An individualized causal framework for learning intercellular communication networks that define microenvironments of individual tumors

Xueer Chen[1,2*], Lujia Chen[1,2*], Cornelius H.L. Kürten[3,4,5*], Fattaneh Jabbari[1,2], Lazar Vujanovic[3,4], Ying Ding[6], Binfeng Lu[7], Kevin Lu[8], Aditi Kulkarni[3], Tracy Tabib[6], Robert Lafyatis[6], Gregory Cooper[1,2,4§], Robert Ferris[3,4§], and Xinghua Lu[1,2,4§]

[1]Department of Biomedical Informatics, University of Pittsburgh
[2] Center for Causal Discovery, University of Pittsburgh
[3]Department of Otolaryngology, University of Pittsburgh
[4] University of Pittsburgh Hillman Cancer Center, University of Pittsburgh
[5] Department of Otorhinolaryngology, Head and Neck Surgery, University Hospital Essen, University Duisburg-Essen,
[6] Department of Biostatistics, University of Pittsburgh
[7] Department of Immunology, University of Pittsburgh
[8] Williamsville North High School, Williamsville, NY 14221

* These authors contributed equally to the work.
§ Corresponding authors

Abstract

Cells within a tumor microenvironment (TME) dynamically communicate and influence each other's cellular states through an intercellular communication network (ICN). In cancers, intercellular communications underlie immune evasion mechanisms of individual tumors. We developed an individualized causal analysis framework for discovering tumor-specific ICNs. Using head and neck squamous cell carcinoma (HNSCC) tumors as a testbed, we first mined single-cell RNA-sequencing data to discover gene expression modules (GEMs) that reflect the states of transcriptomic processes within tumor and stromal single cells. By deconvoluting bulk transcriptomes of HNSCC tumors profiled by The Cancer Genome Atlas (TCGA), we estimated the activation states of these transcriptomic processes in individual tumors. Finally, we applied individualized causal network learning to discover an ICN within each tumor. Our results show that cellular states of cells in TMEs are coordinated through ICNs that enable multi-way communications among epithelial, fibroblast, endothelial, and immune cells. Further analyses of individual ICNs revealed structural patterns that were shared across subsets of tumors, leading to the discovery of 4 different subtypes of networks that underlie disparate TMEs of HNSCC. Patients with distinct TMEs exhibited significantly different clinical outcomes. Our results show that the capability of estimating individual ICNs reveals heterogeneity of ICNs and sheds light on the importance of intercellular communication in impacting disease development and progression.

Communications among cells in a tissue influence their cellular states, leading to a homeostatic state at a given time that impacts the physiological and pathological processes of cells. As an example, cancer cells expressing neoantigens actively modulate the immune environment through intercellular communications[1-6] to evade immune surveillance. The successes of contemporary immune checkpoint inhibitors [7-9] demonstrate the potential for identifying and targeting intercellular communications in treating diseases. Importantly, heterogeneity of genomic alterations among tumors also leads to diverse TMEs [10, 11] and thereby to distinct immune evasion mechanisms and heterogeneous responses to immune therapies. Thus, it is important to understand how cells communicate with each other within each tumor and gain insights into tumor-specific immune evasion mechanisms to help guide personalized immune therapies.

Studying cell-cell communication is a long-standing problem [12-14], and the advent of technologies for profiling single cells [15-20] provides unprecedented opportunities to advance the field. An increasing number of studies concentrate on modeling cell-cell communication through mining single cell transcriptome data [20-26]. However, the majority of current efforts (see recent reviews by Almet et al. [25] and Armingol et al [26]) concentrate on modeling potential interactions of known ligand-receptors, through studying the correlations between the expression values of ligand-receptor pairs. While informative, a knowledge-based approach cannot discover previously unknown interactions; co-expression of a pair of ligand-receptor mRNA does not entail physical interaction between their protein products; and, more importantly, correlation analysis does not provide information regarding the biological consequence of ligand-receptor interactions. Finally, contemporary modeling of cell-cell communication is mostly performed at a population level, where a single model is developed to

explain the data of all cases. Such analysis is not suitable to investigate ICNs underlying the well-appreciated heterogeneity of TMEs observed in tumors [10, 11]. Thus, there is an unmet need for computational methodologies that enable studying ICNs within individual tumors to better understand how cells communicate within a tumor's TME.

An ICN is essentially a causal network, in which a change in the cellular state of a cell (or a subpopulation of cells) can causally influence the cellular states of other cells via communication channels involving physical or paracrine ligand-receptor interactions. If we can accurately detect and represent the state of cellular processes in cells, we can apply principled causal discovery methods to search for causal relationships among the cellular processes across cells, i.e., an ICN. As a discipline of machine learning and artificial intelligence, the field of computational causal discovery has witnessed significant progress in recent decades [27-29]. Particularly, a new family of individualized (or instance-specific) causal discovery methods [30-32] has been developed, which can learn a causal network based on the data of an individual case, be it a single cell or a single tumor.

In this study, we developed and applied an individualized Bayesian causal discovery framework to learn tumor-specific ICNs, and we used HNSCC cancer as a testbed for our study. The framework consists of multiple steps as follows. First step, we applied a statistical model to identify GEMs from single cell data, such that expression status of a GEM would reflect the state of a transcriptomic process regulating its expression in a cell or a tumor. Identification of GEMs lays a foundation for modeling how the change of a cellular process in one type of cell (e.g., cancerous epithelial cells) causally influences the state of another process in another type of cells (e.g., CD8+ lymphocytes). Second, we inferred the activation states of the discovered transcriptomic processes in a large number of tumors, so that we could reliably learn causal

relationships of cellular processes across different types of cells. Third, we employed causal Bayesian network (CBN) learning algorithms to search for ICNs among diverse types of cells at the cohort and the individual tumor levels. Our analyses uncovered ICNs that model well the coordinated expression of GEMs across diverse types of cells. By mining the patterns of ICNs of individual tumors, we discovered distinct patterns of ICN structure that underlie disparate TMEs of HNSCCs that are associated with significantly different patient outcomes. Our methods serve as a general tool for studying intercellular communications that influence disease development and progression.

RESULTS

Identification of context-specific GEMs reflective of the states of transcriptomic processes

We collected single-cell RNA sequencing (scRNAseq) data of 109,879 cells from 18 HNSCC tumors and matched peripheral blood samples. After quality control, we retained data on 78,887 $CD45^+$ cells from tumor and peripheral blood samples of the 18 patients, and we collected 19,107 $CD45^-$ cells from 12 of the tumors (**Figure 1a, Figure S1**). Since different cells (e.g., epithelial vs lymphocytes) have significantly different transcriptomes and regulatory processes, we separated cells into 4 major categories and modeled their transcriptomic regulatory process respectively: immune-related (PTPRC/$CD45^+$; 78,887 cells), epithelial (9,852 cells), fibroblast (3,476 cells), and endothelial (5,779 cells), based on the expression of respective marker genes.

Based on the assumption that the transcriptome of each single cell is a mixture of GEMs controlled by different transcriptomic processes, we applied the nested hierarchical Dirichlet processes (NHDP)[33] model to the scRNAseq data to identify GEMs. NHDP is a non-parametric statistical model that can deconvolve a mixture of discrete variables (e.g., words in a

text documents) by searching for modules (topics) that capture co-occurrence of words in text documents. When applied to scRNAseq data from a population of cells, NHDP can discover sets of co-expressed genes (i.e., GEMs), with each GEM consisting of genes that are repeatedly co-expressed in many cells. As such, the expression status of a GEM in a cell reflects the state of a transcriptomic process that regulates the GEM. NHDP is a hierarchical model that represents the joint distribution of GEMs (i.e., co-occurrence patterns of GEMs in cells) using a tree-structured hierarchy. It can represent the differentiation (evolution) processes of cells in a cell population, such that a GEM close to the root of a tree tends to be expressed broadly among the majority of cells, representing a general transcriptomic process shared by cells of a common lineage. In contrast, a GEM close to the leaves of the tree tends to be expressed in a subpopulation of cells, reflecting the state of transcriptomic processes that lead to differentiation of a subtype of cells.

Applying the NHDP model to pooled peripheral and tumor-infiltrating lymphocyte (TIL) $CD45^+$ cells, we identified 39 GEMs (referred to as immune GEMs) and their signature genes, and we examined the expression patterns of each immune GEM among the $CD45^+$ cells (**Figure 1, Table S1,** and **Figure S2**). As an example, immune GEM1 is broadly expressed in T cells, B cells, and natural killer (NK) cells, and thus, it represents the transcriptomic process that is shared by cells from the lymphoid lineage; immune GEM2 is mainly expressed in cells of myeloid (monocyte/macrophage) lineage; immune GEM3 is concentrated in TILs, potentially reflecting the impact of tissue environment on T cells [34]; GEM14 is mainly expressed in $CD8^+$ T cells, which contain genes commonly expressed in activated cytotoxic T cells (such as *CD8A, GZMB, CXCL13, IFNG, GZMA, GNLY, NKG7,* and *PRF1*), suggesting that this GEM is likely regulated by a T-cell activation processes. Interestingly, certain GEMs can be expressed in cells across conventionally defined immune cell lineages. For example, GEM32 is shared among T

cells and NK cells from peripheral blood (**Figure 1f, Figure S2**), indicating that our approach can discover shared transcriptomic processes among cells that otherwise have distinct characteristics.

In addition to PTPRC/CD45$^+$ cells, we identified GEMs among epithelial, fibroblast, and endothelial cells, respectively, as well. Similarly, some cell-type-specific GEMs represent lineage-defining transcriptomic processes, and some are unique to a small subset of cells reflecting different activation states (**Figure S2**). In summary, the NHDP model can identify GEMs that reflect the state of transcriptomic processes in individual cells, enabling us to represent each cell in the space of transcriptomic processes. This representation not only provides a more concise and robust representation of cellular state than is possible with individual genes, but it also serves as an instrument for modeling the causal relationships of transcriptomic processes across cell boundaries.

Different combinations of active transcriptomic programs define distinct cell subtypes

We set out to examine whether representing cells based on their active transcriptomic processes (expression status of GEMs) can reveal cell subtypes based their cellular states in contrast to representing cells in a gene space. We performed clustering analyses to identify subtypes within each of the following major immune-cell categories: CD8$^+$, CD4$^+$, and T regulatory cells (Treg), B cells, NK cells, monocytes and macrophages (**Figure S3**). As an example, eight subtypes (clusters) of CD8$^+$ T cells were identified (**Figure 2a**), with each exhibiting a distinct composition of GEMs. Pseudo-time analysis [35] of CD8$^+$ T cells represented in GEM space revealed a development path of these CD8$^+$ clusters, starting from "naïve" T cells from peripheral blood, to "primed/activated" tumor infiltrating CD8$^+$ T cells, and further onto "exhausted" CD8$^+$ T cells (**Figure 2b**). Interestingly, the composition of GEMs expressed in

these subtypes can be represented as a tree (**Figure 2c**) that mimics the development path of $CD8^+$ T cells: All subtypes expressed GEM1, which represents a transcriptomic process shared by cells of the lymphoid lineage; expression of GEM3 differentiates tumor infiltrating and peripheral $CD8^+$ T cells; expression of other GEMs further leads to differentiation of cells into subtypes. Unlike conventional clustering analysis based on overall transcriptomics, our approach clearly delineated the transcriptomic processes that contributed to such development. For example, cluster CL6 constitutes a subtype of tumor infiltrating $CD8^+$ T cells that followed a unique trajectory (**Figure 2b**), and the expression of GEM7 appears to underlie the development of this subtype (**Figure 2c**). Thus, inspecting the GEM composition of each cell cluster not only reveals distinct patterns of activation states of transcriptomic processes in cells, but also illustrates how a cell subtype develops. Using the same approach, we have identified cell subtypes (clusters) among each major PTPRC/$CD45^+$ cell category, as well as among epithelial, fibroblast, and endothelial cells (**Figure S3**).

Distinct patterns of active transcriptomic processes and cell composition revealed heterogeneous immune environments of HNSCCs

Based on the assumption that the expression status of a GEM in a tumor would reflect the presence and activation status of transcriptomic process regulating the GEM, we set out to detect the activity of transcriptomic processes by deconvoluting bulk RNA sequencing data of HNSCC tumors profiled by TCGA [36]. Using the top 50 genes of a GEM (ranked according to the probabilities of the genes being assigned to the GEM) as a signature profile of the GEM, we employed gene set variation analysis (GSVA) [37] to determine the expression status of the GEM in each tumor. The GSVA method calculates a sample-wise enrichment score of genes

within a GEM and further estimates the variance of GEM enrichment scores across a cohort of samples. Thus, it infers the relative expression status of each GEM, i.e., the relative activation status of the corresponding transcriptomic process in each tumor (**Table S2a**). We performed GSVA analysis of immune GEMs using the bulk RNA data of 522 HNSCC tumors profiled by TCGA. Similarly, we used the gene expression signature of each subtype of immune cell to infer their relative enrichment scores in the tumors (**Table S2b**).

We then performed consensus clustering analyses on the tumors using as features the GSVA scores of immune GEMs, and alternatively, the cell subtypes. Both analyses revealed 4 distinct immune environments, with each cluster consisting of distinct compositions of GEMs (**Figure 3a**) or cell subtypes (**Figure 3b**), wherein the members of the clusters defined by GEMs and cell subtypes, respectively significantly overlapped. Survival analyses showed significant differences in patient outcomes between the clusters, supporting the impact of the modeled immune environment on disease progression.

To identify the transcriptomic processes impacting survival, we performed Cox proportional hazard analyses, using the GSVA scores of the GEMs (or the cell subtypes) as independent variables. When only considering the impact of an individual GEM (**Table S3**) and immune cell subtype (**Table S4**) on survival, 20 GEMs and 17 cell subtypes were deemed to be significantly associated with overall survival of patients ($p < 0.05$). We further performed a forward-and-backward search to investigate the impact of combinations of these GEMs (or cell types) on survival, which could detect and eliminate redundant GEMs or cell subtypes; 7 GEMs and 6 immune-cell subtypes (**Figure 3c and 3d)** were statistically significantly associated with patient survival in the multi-variate Cox proportional hazard analyses. The results of GEM-based and cell-subtype-based analyses are consistent. For example, the expression of immune

GEM7 is concentrated in cluster CL6 of $CD8^+$ T cells (**Figure 3e** and **Figure 2a**), and the enrichment of the GEM7 or cells from cluster CL6 were deemed as significantly associated with increased hazard ratio; similarly, expression of immune GEM18 is concentrated in cluster CL6 of monocytes (**Figure 3f** and **Figure S3**), and their enrichment in tumors were both significantly associated with an increased hazard ratio. The results support that HNSCC tumors exhibit distinct patterns of immune environments, which are associated with patient outcomes.

Transcriptomic programs of cells are coordinated in TMEs

The capability of detecting the states of transcriptomic processes in cells and individual tumors enabled us to further investigate whether the transcriptomic programs in different cell types are coordinated. We estimated GSVA scores of GEMs of the four major cell types: epithelial, immune, endothelial, and fibroblast cells (**Table S5)** and performed pair-wise correlation analyses (**Figure 4a** and **Figure S4)**. The results showed that expression of GEMs from different types of cells are significantly correlated, supporting the notion that intercellular communication may underlie coordinated activation/inactivation of transcriptomic processes across major cell types. To investigate the impact of cancer and stromal cells on the states of immune cells, we further trained regression models using GSVA scores of the non-immune cell GEMs as independent variables to predict expression of each of the immune GEMs in 10-fold cross-validation experiments. The results show that the expression status (reflected as GSVA scores) of almost all immune GEMs could be accurately predicted (**Figure 4b**) based on those of non-immune cells. The mean and standard deviation (SD) of R^2 for the predicted and observed values of 39 immune GEMs were 0.75 and 0.11, respectively, and *p*-values for all regression R^2 (except GEM33) were less than 0.001. The results provide evidence that transcriptomic

programs of cells in HNSCC TMEs are coordinated, and such coordination may either be the result of causal intercellular communications, or may be confounded by unknown factors, or both, which can be informed by principled causal analyses.

A global ICN of HNSCC identified by causal Bayesian network learning

To go beyond the correlation analysis, we next set out to find direct support for general causal relationships between the transcriptomic processes, i.e., discovering an ICN. We employed the greedy fast causal inference (GFCI) algorithm [38] , which is capable of learning causal relationships from observational data (under assumptions), including the possibility of unmeasured (hidden) confounding when two or more variables are caused by some hidden process. The GFCI algorithm takes a set of measured variables (e.g., GSVA scores) as input data and returns a partial ancestor graph (PAG) [39], which represents an equivalence class of causal models (see Methods). In our case, a direct causal edge in a PAG from GEM_A to GEM_B across cell types indicates that the state of the transcriptomic process regulating GEM_A in a subpopulation of cells is estimated to be causally influencing the state of the transcriptomic process regulating GEM_B in another subpopulations of cells.

Using the GSVA scores of GEMs from TCGA HNSCC tumors as inputs to GFCI algorithm, we first learned a consensus PAG (see Methods). To concentrate on intercellular communication, we inspected the expression of GEMs in cell populations based on Figure S2 and identified GEMs that were co-expressed in subpopulations. We then constrained the GFCI algorithm to exclude the candidate edges between GEMs that were known to be co-expressed in same subpopulation of cells, and thus, the returned PAG mainly captured the causal of relationships reflecting intercellular communication. Figure 4c shows a subgraph of the consensus PAG (see Methods), and the

complete PAG is shown in Figure S5 and Table S6. Figure 4c uses the immune GEM16 to illustrate causal relationships with other GEMs, where it is shown at the center surrounded by a set of neighboring GEMs (parents, children, and parents of children, which are commonly referred to as Markov blanket). We examined the directions of the causal edges between major cell categories to provide an overview of how different cells communicate with each other in tumor microenvironment of HNSCCs (Table 1), according to our model. The results indicate dense and multi-way intercellular communications among cells in TMEs.

Since experimentally validating the discovered causal edges is beyond the scope of this study, we set out to mathematically test that the learned ICN precisely captured statistical relationships among the transcriptomic processes in diverse types of cells. Conditional independence plays a central role in discovering causal relationships [27, 28, 40]. In a causal Bayesian network, the value of a variable is independent of the remainder of the variables in the network conditioned on the values of its Markov boundary; in other words, the Markov boundary variables serve as the smallest sufficient set of variables for optimally predicting the state of the GEM of interest. The better we can predict the expression status of a GEM based on the expression status of GEMs in its Markov boundary, the more support (although not proof) it provides for any causal edges within the Markov boundary (as shown in **Figure 4c**) being valid [41]. For each immune GEM, we identified its Markov boundary GEMs (including those from all cell types) and used their expression status (GSVA scores) as independent variables to train a regression model to predict the expression status of the immune GEM of interest. Most immune GEMs could be predicted well based on the expression of a few neighboring GEMs discovered by GFCI, with R^2 ranging from 0.36 to 0.97 and the mean and standard deviation of the R^2 values being 0.67 and 0.16, respectively (**Figure 4d**), and the p-values of regressions (except GEM23)

are less than 0.001. Thus, the PAG resulting from our analysis provides a parsimonious representation of an ICN that serves as an hypothesis of the communications between the cells expressing specific GEMs.

Tumor-specific ICNs of individual tumors underlie the heterogeneous tumor microenvironments of HNSCC

While an ICN trained at the tumor cohort level provides a working hypothesis of the general causal relationships among transcriptomic processes, it does not capture the idiosyncratic characteristics of intercellular communications of individual tumors, which we conjecture underlie the diverse TMEs, as shown in Figure 3. We therefore applied an individualized causal discovery algorithm previously developed by us, which we call the individualized GFCI (iGFCI) algorithm [31] (see Methods) and estimated a tumor-specific ICN for each of the 522 HNSCC tumors from TCGA.

Inspecting the individual ICNs, we note that while each ICN is unique, certain patterns of network structure were conserved among subpopulations of tumors. We represented each tumor-specific network with a vector of causal edges among GEMs, in which element i =1 if and only if that network contained edge i. We then performed consensus clustering analyses of these networks. Four distinct clusters of networks emerged (**Figure 5a**), with certain edges being distinguishably present or absent from some clusters of tumors (**Figure 5b**). The patients from the 4 subtypes exhibited significantly different survival outcomes (**Figure 5c**).

Our results also showed that the presence/absence of the certain edges were significantly associated with patient outcomes (**Figure 5d**), suggesting that communication between cells may influence the overall immune responses. Of interest, the edge linking immGEM16 and

immGEM19, expressed in Treg cells and dendritic cells (DCs) (**Figure S6**), respectively, is associated with better prognosis for cancer patients. This result corroborates previous knowledge [42]. Another example is the edge between immGEM3 and immGEM35. ImmGEM3 and immGEM35 were highly expressed in two CD8 T cell subtypes, both of which expressed high-levels of immune checkpoint genes such as PD-1, LAG3 and TIM3 (**Figure S6**). This result suggests that communication between these two CD8 T cell subtypes is involved in tumor immune surveillance [43, 44], and such communication is associated with better outcomes.

We further investigated whether the distinct ICN patterns were associated with distinct immune environments, to explore the hypothesis that an immune environment is defined by a subtype type of ICN. We compared the overlap of cluster membership of tumors when they were clustered based on immune GEMs (Figure 3a) and causal network edges (Figure 5a), respectively; we found that the tumor clusters derived from two the clustering analyses share a significant number of tumors (**Figure S7**) (chi-square test p-value < 2.2E-16). Particularly, the Network Cluster # 1 closely aligns with the GEM-based Cluster # 4, and patients assigned to these clusters have the best outcomes compared to patients assigned to the other clusters; similarly, Network Cluster #2 aligns best with the GEM-based Cluster #2, and the patients assigned to these clusters exhibited the worst outcomes compared to patients assigned to the other clusters. The results indicate that certain communication channels between different types of cells are highly conserved among tumors, and it is such multi-way communications that may define a TME.

Discussion

In this study, we have developed a framework for estimating tumor-specific ICNs, which involves two major conceptual steps: Firstly, mining single cell transcriptomes to identify GEMs to serve as surrogate representations of cellular signaling processes. Secondly, modeling intercellular communication as a causal network between these processes to estimate intercellular communication at the level of individual tumors. We anticipate that the capability of inferring ICN, particularly tumor-specific ICNs, will help lay a foundation for future mechanistic investigations of how intercellular communications define the state of cells in a TME, and in turn, how tumors evade the immune surveillance of hosts.

While it is well-appreciated that cells within a TME actively communicate and influence the functional states of each other [19-21], few computational methods have been developed to analyze the fundamental nature of an ICN, especially the causal effects of communication. Our work represents an effort in this direction. Our results show that the inferred ICNs serve as concise statistical models that explain well the coordinated transcriptomic programs across different types of cells and that tumor-specific ICNs provide a new approach to investigate the heterogeneity of immune environments in cancer. Under assumptions (see Methods), the methods we applied identified communication channels (directed edges) of learned ICNs that capture causal effects of intercellular communications, beyond accurately modeling the joint distribution of cellular events (controlled expression of GEMs) in different cells. However, we realize that those assumptions may not always hold in a complex system as the ones we are studying, and thus, we view the learned ICNs as hypotheses, which can be experimentally tested in future.

In this study, we illustrated the capability and utility of our methods in studying ICNs in the cancer (HNSCC) domain. The principled framework introduced here is quite general and can

be applied in wide range of biomedical domains to study communications of cells in a tissue or organ. With the advent of more single-cell omics data, particularly in situ spatial omics studies of tissues [15, 45, 46], the framework we present can be used to study intercellular communications under a wide range of physiological (e.g., embryo development) and pathological (e.g., rheumatism or aplastic anemia) settings besides cancer, which provides many opportunities for future research.

Methods

Single cell RNA sequencing data processing

FASTQ files were processed with the CellRanger software (10X Inc, version 2.0). We aggregated the 46 samples of all the 18 patients, including samples from CD45+/- tumor specimen and peripheral blood. After the aggregation and default depth normalization using the CellRanger, we collected the count matrix of 103,006 cells in total, with a post-normalization mean reads per cell of 31,303 and median genes per cell of 1,005. We then processed the aggregated count matrix in Scanpy [47]. We filtered cells and genes with cells expressing at least 200 and at most 5000 genes. Cells with more than 5 percent of genes belonging to mitochondria genes were removed. Genes were then filtered with dispersion range (minimum mean of 0.0125, maximum mean of 5 and minimum dispersion of 0.1). Scanpy also provides the implementation of feature reduction and visualization called UMAP [48, 49]. We visualized all the cells using in UMAP in which cells are pseudo-colored to mark cell types based canonical gene signatures of general cell types or manual labels.

NHDP modeling on single cell RNAseq data

We applied the NHDP on scRNAseq data to identify the gene modules which potentially indicate different biological processes. The NHDP method allows users to define a hierarchical structure and the algorithm will allocate modules along the tree. In this study, we designed a three-layer hierarchical tree structure, with branching factors of 5, 4, and 3 from the root to the 2nd and 3rd layers. That is, starting from the root, the 1st layer had 5 nodes; each node in the 1st layer is connected to 4 child nodes in the 2nd layer; each node in 2nd layer is connected to 3 child nodes in the 3rd layer. Each node defines a distribution over the space of genes reflecting the information

on which genes are commonly assigned to the module. Input to the NHDP model is a binarized expression matrix of scRNAseq data, where rows are individual cells and columns are genes. An element of the matrix was set to "1" if the corresponding gene is expressed in a given cell, and "0" otherwise. The algorithm generates two outputs: 1) For each node in the hierarchical tree, a distribution over the space of genes is returned. From the distribution, one can extract a ranked list of genes most commonly assigned to the node (GEM). 2) A cell-by-GEM count matrix is returned. An element of the matrix reflects the number of genes expressed in a given cell that are assigned to a specific GEM. This is equivalent to projecting a cell into the space of GEMs, which enabled us to search cell subtypes through clustering analysis. Certain GEMs were not populated by the NHDP, too few genes were assigned to the GEM in cells, and they were excluded from further modeling.

Enrichment of GEMs in bulk RNAseq and tumor subtype identification

We analyzed the enrichment of all the immune and non-immune GEMs in three bulk RNAseq datasets, namely, TCGA HNSCC tumors and two independent validation sets from the GEO database (GSE39366, GSE27020) [50]. Genes within a GEM are ranked according to the probability mass assigned to them by the NHDP model, and we used the top 50 genes of a GEM to represent its signature co-expression profile. We applied the gene set variation analysis (GSVA) method [37] to estimate the enrichment of a GEM in a tumor relative to a cohort of tumor samples. The distribution of the GSVA scores among the input tumor samples generally followed a normal distribution in the range of -3 to 3.

With the estimated GSVA scores for all the immune GEMs and cell subtypes in TCGA HNSCC tumor samples, we applied consensus clustering (implemented in the R package as ConsensusClusterPlus [51]) and identified subtypes of tumor samples with distinct enrichment patterns of immune GEMs. The optimal number of clusters were chosen using the relative change in area under the empirical cumulative density function (CDF) curve [52], which shows the relative increase in consensus and cluster number at which there is no appreciable increase.

In addition to GEMs, we also calculated the GSVA enrichment scores of immune cell subtypes, which were identified in **Figure S4**. The signatures of each cell subtype were curated based on the GEMs that were enriched in the corresponding cell subtypes; in particular, we used the union set of top30 genes of the GEMs enriched in individual cell subtypes. Following the same procedure, we estimated the cell-subtype enrichment GSVA scores in bulk RNA data and performed consensus clustering to identify tumor clusters with distinct composition of cell subtypes.

Evaluate association between patient survival and GEMs or cell subtypes. We performed Cox Proportional-Hazards regression (*coxph* function of R language) using both the immune GEMs and cell subtypes enrichment scores as input to evaluate which GEMs or cell subtypes were significantly associated with patient survival outcomes. We conducted both individual Coxph regression on individual immune GEMs and alternatively on cell types. To do so, we used a stepwise selection of multiple variates to obtain a minimum and optimized set of immune GEMs or cell types that could fit the survival curves well.

Cross-cell-type GEM correlations analysis

We examined the correlations among the enrichment scores of GEMs with both Pearson correlation and multiple regression models. Across all cell categories (immune, epithelial, fibroblast, and endothelial), we performed Pearson correlation analysis of the expression levels between all pairs of GEMs. We excluded the intra-cell-category correlation analysis to avoid potential confounding effects of coordinated cell-type-specific expression programs. We also applied the GLMNET [53] package from the R language to train regression models, which predict each immune GEM using non-immune GEMs from epithelial, endothelial, and fibroblast cells as independent variables. We performed 10-fold cross-validation experiments, in which we used 9 folds of data to train a regression model and used the trained model to predict immune GEMs for the tumors in the held-out fold. We then evaluated the performance of the regression models by performing a correlation analysis using predicted GEM-expression values versus observed GEM expression values and recorded R^2 values.

Model intercellular communication network using causal discovery methods

In this study, we set out to model the intercellular communication (i.e., signal transduction across the cell boundary). Here, we introduce the biological problem (**Figure M1A**), which our methods aim to address and how this problem is transformed into a causal Bayesian network problem. We then introduce the computational methods for learning ICNs in detail.

When the state of signaling pathway X changes in a cell (Cell A), either due to physiological signal transduction (e.g., signal received from environment, W) or pathological perturbation (e.g., a mutation of a gene affecting the pathway, V), its signal may be communicated to another type of cell (Cell B) through paracrine or ligand-receptor interaction at the cell surface. The goal of modeling intercellular communication is essentially to model

whether the state of the pathway *X* in Cell A causally influences the state of the pathway *Y* in Cell B. However, it is usually infeasible to measure the states of pathways (i.e., *X* and *Y* are latent variables), making the task of modeling intercellular communication challenging.

To address this challenge, we hypothesize that if changed states of pathway *X* in Cell A is closely associated with a transcriptomic change of the expression of *GEM-X*, we can use the expression status of *GEM-X* as a surrogate to represent the state of *X.* Similarly, if changed state of pathway *Y* in Cell B is closely associated with transcriptomic changes of *GEM-Y*, we can use expression status of *GEM-Y* as a proxy to represent the state of pathway *Y*. Under such a setting, we can model the causal relationships between *X* and *Y* as a *causal Bayesian network* (CBN) as shown in **Figure M1B**. Here, *V, W* can be any type of measured variable. For example, they can be the mutation status of a gene, or they can be measured concentrations of signaling molecules. More relevant to this study, they can be extracellular signals transmitted to the Cell A from other cells through cell-cell communication in the same fashion as between *X* and *Y*. In the latter case, the states of pathways *V* and *W* in those cells can be represented by their corresponding GEMs. Under such a setting, we can search for a ICN (i.e., intercellular signal transduction) between latent signaling pathways across cells using GEMs as surrogates of the pathway, like the one shown in **Figure S5**.

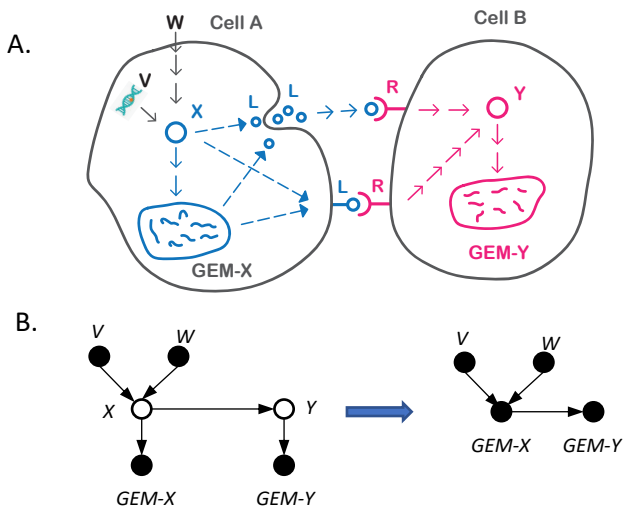


Figure M1. Causal modeling of an intercellular cellular communication network. A. A biological scenario to be modeled by our framework. The activation states of the pathway *X* in Cell A can be affected by physiological signal (e.g., an extracellular signal *W*) or genomic alteration (e.g., *V*). Activation of pathway *X* in Cell A may lead to release or cell-surface presentation of a ligand *L*, which may be produced through *de novo* transcription. Ligand-receptor interaction (LR) leads to activation of pathway *Y* in Cell B. Activation of *X* and *Y* may be deterministically associated with expression of corresponding GEMs. **B. A Causal Bayesian network (CBN) for modeling the signal transduction depicted in A**. The CBN on the left side explicitly represents the relationships of pathways *X* and *Y*, their upstream signals (*V* and *W*), and the GEMs regulated by them. In a CBN, observed variables are represented as filled nodes, and latent variables are represented as open nodes. In general, variables *V* and *W* can be any type of observed variable, including the expression values of GEMs reflecting the states of corresponding latent signaling pathways in other cells. Under the above assumption, we can use *GEM-X* and *GEM-Y* as proxies of latent variables *X* and *Y*, such that the causal edge between *X* and *Y* can be represented and estimated as a causal edge between *GEM-X* and *GEM-Y* as reported in this work.

From a causal modeling perspective, the expression status of *GEM-X* and *GEM-Y* mainly serve as surrogates of the state of pathways (*X* and *Y*) regulating them. Although it is of value to investigate the functions of genes in each GEM to gain biological insights, interpreting the function of genes in GEMs is not required to model the causal relationships among pathways regulating their expression.

Background on causal modeling and discovery

In our investigation, we used CBNs as the underlying representation of causal relationships. Thus, in this section we first define Bayesian networks (BNs) and then define CBNs as a special version of them. Next, we describe a graphical model called a *maximal ancestral graph* (MAG) that can represent CBNs when they contain unmeasured confounding.[1] We provide a brief overview of constraint-based causal discovery methods, which test for conditional independence relationships (constraints) in the data and identify a set of MAGs that are consistent with those constraints. We introduce a graphical model called a *partial ancestral graph* (PAG) that graphically represents the set of MAGs that all share the same constraints. If a causal arc appears in a PAG, then all the MAGs consistent with the data-derived constraints contain that causal arc, and thus, that causal relationship is supported by the data.

Causal modeling methods

A BN consists of a structural model and a set of probabilities [55-61]. The structural model is a directed acyclic graph (DAG) in which nodes represent variables and arcs represent probabilistic dependence. For convenience, we use the terms *node* and *variable* interchangeably in this paper. Each node can represent a continuous or discrete variable, although we focus on using discrete variables in this paper. For each node X_i there is a probability distribution on that node given the state of its parents, which are the nodes with arcs into X_i. The following equation indicates how the node probabilities of a BN specify the joint probability distribution over all *n* nodes that are being modeled:

[1] It is also possible to graphically model and learn selection bias (54. Spirtes P, Meek C, Richardson T, editors. Causal inference in the presence of latent variables and selection bias. The Conference on Uncertainty in Artificial Intelligence; 1995.) although we assume no such bias in the data we analyzed in this study, and thus, in this section we do not describe modeling and learning under selection bias.

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid parents(X_i)). \quad (1)$$

The value of a variable may represent any aspect of the modeled entity. The value of any particular variable may be measured or missing. If the value of a variable is always missing, then we say this is an *unmeasured*, *hidden,* or *latent variable*; otherwise, we say it is a *measured or observed variable*. Figure M2 illustrates the structure of a hypothetical BN involving pneumonia and some of its possible consequences.

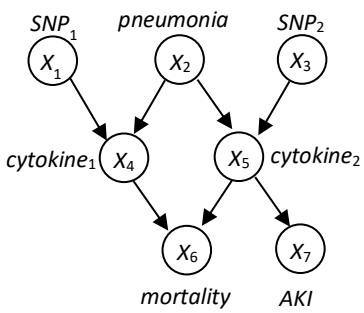


Figure M2. A hypothetical medical Bayesian network structure with 7 variables represented as nodes. This structure is presented for illustration only and is not intended to present an accurate and complete model of the causal relationships involving pneumonia and mortality. The variables *cytokine*$_1$ and *cytokine*$_2$ represent two specific types of circulating cytokines associated with inflammatory disease. *SNP*$_1$ and *SNP*$_2$ are particular single nucleotide polymorphisms (SNPs) that represent genetic point variations. *AKI* is acute kidney injury.

Since a BN encodes a joint probability distribution, it represents all the information needed to compute any marginal or conditional probability on the nodes in the network model [62]. A variety of algorithms have been developed for computing $P(\boldsymbol{A} \mid \boldsymbol{B})$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are arbitrary sets of variables with assigned values [60, 61, 63, 64].

A *causal Bayesian network* (CBN) is a Bayesian network in which each arc is interpreted as a direct causal influence between a parent node (a cause) and a child node (an effect), relative to the other nodes in the network [59, 65-67]. A CBN structure encodes the following primary set of independence relationships, which is called the *local causal Markov condition* (LCMC) [67]: Each node is independent of its non-effects (non-descendants), given the state of its direct causes

(parents). The local causal Markov condition implies the *global causal Markov condition* (GCMC) which is a graphical set of rules (called d-separation rules) that determine whether X is independent of Y given $\mathbf{Z}$, for every pair of nodes X and Y and every set of nodes $\mathbf{Z}$ [67].

An unmeasured confounder is an unmeasured variable that directly causes two or more measured variables. Unmeasured confounders are common in biological data, and thus, modeling them is important. A MAG [39] is an extension of a CBN structure in which $X \leftrightarrow Y$ represents the presence of an unmeasured confounder between nodes X and Y. There is an associated extension of the LCMC that applies to MAGs, as well as an extension of d-separation that is called m-separation [68]. If the node *pneumonia* in Figure M2 was unmeasured, then it would be an unmeasured confounder of $cytokine_1$ and $cytokine_2$; Figure M3 shows the corresponding MAG.

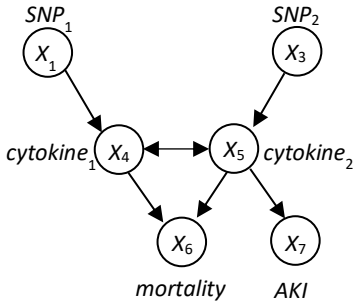


Figure M3. A MAG representing the causal model in Figure M2 when the node *pneumonia* is unmeasured.

Causal Discovery Methods

We are interested in learning causal relationships from observational data [69], such as the data we analyzed for this paper. To do so, we used a causal discovery algorithm that is based on the Fast Causal Inference (FCI) algorithm [59, 70]. FCI is one of the most well studied and frequently applied causal discovery algorithms that models unmeasured confounding. Given an observational dataset, it performs a series of statistical tests of the form "*Is X* independent of *Y*

given **Z**", which are called *constraints*. Based on the result of one test (and of other previous tests), FCI will ask for the result of another test. The algorithm has a highly efficient method for performing a series of statistical tests, such that the total set is relatively small. It also starts with low dimensional sets **Z** in order to maximize statistical testing power. FCI continues to perform such tests until no additional testing will further resolve the causal relationships among the measured variables. It then returns a set **C** of causal models (MAGs) that all satisfy the constraints that were tested. Each of the causal edges $X \rightarrow Y$ that appear in all of the MAGs in **C** are supported by the data as being present. FCI outputs a compact, graphical representation of the edges of the MAGs in **C**, which is called a PAG [39]. There are two primary edge types in PAGs: $X \rightarrow Y$ [2] represents causation, and $X \leftrightarrow Y$ represents unmeasured confounding. There are also two additional edge types in PAGs: X o→ Y represents that the relationship could be either $X \rightarrow Y$ or $X \leftrightarrow Y$, and X o—o Y represents that the relationship could be either $X \rightarrow Y$, $X \leftarrow Y$, or $X \leftrightarrow Y$. The absence of an edge between X and Y denotes that none of these relationships are modeled as existing. Table 2 provides a more detailed description of the interpretation of edges in a PAG; this table originally appeared in an appendix written by Peter Spirtes in the article available at [71].

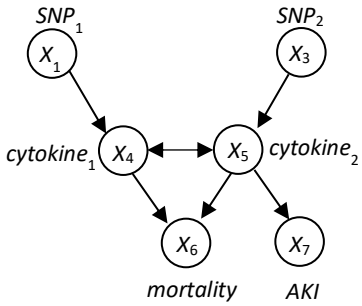
Table 2. A description of the edge types in a PAG and their causal interpretation.

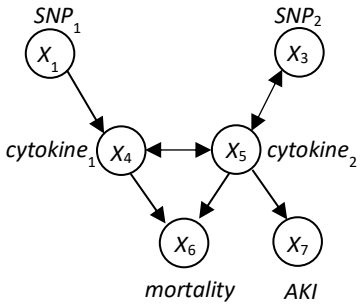
Edge type	Relationships that are present	Relationships that are absent
$X \rightarrow Y$	X is a cause of Y. It may be a direct or indirect cause that may include other measured variables. Also, there may be an unmeasured confounder of X and Y.	Y is not a cause of X.
$X \leftrightarrow Y$	There is an unmeasured variable (call it L) that is a cause of X and Y. There may be measured	X is not a cause of Y. Y is not a cause of X.

[2] $X \leftarrow Y$ is possible as well, but does not illustrate a fundamentally different edge type.

	variables along the causal pathway from *L* to *X* or from *L* to *Y*.	
X o→ *Y*	Either *X* is a cause of *Y*, or there is an unmeasured variable that is a cause of *X* and *Y*, or both.	*Y* is not a cause of *X*.
X o–o *Y*	Exactly one of the following holds: (a) *X* is a cause of *Y,* or (b) *Y* is a cause of *X,* or (c) there is an unmeasured variable that is a cause of *X* and *Y,* or (d) both a and c, or (e) both b and c.	

As an example, consider that the causal process generating the observational data available to us is accurately modeled by a CBN with the structure shown in Figure M2, where pneumonia is an unmeasured variable. Suppose that an analysis of the data yields constraints consistent with that structure. The MAGs consistent with those constraints are shown in Figure M4. The PAG representing those four MAGs is shown in Figure M5. That PAG contains all of the causal relationships among the measured variables in Figure M2, except that the relationship between each SNP and its associated cytokine includes several possibilities; nonetheless, the PAG represents that each SNP <u>may</u> cause its associated cytokine and the cytokine does not cause the SNP.





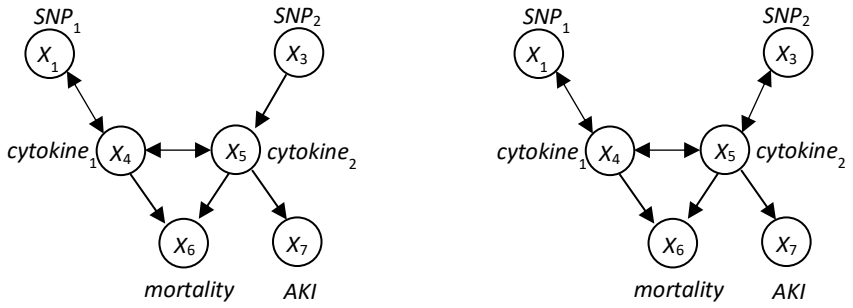


Figure M4. The MAGs consistent with the CBN shown in Figure M2, assuming *pneumonia* is an unmeasured variable.

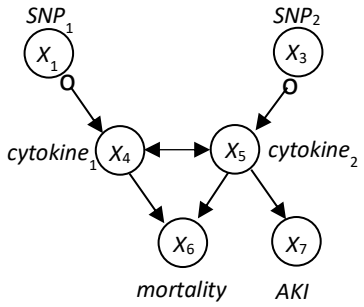


Figure M5. A PAG representing the four MAGs shown in Figure M4.

Even in the best of circumstances (see Causal Discovery Assumptions below), it often is not possible to learn from constraints many of the causal relationships among a set of measured variables. Nonetheless, identifying just a few causal relationships that are novel and valid may be helpful in supporting scientific inquiry. Figure S5 shows that the causal discovery method we applied output many possible causal relationships that serve as causal hypotheses. In general, the FCI algorithm, and many of its extensions, such as GFCI (see below), have been proven under assumptions to discover only and all the causal relationships that can be discovered from valid conditional independence constraints [39], and in that sense they are sound and complete.

Among causal discovery methods in use in biomedicine, the approach used by FCI is very similar to the one used by the more recently developed Causal Inference Test (CIT) method [72, 73]. CIT is a constraint-based method that typically evaluates four constraints on three

variables: a genomic variable (e.g., a SNP), a variable that represents an intermediate cellular state or process (e.g., an expression variable), and a variable that represents a phenotype of interest. A search of Google Scholar on 01/16/2022 using the phrase "causal inference test" yielded over 300 articles, including analyses of observational biomedical data for causal relationships in obesity, rheumatoid arthritis, psoriasis, ovarian cancer, autism, type 2 diabetes, schizophrenia, Alzheimer's disease, and other diseases and conditions. FCI also can be applied using only three variables, in which case it operates very similar to CIT. More generally, however, FCI can handle causal discovery using thousands of variables. In doing so, it identifies potentially complex patterns of constraints among those variables that support causal relationships that cannot be found using only three variables.

In the study reported here, we applied the GFCI algorithm [74], which is an extension of FCI. The first phase of GFCI learns a BN structure (DAG)[3] that does not model unmeasured variables, using a run-time-optimized version [75] of the GES algorithm [76]. That BN is used to constrain the relationships among the measured variables considered by FCI.[4] Although GES is a greedy algorithm, under reasonable assumptions it has been proven to return the correct BN structure if given sufficient data [76]. Although no statistical algorithm can be guaranteed to return a correct model when given finite data, as is always the case in reality, having an algorithm that is provably correct in the large sample limit is generally considered to be an important property for it to have. In the second phase of GFCI, it applies the FCI algorithm, which begins its search relative to the constraints found by GES in the first phase.

[3] More precisely, it returns a Markov equivalence class of BNs that cannot be distinguished by independence constraints.
[4] Traditionally, the FCI algorithm begins its PAG search without any constraints on the causal relationships among the measured variables.

For any method to discover causal relationships from observational data, assumptions are required. For GFCI, the key assumptions are as follows: (1) The causal process being modeled can be represented by a CBN, which possibly contains unmeasured variables. Thus, the structure of the causal relationships can be represented by a DAG. (2) The available observational data can be viewed as having been generated from sampling the joint probability distribution of some CBN. Thus, there is no selection bias due to some types of cases being selectively missing. (3) The independence constraints represented by a CBN structure imply independence constraints in the probability distribution represented by the CBN. 4) The dependence constraints represented by a CBN structure (i.e., where independence constraints do not occur) imply dependence constraints in the probability distribution represented by the CBN. (5) Statistical tests of independence on the available observational data are correct, relative to the independence constraints represented by the data-generating CBN.

Statistical testing on finite data is imperfect, and thus, so are the constraints provided to GFCI. Therefore, as a form of robustness analysis, we applied GFCI multiple times using bootstrapping [77]. For each of *m* bootstrap datasets of the observational data, we ran GFCI to generate a PAG. Over the *m* PAGs produced, for each pair of measured variables we report a causal relationship for that pair if that relationship occurred in at least 20% of the bootstrap samples.

Application of causal discovery methods

We employed the GFCI algorithm to infer from data whether one GEM is supported as causally influencing the expression of another GEM, particularly across cell types. We discretized GSVA scores of a GEM to 0, 1, 2, based on the following percentile brackets: [0 - 25%], (25% - 75%],

and (75% - 100%]. For estimating a population-wide ICN, we applied the discrete version of GFCI algorithm implemented in the Tetrad Package [78]. Unmeasured confounders are likely to be common at the cellular level, and thus, the ability of GFCI to model them is important. On the other hand, there may be patterns of association among GEMs that strongly support that one GEM causes the expression of another; we are particularly interested in analyzing the data for such results.

The Tetrad implementation of the GFCI algorithm allows a user to input prior knowledge regarding the network, such as edges to be included, as well as those to be excluded. To avoid confounding effects of cell-type-specific co-expression of GEMs, we first examined the distribution of GEMs in cells based on the results shown in **Figure S2** to identify a list of pairs of GEMs that were co-expressed in a set of cells. This list was input to GFCI as prior knowledge, so that the algorithm would exclude edges between a pair of GEMs in the list. Using the GEM data of TCGA HNSCC tumors as inputs, we produced 50 bootstrap samples by drawing 90% of the tumors each time with replacement. The first phase of GFCI used the Bayesian information criterion with default settings and a penalty discount of 1.0 (see "sem-bic" in the Tetrad manual [79]), and the second phase used the Fisher-*z*-test with alpha = 0.01. We created a consensus causal PAG by keeping the edges that were conserved more than 20% of the time, according to the bootstrap results. [5]

Discovering tumor-specific intercellular communication networks

[5] There were 15 cases in which both X→ Y and X ← Y occurred more than 20% of the time in the bootstrap results. In those cases, we did not enter these relationships into the consensus PAG.

This section first provides background on the methods we developed for performing individualized causal modeling and discovery. It then describes how we applied those methods to the study data to learn tumor-specific ICNs.

Background on individualized causal modeling and discovery

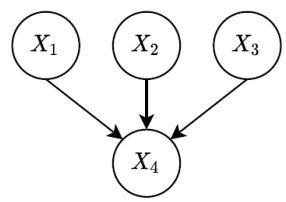
In our investigation, we also used a method that learns causal models that are individualized or specialized to a given tumor. In doing so, we seek to learn more accurate models of the causal relationships in individual tumors and how these relationships vary among tumors. The method we developed [80] and applied considers the values **V** of the variables that characterize a current tumor for which we wish to learn a causal model. In performing constraint-based causal discovery, the method uses those values and a training set of data on other tumors to learn constraints that are individualized to **V**. Those individualized constraints then lead GFCI to learn an individualized causal model; this individualized version of GFCI we call iGFCI. In this section, we provide a high level overview of the iGFCI algorithm; additional details are available in [80].

iGFCI uses variations of the GES and FCI algorithms used in GFCI. As discussed above, GES learns a BN. The individualized version of GES [81], which is used in iGFCI, learns a BN that represents context-specific independence (CSI) [82] wherein the variables that are the parents of a node depend on the values of **V**. Figure M6 shows an example of CSI in a BN. Suppose that $T = \{ X_1 = 1, X_2 = 1, X_3 = 0\}$. For this instance T, Figure M6b indicates that X_4 is independent of X_3 given $X_1 = 1$ and $X_2 = 1$; this independence implies that for this instance, the parents of X_4 are just X_1 and X_2, as shown in Figure M6b. Figure M6c shows the PAG that can be learned in the large sample limit from data generated by the CBN shown in Figure M6b, under the assumptions

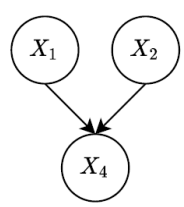
discussed above; with additional measured variables or background knowledge, it may be possible to further resolve the causal relationships among these three variables.

The second phase of iGFCI applies a modified version of FCI. The traditional version of FCI tests whether X is independent of Y for all values of **Z**. In contrast, iGFCI includes a context-specific test of conditional independence that assesses whether variable X is independent of variable Y given **Z** = **z**, where **z** are the values given by **V** for the variables in conditioning set **Z**. In this way, iGFCI learns a PAG that is individualized with respect to the values of the variables that characterize tumor T.

$X1, X_2, X_3$	$P(X_4 \mid X_1, X_2, X_3)$
0, 0, 0	0.90
0, 0, 1	0.90
0, 1, 0	0.90
0, 1, 1	0.90
1, 0, 0	0.32
1, 0, 1	0.78
1, 1, 0	0.13
1, 1, 1	0.13

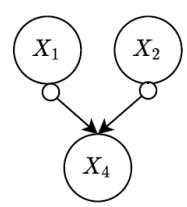


(a)



$T = \{X_1 = 1, X_2 = 1, X_3 = 0\}$

(b)



$T = \{X_1 = 1, X_2 = 1, X_3 = 0\}$

(c)

Figure M6. A simple, hypothetical example of CSI. (**a**) The data-generating CBN for variable X_4 given its parents X_1, X_2, X_3; (**b**) the individualized CBN that represents the CSI relationship $X_4 \perp X_3 | (X_1 = 1, X_2 = 1)$; (**c**) the individualized PAG that is learnable in the large sample limit from observational data generated by the CBN in **b**, given the assumptions described above.

Application of an Individualized Causal Discovery Method

To investigate the mechanisms underlying heterogeneous immune environments among tumors, we applied iGFCI to the GEM variables derived from bulk TCGA HNSCC tumors, as described above. Since the GEMs are continuous variables and iGFCI currently operates on discrete variables, we discretized the GSVA scores of a GEM to 0, 1, 2, based on the following percentile binnings: [0 - 25%], (25% - 75%], (75% - 100%].

We used bootstrap sampling to enhance the robustness of the results generated by iGFCI. In particular, for each tumor, we applied iGFCI to 20 bootstrap datasets that were generated from the original dataset, where each bootstrap dataset was obtained by 90% sampling of the original dataset with replacement. Using the bootstrap results for a given tumor *T*, we derived the conserved edges of type $X \rightarrow Y$. In particular, we kept the edges that were present in more than two bootstrap-derived PAGs from among the 20 individualized PAGs derived for tumor *T*. The union of the retained causal edges yielded an individualized, "consensus" intercellular communication network (ICN) for each tumor *T*. Specifically, an element of the causal-edge ICN vector for a tumor T was set to "1" if the causal edge was present in the tumor-specific intercellular communication network for *T*, and set to "0" otherwise.

Finding patterns among tumor-specific ICNs

We clustered the tumor samples based on their individual ICNs learned using iGFCI to obtain novel subtypes of HNSCC with different cellular communication patterns. Each tumor is represented as a vector over the space of the union of causal edges of the tumor-specific intercellular communication networks as described in the previous subsection. We applied the

ConsensusClusterPlus package from the R Bioconductor Consortium to cluster these vectors and grouped tumors based on their cluster assignments.

Acknowledgements
We thank Peter Spirtes and Bryan Andrews for comments on the text.

References

1. Spranger S, Bao R, Gajewski TF. Melanoma-intrinsic beta-catenin signalling prevents anti-tumour immunity. Nature. 2015;523(7559):231-5. doi: 10.1038/nature14404. PubMed PMID: 25970248.
2. Collins DC, Chenard-Poirier M, Lopez JS. The PI3K Pathway at the Crossroads of Cancer and the Immune System: Strategies for Next Generation Immunotherapy Combinations. Curr Cancer Drug Targets. 2018;18(4):355-64. Epub 2017/10/04. doi: 10.2174/1568009617666170927114440. PubMed PMID: 28969535.
3. Concha-Benavente F, Ferris RL. Oncogenic growth factor signaling mediating tumor escape from cellular immunity. Curr Opin Immunol. 2017;45:52-9. Epub 2017/02/17. doi: 10.1016/j.coi.2017.01.004. PubMed PMID: 28208102; PubMed Central PMCID: PMCPMC5449256.
4. Seton-Rogers S. Oncogenes: Driving immune evasion. Nat Rev Cancer. 2018;18(2):67. Epub 2018/01/26. doi: 10.1038/nrc.2018.5. PubMed PMID: 29368742.
5. Spranger S, Gajewski TF. Mechanisms of tumor cell–intrinsic immune evasion. Annual Review of Cancer Biology. 2018;2:213-8.
6. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su MJ, Melms JC, et al. A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. Cell. 2018;175(4):984-97 e24. Epub 2018/11/06. doi: 10.1016/j.cell.2018.09.006. PubMed PMID: 30388455; PubMed Central PMCID: PMCPMC6410377.
7. Robert C. A decade of immune-checkpoint inhibitors in cancer therapy. Nat Commun. 2020;11(1):3801. Epub 2020/08/01. doi: 10.1038/s41467-020-17670-y. PubMed PMID: 32732879; PubMed Central PMCID: PMCPMC7393098.
8. Hodi FS, O'Day SJ, McDermott DF, Weber RW, Sosman JA, Haanen JB, et al. Improved survival with ipilimumab in patients with metastatic melanoma. N Engl J Med. 2010;363(8):711-23. Epub 2010/06/08. doi: 10.1056/NEJMoa1003466. PubMed PMID: 20525992; PubMed Central PMCID: PMCPMC3549297.
9. Robert C, Ribas A, Wolchok JD, Hodi FS, Hamid O, Kefford R, et al. Anti-programmed-death-receptor-1 treatment with pembrolizumab in ipilimumab-refractory advanced melanoma: a randomised dose-comparison cohort of a phase 1 trial. Lancet. 2014;384(9948):1109-17. Epub 2014/07/19. doi: 10.1016/S0140-6736(14)60958-2. PubMed PMID: 25034862.
10. Chen DS, Mellman I. Elements of cancer immunity and the cancer-immune set point. Nature. 2017;541(7637):321-30. Epub 2017/01/20. doi: 10.1038/nature21349. PubMed PMID: 28102259.
11. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, et al. The Immune Landscape of Cancer. Immunity. 2018;48(4):812-30 e14. Epub 2018/04/10. doi: 10.1016/j.immuni.2018.03.023. PubMed PMID: 29628290; PubMed Central PMCID: PMCPMC5982584.
12. Nishida-Aoki N, Gujral TS. Emerging approaches to study cell-cell interactions in tumor microenvironment. Oncotarget. 2019;10(7):785-97. Epub 2019/02/19. doi: 10.18632/oncotarget.26585. PubMed PMID: 30774780; PubMed Central PMCID: PMCPMC6366828.

13. Schwager SC, Taufalele PV, Reinhart-King CA. Cell-Cell Mechanical Communication in Cancer. Cell Mol Bioeng. 2019;12(1):1-14. Epub 2019/10/01. doi: 10.1007/s12195-018-00564-x. PubMed PMID: 31565083; PubMed Central PMCID: PMCPMC6764766.
14. Song D, Yang D, Powell CA, Wang X. Cell-cell communication: old mystery and new opportunity. Cell Biol Toxicol. 2019;35(2):89-93. Epub 2019/03/01. doi: 10.1007/s10565-019-09470-y. PubMed PMID: 30815784.
15. Burgess DJ. Spatial transcriptomics coming of age. Nature reviews Genetics. 2019;20(6):317. Epub 2019/04/14. doi: 10.1038/s41576-019-0129-z. PubMed PMID: 30980030.
16. Eng CL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. Nature. 2019;568(7751):235-9. Epub 2019/03/27. doi: 10.1038/s41586-019-1049-y. PubMed PMID: 30911168; PubMed Central PMCID: PMCPMC6544023.
17. Lasken RS. Single-cell sequencing in its prime. Nature biotechnology. 2013;31(3):211-2. Epub 2013/03/09. doi: 10.1038/nbt.2523. PubMed PMID: 23471069.
18. Ray LB. Single-cell biology. Cells go solo. Introduction. Science. 2013;342(6163):1187. Epub 2013/12/07. doi: 10.1126/science.342.6163.1187. PubMed PMID: 24311679.
19. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell. 2017;171(7):1611-24 e24. Epub 2017/12/05. doi: 10.1016/j.cell.2017.10.044. PubMed PMID: 29198524.
20. Cillo AR, Kurten CHL, Tabib T, Qi Z, Onkar S, Wang T, et al. Immune Landscape of Viral- and Carcinogen-Driven Head and Neck Cancer. Immunity. 2020;52(1):183-99 e9. Epub 2020/01/12. doi: 10.1016/j.immuni.2019.11.014. PubMed PMID: 31924475; PubMed Central PMCID: PMCPMC7201194.
21. Kumar MP, Du J, Lagoudas G, Jiao Y, Sawyer A, Drummond DC, et al. Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics. Cell Rep. 2018;25(6):1458-68 e4. Epub 2018/11/08. doi: 10.1016/j.celrep.2018.10.047. PubMed PMID: 30404002.
22. Hou R, Denisenko E, Ong HT, Ramilowski JA, Forrest ARR. Predicting cell-to-cell communication networks using NATMI. Nat Commun. 2020;11(1):5011. Epub 2020/10/08. doi: 10.1038/s41467-020-18873-z. PubMed PMID: 33024107; PubMed Central PMCID: PMCPMC7538930.
23. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. Nat Protoc. 2020;15(4):1484-506. Epub 2020/02/28. doi: 10.1038/s41596-020-0292-x. PubMed PMID: 32103204.
24. Camp JG, Sekine K, Gerber T, Loeffler-Wirth H, Binder H, Gac M, et al. Multilineage communication regulates human liver bud development from pluripotency. Nature. 2017;546(7659):533-8. Epub 2017/06/15. doi: 10.1038/nature22796. PubMed PMID: 28614297.
25. Almet AA, Cang Z, Jin S, Nie Q. The landscape of cell–cell communication through single-cell transcriptomics. Current Opinion in Systems Biology. 2021;26:12-23.

26. Armingol E, Officer A, Harismendy O, Lewis NE. Deciphering cell-cell interactions and communication from gene expression. Nature reviews Genetics. 2021;22(2):71-88. Epub 2020/11/11. doi: 10.1038/s41576-020-00292-x. PubMed PMID: 33168968; PubMed Central PMCID: PMCPMC7649713.
27. Pearl J, Glymour M, Jewell N. Causal Inference in Statistics: A Primer Wiley; 2016.
28. Zhang K, Scholkopf B, Spirtes P, Glymour C. Learning causality and causality-related learning: some recent progress. Natl Sci Rev. 2018;5(1):26-9. Epub 2018/07/24. doi: 10.1093/nsr/nwx137. PubMed PMID: 30034911; PubMed Central PMCID: PMCPMC6051411.
29. Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. . Frontiers in Genetics. 2019;10:524.
30. Cai C, Cooper GF, Lu KN, Ma X, Xu S, Zhao Z, et al. Systematic discovery of the functional impact of somatic genome alterations in individual tumors through tumor-specific causal inference. PLoS computational biology. 2019;15(7):e1007088. Epub 2019/07/06. doi: 10.1371/journal.pcbi.1007088. PubMed PMID: 31276486; PubMed Central PMCID: PMCPMC6650088.
31. Jabbari F, Visweswaran S, Cooper GF. Instance-specific Bayesian network structure learning. Proceedings of Machine Learning Research. 2018;72:169-80.
32. Johnson A, Cooper GF, Visweswaran S. Patient-Specific Modeling with Personalized Decision Paths. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2020;2020:602-11. Epub 2021/05/04. PubMed PMID: 33936434; PubMed Central PMCID: PMCPMC8075540.
33. Paisley J, Wang C, Blei D, Jordan. M. Nested hierarchical Dirichlet processes. IEEE transactions on pattern analysis and machine intelligence. 2015;37(2):256-70.
34. Szabo PA, Levitin HM, Miron M, Snyder ME, Senda T, Yuan J, et al. Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. Nat Commun. 2019;10(1):4706. Epub 2019/10/19. doi: 10.1038/s41467-019-12464-3. PubMed PMID: 31624246; PubMed Central PMCID: PMCPMC6797728.
35. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature biotechnology. 2014;32(4):381-6. Epub 2014/03/25. doi: 10.1038/nbt.2859. PubMed PMID: 24658644; PubMed Central PMCID: PMCPMC4122333.
36. TCGA Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015;517(7536):576-82. Epub 2015/01/30. doi: 10.1038/nature14129. PubMed PMID: 25631445; PubMed Central PMCID: PMCPMC4311405.
37. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7. Epub 2013/01/18. doi: 10.1186/1471-2105-14-7. PubMed PMID: 23323831; PubMed Central PMCID: PMCPMC3618321.
38. Ogarrio J, Spirtes P, Ramsey J. A hybrid causal search algorithm for latent variable models. JMLR Workshop and Conference Proceedings. 2016;52:368-79.
39. Zhang J. Causal reasoning with ancestral graphs. Journal of Machine Learning Research. 2008;9:1437-74.

40. Glymour C, Zhang K, Spirtes P. Review of Causal Discovery Methods Based on Graphical Models. Front Genet. 2019;10:524. Epub 2019/06/20. doi: 10.3389/fgene.2019.00524. PubMed PMID: 31214249; PubMed Central PMCID: PMCPMC6558187.
41. Ramsey JD, Zhang J, Spirtes P. Adjacency-Faithfulness and Conservative Causal Inference. 22nd Uncertain in AI; Cambridge, MA2006.
42. Leventhal DS, Gilmore DC, Berger JM, Nishi S, Lee V, Malchow S, et al. Dendritic Cells Coordinate the Development and Homeostasis of Organ-Specific Regulatory T Cells. Immunity. 2016;44(4):847-59. Epub 2016/04/03. doi: 10.1016/j.immuni.2016.01.025. PubMed PMID: 27037189; PubMed Central PMCID: PMCPMC4842258.
43. Guo X, Zhang Y, Zheng L, Zheng C, Song J, Zhang Q, et al. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. Nat Med. 2018;24(7):978-85. Epub 2018/06/27. doi: 10.1038/s41591-018-0045-3. PubMed PMID: 29942094.
44. Yost KE, Satpathy AT, Wells DK, Qi Y, Wang C, Kageyama R, et al. Clonal replacement of tumor-specific T cells following PD-1 blockade. Nat Med. 2019;25(8):1251-9. Epub 2019/07/31. doi: 10.1038/s41591-019-0522-3. PubMed PMID: 31359002; PubMed Central PMCID: PMCPMC6689255.
45. Morf J, Wingett SW, Farabella I, Cairns J, Furlan-Magaril M, Jimenez-Garcia LF, et al. RNA proximity sequencing reveals the spatial organization of the transcriptome in the nucleus. Nature biotechnology. 2019;37(7):793-802. Epub 2019/07/04. doi: 10.1038/s41587-019-0166-3. PubMed PMID: 31267103.
46. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science. 2019;363(6434):1463-7. Epub 2019/03/30. doi: 10.1126/science.aaw1219. PubMed PMID: 30923225; PubMed Central PMCID: PMCPMC6927209.
47. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome biology. 2018;19(1):15.
48. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426. 2018.
49. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IW, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nature biotechnology. 2019;37(1):38.
50. Walter V, Yin X, Wilkerson MD, Cabanski CR, Zhao N, Du Y, et al. Molecular subtypes in head and neck cancer exhibit distinct patterns of chromosomal gain and loss of canonical cancer genes. PloS one. 2013;8(2):e56823.
51. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010;26(12):1572-3.
52. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine learning. 2003;52(1-2):91-118.
53. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33(1):1-22. Epub 2010/09/03. PubMed PMID: 20808728; PubMed Central PMCID: PMCPMC2929880.
54. Spirtes P, Meek C, Richardson T, editors. Causal inference in the presence of latent variables and selection bias. The Conference on Uncertainty in Artificial Intelligence; 1995.

55. Pearl J. Probabilistic Reasoning in Intelligent Systems. Brachman RJ, editor. San Mateo, CA: Morgan Kaufmann; 1988.
56. Neapolitan R. Probabilistic Reasoning in Expert Systems. New York: John Wiley & Sons; 1990.
57. Castillo E, Gutierrez JM, Hadi AS. Expert Systems and Probabilistic Network Models. New York: Spring-Verlag; 1997.
58. Jensen FV. An Introduction to Bayesian Networks. New York: Springer Verlag; 1997.
59. Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search. Cambridge, MA: MIT Press; 2000.
60. Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques. Cambridge, MA: MIT Press; 2009.
61. Darwiche A. Modeling and Reasoning with Bayesian Networks. Cambridge, UK: Cambridge University Press; 2009.
62. Pearl J. Fusion, propagation and structuring in belief networks. Artificial Intelligence. 1986;29(3):241–88.
63. Neapolitan RE. Learning Bayesian Networks: Prentice Hall; 2003.
64. Jensen FV, Niesen TD. Bayesian Networks and Decision Graphs (second edition). New York: Springer; 2007.
65. Heckerman D, Meek C, Cooper GF. A Bayesian approach to causal discovery. In: Sprites P, Glymour C, Scheines R, editors. Causation, Prediction, and Search. Cambridge, MA: MIT Press; 2000.
66. Glymour C, Cooper GF, editors. Computation, Causation, and Discovery. Cambridge, MA: MIT Press; 1999.
67. Pearl J. Causality: Models, Reasoning, and Inference. Cambridge, U.K.: Cambridge University Press; 2009.
68. Richardson T, Spirtes P. Ancestral graph Markov models. The Annals of Statistics. 2002;30(4):962-1030.
69. Kitson NK, Constantinou AC, Guo Z, Liu Y, Chobtham K. A survey of Bayesian Network structure learning. arXiv preprint arXiv:210911415. 2021.
70. Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. Frontiers in Genetics. 2019;10:524.
71. CCD. Greedy Fast Causal Inference (GFCI) Algorithm for Discrete Variables. Available at: https://www.ccd.pitt.edu/pdfs/GFCId.pdf. 2019.
72. Millstein J, Zhang B, Zhu J, Schadt EE. Disentangling molecular relationships with a causal inference test. BMC Genetics. 2009;10(1):1-15.
73. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, GuhaThakurta D, et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nature Genetics. 2005;37(7):710-7.
74. Ogarrio JM, Spirtes P, Ramsey J, editors. A hybrid causal search algorithm for latent variable models. Conference on Probabilistic Graphical Models; 2016: PMLR.
75. Ramsey J, Glymour M, Sanchez-Romero R, Glymour C. A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. International Journal of Data Science and Analytics. 2017;3(2):121-9.

76. Chickering DM. Optimal structure identification with greedy search. Journal of Machine Learning Research. 2002;3(Nov):507-54.
77. Efron B, Tibshirani RJ. An Introduction to the Bootstrap: CRC Press; 1994.
78. Tetrad. Tetrad code repository 2022. Available from: https://cloud.ccd.pitt.edu/nexus/content/repositories/releases/edu/cmu/tetrad-gui/.
79. Tetrad. Tetrad manual 2022. Available from: http://cmu-phil.github.io/tetrad/manual/.
80. Jabbari F, Cooper GF. An instance-specific algorithm for learning the structure of causal Bayesian networks containing latent variables. Proceedings of the 2020 SIAM International Conference on Data Mining: SIAM; 2020. p. 433-41.
81. Jabbari F, Visweswaran S, Cooper GF. Instance-specific Bayesian network structure learning. International Conference on Probabilistic Graphical Models: PMLR; 2018. p. 169-80.
82. Boutilier C, Friedman N, Goldszmidt M, Koller D. Context-specific independence in Bayesian networks. Proceedings of the Conference on Uncertainty in Artificial Intelligence: UAI; 1996. p. 115-23.

Tables

From \ To	ENDO	EPI	FIB	IMM
ENDO	25	63	24	38
EPI	57	96	35	63
FIB	29	34	20	26
IMM	28	41	9	90

Table 1. Statistics of directed causal edges between different cell types. Rows indicate cell types (endothelial, epithelial, fibroblast, and immune cells) from which edges emit and columns represent cell types to which edges point. A number in the table represents the count of edges between two types of GEMs. (summary of edges in Figure S5)

Figures and legends

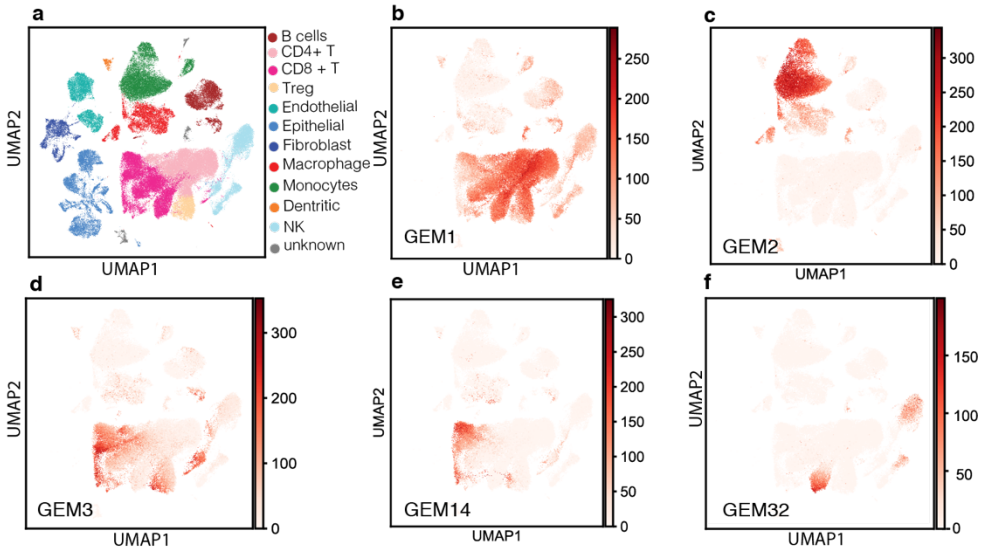


Figure 1. a. UMAP plot of all cells collected from 17 HNSCC tumors. Cells were represented in the original transcriptome and were projected to 2-D space using the UMAP method. Pseudocolors illustrate the cell types. **b – f. Distribution of example immune-cell GEMs among CD45^{+} cells**. The sum of the probabilities of the genes expressed in a cell being a member of an indicated GEM is represented as pseudocolors.

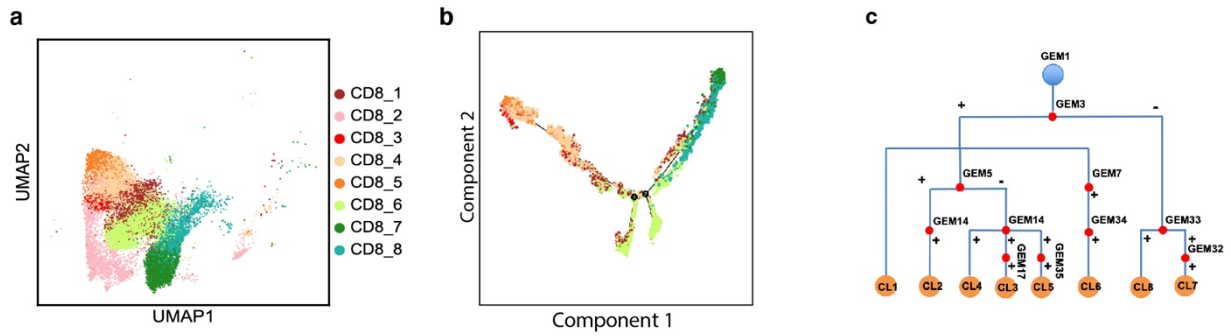


Figure 2. CD8$^+$ T cell subtyping based on GEM compositions. a. UMAP plot of CD8$^+$ T cells. Eight subtypes were identified, which are indicated by pseudo-colors. **b. Pseudo-time plot reflecting evolution of subtypes of CD8$^+$ T cells** (cell subtype indicated with the same pseudo color scheme as in **a**). **c. A hierarchical tree reflecting GEM composition of cell subtypes.**

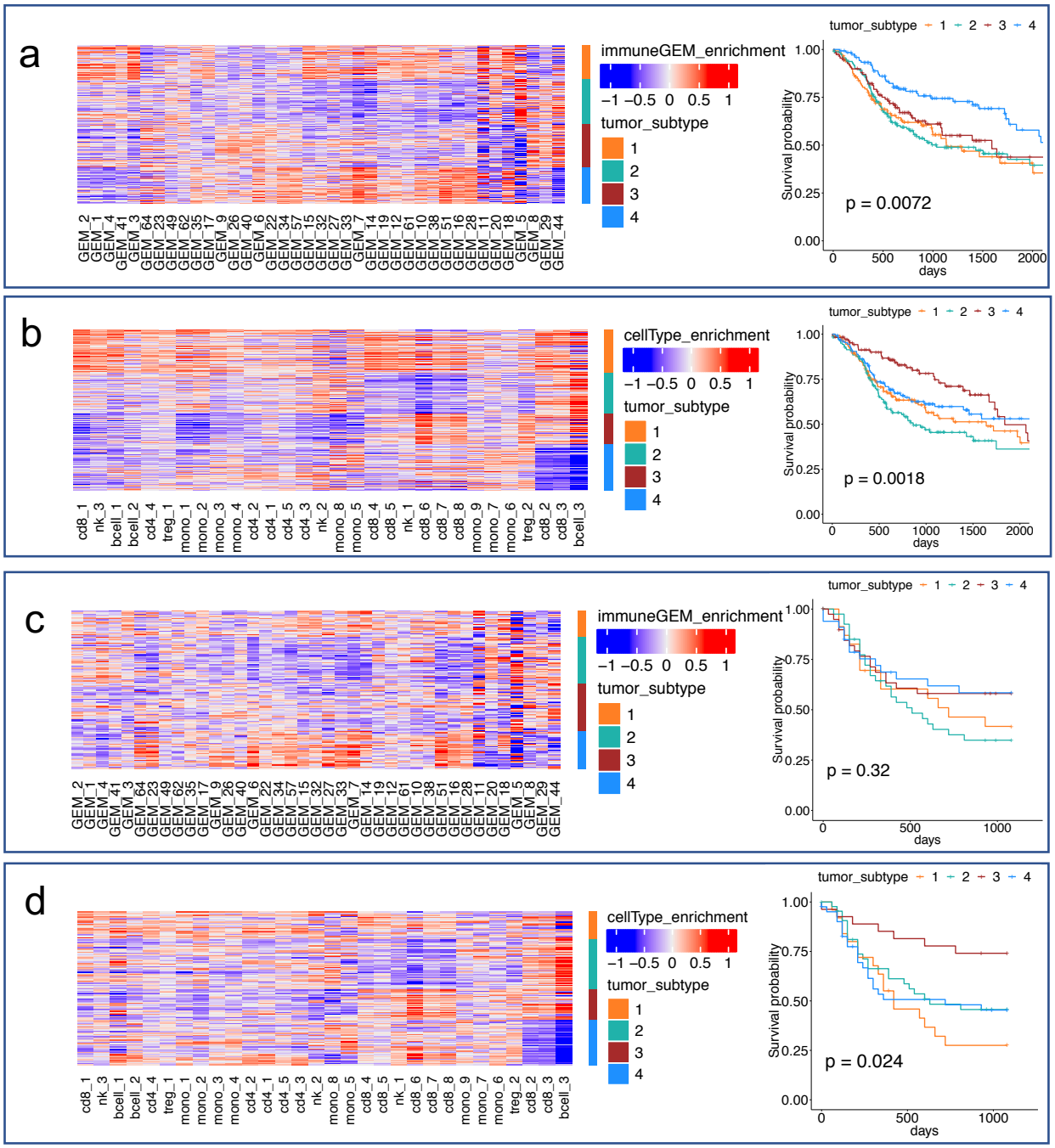


Figure 3. Distinct immune environment of HNSCCs defined by GEMs and cell subtypes. a.

Heatmap reflecting relative expression levels of GEMs (columns) among TCGA tumors (rows).

Tumors are organized according to the clustering results, with color bars that indicate the tumor subtype cluster id. Kaplan-Meier curves of patients belonging to different clusters are shown on the right, using the same color scheme ids. **b.** Heatmap reflecting relative enrichment of immune cell subtypes among TCGA HNSCC tumors. Tumors are organized according to clustering results, and Kaplan-Meier curves of patients belonging to different clusters are shown on the right. **c.** Immune GEMs (with id indicated by the number) that are associated with statistically significant hazard ratios. **d.** Immune cell subtypes (indicated by major cell type, followed by subtype cluster id) that are associated with statistically significant hazard ratios.

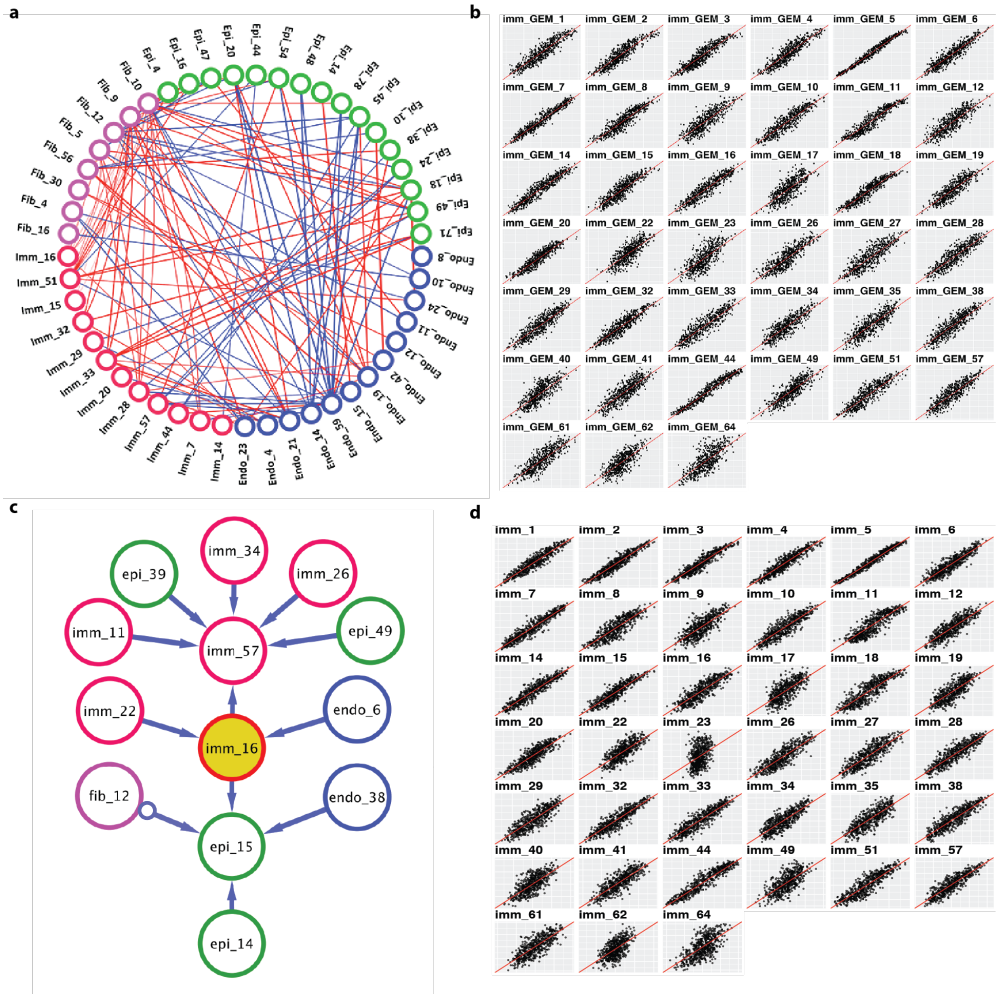


Figure 4. Intercellular communication among cells in HNSCC TMEs. a. Pair-wise correlation between GEMs from different type of cells. GEMs from major cell categories are color-coded and the top 10 strongest correlated GEMs for each cell type are shown as edges connecting the GEMs. Positive and negative correlated pairs are indicated as red and blue edges, respectively. **b. Scatter plot of predicted immune GEM enrichment values using non-immune GEMs as independent variables in a regression model vs observed GEMs values.** The average R^2 is 0.71 . **c. A subgraph of the PAG learned by GFCI algorithm.** GEMs from different cell types are color-coded. A direct edge (A→B) represents a direct causal relationship, i.e., A causes B, and blunt tailed edge (A o→ B) indicates either A causes B, or there is a latent variable causes both A and B, or both situations. **d. Scatter plot of predicted immune GEM enrichment values using the Markov boundary of a GEM as independent variables in a regression model vs observed GEM values.** The mean of R^2 is 0.67, and the standard deviation is 0.16.

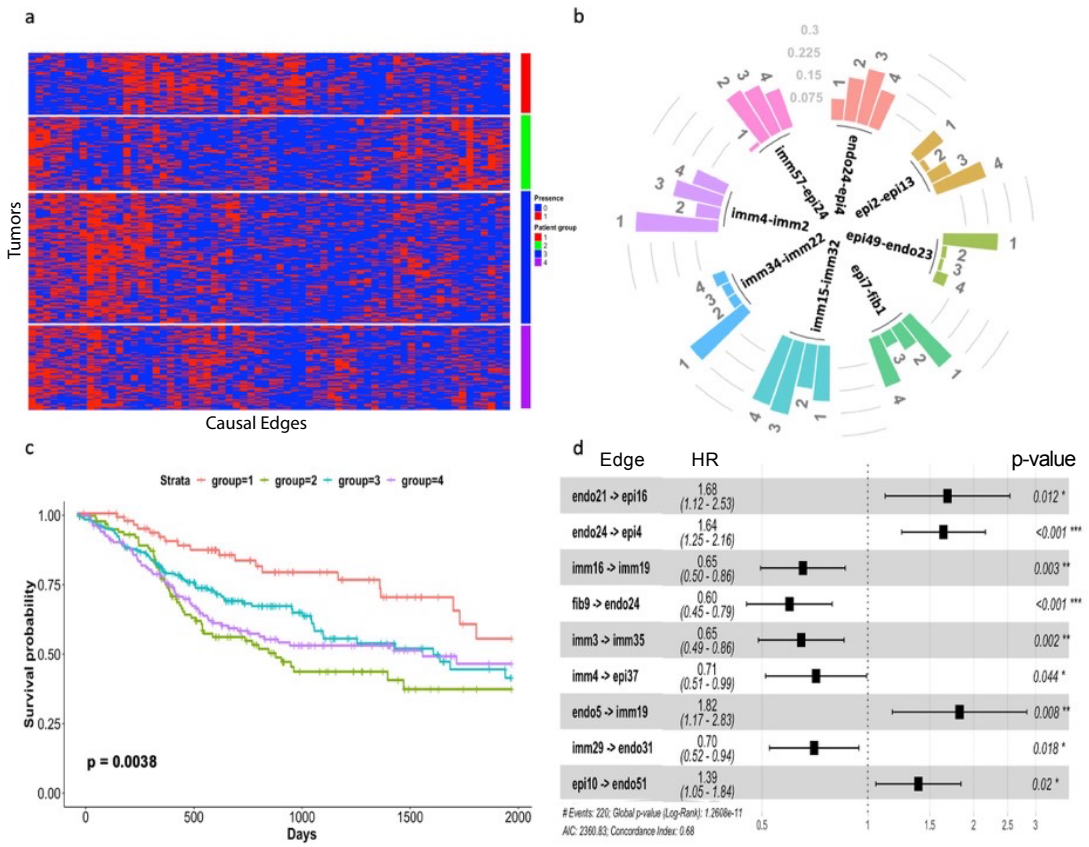


Figure 5. Distinct intercellular communication networks revealed by individualized CBN discovery. a. Clustering of tumor-specific intercellular communication networks. Tumor-specific causal networks returned by the iGFCI algorithm were represented in the space spanned by the union of edges, where the presence/absence of an edge is represented by a binary variable ("1" / "0" respectively). Tumors (rows) were divided into 4 clusters, and the presence/absence of edges (columns) is shown as a heatmap. **b. Examples of causal edges with high variance among tumor clusters.** Eight edges are shown, and the directions of the edges are indexed by the cause-effect GEMs. The height of a bar reflects the ratio (or percentage) of tumors in a cluster (indicated by a number from 1 to 4) that have the edge in their tumor-specific network. **c. Kaplan-Meier survival curves of patients assigned to different clusters. d.**

A list of the causal edges that are associated with a statistically significant proportional hazard ratios according to a Cox proportional hazard analysis.

SUPPLEMENTARY FIGURES

Figure S1

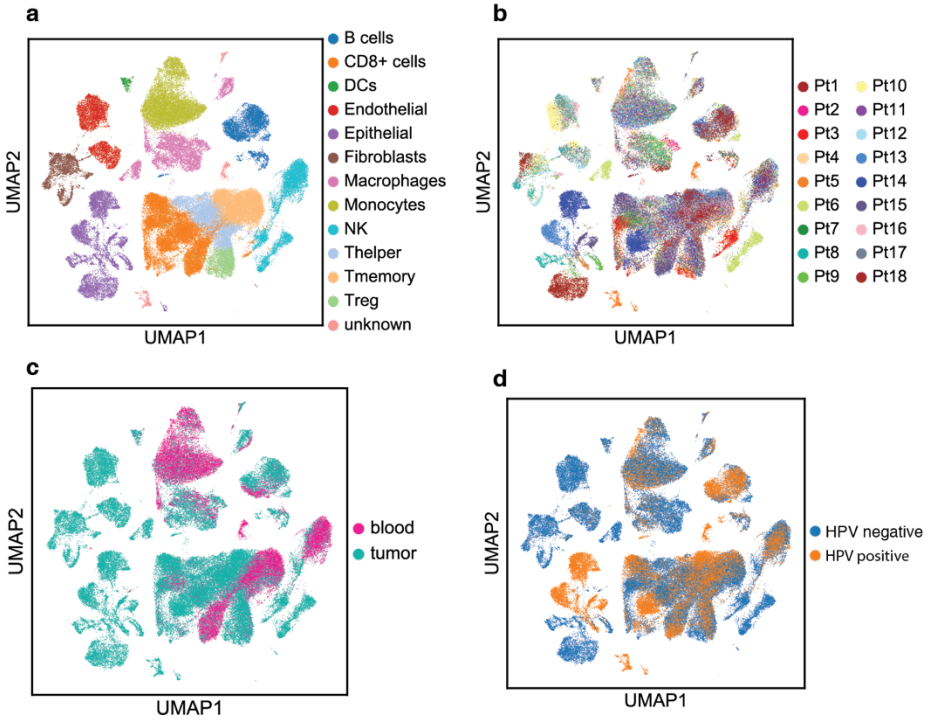


Figure S1. Distributions of cells with respect to cell type and sources. a. UMAP visualization of major types of cells. **b**. Visualization of cells from different tumors in UMAP 2-D space. C. Distribution of cells from tumors and peripheral blood. D. Distribution of cells from tumors with different human papillomavirus status.

Figure S2 is available at the end of this file due to its big size. The figure contains the members of each GEM, and the distribution of cells expressing a GEM is illustrated in UMAP format.

