TITLE

Long : BioDiscViz : a visualization support and consensus signature selector for BioDiscML results

Short : BioDiscViz: visualization tool for BioDiscML

Authors

Sophiane Bouirdene[1], Mickael Leclercq[1], Léopold Quitté[1], Steve Bilodeau[2], Arnaud Droit[1*].

1. Département de Médecine Moléculaire CHU de Québec Research Center, Université Laval, Québec, QC, Canada

2. Département de biologie moléculaire, biochimie médicale et pathologie, Faculté de Médecine, Université Laval, Québec, Canada

** Corresponding author*

Abstract

Machine learning (ML) algorithms are powerful tools to find complex patterns and biomarker signatures where conventional statistical methods may fail to identify them. While the ML field made significant progress, state of the art methodologies to build efficient and non-overfitting models are not always applied in the litterature. To this purpose, automatic programs, such as BioDiscML, have been designed to identify biomarker signatures and correlated features while escaping overfitting using multiple evaluation strategies, such as cross validation, bootstrapping and repeated holdout. To further improve BioDiscML and reach a broader audience, better visualization support and flexibility in choosing the best models and signatures are needed. Thus, to provide researchers with an easily accessible and usable tool for in depth investigation of the results from BioDiscML outputs, we developed a visual interaction tool called BioDiscViz. This tool provides summaries, tables and graphics, in the form of Principal Component Analysis (PCA) plots, heatmaps and boxplots for the best model and the correlated features. Furthermore, this tool also provides visual support to extract a consensus signature from BioDiscML models using a combination of filters. BioDiscViz will be a great visual support for research implying machine learning, hence new opportunities in this field by opening it to a broader community.

Introduction

In recent years, new methods of Artificial Intelligence (AI) have been deployed in bioinformatics research to provide pattern classification, biomarker identification and forecast modeling using omics data. Studying biomarker signatures is an important part of the research process as they are correlated to biological functions. Machine learning and feature selection will identify multivariate associations of biomarkers (i.e., features) and detect complex hidden patterns in the data. Considering the existence of many algorithms for feature selection and classification, multiple models are often generated with different signatures, but inconsistent overlaps between signatures despite equivalent performances are frequent (Li et al. 2016). Furthermore, correlated features may not be retained by the models during their optimization when avoiding redundancy of information. Indeed, selecting a "best model" and its signatures is an equilibrium between decomplexifying the model and getting all valuable biomarkers. Often, various approaches, like ensemble learning or union of overlapping features, tend to find optimized solutions but at the cost of either side of the balance.

A solution to facilitate the generation of multiple models and signatures has been proposed with an automatic ML tool, BioDiscML (Leclercq et al. 2019). BioDiscML is a new generation ML tool which has been demonstrated to be highly efficient in multiple research topics involving the identification of biomarker signatures from various types of data, such as proteomics (Roux-Dalvai et al. 2019), transcriptomics (Rabaglino and Kadarmideen 2020) and multi-omics (metagenomics/metabolomics, metagenomics/lipidomics) (Khorraminezhad et al. 2020; Doré et al. 2022). Furthermore, BioDiscML proposes various conditions for choosing a "best model" from its results, but this is complex to determine as some data are too heterogeneous to propose ideal decision threshold metrics. However, this tool does not provide visualization of the signature, hence limiting a rapid view of the results. Thus, to help in these decisions, we propose a visual tool, BioDiscViz, to support the choice of consensus features within a set of trained classifiers with their corresponding signatures.

Design and implementation

BioDiscViz is a visual Shiny application working on Windows and Unix operating systems to support BioDiscML by presenting an interactive interface and graphs to the researchers which will improve their understanding of the results. The application is based on R (R Core Team 2021). It uses the framework Rshiny (Chang W et al 2021) and its dependency Rshiny Dashboard (Chang W, Borges Ribeiro B 2021) and requires Rstudio (RStudio Team 2020), an integrated development environment for R.

Input

BioDiscViz takes as input a directory containing BioDiscML output in csv format and their summary results. The best model and the classification or regression results are independently accessible. Furthermore, the tool supports multiple BioDiscML outputs in the same directory and allows rapid switching between them.

Layout

The application is divided into four main parts which are : short signature, long signature, attribute distribution and consensus signature.

Short signature presents the results of the best model selected by BioDiscML in the form of different plots representing the features selected by the models.

