World Journal of Clinical Sciences
Print ISSN: 2960-0278
Online ISSN: 2960-0286
DOI: https://doi.org/10.61784/wjcs3020

TRANSCRIPTOMIC ANALYSIS IN INFLAMMATORY DISEASES: PUBLIC DATA RESOURCES, ANALYTICAL METHODS, AND A FRAMEWORK FOR REPRODUCIBLE META-ANALYSIS

BingYao Chen
Xi'an Jiaotong-Liverpool University, Suzhou 215000, Jiangsu, China.

Abstract: Transcriptomic data reveal how gene expression changes in inflammatory diseases. Thousands of these datasets are now public and free to reuse. This working paper reviews the main resources and methods for this field. It uses inflammatory bowel disease as a running example. The paper has four goals. First, it maps the general repositories and the disease specific databases that supply free data. Second, it describes the standard analysis pipeline for bulk and single cell data. Third, it summarizes shared molecular signatures that recur across inflammatory diseases. Fourth, it proposes a reproducible meta-analysis design that combines many small studies into one larger and more reliable analysis. The paper does not present new experimental results. It is a methods and resource synthesis. Its aim is to give a clear starting point for researchers who want to reuse public transcriptomic data in inflammatory disease research.
Keywords: Transcriptomics; Inflammatory diseases; Public data repositories; Reproducible meta-analysis

1 INTRODUCTION

Inflammation is a core process in many human diseases. Inflammatory bowel disease, rheumatoid arthritis, and systemic lupus erythematosus are common examples. These diseases involve abnormal immune activity. They are also heterogeneous. Patients with the same diagnosis often differ in their molecular profiles. Transcriptomics helps us study this variation. It measures the expression of thousands of genes at once. It can compare diseased and healthy states. It can also reveal the biological pathways that drive disease.

Two main technologies generate transcriptomic data. The first is bulk profiling. This includes microarrays and bulk RNA sequencing. Bulk methods measure the average expression across all cells in a sample. The second is single cell RNA sequencing. This method measures expression in each cell separately. It can resolve the immune cell types that bulk methods blend together. Both technologies have produced very large public archives.

Inflammatory diseases place a heavy burden on patients and health systems. They are often chronic. They tend to flare and remit over many years. Current drugs do not work for every patient. Many patients lose their response over time. This gap drives the search for new drug targets. Transcriptomics supports that search. It can rank candidate genes. It can also group patients by molecular subtype. These subtypes may guide treatment choices in the future.

The reuse of public data has clear value. Each new study is expensive and often small. A single laboratory rarely recruits enough patients for strong statistical power. Public archives solve part of this problem. They let researchers combine many studies. They also let researchers validate findings in independent cohorts. Reuse also supports reproducibility. A result that holds across many datasets is more trustworthy than a result from one cohort. Figure 1 shows the general flow from a patient sample to reusable public data.

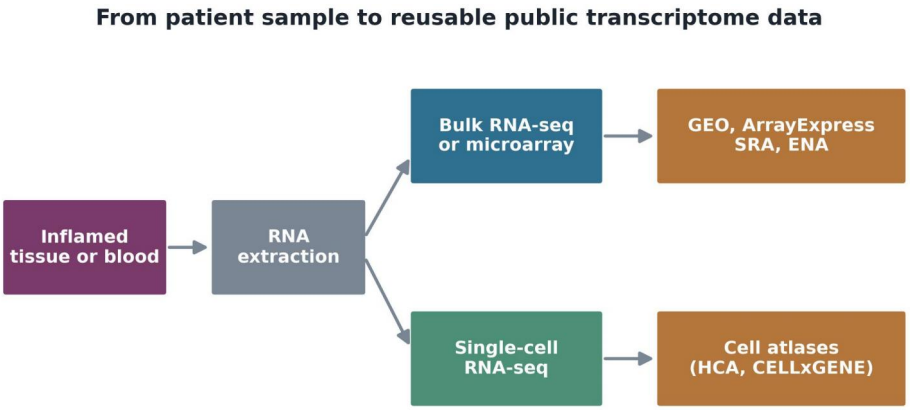

Figure 1 From Patient Sample to Reusable Public Transcriptome Data

This paper is organized as follows. Section 2 reviews the public data resources. Section 3 describes the analytical framework. Section 4 covers machine learning for classification. Section 5 summarizes recurring cross disease

signatures. Section 6 proposes a reproducible meta-analysis design. Section 7 lists curated datasets for reuse. Section 8 discusses limitations. Section 9 concludes.

2 PUBLIC DATA RESOURCES

2.1 General Repositories

Most transcriptomic data are stored in a few large repositories. The Gene Expression Omnibus, known as GEO, is the most widely used. It was established at the National Center for Biotechnology Information [1-2]. GEO now holds more than fifty eight thousand human study series. These contain about two million individual samples. The data span many diseases, tissues, and cell types. ArrayExpress, now part of BioStudies at the European Bioinformatics Institute, is the European counterpart. Many studies are deposited in both archives. The raw sequencing reads are stored separately. They sit in the Sequence Read Archive and the European Nucleotide Archive. A further resource is GTEx. It provides healthy tissue expression as a reference baseline. These archives are free to access. They form the backbone of public transcriptomics. Table 1 lists these general resources.

