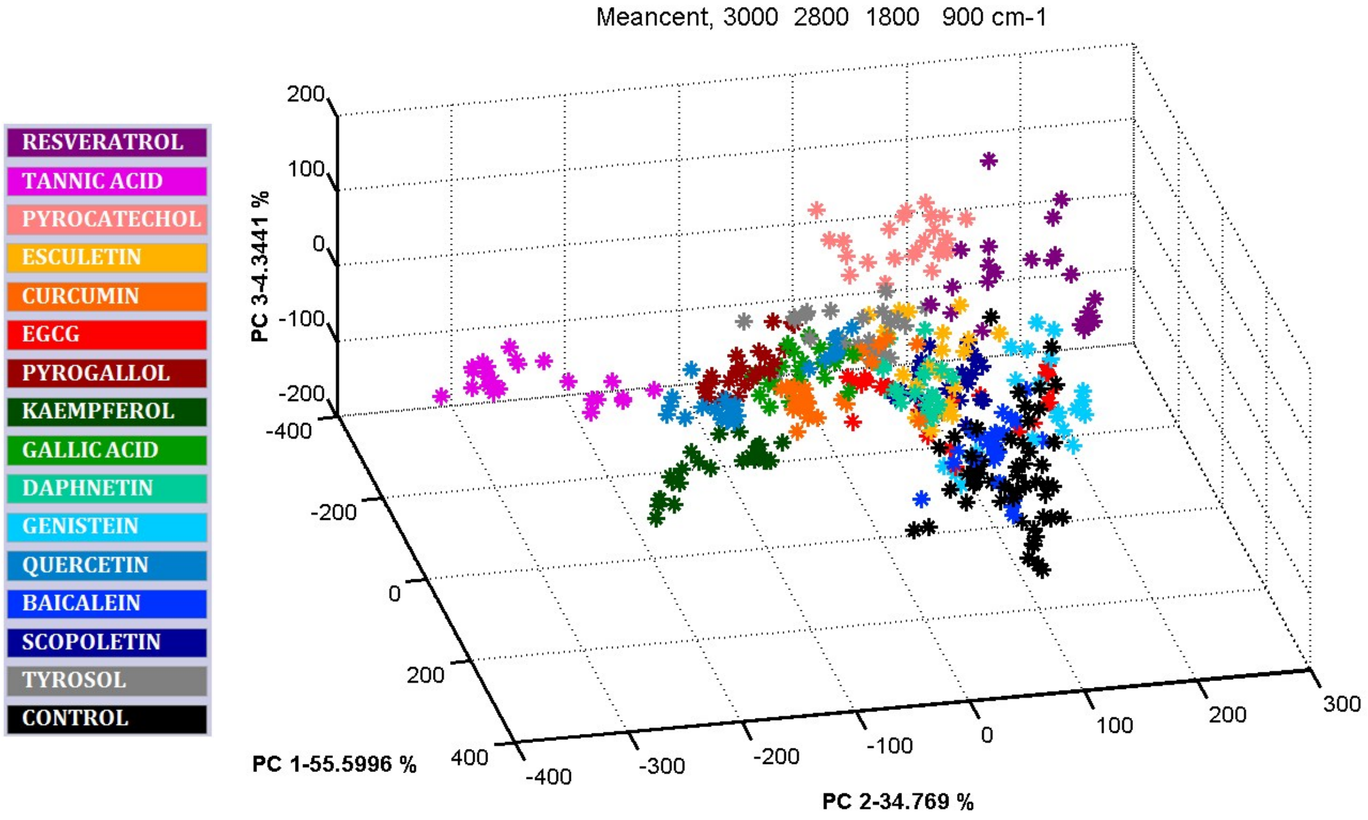
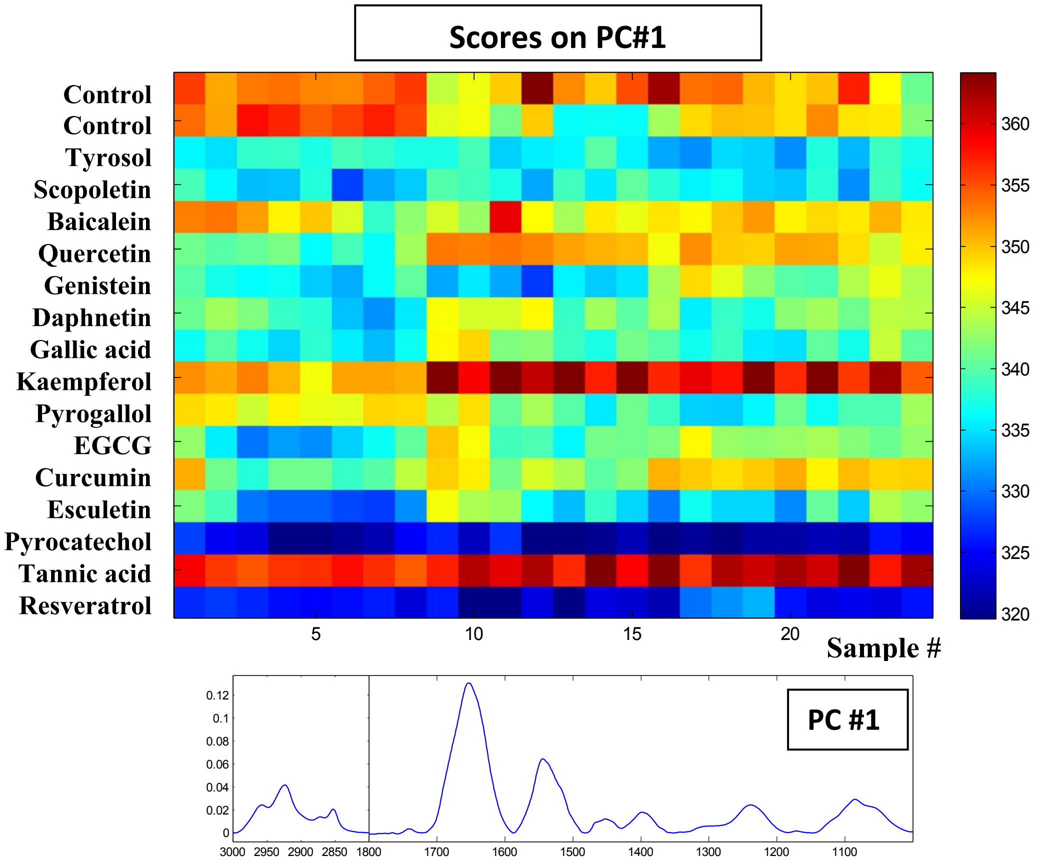
Supplementary data

