High-definition spatial transcriptomics for in situ tissue profiling

Sanja Vickovic1,2,*, Gökcen Eraslan1,2,12, Fredrik Salmén1,2, Johanna Klughammer1,12, Linnea Stenbeck2,12, Denis Schapiro1,3, Tarmo Äijö4, Richard Bonneau5,6, Ludvig Bergensträhle2, José Fernández Navarro2, Joshua Gould1, Gabriel K. Griffin1,6, Åke Borg7, Mostafa Ronaghi8, Jonas Frisén9, Joakim Lundeberg2,10, Aviv Regev1,11 and Patrik L. Ståhl2

Spatial and molecular characteristics determine tissue function, yet high-resolution methods to capture both concurrently are lacking. Here, we developed high-definition spatial transcriptomics, which captures RNA from histological tissue sections on a dense, spatially barcoded bead array. Each experiment recovers several hundred thousand transcript-coupled spatial barcodes at 2-μm resolution, as demonstrated in mouse brain and primary breast cancer. This opens the way to high-resolution spatial analysis of cells and tissues.

Charting cells’ spatial organization and molecular features is essential to understand how they interact in both normal and diseased tissues1,2. Massively parallel single-cell RNA sequencing (scRNA-Seq)3,4 can profile hundreds of thousands of dissociated individual cells, but does not retain their spatial position, and can introduce biases in cell recovery5. Conversely, spatial profiling captures detailed positional information in intact tissue, but current methods require pre-selected markers, rely on nonstandard instrumentation6–11 or have limited spatial resolution, scalability or applicability. In particular, spatial transcriptomics12 (ST) is a spatially barcoded RNA-Seq method providing transcriptome-wide coverage in many systems, but at a resolution of 100 μm (3–30 cells).

To bridge this gap, we developed high-definition spatial transcriptomics (HDST, Fig. 1a) and demonstrate its application to large tissue areas in the mouse brain and human tumors in situ. In HDST, we deposit barcoded poly(d)T oligonucleotides into 2-μm wells and then decode their positions by a sequential hybridization and error-correcting strategy12,13. After a frozen tissue section is placed on the decoded slide, stained and imaged, RNA is captured and then profiled by RNA-Seq.

To produce a high-resolution, high-density bead array, we generated 2,893,865 individual barcoded beads with a split-and-pool approach (Supplementary Fig. 1a), randomly placed them into a hexagonal array of >1.4 million 2-μm wells and then decoded each bead’s location (Fig. 1a)14,15 with several hybridization rounds. Each round hybridizes a set of complementary and labeled decoder oligonucleotides (decoders) (Methods), records fluorescence across the entire slide area and then strips the decoders. The process is repeated log²N times (14 times for the array presented here), where N is the number of sequences to be decoded with three labels used (Supplementary Fig. 1b). In this way, each bead and barcode receives a unique spatial color address11 creating a HDST array in ~3 h total processing time.

To test HDST, we first profiled the main olfactory bulb (MOB) of the mouse brain, whose neurons have traditionally been defined by the presence of neuronal cell bodies across morphological layers16. We assessed whether HDST molecular data can be related to layers and other histological features. We analyzed three replicate sections by HDST and tested its performance in two key tasks: (1) generating high-resolution spatial expression patterns of individual genes, and (2) detecting cell types and assigning them to correct, high-resolution positions.

We confirmed that RNA capture was specific and in overall agreement with bulk RNA-Seq controls, despite the relatively low number of transcripts captured per spatial barcode. We first accounted for barcode redundancy (‘clashing’), decoding efficiency and stringent barcode demultiplexing (Supplementary Fig. 2a and Supplementary Table 1). Next, we observed that at saturating sequencing depth (Supplementary Fig. 2b), 85.6 ± 3.3% of all genes detected were located within the area physically covered by the tissue specimen (without using any lower cutoffs), with almost 160,000 barcodes generating spatially mapped transcripts per assay (n = 3 sections, Supplementary Fig. 2c). Although there were few unique molecular identifiers (UMIs) per barcode location (7.1 ± 6.0 (mean ± s.d.), n = 3 sections), a very distinct spatial in situ profile followed the tissue boundary (Supplementary Fig. 2d,e), suggesting detection specificity. Moreover, a combined ‘bulk’ expression profile of each HDST dataset correlated significantly with published MOB bulk RNA-Seq (Supplementary Fig. 3a; Spearman’s ρ = 0.69 ± 0.02; mean ± s.d.) and across the three replicate experiments (Spearman’s ρ = 0.82 ± 0.06; mean ± s.d.). Most detected genes agreed between the bulk and HDST datasets (Supplementary Fig. 3b). Finally, comparing the HDST capture per bead to that observed from smFISH for three genes (Penk, Slc17a7 and Fabp7), we estimated HDST capture efficiency at 1.3% per bead (Methods and Supplementary Table 2).

1Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. 2Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, Stockholm, Sweden. 3Laboratory of Systems Pharmacology, Harvard Medical School, Boston, MA, USA. 4Center for Computational Biology, Flatiron Institute, New York, NY, USA. 5Center for Data Science, New York University, New York, NY, USA. 6Department of Pathology, Brigham and Women’s Hospital, Boston, MA, USA. 7Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, Lund, Sweden. 8Illumina, Inc., San Diego, CA, USA. 9Department of Cell and Molecular Biology, Karolinska Institute, Stockholm, Sweden. 10Department of Bioengineering, Stanford University, Stanford, CA, USA. 11Howard Hughes Medical Institute and Koch Institute for Integrative Cancer Research, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. 12These authors contributed equally: Gökcen Eraslan, Fredrik Salmén, Johanna Klughammer, Linnea Stenbeck. *e-mail: vickovic@broadinstitute.org; joakim.lundeberg@scilifelab.se
Layer-enriched upregulated DE genes performed a two-sided exact test, Bonferroni adjusted, data (63.5 ± 15.9% mean ± s.d., n = 3 sections) M/T, mitral layer; IPL, internal plexiform layer; GCL-E, GCL-I and GCL-D, granular layers; GL, glomerular layer; EPL, external plexiform layer.

Fig. 1 | HDST. a, HDST workflow. b, Labeling of morphological layers. HDST H&E image of a MOB and matching HDST (x, y) barcodes annotated into nine morphological areas. c, Layer-specific DE patterns in HDST. Shown is the summed normalized expression of positively enriched signature genes significantly (FDR < 0.01, two-sided t-test) associated with each layer as annotated in b. d, e, Enrichment of sn- and sc-like spatial barcodes with assigned cell types (columns) to morphological layers (rows) as in b. Color bar represents −log10(P value) (one-sided Fisher’s exact test, Bonferroni adjusted, P < 0.01) and gray tiles are nonsignificant values. OBNBL1, olfactory neuroblasts; OBINH1-3, inhibitory neurons; EPMB and ACMBS, astroependymal cells; VLMC2, vascular cells; SATG2, satellite glia; OBINBL5, GABAergic neuroblasts; OBDOB, dopaminergic periglomerular neuroblasts; OBINBL2, GGLUT1/2 neuroblasts; SEZ, subependymal zone; ONL, olfactory nerve layer; M/T, mitral layer; IPL, internal plexiform layer; GCL-E, GCL-I and GCL-D, granular layers; GL, glomerular layer; EPL, external plexiform layer.

Next, supervised analysis of HDST data correctly identified layer-specific expression signatures. We first annotated morphological layers (Methods) from the hematoxylin and eosin (H&E) stain of each specimen (Fig. 1b). We reasoned that 24 neighboring wells (r = ~6.5 μm) will likely capture transcripts from the same cell. We thus enhanced the signal by light binning, which pooled reads within a short range (for example, 5 × compared to the 5 × 5 hexagonal wells). On average, each 5 × ‘enhanced’ bin had observations from 5.6 ± 2.7 (mean ± s.d., n = 3 sections) (x, y) decoded beads and 44.4 ± 30.6 (mean ± s.d., n = 3 sections) UMIs. Finally, we assigned each ‘enhanced’ bin to a layer, to robustly identify differentially expressed (DE) genes between morphological layers (Methods). Following a smoothing Gaussian filter on the binned data (63.5 ± 38.6 (mean ± s.d.) UMIs per bin, n = 3 sections), we performed a two-sided t-test (FDR < 0.1), identifying DE signatures specific to morphological layers (Fig. 1c, Supplementary Table 3 and Supplementary Fig. 4). Layer-enriched upregulated DE genes (FDR < 0.05; log2(fold change) > 1.5) were correctly assigned, as assessed by comparing their average and individual signatures to their in situ hybridization (ISH) score from the Allen Brain Atlas (ABA) (Supplementary Fig. 5).

To test spatial assignment of cell types, we developed a multinomial naïve Bayes classifier to map the sparse high-resolution HDST data to cell type annotations by integration with scRNA-Seq (Methods). We first used scRNA-Seq UMI counts to compute the maximum likelihood estimates of the multinomial parameters for each cell type (Supplementary Table 4). We then estimated the likelihood that an expression profile of a given HDST barcode originated from a scRNA-Seq cell type and, using posterior probabilities, assigned cell types to barcode locations (Supplementary Table 5).

