



Cell annotation using scRNA-seq data: A protein-protein interaction network approach



Daniela Senra, Nara Guisoni, Luis Diambra*

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

ARTICLE INFO

Method name:

ORIGINS2

Keywords:

scRNA-seq
Protein-protein interaction networks
Cell annotation
Biological Processes
Breast cancer

ABSTRACT

Pathway analysis is an important step in the interpretation of single cell transcriptomic data, as it provides powerful information to detect which cellular processes are active in each individual cell. We have recently developed a protein-protein interaction network-based framework to quantify pluripotency associated pathways from scRNA-seq data. On this occasion, we extend this approach to quantify the activity of a pathway associated with any biological process, or even any list of genes. A systems-level characterization of pathway activities across multiple cell types provides a broadly applicable tool for the analysis of pathways in both healthy and disease conditions. Dysregulated cellular functions are a hallmark of a wide spectrum of human disorders, including cancer and autoimmune diseases. Here, we illustrate our method by analyzing various biological processes in healthy and cancer breast samples. Using this approach we found that tumor breast cells, even when they form a single group in the UMAP space, keep diverse biological programs active in a differentiated manner within the cluster.

- We implement a protein-protein interaction network-based approach to quantify the activity of different biological processes.
- The methodology can be used for cell annotation in scRNA-seq studies and is freely available as R package.

Specifications table

Subject area:	Bioinformatics
More specific subject area:	Cell annotation from single cell transcriptomics
Name of your method:	ORIGINS2
Name and reference of original method:	Not applicable
Resource availability:	https://github.com/danielasenraoka/ORIGINS2

Method details

Background

Bulk RNA-sequencing (RNA-seq) is focused on quantifying the gene expression profiles across a population of cells to get insight into the biological processes at transcriptional level without taking into consideration cellular heterogeneity. On the other hand,

* Corresponding author.

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

Social media: (L. Diambra)

<https://doi.org/10.1016/j.mex.2023.102179>

Received 31 December 2022; Accepted 9 April 2023

Available online 10 April 2023

2215-0161/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

single-cell RNA sequencing (scRNA-seq) is a high-throughput technology used to study cellular heterogeneity at the transcriptional level and it has been successful to discover novel cell types in cancer and analyze lineages in the context of embryonic development [1,2].

Although scRNA-seq provides high-resolution transcriptomic data that can potentially characterize various phenotypes, the task of identifying true biological differences between cells is challenging due to several technical constraints. Most of the scRNA-seq studies consist of differential gene expression analysis performed on groups of cells, derived by unsupervised clustering methods, to discover genes (features) associated with these cell type clusters (phenotypes). The cell clusters are determined based on the similarity of transcriptional profiles. Additionally, gene set enrichment analysis, originally developed for bulk data, can be used to retrieve a functional profile of the gene sets obtained after clustering and differential expression analysis, in order to better understand the underlying biological processes [3–5].

Due to intrinsic cellular heterogeneity this commonly used approach can lead to excessive partition of a uniform cell sub-population or yield subtle phenotypic changes, without a clear boundary between different phenotypes. Further, due to the sparse and noisy nature of scRNA-seq measurement the differential gene expression analysis on a gene-by-gene basis may result in an inaccurate characterization of the cell population or yield misleading conclusions, even when using well-established tools that exist for such analysis [6,7].

In general, complex interactions between molecules including DNA, RNA, proteins, and small molecules give rise to the biological processes that underlie a specific cell phenotype [8]. Nowadays, more information is being made available on these interactions and the elements integrating different biological function modules. This makes network-based approaches dedicated to detecting gene expression signatures in single cells, resulting from a specific biological process or cell state based on a combined group of genes, feasible. Here, we extend the concept introduced in a recent approach that takes advantage of well-established knowledge, such as the gene ontology annotation associated with a given biological process and the protein-protein interaction network (PPIN) to identify the most highly phenotype-associated cell sub-populations. Originally this approach, named ORIGINS [9], aimed to quantify pluripotency and detect cells that may be the origin of a lineage in differentiation pathways. In this work we provide a generalization of the previous methodology to evaluate several biological processes other than cell differentiation. In order to illustrate the use and application of the ORIGINS extension to discover cells that carry out specific biological programs of interest, we apply our method to assess biological processes such as: stem cells, proliferation, migration, cell cycle, etc. from scRNA-seq data derived from triple negative breast cancer (TNBC) and healthy human breast samples.

