# Spatial Transcriptomics Data Analysis Pipeline & Infrastructure

Published August 6, 2026 34 min read

# Spatial Transcriptomics Data Analysis Pipeline & Infrastructure

## Executive Summary

Spatial transcriptomics has moved from a niche method to a mainstream life-sciences data problem, and the infrastructure built to analyze it now largely determines whether a lab or a pharmaceutical R&D group can extract value from the technology. Annual peer-reviewed output on spatial transcriptomics stayed below 100 papers per year through the end of 2019, then accelerated sharply after the method was named a "Method of the Year," reaching a corpus of 1,467 studies published across 489 journals between 2006 and 2023 (Source: www.ncbi.nlm.nih.gov). The underlying market reflects that trajectory unevenly: MarketsandMarkets values the spatial genomics and transcriptomics market at **$554.5 million in 2024**, rising to **$995.7 million by 2029** at a 12.4% compound annual growth rate (CAGR) (Source: www.marketsandmarkets.com), while Precedence Research puts the narrower spatial transcriptomics segment at $469.36 million in 2025 growing to $1,569.03 million by 2034 at a 14.35% CAGR (Source: www.precedenceresearch.com), and Mordor Intelligence sizes the combined market at $0.75 billion in 2025 growing to $1.35 billion by 2030 (Source: www.mordorintelligence.com). No two major research firms agree closely on market size, and any figure cited from this space should be treated as directional rather than precise.

The core technical story of 2026 is that the data analysis pipeline, not the wet-lab assay, has become the bottleneck. **10x Genomics' Xenium** platform, an imaging-based system, produces raw internal sensor data on the order of **tens of terabytes per sample**, a scale the vendor's own documentation says is "not practically useful for reanalysis or storage" in its raw form (Source: www.10xgenomics.com). Once the onboard pipeline processes that stream, 10x's own public reference datasets show archived output directory sizes ranging from **3.5 GB** for a small tissue subset up to **144 GB** for a dense Xenium Prime sample, and as high as 165 GB for the densest full-section panel runs (Source: www.10xgenomics.com). By contrast, **Visium HD**, 10x's sequencing-based (next-generation sequencing, NGS) whole-transcriptome platform, resolves tissue into a continuous grid of 2 by 2 micron barcoded squares computationally binned up to 8 by 8 microns for analysis (Source: cdn.10xgenomics.com). Processed-output sizes should be estimated from representative datasets rather than inferred from readout type alone ( 10x Genomics. This report walks through the full pipeline architecture, from **Space Ranger** and **Xenium Onboard Analysis** at the vendor layer, through segmentation tools such as **Cellpose**, **StarDist**, and **Baysor**, to downstream frameworks including **Scanpy**, **Seurat**, **Squidpy**, **Giotto**, and the **SpatialData** framework built on Zarr and Parquet formats (Source: www.nature.com).

Infrastructure economics matter as much as software choice. Cloud storage for the resulting datasets ranges from Google Cloud Storage Standard at roughly $0.000027397 per gibibyte-hour (Source: cloud.google.com) to Azure Blob Storage's Hot tier at $0.0184 per GB for the first 50 terabytes (TB) per month (Source: azure.microsoft.com), while AWS HealthOmics prices genomic sequence storage at $0.005769 per gigabase per month in its active storage class (Source: aws.amazon.com) and can scale compute workflows across more than 100,000 concurrent virtual CPUs (vCPUs) (Source: aws.amazon.com). Large public consortia illustrate the scale involved in practice: the Human Tumor Atlas Network's (HTAN) first data release spans 8,425 biospecimens from 2,042 participants profiled with more than 20 molecular assays, with cloud-queryable assay files spanning more than 200 million cells (Source: www.nature.com), while the Human BioMolecular Atlas Program (HuBMAP) data portal hosts more than 9,232 public datasets across 25 data types and 29 organ classes as of mid-2026 (Source: arxiv.org). For life-sciences and AI advisory firms such as **IntuitionLabs**, which does not sell spatial transcriptomics hardware or software but designs data pipelines, ETL processes, and data warehouses for regulated pharmaceutical organizations (Source: intuitionlabs.ai), the practical challenge for adopting organizations is less about choosing a platform and more about building the storage tiering, compute orchestration, and cross-tool interoperability layer that spatial data at this scale demands.

## Introduction and Background

Spatial transcriptomics refers to a family of laboratory methods that measure gene expression while preserving the physical location of that expression within a tissue section, a capability that conventional single-cell RNA sequencing (scRNA-seq) discards when it dissociates tissue into a cell suspension. The method's rise has been unusually fast: publication counts remained under 100 per year through 2019 and then rose sharply, culminating in a corpus of 1,467 peer-reviewed papers and reviews published between 2006 and 2023 across 489 distinct journals, with Nature Communications alone publishing 107 of them (Source: www.ncbi.nlm.nih.gov).

What distinguishes this report's focus, the **data analysis pipeline**, from the underlying wet-lab assay is that spatial transcriptomics generates two fundamentally different classes of data depending on chemistry. Sequencing-based platforms such as **Visium** and **Visium HD** from 10x Genomics produce standard NGS read files (FASTQs) that are aligned and quantified into a spot-by-gene count matrix, an output format familiar to anyone who has worked with bulk or single-cell RNA sequencing. Imaging-based platforms such as **Xenium**, **CosMx Spatial Molecular Imager (SMI)** from Bruker/NanoString, and **MERSCOPE** from Vizgen instead capture terabytes of raw microscopy images that must be processed through cell segmentation and transcript-calling algorithms before they resemble a conventional expression matrix at all, a raw-imaging-data challenge shared with other high-content microscopy fields that have standardized on chunked, cloud-optimized formats (Source: www.ncbi.nlm.nih.gov). This distinction, NGS-based versus imaging-based data readout, is explicit in 10x Genomics' own platform comparison materials, which classify Xenium's readout as imaging-based versus the NGS-based readout of Chromium and Visium (Source: www.10xgenomics.com).

