Multi-modal quantification of pathway activity with MAYA

Yuna Landais, Céline Vallot

Multi-modal quantification of pathway activity with MAYA

Yuna Landais[1], Céline Vallot[2,3,4]

1 One Biosciences, Paris, France

2 CNRS UMR3244, Institut Curie, PSL University, Paris, France

3 Translational Research Department, Institut Curie, PSL University, Paris, France

4 Single Cell Initiative, Institut Curie, PSL University, Paris, France

Abstract

Signaling pathways can be activated through various cascades of genes depending on cell identity and biological context. Single-cell atlases now provide the opportunity to inspect such complexity in health and disease. Yet, existing reference tools for pathway scoring resume activity of each pathway to one unique common metric across cell types. Here, we present MAYA a computational method that enables the automatic detection and scoring of the diverse modes of activation of biological pathways across cell populations. MAYA improves the granularity of pathway analysis by detecting subgroups of genes within reference pathways, each characteristic of a cell population and how it activates a pathway. Using multiple single-cell datasets, we demonstrate the biological relevance of identified modes of activation, the robustness of MAYA to noisy pathway lists and batch effect. MAYA can also predict cell types starting from lists of reference markers in a cluster-free manner. Finally, we show that MAYA reveals common modes of pathway activation in tumor cells across patients, opening the perspective to discover shared therapeutic vulnerabilities.

Introduction

The identification of cell type and function is the driving force of a majority of 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[1–3], which makes the resulting conclusions highly dependent on the chosen algorithm and clustering parameters. In the case of oncogenic datasets, such clustering-based approaches appear inadequate to identify shared transcriptional programs across tumors as cancer cells tend to cluster independently per patient[4–9] rather than group by biological similarities. Several approaches have emerged, bypassing dimensionality reduction and clustering, by proposing to score pathway activity directly in individual cells rather than clusters. Such pooling of gene-based measurements into scores for gene lists has proven extremely powerful for the interpretation of sparse and noisy scRNA-seq datasets[10,11]. A recent benchmark[12] presented Pagoda2[13] and AUCell[14] 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 select significant scores. Nonetheless, both tools compute a unique activity score by 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 a majority of single-cell studies study pathway activation with single scores based on gene lists built from bulk data. Such curated gene lists represent the current reference biological knowledge and are the only available key 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. In order to already inspect existing pathway databases with single-cell resolution, we developed MAYA (Multimodes of pathwAY Activation), a tool that detects for each pathway the different modes of activation across cell types, each mode relying on different subsets of genes. We 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 oncogenic datasets, we show that MAYA can detect cell type specific modes of pathway activation for both the microenvironment and tumor cells, identifying common transcriptional programs across patients.

Results

MAYA method

MAYA enables comprehensive pathway study thanks to multimodal activity scoring of gene lists in individual cells (Fig. 1). Provided a scRNA-seq count matrix and pathway lists, MAYA detects all biologically relevant ways to activate each pathway relying subgroups of genes and summarizes their activity in each cell in a multimodal pathway activity score matrix (Fig. 1a). This activity matrix can then be used to identify groups of cells sharing similar activation of provided pathways and as a dimensionally-reduced dataset for cell visualization (Fig. 1b). As a comparison, reference tools that measure pathway activity, such as AUCell[14] or Pagoda2[13], provide a unique activity score per pathway where MAYA can provide several.

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 whose expression is coordinated and variable across cells, and to simultaneously 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.

However, all detected modes might not reflect a relevant biological pattern in the data and could be driven by outliers, either cells and/or genes, and this probability increases as modes explain less and less variance in the dataset. We 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, which can be determined based on previous knowledge of the underlying biology or set arbitrarily (Supplementary Fig.1a-c). 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, we 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. 1d,e). 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. 1f).

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 gene lists. To this end, we have developed a built-in function to leverage MAYA's scoring and informativity methods to automatically annotate cells in a dataset. This approach is based on activation of the first mode of provided cell type markers lists, using PanglaoDB[16] by default (Methods). 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 minute and 125,000 cells in approximately 15 minutes (Supplementary Fig.1g).

MAYA detects biologically relevant multimodal pathway activity in kidney

The main distinguishing feature of MAYA over existing pathway activity scoring tools is the multimodality of its activity score, which proves useful when studying broad pathways in complex biological systems. We first sought to demonstrate its ability to detect cell-type specific activation modes of hallmark pathways. For that, we ran MAYA on a dataset of normal kidney and immune cells from Young et al.[17], from which we selected cells from 5 distinct subtypes for clarity (n=1,252). We used the MSigDB Hallmark pathways[18] as input gene lists, covering main biological functions. Unsupervised clustering on the multimodal activity matrix shows MAYA detects modes that distinguish different cell populations (Fig. 2a). More specifically, we 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 presentation[19] and gene regulation during myeloid development[20] - 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). 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). Another detailed example is shown in Supplementary Figure 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 specific to the population in which the mode is found to be active: (i) IL6ST is a signal transducer, which dimerizes with IL6R and bound for instance by IL-6, resulting in the activation of downstream cascades in endothelial cells[21], (ii) IL1B is a lymphocyte activating factor produced by monocytes, macrophages and neutrophils, and (iii) IL7R is associated with T cell differentiation.

Altogether, we demonstrate here that MAYA identifies relevant cell-type specific modes of pathway activation from general reference gene lists.

To test both the stability and the ability of MAYA to detect biologically relevant signal in noisy gene lists, we 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 a 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). These results also show the accuracy of our selection method to detect relevant modes, as we 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).

MAYA detects biologically relevant multimodal pathway activity in colon

We 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.[22] - from which we selected cells from 10 distinct cell types (n=1,415) - and using the MSigDB KEGG and REACTOME pathways[23]. 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, we 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 molecules[24], 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 junctions[25,26], 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).

Applying MAYA to the REACTOME pathway *ion channel transport*, we were able to detect different types of ion channels and functions, specific to each cell populations (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-ATPase[27] - that have been shown to participate to the regulation of salt and water absorption from the colon lumen[28,29]. In particular, activation mode 1 captures genes regulating ENaCs and their residence at the apical membrane: *SCNN1A* encodes a subunit of ENaCs[30], *NEDD4L*

participates to ENaCs ubiquitination which leads to their retrieval from cell surface[31] and *SGK1* is known to phosphorylate NEDD4L product, which decreases its binding to ENaCs[32,33]. 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 secretion[34]. 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 ATPases[35] (*ATP6V1B2, ATP6AP1, ATP6V1F, ATP6V0E1, ATP6V0D2*[36]) and Chloride channels[37] (*TTYH3, CLIC2*), a process necessary for phagocytosis. Altogether, as for the kidney, starting from reference databases, MAYA untangles pathway activities specific to each cell type, revealing precise cell functions.

MAYA automatically assigns cell identity

We then leveraged MAYA's scoring and selection ability to automatically and robustly assign cell identity. We applied MAYA to PanglaoDB cell type marker lists and the subsets of kidney and colon datasets used previously (Fig. 4a,d). We demonstrate that MAYA enabled an automated and accurate annotation of each cell in the two datasets, using the initial cell type annotation by authors as a reference (Fig. 4b,e). We compared the accuracy of our predictions with the ones obtained with three other algorithms: AUCell[14], Pagoda2[13] and Cell-ID[38], a cell type identification method based on Multiple Correspondence Analysis (MCA). MAYA presents among the highest rates of recall and precision for both datasets (Fig. 4c,f and Supplementary Fig. 4a,b). We finally tested the scalability of MAYA and its ability to detect rare cell types on a dataset with 16,815 cells from ovarian tumors[6] (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.

Furthermore, as batch effect is a main concern in single-cell analyses, notably for data visualization and cell annotation, we tested whether MAYA was affected by such technical biases. We worked on a dataset containing n=5,179 cells from laryngeal squamous cell carcinoma biopsies of 2 patients with a batch effect between patients[39]. 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). To quantify the inter-patient overlap between clusters of similar cell types, we computed the Shannon Diversity Index (SDI) for both methods as well as for clusters obtained with the reference integration tool Harmony[40] (Supplementary Fig. 4e). MAYA had an average SDI of 0.77 against 0.65 and 0.17 for the integration-based and the gene-based method respectively (Fig 4h). 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.

MAYA detects common modes of pathway activation across cancer patients

Patient-specificity of cancer cells is currently a major limitation for the comprehensive study of oncogenic 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 clusters[4–9]. 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 Harmony[40], are not suited to deal with such cell-type specific effect.

We 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, we show that MAYA identifies several modes of pathway activation shared across patients (Fig. 5a,b and Supplementary Fig. 5a,b) that are associated with known cancer hallmarks. Indeed, top specific modes of epithelial cancer cells reflect the expression of targets of the oncogene KRAS, genes associated with early response to estrogen or the P53 pathway (specificity of 0.63, 0.45 and 0.31 respectively), that all relate to tumor growth and proliferation (Fig. 5b). MAYA also identifies modes of pathway activation specific to tumor microenvironment populations. It notably detects a cell-type specific activation of complement genes in macrophages (specificity of 0.24) and of angiogenesis-related genes in cancer-associated fibroblasts (CAFs) (specificity of 0.40).

MAYA 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. 5c): 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). MAYA identifies a combination of genes that characterizes EMT occurring in epithelial cells, with *LAMA3* and *LAMC2* being exclusive to this cell type (Fig. 5e). 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 motility[41,42]. 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 directly[43] or by increasing the ECM stiffness[44,45]. 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 cancer[46–49]. Two additional modes are detected but are not as cell-type specific as the others (Supplementary Fig. 5a, maximum specificity scores of 0.12).

MAYA also identifies two different modes of activation of the *estrogen response early cascade* (Supplementary Fig. 5c,d), one specific to tumor cells, and one specific to CAFs, consistent with the observation that CAFs can use ER-mediated signaling pathways to promote tumor cell proliferation[50,51]. MAYA also helps to untangle the respective contribution of cancer cells and its microenvironment to the hemostatic imbalance observed in cancer[52,53], by detecting *coagulation* modes with high specificity for CAFs and mesothelial cells (0.31 and 0.32), tumor cells (0.22) and macrophages (0.24) (Supplementary Fig. 5e,f).

