nature portfolio

https://doi.org/10.1038/s41587-022-01440-w

Supplementary information

Removing unwanted variation from large-scale RNA sequencing data with PRPS

In the format provided by the authors and unedited

Removing unwanted variation from large-scale RNA-sequencing data with PRPS

Ramyar Molania[1,2*], Momeneh Foroutan[3], Johann A. Gagnon-Bartsch[4,] Luke C Gandolfo[1,2,5], Aryan Jain[6], Abhishek Sinha[6], Gavriel Olshansky[7,8], Alexander Dobrovic[9], Anthony T Papenfuss[1,2,10,11*^], Terence P Speed[1,5*^]

[1]Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia;
[2]Department of Medical Biology, The University of Melbourne, VIC 3010 Australia;
[3]Biomedicine Discovery Institute and the Department of Biochemistry and Molecular Biology, Monash University, Clayton, VIC 3145, Australia;
[4]Department of Statistics, University of Michigan, Ann Arbor, Michigan, MI 48109, USA;
[5] School of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australia;
[6]Department of Economics and Statistics, Monash University, VIC 3170, Australia;
[7]Metabolomics Laboratory, Baker Heart and Diabetes Institute, Melbourne, VIC 3004 Australia;
[8]Baker Department of Cardiometabolic Health, The University of Melbourne, VIC 3010, Australia;
[9]Department of Surgery, The University of Melbourne, Austin Health, Heidelberg, VIC 3084, Australia;
[10]Peter MacCallum Cancer Centre, Melbourne, VIC 3000, Australia;
[11]Sir Peter MacCallum Department of Oncology, The University of Melbourne, Victoria 3010, Australia.

*Corresponding authors
^These authors contributed equally

Table of Contents

Supplementary Text

Consensus molecular subtypes of the TCGA READ RNA-seq samples

Colorectal cancers (CRCs) can be classified into four widely accepted consensus molecular subtypes (CMS) based on their gene expression profiles [1] . This classification provides a framework for stratifying the treatment of patients with colon and rectum cancer. We used the CMScaller R package [2] to identify the CMS of the TCGA READ RNA-seq samples. The CMScaller provides a classification based on pre-defined cancer-cell intrinsic CMS templates. This classifier uses cosine distance or Pearson correlation coefficient to compute distance between training and test samples. Prediction confidence values (pred-conf) are estimated from gene resampling (n = 1000) and samples with false discovery rate adjusted pred-values > 0.05 are labelled as "not assigned" (NA). The R package CMScaller uses the Camera method from limma R package for the gene set enrichment analysis. We applied that classifier to all samples and also to samples within the key time intervals (2010 and 2011:2014) using the FPKM and FPK.UQ normalized datasets (Supplementary Figure 5). The reason for applying the classifier within each key time interval was to assess the effect of large library differences on the CMS classifications. We also identified CMS using the RUV-III normalized data (Supplementary Figure 6 A and B). We found a high concordance between the CMS obtained using the different strategies and datasets normalized by different methods (Supplementary Figure 6 C). The CMS obtained from the FPKM.UQ normalized data were used to create pseudo-samples for RUV-III normalization.

TCGA COAD RNA-seq study Outline

The TCGA colon adenocarcinoma (COAD) RNA-seq study consists of 479 assays generated across 4 years (Supplementary Figure 13). There were 40 adjacent normal colon tissues profiled after 2010. The median library size of samples assayed in 2010 is less than half that of the rest of the samples (Supplementary Figure 13). The consensus molecular subtypes (CMS) of these samples were identified using the R package CMScaller [2] on the raw counts and differently normalized datasets (Supplementary Figure 13 and 14).

Consensus molecular subtypes of the TCGA COAD RNA-seq samples

As with the TCGA READ RNA-seq data, we used the R package CMScaller to identify the consensus molecular subtypes (CMS) of the TCGA COAD samples. We applied the classifier to the raw counts and the FPKM, FPKM.UQ, and RUV-III normalized datasets (Supplementary Figure 14 and 15 A and B). We also applied the classifier within each key time interval, 2010 and 2011-2014, in the TCGA normalized datasets to assess the effects of the large library size differences between the time intervals on the CMS classifications. The results showed that the proportion of un-classified samples was significantly smaller when the classifier was applied within each key time interval compared with that when applied across all samples together (Supplementary Figure 15 C and D). Supplementary Figure 15 E shows that the CMS obtained using the full set of the FPKM and FPKM.UQ data are associated with library size. We also found the CMS identified in the RUV-III normalized data are highly concordant with the CMS obtained using the TCGA normalized datasets within each key time point. These results show the effect of library size on the identification of CMS in the full sets of FPKM and FPKM.UQ normalized data.

It has been shown that different CMS are associated with various gene signatures [2]. Gene set expression analyses show that the CMS obtained from the FPKM, FPKM.UQ and RUV-III normalization are associated with the known gene signatures. (Supplementary Figure 14, B and 15, B). These results suggested that the different normalizations preserve known biological signals associated with CMS in the data.

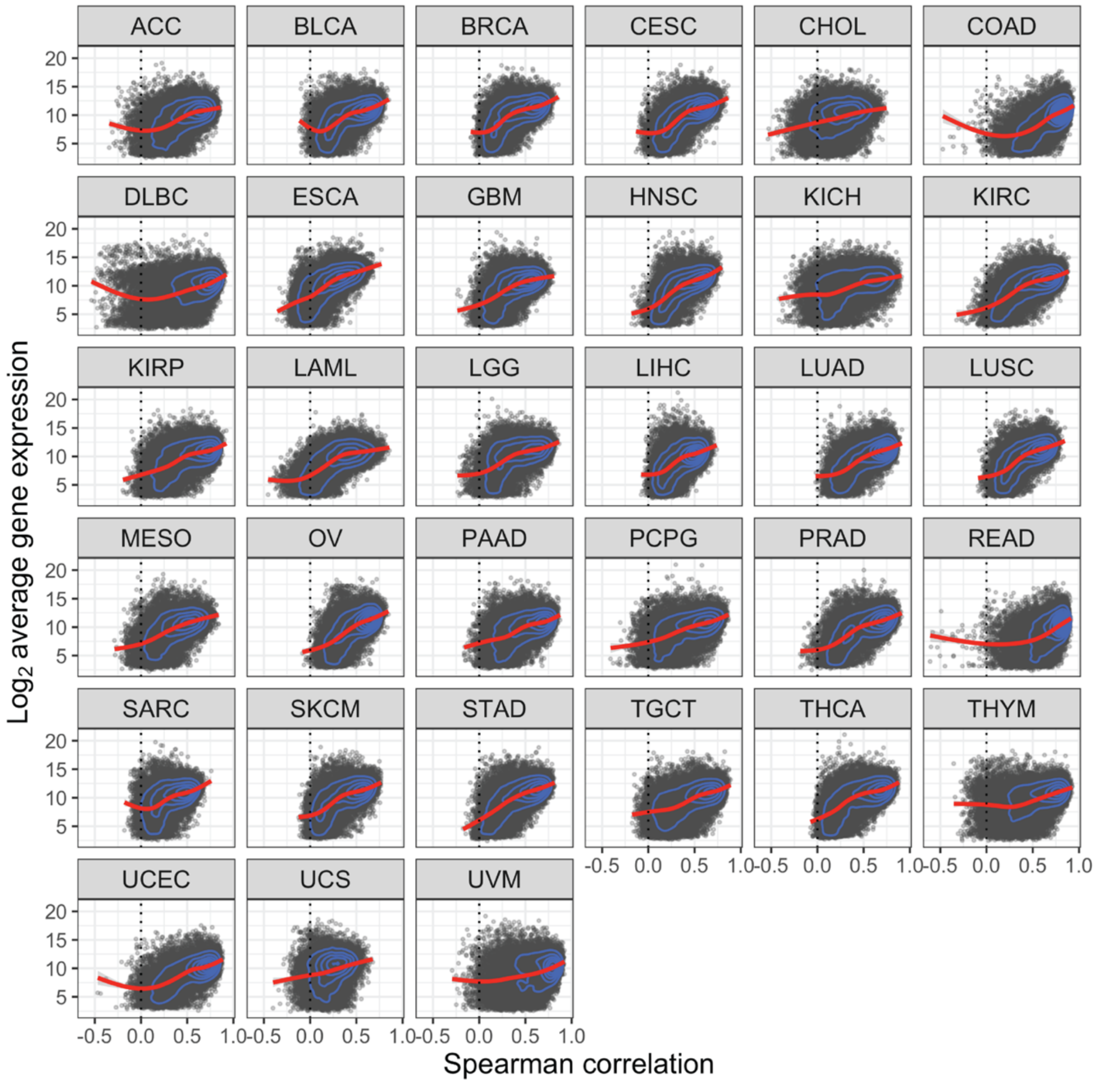
Breast tumour intrinsic subtypes

Tumour intrinsic molecular subtypes given by the prediction analysis of microarray 50 (PAM50) approach were established to better understand the biology and improve clinical outcomes of breast cancer [3]. The intrinsic subtypes of breast cancer are basal-like (Basal), human epidermal growth factor receptor 2 (HER2)–enriched, luminal A (LumA), luminal B (LumB) and normal-like. The function *molecular.subtyping()* in the genefu R/Bioconductor package [4] and a classifier reported by Picornell et al. [5] were used to identify the PAM50 subtypes in the TCGA RNA-sequencing datasets.

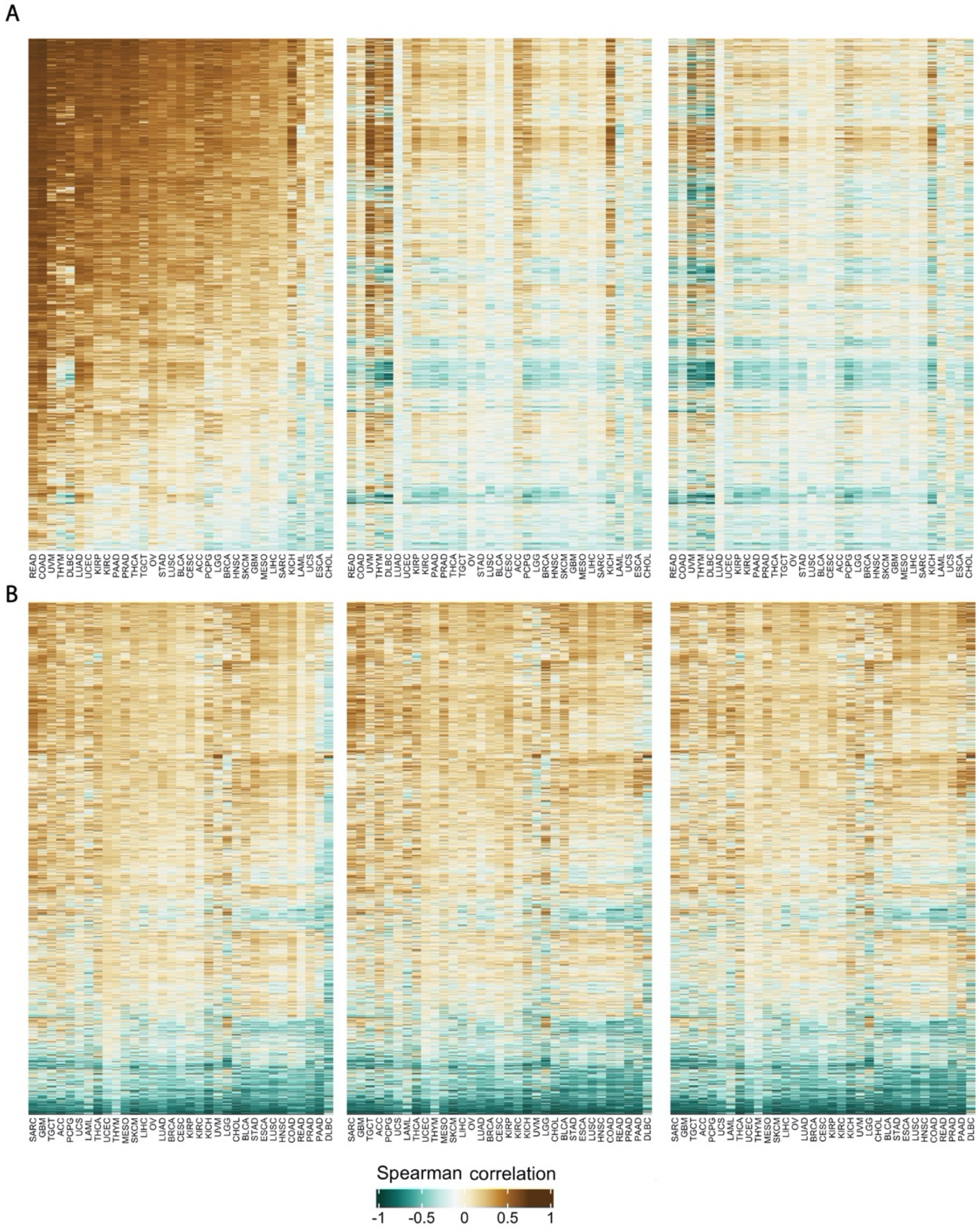
This method was used on both ER-balanced and ordinary gene expression datasets normalized by differed methods. To obtain ER-balanced data, we first selected a subset of cancer samples which are ER status-balanced (the same sample sizes for estrogen receptor-positive and negative values based on clinical measurements) and used the median derived from that subset to centre the entire cohort. Figure 27 A shows the concordance between the PAM50 subtypes identified by the two methods in the different datasets. The PAM50 subtypes called by the genefu on the TCGA FPKM and FPKM.UQ normalized data were used for PRPS.

We also found that the chemistry effects can lead to unreliable gene expression differences between paired primary and metastatic samples assayed across the chemistries. Supplementary Figure 31 showed MA plots of the FPKM.UQ and RUV-III normalized expression values for three primary and metastatic pairs of samples. Samples from two of the pairs were profiled using the same chemistry, and those from the third pair were profiled across the two chemistries. These results suggest that a considerable proportion of the gene expression differences between paired primary and metastatic samples that were profiled across chemistries likely results from the different flow cell chemistries, as we did not see this in the RUV-III normalized data.

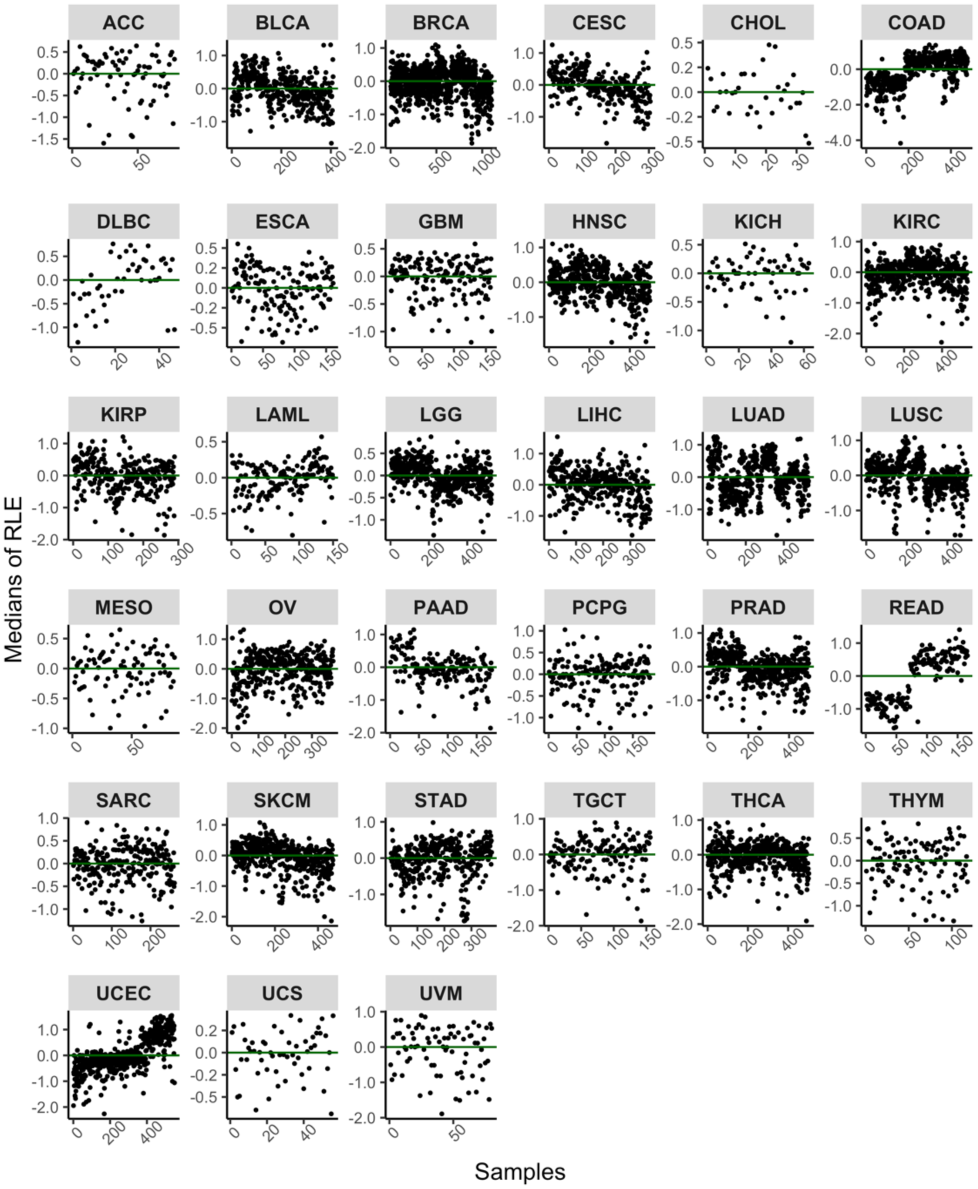
Supplementary Figures



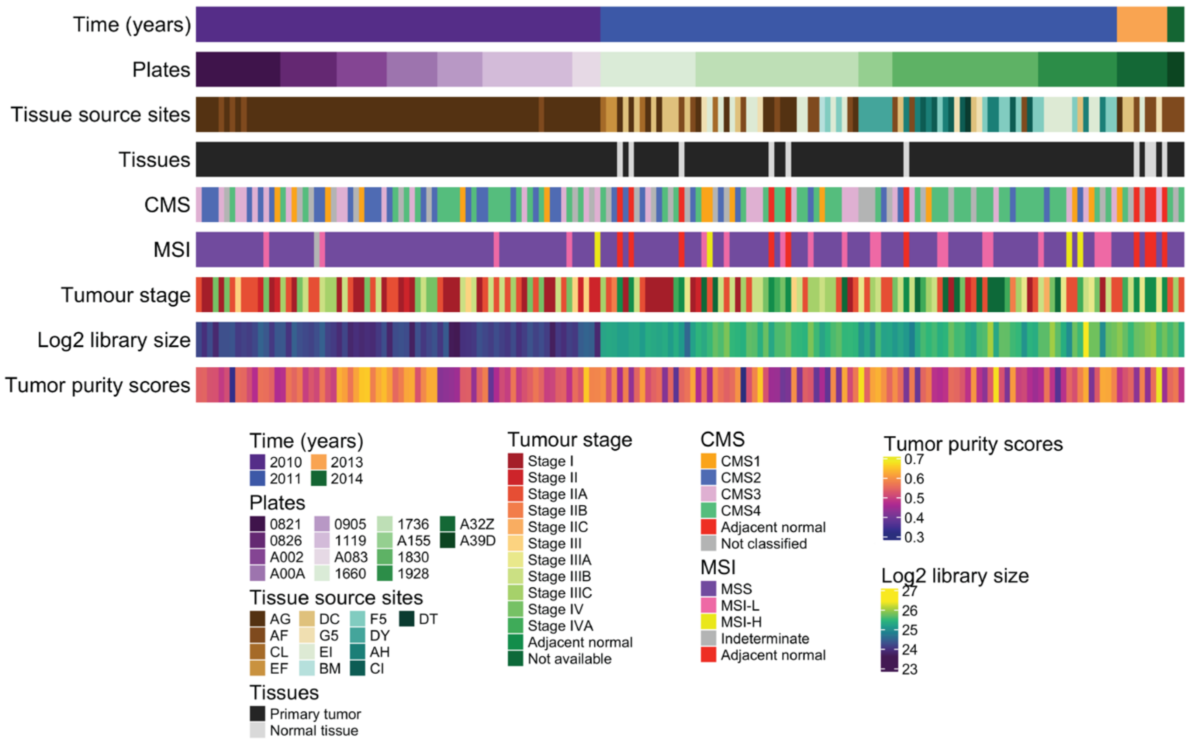
Supplementary Figure 1. Relationships between average gene expression levels and correlation of raw gene-level counts with library size in the TCGA RNA-seq studies. X-axes show Spearman correlation coefficients between $\log_2$ raw gene counts and library size, Y-axes show $\log_2$(average(raw count + 1)) the TCGA RNA-sequencing raw counts. Spearman correlations were computed using $\log_2$ of library size and $\log_2$(gene counts+1).



