Integrating Spatially-Resolved Transcriptomics Data Across Tissues and Individuals: Challenges and Opportunities

Boyi Guo, Wodan Ling, Sang Ho Kwon, Pratibha Panwar, Shila Ghazanfar, Keri Martinowich,* and Stephanie C. Hicks**

Advances in spatially-resolved transcriptomics (SRT) technologies have propelled the development of new computational analysis methods to unlock biological insights. The lowering cost of SRT data generation presents an unprecedented opportunity to create large-scale spatial atlases and enable population-level investigation, integrating SRT data across multiple tissues, individuals, species, or phenotypes. Here, unique challenges are described in the SRT data integration, where the analytic impact of varying spatial and biological resolutions is characterized and explored. A succinct review of spatially-aware integration methods and computational strategies is provided. Exciting opportunities to advance computational algorithms amenable to atlas-scale datasets along with standardized preprocessing methods, leading to improved sensitivity and reproducibility in the future are further highlighted.

1. Introduction

Comprehensive molecular atlases with spatial resolution have the power to provide novel and unique insights into human health and disease, which can transform the future of medicine via improved diagnostics and targeted therapies.[1,2] Recent commercialization has led to broad accessibility and hence substantial amounts of spatially-resolved transcriptomics (SRT) data being collected, signifying a new era for spatial cellular atlases to chart the unknown territory of life science.[3] These technologies enable the mapping

B. Guo, S. C. Hicks
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Baltimore, MD 21205, USA
E-mail: shicks19@jhu.edu

W. Ling
Division of Biostatistics
Department of Population Health Sciences
Weill Cornell Medicine
New York, NY 10065, USA

S. H. Kwon, K. Martinowich
Lieber Institute for Brain Development
Johns Hopkins Medical Campus
Baltimore, MD 21205, USA
E-mail: keri.martinowich@libd.org

S. H. Kwon, K. Martinowich
Solomon H. Snyder Department of Neuroscience
Johns Hopkins School of Medicine
Baltimore, MD 21205, USA

S. H. Kwon
Biochemistry, Cellular, and Molecular Biology Graduate Program
Johns Hopkins School of Medicine
Baltimore, MD 21205, USA

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/smtd.202401194

DOI: 10.1002/smtd.202401194

P. Panwar, S. Ghazanfar
School of Mathematics and Statistics
The University of Sydney
Camperdown, NSW 2006, Australia
E-mail: shila.ghazanfar@sydney.edu.au

P. Panwar, S. Ghazanfar
Sydney Precision Data Science Centre
University of Sydney
Camperdown, NSW 2006, Australia

P. Panwar, S. Ghazanfar
Charles Perkins Centre
The University of Sydney
Camperdown, NSW 2006, Australia

K. Martinowich
Department of Psychiatry and Behavioral Sciences
Johns Hopkins School of Medicine
Baltimore, MD, USA

K. Martinowich
Johns Hopkins Kavli Neuroscience Discovery Institute
Johns Hopkins University
Baltimore, MD 21218, USA

K. Martinowich
Department of Biomedical Engineering
Johns Hopkins University
Baltimore, MD 21218, USA

S. C. Hicks
Center for Computational Biology
Johns Hopkins University
Baltimore, MD 21218, USA

S. C. Hicks
Malone Center for Engineering in Healthcare
Johns Hopkins University
Baltimore, MD 21218, USA

of heterogeneous cell populations in situ to tissue architectures, equipping investigators to study the relationships between structure and biological activities.[4] Computational tools and analytic strategies that can fully exploit the atlas-scale SRT data and increase the power to detect small biological signals are critically needed.[5–7] Especially, spatially-aware integrating multiple tissues,[8] developmental stages,[9,10] species,[11,12] or phenotypes[13] to perform population-level analyses faces new and unique challenges.

In contrast to single-sample analyses,[14] integrative analysis of SRT datasets enables unbiased analysis at the population level and has the potential to identify generalizable spatially-dependent commonalities and differences across disease states or conditions such as Alzheimer's disease,[15,16] schizophrenia,[17] and cancer.[18,19] In addition, integrating SRT datasets from different stages of a disease provides insights into the dynamics of spatial gene expressions, revealing mechanisms of tissue development underlying the disease progression.[20,21] Moreover, integrating massive SRT datasets from different studies pertaining to various body sites, phenotypes, etc. helps establish reference atlases,[22,23] which standardizes the baseline for translational biomedical research. Altogether, large-scale and integrative SRT research is an unstoppable trend, which maximizes the value of SRT datasets and empowers reliable spatial molecular discoveries for a range of scientific advancements.

However, in order to realize this potential, considerable effort needs to be made to overcome challenges brought on by the specifics of SRT technology and datasets.[24–26] SRT datasets are varied in underlying biotechnology and therefore in data structure, and affect the scope and quality of integration applications. Coupling these issues with broader applications across the research community to atlas-scale efforts, there is a need to focus efforts on addressing issues that can lead to more effective harmonization of disparate datasets across distinct SRT technologies.

Here, we discuss the computational challenges involved in analyzing an integrative spatial atlas across tissues and individuals with a focus on the existing computational strategies currently available as well as future opportunities for development. We focus on the challenge of integrating SRT samples where observations are measured at different levels of spatial resolution due to the inherent capabilities and limitations of the employed technologies. We illustrate that varying levels of resolution combined with differences in the features captured can lead to spurious findings in downstream analyses, such as dimensionality reduction, both conceptually and via real-world data analysis. These problems are exacerbated by challenges faced in bulk and single-cell/nucleus RNA-sequencing (sc/snRNA-seq) data, such as sparsity and noise.[27] Finally, we summarize the state-of-the-art methods for integrating multiple SRT samples to perform population-level analyses.

1.1. From Bulk to Single-Cell and Spatial Resolution

Integration has become commonplace as the number of SRT datasets increases. Its value in the reliable identification of shared or distinctive spatial cellular features across individuals or phenotypes has already been demonstrated.[28–30] Unwanted variation across SRT datasets is an inevitable challenge faced during integration.[27,31] The challenge is ubiquitous across most sequencing modalities ranging from bulk to single-cell data,[32,33] and is routinely referred to as batch effects. This undesired heterogeneity usually comes from artifacts such as differences in handling protocols, library preparation, and sequencing platforms. Correcting for batch effects has long been a major goal of genomics data integration. Examples for how to correct for batch effects in bulk RNA-seq include the use of statistical modeling to adjust for sample-level differences[34–36] along with the use of control genes.[37]

In contrast to bulk RNA-seq, which measures gene expression in one sample that is averaged across measured cells, scRNA-seq measures gene expression across thousands to millions of cells and introduces more heterogeneity in the gene expression space. Therefore, as we moved from bulk to single-cell resolution, one type of integration strategy that was developed for scRNA-seq experiments was to identify groups of cells that share similar expression patterns across batches (called *anchors*). Broadly these approaches use similarity-based methods in a reduced dimension space, such as mutual nearest neighbors (MNN),[38] Harmony,[39] and canonical correlation.[40] The key idea is that similar cells should follow a common distribution in the latent space regardless of the batch. As an extension of dimension reduction methods, generative models effectively help capture nonlinear characteristics of batch effects and systematic biological signals, such as improving the exhaustiveness of artifact elimination.[41]

However, a prominent feature of scRNA-seq data is that the measured observation, namely gene expression in one cell, is the same, in principle, across all observations measured in multiple scRNA-seq experiments. With SRT, the observations that we measure within a tissue may be the same, but the resolution of observations across multiple samples may not be the same (**Figure 1**). Therefore, while these integrative methods developed for bulk and scRNA-seq experiments demonstrate significant success when integrating bulk and single-cell data, it remains unclear how well these methods will work for SRT data due to intrinsic differences in experimental protocols and the biological context of generated data. For example, this motivates the use of alternative pieces of information, such as anatomical landmarks,[42,43] to assist in the construction of population-level spatial atlases, but these are not always relevant, for example with cancer tissue.

