Exploring RNA-seq data normalization methods using principal component analysis and KEGG pathway enrichment

Henk J. van Lingen[1]*, Maria Suarez-Diez[1], Edoardo Saccenti[1]
[1]Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen, the Netherlands.
*henk.vanlingen@wur.nl

Introduction

- AIM: to investigate the implications of normalization methods for RNA-seq data using principal component analysis (PCA).

Datasets and normalization methods

Human tumor data

- 3 patients from whom normal and tumour tissue was taken
- Gene expression read counts for 10144 genes

HCT116 human colon cancer cell line data

- 4 HCT116 cell lines of which 2 control and 2 MIB1 knockdown
- Gene expression read counts for 9853 genes

Normalization methods: total-count, upper-quartile, trimmed mean of *M*-values (*i.e.* fold changes; TMM), edgeR TMM, edgeR TMM with singleton paring (TMMwsp), relative log expression, centred log ratio, full-quantile, probabilistic quotient, within-lane GC-content based, within-and-between-lane GC-content based, transcripts per million (TPM), reads per kilobase of transcript per million reads mapped (RPKM) and conditional quantile

Results

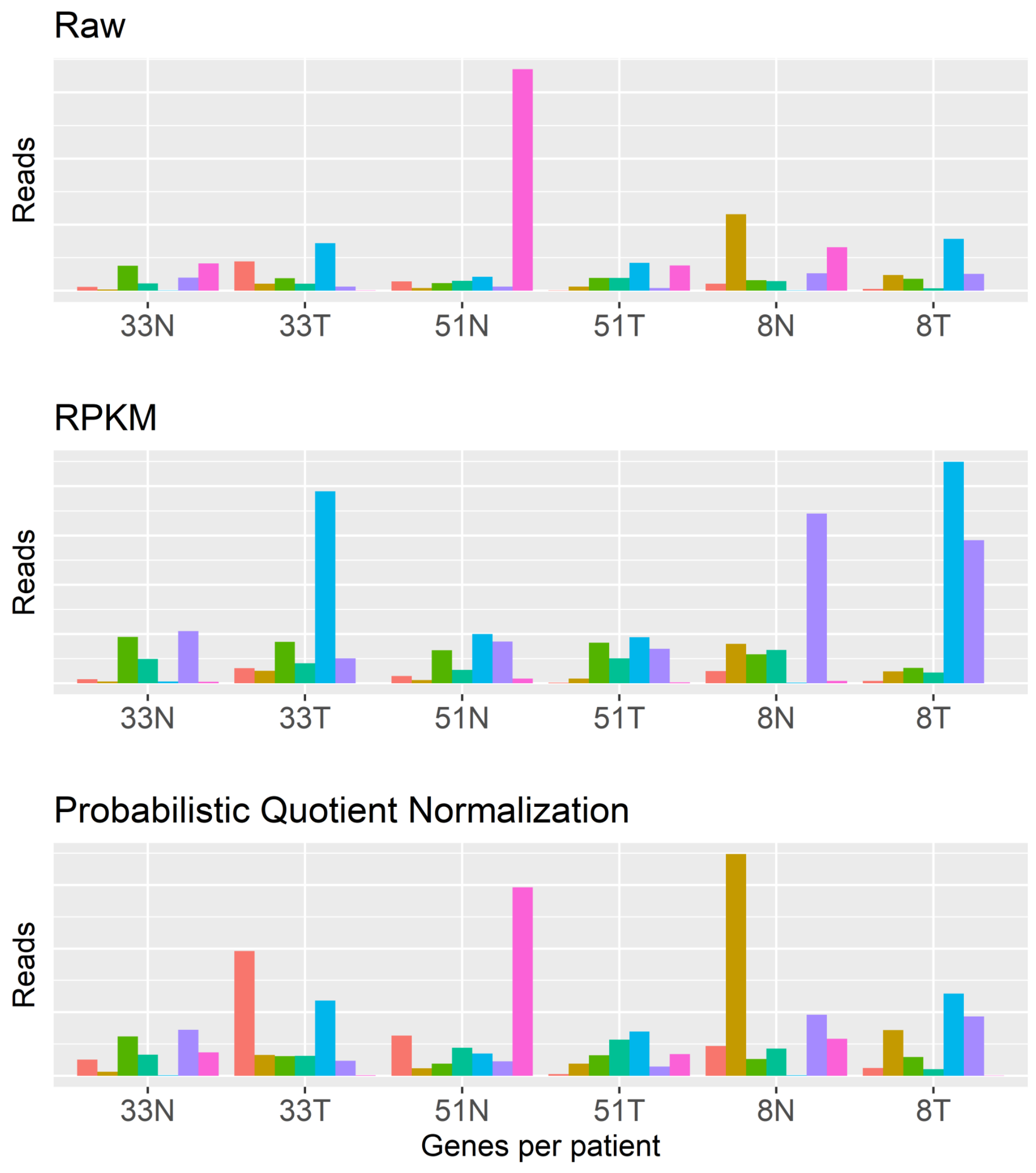

Figure 1. Raw and RPKM and probabilistic quotient normalized read counts for seven randomly selected genes from normal (N) and tumor (T) tissue from the 3 patients (*viz.* 8, 33, 51) of the human tumor data.

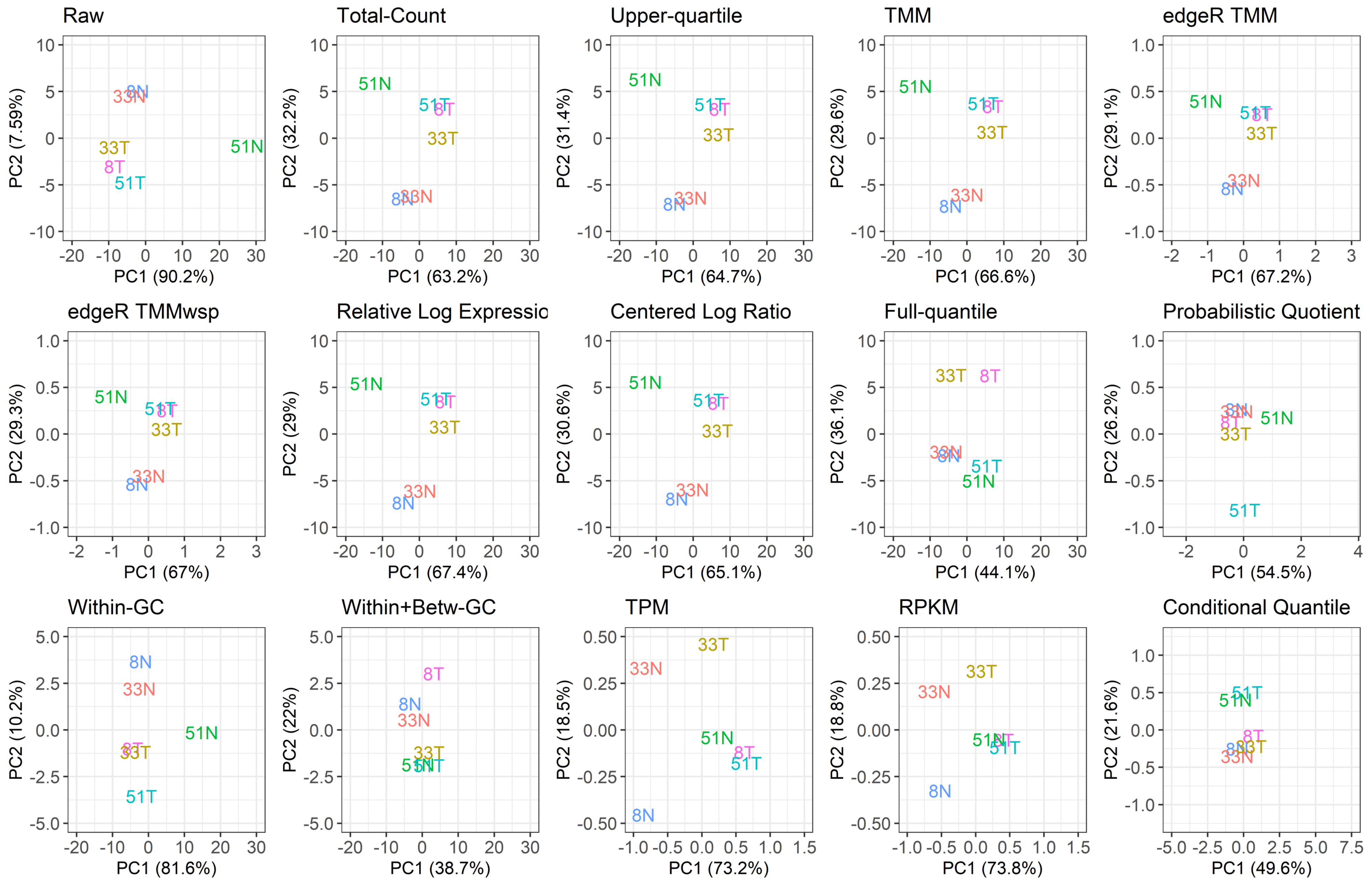

Figure 2. PCA plots for the raw and normalized human tumor data with variance explained for the first two principal components

