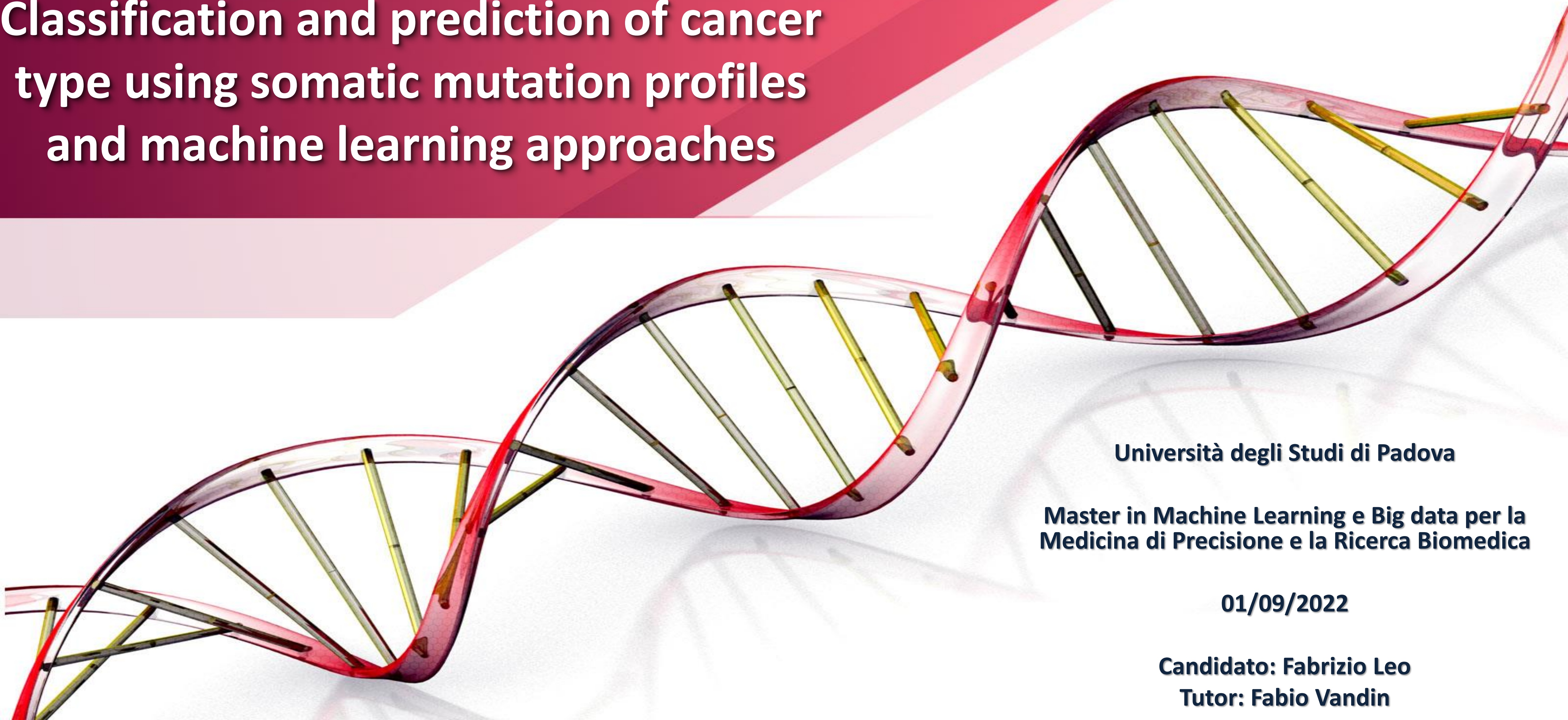
# Classification and prediction of cancer type using somatic mutation profiles and machine learning approaches



Università degli Studi di Padova

Master in Machine Learning e Big data per la Medicina di Precisione e la Ricerca Biomedica

01/09/2022

Candidato: Fabrizio Leo
Tutor: Fabio Vandin

# Data

- Source: The Cancer Genome Atlas (TCGA)