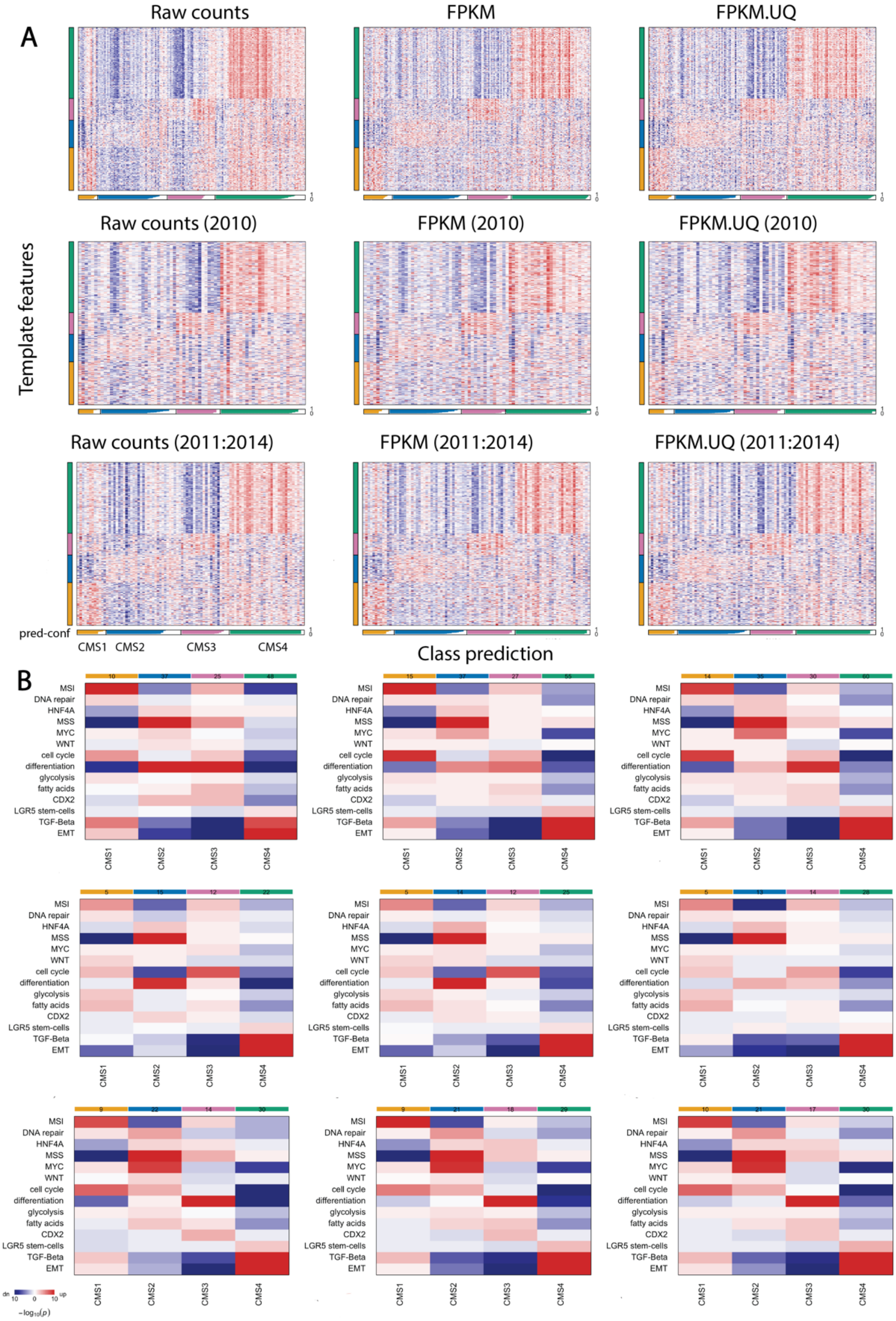
Supplementary Figure 2. Relationships between gene expression and both library size and tumour purity in the TCGA RNA-seq studies. A) Heatmaps show Spearman correlation coefficients between individual gene expression values and library size in the raw (first heatmap), FPKM normalized (second heatmap) and FPKM.UQ normalized (third heatmap) datasets. **B)** Heatmaps show Spearman correlation coefficients between individual gene expression levels and tumour purity scores in raw counts (first heatmap), FPKM normalized (second heatmap), and FPKM.UQ normalized (third heatmap) counts. The order of genes (rows) and studies (columns) are the same across all heatmaps.



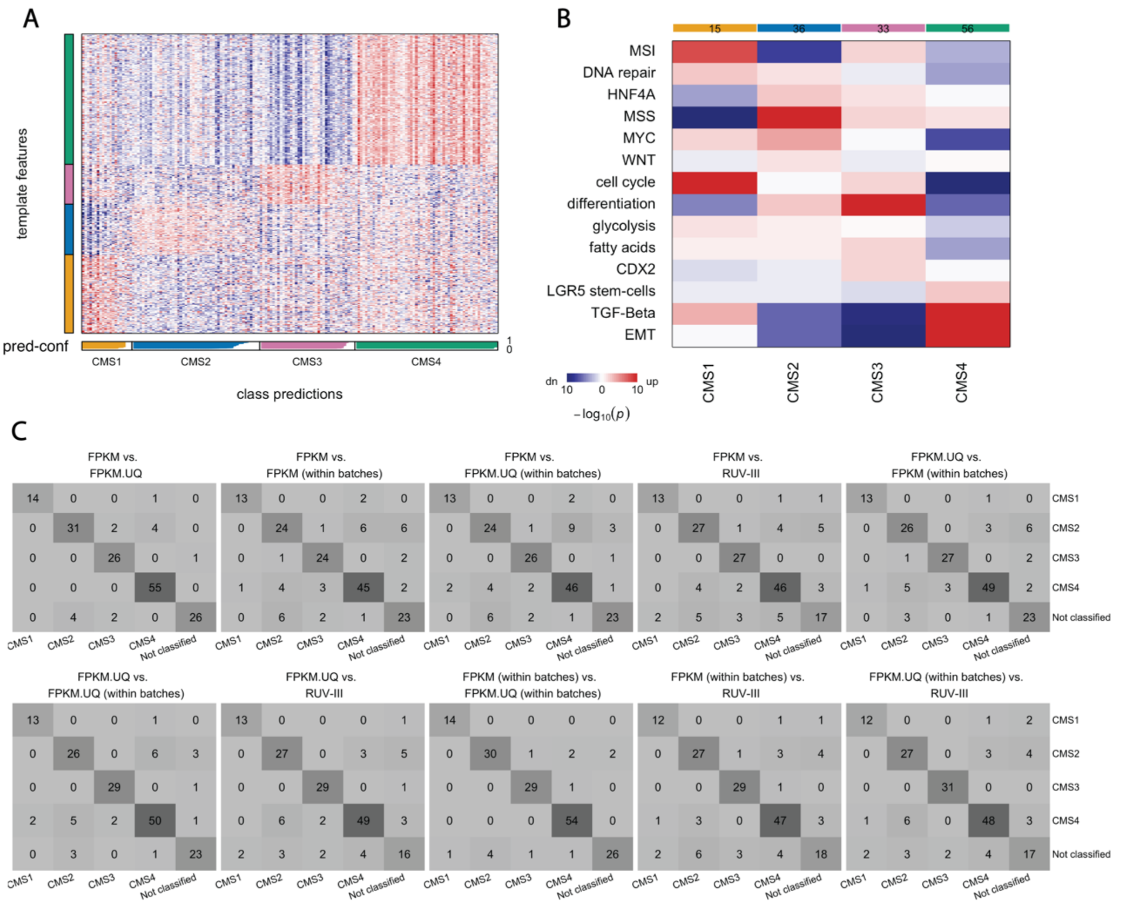
Supplementary Figure 3. Processing date trends in the relative log expression (RLE) plot medians for the TCGA RNA-seq raw counts. Samples in each study are ordered according to the time of running sequencing plates**.** In the absence of unwanted variation, the RLE medians should be centred around zero. Lowly expressed genes were removed prior to creating the RLE plots.



Supplementary figure 4. Outline of the TCGA rectum adenocarcinoma RNA-seq study. A) 176 rectum adenocarcinoma and adjacent normal tissues were collected from 13 tissue source sites (TSS) and distributed across 14 sequencing plates for profiling at 14 time points over a span of 4 years. The consensus molecular subtypes were obtained using the R package CMScaller on the FPKM.UQ normalized data. MSI and tumour stage assignments were obtained from the pathological reports in the TCGA clinical data. The library sizes are calculated after removing lowly expressed genes and $\log_2$ transformed. Tumour purity was estimated using the stromal and immune gene signatures (see Methods).



Supplementary Figure 5. Identification of consensus molecular subtypes (CMS) in the TCGA READ RNA-seq studies using the CMScaller R package. A) Heatmaps show the relative expression levels of CMS marker genes (vertical bar) with classifications indicated below. The height of the white bars gives the prediction confidence values (pred-conf). **B)** Gene set enrichment analysis (Camera) confirms enrichment of known characteristics in each CMS group. Heatmaps show statistical significance with red and blue indicating direction of change.



FPKM vs. FPKM.UQ

	CMS1	CMS2	CMS3	CMS4	Not classified
CMS1	14	0	0	1	0
CMS2	0	31	2	4	0
CMS3	0	0	26	0	1
CMS4	0	0	0	55	0
Not classified	0	4	2	0	26

FPKM vs. FPKM (within batches)

	CMS1	CMS2	CMS3	CMS4	Not classified
CMS1	13	0	0	2	0
CMS2	0	24	1	6	6
CMS3	0	1	24	0	2
CMS4	1	4	3	45	2
Not classified	0	6	2	1	23

FPKM vs. FPKM.UQ (within batches)

	CMS1	CMS2	CMS3	CMS4	Not classified
CMS1	13	0	0	2	0
CMS2	0	24	1	9	3
CMS3	0	0	26	0	1
CMS4	2	4	2	46	1
Not classified	0	6	2	1	23

FPKM vs. RUV-III

	CMS1	CMS2	CMS3	CMS4	Not classified
CMS1	13	0	0	1	1
CMS2	0	27	1	4	5
CMS3	0	0	27	0	0
CMS4	0	4	2	46	3
Not classified	2	5	3	5	17

FPKM.UQ vs. FPKM (within batches)

	CMS1	CMS2	CMS3	CMS4	Not classified
CMS1	13	0	0	1	0
CMS2	0	26	0	3	6
CMS3	0	1	27	0	2
CMS4	1	5	3	49	2
Not classified	0	3	0	1	23

FPKM.UQ vs. FPKM.UQ (within batches)

	CMS1	CMS2	CMS3	CMS4	Not classified
CMS1	13	0	0	1	0
CMS2	0	26	0	6	3
CMS3	0	0	29	0	1
CMS4	2	5	2	50	1
Not classified	0	3	0	1	23

FPKM.UQ vs. RUV-III

	CMS1	CMS2	CMS3	CMS4	Not classified
CMS1	13	0	0	0	1
CMS2	0	27	0	3	5
CMS3	0	0	29	0	1
CMS4	0	6	2	49	3
Not classified	2	3	2	4	16

FPKM (within batches) vs. FPKM.UQ (within batches)

	CMS1	CMS2	CMS3	CMS4	Not classified
CMS1	14	0	0	0	0
CMS2	0	30	1	2	2
CMS3	0	0	29	1	0
CMS4	0	0	0	54	0
Not classified	1	4	1	1	26

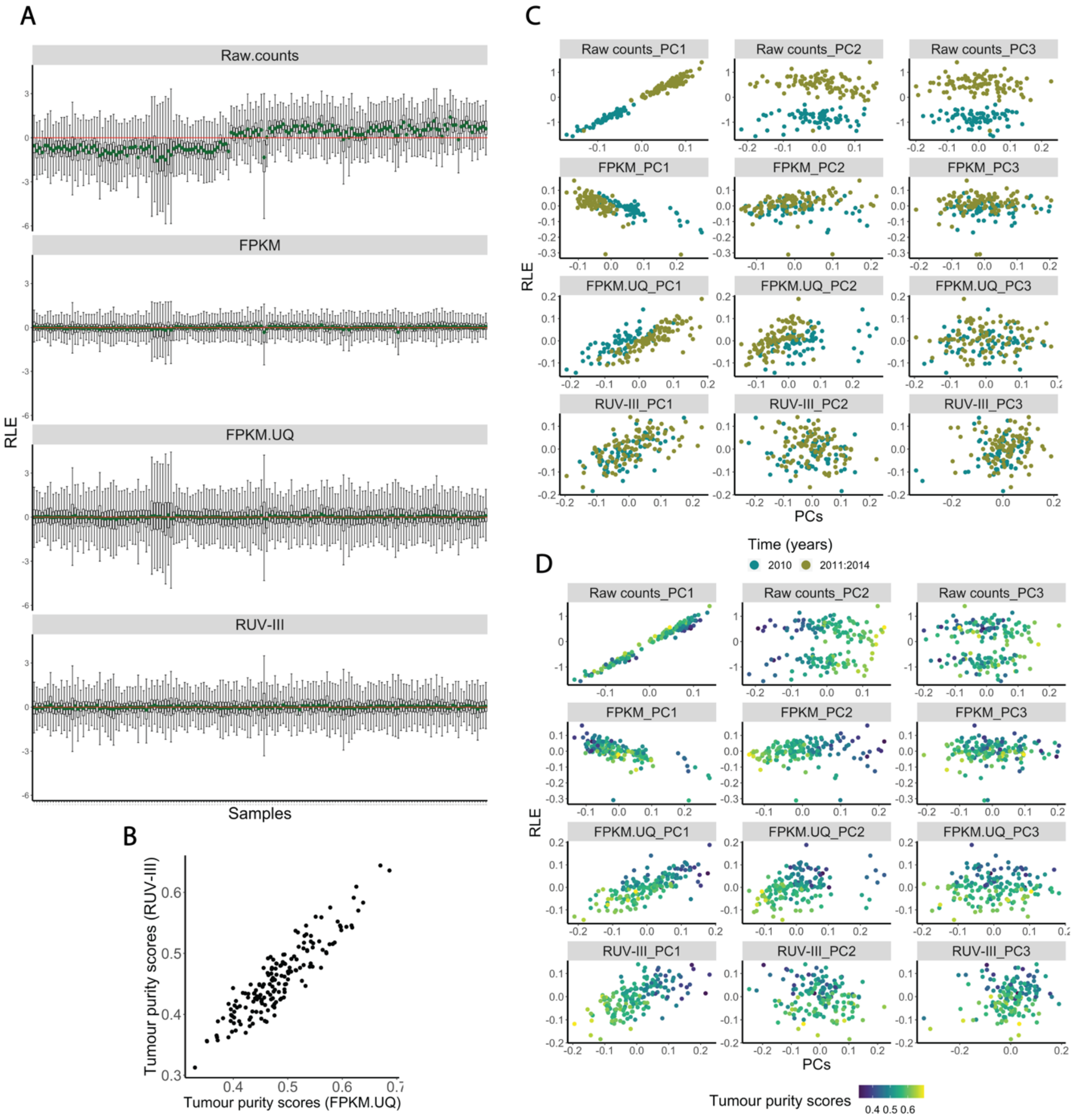
FPKM (within batches) vs. RUV-III

	CMS1	CMS2	CMS3	CMS4	Not classified
CMS1	12	0	0	1	1
CMS2	0	27	1	3	4
CMS3	0	0	29	1	0
CMS4	1	3	0	47	3
Not classified	2	6	3	4	18

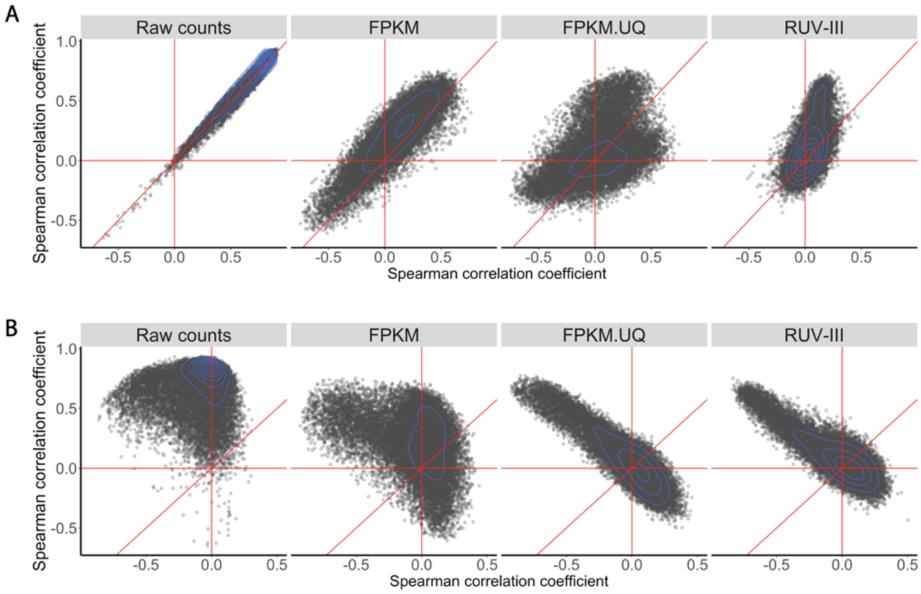
FPKM.UQ (within batches) vs. RUV-III

	CMS1	CMS2	CMS3	CMS4	Not classified
CMS1	12	0	0	1	2
CMS2	0	27	0	3	4
CMS3	0	0	31	0	0
CMS4	1	6	0	48	3
Not classified	2	3	2	4	17

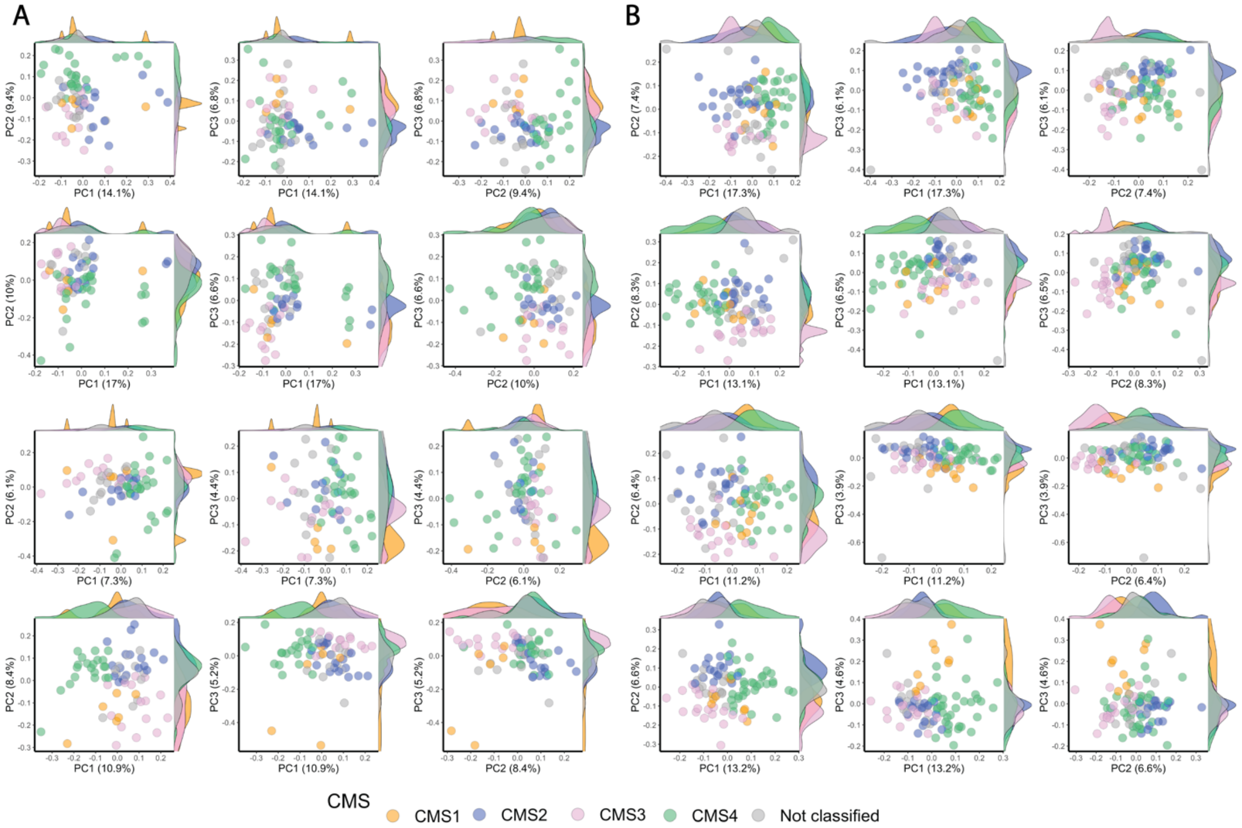
Supplementary Figure 6. Identification of consensus molecular subtypes (CMS) in the RUV-III normalized data using the CMScaller R package. A) Heatmap shows the relative expression levels of CMS marker genes (vertical bar) with classifications indicated below. The height of the white bars gives the prediction confidence values (pred-conf). **B)** Gene set enrichment analysis (Camera) confirms enrichment of known characteristics in each CMS group. Heatmaps show statistical significance with red and blue colours indicating direction of change. **C)** The tables show CMS calls obtained from different strategies and datasets normalized by different methods.



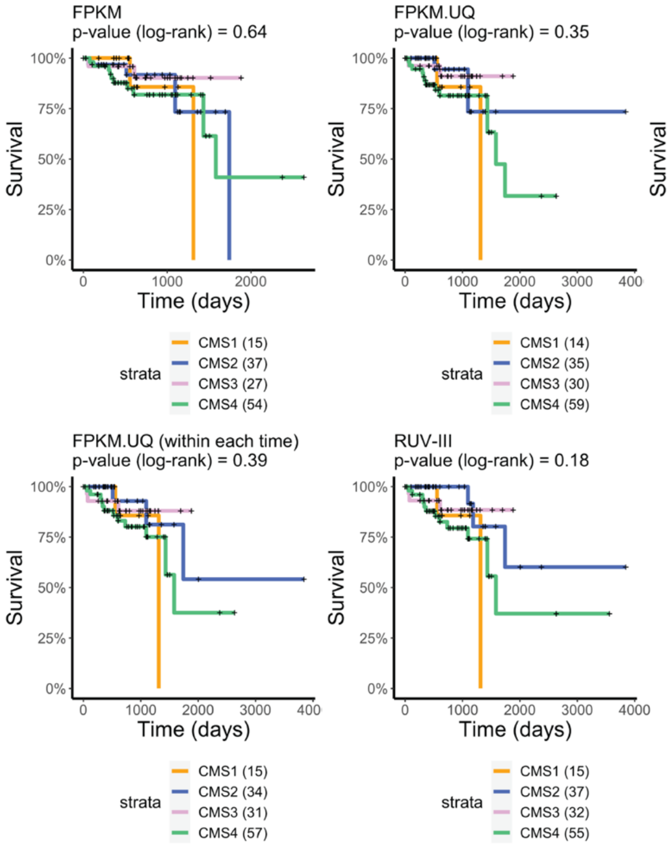
Supplementary Figure 7. RLE plots, tumour purity estimates and relationships between the RLE plot medians and principal components in the TCGA READ RNA-sequencing data. A) RLE plots of raw counts, TCGA and RUV-III normalized data. The heavy middle line represents the median, the box shows the inter-quartile range (IQR), and the upper and lower whiskers extend from the hinges no further than 1.5 ´ IQR. **B)** Scatter plot shows tumour purity scores obtained from the TCGA FPKM.UQ and RUV-III normalized data. Note, no attempt was made to remove tumour purity variation by RUV-III normalization**. C)** The scatter plots show the relationships between the first three principal components and the RLE plot medians of raw counts, TCGA and RUV-III normalized data coloured by key time intervals. **D)** The same as C, coloured by the tumour purity scores.



