iScience, Volume *24*

Supplemental information

Hyperbolic geometry of gene expression

Yuansheng Zhou and Tatyana O. Sharpee

1. Transparent Methods

Metric and non-metric MDS in Euclidean model

Assume there sectionre n objects described by a set of measurements, the dissimilarity of the objects can be obtained by the experimental measurements of the objects. For example, the dissimilarities of two cells can be calculated by the Euclidean distances of the gene expression vectors. Metric MDS approximates the geometric distances d_{ij} to the data dissimilarities δ_{ij}, while non-metric MDS approximates a monotonic transformation of dissimilarities of data. The transformed values are known as disparities $\hat{d_{ij}}$. The loss function S in Euclidean embedding was defined as :

$$S = \sqrt{\frac{S^*}{T^*}} \tag{S1}$$

Where $S^* = \Sigma_{i,j}(d_{ij} - \hat{d_{ij}})^2$, $T^* = \Sigma_{i,j} d_{ij}^2$. In non-metric MDS, $\hat{d_{ij}}$ is determined using the greatest convex minorant method in Kruskal's approach Kruskal (1964). In metric MDS, disparities are equal to dissimilarities: $\hat{d_{ij}} = \delta_{ij}$.

1.1. Non-metric MDS in native hyperbolic model

There are many hyperbolic space representations, we will use the native representation with polar coordinates Krioukov et al. (2010) in our hyperbolic MDS. The angular coordinates in the space are the same as in an Euclidean ball, the radius $R_{\text{model}} \in (0, \infty)$ characterizes the hierarchical depth of the structure, measures the degree of hierarchy in data, and determines how points distribute in the space. The distance of two points d_{ij} is calculated as:

$$\cosh(d_{ij}) = \cosh(r_i)\cosh(r_j) - \sinh(r_i)\sinh(r_j)\cos(\Delta\theta_{ij}) \tag{S2}$$

Where r_i and r_j are the radial coordinates of the two points, and $\Delta\theta_{ij}$ is the angle between them. In D-dimensional HMDS, we initialize the embedding process by uniformly sampling points within radius R_{model} in the native hyperbolic model. The points directions are uniformly sampled around the high-dimensional sphere, and the radial coordinate $r \in (0, R_{\text{model}}]$ follows :

$$\rho(r) \sim \sinh^{D-1} r \tag{S3}$$

We note that there can be merits to sample the points uniformly in the angular variables. Although this does not lead to uniform sampling of points along the sphere Koay (2011), this way of sampling can be particularly advantageous in the situation where the angular variable maps onto periodic variables that correspond to cell cycle or other rhytms. We have used this sampling in our previous publication on olfactory signals produced by fruits and plants Zhou et al. (2018) where it matched developmental processes in the fruit.

During the iteration process, we update both angular and radial coordinates according to the gradient descent of the loss function Eq. (S1), and at the same time set R_{model} as the upper bound of the radial coordinates. The reason of setting a bound is that the coordinates in hyperbolic model are polar coordinates which cannot be normalized after each iteration as performed in Euclidean MDS, so without bound the gradient descent of loss functionsection Eq. (S1) would lead to very large r_i and d_{ij} (since d_{ij} is in the denominator) and hence fail to preserve radial coordinates of data. By setting the upper bound for radial coordinates, the HMDS embedding can well preserve the data distances and precisely detect hyperbolic radius of data R_{data} (Fig. S3).

1.2. Fitting of Shepard diagram

The Shepard diagram is linear if the geometry of input data matches the geometry of embedding space, and otherwise nonlinear. In both EMDS and HMDS, we use the power function below to fit the pairwise distances:

$$y = a(x - x_0)^{\kappa+1} \tag{S4}$$

Where $x_0 = \min(x) - \epsilon$ is an offset representing the distance caused by intrinsic noise of data, a small value ϵ is introduced to avoid zero inputs in the fitting. The convexity κ describes the linearity of the fitting. $\kappa = 0$ indicates Euclidean input in EMDS and means $R_{\text{data}} = R_{\text{model}}$ in HMDS. $\kappa > 0$ means the data is more hyperbolic than the model, and vice versa.

1.3. Hyperbolic t-SNE

Given a data set containing N data points described by D dimensional vectors: $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_N; \mathbf{x}_i \in \mathbb{R}^D\}$. The t-SNE algorithm Maaten and Hinton (2008) defines the similarity of two points $\mathbf{x}_i, \mathbf{x}_j$ as the joint probability p_{ij}:

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)}, \tag{S5}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}. \tag{S6}$$

The similarity of two points $\mathbf{y}_i, \mathbf{y}_j$ in embedding space is defined as the joint probability q_{ij}:

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{m\neq n}(1 + \|\mathbf{y}_m - \mathbf{y}_n\|^2)^{-1}}. \tag{S7}$$

The discrepancy between the similarities of data and embedding points is the loss function, which is defined by Kullback-Leibler (KL) divergence of the joint probability p_{ij} and q_{ij} :

$$L = D_{KL}(P\|Q) = \sum_i \sum_j p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right). \tag{S8}$$

