Removing unwanted variation from large-scale RNAseq data by leveraging pseudo replicates of pseudo samples

Ramyar Molania

Speed Lab Meeting

Bioinformatics Division
Walter and Eliza Hall Institute of Medical Research

07/10/2020

When you recommend to use better normalization methods...

Outline

- Sources of unwanted variation in TCGA RNAseq datasets
- Pseudo replicates of pseudo samples approach (PRPS)
- Unwanted variation in the TCGA BRCA, READ, COAD
- How to use RUV-III with PRPS approach to remove unwanted variation from RNAseq data
- Summary

TCGA biospecimens processing overview

To generate the data for each cancer type, samples were:

- collected from multiple institutions (Tissue Source Sites (TSS))
  - In total 831 institutions
- shipped to genomics centers (Centers) over a span of 6 years (2009:2014)
  - In total 31 genomics centers
- allocated to different sequencing batches (Plates) for profiling
  - 96-well plates

Examples

TCGA Breast Invasive Carcinoma (BRCA) RNA-Seq data:

- 1100 samples
- 40 TSS
- 5 years (2010:2014)
- 39 plates
- 2 different flow cell chemistries (personal communication from TCGA)

TCGA Kidney Chromophobe (KICH) RNA-Seq data:

- 80 samples
- 6 TSS
- 1 year (2014)
- 1 plate

Source: https://docs.gdc.cancer.gov/Encyclopedia

Sources of unwanted variation in TCGA RNA-Seq datasets

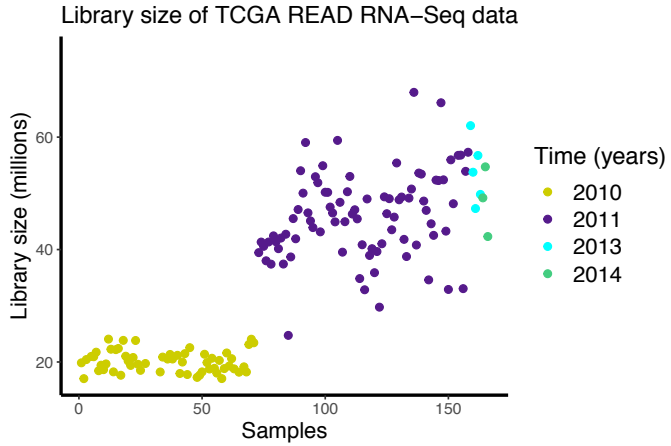
What is unwanted variation?

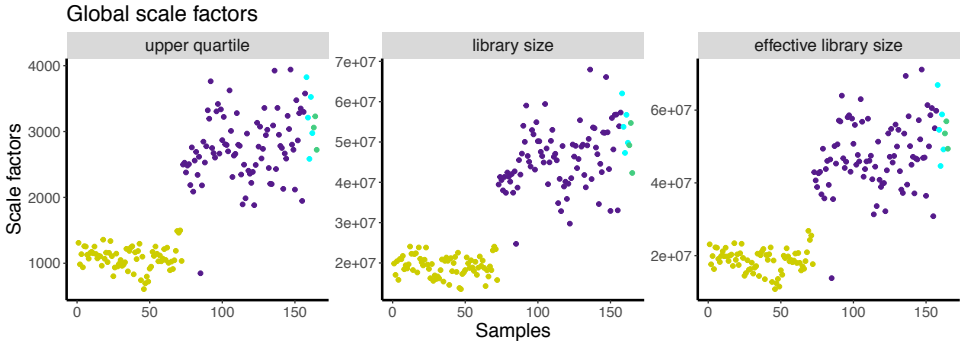
Any variation in the data that should have no effects on down-stream analysis of interest.

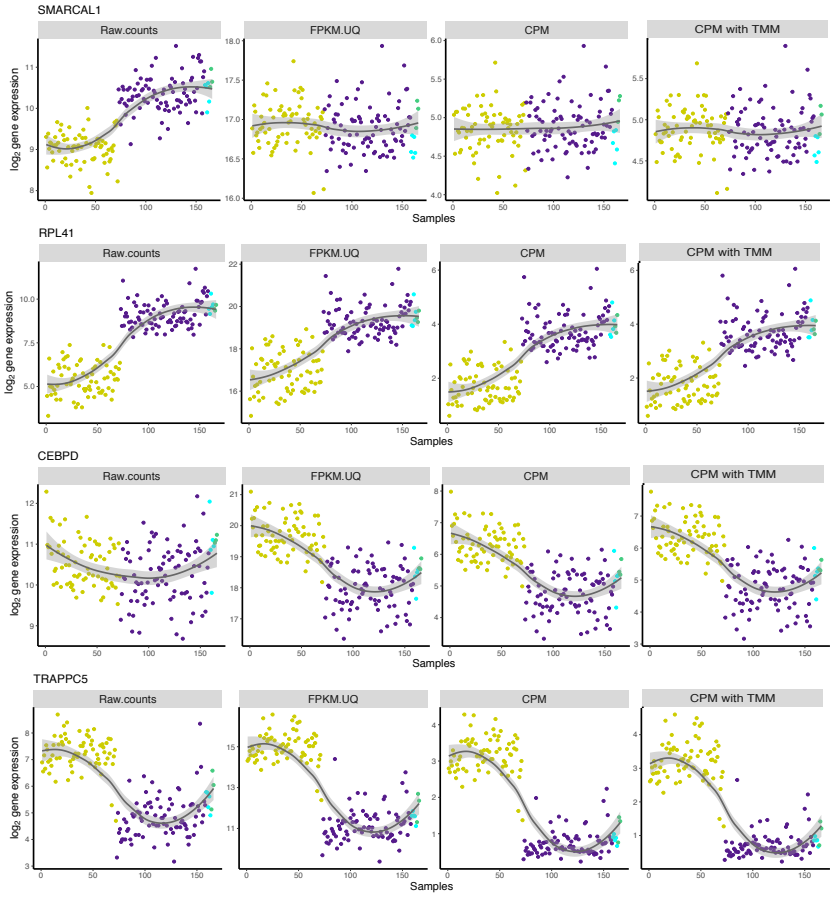
Sources of unwanted variation

- Library size (sequencing depth)
- Tumor purity
- Batch effects
  - different reagents, plates effects, time effects
- Other unknown source(s)

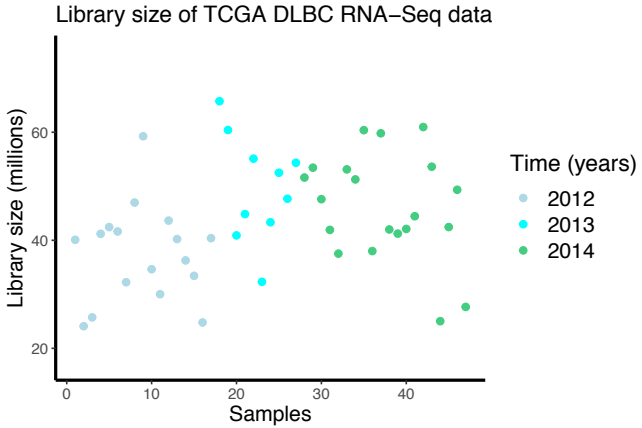
Global scale factors are insufficient to adjust for library size

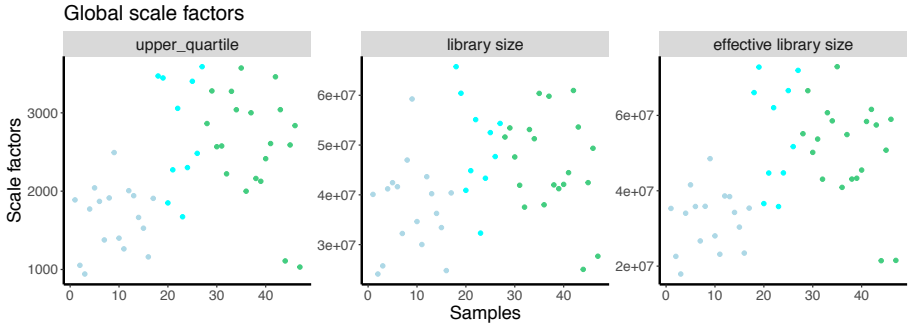


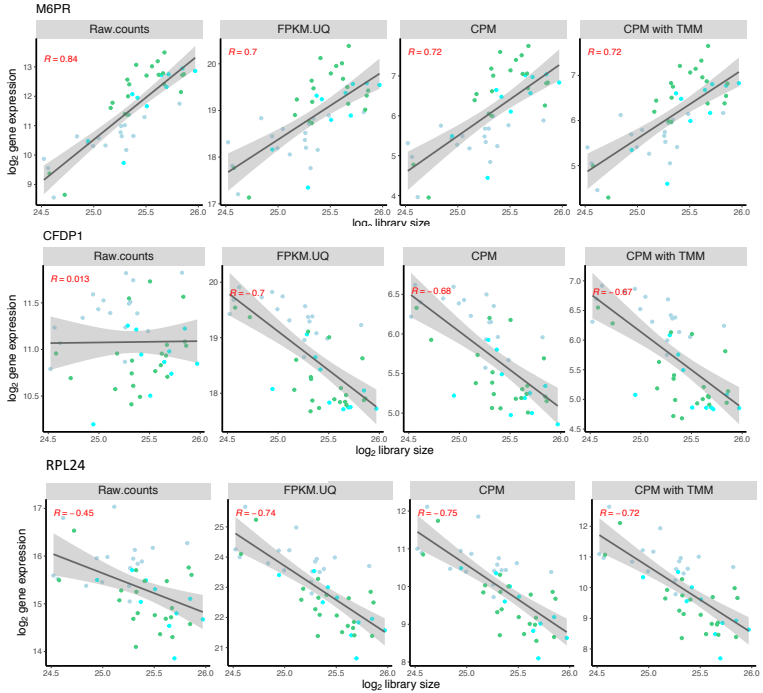




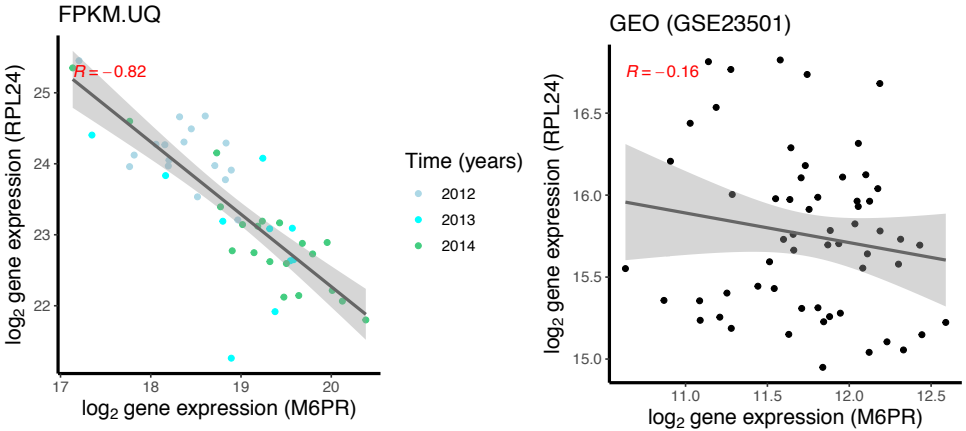
Global scale factors are insufficient to adjust for library size, cont.