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. We 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.

Discussion

MAYA sorts 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. We 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 also performs automated 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 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.

In comparison to previously published methods (AUCell[14], Pagoda2[13], ROMA[54] and UCell[55]), MAYA provides multiple activation scores per pathway, and in a time efficient and user-friendly way. Indeed, running Pagoda2 for example can quickly become computationally intensive; its selection method requires to build a null distribution for each pathway by retrieving variance explained by PC1 for random gene lists of the same pathway size – which drastically increases the number of PCA run to score a single pathway. AUCell computes several bimodality thresholds by pathway, which can also increase computing time, and needs rather advanced users to tune its technical parameters if default ones do not provide satisfying results. With MAYA, we simplified bimodal detection by focusing on inflection points and introducing two biologically interpretable parameters, easily tuned 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.

We 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 we know pathway and cell markers manual curation is very time-consuming. Here, we 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.

We also leveraged our methods of scoring and selection of informative scores to propose a built-in function to automatically annotate cells using PanglaoDB cell type markers lists. This method performs better with MAYA scores than Pagoda2 or AUCell scores and has performance results equivalent to or better than Cell-ID[38], a package specialized in cell type annotation. MAYA is scalable to large datasets (>100,000 cells, in 15 minutes) and it is able to accurately detect and annotate cell populations representing less than 5% of cells. MAYA is therefore an all-in-one tool proposing both cell type identification - like Cell-ID[38], CellTypist[56] and scGate[57] - and multi-modal pathway analysis.

MAYA also enables to identify shared identity expression patterns between cells from the same type across patients, which proves useful in case of batch effect. Indeed, as MAYA focuses on cell identity by looking only at genes considered as markers, it does not detect the variations between patients driven by other sets of genes that are not related to cell type identity and that lead to the formation of different clusters in a classical gene-based analysis.

Finally, MAYA brings particular biological insights 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, Gavish et al.[15] provided the community with clues about 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 we interrogate existing knowledge. 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 ensemble 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.

References

1. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).

2. Pereira, W. J. *et al.* Asc-Seurat: analytical single-cell Seurat-based web application. *BMC Bioinformatics* **22**, 556 (2021).

3. Prieto, C., Barrios, D. & Villaverde, A. SingleCAnalyzer: Interactive Analysis of Single Cell RNA-Seq Data on the Cloud. *Front. Bioinforma.* **2**, (2022).

4. Wu, F. *et al.* Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nat. Commun.* **12**, 2540 (2021).

5. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).

6. Zhang, K. *et al.* Longitudinal single-cell RNA-seq analysis reveals stress-promoted chemoresistance in metastatic ovarian cancer. *Sci. Adv.* **8**, (2022).

7. Maynard, A. *et al.* Therapy-Induced Evolution of Human Lung Cancer Revealed by Single-Cell RNA Sequencing. *Cell* **182**, 1232-1251.e22 (2020).

8. Neftel, C. *et al.* An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell* **178**, 835-849.e21 (2019).

9. Puram, S. V *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624.e24 (2017).

10. Wang, S. *et al.* Single-Cell Transcriptomic Atlas of Primate Ovarian Aging. *Cell* **180**, 585-600.e19 (2020).

11. Ramirez, A. K. *et al.* Single-cell transcriptional networks in differentiating preadipocytes suggest drivers associated with tissue heterogeneity. *Nat. Commun.* **11**, 2117 (2020).

12. Zhang, Y. *et al.* Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data. *Comput. Struct. Biotechnol. J.* **18**, 2953–2961 (2020).

13. Fan, J. *et al.* Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* (2016) doi:10.1038/nmeth.3734.

14. Aibar, S. *et al.* SCENIC: Single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).

15. Gavish, A. *et al.* The transcriptional hallmarks of intra-tumor heterogeneity across a thousand tumors. *bioRxiv* 2021.12.19.473368 (2021) doi:10.1101/2021.12.19.473368.

16. Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019**, 46 (2019).

17. Young, M. D. *et al.* Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science (80-.).* **361**, 594–599 (2018).

18. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* **1**, 417–425 (2015).

19. Nakagawa, T. Y. & Rudensky, A. Y. The role of lysosomal proteinases in MHC class II-mediated antigen processing and presentation. *Immunol. Rev.* **172**, 121–129 (1999).

20. Klemsz, M. J., McKercher, S. R., Celada, A., Van Beveren, C. & Maki, R. A. The macrophage and B cell-specific transcription factor PU.1 is related to the ets oncogene. *Cell* **61**, 113–124 (1990).

21. Kang, S. & Kishimoto, T. Interplay between interleukin-6 signaling and the vascular endothelium in cytokine storms. *Exp. Mol. Med.* **53**, 1116–1123 (2021).

22. Lee, H. O. *et al.* Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat. Genet.* **52**, 594–603 (2020).

23. Liberzon, A. *et al.* Databases and ontologies Molecular signatures database (MSigDB) 3.0. *Bioinforma. Appl. NOTE* **27**, 1739–1740 (2011).

24. Handunnetthi, L., Ramagopalan, S. V, Ebers, G. C. & Knight, J. C. Regulation of major histocompatibility complex class II gene expression, genetic variation and disease. *Genes Immun.* **11**, 99–112 (2010).

25. Tsukita, S., Tanaka, H. & Tamura, A. The Claudins: From Tight Junctions to Biological Systems. *Trends Biochem. Sci.* **44**, 141–152 (2019).

26. Braga, V. Epithelial cell shape: cadherins and small GTPases. *Exp. Cell Res.* **261**, 83–90 (2000).

27. Bibert, S. *et al.* A link between FXYD3 (Mat-8)-mediated Na,K-ATPase regulation and differentiation of Caco-2 intestinal epithelial cells. *Mol. Biol. Cell* **20**, 1132–1140 (2009).

28. Rajendran, V. M., Schulzke, J.-D. & Seidler, U. E. Chapter 58 - Ion Channels of the Gastrointestinal Epithelial Cells. in (ed. Said, H. M. B. T.-P. of the G. T. (Sixth E.) 1363–1404 (Academic Press, 2018). doi:https://doi.org/10.1016/B978-0-12-809954-4.00058-X.

29. Kunzelmann, K. & Mall, M. Electrolyte transport in the mammalian colon: mechanisms and implications for disease. *Physiol. Rev.* **82**, 245–289 (2002).

30. Saxena, A. *et al.* Gene Structure of the Human Amiloride-Sensitive Epithelial Sodium Channel Beta Subunit. *Biochem. Biophys. Res. Commun.* **252**, 208–213 (1998).

31. Zhou, R., Patel, S. V & Snyder, P. M. Nedd4-2 catalyzes ubiquitination and degradation of cell surface ENaC. *J. Biol. Chem.* **282**, 20207–20212 (2007).

32. Lang, F. *et al.* Regulation of channels by the serum and glucocorticoid-inducible kinase - implications for transport, excitability and cell proliferation. *Cell. Physiol. Biochem. Int. J. Exp. Cell. Physiol. Biochem. Pharmacol.* **13**, 41–50 (2003).

33. Snyder, P. M. Minireview: Regulation of Epithelial Na+ Channel Trafficking. *Endocrinology* **146**, 5079–5085 (2005).

34. Gruber, A. D. *et al.* Genomic cloning, molecular characterization, and functional analysis of human CLCA1, the first human member of the family of Ca2+-activated Cl- channel proteins. *Genomics* **54**, 200–214 (1998).

35. Grinstein, S., Nanda, A., Lukacs, G. & Rotstein, O. V-ATPases in phagocytic cells. *J. Exp. Biol.* **172**, 179–192 (1992).

36. Xia, Y. *et al.* The macrophage-specific V-ATPase subunit ATP6V0D2 restricts inflammasome activation and bacterial infection by facilitating autophagosome-lysosome fusion. *Autophagy* **15**, 960–975 (2019).

37. Carraro-Lacroix, L. R., Lessa, L. M. A., Fernandez, R. & Malnic, G. Physiological implications of the regulation of vacuolar H+-ATPase by chloride ions. *Brazilian J. Med. Biol. Res.* **42**, 155–163 (2009).

38. Cortal, A., Martignetti, L., Six, E. & Rausell, A. Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. *Nat. Biotechnol.* **39**, 1095–1102 (2021).

39. Song, L. *et al.* Cellular heterogeneity landscape in laryngeal squamous cell carcinoma. *Int. J. Cancer* **147**, 2879–2890 (2020).

40. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

41. Domogatskaya, A., Rodin, S. & Tryggvason, K. Functional diversity of laminins. *Annu. Rev. Cell Dev. Biol.* **28**, 523–553 (2012).

42. Carpenter, P. M. *et al.* Migration of breast cancer cell lines in response to pulmonary laminin 332. *Cancer Med.* **6**, 220–234 (2017).

43. Wei, S. C. *et al.* Matrix stiffness drives epithelial–mesenchymal transition and tumour metastasis through a TWIST1–G3BP2 mechanotransduction pathway. *Nat. Cell Biol.* **17**, 678–688 (2015).

44. Shintani, Y., Hollingsworth, M. A., Wheelock, M. J. & Johnson, K. R. Collagen I promotes metastasis in pancreatic cancer by activating c-Jun NH(2)-terminal kinase 1 and up-regulating N-cadherin expression. *Cancer Res.* **66**, 11745–11753 (2006).

45. Koenig, A., Mueller, C., Hasel, C., Adler, G. & Menke, A. Collagen type I induces disruption of E-cadherin-mediated cell-cell contacts and promotes proliferation of pancreatic carcinoma cells. *Cancer Res.* **66**, 4662–4671 (2006).

46. Liu, J. *et al.* Association of tumour-associated macrophages with cancer cell EMT, invasion, and metastasis of Kazakh oesophageal squamous cell cancer. *Diagn. Pathol.* **14**, 55 (2019).