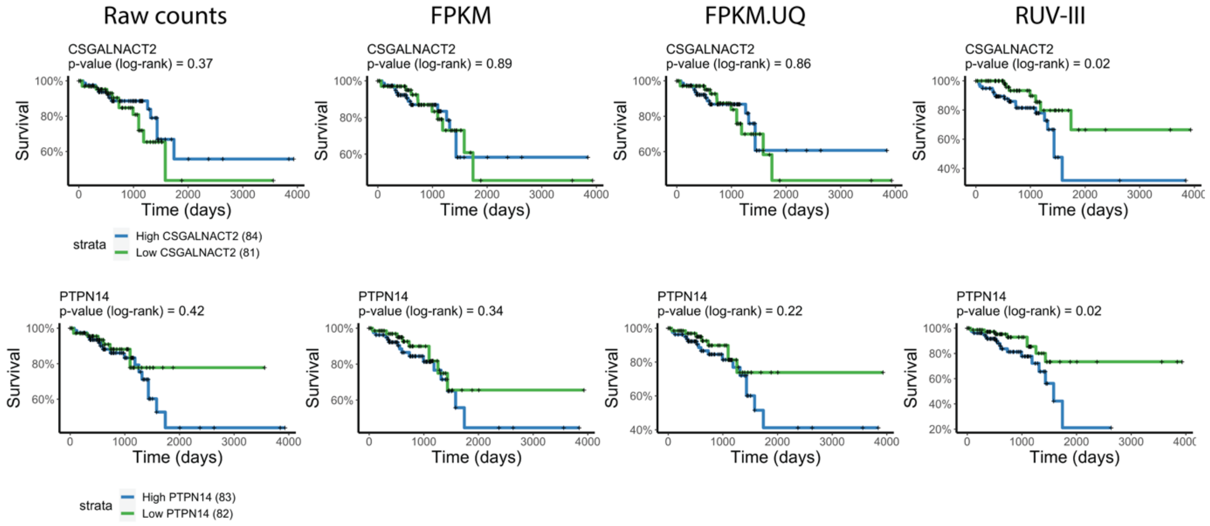
Supplementary Figure 8. Spearman correlation analysis between individual gene expression values and the RLE medians, library size, and tumour purity in the TCGA READ RNA-seq datasets. A) Scatter plots show Spearman correlation coefficients between individual gene expression levels and the RLE plot medians (Y-axis) and library sizes (X-axis) in the different datasets. **B)** Scatter plots display Spearman correlation coefficients between individual gene expression values and RLE plot medians (Y-axis), and tumour purity (X-axis) in the different datasets. The diagonal lines are the 45-degree line.



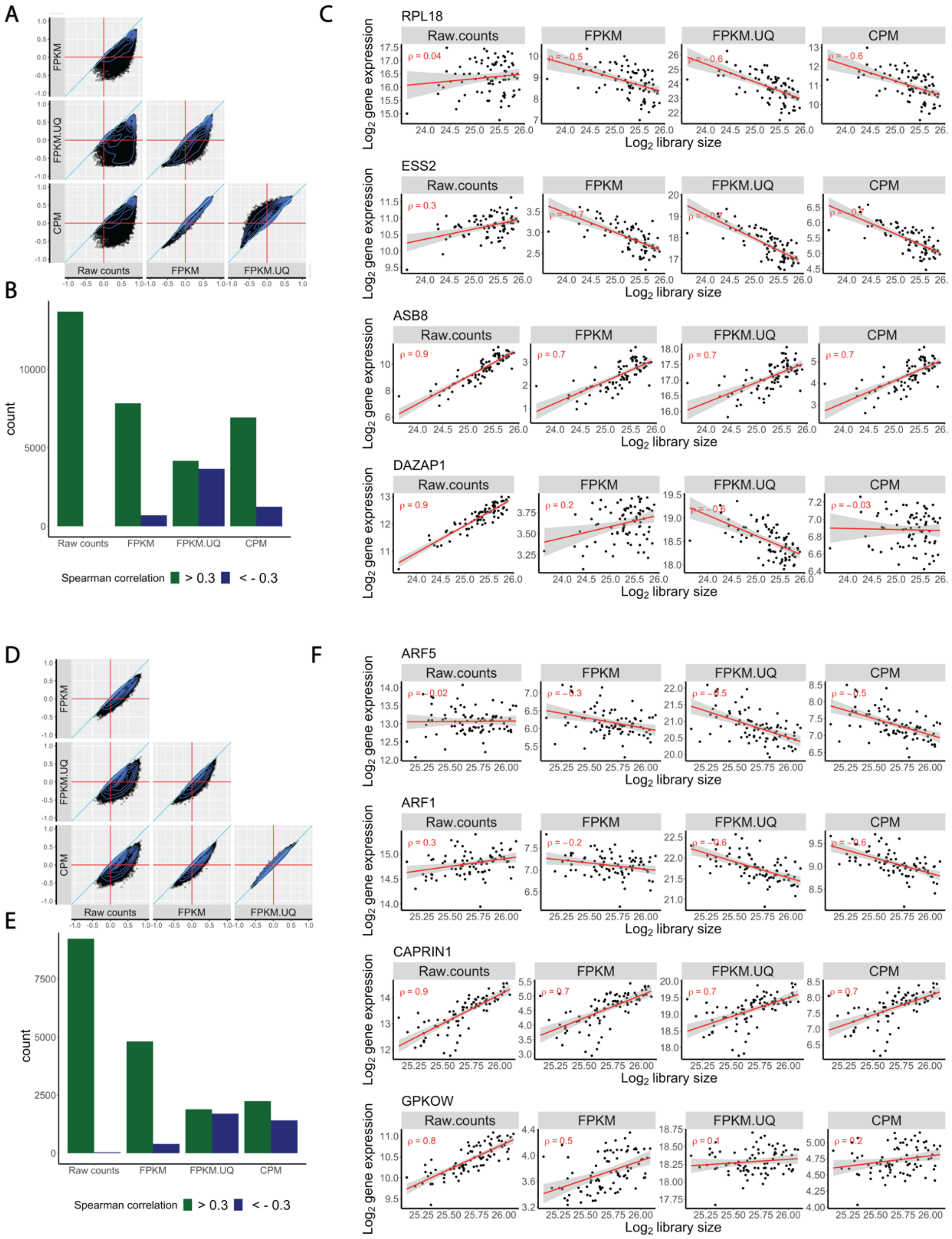
Supplementary Figure 9. PCA plots coloured by CMS with each key time interval in the TCGA READ RNA-seq data normalized by different methods. A) PCA plots of different datasets (from top to bottom: raw counts, FPKM, FPKM.UQ and RUV-III normalized data for samples profiled in 2010. **B)** Same as A, for samples profiled in 2011-2014. The CMS were obtained separately within each time interval.



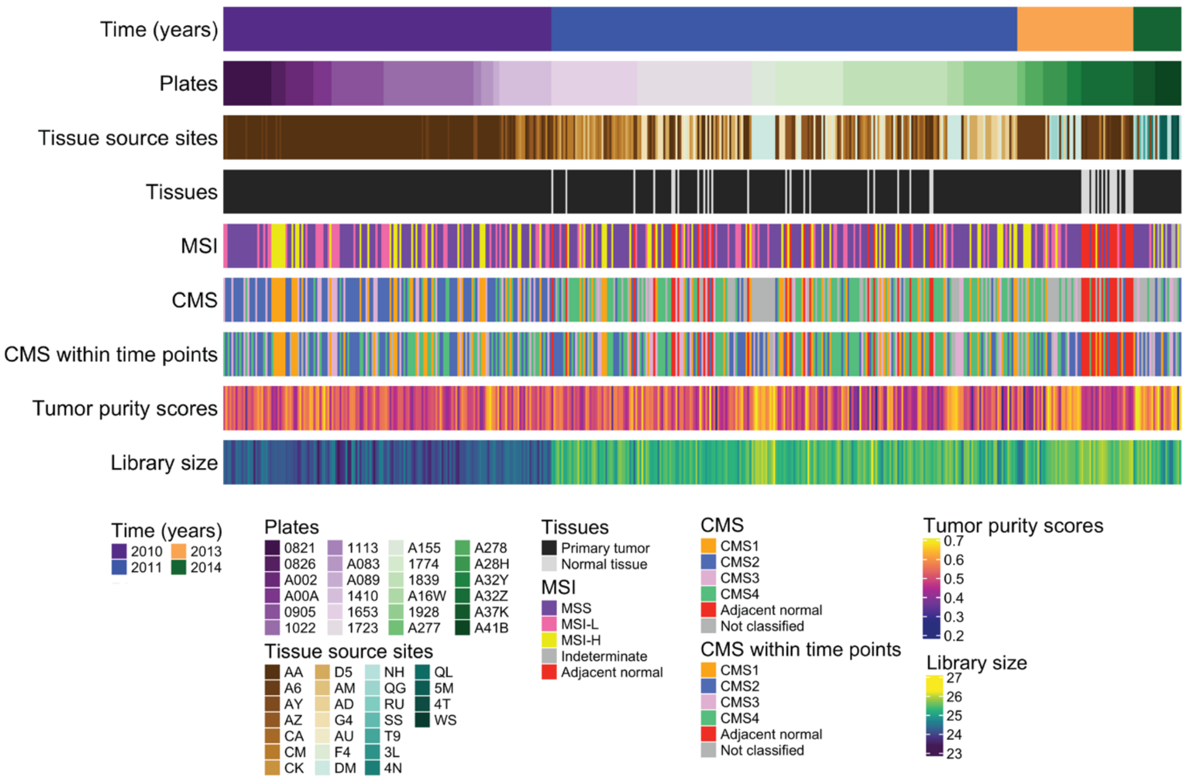
Supplementary Figure 10. Kaplan-Meier survival curves by CMS using the TCGA READ RNA-seq data normalized by different methods. Prognostic value of CMS with Kaplan-Meier survival analysis in FPKM, FPK.UQ and RUV-III normalized data for overall survival. The CMS for FPKM and FPKM.UQ data were identified either in the full set of samples or within each key time interval**.**



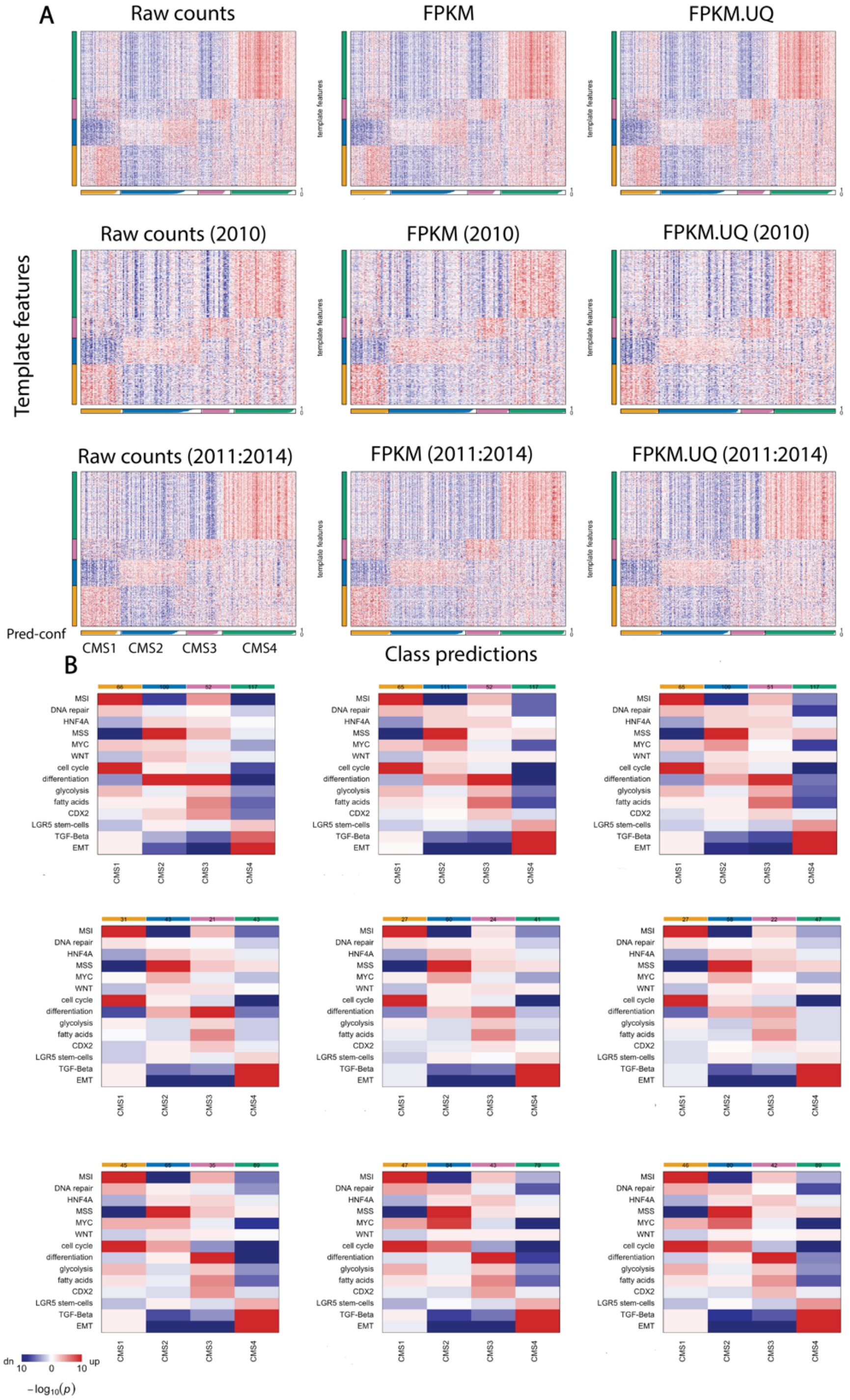
Supplementary Figure 11. Association of the CSGALNACT2 and PTPN14 gene expression levels with overall survival in the differently normalized TCGA READ RNA-seq data.



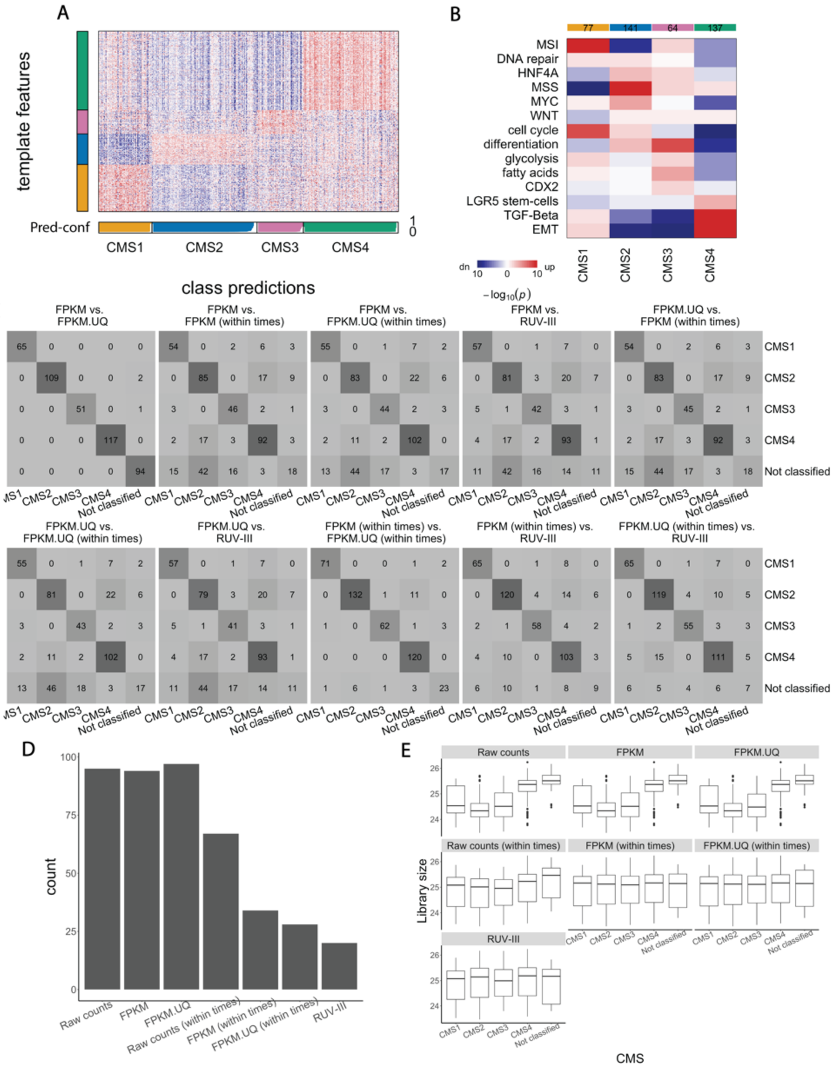
Supplementary Figure 12. Performance of library size normalization in the TCGA UVM and KICH RNA-seq data. A) scatter plots show Spearman correlation coefficients between individual gene expression and log$_2$ of library size in differently normalized TCGA UVM data. **B)** bar plots show the number of genes that show high positive and negative correlations with library size in the TCGA UVM RNA-seq data normalized by different methods. **C)** Relationship of several gene expression with library size in differently normalized TCGA UVM RNA-seq data. Confidence intervals are shown as grey bands. **D**, **E** and **F** are as **A, B** and **C** for the TCGA KICH RNA-seq data. Each of these datasets were generated on a single plate.



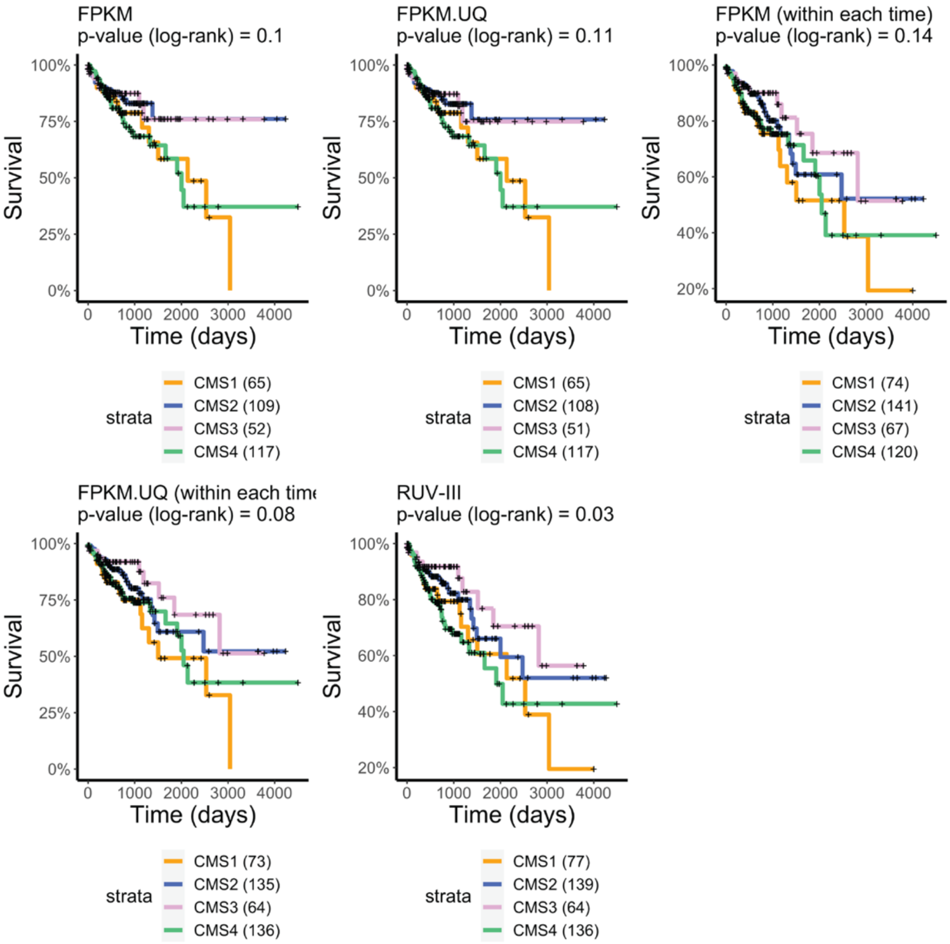
Supplementary Figure 13. Outline of the TCGA colon adenocarcinoma (COAD) RNA-seq study. A) 479 colon adenocarcinoma and adjacent normal tissues were collected from 25 tissue source sites (TSS) and distributed to 24 sequencing-plates for profiling at 20 different times over a span of 4 years. The microsatellite instability (MSI) and tumour stage assignments were obtained from the TCGA pathological reports. The consensus molecular subtypes (CMS) were identified using the CMScaller R package on the FPKM.UQ normalized data in two different ways, across all samples and within key time intervals (2010 and 2010-2014). The sample library sizes are calculated after removing lowly expressed genes and are $\log_2$ transformed. The tumour purity scores are obtained from 1- stromal and immune scores (see Methods).



