APPLICATION OF CLASSIFICATION TREES TO MULTIVARIATE COMPARISON OF HCS DATA

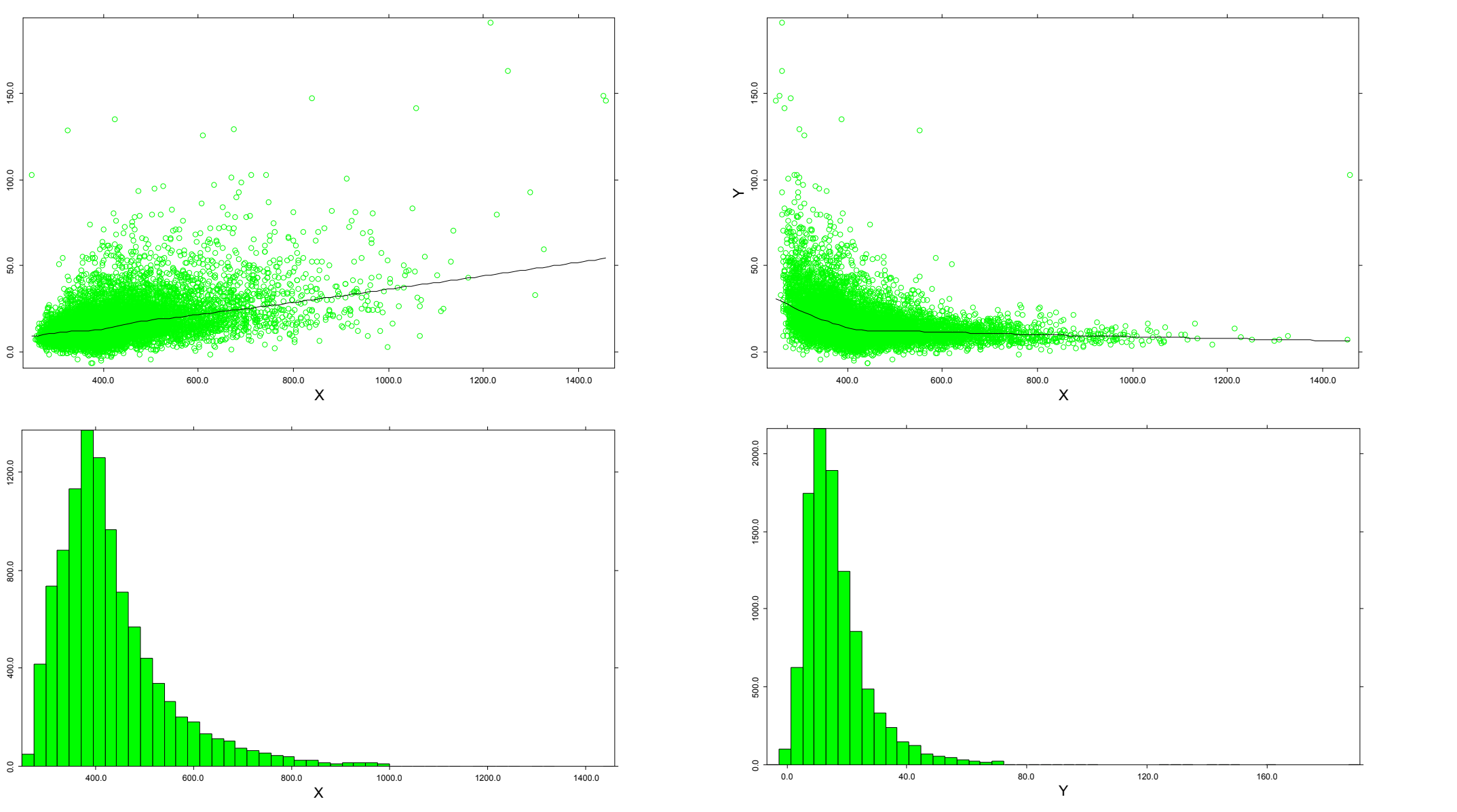
The Chi-Square Works, Inc. (http://chi-square-works.com)

Abstract

HCS data sets are multivariate in nature. All the variables have to be considered JOINTLY to effectively and properly use HCS data for any two-sample tests. This poster demonstrates a novel application of classification trees to HCS data, using dose response analysis as an example. The technique of classification trees has 3 unique advantages in HCS data analysis: 1) it performs multivariate two-sample comparison, 2) it outputs measures of importance for ALL the variables involved, and 3) it gives succinct characterizations of the conditions that drive a cellular phenomenon.

Introduction

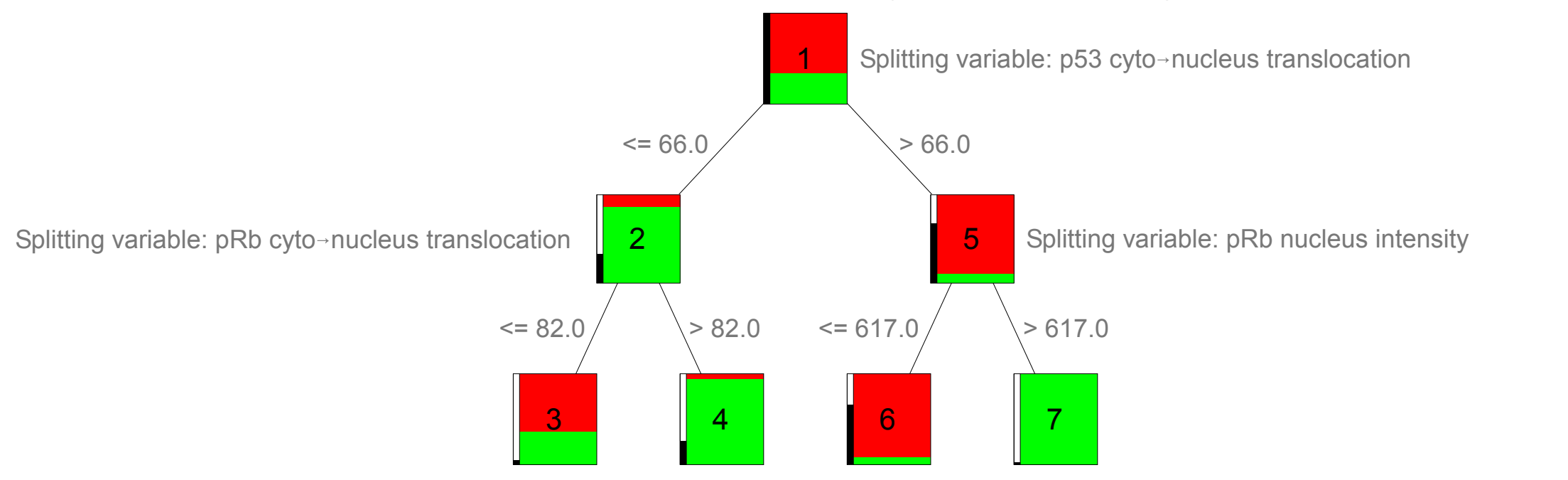
- HCS data are inherently multivariate: Hundreds to thousands of cells in each well of microplates are imaged in multiple fluorescent channels; tens or hundreds parameters are reported for each cell.
- Histograms and Kolmogorov-Smirnov (KS) tests are frequently used to compare HCS (and flow cytometry) data.
- These methods are based on the marginal distribution of a SINGLE variable ONLY and do not take relationships between variables into account. Quite likely, important information is not revealed as a result.
- When comparing 2 samples of multivariate data, similar-looking histograms (hence, nonsignificant KS statistics) for each of the variables do not necessarily imply the same population. The following data come from 2 different populations but have the same X and Y histograms:



- We should examine the JOINT distributions of HCS variables both ANALYTICALLY and GRAPHICALLY. These can be achieved with advanced statistical techniques such as classification trees and multidimensional scaling.

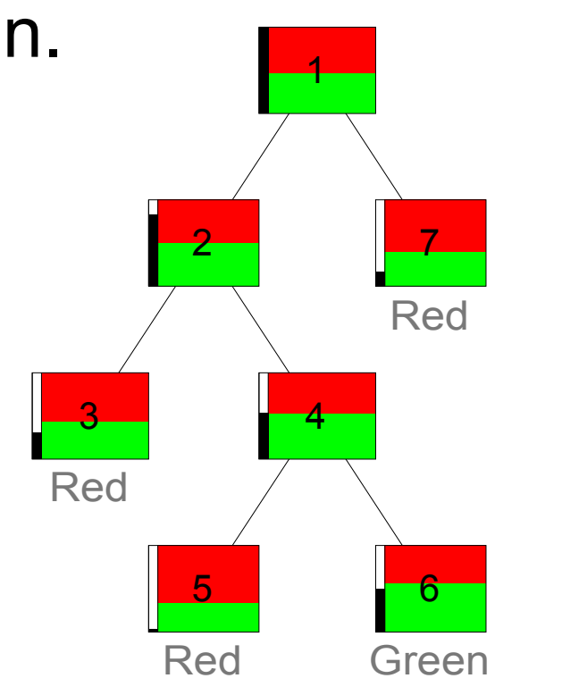
Classification Trees

- Given a set of observations that belong to 2 classes (C_1 and C_2), a classification tree recursively splits the observations based on a variable value test into 2 subsets where the combined "impurity" of the 2 subsets is less than the impurity of the 2 subsets pooled together.
- Impurity of a set of data is defined to be $1 - p^2 - q^2$, where p and q are the proportion of C_1 and C_2 observations in this data set, respectively (hence, p + q = 1).
- Example: 1359 (red) cells treated by etoposide and 720 (green) cells treated by vinblastin.
  - 2 classes: etopside (red) vs. vinblastin (green)
  - 8 variables are used to grow a classification tree; only 3 show up in the final tree.
  - Misclassification rate: 0.089
  - The most important variable: p53 cyto→nucleus translocation
  - The least important variable: p53 cyto intensity

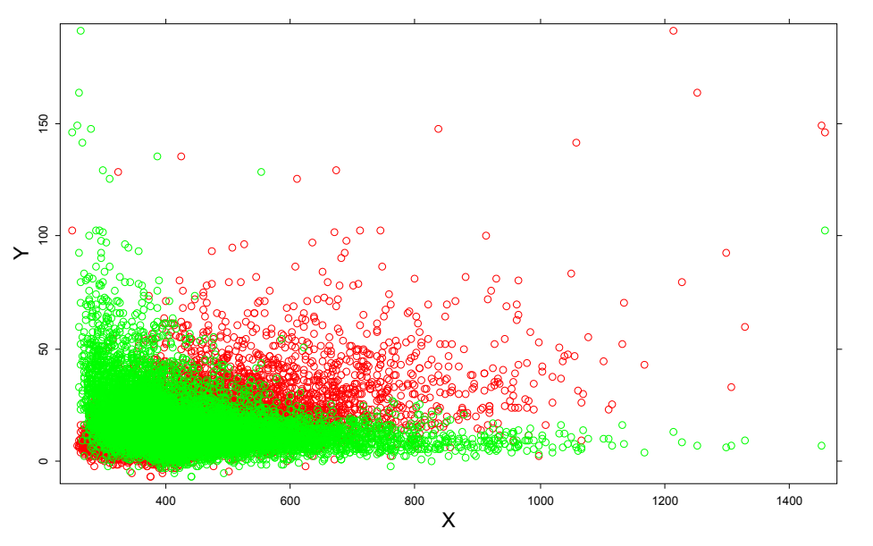

1
Splitting variable: p53 cyto→nucleus translocation
<= 66.0
> 66.0
Splitting variable: pRb cyto→nucleus translocation
2
5
Splitting variable: pRb nucleus intensity
<= 82.0
> 82.0
<= 617.0
> 617.0
3
4
6
7

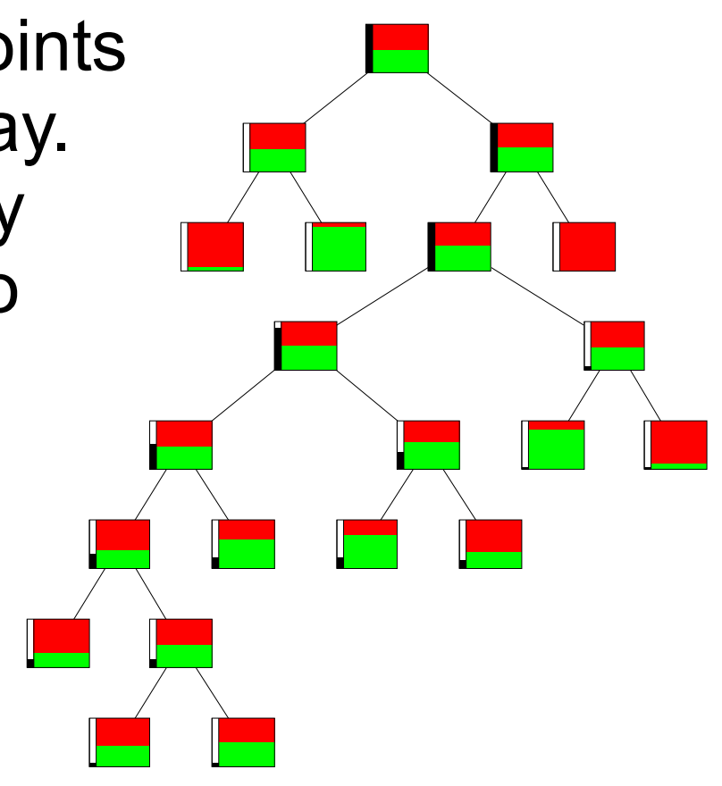
Rationale of 2-Sample Tests by Classification Trees