Table 1 Major General Purpose Public Transcriptomic Repositories

Resource	Content	Data type	Access
GEO	Functional genomics series and samples	Microarray, RNA-seq matrices	ncbi.nlm.nih.gov/geo
ArrayExpress / BioStudies	Bulk and single cell experiments	Processed and raw	ebi.ac.uk/biostudies
SRA	Raw sequencing reads	FASTQ, aligned reads	ncbi.nlm.nih.gov/sra
ENA	Raw sequencing reads (Europe)	FASTQ, assemblies	ebi.ac.uk/ena
GTEx	Healthy tissue reference expression	Bulk RNA-seq	gtexportal.org

2.2 Disease Specific Curated Databases

General repositories are powerful but unstructured. Their metadata are often inconsistent. This makes large searches slow and error prone. Curated databases solve part of this problem. They collect datasets for one disease area. They clean the data. They standardize the metadata. They also add ready made analysis tools.

Inflammatory bowel disease has the richest set of such databases. IBDTransDB is a manually curated resource. It contains thirty four transcriptomic datasets. These cover two thousand nine hundred thirty two samples and one hundred twenty two differential comparisons [3]. It supports queries across tissue, treatment, and cell type. IBD TaMMA is a second resource. It surveys public IBD RNA sequencing datasets. It applies batch correction so that studies can be compared [4]. For autoimmune disease more broadly, the IAAA platform integrates bulk data from nine hundred twenty nine samples across ten diseases. It also integrates single cell data from seven hundred eighty three thousand cells across six diseases [5]. Table 2 summarizes these and related resources.

Table 2 Selected Curated Databases for Inflammatory and Autoimmune Diseases

Database	Disease focus	Data type	Scale	Reference
IBDTransDB	Inflammatory bowel disease	Bulk, curated matrices	2,932 samples	[1]
IBD TaMMA	Inflammatory bowel disease	Bulk, batch corrected	Multi-study	[10]
IAAA	Ten autoimmune diseases	Bulk and single cell	929 + 783k cells	[18]
ImmuNexUT	Immune mediated diseases	Sorted immune cell RNA-seq	28 cell subsets	[12]
DICE	Human immune cell types	Sorted cell expression, eQTL	13 cell types	[17]

Two further resources focus on immune cells. ImmuNexUT profiles sorted immune cell subsets from patients with immune mediated diseases. It links expression to genetic variation [6]. The DICE database profiles sorted human immune cell types from healthy donors. It also reports expression quantitative trait loci [7]. These resources help researchers interpret signals at the level of specific cell types.

2.3 Single Cell Atlases

Single cell data have their own archives. The Human Cell Atlas Data Portal is one major source. CELLxGENE is another. Both let users browse datasets by disease and tissue. Both also let users download standard data files. These atlases are growing fast. They now include large inflammatory disease cohorts. The lupus blood study is one example. It profiled more than one million cells from cases and controls [8]. Such datasets are too large for a single laboratory to generate alone. Public release makes them available to the whole field. A researcher can reuse them to test a hypothesis at no extra sampling cost.

Figure 2 compares the scale of several of these resources. The general archive GEO dwarfs the curated databases in size. The curated databases are smaller. They are also cleaner and easier to use. A practical strategy uses both. Researchers can mine GEO for breadth. They can use curated databases for speed and quality control.

