Epigene Labs | WHITEPAPER

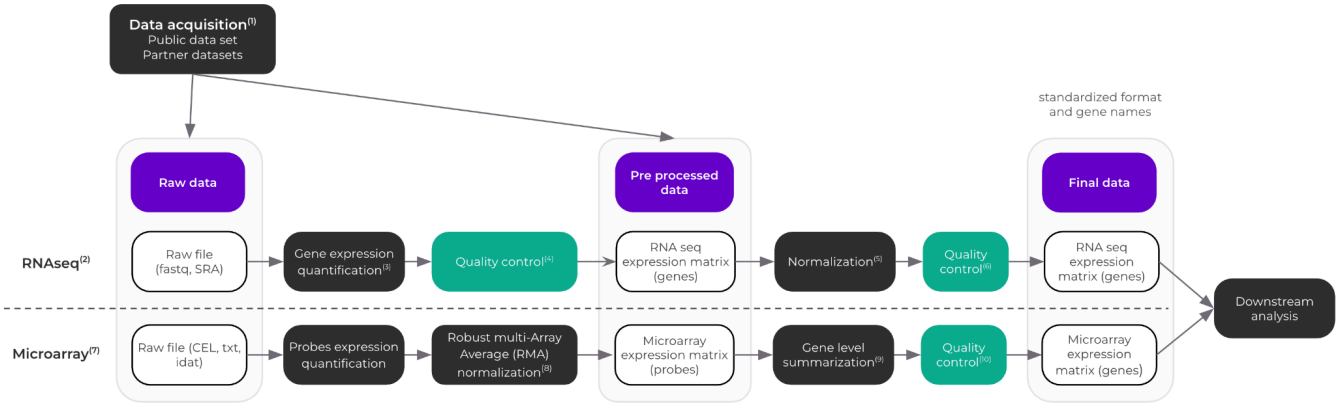
Bulk transcriptomics processing

Guillaume Appé, Maximilien Colange, Léa Meunier, Solène Weill, Abdelkader Behdenna, Akpéli Nordor

September 2025

contact@epigenelabs.com

Graphical abstract



Navigate to the section:

Executive summary

Transcriptomic datasets hold enormous potential for oncology, yet much of this value remains locked due to fragmentation across platforms, formats, and studies. Without harmonization, comparisons are difficult, reproducibility suffers, and large-scale analyses become unreliable.

At Epigene Labs, we have developed a scalable framework to systematically process, harmonize, and analyze bulk transcriptomic data. By unifying diverse datasets under a common structure and quality standard, we address one of the most persistent challenges in translational research: making disparate data comparable and actionable.

Our framework integrates data from major repositories such as GEO[1], GDC[2], and ArrayExpress[1], with a focus on oncology. We apply rigorous curation and metadata harmonization, enabling robust integration across sources. Both microarray and RNA-seq data are supported, with pipelines built on state-of-the-art open-source tools and adapted for scale.

To maximize statistical power and minimize dataset-specific biases, we build custom atlases – large, harmonized collections of transcriptomic data organized across diseases, tissues, and platforms. These atlases make it possible to analyze rare subpopulations, validate biomarkers across heterogeneous cohorts, and support meta-analyses at a scale not achievable with individual datasets.

Applications in precision oncology include:

- Differential gene expression analysis to uncover tumor-specific markers or predictors of therapy response.
- Outcome association studies, linking expression profiles to survival and recurrence data.
- Functional enrichment analyses with tools such as GSEA and Enrichr, revealing dysregulated pathways.
- Biomarker validation and patient stratification, enabled by large, diverse datasets.

This work is underpinned by our open-source ecosystem, notably the Python-based InMoose toolkit, which ensures interoperability, transparency, and reproducibility.

While bulk transcriptomics provides essential system-level insights, it inherently averages out cellular heterogeneity. To complement this, Epigene Labs is extending its pipelines to single-cell RNA-seq. A dedicated white paper highlighting our scRNA-seq methodology will be published soon.

By harmonizing transcriptomic data at scale, we aim to unlock the full potential of existing datasets, accelerate biomarker discovery, and support the development of precision oncology therapeutics.

Introduction

Transcriptomics is the study of the transcriptome, the complete set of RNA molecules produced by an organism. Because RNA sits between DNA and proteins in the central dogma of biology, it captures a uniquely dynamic picture of cellular activity. It reflects not only genetic information but also regulatory states, environmental influences, and disease-driven alterations.

The transcriptome includes both protein-coding RNAs and non-coding RNAs such as lncRNAs, miRNAs, and snoRNAs. These non-coding molecules, once considered biological noise, are now recognized as critical regulators of gene expression and disease mechanisms. Studying the transcriptome, therefore, provides insights into how genomic and epigenomic variations translate into cellular behavior.

While proteomics is the most direct way to study cellular phenotypes, comprehensive protein profiling remains technically demanding and costly due to the chemical complexity of proteins and their modifications. By contrast, advances in next-generation sequencing (NGS) since the late 2000s have made transcriptomics far more accessible and scalable[1]. RNA-based profiling not only costs less but also expands the analytical scope beyond fixed probe sets used in microarrays (Fig.1). Researchers can now explore the entire transcriptome, detect novel isoforms, and quantify expression at unprecedented resolution.

The explosion of transcriptomic data in public and private repositories[2] has transformed the research landscape. Thousands of datasets across tissues, diseases, and experimental conditions are available, fueling *in silico* research for target discovery, biomarker development, and hypothesis generation. Proteomic data, though growing, is still more limited, often context-specific, and harder to integrate across studies. This positions transcriptomics not only as a standalone tool but also as a key complement to proteomics.

