



**HAL**  
open science

# Multi-modal quantification of pathway activity with MAYA

Yuna Landais, Céline Vallot

► **To cite this version:**

Yuna Landais, Céline Vallot. Multi-modal quantification of pathway activity with MAYA. 2022. hal-03864031

**HAL Id: hal-03864031**

**<https://hal.science/hal-03864031>**

Preprint submitted on 22 Nov 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# 1 Multi-modal quantification of pathway activity with MAYA

2

3

4 Yuna Landais<sup>1</sup>, Céline Vallot<sup>2,3,4</sup>

5

6 1 One Biosciences, Paris, France

7 2 CNRS UMR3244, Institut Curie, PSL University, Paris, France

8 3 Translational Research Department, Institut Curie, PSL University, Paris, France

9 4 Single Cell Initiative, Institut Curie, PSL University, Paris, France

## 10 **Abstract**

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

23

24

25

26

27

28

29

30

31

32

33

34

35

36

## 37 Introduction

38 The identification of cell type and function is the driving force of a majority of single-cell studies. Such  
39 approaches are based on lists of canonical marker genes and pathway databases. Standard scRNA-seq  
40 analysis pipelines involve steps of dimensionality reduction and clustering before starting any marker  
41 or pathway analysis<sup>1-3</sup>, which makes the resulting conclusions highly dependent on the chosen  
42 algorithm and clustering parameters. In the case of oncogenic datasets, such clustering-based  
43 approaches appear inadequate to identify shared transcriptional programs across tumors as cancer  
44 cells tend to cluster independently per patient<sup>4-9</sup> rather than group by biological similarities. Several  
45 approaches have emerged, bypassing dimensionality reduction and clustering, by proposing to score  
46 pathway activity directly in individual cells rather than clusters. Such pooling of gene-based  
47 measurements into scores for gene lists has proven extremely powerful for the interpretation of sparse  
48 and noisy scRNA-seq datasets<sup>10,11</sup>. A recent benchmark<sup>12</sup> presented Pagoda2<sup>13</sup> and AUCell<sup>14</sup> as two of  
49 the top performing tools for pathway activity scoring. They are based on different scoring methods -  
50 AUCell estimates the proportion of highly expressed genes in each pathway while Pagoda2 uses the  
51 weights of the first principal component from Principal Component Analysis (PCA) – and each proposes  
52 a way to select significant scores. Nonetheless, both tools compute a unique activity score by pathway  
53 for all cells, implying that genes of a given signaling pathway should have coordinated expression  
54 across cell types.

55 Biological evaluation of pathway activation and more recently single-cell studies have repeatedly  
56 demonstrated the heterogeneity of cell functions depending on the biological context. Yet a majority  
57 of single-cell studies study pathway activation with single scores based on gene lists built from bulk  
58 data. Such curated gene lists represent the current reference biological knowledge and are the only  
59 available key to make biological sense of sparse and noisy scRNA-seq data. Adding more specialized  
60 curated gene lists to databases - detailing cellular functions according to cell identity - is ongoing but  
61 it will take some time to be completed. In order to already inspect existing pathway databases with  
62 single-cell resolution, we developed MAYA (Multimodes of pathWAY Activation), a tool that detects for  
63 each pathway the different modes of activation across cell types, each mode relying on different  
64 subsets of genes. We argue that MAYA could be a way for currently available biological knowledge to  
65 meet the granularity reached by single-cell data and help researchers go deeper in their understanding  
66 of complex cellular mechanisms. Particularly, in the case of oncogenic datasets, we show that MAYA  
67 can detect cell type specific modes of pathway activation for both the microenvironment and tumor  
68 cells, identifying common transcriptional programs across patients.

69

## 70 **Results**

### 71 **MAYA method**

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

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

88 However, all detected modes might not reflect a relevant biological pattern in the data and could be  
89 driven by outliers, either cells and/or genes, and this probability increases as modes explain less and  
90 less variance in the dataset. We thus developed a method to assess the informativity of each mode,  
91 based on two biologically interpretable criteria. First, an informative mode should be more active in a  
92 minimal subset of cells compared with other cells. This is assessed by detecting bimodal distributions  
93 of scores across cells and checking that the group of active cells represents more than a minimum  
94 fraction of the population, which can be determined based on previous knowledge of the underlying  
95 biology or set arbitrarily (Supplementary Fig.1a-c). Second, an informative mode should be driven by  
96 enough genes to be considered as a mode of activation per se and not solely correspond to the  
97 expression of a single outlier gene. To that end, we determined a cutoff for maximal variance of each  
98 gene of a mode, indicative of how much a gene can contribute on its own (Supplementary Fig. 1d,e).  
99 Default cutoff value was chosen to maximize the number of modes detected as informative while  
100 keeping a high average number of genes significantly contributing to each mode (Supplementary Fig.  
101 1f).

102 Although MAYA's main purpose is to detect multimodal activation of pathways, it can also perform  
103 unimodal activity scoring, to detect cell identity from any cell marker gene lists. To this end, we have  
104 developed a built-in function to leverage MAYA's scoring and informativity methods to automatically  
105 annotate cells in a dataset. This approach is based on activation of the first mode of provided cell type  
106 markers lists, using PanglaoDB<sup>16</sup> by default (Methods). This function allows cluster-free cell type  
107 annotation in a timely fashion as it annotates a dataset of around 16,000 cells in less than 1 minute  
108 and 125,000 cells in approximately 15 minutes (Supplementary Fig.1g).

109

### 110 **MAYA detects biologically relevant multimodal pathway activity in kidney**

111 The main distinguishing feature of MAYA over existing pathway activity scoring tools is the  
112 multimodality of its activity score, which proves useful when studying broad pathways in complex  
113 biological systems. We first sought to demonstrate its ability to detect cell-type specific activation  
114 modes of hallmark pathways. For that, we ran MAYA on a dataset of normal kidney and immune cells  
115 from Young et al.<sup>17</sup>, from which we selected cells from 5 distinct subtypes for clarity (n=1,252). We  
116 used the MSigDB Hallmark pathways<sup>18</sup> as input gene lists, covering main biological functions.  
117 Unsupervised clustering on the multimodal activity matrix shows MAYA detects modes that distinguish  
118 different cell populations (Fig. 2a). More specifically, we noticed that modes from the same pathway  
119 were specifically activated in different cell types. As an example, the *Allograft rejection* pathway  
120 presents two modes of activation (Fig. 2b-d): (i) mode 1, driven by the expression of *CTSS* and *SPI1* -  
121 known to have a critical role in antigen presentation<sup>19</sup> and gene regulation during myeloid  
122 development<sup>20</sup> - and specific to monocytes (specificity of 0.57), and (ii) mode 2, driven by *CD2*, *CD3E*  
123 and *CD3D* - coding for T cell surface proteins - and by *CD8A* and *CD8B* - coding for the CD8 antigen -  
124 and specific to CD8 T cells (specificity of 0.88). In contrast, AUCell and Pagoda2 both describe this  
125 pathway with a single score, corresponding to an aggregation of MAYA's mode 1 and 2, or mode 1 only  
126 respectively (Fig. 2e). Another detailed example is shown in Supplementary Figure 2 for the *TNFA*  
127 *signaling via NFKB* pathway, where four activation modes were detected with MAYA based on their  
128 bimodal activity distribution (Supplementary Fig. 2a): one specific to monocytes, one to CD8 T cells  
129 and two to endothelial cells (Supplementary Fig. 2b-d). Interestingly, each mode involves a different  
130 interleukin specific to the population in which the mode is found to be active: (i) *IL6ST* is a signal  
131 transducer, which dimerizes with *IL6R* and bound for instance by IL-6, resulting in the activation of  
132 downstream cascades in endothelial cells<sup>21</sup>, (ii) *IL1B* is a lymphocyte activating factor produced by  
133 monocytes, macrophages and neutrophils, and (iii) *IL7R* is associated with T cell differentiation.

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

136 To test both the stability and the ability of MAYA to detect biologically relevant signal in noisy gene  
137 lists, we added 10, 50, 100 and 200 random genes to the initial 200 genes of the pathways *Allograft*  
138 *rejection* and *TNFA signaling via NFkB*; each experiment was repeated a 100 times. For the *Allograft*  
139 *Rejection* pathway, the two initial activation modes were detected for all modified gene lists with a  
140 high cell-type specificity, whatever the level of added noise (Fig. 2f,g). These results also show the  
141 accuracy of our selection method to detect relevant modes, as we rarely detect additional activation  
142 modes (corresponding to PC3/mode 3) even when randomly increasing the reference gene lists.  
143 Similarly, for the *TNFA signaling* pathway, the first three modes are robust to noise, with a decrease in  
144 sensitivity of detection when adding more than 100 unrelated genes (Supplementary Fig. 2e).

145

#### 146 **MAYA detects biologically relevant multimodal pathway activity in colon**

147 We then illustrated the relevance of the biological insight gained by using multimodal pathway analysis  
148 for another tissue with a dataset of colon and immune cells from Lee et al.<sup>22</sup> - from which we selected  
149 cells from 10 distinct cell types (n=1,415) - and using the MSigDB KEGG and REACTOME pathways<sup>23</sup>.  
150 Both analyses recover cell-type specific activation modes, given the clustering of cells by cell type on  
151 the heatmaps derived from the activity matrix (Supplementary Fig. 3a,c). Focusing on KEGG *cell*  
152 *adhesion molecules* list, we observed that MAYA was able to detect several well-known types of cell-  
153 cell adhesion processes starting from the mixed general reference list (Fig. 3a,b and Supplementary  
154 Fig. 3b): (i) mode 1 driven by the expression of HLA genes coding MHC class II molecules<sup>24</sup>, detected in  
155 antigen-presenting cells - monocytes and dendritic cells - and B cells, with a specificity of 0.29, 0.27  
156 and 0.15 respectively, (ii) mode 2 driven by the expression of genes coding for claudins and cadherins  
157 located at tight junctions<sup>25,26</sup>, specifically activated in epithelial cells (specificity of 0.24 and 0.16 for  
158 enterocytes and goblet cells respectively), and (iii) mode 3 driven by the expression of T cell membrane  
159 molecules, specific to Regulatory T cells (specificity of 0.29).

160 Applying MAYA to the REACTOME pathway *ion channel transport*, we were able to detect different  
161 types of ion channels and functions, specific to each cell populations (Fig. 3c,d and Supplementary Fig.  
162 3d). Mode 1 is specific to colon epithelial cells (specificity of 0.34 and 0.24 for enterocytes and goblet  
163 cells respectively, Fig. 3d) and corresponds to two types of ion channels – Epithelial Sodium Channel  
164 (ENaCs) and Na,K-ATPase<sup>27</sup> - that have been shown to participate to the regulation of salt and water  
165 absorption from the colon lumen<sup>28,29</sup>. In particular, activation mode 1 captures genes regulating ENaCs  
166 and their residence at the apical membrane: *SCNN1A* encodes a subunit of ENaCs<sup>30</sup>, *NEDD4L*

167 participates to ENaCs ubiquitination which leads to their retrieval from cell surface<sup>31</sup> and *SGK1* is known  
168 to phosphorylate NEDD4L product, which decreases its binding to ENaCs<sup>32,33</sup>. Mode 4 is specific to  
169 goblet cells only, driven by the expression of the genes *CLCA1* and *BEST2*. These two genes are  
170 associated with Calcium-activated Chloride Channels (CaCCs) that have been shown to participate in  
171 epithelial secretion<sup>34</sup>. Mode 3 is specific to pericytes and smooth muscle cells (specificity of 0.16 and  
172 0.22 respectively) and is associated with Calcium homeostasis (*ATP2B4*, *PLN*, *CASQ2*) and Na,K-ATPases  
173 (*FXVD1*, *FXVD6*, *ATP1A2*, *ATP1B2*), two important channels for the membrane polarization of  
174 contractile cells. Finally, mode 2, mainly active in monocytes and dendritic cells (specificity of 0.33 and  
175 0.16 respectively), involves genes associated with acidification of intracellular organelles through  
176 colocalization of V-type proton ATPases<sup>35</sup> (*ATP6V1B2*, *ATP6AP1*, *ATP6V1F*, *ATP6V0E1*, *ATP6V0D2*<sup>36</sup>)  
177 and Chloride channels<sup>37</sup> (*TTYH3*, *CLIC2*), a process necessary for phagocytosis. Altogether, as for the  
178 kidney, starting from reference databases, MAYA untangles pathway activities specific to each cell  
179 type, revealing precise cell functions.

