AI-Enhanced RUV Normalization: Current State Analysis

Document Date: November 2025

Context: Analysis of AI/ML applications to RUV-III with PRPS normalization methods for RNA-seq data

Executive Summary

The RUV-III with PRPS (Pseudo-Replicates of Pseudo-Samples) method represents a state-of-the-art approach for removing unwanted variation from large-scale RNA-seq data, as demonstrated by Molania et al. (2023) in *Nature Biotechnology*. However, this sophisticated statistical framework currently relies heavily on manual parameter selection and expert judgment. Our analysis reveals a significant gap: no published work has attempted to harness modern AI/ML techniques to automate or enhance the RUV normalization process. This document outlines the current state of AI/ML applications in related RNA-seq analysis tasks and identifies critical opportunities for innovation.

---

1. Existing RUV Methods and Tools

1.1 Classical RUV Framework

The RUV (Removing Unwanted Variation) family of methods includes several key approaches:

- **RUV-III** (Molania et al., 2019): Requires technical replicates and negative control genes
- **RUV-III with PRPS** (Molania et al., 2023): Extends to datasets without technical replicates
- **RUV-III-NB** (Salim et al., 2022): Adaptation for single-cell RNA-seq using negative binomial GLM
- **RUVSeq** (Risso et al., 2014): Earlier versions including RUVg, RUVs, RUVr
- **RUVnormalize** (Jacob et al., 2012): For unsupervised normalization tasks

1.2 Current Software Ecosystem

RUVprps R package: In development by Molania et al. (expected release 2024-2025)

- Mentions "unsupervised methods for identifying PRPS and negative control genes"
- Provides "comprehensive diagnostic and assessment tools"
- Claims to offer methods "particularly when biological variation is unknown"

Key observation: These "unsupervised methods" appear to be classical statistical approaches (clustering, PCA-based) rather than modern machine learning.

GitHub repositories:

- https://github.com/RMolania/RUVprps (package in development)
- https://github.com/RMolania/TCGA_PanCancer_UnwantedVariation (analysis code)
- https://github.com/drisso/RUVSeq (established Bioconductor package)
- https://github.com/limfuxing/ruvIIInb (single-cell version)

---

2. AI/ML Applications in Related RNA-seq Analysis Tasks

While we found **no AI/ML applications specifically for RUV normalization**, we identified several adjacent areas where AI/ML is being successfully applied. In each subsection below, we document:

1. The traditional non-ML/AI approach being updated
2. The ML/AI method applied
3. How success is assessed
4. Key supporting references

2.1 Gene Selection and Feature Extraction

Traditional Approach

Gene selection has traditionally relied on **differential expression (DE) analysis** using statistical tests (t-tests, ANOVA, or negative binomial models in tools like DESeq2, edgeR). These approaches test each gene independently for differences between conditions and rank genes by statistical significance (p-values or adjusted p-values).

Success Metrics for Traditional DE:

- False discovery rate (FDR)
- Number of genes called significant
- Overlap with known disease-associated genes
- Validation via qPCR or independent cohorts

ML/AI Approaches Applied

1. Random Forests for Gene Ranking

<cite>Wenric and Shemirani (2018)</cite> applied Random Forest (RF) classification to rank genes based on variable importance measures derived from RF classifiers trained on case-control RNA-seq studies. Their approach leverages the permutation importance metric that measures how much classification accuracy drops when a gene's values are shuffled.

Method: Random Forest classification with permutation-based feature importance **Assessment of Success:**

- The RF-based gene rankings were compared to differential expression analysis results
- Both methods were evaluated on their ability to predict survival outcomes in validation cohorts
- **Key Finding:** The RF method outperformed DE analysis in **9 out of 12 cancer datasets** for identifying survival-associated genes

This demonstrates that ML can capture gene-disease associations missed by univariate statistical testing, likely due to RF's ability to model non-linear relationships and gene-gene interactions.

