(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
23 January 2025 (23.01.2025)

WIPO | PCT

(10) International Publication Number
WO 2025/017162 A1

(51) International Patent Classification:
*G16B 5/00* (2019.01) *G16B 25/10* (2019.01)

(21) International Application Number:
PCT/EP2024/070475

(22) International Filing Date:
18 July 2024 (18.07.2024)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
18/355,589 20 July 2023 (20.07.2023) US

(71) Applicants: **ONE BIOSCIENCES** [FR/FR]; 14 rue de l'Ancienne Comédie, 75006 PARIS (FR). **CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE** [FR/FR]; 3 rue Michel-Ange, 75016 PARIS (FR). **INSTITUT CURIE** [FR/FR]; 26 rue d'Ulm, 75005 PARIS (FR). **SORBONNE UNIVERSITE** [FR/FR]; 21 rue de l'Ecole de Médecine, 75006 PARIS (FR).

(72) Inventors: **VALLOT, Céline**; c/o One Biosciences, 14 rue de l'Ancienne Comédie, 75006 PARIS (FR). **LANDAIS, Yuna**; c/o One Biosciences, 14 rue de l'Ancienne Comédie, 75006 PARIS (FR).

(74) Agent: **ICOSA**; 83 avenue Denfert-Rochereau, 75014 Paris (FR).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH,

(54) Title: METHOD AND DEVICE FOR DETERMINING AND QUANTIFYING MODES OF ACTIVATION OF BIOLOGICAL PATHWAYS IN INDIVIDUAL CELLS

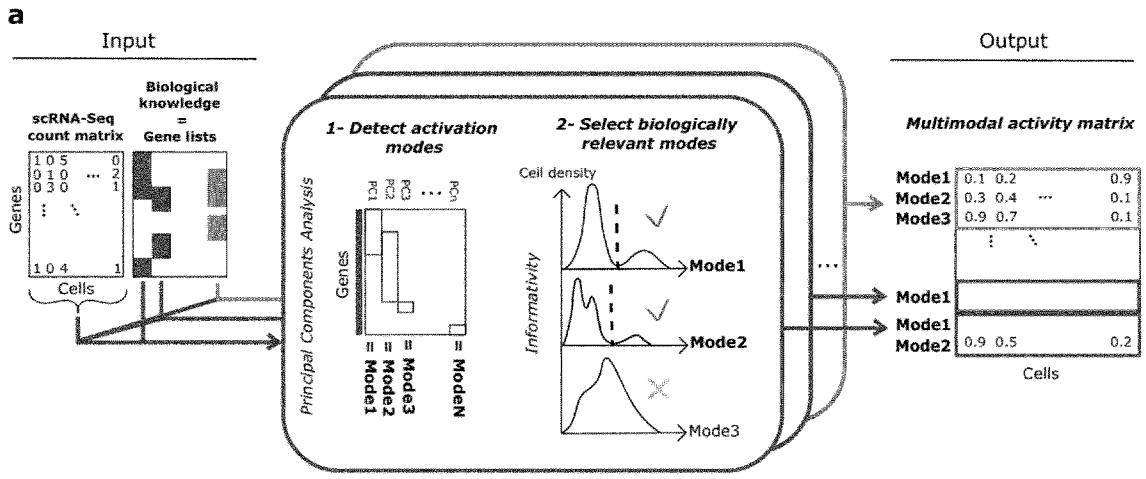



FIG. 1A

(57) Abstract: A computer-implemented method for determining and quantifying modes of activation of biological pathways in individual cells, including: receiving sequencing data obtained by a single-cell RNA sequencing method, and a plurality of gene lists, determining a gene-cell expression matrix based on the sequencing data and the plurality of gene lists, carrying out a principal component analysis (PCA) on said gene-cell expression matrix, so as to determine a plurality of modes of activation of at least one among the biological pathways, selecting a subset of so-called effective modes of activation among the plurality of modes of activation, determining a matrix referred to as activity matrix, the activity matrix including scores quantifying a level of activity of each effective mode of activation among the subset of effective modes of activation within each individual cell among the set of individual cells.

WO 2025/017162 A1

*[Continued on next page]*

TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

**Declarations under Rule 4.17:**

— *of inventorship (Rule 4.17(iv))*

**Published:**

— *with international search report (Art. 21(3))*

— *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*

# METHOD AND DEVICE FOR DETERMINING AND QUANTIFYING MODES OF ACTIVATION OF BIOLOGICAL PATHWAYS IN INDIVIDUAL CELLS

## FIELD OF INVENTION

[0001] The present invention relates to a device and a method to analyze genes in individual cells. More in details the present invention relates to a device and method for determining and quantifying modes of activation of biological pathways in individual cells.

## BACKGROUND

[0002] The identification of cell type and function is the driving force of most single-cell studies. Such approaches are based on lists of canonical marker genes and pathway databases. Standard scRNA-seq analysis pipelines involve steps of dimensionality reduction and clustering before starting any marker or pathway analysis, which makes the resulting conclusions highly dependent on the chosen algorithm and clustering parameters. In the case of cancer datasets, such clustering-based approaches appear inadequate to identify shared transcriptional programs across tumors as cancer cells tend to cluster per patient rather than group by biological similarities. Several approaches have emerged, bypassing dimensionality reduction and clustering, and proposing to score pathway activity directly for individual cells rather than for clusters. Such pooling of several gene-based measurements into scores has proven extremely powerful for the interpretation of sparse and noisy scRNA-seq datasets. A recent benchmark presented Pagoda2 (Fan, J. et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat. Methods) and AUCell (Aibar, S. et al. SCENIC: Single-cell regulatory network inference and clustering. Nat. Methods 14, 1083–1086) as two of the top performing tools for pathway activity scoring. They are based on different scoring methods—AUCell estimates the proportion of highly expressed genes in each pathway while Pagoda2 uses the weights of the first principal component from Principal

Component Analysis (PCA)—and each proposes a way to ubique. Nonetheless, both tools compute a unique activity score per pathway for all cells, implying that genes of a given signaling pathway should have coordinated expression across cell types.

Biological evaluation of pathway activation and more recently single-cell studies have repeatedly demonstrated the heterogeneity of cell functions depending on the biological context. Yet most single-cell studies analyze pathway activation with single scores based on gene lists extracted from bulk data. Such curated gene lists represent the current reference biological knowledge, that the community uses to make biological sense of sparse and noisy scRNA-seq data. Adding more specialized curated gene lists to databases—detailing cellular functions according to cell identity—is ongoing but it will take some time to be completed.

## SUMMARY

[0003] This invention thus relates to a computer-implemented method for determining and quantifying modes of activation of biological pathways in individual cells, comprising:

- receiving a) sequencing data obtained by a single-cell RNA sequencing method, said sequencing data being characterized by a set of genes and a set of individual cells, and b) a plurality of gene lists, each gene list corresponding to a biological pathway or a gene signature of interest,
- determining a gene-cell expression matrix based on said sequencing data and said plurality of gene lists, said gene-expression matrix comprising rows corresponding to genes of said set of genes referred to as pathway genes and belonging to at least one among said gene lists,
- carrying out a principal component analysis (PCA) on said gene-cell expression matrix, so as to determine a plurality of modes of activation of at least one among said biological pathways, each mode of activation corresponding to one among the principal components obtained from the PCA,
- selecting a subset of so-called effective modes of activation among said plurality of modes of activation,
- determining a matrix referred to as activity matrix, said activity matrix comprising

scores quantifying a level of activity of each effective mode of activation among said subset of effective modes of activation within each individual cell among said set of individual cells.

[0004] Advantageously, the method of the present invention allows to sort out the different modes of pathway activation specific to each cell type, by automatically detecting gene subgroups within reference pathways and computing several scores of pathway activation. In addition to pathway analysis, the present method also performs assisted cell typing as a side function, making it an all-in-one tool for both cell type and cell function identification. The present invention proves particularly useful for single-cell cancer datasets, by (i) identifying common modes of pathway activations across patients in tumor cells, and also by (ii) dissecting the contribution of each population – fibroblast, immune & tumor cell – to the activation of a given pathway. Advantageously, the method of the present invention is insensitive to batch effect, enabling the production of embeddings free of any batch effect arising from technological or experimental bias.

[0005] Some examples of gene lists that may be used in the present invention may be chosen among: KEGG database https://www.genome.jp/kegg/pathway.html, MSigDB database https://www.gsea-msigdb.org/gsea/msigdb, or any custom gene list.

[0006] According to other advantageous aspects of the invention, the device comprises one or more of the features described in the following embodiments, taken alone or in any possible combination.

[0007] According to one embodiment, for each cell, said score is calculated as the coordinate of the cell on the principal component axis related to the mode of activation of the pathway.

[0008] According to one embodiment, the method uses an initial cell type annotation by users as a reference, which advantageously enables an assisted and accurate annotation of cell identity in single-cell datasets.

[0009] According to one embodiment, each mode of activation among said plurality of modes of activation is defined by a specific subset of genes among said pathway genes.

[0010] According to one embodiment, the step of selecting said subset of modes of activation comprises: detecting bimodal distributions of scores in said activity matrix and verifying that the number of active cells is superior to a predefined minimum fraction of the set of individual cells.

[0011] According to one embodiment, the step of selecting said subset of modes of activation further comprises selecting modes of activation for which said specific subset of genes comprises a number of genes higher than a predefined threshold.

[0012] According to one embodiment, said biological pathways comprise at least one hallmark pathway.

[0013] According to one embodiment, the gene dataset is a curated gene dataset.

[0014] According to one embodiment, the method allows to detect subgroups of genes within biological pathways.

[0015] According to one embodiment, the method allows to identify common modes of activation of biological pathways in tumor cells shared across patients.

[0016] According to one embodiment, the method allows advantageously to identify common modes of activation of biological pathways shared across tumor cells.

[0017] According to one embodiment, the method allows advantageously to identify modes of activation of biological pathways specific of cell types.

[0018] According to one embodiment, the method allows advantageously to identify modes of activation of biological pathways specific to the tumor microenvironment.

[0019] According to one embodiment, the method allows advantageously to identify cell types.

[0020] According to one embodiment, the method allows advantageously to identify rare cell types.

[0021] According to one embodiment, the method allows advantageously to identify tumor cells.

[0022] According to one embodiment, the method allows advantageously to identify cell function.

[0023] According to one embodiment, the method allows advantageously to identify therapeutic targets.

[0024] According to one embodiment, the method allows advantageously to detect relevant biological signal from noisy gene list.

[0025] According to one embodiment, the method allows advantageously to detect relevant biological signal independently of batch effect.

[0026] The present invention further relates to a device for obtaining and quantifying modes of activation of biological pathways in individual cells to determine therapeutic targets for cancer therapy, comprising:

- at least one input interface configured to receive a) sequencing data obtained by a single-cell RNA sequencing method, said sequencing data being characterized by a set of genes and a set of individual cells, and b) a plurality of gene lists, each gene list corresponding to a biological pathway,
- at least one processor configured to:
  - determine a gene-cell expression matrix based on said sequencing data and said plurality of gene lists, said gene-expression matrix comprising rows corresponding to genes of said set of genes referred to as pathway genes and belonging to at least one among said gene lists,
  - carry out a principal component analysis (PCA) on said gene-expression matrix, so as to obtain a plurality of modes of activation of at least one among said biological pathways, each mode of activation corresponding to one principal component obtained from the PCA,
  - selecting a subset of modes of activation among said plurality of modes of activation,
  - determining a matrix referred to as activity matrix, said activity matrix comprising

scores quantifying a level of activity of each mode of activation among said subset of modes of activation within each individual cell among said set of individual cells,
- at least one output interface configured to output said activity matrix.

[0027] In addition, the disclosure relates to a computer program comprising software code adapted to perform a method for determining and quantifying modes of activation of biological pathways in individual cells compliant with any of the above execution modes when the program is executed by a processor.

[0028] The present disclosure further pertains to a non-transitory program storage device, readable by a computer, tangibly embodying a program of instructions executable by the computer to perform a method for determining and quantifying modes of activation of biological pathways in individual cells, compliant with the present disclosure.

## BRIEF DESCRIPTION OF THE DRAWINGS

[0029] The present disclosure will be better understood, and other specific features and advantages will emerge upon reading the following description of particular and non-restrictive illustrative embodiments, the description making reference to the annexed drawings wherein:

[0030] **Figures 1A and 1B** show an overview of a process MAYA according to some embodiments. 1A shows that MAYA takes as input a scRNA-Seq dataset and reference gene lists, and produces as output an activity matrix, with for each cell its activity score for each mode of every reference gene list. 1B shows an example of MAYA outputs: a heatmap to visualize the modes of activation of reference pathways, or a Uniform Manifold Approximation and Projection (UMAP) of the activity matrix to visualize cells according to any annotation (activity scores for different modes, predicted cell type or any user annotation).

[0031] **Figures 2A to 2G** represents graphs relating to activation modes of Hallmark pathways in kidney with MAYA. 2A shows a heatmap of activity matrix computed on kidney dataset with MSigDB Hallmark pathways, initial author annotation is indicated above heatmap. The two activation modes of Allograft Rejection are highlighted in bold

and further described in the subsequent panels, and the four modes of activation of TNFA signaling via NFKB are further described in Supplementary Fig. 2. F2B shows a scatterplot of Mode 2 versus Mode 1cell activity scores. Associated density histograms are indicated on the sides of the graph. 2C shows a heatmap of scaled gene expression for top 10 contributing genes for Mode1 (top) and Mode2 (bottom) of Allograft Rejection pathway, ordered by decreasing contribution for each. 2D shows a UMAP representation of activity matrix of Hallmark pathways, cells are colored according to author annotation, or activity scores of modes 1 and 2 of Allograft rejection pathway. Specificity score of cell populations is displayed next to relevant clusters. 2E shows a heatmap of activity scores computed by Pagoda2, AUCell and MAYA for Allograft Rejection pathway, cells are grouped according to author annotation. 2F shows a barplot representation of the detection rate of modes 1 to 3 for the pathway Allograft Rejection when adding various numbers of random genes to the pathway gene list (n = 100 experiments each). Barplots are colored according to the cell population with the highest specificity score for the identified mode. 2G shows a jitter representation of specificity scores of modes 1 and 2 grouped by level of added noise, data points are colored according to author annotation. Specificity obtained for each mode with initial gene list is represented with a dashed line.

[0032] **Figures 3A to 3D** are graphs pertaining to KEGG and REACTOME activity in colon with MAYA. 3A shows a heatmap of scaled gene expression for top10 contributing genes for the three activation modes of KEGG Cell Adhesion Molecules pathway, ordered by decreasing contribution. 3B shows a UMAP representation of activity matrix of KEGG pathways, cells are colored according to author annotation, or activity scores of the three modes of Cell Adhesion Molecules pathway. Specificity score of cell populations is displayed next to relevant clusters. 3C shows a heatmap of scaled gene expression for top10 contributing genes for the four activation modes of the Ion Channel Transport pathway, ordered by decreasing contribution. Figure 3D shows a UMAP representation of activity matrix of REACTOME pathways, cells are colored according to author annotation, or activity scores of the four modes of Ion Channel Transport pathway. Specificity score of cell populations is displayed next to relevant clusters.

[0033] **Figures 4A to 4H** show several results pertaining to the annotation of cell types with MAYA. 4A shows a gene-based UMAP representation of kidney dataset, cells are colored according to author annotation. 4B shows a heatmap representing for each author annotation (rows) the fraction of cells labeled with each MAYA annotation (columns) for the kidney dataset. 4C shows an overlaid jitter and boxplot representation of F1-scores for automatic annotation of the kidney dataset using Cell-ID, SCINA and MAYA, data points are colored according to author annotation; n = 5 cell types; Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; Adjusted p-values from two-sided Wilcoxon test are symbolized with: *: <0.05, **: <0.01, ***: <0.001, ****: <0.0001 (pval = 0.011 for Cell-ID and for SCINA). 4D shows a gene-based UMAP representation of colon dataset, cells are colored according to author annotation. 4E shows a heatmap representing for each author annotation (rows) the fraction of cells labeled with each MAYA annotation (columns) for the colon dataset. 4F shows an overlaid jitter and boxplot representation of F1-scores for automatic annotation of the colon dataset using Cell-ID, SCINA and MAYA, data points are colored according to author annotation; n = 10 cell types; Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; Adjusted p-values from two-sided Wilcoxon test are not significant. 4G shows a UMAP representation of the larynx dataset, either gene-based or based on activity matrix of PanglaoDB cell-type markers lists, cells are colored according to cell type or to patient. 4H shows an overlaid jitter and boxplot representation of Shannon Diversity Index (SDI), for clusters derived from gene-based dimensionality reduction, Harmony dimensionality reduction, CCA dimensionality reduction and MAYA activity matrix of the larynx dataset; n = 12, 11, 7, and 7 independent clusters respectively; Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; Adjusted p-values from two-sided Wilcoxon test are symbolized with: *: <0.05, **: <0.01, ***: <0.001, ****: <0.0001. Adjusted p-value is 0.026 for Gene-based relative to MAYA and non-significant for other comparisons.

[0034] **Figures 5A to 5H** show results illustrating how MAYA detects pathway activation in tumors across patients. 5A shows a UMAP representation of activity matrix of Hallmark pathways, cells are colored according to author annotation and patient.

Clusters derived from activity matrix are displayed next to relevant groups of cells. 5B shows an overlaid jitter and boxplot representation of Shannon Diversity Index (SDI), for clusters derived from gene-based dimensionality reduction and MAYA activity matrix of the ovary dataset. Clusters corresponding to tumor cells are colored in pink; n = 11 and 4 independent clusters respectively; Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range. 5C shows a barplot representation of specificity scores of the top5 specific modes for the four most prevalent populations in the dataset. 5D shows a heatmap of activity scores of the three modes of the Hallmark Epithelial Mesenchymal Transition (EMT) pathway, initial author annotation is indicated above heatmap. Heatmap of scaled gene expression for top10 contributing genes for the three modes of EMT, ordered by decreasing contribution. 5E shows a UMAP representation of activity matrix of Hallmark pathways, cells are colored according to activity scores of the three EMT modes. Specificity score of cell populations is displayed next to relevant clusters. Violin plots of activity scores for corresponding modes, grouped by author annotation (adjusted p-values from two-sided Wilcoxon test are symbolized with: *: <0.05, **: <0.01, ***: <0.001, ****: <0.0001). Heatmap of the activity scores for the modes of 5F-Hallmark Estrogen Response Early pathway, 5G-Hallmark Coagulation pathway and 5H- KEGG Wnt pathway, cells are grouped according to author annotation. Heatmap of scaled gene expression for top10 contributing genes for corresponding modes, ordered by decreasing contribution.

[0035] **Figure 6A** shows a scatterplot representing the number of contributing genes versus the maximum gene contribution, for the first five modes of all pathways from the KEGG pathway list on the kidney dataset. Error band represents the 95% confidence interval for locally estimated scatterplot smoothing (LOESS).

[0036] **Figure 6B** shows a scatterplots of the average mode specificity, the number of informative modes and the number of informative pathways according to the maximum single-gene contribution. Default cut-off of maximum single gene variance (0.4) was chosen to maximize the specificity of the modes of activation and is indicated as a vertical dashed line. Error bands represent the 95% confidence interval for locally estimated scatterplot smoothing (LOESS).

[0037] **Figure 7** represents heatmap of activity matrices for different cut-off of single-gene contribution: 0.2, 0.4 and 0.9.

[0038] **Figure 8** is a block diagram representing schematically a particular mode of a device for determining and quantifying modes of activation of biological pathways in individual cells, compliant with the present disclosure.

[0039] **Figure 9** is a flow chart showing successive steps executed with the device for determining and quantifying modes of activation of biological pathways in individual cells of figure 8.

## DEFINITIONS

[0040] In the present invention, the following terms have the following meanings:

[0041] The terms "**adapted**" and "**configured**" are used in the present disclosure as broadly encompassing initial configuration, later adaptation or complementation of the present device, or any combination thereof alike, whether effected through material or software means (including firmware).

[0042] The term "**processor**" should not be construed to be restricted to hardware capable of executing software, and refers in a general way to a processing device, which can for example include a computer, a microprocessor, an integrated circuit, or a programmable logic device (PLD). The processor may also encompass one or more Graphics Processing Units (GPU), whether exploited for computer graphics and image processing or other functions. Additionally, the instructions and/or data enabling to perform associated and/or resulting functionalities may be stored on any processor-readable medium such as, e.g., an integrated circuit, a hard disk, a CD (Compact Disc), an optical disc such as a DVD (Digital Versatile Disc), a RAM (Random-Access Memory) or a ROM (Read-Only Memory). Instructions may be notably stored in hardware, software, firmware or in any combination thereof.

## DETAILED DESCRIPTION

[0043] The present description illustrates the principles of the present disclosure. It will thus be appreciated that those skilled in the art will be able to devise various arrangements that, although not explicitly described or shown herein, embody the principles of the disclosure and are included within its scope.

[0044] All examples and conditional language recited herein are intended for educational purposes to aid the reader in understanding the principles of the disclosure and the concepts contributed by the inventor to furthering the art, and are to be construed as being without limitation to such specifically recited examples and conditions.

[0045] Moreover, all statements herein reciting principles, aspects, and embodiments of the disclosure, as well as specific examples thereof, are intended to encompass both structural and functional equivalents thereof. Additionally, it is intended that such equivalents include both currently known equivalents as well as equivalents developed in the future, i.e., any elements developed that perform the same function, regardless of structure.

[0046] Thus, for example, it will be appreciated by those skilled in the art that the block diagrams presented herein may represent conceptual views of illustrative circuitry embodying the principles of the disclosure. Similarly, it will be appreciated that any flow charts, flow diagrams, and the like represent various processes which may be substantially represented in computer readable media and so executed by a computer or processor, whether or not such computer or processor is explicitly shown.

[0047] The functions of the various elements shown in the figures may be provided through the use of dedicated hardware as well as hardware capable of executing software in association with appropriate software. When provided by a processor, the functions may be provided by a single dedicated processor, a single shared processor, or a plurality of individual processors, some of which may be shared.

[0048] It should be understood that the elements shown in the figures may be implemented in various forms of hardware, software or combinations thereof. Preferably, these elements are implemented in a combination of hardware and software on one or

more appropriately programmed general-purpose devices, which may include a processor, memory and input/output interfaces.

[0049] The present disclosure will be described in reference to a particular functional embodiment of a device 1 for determining and quantifying modes of activation of biological pathways in individual cells, as illustrated on Figure 8.

[0050] The device 1 is adapted to produce an activity matrix 31 (also called multimodal score matrix) providing multimodal activity scores per pathway.

[0051] The input sequencing data 21 may be derived from a single-cell RNA sequencing method, said sequencing data being characterized by a set of genes and a set of individual cells. The input gene lists 22 may be derived from a gene dataset which is a curated gene dataset.

[0052] In one example gene lists 22 may be derived from one of the following databases KEGG database https://www.genome.jp/kegg/pathway.html, MSigDB database https://www.gsea-msigdb.org/gsea/msigdb, or any custom gene list. Each gene list of the plurality of gene lists corresponding to a biological pathway or a gene signature of interest.

[0053] Though the presently described device 1 is versatile and provided with several functions that can be carried out alternatively or in any cumulative way, other implementations within the scope of the present disclosure include a device having only parts of the present functionalities.