180

#### 181 **MAYA automatically assigns cell identity**

182 We then leveraged MAYA's scoring and selection ability to automatically and robustly assign cell  
183 identity. We applied MAYA to PanglaoDB cell type marker lists and the subsets of kidney and colon  
184 datasets used previously (Fig. 4a,d). We demonstrate that MAYA enabled an automated and accurate  
185 annotation of each cell in the two datasets, using the initial cell type annotation by authors as a  
186 reference (Fig. 4b,e). We compared the accuracy of our predictions with the ones obtained with three  
187 other algorithms: AUCell<sup>14</sup>, Pagoda2<sup>13</sup> and Cell-ID<sup>38</sup>, a cell type identification method based on Multiple  
188 Correspondence Analysis (MCA). MAYA presents among the highest rates of recall and precision for  
189 both datasets (Fig. 4c,f and Supplementary Fig. 4a,b). We finally tested the scalability of MAYA and its  
190 ability to detect rare cell types on a dataset with 16,815 cells from ovarian tumors<sup>6</sup> (Supplementary  
191 Fig. 4c). Overall, MAYA had an average precision of 51% and recall of 68%. Notably, B cells were  
192 identified with a precision and recall of 98% when they represent only 4.9% of the dataset and  
193 endothelial cells with a precision of 100% and recall of 85% when they represent 0.2% of cells in the  
194 dataset (Supplementary Fig. 4d). Lower precision is achieved for some types probably due to overlap  
195 between cell type markers in PanglaoDB, such as between NK cells and T cells (28 shared markers out  
196 of 80 and 95 markers respectively), dendritic cells and macrophages (34 shared out of 121 and 128  
197 markers respectively), and endothelial cells and fibroblasts (13 shared out of 187 and 171 markers  
198 respectively). All three pairs of cell types share more genes than with any other type from the  
199 PanglaoDB lists.



200 Furthermore, as batch effect is a main concern in single-cell analyses, notably for data visualization  
201 and cell annotation, we tested whether MAYA was affected by such technical biases. We worked on a  
202 dataset containing  $n=5,179$  cells from laryngeal squamous cell carcinoma biopsies of 2 patients with a  
203 batch effect between patients<sup>39</sup>. Using standard gene-based scRNA-seq matrix processing, cells from  
204 the same cell types – whether cells from the microenvironment or the tumors – indeed cluster by  
205 patient whereas clustering on the MAYA activity matrix groups cells by cell type, with cells from both  
206 patients within the same cluster (Fig. 4g). To quantify the inter-patient overlap between clusters of  
207 similar cell types, we computed the Shannon Diversity Index (SDI) for both methods as well as for  
208 clusters obtained with the reference integration tool Harmony<sup>40</sup> (Supplementary Fig. 4e). MAYA had  
209 an average SDI of 0.77 against 0.65 and 0.17 for the integration-based and the gene-based method  
210 respectively (Fig 4h). In addition to pathway scoring, MAYA can perform accurate cell type annotation  
211 independently of batch effect, making it an all-in-one tool to address both cell identity and function.  
212

### 213 **MAYA detects common modes of pathway activation across cancer patients**

214 Patient-specificity of cancer cells is currently a major limitation for the comprehensive study of  
215 oncogenic scRNA-seq datasets. Cells of the microenvironment coming from different patients can  
216 easily group together, showing the absence of a major batch effect between samples, while tumor  
217 cells form distinct clusters<sup>4-9</sup>. Such behavior is thought to be due in part to the genetic variations across  
218 tumor cells from different patients, notably copy-number variations. Integration methods, correcting  
219 for general batch effect in samples, such as Harmony<sup>40</sup>, are not suited to deal with such cell-type  
220 specific effect.

221 We demonstrate here that MAYA can be an alternative to gene-based or integration-based methods  
222 to identify common transcriptional features between cancer cells across patients. Using an ovarian  
223 cancer dataset, we show that MAYA identifies several modes of pathway activation shared across  
224 patients (Fig. 5a,b and Supplementary Fig. 5a,b) that are associated with known cancer hallmarks.  
225 Indeed, top specific modes of epithelial cancer cells reflect the expression of targets of the oncogene  
226 KRAS, genes associated with early response to estrogen or the P53 pathway (specificity of 0.63, 0.45  
227 and 0.31 respectively), that all relate to tumor growth and proliferation (Fig. 5b). MAYA also identifies  
228 modes of pathway activation specific to tumor microenvironment populations. It notably detects a  
229 cell-type specific activation of complement genes in macrophages (specificity of 0.24) and of  
230 angiogenesis-related genes in cancer-associated fibroblasts (CAFs) (specificity of 0.40).

231 MAYA multimodality allows to untangle several cell-type specific modes of activation for biological  
232 phenomena that are commonly difficult to sort out between cell populations within the tumors and  
233 their microenvironment. For example, MAYA detects different modes of *epithelial-to-mesenchymal*

234 *transition* (EMT) (Fig. 5c): mode 1 specific to CAFs/mesothelial cells (specificity of 0.47 and 0.36  
235 respectively), mode 2 specific to tumor cells (specificity of 0.30) and mode 3 to macrophages  
236 (specificity of 0.19) (Fig. 5d). MAYA identifies a combination of genes that characterizes EMT occurring  
237 in epithelial cells, with *LAMA3* and *LAMC2* being exclusive to this cell type (Fig. 5e). These two genes  
238 expressed by basal epithelium code for two subunits of laminin 332, an essential component of  
239 epithelial basement membrane that promotes tumor cell motility<sup>41,42</sup>. In CAFs, MAYA detects EMT as  
240 driven mainly by genes encoding proteins from the extracellular matrix (ECM) including collagens,  
241 which have been shown to promote EMT in the tumor microenvironment directly<sup>43</sup> or by increasing  
242 the ECM stiffness<sup>44,45</sup>. A third mode of EMT, characterized by the expression of the gene *SPP1*, is found  
243 in macrophages; macrophages have indeed been shown to be involved in EMT induction in various  
244 types of cancer<sup>46-49</sup>. Two additional modes are detected but are not as cell-type specific as the others  
245 (Supplementary Fig. 5a, maximum specificity scores of 0.12).

246 MAYA also identifies two different modes of activation of the *estrogen response early cascade*  
247 (Supplementary Fig. 5c,d), one specific to tumor cells, and one specific to CAFs, consistent with the  
248 observation that CAFs can use ER-mediated signaling pathways to promote tumor cell proliferation<sup>50,51</sup>.

249 MAYA also helps to untangle the respective contribution of cancer cells and its microenvironment to  
250 the hemostatic imbalance observed in cancer<sup>52,53</sup>, by detecting *coagulation* modes with high specificity  
251 for CAFs and mesothelial cells (0.31 and 0.32), tumor cells (0.22) and macrophages (0.24)  
252 (Supplementary Fig. 5e,f).

253 Altogether, MAYA appears extremely powerful to detect modes of pathway activation across tumor  
254 cells from different patients as well as within the microenvironment - novel combinations of genes  
255 within known global reference gene lists. We see with these examples that MAYA can discover refined  
256 gene lists, specific to each population, matching the biological interpretation of pathway activation to  
257 the granularity of the single-cell measurements.

## 258 Discussion

259 MAYA sorts out the different modes of pathway activation specific to each cell type, by automatically  
260 detecting gene subgroups within reference pathways and computing several scores of pathway  
261 activation. We show that MAYA leverages existing biological knowledge to extract cell-type specific  
262 ways of activating pathways from single-cell datasets. In addition to pathway analysis, MAYA also  
263 performs automated cell typing as a side function, making it an all-in-one tool for both cell type and  
264 cell function identification. MAYA proves particularly useful for single-cell cancer datasets, by (i)  
265 identifying common modes of pathway activations across patients in tumor cells, and also by (ii)  
266 dissecting the contribution of each population – fibroblast, immune & tumor cell – to the activation of  
267 a given pathway.

268 In comparison to previously published methods (AUCell<sup>14</sup>, Pagoda2<sup>13</sup>, ROMA<sup>54</sup> and UCell<sup>55</sup>), MAYA  
269 provides multiple activation scores per pathway, and in a time efficient and user-friendly way. Indeed,  
270 running Pagoda2 for example can quickly become computationally intensive; its selection method  
271 requires to build a null distribution for each pathway by retrieving variance explained by PC1 for  
272 random gene lists of the same pathway size – which drastically increases the number of PCA run to  
273 score a single pathway. AUCell computes several bimodality thresholds by pathway, which can also  
274 increase computing time, and needs rather advanced users to tune its technical parameters if default  
275 ones do not provide satisfying results. With MAYA, we simplified bimodal detection by focusing on  
276 inflection points and introducing two biologically interpretable parameters, easily tuned by users: (i)  
277 minimum proportion of cells that should activate a mode for the mode to be considered relevant and  
278 (ii) a maximum contribution to a mode that a single gene can have.

279 We have also challenged the robustness to noise of our scoring and informativity methods and showed  
280 MAYA can detect relevant biological signal from noisy pathway lists. It can prove very useful as we  
281 know pathway and cell markers manual curation is very time-consuming. Here, we argue that MAYA  
282 can take as input non-curated and potentially very exhaustive pathway or cell type lists and detect  
283 biological signal if they contain any.

284 We also leveraged our methods of scoring and selection of informative scores to propose a built-in  
285 function to automatically annotate cells using PanglaoDB cell type markers lists. This method performs  
286 better with MAYA scores than Pagoda2 or AUCell scores and has performance results equivalent to or  
287 better than Cell-ID<sup>38</sup>, a package specialized in cell type annotation. MAYA is scalable to large datasets  
288 (>100,000 cells, in 15 minutes) and it is able to accurately detect and annotate cell populations  
289 representing less than 5% of cells. MAYA is therefore an all-in-one tool proposing both cell type  
290 identification - like Cell-ID<sup>38</sup>, CellTypist<sup>56</sup> and scGate<sup>57</sup> - and multi-modal pathway analysis.