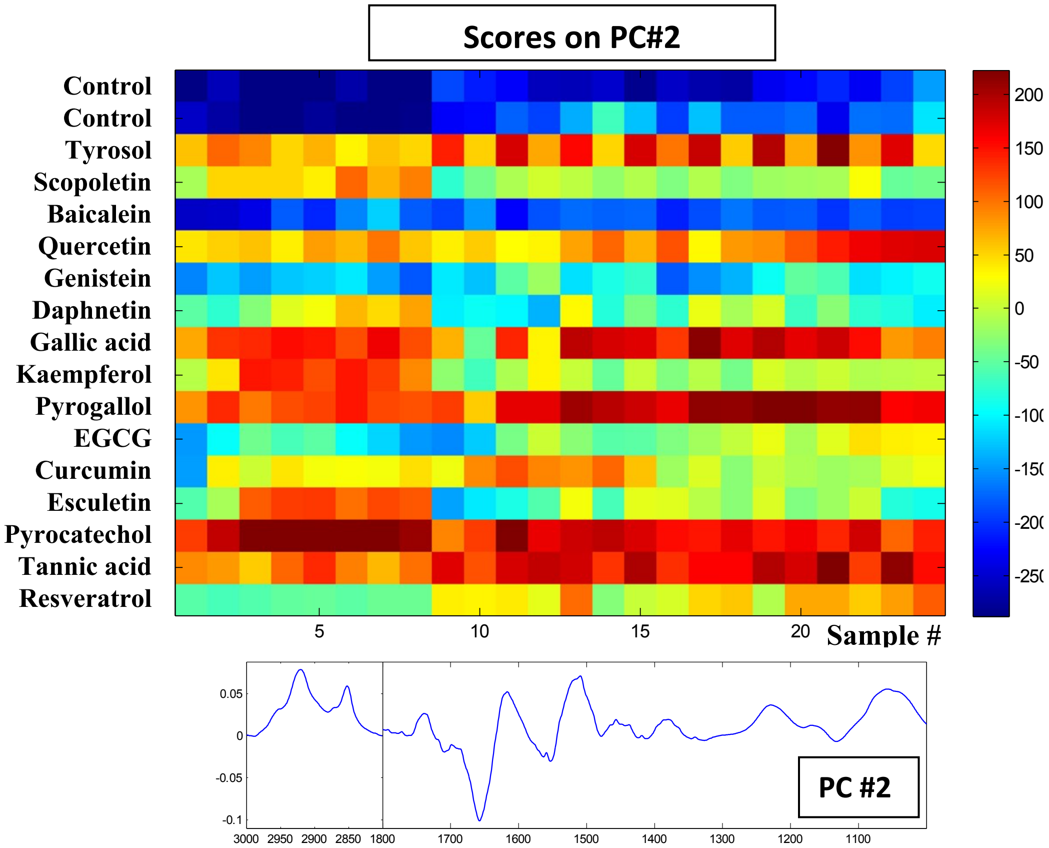
Score plot obtained after principal component analysis of 24 spectra of MDA-MD-231 cells exposed to each polyphenol PCA was computed between 3000-2800 and 1800-900 cm^{-1}.