47. Su, S. *et al.* A Positive Feedback Loop between Mesenchymal-like Cancer Cells and Macrophages Is Essential to Breast Cancer Metastasis. *Cancer Cell* **25**, 605–620 (2014).

48. Jia, Z., Zhang, Y., Xu, Q., Guo, W. & Guo, A. miR-126 suppresses epithelial-to-mesenchymal transition and metastasis by targeting PI3K/AKT/Snail signaling of lung cancer cells. *Oncol. Lett.* **15**, 7369–7375 (2018).

49. Fu, X.-T. *et al.* Macrophage-secreted IL-8 induces epithelial-mesenchymal transition in hepatocellular carcinoma cells by activating the JAK2/STAT3/Snail pathway. *Int J Oncol* **46**, 587–596 (2015).

50. Rothenberger, N. J., Somasundaram, A. & Stabile, L. P. The Role of the Estrogen Pathway in the Tumor Microenvironment. *Int. J. Mol. Sci.* **19**, 611 (2018).

51. Subramaniam, K. S. *et al.* Cancer-associated fibroblasts promote proliferation of endometrial cancer cells. *PLoS One* **8**, e68923–e68923 (2013).

52. Galmiche, A., Rak, J., Roumenina, L. T. & Saidak, Z. Coagulome and the tumor microenvironment: an actionable interplay. *Trends in Cancer* **8**, 369–383 (2022).

53. Mitrugno, A., Tormoen, G. W., Kuhn, P. & McCarty, O. J. T. The prothrombotic activity of cancer cells in the circulation. *Blood Rev.* **30**, 11–19 (2016).

54. Martignetti, L., Calzone, L., Bonnet, E., Barillot, E. & Zinovyev, A. ROMA: Representation and Quantification of Module Activity from Target Expression Data. *Front. Genet. | www.frontiersin.org* **7**, 18 (2016).

55. Andreatta, M. & Carmona, S. J. UCell: Robust and scalable single-cell gene signature scoring. *Comput. Struct. Biotechnol. J.* **19**, 3796–3798 (2021).

438 56. Domínguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in
439 humans. *Science* **376**, eabl5197 (2022).

440 57. Andreatta, M., Berenstein, A. J. & Carmona, S. J. scGate: marker-based purification of cell
441 types from heterogeneous single-cell RNA-seq datasets. *Bioinformatics* **38**, 2642–2644 (2022).

442

443

444

445

Methods

Code availability: MAYA is available as an R package on GitHub at https://github.com/One-Biosciences/MAYA/. Requires R >= 4.0.5.

Data availability:

Kidney dataset: The count matrices were downloaded from Supplementary data S1 from Young et al. and metadata was built by combining table S11 providing a cell manifest with table S2 providing author's cell type annotation. Only protein-coding genes were kept for downstream analysis. Data was provided for 125,139 cells, with 72,502 cells passing the author's quality control criteria. MAYA automatic annotation function was run on the dataset before and after QC filtering to evaluate its scalability to large datasets. For our detailed pathway analysis, only normal kidney cells were selected based on author's annotation (categories "Normal_mature_kidney" and "Normal_mature_kidney_immune"). Cells from 5 distinct cell types out of 28 were selected after default Seurat processing and clustering (aliases 8T, AV2, MNP1, G and M) for a total 1,252 cells.

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. Only protein-coding genes were kept for downstream analysis. MAYA automatic annotation function was run on this full dataset - including normal, tumor and border cells - to evaluate its scalability to large datasets. For our detailed pathway analysis, cells from Class "Normal" and from 10 out of the 35 cell types identifies by the authors were selected, representing a total of 1,415 cells.

Ovary dataset: Count data were downloaded from the NCBI Gene Expression Omnibus (GEO) database with accession code GSE165897. Only cells labelled as treatment-naïve for the treatment phase metadata field were kept for downstream analysis, representing a total of 16,815 cells.

Larynx dataset: Count data were downloaded from the NCBI Gene Expression Omnibus (GEO) database with accession code GSE150321 (2 files, one for each patient), for a total of 5,179 cells.

Reference databases:

PanglaoDB was downloaded from the website (https://panglaodb.se/) and loaded in R with the provided command line. Markers lists are categorized by organs. Some can be considered as generic organs that should always be tested for a dataset (connective tissue, smooth muscle, immune system, vasculature, blood, epithelium, skeletal muscle), others are more specific such as kidney or lungs and can be loaded on demand. The full Panglao gene list can be loaded as well. Kidney related lists were

loaded for the kidney dataset, GI tract related lists for the colon dataset, and finally no other list than generic types for the larynx and ovary datasets.

MSigDB gene lists (Hallmark, KEGG and REACTOME) were downloaded from the Broad Institute website (http://www.gsea-msigdb.org/gsea/msigdb/collections.jsp) in their version 7.4. For the Reactome database, only pathways comprising between 100 and 300 genes were kept for efficiency purposes, which represents 165 pathways kept over 1615.

Matrix preprocessing: 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 2,000 variable features were found using "vst" method. PCA and UMAP 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 5,179 cells. The authors did not provide their annotation, so we followed the default Seurat pipeline on each individual count matrix, performed PCA and default clustering. We then annotated clusters based on expression of cell type markers described in the publication.

Detailed description of MAYA algorithm:

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 directions for which the absolute value of gene contribution is the highest. Each mode is scaled between 0 and 1. An iterative process then begins: we evaluate the informativity of each successive PC starting from PC1. If a PC is found uninformative, the iteration stops, and we do not interrogate further PCs. There is however an exception for PC1: we 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.

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 2 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, we start with an adjustment meant to detect global variations of distributions and if none are detected we 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, we 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 we 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, we 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, we 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, we 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 we know the following will explain even less variance and is more likely to be noise.

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.

Comparison with other tools:

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.

Pagoda2, AUCell and Cell-ID to compare cell type prediction with MAYA: The three tools were provided the same PanglaoDB cell type marker lists as MAYA. Pagoda2 was run with default settings, following the vignette. AUCell was run using default settings, with log-normalized counts as input. We used AUCell_exploreThresholds function to select the cell type lists that were activated in at least one cell. MAYA's procedure of clustering and cell type attribution was performed on AUCell and Pagoda2 activity matrix as they do not have an integrated function for cell annotation. 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.

Integration with Harmony: Harmony was run through Seurat v3 with default settings.

Metrics:

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

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

With c the cluster in which we compute the SDI, N the number of different possible identities (patients in our case) and p_i 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.

Specificity metric: For a mode, we 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.

$$S_{m,c} = \frac{{a_{m,c}}^2}{\sum_{p=1}^{N} {a_{m,p}}^2}$$

569 $$\sum_{p=1}^{N} {S_{m,p}}^2 = 1$$

570 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
571 number of clusters.

572 We consider that specificity is significant for a cluster when it is 50% above expected value of 1/N
573 (specificity score when all cells across all clusters have the same activity).

574 **Precision, recall, F1-score**

575 $\text{Precision} = \frac{TP}{TP+FP}$; $\text{Recall} = \frac{TP}{TP+FN}$; $\text{F1\_score} = \frac{2*Precison*Recall}{Precision+Recall}$

576 Where *TP* is the number of true positives, *FP* the number of false positives and *FN* the number of false
577 negatives. F1-score of 1 means perfect precision and recall.

578 **Matching PanglaoDB cell types with author annotation for precision and recall assessment:**

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

582 Kidney: Monocytes=c("Monocytes"), Endothelial cells=c("Endothelial cells"),
583 Mesangial_cells=c("Mesangial cells","Smooth muscle cells"), Podocytes=c("Podocytes"), TCD8 =c("T
584 cells","T memory cells", "T helper cells")

585 Colon: `Mature Enterocytes`=c("Enterocytes"), `Goblet cells`=c("Goblet cells"),
586 Pericytes=c("Pericytes"), `Smooth muscle cells`=c("Smooth muscle cells"), cDC=c("Dendritic cells"),
587 Proliferating monocytes=c("Monocytes","Macrophages"), `NK cells`=c("NK cells","Natural killer T
588 cells"), `Regulatory T cells`=c("T regulatory cells","T cells","T memory cells","T helper cells","T follicular
589 helper cells","T cytotoxic cells"), `CD19+CD20+ B`=c("B cells","B cells naive","B cells memory"), `Mast
590 cells`=c("Mast cells")

591

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

593

594

Contributions

Y.L and C.V, as scientific advisor for One Biosciences, conceived the algorithm. Y.L implemented the code, C.V supervised the work. Both authors wrote the manuscript.

Corresponding author

Correspondence to Céline Vallot – celine.vallot@onebiosciences.fr.

Competing interests

C.V. is a founder and equity holder of One Biosciences. The remaining author declares no competing interests.