Long signature gives a representation of the correlated biomarkers.

Attribute distribution is an addition of the visualizer which allows to interactively visualize the features that were used the most by the different classifiers tested. The user is able to select different thresholds for the classifiers using the Matthew correlation coefficient and standard deviation metrics. Furthermore, the user determines the number of attributes to match the experimental design.

Consensus signature gives a representation of the different signatures which were called by the most classifiers under user-determined parameters.

Representations

There is a heatmap, a PCA graph and a boxplot to represent the short, long and consensus signatures. The heatmap was made using ComplexHeatmap (Gu, Eils, and Schlesner 2016), an user-friendly package for better representation of heatmaps. The PCA and boxplots were built using FactoExtra (Kassambara 2020) and ggplot2 (Wickham H 2016) respectively.

The attribute distribution is represented under the form of a UpsetR plot (Conway, Lex, and Gehlenborg 2017) (Fig1.B). UpsetR is a R package generating static upset plots to visualize the intersections between the different features in the different classifiers.

BioDiscViz also gives access to the summary details for the short signature and an interactive table of the data used for the short and long signature.

Considering that non-numerical features cannot be easily integrated into PCA and heatmap with other numerical values, a particularity of BioDiscViz is the transformation of categorical features into numerical ones. This form allows one to simply annotate them on the side of the heatmaps to integrate the information contained by these features into the clustering of PCA.

Outputs

BioDiscViz also possesses different functions to facilitate use and export of the results for archiving, sharing and publication. The first one is the creation of a report of the different graphs represented in the application and takes into account the modifications carried out by the user. The second functionality is to be able to download a sub dataset containing the information for the selected features in the “attribute distribution”.

The study of consensus signatures has a great interest as it allows researchers to identify new molecular targets of interest. If the best model provides a vision of which useful data were selected, the model does not necessarily use the features providing the most information. As we consider that the most frequently used signatures by the classifiers contain important information for our problem. Forcing BioDiscML’s models to use these consensus signatures could potentially improve the initial results obtained by the previous best model.

Results

To demonstrate the functionalities of BioDiscViz, we used a colon cancer dataset (Alon et al. 1999) which was used for the BioDiscML publication and which is available on BioDiscViz gitlab. This dataset contains gene expression in 40 tumor and 22 normal colon tissue samples.

Visualize the best signature

The identification of the best signatures were studied from two perspectives. First, the signatures from the best model followed by the consensus signatures.

For signatures retrieved from the best model, different plots were generated. In this case, two classes were cleanly separated on the PCA and the heatmap (Fig 1.A), showing that they provide enough information to the model to correctly predict tumor tissues and healthy tissues. Then, differential expressions of genes identified in the model were visualized using boxplots (Fig.2). Interestingly, all the signatures showed promising results as there is a clear difference for each gene between the two classes.

Consensus signature

BioDiscML uses ensemble methods to create an association of signatures, but it does not take advantage of all generated models during its learning stage. Ensemble methods also keep all features of the model's signatures, without any optimization, thus complexifying the model. We made the hypothesis that the features which were called the most frequently by the different models were containing useful information for the problem treated.These signatures, which we call consensus signatures, could represent an interesting option to rebuild a new signature and improve or simplify the model. To look further into the signatures, it is possible to generate a dataset of consensus signatures directly with BioDiscViz. After different tests, it was decided to focus on the 10 best signatures obtained from the classifiers which were passing the threshold of MCC>=0.76 (Matthew's Correlation Coefficient) and STD_MCC<=0.15 (standard deviation of MCC) (Fig.1B). The quality of the selection was assessed using the heatmap (Fig.3A) and PCA (Fig. 3B), which presented a better separation between the classes than the best model identified by BioDiscML. Compared to the best model signatures, these consensus signatures consist of 3 genes overlapping with the best signature, and 7 genes have been newly added. To further look into these new signatures it was possible to use the boxplot to only select genes which seemed to have a clear separation in the distributions between the tissues with and without cancer.

