Raman2RNA: Live-cell label-free prediction of single-cell RNA expression profiles by Raman microscopy

Koseki J. Kobayashi-Kirschvink[1,2,‡], Shreya Gaddam[1,9], Taylor James-Sorenson[1], Emanuelle Grody[1,3], Johain R. Ounadjela[1,3], Baoliang Ge[4], Ke Zhang[5], Jeon Woong Kang[2], Ramnik Xavier[1,6], Peter T. C. So[2,4], Tommaso Biancalani[1,9,†,‡], Jian Shu[1,3,5,†,‡], Aviv Regev[1,7,8,9,†,‡]

[1]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

[2]Laser Biomedical Research Center, G. R. Harrison Spectroscopy Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[3]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA

[4]Department of Mechanical and Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[5]Cutaneous Biology Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02129, USA

[6]Center for Computational and Integrative Biology and Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

[7]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[8]Howard Hughes Medical Institute, Cambridge, MA 02142, USA

[9]Present address: Genentech, 1 DNA Way, South San Francisco, CA 94080, USA

[†]These authors contributed equally

[‡]Correspondence: kkobayas@broadinstitute.org; aviv.regev.sc@gmail.com; jian.shu@mgh.harvard.edu; tbiancal@broadinstitute.org;

Error! Hyperlink reference not valid.Single cell RNA-Seq (scRNA-seq) and other profiling assays have opened new windows into understanding the properties, regulation, dynamics, and function of cells at unprecedented resolution and scale. However, these assays are inherently destructive, precluding us from tracking the temporal dynamics of live cells, in cell culture or whole organisms. Raman microscopy offers a unique opportunity to comprehensively report on the vibrational energy levels of molecules in a label-free and non-destructive manner at a subcellular spatial resolution, but it lacks in genetic and molecular interpretability. Here, we developed Raman2RNA (R2R), an experimental and computational framework to infer single-cell expression profiles in live cells through label-free hyperspectral Raman microscopy images and multi-modal data integration and domain translation. We used spatially resolved single-molecule RNA-FISH (smFISH) data as anchors to link scRNA-seq profiles to the paired spatial hyperspectral Raman images, and trained machine learning models to infer expression profiles from Raman spectra at the single-cell level. In reprogramming of mouse fibroblasts into induced pluripotent stem cells (iPSCs), R2R accurately (r>0.96) inferred from Raman images the expression profiles of various cell states and fates, including iPSCs, mesenchymal-epithelial transition (MET) cells, stromal cells, epithelial cells, and fibroblasts. R2R outperformed inference from brightfield images, showing the importance of spectroscopic content afforded by Raman microscopy. Raman2RNA lays a foundation for future investigations into exploring single-cell genome-wide molecular dynamics through imaging data, *in vitro* and *in vivo*.

Keywords: Raman microscopy, single-cell transcriptomics, multi-domain translation

Main

Cellular states and functions are determined by a dynamic balance between intrinsic and extrinsic programs. Dynamic processes such as cell growth, stress responses, differentiation, and reprogramming are not determined by a single gene, but by the orchestrated temporal expression and function of multiple genes organized in programs and their interactions with other cells and the surrounding environment[1]. To understand how cells change their states in physiological and pathological conditions it is essential to decipher the dynamics of the underlying gene programs.

Despite major advances in single cell genomics and microscopy, we still cannot track live cells and tissues at the genomic level. On the one hand, single cell and spatial genomics have provided a view of gene programs and cell states at unprecedented scale and resolution[1], but these measurement methods are destructive, and involve tissue fixation and freezing and/or cell lysis, precluding us from directly tracking the dynamics of full molecular profiles in live cells or organisms. While advanced computational methods, such as pseudo-time algorithms (*e.g.*, Monocle[2], Waddington-OT[3]) and velocity-based methods (*e.g.*, velocyto[4], scVelo[5]), can infer dynamics from snapshots of molecular profiles, they rely on assumptions that remain challenging to verify experimentally[6]. On the other hand, fluorescent reporters can be used to monitor the dynamics of individual genes and programs within live cells, but are limited in the number of targets they can report[7], must be chosen ahead of the experiment and often involve genetically engineered cells. Moreover, the vast majority of dyes and reporters require fixation or can interfere with nascent biochemical processes and alter the natural state of the gene of interest[7]. Therefore, it remains technically challenging to dynamically monitor the activity of a large number of genes simultaneously.

68 Raman microscopy opens a unique opportunity for monitoring live cells and tissues, as it
69 collectively reports on the vibrational energy levels of molecules in a label-free and non-
70 destructive manner at a subcellular spatial resolution, thus providing molecular fingerprints of
71 cells[8]. Pioneering research has demonstrated that Raman microscopy can be used for
72 characterizing cell types and cell states[8], non-destructively diagnosing pathological specimens
73 such as tumors[9], characterizing the developmental states of embryos[10], and identifying bacteria
74 with antibiotic resistance[11]. However, the complex and high-dimensional nature of the spectra, the
75 spectral overlaps of biomolecules such as proteins and nucleic acids, and the lack of unified
76 computational frameworks have hindered the decomposition of the underlying molecular
77 profiles[7,8].

78 To address this challenge and leverage the complementary strengths of Raman microscopy and
79 scRNA-Seq, we developed Raman2RNA (R2R), an experimental and computational framework
80 for inferring single-cell RNA expression profiles from label-free non-destructive Raman
81 hyperspectral images (**Fig. 1**). R2R takes as input spatially resolved hyperspectral Raman images
82 from live cells, smFISH data of selected markers from the same cells, and scRNA-seq from the
83 same biological system. R2R then uses the smFISH data as an anchor to learn a model that links
84 spatially resolved hyperspectral Raman images to scRNA-seq. Finally, from this model, R2R then
85 computationally infers the anchor smFISH measurements from hyperspectral Raman images and
86 then the single-cell expression profiles. The result is a label-free live-cell inference of single-cell
87 expression profiles.

88 To facilitate data acquisition, we developed a high-throughput multi-modal spontaneous Raman
89 microscope that enables automated acquisition of Raman spectra, brightfield, and fluorescent
90 images. In particular, we integrated Raman microscopy optics to a fluorescence microscope, where

91 high-speed galvo mirrors and motorized stages were combined to achieve a large field of view
92 (FOV) scanning, and where dedicated electronics automate measurements across multiple
93 modalities (**Extended Data Fig. 1-2, Methods**).

94 We first demonstrated that R2R can infer profiles of two distinct cell types: mouse induced
95 pluripotent stem cells (iPSCs) expressing an endogenous *Oct4*-GFP reporter and mouse
96 fibroblasts[12]. To this end, we mixed the cells in equal proportions, plated them in a gelatin-coated
97 quartz glass-bottom Petri dish, and performed live-cell Raman imaging, along with fluorescent
98 imaging of live-cell nucleus staining dye (Hoechst 33342) for cell segmentation and image
99 registration, and an iPSC marker gene, *Oct4*-GFP (**Fig. 2a**). The excitation wavelength for our
100 Raman microscope (785 nm) was distant enough from the GFP Stokes shift emission, such that
101 there was no interference with the cellular Raman spectra (**Extended Data Fig. 3**). Furthermore,
102 there was no notable photo-toxicity induced in the cells. After Raman and fluorescence imaging,
103 we fixed and permeabilized the cells and performed smFISH (with hybridization chain reaction
104 (HCR[13]), **Methods**) of marker genes for mouse iPSCs (*Nanog*) and fibroblasts (*Col1a1*). We
105 registered the nuclei stains, GFP images, HCR images, and Raman images through either
106 polystyrene control bead images or reference points marked under the glass bottom dishes
107 (**Extended Data Fig. 4**, **Methods**).

108 The Raman spectra distinguished the two cell populations in a manner congruent with the
109 expression of their respective reporter (measured live or by smFISH in the same cells), as reflected
110 by a low-dimensional embedding of hyperspectral Raman data (**Fig. 2b**). Specifically, we focused
111 on the fingerprint region of Raman spectra (600-1800 cm^{-1}, 930 of the 1,340 features in a Raman
112 spectrum), where most of the signatures from various key biomolecules, such as proteins, nucleic
113 acids, and metabolites, lie[8]. After basic preprocessing, including cosmic-ray and background

114 removal and normalization, we aggregated Raman spectra that are confined to the nuclei, obtaining
115 a 930-dimensional Raman spectroscopic representation for each cell's nucleus. We then visualized
116 these Raman profiles in an embedding in two dimensions using Uniform Manifold Approximation
117 and Projection (UMAP)[14] and labeled cells with the gene expression levels that were concurrently
118 measured by either an *Oct4*-GFP reporter or smFISH (**Fig. 2b**). The cells separated clearly in their
119 Raman profiles in a manner consistent with their gene expression characteristics, forming two
120 main subsets in the embedding, one with cells with high *Oct4* and *Nanog* expression (iPSCs
121 markers) and another with cells with relatively high *Col1a1* expression (fibroblasts marker),
122 indicating that Raman spectra reflect cell-intrinsic expression differences (**Fig. 2b**).

123 We further successfully trained a classifier to classify the 'on' or 'off' expression states of *Oct4*,
124 *Nanog* and *Col1a1* in each cell based on its Raman profile (**Methods**). We trained a logistic
125 regression classifier with 50% of the data and held out 50% for testing. We predicted *Oct4* and
126 *Nanog* expression states with high accuracy on the held-out test data (area under the receiver
127 operating characteristic curve (AUROC) = 0.98 and 0.95, respectively; **Fig. 2c**), indicating that
128 expression of iPSC markers can be predicted confidently from Raman spectra of live, label-free
129 cells. We also successfully classified the expression state of the fibroblast marker *Col1a1*
130 (AUROC = 0.87; **Fig. 2c**), albeit with lower confidence, which is consistent with the lower contrast
131 in *Col1a1* expression (**Fig. 2b**) between iPSC (*Oct4*+ or *Nanog*+ cells) *vs*. non-iPSCs, compared
132 to *Oct4* or *Nanog*. Most misclassifications occurred when the ground truth expression levels were
133 near the threshold of the classifier, showing that misclassifications were likely due to the
134 uncertainty in the ground truth expression level (**Extended Data Fig. 5**).

135 Next, we asked if the Raman images could predict entire expression profiles non-destructively at
136 single-cell resolution. To this end, we aimed to reconstruct scRNA-seq profiles from Raman

137 images by multi-modal data integration and translation, using multiplex smFISH data to anchor
138 between the Raman images and scRNA-seq profiles (**Fig. 3a**). As a test case, we focused on the
139 mouse iPSC reprogramming model system, where we have previously generated ~250,000
140 scRNA-seq profiles at ½ day intervals throughout an 18 day, 36 time point time course of
141 reprogramming[3] (**Methods**). We used Waddington-OT[3] (WOT) to select from the scRNA-seq
142 profiles nine anchor genes that represent diverse cell types that emerge during reprogramming
143 (iPSCs: *Nanog*, *Utf1* and *Epcam*; MET and neural: *Nnat* and *Fabp7*; epithelial: *Krt7* and *Peg10*;
144 stromal: *Bgn* and *Col1a1*; **Methods**). We performed live-cell Raman imaging from day 8 of
145 reprogramming, in which distinct cell types begin to emerge[3], up to day 14.5, at half-day intervals,
146 totaling 14 time points (**Methods**). We imaged ~500 cells per plate at 1μm spatial resolution.
147 Finally, we fixed cells immediately after each Raman imaging time point followed by smFISH on
148 the 9 anchor genes (**Methods**).

149 Strikingly, a low dimensional representation of the Raman profiles showed that they encoded
150 similar temporal dynamics to those observed with scRNA-seq during reprogramming (**Fig. 3b,c,**
151 **Extended Data Fig. 6**), indicating that they may qualitatively mirror scRNA-seq.

152 Integrating Raman and scRNA-seq profiles (**Methods**), R2R then learned a model that can infer
153 an scRNA-seq profile for each Raman imaged cell, by first predicting smFISH anchors from the
154 Raman profiles using Catboost[15] (**Methods**) and then using our Tangram[16] method to map from
155 the anchors to full scRNA-seq profiles (**Fig. 1**, **Fig. 3d-f**). In the first step, we averaged the smFISH
156 signal within a nucleus to represent a single nucleus's expression level. As we conducted smFISH
157 of 9 genes, the result was a 9-dimensional smFISH profile for each single nucleus. Then, Raman
158 profiles were translated to these 9-dimensional profiles with Catboost[15], a non-linear regression
159 model, using 50% of the Raman and smFISH profiles as training data.