Approximately 49.4 ± 15.9% (mean ± s.d., n = 3 sections) of spatially barcoded HDST (1x) profiles were confidently assigned to a single cell type. We then leveraged the matched H&E images in HDST to segment single-cell nuclei based on the nuclear stain (Methods), related beads within nuclei and then used this aggregated expression information to also perform cell typing. To estimate our cell assignment’s sensitivity to read depth and spatial resolution, we decreased the resolution using segmenting and binning (Fig. 1d,e) and compared the assigned data to a ST dataset (Supplementary Fig. 6a). The posterior probabilities of cell type assignments increased in the aggregated data (Supplementary Fig. 6b,c and Supplementary Table 6), with a cell type confidently predicted...
in 58.1 ± 5.3% (mean ± s.d., n = 3 sections) of segmented and 61.3 ± 3.7% (mean ± s.d., n = 3 sections) of all (x,y) positions (Supplementary Fig. 6d), compared to 0.4% of (x,y) positions in ST data. DE markers drove the assignment task (Supplementary Fig. 6e and Supplementary Table 7).

Collecting H&E stains jointly with HDST data allowed us to further relate high-resolution barcodes to sub-cellular features. To demonstrate this, we performed nuclear segmentation and identified transcripts with preferential nuclear localization, by comparing RNAs associated with barcodes within or outside segmented nuclei (Supplementary Fig. 6e,f and Supplementary Table 8). Most of the 186 genes identified as nucleus specific by both HDST and single-nucleus RNA-Seq (Methods) were protein coding. Furthermore, HDST barcodes overlapping within segmented nuclei showed significantly higher (P < 0.05, one-sided unpaired Welch’s t-test) ratios of intronic versus exonic reads. This analysis can be extended to other sub-cellular features imaged with dedicated stains (for example, dendrites).

We related spatially assigned cell types to morphological layers (Fig. 1d,e, see Methods), finding layer-specific patterns for 15 of 63 tested cell types, typically consistent between segments and lightly binned data. For example, an olfactory neuroblast population was enriched in the mitral (M/T) and external plexiform layers (EPL), inhibitory neurons in the deep granular zone (GCL-D), astroependymal cells in the subependymal zone (SEZ) and olfactory ensheathing cells, vascular cells and satellite glia in the olfactory nerve layer (ONL). GABAergic, dopaminergic periglomerular and VGLUT1/2 neuroblasts were found in the glomerular layer. Many of these associations and classifications have previously been reported15–17, with inhibitory neurons dominating the granular (GCL-E, GCL-I and GCL-D) and internal plexiform (IPL) layers.

Fig. 2 | HDST distinguishes cell types and niches in a breast cancer resection. a, Labeling of morphological layers. HDST H&E image (left) of a breast cancer section and matching HDST (x,y) barcodes annotated into 13 morphological areas (right, color code). b, Layer-specific spatial DE patterns in HDST. Summed normalized expression of positively enriched signature genes significantly (FDR < 0.1, two-sided t-test) associated with each layer as in a. c, Cell-type assignments by single nuclei as in a. Two enlarged regions (black and red squares) with H&E and color-coded segments.

Relating histopathology and transcriptional profiles could help improve our understanding of disease biology and patient diagnosis and treatment. We assessed HDST’s clinical potential in a tumor section from a histological grade 3 breast HER2+ cancer patient (Fig. 2a, see Methods). We annotated clinically relevant morphological features in an H&E stain, and performed segmentation, differential expression analysis and cell typing, leveraging published auxiliary breast cancer scRNA-Seq data18 (Supplementary Table 9). DE genes between morphological areas (Fig. 2b and Supplementary Fig. 7a,b) using smoothed and binned data (38.03 ± 23.6 UMIs per bin containing 6.1 ± 3.1 beads at 91% library saturation) revealed that invasive cancer-specific areas were high in KRT19 and ERBB2, as expected, but also in TMSB10, a marker promoting migration of breast cancer cells19 (Supplementary Tables 10 and 11). A single cell type could be assigned to 59.8% of segments and 75.5% of bins driven by DE genes (Fig. 2c, Supplementary Fig. 7c and Supplementary Tables 12 and 13).

In conclusion, HDST is a high-definition method to measure in situ spatial information, at 1,400-fold higher resolution than ST, in healthy and pathological tissue. HDST is readily deployable as it relies on robust and commoditized tissue, molecular, bead-array and imaging modular tasks. Recently, Slide-Seq, a spatial RNA-Seq method with comparably low capture rates was developed20 with related features. However, Slide-Seq does not include histology, provides 25× lower resolution than HDST and has a higher rate of measurements confounded by signals from multiple cells. While HDST data is currently relatively sparse, signals are highly specific and interpretable by computational integration with morphological features and single-cell profiles. Further HDST development will improve understanding of tissue organization and function in health and disease.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41592-019-0548-y.

