Detection of Patlent-Level distances from single cell genomics and pathomics data with Optimal Transport (PILOT)

Mehdi Joodaki[1], Mina Shaigan[1], Victor Parra[1], Roman D. Bülow[2], Christoph Kuppe[3], David L. Hölscher[2], Mingbo Cheng[1], James S. Nagai[1], Nassim Bouteldja[2], Vladimir Tesar[5], Jonathan Barratt[6,7], Ian S.D. Roberts[8], Rosanna Coppo[9], Rafael Kramann[3,4], Peter Boor[2,*], and Ivan G. Costa[1,*]

[1]Institute for Computational Genomics, Joint Research Center for Computational Biomedicine, RWTH Aachen University Medical School
[2]Institute of Pathology, Laboratory of Nephropathology, RWTH Aachen University Medical School
[3]Institute of Experimental Medicine and Systems Biology, RWTH Aachen University
[4]Department of Internal Medicine, Nephrology and Transplantation, Erasmus Medical Center
[5]Department of Nephrology, 1st Faculty of Medicine and General University Hospital, Charles University, Prague, Czech Republic
[6]John Walls Renal Unit, University Hospital of Leicester National Health Service Trust, Leicester, United Kingdom
[7]Department of Cardiovascular Sciences, University of Leicester, Leicester, United Kingdom
[8]Department of Cellular Pathology, Oxford University Hospitals National Health Services Foundation Trust, Oxford, United Kingdom
[9]Fondazione Ricerca Molinette. Regina Margherita Children's University Hospital, Torino, Italy
[*]pboor@ukaachen.de, ivan.costa@rwth-aachen.de

ABSTRACT

Although clinical applications represent the next challenge in single-cell genomics and digital pathology, we are still lacking computational methods to analyse single-cell and pathomics data at a patient level for finding patient trajectories associated with diseases. This is challenging as a single-cell/pathomics data is represented by clusters of cells/structures, which cannot be easily compared with other samples. We here propose Patlent Level analysis with Optimal Transport (PILOT). PILOT uses optimal transport to compute the Wasserstein distance between two single single-cell experiments. This allows us to perform unsupervised analysis at the sample level and to uncover trajectories associated with disease progression. Moreover, PILOT provides a statistical approach to delineate non-linear changes in cell populations, gene expression and tissues structures related to the disease trajectories. We evaluate PILOT and competing approaches in disease single-cell genomics and pathomics studies with up to 1.000 patients/donors and millions of cells or structures. Results demonstrate that PILOT detects disease-associated samples, cells, structures and genes from large and complex single-cell and pathomics data.

Introduction

Single-cell genomics and digital pathology methods are revolutionary technologies, which in principle allow researchers to computationally dissect molecular, cellular and structural changes in human tissues[1,2]. However, clinical application of single-cell sequencing, i.e. finding cells and their markers for patient stratification and personalized treatments[3], is still in its infancy. Recent efforts include the use of single-cell and spatial transcriptomics to dissect the progression of acute human myocardial infarction[4], to characterize COVID-19 patient's disease severity[5] and the development of pancreatic ductal adenocarcinomas[6]. Computational analysis of these datasets mostly leverage standard methods for single-cell sequencing analysis, i.e. finding genes with differential expression in control vs. disease for each cell cluster. These method approaches require a-priori patient classifications, i.e. control vs. disease. Therefore, they cannot be used to find disease progression trajectories and respective molecular markers. Particularly challenging is the unstructured nature of single-cell experiments, i.e. identification of thousands of cells per patient and each cell is described by thousands of molecular features. Advanced methods addressing the problem of patient clustering and trajectory inference are currently lacking. Pathomics, i.e. the use of machine learning methods to extract morphological structures in histology slides, is likely to revolutionize pathology by

producing reproducible and fast quantification of histological slides for disease diagnosis, as recently shown in kidney[7]. Similar to single-cell sequencing, pathomics data of an individual slide is represented by thousands of anatomical structures, which are described with morphometric features, e.g. their individual shape and size. Similar to single cell genomics, no analytical method to compare histology of slides based on clustered structures is available.

Only a few methods allow the analysis of single-cell genomics datasets at a sample level. PhEMD[8] is based on earth moving distance (EMD) to measure the distance between specimens (single-cell samples), where the distance between specimens was based on clustering representations from a diffusion-based space. This method was successfully used to explore the response of cell lines to drug effects. Nevertheless, it is based on a diffusion maps and pseudotime estimates on the cell level. This explicitly assumes the presence of a cellular continuum (cell differentiation/activation) between all cells in the scRNA-seq experiments. Therefore, it is not suitable for the analysis of single-cell experiments measured in heterogeneous/whole organs samples. Moreover, PhEMD lacks methods for interpretation of the results, i.e. detection of molecular and cellular features explaining predictions. Ravindra and colleagues [9] proposes the use of graph attention networks for classification of scRNA-seq samples and applied this to predict the disease states of multiple Sclerosis[10]. More recently, SCANCell[11], which explores association networks across cell clusters, was proposed and applied for the analysis of systemic lupus erythematodes. Finally, multiscale PHATE[12] explores non-linear embeddings and multi-resolution representation to find the individual sample representation. This method uses supervised filters to find groups of cells related to COVID-19 mortality. With exception of PhEMD, all other related methods[9,11,12] require labels of patients for their analysis and cannot be used in the unsupervised analysis of patient-level single-cell experiments.

Here, we propose PILOT, a Patient level distance with Optimal Transport (OT) to find similarities between samples measured with single-cell or pathomics data with the optimal transport based Wasserstein distance[13] (Fig. 1A). This can be used to infer patient-level disease progression trajectories or to cluster patients regarding potentially novel disease sub-types. For this, PILOT encodes clustering of single-cell or pathomics data[14] as sample conditional probability distributions. It next uses OT to compare the cluster distribution between two samples by considering the distance between the cells as transport cost. Due to the stochastic nature of single-cell data, i.e. cellular recovery can be distinct across samples or be affected by the cellular composition of the biopsy, we use a Dirichlet prior and maximum a posteriori to regularize cluster distribution estimates. Next, PILOT explore diffusion maps[15] to infer sample level disease progression trajectories. Finally, PILOT explores robust non-linear regression models[16] to find cell types, genes or morphological features associated with disease progression. Thus PILOT represents the first approach to detect unknown patient trajectories and uncover cellular, molecular and morphometric features associated with disease progression.

Results

Evaluation of methods for patient-level analysis

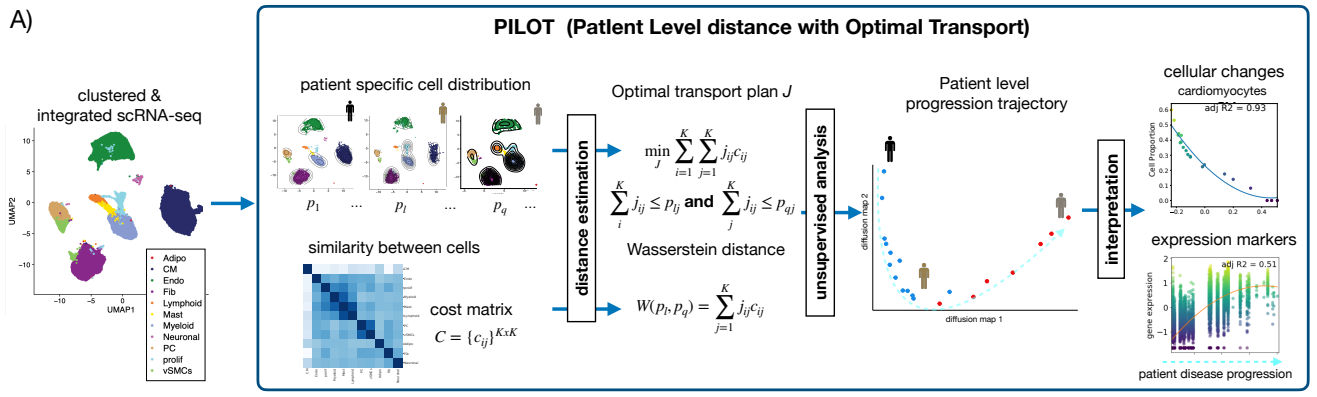
We compare the results of PILOT and PhEMD[8] in the recovery of known patient groups of four single-cell RNA data sets (PBMCs from systemic lupus erythematodes[17], pancreatic ductal adenocarcinoma [PADC][6], acute myocardial infarction[4], PBMC of patients with COVID-19 infection[5]) and four pathomics data with morphological features of glomeruli and tubules from two distinct disease cohorts (kidney IgA nephropathy of the VALIGA study [IgAN][18] and kidney biopsies of the Aachen cohort [AC][7]). For most of these data, we use disease status (control and case) as labels for the clustering task. For the Kidney IgAN, patients/donors were labeled by their estimated glomerular filtration rate (eGFR; normal, reduced and low), while for the COVID-19 samples were labeled as control, mild and severe cases (Fig.1:B). As a baseline, pseudo-bulk libraries for each patient were estimated and a standard RNA-seq pipeline was performed (see method sections for details). For scRNA-seq data, we used the clustering and normalization as provided in the original manuscript as input for PILOT. For pathomics data, structure segmentation was performed with FLASH followed by graph based clustering as proposed before[7].

First, we evaluate the performance of distance metrics of PILOT and competing methods with the Silhouette index[19] (Fig. 1C; Sup. Table S1). PILOT had highest values in 6 out 8 data sets and obtained a higher average silhouette than PhEMD (Fig. 1D; p-value < 0.05; Friedman-Nemenyi test). Next, we used graph-based clustering[20] on these distance matrices and evaluated results with the adjusted Rand index (ARI)[21] (Fig. 1C; Supplementary Table S1). PILOT obtained highest ARI score in all eight data sets (Fig. 1D; p-value < 0.01; Friedman-Nemenyi test). Due to the large size of the data sets, the computing time is also a relevant aspect. In the largest data set (Lupus PBMC), PILOT required 58.12 seconds versus 2.14 seconds of the simple pseudo-bulk and 16.50 minutes of PhEMD [1]. These indicates that PILOT (and competing methods) can be efficiently run in large data with millions of cells or tissue structures.

PILOT reveals disease progression trajectories

We observe that diffusion maps estimated with PILOT recovered trajectory like structures in all analysed data sets (Sup. Fig. S1-S2). Therefore, we modelled these trajectories to obtain disease progression scores. We observe that a disease

[1] This computing time of PhEMD also includes clustering of single cells, as this is an require step in PhEMD framework



B)

Data	Type	# cells	# samples	# Cell Types	# Classes
COVID-19 PBMC	scRNA-seq	993171	151	10	3
Myocardial Infraction	scRNA-seq	115517	20	11	2
Lupus PBMC	scRNA-seq	1263676	261	11	2
Pancreas (PDAC)	scRNA-seq	57530	35	10	2
Kidney IgAN (Glomeruli)	Pathomics	24227	634	14	3
Kidney IgAN (Tubule)	Pathomics	64493	634	15	3
Kidney AC (Glomeruli)	Pathomics	4731	57	28	2
Kidney AC (Tubule)	Pathomics	56998	57	7	2

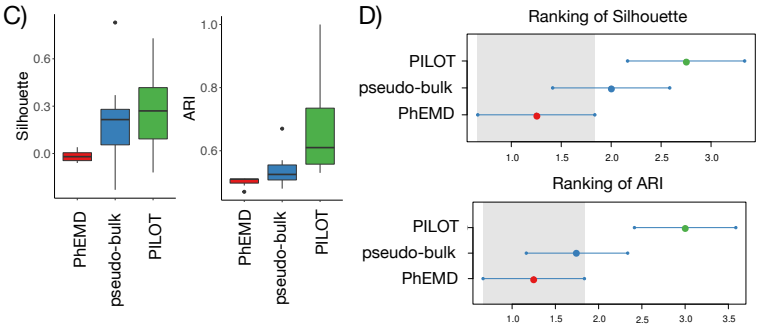


Figure 1. PILOT schematic and benchmarking: **A.** PILOT receives as input a clustered and integrated scRNA-seq data. Next, it estimates patient specific clusters distribution and a cost function (similarity between clusters). These are used as input for an optimal transport algorithm[13]. The transport costs provides a distance (Wasserstein), which can be used for unsupervised analysis such as diffusion map and disease progression analysis. Finally, PILOT uses non-linear regression to find cellular, molecular or morphological features associated with disease progression. **B.** Description of data sets (pathomics and scRNA-seq) used to benchmark PILOT. **C.** Box plot with Silhouette and ARI values (y-axis) for distinct evaluated methods (x-axis). **D.** Ranking distributions (x-axis) for distinct methods regarding Silhouette or ARI values. Higher ranks indicates better performance.