Following the identification of the consensus signature, we re-runned BioDiscML to find an optimal machine learning classifier with the full signature, without any feature selection. The best classifier was a Kstar model with 6 attributes signature which had a MCC of 0.776 with a standard deviation across (STD) all evaluation procedures of 0.037, which is a reasonable performance considering past work on MCC evaluations (Schober, Boer, and Schwarte 2018). Furthermore, the model had an accuracy of 0.8571 which is comparable or even better compared to the literature for this type of data ("Detection of Effective Genes in Colon

Cancer: A Machine Learning Approach" 2021). With the consensus signature, obtained a Fuzzy Lattice Reasoning model with a MCC of 0.791 (STD 0.032) which is slightly better than the previous best model (MCC increased by 1,9% and STD decreased by 15,6%).

In Conclusion our tool is able to give visual support to BioDiscML and new insights outside of the best model by looking into the consensus signatures. Furthermore, these consensus signatures could be used to rerun BioDiscML and may enhance the quality of the model.

Availability and Future Directions

Biodisc-Viz is directly implemented in R and can be downloaded from Gitlab (https://gitlab.com/SBouirdene/biodiscviz.git)

References

Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. 1999. "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays." *Proceedings of the National Academy of Sciences of the United States of America* 96 (12): 6745–50.

Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B (2021). _shiny: Web Application Framework for R_. R package version 1.7.1

Chang W, Borges Ribeiro B (2021). _shinydashboard: Create Dashboards with 'Shiny'_. R package version 0.7.2

Conway, Jake R., Alexander Lex, and Nils Gehlenborg. 2017. "UpSetR: An R Package for the Visualization of Intersecting Sets and Their Properties." *Bioinformatics* 33 (18): 2938–40.

"Detection of Effective Genes in Colon Cancer: A Machine Learning Approach." 2021. *Informatics in Medicine Unlocked* 24 (January): 100605.

Doré, Etienne, Charles Joly-Beauparlant, Satoshi Morozumi, Alban Mathieu, Tania Lévesque, Isabelle Allaeys, Anne-Claire Duchez, et al. 2022. "The Interaction of Secreted Phospholipase A2-IIA with the Microbiota Alters Its Lipidome and Promotes Inflammation." *JCI Insight* 7 (2). https://doi.org/10.1172/jci.insight.152638.

Gu, Zuguang, Roland Eils, and Matthias Schlesner. 2016. "Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data." *Bioinformatics* 32 (18): 2847–49.

Kassambara, A. and Mundt, F. (2020) Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R Package Version 1.0.7.

Khorraminezhad, Leila, Mickael Leclercq, Sarah O'Connor, Pierre Julien, S. John Weisnagel, Claudia Gagnon, Arnaud Droit, and Iwona Rudkowska. 2020. "Dairy Product Intake Modifies Gut Microbiota Composition among Hyperinsulinemic Individuals." *European*

Journal of Nutrition 60 (1): 159–67.

Leclercq, Mickael, Benjamin Vittrant, Marie Laure Martin-Magniette, Marie Pier Scott Boyer, Olivier Perin, Alain Bergeron, Yves Fradet, and Arnaud Droit. 2019. “Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional OMICs Data.” *Frontiers in Genetics* 10 (May): 452.

Li, Ying Hong, Jing Yu Xu, Lin Tao, Xiao Feng Li, Shuang Li, Xian Zeng, Shang Ying Chen, et al. 2016. “SVM-Prot 2016: A Web-Server for Machine Learning Prediction of Protein Functional Families from Sequence Irrespective of Similarity.” *PLOS ONE*. https://doi.org/10.1371/journal.pone.0155290.

Rabaglino, Maria B., and Haja N. Kadarmideen. 2020. “Machine Learning Approach to Integrated Endometrial Transcriptomic Datasets Reveals Biomarkers Predicting Uterine Receptivity in Cattle at Seven Days after Estrous.” *Scientific Reports* 10 (1): 1–10.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Roux-Dalvai, Florence, Clarisse Gotti, Mickaël Leclercq, Marie-Claude Hélie, Maurice Boissinot, Tabiwang N. Arrey, Claire Dauly, et al. 2019. “Fast and Accurate Bacterial Species Identification in Urine Specimens Using LC-MS/MS Mass Spectrometry and Machine Learning.” *Molecular & Cellular Proteomics: MCP* 18 (12): 2492–2505.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA

Schober, Patrick, Christa Boer, and Lothar A. Schwarte. 2018. “Correlation Coefficients: Appropriate Use and Interpretation.” *Anesthesia and Analgesia* 126 (5): 1763–68.