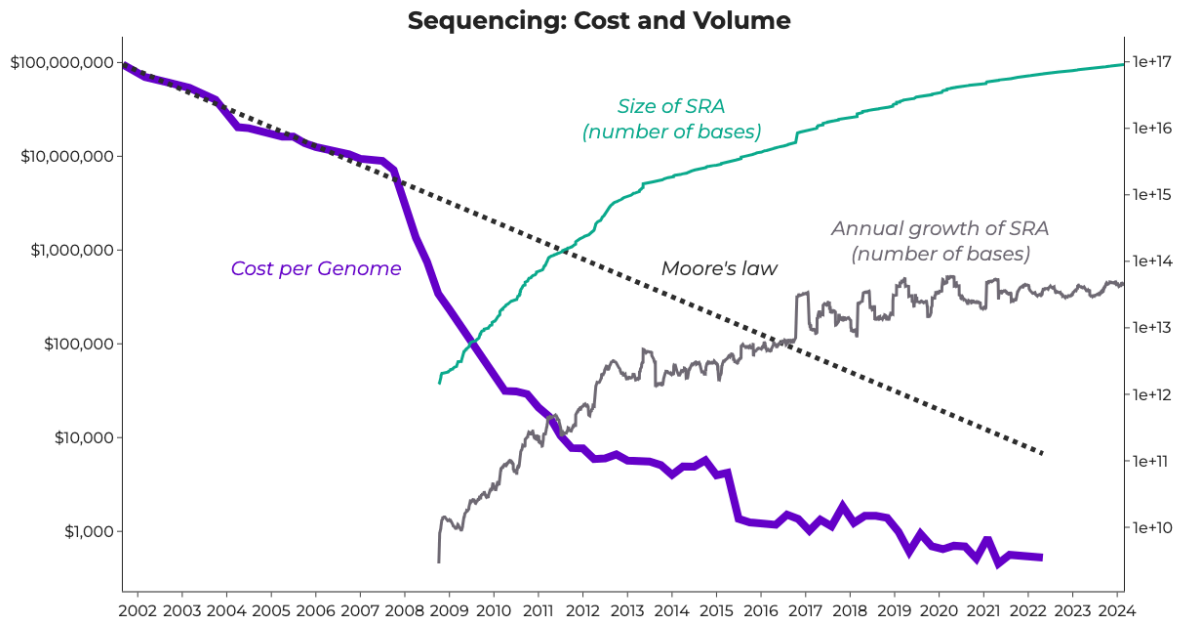
Today, bulk transcriptomics is typically performed with three main technologies:

- **Microarrays**: targeted detection of transcripts using predefined probes, limited to known sequences, but cost-effective and widely available.
- **RNA-seq**: unbiased sequencing of all transcripts in a sample, enabling broader discovery and higher sensitivity.
- **Targeted assays** (e.g., qRT-PCR, NanoString): often used for validation or clinical settings.

In oncology, the value of transcriptomics lies in its ability to uncover gene expression signatures that inform prognosis, stratify patients, and reveal therapeutic opportunities. However, this potential depends on the reliability of both molecular and clinical data. Without consistent preprocessing and harmonization, even the largest datasets can yield misleading results.

This white paper focuses on the processing of bulk transcriptomic data, microarray and RNA-seq, detailing how Epigene Labs addresses the technical challenges of data acquisition, normalization, and integration. Clinical data curation, equally critical for interpretability, will be the subject of a dedicated companion paper.

Figure 1. Sequencing cost and volume evolution (sources: NHGRI and SRA)



Data acquisition

Transcriptomic research depends not only on high-quality molecular data but also on consistent and accessible sources. The reliability of downstream analyses hinges on how datasets are acquired, curated, and prepared. Public repositories have made thousands of transcriptomic datasets available to the community, but their diversity in format, scope, and quality makes standardization essential.

At Epigene Labs, we prioritize oncology-focused datasets and integrate them into harmonized pipelines that ensure reproducibility and comparability across studies. This begins with understanding where the data comes from, how it is formatted, and what constraints govern its use.

Key databases

Several major repositories provide access to transcriptomic datasets, each with its own focus and data structures. Our pipelines are designed to interact with these resources efficiently, using dedicated tools and APIs.

Genomic Data Commons[3] (GDC)

The GDC is a cornerstone for oncology research. It aggregates data from flagship cancer projects such as TCGA, CPTAC, TARGET, and CMI, covering more than 30 cancer types. All data are processed with standardized pipelines and reference files, ensuring a high level of homogeneity. It provides a reliable anchor for large-scale oncology studies.

Database of Genotypes and Phenotypes[5] (dbGaP)

The National Institute of Health (NIH) maintains dbGaP, a repository for genetic and phenotypic data spanning multiple omics. While it is invaluable for comprehensive studies, the platform is less suited for rapid integration of exploratory datasets due to access restrictions.

Gene expression omnibus[6] (GEO)

The GEO, run by the National Center for Biotechnology Information (NCBI), is one of the most widely used repositories. It stores data from microarrays, RNA-seq, and other experimental platforms.

A notable recent development is NCBI's initiative to reprocess RNA-seq data uniformly from FASTQ files in the Sequence Read Archive (SRA). Since 2024, this has significantly improved the quality and comparability of public RNA-seq datasets.

ArrayExpress[8]

Developed by EMBL-EBI, ArrayExpress complements GEO and GDC by hosting functional genomics data.

	GDC	dbGaP	GEO	ArrayExpress
Data types	Mix of raw and processed data	Raw data	Mix of raw and processed data	Mix of raw and processed data
Searchability	By projects and experimental strategies	By projects or filters	MeSH terms	Experiment Factor Ontology terms
Access	Mix of controlled and open access data	Mix of controlled and open access data	Mix of controlled and open-access data	Mix of controlled and open access data
Integration	Through dbGaP and TCGAbiolinks R package[4]	Manual file card download and scripting with SRA toolkit	GEOparse Python package[7]	R package[9] or direct API queries

Data formats

Transcriptomic datasets can be shared in various formats, reflecting different stages of the processing pipeline. Understanding these formats is crucial for integration.

- Raw data are stored as fluorescence intensity values files for microarray, and as FASTQ (raw reads) or BAM files (aligned reads) for RNA-seq. Access to raw data is often controlled due to privacy concerns.
- Pre-processed data are available as expression matrices, usually in open access. These matrices summarize transcript or gene expression across samples. Depending on the source, they may be normalized or unnormalized.

Epigene Labs integrates both raw and preprocessed data. Raw files allow maximum control but require significant computation, while preprocessed matrices make large-scale integration faster and more scalable.

Our approach

By systematically combining multiple sources, we ensure that:

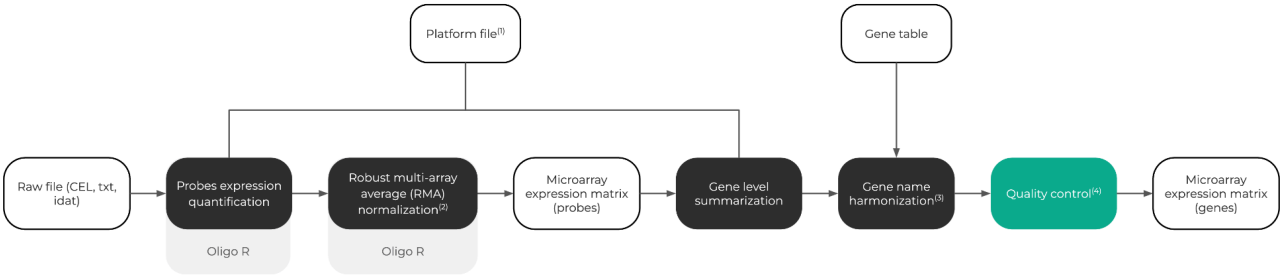
1. Coverage is maximized – large datasets spanning diverse cancers and conditions.
2. Quality is controlled – harmonized pipelines correct for differences in platforms and annotation versions.
3. Metadata is curated – enabling datasets to be linked meaningfully with clinical outcomes.