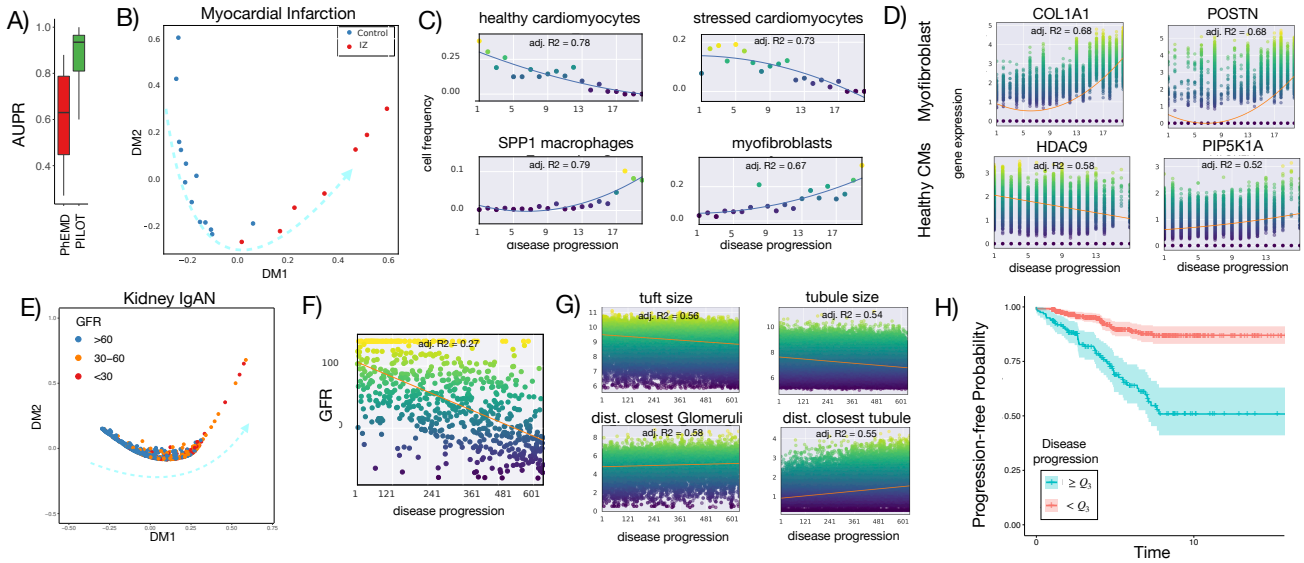


Figure 2. Sample level analysis of myocardial infarction and kidney IgAN : **A** Distribution of AUPR values of disease progression scores for all benchmarking data sets for PhEMD and PILOT. **B** Diffusion map and disease progression trajectory in myocardial infarction. **C** Fraction (y-axis) of top four cells (highest R2) vs. disease progression (x-axis). **D** Expression (y-axis) of selected genes with significant changes in disease progression (x-axis) for healthy cardiomyocytes and myofibroblast cells. **E** Diffusion map and disease progression trajectory in kidney IgAN. **F** GFR values (y-axis) versus disease progression (x-axis). **G** Morphometric values (y-axis) of kidney structures vs. disease progression (x-axis). **H** Kaplan-meyer plot with the time to kidney failure from patients within the upper quartile (high progression score) vs. other samples (low progression score).

progression score recapitulates the severity of diseases in the data sets with disease severity associated labels (Sup. Fig. S3). PILOT obtains an average 0,86 area under the precision recall curves (AUPR), where samples are ranked by disease progression and contrasted with their disease label (Fig. 2A; Supp. Table. S2). Moreover, PILOT outperforms PhEMD in all evaluated scenarios (Sup. Table S2; p-value > 0.005; Friedman-Nemenyi test)[2].

Next, we analyze a trajectory from normal heart tissues towards ischaemic zone (IZ) tissues in acute myocardial infarction. The trajectory reflected the known annotation of samples with exception of two cases in the decision boundary between controls/IZ samples (Fig. 2B). We use non-linear regression methods to find cellular and molecular changes associated with the disease progression score. Major cellular changes, as indicated by highest adjusted R^2 values, include a quadratic increase of SPP1 positive macrophages and myofibroblasts; a quadratic decrease of healthy cardiomyocytes followed by a latter and smoother decrease of stressed cardiomyocytes during disease progression (Fig. 2C; Sup. Fig. S4). These patterns are in accordance with cellular changes in early myocardial infarction, which include loss of myogenic tissue, inflammation and fibrosis. At the gene level, PILOT identifies genes related to extracellular matrix to be associated in myofibroblasts (Supp. Fig. S5). These include COL1A1 and POSTN, which have an overall quadratic increase in gene expression upon disease progression (Fig. 2D;Supp. Fig. S7). We observe genes associated with muscle contraction and cardiomyopathy as top ranked in healthy cardiomyocytes (Supp. Fig. S5). One of these genes with reduced expression is HDAC9, which functions as a signal-responsive repressor of cardiac hypertrophy by inactivating MEF2 transcription factors[22]. A gene with increase in gene expresison is PIP5K1A, which is linked to activation of myogenic differentiation via AKT activation and calcium release in cardiomyocytes (Fig. 2D; Supp. Fig. S6)[23]. Altogether, these results display how PILOT is able to uncover molecular and cellular changes related to the onset of myocardial infarction.

Next, we performed a trajectory analysis of the pathomics data of kidney IgAN biopsies. As distinct morphological features are used to describe tubules and glomeruli, it is not possible to cluster these structures together so these data are independently analysed as in[7]. We combine the two disease progression scores by their sum, which yields similar or higher AUPR scores than the trajectories considering one structure at a time (Fig. 2E; Supp. Fig. S8). This trajectory is associated with a linear decrease in eGFR (Fig. 2F), which is the current surrogate used in clinical practice to estimate kidney function. Regarding morphometric features, PILOT detects a significant linear decrease of glomerular tuft sizes and tubule sizes; and linear increase in distance between glomeruli and tubules as major characteristics of disease progression (Fig. 2G; Supp. Fig. S9- S10). These are indicative for interstitial fibrosis, tubular atrophy and glomerulosclerosis, which are all hallmarks of kidney function decline. Finally, we made use of a prognostic variable from the kidney IgAN biopsies, which indicates if patients progressed to kidney

[2]No trajectory analyse was possible in pseudo-bulk due to the absence of linear distributions of samples in the diffusion maps

failure. PILOT progression score is more associated with kidney failure (p-value of 2.4e-11; likelihood ratio test) than a multivariate model combining all morphometric variables (p-value of 7e-05; likelihood ratio test) or the use of individual morphometric variables (Supp. Table S3). Indeed, samples with higher disease progression scores (top 75 quartile) have higher chance of kidney failure than other samples (Fig. 2H). Altogether, these results reinforce the power of PILOT in discriminating disease outcome from pathomics data.

Discussion

The technological improvements in single cell genomics and digital pathology are providing us with clinically rich and large sized data describing cellular and morphological changes in diseases. We present here PILOT—the first computational pipeline for detection and feature characterization of disease trajectories from single cell genomics or pathomics data. By using a comprehensive benchmarking with eight data sets, we show that PILOT is superior to the competing approaches in both clustering and disease trajectory prediction. Another important aspect, which is only addressed by PILOT, is the interpretation of predictions. PILOT uses a robust non-linear regression model, which can predict cellular populations, marker genes or morphological features associated with the disease progression.

We revisited the analysis of the largest single cell genomics data on myocardial infarction[4], where PILOT could successfully predict a trajectory from controls towards ischemic samples. This allowed us to find non-linear changes in cell composition during cardiac remodelling, i.e. quadratic decrease of healthy cardiomyocytes and quadratic increase of myofibroblast and macrophages cells. Similarly, PILOT dissected gene expression programs associated with these changes, such as non-linear increase in expression of extra cellular matrix related genes in myofibroblasts and increase of genes associated with cardiomyopathy in healthy cardiomyocytes. We also evaluated the power of PILOT in inferring a disease trajectory of pathomics data from patients with kidney IgAN. We show that disease progression as estimated by PILOT is related to eGFR and provides a better predictor for future kidney failure than the use of the morphometric features alone or together. These highlights the power of PILOT in finding non-linear manifolds to model disease progression. Molecular studies of diseases are increasingly based on multi-modal measurements; gene expression, protein abundances, chromatin accessibility and histology images at either single cell and/or spatial level[24]. Future work of sample level analysis will require methods considering multimodal and spatial nature of these data into account.

Methods

PILOT

We formalize main steps of PILOT[3]. In analyzing single-cell experiments, assume there are L single-cell matrices $\mathscr{X} = \{\mathbf{X}_1, ..., \mathbf{X}_l, ..., \mathbf{X}_L\}$, where $\mathbf{X}_l \in \{\mathbb{R}\}^{N_l \times D}$. Where, N_l represents the number of cells in sample l and D is the number of features (genes), which is common between all experiments. In practice, PILOT receives as input an single matrix $\mathbf{X^I} \in \{\mathbb{R}\}^{N \times D}$ after integration of the matrices in $\mathscr{X}$, where N represents the total number of cells. To keep the sample information, we define an vector indicating the sample identity of cell i, i.e. $s = \{s_1, ..., s_N\}$, where $s_i \in \{1, ..., L\}$ indicates to which sample (patient) the cells belongs to. Moreover, a standard way to summarise the information in cells is to group then using a clustering algorithm, which will output as a result a vector $y = y_1, ..., y_N$, where $y_i \in \{1, ..., K\}$ indicates the group of cell i.

Our main problem is to estimate a distance between scRNA-seq data measured over two distinct samples (patients), where a scRNA-seq is represented as a set of clustered cells. Here we explore the concepts of the optimal transport based Wasserstein method[25] to compare two samples by representing samples as distributions of cells. Let the probability to have a cluster k for patient l:

$$p_{lk} = P(y = k | s = l). \tag{1}$$

This can be used to define a probability distribution vector $p_l = (p_{l1}, ..., p_{lK})$.

From these, we use optimal transport to find the optimal transport plan $J = \{j_{ij}\}^{KxK}$ between two distributions p_l and p_q describing samples l and q by considering that there is a cost c_{ij} associated with move some mass between cluster i to cluster j:

$$\min_J \sum_{i=1}^{K} \sum_{j=1}^{K} j_{ij} c_{ij}, \tag{2}$$

such that $j_{ij} \geq 0$, $\sum_i^K j_{ij} \leq p_{lj}$ and $\sum_j^K j_{ij} \leq p_{qj}$ and $\sum_{i=1}^K \sum_{j=1}^K j_{ij} = 1$.

[3]For simplicitly, we focus here on single cell data, but the same formalism applies to phathomics data

For a given optimal transport plan J, the Wasserstein distance is calculated as:

$$W(p_l, p_q) = \sum_{j=1}^{K} j_{ij} c_{ij}. \tag{3}$$

The optimal transport matrix J can be estimated using a minimum flow cost algorithm[25]. By estimating the Wasserstein distance between all pairs of samples, we obtain a distance matrix W between all samples.

This framework, which is also denoted Earth Moving Distance[26], is also explored by PhEMD. PILOT however explores important issues, which are crucial in noisy and large nature of single cell and pathomics data. These are namely: (1) how to obtain robust estimates of probability distributions due to potential cell content bias and low cell coverage of individual samples; (2) how to estimate the cost matrix C on the large matrices $\mathbf{X^T}$; and (3) by offering statistical models to select features associated with disease progression. These are described below.

Robust estimation of sample probability distributions

The coverage of cells per cluster can vary across distinct samples. This effect is potentially higher in diseased samples, due to their lower cell viability. Let $z_l = (z_{l1}, ...z_{li}, ..., z_{lK})$ be a vector, where z_{lk} is the number of cells in a cluster k and sample l. We define the following hierarchical model:

$$\begin{aligned} z_l &\sim Multi(N_l, \Theta) \\ \Theta &\sim Dir(\alpha) \end{aligned} \tag{4}$$

where N_l is the number of cells in the sample l, Θ is a random variable representing distributions p and α is the hyper-parameter of a Dirichlet distribution. The posterior distribution can be re-written as

$$p(\Theta | z_i, \alpha) \sim Dir(z_i + \alpha) \tag{5}$$

We can used this to obtain maximum a posteriori estimates of Θ, i.e.

$$\hat{p}_{lk} = \frac{n_{kl+\alpha_k}}{n_l + \sum_{i=1}^{K} \alpha_i}. \tag{6}$$

Here, we use the following parametrization of the prior $\alpha_k = N_k / N * c$, where c is set as 0.1 as default. This prior adds "pseudo cell counts", which are weighted by the cell distribution of all samples. This prior mitigate the fact rare cells might not be observed in samples with low cell coverage.

