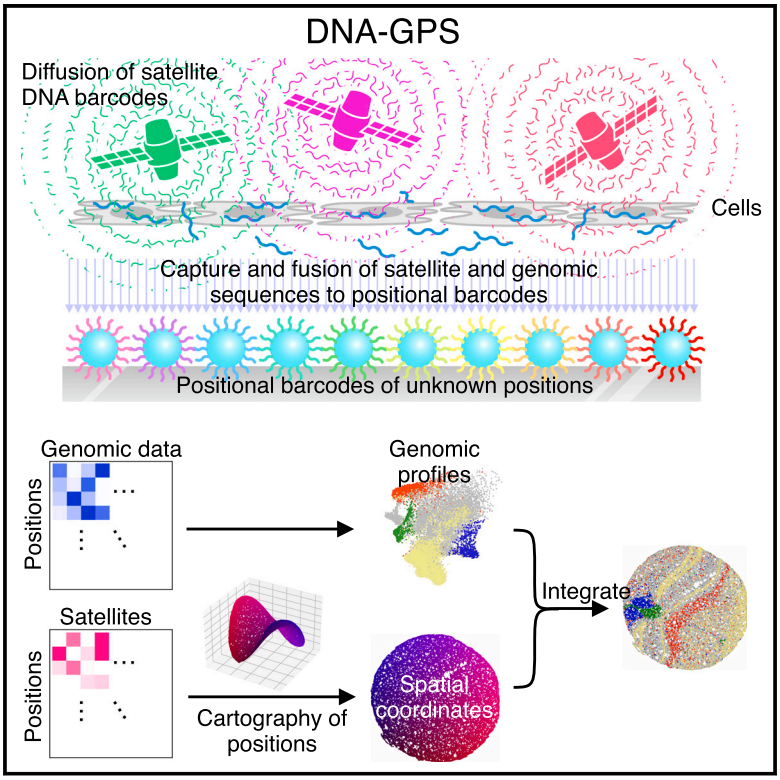
Cell Systems

DNA-GPS: A theoretical framework for optics-free spatial genomics and synthesis of current methods

Graphical abstract



Authors

Laura Greenstreet, Anton Afanassiev, Yusuke Kijima, ..., Samuel King, Nozomu Yachie, Geoffrey Schiebinger

Correspondence

nozomu.yachie@ubc.ca (N.Y.), geoff@math.ubc.ca (G.S.)

In brief

While single-cell sequencing methods capture genomic profiles at the cellular level, they lose spatial context. In the past decade, over a dozen spatial transcriptomics methods have been developed. Greenstreet et al. highlight trade-offs and synergies in existing methods and propose DNA-GPS, a framework for large-scale optics-free spatial transcriptomics.

Highlights

- Spatial transcriptomics methods exhibit trade-offs in resolution and scale
- They can be split into optical imaging, positional indexing, and mathematical cartography
- DNA-GPS combines ideas from positional indexing and mathematical cartography
- It provides a mathematical framework for large-scale optics-free spatial transcriptomics

Greenstreet et al., 2023, Cell Systems *14*, 844–859
October 18, 2023
https://doi.org/10.1016/j.cels.2023.08.005

CellPress

CellPress
OPEN ACCESS

Cell Systems

Synthesis

DNA-GPS: A theoretical framework for optics-free spatial genomics and synthesis of current methods

Laura Greenstreet,[1,7] Anton Afanassiev,[1,7] Yusuke Kijima,[2,3,7] Matthieu Heitz,[1,7] Soh Ishiguro,[2] Samuel King,[2] Nozomu Yachie,[2,4,5,6,8,*] and Geoffrey Schiebinger[1,2,9,10,*]

[1]Department of Mathematics, The University of British Columbia, Vancouver, BC, Canada
[2]School of Biomedical Engineering, The University of British Columbia, Vancouver, BC, Canada
[3]Department of Aquatic Bioscience, The University of Tokyo, Tokyo, Japan
[4]Research Center for Advanced Science and Technology, The University of Tokyo, Tokyo, Japan
[5]Premium Research Institute for Human Metaverse Medicine (WPI-PRIMe), Osaka University, Suita, Osaka, Japan
[6]Graduate School of Media and Governance, Keio University, Fujisawa, Japan
[7]These authors contributed equally
[8]X (formerly Twitter): @nzmyachie
[9]X (formerly Twitter): @geoffschieb
[10]Lead contact
*Correspondence: nozomu.yachie@ubc.ca (N.Y.), geoff@math.ubc.ca (G.S.)
https://doi.org/10.1016/j.cels.2023.08.005

SUMMARY

While single-cell sequencing technologies provide unprecedented insights into genomic profiles at the cellular level, they lose the spatial context of cells. Over the past decade, diverse spatial transcriptomics and multi-omics technologies have been developed to analyze molecular profiles of tissues. In this article, we categorize current spatial genomics technologies into three classes: optical imaging, positional indexing, and mathematical cartography. We discuss trade-offs in resolution and scale, identify limitations, and highlight synergies between existing single-cell and spatial genomics methods. Further, we propose DNA-GPS (global positioning system), a theoretical framework for large-scale optics-free spatial genomics that combines ideas from mathematical cartography and positional indexing. DNA-GPS has the potential to achieve scalable spatial genomics for multiple measurement modalities, and by eliminating the need for optical measurement, it has the potential to position cells in three-dimensions (3D).

INTRODUCTION

Understanding the cellular architecture of tissues is a major challenge with tremendous potential impacts across the life sciences. Single-cell measurement technologies have enabled high-dimensional, unbiased measurements of transcriptomic, proteomic, genomic, or epigenomic information in large populations of cells.[1–4] However, these technologies lose spatial context. Spatial genomics has made it possible to chart the spatial distribution of transcriptional states,[5] chromatin accessibility,[6] and other elements of cell state.[7,8] This article first categorizes current spatial genomics technologies into three groups, identifying gaps and synergies, then introduces a theoretical framework for optics-free spatial genomics called DNA-GPS (global positioning system) that combines ideas from mathematical cartography and positional indexing to overcome the gaps in the current technologies. While DNA-GPS could be implemented with small modifications to existing positional indexing methods, it has the potential for further extensions to multi-modal measurement and three-dimensional (3D) volumes.

Optical imaging

Since Hooke's discovery of cells in the 1660s, the microscope has been an indispensable tool for biological discovery. Today, gene expression can be imaged through a variety of optics-based approaches (Figure 1A).

In situ hybridization (ISH) is an imaging system where reporter probes hybridize to a specific target nucleotide sequence, enabling the spatial localization of the target sequence in a histologic section under a microscope.[29] Expression of a target gene over a tissue section can be quantified from the intensity of the reporter. Fluorescent ISH (FISH) using fluorescent probes has enabled multiplexed measurements,[30] but its scalability has classically been limited to the number of distinct fluorescent probes that can be imaged simultaneously. Recent advances in single-molecule FISH (smFISH) with super-resolution imaging have enabled localization and counting of individual RNA molecules over the tissue space, and combinatorial indexing strategies have expanded the number of targets to the genomic scale.[9–11] In the multiplexed smFISH approach, each RNA molecule is imaged multiple times over cycles of different probe hybridizations and

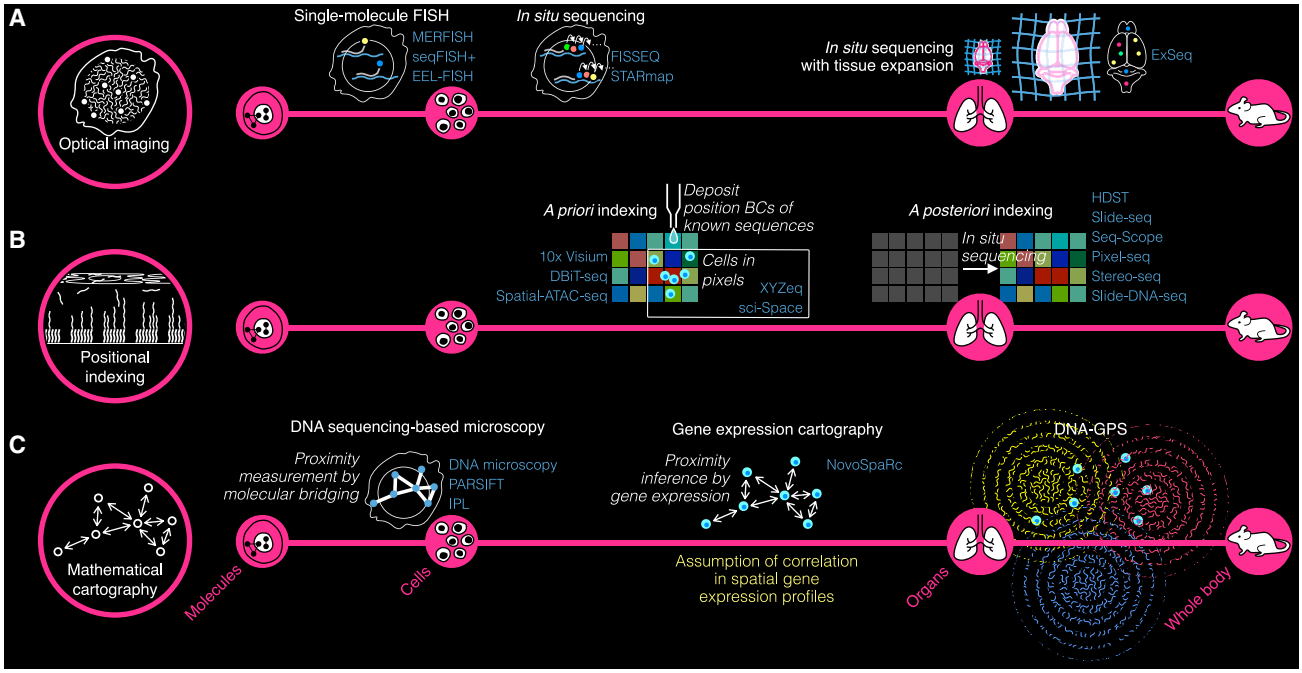


Figure 1. Overview of spatial genomics

(A) Optical imaging. MERFISH,[9] seqFISH+,[10] and EEL-FISH[11] fall into smFISH category. FISSEQ,[12] STARmap,[13] and ExSeq[14] fall into *in situ* sequencing category.
(B) Positional indexing. 10× Visium,[15] DBiT-seq,[16] Spatial-ATAC-seq,[6] XYZeq,[17] and sci-Space[18] fall into *a priori* indexing category. HDST,[19] Slide-seq,[20] Seq-Scope,[21] Pixel-seq,[22] Stereo-seq,[23] and Slide-DNA-seq[24] fall into *a posteriori* indexing category.
(C) Mathematical cartography. DNA microscopy,[25] PARSIFT,[26] and IPL[27] fall into DNA sequencing-based microscopy category. NovoSpaRc[28] is a gene expression cartography method.

fluorophore quenching, where a unique fluorescent color code appearing through the imaging cycle is assigned to each RNA species.[9] Scalable fluorescent indexing of many RNA species has been enabled by two-step smart encoding approaches, where the first gene specific probes with common anchor sequences are further probed by secondary common fluorophore probes, minimizing the cost of fluorescent probes.[9–11] However, due to its reliance on super-resolution imaging, smFISH greatly sacrifices field of view to observe large numbers of genes. A new technology called EEL-FISH, which transfers tissue RNAs onto a positively charged glass slide by electrophoresis, has improved scalability by reducing imaging effort. Yet, it still takes more than 2 days to process ~500 genes over 1 cm^2.[11] Furthermore, smFISH probes designed for one species cannot be utilized for other species.

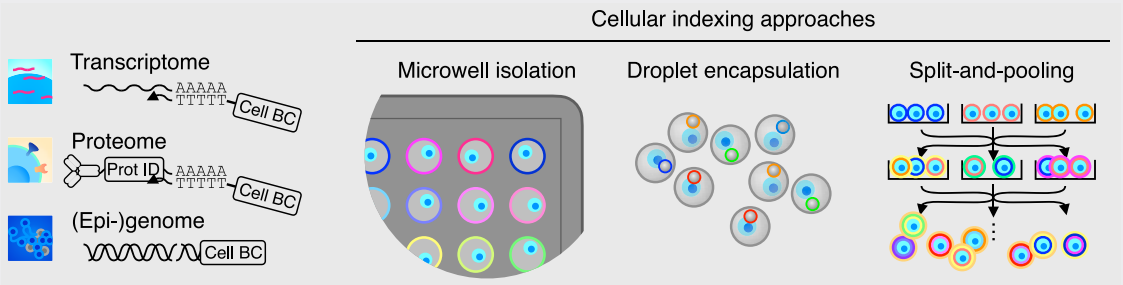
In situ sequencing has been coupled with super-resolution imaging to enable genome-scale spatial transcriptomics.[12–14] In this approach, after fixing a tissue section, RNAs are reverse transcribed into complementary (c)DNAs, which are then circularized and amplified by rolling circle amplification (RCA). Next, the amplified cDNA sequences are sequenced *in situ* by sequencing-by-ligation (where specific fluorescent signals are spatially obtained over cycles along with the template directed elongation of sequence strand by short probe ligation). Similar to smFISH, this approach requires high-resolution imaging, resulting in a small field of view. Expansion microscopy (ExM), a technology to physically expand tissue using a swellable polymer gel,[31] has also been applied to *in situ* transcriptomics and achieved high scalability with better signal separation and reagent penetration. Coupling *in situ* sequencing of short gene regions as well as *ex situ* sequencing of longer gene regions, this expansion sequencing (ExSeq) approach has enabled spatial mapping of long sequencing reads.[14] As the expansion process preserves the intact tissue structure, ExSeq has also demonstrated spatial transcriptomics of a 3D volume; however, the expansion approach has only be demonstrated in soft tissues.[32]

Scalable optical imaging has also been developed for other non-transcriptomic modalities. Multiplexed smFISH and *in situ* sequencing have been applied to locate genomic sequences together with epigenetic marks at molecular resolution, enabling the identification of high-resolution chromatin structures, but they are limited to few thousands of cells.[33–35] Tissue clarification approaches and light-sheet microscopy have also enabled multiplexed 3D localization of proteins at whole organ scale using fluorescent-conjugated antibodies, but for a limited number of protein species.[36] In summary, optical imaging technologies can profile the 3D distribution of molecules across cells or tissues, but the use of optics introduces fundamental trade-offs between field of view, resolution, and the number of molecular species to be profiled.