Rights and permissions

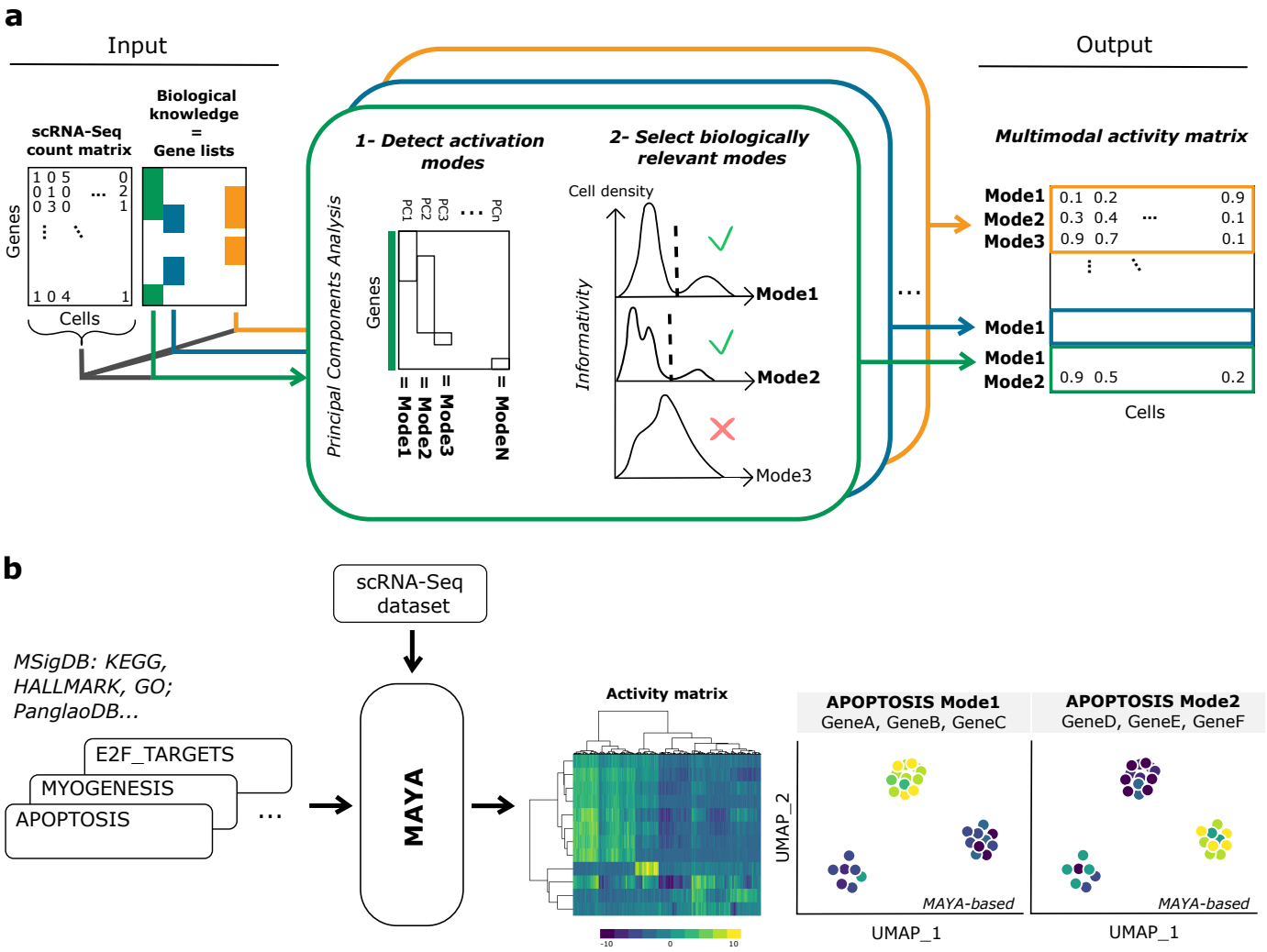
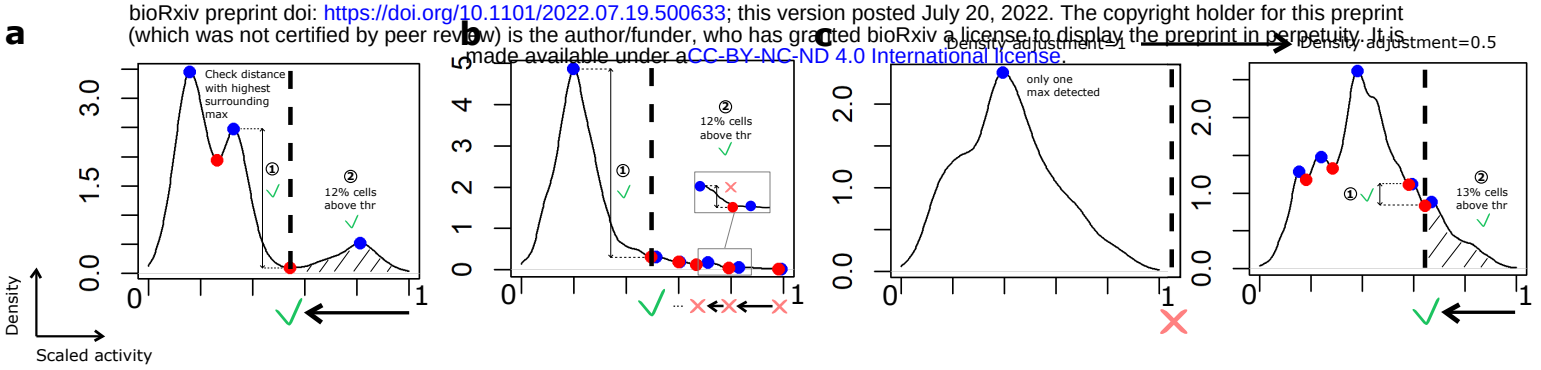
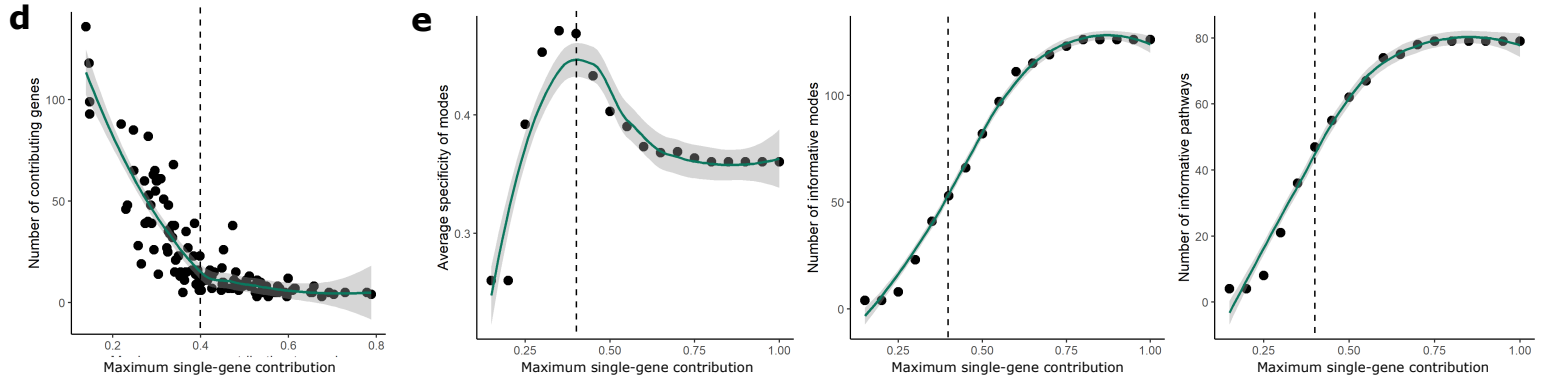
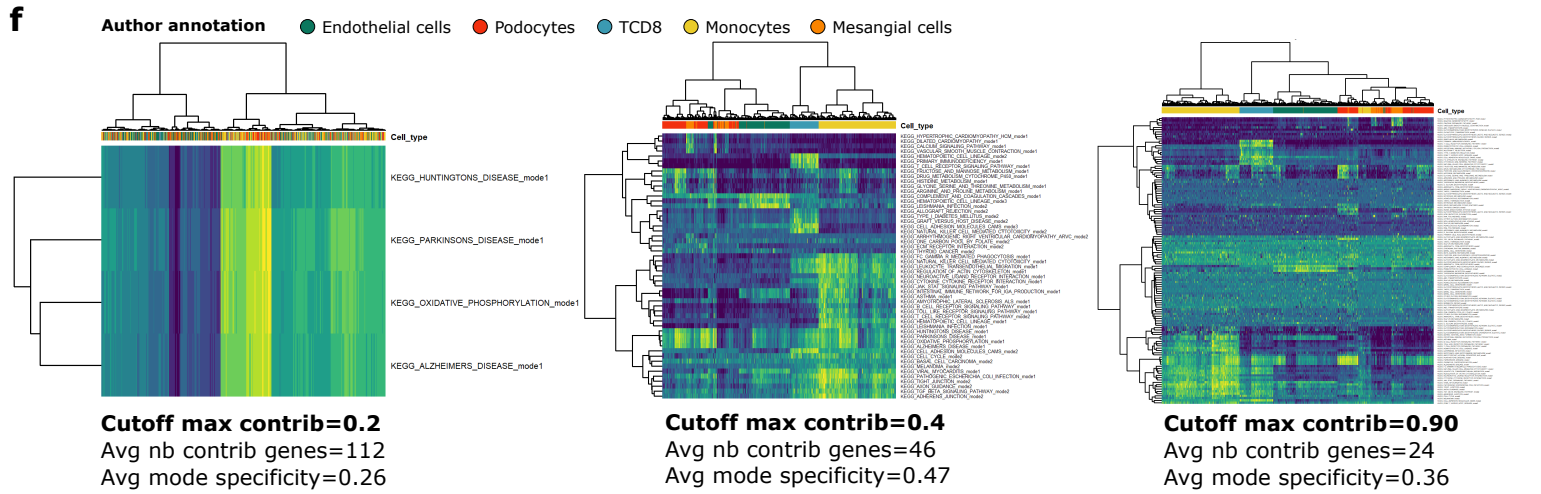


Fig.1: MAYA overview

(a) 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 lists. **(b)** 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).