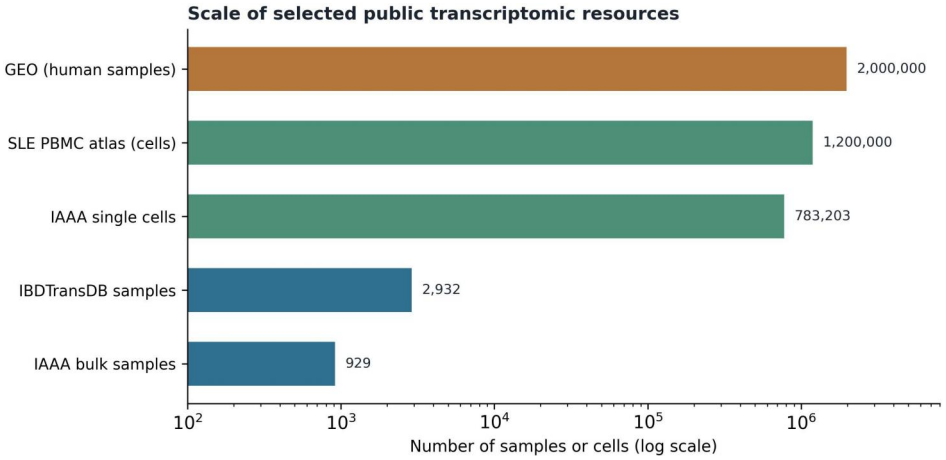

Figure 2 Scale of Selected Public Transcriptomic Resources

3 ANALYTICAL FRAMEWORK

3.1 The Bulk Analysis Pipeline

Bulk analysis follows a standard sequence. Figure 3 shows the main steps. The first step is to obtain a count matrix. This matrix has genes in rows and samples in columns. The second step is quality control. Low count genes are removed. They carry little signal and add noise. Outlier samples are also flagged. A sample may be an outlier for technical reasons. Principal component analysis helps spot such samples. The third step is normalization. This corrects for differences in sequencing depth between samples. Without this step, deeper samples would look more active by accident. Common methods include the median of ratios method and the trimmed mean of M values method. The fourth step is differential expression analysis. Three tools dominate this step. DESeq2 models counts with a negative binomial distribution [9]. edgeR uses a related model and is well suited to small sample sizes [10]. limma was first built for microarrays and now also handles RNA sequencing through its voom method [11].

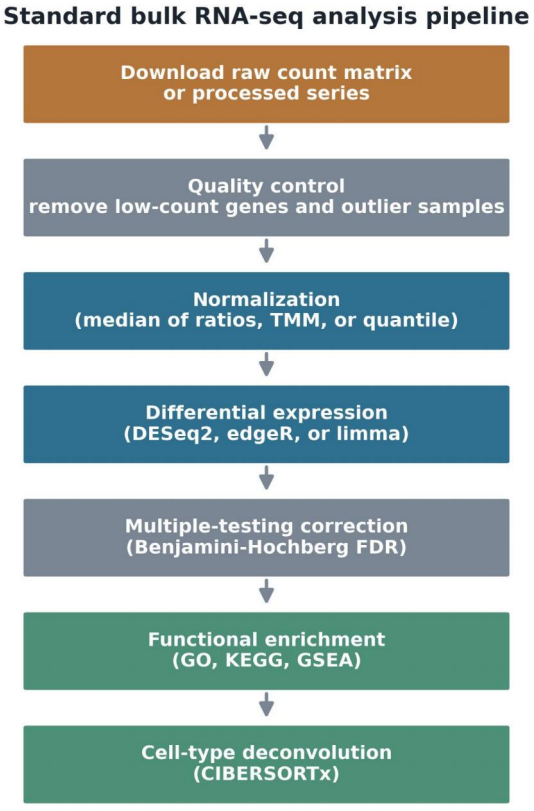

Figure 3 Standard Bulk RNA-seq Analysis Pipeline

The output of this step is a list of differentially expressed genes. Each gene has a fold change and a p value. Because many genes are tested at once, the p values must be corrected. The Benjamini and Hochberg method controls the false discovery rate. A common threshold uses an adjusted p value below 0.05. A fold change cutoff is often added. Table 3 lists thresholds and choices that recur in the literature.

Table 3 Common Analysis Choices and Thresholds in Bulk Transcriptomic Studies

Step	Typical choice	Purpose
Gene filtering	Remove genes with very low counts	Reduce noise and testing burden
Normalization	Median of ratios or TMM	Correct sequencing depth
Significance	Adjusted p value below 0.05	Control false discovery rate
Effect size	Absolute log2 fold change above 1	Focus on larger changes
Batch handling	ComBat or model covariate	Remove study specific bias

After the gene list is ready, two further analyses are common. The first is functional enrichment. This step asks which biological pathways are over represented. Gene Set Enrichment Analysis is a widely used method for this purpose [12]. The second is cell type deconvolution. Bulk samples are mixtures of many cell types. Deconvolution estimates the fraction of each cell type from the bulk profile. CIBERSORTx is a popular tool for this task [13]. Deconvolution is useful in inflammatory diseases. A change in bulk expression may reflect a shift in cell composition rather than a change within cells.

3.2 Single Cell Analysis

Single cell data require a different pipeline. The data are sparse and noisy. Each cell yields only a partial view of its transcriptome. The analysis usually starts with quality control on each cell. It then normalizes the data and reduces dimensions. Cells are then clustered into groups. Each group is annotated with a cell type. Two software ecosystems dominate this work. Seurat is an R package for multimodal single cell analysis [14]. SCANPY is a Python package built for large datasets [15].

Figure 4 shows why single cell methods matter for inflammation. Bulk methods give one averaged profile per sample. Single cell methods separate the immune cell types. This separation is important because inflammatory disease often changes the balance of cell types. A study of lupus blood made this clear. It profiled more than one million peripheral blood cells from cases and controls. It found a strong interferon signature in monocytes. It also found shifts in T cell populations [8].