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

296 Finally, MAYA brings particular biological insights when studying single-cell datasets from cancer  
297 patients that do not suffer from batch effect on all cell types but from patient-specificity for tumor  
298 cells. There is currently no standard way to address this challenge for data interpretation and a growing  
299 need to understand common cancer features across patients. Recently, Gavish et al.<sup>15</sup> provided the  
300 community with clues about shared transcriptional programs across patient and tumor types by  
301 describing 41 “meta-programs” grouped in 11 hallmarks of intra-tumor heterogeneity. These “meta-  
302 programs” were inferred de novo by studying scRNA-seq from multiple tissues and cancer types. This  
303 approach is very complementary to ours, where we interrogate existing knowledge. MAYA identifies  
304 common modes of activation across tumor cells, which could be compared to such tumor meta-  
305 programs. In addition, MAYA deciphers the respective contribution of each cell population to the  
306 activation of a given pathway, by defining the ensemble of genes that drive the pathway activity in  
307 each contributing population. Both inter and intra-patient features of MAYA will enable the  
308 identification of shared therapeutic vulnerabilities across patients, as well as various strategies to  
309 target them within the tumor eco-system.

310

311

312

313

314

## 315 References

- 316 1. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial.  
317 *Mol. Syst. Biol.* **15**, e8746 (2019).
- 318 2. Pereira, W. J. *et al.* Asc-Seurat: analytical single-cell Seurat-based web application. *BMC*  
319 *Bioinformatics* **22**, 556 (2021).
- 320 3. Prieto, C., Barrios, D. & Villaverde, A. SingleCAnalyzer: Interactive Analysis of Single Cell RNA-  
321 Seq Data on the Cloud. *Front. Bioinforma.* **2**, (2022).
- 322 4. Wu, F. *et al.* Single-cell profiling of tumor heterogeneity and the microenvironment in  
323 advanced non-small cell lung cancer. *Nat. Commun.* **12**, 2540 (2021).
- 324 5. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell  
325 RNA-seq. *Science* **352**, 189–196 (2016).
- 326 6. Zhang, K. *et al.* Longitudinal single-cell RNA-seq analysis reveals stress-promoted  
327 chemoresistance in metastatic ovarian cancer. *Sci. Adv.* **8**, (2022).
- 328 7. Maynard, A. *et al.* Therapy-Induced Evolution of Human Lung Cancer Revealed by Single-Cell  
329 RNA Sequencing. *Cell* **182**, 1232-1251.e22 (2020).
- 330 8. Neftel, C. *et al.* An Integrative Model of Cellular States, Plasticity, and Genetics for  
331 Glioblastoma. *Cell* **178**, 835-849.e21 (2019).
- 332 9. Puram, S. V *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor  
333 Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624.e24 (2017).
- 334 10. Wang, S. *et al.* Single-Cell Transcriptomic Atlas of Primate Ovarian Aging. *Cell* **180**, 585-  
335 600.e19 (2020).
- 336 11. Ramirez, A. K. *et al.* Single-cell transcriptional networks in differentiating preadipocytes  
337 suggest drivers associated with tissue heterogeneity. *Nat. Commun.* **11**, 2117 (2020).
- 338 12. Zhang, Y. *et al.* Benchmarking algorithms for pathway activity transformation of single-cell  
339 RNA-seq data. *Comput. Struct. Biotechnol. J.* **18**, 2953–2961 (2020).
- 340 13. Fan, J. *et al.* Characterizing transcriptional heterogeneity through pathway and gene set  
341 overdispersion analysis. *Nat. Methods* (2016) doi:10.1038/nmeth.3734.
- 342 14. Aibar, S. *et al.* SCENIC: Single-cell regulatory network inference and clustering. *Nat. Methods*  
343 **14**, 1083–1086 (2017).
- 344 15. Gavish, A. *et al.* The transcriptional hallmarks of intra-tumor heterogeneity across a thousand  
345 tumors. *bioRxiv* 2021.12.19.473368 (2021) doi:10.1101/2021.12.19.473368.
- 346 16. Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of  
347 mouse and human single-cell RNA sequencing data. *Database* **2019**, 46 (2019).
- 348 17. Young, M. D. *et al.* Single-cell transcriptomes from human kidneys reveal the cellular identity  
349 of renal tumors. *Science (80-. )*. **361**, 594–599 (2018).
- 350 18. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.*  
351 **1**, 417–425 (2015).
- 352 19. Nakagawa, T. Y. & Rudensky, A. Y. The role of lysosomal proteinases in MHC class II-mediated  
353 antigen processing and presentation. *Immunol. Rev.* **172**, 121–129 (1999).

- 354 20. Klemsz, M. J., McKercher, S. R., Celada, A., Van Beveren, C. & Maki, R. A. The macrophage and  
355 B cell-specific transcription factor PU.1 is related to the ets oncogene. *Cell* **61**, 113–124  
356 (1990).
- 357 21. Kang, S. & Kishimoto, T. Interplay between interleukin-6 signaling and the vascular  
358 endothelium in cytokine storms. *Exp. Mol. Med.* **53**, 1116–1123 (2021).
- 359 22. Lee, H. O. *et al.* Lineage-dependent gene expression programs influence the immune  
360 landscape of colorectal cancer. *Nat. Genet.* **52**, 594–603 (2020).
- 361 23. Liberzon, A. *et al.* Databases and ontologies Molecular signatures database (MSigDB) 3.0.  
362 *Bioinforma. Appl. NOTE* **27**, 1739–1740 (2011).
- 363 24. Handunnetthi, L., Ramagopalan, S. V, Ebers, G. C. & Knight, J. C. Regulation of major  
364 histocompatibility complex class II gene expression, genetic variation and disease. *Genes*  
365 *Immun.* **11**, 99–112 (2010).
- 366 25. Tsukita, S., Tanaka, H. & Tamura, A. The Claudins: From Tight Junctions to Biological Systems.  
367 *Trends Biochem. Sci.* **44**, 141–152 (2019).
- 368 26. Braga, V. Epithelial cell shape: cadherins and small GTPases. *Exp. Cell Res.* **261**, 83–90 (2000).
- 369 27. Bibert, S. *et al.* A link between FXD3 (Mat-8)-mediated Na,K-ATPase regulation and  
370 differentiation of Caco-2 intestinal epithelial cells. *Mol. Biol. Cell* **20**, 1132–1140 (2009).
- 371 28. Rajendran, V. M., Schulzke, J.-D. & Seidler, U. E. Chapter 58 - Ion Channels of the  
372 Gastrointestinal Epithelial Cells. in (ed. Said, H. M. B. T.-P. of the G. T. (Sixth E.) 1363–1404  
373 (Academic Press, 2018). doi:<https://doi.org/10.1016/B978-0-12-809954-4.00058-X>.
- 374 29. Kunzelmann, K. & Mall, M. Electrolyte transport in the mammalian colon: mechanisms and  
375 implications for disease. *Physiol. Rev.* **82**, 245–289 (2002).
- 376 30. Saxena, A. *et al.* Gene Structure of the Human Amiloride-Sensitive Epithelial Sodium Channel  
377 Beta Subunit. *Biochem. Biophys. Res. Commun.* **252**, 208–213 (1998).
- 378 31. Zhou, R., Patel, S. V & Snyder, P. M. Nedd4-2 catalyzes ubiquitination and degradation of cell  
379 surface ENaC. *J. Biol. Chem.* **282**, 20207–20212 (2007).
- 380 32. Lang, F. *et al.* Regulation of channels by the serum and glucocorticoid-inducible kinase -  
381 implications for transport, excitability and cell proliferation. *Cell. Physiol. Biochem. Int. J. Exp.*  
382 *Cell. Physiol. Biochem. Pharmacol.* **13**, 41–50 (2003).
- 383 33. Snyder, P. M. Minireview: Regulation of Epithelial Na<sup>+</sup> Channel Trafficking. *Endocrinology* **146**,  
384 5079–5085 (2005).
- 385 34. Gruber, A. D. *et al.* Genomic cloning, molecular characterization, and functional analysis of  
386 human CLCA1, the first human member of the family of Ca<sup>2+</sup>-activated Cl<sup>-</sup> channel proteins.  
387 *Genomics* **54**, 200–214 (1998).
- 388 35. Grinstein, S., Nanda, A., Lukacs, G. & Rotstein, O. V-ATPases in phagocytic cells. *J. Exp. Biol.*  
389 **172**, 179–192 (1992).
- 390 36. Xia, Y. *et al.* The macrophage-specific V-ATPase subunit ATP6V0D2 restricts inflammasome  
391 activation and bacterial infection by facilitating autophagosome-lysosome fusion. *Autophagy*  
392 **15**, 960–975 (2019).
- 393 37. Carraro-Lacroix, L. R., Lessa, L. M. A., Fernandez, R. & Malnic, G. Physiological implications of  
394 the regulation of vacuolar H<sup>+</sup>-ATPase by chloride ions. *Brazilian J. Med. Biol. Res.* **42**, 155–163  
395 (2009).

- 396 38. Cortal, A., Martignetti, L., Six, E. & Rausell, A. Gene signature extraction and cell identity  
397 recognition at the single-cell level with Cell-ID. *Nat. Biotechnol.* **39**, 1095–1102 (2021).
- 398 39. Song, L. *et al.* Cellular heterogeneity landscape in laryngeal squamous cell carcinoma. *Int. J.*  
399 *Cancer* **147**, 2879–2890 (2020).
- 400 40. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony.  
401 *Nat. Methods* **16**, 1289–1296 (2019).
- 402 41. Domogatskaya, A., Rodin, S. & Tryggvason, K. Functional diversity of laminins. *Annu. Rev. Cell*  
403 *Dev. Biol.* **28**, 523–553 (2012).
- 404 42. Carpenter, P. M. *et al.* Migration of breast cancer cell lines in response to pulmonary laminin  
405 332. *Cancer Med.* **6**, 220–234 (2017).
- 406 43. Wei, S. C. *et al.* Matrix stiffness drives epithelial–mesenchymal transition and tumour  
407 metastasis through a TWIST1–G3BP2 mechanotransduction pathway. *Nat. Cell Biol.* **17**, 678–  
408 688 (2015).
- 409 44. Shintani, Y., Hollingsworth, M. A., Wheelock, M. J. & Johnson, K. R. Collagen I promotes  
410 metastasis in pancreatic cancer by activating c-Jun NH(2)-terminal kinase 1 and up-regulating  
411 N-cadherin expression. *Cancer Res.* **66**, 11745–11753 (2006).
- 412 45. Koenig, A., Mueller, C., Hasel, C., Adler, G. & Menke, A. Collagen type I induces disruption of E-  
413 cadherin-mediated cell-cell contacts and promotes proliferation of pancreatic carcinoma  
414 cells. *Cancer Res.* **66**, 4662–4671 (2006).
- 415 46. Liu, J. *et al.* Association of tumour-associated macrophages with cancer cell EMT, invasion,  
416 and metastasis of Kazakh oesophageal squamous cell cancer. *Diagn. Pathol.* **14**, 55 (2019).
- 417 47. Su, S. *et al.* A Positive Feedback Loop between Mesenchymal-like Cancer Cells and  
418 Macrophages Is Essential to Breast Cancer Metastasis. *Cancer Cell* **25**, 605–620 (2014).
- 419 48. Jia, Z., Zhang, Y., Xu, Q., Guo, W. & Guo, A. miR-126 suppresses epithelial-to-mesenchymal  
420 transition and metastasis by targeting PI3K/AKT/Snail signaling of lung cancer cells. *Oncol.*  
421 *Lett.* **15**, 7369–7375 (2018).
- 422 49. Fu, X.-T. *et al.* Macrophage-secreted IL-8 induces epithelial-mesenchymal transition in  
423 hepatocellular carcinoma cells by activating the JAK2/STAT3/Snail pathway. *Int J Oncol* **46**,  
424 587–596 (2015).
- 425 50. Rothenberger, N. J., Somasundaram, A. & Stabile, L. P. The Role of the Estrogen Pathway in  
426 the Tumor Microenvironment. *Int. J. Mol. Sci.* **19**, 611 (2018).
- 427 51. Subramaniam, K. S. *et al.* Cancer-associated fibroblasts promote proliferation of endometrial  
428 cancer cells. *PLoS One* **8**, e68923–e68923 (2013).
- 429 52. Galmiche, A., Rak, J., Roumenina, L. T. & Saidak, Z. Coagulome and the tumor  
430 microenvironment: an actionable interplay. *Trends in Cancer* **8**, 369–383 (2022).
- 431 53. Mitrugno, A., Tormoen, G. W., Kuhn, P. & McCarty, O. J. T. The prothrombotic activity of  
432 cancer cells in the circulation. *Blood Rev.* **30**, 11–19 (2016).
- 433 54. Martignetti, L., Calzone, L., Bonnet, E., Barillot, E. & Zinovyev, A. ROMA: Representation and  
434 Quantification of Module Activity from Target Expression Data. *Front. Genet. |*  
435 *www.frontiersin.org* **7**, 18 (2016).
- 436 55. Andreatta, M. & Carmona, S. J. UCell: Robust and scalable single-cell gene signature scoring.  
437 *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



## 446 **Methods**

447 **Code availability:** MAYA is available as an R package on GitHub at <https://github.com/One->  
448 [Biosciences/MAYA/](https://github.com/One-Biosciences/MAYA/). Requires R >= 4.0.5.