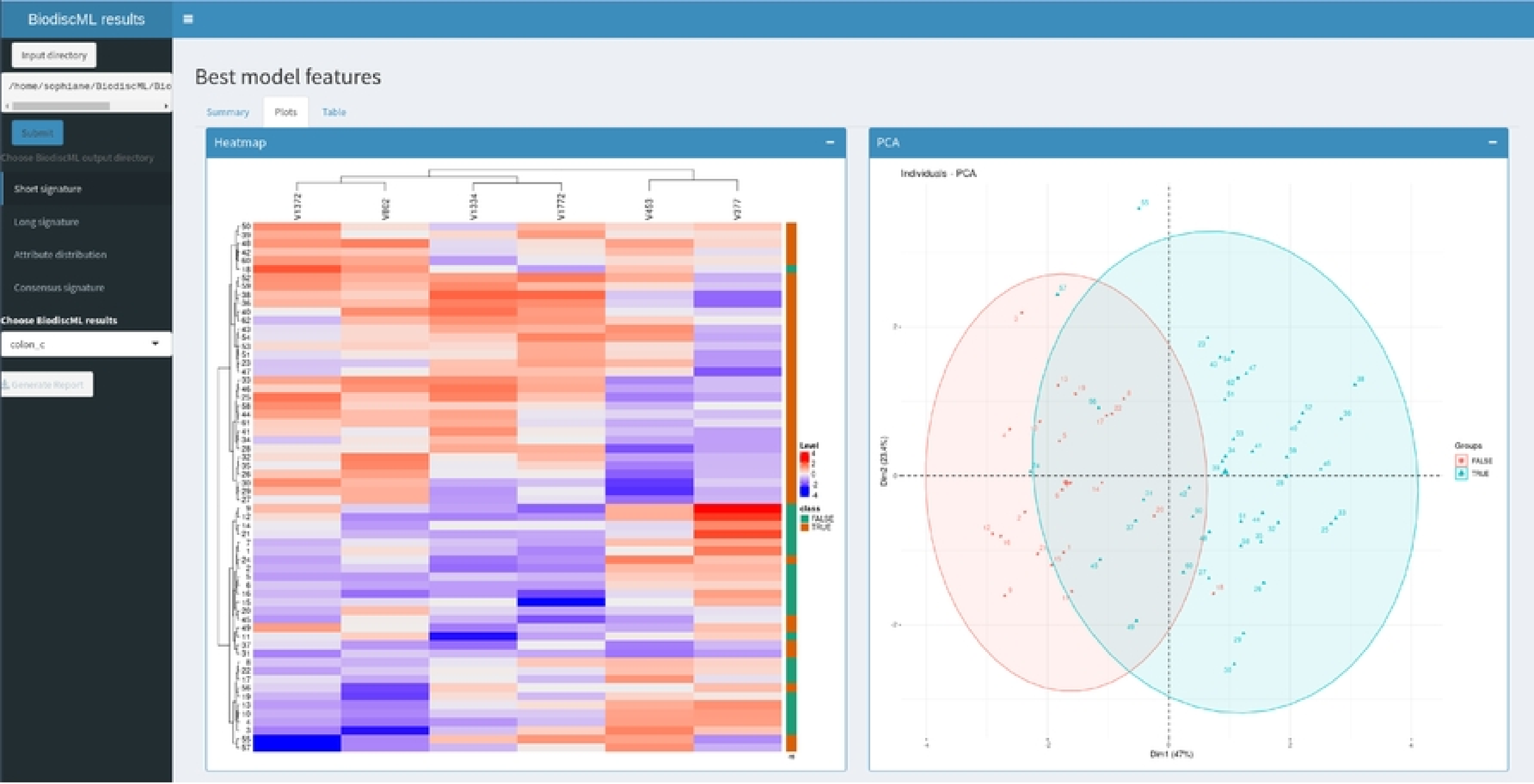
Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4.

Figures

Figure 1: A. Heatmap and PCA by BiodiscViz on the best model found by BiodiscML, Kstar model,on the colon cancer dataset. We observe in cyan color the tumor tissues and in red the normal ones. **B.** Selection of the consensus signatures with BiodiscViz on the colon cancer dataset. Here were selected the 10 attributes most frequently called by the classifiers passing the threshold of a Matthew Correlation Coefficient >= 0.75 and a Standard Deviation <= 0.15.