Figure S3

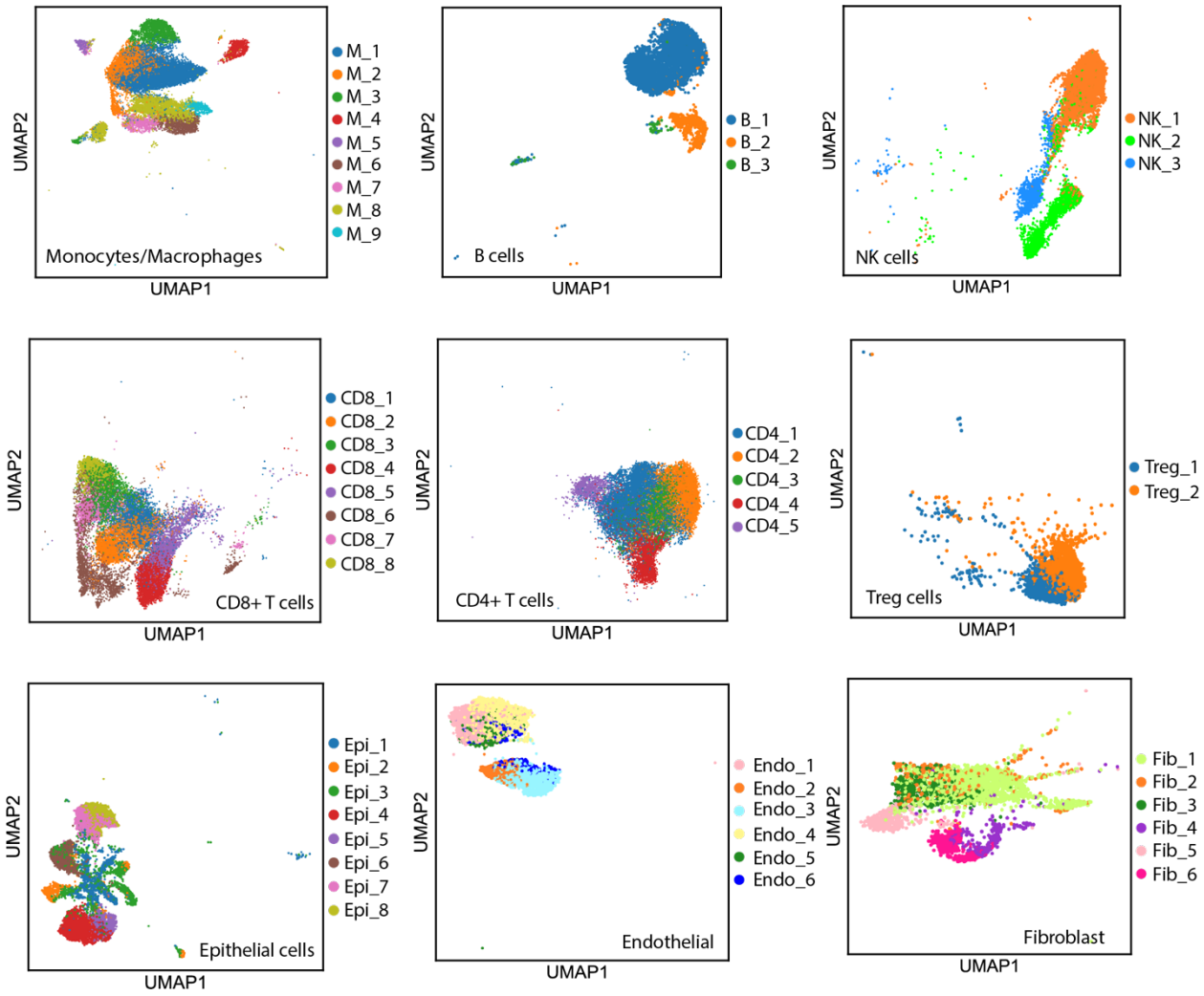


Figure S4. Cell subtypes within major cell categories. Cells were first divided into 9 major categories based on the expression of markers. Within each major category, cells were represented in the GEM space and analyzed using consensus clustering. Each block shows the subtyping within one of 9 major categories.

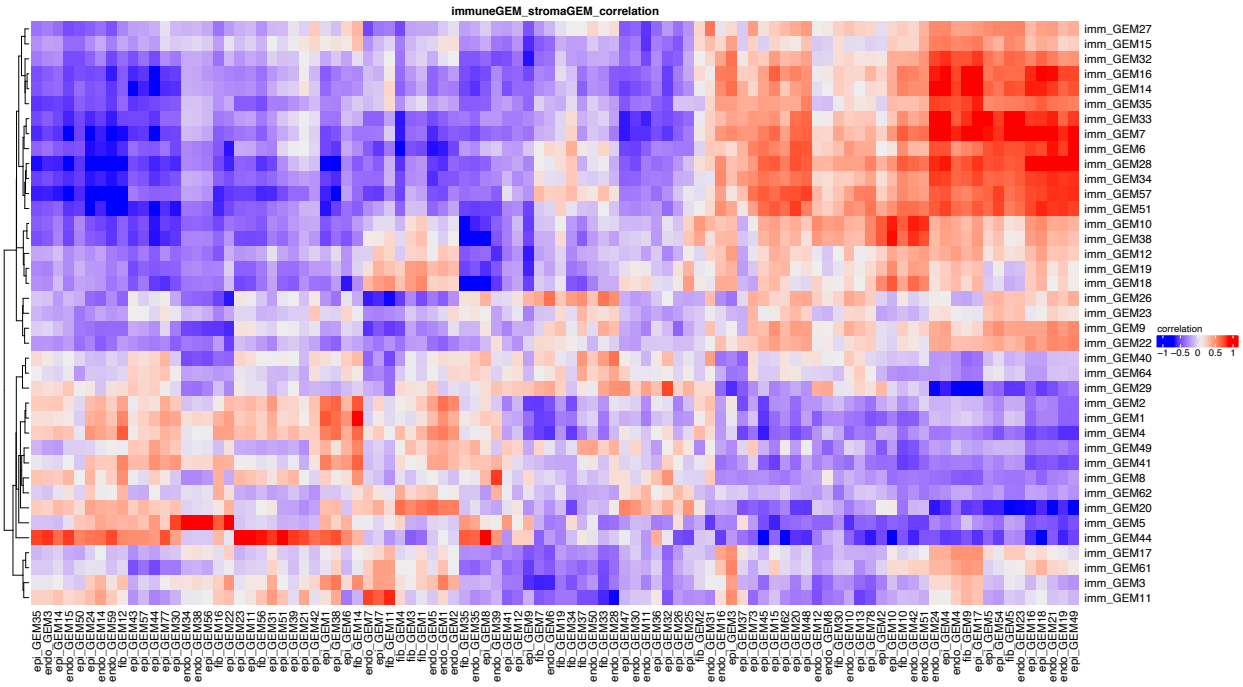
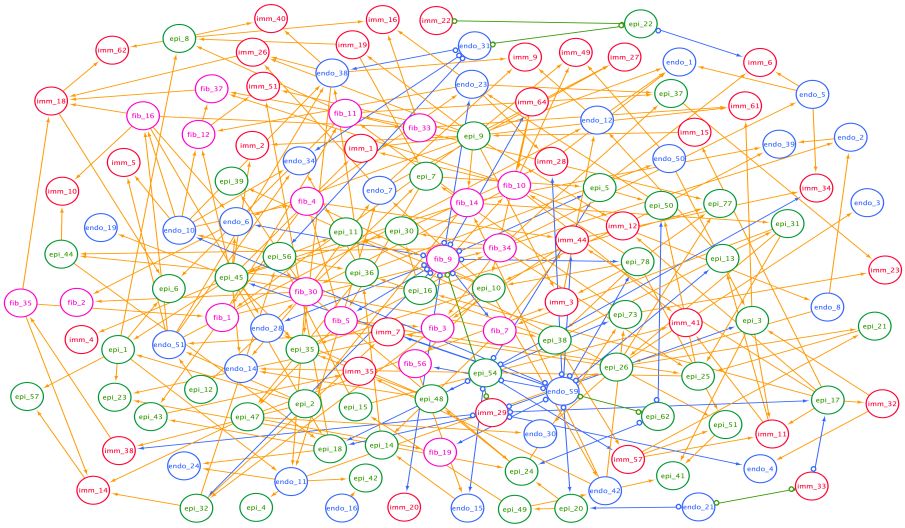


Figure S4. Heatmap representing significant correlation coefficients of pair-wise correlation analyses between non-immune GEMs and immune GEMs. Value of correlation coefficients are color-coded.



A → B Causal relationship and no latent confounder from A to B
A o→ B Either A is a cause of B or there is an unmeasured confounder of A and B or both
A o—o B Either (1) A is a cause of B or B is a cause of A, or (2) there is an unmeasured confounder of A and B, or both 1 and 2 hold

Figure S5. The cohort-wise PAG learned by GFCI algorithm.

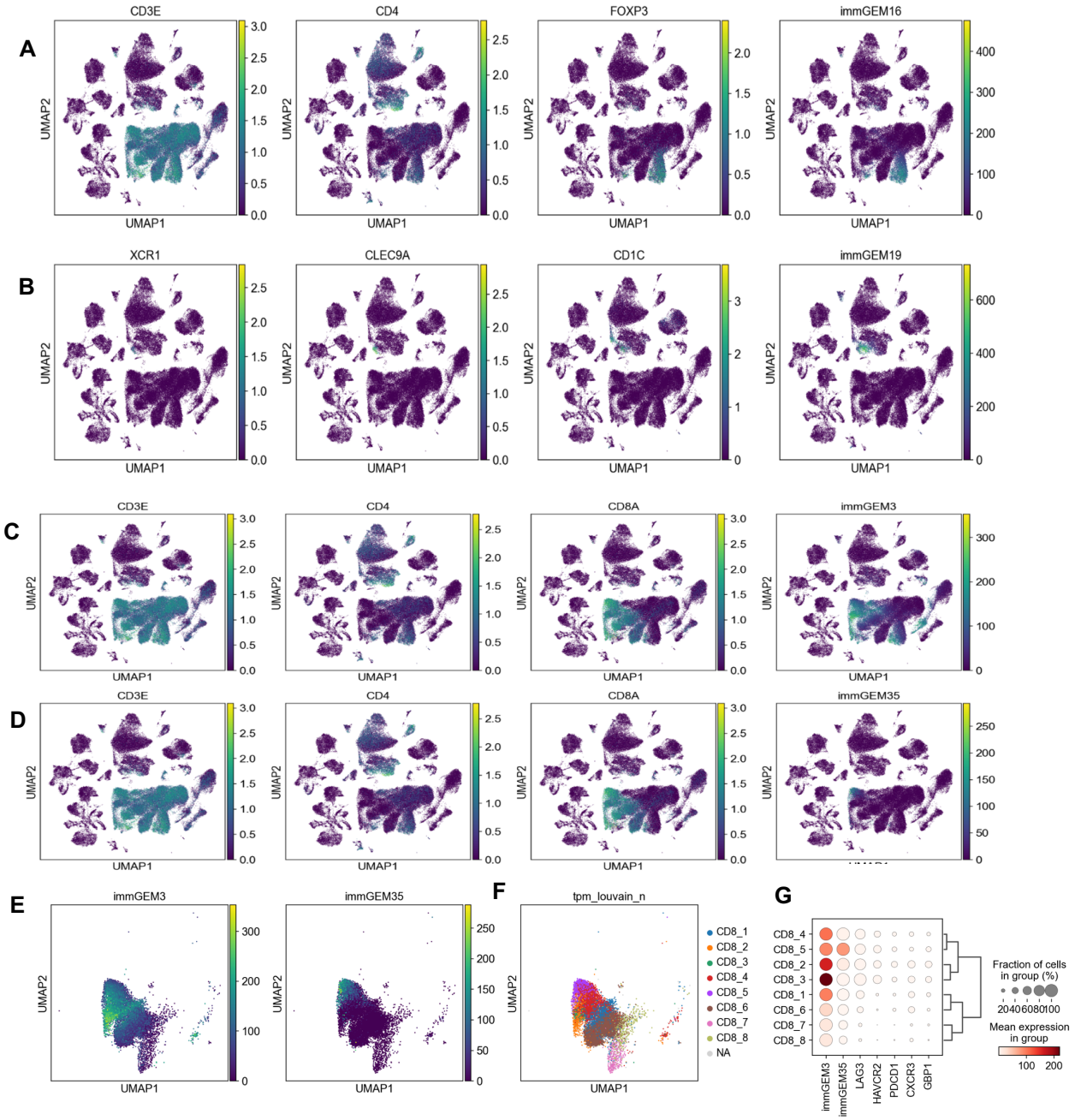


Figure S6. Cell types expressing different GEMs. **A.** UMAP projections of Treg markers (*CD3E, CD4, FOXP3*) and immGEM16. **B.** UMAP projections of dendric cell markers (*XCR1, CLEC9A, CD1C*) and immGEM19. **C** & **D**. UMAP projection of CD8 cell markers (CD3E, CD4, CD8A), immGEM3 and immGEM35 respectively. **E**. UMAP projections comparing distributions of immGEM3 and immGEM35 among CD8+ cells. **F**. CD8+ cell clusters. **G.** Dot plot show the enrichment of immGEM3 and immGEM35 among the subtypes of CD8+ cells.

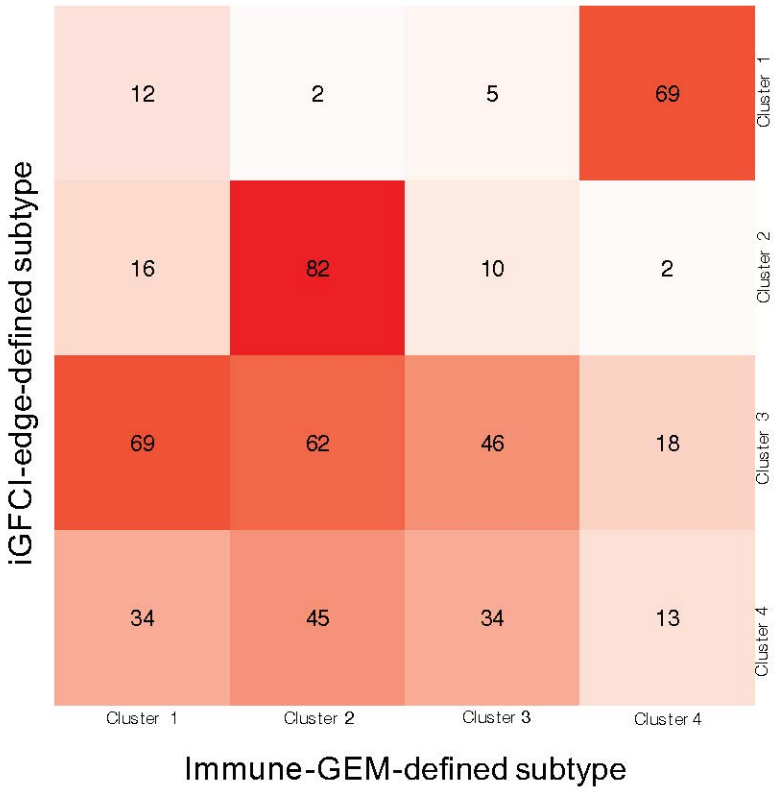
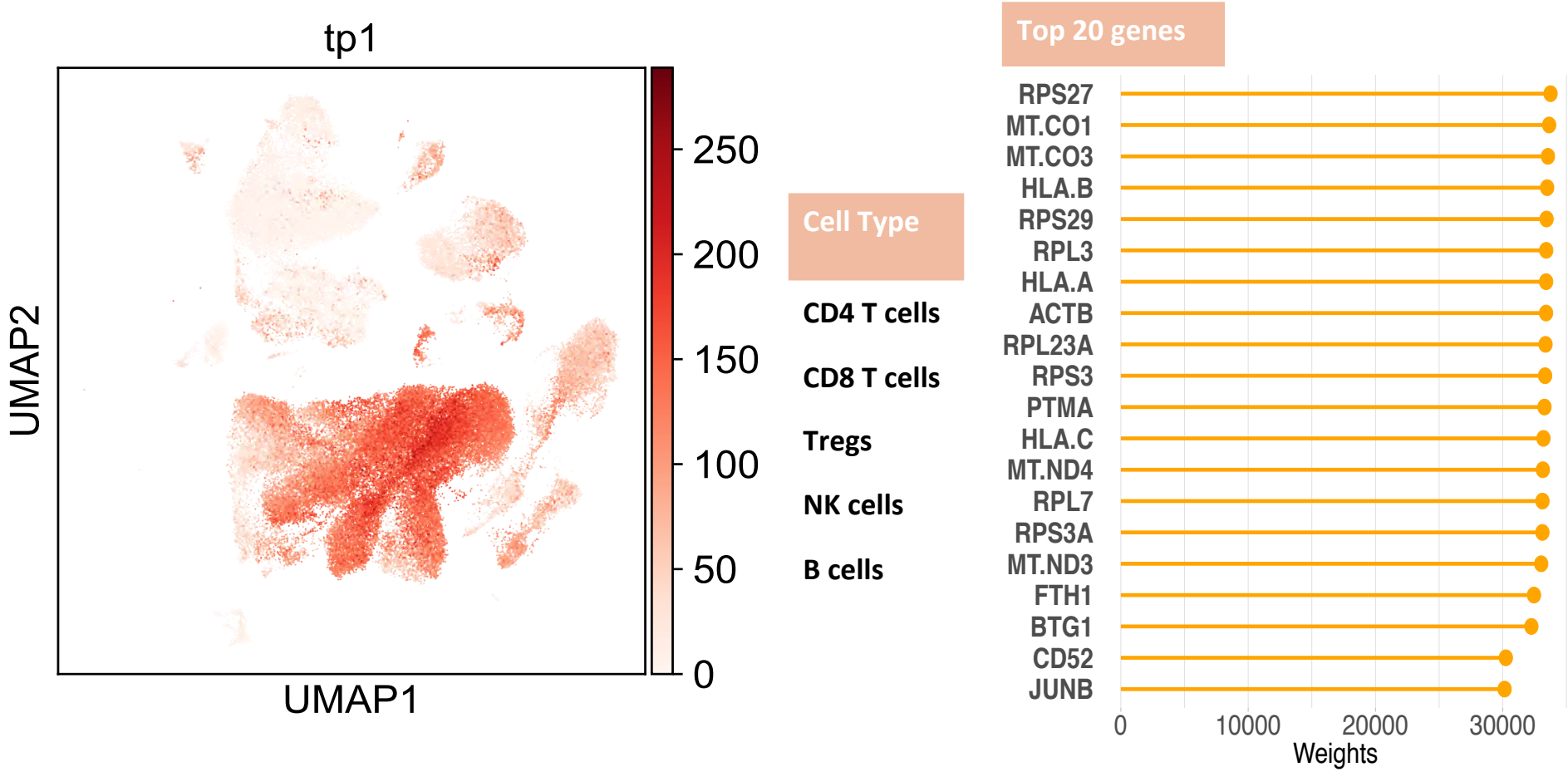


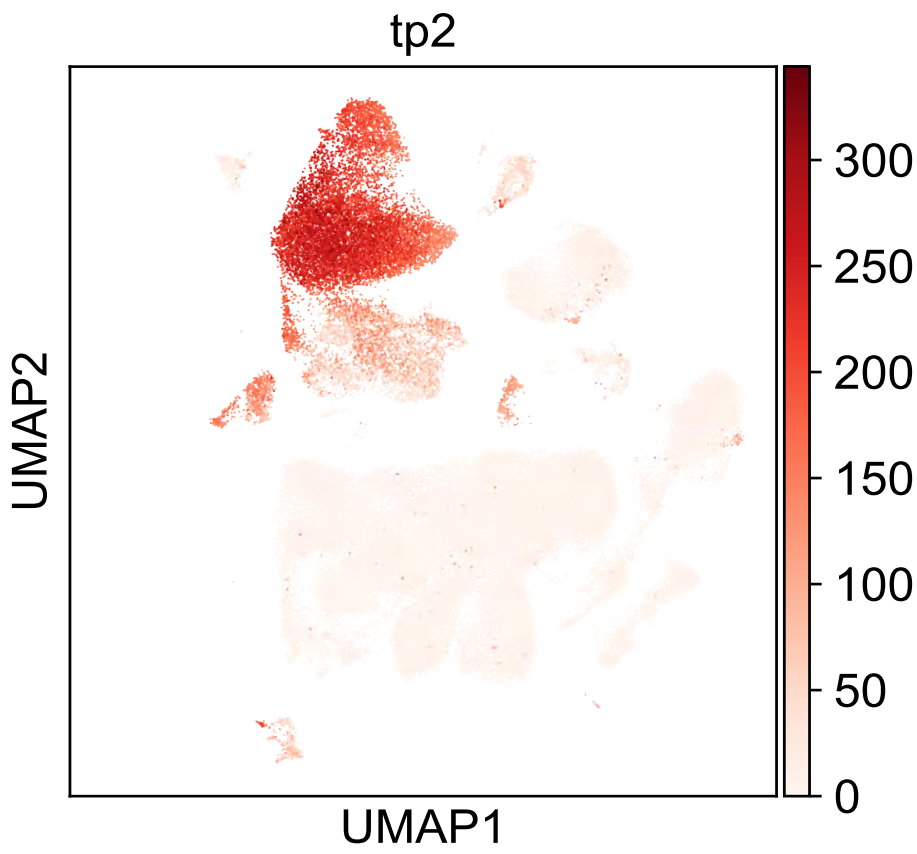
Figure S7. Heat map representing agreements of tumor cluster assignment by comparing clustering results based on GEM expression (columns) and on intercellular communication edges (rows).