### 449 **Data availability:**

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

459 Colon dataset: Raw count matrix and cell annotations were downloaded from the NCBI Gene  
460 Expression Omnibus (GEO) database under the accession code GSE144735 for the KUL3 cohort. Only  
461 protein-coding genes were kept for downstream analysis. MAYA automatic annotation function was  
462 run on this full dataset - including normal, tumor and border cells - to evaluate its scalability to large  
463 datasets. For our detailed pathway analysis, cells from Class "Normal" and from 10 out of the 35 cell  
464 types identifies by the authors were selected, representing a total of 1,415 cells.

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

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

### 470 Reference databases:

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

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

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

482

483 **Matrix preprocessing:** All count matrices were processed with Seurat v3 to get the gene-based cell  
484 embeddings and check the consistency of author's annotations. Matrices were log-normalized using  
485 scale factor 10,000. Top 2,000 variable features were found using "vst" method. PCA and UMAP  
486 computed with default settings, using first 10 PCs for UMAP, which constitutes the "gene-based  
487 UMAP". For the larynx dataset, the two datasets were read separately and merged in a unique Seurat  
488 object of 5,179 cells. The authors did not provide their annotation, so we followed the default Seurat  
489 pipeline on each individual count matrix, performed PCA and default clustering. We then annotated  
490 clusters based on expression of cell type markers described in the publication.

491

492 **Detailed description of MAYA algorithm:**

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

503 **Informativity:** For each successive mode, a density curve is drawn from the distribution to get local  
504 maxima and minima. A bimodal curve is expected to have at least one minimum that will be low  
505 enough relative to its surrounding maxima on the y-axis to mark a clear distinction between 2 groups  
506 of cells (difference of at least 10% of global maximum density). Only local minima with abscissa  
507 superior to the one of the global maximum are considered and iteratively evaluated in decreasing

508 order as the point is to detect extreme behaviors and activation patterns that potentially occur in rare  
509 populations. The iteration stops when a potential minimum meets the criteria, or none was found. As  
510 this process relies on the detection of inflection points that depends itself on the adjustment of the  
511 density curve to the distribution, we start with an adjustment meant to detect global variations of  
512 distributions and if none are detected we test a more fitted adjustment to ensure no significant local  
513 variation was missed. Then follow two additional checks to ensure the biological relevance of the  
514 detected mode. First, we filter out modes that are activated in very few cells as they could be outliers.  
515 The user can adjust this parameter based on what he expects to observe in the dataset or the number  
516 of cells from rarer cell type or set it to default 5%. The second biological check is based on the number  
517 of genes potentially contributing to the mode. However, it is hard to set a definition of what is a  
518 contributing gene to PCA; here we consider that contributing genes contribute more than they would  
519 be expected i.e. if all genes from the pathway contributed the same ( $1/\text{number of genes in the}$   
520  $\text{pathway}$ ). Given that pathways have various sizes, it is difficult to set a hard cutoff on this number of  
521 genes contributing to the mode. Instead, we chose to set a cut-off on the maximum contribution of a  
522 gene to a mode. As the sum of squared gene contributions is equal to 1, if a gene contributes to up to  
523 0.8, there is not much contribution left for other genes to share and this mode is probably driven by  
524 this unique gene. As a mode should represent joint expression of groups of genes, we do not consider  
525 these monogenic modes biologically significant. Setting a threshold of 0.4 allows to remove monogenic  
526 modes while keeping a relatively large number of modes with higher cell type specificity. This  
527 parameter can also be changed by the user depending on the tolerance to probable monogenic  
528 pathways. Finally, we chose to test the informativity of each pathway mode in decreasing order of  
529 variance explained in the dataset and to stop when a mode is found uninformative after mode 2 as we  
530 know the following will explain even less variance and is more likely to be noise.

531 **Predict cell type:** Once the activity matrix generated, a k-Nearest Neighbors matrix with  $k=20$  is  
532 computed, then an adjacency matrix using Jaccard distance and finally transformed as a weighted  
533 graph using igraph function `graph.adjacency`. Clustering is then performed using `leiden_find_partition`  
534 from `leidenbase` package with `ModularityVertexPartition` as partition type and a maximum number of  
535 iterations of 2. The average activity score is computed by cell type and by cluster. Each cluster is  
536 attributed the cell type for which the activity score is the highest, if it passes a threshold of default  
537 value 0, otherwise it is labeled as unassigned. This value can be modified by the user, depending on  
538 the level of confidence needed for annotation.

539

540 **Comparison with other tools:**

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

544 **Pagoda2, AUCell and Cell-ID to compare cell type prediction with MAYA:** The three tools were  
545 provided the same PanglaoDB cell type marker lists as MAYA. Pagoda2 was run with default settings,  
546 following the vignette. AUCell was run using default settings, with log-normalized counts as input. We  
547 used AUCell\_exploreThresholds function to select the cell type lists that were activated in at least one  
548 cell. MAYA's procedure of clustering and cell type attribution was performed on AUCell and Pagoda2  
549 activity matrix as they do not have an integrated function for cell annotation. Cell-ID was applied on a  
550 Seurat object following standard procedure, computing MCA and then performing hypergeometric  
551 test with gene lists. Each cell was attributed the cell type for which  $-\log_{10}(p\text{-value})$  was the highest.  
552 When the value was inferior to 2, the cell was labeled unassigned.

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

554

555 **Metrics:**

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

559 
$$SDI_c = \frac{(-1) * \sum_{i=1}^N p_i * \log(p_i)}{\log(N)}$$

560 With  $c$  the cluster in which we compute the SDI,  $N$  the number of different possible identities (patients  
561 in our case) and  $p_i$  is the proportion of cells from the cluster corresponding to identity  $i$ . SDI of 1  
562 indicates that cells constituting the cluster come equally from all possible identities i.e. the cluster  
563 displays high identity diversity.

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

568 
$$S_{m,c} = \frac{a_{m,c}^2}{\sum_{p=1}^N a_{m,p}^2}$$

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

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

595 **Contributions**

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

598

599 **Corresponding author**

600 Correspondence to Céline Vallot – [celine.vallot@onebiosciences.fr](mailto:celine.vallot@onebiosciences.fr).

601

602 **Competing interests**

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

605

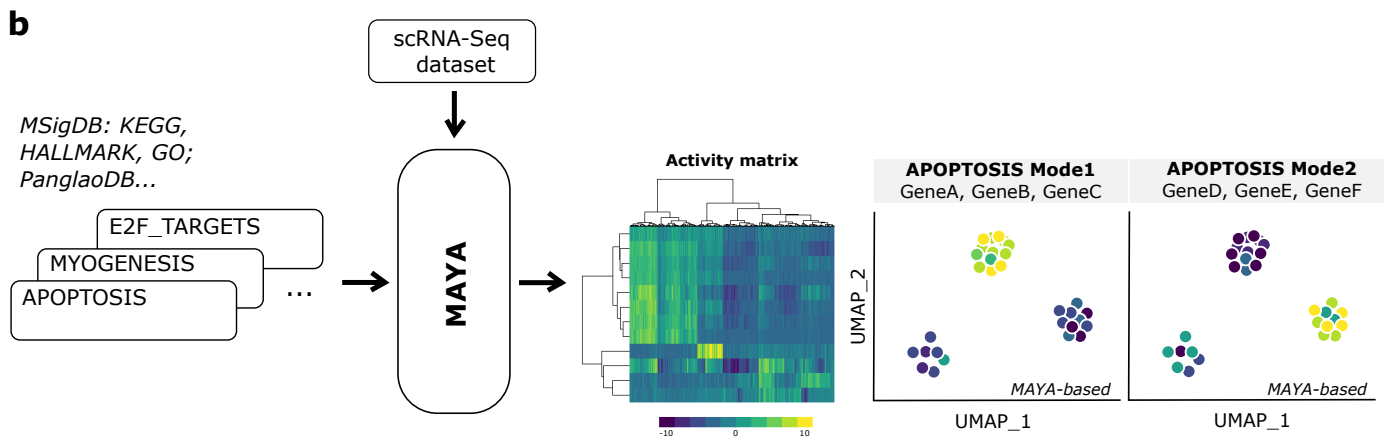
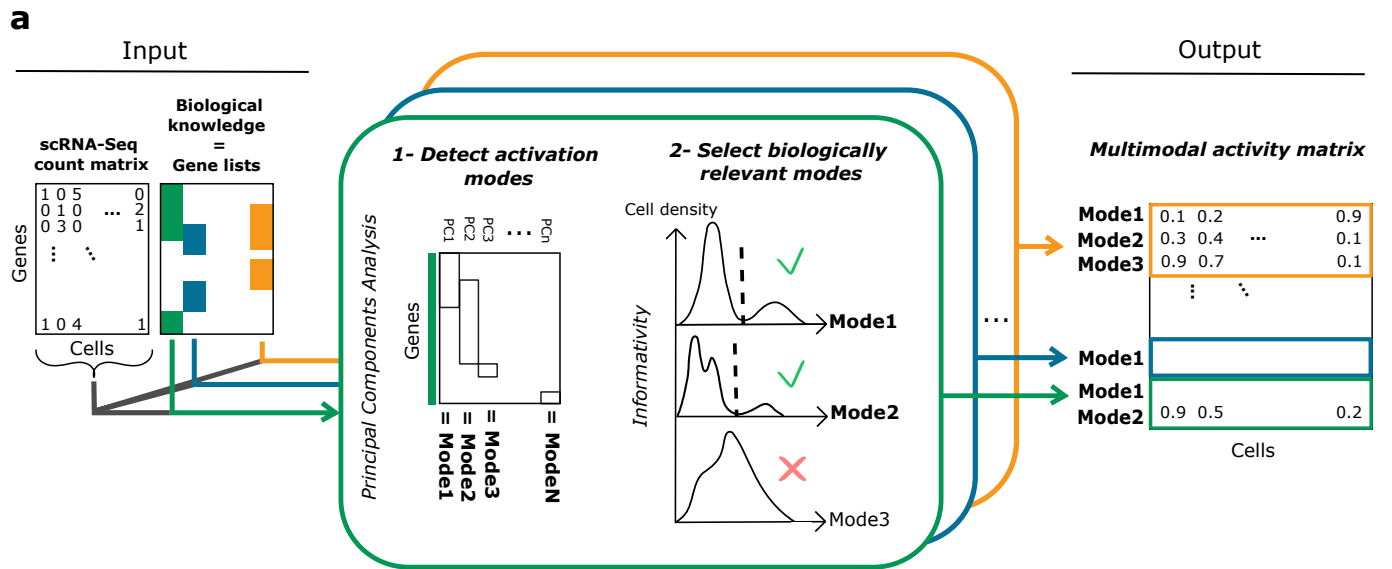
606 **Rights and permissions**

607 **Open Access.** MAYA is available on GitHub at <https://github.com/One-Biosciences/MAYA/> and  
608 licensed by One Biosciences under a GNU Affero General Public Licence v3.0. To view a copy of this  
609 license, visit <https://www.gnu.org/licenses/agpl-3.0-standalone.html> .