This foundation allows us to move from fragmented data silos toward coherent, disease-specific atlases that empower oncology research.

Data processing

Microarray

Figure 2. Microarray processing pipeline steps



Navigate to the section:

Principle of microarray technology

Microarrays were one of the first scalable methods to measure gene expression across thousands of genes simultaneously. Each microarray platform targets a predefined set of transcripts, often up to 20,000 genes.

The workflow begins with RNA extracted from the sample and converted into fluorescently labeled cDNA. This cDNA is then hybridized to the microarray chip, which contains DNA probes fixed at specific positions. When the cDNA binds to its complementary probe, it generates a fluorescence signal whose intensity reflects the abundance of the corresponding transcript in the sample. The immediate output of this process is a matrix of fluorescence intensities. To make the data biologically meaningful, these raw signals must be processed into gene-level expression values[10]. At Epigene Labs, our pipelines perform this transformation, producing high-quality, harmonized expression matrices that are ready for downstream analysis (Fig. 2).

Platforms and probe annotation

Each microarray platform comes with a Chip Description File (CDF), which maps probes to specific genes. However, annotations evolve as biological knowledge improves. To maintain accuracy, we use updated mappings from trusted repositories like Brainarray, as well as annotation tools from Bioconductor[11] and BioMart[12]. These are updated more regularly than manufacturer-provided CDFs and ensure robust, reproducible gene mappings.[13], [14]

Because our analyses focus on the gene level, not individual probes, consistent probe-to-gene mapping is critical. This allows us to merge data across platforms and compare results with other omics types.

Manufacturers and protocols

The dominant commercial platforms are produced by Illumina, Affymetrix, and Agilent. Each uses slightly different designs:

- Single-channel arrays (e.g., Affymetrix GeneChip) measure one sample per array.
- Two-channel arrays compare two samples on the same array by measuring fluorescence ratios.

At Epigene Labs, we focus on single-channel microarrays since they are more abundant in public repositories, and the corresponding analytical tools are more mature and widely validated. The MAQC project additionally demonstrated that single-channel data yield more stable measurements compared to two-channel approaches.

Normalization and quality control

Processing microarray data requires multiple steps to remove noise and make datasets comparable.

Step 1: Background correction and normalization

For Affymetrix arrays, we use the *Oligo* R package[15]. The pipeline applies Robust Multi-array Average (RMA) normalization[16], which:

1. Removes background noise.
2. Applies a log transformation.
3. Performs quantile normalization to ensure comparability across arrays.

The output is a probe-level expression matrix.

Step 2: Probe-to-gene mapping

We then filter probes to keep only those linked to known HUGO gene symbols. Multiple probes mapping to the same gene are aggregated into a single gene-level value, typically using the maximum observed intensity per gene. This ensures that gene expression is not underestimated.

Step 3: Quality control checks

Our QC pipeline ensures:

- No missing, empty, or negative values remain.
- The number of genes detected aligns with expectations for the given platform.
- Expression values are within biologically plausible ranges.

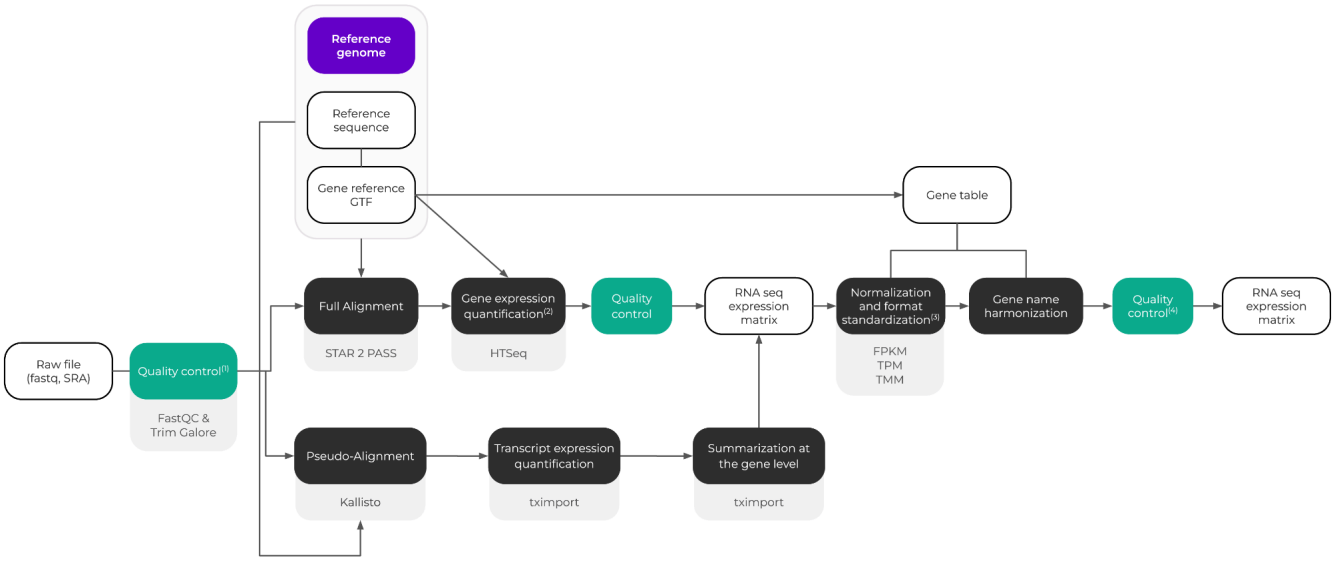
Cross-platform integration

Experiments that include multiple platforms require harmonization. We process each platform separately and then adjust for technical differences using pyCombat[17], our Python port of the well-established ComBat[18].

The result is a clean, gene-level expression matrix, suitable for integration into larger atlases.