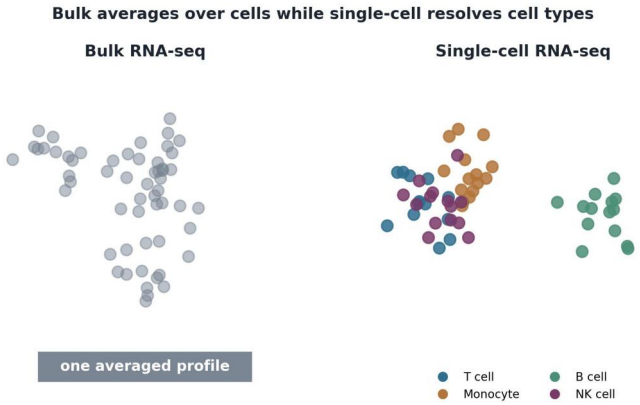

Figure 4 Bulk Averages over Cells While Single-Cell Resolves Cell Types

Table 4 Common Software for Transcriptomic Analysis

Tool	Purpose	Setting	Reference
DESeq2	Differential expression	Bulk RNA-seq	[9]
edgeR	Differential expression	Bulk RNA-seq	[15]
limma	Differential expression	Microarray, RNA-seq	[14]
ComBat	Batch effect correction	Multi-study	[8]
GSEA	Pathway enrichment	Gene lists	[19]
CIBERSORTx	Cell type deconvolution	Bulk mixtures	[11]
Seurat	Single cell analysis	scRNA-seq (R)	[7]
SCANPY	Single cell analysis	scRNA-seq (Python)	[21]

A good analysis also records its design choices. The choice of reference group matters. The choice of covariates matters too. Age, demographic characteristics, and treatment can all affect expression. These factors should be included in the model when they are known. Clear records make the analysis easier to reproduce. They also make it easier for reviewers to judge. Table 4 lists the main software discussed above.

4 MACHINE LEARNING FOR PATIENT CLASSIFICATION AND BIOMARKER DISCOVERY

4.1 Why Machine Learning

Differential expression finds genes that change on average. It does not build a predictor. Many clinical goals need a predictor. We may want to separate patients from controls. We may want to predict response to a drug. We may want to assign a molecular subtype. Machine learning is built for these goals. It learns a rule from labeled data. It then applies the rule to new samples. Recent IBD studies use this approach. One blood study built diagnostic models with regularized regression and support vector machines [16]. A second study used machine learning to define molecular subtypes across large RNA sequencing cohorts [17].

4.2 Feature Selection by Regularized Regression

Transcriptomic data have a hard shape. They have many genes and few samples. A typical study has tens of thousands of genes. It may have only tens or hundreds of samples. This setting is called high dimensional. Ordinary regression fails here. It overfits the training data. Regularization solves this problem. It adds a penalty that shrinks the coefficients. The most common choice is the LASSO [18]. It uses an L1 penalty. The penalty forces many coefficients to exactly zero. This performs feature selection at the same time as fitting.
For a binary outcome, the penalized logistic regression objective is

$$\hat{\boldsymbol{\beta}}=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{-\frac{1}{n}\sum_{i=1}^{n}\left[y_i\log p_i+(1-y_i)\log(1-p_i)\right]+\lambda\left(\alpha\,/\!/\,\boldsymbol{\beta}\,/\!/_1+\frac{1-\alpha}{2}\,/\!/\,\boldsymbol{\beta}\,/\!/_2^2\right)\right\}, \quad (1)$$

where $p_i = 1/\left(1+e^{-\mathbf{x}_i^{\top}\boldsymbol{\beta}}\right)$ is the predicted probability for sample i. The term y_i is the class label. The vector $\mathbf{x}_i$ holds the gene expression values. The parameter λ controls the strength of the penalty. The parameter α mixes the two penalty types. Setting $\alpha = 1$ gives the LASSO. Setting $\alpha = 0$ gives ridge regression. A value between the two gives the elastic net. The elastic net is useful when genes are correlated. It tends to keep groups of related genes together. Figure 5 shows how the coefficients shrink as the penalty grows.

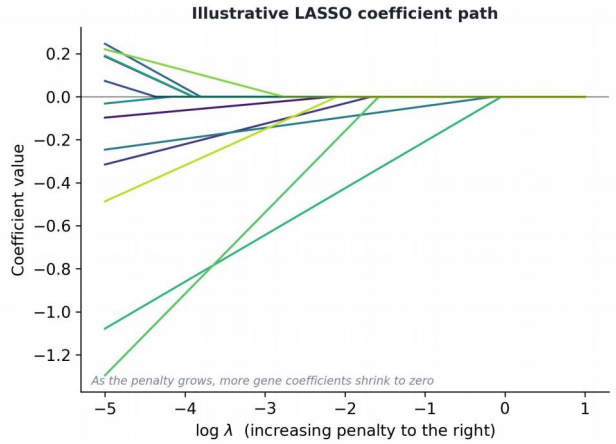

Figure 5 Illustrative LASSO Coefficient Path

4.3 Classifiers

Table 5 Common Machine Learning Models for Transcriptomic Classification

Model	Type	Strength	Weakness
LASSO logistic	Linear, sparse	Interpretable; selects genes	Misses nonlinear effects
Elastic net	Linear, sparse	Handles correlated genes	Two parameters to tune
SVM (RBF kernel)	Margin based	Models nonlinear patterns	Harder to interpret
Random forest	Tree ensemble	Robust; ranks features	Larger models; less sparse
Gradient boosting	Tree ensemble	High accuracy	Sensitive to tuning
Autoencoder	Neural network	Learns compact features	Needs many samples