1.2. Inconsistent Spatial and Biological Resolutions Challenge Data Integration

"Spatially-resolved transcriptomics",[44] is often used as an umbrella term for distinct technologies that measure gene expression with spatial resolution.[26] However, due to intrinsic differences in these technologies, for example, sequencing-based versus imaging-based, data generated by these technologies have various resolutions, leading to unique computational and biological properties that make using integration strategies developed for bulk or scRNA-seq data analysis challenging. For example, the units in which we measure observations, namely individual cells or groups of cells, referred to as observational units,

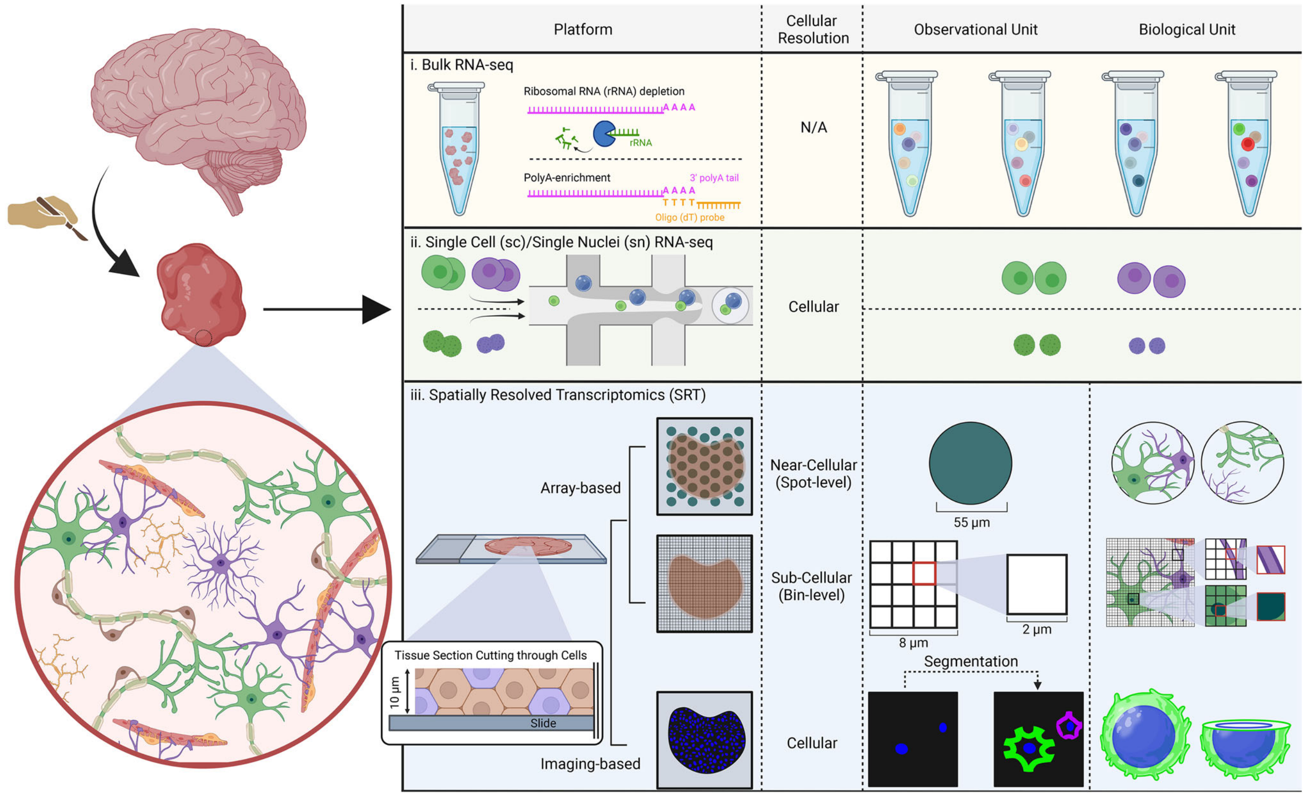


Figure 1. Schematic of experimental platforms and cellular resolutions across transcriptomics technologies. Considering three different experimental platforms i) bulk RNA-sequencing, ii) single-cell/nucleus RNA-sequencing, and iii) spatially-resolved transcriptomics, each of these can profile gene expression at different cellular resolutions, including cellular, near-cellular, and sub-cellular. Differences in experimental platforms also have differences in the units being measured, including observational units and biological units, where observational units describe the observations that we measure and the biological units describe the cellular structure that the observation unit captures.

vary substantially across the SRT platforms. In image-based, targeted, in situ, transcriptomic profiling, such as MERFISH[45] or Xenium,[46] gene expression is captured from a targeted subset of genes at the molecule-level resolution, where the molecules are aggregated together to computationally infer the "cellular" observational unit using cell segmentation algorithms. In contrast, non-targeted RNA capture and sequencing profiling, such as Slide-seq[47] or Visium,[48] captures RNA on an array-based platform at different resolutions, including "near-cellular" (such as 55 µm spots on the Visium platform) or "sub-cellular" (such as 2 µm grids on the VisiumHD platform).[49] Integration of data generated across technologies with different observational units requires special attention.

Unlike the concept of the observational unit whose distinction across SRT technologies is well acknowledged,[3] the heterogeneity in biological content being profiled across observations generated from the same SRT technology is often overlooked. Sc/snRNA-seq protocols often employ cell dissociation techniques to isolate individual cells or nuclei. When processed correctly, data observations have a uniform and biologically meaningful unit (referred to as biological unit hereafter), cell or nuclei, across samples and studies. However, because the profiling happens in situ, this property is often missing in SRT data. For example, with sequencing-based SRT technologies, the profiling within the observational unit is constrained by physical size, for example, spots or grids, so the generated data observation frequently does not maintain a uniform biological content, and hence the biological unit of data observations varies widely. Specifically, the cellular structures being captured across observations could include both the soma of cells and the extracellular space between multiple cell bodies (Figure 1). Inconsistency in biological units in SRT datasets greatly challenges the fundamental assumption that many integration methods for sc/snRNA seq data depend on, namely that each observation is an individual cell. This can lead to spurious results or biases in fundamental data preprocessing steps, such as data normalization,[50] quality control,[51] and in turn, propagate through downstream integration steps.

Even single-cell resolution image-based SRT technologies may suffer from the inconsistency of biological units across data observations. Despite individual SRT tissue sections being conceptually treated as 2D objects, each tissue section has a 3D structure, meaning that the tissue section has some dimension into the *Z*-plane. Depending on their orientation, it is possible for cells to be bisected during tissue sectioning. In this scenario, a cell will not maintain full integrity since only a portion of the cell structure is captured (Figure 1). Moreover, many image-based SRT technologies require iterative imaging of small regions of a tissue section. This iterative imaging procedure creates cells that are located at the boundary of images and hence only partially profiled, resulting in variation in captured genes.[50,52] Although the degree of variation in biological units is smaller than sequencing-based SRT data, further research is necessary to understand the downstream impact of these confounders in data analysis and integration.

Another significant challenge in integrating across different SRT datasets is to mitigate the inclusion of divergent sets of assayed genes across platforms or studies. While targeted profiling technologies provide better spatial resolution, owing to their basis in microscopy, they are often limited by the identity and number of genes profiled via targeted gene panels.[53] Specifically, targeted panels focus on measuring pre-selected sets of genes that are often tissue- or disease-specific, occasionally with some additional genes that are customized to individual studies. For example, a recent study in breast cancer[46] performed joint profiling of a single tumor tissue using the 10x Genomics Xenium platform of 313 genes, alongside the 10x Genomics Visium platform of ≈20 000 genes. While recent commercial offerings of targeted-panel-based SRT have expanded to thousands of genes, this is still not to the full-transcriptome level. Unlike transcriptome-wide sequencing technologies that allow for reads to be mapped to a consistent set of genes across studies, the divergent sets of assayed genes from targeted panels lead to missing gene features when integrating data collected from different studies. Across many such datasets, this could lead to a vanishingly small set of fully intersecting genes, therefore affecting the choice of inclusion of datasets alongside other analytical steps such as normalization. Also, the mismatching of gene profiles due to the frequent missing genes issue of the targeted technologies prevents the direct adoption of scRNA-seq methods to integrate data generated from targeted and non-targeted platforms, since integration methods often require non-missing input data.