RNA-Sequencing

Figure 3. RNA-seq processing pipeline steps



Navigate to the section:

1. Quality control
2. Gene expression quantification
3. Normalization and format standardization
4. Quality control

Overview

RNA sequencing (RNA-seq) has become the dominant technology for transcriptomics, thanks to its ability to capture the entire transcriptome with high sensitivity. Unlike microarrays, which are limited to predefined probes, RNA-seq can detect novel isoforms, rare transcripts, and subtle expression differences, making it indispensable in oncology research.

The workflow begins with RNA extraction, proceeds through sequencing and quality control, and ends with a gene-level expression matrix that can be compared across samples and studies. At Epigene Labs, we use standardized, scalable pipelines that ensure reproducibility while accounting for the complexity and variability of RNA-seq data (Fig. 3).

From sample to raw data

The RNA-seq workflow begins with RNA extraction from biological samples. Depending on the study design, enrichment strategies may be applied at this stage—for example, poly(A) selection to isolate mRNA, or ribosomal RNA depletion to capture both coding and non-coding RNAs. Once extracted, the RNA is fragmented into pieces of roughly 200–300 base pairs, converted into cDNA for stability, and then amplified by PCR to prepare it for sequencing[19]. High-throughput sequencers subsequently generate tens of millions of short reads per sample, each typically a few dozen bases in length. The resulting datasets are enormous, often reaching tens of gigabytes for a single experiment. These raw reads are stored either in FASTQ format (unprocessed) or in BAM format (aligned), and access is often restricted due to their sensitive nature.

Quality Control of raw reads

Raw reads must be evaluated before downstream analysis. Our pipeline includes:

- *FastQC*[20]: generates reports on base quality scores, GC content, sequence duplication, and adapter contamination.
- *Trim Galore*[21]: trims adapters and low-quality bases, producing cleaner input.
- *MultiQC*[22]: aggregates results from multiple tools into a single interactive report for easier inspection.

This stage ensures that sequencing errors or contamination do not propagate into later analyses.

Alignment and Quantification

Alignment

Reads must be mapped back to a reference genome to determine their origin. This step, known as alignment, is computationally intensive but critical for accuracy.

Full alignment. Tools like STAR[23] perform base-by-base alignment, handling insertions, deletions, and complex mappings. This is our preferred method due to its interpretability.

Pseudo-alignment. Faster approaches avoid full base matching, but at the cost of accuracy, especially in experimental datasets where indels or unannotated regions matter[24].

At Epigene Labs, we generally favor full alignment over pseudo-alignment, unless speed is an overriding requirement set by partners.

Gene expression quantification

Once the reads are aligned, a GTF annotation file is used to map them to exons and genes. From this mapping, the number of reads associated with each gene is counted, producing a gene expression vector for every sample. These vectors are then combined into an expression matrix that typically spans around 60,000 genes across multiple samples. To ensure consistency, Epigene Labs relies on the GDC reference pipeline[25], which uses STAR for alignment and HTSeq[26] for quantification. This choice not only standardizes processing across datasets but also aligns our analyses with TCGA, the oncology gold standard, ensuring reliable comparability.

Count matrices from public databases

Because raw RNA-seq processing is computationally expensive, we frequently integrate preprocessed count matrices from public repositories. These represent the direct output of quantification pipelines.

To validate them, we retrieve metadata detailing the reference genome and gene annotation used, then cross-check gene indices against known models. This step safeguards against annotation mismatches that could distort downstream analyses.

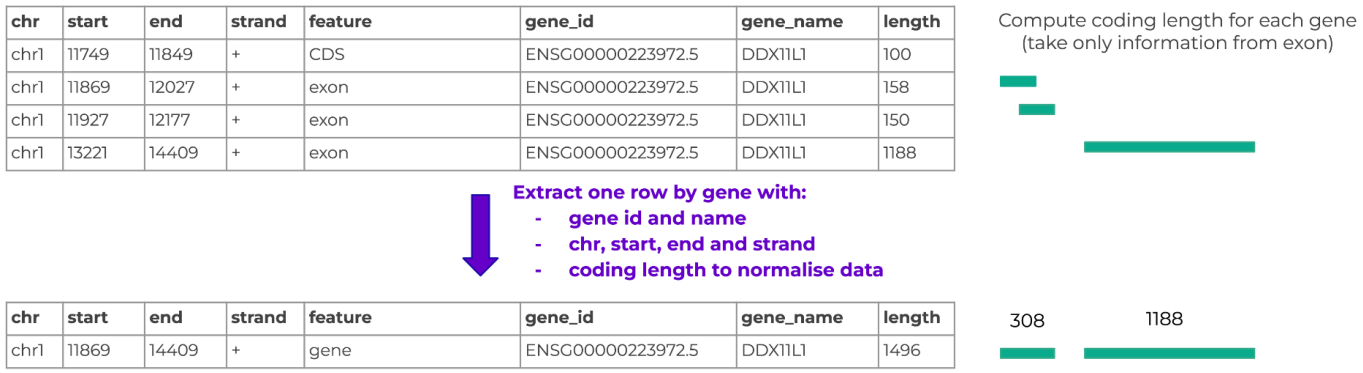
Normalization

Raw counts are not directly comparable across samples due to differences in sequencing depth, gene length, and library composition. Normalization corrects for these biases and is carefully selected depending on downstream analysis goals.[27]

Normalization	Adjust for	Recommandation
TMM[28] (Trimmed Mean of M-values)	• Sequencing depth • Library composition • Gene length	• Between-sample normalization • Comparison of samples across conditions
DESeq2's median-of-ratios[29]	• Sequencing depth • Library composition	• Between-sample normalization • Comparison of samples across conditions
TPM	• Sequencing depth • Gene length	• Within-sample normalization • Comparison of the expression level within a sample • Between-sample comparison under specific conditions
FPKM	• Sequencing depth • Gene length	• Within-sample normalization • Comparison of the expression level within a sample

In most cases, normalized values are log-transformed (with a small offset to avoid zeros) to reduce variance bias and stabilize statistical behavior.

Figure 4. Gene length computation method



Quality control at the expression matrix level

After quantification and normalization, we apply several layers of quality control to ensure that datasets are both technically and biologically consistent. First, we verify sequencing depth, as robust differential expression analysis requires at least five million mapped reads per human sample. Next, we perform biological consistency checks using in-house machine learning classifiers to validate metadata. For example, a sex classifier trained on Y-chromosome transcripts achieves 99% accuracy, reliably confirming that the declared sample sex matches the molecular profile. Similar models extend this approach to predict cancer type and tissue of origin, further protecting against mislabeled or misclassified data. Finally, we conduct technical checks to validate the structure of the expression matrices. Raw counts must be stored as integers and normalized values as floats, with no duplicates, missing entries, or invalid values present.