Reference: Wenric, S. & Shemirani, R. (2018). Using Supervised Learning Methods for Gene Selection in RNA-Seq Case-Control Studies. *Frontiers in Genetics*, 9:297. https://doi.org/10.3389/fgene.2018.00297

---

2. Variational Autoencoders (VAEs) for Gene Selection

The same study by <cite>Wenric and Shemirani (2018)</cite> introduced the **Extreme Pseudo-Samples (EPS) pipeline**, which uses Variational Autoencoders to generate synthetic samples and extract gene rankings.

Method: VAEs create pseudo-samples in latent space that represent "extreme" cases and controls. Gene weights from these synthetic samples are used to rank gene importance.

Traditional Alternative: Creating synthetic samples was not previously done in differential expression; the closest analog would be bootstrapping or resampling, which maintains the original data distribution rather than exploring extreme points in latent space.

Assessment of Success:

- EPS gene rankings were compared to differential expression analysis
- Evaluation via survival analysis on independent validation cohorts
- **Key Finding:** EPS outperformed DE analysis in **8 out of 12 cancer datasets**

Why VAEs Work: VAEs learn a meaningful latent representation of the data where one can sample new points that closely follow real sample statistics but explore extreme regions of biological variation that may be underrepresented in limited datasets.

Reference: Wenric, S. & Shemirani, R. (2018). *Frontiers in Genetics*, 9:297.

Additional VAE Applications:

- <cite>Grønbech et al. (2020)</cite> developed **scVAE**, demonstrating VAEs can model single-cell and bulk RNA-seq data with different likelihood functions (zero-inflated negative binomial for scRNA-seq). Their model achieved higher log-likelihood lower bounds than factor analysis models and showed clear separation of cell types in latent space.
    - **Reference:** Grønbech, C.H., Vording, M.F., Timshel, P.N., Sønderby, C.K., Pers, T.H., & Winther, O. (2020). scVAE: Variational auto-encoders for single-cell gene expression data. *Bioinformatics*, 36(16):4415-4422. https://doi.org/10.1093/bioinformatics/btaa293
- <cite>Way and Greene (2017)</cite> applied VAEs to TCGA pan-cancer RNA-seq data ("Tybalt" model), showing that decoder weights captured gene contributions to learned features and that the latent space encoded biologically meaningful patterns including cancer type, tumor purity, and pathway activations.
    - **Reference:** Way, G.P. & Greene, C.S. (2017). Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pacific Symposium on Biocomputing*, 23:80-91. https://doi.org/10.1142/9789813235533_0008

---

3. Deep Learning Feature Selection

<cite>Multiple studies</cite> have applied neural networks with embedded feature selection layers to identify prognostic genes directly from expression data.

Traditional Approach: Cox proportional hazards regression with LASSO or elastic net regularization for survival-associated gene selection.

ML Approach: Deep neural networks with:

- Integrated feature selection layers (attention mechanisms or learned feature weights)
- End-to-end training on survival outcomes
- Automatic capture of non-linear gene interactions

Assessment of Success: Comparison of survival prediction accuracy (concordance index), discrimination ability (AUC), and clinical utility.

Example Studies:

- <cite>Ching et al. (2018)</cite> review multiple applications showing DNNs outperform traditional methods for survival prediction across multiple cancer types.
    - **Reference:** Ching, T., Zhu, X., & Garmire, L.X. (2018). Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Computational Biology*, 14(4):e1006076.

2.2 Single-Cell RNA-seq Analysis

Single-cell RNA-seq (scRNA-seq) presents unique challenges compared to bulk RNA-seq, including high dimensionality, extreme sparsity (dropout events), and the need to identify rare cell populations. Traditional methods and ML/AI innovations are documented below.

Cell Type Annotation

Traditional Approach: Manual cluster annotation based on:

- Visual inspection of UMAP/t-SNE plots
- Identification of marker genes (differentially expressed genes in each cluster)
- Comparison to known cell type markers from literature
- Expert biological knowledge

This process is subjective, time-consuming, and difficult to reproduce across labs.