[0054] The device 1 is advantageously an apparatus, or a physical part of an apparatus, designed, configured and/or adapted for performing the mentioned functions and produce the mentioned effects or results. In alternative implementations, the device 1 is embodied as a set of apparatus or physical parts of apparatus, whether grouped in a same machine or in different, possibly remote, machines. The device 1 may e.g. have functions distributed over a cloud infrastructure and be available to users as a cloud-based service, or have remote functions accessible through an API.

[0055] In what follows, the modules are to be understood as functional entities rather than material, physically distinct, components. They can consequently be embodied either as grouped together in a same tangible and concrete component, or distributed into several

such components. Also, each of those modules is possibly itself shared between at least two physical components. In addition, the modules are implemented in hardware, software, firmware, or any mixed form thereof as well. They are preferably embodied within at least one processor of the device 1.

[0056] The device 1 comprises a module 11 for receiving the sequencing data 21 and the plurality of gene lists 22, stored in one or more local or remote database(s) 10. The latter can take the form of storage resources available from any kind of appropriate storage means, which can be notably a RAM or an EEPROM (Electrically-Erasable Programmable Read-Only Memory) such as a Flash memory, possibly within an SSD (Solid-State Disk). Alternatively, the sequencing data 21 and the plurality of gene lists 22are received from a communication network.

[0057] A gene signature of interest may be obtained from genes identified as up-regulated in a cell population in a former experiment, or genes up-regulated following genetic manipulation in a model system. For example, the gene signature is associated to a cancer cell population and has been previously identified as up-regulated in this specific cancer cell population.

[0058] The device 1 further comprises optionally a module for preprocessing the received the sequencing data 21 and the plurality of gene lists 22.

[0059] The device 1 may also comprise a module 12 for determining a gene-cell expression matrix based. Module 12 is configured to determining a gene-cell expression matrix based on said sequencing data 21 and said plurality of gene lists 22, said gene-expression matrix comprising rows corresponding to genes of said set of genes referred to as pathway genes and belonging to at least one among said gene lists.

[0060] The device 1 may also comprise a module 13 carrying out a PCA on said gene-cell expression matrix, so as to determine a plurality of modes of activation of at least one among said biological pathways, each mode of activation corresponding to one among the principal components obtained from the PCA. Each mode of activation among said plurality of modes of activation is defined by a specific subset of genes among said pathway genes.

[0061] The device 1 may comprise as well as a module 14 configured to select a subset of so-called effective modes of activation among said plurality of modes of activation. This module 14 may for example perform a step of detecting bimodal distributions of scores in said activity matrix and a step of verifying that the number of active cells is superior to a predefined minimum fraction of the set of individual cells. The minimum fraction may be set to 0.05, expecting more of 5% of cells of the studied dataset to activate a pathway. Depending on the dataset and the frequency of the population of interest, such threshold can be tuned, and for example set to 0 for maximal sensitivity.

[0062] Module 14 may also comprise selecting modes of activation for which said specific subset of genes comprises a number of genes higher than a predefined threshold.

[0063] The device 1 further comprises a module 15 for determining a matrix referred to as activity matrix, said activity matrix comprising scores quantifying a level of activity of each effective mode of activation among said subset of effective modes of activation within each individual cell among said set of individual cells. The scores are calculated as the coordinate of the cell on the principal component axis related to the mode of activation of the pathway.

[0064] The device 1 is interacting with a user interface 16, via which information can be entered and retrieved by a user. The user interface 16 includes any means appropriate for entering or retrieving data, information, or instructions, notably visual, tactile and/or audio capacities that can encompass any or several of the following means as well known by a person skilled in the art: a screen, a keyboard, a trackball, a touchpad, a touchscreen, a loudspeaker, a voice recognition system.

In its automatic actions, the device 1 may for example execute the following process (Figure 9):

- receiving sequencing data 21 and the plurality of gene lists 22 (step 41),
- determining a gene-cell expression matrix based on said sequencing data and said plurality of gene lists (step 42),
- carrying out a principal component analysis (PCA) on said gene-cell expression matrix (step 43),

- selecting a subset of so-called effective modes of activation among said plurality of modes of activation (step 44),
- determining a matrix referred to as activity matrix 31 (step 45).

[0065] A particular apparatus, is embodying the device 1 described above. It corresponds for example to a workstation, a laptop, a tablet, a smartphone, or a head-mounted display (HMD).

[0066] It comprises the following elements, connected to each other by a bus of addresses and data that also transports a clock signal:

- a microprocessor (or CPU);
- a graphics card comprising several Graphical Processing Units (or GPUs) and a Graphical Random Access Memory (GRAM);
- a non-volatile memory of ROM type;
- a RAM;
- one or several I/O (Input/Output) devices such as for example a keyboard, a mouse, a trackball, a webcam; other modes for introduction of commands such as for example vocal recognition are also possible;
- a power source; and
- a radiofrequency unit.

[0067] According to a variant, the power supply is external to the apparatus 9.

[0068] The apparatus also comprises a display device of display screen type directly connected to the graphics card 92 to display synthesized images calculated and composed in the graphics card. According to a variant, a display device is external to apparatus 9 and is connected thereto by a cable or wirelessly for transmitting the display signals. The apparatus, for example through the graphics card, comprises an interface for transmission or connection adapted to transmit a display signal to an external display means such as for example an LCD or plasma screen or a video-projector.

[0069] When switched-on, the microprocessor 91 loads and executes the instructions of the program contained in the RAM 97.

[0070] As will be understood by a skilled person, the presence of the graphics card 92 is not mandatory, and can be replaced with entire CPU processing and/or simpler visualization implementations.

[0071] In variant modes, the apparatus may include only the functionalities of the device 1. In addition, the device 1 may be implemented differently than a standalone software, and an apparatus or set of apparatus comprising only parts of the apparatus may be exploited through an API call or via a cloud interface.

## EXAMPLE

[0072] The present invention is further illustrated by the following examples.

[0073] To inspect existing pathway databases with single-cell resolution, the inventors developed the method of the present invention, also named MAYA (Multimodes of pathwAY Activation), a tool that detects—for each pathway—the different modes of activation across cell types, each mode relying on a different subset of genes. The inventors argue that MAYA could be a way for currently available biological knowledge to meet the granularity reached by single-cell data and help researchers go deeper in their understanding of complex cellular mechanisms. Particularly, in the case of cancer datasets, the inventors show that MAYA can detect cell type specific modes of pathway activation for both the microenvironment and tumor cells, defining common expression programs across patients, in line with recently identified tumor “meta-programs”.

### Results

[0074] MAYA method

[0075] MAYA enables comprehensive pathway study thanks to multimodal scoring of gene lists in individual cells (Fig. 1).

[0076] Figure 1 shows an overview of MAYA. As shown on Figure 1a, MAYA takes as input a scRNEQ-Seq dataset and reference gene lists, and produces as output an activity matrix, with for each cell its activity score for each mode of every reference gene list. Figure 1b shows an example of MAYA outputs: a heatmap to visualize the modes of activation of reference pathways, or a Uniform Manifold Approximation and Projection (UMAP) of the activity matrix to visualize cells according to any annotation (activity scores for different modes, predicted cell type or any user annotation).

[0077] Provided a scRNA-seq count matrix and pathway lists, MAYA detects all biologically relevant ways to activate each pathway based on subgroups of genes and summarizes their activity in each cell in a multimodal score matrix (Fig. 1a). This activity matrix can then be used to identify groups of cells sharing similar activation of provided pathways and to visualize datasets in lower dimensions (Fig. 1b). As a comparison, reference tools that measure pathway activity, such as AUCell or Pagoda2, provide a unique activity score per pathway where MAYA can provide several.

[0078] MAYA is built on two main functions that are applied to each provided gene list: the detection of activation modes and the selection of biologically relevant ones. Detection of modes is performed thanks to a PCA on a normalized gene-cell expression matrix restricted to pathway genes (Fig. 1a). The purpose of such decomposition of the matrix is to find, within the pathway, genes which expression is coordinated and variable across cells, and to score their activity in individual cells. Each principal component (PC) represents a possible mode of activation of the pathway, that is characterized by the genes that contribute the most to the PC, and by a score that corresponds to the cell coordinate on the PC. Each gene can contribute to several PCs and therefore to several modes.

[0079] However, not all detected modes reflect a relevant biological pattern in the data and some could be driven by outliers, either cells and/or genes; this probability increases as modes explain less and less variance in the dataset (Supplementary Fig. 1a). The inventors thus developed a method to assess the informativity of each mode, based on two biologically interpretable criteria. First, an informative mode should be more active in a minimal subset of cells compared with other cells. This is assessed by detecting bimodal distributions of scores across cells and checking that the group of active cells

represents more than a minimum fraction of the population. This fraction can be determined based on previous knowledge of the underlying biology or set arbitrarily (Supplementary Fig. 1b–d). Second, an informative mode should be driven by enough genes to be considered as a mode of activation per se and not solely correspond to the expression of a single outlier gene. To that end, the inventors determined a cutoff for maximal variance of each gene of a mode, indicative of how much a gene can contribute on its own (Supplementary Fig. 1e, f). Default cutoff value was chosen to maximize the number of modes detected as informative while keeping a high average number of genes significantly contributing to each mode (Supplementary Fig. 1g). This method is robust across single-cell technologies, whether 10X Chromium or Smart-Seq2, as shown with a PBMC dataset generated using both technologies (Supplementary Fig. 1h).

[0080] Although MAYA's main purpose is to detect multimodal activation of pathways, it can also perform unimodal activity scoring to detect cell identity from any cell marker lists. To this end, the inventors have developed a built-in function to leverage MAYA's scoring and informativity methods to annotate cells in a dataset. This approach is based on activation of the first mode using PanglaoDB as input gene lists by default (see Methods' paragraph). This function allows cluster-free cell type annotation in a timely fashion as it annotates a dataset of around 16,000 cells in less than 1 min and 125,000 cells in ~15 min (Supplementary Fig. 1i).

[0081] **MAYA detects biologically relevant multimodal pathway activity in kidney**

[0082] The main distinguishing feature of MAYA over existing pathway scoring tools is the multimodality of its activity score, which proves useful when studying broad pathways in complex biological systems. The inventors first sought to demonstrate its ability to detect cell-type specific activation modes of hallmark pathways. For that, the inventors ran MAYA on a dataset of normal kidney and immune cells, from which the inventors selected cells from five distinct subtypes for clarity (n = 1252). The inventors used the MSigDB Hallmark pathways as input gene lists, covering main biological functions. Figure 2 shows the activation modes of Hallmark pathways in kidney with MAYA. Unsupervised clustering on the multimodal activity matrix shows MAYA detects modes that distinguish different cell populations (Fig. 2a). Figure 2a shows a heatmap of

activity matrix computed on kidney dataset with MSigDB Hallmark pathways, initial author annotation is indicated above heatmap. The two activation modes of Allograft Rejection are highlighted in bold and further described in the subsequent panels, and the four modes of activation of TNFA signaling via NFKB are further described in Supplementary Fig. 2. More specifically, the inventors noticed that modes from the same pathway were specifically activated in different cell types. As an example, the Allograft rejection pathway presents two modes of activation (Fig. 2b–d): (i) mode 1, driven by the expression of CTSS and SPI1—known to have a critical role in antigen presentation20 and gene regulation during myeloid development21—and specific to monocytes (specificity of 0.57), and (ii) mode 2, driven by CD2, CD3E and CD3D—coding for T cell surface proteins—and by CD8A and CD8B - coding for the CD8 antigen—and specific to CD8 T cells (specificity of 0.88). Figure 2b shows a scatterplot of Mode 2 versus Mode 1 cell activity scores. Associated density histograms are indicated on the sides of the graph. Figure 2c shows a heatmap of scaled gene expression for top 10 contributing genes for Mode1 (top) and Mode2 (bottom) of Allograft Rejection pathway, ordered by decreasing contribution for each. Figure 2d shows a UMAP representation of activity matrix of Hallmark pathways, cells are colored according to author annotation, or activity scores of modes 1 and 2 of Allograft rejection pathway. Specificity score of cell populations is displayed next to relevant clusters.

[0083] In contrast, AUCell and Pagoda2 both describe this pathway with a single score, corresponding to an aggregation of MAYA's mode 1 and 2, or mode 1 only respectively (Fig. 2e). Figure 2e shows a h Heatmap of activity scores computed by Pagoda2, AUCell and MAYA for Allograft Rejection pathway, cells are grouped according to author annotation. Another detailed example is shown in Supplementary Fig. 2 for the TNFA signaling via NFKB pathway, where four activation modes were detected with MAYA based on their bimodal activity distribution (Supplementary Fig. 2a): one specific to monocytes, one to CD8 T cells and two to endothelial cells (Supplementary Fig. 2b–d). Interestingly, each mode involves a different interleukin, each specific to the population in which the mode is found to be active: (i) IL6ST is a signal transducer, which dimerizes with IL6R and is bound for instance by IL-6, resulting in the activation of downstream cascades in endothelial cells22, (ii) IL1B is a lymphocyte activating factor produced by

monocytes, macrophages and neutrophils, and (iii) IL7R is associated with T cell differentiation. Altogether, the inventors demonstrate here that MAYA identifies relevant cell-type specific modes of pathway activation from broad reference gene lists.

[0084] To test both the stability and the ability of MAYA to detect biologically relevant signal from noisy gene lists, the inventors added 10, 50, 100, and 200 random genes to the initial 200 genes of the pathways Allograft rejection and TNFA signaling via NFKB; each experiment was repeated 100 times. For the Allograft Rejection pathway, the two initial activation modes were detected for all modified gene lists with a high cell-type specificity, whatever the level of added noise (Fig. 2f, g). Figure 2f shows a barplot representation of the detection rate of modes 1 to 3 for the pathway Allograft Rejection when adding various numbers of random genes to the pathway gene list (n = 100 experiments each). Barplots are colored according to the cell population with the highest specificity score for the identified mode. Figure 2g shows a jitter representation of specificity scores of modes 1 and 2 grouped by level of added noise, data points are colored according to author annotation. Specificity obtained for each mode with initial gene list is represented with a dashed line. These results also show the accuracy of our selection method to detect relevant modes, as the inventors rarely detect additional activation modes (corresponding to PC3/mode 3) even when randomly increasing the reference gene lists. Similarly, for the TNFA signaling pathway, the first three modes are robust to noise, with a decrease in sensitivity of detection when adding more than 100 unrelated genes (Supplementary Fig. 2e).

[0085] **MAYA detects biologically relevant multimodal pathway activity in colon**

[0086] The inventors then illustrated the relevance of the biological insight gained by using multimodal pathway analysis for another tissue with a dataset of colon and immune cells from Lee et al. (Lee, H. O. et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Nat. Genet. 52, 594–603 (2020).) —from which the inventors selected cells from ten distinct cell types (n = 1415)—and using the MSigDB KEGG and REACTOME pathways24. Both analyses recover cell-type specific activation modes, given the clustering of cells by cell type on the heatmaps derived from the activity matrix (Supplementary Fig. 3a, c). Focusing on KEGG cell

adhesion molecules list, the inventors observed that MAYA was able to detect several well-known types of cell-cell adhesion processes starting from the mixed general reference list (Fig. 3a, b and Supplementary Fig. 3b): (i) mode 1 driven by the expression of HLA genes coding MHC class II molecules25, detected in antigen-presenting cells—monocytes and dendritic cells—and B cells, with a specificity of 0.29, 0.27 and 0.15 respectively, (ii) mode 2 driven by the expression of genes coding for claudins and cadherins located at tight junctions26,27, specifically activated in epithelial cells (specificity of 0.24 and 0.16 for enterocytes and goblet cells respectively), and (iii) mode 3 driven by the expression of T cell membrane molecules, specific to Regulatory T cells (specificity of 0.29). Figure 3a shows a heatmap of scaled gene expression for top10 contributing genes for the three activation modes of KEGG Cell Adhesion Molecules pathway, ordered by decreasing contribution. Figure 3b shows a UMAP representation of activity matrix of KEGG pathways, cells are colored according to author annotation, or activity scores of the three modes of Cell Adhesion Molecules pathway. Specificity score of cell populations is displayed next to relevant clusters.

[0087] Applying MAYA to the REACTOME pathway ion channel transport, the inventors were able to detect different types of ion channels and functions, specific to each cell population (Fig. 3c, d and Supplementary Fig. 3d). Mode 1 is specific to colon epithelial cells (specificity of 0.34 and 0.24 for enterocytes and goblet cells respectively, Fig. 3d) and corresponds to two types of ion channels—Epithelial Sodium Channel (ENaCs) and Na,K-ATPase28 - that have been shown to participate to the regulation of salt and water absorption from the colon lumen29,30. In particular, activation mode 1 captures genes regulating ENaCs and their residence at the apical membrane: SCNN1A encodes a subunit of ENaCs, NEDD4L participates to ENaCs ubiquitination which leads to their retrieval from cell surface32 and SGK1 is known to phosphorylate NEDD4L product, which decreases its binding to ENaCs33,34. Mode 4 is specific to goblet cells only, driven by the expression of the genes CLCA1 and BEST2. These two genes are associated with Calcium-activated Chloride Channels (CaCCs) that have been shown to participate in epithelial secretion35. Mode 3 is specific to pericytes and smooth muscle cells (specificity of 0.16 and 0.22 respectively) and is associated with Calcium homeostasis (ATP2B4, PLN, CASQ2) and Na,K-ATPases (FXYD1, FXYD6, ATP1A2,

ATP1B2), two important channels for the membrane polarization of contractile cells. Finally, mode 2, mainly active in monocytes and dendritic cells (specificity of 0.33 and 0.16 respectively), involves genes associated with acidification of intracellular organelles through colocalization of V-type proton ATPases36 (ATP6V1B2, ATP6AP1, ATP6V1F, ATP6V0E1, ATP6V0D237) and Chloride channels38 (TTYH3, CLIC2), a process necessary for phagocytosis. Altogether, as for the kidney, starting from broad reference databases, MAYA untangles pathway activities specific to each cell type, revealing precise cell functions. Figure 3c shows a heatmap of scaled gene expression for top10 contributing genes for the four activation modes of the Ion Channel Transport pathway, ordered by decreasing contribution. Figure 3d shows a UMAP representation of activity matrix of REACTOME pathways, cells are colored according to author annotation, or activity scores of the four modes of Ion Channel Transport pathway. Specificity score of cell populations is displayed next to relevant clusters.

[0088] **MAYA assigns cell identity**

[0089] The inventors then leveraged MAYA's scoring and selection ability to robustly assign cell identity. The inventors applied MAYA to PanglaoDB cell marker lists and the subsets of kidney and colon datasets used previously (Fig. 4a, d). Figure 4a shows a gene-based UMAP representation of kidney dataset, cells are colored according to author annotation. Figure 4d shows a gene-based UMAP representation of colon dataset, cells are colored according to author annotation.

[0090] The inventors demonstrate that MAYA enabled an assisted and accurate annotation of each cell in the two datasets, using the initial cell type annotation by authors as a reference (Fig. 4b, e). Figure 4b shows a heatmap representing for each author annotation (rows) the fraction of cells labeled with each MAYA annotation (columns) for the kidney dataset. Figure 4e shows a heatmap representing for each author annotation (rows) the fraction of cells labeled with each MAYA annotation (columns) for the colon dataset. The inventors compared the accuracy of our predictions with the ones obtained with two other cell type identification methods: Cell-ID39, based on Multiple Correspondence Analysis (MCA), and SCINA40, based on an expectation-maximization model. MAYA presents among the highest rates of recall and precision for both datasets

(Fig. 4c, f and Supplementary Fig. 4a, b). Figure 4c shows an overlaid jitter and boxplot representation of F1-scores for automatic annotation of the kidney dataset using Cell-ID, SCINA and MAYA, data points are colored according to author annotation; n = 5 cell types; Center line, median; box limits, upper and lower quartiles; whiskers, 1.5×interquartile range; Adjusted p-values from two-sided Wilcoxon test are symbolized with: *: <0.05, **: <0.01, ***: <0.001, ****: <0.0001 (pval = 0.011 for Cell-ID and for SCINA). Figure 4f shows an overlaid jitter and boxplot representation of F1-scores for automatic annotation of the colon dataset using Cell-ID, SCINA and MAYA, data points are colored according to author annotation; n = 10 cell types; Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; Adjusted p-values from two-sided Wilcoxon test are not significant.

[0091] The inventors finally tested the scalability of MAYA and its ability to detect rare cell types on a dataset with 16,815 cells from ovarian tumors6 (Supplementary Fig. 4c). Overall, MAYA had an average precision of 51% and recall of 68%. Notably, B cells were identified with a precision and recall of 98% when they represent only 4.9% of the dataset and endothelial cells with a precision of 100% and recall of 85% when they represent 0.2% of cells in the dataset (Supplementary Fig. 4d). Lower precision is achieved for some types probably due to overlap between cell type markers in PanglaoDB, such as between NK cells and T cells (28 shared markers out of 80 and 95 markers respectively), dendritic cells and macrophages (34 shared out of 121 and 128 markers respectively), and endothelial cells and fibroblasts (13 shared out of 187 and 171 markers respectively). All three pairs of cell types share more genes than with any other type from the PanglaoDB lists.

[0092] Furthermore, as batch effect is a main concern in single-cell analyses, the inventors tested whether MAYA was affected by such technical biases. The inventors worked on a dataset containing n = 5179 cells from laryngeal squamous cell carcinoma biopsies of two patients with a batch effect between patients41. Using standard gene-based scRNA-seq matrix processing, cells from the same cell types – whether cells from the microenvironment or the tumors—indeed cluster by patient whereas clustering on the MAYA activity matrix groups cells by cell type, with cells from both patients within the

same cluster (Fig. 4g). Figure 4g shows UMAP representation of the larynx dataset, either gene-based or based on activity matrix of PanglaoDB cell-type markers lists, cells are colored according to cell type or to patient.

[0093] To quantify the inter-patient overlap between clusters of similar cell types, the inventors computed the Shannon Diversity Index (SDI) for both methods as well as for clusters obtained with the reference integration tool Harmony42 and an integration method based on Canonical Correlation Analysis43 (CCA) (Supplementary Fig. 4e). MAYA had an average SDI of 0.77 against 0.76, 0.63, and 0.23 for the integration based on Harmony, on CCA and the gene-based method respectively (Fig. 4h). Figure 4h shows an overlaid jitter and boxplot representation of Shannon Diversity Index (SDI), for clusters derived from gene-based dimensionality reduction, Harmony dimensionality reduction, CCA dimensionality reduction and MAYA activity matrix of the larynx dataset; n = 12, 11, 7, and 7 independent clusters respectively; Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; Adjusted p-values from two-sided Wilcoxon test are symbolized with: *: <0.05, **: <0.01, ***: <0.001, ****: <0.0001. Adjusted p-value is 0.026 for Gene-based relative to MAYA and non-significant for other comparisons.

[0094] To confirm these observations, the inventors compared the four methods on a pancreas dataset containing n = 14,890 cells generated with five different single cell technologies44 (Supplementary Fig. 4e, f). Similarly, the gene-based method generated technology-specific clusters whereas MAYA along with the two other integration methods consistently mixed cells from different technologies. In addition to pathway scoring, MAYA can perform accurate cell type annotation independently of batch effect, making it an all-in-one tool to address both cell identity and function.

