Annotation of large-scale transcriptomic data using machine learning techniques

INTRODUCTION

Publicly available mRNA expression profiles provide a treasure trove of genetic information to be explored.
Ideally all of this data would be accompanied by annotation information, but in reality this information is often incomplete or outright missing. To combat this issue, one would need to find a way to infer or predict annotations based on information that *is* available, but this presents a new challenge. Samples in these public datasets are often derived from biopsies where various signals (e.g., biological pathways, non-cancer tissues, nonbiological effects) are intermingled in the final mRNA expression profile. An annotation method like Guilt By Association would therefore be significantly influenced by a dominating signal, producing a biased annotation.

The Fehrmann group at the UMCG collected RNA-seq data from The Cancer Genome Atlas and applied Independent Component Analysis (ICA) to disentangle the intermingled signals in the mRNA expression profiles. This process resulted in a set of independent components called Transcriptional Components (TCs) and a Mixing Matrix (MM) representing the activity of TCs in a given set of samples. A yet to be published correction was applied to the MM to rid the data of platform specific and batch effects.

Various machine learning based methods for the prediction of cancer type annotations were applied to the corrected MM, with the goal of producing a model that can accurately predict cancer type annotations for samples of various origins.

METHODS

The Fehrman group collected 10,817 samples from TCGA and, after pre-processing and quality control, applied consensus-ICA to the dataset. A multinomial logistic regression model, conditional inference trees (Ctrees), and random forests using Ctrees as base learners (Cforests) were trained on a training set of the resulting MM using functions from the `glmnet` and `partykit` R packages.
The models were assessed by predicting cancer types for a testing set and calculating AUC-ROC, top-3 accuracy, Adjusted Rand Index, and Matthews Correlation Coefficient (MCC), with the latter of these being decisive.
The best performing machine learning method was used to carry out a grid search to find the best hyperparameters for this model. The performance of the best of the grid search models was validated using mRNA expressions from microarray (MA) data acquired from the GEO platform GPL570 by the Fehrmann group. They projected the TCs of the TCGA dataset onto the expression profiles of the MA dataset, resulting in an MA MM, to which a correction was also applied. This allowed for validation using data sourced from a platform different from the model's training data.
After examining performance, new models were trained on subsets of the TCGA MM and the MA MM excluding cancer types with fewer than 100 samples. These models were subsequently validated with their opposing dataset.
Lastly, variable importance scores were calculated for both models to gain insight into the models' inner workings.