Minimizing the loss function L with respect to the embedding coordinates $\mathbf{y}_i$ by gradient descent gives:

$$\frac{\partial L}{\partial \mathbf{y}_i} = 4\sum_j (p_{ij} - q_{ij})(\mathbf{y}_i - \mathbf{y}_j)(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}. \tag{S9}$$

The original definitions of similarities Eqs. (S5-S7) in t-SNE are sensitive to small pairwise distances among neighboring points but not to large distances between distant points. To preserve large

distances, Zhou et al. Zhou and Sharpee (2018) proposed global t-SNE algorithm that introduced global similarity terms $\hat{p}_{ij}$ and $\hat{q}_{ij}$ which are primarily sensitive to large distance values:

$$\begin{aligned} \hat{p}_{ij} &= \frac{1 + \|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sum_{m \neq n} (1 + \|\mathbf{x}_m - \mathbf{x}_n\|^2)} \\ \hat{q}_{ij} &= \frac{1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2}{\sum_{m \neq n} (1 + \|\mathbf{y}_m - \mathbf{y}_n\|^2)} \end{aligned} \tag{S10}$$

And they defined the global loss function $\hat{L}$ as:

$$\hat{L} = D_{KL}(\hat{P} \| \hat{Q}) = \sum_i \sum_j \hat{p}_{ij} \log \left(\frac{\hat{p}_{ij}}{\hat{q}_{ij}} \right) \tag{S11}$$

The total loss function L_{total} was then defined by combining the two loss functions using a weight parameter λ:

$$L_{\text{total}} = L + \lambda \hat{L} \tag{S12}$$

The gradient of the total loss function L_{total} gives:

$$\frac{\partial L_{\text{total}}}{\partial \mathbf{y}_i} = 4 \sum_j [(p_{ij} - q_{ij}) - \lambda(\hat{p}_{ij} - \hat{q}_{ij})] \cdot (\mathbf{y}_i - \mathbf{y}_j)(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}, \tag{S13}$$

where the weight λ of the global loss function controls the balance between the local clustering and global organization of the data. Large λ values lead to more robust global distribution of clusters, but less clear classifications. Small λ moves back to approximate the traditional t-SNE, and will be exactly the same when $\lambda = 0$. In hyperbolic t-SNE, we still use native representation parametrized by R_{model} as in HMDS, but here R_{model} is only used to determine the initial radial distribution, not to set the upper bound. We substitute the Euclidean distances in global similarity terms Eq. (S10) by hyperbolic distances d^h_{ij} defined in Eq. (S2), and change Cartesian coordinate system to polar one for all the distance calculations. Then the gradient of total loss function with respect to polar coordinates would be:

$$\begin{aligned} \frac{\partial L_{total}}{\partial r_i} &= 4 \sum_j [(p_{ij} - q_{ij}) \cdot d^e_{ij} \cdot \frac{\partial d^e_{ij}}{\partial r_i} (1 + (d^e_{ij})^2)^{-1} \\ &\quad - \lambda(\hat{p}_{ij} - \hat{q}_{ij}) \cdot d^h_{ij} \cdot \frac{\partial d^h_{ij}}{\partial r_i} (1 + (d^h_{ij})^2)^{-1}] \\ \frac{\partial L_{total}}{\partial \boldsymbol{\theta}_i} &= 4 \sum_j [(p_{ij} - q_{ij}) \cdot d^e_{ij} \cdot \frac{\partial d^e_{ij}}{\partial \boldsymbol{\theta}_i} (1 + (d^e_{ij})^2)^{-1} \\ &\quad - \lambda(\hat{p}_{ij} - \hat{q}_{ij}) \cdot d^h_{ij} \cdot \frac{\partial d^h_{ij}}{\partial \boldsymbol{\theta}_i} (1 + (d^h_{ij})^2)^{-1}] \end{aligned} \tag{S14}$$

Where d_{ij}^e is the Euclidean pairwise distance in polar coordinates, d_{ij}^h is the hyperbolic pairwise distance obtained from Eq. (S2). p_{ij}, q_{ij}, $\hat{p}_{ij}$ and $\hat{q}_{ij}$ are defined by Eqs. (S5-S7) and Eq. (S10) with polar coordinates. When implementing the algorithm, we substitute the radial coordinates with their exponential transformation to avoid negative radii during the iteration:

$$r_{exp} = e^r \tag{S15}$$

The derivative of distances with respect to the new variable would be:

$$\frac{\partial}{\partial r_{exp}} = \frac{\partial}{\partial r} \cdot \frac{\partial r}{\partial r_{exp}} = \frac{1}{r_{exp}} \cdot \frac{\partial}{\partial r} \tag{S16}$$

When the iterations converge, we make logarithm transformation of r_{exp} to get the real radial coordinates.

