters were revealed.

Clustering. We performed clustering using the Seurat functions FindNeighbors() and FindClusters(). In short, it includes building a KNN graph using the euclidean distance in the PCA space and applying the Louvain algorithm that optimizes the standard modularity function.

Differential expression. We carried out differential expression analysis and obtained the gene signatures of all the clusters. We executed non-parametric Wilcoxon rank sum test and a ROC test that returns the power of a classifier. In Fig. S4A we display some markers in the UMAP space and in Fig. S4B a heatmap is shown to portray the differential expression among clusters.

Cell annotation. The previous step provided a gene signature for each cell cluster. Three epithelial cell types were identified. The orange/pink cluster in Fig. 3 differentially expressed keratin coding genes such as KRT14 (Fig. S4A KRT14), KRT5 and KRT17 [20,40,41]. This cluster also displayed smooth muscle related genes, e.g., ACTA2 and MYLK and was labeled as a myoepithelial basal cluster [20,42]. The two remaining cell clusters were both positive to KRT18 gene expression and identified as luminal cell types [20,40]. The blue cluster in Fig. 3 exhibited high expression levels of SLPI and LTF (Figs. S4A SLPI and LTF), which are the typical Luminal Progenitor markers [20,43]. Thus, we annotated this cluster as Luminal 1 (L1). The green cluster in Fig. 3 could be identified as a luminal mature cell type because of the differentially expressed ANKRD30A gene (Fig. S4A ANKRD30A) [20, 44]. Another gene marker highly expressed was AREG, a central factor in estrogen action and ductal development of the mammary glands [44,45]. AGR2, which is a hormone responsive gene [20,46], was over-expressed too. Overall, this cluster was labeled as Luminal 2 (L2) and is associated with a hormone responsive function.

Pluripotency quantification using ORIGINS

Differentiation PPI network activity was computed over the gene expression matrix. The normalized expression matrix used to calculate the activity was not the one provided by Seurat because this procedure returns a matrix that has null elements. Since our goal is to compare our algorithm performance with other methods, and LandSCENT does not accept expression matrices with null elements as input, we applied a different normalization. We followed the normalization on the LandSCENT tutorial which sets an offset value of 1.1 before log-transformation to avoid having zero values.

Differentiation activity can be visualized in the PCA and UMAP spaces in Figs. 4A and 1A. The highest levels of activity were found within the basal group. This is in agreement with the results of the work in which the data was originally published, as the authors found a group of basal cells with stemness capacity [20]. Similarly, in the LandSCENT publication, the authors found a higher percentage of multipotent cells within the basal cluster, although they also observed high levels of SR in luminal cells in close proximity to the basal cluster [1].

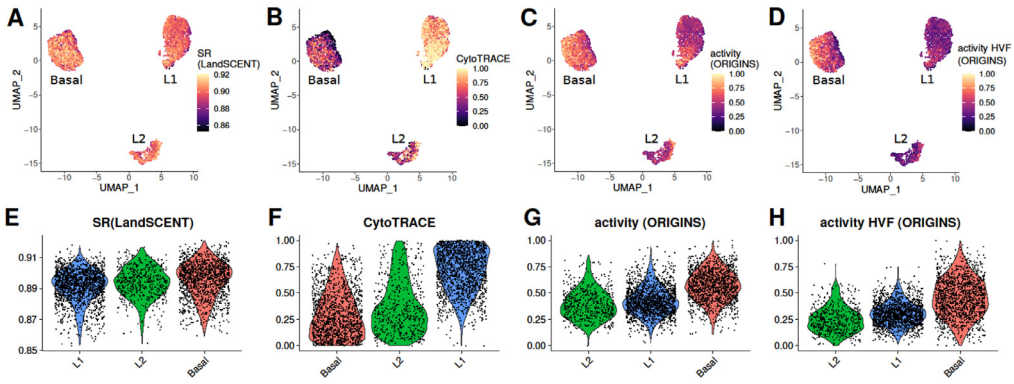

Fig. 1. A–D: UMAP representation of the breast sample colored by the scores calculated by LandSCENT, CytoTRACE, ORIGINS using all genes and the top highly variable features (HVF). **E–H**: Violin plots of the pluripotency scores per cell type sorted according to increasing values of the mean scores.

Fig. 2. Correlation matrices between all applied methodologies, SR (LandSCENT), CytoTRACE, activity (ORIGINS) and activity HVF (ORIGINS) for all data sets.

