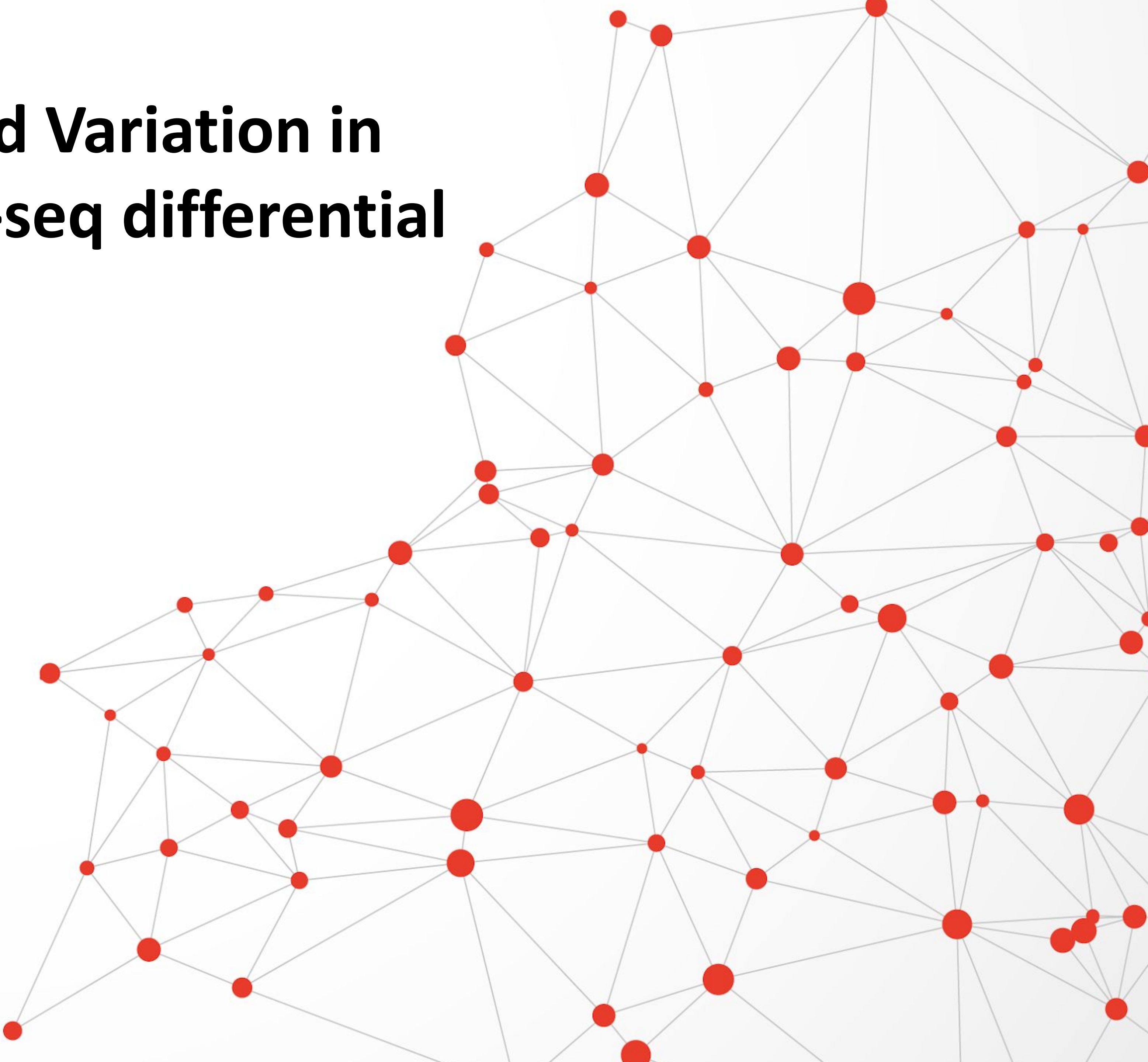
Removing Unwanted Variation in pseudo-bulk scRNA-seq differential expression

Sofia Prieto - Hasselt University

WWW.UHASSELT.BE/DSI



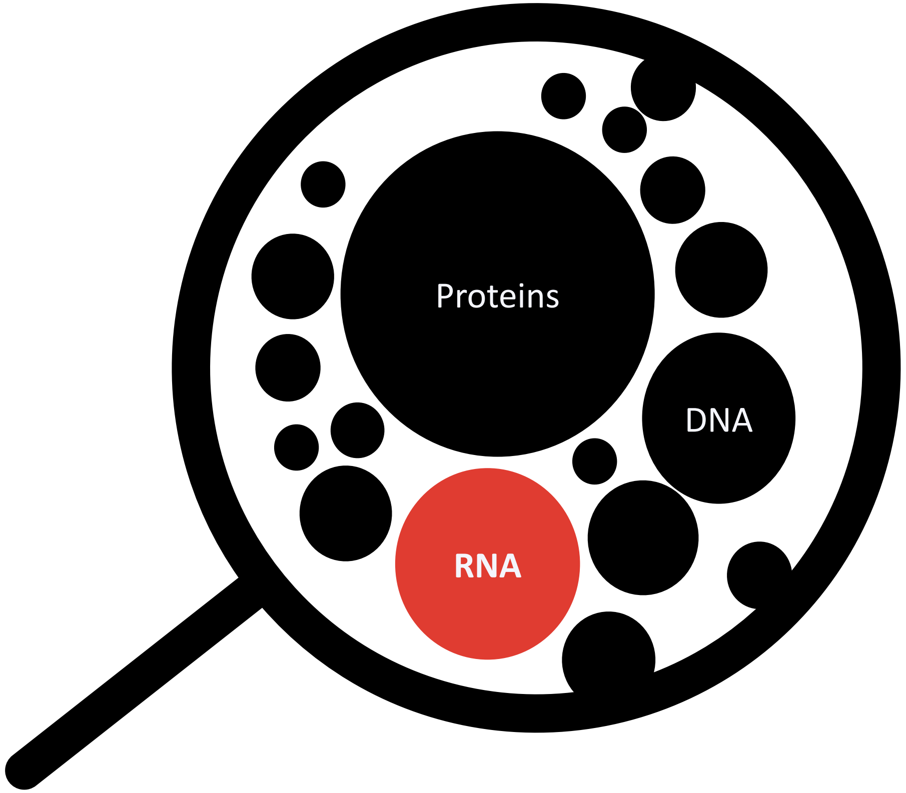
The team

Prof. Olivier Thas - Hasselt University
Prof. Helena Geys - Janssen Pharmaceuticals
Dr. Koen Van den Berge - Janssen Pharmaceuticals
Dr. Marjolein Crabbe - Janssen Pharmaceuticals
Dr. Ewoud De Troyer - Janssen Pharmaceuticals
Dr. Davit Sargsyan - Janssen Pharmaceuticals

WWW.UHASSELT.BE/DSI

THE OMICS FIELD

Quantify characteristics of cells substances



- **Granularity**

  Bulk studies vs **Single-cell studies**

THE SMOOTHIE ANALOGY

From Bulk to Single-Cell

- A blend of various ingredients
- You might detect the stronger flavours, but some go unnoticed

- Taste the flavours separately
- You know the exact proportions of each fruit

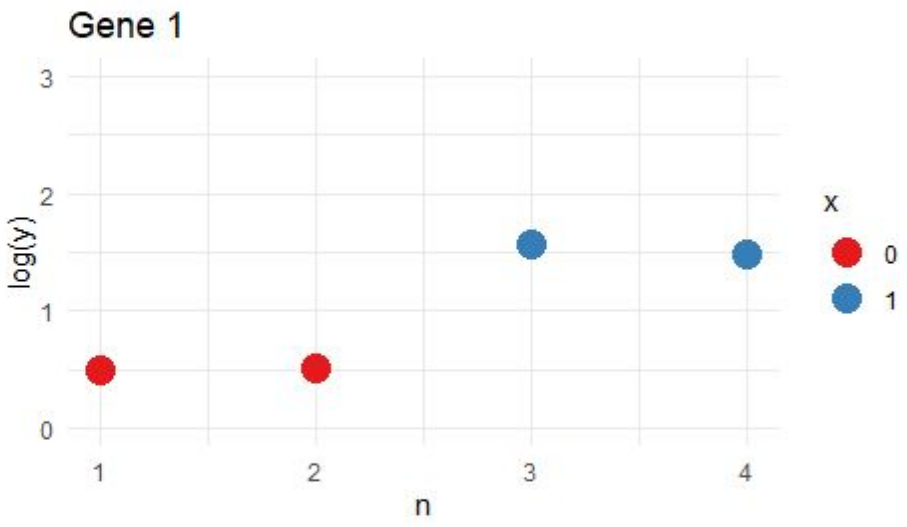
THE SMOOTHIE ANALOGY

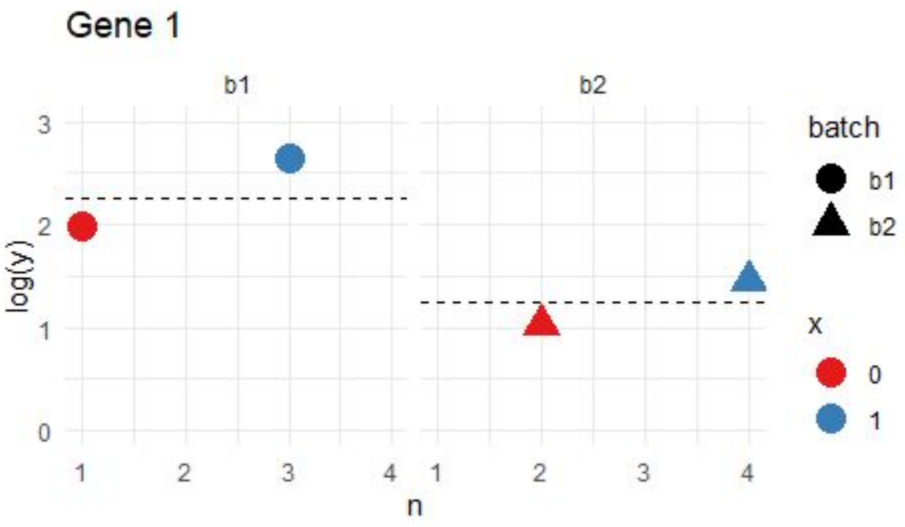
Pseudo-Bulk

- Blends fruits of the same kind
- Taste the flavours separately

CONCEPT

Differential Expression Analysis (DEA) between samples with the same cell-type





- Mean differences between groups

- Unwanted factors such as batches alter the gene expression

CONCEPT

Removing Unwanted Variation (RUV)

DEA

$$\underset{\text{Log transformed counts}}{\log(Y)} = \underset{\text{Intercept}}{\beta_0} + \overset{\text{Hidden UV factors}}{W}\underbrace{\alpha + \overset{\text{Factor of interest}}{X}\beta_1}_{\text{Model parameters}} + \underset{\text{error term}}{\epsilon}$$