Several classifiers are common in this field. Regularized logistic regression is the simplest. It is easy to interpret. Its coefficients show the direction of each gene effect. The support vector machine is a second option. It finds the boundary with the widest margin between classes. With a kernel it can model nonlinear patterns. Tree based methods are a third option. The random forest builds many decision trees and averages them [19]. Gradient boosting builds trees in sequence and corrects past errors. XGBoost is a fast and widely used boosting library [20]. Tree methods handle interactions well. They are also robust to feature scaling. Table 5 compares these models.

4.4 Evaluation without Leakage

Evaluation must be honest. The biggest risk is information leakage. Leakage happens when test data influence training. Feature selection is a common source. If genes are selected using all samples, the test set is no longer independent. The fix is nested cross-validation. The outer loop estimates performance. The inner loop selects features and tunes parameters. Feature selection stays inside the inner loop. Figure 6 shows this design. The main performance metric is the area under the ROC curve. This metric is often called the AUC. It measures how well the model ranks cases above controls. A value of 0.5 means random guessing. A value near 1.0 means strong separation. Figure 7 compares several models on illustrative curves. Sensitivity and specificity are also reported. They describe the trade off at a chosen threshold.

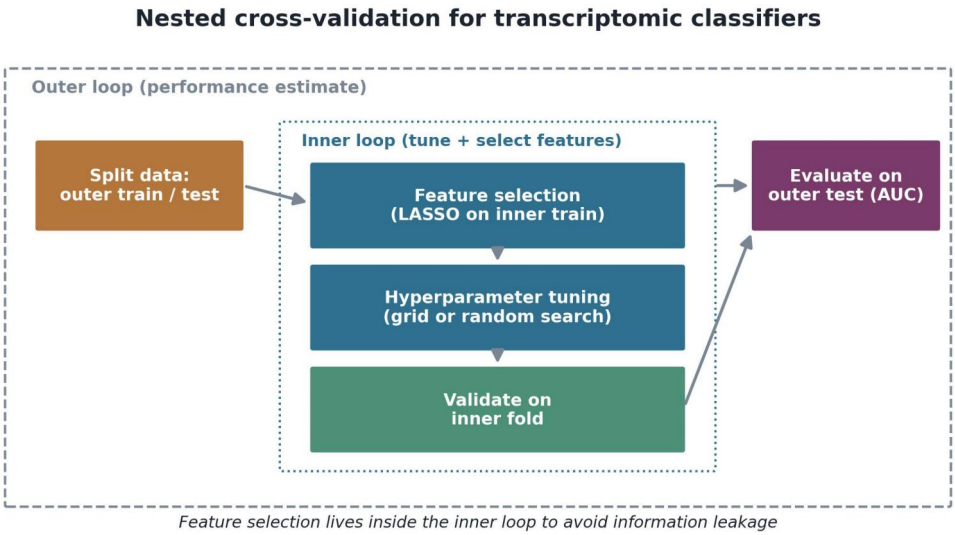

Figure 6 Nested Cross-Validation for Transcriptomic Classifiers

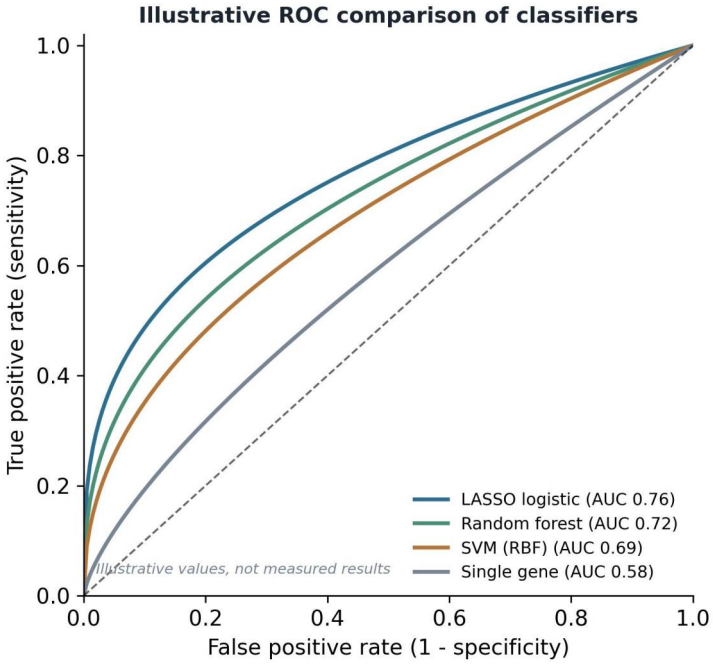

Figure 7 Illustrative ROC Comparison of Classifiers

4.5 Deep Learning and Its Limits

Deep learning is also used in this field. Autoencoders can learn compact representations of expression data. Variational autoencoders can model the structure of single cell data. These methods are powerful. They also need many samples. Most single study cohorts are too small for deep models. Overfitting is a serious risk. For many tasks, regularized linear models remain competitive. They are also easier to explain to clinicians. A practical rule is to start simple. A more complex model is justified only when it clearly beats the simple one on held out data.

5 RECURRING SIGNATURES ACROSS INFLAMMATORY DISEASES

Transcriptomic studies reveal patterns that recur across diseases. The interferon signature is the clearest example. It appears in lupus blood cells [8]. It also appears in other immune mediated diseases profiled by sorted cell resources [6]. This shared signature suggests common pathways. It also suggests that some drugs may help across several diseases.
