# 1. Introduction

To gain molecular insight into cellular function, researchers analyze various biological layers, including genomic DNA, chromatin structure, messenger RNA (mRNA), non-coding RNAs, proteins, post-translational modifications, and metabolites. Since the discovery of RNA’s role as the intermediary between the genome and the proteome, RNA sequencing (RNA-seq) has become a cornerstone of transcriptomic research. By quantifying mRNA levels across the genome, RNA-seq enables comprehensive profiling of gene activity, offering critical insights into cellular responses, disease mechanisms, and therapeutic effects (Byron et al., 2016).

In parallel, metabolomics, the large-scale study of small molecules, or metabolites, captures the downstream products of gene expression and cellular processes. While RNA-seq captures the "blueprint" of cellular activity, metabolomics reflects its actual outcome, providing a phenotype-level readout. Integrating these two omics layers enables more robust biological interpretation, linking changes in gene regulation to shifts in biochemical pathways and metabolic function (Hasin et al., 2017).

## Research Context: Radiation-Induced Salivary Gland Injury

This paper centers its analytical walkthrough on a publicly available multi-omics dataset (GSE155902), which profiles radiation-induced dysfunction in mouse salivary glands. Ionizing radiation, commonly used in head and neck cancer therapy, often causes irreversible salivary gland damage, leading to xerostomia (dry mouth) and impaired quality of life. The dataset includes RNA-seq and metabolomics data from control and irradiated mouse parotid glands, offering a biologically meaningful context in which to explore differential gene and metabolite expression, pathway perturbations, and multi-omics integration.

By situating the workflow within this disease-relevant model, the paper not only demonstrates computational techniques but also highlights how multi-omics analyses can uncover molecular mechanisms underlying tissue injury and repair.

## Scope of the Paper

This paper presents a structured, technically sound approach to analyzing RNA-seq data and integrating it with metabolomic profiles. The project simulates a teaching environment in which the reader is guided through the complete analysis pipeline — from raw gene expression counts and metabolite abundance matrices to pathway-based feature selection and enrichment. Both R and Python implementations are discussed, allowing comparison of analytic workflows and accessibility for a broader range of bioinformatics learners.

A major emphasis is placed on addressing common challenges in omics data analysis, particularly the “big P, small N” problem — where the number of measured variables (genes, metabolites) greatly exceeds the number of samples. This issue introduces risks of overfitting, instability in feature selection, and inflated false discovery rates. Throughout the analysis, strategies such as dimension reduction, pathway-level aggregation, and regularized modeling are applied to mitigate these risks and promote reproducible findings.

Ultimately, this white paper serves not only as a reference for multi-omics integration techniques, but also as a step-by-step educational guide for students and researchers entering the field of transcriptomics and bioinformatics.

# 2. Data Types & Study Design

https://arxiv.org/abs/1302.3685

https://academic.oup.com/bioinformatics/article/39/9/btad547/7260507?login=false

## R Packages

### **Core Packages – Metabolomics**

| **Package** | **Purpose** |
| --- | --- |
| **MetaboAnalystR** | Core framework: preprocess metabolomics data, run stats, enrichment, and pathway analysis. Works directly with your .csv abundance files. |
| **pathview** | Optional. Visualize metabolite/gene expression on KEGG pathways (overlay data). |

### **Core Packages – RNA-Seq**

| **Package** | **Purpose** |
| --- | --- |
| **DESeq2** | Normalize RNA-seq count data and identify differentially expressed genes (DEGs). |
| **biomaRt** | Retrieve gene annotations (e.g., Ensembl → gene symbols, KEGG IDs). |
| **org.Mm.eg.db** | Mouse-specific gene annotation database (for mapping IDs). |
| **EnhancedVolcano** | Visualize DE results as volcano plots. |
| **pheatmap** | Create expression heatmaps of top DEGs. |

### **For Batch Effect Correction**

| **Package** | **Purpose** |
| --- | --- |
| **sva** | Implements **ComBat**, a popular batch correction method. |
| **EigenMS** | Uses singular value decomposition for batch effect removal (commonly used in proteomics). |
| **ruv**, **limma**, **waveICA** | Additional options for removing unwanted variation. Use only if strong batch effects are suspected. |

Will be following the piple from:  
https://www.mdpi.com/2218-1989/10/5/186?utm\_source

## Working on a mouse line with heterogeneity would provide consisten results

You wouldn't catch specific pathway differences humans are very heterogeneous need much larger numbers to.capture patterns

Take away as many variables to see whether analysis can do this how close can I get with my analysis

How do.i broaden it to capote subtle difference s

How translatavl3 is? Answer may be not at all.