Figure S2. GEMs derived from different cells



Imm GEM 1

Imm GEM 2

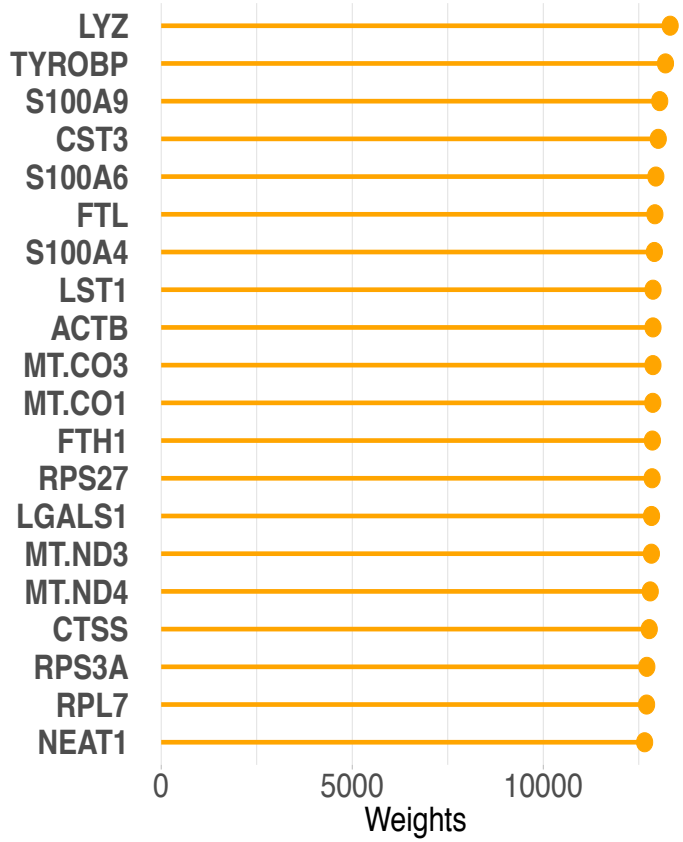


Cell Type

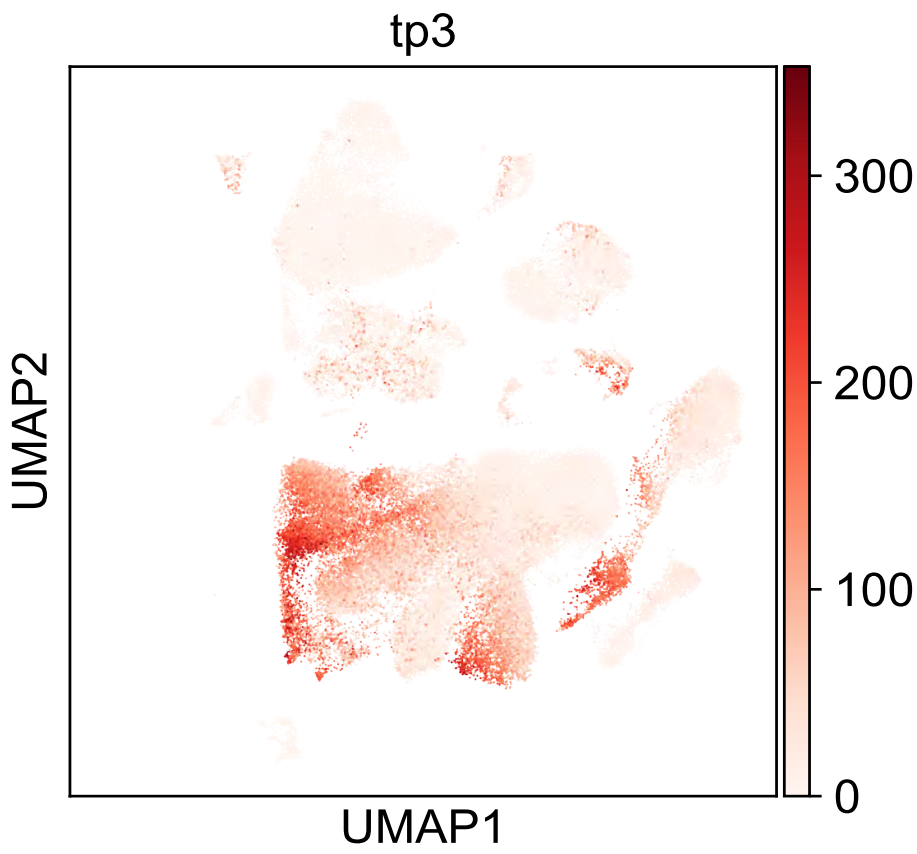
Monocytes

Macrophage

Top 20 genes



Imm GEM 3



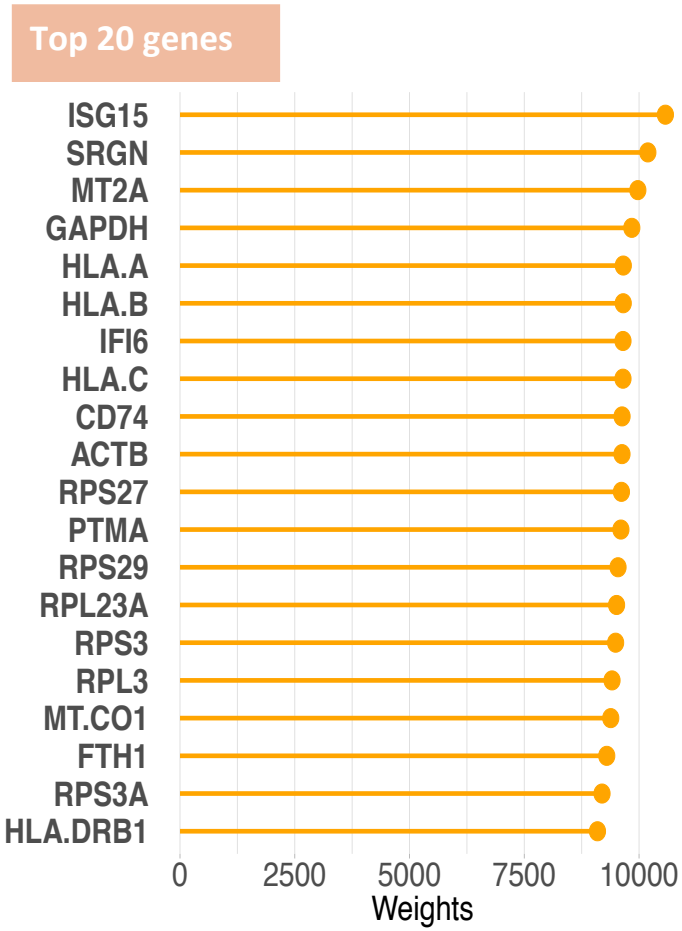
Cell Type

CD8 T cells

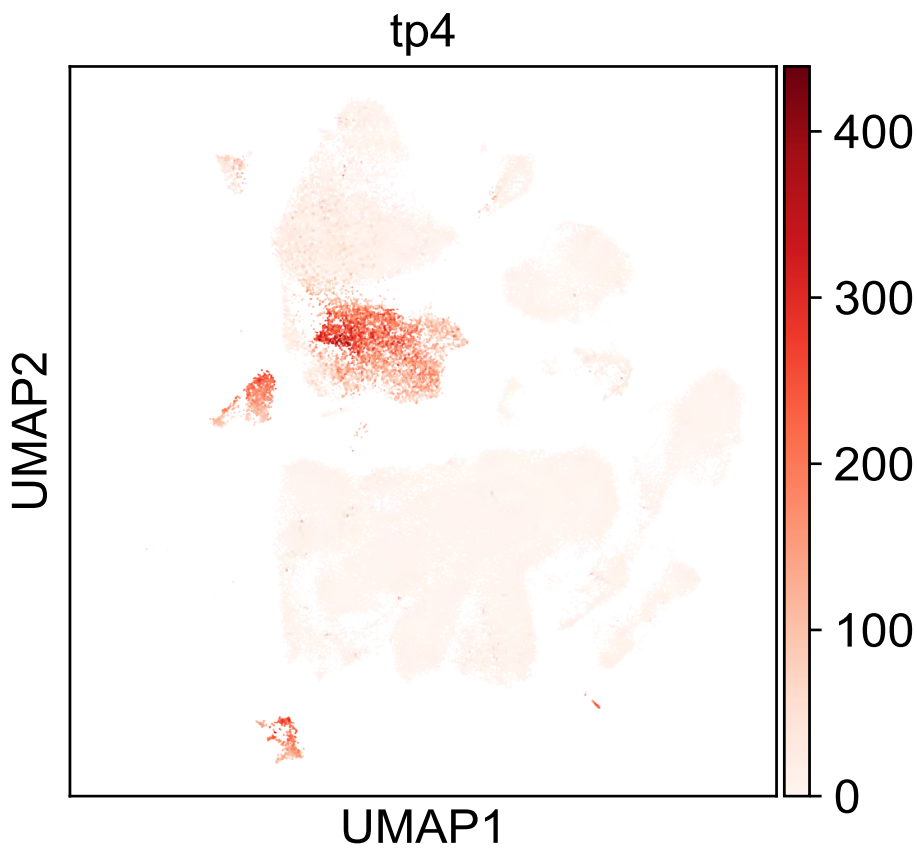
CD4 T cells

Tregs

NK cells



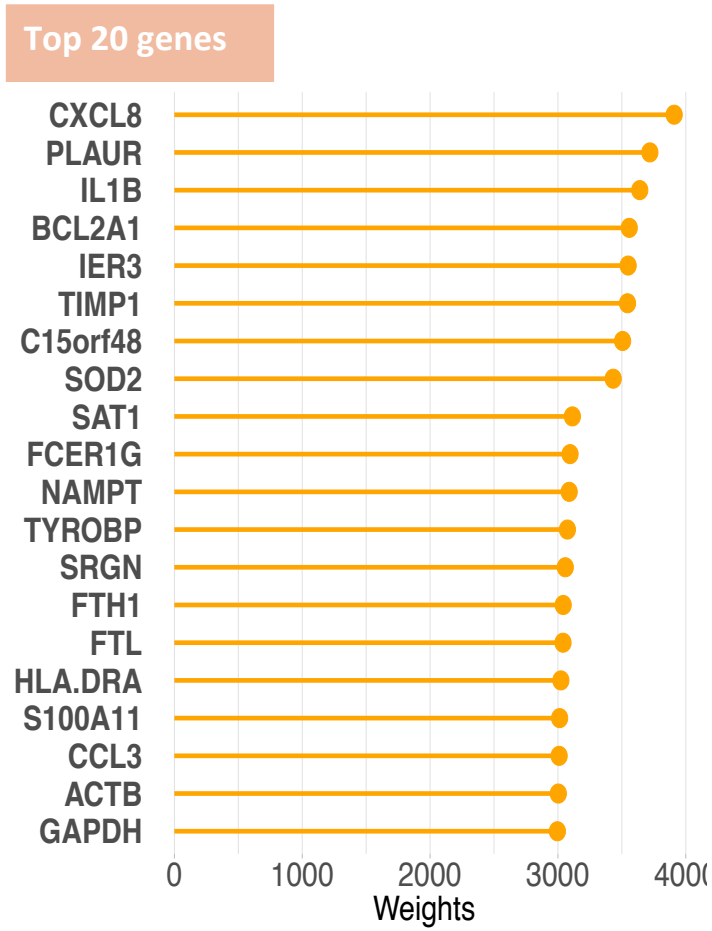
Imm GEM 4



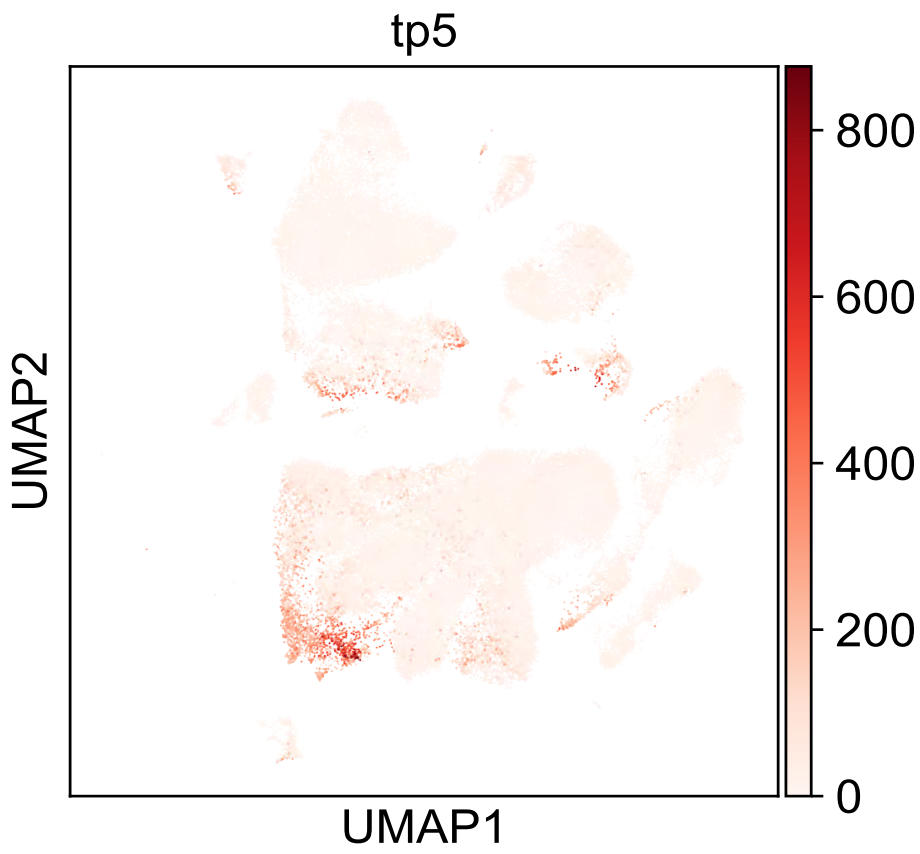
Cell Type

Macrophage

unknown

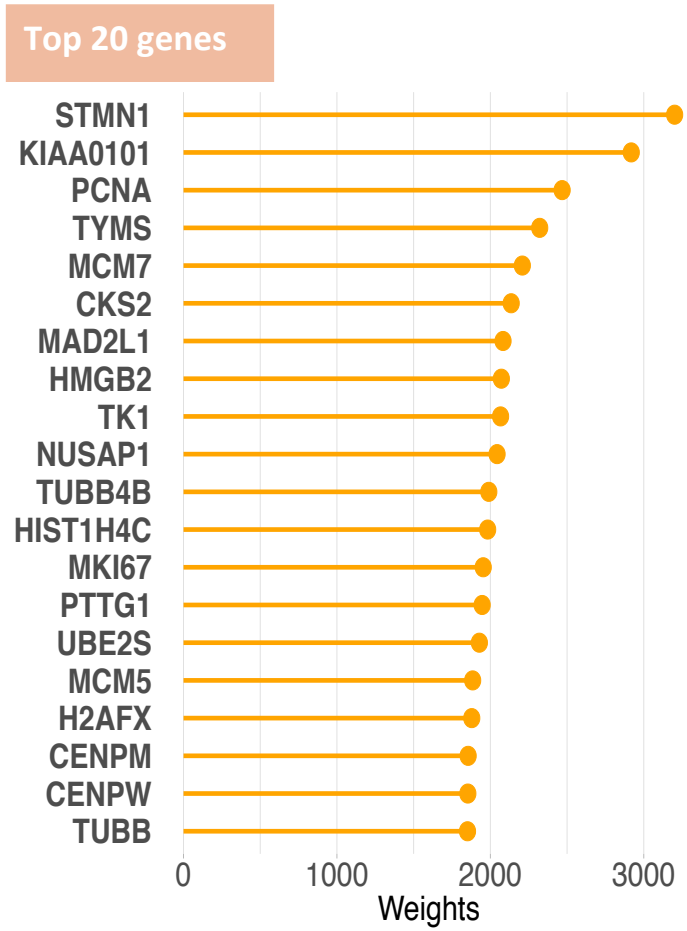


Imm GEM 5

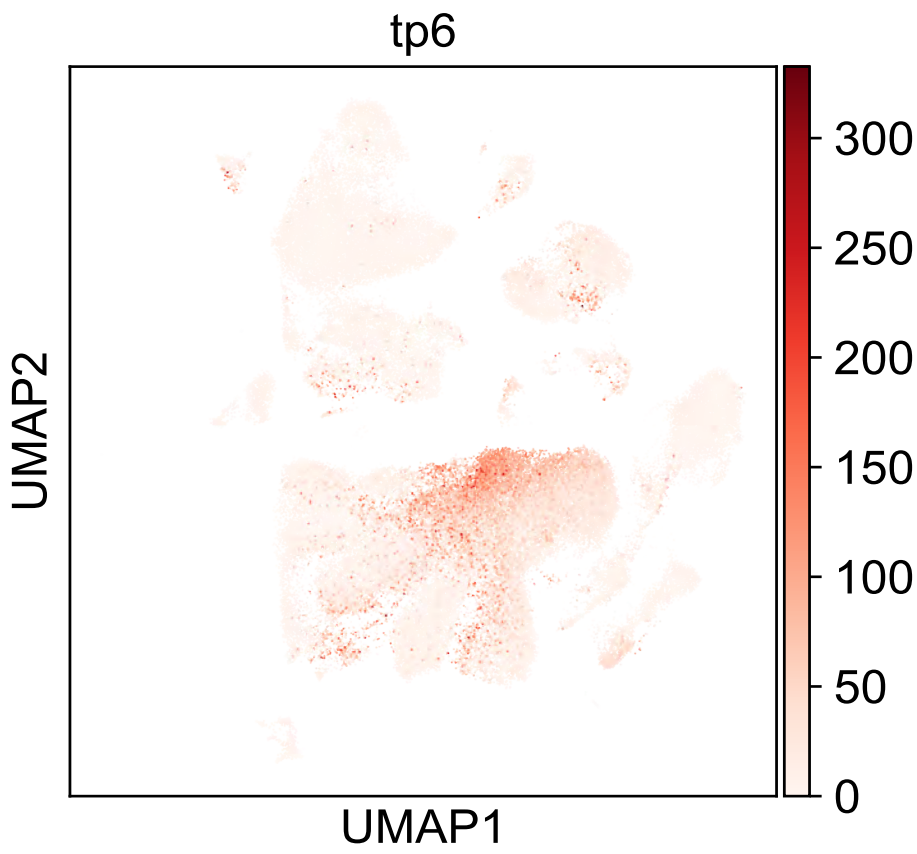


Cell Type

CD8 T cells



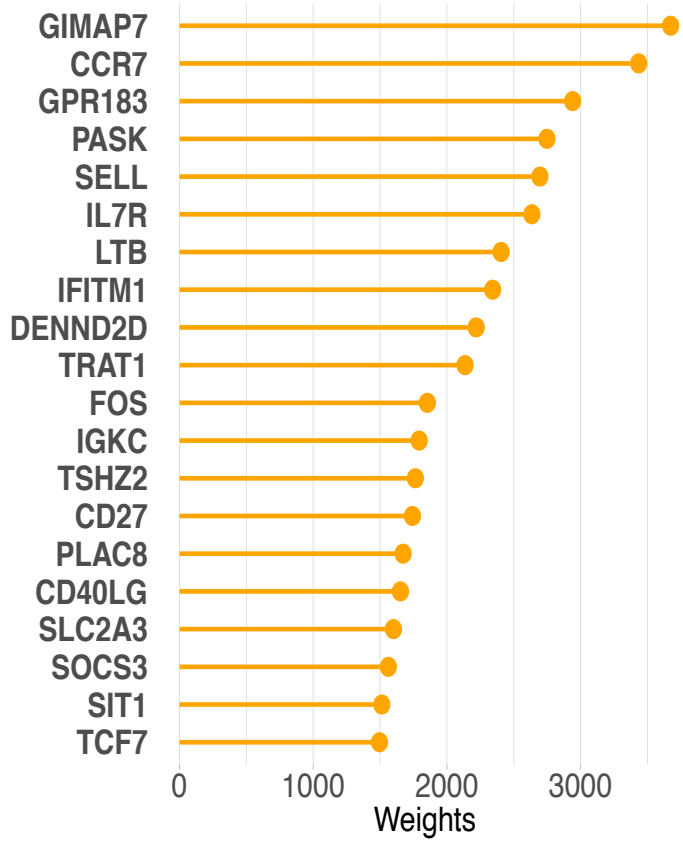
Imm GEM 6



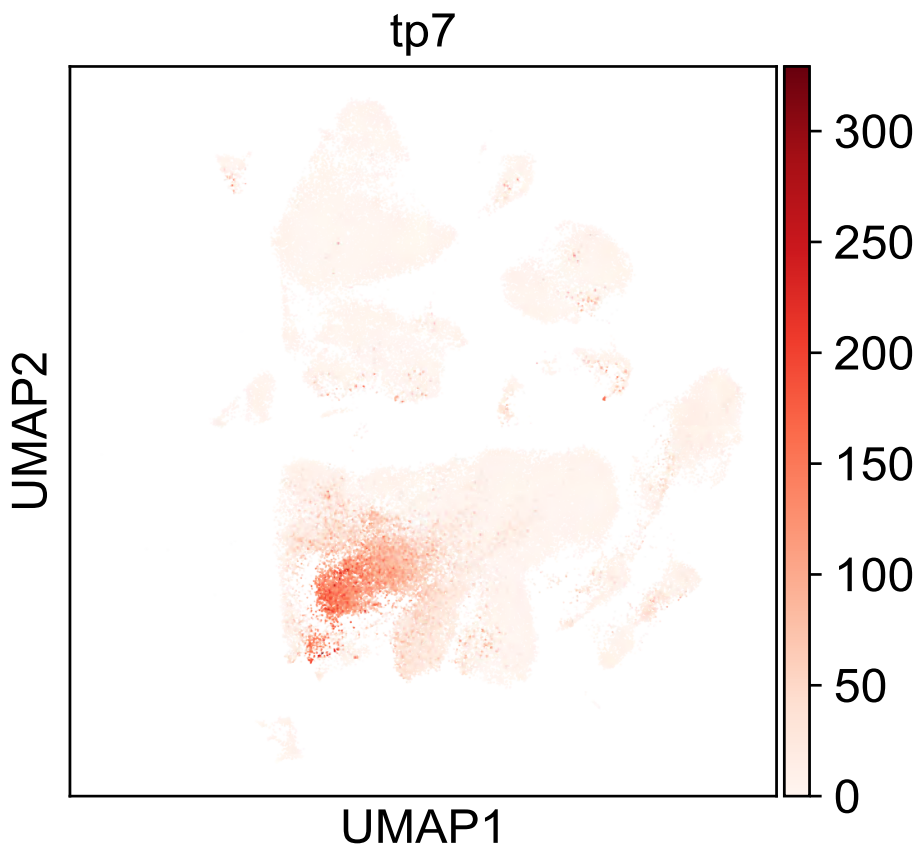
Cell Type

CD4 T cells

Top 20 genes



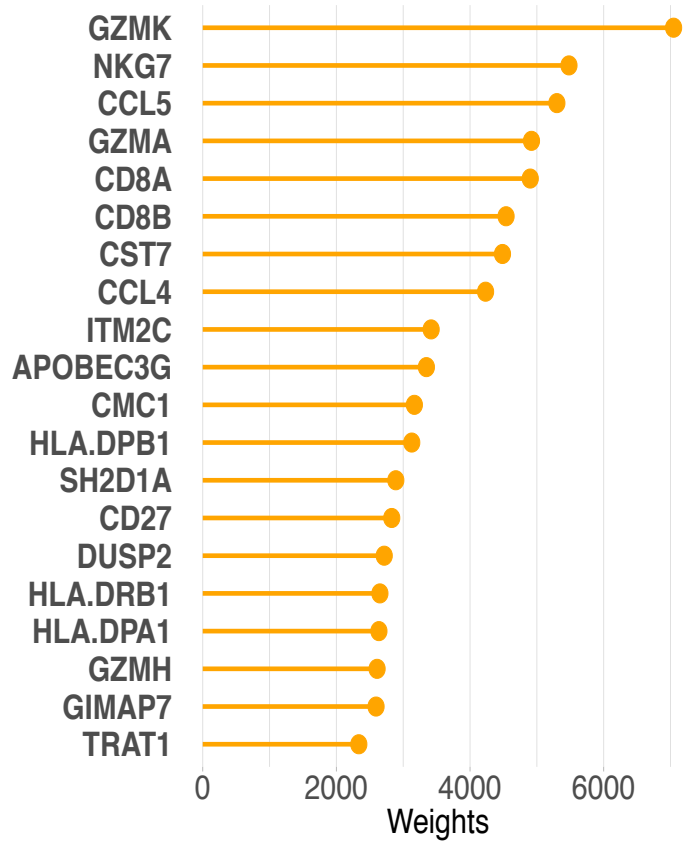
Imm GEM 7



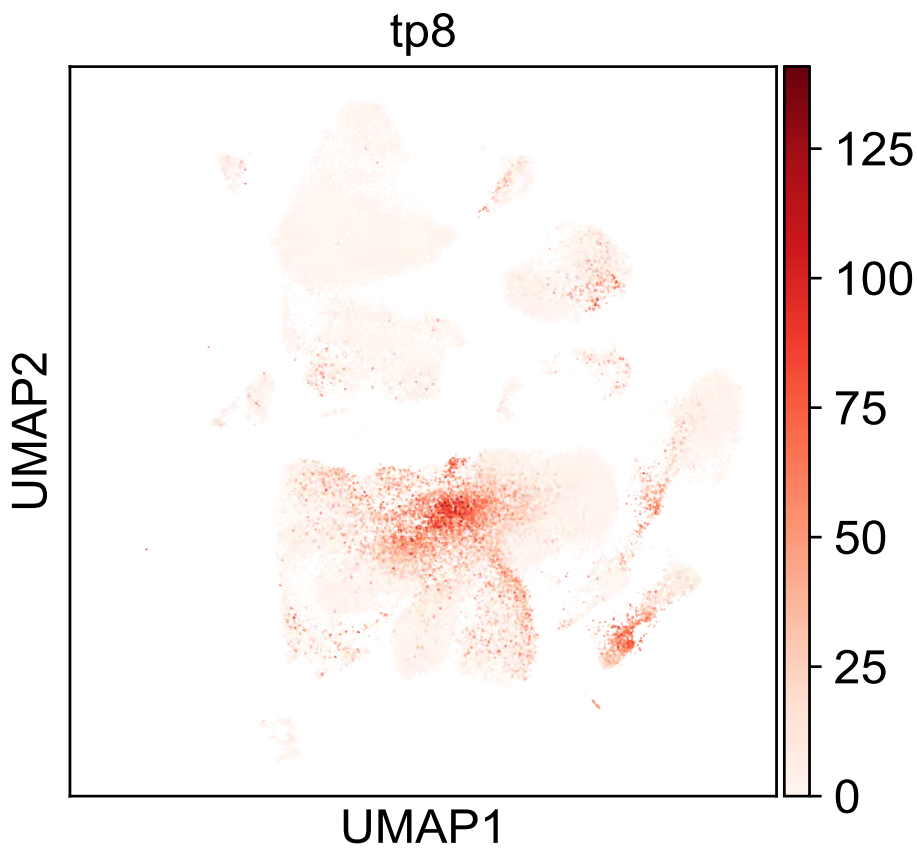
Cell Type

CD8 T cells

Top 20 genes



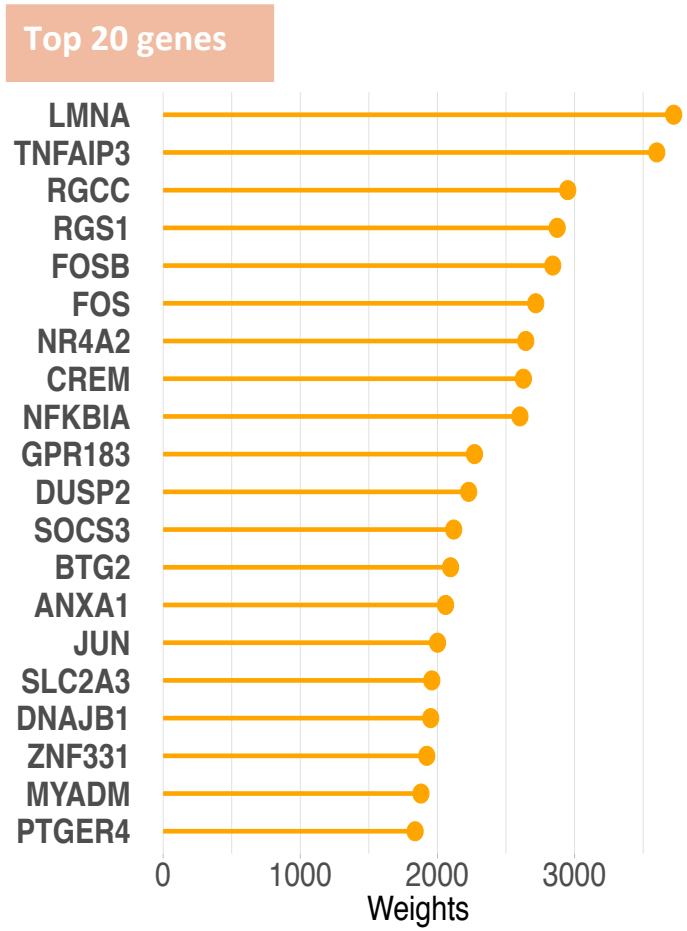
Imm GEM 8



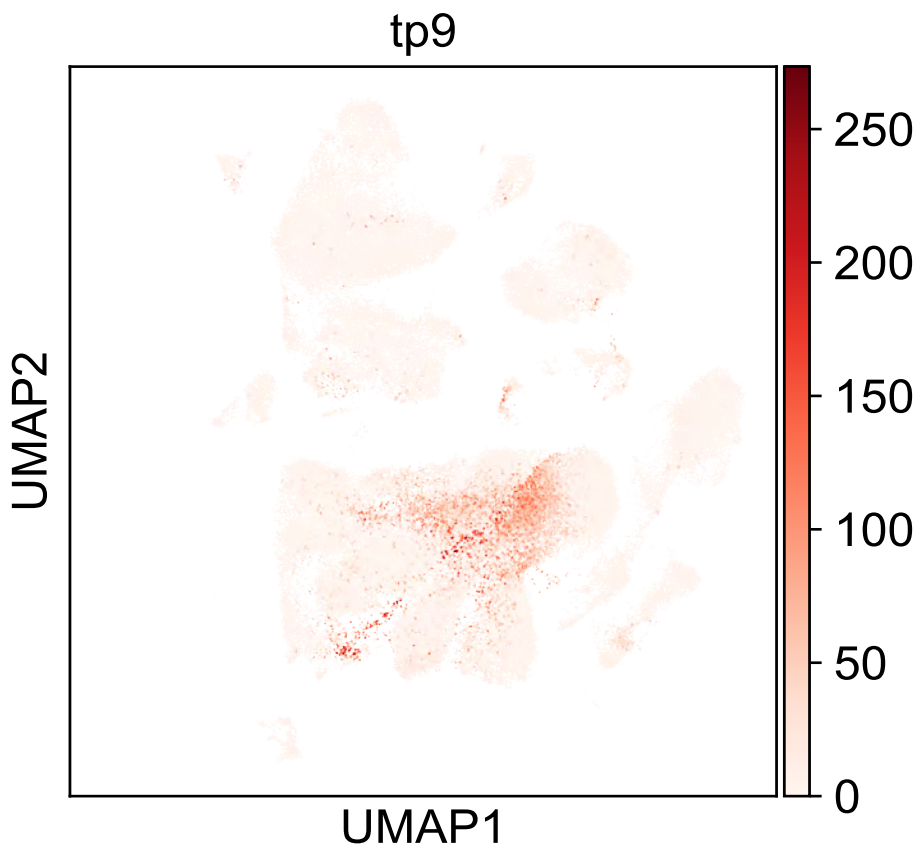
Cell Type

CD4 T cells

NK cells



Imm GEM 9

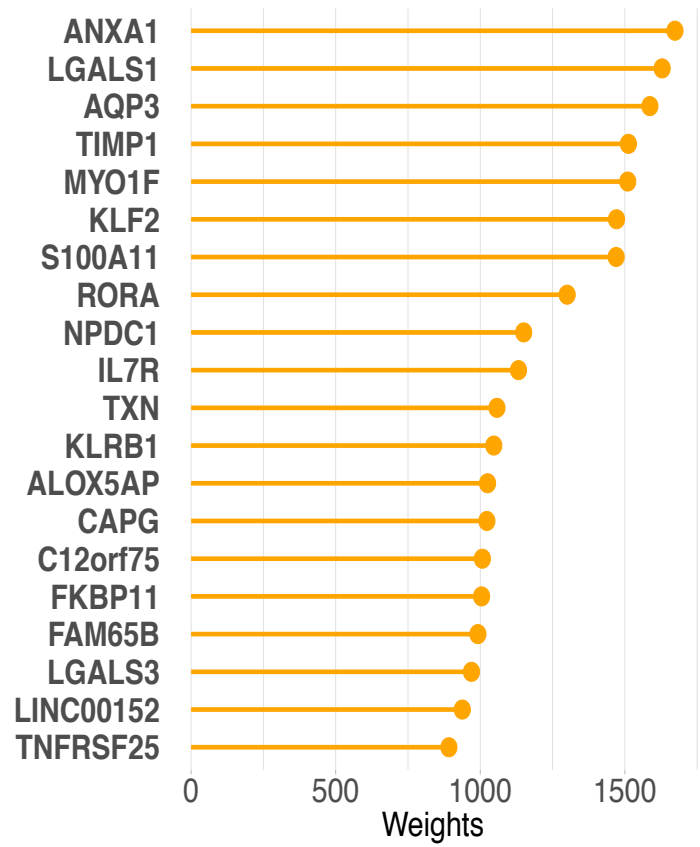


Cell Type

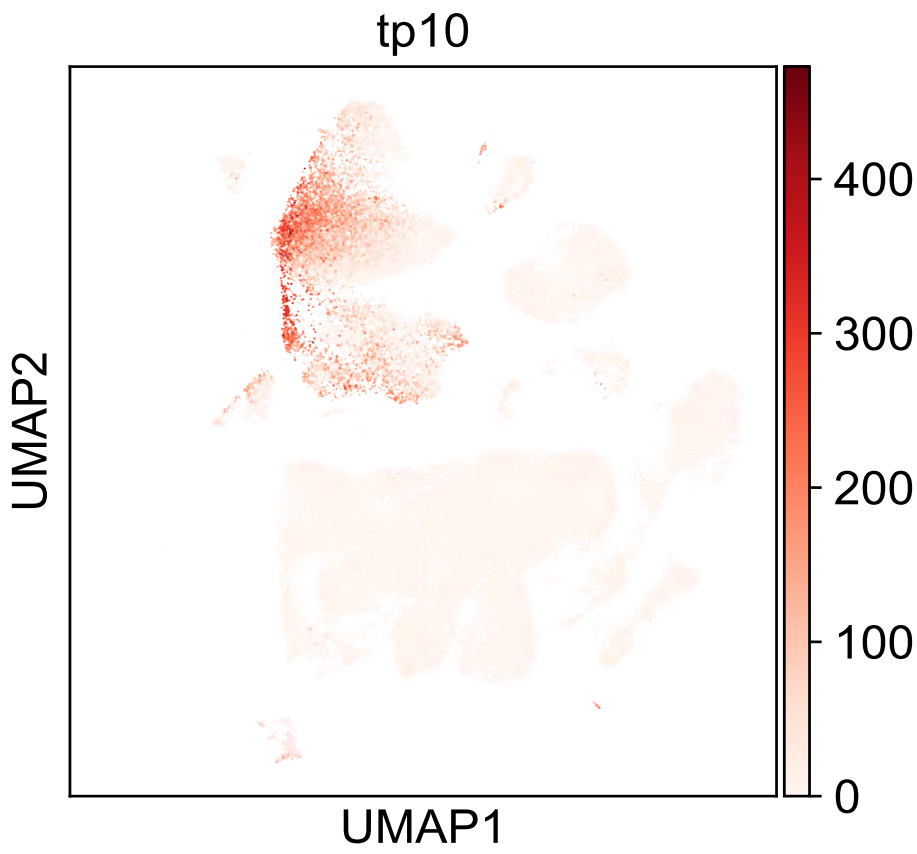
CD4 T cells

CD8 T cells

Top 20 genes



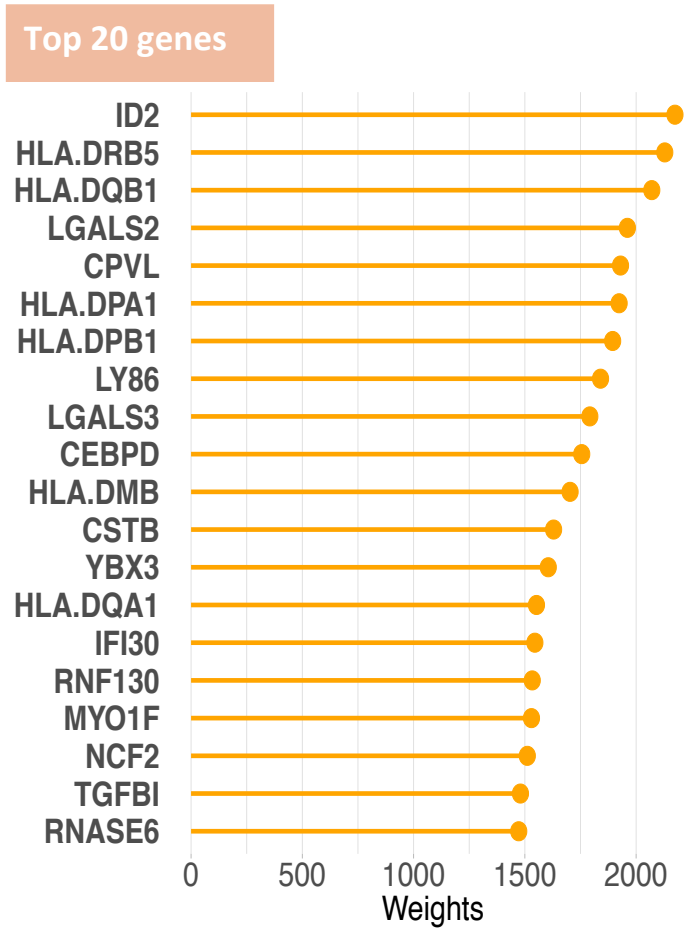
Imm GEM 10



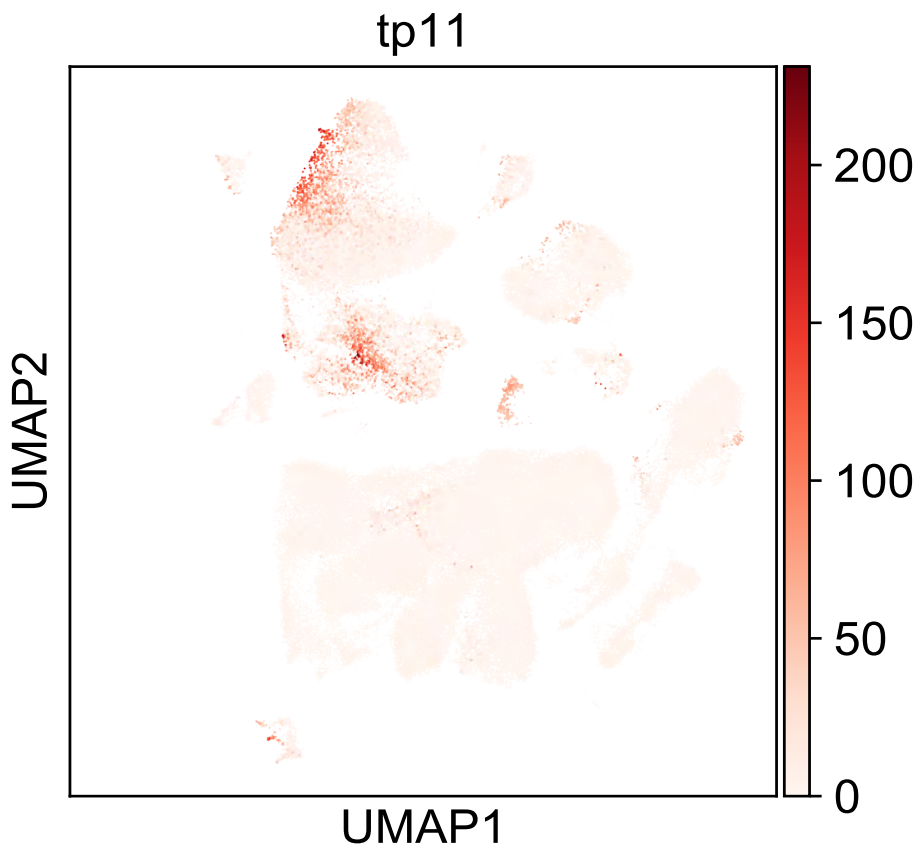
Cell Type

Monocytes

Macrophage



Imm GEM 11

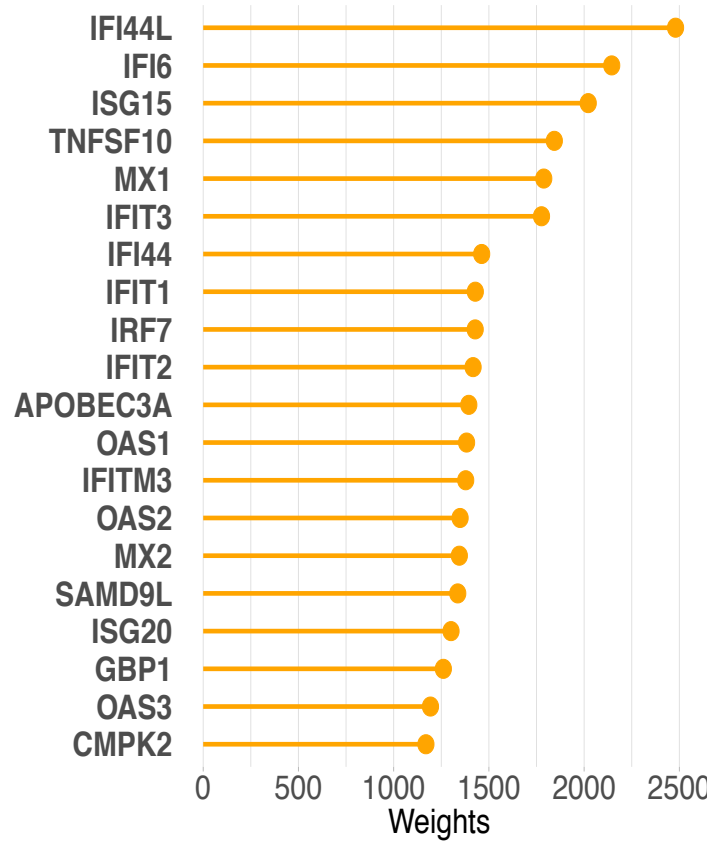


Cell Type

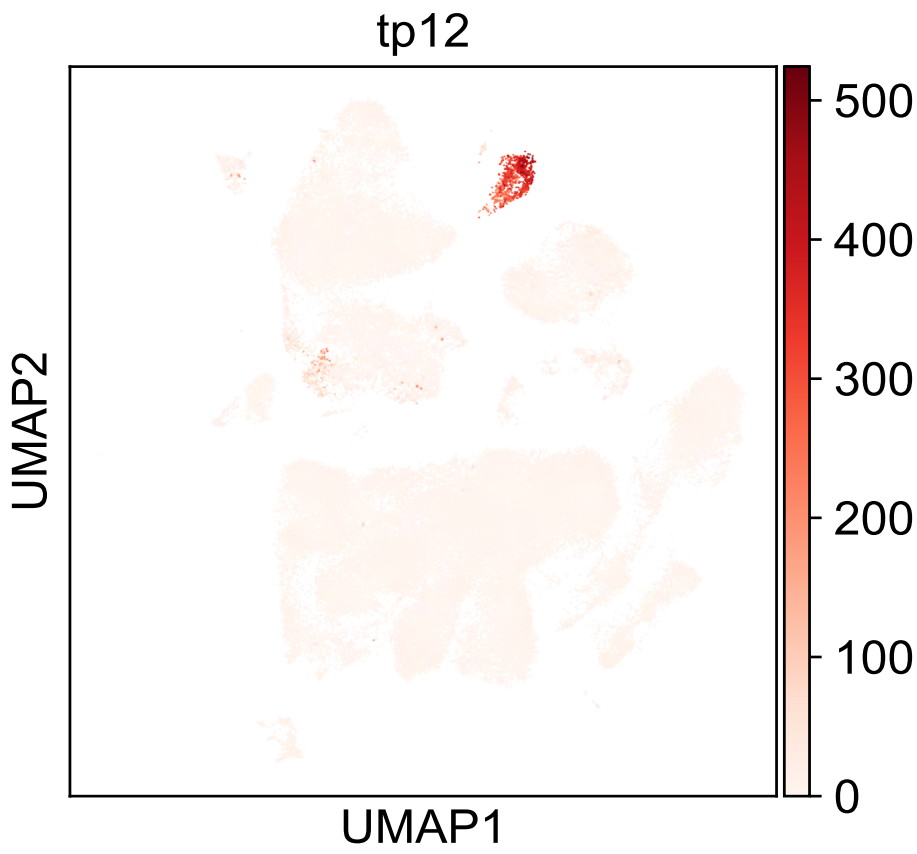
Monocytes

Macrophage

Top 20 genes

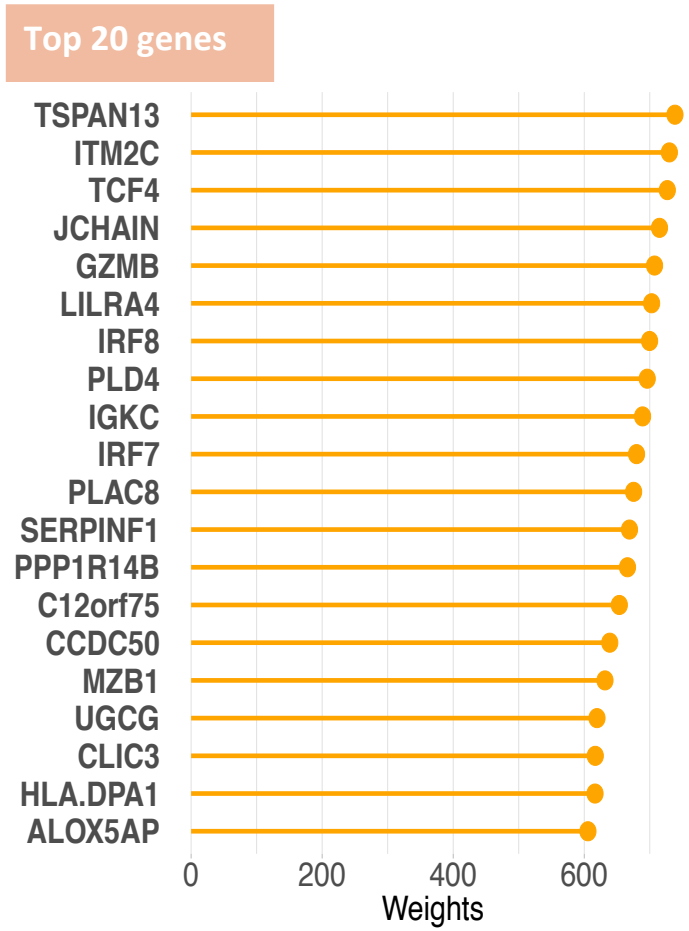


Imm GEM 12

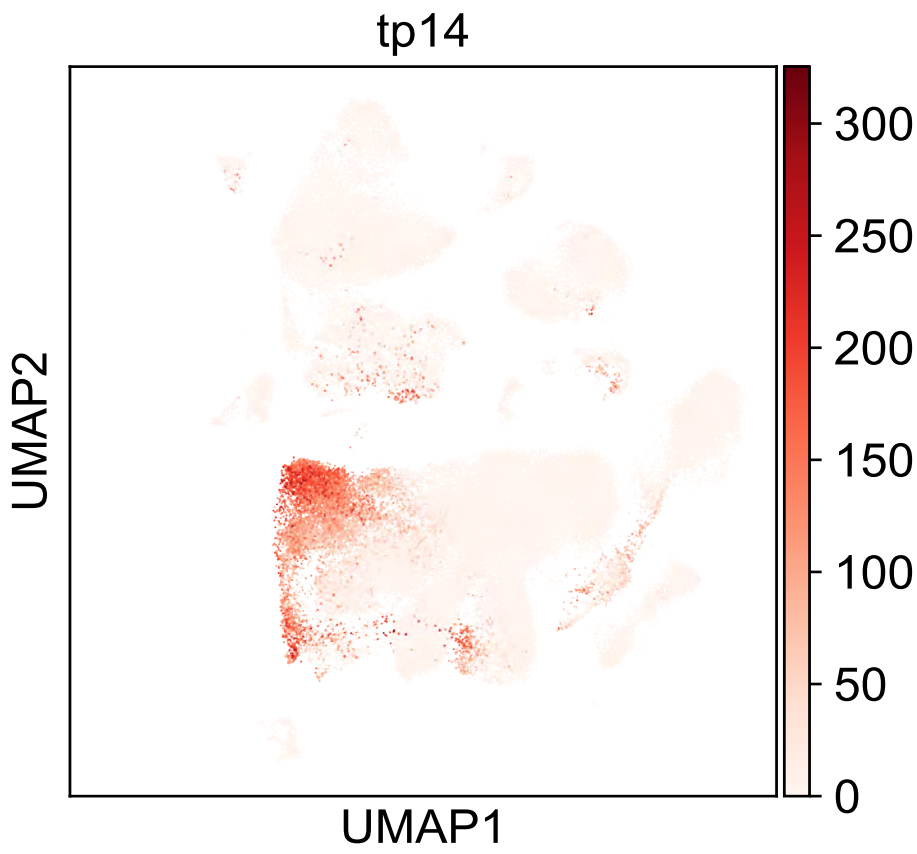


Cell Type

unknown



Imm GEM 14



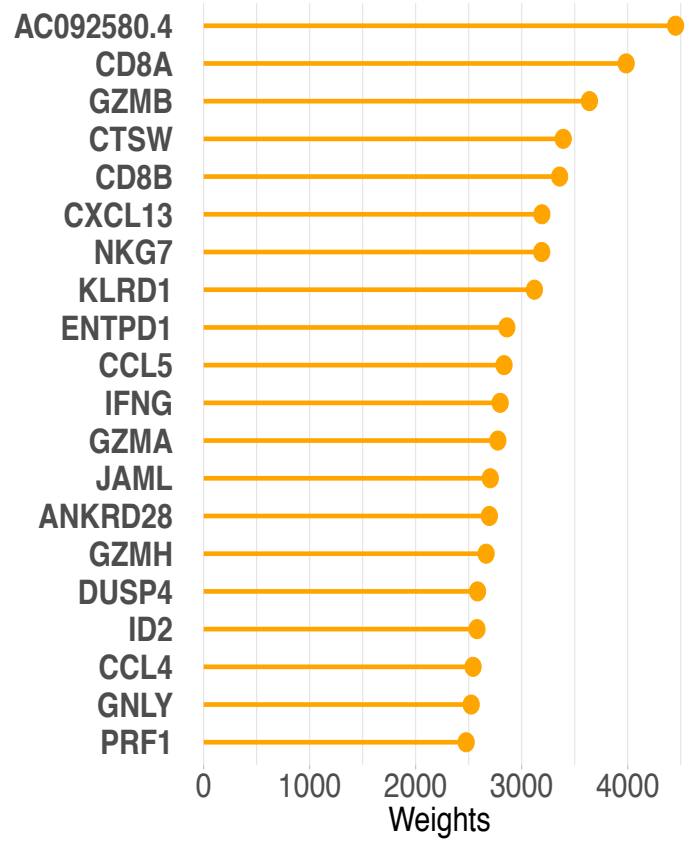
Cell Type

CD8 T cells

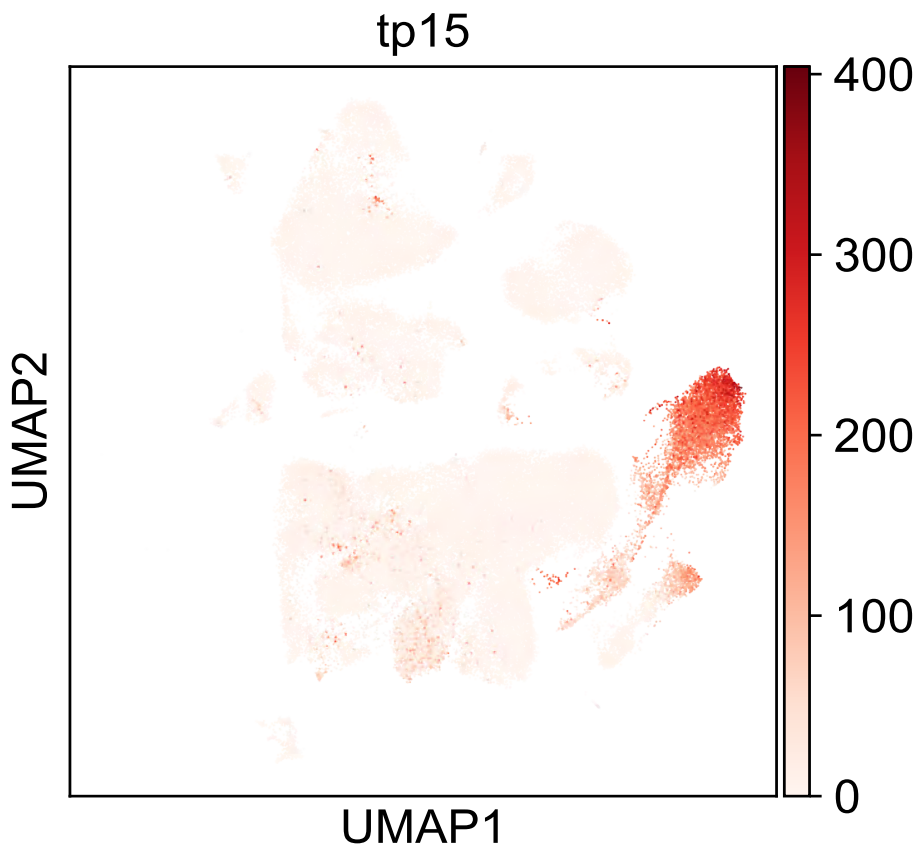
Tregs

NK cells

Top 20 genes

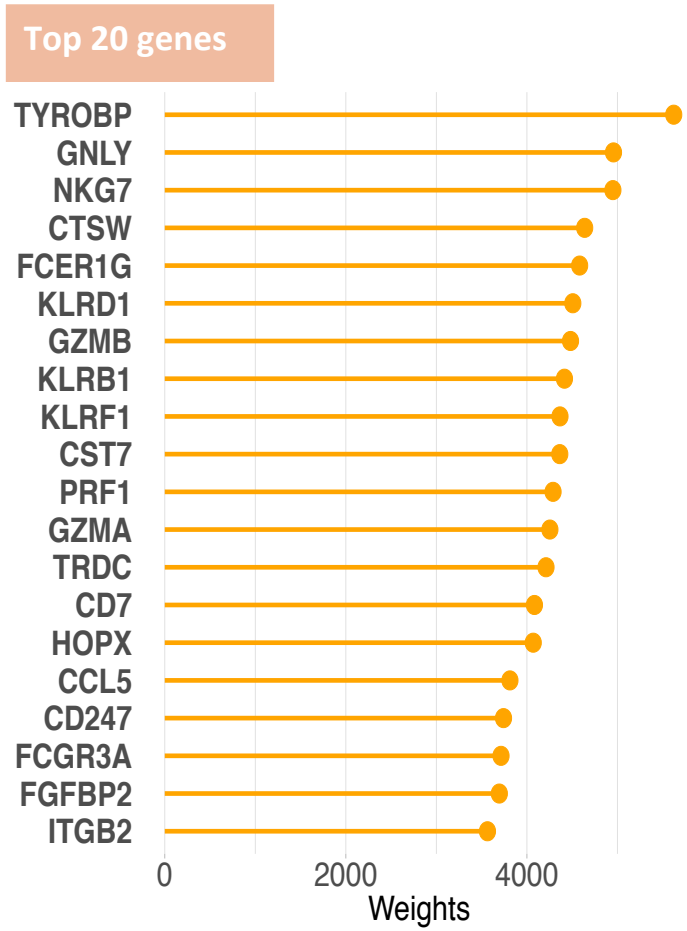


Imm GEM 15

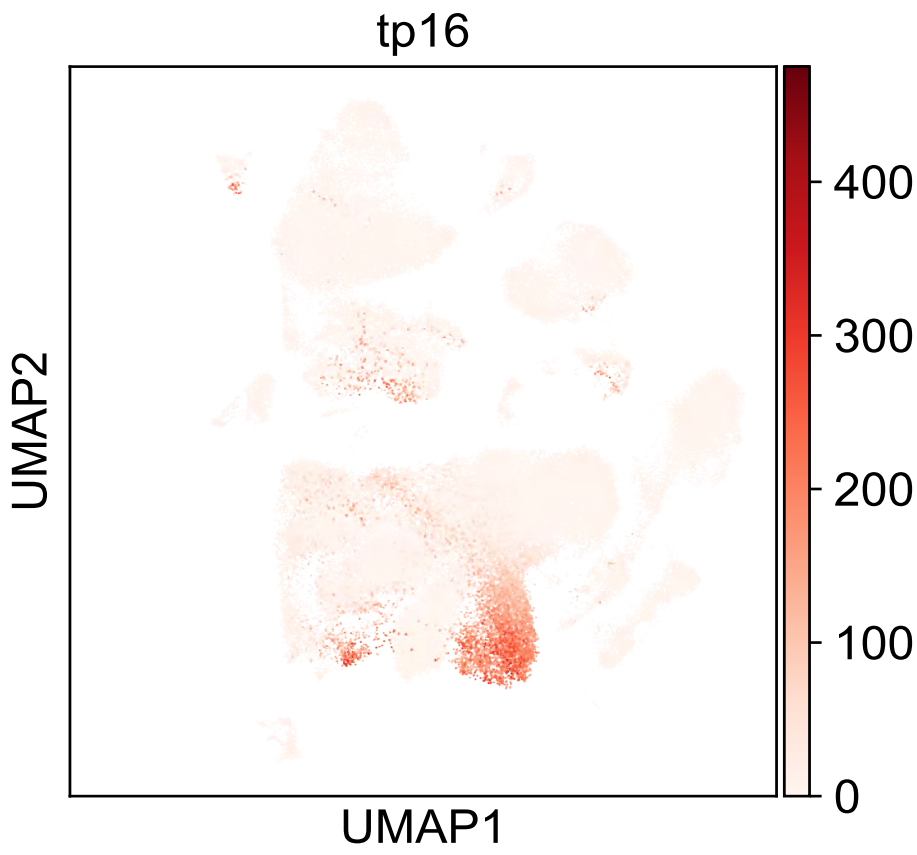


Cell Type

NK cells



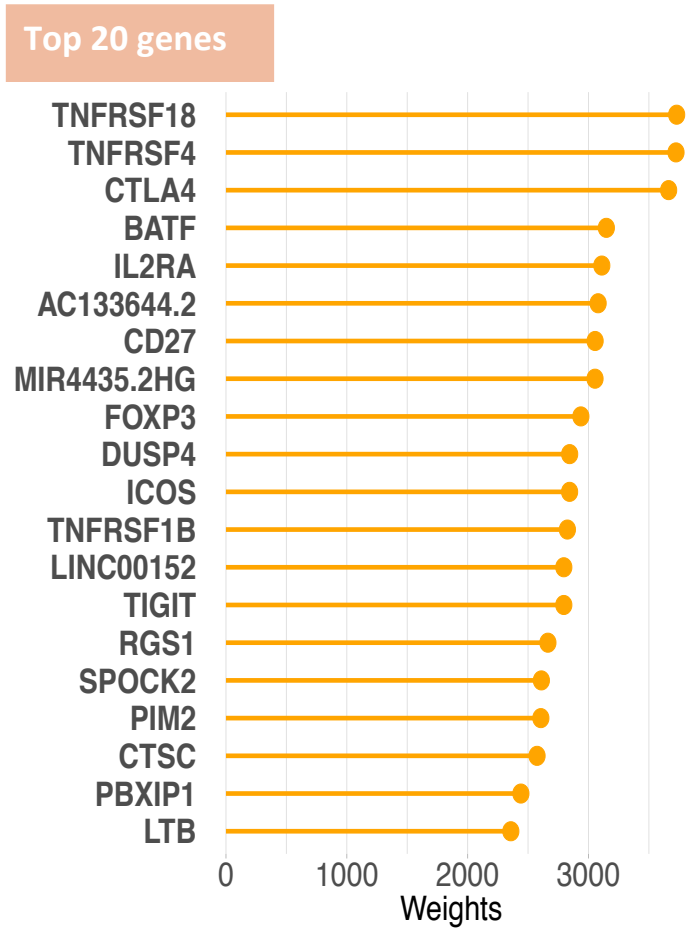
Imm GEM 16



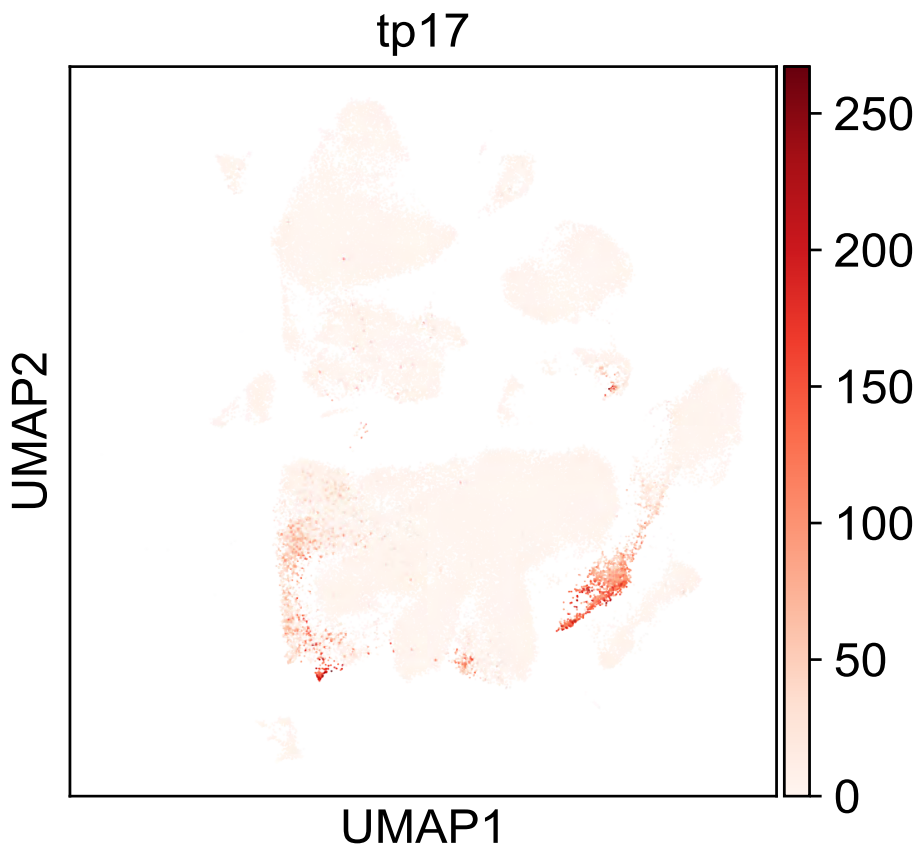
Cell Type

Tregs

CD4 T cells



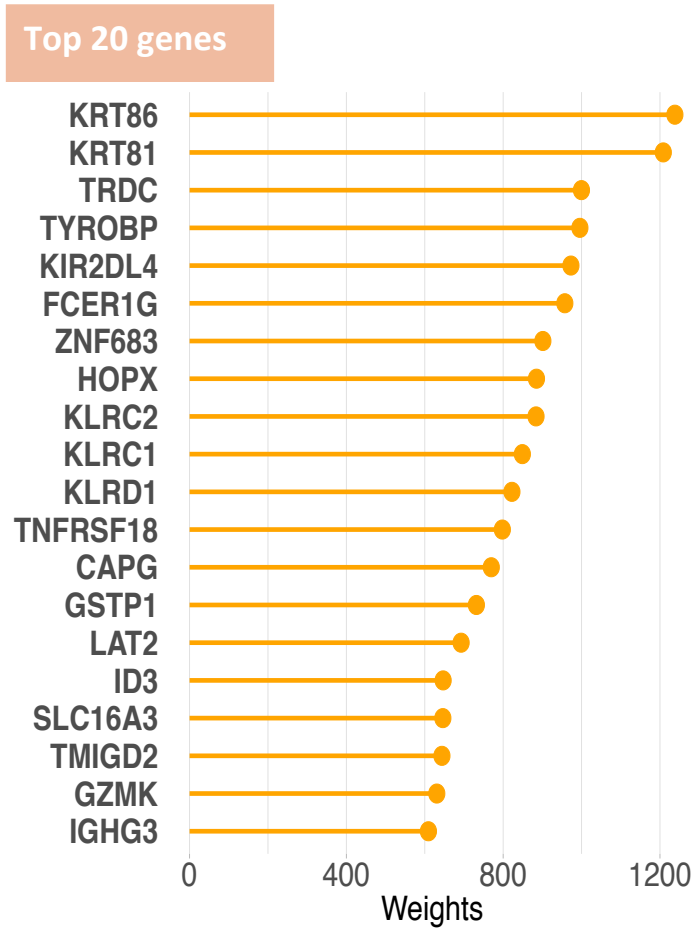
Imm GEM 17



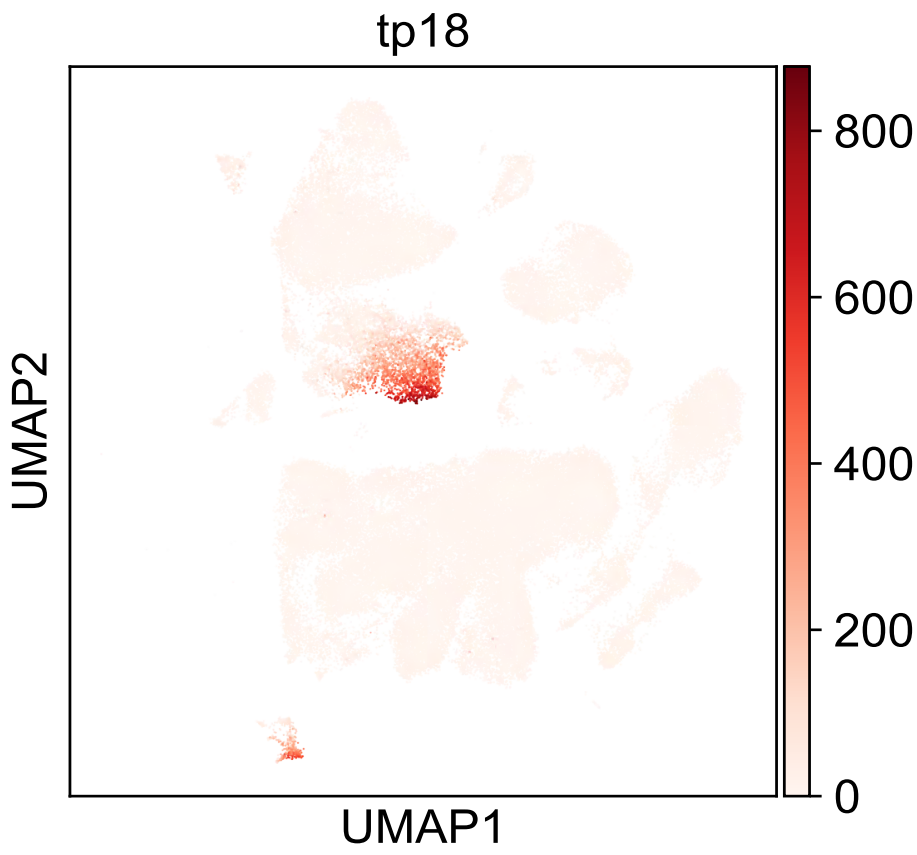
Cell Type

NK cells

CD8 T cells

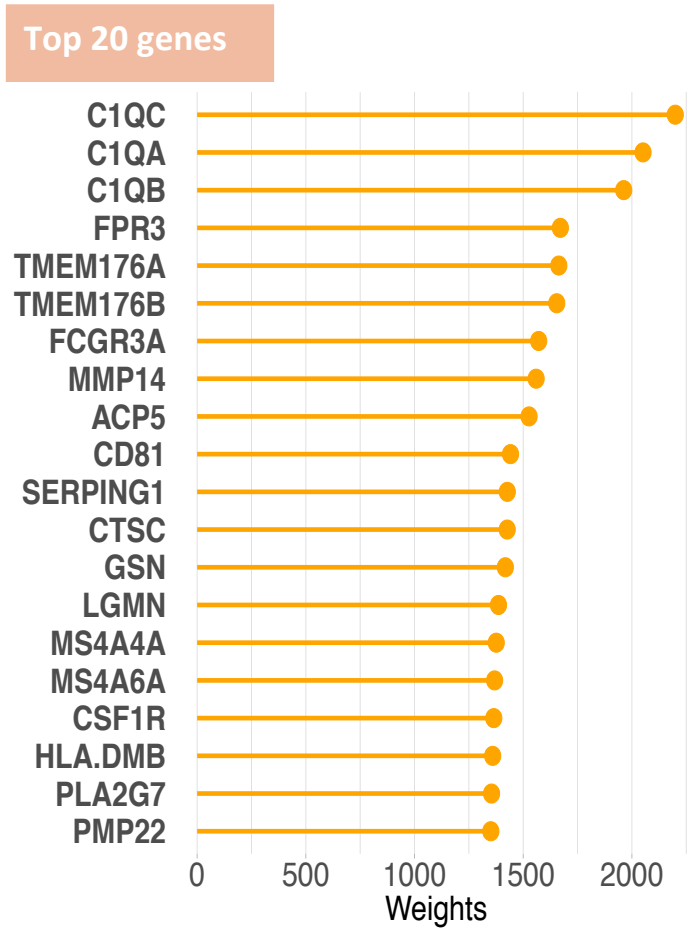


Imm GEM 18

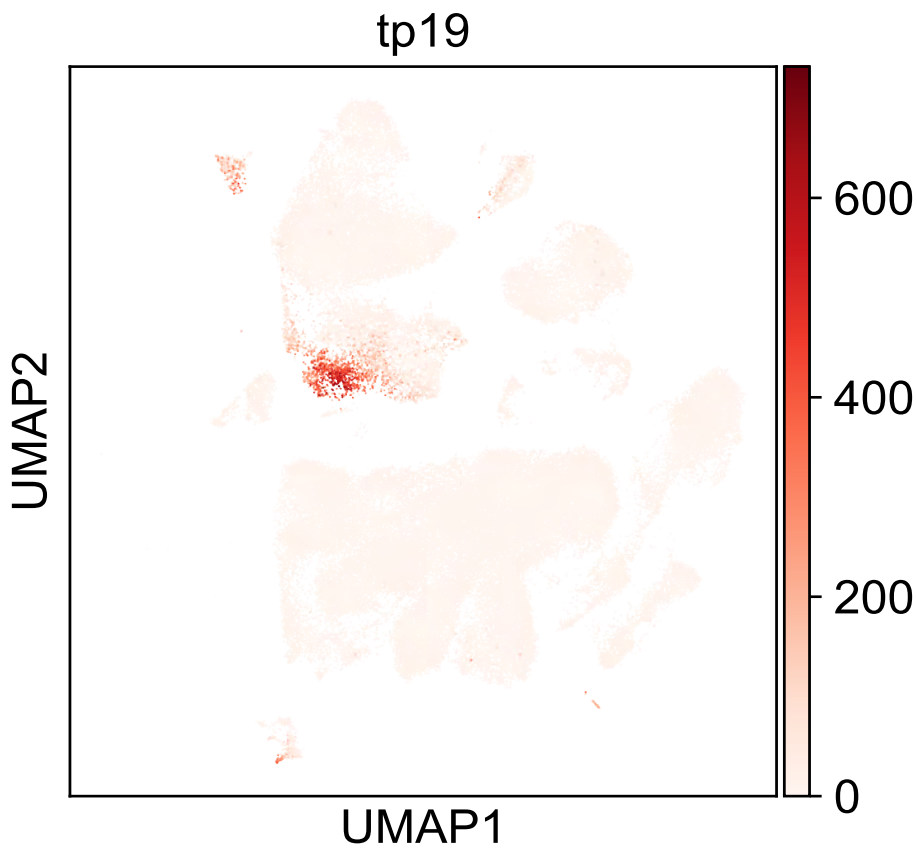


Cell Type

Macrophage



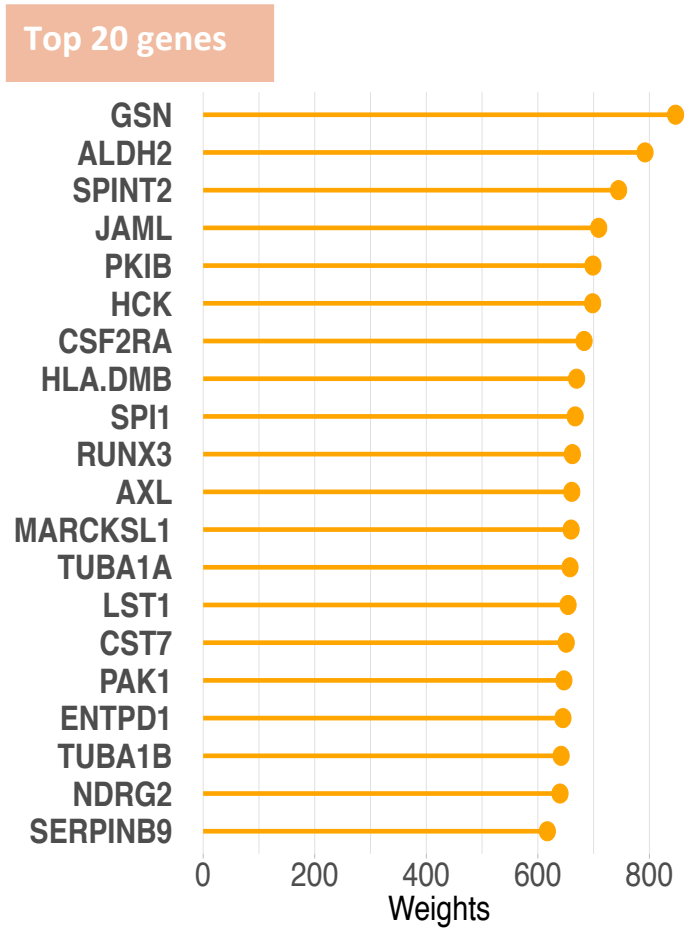
Imm GEM 19



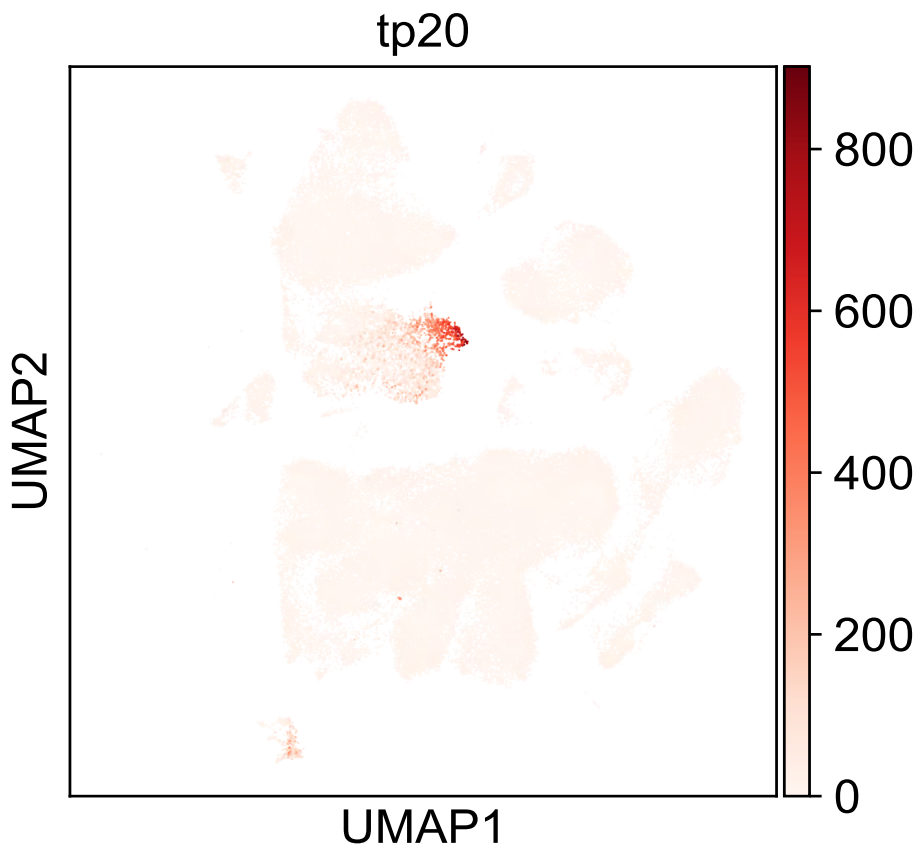
Cell Type

Macrophage

Dendritic

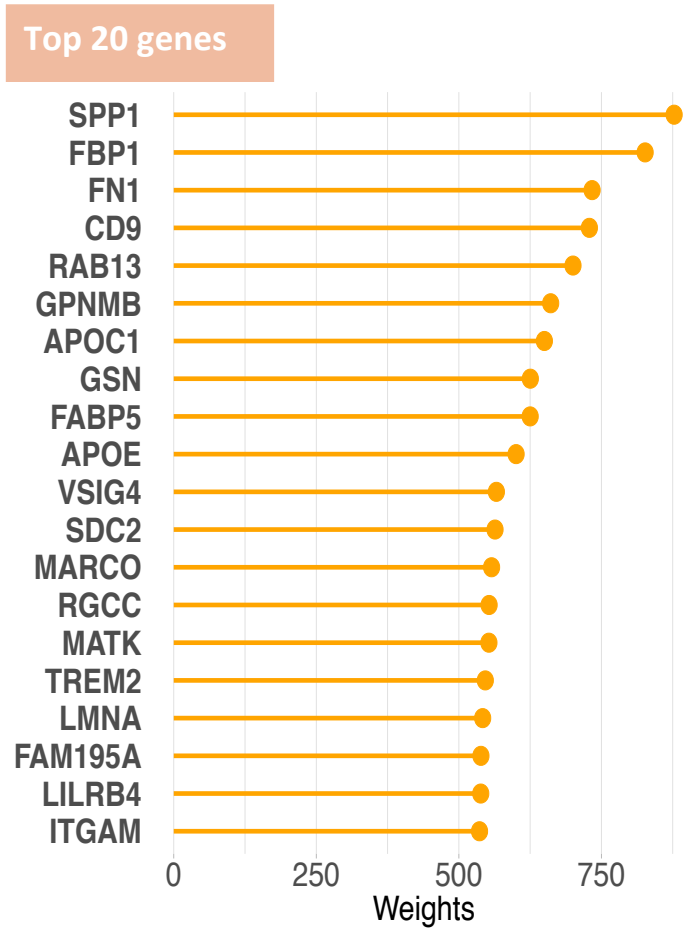


Imm GEM 20

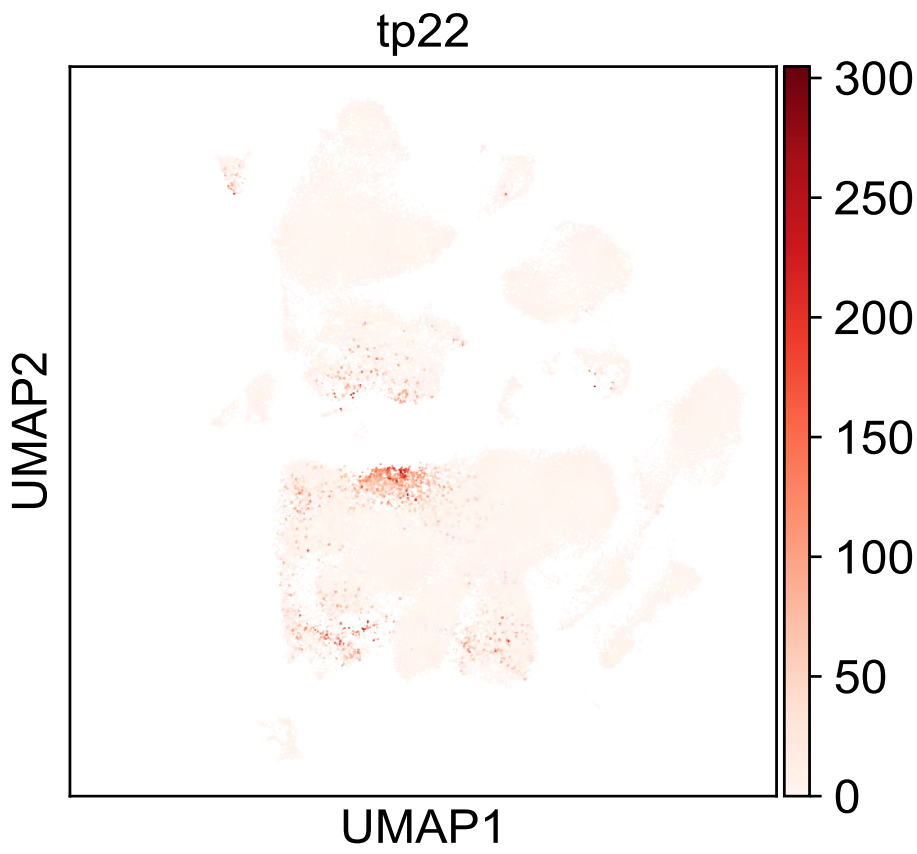


Cell Type

Macrophage



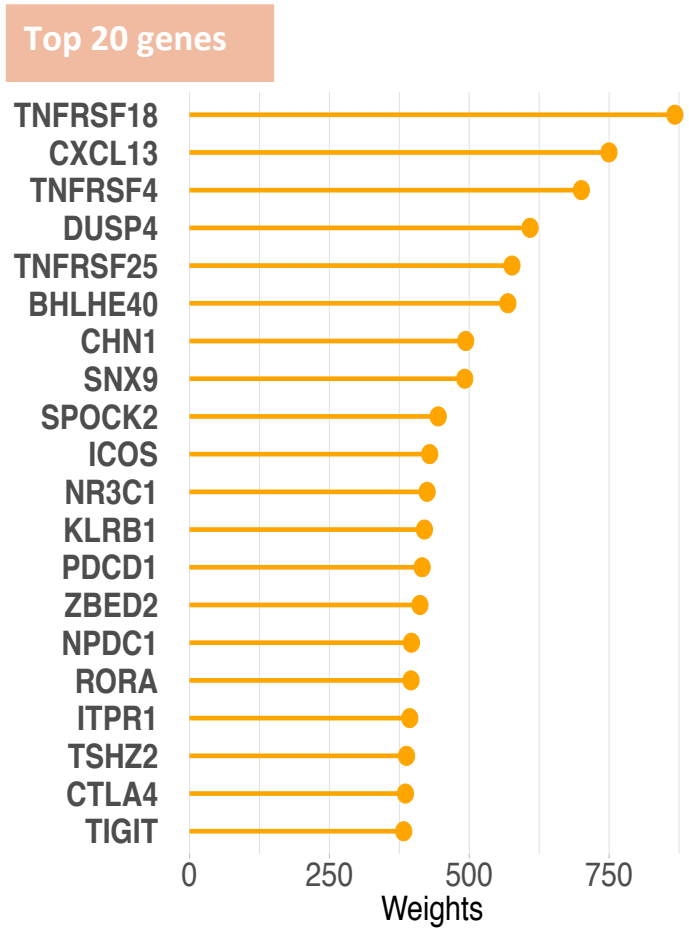
Imm GEM 22



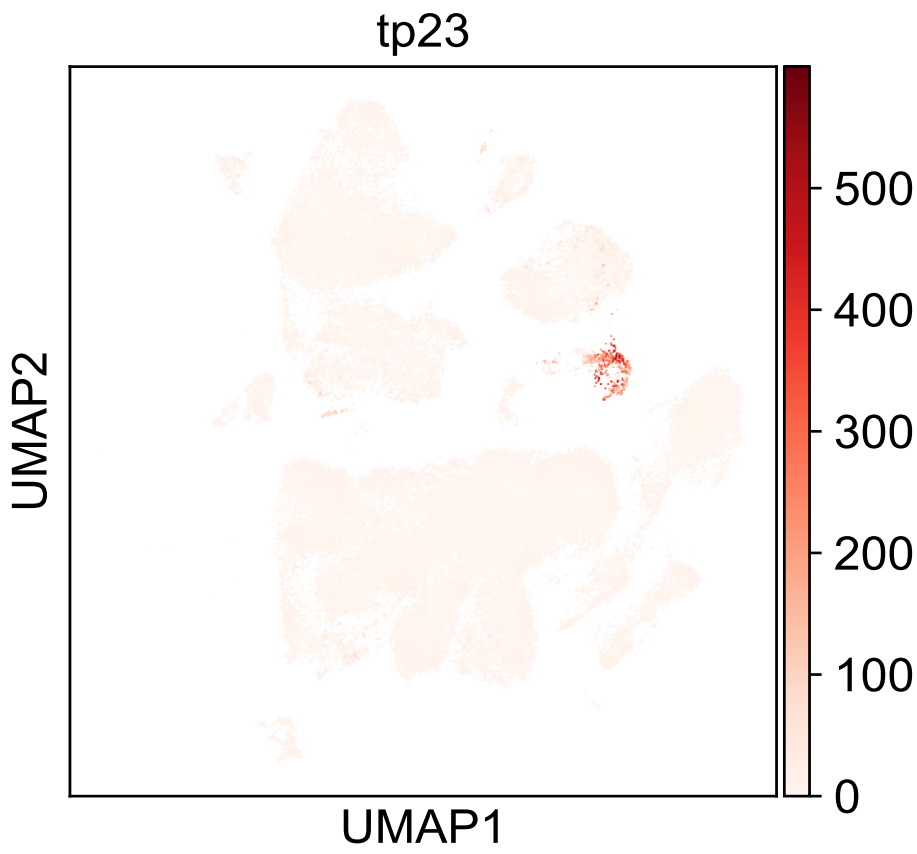
Cell Type

CD4 T cells

CD8 T cells



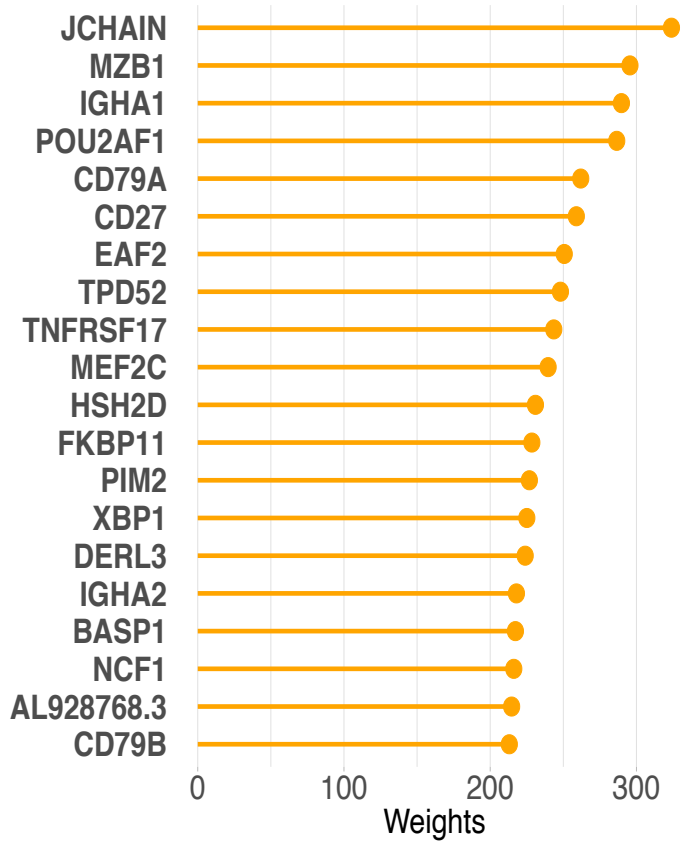
Imm GEM 23



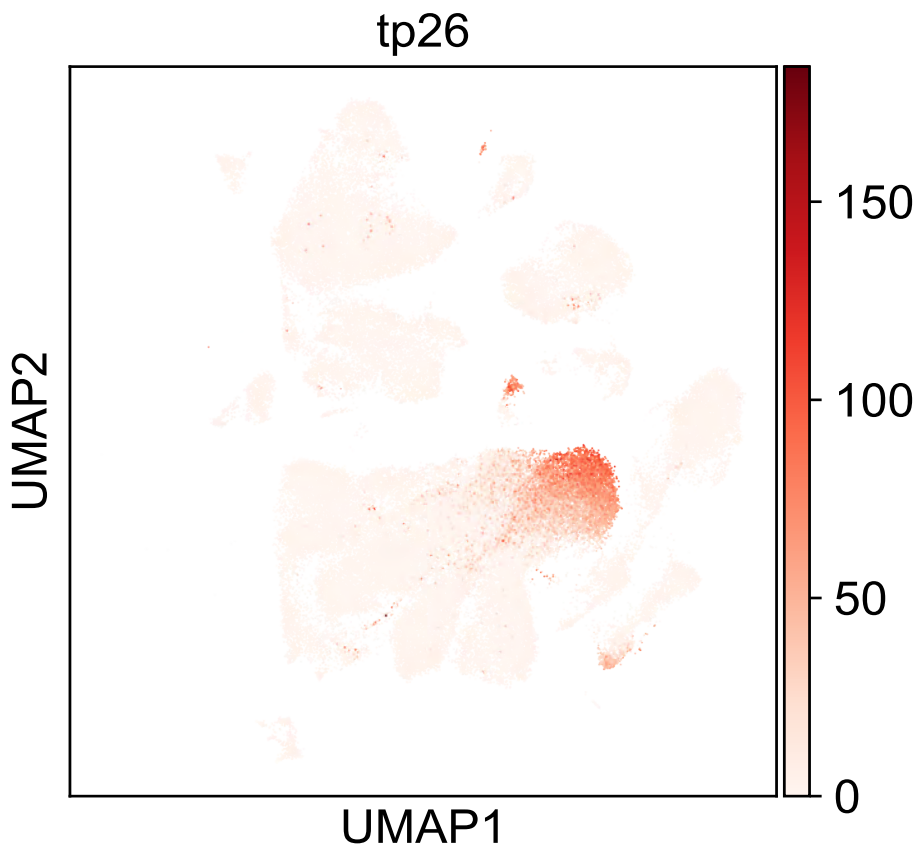
Cell Type

B cells

Top 20 genes



Imm GEM 26

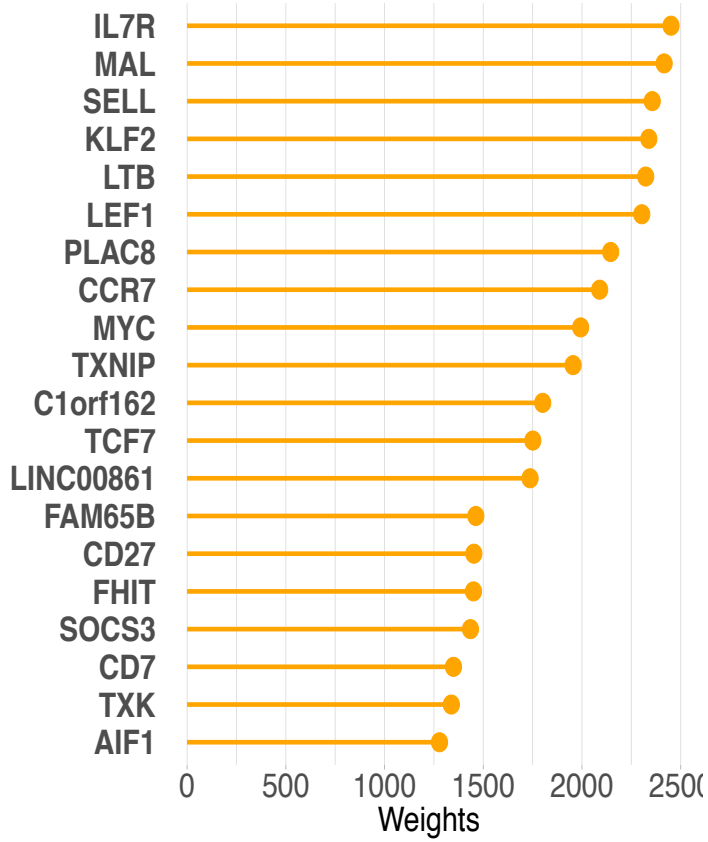


Cell Type

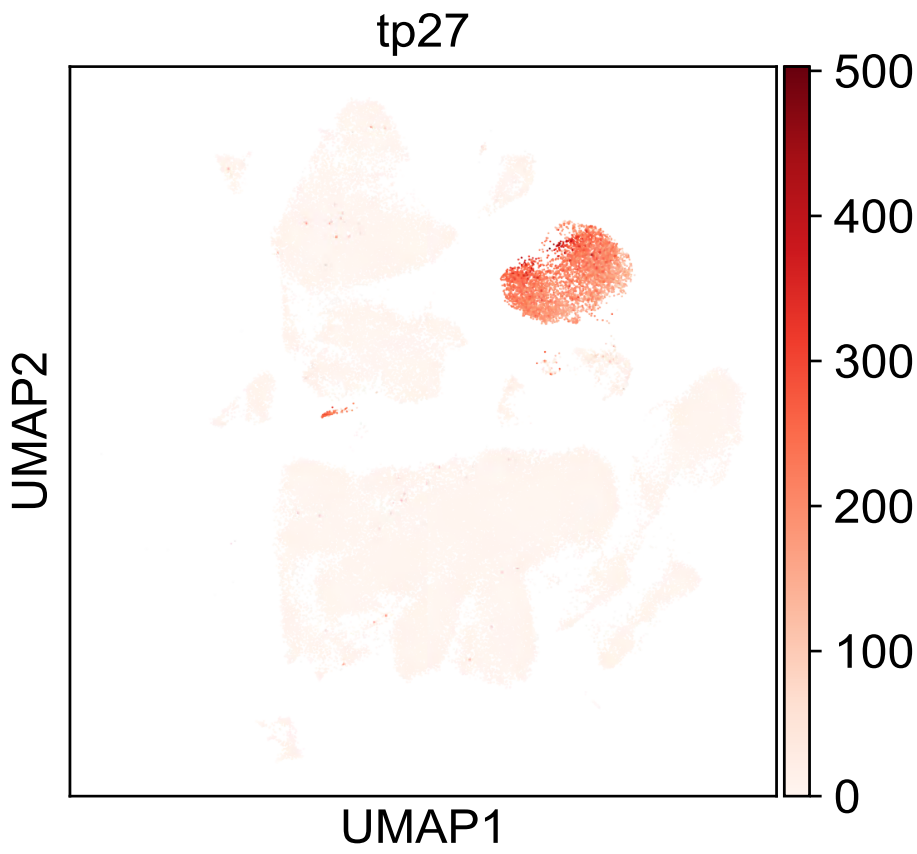