In addition to the extension to assess different biological modules, we made technical adjustments to the previous method. Modifications include the removal of proteins from the PPIN that act as negative regulators of the module under study, as overexpression of negative regulators would inaccurately increase the activity of the corresponding PPIN. To further enhance the efficiency of our method, we have modified the program to eliminate duplicated interactions in the PPIN, thereby reducing computing time of the algorithm. Furthermore, we provide a new functionality to build PPINs associated with any biological process or set of genes of interest. The method is implemented as an R package called **ORIGINS2**, which also includes the PPINs used in this work. The package is open-source and publicly available at <https://github.com/danielasenaoka/ORIGINS2>, where a user guide to download and use the program is included.

Data description

We used publicly available scRNA-seq data from two published studies [1,10]. From the first study we take one sample corresponding to normal breast tissue (ind4), while we used a TNBC sample from the second dataset (TN-0135). The criteria to choose this samples were based on the large number of cells assessed in these cases.

PPIN activity computation

A PPIN is a collection of nodes (proteins) interconnected by edges (interactions). The edges convey information about the links between the nodes (weights) and in this case edges have no direction, that is, the PPIN is an undirected graph. All the PPINs used in this paper were built by considering the whole biochemical interactions listed in Pathways Commons (version 12), that integrate 2424,055 interactions from 22 databases [11]. From this set of interactions we disregard the interactions involving not-protein molecules, such as chemical compounds. From this whole network we select the protein-protein interactions associated with the gene products involved in different biological process (BP) without considering the genes acting as negative regulators of the BP and build the PPIN associated with this BP. The protein lists from *H. sapiens* were downloaded from the QuickGo database [12] for each biological process. The edges of this PPIN correspond to biochemical reactions and contribute to the activity of the module P , following mass-action law, proportionally to their amount [9]:

$$P = \sum_{i,j=1}^{N_p} A_{ij} x_i x_j, \quad (1)$$

where A_{ij} is the upper triangular adjacency matrix (i.e., the element A_{ij} is one when proteins i and j interact and zero otherwise), x_i and x_j are the expression levels associated with genes I and J , respectively; and N_p is the number of genes in the pathway p . In contrast to the previous algorithm, here we use the upper triangular adjacency matrix as we consider only one interaction between each pair of proteins instead of the complete adjacency matrix. Since the adjacency matrices are symmetric, that is, A interacts with

B and B with A, we decided to eliminate this duplication of edges, only considering one for each pair of proteins. In this way, the computation time is reduced, obtaining the same result.

Moreover, in the original methodology, we considered all the proteins involved in cell differentiation, which was the biological process for which the previous method was developed. Here we decided to remove all the proteins that are negative regulators of the module, as they would contribute to activity inaccurately. The negative regulators that were discarded were obtained by selecting negative regulation of the corresponding process in the QuickGo database [12]. Some negative regulators act as positive regulators simultaneously, therefore only the proteins that exclusively act as negative regulators were eliminated from the network.

The earliest version of the software was developed to quantify activity of the PPIN associated with cell differentiation BP, which is particularly useful to identify pluripotent cells to use as root when performing trajectory inference. The main contribution of ORIGINS2 lies in its capability to assess activity of the PPIN associated with any biological process or set of genes. The inclusion of pre-built PPINs within the software package that are associated with a diverse range of biological processes is a major novelty of ORIGINS2, and serves to greatly expand the scope of analyses that can be performed using the software. In this regard we supply the PPINs associated with several biological processes, including but not limited to: cell differentiation (GO:0030154), cell cycle (GO:0007049), mitotic cell cycle (GO:1903047), cell population proliferation (GO:0008283), DNA repair (GO:0006281), DNA replication (GO:0006260), immune response (GO:0006955), inflammatory response (GO:0006954), cell migration (GO:0016477), stem cell proliferation (GO:0072089), stem cell differentiation (GO:0048863), hormone secretion (GO:0046879), humoral immune response (GO:0006959), acute inflammatory response (GO:0002526), chronic inflammatory response (GO:0002544), response to tumor cell (GO:0002347), innate immune response (GO:0045087) and adaptive immune response (GO:0002250).