The practical consequence is that "spatial transcriptomics data analysis pipeline" is not a single workflow but a layered stack: on-instrument vendor software (Space Ranger, Xenium Onboard Analysis), image and segmentation tooling (Cellpose, StarDist, Baysor), downstream statistical frameworks (Scanpy, Seurat, Squidpy, Giotto), interoperable storage formats (OME-Zarr, SpatialData), and finally the cloud or on-premises compute and storage infrastructure tying it all together. Each layer has matured independently and at a different pace, which is why a lab that has successfully run a Visium experiment cannot assume its infrastructure will handle a Xenium or CosMx dataset without substantial re-architecture.

This report is organized to answer that infrastructure question directly. It first surveys the major spatial transcriptomics platforms and the raw data volumes they generate, then walks the analysis pipeline from raw signal to biological insight, then quantifies storage and cloud computing requirements with current provider pricing, then places the software ecosystem in market and adoption context, and finally examines named real-world deployments at academic medical centers and multi-institution atlas consortia. Readers evaluating a build-versus-buy decision for spatial omics infrastructure, or benchmarking an existing pipeline against field practice as of August 2026, should find each secondary question common to this space answered in a dedicated section below.

## Spatial Transcriptomics Platforms and Data Generation

The platform a lab chooses determines the shape of its data problem more than any downstream software decision. Four platforms dominate current deployments: 10x Genomics' **Visium HD** and **Xenium**, Bruker/NanoString's **CosMx SMI**, and Vizgen's **MERSCOPE**. Each trades off spatial resolution, gene panel breadth, and raw data volume differently.

**Visium HD** replaced the original Visium's 55-micron spots with a continuous lawn of barcoded oligonucleotides arranged in a grid of 2 by 2 micron squares, described by 10x as mapping whole-transcriptome messenger RNA (mRNA) expression at single-cell scale, and computationally binned up to an 8 by 8 micron analysis bin by default (Source: cdn.10xgenomics.com). The vendor's own platform-comparison table describes the assigned resolution simply as "transcripts assigned to 2 micron areas" prior to binning (Source: www.10xgenomics.com). Because Visium HD remains NGS-based, its capture areas measure 6.5 by 6.5 millimeters (mm) in the standard format and up to 11 by 11 mm with the CytAssist workflow that allows tissue placement on standard slides (Source: research.bidmc.org), and its original spot size for the earlier, non-HD product was 55 microns.

**Xenium** takes the opposite approach: it is described by 10x as delivering "high-plex in situ at subcellular resolution with nanometer precision," using targeted fluorescent probe hybridization and imaging rather than sequencing (Source: www.10xgenomics.com). Its largest commercial panel, the **Xenium Prime 5K**, targets 5,006 genes on a single panel and processes up to 472 square millimeters ($mm^2$) of tissue across two slides in under six days (Source: www.10xgenomics.com). A separate multiomic Xenium configuration adds a customizable protein panel of up to 480 genes and 27 proteins, distinct from the pure targeted gene-expression panels that scale to 5,000 genes (Source: www.10xgenomics.com).

**CosMx SMI**, now marketed by Bruker following its acquisition of NanoString's spatial business, enables high-plex analysis of over 18,000 genes from a single formalin-fixed paraffin-embedded (FFPE) slide, alongside a protein assay covering up to 68 proteins per slide (Source: brukerspatialbiology.com). NanoString's commercially launched **CosMx 6K Discovery** panel, introduced in February 2024, lets researchers measure over 6,000 RNA targets representing nearly every human biological pathway (Source: nanostring.com), specifically 6,175 RNA targets including more than 3,350 ligand-receptor pairs (Source: brukerspatialbiology.com), while a whole-transcriptome prototype panel reaches almost 19,000 genes at true single-cell and subcellular resolution (Source: nanostring.com). Bruker's own reference datasets illustrate typical scale: a public mouse brain dataset totals over 90,000 single cells and roughly 225 million transcripts, and a human frontal cortex dataset covering roughly 100 $mm^2$ totals over 194,000 single cells and about 331 million transcripts (Source: brukerspatialbiology.com).

**MERSCOPE Ultra** from Vizgen, built on multiplexed error-robust fluorescence in situ hybridization (MERFISH), is marketed as offering a tissue-wide view of up to 1,000 custom genes at single-cell resolution, with a lateral optical resolution specification of 100 nanometer (nm) pixel size (Source: vizgen.com), using a 27-bit codebook that supports up to 1,000 genes (Source: vizgen.com). Vizgen describes its underlying MERFISH 2.0 chemistry as localizing transcripts with nanometer precision to reveal spatial organization within tissue (Source: vizgen.com).

*Table 1* below summarizes the four platforms' resolution, panel breadth, and readout type side by side.

| PLATFORM | VENDOR | RESOLUTION | MAXIMUM PANEL SIZE | DATA READOUT TYPE |
|---|---|---|---|---|
| **Visium HD** | 10x Genomics | 2×2 µm barcoded squares, binned to 8×8 µm for analysis (Source: cdn.10xgenomics.com) | Whole transcriptome | NGS-based (sequencing) |
| **Xenium** | 10x Genomics | Subcellular, nanometer precision | Up to 5,006 genes (Xenium Prime 5K) | Imaging-based (in situ hybridization) (Source: www.10xgenomics.com) |
| **CosMx SMI** | Bruker / NanoString | Subcellular, single-cell | Over 18,000 genes (WTx) / 6,175-gene 6K Discovery (Source: brukerspatialbiology.com) | Imaging-based (in situ hybridization) |
| **MERSCOPE Ultra** | Vizgen | 100 nm pixel size, single-cell (Source: vizgen.com) | Up to 1,000 genes (27-bit codebook) | Imaging-based (MERFISH) |

The pattern in *Table 1*, whose figures are drawn from the vendor citations given in the prose above, is consistent across vendors: imaging-based platforms (Xenium, CosMx, MERSCOPE) trade sequencing depth for subcellular spatial precision, and that precision is exactly what inflates raw data volume, since every field of view must be captured, stitched, and stored as high-resolution microscopy imagery before any gene expression signal can be extracted. Sequencing-based Visium HD, in contrast, produces its spatial signal as sequencing reads, a data type the genomics field has spent two decades building infrastructure to move, store, and compress efficiently, a substantial and under-appreciated advantage when planning storage budgets.