CD4 T cells

NK cells

Top 20 genes



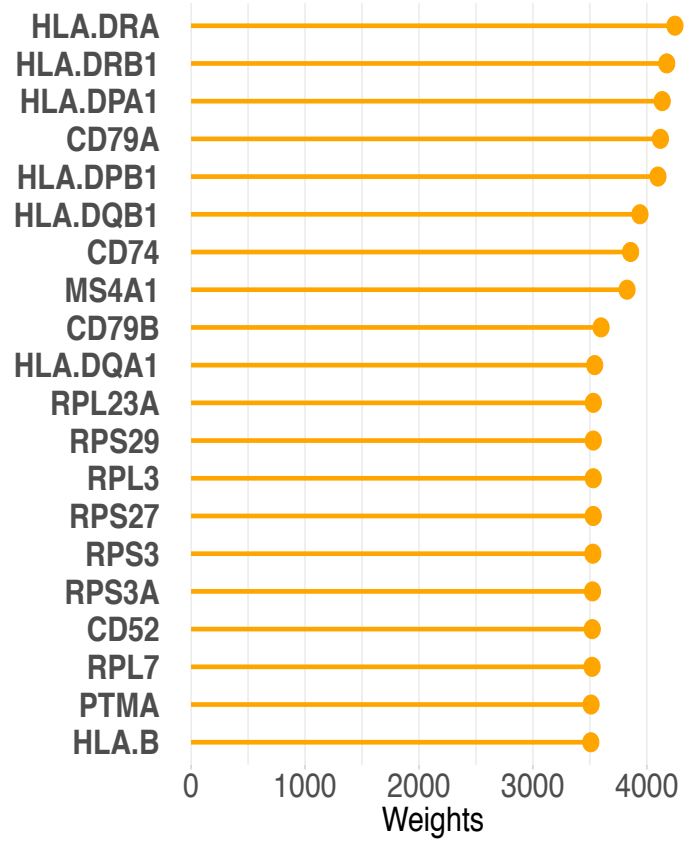
Imm GEM 27



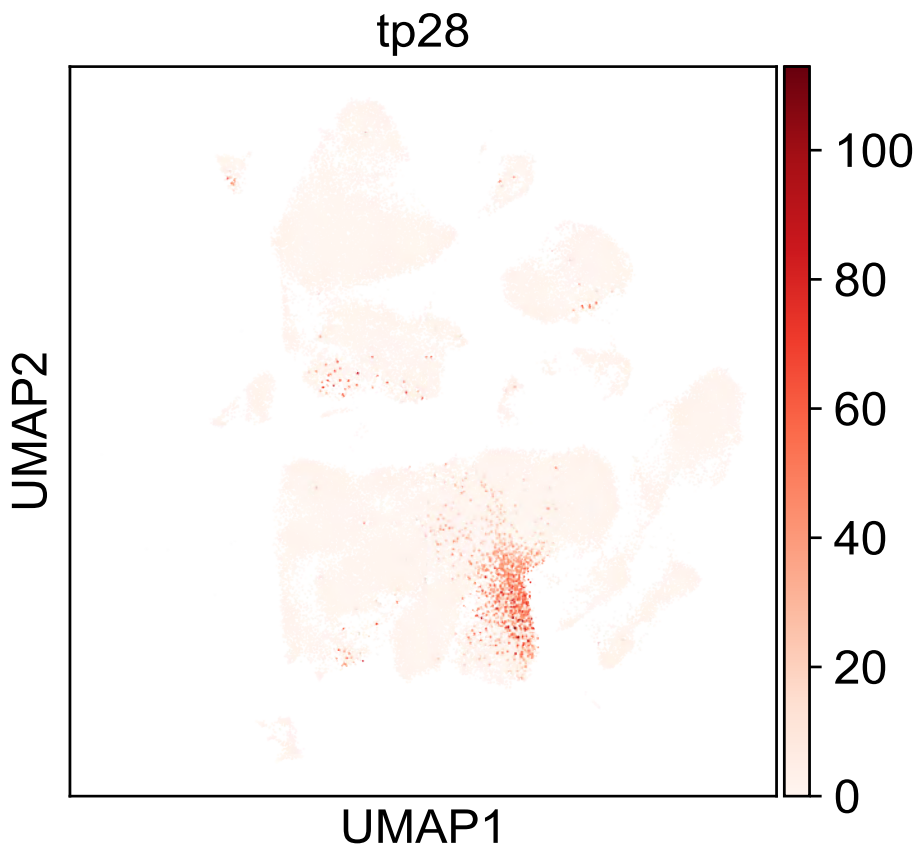
Cell Type

B cells

Top 20 genes



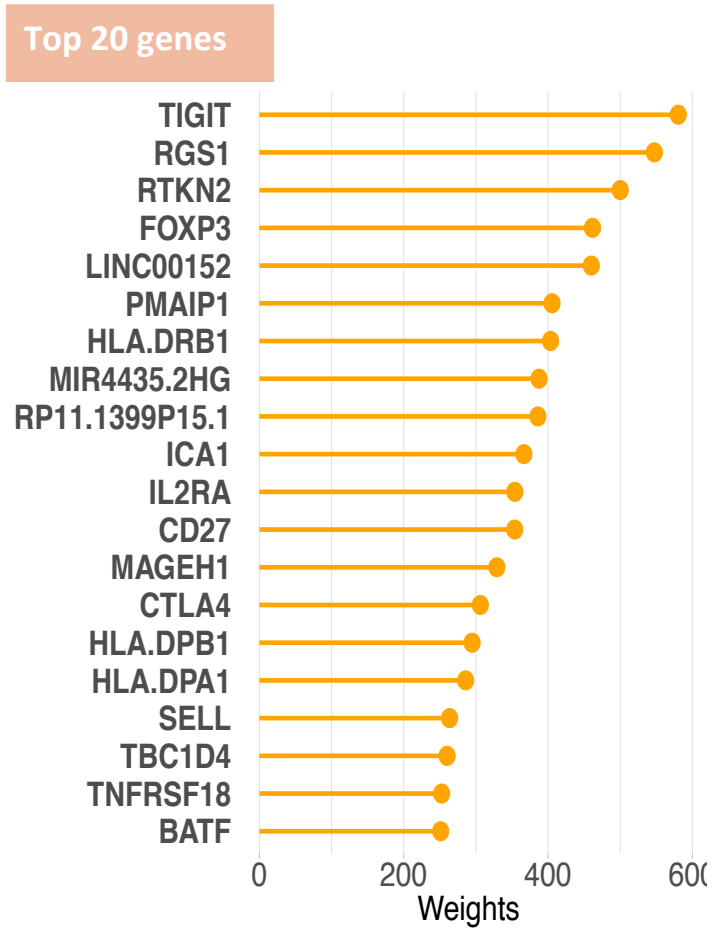
Imm GEM 28



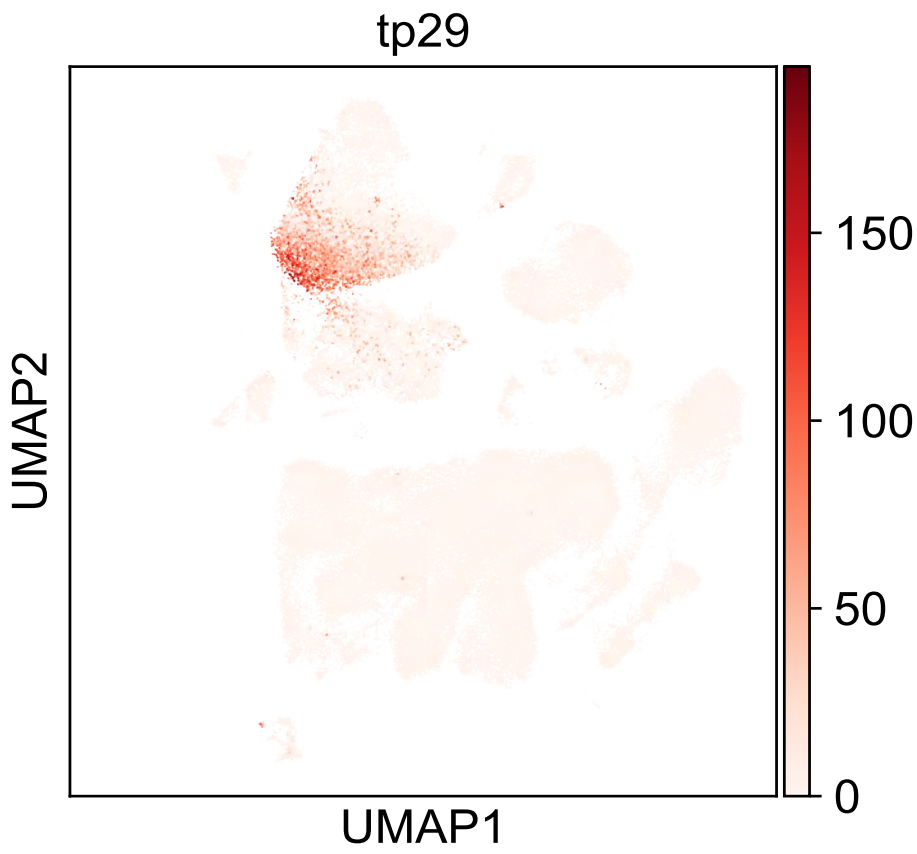
Cell Type

CD4 T cells

Tregs

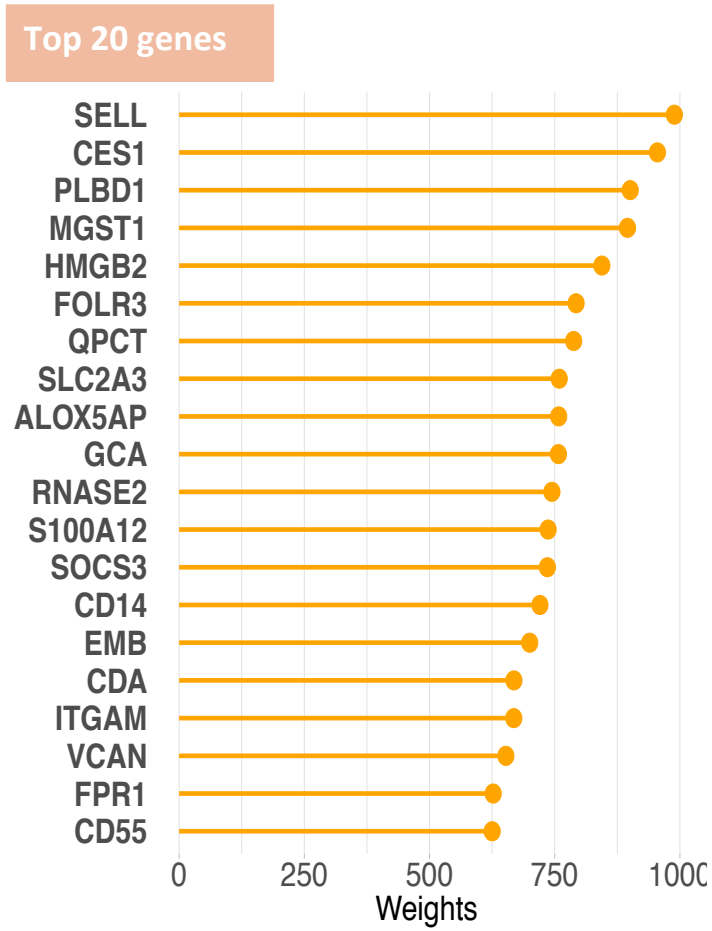


Imm GEM 29

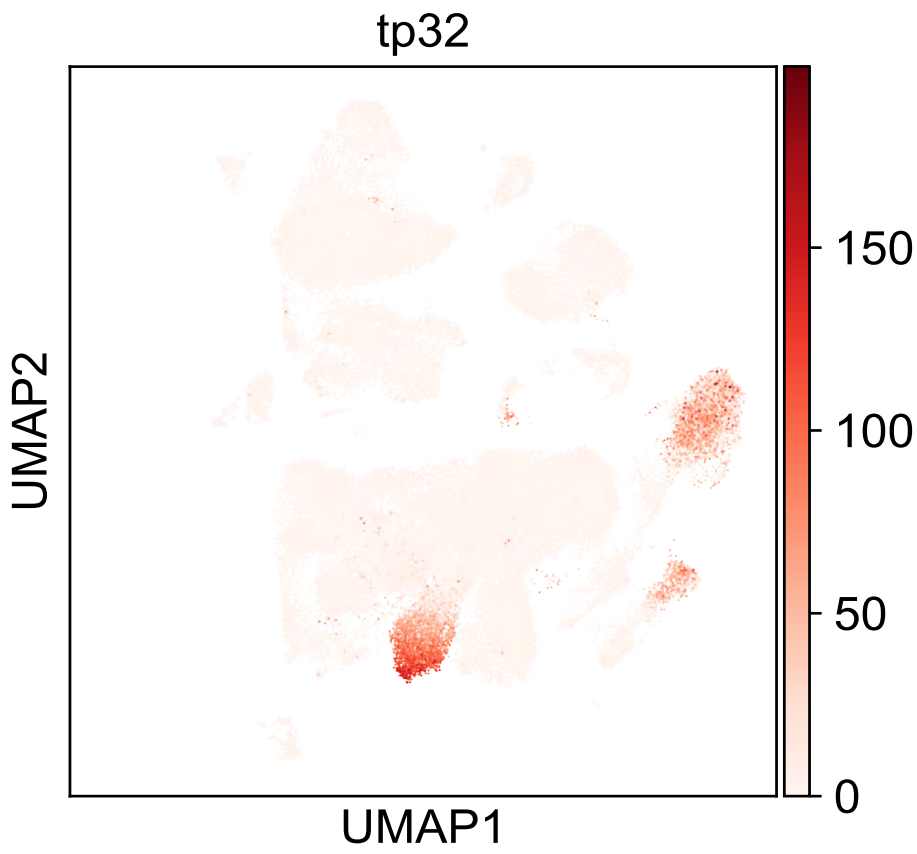


Cell Type

Monocytes



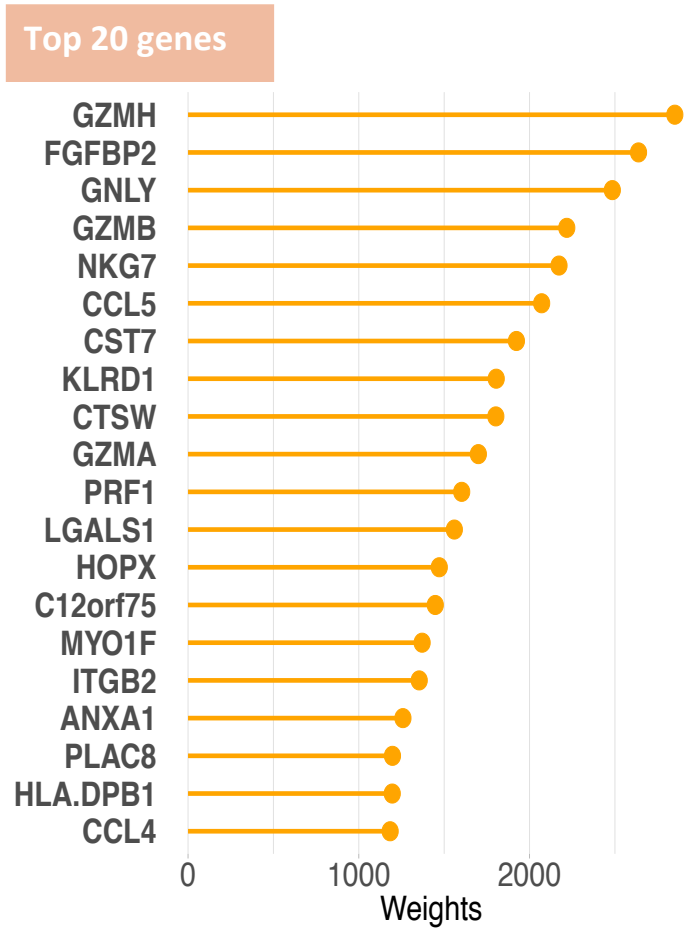
Imm GEM 32



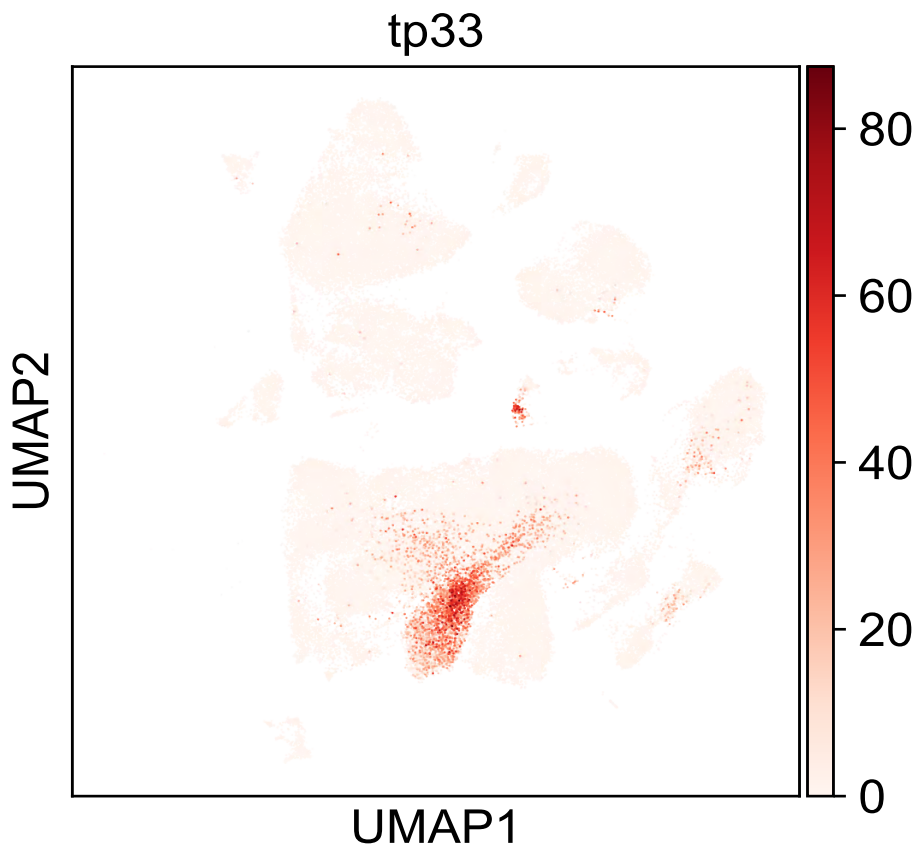
Cell Type

CD8 T cells

NK cells

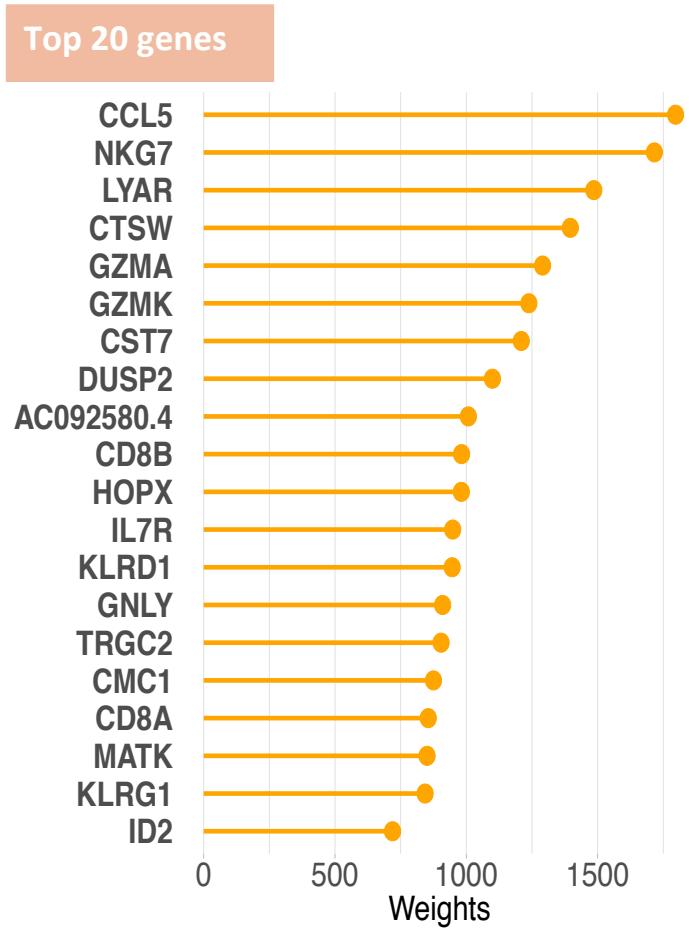


Imm GEM 33

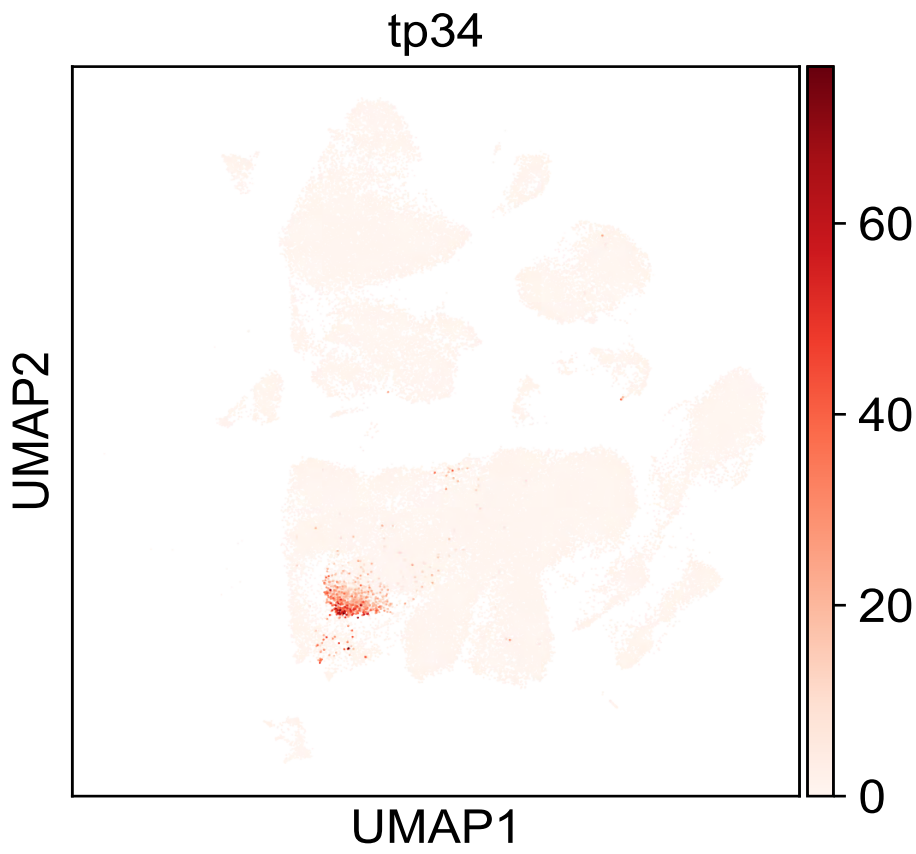


Cell Type

CD8 T cells

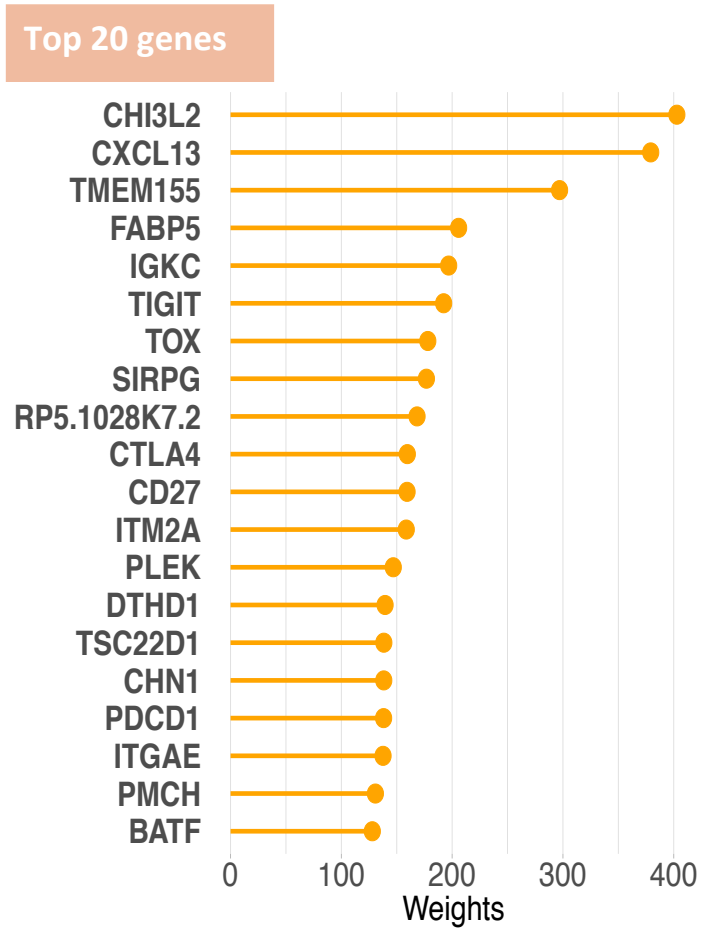


Imm GEM 34

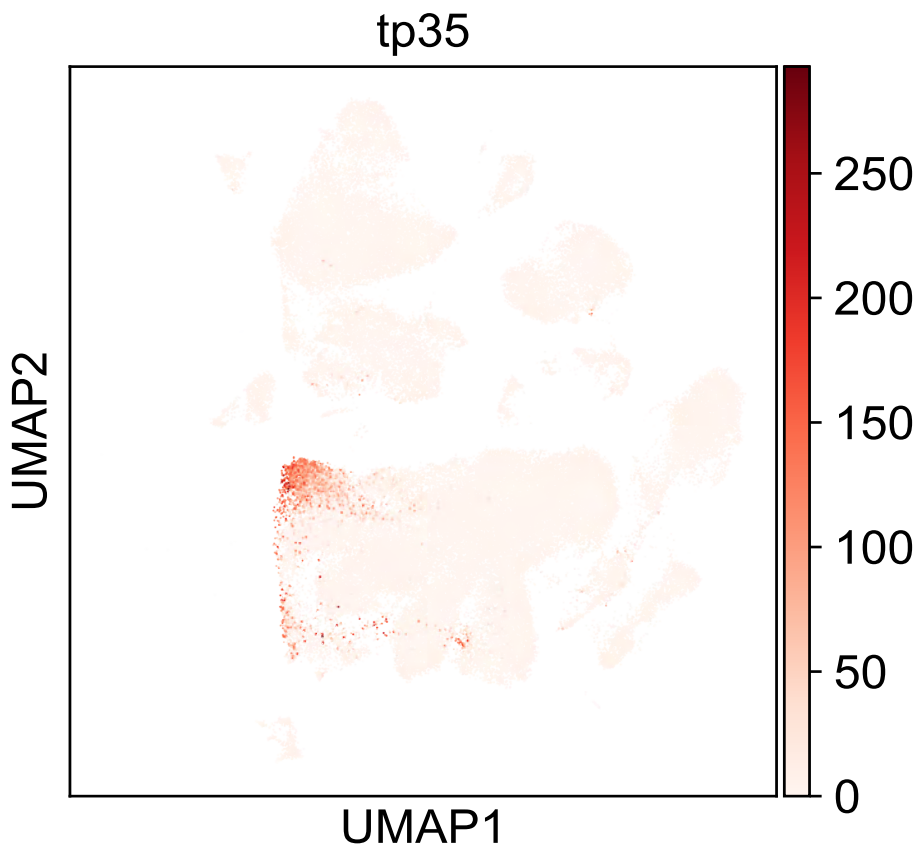


Cell Type

CD8 T cells

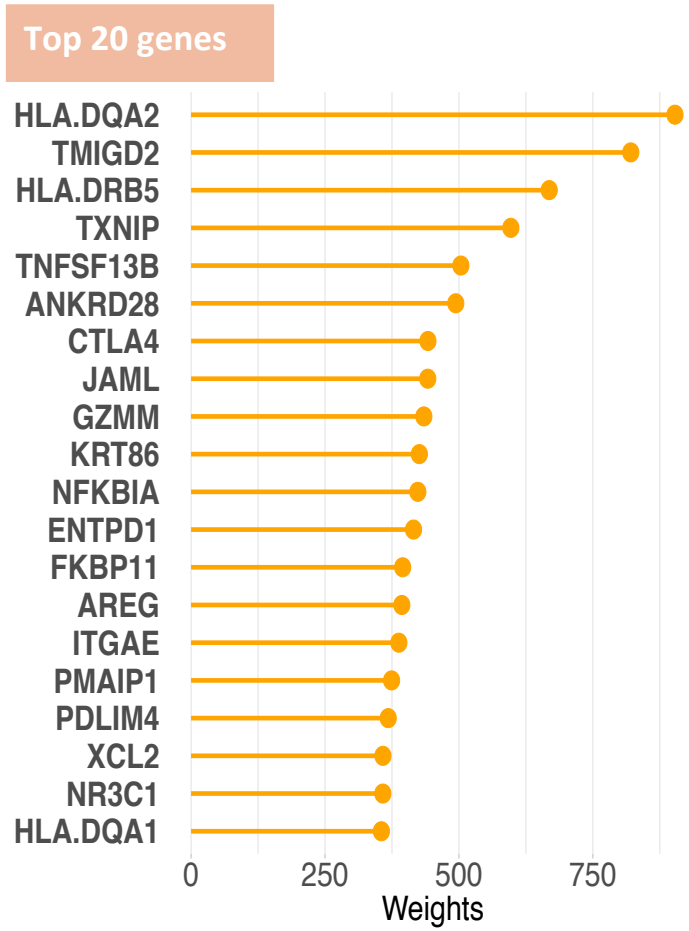


Imm GEM 35

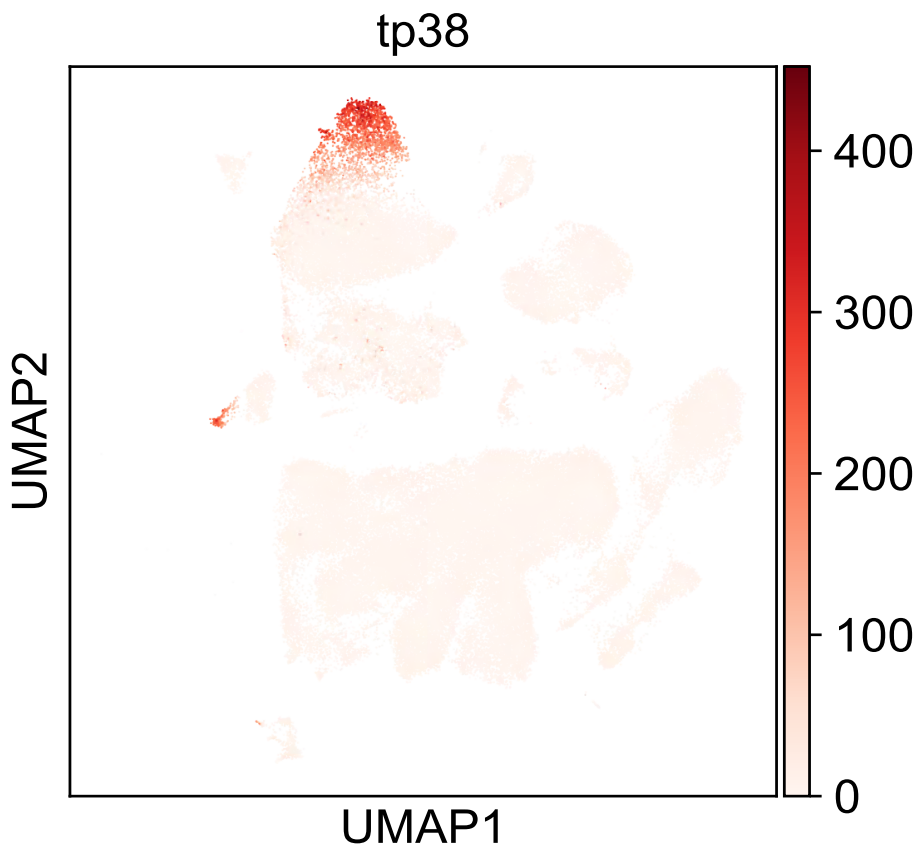


Cell Type

CD8 T cells

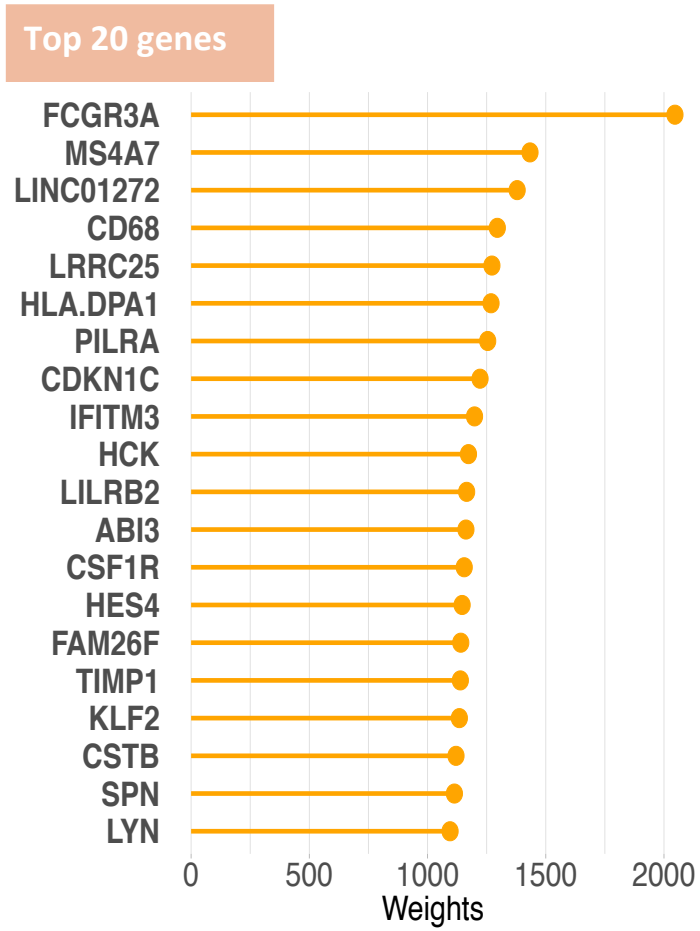


Imm GEM 38

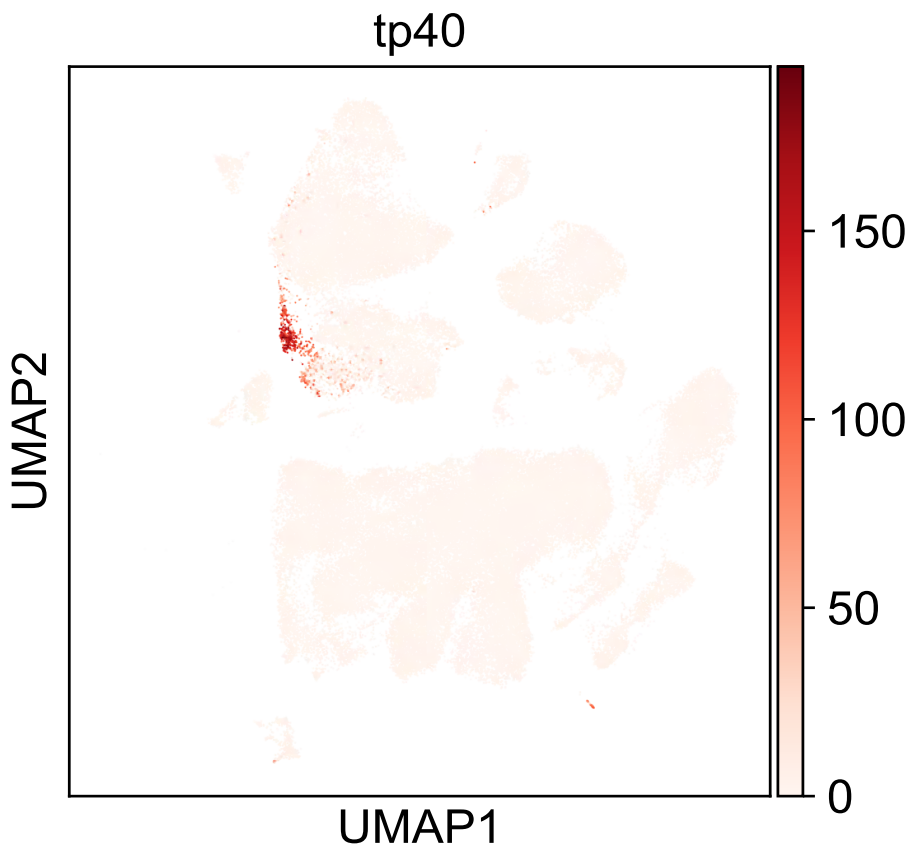


Cell Type

Monocytes

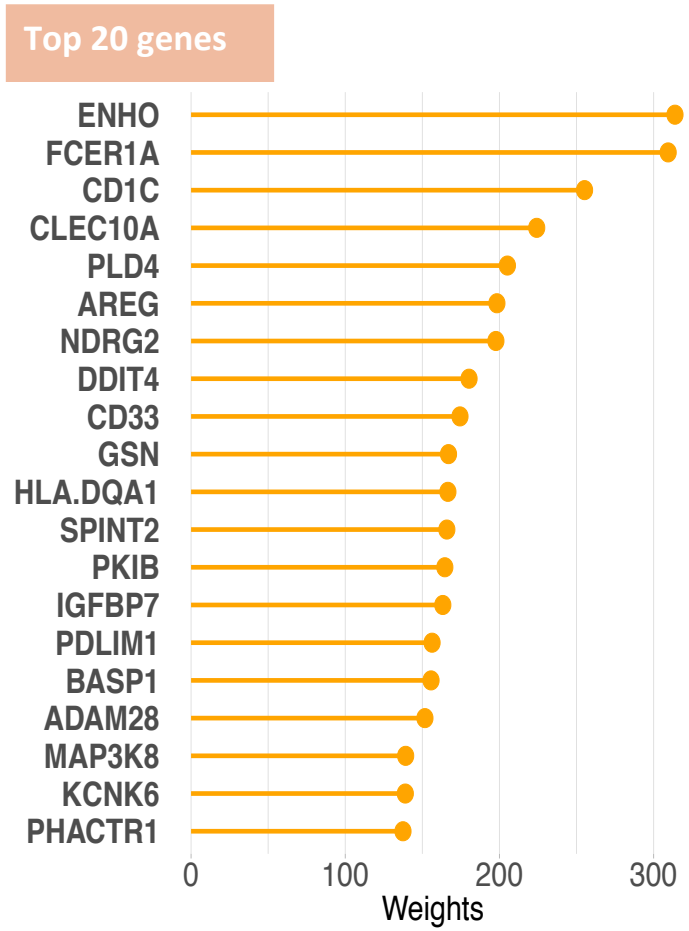


Imm GEM 40

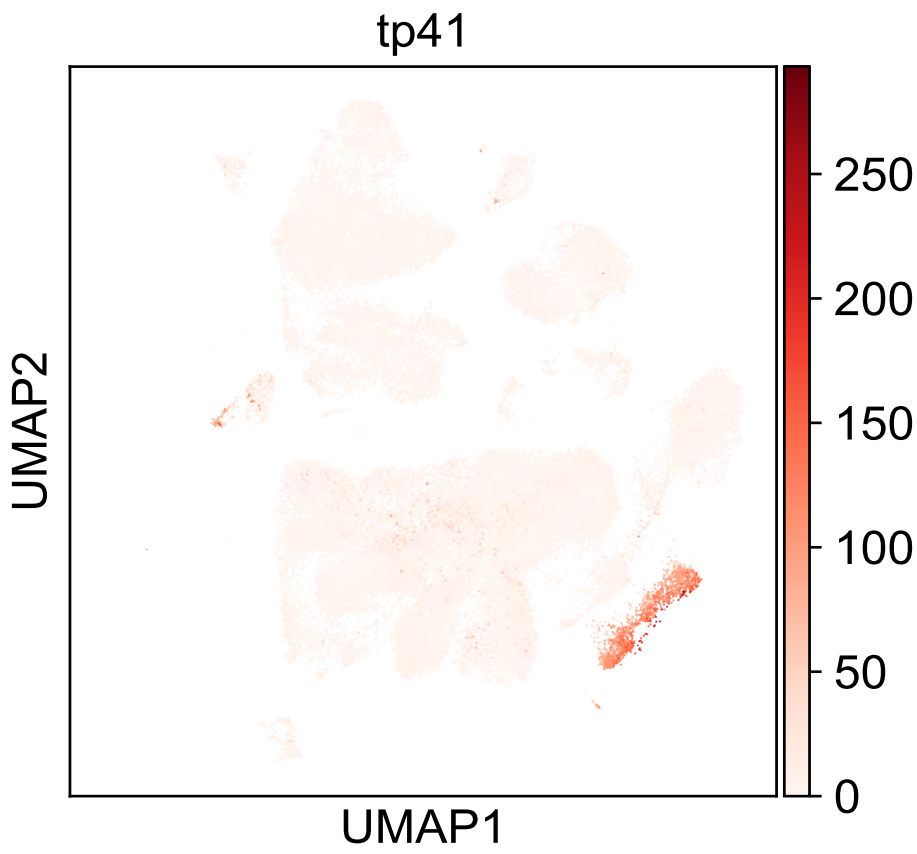


Cell Type

Macrophage

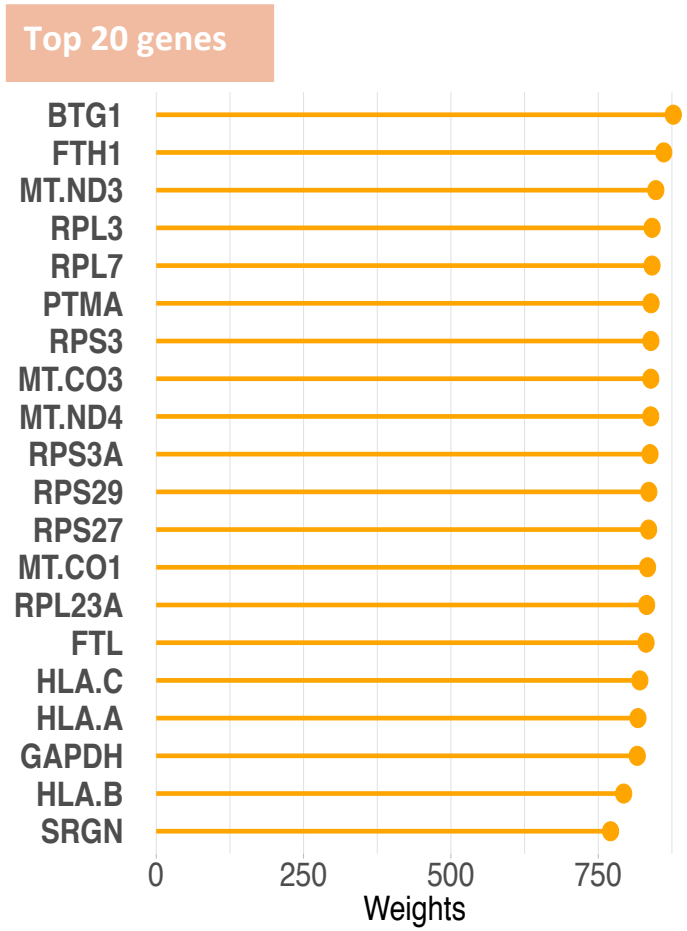


Imm GEM 41

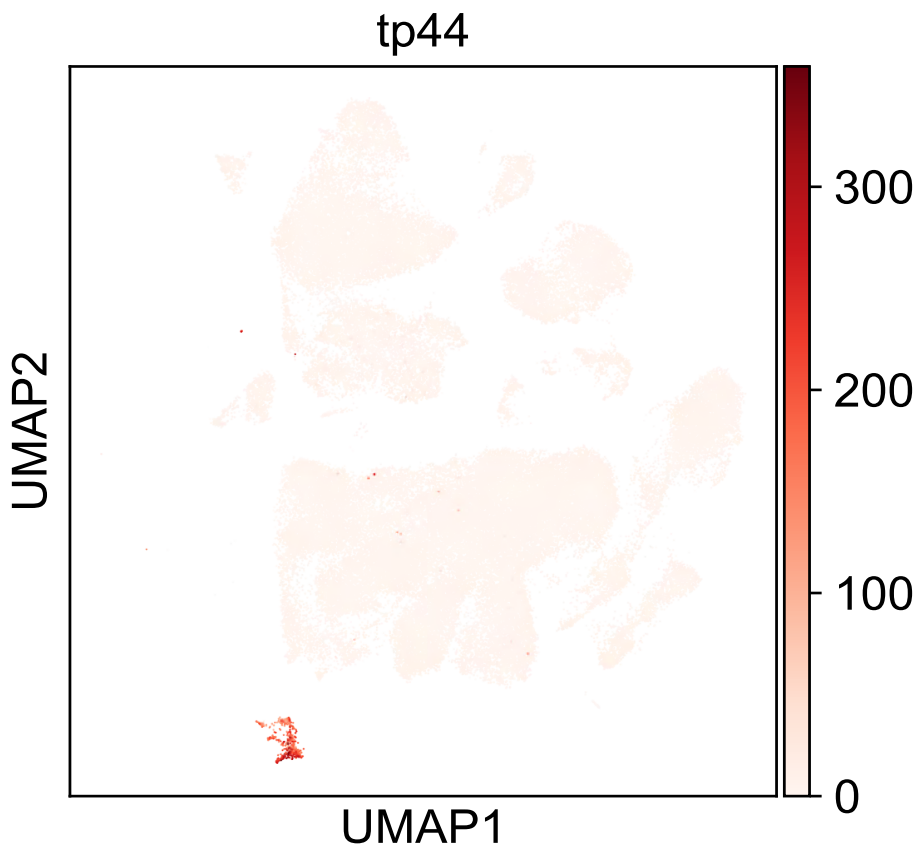


Cell Type

NK cells

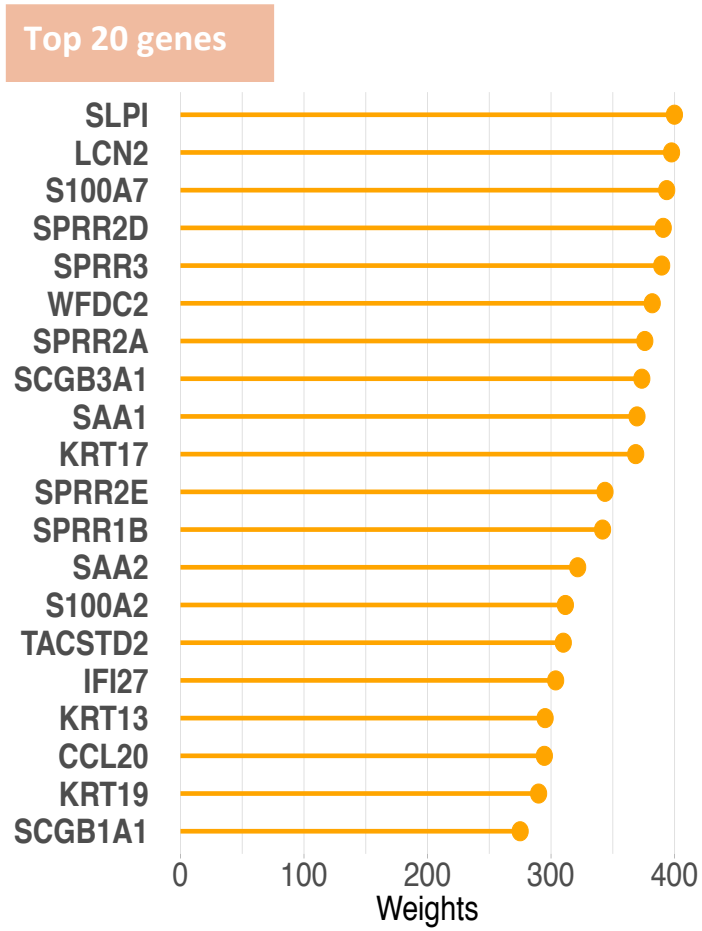


Imm GEM 44

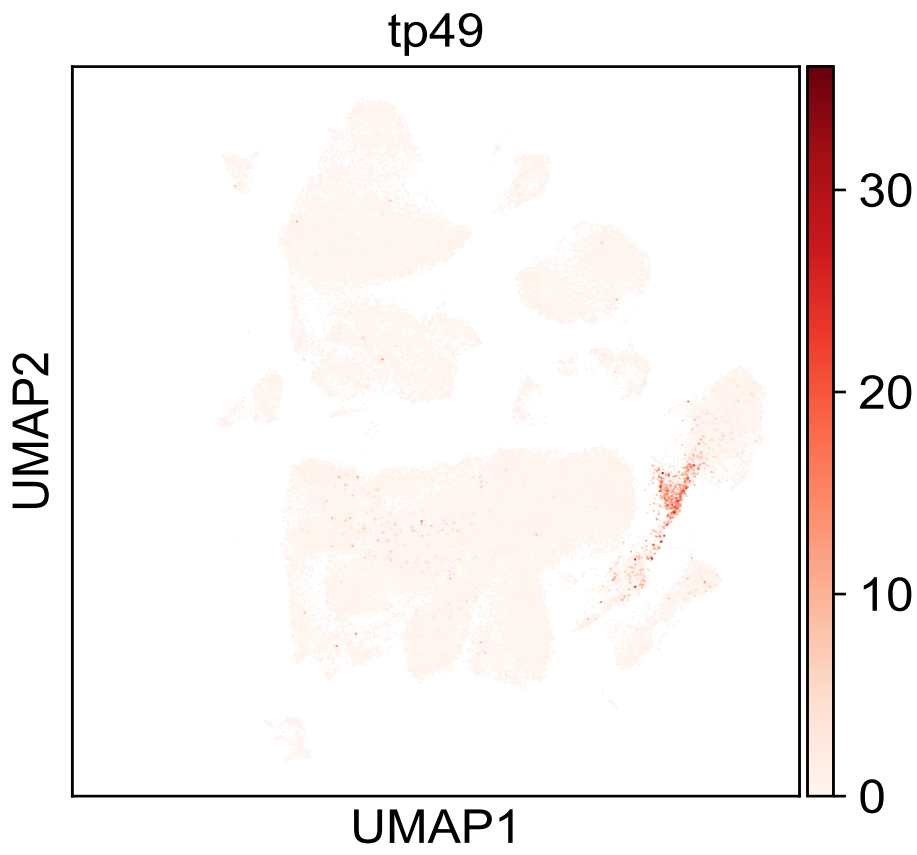


Cell Type

unknown

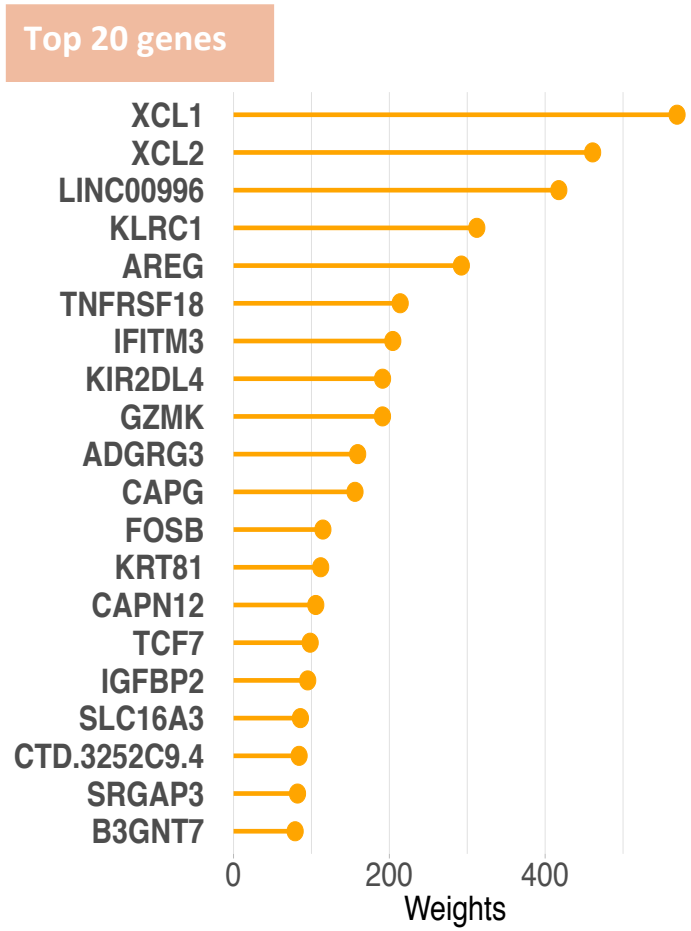


Imm GEM 49

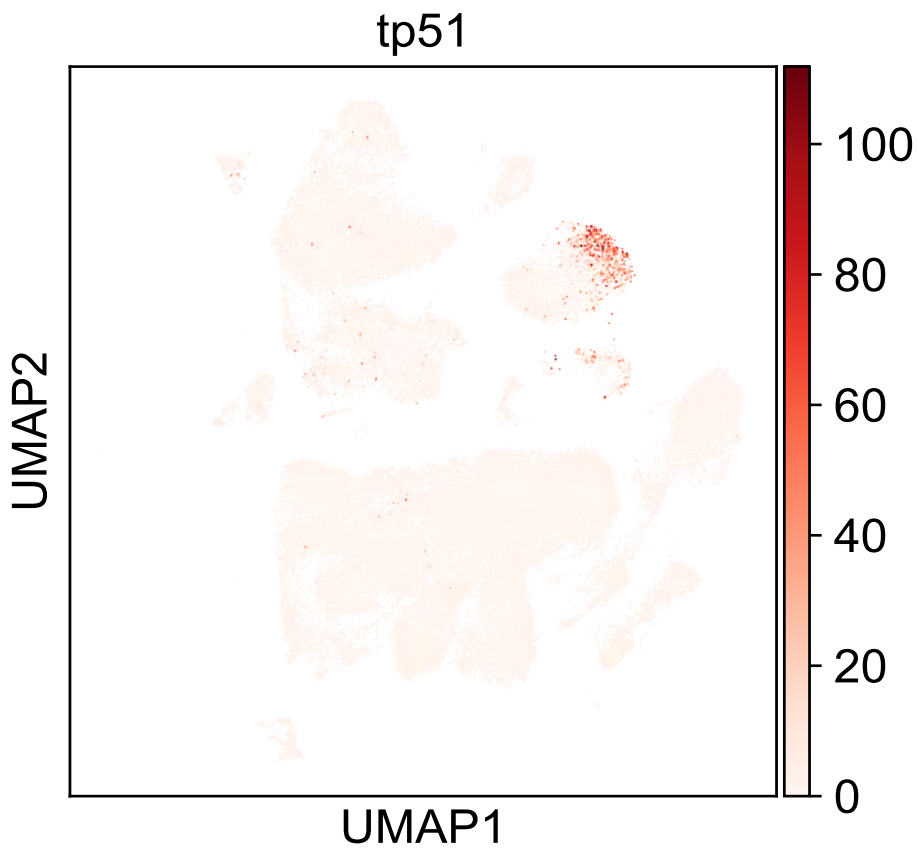


Cell Type

NK cells



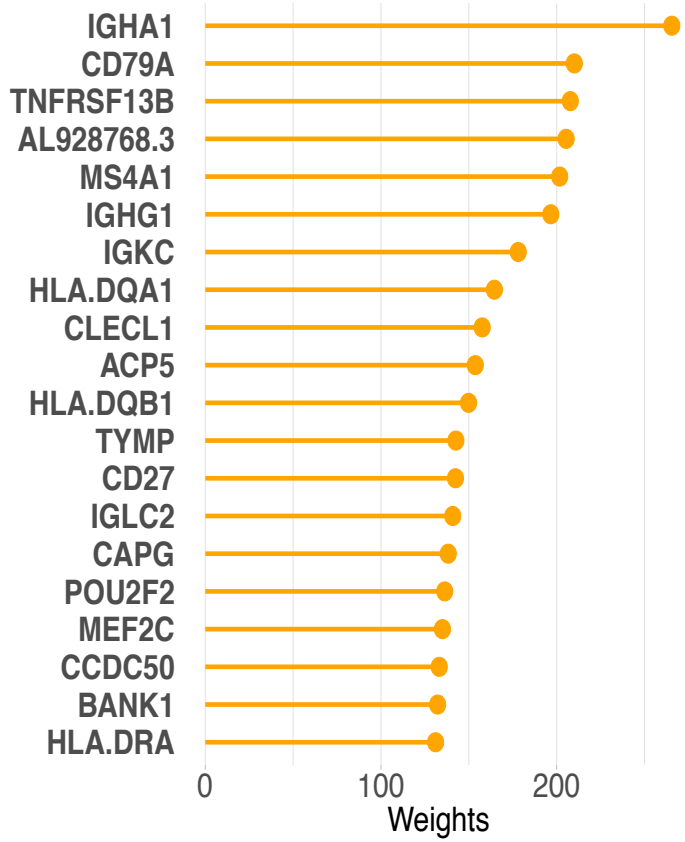
Imm GEM 51



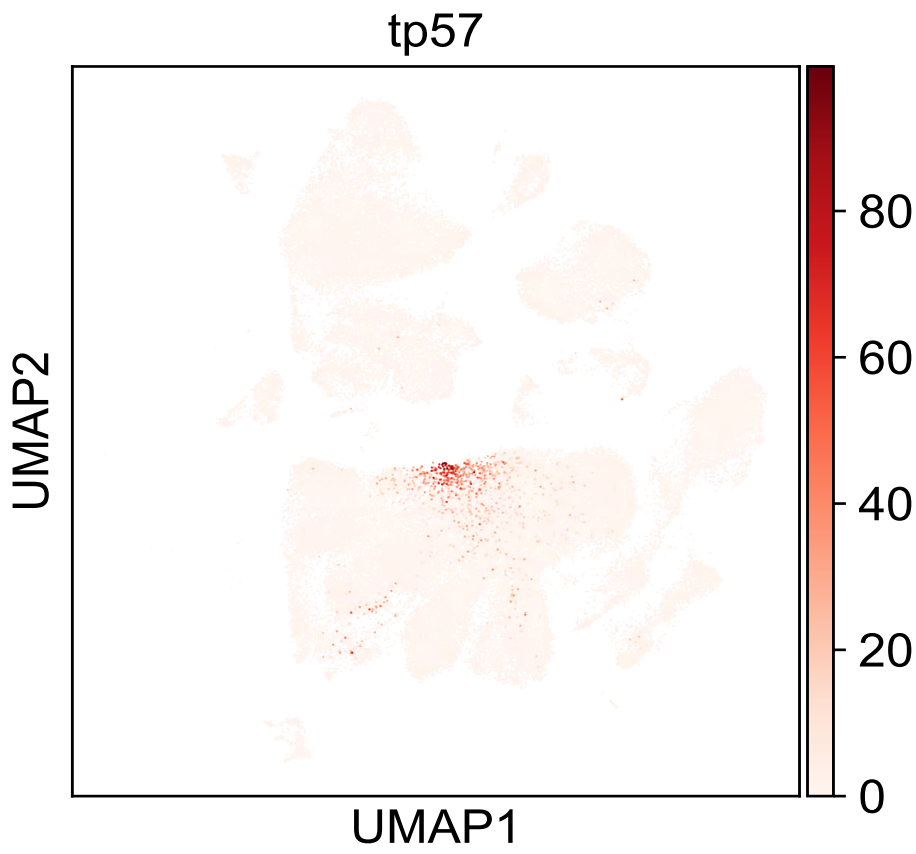
Cell Type

B cells

Top 20 genes



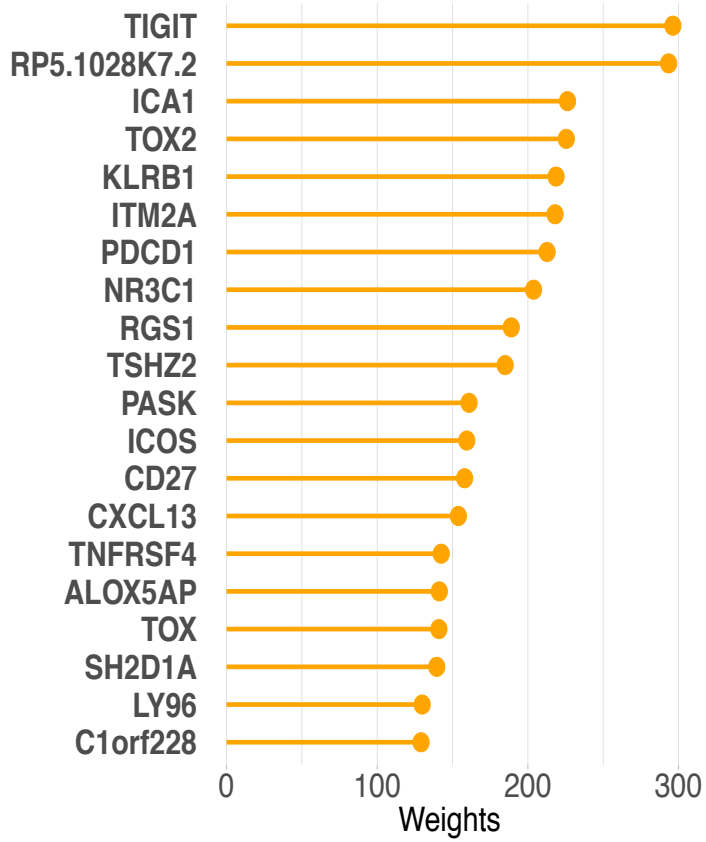
Imm GEM 57



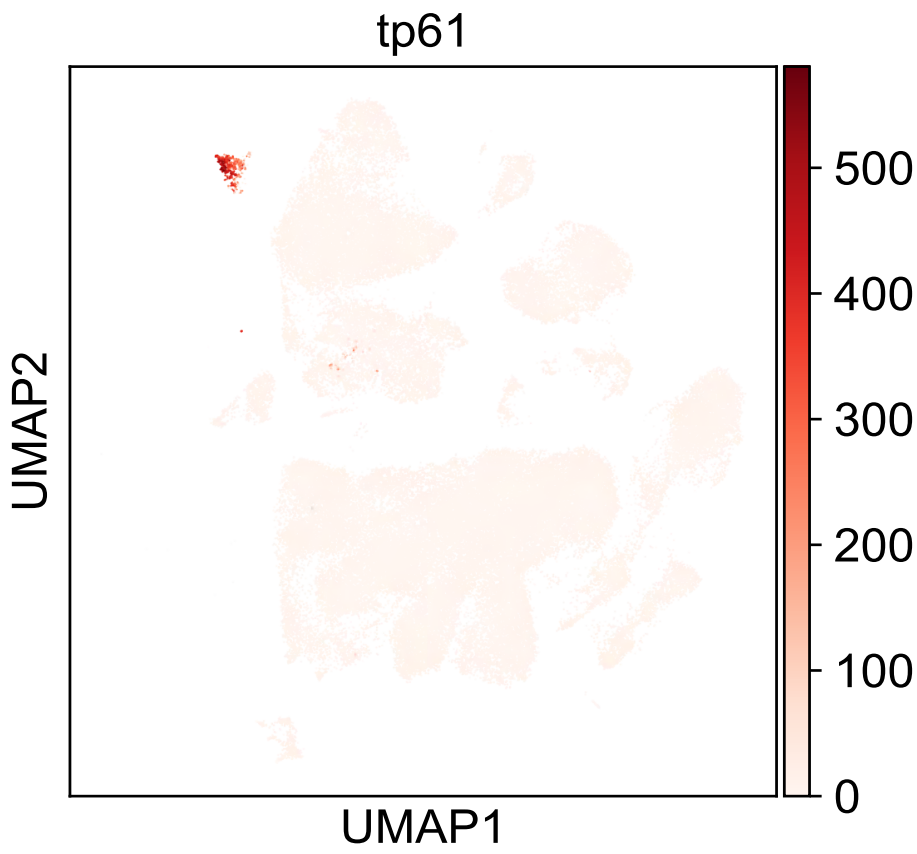
Cell Type

CD4 T cells

Top 20 genes

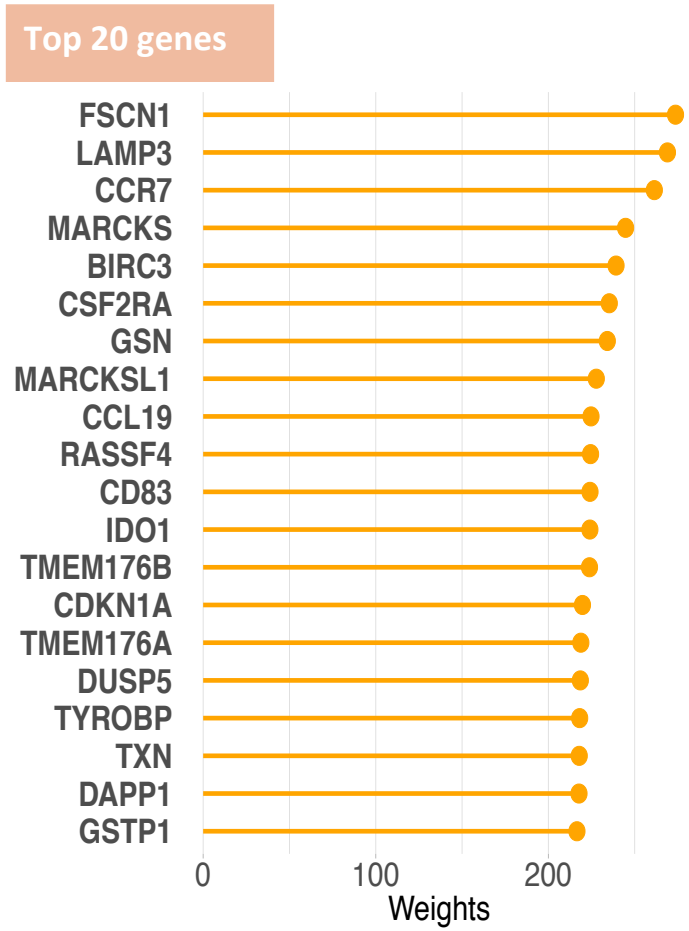


Imm GEM 61

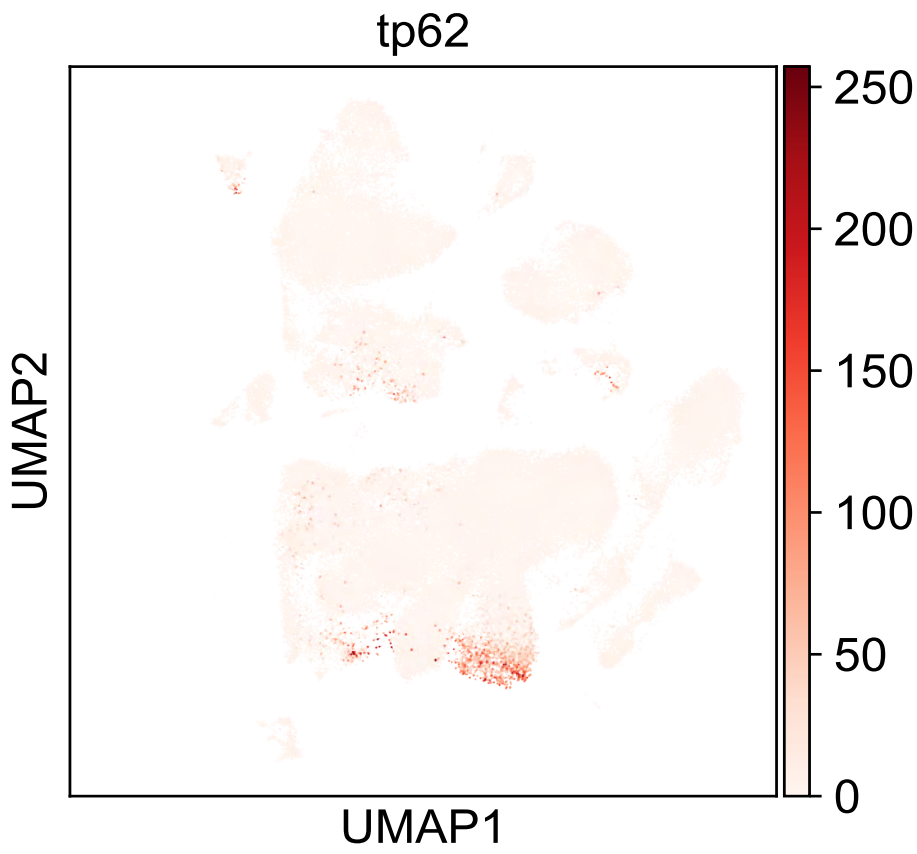


Cell Type

Dendritic



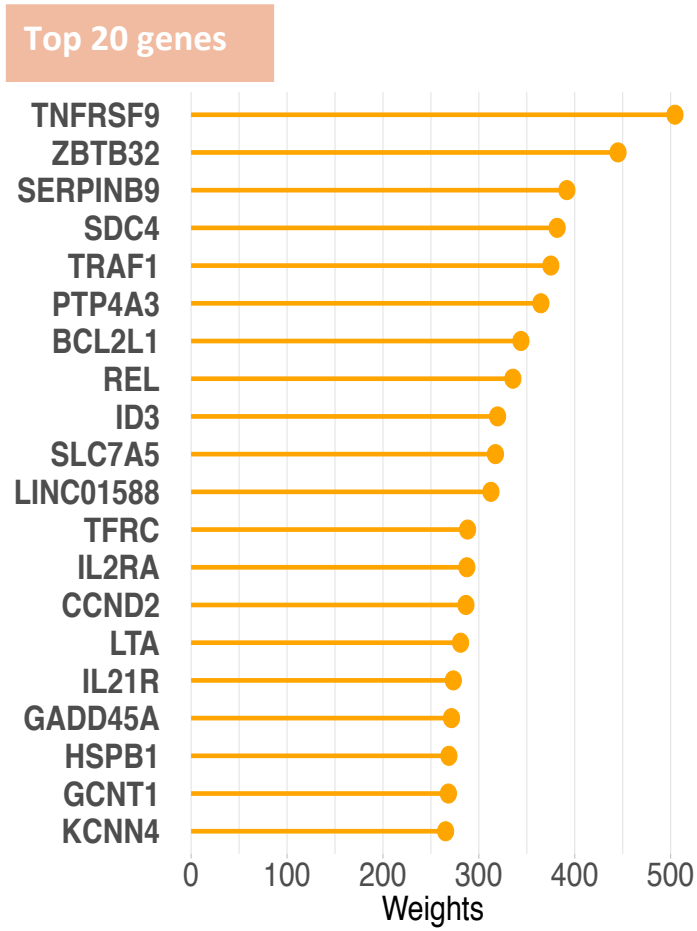
Imm GEM 62



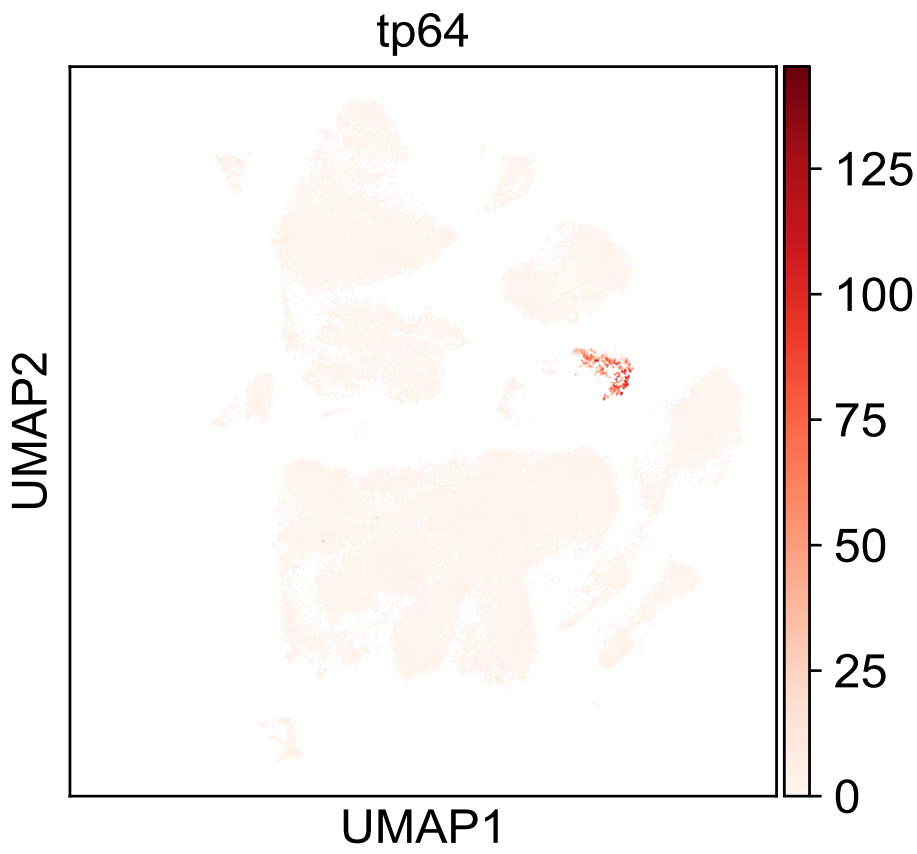
Cell Type

Tregs

CD8 T cells



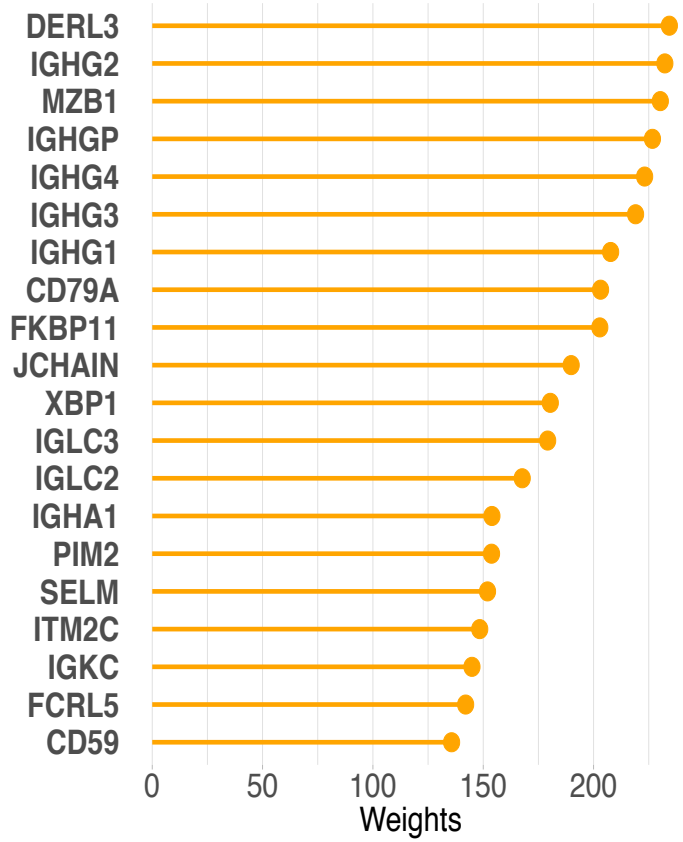
Imm GEM 64



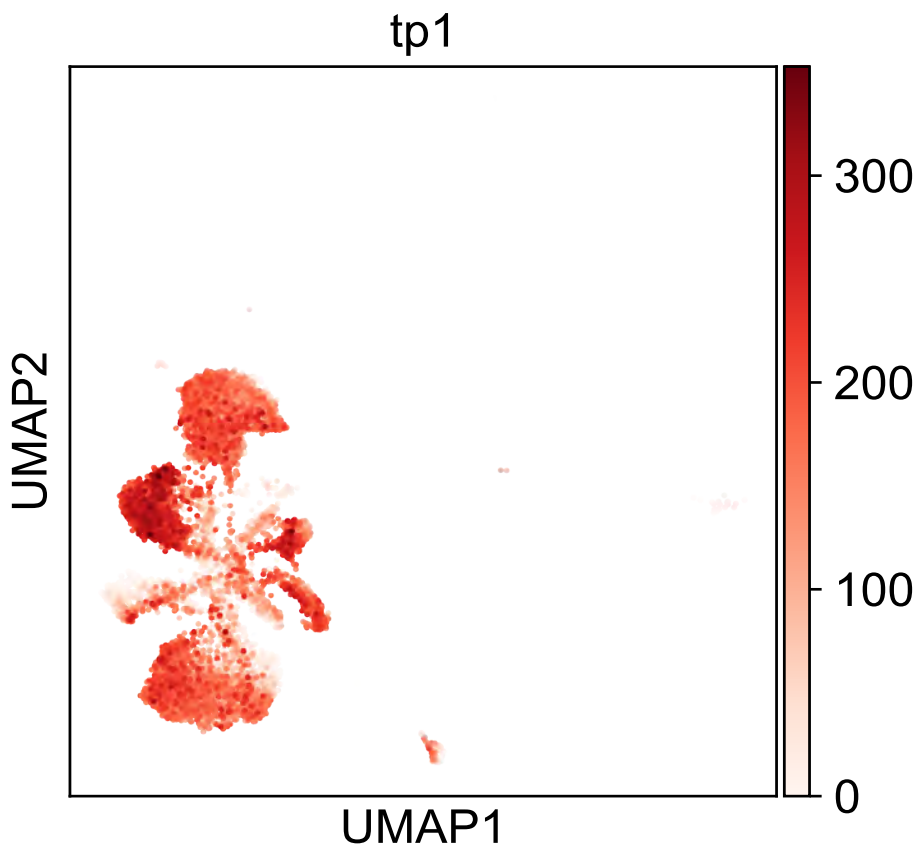
Cell Type

B cells

Top 20 genes

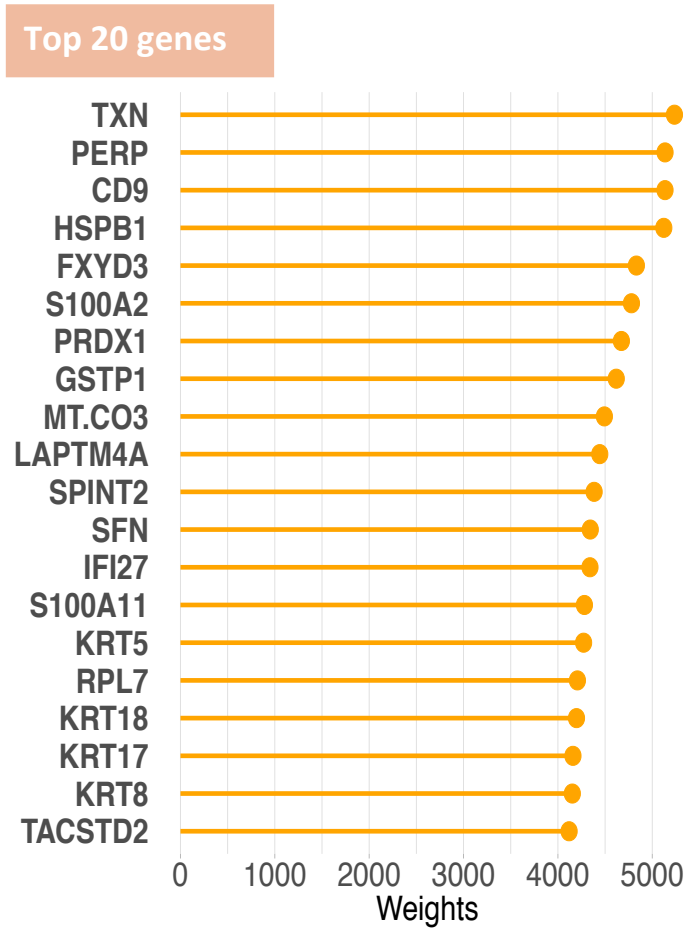


Epi GEM 1

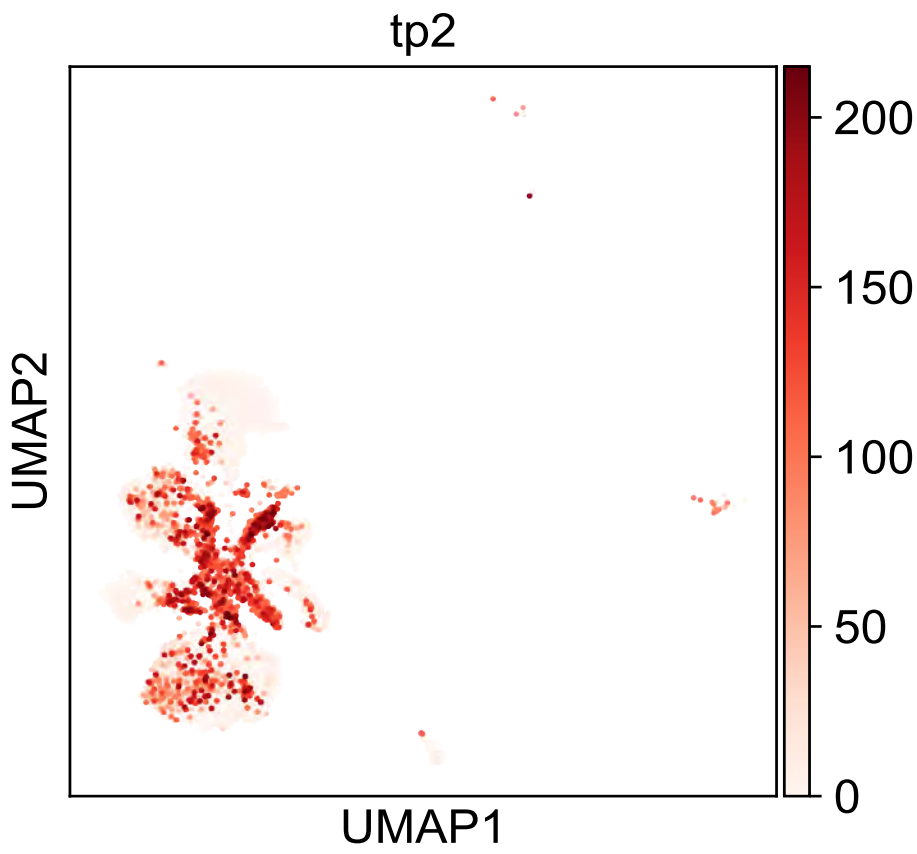


Cell Type

- Epi_2
- Epi_3
- Epi_4
- Epi_5
- Epi_6
- Epi_7
- Epi_8



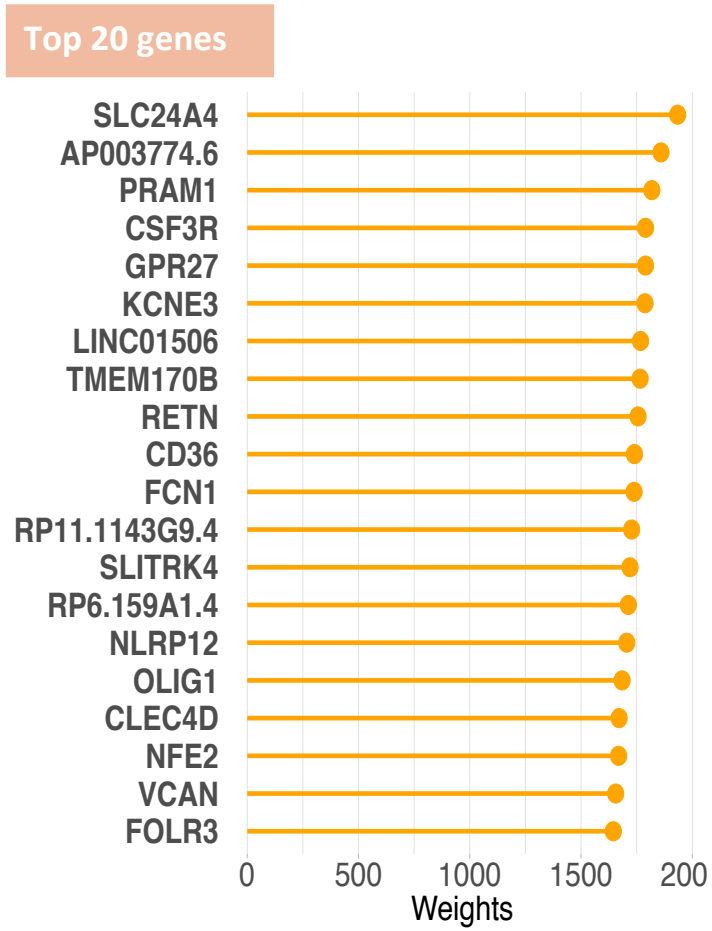
Epi GEM 2



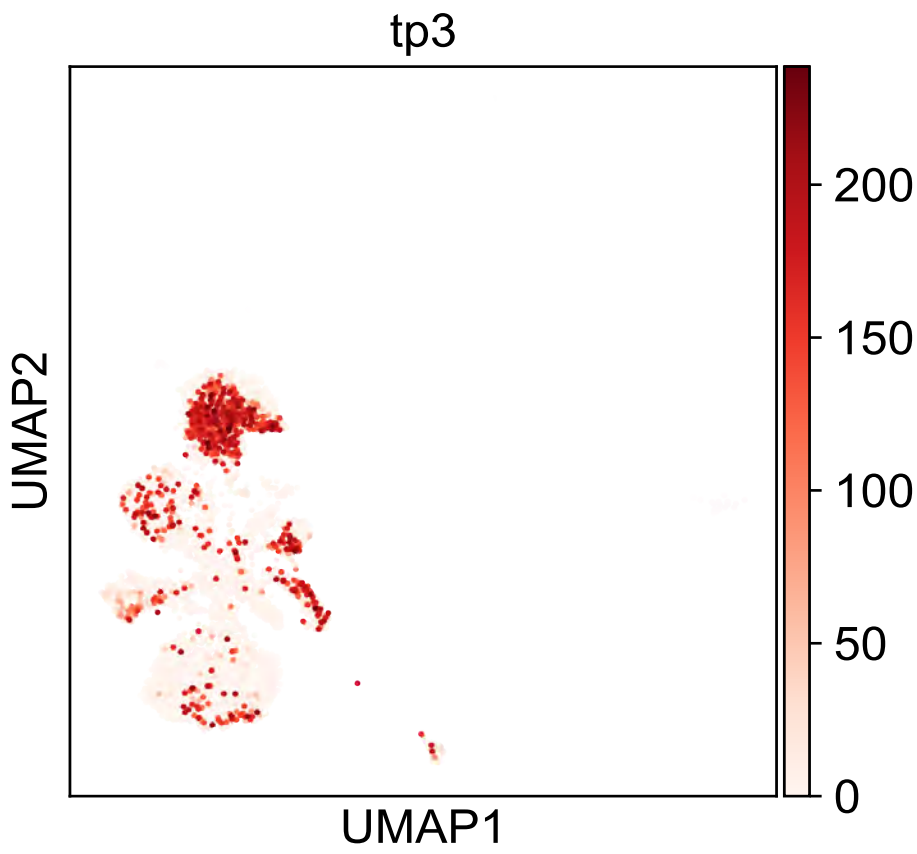
Cell Type

Epi_1

Epi_4



Epi GEM 3

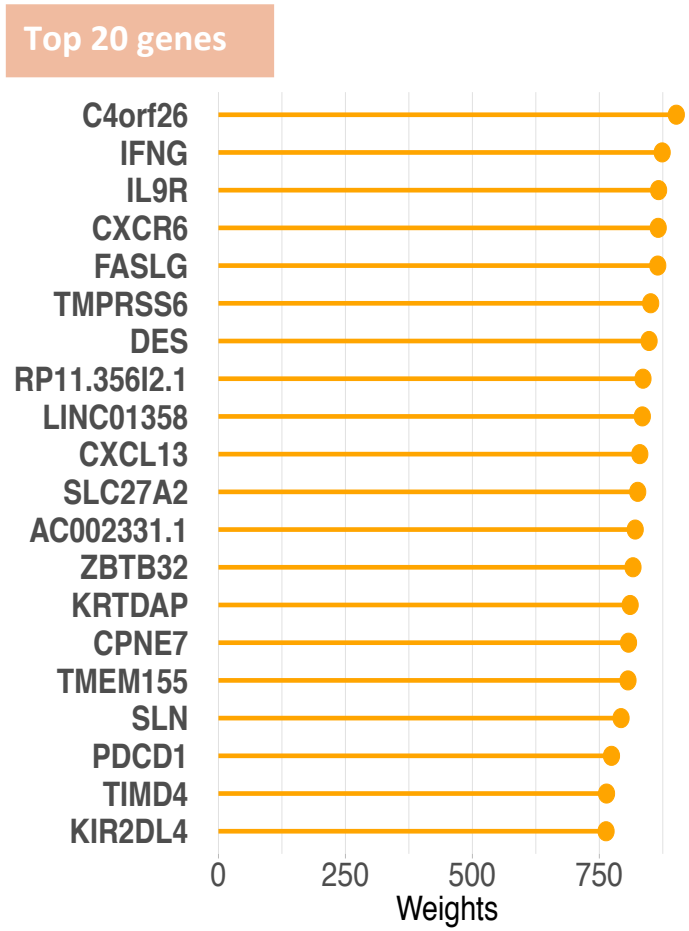


Cell Type

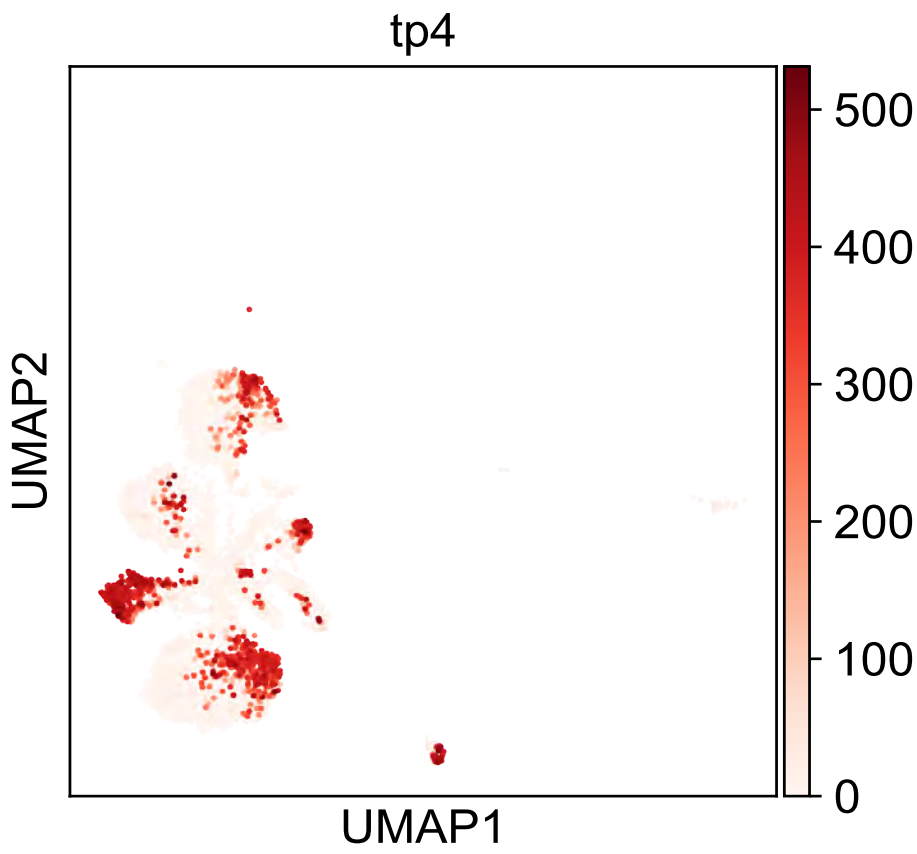
Epi_1

Epi_2

Epi_3



Epi GEM 4

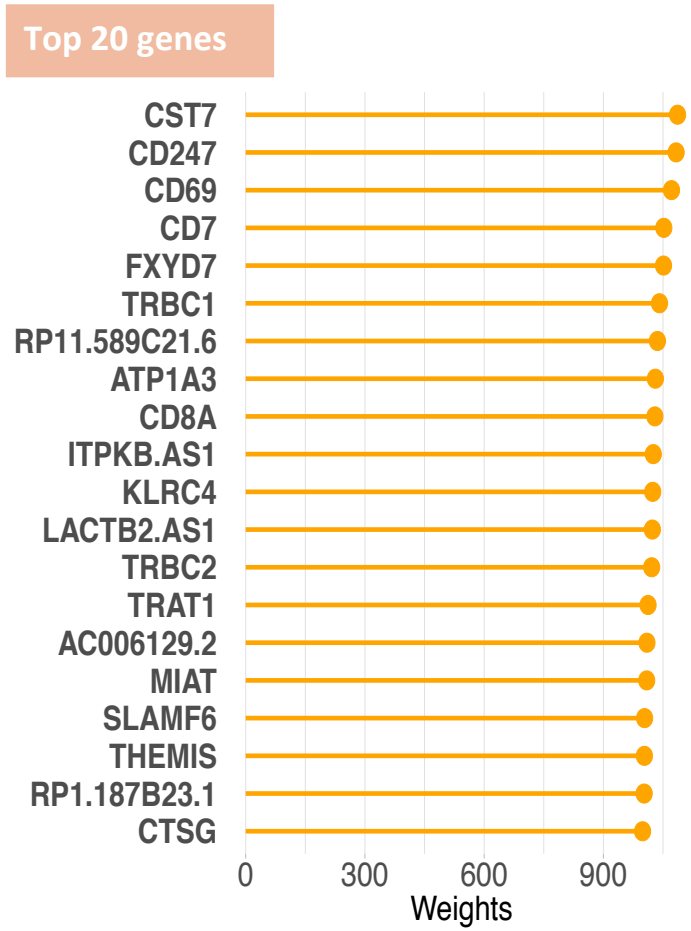


Cell Type

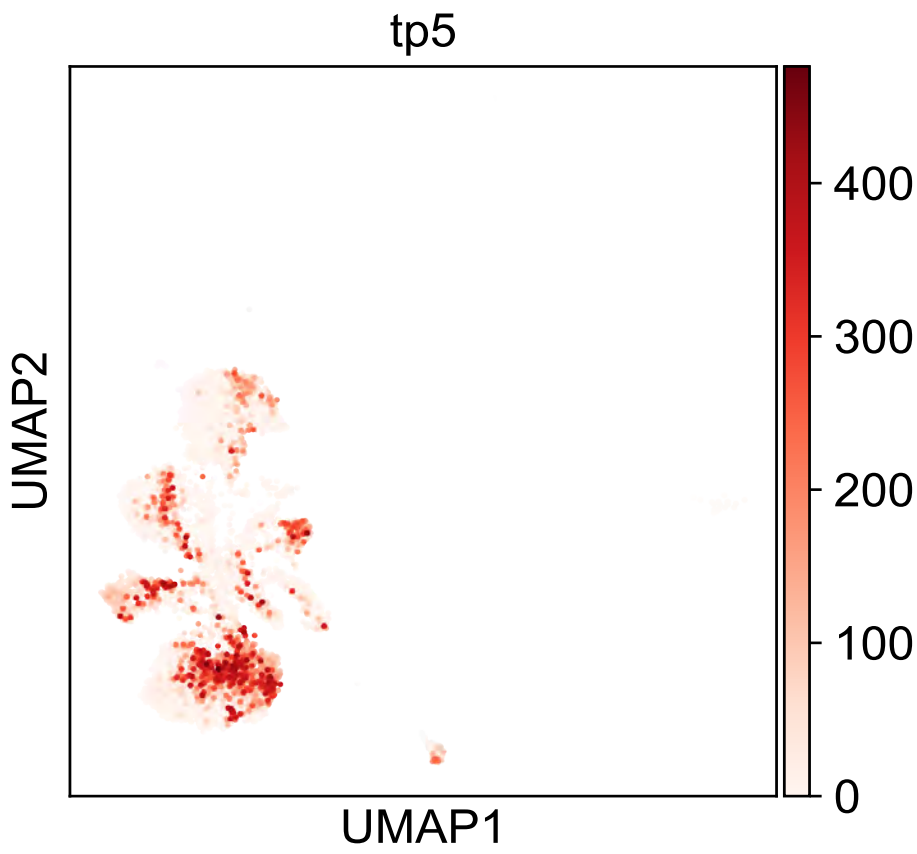
Epi_5

Epi_2

Epi_8

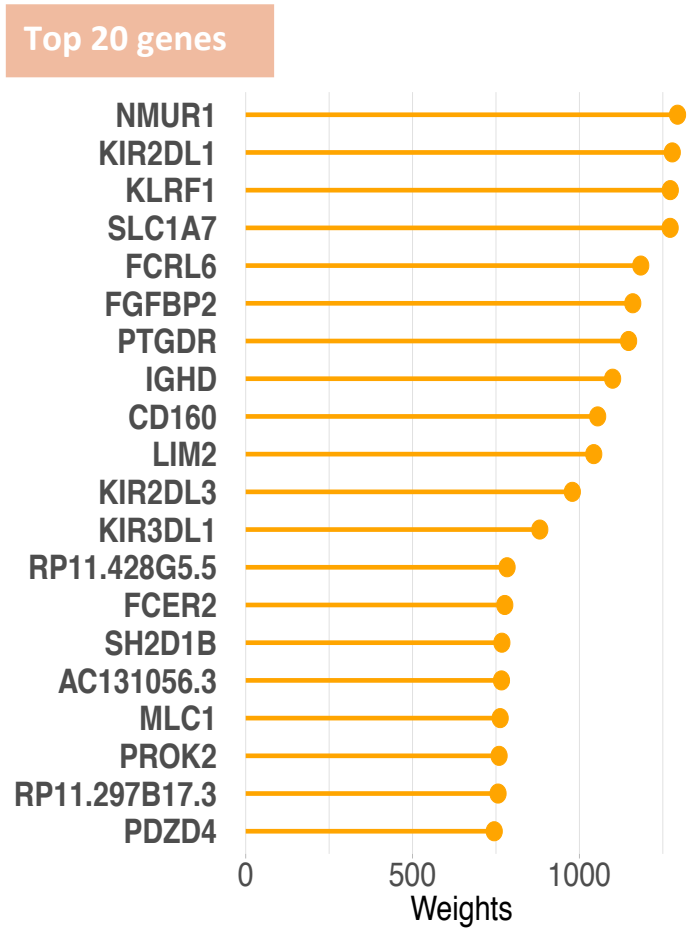


Epi GEM 5

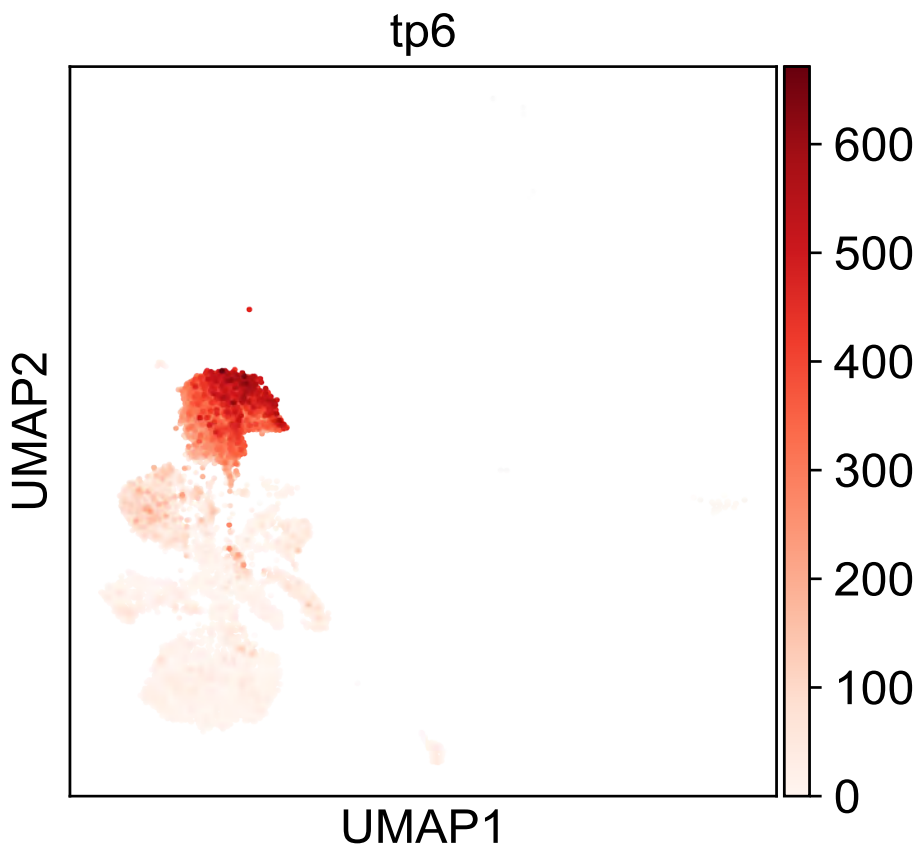


Cell Type

- Epi_1
- Epi_2