Importantly, we avoid imputation: when data are missing, the corresponding samples or genes are excluded rather than artificially filled, preventing the introduction of bias.

Integration pipeline performance

Our RNA-seq integration pipeline is optimized for both speed and scalability.

Starting from preprocessed count matrices, integration typically completes in ~0.5 hours for GEO datasets and ~2.5 hours for custom datasets.

Starting from raw reads, full alignment typically runs in 5 hours on a 64 GB RAM machine, while pseudo-alignment runs in about 2 hours, but with reduced interpretability.

Whenever possible, we favor preprocessed count matrices, as they offer the most stable and scalable approach to reanalyzing large volumes of public data.

Technology comparison

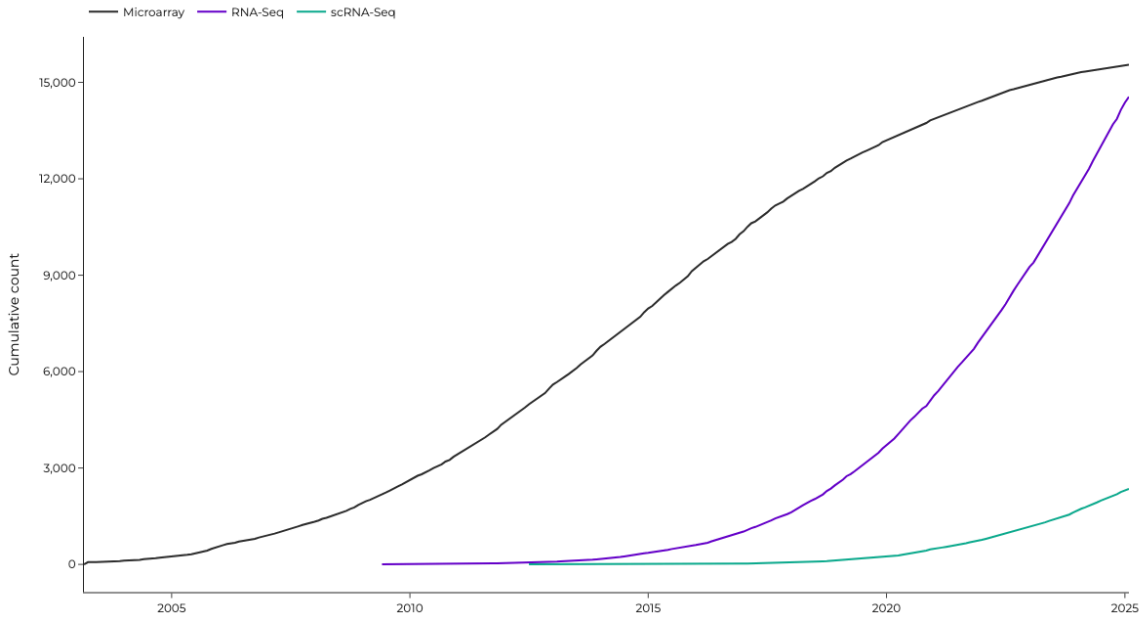
Complementary strengths

RNA-seq and microarrays are two distinct ways of measuring gene expression. Both have played transformative roles in transcriptomics, and each carries specific advantages and limitations.

While RNA-seq has now overtaken microarrays as the dominant method, the enormous legacy of microarray data ensures it will remain valuable for years to come (Fig. 5).[30], [31]

	Microarray	RNA-Seq
Scope	Limited to predefined probe sets, typically ~20,000 genes	Whole-transcriptome coverage; captures both known and novel transcripts, isoforms, and rare RNAs
Discovery potential	Restricted to targeted genes; cannot detect unknown sequences	Reference-free, enabling identification of new transcripts
Quantification	Relative: fluorescence intensities serve as proxies	Absolute: read counts are directly proportional to transcript abundance
Sensitivity & dynamic range	Lower sensitivity, higher background noise, dynamic range around ~1,000	High sensitivity, low background noise, dynamic range up to ~10,000
Cost & accessibility	Lower per-sample cost; large open-access archives (e.g., GEO)	More expensive per sample; raw data often under controlled access (FASTQ/BAM)
Statistical behavior	Approximates normal distribution, allowing simpler and lighter analyses	Negative binomial distribution, requiring more complex models
Computation	Downstream analyses require fewer resources	Alignment and processing are resource-intensive
Data formats	Highly standardized and homogeneous across repositories	Heterogeneous across public sources, complicating integration

Figure 5. Number of cancer series on GEO, per technology



Why both matter

Despite its disadvantages, microarray data remain invaluable:

- The volume of historical data is immense and irreplaceable, especially for tissues no longer accessible (Fig. 5). So far, Epigene Labs has identified 677,454 microarray samples with its capabilities.
- Well-annotated platforms (e.g., GPL570) are still relevant, as many targeted genes remain central to oncology research.
- Microarrays provide an orthogonal validation layer for RNA-seq findings, strengthening conclusions in integrative studies.

Conversely, RNA-seq continues to expand the frontier of discovery:

- Capturing the entire transcriptome enables isoform-level analyses and the study of non-coding RNA.
- Increasing affordability and standardization are making it the long-term foundation of transcriptomics. To date, Epigene Labs has already identified 436,081 RNAseq samples.

Statistical incompatibility

Although both technologies measure gene expression, their outputs are not directly comparable. The key issue lies in their underlying distributions: microarray intensities approximate normal distributions, while RNA-seq counts follow a negative binomial distribution.

Pooling data across the two technologies without appropriate statistical adjustments risks introducing bias and distorting biological signals.

Methods such as Voom[32] (limma package) or quantile normalization have been proposed to bridge the gap, but they lack strong mathematical foundations for general-purpose use. Moreover, RNA-seq datasets often contain more than 50,000 genes, while microarrays typically cover fewer than 20,000. This mismatch further complicates integration.

For these reasons, Epigene Labs deliberately avoids merging RNA-seq and microarray data into single datasets. Instead, we build separate atlases for each technology and combine results through meta-analysis, which preserves robustness without forcing incompatible data together.

Atlases creation

Why atlases are needed

Public repositories host a vast amount of transcriptomic data, but these datasets are rarely designed for direct comparison. Most originate from specific contexts, a clinical trial, a research consortium, or a regional study.