Positional indexing

Positional indexing (Figure 1B) is a category of spatial genomics methods whose concept has been derived from the widely applied idea of cellular indexing in single-cell genomics technologies (Box 1). In cellular indexing, cellular materials are

Box 1. Cellular indexing approaches for single-cell genomics



Single-cell RNA sequencing (scRNA-seq) was originally demonstrated in microwells, where individual cells are physically isolated in wells. In each well, polyadenylated (poly(A)ed) RNA products are captured and reverse transcribed by primers having poly(thymine) (poly(T)) on their 3′ ends together with well-specific barcodes. This reverse transcription (RT) reactions fuse the producing complementary (c)DNA products to the well-specific barcodes. After high-throughput sequencing, single-cell gene expression profiles can be reconstructed from read counts of genes sorted by the well-specific barcodes.[37–39] In 2015, water-in-oil droplet-based approaches revolutionized scRNA-seq, where single cells are each encapsulated into droplets together with microbead anchoring poly(T) RT primers encoding a unique bead-specific barcode.[40,41] Following microfluidic high-throughput generation of droplets, single-cell transcriptomes are captured on the beads, and their 3′ sequences are fused to the bead-specific (i.e., cell-specific) barcodes by RT for high-throughput sequencing. Similar to the microwell-based approaches, the total sequencing reads for gene expression counts are sorted into single cells according to the unique cell-specific barcodes fused to them. This approach has been widely adopted by the community today and is still under significant development.[42–45]

Recently, split-and-pooling was proposed to achieve scalable single-cell genomics with no single-cell compartmentalization.[4,46–48] In split-and-pooling, single cells are first fixed with their genomic materials and split into a handful of wells where their transcriptome sequences are reverse transcribed and conjugated to well-specific DNA barcodes *in situ*. The cell populations are then pooled and split again into multiple wells where their transcriptome sequences are further conjugated to barcodes specific to the second-round wells. The iterations of this split-and-pooling procedure enable transcripts of every single cell to be tagged with a unique combinatorial array of barcodes. This approach has been demonstrated to achieve profiling multiple millions of single cells.[46,49]

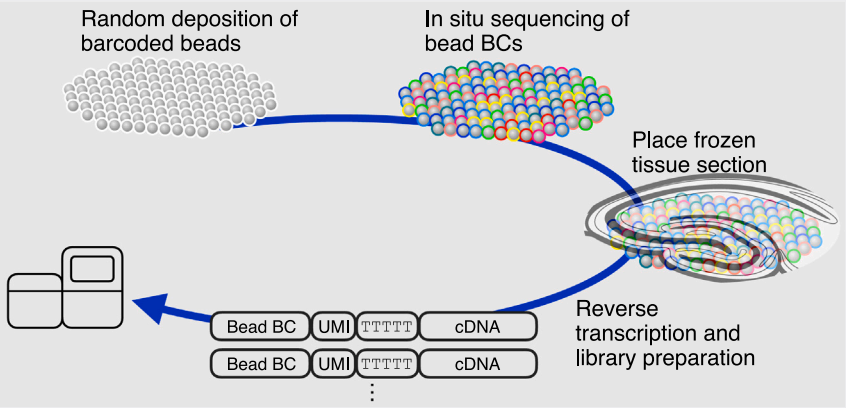
Single-cell measurements of molecular profiles that can be converted into a form of DNA sequences have been enabled by these cellular indexing approaches. For example, similar to the modality of transcriptome measurements achieved by converting RNA to DNA with the conjugation to cell-specific barcodes, single-cell surface proteome measurement has also been enabled by labeling surface proteins using a cocktail of antibodies tethered to their specific DNA barcodes followed by poly(T) sequence.[2] This enabled the simultaneous single-cell measurement of transcriptome and surface proteome profiles. Genomic sequences,[3] epigenetic modifications,[50,51] and chromatin structures[4,52] have also been measured at the single-cell resolution with the same concept.

translated into a form of DNA, if needed, and fused to cell-specific DNA barcodes. The barcoded DNA sequences are identified by high-throughput sequencing. DNA sequences associated with single cells are later sorted computationally according to the cell-specific DNA barcodes.

In positional indexing for spatial genomics, DNA barcodes are used to index the spatial pixels of an image. For example, in spatial transcriptomics a tissue slice is first stamped on a pixelated surface of barcoded poly(thymine) (poly(T)) primers for reverse transcription (RT) of polyadenylated (poly(A)ed) RNAs. Tissue RNAs are captured and reverse transcribed by their proximal RT primers, resulting in the fusion of positional barcodes to cDNAs. After sequencing, the gene expression profile of each pixel is reconstructed computationally from the positional barcodes. This approach has great potential scalability because it allows for reading out the transcriptome in a highly multiplexed manner by high-throughput DNA sequencing without the need for a specific probe set or in-tissue optical imaging. Some methods in this category have already reached single-cell resolution at whole-body scale for organisms such as mice.[23]

There are two types of indexing approaches to encode positional information to DNA barcodes: *a priori* indexing and *a posteriori* indexing. In *a priori* positional indexing, as seen in 10× Genomics Visium,[15] a set of known DNA barcodes are deposited to preassigned positions. After applying a tissue slice on the barcoded surface, DNA sequences representing molecular profiles are fused to the positional barcodes and analyzed by sequencing. The density, size, and number of barcoded spots determine the field of view and resolution of the spatial genomics image being obtained. While its scalability has been limited to the number of DNA barcodes that can be synthesized separately and deposited to unique spatial positions, combinatorial indexing seen in DBiT-seq has broken this linear scalability.[16] DBiT-seq uses a microfluidics device that can create two sets of unidirectional massively parallel flows on a tissue slice. The first set of flows provides DNA barcodes of primary (row) coordinates to the tissue-derived DNA. The second set of flows are 90°-rotated from the first and provide secondary (column) coordinates. The two barcodes are ligated together to encode both row and column information. This combinatorial row-column

Box 2. Slide-seq



Slide-seq[20,53] is a pioneering method for spatial transcriptomics using *a posteriori* positional indexing strategy that has many similarities with Drop-seq,[40] a droplet-based method for scRNA-seq (Box 1). Slide-seq also employs barcoded beads that are 10 μm diameter, made of polystyrene, and surfaced by poly(T) RT primers encoding unique bead-specific barcodes. The barcoded beads are first densely and randomly deposited on a class surface. The bead layer is then subjected to optical *in situ* sequencing of a sequencing-by-ligation method to identify barcode sequences and their locations. After the identification of barcodes and their coordinates, tissue section is applied on their surface, where released RNAs are captured by their proximal barcoded RT primers on the beads. Finally, the sequencing library is prepared by RT, followed by PCR amplification. The sequencing reads are analyzed similarly to that of scRNA-seq, where gene expression profiles of positional barcodes are sorted with their two-dimensional coordinates. Slide-seq has been demonstrated for a range of mouse tissues including hippocampus, cerebellum, liver, and kidney and revealed the spatial cell state distribution and their molecular profiles across the tissue substructures.

The field of view of Slide-seq is a 3-mm diameter circle since the current protocol uses a 3-mm gasket to distribute barcoded beads. The same group reported an updated version, Slide-seq V2, in 2020 with an improved bead synthesis strategy, sequencing method, and library preparation protocol.[53] Although it largely improved the transcriptome yields compared with the initial version, the limitation in scale remains. Additionally, like other positional indexing approaches, transcripts from multiple cells may be captured by the same bead.

indexing of spatial genomic products has achieved more scalable position indexing with minimized barcode members and a spot size of ~10 μm square per side.

In *a posteriori* positional indexing, clusters of unidentified monoclonal DNA barcodes are first distributed on a two-dimensional (2D) surface and then later identified by *in situ* sequencing or smFISH. In Slide-seq[20,53] (Box 2) and HDST,[19] barcoded poly(T) RT primer beads are randomly distributed in space and then identified by *in situ* sequencing and smFISH, respectively. Seq-Scope[21] and Pixel-seq[22] use an approach similar to Illumina sequencing: spatially distributed monoclonal DNA barcode clusters are prepared by bridge PCR amplification of sparsely immobilized DNA barcode molecules on a 2D surface. The sequences of clonal sequence clusters are then identified via sequencing by synthesis. Similarly, in Stereo-seq,[23] barcoded DNA nanoballs are generated by RCA, deposited spatially, and then identified by DNBSEQ sequencing.[54] Stereo-seq has achieved unparalleled scalability, positioning reads over a 13.2 cm × 13.2 cm area at subcellular resolution.

Positional indexing links genomic products to a single spatial pixel, but these pixels can capture information from multiple cells because the pixel boundaries do not align with cell boundaries. Further, genomic products of single cells are likely to diffuse before they bind to positional barcodes, blurring the image. To address this issue, several approaches have been developed to extend cellular indexing from single-cell genomics to spatial genomics. In XYZeq[17] and sci-Space,[18] a tissue sample is first stamped on a surface of positional barcodes prepared by *a priori* indexing. In this approach, the positional barcode molecules are not fused to the genomic products of cells but to the cells themselves. Cells are then dissociated, and their genomic products and positional barcodes are both indexed by cellular barcodes using split-and-pooling indexing (Box 1). Both the spatial positions and genomic profiles of cells can be decoded by high-throughput sequencing. However, the cells are not localized with extreme precision: sci-Space has achieved a higher resolution with a pixel size of 73.2 μm compared with XYZeq with a pixel size of 500 μm. For precise localization of single cells, this approach would require smaller positional barcode pixels. Further, because split-and-pooling loses a significant fraction of cells, this approach cannot exhaustively scan single-cell genomic profiles across a tissue section.

Positional indexing has also enabled the measurement of diverse modalities of molecular profiles in a spatial context. For example, DBiT-seq has demonstrated simultaneous spatial profiling of the transcriptome and cell surface proteome.[7,16] Slide-seq has enabled spatial genome sequencing by applying Tn5 transposase that fragments genomic DNA and concatenates

the fragments to adapter sequences.[24] Similarly, DBiT-seq has also recently been demonstrated for different genomic modalities to identify open chromatin regions, chromatin modifications, and their combinations with gene expression profiles across a tissue space.[6,8,55]

Positional indexing has scaled 2D spatial genomics to the level of whole organs and small organisms. *A priori* approaches are likely reaching the limit of their scalability due to technical limitations in precisely depositing many positional barcodes to corresponding small areas. While *a posteriori* indexing approaches have the potential to scale with the growth of massively parallel sequencing technologies, they require optical *in situ* sequencing to retrospectively identify positional barcodes. Therefore, these positional indexing approaches cannot benefit from advances in non-optics-based DNA sequencers, such as nanopore sequencers. Further, the need for optical imaging limits positional indexing approaches to 2D tissue sections.

Mathematical cartography

The last emerging family of spatial genomics techniques is mathematical cartography (Figure 1C), where an image is reconstructed by solving an inverse problem. This eases the experimental burden of directly profiling transcriptional states at precise positions over large areas and relies only on DNA sequencing. Mathematical cartography approaches have the potential to elegantly increase in scale as DNA sequencing technologies improve, but none of them currently have practical large-scale demonstrations. In principle, these techniques are not limited to two dimensions. Currently, this class is represented by sequencing-based microscopy and gene expression cartography methods. However, we present a theoretical framework for mathematical cartography, DNA-GPS, which is more scalable than existing approaches and synergizes with positional indexing methods while not relying on optical imaging.

DNA microscopy[25] offers an alternative to both optical imaging and positional indexing approaches. Similar to positional indexing, it is a sequencing-based approach. However, while positional indexing technologies require an experimental procedure to determine their spatial locations, in DNA microscopy, the locations of spatially distributed molecules are inferred only from DNA sequencing data where each sequencing read encodes the information needed to reconstruct its position. Its first implementation reconstructed the spatial distribution of RNA molecules over an estimated 250–700 cells. After fixing cells, target RNA species are reverse transcribed *in situ* with primer molecules encoding random short sequences as unique molecular identifiers (UMIs). The resulting UMI-tagged cDNAs are then amplified *in situ* by PCR and their products are allowed to diffuse spatially. As this *in situ* PCR step is performed with primers encoding common sequences and random nucleotides, the cDNA amplicons are fused with proximal amplicon products through overlap-extension, resulting in two UMI-tagged gene products fused with primer-derived random nucleotide segments at the junction. The resulting random nucleotide pair at the junction is called a unique event identifier (UEI). The cDNA sequences, UMIs, and UEIs are readout by high-throughput DNA sequencing. This sequencing data collectively represent the physical proximities between every single RNA molecule, which originally served as a source template for spatially diffused cDNA amplicons. The spatial positions of the RNA molecules can be reconstructed with a maximum likelihood procedure, based on pairwise proximities.