1.3. Thought Experiment: Cross-Platform Integration Using Cell Type-Based Anchors

In the following section, we highlight a few examples of how the unique properties of SRT data generate computational challenges for integrating multiple samples.

Normalization is a critical precursor in processing transcriptomics data to remove variation due to technical noise within each dataset. Without proper normalization, artifacts or biases may not be effectively aligned across datasets during integration. Therefore, normalization is considered an integral part of integration in a broader sense and is discussed here as the initial step of the cross-platform integration examples. Current normalization practices for SRT data, regardless of platform, are directly adopted from the scRNA-seq pipeline. However, whether this practice is uniformly appropriate for the diverse types of SRT data remains unclear. A common practice is to normalize the expression of each gene according to the total number of transcripts detected, often referred to as library size normalization. The library size normalization is based on the assumption that the variation in library size across samples is due to technical reasons. However, due to the inconsistent biological unit across samples, library size could reflect variation attributed to the differences in underlying biology. Hence, library size normalization can overcorrect the technical variation and potentially reduce the biological signal. As a result, downstream tasks, such as spatial clustering to establish functional regions, are significantly impacted.[54] Moreover, Atta et al. recently demonstrated that applying library size normalization to targeted SRT data could result in false positive and false negative findings in differential expression testing and spatially variable gene detection.[50] Relevantly, many QC methods rely on descriptive metrics such as library size, total gene detected, which is not robust to the inconsistent biological unit unique to SRT data. Totty et al.[51] recently showed that scRNA-seq-inspired quality control methods could result in differential removal of data observations across multiple biological functional regions in an undesirable way. Additionally, cell types, often used as anchors to harmonize multiple datasets in sc/snRNA-seq, could be substantially challenging to be properly defined from both the computational perspective and philosophical perspective. While the main intuition for cell type annotation is that the difference in gene signature between data observations, that is, cells, is driven by the difference of the cell types, the implicit assumption here is that the data observations are single cells. Nevertheless, for near-cellular and sub-cellular resolution SRT data, such an assumption is often violated. Mapping observations with different biological units to the common latent space can create dubious clusters that lack biological meanings and confound cell type-driven anchors for cross-study integration. For example, in near cellular resolution platforms, each observation could contain a homogeneous or heterogeneous cell population, resulting in distinction in biological units across observations beyond simply capturing different numbers of cells. This creates challenges to define cell type-driven anchors and leads to extra cell type clusters, which, in fact, should be merged with existing clearly-defined cell types (**Figure** 2B). In another case, targeted platforms can miss important marker genes. Thus, when integrated with data generated from transcriptome-wide platforms that have a full spectrum of genes, the anchors cannot be accurately established such that some cell types cannot be successfully differentiated (Figure 2C). Analytically, the SRT technologies provide an unprecedented opportunity to study the molecular mechanisms underpinning the heterogeneities in functions across tissue regions. Many research questions of interest, investigated using SRT platforms, focus on heterogeneity in gene expression associated with functional regions instead of cell types, requiring a switch of thinking from the cell-type centric to the tissue-centric.[55] As a result, the integration tools and strategies that account for both gene expression space and physical space are highly motivated to establish a common coordinate framework.[43]

1.4. Case Study: Joint Analysis of Visium and MERFISH in Adult Mouse Brain

To further demonstrate the challenges in the cross-technology integration, we conducted a preliminary analysis where two datasets of the adult mouse brain, assayed using the sequencing-based Visium and imaging-based MERFISH respectively, are jointly analyzed following the single-cell RNA-seq pipeline. The preliminary analysis highlights the drastic differences in various aspects of the generated data due to heterogeneity in observational and biological units, and their implication in the

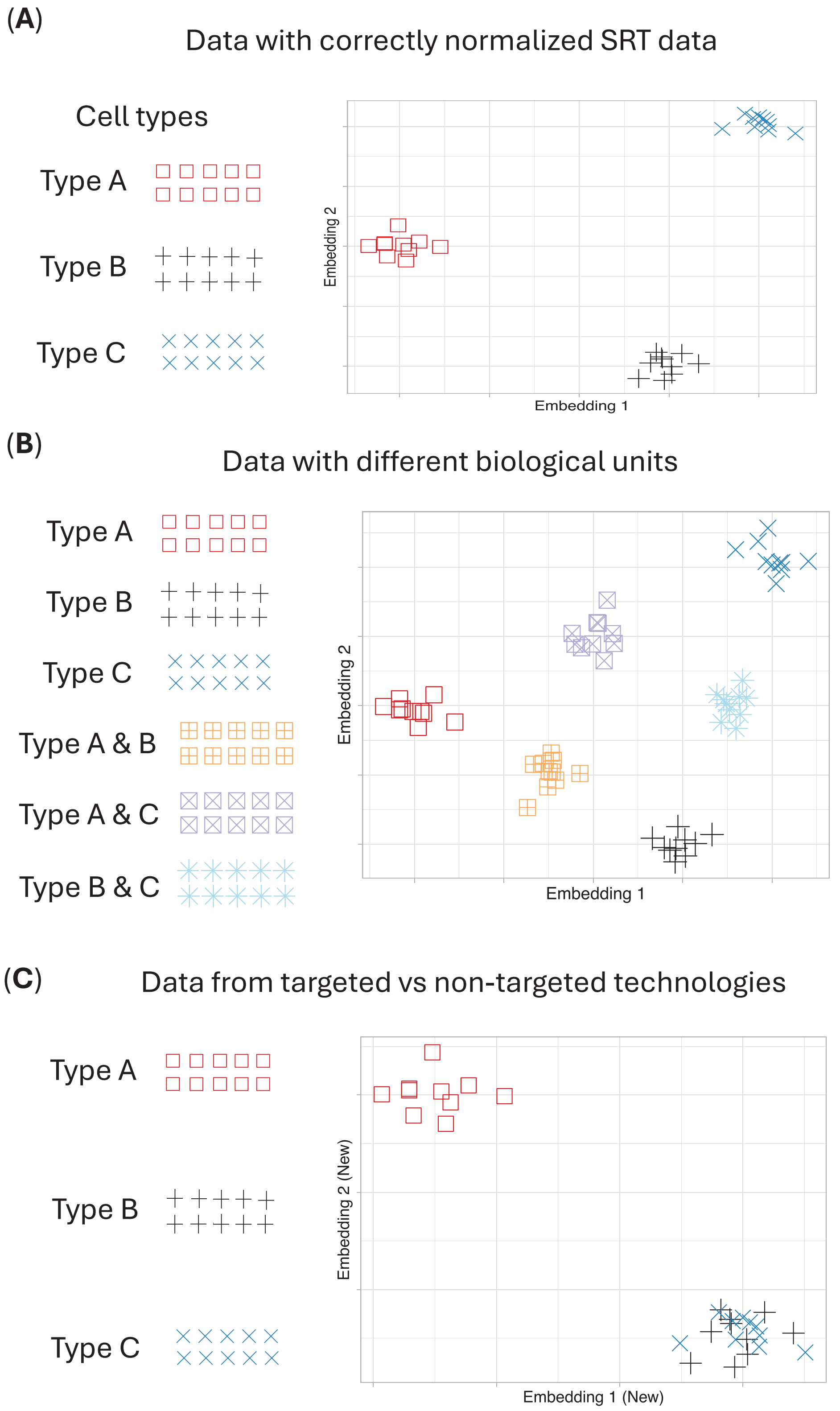


Figure 2. Schematic of cell-type driven integration in gene expression space across multiple SRT technologies. A) With accurate normalization removing technical variation, integration of image-based SRT follows single-cell practice. B) Mapping observations that have different biological units to a common latent space results in dubious clusters that lack biological meanings and confound cell-type driven anchors for cross-study integration. C) Integrating SRT datasets generated with different gene panels (targeted vs non-targeted) creates challenges to computationally define gene expression space where cell-type-based anchors cannot be clearly defined due to missing marker genes in targeted platforms.

data-driven cell typing. These findings echo the concepts introduced in previous sections and demonstrate the urgent need for specialized analytic strategies and advanced computational methods for SRT data integration. While the following case study highlights challenges in cross-technology integration, similar challenges persist when integrating data generated in the same or similar technologies.