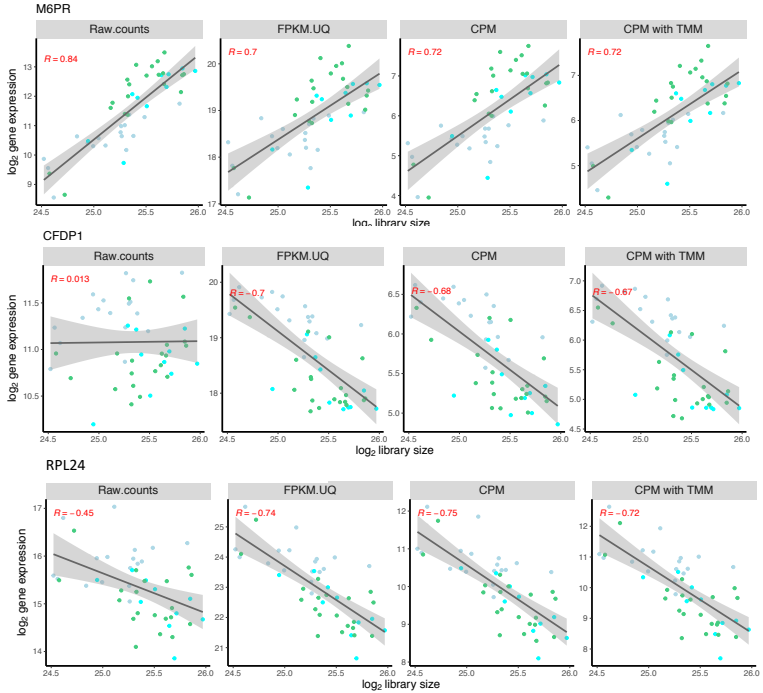






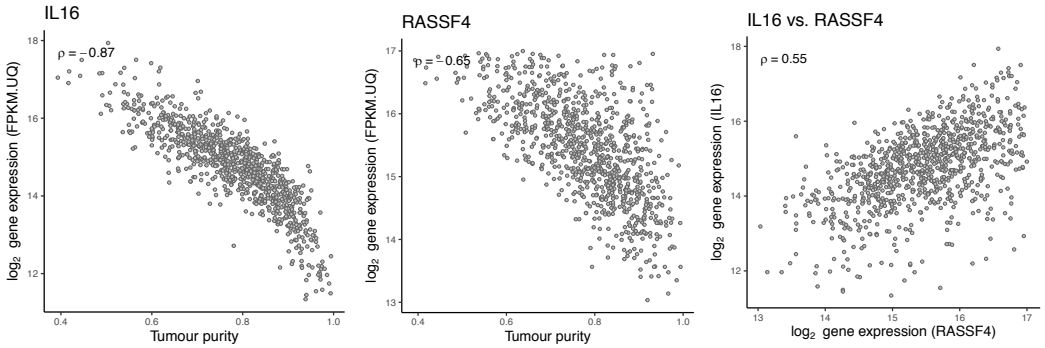
Unwanted variation in the TCGA DLBC RNA-Seq data



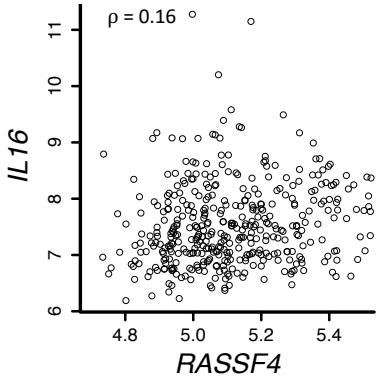


Tumor purity compromises gene co-expression analysis

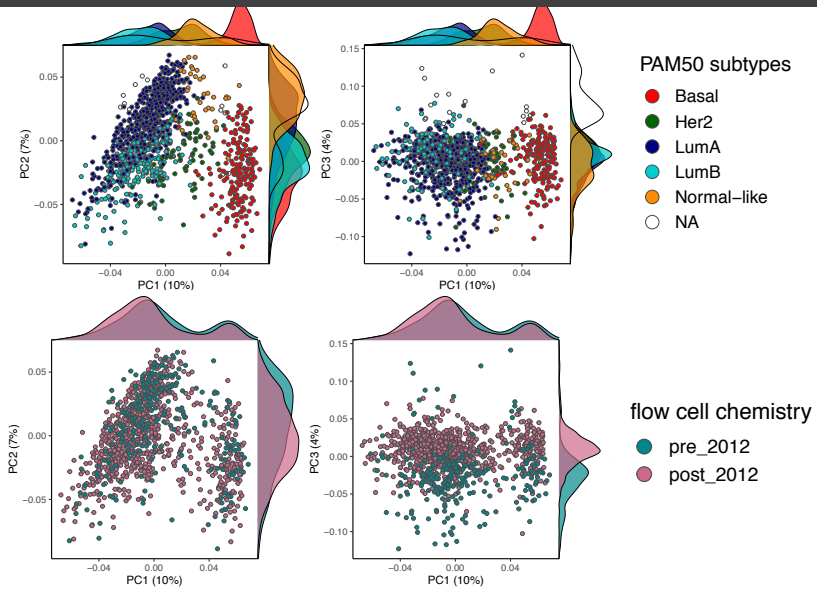
TCGA beast cancer RNA-Seq (FPKM.UQ)

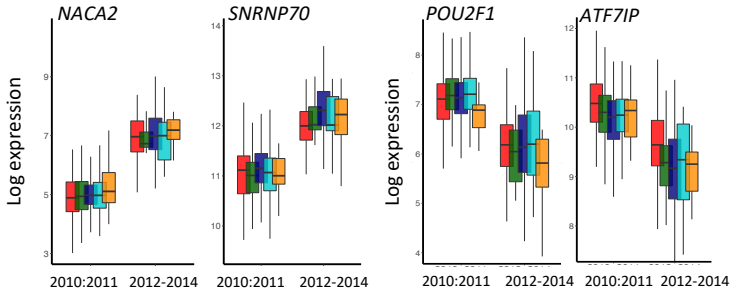


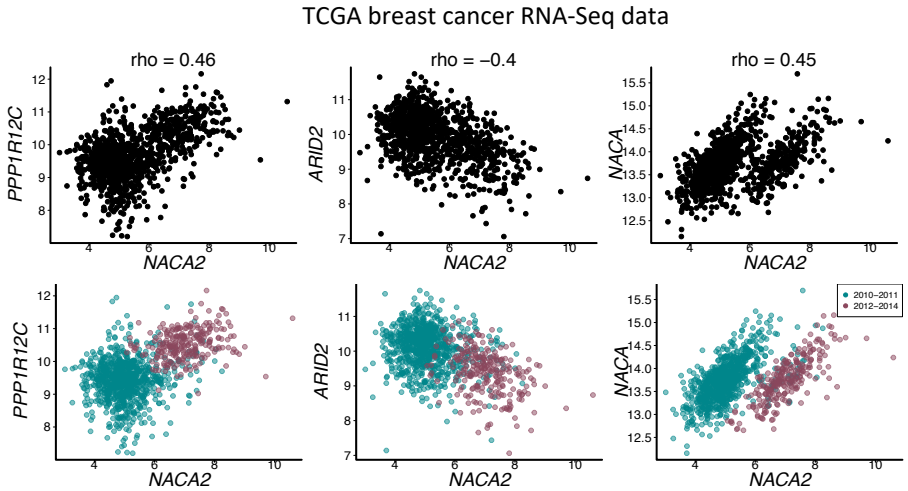
Laser Capture Microdissection (LCM) is a method for isolating specific cells of interest from microscopic regions of tissue

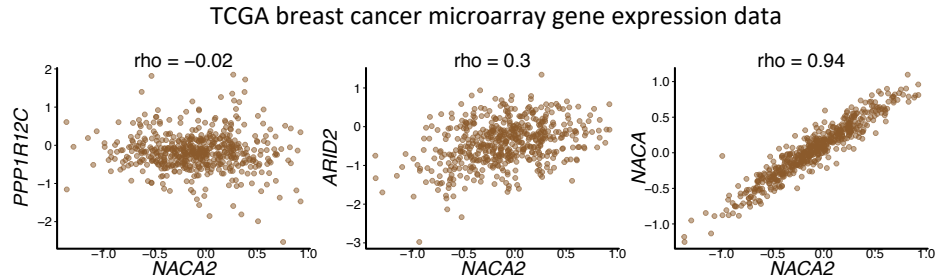


Using different flow cell chemistries introduced batch effects in TCGA BRCA RNAseq data









Challenges of using RUV-III to normalize TCGA RNA-Seq data

Challenge 1:

The TCGA RNA-seq datasets do NOT have any technical replicates?!!

Solution 1:

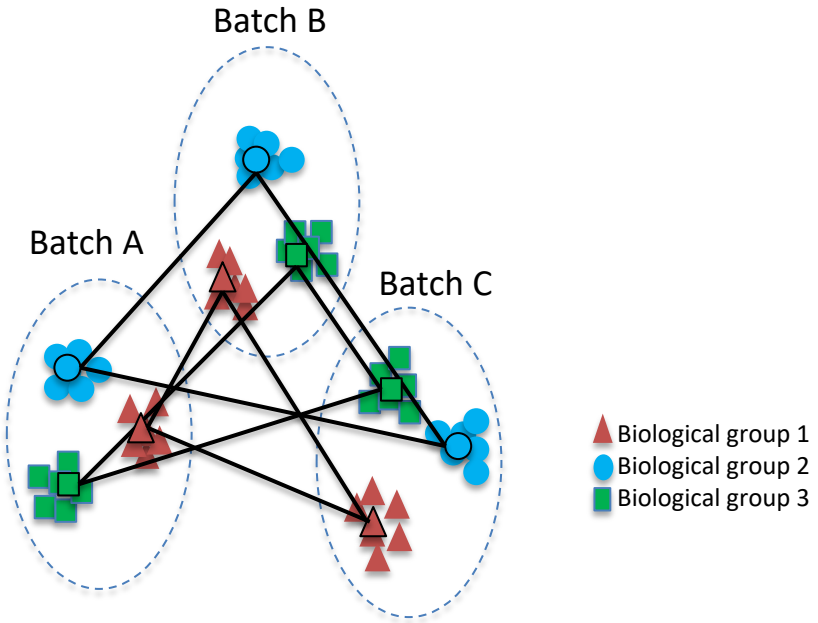
RUV-III based on our **pseudo-replicate approach**[1]
(this approach has been used in the scMerge method for scRNA-Seq data[2])

Comments:

- In cancer RNAseq data, pseudo-replicates may capture sample to sample differences.
- Technical replicates can not be used to estimate unwanted variation caused by tumor purity as a pair of replicate has the same amount of tumor purity.