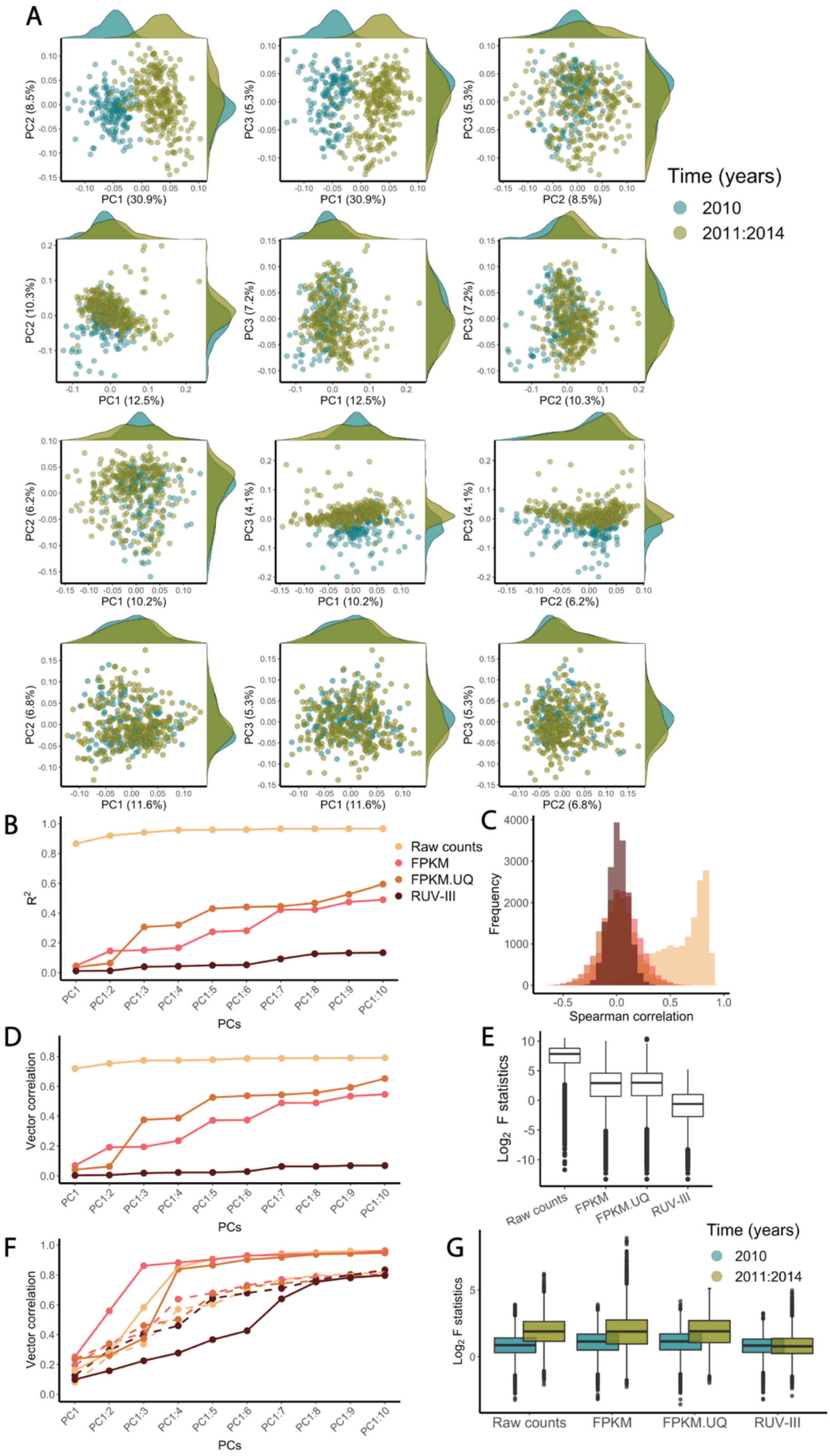
Supplementary Figure 14. Identification of consensus molecular subtypes in the TCGA COAD RNA-seq data using the CMScaller R package. A) Heatmaps show the relative expression levels of subtype marker genes (vertical bar) with classifications. CMS1-CMS4, designated below. The height of the white bars gives the prediction confidence values (pred-conf). The CMS classification were performed across all samples and within each key time intervals (2010 and 2010-2014) to assess the impact of batch effects. **B)** Gene set testing (Camera) confirms enrichment of known characteristics in each CMS group. Heatmaps show statistical significance and red and blue colours indicating direction of change.



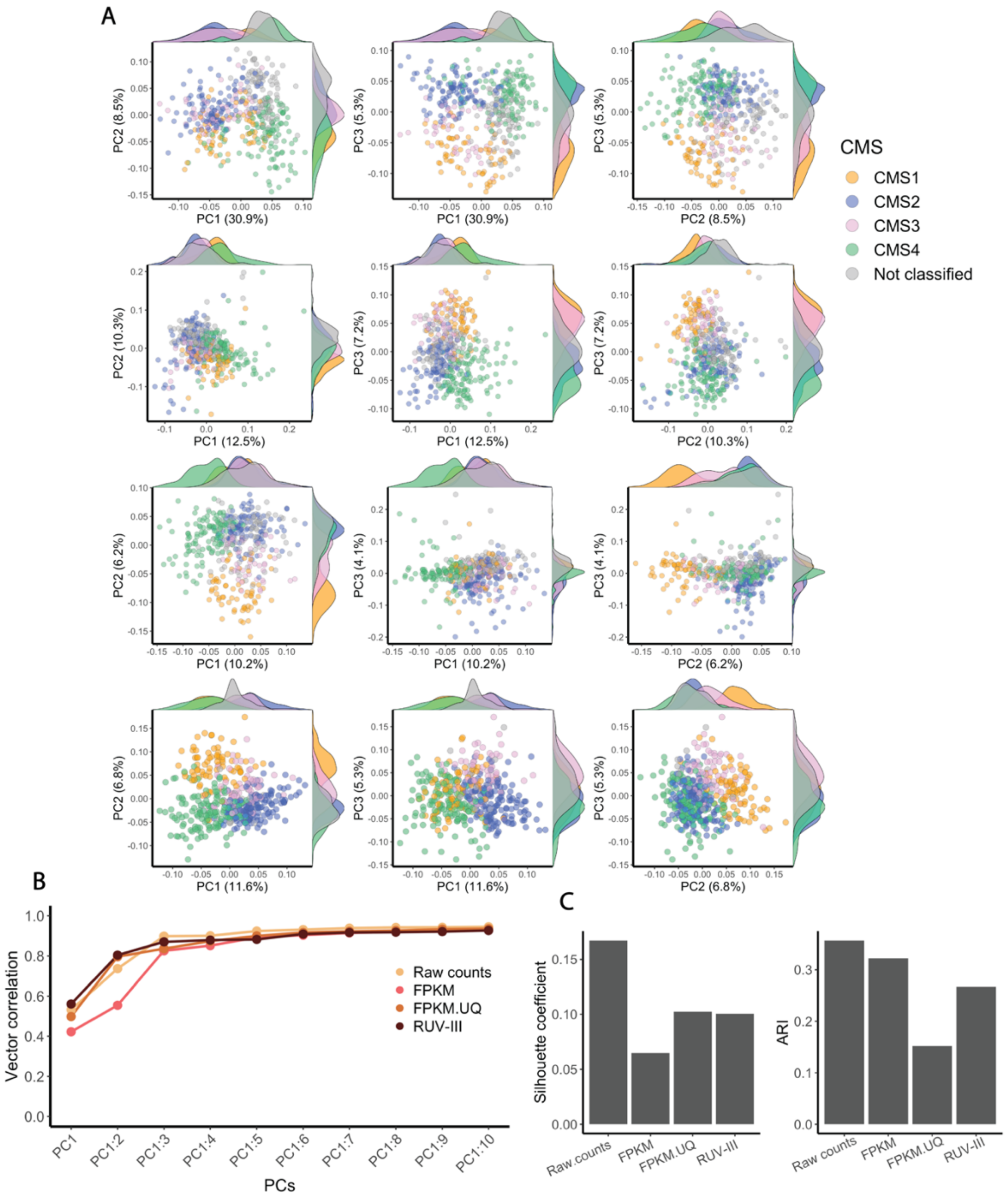
Supplementary figure 15. Identification of consensus molecular subtypes (CMS) in the TCGA COAD RNA-sequencing study using the RUV-III normalized data and comparing CMS obtained after normalizing by different methods. A) Heatmap shows the relative expression levels of subtype marker genes (vertical bar) with classifications. CMS1-CMS4, designated below. The height of the white bars gives the prediction confidence values (pred-conf). B**)** Gene set testing (Camera) confirms enrichment of known characteristics in each CMS group. Heatmaps show statistical significance with red and blue colours indicating direction of change**. C)** Comparisons between CMS calls obtained from datasets from different normalization and conditions. **D)** Bar plot shows the number of un-classified samples obtained by CMScaller in different datasets. **E)** Plots show the relationship between library size ($\log_2$) and CMS calls in different datasets. The heavy middle line represents the median, the box shows the inter-quartile range (IQR), the upper and lower whiskers extend from the hinges no further than 1.5 ´ IQR, and any outliers beyond the whiskers are shown as points. The number of samples in each subtype and differently normalized datasets is shown in C.



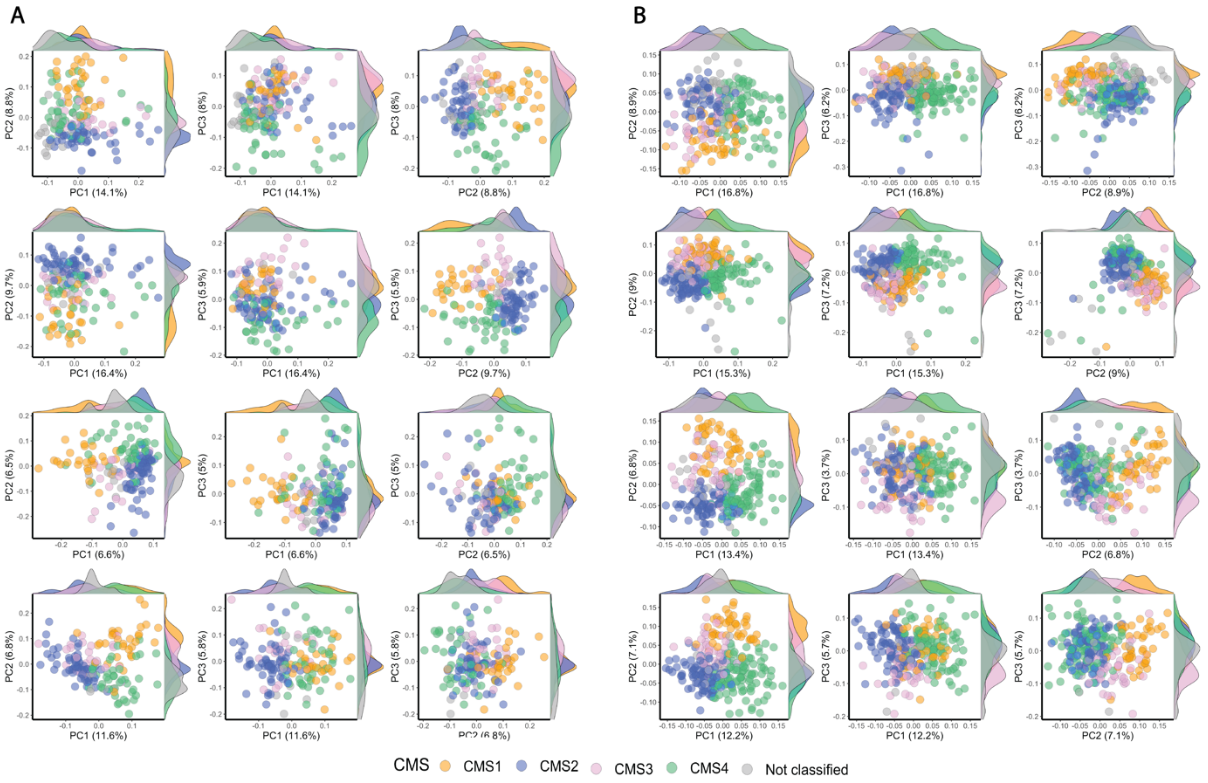
Supplementary figure 16. Kaplan-Meier survival analyses by consensus molecular subtype in the TCGA COAD RNA-seq normalized by different methods.



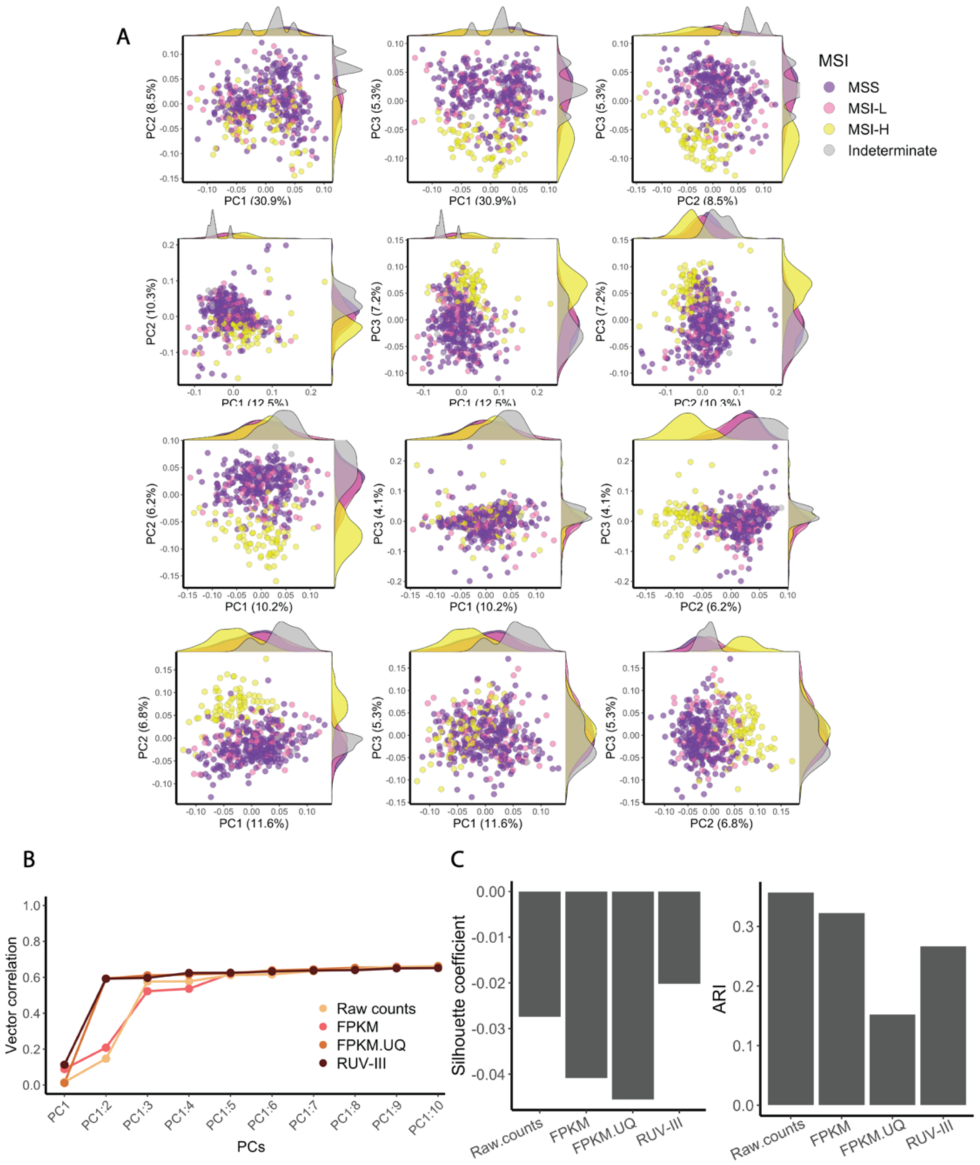
Supplementary figure 17. Performance of different normalization methods in removing library size and plate effects in the TCGA COAD RNA-seq data. A) From top row to bottom: Scatter plots of first three principal components for the raw counts, FPKM, FPKM.UQ and RUV-III normalized data coloured by key time intervals (2010, 2011-2014). **B)** A plot showing the R-squared (R^2) of linear regression between library size and up to the first 10 principal components (cumulatively) of different datasets. **C)** Histograms of Spearman correlation coefficients between the gene expression levels and library size. **D)** A plot showing the vector correlation between the key time intervals and up to the first 10 principal components (cumulatively). **E)** Boxplots of $\log_2$ F statistics obtained from ANOVA between gene expression with key time intervals being the factor (n=16,479 genes). **F)** A plot showing the vector correlation between plates and up to the first 10 principal components (cumulatively) within each key time interval. **G)** Boxplots of $\log_2$ F statistics obtained from ANOVA between gene expression and plates within key time intervals (n=16,479 genes for each boxplot). In the boxplots (panels E and G), the heavy middle line represents the median, the box shows the inter-quartile range (IQR), the upper and lower whiskers extend from the hinges no further than 1.5 × IQR, and any outliers beyond the whiskers are shown as points.



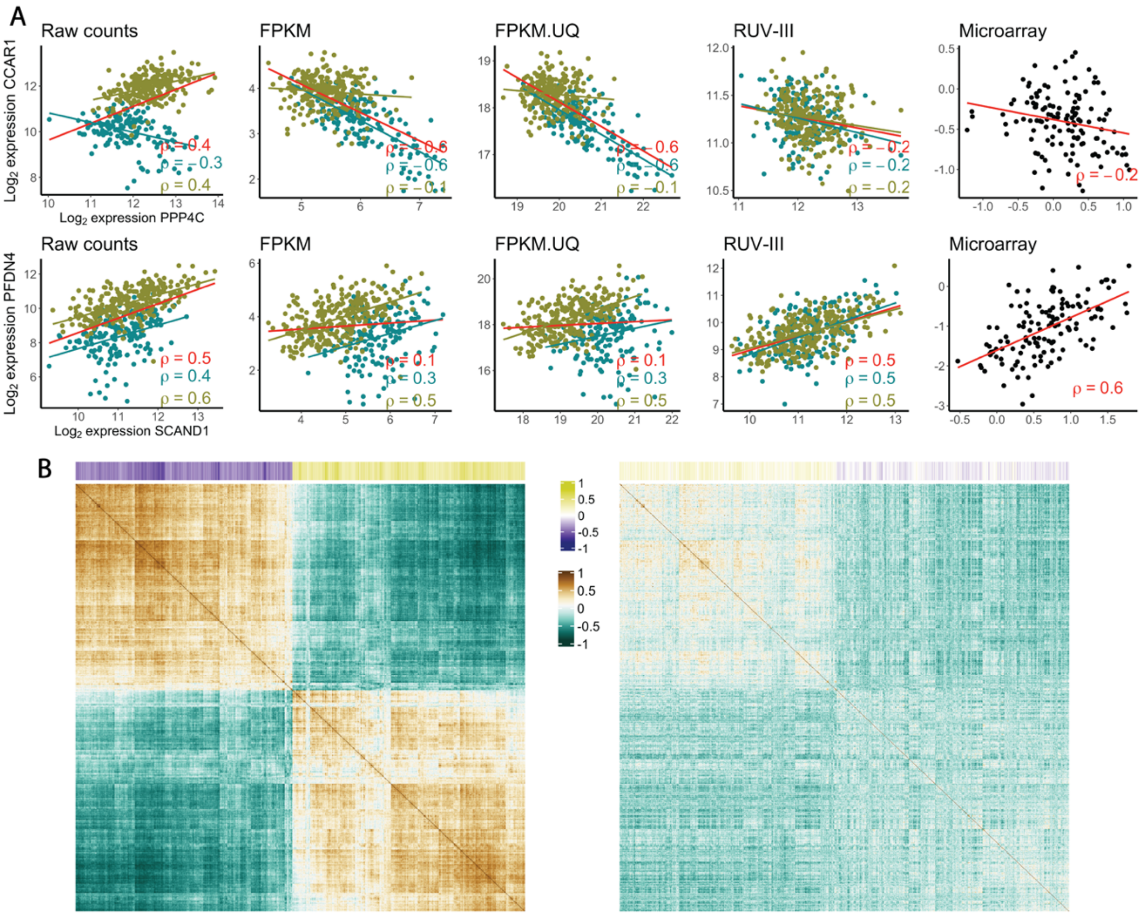
Supplementary Figure 18. Performance of the different normalization methods in preserving the consensus molecular subtypes in the TCGA COAD RNA-seq data. A) From top row to bottom: scatter plots of the first three principal components colored by CMS for the TCGA COAD raw counts, FPKM, FPKM.UQ and RUV-III normalized data. The CMS were obtained for each dataset separately. **B)** A plot showing the vector correlation between CMS and up to the first 10 principal components (cumulatively) in the differently normalized datasets. **C)** Silhouette coefficients and ARI index for showing the separation of CMS subtypes in the differently normalized datasets.



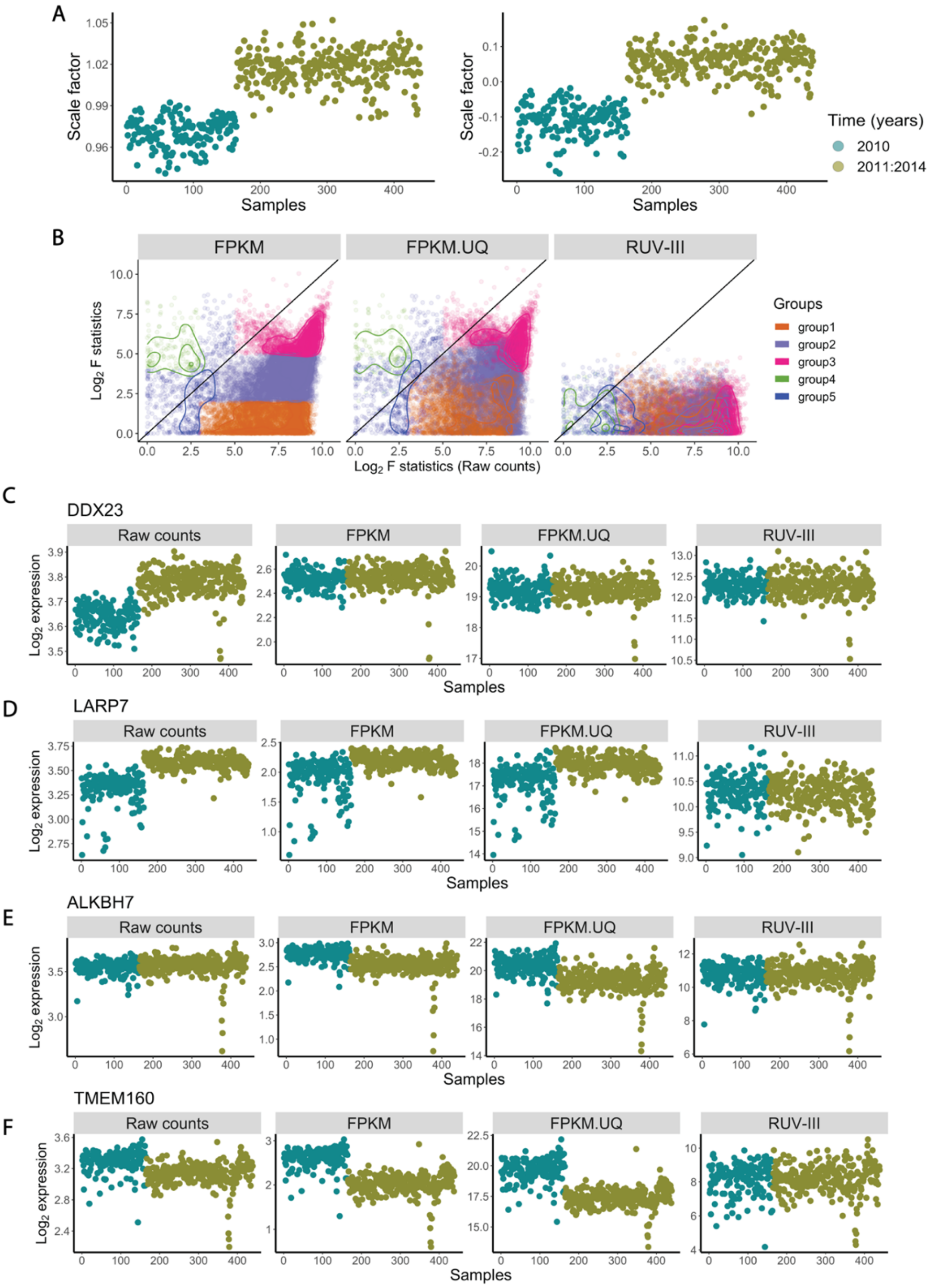
Supplementary figure 19. PCA plots coloured by CMS within each key time interval in the TCGA COAD RNA-seq data normalized by different methods. A) Scatter plots of the first three principal components obtained from samples profiled in 2010 (from top to bottom: the TCGA raw counts, FPKM, FPKM.UQ, and RUV-III normalized datasets). **B)** Same as A, for samples profiled in 2011-2014.