Fundamental differences in molecular profiling techniques result in variations across multiple aspects of the generated data, ranging from the morphological level to the transcriptomics level. Visium assay, limited by its 6.5 × 6.5 mm capture area, profiles one hemisphere of the mouse brain; in contrast, the MERFISH assay, with a one square-centimeter capture area, profiles both hemispheres of the brain (**Figure** 3A). Despite mouse brains having more consistent structure across individuals compared to other tissues such as cancer, subtle differences in the anatomical structure of the two brains could benefit from spatial registration, especially when inquiring about the differential abundance of cell populations across samples. At the transcriptome level, the two datasets, when concatenated together, present a mosaic pattern (Figure 3B). While providing single-cell resolution and hence an extremely large number of data observations, MERFISH assays a limited number of genes. The overlapping genes between the two assays exist, but the number is disproportionately small. It is not hard to foresee a situation where two datasets, likely both assayed using targeted probes, contain completely disjoint sets of genes, rendering feature-space integration extremely challenging if not impossible.

The differences in the observational and biological units between the two assays lead to distributional differences in transcriptomic profiles of data observations. To minimize the impact of tissue difference, spatial registration was performed such that data observations from the two assays can be directly compared. The total number of profiled transcripts, also known as library size, are different, as anticipated, between the two assays, largely attributed to the different numbers of genes being profiled (Figure 3C). While limiting the genes to only the overlapping ones brings the library size to a similar scale, the distributions are still different. The multimodal distribution of the MERFISH data could indicate nuanced variance across biological units or cell populations being captured due to the single-cell resolution. Similarly, distributional differences are observed for individual genes, for example, *Baiap2*, even after library size normalization (Figure 3D). The distributional difference in normalized gene expression between the two assays suggests that library size normalization does not completely address non-biological variation, highlighting the opportunity to advance cross-technology normalization methods.

These key differences in technologies and undressed computational challenges impact fundamental biological investigations, such as cell typing. Outstanding batch effect exists between the two assays. (Figure 3E) When projecting MERFISH and Visium data to a common latent space, the Visium observations form a tight cluster among different clusters of MERFISH observations. This suggests the biological variations captured in Visium and MERFISH are on drastically distinct scales and require additional attention. Additionally, when integrating with MERFISH, Visium data needs to reduce its genes to the overlapping ones, whose size is modest. This reduction of genes limits the ability to detect nuanced cell populations when compared to using transcriptome-wide information. This is demonstrated in the data-driven clustering (graph-based with default settings) of Visium observations using overlapping genes and 2000 highly variable genes respectively. (Figure 3F) When necessary, annotating cluster labels across multiple datasets creates additional burden and hence hardly practical when integrating across a large number of datasets (Figure 3G).

1.5. State-of-the-Art Methods

Broadly, methods developed for bulk or scRNA-seq are being widely applied to spatial data, despite the problems outlined above. However, new methods to integrate multiple samples for spatial transcriptomics data have recently been developed. In this section, we outline the modern methods specifically designed for spatial data and give recommendations to data analysts and users of these methods.

1.6. Integration in a Physical Space

The first category that we consider is to integrate multiple samples in a physical space. Within this category, we further distinguish approaches based on the type of data being integrated including i) the alignment of two tissue slices from the same tissue block or from different tissue blocks, but both profiling the transcriptome in a 2D space and ii) the registration of a set of dissociated single cells to one tissue slice profiling the transcriptome in a 2D space.

Early work of spatial alignment was computer-assisted, requiring human input, such as manually defined anatomical landmarks, and computationally relies on the affine transformation, for example, using iterative closest point algorithm,[57] of high-resolution images of samples, for example, hematoxylin and eosin (H&E) or immunofluorescent images, to address rotations and shifting. Then, various methods were developed to address possible nonlinear distortion, leveraging thin plate spline,[58] Gaussian process,[59] diffeomorphic metric mapping.[56] Because spatial alignment of tissue images normally requires different degrees of involvement in manual labor, a significant challenge is how to scale it to atlas-scale data sets that contain hundreds of samples. Considerable approaches have been proposed to address this challenge, most involving modeling the entire gene expression profiles accounting for the global structure of the spatial unit arrangement, including the two-layer Gaussian process model,[59] diffeomorphic metric mapping,[56] optimal transport,[60] and a graph adversarial matching strategy.[61] These methods seem to be well motivated for the alignment of i) samples with partial matching, also referred to as spatially heterogeneous samples, ii) spatial alignment to a reference or template, such as a reference include a predefined Brain atlas,[62,63] or iii) samples across different SRT technologies or possible of various phenotype readouts, such as gene expression and protein expression.

In addition, the spatial registration of single cells to a 2D tissue section provides another venue for spatial integration that mitigates analytic challenges due to morphological variation. Specif-

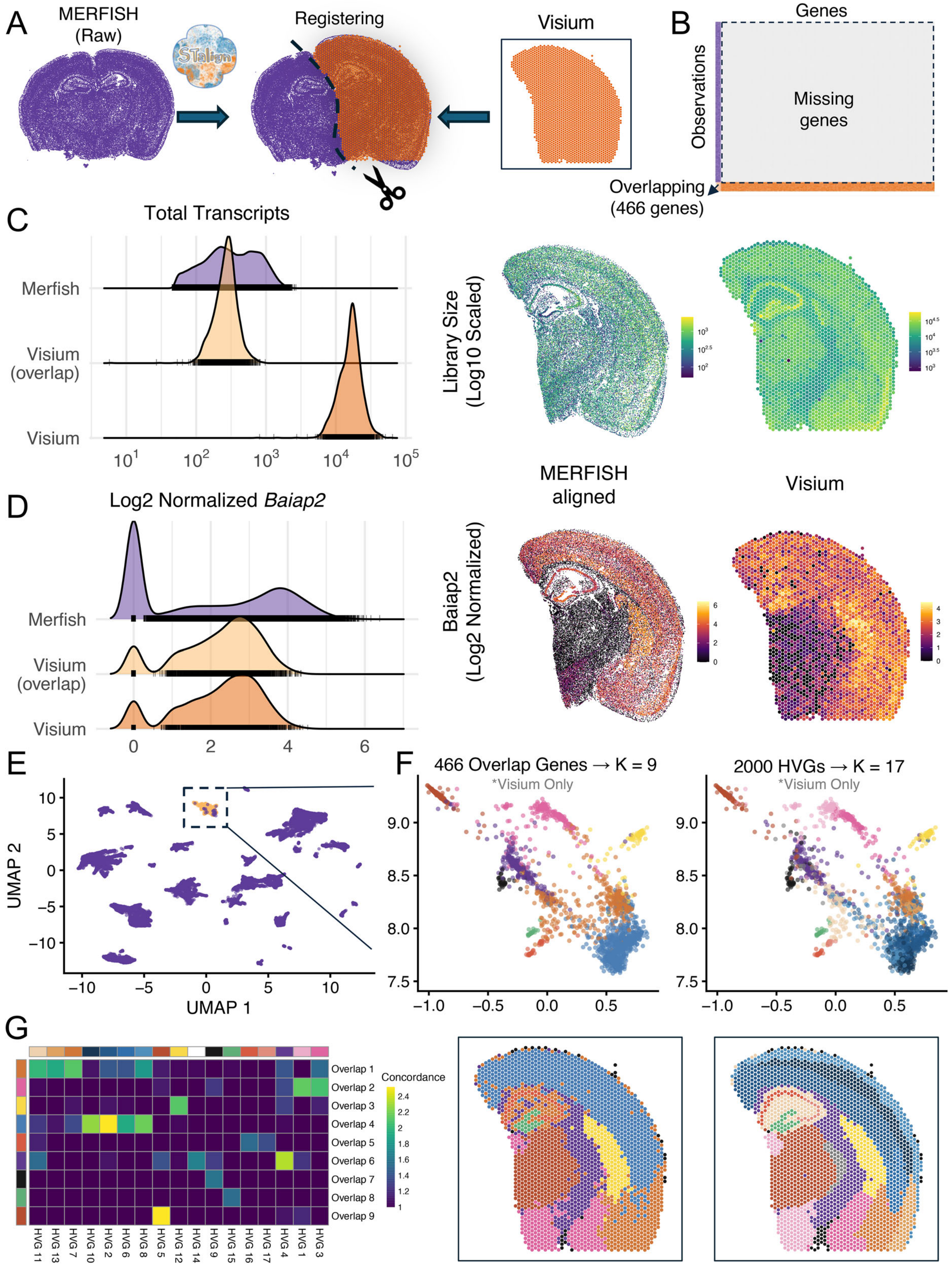


Figure 3. Computational challenges in cross-technology integration – a joint analysis of Visium and MERFISH datasets of coronal sections of the adult mouse brain. A) Variation in tissue morphology exists between the Visium sample (orange) and the MERFISH sample (purple) and is addressed by spital registration using STalign.[56] B) Heatmap of gene expression shows the inconformable dimensions of data generated via Visium and MERFISH, highlighting the small number of overlapping genes (light orange) between technologies and substantial missing genes (grey) in MERFISH. C) Ridge plot and spatial plot visualize the differences in the distributions of total transcripts profiled across observations of MERFISH (purple), Visium (orange), and