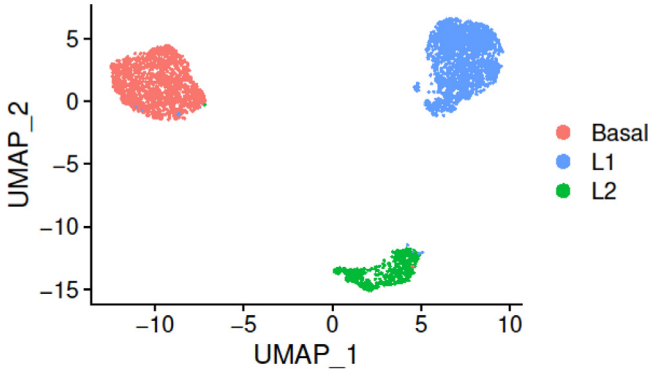

Fig. 3. Breast sc-RNAseq data in the UMAP space colored by cell type.

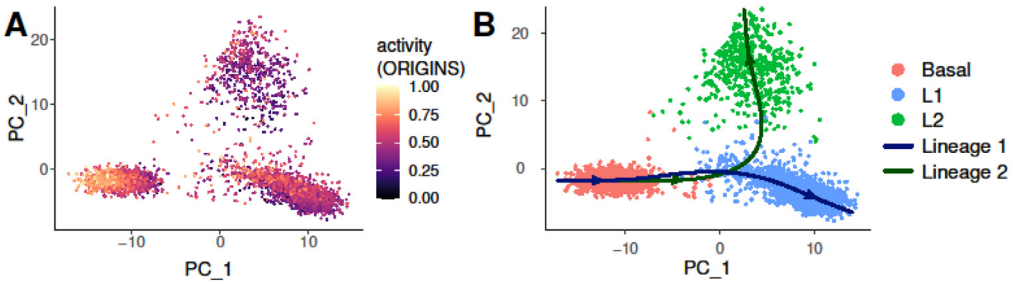

Fig. 4. **A:** Breast sc-RNAseq data representation in the PCA space. . **B:** Breast sc-RNAseq data representation in the PCA space colored by cell types. Performing trajectory inference revealed a bifurcating pathway. The first branch, Lineage 1, leads to the L1 cell type. The second branch, Lineage 2, ends at the L2 cell cluster.

Trajectory inference

Trajectory inference are single cell transcriptomics techniques that infer lineages structures. Essentially, these computational methods order cells based on their expression similarities along a "temporal" variable called pseudotime. In agreement with Saelens et al. [4] we consider that, among the large number of tools available, Slingshot [47] is one of the simplest, most robust and best documented. For this reason, we run Slingshot, an R package, on the breast data set. Slingshot works in two steps. First, it infers the global lineage structure using a cluster-based minimum spanning tree. In a second instance, it infers pseudotime for each lineage using simultaneous principal curves. Like many methods, Slingshot requires the prior definition of the trajectory origin, namely, the stem or progenitor cells.

We set the root as the cells with the highest differentiation PPI network activity, located within the basal cluster. We discovered that the trajectory progresses from the basal starting point and bifurcates into the L1 and L2 cell types, going through an intermediate state (Fig. 4B). Furthermore, this group of luminal precursor cells located within the L1 cluster has moderately high activity levels (Fig. 4A). At this point the trajectory branches and heads towards the terminal cells L1 or L2 into two separate lineages. The results obtained by performing trajectory inference supports previous works where stem/progenitor breast cells where suggested as bi-potent cells that can originate differentiated basal and luminal cell types [1,20]. We refer to the lineage starting at the basal cluster and ending at cluster L1 as Lineage 1 and the one ending at L2 as Lineage 2.

We next performed differential expression analysis along the inferred trajectory to both lineages. Thereby, we obtained the genes whose expression changed the most along the trajectory, which can

be visualized by the heatmaps in Figs. S5A and B. For that, we used the R Bioconductor package tradeSeq [48]. Briefly, tradeSeq fits a negative binomial generalized additive model (GAM) to model the relation between gene expression and pseudotime and then tests for significant relationships between gene expression and pseudotime.

In summary, we show a complete workflow to perform trajectory inference with sc-RNAseq data. We first identified the root cells using our algorithm (ORIGINS) and then used SLINGSHOT to unveil the differentiation process of the breast epithelium. SLINGSHOT provided a bifurcated trajectory, supporting the theory that breast stem cells are bi-potent. Furthermore, our analysis allowed us to identify the genes that drive the differentiation process.

Data availability

Data analyzed in this work is publicly available on the GEO database under the accession codes: GSE113197 for breast epithelial cells, GSE125970 for colon epithelial cells, GSE117498 for hematopoietic bone marrow cells and GSE161383 for lung cells.

Code availability

ORIGINS is freely available as an open source R package from the GitHub repository: https://github.com/danielasenraoka/ORIGINS.

Funding

This research was supported by CONICET, Argentina (grant number PIP 1748).