- Epi_3



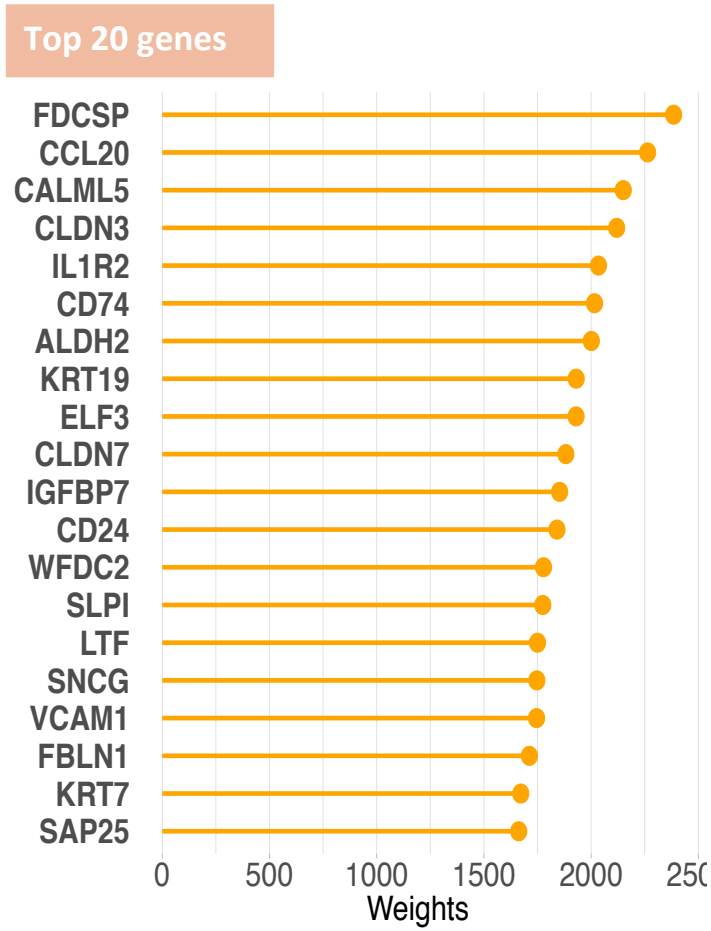
Epi GEM 6



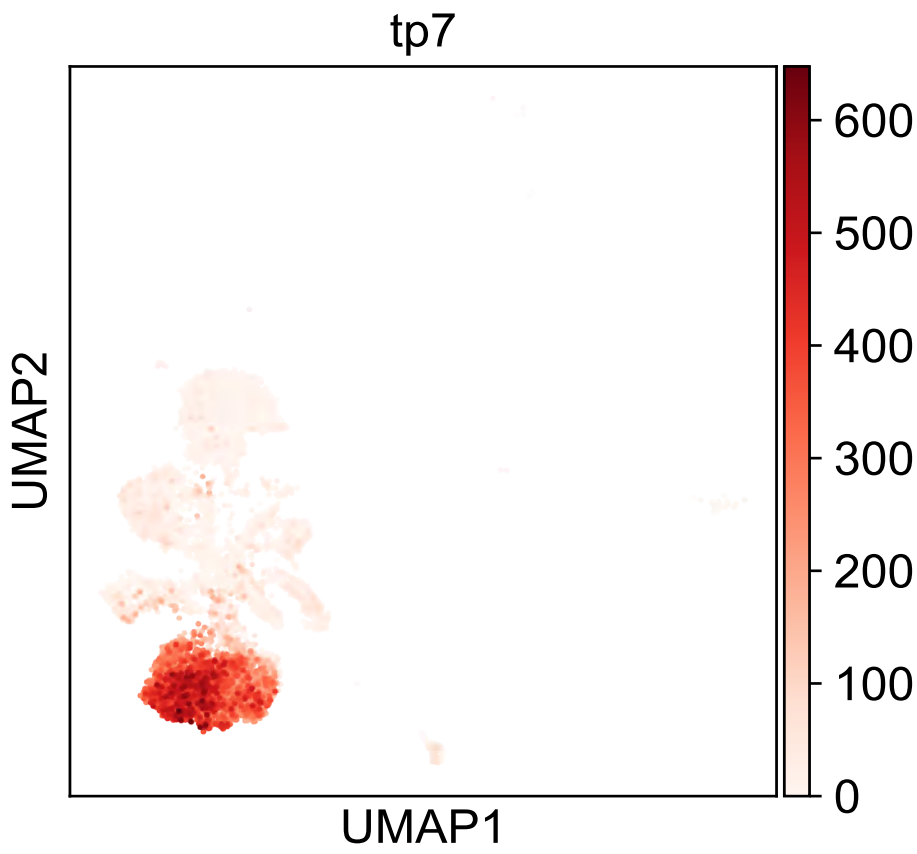
Cell Type

Epi_7

Epi_8



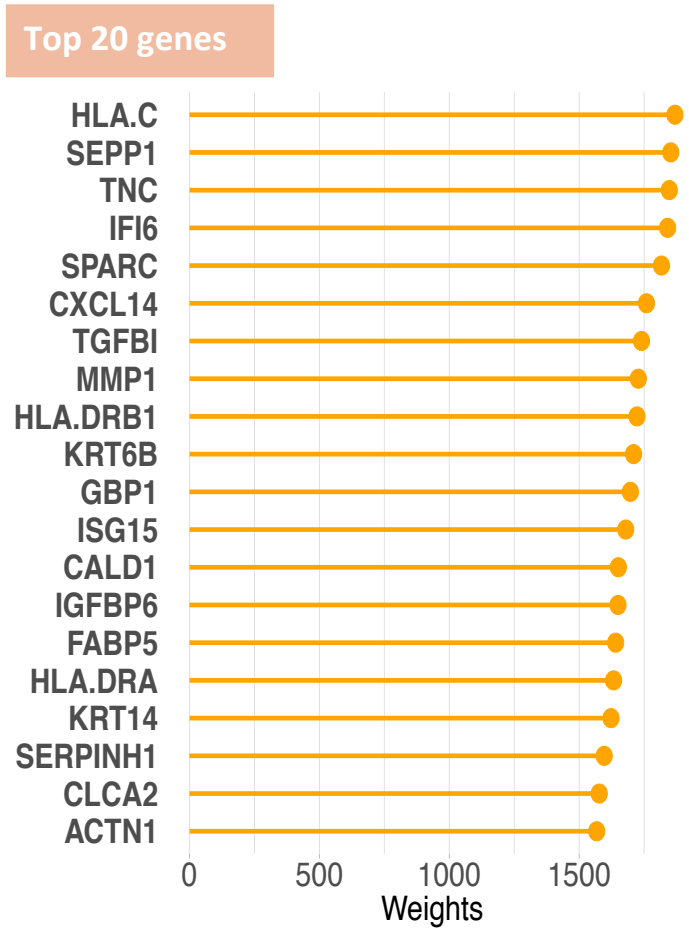
Epi GEM 7



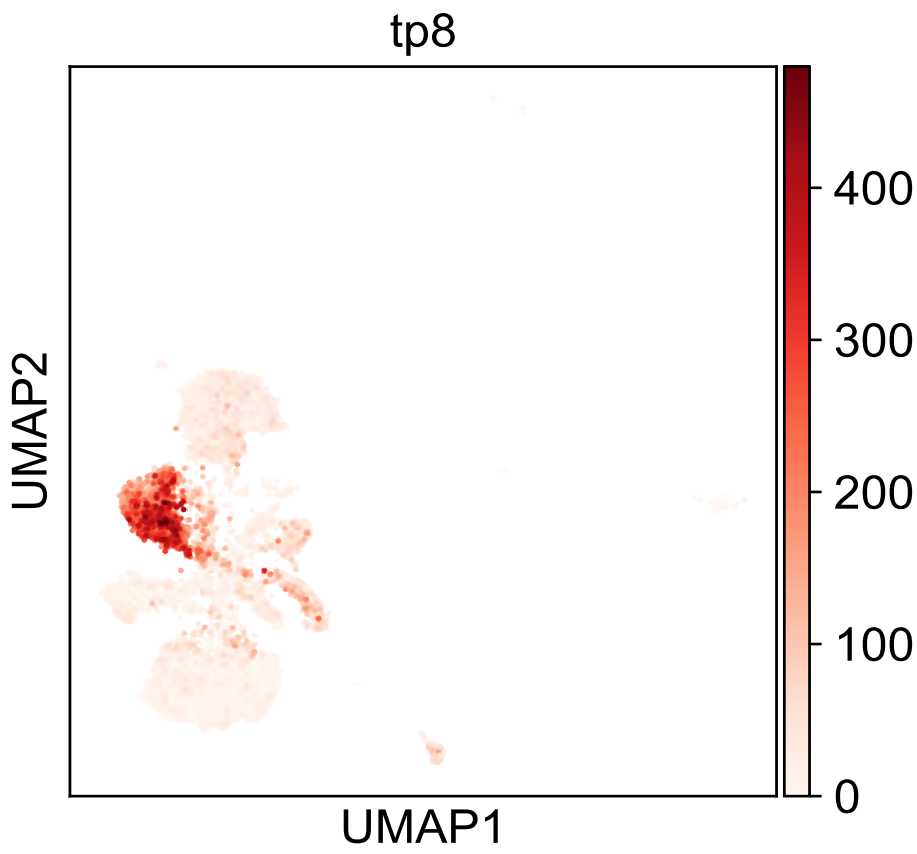
Cell Type

Epi_4

Epi_5

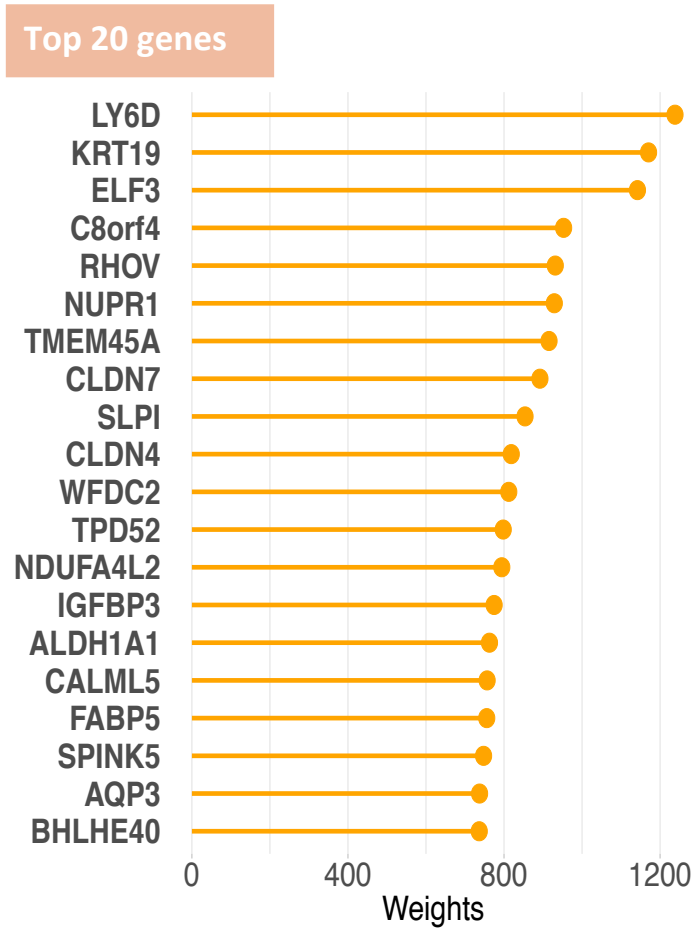


Epi GEM 8

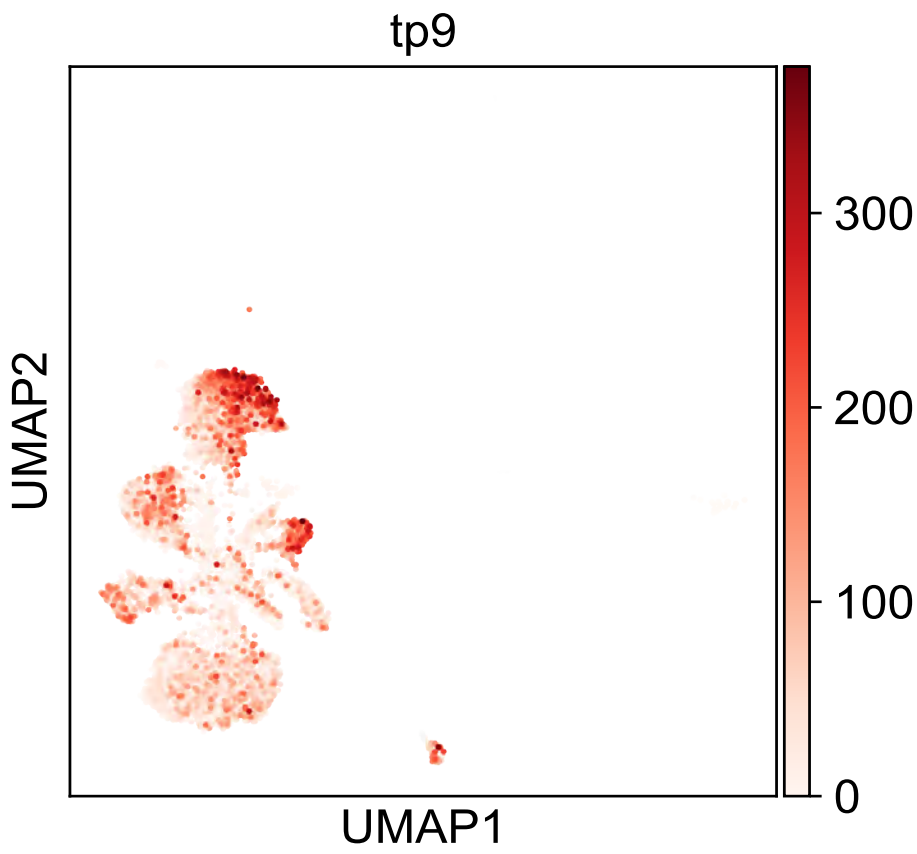


Cell Type

Epi_6



Epi GEM 9

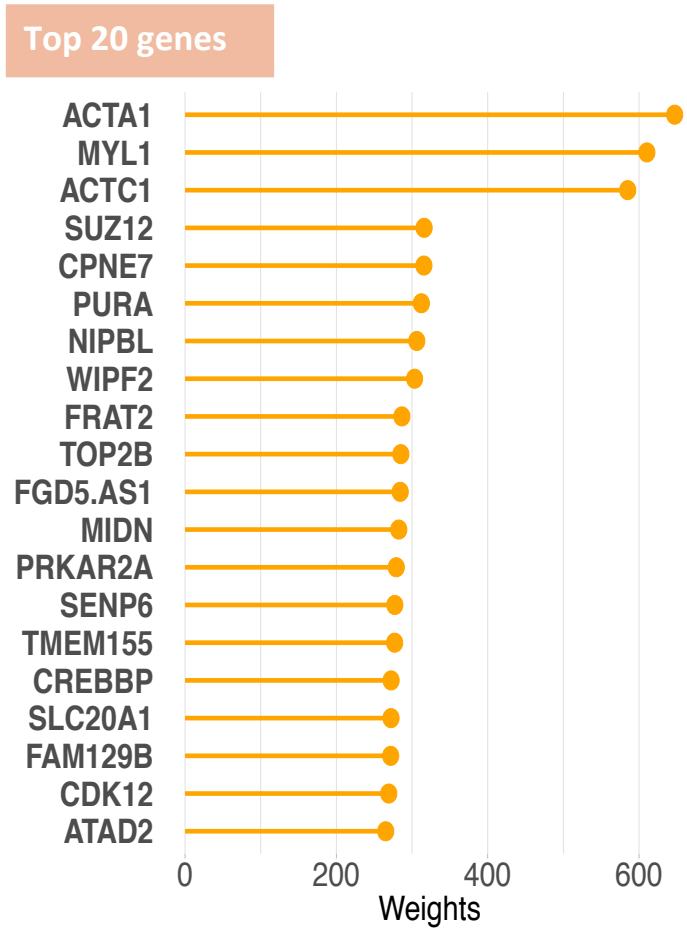


Cell Type

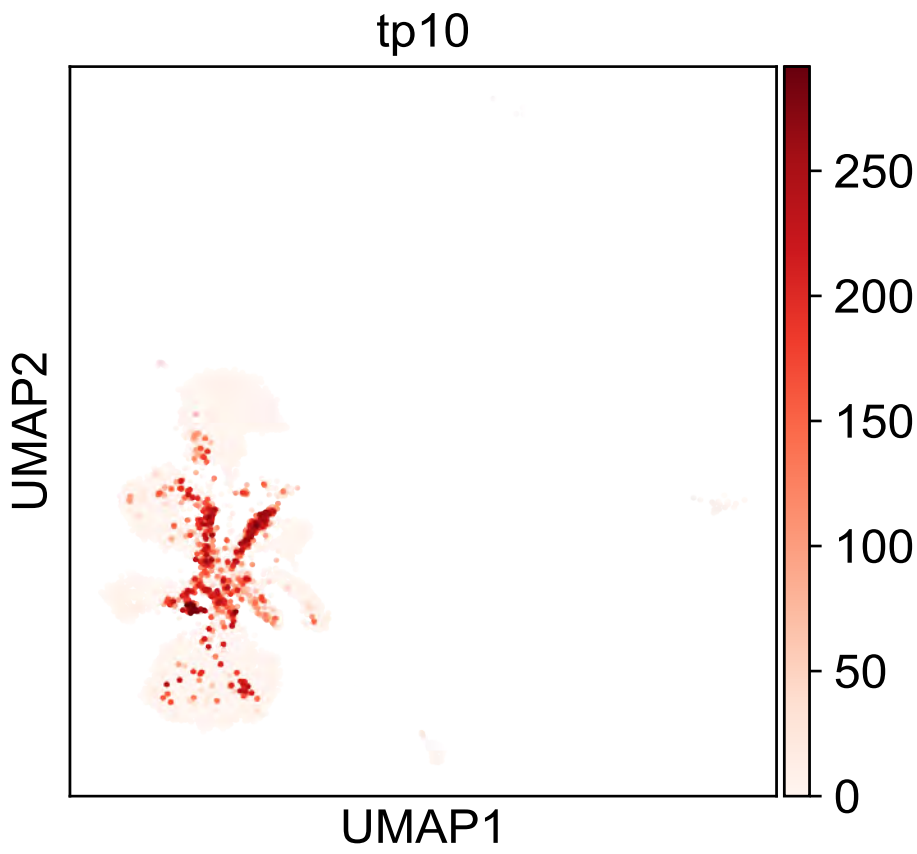
Epi_8

Epi_2

Epi_6

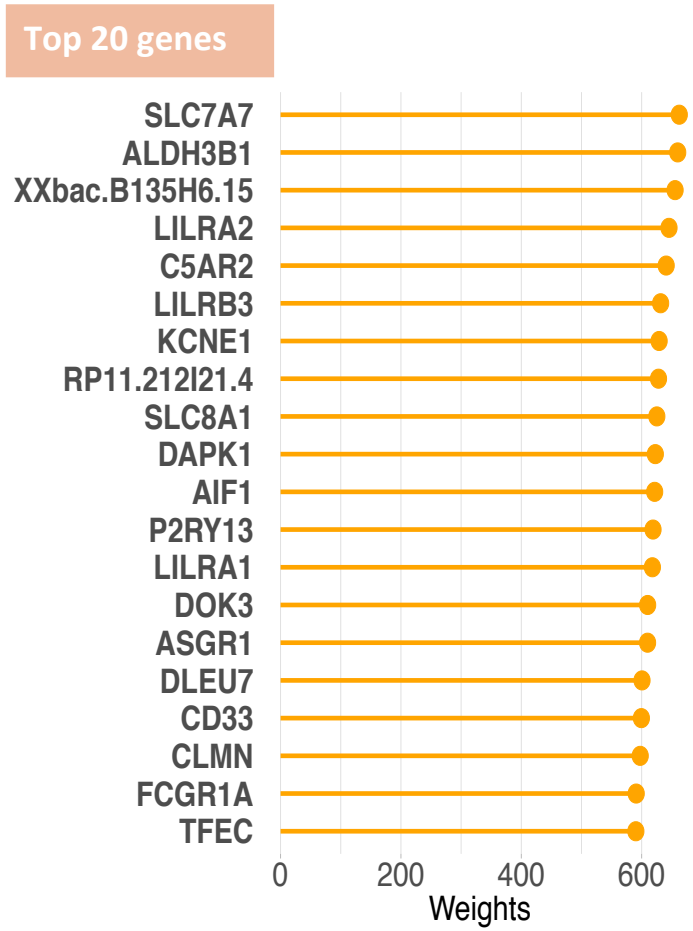


Epi GEM 10

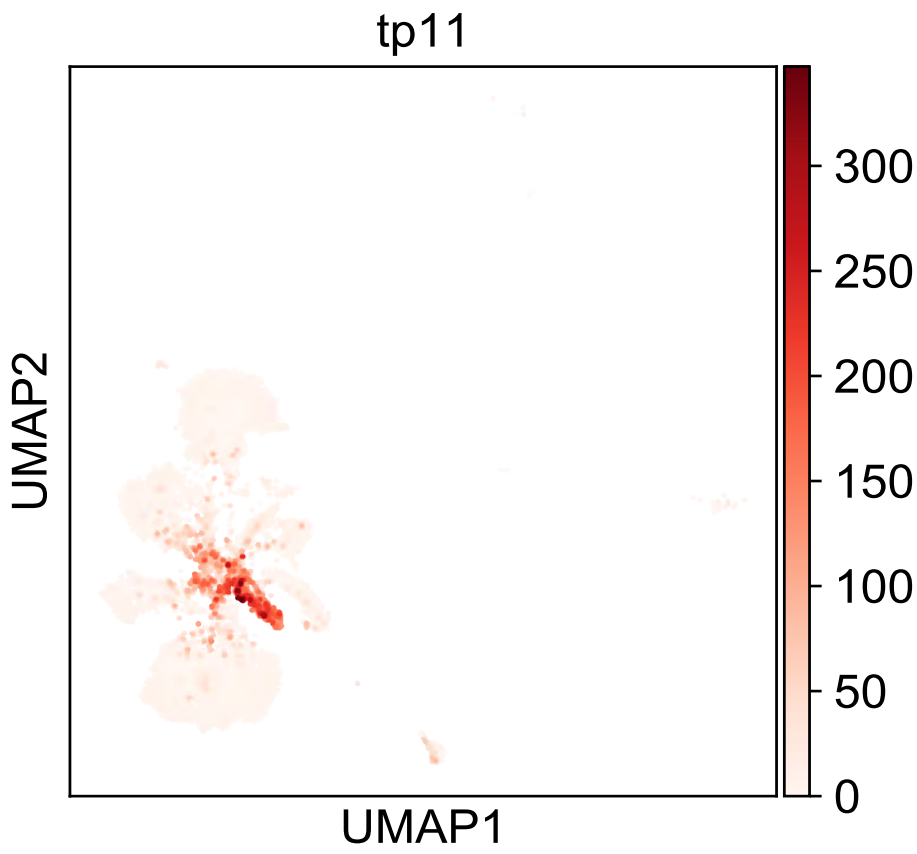


Cell Type

Epi_1

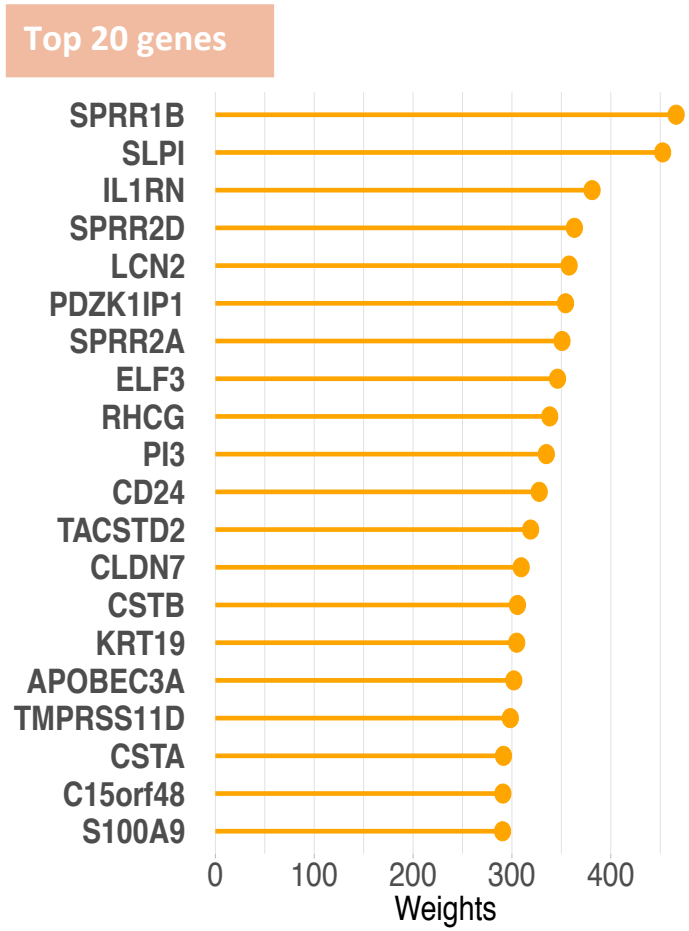


Epi GEM 11

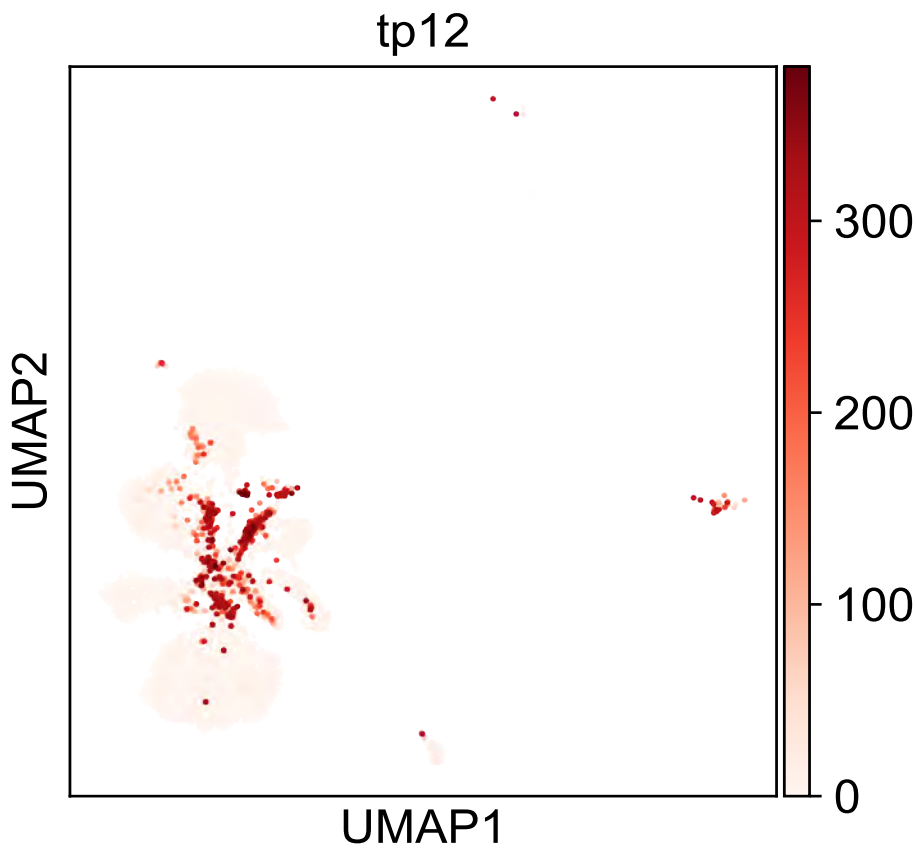


Cell Type

Epi_1

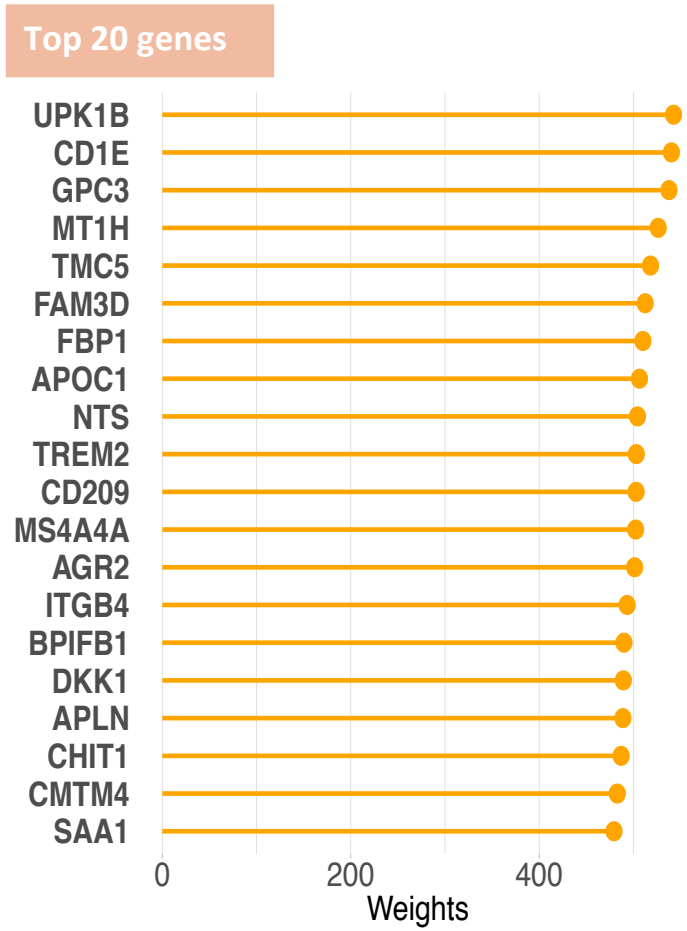


Epi GEM 12

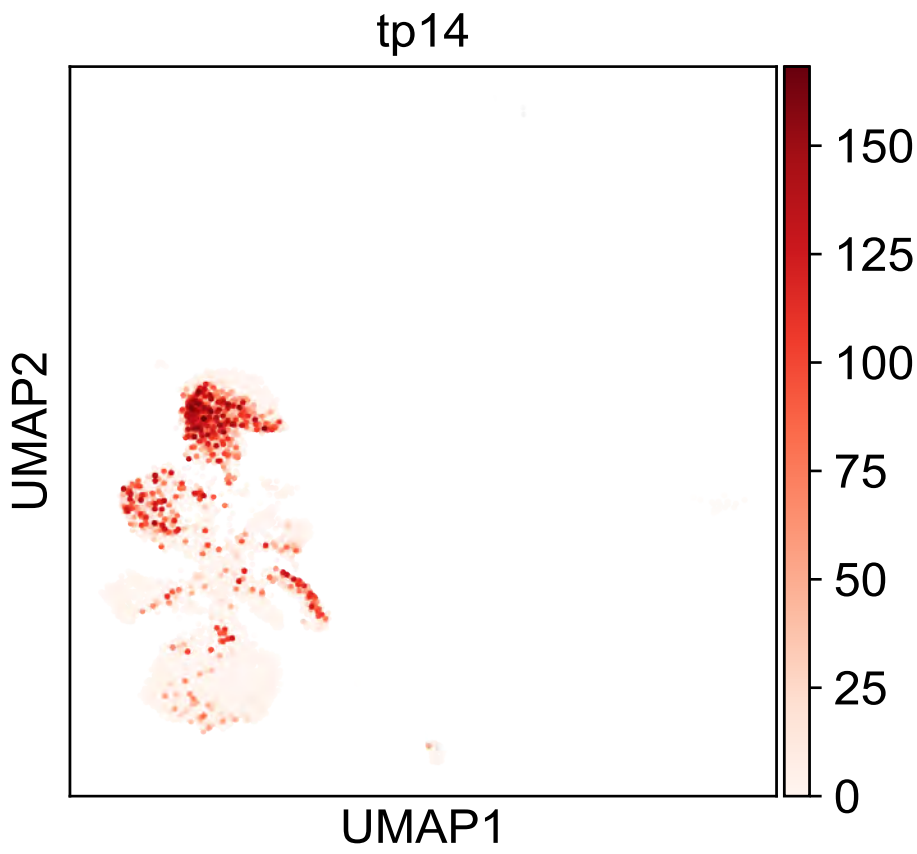


Cell Type

Epi_1



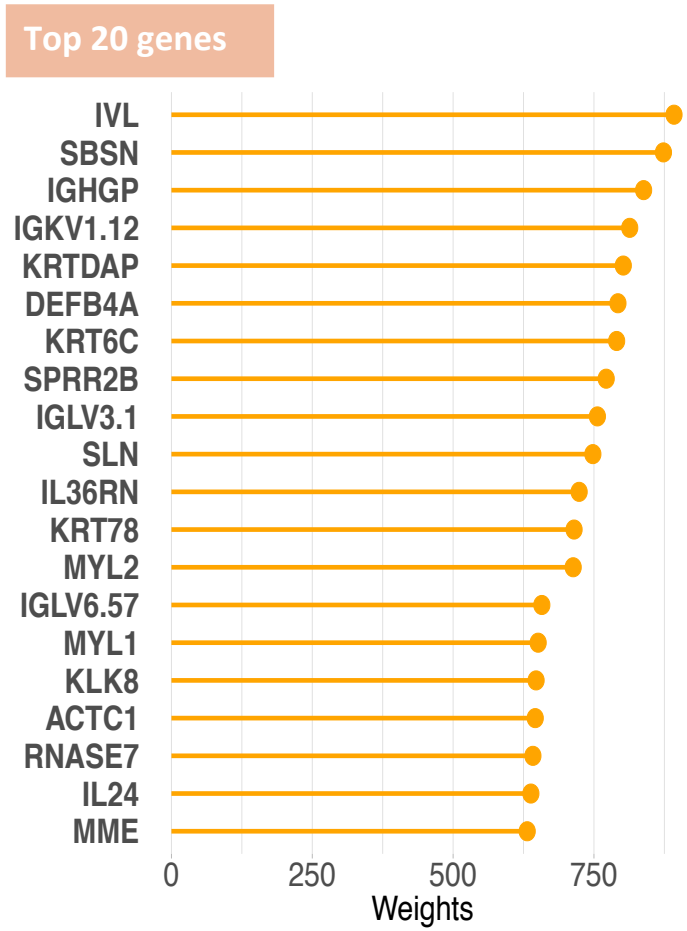
Epi GEM 14



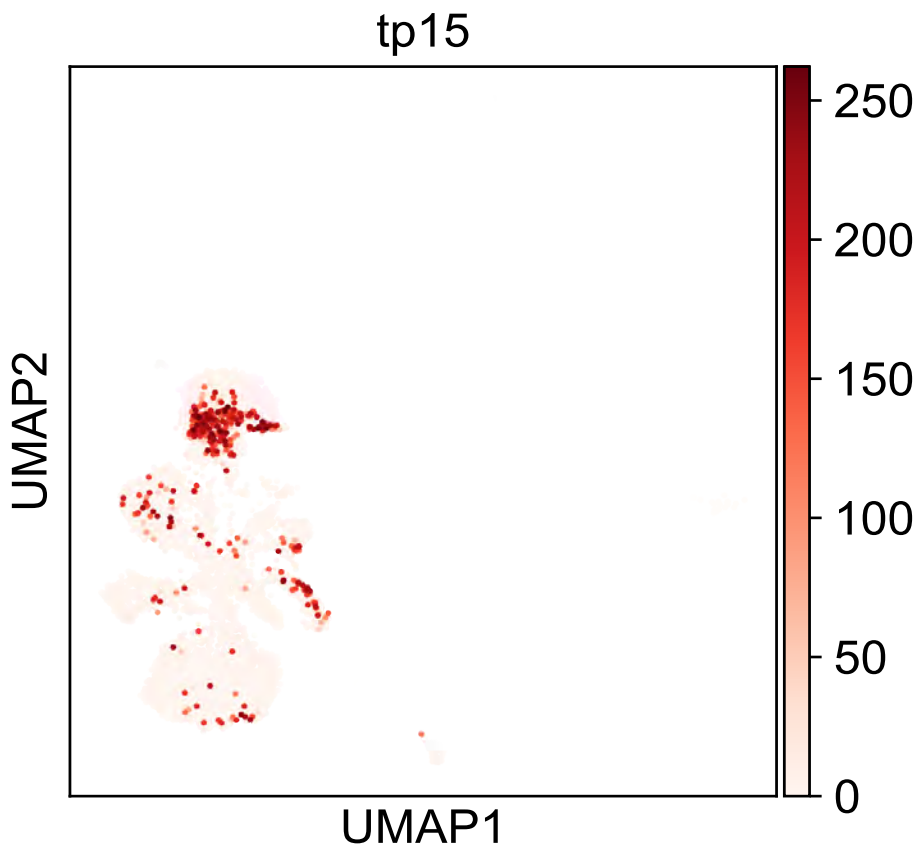
Cell Type

Epi_3

Epi_7



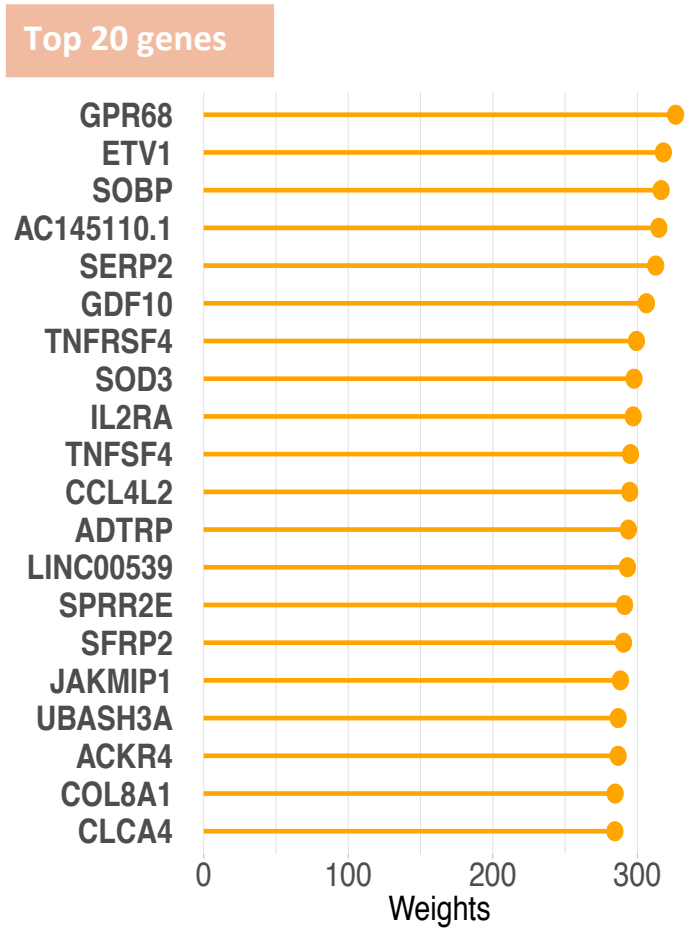
Epi GEM 15



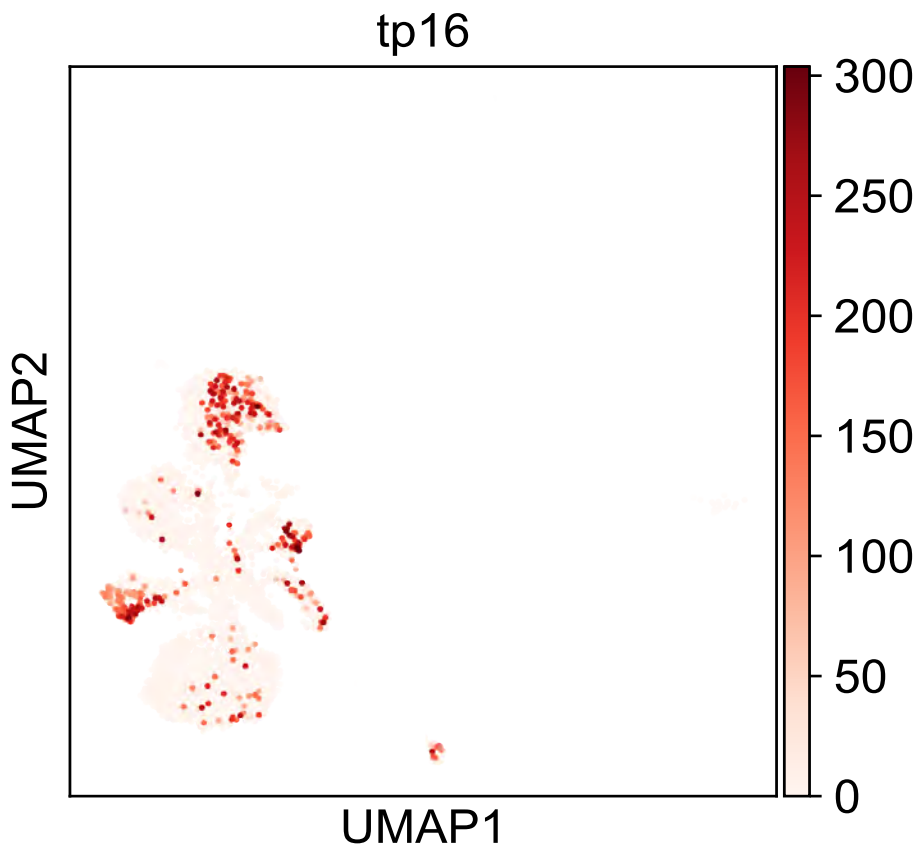
Cell Type

Epi_3

Epi_7



Epi GEM 16

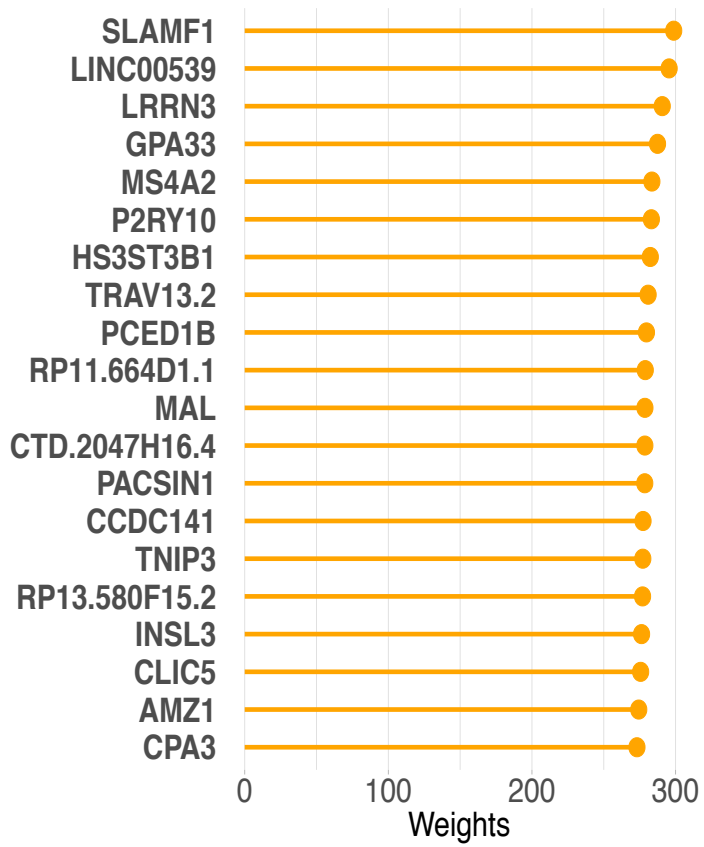


Cell Type

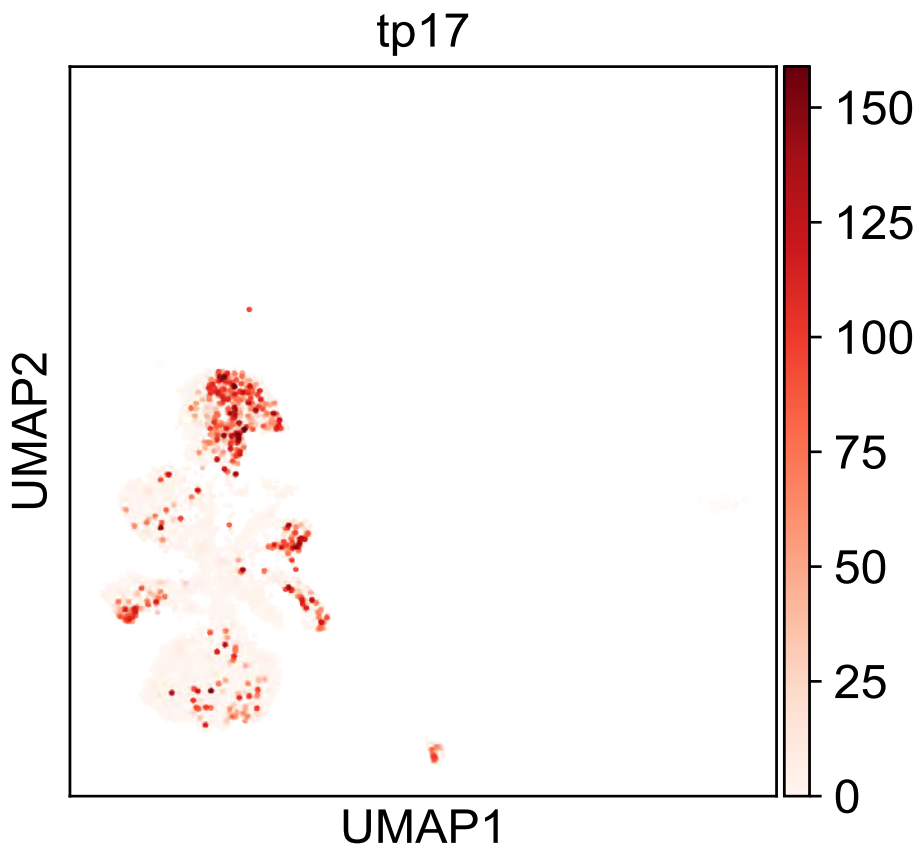
Epi_3

Epi_7

Top 20 genes



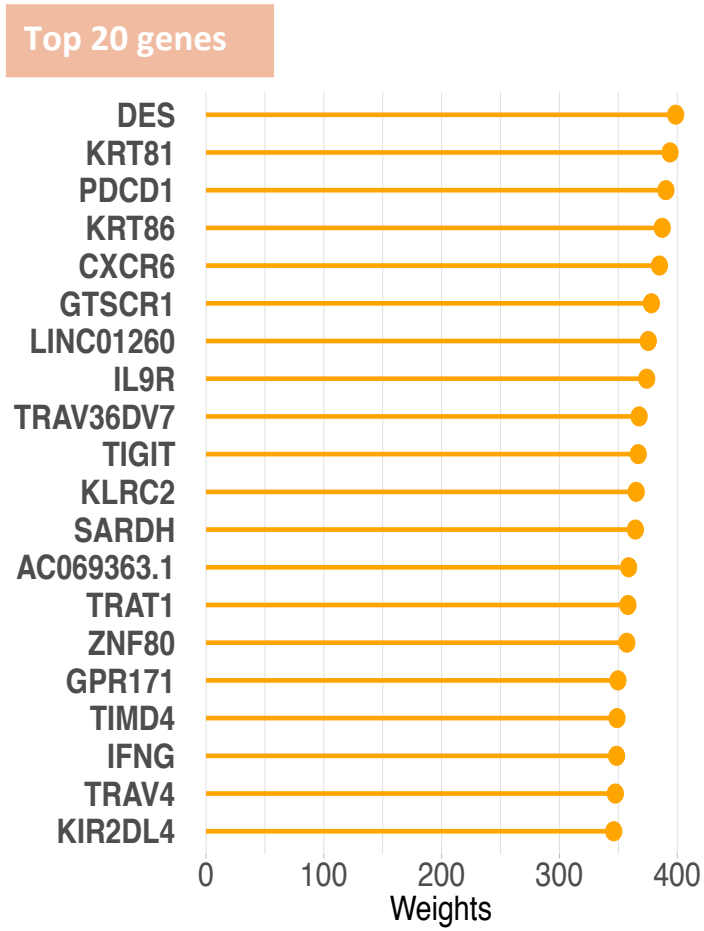
Epi GEM 17



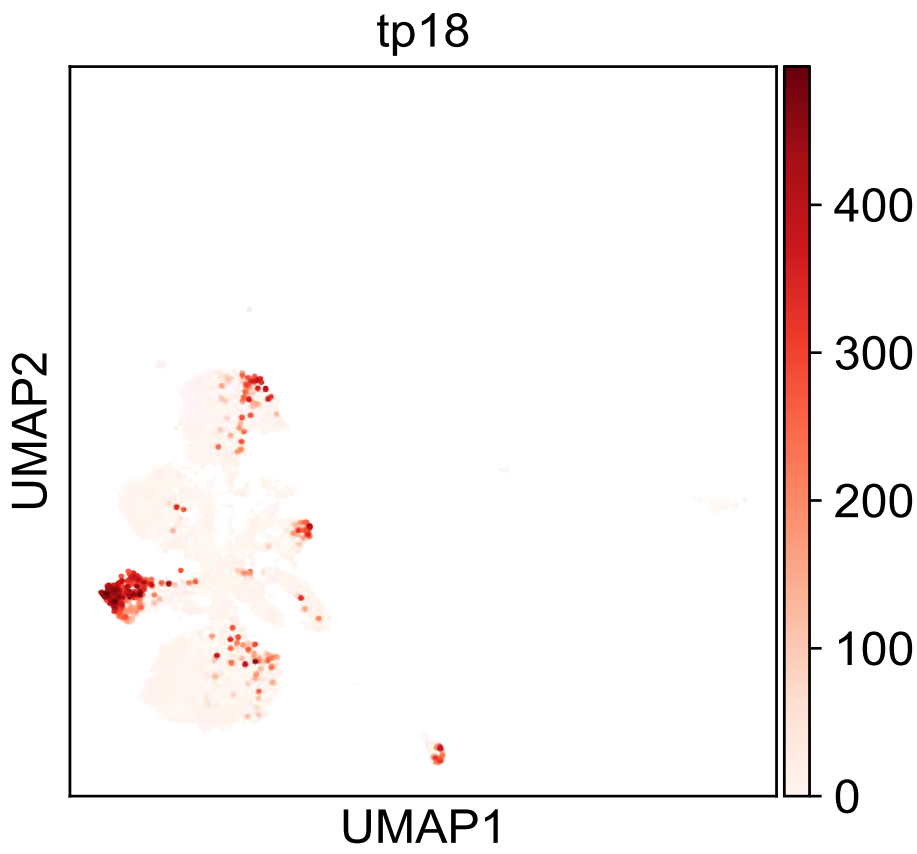
Cell Type

Epi_3

Epi_7

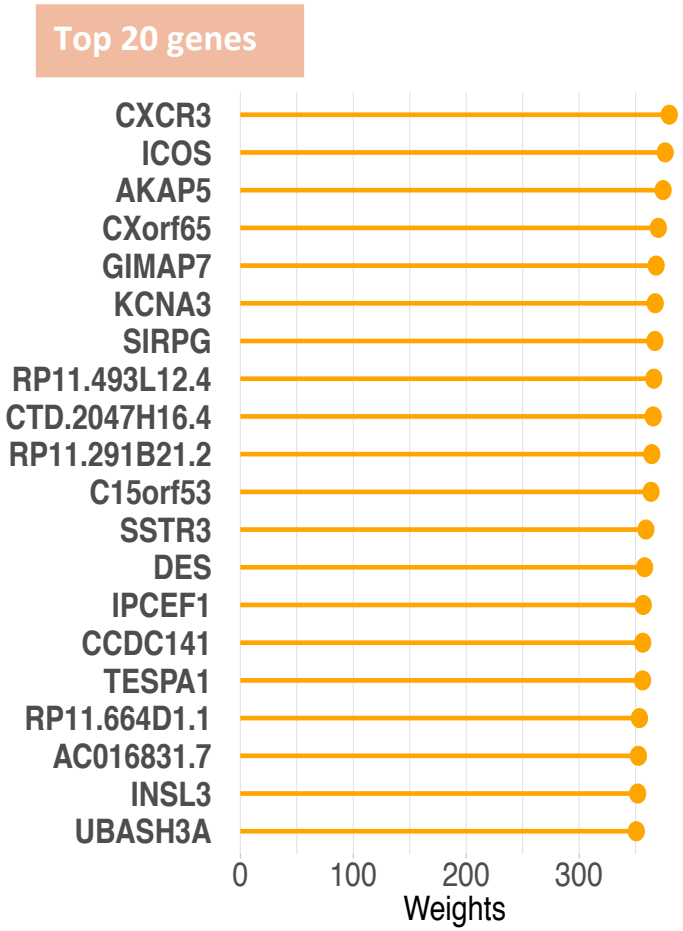


Epi GEM 18

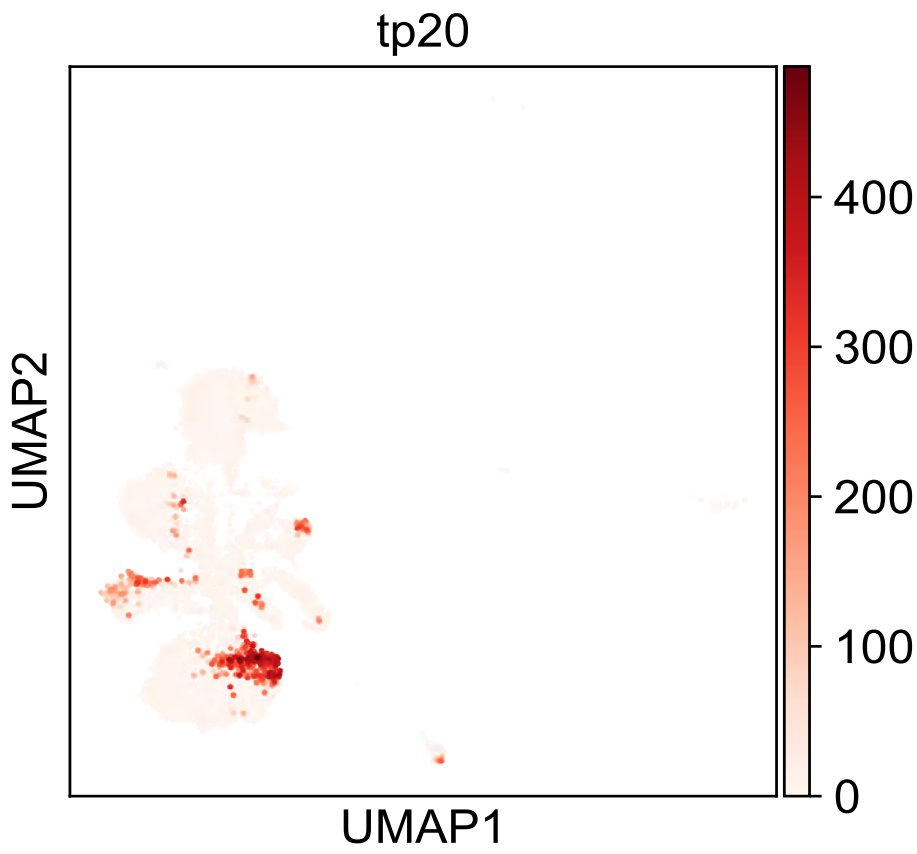


Cell Type

Epi_2



Epi GEM 20

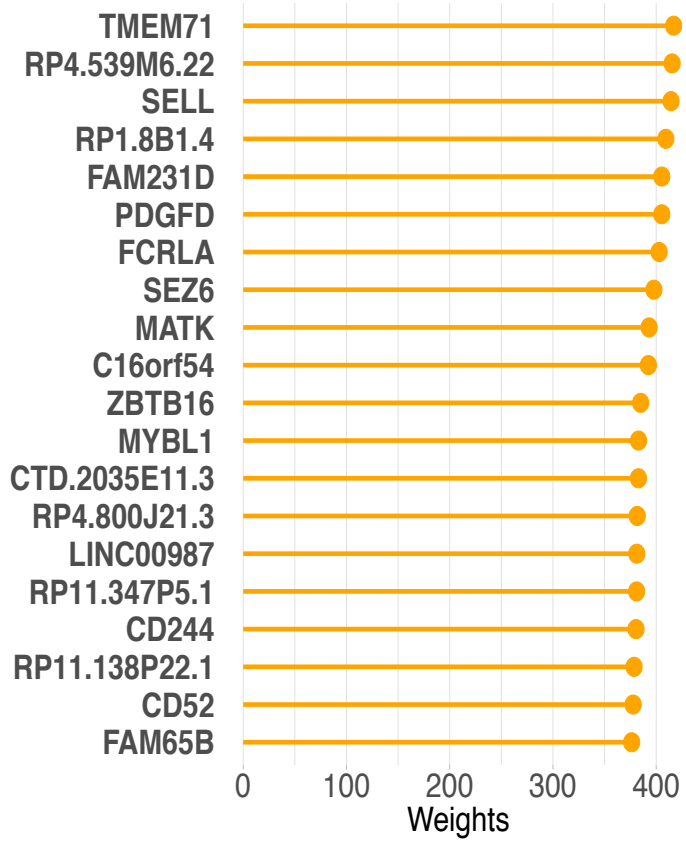


Cell Type

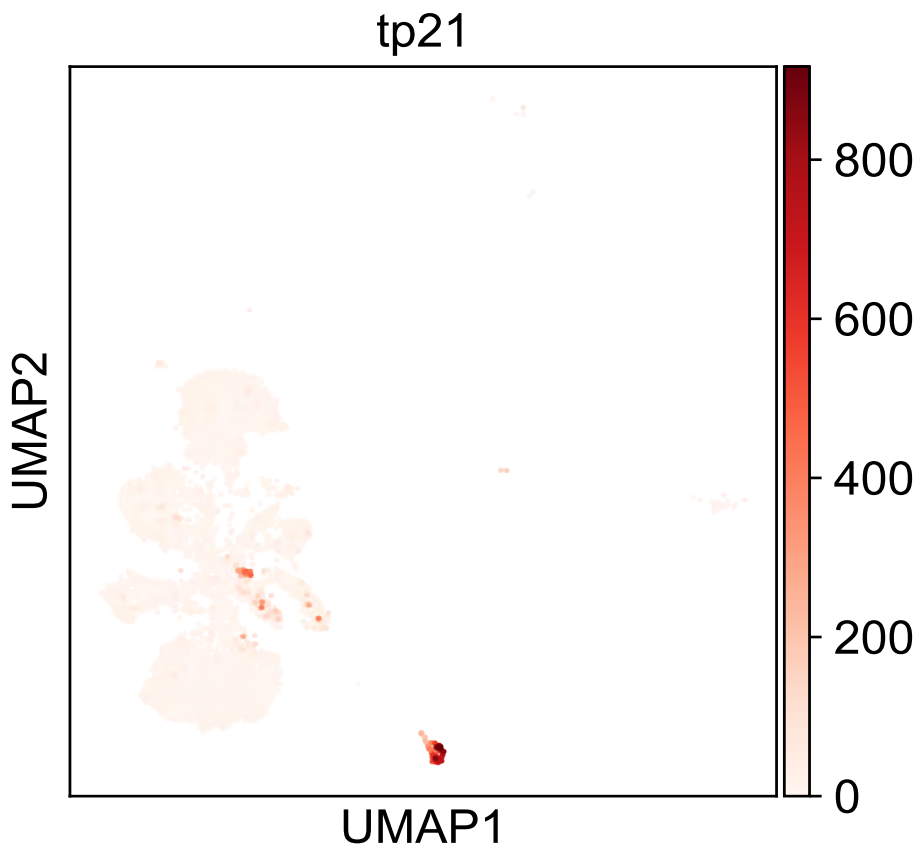
Epi_5

Epi_2

Top 20 genes



Epi GEM 21

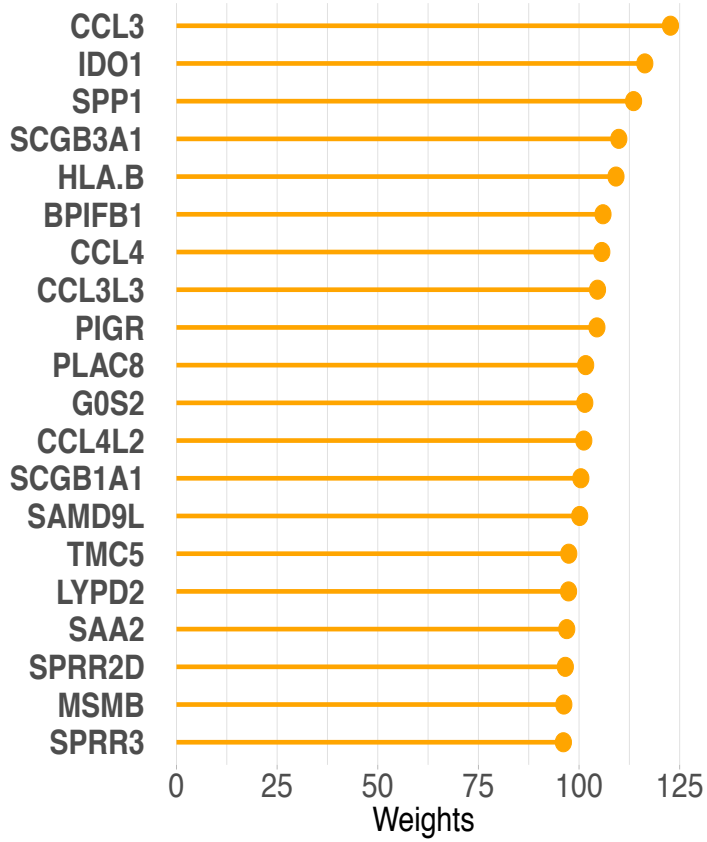


Cell Type

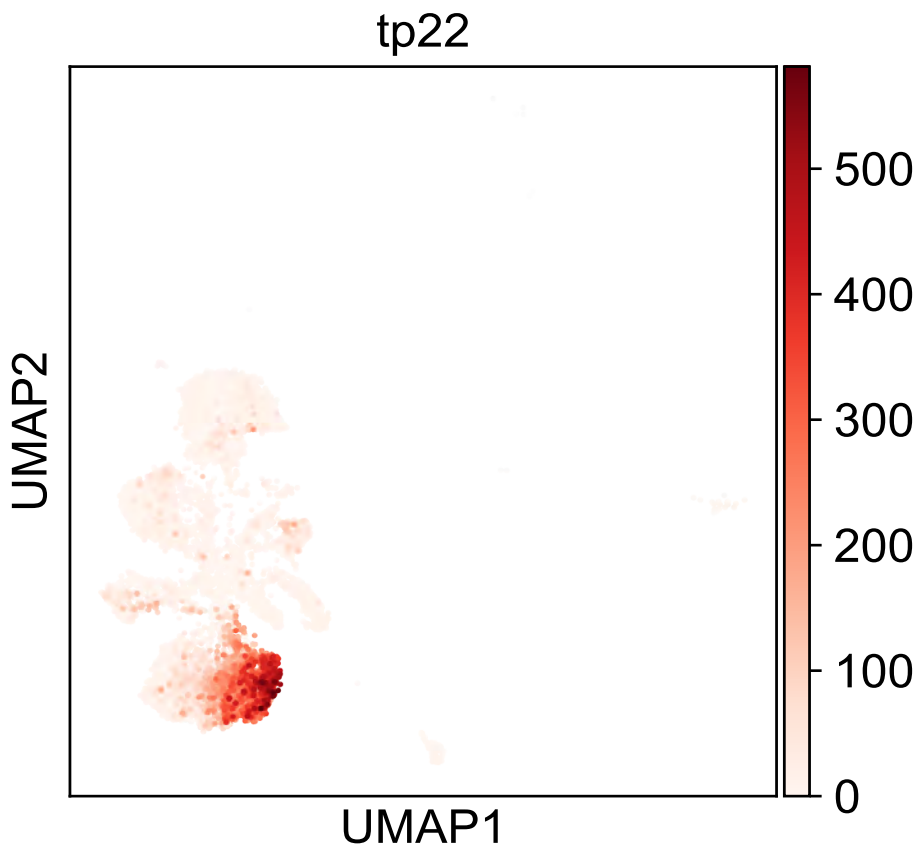
Epi_2

Epi_3

Top 20 genes

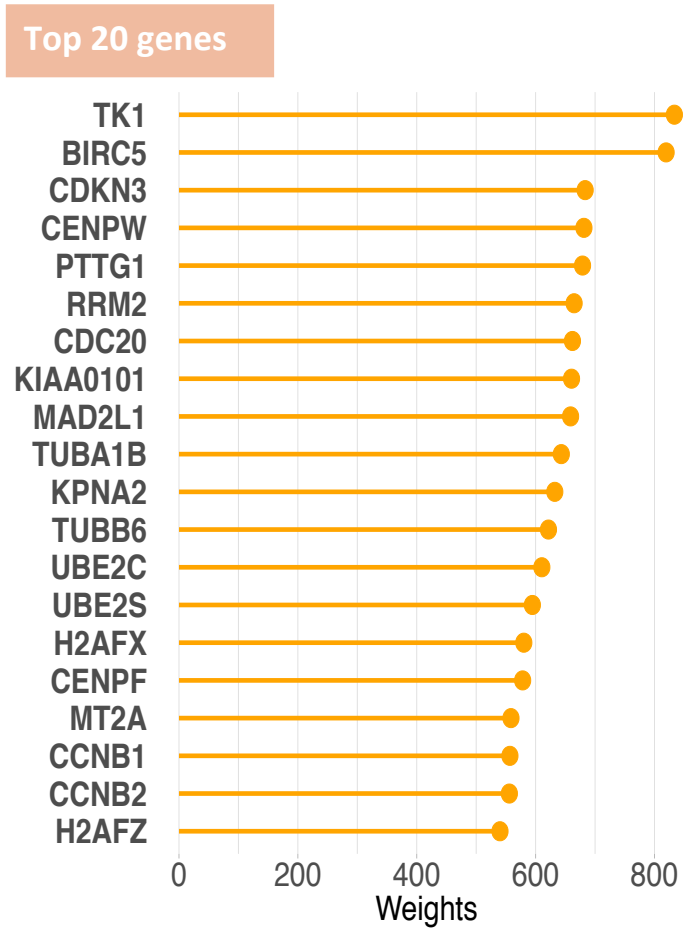


Epi GEM 22

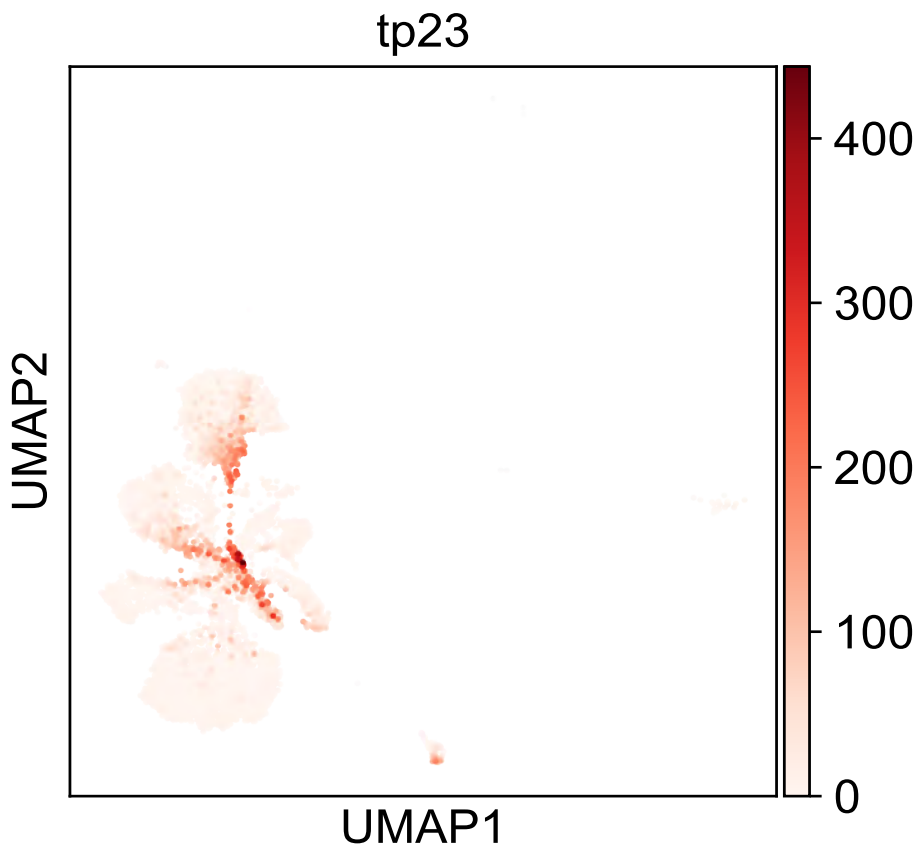


Cell Type

Epi_5

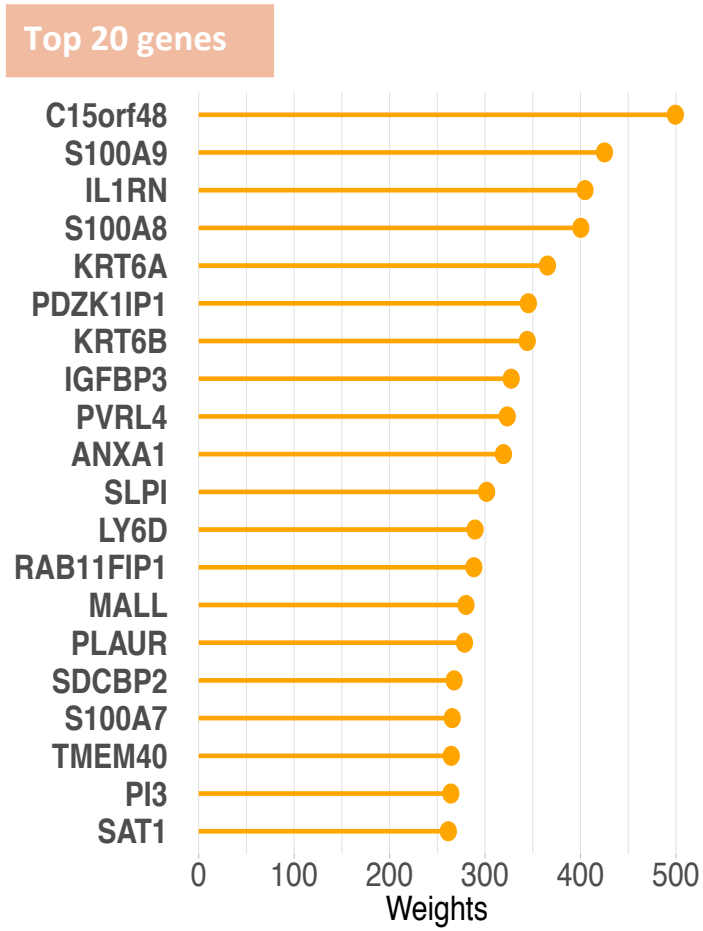


Epi GEM 23

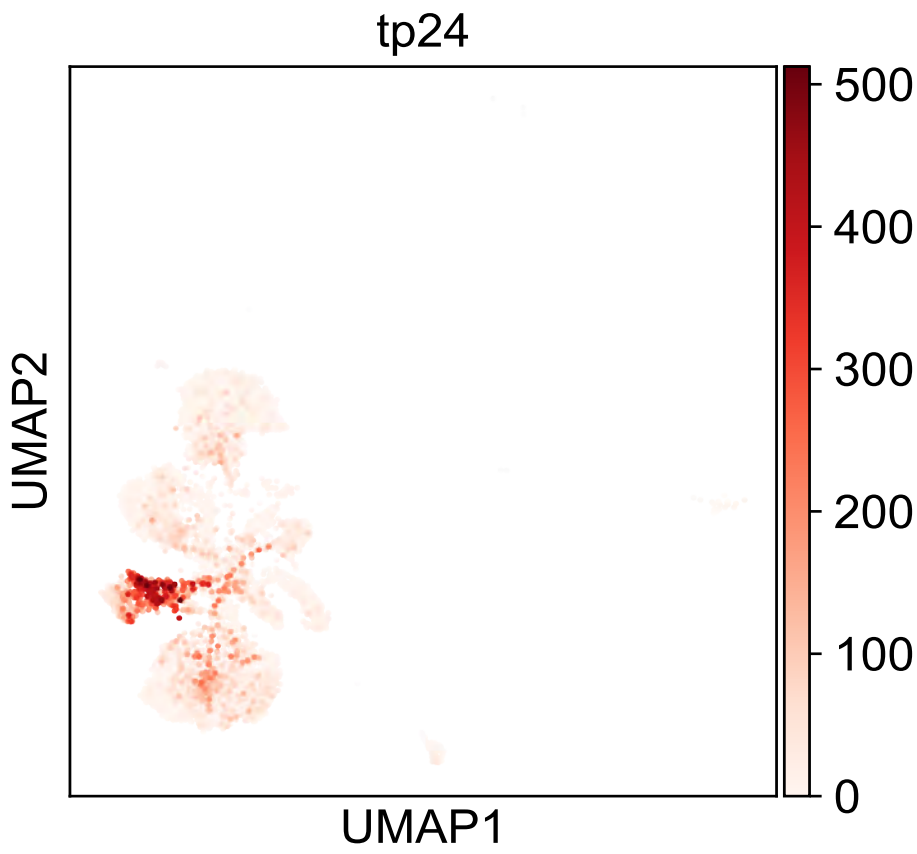


Cell Type

Epi_1



Epi GEM 24

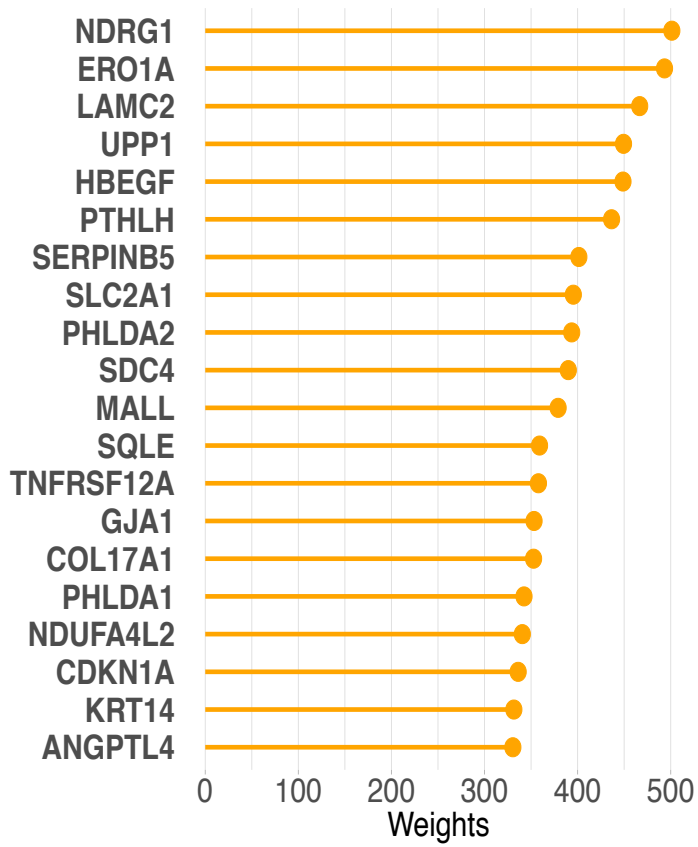


Cell Type

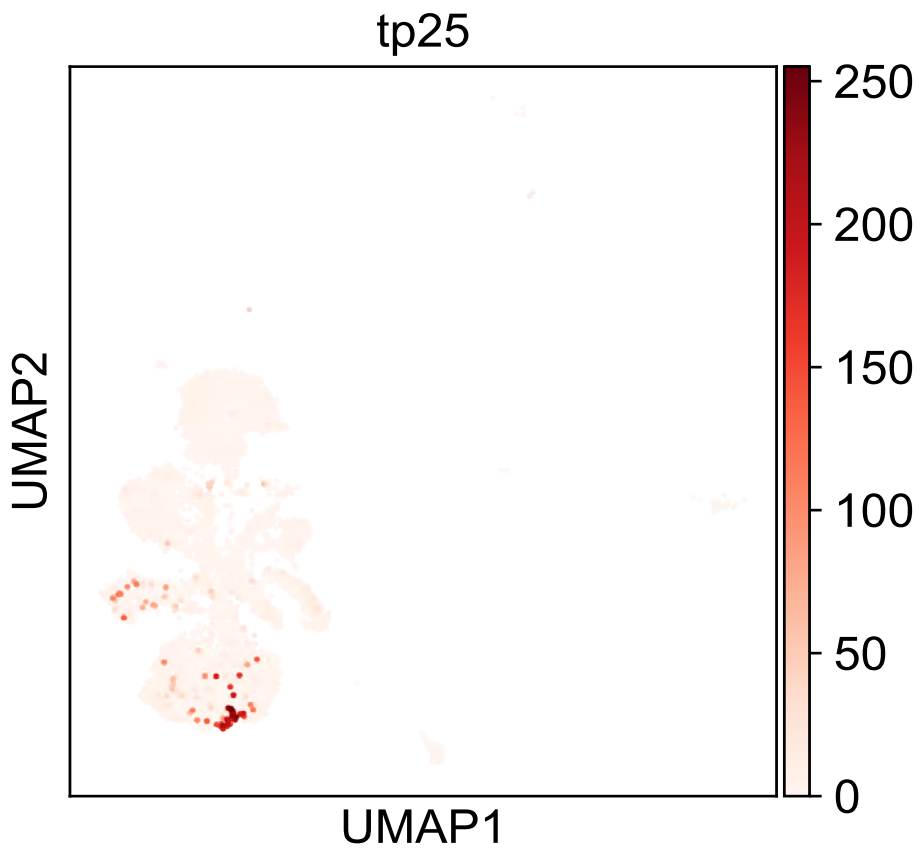
Epi_2

Epi_3

Top 20 genes

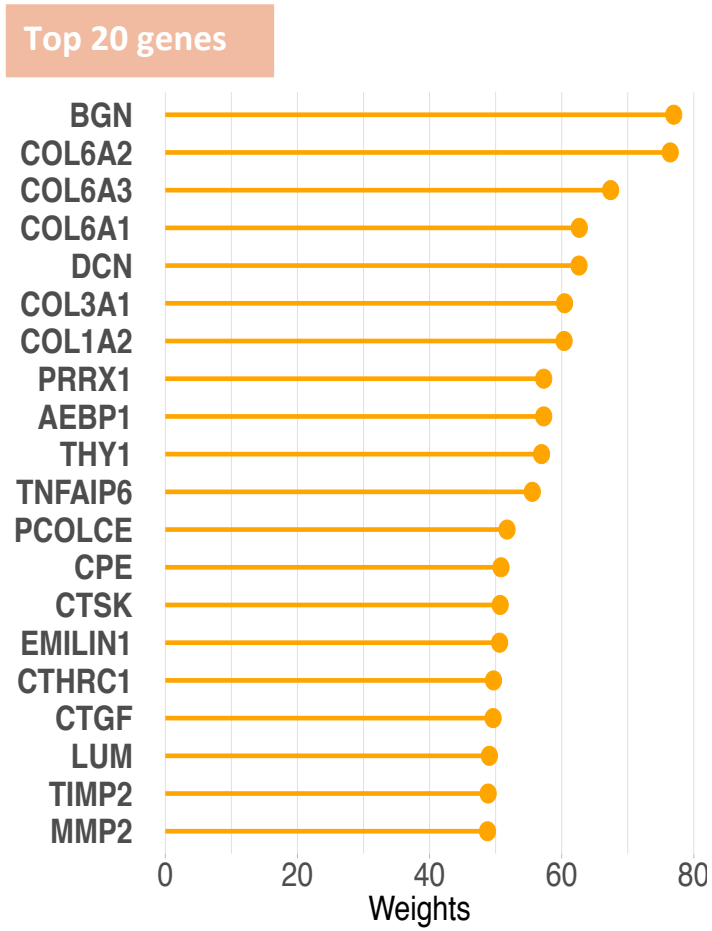


Epi GEM 25

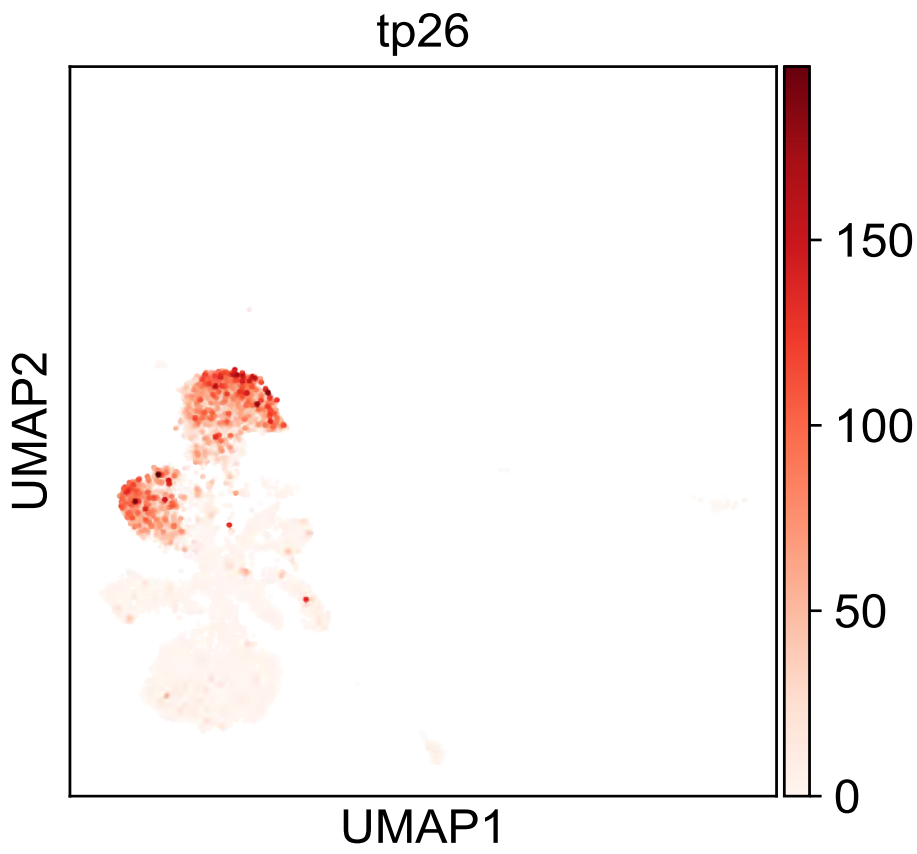


Cell Type

Epi_4



Epi GEM 26

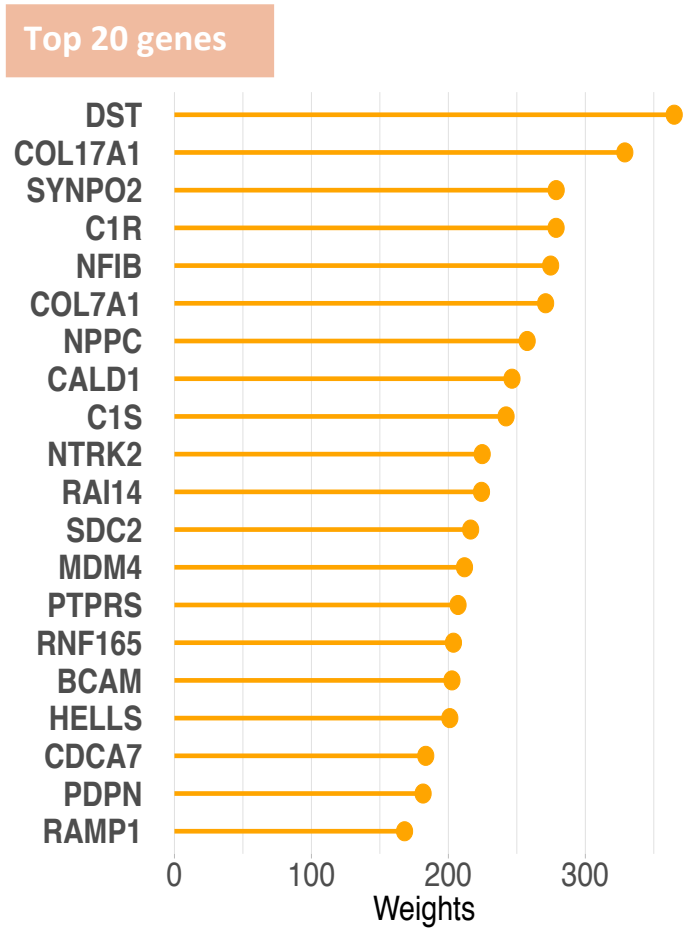


Cell Type

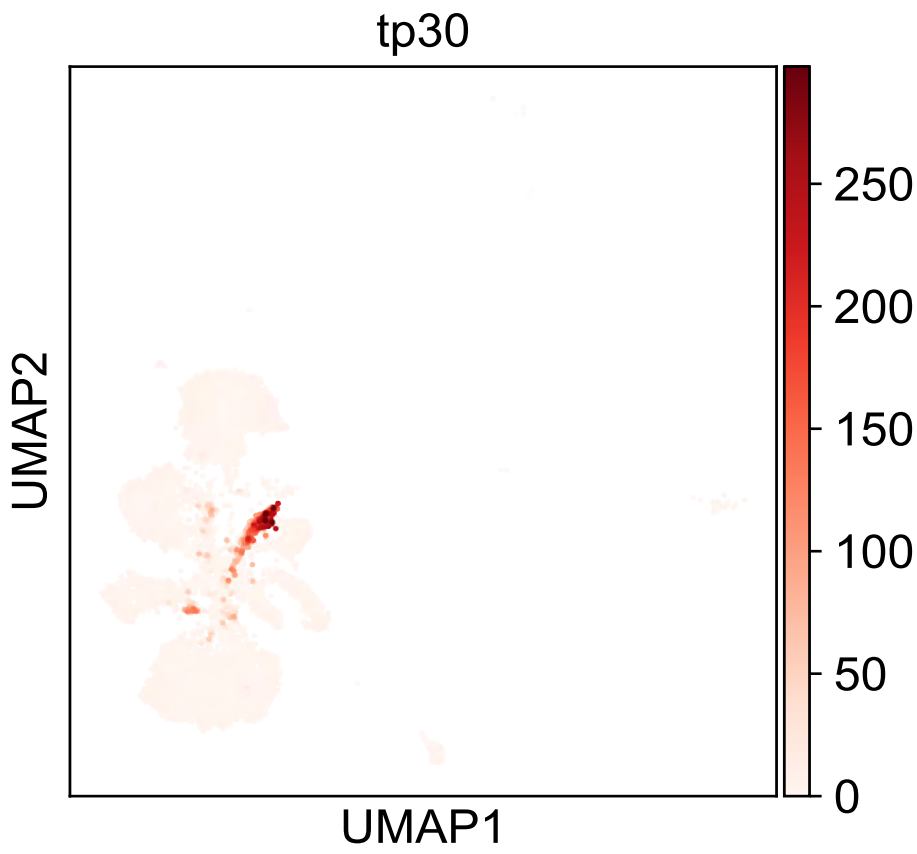
Epi_6

Epi_7

Epi_8

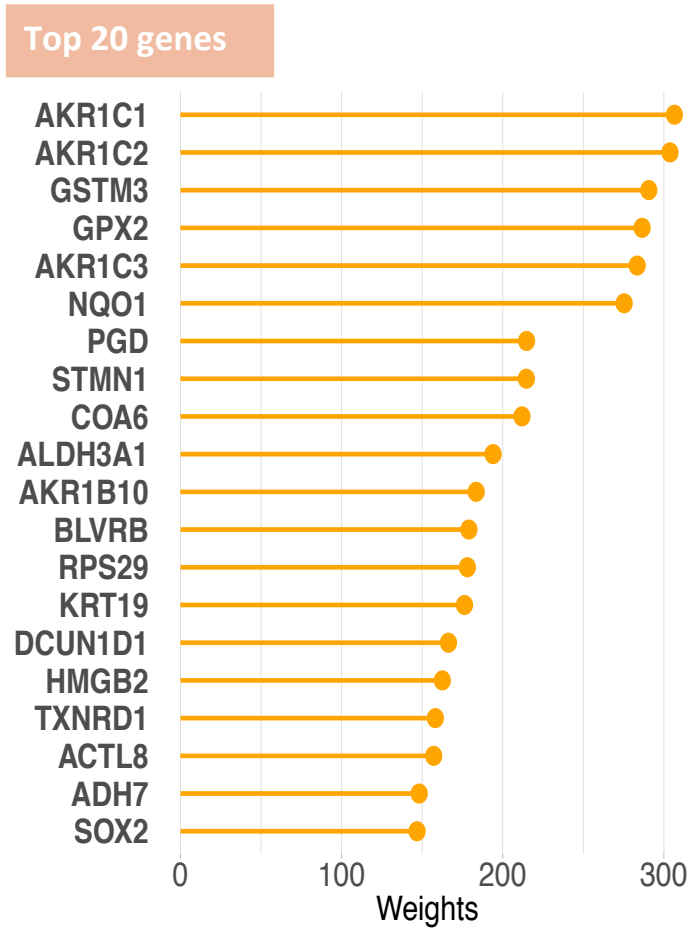


Epi GEM 30

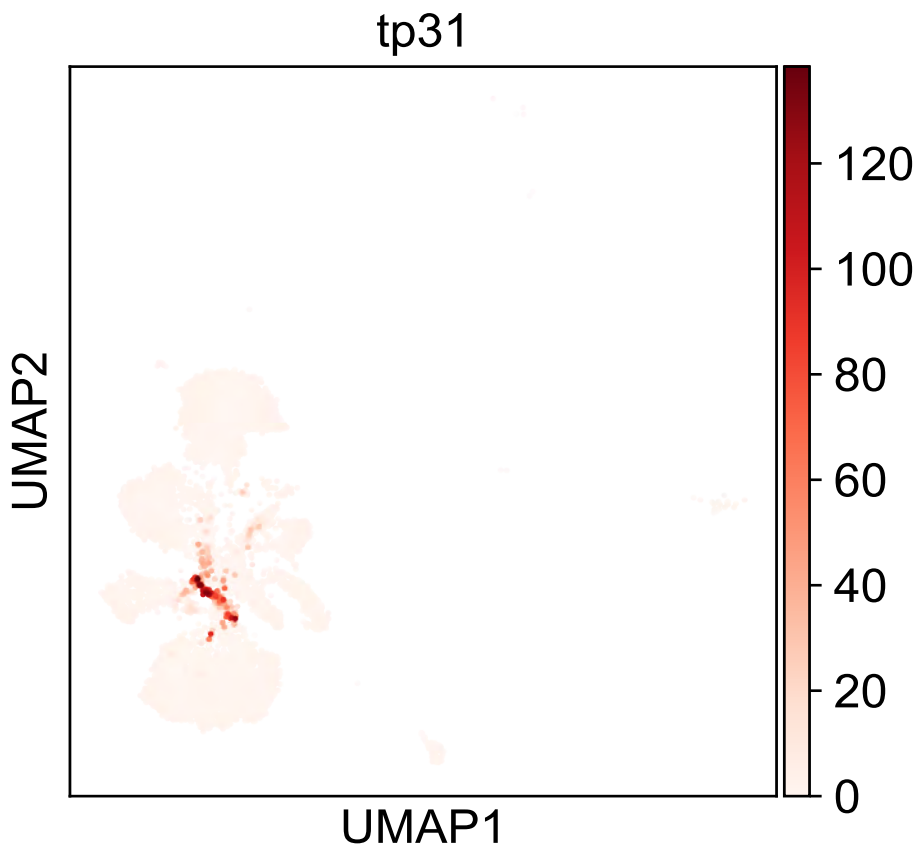


Cell Type

Epi_1

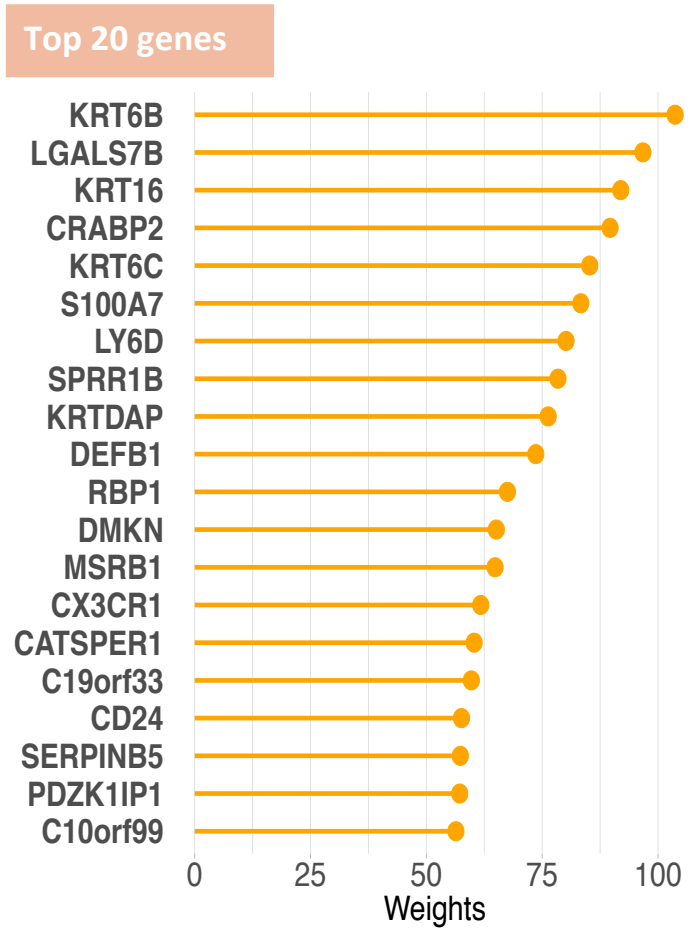


Epi GEM 31

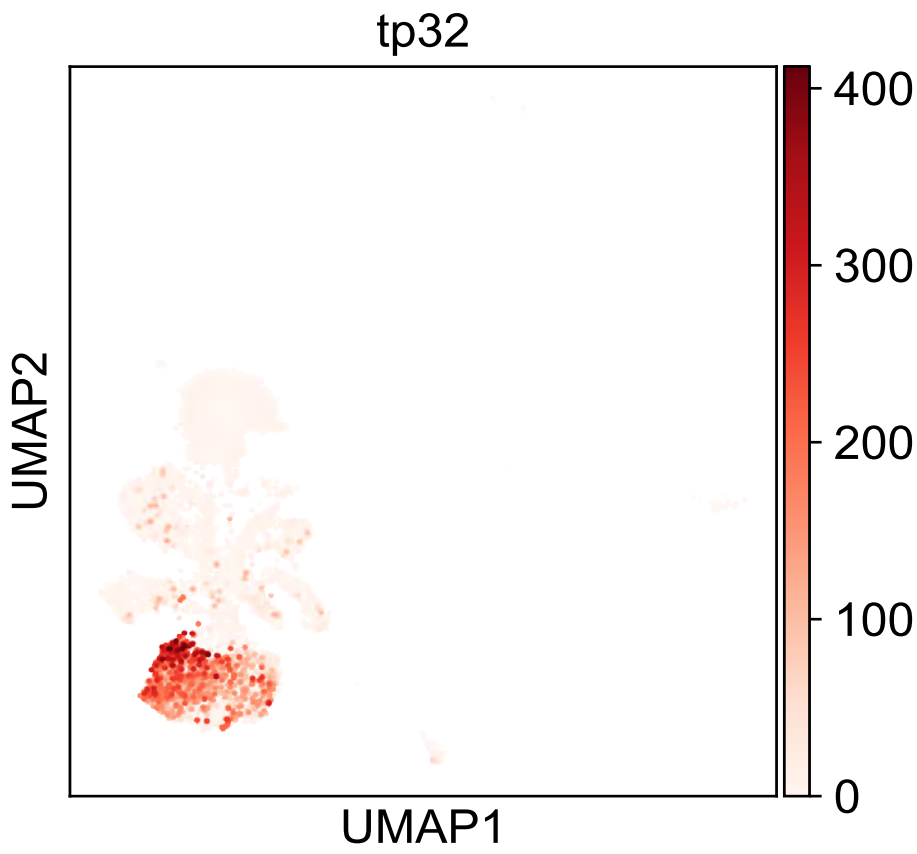


Cell Type

Epi_1



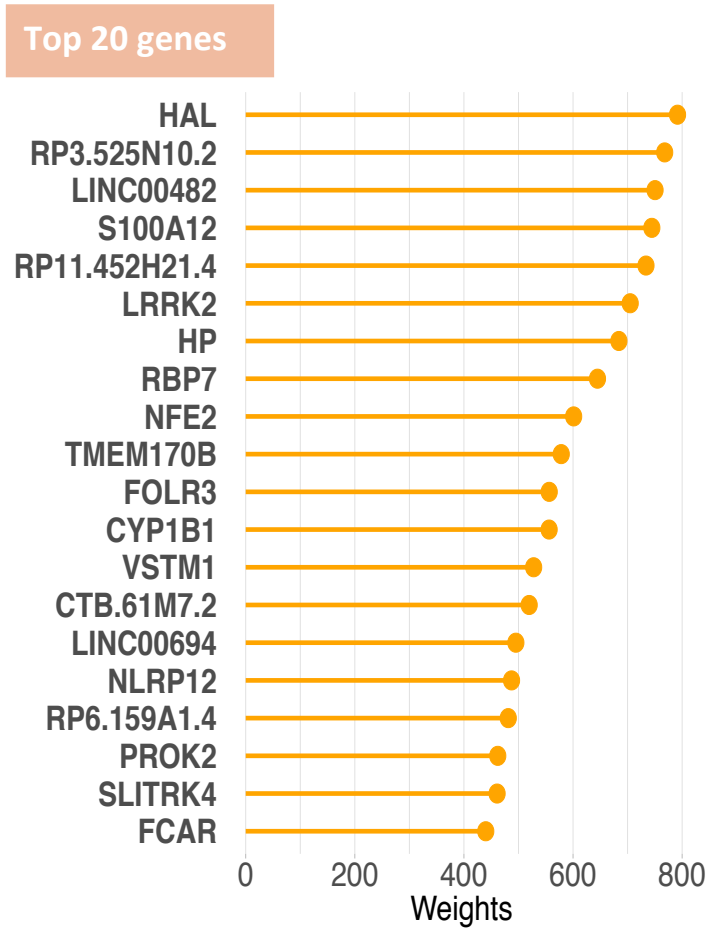
Epi GEM 32



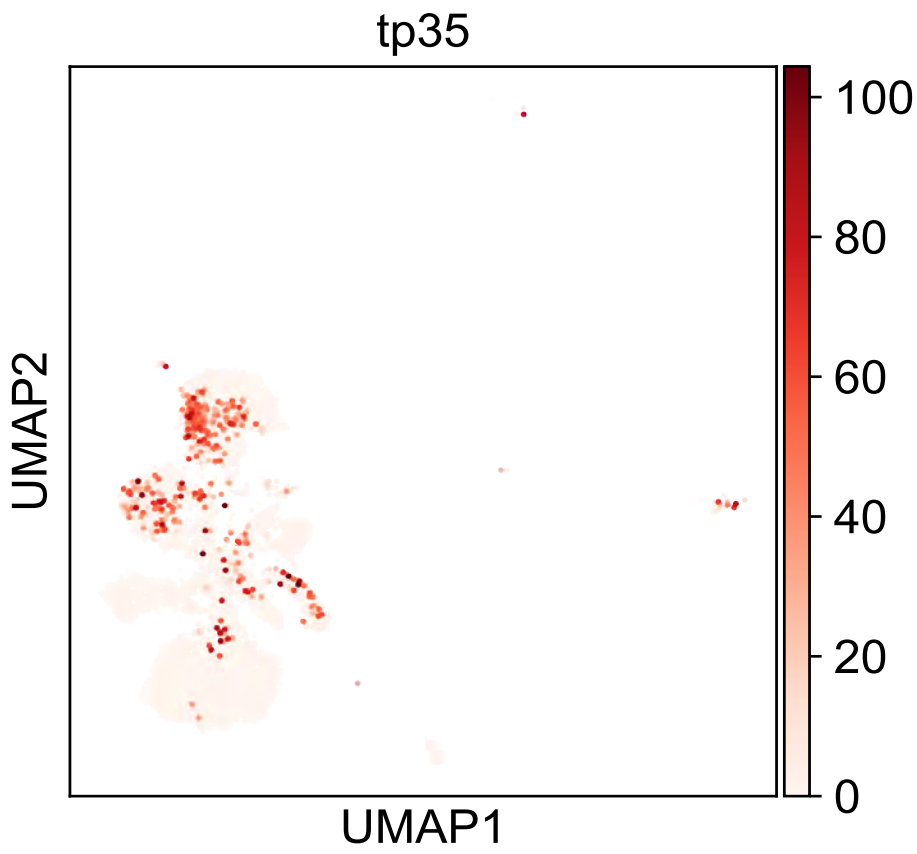
Cell Type

Epi_4

Epi_5



Epi GEM 35

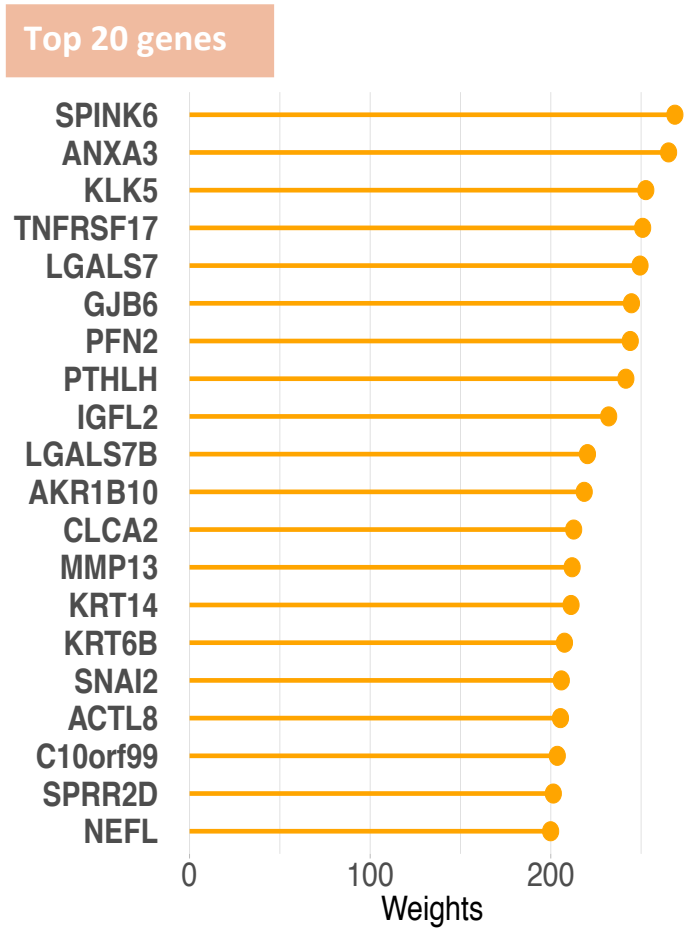


Cell Type

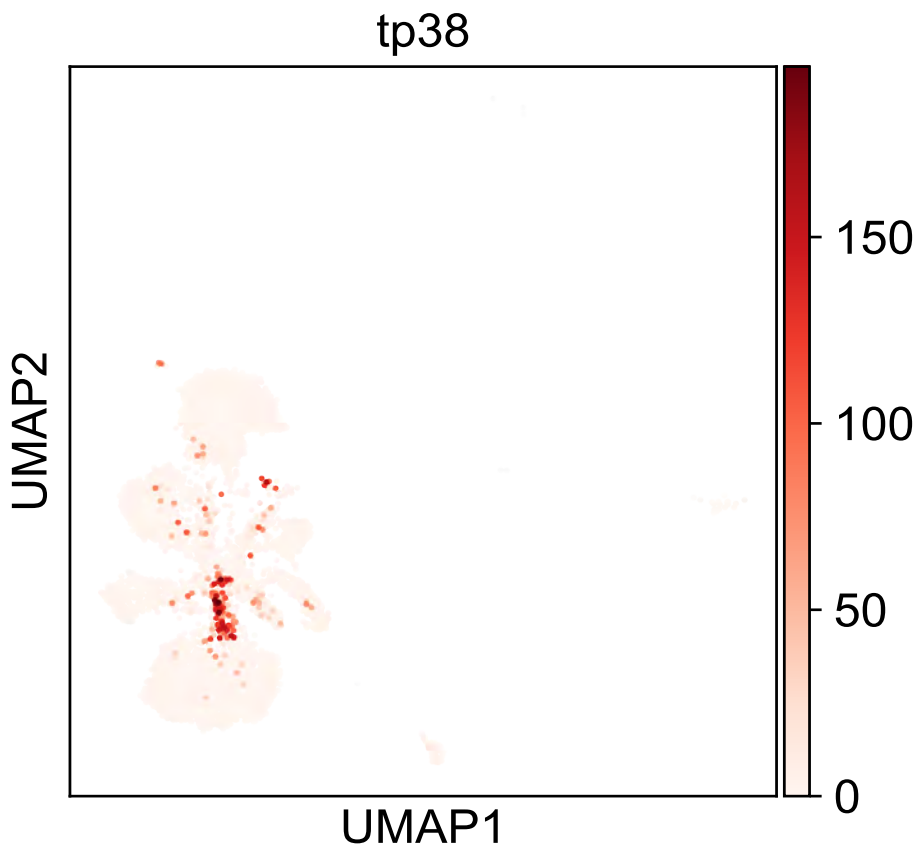
Epi_7