- If 2 samples do not differ from each other, a classification tree will give a misclassification rate close to that of majority vote.
  - Example: 1605 cells in the same well treated by etoposide are randomly assigned to 2 groups: red and green.
    - 775 green cells vs. 830 red cells
    - Majority vote (every cell is red) with misclassification rate 0.483, which is 775 / (775 + 830).
    - Misclassification rate of a classification tree grown with 11 variables: 0.463



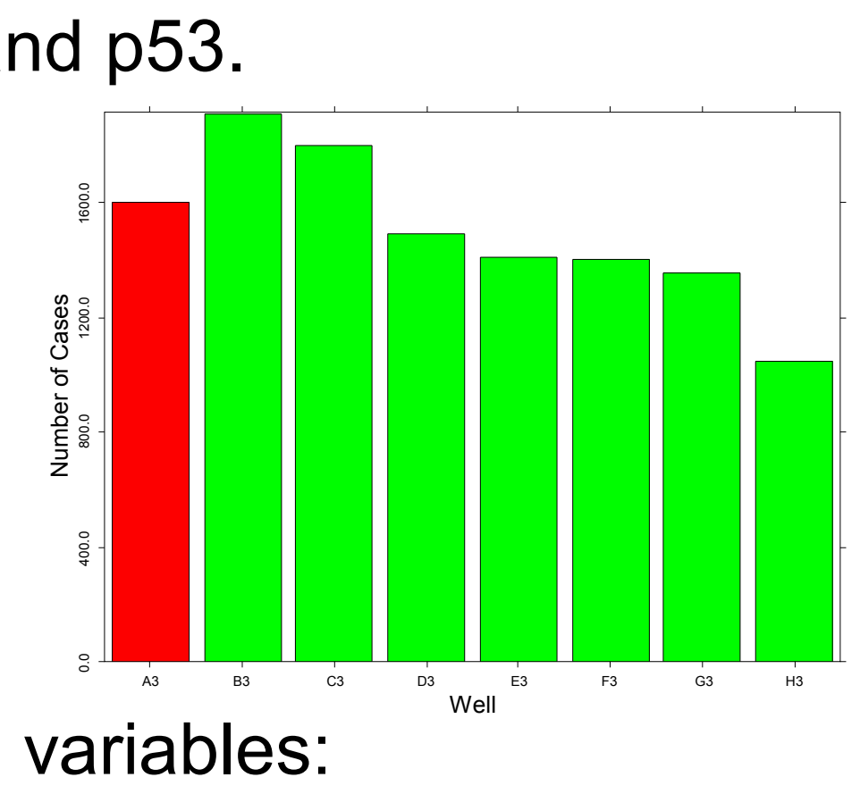
- If 2 samples are different, a classification tree can separate them out with a misclassification rate much lower than that of majority vote.
  - Example: Paint one of the above 2 scatterplots red and pool all the data together.
    - 10283 points each color.
    - Misclassification rate of majority vote: 0.5
    - Misclassification rate of the classification tree grown with X and Y: 0.34.
    - Red points are from an HCS experiment; green points are generated from red points by shuffling the Y values in a certain way. The tree growing algorithm successfully uncovers this pattern and identifies Y to be more important than X.



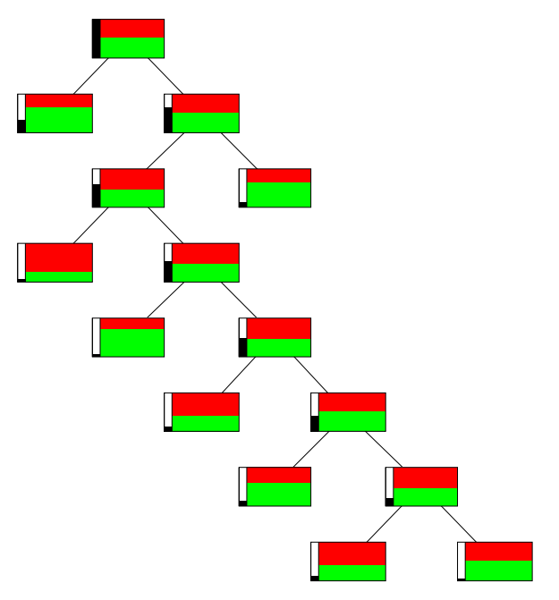


Etoposide Dose Response of U-2 OS Cells

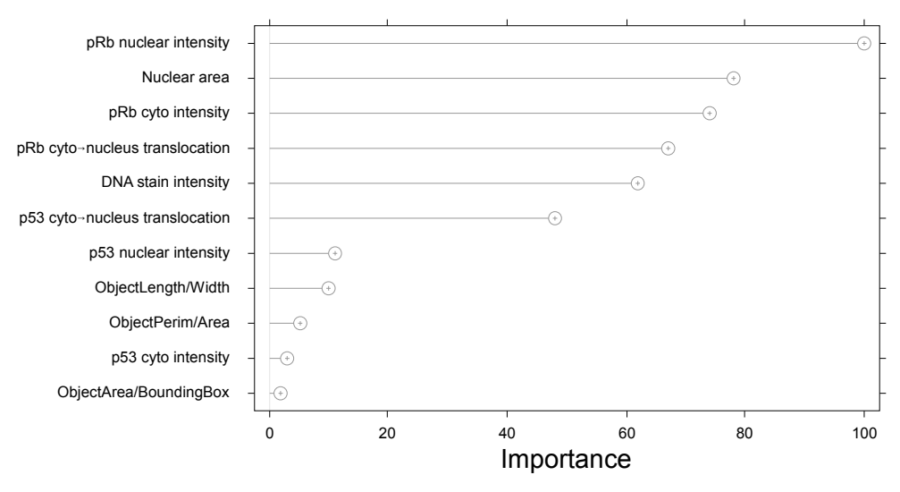
- Comparing the effects of etoposide on U-2 OS cells.
- Cellular targets monitored: DNA, pRb, and p53.
- No etoposide in well A3. Concentrations of etoposide increase with a common ratio of 3 from well B3 to well H3.
- To test for any concentration effect, 7 classification trees are grown to compare the "red" well (A3) with each of the 7 "green" wells.