Estimation of Cost Matrix

The optimal transport also considers the cost c_{ij} of transporting a distribution mass from a a cluster i to a cluster j. For this, we generate a median representation for each cluster (in the PCA space) and estimate the cosine distance between the median value of clusters i and j as the cost to transport masses. The cosine distance between the median of cluster i (M_i) and cluster j (M_j) is defined as:

$$Cosine(M_i, M_i) = 1 - \frac{M_i \cdot M_j}{\|M_i\|_2 \left\|M_j\right\|_2} \tag{7}$$

Of note, PhEMD uses centroids and Euclidean distances in non-linear embedding to estimated the cost matrices. The non-linear embedding assumes as cellular continuum between all cells in the data, which is not present in whole tissue single cell or pathomics data. Moreover, median values reduces the effects of outliers.

Clustering and disease trajectory estimation

The matrix W (Eq. 3) provides the distance between all samples. Clustering analysis can be performed by providing W as input to a Leiden clustering algorithm[20]. Another important feature is the inference of disease trajectories. For this, we obtain a similarity from the matrix (or kernel) $W^s = 1 - W$ and estimate the graph Laplacian

$$L = D^{-1} W^s D^{-1}, \tag{8}$$

where D is a diagonal matrix with $d_{ii} = \sum_j W^s_{ij}$. Next, we normalize it:

$$M = (D^L)^{-1} L, \tag{9}$$

where D^L is a diagonal matrix with $d^L_{ii} = \sum_j l_{ij}$. Finally, we define a t-step random walk over matrix M, where M^t_{ij} is the probability of reaching sample j from i at time t is

$$P(X_t = j | X_{t-1} = i) = M^t_{ij}. \tag{10}$$

Then, we perform the spectral decomposition of matrix M^t as:

$$M = D^{-1/2} V \wedge V^T D^{1/2}, \tag{11}$$

which $\wedge$ and V are the eigenvalue and eigenvectors matrices. Ultimately we order the eigenvalues (largest to smallest) and use eingevectors with highest values for obtaining diffusion maps. Moreover, we use the eigenvector with highest eigenvalue to rank the samples and to obtain a disease progression score $t = t_1, \ldots, t_L$, where t_l is the ranking of the sample l.

Identification of molecular and structural features associated with disease progression

PILOT uses step-wise regression models to identify features (cellular abundances, gene expression or structural properties), whose values are consistent changing across disease trajectories. For this, it fits regression models with linear, quadratic, and linear-quadratic terms for each feature and use statistical tests to determine the goodness of fit. We then report the most significant model for each feature.

Let x_{ij} represent the expression of a a feature (gene) j in a cell i, and p_i the pseudo-time variable (associated with the pseudo-time of its sample, $p_i = t_{s_i}$). In short, we fit the three regression models to the jth feature:

$$\begin{aligned} \hat{x}_{\cdot j} &= \mu_j + \beta_{1j} p, \\ \hat{x}_{\cdot j} &= \mu_j + \beta_{1j} p^2, \\ \hat{x}_{\cdot j} &= \mu_j + \beta_{1j} p + \beta_{2j} p^2. \end{aligned} \tag{12}$$

where μ is the models' intercept, and β are the models' coefficients. The models are ranked based on r-squared (R^2) and we only consider the model with highest R-squared and at a significance level (p-values < 0.05). The same approach can be performed for distinct features such as the proportion of cell types in the sample or morphological features.

In the case of gene expression, one is mostly interested in finding cell type (cluster) specific genes. Therefore, we only consider cells i belonging to cluster k. Due to the sparsity of single cell data, i.e. dropout events, we observed a single side tailed distributions of residuals. Therefore, to improve the robustness of our regression models, we use the Huber least square criterion (Eq. 13)[27].

$$H(e_i) = \begin{cases} \frac{1}{2} e_i^2 & |e_i| <= \delta \\ \delta(|e_i| - \frac{1}{2}\delta) & |e_i| > \delta \end{cases} \tag{13}$$

where $e_i = x_{\cdot j} - \hat{x}_{\cdot j}$. The Huber regression lessens the effects of the outliers by using a term δ (default of 1.35). This defines residual values associated with outliers. For large δ, the regression is equal to ridge regression. Of note, the previous formulation requires a re-definition of the R-squared to consider the Huber regression penalization, that is:

$$\begin{aligned} R^2_{mod} &= 1 - \frac{SS_{res}}{SS_{tot}} \\ SS_{res} &= \sum_{j \left| x_{\cdot j} - \hat{x}_{\cdot j} \right| <= \delta} (x_{\cdot j} - \hat{x}_{\cdot j})^2 + \sum_{j \left| x_{\cdot j} - \hat{x}_{\cdot j} \right| > \delta} \delta(|x_{\cdot j} - \hat{x}_{\cdot j}| - \frac{1}{2}\delta) \\ SS_{tot} &= \sum_j (x_{\cdot j} - \bar{x}_{\cdot j})^2 \end{aligned} \tag{14}$$

Materials

Data sets

We use here public single cells and pathomics data sets for benchmarking of proposed methods (see Fig.1B).

Single cell data sets

Peng and colleagues[6] performed a single cell study on pancreatic ductal adenocarcinoma (PDAC). They characterized 11 healthy and 24 PDAC samples with a total of 57,530 scRNA-seq annotated in 10 major pancreas cell-types. As only raw data was provided (obtained from GSA: CRA001160 https://ngdc.cncb.ac.cn/gsa/browse/CRA001160), we reanalyzed the data using Seurat and re-annotated cell types using the same marker genes (Sup. Fig. S11).

Systemic lupus erythematosus is a common type of lupus that in the immune system pounds its tissues, yielding overall rash and tissue injury in the acted organs[17]. This single cell study characterized the gene expression of peripheral blood mononuclear cells (PBMCs) of 261 donors over 1,263,676 cells. Of these, 99 were healthy controls and 162 were disease patients. We used the normalized and clustered (11 groups) data sets provided in GEO (GSE174188; GSE174188_CLUES1_adjusted.h5ad.gz).

We have recently proposed a large study to characterize myocardial infarction[4]. We recovered 115,517 high quality cells, which were clustered in 33 cellular sub-types. We focus here on samples associated with healthy heart tissues (healthy controls and remote zones; n=13) and ischemic zone (IZ; n=7). We exclude one sample due to low quality (less than 1000 genes (average) per cell). A final dataset is provided in Zenodo (see below).

Finally, we obtained a study with PBMC cells from patients infected with Covid-19[28]. We only considered here patients with PMBCs (frozen or fresh cells). This totals 151 samples classified as either severe infection (n=70), mild infection (n=61) and control (n=20). For these samples, we have 993,171 cells, which were grouped in 10 major cell-types. The data was obtained in GSE158055.

Pathomics data sets

The VALIGA study is a European cohort with kidney biopsies and accompanying clinical data of patients with IgA nephropathy[18]. We used a pathomics pipeline developed by us[7] to detect and measure 3 morphometric features of 65,483 tubules and 14 features associated with 24,227 glomeruli structures, which could be detected in 634 biopsies. Patients were classified regarding their glomerular filtration rate (GFR): normal (GFR>60; n=400), reduced (30<GFR<=60; n=177) or low (GFR<=30; n=57). Lower GFR indicates lower kidney function. A second pathomics data set is the Aachen Cohort data, which includes 57 samples with healthy controls and distinct diseases/co-morbidity associated to lower kidney function[7]. We analysed the histology slides as before[7], which provided 4,731 glomeruli and 46,999 tubules, each quantified with 3/14 morphometric features. Patients were classified as either being healthy controls (n=17) or diseased (n=40).

We employed a uniform pre-processing pipeline utilizing Seurat[29] for the normalization and clustering of structures. First, we normalized the data with the function NormalizeData of Seurat, and next ran the ScaleData function. Next, we perform dimension reduction (RunPCA) and kept the 10 and 2 main components for glomeruli and tubules respectivelly. Next, clusters were found with the lieden algorithm by using the FindNeighbors and FindClusters functions, respectively. The same pipeline was performed on the kidney IgAN (VALIGA) and Aachen cohort data. As morphometric features are not comparable between glomeruli and tubules, these data are analysed independently, which results in four distinct data sets. At the patient level, PILOT also allowed the combination of the Wasserstein distance for glomeruli and tubuli, which provided an unique trajectory for each cohort.

Competing methods

Pseudo-bulk

We explore here pseudo-bulk as a baseline method for sample level analysis. For this, we take the original single cell and pathomics data and calculate the average of gene expressions and morphological features per sample. For pathomics data, we scale values to be between 0 and 1. Then, the principal components(PCA) for samples are computed. We detected knees in the variance plots and keep the PCs with highest variances. We computed the cosine distance, which was then used as input for diffusion maps and Leiden clustering.

PhEMD

PhEMD required the use of Monocle 2[30] to perform normalization, dimension reduction and clustering of cells. The inputs of PhEMD for all data sets are original count matrix for gene expressions and original pathomic measurements. An exception was the COVID-19 data set, We were not able to run PhEMD in a computer with 2TB of memory. Therefore, we provided the dimension reduced version with 50 first PCs. Among other parameters, PhEMD/Monocle2 requires the definition of a distribution function for expression values. We use as dafault Negative Binomial distributions to model single cell data sets as advised in the tutorials. For PBMC COVID-19 and Lupus data, we used Gaussianff, due to the fact no cluster was found with the default model. For pathomics data, we used truncated normal distributions, due to its best performance. We then used the distance marices provided by PhEMD as input for a diffusion map analysis as implemented in PILOT.

Acknowledgements

We thank the European Validation Study of the Oxford Classification of IgAN (VALIGA), funded by the Immunonephrology Working Group of the European Renal Association for access to VALIGA data. This work was funded by grants of the Deutsche Forschungsgemeinschaft (DFG-GE 2811/3) to I.G.C. and the clinical research unit KFO5011 for I.G.C., P.B., R.K. and C.K. and by the Bundesministerium für Bildung und Forschung (BMBF e:Med Consortia Fibromap) to I.G.C. and R.K. PB is supported by the German Research Foundation (DFG, Project IDs 322900939, 454024652, 432698239 445703531), European Research Council (ERC Consolidator Grant No 101001791), and the Federal Ministry of Education and Research (BMBF, STOP-FSGS-01GM2202C).

Author contributions statement

M.J., M.S. and V.P. developed PILOT. R.B., C.K., D.H., M.G., J.N., N. B. and I.C. performed analysis and interpretation of single cell and pathomics data. M. J., R.K., I.C. conceived the experiments. M.J., M.S. and I.C. wrote the manuscript. All authors reviewed the manuscript.

Data and code availability

PILOT code including documentation, tutorials and scripts for replicating experiments are found in https://github.com/CostaLab/PILOT.

