Normalization and correction for batch effects via RUV for RNA-seq data: practical implications for Breast Cancer Research

A. Debit[1], S. Wenric[2], C. Josse[2], V. Bours[2], K. Van Steen[1]
[1] *GIGA-R Medical Genomics, BIO3, University of Liege (Belgium)*
[2] *GIGA-R, Unit of Human Genetics, University of Liege (Belgium)*

BACKGROUND

- RNA-seq raw data are usually colonized by within-sample and between-sample unwanted variation
- Within sample biases: gene length, GC% content.
- Between sample biases: library size, library composition.
- Between sample biases are often related to the known and potential hidden effects
- A good normalization method and correction for batch effects (even known or hidden) must be carefully selected and applied to get good results, increase the accuracy of statistical inference, and improve data interpretation (reduction of type I error)

MOTIVATION

We cannot entirely know the exact causes of unwanted variation, and even we know it, we cannot precisely give a correct measurement of it. Researchers have shown that these unwanted sources of heterogeneity could dramatically reduce the accuracy of statistical inference in genomic data analysis [2].

- Normalization does not remove batch effects, which affect specific subsets of genes and may affect different genes in different ways [1]
- Most of the normalization methods proposed in the literature don't correct for unknown batch effects: RPKM, TMM, UQ, DESeq.
- Therefore, we propose a two-stage normalization method including EDASeq and RUVg for RNA-seq normalization and correction for known and potential hidden effects.

MATERIALS AND METHODS

- **Data:** Data coming from paired-end Illumina sequencing of 22 ER+ human breast cancer and 22 normal matched tissues have been processed in a classical pipeline including the quality control of the data, the mapping to the human reference genome (Homo_sapiens.GRCh38) from Ensembl, and the summarization of the reads per biotype annotation per sample using HTSeq.