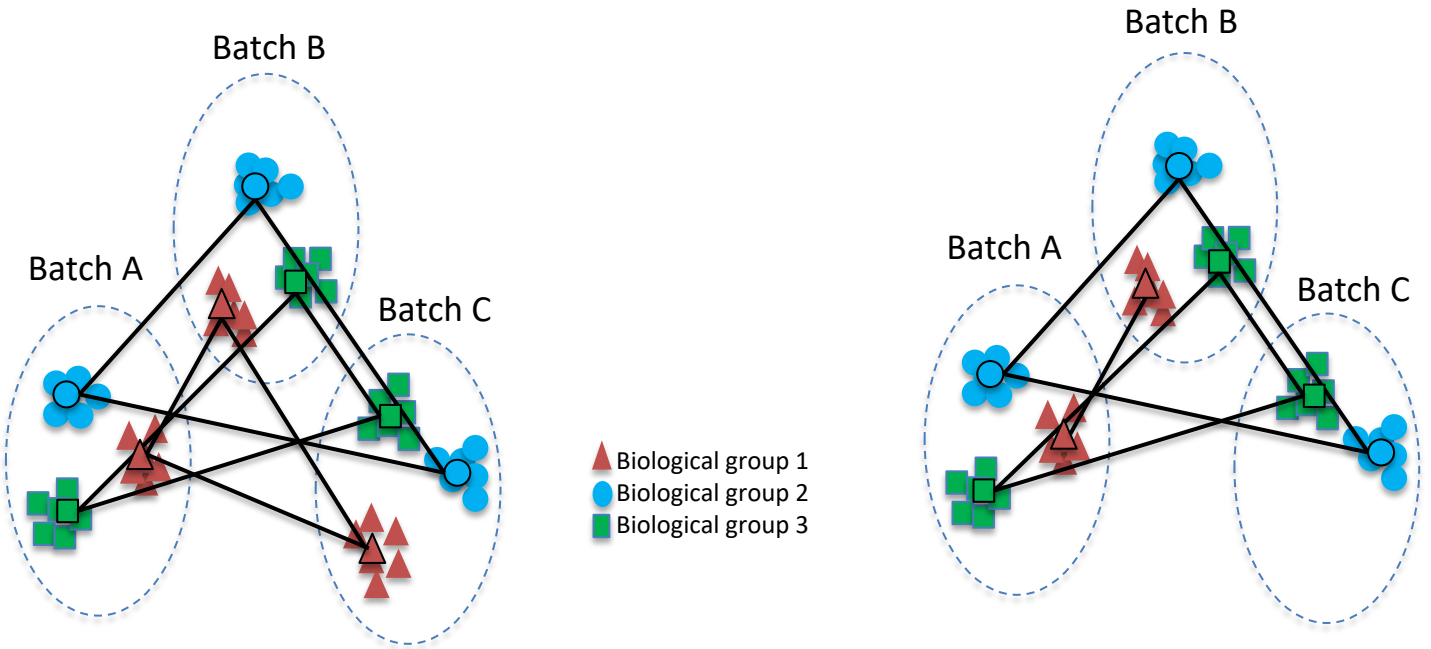
Solution 2:

pseudo-replicate of pseudo samples approach

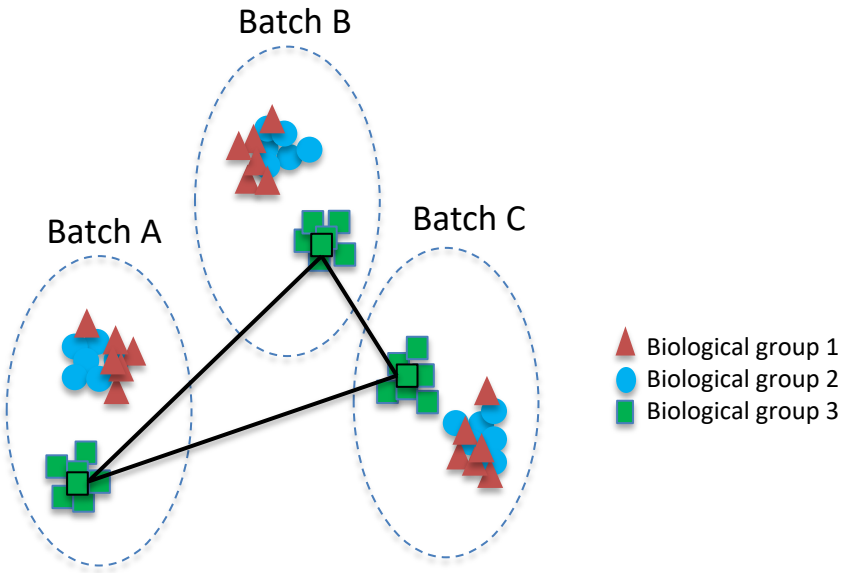


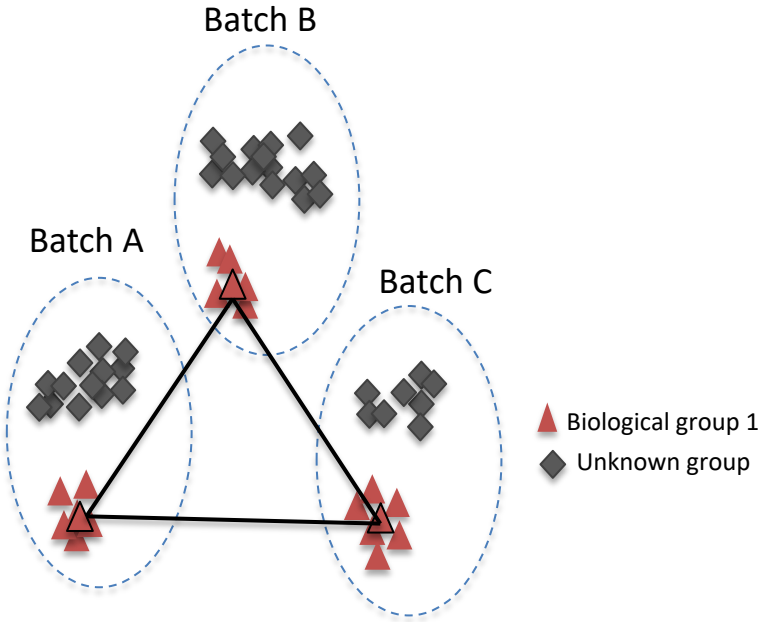
[1]R.Molania, *et.al. Nucleic Acid Research*, 2019.
[2]Yingxin Lin, *PNAS*, 2019.

Pseudo-replicate of pseudo samples approach



Pseudo-replicate of pseudo samples approach, cont.



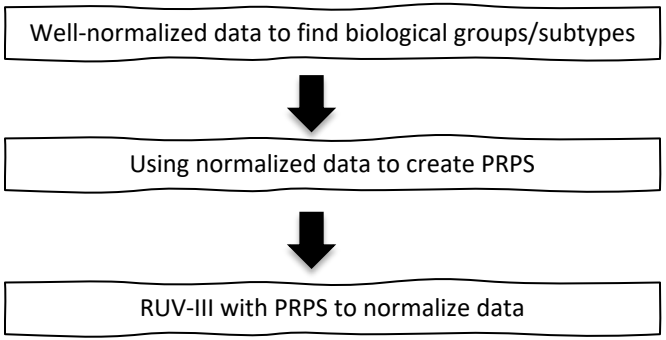


Breast cancer PAM50 subtypes: **Basal**, Her2, LumA and LumB

Challenges of using RUV-III to normalize TCGA RNA-Seq data, cont.

Challenge 2:

How to identify biological groups/subtypes in TCGA RNAseq data.



Solution 1: iterative normalization

- using normal tissues to create PRPS
- running RUV-III with the PRPS
- assessing the performance of RUVIII
- identifying biological groups using RUVIII adjusted data
- creating a new set (larger) of PRPS
- running RUV-III with the new set of PRPS
- assessing the performance of RUVIII
- Obtaining well-normalized data

Problems:

- Normal tissues were not distributed across batches.
- Normal tissues are not appropriate to remove tumour purity.

Solution 2: biological-independent PRPS

- Identifying sources of batch effects
- creating PRPS across batches without considering biological groups within each batch
- running RUV-III with the PRPS
- assessing the performance of RUVIII
- identifying biological groups using RUVIII adjusted data
- creating a new set (larger) of PRPS
- running RUV-III with the new set of PRPS
- assessing the performance of RUVIII
- obtaining well-normalized data

Challenges of using RUV-III to normalize TCGA RNA-Seq data, cont.

Challenge 3:

Selection of appropriate negative control genes.

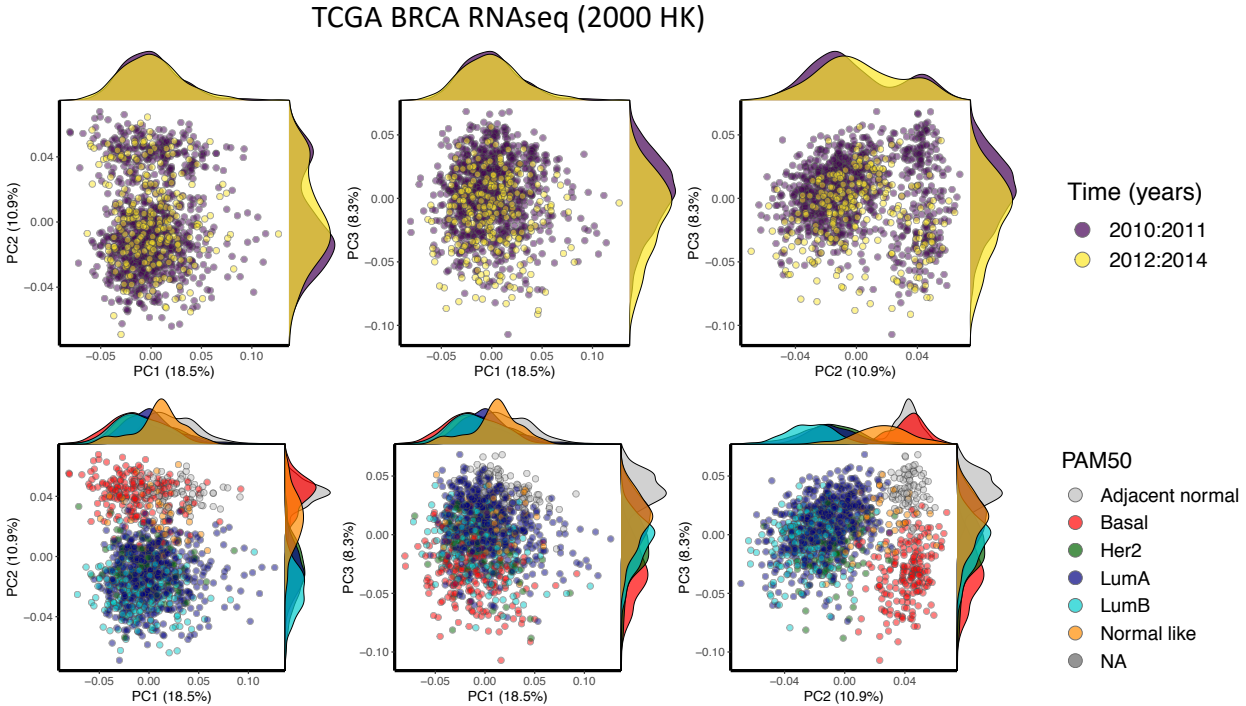
The current lists of stable genes are NOT appropriate for all TCGA cancer types.

Solution 1:

Data-driven stable gene.