Normalization

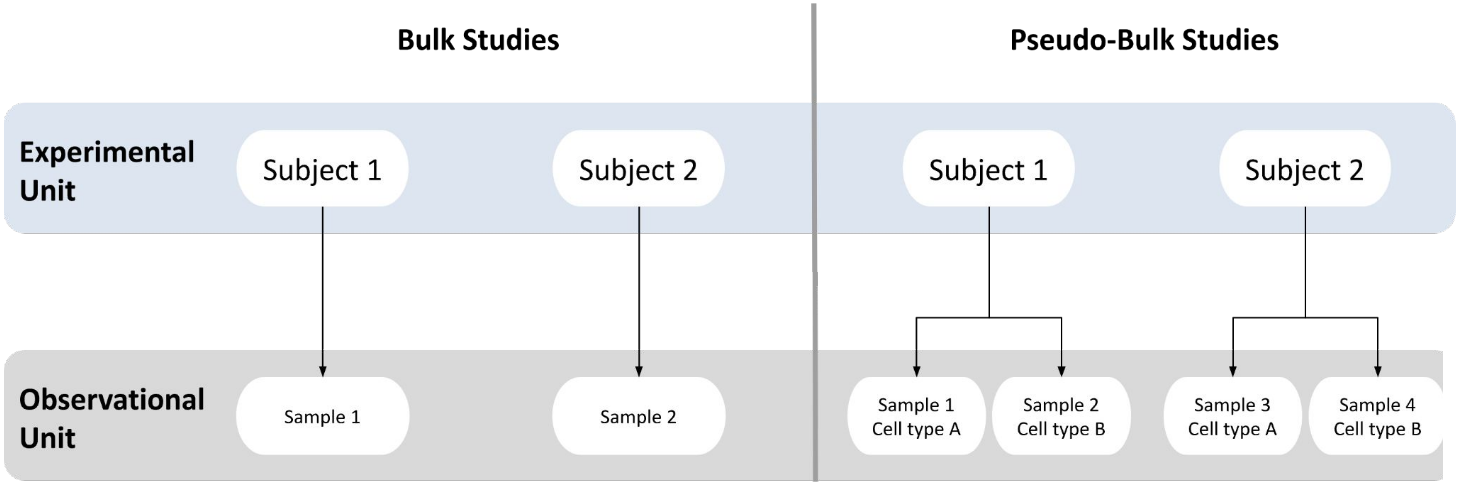
$$\underset{\text{Normalized log-counts}}{\log(Y)^*} = \log(Y) - \widehat{W\alpha}$$

Not all samples are affected in the same way

Each sample has different W values

CONCEPT

Approaches to Pseudo-Bulk data



Type 1

Entire pseudobulk dataset

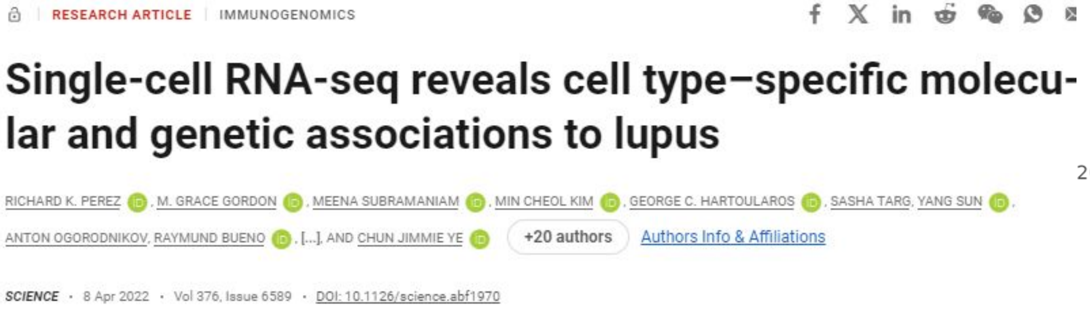
Type 2

Only samples from the same cell type

Type 3

Unwanted factors at a subject level

STUDY CASE



RESEARCH ARTICLE | IMMUNOGENOMICS

Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus

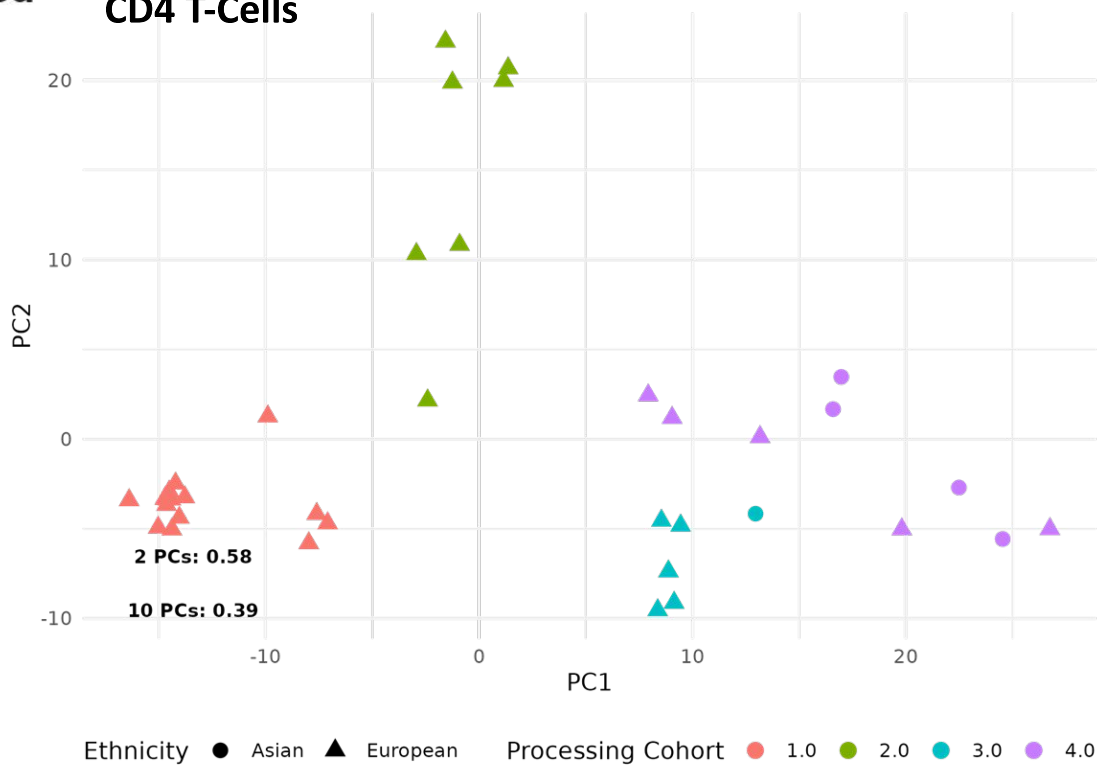
RICHARD K. PEREZ, M. GRACE GORDON, MEENA SUBRAMANIAM, MIN CHEOL KIM, GEORGE C. HARTOULAROS, SASHA TARG, YANG SUN, ANTON OGORODNIKOV, RAYMUND BUENO, [...], AND CHUN JIMMIE YE +20 authors Authors Info & Affiliations

SCIENCE · 8 Apr 2022 · Vol 376, Issue 6589 · DOI: 10.1126/science.abf1970

Healthy controls subsample

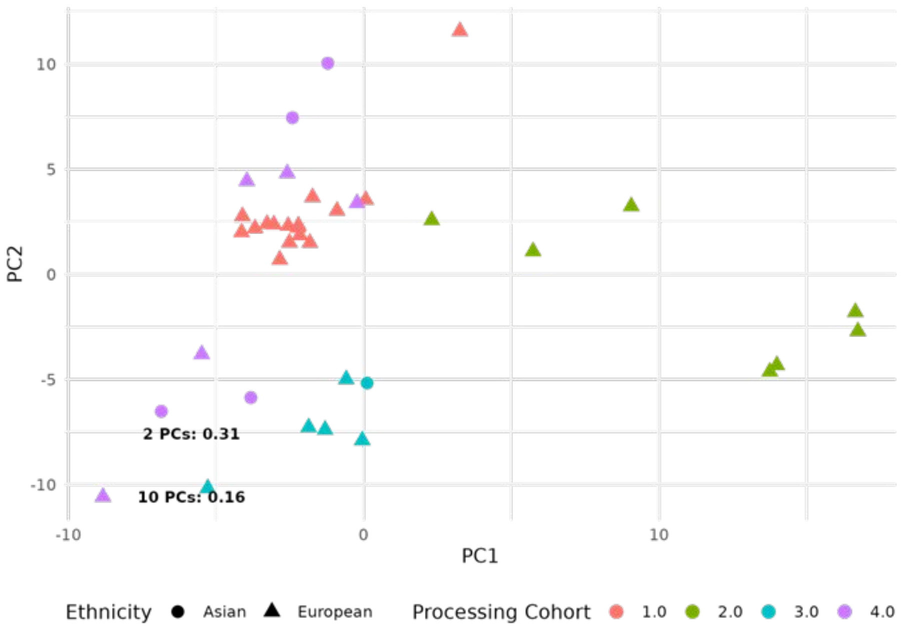
- 38 samples from 30 Individuals
- Biological variables: Age, ethnicity
- Technical variables: Laboratory, processing cohort
- Technical replicates available

CD4 T-Cells

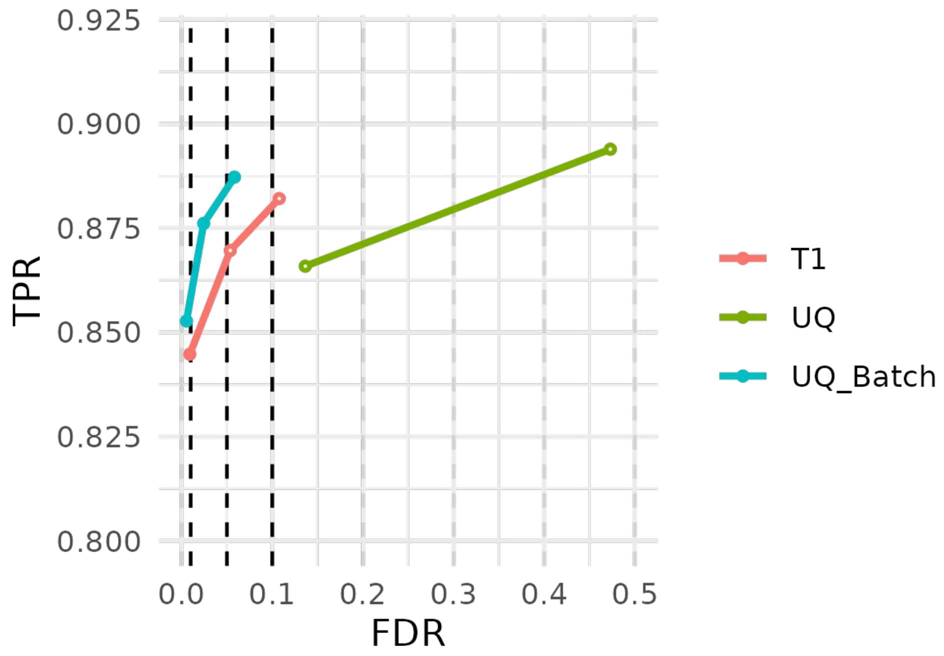


Main source of unwanted variation known: Processing Cohorts

Type 1 approach CD4 T-Cells



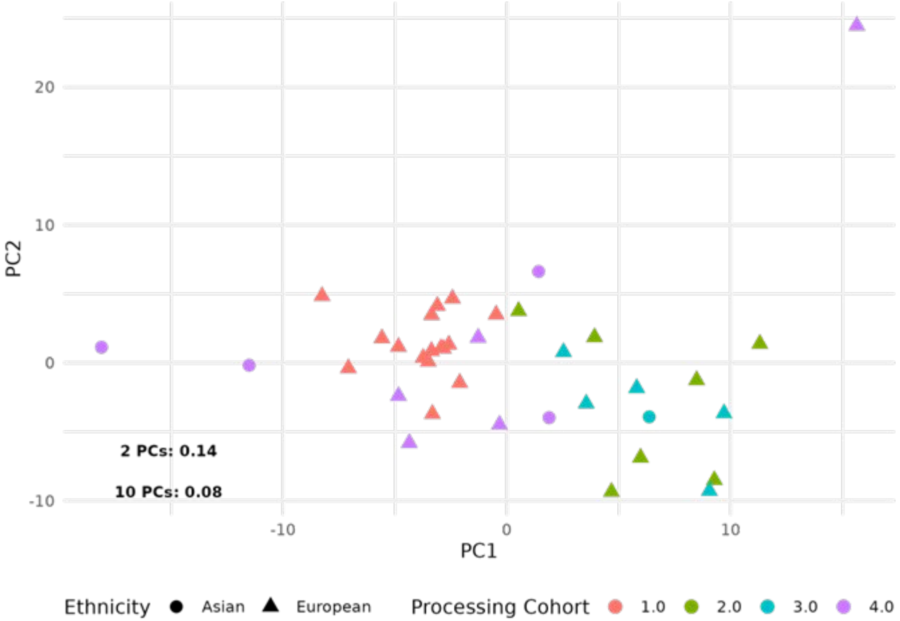
Does not removes the processing cohort effect