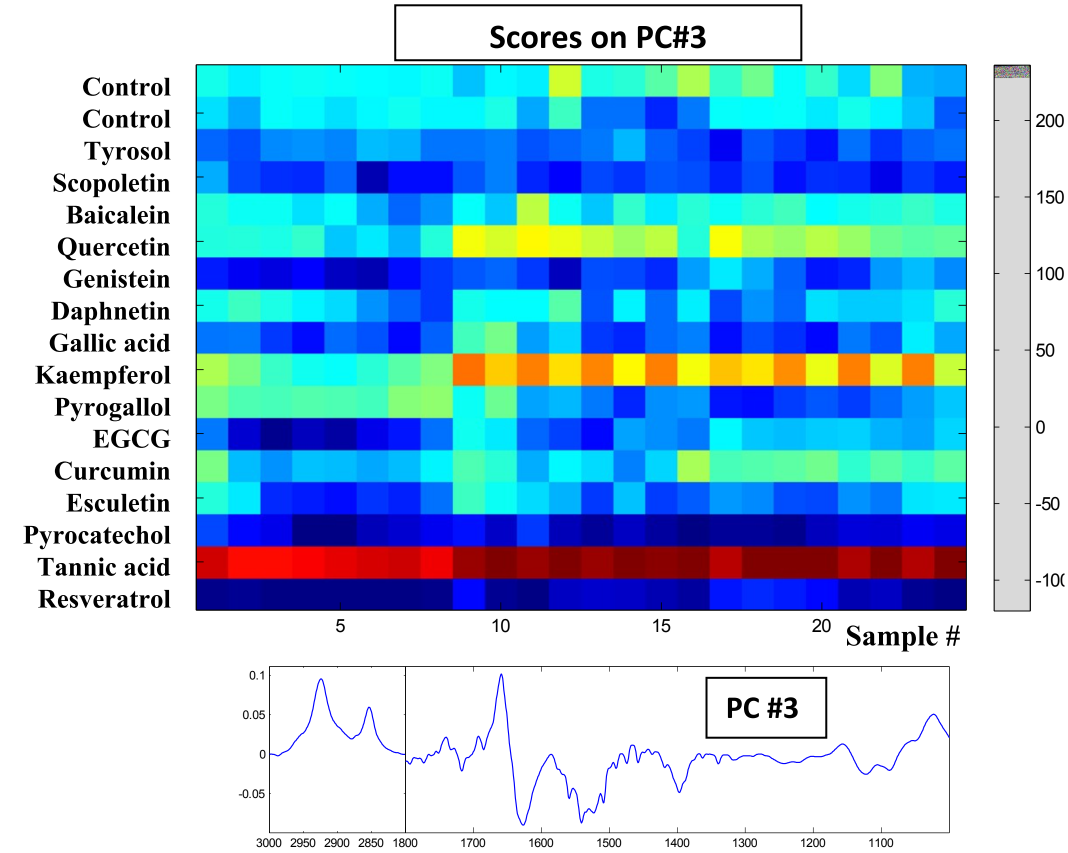


<u>Figure S1</u>: Principal Component Analysis (PCA) of 416 pre-processed and baseline corrected spectra (see Material and Method section) from the breast cancer cell line MDA-MB-231 exposed to 15 polyphenols at their respective IC$_{50}$ during 24h. *24 spectra were obtained for each polyphenol treatment and 56 spectra of non- treated cells (control 0.1% EtOH)*