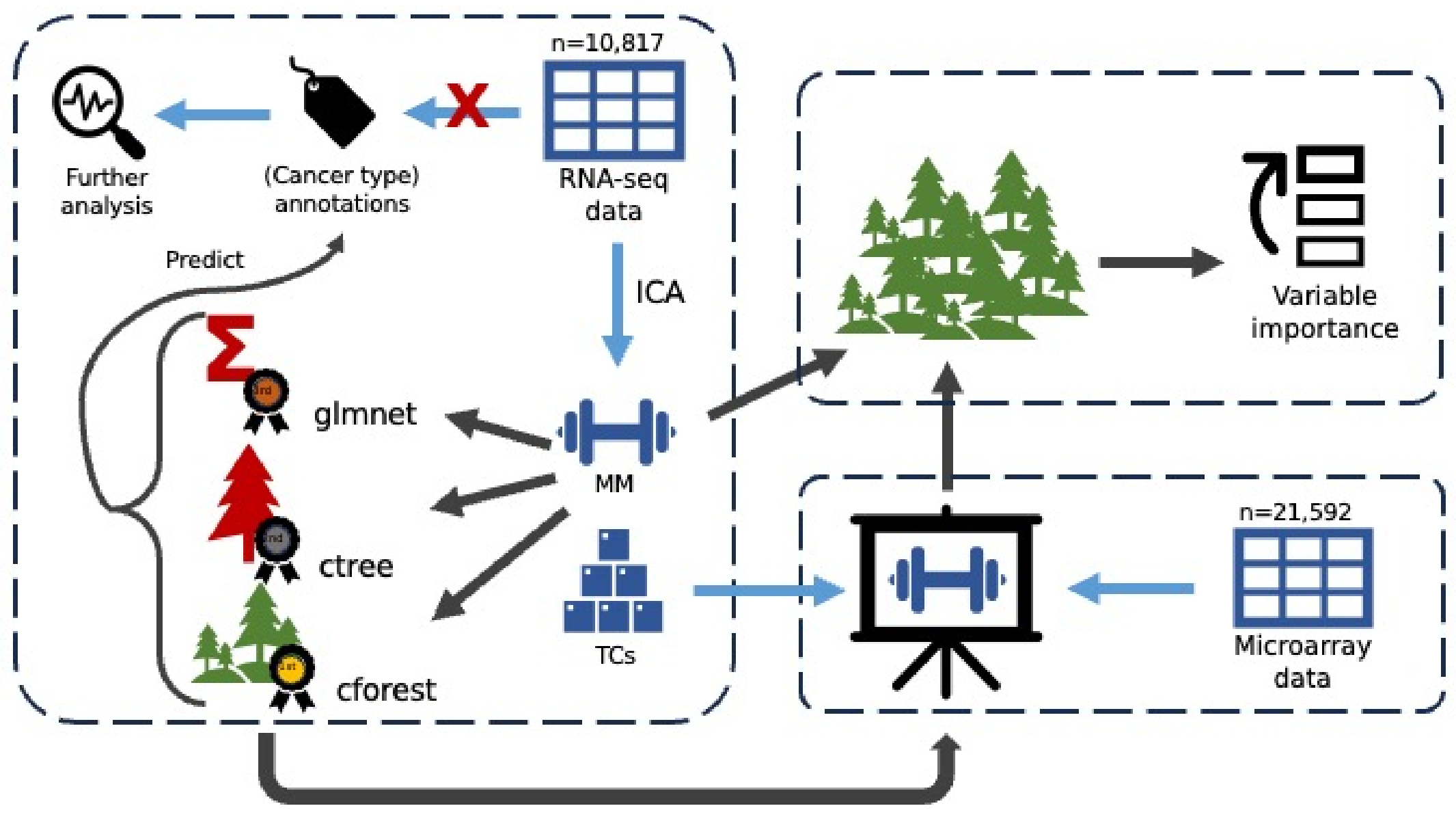

Figure 1 - Visual representation showcasing the methods described above. Blue arrows indicate work done by the Fehrmann group, while grey arrows indicate work done for this project.

CONCLUSION

A machine learning model that can predict cancer type annotations reasonably well was trained on RNA-seq data. This demonstrates the feasibility of this type of prediction, opening the door to predicting other annotations.

RESULTS

After an iterative process of hyperparameter optimalisation, the multinomial logistic regression model produced good results but only converged when trained on a select subset of the complete dataset. A set of Ctree and Cforest models were trained with a given set of hyperparameters under equivalent conditions. The best-performing Cforest model trained with ten trees outperformed the best-performing Ctree model, with the former reaching an MCC of 0.9122 versus the latter's MCC of 0.8687. This gap becomes even more prominent for a Cforest model trained with fifty trees, achieving an MCC of 0.9290.

A grid search was performed to find the optimal hyperparameters for a Cforest model, the results of which can be found in Figure 3 and Figure 4. Figure 3 shows that lower values for both the `minsplit` and the `minbucket` hyperparameters tend to result in higher MCC scores. Figure 4 shows how a higher value for the `ntrees` hyperparameter results in a higher MCC until a point of stagnation is reached after 100 trees. Additionally, setting the `"Univariate"` value for the `testtype` hyperparameter seems to outperform the `"Bonferroni"` setting, although the difference is very slight.

The Cforest model in this grid search, achieving the highest MCC of 0.9329, was selected as the most capable model with the following hyperparameters:

```
testtype="Univariate",
ntrees=1000,
minsplit=10,
minbucket=3.
```

The predicted probabilities of a given sample being assigned a particular cancer type are showcased in the heatmap in Figure 2, providing insight into the model's behaviour.
Validating the model's performance on the MA MM resulted in a much lower MCC of 0.2114 while maintaining a top-3 accuracy of 0.9268 and an AUC-ROC of 0.9796.

