CAGE data analysis of HeLa-Yeast spike-in

Enrichetta Mileti, Marco Salvatore, Yunzhang Wang and Olof Emanuelsson
Karolinska Institutet, Stockholm University and KTH

Here we present a pipeline to analyze CAGE data and three different normalizations. The major change is to identify reads that map on both human and yeast (black-list), in order to exclude those from the final mapping.

Background

- Cap Analysis of Gene Expression (CAGE) is based on a series of full length cDNA technologies [1].
- The spike-in is a transcript used to calibrate measurement in sequencing experiments [2].

Material

- CAGE libraries from HeLa cell lines with 0.1% yeast and 1% yeast respectively.

Results

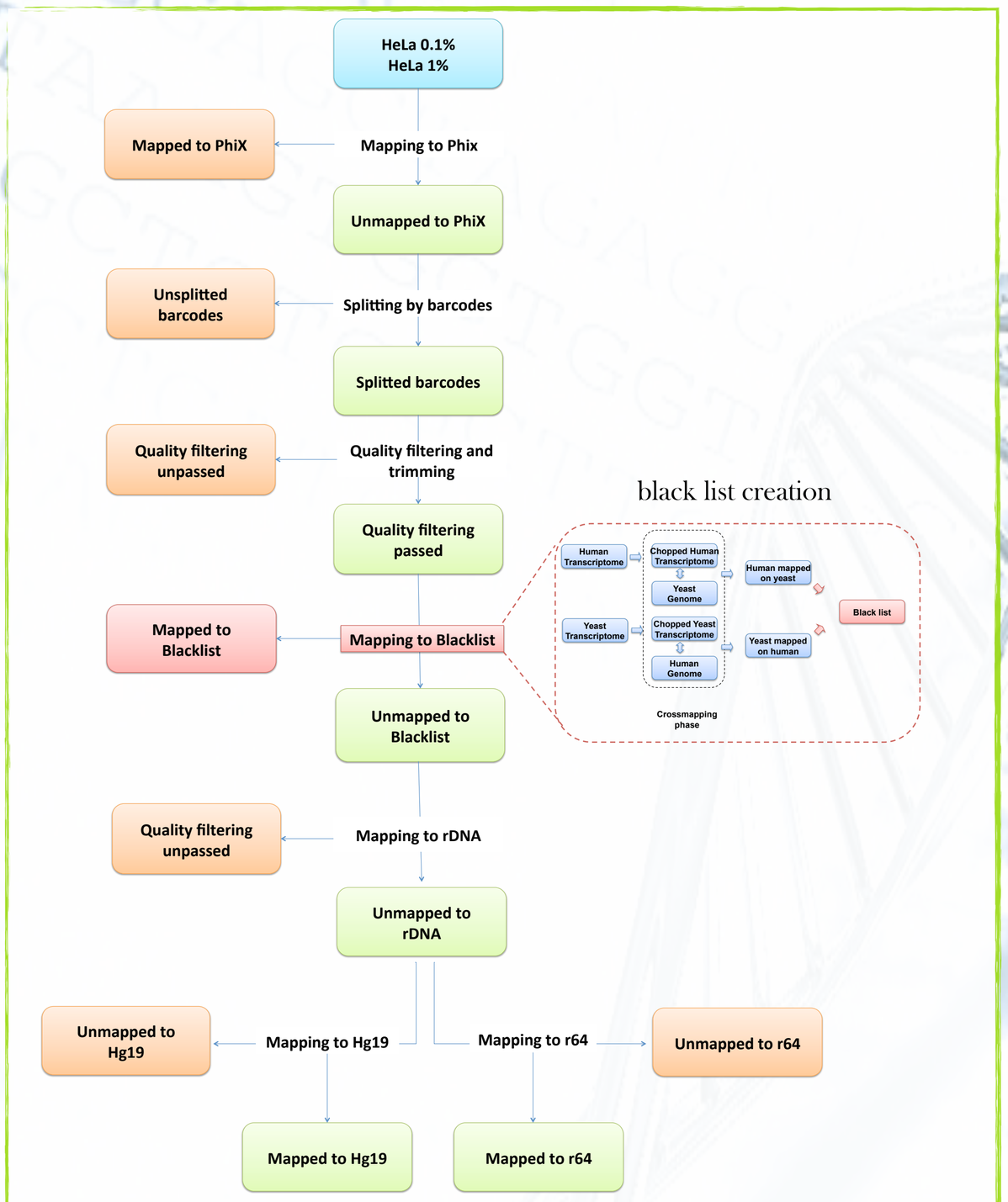

Figure 1: CAGE data analysis pipeline plus the creation of black list

- No influence of the spike-in concentration
- With the black-list we lose some tag cluster, 1 for human and 12 for yeast
- We observed differences in using the black-list in terms of mapped reads.