Figure 2: Boxplot of the best model obtained by BioDiscML

Figure 3: Graphical representation of the 10 best consensus signatures **A.** Heatmap. **B.** PCA

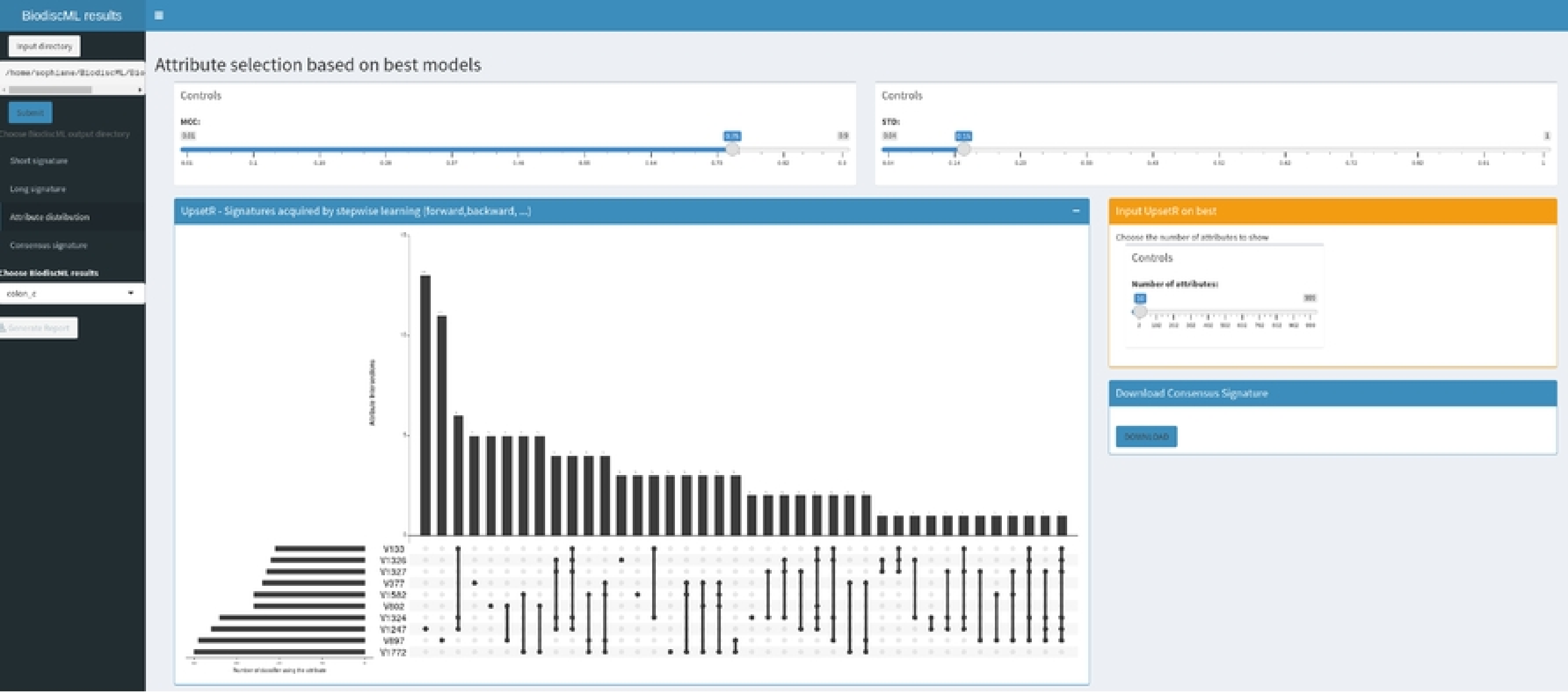


/home/sophiane/BiodiscML/Bio

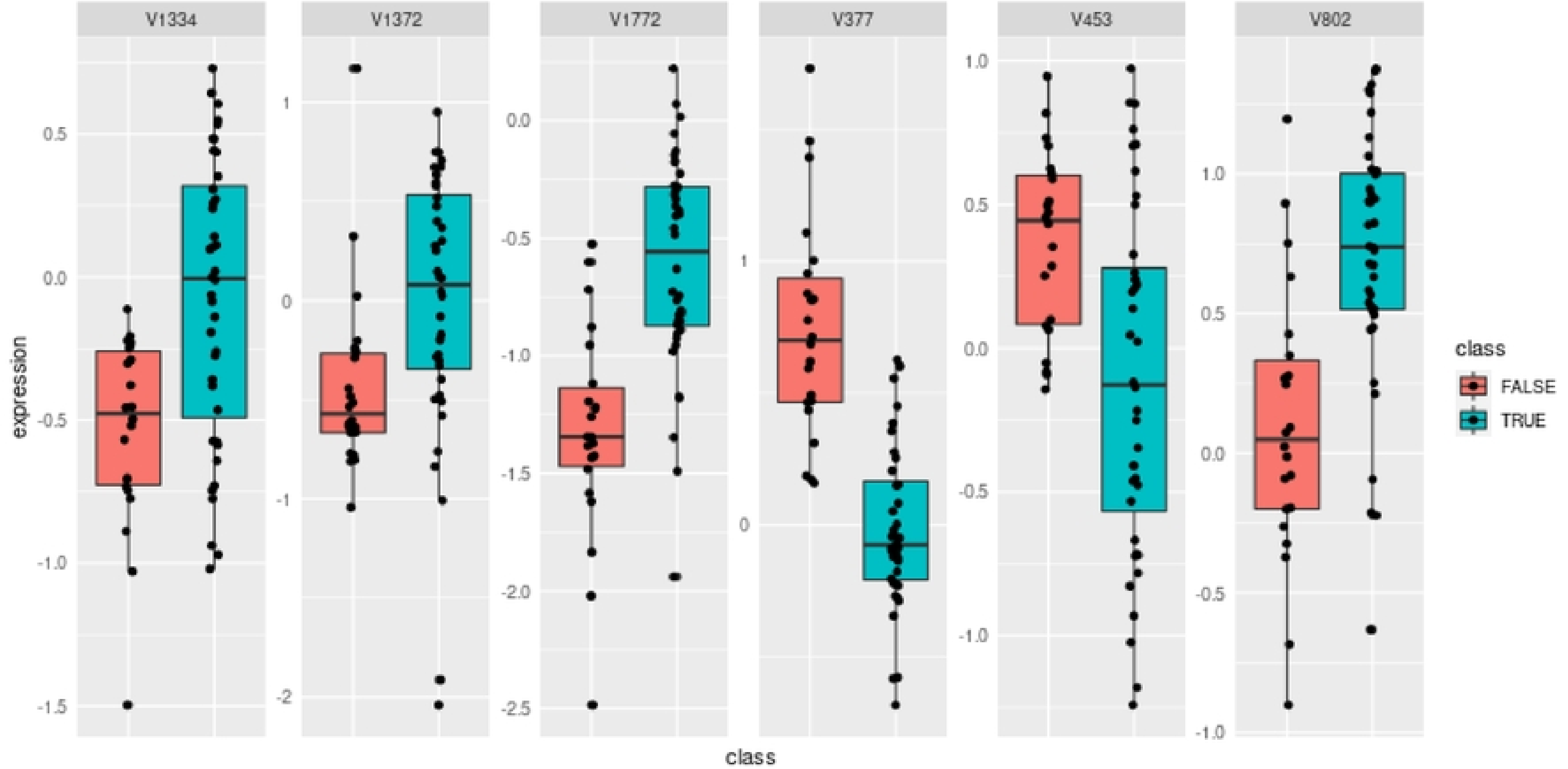
colon_c

Best model features

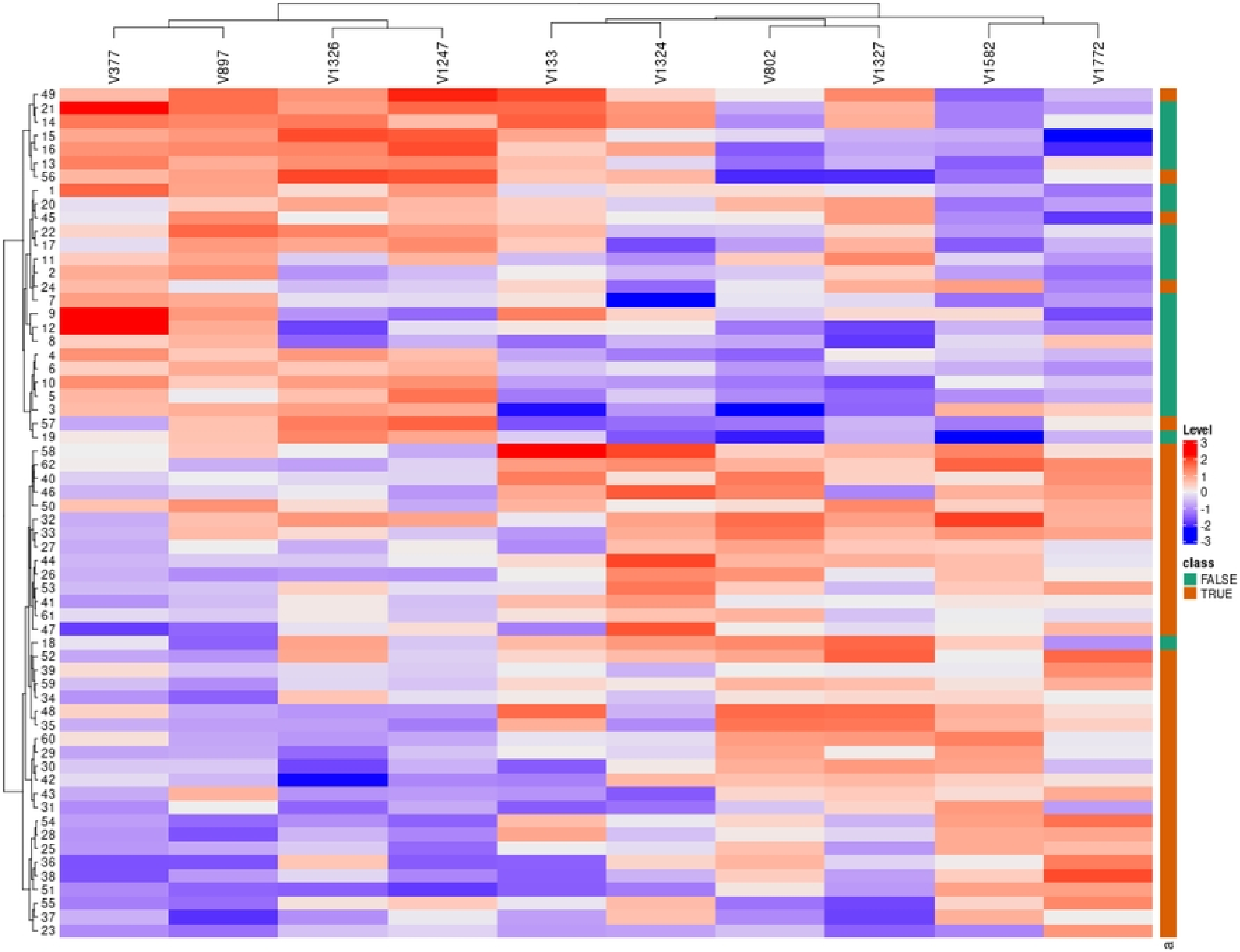
Figure

BiodiscML results
Input directory
/home/sophiane/BiodiscML/Bio
Submit
Choose BiodiscML output directory
Short signature
Long signature
Attribute distribution
Consensus signature
Choose BiodiscML results
colon_c
Generate Report
Attribute selection based on best models
Controls
MCC:
0.01
Controls
STD:
0.01
1
UpsetR - Signatures acquired by stepwise learning (forward,backward, ...)
Attribute Intersections
V133
V1326
V1327
V377
V1582
V802
V1324
V1247
V897
V1772
Number of classifier using the attribute
Input UpsetR on best
Choose the number of attributes to show
Controls
Number of attributes:
Download Consensus Signature
DOWNLOAD