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

g

Dataset	Number of cells	Gene list	Comput. time Activity matrix	Comput. time Cell annot	Total comput. time
Kidney (subset)	1,252	PanglaoDB Kidney (57 cell types)	3.2 secs	0.19 secs	~ 3 secs
Kidney (pre-filtering)	125,139	PanglaoDB Kidney (57 cell types)	6.1 mins	9.7 mins	~ 15 mins
Kidney (post-filtering)	72,501	PanglaoDB Kidney (57 cell types)	3.1 mins	2.4 mins	~ 5 mins
Colon (subset)	1,415	PanglaoDB GI tract (60 cell types)	2.1 secs	0.12 secs	~ 2 secs
Colon	27,414	PanglaoDB GI tract (60 cell types)	1.0 min	12 secs	~ 1 min
Larynx	5,179	PanglaoDB Basic (47 cell types)	9.9 secs	1.1 secs	~ 11 secs
Ovary	16,815	PanglaoDB Basic (47 cell types)	38 secs	4.6 secs	~ 45 secs

Supplementary Fig.1:
(a,b) Examples of density curve of activity scores for one mode of activation. Detected maxima in density are colored in blue and minima in red. MAYA selects a mode as relevant when it has a local density minimum that (i) is low enough compared with surrounding highest maximum and that (ii) splits the datasets into two fractions that are of a minimal size (Methods). Minima are screened in decreasing order on the x-axis and MAYA stops either when a minimum meets the criteria or when it is to the left of the highest density maximum. In (a) the first minimum at the right meets the two criteria and for (b) the fifth. They are marked by a vertical dashed line. **(c)** When no minima are detected with the first density adjustment parameter, a more fitted adjustment is tested. If minima are found, the procedure described in (a,b) is applied. **(d)** 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. **(e)** 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. **(f)** Heatmap of activity matrices for different cut-off of single-gene contribution: 0.2, 0.4 and 0.9. **(g)** Computing time on different datasets for the two main modules of the function MAYA_predict_cell_type (building activity matrix and annotating cells), using PanglaoDB (44 markers on average per cell type) restricted to cell types expected in the tissue corresponding to the datasets.

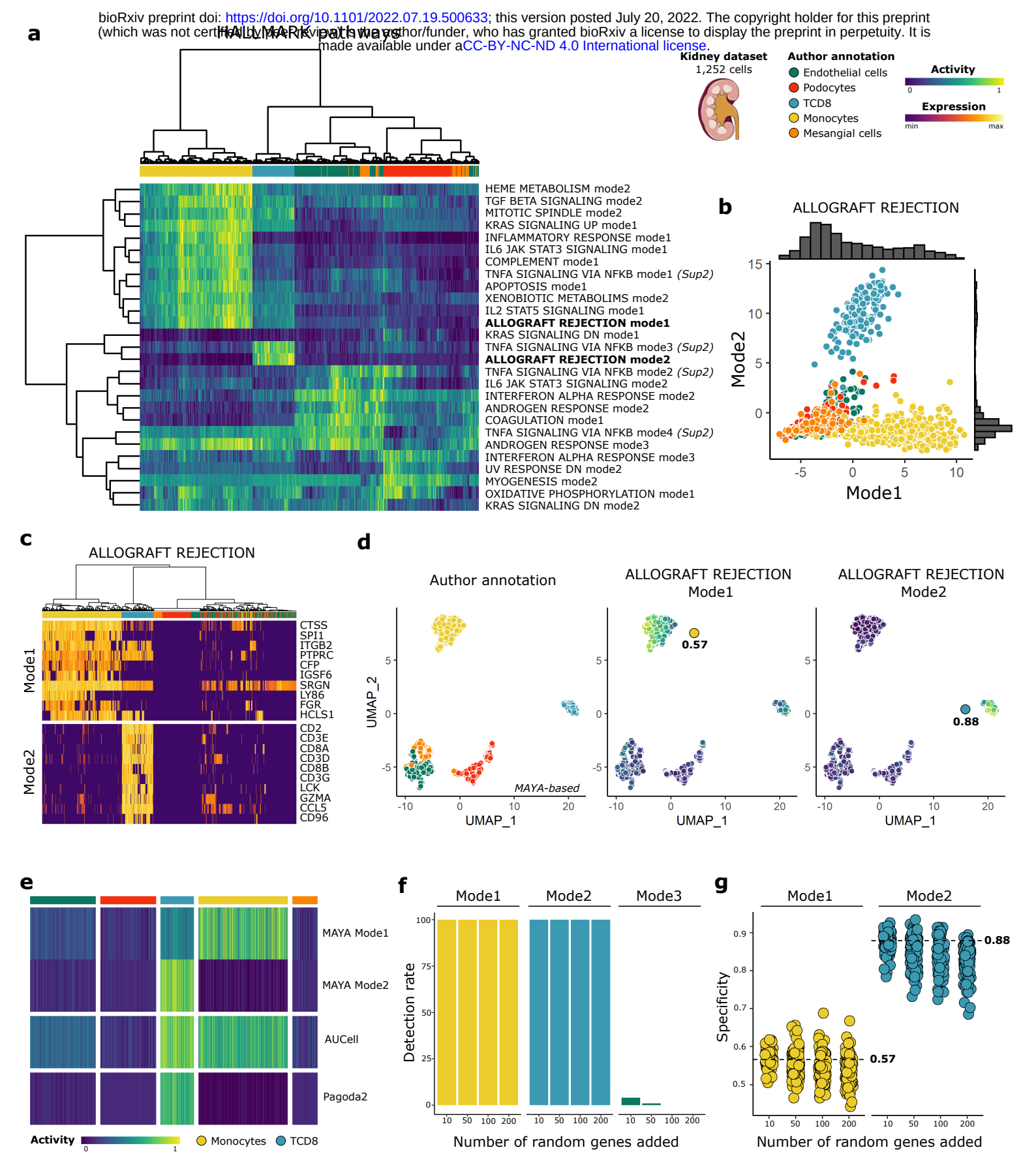
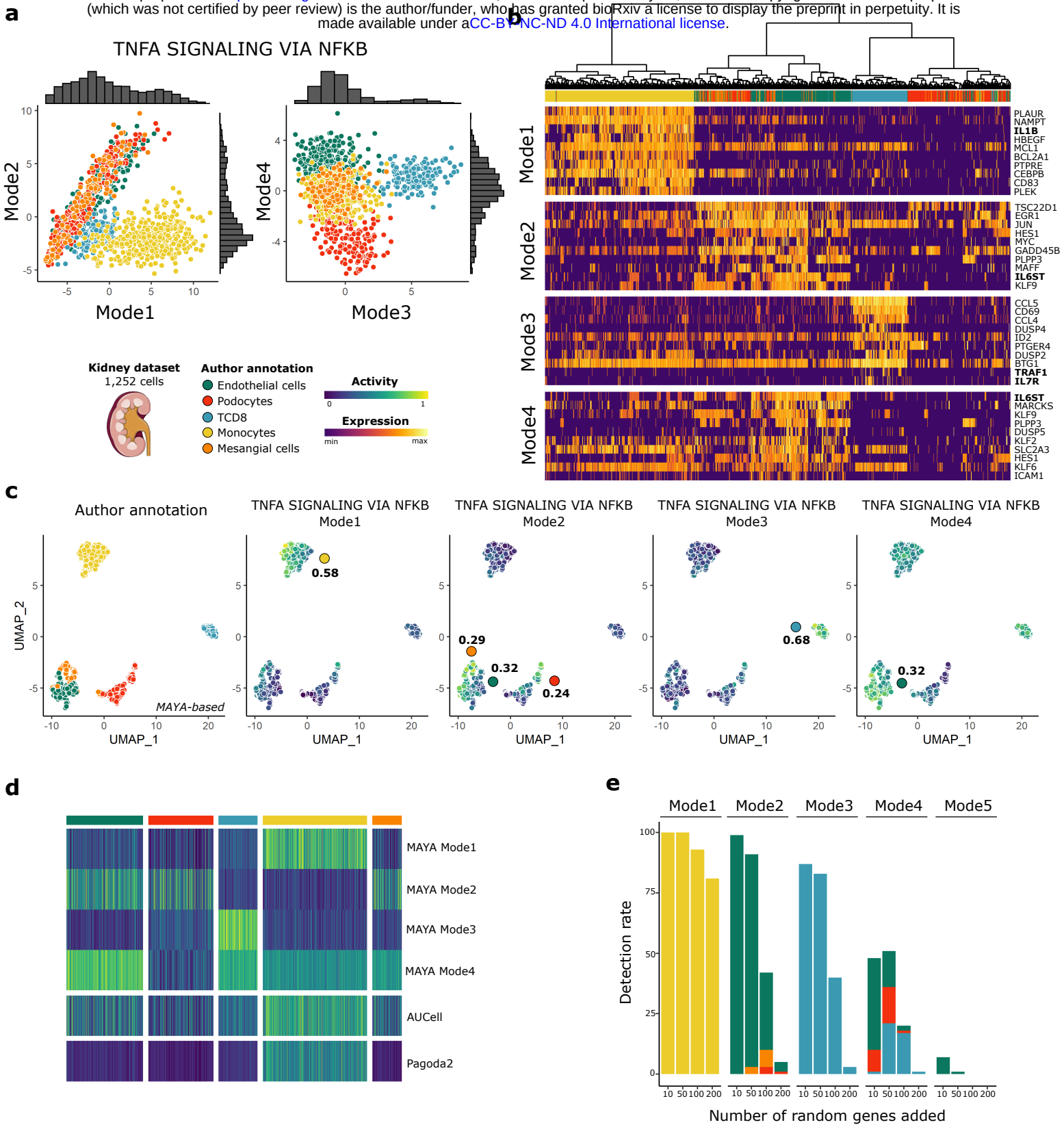


Fig.2: Activation modes of Hallmark pathways in kidney with MAYA

(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. **(b)** Scatterplot of Mode 2 versus Mode 1 cell activity scores. Associated density histograms are indicated on the sides of the graph. **(c)** 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. **(d)** 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. **(e)** Heatmap of activity scores computed by Pagoda2, AUCell and MAYA for *Allograft Rejection* pathway, cells are grouped according to author annotation. **(f)** 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. **(g)** Jitter representation of specificity scores of modes 1 and 2 grouped by level of added noise, datapoints are colored according to author annotation. Specificity obtained for each mode with initial gene list is represented with a dashed line.



Supplementary Fig.2:

(a) Scatterplot of Mode 2 versus Mode 1 and Mode 4 versus Mode 3 cell activity scores, for the pathway *TNFA signaling via NFKB* on the kidney dataset. Associated density histograms are indicated on the sides of the graphs. **(b)** Heatmap of scaled gene expression for top10 contributing genes for the four activation modes of *TNFA signaling via NFKB* pathway, ordered by decreasing contribution. **(c)** UMAP representation of activity matrix of Hallmark pathways, cells are colored according to author annotation, or activity scores of the four modes of *TNFA signaling via NFKB* pathway. Specificity score of cell populations is displayed next to relevant clusters. **(d)** Heatmap of activity scores computed by Pagoda2, AUCell and MAYA for *TNFA signaling via NFKB* pathway, cells are grouped according to author annotation. **(e)** Barplot representation of the detection rate of modes 1 to 5 for the pathway *TNFA signaling via NFKB* 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.