Received: 12 March 2019; Accepted: 2 August 2019; Published online: 9 September 2019

References

1. Oh, S. W. et al. Nature 508, 207–214 (2014).
2. Lein, E., Borm, L. E. & Linnarsson, S. Science 358, 64–69 (2017).
3. Maccosko, E. Z. et al. Cell 161, 1202–1214 (2015).
4. Zheng, G. Y. Y. et al. Nat. Commun. 8, 14049 (2017).
5. van den Brink, S. C. et al. Nat. Methods 14, 935–936 (2017).
6. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Science 348, aaa6090 (2015).
7. Wang, X. et al. Science 361, eaat5691 (2018).
8. Coskun, A. F. & Cai, L. Nat. Methods 13, 657–660 (2016).
9. Beliveau, B. J. et al. Nat. Commun. 6, 7147 (2015).
10. Lee, J. H. et al. Science 343, 1360–1363 (2014).
11. Ke, R. et al. Nat. Methods 10, 857–860 (2013).
12. Stahl, P. L. et al. Science 353, 78–82 (2016).
13. Michael, K. L., Taylor, L. C., Schultz, S. L. & Walt, D. R. Anal. Chem. 70, 1242–1248 (1998).
14. Gundersen, K. L. et al. Genome Res. 14, 870–877 (2004).
15. Nagayama, S., Homma, R. & Imamura, F. Front. Neural Circuits 8, 98 (2014).
16. Lein, E. S. et al. Nature 445, 168–176 (2007).
17. Zeisel, A. et al. Cell 174, 999–1014 (2018).
18. Karayayaz, M. et al. Nat. Commun. 9, 3588 (2018).
19. Zhang, X. et al. Breast Cancer Res. 19, 15 (2017).
20. Rodrigues, S. G. et al. Science 363, 1463–1467 (2019).
Acknowledgements
We thank the NGI Stockholm, SciLifeLab, Flatiron Institute and Simons Foundation for providing infrastructure support. We thank L. Gaffney, A. Hupalowska and J. Rood for help with manuscript preparation. Work was supported by the Knut and Alice Wallenberg Foundation, Swedish Foundation for Strategic Research, Swedish Cancer Society and the Swedish Research Council (to S.V., J.L. and P.L.S.), the EMBO long-term fellowship (no. ALTF 738-2017) to J.K., Early Postdoc Mobility fellowship (no. P2ZHP3_181475) to D.S., the NIH HuBMAP HIVE project and the Klarman Cell Observatory and HHMI (to A.R.). S.V is supported as a Wallenberg Fellow at the Broad Institute of MIT and Harvard.

Author contributions
S.V., J.F., J.L. and P.L.S. designed the experiments. S.V., F.S. and L.S. performed the experiments. S.V., G.E., J.K. and A.R. designed the analysis approaches. G.E. and J.K. devised and conducted the analyses with help from S.V., D.S., T.A., R.B. and J.F.N. Å.B. provided cancer samples. S.V. and J.G. implemented the annotation tool. G.K.G. annotated images. S.V., F.S., J.L and P.L.S. co-developed the high-definition arrays with M.R. S.V. and A.R. wrote the manuscript with input from all the authors. All authors discussed the results.

Competing interests
F.S., J.F., J.L. and P.L.S. are authors on patents PCT/EP2012/056823 (WO2012/140224), PCT/EP2013/071645 (WO2014/060483) and PCT/EP2016/057355 applied for by Spatial Transcriptomics AB (10x Genomics) covering the described technology. M.R. is employed by Illumina Inc. A.R. is a founder and equity holder of Celsius Therapeutics and an SAB member of Syros Pharmaceuticals and ThermoFisher Scientific.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0548-y.

Reprints and permissions information is available at www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.V. or J.L.

Peer review information: Nicole Rusk was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019
Methods

Bead preparation. We used a split-and-pool approach to generate a total of 2,893,865 different quality controlled 2 μm silica beads. A primer precursor containing (1) the T7 promoter, (2) an Illumina sequencing handle and (3) 15 base pair (bp) Spatial barcode_Pool1 (GenBank: KM416978) and (4) a 7 bp ‘bridge’ oligonucleotide sequence (/AmC6/UUUUUGACTGTAATACGACTCACTATAGGGACACGA CGCTTGCTCCGATCT-Spatialbarcode_Pool1-Bridge1) (IDT) were linked to the bead surface using amine chemistry. To increase the bead pool size, we pursued two additional ligation steps adding 14- and 15-bp pools of spatial barcode sequences; for example, Bridge2-Spatialbarcode_Pool2-Bridge3 was ligated to Bridge1 through a complementary Bridge1 Bridge2 14-bp helper sequence. T4 DNA ligase (NEB), following the manufacturer’s protocol, was used to couple the second spatial oligonucleotide construct, which was added in a 2:1 ratio to the precursor oligonucleotide. In the second ligation, the ligated sequence ending with the Bridge3 sequence acted as the precursor for the next spatial barcode pool (Bridge3-Spatial_barcode_Pool2-Bridge4) which was ligated to Bridge2 through Bridge3 ligated Spatial_barcode_Pool3 was followed by a 5-bp unique molecular identifier and 20 poly(d)VTN. The ligated Bridge1 Bridge2 sequences read GACTTGTCTAGAGC TGATGCCACACTACTC. All sequences used in the split-and-pool ligation steps (except the first precursor oligonucleotide containing the Spatial_barcode_Pool1) were synthesized on Illumina’s ‘Big Bird’ high-throughput oligonucleotide synthesis platform using phosphoramidite synthesis chemistry.