Group information

Validation Study of the Oxford Classification of IgAN (VALIGA) investigators:

M.L. Russo (MA, PhD, Fondazione Ricerca Molinette, Torino, Italy); S. Troyanov (MD, Division of Nephrology, Department of Medicine, Hopital du Sacre-Coeur de Montreal, Montreal, Quebec, Canada); H.T. Cook (MD, Centre for Complement and Inflammation Research, Department of Medicine, Imperial College, London, England); I. Roberts (MD, Department of Cellular Pathology, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, United Kingdom); V. Tesar, (MD, Department of Nephrology, 1st Faculty of Medicine and General University Hospital, Charles University, Prague, Czech Republic); D. Maixnerova (MD, Department of Nephrology, 1st Faculty of Medicine and General University Hospital, Charles University, Prague, Czech Republic); S. Lundberg (MD, Nephrology Unit, Department of Clinical Sciences, Karolinska Institute, Stockholm, Sweden); L. Gesualdo (MD, Department of Nephrology, Emergency and Organ Transplantation, University of Bari "Aldo Moro," Foggia-Bari, Italy); F. Emma (MD, Division of Nephrology, Department of Pediatric Subspecialties, Bambino Gesù Children's Hospital IRCCS, Rome, Italy); F. Diomedi (MD, Division of Nephrology, Department of Pediatric Subspecialties, Bambino Gesù Children's Hospital IRCCS, Rome, Italy); G. Beltrame (MD, Nephrology and Dialysis Unit, San Giovanni Bosco Hospital, and University of Turin,Turin, Italy); C. Rollino (MD, Nephrology and Dialysis Unit, San Giovanni Bosco Hospital, and University of Turin,Turin, Italy); A. Amore (MD, Nephrology Unit, Regina Margherita Children's Hospital,Turin, Italy); R. Camilla (MD Nephrology Unit, Regina Margherita Children's Hospital, Turin, Italy); L. Peruzzi (MD, Nephrology Unit, Regina Margherita Children's Hospital, Turin, Italy); M. Praga (MD, Nephrology Unit, Hospital 12 de Octubre,Madrid, Spain); S. Feriozzi (MD, Nephrology Unit, Belcolle Hospital, Viterbo, Italy), R. Polci, (MD, Nephrology Unit, Belcolle Hospital,Viterbo, Italy); G. Segoloni, (MD, Division of Nephrology Dialysis and Transplantation, Department of Medical Sciences, Città della Salute e della Scienza Hospital and University of Turin, Turin, Italy); L.Colla (MD, Division of Nephrology Dialysis and Transplantation, Department of Medical Sciences, Città della Salute e della Scienza Hospital and University of Turin,Turin, Italy); A. Pani (MD, Nephrology Unit, G. Brotzu Hospital, Cagliari, Italy); D. Piras (MD, Nephrology Unit, G. Brotzu Hospital, Cagliari, Italy), A. Angioi (MD, Nephrology Unit, G. Brotzu Hospital, Cagliari, Italy); G. Cancarini, (MD, Nephrology Unit, Spedali Civili University Hospital, Brescia, Italy); S. Ravera (MD, Nephrology Unit, Spedali Civili University Hospital, Brescia, Italy); M. Durlik (MD, Department of Transplantation Medicine, Nephrology, and Internal Medicine, Medical University ofWarsaw,Warsaw, Poland); E. Moggia (Nephrology Unit, Santa Croce Hospital, Cuneo, Italy); J. Ballarin (MD, Department of Nephrology, Fundacion Puigvert, Barcelona, Spain); S. Di Giulio (MD, Nephrology Unit, San Camillo Forlanini Hospital, Rome, Italy); F. Pugliese (MD, Department of Nephrology, Policlinico Umberto I University Hospital, Rome, Italy); I. Serriello (MD, Department of Nephrology, Policlinico Umberto I University Hospital, Rome, Italy); Y. Caliskan (MD, Division of Nephrology, Department of Internal Medicine, Istanbul Faculty of Medicine, Istanbul University, Istanbul, Turkey); M. Sever (MD, Division of Nephrology, Department of Internal Medicine, Istanbul Faculty of Medicine, Istanbul University, Istanbul, Turkey); I. Kilicaslan (MD, Department of Pathology, Istanbul Faculty of Medicine, Istanbul University, Istanbul, Turkey); F. Locatelli (MD, Department of Nephrology and Dialysis, Alessandro Manzoni Hospital, ASST Lecco, Italy); L. Del Vecchio (MD, Department of Nephrology and Dialysis, Alessandro Manzoni Hospital, ASST Lecco, Italy); J.F.M.Wetzels (MD, Departments of Nephrology, Radboud University Medical Center, Nijmegen,

the Netherlands); H. Peters (MD, Departments of Nephrology, Radboud University Medical Center, Nijmegen, the Netherlands); U. Berg (MD, Division of Pediatrics, Department of Clinical Science, Intervention and Technology, Huddinge, Sweden); F. Carvalho (MD, Nephrology Unit, Hospital de Curry Cabral, Lisbon, Portugal); A.C. da Costa Ferreira (MD, Nephrology Unit, Hospital de Curry Cabral, Lisbon, Portugal); M. Maggio (MD, Nephrology Unit, Hospital Maggiore di Lodi, Lodi, Italy); A. Wiecek (MD, Department Nephrology, Endocrinology and Metabolic Diseases, Silesian University of Medicine, Katowice, Poland); M. Ots-Rosenberg(MD, Nephrology Unit, Tartu University Clinics, Tartu,Estonia); R. Magistroni (MD, Department of Nephrology, Policlinic of Modena and Reggio Emilia; Modena, Italy); R. Topaloglu (MD, Department of Pediatric Nephrology and Rheumatology, Hacettepe University, Ankara, Turkey); Y. Bilginer (MD, Department of Pediatric Nephrology and Rheumatology, Hacettepe University, Ankara, Turkey); M. D'Amico (MD, Nephrology Unit, S. Anna Hospital, Como, Italy); M. Stangou (MD, Department of Nephrology, Hippokration General Hospital, Aristotle University of Thessaloniki, Thessaloniki, Greece); F. Giacchino (MD, Nephrology Unit, Ivrea Hospital, Ivrea, Italy); D. Goumenos (MD Department of Nephrology, University Hospital of Patras, Patras, Greece); P. Kalliakmani (MD Department of Nephrology, University Hospital of Patras, Patras, Greece); M. Papasotiriou (MD Department of Nephrology, University Hospital of Patras, Patras, Greece); K. Galesic (MD, Department of Nephrology, University Hospital Dubrava, Zagreb, Croatia); C. Geddes (MD, Renal Unit,Western Infirmary Glasgow, Glasgow, United Kingdom); K. Siamopoulos (MD, Nephrology Unit, Medical School University of Ioanina, Ioannina, Greece); O. Balafa (MD, Nephrology Unit,Medical School University of Ioanina, Ioannina, Greece); M. Galliani (MD, Nephrology Unit, S.Pertini Hospital, Rome, Italy); P. Stratta (MD, Department of Nephrology, Maggiore della Carità Hospital, Piemonte Orientale University, Novara, Italy); M. Quaglia (MD, Department of Nephrology, Maggiore della Carità Hospital, Piemonte Orientale University, Novara, Italy); R. Bergia (MD, Nephrology Unit, Degli Infermi Hospital, Biella,Italy); R. Cravero (MD, Nephrology Unit, Degli Infermi Hospital, Biella, Italy); M. Salvadori, (MD, Department of Nephrology, Careggi Hospital, Florence, Italy); L. Cirami (MD, Department of Nephrology, Careggi Hospital, Florence, Italy); B. Fellstrom (MD, Renal Department, University of Uppsala, Uppsala, Sweden); H. Kloster Smerud (MD, Renal Department, University of Uppsala, Uppsala, Sweden); F. Ferrario (MD, Nephropathology Unit, San Gerardo Hospital, Monza, Italy); T. Stellato (MD, Nephropathology Unit, San Gerardo Hospital, Monza, Italy); J. Egido (MD, Department of Nephrology, Fundacion Jimenez Diaz, Madrid, Spain); C. Martin (MD, Department of Nephrology, Fundacion Jimenez Diaz, Madrid, Spain); J. Floege (MD, Nephrology and Immunology, Medizinische Klinik II, University of Aachen, Aachen, Germany); F. Eitner (MD, Nephrology and Immunology, Medizinische Klinik II, University of Aachen, Aachen, Germany); A. Lupo (MD, Department of Nephrology, University of Verona, Verona, Italy); P. Bernich (MD, Department of Nephrology, University of Verona, Verona, Italy); P. Menè (Department of Nephrology,S. Andrea Hospital, Rome, Italy); M. Morosetti (Nephrology Unit, Grassi Hospital, Ostia, Italy); C. van Kooten, (MD, Department of Nephrology, Leiden University Medical Centre, Leiden, The Netherlands); T. Rabelink (MD, Department of Nephrology, Leiden University Medical Centre, Leiden, The Netherlands); M.E.J. Reinders (MD, Department of Nephrology, Leiden University Medical Centre, Leiden, The Netherlands); J.M. Boria Grinyo (Department of Nephrology, Hospital Bellvitge, Barcelona, Spain); S. Cusinato (MD, Nephrology Unit, Borgomanero Hospital, Borgomanero, Italy); L. Benozzi (MD, Nephrology Unit, Borgomanero Hospital, Borgomanero, Italy); S. Savoldi, (MD, Nephrology Unit, Civile Hospital, Ciriè, Italy); C. Licata (MD, Nephrology Unit, Civile Hospital, Ciriè, Italy); M. Mizerska-Wasiak (MD, Department of Pediatrics, Medical University ofWarsaw,Warsaw, Poland); G. Martina (MD, Nephrology Unit, Chivasso Hospital, Chivasso, Italy); A. Messuerotti (MD, Nephrology Unit, Chivasso Hospital, Chivasso, Italy); A. Dal Canton (MD, Nephrology Unit, S. Matteo Hospital, Pavia, Italy); C. Esposito (MD, Nephrology Unit, Maugeri Foundation, Pavia, Italy); C. Migotto (MD, Nephrology Unit, Maugeri Foundation, Pavia, Italy); G. Triolo MD, Nephrology Unit CTO, Turin, Italy); F.Mariano (MD, Nephrology Unit CTO, Turin, Italy); C. Pozzi (MD, Nephrology Unit, Bassini Hospital, Cinisello Balsamo, Italy); R. Boero (MD, Nephrology Unit, Martini Hospital, Turin, Italy);

VALIGA pathology investigators: S. Bellur (MD, Department of Cellular Pathology, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, United Kingdom); G.Mazzucco (MD, Pathology Department, University of Turin, Turin, Italy); C. Giannakakis (MD, Pathology Department, La Sapienza University, Rome, Italy); E. Honsova (MD, Department of Clinical and Transplant Pathology, Institute for Clinical and Experimental Medicine, Prague, Czech Republic); B. Sundelin (MD Department of Pathology and Cytology, Karolinska University Hospital, Karolinska Institute, Stockholm, Sweden); A.M. Di Palma (Nephrology Unit, Aldo Moro University, Foggia-Bari, Italy); F. Ferrario (MD, Nephropathology Unit, San Gerardo Hospital, Monza, Italy); E. Gutiérrez (MD, Renal, Vascular and Diabetes Research Laboratory, Fundación Instituto de Investigaciones Sanitarias-Fundación Jiménez Díaz, Universidad Autónoma de Madrid, Madrid, Spain); A.M. Asunis (MD, Department of Pathology, Brotzu Hospital, Cagliari, Italy); J. Barratt (MD, The JohnWalls Renal Unit, Leicester General Hospital, Leicester, United Kingdom); R. Tardanico (MD, Department of Pathology, Spedali Civili Hospital, University of Brescia, Brescia, Italy); A. Perkowska-Ptasinska (MD, Department of Transplantation Medicine, Nephrology and Internal Medicine, Medical University of Warsaw, Warsaw, Poland); J. Arce Terroba (MD, Pathology Department, Fundació Puigvert, Barcelona, Spain); M. Fortunato (MD, Pathology Department, S. Croce Hospital, Cuneo, Italy); A. Pantzaki (MD, Department of Pathology, Hippokration Hospital, Thessaloniki, Greece); Y. Ozluk (MD, Department of Pathology, Istanbul University,