## The Data Analysis Pipeline: From Raw Signal to Biological Insight

Every spatial transcriptomics platform ships with a vendor-authored primary analysis pipeline that performs the irreducible first pass over raw instrument data: image registration, cell or nucleus segmentation, transcript or read alignment, and count matrix generation. Downstream, open-source segmentation and analysis packages take over for biological interpretation.

For **Visium**, 10x Genomics' **Space Ranger** pipeline is the standard entry point. Its core `spaceranger count` command takes a reference genome, a microscope slide image, and sequencing read files (FASTQs) as inputs and produces feature-barcode matrices, clustering, and differential expression results (Source: www.10xgenomics.com). Space Ranger's image-processing stage solves two problems inherent to spatial NGS assays: deciding where tissue has actually been placed on the slide, and aligning the printed fiducial spot pattern that anchors sequencing barcodes to physical coordinates, using (since version 3.1) a feature-matching algorithm refined by a Max Mutual Information (MMI) method borrowed from the Insight Toolkit (ITK) (Source: www.10xgenomics.com). For transcript quantification, Space Ranger uses the widely used **STAR** aligner to perform splicing-aware alignment of transcript reads to the reference genome (Source: www.10xgenomics.com). Since version 4.0, a dedicated `spaceranger segment` pipeline adds nucleus and cell segmentation for Visium HD hematoxylin and eosin (H&E) images, outputting masks in TIFF and GeoJSON formats compatible with third-party tools (Source: www.10xgenomics.com).

For **Xenium**, the on-instrument software pipeline is **Xenium Onboard Analysis (XOA)**, which 10x describes as simultaneously collecting and processing Xenium In Situ Gene Expression data in real time as imaging occurs (Source: www.10xgenomics.com). XOA runs a custom neural network for nucleus segmentation trained on thousands of manually labeled 4',6-diamidino-2-phenylindole (DAPI) image patches spanning many tissue types, and when additional boundary and membrane stains are available it prioritizes a multimodal segmentation algorithm using custom deep-learning models to analyze the multi-channel stain images directly rather than simply expanding nuclei; in DAPI-only workflows, the fallback method instead expands nucleus boundaries by 5 micrometers, or until they encounter another cell's boundary, using a Voronoi-like tessellation (Source: www.10xgenomics.com). XOA's outputs include a cell-feature matrix restricted to molecule counts of gene features meeting a Phred-scaled quality score (Q-Score) of 20 or higher, plus secondary analysis such as clustering and differential expression (Source: www.10xgenomics.com). For labs revisiting an already-imaged sample, **Xenium Ranger** is 10x's off-instrument reanalysis tool, letting users relabel transcripts, resegment cells, or import third-party segmentation results without rerunning the instrument (Source: www.10xgenomics.com).

Vendor segmentation is not always the final word. Three open-source tools dominate independent segmentation work in spatial omics. **Cellpose**, published in Nature Methods, is a generalist deep-learning segmentation algorithm trained on a dataset of over 70,000 segmented objects, explicitly designed so it "does not require model retraining or parameter adjustments" across new image types (Source: www.nature.com). **StarDist**, introduced at the Medical Image Computing and Computer Assisted Intervention (MICCAI) conference in 2018, instead localizes cell nuclei via star-convex polygons, a shape representation its authors argue is much better suited to nuclei than conventional bounding boxes (Source: arxiv.org). **Baysor**, published in Nature Biotechnology, takes a different approach: rather than segmenting cells from an image first, it optimizes two-dimensional (2D) or three-dimensional (3D) cell boundaries by jointly considering the likelihood of transcriptional composition and cell morphology together, and its authors report this transcript-aware approach can, in some cases, nearly double the number of cells recovered compared to existing nucleus-based tools (Source: www.nature.com).

Once a count matrix exists, the field converges on a small set of downstream analysis frameworks. **Scanpy**, built jointly with the AnnData data structure, is described by its maintainers as a scalable toolkit for analyzing single-cell gene expression data (Source: scanpy.readthedocs.io). **Seurat**, the equivalent R package from the Satija Lab, is designed for quality control (QC), analysis, and exploration of single-cell RNA sequencing data and now ships dedicated vignettes for both sequencing-based and imaging-based spatial data (Source: satijalab.org). **Squidpy**, part of the scverse Python ecosystem, adds spatial-specific statistics such as neighborhood enrichment; its authors report their implementation running roughly tenfold faster than a comparable implementation in Giotto on one internal benchmark (Source: www.nature.com). **Giotto Suite**, the R-based alternative, provides tools to process, analyze, and visualize spatial multi-omics data "at all scales and multiple resolutions," spanning in situ hybridization, sequencing, and imaging-based technologies within one framework (Source: giottosuite.com).

Interoperability across these tools is increasingly handled by **SpatialData**, a scverse project its developers describe as "a data framework, a schema and a serialization format for uni- and multi-modal spatial omics datasets," with client libraries in Python, R, and JavaScript (Source: spatialdata.scverse.org). For labs seeking reproducible, containerized pipelines, the Nextflow-based **nf-core/spatialvi** community pipeline can process Visium data either directly from raw sequencing files by invoking Space Ranger internally, or from pre-existing Space Ranger output (Source: nf-co.re), and the broader nf-core ecosystem includes reusable modules such as a GPU-accelerated Cellpose module built to perform deep-learning-based cell segmentation on microscopy images at pipeline scale (Source: nf-co.re).