Declaration of Competing Interest

The authors decSlare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The code is available on https://github.com/danielasenraoka/ORIGINS. The data used is publicly available under the accession codes specified in the manuscript.

CRediT authorship contribution statement

Daniela Senra: Conceptualization, Methodology, Software, Validation, Data curation, Writing – original draft, Visualization. **Nara Guisoni:** Conceptualization, Methodology, Writing – review & editing, Visualization. **Luis Diambra:** Project administration, Supervision, Conceptualization, Methodology, Writing – original draft, Visualization.

Acknowledgments

We thank the community of Women in Bioinformatics and Data Science (WBDS - Latin America) for the invitation to contribute to this Special Issue. NG and LD are researchers and DS is a PhD fellow at CONICET, Argentina.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.mex.2022.101778.

References

[1] W. Chen, S.J. Morabito, K. Kessenbrock, T. Enver, K.B. Meyer, A.E. Teschendorff, Single-cell landscape in mammary epithelium reveals bipotent-like cells associated with breast cancer risk and outcome, Commun. Biol. 2 (1) (2019) 1–13.
[2] G.S. Gulati, et al., Single-cell transcriptional diversity is a hallmark of developmental potential, Science 367 (6476) (2020) 405–411.
[3] X. Zhang, K. Powell, L. Li, Breast cancer stem cells: Biomarkers, identification and isolation methods, regulating mechanisms, cellular origin, and beyond, Cancers 12 (12) (2020) 3765.
[4] W. Saelens, R. Cannoodt, H. Todorov, Y. Saeys, A comparison of single-cell trajectory inference methods, Nat. Biotechnol. 37 (5) (2019) 547–554.
[5] M. Setty, V. Kiseliovas, J. Levine, A. Gayoso, L. Mazutis, D. Pe'er, Characterization of cell fate probabilities in single-cell data with palantir, Nat. Biotechnol. 37 (4) (2019) 451–460.
[6] H. Chen, et al., Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM, Nat. Commun. 10 (1) (2019) 1–14.
[7] S.V. Stassen, G.G. Yip, K.K. Wong, J.W. Ho, K.K. Tsia, Generalized and scalable trajectory inference in single-cell omics data with VIA, Nat. Commun. 12 (1) (2021) 1–18.
[8] F.A. Wolf, et al., PAGA: Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells, Genome Biol. 20 (1) (2019) 1–9.
[9] J. Cao, et al., The single-cell transcriptional landscape of mammalian organogenesis, Nature 566 (7745) (2019) 496–502.
[10] C.R. Banerji, et al., Cellular network entropy as the energy potential in waddington's differentiation landscape, Sci. Rep. 3 (1) (2013) 1–7.
[11] A.E. Teschendorff, P. Sollich, R. Kuehn, Signalling entropy: A novel network-theoretical framework for systems analysis and interpretation of functional omic data, Methods 67 (3) (2014) 282–293.
[12] A.E. Teschendorff, T. Enver, Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome, Nat. Commun. 8 (1) (2017) 1–15.
[13] D. Grün, et al., De novo prediction of stem cell identity using single-cell transcriptome data, Cell Stem Cell 19 (2) (2016) 266–277.
[14] M. Guo, E.L. Bao, M. Wagner, J.A. Whitsett, Y. Xu, SLICE: determining cell differentiation and lineage based on single cell entropy, Nucleic Acids Res. 45 (7) (2017) e54.
[15] N.P. Palmer, P.R. Schmid, B. Berger, I.S. Kohane, A gene expression profile of stem cell pluripotentiality and differentiation is conserved across diverse solid and hematopoietic cancers, Genome Biol. 13 (8) (2012) 1–13.
[16] A.L. Barabasi, Z.N. Oltvai, Network biology: Understanding the cell's functional organization, Nat. Rev. Genet. 5 (2) (2004) 101–113.
[17] T.G.O. Consortium, The gene ontology resource: 20 years and still GOing strong, Nucleic Acids Res. 47 (D1) (2018) D330–D338, doi:10.1093/nar/gky1055.
[18] D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'Donovan, R. Apweiler. QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics. 25 (22) (2009) 3045-3046.
[19] I. Rodchenkov, et al., Pathway commons 2019 update: integration, analysis and exploration of pathway data, Nucleic Acids Res. 48 (D1) (2019) D489–D497, doi:10.1093/nar/gkz946.
[20] Q.H. Nguyen, et al., Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity, Nat. Commun. 9 (1) (2018) 1–12.
[21] Y. Wang, et al., Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine, J. Exp. Med. 217 (2) (2020).
[22] D. Pellin, et al., A comprehensive single cell transcriptional landscape of human hematopoietic progenitors, Nat. Commun. 10 (1) (2019) 1–15.
[23] A. Wang, et al., Single-cell multiomic profiling of human lungs reveals cell-type-specific and age-dynamic control of SARS-CoV2 host genes, Elife 9 (2020) e62522.
[24] S. Umar, Intestinal stem cells, Curr. Gastroenterol. Rep. 12 (5) (2010) 340–348.
[25] S. Cui, P.Y. Chang, Current understanding concerning intestinal stem cells, World J. Gastroenterol. 22 (31) (2016) 7099.
[26] N. Barker, Adult intestinal stem cells: critical drivers of epithelial homeostasis and regeneration, Nat. Rev. Mol. Cell Biol. 15 (1) (2014) 19–33.
[27] A. Birbrair, P.S. Frenette, Niche heterogeneity in the bone marrow, Ann. N.Y. Acad. Sci. U. S. A. 1370 (1) (2016) 82–96.
[28] S.J. Szilvassy, The biology of hematopoietic stem cells, Arch. Med. Res. 34 (6) (2003) 446–460.
[29] M. Zhao, L. Li, Regulation of hematopoietic stem cells in the niche, Sci. China Life Sci. 58 (12) (2015) 1209–1215.
[30] C. Baumgartner, et al., An ERK-dependent feedback mechanism prevents hematopoietic stem cell exhaustion, Cell Stem Cell 22 (6) (2018) 879–892.
[31] C. Dussiau, et al., Hematopoietic differentiation is characterized by a transient peak of entropy at a single-cell level, BMC Biol. 20 (1) (2022) 1–15.
[32] K. Wiesner, J. Teles, M. Hartnor, C. Peterson, Haematopoietic stem cells: entropic landscapes of differentiation, Interface Focus 8 (6) (2018) 20180040.
[33] A. Ciechanowicz, Stem cells in lungs, Stem Cells (2019) 261–274.
[34] S.D. Reynolds, A.M. Malkinson, Clara cell: progenitor for the bronchiolar epithelium, Int. J. Biochem. Cell Biol. 42 (1) (2010) 1–4.
[35] E.L. Rawlins, et al., The role of Scgb1a1+ clara cells in the long-term maintenance and repair of lung airway, but not alveolar, epithelium, Cell Stem Cell 4 (6) (2009) 525–534.
[36] R.G. Crystal, Airway basal cells. The 'smoking gun' of chronic obstructive pulmonary disease, Am. J. Respir. Crit. Care Med. 190 (12) (2014) 1355–1362.
[37] K.U. Hong, S.D. Reynolds, S. Watkins, E. Fuchs, B.R. Stripp, In vivo differentiation potential of tracheal basal cells: evidence for multipotent and unipotent subpopulations, Am. J. Physiology Lung Cell. Mole. Physiol. 286 (4) (2004) L643–L649.