160 In the second step, we mapped these anchor smFISH profiles to full scRNA-seq profiles using
161 Tangram, yielding well-predicted single cell RNA profiles, as supported by several lines of
162 evidence. First, we performed leave-one-out cross-validation (LOOCV) analysis, in which we used
163 eight out of the nine anchor genes to integrate Raman with scRNA-seq, and compared the predicted
164 expression of the remaining genes to its smFISH measurements. The predicted left-out genes based
165 on scRNA-seq showed a significant correlation with the measured smFISH expression for any left-
166 out gene (Pearson r~0.7, p-value<10^{-100}, **Fig. 3d**). Notably, when we analogously applied a
167 modified U-net[17] to infer smFISH profiles from brightfield (**Extended Data Fig. 15**, **Methods**),
168 we observed a poor, near-random prediction of expression profiles for all 9 genes in leave-one-out
169 cross-validation (r<0.15), indicating that, unlike Raman spectra, brightfield z-stack images either
170 do not have the necessary information to infer expression profiles, or require more data. Second,
171 we compared the real (scRNA-seq measured) and R2R predicted expression profiles averaged
172 across cells of the same cell type ("pseudobulk" for each of iPSCs, epithelial cells, stromal cells,
173 and MET). Here, we obtained the "ground truth" cell types of the R2R profiles by transferring
174 scRNA-seq annotations to the matching smFISH profiles using Tangram's label transfer function.
175 Then, based on the labels, we averaged R2R's predicted profiles across the cells of a single cell
176 type. The two profiles (R2R-inferred and scRNA-seq pseudo-bulk per cell type) showed high
177 correlations (Pearson's r>0.96) (**Fig. 3e,f**, **Extended Data Fig. 7**), demonstrating the accuracy of
178 R2R at the cell type level. Furthermore, projecting the R2R predicted profiles of each cell onto an
179 embedding learned from the real scRNA-seq shows that the predicted profiles span the key cell
180 types as captured in real profiles (**Fig. 3g-j, Extended Data Fig. 8-12**). We note that the predicted
181 profiles had lower variance compared to real scRNA-seq. As this is observed even when co-
182 embedding only smFISH and scRNA-seq measurements (with no Raman data or projection,

183 **Extended Data Fig. 13**), we believe it mostly reflects the limited number and domain
184 maladaptation of the smFISH anchor genes used for integration. Given the similarity of the
185 separate embeddings of Raman and scRNA-seq profiles, future studies without anchors could
186 address this.

187 Lastly, we calculated feature importance scores in R2R predictions (**Methods**) and identified
188 Raman spectral features correlated with expression levels (**Fig. 3k**, **Extended Data Fig. 14**). For
189 example, Raman bands at approximately 752cm^{-1} (C-C, Try, cytochrome), 1004 cm^{-1} (CC, Phe,
190 Tyr), and 1445 cm^{-1} (CH_2, lipids) contributed to predicting iPSCs-related expression profiles,
191 which is consistent with previous research that employed single cell Raman spectra to identify
192 mouse embryonic stem cells (ESCs)[18] (**Fig. 3k**). The contributions of these bands were either
193 suppressed or increased for other cell types, such as stromal or epithelial cells (**Extended Data**
194 **Fig. 14**).

195 In conclusion, we reported R2R, a label-free non-destructive framework for inferring expression
196 profiles at single-cell resolution from Raman spectra of live cells, by integrating Raman
197 hyperspectral images with scRNA-seq data through paired smFISH measurements and multi-
198 modal data integration and translation. We inferred single-cell expression profiles with high
199 accuracy, based on both averages within cell types and co-embeddings of individual profiles. We
200 further showed that predictions using brightfield z-stacks had poor performance, indicating the
201 importance of Raman microscopy for predicting expression profiles.

202 R2R can be further developed in several ways. First, the throughput of single-cell Raman
203 microscopy is still limited. In this pilot study, we profiled ~6,000 cells in total. By using emerging
204 vibrational spectroscopy techniques, such as Stimulated Raman Scattering microscopy[19] or photo-

thermal microscopy[20,21], we envision increasing throughput by several orders of magnitude, to match the throughput of massively parallel single cell genomics. Second, because molecular circuits and gene regulation are structured, with strong co-variation in gene expression profiles across cells, we can leverage the advances in computational microscopy to infer high-resolution data from low-resolution data, such as by using compressed sensing, to further increase throughput[22]. Third, increasing the number of anchor genes (*e.g.*, by seqFISH[23], merFISH[24], STARmap[25], or ExSeq[26]) can increase our prediction accuracy and capture more single-cell variance. Additionally, with single-cell multi-omics, we can project other modalities, such as scATAC-seq from Raman spectra. Finally, given the similarity in the overall independent embedding of Raman and scRNA-seq profiles, we expect computational methods such as multi-domain translation[27] to allow mapping between Raman spectra and molecular profiles without measuring any anchors *in situ*. Overall, with further advances in single-cell genomics, imaging, and machine learning, Raman2RNA could allow us to non-destructively infer omics profiles at scale *in vitro*, and possibly *in vivo* in living organisms.

220 **Materials and Methods**

221 **Mouse fibroblast reprogramming**

222 OKSM secondary mouse embryonic fibroblasts (MEFs) were derived from E13.5 female embryos
223 with a mixed B6;129 background. The cell line used in this study was homozygous for ROSA26-
224 M2rtTA, homozygous for a polycistronic cassette carrying *Oct4*, *Klf4*, *Sox2*, and *Myc* at the
225 *Col1a1* 3' end, and homozygous for an EGFP reporter under the control of the *Oct4* promoter.
226 Briefly, MEFs were isolated from E13.5 embryos from timed-matings by removing the head,
227 limbs, and internal organs under a dissecting microscope. The remaining tissue was finely minced
228 using scalpels and dissociated by incubation at 37°C for 10 minutes in trypsin-EDTA
229 (ThermoFisher Scientific). Dissociated cells were then plated in MEF medium containing DMEM
230 (ThermoFisher Scientific), supplemented with 10% fetal bovine serum (GE Healthcare Life
231 Sciences), non-essential amino acids (ThermoFisher Scientific), and GlutaMAX (ThermoFisher
232 Scientific). MEFs were cultured at 37°C and 4% CO_2 and passaged until confluent. All procedures,
233 including maintenance of animals, were performed according to a mouse protocol (2006N000104)
234 approved by the MGH Subcommittee on Research Animal Care[3].

235 For the reprogramming assay, 50,000 low passage MEFs (no greater than 3-4 passages from
236 isolation) were seeded in 14 3.5cm quartz glass-bottom Petri dishes (Waken B Tech) coated with
237 gelatin. These cells were cultured at 37°C and 5% CO_2 in reprogramming medium containing
238 KnockOut DMEM (GIBCO), 10% knockout serum replacement (KSR, GIBCO), 10% fetal bovine
239 serum (FBS, GIBCO), 1% GlutaMAX (Invitrogen), 1% nonessential amino acids (NEAA,
240 Invitrogen), 0.055 mM 2-mercaptoethanol (Sigma), 1% penicillin-streptomycin (Invitrogen) and
241 1,000 U/ml leukemia inhibitory factor (LIF, Millipore). Day 0 medium was supplemented with 2

242 mg/mL doxycycline Phase-1 (Dox) to induce the polycistronic OKSM expression cassette. The
243 medium was refreshed every other day. On day 8, doxycycline was withdrawn. Fresh medium was
244 added every other day until the final time point on day 14. One plate was taken every 0.5 days
245 after day 8 (D8-D14.5) for Raman imaging and fixed with 4% formaldehyde immediately after for
246 HCR.

247 **High-throughput multi-modal Raman microscope**

248 Due to the lack of commercial systems, we developed an automated high-throughput multi-modal
249 microscope capable of multi-position and multi-timepoint fluorescence imaging and point
250 scanning Raman microscopy (**Extended Data Fig. 1**). A 749 nm short-pass filter was placed to
251 separate brightfield and fluorescence from Raman scattering signal, and the fluorescence and
252 Raman imaging modes were switched by swapping dichroic filters with auto-turrets. To realize a
253 high-throughput Raman measurement, galvo mirror-based point scanning and stage scanning was
254 combined to acquire each FOV and multiple different FOVs, respectively.

255 To realize this in an automated fashion, a MATLAB (2020b) script that communicates with Micro-
256 manager[28], a digital acquisition (DAQ) board, and Raman scattering detector (Princeton
257 Instruments, PIXIS 100BR eXcelon) was written (**Extended Data Fig. 2**). A 2D point scan Raman
258 imaging sequence was regarded as a dummy image acquisition in Micro-manager, during which
259 the script communicated via the DAQ board with 1. the detector to read out a spectrum, 2. the
260 mirror to update the mirror angles, and 3. shutters to control laser exposure. All communications
261 were realized using transistor-transistor logic (TTL) signaling. Updating of the galvo mirror angles
262 was conducted during the readout of the detector. While the script ran in the background, Micro-

263 manager initiated a multi-dimensional acquisition consisting of brightfield, DAPI, GFP, and
264 dummy Raman channel at multiple positions and z-stacks.

265 An Olympus IX83 fluorescence microscope body was integrated with a 785 nm Raman excitation
266 laser coupled to the backport, where the short-pass filter deflected the excitation to the sample
267 through an Olympus UPLSAPO 60X NA 1.2 water immersion objective. The backscattered light
268 was collimated through the same objective and collected with a 50 μm core multi-mode fiber,
269 which was then sent to the spectrograph (Holospec f/1.8i 785 nm model) and detector. The
270 fluorescence and brightfield channels were imaged by the Orca Flash 4.0 v2 sCMOS camera from
271 Hamamatsu Photonics. The exposure time for each point in the Raman measurement was 20 msec,
272 and laser power at the sample plane was 212 mW. Each FOV was 100x100 pixels, with each pixel
273 corresponding to about 1 μm. The laser source was a 785 nm Ti-Sapphire laser cavity coupled to
274 a 532 nm pump laser operating at 4.7W.

275 The time to acquire Raman hyperspectral images was roughly 8 minutes per FOV. With 8 minutes,
276 it is unrealistic to image an entire glass-bottom plate. Therefore, we visually chose representative
277 FOVs that cover all representative cell types including iPSC-like, epithelial-like, stromal-like and
278 MET cells. 20 FOVs were chosen for each plate, where roughly 15 FOVs were from the boundaries
279 of colonies, five from non-colonies, and one from non-cells to use for background correction.

280 Due to the extended Raman imaging time, evaporation of the immersion water was no longer
281 negligible. Therefore, we developed an automated water immersion feeder using syringe pumps
282 and syringe needles glued to the tip of the objective lens. Here, water was supplied at a flow rate
283 of 1 μL/min.

284 **iPSC and MEF mixture experiment**

285 Low passage iPSCs were first cultured in N2B27 2i media containing 3 mM CHIR99021, 1 mM
286 PD0325901, and LIF. On the day of the experiment, 750,000 iPSCs and 750,000 MEFs were plated
287 on the same gelatin-coated 3.5cm quartz glass-bottom Petri dish. Cells were plated in the same
288 reprogramming medium as previously described (with Dox) with the exception of utilizing DMEM
289 without phenol red (Gibco) instead of KnockOut DMEM. 6 hours after plating, the quartz dishes
290 were taken for Raman imaging and fixed with 4% formaldehyde immediately after for HCR.

291 **Anchor gene selection by Waddington-OT**

292 To select anchor genes for connecting spatial information to the full transcriptome data,
293 Waddington-OT (WOT)[3], a probabilistic time-lapse algorithm that can reconstruct developmental
294 trajectories, was used. We applied WOT to mouse fibroblast reprogramming scRNA-seq data
295 collected at matching time-points and culture condition (day 8-14.5 at ½ day intervals)[3]. For each
296 cell fate, we calculated the transition probabilities of each cell and selected the top 10 percentile
297 cells per time point (**Extended Data Fig. 6**). Based on this, we ran the *FindMarker* function in
298 Seurat[29] to find genes differentially expressed in these cell subsets per time point. Through this
299 approach, we chose two genes per cell type that are both found by Seurat and commonly used for
300 these cell types (iPSCs: *Nanog*, *Utf1*; epithelial: *Krt7*, *Peg10*; stromal: *Bgn*, *Col1a1*; MET and
301 neural: *Fabp7*, *Nnat*), along with one gene that is an early marker of iPSCs, *Epcam*.

302 **smRNA-FISH by hybridization chain reaction (HCR)**

303 Fixed samples were prepared for imaging using the HCR v3.0 protocol for mammalian cells on a
304 chambered slide, incubating at the amplification step for 45 minutes in the dark at room

305 temperature. Three probes with amplifiers conjugated to fluorophores Alexa Fluor 488, Alexa
306 Fluor 546, and Alexa Fluor 647 were used. Samples were stained with DAPI prior to imaging.
307 After imaging, probes were stripped from samples by washing samples once for 5 minutes in 80%
308 formamide at room temperature and then incubating three times for 30 minutes in 80% formamide
309 at 37ºC. Samples were washed once more with 80% formamide, then once with PBS, and reprobed
310 with another panel of probes for subsequent imaging.

311 **Image registration of Raman hyperspectral images and fluorescence/smFISH images**

312 Brightfield and fluorescence channels including DAPI and GFP, along with corresponding Raman
313 images, were registered by using 5 μm polystyrene beads deposited on quartz glass-bottom Petri
314 dishes (SF-S-D12, Waken B Tech) for calibration. The brightfield and fluorescence images of the
315 beads were then registered by the scale-invariant template matching algorithm of the OpenCV
316 (https://github.com/opencv/opencv) *matchTemplate* function followed by manual correction.
317 For the registration of smFISH and Raman images, four marks inscribed under the glass-bottom
318 Petri dishes were used as reference points (**Extended Data Fig. 4**). As the Petri dishes are
319 temporarily removed from the Raman microscope after imaging to do smFISH measurements, the
320 dishes cannot be placed back at the same exact location on the microscope. Therefore, the
321 coordinates of these reference points were measured along with the different FOVs. When the
322 dishes were placed again after smFISH measurements, the reference mark coordinates were
323 measured, and an affine mapping was constructed to calculate the new FOV coordinates. Lastly,
324 as smFISH consisted of 3 rounds of hybridization and imaging, the following steps were performed
325 to register images across different rounds with a custom MATLAB script:
326 1. Maximum intensity projection of nuclei stain and RNA images

2. Automatic registration of round 1 images to rounds 2 and 3 based on nuclei stain images and MATLAB function *imregtform*. First, initial registration transformation functions were obtained with a similarity transformation model passing the 'multimodal' configuration. Then, those transformations were used as the initial conditions for an affine model-based registration with the *imregtform* function. Finally, this affine mapping transformation was applied to all the smFISH (RNA) images.

