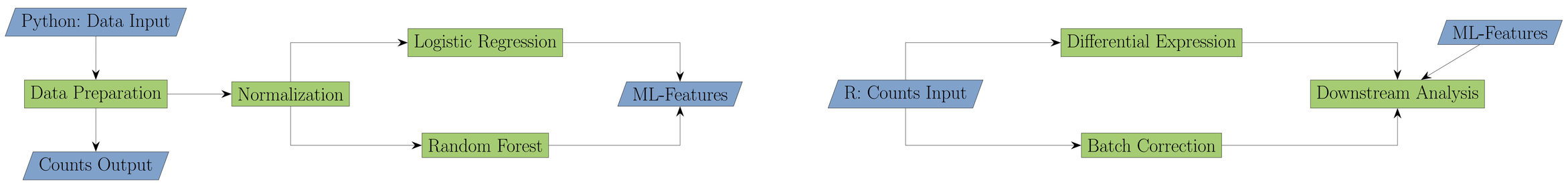
Robustness of machine learning against Batch effects in RNA-seq data

Practical phase of Laszlo Lang
Supervised by Dr. Maximilian Sprang – The Mayer Lab Mainz

Introduction

RNA sequencing (RNA-seq) enables the systematic study of gene expression and provides insights into biological processes at the molecular level. However, RNA-seq data are often affected by batch effects. These arise from systematic technical variation, for example, due to sequencing runs, sample preparation, or laboratory conditions. Such variation can obscure true biological signals and complicate downstream analyses. Machine learning (ML) methods are powerful tools for detecting complex patterns in high-dimensional data. Yet, in the context of RNA-seq, it remains unclear whether ML models capture genuine biological variation or adapt to technical artifacts.
In this project, we investigate the robustness of selected ML methods against batch effects in RNA-seq data. Classical differential expression analysis is compared with ML-based approaches, and the resulting gene lists are interpreted in biological context using over-representation analysis (ORA) and correlation analysis. This allows us to highlight differences between methods and assess which biological pathways remain consistently detectable despite batch effects.

Workflow & Methodology



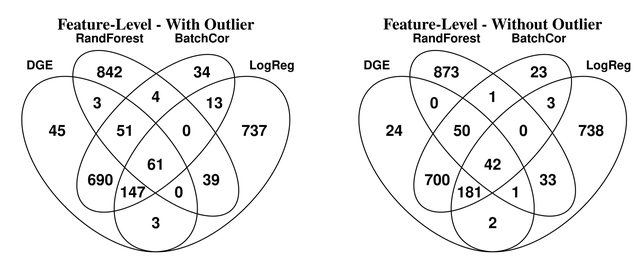
An HCC RNA-seq dataset was analyzed with four feature selection strategies: classical Differential Gene Expression analysis (DGE), batch-corrected DGE (BatchCor), Logistic Regression (LogReg) and Random Forest Classification (RandForest). Each method was used to identify the top 1000 features, which were then used for over-representation analysis (ORA) to extract the top 100 pathways.
Overlaps between feature lists and pathway lists were visualized with Venn diagrams. To evaluate the stability of selected signatures under batch effects, two correlation analyses were performed.
One compares each dataset with the feature lists identified by a method with and without the outlier. The other compares each dataset with method-specific consistent features (intersection, $\cap$, found in both conditions) versus variable features (symmetric difference, Δ, found only in one condition), focusing on features uniquely detected by the respective method.

Downstream Analysis Results

Feature-Level Overlaps (Top 1000 Features):

- **With Outlier**: 61 features shared across all methods; most are method-specific (RandForest: 842; LogReg: 737; DGE: 45; BatchCor: 34).
- **Without Outlier**: Common overlap drops to 42; DGE/BatchCor lose many unique features (45 → 24; 34 → 23).
- Random Forest and Logistic Regression remain similar (873 and 738 unique), but largely method-exclusive.
- Only 18 features are shared in both conditions ⇒ even the common core is unstable.

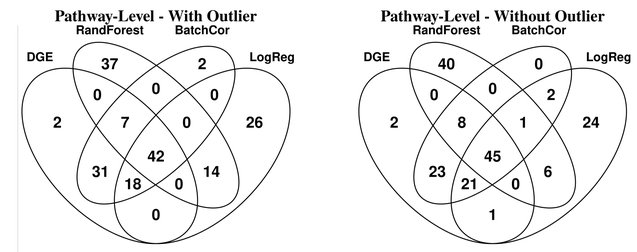
Interpretation: ML-based methods yield stable counts but mostly method-specific feature sets with limited overlap to each other and to classical approaches. In contrast, DGE and BatchCor show the largest overlap, consistent with their shared methodology. Overall, classical methods converge on more similar signatures, whereas ML approaches produce distinct yet internally stable sets, raising concerns about the robustness of feature selection.



Pathway-Level Overlaps (Top 100 Pathways):

- **With Outlier**: 42 pathways shared across all methods.
- **Without Outlier**: Overlap rises slightly to 45.
- Only 29 pathways are identical in both conditions ⇒ shared core is limited.
- Unique pathways: RandForest 37 → 40; LogReg 26 → 24; DGE 2 → 2; BatchCor 2 → 0.