610

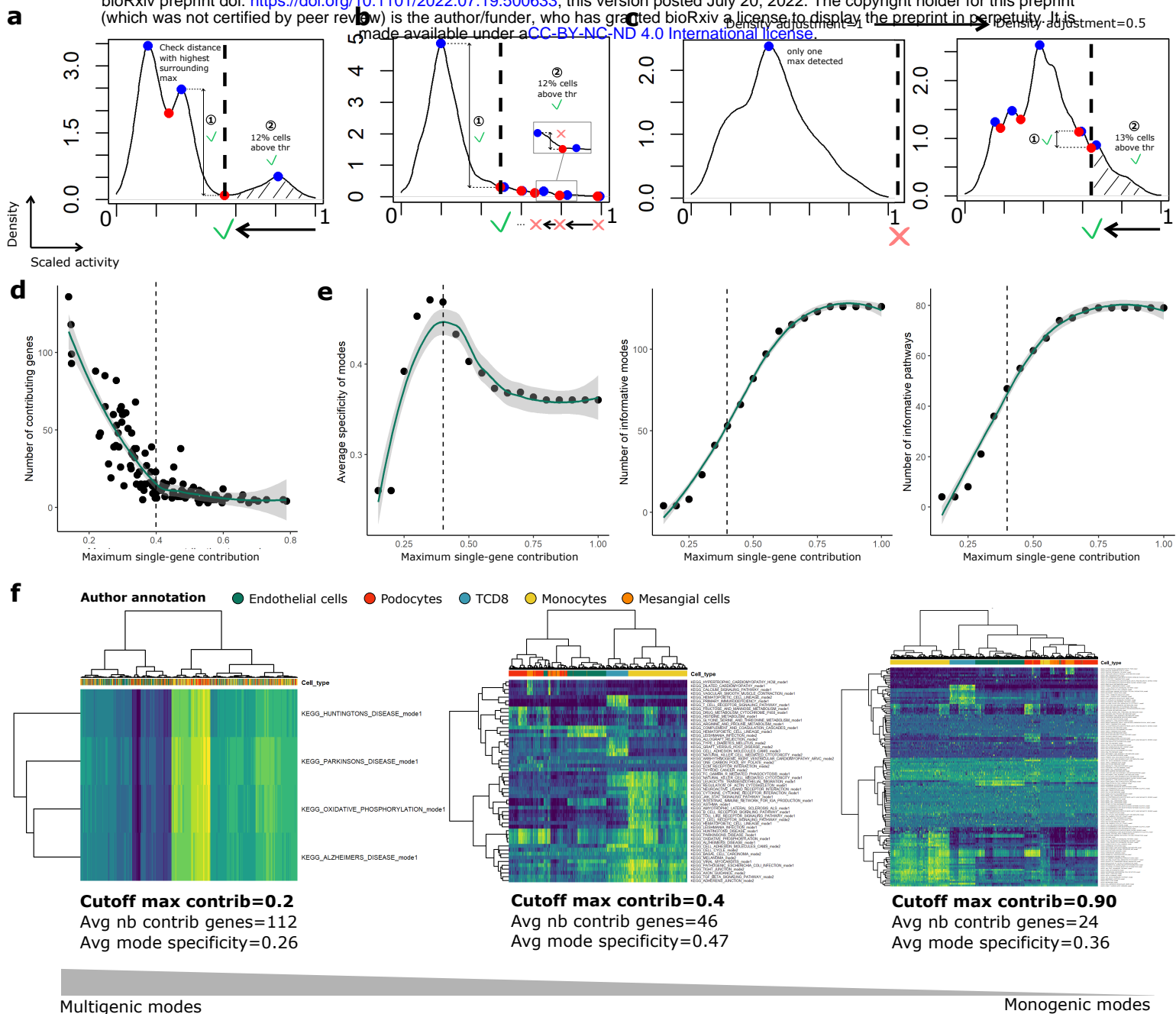
611

612



**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).



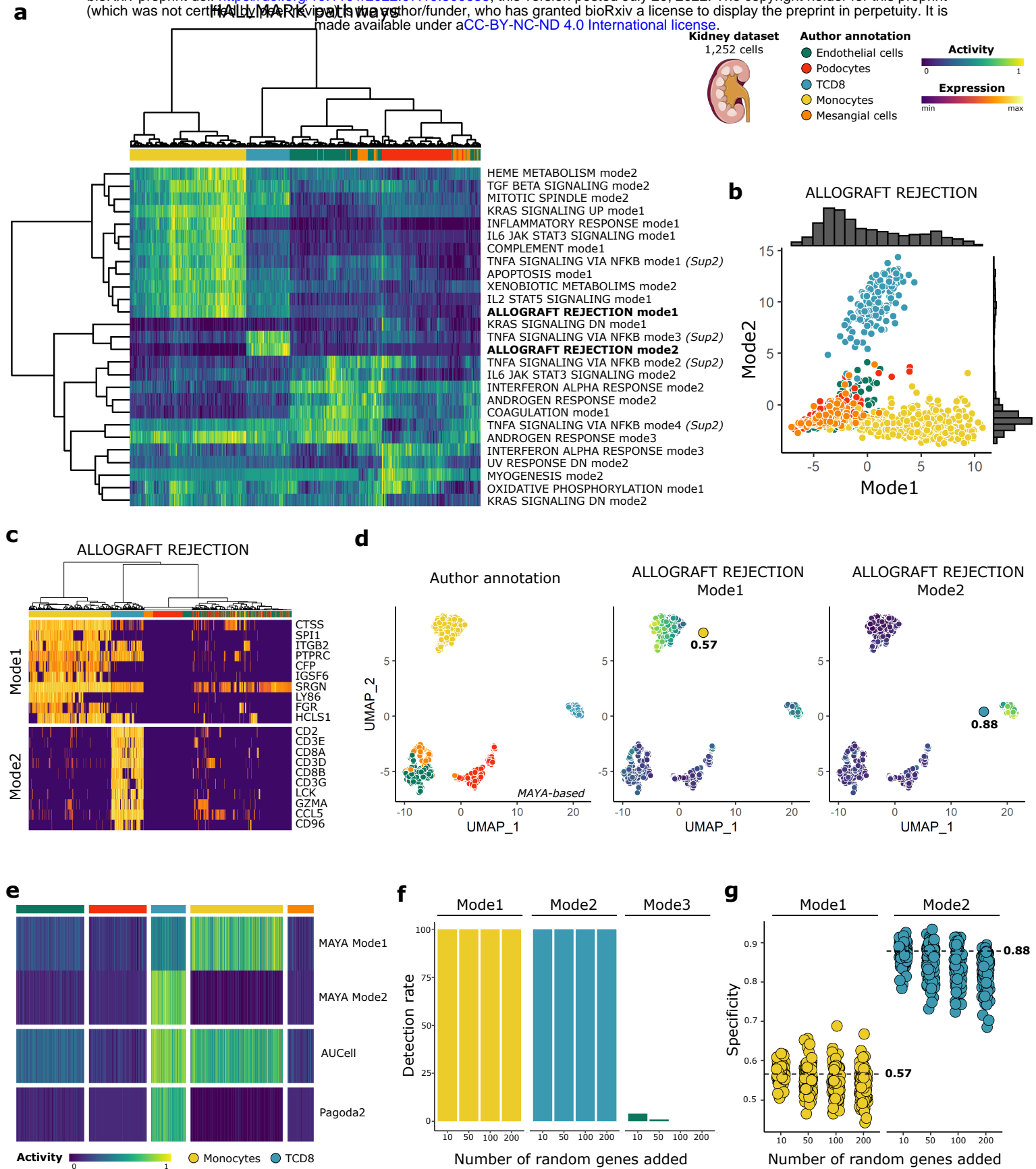
**g**

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:**

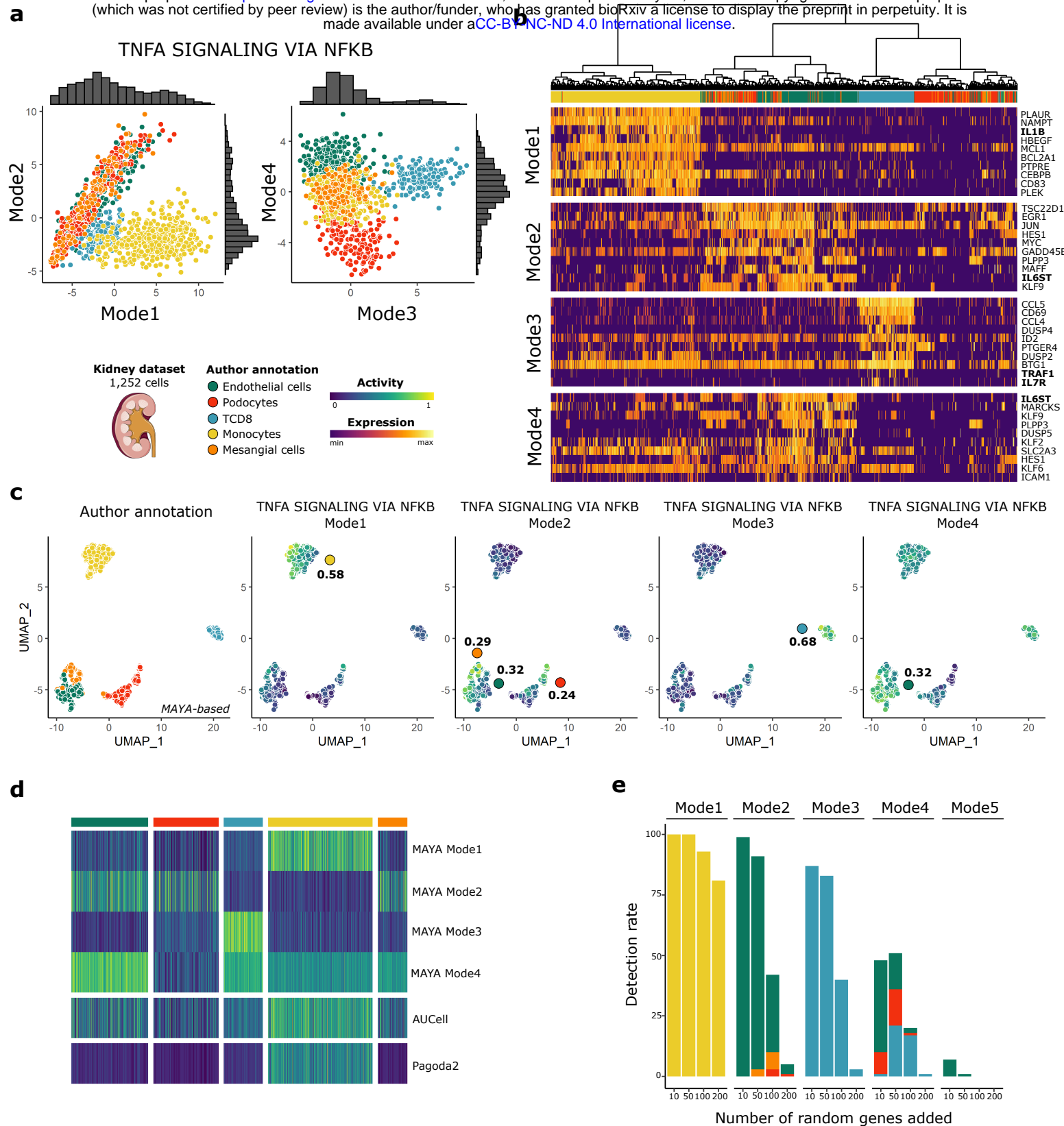
(a,b) Examples of density curve of activity scores for one mode of activation. Detected maxima in density are colored in blue and minima in red. MAYA selects a mode as relevant when it has a local density minimum that (i) is low enough compared with surrounding highest maximum and that (ii) splits the 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.