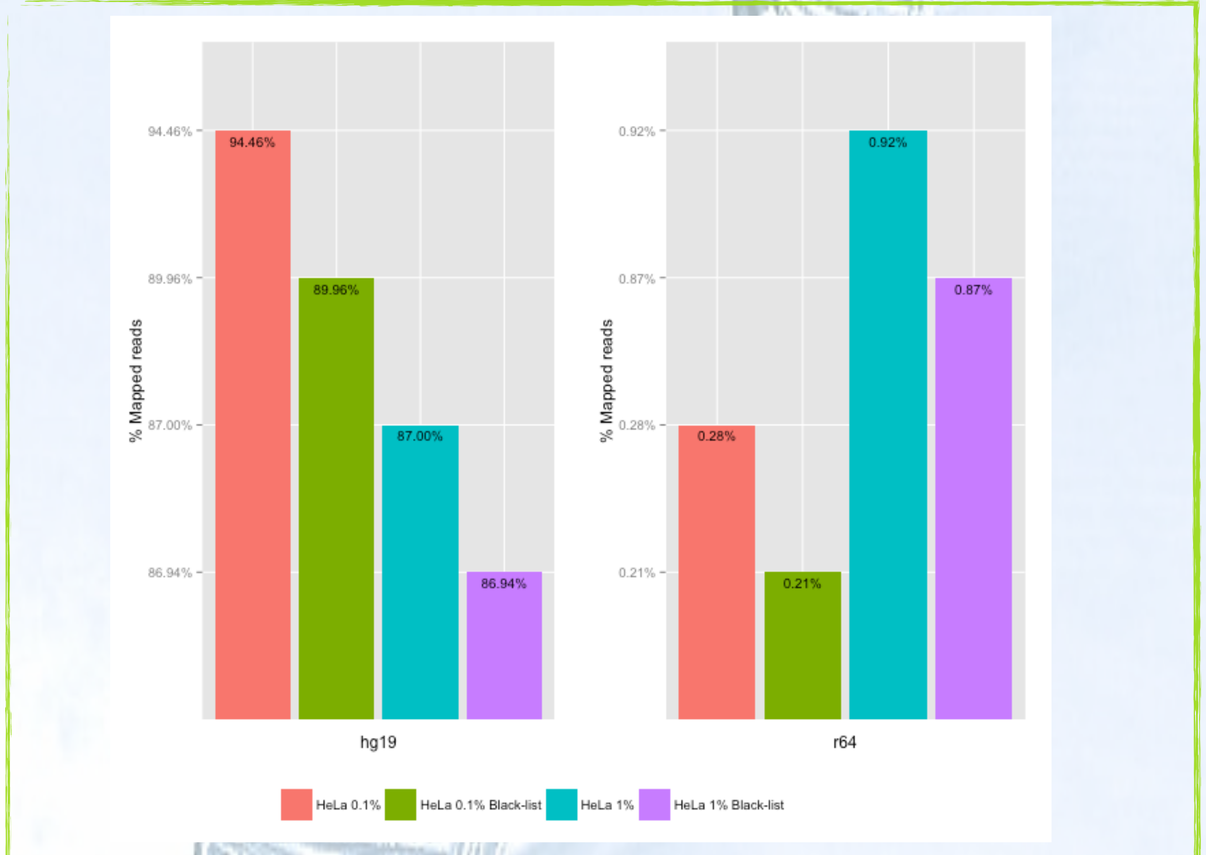

Figure 2: Number of mapped reads in both human and yeast. The left plot show the percentage of read that mapped on human; the right plot show the percentage of read that mapped on yeast.