[0095] **MAYA detects common modes of pathway activation across cancer patients**

[0096] Patient-specificity of cancer cells is currently a major limitation for the comprehensive study of cancer scRNA-seq datasets. Cells of the microenvironment coming from different patients can easily group together, showing the absence of a major batch effect between samples, while tumor cells form distinct clusters4–9. Such behavior

is thought to be due in part to the genetic variations across tumor cells from different patients, notably copy-number variations. Integration methods, correcting for general batch effect in samples, such as Harmony42, are not suited to deal with such cell-type specific effect.

[0097] The inventors demonstrate here that MAYA can be an alternative to gene-based or integration-based methods to identify common transcriptional features between cancer cells across patients. Using an ovarian cancer dataset, the inventors show that MAYA identifies several modes of pathway activation shared across patients (Fig. 5a–c and Supplementary Fig. 5a, b) that are associated with known cancer hallmarks. Indeed, top specific modes of epithelial cancer cells reflect the expression of genes associated with early response to estrogen or the P53 pathway (specificity of 0.45 and 0.31 respectively), that relate to tumor growth and proliferation (Fig. 5c). Such tumor-cell specific activation would not have been found with a classical GSEA approach (Supplementary Fig. 5c). Figure 5a shows UMAP representation of activity matrix of Hallmark pathways, cells are colored according to author annotation and patient. Clusters derived from activity matrix are displayed next to relevant groups of cells. Figure 5b shows an overlaid jitter and boxplot representation of Shannon Diversity Index (SDI), for clusters derived from gene-based dimensionality reduction and MAYA activity matrix of the ovary dataset. Clusters corresponding to tumor cells are colored in pink; n = 11 and 4 independent clusters respectively; Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range. Figure 5c shows a barplot representation of specificity scores of the top5 specific modes for the four most prevalent populations in the dataset.

[0098] MAYA also identifies modes of pathway activation specific to the tumor microenvironment, e.g a cell-type specific activation of complement genes in macrophages (specificity of 0.24) and of angiogenesis in cancer-associated fibroblasts (CAFs) (specificity of 0.40).

[0099] MAYA's multimodality allows to untangle several cell-type specific modes of activation for biological phenomena that are commonly difficult to sort out between cell populations within the tumors and their microenvironment. For example, MAYA detects different modes of epithelial-to-mesenchymal transition (EMT) (Fig. 5d, e): mode 1

specific to CAFs/mesothelial cells (specificity of 0.47 and 0.36 respectively), mode 2 specific to tumor cells (specificity of 0.30) and mode 3 to macrophages (specificity of 0.19) (Fig. 5d). Figure 5d shows Heatmap of activity scores of the three modes of the Hallmark Epithelial Mesenchymal Transition (EMT) pathway, initial author annotation is indicated above heatmap. Heatmap of scaled gene expression for top10 contributing genes for the three modes of EMT, ordered by decreasing contribution.

[0100] MAYA identifies a combination of genes that characterizes EMT occurring in epithelial cells, with LAMA3 and LAMC2 being exclusive to this cell type (Fig. 5d). These two genes expressed by basal epithelium code for two subunits of laminin 332, an essential component of epithelial basement membrane that promotes tumor cell motility45,46. In CAFs, MAYA detects EMT as driven mainly by genes encoding proteins from the extracellular matrix (ECM) including collagens, which have been shown to promote EMT in the tumor microenvironment directly47 or by increasing the ECM stiffness48,49. A third mode of EMT, characterized by the expression of the gene SPP1, is found in macrophages; macrophages have indeed been shown to be involved in EMT induction in various types of cancer50–53. Two additional modes are detected but are not as cell-type specific as the others (Supplementary Fig. 5a, maximum specificity scores of 0.12). Interestingly, EMT modes specific to CAFs and to tumor cells were also detected in two other cancer datasets (breast54 and lung55, Supplementary Fig. 5d). When evaluated in the same cells, the ovary, breast and lung modes display high pair-wise Pearson correlation coefficients between datasets and high cell type specificity for both the CAF and tumor-specific modes (Supplementary Fig. 5e).

[0101] Figures 5f, 5g and 5h show heatmaps of the activity scores for the modes of (f) Hallmark Estrogen Response Early pathway, (g) Hallmark Coagulation pathway and (h) KEGG Wnt pathway, cells are grouped according to author annotation. Heatmap of scaled gene expression for top10 contributing genes for corresponding modes, ordered by decreasing contribution.

[0102] MAYA also identifies two different modes of activation of the early estrogen response (Fig. 5f and Supplementary Fig. 6a), one specific to tumor cells, and another specific to CAFs, consistent with the observation that CAFs can use ER-mediated

signaling to promote tumor cell proliferation56,57. MAYA also helps to untangle the respective contribution of cancer cells and its microenvironment to the hemostatic imbalance observed in cancer58,59. It detects coagulation modes with high specificity for CAFs and mesothelial cells (0.31 and 0.32), tumor cells (0.22) and macrophages (0.24) (Fig. 5g and Supplementary Fig. 6b). In addition, when running MAYA with MSigDB KEGG gene lists, the inventors detect two modes of activation of the WNT pathway, one specific to tumor cells and the other specific to CAFs (Fig. 5h). It is indeed known from the literature that Wnt signaling is a complex pathway involving β-catenin in its canonical form, but at least three other non-canonical Wnt-mediated pathways have been proposed to function independently of β-catenin60. One of them involves activation of calcium/calmodulin-dependent kinase II (CamKII), protein kinase C and phosphatase CaN (PPP3CC), and is referred to as the "Wnt Ca2+ pathway" 61. Inspecting the top contributing genes of the MAYA modes, the inventors observed that mode 1 contains solely genes from the canonical WNT pathway, driven potentially by the ligand WNT7A, while mode 2 contains genes implicated in both canonical and non-canonical Wnt Ca2+ pathway, in particular through the ligand WNT5A62,63. This example illustrates how MAYA can untangle different ways to activate the same biological function.

[0103] Finally, it was recently proposed a method based on Non-negative Matrix Factorization (NMF) to identify de novo tumor "meta-programs" shared across patients (see Methods). Applied to epithelial ovarian cancer cells (Supplementary Fig. 6c–e), this approach identifies 'meta-programs' related to functions that were also detected by MAYA (E2F targets, hypoxia) and others that were not found as specific to tumor cells by MAYA as TNFA signaling. On the other hand, no "meta-program" was found to be associated with estrogen response— in contrast to MAYA—as it is activated in all tumor cells, showing the complementarity between the two approaches, one able to detect shared activation programs across all cells or a subset, and the other focusing on intra-tumoral heterogeneity.

[0104] Altogether, MAYA appears extremely powerful to detect modes of pathway activation across tumor cells from different patients as well as within the microenvironment - novel combinations of genes within known global reference gene

lists. The inventors see with these examples that MAYA can discover refined gene lists, specific to each population, matching the biological interpretation of pathway activation to the granularity of the single-cell measurements. For the study of tumor cells specifically, MAYA appears complementary to other recently proposed methods based on the discovery of de novo consensus transcriptomic programs.

[0105] **Discussion**

[0106] MAYA sorts out the different modes of pathway activation specific to each cell type, by automatically detecting subgroups of genes within reference pathways, and computing several scores of pathway activation. The inventors show that MAYA leverages existing biological knowledge to extract cell-type specific ways of activating pathways from single-cell datasets. In addition to pathway analysis, MAYA performs assisted cell typing as a side function, making it an all-in one tool for both cell type and cell function identification. MAYA proves particularly useful for single-cell cancer datasets, by (i) identifying common modes of pathway activation across patients in tumor cells, and also by (ii) dissecting the contribution of each population—fibroblast, immune & tumor cell—to the activation of a given pathway.

[0107] In comparison to previously published methods (AUCell14, Pagoda213, ROMA64, and UCell65), MAYA provides multiple activation scores per pathway. With MAYA, the inventors simplified bimodal detection by focusing on inflection points and introducing two biologically interpretable parameters, easily tunable by users: (i) a minimum proportion of cells that should activate a mode for the mode to be considered relevant and (ii) a maximum contribution to a mode that a single gene can have. MAYA will detect activation modes for a pathway given that the provided datasets present both cells that activate and cells that do not activate such pathway.

[0108] The inventors have also challenged the robustness to noise of our scoring and informativity methods and showed MAYA can detect relevant biological signal from noisy pathway lists. It can prove very useful as the inventors know pathway and cell markers manual curation is very time-consuming. Here, the inventors argue that MAYA

can take as input non-curated and potentially very exhaustive pathway or cell type lists and detect biological signal if they contain any.

[0109] The inventors also leveraged our methods of scoring and selection of informative scores to propose a built-in function to annotate cells using the first mode of activation of PanglaoDB cell type markers lists. This method has performance results equivalent to Cell-ID39 and SCINA40, two packages specialized in cell type annotation. MAYA is scalable to large datasets (>100,000 cells, in 15 min), unsensitive to batch effect, and is able to accurately detect and annotate cell populations representing less than 5% of cells. However, MAYA might not be suited to annotate cell types using few marker genes—the inventors recommend using MAYA with lists containing at least ten genes. In addition, for cell types which share many markers in reference databases, users will need additional expertise to validate the assisted annotation.

[0110] Finally, MAYA appears particularly useful when studying single-cell datasets from cancer patients that do not suffer from batch effect on all cell types but from patient-specificity for tumor cells. There is currently no standard way to address this challenge for data interpretation and a growing need to understand common cancer features across patients. Recently, the community was provided with recurrent shared transcriptional programs across patient and tumor types by describing 41 "meta-programs" grouped in 11 hallmarks of intra-tumor heterogeneity. These "meta-programs" were inferred de novo by studying scRNA-seq from multiple tissues and cancer types. This approach is very complementary to ours, where the inventors interrogate existing knowledge instead of performing de novo identification. MAYA identifies common modes of activation across tumor cells, which could be compared to such tumor meta-programs. In addition, MAYA deciphers the respective contribution of each cell population to the activation of a given pathway, by defining the group of genes that drive the pathway activity in each contributing population. Both inter and intra-patient features of MAYA will enable the identification of shared therapeutic vulnerabilities across patients, as well as various strategies to target them within the tumor eco-system.

[0111] **Methods**

[0112] **Matrix preprocessing**

[0113] All count matrices were processed with Seurat v3 to get the gene-based cell embeddings and check the consistency of author's annotations. Matrices were log-normalized using scale factor 10,000. Top 2000 variable features were found using "vst" method. PCA and UMAP were computed with default settings, using first 10 PCs for UMAP, which constitutes the "gene-based UMAP". For the larynx dataset, the two datasets were read separately and merged in a unique Seurat object of 5179 cells. The authors did not provide their annotation, so the inventors followed the default Seurat pipeline on each individual count matrix, performed PCA and default clustering. The inventors then annotated clusters based on expression of cell type markers described in the publication.

[0114] **Detailed description of MAYA algorithm**

[0115] **Building count matrix.** For a provided gene list, the log-normalized CPM matrix is subsetted to keep all cells but only genes from the list. Rows of the matrix are then scaled so that more highly expressed genes do not weight more than the others in the PCA that is later performed. The sign of each principal component is then chosen to favor the direction for which the absolute value of gene contribution is the highest. Each mode is scaled between 0 and 1. An iterative process then begins: the inventors evaluate the informativity of each successive PC starting from PC1. If a PC is found uninformative, the iteration stops, and the inventors do not interrogate further PCs. There is however an exception for PC1: the inventors interrogate PC2 even if PC1 is uninformative, as PC2 can still explain a significance part of the variance. The final activity matrix is built by gathering all modes from all gene lists in a single matrix with modes as rows and cells as columns.

[0116] **Informativity.** For each successive mode, a density curve is drawn from the distribution to get local maxima and minima. A bimodal curve is expected to have at least one minimum that will be low enough relative to its surrounding maxima on the y-axis to mark a clear distinction between two groups of cells (difference of at least 10% of global maximum density). Only local minima with abscissa superior to the one of the global

maximum are considered and iteratively evaluated in decreasing order as the point is to detect extreme behaviors and activation patterns that potentially occur in rare populations. The iteration stops when a potential minimum meets the criteria, or none was found. As this process relies on the detection of inflection points that depends itself on the adjustment of the density curve to the distribution, the inventors start with an adjustment meant to detect global variations of distributions and if none are detected the inventors test a more fitted adjustment to ensure no significant local variation was missed. Then follow two additional checks to ensure the biological relevance of the detected mode. First, the inventors filter out modes that are activated in very few cells as they could be outliers. The user can adjust this parameter based on what he expects to observe in the dataset or the number of cells from rarer cell type or set it to default 5%. The second biological check is based on the number of genes potentially contributing to the mode. However, it is hard to set a definition of what is a contributing gene to PCA; here the inventors consider that contributing genes contribute more than they would be expected i.e., if all genes from the pathway contributed the same (1/number of genes in the pathway). Given that pathways have various sizes, it is difficult to set a hard cutoff on this number of genes contributing to the mode. Instead, the inventors chose to set a cut-off on the maximum contribution of a gene to a mode. As the sum of squared gene contributions is equal to 1, if a gene contributes to up to 0.8, there is not much contribution left for other genes to share and this mode is probably driven by this unique gene. As a mode should represent joint expression of groups of genes, the inventors do not consider these monogenic modes biologically significant. Setting a threshold of 0.4 allows to remove monogenic modes while keeping a relatively large number of modes with higher cell type specificity. This parameter can also be changed by the user depending on the tolerance to probable monogenic pathways. Finally, the inventors chose to test the informativity of each pathway mode in decreasing order of variance explained in the dataset and to stop when a mode is found uninformative after mode 2 as the inventors know the following will explain even less variance and is more likely to be noise.

[0117] **Predict cell type.** Once the activity matrix generated, a k-Nearest Neighbors matrix with k = 20 is computed, then an adjacency matrix using Jaccard distance and finally transformed as a weighted graph using igraph function graph.adjacency.

Clustering is then performed using leiden_find_partition from leidenbase package with ModularityVertexPartition as partition type and a maximum number of iterations of 2. The average activity score is computed by cell type and by cluster. Each cluster is attributed the cell type for which the activity score is the highest, if it passes a threshold of default value 0, otherwise it is labeled as unassigned. This value can be modified by the user, depending on the level of confidence needed for annotation.

[0118] MAYA presents two tunable parameters, with default values: (i) a minimum proportion of cells that should activate a mode for the mode to be considered relevant (5%) and (ii) a maximum contribution to a mode that a single gene can have. To that end, we determined a cutoff for maximal variance of each gene of a mode, indicative of how much a gene can contribute on its own (Figure 6A, B). Default cutoff value was chosen to maximize the number of modes detected as informative while keeping a high average number of genes significantly contributing to each mode (Figure 7).

[0119] **Pagoda2 and AUCell to compare pathway activity scoring with MAYA.** Pagoda2 was run with default settings, following the vignette. AUCell was run using default settings, with log-normalized counts as input. Pagoda2 and AUCell were provided the same pathway lists as MAYA.

[0120] **Cell-ID and SCINA to compare cell type prediction with MAYA.** The two tools were provided the same PanglaoDB cell type marker lists as MAYA. Cell-ID was applied on a Seurat object following standard procedure, computing MCA and then performing hypergeometric test with gene lists. Each cell was attributed the cell type for which –log10(p-value) was the highest. When the value was inferior to 2, the cell was labeled unassigned. SCINA was run with default parameters, except for the sensitivity cutoff that was lowered to 0.9 and rm_overlap=FALSE as the input gene lists could be partially redundant.

[0121] **Integration with harmony.** Harmony was run through Seurat v3 with default settings.

[0122] **Integration using Canonical Correlation Analysis.** CCA was performed using Seurat and more specifically the functions FindIntegrationAnchors followed by

IntegrateData. This function aims at identifying directions that account for the most co-variance between different datasets and is another popular method to perform batch effect correction.

[0123] **GSEA.** Fold change is computed using Seurat Function "FoldChange" with default parameters and comparing the population of interest with all other cells from the dataset. The package fgsea was used to compute the adjusted p-value of the enrichment of Hallmark pathway lists in differentially expressed genes, setting a threshold at 0.01 for significance.

[0124] **Non-negative Matrix Factorization.** The inventors followed the method described by Gavish et al. (Gavish, A. et al. The transcriptional hallmarks of intra-tumor heterogeneity across a thousand tumors. bioRxiv 2021.12.19.473368, https://doi.org/10.1101/2021.12.19.473368 (2021)) to first identify the intra-tumoral heterogeneity programs in each patient individually using NMF on the subset of tumor cells from each patient, with different ranks. Only robust programs identified with different ranks were kept. The inventors then merged into 'meta-programs' the robust individual programs that share the most similarity across patients based on their top50 genes, by taking the union of their contributing genes.

[0125] **Metrics**

[0126] **Statistics**. A Wilcoxon rank sum test is systematically used to compare the distributions of two groups of data points and the inventors provide a two-sided p-value to assess the significance of the test.

[0127] **Shannon Diversity Index.** It measures in each predefined cluster the diversity of cells in terms of patient identity, batch or cell type. Here the inventors use it to measure the diversity of patients found in each Leiden cluster computed on the activity matrix.

$$SDI_c = \frac{(-1) * \sum_{i=1}^{N} p_i * \log(p_i)}{\log(N)} \qquad (1)$$

[0128] With c the cluster in which the inventors compute the SDI, N the number of different possible identities (patients in our case) and pi is the proportion of cells from the

cluster corresponding to identity i. SDI of 1 indicates that cells constituting the cluster come equally from all possible identities i.e., the cluster displays high identity diversity.

[0129] **Specificity metric.** For a mode, the inventors can compute for each predetermined cluster of cells (cells grouped by cell type in our case) a specificity score. As the sum of scores across clusters for a mode equals 1, the maximum value of specificity across cells reflects the repartition of high activity scores between clusters.

$$s_{m,c} = \frac{a_{m,c}^2}{\sum_{p=1}^{N} a_{m,p}^2} \qquad (2)$$

$$\sum_{p=1}^{N} s_{m,p}^2 = 1 \qquad (3)$$

with $s_{m,c}$ the specificity of mode m in cluster c, $a_{m,c}$ the average activity score of m in c, and N the number of clusters.

[0130] The inventors consider that specificity is significant for a cluster when it is 50% above expected value of 1/N (specificity score when all cells across all clusters have the same activity).

[0131] **Precision, recall, F1-score**

$$Precision = \frac{TP}{TP + FP} \qquad (4)$$

$$Recall = \frac{TP}{TP + FN} \qquad (5)$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \qquad (6)$$

[0132] where TP is the number of true positives, FP the number of false positives and FN the number of false negatives. F1-score of 1 means perfect precision and recall.

[0133] **Matching PanglaoDB cell types with author annotation for precision and recall assessment**

[0134] To assess precision and recall of cell-type annotation tools, the inventors had to find equivalents of cell types described by authors in the PanglaoDB and chose the closest type or multiple types when PanglaoDB included several subtypes.

[0135] **Kidney.** Monocytes=c("Monocytes"), Endothelial cells=c("Endothelial cells"), Mesangial_cells=c("Mesangial cells","Smooth muscle cells"), Podocytes=c("Podocytes"), TCD8 = c("T cells","T memory cells", "T cytotoxic cells").

[0136] **Colon.** 'Mature Enterocytes'=c("Enterocytes"), 'Goblet cells'=c("Goblet cells"), Pericytes=c("Pericytes"), 'Smooth muscle cells'=c("Smooth muscle cells"), cDC=c("Dendritic cells"), Proliferating monocytes= c("Monocytes","Macrophages"), 'NK cells'=c("NK cells","Natural killer T cells"), 'Regulatory T cells'=c("T regulatory cells","T cells","T memory cells"), 'CD19 + CD20 + B' = c("B cells","B cells naive","B cells memory"), 'Mast cells'=c("Mast cells").

[0137] **Performances.** All tests were run with CPU: 6 cores/12 threads@ 2.6 GHz.

[0138] **Data availability**

[0139] Kidney dataset: The count matrices were downloaded from Supplementary data S1 from Young et al. [https://www.science.org/doi/suppl/10.1126/science.aat1699/suppl_file/aat1699_datas1.gz.zip].

[0140] Colon dataset: Raw count matrix and cell annotations were downloaded from the NCBI Gene Expression Omnibus (GEO) database under the accession code GSE144735 for the KUL3 cohort.

[0141] Ovary dataset: Count data were downloaded from the NCBI Gene Expression Omnibus (GEO) database with accession code GSE165897.

[0142] Larynx dataset: Count data were downloaded from the NCBI Gene Expression Omnibus (GEO) database with accession code GSE150321.

[0143] Pancreas dataset: A Seurat object was directly loaded from the package SeuratData [https://github.com/satijalab/seurat-data] (panc8 v3.0.2). Breast dataset: Count matrix and metadata the dataset from Qian et al were retrieved from https://www.weizmann.ac.il/sites/3CA/breast as formatted by Gavish et al [https://www.dropbox.com/sh/nbx7v3om85wkfoq/AACpeZEZ4RNQwMW37Q7AHxExa?dl=1]. Lung dataset: Count matrix and metadata the dataset from Kim et al were retrieved from https://www.weizmann.ac.il/sites/3CA/lung as formatted by Gavish et al [https://www.dropbox.com/sh/byext689ffg77pj/AACp5jI2RRxndurKn2B0T-VWa?dl=1]. PBMC dataset: A Seurat object was directly loaded from the package

SeuratData [https://github.com/satijalab/seurat-data] (pbmcsca v3.0.0). Reference databases: PanglaoDB was downloaded from the website (https://panglaodb.se/) [https://panglaodb.se/markers/PanglaoDB_markers_27_Mar_2020.tsv.gz]. MSigDB gene lists (Hallmark, KEGG and REACTOME) were downloaded from the Broad Institute website (http://www.gsea-msigdb.org/gsea/msigdb/collections.jsp) [https://data.broadinstitute.org/gsea-msigdb/msigdb/release/7.4/h.all.v7.4.symbols.gmt, https://data.broadinstitute.org/gseamsigdb/msigdb/release/7.4/c2.cp.kegg.v7.4.entrez.gmt,https://data.broadinstitute.org/gseamsigdb/msigdb/release/7.4/c2.cp.reactome.v7.4.symbols.gmt] in their version 7.4. Source data are provided with this paper.

# CLAIMS

**1.** A computer-implemented method for determining and quantifying modes of activation of biological pathways in individual cells, comprising:

- receiving a) sequencing data obtained by a single-cell RNA sequencing method, said sequencing data being characterized by a set of genes and a set of individual cells, and b) a plurality of gene lists, each gene list corresponding to a biological pathway or a gene signature of interest,

- determining a gene-cell expression matrix based on said sequencing data and said plurality of gene lists, said gene-expression matrix comprising rows corresponding to genes of said set of genes referred to as pathway genes and belonging to at least one among said gene lists,

- carrying out a principal component analysis (PCA) on said gene-cell expression matrix, so as to determine a plurality of modes of activation of at least one among said biological pathways, each mode of activation corresponding to one among the principal components obtained from the PCA,

- selecting a subset of so-called effective modes of activation among said plurality of modes of activation,

- determining an activity matrix comprising scores quantifying a level of activity of each effective mode of activation among said subset of effective modes of activation within each individual cell among said set of individual cells.

**2.** The method according to claim **1**, wherein each mode of activation among said plurality of modes of activation is defined by a specific subset of genes among said pathway genes.

