Supporting Information

How many cells are enough for single-cell infrared spectroscopy?

Yadi Wang, ‡[abg] Wentao Dai, ‡[cd] Yue Wang,[e] Jixiang Liu, [d] Zhixiao Liu,[a] Yuanyuan Li,[cd] Xueling Li*f, Jun Hu[ab] and Junhong Lü*[abe]

[a.] Key Laboratory of Interfacial Physics and Technology, Shanghai Institute of Applied Physics, Chinese Academy of Science, Shanghai 201800, China.

[b.] Shanghai Synchrotron Radiation Facility, Zhangjiang Laboratory; Shanghai Advanced Research Institute, Chinese Academy of Science, Shanghai 201203, China.

[c]Department of Surgery, Shanghai Key Laboratory of Gastric Neoplasms, Shanghai Institute of Digestive Surgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[d.] Shanghai Center for Bioinformation Technology & Shanghai Engineering Research Center of Pharmaceutical Translation, Shanghai Industrial Technology Institute, Shanghai 201203, China.

[e.] College of Pharmacy, Binzhou Medical University, Yantai 264003, China.

[f.] Shanghai University of Medicine and Health Sciences, National Engineering Research Center for Nanotechnology, Shanghai 201318, China.

[g.] University of Chinese Academy of Sciences, Beijing 100049, China.

Corresponding authors

* lujunhong@zjlab.org.cn

* xlli@nercn.com.cn

Author Contributions

‡These authors contributed equally to this work.

Materials and Methods

1 Cell culture and sample preparation

Primary human mesenchymal stem cells (hMSCs, obtained from Medical College of Shandong University) were cultured to the eighth generation in complete medium (low glucose DMEM containing 10% fetal bovine serum, 100 U/ml penicillin and 100 μg/ml streptomycin; GIBCO). Primary human umbilical vein endothelial cells (HUVECs, obtained from Allcells, Shanghai) were cultured in complete medium (Allcells) and passaged to no more than 5 generations. All cells were incubated at 37 ℃ with 5% CO_2 saturated humidity. Lipopolysaccharide (LPS; from *Escherichia coli* O55:B5) with a final concentration of 1 μg/ml was added to stimulate HUVECs for 6 hours. To collect cells, the cells were detached from the petri dish using a cell scraper and then transferred into centrifuge tubes. After centrifugation, the supernatant was abandoned, and cells were resuspended and fixed with 4% paraformaldehyde for 20 minutes. The fixed cells were then washed three times using ultrapure water and adjusted to 10^4 - 10^5 single cells per milliliter, then preserved at 4 ℃ before experiments. Before FTIR measurements, 10 μl cell suspension was dropped on a barium fluoride (BaF_2) window and dried completely under room temperature for at least 30 minutes, single cells would disperse on the window.

2 Synchrotron Fourier transform infrared (SR-FTIR) microspectroscopy

Single-cell SR-FTIR measurements were carried out at beamline BL01B of Shanghai Synchrotron Radiation Facility (SSRF) which equipped with Nicolet 6700 Fourier transform infrared spectrometer, Continuμm infrared microscope and 32× infrared objective. Spectra in mid-infrared region were collected within the wavenumber range from 4000 to 600 cm^{-1}. When collecting spectra of hMSCs, the aperture size was set to 20 × 20 μm, spectral resolution was 4 cm^{-1} and 256 co-added scans each spectrum. When collecting single HUVECs, the aperture size was set 10 ×10 μm and 128 co-added scans. Original spectra were smoothed (9-point) and automatic baseline corrected, and their second-derivative spectra were further calculated. Data collection and pre-treatment were all operated on OMNIC (Thermo Fisher Scientific Inc.).

3 Constructing single-cell infrared spectral absorption matrices

The pre-treated second-derivative spectra were further processed for extracting their minimal values (corresponding to absorption peak positions of the original spectra). For hMSCs, 16 spectral peaks were collected, while 17 peaks were collected for HUVECs, with their band positions and corresponding functional groups listed in Table 1 and Table 2, respectively. Each cell can be characterized as a multidimensional vector whose dimensions were equal to the number of feature peaks (spectral absorption peaks). Then, an infrared spectral matrix was formed based on absorption peak values of all the cells in the same condition.