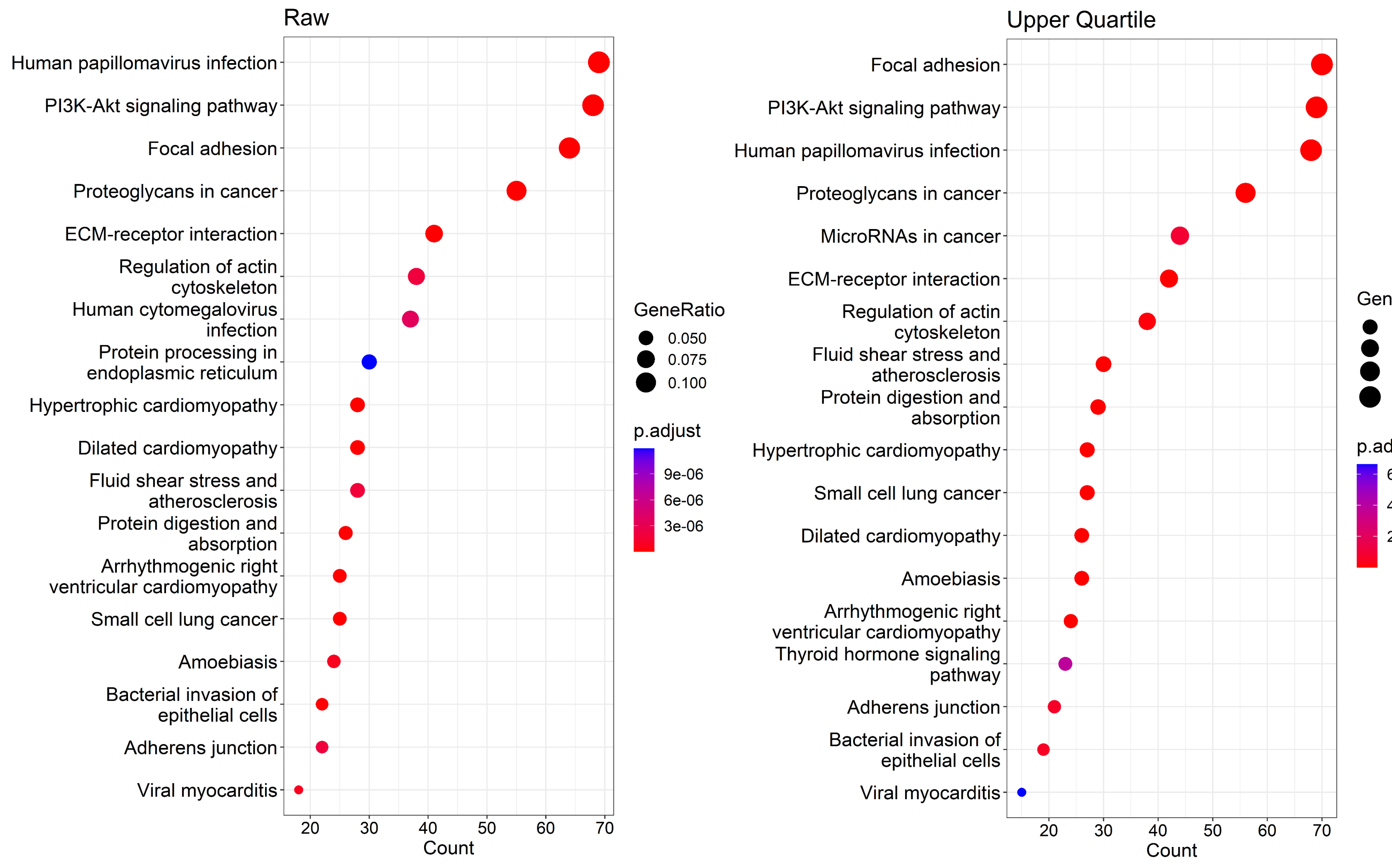

Figure 3. KEGG pathway enrichment analysis dotplots using the 1000 most influential genes indicted by the sum of the loadings multiplied by the variance explained for the first 3 principal components for the raw and upper quartile normalized human tumor data.