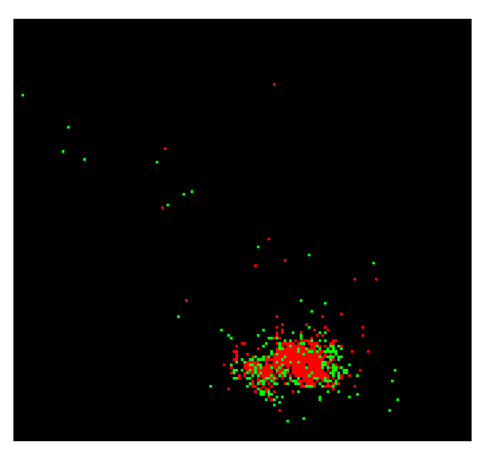


- Each classification tree is grown with 11 variables:
  - DNA stain intensity, nuclear area
  - 3 variables characterizing nucleus shape
  - pRb & p53: cytoplasmic intensity, nuclear intensity, and cytoplasma-to-nucleus translocation.
- For each of the 7 classification trees, an MST planing is done to visualize the joint distribution of the 11 variables and a *multivariate* Kolmogorov-Smirnov test is done as a reference.
- Result:
  - A3 vs. B3:

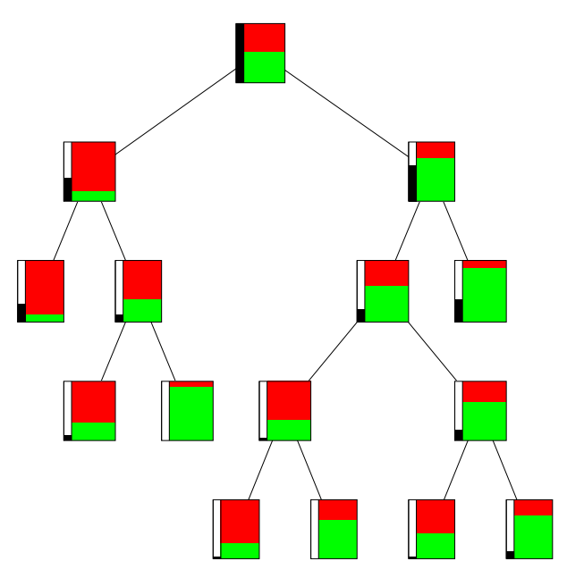


Misclassification rate: 0.400 (s.d. 0.008)
Misclassification rate of majority vote: 0.456
Max absolute deviation of multivariate KS test: 0.045
P-value of multivariate KS test: 0.06

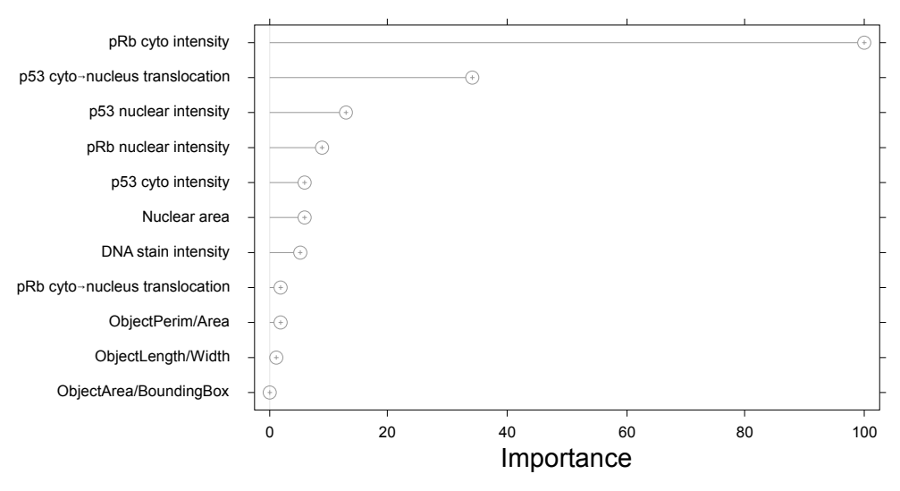


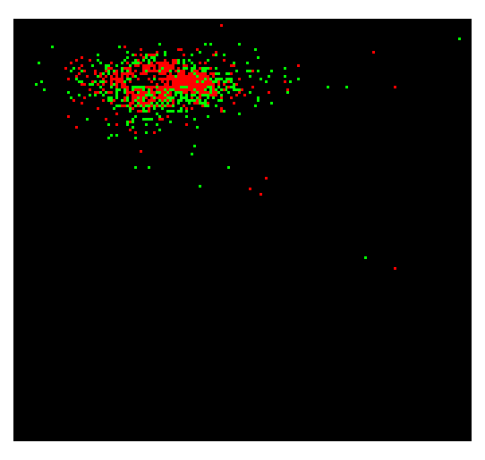


  - A3 vs. C3:

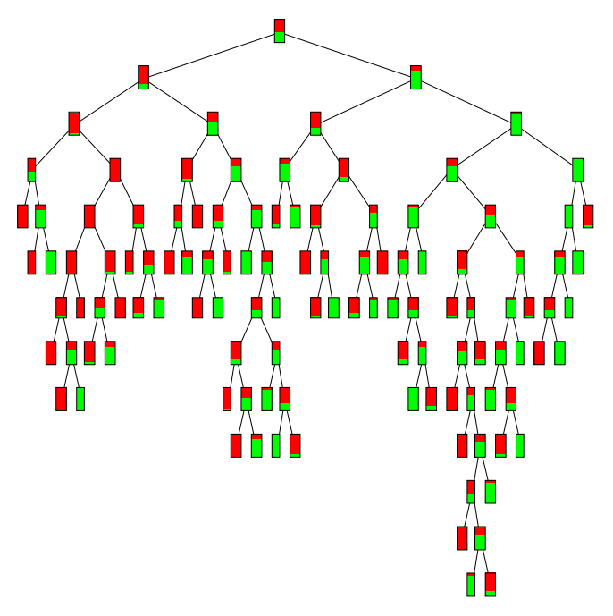


Misclassification rate: 0.192 (s.d. 0.007)
Misclassification rate of majority vote: 0.471
Max absolute deviation of multivariate KS test: 0.119
P-value of multivariate KS test: 0.0