Cell composition changes are a second recurring theme. Inflammatory bowel disease blood shows altered immune cell fractions. Reported changes include more monocytes and regulatory T cells. They also include fewer memory B cells and activated natural killer cells [16]. Such patterns are detected both by deconvolution of bulk data and by single cell profiling. Table 6 summarizes representative findings. The table is illustrative. It is not a complete survey.

Table 6 Representative Transcriptomic Findings in Inflammatory Diseases

Disease	Sample	Reported signature	Reference
Lupus	Blood, single cell	Interferon signature in monocytes; T cell shifts	[8]
Inflammatory bowel disease	Whole blood	Altered immune cell fractions; candidate biomarkers	[16]
Inflammatory bowel disease	Gut tissue, multi-study	Reproducible disease signatures after batch correction	[4]
Immune mediated diseases	Sorted immune cells	Cell type specific regulation linked to genetic risk	[6]

These findings carry a clear message. Disease signals are often cell type specific. A bulk analysis may miss them. A cell aware analysis can recover them. This is why deconvolution and single cell methods are now central to the field.
Shared signatures also carry a practical message. They suggest opportunities for drug repurposing. A drug that blocks the interferon pathway may help in more than one disease. A target found in lupus may also matter in another autoimmune condition. Transcriptomic atlases support this reasoning. They let researchers compare diseases on a common molecular scale. They also help match a drug mechanism to the right patient group. This logic is now a major theme in translational research.

6 A FRAMEWORK FOR REPRODUCIBLE META-ANALYSIS

Single studies are often too small to be reliable. Meta-analysis offers a solution. It pools many studies into one analysis. This raises statistical power. It also tests whether a finding holds across cohorts. Figure 8 shows a proposed design. The design is built for reproducibility.

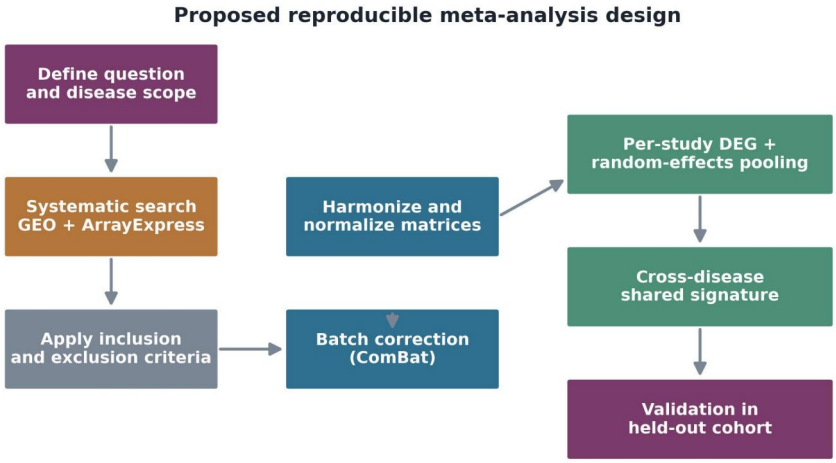

Figure 8 Proposed Reproducible Meta-Analysis Design

The design has six stages. The first stage defines the question and the disease scope. A narrow scope is easier to defend. The second stage runs a systematic search of GEO and ArrayExpress. The search uses disease terms, the organism filter, and the assay type. The third stage applies clear inclusion and exclusion criteria. These criteria should be fixed before the search results are read. Table 7 gives an example set of criteria.