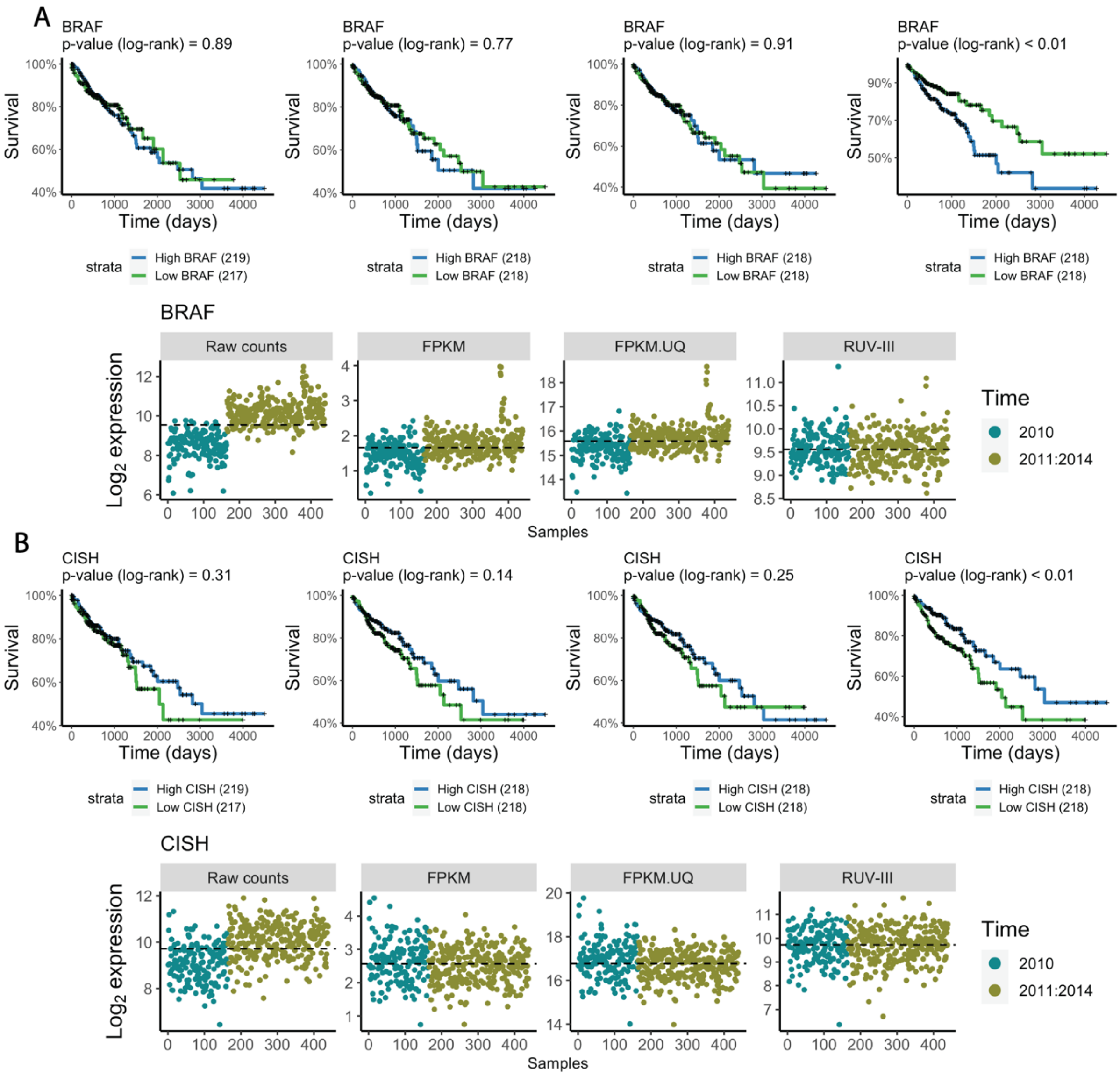
Supplementary figure 20. Performance of different normalization methods in preserving the microsatellite instability (MSI) clusters in the TCGA COAD RNA-seq data. A) Scatter plots of the first three principal components colored by MSI for the TCGA COAD raw counts, FPKM, FPKM.UQ and RUV-III normalized data. The MSI details (MSS: microsatellite stable, MSI-L = microsatellite instability-low, MSI-H = microsatellite instability-high) were obtained from the TCGA pathological data. **B)** Plot showing the vector correlation coefficient between MSI and up to the first ten principal components (cumulatively). **C)** Silhouette coefficients and ARI index for exhibiting the separation of MSI subtypes in different datasets



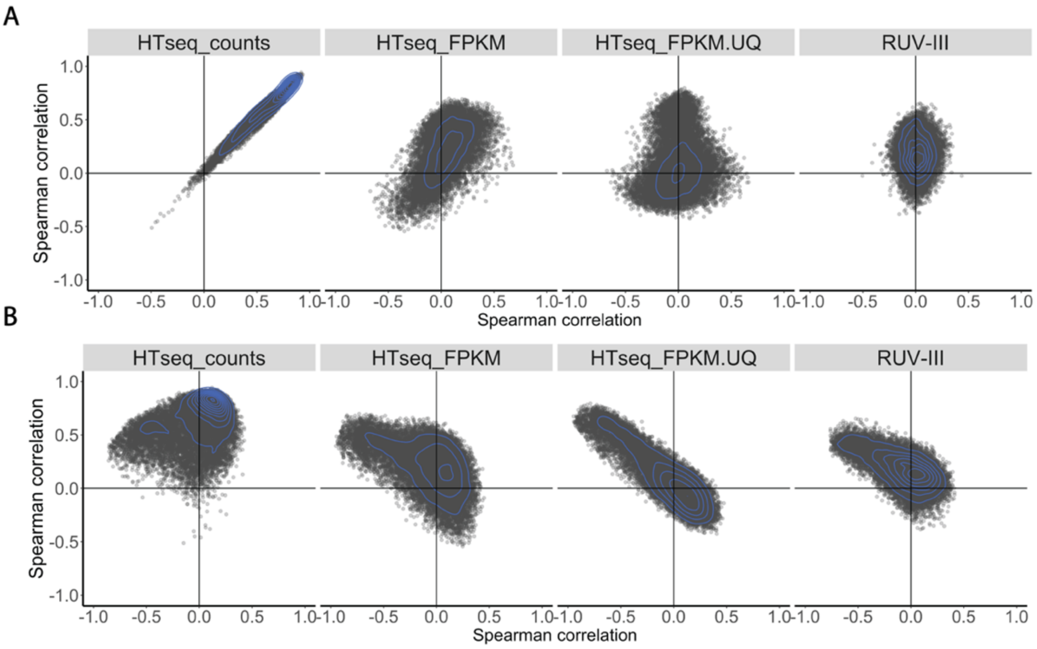
Supplementary figure 21. Impact of unwanted variation on gene co-expression analyses in the TCGA COAD RNA-seq data. A) First row: scatter plots between the expression levels of the *CCAR1* and *PPP4C* genes in the raw and differently normalized counts. The red line indicates overall association between the expression levels, while the green and yellow lines show the associations seen in the 2010 samples and the rest, respectively. **B)** First heatmap shows correlation matrix of the 1300 genes whose expression levels have the highest correlation with library size in the FPKM.UQ normalized data. The coloured bars at the top of heatmaps show the correlation of individual gene expression levels with library size. Second heatmap: same as the first one, for the RUV-III normalized data. The order of rows and columns in both correlation matrix are the same.



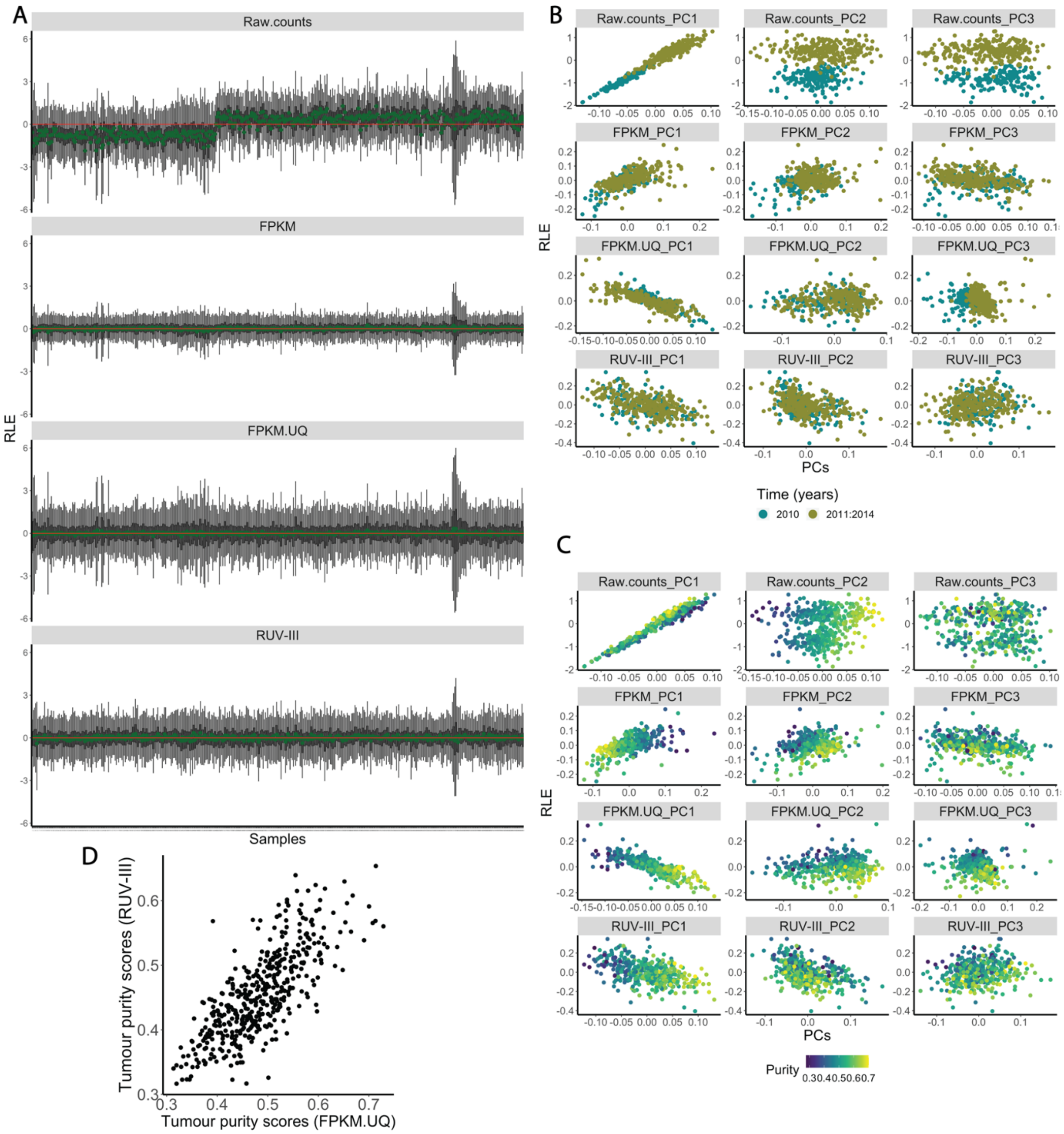
Supplementary figure 22. Relationship between gene level ($\log_2$) counts and ($\log_2$) library size in the TCGA COAD RNA-seq. A) Plots shows the global scale factors obtained by sample library sizes (first plot) and upper-quartiles (second plot) of the COAD raw counts against time. **B)** Scatter plots of the $\log_2$ F-value obtained from ANOVA analyses of gene expression levels with the major time variation: 2010 vs all other years; ($\log_2$) raw READ counts on the horizontal axes of all plots and differently normalized counts vertically. **C)** Expression patterns of four genes whose counts have different relationships with the global scaling factors calculated from the COAD raw count data.



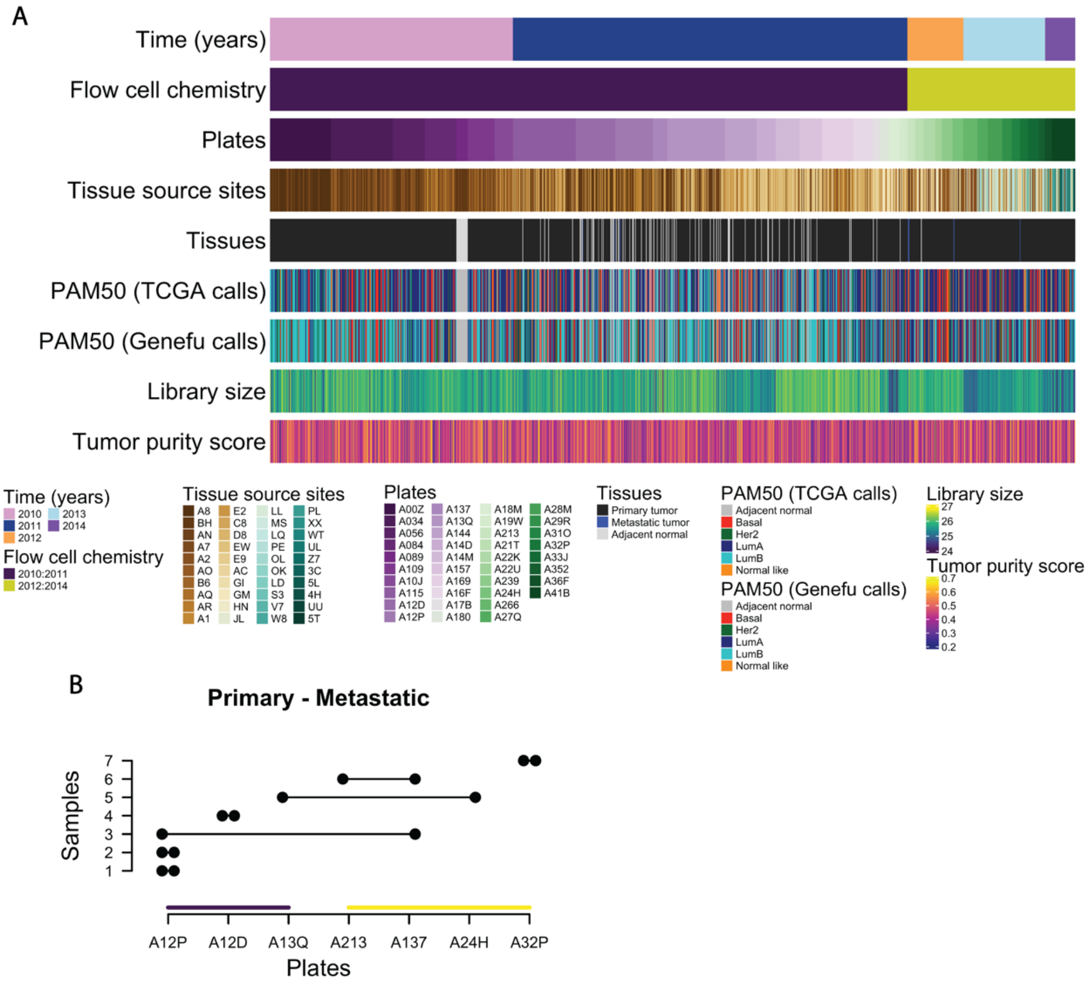
Supplementary figure 23. Association of the BRAF and CISH gene expression levels with overall survival in the differently normalized TCGA COAD datasets. A) First row: Kaplan Meier survival analysis shows the association between the BRAF gene expression and overall survival in differently normalized TCGA COAD datasets. Second row: expression patterns of the BRAF gene expression across samples in differently normalized TCGA COAD datasets. **B)** same as A, for the CISH gene**.**



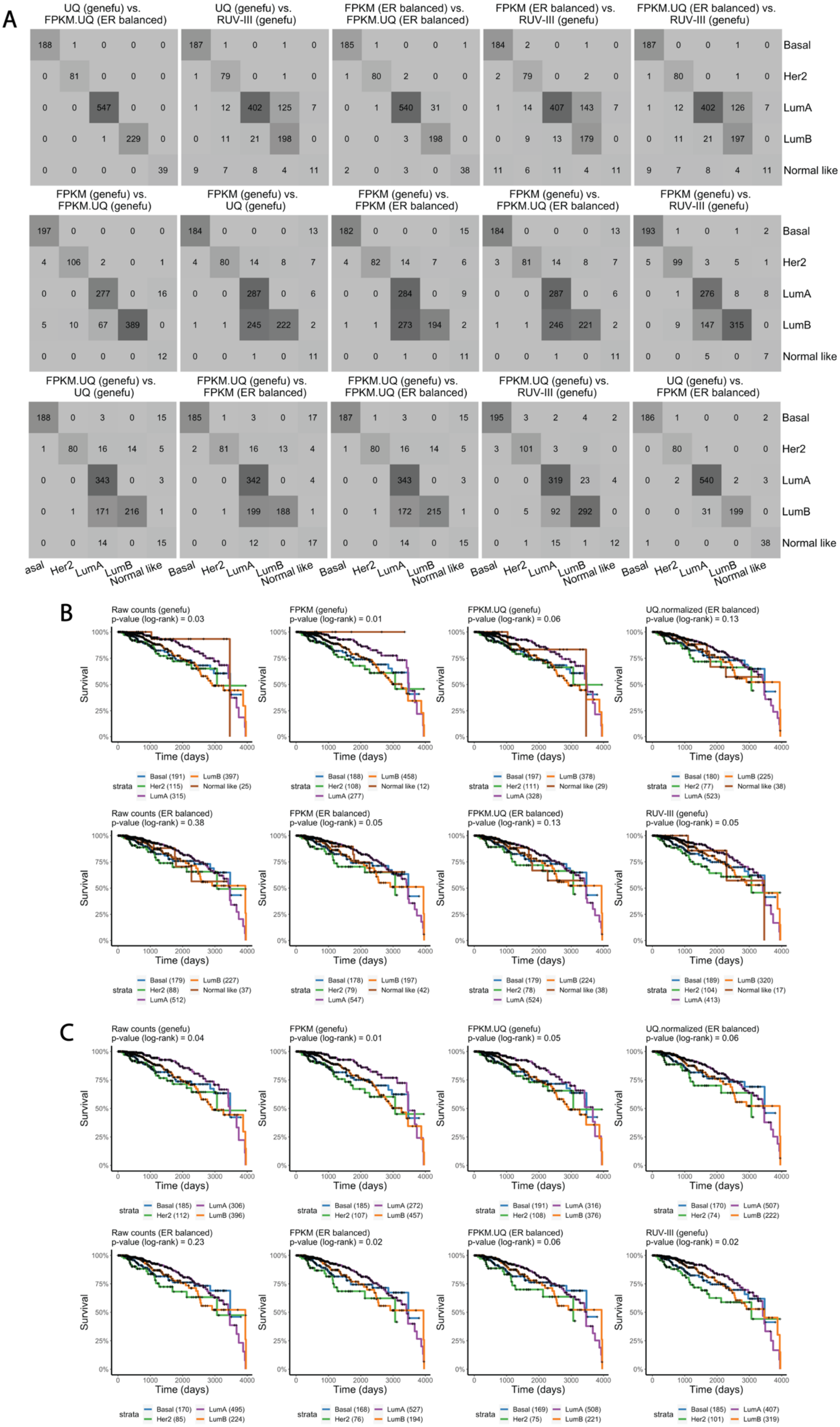
Supplementary figure 24. Spearman correlation analyses between gene expression levels and RLE plot medians, library size, and tumour purity in the TCGA COAD raw and differently normalized counts. A) Scatter plots of Spearman correlations between individual gene expression levels with RLE plot medians (Y-axis) and library size (X-axis) in the different datasets. **B)** Scatter plot of Spearman correlation coefficients between individual gene expression levels with RLE plot medians (Y-axis) and tumour purity (X-axis) in the different datasets.



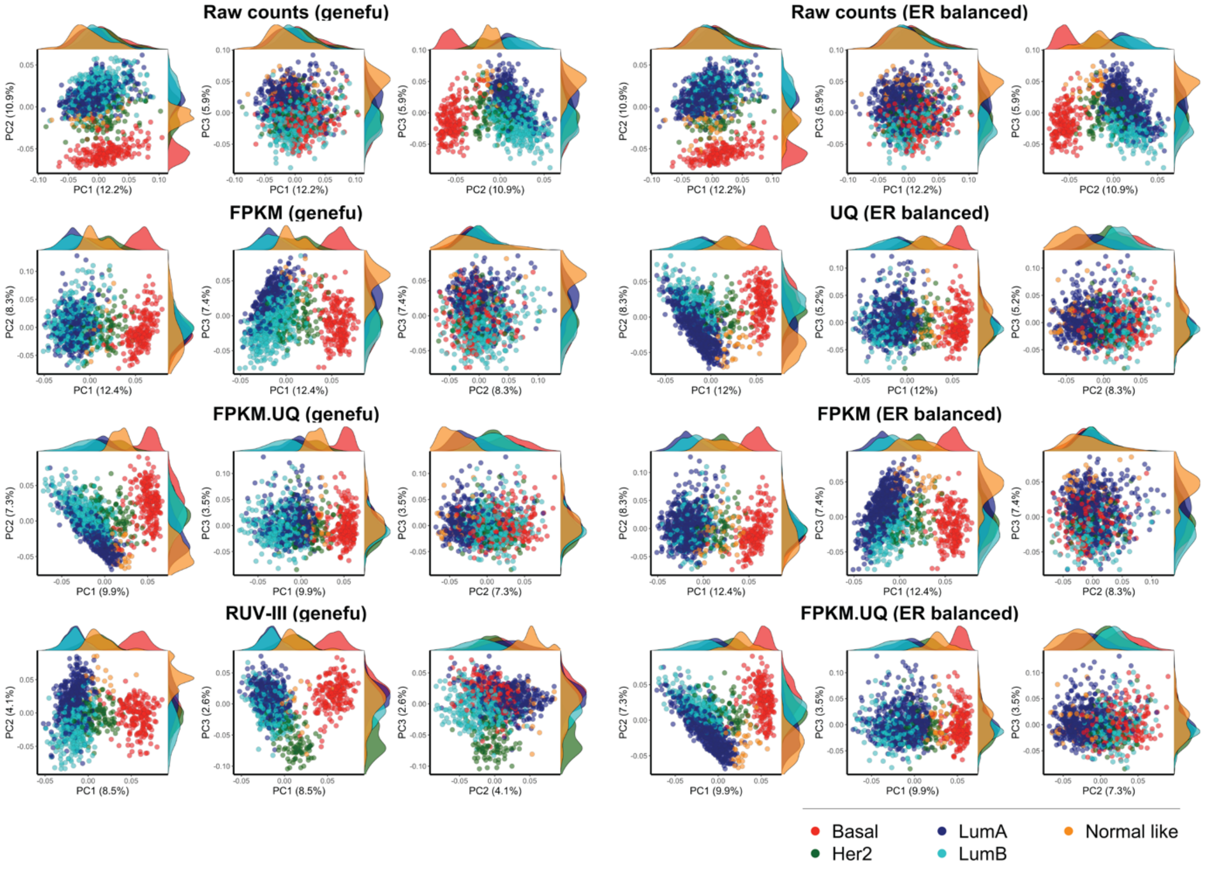
Supplementary figure 25. RLE plots and relationships between the RLE plot medians and principal components in the raw and differently normalized counts of the TCGA COAD RNA-seq data. A) RLE plots of raw counts, TCGA and RUV-III normalized datasets. The heavy middle line represents the median, the box shows the inter-quartile range (IQR), and the upper and lower whiskers extend from the hinges no further than 1.5 ´ IQR. **B)** Scatter plots show the relationships between the first three principal components and the RLE plot medians for raw and differently normalized counts, coloured by different times. **C)** As for B, coloured by tumour purity. **D)** Tumour purity estimates for the TCGA FPKM.UQ and RUV-III normalized data. Note, that the RUV-III normalization did not aim to remove the effect of tumour purity variation.