Furthermore, we provide a new feature that builds a PPIN associated with any module of interest. To create a custom-made PPIN the *build_ppin* R function should be used, the inputs are the gene ID list associated with the corresponding biological process and the full human PPIN, included within the package. Essentially, this function operates by overlaying the gene list onto the full human PPIN, subsetting the protein-protein interactions (edges) from the full human PPIN whose nodes are included in the gene list provided by the user. The output is an adjacency list that contains a subset of the full human PPIN that is specific to the biological process under study. This feature enables the user to determine the activity of the PPIN associated with any desired biological process. To accomplish this, the previously created PPIN and the expression matrix must be entered as inputs into the *activity* function. For further information users may refer to the detailed documentation provided by the R package.

Method validation

Application to healthy breast tissue

In our previous version we do not distinguish if any node is acting as a negative or positive regulator. However, accounting for the expression of negative regulators in Eq.(1) could lead to an overestimation of the actual activity P of a given biological module. To circumvent this drawback, the new approach does not take into account any of the genes that act as negative regulators of the BP in question. To illustrate the effect of this improvement, we compare the performance of both computations for a healthy breast sample [10] by considering two BPs: cell differentiation (Fig.1) and cell population proliferation (Fig. 2). These UMAP representations contain 3 clusters: luminal 1 (L1), luminal 2 (L2) and basal cells, as we determined previously [9]. Activity visualization on the UMAP space, with and without negative regulators, yielded similar relative results on the color scale. Although, as expected, lower absolute values are observed when negative regulators are not considered, since the new form of calculation involves a network of fewer genes. However, when comparing the distributions of the activities associated with cell differentiation and cell proliferation in each population, we observe that there are differences between them, that may or may not be significant depending on the significance level given, depending on whether or not PPIN considers negative regulators. For example, Fig. 2 shows that the cell proliferation-associated activity distributions of basal and L2 cells are significantly different in the overestimated calculation, whereas when the activity is calculated without adding the negative regulators are not (at the 0.01 significance level).

By considering this modification in the PPIN activity calculation, we also compute the activities associated with several other biological processes over this healthy breast sample. Fig. 3A depicts the UMAP representation of a healthy breast sample colored by the activity associated with acute inflammatory response. In this case, the highest activity levels are concentrated in the upper region of the L1 cells in the UMAP space. There are also few L2 cells showing acute inflammatory response to the left of the L2 cluster, also visible in the distribution (green plot in Fig. 3B). From this example, it is evident that even from a clearly delimited cluster of cells, the activity level of some biological module can be heterogeneously distributed among the population of cells; and that the method proposed here is useful to identify intra-cluster cells that perform a particular BP. Fig. 3C depicts Pearson correlation coefficients between several BP computed over all cells in this data set. The BPs analyzed can be mainly clustered into two groups: one including inflammatory and immune responses and the other including cell replication and differentiation. These groups of BPs appear to be anti-correlated in this healthy breast sample.

Application to triple negative breast cancer

Healthy samples exhibit mixed cell types, which can be identified based on differences in their transcriptomic profiles. The situation is rather more complex in cancer research, where characteristic tumor signatures dilute intra-tumoral heterogeneity of tumor cells. To overcome this drawback, our technique can be used to identify biological modules operating in different regions of the cluster corresponding to tumor cells. Fig. 4 illustrates the UMAP representations of the activities associated with nine BPs