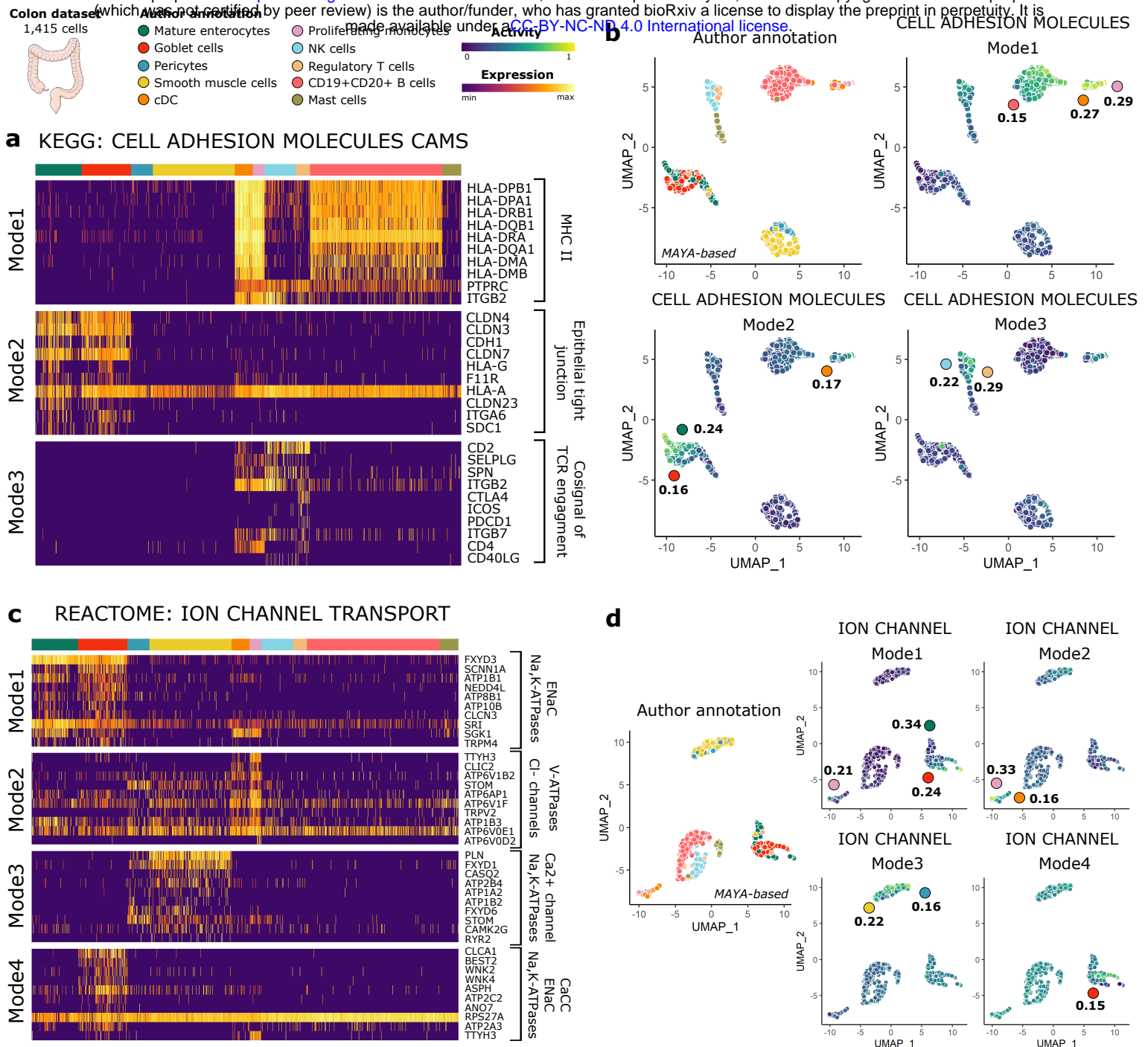
**Fig.2: Activation modes of Hallmark pathways in kidney with MAYA**

(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:**

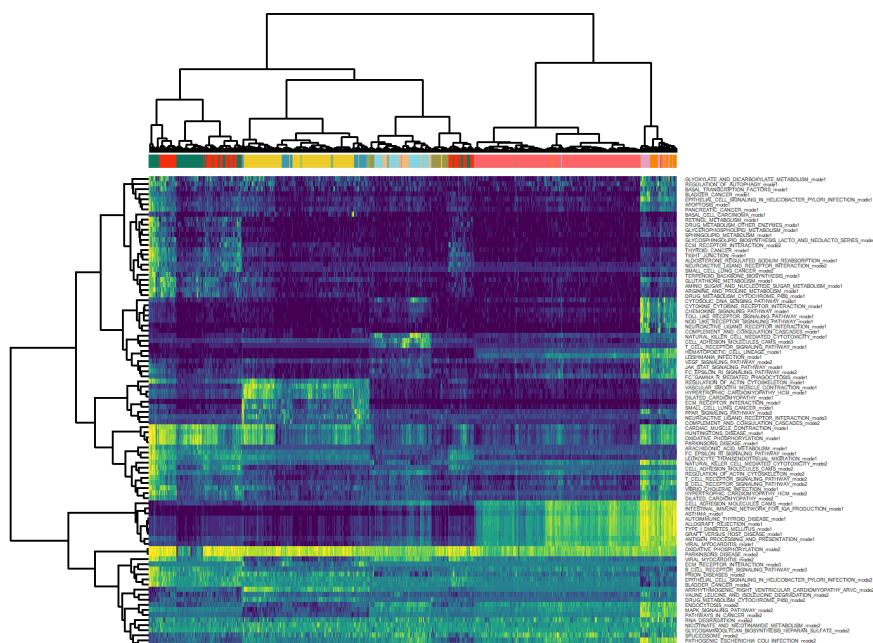
**(a)** Scatterplot of Mode 2 versus Mode 1 and Mode 4 versus Mode 3 cell activity scores, for the pathway *TNFA signaling via NFKB* on the kidney dataset. Associated density histograms are indicated on the sides of the graphs. **(b)** Heatmap of scaled gene expression for top10 contributing genes for 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.



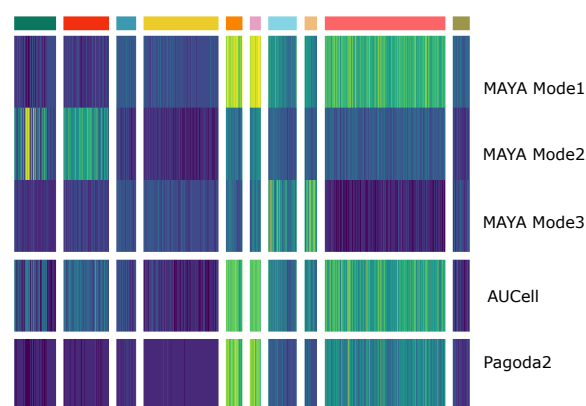
**Fig.3: KEGG and REACTOME activity in colon with MAYA**

(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.