- **Methods:**

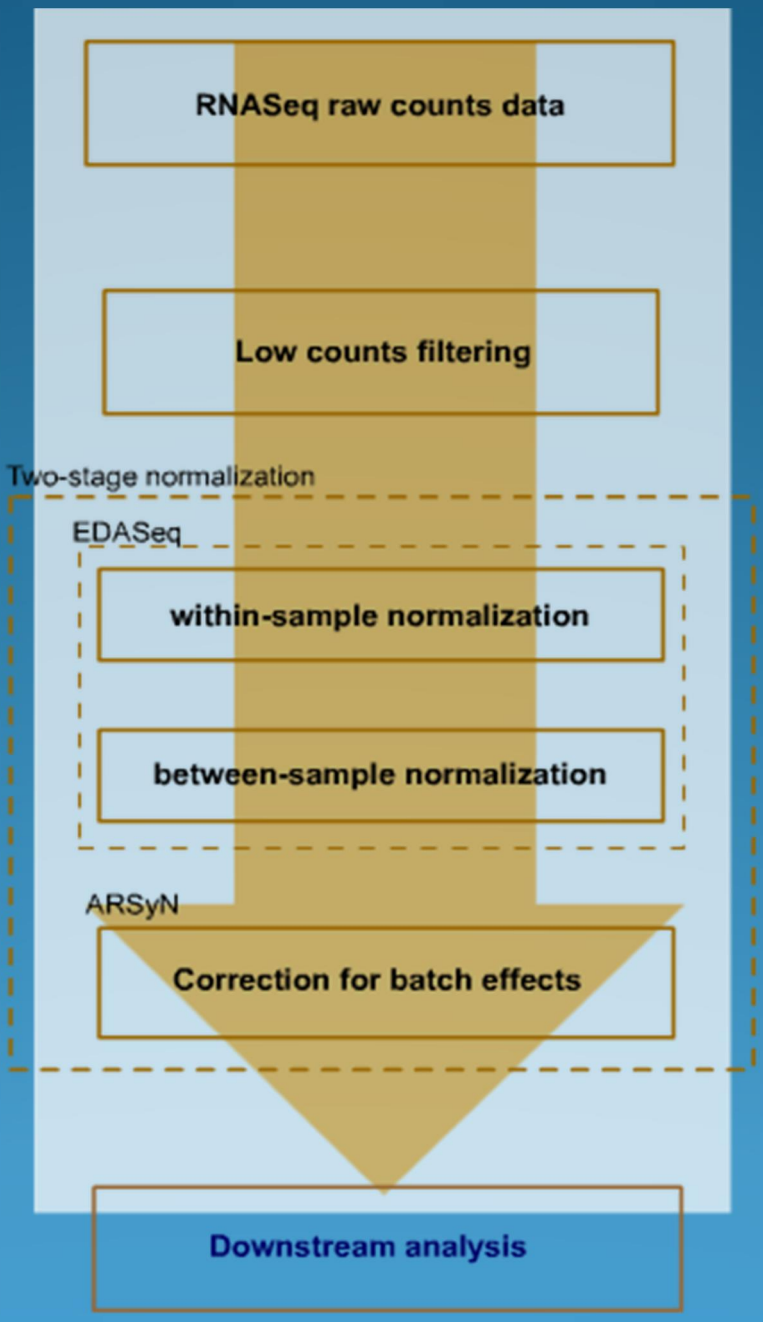

Fig 1. Normalization pipeline

RESULTS

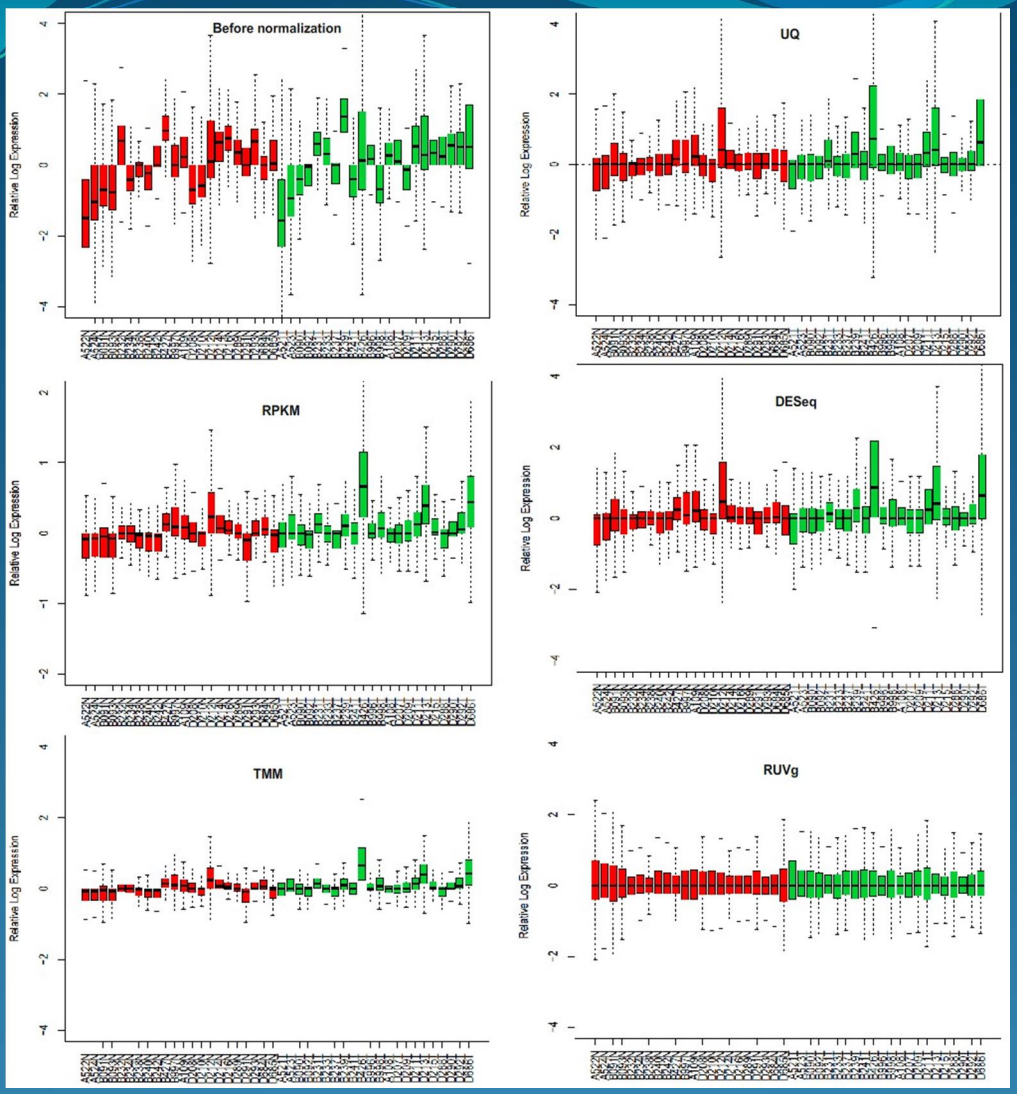

Fig 2. RLE Boxplot of the 5 normalization methods: UQ, RPKM, DESeq, TMM, and RUVg (following the proposed two-stages normalization)