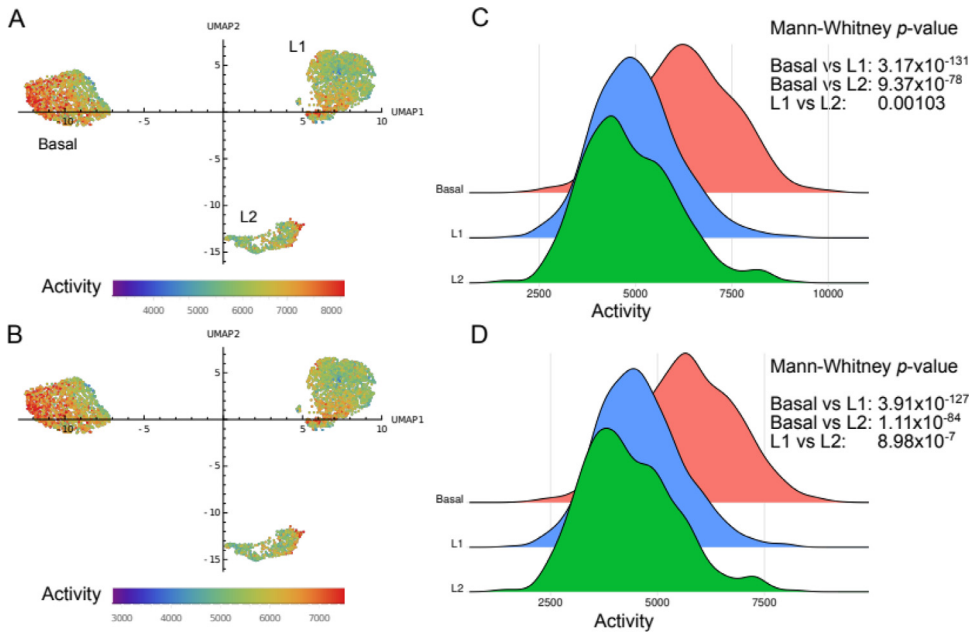


Fig. 1. A: UMAP representation of the healthy breast sample colored by the activity of the PPIN associated with all genes involved in the cell differentiation BP (GO:0030154). B: Activity computed by excluding the genes that negatively regulate the process above. C: Probability density estimation of cell differentiation activity per cell type computed with all genes, as panel A. D: Probability density estimation of cell differentiation activity computed excluding genes that negatively regulate the process, as panel B. Results from the Mann-Whitney test among different distribution are displayed at right. Luminal 1 cells (L1), luminal 2 cells (L2).

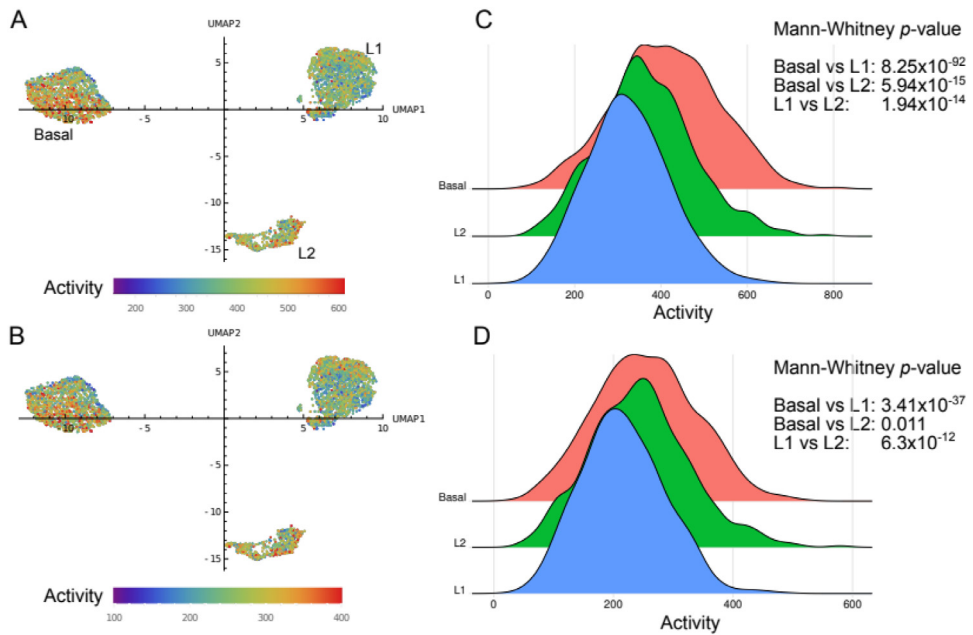


Fig. 2. A: UMAP representation of the healthy breast sample colored by the activity associated with all genes involved in the cell proliferation BP (GO:0008283). B: Activity computed by excluding the genes that negatively regulate the process above. C: Probability density estimation of cell proliferation activity per cell type computed with all genes, as panel A. D: Probability density estimation of cell proliferation activity computed excluding genes that negatively regulate the process, as panel B. Results from the Mann-Whitney test among different distributions are displayed at right. Luminal 1 cells (L1), luminal 2 cells (L2).