Solution 2:

Using all genes



[1]R.Molania, *et.al. Nucleic Acid Research*, 2019.

[2]Yingxin Lin, *PNAS*, 2019.

The performance of RUV-III with biological-independent PRPS and using all genes as negative control genes

RUV-III:

- biological-independent PRPS
- All genes as negative control genes

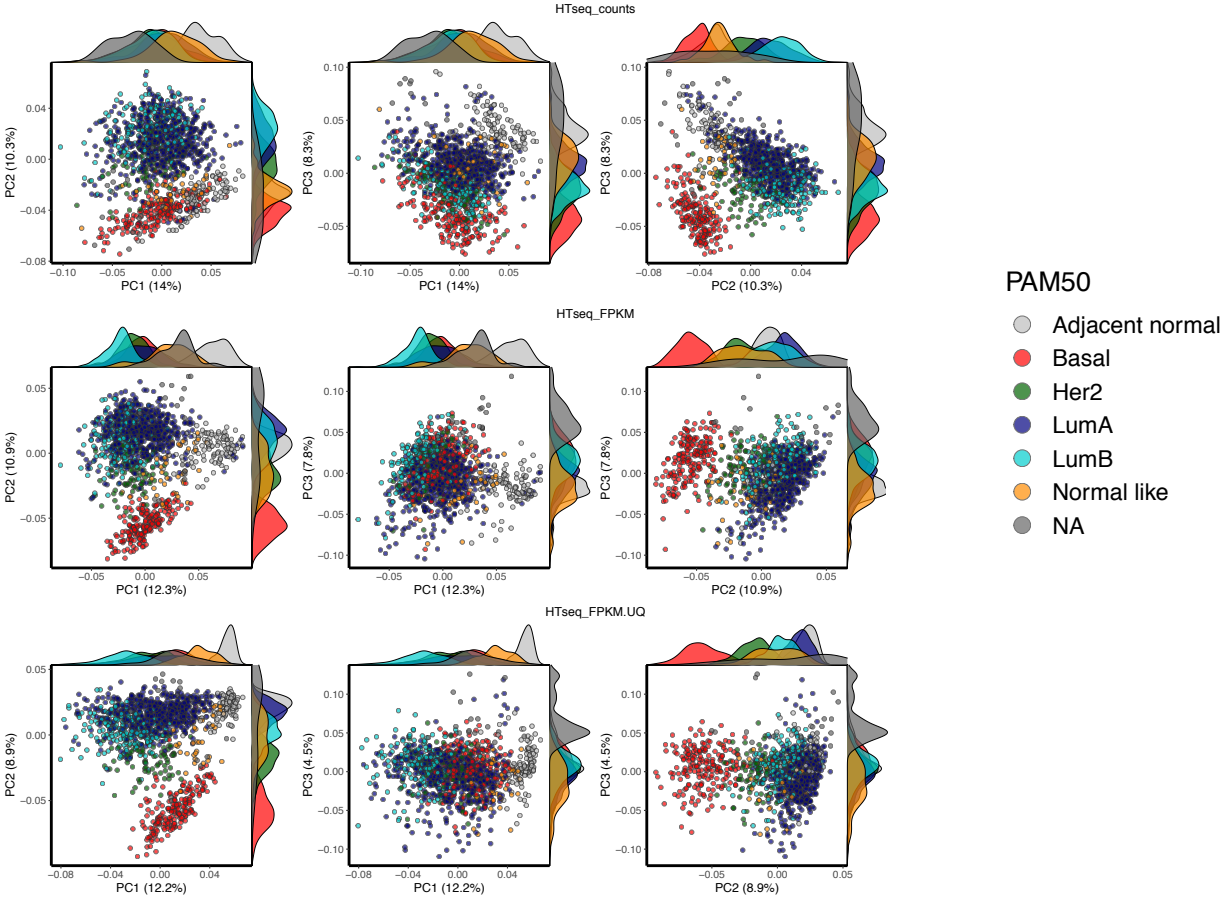
Datasets

- TCGA BRCA RNAseq data
- TCGA READ RNAseq data
- TCGA COAD RNAseq data

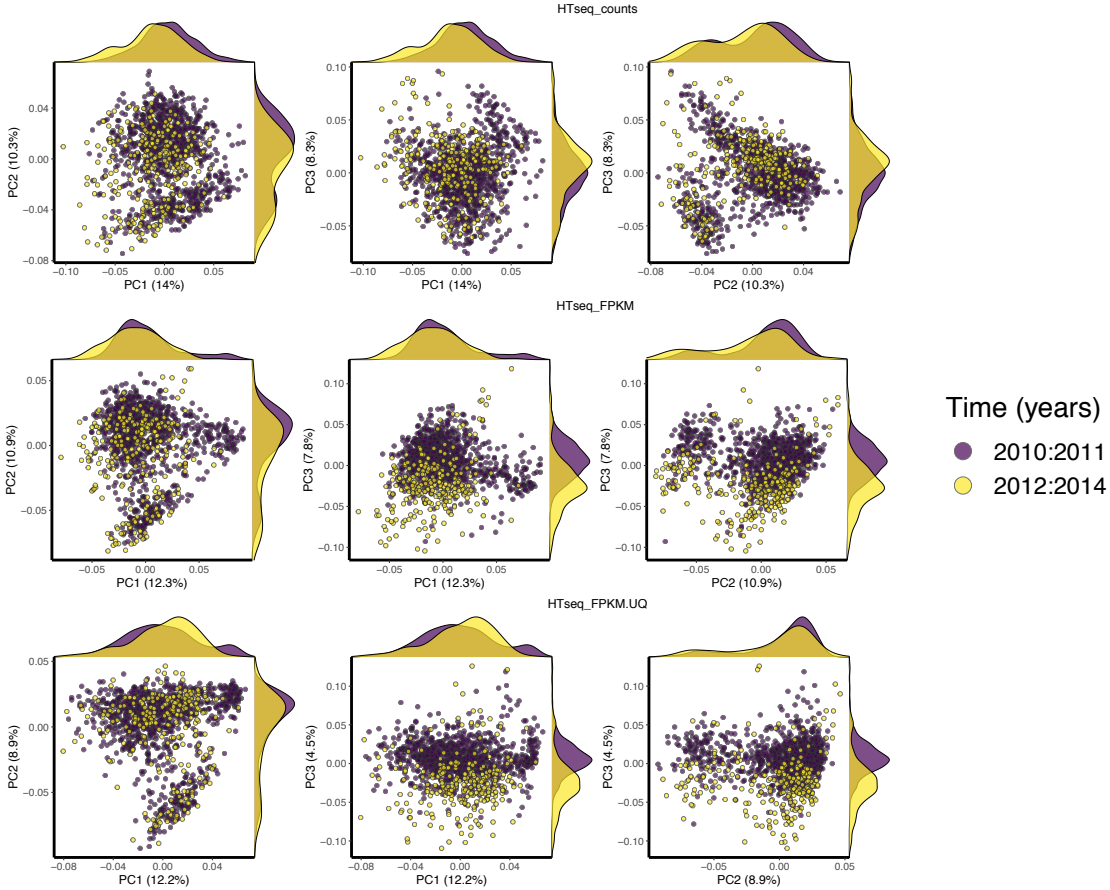
Workflow:

- Identifying sources of batch effects
- creating PRPS across batches without considering biological groups within each batch
- running RUV-III with the PRPS
- assessing the performance of RUVIII

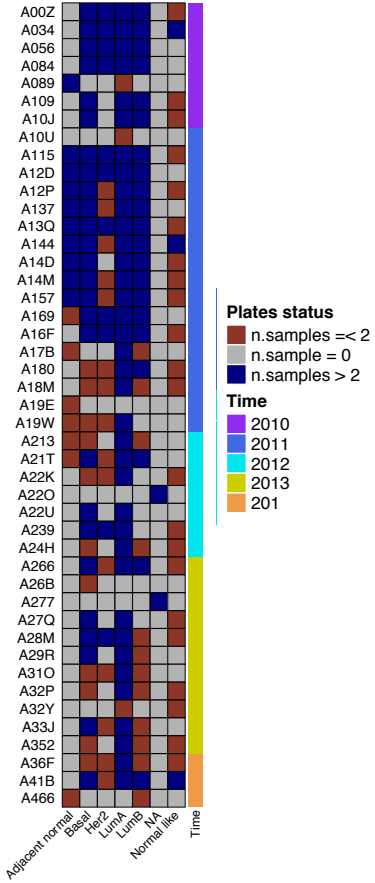
The PAM50 subtypes of TCGA BRCA RNAseq data

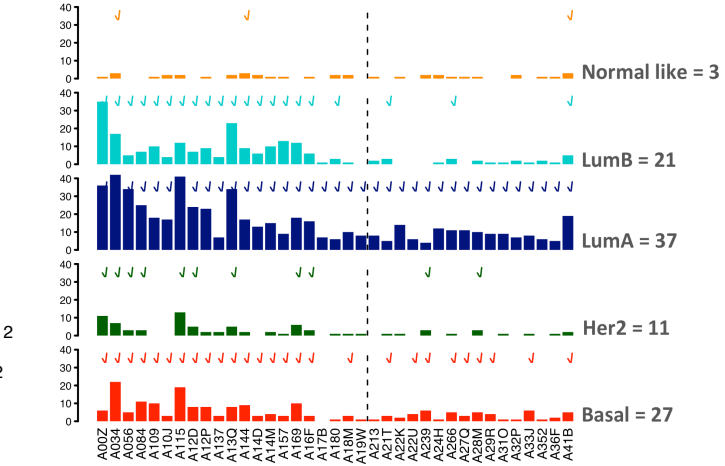


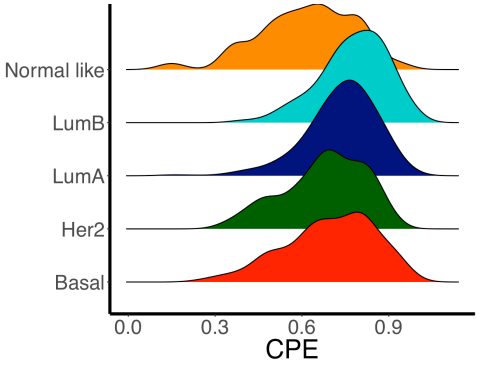
Batch effects in the TCGA BRCA RNAseq data

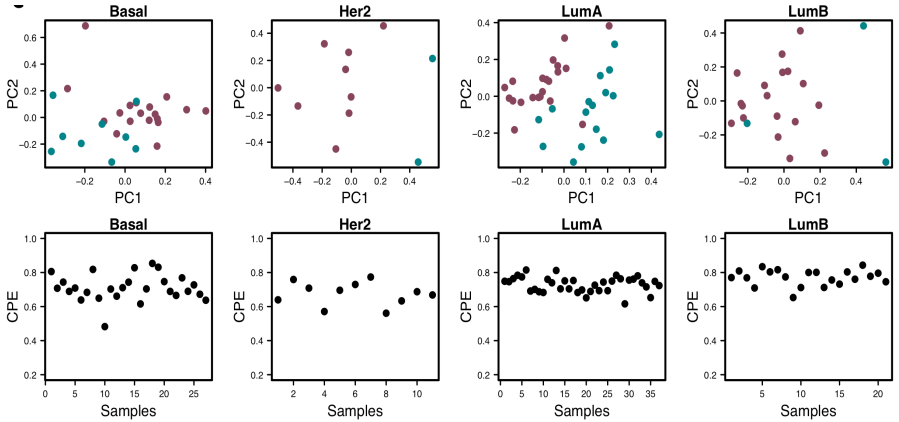


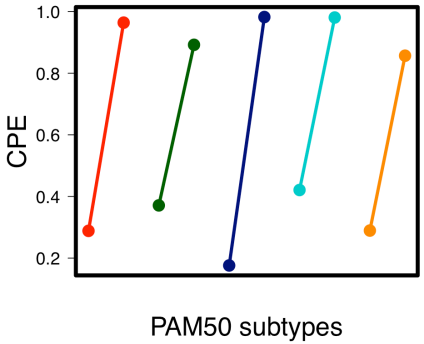
Appropriate approach to create pseudo replicates of pseudo samples