In theory, this reaction can be performed at a tissue scale with a single readout of high-throughput sequencing. However, this molecular cartography has only been demonstrated to date with images consisting of less than a thousand cells due to the required sequencing depth. Since molecules must be densely sequenced to reconstruct positions using collisions, both sequencing and computational efficiency would be needed to apply this method at a large scale. For example, with $\sim 10^5$ molecules per cell and $\sim 10^6$ cells/cm^2, reconstructing 100 cm^2 of tissue would require sequencing 10^{13} molecules and, even if a sparse binary matrix could be constructed with only 10 neighbors per molecule, would still require storing and processing 100 terabyte (TB) of data. This illustrates the benefit of localizing cells instead of molecules: while single-molecule resolution may eventually be useful, there will still be much to be gained from studying tissue architecture at the level of cells, rather than individual molecules within cells. Indeed, some recent single-molecule studies have analyzed their data by aggregating them to the level of cells.[10,23]

PARSIFT[26] and IPL[27] are two theoretical mathematical cartography approaches.[56] In these approaches, a 2D surface of DNA sequences is charted by planar graph embedding where the nodes of the graph represent DNA sequence locations and adjacent sequences are connected by an edge. The two methods propose different methods for measuring this adjacency graph, roughly based on concatenating adjacent barcodes and sequencing. Similar to the first method, both of these methods would require tremendous sequencing depth to cover large spatial areas. Further, practical implementations of spatial genomics based on these approaches remain to be proposed.

Gene expression cartography is a mathematical framework that predicts cellular locations directly from single-cell gene expression profiles obtained by single-cell RNA sequencing (scRNA-seq). One of the earliest methods in this category is Seurat,[57] which has been demonstrated on zebrafish embryos. However, it requires additional FISH spatial expression patterns of several marker genes collected from independent embryos of the same developmental stage. Therefore, this approach can be used only when rich spatial gene expression datasets of the same species at the same biological condition are available. NovoSpaRc[28] is another approach that does not theoretically rely on auxiliary FISH datasets. Instead, this method relies on the strong assumption that cells located in spatially close positions have similar gene expression profiles. This assumption enables setting up an optimization problem to identify the locations of cells from their gene expression profiles. Although NovoSpaRc demonstrated reconstructions of spatial gene expression profiles from a range of datasets to some degree, images that were similar to the biological ground-truth samples could not be obtained unless spatial gene expression patterns for marker genes were provided to constrain the image reconstruction process. This is likely because the key assumption of nearby cells sharing similar transcription profiles fails to hold in some locations, such as at tissue boundaries. Accordingly, gene expression cartography is a scalable approach but still

requires direct measurement of spatial expression patterns for at least several genes with FISH.

DNA GPS

We now propose DNA-GPS, a theoretical framework for large-scale optics-free spatial genomics. Similar to other mathematical cartography methods, DNA-GPS localizes positions using sequencing information alone. However, DNA-GPS greatly increases scalability by incorporating positional barcodes instead of resolving individual molecules. These barcodes can be associated with spatial pixels, as is done in positional indexing, or with the cellular barcodes used in single-cell genomics. DNA-GPS localizes positional barcodes using manifold learning (Box 3). The data required for this spatial genomics can be collected entirely through DNA sequencing, without optics at any stage. Based on simulations presented below, only 30–500 additional reads per barcode would need to be sequenced to position each barcode at 10–30 μm resolution. The majority of mammalian cells are 10–100 μm in diameter,[58–70] though a few cell types are slightly below 10 μm.[71–76] Therefore, in principle, DNA-GPS should scale as a small overhead on total sequencing at the near-single-cell resolution as DNA sequencing technologies improve.

Our key idea is to introduce an artificial set of "satellite DNA barcodes," each of which diffuses in a spatially coherent way from a random initial position (Figure 2A). The spatially distributed satellite barcodes can be captured in the same way as the cellular genomic contents and tagged by positional barcodes. Therefore, the positional barcodes provide "information anchoring" between the satellite barcodes and the cellular genomic contents from the same position. For example, DNA-GPS can determine the positions and sequences of positional bead barcodes used in Slide-seq (Box 2) with no *a priori* or *a posteriori* indexing process. Below, we explain and demonstrate the idea through simulations. In these simulations, we focus on one concrete realization of DNA-GPS related to Slide-seq. We then discuss the broad applicability of DNA-GPS and its potential synergies with other spatial and singe-cell genomics methods.

To apply DNA-GPS to determine locations of positional bead barcodes, similar to those used in Slide-seq, tissue transcriptome and satellite barcodes would both be captured by their proximal RT primers encoding positional barcodes on spatially distributed beads and fused to the positional barcode sequences by RT and polymerase extension, respectively (Figure 2B). The locations of bead barcodes could then be reconstructed using manifold learning to learn the underlying 2D positions of beads from their satellite barcode count profiles. This is similar to the GPS used in navigation systems for cars and cell phones, where devices determine their precise location on Earth using their distance to multiple satellites. The core principle of DNA-GPS is that neighboring points in physical space will collect similar counts of satellite barcodes because they are distributed in a spatially coherent way. In mathematical terms, the measurement process can be understood to embed the physical 2D positions of bead barcodes into a high-dimensional "satellite barcode space," where each point is represented by its vector of satellite barcode counts, similar to how a transcriptome profile is represented by a high-dimensional gene expression vector (Figure 2C). The embedded bead layer forms a low-dimensional manifold, or surface, in high-dimensional satellite barcode space. We provide a precise theorem statement in the supplemental information (see Method S1 and

Box 3. Manifold learning

Manifold learning aims to learn a low-dimensional structure embedded in a higher dimensional space. It can be thought of as a form of dimensionality reduction, with the stronger assumption of an underlying low-dimensional structure.

Principal-component analysis (PCA) can exactly capture linear structures such as lines and planes by decomposing the data into linear, uncorrelated components.[77] PCA has high computational complexity, as it is calculated by performing a spectral decomposition on the covariance matrix. However, a small number of linear components can be found more efficiently by performing truncated singular value decomposition (SVD). Classic or metric multidimensional scaling (MDS) is another popular linear manifold learning approach that uses the pairwise distances between data points.[78] It similarly has a high computational complexity by relying on a spectral decomposition.

Many times, we want to capture non-linear structures. Non-linear spectral methods generally share a three-step process of creating a graph on the data by connecting neighboring points, converting the graph to a matrix, and taking a spectral embedding of the resulting matrix via a spectral decomposition, where the methods differ in the process used to convert the graph to a matrix.[79] Spectral methods include locally linear embeddings (LLEs), Isomap, and Laplacian and Hessian eigenmaps.[80–83] The spectral decomposition is computationally intensive, but some methods, such as LLE, are more scalable as they consider small areas instead of global structure.

Force-directed layout embedding methods offer more scalability, as naively they only require pairwise comparisons and using approximations can be scaled to millions of data points. t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP) are two popular force-directed layout methods. t-SNE represents the similarity between points as probabilities and minimizes the divergence between the high and low-dimensional probability distributions.[84,85] UMAP uses topological structures to capture properties of the data in the high-dimensional space and then uses a force-based method to find a low-dimensional Euclidean representation that approximates the high-dimensional topological structure.[86]

Autoencoders are a class of deep learning models for dimensionality reduction that can capture complex non-linear relationships.[87] Autoencoders are neural networks with internal layers that are a lower dimension than the original data, forcing the network to learn a low-dimensional representation of the data to reconstruct the original data as output. Autoencoders are highly scalable. However, it can be difficult to interpret the embeddings with respect to the original features.

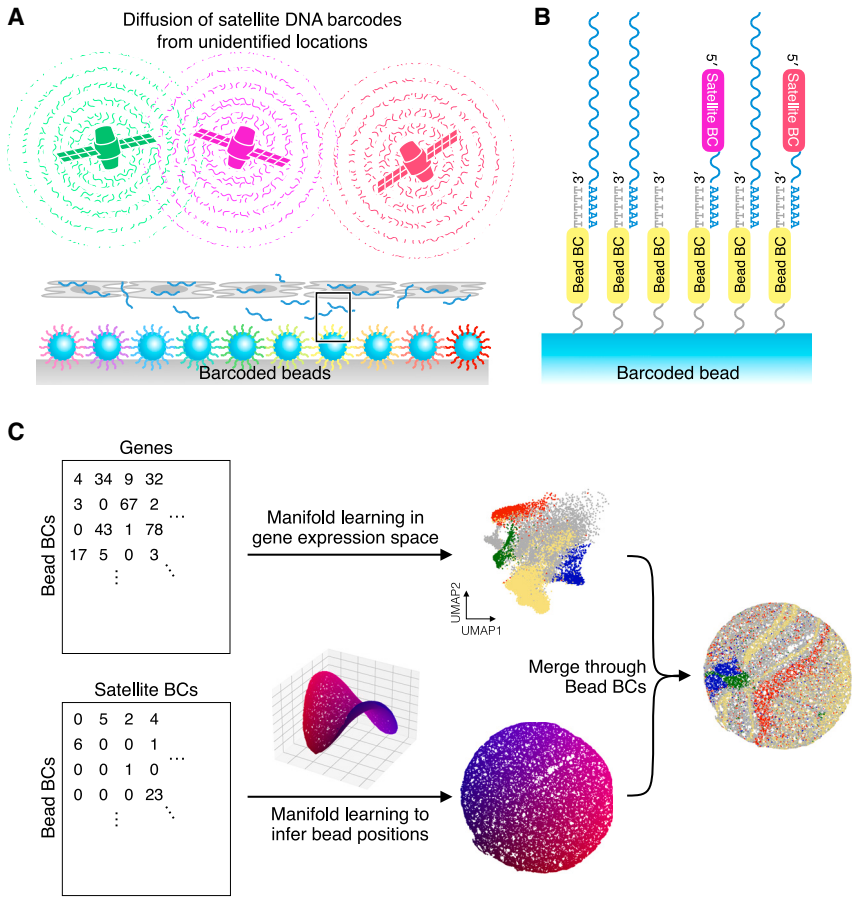


Figure 2. DNA-global positioning system (DNA-GPS)

(A) Conceptual diagram of applying DNA-GPS to locate spatial positions of Slide-seq barcoded beads. DNA-GPS infers the spatial coordinates of beads having unique bead barcodes (BCs). Satellite barcodes (BCs) released from satellite devices are concatenate to bead BCs. The satellite BC counts fused to each bead BC reflect proximities between the bead and satellite devices and can be used to infer the locations of bead BCs. Poly(A)ed RNAs released from a tissue section are also captured, reverse transcribed and concatenated to bead BCs so that the transcription profiles can be mapped onto the reconstructed space.

(B) Hypothetical design to fuse satellite BCs to bead BCs together with poly(A)ed RNAs (a close-up view of A). The poly(T) primers encoding bead BCs capture both poly(A)ed RNAs from nearby cells and poly(A)-tailed satellite (sBC) from nearby satellite devices. Reverse transcription and polymerase extension fuse RNA and satellite BC sequences to the bead BCs.

(C) Proposed workflow to obtain spatial transcriptomics image using only sequencing data. Each bead BC is associated with two count profiles: a gene expression profile by RNA sequence counts and a satellite BC count profile. The gene expression profiles are stored as rows of a gene expression matrix (top) and the satellite BC count profiles are stored as rows of a satellite BC count matrix (bottom). Because bead BCs that are nearby in physical space collect similar counts of satellite BCs, we can use the similarities between satellite BC profiles to reconstruct locations of bead BCs with manifold learning. Gene expression profiles of bead BCs are then mapped to the spatial positions.

Figure S1). The method is not limited to 2D reconstructions: if a method were devised to distribute beads and satellites in 3D, then manifold learning could be applied to reconstruct the 3D positions. In summary, DNA-GPS uses manifold learning to learn the low-dimensional embedding corresponding to the original bead positions from the high-dimensional vectors of satellite barcode counts.

Manifold learning methods aim to find a low-dimensional surface, or manifold, embedded in a high-dimensional space. They can be thought of as a subset of dimensionality reduction methods, with the stronger assumption of an underlying low-dimensional structure. For example, t-distributed stochastic neighbor embedding (t-SNE)[84,85] and uniform manifold approximation and projection (UMAP)[86] are frequently used in scRNA-seq to create visualizations of high-dimensional transcription profiles[88] (Figure 2C). In these visualizations, the low-dimensional structure is an abstract "manifold of cell states." Manifold learning has great potential to recover the very concrete manifolds of underlying 2D or 3D positions of cells or molecules in space from higher dimensional data, such as molecular interactions. In the Slide-seq example, the goal is to recover the positions of bead barcodes on a 2D surface from the satellite barcode counts. A large number of manifold learning algorithms exist that vary in their assumptions, limiting what structures they can recover, and their computational complexity (Box 3). In DNA-GPS, we use local distances between positional barcodes to reconstruct their positions because distances plateau for points that share no reads from a common satellite. This results in a non-linear manifold structure. As we require the method to scale to hundreds of thousands or millions of beads, we focused on force-based methods over spectral approaches.

RESULTS

Simulation design

We performed extensive simulations to test the feasibility of DNA-GPS and to guide the design of satellite barcode systems. For our initial simulations, we modeled satellite barcodes as diffusing from point-sources, producing Gaussian profiles (Figure 3A). In subsequent simulations, we tested other non-Gaussian satellite profiles, motivated by possible experimental implementations of satellite barcodes. We modeled positional barcodes with Slide-seq beads on which satellite barcodes are captured. We simulated bead positions using (1) ground-truth positions from the Slide-seq datasets[20] as well as (2) regular grids of hundreds of thousands of densely packed 10-μm beads to test scalability. Each bead captures reads from nearby satellite barcodes, where the number of reads is proportional to the intensity of the satellite barcode's Gaussian profile at the bead's position. We tested the robustness of image reconstruction to multiple parameters including sequencing depth and diffusion level of satellite barcodes.