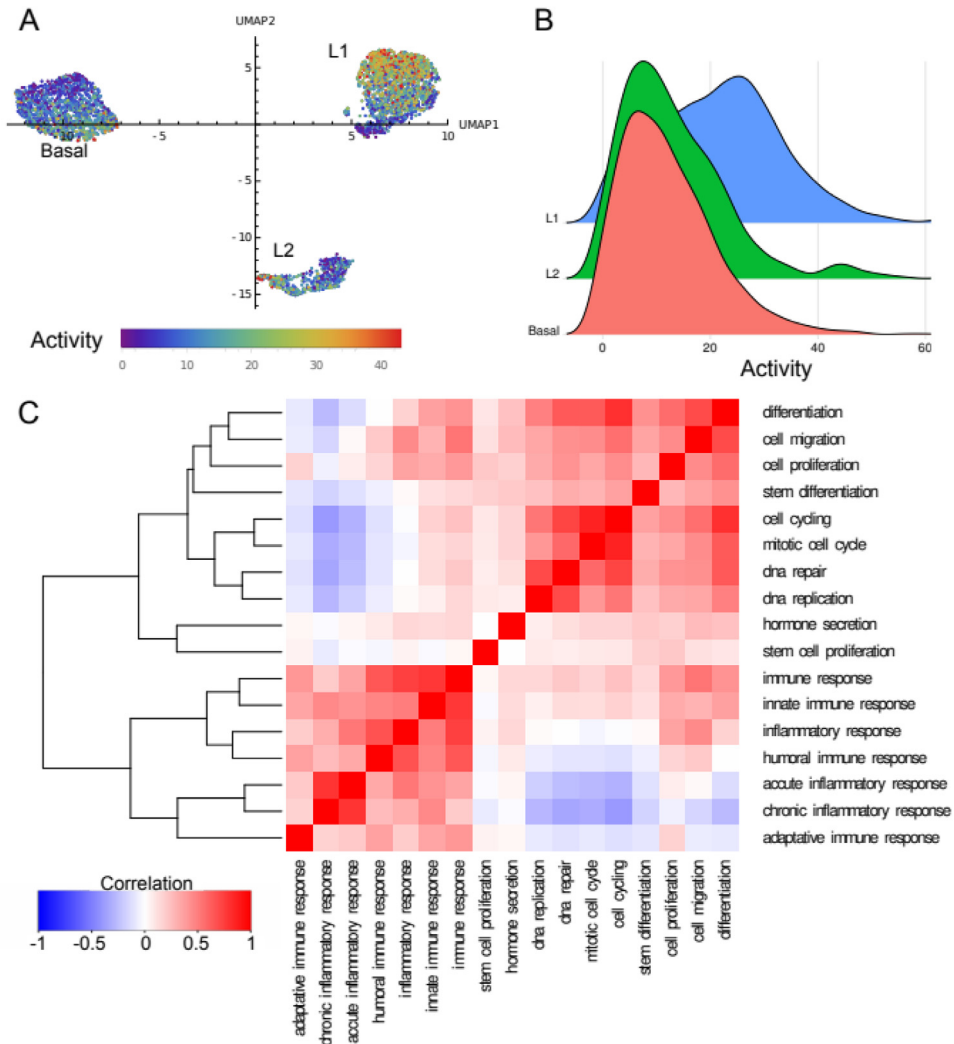


Fig. 3. A: UMAP representation of the healthy breast sample colored by the activity associated with acute inflammatory response BP (GO:0002526) computed by excluding the genes that negatively regulate this process. B: Probability density estimation of acute inflammatory response activity per cell type. C: Correlations among activities associated with different BP computed over the same healthy breast sample. Luminal 1 cells (L1), luminal 2 cells (L2).

of interest in a tumoral breast sample. These UMAP representations contain four cell clusters: tumor cells (TC), fibroblasts (CAF), macrophages (TAM) and T-cells. It can be observed that the tumor cells cluster presents high heterogeneity, i.e., different regions in the UMAP space where biological modules are operating. For example, while cell differentiation activity appears in a vast number of TC, the cell migration process is active in a small number of cells. On the other hand, mitotic cell cycling, DNA replication and DNA repair processes appear quite correlated, as expected, and highly activated in the lower-left region of the tumor cells. This is in agreement with the observations made by [1], where authors found out that all the breast cancer subtypes studied in their work [1], in particular TNBC, always showcase a population of cycling TC. Further, there is a small group of cells which present an important stem cell differentiation activity within the TC cluster, as well as fibroblasts. Finally, while the innate immune response affects TC and macrophages, the adaptive immune response activity is observed almost exclusively in macrophages.

Certainly, the activity levels of many of these processes can be correlated, as shown previously for the healthy sample in Fig. 3C. Fig. 5 depicts Pearson correlation coefficient between several BPs computed over all cells in this cancer breast sample. In contrast to the healthy sample, most BPs appear to be correlated or poorly correlated but not anti-correlated.

The method proposed here is particularly useful for efficiently translating scRNA-seq transcriptomes into biological discoveries associated with a functional network of genes of interest. The ultimate goal is to examine whether the most significant biological features can be further elucidated by an alternative approach.

