Summary: Removing Unwanted Variation from Large-Scale RNA Sequencing Data with PRPS

Citation: Molania, R., Foroutan, M., Gagnon-Bartsch, J.A., Gandolfo, L.C., Jain, A., Sinha, A., Olshansky, G., Dobrovic, A., Papenfuss, A.T., and 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

---

Executive Summary

This paper presents **PRPS (Pseudo-Replicates of Pseudo-Samples)**, a strategy for deploying the RUV-III normalization method to remove unwanted variation from large-scale RNA-seq data, particularly from The Cancer Genome Atlas (TCGA). The method addresses critical limitations of standard normalization approaches (FPKM, FPKM-UQ) that fail to adequately remove technical variation while preserving biological signal.

Key Problem

RNA-seq data from large studies like TCGA are significantly compromised by unwanted variation from sources including:

- **Library size differences** - Variation in sequencing depth between samples
- **Tumor purity variation** - The proportion of cancer cells versus stromal/immune cells in tumor samples
- **Batch effects** - Technical variation from plates, time, sequencing chemistry, and facilities

These variations can affect:

- Cancer subtype identification
- Gene expression-survival associations
- Gene co-expression analyses

Standard normalizations like FPKM and FPKM-UQ failed to adequately remove these variations.

Main Findings

1. Sources of Unwanted Variation in TCGA Data

The authors systematically characterized unwanted variation across TCGA RNA-seq datasets (~11,000 samples from 33 cancer types):

- **Library size:** The first five principal components were strongly associated with log library size in raw gene counts. FPKM and FPKM-UQ reduced but did not eliminate these effects in several cancer types.
- **Tumor purity:** Substantial variation in tumor purity was present, and FPKM and FPKM-UQ normalizations could not correct for this.
- **Plate/batch effects:** Vector correlation and ANOVA revealed plate effects persisted in raw counts, FPKM, and FPKM-UQ normalized datasets.

2. Critical Gene-Level Discovery

A key finding: Gene-level counts are not uniformly proportional to library size. The authors identified four groups of genes with different relationships to global scaling factors:

1. **Proportional genes** - Expression scales with library size (FPKM/FPKM-UQ work well)
2. **Over-expressed genes** - Expression greater than expected (FPKM/FPKM-UQ insufficient)
3. **Uncorrelated genes** - No association with library size (FPKM/FPKM-UQ introduce artifacts)
4. **Inversely correlated genes** - Negative association (FPKM/FPKM-UQ exacerbate problems)

This heterogeneity explains why single scaling factor normalizations fail for many genes.

3. RUV-III with PRPS Solution

How PRPS works:

- Creates in silico "pseudo-samples" from small groups of biologically similar samples that differ in unwanted variation
- Groups pseudo-samples with the same biology as "pseudo-replicates"
- Gene expression differences between pseudo-replicates largely represent unwanted variation
- RUV-III uses these differences plus negative control genes to estimate and remove unwanted variation

Requirements:

- Identification of major biological subpopulations in the data
- Selection of negative control genes (not affected by biology but affected by technical variation)
- Determination of dimensionality K for unwanted variation

Detailed Results by Dataset

TCGA READ (Rectum Adenocarcinoma) RNA-seq Study

Study characteristics:

- 176 samples across 14 plates over 4 years
- Major library size variation between 2010 samples vs. 2011-2014 samples
- Biological populations: 4 consensus molecular subtypes (CMS)

Key improvements with RUV-III:

1. **Library size and batch removal:**
   - RUV-III effectively removed library size variation that persisted in FPKM/FPKM-UQ
   - Better mixing of samples across time intervals (measured by silhouette coefficients and ARI)
   - Reduced spurious differential expression between high and low library size samples
2. **Better subtype separation:**
   - Improved separation of CMS clusters in PCA plots
   - Vector correlation analyses confirmed better alignment between PCs and biological subtypes
   - CMS survival differences (CMS2 vs CMS4) were clearer with RUV-III data
3. **Improved gene co-expression:**
   - **Removed spurious correlations:** TMF1 and BCLAF1 genes showed correlation of ϱ=0.7-0.8 in FPKM/FPKM-UQ but no correlation in RUV-III (validated by microarray data)
   - **Revealed true correlations:** MDH2 and EIF4H genes showed overall correlation ϱ=-0.05 in FPKM-UQ but ϱ=0.7 in RUV-III (validated by microarray data)
   - Analysis of 500 genes most correlated with library size showed numerous spurious correlations in FPKM-UQ that were eliminated by RUV-III

4. **Enhanced survival associations:**
    - RAB18 and FBXL14 gene associations with overall survival were revealed by RUV-III but obscured in TCGA normalizations
    - In FPKM/FPKM-UQ, median expression splits mainly separated high vs. low library size groups rather than biologically meaningful groups

"Money Shot" Figure: Figure 2 demonstrates RUV-III improvements across multiple panels:

- PCA plots showing batch mixing (colored by time) and biological separation (colored by CMS)
- R^2 values for library size associations (reduced in RUV-III)
- Silhouette coefficients and ARI scores (improved in RUV-III)
- Spearman correlations between gene expression and library size (dramatically reduced in RUV-III)

TCGA BRCA (Breast Cancer) RNA-seq Study

Study characteristics:

- 1,180 samples from 40 tissue source sites across 38 plates over 5 years (2010-2014)
- Samples processed using two different flow cell chemistries
- 94 adjacent normal samples, 7 paired primary-metastatic samples
- Biological populations: 5 PAM50 subtypes (HER2-enriched, basal-like, luminal A, luminal B, normal-like)

Key improvements with RUV-III:

1. **Tumor purity removal:**
    - RUV-III substantially removed tumor purity variation within PAM50 subtypes (measured by linear regression R^2)
    - Spearman correlations between individual genes and tumor purity dramatically reduced
    - Differential expression analysis between low/high purity samples showed uniform p-value distributions in RUV-III (indicating proper normalization)
    - Tumor purity score variance significantly smaller in RUV-III data
2. **Flow cell chemistry and unknown batch removal:**
    - RUV-III effectively removed known flow cell chemistry effects
    - **Discovery of unknown variation:** Heat map revealed two clusters within first flow cell chemistry samples, suggesting additional unknown sources
    - Batch scoring identified 4 groups in FPKM-UQ that were eliminated in RUV-III
    - ANOVA F-statistics for flow cell chemistry effects dramatically reduced
    - Silhouette analyses and ARI confirmed better sample mixing
3. **Corrected gene co-expression:**
    - **Tumor purity artifacts:** ZEB2 and ETS1 showed high correlation in FPKM-UQ (both correlated with tumor purity) but no correlation in RUV-III or laser capture microdissection (LCM) data
    - Analysis of 1,300 genes highly correlated with tumor purity showed many spurious pair-wise correlations in FPKM-UQ that disappeared when adjusting for purity
    - **Batch artifacts:** ESRRA and MAP3K2 showed positive correlation in FPKM-UQ but no correlation in RUV-III or TCGA microarray data
    - **Simpson's paradox example:** E2F4 and CNOT1 showed overall correlation ϱ=0.1 but within-batch average ϱ=0.4 in FPKM-UQ; RUV-III and microarray showed high positive correlation
4. **Enhanced survival associations:**
    - ZEB2 high expression associated with poor outcome revealed by RUV-III (obscured by tumor purity in FPKM-UQ)
    - STAB1 gene survival association only evident after removing tumor purity variation
    - FOXA1 expression within luminal B subtype associated with poorer outcome (obscured by tumor purity in FPKM-UQ)
5. **Improved PAM50 subtype separation:**
    - Better PCA clustering of PAM50 subtypes