4 Simulated Data

We used random bootstrap sampling methods for simulating single-cell infrared spectral data based on its corresponding raw datasets. Under four cell number conditions (5, 10, 15 and 20 single cells), five single-cell IR datasets were generated each.

5 Similarity measure and distance matrix

The similarities among cells were calculated by two different measure methods. The first similarity distance measure is Euclidean distance:

$$D_{Eu}(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (1)$$

In this formula, p and q are two cells represented by n dimensional vectors. $D_{Eu}(p,q)$ is the Euclidean distance between these two cells, whose value is negative correlation with their similarity.

The other similarity distance measure is Chebyshev distance, whose formula is as follows:

$$D_{Chebyshev}(p,q) = \max_{i \in [1,n)} (|p_i - q_i|) \quad (2)$$

$D_{Chebyshev}(p,q)$ is the chebyshev distance between two cells p and q. It is the greatest different distance of these cells along any dimension. Its value is also negative correlation with cell similarity.

Based on the similarity distance among cells in each single-cell IR dataset, we could construct the distance matrix corresponding to the specific state data. For example, according to five single-cell IR datasets in 20 cells group, we could generate five corresponding distance matrices on each similarity measure.

6 Computation of variation on batch effects

To quantitatively evaluate batch effects under different experimental conditions, the coefficient of variation (CV) were used to express the variation of similarity measure based on single-cell IR spectra. The CV decreases with less variations and batch effects showing more reliable analysis results. In this study, the CV was calculated as follows:

$$CV_{EG} = \mu / \sigma \quad (3)$$

The μ and σ represented the mean and standard deviation of mean values of the upper triangular distance matrix for a specific experimental group (EG), respectively. The EG represented an experimental group which has parameters such as cell types or conditions, cell numbers, feature peaks and similarity measures.

Taking single-cell IR detection on hMSCs as an example, each cell was represented as a 16-

dimensional vector (16 characteristic absorption peaks). A total of 5 experiments were performed to get 5 group data, each group containing 10 cells data. Then five 16×10 dimensional cell similarity distance matrices were obtained after calculating the Euclidean distance among cells. In units of matrices，the mean values of the upper triangular matrix elements for the above five matrices were calculated separately. According to the five values, the corresponding mean, standard deviation and CV were obtained as the above equation (3). The batch effects of these five MSC single-cell IR experiments were evaluated by the CV value.

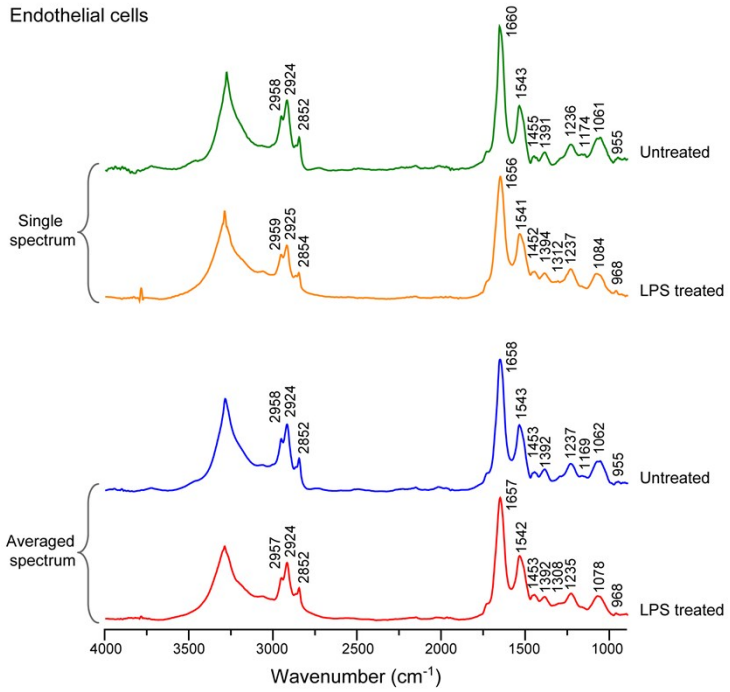

Figure S1 Comparison of single-cell spectrum and the averaged spectrum with 20 cells