Epi_3

Epi_6



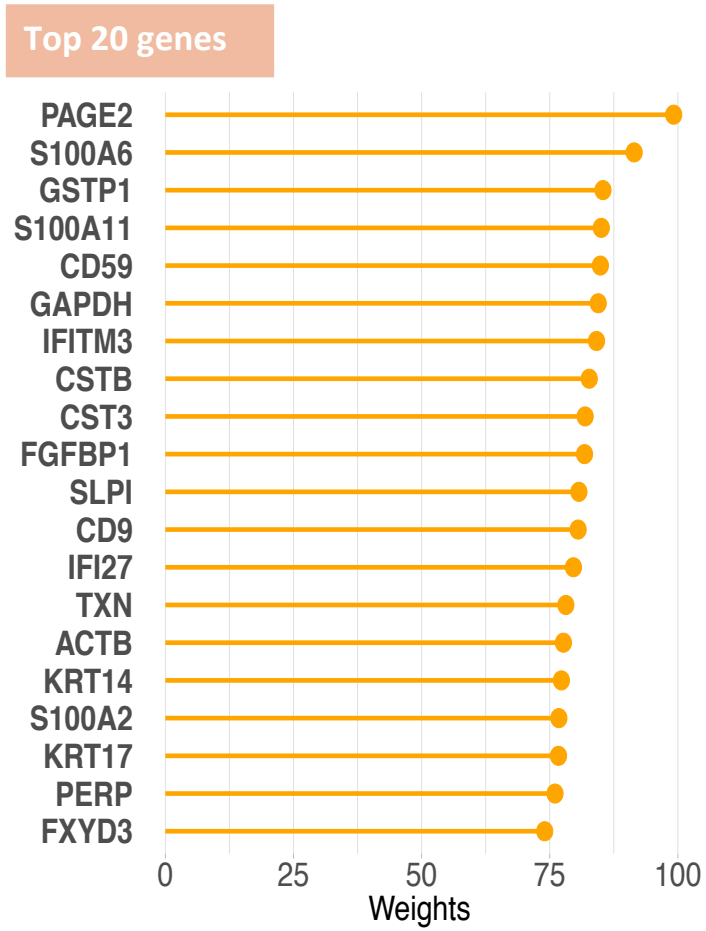
Epi GEM 38



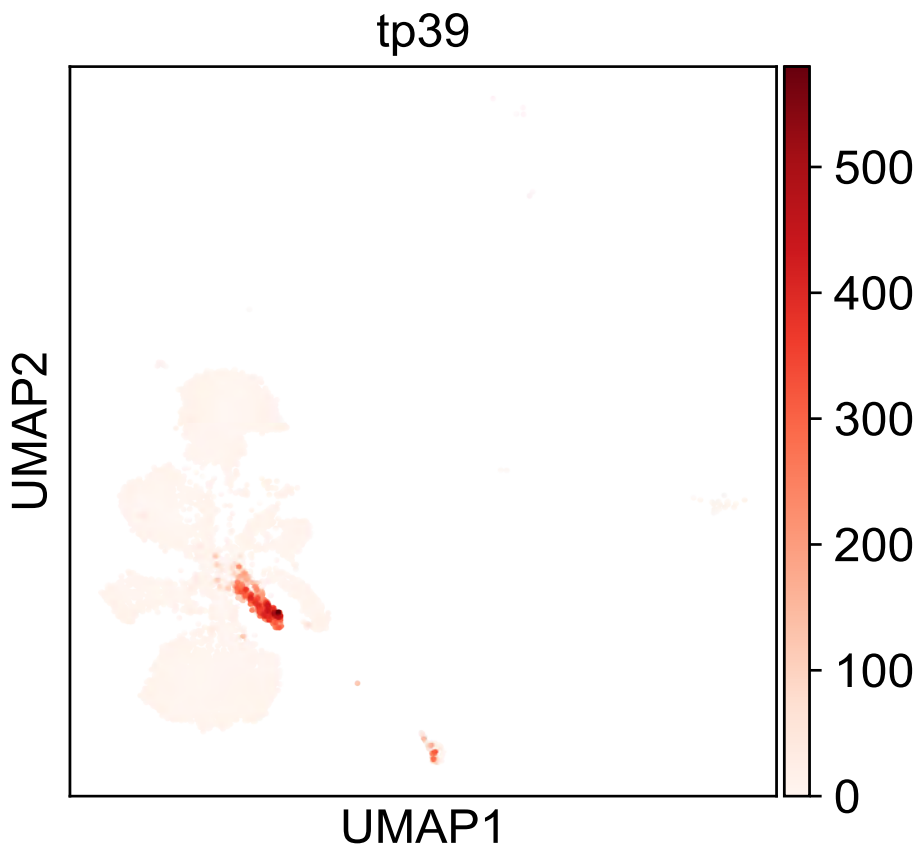
Cell Type

Epi_1

Epi_3

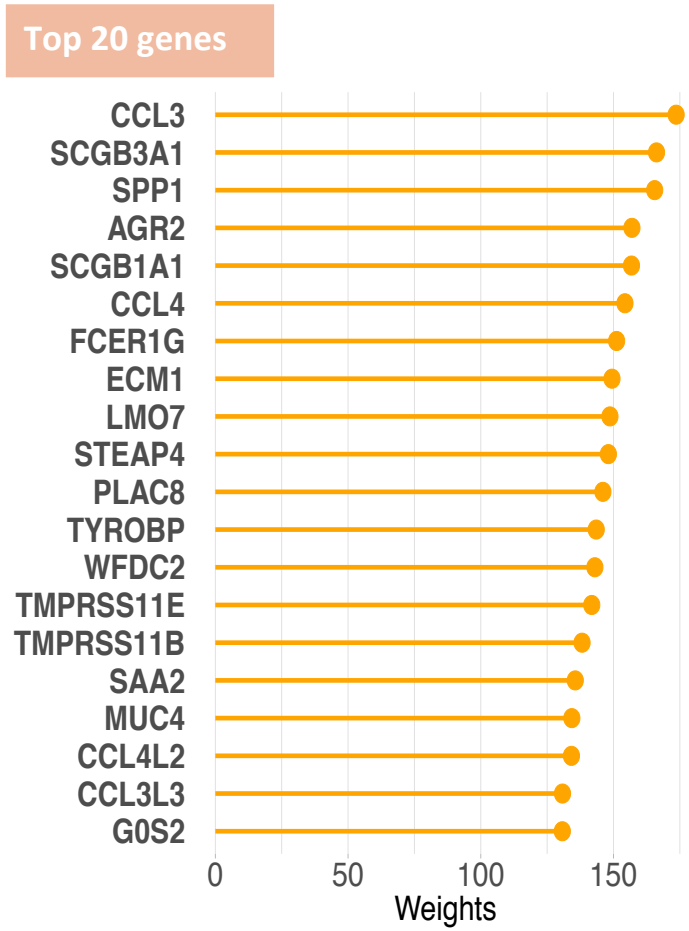


Epi GEM 39

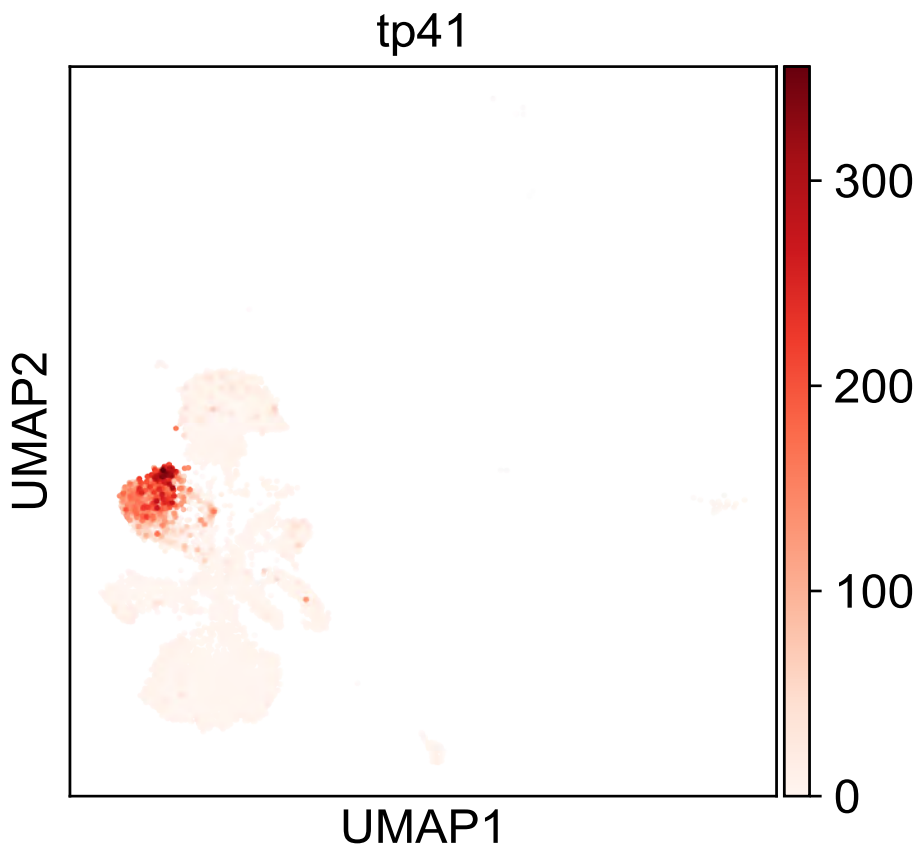


Cell Type

Epi_1

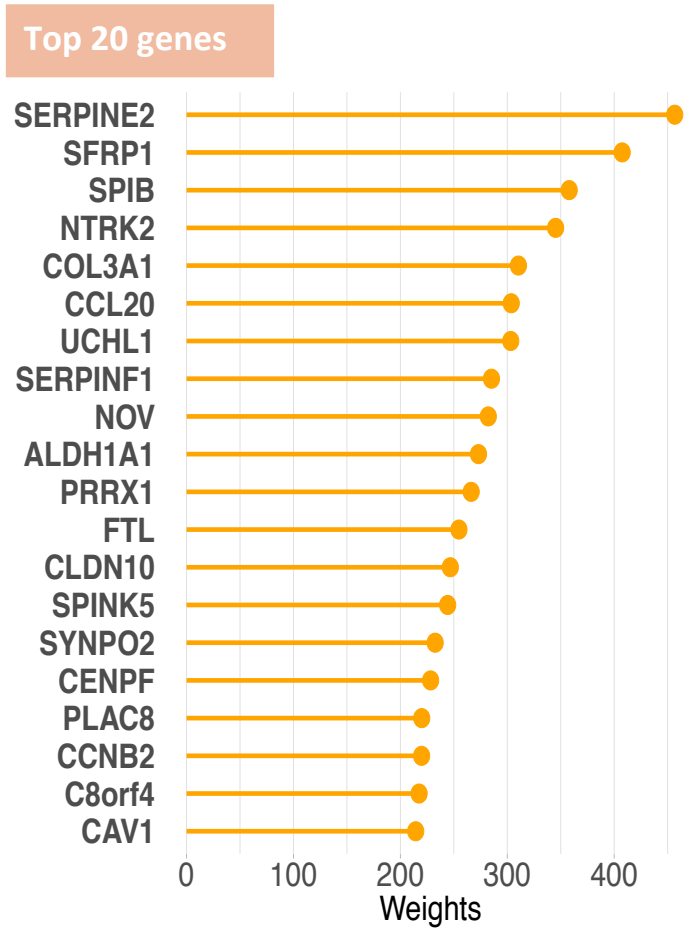


Epi GEM 41

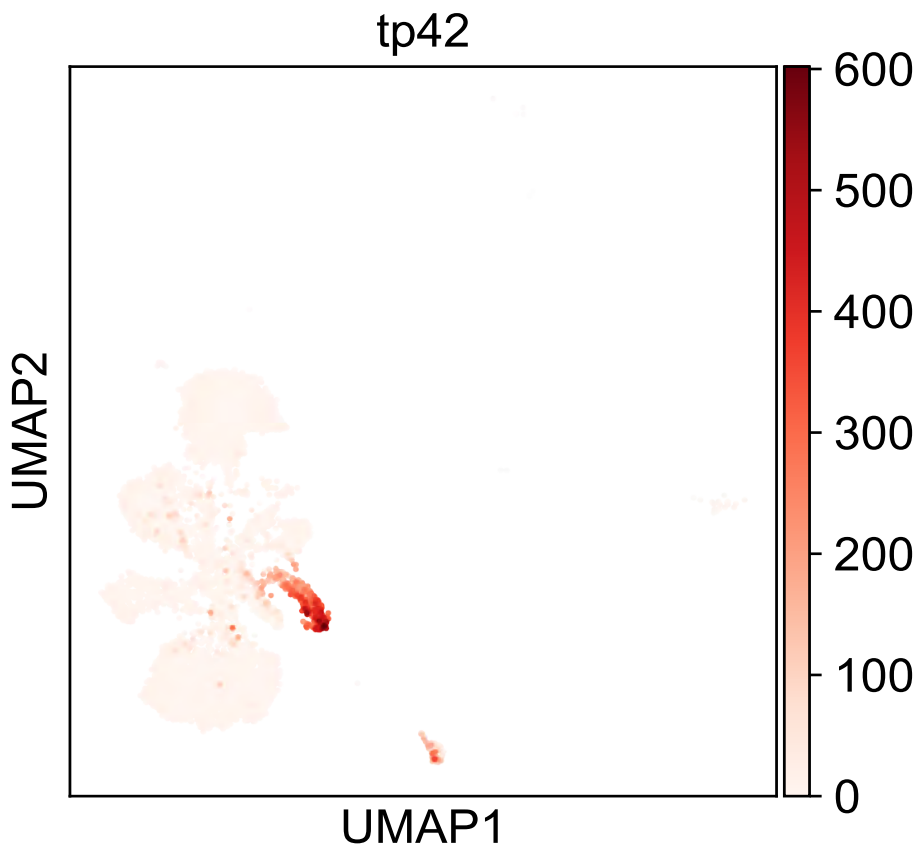


Cell Type

Epi_6

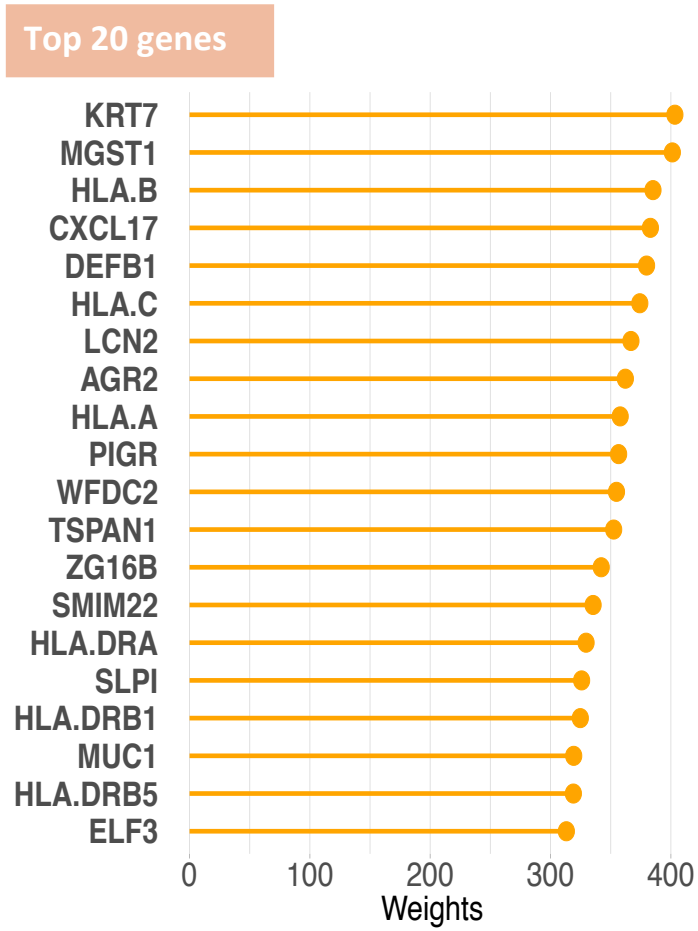


Epi GEM 42

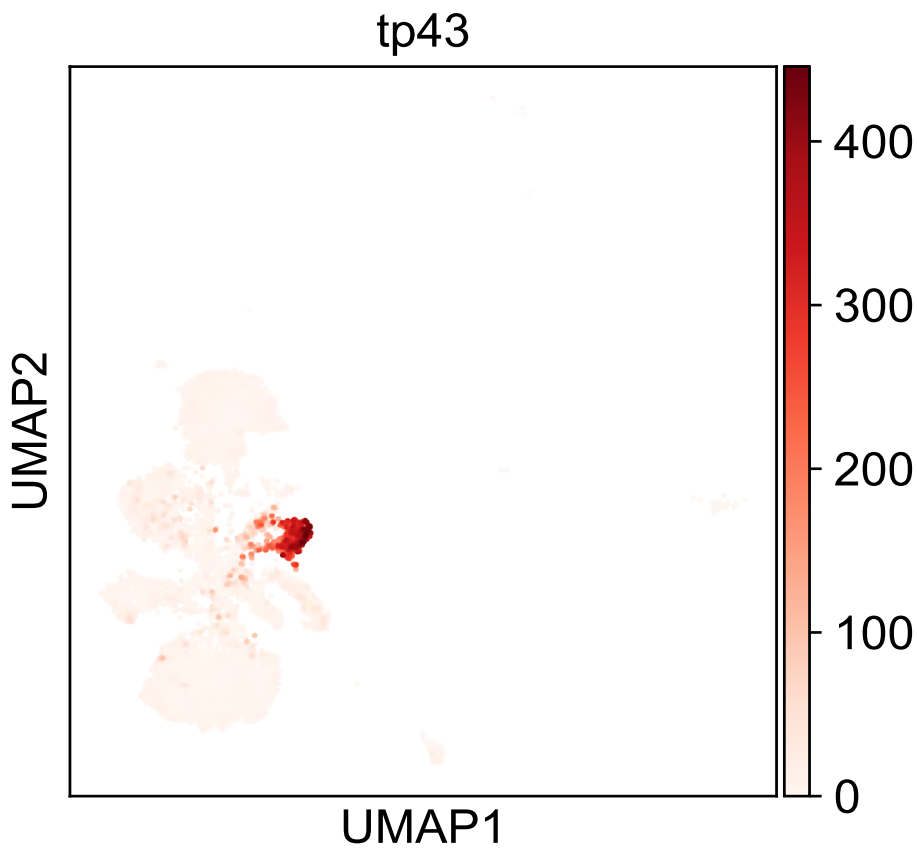


Cell Type

Epi_3



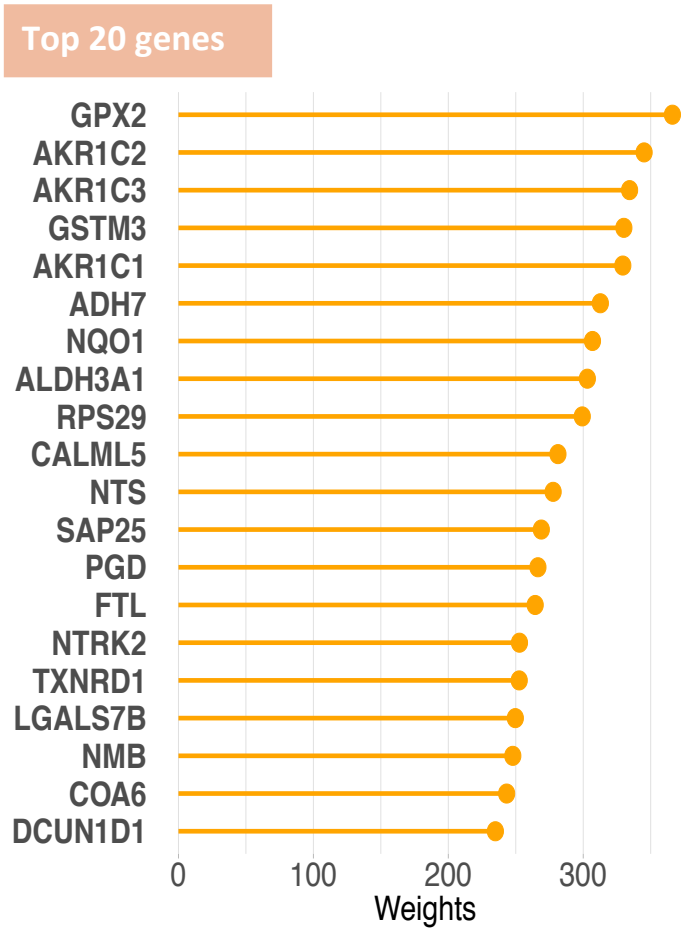
Epi GEM 43



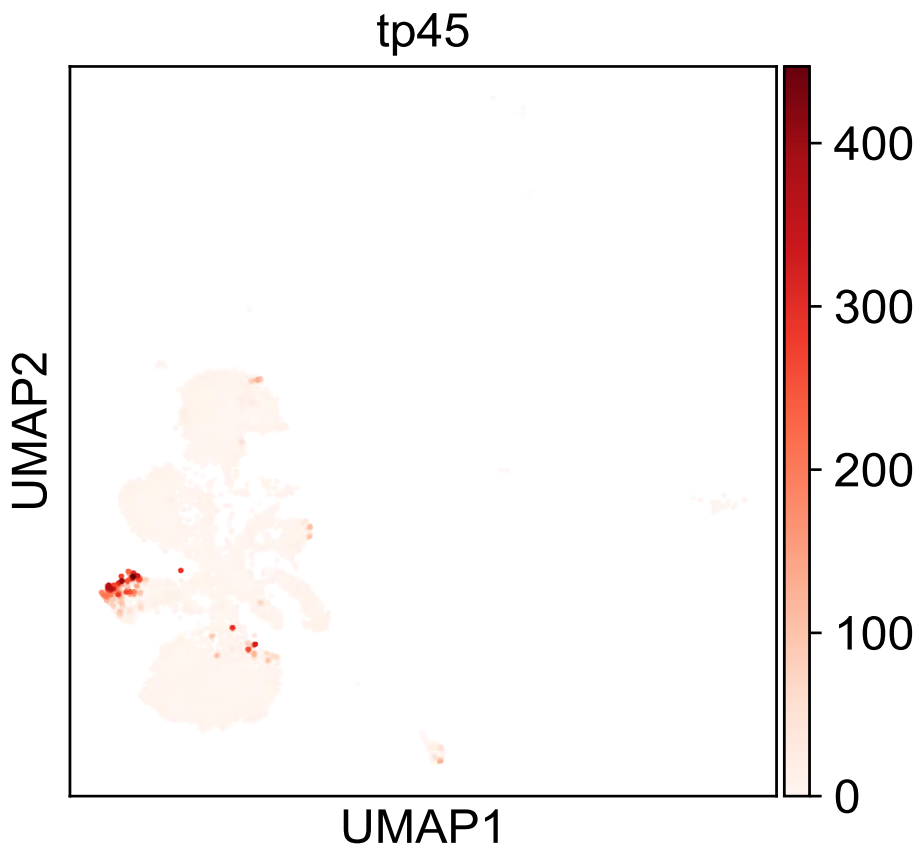
Cell Type

Epi_3

Epi_2

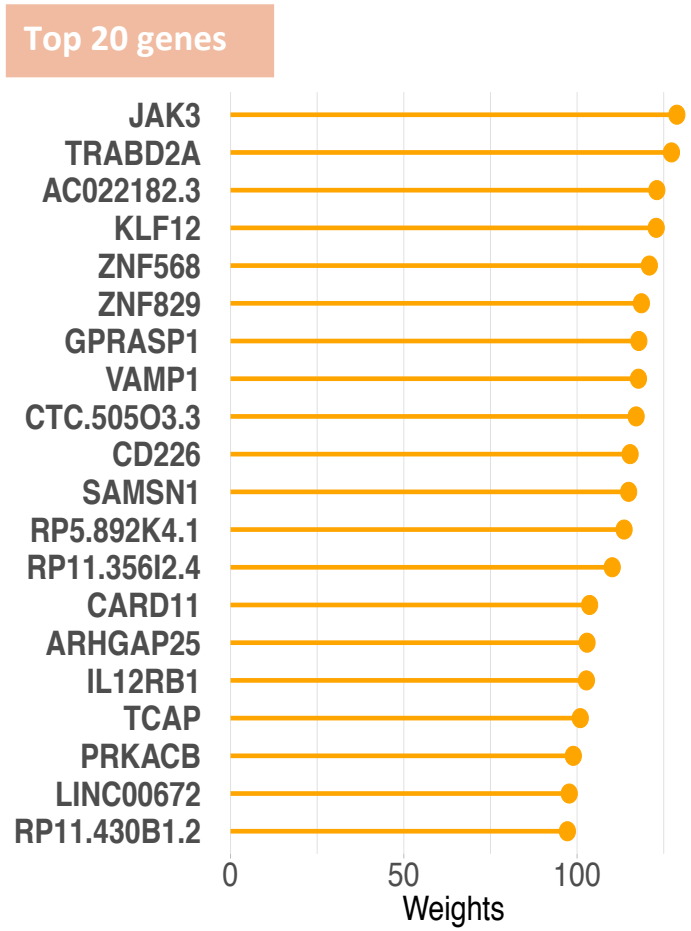


Epi GEM 45

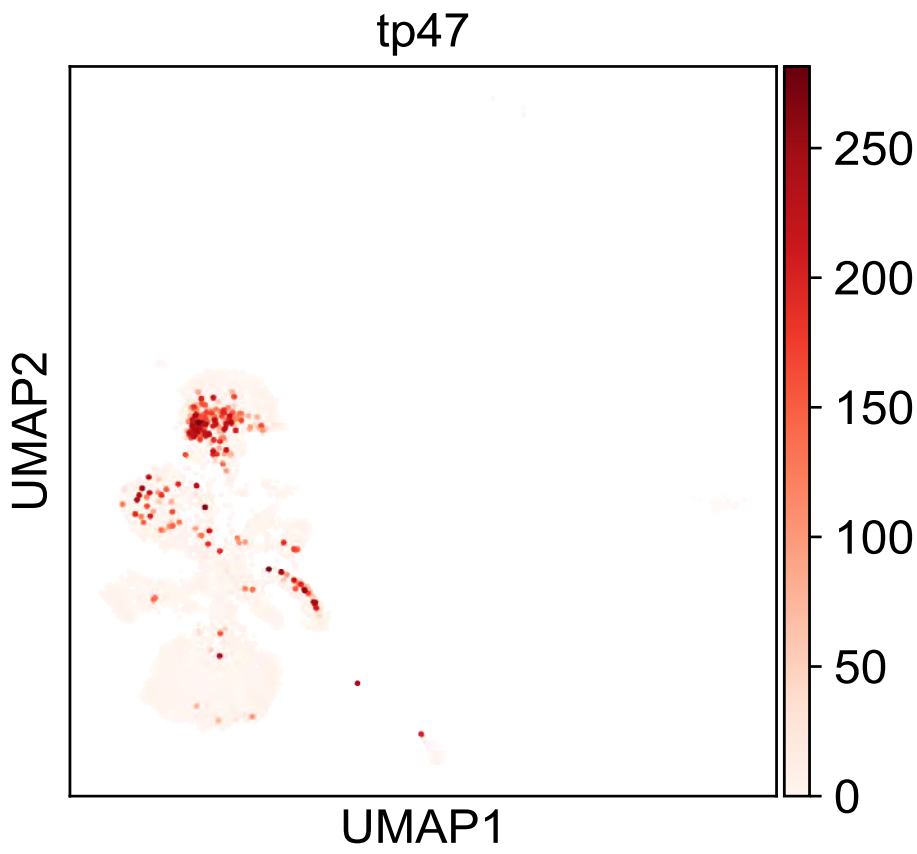


Cell Type

Epi_2



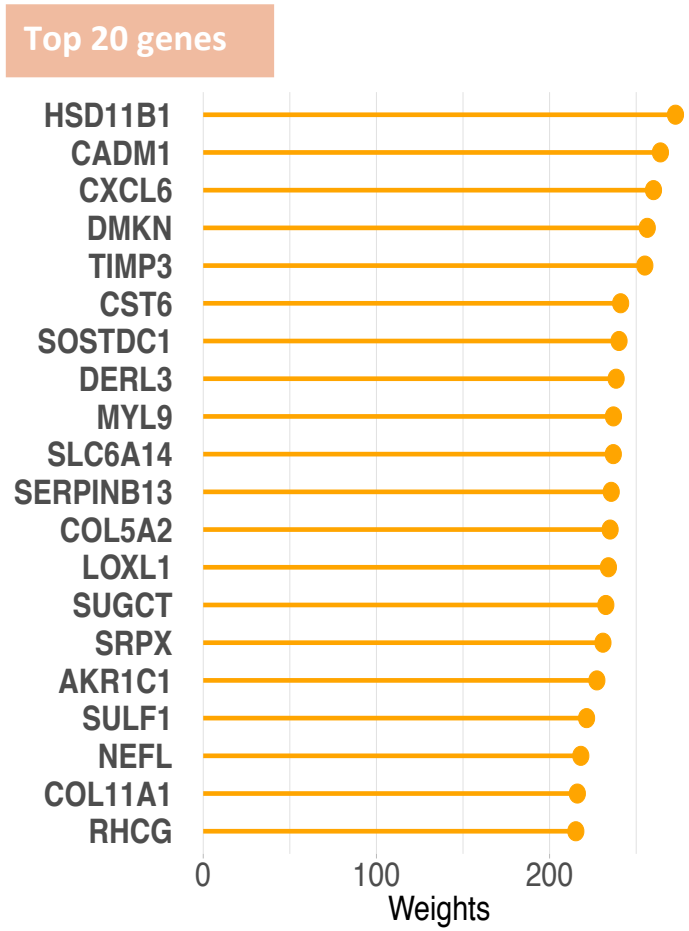
Epi GEM 47



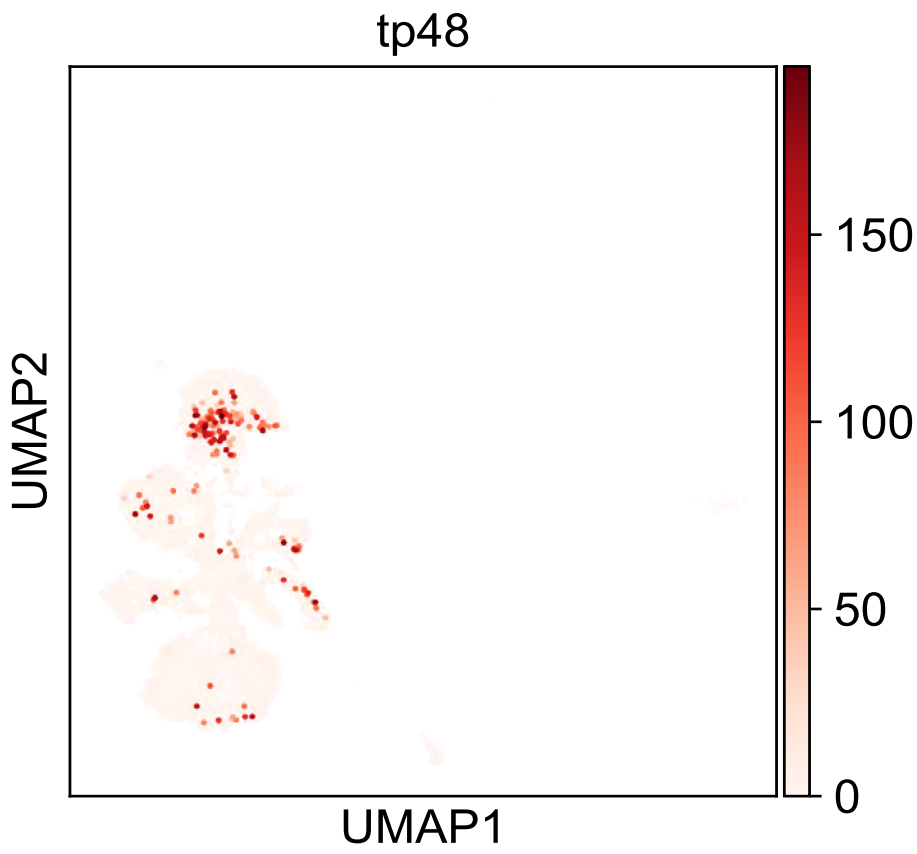
Cell Type

Epi_7

Epi_3



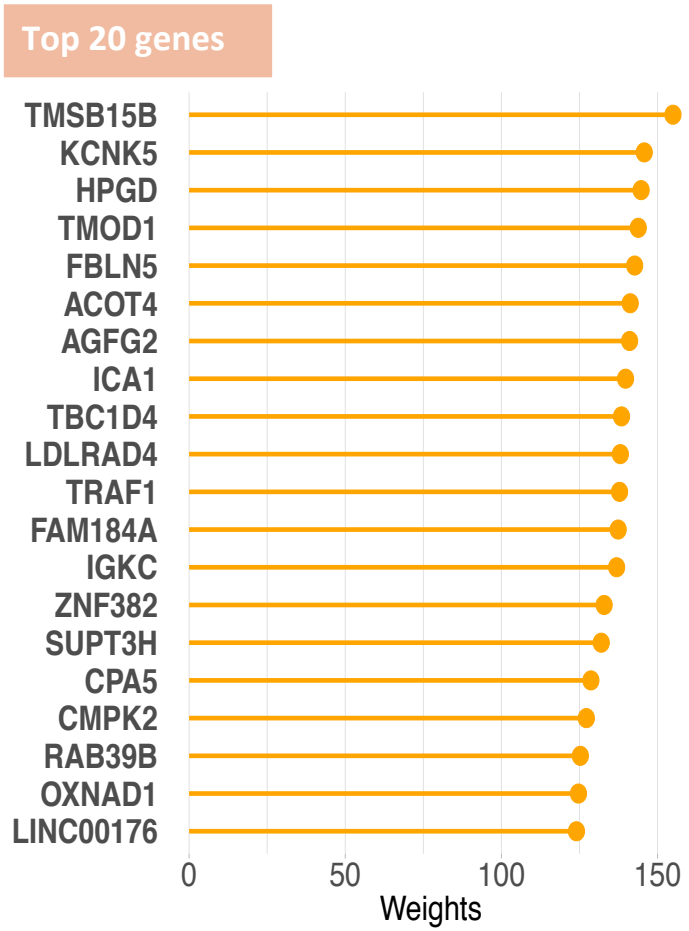
Epi GEM 48



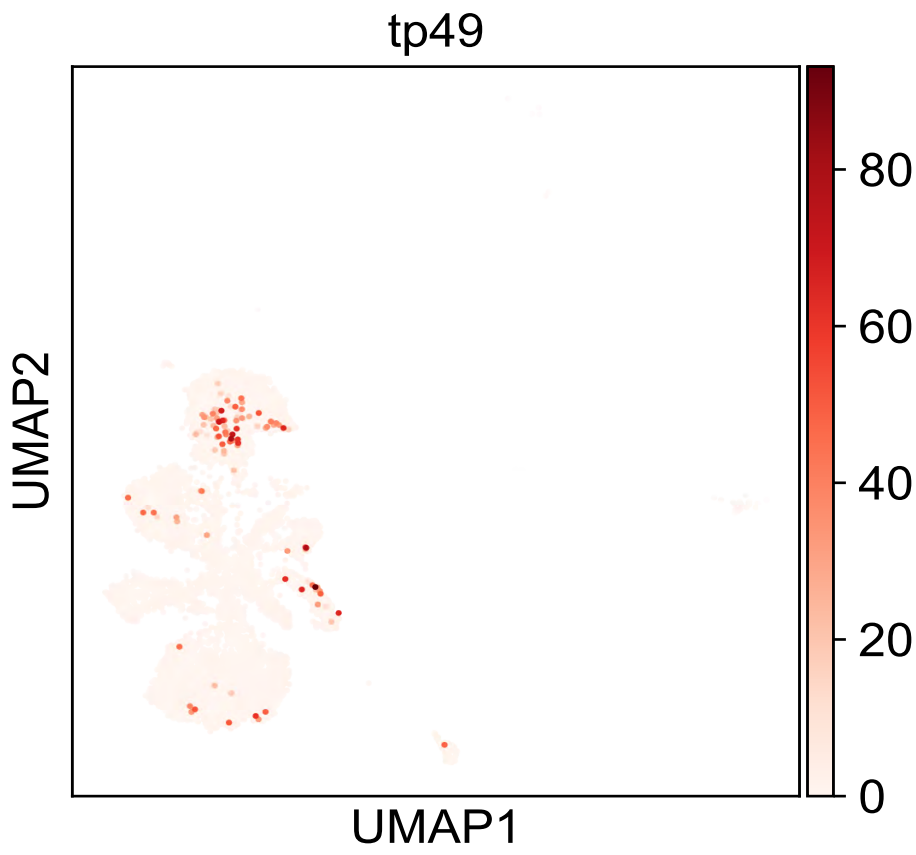
Cell Type

Epi_7

Epi_3



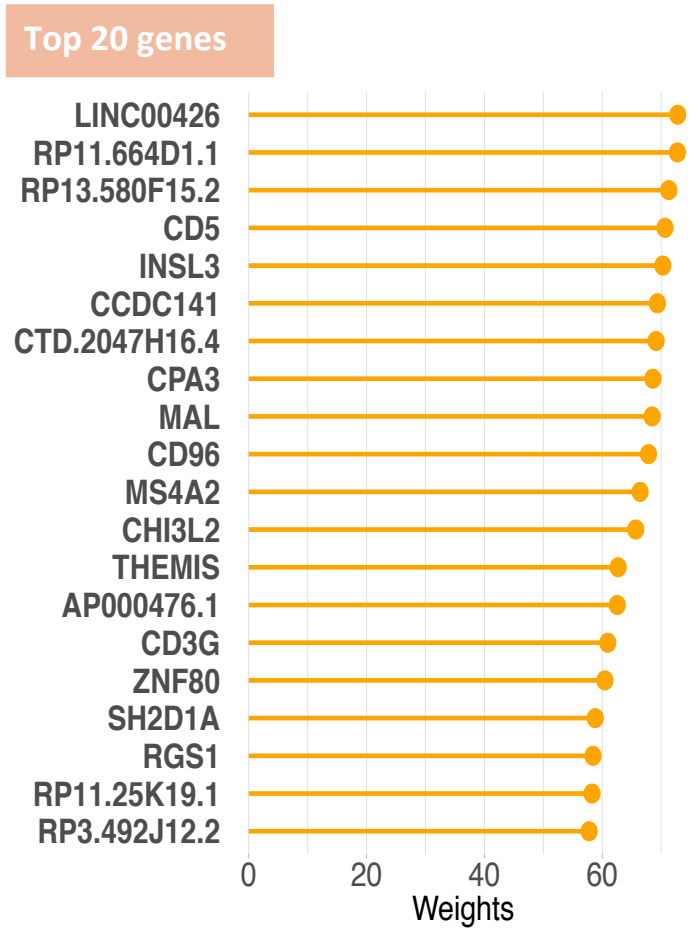
Epi GEM 49



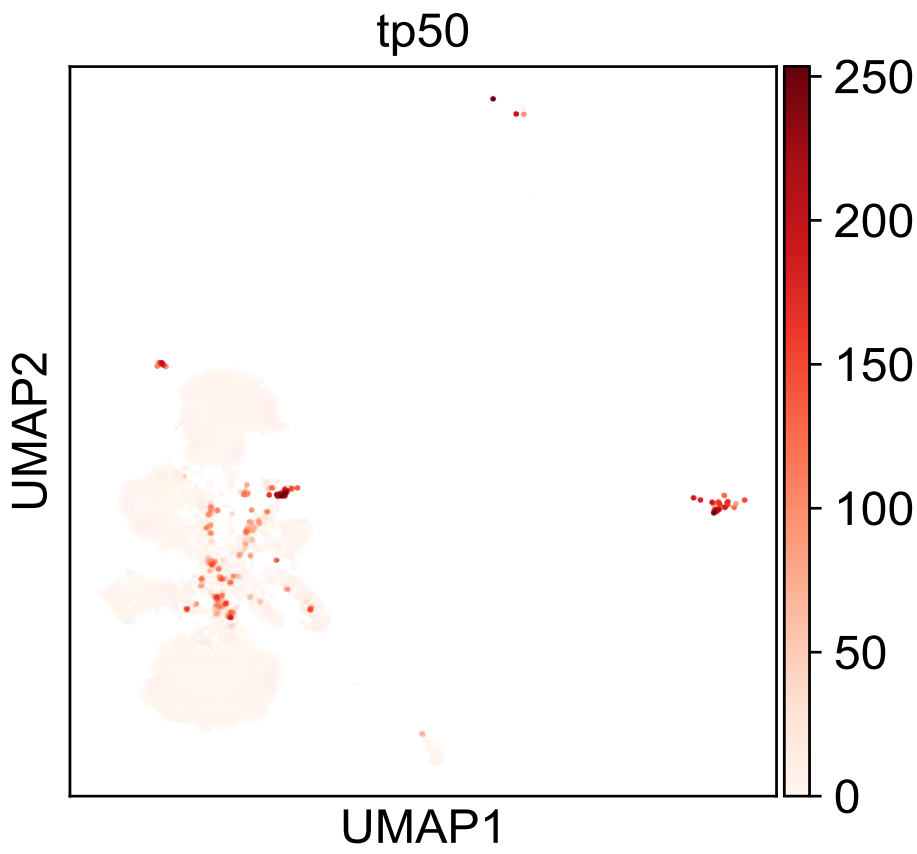
Cell Type

Epi_7

Epi_3

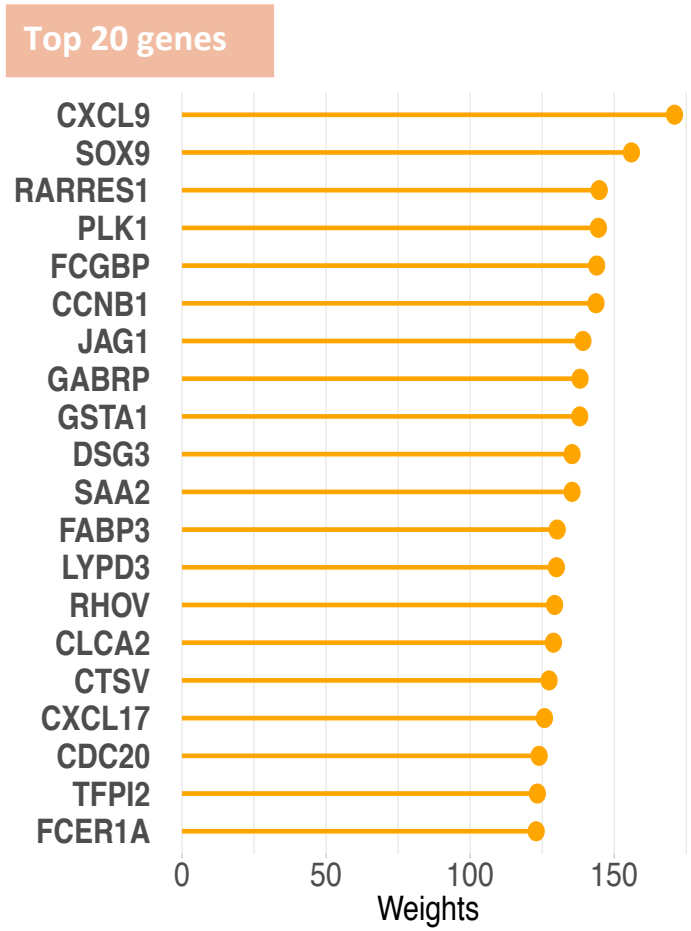


Epi GEM 50

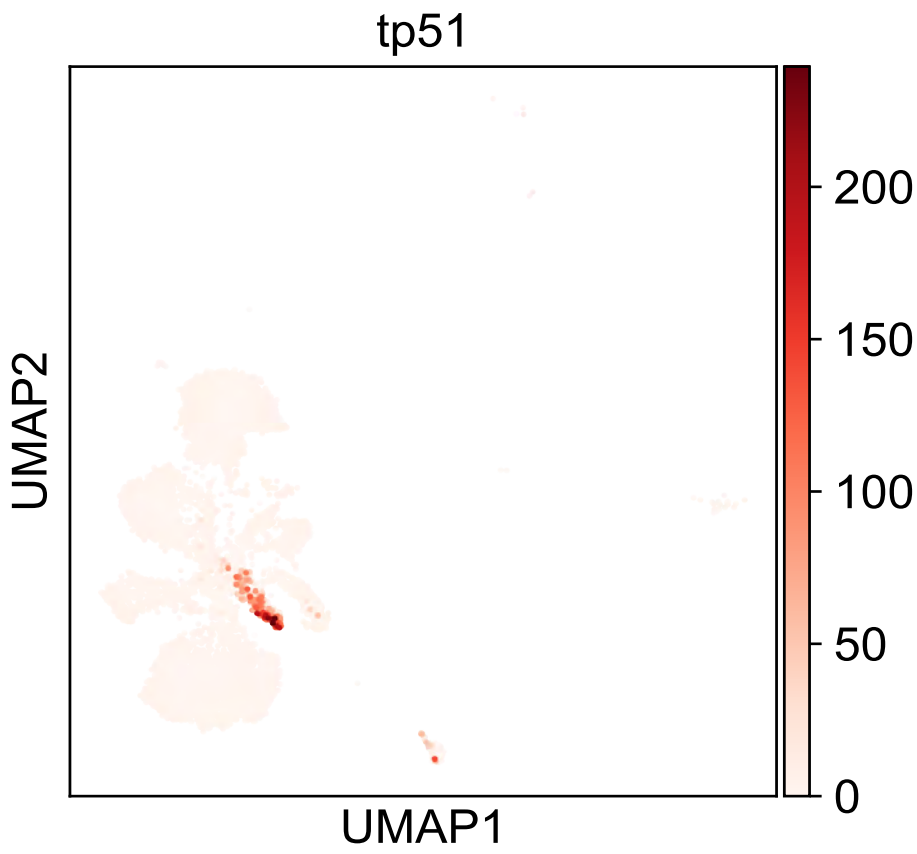


Cell Type

Epi_1

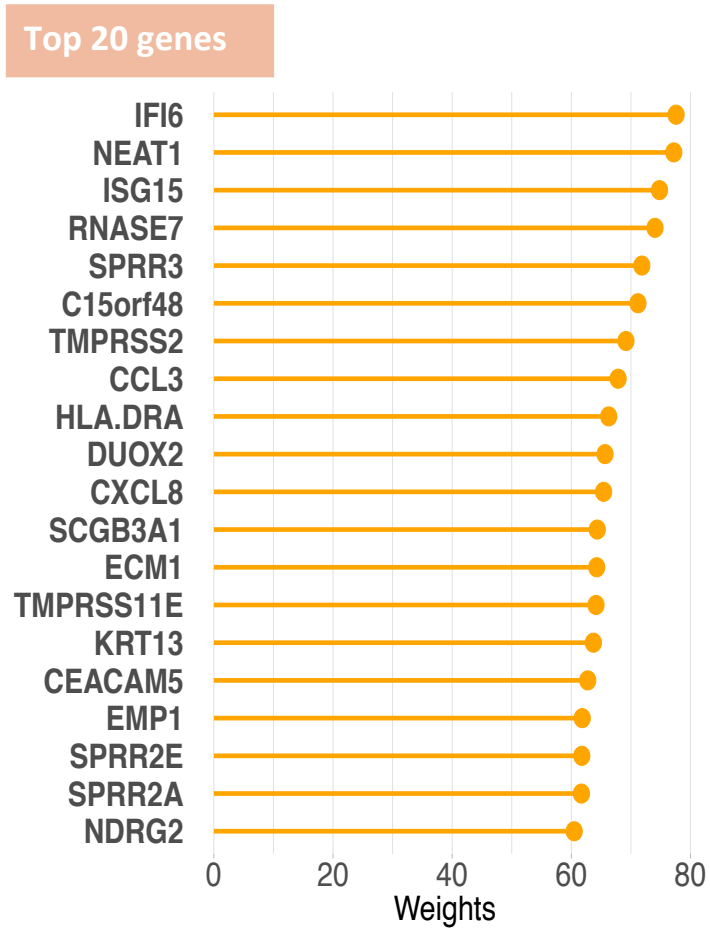


Epi GEM 51

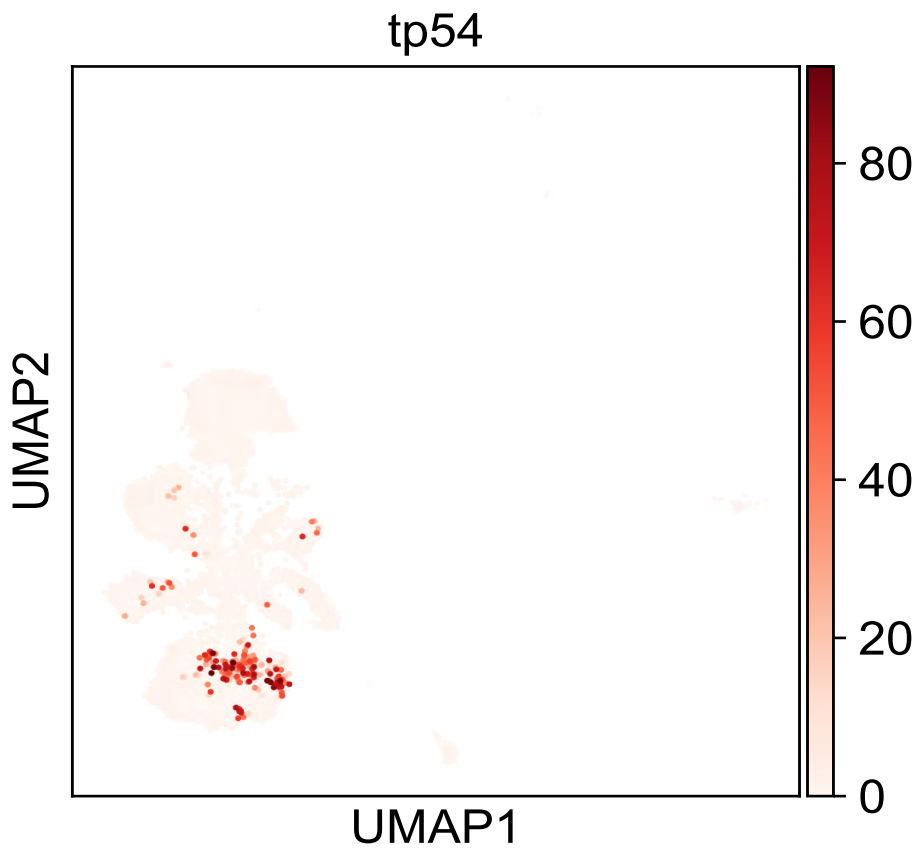


Cell Type

Epi_1

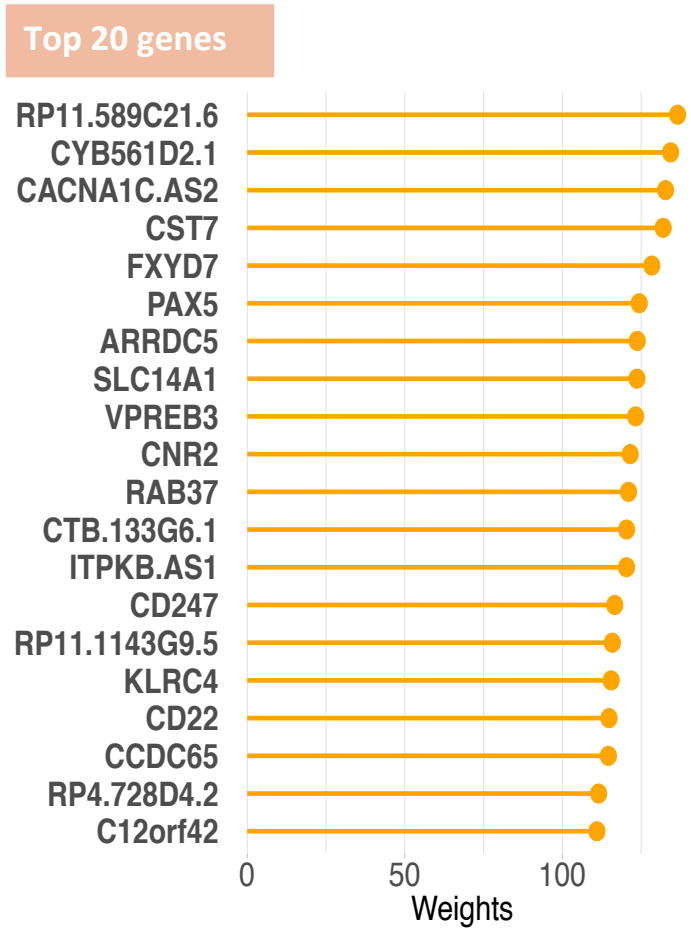


Epi GEM 54

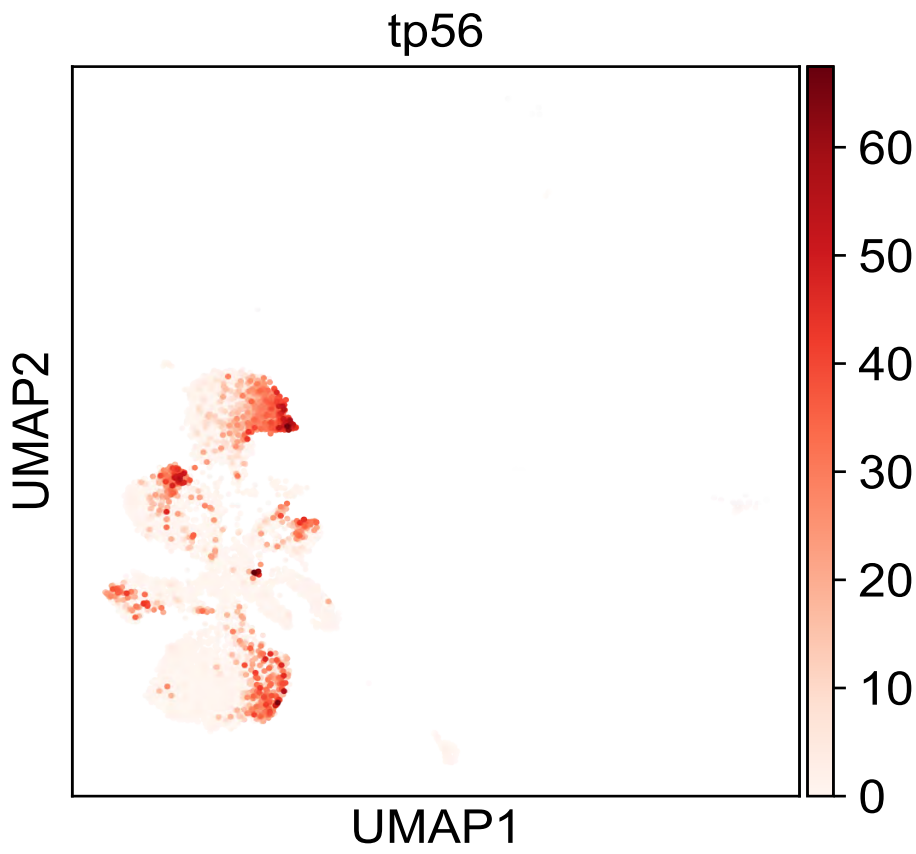


Cell Type

Epi_5



Epi GEM 56

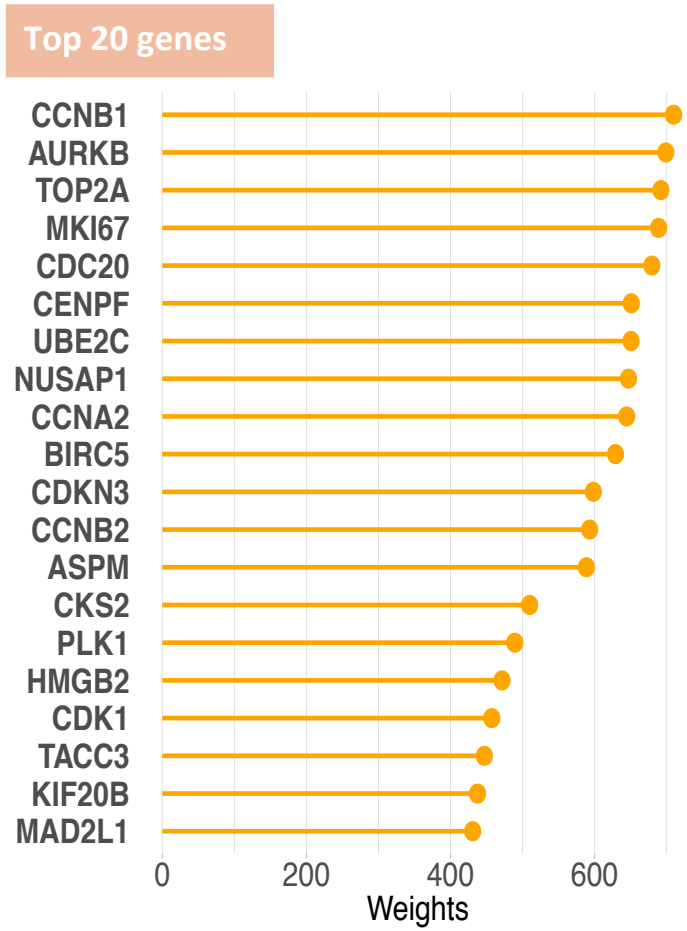


Cell Type

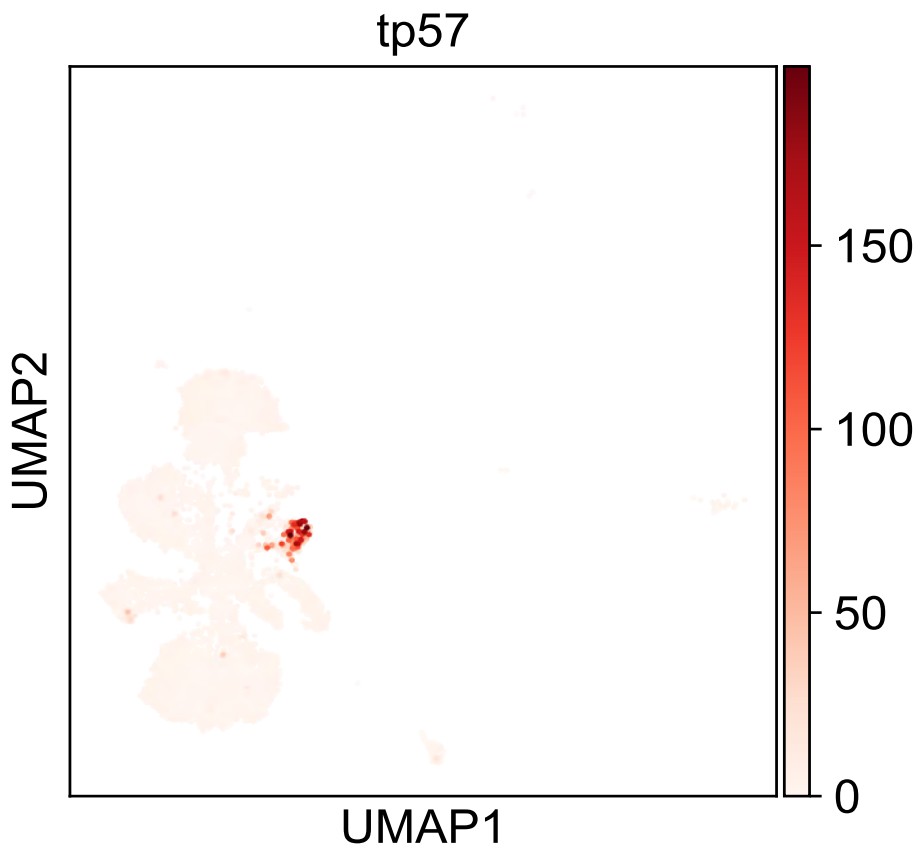
Epi_8

Epi_5

Epi_2

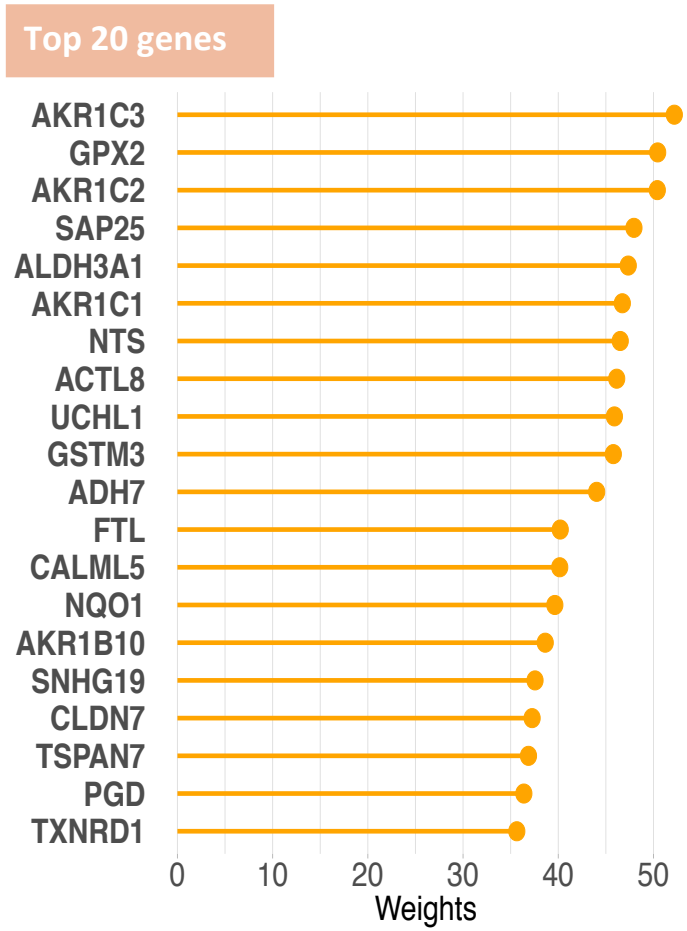


Epi GEM 57

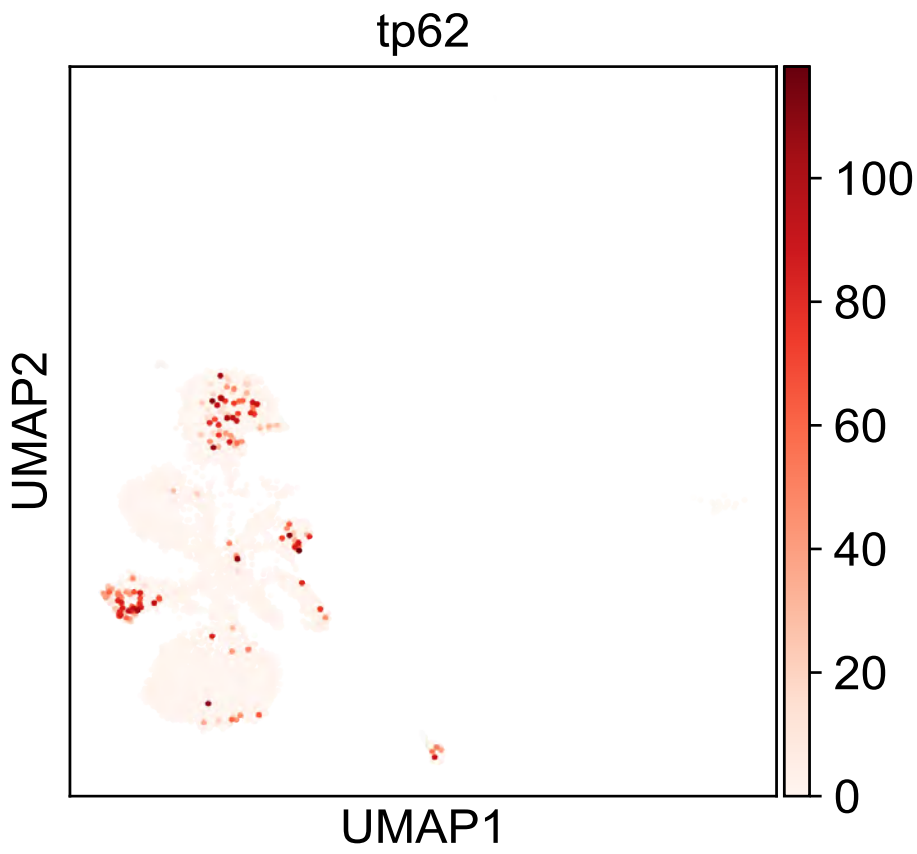


Cell Type

Epi_2



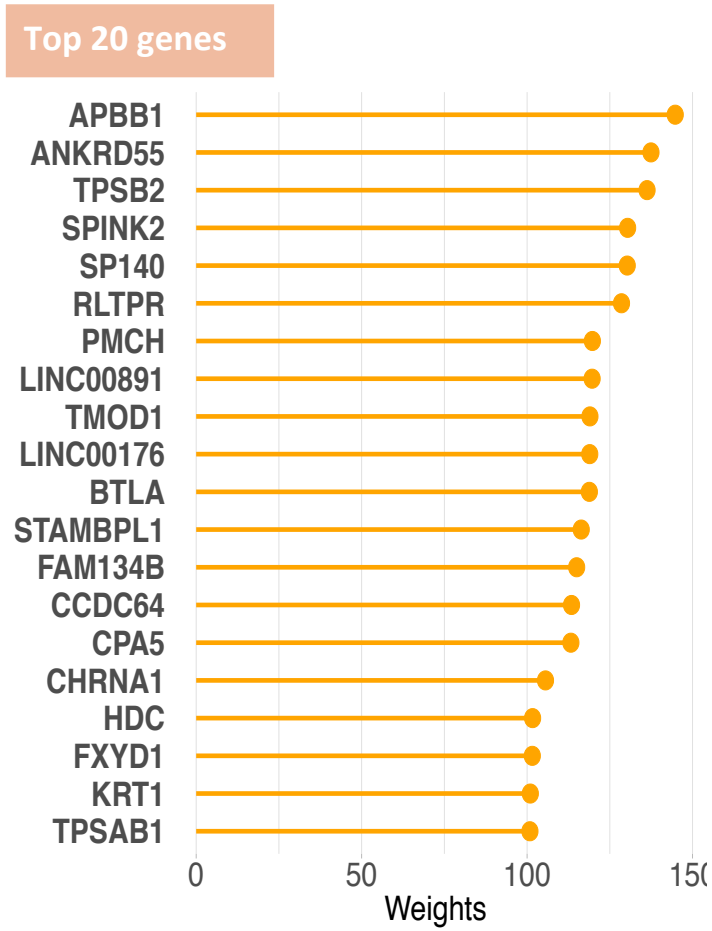
Epi GEM 62



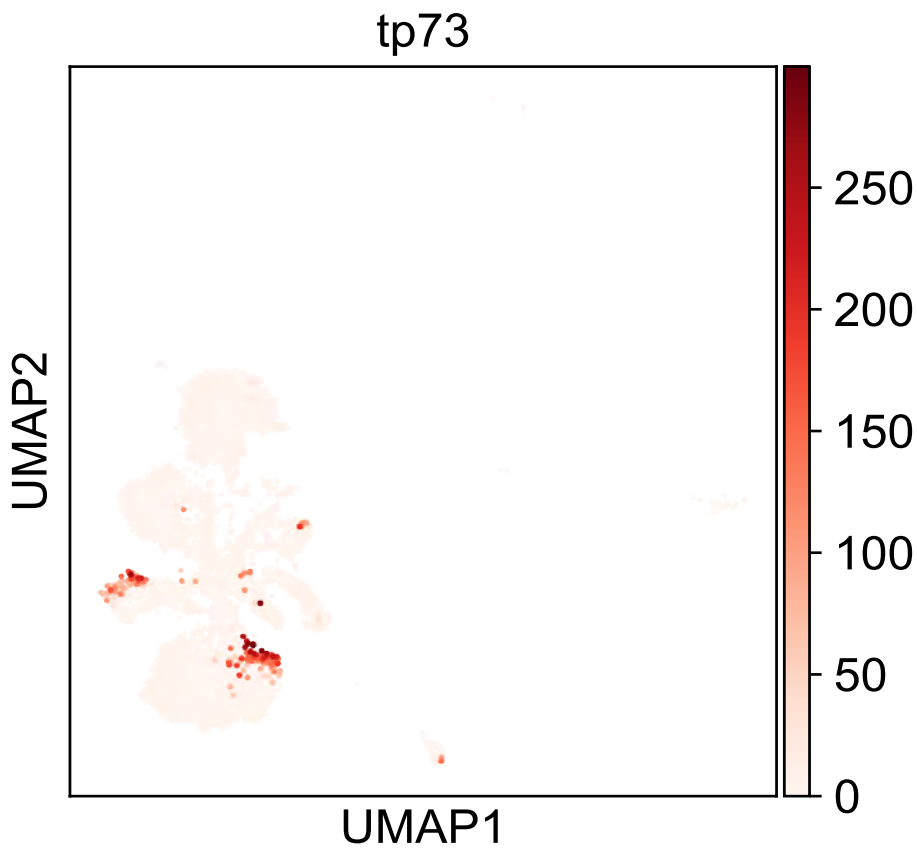
Cell Type

Epi_8

Epi_2



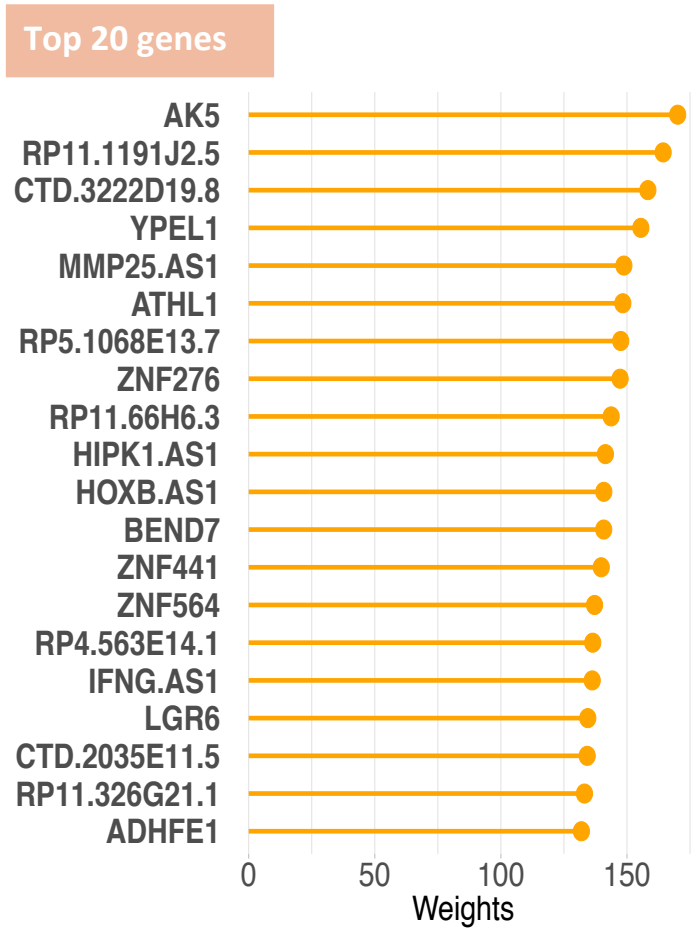
Epi GEM 73



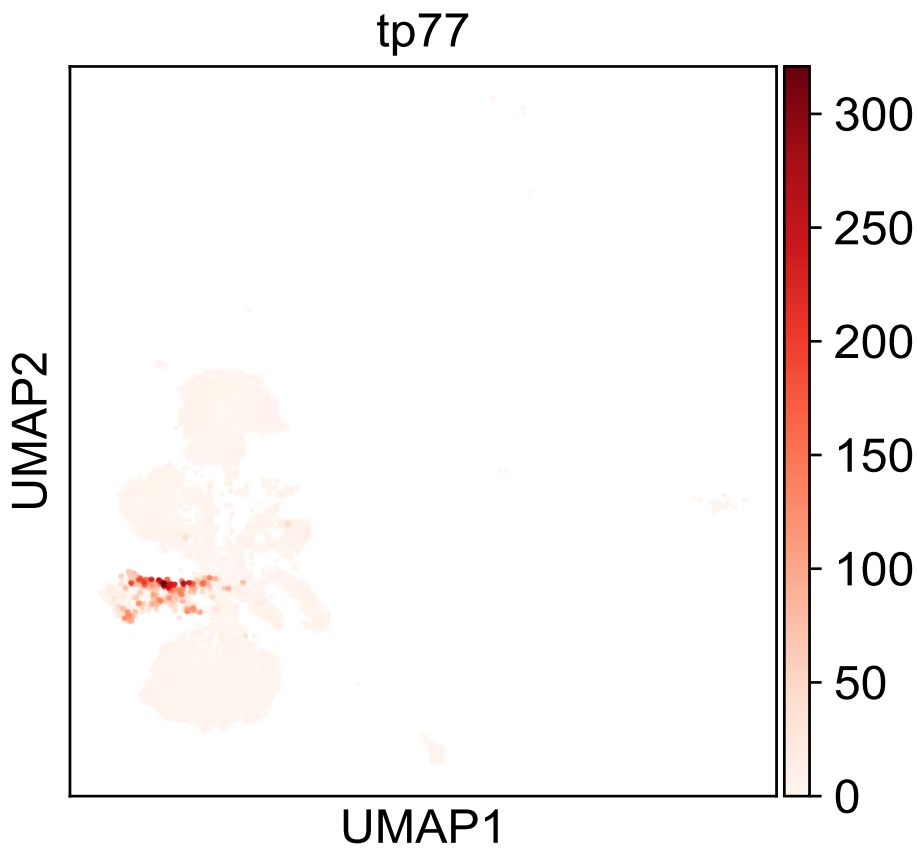
Cell Type

Epi_2

Epi_5



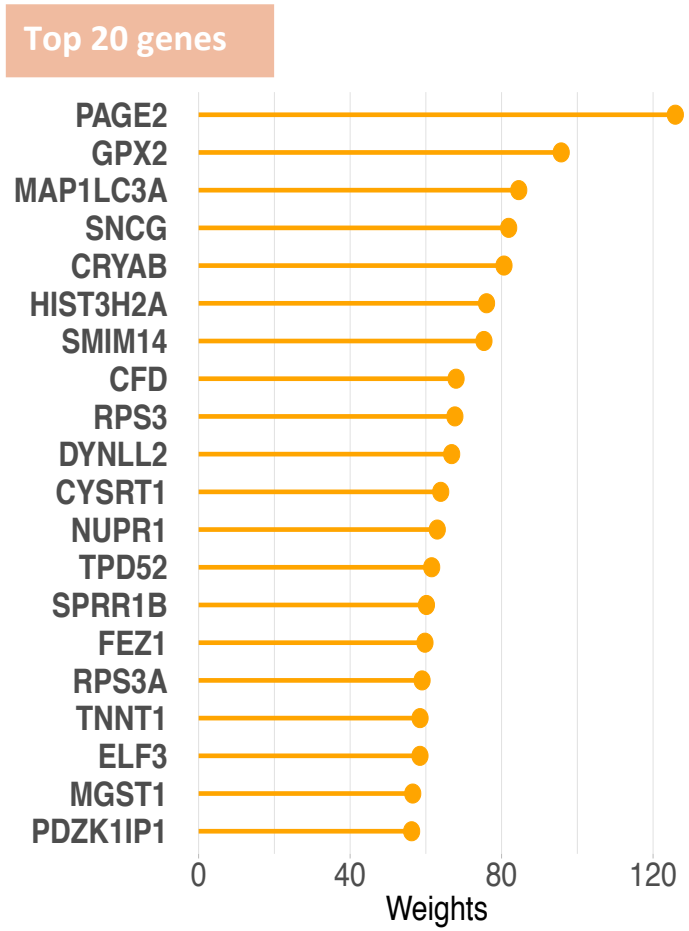
Epi GEM 77



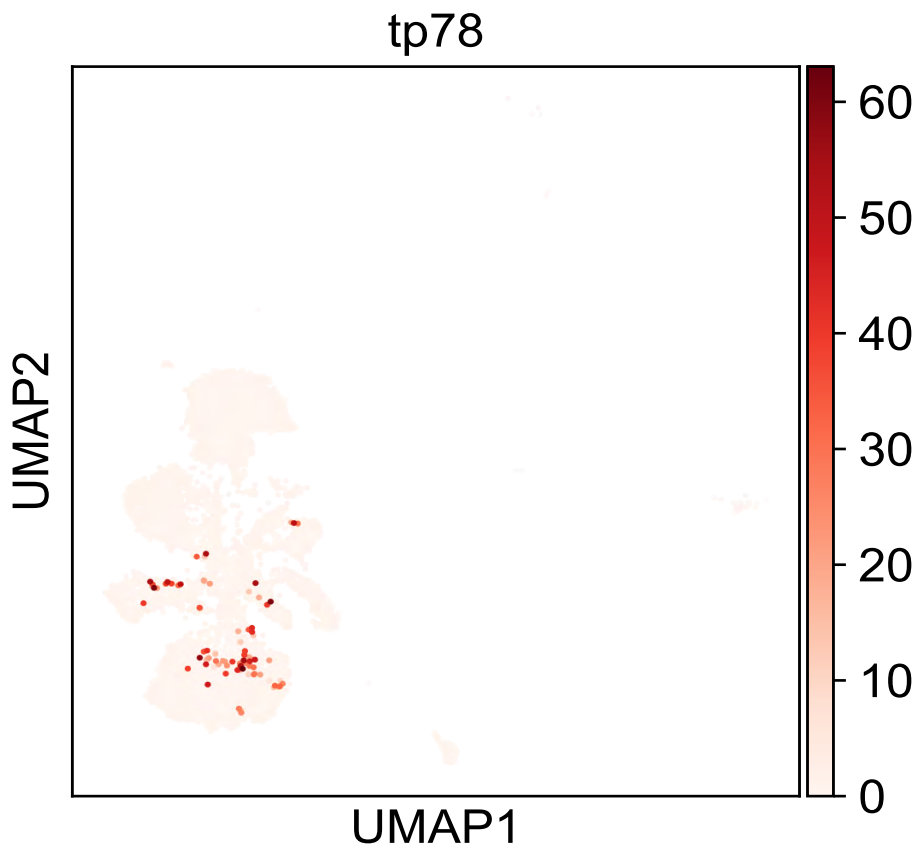
Cell Type

Epi_2

Epi_3



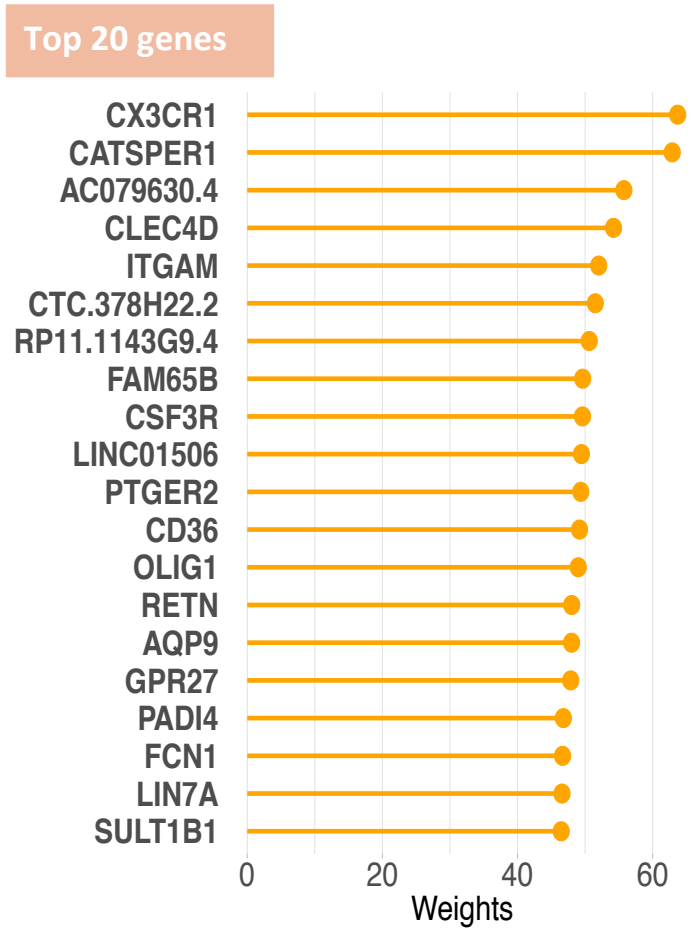
Epi GEM 78



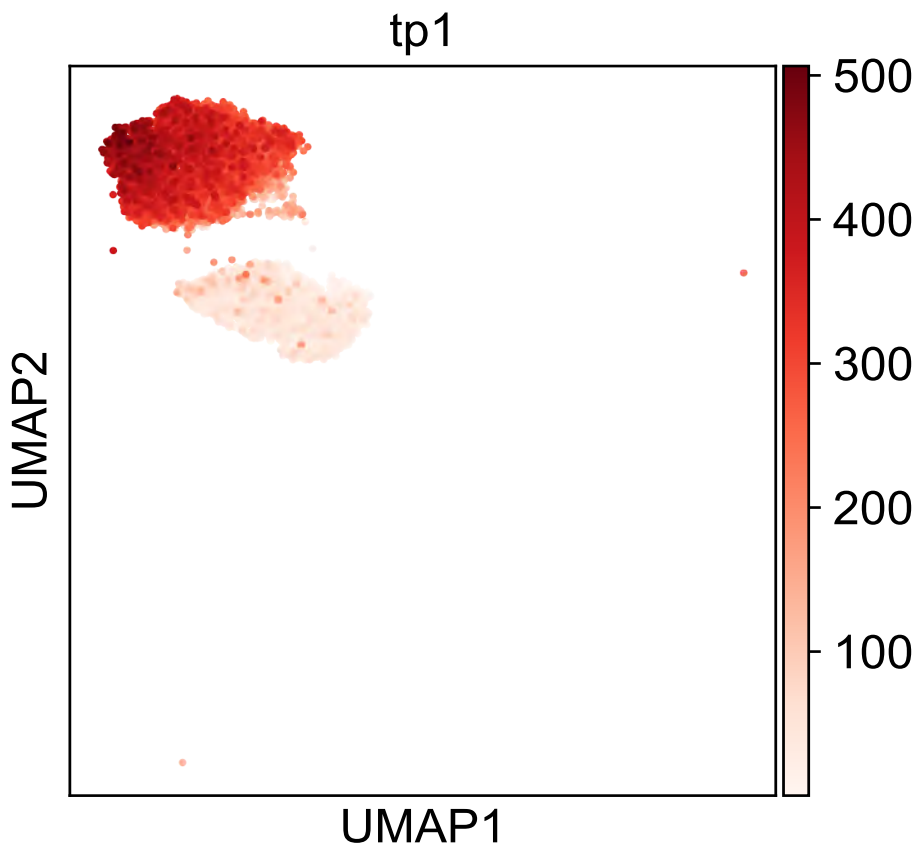
Cell Type

Epi_4

Epi_5

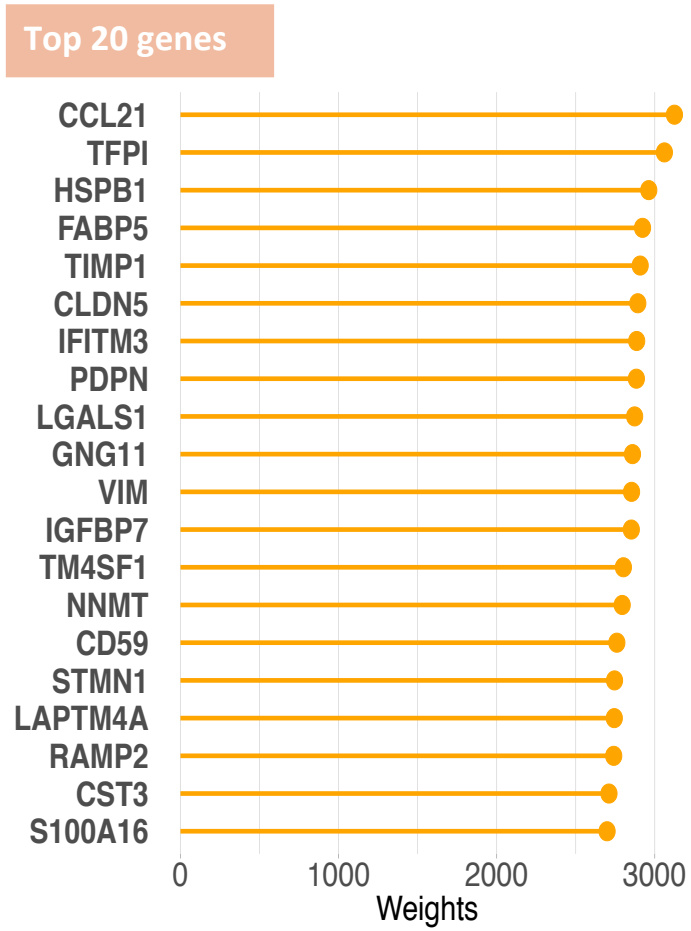


Endo GEM 1

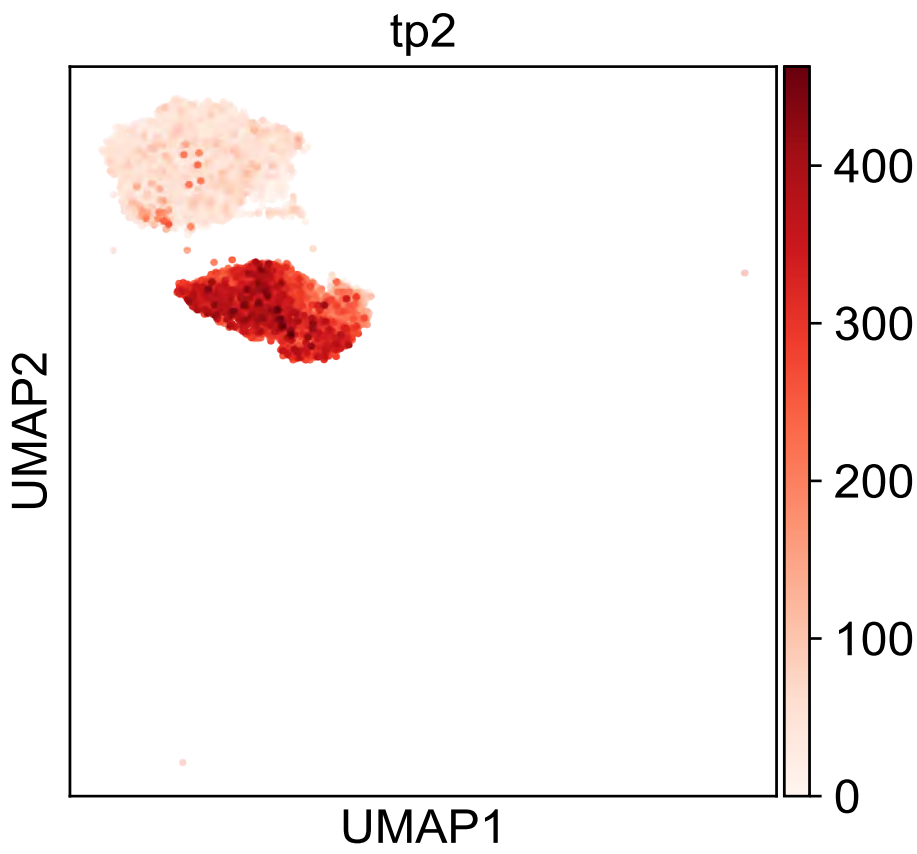


Cell Type

- Endo_1
- Endo_4
- Endo_5
- Endo_6



Endo GEM 2

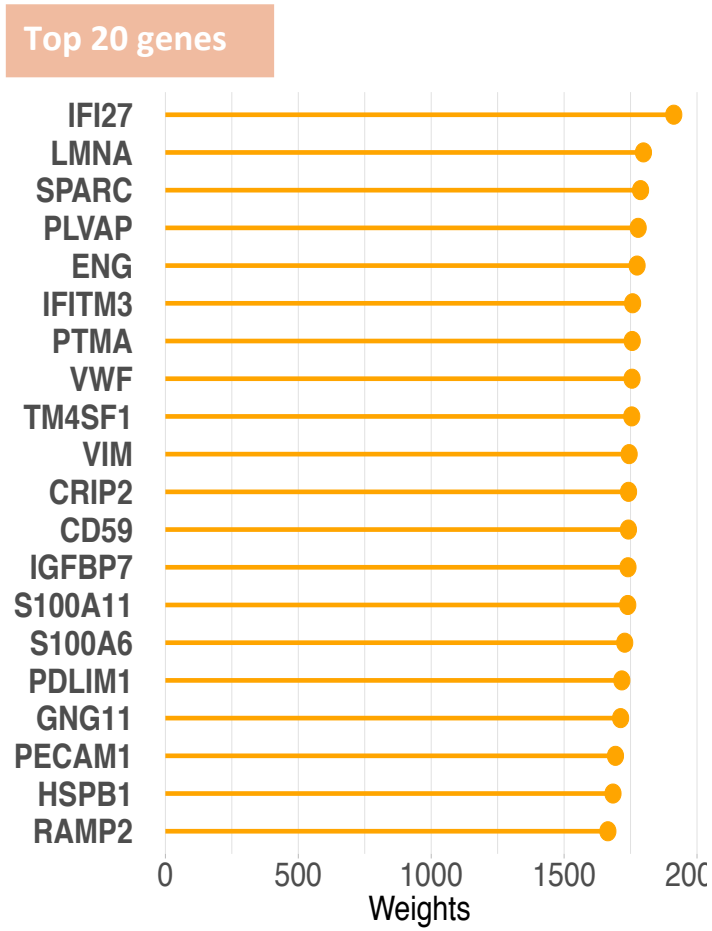


Cell Type

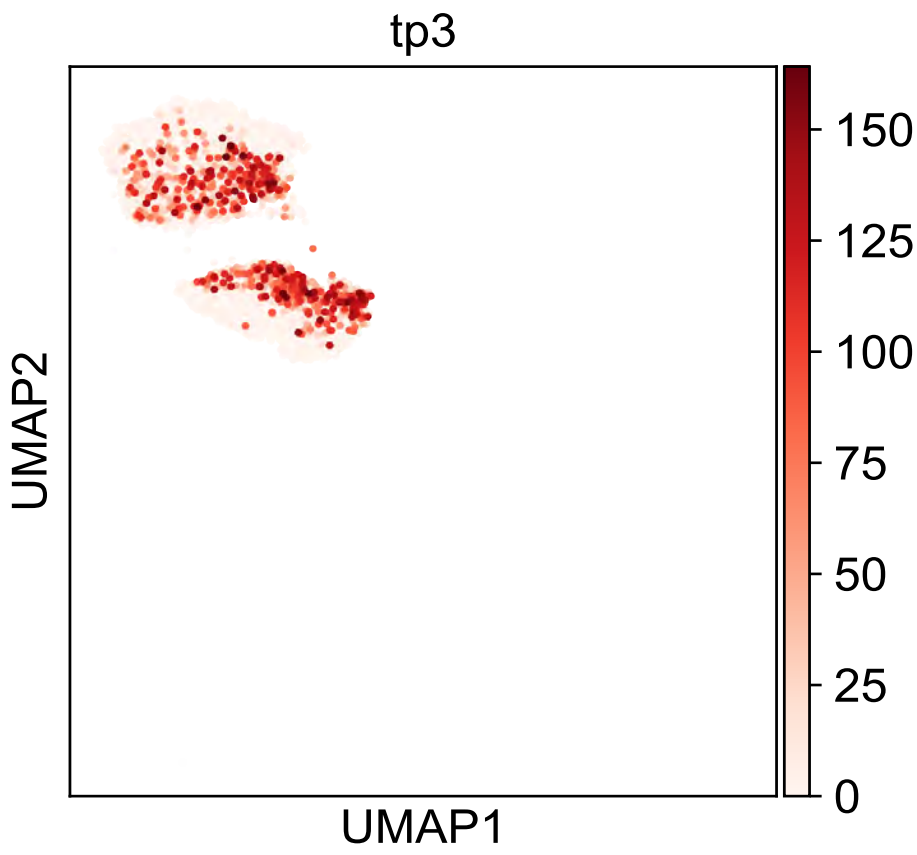
Endo_2

Endo_3

Endo_6

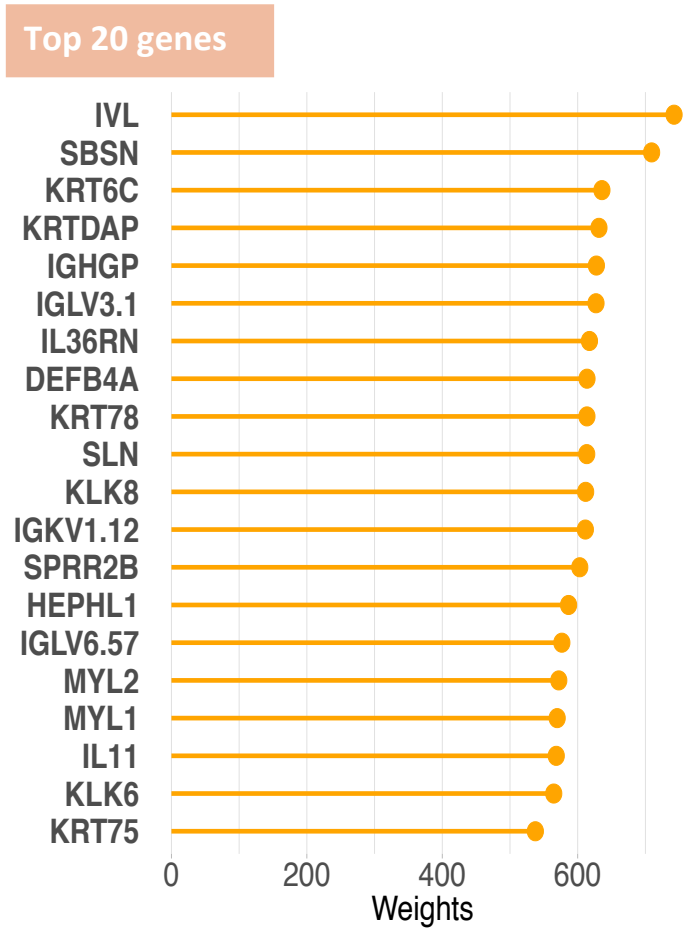


Endo GEM 3

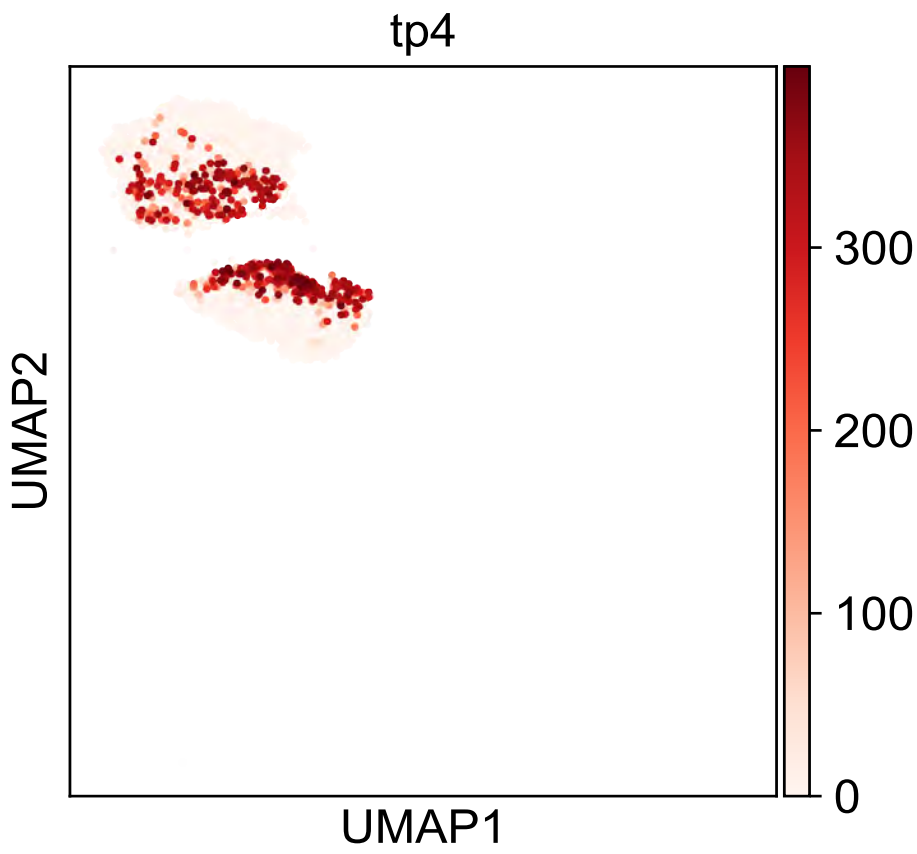


Cell Type

Endo_6



Endo GEM 4

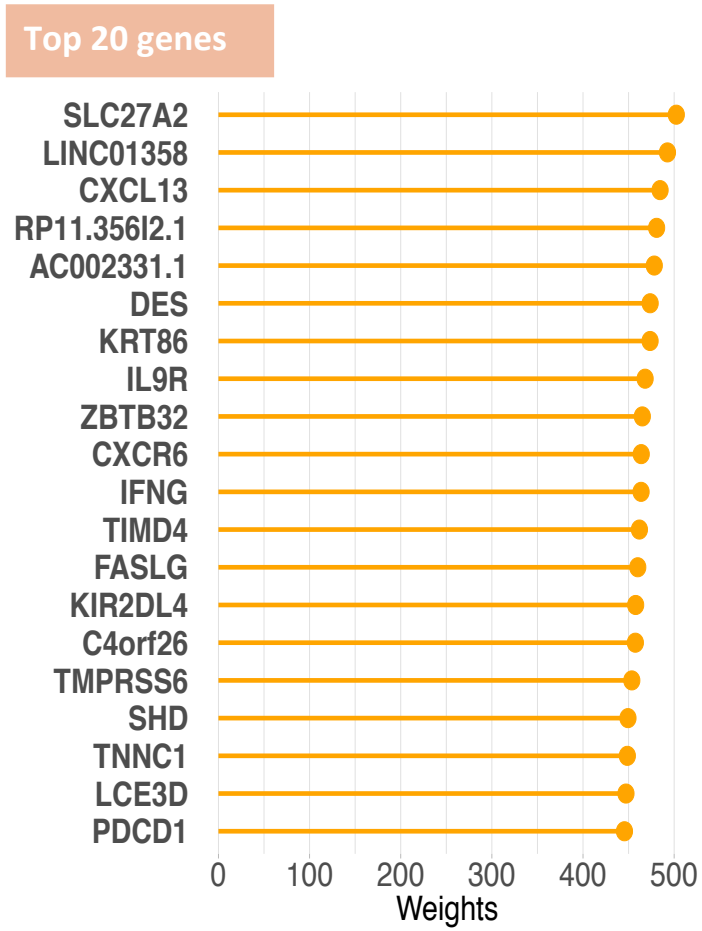


Cell Type

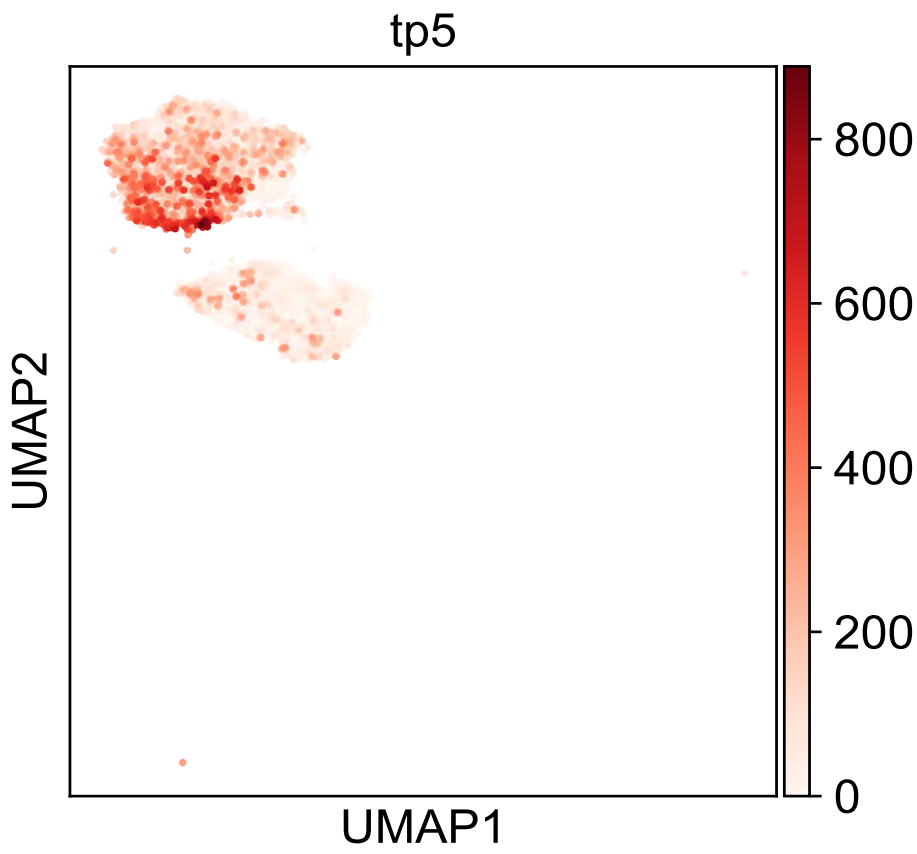
Endo_1

Endo_5

Endo_6



Endo GEM 5

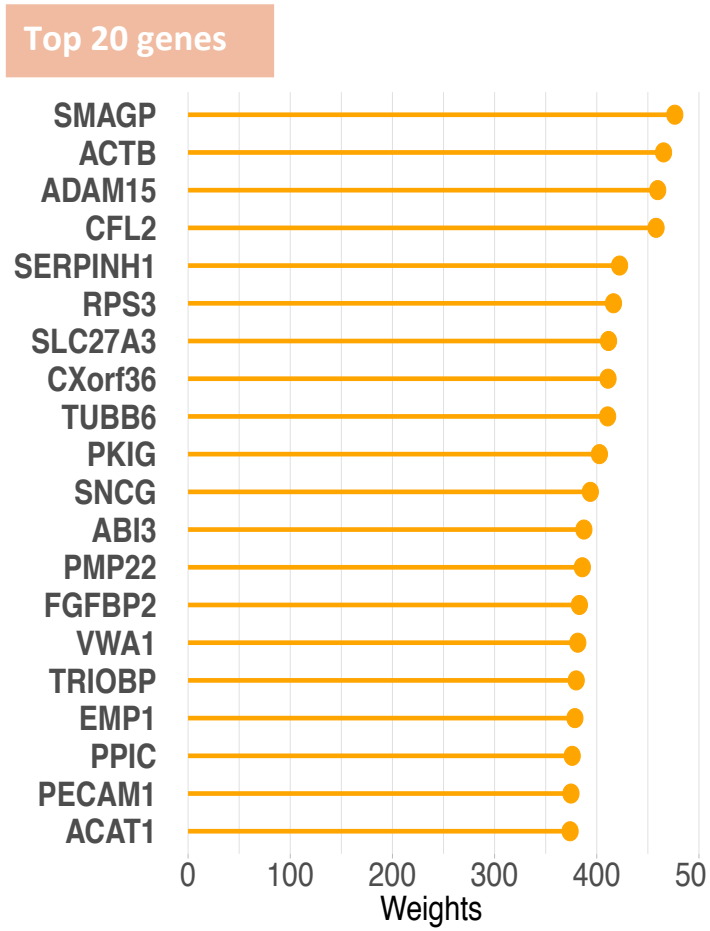


Cell Type

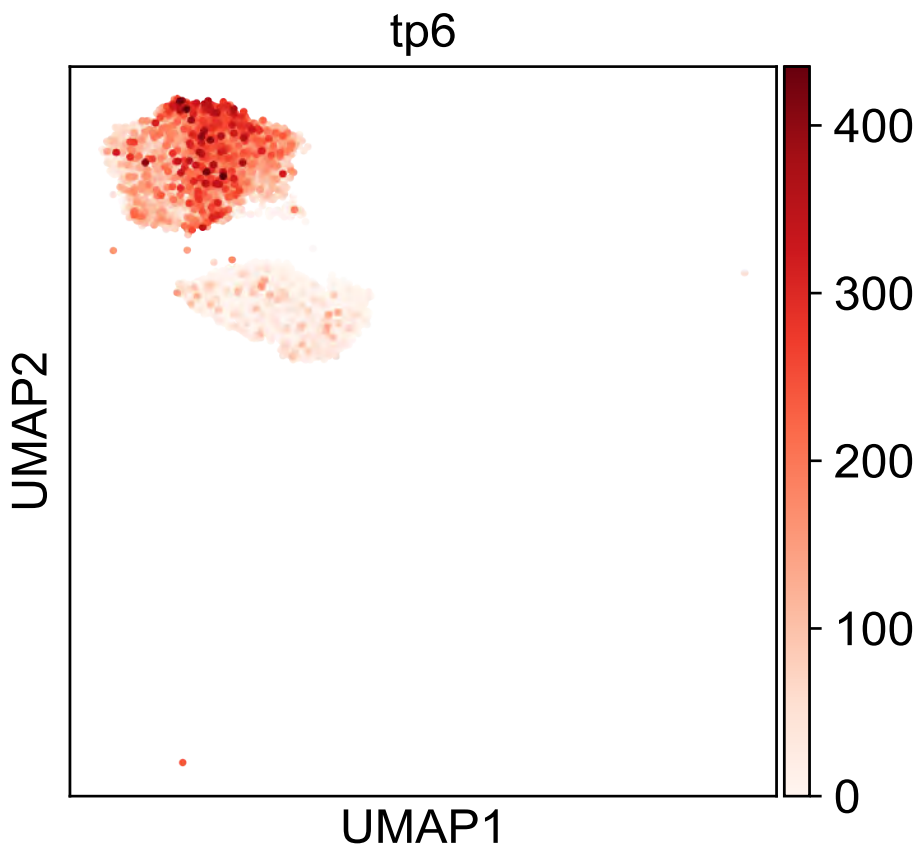
Endo_1

Endo_5