1.4. Parameters in visualization algorithms

For Lukk et al. Lukk et al. (2010) data, we set $\lambda = 8$ and $R_{\text{model}} = 1$ in h-SNE. We select the result that best preserves data distances from 30 repeats. After obtaining the embedded points, we transform the points from native representation to Poincaré disk model by performing the transformation on radial coordinates:

$$r_{\text{Poincare}} = \tanh(\frac{r_{\text{native}}}{2}) \tag{S17}$$

In g-SNE, the parameter is: $\lambda = 20$. In PCA, we use the first two principal components for visualization. In UMAP, we screen a wide range of the combination of two key parameters: number of neighbors $\in \{5, 10, 20, 50, 100\}$ and minimal distance $\in \{0.001, 0.01, 0.1, 0.5, 0.8\}$, each of the 25 combinations was repeated 30 times. The optimal combination of parameters that leads to largest distance correlation of Shepard diagram is: number of neighbors = 100 and minimal distance = 0.5, and the corresponding result is shown in Fig. 6.For Moignard et al. data Moignard et al. (2015), the parameters for h-SNE are: $\lambda = 10$, $R_{\text{model}} = 1$. The root node index is 1800. When plotting Poincaré map, we directly use the embedding positions provided in Klimovskaia et al. Klimovskaia et al. (2020).

1.5. Evaluation of embedding

The Pearson correlation coefficient of Shepard diagram (embedding distances versus data distances) is used to measure the preservation of distances and global structure. For local clusters, we apply silhouette score Rousseeuw (1987) to our embedding results. Silhouette score measures the quality of data partitioning and clustering in graphical representation of objects, which in our case can be used to measure the consistency of the data configuration in 2D embeddings with the "ground truth" cluster labels. The higher score indicates better consistency with data labels. We consider all the three types of labels available – six hematopoietic properties, four malignancy properties and fifteen subtypes, and calculate the geometric mean of silhouette scores obtained by using these three labels:

$$s = \sqrt[3]{s_1 s_2 s_3} \tag{S18}$$

Where s_1, s_2, s_3 represent the silhouette scores by using the three types of labeling respectively. The mean score s is used to quantify the local structure preservation for the five visualization algorithms.

1.6. Data preprocessing and analysis

No pre-processing was done for the microarray dataset from human samples Lukk et al. (2010).

For scRNA-seq dataset from Han et al. (2018), we use Seurat packages Butler et al. (2018) to perform normalization, feature selection and scaling for the data, and to select the top 50 principal components for analyses. These results are reported in Figure 5.The results without pre-processing (and keeping all 1000 principle components) were qualitatively similar but had slightly reduced hyperbolic radii: $R_{\text{mouse brain}} = 1.62 \pm 0.02$, $R_{\text{mouse lung}} = 1.47 \pm 0.03$, $R_{\text{mouse kidney}} = 1.45 \pm 0.03$, and $R_{\text{mouse embryo}} = 0.98 \pm 0.02$ (compare with number in Fig. 5). These obervations are consistent with the observation that noise reduction done during pre-processing makes hyperbolic effects stronger and more apparent.

For mouse hematopoiesis data, we use the processed data from Klimovskaia et al. Klimovskaia et al. (2020). The Seurat analysis, violin plots and linear regression were performed using R version 3.6.2, the other analyses were performed using MATLAB R2017a.