Simplify it so it's understandable for biologists and data scientists

May not have right identities of metaboloties. Sensitivity of my test is depe dent on detecting metabolites that changes depending on proces stats put into searchable pathway

Inherent weaknesses based on data and collection.

Metaboanalyst used for human pathway analysis. Has the map to keeg Kyoto mapped targets. Has to be known to be present in that pathway analysis if unknown not part of keeg analysis

Look at what the weakness is for dseq2

Ipa software might be better

# 9. References

Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D., & Craig, D. W. (2016). Translating RNA sequencing into clinical diagnostics: Opportunities and challenges. *Nature Reviews Genetics, 17*(5), 257–271. <https://doi.org/10.1038/nrg.2016.10>

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology, 17*, 13. <https://doi.org/10.1186/s13059-016-0881-8>

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology, 18*, 83. <https://doi.org/10.1186/s13059-017-1215-1>

Meeks, L. M., de Oliveira Pessoa, D., Martinez, J. A., Limesand, K. H., & Padi, M. (2021). *Integrating metabolomics and transcriptomics reveals convergent pathways driving radiation-induced salivary gland dysfunction* [Data set]. NCBI Gene Expression Omnibus.<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE155902>​[NCBI+1PMC+1](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE155902&utm_source=chatgpt.com)

Meeks, L. M., de Oliveira Pessoa, D., Martinez, J. A., Limesand, K. H., & Padi, M. (2020). *Supplementary data for integrating metabolomics and transcriptomics reveals convergent pathways driving radiation-induced salivary gland dysfunction* [Data set]. Zenodo.<https://doi.org/10.5281/zenodo.4391402>

## 

## 

### 

### 

### 

### 

### 

### 

# Project Scope

Goal:

To guide students through the complete workflow of integrating RNA-sequencing and metabolomic data, with a focus on feature selection and pathway-based analysis.

Learning Objectives:

* Understand the biological and analytical foundation of RNA-seq and metabolomics.
* Preprocess, normalize, and explore multi-omics data.
* Perform feature selection using statistical and pathway-based methods.
* Interpret findings within a biological context (e.g., radiation-induced responses).
* Compare analysis workflows in both R and Python.

### 

# WhitePaper Outline

### **1. Introduction**

* Purpose and scope of RNA-seq and metabolomic integration
* Research context (e.g., radiation-induced tissue damage)

### **2. Data Types & Study Design**

* RNA-seq read counts and metabolomic abundance profiles
* Experimental groups (Control vs Irradiated)
* Sample metadata overview

### **3. RNA-seq Analysis Pipeline**

* Quality control (brief mention, optional if preprocessed)
* Normalization methods (e.g., DESeq2)
* Differential expression (DEG) analysis
* Feature selection considerations for big-P, small-N

### **4. Metabolomics Analysis Pipeline**

* Normalization and filtering
* Identifying differentially abundant metabolites
* Linking features to metabolite identities

### **5. Pathway Mapping and Enrichment**

* How to map genes and metabolites to KEGG/Reactome pathways
* Enrichment analysis techniques
* Tools used (Pathview, MetaboAnalyst, etc.)

### **6. Data Integration Methods**

* Correlation analysis between omics
* Multivariate approaches (e.g., mixOmics, sPLS)
* Case study: Top pathways affected in radiation

### **7. Comparative Implementation in R and Python**

* Code examples or table comparing steps in both languages
* Strengths/limitations of tools in each

### **8. Conclusion and Teaching Considerations**

* Best practices when teaching these pipelines
* Key challenges (e.g., annotation, pathway overlap, sample size)

### **9. References**