**3.** The method according to any one of claims **1** or **2**, wherein selecting said subset of modes of activation comprises: detecting bimodal distributions of scores in said activity matrix and verifying that the number of active cells is superior to a predefined minimum fraction of the set of individual cells.

**4.** The method according to claims **2** and **3**, wherein selecting said subset of modes of activation further comprises selecting modes of activation for which said specific subset of genes comprises a number of genes higher than a predefined threshold.

**5.** The method according to any one of claims **1** to **4**, wherein said biological pathways comprise at least one hallmark pathway.

**6.** The method according to any one of claims **1** to **5**, wherein the gene dataset is a curated gene dataset.

**7.** The method according to any one of claims **1** to **6**, wherein the method allows to detect subgroups of genes within biological pathways.

**8.** The method according to any one of claims **1** to **7**, wherein the method allows to identify common modes of activation of biological pathways in tumor cells shared across patients.

**9.** The method according to any one of claims **1** to **8**, wherein the method allows to identify common modes of activation of biological pathways shared across tumor cells.

**10.** The method according to any one of claims **1** to **9**, wherein the method allows to identify modes of activation of biological pathways specific of cell types.

**11.** The method according to any one of claims **1** to **10**, wherein the method allows to identify modes of activation of biological pathways specific to the tumor microenvironment.

**12.** The method according to any one of claims **1** to **11**, wherein the method allows to identify cell types.

**13.** The method according to any one of claims **1** to **12**, wherein the method allows to identify rare cell types.

**14.** The method according to claim **4**, wherein the method allows to identify tumor cells.

**15.** The method according to any one of claims **1** to **14**, wherein the method allows to identify cell function.

**16.** The method according to any one of claims **1** to **15**, wherein the method allows to detect relevant biological signal from noisy gene list.

**17.** The method according to any one of claims **1** to **16**, wherein the method allows to detect relevant biological signal independently of batch effect.

**18.** A device for obtaining and quantifying modes of activation of biological pathways in individual cells to determine therapeutic targets for cancer therapy, comprising:
- at least one input interface configured to receive a) sequencing data obtained by a single-cell RNA sequencing method, said sequencing data being characterized by a set of genes and a set of individual cells, and b) a plurality of gene lists, each gene list corresponding to a biological pathway,
- at least one processor configured to:
    - determine a gene-cell expression matrix based on said sequencing data and said plurality of gene lists, said gene-expression matrix comprising rows corresponding to genes of said set of genes referred to as pathway genes and belonging to at least one among said gene lists,
    - carry out a principal component analysis (PCA) on said gene-expression matrix, so as to obtain a plurality of modes of activation of at least one among said biological pathways, each mode of activation corresponding to one principal component obtained from the PCA,
    - selecting a subset of modes of activation among said plurality of modes of activation,
    - determining an activity matrix comprising scores quantifying a level of activity of each mode of activation among said subset of modes of activation within each individual cell among said set of individual cells,
- at least one output interface configured to output said activity matrix.

**19.** A non-transitory computer-readable storage medium comprising instructions which, when executed by a computer, cause the computer to carry out the method according to any one of claims **1** to **17**.

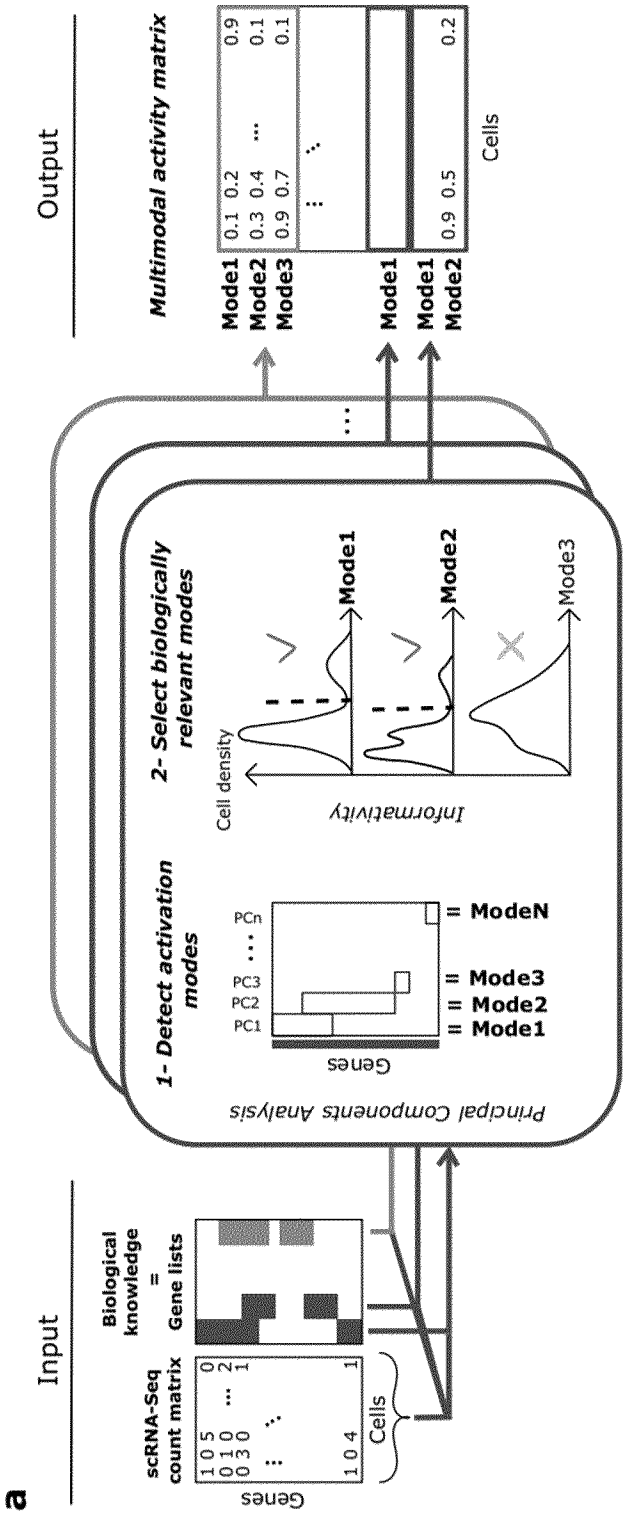

a
Input
Output
scRNA-Seq count matrix
Genes
Cells
Biological knowledge = Gene lists
Principal Components Analysis
1- Detect activation modes
PC1
PC2
PC3
PCn
Genes
= Mode1
= Mode2
= Mode3
= ModeN
2- Select biologically relevant modes
Cell density
Informativity
Mode1
Mode2
Mode3
Multimodal activity matrix
Mode1
Mode2
Mode3
Cells

FIG. 1A

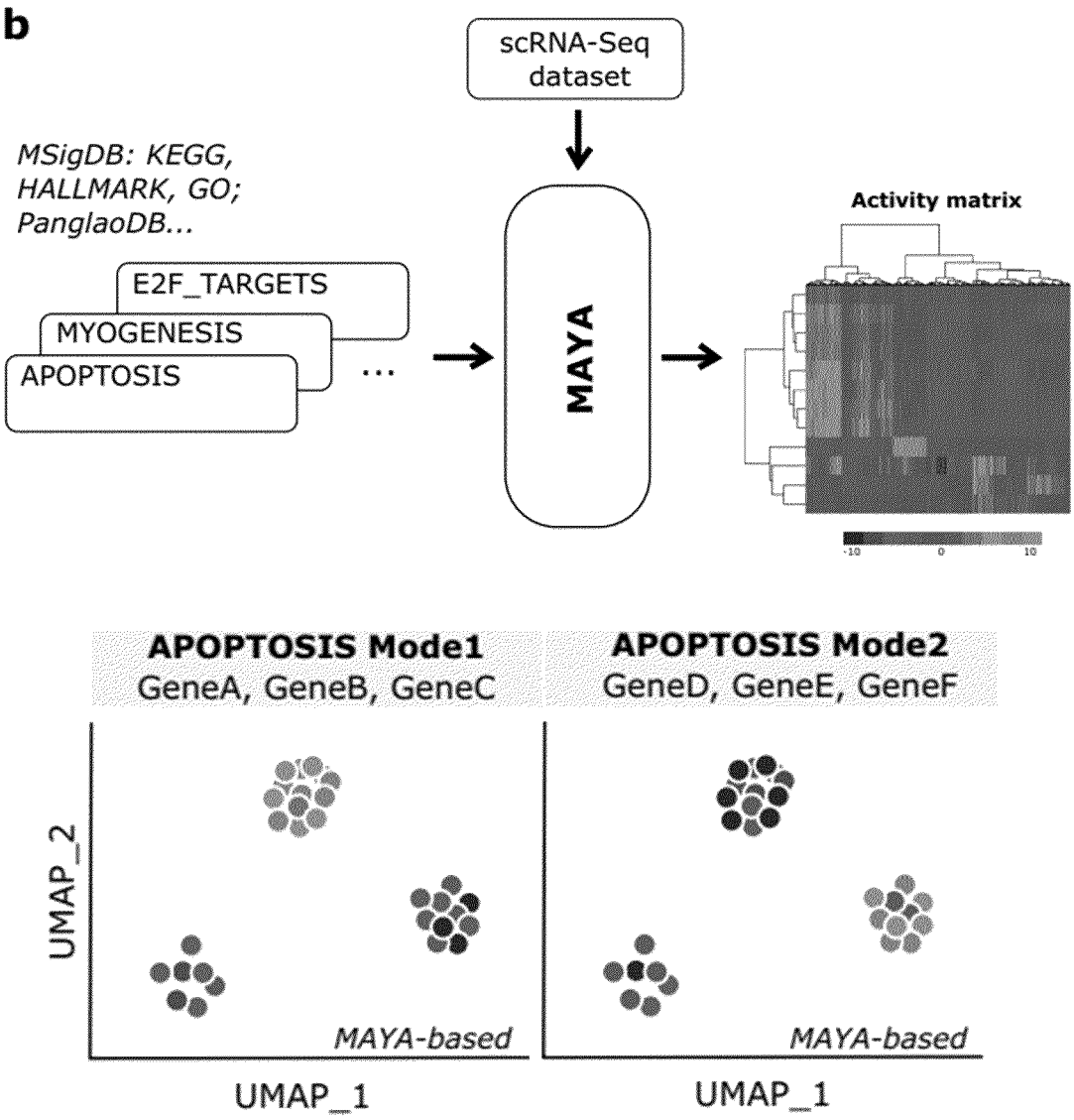

b
scRNA-Seq dataset
MSigDB: KEGG, HALLMARK, GO; PanglaoDB...
E2F_TARGETS
MYOGENESIS
APOPTOSIS
...
MAYA
Activity matrix
APOPTOSIS Mode1
GeneA, GeneB, GeneC
APOPTOSIS Mode2
GeneD, GeneE, GeneF
UMAP_2
MAYA-based
UMAP_1
MAYA-based
UMAP_1

FIG. 1B

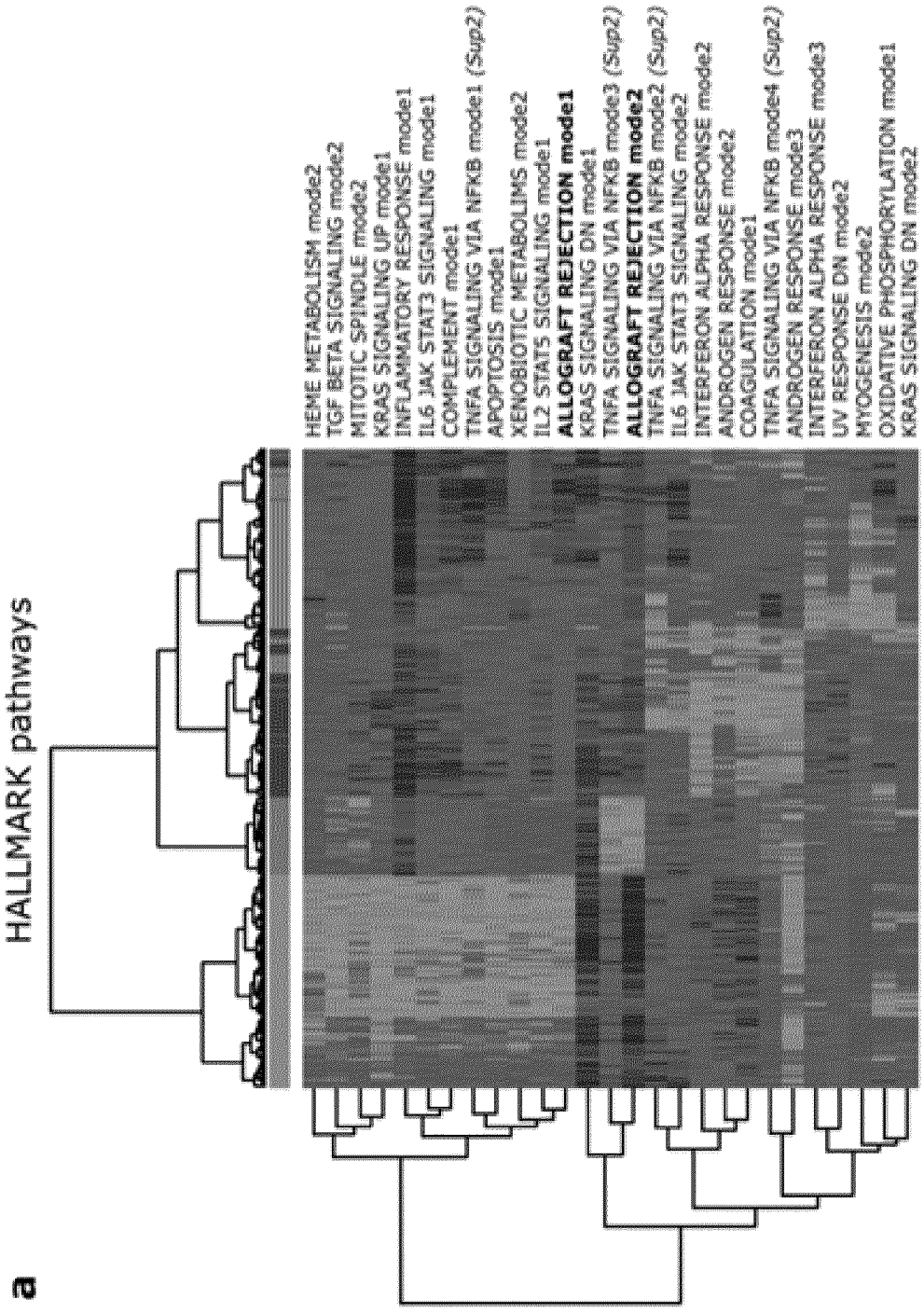
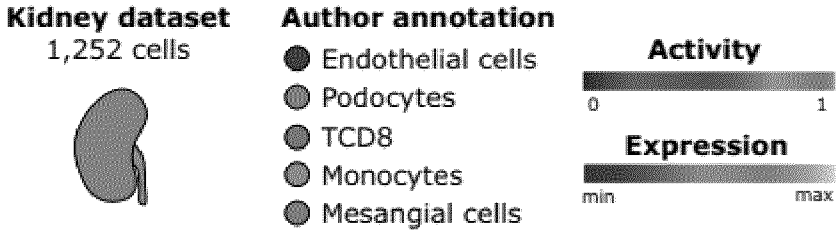

a
HALLMARK pathways
HEME METABOLISM mode2
TGF BETA SIGNALING mode2
MITOTIC SPINDLE mode2
KRAS SIGNALING UP mode1
INFLAMMATORY RESPONSE mode1
IL6 JAK STAT3 SIGNALING mode1
COMPLEMENT mode1
TNFA SIGNALING VIA NFKB mode1 (Sup2)
APOPTOSIS mode1
XENOBIOTIC METABOLISM mode2
IL2 STAT5 SIGNALING mode1
ALLOGRAFT REJECTION mode1
KRAS SIGNALING DN mode1
TNFA SIGNALING VIA NFKB mode3 (Sup2)
ALLOGRAFT REJECTION mode2
TNFA SIGNALING VIA NFKB mode2 (Sup2)
IL6 JAK STAT3 SIGNALING mode2
INTERFERON ALPHA RESPONSE mode2
ANDROGEN RESPONSE mode2
COAGULATION mode1
TNFA SIGNALING VIA NFKB mode4 (Sup2)
ANDROGEN RESPONSE mode3
INTERFERON ALPHA RESPONSE mode3
UV RESPONSE DN mode2
MYOGENESIS mode2
OXIDATIVE PHOSPHORYLATION mode1
KRAS SIGNALING DN mode2

FIG. 2A


Kidney dataset
1,252 cells
Author annotation
Endothelial cells
Podocytes
TCD8
Monocytes
Mesangial cells
Activity
0
1
Expression
min
max

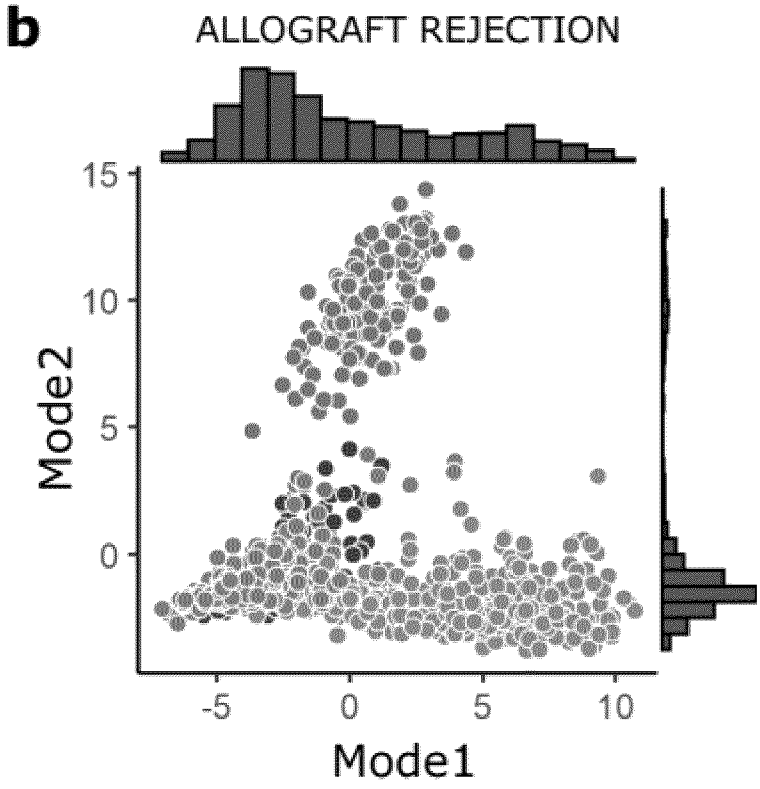

b
ALLOGRAFT REJECTION
Mode2
15
10
5
0
-5
0
5
10
Mode1

FIG. 2B

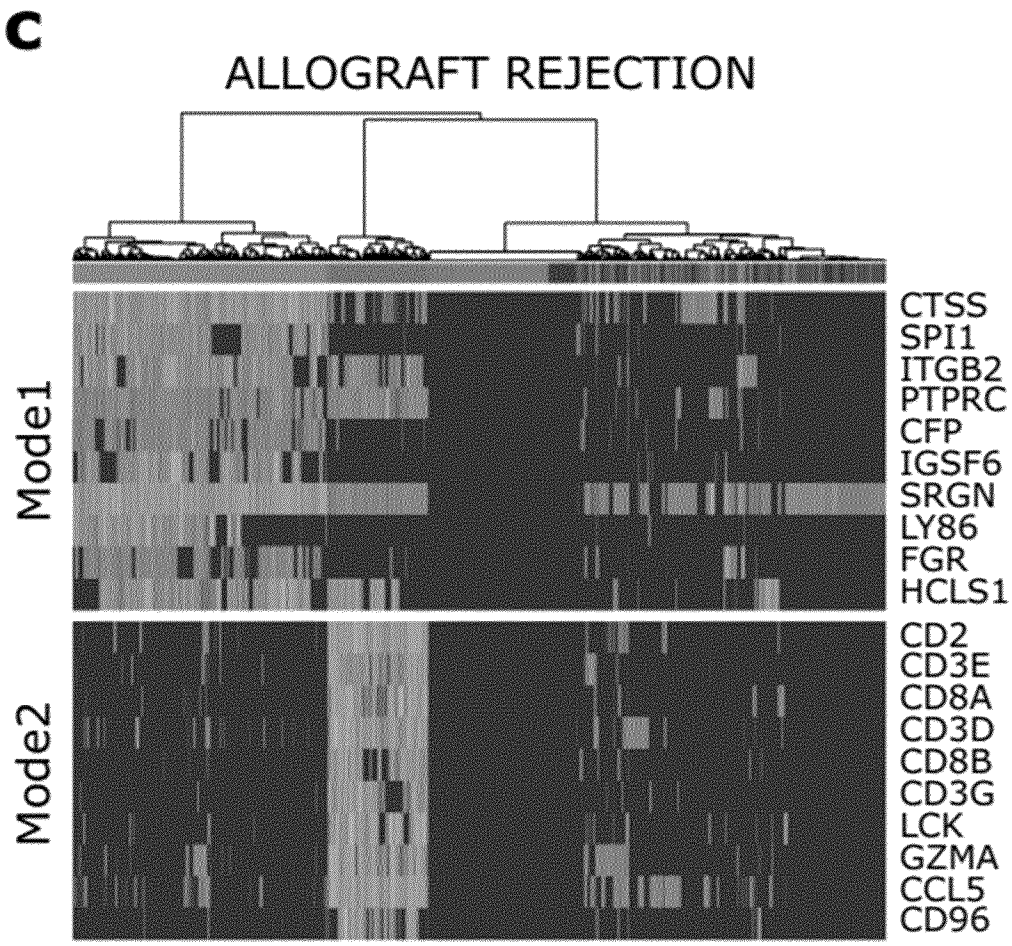

c
ALLOGRAFT REJECTION
Mode1
CTSS
SPI1
ITGB2
PTPRC
CFP
IGSF6
SRGN
LY86
FGR
HCLS1
Mode2
CD2
CD3E
CD8A
CD3D
CD8B
CD3G
LCK
GZMA
CCL5
CD96