3. Use the protocol in (2) to register nuclei stain images obtained from the multimodal Raman microscope and the 1st round of images used for smFISH. Then, apply the transformation to the remaining 2nd and 3rd rounds.

4. Manually remove registration outliers in (3).

Fibroblast cells were mobile during the 2-class mixture experiment so that by the time Raman imaging finished, cells had moved far enough from their original position that the above semi-automated strategy could not be applied. Thus, we manually identified cells present in both nuclei stain images before and after the Raman imaging.

Hyperspectral Raman image processing

Each raw Raman spectrum has 1,340 channels. Of those channels, we extracted the fingerprint region (600-1800 cm^{-1}), which resulted in a total of 930 channels per spectrum. Thus, each FOV is a 100x100x930 hyperspectral image. The hyperspectral images were then preprocessed by a python script as follows:

1. Cosmic ray removal. Cosmic rays were detected by subtracting the median filtered spectra from the raw spectra, and any feature above 5 was classified as an outlier and replaced with the median value. The kernel window size for the median filter was 7.
2. Autofluorescence removal. The *baseline* function in *rampy* (https://github.com/charlesll/rampy), a python package for Raman spectral preprocessing, was used with the alternating least squares algorithm *'als'*.
3. Savitzky-Golay smoothing. The *scipy.signal.savgol_filter* function was used with window size 5 and polynomial order 3.
4. Averaging spectra at the single-cell level. Nuclei stain images were segmented using *NucleAIzer* (https://github.com/spreka/biomagdsb) and averaged pixel-level spectra that fall within each nucleus.
5. Spectra standardization. Spectra were standardized to a mean of 0 and a standard deviation of 1.

Inferring anchor smFISH from Raman spectra or brightfield z-stacks

For the two-class mixture and reprogramming experiment, we trained a decision tree-based non-linear regression, *Catboost*[15], to predict the 'on' or 'off' expression states for each anchor gene from Raman spectra. We used 80% of the data as training and the remaining 20% as test data. The early stopping parameter was set to 5.

For the brightfield z-stack to smFISH inference, we applied deep learning to the whole image level. We trained a modified U-net with skip connections and residual blocks to estimate the corresponding smFISH image[17]. Due to the small size of the available training dataset, we augmented the data by rotation and flipping. Furthermore, a subsample of each brightfield image

368 was taken due to memory constraints (50x50 pixel region). Training was carried out on an NVIDIA
369 Tesla P100 GPU, the number of epochs was 100, the learning rate was 0.01, and the batch size
370 was 400. For each smFISH prediction, we chose the epoch that gave the best validation score.

371 **Inferring expression profiles from Raman images**

372 To infer expression profiles from Raman images, we used Tangram[16]. Tangram enables the
373 alignment of spatial measurements of a small number of genes to scRNA-seq measurements. After
374 using Catboost to infer anchor expression levels from Raman profiles, we aligned the inferred
375 expression levels to scRNA-seq profiles using the *map_cells_to_space* function
376 (learning_rate=0.1, num_epochs=1000) on an Nvidia Tesla P100 GPU, followed by the
377 *project_genes* function in Tangram.
378 When comparing different pseudo-bulk transcriptome predictions with the real scRNA-seq data,
379 we first transferred labels of annotated scRNA-seq profiles to the ground truth smFISH profiles
380 using Tangram's label transfer function *project_cell_annotations*. Then, the average expression
381 profiles across cells of a cell type were calculated by referring to the transferred labels and
382 compared with those from the real scRNA-seq data[3].

383 **Dimensionality reduction, embedding and projection**

384 For dimension reduction and visualization of Raman and scRNA-seq profiles, we performed
385 forced layout embedding (FLE) using the *Pegasus* pipeline (https://github.com/klarman-cell-
386 observatory/pegasus). First, we performed principal component analysis on both Raman and
387 scRNA-seq profiles independently, calculated diffusion maps on the top 100 principal
388 components, and performed an approximated FLE graph using Deep Learning by *pegasus.net_fle*
389 with default parameters.

To project Raman profiles to a scRNA-seq embedding, we calculated a k-nearest neighbor graph (*k*-NN, *k*=15) on the scRNA-seq top 50 principal components with the cosine metric, and UMAP with the *scanpy.tl.umap* function in Scanpy[30] version 1.7.2 with default parameters. Then, the Raman predicted expression profiles were projected on to the scRNA-seq UMAP embedding by *scanpy.tl.ingest* using k-NN as the labeling method and default parameters.

Feature importance analysis

To evaluate the contributions of Raman spectral features to expression profile prediction, we used the *get_feature_importance* function in Catboost with default parameters. As the dimensions of Raman spectra were reduced by PCA prior to Catboost, feature importance scores were calculated for each principal component, and the weighted linear combination of the Raman PCA eigen vectors with feature scores as the weight were calculated to obtain the full spectrum.

Author contributions

KJKK, JS, TB and AR conceived the research and developed the methodology. JS, TB and AR funded and supervised research. KJKK, JS, JO performed reprogramming experiments. KJKK developed the multi-modal Raman microscope and control software with supervision from JWK and PS. KJKK, EG, and KZ performed smFISH. KJKK, SG, TJS, and TB developed the Raman spectral preprocessing and classification pipeline. KJKK developed the image registration pipeline, and performed Waddington-OT, Tangram and feature importance analysis. KJKK and BG performed U-net. KJKK, JS, and AR wrote the manuscript with input from all the authors.

409 **Competing interests statement**

410 AR is a co-founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas, and
411 was a scientific advisory board member of ThermoFisher Scientific, Syros Pharmaceuticals,
412 Neogene Therapeutics and Asimov until 31 July 2020. AR, TB, and SG are employees of
413 Genentech from August 1, 2020, respectively. A patent application has been filed related to this
414 work.

415 **Acknowledgements**

416 KJKK was supported by the Japan Society for the Promotion of Science Postdoctoral Fellowship
417 for Overseas Researchers, and the Naito Foundation Overseas Postdoctoral Fellowship. BG was
418 supported by the MathWorks Fellowship. JS was supported by the Helen Hay Whitney Foundation
419 and NIH Pathway to Independence Award (1K99HD096049-01, 5K99HD096049-02,
420 4R00HD096049-03), and funds from the Broad Institute of MIT and Harvard and Massachusetts
421 General Hospital. This research was funded by NIH National Institute of Biomedical Imaging and
422 Bioengineering, grant P41EB015871 (JWK, PS), NIH grant U19 MH114821 (TB), HubMap
423 UH3CA246632 (TB), and HHMI and the Klarman Cell Observatory (AR). AR was a Howard
424 Hughes Medical Institute Investigator when this work was initiated. We thank Eric Lander, Rudolf
425 Jaenisch, Doeke Hekstra, Joseph Kirschvink for their helpful discussion and insights. We thank
426 Leslie Gaffney for creating and editing figures.

427

428 **References**

429 1. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature*
430 **541**, 331–338 (2017).
431 2. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by
432 pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
433 3. Schiebinger, G. *et al.* Optimal-Transport Analysis of Single-Cell Gene Expression Identifies
434 Developmental Trajectories in Reprogramming. *Cell* **176**, 928-943.e22 (2019).
435 4. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
436 5. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient
437 cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
438 6. Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and
439 challenges. *Nat. Rev. Genet.* **21**, 410–427 (2020).
440 7. Wei, L. *et al.* Super-multiplex vibrational imaging. *Nature* **544**, 465–470 (2017).
441 8. Kobayashi-Kirschvink, K. J. *et al.* Linear Regression Links Transcriptomic Data and Cellular
442 Raman Spectra. *Cell Systems* vol. 7 104-117.e4 (2018).
443 9. Singh, S. P. *et al.* Label-free characterization of ultra violet-radiation-induced changes in skin
444 fibroblasts with Raman spectroscopy and quantitative phase microscopy. *Sci. Rep.* **7**, 10829 (2017).
445 10. Ichimura, T. *et al.* Visualizing cell state transition using Raman spectroscopy. *PLoS One* **9**, e84478
446 (2014).
447 11. Ho, C.-S. *et al.* Rapid identification of pathogenic bacteria using Raman spectroscopy and deep
448 learning. *Nat. Commun.* **10**, 4927 (2019).
449 12. Stadtfeld, M., Maherali, N., Borkent, M. & Hochedlinger, K. A reprogrammable mouse strain from
450 gene-targeted embryonic stem cells. *Nat. Methods* **7**, 53–55 (2010).
451 13. Choi, H. M. T. *et al.* Third-generation in situ hybridization chain reaction: multiplexed, quantitative,
452 sensitive, versatile, robust. *Development* **145**, (2018).

453 14. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and
454 Projection. *J. Open Source Softw.* **3**, 861 (2018).
455 15. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased
456 boosting with categorical features.
457 16. Biancalani, T. *et al.* Deep learning and alignment of spatially resolved single-cell transcriptomes
458 with Tangram. *Nat. Methods* **18**, 1352–1362 (2021).
459 17. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *2016 IEEE*
460 *Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2016).
461 doi:10.1109/cvpr.2016.90.
462 18. Germond, A., Panina, Y., Shiga, M., Niioka, H. & Watanabe, T. M. Following Embryonic Stem
463 Cells, Their Differentiated Progeny, and Cell-State Changes During iPS Reprogramming by Raman
464 Spectroscopy. *Anal. Chem.* **92**, 14915–14923 (2020).
465 19. Freudiger, C. W. *et al.* Label-free biomedical imaging with high sensitivity by stimulated Raman
466 scattering microscopy. *Science* **322**, 1857–1861 (2008).
467 20. Bai, Y. *et al.* Ultrafast chemical imaging by widefield photothermal sensing of infrared absorption.
468 *Sci Adv* **5**, eaav7127 (2019).
469 21. Tamamitsu, M., Toda, K., Horisaki, R. & Ideguchi, T. Quantitative phase imaging with molecular
470 vibrational sensitivity. *Opt. Lett.* **44**, 3729–3732 (2019).
471 22. Cleary, B., Cong, L., Cheung, A., Lander, E. S. & Regev, A. Efficient Generation of Transcriptomic
472 Profiles by Random Composite Measurements. *Cell* **171**, 1424-1436.e18 (2017).
473 23. Eng, C.-H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature*
474 **568**, 235–239 (2019).
475 24. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially
476 resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
477 25. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states.
478 *Science* (2018) doi:10.1126/science.aat5691.

26. Alon, S. *et al.* Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. *Science* **371**, (2021).

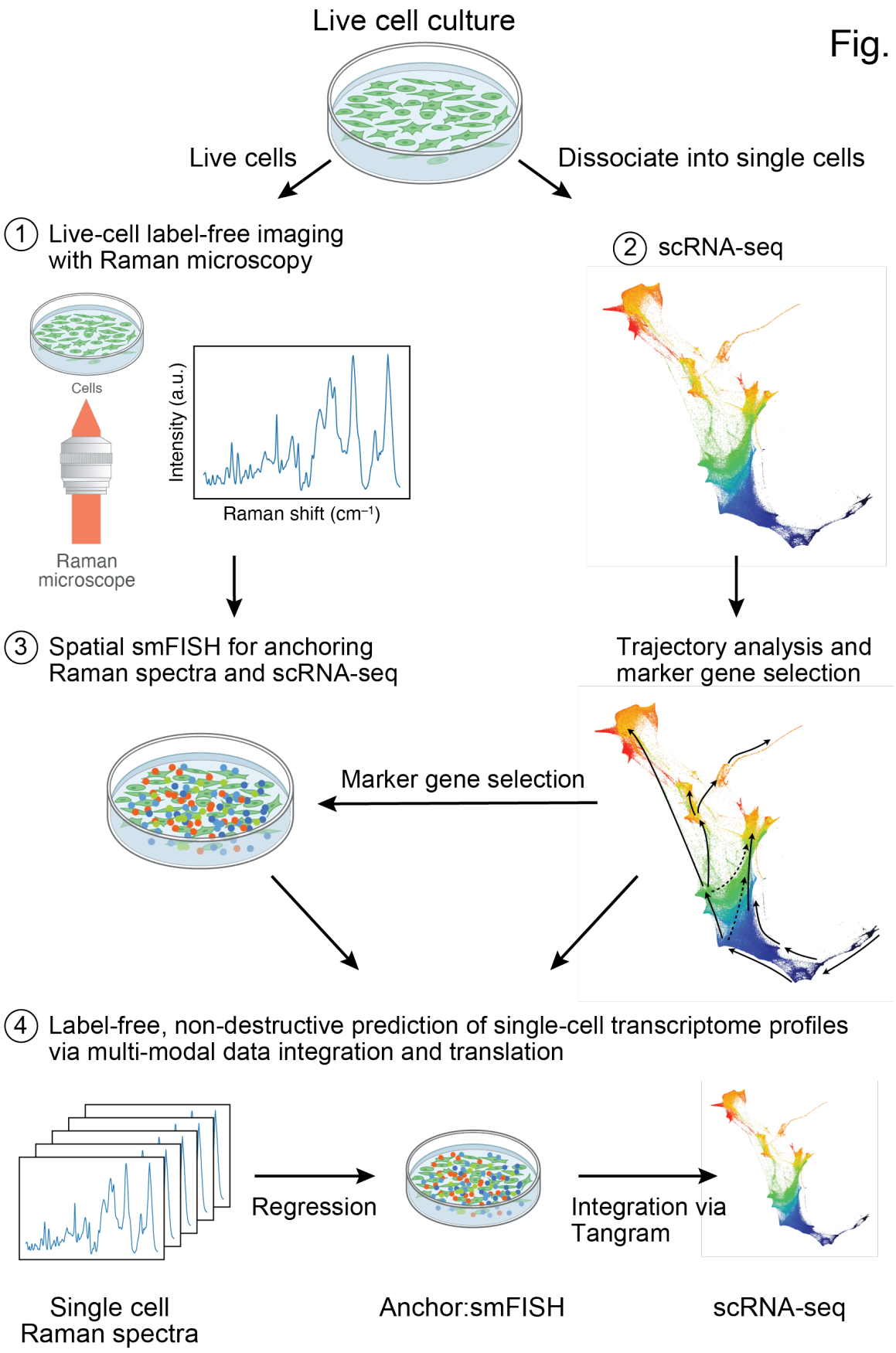
27. Yang, K. D. *et al.* Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nat. Commun.* **12**, 31 (2021).

