# Rank-based Bayesian inference for transcriptomic analyses in cancer

Emilie Ødegaard and Valeria Vitelli
joint work with: Thomas Fleischer, Arnoldo Frigessi, Manuela Zucknick

Whole-genome (-omics) profiling of tumors has brought unprecedented information and knowledge of the characteristics of the cancer disease, allowing investigation of the underlying DNA (mutations, copy number alterations and epigenetic alterations) and the phenotype represented by RNA profiling and protein profiling; however, the analysis and interpretation of these data are still a challenge for the field. Pan-cancer studies may aim to identify key molecular characteristics distinguishing subtypes of cancer, and to explore differences and commonalities of tumors originating from different tissues.

To the aims of performing a pan-cancer analysis, a clustering method capable of jointly handling heterogeneous data sources (across tissues) is required. Moreover, Bayesian methods are to be preferred, to allow for the integration and propagation of both data and model uncertainties in the final results. A typical drawback of Bayesian methods, and the reason why their use in integrative genomics is limited, is the computational burden often associated with such procedures.

We first propose to use a Bayesian rank-based approach, BayesMallows (Vitelli et al., 2018; Sørensen et al., 2020), that allows the analysis of RNA-seq data for thousands of tumor samples and genes, performing data integration via the use of ranks, and providing full Bayesian uncertainty propagation via prior and modelling choices. We employ this method for the molecular analysis of 3,299 tumors from 12 cancer types using around 500 selected functional events, previously reported as pan-cancer associated (Ciriello et al., 2013).

Even if the results of such pan-cancer analysis showed quite interesting insights as compared to similar studies, relying on a previously reported selection of relevant genes is a major drawback of BayesMallows. Indeed, inference with the BayesMallows model is computationally intensive when data are very large (the model cannot scale to the data dimension typical of -omics applications – thousands of patients and hundreds of thousands of genes), and a selection of a sub-set of the genes needs to be performed prior to the analysis. In the second part of the talk, I therefore describe a possible way of performing inference on a reduced-dimensional Mallows model, starting from high-dimensional data. This is relevant for some reasons:

(i) computational advantages, since by reducing the dimension of the parameter space, inference is much easier and faster;

(ii) better modelling, since the reduced-dimensional Mallows allows for automatically selecting the items that are "relevant enough" to be included in the distance computation;

(iii) better results interpretation, since a model with fewer parameters enforces a better distinction between noisy and relevant items.

All the above mentioned aspects have a quite substantial importance in -omics applications. The resulting rank-based Bayesian variable selection method is still in its infancy, but it has provided meaningful results on RNA-seq data from ovarian tumor samples in the TCGA database. Extensive testing of the method on synthetic data has also been performed. The inferential approach has yet to be modified for including clustering, and for handling missing data.

## References

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., & Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nature genetics*, *45*(10), 1127-1133.

Sørensen, Ø., Crispino, M., Liu, Q., & Vitelli, V. (2020). BayesMallows: An R package for the Bayesian mallows model. *The R Journal*, *12*(1), 324-342.

Vitelli, V., Sørensen, Ø., Crispino, M., Frigessi Di Rattalma, A., & Arjas, E. (2018). Probabilistic preference learning with the Mallows rank model. *Journal of Machine Learning Research*, *18*(158), 1-49.