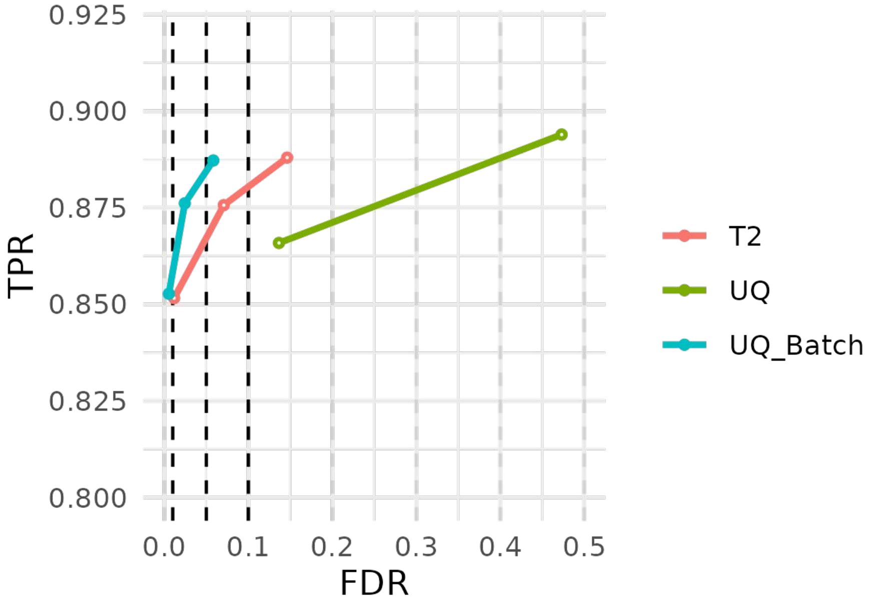


Improves the FDR

RESULTS

Type 2 approach CD4 T-Cells



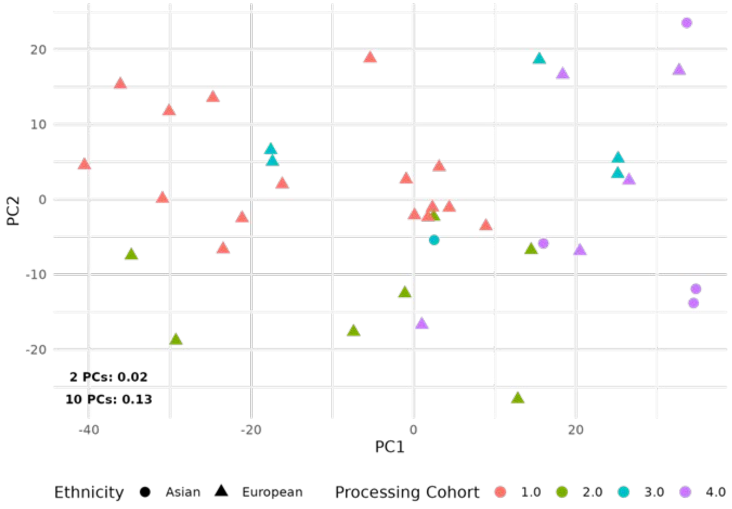


Removes processing cohort effect

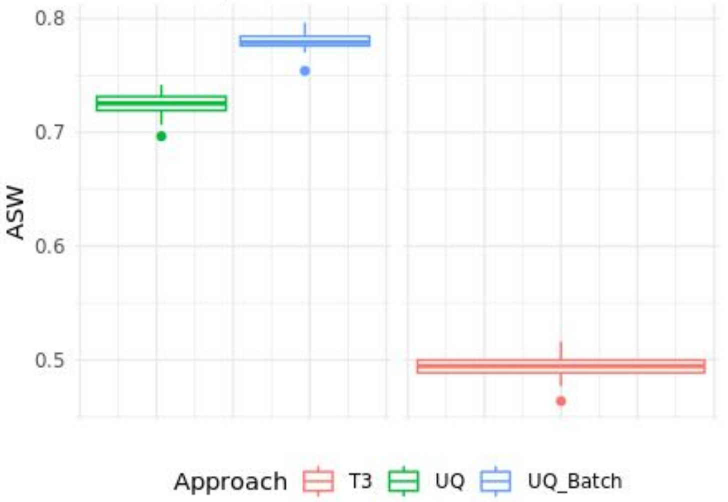
Improves the FDR

STUDY CASE

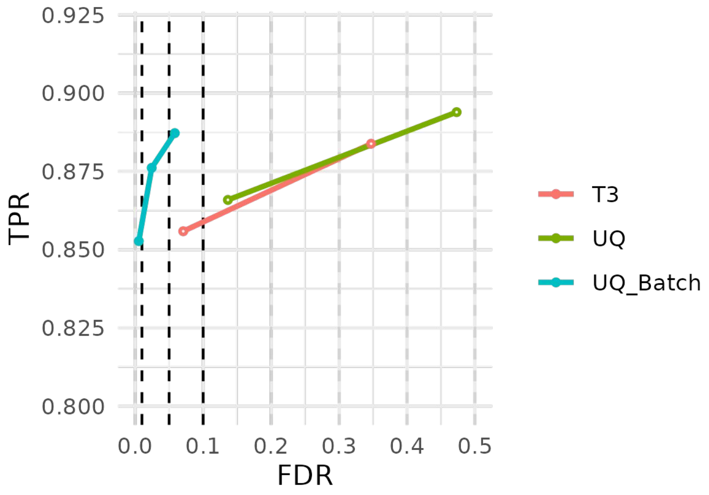
Type 3 approach CD4 T-Cells



Removes the processing cohort effect



Removes biological information



Decreases the TPR

References

- Gagnon-Bartsch, J. A., & Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. Biostatistics, 13(3), 539-552.

- Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., ... & Jaffrézic, F. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in bioinformatics, 14(6), 671-683.

- Gagnon-Bartsch, J. A., Jacob, L., & Speed, T. P. (2013). Removing unwanted variation from high dimensional data with negative controls. Berkeley: Tech Reports from Dep Stat Univ California, 1-112.

- 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(9), 896-902.

References

Peixoto, L., Risso, D., Poplawski, S. G., Wimmer, M. E., Speed, T. P., Wood, M. A., & Abel, T. (2015). How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets. Nucleic acids research, 43(16), 7664-7674.

Wang, J., Zhao, Q., Hastie, T., & Owen, A. B. (2017). Confounder adjustment in multiple hypothesis testing. Annals of statistics, 45(5), 1863.

Molania, R., Gagnon-Bartsch, J. A., Dobrovic, A., & Speed, T. P. (2019). A new normalization for Nanostring nCounter gene expression data. Nucleic acids research, 47(12), 6073-6083.

Deeke, J. M., & Gagnon-Bartsch, J. A. (2020). Stably expressed genes in single-cell RNA sequencing. Journal of Bioinformatics and Computational Biology, 18(01), 2040004.

References

- Gerard, D., & Stephens, M. (2021). Unifying and generalizing methods for removing unwanted variation based on negative controls. Statistica Sinica, 31(3), 1145.

- Salim, A., Molania, R., Wang, J., De Livera, A., Thijssen, R., & Speed, T. P. (2022). RUV-III-NB: normalization of single cell RNA-seq data. Nucleic Acids Research, 50(16), e96-e96.

- Molania, R., Foroutan, M., Gagnon-Bartsch, J. A., Gandolfo, L. C., Jain, A., Sinha, A., ... & Speed, T. P. (2023). Removing unwanted variation from large-scale RNA sequencing data with PRPS. Nature Biotechnology, 41(1), 82-95.

Complementary information

CONCEPT

Removing Unwanted Variation (RUV)

The nuisance technical effects, source of the Unwanted Variation (UV)

- **Different batches**
  plate and time effects
- **Library preparation**

Not all samples are affected in the same way

Each sample has different W values

DEA

Hidden UV factors

Factor of interest

$$\log(Y) = \beta_0 + W\alpha + X\beta_1 + \epsilon$$

Log transformed counts

Intercept

Model parameters

error term

Normalization

$$\log(Y)^* = \log(Y) - \widehat{W\alpha}$$

Normalized log-counts

CONCEPT

The Hidden W factors

$$\log(y_{ng}) = \beta_{0g} + \boxed{w_n}\alpha_g + x_n\beta_{1g} + \epsilon_{ng}$$

Estimation via Exploratory Factor Analysis

Must meet at least one condition

● **Negative control genes (g)**

A set of genes for which the counts are not influenced by the covariates of interest

● **Residuals (r)**

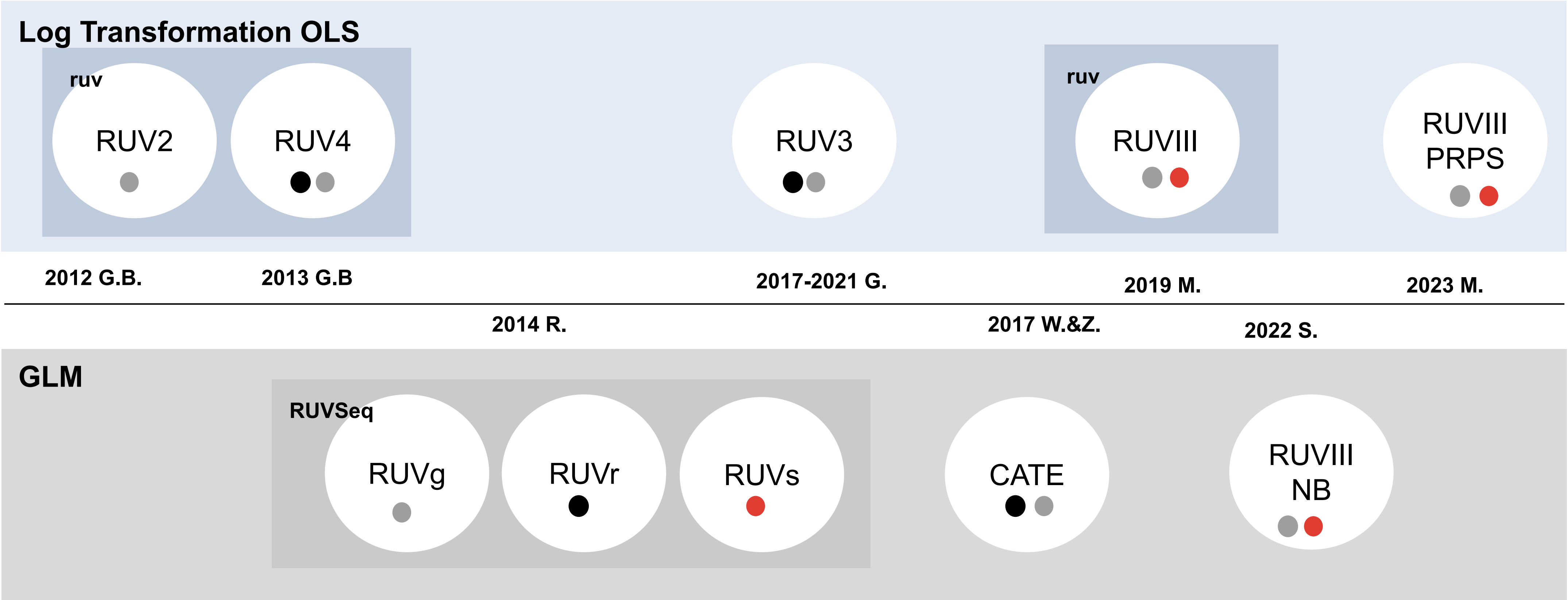
The matrix X of biological covariates is known and is not correlated with the W factors

● **Negative control samples (s)**

A set of replicates for which the biological covariates of interest are constant, and the W factors are uncorrelated with the variables of interest.