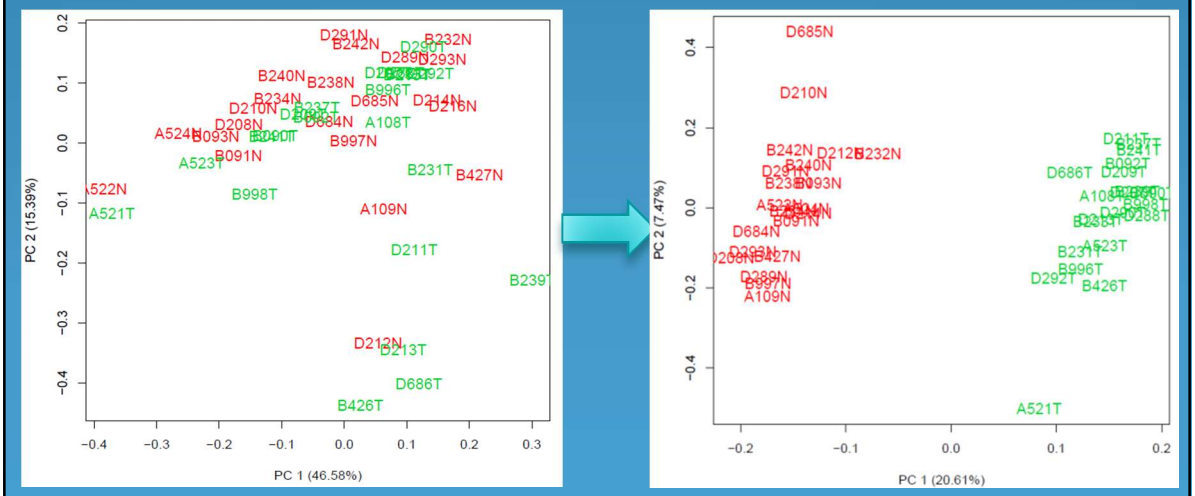

Fig 3. PCA Plot of the raw data before normalization and after applying the proposed normalization pipeline (red: Normal samples, green: Tumor samples)

DISCUSSION AND CONCLUSION

- Inspecting the RLE boxplots for all the methods on the full dataset, our framework EDASeq + RUVg reduces clearly the within-sample variability confirming the conclusion about RUV in [5], but also the between-sample variability. Our framework EDASeq + RUVg performs well and leads to a perfect stabilization of read count distributions, and RLE values centered at 0 across the samples under comparison. The PCA plot shows a clear separation of clusters according to the biological factor of interest (tissue status: Normal vs Tumor). As a comparison to other methods, EDASeq + RUVg framework outperforms all the most used methods like RPKM, TMM, and DESeq.

- The two-stage normalization method involving EDASeq and RUVg performs well and may have been specific to our real dataset, and that much more can be done to assess how this method performs in other cases (other datasets) before picking conclusion about this method.

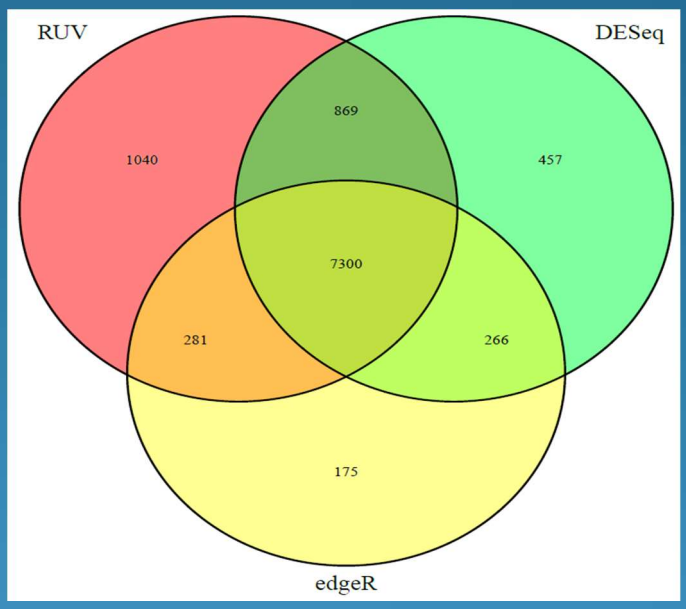

Fig 4. Venn diagram showing the overlapped DEGs between the methods under comparison. The number of DEGs discovered by RUV-based DE analysis is higher than those discovered by the other methods.

REFERENCES

- [1] Tackling the widespread and critical impact of batch effects inhigh-throughput data, Jeffrey T. Leek et al. 2010
- [2] Evaluation of Methods in Removing Batch Effects on RNA-seq Data, Qian Liu and Marianthi Markatou, 2016
- [3] Normalization of RNA-seq data using factor analysis of control genes or samples, Davide Risso et al. 2014
- [4] Removing Unwanted Variation from High Dimensional Data with Negative Controls, Johann A. Gagnon-Bartsch et al. 2013
- [5] A Comparison of Methods: Normalizing High-Throughput RNA Sequencing Data, Rahul Reddy, 2015

ACKNOWLEDGMENTS

Funding was provided by Region Wallonne WalInnov-NACATS (Belgium)