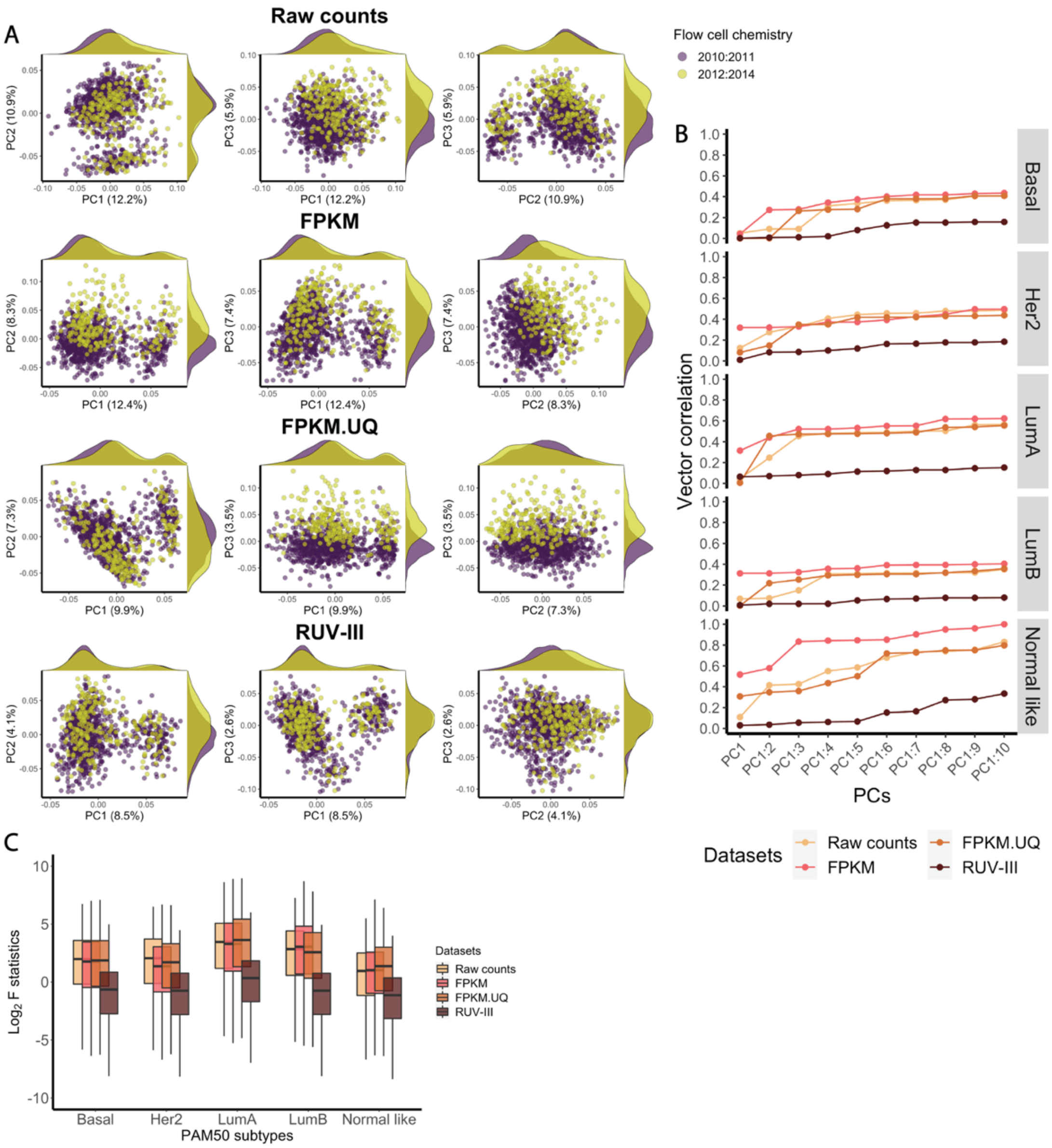
Supplementary figure 26. a) 1180 cancer and adjacent normal breast tissues were collected from 40 tissue source sites (TSS) and distributed unevenly to 38 sequencing plates for profiling at 40 different time points over a span of 5 years from 2010 to 2014. Most samples (958 out of 1212 samples) were profiled in the period 2010-2011 using the first flow cell chemistry, and the rest of the samples were profiled in 2012-2014 using another flow cell chemistry. PAM50 subtypes were called using two different approaches. **b)** The distributions of 7 paired primary-metastatic samples across different sequencing plates and the two flow cell chemistries. Solid points connected by a line show primary-metastasis pairs**.**


A
UQ (genefu) vs. FPKM.UQ (ER balanced)
188 1 0 0 0
0 81 0 0 0
0 0 547 0 0
0 0 1 229 0
0 0 0 0 39
UQ (genefu) vs. RUV-III (genefu)
187 1 0 1 0
1 79 0 1 0
1 12 402 125 7
0 11 21 198 0
9 7 8 4 11
FPKM (ER balanced) vs. FPKM.UQ (ER balanced)
185 1 0 0 1
1 80 2 0 0
0 1 540 31 0
0 0 3 198 0
2 0 3 0 38
FPKM (ER balanced) vs. RUV-III (genefu)
184 2 0 1 0
2 79 0 2 0
1 14 407 143 7
0 9 13 179 0
11 6 11 4 11
FPKM.UQ (ER balanced) vs. RUV-III (genefu)
187 0 0 1 0
1 80 0 1 0
1 12 402 126 7
0 11 21 197 0
9 7 8 4 11
FPKM (genefu) vs. FPKM.UQ (genefu)
197 0 0 0 0
4 106 2 0 1
0 0 277 0 16
5 10 67 389 0
0 0 0 0 12
FPKM (genefu) vs. UQ (genefu)
184 0 0 0 13
4 80 14 8 7
0 0 287 0 6
1 1 245 222 2
0 0 1 0 11
FPKM (genefu) vs. FPKM (ER balanced)
182 0 0 0 15
4 82 14 7 6
0 0 284 0 9
1 1 273 194 2
0 0 1 0 11
FPKM (genefu) vs. FPKM.UQ (ER balanced)
184 0 0 0 13
3 81 14 8 7
0 0 287 0 6
1 1 246 221 2
0 0 1 0 11
FPKM (genefu) vs. RUV-III (genefu)
193 1 0 1 2
5 99 3 5 1
0 1 276 8 8
0 9 147 315 0
0 0 5 0 7
FPKM.UQ (genefu) vs. UQ (genefu)
188 0 3 0 15
1 80 16 14 5
0 0 343 0 3
0 1 171 216 1
0 0 14 0 15
FPKM.UQ (genefu) vs. FPKM (ER balanced)
185 1 3 0 17
2 81 16 13 4
0 0 342 0 4
0 1 199 188 1
0 0 12 0 17
FPKM.UQ (genefu) vs. FPKM.UQ (ER balanced)
187 1 3 0 15
1 80 16 14 5
0 0 343 0 3
0 1 172 215 1
0 0 14 0 15
FPKM.UQ (genefu) vs. RUV-III (genefu)
195 3 2 4 2
3 101 3 9 0
0 0 319 23 4
0 5 92 292 0
0 1 15 1 12
UQ (genefu) vs. FPKM (ER balanced)
186 1 0 0 2
0 80 1 0 0
0 2 540 2 3
0 0 31 199 0
1 0 0 0 38
Basal
Her2
LumA
LumB
Normal like
asal Her2 LumA LumB Normal like
B
Raw counts (genefu) p-value (log-rank) = 0.03
strata Basal (191) Her2 (115) LumA (315) LumB (397) Normal like (25)
FPKM (genefu) p-value (log-rank) = 0.01
strata Basal (188) Her2 (108) LumA (277) LumB (458) Normal like (12)
FPKM.UQ (genefu) p-value (log-rank) = 0.06
strata Basal (197) Her2 (111) LumA (328) LumB (378) Normal like (29)
UQ.normalized (ER balanced) p-value (log-rank) = 0.13
strata Basal (180) Her2 (77) LumA (523) LumB (225) Normal like (38)
Raw counts (ER balanced) p-value (log-rank) = 0.38
strata Basal (179) Her2 (88) LumA (512) LumB (227) Normal like (37)
FPKM (ER balanced) p-value (log-rank) = 0.05
strata Basal (178) Her2 (79) LumA (547) LumB (197) Normal like (42)
FPKM.UQ (ER balanced) p-value (log-rank) = 0.13
strata Basal (179) Her2 (78) LumA (524) LumB (224) Normal like (38)
RUV-III (genefu) p-value (log-rank) = 0.05
strata Basal (189) Her2 (104) LumA (413) LumB (320) Normal like (17)
C
Raw counts (genefu) p-value (log-rank) = 0.04
strata Basal (185) Her2 (112) LumA (306) LumB (396)
FPKM (genefu) p-value (log-rank) = 0.01
strata Basal (185) Her2 (107) LumA (272) LumB (457)
FPKM.UQ (genefu) p-value (log-rank) = 0.05
strata Basal (191) Her2 (108) LumA (316) LumB (376)
UQ.normalized (ER balanced) p-value (log-rank) = 0.06
strata Basal (170) Her2 (74) LumA (507) LumB (222)
Raw counts (ER balanced) p-value (log-rank) = 0.23
strata Basal (170) Her2 (85) LumA (495) LumB (224)
FPKM (ER balanced) p-value (log-rank) = 0.02
strata Basal (168) Her2 (76) LumA (527) LumB (194)
FPKM.UQ (ER balanced) p-value (log-rank) = 0.06
strata Basal (169) Her2 (75) LumA (508) LumB (221)
RUV-III (genefu) p-value (log-rank) = 0.02
strata Basal (185) Her2 (101) LumA (407) LumB (319)
Survival
100% 75% 50% 25% 0%
Time (days)
0 1000 2000 3000 4000

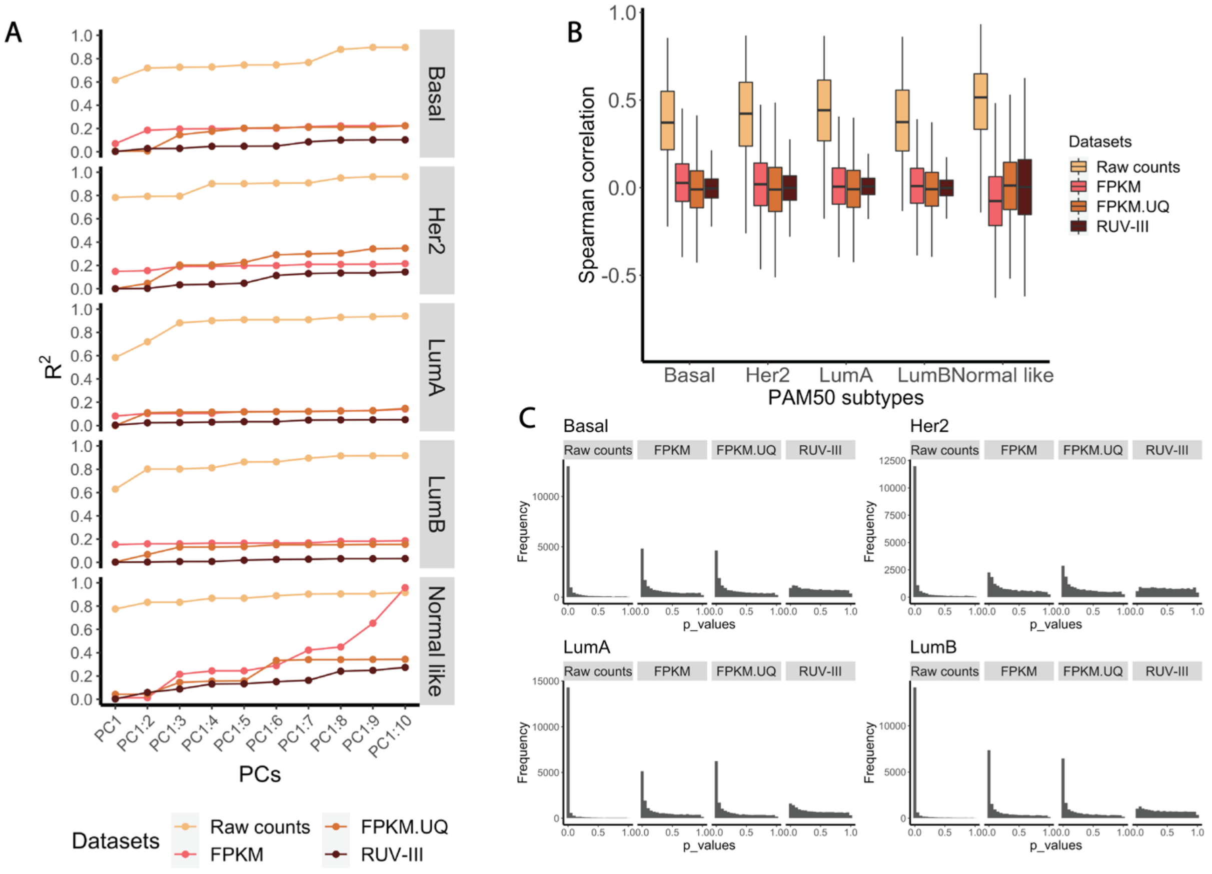
Supplementary Figure 27. PAM50 subtype identification using two different methods in the TCGA BRCA RNA-sequencing data. A) Tables show the concordance between the PAM50 subtypes obtained using two methods on different datasets. UQ is upper-quartile normalization of the TCGA BRCA raw counts data. **B)** Kaplan-Meier survival curves show the association between the PAM50 subtypes and overall survival in the TCGA BRCA RNA-sequencing data. **C)** same as B without the normal-like subtypes.



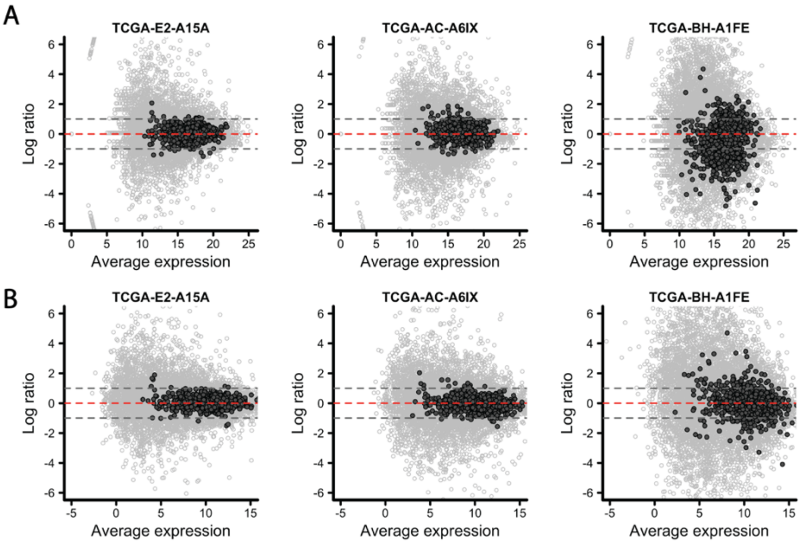
Supplementary Figure 28. PCA plots coloured the PAM50 subtypes of the TCGA BRCA RNA-sequencing datasets. Each panel shows the scatter plots for two of the first three PC for the raw data and several normalized versions of the TCGA BRCA RNA-sequencing data. Samples are coloured based on the PAM50 calls obtained using the different approaches. UQ is upper-quartile normalization of the TCGA BRCA raw counts data.



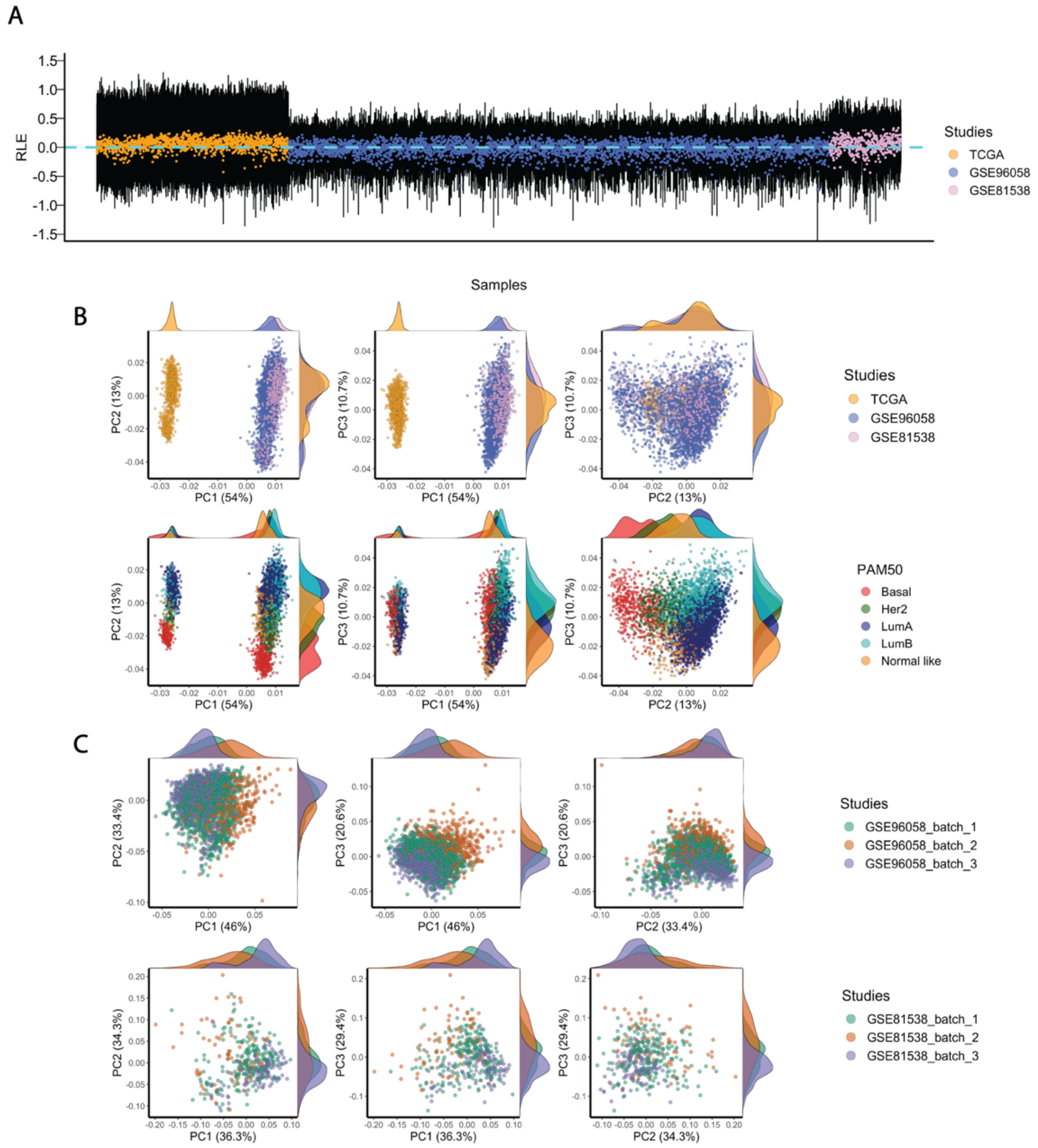
Supplementary figure 29. RUV-III removes the effects of different flow cell chemistries in the TCGA BRCA RNA-sequencing data. A) The first three PC coloured by flow cell chemistry in the differently normalized RNA-seq data. **B)** Vector correlation analysis between the first ten PCs cumulatively and flow cell chemistries within each of the PAM subtypes of the differently normalized TCGA BRCA RNA-seq data. C) Boxplots of $\log_2$ F statistics obtained from ANOVA between gene expression with key time intervals being the factor (n=16,537 genes). The heavy middle line represents the median, the box shows the inter-quartile range (IQR), the upper and lower whiskers extend from the hinges no further than 1.5 × IQR, and any outliers beyond the whiskers are shown as points.