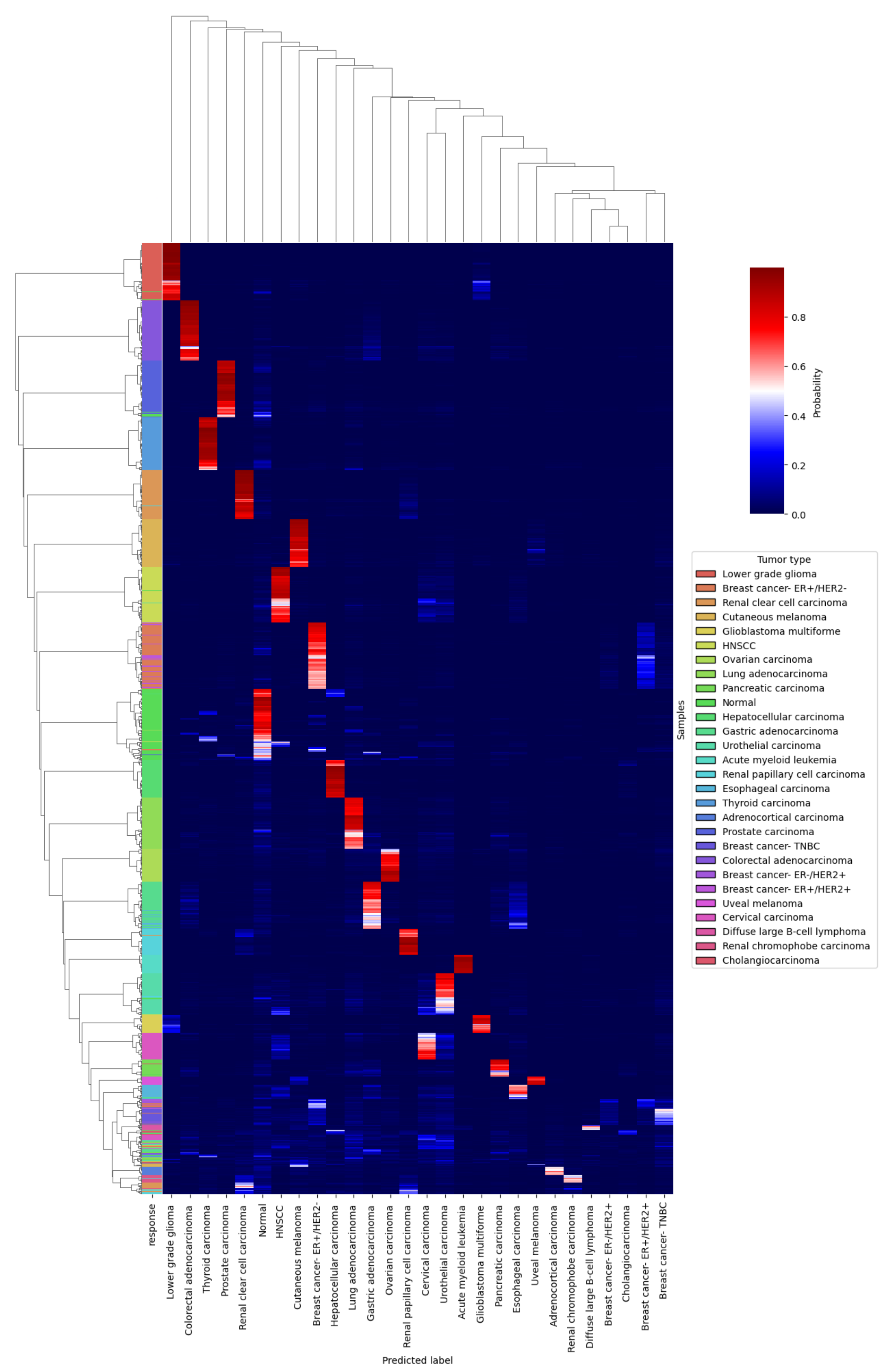

Figure 2 - Heatmap showing the predicted probabilities of a predicted cancer type (x-axis) for a given set of samples (y-axis). Reds indicate high predicted probabilities, while blues indicate low predicted probabilities. Samples and predicted cancer types were clustered with the ward.d2 method using a euclidean distance metric. True cancer type labels are presented in the coloured bar on the left hand side of the heatmap.