FIG. 2C

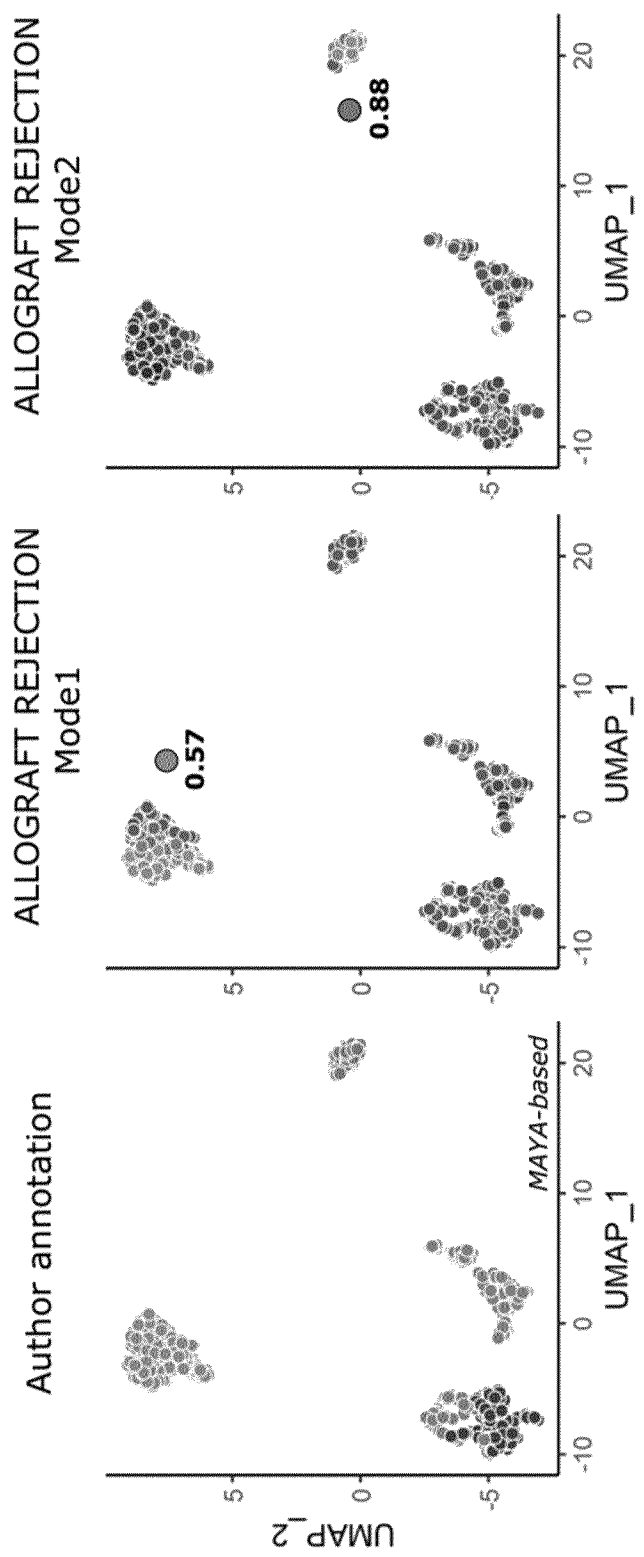

Author annotation
ALLOGRAFT REJECTION
Mode1
ALLOGRAFT REJECTION
Mode2
MAYA-based
0.57
0.88
UMAP_1
UMAP_2

FIG. 2D

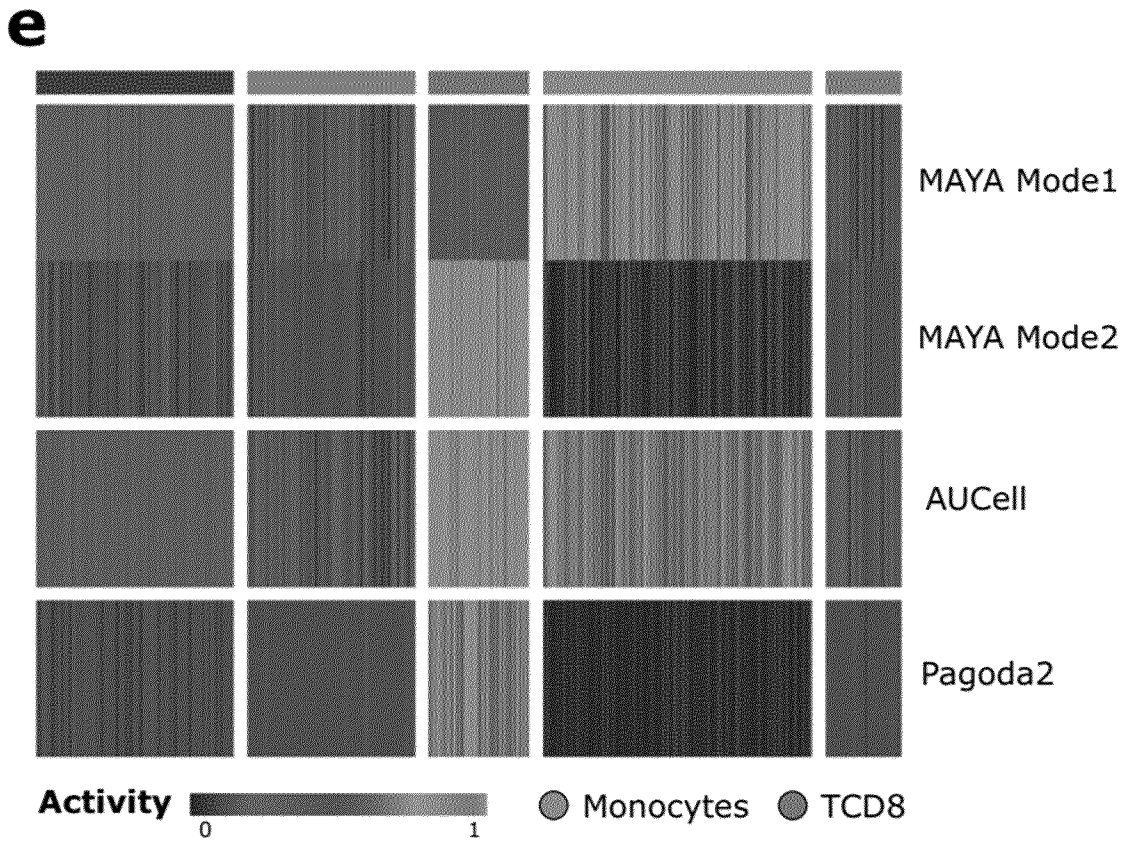

e
MAYA Mode1
MAYA Mode2
AUCell
Pagoda2
Activity
0
1
Monocytes
TCD8

FIG. 2E

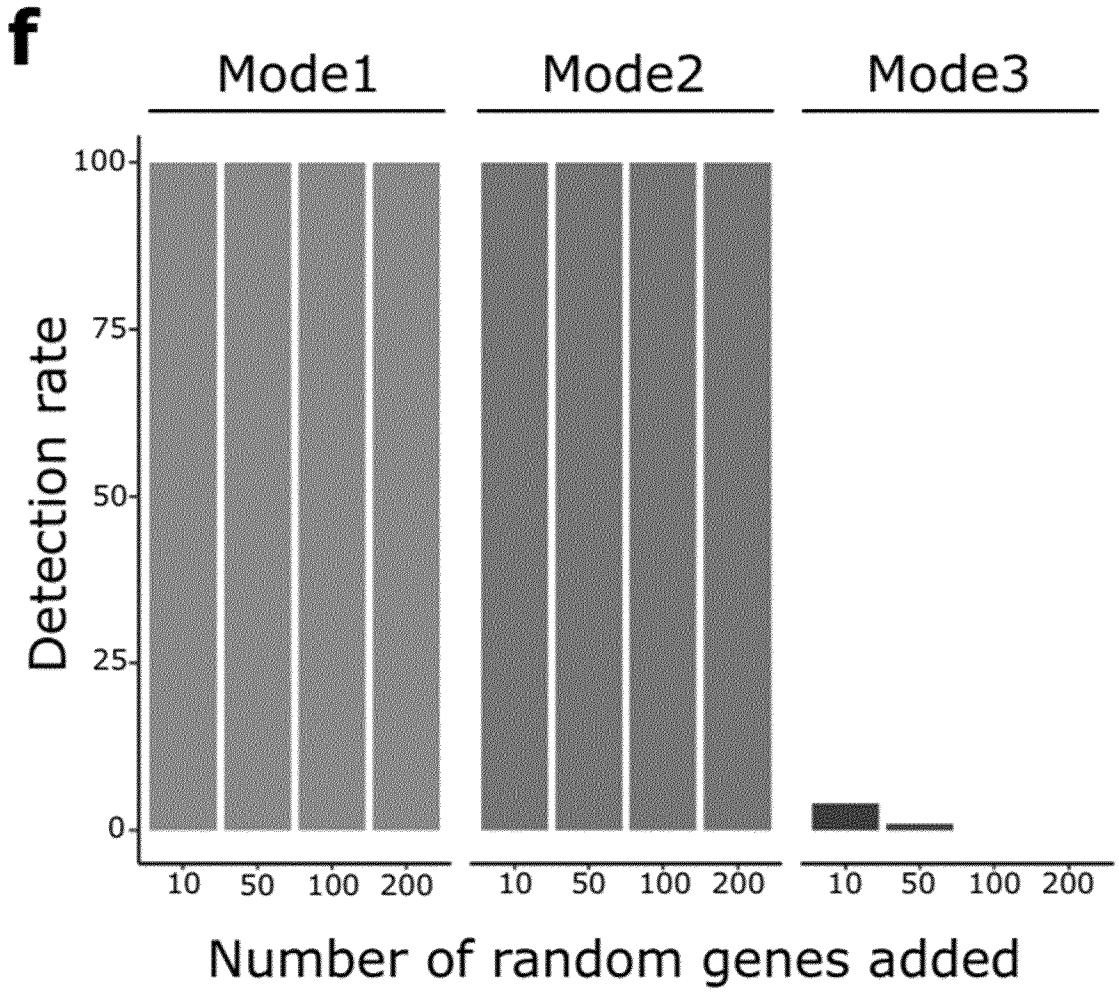

f
Mode1
Mode2
Mode3
Detection rate
100
75
50
25
0
10
50
100
200
Number of random genes added

FIG. 2F

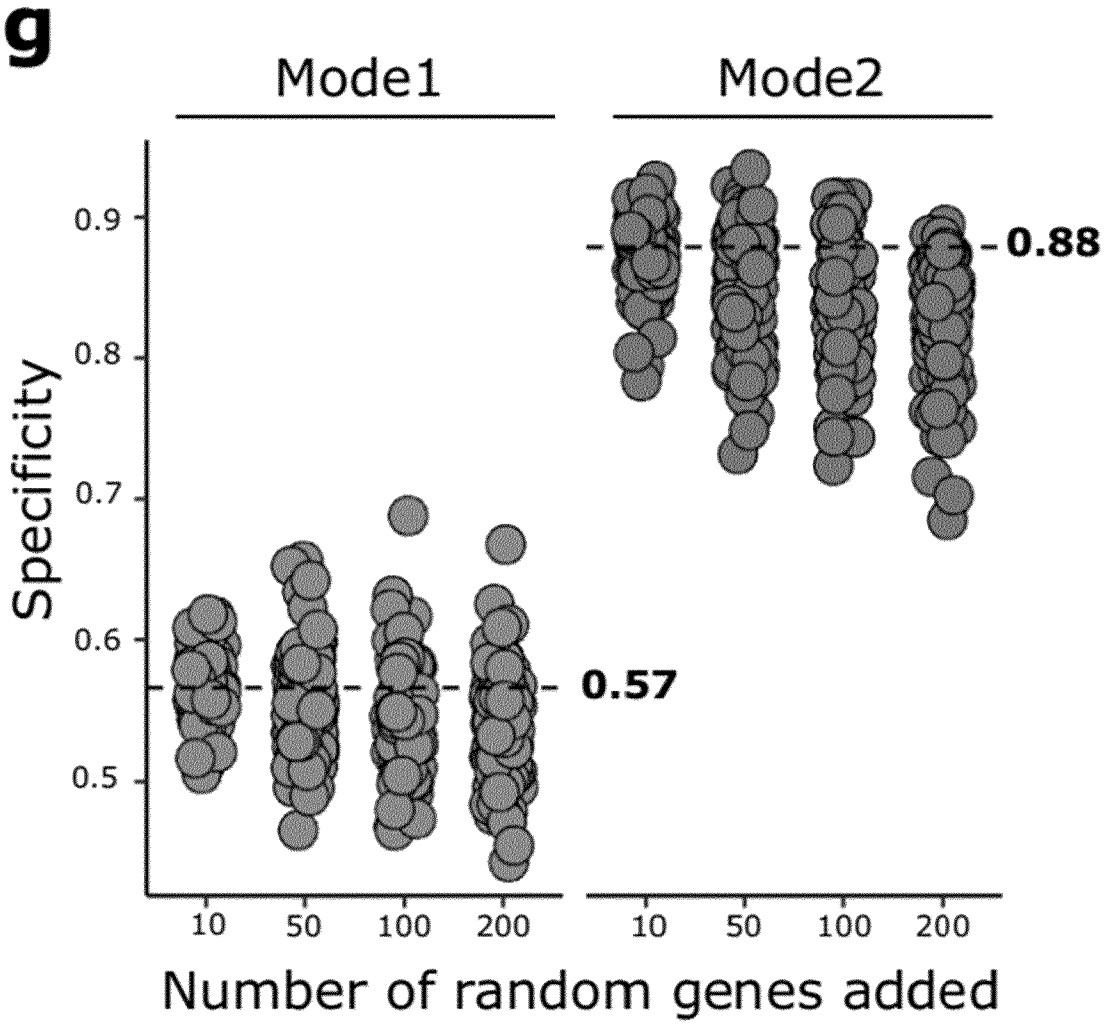

g
Mode1
Mode2
0.9
0.8
0.7
0.6
0.5
Specificity
0.88
0.57
10
50
100
200
10
50
100
200
Number of random genes added

FIG. 2G

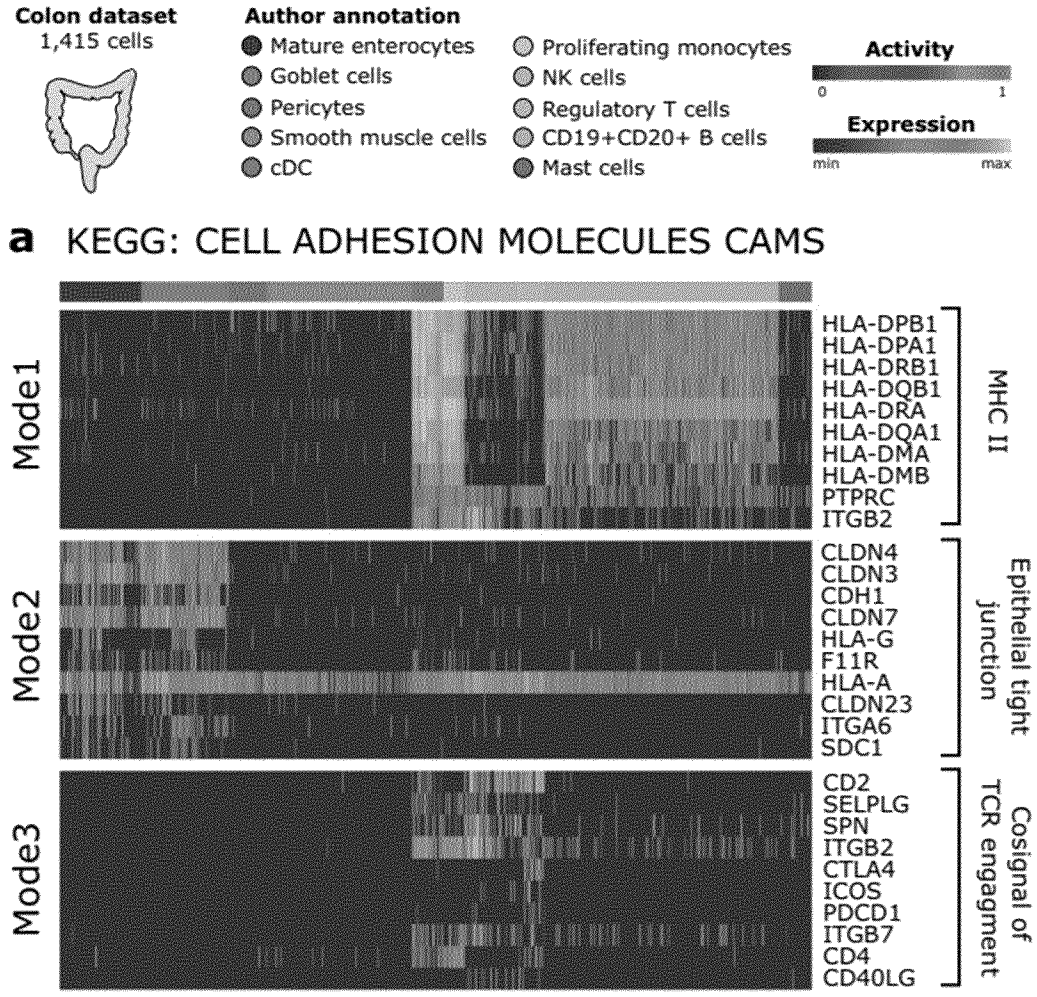

Colon dataset
1,415 cells
Author annotation
Mature enterocytes
Goblet cells
Pericytes
Smooth muscle cells
cDC
Proliferating monocytes
NK cells
Regulatory T cells
CD19+CD20+ B cells
Mast cells
Activity
0
1
Expression
min
max
a KEGG: CELL ADHESION MOLECULES CAMS
Mode1
Mode2
Mode3
HLA-DPB1
HLA-DPA1
HLA-DRB1
HLA-DQB1
HLA-DRA
HLA-DQA1
HLA-DMA
HLA-DMB
PTPRC
ITGB2
MHC II
CLDN4
CLDN3
CDH1
CLDN7
HLA-G
F11R
HLA-A
CLDN23
ITGA6
SDC1
Epithelial tight junction
CD2
SELPLG
SPN
ITGB2
CTLA4
ICOS
PDCD1
ITGB7
CD4
CD40LG
Cosignal of TCR engagment

FIG. 3A

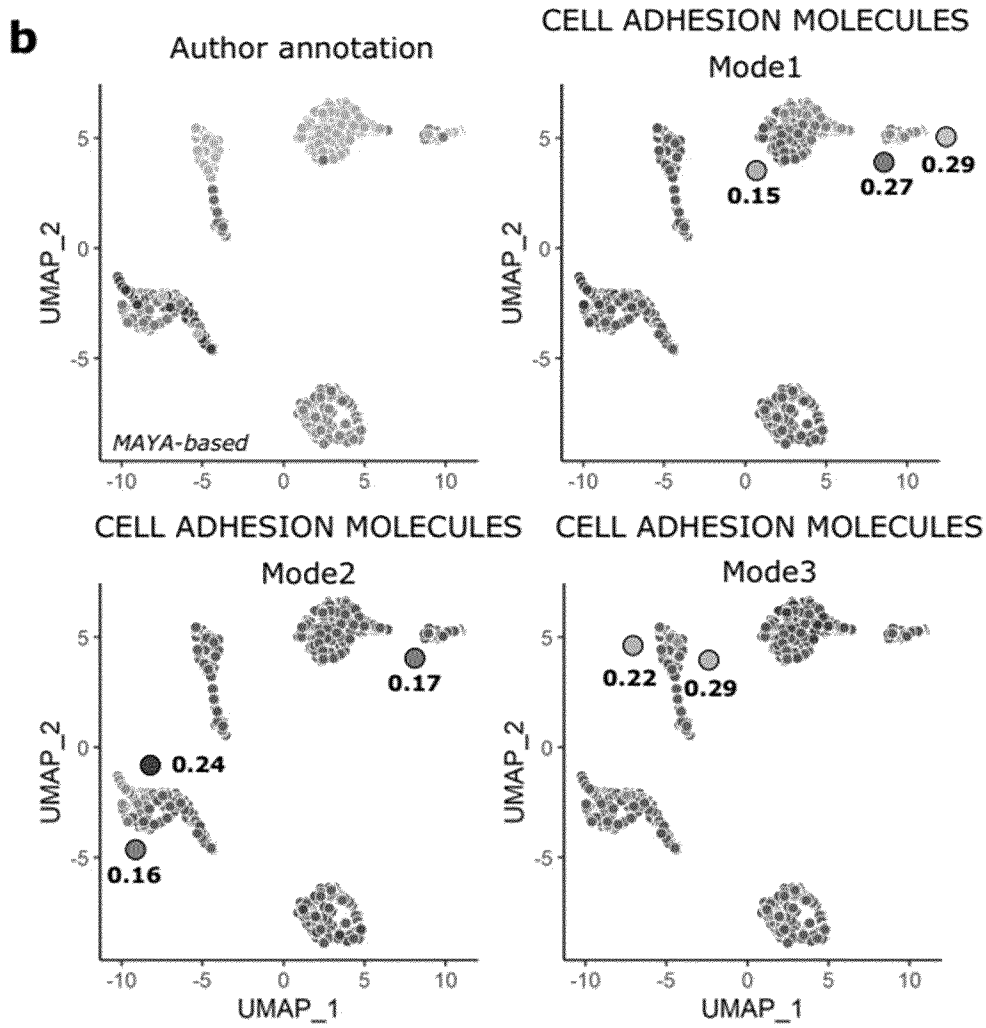

b
Author annotation
MAYA-based
CELL ADHESION MOLECULES
Mode1
0.15
0.27
0.29
CELL ADHESION MOLECULES
Mode2
0.24
0.16
0.17
CELL ADHESION MOLECULES
Mode3
0.22
0.29
UMAP_1
UMAP_2

FIG. 3B

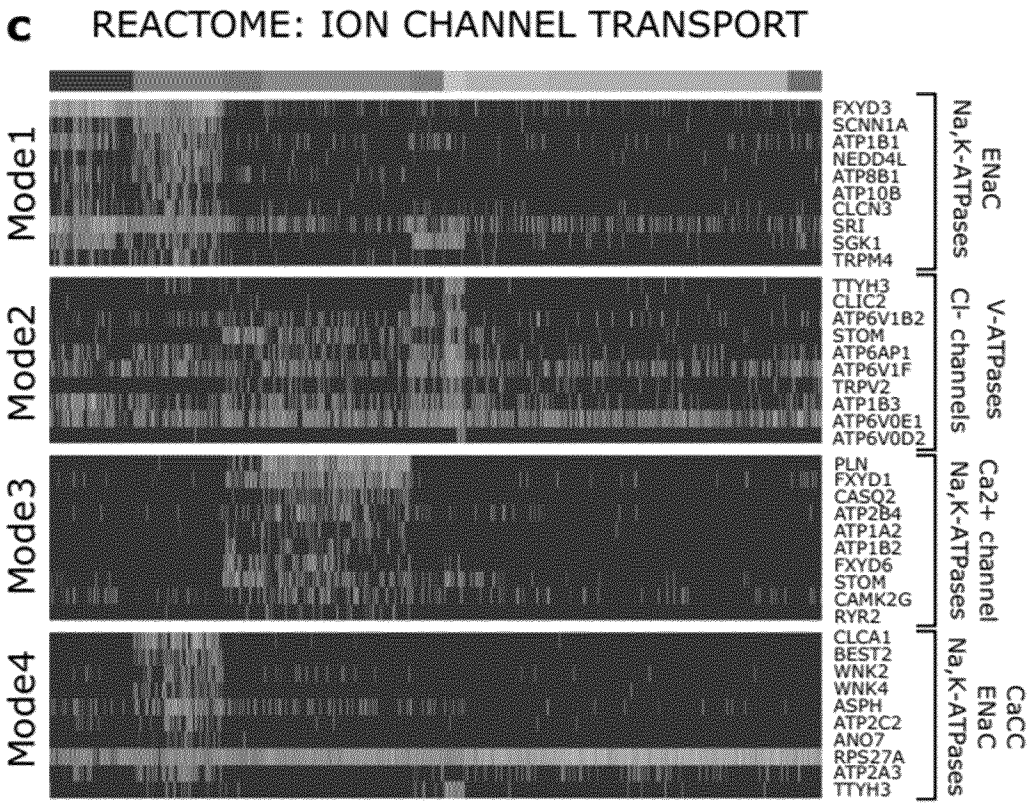

c
REACTOME: ION CHANNEL TRANSPORT
Mode1
Mode2
Mode3
Mode4
FXYD3
SCNN1A
ATP1B1
NEDD4L
ATP8B1
ATP10B
CLCN3
SRI
SGK1
TRPM4
TTYH3
CLIC2
ATP6V1B2
STOM
ATP6AP1
ATP6V1F
TRPV2
ATP1B3
ATP6V0E1
ATP6V0D2
PLN
FXYD1
CASQ2
ATP2B4
ATP1A2
ATP1B2
FXYD6
STOM
CAMK2G
RYR2
CLCA1
BEST2
WNK2
WNK4
ASPH
ATP2C2
ANO7
RPS27A
ATP2A3
TTYH3
ENaC
Na,K-ATPases
V-ATPases
Cl- channels
Ca2+ channel
Na,K-ATPases
CaCC
ENaC
Na,K-ATPases