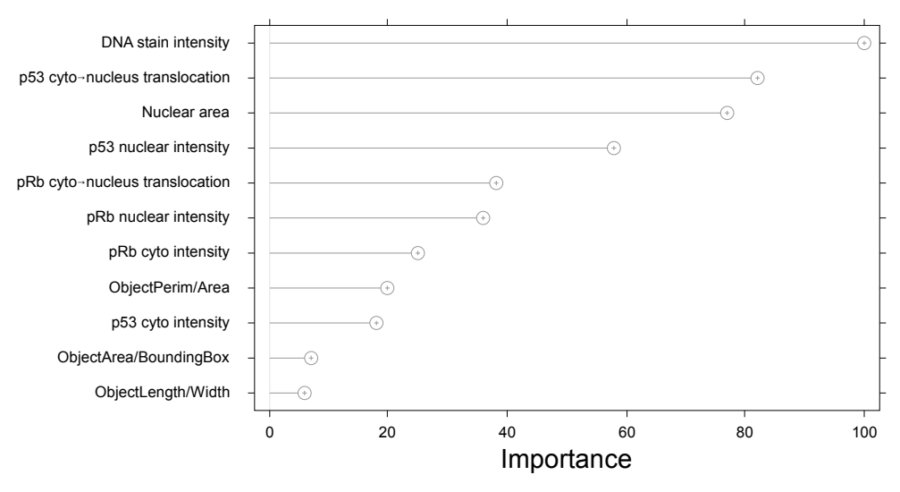


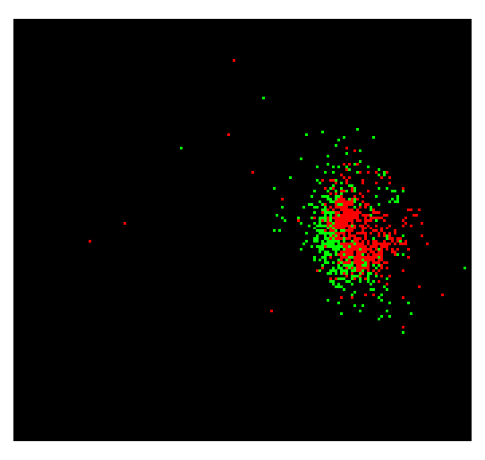


  - A3 vs. D3:

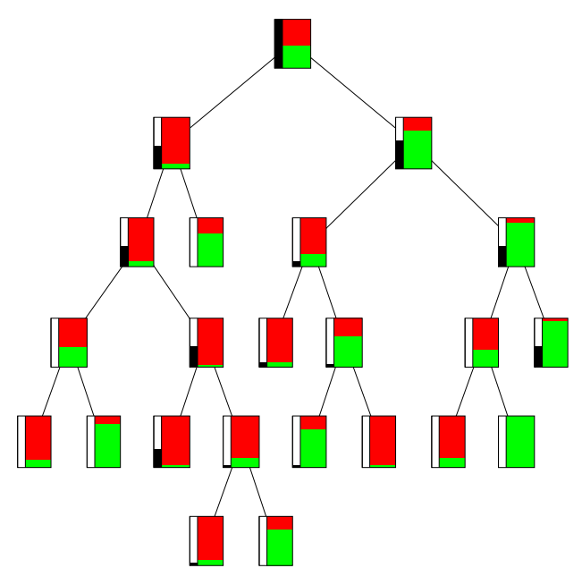


Misclassification rate: 0.091 (s.d. 0.005)
Misclassification rate of majority vote: 0.483
Max absolute deviation of multivariate KS test: 0.439
P-value of multivariate KS test: 0.0

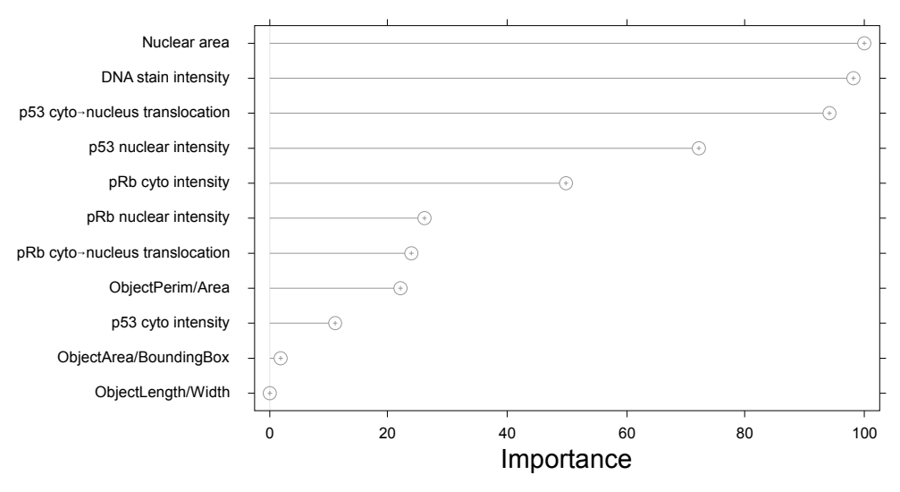


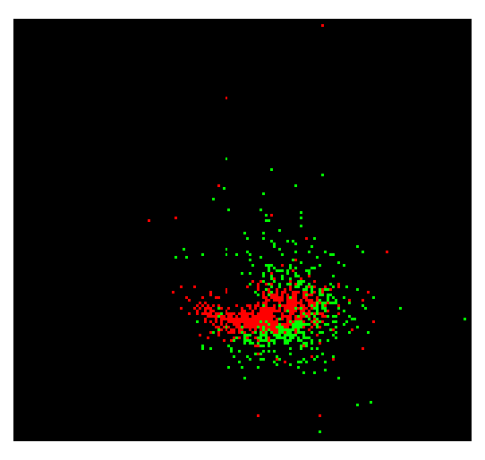


  - A3 vs. E3:

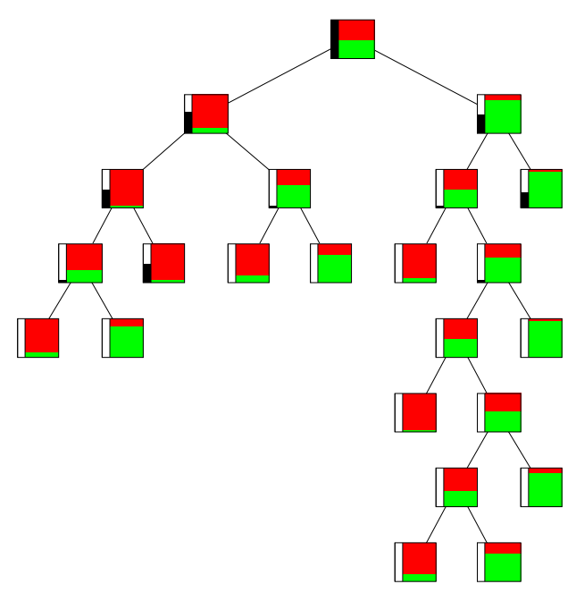


Misclassification rate: 0.091 (s.d. 0.005)
Misclassification rate of majority vote: 0.469
Max absolute deviation of multivariate KS test: 0.646
P-value of multivariate KS test: 0.0