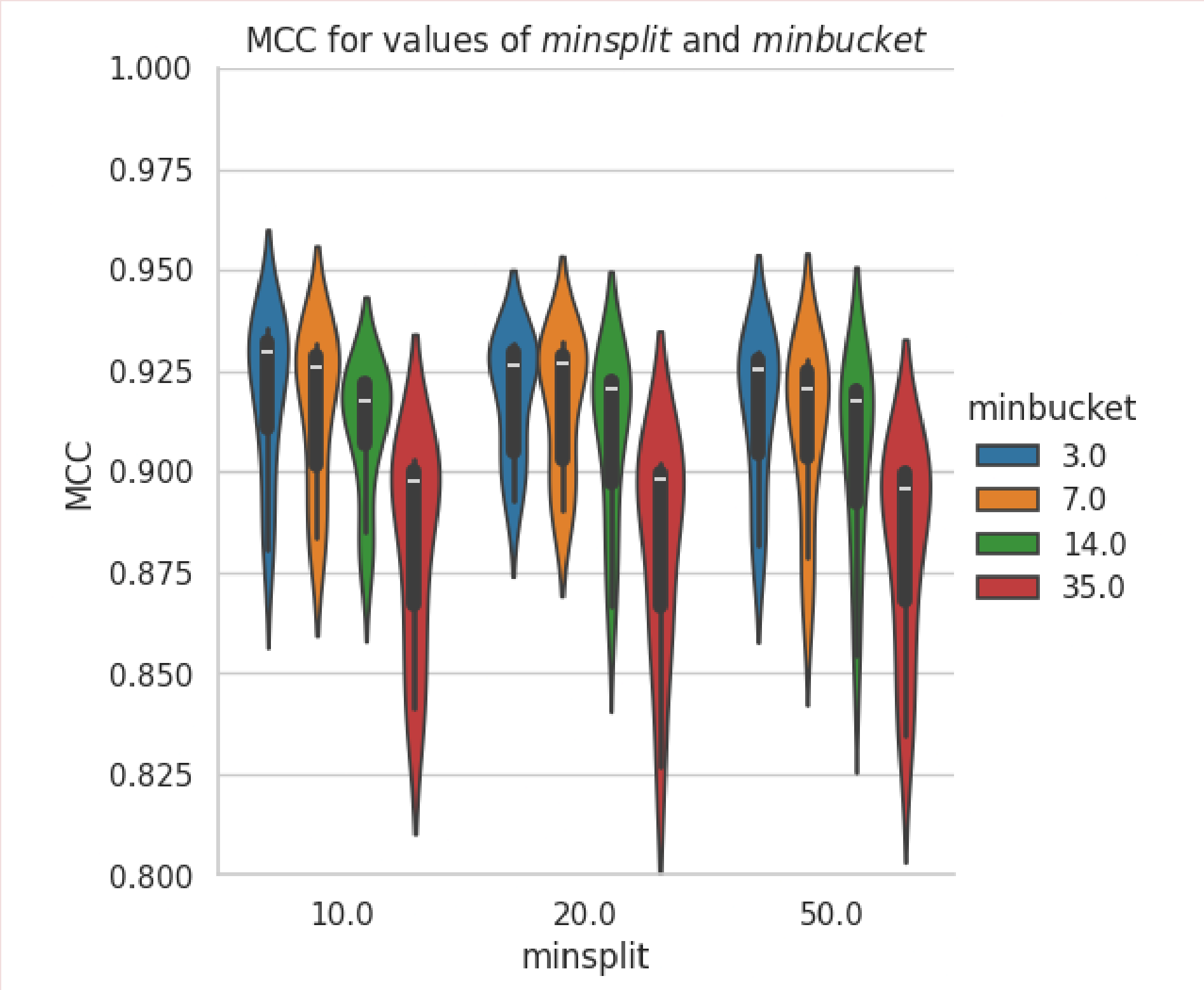

Figure 3 - Results of the grid search to file the optimal parameters for Cforest. Varying values for `minsplit` (x-axis) and `minbucket` (colouring) are compared by their MCC scores (y-axis).