The typical pipeline to identify subpopulations from single-cell transcriptomic data involves several serial steps. First, unsupervised clustering is performed to define cell clusters. Then, differential expression analysis is carried out to obtain the marker genes for each

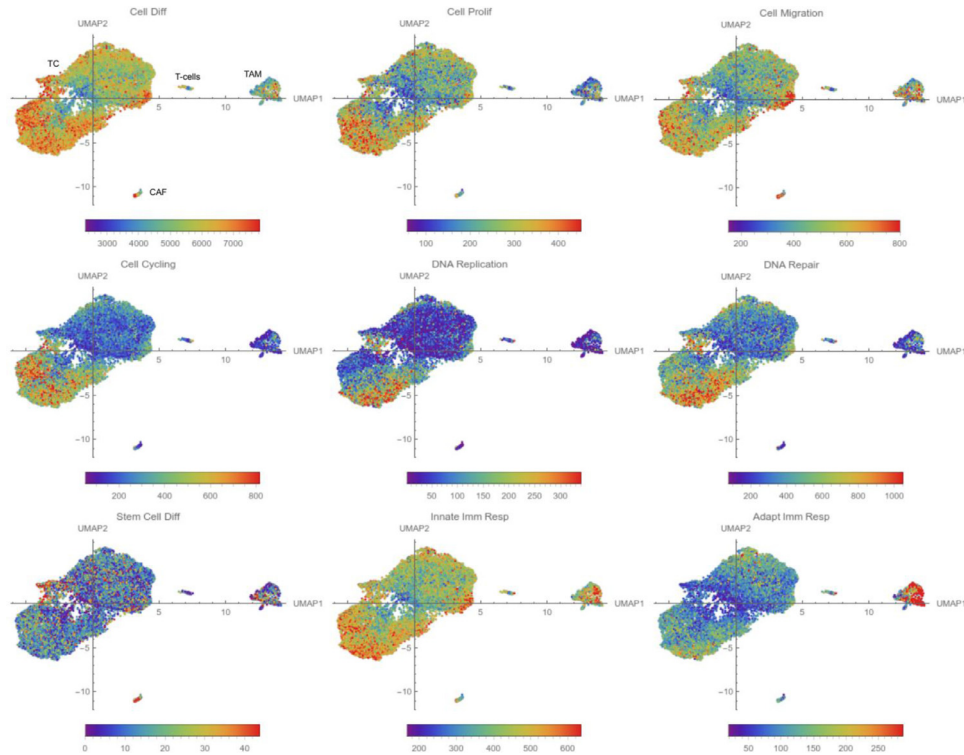


Fig. 4. UMAP representation of the breast cancer sample colored by the activity associated with nine BP of interest: cell differentiation (GO:0030154), cell population proliferation (GO:0008283), cell migration (GO:0016477), mitotic cell cycle (GO:1903047), DNA replication (GO:0006260), DNA repair (GO:0006281), stem cell differentiation (GO:0048863), innate immune response (GO:0045087) and adaptive immune response (GO:0002250). TC: tumor cell, TAM: tumor-associated macrophage, CAF: cancer-associated fibroblast.

cluster. Finally, the enrichment of these markers in known cell types and pathways is assessed to gain insight into the biological significance of each cell cluster. However, this traditional approach is highly dependent on the preceding steps, specially on clustering, as the technique identifies the active pathways in each cell cluster. In this sense, this widely used methodology does not provide information about the heterogeneity in different biological pathways at single cell level, which might be critical in samples containing rare cell types. In this way, ORIGINS2 can quantify how active the PPIN associated with a certain biological process is for each cell, without assuming the number of cell types or cell states within a sample.

Comparison with other methods

In order to compare ORIGINS2 with other methodologies that perform pathway enrichment we evaluated the mean expression (Average) of the gene set associated with the biological process of interest, a common practice when scoring expression levels of a gene set [13,14]. In addition, we also assessed AUCell [15], a computational method that determines the enrichment of user-defined gene sets within each cell. To evaluate enrichment, AUCell utilizes the Area Under the Curve (AUC) approach to determine if a critical subset of the input gene set is overrepresented among the expressed genes in each cell. Specifically, we used gene sets that were associated with the biological processes studied in our investigation using ORIGINS2. By providing the scRNA-seq expression matrix and the gene sets of interest to AUCell, the software calculates a score that indicates the activity level of the pathway of interest in each individual cell. We computed the three scores using the TNBC sample used in the section *Application to triple negative breast cancer*.