Array generation. The complete bead pool was used to load a total of 1,467,270 individual hexagonal wells covering a 13.7 mm² area (5.7 × 2.4 mm²). The wells were etched using a weak acid in a planar silica slide and polished to 1 μm height. A total of 14 cycles were made on each slide. The bead pool (~120 mg) was loaded in ethanol onto the planar slides with shaking. To ensure only one 2 μm bead would fit one well, the wells were etched at a diameter of 2.05 μm yielding a single bead per well coverage in over 99% of the wells.

Array decoding. Two sets of complementary and fluorescently labeled (FAM and Cy3) oligonucleotides were synthesized, denatured and purified. An additional set of unlabelled but still complementary probes was made. Each decoder set represented an individual decoder pool (10 μm). A total of 14 different colored (red, green and dark) pools were made. For example, if decoding 65 spatial oligonucleotide barcodes (Spatial_barcode_Pool1), the first decoder pool contained oligonucleotides complementary to spatial barcodes 1–27 labeled with FAM, barcodes 28–54 labeled with Cy3 and finally oligonucleotides 55–65 with no color label attached. In the second decoding cycle, oligonucleotides complementary to spatial barcodes 1–9, 28–36 and 55–63 were labeled with FAM, oligonucleotides complementary to barcodes 10–18, 37–45 and 64–65 were labeled with Cy3, and the rest were unlabeled. The color scheme was cycled for another two cycles for decoding 65 oligonucleotides in a total of four cycles. The same approach was then repeated to decode 211 barcoded oligonucleotides (Spatial_barcode_Pool2) with five cycles and another five cycles to decode 211 barcoded oligonucleotides in Spatial_barcode_Pool3 for a total of 14 cycles. This decoding approach was conducted as previously published resulting in each (x,y) position encoded per (x,y) barcode. For ST, more than one tag could be assigned per 2D spatial barcode, per region corresponding to 100 μm², a hexagonal well matrix distribution. We then translated pixel coordinates into fixed centroid (x,y) coordinates using the total detected area of the array. Well coordinates detected under the tissue boundaries were used in further analysis.

Histological image processing. To relate the histological image and the counts matrix, we assigned image pixel coordinates to the centroid of each bead well. First, we detected the arrays’ boundaries and corners, and assumed a perfect hexagonal shape. We then translate pixel coordinates into fixed centroid (x,y) coordinates using the total detected area of the array. Well coordinates detected under the tissue boundaries were used in further analysis.

Image annotation. Images used in this study were annotated with an interactive user interface for selecting spatial barcodes and their (x,y) coordinates based on the tissue morphology. Each (x,y) barcode position could be assigned to one or more of the nine distinct regions in the mouse olfactory bulb: ONL, granular cell layer external (GCL-E), granular cell layer internal (GCL-I), granular cell zone deep (GCL-D), EPL, M/T, IPL, SEZ and GL. For HDST MOB annotation, exactly one regional tags was assigned to one (x,y) spatial barcode. For ST, more than one spatial barcode was assigned per (x,y) position. For breast cancer, in pathological sections, we used six areas: invasive cancer, fatty tissue, fibrous tissue, normal glands, vascular space and immune/lymphoid cells with allowance for multiple region assignments. If a barcode position was not covered by an annotation polygon, the position was assigned to the closest polygon in cases where the unassigned barcode was within a 5 pixel distance.

H&E image segmentation for single-cell identification. Single-cell based segmentation on the H&E image was done by combining Ilastik v.1.3.2 (ref. 29) and CellProfiler v.3.1.8 (ref. 30). In Ilastik, we trained a random forest classifier to identify two distinct classes (nuclei and background). Based on this, we were able to segment and export the (x,y) files. Decoded pixel coordinates were then input into CellProfiler to segment the probability maps and identify single-cell masks for downstream analysis.