FIG. 3C

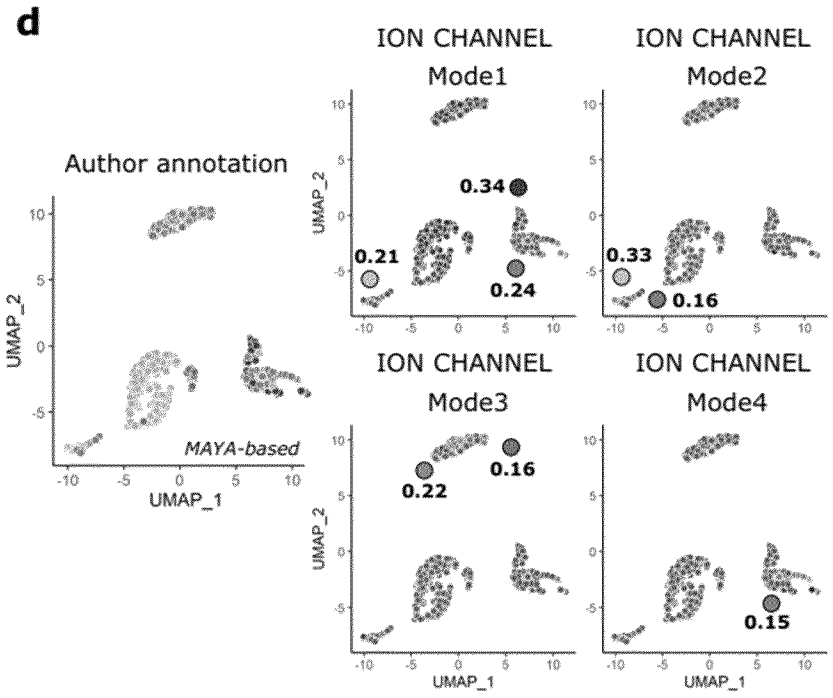

d
Author annotation
MAYA-based
UMAP_1
UMAP_2
ION CHANNEL
Mode1
0.34
0.21
0.24
ION CHANNEL
Mode2
0.33
0.16
ION CHANNEL
Mode3
0.22
0.16
ION CHANNEL
Mode4
0.15

FIG. 3D

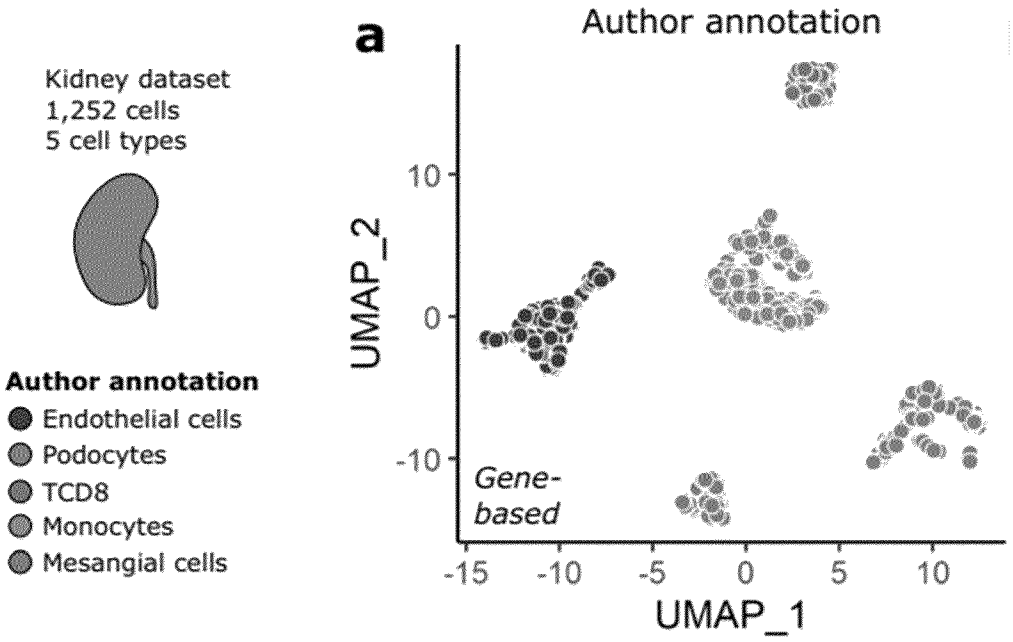

a
Author annotation
Kidney dataset
1,252 cells
5 cell types
Author annotation
Endothelial cells
Podocytes
TCD8
Monocytes
Mesangial cells
UMAP_2
UMAP_1
10
0
-10
-15
-10
-5
0
5
10
Gene-based

FIG. 4A

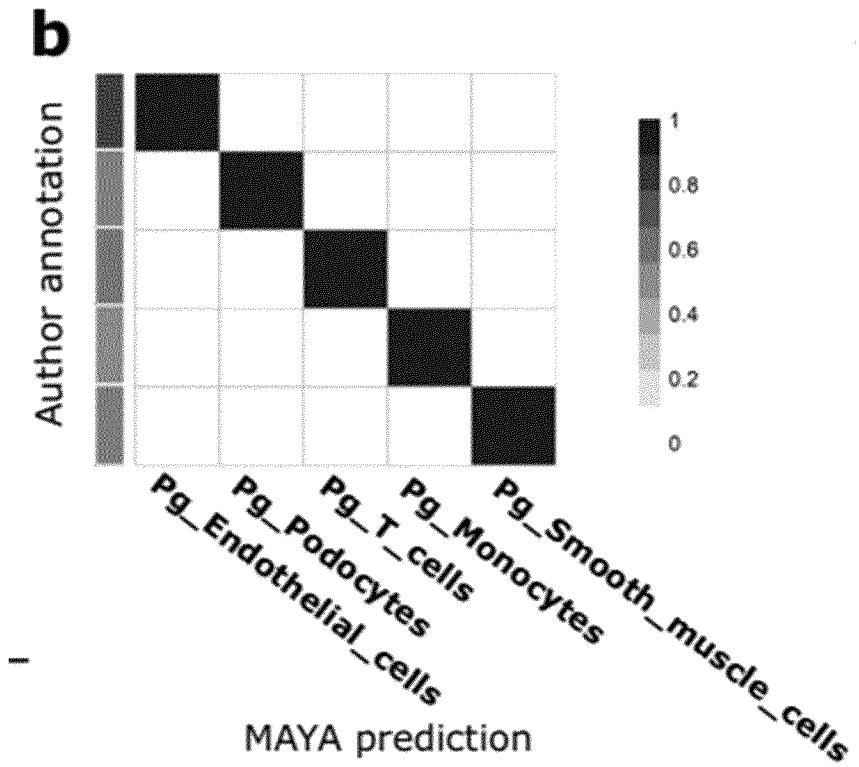

b
Author annotation
1
0.8
0.6
0.4
0.2
0
Pg_Endothelial_cells
Pg_Podocytes
Pg_T_cells
Pg_Monocytes
Pg_Smooth_muscle_cells
MAYA prediction

FIG. 4B

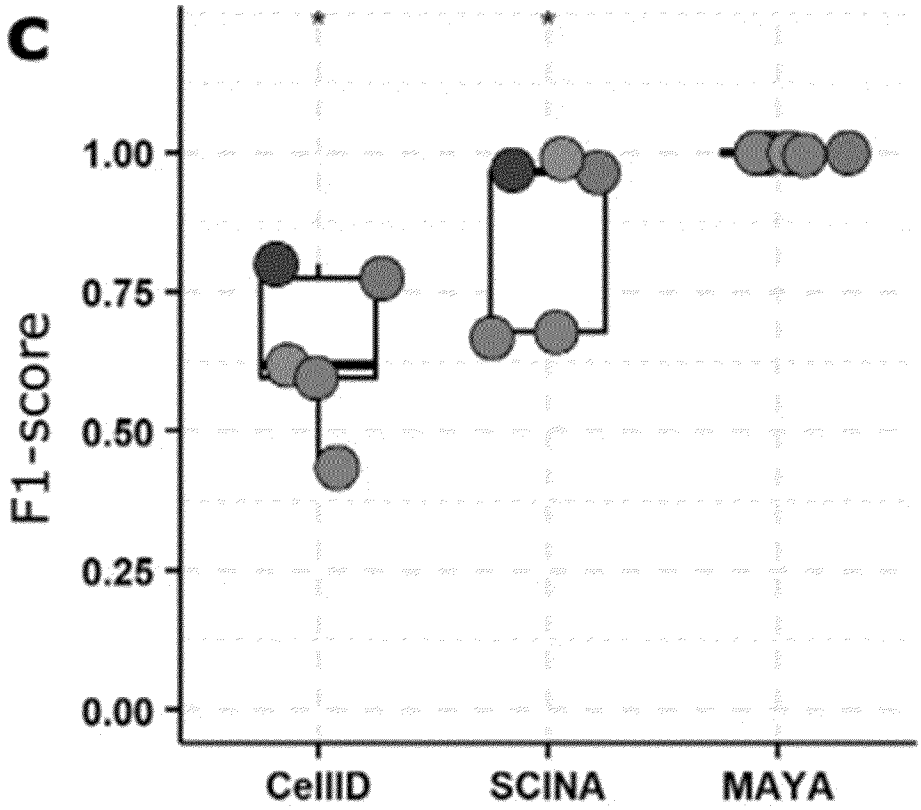



FIG. 4C

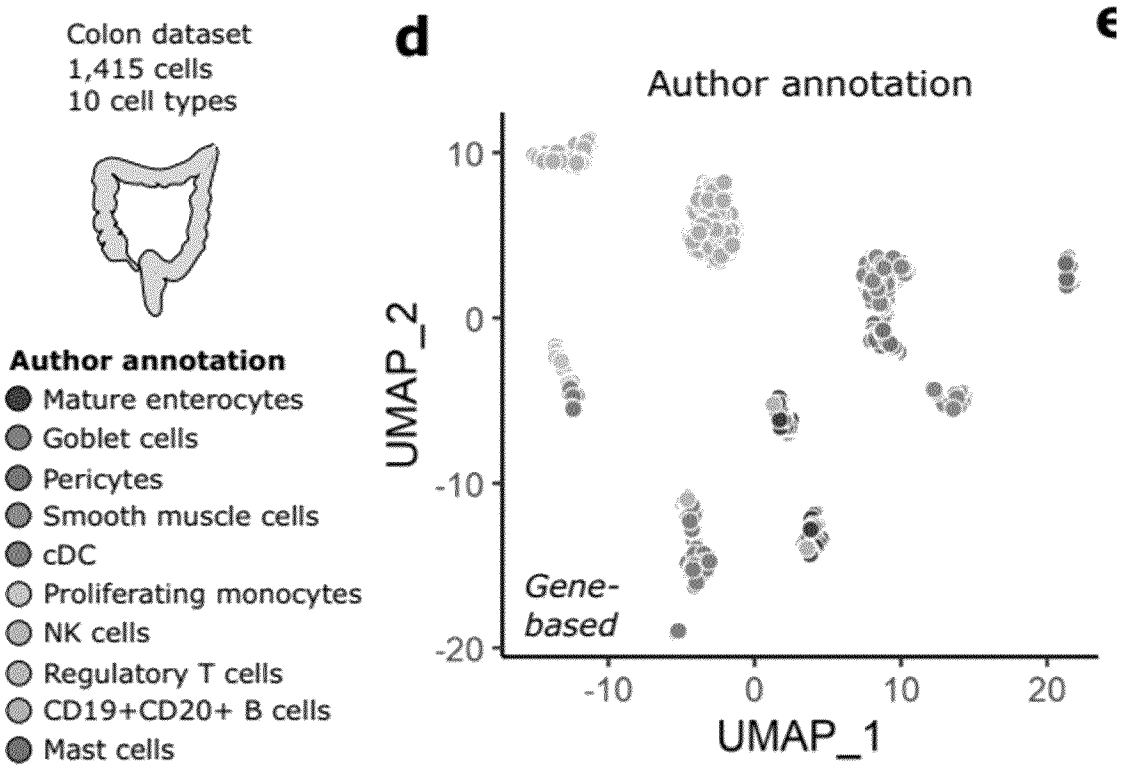

Colon dataset
1,415 cells
10 cell types
Author annotation
Mature enterocytes
Goblet cells
Pericytes
Smooth muscle cells
cDC
Proliferating monocytes
NK cells
Regulatory T cells
CD19+CD20+ B cells
Mast cells
d
Author annotation
UMAP_2
UMAP_1
10
0
-10
-20
-10
0
10
20
Gene-based

FIG. 4D

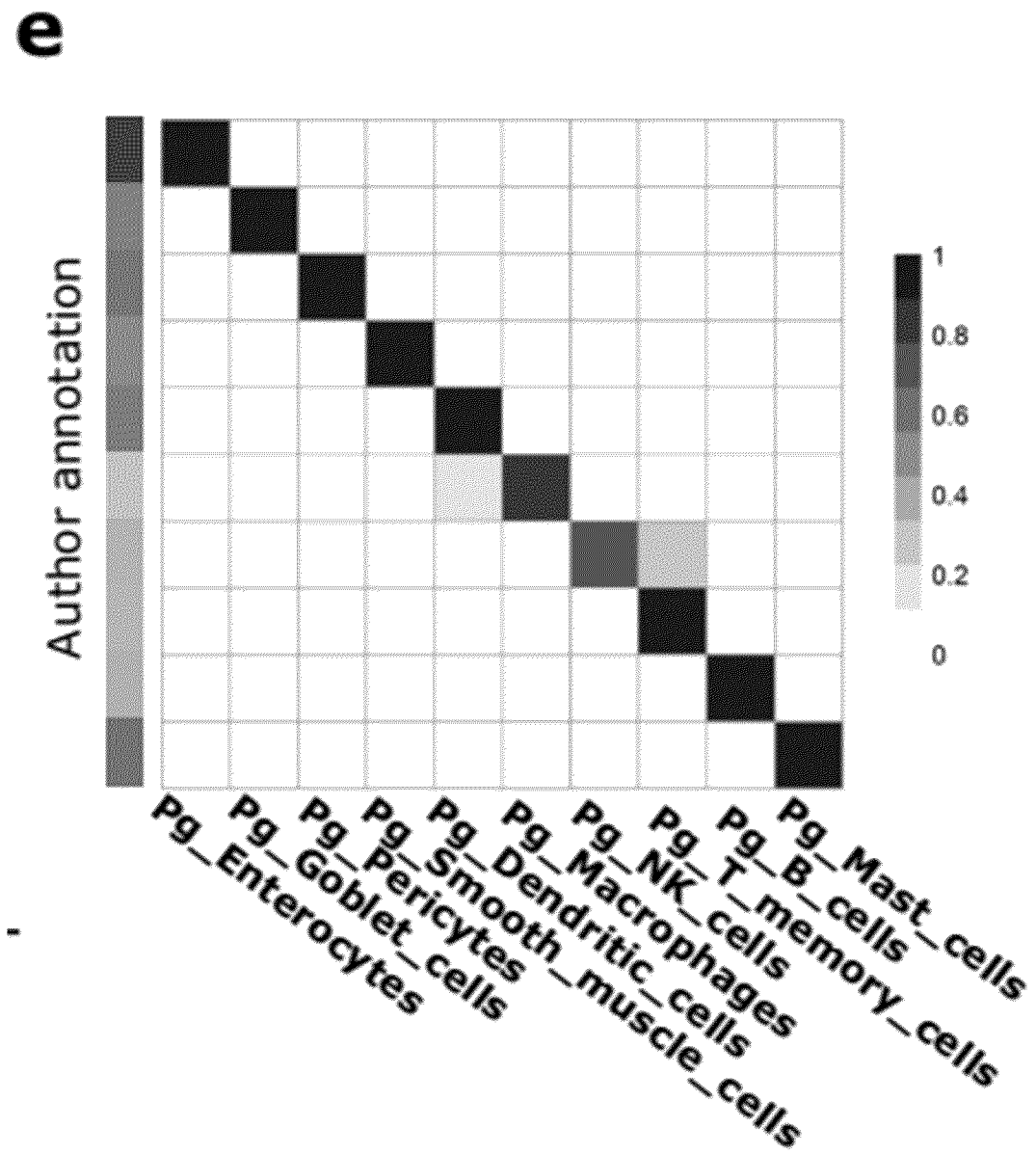

e
Author annotation
1
0.8
0.6
0.4
0.2
0
Pg_Enterocytes
Pg_Goblet_cells
Pg_Pericytes
Pg_Smooth_muscle_cells
Pg_Dendritic_cells
Pg_Macrophages
Pg_NK_cells
Pg_T_memory_cells
Pg_B_cells
Pg_Mast_cells

FIG. 4E

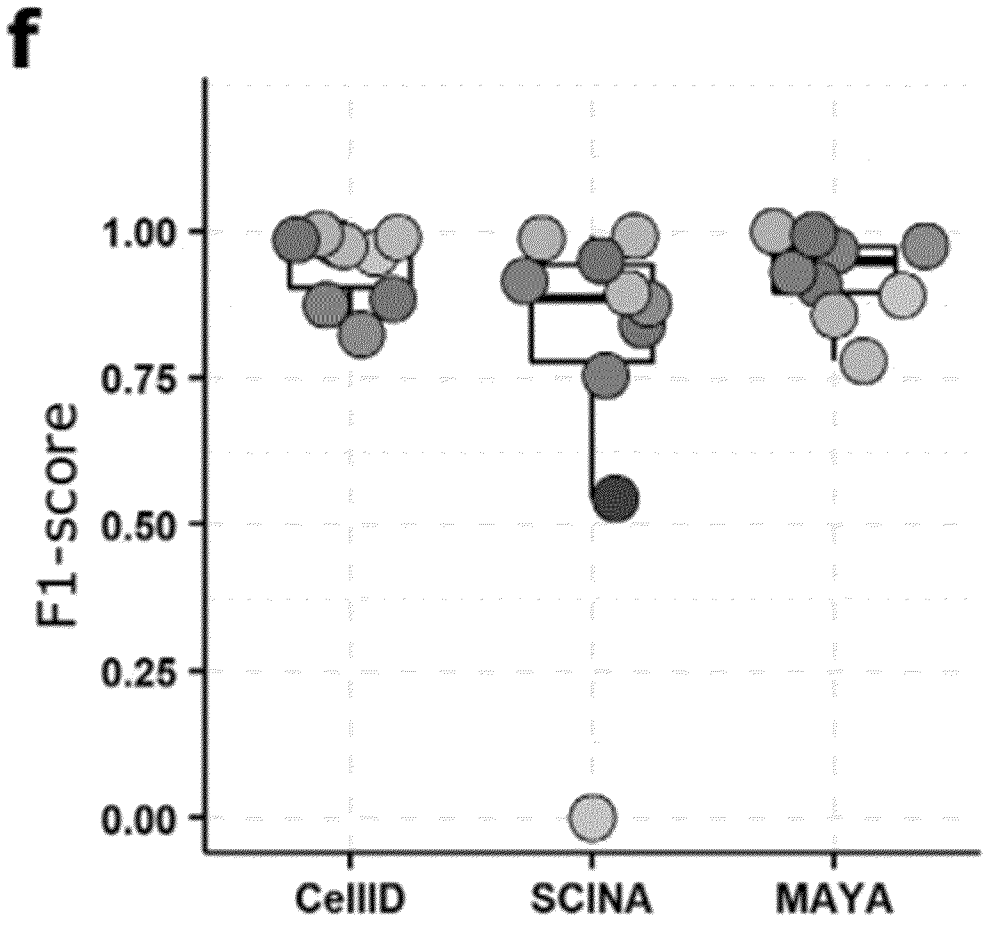



FIG. 4F

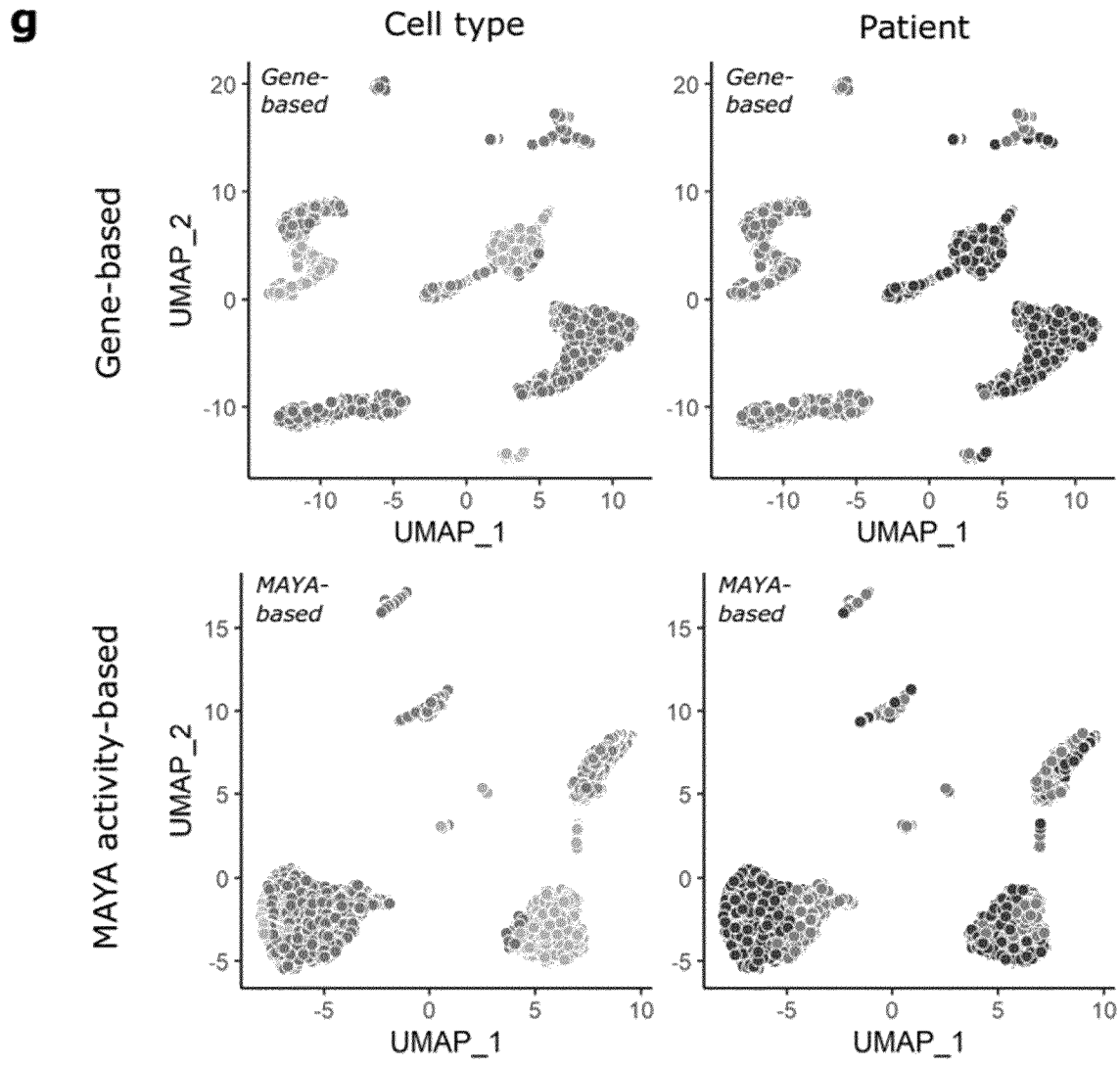
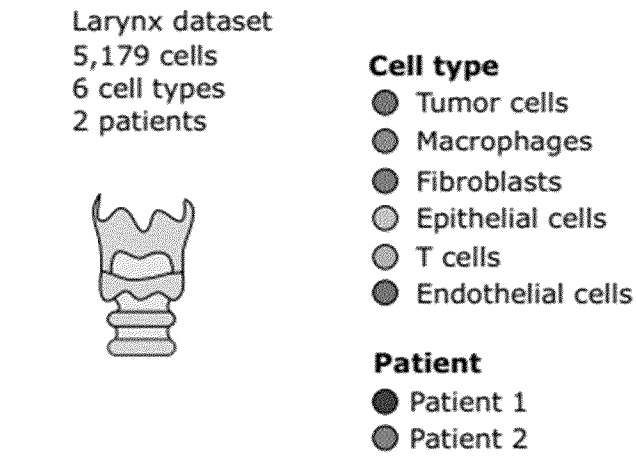

g
Cell type
Patient
Gene-based
MAYA activity-based
Gene-based
MAYA-based
UMAP_1
UMAP_2