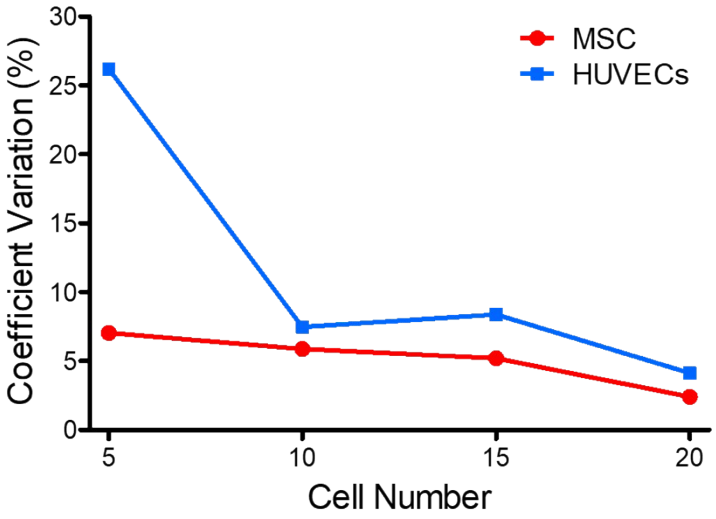

Figure S2 The coefficient variation(CV) for single-cell infrared spectra among mesenchymal stem cells(MSCs) and endothelial cells (HUVECs).

Table S1 Absorption bands of hMSCs.

Number	Macromolecules	Functional groups	Frequency (cm^{-1})
1	**Lipids**	$\nu_{as}(CH_3)$	2950-2970
2		$\nu_{as}(CH_2)$	2917-2937
3		$\nu_s(CH_3)$	2863-2883
4		$\nu_s(CH_2)$	2842-2862
5	**Proteins**	ν_s(C=O) of esters	1729-1749
6		N-H bending of Amide I	1678-1698
7		α-helical of Amide I	1644-1664
8		Amide II	1538-1558
9		‘‘Tyrosine’’ band	1504-1524
10		$\delta_s(CH_2)$	1454-1474
11		$\nu_s(COO^-)$	1372-1392
12	**Nucleic acids and sugars**	$\nu_{as}(PO_2^-)$	1230-1242
13		ν_s(C-O), ν_{as}(CO-O-C)	1166-1176
14		ν_s(C-O)	1147-1157
15		ν_s(C-O)	1120-1130
16		$\nu_s(PO_2^-)$	1075-1095

Table S2 Absorption bands of HUVECs.

Number	Macromolecules	Functional groups	Frequency (cm^{-1})
1	**Lipids**	$\nu_{as}(CH_3)$	2949-2972
2		$\nu_{as}(CH_2)$	2924-2934
3		$\nu_s(CH_3)$	2866-2879
4		$\nu_s(CH_2)$	2845-2857
5	**Proteins**	N-H bending of Amide I	1677-1696
6		α-helical of Amide I	1647-1654
7		Amide I of b-pleated sheet structures	1619-1633
8		Amide II	1526-1549
9		‘‘Tyrosine’’ band	1509-1516
10		$\delta_s(CH_2)$, $\nu_s(C=C)$	1489-1500
11		$\nu_s(COO^-)$	1378-1390
12	**Nucleic acids and sugars**	$\nu_{as}(PO_2^-)$	1237-1244
13		$\nu_s(C-O)$, $\nu_{as}(CO-O-C)$	1166-1175
14		$\nu_s(C-O)$	1155-1161
15		C-OH	1143-1148
16		ν(C-O)	1123-1131
17		$\nu_s(PO_2^-)$	1074-1083