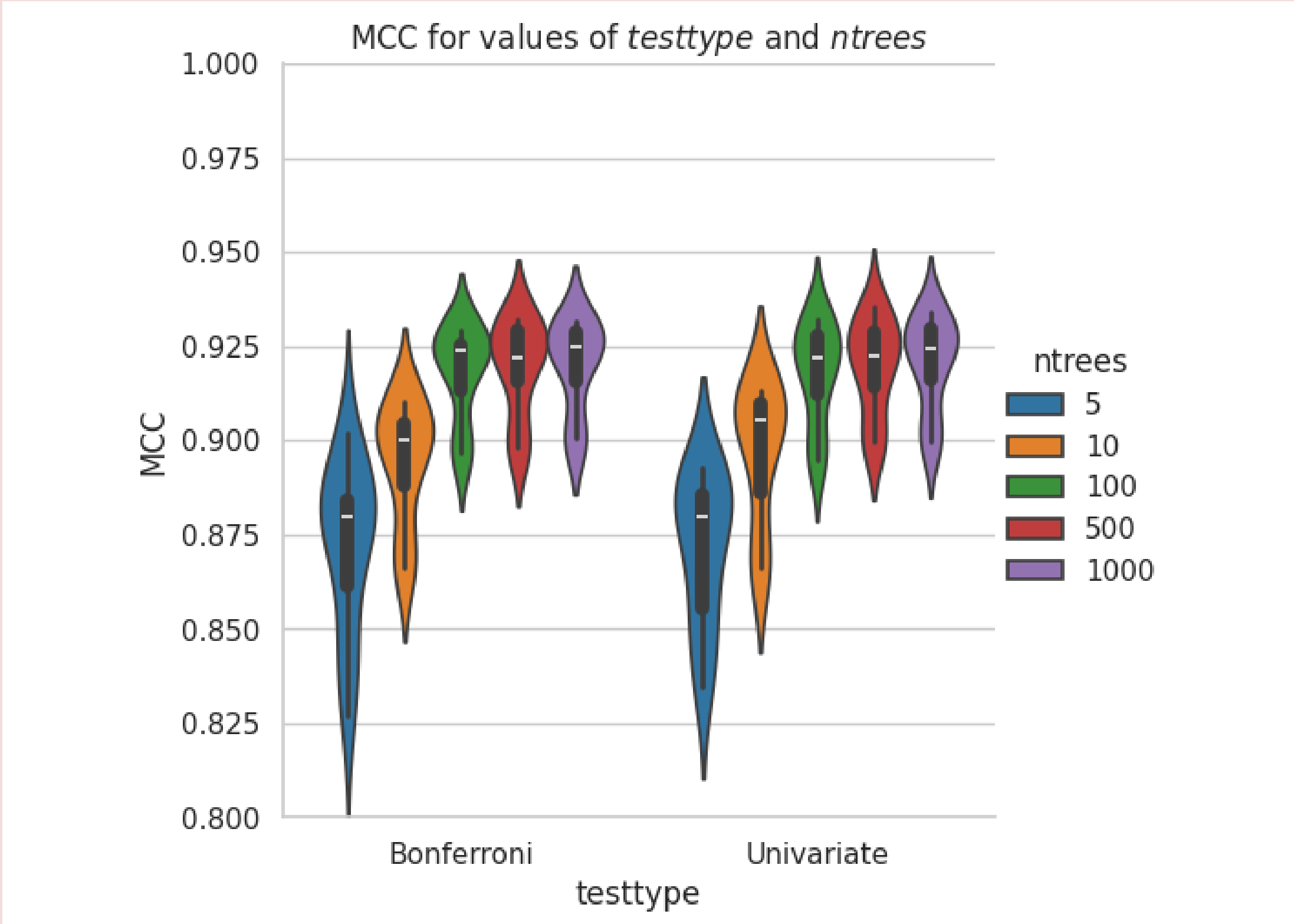

Figure 4 - Results of the grid search to file the optimal parameters for Cforest. Varying values for `testtype` (x-axis) and `ntrees` (colouring) are compared by their MCC scores (y-axis).