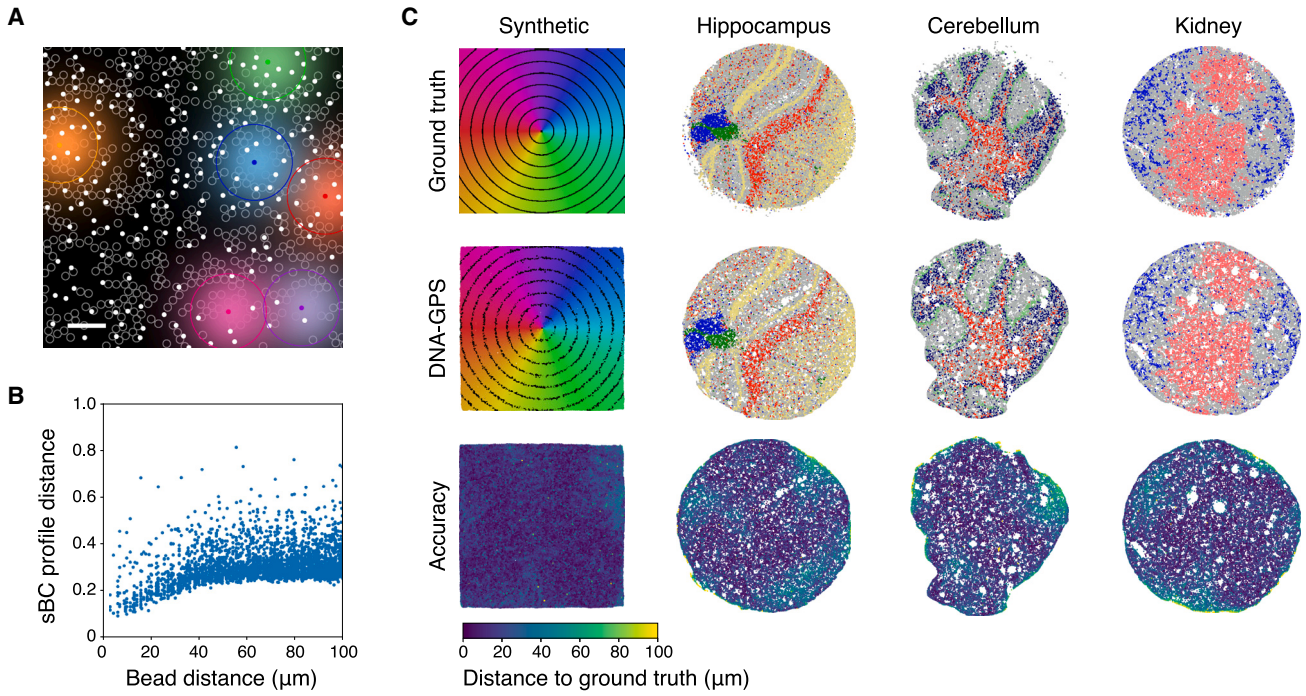


Figure 3. Simulation of spatial transcriptomics by DNA-GPS

(A) Simulation of satellite devices (filled dots) distributed over barcoded beads (white open circles; 10 μm diameter) based on the density observed in a Slide-seq dataset. Arbitrarily selected satellites are colored with the Gaussian diffusion of their satellite BCs (blurred pattern) with a standard deviation σ of 50 μm represented by radius circles of the corresponding colors. Scale bars: 50 μm.

(B) Correlation between ground-truth distance and Euclidean distance between normalized satellite BC count profiles for 10^5 randomly selected pairs of beads from a synthetic square dataset (C). While the distances are initially correlated, the sBC profile distance plateaus once the two beads share no satellites in common. This motivates the use of a manifold learning method that preserves local distances.

(C) DNA-GPS reconstructions on a simulated dataset with 10^5 densely packed beads and three sets of ground-truth bead positions from the Slide-seq dataset. The top row shows ground-truth positions artificially colored for the simulated data and colored by cell states for the Slide-seq datasets. The middle row shows the DNA-GPS reconstruction using the same coloring as the top row. The bottom row shows the DNA-GPS reconstruction where each bead is colored by the distance between the ground truth and reconstructed position after alignment through a simple linear transformation. All reconstructions achieve a median reconstruction distance below 20 μm using 100,000 satellites/cm^2, 50 μm diffusion, and 130 UMIs per bead (556 total reads per bead).

In practice, beads can vary in quality, with higher quality beads capturing more reads and vice versa. To mimic an experimental distribution of bead qualities, we used the distribution of read per bead counts from the Slide-seq kidney dataset, downsampling to achieve lower sequencing depths (Figures S2A and S2B). Further, multiple satellites may share the same barcode, due to finite satellite barcode complexity and the non-uniform distribution of barcodes, and reads from satellites sharing the same barcode are indistinguishable (Figure S2C). We accounted for this in our simulations by randomly selecting a barcode for each satellite device from a pool of possible barcodes according to an experimental distribution of barcode frequencies (Figure S2C). This corresponds mathematically to a random projection in satellite barcode space. However, random projections are known to preserve pairwise distances in high dimensions.[89] Indeed, when we examined the Euclidean distance between beads in "projected" satellite barcode count space, we found that beads that are spatially close together share similar satellite barcode vectors and were close together in high-dimensional space (Figure 3B). This correlation understandably dropped off at large-scale distances because satellite barcodes only diffuse to a limited extent and any pair of beads that share no satellites in common are roughly the same distance apart in the satellite barcode count space.

We found that the UMAP algorithm[86] can leverage local distances in satellite barcode count space to reconstruct the positions of positional barcodes at 10–30 μm resolution (Figure 3C). Although UMAP is known to not exactly preserve structures,[90] we demonstrate in simulations that it could effectively reconstruct bead barcode positions at the near-single-cell resolution. We used a linear transformation to align the output of UMAP to the ground truth for quantitative comparison, as the output of UMAP may be scaled, rotated, or flipped relative to the original embedding, and computed the distance of each reconstructed position to the ground truth. However, in practice the reconstructions would not need to be aligned. For both the Slide-seq bead positions and grids of densely packed beads, we performed a grid of simulations testing all combinations of 6 diffusion levels, 4 satellite barcode densities, 11 sequencing depths, and 4 UMAP hyperparameter settings (STAR Methods).

Simulation results

In our simulations on densely packed beads, DNA-GPS achieved reconstructions with as high as 10-μm resolution,

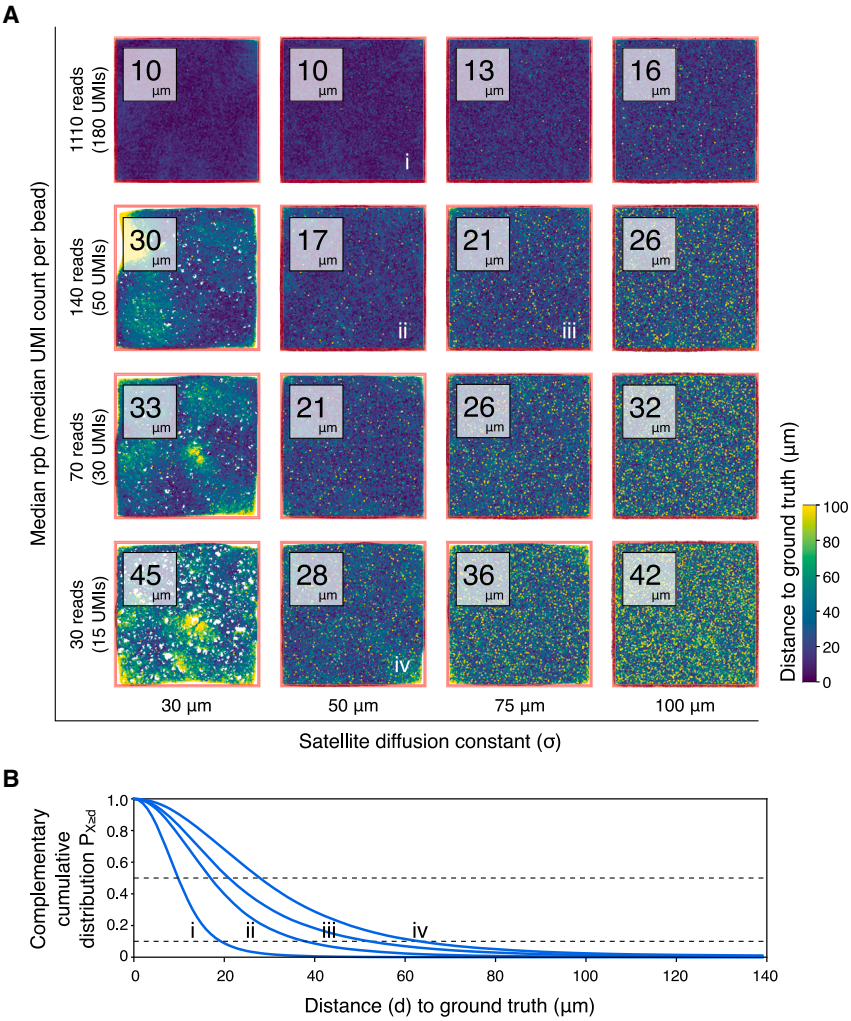


Figure 4. Single-cell resolution is possible over a range of physical parameters

(A) The best reconstruction across satellite densities and UMAP parameters on the synthetic dataset for each combination of sequencing depth and diffusion level. Images show the distance to the ground truth for each bead with the median distance inset. DNA-GPS can achieve reconstructions with as high as 10 μm resolution, with the flexibility to trade a small decrease in resolution for a large decrease in sequencing depth. The reconstructions were successful over a wide range of sequencing depths and diffusion levels.

(B) Complement of the cumulative distribution of bead distances to ground truth for each of the four reconstructions (i–iv) indicated in (A). Dashed lines indicate the 50th and 90th percentiles of beads. These distributions showed that, while we reported median values for reconstruction resolution, there were not many outlier beads that the method could not place, with >90% of beads within twice the median distance.

See also Figures S2–S7.

measured as median distance of a bead in the aligned reconstruction to the ground truth, with >90% of beads within 20 μm (Figures 4A and 4B). For the Slide-seq simulations, the best reconstructions achieved a resolution of 15 μm at as low as 130 UMI (556 reads per bead [rpb]) (Figure S9). As our model depends on UMI counts, while sequencing cost depends on reads, we provide both values in our results (Figures S5, S6, and S9).

On the dense grids of beads, DNA-GPS achieved a resolution of 10 μm at as low as 180 UMI (1,110 rpb). However, sequencing depth can be significantly reduced with only a moderate decrease in resolution, with DNA-GPS achieving reconstructions under 20 μm at as few as 40 UMI (90 rpb) and 30 μm resolution with as few as 15 UMIs per bead (30 rpb).

Further, the reconstructions are robust to variation in physical parameters and non-Gaussian diffusion (Figures S4–S9). DNA-GPS achieved reconstructions under 20 μm for all but one combination of parameters with diffusion ≥30 μm and all satellite barcode densities, requiring no more than 180 UMI (1,110 rpb) (Figure S4). Similarly, 30 μm resolution could be achieved for all but one combination of parameters at no more than 85 UMI (280 rpb). For the Slide-seq positions, DNA-GPS achieved near-single-cell resolution (<30 μm) at all diffusion levels ≥30 μm and all satellite barcode density levels at read depths as low as 25 UMI (50 rpb). Further, simulating diffusion from bacterial colonies (STAR Methods), we found that the simulations were robust to non-Gaussian diffusion, achieving peak resolution of 13 and 30 μm resolution at as low as 25 UMI (50 rpb) (Figure S8).

Our simulations suggest DNA-GPS could achieve resolutions as high as 10 μm, with the flexibility to trade a moderate decrease in resolution for a significant decrease in required sequencing depth. Using experimental values for ground-truth bead positions, satellite barcode redundancy, and bead quality, we found that our method can achieve near-single-cell resolution over satellite densities from 25,000to 250,000 satellite barcodes per cm^2, and diffusion levels from 30 to 100 μm, with the best performance occurring with 50 μm diffusion. These results can guide the development of experimental systems of satellite barcodes and demonstrate the method is easily scalable to hundreds of thousands of beads.

Comparison to other spatial genomics technologies

We next sought to compare the potential resolution and scalability of DNA-GPS to other state-of-the-art spatial transcriptomics technologies. Using our simulation results and an exponential model, we modeled the performance of DNA-GPS as a function of how deeply we sequence satellite barcodes, as we are able to localize beads more precisely as barcodes are sequenced more deeply (Figure 5A). Using this model, we estimated the resolution (μm) and field-of-view (image width of a square in mm) achieved by DNA-GPS for different sequencing depths (of satellite barcodes) and compared them with seven leading methods (Figure 5B). Intuitively, the number of rpb varies along each curve.