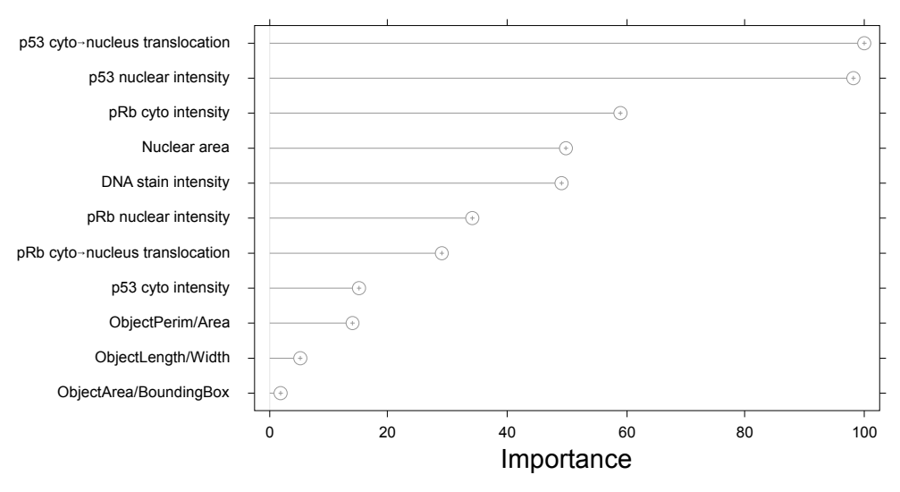


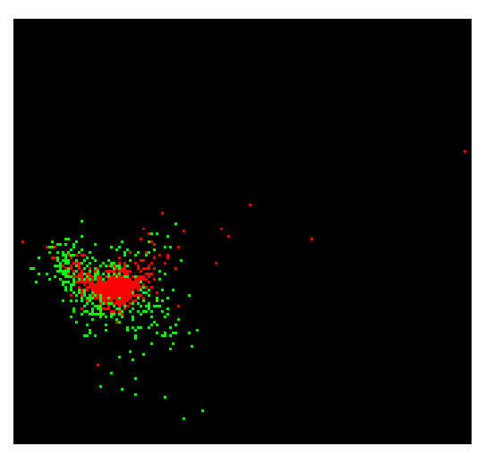


  - A3 vs. F3:

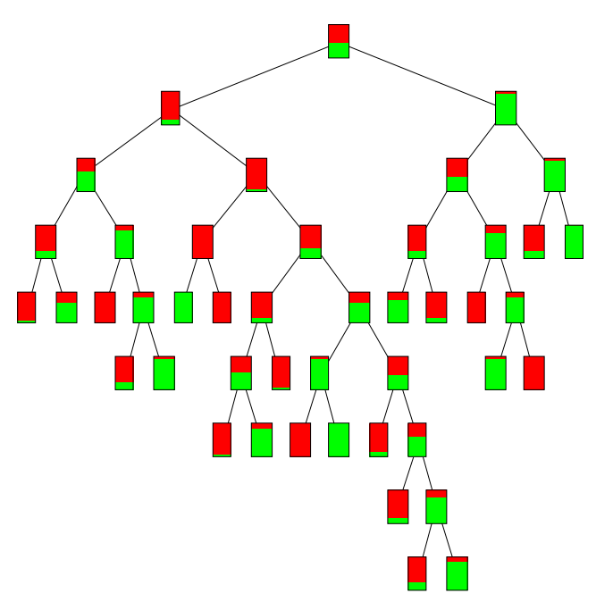


Misclassification rate: 0.086 (s.d. 0.005)
Misclassification rate of majority vote: 0.467
Max absolute deviation of multivariate KS test: 0.786
P-value of multivariate KS test: 0.0

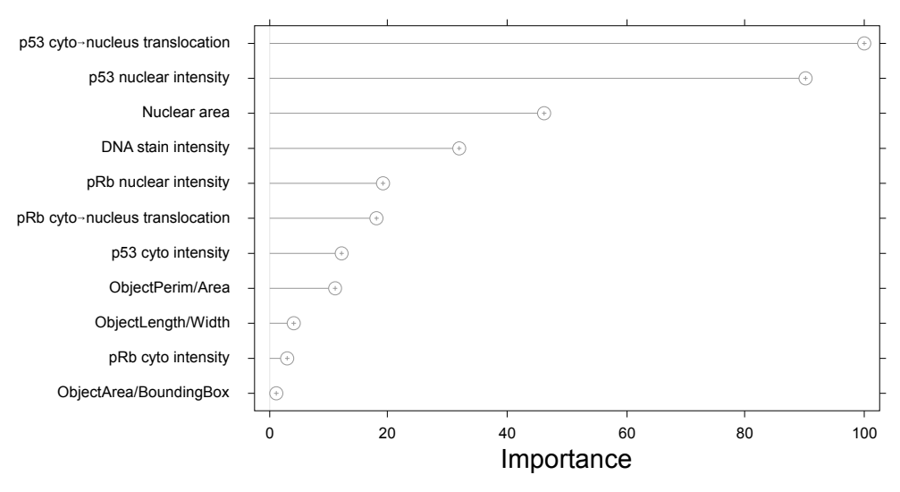


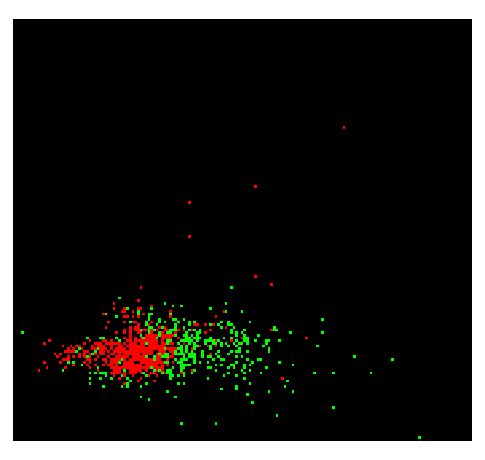


  - A3 vs. G3:

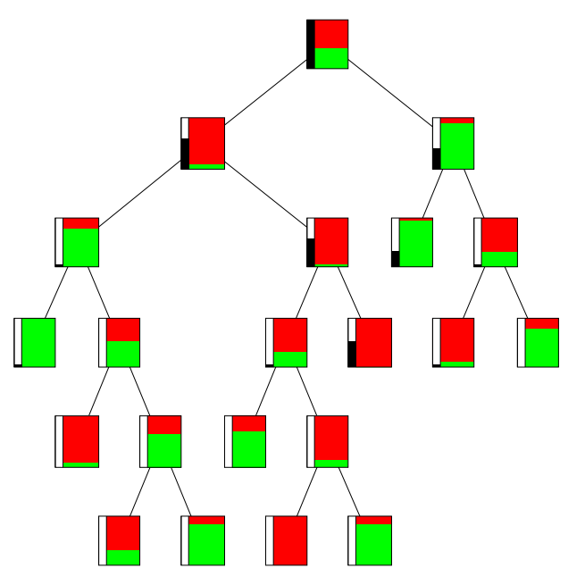


Misclassification rate: 0.071 (s.d. 0.005)
Misclassification rate of majority vote: 0.459
Max absolute deviation of multivariate KS test: 0.785
P-value of multivariate KS test: 0.0

