

# Feature Selection (FS) workflow report

*May 18, 2017*

## Introduction

The report summarizing the Feature Selection pipeline results.

## Feature Selection workflow

Univariate canonical correlation (X2) with Recursive Feature Elimination (RFE) wrapped with Random Forest (RF).

## Dataset

Triple-Negative Breast Cancer (TNBC) proteome. Label-free deep proteome analysis of 44 (samples and technical replicates) human breast specimens.

## Summary stats from training phase

Table 1: Best model metrics from 10-folds cross-validation resampling.

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.575	0.4511	0.3104	0.3857
2	0.5167	0.3934	0.2908	0.3356
3	0.7667	0.6907	0.235	0.3094
4	0.6917	0.5311	0.2637	0.42
5	0.7417	0.667	0.2952	0.3793
6	0.6917	0.567	0.2888	0.4122
7	0.8667	0.8289	0.2194	0.2817
8	0.7833	0.6861	0.2838	0.418
9	0.8417	0.7597	0.1819	0.3203
10	0.8917	0.8667	0.1845	0.222
15	0.9	0.8762	0.1748	0.2096
20	0.925	0.9095	0.1687	0.1988
25	0.8917	0.8597	0.1845	0.2417
30	0.8583	0.8168	0.1927	0.2519
35	0.8583	0.8168	0.1927	0.2519
40	0.8583	0.8168	0.1927	0.2519
45	0.8583	0.8168	0.1927	0.2519
50	0.8833	0.8502	0.1532	0.1968
60	0.8833	0.8476	0.1532	0.1993
70	0.9083	0.8835	0.1493	0.1904
80	0.8833	0.8476	0.1532	0.1993
90	0.9167	0.893	0.1361	0.1748
100	0.9167	0.8905	0.1361	0.1782
1944	0.8833	0.8609	0.2194	0.2527

## Summary stats from testing phase

Table 2: Classification metrics from twenty class-balanced and randomized runs.

run	Variables	Accuracy	Kappa	AccuracyPValue
1	40	1	1	2.216e-07
2	40	0.8462	0.803	9.42e-05
3	1944	0.9231	0.9008	6.703e-06
4	80	0.9231	0.8992	6.703e-06
5	100	0.9231	0.8992	6.703e-06
6	100	1	1	2.216e-07
7	1944	1	1	2.216e-07
8	70	1	1	2.216e-07
9	80	1	1	2.216e-07
10	50	1	1	2.216e-07
11	35	1	1	2.216e-07
12	1944	1	1	2.216e-07
13	1944	0.9231	0.8992	6.703e-06
14	1944	0.8462	0.7969	9.42e-05
15	30	1	1	2.216e-07
<b>16</b>	<b>20</b>	<b>1</b>	<b>1</b>	<b>2.216e-07</b>
17	15	0.9231	0.9015	6.703e-06
18	90	0.9231	0.9008	6.703e-06
19	25	1	1	2.216e-07
20	35	0.9231	0.9008	6.703e-06

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.9577	0.05279	1

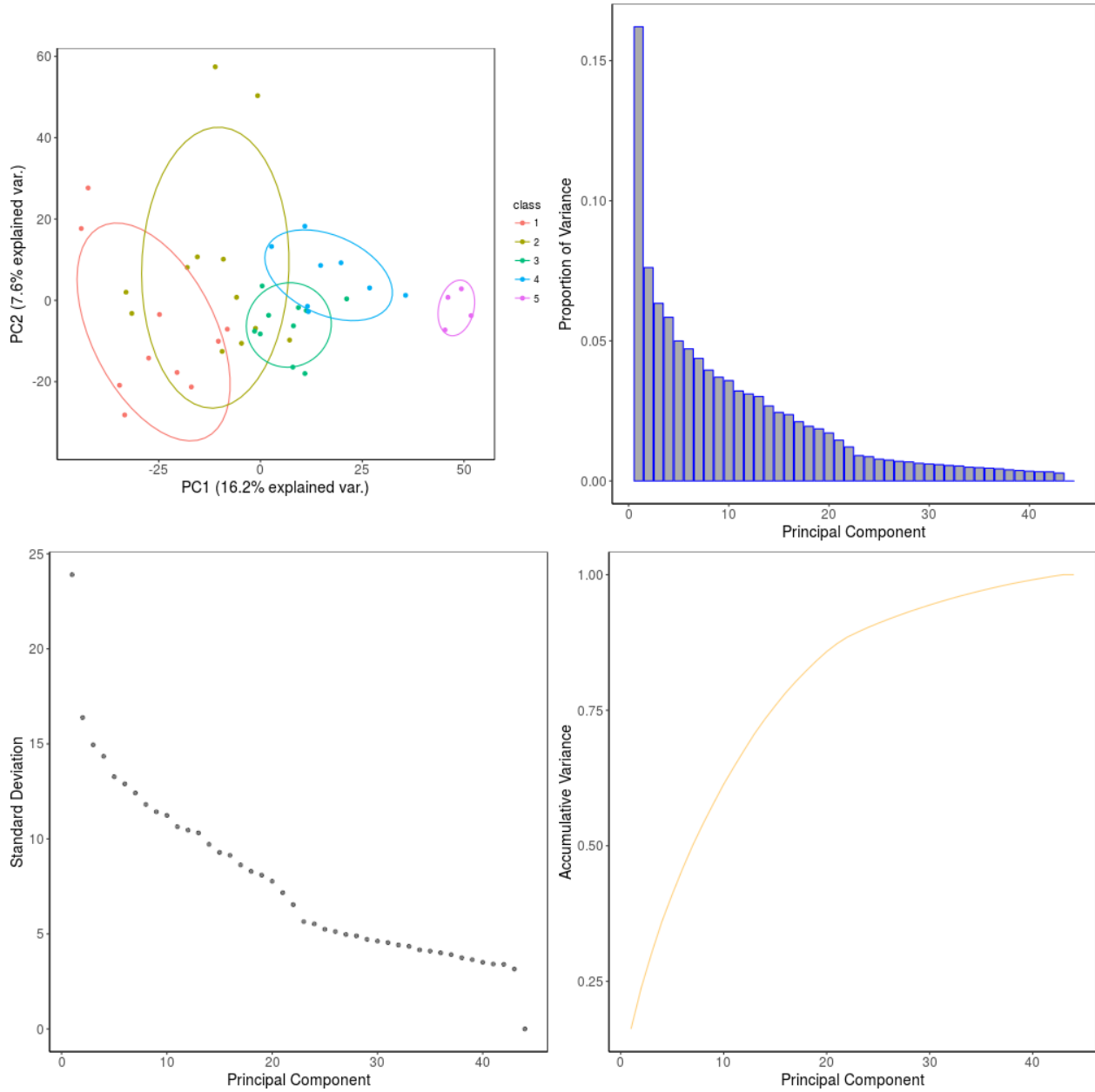
## Workflow runtime

4.949 minutes

## Plots

### Visualization of the classification using PCA

- Groups distribution on the first two Principal Components (PC1 and PC2) from the original data (without apply any FS method).



- Groups distribution on the first two Principal Components (PC1 and PC2) after to apply the FS workflow.