smFISH data processing. Previously published and processed MOB smFISH data was used in the analysis. Briefly, a 10 μm section was attached to a cover slip and the tissue was stained with 250 nM fluorescent label probes (LGC Biosearch Technologies) for three of the genes (Perk, Slc17a7 and Fabp7) diluted in staining buffer and a counterstain with Hoescht applied. z-stacks were imaged at 0.3 μm distance on a Nikon Ti-E. The images were stitched in Fiji. Three replicate regions of interest corresponding to 100 μm (in diameter) were randomly placed in three layers of the olfactory bulb: granular layer, glomerular layer and ONL. To estimate the average number of nuclei present per one region of interest in each layer, we used a grid approach on our histological image and counted detected nuclei. We then estimated the number of RNA spots per nucleus for each region and layer. These averaged RNA counts were compared allowing a Hamming distance of four mismatches. The forward read contained both the barcode the bridge sequences and were trimmed retaining the following bases: 1–15, 31–45 and 61–76. This created a forward read containing only a spatial barcode followed by a UMI. R2 transcripts were mapped with STAR24 to the mm10 mouse or GRCh38.79 human reference genome. Mapped reads were counted using the HTseq count tool (ref. 31) producing a UMI file per each tissue and region (normal and cancer).

Binning of HDST data. We divided the total area of each HDST array into nonoverlapping bins, each bin being 37 μm in diameter (100 μm X, 100 μm Y), and summed the UMI counts of beads within each spatial bin. To ensure appropriate bin sizes, we first considered all manufactured wells as a 1,918 × 765 matrix. On average, around 1,370 (x,y) wells filled with beads would size up to one ST spot (100 μm X, X = 38) when taking into account the center-to-center distance between two wells. We took the binned data containing 1,370 wells per bin and took every second bin into account in order to prevent overlapping. We then estimated the number of RNA spots per ST spot using a correction due to the Hamming distance between two ST spots would be accounted for. This bin size was called ‘ST-like’ in all further analyses. We did not take into consideration that this bin actually...
represents 63% of the transcriptome profiled per ST spot due to space between two hexagonal wells. Second, we made bins with fewer wells per bin in a logarithmic manner until reaching the smallest bin (5x). The 5x bin was referred to as single cell like of 'sc-like'.

**Spatial differential expression analysis.** Binned 5x data was smoothed using a Gaussian kernel (5x 5) with 0.5 standard deviations equally in both x and y directions. The smoothed binned data was then scaled for purposes of visualization such that the maximum expression value stayed the same. We performed a two-sided t-test to identify DE genes for each HDST morphological region (one versus the rest). The genes with a log(fold change) > 1.5 and FDR < 10% were identified as DE and used in further analyses. The Scapy package was used for visualization and differential expression analysis23.

**Validation of DE genes.** To validate layer-specific gene expression in the HDST data, we performed enrichment analysis using layer-specific gene sets from the ABA as reference. Layers annotated in both datasets were used in the analysis with all HDST granular layers merged into one instance to be comparable to the data provided in the ABA. Genes with a layer-specific log(fold change) > 1.5 (indicating upregulation) and FDR < 5% were tested for enrichment in the layer-specific gene sets (expression fold change greater than 1.5) from the ABA. Only genes passing the respective fold-change thresholds in both datasets were included in the analysis. For the top gene present in each layer were downloaded from ABAs High-Resolution Image Viewer and stitched using Fiji32.

**Assessing nuclear RNAs in HDST data.** Single-cell and single-nucleus data from the mouse (10x Chromium 3.2.v2 sequencing) from the M1 region on the mouse brain was downloaded from www.bicnc.org. BICNC data, tools and resources are released under the Creative Commons Attribution 4.0 International (CC BY 4.0, https://creativecommons.org/licenses/by/4.0/legalcode) License. The single-cell dataset was published from the U19 Zeng team (1U19MH114830-01) and the single-nucleus dataset from the U19 Huang team (1U19MH114832-01). Then, 50,000 randomly selected single nuclei and cells were used in downstream analysis. HDST data were split into two sets, based on whether the respective (xy) coordinate overlapped or not the segmented nucleus. We then identified those genes in each of the subsets that are present in either nuclei or cells. We observed 186 genes (128 protein coding, 58 noncoding) that are expressed exclusively in genes in each of the subsets that are present in either nuclei or cells. We observed 0.04807 ± 0.1507 (mean ± s.d.), nonnuclear barcodes: 0.0461 ± 0.1475 (mean ± s.d.), P = 0.017, one-sided unpaired Welch’s t-test.

**Cell type assignment to HDST barcodes.** For analysis of MOB data, we downloaded the matrix containing mean expression values x per cell type j from Zeisel et al. using the loompy package (https://github.com/tinnarson-lab/loompy). We subselected the matrix to contain only cell types annotated in the olfactory bulb and nonneuronal cell types for a total of 63 cell types. For analysis of HDST breast cancer data, we downloaded the expression matrix for a triple negative breast cancer single-cell RNA-Seq dataset18 from GEO (GSE118389) and calculated mean expression values x per cell type j for all genes contained in the matrix. Cell type annotations were obtained from the study’s GitHub repository (https://github.com/Michorlab/mac_scnrsae/blob/master/data/cell_types_tab_s9.txt).