Istanbul Faculty of Medicine, Istanbul, Turkey); E. Steenbergen (MD, Radboud University Medical Center, Department of Pathology, Nijmegen, The Netherlands); M. Soderberg (MD, Department of Pathology, Drug Safety and Metabolism, Huddinge, Sweden); Z. Riispere (MD, Department of Pathology, University of Tartu, Tartu, Estonia); L. Furci (MD, Pathology Department, University of Modena, Italy); D. Orhan (MD, Department of Pediatrics, Division of Rheumatology, Hacettepe University Faculty of Medicine, Ankara, Turkey); D. Kipgen (MD,Pathology Department, Queen Elizabeth University Hospital, Glasgow, United Kingdom); D. Casartelli (Pathology Department, Manzoni Hospital, Lecco, Italy); D. Galesic Ljubanovic (MD, Nephrology Department, University Hospital, Zagreb, Croatia; Zagreb, Croatia); H Gakiopoulou (MD, Department of Pathology, National and Kapodistrian University of Athens, Athens, Greece); E. Bertoni (MD, Nephrology Department, Careggi Hospital, Florence, Italy); P. Cannata Ortiz (MD, Pathology Department, IIS-Fundacion Jimenez Diaz UAM, Madrid, Spain); H. KarkoszkaMD, (Nephrology, Endocrinology and Metabolic Diseases, Medical University of Silesia, Katowice, Katowice, Poland); H.J. Groene (MD, Cellular and Molecular Pathology, German Cancer Research Center, Heidelberg, Germany); A. Stoppacciaro (MD, Surgical Pathology Units, Department of Clinical and Molecular Medicine, Ospedale Sant'Andrea, Sapienza University of Rome, Rome, Italy); I. Bajema (MD, Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands); J. Bruijn (MD, Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands); X. Fulladosa Oliveras (MD, Nephrology Unit, Bellvitge University Hospital, Hospitalet de Llobregat, Barcelona, Spain); J. Maldyk (MD, Division of Pathomorphology, Children's Clinical Hospital, Medical University ofWarsaw,Warsaw, Poland); and E. Ioachim (MD, Department of Pathology, Medical School, University of Ioannina, Ioannina, Greece); the Oxford derivation and North American validation investigators: Bavbek N (MD, Department of Pathology, Vanderbilt University, Nashville, Tennessee); Cook T (MD, Imperial College, London, England), Troyanov S (MD, Division of Nephrology, Department of Medicine, Hopital du Sacre-Coeur de Montreal, Montreal, Quebec, Canada); Alpers C (MD, Department of Pathology, University of Washington Medical Center, Seattle,Washington), Amore A (MD, Nephrology, Dialysis and Transplantation Unit, Regina Margherita Children's Hospital, University of Turin, Turin, Italy), Barratt J (MD, The John Walls Renal Unit, Leicester General Hospital, Leicester, England); Berthoux F (MD, Department of Nephrology, Dialysis, and Renal Transplantation, Hôpital Nord, CHU de Saint-Etienne, Saint-Etienne, France); Bonsib S (MD, Department of Pathology, LSU Health Sciences Center, Shreveport, Los Angeles); Bruijn J (MD, Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands); D'Agati V (MD, Department of Pathology, Columbia University College of Physicians Surgeons, New York, New York); D'Amico G (MD, Fondazione D'Amico per la Ricerca sulle Malattie Renali, Milan, Italy); Emancipator S (MD, Department of Pathology, Case Western Reserve University, Cleveland, Ohio); Emmal F (MD, Division of Nephrology and Dialysis, Department of Nephrology and Urology, Bambino Gesù Children's Hospital and Research Institute, Piazza S Onofrio, Rome, Italy); Ferrario F (MD, Renal Immunopathology Center, San Carlo Borromeo Hospital, Milan, Italy); Fervenza F (MD PhD, Division of Nephrology and Hypertension, Mayo Clinic, Rochester); Florquin S (MD, Department of Pathology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands); Fogo A (MD, Department of Pathology, Vanderbilt University, Nashville, Tennessee); Geddes C (MD, The Renal Unit, Western Infirmary, Glasgow, Scotland); Groene H (MD, Department of Cellular and Molecular Pathology, German Cancer Research Center, Heidelberg, Germany); Haas M(MD, Department of Pathology and Laboratory Medicine, Cedars-Sinai Medical Center, Los Angeles, California); Hill P (MD, St Vincent's Hospital, Melbourne, Australia); Hogg R (MD, Scott and White Medical Center, Temple, Texas (retired)); Hsu S (MD, Division of Nephrology, Hypertension and Renal Transplantation, College of Medicine, University of Florida, Gainesville, Florida); Hunley T (MD, Department of Pathology, Vanderbilt University, Nashville, Tennessee); Hladunewich (MD, Division of Nephrology, Sunnybrook Health Science Center, University of Toronto, Ontario, Canada M); Jennette C (MD, Department of Pathology and Laboratory Medicine, University of North Carolina, Chapel Hill, North Carolina); Joh K (MD, Division of Immunopathology, Clinical Research Center Chiba, East National Hospital, Chiba, Japan); Julian B (MD, Department of Medicine, University of Alabama at Birmingham, Birmingham, Alabama); Kawamura T (MD, Division of Nephrology and Hypertension, Jikei University School of Medicine, Tokyo, Japan); Lai F (MD, The Chinese University of Hong Kong, Hong Kong); Leung C (MD, Department of Medicine, Prince of Wales Hospital, Chinese University of Hong Kong, Hong Kong); Li L (MD, Research Institute of Nephrology, Jinling Hospital, Nanjing University School of Medicine, Nanjing, China); Li P (MD, Department of Medicine, Prince of Wales Hospital, Chinese University of Hong Kong, Hong Kong); Liu Z (MD, Research Institute of Nephrology, Jinling Hospital, Nanjing University School of Medicine, Nanjing, China); Massat A (MD, Division of Nephrology and Hypertension, Mayo Clinic, Rochester, Minnesota); Mackinnon B (MD, The Renal Unit,Western Infirmary, Glasgow, Scotland); Mezzano S (MD, Departamento de Nefrología, Escuela de Medicina, Universidad Austral, Valdivia, Chile); Schena F (MD, Renal, Dialysis and Transplant Unit, Policlinico, Bari, Italy); Tomino Y (MD, Division of Nephrology, Department of Internal Medicine, Juntendo University School of Medicine, Tokyo, Japan); Walker P (MD, Nephropathology Associates, Little Rock, Arkansas);Wang H (MD, Renal Division of Peking University First Hospital, Peking University Institute of Nephrology, Beijing, China (deceased)); Weening J (MD, Erasmus Medical Center, Rotterdam, The Netherlands); and Yoshikawa N (MD, Department of Pediatrics, Wakayama Medical University, Wakayama City, Japan); the International investigators: Cai-Hong Zeng (MD, Nanjing University School of Medicine, Nanjing, China); Sufang Shi (MD,

Peking University Institute of Nephrology, Beijing, China); C.Nogi (MD, Juntendo University, Faculty of Medicine, Tokyo, Japan); H.Suzuki (MD, Juntendo University, Faculty of Medicine, Tokyo, Japan); K. Koike (MD, Division of Nephrology and Hypertension, Department of Internal Medicine, Jikei University School of Medicine, Tokyo, Japan); K. Hirano (MD, Division of Nephrology and Hypertension, Department of Internal Medicine, Jikei University School of Medicine, Tokyo, Japan); T. Kawamura (MD, Division of Nephrology and Hypertension, Department of Internal Medicine, Jikei University School of Medicine, Tokyo, Japan); T. Yokoo (MD, Division of Nephrology and Hypertension, Department of Internal Medicine, Jikei UniversitySchool of Medicine, Tokyo, Japan); M. Hanai (MD, Division of Nephrology, Department of Medicine, Kurume University School of Medicine, Fukuoka, Japan); K. Fukami (MD, Division of Nephrology, Department of Medicine, Kurume University School of Medicine, Fukuoka,, Japan); K. Takahashi (MD, Department of Nephrology, Fujita Health University School of Medicne, Aichi, Japan); Y. Yuzawa (MD, Department of Nephrology, Fujita Health University School of Medicine, Aichi, Japan); M. Niwa (MD, Department of Nephrology, Nagoya University Graduate School of Medicine, Aichi, Japan); Y. Yasuda (MD, Department of Nephrology, Nagoya University Graduate School of Medicine, Aichi, Japan); S. Maruyama (MD, Department of Nephrology, Nagoya University Graduate School of Medicine, Aichi, Japan); D. Ichikawa (MD, Division of Nephrology and Hypertension, Department of Internal Medicine, St Marianna University School of Medicine, Kanagawa, Japan); T. Suzuki (MD, Division of Nephrology and Hypertension, Department of Internal Medicine, St Marianna University School of Medicine, Kanagawa, Japan); S. Shirai (MD, Division of Nephrology and Hypertension, Department of Internal Medicine, St Marianna University School of Medicine, Kanagawa, Japan); A. Fukuda (MD, First Department of Internal Medicine, Faculty of Medicine, University of Miyazaki, Miyazaki, Japan); S. Fujimoto (MD, Department of Hemovascular Medicine and Artificial Organs, Faculty of Medicine, University of Miyazaki, Miyazaki, Japan); H. Trimarchi (MD, Division of Nephrology, Hospital Britanico, Buenos Aires, Argentina).

References

1. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, eaba7721 (2020).
2. Consortium*, T. S. *et al.* The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
3. Sklavenitis-Pistofidis, R., Getz, G. & Ghobrial, I. Single-cell rna sequencing: one step closer to the clinic. *Nat. Medicine* **27**, 375–376 (2021).
4. Kuppe, C. *et al.* Spatial multi-omic map of human myocardial infarction. *Nature* 1–12 (2022).
5. Stephenson, E. *et al.* Single-cell multi-omics analysis of the immune response in covid-19. *Nat. medicine* **27**, 904–916 (2021).
6. Peng, J. *et al.* Single-cell rna-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell research* **29**, 725–738 (2019).
7. Boor, P. *et al.* Next-generation morphometry for pathomics-data mining in histopathology, DOI: 10.21203/rs.3.rs-1609168/v1 (2022).
8. Chen, W. S. *et al.* Uncovering axes of variation among single-cell cancer specimens. *Nat. Methods* **17**, 302–310, DOI: 10.1038/s41592-019-0689-z (2020).
9. Ravindra, N., Sehanobish, A., Pappalardo, J. L., Hafler, D. A. & van Dijk, D. Disease state prediction from single-cell data using graph attention networks. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, 121–130, DOI: 10.1145/3368555.3384449 (ACM, New York, NY, USA, 2020). 2002.07128.
10. Lublin, F. D., Reingold, S. C. *et al.* Defining the clinical course of multiple sclerosis: results of an international survey. *Neurology* **46**, 907–911 (1996).
11. Zhang, L. *et al.* SCANCell reveals diverse inter-cluster interaction patterns in systemic lupus erythematosus across the disease spectrum. *Bioinformatics* **38**, 1361–1368, DOI: 10.1093/bioinformatics/btab713 (2022).
12. Kuchroo, M. *et al.* Multiscale phate identifies multimodal signatures of covid-19. *Nat. Biotechnol.* **40**, 681–691 (2022).
13. Peyré, G. & Cuturi, M. Computational optimal transport. *Foundations Trends Mach. Learn.* **11**, 1–257, DOI: 10.1561/2200000073 (2019). 1803.00567.
14. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. statistical mechanics: theory experiment* **2008**, P10008 (2008).
15. Coifman, R. R. & Lafon, S. Diffusion maps. *Appl. computational harmonic analysis* **21**, 5–30 (2006).
16. Huber, P. J. A robust version of the probability ratio test. *The Annals Math. Stat.* 1753–1758 (1965).

17. Perez, R. K. *et al.* Single-cell rna-seq reveals cell type–specific molecular and genetic associations to lupus. *Science* **376**, eabf1970 (2022).

18. Coppo, R. *et al.* Validation of the oxford classification of iga nephropathy in cohorts with different presentations and treatments. *Kidney international* **86**, 828–836 (2014).

19. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. computational applied mathematics* **20**, 53–65 (1987).

20. Traag, V. A., Waltman, L. & Van Eck, N. J. From louvain to leiden: guaranteeing well-connected communities. *Sci. reports* **9**, 1–12 (2019).

21. Rand, W. M. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. association* **66**, 846–850 (1971).

22. Chang, S. *et al.* Histone deacetylases 5 and 9 govern responsiveness of the heart to a subset of stress signals and play redundant roles in heart development. *Mol. cellular biology* **24**, 8467–76, DOI: 10.1128/MCB.24.19.8467-8476.2004 (2004).

23. Chen, X., Wan, J., Yu, B., Diao, Y. & Zhang, W. PIP5K1α promotes myogenic differentiation via AKT activation and calcium release. *Stem cell research therapy* **9**, 33, DOI: 10.1186/s13287-018-0770-z (2018).

24. Marx, V. How single-cell multi-omics builds relationships. *Nat. methods* **19**, 142–146, DOI: 10.1038/s41592-022-01392-8 (2022).

25. Bonneel, N., Van De Panne, M., Paris, S. & Heidrich, W. Displacement interpolation using lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*, 1–12 (2011).

26. Rubner, Y., Tomasi, C. & Guibas, L. J. The earth mover's distance as a metric for image retrieval. *Int. journal computer vision* **40**, 99–121 (2000).

27. Huber, P. J. Robust estimation of a location parameter. In *Breakthroughs in statistics*, 492–518 (Springer, 1992).

28. Ren, X. *et al.* Covid-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* **184**, 1895–1913 (2021).

29. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).

30. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. biotechnology* **32**, 381–386 (2014).

Supplementary Tables and Figures

Table S1. Benchmarking of clustering results based on ARI and Silhouette Scores. Results in bold indicate the best value per method.