It can be observed that there is some grouping of the cell spectra according to treatment but, in the PC1, PC2 and PC3 space, separation is not complete. As 16 conditions have been tested, the number of independent variations can be larger than 3 and the 16 conditions may require more than 3 PCs to be correctly described. In order to observe a larger dimensional space, Figure S2 presents score plots on 4 PCs (PC1, PC2, PC3 and PC 6, selected as an example) and shed some light on how discrimination among polyphenol effect occurs. In Figure S2, the scores are color coded. Each square in a line represents one spectrum of MDA-MB-231 cells treated with one polyphenol. The different replicates described in Methods result in 24 spectra for each polyphenol treatment. All spectra corresponding to 1 experimental condition are present in one line. For instance, pyrocatechol and resveratrol have unique but similar score values on PC1. Yet, the scores for pyrocatechol and resveratrol are significantly different on PC2, resulting in a unique identification. Similarly, using more PCs, it is possible to find a unique pattern of scores that identify uniquely each polyphenol effect. This is the purpose of the decision tree presented in Figure S3.


Scores on PC#1
Control
Control
Tyrosol
Scopoletin
Baicalein
Quercetin
Genistein
Daphnetin
Gallic acid
Kaempferol
Pyrogallol
EGCG
Curcumin
Esculetin
Pyrocatechol
Tannic acid
Resveratrol
5
10
15
20
Sample #
360
355
350
345
340
335
330
325
320
PC #1
0.12
0.1
0.08
0.06
0.04
0.02
0
3000
2950
2900
2850
1800
1700
1600
1500
1400
1300
1200
1100


Scores on PC#2
Control
Control
Tyrosol
Scopoletin
Baicalein
Quercetin
Genistein
Daphnetin
Gallic acid
Kaempferol
Pyrogallol
EGCG
Curcumin
Esculetin
Pyrocatechol
Tannic acid
Resveratrol
5
10
15
20
Sample #
200
150
100
50
0
-50
-10(
-15(
-20(
-25(
PC #2
0.05
0
-0.05
-0.1
3000
2950
2900
2850
1800
1700
1600
1500
1400
1300
1200
1100


Scores on PC#3
Control
Control
Tyrosol
Scopoletin
Baicalein
Quercetin
Genistein
Daphnetin
Gallic acid
Kaempferol
Pyrogallol
EGCG
Curcumin
Esculetin
Pyrocatechol
Tannic acid
Resveratrol
5
10
15
20
Sample #
200
150
100
50
0
-50
-10(
PC #3
0.1
0.05
0
-0.05
3000
2950
2900
2850
1800
1700
1600
1500
1400
1300
1200
1100