ically, isolated single cells, or possibly spots, are computationally mapped to a spatial common coordinate system, for example a spatial template, based on their molecular signature. By mapping all cells isolated from a tissue (either experimentally through scRNA-seq or computationally with SRT data) to a reference or template tissue with uniform morphology, the tissue slices are hence aligned in the physical space with conformable shapes. Developed for the spatial reference assayed with low-throughput technologies, early methods use a set of pre-specified marker genes to anchor cells to a limited number of spatial positions, such as the tessellation of a 2D surface, in Gaussian mixture models[64] or Monte Carlo simulations.[65] To allow non-informative mapping without external reference and improve the spatial resolution of the mapping, advanced machine learning frameworks are adapted, including multiple variations of optimal transport algorithms,[66,67] convex optimization using the Jonker–Volgenant shortest augmenting path algorithm,[68] and deep neural networks.[69,70] Some spatially-aware deconvolution methods can be also used for the spatial registration of cells.[71]

While these methods are promising, it is important to be cautious in the interpretation because it is possible that spurious gene expression correlations are found due to known problems such as "double dipping".[72] Given the fast-paced nature of this research area, multi-modal integration approaches might also be useful to avoid the circular problem of double dipping and directly use different data modalities.

1.7. Integration in a Latent Space

The second approach ignores (for the most part) the physical space, and focuses on integrating the samples in the latent space. Recent examples are the use of deep learning models,[73–75] which combine spatial neighbor networks and graph auto-encoders to learn latent embeddings. In contrast to integration in the physical space, integration in the latent space is anchored in the feature (gene expression) space, while also borrowing information from nearby, physically adjacent cells/spots. Feature space integration has a long history in scRNA data analysis, such as the MNN, Harmony, and canonical correlation methods mentioned in discussing the evolution of genomics data integration. The fundamental idea is to project gene expression data into the same latent space accounting for batch effects and the downstream investigation is completed in the shared latent space. However, integration that ignores the physical relationships between cells is vulnerable to noise in the data, particularly when the biological signal is small. The spatial information is introduced to the latent space integration to remove the noise, such as in the form of dimension reduction, or possibly in the form of clustering. Distinctly, PRECAST[76] models any arbitrary tissue architecture across multiple tissue samples by factorizing the input into a latent factor with a shared distribution for cell, or domain,i and then using an intrinsic conditional autoregressive component to capture the spatial dependence for spatial clustering. In addition, the spatial information can be also used to smooth out any possible noise. For example, BayesSpace[77] uses majority voting to accomplish spatial smoothing of cluster membership. Once these spatial domains are identified, which normally aligns with anatomical definitions, cells/spots are pseudo-bulked across individual samples, followed by nominal analysis, such as differential expression analysis. However, this pseudo-bulking analysis normally lacks sensitivity for local spatial signals and hence would not be helpful for any granular analysis to study the micro-environment.

1.8. Integration Using Pseudobulking Approaches

Given the flourishing development of pseudobulking approaches in the integration of scRNAseq data,[78,79] it is natural to extend this for SRT data where gene expression is aggregated across spots within a spatial domain and a tissue section. As the analysis is conducted at the observational unit of a spatial domain and tissue section, there is no need to address morphological variation. Furthermore, this approach enables existing methods, designed for bulk RNA-sequencing to be used in this setting. For example, pseudobulking SRT data can be used to identify differentially expressed genes (DEGs) across spatial domains with multiple tissue blocks or individuals, which has been successfully applied in human brain tissue including dorsolateral prefrontal cortex,[30,80] locus coeruleus,[81] and the hippocampus.[82] In addition, pseudobulking provides a scalable solution to analyzing atlas-scale SRT data, motivated by the underlying "divide-and-conquer" or "map-and-reduce" philosophy.

However, this approach is not necessarily appropriate for all research questions and methods for integrative analysis using SRT data. While this approach sidesteps the challenges due to morphological variation, it also ignores variation due to gene expression patterns varying within a spatial domain as that information is aggregated together into one summary statistic. This loss of information might be particularly important for some downstream analyses, such as identifying spatially variable genes, cellular deconvolution, and cell–cell communication. Pseudobulking or other approaches that summarize features at the sample-level[83] might not be appropriate in these cases.

Furthermore, there are open opportunities to improve the sensitivity and robustness of the statistical models used to perform integrative analyses using pseudobulk data. Current approaches use linear models with empirical Bayes techniques to identify DEGs where the spatial domains are assumed to be discrete.

Visium with overlapping genes only (light orange), highlighting the heterogeneity in both the observation unit and biological unit across technologies. D) The distributions of library-size normalized Baiap2 expression differ between MERFISH and Visium, suggesting the immense need for advanced normalization methods that account for differences in observation and biological units. E) Naive analysis by simply concatenating gene expression (Panel B) without computational integration highlights the outstanding batch effects between technologies. The variation between cell populations in MERFISH is considerably large in reference to the total variation within Visium technology, quantified by Uniform Manifold Approximation and Projection (UMAP). F,G) Data-driven clustering of Visium observations using the 466 MERFISH-overlapping genes ($K = 9$) greatly limits the potential to identify nuanced cell types, compared to clustering results ($K = 17$) using 2000 highly variable genes (HVGs) identified from the transcriptome-wide gene set (Cluster legends are shown in Panel G). G) Heatmap displays the concordance between data-driven clusters identified using 466 genes and 2000 HVGs.

However, more modern models could be considered where the spatial domains are more continuous across a 2D space. In addition, this approach requires labor-intensive human intervention, for example, the need to harmonize the labels corresponding to the spatial domains from unsupervised clustering algorithms. Specifically, the spatial alignment can be addressed via a manual operation, but this can be time-consuming and is prone to human error, leading to potentially unreliable and unreplicable results.

2. Discussion

The aim of this work is to both describe the historical context and summarize state-of-the-art strategies to perform population-level analyses that can overcome potential systematic biases in SRT data. The three approaches can broadly be summarized as i) integration in a physical space, ii) integration in a latent space, and iii) integration using pseudobulking or similar approaches. Despite this, approaches developed for scRNA-seq remain widely used in practice. While it remains unclear which type of approach is best to integrate SRT data, there are many ongoing and active efforts to begin comparing these approaches through robust benchmark evaluations.

Furthermore, while much progress has been made toward early strategies, there remain important open challenges to be addressed. For example, when using approaches to integrate tissue sections in a physical space, it remains unclear how to identify the partial overlap and quantify how much area is needed for successful alignment. In addition, smoothing of gene expression is often used implicitly or explicitly as a step in the alignment process before passing to downstream analysis. Further validation to avoid potential over-smoothing that eliminates nuanced biological signals in microenvironments would be greatly encouraged to avoid introducing computational artifacts. These challenges also relate to potential differences in the amount of biological tissue measured, where one can imagine new spatial platforms capturing a larger tissue area compared to older spatial platforms.

There also remain challenges with the current state-of-the-art strategies, such as the accuracy of cell segmentation, which remains one of the largest challenges with SRT data. Also, while pseudobulking enables the integration across datasets with different observational units, this approach also potentially masks important spatial variation within a given spatial domain. Therefore, we imagine new computational tools being developed that can integrate multiple samples measured with different observational units to take advantage of the full rich information provided by multi-sample SRT datasets.