THE RUV MULTIVERSE

● nc genes
● residuals
● nc samples



Log Transformation OLS

ruv

RUV2 ●

RUV4 ●●

RUV3 ●●

ruv

RUVIII ●●

RUVIII PRPS ●●

2012 G.B.

2013 G.B

2017-2021 G.

2019 M.

2023 M.

2014 R.

2017 W.&Z.

2022 S.

GLM

RUVSeq

RUVg ●

RUVr ●

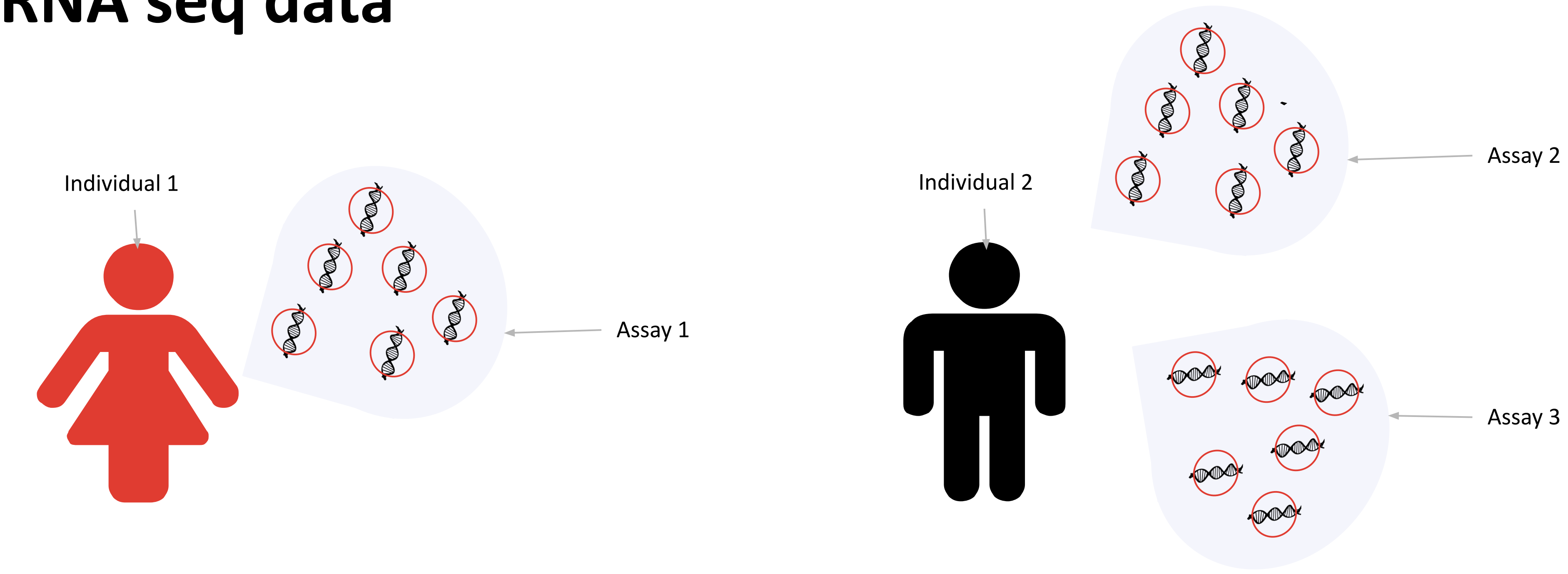
RUVs ●

CATE ●●

RUVIII NB ●●

What is the best way to apply RUV methods to Pseudobulk studies?

Sc-RNA seq data

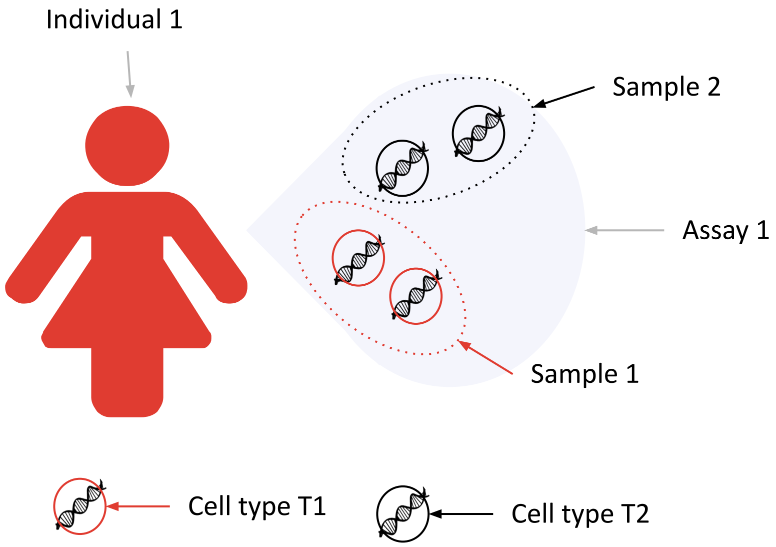


- **Assay:** Collection of single-cell measurements from an individual

- **Technical Replicate**： Multiple assays may be taken from the same individual

THE DATA

ScRNA-seq to Pseudobulk



Pseudobulk sample

Aggregation at Assay x Cell type

Sample 1
cells ∈ Assay 1 & Cell type 1

Cells\Genes	G1	G2	G3
C1	y_{11}	y_{12}	y_{13}
C2	y_{21}	y_{22}	y_{23}
S1	$y_{.1}$	$y_{.2}$	$y_{.3}$

Sample 2
cells ∈ Assay 1 & Cell type 2

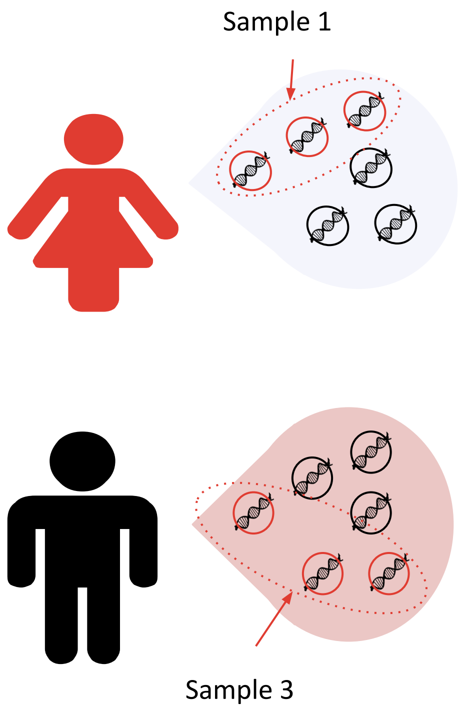
Cells\Genes	G1	G2	G3
C3	y_{31}	y_{32}	y_{33}
C4	y_{41}	y_{42}	y_{43}
S2	$y_{.1}$	$y_{.2}$	$y_{.3}$

Pseudobulk Matrix Y_p

Sample\Genes	G1	G2	G3
S1	$y\square_{11}$	$y\square_{12}$	$y\square_{13}$
S2	$y\square_{21}$	$y\square_{22}$	$y\square_{23}$

THE DATA

Pseudobulk matrices



Pseudobulk counts matrix Y_p

Sample	G1	G2	G3
S1	y□$_{11}$	y□$_{12}$	y□$_{13}$
S2	y□$_{21}$	y□$_{22}$	y□$_{23}$
S3	y□$_{31}$	y□$_{32}$	y□$_{33}$
S4	y□$_{41}$	y□$_{42}$	y□$_{43}$

Pseudobulk counts matrix aggregated only by assay Y_N

Assay	G1	G2	G3
N1	y□$_{11}$	y□$_{12}$	y□$_{13}$
N2	y□$_{21}$	y□$_{22}$	y□$_{23}$

Pseudobulk counts matrix from Samples with cell type t=T1 Y_t

Sample	G1	G2	G3
S1	y□$_{11}$	y□$_{12}$	y□$_{13}$
S3	y□$_{21}$	y□$_{22}$	y□$_{23}$

CONCEPT

Finding W in the Pseudobulk context

We compare 3 approaches to estimate W

Type 1 $E(\log(Y_p)|W, Z) = W\alpha + Z\gamma$

Using the entire Pseudobulk dataset, and including the interactions between cell types and factors of interest

→ Assumes independence between samples, which is not true.

Type 2 $E(\log(Y_t)|W_t, X) = W_t\alpha + X\beta$

Using only samples from the same cell type t

→ Does not make use of information available from other samples

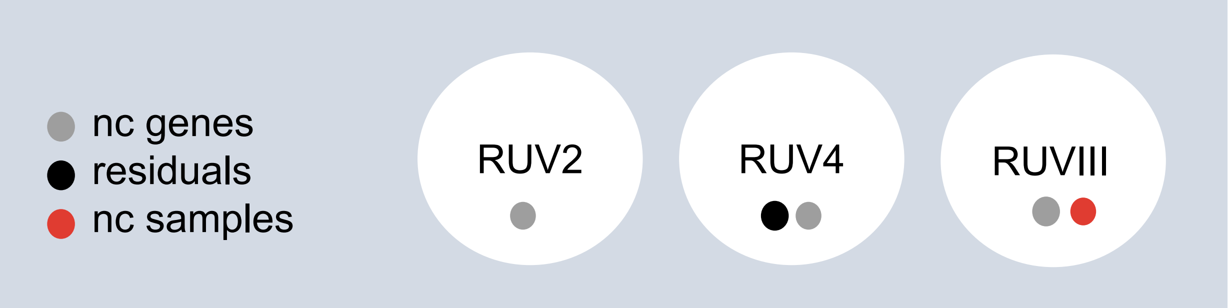
Type 3 $E(\log(Y_N)|W_N, X) = W_N\alpha + X\beta$

Using the entire dataset aggregated at an assay level. Cell types are not considered

→ Assumes all samples from the same assay have the same W values, i.e., there is no cell type specific technical variation.