- Information about cancer patient somatic mutations and type of cancer for each patient

  a) «samples_labels.txt» -> list of patients with their cancer type
  b) «snvs.tsv» -> list of mutated genes for each patient
  c) «Compendium_Cancer_Genes.txt» -> list of genes considered important

# Pre-processing

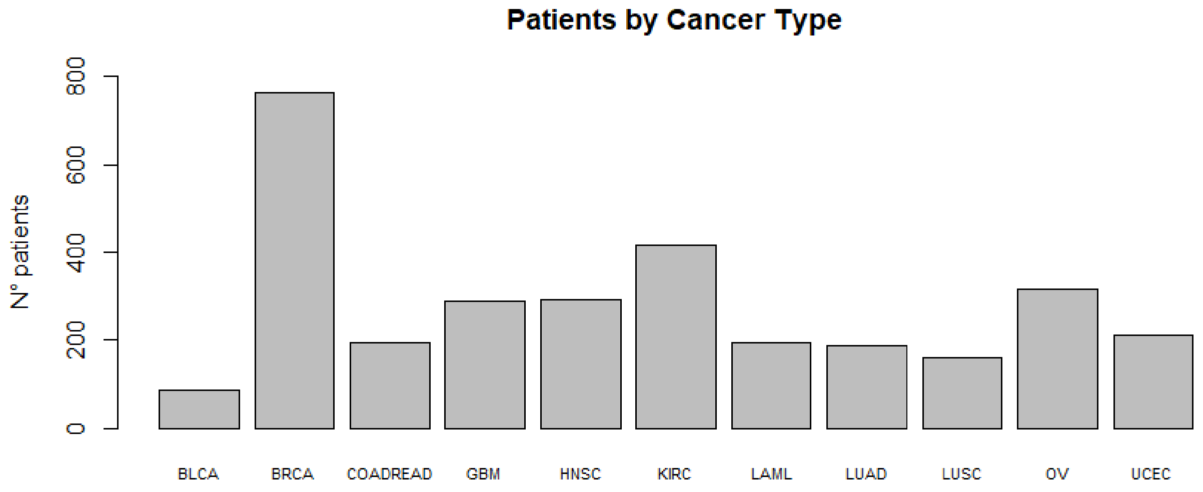
- I only considered, for computational reasons, the list of 568 genes considered as important
- Built binary mutation matrix considering these genes
- Removed 42 genes not mutated in any patient

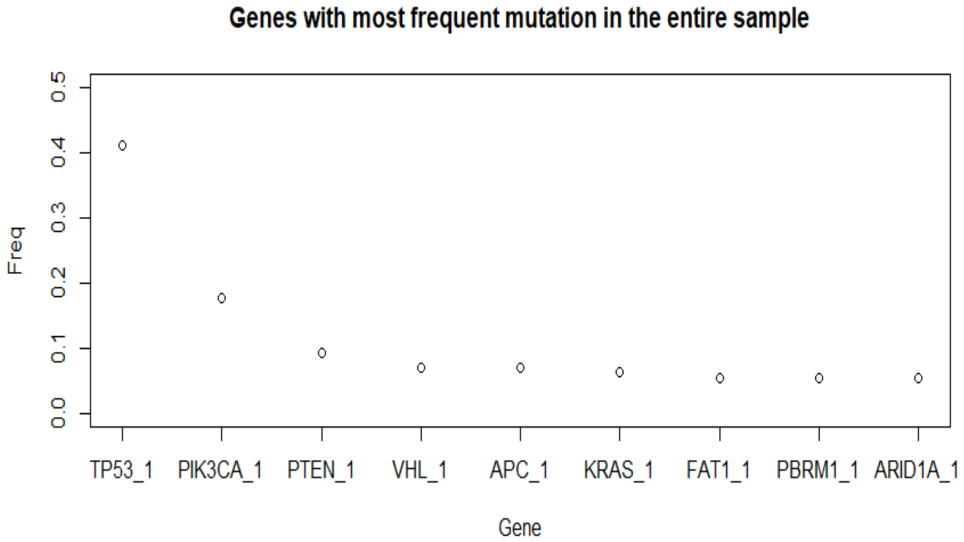
- Therefore, I obtained a mutation matrix 3109 (patients) x 526 (genes)

# Outline

- Is it possible to predict cancer type based on genes with somatic mutation in a patient?
- Is there a «small» set of genes having a good predictive power, or at least as good as the entire set of genes?
- Does the patient grouping based on similarity of mutated genes reflect the grouping based on cancer type?