*Table 2* below organizes this software landscape by category, ecosystem, and function.

| TOOL | CATEGORY | LANGUAGE / ECOSYSTEM | PRIMARY FUNCTION |
|---|---|---|---|
| **Space Ranger** | Vendor primary pipeline | Standalone (10x Genomics) | Visium image registration, alignment, count matrix generation (Source: www.10xgenomics.com) |
| **Xenium Onboard Analysis / Xenium Ranger** | Vendor primary pipeline | Standalone (10x Genomics) | Real-time segmentation, transcript calling, off-instrument reanalysis (Source: www.10xgenomics.com) |
| **Cellpose** | Segmentation | Python | Generalist deep-learning cell/nucleus segmentation (Source: www.nature.com) |
| **StarDist** | Segmentation | Python | Star-convex polygon nucleus segmentation (Source: arxiv.org) |
| **Baysor** | Segmentation | Julia | Transcript-and-morphology joint-likelihood boundary optimization (Source: www.nature.com) |
| **Scanpy / AnnData** | Downstream analysis | Python | Data structure, clustering, differential expression (Source: scanpy.readthedocs.io) |
| **Seurat** | Downstream analysis | R | QC, clustering, sequencing- and imaging-based spatial vignettes (Source: satijalab.org) |
| **Squidpy** | Downstream analysis | Python (scverse) | Spatial statistics, neighborhood enrichment (Source: www.nature.com) |
| **Giotto Suite** | Downstream analysis | R | Multi-omic spatial toolbox across technologies (Source: giottosuite.com) |
| **SpatialData** | Data framework | Python / R / JavaScript (scverse) | Zarr/Parquet-based unified spatial data model (Source: spatialdata.scverse.org) |
| **nf-core/spatialvi** | Workflow manager | Nextflow | Reproducible, containerized Visium pipeline (Source: nf-co.re) |

As *Table 2* makes clear, drawing on the tool-by-tool citations given above, no single tool spans the full pipeline. A typical production pipeline in 2026 chains a vendor primary tool, an independent segmentation package when nucleus-only segmentation is insufficient, a scverse or Seurat-based downstream framework, and increasingly a workflow manager such as Nextflow to make the whole chain reproducible and containerized for regulated environments. This modularity is a deliberate response to how fast individual algorithms have improved: Cellpose, StarDist, and Baysor were all published within a roughly three-year window and each remains under active development, so pipelines hard-coded to a single segmentation method risk falling behind accuracy improvements that a modular, swappable architecture would have captured automatically.

## Data Volume, Storage, and Cloud Infrastructure Requirements

Raw data volume is the single most consequential planning variable for a spatial transcriptomics program, and it varies by an order of magnitude or more depending on platform and panel size. 10x Genomics' own Xenium documentation states plainly that the raw internal sensor data generated during imaging is on the order of tens of terabytes per sample and is "not practically useful for reanalysis or storage" in that raw form (Source: www.10xgenomics.com). After Xenium Onboard Analysis processes that raw sensor stream, 10x's own published example datasets show archived directory sizes spanning a small tissue subset at 3.5 GB up to a dense Xenium Prime sample at 144 GB (Source: www.10xgenomics.com). At the small end, 10x's estimation tables show a core needle biopsy covering just 0.01 cm² producing an RNA-only output of approximately 0.3 GB, while a

full 2.35 cm² Xenium Prime 5K tissue section at high transcript density can archive to as much as 165 GB (Source: www.10xgenomics.com), typically stored as Parquet transcript tables alongside pyramidal OME-TIFF morphology imagery, mirroring the Zarr-and-Parquet convention the SpatialData framework formalizes (Source: www.nature.com).

That imagery is where storage volume compounds fastest when it is not managed with cloud-optimized formats. The **OME-NGFF (Next Generation File Format)** specification, commonly implemented as **OME-Zarr**, addresses this by structuring large microscopy datasets as chunked, multi-resolution arrays; a peer-reviewed description of the format notes that its top-level directory abstraction "can represent an entire 1 Terabyte plate" of high-content imaging data within a single hierarchical structure (Source: www.ncbi.nlm.nih.gov). The same paper cites a full-resolution mouse brain dataset that consumes about 33 TB on disk, and notes that high-content cell painting screens commonly reach 100 terabytes, a scale the Image Data Resource addresses by converting to OME-Zarr (Source: www.ncbi.nlm.nih.gov). HuBMAP made the same migration, moving from OME-TIFF to OME-Zarr specifically to enable flexible, client-side browsing of multi-terabyte datasets directly within its data portal, using the Viv visualization library (Source: www.ncbi.nlm.nih.gov). The **SpatialData** framework standardizes on the same Zarr foundation, using an on-disk representation based on Zarr and Parquet data formats that allows datasets to be either downloaded in full or accessed directly from public cloud storage such as Amazon Simple Storage Service (S3), without requiring a local copy at all (Source: www.nature.com).