As a result, each dataset carries its own built-in biases:

- **Population biases**: e.g., a study of U.S. Army veterans may skew by age, sex, or occupation.
- **Geographic or environmental biases**: datasets from specific countries reflect local health, environment, or healthcare systems.
- **Technical biases:** differences in sequencing machines, lab protocols, or operators can shift expression values.
- **Sample size limitations**: many studies are small, reducing statistical power.
- **Narrow design**: most datasets are built to answer a specific research question, not to support broad exploratory analyses.

Together, these factors make it difficult to reproduce results across datasets or to detect subtle biological signals in rare subpopulations.

Atlases solve this fragmentation problem. An atlas is a curated collection of samples pooled from multiple datasets, designed to maximize statistical power, reduce dataset-specific bias, and capture a broader range of biological variability. Atlases are already common in multi-center clinical trials[33]; Epigene Labs extends this principle to large-scale transcriptomics.

Principles of Atlas Design

Building a biologically meaningful and statistically robust atlas requires careful design and rigorous harmonization. The first priority is ensuring that molecular data are compatible. Datasets must share a common gene space: in microarrays this means identifying the intersection of genes across different platforms, while in RNA-seq it requires consistency in the reference genome and gene annotation versions. Once compatibility is established, batch effect correction becomes essential.

Even when datasets share genes, expression values can still diverge due to technical factors such as sequencing depth, lab protocols, or instrument variability. Correction methods are designed to separate true biological variation – linked to covariates like age, tissue, or tumor type – from technical noise introduced by processing differences. To be valid, these corrections must preserve biologically relevant covariates while removing only the unwanted technical components.

A related challenge is avoiding confounding. When a biological covariate is perfectly aligned with dataset origin, it becomes impossible to distinguish technical effects from genuine biology.

For example, if one dataset contains only male samples and another only female samples, sex and dataset source become indistinguishable. The solution in such cases is to introduce pivot datasets that contain mixed covariates, helping to disentangle confounded variables.

Beyond molecular data, clinical metadata must also be harmonized. Consistent information on age, sex, tumor type, and outcomes is critical for interpretation and often serves as covariates in correction models. Poorly curated or inconsistent metadata can reintroduce bias, limit the number of usable samples, and compromise analytical potential.

Finally, diversity must be balanced carefully against statistical significance. Expanding the range of datasets reduces systematic biases, but too much heterogeneity can overwhelm the biological signal with technical or contextual noise. Selecting datasets that are broad enough to enhance robustness, yet focused enough to preserve interpretability, is key to building a reliable and biologically meaningful atlas.

Microarray Atlases

Microarray atlas construction poses a trade-off: adding more platforms increases the sample pool, but the common gene set shrinks as platform diversity grows. The challenge is to maximize both sample size and gene coverage. At Epigene Labs, we apply brute-force algorithms that test combinations of platforms and select the optimal balance—large enough sample sets without sacrificing too many genes.

RNA-Seq Atlases

Unlike microarray data, which is bound to fixed chip designs, RNA-seq offers far greater flexibility across platforms. Yet this flexibility comes with complexity. Sequencing protocols differ widely and these differences can strongly influence the resulting data. For instance, poly(A) enrichment selectively captures polyadenylated RNAs, such as mRNAs, while total RNA sequencing or ribosomal depletion strategies also retain non-polyadenylated transcripts, including small nucleolar and small nuclear RNAs. Long-read RNA-seq technologies, such as PacBio or Oxford Nanopore, add another dimension by capturing full-length transcripts and isoforms directly, eliminating the need for computational reconstruction. However, their structure and quantification metrics differ so markedly from short-read sequencing that merging the two within a single atlas is not feasible. For this reason, atlas construction works best when limited to datasets generated with consistent approaches—in our case, short-read RNA-seq with uniform RNA selection methods.

Beyond protocol choice, two additional factors exert a major influence on dataset comparability: the reference genome used for alignment and the gene annotation version used for quantification. While the genome provides the sequence framework, the annotation defines gene identity, structure, and location, including exon-intron boundaries and transcript isoforms. Because annotations are updated frequently, changing names, models, and boundaries, even datasets aligned to the same genome build can become incompatible if different annotation versions were applied.

To address this, Epigene Labs processes all raw RNA-seq data with a standardized pipeline using the same reference genome and annotation version, ensuring consistency across studies. When working with external datasets, we often encounter mismatched identifiers due to annotation updates. To resolve this, we employ custom harmonization tools that align gene names across versions, enabling safe integration.

For datasets that are otherwise compatible, batch effect correction remains essential to mitigate technical biases. These methods rely on meaningful clinical covariates to distinguish biological signal from noise. When such covariates are missing or fully confounded with dataset origin, we adopt a fallback strategy: reprocessing all data through a uniform alignment and quantification pipeline. While this does not eliminate technical variation entirely, it reduces variability introduced by computational pipelines. Nonetheless, we regard this as a last resort, favoring proper batch correction whenever possible.

Benefits and caveats

Atlases significantly expand the analytical possibilities of transcriptomic research. By pooling samples across studies, they make it possible to analyze rare populations that would be too small to study within individual datasets. Their consistency also supports reproducibility and enables reliable meta-analyses, while their flexibility allows researchers to define custom inclusion criteria tailored to specific projects.

At the same time, atlases come with important caveats. Poorly curated clinical metadata can undermine their reliability, excessive heterogeneity across datasets can dilute statistical power, and confounded covariates may prevent effective correction of technical biases.

When built with rigorous curation and careful dataset selection, however, atlases become a powerful tool for secondary transcriptomic analyses, enabling discoveries that no single dataset could ever achieve on its own.

Use cases

Differential expression analysis

Differential expression analysis is one of the core applications of transcriptomics[19]. The principle is straightforward: compare gene expression between two (or more) groups of samples to identify genes that are significantly over- or under-expressed.

At Epigene Labs, we rely on the Log Fold Change (logFC) to measure the magnitude of differences, and on False Discovery Rate (FDR) thresholds (default: logFC ≥ 1, FDR ≤ 0.05) to ensure statistical robustness.

These analyses yield lists of Differentially Expressed Genes (DEGs) that can point to tumor-specific markers (e.g., genes upregulated in cancer vs. healthy tissue), or predictors of therapy response (e.g., genes linked to sensitivity or resistance).

Because our pipelines can merge data across multiple datasets, we can extend these analyses to rare conditions or stratify by clinical variables that are often underpowered in single studies.

Tools used

- RNA-seq: DESeq2[34] or edgeR[35] (applied to raw counts).
- Microarray: limma[36] (applied to RMA-normalized log-transformed intensities).
- All with Benjamini–Hochberg correction for multiple testing.