ML/AI Approach: Automated Reference-Based Classification

SingleR (<cite>Aran et al., 2019</cite>) represents a breakthrough in automated cell type annotation:

Method:

- Given a reference dataset (bulk or single-cell) with known cell type labels
- Computes Spearman correlations between test cells and reference cell types
- Uses marker genes specific to each cell type
- Performs fine-tuning rounds to increase confidence in assignments

Assessment of Success:

- Concordance with manual annotations (when available)
- Consistency across multiple reference datasets
- Ability to handle unseen cell types (rejection of low-confidence predictions)
- Speed and scalability to large datasets

Key Advantage: Transfers biological expertise embedded in reference datasets to new data automatically, eliminating the need for repeated manual interpretation of clusters.

Reference: Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., Butte, A.J., & Bhattacharya, M. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, 20(2):163-172. https://doi.org/10.1038/s41590-018-0276-y

Benchmark Studies:

- <cite>Sun et al. (2020)</cite> evaluated 10 cell type annotation methods (including SingleR, Seurat, scmap, CHETAH, SingleCellNet) on diverse scRNA-seq datasets. SingleR performed among the top methods for inter-dataset prediction accuracy while maintaining robustness against gene filtering and downsampling.
    - **Reference:** Sun, Q., Zhao, X., Li, R., Liu, X., Zhang, X., Wang, C., & Sun, X. (2020). Evaluation of Cell Type Annotation R Packages on Single-cell RNA-seq Data. *Genomics, Proteomics & Bioinformatics*, 18(6):711-724. https://doi.org/10.1016/j.gpb.2020.07.004

Clustering with Dropout Modeling

Traditional Approach:

- k-means clustering on PCA-reduced dimensions
- Hierarchical clustering
- Graph-based clustering (Louvain, Leiden algorithms)

Challenge: These methods treat zero counts as true zeros, but scRNA-seq has pervasive "dropout events" — biological transcripts present in cells but not captured due to technical limitations, resulting in false zero observations.

ML/AI Approach: Model-Based Deep Learning

<cite>Tian et al. (2019)</cite> developed **scDeepCluster**, combining:

- Deep embedded clustering (learns representations and clusters simultaneously)
- Explicit modeling of dropout events via zero-inflated negative binomial distribution
- Autoencoder architecture that reconstructs expression while discovering clusters

Method Details:

- Encoder: Compresses expression data to low-dimensional latent representation
- Decoder: Reconstructs expression from latent space
- Clustering layer: Learns cluster assignments in latent space
- Joint optimization: Minimizes reconstruction loss + clustering loss

Traditional Comparisons: The authors compared scDeepCluster against:

- PCA + k-means
- t-SNE + k-means
- SIMLR (similarity-based learning)
- CIDR (clustering through imputation)

Assessment of Success (Metrics Used):

- **Normalized Mutual Information (NMI):** Measures agreement between predicted and true labels (when available)
- **Adjusted Rand Index (ARI):** Clustering accuracy corrected for chance
- **Clustering Accuracy (CA):** Proportion of correctly assigned cells
- **Scalability:** Computational time as a function of sample size

Key Findings:

- scDeepCluster **outperformed all competing methods** on NMI, ARI, and CA across 4 real datasets and extensive simulations
- Computational time scaled **linearly with sample size** (critical for large datasets)
- Better separation of cell types in latent space visualizations

Reference: Tian, T., Wan, J., Song, Q., & Wei, Z. (2019). Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nature Machine Intelligence*, 1(4):191-198. https://doi.org/10.1038/s42256-019-0037-0

2.3 Batch Correction (Non-RUV Methods)

Batch effects — systematic technical variation between experimental batches — are pervasive in RNA-seq. Numerous methods exist, but **none specifically enhance RUV with AI/ML**.