**Cloud infrastructure** providers have responded with genomics-specific storage and compute products, though pricing structures differ across the three major hyperscalers. **AWS HealthOmics** prices its Sequence Store active storage class at $0.005769 per gigabase per month, while its Variant and Annotation Store charges $0.035 per GB per month with a 30-day minimum retention (Source: aws.amazon.com), and the service can scale bioinformatics workflows across more than 100,000 concurrent vCPUs (Source: aws.amazon.com). **Azure Blob Storage**, commonly used for spatial imaging archives, prices its Hot access tier at $0.0184 per GB for the first 50 TB stored per month in at least one published region tier (Source: azure.microsoft.com). **Google Cloud Storage** prices its Standard class at approximately $0.000027397 per gibibyte-hour, roughly two cents per GB per month (Source: cloud.google.com); Google deprecated its dedicated Cloud Life Sciences product, with Google Cloud Batch now positioned as "a comprehensive successor" for genomics workloads (Source: docs.cloud.google.com). Microsoft's **Microsoft Genomics** service is described by Microsoft as "secure, compliant (ISO certified, HIPAA compliant), and covered under Microsoft BAA," and performs secondary analysis such as Burrows-Wheeler Aligner (BWA) alignment and Genome Analysis Toolkit (GATK) variant calling (Source: www.microsoft.com), while **Cromwell on Azure** orchestrates dynamic provisioning of compute resources via Azure Batch, integrating with Blob storage for workflow data (Source: azure.microsoft.com). Microsoft has separately projected that genomics data volumes and associated computation will require "tens of exabytes and trillions of core hours" within a five-year horizon, a forecast that implicitly includes the imaging-heavy spatial omics workloads discussed throughout this report (Source: azure.microsoft.com).

*Table 3* below summarizes the three major cloud providers' relevant storage and compute pricing as of the dates their respective pages were accessed in August 2026.

| PROVIDER | SERVICE | PRICE | NOTES |
|---|---|---|---|
| **AWS** | HealthOmics Sequence Store (active storage) | $0.005769 per gigabase/month (Source: aws.amazon.com) | Genomics-specific managed store |
| **AWS** | HealthOmics Variant & Annotation Store | $0.035 per GB/month, 30-day minimum | Structured variant data |
| **AWS** | HealthOmics compute | Scales to 100,000+ concurrent vCPUs | Managed bioinformatics workflow execution |
| **Microsoft Azure** | Blob Storage, Hot tier | $0.0184 per GB for first 50 TB/month | General-purpose object storage, one published region tier |
| **Microsoft Azure** | Microsoft Genomics | Secure, compliant (ISO certified, HIPAA compliant), Microsoft BAA-covered (Source: www.microsoft.com) | Secondary analysis: alignment, variant calling |
| **Google Cloud** | Cloud Storage, Standard class | ~$0.000027397 per GiB-hour (~$0.02/GB-month) | General-purpose object storage |
| **Google Cloud** | Cloud Batch | Successor to deprecated Cloud Life Sciences (Source: docs.cloud.google.com) | Batch compute for genomics workflows |

Table 3 presents service-specific list-price metrics, not a normalized cross-cloud storage comparison. AWS HealthOmics Sequence Store is billed per gigabase of biological sequence data, while Google Cloud Storage and Azure Blob Storage are object-storage services whose rates vary by region and configuration; request, retrieval, replication, and data-transfer charges are additional. A provider choice therefore requires a workload- and region-specific estimate. What the table does not capture, because no vendor publishes it directly, is compute cost for image-heavy segmentation. Cellpose's own documentation states plainly that "runtime and memory usage increases with the data size," and recommends reducing the `batch_size` parameter, whose default value is 8, specifically to avoid exhausting graphics processing unit (GPU) memory on large whole-slide images (Source: cellpose.readthedocs.io). Compute cost should be estimated from a specific workload and cloud configuration; runtime and memory usage alone do not establish a compute-to-storage cost relationship.

Large multi-institution consortia offer the clearest evidence of what these demands look like at scale. The Human Tumor Atlas Network integrates its data storage with external cloud providers, specifically AWS S3 and Google Cloud Storage, explicitly to minimize data egress costs, and makes more than 850 spatial and single-cell assay files queryable through Google BigQuery rather than requiring bulk download (Source: www.nature.com). This cloud-native, query-first architecture, rather than a traditional download-and-analyze-locally model, is becoming the default pattern for spatial omics data management precisely because the underlying files are simply too large to move casually.

## Analysis of Key Segments: Market Structure and Tool Adoption

The spatial transcriptomics ecosystem splits into three commercial and technical segments: instrument and consumables vendors, downstream software (open-source and vendor-proprietary), and cloud infrastructure providers, each with a different revenue and adoption dynamic, as the market-sizing disagreement discussed below illustrates (Source: www.marketsandmarkets.com).

On the instrument side, 10x Genomics' public financial disclosures provide a company-reported view of its spatial business, but do not establish its market share relative to competitors. The company's fourth-quarter and full-year 2025 financial statement lists spatial consumables revenue of $143.977 million for full-year 2025 versus $121.124 million for 2024, and spatial instruments revenue of $34.108 million for full-year 2025 versus $57.503 million for 2024, part of total full-year 2025 revenue of $642.8 million, a 5% increase over 2024's $610.8 million (Source: www.prnewswire.com) (Source: www.prnewswire.com). For fiscal year 2026, the company has guided total revenue to $600 to $625 million (Source: www.prnewswire.com), and first-quarter 2026 revenue was $150.8 million, a 3% year-over-year decrease the company attributed largely to a one-time prior-year licensing item, alongside continued double-digit growth in spatial consumables (Source: www.prnewswire.com). Notably, 10x's Q1 2026

earnings materials disclose an upcoming platform called **Atera**, described as "engineered to deliver spatial whole-transcriptome analysis," slated to ship in the second half of 2026 (Source: s28.q4cdn.com), a signal that the resolution-versus-throughput tradeoff described earlier is itself a moving target.