FIG. 4G


Larynx dataset
5,179 cells
6 cell types
2 patients
Cell type
Tumor cells
Macrophages
Fibroblasts
Epithelial cells
T cells
Endothelial cells
Patient
Patient 1
Patient 2

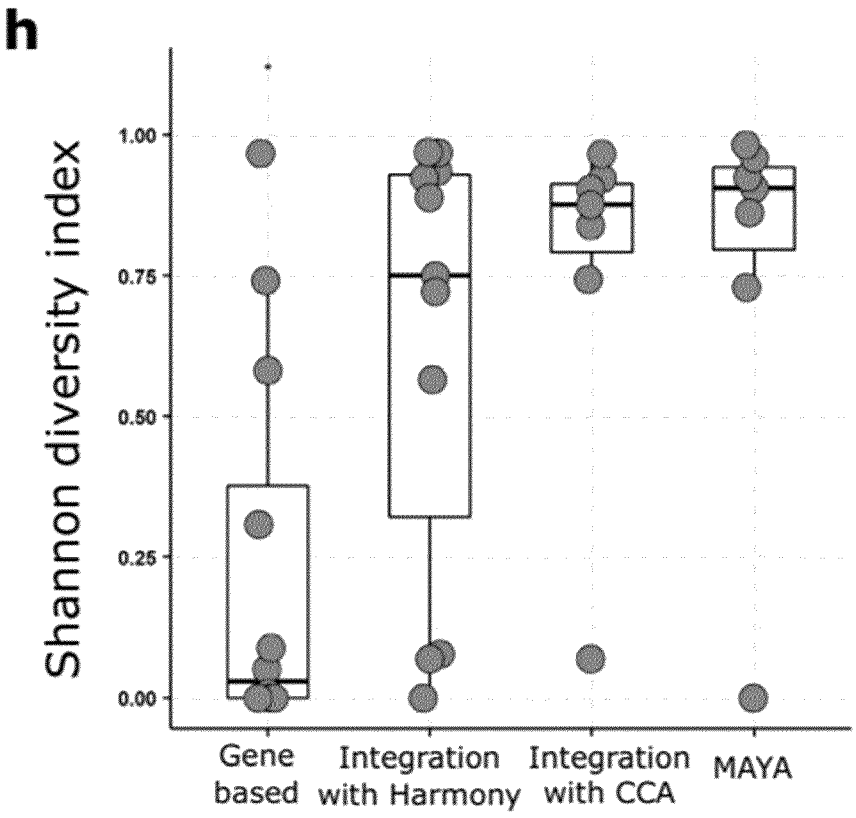

h
Shannon diversity index
1.00
0.75
0.50
0.25
0.00
Gene based
Integration with Harmony
Integration with CCA
MAYA

FIG. 4H

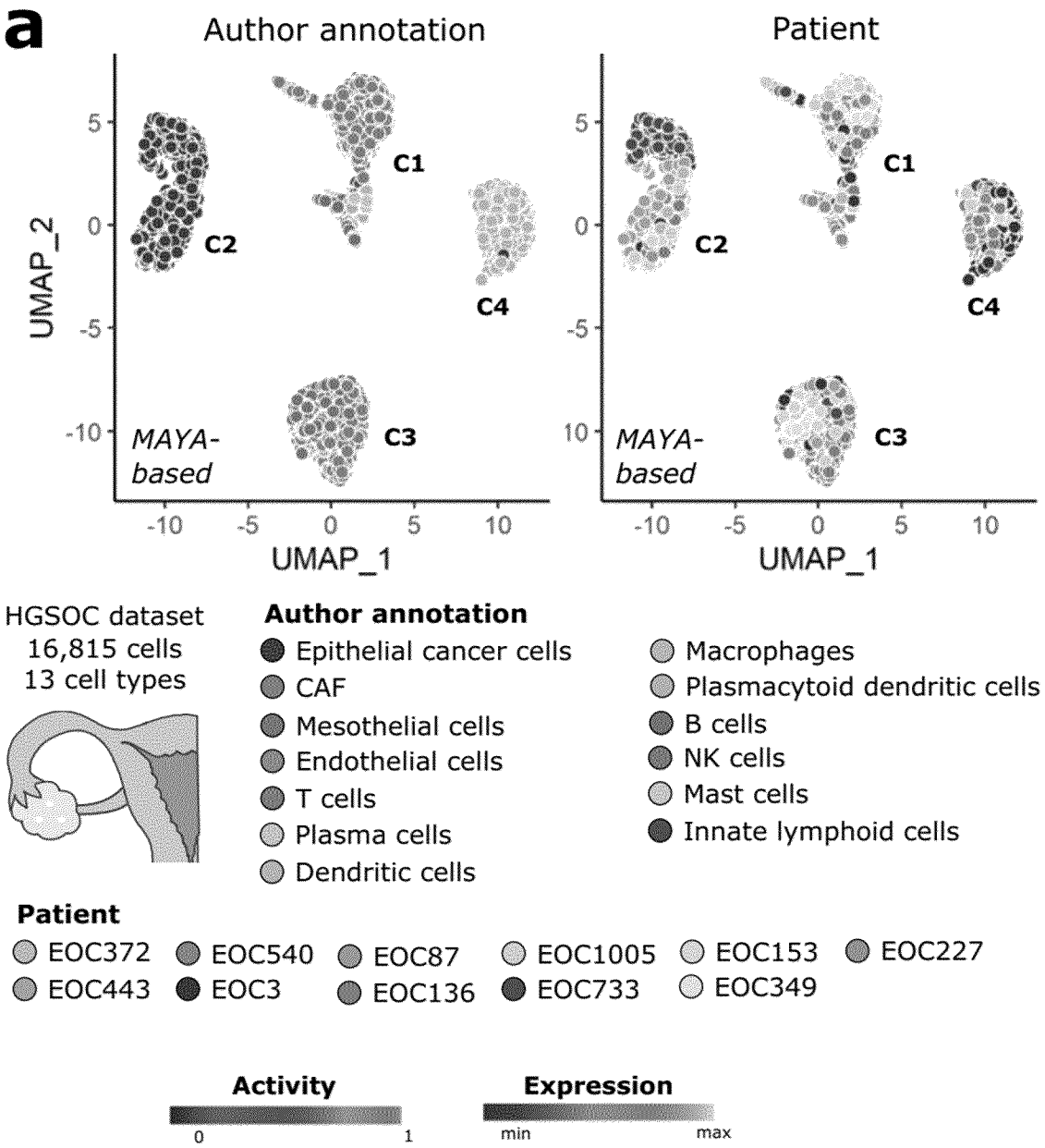

a
Author annotation
Patient
UMAP_2
UMAP_1
C1
C2
C3
C4
MAYA-based
HGSOC dataset
16,815 cells
13 cell types
Author annotation
Epithelial cancer cells
CAF
Mesothelial cells
Endothelial cells
T cells
Plasma cells
Dendritic cells
Macrophages
Plasmacytoid dendritic cells
B cells
NK cells
Mast cells
Innate lymphoid cells
Patient
EOC372
EOC540
EOC87
EOC1005
EOC153
EOC227
EOC443
EOC3
EOC136
EOC733
EOC349
Activity
0
1
Expression
min
max

FIG. 5A

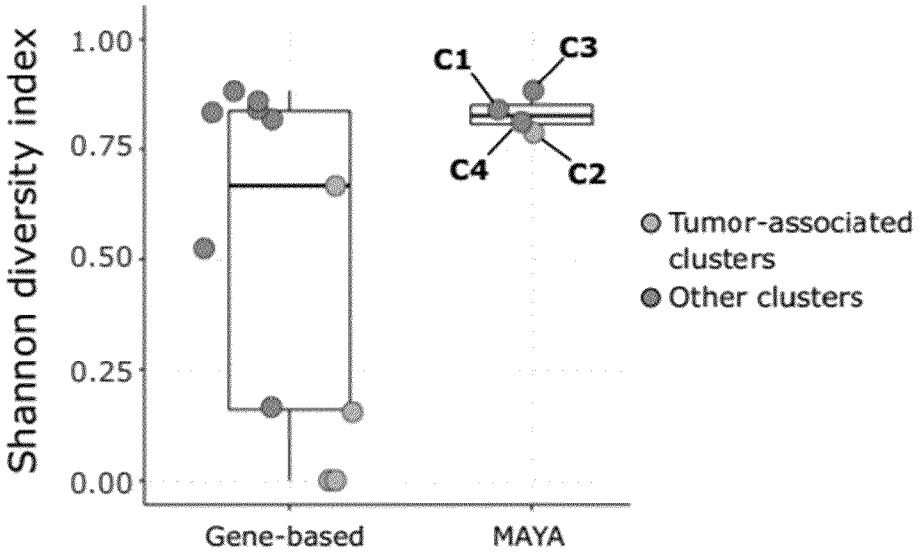

Shannon diversity index
1.00
0.75
0.50
0.25
0.00
Gene-based
MAYA
C1
C3
C4
C2
Tumor-associated clusters
Other clusters

FIG. 5B

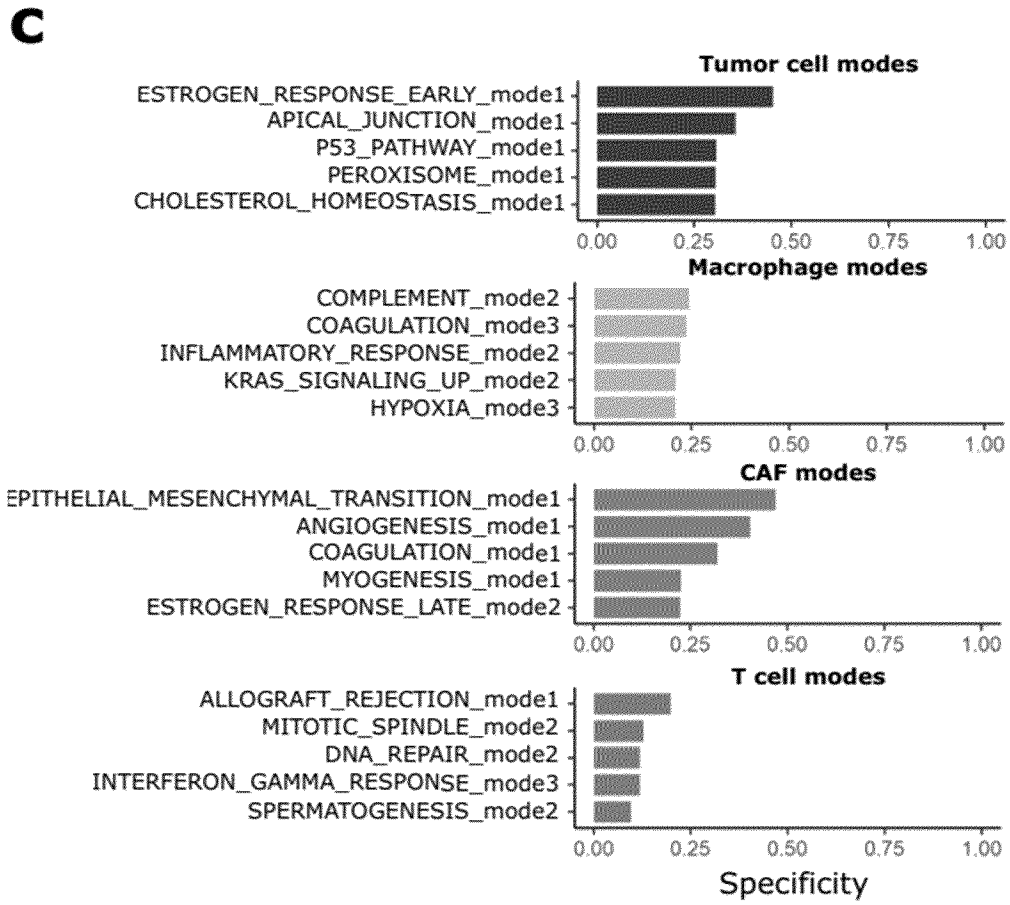

C
Tumor cell modes
ESTROGEN_RESPONSE_EARLY_mode1
APICAL_JUNCTION_mode1
P53_PATHWAY_mode1
PEROXISOME_mode1
CHOLESTEROL_HOMEOSTASIS_mode1
Macrophage modes
COMPLEMENT_mode2
COAGULATION_mode3
INFLAMMATORY_RESPONSE_mode2
KRAS_SIGNALING_UP_mode2
HYPOXIA_mode3
CAF modes
EPITHELIAL_MESENCHYMAL_TRANSITION_mode1
ANGIOGENESIS_mode1
COAGULATION_mode1
MYOGENESIS_mode1
ESTROGEN_RESPONSE_LATE_mode2
T cell modes
ALLOGRAFT_REJECTION_mode1
MITOTIC_SPINDLE_mode2
DNA_REPAIR_mode2
INTERFERON_GAMMA_RESPONSE_mode3
SPERMATOGENESIS_mode2
0.00
0.25
0.50
0.75
1.00
Specificity

FIG. 5C

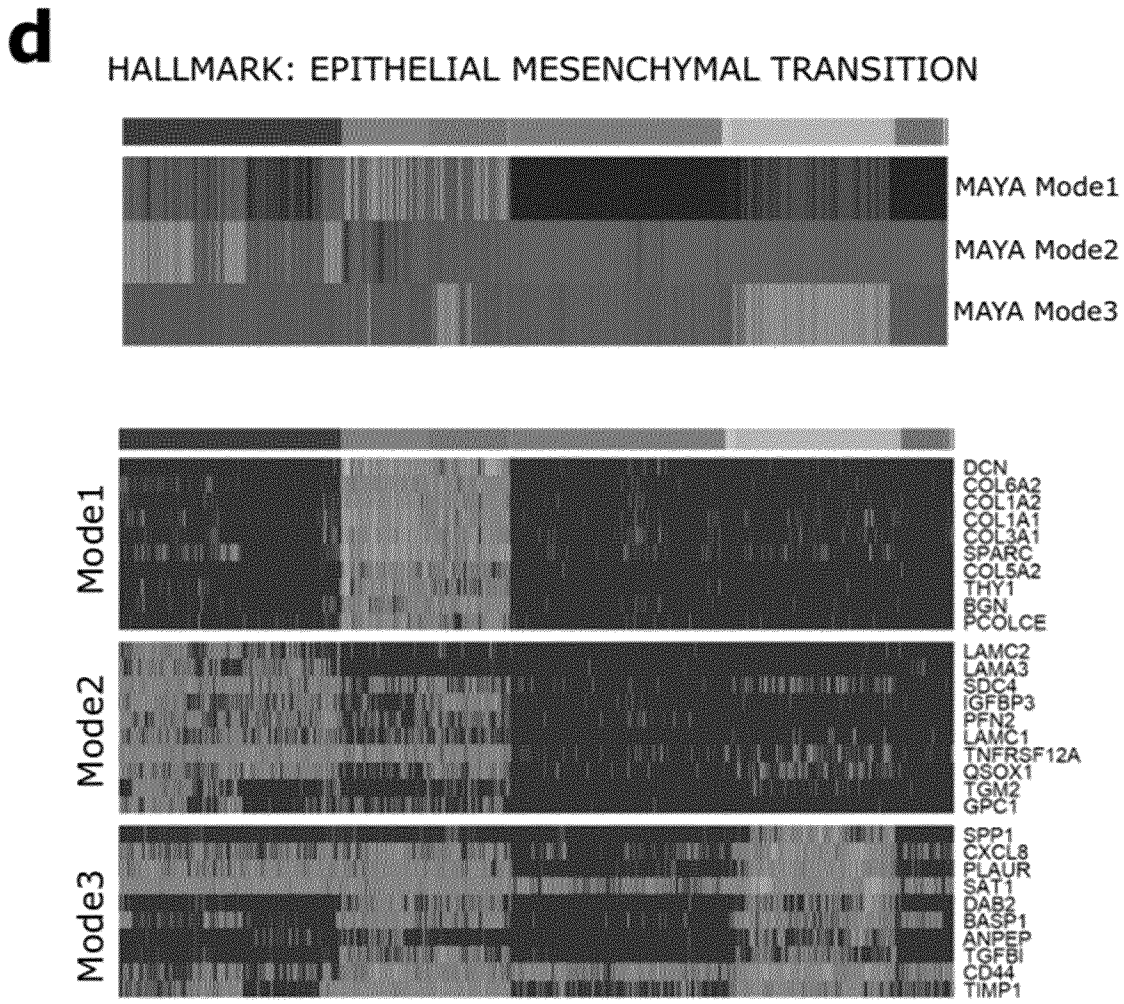

d
HALLMARK: EPITHELIAL MESENCHYMAL TRANSITION
MAYA Mode1
MAYA Mode2
MAYA Mode3
Mode1
DCN
COL6A2
COL1A2
COL1A1
COL3A1
SPARC
COL5A2
THY1
BGN
PCOLCE
Mode2
LAMC2
LAMA3
SDC4
IGFBP3
PFN2
LAMC1
TNFRSF12A
QSOX1
TGM2
GPC1
Mode3
SPP1
CXCL8
PLAUR
SAT1
DAB2
BASP1
ANPEP
TGFBI
CD44
TIMP1

FIG. 5D

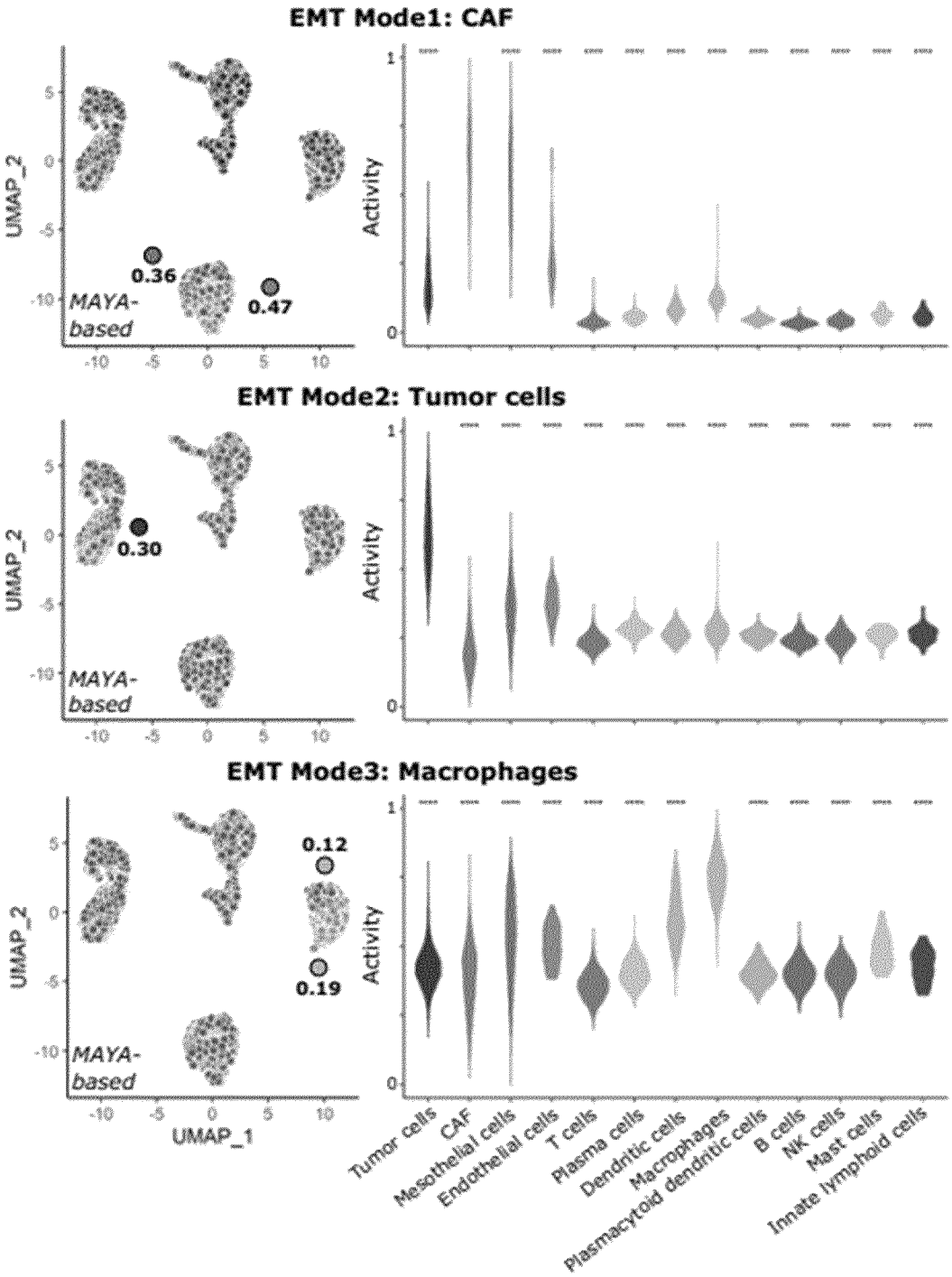

EMT Mode1: CAF
UMAP_2
Activity
0.36
0.47
MAYA-based
EMT Mode2: Tumor cells
0.30
EMT Mode3: Macrophages
0.12
0.19
UMAP_1
Tumor cells
CAF
Mesothelial cells
Endothelial cells
T cells
Plasma cells
Dendritic cells
Macrophages
Plasmacytoid dendritic cells
B cells
NK cells
Mast cells
Innate lymphoid cells