CONCEPT

OLS methods



⚪ nc genes
⚫ residuals
🔴 nc samples

RUV2 ⚪

RUV4 ⚫ ⚪

RUVIII ⚪ 🔴

CONCEPT

Methods with no RUV normalisation

$$E(\log(Y_t)|X) = X\beta$$

- **Upper-quartile normalization (UQ)**
  - X only has the treatment information

- **UQ Batch**
  - X has the treatment information and the batch information

STUDY CASE

RESEARCH ARTICLE | IMMUNOGENOMICS

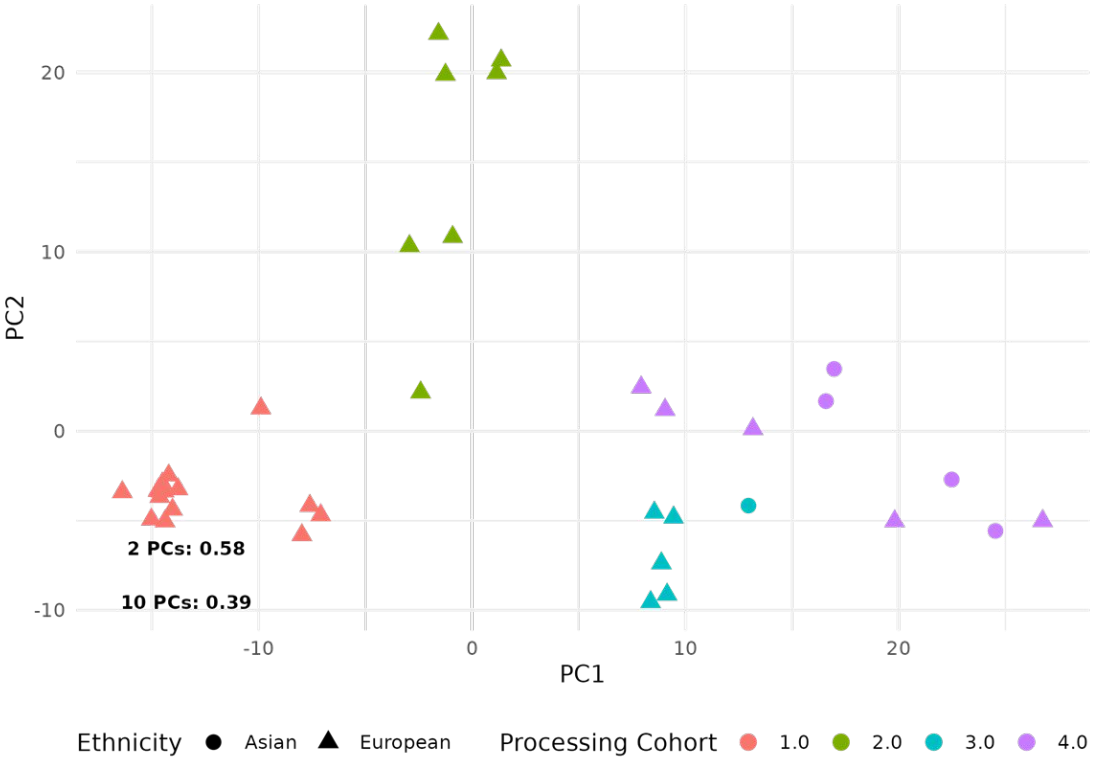
Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus

RICHARD K. PEREZ, M. GRACE GORDON, MEENA SUBRAMANIAM, MIN CHEOL KIM, GEORGE C. HARTOULAROS, SASHA TARG, YANG SUN, ANTON OGORODNIKOV, RAYMUND BUENO, [...], AND CHUN JIMMIE YE +20 authors Authors Info & Affiliations

SCIENCE • 8 Apr 2022 • Vol 376, Issue 6589 • DOI: 10.1126/science.abf1970

Healthy controls subsample

- 38 scRNA-Seq assays from 30 Individuals
- Ages from 25 to 28
- 2 ethnicities
- 3 laboratories
- Samples processed in 4 cohorts
- 8 cell types
- Technical replicates available

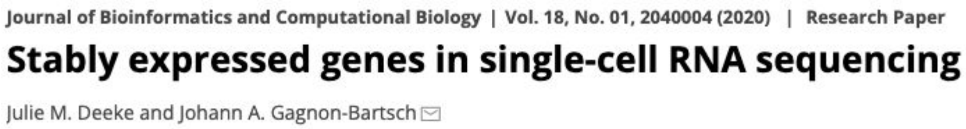


Main source of unwanted variation known: Processing Cohorts

STUDY CASE

Assumptions

Negative control genes (g)



Journal of Bioinformatics and Computational Biology | Vol. 18, No. 01, 2040004 (2020) | Research Paper

Stably expressed genes in single-cell RNA sequencing

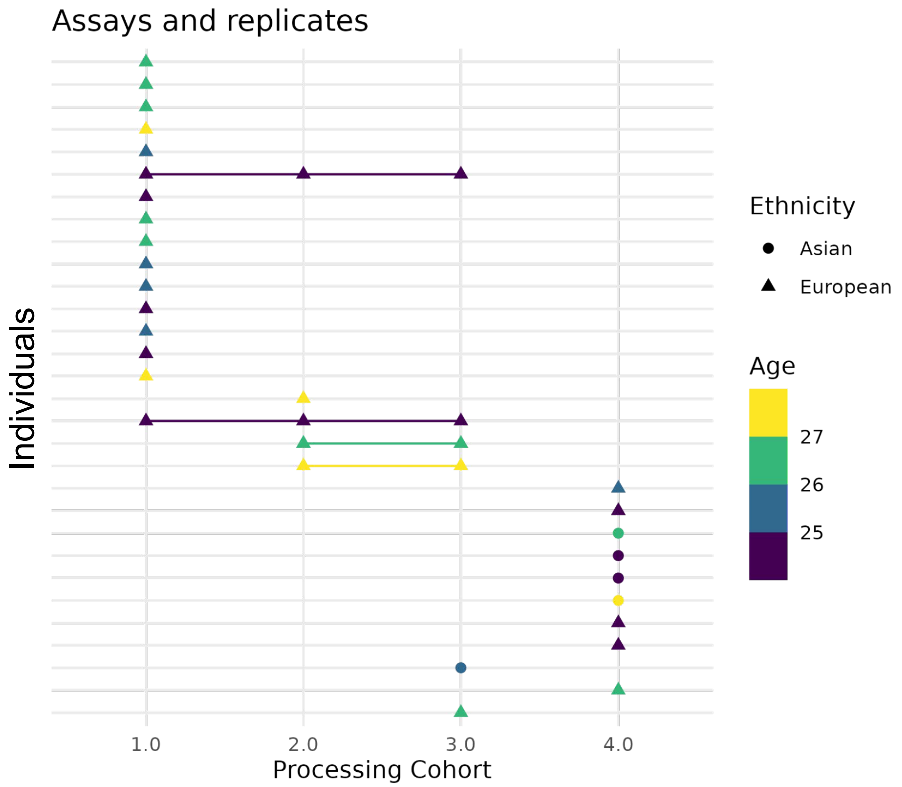
Julie M. Deeke and Johann A. Gagnon-Bartsch

Residuals (r)

The factor of interest is a mock treatment, randomly assigned and independent from the Processing Cohorts.

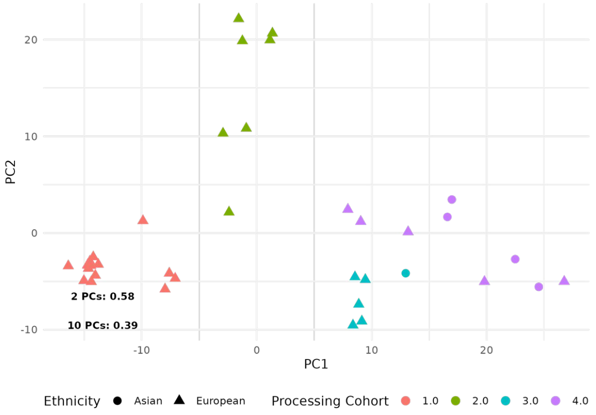
Negative control samples (s)

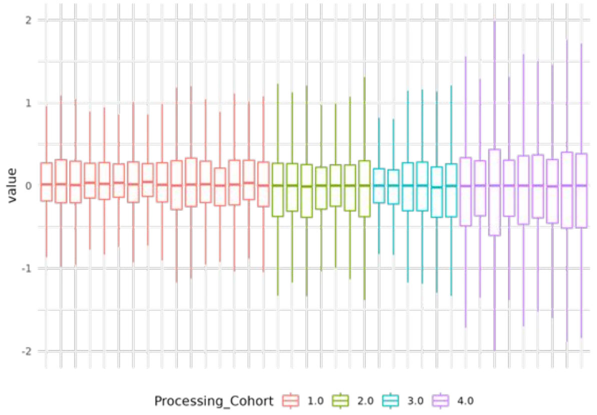
Technical replicates are available, but not present in the processing cohort 4

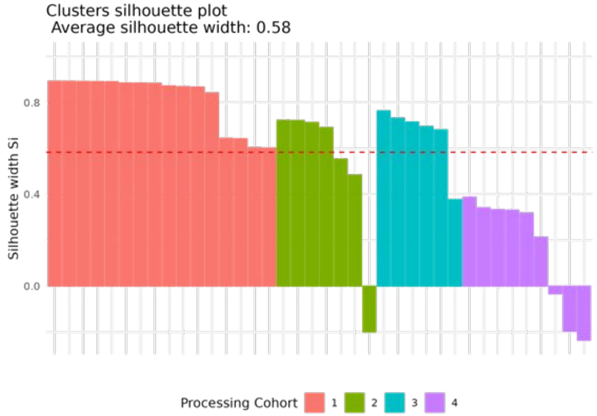


Diagnostics over the normalized matrices

$$\log(Y_t)^* = \log(Y_t) - \widehat{W_t\alpha}$$







PCA plot

3 first PCAs colored by the biological and technical variables

Relative Log Expression (RLE)

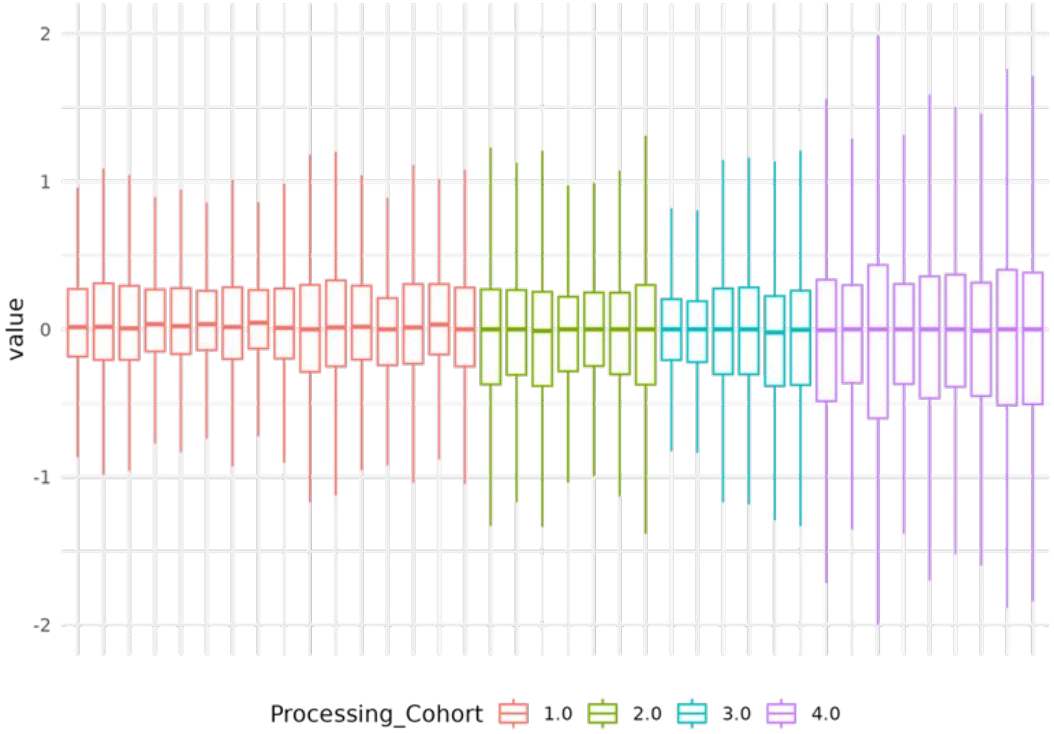
Differences between the log-counts and its respective gene median. Summarised in a boxplot by sample

Average Silhouette Width (ASW)