Market-sizing estimates vary widely, a discrepancy worth stating plainly. MarketsandMarkets values the spatial genomics and transcriptomics market at $554.5 million in 2024, projecting $995.7 million by 2029 at a 12.4% CAGR. Precedence Research puts the spatial transcriptomics segment specifically at $469.36 million in 2025, forecasting $1,569.03 million by 2034 at a 14.35% CAGR (Source: www.precedenceresearch.com) (Source: www.precedenceresearch.com). Mordor Intelligence sizes the 2025 combined market at $0.75 billion growing to $1.35 billion by 2030, and reports spatial transcriptomics specifically led with 54.80% of that market's 2025 revenue share (Source: www.mordorintelligence.com). No two of these three estimates agree on baseline size or CAGR, likely reflecting differing scope definitions as much as differing methodology; readers should treat any single figure as directional rather than precise.

Software adoption is harder to quantify than instrument revenue because open-source tool usage is rarely metered. The clearest available signal is publication growth itself: annual output remained below 100 papers through 2019, while the bibliometric study's corpus contained **1,467 studies cumulatively across 2006–2023** (Source: www.ncbi.nlm.nih.gov). That growth coincides with the release dates of Cellpose (2020), Squidpy (2021) (Source: www.nature.com), and Baysor (2021). Trade publication GenomeWeb has separately run a reader survey of spatial biology technology usage, finding that "awareness and usage of different spatial biology assays is broad" across the surveyed community, though the full respondent count and platform-by-platform breakdown sit behind a subscription paywall (Source: www.genomeweb.com). Cost at the point of use is often more decision-relevant than list price, and is rarely published by vendors: the University of Chicago's genomics core facility publishes a full-service Xenium assay rate of $2,500 per slide for internal researchers (Source: voices.uchicago.edu), a concrete anchor for budgeting a pilot study even though 10x Genomics does not publish list pricing on its own site.

## Data Analysis and Evidence

Three quantitative threads, already introduced above, anchor any infrastructure planning exercise for a new spatial transcriptomics program: raw data volume per sample, the growth trajectory of the underlying research field, and the fragmented state of market-size forecasting.

On data volume, the roughly fortyfold range between a small public Xenium example dataset and a full, dense Xenium Prime section, before even accounting for the tens-of-terabytes raw sensor data most labs never retain, is not unique to spatial transcriptomics. Independent confirmation of that order of magnitude comes from adjacent high-content imaging fields, where a peer-reviewed OME-NGFF paper documents comparable full-resolution datasets reaching 33 TB for a single mouse brain and up to 100 TB for screening campaigns (Source: www.ncbi.nlm.nih.gov), reinforcing that spatial transcriptomics is a specific instance of a broader raw-imaging-data problem the field has been solving for years with formats like OME-Zarr.

On field growth, the bibliometric analysis of 1,467 studies published from 2006 through 2023 across 489 journals provides a rare, methodologically transparent measure of adoption momentum, independent of any single vendor's disclosures (Source: www.ncbi.nlm.nih.gov). The inflection point after 2019 coincides almost exactly with the commercial launch window for Visium (2019) and the subsequent launches of Xenium (2022) and the CosMx and MERSCOPE imaging platforms, suggesting instrument commercialization and the open-source software ecosystem grew in lockstep rather than software lagging hardware.

On market forecasting, the three-way disagreement among MarketsandMarkets, Precedence Research (Source: www.precedenceresearch.com), and Mordor Intelligence should itself be treated as evidence: it indicates a market still young enough that research firms have not converged on shared scope definitions. 10x Genomics disclosed approximately $178 million in 2025 spatial instrument and consumables revenue, but that company disclosure does not establish market share or support a conclusion about overall vendor fragmentation ( 10x Genomics 2025 Form 10-K.

## Case Studies and Real-World Examples

Five named deployments illustrate how the pipeline architecture and infrastructure choices described above play out at production scale, spanning federal atlas consortia, academic medical center core facilities, and industry-academia drug discovery collaborations, several of which build on the CosMx and Xenium platforms profiled earlier (Source: brukerspatialbiology.com).

## The Human Tumor Atlas Network (HTAN)

HTAN, a National Cancer Institute-funded, multi-institution consortium, released its first public dataset comprising 8,425 biospecimens from 2,042 research participants, profiled with more than 20 distinct molecular assays including spatial transcriptomics (Source: www.nature.com). As of September 2024, the network's scope had grown to two pilot projects, ten atlases, and four trans-network projects (Source: www.nature.com). Rather than distributing raw files for local download by default, the consortium makes hundreds of spatial and single-cell assay files, collectively spanning more than 200 million cells, directly queryable through Google BigQuery (Source: www.nature.com). Notably, the consortium's own published methods paper states that individual HTAN research centers had to build custom bioinformatics pipelines because spatial profiling assays are cutting-edge and non-standardized (Source: www.nature.com). HTAN's data has already generated 113 dbGaP (database of Genotypes and Phenotypes)-approved data-use plans from outside researchers reusing the dataset, a concrete measure of downstream demand for pre-processed spatial data.

## The Human BioMolecular Atlas Program (HuBMAP)

HuBMAP, a National Institutes of Health (NIH)-funded initiative mapping the human body at cellular resolution, operates a public data portal that, as of June 2026, hosts 9,232 public datasets spanning 25 distinct data types across 29 organ classes and 498 donors (Source: arxiv.org). The consortium enforces uniform, data-type-specific processing pipelines and rigorous quality control to ensure cross-lab, cross-organ, and cross-donor comparability, a design choice made necessary by the sheer diversity of assay types and originating labs feeding into a single shared portal (Source: arxiv.org). The live portal invites users to "filter your way through 9,000+ datasets across 40+ facets," and pairs its catalog with cloud-backed JupyterLab-based analysis workspaces so researchers can compute directly against hosted data rather than downloading it locally first (Source: portal.hubmapconsortium.org).