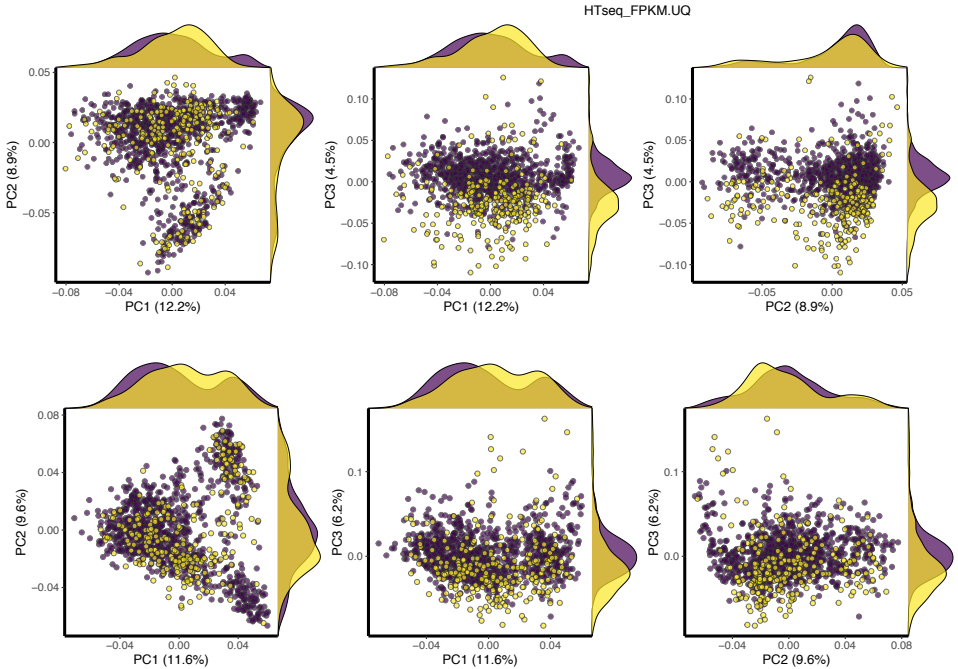


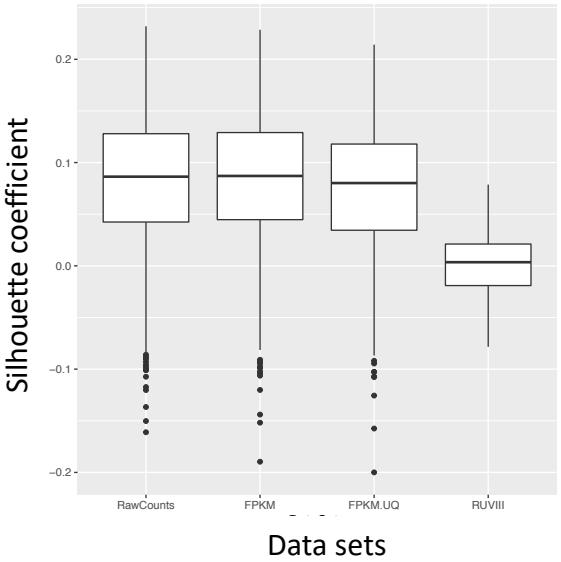




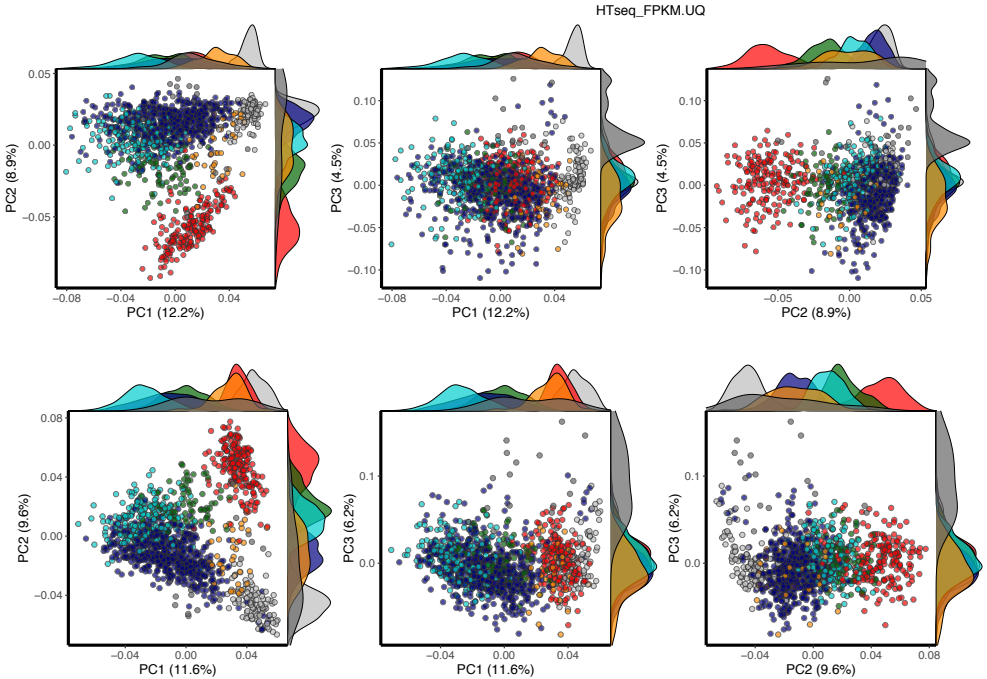


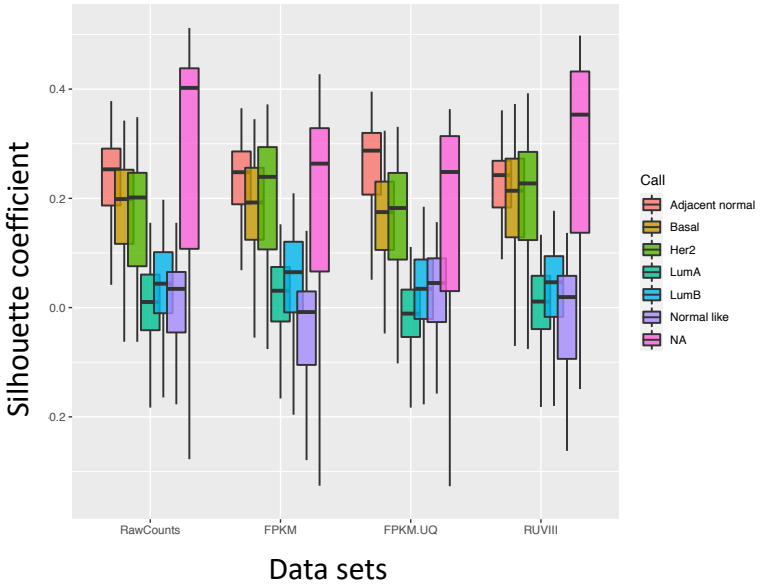
TCGA BRCA RNAseq, RUV-III normalization



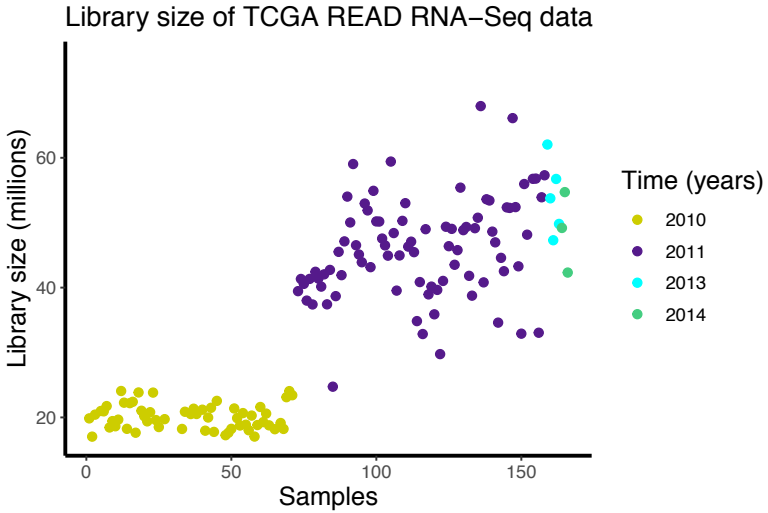


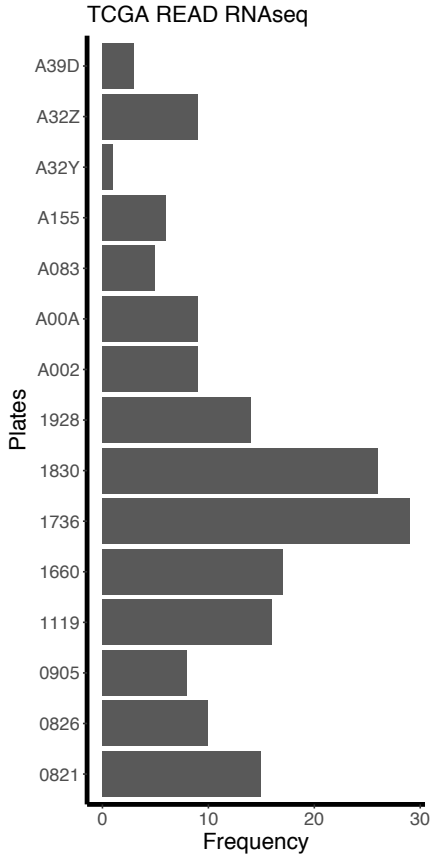
TCGA BRCA RNAseq, RUV-III normalization





Example 2, TCGA READ RNAseq data



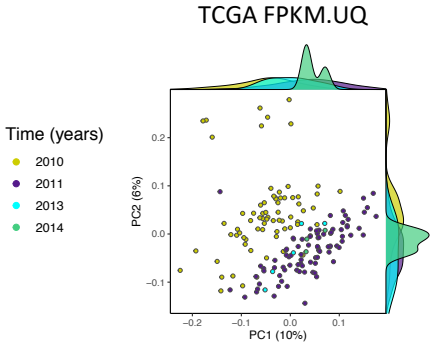


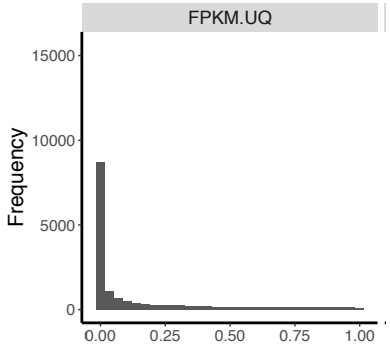
RUV-III:

PRPS: Averaging samples per individual plates

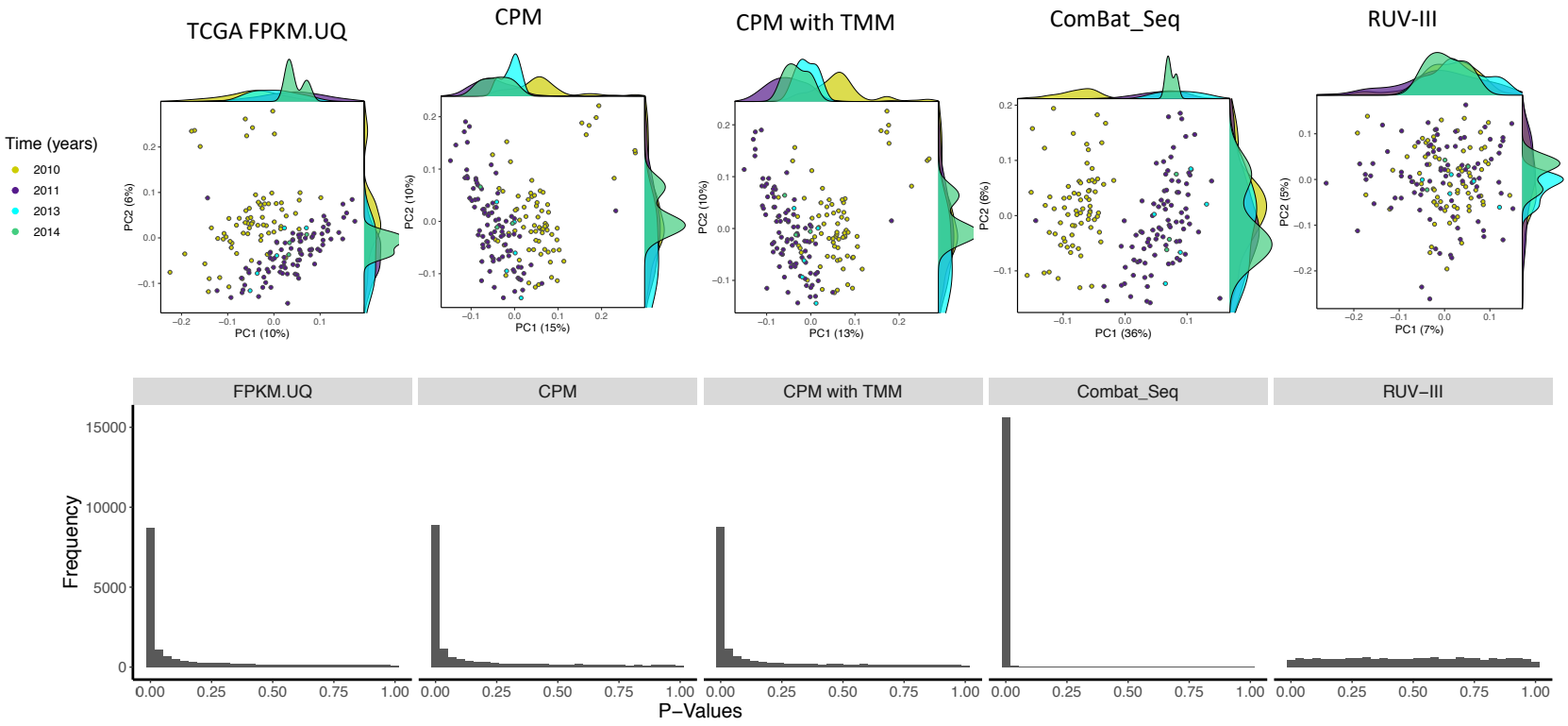
Negative control genes: using all genes

RUV-III removes the library size effects in TCGA READ RNA-Seq data





RUV-III removes the library size effects in TCGA READ RNA-Seq data



RUV-III preserves biological variation in the TCGA READ RNA-Seq data

Martinez-Romero *et al. BMC Genomics* 2018, **19**(Suppl 8):857
https://doi.org/10.1186/s12864-018-5193-9

BMC Genomics

RESEARCH Open Access

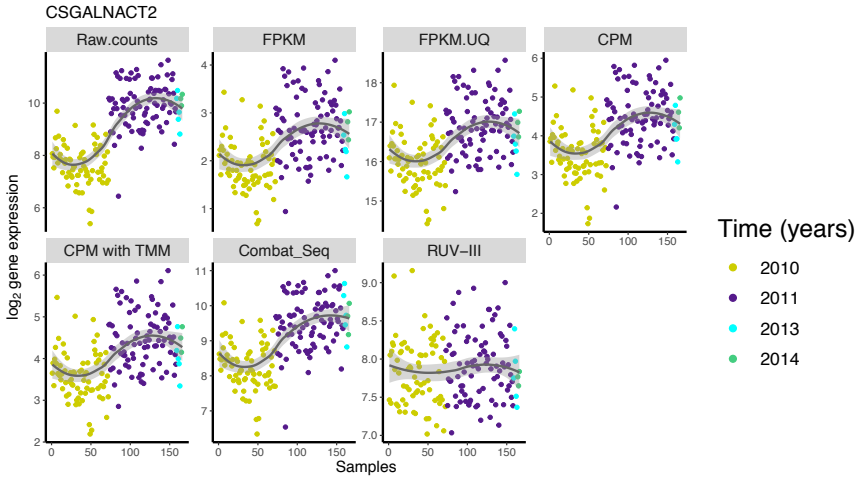
Survival marker genes of colorectal cancer derived from consistent transcriptomic profiling

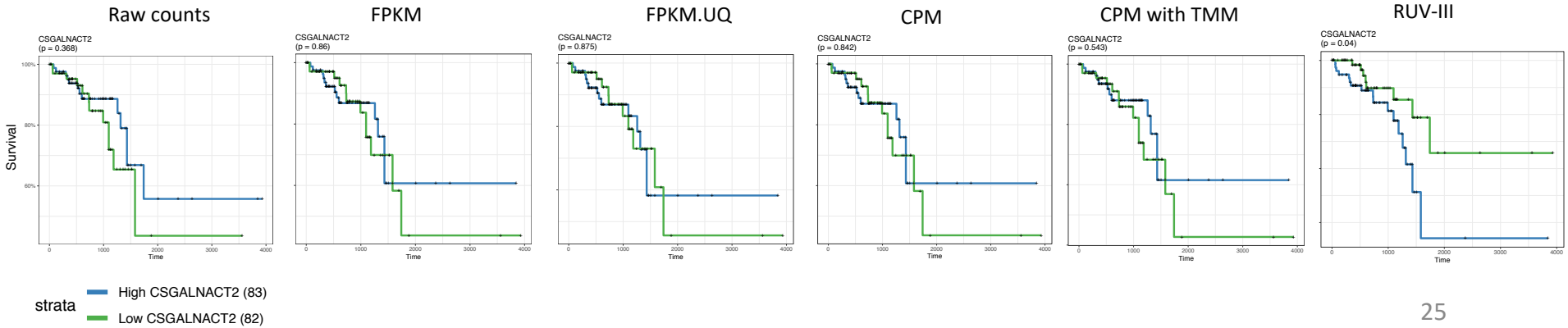
Jorge Martinez-Romero[1,2†], Santiago Bueno-Fortes[1†], Manuel Martín-Merino[1,3], Ana Ramirez de Molina[2] and Javier De Las Rivas[1*]

~1200 samples (microarray gene expression)

CSGALNACT2(chondroitin sulfate N-acetylgalactosaminyltransferase 2)

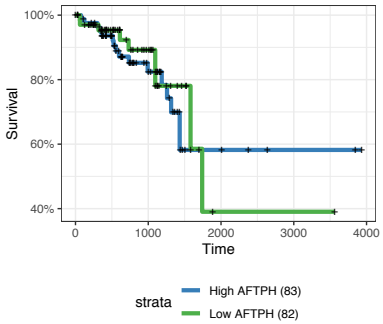
The lower gene expression, the longer survival