Table 1 displays the Pearson correlation coefficient between ORIGINS2, AUCell and Average for all the gene sets used in Fig. 5A. All three methods are positively correlated, with mean Pearson correlation coefficients ranging from 0.31 to 0.96. The mean correlation coefficients are $\langle \rho_{ORIGINS2,AUCell} \rangle = 0.59$, $\langle \rho_{ORIGINS2,Average} \rangle = 0.79$ and $\langle \rho_{AUCell,Average} \rangle = 0.66$. The different scores for gene sets containing a small number of genes may be sensible to dropouts and less stable. Thus, we calculated the mean correlation coefficients disregarding the BP with less than 200 genes $\langle \rho_{ORIGINS2,AUCell}^{N_{genes}>200} \rangle = 0.63$, $\langle \rho_{ORIGINS2,Average}^{N_{genes}>200} \rangle = 0.86$ and $\langle \rho_{AUCell,Average}^{N_{genes}>200} \rangle = 0.66$.

In Fig. S1 we show the AUCell score (panels A-C) for the breast cancer data set compared to the ORIGINS2 activity (panels D-F) associated with three biological processes: cell differentiation, cell proliferation and cell migration. Overall, both scores are comparable although the AUCell score presents activities in a more concentrated group of cells. In this figure, it is observed that as

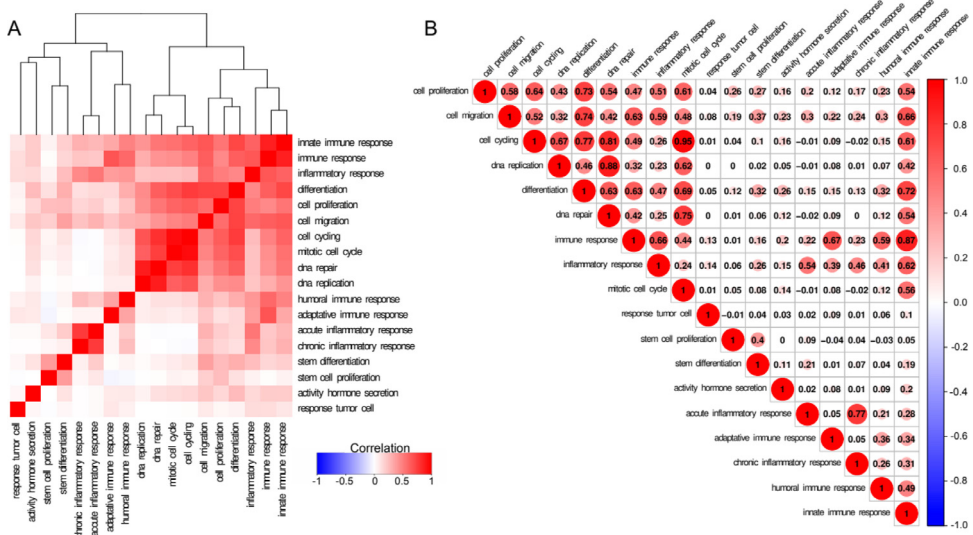


Fig. 5. A: Heat-map of correlations among activities associated with different BPs computed over the same tumoral breast sample. **B:** Alternative representation of the correlation matrix among the same activities.

Table 1

Pearson correlation coefficient between ORIGINS2, AUCell and Average gene expression and number of genes (N_{genes}) of different biological processes studied.

Gene Ontology Biological Process	$\rho_{ORIGINS2,AUCell}$	$\rho_{ORIGINS2,Average}$	$\rho_{AUCell,Average}$	N_{genes}
Cell population proliferation	0.53	0.84	0.49	656
Cell differentiation	0.39	0.94	0.35	3704
Mitotic cell cycle	0.68	0.93	0.70	494
Cell migration	0.60	0.83	0.62	914
Immune response	0.66	0.89	0.63	2414
Stem cell differentiation	0.46	0.54	0.50	179
Cell cycle	0.68	0.96	0.67	1367
DNA replication	0.69	0.90	0.73	332
DNA repair	0.62	0.94	0.64	824
Inflammatory response	0.75	0.84	0.75	581
Stem cell proliferation	0.44	0.52	0.61	64
Hormone secretion	0.31	0.55	0.61	100
Acute inflammatory response	0.65	0.66	0.76	103
Chronic inflammatory response	0.70	0.77	0.90	13
Adaptive immune response	0.84	0.81	0.80	867
Humoral immune response	0.54	0.62	0.88	415
Innate immune response	0.53	0.86	0.60	1005

the correlation coefficients between ORIGINS2 and AUCell increase, the UMAP representations of the different biological processes more closely resemble each other.