- We first used the classical normalization for CAGE data such as Tags Per Million (TPM) and Relative Log Expression (RLE)
- Afterwards we normalized the data using the yeast spike-in

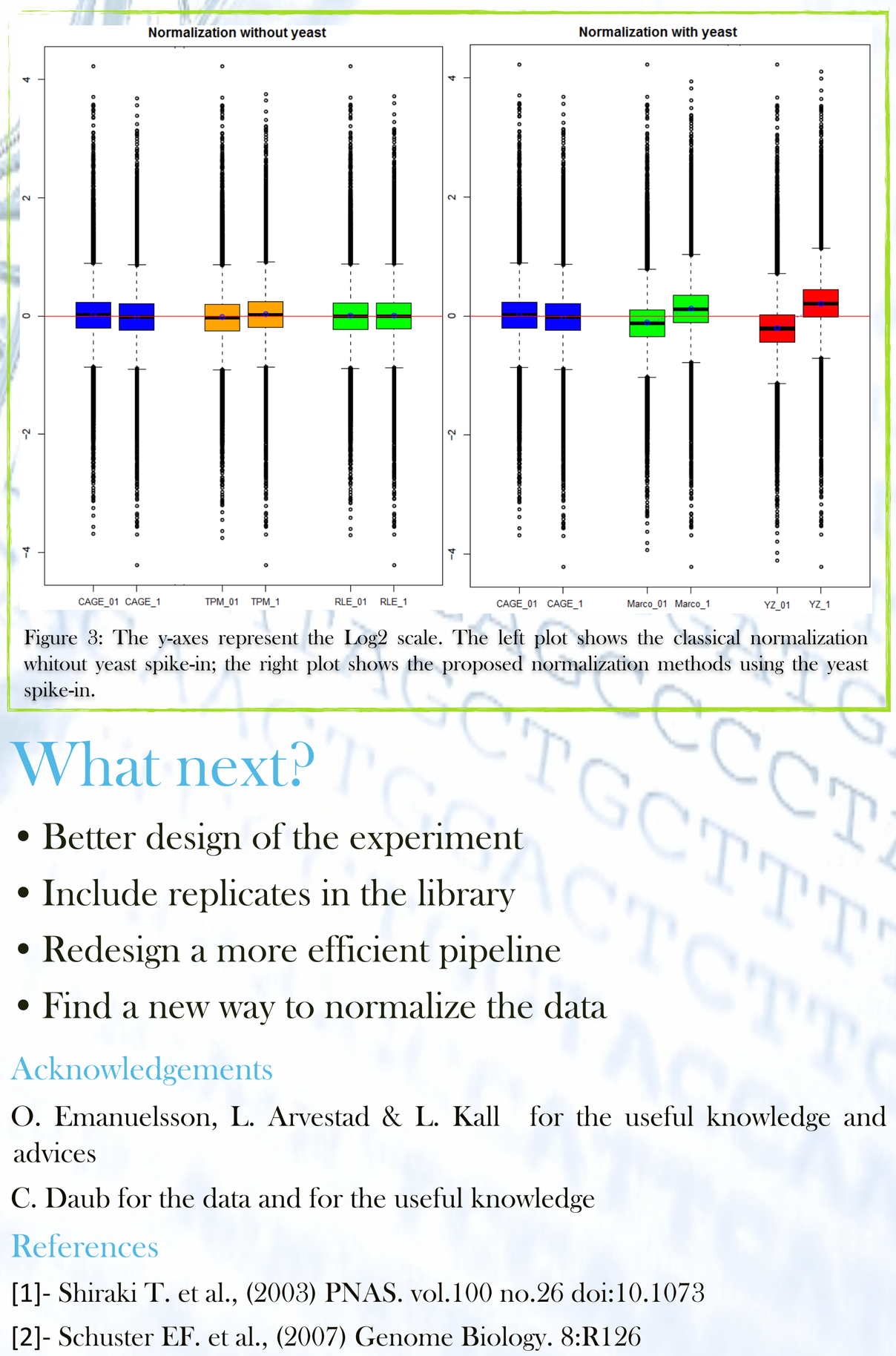

Figure 3: The y-axes represent the Log2 scale. The left plot shows the classical normalization whitout yeast spike-in; the right plot shows the proposed normalization methods using the yeast spike-in.

What next?

- Better design of the experiment
- Include replicates in the library
- Redesign a more efficient pipeline
- Find a new way to normalize the data

Acknowledgements

O. Emanuelsson, L. Arvestad & L. Kall for the useful knowledge and advices

C. Daub for the data and for the useful knowledge

References

[1]- Shiraki T. et al., (2003) PNAS. vol.100 no.26 doi:10.1073

[2]- Schuster EF. et al., (2007) Genome Biology. 8:R126

[3]- Balwierz P.J. et al., (2009) Genome Biology. 10:R79