3. Data Analysis

STalign[56] is used to spatially register the Visium and MERFISH mouse brain sections, where a threshold of 0.95 is used to identify the overlapping hemisphere between Visium and MERFISH. The Visium dataset, using transcriptome-wide genes and 466 MERFISH-overlapping genes respectively, and the MERFISH dataset are library size normalized and log2 transformed, using scuttle::logNormCounts.[84] To select a subset of genes from the transcriptomic-wide dataset, we used scran::getTopHVGs[85] to identify top 2000 highly variable genes (HVG). To identify cell populations of the Visium data in a data-driven manner, the Visium datasets, the 2000-HVG and 466 MERFISH-overlapping version, respectively underwent the principal component analysis (PCA) using scran::fixedPCA, followed by graph-based clustering, implemented using scran::clusterCells with default parameters. To create a common latent embedding of Visium and MERFISH datasets, the concatenated dataset of MERFISH and Visium (only 466 genes) are log normalized jointly, followed by PCA and Uniform Manifold Approximation and Projection using scater::runUMAP. Spatial visualization of Visium and MERFISH is constructed using escheR package.[86]

Acknowledgements

B.G. and W.L. contributed equally to this work. The authors would like to thank Kasper Hansen and members of the Hicks Lab for their feedback on this commentary. The authors would also like to thank the collaborators at the Lieber Institute for Brain Development for their input and feedback. This project was supported by the National Institute of Mental Health [R01MH126393 to B.G., S.H.K., K.M., S.C.H.], and the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation [DAF2023-323340 to S.C.H., P.P., and S.G.], the National Institute of General Medical Sciences [R01GM151301, R01GM155734 to W.L.], and Australian Research Council Discovery Early Career Researcher Awards (DE220100964) funded by the Australian Government to S.G.; Chan Zuckerberg Initiative Single Cell Biology Data Insights grant (2022-249319) to S.G. and P.P. All funding bodies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

integrative analysis, multi-sample, population-level, spatial alignment, spatial registration, spatially-resolved transcriptomics

Received: August 1, 2024
Revised: December 13, 2024
Published online: February 11, 2025

[1] J. E. Rood, A. Maartens, A. Hupalowska, S. A. Teichmann, A. Regev, *Nat. Med.* **2022**, *28*, 2486.

[2] J. Park, R. Gregorio, E. Hissong, S. Patel, B. Robinson, F. Socciarelli, E. Metzger, Y. Liang, J. Reeves, J. Beechem, O. Elemento, A. Alonso, S. Houlihan, R. Schwartz, C. E. Mason, *Cancer Res.* **2024**, *84*, 4900.

[3] L. Moses, L. Pachter, *Nat. Methods* **2022**, *19*, 534.

[4] A. Rao, D. Barkley, G. S. França, I. Yanai, *Nature* **2021**, *596*, 211.

[5] M. Cheng, Y. Jiang, J. Xu, A.-F. A. Mentis, S. Wang, H. Zheng, S. K. Sahu, L. Liu, X. Xu, *J. Genet. Genomics* **2023**, *50*, 625.

[6] M. Piwecka, N. Rajewsky, A. Rybak-Wolf, *Nat. Rev. Neurol.* **2023**, *19*, 346.

[7] M. Asp, J. Bergenstråhle, J. Lundeberg, *BioEssays* **2020**, *42*, 1900221.

[8] S. Jain, L. Pei, J. M. Spraggins, M. Angelo, J. P. Carson, N. Gehlenborg, F. Ginty, J. P. Gonçalves, J. S. Hagood, J. W. Hickey, N. L. Kelleher, L. C. Laurent, S. Lin, Y. Lin, H. Liu, A. Naba, E. S. Nakayasu, W.-J. Qian,

A. Radtke, P. Robson, B. R. Stockwell, R. Van de Plas, I. S. Vlachos, M. Zhou, K. J. Ahn, J. Allen, D. M. Anderson, C. R. Anderton, C. Curcio, A. Angelin, et al., *Nat. Cell Biol.* **2023**, *25*, 1089.

[9] M. E. Ardini-Poleske, R. F. Clark, C. Ansong, J. P. Carson, R. A. Corley, G. H. Deutsch, J. S. Hagood, N. Kaminski, T. J. Mariani, S. S. Potter, G. S. Pryhuber, D. Warburton, J. A. Whitsett, S. M. Palmer, N. Ambalavanan, *Am. J. Physiol. Lung Cell. Mol. Physiol.* **2017**, *313*, L733.

[10] T. Lohoff, S. Ghazanfar, A. Missarova, N. Koulena, N. Pierson, J. A. Griffiths, E. S. Bardot, C.-H. L. Eng, R. C. V. Tyser, R. Argelaguet, C. Guibentif, S. Srinivas, J. Briscoe, B. D. Simons, A.-K. Hadjantonakis, B. Göttgens, W. Reik, J. Nichols, L. Cai, J. C. Marioni, *Nat. Biotechnol.* **2022**, *40*, 74.

[11] M. Lotfollahi, Y. Hao, F. J. Theis, R. Satija, *Cell* **2024**, *187*, 2343.

[12] Y. Rosen, M. Brbić, Y. Roohani, K. Swanson, Z. Li, J. Leskovec, *Nat. Methods* **2024**, *21*, 1492.

[13] T. Kumar, K. Nee, R. Wei, S. He, Q. H. Nguyen, S. Bai, K. Blake, M. Pein, Y. Gong, E. Sei, M. Hu, A. K. Casasent, A. Thennavan, J. Li, T. Tran, K. Chen, B. Nilges, N. Kashikar, O. Braubach, B. Ben Cheikh, N. Nikulina, H. Chen, M. Teshome, B. Menegaz, H. Javaid, C. Nagi, J. Montalvan, T. Lev, S. Mallya, D. F. Tifrea, et al., *Nature* **2023**, *620*, 181.

[14] L. Atta, J. Fan, *Nat. Commun.* **2021**, *12*, 5283.

[15] E. Miyoshi, S. Morabito, C. M. Henningfield, N. Rahimzadeh, S. Kiani Shabestari, S. Das, N. Michael, F. Reese, Z. Shi, Z. Cao, et al., *BioRxiv* **2023**.

[16] C. Wang, D. Acosta, M. McNutt, J. Bian, A. Ma, H. Fu, Q. Ma, *Nat. Commun.* **2024**, *15*, 4710.

[17] J. Leon, S. Yoshinaga, M. Hino, A. Nagaoka, Y. Ando, J. Moody, M. Kojima, A. Kitazawa, K. Hayashi, K. Nakajima, et al., *BioRxiv* **2024**.

[18] R. Arora, C. Cao, M. Kumar, S. Sinha, A. Chanda, R. McNeil, D. Samuel, R. K. Arora, T. W Matthews, S. Chandarana, R. Hart, J. C. Dort, J. Biernaskie, P. Neri, M. D. Hyrcza, P. Bose, *Nat. Commun.* **2023**, *14*, 5029.

[19] E. Denisenko, L. de Kock, A. Tan, A. B. Beasley, M. Beilin, M. E. Jones, R. Hou, D. Ó. Muirí, S. Bilic, G. R. K. A. Mohan, S. Salfinger, S. Fox, K. P. W. Hmon, Y. Yeow, Y. Kim, R. John, T. S. Gilderman, E. Killingbeck, E. S. Gray, P. A. Cohen, Y. Yu, A. R. R. Forrest, *Nat. Commun.* **2024**, *15*, 2860.

[20] S. Huang, L. Ouyang, J. Tang, K. Qian, X. Chen, Z. Xu, J. Ming, R. Xi, *CCB* **2024**, *3*, 13.

[21] R. Zhou, G. Yang, Y. Zhang, Y. Wang, *Mol. Biomed.* **2023**, *4*, 32.

[22] O. Rozenblatt-Rosen, M. J. T. Stubbington, A. Regev, S. A. Teichmann, *Nature* **2017**, *550*, 451.