Endo_6

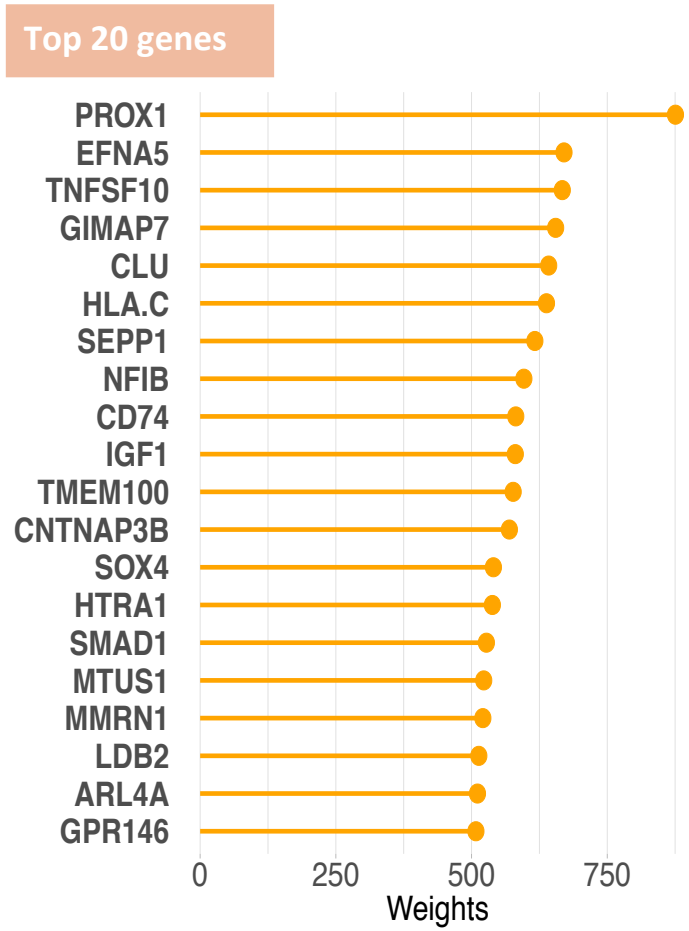


Endo GEM 6

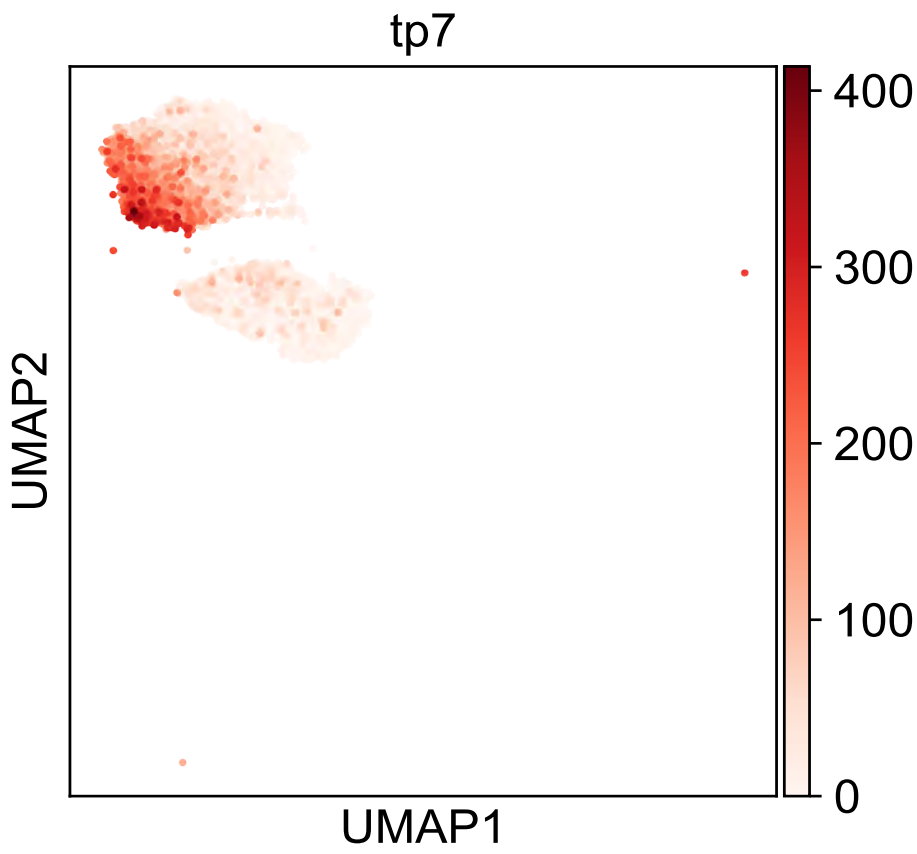


Cell Type

- Endo_1
- Endo_4
- Endo_5
- Endo_6



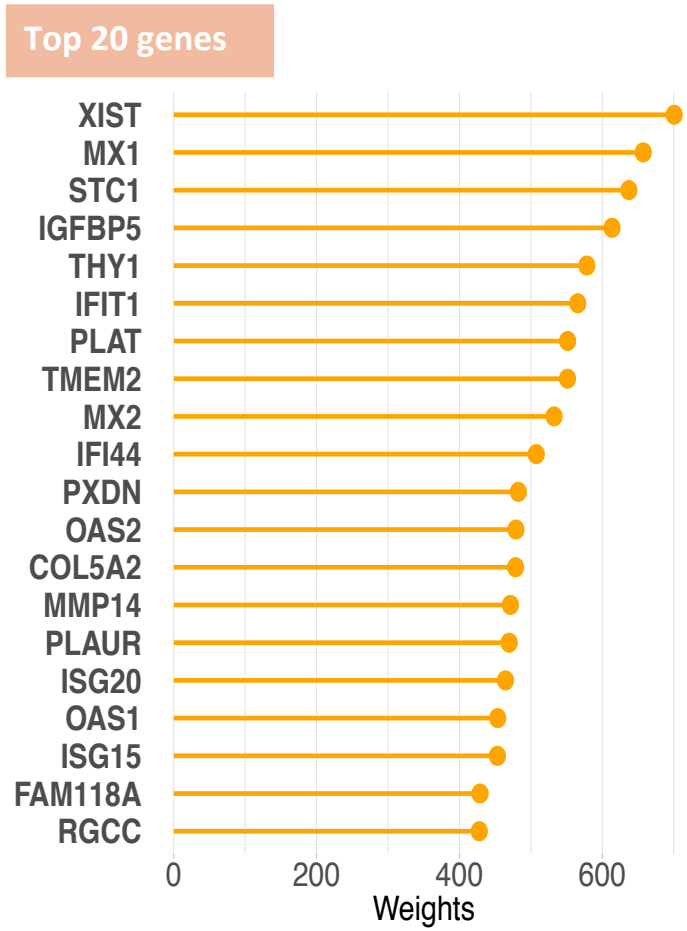
Endo GEM 7



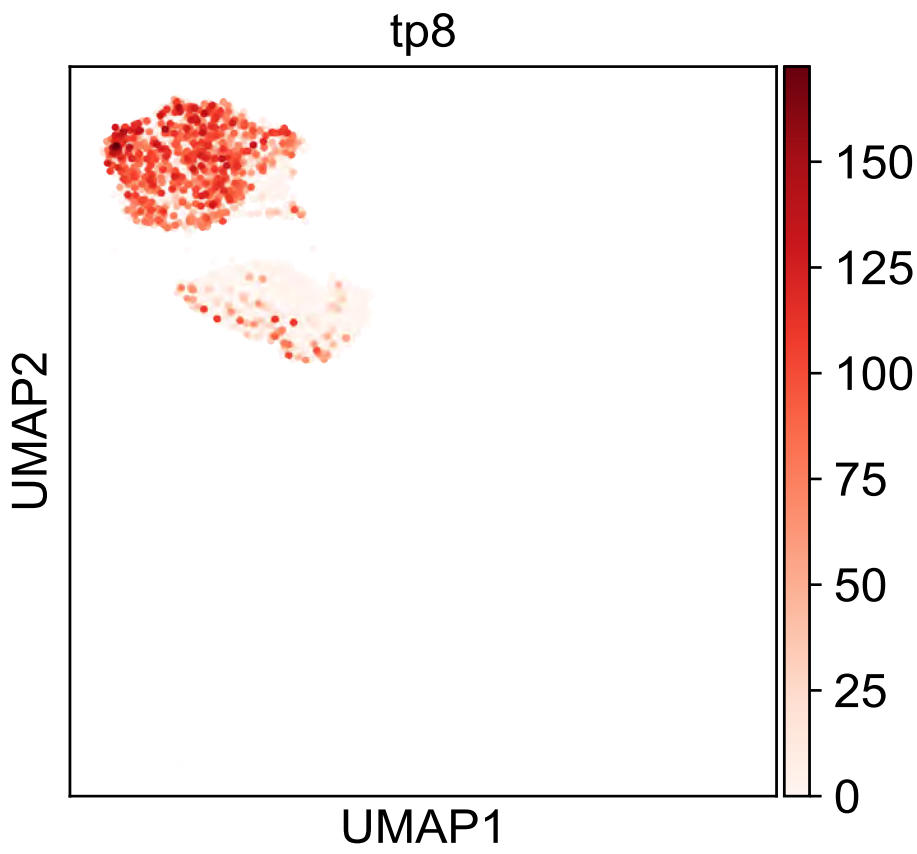
Cell Type

Endo_1

Endo_5



Endo GEM 8



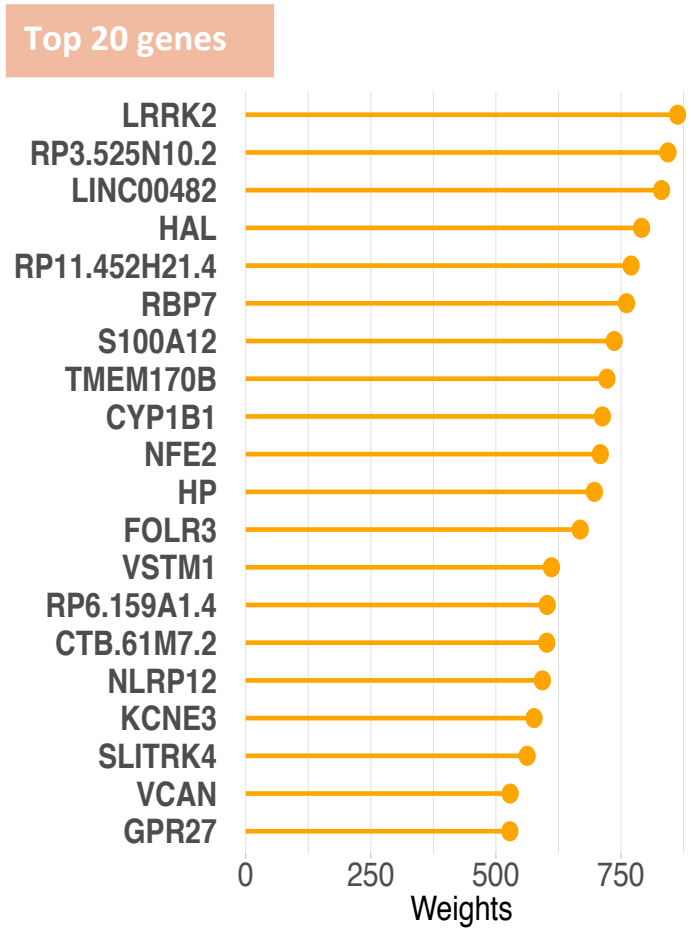
Cell Type

Endo_1

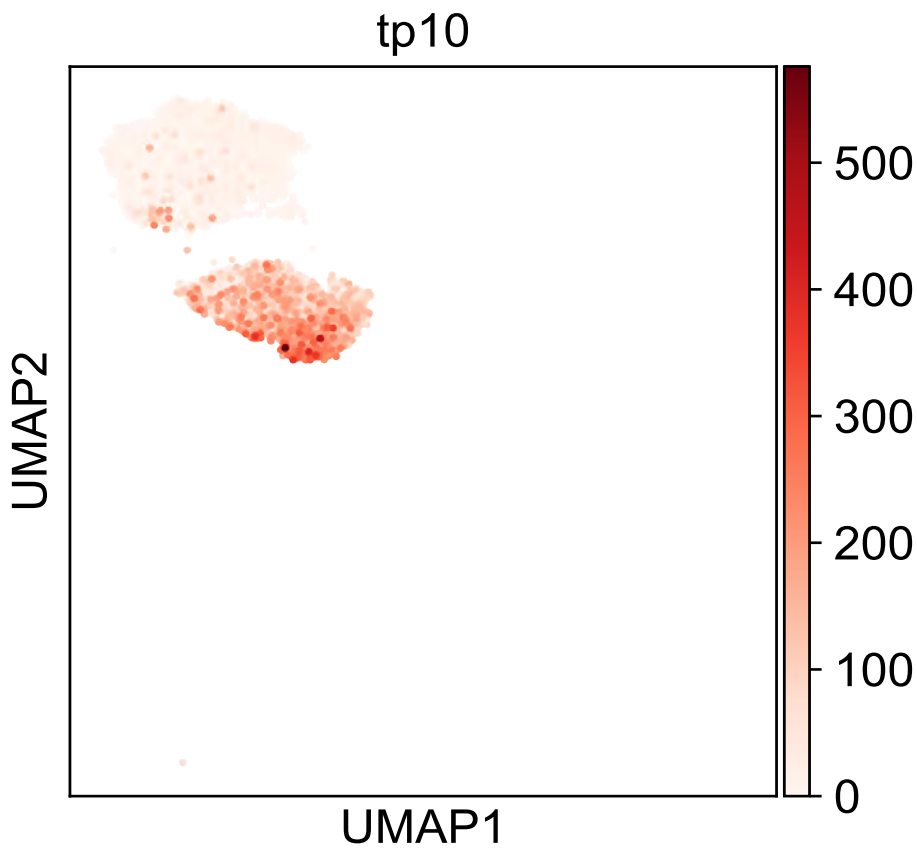
Endo_4

Endo_5

Endo_6



Endo GEM 10

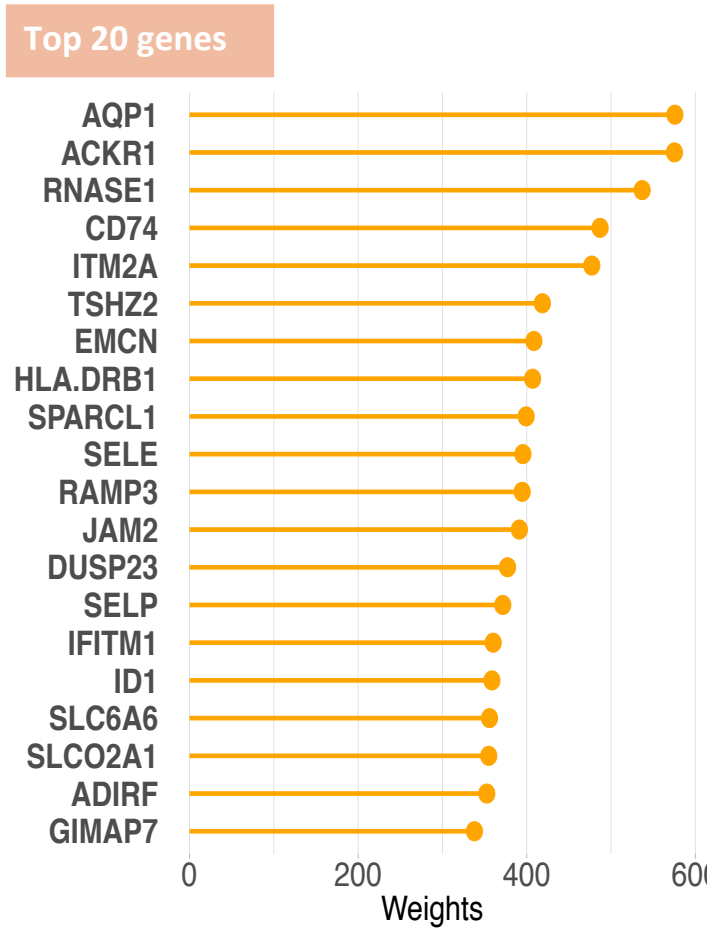


Cell Type

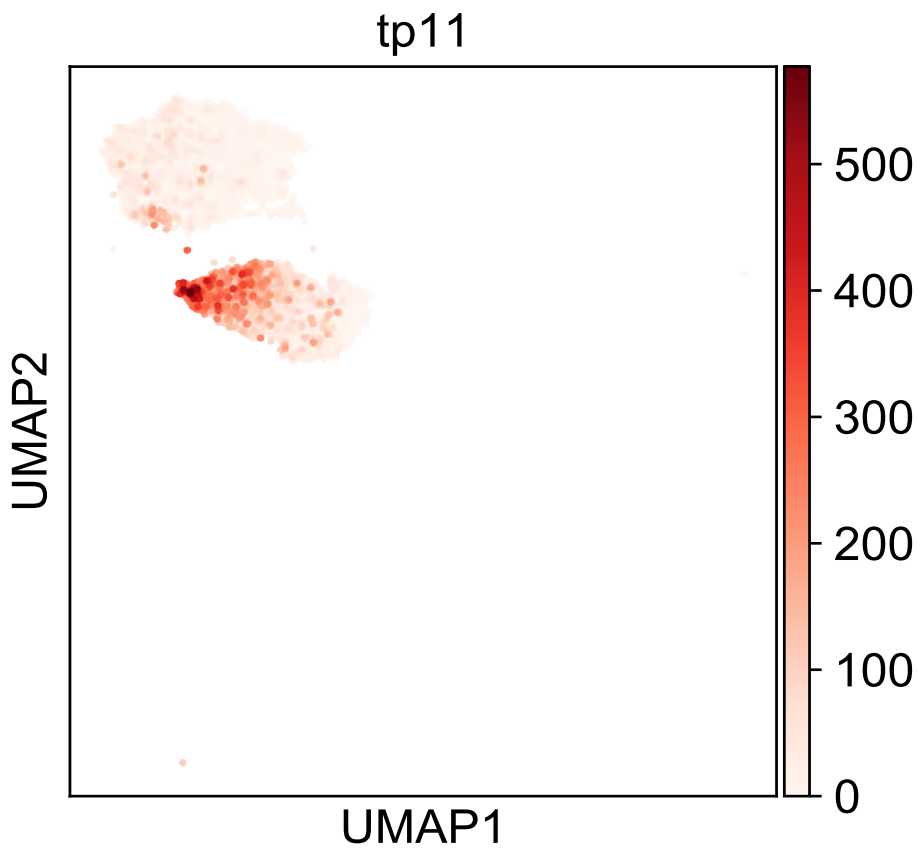
Endo_2

Endo_3

Endo_6



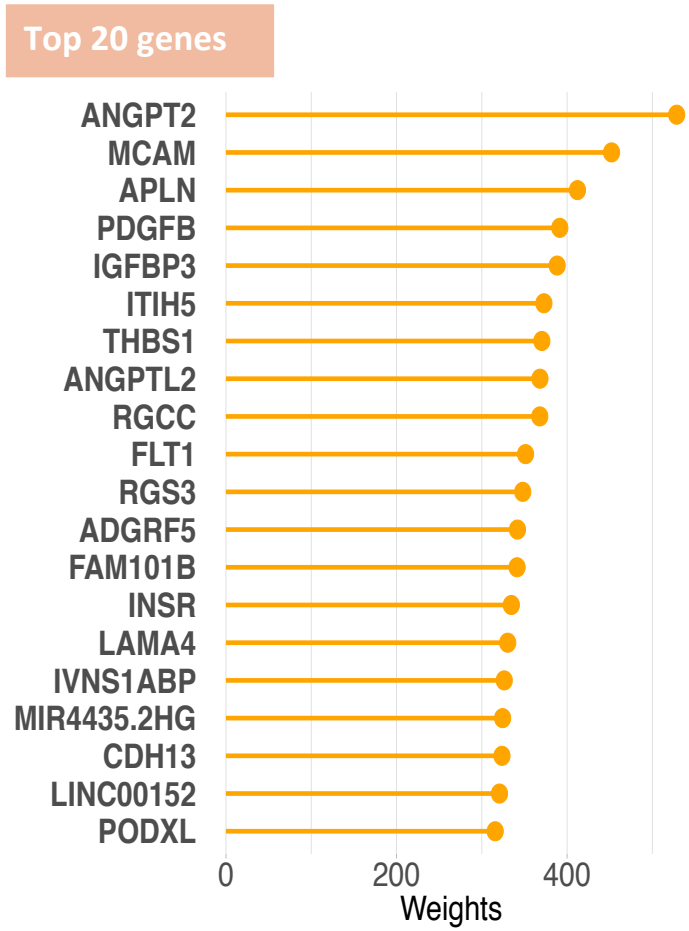
Endo GEM 11



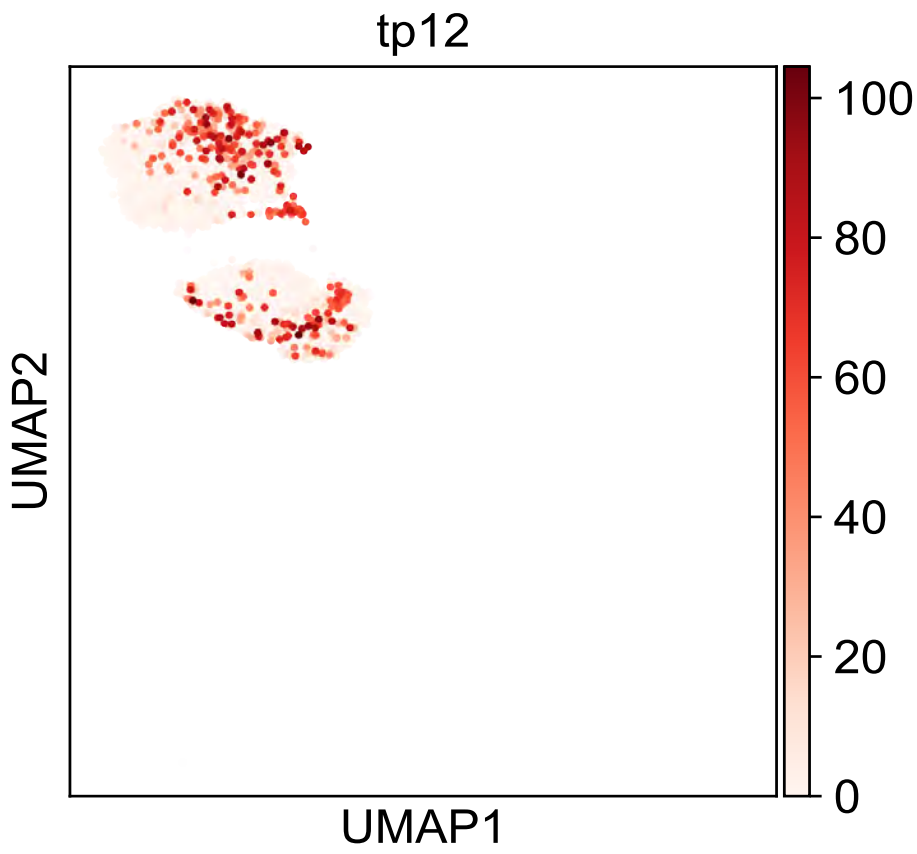
Cell Type

Endo_2

Endo_3



Endo GEM 12

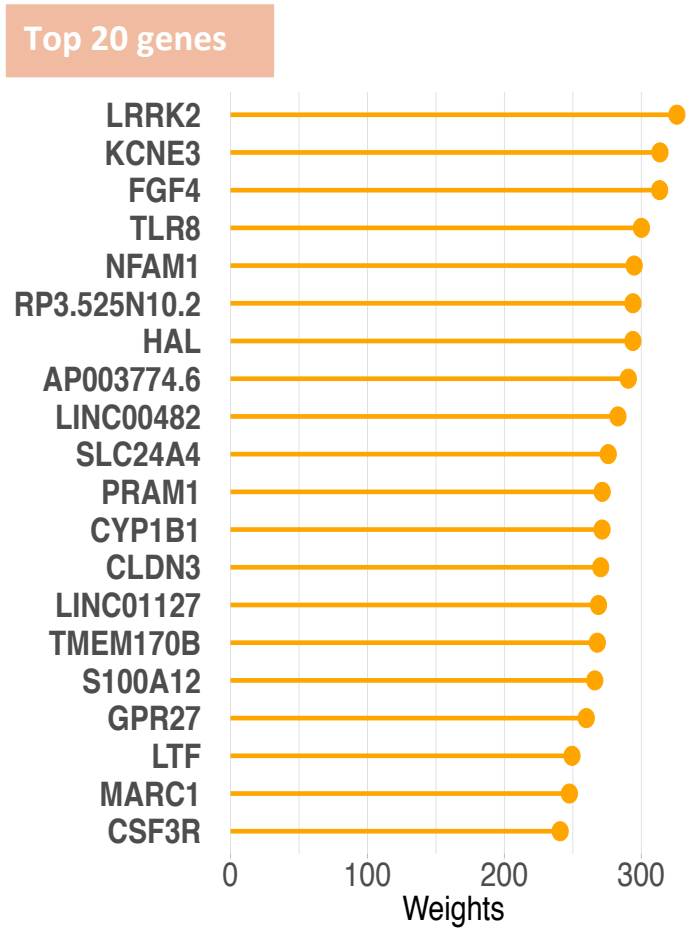


Cell Type

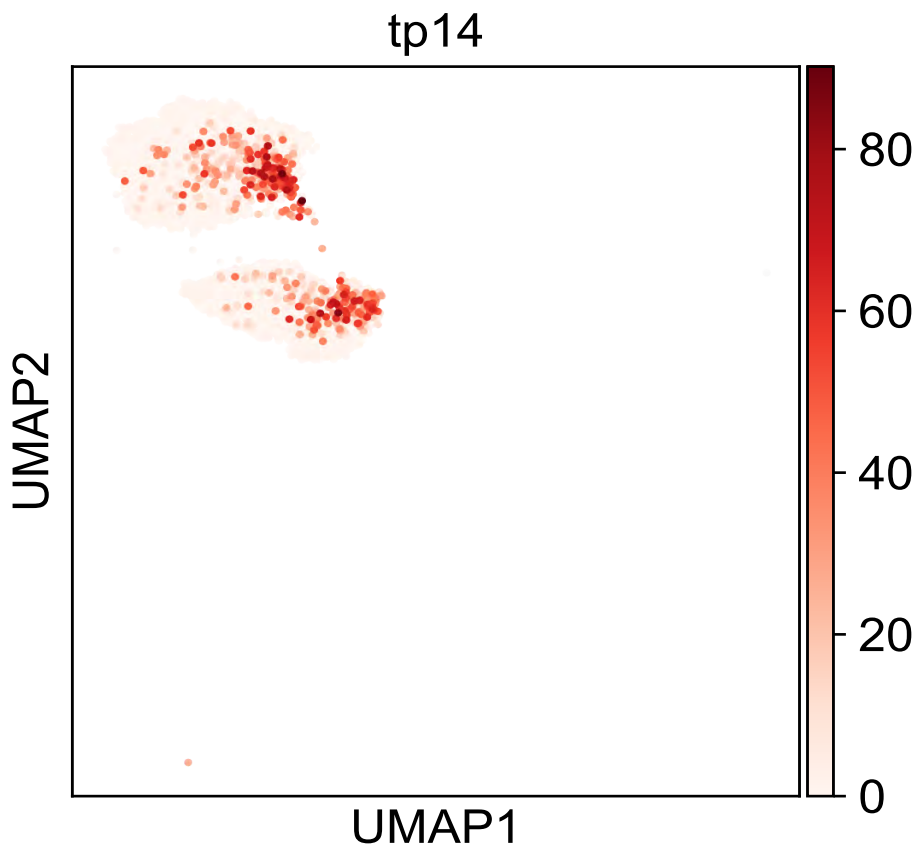
Endo_2

Endo_3

Endo_4



Endo GEM 14

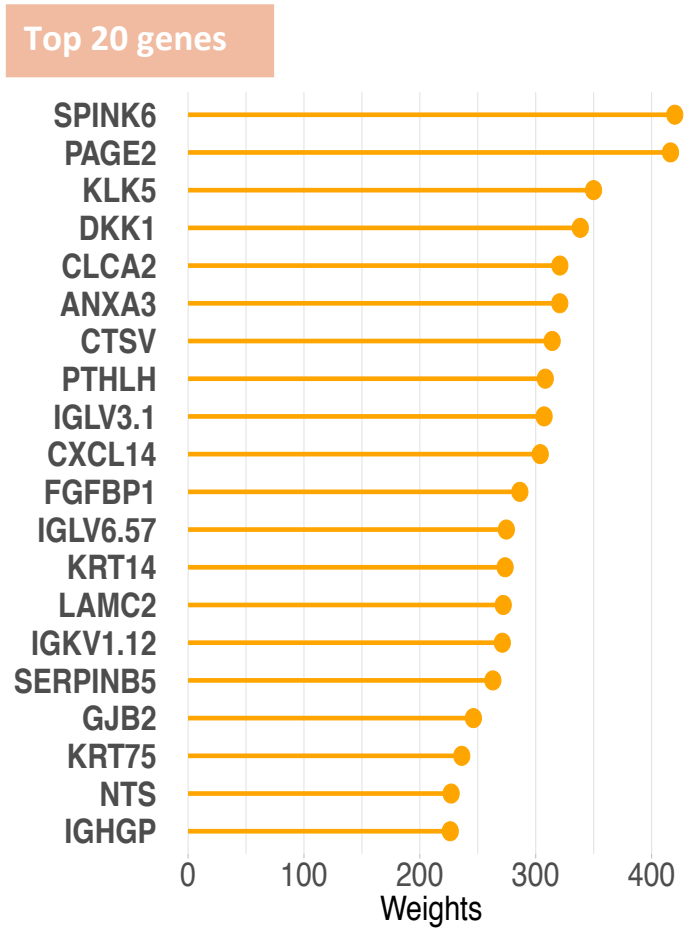


Cell Type

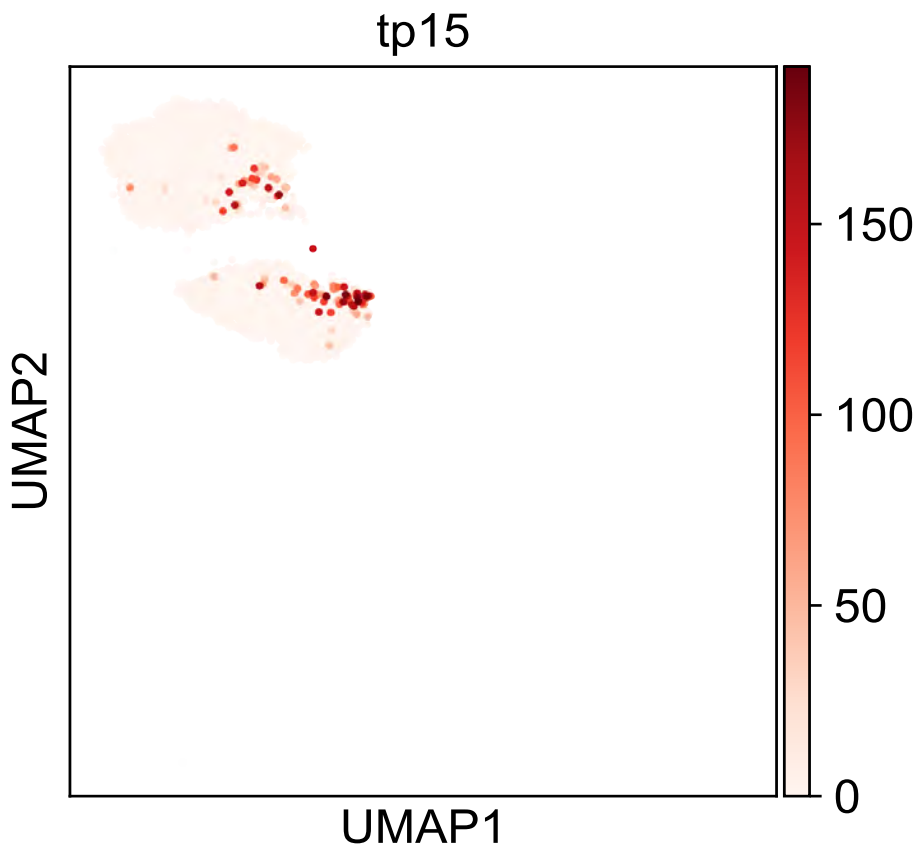
Endo_3

Endo_4

Endo_6

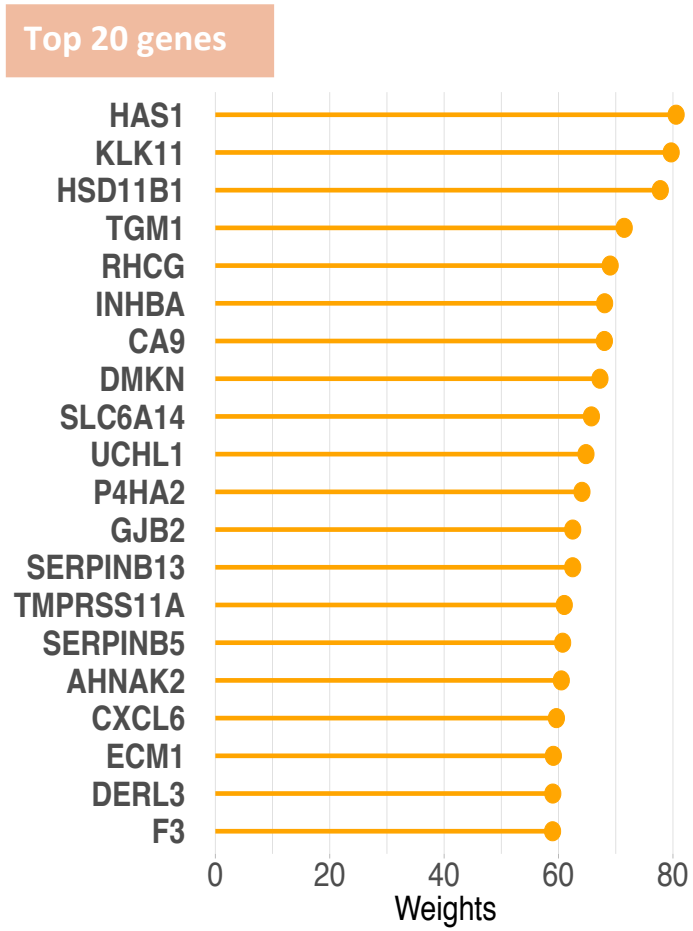


Endo GEM 15

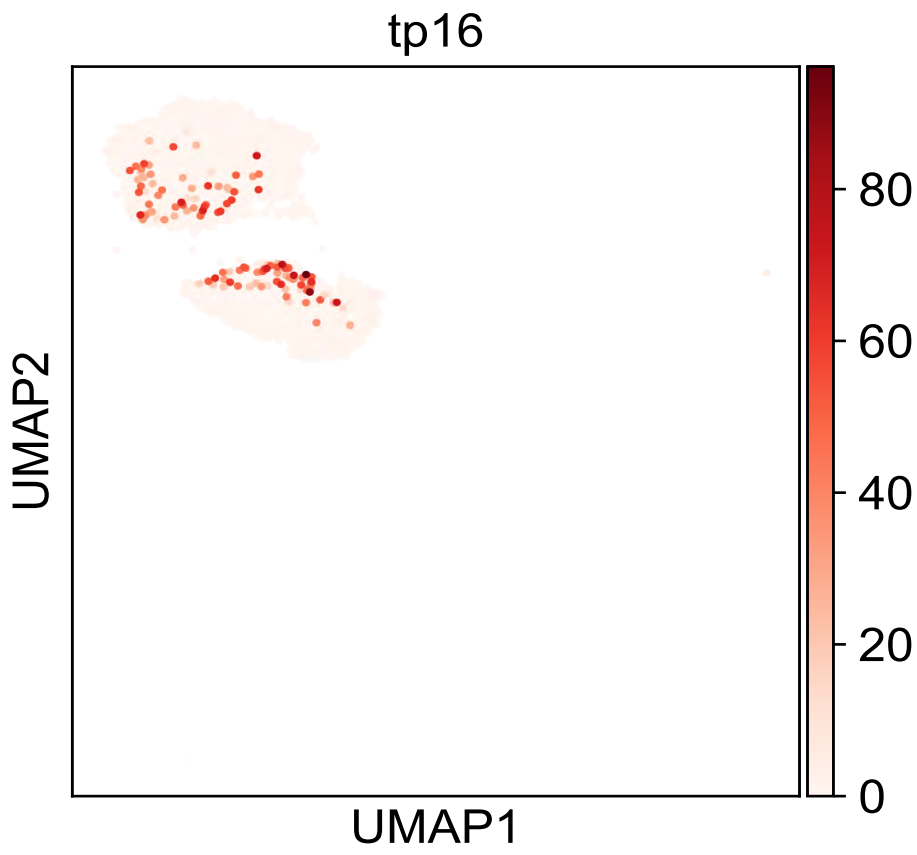


Cell Type

Endo_6



Endo GEM 16

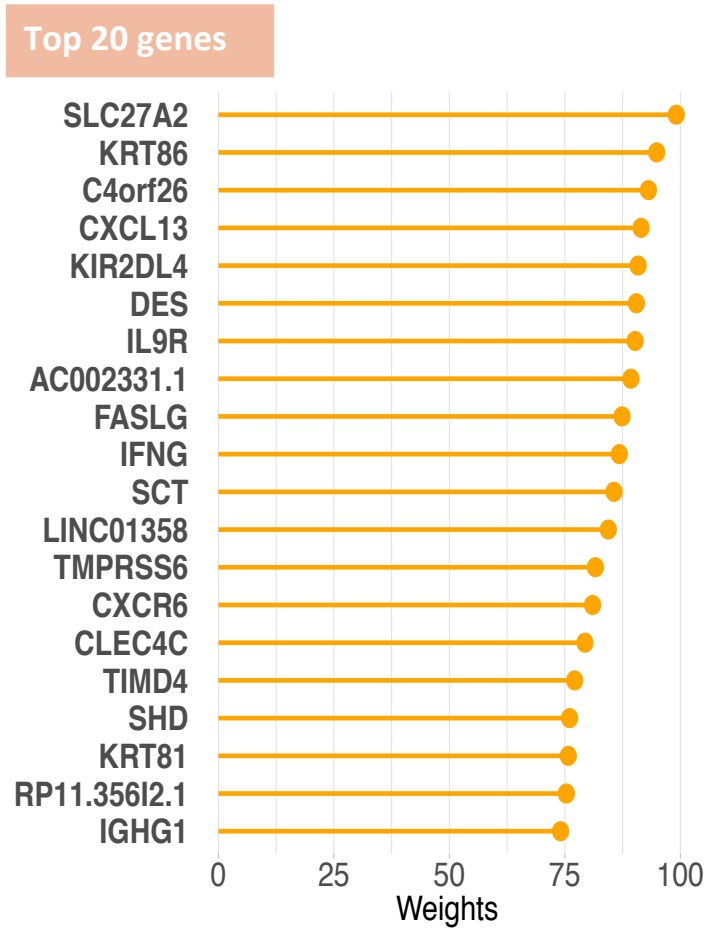


Cell Type

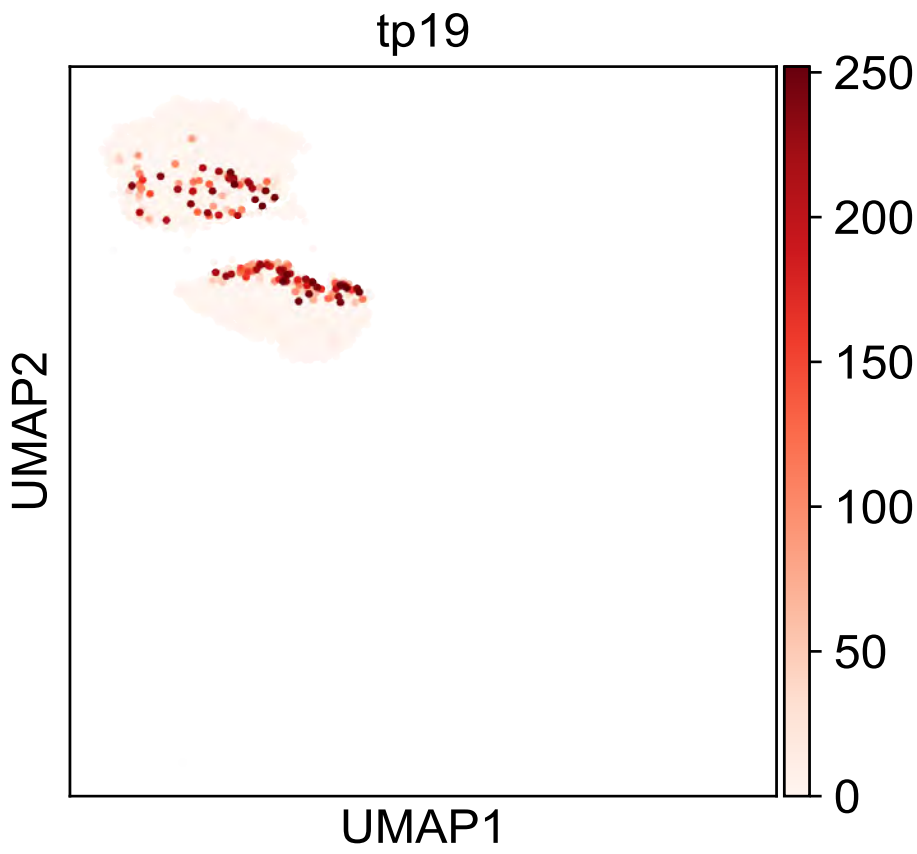
Endo_1

Endo_5

Endo_6



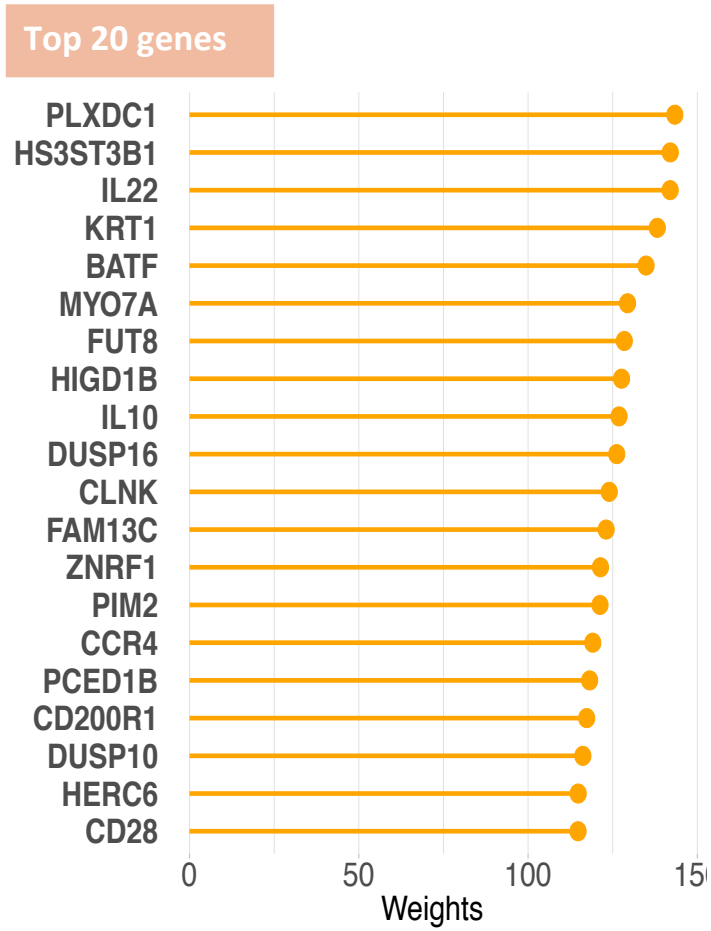
Endo GEM 19



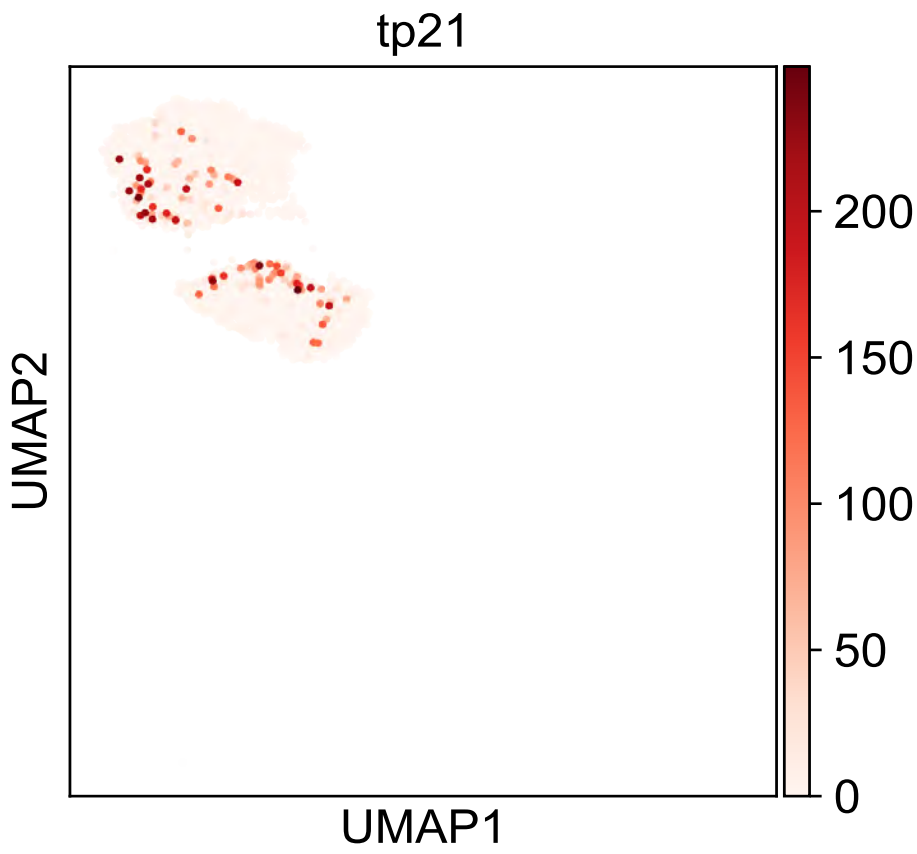
Cell Type

Endo_5

Endo_6



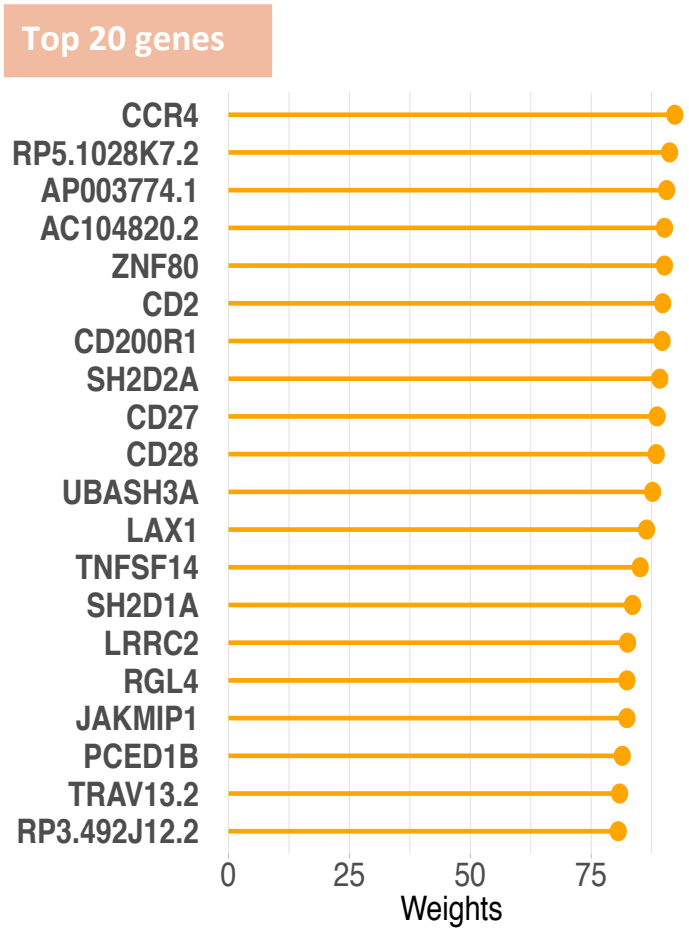
Endo GEM 21



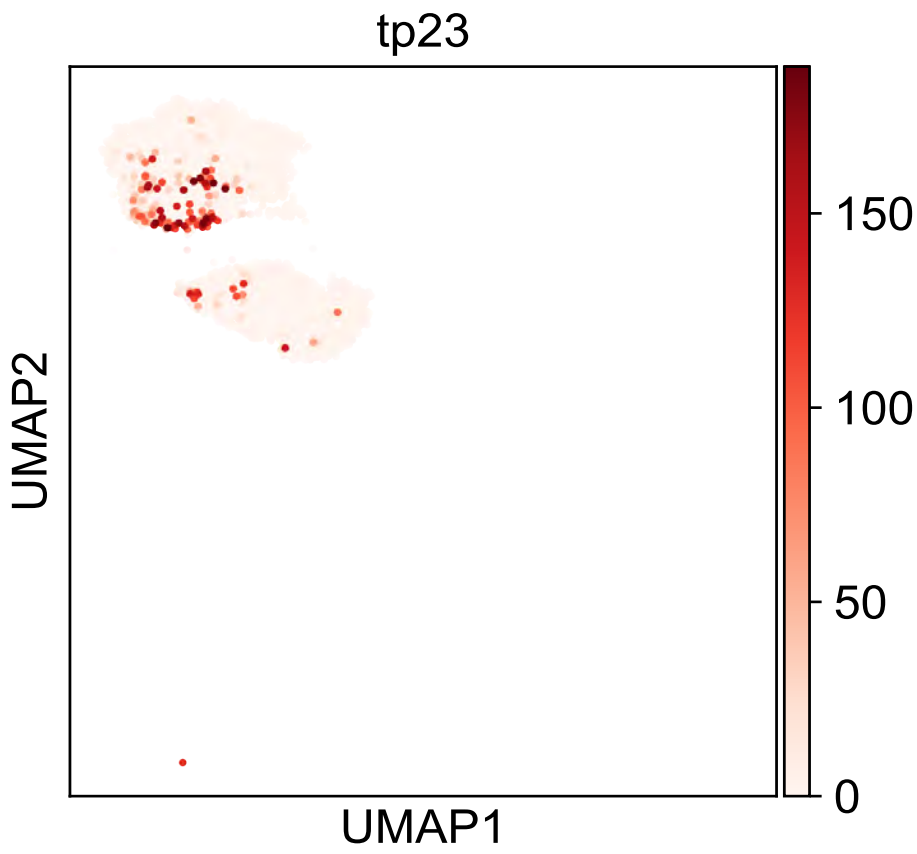
Cell Type

Endo_1

Endo_6

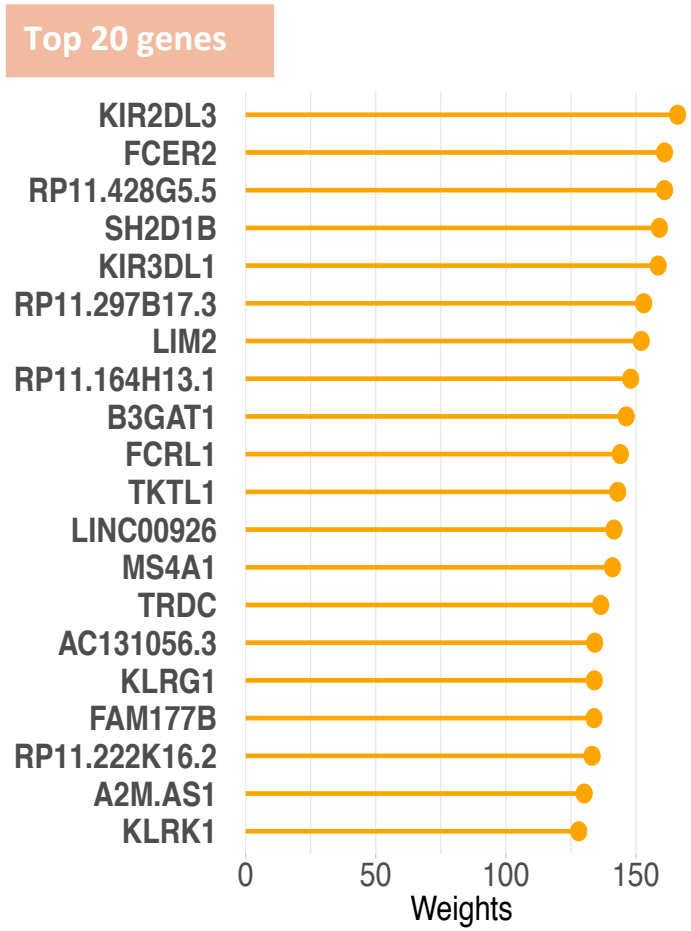


Endo GEM 23

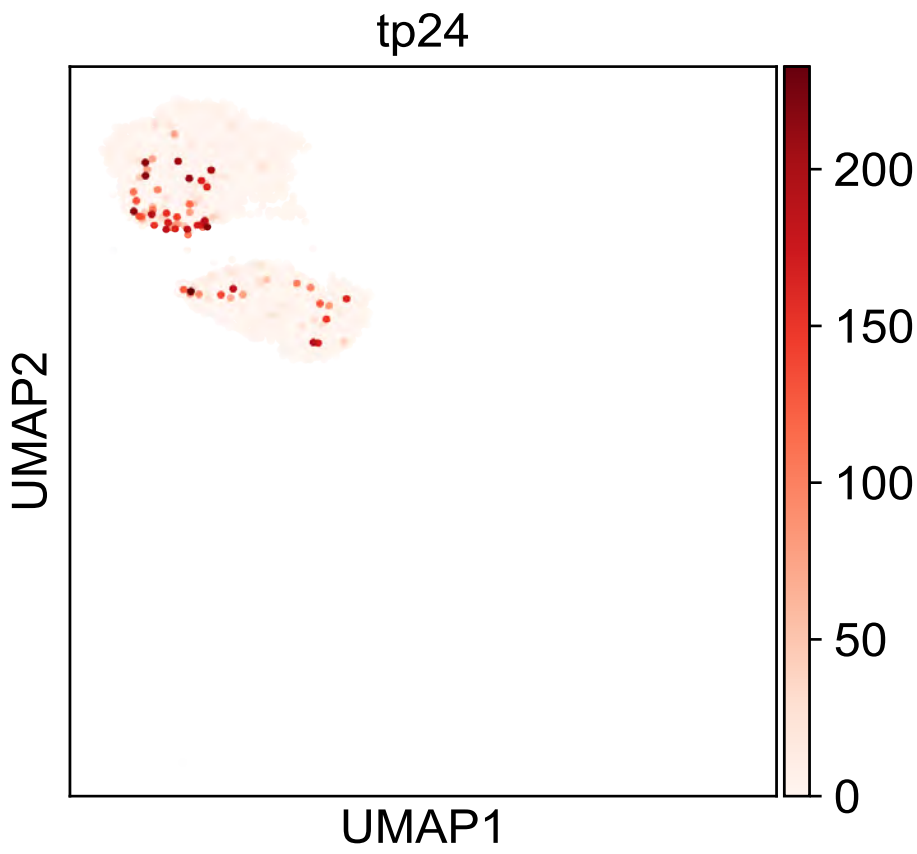


Cell Type

Endo_5



Endo GEM 24

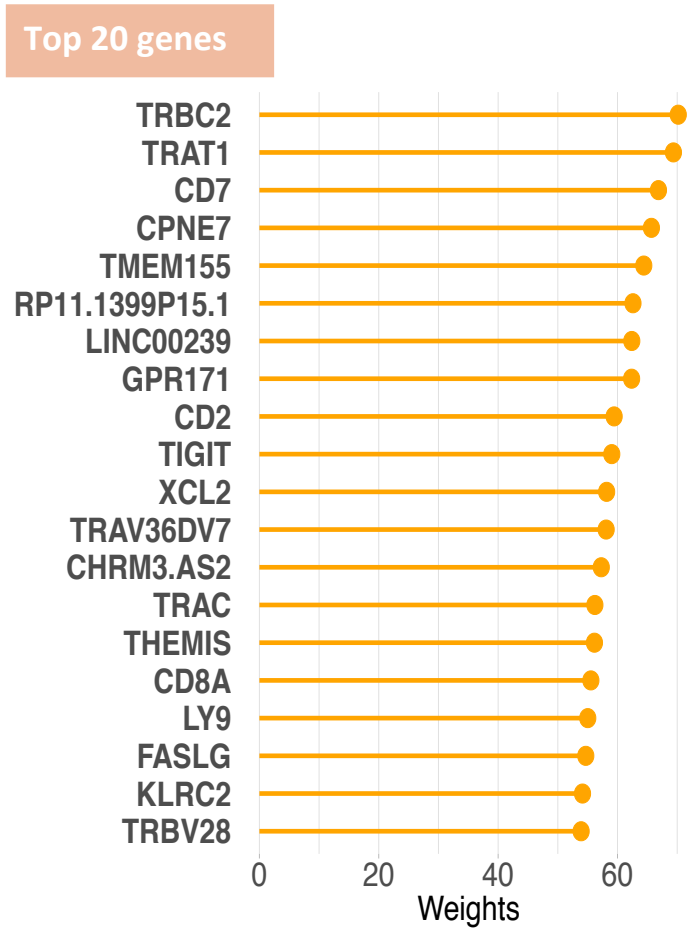


Cell Type

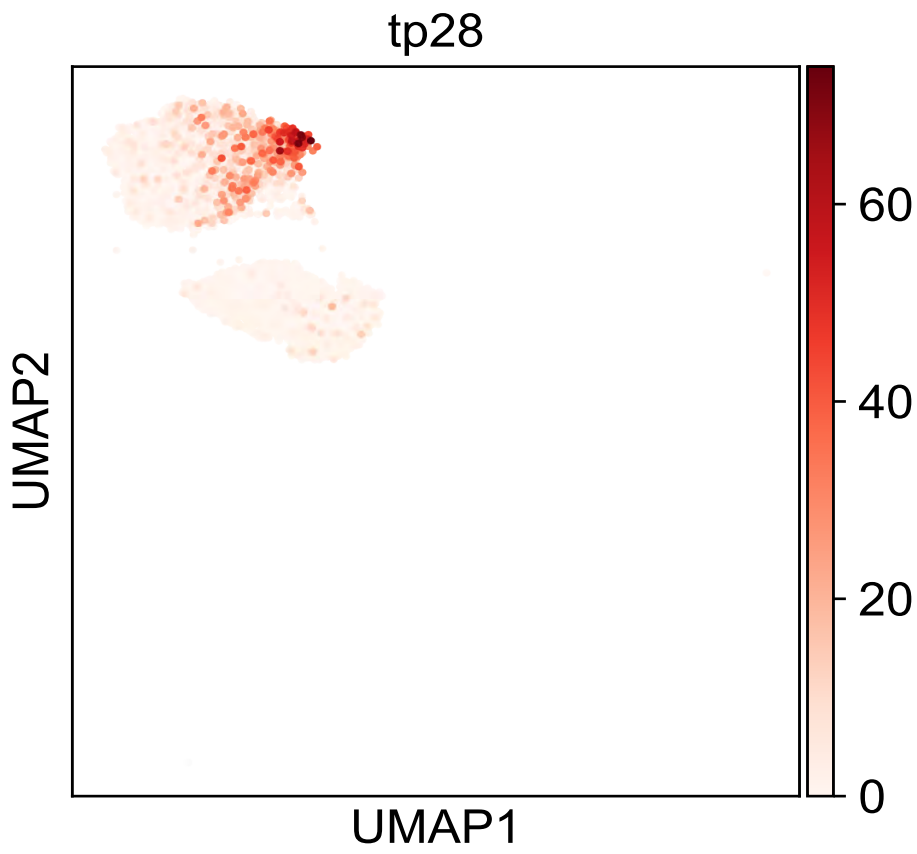
Endo_1

Endo_3

Endo_5



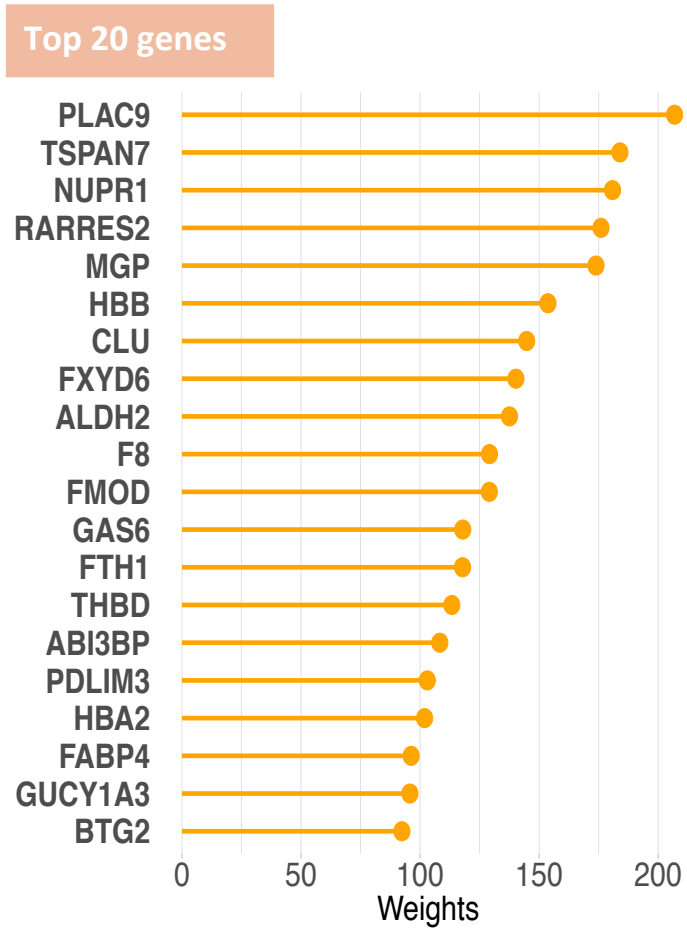
Endo GEM 28



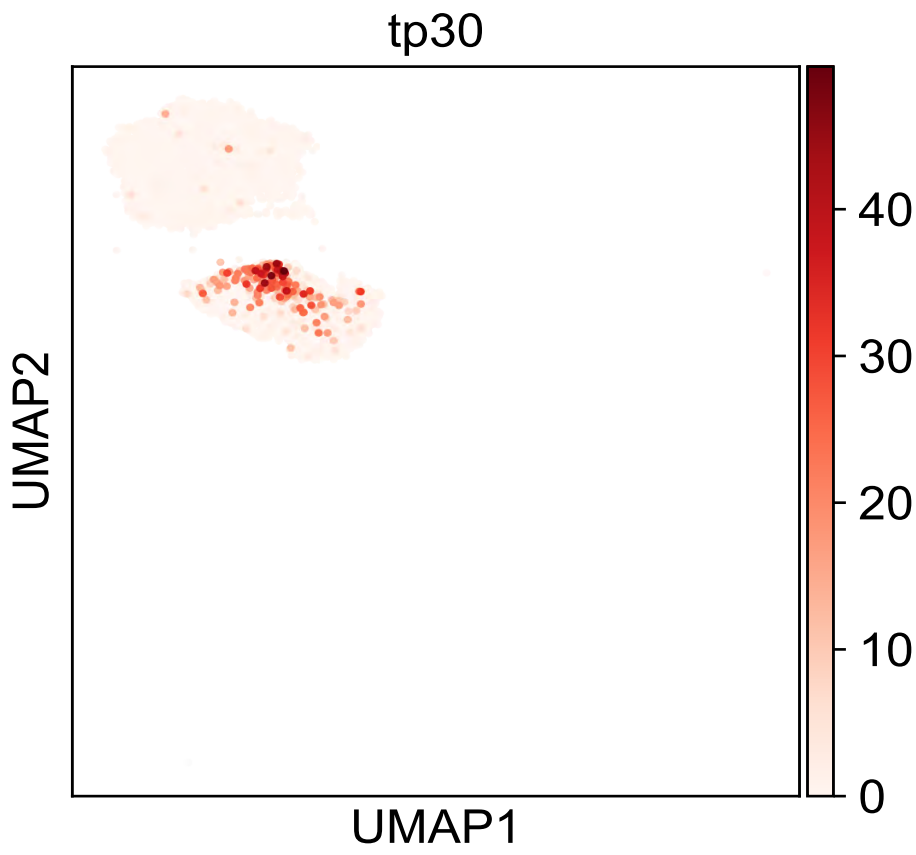
Cell Type

Endo_4

Endo_5

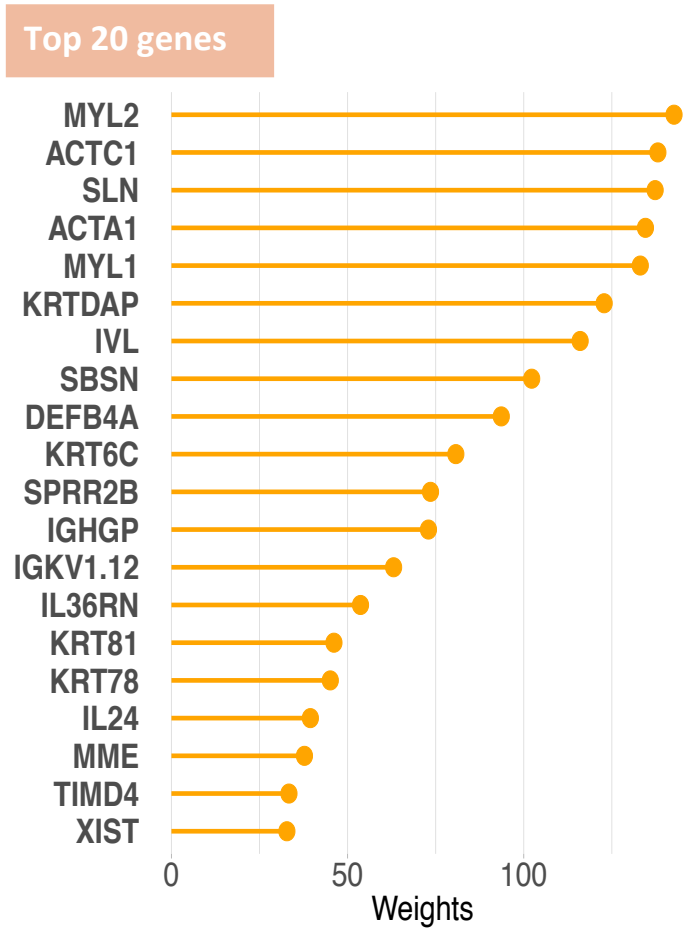


Endo GEM 30

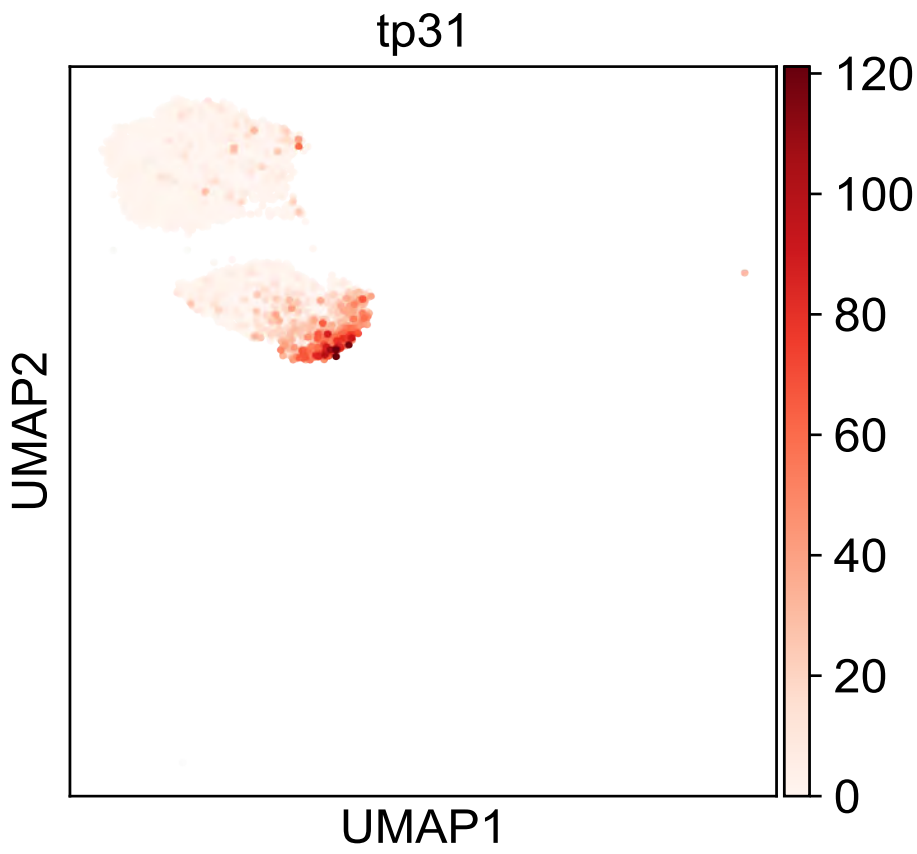


Cell Type

Endo_6

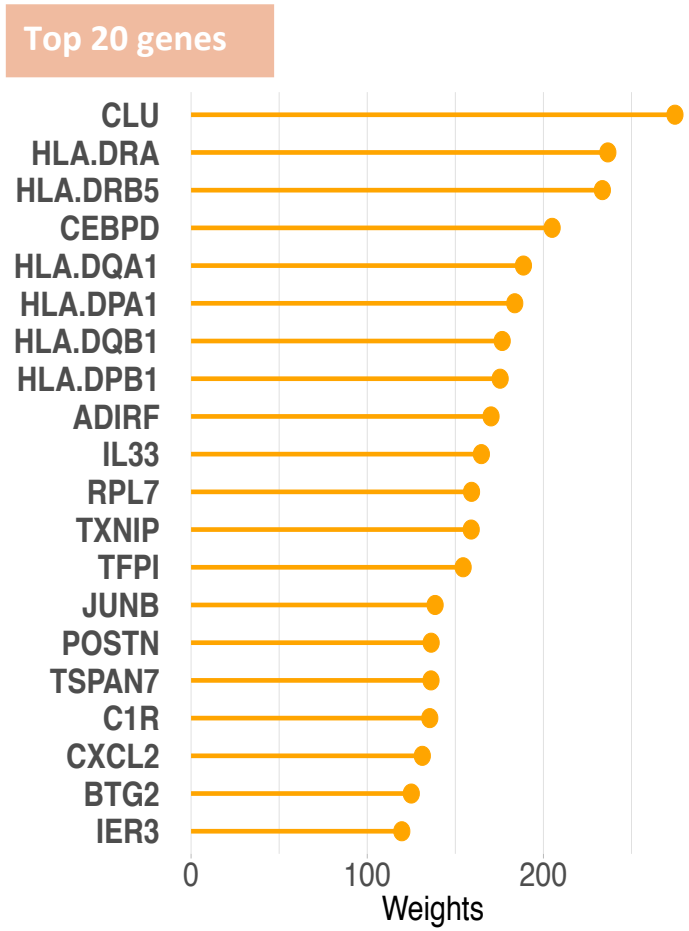


Endo GEM 31

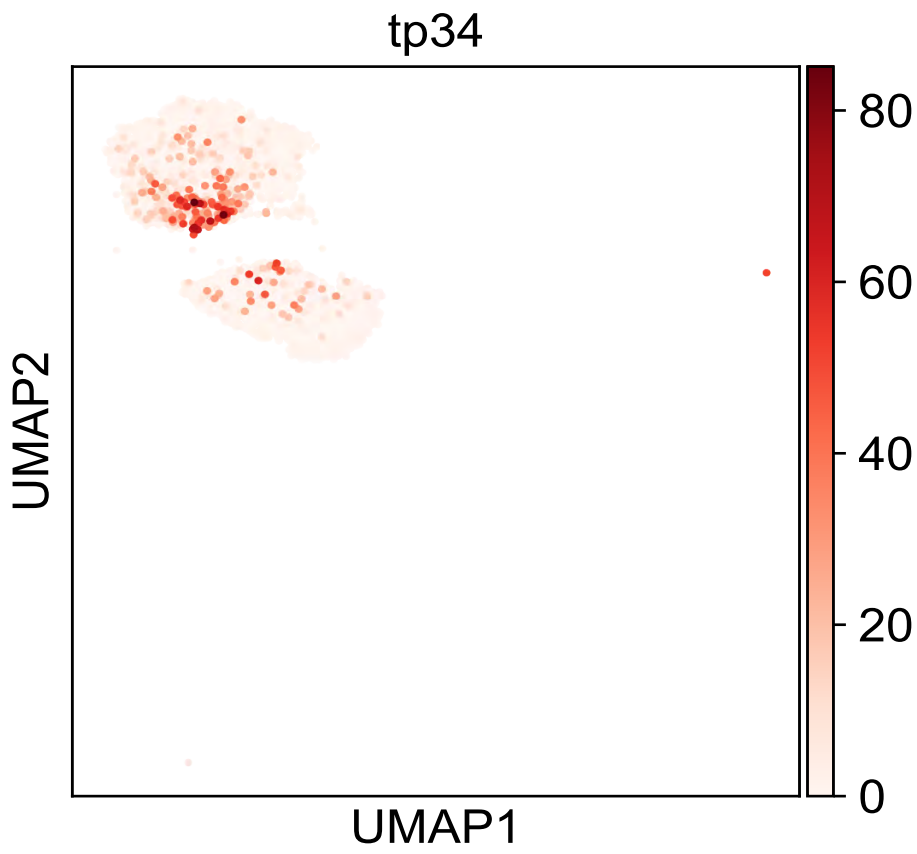


Cell Type

Endo_3



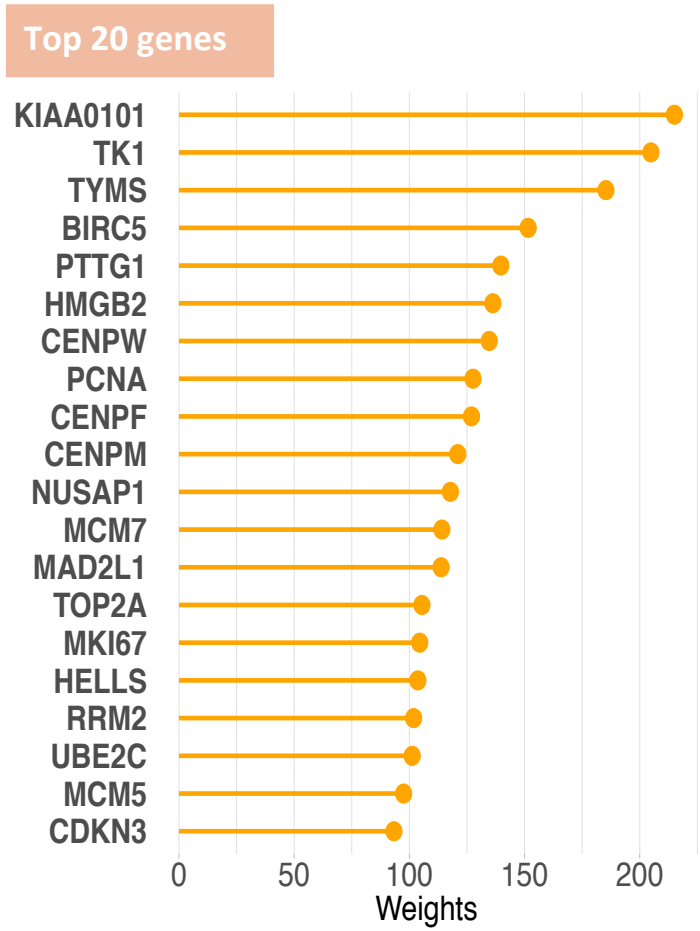
Endo GEM 34



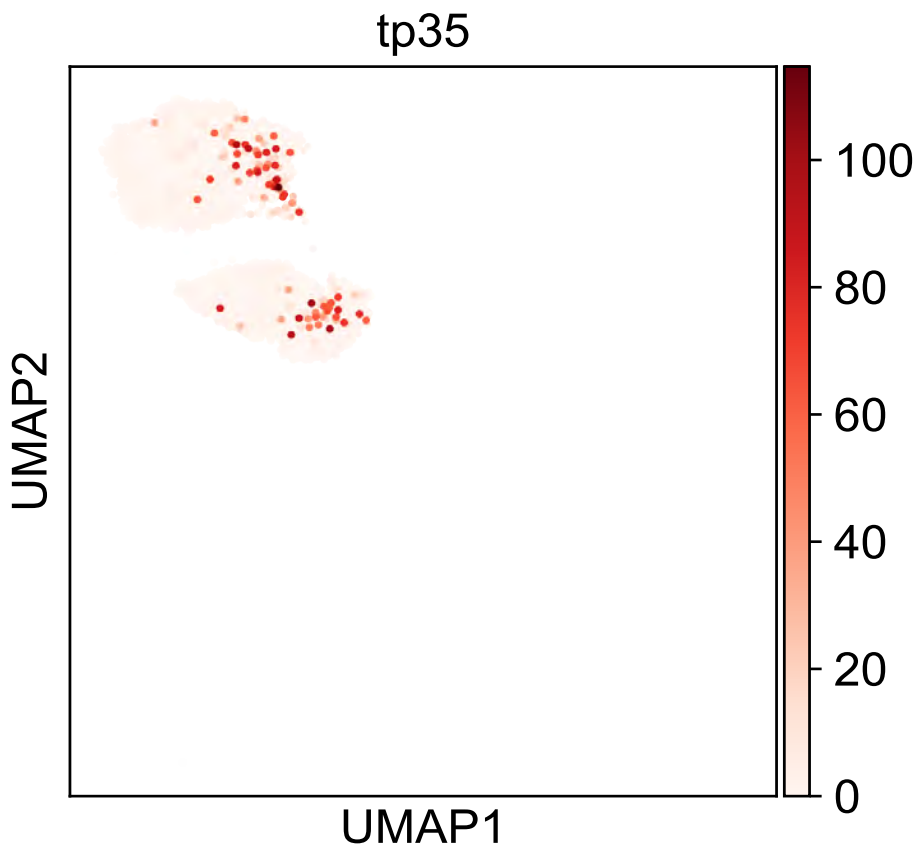
Cell Type

Endo_5

Endo_6



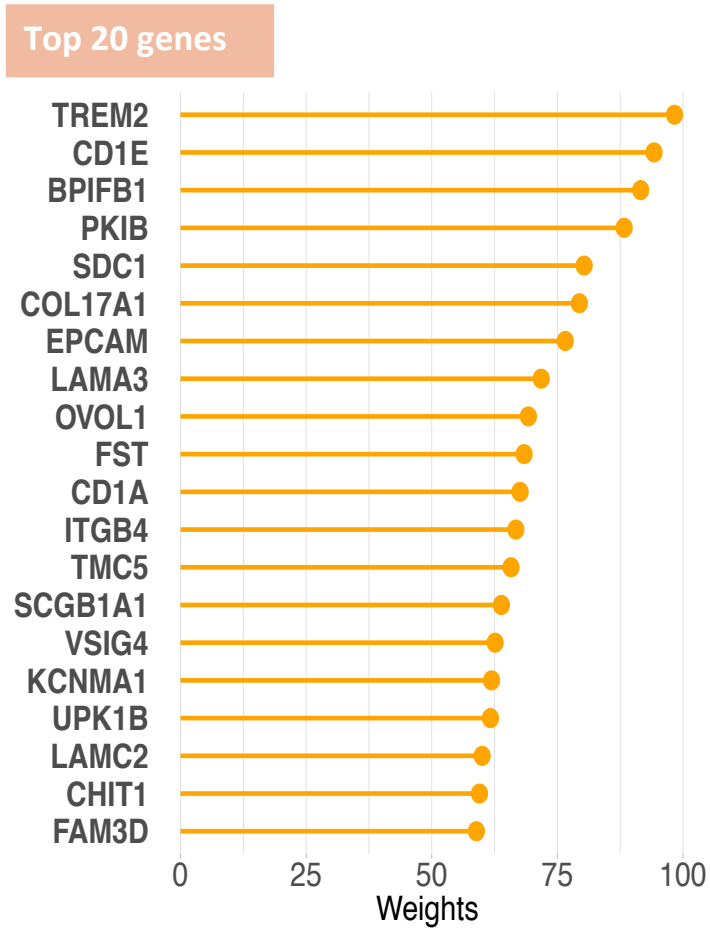
Endo GEM 35



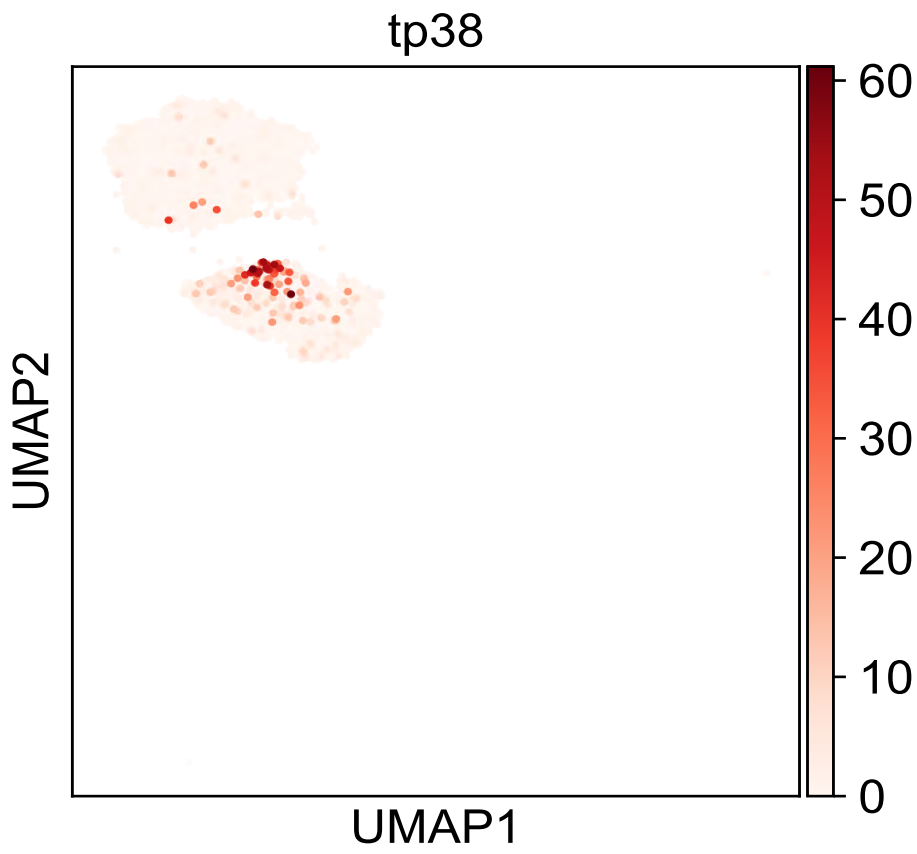
Cell Type

Endo_3

Endo_4



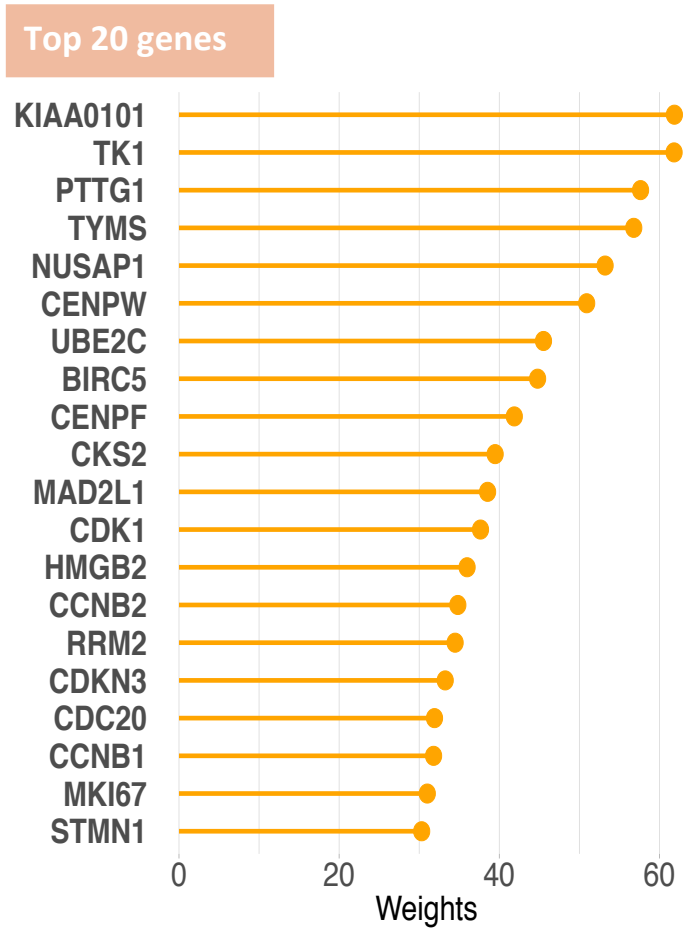
Endo GEM 38



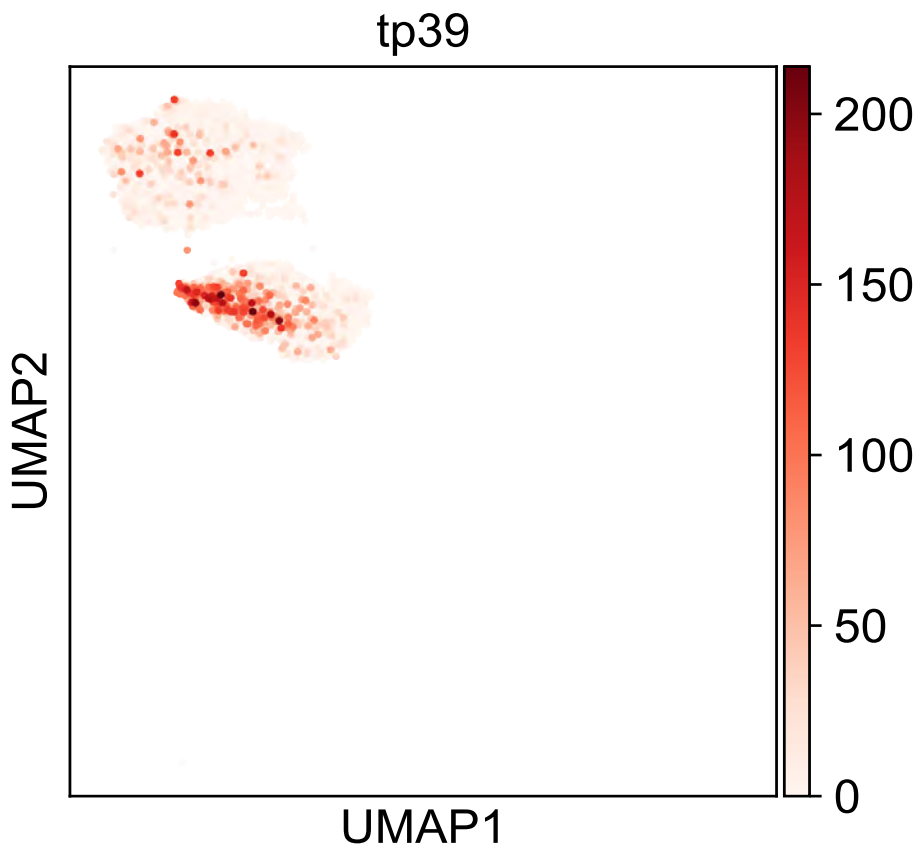
Cell Type

Endo_3

Endo_6



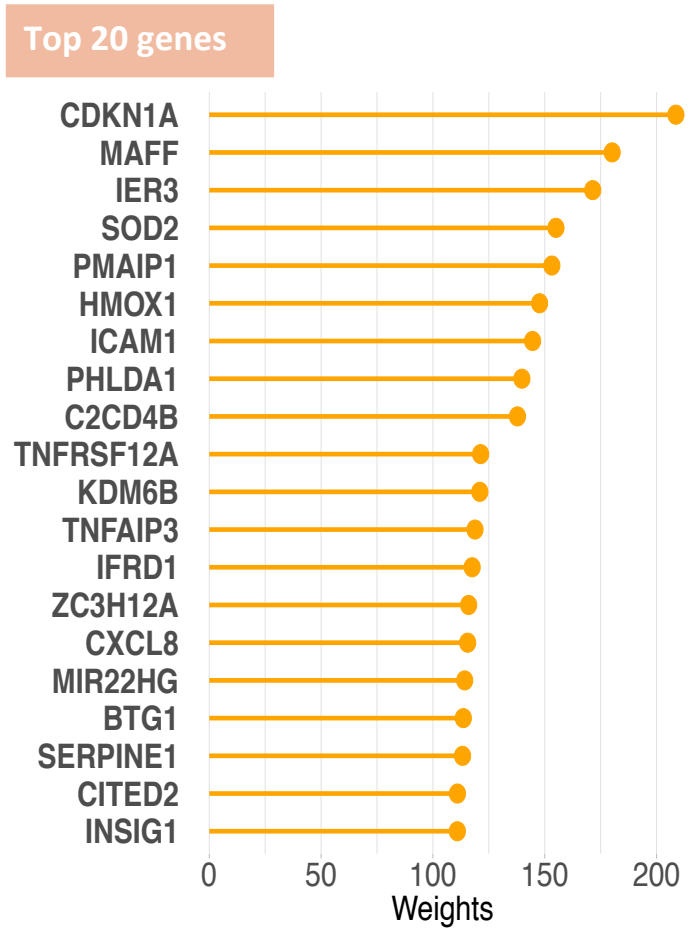
Endo GEM 39



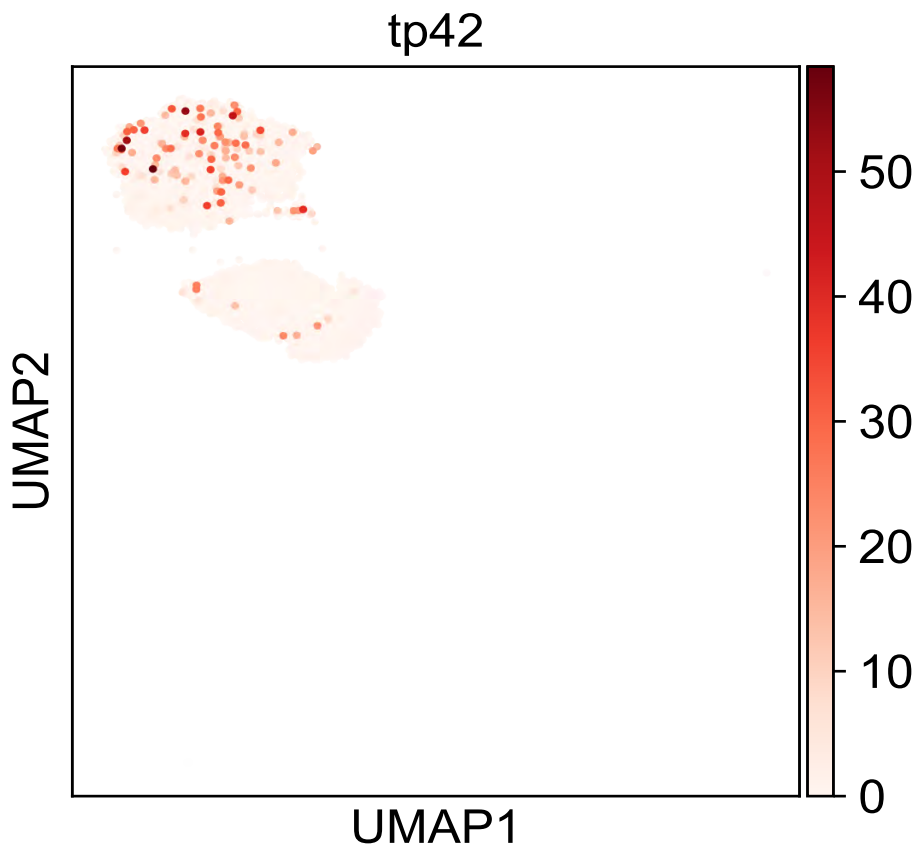
Cell Type

Endo_2

Endo_3



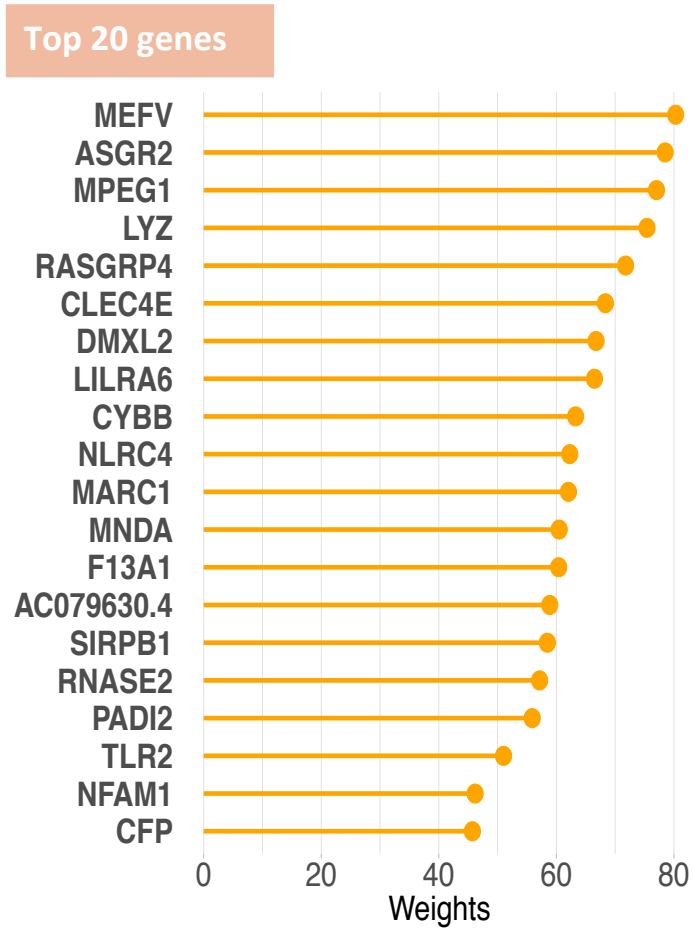
Endo GEM 42



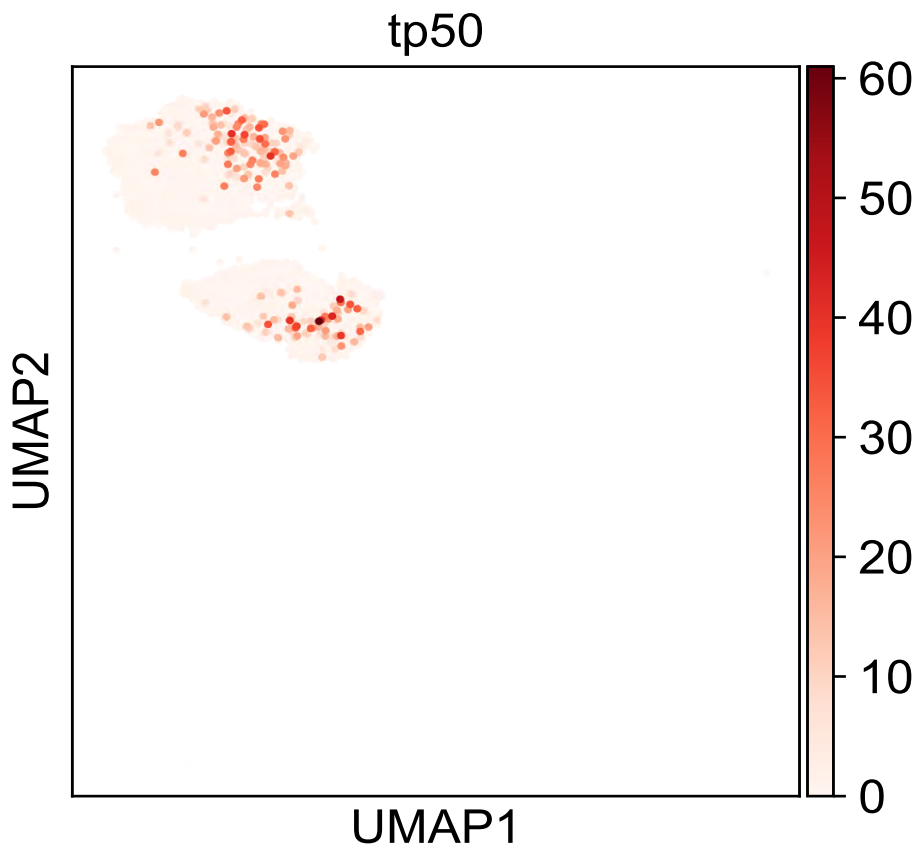
Cell Type

Endo_1

Endo_4



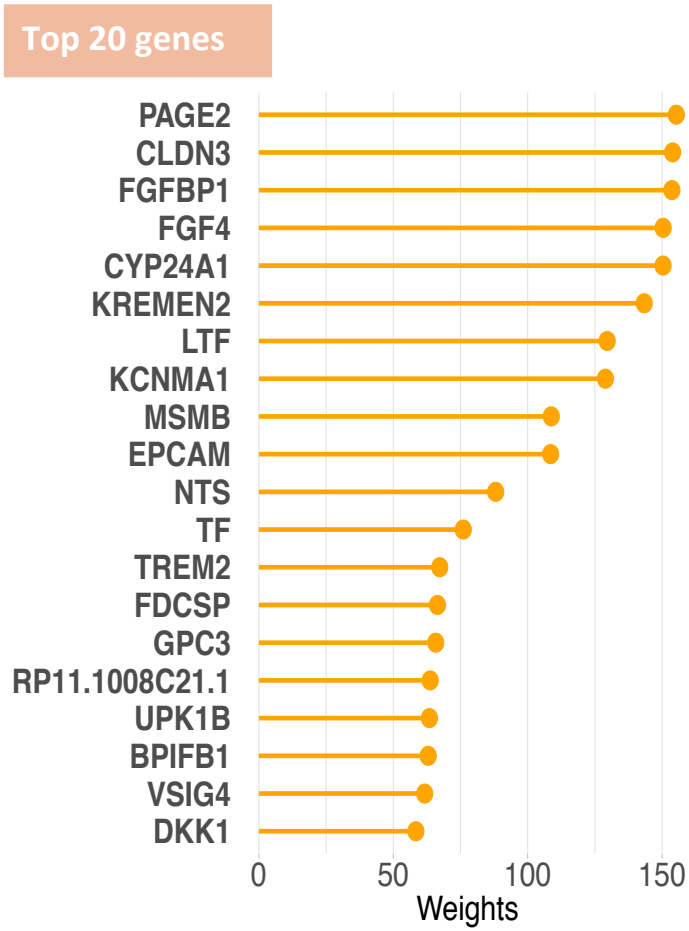
Endo GEM 50



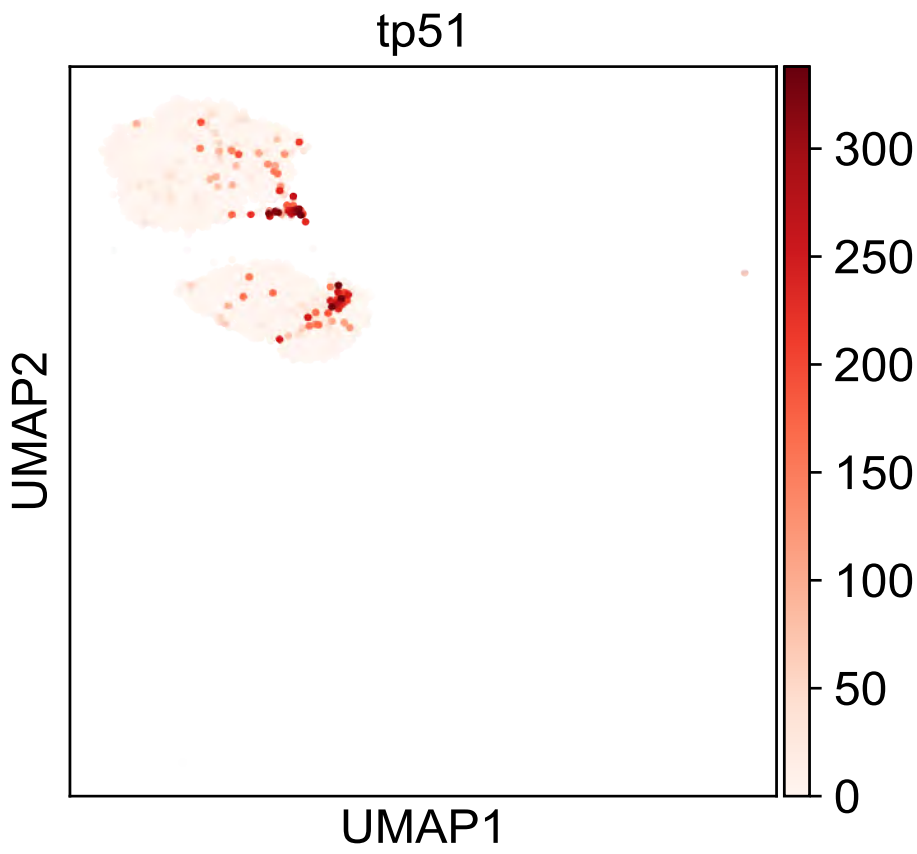
Cell Type

Endo_3

Endo_4



Endo GEM 51