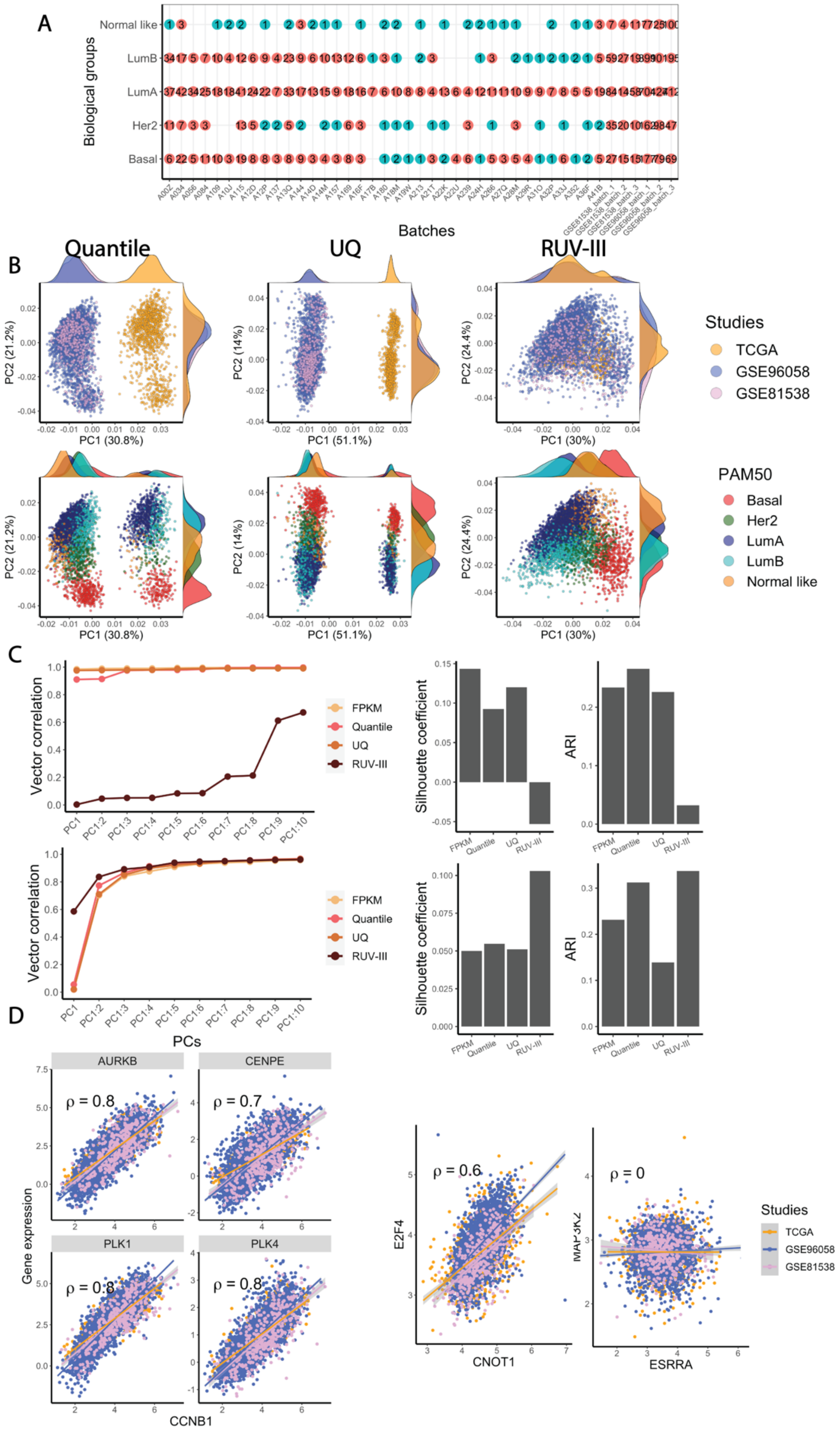
Supplementary figure 30. RUV-III removes library size from the TCGA BRCA RNA-seq data. A) Plots show R^2 calculated using linear regression between the first 10 principal components cumulatively and library size in the TCGA and RUV-III normalized datasets. **B)** The boxplots of the Spearman correlation coefficients between the individual gene expression levels and library size within the PAM50 subtypes in differently normalized datasets (n=16,537 genes). The heavy middle line represents the median, the box shows the inter-quartile range (IQR), the upper and lower whiskers extend from the hinges no further than 1.5 × IQR, and any outliers beyond the whiskers are shown as points. **C)** The histograms of uncorrected p-values obtained from differential expression analyses using Wilcoxon signed-rank test between samples with high and low library sizes within each PAM50 subtype in differently normalized datasets.



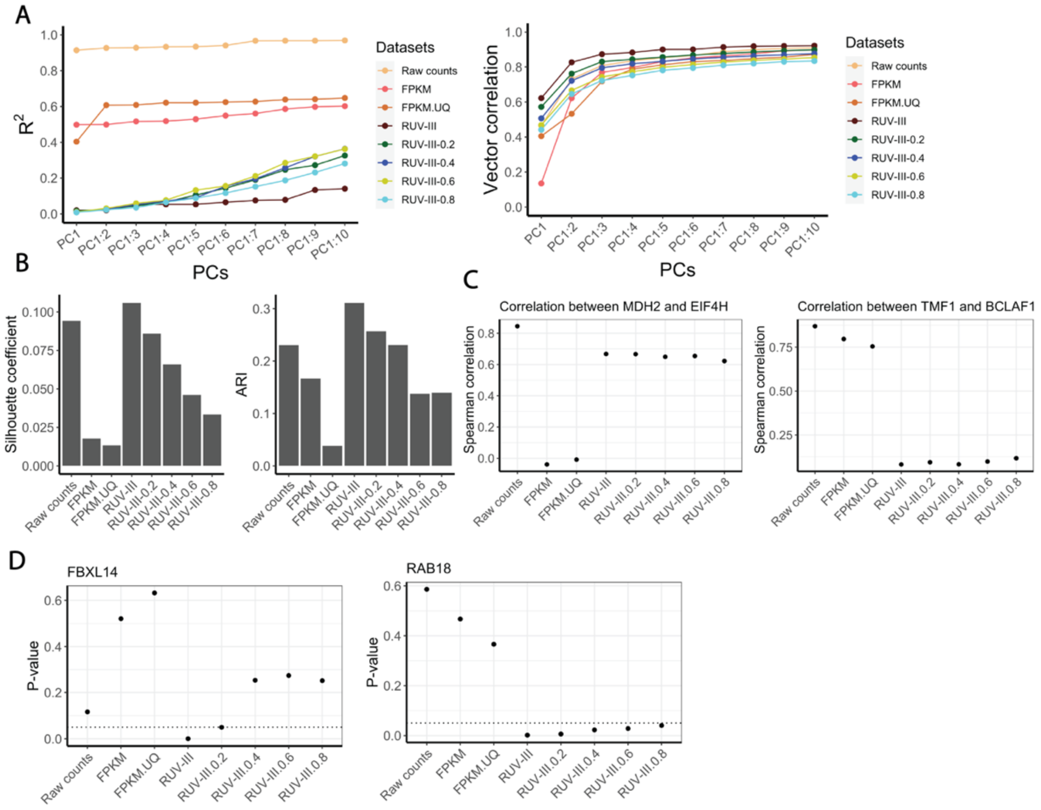
Supplementary Figure 31. Impact of flow cell chemistries on gene expression differences between paired primary and metastatic samples in the TCGA BRCA RNA-sequencing data. MA plot of three paired primary and metastatic samples profiled within (the first two plots of top row) and across flow cell chemistries (third plot of top row) in the TCGA FPKM.UQ data. Second row: same a first row for the RUV-III normalized data. The solid black points in the MA plots represent the genes highly affected by flow cell chemistries.



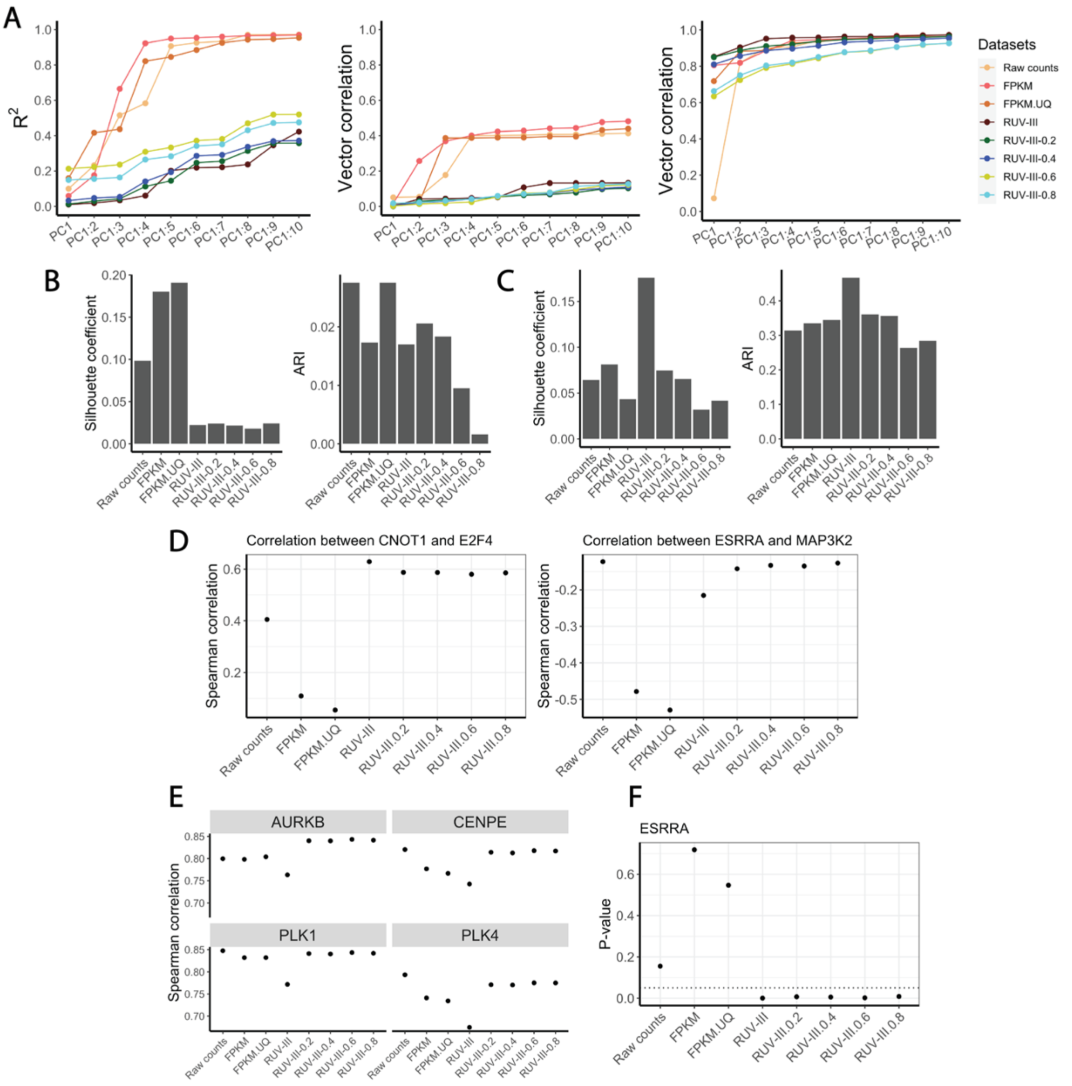
Supplementary Figure 32. Unwanted variation within and between three large breast cancer RNA-seq studies. A) The RLE plots of combined FPKM counts of the studies**. The medians of boxplots are coloured based on different studies.** The box shows the inter-quartile range (IQR) and the upper and lower whiskers extend from the hinges no further than 1.5 ´ IQR. **B)** Scatter plots of the first three PCs of the combined FPKM counts coloured by the studies (first row) and the PAM50 subtypes (second row). **C)** PCA plots using negative control genes (RNA-seq housekeeping genes) within each study coloured by the batches identified using the RLE medians.



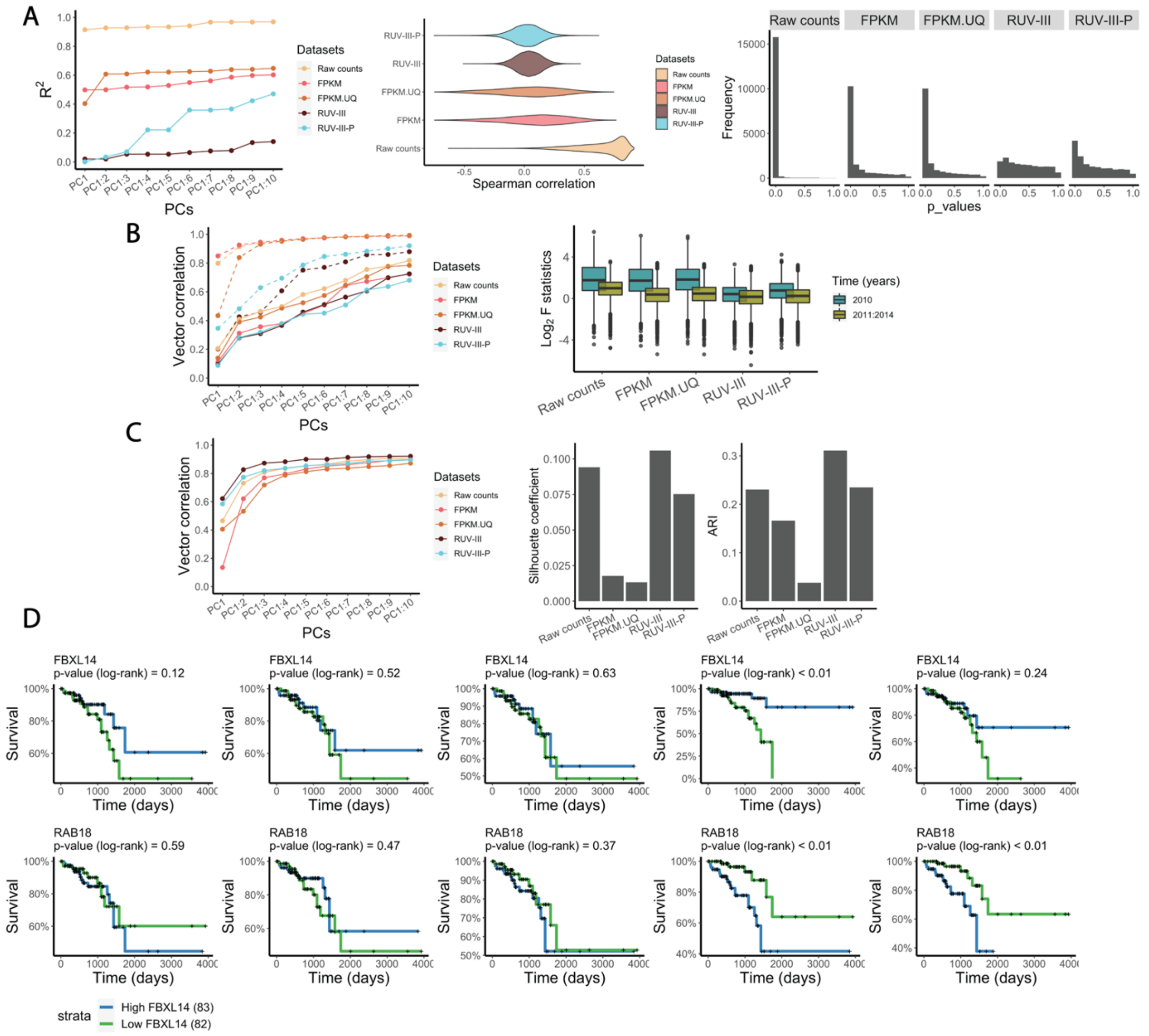
Supplementary Figure 33. Normalization of three large breast cancer RNA-Seq studies. A) Plot showing the sample sizes of the PAM50 subtypes across batches across all the studies. The batches for the non-TCGA studies were identified using the RLE medians within each study. **B)** The scatter plots of the first two PCs of the data normalized by different methods. These plots are coloured by the studies (top row) and the PAM50 subtypes (bottom row). **C)** Top row: A plot showing the vector correlation coefficient between studies and up to the first 10 principal components. Silhouette coefficients and ARI index for mixing samples from different studies. Bottom row: same as the top row for the PAM50 subtypes. **D)** First panel: scatter plots show correlation between the CCNB1 (X-axes) and other genes, including AURKB, CENPE, PLK1 and PLK4 (Y-axes) in the RUV-III normalized data. Second panel: scatter plots show correlation between the two pairs of genes, CNOT1_E2F4 and ESRRA_MAP3K2 in the RUV-III normalized data. Confidence intervals are shown as grey bands.



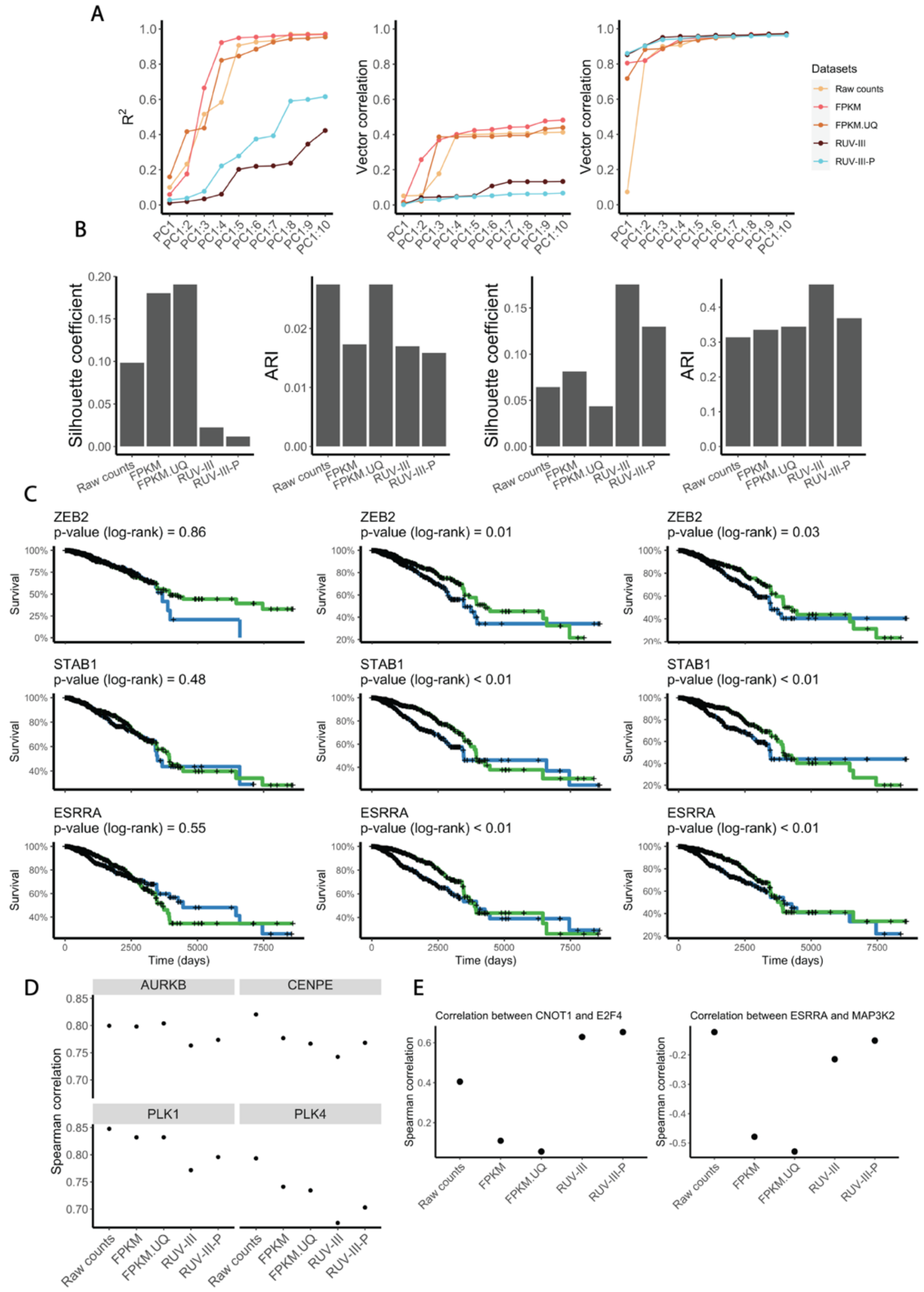
Supplementary Figure 34. Performance of RUV-III with poorly chosen PRPS in the TCGA READ RNA-seq data. A) First plot: R^2 obtained from linear regression between the first 10 principal components (cumulatively) and library size in the differently normalized datasets. We randomly shuffled various proportions 20%, 40%, 60% and 80% of the CMS subtypes that were initially used to create PRPS for RUV-III. Second plot: vector correlation between the first 10 PC (cumulatively) and the CMS subtypes in differently normalized datasets. **B)** Bar charts of silhouette coefficients and ARI showing the performance of different normalization methods in separating the CMS subtypes. **C)** Plots show Spearman correlation between two pair of genes in differently normalized data. **D)** Plots show p-values obtained from Kaplan-Meier survival analyses between the genes expression and survival outcomes. The dashed line shows 0.05.