    - Vector correlation between PCs and subtypes improved
    - Silhouette coefficients and ARI increased
    - Kaplan-Meier survival analysis showed significant associations with patient outcomes
    - Important note: Normal-like subtype in FPKM-UQ data was compromised by low tumor purity; LCM data showed no normal-like subtype

"Money Shot" Figure: Figure 6 comprehensively demonstrates RUV-III improvements:

- R^2 for tumor purity associations within PAM50 subtypes (reduced in RUV-III)
- Box plots of gene-tumor purity correlations (dramatically reduced in RUV-III)
- P-value histograms from differential expression analysis (uniform in RUV-III, non-uniform in FPKM-UQ)
- Vector correlations with flow cell chemistry (reduced in RUV-III)
- Heat map of 400 genes affected by batch effects
- Batch score distributions (4 groups in FPKM-UQ, eliminated in RUV-III)
- Gene-batch score correlations (dramatically reduced in RUV-III)

Removing Unwanted Variation from Unknown Sources

RUV-III's Design for Unknown Factors

The RUV-III method is a linear model through which the presence and impact of **both known and unknown** unwanted factors can be inferred via technical replicates (or pseudo-replicates) and negative control genes.

Discovery of Unknown Source in BRCA Data

The most compelling example of handling unknown variation comes from the BRCA dataset:

- A heat map of genes highly affected by flow cell chemistries revealed two clusters within samples processed by the first flow cell chemistry
- This suggested additional sources of unwanted variation of **unknown origin** within each flow cell chemistry
- Batch scores identified this unknown variation, dividing samples into four groups in FPKM-UQ data
- These groups were **not visible** in RUV-III normalized data, indicating successful removal
- A surprising number of genes showed high correlations with batch scores in FPKM-UQ, whereas these correlations were much lower in RUV-III

General Strategy Using RLE Medians

RLE (Relative Log Expression) medians serve as a tool for identifying unknown sources of unwanted variation:

- In the absence of unwanted variation, RLE medians should be centered around zero
- Any deviation from zero indicates the presence of unwanted variation
- The authors used RLE medians to identify unknown batches within studies where batch information wasn't explicitly documented
- For multi-study normalization, RLE medians were clustered into groups to define batches

Key Principle

The power of RUV-III lies in its ability to **estimate unwanted variation empirically** from the data structure itself, without requiring complete prior knowledge of all sources. It captures residual variation patterns through:

- Pseudo-replicates (capturing differences between similar biological groups)
- Negative control genes (capturing technical but not biological variation)
- Spectral decomposition of residual variation

This makes RUV-III effective against both known and unknown technical artifacts.

Robustness Testing

Performance with Poorly Chosen PRPS

The authors tested robustness by randomly shuffling 20%, 40%, 60%, and 80% of biological labels:

READ dataset:

- RUV-III outperformed FPKM/FPKM-UQ even with poorly chosen PRPS
- Library size removal and CMS cluster preservation maintained
- Gene-gene correlations preserved
- RAB18 survival association identified in all RUV-III datasets
- FBXL14 survival association found only with 20% shuffled labels

BRCA dataset:

- Satisfactory performance with poorly chosen PRPS for removing flow cell chemistry and tumor purity
- Slightly lower PAM50 separation with 60% and 80% shuffled labels
- Gene-gene correlations and survival associations still well-maintained

Performance with Partially Known Biology

Testing with only subset of biological subtypes to create PRPS:

READ dataset:

- Used only CMS4 subtype (present in 8 of 14 plates)
- Very satisfactory normalization achieved
- Library size and plate effects removed
- CMS clusters preserved

BRCA dataset:

- Used only basal and luminal A subtypes
- Performance largely similar to using all PAM50 subtypes

Conclusion: RUV-III works well even with incomplete biological knowledge.