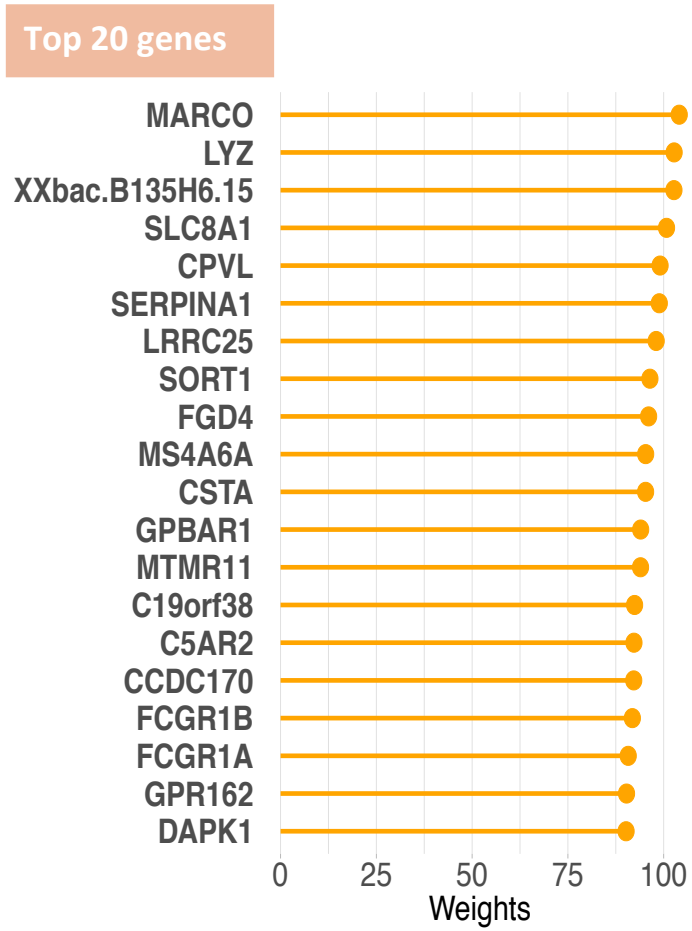


Cell Type

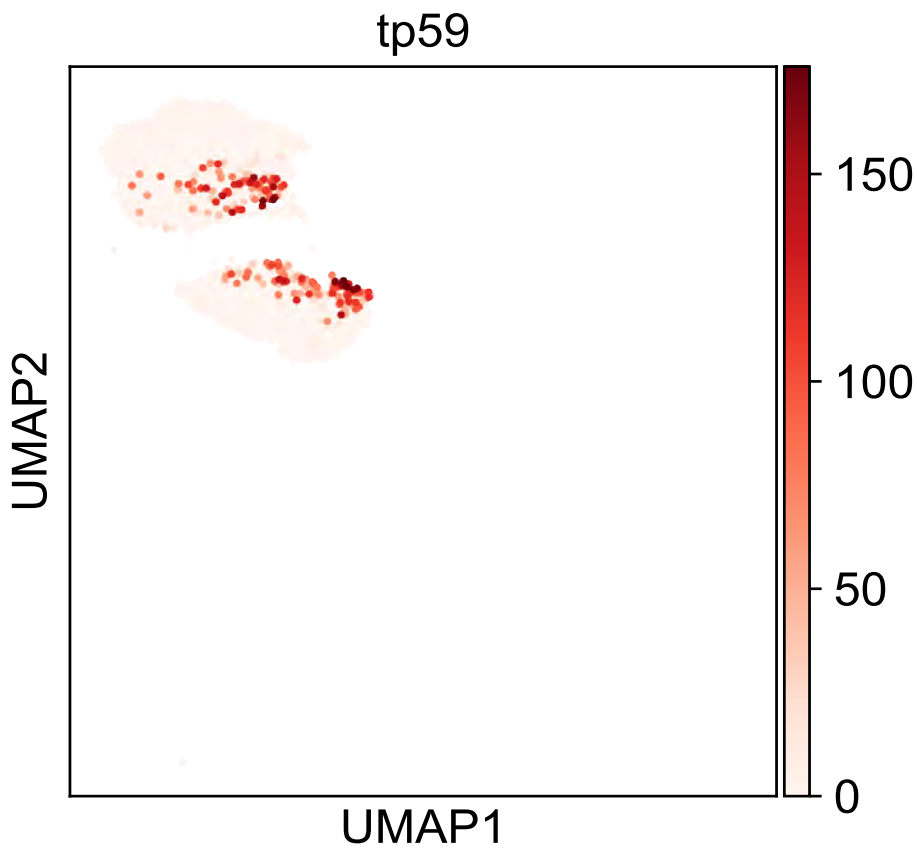
Endo_3

Endo_4

Endo_6



Endo GEM 59

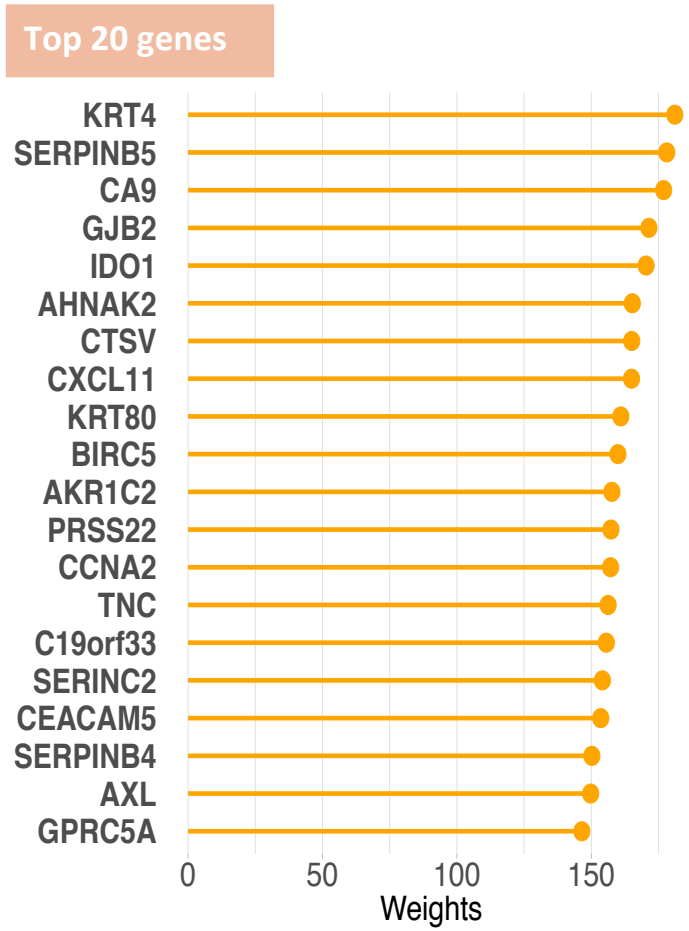


Cell Type

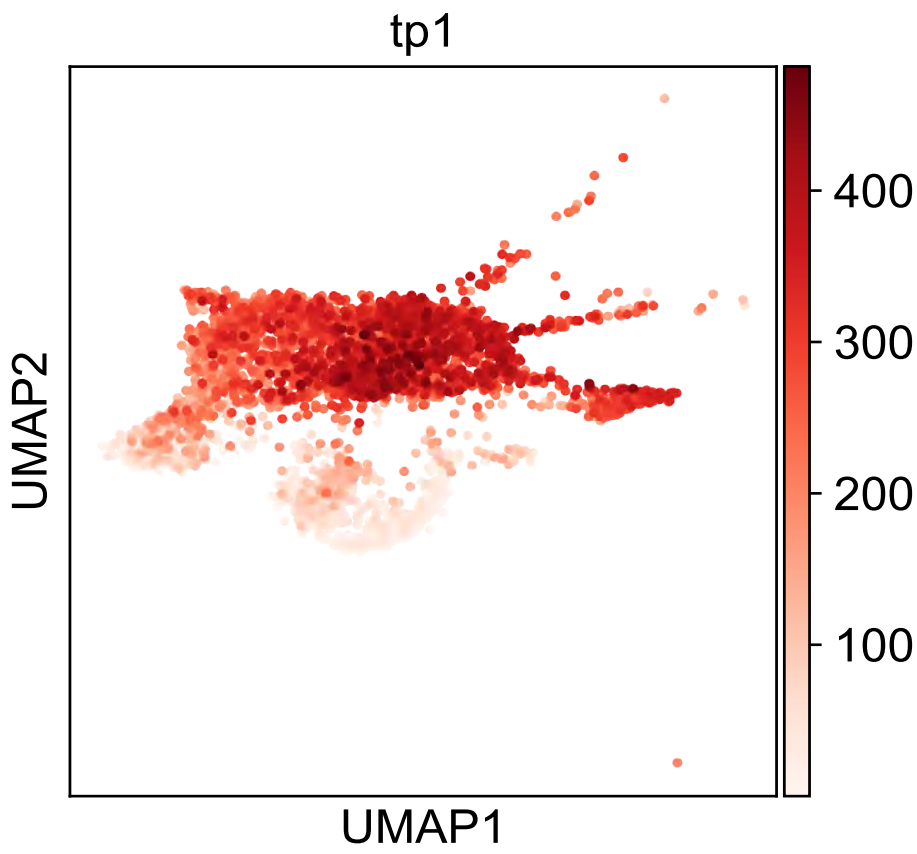
Endo_3

Endo_5

Endo_6



Fib GEM 1

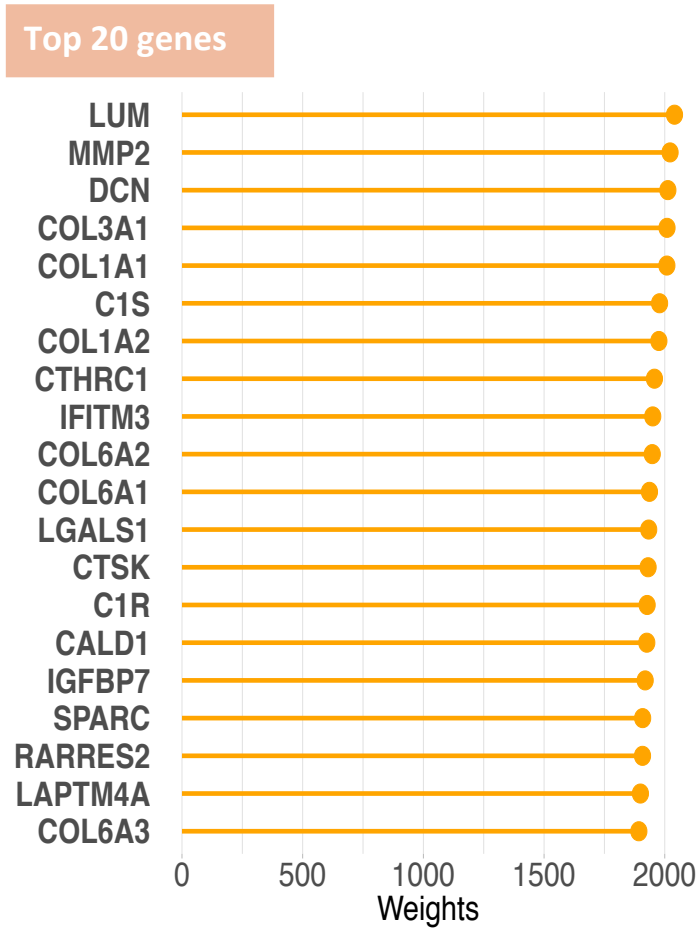


Cell Type

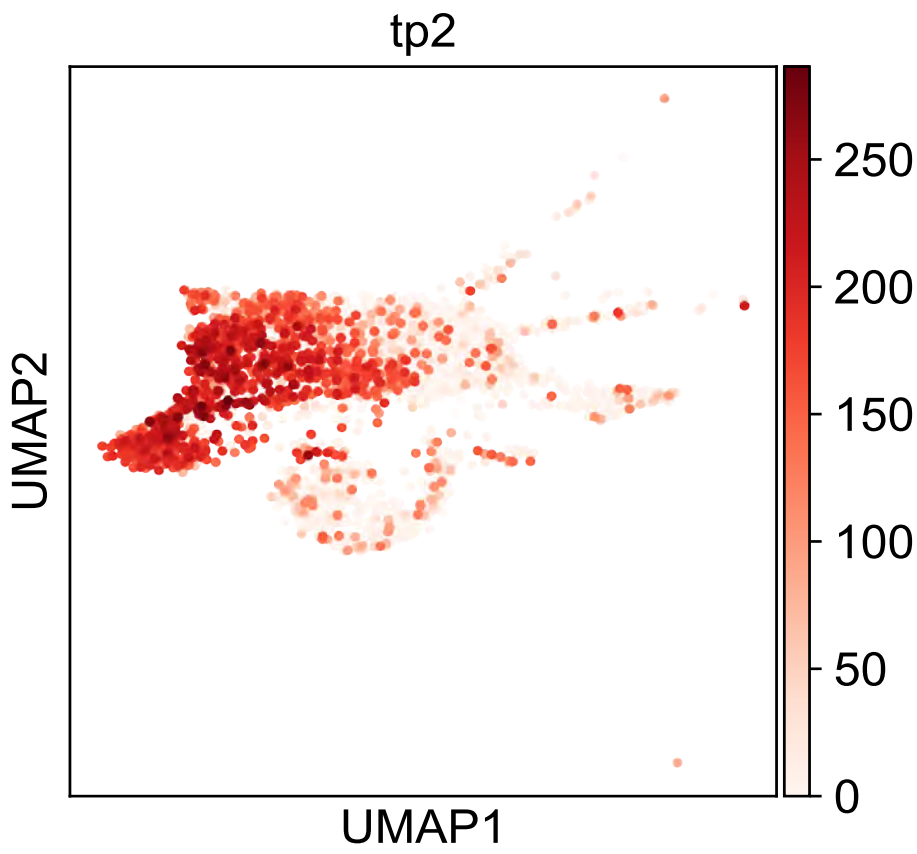
Fib_1

Fib_2

Fib_3



Fib GEM 2

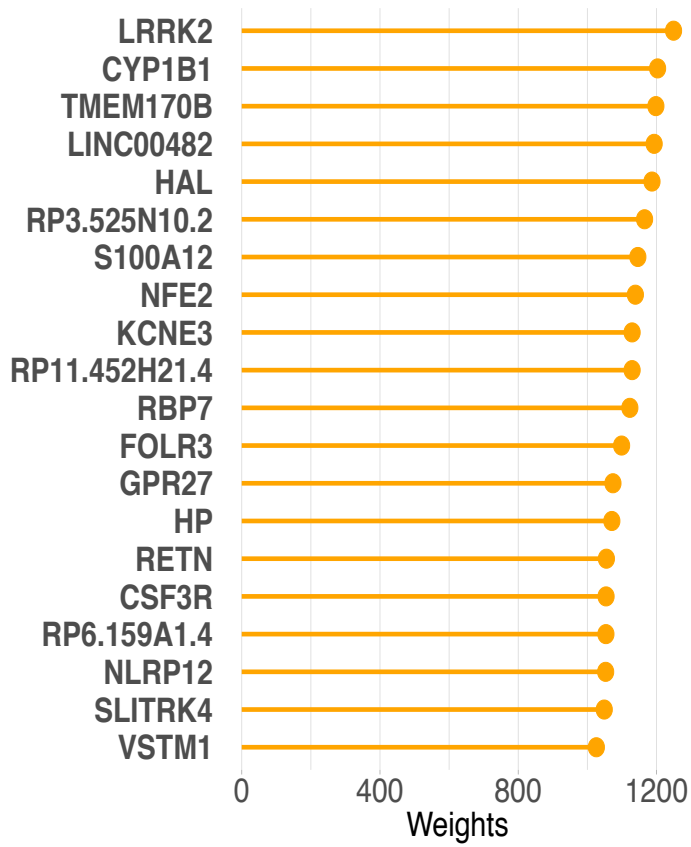


Cell Type

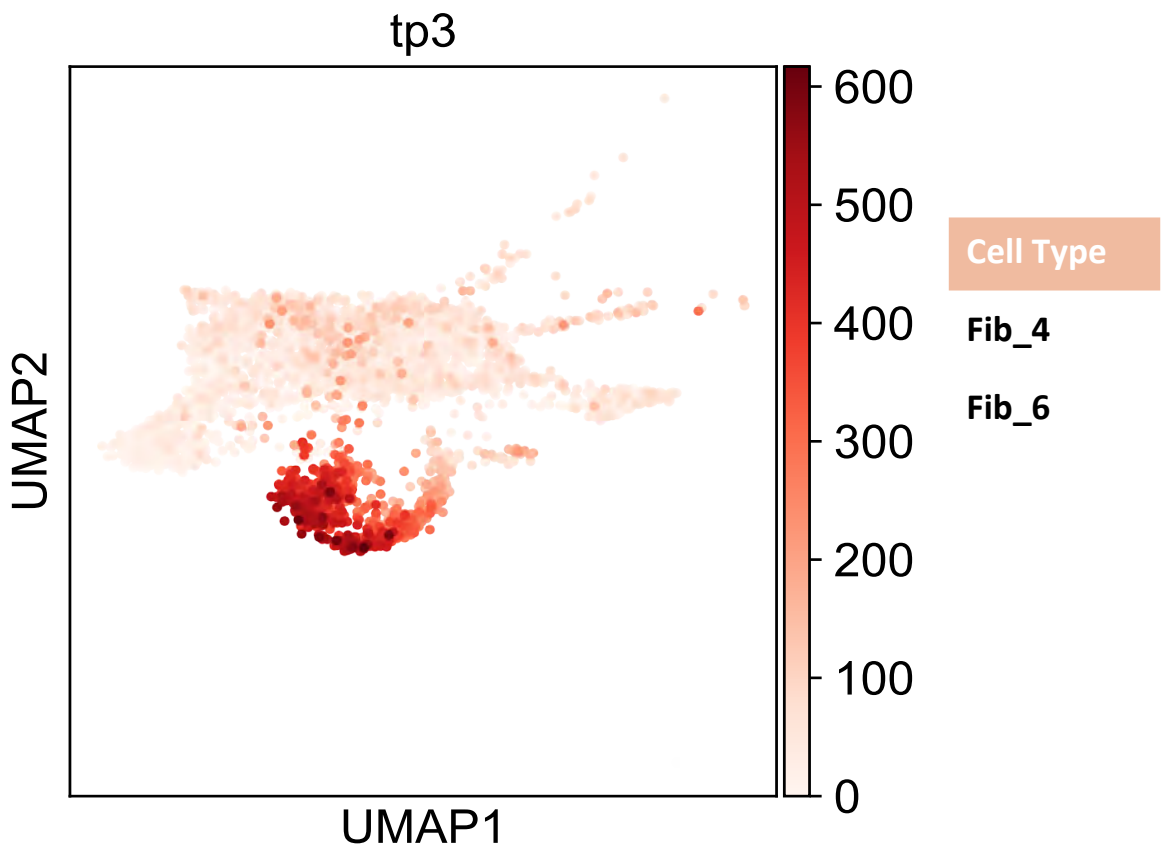
Fib_5

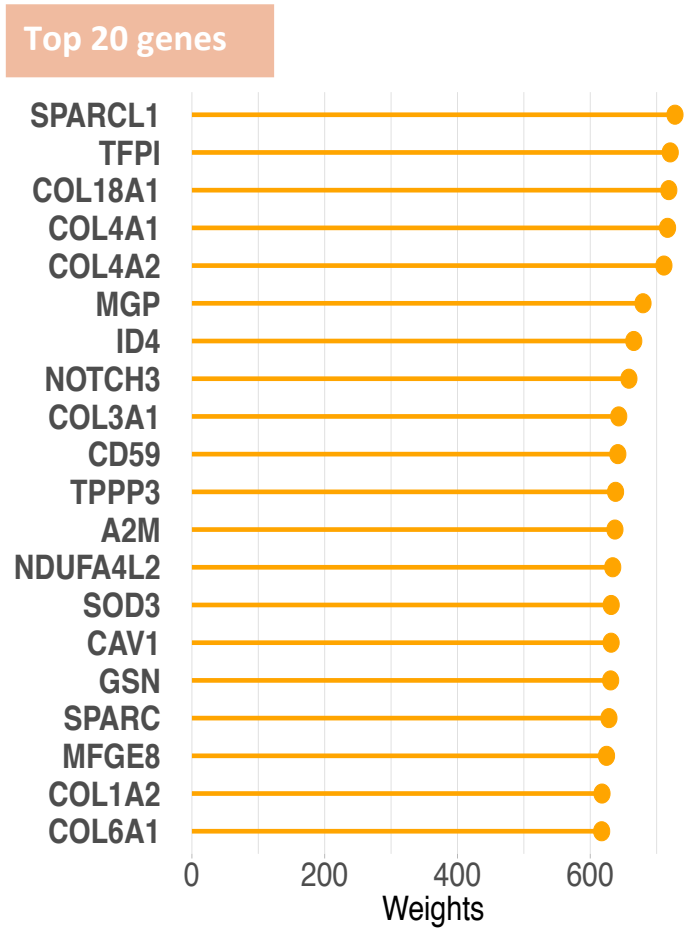
Fib_3

Top 20 genes

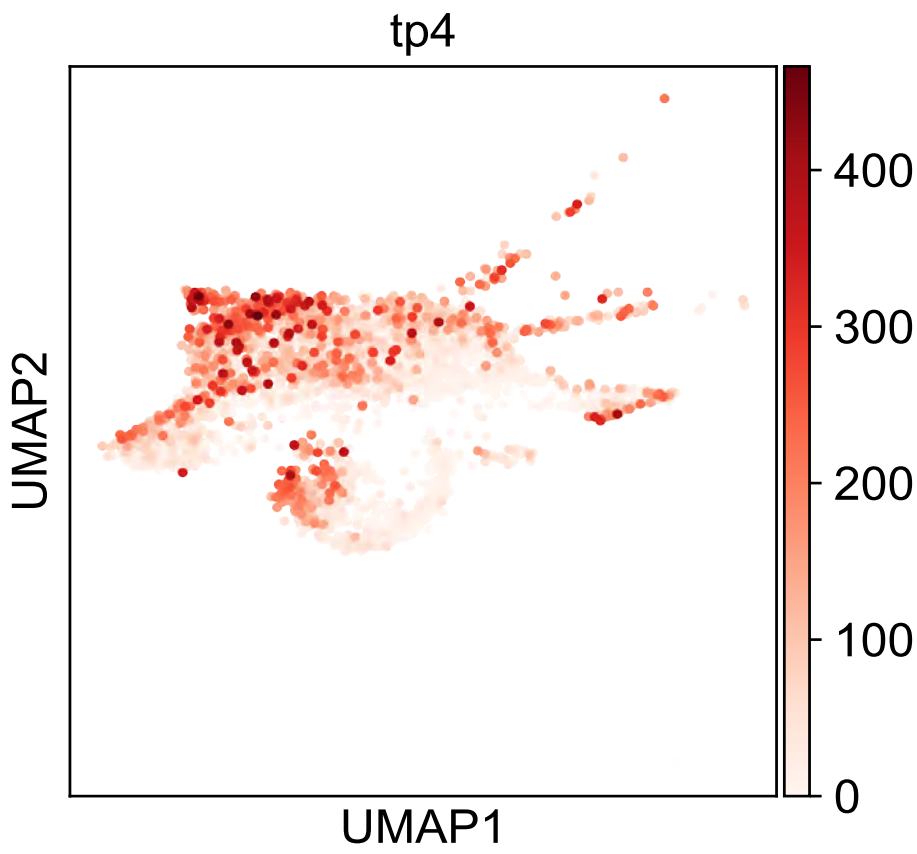


Fib GEM 3


tp3
UMAP2
UMAP1
600
500
400
300
200
100
0
Cell Type
Fib_4
Fib_6

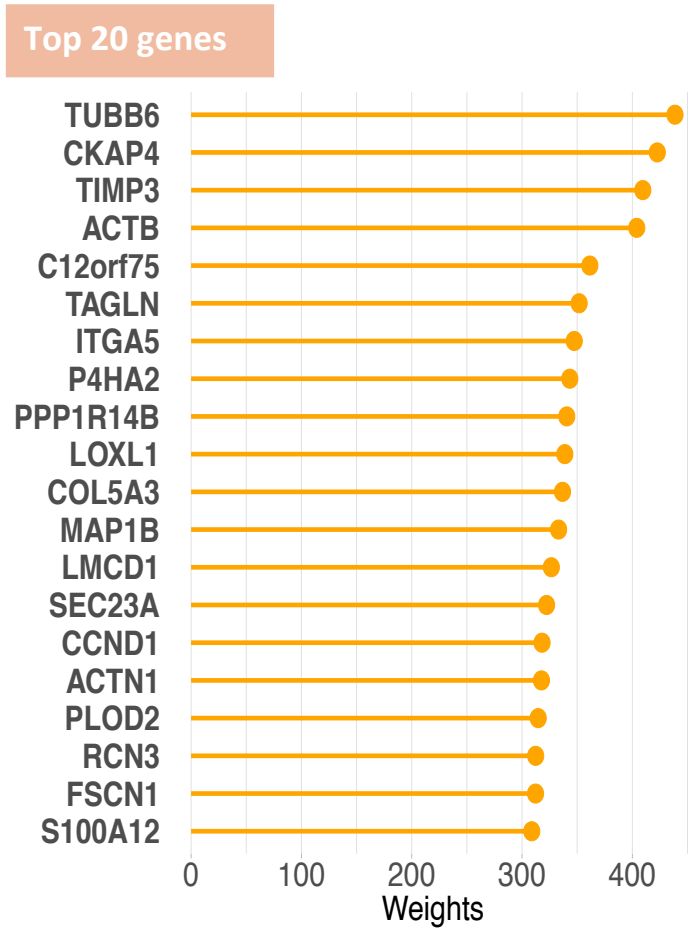

Top 20 genes
SPARCL1
TFPI
COL18A1
COL4A1
COL4A2
MGP
ID4
NOTCH3
COL3A1
CD59
TPPP3
A2M
NDUFA4L2
SOD3
CAV1
GSN
SPARC
MFGE8
COL1A2
COL6A1
0
200
400
600
Weights

Fib GEM 4

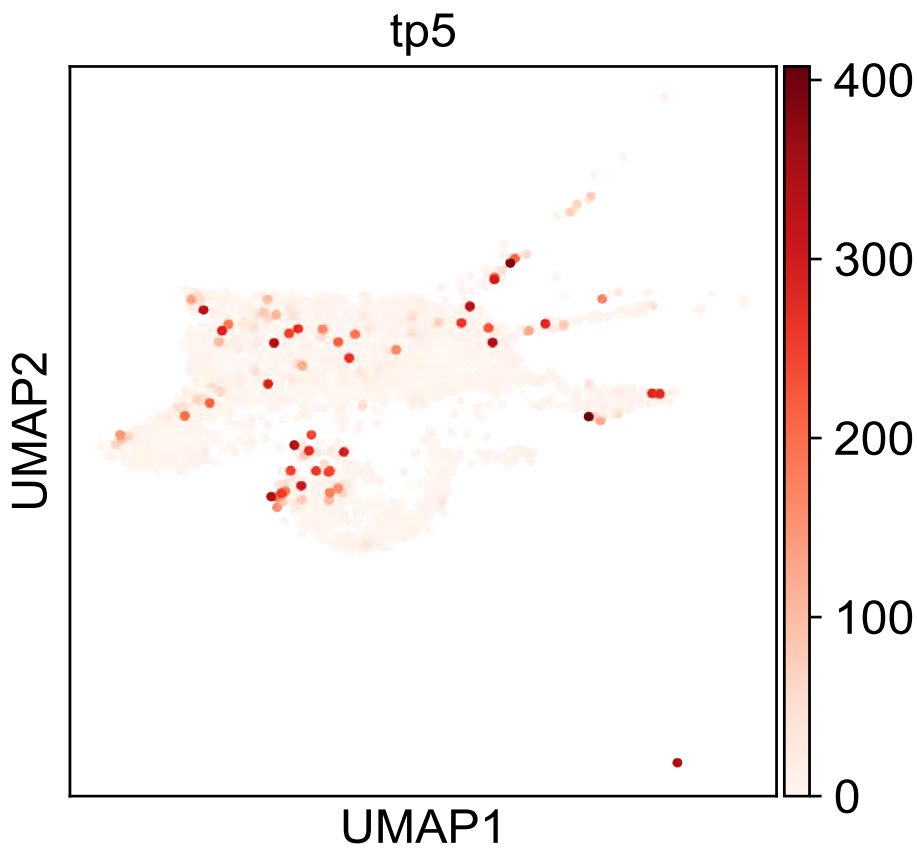


Cell Type

- Fib_5
- Fib_1
- Fib_6
- Fib_2
- Fib_3



Fib GEM 5

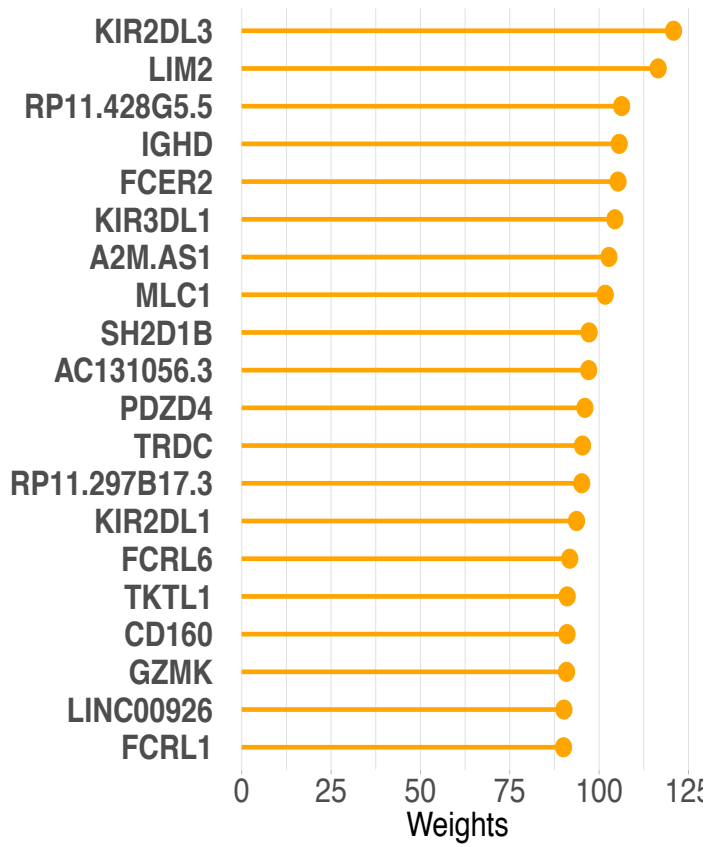


Cell Type

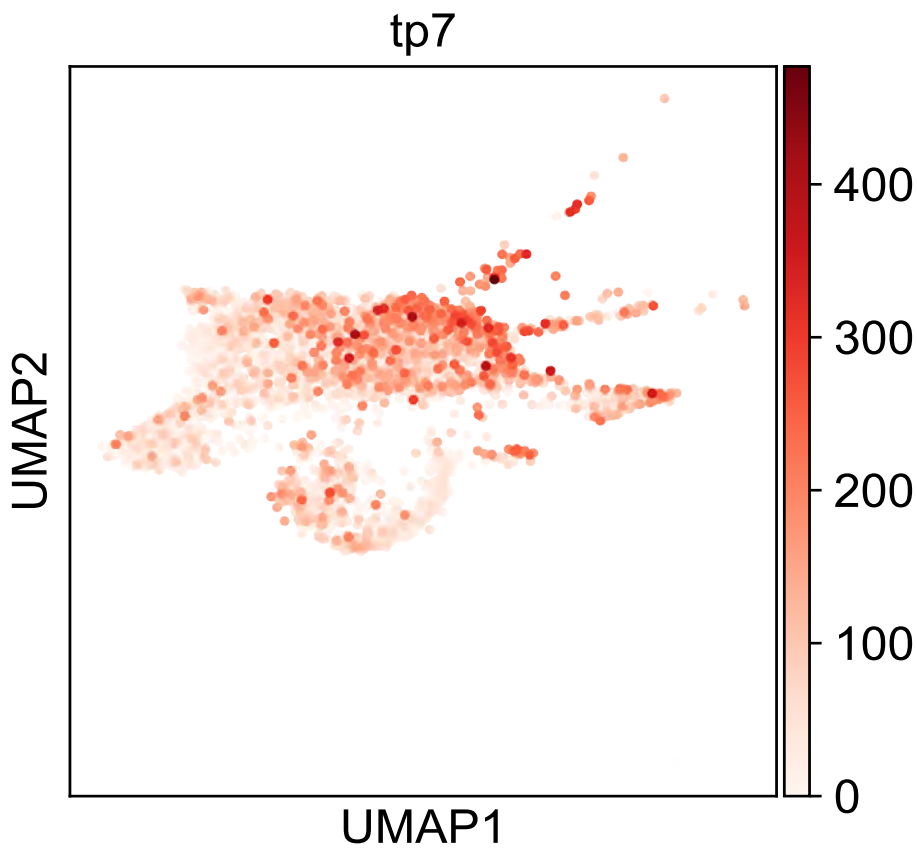
Fib_2

Fib_6

Top 20 genes



Fib GEM 7

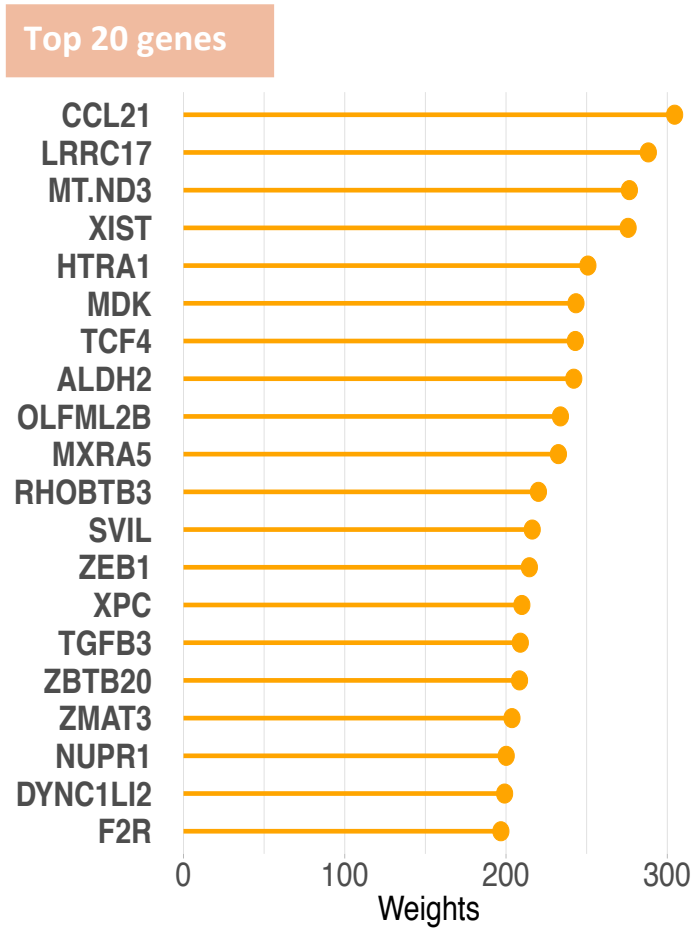


Cell Type

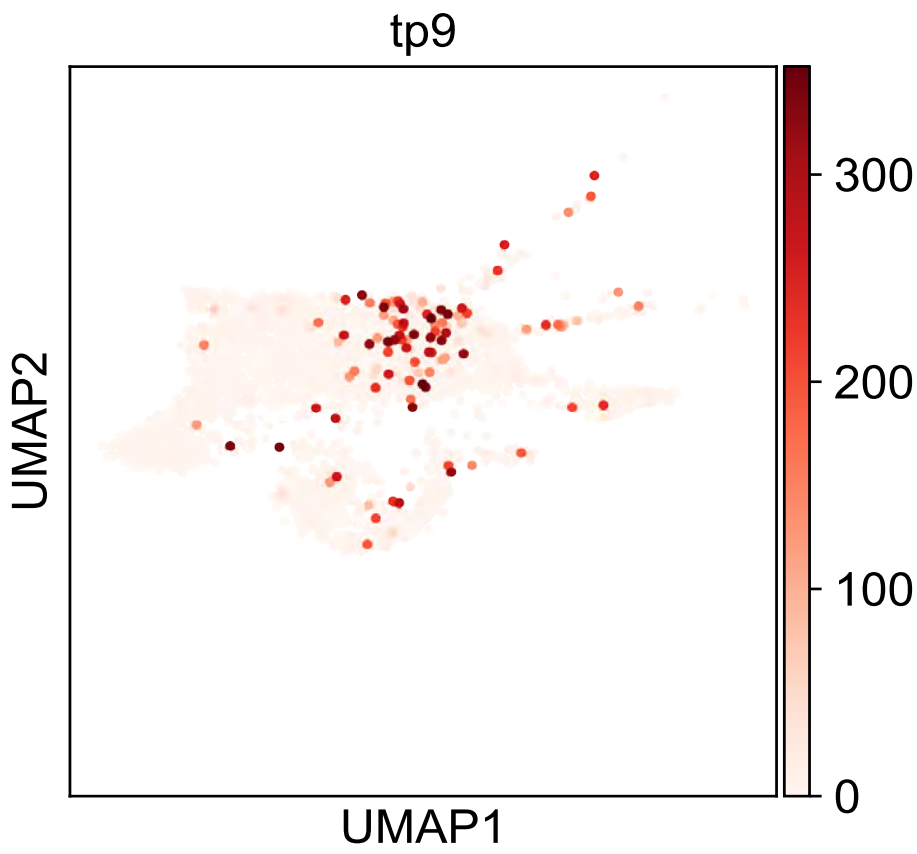
Fib_1

Fib_2

Fib_6



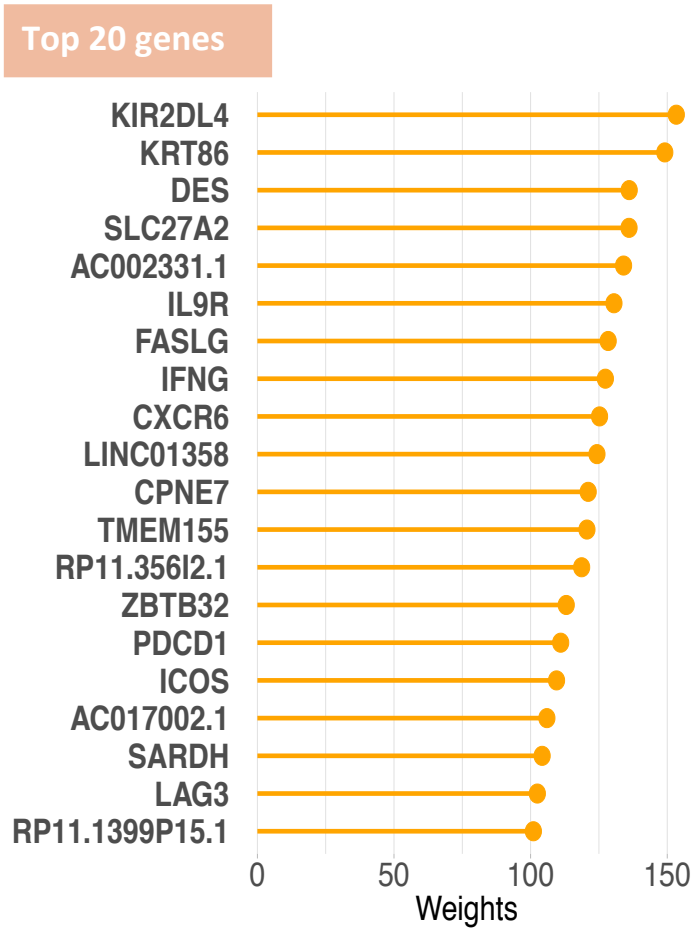
Fib GEM 9



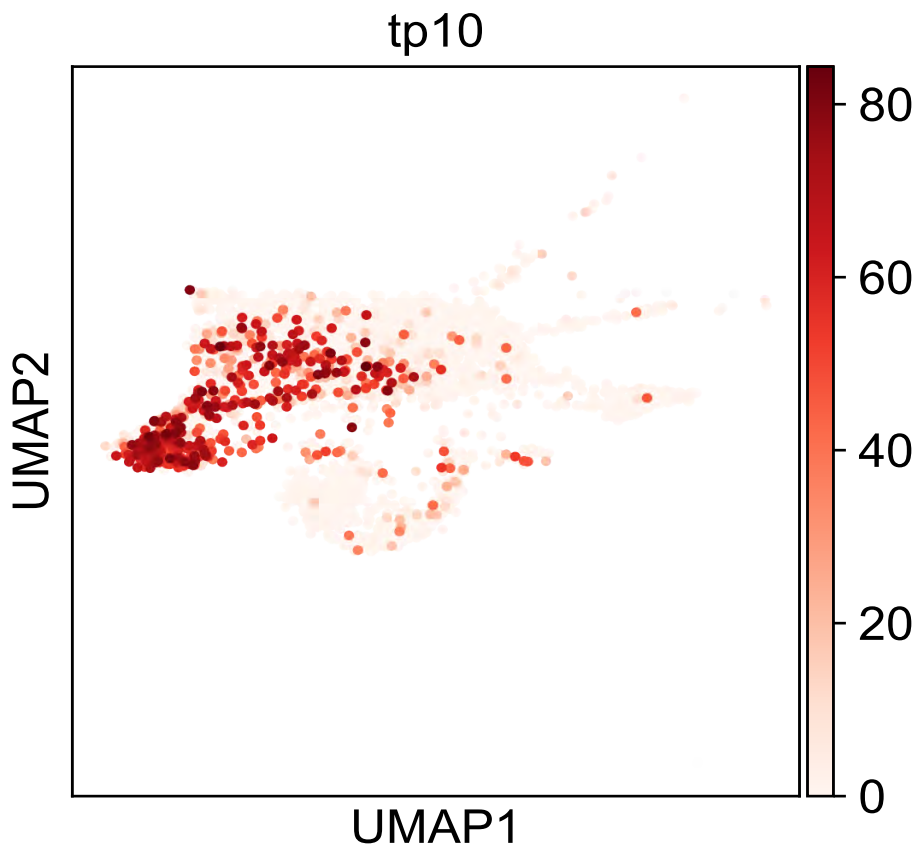
Cell Type

Fib_1

Fib_4



Fib GEM 10

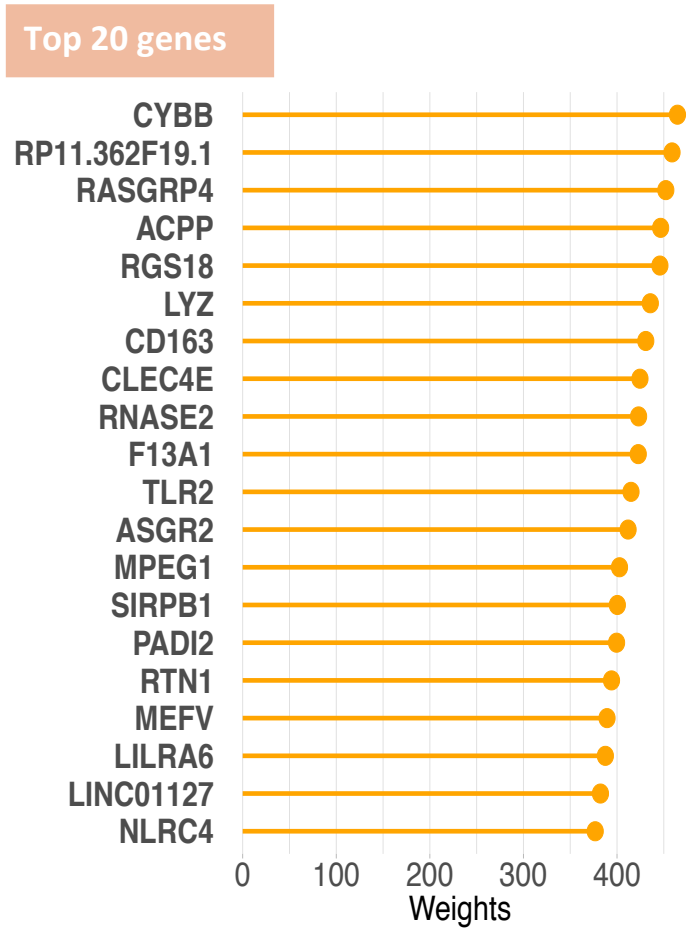


Cell Type

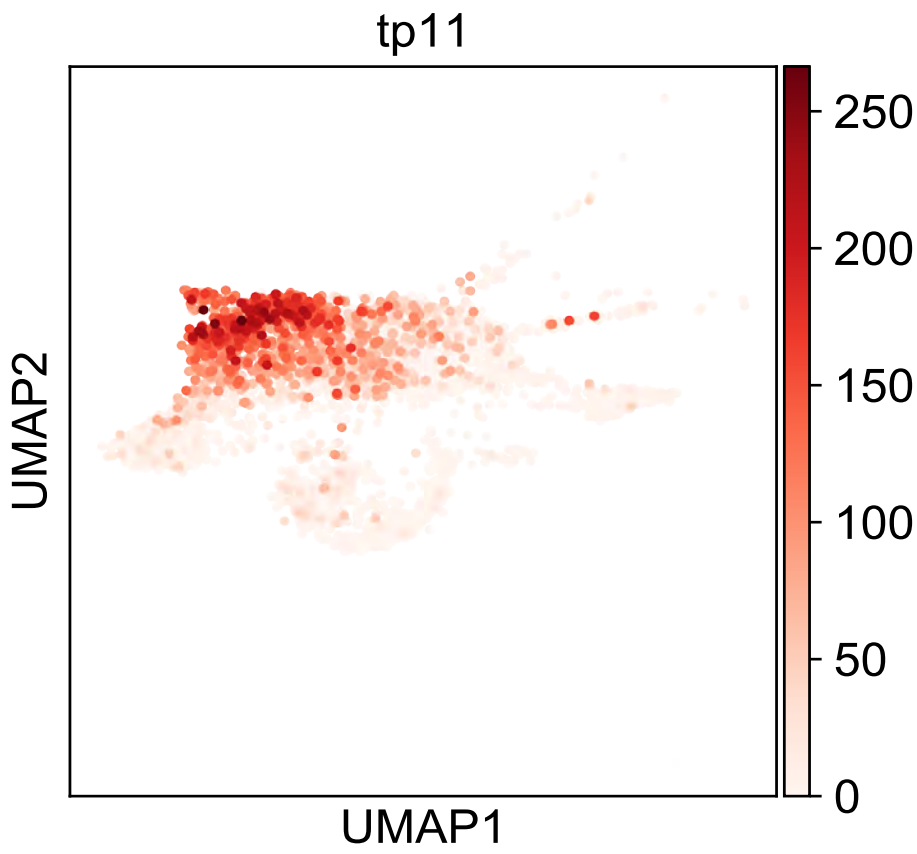
Fib_5

Fib_1

Fib_3



Fib GEM 11

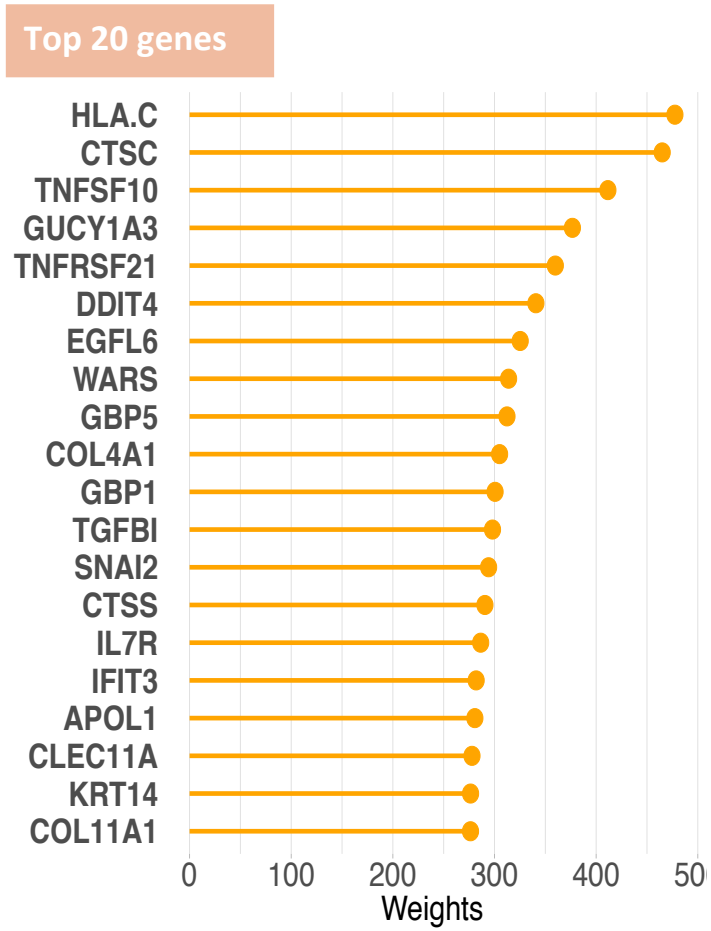


Cell Type

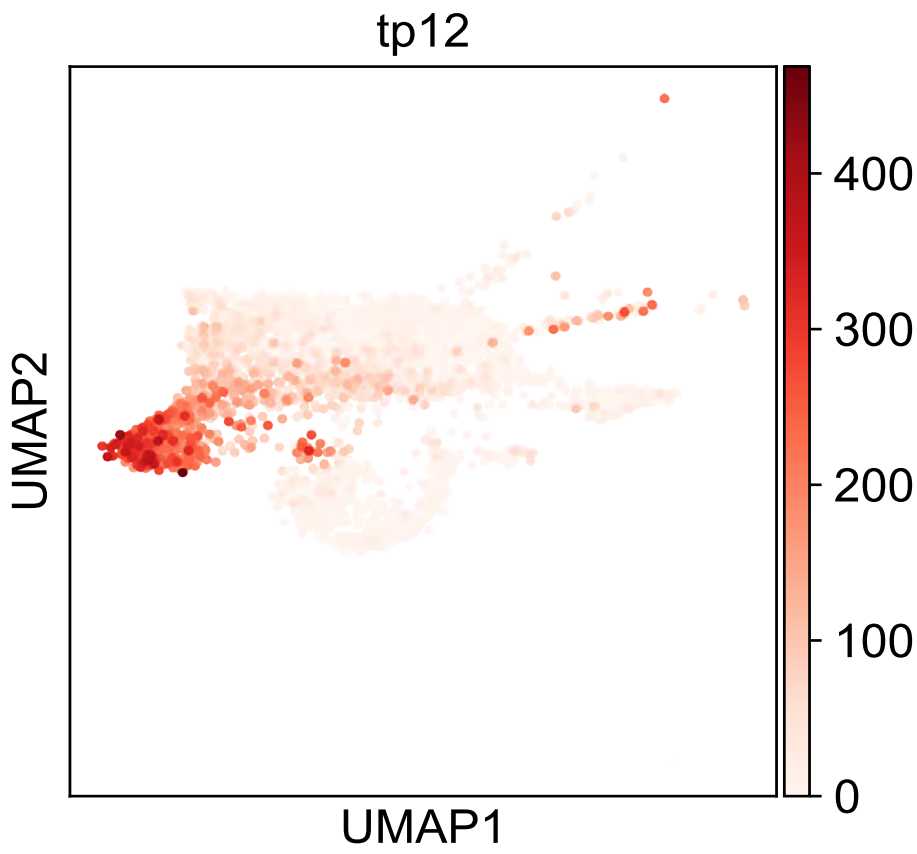
Fib_1

Fib_2

Fib_3



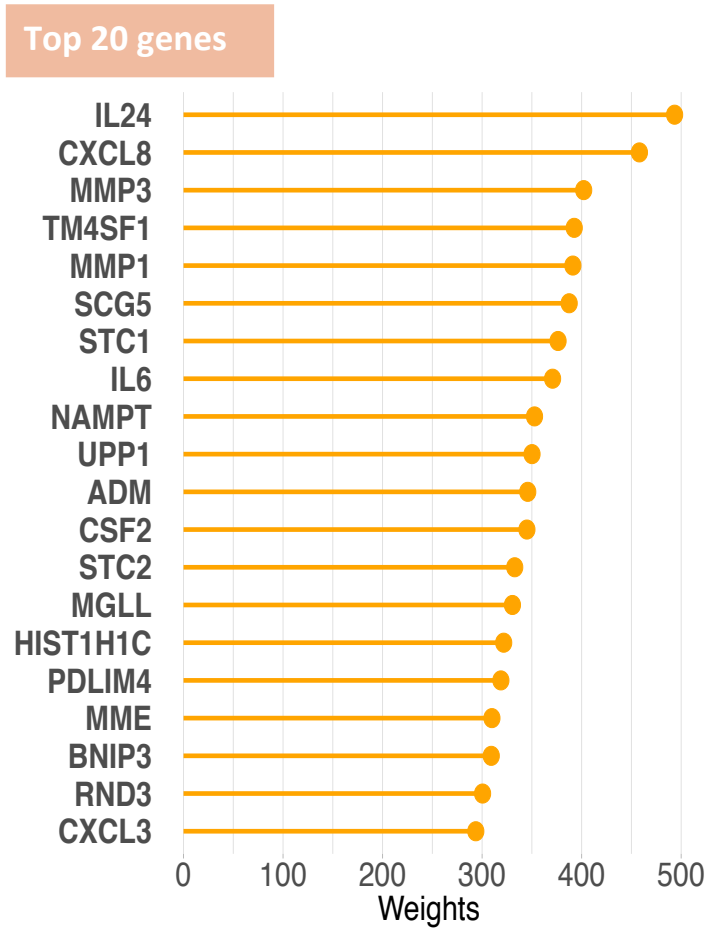
Fib GEM 12



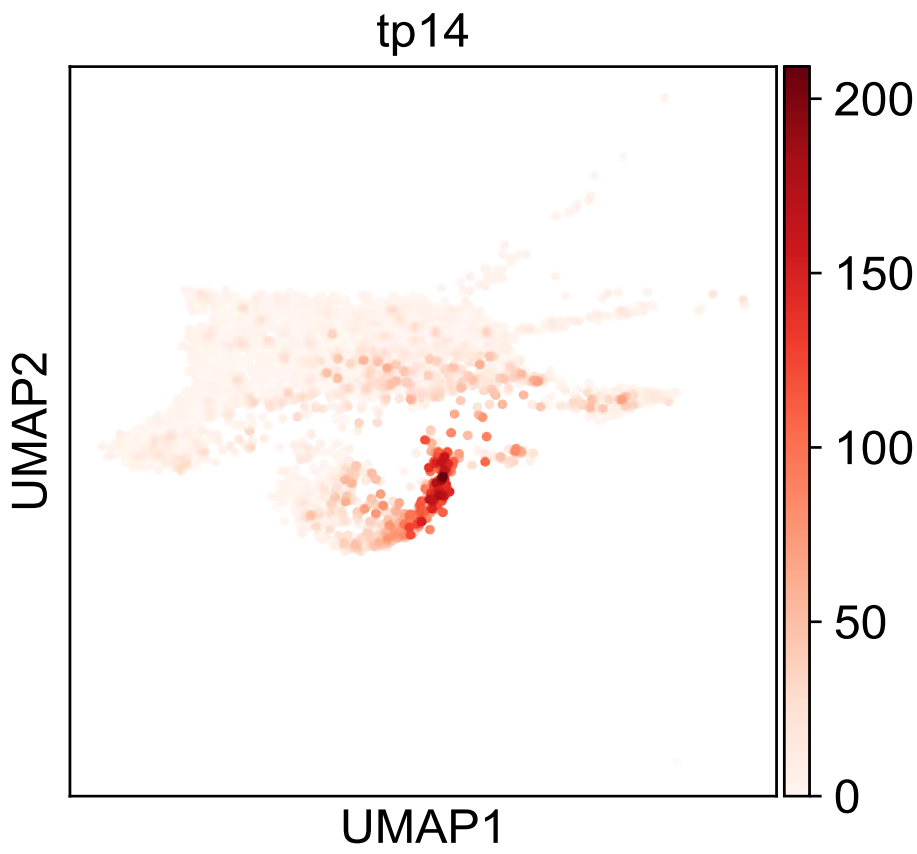
Cell Type

Fib_5

Fib_1

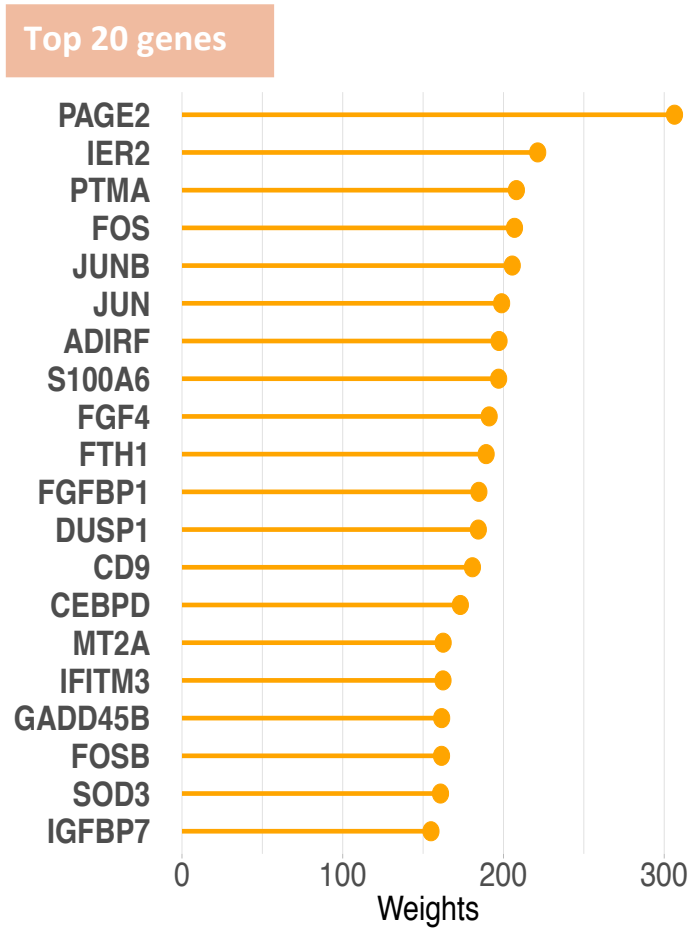


Fib GEM 14

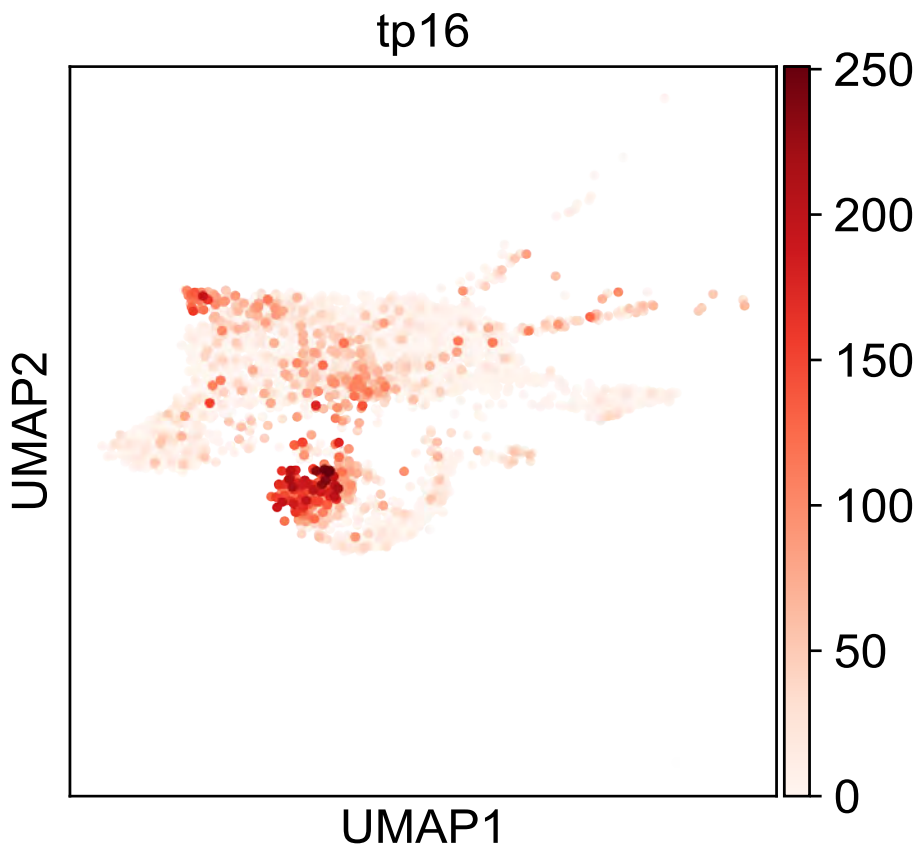


Cell Type

Fib_4



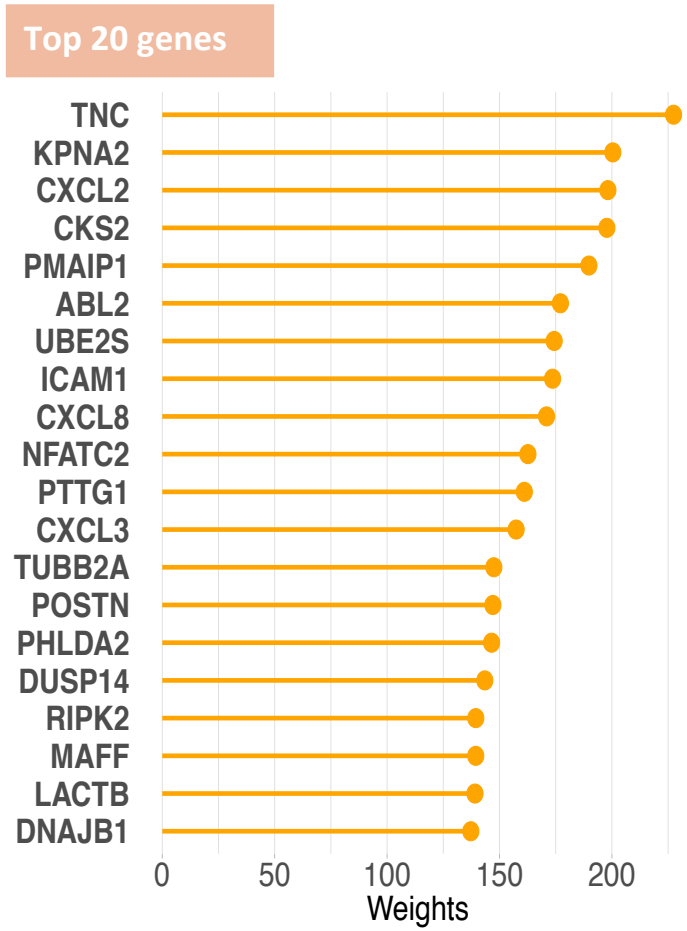
Fib GEM 16



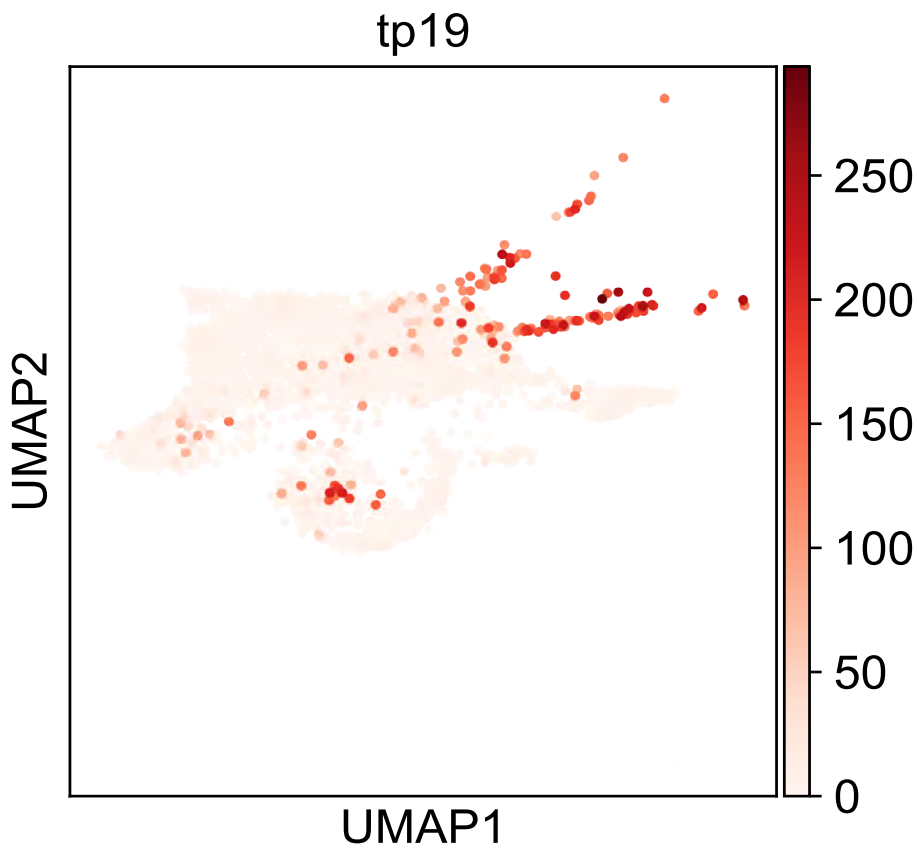
Cell Type

Fib_6

Fib_2



Fib GEM 19

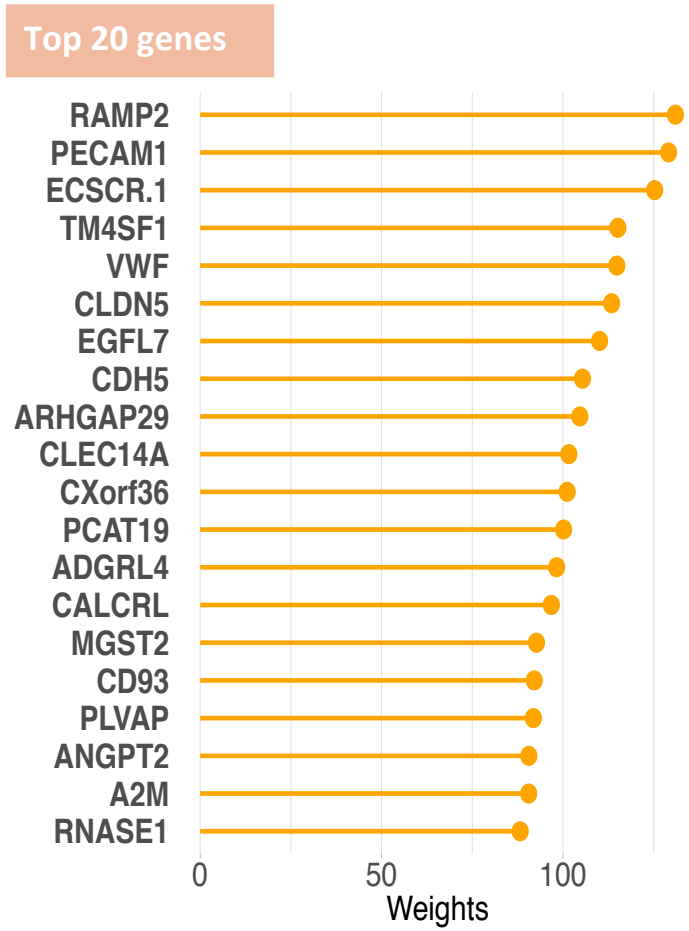


Cell Type

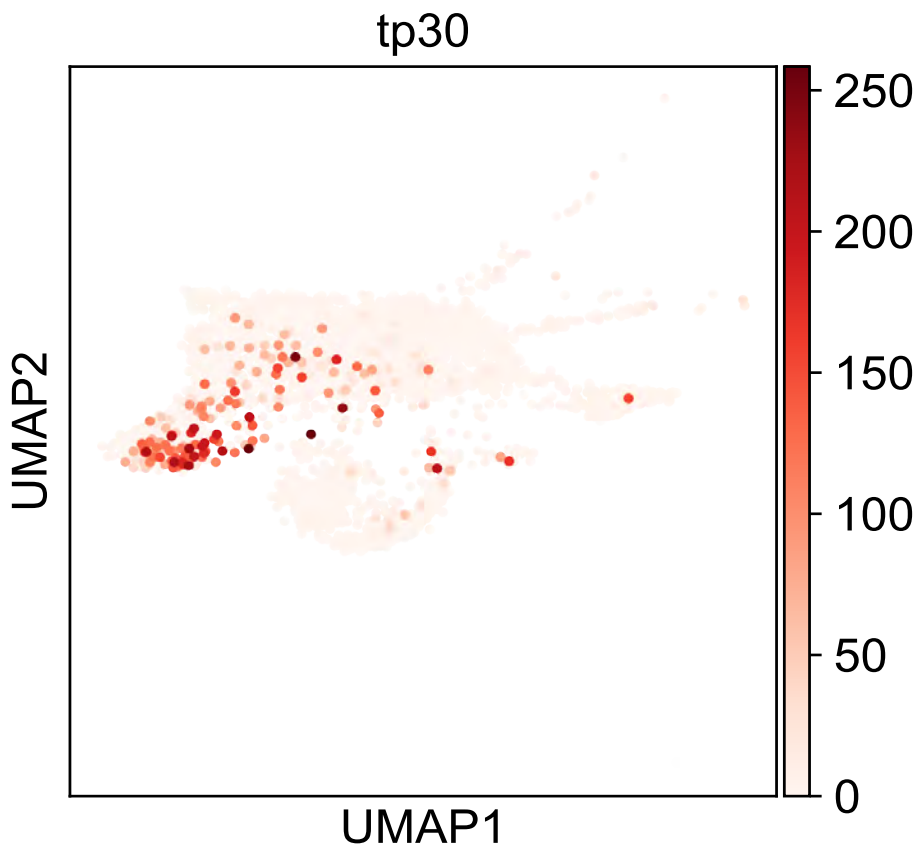
Fib_1

Fib_2

Fib_4



Fib GEM 30

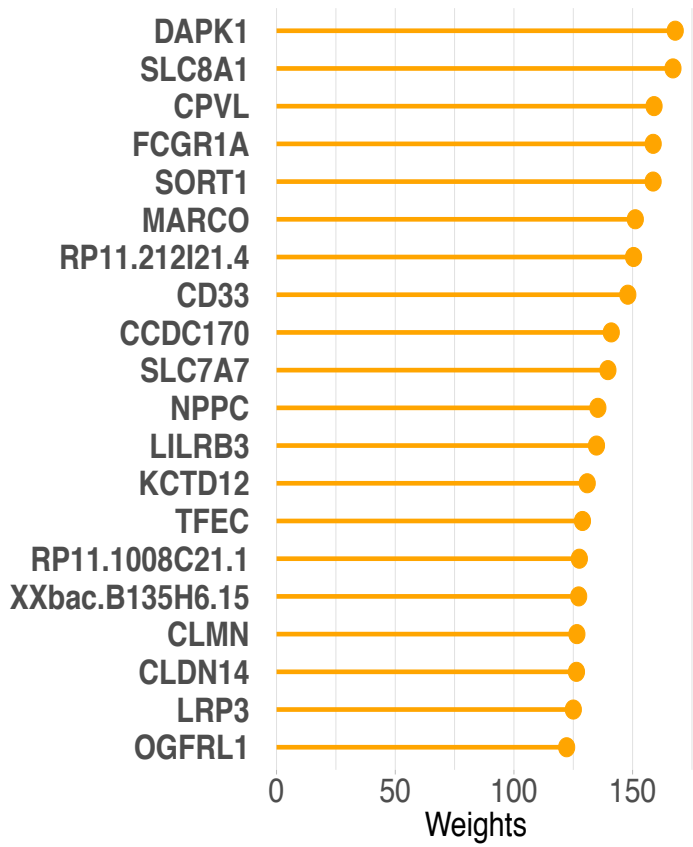


Cell Type

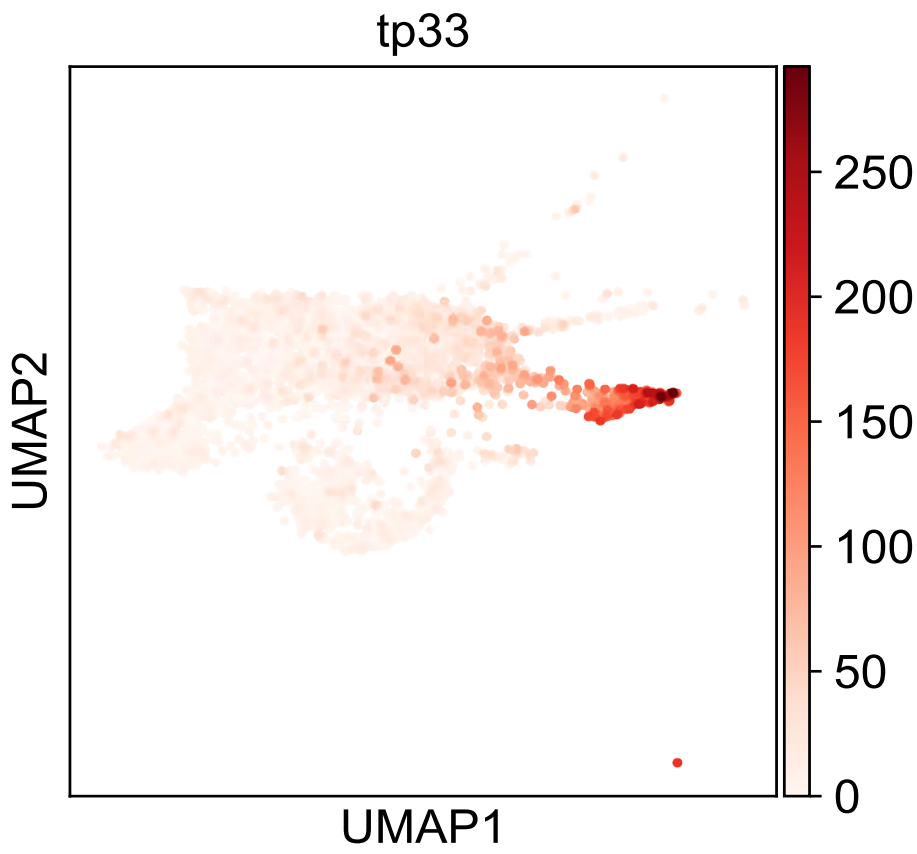
Fib_5

Fib_3

Top 20 genes



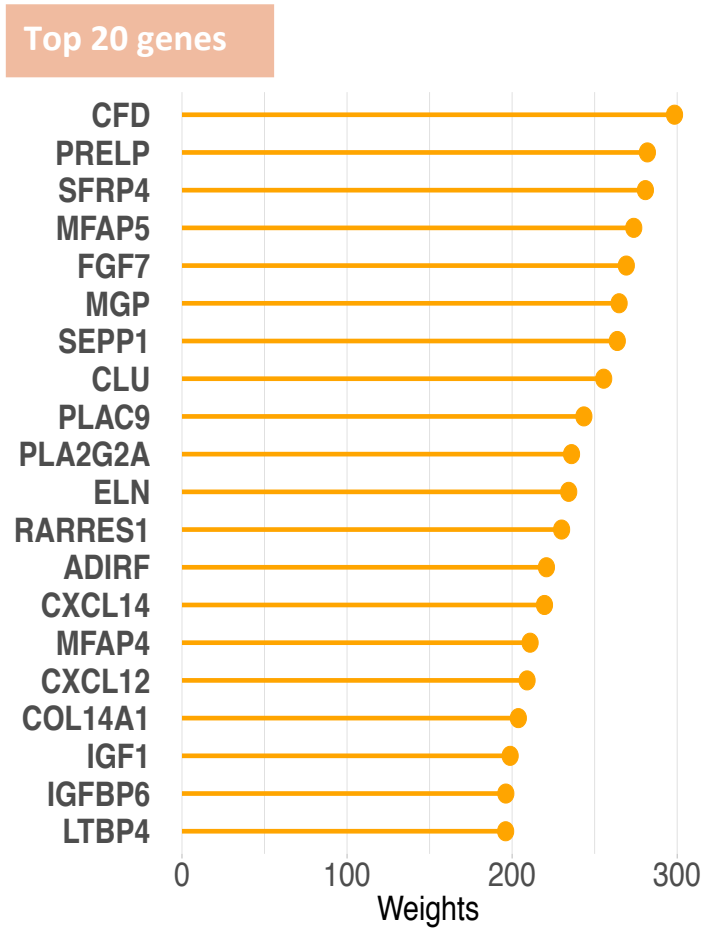
Fib GEM 33



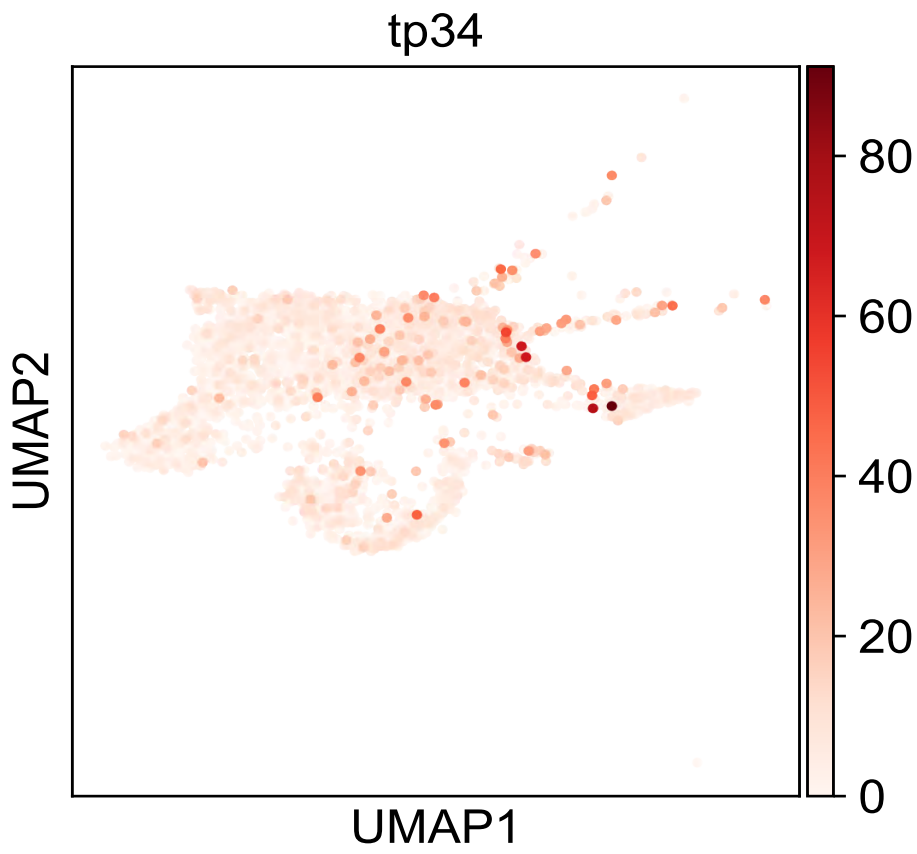
Cell Type

Fib_1

Fib_2



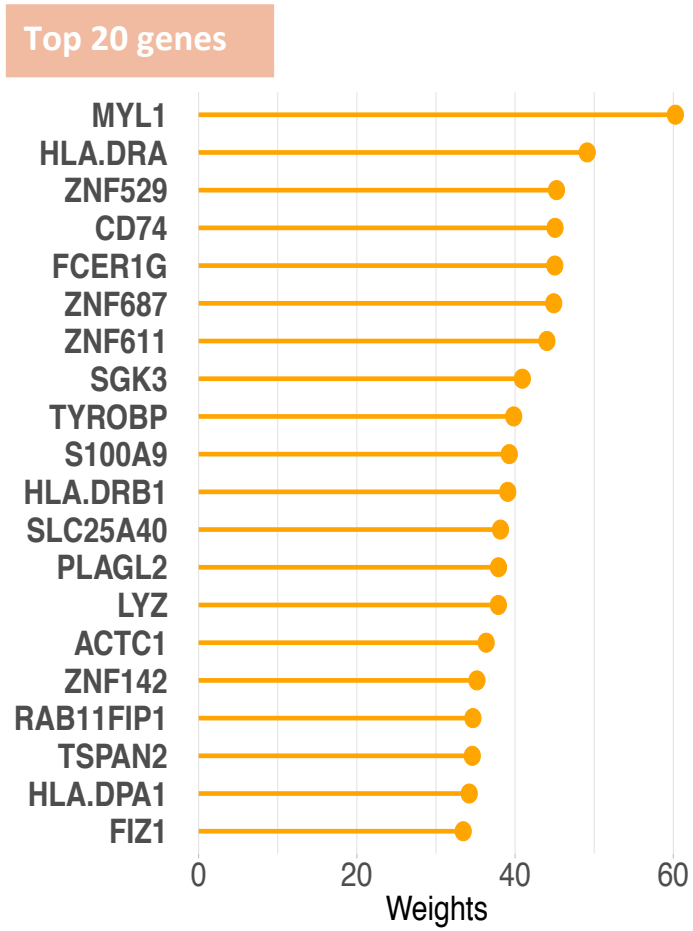
Fib GEM 34



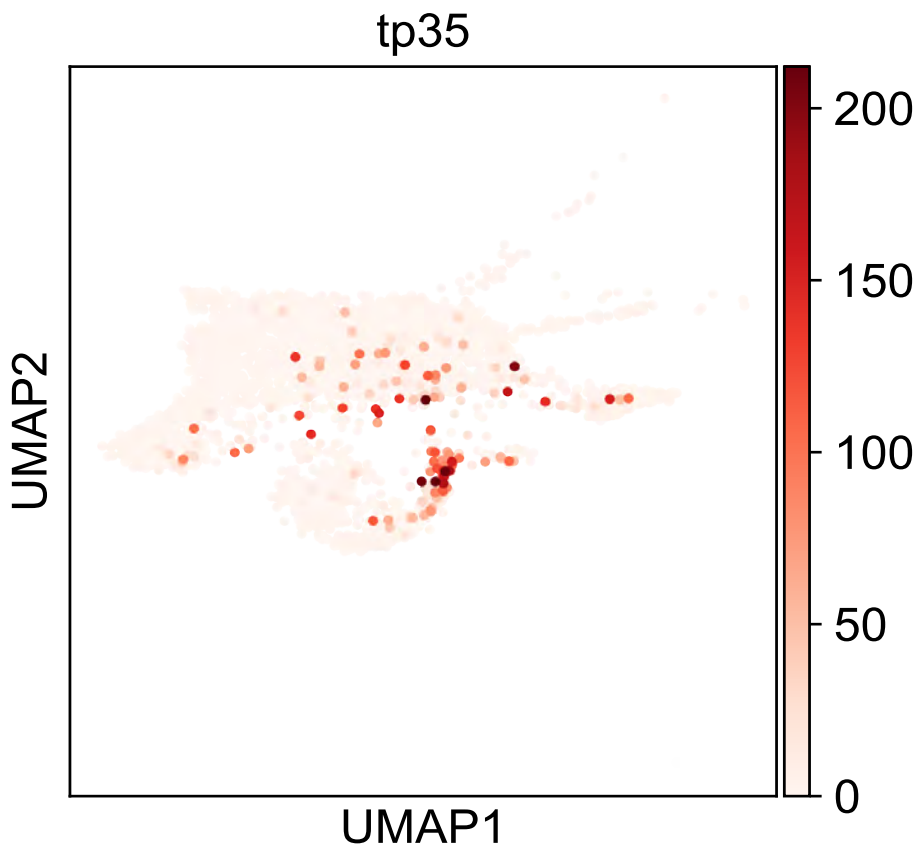
Cell Type

Fib_1

Fib_2



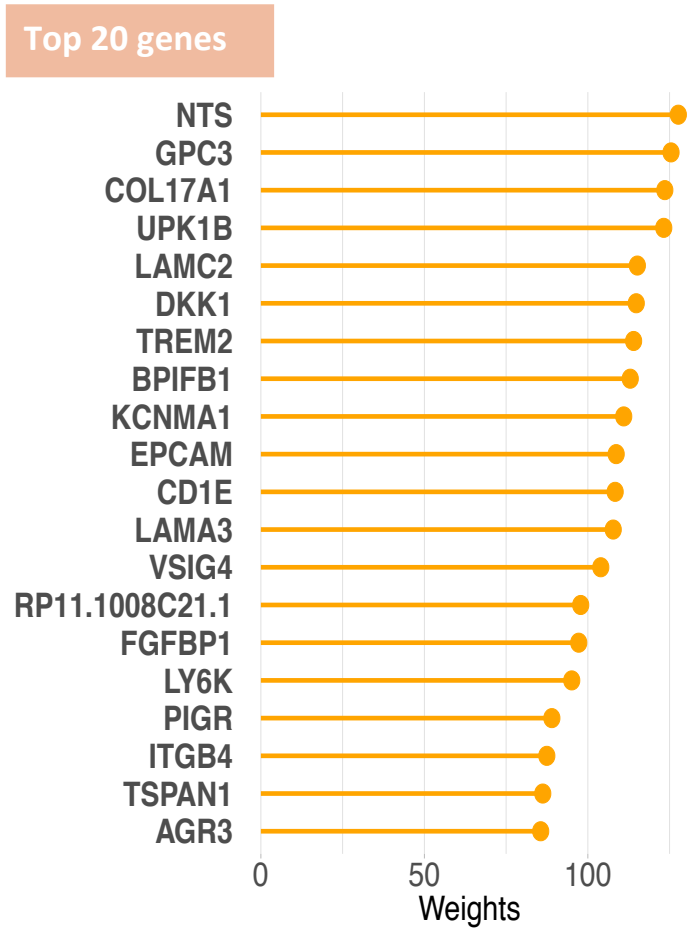
Fib GEM 35



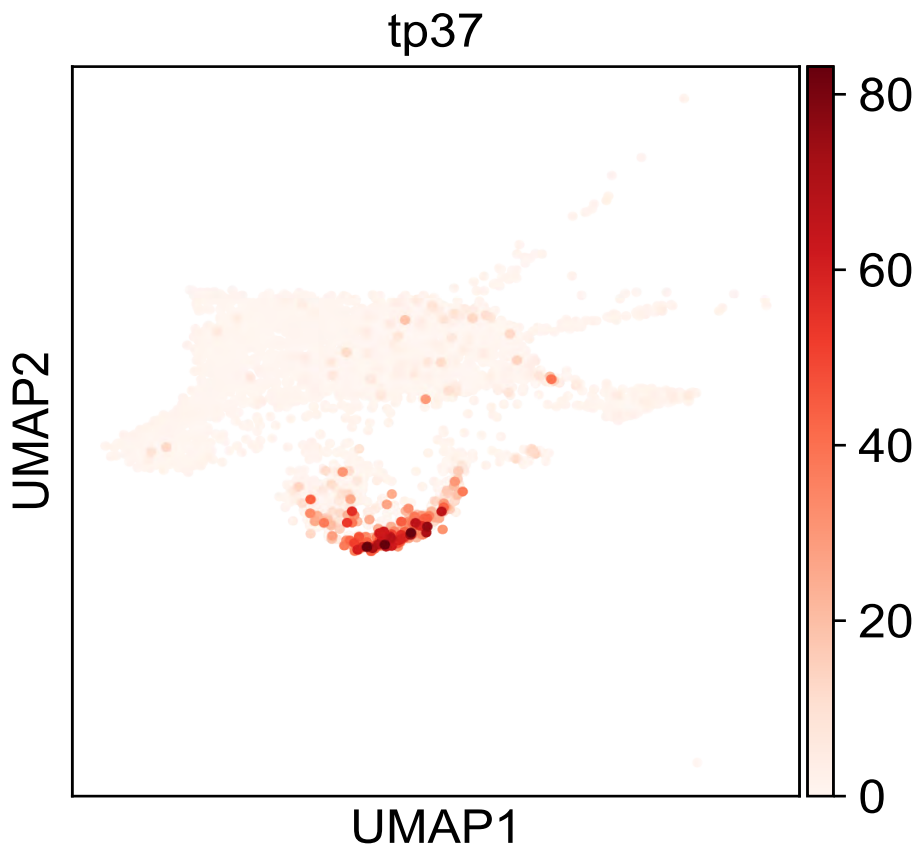
Cell Type

Fib_4

Fib_1

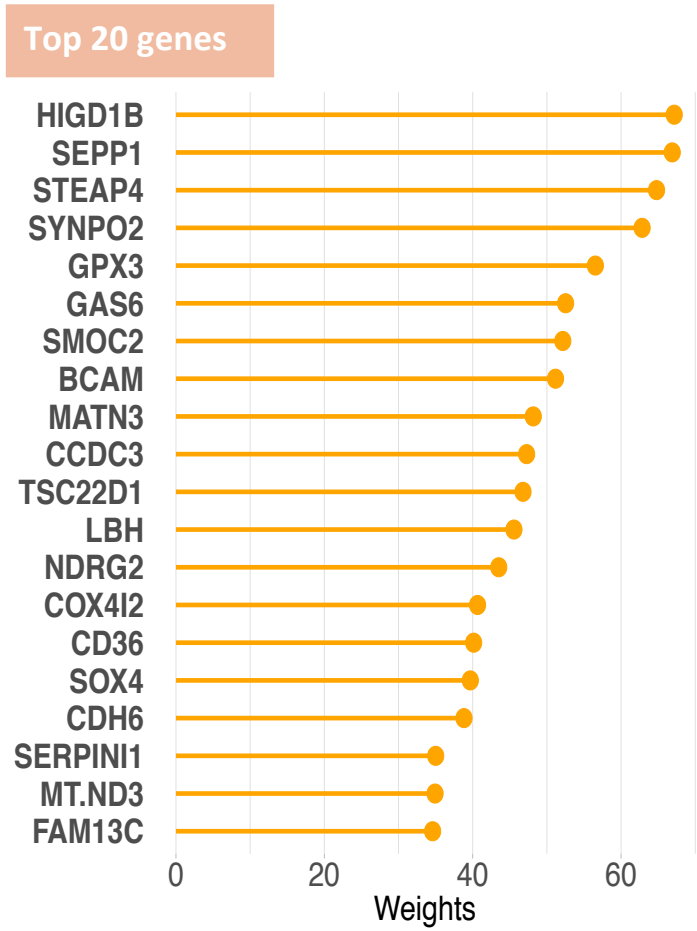


Fib GEM 37

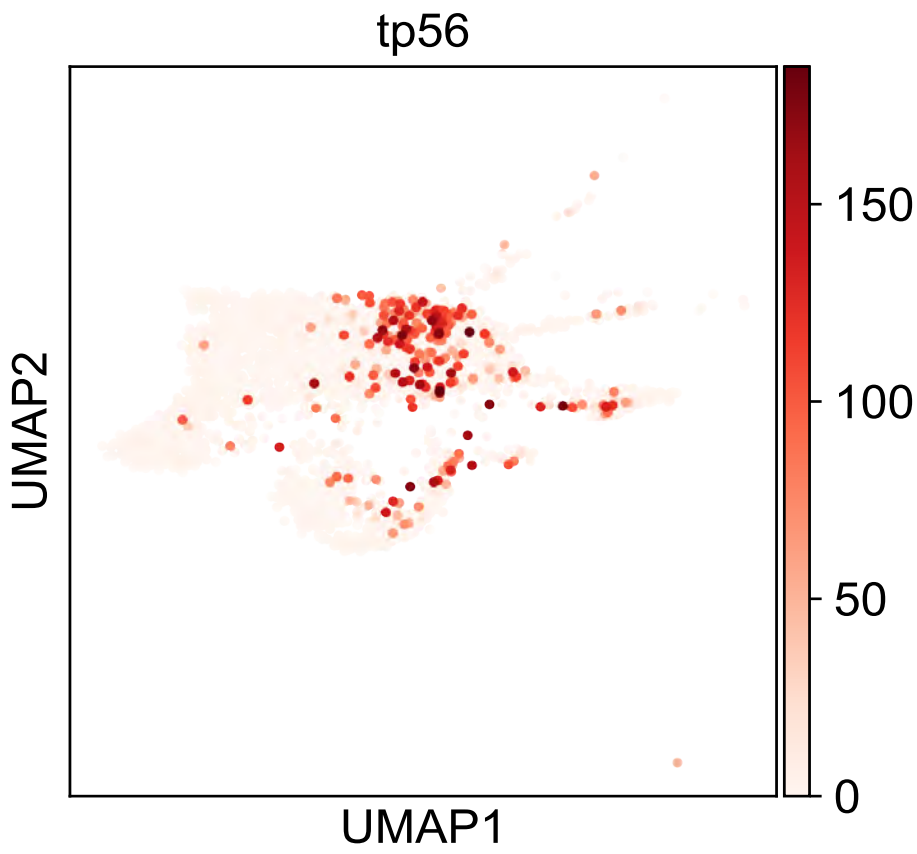


Cell Type

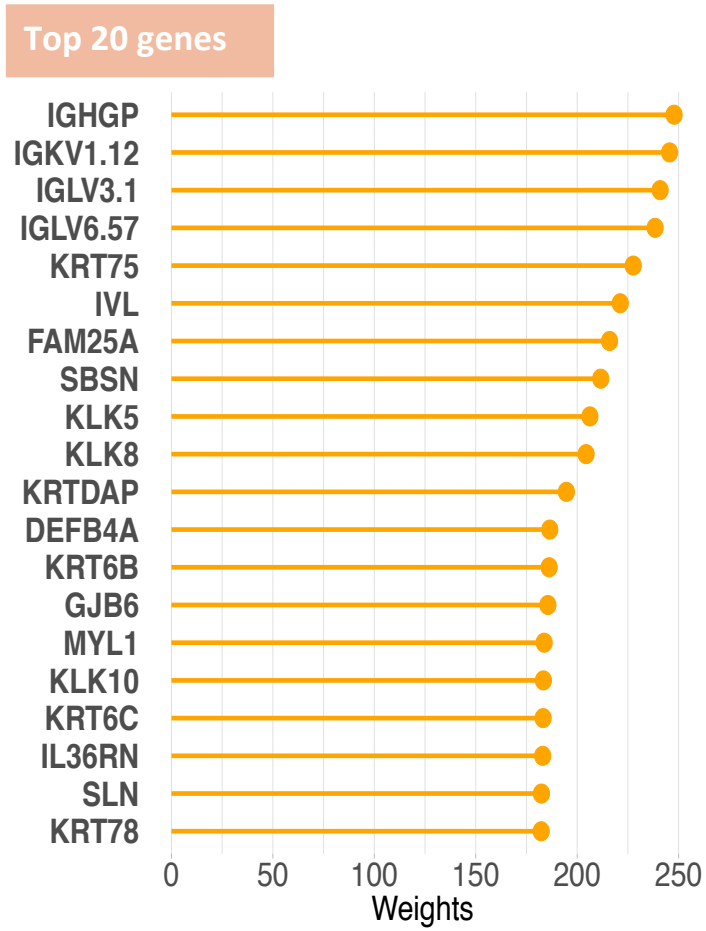
Fib_6



Fib GEM 56



Cell Type

Fib_1

Fib_4