2. Supplemental figures

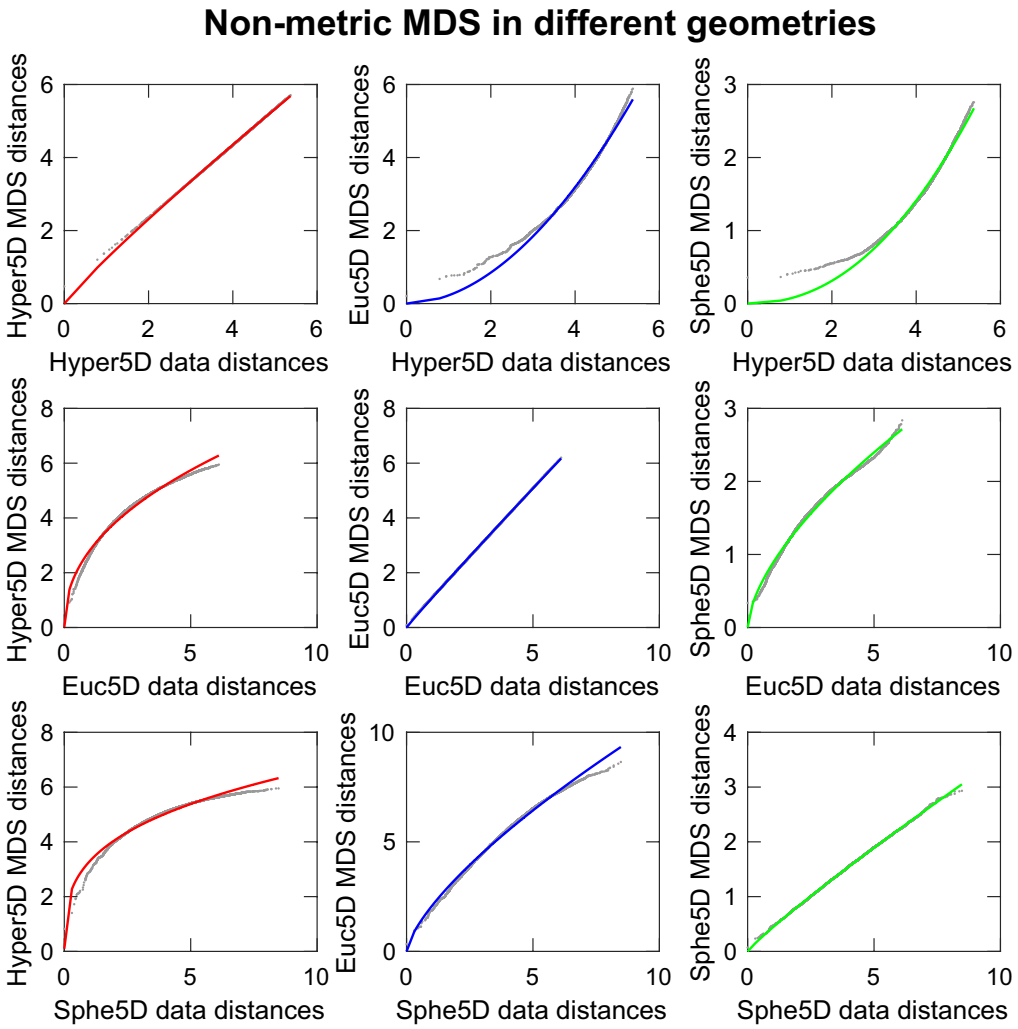

Figure S1: **Illustration of non-metric MDS embedding in different geometries. Related to Figure 2.** 100 synthetic points in 5D hyperbolic (top), Euclidean (middle) and spherical (bottom) space are embedded into 5D hyperbolic (left), Euclidean (middle) and spherical (right) space respectively. The hyperbolic radius is $R = 5$ for both data and model. The fitting methods are the same as in Figure 2 in the manuscript.

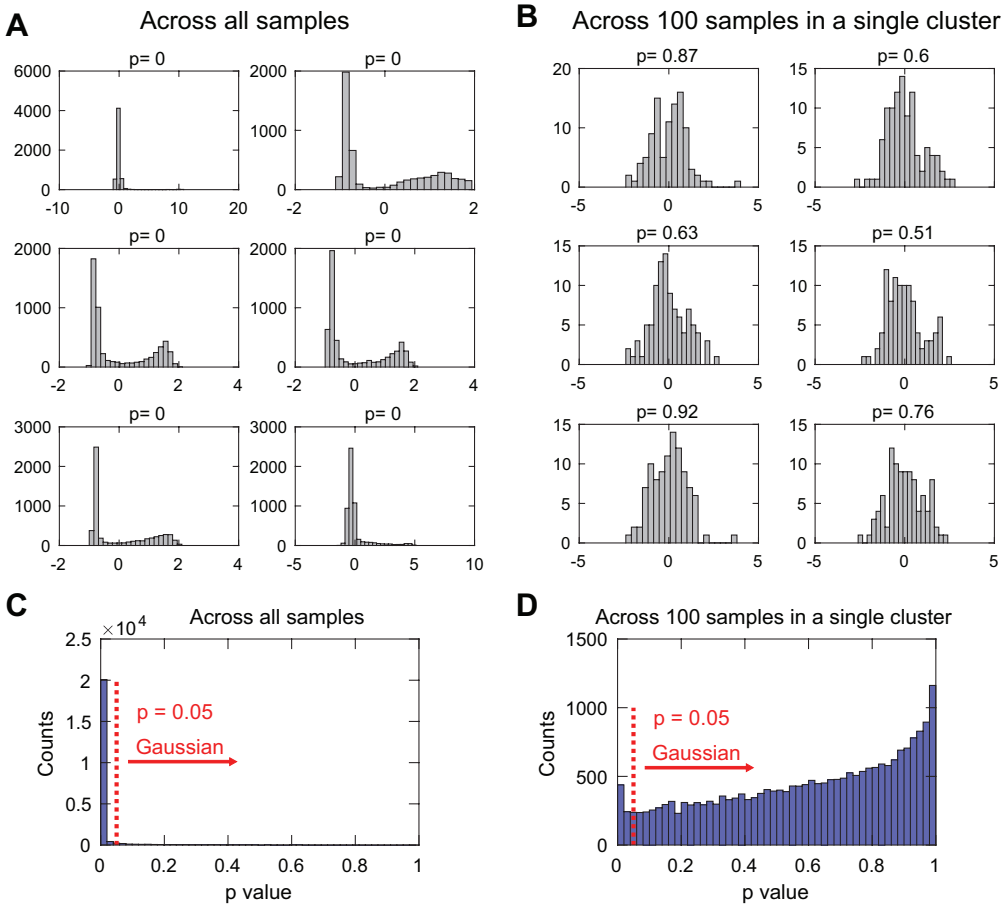

Figure S2: **Gaussianity of normalized gene expressions across the whole samples and from a single cluster. Related to Figure 3.** (A) Gene expression distributions of the six most non-Gaussian distributed probes across all the samples, p values were given by one-sample Kolmogorov-Smirnov test for Gaussianity, the null hypothesis is that the normalized distribution is standard normal distribution. (B) Gene expression distributions of the same six non-Gaussian probes as in (A) across 100 samples in one of the k-means ($k = 80$) cluster. (C) p value distributions of all the probes across the whole samples. (D) p value distributions of all the probes across 100 samples in one of the k-means cluster.