28. Edelstein, A., Amodaj, N., Hoover, K., Vale, R. & Stuurman, N. Computer control of microscopes using µManager. *Curr. Protoc. Mol. Biol.* **Chapter 14**, Unit14.20 (2010).

29. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

30. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

Fig. 1


Live cell culture
Live cells
Dissociate into single cells
1 Live-cell label-free imaging with Raman microscopy
Cells
Raman microscope
Intensity (a.u.)
Raman shift (cm⁻¹)
2 scRNA-seq
3 Spatial smFISH for anchoring Raman spectra and scRNA-seq
Trajectory analysis and marker gene selection
Marker gene selection
4 Label-free, non-destructive prediction of single-cell transcriptome profiles via multi-modal data integration and translation
Regression
Integration via Tangram
Single cell Raman spectra
Anchor:smFISH
scRNA-seq

490 **Fig. 1 | Raman2RNA.** Live cells are cultured on gelatin-coated quartz glass-bottom plates (top) and
491 Raman spectra are then measured at each pixel (at spatial sub-cellular resolution) within an image frame
492 (1), followed by smFISH imaging in the same area (3). From parallel plates, cells are dissociated into a
493 single cell suspension and profiled by scRNA-seq (2). scRNA-seq profiles are used to select 9 marker
494 genes for 5 major cell clusters, and those are measured with spatial smFISH (3). Lastly, a regression
495 model is trained (4) to predict anchor smFISH profiles from Raman spectra, followed by integration via
496 Tangram[16] to predict whole single-cell transcriptome profiles from smFISH profiles.

Fig. 2

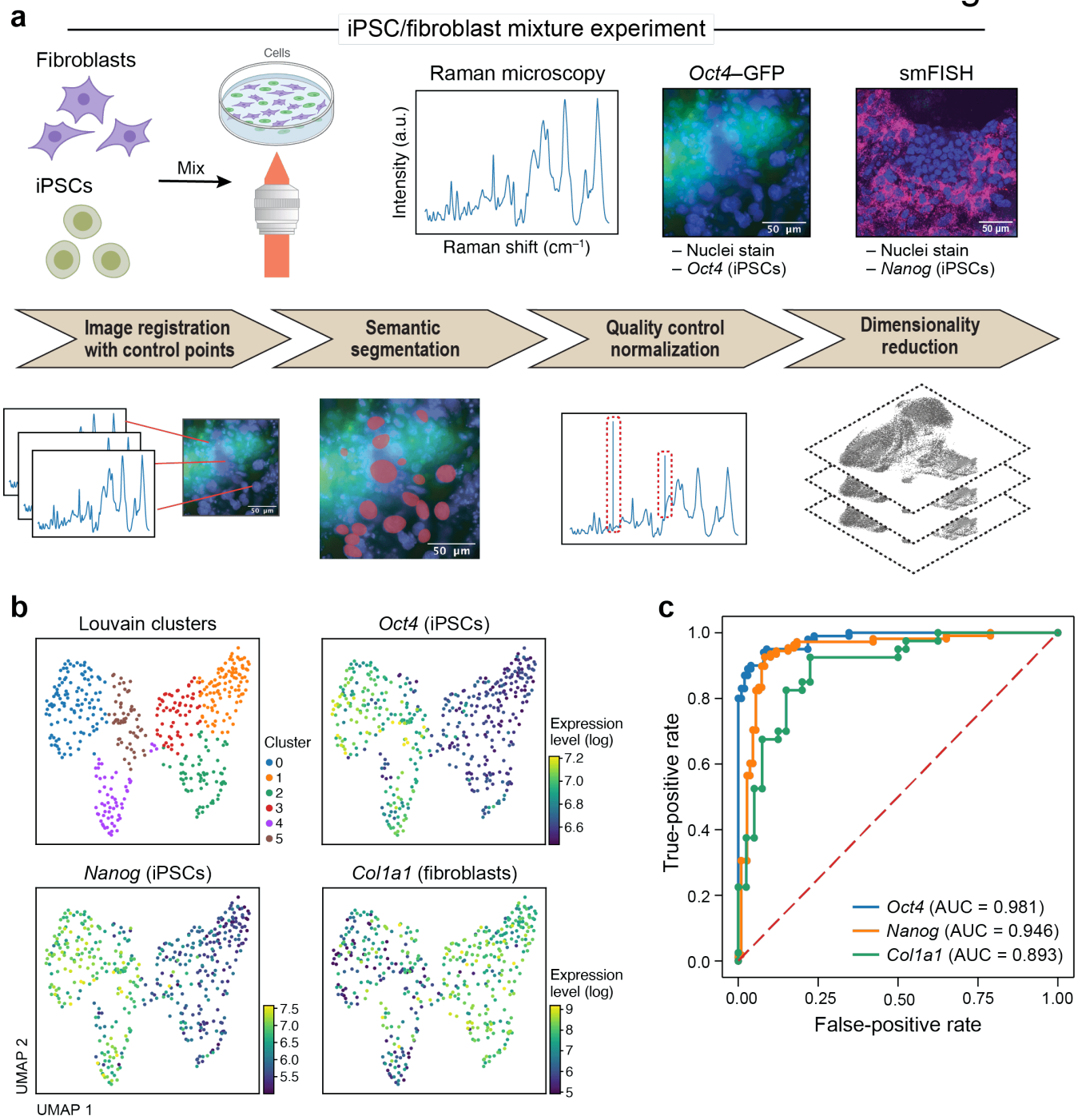


Fig. 2 | Raman2RNA accurately distinguishes cell types and predicts binary expression of marker genes in a mixture of mouse fibroblasts and iPSCs. a. Overview. Top: Experimental procedures. Mouse fibroblasts and iPSCs were mixed 1:1 and plated on glass-bottom plates, followed by Raman imaging of live cells, nuclei staining and measurement of endogenous *Oct4*-GFP (iPSC marker) reporter) by fluorescence imaging, and cell fixation and processing for smFISH with DAPI and probes for *Nanog*

503 (iPSCs, magenta) and *Col1a1* (fibroblasts). Bottom: Preprocessing and analysis. From left: Image
504 registration with control points (**Methods**), was followed by semantic cell segmentation, outlier
505 removal/normalization and dimensionality reduction. **b.** Raman2RNA distinguishes cell states from
506 Raman spectra. 2D UMAP embedding of single-cell Raman spectra (dots) colored by Louvain clustering
507 labels (top left) or smFISH measured expression of *Oct4* (top right), *Nanog* (bottom left) and *Col1a1*
508 (bottom right). **c.** Raman2RNA accurately predicts binary (on/off) expression of marker genes. Receiver
509 operating characteristic (ROC) plots and area under the curve (AUC) obtained by classifying the 'on' and
510 'off' states of *Oct4* (blue), *Nanog* (orange) and *Col1a1* (green).

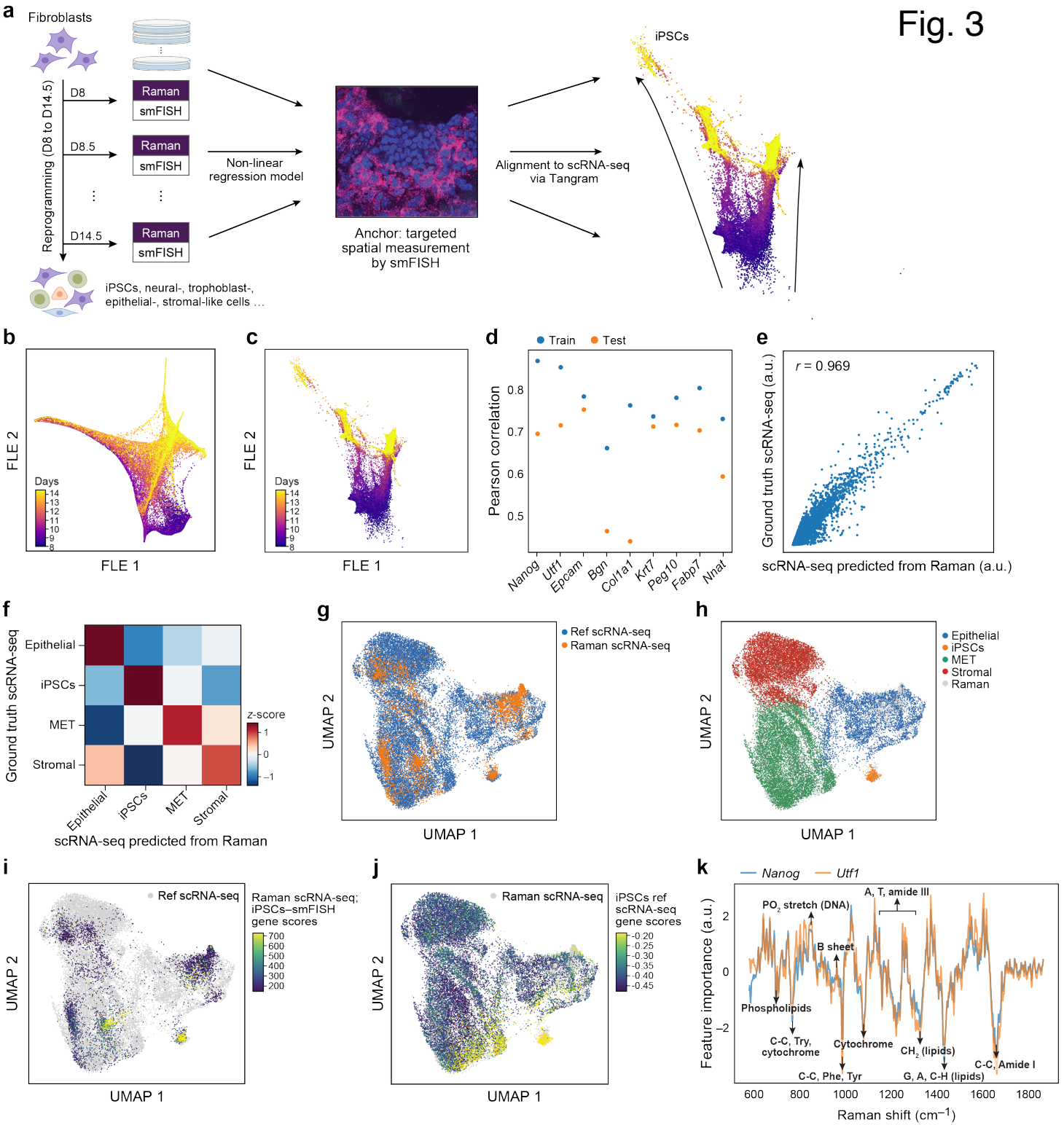
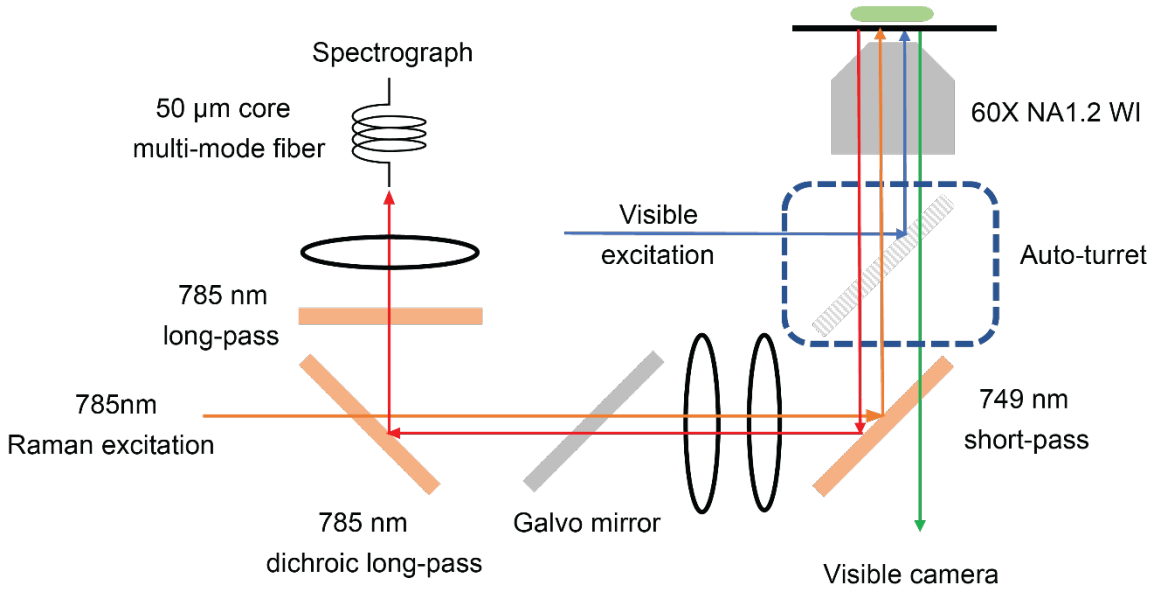