Data sets	**PDAC**		**Myoc. Infarc.**		**COVID-19**		**Lupus PBMC**		**Kidney AC T**		**Kidney AC G**		**Kidney IgAN T**		**Kidney IgAN G**	
Methods/Metrics	*ARI*	*Silh*	ARI	*Silh*	*ARI*	*Silh*	*ARI*	*Silh*	ARI	Silh	*ARI*	Silh	*ARI*	Silh	*ARI*	Silh
PILOT	**1.00**	**0.59**	**0.90**	0.73	**0.56**	-0.13	**0.58**	**0.24**	**0.68**	**0.30**	**0.64**	**0.36**	**0.53**	**0.10**	**0.55**	**0.07**
PhEMD	0.51	0.04	0.47	-0.06	0.51	**-0.06**	0.51	0.02	0.51	0.00	0.51	-0.01	0.49	-0.03	0.50	-0.04
Pseudo Bulk	0.67	0.37	0.52	**0.83**	0.52	-0.26	0.55	0.22	0.57	0.21	0.51	0.25	0.48	0.07	0.50	0.01

Table S2. Benchmarking of disease progression results based on area under the PR curve (AUPR). Results in bold indicate the best value per method.

Data sets	**PDAC**	**Myoc. Infarc.**	**COVID-19**	**Lupus PBMC**	**Kidney AC G**	**Kidney AC T**	**Kidney IgAN T**	**Kidney IgAN G**
PILOT	**1.00**	**0.98**	**0.94**	**0.88**	**0.96**	**0.93**	**0.60**	**0.60**
PhEMD	0.87	0.27	0.88	0.57	0.69	0.76	0.48	0.35

Table S3. Log likelihood test and hazard ratios (HR) of cox regression for disease progression and morphometric variables

Features	**p_value**	**HR**	**HR_lower_95**	**HR_higher_95**
Disease_Progression	2.38E-11	14.22285964	6.133761376	32.97972059
Tuft-Area-Fraction	6.55E-06	0.006096	0.000754147	0.049275825
Age	0.000161201	1.026118273	1.012515212	1.03990409
Tuft Elongation	0.035174436	0.995855111	0.991024645	1.000709122
Tuft Area	0.059838632	0.99999991	0.999999796	1.000000024
Glomerular Area	0.084688986	0.99999995	0.999999883	1.000000017
Tuft Circularity	0.120647357	0.997847514	0.99474223	1.000962492
Tuft Eccentricity	0.193320223	0.99911608	0.99765839	1.0005759
Bowman Area	0.285387443	0.999999915	0.999999741	1.000000088
Glomerular Circularity	0.292065233	0.998989099	0.996929139	1.001053316
Glomerular Diameter	0.295112885	0.999996112	0.999988267	1.000003957
Tubular Distance	0.334379046	0.999806135	0.999357973	1.000254498
Tubular Area	0.361417305	1.000000128	0.999999865	1.000000392
Tuft Solidity	0.468203891	0.999443691	0.997818499	1.001071531
Glomerular Distance	0.55738282	1.000000526	0.999998905	1.000002148
Glomerular Solidity	0.594992091	0.999714882	0.998620471	1.000810492
Glomerular Eccentricity	0.721397871	1.000225975	0.999008779	1.001444655
Glomerular Elongation	0.944083375	0.999862207	0.995995708	1.003743715
Tubular Diameter	0.954027614	0.999999227	0.999972858	1.000025597

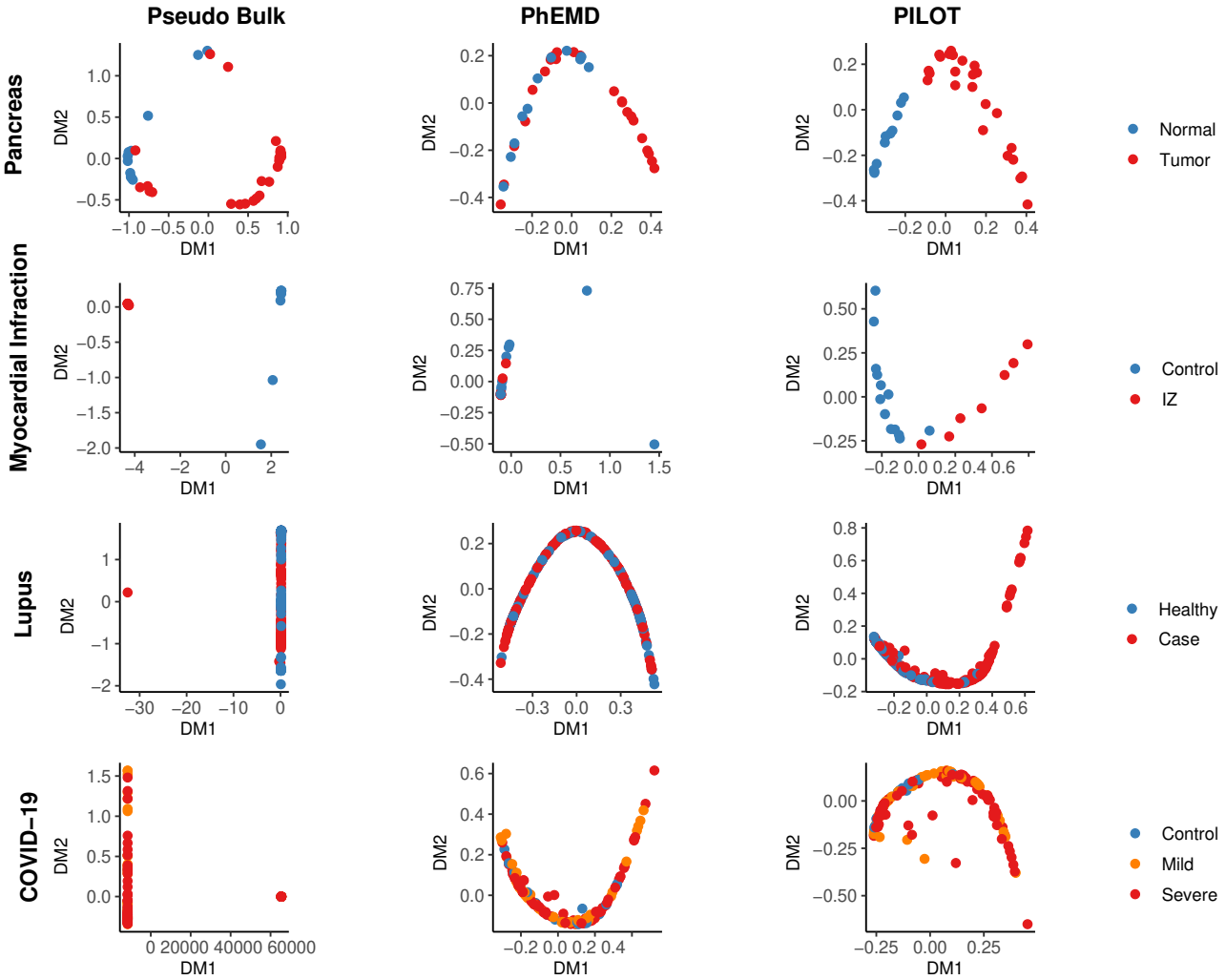


Figure S1. Diffusion maps of single cell data sets (rows) for Pseudo Bulk, PhEMD, and PILOT (columns).

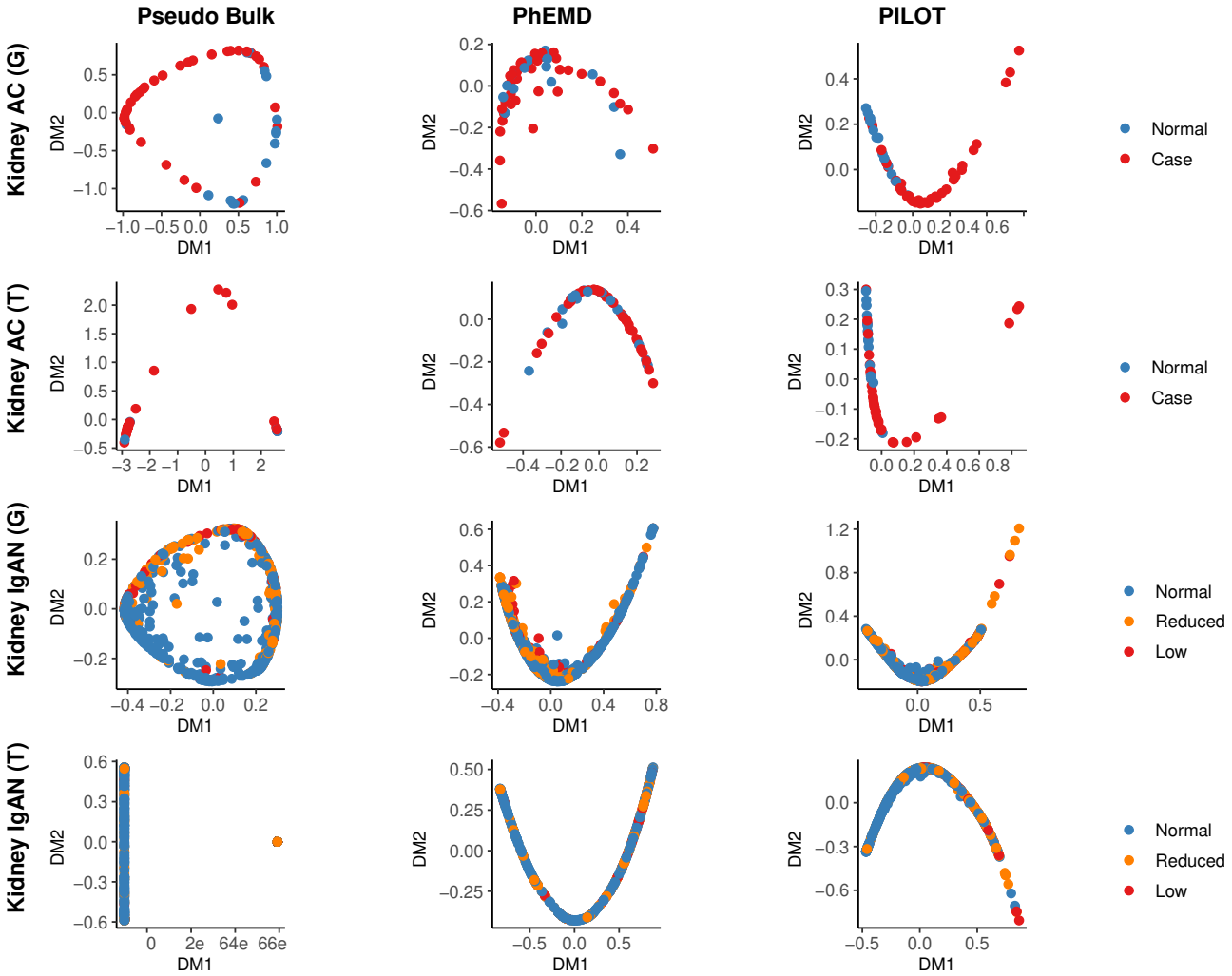


Figure S2. Diffusion maps of pathomics data sets (rows) for Pseudo Bulk, PhEMD, and PILOT (columns).