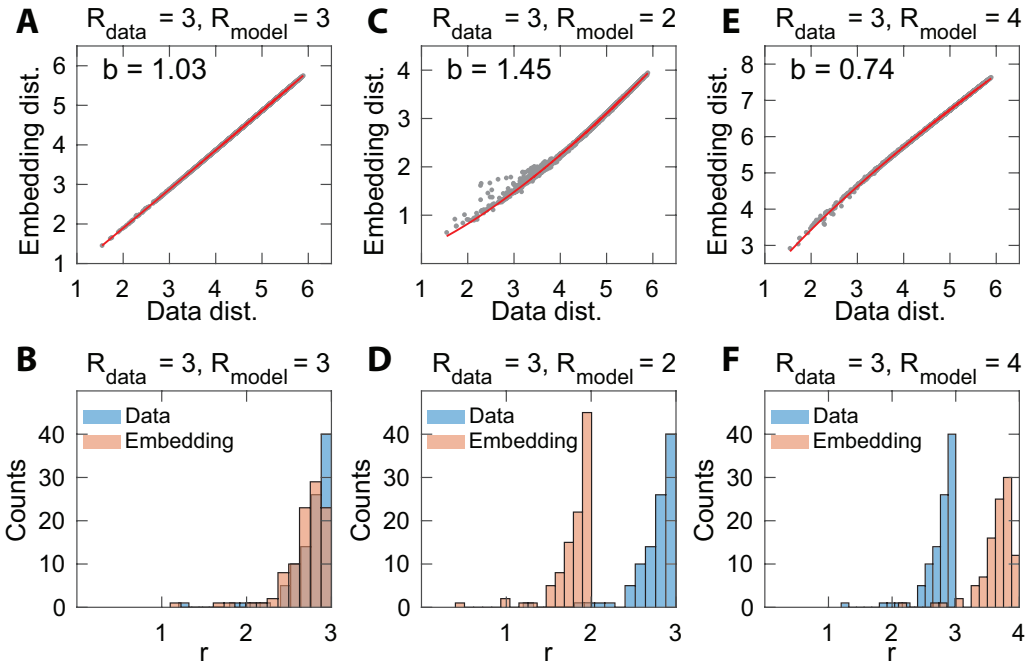

Figure S3: **HMDS embedding of hyperbolic data with different R_{model}. Related to Figure 5.** 100 points are sampled in hyperbolic space with $D = 5$, $R_{\text{data}} = 3$. The embedding dimension is $D = 5$ in (A-F). The Shepard diagram convexity κ is shown in panel (A,C,E). (A) Shepard diagram of HMDS embedding of the samples to 5D hyperbolic space with $R_{\text{model}} = 3$. (B) Histogram of radial coordinates r of 100 sample points and model points after HMDS embedding with $R_{\text{model}} = 3$. (C-D) $R_{\text{model}} = 2$. (E-F) $R_{\text{model}} = 4$.