**a**

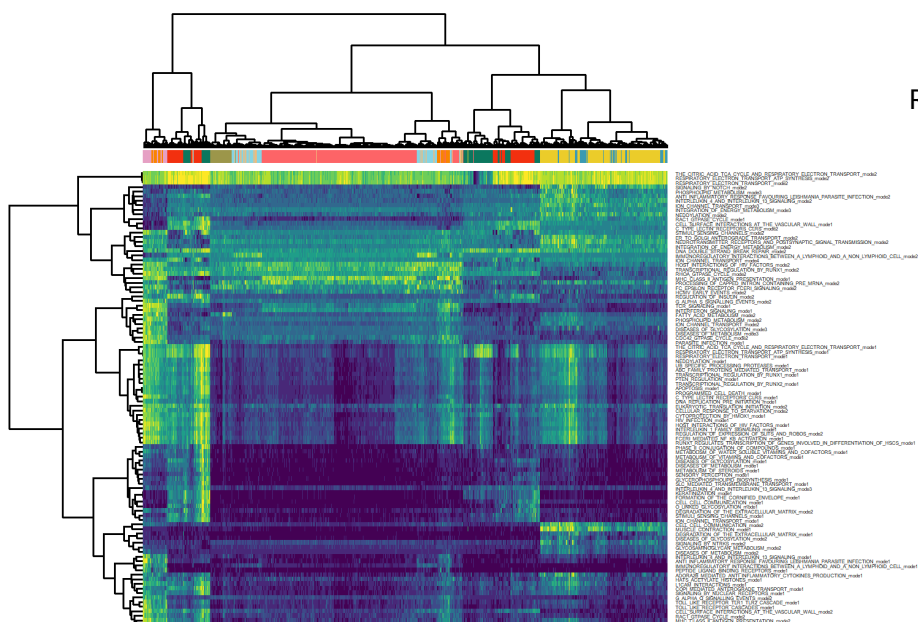


**KEGG: CELL ADHESION MOLECULES**



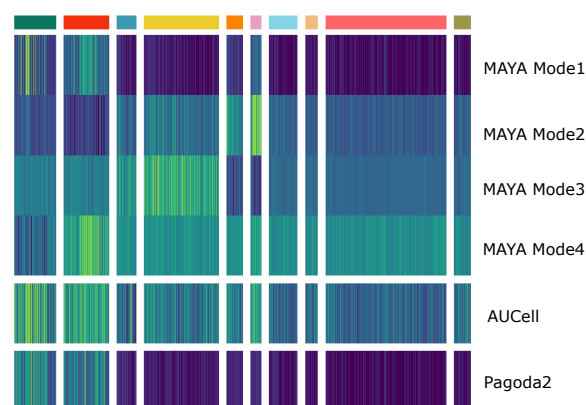
**c**

**REACTOME pathways**

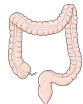


**d**

**REACTOME: ION CHANNEL TRANSPORT**



**Colon dataset**  
1,415 cells



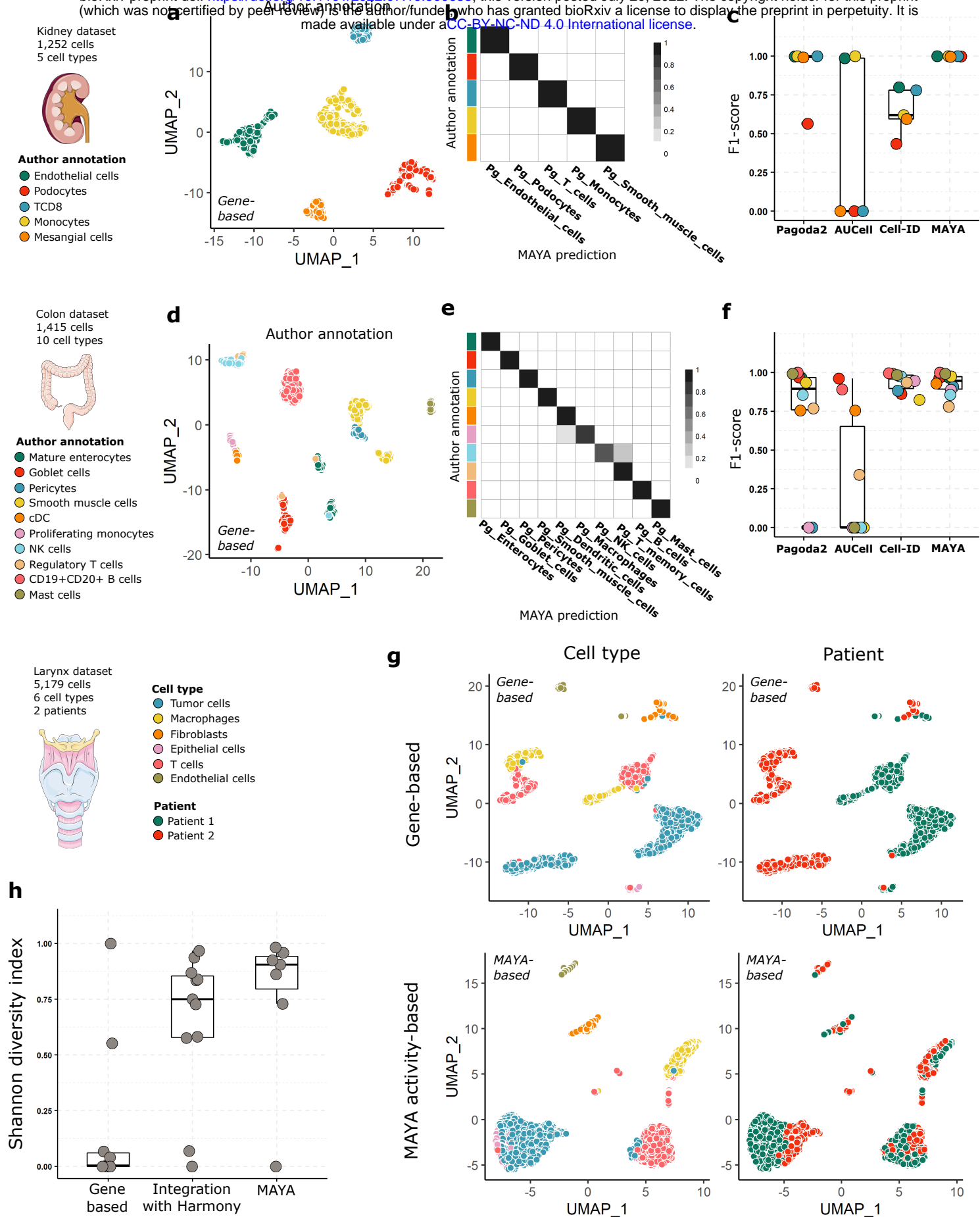
**Author annotation**

- Mature enterocytes
- Proliferating monocytes
- Goblet cells
- NK cells
- Pericytes
- Regulatory T cells
- Smooth muscle cells
- CD19+CD20+ B cells
- cDC
- Mast cells



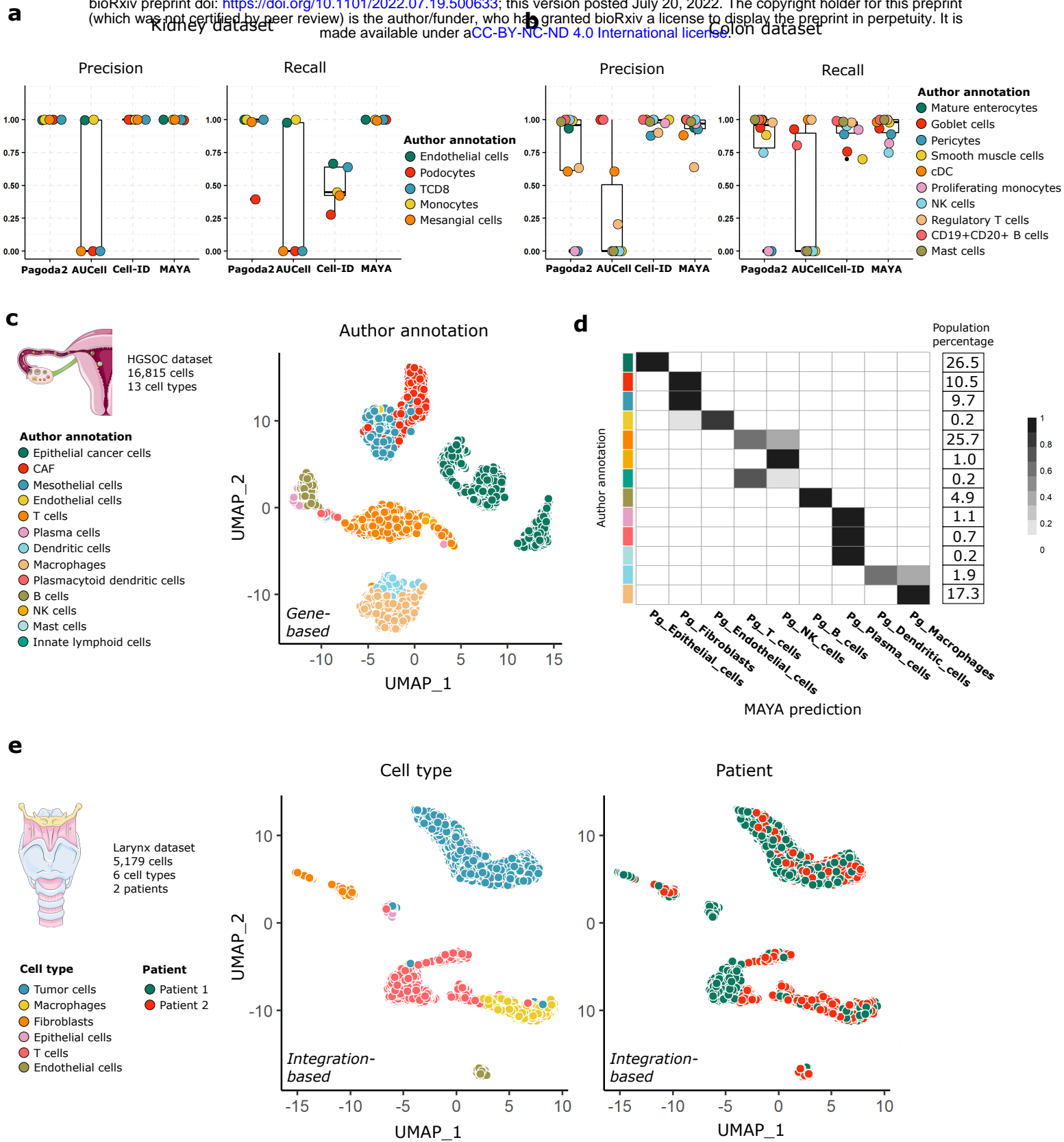
**Supplementary Fig.3:**

**(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.



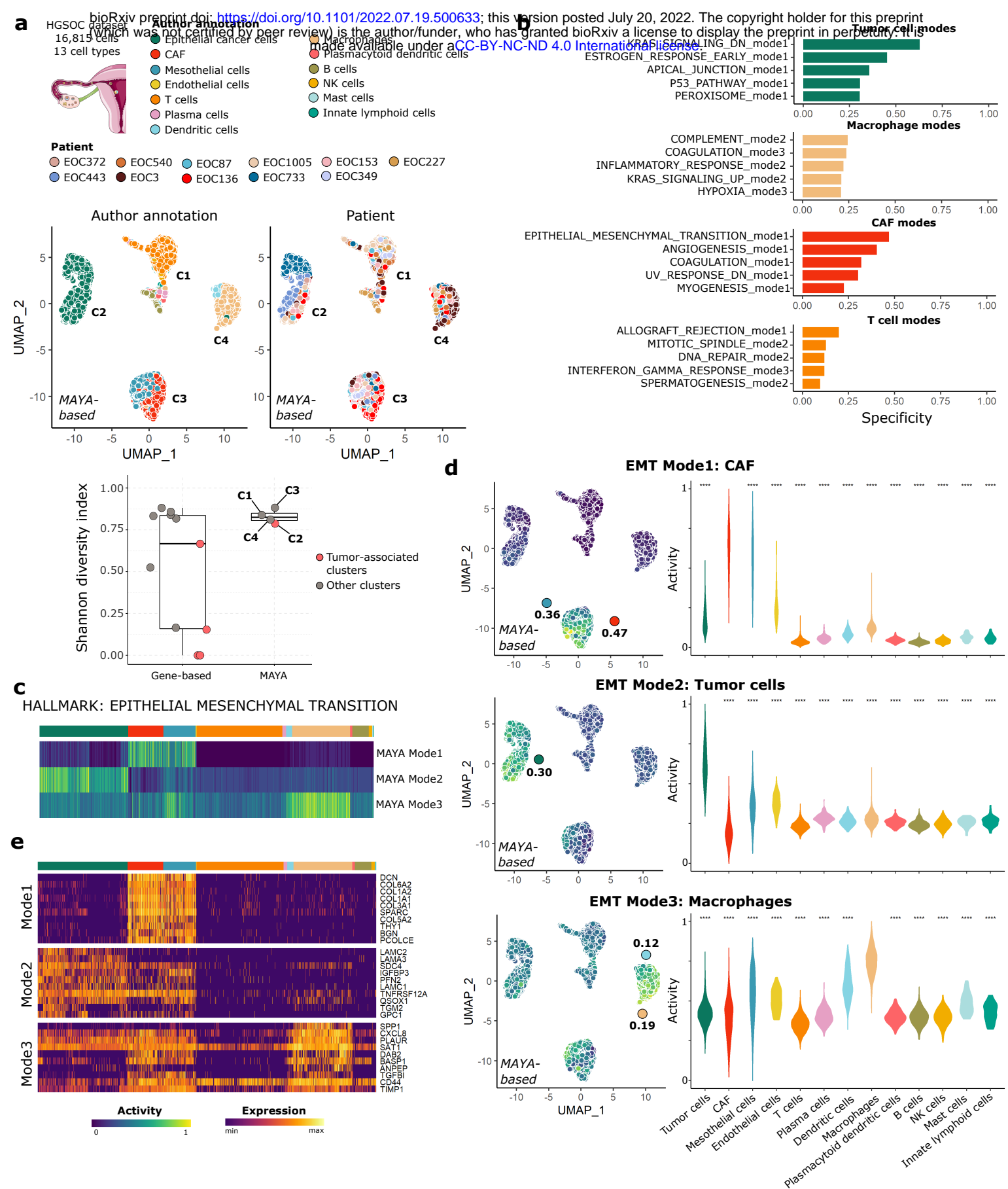
**Fig.4: MAYA automatically annotates cell type**

(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.