## Spatial Atlas of Human Anatomy (SAHA)

Positioned by its authors as filling a gap left by dissociated-cell atlases, the Spatial Atlas of Human Anatomy (SAHA), led by researchers at Weill Cornell Medicine and collaborating institutions, profiled 15,915,616 cells from more than 100 healthy donors at a spatial resolution of 50 nanometers, combining 9.5 million tissue-resident cells with 6 million circulating blood cells (Source: www.ncbi.nlm.nih.gov). Its CosMx RNA component alone, covering 94 tissue samples, profiled 2.9 million cells and captured over 524 million individual transcript measurements (Source: www.ncbi.nlm.nih.gov). The study's authors explicitly justify SAHA's necessity by noting that "existing atlases like the Human Cell Atlas... [and] Human Tumor Atlas Network... are largely derived from dissociated cells," meaning they discard the spatial architecture that platforms like CosMx, Xenium, and MERSCOPE were built to preserve (Source: www.ncbi.nlm.nih.gov).

## Academic Medical Center Core Facilities: MD Anderson, MSKCC, and Karolinska Institutet

At the institutional level, dedicated spatial biology core facilities show how individual cancer centers have operationalized these pipelines. The University of Texas MD Anderson Cancer Center's **Advanced Spatial Genomics (ASG) Core** was established specifically to provide spatial transcriptomics and genomics services to cancer researchers across the institution (Source: www.mdanderson.org), funded by a $3 million grant from the Cancer Prevention and Research Institute of Texas (CPRIT, grant RP240497) alongside institutional funding (Source: www.mdanderson.org). At Memorial Sloan Kettering Cancer Center (MSKCC), the **Integrated Genomics Operation (IGO)** core has offered spatial transcriptomics services since 2018 alongside single-cell genomics and long-read sequencing, and states its own software engineers provide end-to-end bioinformatics pipeline support (Source: www.mskcc.org). In Sweden, Karolinska Institutet's **KIGene Spatial Biology** core runs both the CosMx SMI and GeoMx Digital Spatial Profiler (DSP) platforms, enabling researchers to analyze up to the whole transcriptome on fixed and FFPE tissue depending on assay (Source: ki.se). Together, these institutions show a consistent pattern: even well-resourced academic medical centers treat spatial transcriptomics analysis as requiring dedicated, in-house bioinformatics staff rather than something individual labs can absorb into existing pipelines without additional investment.

## Industry Collaborations: TISHUMAP and PharosAI

Two 2025-era industry-academia collaborations illustrate spatial transcriptomics' expanding role in pharmaceutical drug discovery. 10x Genomics and Singapore's A*STAR Genome Institute announced the **TISHUMAP** study in July 2025 to analyze up to 2,500 FFPE clinical tumor tissue samples, including gastric, liver, and colorectal cancers, on the Xenium platform for AI-driven drug target discovery (Source: www.prnewswire.com). The collaboration's scope explicitly includes co-developing custom gene panels alongside "smart software pipelines designed to handle the massive

datasets" the study will generate, an acknowledgment from the vendor that off-the-shelf software is insufficient at this scale (Source: www.prnewswire.com). Separately, a United Kingdom research consortium branded **PharosAI**, spanning King's College London, Queen Mary University of London, and NHS trusts, announced it will use the Xenium spatial platform to build multimodal cancer datasets from decades of archived National Health Service (NHS) patient samples for pairing with AI drug-discovery models (Source: www.genomeweb.com). Both collaborations underscore a pattern visible in the core-facility examples above: as spatial datasets scale into the hundreds or thousands of samples, the software pipeline itself becomes a named deliverable of the project, not an afterthought to instrument procurement.

## Implications and Future Directions

Several structural trends emerge from the evidence assembled above. First, raw data volume growth is outpacing the maturity of standardized storage tooling. OME-Zarr and SpatialData address the imaging and multi-modal integration problems reasonably well for labs willing to adopt scverse-aligned tooling, but the fact that HTAN, HuBMAP, and individual core facilities each describe building custom bioinformatics pipelines rather than adopting a single shared standard suggests the field has not yet converged the way, for example, BAM/CRAM file formats standardized bulk sequencing a decade earlier. Organizations planning multi-year spatial omics programs should expect to allocate engineering effort to pipeline interoperability, not just instrument procurement, and should budget for the likelihood that today's chosen segmentation tool (Cellpose, StarDist, or Baysor) will be supplemented or replaced within a two-to-three-year window given how recently all three were published (Source: www.nature.com).

Second, the sequencing-versus-imaging split described earlier is a durable architectural fault line, not a temporary one. Visium HD's NGS-based readout inherits decades of genomics storage and compute tooling, while Xenium, CosMx, and MERSCOPE's imaging-based readouts inherit the OME-Zarr and high-content screening tooling built for microscopy (Source: www.ncbi.nlm.nih.gov), a fundamentally different software lineage. 10x Genomics' announced **Atera** platform, positioned for shipping in the second half of 2026 as a spatial whole-transcriptome system, suggests vendors are actively trying to blur this line, which will likely push raw data volumes higher still rather than lower.

Third, cloud infrastructure economics favor a query-first, storage-tiered architecture over bulk local download, as demonstrated concretely by HTAN's BigQuery-queryable assay files and SpatialData's direct cloud-native S3 access. For organizations without in-house bioinformatics infrastructure teams, particularly pharmaceutical and life-sciences companies evaluating whether to build spatial omics data pipelines internally or engage outside expertise, the relevant lesson from HTAN, HuBMAP (Source: arxiv.org), and the academic core facilities profiled above is that pipeline engineering, ETL (extract, transform, load) design, and cloud storage tiering are now first-class requirements alongside the wet-lab assay itself, not secondary IT concerns. This is precisely the kind of data engineering and integration work, spanning data pipeline development, data warehousing, and multi-source integration across platforms such as Databricks and Snowflake, that consultancies like **IntuitionLabs** provide to pharmaceutical and life-sciences organizations, without being a spatial transcriptomics instrument or software vendor themselves (Source: intuitionlabs.ai). A life-sciences organization evaluating whether to route a new spatial omics program through a dedicated bioinformatics hire, a core facility partnership, or an external data engineering advisor should weigh the sustained, multi-year infrastructure commitment documented across every case study in this report, since none of the profiled institutions treated spatial data management as a one-time setup task.