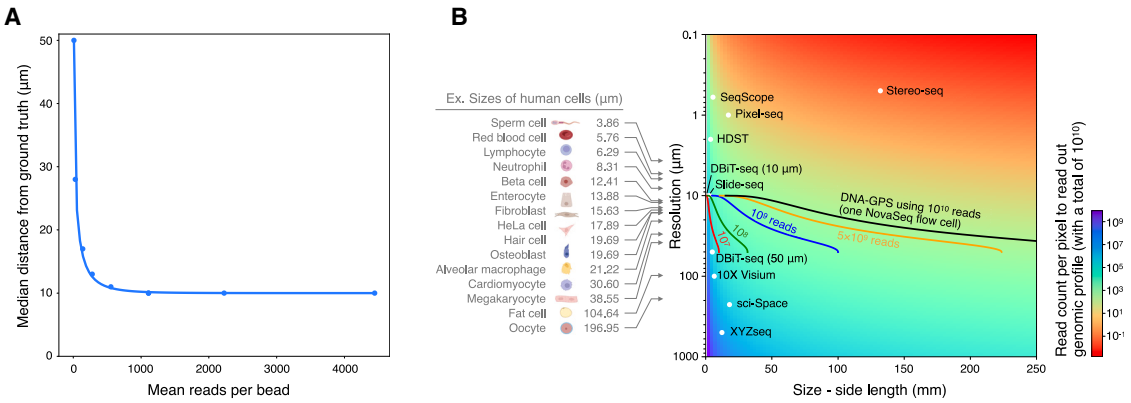


Figure 5. Comparison of DNA-GPS applied to Slide-seq with other spatial transcriptomics technologies

(A) Best median reconstruction resolution versus mean reads per bead (rpb) obtained from the simulated data sweep with an exponential line of best fit.

(B) Performance of DNA-GPS and current spatial transcriptomics technologies. Lines indicate interpolated DNA-GPS performance estimated for different read depths for sequencing satellite BCs using the curve fit in (A) and assuming densely packed 10 μm beads. For DNA-GPS, resolution is given by the median distance to the ground truth. For the other technologies (10× Visium,[15] DBiT-seq,[16] HDST,[19] Slide-seq,[20] Seq-Scope,[21] Pixel-seq,[22] Stereo-seq,[23] XYZeq,[17] and sci-Space[18]), resolution is given by the distance between adjacent pixels (or beads) to capture transcriptomes and values are shown based on the largest experimental demonstration in their respective publications. The background color represents read count per pixel (or bead) to readout transcriptome (when a total read count of 10^{10} is given). Cellular sizes were estimated as spherical diameters of their volumes collected from BioNumbers.[76]

With 10 billion reads (one NovaSeq run), DNA-GPS can localize 10^8 beads to generate a 10 cm by 10 cm image at the resolution of individual cells (17 μm).

Current large-scale RNA-seq studies often profile 10^6 cells at depths of 5,000 reads per cell.[46] Sequencing an additional 500 reads per cell, our simulations suggest that DNA-GPS could reconstruct cell positions as accurately as 10 μm resolution. More generally, the cost of DNA-GPS can be thought of as a percent increase in the original sequencing budget that depends on the desired sequencing depth for the transcriptome and desired resolution, where near-single-cell resolution reconstructions can be achieved by increasing the budget by <10%. As sequencing costs drop and studies are able to profile larger cell populations, DNA-GPS could continue to scale as a small overhead on the original sequencing budget.

Toward an experimental implementation of DNA-GPS

An experimental implementation of DNA-GPS could be achieved by multiple approaches, where the key components are the following.

Satellite formation: a large number of satellite barcodes should be densely distributed to cover the target space, in which satellite barcodes of a unique sequence are produced in a single point source. Our simulations suggest there can be some level of barcode redundancy, consistent with experimentally observed levels (STAR Methods).

Satellite transmission: When transmitting satellite barcodes to the tissue sample, some diffusion is beneficial so that spatial beads (or pixels) receive satellite barcodes from multiple neighboring satellites. Our simulations suggest the optimal diffusion level is ~50 μm for densely packed 10 μm beads (Figure 4A).

Information anchoring: Satellite barcodes must be captured together with genomic material from cells, and both are fused to positional barcodes. This information anchoring allows a spatial genomics image to be formed after satellite barcodes are used to localize positional barcodes.

There are multiple potential implementations for each of these components, which are based on existing methods (Figure 6A). For example, the satellite formation step could be achieved by culturing engineered bacterial colonies expressing unique satellite barcodes or by forming "polonies" of satellite barcodes through in-gel PCR (Box 4). Satellite barcodes could then be transmitted to the surface of positional barcodes or tissue by either free diffusion or electrophoresis blotting, where diffusion level could be tuned during this step. Finally, spatially distributed satellite barcodes and genomic materials could be anchored to either positional barcodes of pixel-based spatial genomics, or single-cell barcodes of single-cell genomics.

Thus, one potential end-to-end implementation could stamp *E. coli* colonies expressing unique, poly(A)ed satellite RNA barcodes onto a surface of Slide-seq barcoded beads of unidentified locations, followed by the deposition of tissue sample (Figure 6B). Satellite barcodes and local transcriptomes would then be captured by positional barcodes on the beads and sequenced together with the positional barcode sequences. Spatial transcriptome profiles would then be associated with barcoded beads' positions (which are identified using satellite barcode profiles by DNA-GPS).

To highlight the flexibility of DNA-GPS, we present a second potential implementation. Instead of using beads, satellite barcodes could be directly deposited onto tissue, potentially by using electrophoresis to improve transfer (Figure 6C).The tissue is then applied to the paper, where single-stranded satellite barcode molecules stick to cells (and nuclei).[18] After dissociation of the satellite-attached tissue into single cells, any cellular indexing single-cell genomics methods (e.g., split-and-pooling and droplet encapsulation) can conjugate single-cell RNAs (or DNAs) and their associated satellite barcode profiles to the single-cell barcodes, enabling scalable spatial genomics with

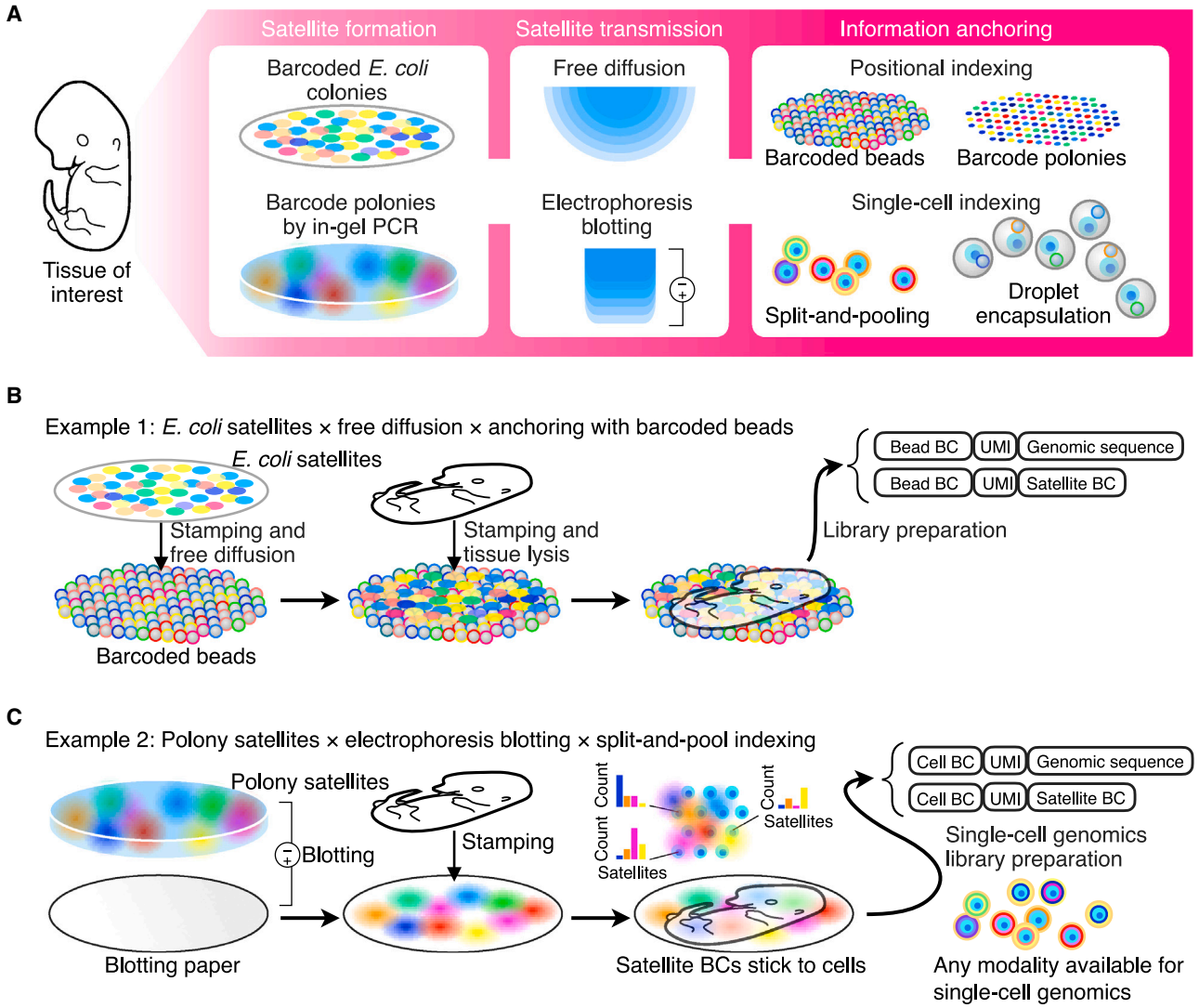


Figure 6. Possible implementations of DNA-GPS for diverse modalities of spatial genomics

(A) The potential implementations of DNA-GPS. Any DNA-GPS implementation requires three components: (1) satellite formation, (2) satellite transmission and (3) information anchoring. We envision that satellite formation could be implemented by either engineering bacterial colonies each expressing unique satellite BCs or seeding a gel with unique satellite BC DNA molecules to form polonies (see Box 4). Satellite transmission could be achieved by either free diffusion or electrophoresis blotting. Finally, information anchoring could be implemented by either capturing RNA on positional ID clusters such as barcoded beads or with standard single-cell genomics methods.

(B) Example 1. *E. coli* colonies serve as satellite devices and barcoded beads of unidentified bead BCs serve as image pixels. The colonies and then the tissue sample are stamped onto the beads.

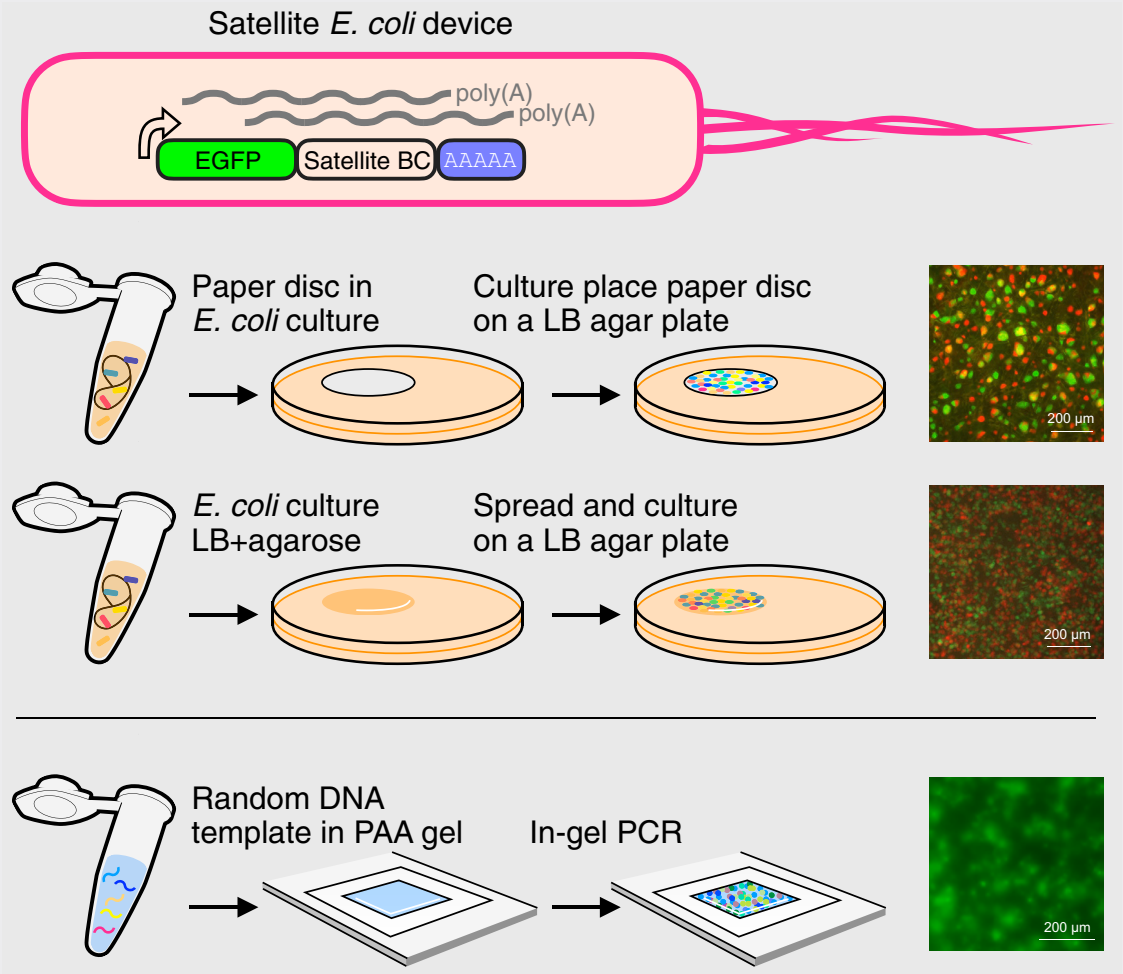
(C) Example 2. Diffused satellite barcodes can be directly deposited to a target tissue section, where single-stranded satellite BCs stick to cells (or nuclei). After dissociation of the satellite-attached tissue into single cells, any of the single-cell genomics platforms can conjugate single-cell genomic materials and their satellite BCs to the single-cell identifiers (cell BCs), enabling scalable spatial transcriptomics at the resolution of individual cells.

single-cell resolution. Theoretically, any modality of high-coverage single-cell genomics could be transformed into spatial genomics with this strategy, as long as satellite barcodes can be captured by single-cell identifiers together with genomic material of interest.