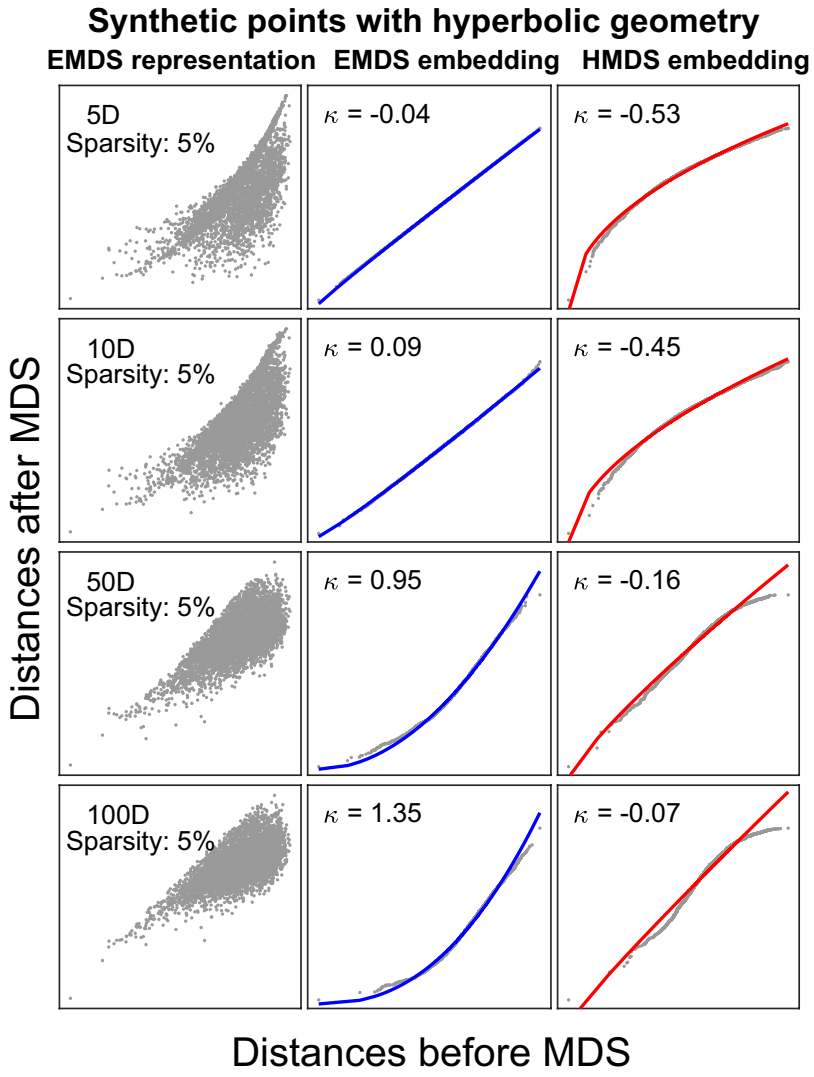

Figure S4: **Robustness to changes in data sparseness on geometry detection. Related to Figure 2 and Figure 5.** In Euclidean representation after EMDS, the coordinates of all the points are thresholded by fifth percentile of all the coordinate values to simulate the sparse RNA-seq values of cells. The left column shows the plots of thresholded embedding distances versus distances before MDS. The fitting plots and inserted convexity values in the rest columns have same meanings as in Figure 2 in the manuscript.

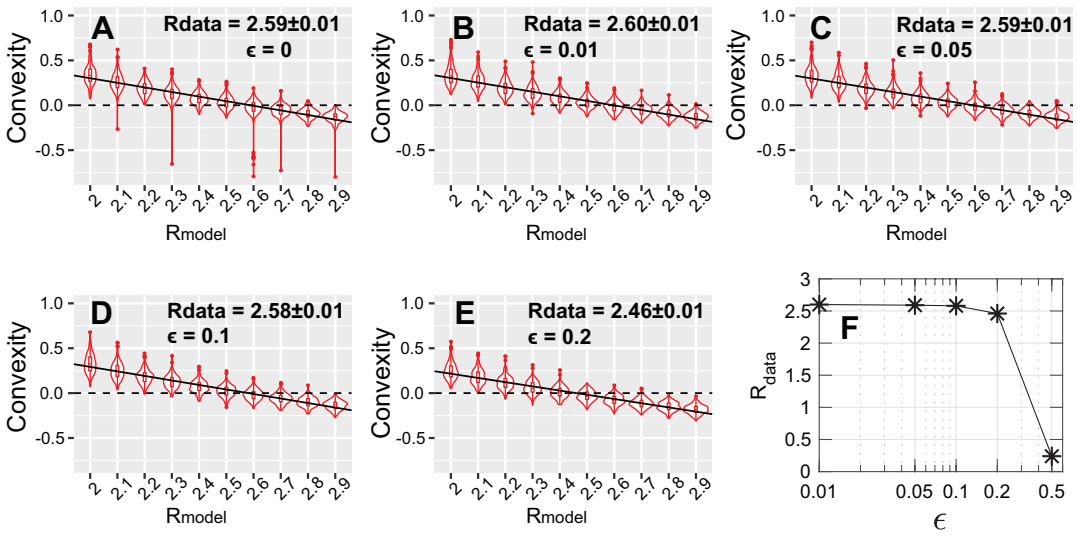

Figure S5: **Screening $R_{\mathbf{data}}$ of Lukk data with different magnitude of noise added by doing HMDS. Related to Figure 5.** The embedding dimension is $D = 5$. The noise ϵ is added as multiplicative Gaussian noise: $M_{noise} = M[1 + \epsilon \cdot N(0, 1)]$. (A-E) Fitting of R_{data} under different ϵ. (F) Plot R_{data} as the function of ϵ.