Interpretation: A moderate pathway core is consistently recovered, but many remain method-specific. DGE and BatchCor show strong overlap, whereas ML methods yield more distinct results. Thus, enrichment outcomes are partly robust but also shaped by method-related artifacts, which limits the reliability of unique findings.



Correlation Analysis:

Method-Derived Feature Sets:
Features identified by each method showed very high correlations across most datasets, confirming stable expression patterns.
The outlier dataset GSE144269 was a clear exception, where correlations dropped significantly, confirming its disruptive effect.
Correlation levels varied slightly between methods, reflecting differences in feature selection strategies.

Method-Specific Feature Sets:

- **DGE:** Produces many more variable than consistent features; the consistent set shows markedly lower correlations, in some cases approaching zero or negative values.
- **Batch Correction:** Nearly perfect correlations ($\sim$ 0.99 - 1.0), suggesting effective batch adjustment and robustness.
- **Random Forest:** Very high correlations for both consistent and variable features ($\sim$ 0.97 - 0.99), indicating strong stability.
- **Logistic Regression:** Solid correlations ($\sim$ 0.84 - 0.96), somewhat lower but still robust.

Interpretation: Correlation analysis suggests that features consistently detected across methods represent robust biological signals. Random Forest and Batch Correction appear to produce the most stable signatures, while DGE seems sensitive to dataset variability and outliers.

Method-Derived Feature Sets - With (A) vs. Without (B) Outlier

	GSE25599	GSE77314	GSE82177	GSE101432	GSE105130	GSE114564	GSE144269	GSE148355
DGE (A)	0.87	0.95	0.73	0.95	0.93	0.96	0.33	0.95
DGE (B)	0.69	0.86	0.76	0.9	0.8	0.91	0.62	0.91
BatchCor (A)	0.87	0.95	0.73	0.95	0.93	0.96	0.33	0.95
BatchCor (B)	0.76	0.9	0.79	0.93	0.86	0.93	0.52	0.93
LogReg (A)	0.76	0.94	0.73	0.93	0.82	0.94	0.7	0.93
LogReg (B)	0.74	0.89	0.71	0.94	0.92	0.93	0.69	0.95
RandForest (A)	0.96	0.9	0.74	0.96	0.88	0.86	0.39	0.86
RandForest (B)	0.95	0.96	0.97	0.99	0.98	0.99	0.97	0.99
Common (A∩B)	0.99	0.98	1	0.99	0.99	0.99	0.079	0.99
Uncommon (AΔB)	0.98	0.96	0.9	0.98	0.95	0.95	0.95	0.98

Method-Specific Feature Sets - Consistently (A) vs. Variably (B) Identified

	GSE25599	GSE77314	GSE82177	GSE101432	GSE105130	GSE114564	GSE144269	GSE148355
DGE (A)	0.032	-0.27	0.99	1	-0.024	0.9	1	0.98
DGE (B)	0.66	0.64	0.82	0.84	0.67	0.92	0.61	0.92
BatchCor (A)	1	0.99	0.97	1	1	0.99	1	0.98
BatchCor (B)	0.99	0.99	0.99	1	0.99	1	0.096	1
LogReg (A)	0.84	0.87	0.68	0.92	0.77	0.93	0.89	0.91
LogReg (B)	0.88	0.92	0.79	0.94	0.94	0.95	0.84	0.96
RandForest (A)	0.97	0.95	0.92	0.97	0.93	0.98	0.33	0.99
RandForest (B)	0.95	0.94	0.96	0.99	0.95	0.99	0.96	0.99

Abstract / Conclusion

This work evaluates the robustness of machine learning versus classical approaches for RNA-seq feature selection under batch effects.

- **Feature-Level:** Overlaps between methods are limited (61 features with the outlier, 42 without; 18 identical), with most features being method-specific. DGE and BatchCor lose many unique features once the outlier is removed, yet they still show the largest mutual overlap overall. In contrast, RandForest and LogReg remain relatively stable but largely distinct from the classical approaches.
- **Pathway-Level:** A moderate overlap is observed (42 pathways with the outlier, 45 without; 29 identical). DGE and BatchCor show the strongest agreement, while ML approaches contribute many method-exclusive pathways, questioning the reliability of unique results.
- **Correlation:** Consistent features show stable expression patterns across datasets, with the outlier being the main exception. Among methods, Random Forest and BatchCor yield the most stable correlation profiles, whereas DGE appears more sensitive to dataset variation.

Limitations: The analysis is based on a single dataset (HCC) with strong expression signals, and the true ground truth is unknown.
Improvment: Validation on additional datasets (e.g., glioblastoma or diseases with weaker DEG signals) and on simulated data with defined ground truth would help confirm these findings.

Final assessment: The results suggest that machine learning methods such as Random Forest may be less affected by batch variation, while classical methods converge more strongly on shared results.
Both strategies reveal a partly stable biological core at the pathway level, but substantial method-specific differences remain.
Broader validation is needed to assess whether ML models capture robust biology or adapt to technical artifacts.