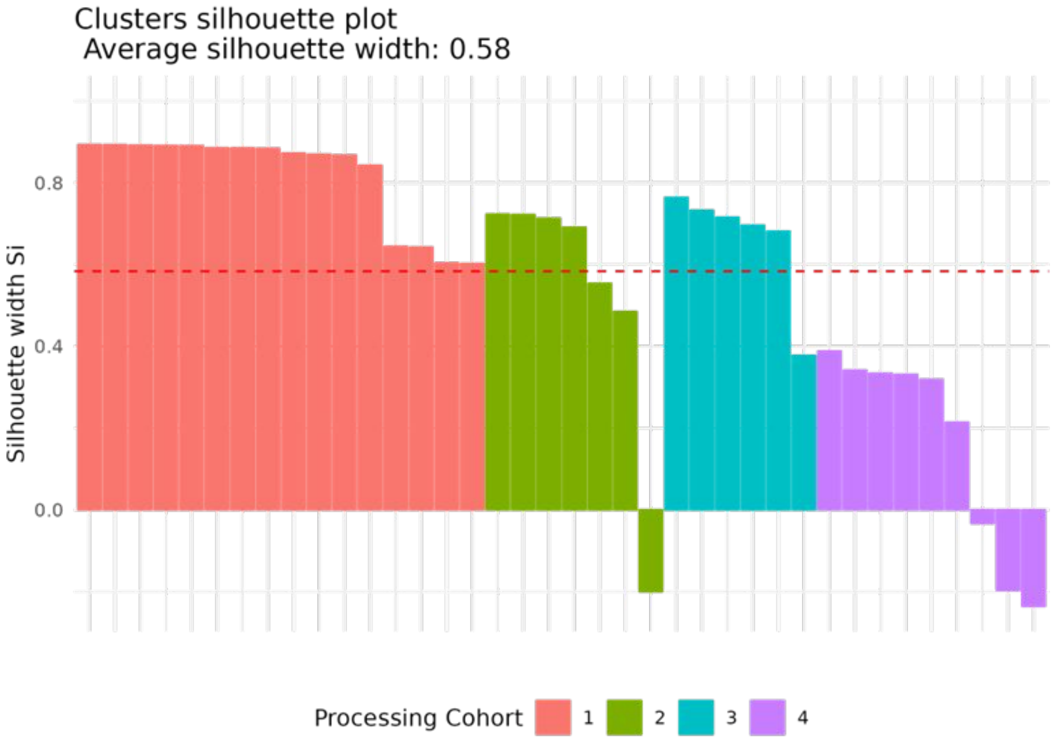
Average of the silhouettes of each sample using the known biological and technical factors as clustering labels.