Traditional Statistical Batch Correction Methods

1. ComBat and ComBat-seq

<cite>Johnson et al. (2007)</cite> developed ComBat using **empirical Bayes** to adjust for batch effects:

- Models batch effects as additive and multiplicative factors
- Shrinks batch effect estimates toward overall batch means using empirical Bayes
- Originally for microarrays, extended to RNA-seq by <cite>Zhang et al. (2020)</cite>

Not ML: ComBat uses statistical modeling and empirical Bayes shrinkage, but does not employ machine learning techniques like neural networks, gradient descent optimization, or representation learning.

References:

- Johnson, W.E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118-127. https://doi.org/10.1093/biostatistics/kxj037
- Zhang, Y., Parmigiani, G., & Johnson, W.E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics*, 2(3):lqaa078. https://doi.org/10.1093/nargab/lqaa078

---

2. Harmony

<cite>Korsunsky et al. (2019)</cite> developed Harmony using **iterative clustering** to integrate datasets:

- Applies soft k-means clustering in PCA space
- Adjusts cluster centroids to remove batch effects
- Iterates until convergence

Not Deep Learning: While iterative and optimization-based, Harmony does not use neural networks or deep learning architectures.

Reference: Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289-1296. https://doi.org/10.1038/s41592-019-0619-0

---

3. Seurat Integration

<cite>Stuart et al. (2019)</cite> developed Seurat v3 integration using **canonical correlation analysis (CCA)** and **mutual nearest neighbors (MNN)**:

- Identifies shared biological variation across datasets via CCA
- Finds "anchors" (mutual nearest neighbors) between datasets
- Uses anchors to harmonize datasets

Not ML: Graph-based method using linear dimensionality reduction and nearest neighbor search.

Reference: Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M. 3rd, Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031

---

4. scVI — A Neural Network Approach

<cite>Lopez et al. (2018)</cite> developed **scVI** (single-cell Variational Inference), which DOES use neural networks:

Method:

- Variational autoencoder architecture
- Models gene expression with zero-inflated negative binomial likelihood
- Learns latent representations that capture biological variation
- **Importantly:** Includes batch as a known covariate in the model, but batch correction is achieved through the learned latent representation, not through a specific batch correction module with learnable parameters

Partial ML for Batch Correction: scVI uses VAEs for data modeling and dimensionality reduction, which indirectly aids batch integration by learning shared biological variation. However, the batch correction itself is not the primary ML

innovation — rather, it's a consequence of learning good latent representations.

Reference: Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053-1058. https://doi.org/10.1038/s41592-018-0229-2

Summary for Batch Correction:

- **ComBat-seq, Harmony, Seurat:** Use empirical Bayes, iterative clustering, and graph methods respectively — not ML/AI
- **scVI:** Uses deep learning (VAE) for modeling, which aids integration, but batch correction is an indirect benefit rather than a targeted ML module
- **Critical Gap:** No method applies ML/AI specifically to optimize **RUV-based** batch correction

---

2.4 General RNA-seq Processing Tasks

Expression Prediction from Epigenetic Data

Traditional Approach: Linear regression models relating histone modifications, DNA methylation, and chromatin accessibility to gene expression.

ML Approach: <cite>Singh et al. (2016)</cite> and others have applied:

- Deep neural networks
- Gradient boosting machines
- Support vector machines with non-linear kernels

Assessment of Success: Correlation between predicted and observed expression (Pearson r, Spearman ϱ), prediction accuracy on held-out chromosomes.

Reference Example: Singh, R., Lanchantin, J., Robins, G., & Qi, Y. (2016). DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, 32(17):i639-i648.

Transcript Quantification

Traditional Approach: Alignment-based quantification (RSEM, Salmon) using expectation-maximization algorithms.

ML Approach: Neural network-based alignment-free quantification methods that learn to predict transcript abundances directly from k-mer counts.

Success Assessment: Accuracy vs gold-standard simulated data, computational speed.

Quality Filtering

Traditional Approach: Hard thresholds on metrics like:

- Number of reads per sample
- Percentage of mitochondrial genes (for scRNA-seq)
- Library size

ML Approach: <cite>Mangiola et al. (2021)</cite> developed probabilistic outlier identification using mixture models and machine learning-based anomaly detection.