# Explorative analysis

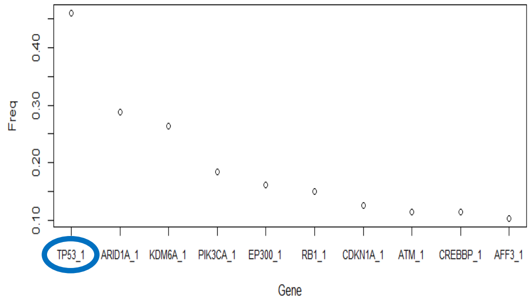




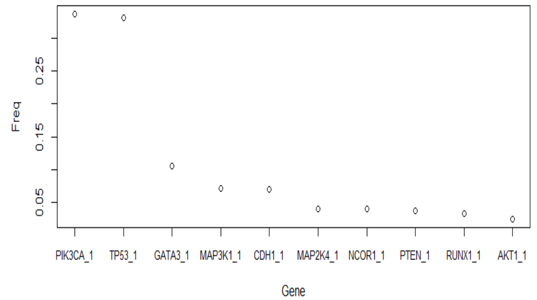
- 763 cases BRCA (24.5%)
- 87 cases BLCA (2.8%)

# Explorative analysis

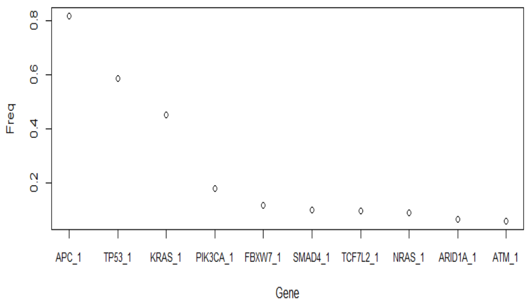
Genes with most frequent mutation in the BLCA Cancer Type



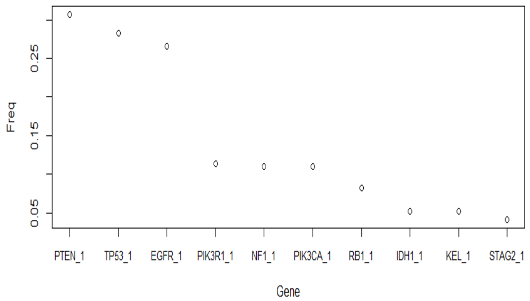
Genes with most frequent mutation in the BRCA Cancer Type



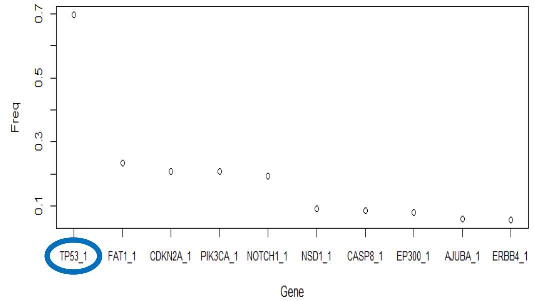
Genes with most frequent mutation in the COADREAD Cancer Type



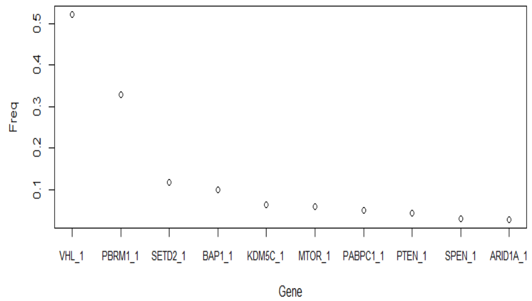
Genes with most frequent mutation in the GBM Cancer Type