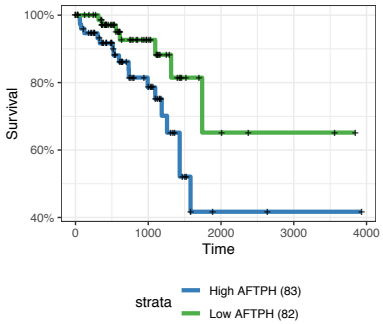


RUV-III preserves biological variation in the TCGA READ RNA-Seq data

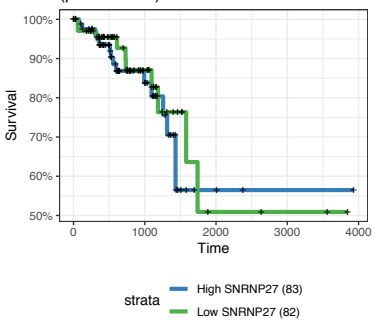
FPKM.UQ



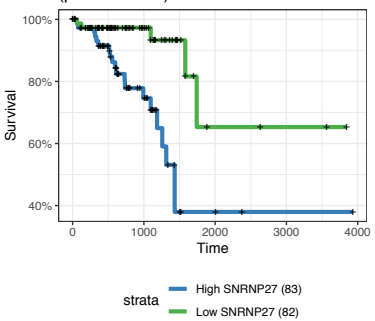
RUV-III



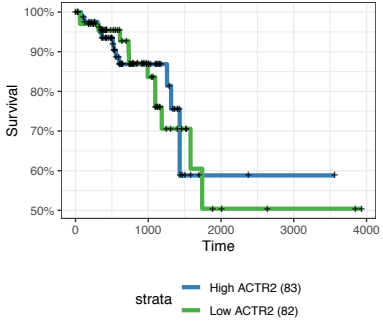
FPKM.UQ



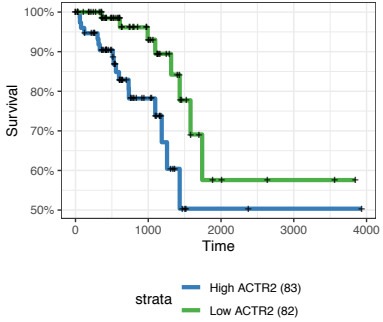
RUV-III



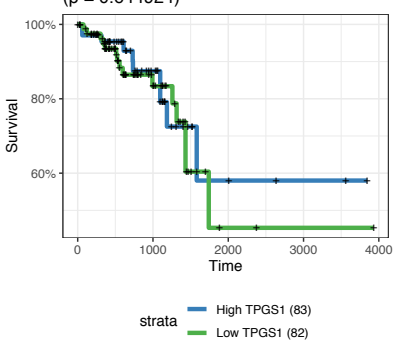
FPKM.UQ



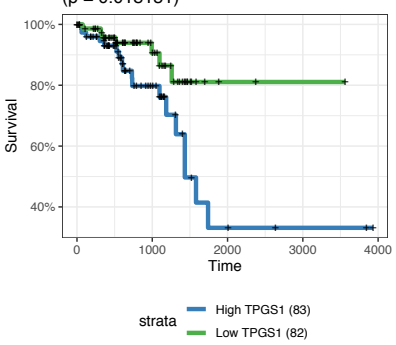
RUV-III



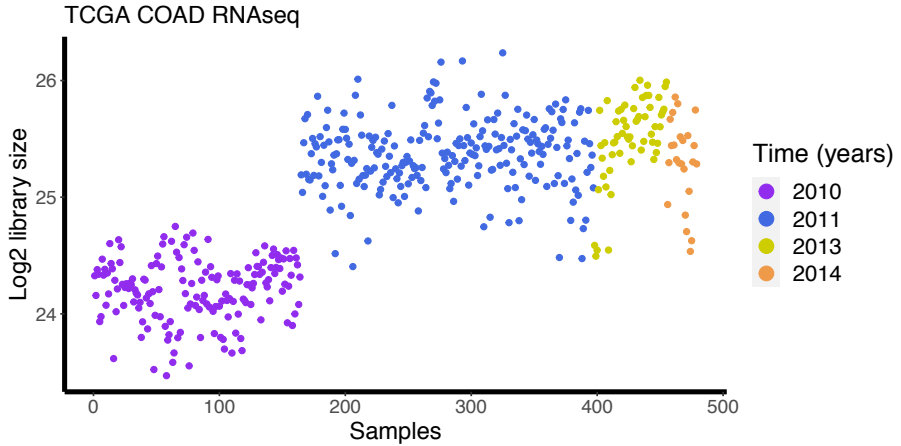
FPKM.UQ

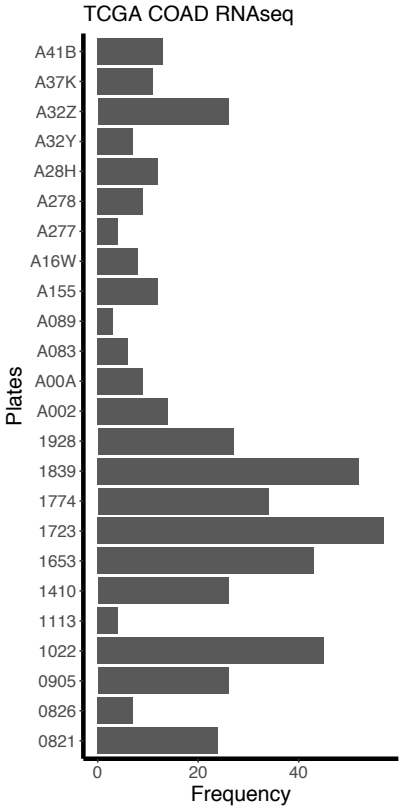


RUV-III



Example 3, TCGA COAD RNAseq data

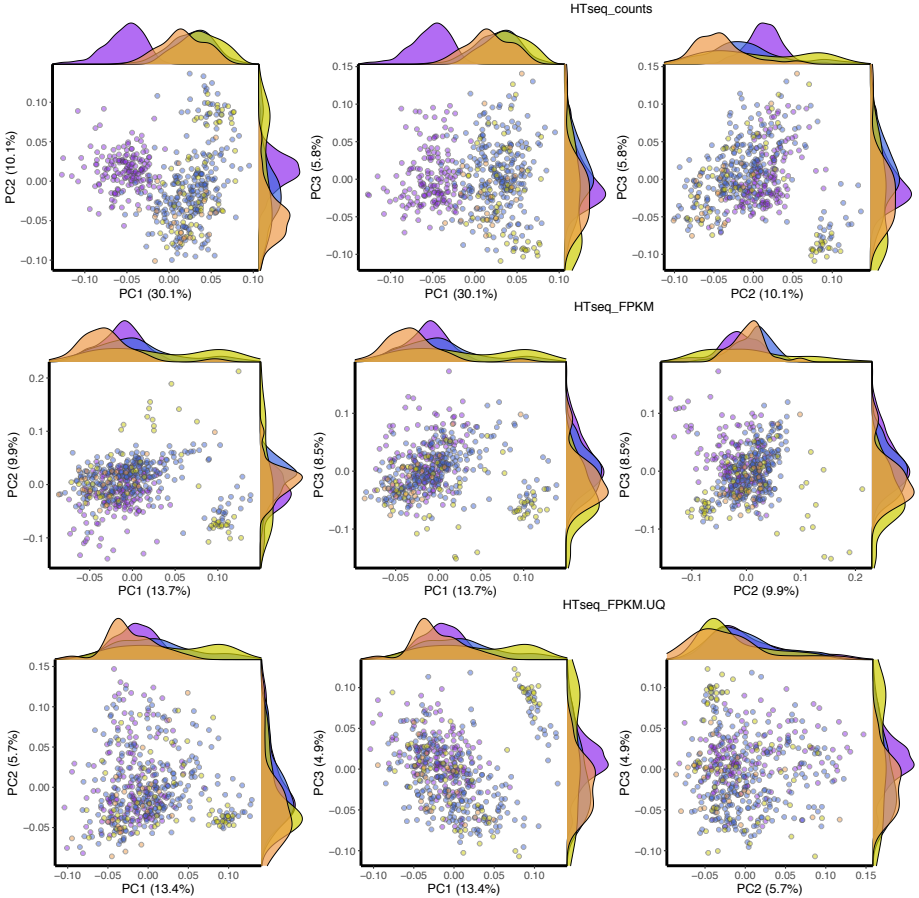


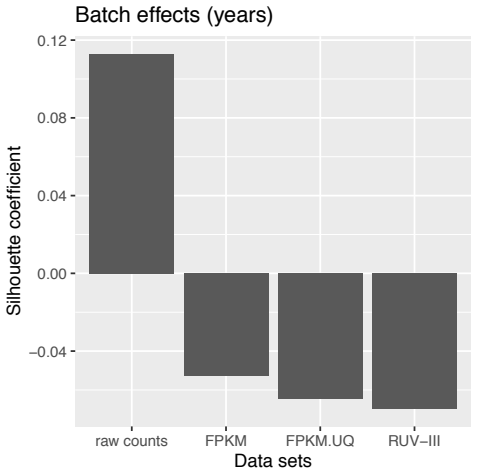


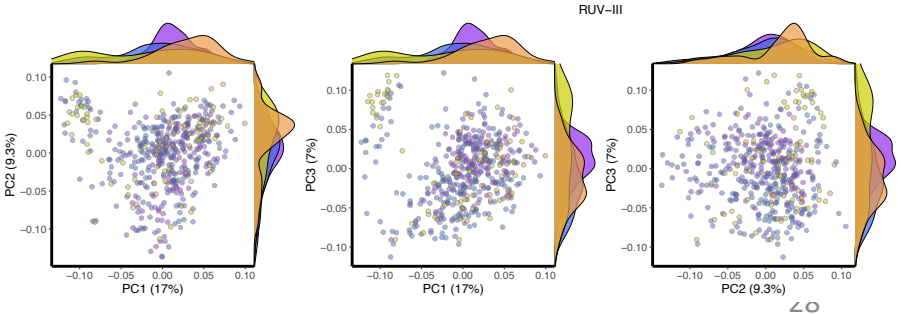
RUV-III:

PRPS: Averaging samples per individual plates
Negative control genes: using all genes

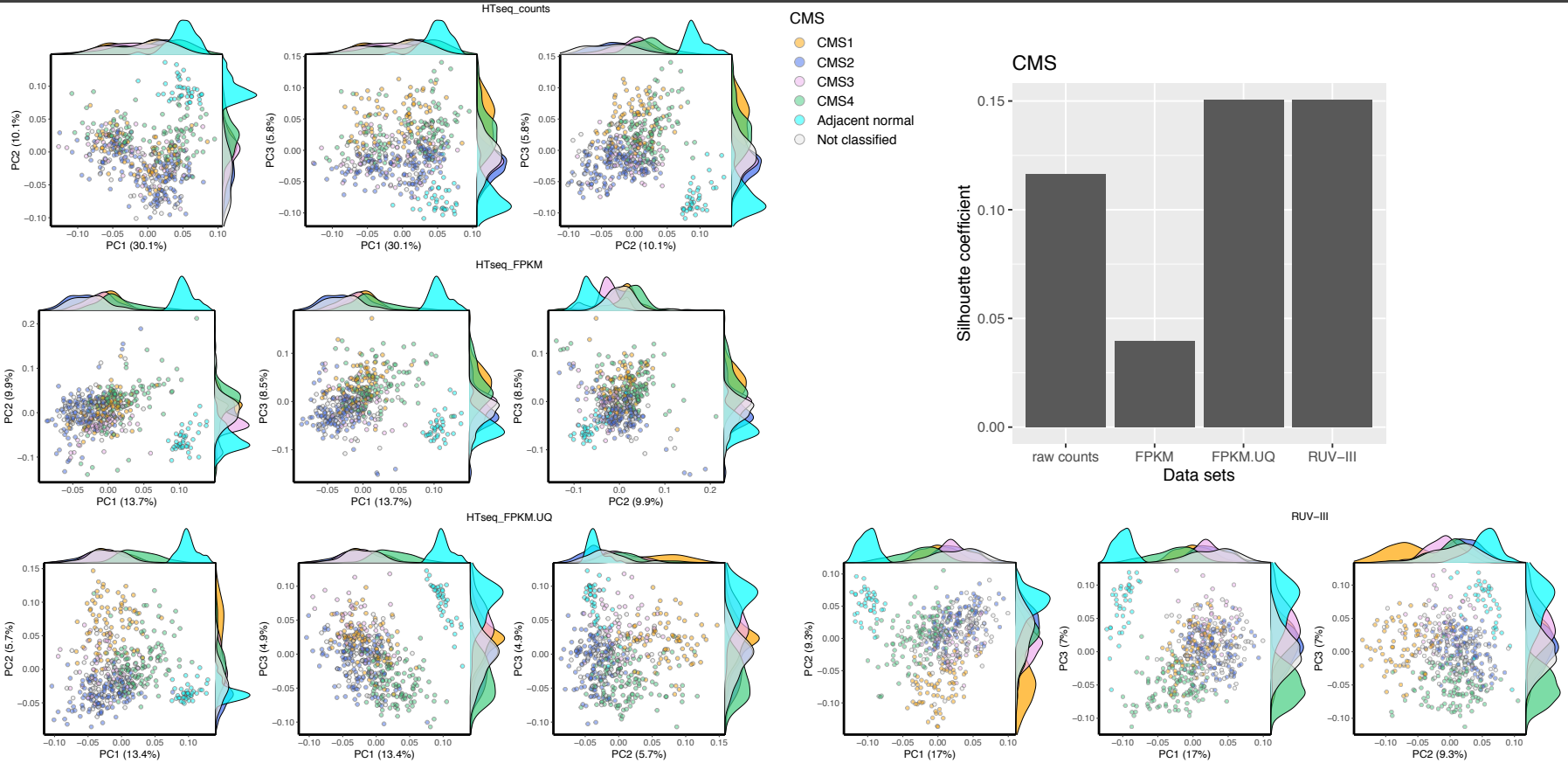
Time effects in the TCGA COAD RNAseq data