Success Assessment: Concordance with manual QC decisions, downstream analysis quality (e.g., reduced batch effects, clearer clustering).

Reference: Mangiola, S., Papenfuss, A.T., & Thomas, T. (2021). Probabilistic outlier identification for RNA sequencing generalized linear models. *NAR Genomics and Bioinformatics*, 3(1):lqaa107. https://doi.org/10.1093/nargab/lqaa107

---

3. Critical Gaps: What's Missing

Based on comprehensive literature search (PubMed, Google Scholar, bioRxiv, GitHub), we found **zero publications or preprints** applying modern AI/ML to:

3.1 Automated Negative Control Gene (NCG) Selection

- **Current practice:** Manual selection based on biological knowledge (housekeeping genes) or statistical filters (low variance in biology, high variance in batch)
- **Opportunity:** Train supervised ML models to predict NCG suitability from gene characteristics (expression level, variance patterns, GO annotations, evolutionary conservation)

3.2 Intelligent PRPS Construction

- **Current practice:** Users manually define biological subpopulations (e.g., cancer subtypes), then create pseudo-samples by averaging
- **Opportunity:** Unsupervised deep learning (e.g., deep clustering, graph neural networks) to discover optimal biological groupings automatically

3.3 K Parameter Optimization

- **Current practice:** Try multiple K values, manually evaluate RLE plots and PCA, select "best" K subjectively
- **Opportunity:** Reinforcement learning or Bayesian optimization to automatically select optimal K based on composite quality metrics

3.4 Unknown Batch Detection

- **Current practice:** Visual inspection of RLE plots, heatmaps, and PCA; ad-hoc clustering of RLE medians
- **Opportunity:** Deep learning anomaly detection (VAEs, isolation forests) to systematically identify latent batch effects

3.5 Cross-Study Transfer Learning

- **Current practice:** Normalize each study independently
- **Opportunity:** Pre-train models on large datasets (TCGA) to improve normalization quality in small studies via transfer learning

3.6 End-to-End Automated Pipelines

- **Current practice:** Multi-step manual process requiring expert judgment at each stage
- **Opportunity:** Integrated AI system making principled, automated decisions across all RUV steps

---

4. Why This Gap Exists

Possible Reasons:

1. **Recency:** RUV-III with PRPS only published in 2023; AI applications may be in progress but not yet published
2. **Complexity:** RUV methods are sophisticated statistical frameworks; integrating ML requires deep understanding of both domains
3. **Conservative Field:** Bioinformatics often adopts new computational methods slowly due to reproducibility and interpretability concerns
4. **Data Requirements:** Training AI models requires large annotated datasets of "good" vs "bad" normalizations — difficult to obtain
5. **Interpretability Needs:** Biologists prefer interpretable methods; black-box AI faces resistance
6. **Success of Current Methods:** RUV-III works well; less pressure to innovate when existing methods are effective

---

5. Recent Trends Suggesting Readiness

Factors Indicating Opportunity:

1. **Software Development:** Active RUVprps package development shows continued community interest
2. **Single-Cell Success:** ML applications in scRNA-seq (scDeepCluster, scVI, SingleR) demonstrate feasibility and acceptance
3. **Computational Biology ML Explosion:** Dramatic increase in ML applications (AlphaFold, protein language models, etc.) shows field openness
4. **Data Availability:** TCGA and other large datasets provide training opportunities
5. **Community Openness:** Molania et al. made code, data, and methods publicly available, facilitating follow-up work

---

6. Stakeholder Analysis

Who Would Benefit:

- **Cancer researchers:** Using TCGA or multi-site clinical trial data
- **Single-cell genomicists:** Dealing with complex batch structures
- **Pharmaceutical companies:** Normalizing multi-center drug trial transcriptomics
- **Bioinformaticians:** Seeking automated, reproducible normalization pipelines
- **Precision medicine initiatives:** Requiring robust cross-platform integration