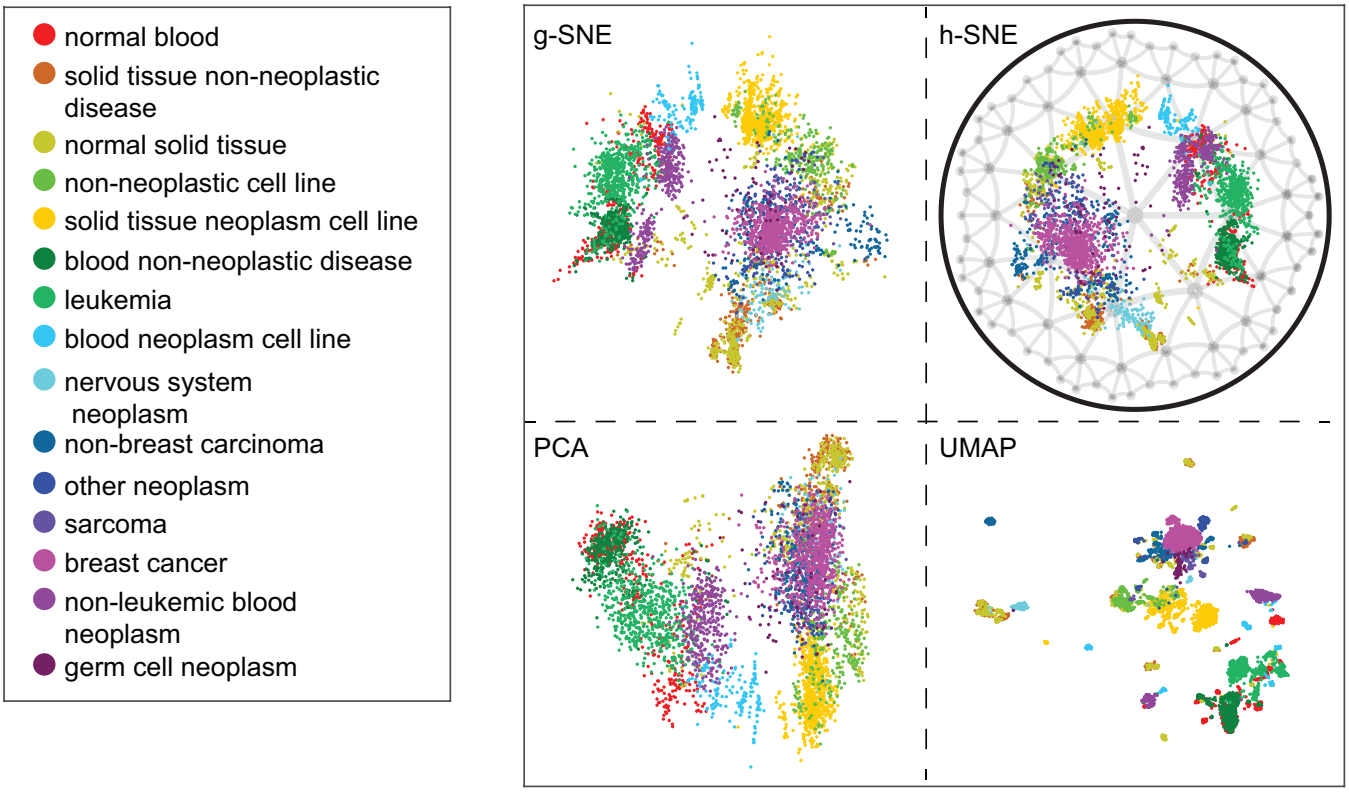

Figure S6: **Comparison of two-dimensional visualizations of human expression data using g-SNE, h-SNE, PCA and UMAP. Related to Figure 6.** The samples are colored according to the 15 tissue and disease types.