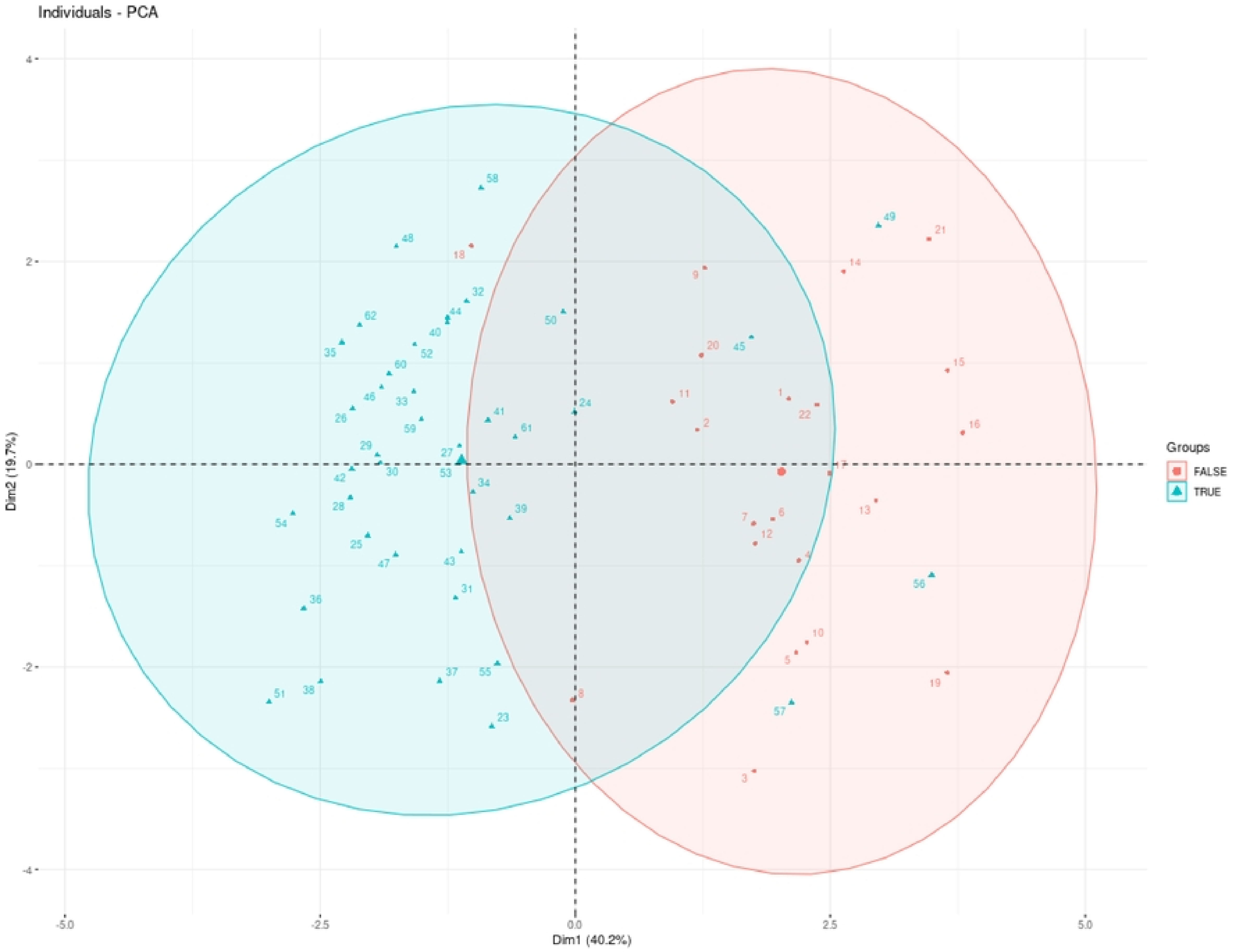
Figure


V1334
V1372
V1772
V377
V453
V802
expression
class
class
FALSE
TRUE
0.5
0.0
-0.5
-1.0
-1.5
1
0
-1
-2
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
1
0
1.0
0.5
0.0
-0.5
-1.0
1.0
0.5
0.0
-0.5
-1.0

Figure


V377
V897
V1326
V1247
V133
V1324
V802
V1327
V1582
V1772
49
21
14
15
16
13
56
1
20
45
22
17
11
2
24
7
9
12
8
4
6
10
5
3
57
19
58
62
40
46
50
32
33
27
44
26
53
41
61
47
18
52
39
59
34
48
35
60
29
30
42
43
31
54
28
25
36
38
51
55
37
23
Level
3
2
1
0
-1
-2
-3
class
FALSE
TRUE
a

Figure


Individuals - PCA
Dim2 (19.7%)
Dim1 (40.2%)
4
2
0
-2
-4
-5.0
-2.5
0.0
2.5
5.0
Groups
FALSE
TRUE
58
49
21
48
18
14
9
32
44
62
50
40
35
52
20
45
15
60
33
46
11
1
26
41
24
22
59
61
2
16
29
27
53
30
42
17
34
39
28
6
7
54
12
13
25
4
47
43
56
31
36
10
5
37
55
19
51
38
8
57
23
3

Figure