[38] K.U. Hong, S.D. Reynolds, S. Watkins, E. Fuchs, B.R. Stripp, Basal cells are a multipotent progenitor capable of renewing the bronchial epithelium, Am. J. Pathol. 164 (2) (2004) 577–588.
[39] D.J. Weiss, et al., Stem cells and cell therapies in lung biology and lung diseases, Proc. Am. Thorac. Soc. 8 (3) (2011) 223–272.
[40] E. Lim, et al., Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways, Breast Cancer Res. 12 (2) (2010) 1–14.
[41] R. Villadsen, et al., Evidence for a stem cell hierarchy in the adult human breast, J. Cell Biol. 177 (1) (2007) 87–101.
[42] B. Pal, et al., A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast, EMBO J. 40 (11) (2021) e107333.
[43] P. Bhat-Nakshatri, et al., A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells, Cell Reports Medicine 2 (3) (2021) 100219.
[44] K. Saeki, et al., Mammary cell gene expression atlas links epithelial cell remodeling events to breast carcinogenesis, Commun. Biol. 4 (1) (2021) 1–16.
[45] C. Mukhopadhyay, X. Zhao, D. Maroni, V. Band, M. Naramura, Distinct effects of EGFR ligands on human mammary epithelial cell differentiation, PLoS One 8 (10) (2013) e75907.
[46] M.L. Salmans, F. Zhao, B. Andersen, The estrogen-regulated anterior gradient 2 (AGR2) protein in breast cancer: a potential drug target and biomarker, Breast Cancer Res. 15 (2) (2013) 1–14.
[47] K. Street, et al., “Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics,”, BMC Genom. 19 (1) (2018) 1–16.
[48] K. Van den Berge, et al., Trajectory-based differential expression analysis for single-cell sequencing data, Nat. Commun. 11 (1) (2020) 1–13.