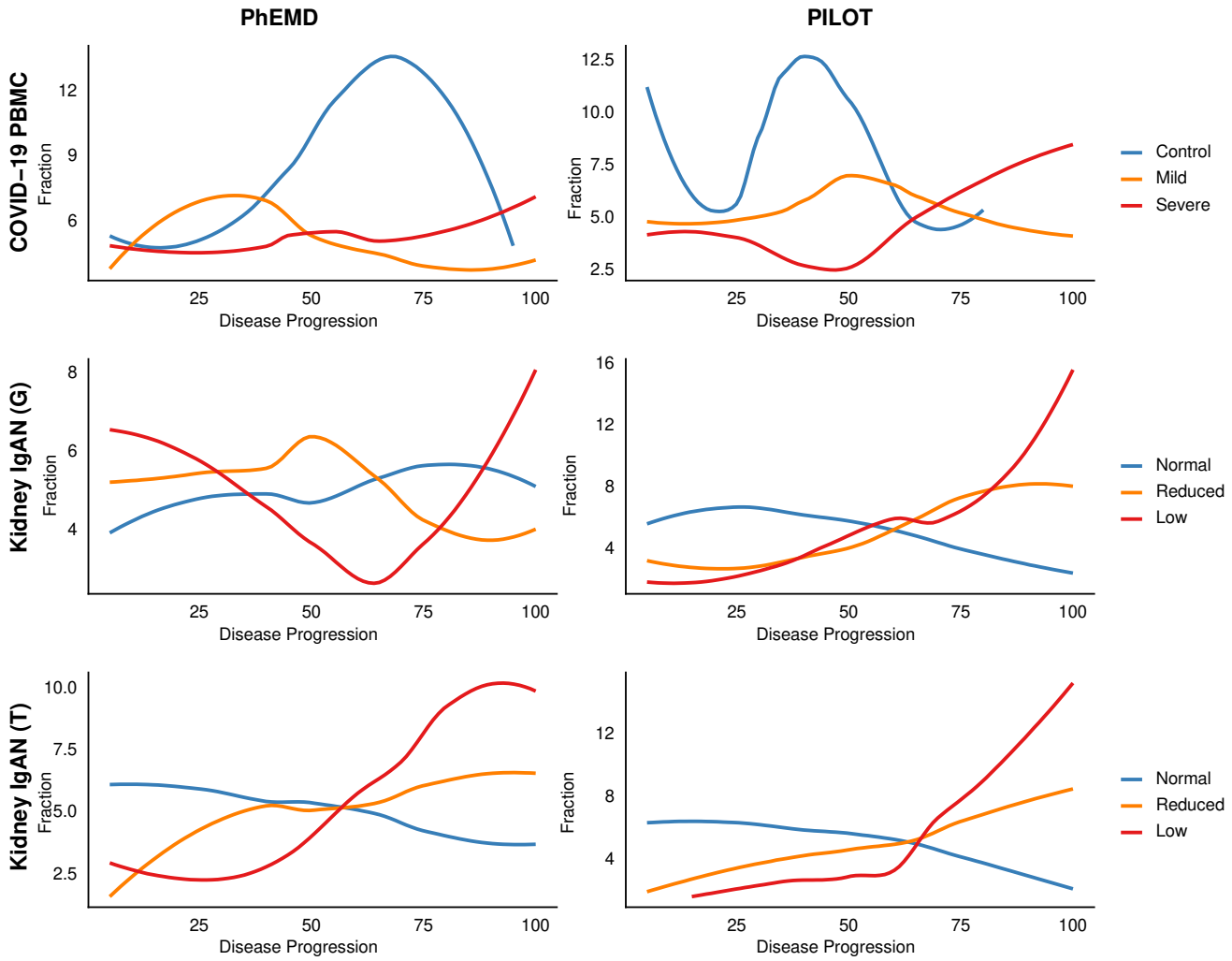


Figure S3. Fraction samples (y-axis) for distinct samples labels over the pseudo-time (x-axis) for disease progression trajectories of PILOT and PhEMD. Trajectories are ordered from normal to case/disease. No disease progression could be estimated from the pseudo-bulk due to the inability of the diffusion maps to find trajectories.

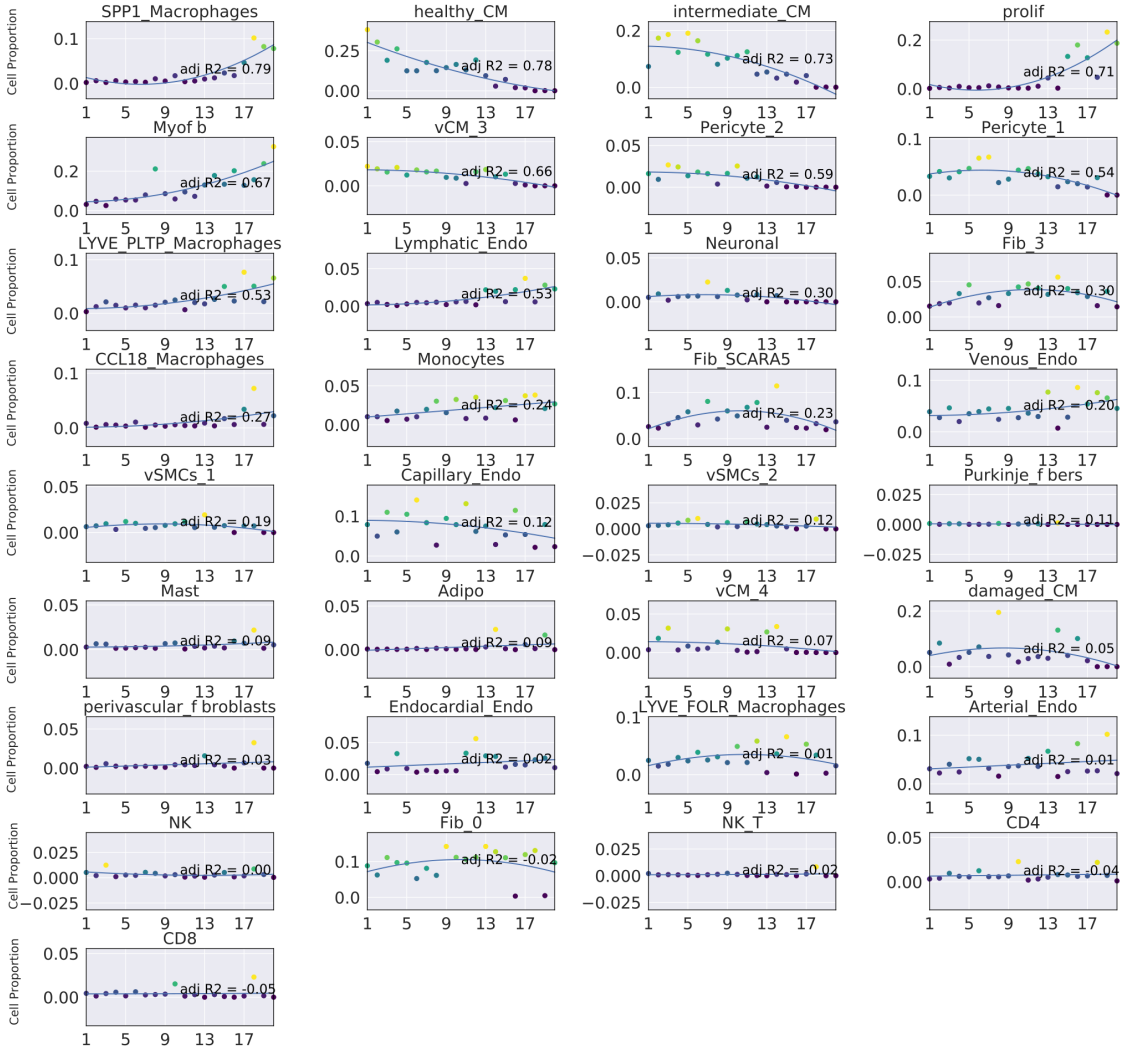


Figure S4. Cell type frequency (y-axis) vs. PILOT disease progression (x-axis) for Myocardial Infarction scRNA-seq data.

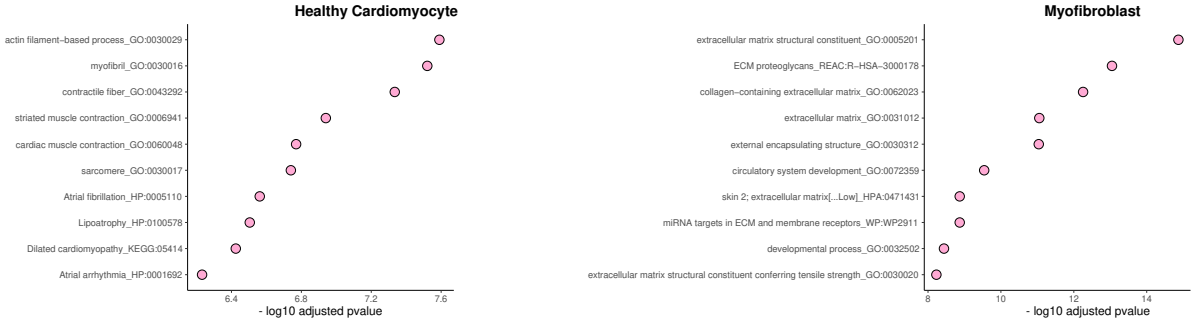


Figure S5. Top 10 terms of the enrichment analysis by using 100 genes with highest adjusted R^2 for Healthy cardiomyocytes and myofibroblasts. The analysis was performed with g:profiler https://biit.cs.ut.ee/gprofiler/gost.

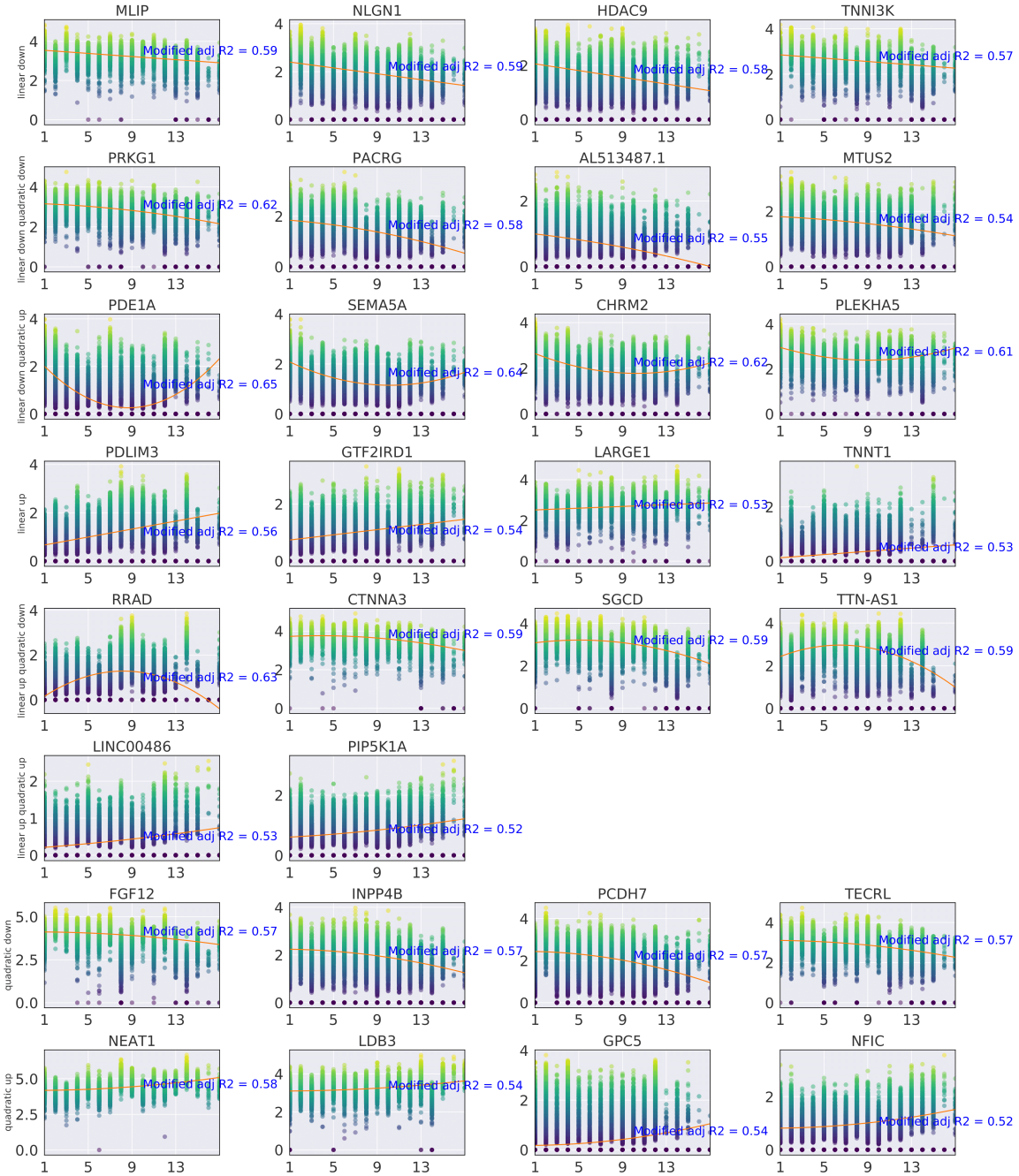


Figure S6. Top genes associated (ranked by adjusted R2) from Healthy cardiomiocytes of the myocardial infarction scRNA-seq. Every line corresponds to distinct models (linear, quadratic, linear-quadratic) and patterns (up vs. down). Only significant genes (p-value > 0.01) are shown.