The vector of probabilities Θj of each gene being captured in a cell type j is defined as follows:

\[ \Theta_j = [\theta_{j,1}, \ldots, \theta_{j,23}]^T \]

where θj is the gene, j is the cell type, k is the total number of genes and x is the mean expression. We calculate the likelihood L of the cell type j specified by Θj, given the observed UMIs b per gene for a HDST (xy) barcode as follows:

\[ L(\Theta_j | b) = \frac{n!}{b_1! \ldots b_{23}!} \times \theta_{j,1}^{b_1} \ldots \theta_{j,23}^{b_{23}} \]

where n is the vector of UMI counts per gene for an individual HDST (xy) profile and n is the total number of UMIs for an individual HDST (xy) profile.

For each HDST (xy) and cell type j, we calculated the ratio between the likelihood of that cell type L and the likelihood of the most likely cell type Lmax as a measure to assess how good the secondary cell type assignments are compared to the most likely cell type (primary assignment):

\[ L_k \frac{L_j}{L_{\text{max}}} \]

where \( L_k \) denotes the uniform prior for cell type k, \( P(c_k | b) \) and \( P(b | c_k) \) represent the posterior probability and the likelihood, respectively. \( P(b) \), the evidence term, is defined as \( \sum_\theta P(b | \theta) P(\theta) \) and used for the normalization.

Finally, to test against the null-hypothesis that HDST (xy) expression profiles originate from random expression profiles for each cell type j and respective Θj, we retained only nonzero elements of each Θj and shuffled them in 1,000 iterations while keeping the distribution of UMIs b as in the corresponding HDST (xy) expression profile. We then calculated the randomized likelihood Lrand, for each HDST (xy), cell type and iteration. Finally, an empirical P value \( p_{emp} \) was calculated for each HDST (xy) and cell type assignment as the fraction of \( L_{\text{max}} \) that yielded a likelihood higher or equal to the cell type likelihood L multiplied by the probability of drawing only nonzero values from Θ, given b, with correction for multiple testing (Benjamini–Hochberg). For each HDST (xy) the cell type with the highest likelihood \( L_{\text{max}} \) was considered the primary assignment and all cell types with \( L_j \geq 0.8 \) were considered secondary assignments. Finally, for a cell type assignment to be considered valid, we required \( L_j > 0.1 \) and \( p_{emp} < 0.01 \) for MOB and \( L_j > 0.7 \) and \( p_{emp} < 0.05 \) for breast cancer. For further analysis only HDST (xy) with exactly one valid cell type assignment were considered. Cell type assignments, ratios and the empirical P values for all HDST (xy) and binned MOB profiles have been reported in Supplementary Tables 5 and 6. Differential expression analysis was carried out between the cell types using two-sided t-test (log(fold change) > 1.5 and FDR < 10%).

**Auxiliary data pre-processing.** Public bulk RNA-Seq datasets19 were downloaded from NCBI’s GEO SRA with accession PRJNA316857, mapped to the mm10 reference and UMI filtered using the ST Pipeline v1.3.1. Averaged and naively adjusted34 bulk gene expression signatures were compared to those of the three replicates created with HDST and normalized the same way. ABA gene lists were downloaded from the ABA API using the ConnectedServices module of the allensdk Python package v0.16.0. The standard ST data as a counts matrix was downloaded from http://www.spatialtranscriptomicsresearch.org/.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

The raw mouse data have been deposited to NCBI’s GEO archive GSE139682. Raw files for the breast cancer sample are available through an MTA with Å. Borg (ake.borg.med.lu.se). All processed data is available at the Single Cell Portal (https://portals.broadinstitute.org/single_cell/study/SCP420).

**Code availability**

All code has been deposited on GitHub at https://github.com/kllarmann-cell-observatory/hdst.