Table 7 Example Inclusion and Exclusion Criteria for a Meta-Analysis

Criterion	Include	Exclude
Organism	Human samples	Animal models
Design	Disease and control groups present	Case only studies
Assay	Bulk RNA-seq or microarray	Targeted panels
Sample size	At least ten per group	Very small pilots
Data access	Raw or normalized matrix available	Summary statistics only

The fourth stage harmonizes and normalizes the matrices. Gene identifiers are mapped to a common scheme. Expression values are placed on a common scale. The fifth stage removes batch effects. Each study adds its own technical bias. This bias can dominate the biological signal. The ComBat method corrects for it with an empirical Bayes

model [21]. Figure 9 shows the idea as a schematic. Before correction, samples cluster by study. After correction, samples cluster by phenotype.

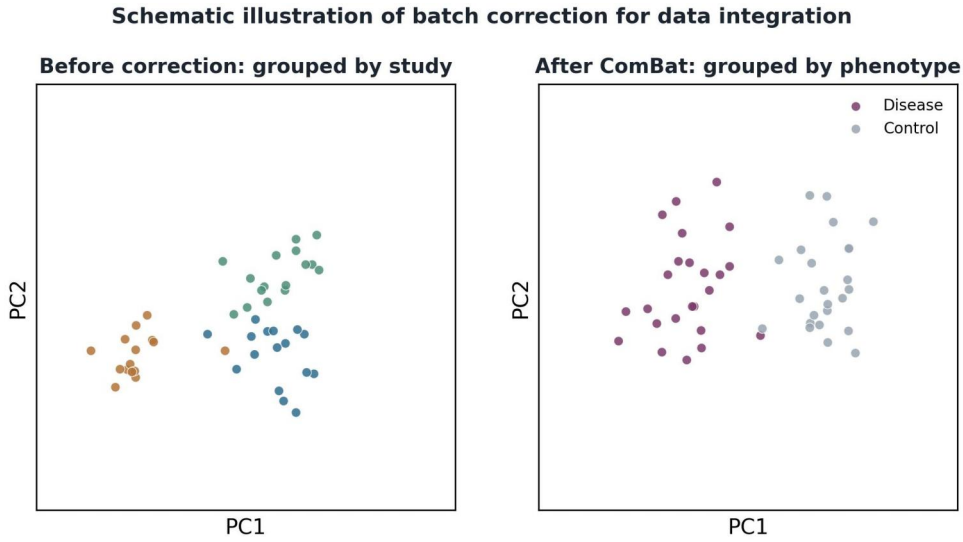

Figure 9 Schematic Illustration of Batch Correction for Data Integration

The sixth stage performs the pooled analysis. There are two common approaches. One approach runs differential expression in each study and then pools the effect sizes with a random effects model. This approach respects the boundaries between studies. It also models the variation between them. The other approach merges the corrected matrices and analyzes them together. This approach gains power from the larger sample. It depends more heavily on good batch correction. Both approaches end with validation. A held out cohort tests whether the signature generalizes. This final step protects against overfitting. A signature that holds in new data is far more useful than one that fits only the training set.

7 CURATED DATASETS FOR REUSE

A working paper needs concrete data. Table 8 lists verified public IBD datasets. Each entry gives a GEO accession number. The table mixes RNA sequencing and microarray sets. It also mixes tissue and blood samples. This mix supports several study designs. A large tissue cohort can serve as the primary set. A blood cohort can support a non-invasive aim. Smaller sets can serve as independent validation. The full list, with direct download links and code snippets, is provided as a companion file.

Table 8 Verified Public Transcriptomic Datasets for Inflammatory Bowel Disease

Accession	Disease	Tissue	Assay	Role
GSE193677	UC, CD, control	Intestine	RNA-seq	Large primary cohort (about 2,490 patients)
GSE186507	IBD	Blood	RNA-seq	Blood biomarker cohort (about 1,030 patients)
GSE57945	CD, non-IBD	Ileum	RNA-seq	RISK pediatric reference set
GSE117993	CD, non-IBD	Rectum	RNA-seq	RISK rectal companion set
GSE16879	CD, UC	Mucosa	Microarray	Anti-TNF response (infliximab)
GSE75214	UC, CD, control	Mucosa	Microarray	Active vs inactive validation set
GSE36807	UC, CD, control	Colon	Microarray	Small clean case-control set
GSE9686	UC, CD, control	Colon	Microarray	Independent microarray set
GSE137344	IBD	Intestine	RNA-seq	Validation cohort
GSE235236	IBD	Intestine	RNA-seq	Validation cohort

Note: Sample counts are approximate and should be confirmed on the GEO record.

The choice of sets should match the question. A diagnostic model needs cases and controls. A response model needs treated patients with known outcomes. GSE16879 fits the response question well. It contains samples before and after infliximab. It also labels responders and non-responders. A subtype study needs a large and diverse cohort. GSE193677 fits that aim. The general rule is simple. Use one large set to train. Use one or more independent sets to validate. Keep the validation sets untouched until the end.