Fourth, market forecasting immaturity, visible in the wide disagreement among MarketsandMarkets, Precedence Research, and Mordor Intelligence (Source: www.mordorintelligence.com), is itself informative for procurement planning: organizations should not anchor multi-year infrastructure budgets to any single third-party market forecast, and should instead track platform-level signals, such as 10x Genomics' own quarterly spatial consumables revenue trend, publication growth in their specific disease or tissue area of interest, and core-facility pricing at peer institutions, as more reliable near-term indicators of where the field is actually heading.

## Frequently Asked Questions (FAQs)

**What is a spatial transcriptomics data analysis pipeline?** It is the layered software stack that converts raw instrument output, either sequencing reads for NGS-based platforms like Visium HD or microscopy images for imaging-based platforms like Xenium, CosMx, and MERSCOPE, into a spatially resolved gene expression matrix ready for biological interpretation. It typically includes vendor primary analysis software such as Space Ranger (Source: www.10xgenomics.com), a segmentation stage such as StarDist (Source: arxiv.org), and downstream analysis frameworks such as Scanpy, Seurat, Squidpy, or Giotto.

**How much data does a spatial transcriptomics experiment generate?** It varies enormously by platform and panel size. A single Xenium slide's archived output, per 10x Genomics' own published example datasets, ranges from about 3.5 GB for a small tissue subset to 144 GB for a dense sample, while the pre-processed raw sensor data underlying that archive can reach tens of terabytes per sample (Source: www.10xgenomics.com).

**What is the difference between Xenium and Visium data?** Xenium is imaging-based, producing subcellular-resolution transcript locations from targeted fluorescent probes, while Visium and Visium HD are NGS-based, producing standard sequencing reads aligned into a spot- or bin-level count matrix (Source: www.10xgenomics.com). Archived output size is workflow- and sample-dependent: 10x states that Xenium output size varies with tissue shape, cell count, decoded transcript count, and the proportion of high-quality transcripts ( 10x Genomics. A comparison with Visium requires matched tissue area, analysis binning, and retained output components.

**What cloud infrastructure is used for spatial transcriptomics data?** All three major hyperscalers offer relevant services: AWS HealthOmics for genomics-specific storage and compute (Source: aws.amazon.com), Azure Blob Storage plus Azure Batch and Cromwell for workflow execution, and Google Cloud Storage plus Google Cloud Batch, each detailed with current pricing in the infrastructure section above. Large consortia such as HTAN use a mix of these providers alongside Google BigQuery to make data directly queryable.

**What software tools are used to analyze spatial transcriptomics data?** The most widely used downstream frameworks are Scanpy and AnnData in Python (Source: scanpy.readthedocs.io), Seurat in R (Source: satijalab.org), and the spatial-specific Squidpy and Giotto packages built on top of them (Source: giottosuite.com), with SpatialData increasingly serving as the shared interoperability layer between them.

**How is spatial omics data managed and shared at scale?** Large programs increasingly favor cloud-native, query-first architectures over bulk file download. HuBMAP, for example, pairs its 9,232-dataset public portal with cloud-backed analysis workspaces (Source: arxiv.org), and HTAN makes hundreds of assay files directly queryable via BigQuery rather than requiring local copies.

## Conclusion

Spatial transcriptomics data analysis in 2026 sits at the intersection of two previously separate infrastructure disciplines: genomics sequencing pipelines and high-content microscopy image management, and neither discipline's existing tooling was built to handle both simultaneously at the scale the field now requires. The evidence assembled in this report shows a field whose publication growth accelerated sharply after 2020, whose flagship imaging platforms generate raw data on the order of tens of terabytes per sample even as their archived, analysis-ready outputs shrink to tens of gigabytes, and whose three leading commercial market forecasts disagree with each other by a wide enough margin that no single figure should anchor a procurement decision. What is consistent across every named deployment examined, from the Human Tumor Atlas Network and the Human BioMolecular Atlas Program down to individual academic medical center core facilities at MD Anderson, Memorial Sloan Kettering, and Karolinska Institutet, is that none of them treated the analysis pipeline as a solved, off-the-shelf problem; each built or substantially customized its own bioinformatics infrastructure. For organizations evaluating a spatial transcriptomics program, the platform decision (Visium HD versus Xenium versus CosMx versus MERSCOPE) is only the first of several major infrastructure decisions, and the storage tiering, segmentation tooling, downstream analysis framework, and cloud compute architecture chosen alongside it will likely matter more to the program's long-term success than the instrument itself.

---

Tags: spatial transcriptomics, data analysis pipeline, 10x genomics xenium, visium hd, spatial omics, bioinformatics software, cloud computing infrastructure, single cell genomics, data storage requirements

## IntuitionLabs - Industry Leadership & Services

**North America's #1 AI Software Development Firm for Pharmaceutical & Biotech:** IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

**Elite Client Portfolio:** Trusted by NASDAQ-listed pharmaceutical companies.

**Regulatory Excellence:** Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

**Founder Excellence:** Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

**Custom AI Software Development:** Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

**Private AI Infrastructure:** Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

**Document Processing Systems:** Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

**Custom CRM Development:** Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

**AI Chatbot Development:** Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

**Custom ERP Development:** Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

**Big Data & Analytics:** Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

**Dashboard & Visualization:** Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

**Leading Claude & ChatGPT AI Enablement and Training:** Help biotech and pharmaceutical teams adopt Claude and ChatGPT through practical training, governed workflows, use-case development, and implementation designed for regulated environments.

**AI Consulting & Training:** Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at intuitionlabs.ai/contact for a consultation.

## DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.