[23] K. Siletti, R. Hodge, A. Mossi Albiach, K. W Lee, S.-L. Ding, L. Hu, P. Lönnerberg, T. Bakken, T. Casper, M. Clark, N. Dee, J. Gloe, D. Hirschstein, N. V. Shapovalova, C. D Keene, J. Nyhus, H. Tung, A. M. Yanny, E. Arenas, E. S. Lein, S. Linnarsson, *Science* **2023**, *382*, eadd7046.

[24] D. Bressan, G. Battistoni, G. J. Hannon, *Science* **2023**, *381*, eabq4964.

[25] L. Liu, A. Chen, Y. Li, J. Mulder, H. Heyn, X. Xu, *Cell* **2024**, *187*, 4488.

[26] T. Alexandrov, J. Saez-Rodriguez, S. K. Saka, *Mol. Syst. Biol.* **2023**, *19*, e10571.

[27] D. Lähnemann, J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz, L. Pinello, P. Skums, A. Stamatakis, C. S.-O. Attolini, S. Aparicio, J. Baaijens, M. Balvert, B. D. Barbanson, A. Cappuccio, G. Corleone, B. E. Dutilh, M. Florescu, V. Guryev, R. Holmer, K. Jahn, T. J. Lobo, E. M. Keizer, I. Khatri, S. M. Kielbasa, J. O. Korbel, et al., *Genome Biol.* **2020**, *21*, 31.

[28] R. Dries, Q. Zhu, R. Dong, C.-H. L. Eng, H. Li, K. Liu, Y. Fu, T. Zhao, A. Sarkar, F. Bao, R. E. George, N. Pierson, L. Cai, G.-C. Yuan, *Genome Biol.* **2021**, *22*, 78.

[29] B. He, L. Bergenstråhle, L. Stenbeck, A. Abid, A. Andersson, Å. Borg, J. Maaskola, J. Lundeberg, J. Zou, *Nat. Biomed. Eng.* **2020**, *4*, 827.

[30] K. R. Maynard, L. Collado-Torres, L. M. Weber, C. Uytingco, B. K. Barry, S. R. Williams, J. L. Catallini, M. N. Tran, Z. Besich, M. Tippani, J. Chew, Y. Yin, J. E. Kleinman, T. M. Hyde, N. Rao, S. C. Hicks, K. Martinowich, A. E. Jaffe, *Nat. Neurosci.* **2021**, *24*, 425.

[31] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, R. A. Irizarry, *Nat. Rev. Genet.* **2010**, *11*, 733.

[32] S. C. Hicks, F. W. Townes, M. Teng, R. A. Irizarry, *Biostatistics* **2018**, *19*, 562.

[33] M. D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, F. J. Theis, *Nat. Methods* **2022**, *19*, 41.

[34] C. W. Law, Y. Chen, W. Shi, G. K. Smyth, *Genome Biol.* **2014**, *15*, R29.

[35] M. D. Robinson, D. J. McCarthy, G. K. Smyth, *Bioinformatics* **2010**, *26*, 139.

[36] M. I. Love, W. Huber, S. Anders, *Genome Biol.* **2014**, *15*, 550.

[37] D. Risso, J. Ngai, T. P. Speed, S. Dudoit, *Nat. Biotechnol.* **2014**, *32*, 896.

[38] L. Haghverdi, A. T. L. Lun, M. D. Morgan, J. C. Marioni, *Nat. Biotechnol.* **2018**, *36*, 421.

[39] I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-R. Loh, S. Raychaudhuri, *Nat. Methods* **2019**, *16*, 1289.

[40] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, *Nat. Biotechnol.* **2018**, *36*, 411.

[41] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, N. Yosef, *Nat. Methods* **2018**, *15*, 1053.

[42] M. Ekvall, L. Bergenstråhle, A. Andersson, P. Czarnewski, J. Olegård, L. Käll, J. Lundeberg, *Nat. Methods* **2024**, *21*, 673.

[43] J. E. Rood, T. Stuart, S. Ghazanfar, T. Biancalani, E. Fisher, A. Butler, A. Hupalowska, L. Gaffney, W. Mauck, G. Eraslan, J. C. Marioni, A. Regev, R. Satija, *Cell* **2019**, *179*, 1455.

[44] V. Marx, *Nat. Methods* **2021**, *18*, 9.

[45] K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, X. Zhuang, *Science* **2015**, *348*, aaa6090.

[46] A. Janesick, R. Shelansky, A. D. Gottscho, F. Wagner, S. R. Williams, M. Rouault, G. Beliakoff, C. A. Morrison, M. F. Oliveira, J. T. Sicherman, A. Kohlway, J. Abousoud, T. Y. Drennon, S. H. Mohabbat, S. E. B. Taylor, *Nat. Commun.* **2023**, *14*, 8353.

[47] S. G. Rodriques, R. R. Stickels, A. Goeva, C. A. Martin, E. Murray, C. R. Vanderburg, J. Welch, L. M. Chen, F. Chen, E. Z. Macosko, *Science* **2019**, *363*, 1463.

[48] P. L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, A. Mollbrink, S. Linnarsson, S. Codeluppi, Å. Borg, F. Pontén, P. I. Costea, P. Sahlén, J. Mulder, O. Bergmann, J. Lundeberg, J. Frisén, *Science* **2016**, *353*, 78.

[49] M. F. Oliveira, J. P. Romero, M. Chung, S. Williams, A. D. Gottscho, A. Gupta, S. E. Pilipauskas, S. Mohabbat, N. Raman, D. Sukovich, et al., *BioRxiv* **2024**.

[50] L. Atta, K. Clifton, M. Anant, G. Aihara, J. Fan, *Genome Biol.* **2024**, *25*, 153.

[51] M. Totty, S. C. Hicks, B. Guo, *BioRxiv* **2024**.

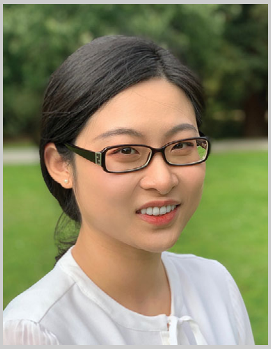
[52] S. Cervilla, D. Grases, E. Perez, F. X. Real, E. Musulen, M. Esteller, E. Porta-Pardo, *BioRxiv* **2024**.

[53] J. R. Moffitt, E. Lundberg, H. Heyn, *Nat. Rev. Genet.* **2022**, *23*, 741.

[54] D. D. Bhuva, C. W. Tan, A. Salim, C. Marceaux, M. A. Pickering, J. Chen, M. Kharbanda, X. Jin, N. Liu, K. Feher, G. Putri, W. D. Tilley, T. E. Hickey, M.-L. Asselin-Labat, B. Phipson, M. J. Davis, *Genome Biol.* **2024**, *25*, 99.