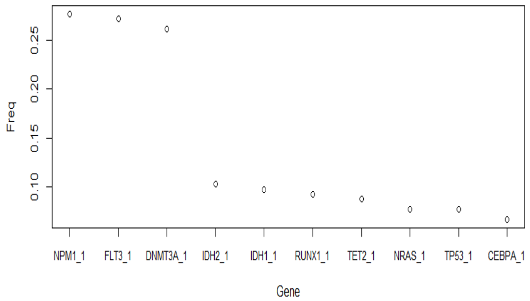
Genes with most frequent mutation in the HNSC Cancer Type



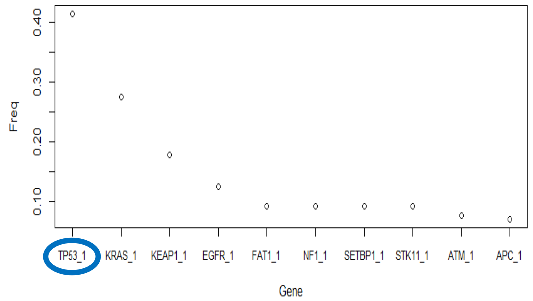
Genes with most frequent mutation in the KIRC Cancer Type



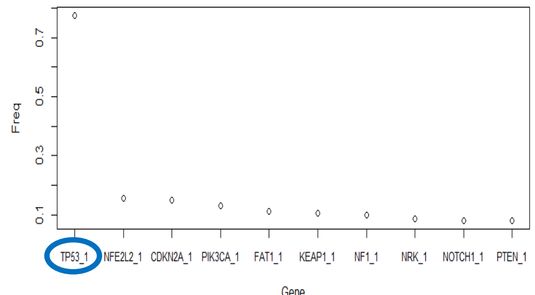
Genes with most frequent mutation in the LAML Cancer Type



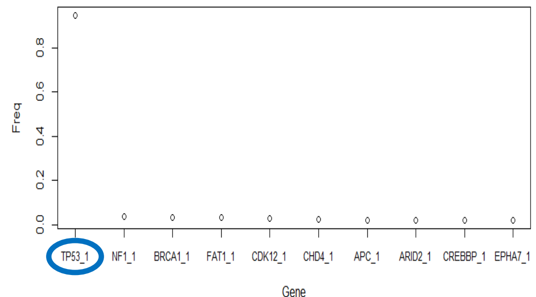
Genes with most frequent mutation in the LUAD Cancer Type



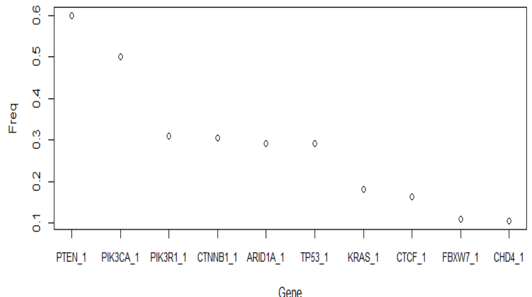
Genes with most frequent mutation in the LUSC Cancer Type