DISCUSSION

The past decade has witnessed the development of numerous spatial genomics technologies that can be broadly divided into optical imaging, positional indexing, and mathematical cartography methods. Optical imaging technologies can profile the 3D distribution of molecules across cells or tissues, but the use of optics introduces fundamental trade-offs between field-of-view, resolution, and the number of molecular species to be profiled. Positional indexing technologies have scaled 2D spatial genomics to the level of whole organs and bodies. However, *a priori* approaches are limited in scale, and *a posteriori* approaches require optical *in situ* sequencing to retrospectively identify positional barcodes. Both strategies limit the specimens

Box 4. Potential implementations to realize satellite barcodes



BACTERIAL COLONIES

Sparsely spread bacterial cells form clonal colonies on an agar gel surface. This phenomenon has widely been adopted in molecular cloning to isolate clonal DNA plasmid products. We propose the use of spatially distributed *Escherichia coli* colonies as an option to produce many locally distributed satellite barcodes, where each *E. coli* cell serves as a satellite device to produce poly(A)ed satellite RNA barcodes from its plasmid DNA. Upon cell lysis, the unique satellite RNA barcode signals are diffused from the positions of their source colonies. The challenge is making uniform sizes of colonies as small and dense as possible. We found that such colonies could be generated by placing a filter paper soaked with a barcoded cell culture on an agar media and culturing overnight. We also found that cells embedded in a soft agarose with culture media yielded spatially segregated, tiny colonies.

POLONIES

PCR amplification of sparsely distributed DNA molecules embedded in polyacrylamide gel confers polymerase colonies (polonies). Clonal DNA amplicons are produced at proximal but not identical positions to their templates and therefore form diffused patterns from the original positions along with PCR cycles. Since each polony can consist of unique clonal DNA barcode molecules, polonies can also be used as satellite devices. Overall size and density of polonies are controllable by with gel concentrations, PCR conditions and the number of template molecules.

to 2D tissue slices because there has been no efficient approach to acquire z sections by stamping tissue slices on a surface of positional barcodes. 3D spatial transcriptomics has been achieved by *in situ* sequencing of a small tissue block. However, this approach is time consuming. Indeed, ExSeq takes 10 h to localize spatial barcodes over 1 field of view (around 0.01 mm^2) across 150 z sections.[14] In addition, all the *in situ*-sequencing-based approaches cannot take advantage of advances in non-optical sequencing technologies. While computational cartography technologies offer the possibility of localizing positions from sequencing data alone, they have yet to achieve large-scale demonstrations. DNA-GPS is a theoretical framework for optics-free mathematical cartography, which allows computational localization of positional barcodes by applying manifold learning to satellite barcodes.

One of the biggest challenges to implementing DNA-GPS would be controlling the satellite barcode diffusion pattern. While we only found small decreases in resolution for non-Gaussian diffusion, we found these reconstructions were more sensitive to physical parameters such as satellite device density and diffusion distance (Figure S8). Further, it is possible that chaotic liquid flows could be induced by stamping satellite barcodes onto a tissue or onto positional barcodes. These chaotic flows could make reconstruction impossible because satellite barcodes would be distributed incoherently across positional barcodes. It is worth noting that chaotic flow could potentially affect other spatial omics technologies that stamp a tissue slice on an array of pixels because they also capture mRNA molecules in a buffer. Although several studies investigated the effect of molecular diffusion, most of them have only assessed local diffusion profiles by the distance between observed and expected locations, and the potential perturbation by the chaotic liquid flow remains to be addressed.[20,23] Thus, understanding the nature of the diffusion and optimizing the transfer protocol for better reconstruction by DNA-GPS would also benefit spatial omics technology development.

While we have presented DNA-GPS as applied to transcriptomic sequencing, the core framework could also be applied to other modalities. As seen in the recent extensions of spatial technologies to various genomic modalities,[6–8,16,24,55] the satellite barcode approach of DNA-GPS should be adaptable to most omics modalities at scale with proper experimental design. For example, an ATAC-seq extension would be possible by treating tissue genomic DNA with Tn5 transposases so that the genomic fragments from open chromatin regions are captured by the anchor barcodes. Further, while positional indexing approaches are restricted to profiling flat, 2D slices of tissues, DNA-GPS could in principle profile 3D volumes in an unbiased way, if one could distribute satellite barcodes uniformly over the volume. A somewhat easier starting point might be curved 2D surfaces, which could be achieved by flowing satellite barcodes over the surface an organ or through the interior of the network of blood vessels or lymph ducts could provide a window into full organisms by flowing satellite barcodes through these natural conduits.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2023.08.005.

ACKNOWLEDGMENTS

We thank members of the Schiebinger lab and Yachie lab for valuable discussions and critical assessment of the work, especially Rebecca Bonham-Carter for the help with the theorem proof. This study was supported by the Pilot Innovation Fund program (PIF003) of the Genome British Columbia (Genome BC) (to N.Y. and G.S.), the Canada Research Chair program by the Canadian Institutes for Health Research (CIHR) (to N.Y.), the Canada Foundation for Innovation (CFI) (to N.Y.), the Multiscale Human Program of the Canadian Institute for Advanced Research (CIFAR) (to N.Y.), the Allen Distinguished Investigator Award (to N.Y.), and the Career Award at the Scientific Interface from the Burroughs Wellcome Fund (to G.S.). Y.K. and S.I. were supported by the Japan Society for the Promotion of Science (JSPS) Research Fellowships. The high-throughput sequencing data analysis was partly performed with the SHIROKANE Supercomputer at the University of Tokyo Human Genome Center.

AUTHOR CONTRIBUTIONS

N.Y. and Y.K. conceived the high-level concept of DNA-GPS. G.S., L.G., A.A., Y.K., and M.H. conceived the implementation. G.S. and N.Y. designed the study. L.G., A.A., Y.K., and M.H. performed the analyses. S.I. and S.K. supported and performed the generation satellite barcodes. L.G., A.A., Y.K., N.Y., and G.S. wrote the manuscript.

DECLARATION OF INTERESTS

L.G., A.A., Y.K., M.H., N.Y., and G.H. have filed a provisional patent related to DNA-GPS.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: April 27, 2022
Revised: April 19, 2023
Accepted: August 25, 2023
Published: September 25, 2023

REFERENCES

1. Svensson, V., Vento-Tormo, R., and Teichmann, S.A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. Nat. Protoc. *13*, 599–604.

2. Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods *14*, 865–868. https://doi.org/10.1038/nmeth.4380.

3. Gawad, C., Koh, W., and Quake, S.R. (2016). Single-cell genome sequencing: current state of the science. Nat. Rev. Genet. *17*, 175–188. https://doi.org/10.1038/nrg.2015.16.

4. Cusanovich, D.A., Daza, R., Adey, A., Pliner, H.A., Christiansen, L., Gunderson, K.L., Steemers, F.J., Trapnell, C., and Shendure, J. (2015). Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science *348*, 910–914. https://doi.org/10.1126/science.aab1601.

5. Rao, A., Barkley, D., França, G.S., and Yanai, I. (2021). Exploring tissue architecture using spatial transcriptomics. Nature *596*, 211–220. https://doi.org/10.1038/s41586-021-03634-9.

6. Deng, Y., Bartosovic, M., Ma, S., Zhang, D., Kukanja, P., Xiao, Y., Su, G., Liu, Y., Qin, X., Rosoklija, G.B., et al. (2022). Spatial profiling of chromatin accessibility in mouse and human tissues. Nature *609*, 375–383. https://doi.org/10.1038/s41586-022-05094-1.

7. Liu, Y., DiStasio, M., Su, G., Asashima, H., Enninful, A., Qin, X., Deng, Y., Nam, J., Gao, F., Bordignon, P., et al. (2023). High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial CITE-seq. Nat. Biotechnol. https://doi.org/10.1038/s41587-023-01676-0.

8. Deng, Y., Bartosovic, M., Kukanja, P., Zhang, D., Liu, Y., Su, G., Enninful, A., Bai, Z., Castelo-Branco, G., and Fan, R. (2022). Spatial-CUT&Tag: spatially resolved chromatin modification profiling at the cellular level. Science *375*, 681–686. https://doi.org/10.1126/science.abg7216.

9. Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S., and Zhuang, X. (2015). RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science *348*, aaa6090. https://doi.org/10.1126/science.aaa6090.

10. Eng, C.L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G.C., et al. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. Nature *568*, 235–239. https://doi.org/10.1038/s41586-019-1049-y.

11. Borm, L.E., Mossi Albiach, A., Mannens, C.C.A., Janusauskas, J., Özgün, C., Fernández-García, D., Hodge, R., Castillo, F., Hedin, C.R.H., Villablanca, E.J., et al. (2023). Scalable in situ single-cell profiling by electrophoretic capture of mRNA using EEL FISH. Nat. Biotechnol. *41*, 222–231. https://doi.org/10.1038/s41587-022-01455-3.

12. Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Yang, J.L., Ferrante, T.C., Terry, R., Jeanty, S.S., Li, C., Amamoto, R., et al. (2014). Highly multiplexed subcellular RNA sequencing in situ. Science *343*, 1360–1363. https://doi.org/10.1126/science.1250212.

13. Wang, X., Allen, W.E., Wright, M.A., Sylwestrak, E.L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., et al. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science *361*. https://doi.org/10.1126/science.aat5691.

14. Alon, S., Goodwin, D.R., Sinha, A., Wassie, A.T., Chen, F., Daugharthy, E.R., Bando, Y., Kajita, A., Xue, A.G., Marrett, K., et al. (2021). Expansion sequencing: spatially precise in situ transcriptomics in intact biological systems. Science *371*. https://doi.org/10.1126/science.aax2656.

15. Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science *353*, 78–82. https://doi.org/10.1126/science.aaf2403.

16. Liu, Y., Yang, M., Deng, Y., Su, G., Enninful, A., Guo, C.C., Tebaldi, T., Zhang, D., Kim, D., Bai, Z., et al. (2020). High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. Cell *183*, 1665–1681.e18. https://doi.org/10.1016/j.cell.2020.10.026.

17. Lee, Y., Bogdanoff, D., Wang, Y., Hartoularos, G.C., Woo, J.M., Mowery, C.T., Nisonoff, H.M., Lee, D.S., Sun, Y., Lee, J., et al. (2021). XYZeq: spatially resolved single-cell RNA sequencing reveals expression heterogeneity in the tumor microenvironment. Sci. Adv. *7*. https://doi.org/10.1126/sciadv.abg4755.

18. Srivatsan, S.R., Regier, M.C., Barkan, E., Franks, J.M., Packer, J.S., Grosjean, P., Duran, M., Saxton, S., Ladd, J.J., Spielmann, M., et al. (2021). Embryo-scale, single-cell spatial transcriptomics. Science *373*, 111–117. https://doi.org/10.1126/science.abb9536.

19. Vickovic, S., Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Schapiro, D., Äijö, T., Bonneau, R., Bergenstråhle, L., Navarro, J.F., et al. (2019). High-definition spatial transcriptomics for in situ tissue profiling. Nat. Methods *16*, 987–990. https://doi.org/10.1038/s41592-019-0548-y.

20. Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., Welch, J., Chen, L.M., Chen, F., and Macosko, E.Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science *363*, 1463–1467. https://doi.org/10.1126/science.aaw1219.

21. Cho, C.S., Xi, J., Si, Y., Park, S.R., Hsu, J.E., Kim, M., Jun, G., Kang, H.M., and Lee, J.H. (2021). Microscopic examination of spatial transcriptome using Seq-Scope. Cell *184*, 3559–3572.e22. https://doi.org/10.1016/j.cell.2021.05.010.

22. Fu, X., Sun, L., Dong, R., Chen, J.Y., Silakit, R., Condon, L.F., Lin, Y., Lin, S., Palmiter, R.D., and Gu, L. (2022). Polony gels enable amplifiable DNA stamping and spatial transcriptomics of chronic pain. Cell *185*, 4621–4633.e17. https://doi.org/10.1016/j.cell.2022.10.021.

23. Chen, A., Liao, S., Cheng, M., Ma, K., Wu, L., Lai, Y., Qiu, X., Yang, J., Xu, J., Hao, S., et al. (2022). Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. Cell *185*, 1777–1792.e21. https://doi.org/10.1016/j.cell.2022.04.003.

24. Zhao, T., Chiang, Z.D., Morriss, J.W., LaFave, L.M., Murray, E.M., Del Priore, I., Meli, K., Lareau, C.A., Nadaf, N.M., Li, J., et al. (2022). Spatial genomics enables multi-modal study of clonal heterogeneity in tissues. Nature *601*, 85–91. https://doi.org/10.1038/s41586-021-04217-4.

25. Weinstein, J.A., Regev, A., and Zhang, F. (2019). DNA microscopy: optics-free spatio-genetic imaging by a stand-alone chemical reaction. Cell *178*, 229–241.e16. https://doi.org/10.1016/j.cell.2019.05.019.

26. Hoffecker, I.T., Yang, Y., Bernardinelli, G., Orponen, P., and Högberg, B. (2019). A computational framework for DNA sequencing microscopy. Proc. Natl. Acad. Sci. USA *116*, 19282–19287. https://doi.org/10.1073/pnas.1821178116.

27. Boulgakov, A.A., Xiong, E., Bhadra, S., Ellington, A.D., and Marcotte, E.M. (2018). From space to sequence and back again: iterative DNA proximity ligation and its applications to DNA-based imaging. Preprint at bioRxiv. https://doi.org/10.1101/470211.