[55] R. O. Ramirez Flores, J. D. Lanzer, D. Dimitrov, B. Velten, J. Saez-Rodriguez, *eLife* **2023**, *12*, 93161.
[56] K. Clifton, M. Anant, G. Aihara, L. Atta, O. K. Aimiuwu, J. M. Kebschull, M. I. Miller, D. Tward, J. Fan, *Nat. Commun.* **2023**, *14*, 8123.
[57] J. Bergenstråhle, L. Larsson, J. Lundeberg, *BMC Genomics* **2020**, *21*, 482.
[58] A. Andersson, Ž. Andrusivová, P. Czarnewski, X. Li, E. Sundström, J. Lundeberg, *BioRxiv* **2021**.
[59] A. Jones, F. W. Townes, D. Li, B. E. Engelhardt, *BioRxiv* **2022**.
[60] R. Zeira, M. Land, A. Strzalkowski, B. J. Raphael, *Nat. Methods* **2022**, *19*, 567.
[61] C.-R. Xia, Z.-J. Cao, X.-M. Tu, G. Gao, *Nat. Commun.* **2023**, *14*, 7236.
[62] Allen Reference Atlases :: Atlas Viewer, Available at: http://atlas.brain-map.org (accessed: July 2024).
[63] ISH Data :: Allen Brain Atlas: Mouse Brain Available at: http://mouse.brain-map.org (accessed: July 2024).
[64] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, A. Regev, *Nat. Biotechnol.* **2015**, *33*, 495.
[65] K. Achim, J.-B. Pettit, L. R. Saraiva, D. Gavriouchkina, T. Larsson, D. Arendt, J. C. Marioni, *Nat. Biotechnol.* **2015**, *33*, 503.
[66] Z. Cang, Q. Nie, *Nat. Commun.* **2020**, *11*, 2084.
[67] M. Nitzan, N. Karaiskos, N. Friedman, N. Rajewsky, *Nature* **2019**, *576*, 132.
[68] M. R. Vahid, E. L. Brown, C. B. Steen, W. Zhang, H. S. Jeon, M. Kang, A. J. Gentles, A. M. Newman, *Nat. Biotechnol.* **2023**, *41*, 1543.
[69] T. Biancalani, G. Scalia, L. Buffoni, R. Avasthi, Z. Lu, A. Sanger, N. Tokcan, C. R. Vanderburg, Å. Segerstolpe, M. Zhang, I. Avraham-Davidi, S. Vickovic, M. Nitzan, S. Ma, A. Subramanian, M. Lipinski, J. Buenrostro, N. B. Brown, D. Fanelli, X. Zhuang, E. Z. Macosko, A. Regev, *Nat. Methods* **2021**, *18*, 1352.
[70] Q. Zhang, S. Jiang, A. Schroeder, J. Hu, K. Li, B. Zhang, D. Dai, E. B. Lee, R. Xiao, M. Li, *Nat. Commun.* **2023**, *14*, 4050.
[71] V. Kleshchevnikov, A. Shmatko, E. Dann, A. Aivazidis, H. W. King, T. Li, R. Elmentaite, A. Lomakin, V. Kedlian, A. Gayoso, M. S. Jain, J. S. Park, L. Ramona, E. Tuck, A. Arutyunyan, R. Vento-Tormo, M. Gerstung, L. James, O. Stegle, O. A. Bayraktar, *Nat. Biotechnol.* **2022**, *40*, 661.
[72] L. L. Gao, J. Bien, D. Witten, *J. Am. Stat. Assoc.* **2024**, *119*, 332.
[73] Y. Long, K. S. Ang, M. Li, K. L. K. Chong, R. Sethi, C. Zhong, H. Xu, Z. Ong, K. Sachaphibulkij, A. Chen, L. Zeng, H. Fu, M. Wu, L. H. K. Lim, L. Liu, J. Chen, *Nat. Commun.* **2023**, *14*, 1155.
[74] K. Dong, S. Zhang, *Nat. Commun.* **2022**, *13*, 1739.
[75] C. Xu, X. Jin, S. Wei, P. Wang, M. Luo, Z. Xu, W. Yang, Y. Cai, L. Xiao, X. Lin, H. Liu, R. Cheng, F. Pang, R. Chen, X. Su, Y. Hu, G. Wang, Q. Jiang, *Nucleic Acids Res.* **2022**, *50*, e131.
[76] W. Liu, X. Liao, Z. Luo, Y. Yang, M. C. Lau, Y. Jiao, X. Shi, W. Zhai, H. Ji, J. Yeong, J. Liu, *Nat. Commun.* **2023**, *14*, 296.
[77] E. Zhao, M. R. Stone, X. Ren, J. Guenthoer, K. S. Smythe, T. Pulliam, S. R. Williams, C. R. Uytingco, S. E. B. Taylor, P. Nghiem, J. H. Bielas, R. Gottardo, *Nat. Biotechnol.* **2021**, *39*, 1375.
[78] H. L. Crowell, C. Soneson, P.-L. Germain, D. Calini, L. Collin, C. Raposo, D. Malhotra, M. D. Robinson, *Nat. Commun.* **2020**, *11*, 6077.
[79] A. E. Murphy, N. G. Skene, *Nat. Commun.* **2022**, *13*, 7851.
[80] L. Huuki-Myers, A. Spangler, N. Eagles, K. D. Montgomery, S. H. Kwon, B. Guo, M. Grant-Peters, H. R. Divecha, M. Tippani, C. Sriworarat, et al., *BioRxiv* **2023**.
[81] L. M. Weber, H. R. Divecha, M. N. Tran, S. H. Kwon, A. Spangler, K. D. Montgomery, M. Tippani, R. Bharadwaj, J. E. Kleinman, S. C. Page, et al., *eLife* **2024**, *12*.
[82] E. D. Nelson, M. Tippani, A. D. Ramnauth, H. R. Divecha, R. A. Miller, N. J. Eagles, E. A. Pattie, S. H. Kwon, S. V. Bach, U. M. Kaipa, *BioRxiv* **2024**.
[83] Y. Cao, Y. Lin, E. Patrick, P. Yang, J. Y. H. Yang, *Bioinformatics* **2022**, *38*, 4745.
[84] D. J. McCarthy, K. R. Campbell, A. T. L. Lun, Q. F. Wills, *Bioinformatics* **2017**, *33*, 1179.
[85] A. T. L. Lun, D. J. McCarthy, J. C. Marioni, *F1000Research* **2016**, *5*, 2122.
[86] B. Guo, L. A. Huuki-Myers, M. Grant-Peters, L. Collado-Torres, S. C. Hicks, *Bioinf. Adv.* **2023**, *3*, vbad179.

Boyi Guo is a postdoctoral fellow, applied statistician, and biomedical data scientist working at the intersection of machine learning, computational omics, and population health. His research concentrates on developing statistically rigorous and computationally scalable machine-learning methods, as well as open-source software, that integrates population-scale multi-omics data to uncover functional mechanisms that explain disease heterogeneity.



Wodan Ling is an assistant professor in the Division of Biostatistics within the Department of Population Health Sciences at Weill Cornell Medicine. She works on statistical issues that arise in complex and structured -omics data, including genetics/genomics, microbiome, metagenomics, metabolomics, etc. Her research focuses on developing robust and powerful quantile regression and machine learning (including deep learning) methods to harness the inherent characteristics of -omics data for an improved understanding of health and disease.

Sang Ho Kwon is a graduate student in the Biochemistry, Cellular and Molecular Biology Program at Johns Hopkins University. His research centers on exploring the cellular and molecular mechanisms underlying complex brain disorders including Alzheimer's disease and schizophrenia, with an emphasis on building expertise in advanced tools and models such as multi-modal omics and iPSC-based organoid disease systems to support drug discovery and development.

Pratibha Panwar is a postdoctoral research associate at the School of Mathematics and Statistics, University of Sydney, and is affiliated with the Sydney Precision Data Science Center and Charles Perkins Center. Her research focuses on developing computational methods for the analysis of spatial transcriptomics data, and she is particularly interested in the analysis of biomedical single-cell and spatial transcriptomics data. Pratibha obtained her Ph.D. from the University of New South Wales, where she utilized bioinformatics approaches to study an Antarctic lake microbiome to understand the impact of climate change on microbial populations.

Shila Ghazanfar is an Australian Research Council DECRA Fellow and Senior Lecturer at the University of Sydney and is an expert in statistical and computational analysis of spatial transcriptomics and single-cell RNA-seq data. Dr. Ghazanfar completed her undergraduate and Ph.D. studies in statistics and statistical bioinformatics at The University of Sydney, before completing a Royal Society Newton International Fellowship at The University of Cambridge in computational biology.

Keri Martinowich is a senior investigator and director of translational Neuroscience at the Lieber Institute for Brain Development and a Professor of Psychiatry and Neuroscience at Johns Hopkins University School of Medicine. She uses a cross-species approach to study how programs of gene expression in cell type- and spatially-defined neuronal populations contribute to circuit function and control of behaviors that are relevant for neuropsychiatric and neurodevelopmental disorders. Specifically, she uses genetic manipulation and viral transgenesis in combination with molecular, cellular, and systems-level recording techniques in animals, and integrates these data with cell type-specific and spatially-resolved transcriptomic data generated in postmortem human brain tissue.

Stephanie C. Hicks is an associate professor in the Department of Biostatistics and in the Department of Biomedical Engineering at Johns Hopkins University. She is an applied statistician working at the intersection of genomics and biomedical data science. Her research addresses computational challenges in single-cell genomics, epigenomics, and spatial transcriptomics leading to an improved understanding of human health and disease. Specifically, she develops computational methods using statistics and machine learning. She implements these methods in open-source software for the analysis of biomedical data.