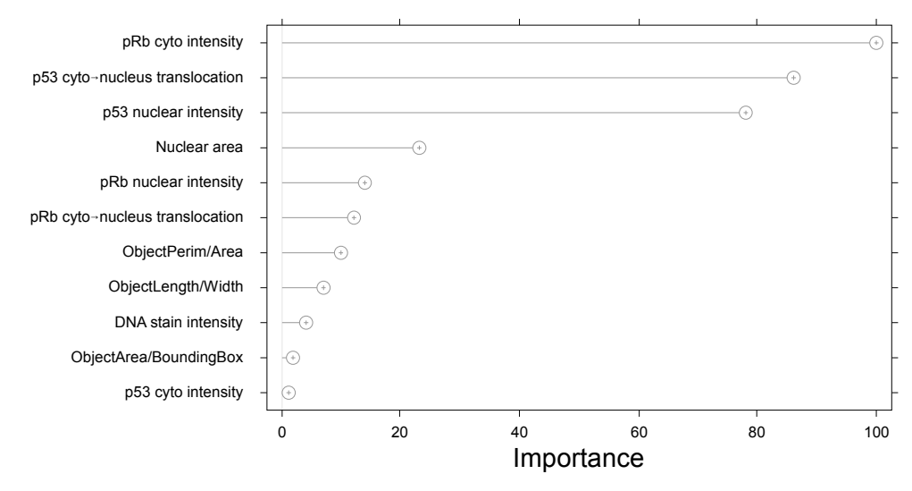


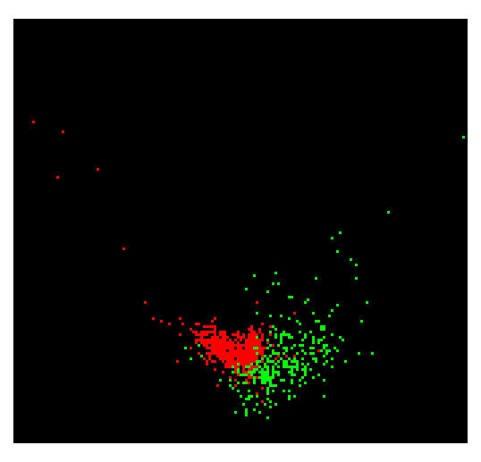


  - A3 vs. H3:

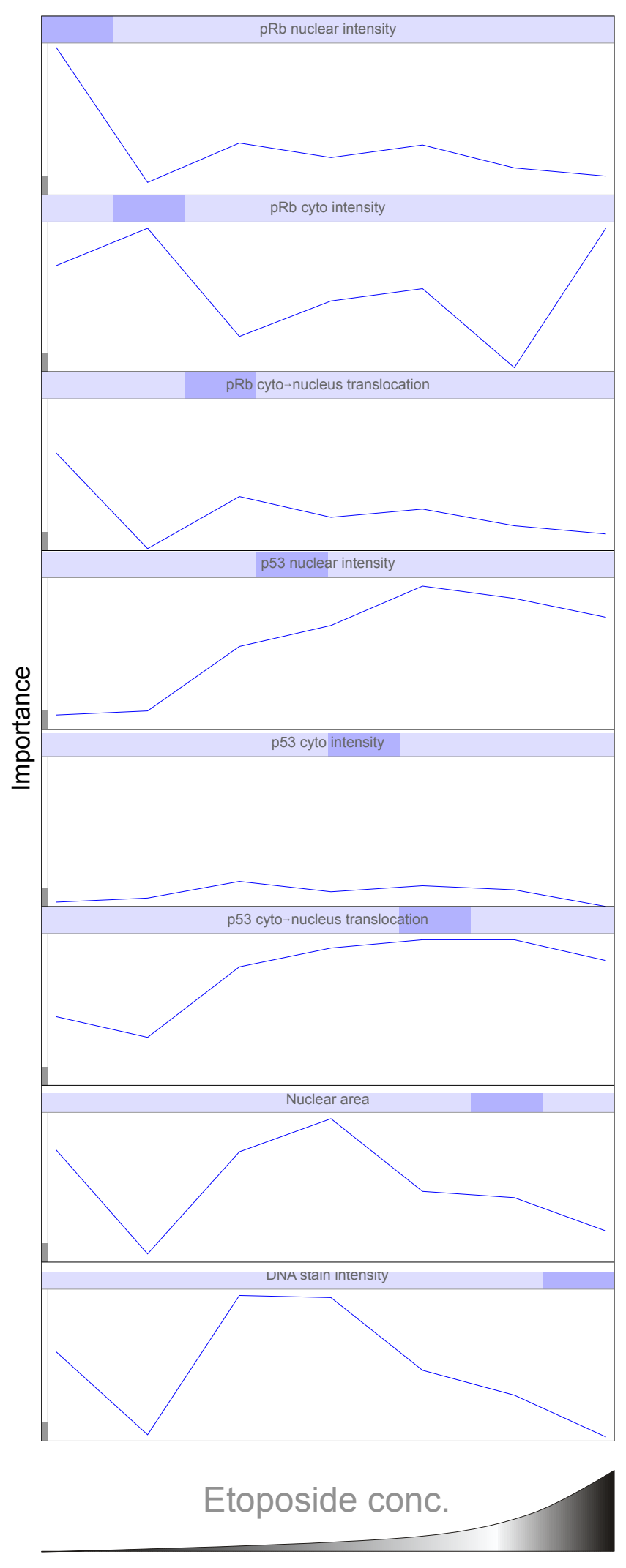


Misclassification rate: 0.049 (s.d. 0.004)
Misclassification rate of majority vote: 0.396
Max absolute deviation of multivariate KS test: 0.758
P-value of multivariate KS test: 0.06