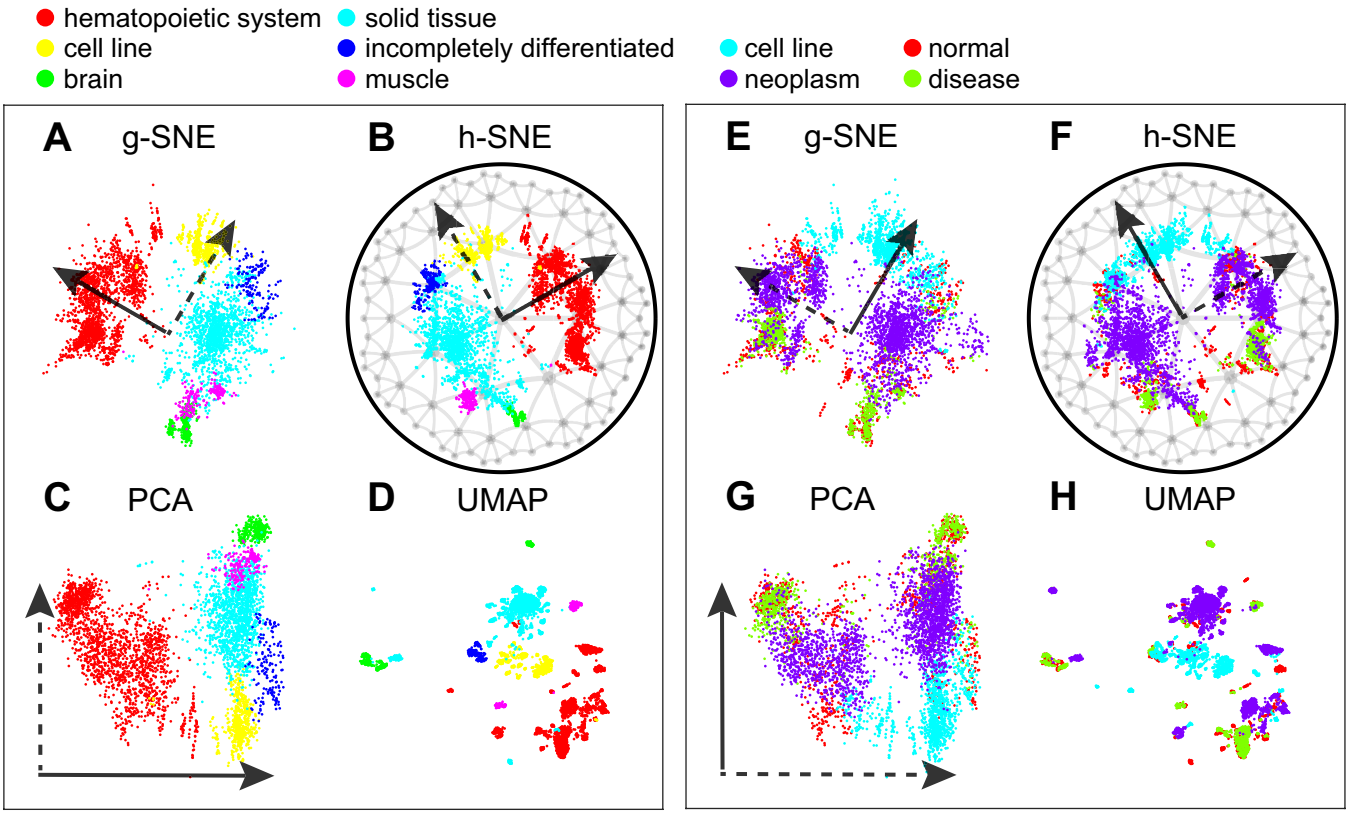

Figure S7: **Comparison of two-dimensional visualizations of human gene expression data in different algorithms. Related to Figure 6.** (A-D) g-SNE, h-SNE, PCA and UMAP embeddings of human samples classified by hematopoietic properties, these labels also represent the six major clusters identified by Lukk et al. The hematopoietic axes are shown in solid lines in g-SNE(A), h-SNE(B) and PCA(C). (E-H) g-SNE, h-SNE, PCA and UMAP embeddings of human samples classified by malignancy properties. The malignancy axes are shown in solid lines in g-SNE (E), h-SNE(F) and PCA(G). The six major clusters, the hematopoietic axis and the malignancy axis are hard to identify in UMAP.

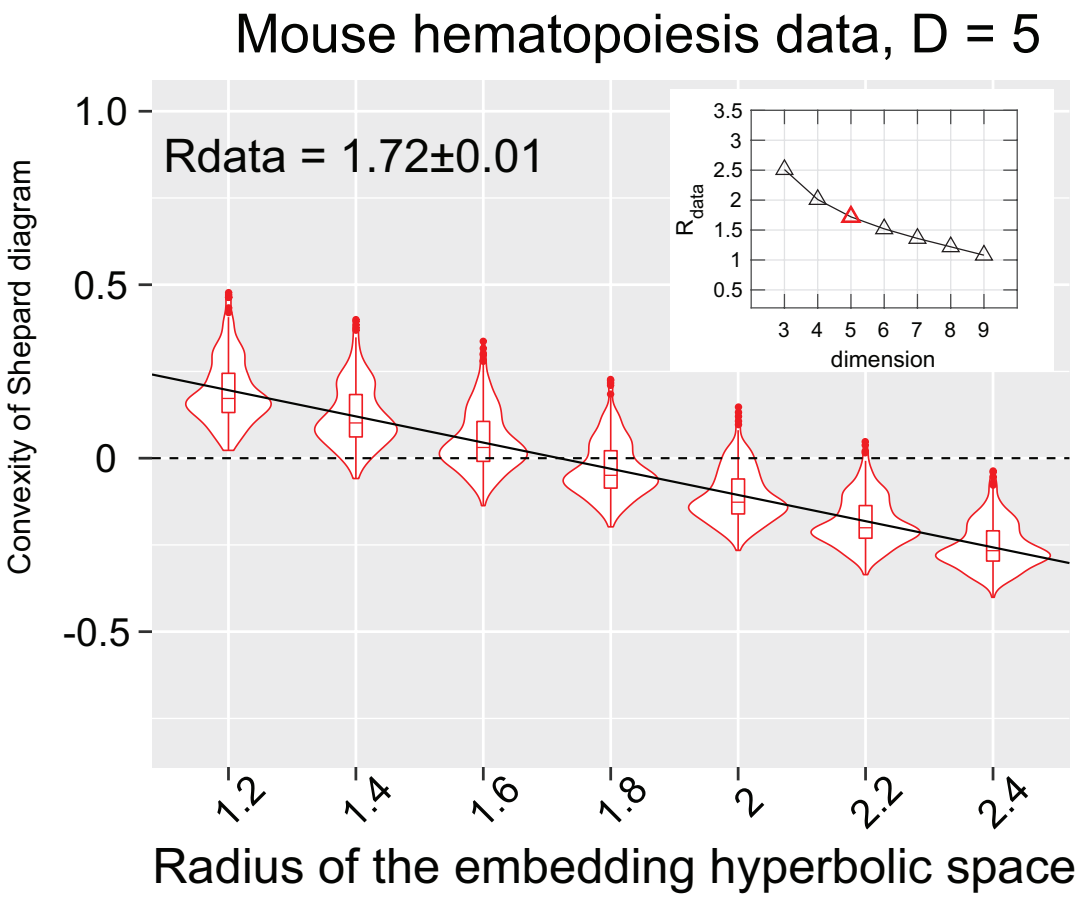

Figure S8: **Screening $R_{\mathbf{data}}$ of mice hematopoiesis data in Moignard et al. by doing HMDS. Related to Figure 7.** The inset shows the fitted R_{data} as the function of the embedding dimension.