Key Contacts:

- **Ramyar Molania:** molania.r@wehi.edu.au (now at Dana-Farber Cancer Institute)
- **Terence P. Speed:** speed@wehi.edu.au (Walter & Eliza Hall Institute)
- **Johann Gagnon-Bartsch:** jgagnon@umich.edu (University of Michigan Statistics)
- **Anthony Papenfuss:** papenfuss@wehi.edu.au (Walter & Eliza Hall Institute)

---

7. Conclusion

This analysis reveals a significant **untapped opportunity**: while AI/ML has successfully been applied to many RNA-seq analysis tasks — gene selection, cell type annotation, clustering, and quality control — **no work has yet leveraged these techniques for RUV-based normalization**. Given the success of ML in adjacent areas and the manual, expert-intensive

nature of current RUV workflows, AI-enhanced RUV normalization represents a high-impact research direction with clear unmet need and technical feasibility.

The next document (Part 2) will detail a comprehensive research proposal to develop these AI-enhanced RUV methods.

References

Primary RUV Publications

1. Molania, R., Gagnon-Bartsch, J.A., Dobrovic, A., & Speed, T.P. (2023). Removing unwanted variation from large-scale RNA sequencing data with PRPS. *Nature Biotechnology*, 41:82–95. https://doi.org/10.1038/s41587-022-01440-w
2. Molania, R., Foroutan, M., Gagnon-Bartsch, J.A., Gandolfo, L.C., Jain, A., Sinha, A., Olshansky, G., Dobrovic, A., Papenfuss, A.T., & Speed, T.P. (2019). A new normalization for Nanostring nCounter gene expression data. *Nucleic Acids Research*, 47(12):6073–6083. https://doi.org/10.1093/nar/gkz433
3. Salim, A., Gagnon-Bartsch, J.A., Speed, T.P., & Bahlo, M. (2022). RUV-III-NB: Normalization of single cell RNA-seq data. *Nucleic Acids Research*, 50(16):e96. https://doi.org/10.1093/nar/gkac486
4. Risso, D., Ngai, J., Speed, T.P., & Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32:896–902. https://doi.org/10.1038/nbt.2931
5. Gagnon-Bartsch, J.A. & Speed, T.P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552. https://doi.org/10.1093/biostatistics/kxr034
6. Jacob, L., Gagnon-Bartsch, J.A., & Speed, T.P. (2012). Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics*, 17(1):16-28.

AI/ML in Gene Selection

7. Wenric, S. & Shemirani, R. (2018). Using Supervised Learning Methods for Gene Selection in RNA-Seq Case-Control Studies. *Frontiers in Genetics*, 9:297. https://doi.org/10.3389/fgene.2018.00297
8. Ching, T., Zhu, X., & Garmire, L.X. (2018). Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Computational Biology*, 14(4):e1006076. https://doi.org/10.1371/journal.pcbi.1006076
9. Swan, A.L., Mobasheri, A., Allaway, D., Liddell, S., & Bacardit, J. (2013). Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *OMICS*, 17(12):595-610. https://doi.org/10.1089/omi.2013.0017

AI/ML with Variational Autoencoders

10. Grønbech, C.H., Vording, M.F., Timshel, P.N., Sønderby, C.K., Pers, T.H., & Winther, O. (2020). scVAE: Variational auto-encoders for single-cell gene expression data. *Bioinformatics*, 36(16):4415-4422. https://doi.org/10.1093/bioinformatics/btaa293
11. Way, G.P. & Greene, C.S. (2017). Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pacific Symposium on Biocomputing*, 23:80-91. https://doi.org/10.1142/9789813235533_0008
12. Kingma, D.P. & Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint* arXiv:1312.6114. https://arxiv.org/abs/1312.6114
13. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053-1058. https://doi.org/10.1038/s41592-018-0229-2
14. Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K.L., Streets, A., & Yosef, N. (2021). Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18(3):272-282. https://doi.org/10.1038/s41592-020-01050-x
15. Ahlmann-Eltze, C. & Huber, W. (2023). Comparison of transformations for single-cell RNA-seq data. *Nature Methods*, 20(5):665-672. https://doi.org/10.1038/s41592-023-01814-1