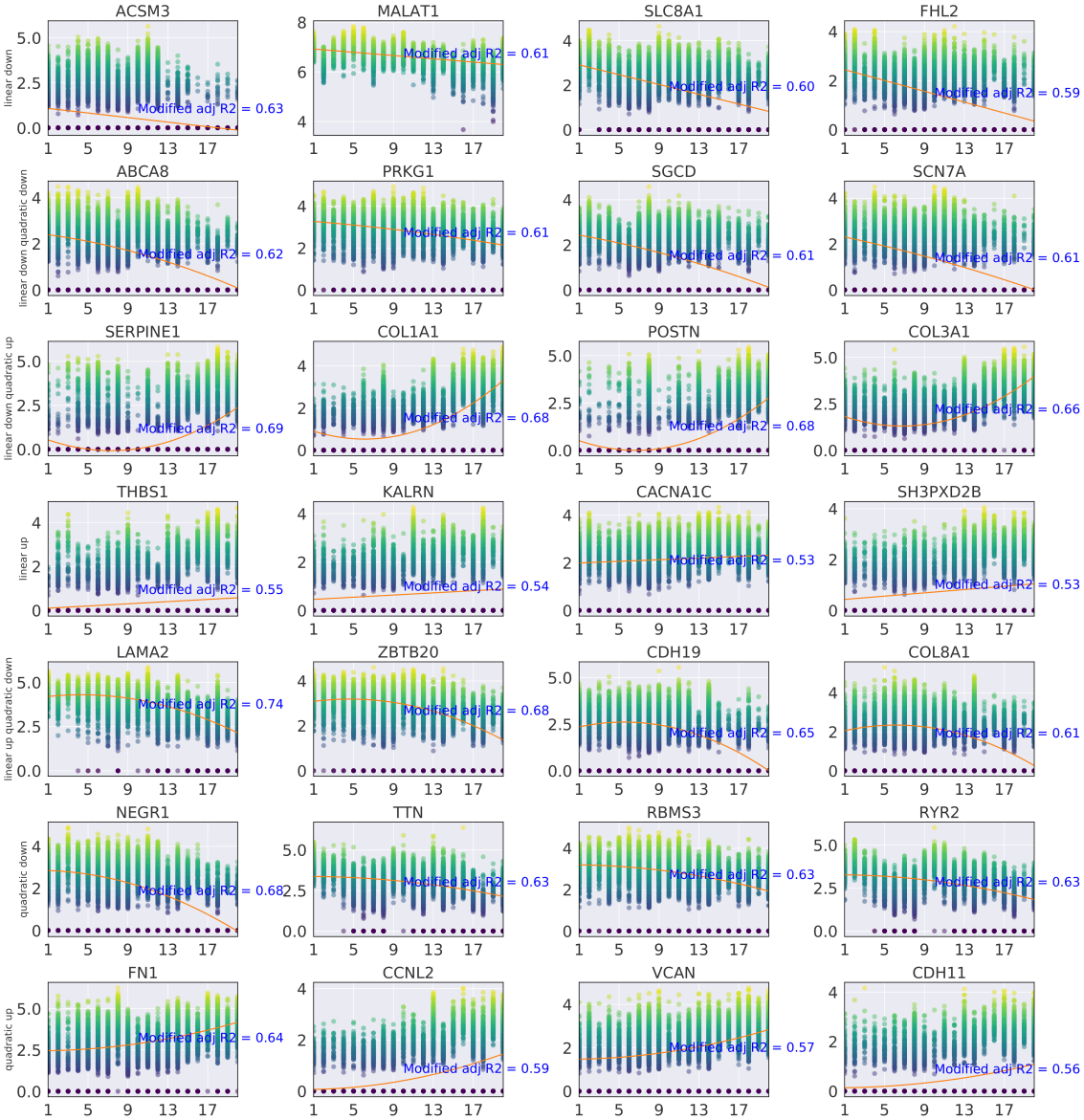


Figure S7. Top genes associated (ranked by adjusted R2) from Myofibroblasts of the myocardial infarction scRNA-seq. Every line corresponds to distinct models (linear, quadratic, linear-quadratic) and patterns (up vs. down). Only significant genes (p-value > 0.01) are shown.

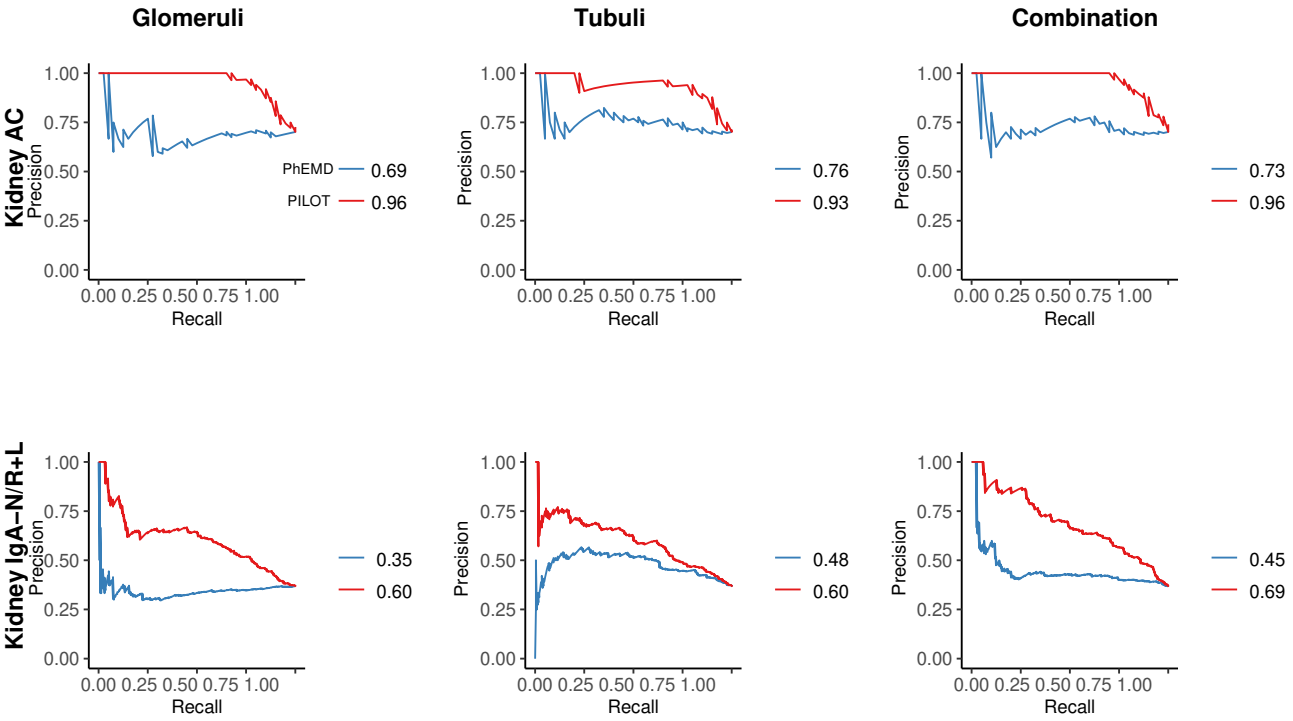


Figure S8. AUCPR plots for Glomeruli, Tubule and both (combined) for Kidney AC, Kidney IgAN pathomics data.

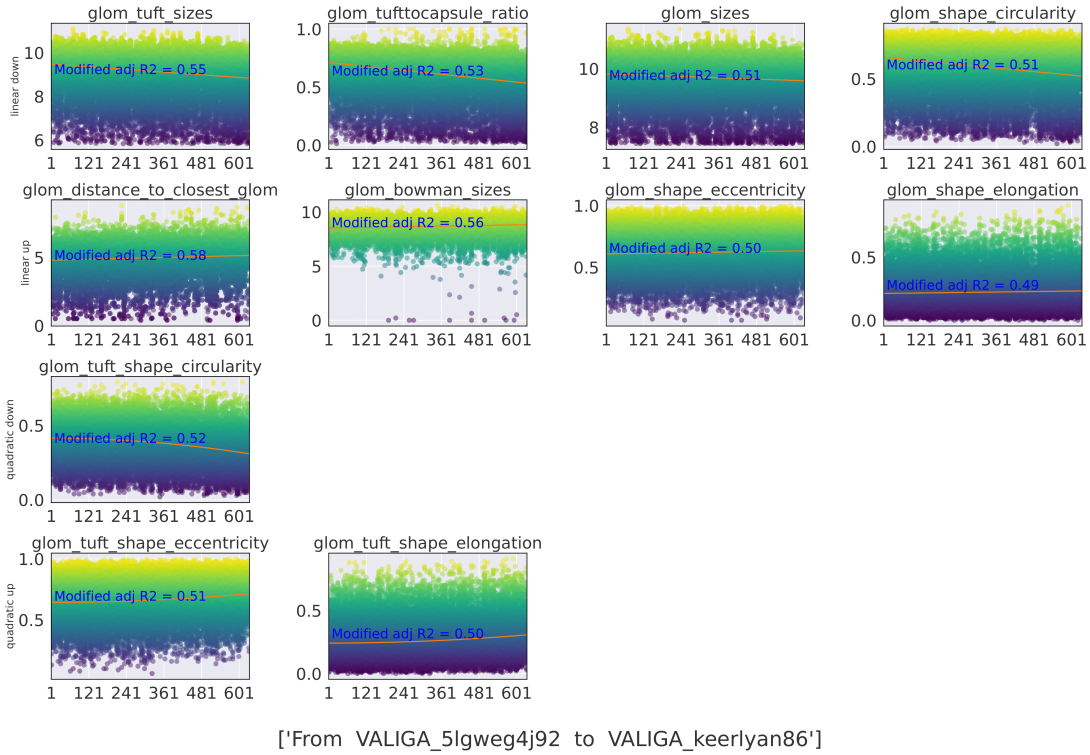


Figure S9. Top morphological features associated (ranked by adjusted R2) from Kidney IgA gromeruli. Every line corresponds to distinct models (linear, quadratic, linear-quadratic) and patterns (up vs. down). Only significant features (p-value > 0.01) are shown.

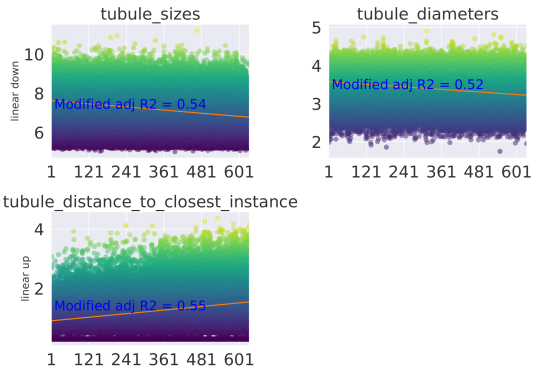


Figure S10. Top morphological features associated (ranked by adjusted R2) from Kidney IgA tubule. Every line corresponds to distinct models (linear, quadratic, linear-quadratic) and patterns (up vs. down). Only significant features (p-value > 0.01) are shown.

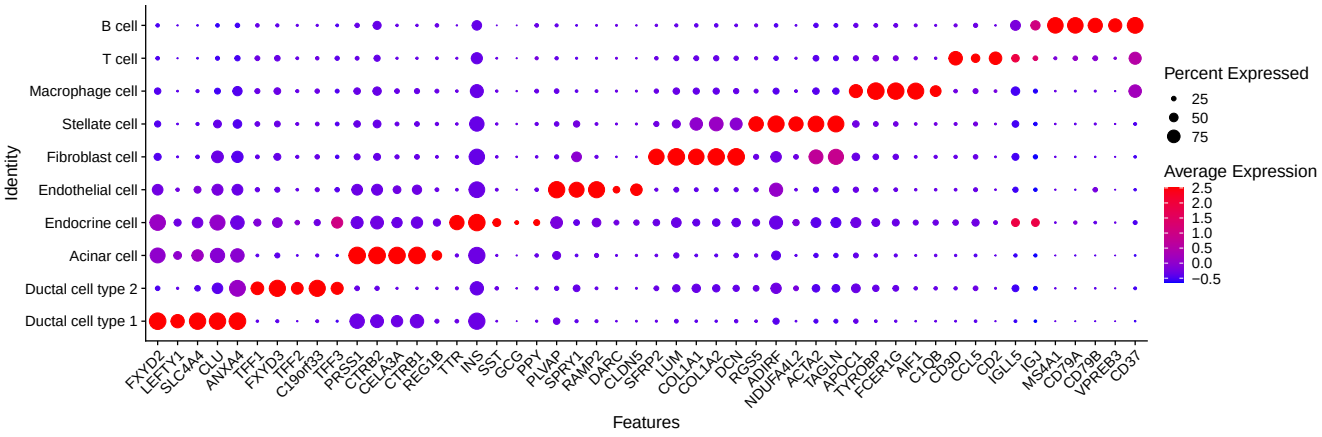


Figure S11. Markers used to characterize clusters in the PDAC single cell data set.