Fig. 3 | Raman2RNA predicts single-cell RNA profiles across cell types during reprogramming of mouse fibroblasts to iPSCs. a. Approach overview. From left: Mouse fibroblasts were reprogrammed

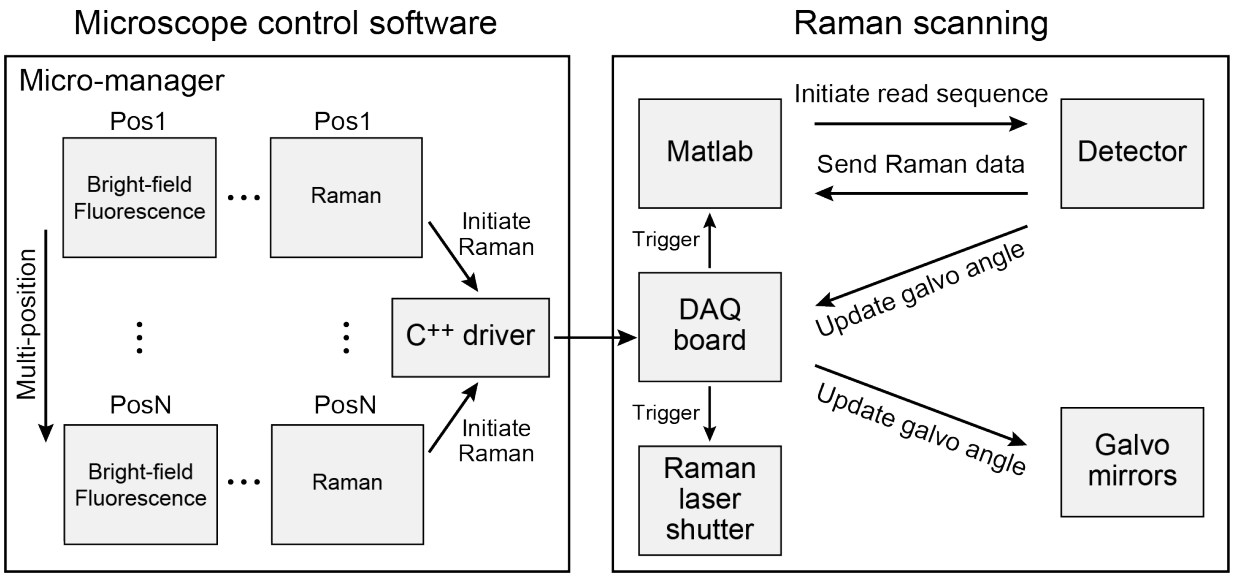
514 into induced pluripotent stem cells (iPSCs) over the course of 14.5 days ('D'), and, at half-day intervals
515 from days 8 to 14.5, spatial Raman spectra, smFISH for nine anchor genes, and nuclei stain by
516 fluorescence imaging were measured for each plate. Machine learning and multi-modal data integration
517 methods (Catboost and Tangram) were used to predict single-cell RNA-seq profiles from Raman spectra
518 using smFISH as anchor. **b,c.** Low dimensionality embedding of single-cell Raman spectra captures
519 progress in reprogramming. Force-directed layout embedding (FLE) of Raman spectra (b, dots) or
520 scRNA-seq (c, dots) colored by days of measurement (colorbar). **d.** Correct prediction of smFISH anchors
521 from Raman spectra. Pearson correlation coefficient (*y* axis) between measured (smFISH) and Raman-
522 predicted levels for each smFISH anchor (*x* axis) in leave-one-out cross-validation where 8 out of 9
523 smFISH anchor genes were used for training, and the left-out gene was predicted. **e.f.** Raman2RNA
524 accurately predicts pseudo-bulk expression profiles of major cell types. **e.** scRNA-seq measured (y axis)
525 and R2R-predicted (x axis) for each gene (dot) in pseudo-bulk RNA profiles averaged across iPSCs. **f.**
526 Pair-wise correlation (color bar) between Raman-predicted and scRNA-seq measured pseudo-bulk
527 profiles in each cell types (rows, columns). **g-j.** Co-embedding highlights agreement between real and
528 R2R inferred single cell profiles. UMAP co-embedding of Raman predicted RNA profiles and measured
529 scRNA-seq profiles (dots) colored by data source (**g,** Raman predicted in orange; measured scRNA-seq in
530 blue), cell type annotations (**h**) or by iPSC gene signature scores (calculated by averaging expression of
531 genes *Nanog* and *Utf1*, and subtracting the average of a randomly selected set of reference genes;
532 **Methods**) of Raman-predicted profiles (**i**) or of real scRNA-seq (**j**). **k.** Feature importance scores of
533 Raman spectra in predicting expression profiles. Feature scores for iPSC related marker genes (y axis)
534 along the Raman spectrum (x axis). Known Raman peaks[18] were annotated.
535

Supp. Fig. 1



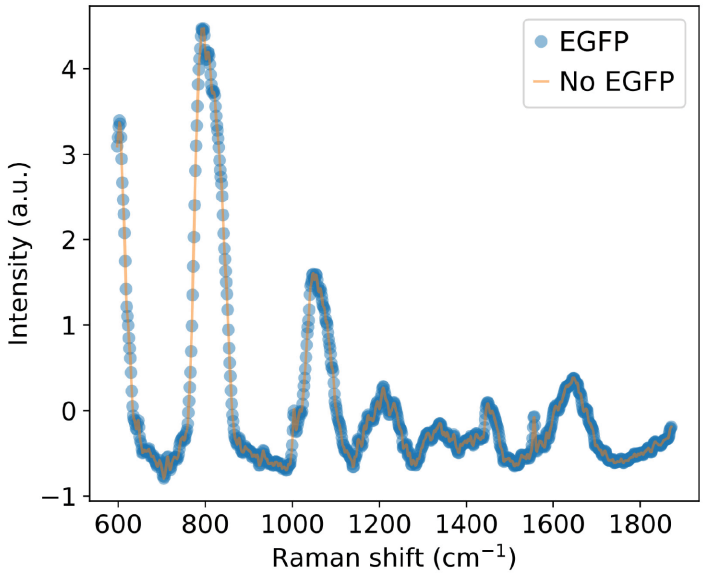
Extended Data Fig. 1 | A multi-modal Raman microscope capable of fluorescence imaging and Raman microscopy. Schematic of a Raman microscope integrated with a wide-field fluorescence microscope for simultaneous detection of nuclei staining, bright field, fluorescence channels, and Raman images.

Supp. Fig. 2



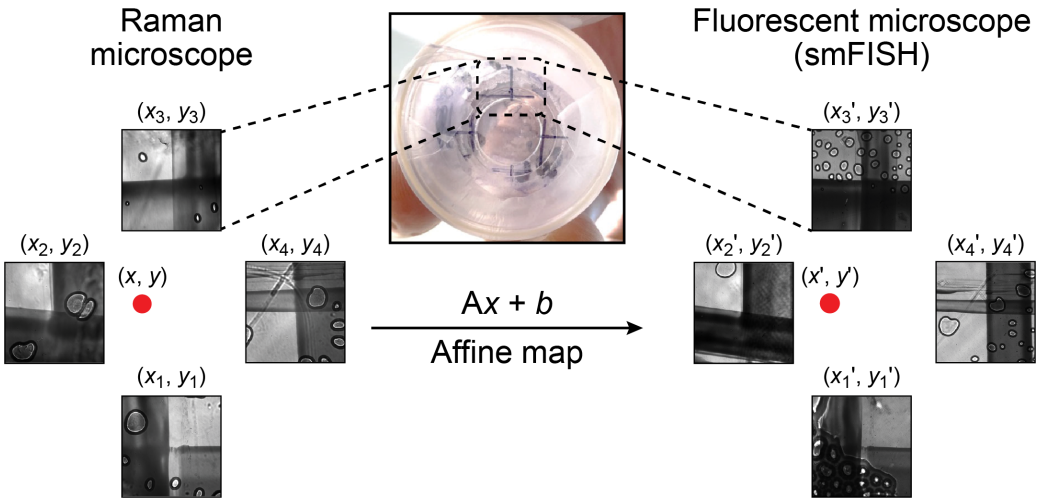
Extended Data Fig. 2 | Overview of high-throughput Raman imaging software used in the study. A general-purpose microscope control software Micro-manager and a custom MATLAB script were combined to enable automated multi-modal measurements. Under Micro-manager, a Raman channel was registered as a 'dummy' channel along with brightfield and fluorescence channels. Micro-manager was responsible for changing the field of view (FOV) and imaging modality. During the Raman sequence, Micro-manager communicated with a digital acquisition (DAQ) board, through which a transistor-to-transistor logic (TTL) signal was generated to initiate the scanning sequence. Upon detection of the TTL signal, the MATLAB script controlled the Raman detector, laser shutter, and updated the galvo mirror angles through the DAQ board.

Supp. Fig. 3



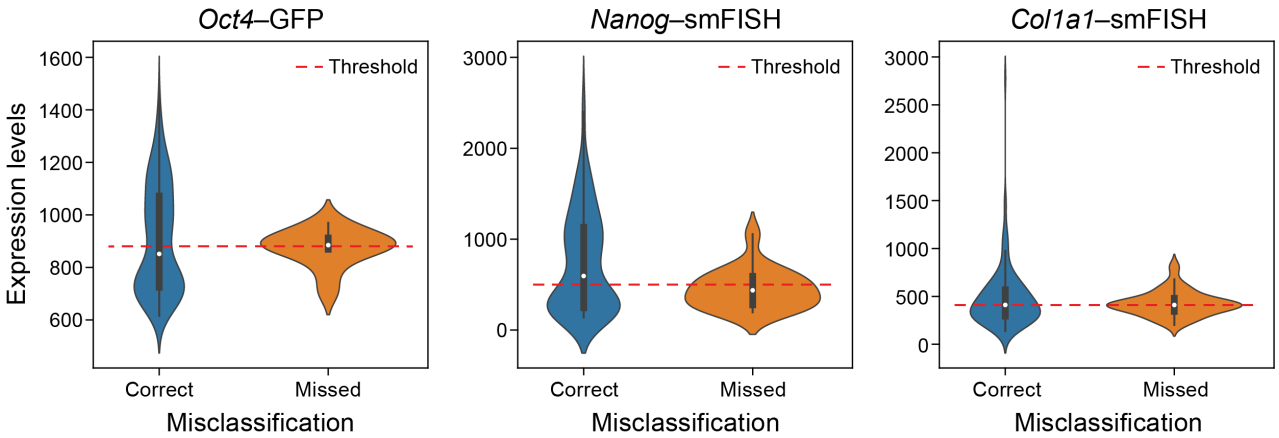
Extended Data Fig. 3 | GFP does not interfere in Raman spectra measurement. Raman spectra of culture media with (blue) and without (orange) GFP at physiological concentration.

Supp. Fig. 4



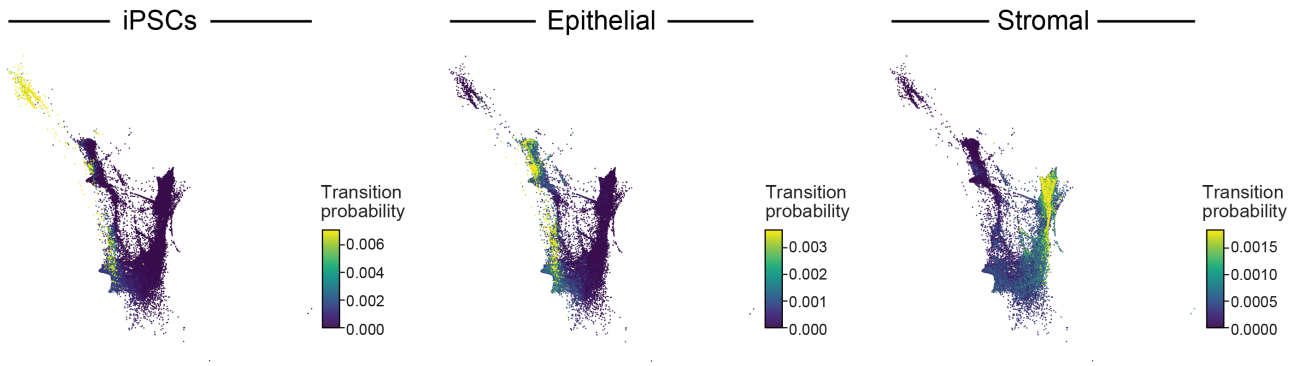
Extended Data Fig. 4 | Image registration between the Raman and smFISH microscope using control points. Control points were inscribed under petri dishes with permanent markers and the coordinates were measured prior to any data acquisition. After Raman measurement and smFISH processing, samples were placed back to the microscope and control point coordinates were remeasured. Then, affine mapping was used to update the FOV coordinates to locate the exact same cells.