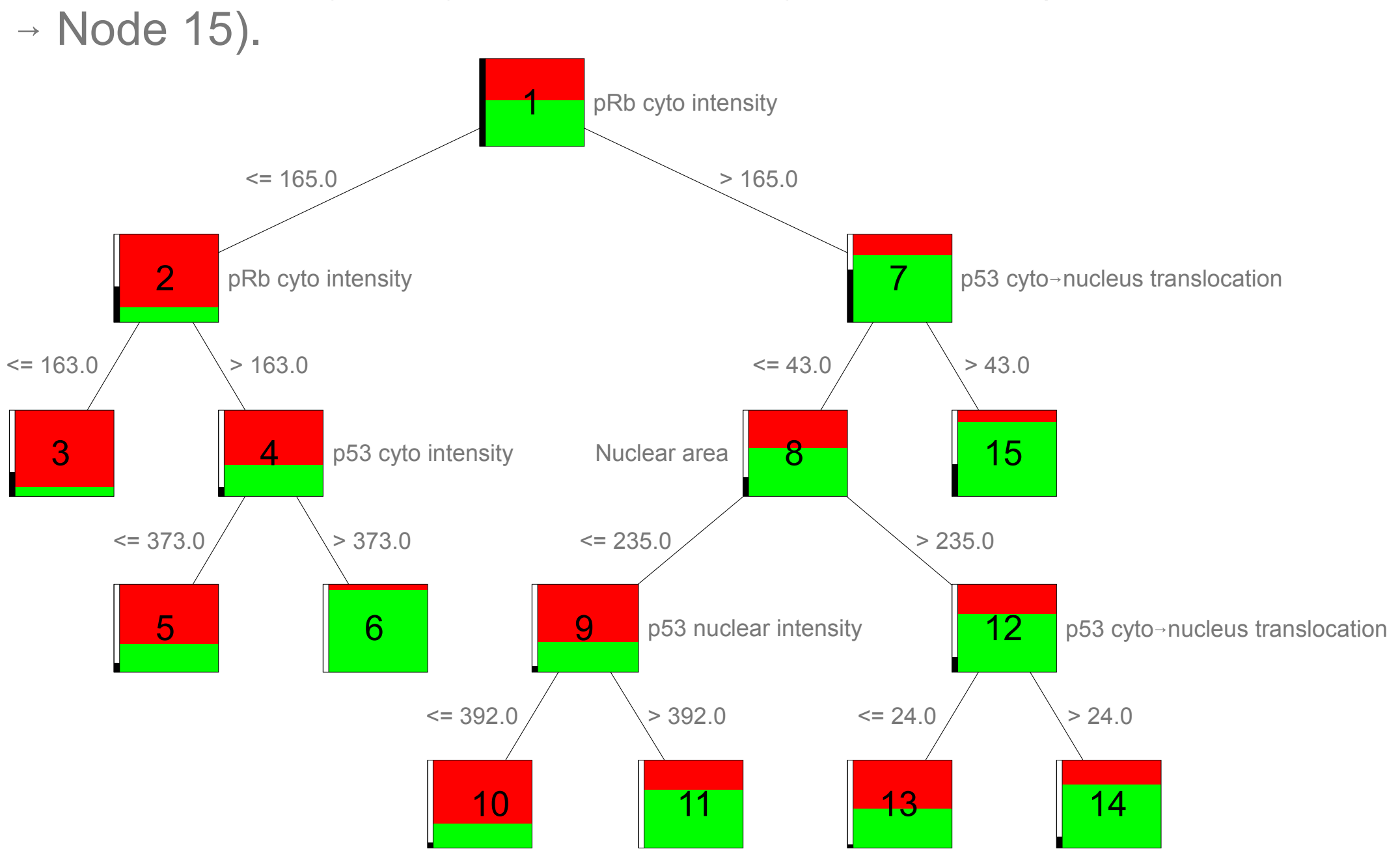




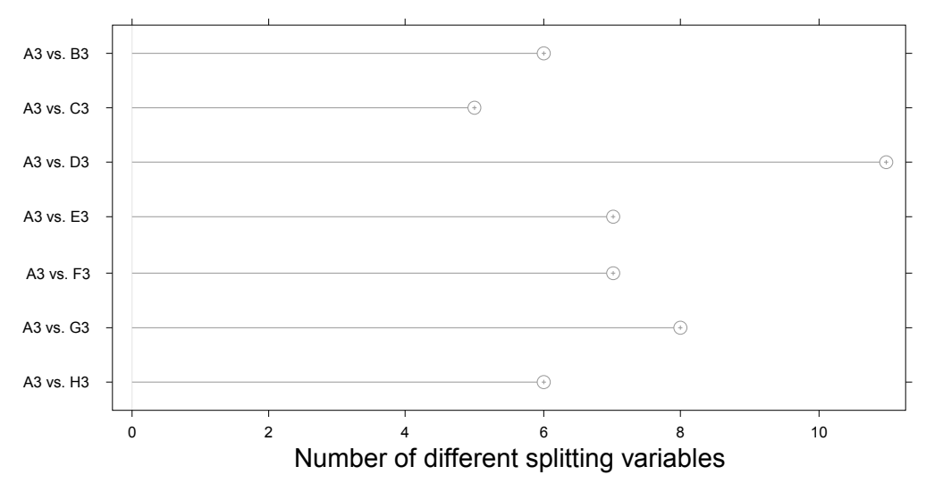
- Let R denote the misclassification rate of a classification tree and R_{mv} the misclassification rate of majority vote. The A3-vs.-B3 comparison exhibits the smallest R_{mv} - R: 0.056, which is 7 times the standard deviation of the R for the A3-vs.-B3 classification tree. This alone should convince us that these 2 samples are different (that is, etoposide affects cells at this lowest concentrations level). Bootstrapping shows these 2 samples are different with a p-value less than 0.002.
- Nuclear intensity and cyto→nucleus translocation of pRb are more important than those of p53 at lower etoposide concentrations; however, the reverse is true at higher etopside concentrations.
- pRb cyto intensity is uniformly more important than p53 cyto intensity at all etoposide concentrations. pRb cyto intensity is the most important variable twice among the 7 classification trees.
- The 3 variables characterizing nucleus shape are always among the 4 least important variables except for the A3-vs.-H3 comparison, where they are among the 5 least important variables.



- Due to space limitation and the static nature of a poster, only minimal information is displayed in each of the 7 classification trees. With the aid of dynamic graphics on a computer screen, much information is just a few mouse clicks away. For example, we can enlarge the A3-vs.-C3 tree to reveal the splitting variable and the splitting value at each node. These additional pieces of information allow us to understand the conditions that determine when a cell is in one class rather than another. For example,
  - If pRb cyto intensity is <= 163.0, a cell is very likely to be untreated by etoposide (Node 1 → Node 2 → Node 3).
  - If 163.0 < pRb cyto intensity <= 165.0 and p53 cyto intensity > 373.0, a cell is very likely to be treated by etoposide (Node 1 → Node 2 → Node 4 → Node 6).
  - If pRb cyto intensity > 165.0 and p53 cyto→nucleus translocation > 43.0, a cell is very likely to be treated by etoposide (Node 1 → Node 7 → Node 15).



- Not all variables supplied to the tree growing algorithm are chosen as splitting variables; only important ones are chosen.



Summary

- HCS data are inherently multivariate.
- Analyzing multivariate data using methods univariate in nature (histograms, the KS test) runs the risk of missing important content of high-content screening data sets.
- Nonparametric methods are required to properly decipher HCS data sets.
- A classification tree is a versatile tool:
  - It can do multivariate two-sample comparison. For screening, it provides objective ways (R, R_{mv} / R, (R_{mv} - R) / s.d. of R or p-value) to compare 2 HCS samples; no more need to squint at a bunch of heat maps.
  - It gives us a clear idea of which variables are important.
  - It enables us to understand what variables or interactions of variables drive a cellar phenomenon.
- All data analysis and plots in this poster were done with Panmo, a dynamic graphics system for exploring HCS data.