28. Nitzan, M., Karaiskos, N., Friedman, N., and Rajewsky, N. (2019). Gene expression cartography. Nature *576*, 132–137. https://doi.org/10.1038/s41586-019-1773-3.

29. Jensen, E. (2014). Technical review: in situ hybridization. Anat. Rec. (Hoboken) *297*, 1349–1353. https://doi.org/10.1002/ar.22944.

30. Cui, C., Shu, W., and Li, P. (2016). Fluorescence in situ hybridization: cell-based genetic diagnostic and research applications. Front. Cell Dev. Biol. *4*, 89. https://doi.org/10.3389/fcell.2016.00089.

31. Chen, F., Tillberg, P.W., and Boyden, E.S. (2015). Optical imaging. Expansion microscopy. Science *347*, 543–548. https://doi.org/10.1126/science.1260088.

32. Gao, R., Asano, S.M., and Boyden, E.S. (2017). Q&A: expansion microscopy. BMC Biol. *15*, 50. https://doi.org/10.1186/s12915-017-0393-3.

33. Takei, Y., Zheng, S., Yun, J., Shah, S., Pierson, N., White, J., Schindler, S., Tischbirek, C.H., Yuan, G.C., and Cai, L. (2021). Single-cell nuclear architecture across cell types in the mouse brain. Science *374*, 586–594. https://doi.org/10.1126/science.abj1966.

34. Takei, Y., Yun, J., Zheng, S., Ollikainen, N., Pierson, N., White, J., Shah, S., Thomassie, J., Suo, S., Eng, C.L., et al. (2021). Integrated spatial genomics reveals global architecture of single nuclei. Nature *590*, 344–350. https://doi.org/10.1038/s41586-020-03126-2.

35. Payne, A.C., Chiang, Z.D., Reginato, P.L., Mangiameli, S.M., Murray, E.M., Yao, C.C., Markoulaki, S., Earl, A.S., Labade, A.S., Jaenisch, R., et al. (2021). In situ genome sequencing resolves DNA sequence and structure in intact biological samples. Science *371*. https://doi.org/10.1126/science.aay3446.

36. Tainaka, K., Kubota, S.I., Suyama, T.Q., Susaki, E.A., Perrin, D., Ukai-Tadenuma, M., Ukai, H., and Ueda, H.R. (2014). Whole-body imaging with single-cell resolution by tissue decolorization. Cell *159*, 911–924. https://doi.org/10.1016/j.cell.2014.10.034.

37. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-seq whole-transcriptome analysis of a single cell. Nat. Methods *6*, 377–382. https://doi.org/10.1038/nmeth.1315.

38. Ramsköld, D., Luo, S., Wang, Y.C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I., Loring, J.F., Laurent, L.C., et al. (2012). Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. Nat. Biotechnol. *30*, 777–782. https://doi.org/10.1038/nbt.2282.

39. Picelli, S., Björklund, Å.K., Faridani, O.R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat. Methods *10*, 1096–1098. https://doi.org/10.1038/nmeth.2639.

40. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002.

41. Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201. https://doi.org/10.1016/j.cell.2015.04.044.

42. Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049. https://doi.org/10.1038/ncomms14049.

43. Datlinger, P., Rendeiro, A.F., Boenke, T., Senekowitsch, M., Krausgruber, T., Barreca, D., and Bock, C. (2021). Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. Nat. Methods *18*, 635–642. https://doi.org/10.1038/s41592-021-01153-z.

44. De Rop, F.V., Ismail, J.N., Bravo González-Blas, C., Hulselmans, G.J., Flerin, C.C., Janssens, J., Theunis, K., Christiaens, V.M., Wouters, J., Marcassa, G., et al. (2022). Hydrop enables droplet-based single-cell ATAC-seq and single-cell RNA-seq using dissolvable hydrogel beads. eLife *11*. https://doi.org/10.7554/eLife.73971.

45. Clark, I.C., Fontanez, K.M., Meltzer, R.H., Xue, Y., Hayford, C., May-Zhang, A., D'Amato, C., Osman, A., Zhang, J.Q., Hettige, P., et al. (2023). Microfluidics-free single-cell genomics with templated emulsification. Nat. Biotechnol. https://doi.org/10.1038/s41587-023-01685-z.

46. Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496–502. https://doi.org/10.1038/s41586-019-0969-x.

47. Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N., Steemers, F.J., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. Science *357*, 661–667. https://doi.org/10.1126/science.aam8940.

48. Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L.T., Peeler, D.J., Mukherjee, S., Chen, W., et al. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science *360*, 176–182. https://doi.org/10.1126/science.aam8999.

49. Cao, J., O'Day, D.R., Pliner, H.A., Kingsley, P.D., Deng, M., Daza, R.M., Zager, M.A., Aldinger, K.A., Blecher-Gonen, R., Zhang, F., et al. (2020). A human cell atlas of fetal gene expression. Science *370*. https://doi.org/10.1126/science.aba7721.

50. Wu, S.J., Furlan, S.N., Mihalas, A.B., Kaya-Okur, H.S., Feroze, A.H., Emerson, S.N., Zheng, Y., Carson, K., Cimino, P.J., Keene, C.D., et al. (2021). Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. Nat. Biotechnol. *39*, 819–824. https://doi.org/10.1038/s41587-021-00865-z.

51. Bartosovic, M., Kabbe, M., and Castelo-Branco, G. (2021). Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. Nat. Biotechnol. *39*, 825–835. https://doi.org/10.1038/s41587-021-00869-9.

52. Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature *502*, 59–64. https://doi.org/10.1038/nature12593.

53. Stickels, R.R., Murray, E., Kumar, P., Li, J., Marshall, J.L., Di Bella, D.J., Arlotta, P., Macosko, E.Z., and Chen, F. (2021). Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Nat. Biotechnol. *39*, 313–319. https://doi.org/10.1038/s41587-020-0739-1.

54. Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., et al. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science *327*, 78–81. https://doi.org/10.1126/science.1181498.

55. Zhang, D., Deng, Y., Kukanja, P., Agirre, E., Bartosovic, M., Dong, M., Ma, C., Ma, S., Su, G., Bao, S., et al. (2023). Spatial epigenome-transcriptome co-profiling of mammalian tissues. Nature *616*, 113–122. https://doi.org/10.1038/s41586-023-05795-1.

56. Boulgakov, A.A., Ellington, A.D., and Marcotte, E.M. (2020). Bringing microscopy-by-sequencing into view. Trends Biotechnol. *38*, 154–162. https://doi.org/10.1016/j.tibtech.2019.06.001.

57. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502. https://doi.org/10.1038/nbt.3192.

58. Ginzberg, M.B., Kafri, R., and Kirschner, M. (2015). Cell biology. On being the right (cell) size. Science *348*, 1245075. https://doi.org/10.1126/science.1245075.

59. Ocqueteau, C., Cury, M., Becker, L., Morgado, E., González, U., Muxica, L., and Gunther, B. (1989). Three-dimensional morphometry of mammalian cells. I. Diameters. Arch. Biol. Med. Exp. *22*, 89–95.

60. Mitsui, Y., and Schneider, E.L. (1976). Relationship between cell replication and volume in senescent human diploid fibroblasts. Mech. Ageing Dev. *5*, 45–56. https://doi.org/10.1016/0047-6374(76)90007-5.

61. Finegood, D.T., Scaglia, L., and Bonner-Weir, S. (1995). Dynamics of beta-cell mass in the growing rat pancreas. Estimation with a simple mathematical model. Diabetes *44*, 249–256. https://doi.org/10.2337/diab.44.3.249.

62. Wiśniewski, J.R., Ostasiewicz, P., Duś, K., Zielińska, D.F., Gnad, F., and Mann, M. (2012). Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. Mol. Syst. Biol. *8*, 611. https://doi.org/10.1038/msb.2012.44.

63. Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J., and Aebersold, R. (2011). The quantitative proteome of a human cell line. Mol. Syst. Biol. *7*, 549. https://doi.org/10.1038/msb.2011.82.

64. Zhao, L., Kroenke, C.D., Song, J., Piwnica-Worms, D., Ackerman, J.J., and Neil, J.J. (2008). Intracellular water-specific MR of microbead-adherent cells: the HeLa cell intracellular water exchange lifetime. NMR Biomed. *21*, 159–164. https://doi.org/10.1002/nbm.1173.

65. Géléoc, G.S., Casalotti, S.O., Forge, A., and Ashmore, J.F. (1999). A sugar transporter as a candidate for the outer hair cell motor. Nat. Neurosci. *2*, 713–719. https://doi.org/10.1038/11174.

66. Calvillo, L., Latini, R., Kajstura, J., Leri, A., Anversa, P., Ghezzi, P., Salio, M., Cerami, A., and Brines, M. (2003). Recombinant human erythropoietin protects the myocardium from ischemia-reperfusion injury and promotes

beneficial remodeling. Proc. Natl. Acad. Sci. USA *100*, 4802–4806. https://doi.org/10.1073/pnas.0630444100.

67. Harker, L.A., Roskos, L.K., Marzec, U.M., Carter, R.A., Cherry, J.K., Sundell, B., Cheung, E.N., Terry, D., and Sheridan, W. (2000). Effects of megakaryocyte growth and development factor on platelet production, platelet life span, and platelet function in healthy human volunteers. Blood *95*, 2514–2522.
68. Krombach, F., Münzing, S., Allmeling, A.M., Gerlach, J.T., Behr, J., and Dörger, M. (1997). Cell size of alveolar macrophages: an interspecies comparison. Environ. Health Perspect. *105* (*Suppl 5*), 1261–1263. https://doi.org/10.1289/ehp.97105s51261.
69. Goyanes, V.J., Ron-Corzo, A., Costas, E., and Maneiro, E. (1990). Morphometric categorization of the human oocyte and early conceptus. Hum. Reprod. *5*, 613–618. https://doi.org/10.1093/oxfordjournals.humrep.a137155.
70. Livingston, J.N., Lerea, K.M., Bolinder, J., Kager, L., Backman, L., and Arner, P. (1984). Binding and molecular weight properties of the insulin receptor from omental and subcutaneous adipocytes in human obesity. Diabetologia *27*, 447–453. https://doi.org/10.1007/BF00273909.
71. Diez-Silva, M., Dao, M., Han, J., Lim, C.T., and Suresh, S. (2010). Shape and biomechanical characteristics of human red blood cells in health and disease. MRS Bull. *35*, 382–388. https://doi.org/10.1557/mrs2010.571.
72. Gilmore, J.A., McGann, L.E., Liu, J., Gao, D.Y., Peter, A.T., Kleinhans, F.W., and Critser, J.K. (1995). Effect of cryoprotectant solutes on water permeability of human spermatozoa. Biol. Reprod. *53*, 985–995. https://doi.org/10.1095/biolreprod53.5.985.
73. Schmid-Schönbein, G.W., Shih, Y.Y., and Chien, S. (1980). Morphometry of human leukocytes. Blood *56*, 866–875.
74. Ballas, S.K. (1987). Erythrocyte concentration and volume are inversely related. Clin. Chim. Acta *164*, 243–244. https://doi.org/10.1016/0009-8981(87)90078-7.
75. Rosengren, S., Henson, P.M., and Worthen, G.S. (1994). Migration-associated volume changes in neutrophils facilitate the migratory process in vitro. Am. J. Physiol. *267*, C1623–C1632. https://doi.org/10.1152/ajpcell.1994.267.6.C1623.
76. Milo, R., Jorgensen, P., Moran, U., Weber, G., and Springer, M. (2010). BioNumbers–the database of key numbers in molecular and cell biology. Nucleic Acids Res. *38*, D750–D753. https://doi.org/10.1093/nar/gkp889.
77. Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. *24*, 417–441.
78. Torgerson, W.S. (1952). Multidimensional scaling: I. Theory and method. Psychometrika *17*, 401–419.
79. Izenman, A.J. (2012). Introduction to manifold learning. Wiley Interdiscip. Rev. Comput. Stat. *4*, 439–446.
80. Roweis, S.T., and Saul, L.K. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science *290*, 2323–2326.
81. Tenenbaum, J.B., de Silva, V., and Langford, J.C. (2000). A global geometric framework for nonlinear dimensionality reduction. Science *290*, 2319–2323. https://doi.org/10.1126/science.290.5500.2319.
82. Belkin, M.a.N.P., and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. *15*, 1373–1396. https://doi.org/10.1162/089976603321780317.
83. Donoho, D.L., and Grimes, C. (2003). Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. Proc. Natl. Acad. Sci. USA *100*, 5591–5596.
84. Van Der Maaten, L. (2014). Accelerating t-SNE using tree-based algorithms. J. Mach. Learn. Res. *15*, 3221–3245.
85. Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.
86. McInnes, L., Healy, J., and Melville, J. (2018). Umap: uniform manifold approximation and projection for dimension reduction. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.03426.
87. Tschannen, M., Bachem, O., and Lucic, M. (2018). Recent advances in autoencoder-based representation learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.1812.05069.
88. Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. Published online December 3, 2018. Nat. Biotechnol. https://doi.org/10.1038/nbt.4314.
89. Cleary, B., Cong, L., Cheung, A., Lander, E.S., and Regev, A. (2017). Efficient generation of transcriptomic profiles by random composite measurements. Cell *171*, 1424–1436.e18. https://doi.org/10.1016/j.cell.2017.10.023.
90. Damrich, S., and Hamprecht, F.A. (2021). On UMAP's true loss function. Adv. Neural Inf. Process. Syst. *34*, 5798–5809.
91. Nishimasu, H., Shi, X., Ishiguro, S., Gao, L., Hirano, S., Okazaki, S., Noda, T., Abudayyeh, O.O., Gootenberg, J.S., Mori, H., et al. (2018). Engineered CRISPR-Cas9 nuclease with expanded targeting space. Science *361*, 1259–1262. https://doi.org/10.1126/science.aas9129.
92. Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One *11*, e0163962. https://doi.org/10.1371/journal.pone.0163962.
93. Kijima, Y., Evans-Yamamoto, D., Toyoshima, H., and Yachie, N. (2023). A universal sequencing read interpreter. Sci. Adv. *9*, eadd2793. https://doi.org/10.1126/sciadv.add2793.
94. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*. https://doi.org/10.1093/gigascience/giab008.
95. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421. https://doi.org/10.1186/1471-2105-10-421.
96. Umeyama, S. (1991). Least-squares estimation of transformation parameters between two point patterns. IEEE Trans. Pattern Anal. Mach. Intell. *13*, 376–380.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
Phusion High-fidelity DNA polymerase	NEB	NEB M0530
KAPA Library Quantification Kit Illumina	KAPA Bioscience	KK4824
5x Phusion HF Buffer	NEB	B0518S
dNTP	NEB	N0447
Critical commercial assays		
FastGene PCR/Gel Extraction Kit	Nippon Genetics	FG-91302
Deposited data		
Slide-seq positions	Rodriques et al.[20]	Single Cell Portal: https://singlecell.broadinstitute.org/single_cell/study/SCP354/slide-seq-study
Raw amplicon sequencing data for estimating barcode overlap	This study	NCBI BioProject: PRJNA 993711
Oligonucleotides		
SI#1275 PS1.0-PAMlib-180720-FW TAACTTACGGAGTCGCTCTACGGC CTGCAGGTCGACTCTAGAGGA	This study	N/A
SI#1274 PS2.0-PAMlib-180720-RV GGATGGGATTCTTTAGGTCCTGG TTGTAAAACGACGGCCAGTGAA	This study	N/A
Recombinant DNA		
Barcoded plasmid library	Nishimasu et al.[91]	N/A
Software and algorithms		
Samtools v.1.9.74-gf69e678	Genome Research Limited	http://www.htslib.org/
Seqkit v0.10.1	Shen et al.[92]	https://bioinf.shenwei.me/seqkit/download/
NCBI Blast+ v.2.6.0	NIH National Laboratory of Medicine	https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
INTERSTELLAR v.1.0.0	Kijima et al.[93]	https://github.com/yachielab/Interstellar
CellRanger v3.0.1	10x Genomics	https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest
R v. 4.1.1	R Project	https://cran.r-project.org/
Python v.3.6.10	Python Software Foundation	https://www.python.org/downloads/
UMAP v.0.4.2	Anaconda Org	https://anaconda.org/conda-forge/umap-learn
Other		
Illumina HiSeq 2000 sequencer	Illumina	N/A
PhiX Control v3	Illumina	FC-110-3001
Simulation codes and resources	This manuscript	GitHub: https://github.com/schiebingerlab/DNA-GPS (version of record deposited at Zenodo: https://doi.org/10.5281/zenodo.8088532)