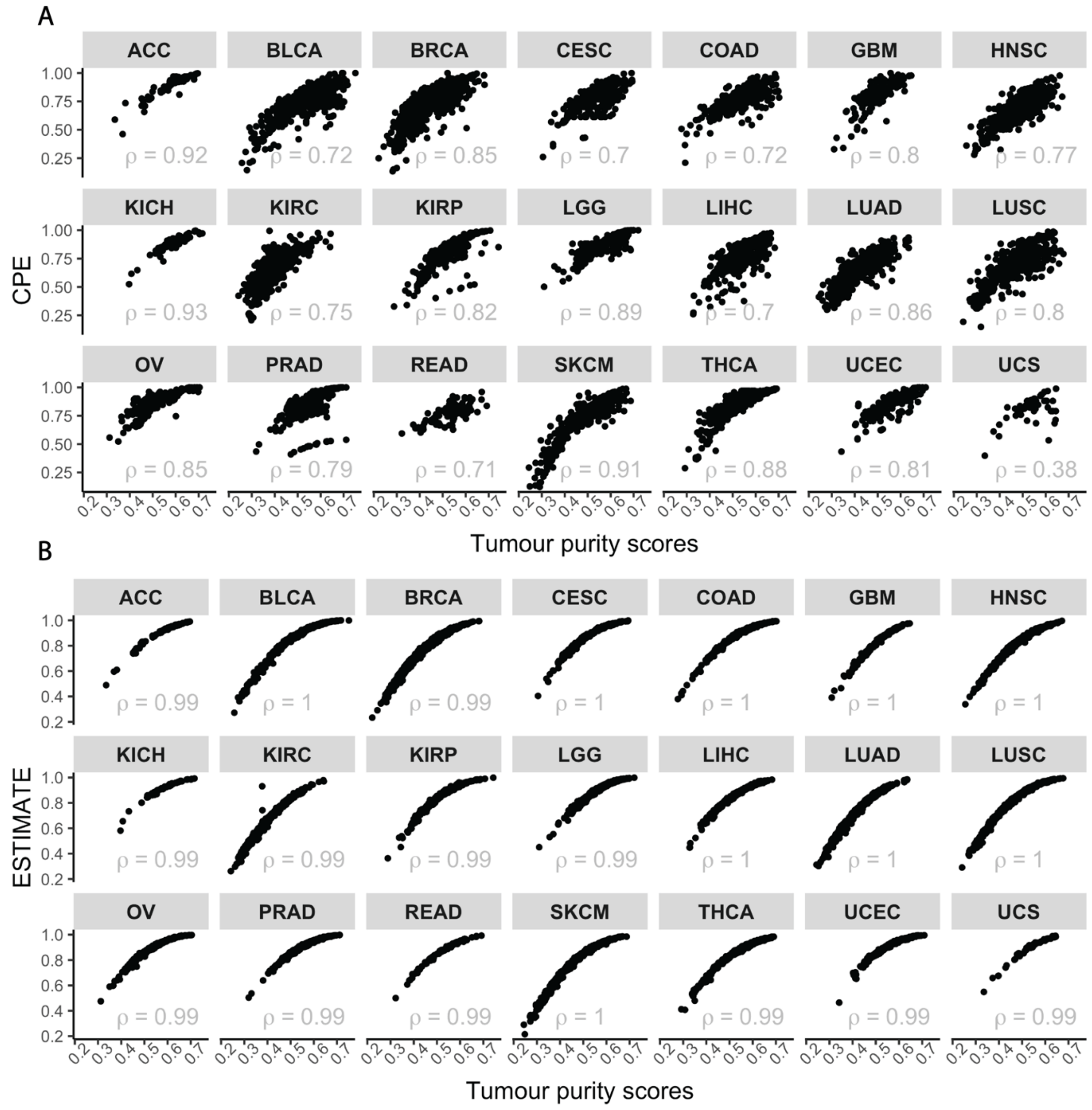
Supplementary Figure 35. Performance of RUV-III with poorly chosen PRPS in the TCGA BRCA RNA-seq data. A) First plot: R^2 obtained from linear regression between the first 10 principal components (cumulatively) and tumour purity in the differently normalized datasets. We randomly shuffled various proportions 20%, 40%, 60% and 80% of the PAM50 subtypes that were initially used to create PRPS for RUV-III. Second plot: Vector correlation between the first 10 PC (cumulatively) and the flow cell chemistry effects in differently normalized datasets. Third plot, similar to the second plot for the PAM50 subtype. **B)** Bar charts of silhouette coefficients and ARI indices showing the performance of different normalization methods in mixing samples from different flow cell chemistries. **C)** Same as B for the PAM50 subtypes. **D)** Plots show the Spearman correlation between two pair of genes in differently normalized data. **E)** Plots show Spearman correlation between the expression of CCNB1 and other genes, including AURKB, CENPE, PLK1 and PLK4. **F)** Plots show p-values obtained from Kaplan-Meier survival analyses between ESRRA gene expression and survival outcomes. The dashed line shows 0.05.



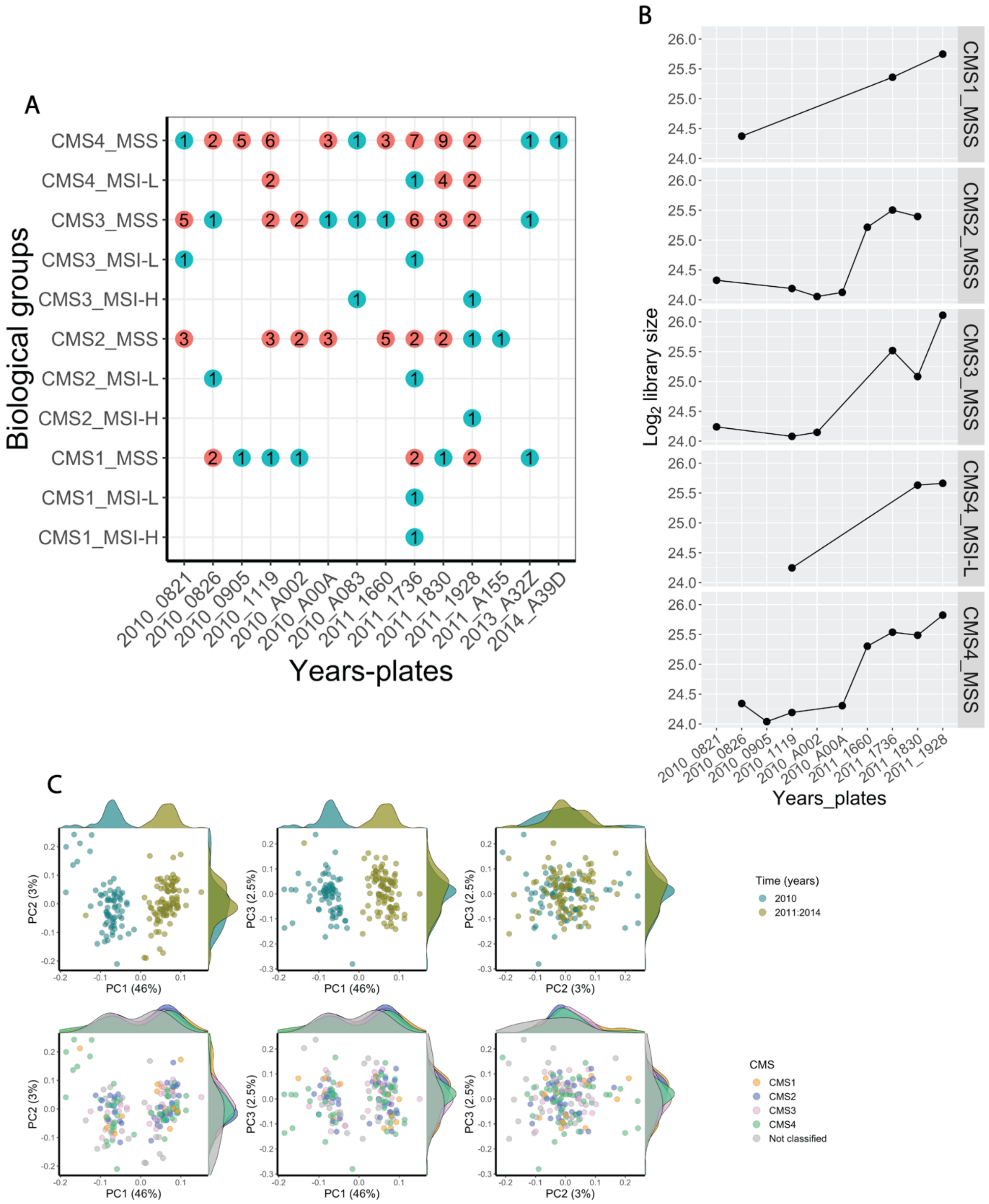
Supplementary Figure 36. Performance of RUV-III normalization with using only the CMS4 subtype for creating PRPS in the TCGA READ RNA-seq data. A) First plot: R^2 obtained from linear regression between the first 10 principal components (cumulatively) and library size in the differently normalized datasets. Second plot: violin plots of Spearman correlation between individual gene expression and library size in the normalized datasets. Third plot, histograms of unadjusted *p*-values obtained from differential gene expression analysis using Wilcoxon signed-rank test between samples with low and high library sizes in the TCGA READ RNA-seq data. **B)** First plot: A plot showing the vector correlation coefficient between plates and the first 10 principal components within each time points. Second plot: boxplots of log2 F statistics obtained from ANOVA within each time points for gene expression with plate as a factor (n = 16,327). In the boxplots, the heavy middle line represents the median, the box shows the inter-quartile range (IQR), the upper and lower whiskers extend from the hinges no further than 1.5 ´ IQR, and any outliers beyond the whiskers are shown as points. **C)** First plot: A plot showing the vector correlation coefficient between CMS subtypes and up to the first 10 principal components. Second plot: Bar charts of silhouette coefficients and ARI index for measuring the separation of CMS subtypes. **D)** Kaplan–Meier curves for samples with low (below median) and high (above median) expression of the FBXL14 and RAB18.



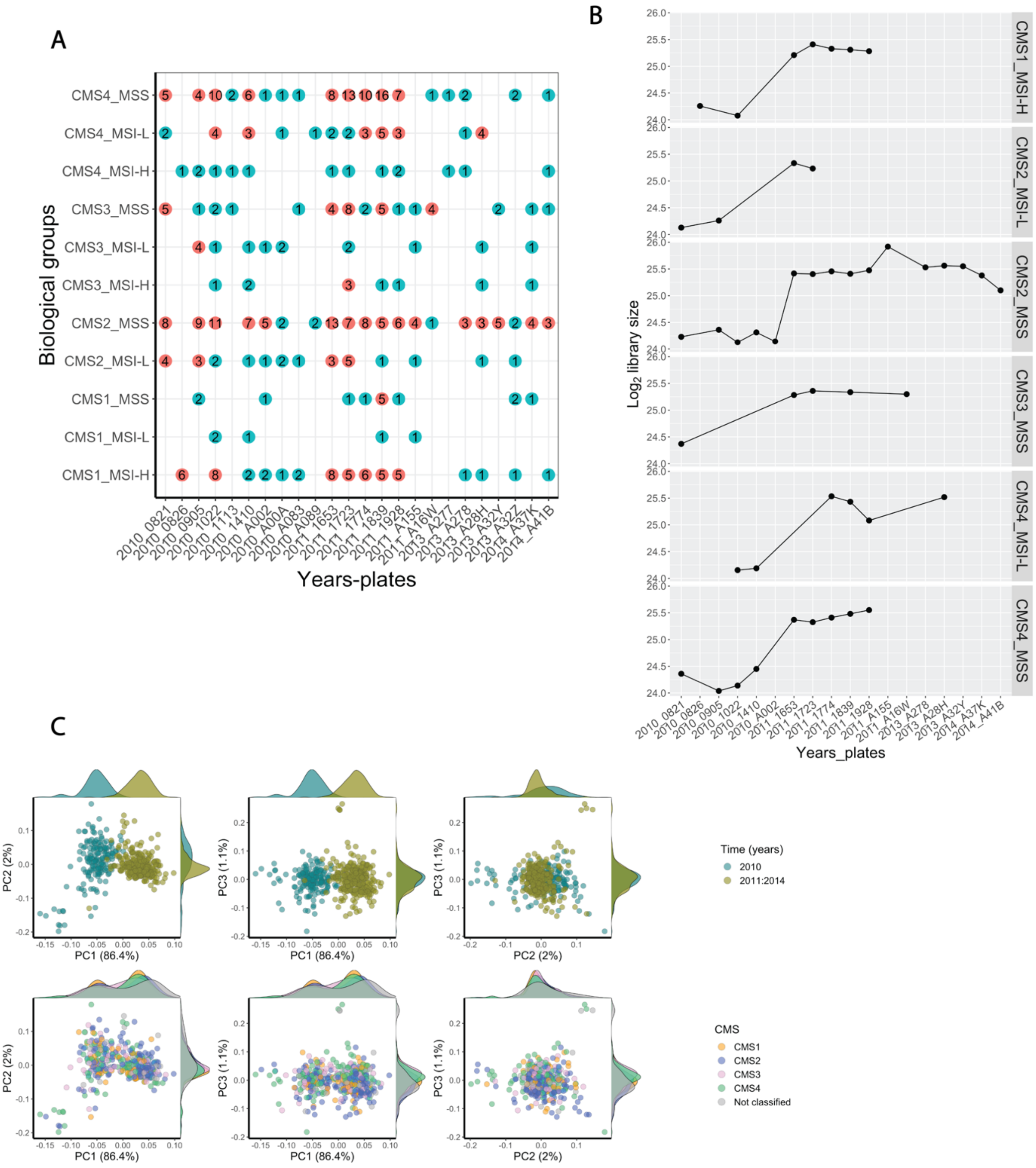
Supplementary Figure 37. Performance of RUV-III normalization with using the basal and lumA subtypes for creating PRPS in the TCGA BRCA RNA-seq data. A) First plot: R^2 obtained from linear regression between the first 10 principal components (cumulatively) and purity in the differently normalized datasets. Second plot: a plot showing the vector correlation coefficient between flow cell chemistries and the first 10 principal components. Third plot: same as the second plot for the PAM50 subtypes. **B)** The first two bar charts: silhouette coefficients and ARI index for measuring the mixing of the flow cell chemistries. The second two bar charts: silhouette coefficients and ARI index for measuring the separation of the PAM50 subtypes. **C)** Kaplan–Meier curves for samples with low (below median) and high (above median) expression of the ZEB2, STAB1, and ESRRA. **D)** Plots show Spearman correlation between the CCNB1 and other genes, including AURKB, CENPE, PLK1 and PLK4. **E)** Plots show the Spearman correlation between two pairs of genes in differently normalized data.



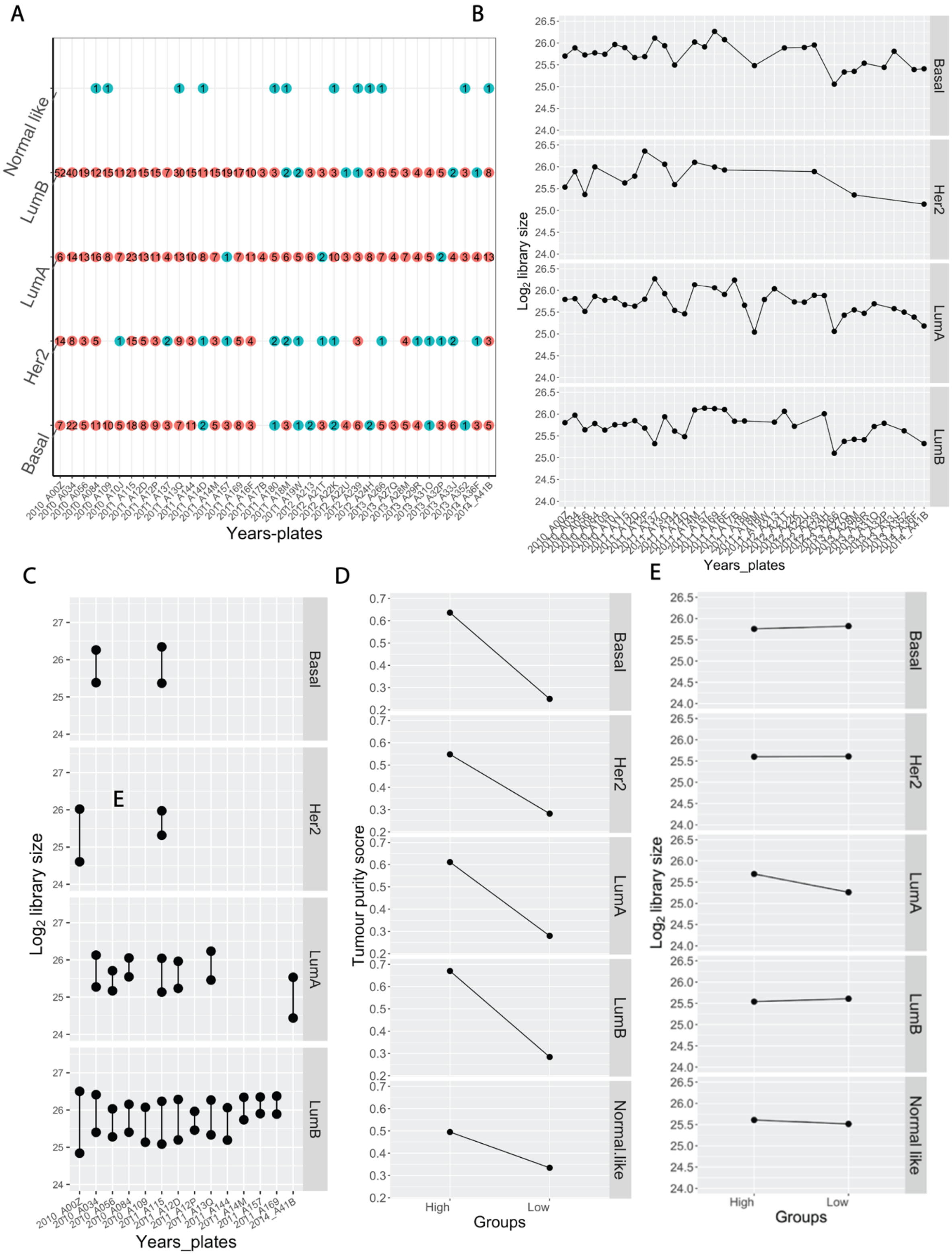
Supplementary Figure 38. Tumour purity estimates in TCGA RNA-sequencing datasets. A) The scatter plots show the relationship between CPE and tumour purity scores. **B)** The scatter plots show relationship between the tumour purity calculated by the ESTIMATE method and consensus measurement of purity estimations (CPE) approach in the TCGA RNA-seq datasets.



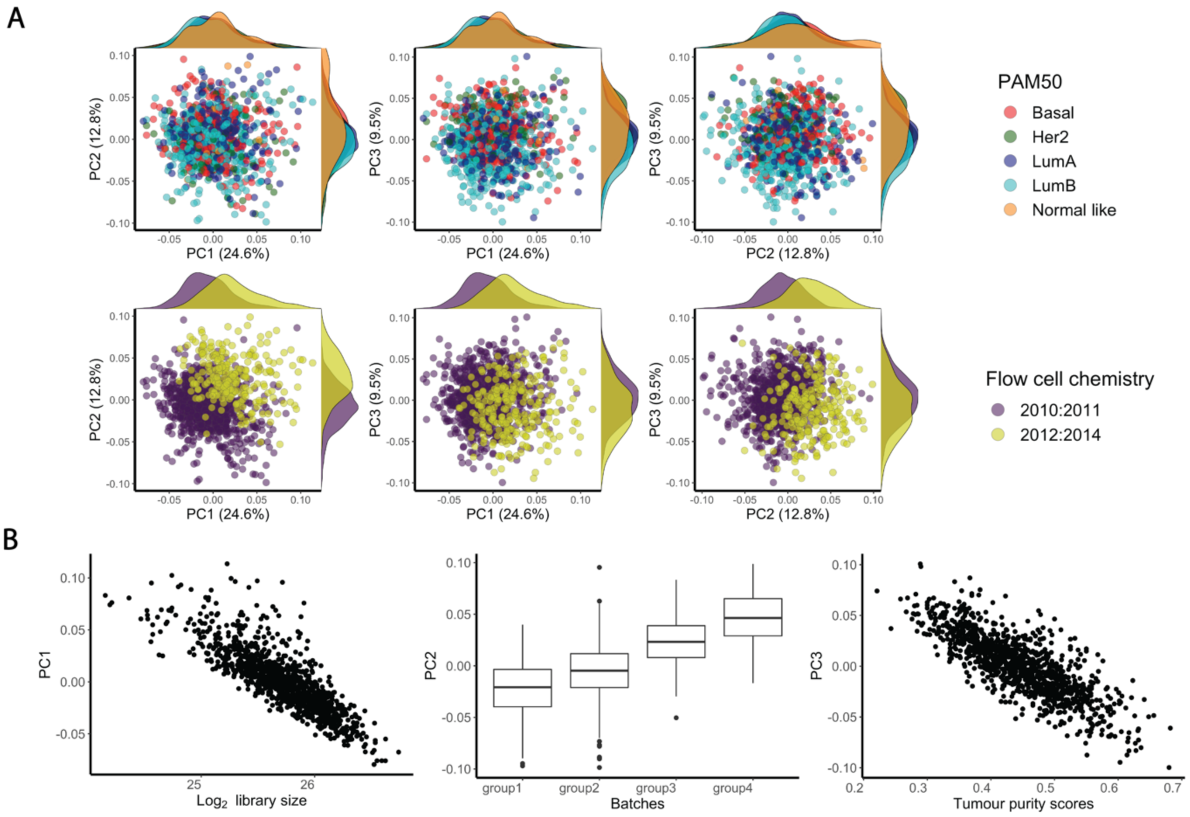
Supplementary Figure 39. PRPS for RUV-III normalization of the TCGA READ RNA-sequencing data. A) Plot showing the sample sizes of the major biological groups across plates in the TCGA READ RNA-seq data. **B)** Library sizes of pseudo-samples**. C)** First row: scatter plots show the first three principal components coloured by key time intervals of only 919 negative control genes in the TCGA raw counts RNA-seq data. Second row, same as first row, coloured by CMS.



Supplementary Figure 40. PRPS for RUV-III normalization of the TCGA COAD RNA-sequencing data. A) Plot showing the sample sizes of the major biological groups across plates in the TCGA COAD RNA-seq data. **B)** Library sizes of pseudo-samples**. C)** First row: scatter plots show the first three principal components coloured by key time intervals of only 262 negative control genes in the TCGA raw counts RNA-seq data. Second row, same as first row, coloured by CMS.



Supplementary Figure 41. PRPS for RUV-III normalization of the TCGA BRCA RNA-sequencing data. A) Plot showing the sample sizes of the major biological groups across plates in the TCGA BRCA RNA-seq data. **B)** Library sizes of pseudo-samples created for removing plate effects. **C)** Library sizes of pseudo-samples created for removing plate library sizes. **D)** Tumour purity score of pseudo-samples created for removing tumour purity variation. **E)** Library sizes of pseudo-samples created for removing tumour purity variation.



Supplementary Figure 42. Negative control genes for RUV-III normalization of the TCGA BRCA RNA-sequencing data. A) first row: the first three PC coloured by PAM50 subtypes of the TCGA BRCA raw counts using a set of negative control genes. Second row: same a first row, coloured by flow cell chemistry. **B)** Scatter plots show relationship between the first three PC of the negative control genes with library size, batches and tumour purity scores (n_{group1} = 117, n_{group2} = 741, n_{group3} = 169, n_{group4} = 59). In the boxplots, the heavy middle line represents the median, the box shows the inter-quartile range (IQR), the upper and lower whiskers extend from the hinges no further than 1.5 × IQR, and any outliers beyond the whiskers are shown as points.

References

1. Guinney, J., et al., *The consensus molecular subtypes of colorectal cancer.* Nat Med, 2015. **21**(11): p. 1350-6.
2. Eide, P.W., et al., *CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models.* Sci Rep, 2017. **7**(1): p. 16618.
3. Parker, J.S., et al., *Supervised risk predictor of breast cancer based on intrinsic subtypes.* J Clin Oncol, 2009. **27**(8): p. 1160-7.
4. Gendoo, D.M., et al., *Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer.* Bioinformatics, 2016. **32**(7): p. 1097-9.
5. Picornell, A.C., et al., *Breast cancer PAM50 signature: correlation and concordance between RNA-Seq and digital multiplexed gene expression technologies in a triple negative breast cancer series.* BMC Genomics, 2019. **20**(1): p. 452.