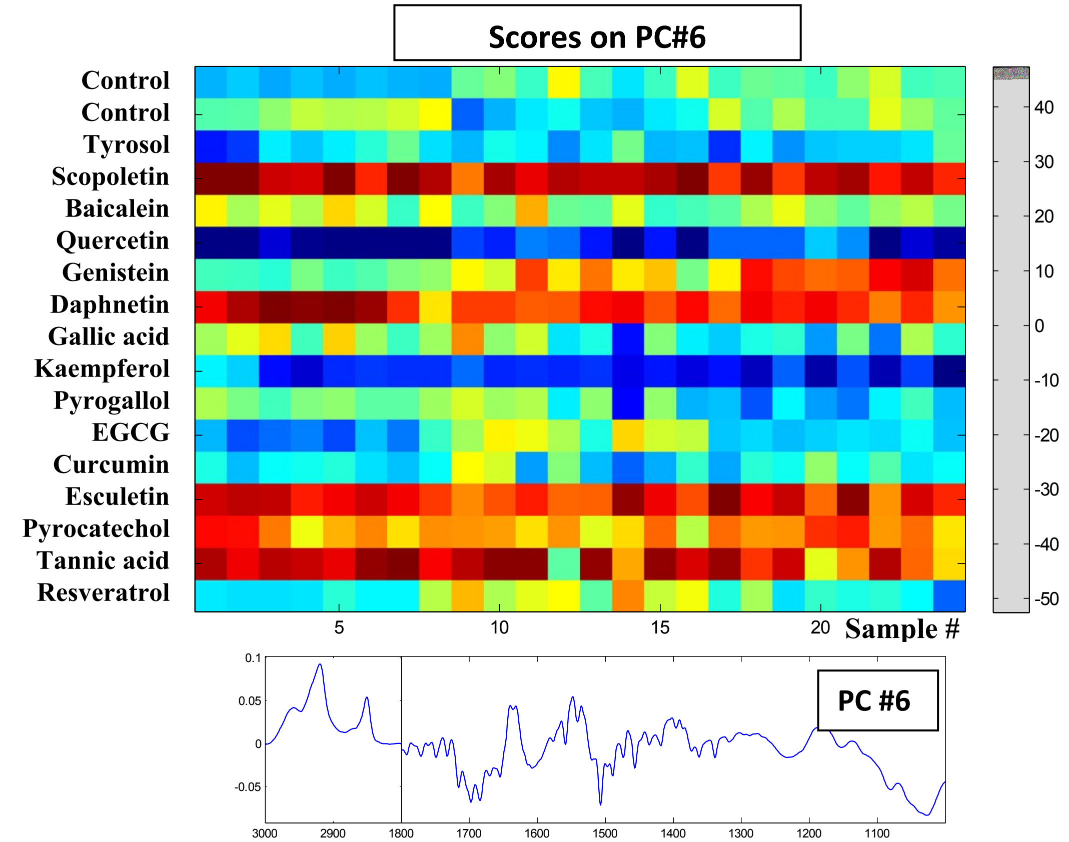

Scores on PC#6
Control
Control
Tyrosol
Scopoletin
Baicalein
Quercetin
Genistein
Daphnetin
Gallic acid
Kaempferol
Pyrogallol
EGCG
Curcumin
Esculetin
Pyrocatechol
Tannic acid
Resveratrol
5
10
15
20
Sample #
40
30
20
10
0
-10
-20
-30
-40
-50
PC #6
0.1
0.05
0
-0.05
3000
2900
1800
1700
1600
1500
1400
1300
1200
1100

***Figure S2*: Color-coded scores.** *Scores on 4 PCs for MDA-MD-231 cells exposed to 16 experimental conditions indicated in the left margin. Every square on the image indicates the value of the score which is color coded as indicated by the color bar. The shape of the corresponding principal component is provided below. The presence of some water vapor contribution can be noted in PC3 and PC6.*

The decision tree describes score threshold that separate polyphenols or polyphenol groups. For instance, at the first node, when PC1<175.47, only control and baicalein treated cells are found. All other conditions are characterized by scores on PC1>175.47. At the second node, on the left, all cells exposed to tannic acid have a score on PC1>175.47 and a score on PC2>226.5. More scores on other PCs must be used to discriminate the other treatments. The leaves of the tree represent one treatment. Some level of error is yet still present. There are 19 leaves while only 16 conditions exist because the controls (ctr) are found in triplicate and tyrosol (tyr) is duplicated, suggesting some level of overfitting. A better evaluation of the prediction error after applying the rules of the decision tree can be obtained by cross-validation. For cross-validation, the rules are applied on a subset of the data selected randomly and applied on another subset. The evolution of the error with respect to the number of terminal nodes is reported in Figure S4