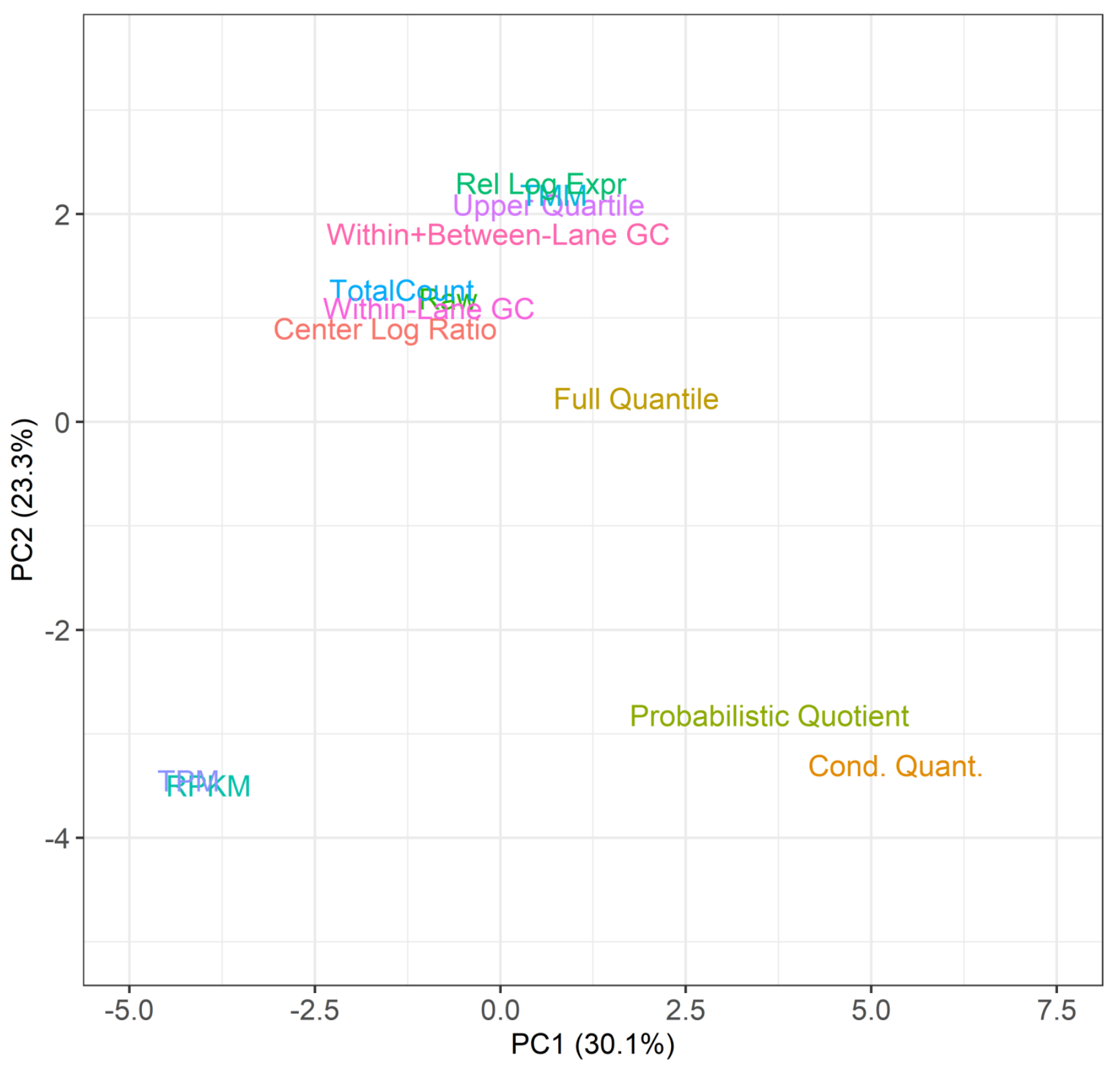

Figure 4. PCA plot on the matrix containing zeros and ones that indicated if a KEGG unit could be obtained from the 1000 most influential genes per normalization method applied to the human tumor data.