FIG. 5E

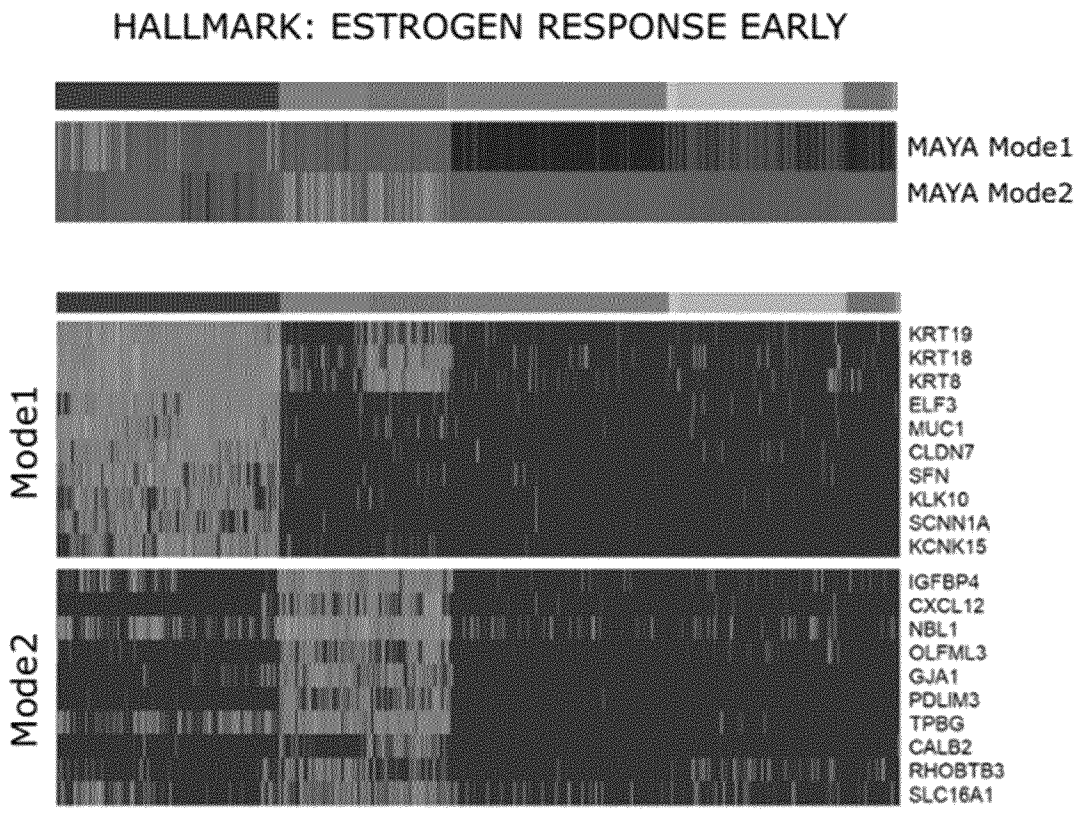

HALLMARK: ESTROGEN RESPONSE EARLY
MAYA Mode1
MAYA Mode2
Mode1
KRT19
KRT18
KRT8
ELF3
MUC1
CLDN7
SFN
KLK10
SCNN1A
KCNK15
Mode2
IGFBP4
CXCL12
NBL1
OLFML3
GJA1
PDLIM3
TPBG
CALB2
RHOBTB3
SLC16A1

FIG. 5F

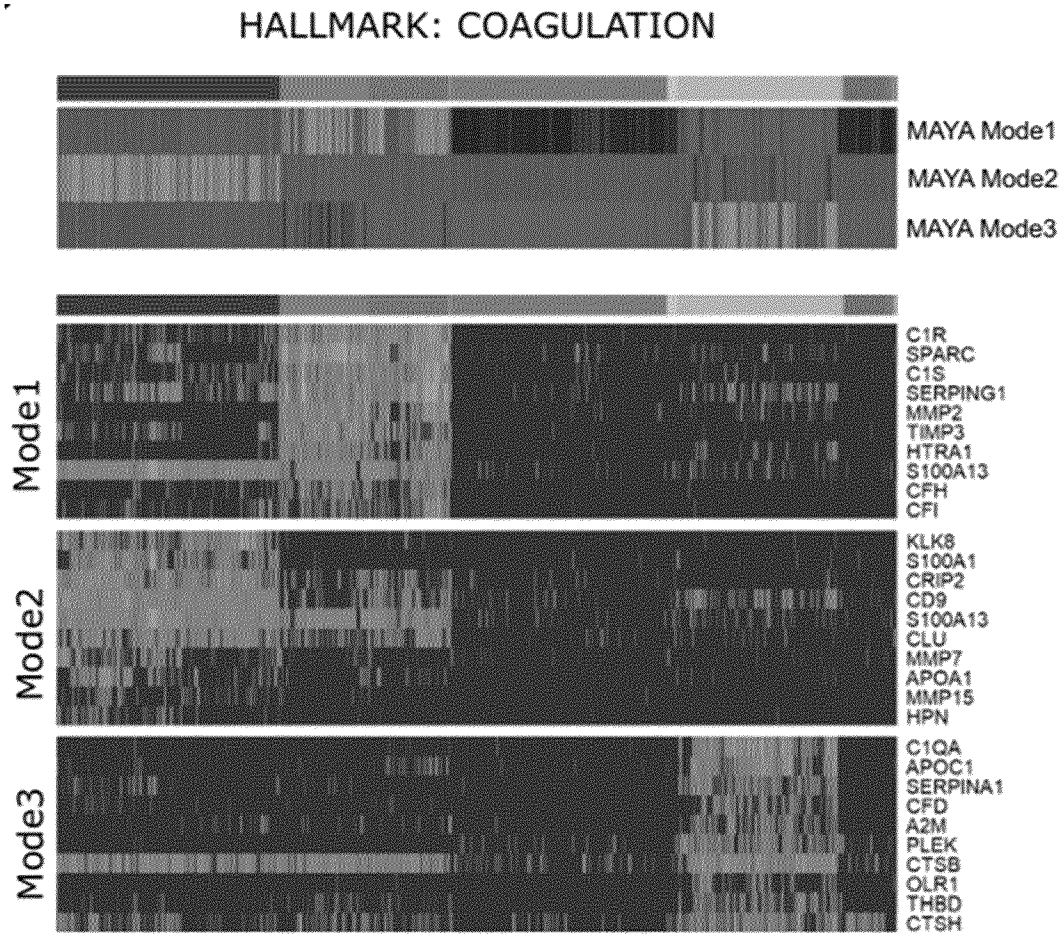

HALLMARK: COAGULATION
MAYA Mode1
MAYA Mode2
MAYA Mode3
Mode1
C1R
SPARC
C1S
SERPING1
MMP2
TIMP3
HTRA1
S100A13
CFH
CFI
Mode2
KLK8
S100A1
CRIP2
CD9
S100A13
CLU
MMP7
APOA1
MMP15
HPN
Mode3
C1QA
APOC1
SERPINA1
CFD
A2M
PLEK
CTSB
OLR1
THBD
CTSH

FIG. 5G

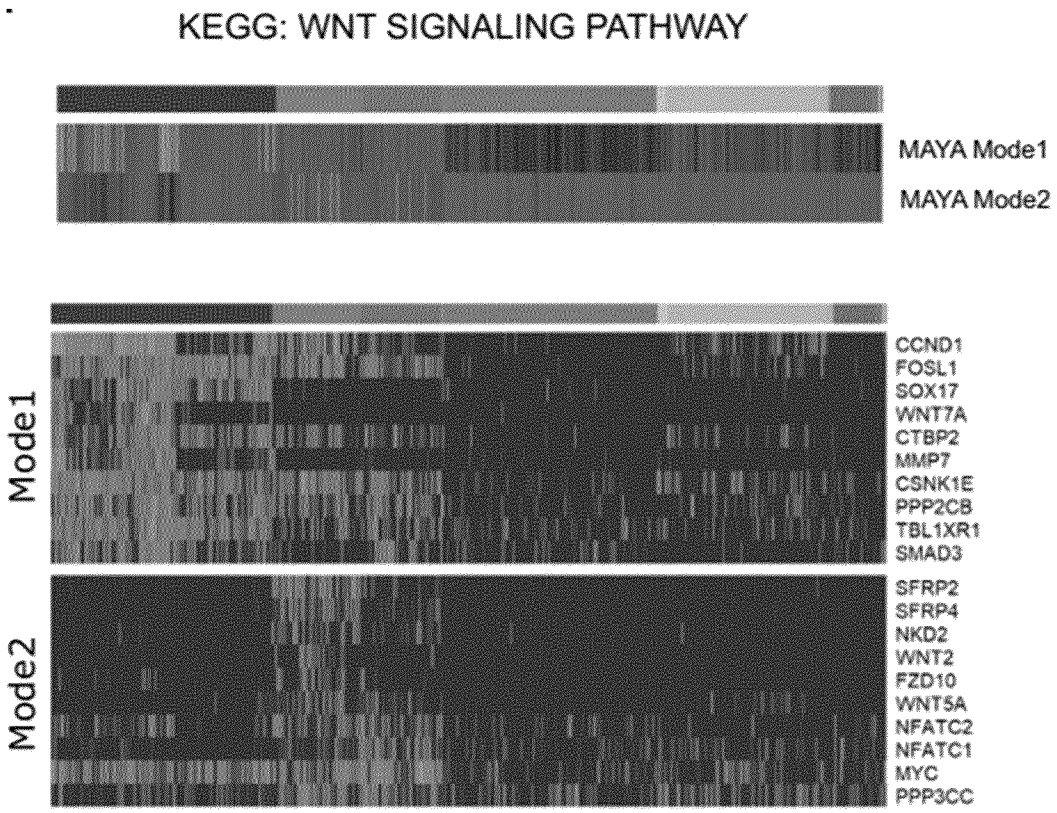

KEGG: WNT SIGNALING PATHWAY
MAYA Mode1
MAYA Mode2
Mode1
CCND1
FOSL1
SOX17
WNT7A
CTBP2
MMP7
CSNK1E
PPP2CB
TBL1XR1
SMAD3
Mode2
SFRP2
SFRP4
NKD2
WNT2
FZD10
WNT5A
NFATC2
NFATC1
MYC
PPP3CC

FIG. 5H

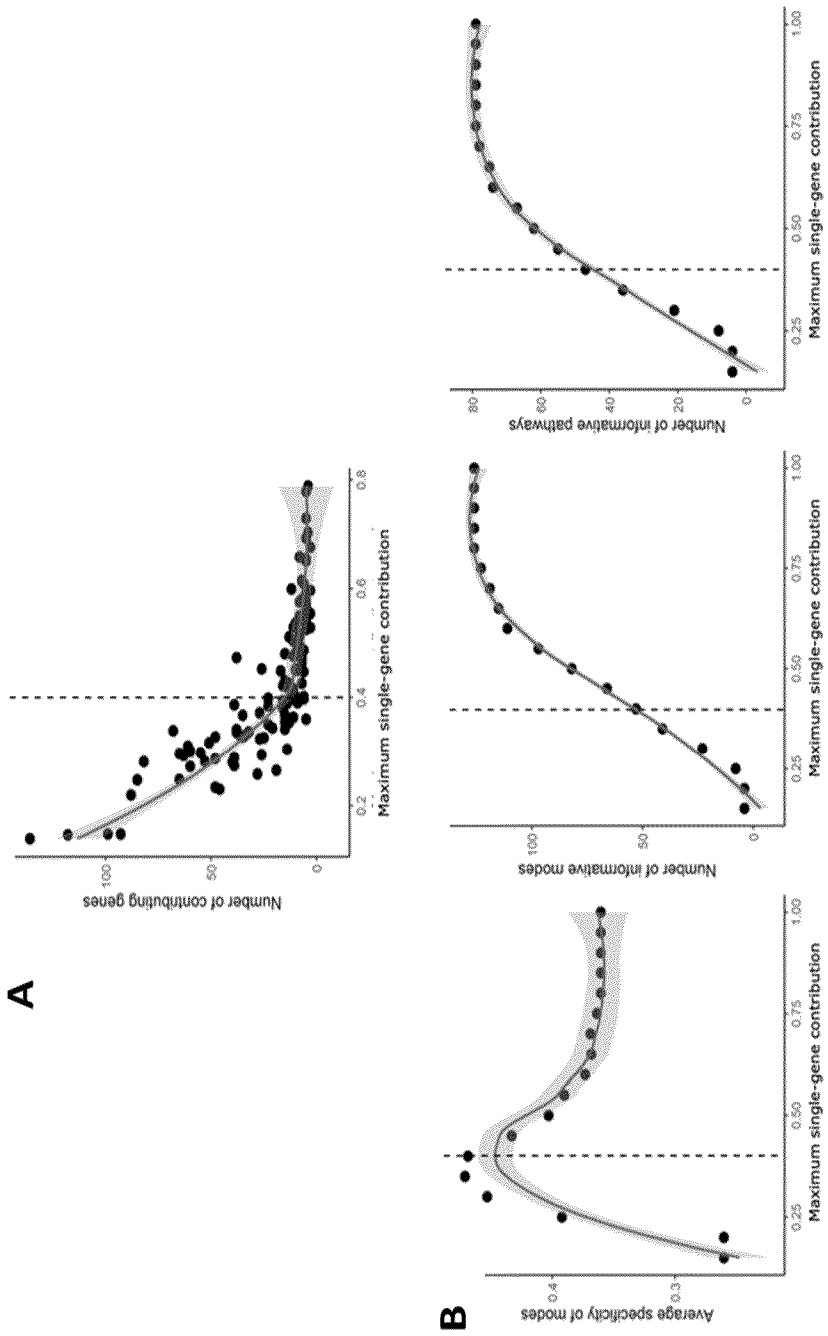

A
Number of contributing genes
Maximum single-gene contribution
B
Average specificity of modes
Number of informative modes
Number of informative pathways
Maximum single-gene contribution

FIG. 6

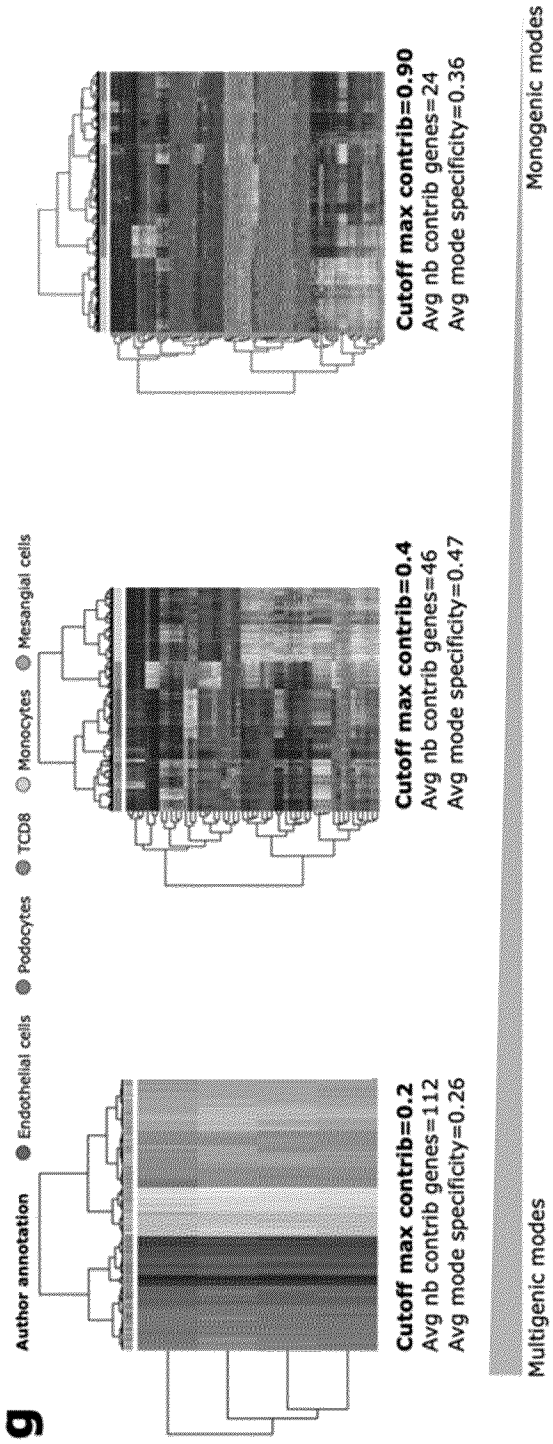

g
Author annotation
Endothelial cells
Podocytes
TCD8
Monocytes
Mesangial cells
Cutoff max contrib=0.2
Avg nb contrib genes=112
Avg mode specificity=0.26
Cutoff max contrib=0.4
Avg nb contrib genes=46
Avg mode specificity=0.47
Cutoff max contrib=0.90
Avg nb contrib genes=24
Avg mode specificity=0.36
Multigenic modes
Monogenic modes

FIG. 7

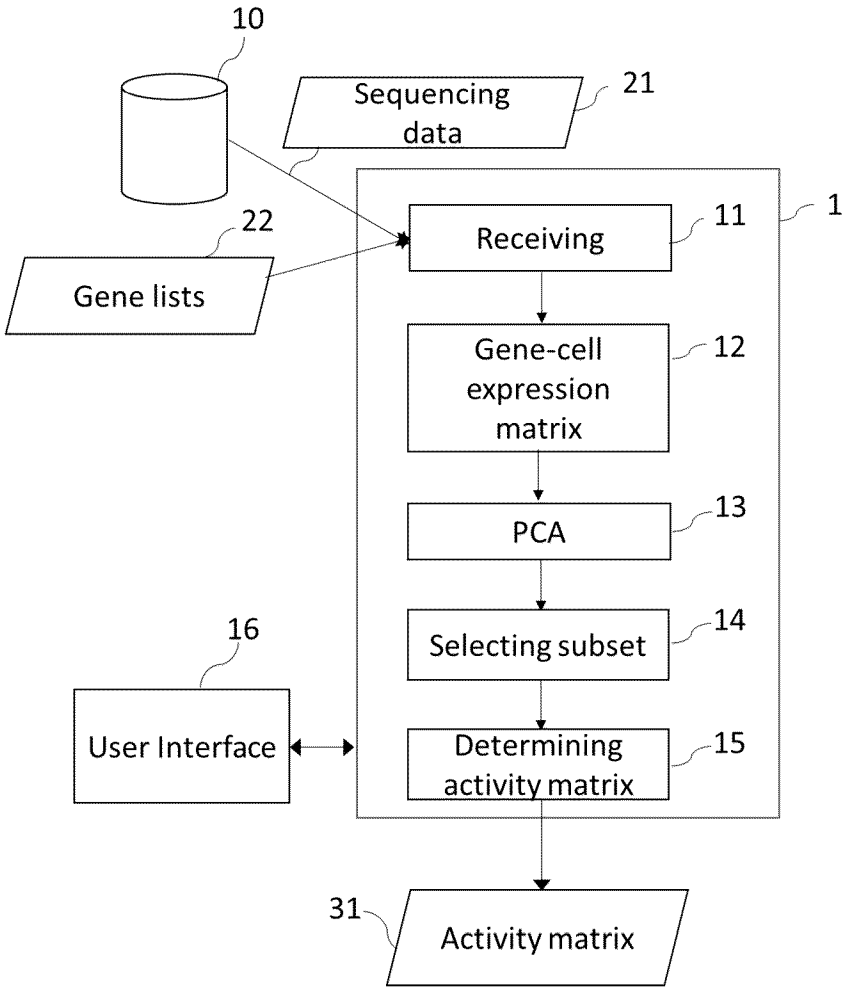

10
Sequencing data
21
22
Gene lists
Receiving
11
1
Gene-cell expression matrix
12
PCA
13
16
Selecting subset
14
User Interface
Determining activity matrix
15
31
Activity matrix

FIG. 8

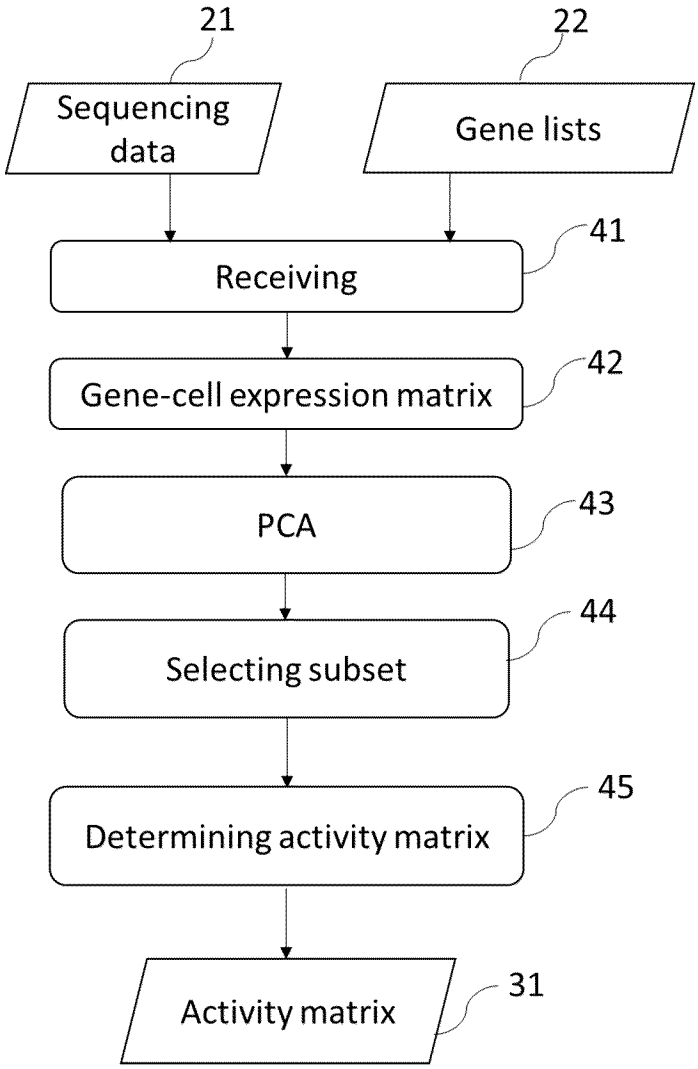

21
22
Sequencing data
Gene lists
41
Receiving
42
Gene-cell expression matrix
43
PCA
44
Selecting subset
45
Determining activity matrix
31
Activity matrix

FIG. 9

# INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2024/070475

**A. CLASSIFICATION OF SUBJECT MATTER**
INV. G16B5/00 G16B25/10
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)
G16B

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| X | LANDAIS YUNA ET AL: "Multi-modal quantification of pathway activity with MAYA", NATURE COMMUNICATIONS, vol. 14, no. 1, 25 March 2023 (2023-03-25) , XP093212979, UK ISSN: 2041-1723, DOI: 10.1038/s41467-023-37410-2 the whole document ----- | 1-19 |

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance;; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance;; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 9 October 2024 | 17/10/2024 |

| Name and mailing address of the ISA/ | Authorized officer |
|---|---|
| European Patent Office, P.B. 5818 Patentlaan 2<br>NL - 2280 HV Rijswijk<br>Tel. (+31-70) 340-2040,<br>Fax: (+31-70) 340-3016 | Vanmontfort, D |



Form PCT/ISA/210 (second sheet) (April 2005)