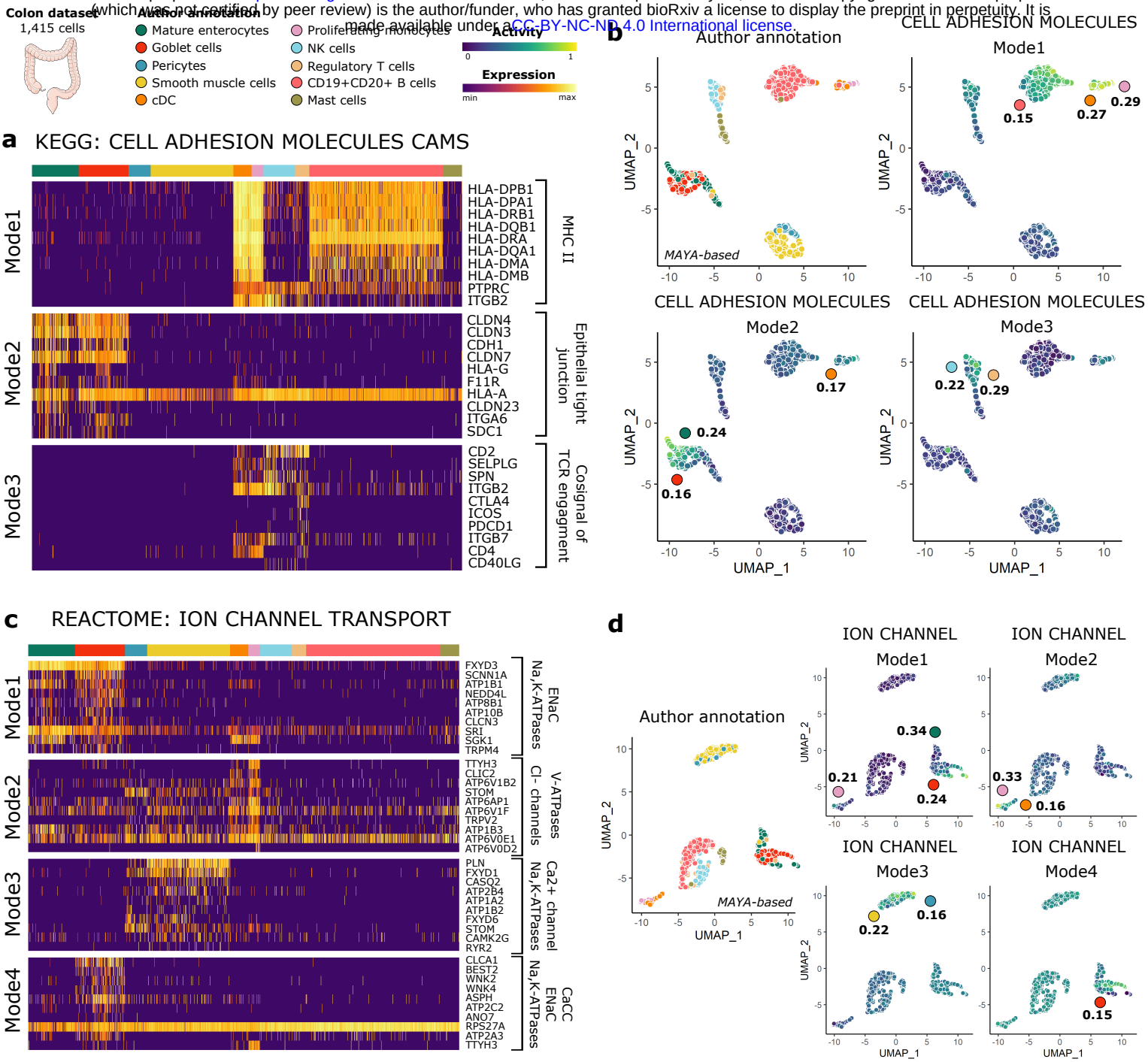
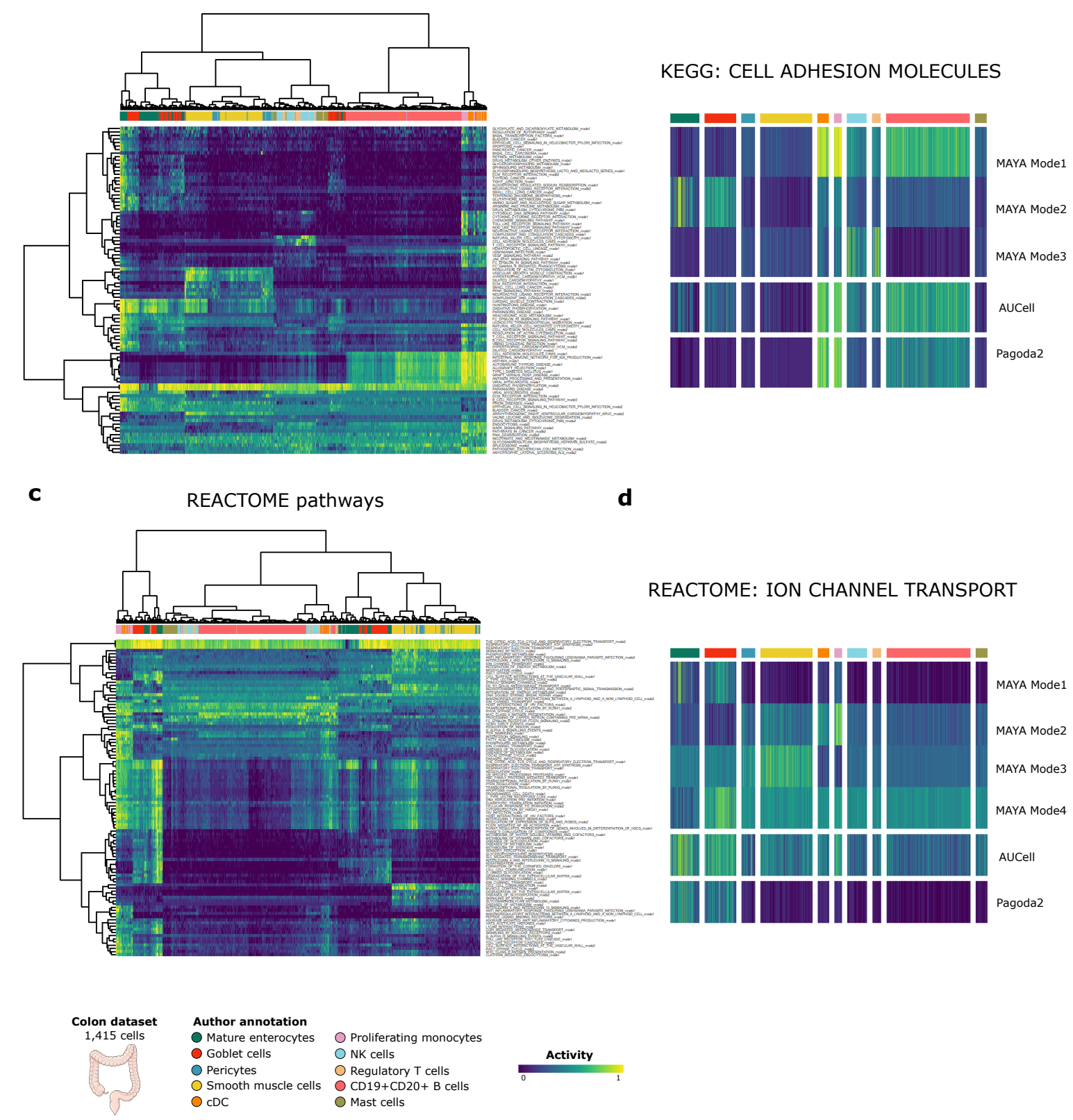


Fig.3: KEGG and REACTOME activity in colon with MAYA

(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. **(b)** 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. **(c)** Heatmap of scaled gene expression for top10 contributing genes for the four activation modes of the *Ion Channel Transport* pathway, ordered by decreasing contribution. **(d)** 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.



Supplementary Fig.3:
(a) Heatmap of activity matrix computed on colon dataset with MSigDB KEGG pathways, initial author annotation is indicated above heatmap. **(b)** Heatmap of activity scores computed by Pagoda2, AUCell and MAYA for KEGG *Cell Adhesion Molecules* pathway, cells are grouped according to author annotation. **(c)** Heatmap of activity matrix computed on colon dataset with MSigDB REACTOME pathways, initial author annotation is indicated above heatmap. **(d)** Heatmap of activity scores computed by Pagoda2, AUCell and MAYA for REACTOME *Ion Channel Transport* pathway, cells are grouped according to author annotation.