STUDY CASE

CD4 T-Cells

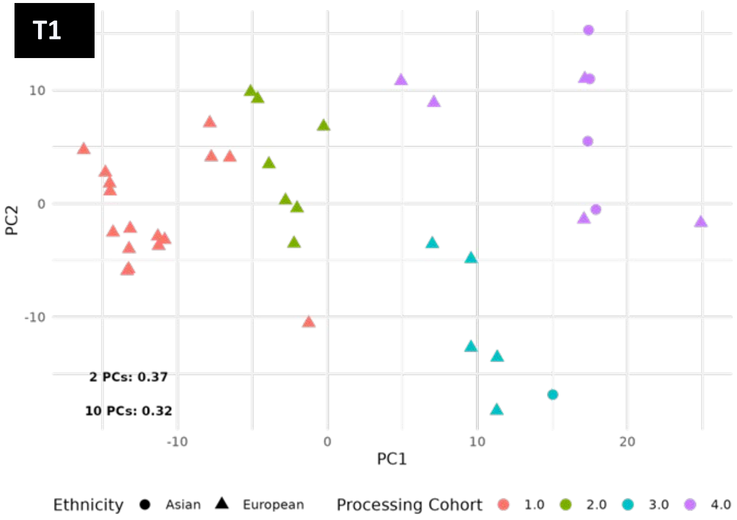


Bigger differences in Processing Cohort 4



High Silhouette widths in the first 3 cohorts

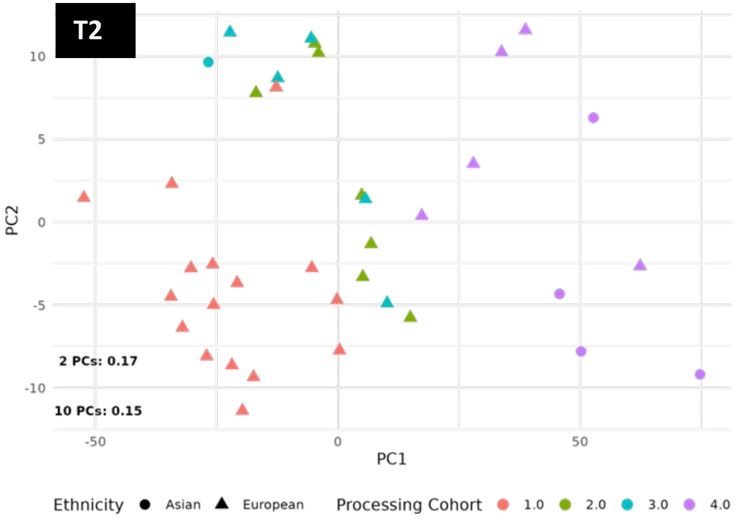
PCA and ASW CD4 T-Cells RUVIII



T1

2 PCs: 0.37

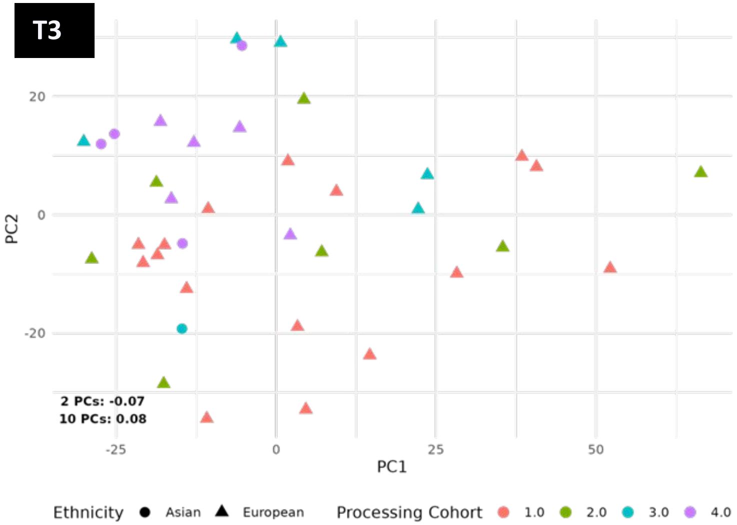
10 PCs: 0.32



T2

2 PCs: 0.17

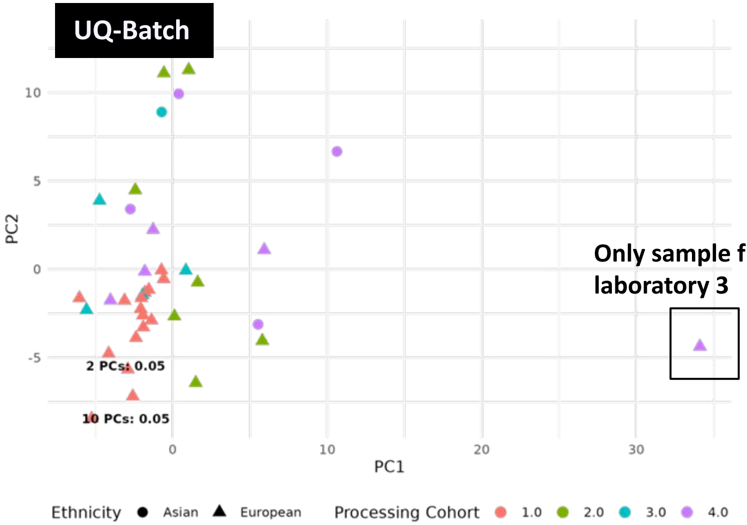
10 PCs: 0.15



T3

2 PCs: -0.07

10 PCs: 0.08



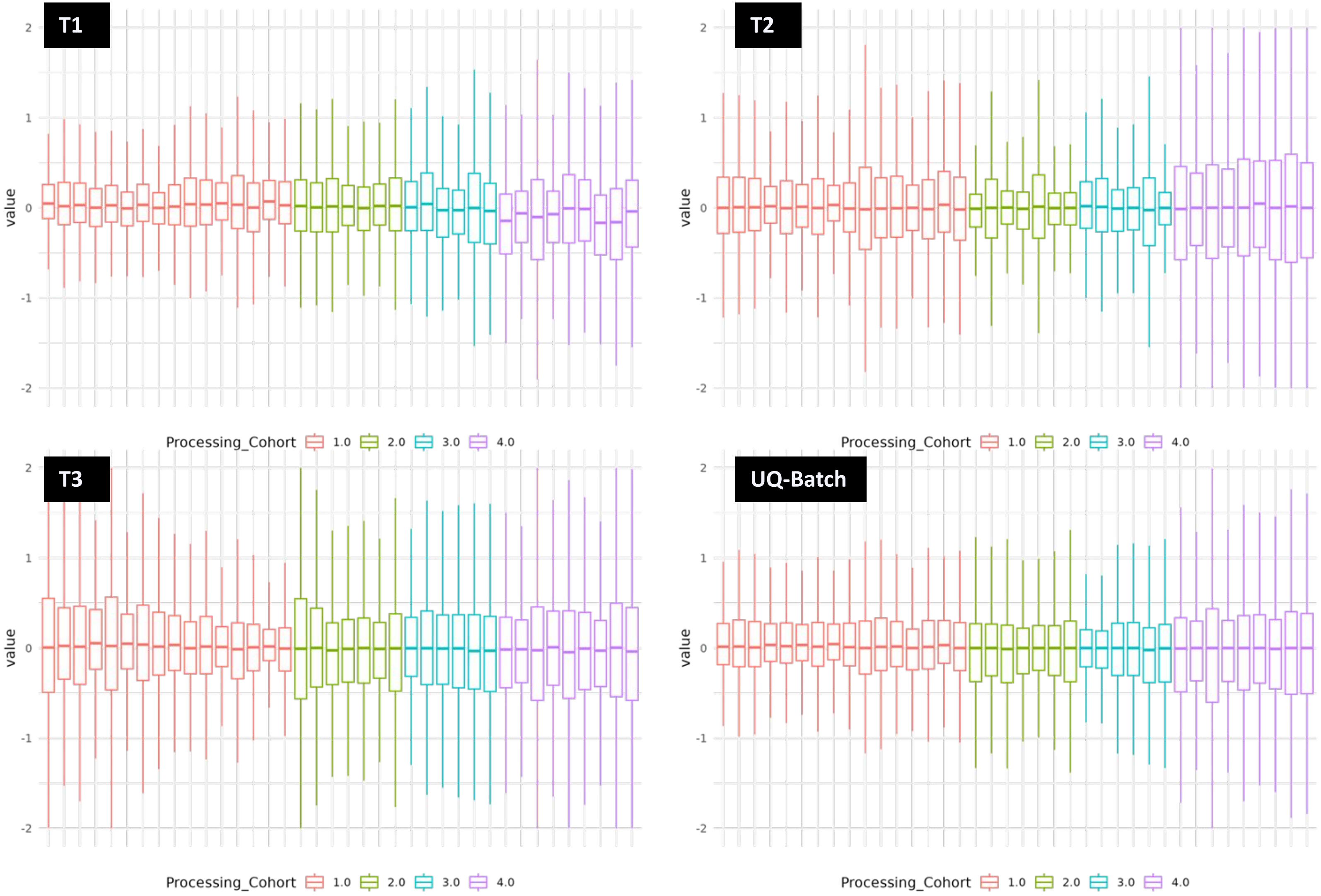
UQ-Batch

2 PCs: 0.05

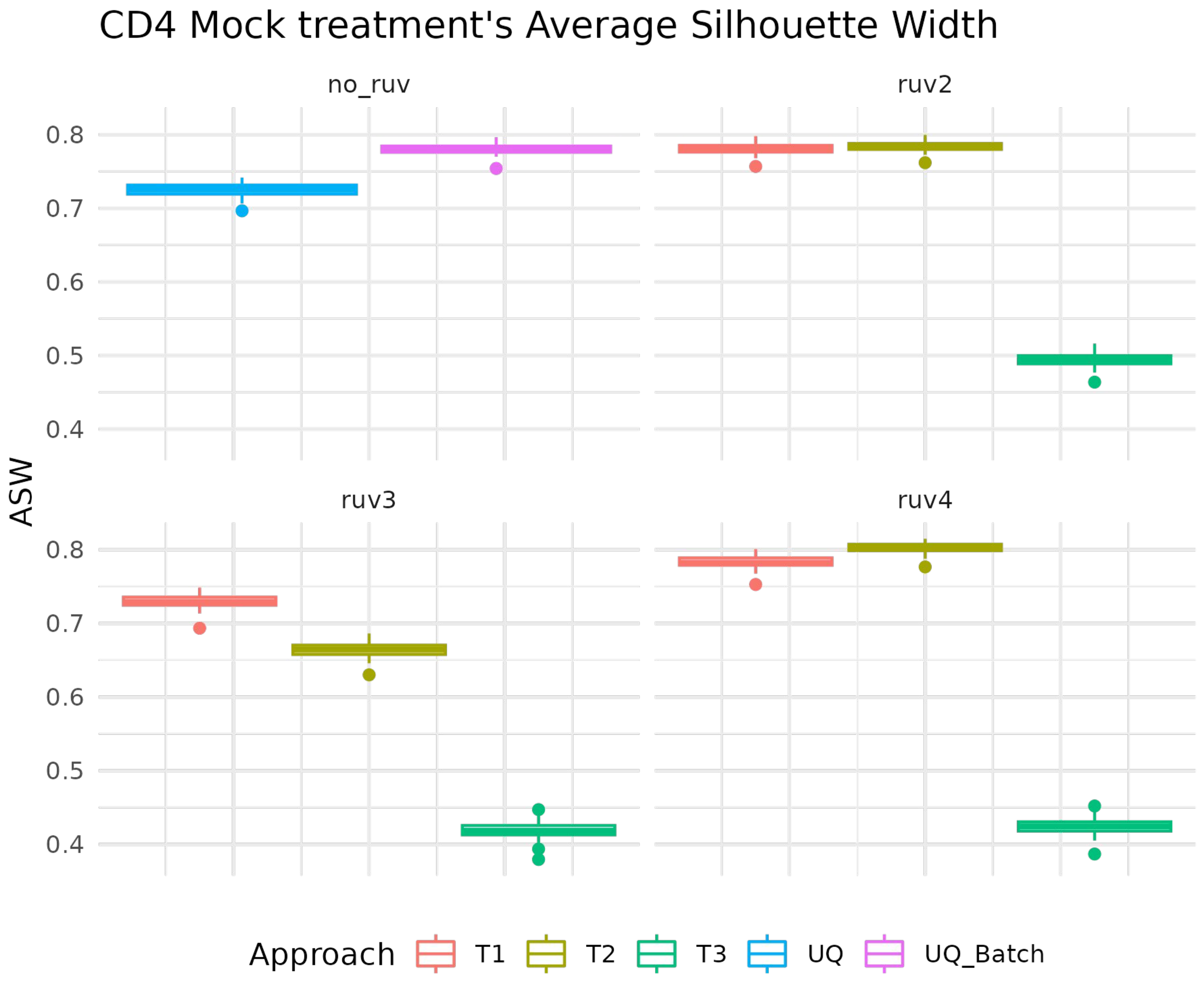
10 PCs: 0.05

Only sample from laboratory 3

RLE Plots
CD4 T-Cells
RUVIII

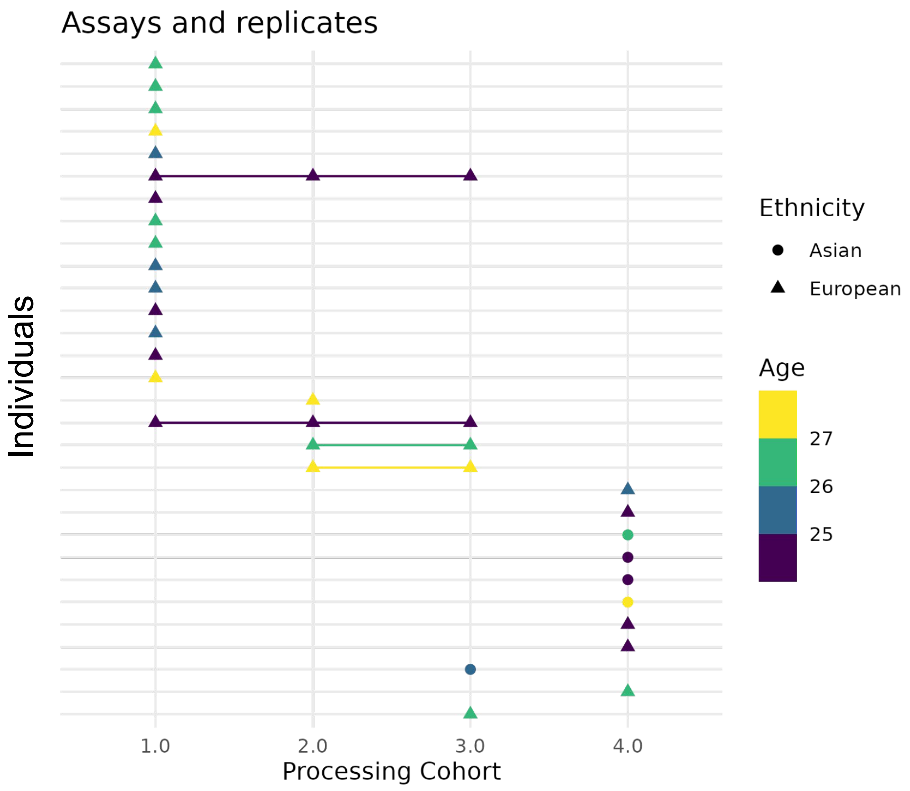


ASW boxplots CD4 T-Cells



STUDY CASE

Differential Expression Analysis: Mock treatment



Design A

Treatment randomly assign to Individuals with equal probabilities.

35% of Individuals with mock treatment **A** belong to PC4
20% of Individuals with mock treatment B belong to PC4

Design B

Increased probability of receiving treatment A (to 90%) for Individuals from Processing Cohort 4.

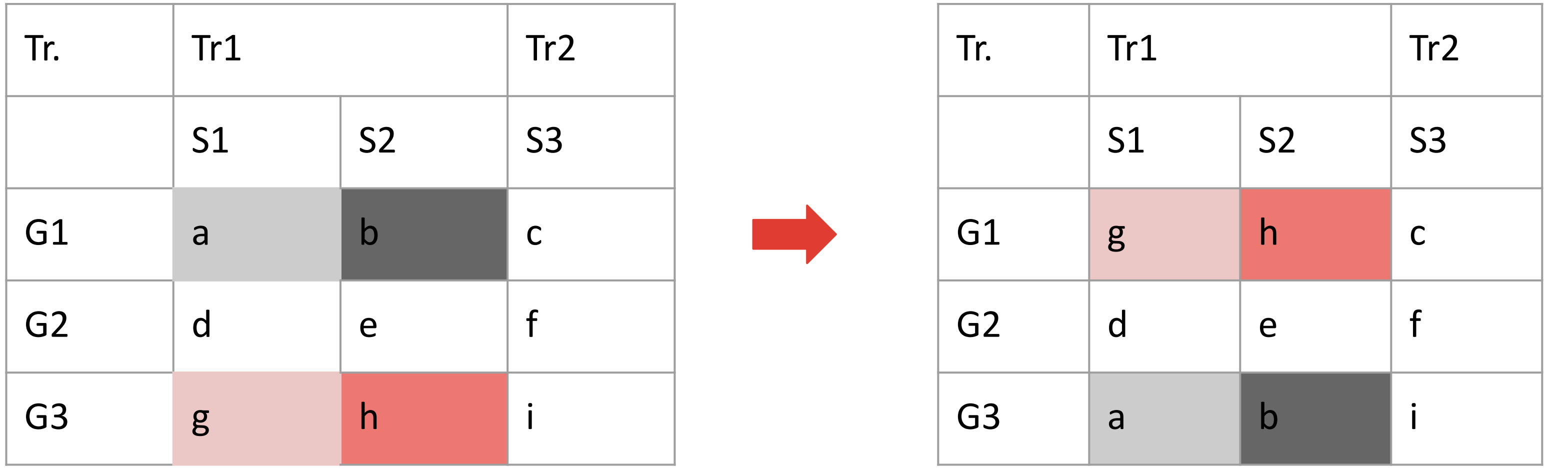
47% of Individuals with mock treatment **A** belong to PC4
13% of Individuals with mock treatment B belong to PC4

SIMULATIONS

Differential Gene Expression Simulation

Randomly scramble 10% of the features (genes) within one experimental group to generate differential expression.

R pkg swapper

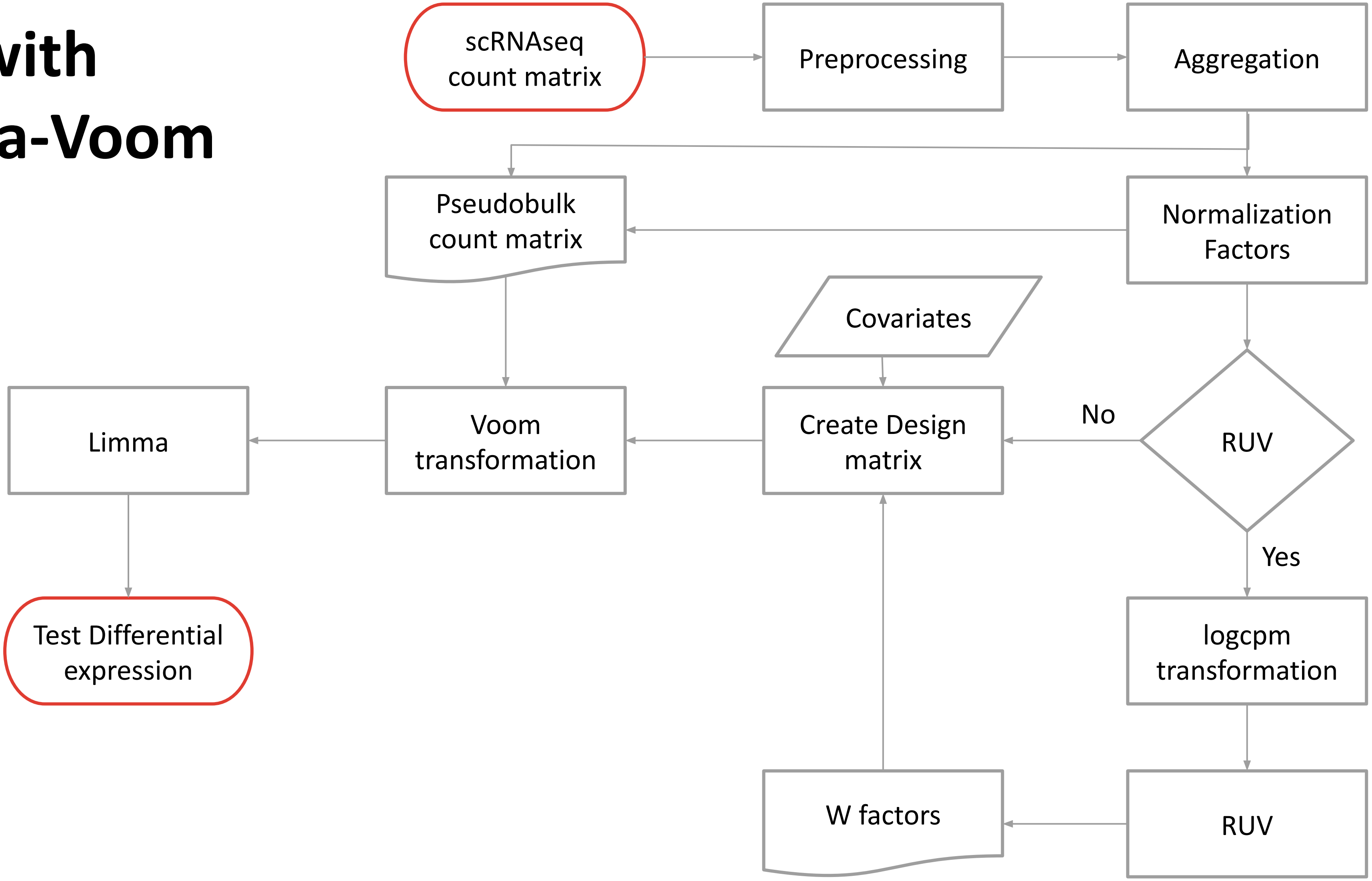


Tr.	Tr1		Tr2
	S1	S2	S3
G1	a	b	c
G2	d	e	f
G3	g	h	i

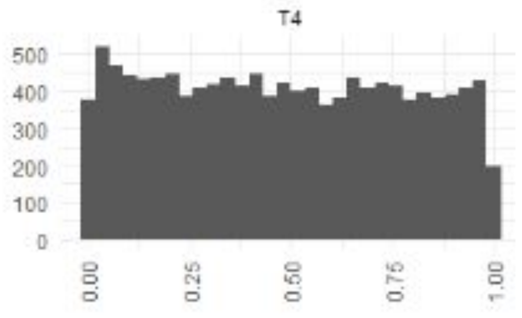
Tr.	Tr1		Tr2
	S1	S2	S3
G1	g	h	c
G2	d	e	f
G3	a	b	i