Multi-Study Integration

RUV-III with PRPS successfully normalized three large breast cancer RNA-seq datasets:

- TCGA BRCA
- Two Brueffer et al. cohorts

Process:

- Used RLE medians to identify batches within each study
- Created PRPS using PAM50 subtypes as biological populations
- Successfully removed between-study and within-study variations
- Preserved PAM50 clusters

- Well-known gene-gene correlations maintained
- Outperformed quantile and upper quartile normalizations

Methodology Highlights

The RUV-III Linear Model

The method uses a linear model:

$$Y = 1\mu + MX\beta + W\alpha + \varepsilon$$

Where:

- Y = gene expression data (m assays × n genes)
- M = mapping matrix (assays to samples)
- X = design matrix for biological factors
- β = biological coefficients
- W = matrix capturing unwanted variation (m × k dimensions)
- α = unwanted variation coefficients
- ε = random error

Key innovation: Uses residual projector R_M to extract information about unwanted variation from differences between replicates (or pseudo-replicates).

Negative Control Genes

Selection strategies varied by dataset but generally included:

- Genes with low F-statistics for biological factors (not affected by biology)
- Genes with high correlations with known unwanted factors
- Housekeeping genes with stable expression
- Genes showing clear unwanted but no biological variation

Pragmatic approach: If using a gene set as negative controls helps remove unwanted variation (evaluated by metrics), they're acceptable controls.

Choice of K (Dimensionality)

- Repeated analysis with range of K values
- Evaluated quality using statistical metrics and biological knowledge
- Generally robust to overestimating K

Comparison with Standard TCGA Normalizations

FPKM (Fragments Per Kilobase Million)

Limitations identified:

- Assumes all gene counts proportional to single scaling factor
- Fails for genes with no or negative correlation to library size
- Can introduce rather than remove library size variation for certain genes
- Cannot address tumor purity variation
- Sample-wise normalization misses gene-specific batch effects

FPKM-UQ (Upper Quartile Normalization)

Limitations identified:

- Similar issues to FPKM
- Slightly better performance but still substantial residual effects
- Shows shortcomings in multiple cancer types
- Principal components remain correlated with unwanted factors

Why RUV-III Succeeds

1. **Gene-wise rather than sample-wise:** Can remove library size effects only from affected genes
2. **Empirical estimation:** Uses data structure to infer unwanted variation
3. **Flexible:** Can target specific sources while preserving others (e.g., remove library size but keep tumor purity)
4. **Multi-dimensional:** Captures multiple orthogonal sources of unwanted variation
5. **Uses negative controls:** Leverages genes unaffected by biology to estimate technical variation

---

Validation Approaches

The authors used multiple validation strategies:

1. Statistical Metrics

- R^2 from linear regression (PCs vs. unwanted factors)
- Spearman correlations (gene expression vs. unwanted factors)
- Vector correlations (PCs vs. categorical factors)
- Silhouette coefficients (cluster separation/mixing)
- Adjusted Rand Index (cluster agreement)
- ANOVA F-statistics (gene-wise batch effects)
- P-value histograms from differential expression

2. Biological Validation

- PAM50/CMS subtype separation
- Kaplan-Meier survival analyses
- Gene set enrichment analyses

3. Independent Data Validation

- TCGA microarray data (same samples, different platform)
- Laser capture microdissection (LCM) data (pure cancer cells)
- Cross-study comparisons

4. Negative/Positive Controls

- RLE plots (should center at zero)
- Known gene-gene correlations
- Known survival associations
- Technical replicate concordance

Important Practical Considerations

When to Use RUV-III with PRPS

Best suited for:

- Large studies without technical replicates
- Studies with tumor purity variation to remove
- Multi-batch/multi-site studies
- When unwanted variation confounded with biology

Requirements:

- At least one homogeneous biological population across unwanted variation sources
- Ability to identify or estimate major biological subpopulations
- Selection of appropriate negative control genes

Recommendations for Study Design

From the Discussion section:

1. **Include technical replicates across sources of unwanted variation** - Positive controls for normalization assessment
2. **Distribute biology across batches** - Avoids complete confounding
3. **Consider homogeneous populations** (e.g., adjacent normal tissue) - But must be distributed across batches
4. **Document all batch/technical factors** - Even if they seem unimportant

Limitations Acknowledged

- Performance depends on identifying major biological populations (though rough identification sufficient)
- Cannot use adjacent normal tissue if not distributed across batches
- Tumor purity same across technical replicates (hence need for PRPS)
- Requires computational expertise for proper implementation

Software and Data Availability

Code

- GitHub repository: https://github.com/RMolania/TCGA_PanCancer_UnwantedVariation

- RShiny application for exploring unwanted variation
- tcgaCleaneR R package
- Comprehensive vignettes for READ and BRCA processing

Data

- RUV-III normalized READ, COAD, and BRCA data: Zenodo (record 6459560)
- TCGA summarized experiment objects: Zenodo (record 6326542)
- Original TCGA data: Available through TCGAbiolinks R package

---

Clinical and Biological Impact

Consequences of Inadequate Normalization

Gene co-expression networks:

- Spurious correlations can suggest false biological relationships
- Obscured true correlations miss important co-expression networks
- Example: MDH2-EIF4H high correlation revealed by RUV-III may indicate novel co-expression network in cancer

Survival biomarkers:

- Library size/purity confounding can mask true survival associations
- Can create false survival associations
- Examples: RAB18, FBXL14, ZEB2, STAB1, FOXA1 associations revealed or clarified

Cancer subtype identification:

- Tumor purity can compromise PAM50 subtype calling
- Normal-like subtype may be artifact of low purity samples
- Batch effects can obscure true biological subtypes

Downstream analyses affected:

- Drug target identification
- Biomarker discovery
- Therapeutic stratification
- Mechanistic studies

Key Figures Summary

Figure 1: Pan-cancer unwanted variation assessment

Shows library size, tumor purity, and plate effects across all 33 TCGA cancer types in raw counts, FPKM, and FPKM-UQ data.

Figure 2: READ dataset normalization comparison

"Money shot" for READ - shows PCA plots, R^2 values, correlations, silhouette coefficients, and ARI comparing raw counts, FPKM, FPKM-UQ, and RUV-III.

Figure 3: Gene co-expression improvements in READ

Shows spurious and obscured correlations in FPKM-UQ vs. true correlations in RUV-III, validated by microarray data.

Figure 4: Survival analysis improvements in READ

Kaplan-Meier plots for RAB18 and FBXL14 genes showing library size confounding in FPKM-UQ vs. true survival associations in RUV-III.

Figure 5: Gene-level library size relationships

Demonstrates four groups of genes with different relationships to scaling factors - explains why global scaling fails.

Figure 6: BRCA dataset normalization comparison

"Money shot" for BRCA - comprehensive assessment of tumor purity, flow cell chemistry, and unknown batch removal.

Figure 7: Gene co-expression and survival in BRCA

Shows tumor purity and batch-induced spurious correlations in FPKM-UQ vs. true relationships in RUV-III, validated by LCM microarray and TCGA microarray data.

Extended Data Figure 1: PAM50 improvements in BRCA

PCA plots, vector correlations, silhouette coefficients, and survival analyses showing improved PAM50 subtype identification.

Extended Data Figure 2: RUV-III with PRPS workflow

Comprehensive workflow diagram for identifying unwanted variation and applying RUV-III normalization.