Supp. Fig. 5



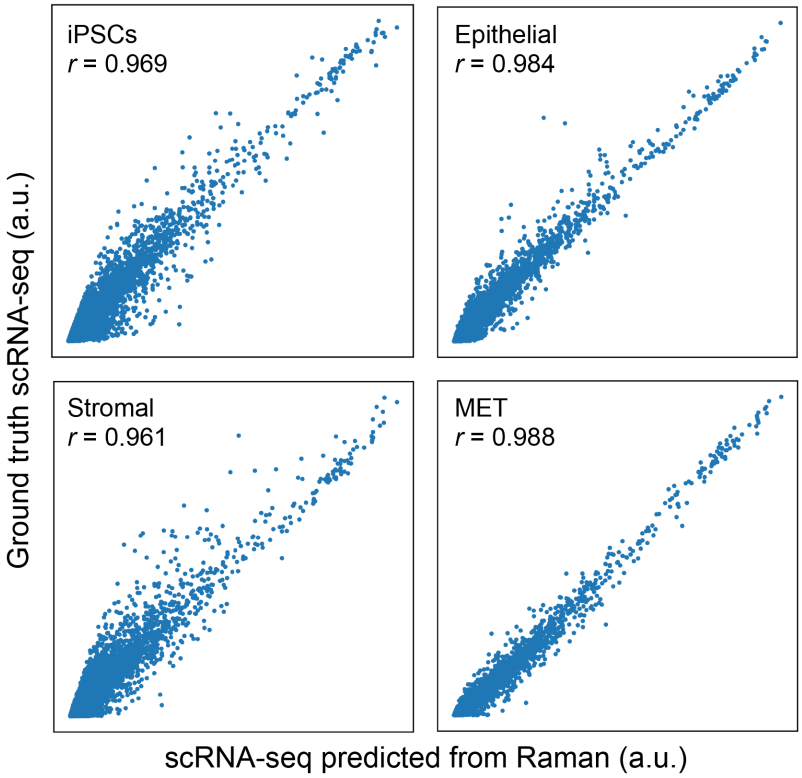
Extended Data Fig. 5 | Misclassification of genes in the cell mixture classification experiment occurs when the ground truth smFISH is near the expression threshold. Distribution of measured smFISH expression level (y axis) for cells correctly (blue) or incorrectly (orange) classified by their Raman spectra for the expression of that gene. Horizontal line: an example threshold used for the logistic regression classifier.

632
633 Supp. Fig. 6
634
635
636



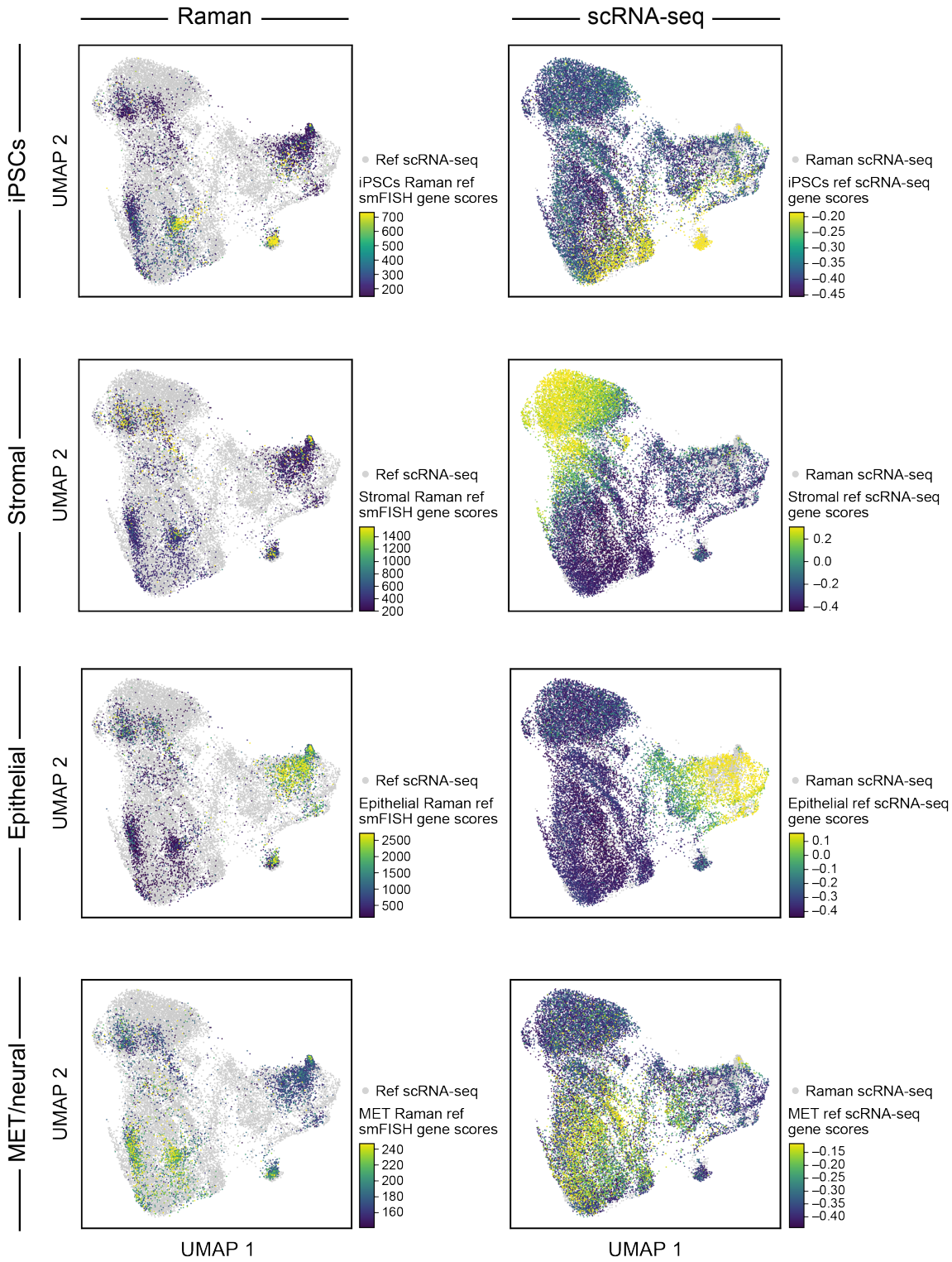
637
638
639 **Extended Data Fig. 6 | Cell transition probabilities inferred by Waddington-OT from scRNA-seq**
640 **during reprogramming.** Force-directed layout embedding (FLE) of scRNA-seq profiles (dots) from
641 days 8 to 14.5 of reprogramming (dots) colored by the transition probability of each cell as inferred by
642 Waddington-OT to be an ancestor of iPSCs (left), epithelial cells (middle) or stromal cells (right) at day
643 14.5.
644
645

Supp. Fig. 7



Extended Data Fig. 7 | Raman-predicted and scRNA-seq measured pseudo-bulk profiles are well correlated across cell types. ScRNA-seq measured (y axis) and R2R-predicted (x axis) expression for each gene (dot) in pseudo-bulk RNA profiles averaged across cells labeled as iPSC (top left), epithelial (top right), stromal (bottom left) and MET (bottom right). Pearson's r is denoted at the top left corner.

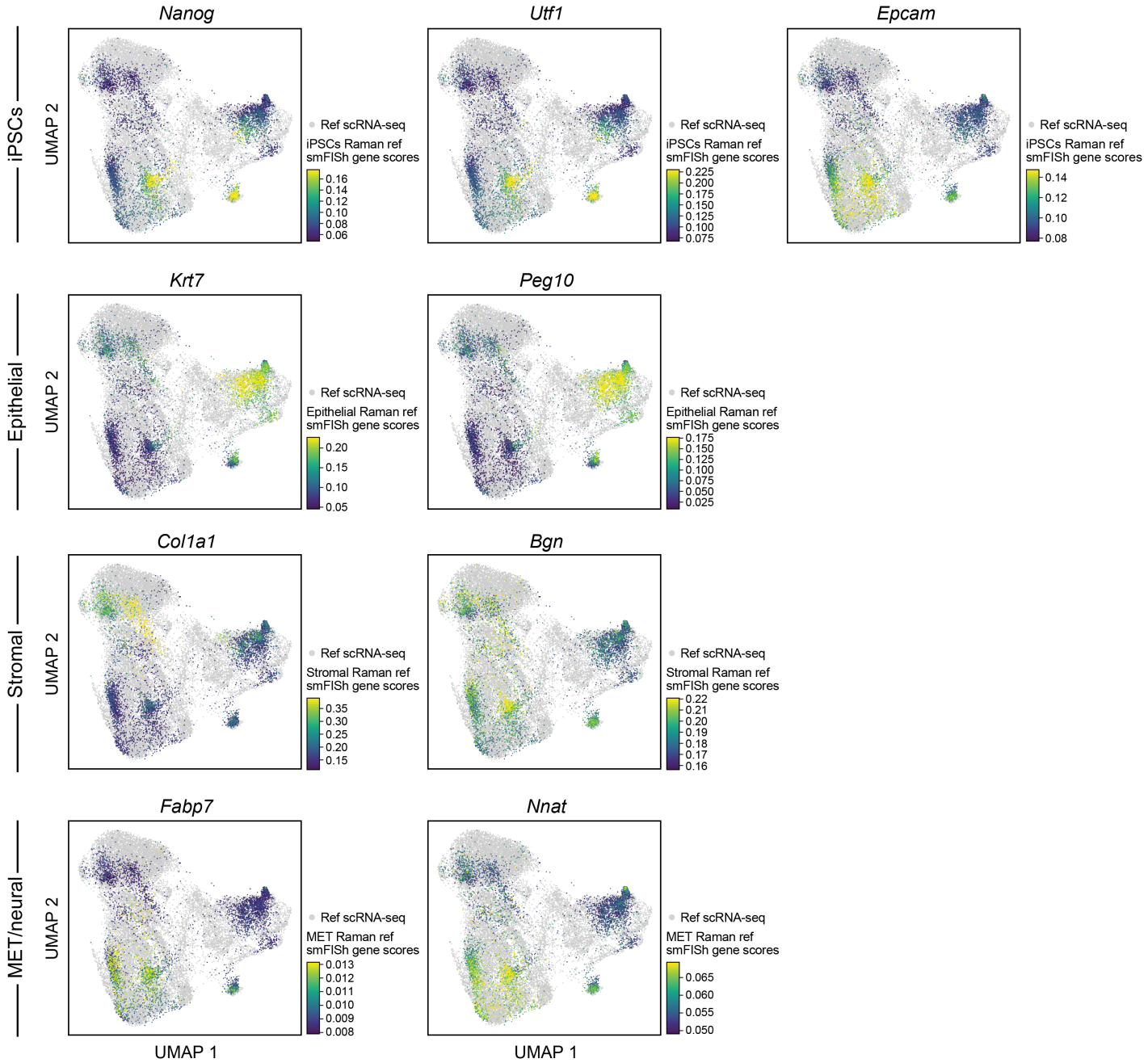
Supp. Fig. 8



Extended Data Fig. 8 | Measured and Raman-predicted single cell profiles co-embed well as reflected by gene scores for each cell type. UMAP co-embedding of Raman predicted RNA profiles and measured scRNA-seq profiles (dots) colored by scores of marker gene for different cell types (rows)

678 determined by smFISH measurements (left, for cells with Raman-predicted profiles) or real scRNA-seq
679 measurements (right, for cells with scRNA-seq profiles).