Outcome analysis

At Epigene Labs, bulk transcriptomic data are a cornerstone for time-to-event analyses, where clinical outcomes such as overall survival, relapse, or recurrence are linked to molecular and clinical features. By exploring these relationships, we can uncover signals in high-dimensional data and identify key features that explain or predict patient outcomes.

Our work combines both univariate and multivariate approaches. Univariate analyses provide a straightforward means of validating biomarkers, discovering new candidates, or stratifying patients into groups based on single-gene associations.

Multivariate analyses, by contrast, capture the combined effects of multiple features (*e.g.* genes, clinical information), allowing us to compare biomarker specificity, identify novel gene signatures, and refine patient stratification strategies. Microarray data are normalized as described earlier in the pipeline, while RNA-seq data are normalized using methods such as TMM to ensure comparability across samples.

The creation of large, harmonized atlases further strengthens these analyses. By pooling data across studies, we assemble larger cohorts that support the application of more powerful algorithms, which often require substantial sample sizes to converge. This broader biological signal not only increases statistical power but also makes the resulting insights more robust and reproducible, ultimately enhancing the reliability of outcome-driven discoveries in oncology.

Functional analysis

Functional analysis, also known as gene-annotation enrichment analysis, provides a way to connect the results of high-throughput studies to biological meaning. Large gene lists – whether derived from multivariate outcome models or differential expression analyses – can be difficult to interpret on their own. Enrichment methods help transform these lists into insights about the pathways and processes they represent.

Two tools form the backbone of our functional analysis workflows: *Enrichr*[37] and *GSEA*[38]. *Enrichr* evaluates whether a set of genes, such as a biomarker signature or the output of a statistical test, is statistically overrepresented in curated libraries. These libraries may group genes by biological function, pathway membership, or other annotations, allowing investigators to identify processes enriched in their gene set of interest. In contrast, *GSEA* typically operates on larger, ranked lists of genes. Instead of focusing on a discrete set, it tests whether predefined gene libraries appear disproportionately at the top or bottom of a ranking, where genes are ordered by metrics such as log fold change, adjusted p-values, or false discovery rates.

Although the two methods differ in approach, they serve the same overarching purpose: to place gene-level findings into a broader biological framework. By linking molecular signatures back to established pathways and functions, enrichment analysis adds interpretability and context, turning raw outputs into mechanistic insights that can guide further research and clinical applications.

Why this matters for oncology

These use cases form the foundation of precision oncology applications:

- Target discovery: DEGs and enriched pathways can identify new therapeutic targets.
- Biomarker validation: outcome-linked signatures help select patients most likely to benefit from a treatment.
- Patient stratification: functional insights enable segmentation of cohorts into clinically meaningful subgroups.

By combining harmonized datasets, robust normalization, and multi-method analyses, Epigene Labs ensures that results are not only statistically sound but also biologically meaningful and clinically actionable.

Software ecosystem

The rise of transcriptomics has created a vast ecosystem of computational tools. Many of the most established packages were originally developed in R, such as *sva*[39] for surrogate variable analysis, *edgeR*[35] and *DESeq2*[34] for differential expression, *limma*[36] for microarray analysis.

These tools are widely trusted and have shaped current best practices. However, modern research environments require more than isolated packages. Transcriptomic analysis increasingly needs to connect with cloud storage systems, machine learning frameworks, and web-based or graphical interfaces for accessibility. This integration challenge calls for a more unified and flexible software environment.

Why Epigene Labs chooses Python

At Epigene Labs, we believe a single mainstream programming language helps reduce fragmentation and lowers the barrier for collaboration.

We have chosen Python as our primary ecosystem for three reasons:

1. **Interoperability:** Python integrates easily with machine learning libraries (e.g., scikit-learn, PyTorch, TensorFlow), cloud APIs, and web frameworks.
2. **Community trends**: The bioinformatics community is increasingly shifting toward Python, as illustrated by the success of scverse[40], the open-source ecosystem for single-cell omics.
3. **Collaboration and accessibility**: Python is widely adopted outside of bioinformatics, making it easier for cross-disciplinary teams – data scientists, software engineers, biologists – to work together.

By consolidating around Python, we aim to federate a more consistent, collaborative environment for bulk transcriptomics.

InMoose: our open-source contribution

To accelerate this shift, we contribute to the community through InMoose[41], our open-source Python package available on PyPI. InMoose implements Python ports of gold-standard R tools: ComBat and ComBat-Seq (batch effect correction), limma, edgeR and DESeq2 (differential gene expression). It also provides unique features to assess the quality of an atlas.

By bringing trusted methods into the Python ecosystem, InMoose makes high-quality tools accessible to any researcher working in Python. It also ensures reproducibility and scalability across Epigene Labs pipelines.

InMoose is just the starting point. Our long-term vision is to build a state-of-the-art Python ecosystem for bulk transcriptomics, open to the community and designed to grow collaboratively.

Commitment to open science

Our approach reflects a broader principle: science advances faster when tools are open, reproducible, and shared. By open-sourcing components of our pipeline, we encourage transparency and peer validation. By standardizing on Python, we reduce silos between disciplines. By contributing actively to the community, we help ensure that transcriptomic research is both scalable and accessible.

Envision and conclusion

Throughout this white paper, we have detailed Epigene Labs' approach to handling bulk transcriptomic data for microarrays and RNA sequencing. Our framework covers every stage of processing:

- Data acquisition from major repositories such as GEO, GDC, and ArrayExpress.
- Preprocessing pipelines tailored to each technology, ensuring noise reduction, normalization, and rigorous quality control.
- Integration strategies that account for platform-specific biases and harmonize data across studies.
- Atlas construction to overcome sample size limitations, reduce dataset-specific distortions, and unlock the potential of pooled analyses.
- Use cases spanning differential expression, outcome analysis, and functional enrichment—applications directly supporting biomarker discovery, patient stratification, and target identification.
- Software contributions, including our Python-based toolkit InMoose, designed to ensure reproducibility, scalability, and accessibility.

Together, these elements form a comprehensive, scalable framework that transforms fragmented data into coherent resources for oncology research.

Looking Beyond Bulk Data

Bulk transcriptomics remains an indispensable tool. Its affordability, scalability, and compatibility with vast archives of existing data make it uniquely suited for building disease atlases and conducting meta-analyses at population scale. However, bulk data inherently averages out cellular heterogeneity, masking the roles of individual cell types within complex tissues[42].