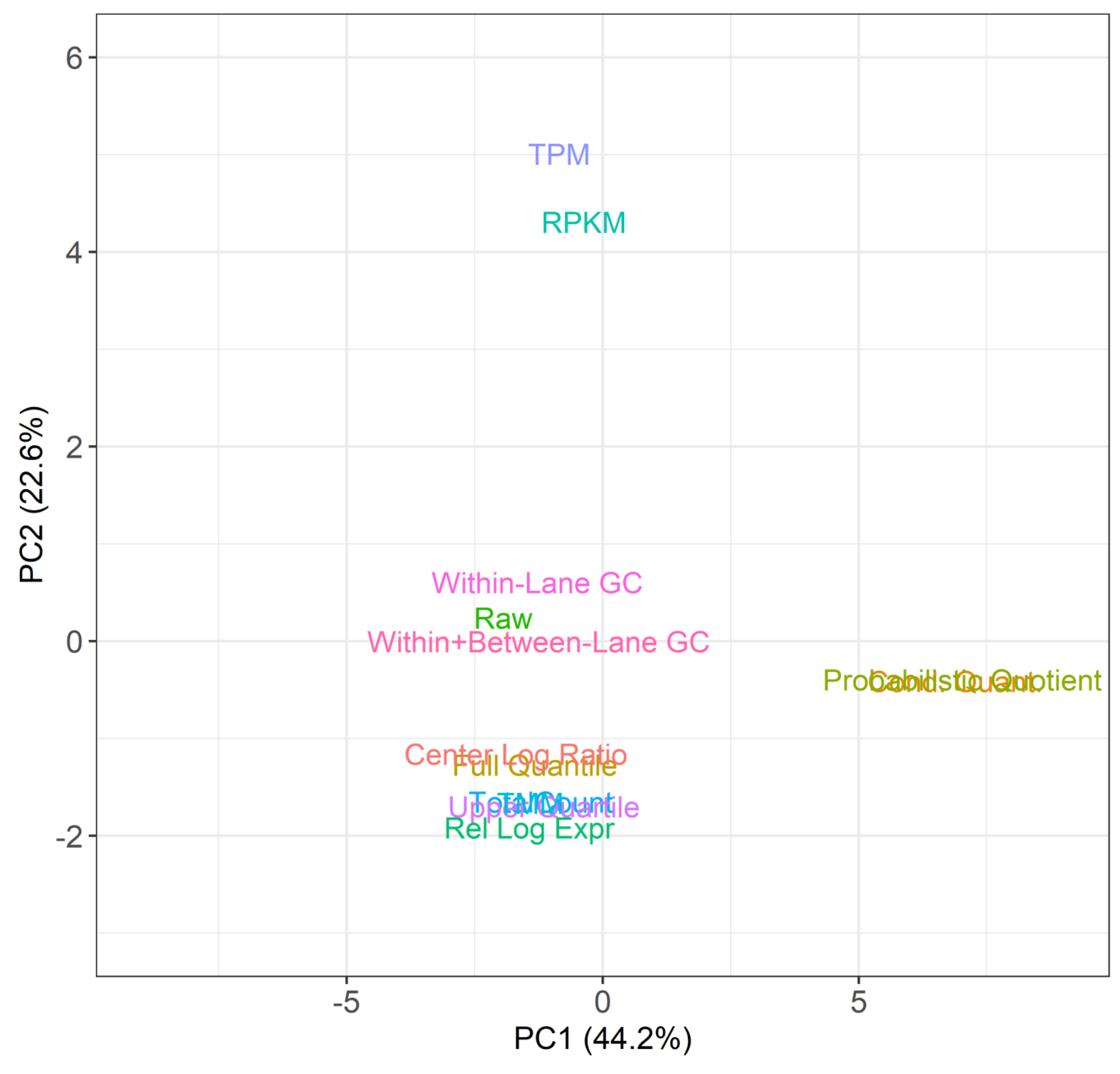

Figure 5. PCA plot on the matrix containing zeros and ones that indicated if a KEGG unit could be obtained from the 1000 most influential genes per normalization method applied to the human colon cancer cell line data

Conclusion

- Selecting the top 1000 most influential genes is relatively arbitrary, but still indicates differences between the various normalization methods.
- PCA indicates normalization method selection is not trivial as these methods have implications for the biology based on the KEGG pathways that were enriched.

References

- Park, J. and Seo, S. "Effect of depletion of MIB1 in HCT116 cells". *Gene Expression Omnibus accession GSE218399*, 2022.
- Tuch, B.B. et al. (2010). "Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations." *PLoS ONE* 5, e9317, 2010.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 953152