Genes with most frequent mutation in the OV Cancer Type



Genes with most frequent mutation in the UCEC Cancer Type



# Outline

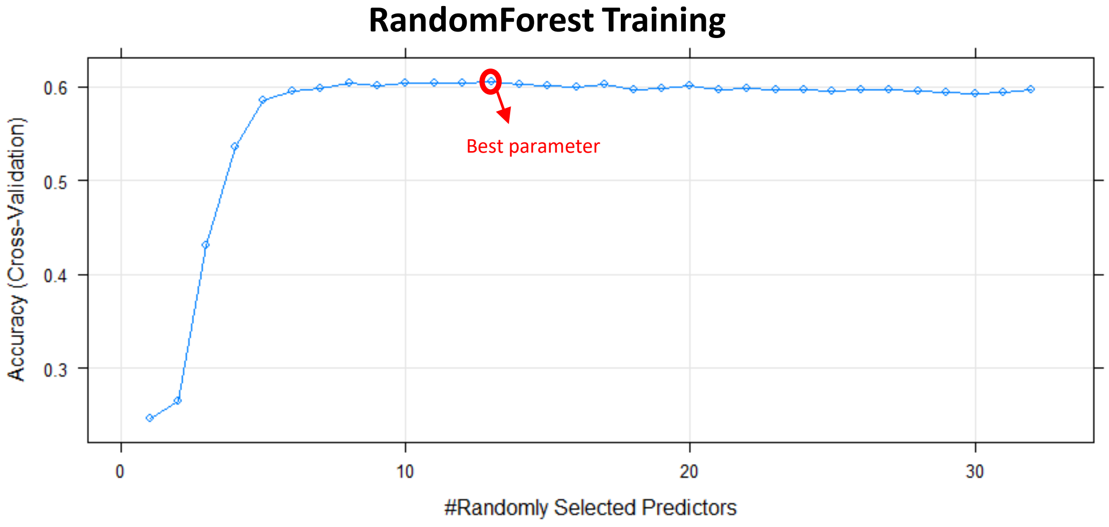
- Is it possible to predict cancer type based on genes with somatic mutation in a patient?

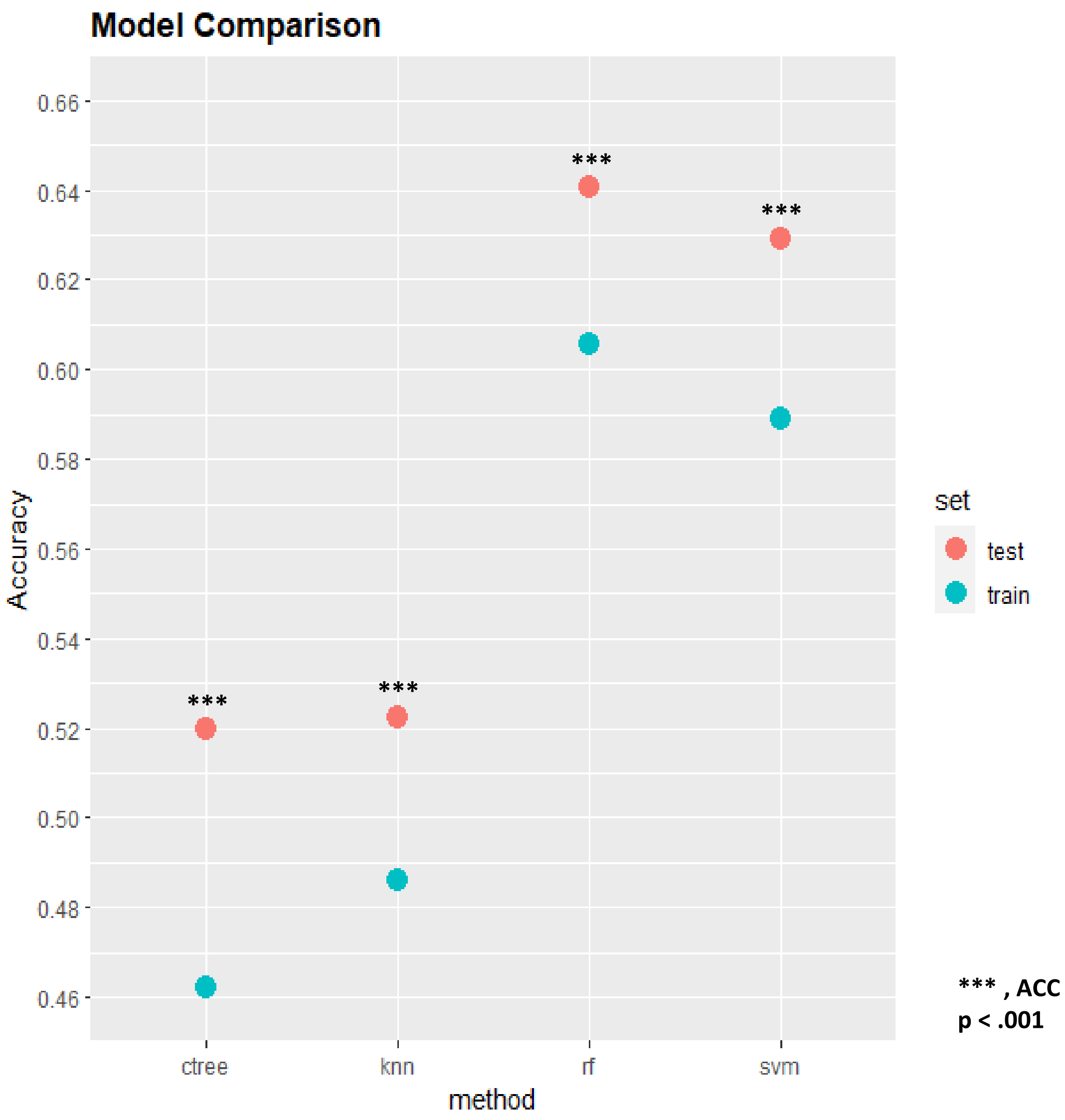
- Is there a «small» set of genes having a good predictive power, or at least as good as the entire set of genes?