**References**

21. Salmen, F. et al. Nat. Protoc. 13, 2501–2534 (2018).
22. Jemt, A. et al. Sci. Rep. 6, 37137 (2016).
23. Navarro, J. F., Sjostrand, J., Salmen, F., Lundeberg, J. & Stahl, P. L. Bioinformatics 33, 2591–2593 (2017).
24. Dobin, A. et al. Bioinformatics 29, 15–21 (2012).
25. Anders, S., Pyl, P. T. & Huber, W. Bioinformatics 31, 166–169 (2015).
26. Costea, P. I., Lundeberg, J. & Akan, P. PloS ONE 8, e57521 (2013).
27. Gauteret, D., Poirot, O., Lopez, F., Audic, S. & Claverie, J. M. Genome Res. 8, 530–531 (1998).
28. Wong, K., Navarro, J. F., Bergenstrahle, L., Stahl, P. L. & Lundeberg, J. Bioinformatics 34, 1966–1968 (2018).
29. Sommer, C., Straelen, C., Kothe, U. & Hamplefert, F. A. In IEEE International Symposium on Biomedical Imaging: From Nano to Macro 230–233 (IEEE, 2011).
30. Lamprecht, M. R., Sabatini, D. M. & Carpenter, A. E. J. Biotechniques 42, 71–75 (2007).
31. Lyubimova, A. et al. Nat. Protoc. 8, 1743–1758 (2013).
32. Schindelin, J. et al. Nat. Methods 9, 676–682 (2012).
33. Wolf, F. A., Angerer, P. & Theis, F. J. Genome Biol. 19, 15 (2018).
34. Svensson, V., Teichmann, S. A. & Stegle, O. Nat. Methods 15, 343–346 (2018).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  - Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Standard Illumina processing pipelines were used to collect pair-end sequencing data on a Nextseq 550 instrument. The BCL files from the sequencer were deposited in a NAS server where they were encrypted and sent to a backup server. The BCL files were then sent to a demultiplexing server where they were demultiplexed with bcl2fastq v2.17.1.14. The demultiplexed files were then sent to an analysis server where they merged by index prior analysis with the ST Pipeline (v1.5.1).

Data analysis

For initial ST data processing, ST pipeline v1.5.1 was used. The pipeline includes the following packages: nvoke, argparse, cython>=0.19, numpy, pandas, scipy, sqlite3, regex, taggd>=0.3.1 HTSeq>=0.7.1, pysam>=0.7.4, setuptools. For segmentation analysis, we used Ilastik v1.3.2 and CellProfiler v3.1.8. All computer code relevant to the file post processing, image alignment, DE and cell typing analysis has been deposited on GitHub at https://github.com/broadinstitute/hdst.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The raw mouse data has been deposited to NCBI's GEO archive GSE130682. Raw files for the breast cancer sample are available through MTA with Åke Borg (ake.borg@med.lu.se). Please note all confidential decoder sequences as proprietary to Illumina Inc were removed from raw files and substituted with a random string of the same length. All processed data is available on the Single Cell Portal (https://portals.broadinstitute.org/single_cell/study/SCP420/).
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- [ ] Life sciences
- [ ] Behavioural & social sciences
- [ ] Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size
Three mice of the same age and strain were used in the analysis. Three replicate sections from the same human sample were used in the analysis. In previous experimental designs and studies, we observed minimal variability for the obtained phenotypes among different samples on individually performed experiments using individual ST slides so we reasoned n=3 ( "n" indicates distinct samples ie HDST arrays and tissues) will be sufficient to determine method robustness as described in the article.

Data exclusions
No data was excluded from the analysis.

Replication
Three mice of the same age and strain were used in the analysis. Three replicate sections from the same human sample were used in the analysis. The experiments have been independently experimentally repeated a minimum of three times. Each experimental repetition have been performed on at least two independent slides. In all the attempts at repetition for each specific condition we observed similar behavior concerning the spatial localization of the markers analyzed.

Randomization
Not relevant for the study.

Blinding
Not relevant for the study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

| Materials & experimental systems       | Methods                        |
|---------------------------------------|--------------------------------|
| n/a                                   | Involved in the study          |
| - [x] Antibodies                    | [x] Involved in the study      |
| - [x] Eukaryotic cell lines          | [x] ChiP-seq                   |
| - [x] Palaeontology                  | - [ ] Flow cytometry           |
| - [x] Animals and other organisms    | - [x] MRI-based neuroimaging   |
| - [x] Human research participants    |                                |
| - [x] Clinical data                  |                                |

Animals and other organisms

Policy information about studies involving animals, ARRIVE guidelines recommended for reporting animal research

Laboratory animals
Adult C57BL/6J mice (at 12 weeks age). Replicate#1 was male (mouse#1) and Replicate#2 and Replicate#3 were females (mouse #2). The metadata associated with all samples have been reported in the GEO metadata format and is reported with the study.

Wild animals
The study did not involve wild animals.

Field-collected samples
The study did not involve samples collected from the field.

Ethics oversight
This study complied with all relevant ethical regulations regarding experiments involving animal tissue samples. Ethical permit (N155-16) for animal work was granted by the Regional Ethics Committee of Sweden.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics
Three replicate sections from the same human sample were used in the analysis. Patient had the following metadata associated:
Population characteristics

T1 (<20mm), lymph node-positive, histological grade 3, ER+, HER2+. The patient was a 70 years old female.

Recruitment

Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.

Ethics oversight

This study complied with all relevant ethical regulations regarding experiments involving human tissue samples and samples were collected with informed consent. Ethical permission (Dnr 2009/658 and Dnr LU240-01) for the human sample used in this study was granted by the Regional Ethics Committee of Sweden at the Lund University Hospital.

Note that full information on the approval of the study protocol must also be provided in the manuscript.