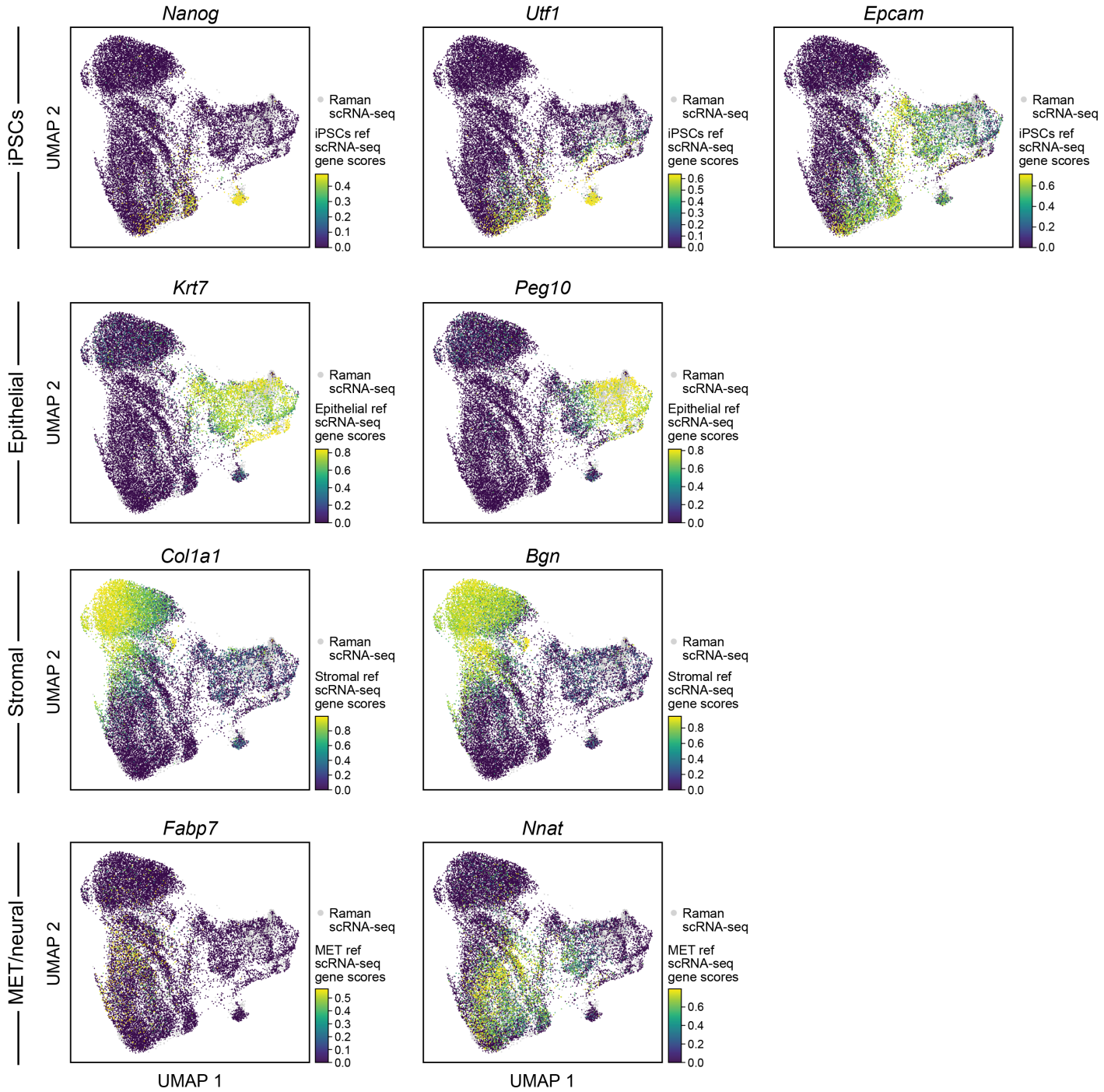
Supp. Fig. 9



Extended Data Fig. 9 | Measured and Raman-predicted single cell profiles co-embed well as reflected by smFISH measurement of Raman cells. UMAP co-embedding of Raman predicted RNA profiles and measured scRNA-seq profiles (dots) where the Raman cells are colored by smFISH measurement of each of nine anchor genes.

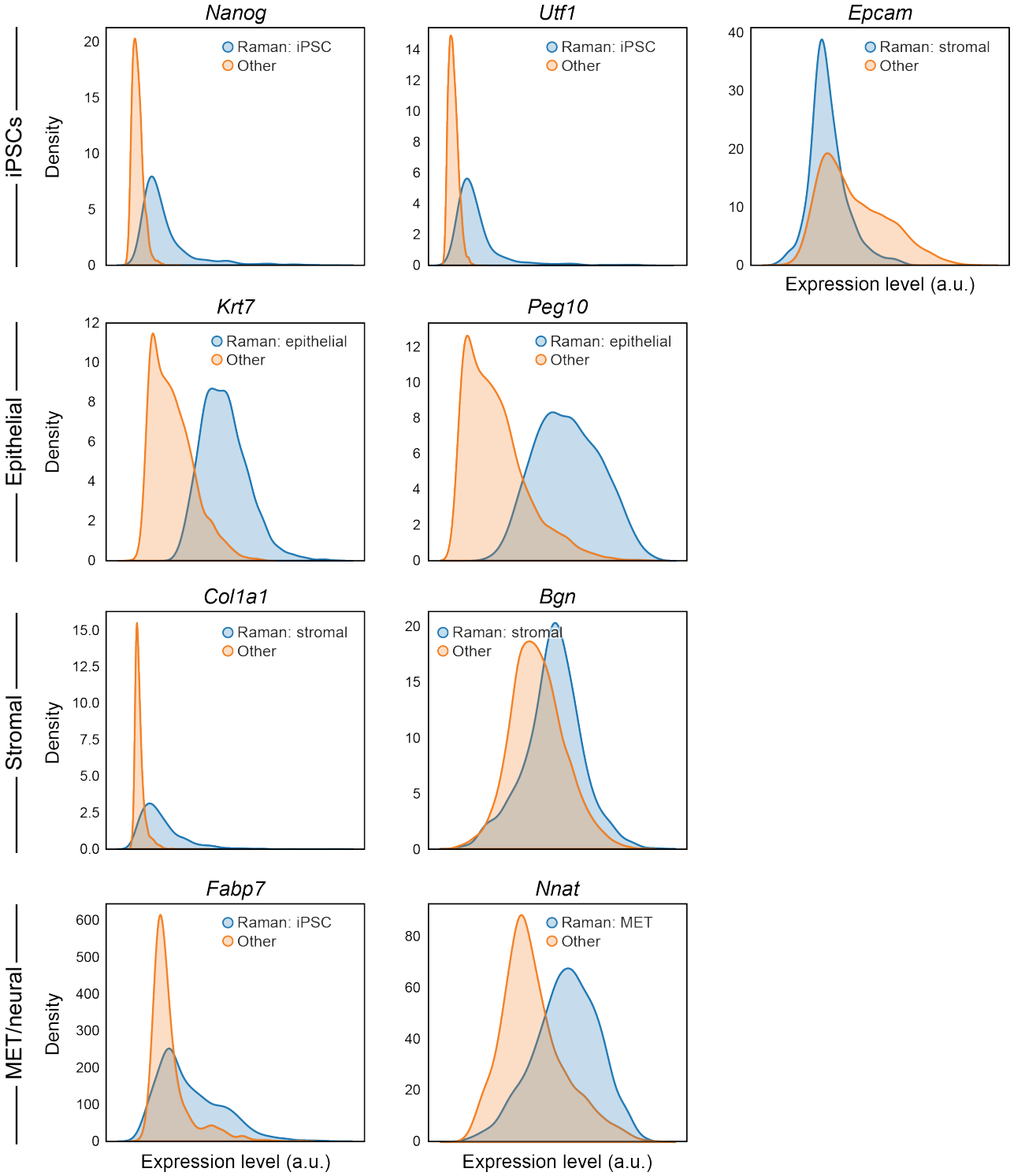
688

Supp. Fig. 10

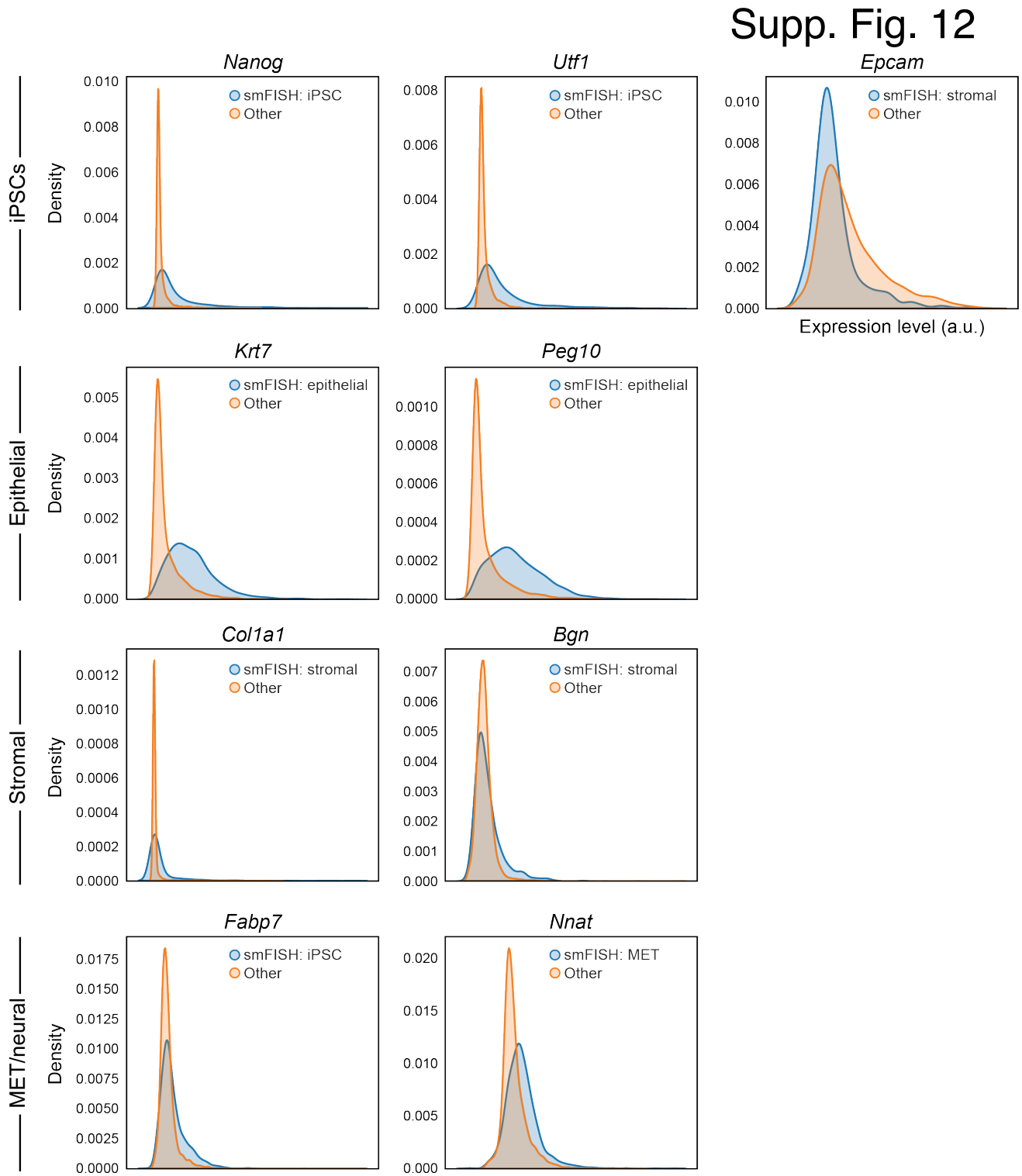


689 **Extended Data Fig. 10 | Measured and Raman-predicted single cell profiles co-embed well as**
690 **reflected by scRNA-seq based expression of nine anchor genes.** UMAP co-embedding of Raman
691 predicted RNA profiles and measured scRNA-Seq profiles (dots) where the scRNA-seq profiled cells are
692 colored by scRNA-seq measured expression of each of nine anchor genes.

Supp. Fig. 11



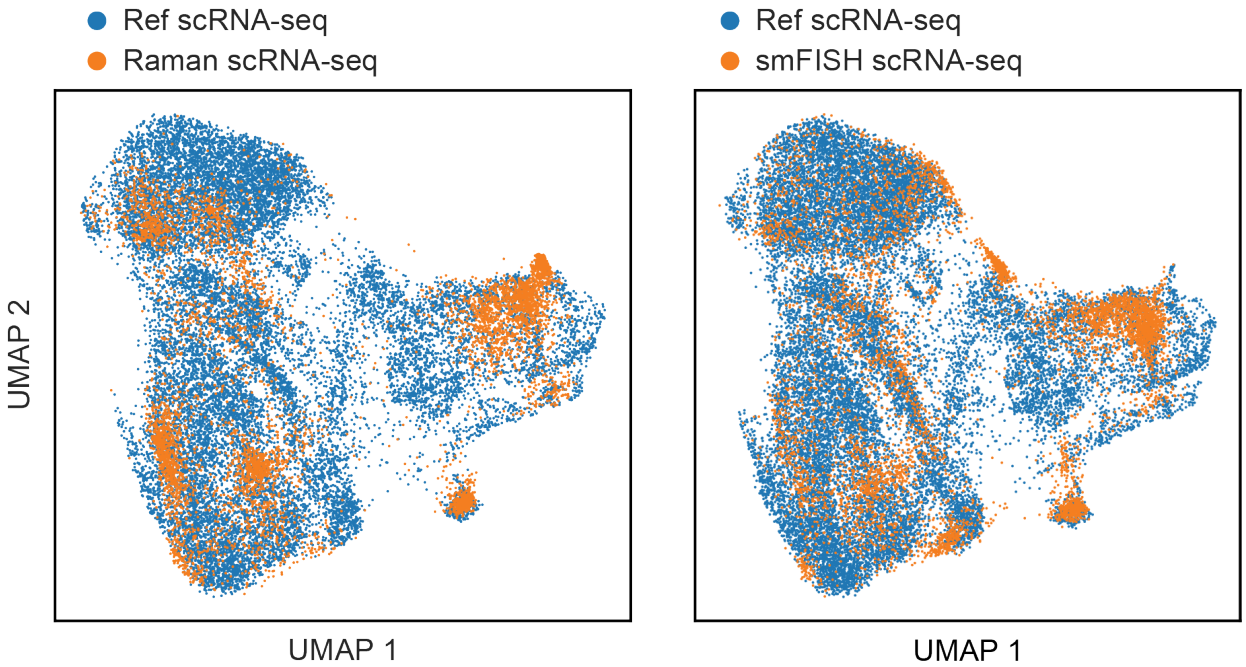
Extended Data Fig. 11 | Distributions of expression of marker genes based on R2R-predicted profiles. Distributions (density plots) of the predicted expression in Raman2RNA inferred profiles for each marker gene (panel) in its expected corresponding cell type (blue, based on the predicted expression profiles) and all other cells (orange).