8 LIMITATIONS AND PITFALLS

Public data reuse has real limits. The first limit is metadata quality. Many records have incomplete clinical labels. Disease subtype, treatment, and tissue are often unclear. This problem motivates the curated databases described above. The second limit is batch effects. They are pervasive in multi-study work. They must be modeled with care. Over correction can remove real biology. Under correction can leave false signals. The third limit is platform difference.

Microarray and RNA sequencing data are not directly comparable. They need careful harmonization before they are combined.

A fourth pitfall is the confusion of composition and regulation. A change in bulk expression has two possible causes. The cells may change their expression. The mix of cells may change instead. Deconvolution and single cell methods help separate these causes. A fifth pitfall is multiple testing. Transcriptomic studies test thousands of genes. Proper false discovery control is essential. A final pitfall is weak validation. A signature found in training data may not generalize. Independent validation is the standard safeguard.

Data licensing and consent also deserve attention. Most archives are open access. Some datasets carry use restrictions. Researchers should check the terms before reuse. Patient privacy is a further concern. Raw sequencing reads can carry identifying information. Many human datasets therefore sit behind controlled access. A clear data management plan helps a project stay compliant. It also makes later publication smoother.

9 CONCLUSION

Public transcriptomic data are abundant and free. They are a strong foundation for inflammatory disease research. General repositories supply breadth. Curated databases supply quality and speed. Standard pipelines turn raw counts into gene lists, pathways, and cell type estimates. Single cell methods add resolution that bulk methods lack. Recurring signals, such as the interferon signature, point to shared biology across diseases. A careful meta-analysis can combine many small studies into one reliable result. The design proposed here aims to make that process reproducible.

The field is moving quickly. Single cell datasets are growing each year. Spatial methods are now adding tissue location to the picture. These advances will deepen our view of inflammation. They will also raise the bar for analysis. Strong methods and careful validation will matter more than ever. The principles in this paper still apply. Use reliable data. Correct for batch effects. Validate in new cohorts. These habits protect the quality of any study.

The next step for an empirical paper is to apply this design to a chosen disease. Inflammatory bowel disease is a strong candidate. It has the deepest pool of public data and the most mature curated resources. A focused meta-analysis of this disease could test a clear hypothesis. It could also deliver a validated molecular signature. Such a result would be a solid contribution to the field.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Barrett T, Wilhite S E, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets, update. Nucleic Acids Research, 2013, 41(D1): D991-D995.

[2] Edgar R, Domon M, Lash A E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research, 2002, 30(1): 207-210.

[3] Avram V, Yadav S, Sahasrabudhe P, et al. IBDTransDB: a manually curated transcriptomic database for inflammatory bowel disease. Database (Oxford), 2024, 2024: baae026.

[4] Massimino L, Lamparelli L A, Houshyar Y, et al. The Inflammatory Bowel Disease Transcriptome and Metatranscriptome Meta-Analysis (IBD TaMMA) framework. Nature Computational Science, 2021, 1: 511-515.

[5] Shen Z, Fang M, Sun W, et al. A transcriptome atlas and interactive analysis platform for autoimmune disease. Database (Oxford), 2022, 2022: baac050.

[6] Ota M, Nagafuchi Y, Hatano H, et al. Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. Cell, 2021, 184(11): 3006-3021.

[7] Schmiedel B J, Singh D, Madrigal A, et al. Impact of genetic polymorphisms on human immune cell gene expression. Cell, 2018, 175(6): 1701-1715.

[8] Perez R K, Gordon M G, Subramaniam M, et al. Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. Science, 2022, 376(6589): eabf1970.

[9] Love M I, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 2014, 15(12): 550.

[10] Robinson M D, McCarthy D J, Smyth G K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 2010, 26(1): 139-140.

[11] Ritchie M E, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 2015, 43(7): e47.

[12] Subramanian A, Tamayo P, Mootha V K, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 2005, 102(43): 15545-15550.

[13] Newman A M, Steen C B, Liu C L, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature Biotechnology, 2019, 37(7): 773-782.

[14] Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. Cell, 2021, 184(13): 3573-3587.

[15] Wolf F A, Angerer P, Theis F J. SCANPY: large-scale single-cell gene expression data analysis. Genome Biology, 2018, 19: 15.
[16] Derakhshan Nazari M H, Ghorbaninejad M, Shahrokh S, et al. Biomarker discovery for non-invasive diagnosis of inflammatory bowel disease using blood transcriptomics. Frontiers in Immunology, 2025, 16: 1570374.
[17] Saini N, Acharjee A. Identifying inflammatory bowel disease subtypes: a comprehensive exploration of transcriptomic data and machine learning-based approaches. Therapeutic Advances in Gastroenterology, 2025, 18: 17562848251362391.
[18] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 1996, 58(1): 267-288.
[19] Breiman L. Random forests. Machine Learning, 2001, 45(1): 5-32.
[20] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 785-794.
[21] Johnson W E, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 2007, 8(1): 118-127.