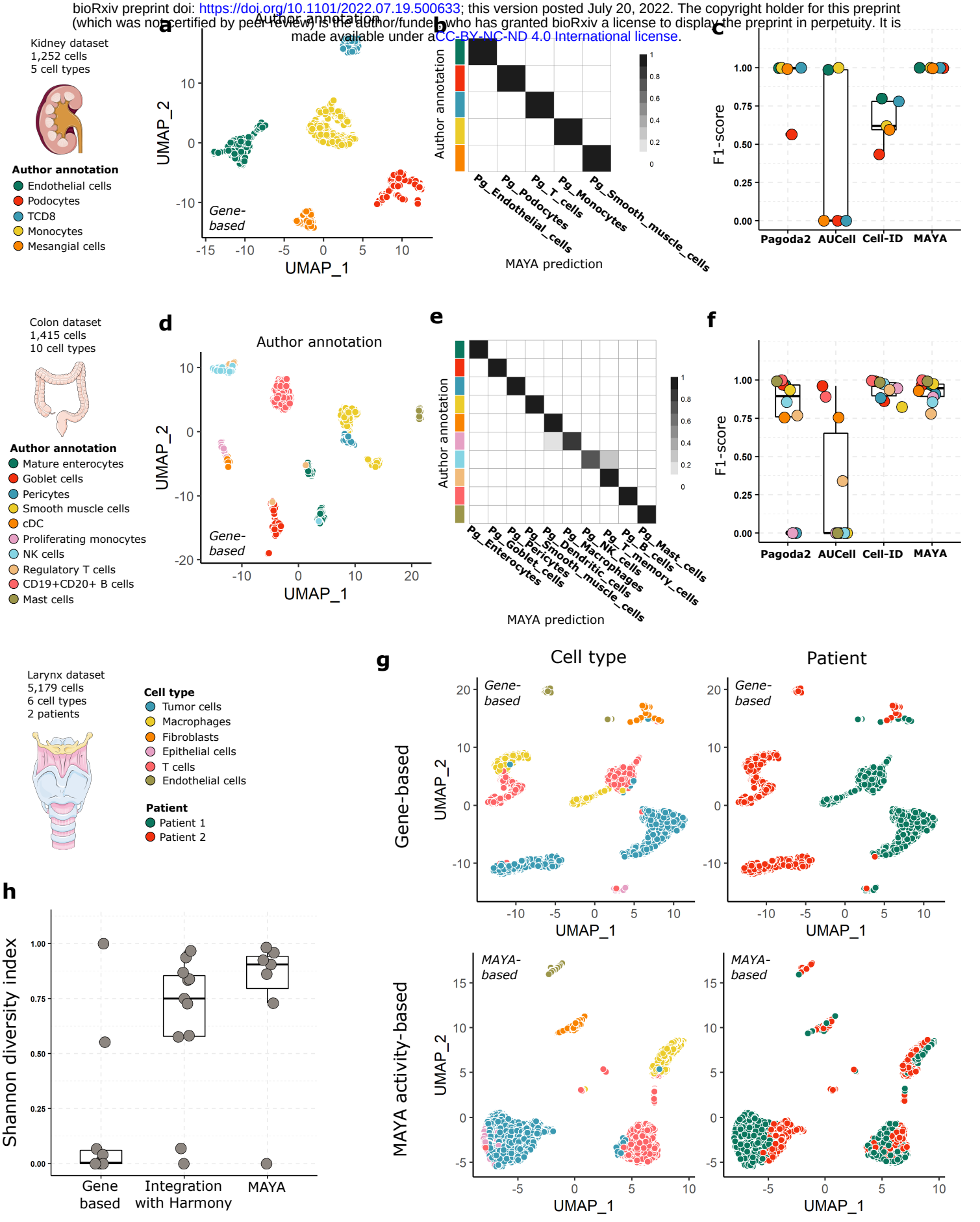
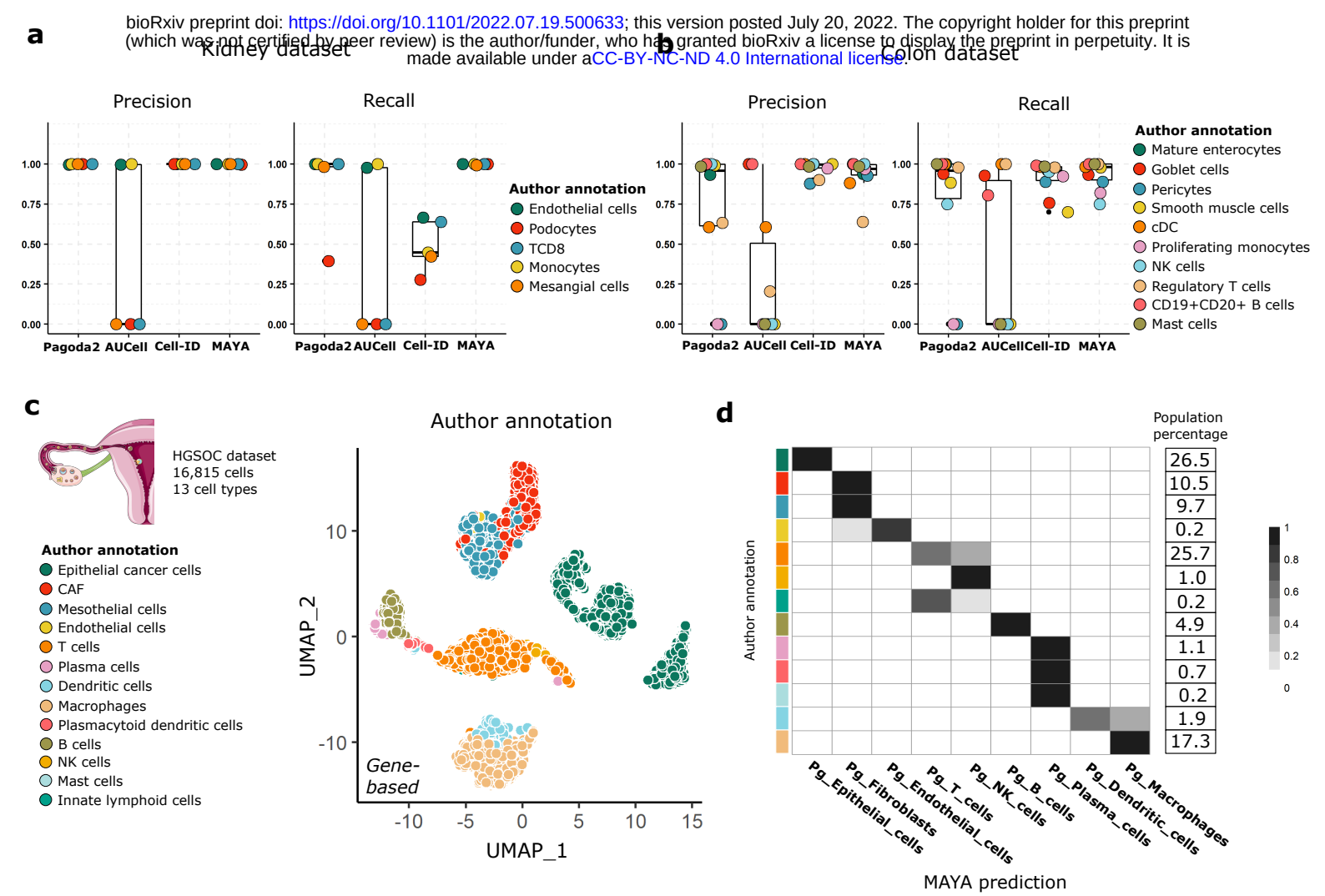
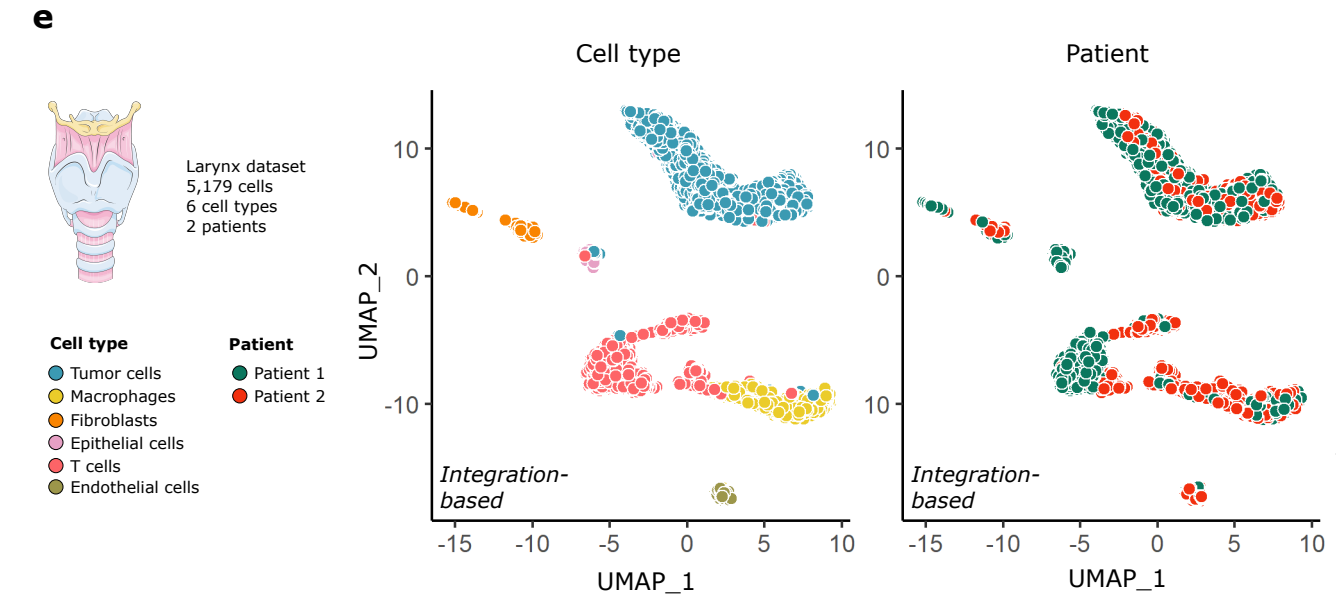


Fig.4: MAYA automatically annotates cell type

(a) Gene-based UMAP representation of kidney dataset, cells are colored according to author annotation. **(b)** Heatmap representing for each author annotation (rows) the fraction of cells labelled with each MAYA annotation (columns) for the kidney dataset. **(c)** Overlaid jitter and boxplot representation of F1-scores for automatic annotation of the kidney dataset using Pagoda2, AUCell, Cell-ID and MAYA, datapoints are colored according to author annotation. **(d)** Gene-based UMAP representation of colon dataset, cells are colored according to author annotation. **(e)** Heatmap representing for each author annotation (rows) the fraction of cells labelled with each MAYA annotation (columns) for the colon dataset. **(f)** Overlaid jitter and boxplot representation of F1-scores for automatic annotation of the colon dataset using Pagoda2, AUCell, Cell-ID and MAYA, datapoints are colored according to author annotation. **(g)** 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. **(h)** Overlaid jitter and boxplot representation of Shannon Diversity Index (SDI), for clusters derived from gene-based dimensionality reduction, Harmony dimensionality reduction and MAYA activity matrix of the larynx dataset.