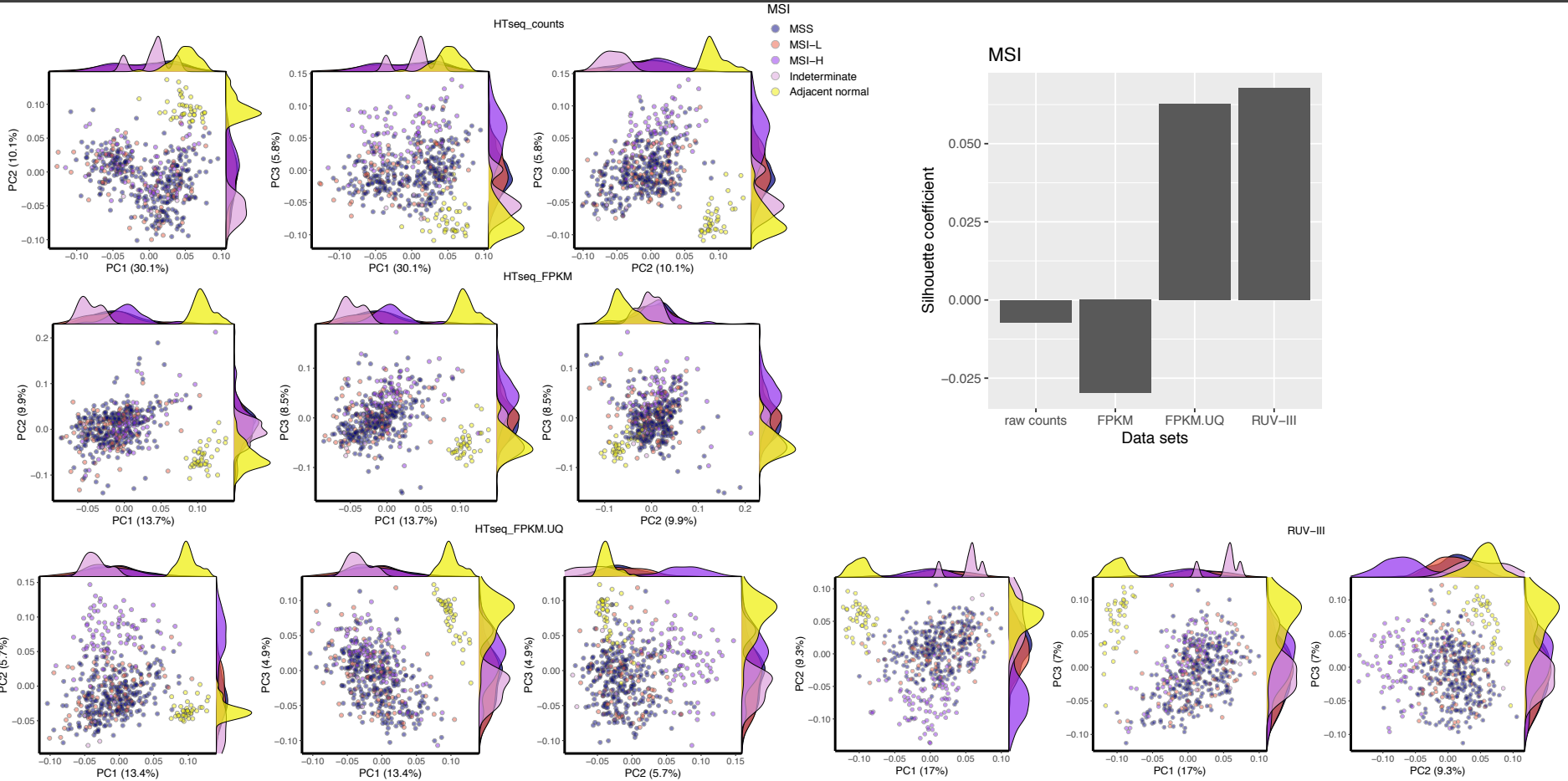




Consensus molecular subtypes of the TCGA COAD RNAseq data

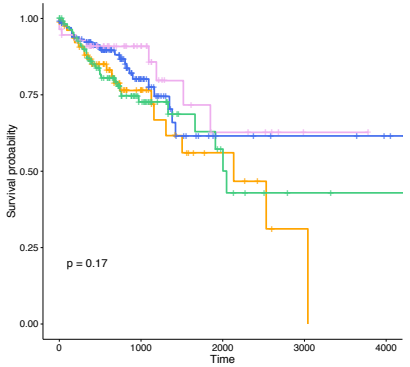


Microsatellite instability in the TCGA COAD RNAseq data

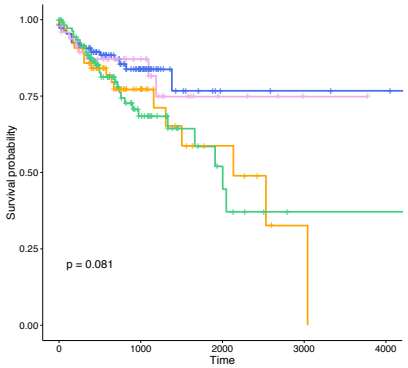


Survival analysis of the CMS in the TCGA COAD RNAseq data

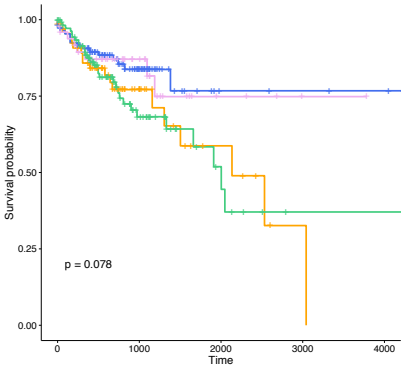
TCGA raw counts



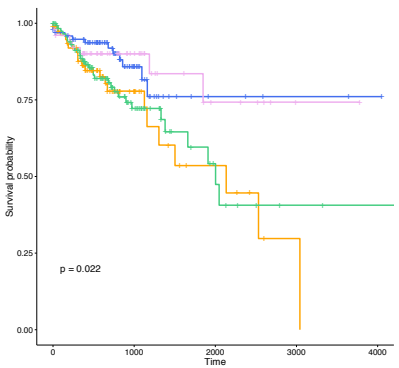
TCGA FPKM



TCGA FPKM.UQ

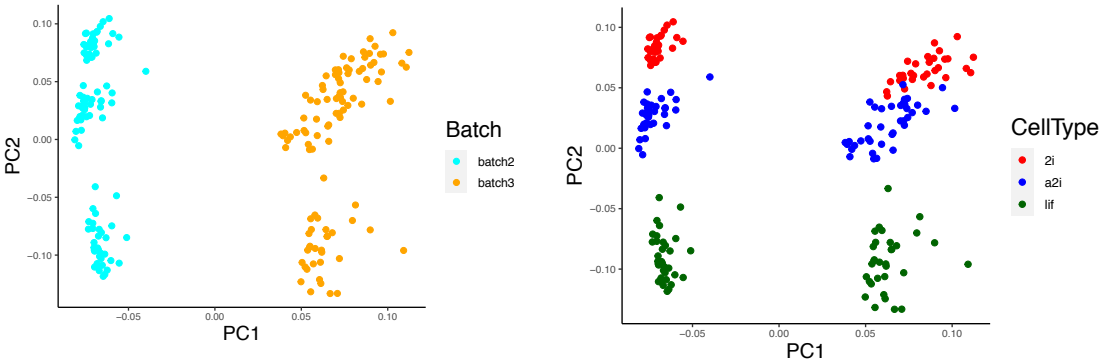


RUV-III normalization

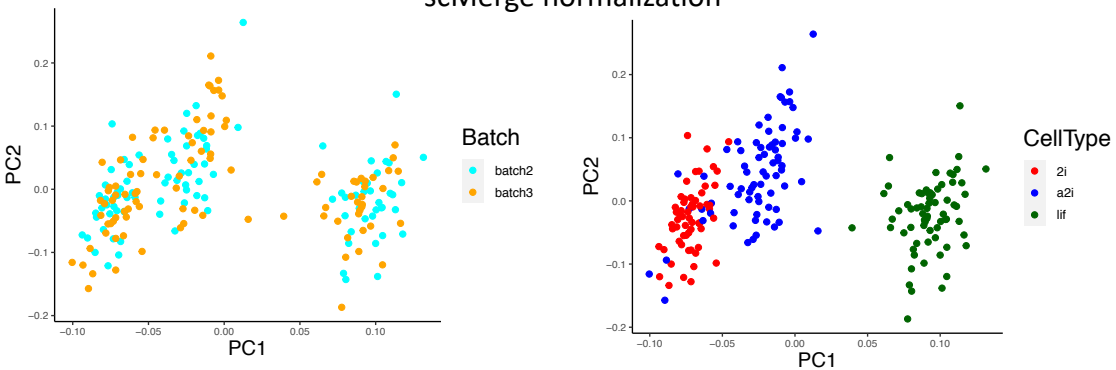


RUV-III with PRPS in single cell RNAseq data

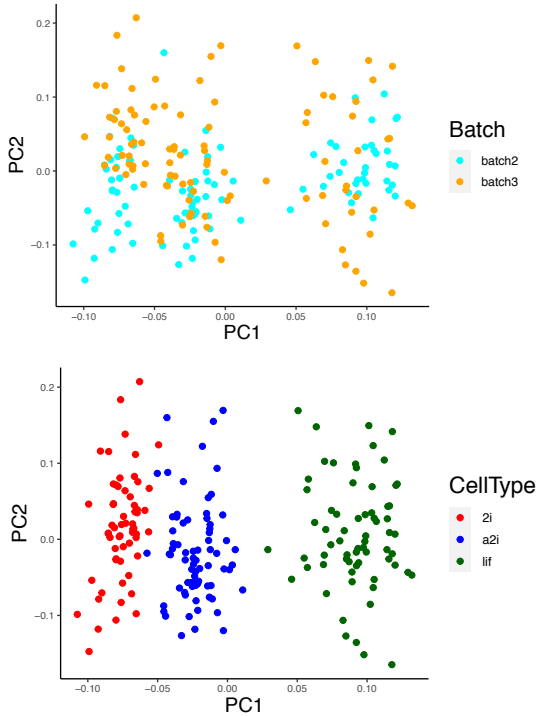
Raw data



scMerge normalization



RUV-III with only 1 pair of PRPS



Source:https://bioconductor.org/packages/release/bioc/vignettes/scMerge/inst/doc/scMerge.html

Current challenge and Summary

Challenge:

- integration of different TCGA RNAseq datasets. e.g. COAD and READ.
- Using normal tissues to create pseudo replicate of pseudo samples across different TCGA datasets.

Summary

- Library size, tumor purity and batch effects are main sources of unwanted variation in TCGA RNA-Seq datasets and they can compromise down-stream analysis.
- RUV-III with PRPS promises to be valuable to remove unwanted variation from large-scale RNAseq data.
- RUV-III is robust to selection of PRPS and negative control samples

Acknowledgement

WEHI bioinformatics

Terry Speed

Tony Papenfuss

Luke Gandolfo

Gavriel Olshansky

Monash University

Momeneh (Sepideh) Foroutan

Translational Genetics and Epigenetics Group – ONJCRI

Alex Dobrovic

Department of Statistics - University of Michigan

Johann Gagnon-Bartsch