- Does the patient grouping based on similarity of mutated genes reflect the grouping based on cancer type?

# Is it possible to predict cancer type?

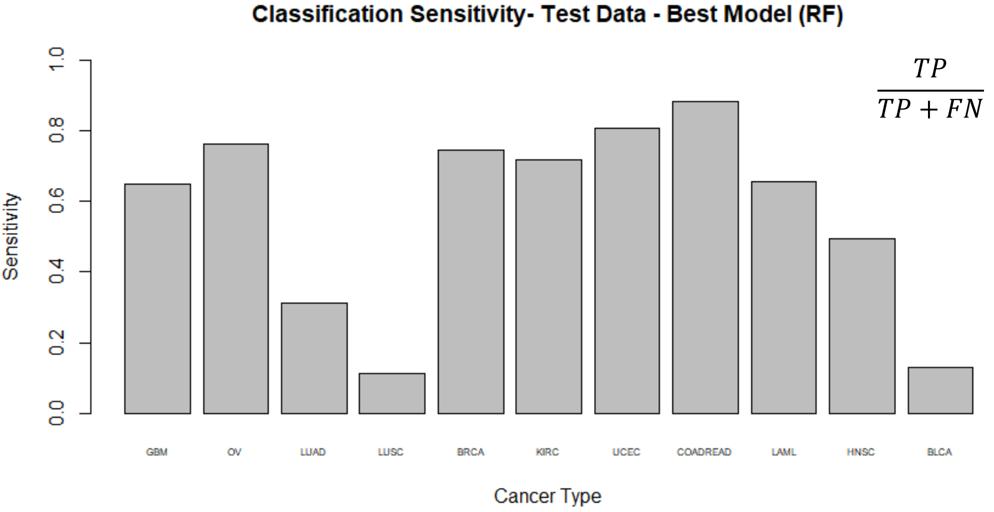
- Compared four models: CIDT, KNN, SVM and RandomForest
- 5-folds CV (25% data test)
- Tuning parameters: mincriterion, k, cost and mtry