SIMULATIONS

DEA with Limma-Voom

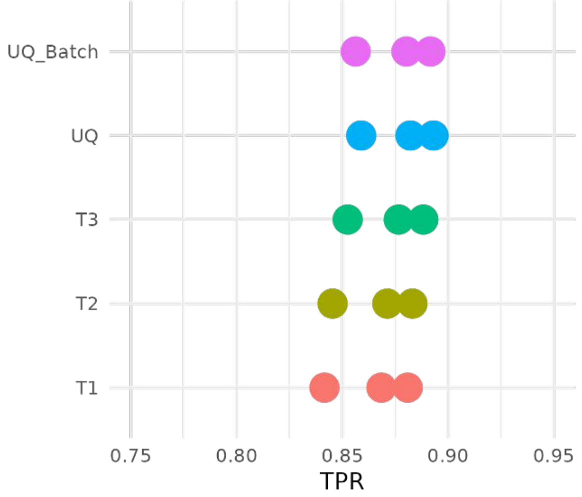


Diagnostics for the Differential Expression Analysis



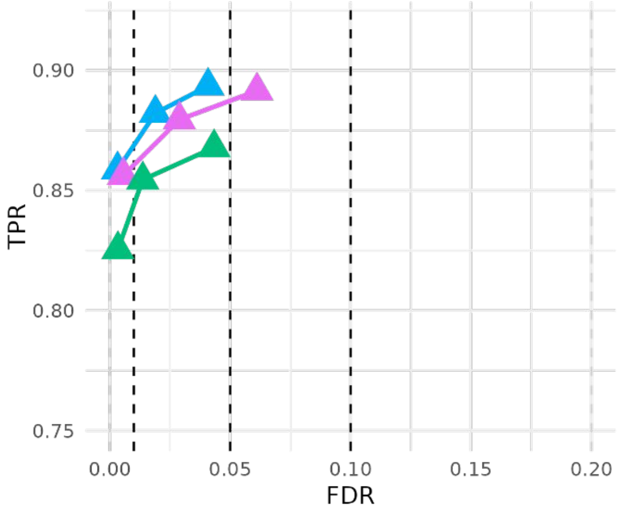
P-values histogram

Histogram of p-values under no differential expression



True Positive Rate TPR

Percentage of true positives at 3 FDR nominal values: 0.1, 0.05 and 0.01

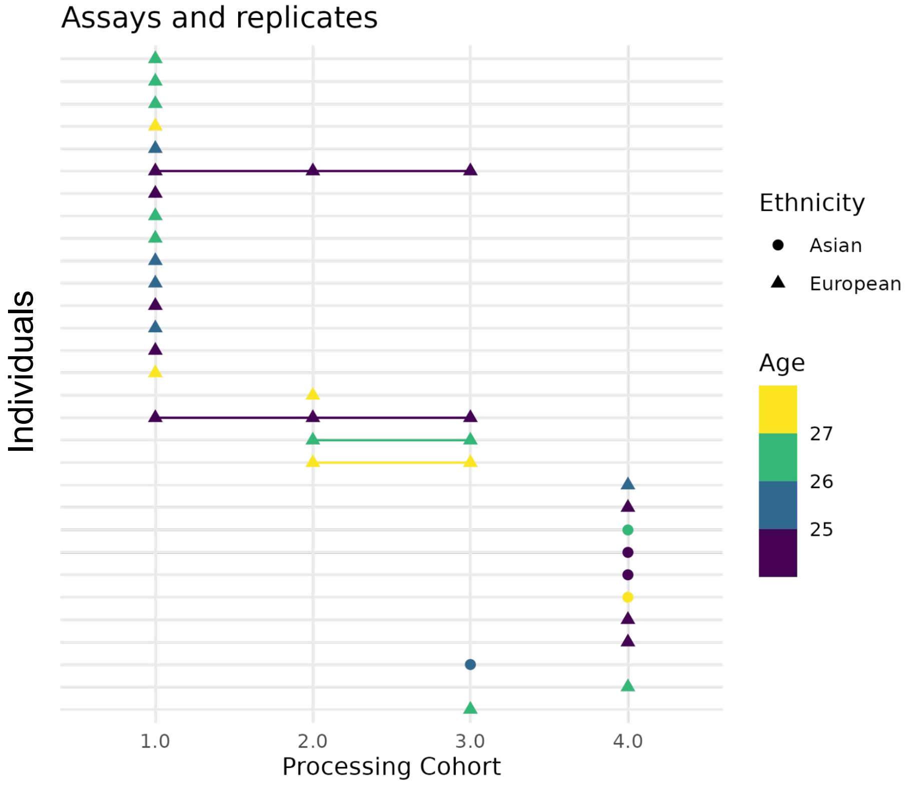


False Discovery Rate vs TPR

Percentage of false positives at 3 FDR nominal values: 0.1, 0.05 and 0.01 vs the TPR

STUDY CASE

Differential Expression Analysis: Mock treatment



Design A

Treatment randomly assign to Individuals with equal probabilities.

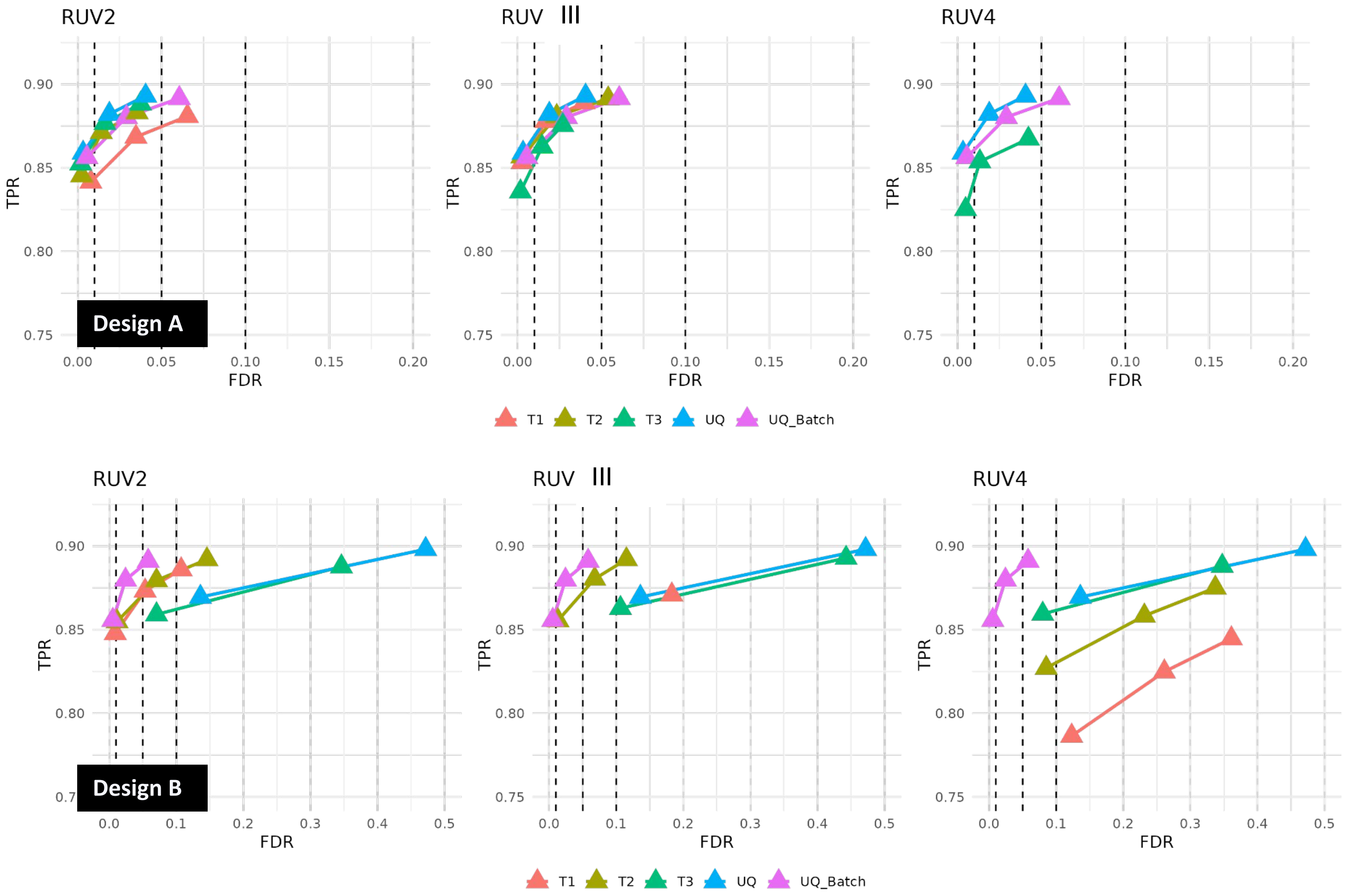
35% of Individuals with mock treatment **A** belong to PC4
20% of Individuals with mock treatment B belong to PC4

Design B

Increased probability of receiving treatment A (to 90%) for Individuals from Processing Cohort 4.

47% of Individuals with mock treatment **A** belong to PC4
13% of Individuals with mock treatment B belong to PC4

FDR vs TPR
CD4 T-Cells



Discussion

- RUV2 has its best performance using the T2 Approach on the design A
- RUV4 has an overall poor performance.
- Lack of replicates in Processing Cohort 4 affects the RUVIII results.

D.A: Design A

D.B: Design B

++: Better than UQ+Batch

+: Better than UQ

-: Worse than UQ

- -: Worst

	RUV2	RUVIII	RUV4
T1	ASW + RLE ++ D.A TPR - D.A FPR D.B TPR D.B FPR ++	ASW RLE + D.A TPR - D.A FPR D.B TPR D.B FPR -	ASW + RLE D.A TPR D.A FPR - - D.B TPR - D.B FPR -
T2	ASW ++ RLE ++ D.A TPR - D.A FPR D.B TPR D.B FPR ++	ASW + RLE - D.A TPR - D.A FPR D.B TPR D.B FPR ++	ASW ++ RLE D.A TPR D.A FPR - - D.B TPR - D.B FPR -
T3	ASW + RLE - D.A TPR - D.A FPR D.B TPR D.B FPR -	ASW ++ RLE - D.A TPR - D.A FPR D.B TPR D.B FPR -	ASW + RLE - D.A TPR - D.A FPR - D.B TPR - D.B FPR -