697

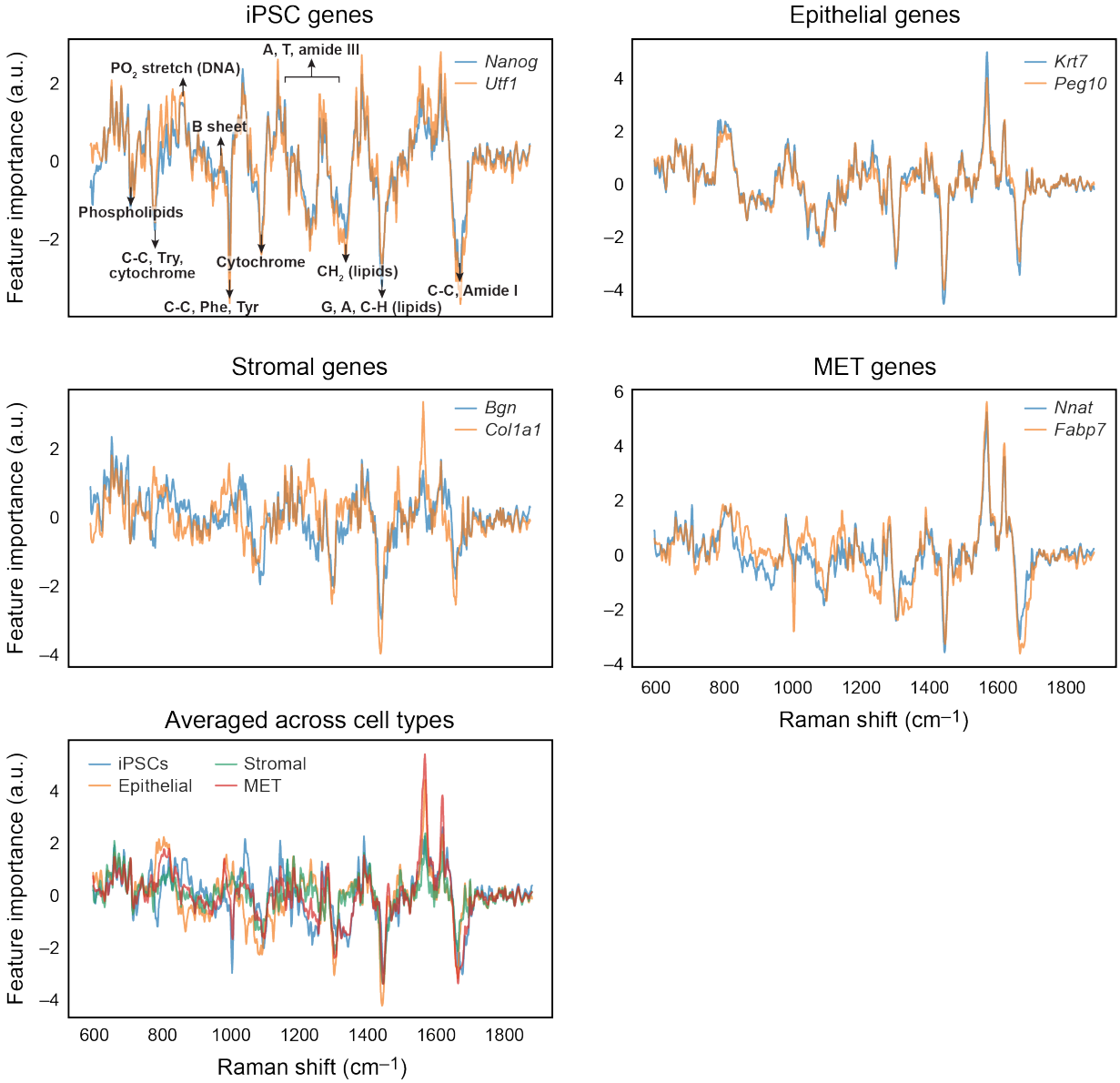
698 **Extended Data Fig. 12 | Distributions of expression of marker genes based on real smFISH profiles.**
699 Distributions (density plots) of the real smFISH profiles for each marker gene (panel) in its expected
700 corresponding cell type (blue, based on the R2R *predicted* expression profiles) and all other cells
701 (orange).

Supp. Fig. 13



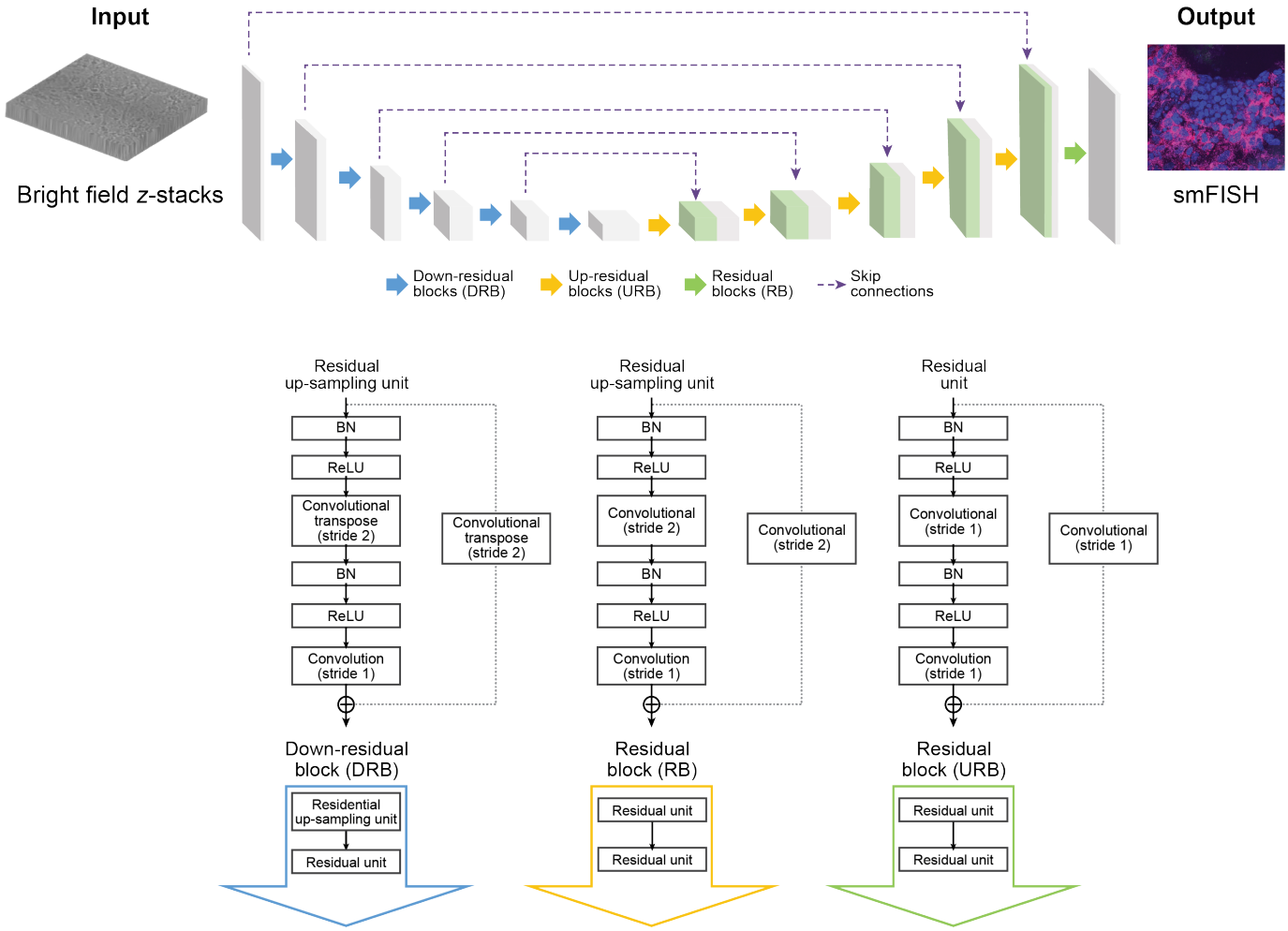
Extended Data Fig. 13 | RNA profiles predicted directly from 9 anchor smFISH measurements lead to reduced variance compared to scRNA-seq. UMAP co-embedding of cells from scRNA-seq (blue) and Raman (orange) experiments, with the latter based on either the Raman-predicted RNA profiles (left) or only smFISH-predicted RNA profiles (right).

Supp. Fig. 14



Extended Data Fig. 14 | Raman spectral feature importance scores for each smFISH anchor gene and its average across all genes for a cell type. Feature importance scores (y axis) for marker genes of each cell type (top two rows), and for all cell types (bottom row), along the Raman spectrum (x axis). Known signals[18] are annotated in the top left panel (identical to **Fig. 3k**).

Supp. Fig. 15



Extended Data Fig. 15 | Neural network-based prediction of smFISH using brightfield z-stacks.

References

1. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).
2. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
3. Schiebinger, G. *et al.* Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell* **176**, 928-943.e22 (2019).
4. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
5. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
6. Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* **21**, 410–427 (2020).
7. Wei, L. *et al.* Super-multiplex vibrational imaging. *Nature* **544**, 465–470 (2017).
8. Kobayashi-Kirschvink, K. J. *et al.* Linear Regression Links Transcriptomic Data and Cellular Raman Spectra. *Cell Systems* vol. 7 104-117.e4 (2018).
9. Singh, S. P. *et al.* Label-free characterization of ultra violet-radiation-induced changes in skin fibroblasts with Raman spectroscopy and quantitative phase microscopy. *Sci. Rep.* **7**, 10829 (2017).
10. Ichimura, T. *et al.* Visualizing cell state transition using Raman spectroscopy. *PLoS One* **9**, e84478 (2014).
11. Ho, C.-S. *et al.* Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. *Nat. Commun.* **10**, 4927 (2019).
12. Stadtfeld, M., Maherali, N., Borkent, M. & Hochedlinger, K. A reprogrammable mouse strain from gene-targeted embryonic stem cells. *Nat. Methods* **7**, 53–55 (2010).
13. Choi, H. M. T. *et al.* Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. *Development* **145**, (2018).

747 14. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and
748 Projection. *J. Open Source Softw.* **3**, 861 (2018).
749 15. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased
750 boosting with categorical features.
751 16. Biancalani, T. *et al.* Deep learning and alignment of spatially resolved single-cell transcriptomes
752 with Tangram. *Nat. Methods* **18**, 1352–1362 (2021).
753 17. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *2016 IEEE*
754 *Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2016).
755 doi:10.1109/cvpr.2016.90.
756 18. Germond, A., Panina, Y., Shiga, M., Niioka, H. & Watanabe, T. M. Following Embryonic Stem
757 Cells, Their Differentiated Progeny, and Cell-State Changes During iPS Reprogramming by Raman
758 Spectroscopy. *Anal. Chem.* **92**, 14915–14923 (2020).
759 19. Freudiger, C. W. *et al.* Label-free biomedical imaging with high sensitivity by stimulated Raman
760 scattering microscopy. *Science* **322**, 1857–1861 (2008).
761 20. Bai, Y. *et al.* Ultrafast chemical imaging by widefield photothermal sensing of infrared absorption.
762 *Sci Adv* **5**, eaav7127 (2019).
763 21. Tamamitsu, M., Toda, K., Horisaki, R. & Ideguchi, T. Quantitative phase imaging with molecular
764 vibrational sensitivity. *Opt. Lett.* **44**, 3729–3732 (2019).
765 22. Cleary, B., Cong, L., Cheung, A., Lander, E. S. & Regev, A. Efficient Generation of Transcriptomic
766 Profiles by Random Composite Measurements. *Cell* **171**, 1424-1436.e18 (2017).
767 23. Eng, C.-H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature*
768 **568**, 235–239 (2019).
769 24. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially
770 resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
771 25. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states.
772 *Science* (2018) doi:10.1126/science.aat5691.

773 26. Alon, S. *et al.* Expansion sequencing: Spatially precise in situ transcriptomics in intact biological
774 systems. *Science* **371**, (2021).
775 27. Yang, K. D. *et al.* Multi-domain translation between single-cell imaging and sequencing data using
776 autoencoders. *Nat. Commun.* **12**, 31 (2021).
777 28. Edelstein, A., Amodaj, N., Hoover, K., Vale, R. & Stuurman, N. Computer control of microscopes
778 using µManager. *Curr. Protoc. Mol. Biol.* **Chapter 14**, Unit14.20 (2010).
779 29. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
780 30. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data
781 analysis. *Genome Biol.* **19**, 15 (2018).
782