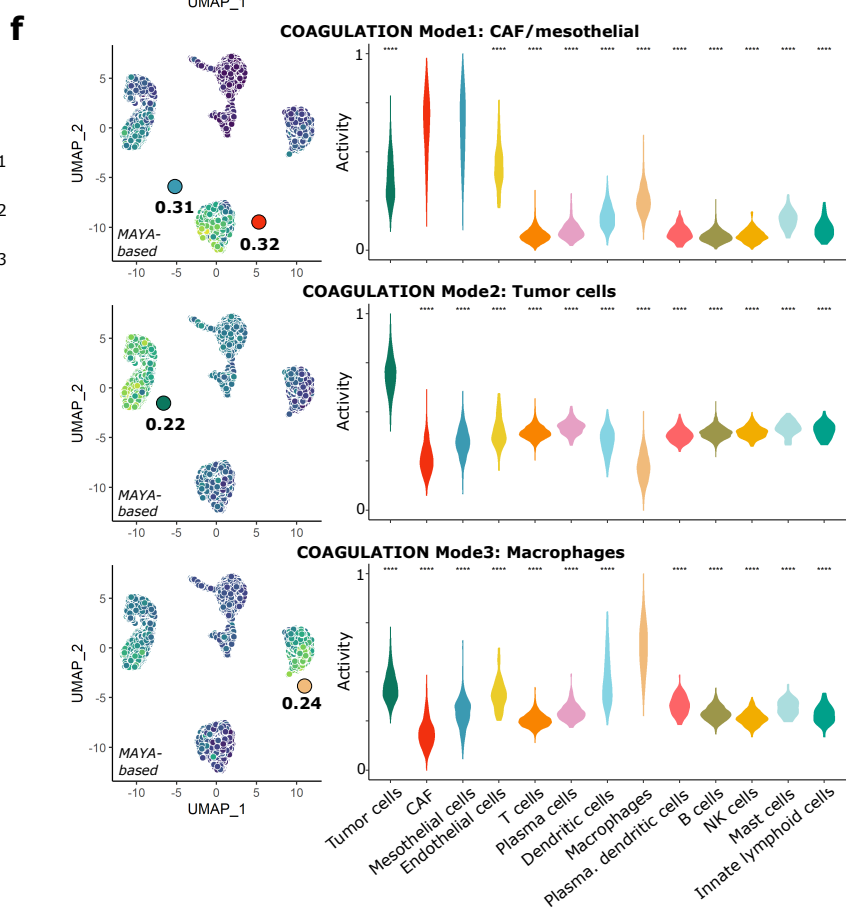
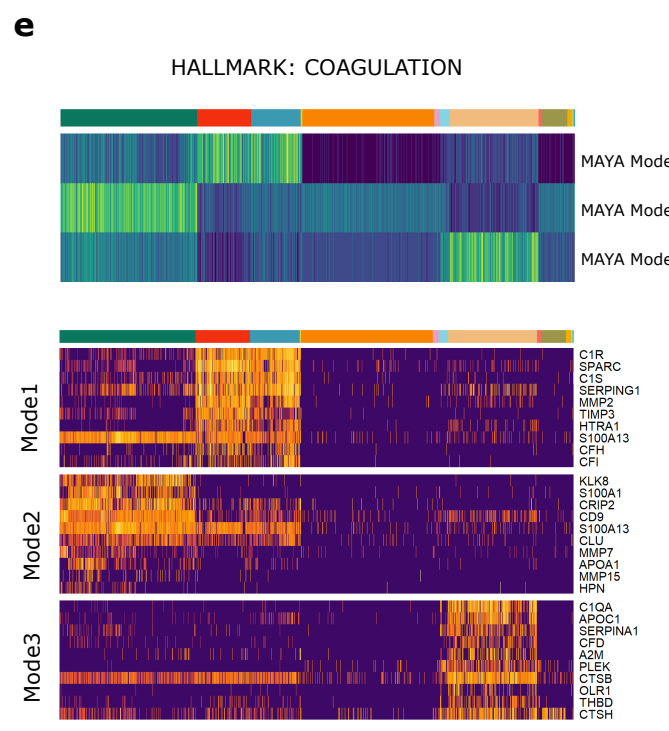
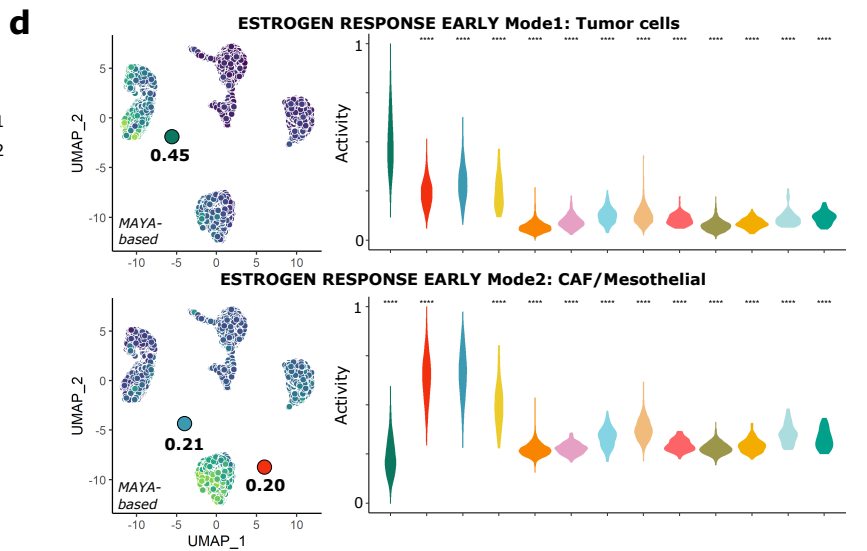
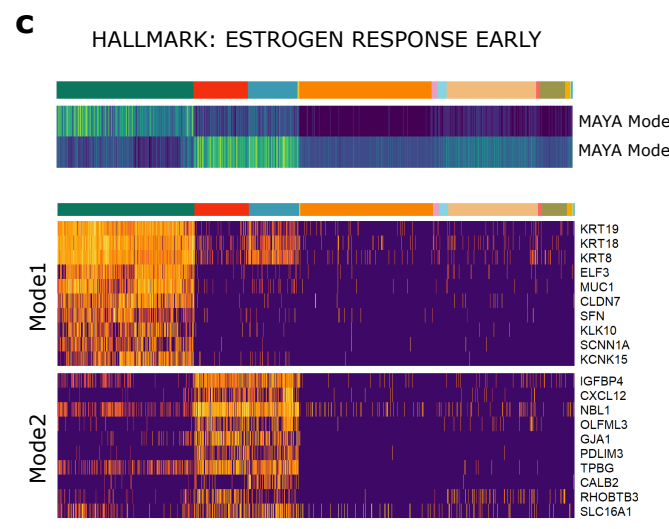
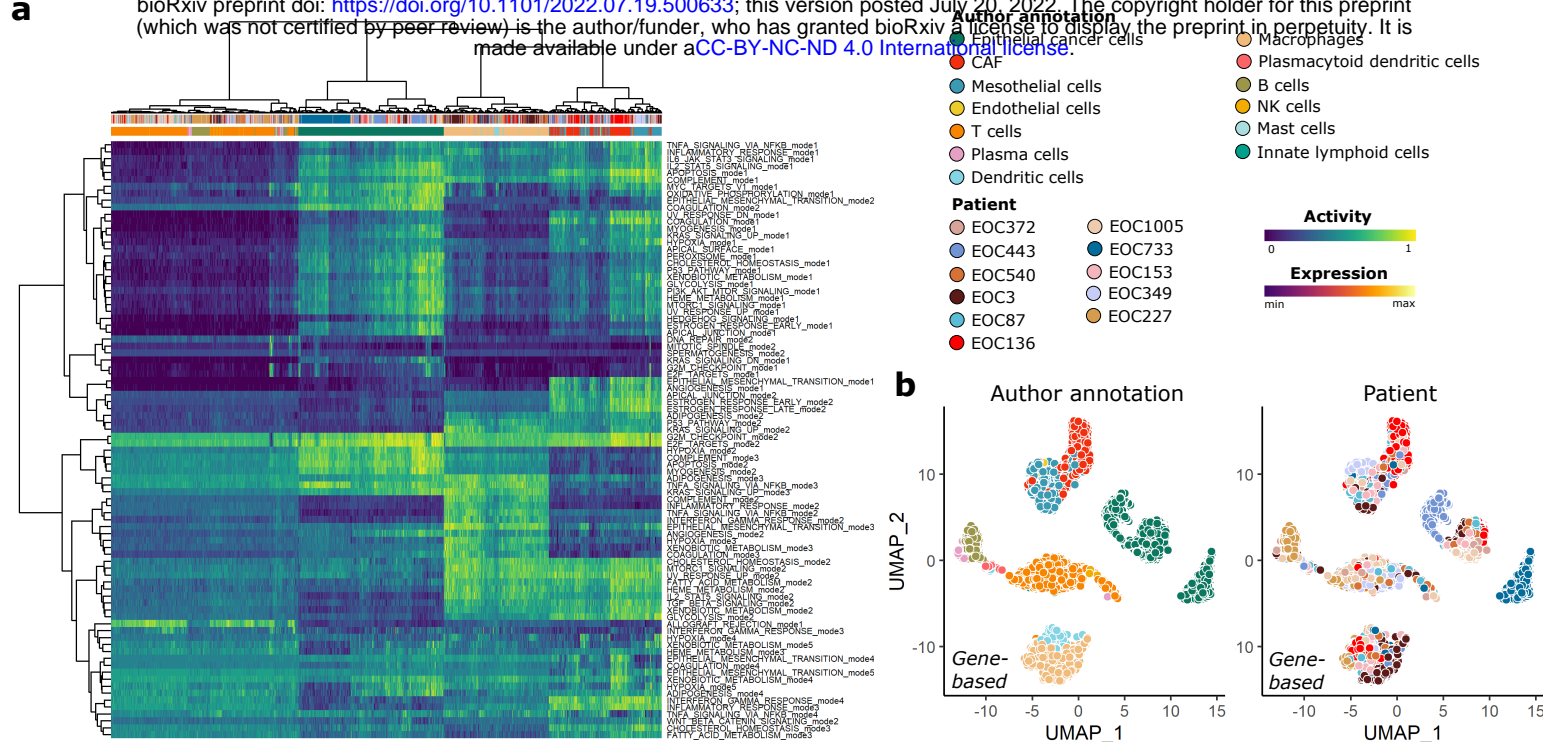
**Supplementary Fig.4:**

(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.





bioRxiv preprint doi: <https://doi.org/10.1101/2022.07.19.500633>; this version posted July 20, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**Supplementary Fig. 5**

**(a)** Heatmap of activity matrix computed on every dataset with 150 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).