Single-Cell Clustering with Deep Learning

16. Tian, T., Wan, J., Song, Q., & Wei, Z. (2019). Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nature Machine Intelligence*, 1(4):191-198. https://doi.org/10.1038/s42256-019-0037-0
17. Tian, T., Zhang, J., Lin, X., Wei, Z., & Hakonarson, H. (2021). Model-based deep embedding for constrained clustering analysis of single cell RNA-seq data. *Nature Communications*, 12:1873. https://doi.org/10.1038/s41467-021-22008-3
18. Xie, J., Girshick, R., & Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. *Proceedings of the 33rd International Conference on Machine Learning*, PMLR 48:478-487.
19. Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S., & Theis, F.J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, 10:390. https://doi.org/10.1038/s41467-018-07931-2

Cell Type Annotation

20. Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., Butte, A.J., & Bhattacharya, M. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, 20(2):163-172. https://doi.org/10.1038/s41590-018-0276-y
21. Sun, Q., Zhao, X., Li, R., Liu, X., Zhang, X., Wang, C., & Sun, X. (2020). Evaluation of Cell Type Annotation R Packages on Single-cell RNA-seq Data. *Genomics, Proteomics & Bioinformatics*, 18(6):711-724. https://doi.org/10.1016/j.gpb.2020.07.004
22. Alquicira-Hernandez, J., Sathe, A., Ji, H.P., Nguyen, Q., & Powell, J.E. (2019). scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biology*, 20:264. https://doi.org/10.1186/s13059-019-1862-5
23. Tan, Y. & Cahan, P. (2019). SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. *Cell Systems*, 9(2):207-213.e2. https://doi.org/10.1016/j.cels.2019.06.004
24. de Kanter, J.K., Lijnzaad, P., Candelli, T., Margaritis, T., & Holstege, F.C.P. (2019). CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Research*, 47(16):e95. https://doi.org/10.1093/nar/gkz543

Batch Correction Methods

25. Johnson, W.E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118-127. https://doi.org/10.1093/biostatistics/kxj037
26. Zhang, Y., Parmigiani, G., & Johnson, W.E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics*, 2(3):lqaa078. https://doi.org/10.1093/nargab/lqaa078
27. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289-1296. https://doi.org/10.1038/s41592-019-0619-0
28. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M. 3rd, Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031
29. Haghverdi, L., Lun, A.T.L., Morgan, M.D., & Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36:421–427. https://doi.org/10.1038/nbt.4091

Quality Control

30. Mangiola, S., Papenfuss, A.T., & Thomas, T. (2021). Probabilistic outlier identification for RNA sequencing generalized linear models. *NAR Genomics and Bioinformatics*, 3(1):lqaa107. https://doi.org/10.1093/nargab/lqaa107
31. Ilicic, T., Kim, J.K., Kolodziejczyk, A.A., Bagger, F.O., McCarthy, D.J., Marioni, J.C., & Teichmann, S.A. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*, 17:29. https://doi.org/10.1186/s13059-016-0888-1

Comprehensive ML Reviews in Genomics

32. Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F.J. (2019). Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20:389–403. https://doi.org/10.1038/s41576-019-0122-6
33. Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2019). A primer on deep learning in genomics. *Nature Genetics*, 51:12–18. https://doi.org/10.1038/s41588-018-0295-5
34. Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C., & Collins, J.J. (2018). Next-Generation Machine Learning for Biological Networks. *Cell*, 173(7):1581-1592. https://doi.org/10.1016/j.cell.2018.05.015

Document Status: Complete - Part 1 of 2
Next Document: Part 2 - Research Proposal for AI-Enhanced RUV Normalization