Although the scores were found to be correlated, it is important to acknowledge that the underlying methodologies for the parameter calculation strongly differ. AUCell and the average method exclusively use gene expression information (expression matrix) to provide a parameter that is high when the set of genes of interest is overrepresented and vice versa. In the case of the algorithm proposed in this study, in addition to inputting the expression matrix, protein-level information is utilized by providing PPINs. Thus, the parameter calculated with ORIGINS2 reflects the level of interaction between genes that encode proteins that are reported to interact with each other.

Furthermore, it is worth noting that ORIGINS2, together with the other evaluated methods in this work, exhibits reduced robustness in scenarios where the set of genes of interest is small. This phenomenon is attributable to the technical limitations inherent to scRNA-seq experiments, which are characterized by high dropout rates.

Our findings demonstrate that this approach is useful in identifying pluripotent cells [9] and annotating cells. To expand the applicability of the tool, several approaches can be pursued in the future, such as extending the PPINs to a greater number of biological processes and calculating activities related to additional biological processes that may be of interest. In addition, ORIGINS2 could be used to map scRNA-seq data onto a space of activities associated with relevant biological processes for each particular dataset. On this space of activities, standard techniques for scRNA-seq, including dimensionality reduction, visualization and trajectory inference, could be performed, among other promising avenues of investigation.

Code availability

ORIGINS2 is freely available as an open source R package from the GitHub repository: <https://github.com/danielasenraoka/ORIGINS2>.

Data availability

Data analyzed in this work is publicly available on the GEO database under the accession codes: GSE113197 for breast epithelial cells, and GSE161529 for the TNBC sample.

CRedit authorship contribution statement

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

Acknowledgments

NG and LD are researchers and DS is a PhD fellow at CONICET, Argentina.

Funding: This research was supported by grants: [PIP 1748 \(CONICET\)](#) and [PICT 2018–03713 \(FONCYT\)](#).

Declaration of interests

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

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:[10.1016/j.mex.2023.102179](https://doi.org/10.1016/j.mex.2023.102179).

References

- [1] 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.
- [2] Q. Yu, et al., Charting human development using a multi-endodermal organ atlas and organoid models, *Cell* 184 (2021) 3281–3298.e22.
- [3] B. Hwang, J.H. Lee, D. Bang, Single-cell RNA sequencing technologies and bioinformatics pipelines, *Exp. Mol. Med.* 50 (8) (2018) 1–14.
- [4] M.D. Luecken, F.J. Theis, Current best practices in single-cell RNA-seq analysis: a tutorial, *Mol. Syst. Biol.* 15 (6) (2019) e8746.
- [5] T. Stuart, et al., Comprehensive integration of single-cell data, *Cell* 177 (7) (2019) 1888–1902.
- [6] M.K. Jaakkola, F. Seyednasrollah, A. Mehmood, L.L. Elo, Comparison of methods to detect differentially expressed genes between single-cell populations, *Brief. Bioinformatics* 18 (5) (2017) 735–743.
- [7] L. Diambra, A.M. Alonso, S. Sookoian, C.J. Pirola, Single cell gene expression profiling of nasal ciliated cells reveals distinctive biological processes related to epigenetic mechanisms in patients with severe COVID-19, *Comput. Biol. Med.* 148 (June) (2022) 105895, doi:[10.1016/j.combiomed.2022.105895](https://doi.org/10.1016/j.combiomed.2022.105895).
- [8] A.-L. Barabasi, Z.N. Oltvai, Network biology: understanding the cell's functional organization, *Nature Rev. Genetics* 5 (2) (2004) 101–113.
- [9] D. Senra, N. Guisoni, L. Diambra, ORIGINS: a protein network-based approach to quantify cell pluripotency from scRNA-seq data, *MethodsX* 9 (2022) 101778.
- [10] 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.
- [11] I. Rodchenkov, et al., Pathway commons 2019 update: integration, analysis and exploration of pathway data, *Nucleic Acids Res.* 48 (D1) (2020) D489–D497.
- [12] 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.
- [13] 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.
- [14] I. Tirosh, et al., Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq, *Science* 352 (6282) (2016) 189–196.
- [15] S. Aibar, et al., SCENIC: single-cell regulatory network inference and clustering, *Nat. Methods* 14 (11) (2017) 1083–1086.