Supplementary Fig.4:
(a,b) Overlaid jitter and boxplot representation of precision and recall for automatic annotation of the kidney and colon datasets using Pagoda2, AUCell, Cell-ID and MAYA, datapoints are colored according to author annotation. **(c)** Gene-based UMAP representation of the ovary dataset, cells are colored according to author annotation. **(d)** Heatmap representing for each author annotation (rows) the fraction of cells labelled with each MAYA annotation (columns) for the ovary dataset. The proportion of each author annotation in the dataset is indicated on the right side of the heatmap. **(e)** UMAP representation of larynx dataset integrated using Harmony, cells are colored according to cell type or patient.

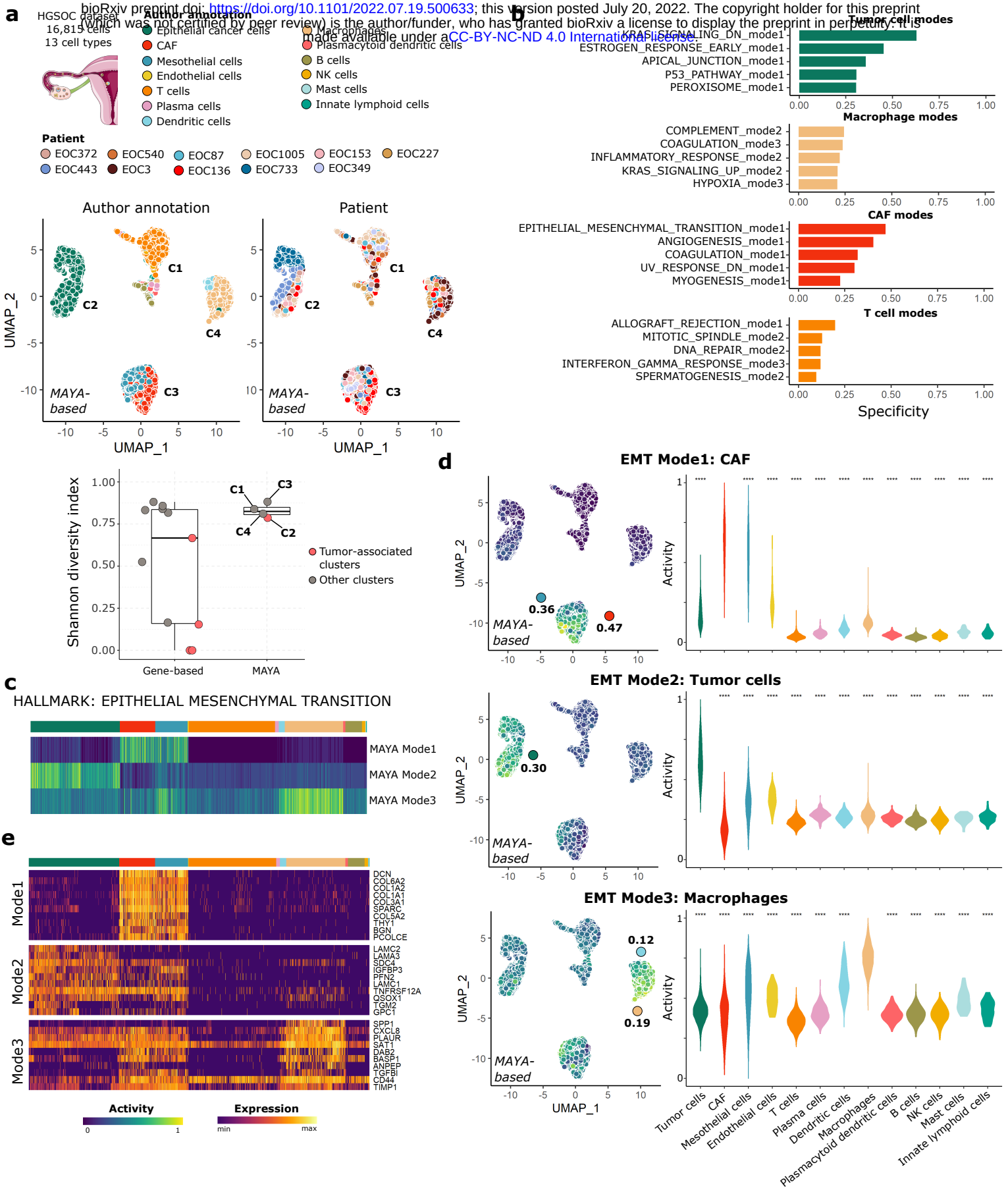
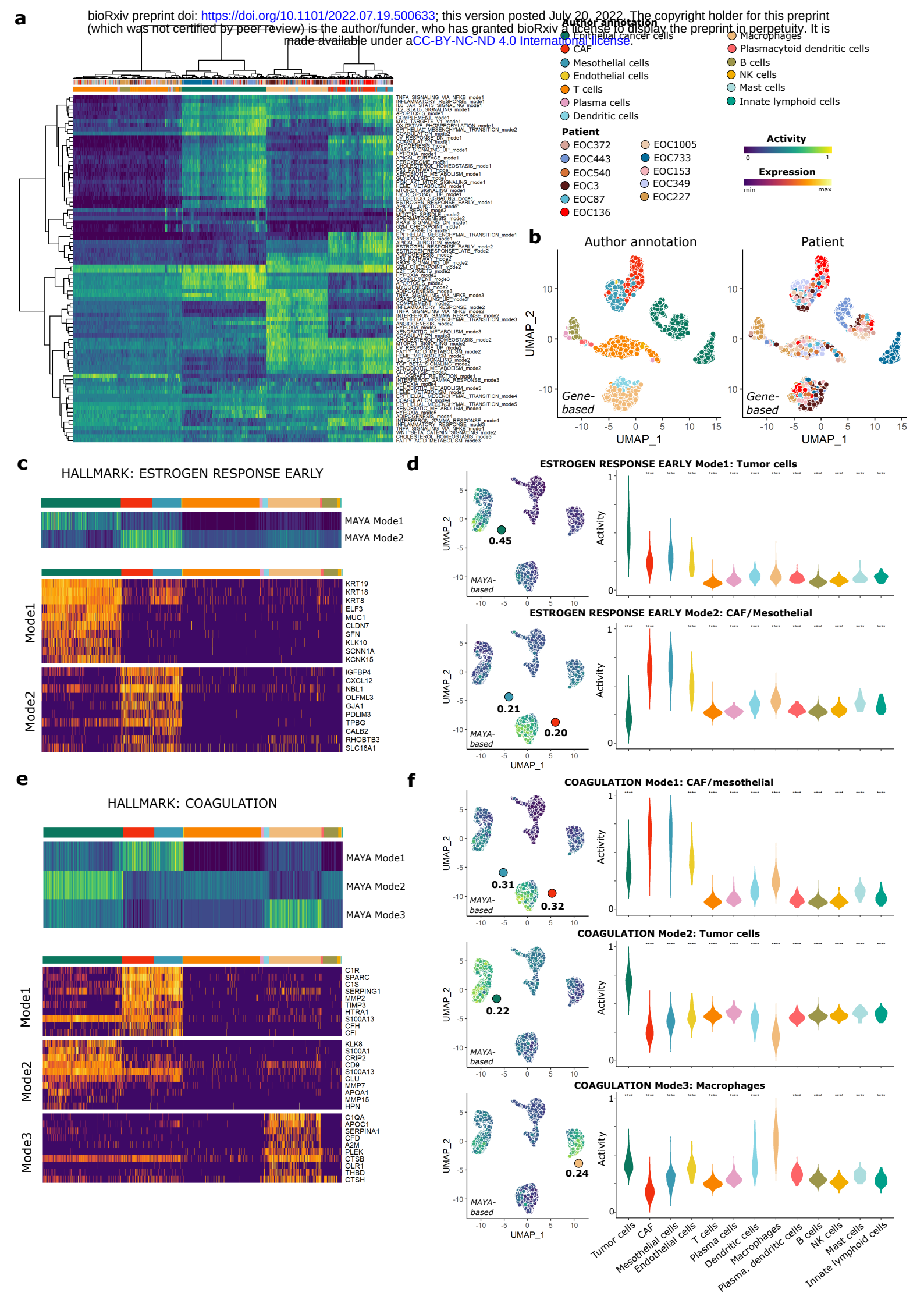


Fig.5: MAYA detects pathway activation in tumors across patients

(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. 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. **(b)** Barplot representation of specificity scores of the top5 specific modes for the four most prevalent populations in the dataset. **(c)** Heatmap of activity scores of the three modes of the Hallmark *Epithelial Mesenchymal Transition* (EMT) pathway, initial author annotation is indicated above heatmap. **(d)** 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 Wilcoxon test are symbolized with: * : <0.05, ** : <0.01, *** : <0.001, **** : <0.0001). **(e)** Heatmap of scaled gene expression for top10 contributing genes for the three modes of EMT, ordered by decreasing contribution.



Supplementary Fig. 5: **(a)** Heatmap of activity matrix computed on ovary dataset with MSigDB Hallmark pathways, initial author annotation is indicated above heatmap. **(b)** Gene-based UMAP representation of expression matrix, cells are colored according to author annotation and patient. **(c,e)** Heatmap of the activity scores for the two modes of Hallmark *Estrogen Response Early* pathway (respectively Hallmark *Coagulation* pathway), cells are grouped according to author annotation. Heatmap of scaled gene expression for top10 contributing genes for corresponding modes, ordered by decreasing contribution. **(d,f)** UMAP representation of activity matrix of Hallmark pathways, cells are colored according to activity scores of the two *Estrogen Response Early* modes (respectively three *Coagulation* 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 Wilcoxon test are symbolized with: * : <0.05, ** : <0.01, *** : <0.001, **** : <0.0001).