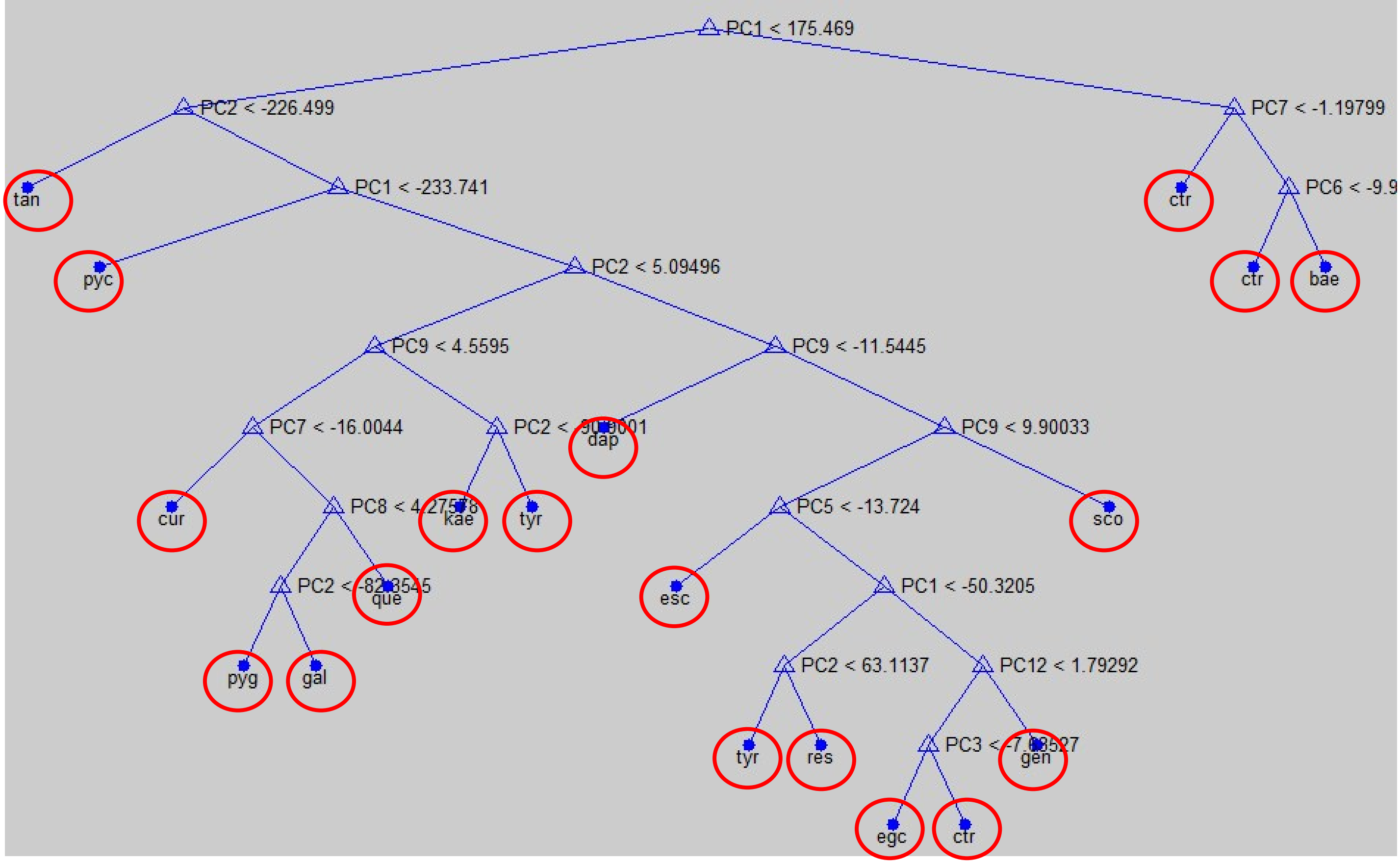


***Figure S3* Decision tree.** *24 spectra were obtained for each experimental conditions, 12 PCs were used in this analysis. At each node, the score of the indicated PC indicates how separation is obtained. Polyphenols are identified by the first 3 letters of their name shown in Figure S1.*

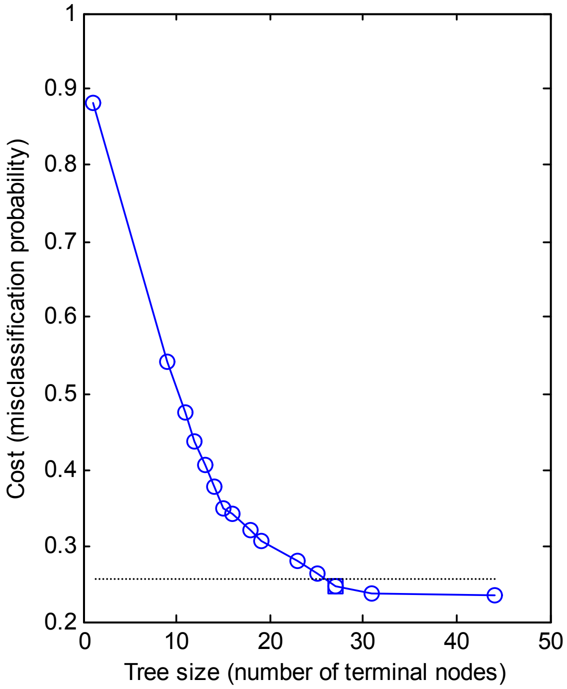


Figure S4 ***Prediction error as a function of tree size.*** *Cross-validation was used to evaluate the prediction error on the rules defined by the decision tree presented in Figure S3.*

In Figure S4, the cost of a node is the sum of the misclassification observed in that node, the global cost presented is the sum obtained on all terminal nodes in a 10-fold cross-validation. The best tree size which produces the smallest tree within one standard error of the minimum-cost subtree is indicated by the dotted line. It must be stressed that the error may be underestimated because a completely independent test set was not used. Yet, the tree presented in Figure S3 visually indicates how scores on more than 3 PCs (as presented in Figure S1) can be useful to separate the response of MDA-MD-231 cancer cells to 16 experimental conditions studied in this paper. Figure S4 shows the limitations of this approach as a supervised method to discriminate treatments.