To address this limitation, Epigene Labs is expanding into single-cell RNA sequencing (scRNA-seq). Our internal pipelines follow the best practices[43] and already cover quality control, normalization, batch correction and cell type annotation.

scRNA-seq complements bulk approaches by revealing cellular diversity and the structure of the tumor microenvironment, offering insights critical for immuno-oncology and next-generation therapeutic strategies. A dedicated white paper on this methodology will be published soon.

Closing thoughts

The challenge of data fragmentation has long hindered the promise of transcriptomics. By harmonizing diverse datasets, curating metadata, and standardizing processing pipelines, Epigene Labs is addressing this challenge head-on. The result is a foundation that enables:

- More reliable discoveries, free from dataset-specific artifacts.
- Faster translation of insights, reducing months of research into days.
- Broader impact, making existing data archives as valuable as new experimental datasets.

As oncology research pushes deeper into complexity, from tumor heterogeneity to immune interactions, harmonized transcriptomic data will remain a cornerstone of discovery. At Epigene Labs, we are committed to building the tools and frameworks that turn data into actionable insight—accelerating progress toward precision oncology.

References

1. Buzdin, A., Moshkovskii, S., Li, X. & Belalov, I. Chapter 15 - Transcriptomics and quantitative proteomics: Competition or symbiosis? in *Handbook of Translational Transcriptomics* (ed. Buzdin, A.) 429–447 (Academic Press, 2025). doi:10.1016/B978-0-443-19110-7.00011-2.
2. Lowe, R., Shirley, N., Bleackley, M., Dolan, S. & Shafee, T. Transcriptomics technologies. *PLOS Comput. Biol.* **13**, e1005457 (2017).
3. GDC. https://portal.gdc.cancer.gov/.
4. Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**, e71–e71 (2016).
5. Home - dbGaP - NCBI. https://www.ncbi.nlm.nih.gov/gap/.
6. GEO Overview - GEO - NCBI. https://ncbi.nlm.nih.gov/geo/info/overview.html.
7. GEOparse — GEOparse 1.2.0 documentation. https://geoparse.readthedocs.io/en/latest/introduction.html.
8. BioStudies. ArrayExpress. https://www.ebi.ac.uk/biostudies/arrayexpress.
9. Kauffmann, A. *et al.* Importing ArrayExpress datasets into R/Bioconductor. *Bioinforma. Oxf. Engl.* **25**, 2092–2094 (2009).
10. Lowe, R., Shirley, N., Bleackley, M., Dolan, S. & Shafee, T. Transcriptomics technologies. *PLOS Comput. Biol.* **13**, e1005457 (2017).
11. Bioconductor. https://www.bioconductor.org/.
12. HGNC BioMart. https://biomart.genenames.org/.
13. Lu, X. & Zhang, X. The effect of GeneChip gene definitions on the microarray study of cancers. *BioEssays* **28**, 739–746 (2006).
14. Sandberg, R. & Larsson, O. Improved precision and accuracy for microarrays using updated probe set definitions. *BMC Bioinformatics* **8**, 48 (2007).
15. Carvalho, B. S. & Irizarry, R. A. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* **26**, 2363–2367 (2010).
16. Irizarry, R. A. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, 15e–115 (2003).
17. Behdenna, A., Haziza, J., Azencott, C.-A. & Nordor, A. *pyComBat, a Python Tool for Batch Effects Correction in High-Throughput Molecular Data Using Empirical Bayes Methods*. http://biorxiv.org/lookup/doi/10.1101/2020.03.17.995431 (2020) doi:10.1101/2020.03.17.995431.

18. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

19. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).

20. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

21. Krueger, F. *et al.* FelixKrueger/TrimGalore: v0.6.10 - add default decompression path. Zenodo https://doi.org/10.5281/ZENODO.7598955 (2023).

22. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinforma. Oxf. Engl.* **32**, 3047–3048 (2016).

23. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* **29**, 15–21 (2013).

24. Srivastava, A. *et al.* Alignment and mapping methodology influence transcript abundance estimation. *Genome Biol.* **21**, 239 (2020).

25. Bioinformatics Pipeline: mRNA Analysis - GDC Docs. https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/.

26. Putri, G. H., Anders, S., Pyl, P. T., Pimanda, J. E. & Zanini, F. Analysing high-throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics* **38**, 2943–2945 (2022).

27. Evans, C., Hardin, J. & Stoebel, D. M. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform.* **19**, 776–792 (2018).

28. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).

29. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

30. Mantione, K. J. *et al.* Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq. *Med. Sci. Monit. Basic Res.* **20**, 138–141 (2014).

31. Zhang, W. *et al.* Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biol.* **16**, 133 (2015).

32. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).

33. Tierney, J. F., Fisher, D. J., Burdett, S., Stewart, L. A. & Parmar, M. K. B. Comparison of aggregate and individual participant data approaches to meta-analysis of randomised trials: An observational study. *PLoS Med.* **17**, e1003019 (2020).

34. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for

RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

35. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

36. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).

37. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).

38. GSEA. https://www.gsea-msigdb.org/gsea/index.jsp.

39. gmail.com>, J. T. L. <jtleek at *et al.* sva: Surrogate Variable Analysis. Bioconductor version: Release (3.17) https://doi.org/10.18129/B9.bioc.sva (2023).

40. Virshup, I. *et al.* The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nat. Biotechnol.* **41**, 604–606 (2023).

41. Colange, M. *et al.* Bridging the gap between R and Python in bulk transcriptomic data analysis with InMoose. *Sci. Rep.* **15**, 18104 (2025).

42. Angerer, P. *et al.* Single cells make big data: New challenges and opportunities in transcriptomics. *Curr. Opin. Syst. Biol.* **4**, 85–91 (2017).

43. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).

About Epigene Labs

Epigene Labs creates intelligence-augmenting solutions to **accelerate precision oncology research and drug development**. By harnessing artificial intelligence and next-generation bioinformatics, its mCUBE platform integrates fragmented multi-omic datasets into comprehensive disease atlases with unprecedented depth and breadth. As a result, mCUBE has significantly impacted various R&D programs on immuno-oncology, often reducing months of research to mere days. Applications include target discovery, biomarker identification, and patient selection at leading cancer research institutes and biopharmaceutical companies. Based in Paris and Boston, Epigene Labs was initially incubated at the Harvard Innovation Labs and later launched in France with the backing of prominent European investors. The company has also secured backing from the European Commission through the prestigious multi-million-euro EIC Accelerator program.

Contact us