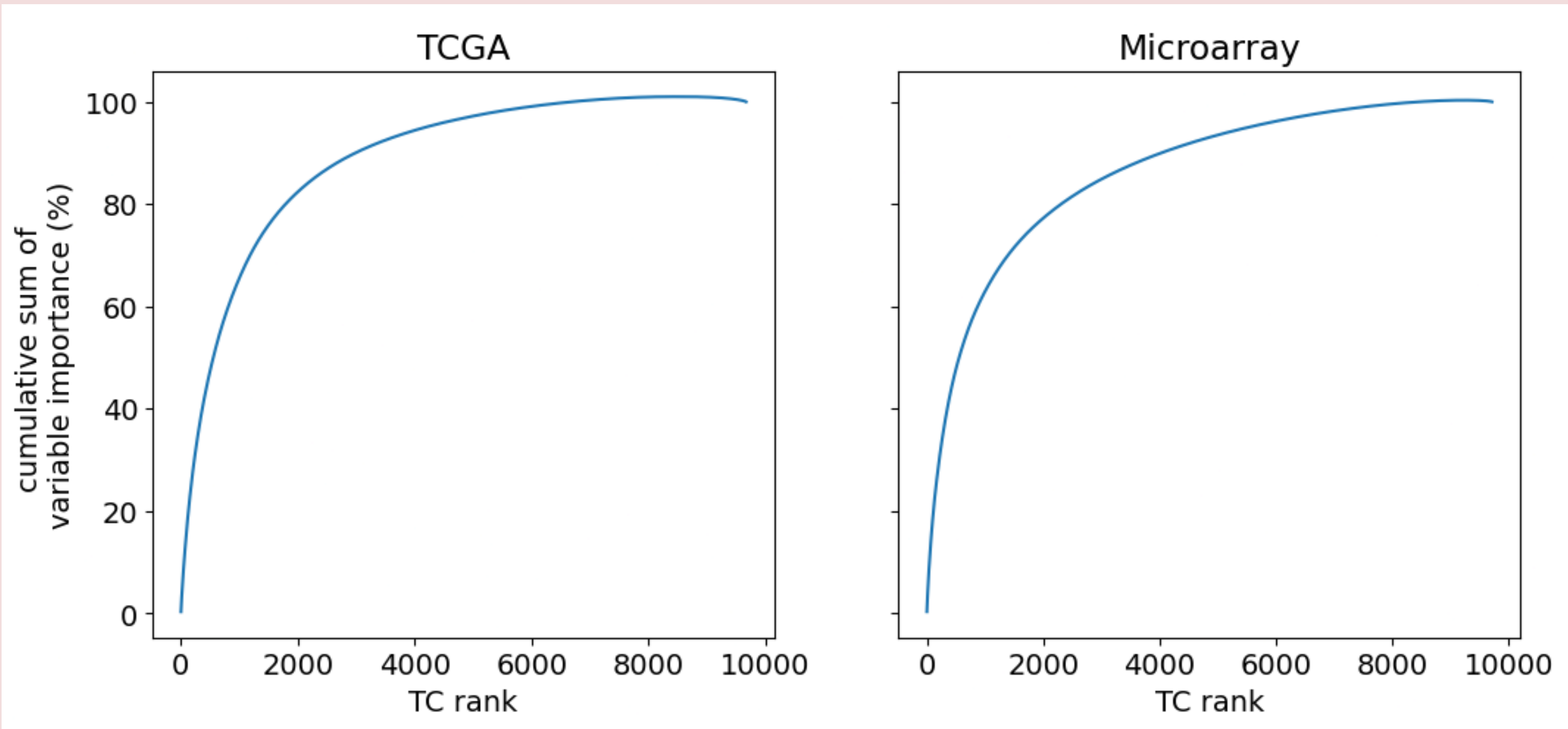

Figure 5 - Result of the variable importance calculations for a Cforest model trained on a subset of the TCGA MM (left) and a subset of the MA MM (right). The line represents the percentage of the cumulative sum of variable importance scores (y-axis) for a given set of TCs sorted by variable importance in descending order (x-axis).

Cforest models were trained on TCGA MM and MA MM subsets using the hyperparameters found in the grid search mentioned above, achieving MCC scores of 0.8260 and 0.8387, respectively. Variable importance scores were calculated for both models following the permutation principle of the mean decrease in accuracy. The percentage of variable importance explained by a given number of TCs ranked from most to least important is showcased in Figure 5, revealing that a relatively small number of TCs have a large impact on the model's outcome.

Dennis Wiersma - denniswiersma@protonmail.com

In collaboration with:
A. Bhattacharya
S. Loipfinger
R.S.N. Fehrmann

Bhattacharya, A., Bense, R. D., Urzúa-Traslaviña, C. G., de Vries, E. G. E., van Vugt, M. A. T. M., and Fehrmann, R. S. N. (2020). Transcriptional effects of copy number alterations in a large set of human cancers. Nature Communications, 11(1):715.

Friedman, J., Tibshirani, R., and Hastie, T. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1):1–22.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics, 15(3):651–674.

Hanzehogeschool Groningen
University of Applied Sciences

umcg