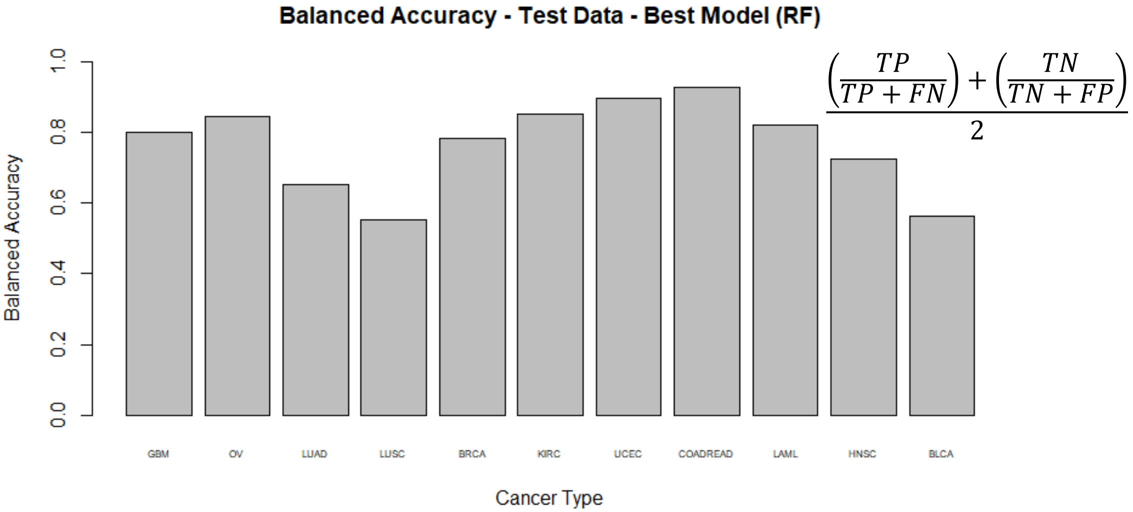
*** , ACC > NIR
p < .001

# Is it possible to predict cancer type?



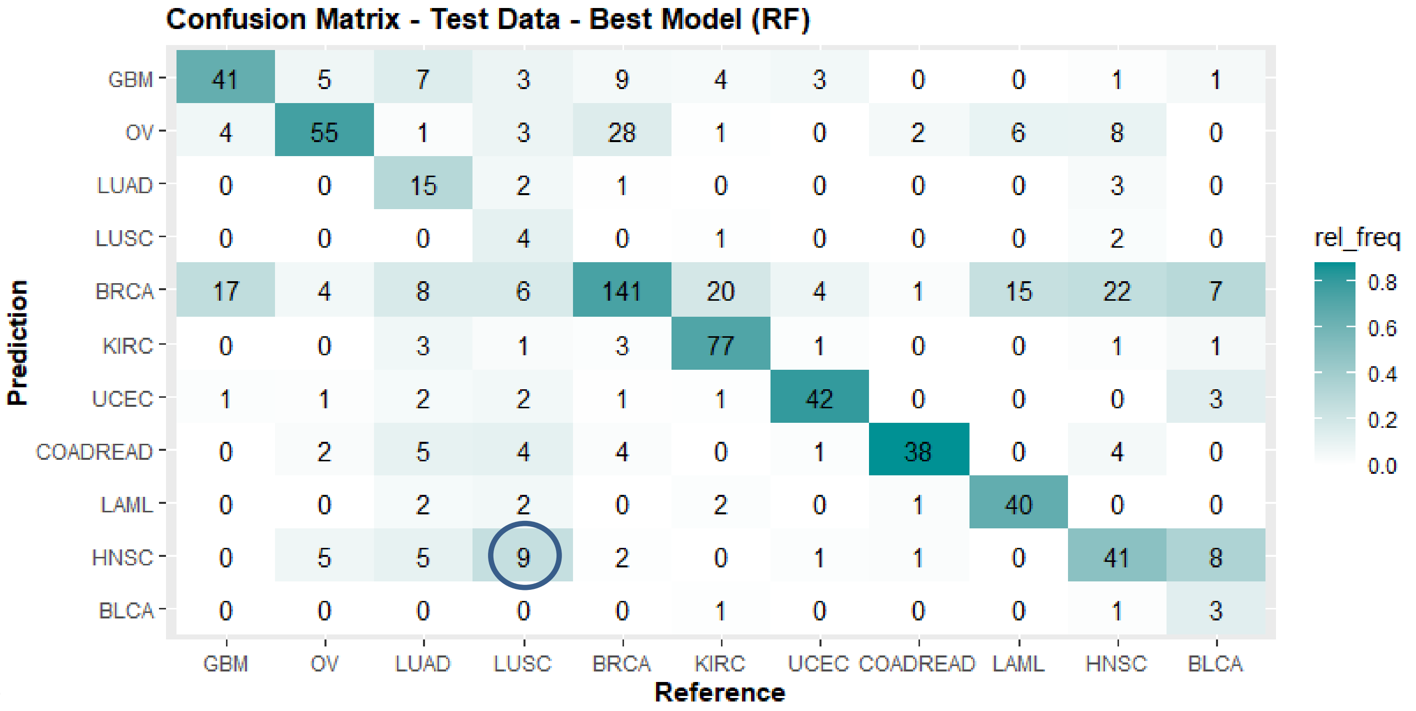
Classification Sensitivity- Test Data - Best Model (RF)

$$\frac{TP}{TP+FN}$$



Balanced Accuracy - Test Data - Best Model (RF)