Conclusions

Main Contributions

1. **Methodological:** PRPS approach enables RUV-III to work without technical replicates and to remove tumor purity variation
2. **Empirical:** Comprehensive characterization of unwanted variation across TCGA data (~11,000 samples)
3. **Discovery:** Gene-level counts not uniformly proportional to library size - explains failure of global scaling methods
4. **Validation:** Extensive demonstration that removing unwanted variation improves biologically meaningful analyses
5. **Practical:** Robust method that works with approximate biological labels and can handle unknown sources of variation

Impact

- Provides principled approach for normalizing large, complex RNA-seq studies
- Can improve cancer subtype identification, biomarker discovery, and survival analyses
- Applicable beyond TCGA to any large transcriptomic studies
- Software and data made publicly available for reproducibility

Future Directions

The methods can be extended to:

- Single-cell RNA-seq normalization
- Other high-throughput genomic platforms
- Integration of multi-omic data
- Prospective clinical trial data normalization

Critical Assessment

Strengths

1. Comprehensive evaluation across multiple cancer types
2. Multiple independent validation strategies
3. Clear demonstration of biological impact
4. Robust to imperfect inputs
5. Handles unknown sources of variation
6. Makes no assumptions about biology-batch orthogonality

Considerations

1. Requires biological subpopulation identification (though rough approximation sufficient)
2. Computationally more intensive than simple scaling methods
3. Choice of negative control genes requires careful consideration
4. K parameter selection needs evaluation
5. Performance depends somewhat on study design (batch-biology distribution)

Comparison to Alternatives

Paper explicitly excludes some methods from comparison:

- SVAseq, ComBat-seq, RUVg: Designed for batch correction when unwanted variation orthogonal to biology (rarely known in advance)
- RUVs: Requires true technical replicates (missing from TCGA)

RUV-III with PRPS fills gap for scenarios where:

- No technical replicates available
- Unwanted variation may be confounded with biology
- Tumor purity needs to be removed
- Multiple complex sources of unwanted variation present

Supplementary Notes

Gene Examples Mentioned

Validated positive controls (known cancer roles):

- BCLAF1: Pro-tumorigenic in colon cancer
- MDH2: Important in cancer growth and metastasis

- EIF4H: Important in cancer metastasis
- RAB18: Cell proliferation and metastasis, poor survival marker
- FBXL14: EMT inhibitor, suppresses metastasis
- ZEB2: EMT regulator, invasion and survival marker
- FOXA1: PAM50 signature gene, tumor purity correlated

Technical genes (negative controls):

- Housekeeping genes (stable expression)
- Genes with low F-statistics for biological factors
- Genes uncorrelated with biology but correlated with technical factors

Statistical Methods Used

- Principal Component Analysis (PCA)
- Linear regression / R^2
- Vector correlation (Rozeboom squared vector correlation)
- Spearman correlation (non-parametric)
- ANOVA / F-statistics
- Wilcoxon signed-rank test
- Silhouette coefficient
- Adjusted Rand Index (ARI)
- Kaplan-Meier survival analysis
- Gene Set Enrichment Analysis (GSEA)

R Packages Utilized

- TCGAbiolinks: TCGA data download
- CMScaller: Colorectal cancer CMS identification
- genefu: PAM50 subtype identification
- singscore: Sample scoring against gene sets
- biomaRt: Gene annotation
- ppcor: Partial correlation
- cluster: Silhouette analysis
- EDAseq: PCA calculations

Glossary of Terms

PRPS: Pseudo-Replicates of Pseudo-Samples - in silico samples created by averaging expression from biologically similar samples

RUV-III: Removing Unwanted Variation III - normalization method using replicates and negative control genes

FPKM: Fragments Per Kilobase Million - standard RNA-seq normalization

FPKM-UQ: FPKM with Upper Quartile normalization - TCGA's standard normalization

RLE: Relative Log Expression - log ratio of gene expression to median, used for quality assessment

CMS: Consensus Molecular Subtypes - colorectal cancer classification (4 subtypes)

PAM50: Prediction Analysis of Microarray 50 - breast cancer classification (5 subtypes)

Tumor purity: Proportion of cancer cells in tumor sample (vs. stromal/immune cells)

Negative control genes: Genes affected by technical but not biological variation

Batch effects: Technical variation from processing in different groups/times/locations

LCM: Laser Capture Microdissection - isolates pure cancer cells

---

Report generated: 2025 Summary compiled from published article in Nature Biotechnology