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Geoffrey Schiebinger (geoff@math.ubc.ca).

Materials Availability

This study did not generate new materials.

Data and code availability

- Sequencing data for the barcode count distribution has been deposited to NCBI BioProject (NCBI BioProject: PRJNA993711).
- This paper analyzes existing, publicly available data. The access links to the datasets are listed in the key resources table.
- All original code has been deposited at https://github.com/schiebingerlab/DNA-GPS and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Barcode distribution data is available at https://github.com/schiebingerlab/DNA-GPS/data.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contacts upon request.

METHOD DETAILS

Data generation

To generate data, we chose ground truth positions for beads and randomly generated satellite positions extending slightly beyond the beads to ensure uniform reads. We used ground truth positions from the Slide-seq kidney, hippocampus, and cerebellum 2 datasets as well as dense grids of 100,000 beads to test scalability. For most simulations, we modelled the reads from each satellite as following a Gaussian distribution where the width of the Gaussian determined the level of diffusion (See also Method S1). However, we also performed simulations exploring non-Gaussian diffusion from simulated bacterial colonies (See Robustness to alternate diffusion schemes). In addition to sBC diffusion, the density of satellite devices was a hyperparameter for data generation.

Bead quality and barcode overlap

To sample an sBC read matrix, C, we needed to define the distributions of satellite barcode redundancy and bead quality. To account for varying bead qualities, we utilized the distribution present in the Slide-seq kidney dataset (See Figure S2B). Slide-seq datasets were downloaded from Single Cell Portal. cDNA sequences were extracted from the BAM file and stored as FASTQ format using samtools.[94] The reads were converted into 10x Chromium V3 format using INTERSTELLAR v1.0.0,[93] and the formatted FASTQ file was downsampled using seqkit.[92] Each downsampled file was mapped to the reference genome version of mm10 using CellRanger v3.0.1 to obtain a UMI count matrix. This resulted in the mean rpb to UMI curves shown in Figure S2. We found that the read to UMI curve differed between different Slide-seq samples. As our model depends on UMI while sequencing cost depends on reads, we provide both values in our results.

Barcodes do not occur uniformly in practice leading to some satellites sharing common barcodes. We quantified the extent of this overlap by conducting the following experiment. We used a barcoded plasmid library that has been previously generated.[91] To generate the high-throughput sequencing library, the plasmid pUC19 encoding eight randomized nucleotides (4^8 possible barcode combinations) was subjected to PCR amplification. The 1st PCR was performed with 20 μL reaction volume, composed of 1 μL of plasmid library, 0.5 μL each of 20 μM forward (SI#1275) and 20 μM reverse (SI#1274) primers, 0.2 μL of Phusion High-fidelity DNA Polymerase (NEB M0530), 4 μL of 5x Phusion HF Buffer (NEB #B0518S), and 2 μL of 2 mM dNTPs (NEB #N0447) with the following thermal cycle condition: 98°C for 30 s, 30 cycles of 98°C for 10 s, 60°C for 10 s, and 72°C for 60 s, and then 72°C for 5 min for the final extension. The 1st PCR product was then purified using FastGene PCR/Gel Extraction Kit (Nippon Genetics #FG-91302). For index PCR to attach Illumina P5 and P7 index sequences, the 2nd PCR was performed with 20 μL reaction volume, composed of 1 μL of plasmid library, 1 μL each of 10 μM P5 and 10 μM P7 primers, 0.2 μL of Phusion High-fidelity DNA Polymerase (NEB M0530), 4 μL of 5x Phusion HF Buffer (NEB #B0518S), and 2 μL of 2 mM dNTPs (NEB #N0447) with the following thermal cycle condition: 98°C for 30 s, 15 cycles of 98°C for 10 s, 60°C for 10 s, and 72°C for 60 s, and then 72°C for 5 min for the final extension. The 2nd PCR product was then purified using FastGene PCR/Gel Extraction Kit (Nippon Genetics #FG-91302) and pooled. The pooled sequencing sample was quantified using a KAPA Library Quantification Kit Illumina (KAPA BIOSYSTEMS #KK4824) and analyzed by paired-end sequencing using Illumina HiSeq2000 with 20% PhiX spike-in control (Illumina #FC-110-3001). The sequencing reads were demultiplexed according to the sample indices and constant sequences using NCBI Blast+ (version 2.6.0)[95] with the blast-short option. Using read alignment information, 8-mer barcodes were extracted to count each barcode abundance. The resulting barcode occurrence distribution is shown in Figure S2C. We sampled satellite barcodes from this distribution for our simulations.

Forming the sBC count matrix

To create the sBC count matrix, we first form a matrix with no barcode redundancy where for each bead we sample a bead quality from the Slide-seq kidney distribution and then sample the specified number of reads from the sBC distribution determined by the bead's position. We then generate a barcode for each satellite, following the experimental distribution, and combine columns of the initial matrix sharing a barcode. Our data generation process has three hyperparameters: (1) density of satellites, (2) diffusion level, and (3) sequencing depth. We tested a grid of four satellite densities, sBCs/cm^2 $\in$ [25000, 50000, 100000, 250000], six diffusion levels, $\sigma \in$ [10, 20, 25, 30, 50, 75, 100] μm, and 11 downsampling levels, mean rpb $\in$ [10, 30, 50, 70, 90, 140, 280, 555, 1100, 2225, 4445].

Reconstructions

UMAP is a dimensionality reduction technique that is frequently used with single-cell data to reduce a gene expression matrix to two dimensions for visualization. However, by applying it to our sBC matrix we can recover the underlying two-dimensional manifold. As UMAP allows an arbitrary distance function, we tested Euclidean distance and cosine similarity on raw counts, row-normalized, and log-tpm normalized data and found the row-normalized Euclidean distance resulted in the best reconstructions.

Hyperparameter selection

UMAP introduces two hyperparameters: (1) the number of neighbors to consider for each bead, *n_neighbors*, and (2) the minimum distance between points in the low-dimensional space, *min_dist*. We tested a grid of four values for each hyperparameter, *min_dist* $\in$ [0.25, 0.5, 0.75, 1] and *n_neighbors* $\in$ [25, 50, 75, 100], for every combination of hyperparameters from our data simulation. As we consistently achieved good reconstructions using 60 neighbors, we fixed *n_neighbors* to 60 for the figures and values reported in the text. We found the best value of *min_dist* to vary with diffusion level and sequencing depth, with a value of 0.25 performing best for 30 μm diffusion and a value of 1 performing better for diffusion $\geq$50 μm. Setting these two value for *min_dist* based only on diffusion, we achieved reconstructions are within 5 μm of best reconstruction in more than 80% of cases (See Figure S6).

Aligning reconstructions

As distances passed to UMAP are relative, the embedding may be a rotation, reflection, scaling, or translation of the ground truth. We aligned each embedding to the ground truth using the Kabsch-Umeyama algorithm,[96] which finds the rigid transformation (rotation, reflection, scaling, and translation) that minimizes the sum of distances between corresponding points in the embedding and the ground truth.

Robustness to physical parameters

We achieved mean alignment distances as low as 10 μm and could achieve 20 μm alignment distances for all but one combination of sBC density and diffusion level from 30-100 μm (see Figure S4). In these ranges, physical parameters do not limit whether a reconstruction is possible. Rather, some combinations result in higher resolution reconstructions for the same sequencing cost. While we achieved successful reconstructions down to 20 μm of diffusion, they had high alignment errors even at the highest sequencing depth (see Figure S7). We did not achieve successful reconstructions for diffusion below 20 μm. As preliminary experiments had already generated $\sim$20,000 bacterial colonies/cm^2, we did not test lower sBC densities.

While the median alignment distance is a convenient statistic to compare between reconstructions, it is also important that there are not large number of outliers with high alignment distances, corresponding to beads that cannot be accurately placed. Figure S5 shows that in most cases >90% of beads are no more than twice the median alignment distance.

Robustness to bead quality

We repeated the simulations assuming uniform bead quality. As expected, the reconstructions improved compared to the experimental bead quality distribution. However, the improvements were moderate, with the best reconstruction only improving from 10 to 7 μm and the best alignment error per UMI curves closely tracking (see Figure S9A). This suggests the method is robust to changes in the distribution of bead qualities. Further, we believe the experimentally obtained bead quality distribution is a conservative estimate of what could be achieved in practice. For example, the 10x protocol achieves a less extreme distribution than Slide-seq and Drop-seq which employ similar library preparation protocols (See Figure S9B).

Robustness to alternate diffusion schemes

As one potential implementation of satellite devices is through E. coli colonies, we wanted to understand the potential impact of non-uniform diffusion from non-point source satellite devices on reconstruction quality. To simulate bacterial colonies, we generated a 10 μm grid and randomly assigned a subset of pixels barcodes from the experimental barcode distribution, where the size of the subset determined the satellite device density and was a hyperparameter for the simulation. We then simulated the colonies by defining the probability of each pixel spreading to an adjacent empty pixel at a single step, and then repeatedly simulating steps until we reached a target density of pixels with assigned barcodes. We simulated Gaussian diffusion from each point in the final grid, resulting in irregular diffusion patterns due to the irregular colony shapes. For our simulations, we used a final density of 90%, so 10% of anchors relied solely on diffusion to receive sBC reads (See Figure S8A). As a colony must consist of multiple E. coli cells and colonies cannot overlap, we tested lower sBC densities of 5000, 10000, 20000, and 30000 sBCs/cm^2. For all other parameters, we tested the same values used in the sweep with uniform Gaussian diffusion.

With such irregular colonies, DNA-GPS is able to achieve a resolution of 13 μm, only a 3 μm degradation from the uniform Gaussian simulations and was still able to achieve under 30 μm resolution at as low as 25 UMI (50 rpb) (See Figure S8B). However, the robustness to physical parameters decreased, with the method only being able to achieve 20 μm resolution at 7/16 diffusion and sBC density parameter combinations and 30 μm resolution at 12/16 combinations, compared to 15/16 with uniform Gaussian diffusion.

Slide-seq reconstructions

Using bead positions from the Slide-seq kidney, cerebellum 2, and hippocampus datasets we were able to achieve alignment distances as low as 15 μm and could achieve 30 μm alignment distances for all but one combination of number of satellites and diffusion levels from 30-100 μm (See Figures S3 and S7). We believe that the reconstructions were slightly worse for the Slide-seq bead positions because only beads positioned over the tissue were included in the dataset, leading to regions without beads. UMAP assumes that sampled points are uniformly distributed, so our reconstructions perform best when beads have close to uniform density. In practice, all beads would be sequenced leading to roughly uniform density and making these tests a conservative estimate of reconstruction quality. Despite this limitation, the reconstructions with the Slide-seq positions were still at the low end of the single-cell size range.