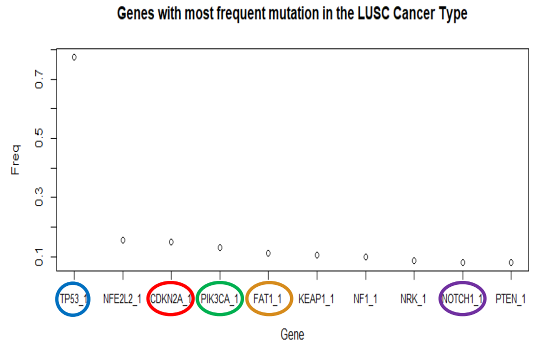
$$\frac{\left(\frac{TP}{TP+FN}\right)+\left(\frac{TN}{TN+FP}\right)}{2}$$



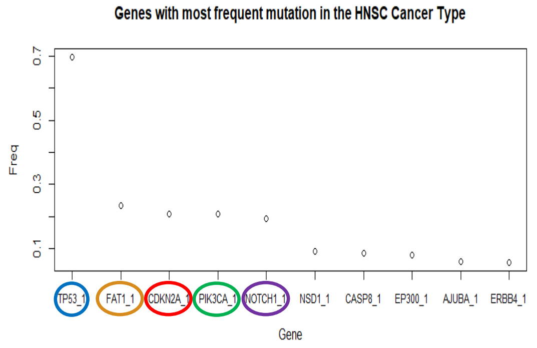
Confusion Matrix - Test Data - Best Model (RF)

| Prediction \ Reference | GBM | OV | LUAD | LUSC | BRCA | KIRC | UCEC | COADREAD | LAML | HNSC | BLCA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| GBM | 41 | 5 | 7 | 3 | 9 | 4 | 3 | 0 | 0 | 1 | 1 |
| OV | 4 | 55 | 1 | 3 | 28 | 1 | 0 | 2 | 6 | 8 | 0 |
| LUAD | 0 | 0 | 15 | 2 | 1 | 0 | 0 | 0 | 0 | 3 | 0 |
| LUSC | 0 | 0 | 0 | 4 | 0 | 1 | 0 | 0 | 0 | 2 | 0 |
| BRCA | 17 | 4 | 8 | 6 | 141 | 20 | 4 | 1 | 15 | 22 | 7 |
| KIRC | 0 | 0 | 3 | 1 | 3 | 77 | 1 | 0 | 0 | 1 | 1 |
| UCEC | 1 | 1 | 2 | 2 | 1 | 1 | 42 | 0 | 0 | 0 | 3 |
| COADREAD | 0 | 2 | 5 | 4 | 4 | 0 | 1 | 38 | 0 | 4 | 0 |
| LAML | 0 | 0 | 2 | 2 | 0 | 2 | 0 | 1 | 40 | 0 | 0 |
| HNSC | 0 | 5 | 5 | 9 | 2 | 0 | 1 | 1 | 0 | 41 | 8 |
| BLCA | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 3 |

Prediction
Reference
rel_freq
0.8
0.6
0.4
0.2
0.0



Genes with most frequent mutation in the LUSC Cancer Type



Genes with most frequent mutation in the HNSC Cancer Type

# Outline

- Is it possible to predict cancer type based on genes with somatic mutation in a patient?

- Is there a «small» set of genes having a good predictive power, or at least as good as the entire set of genes?

- Does the patient grouping based on similarity of mutated genes reflect the grouping based on cancer type?

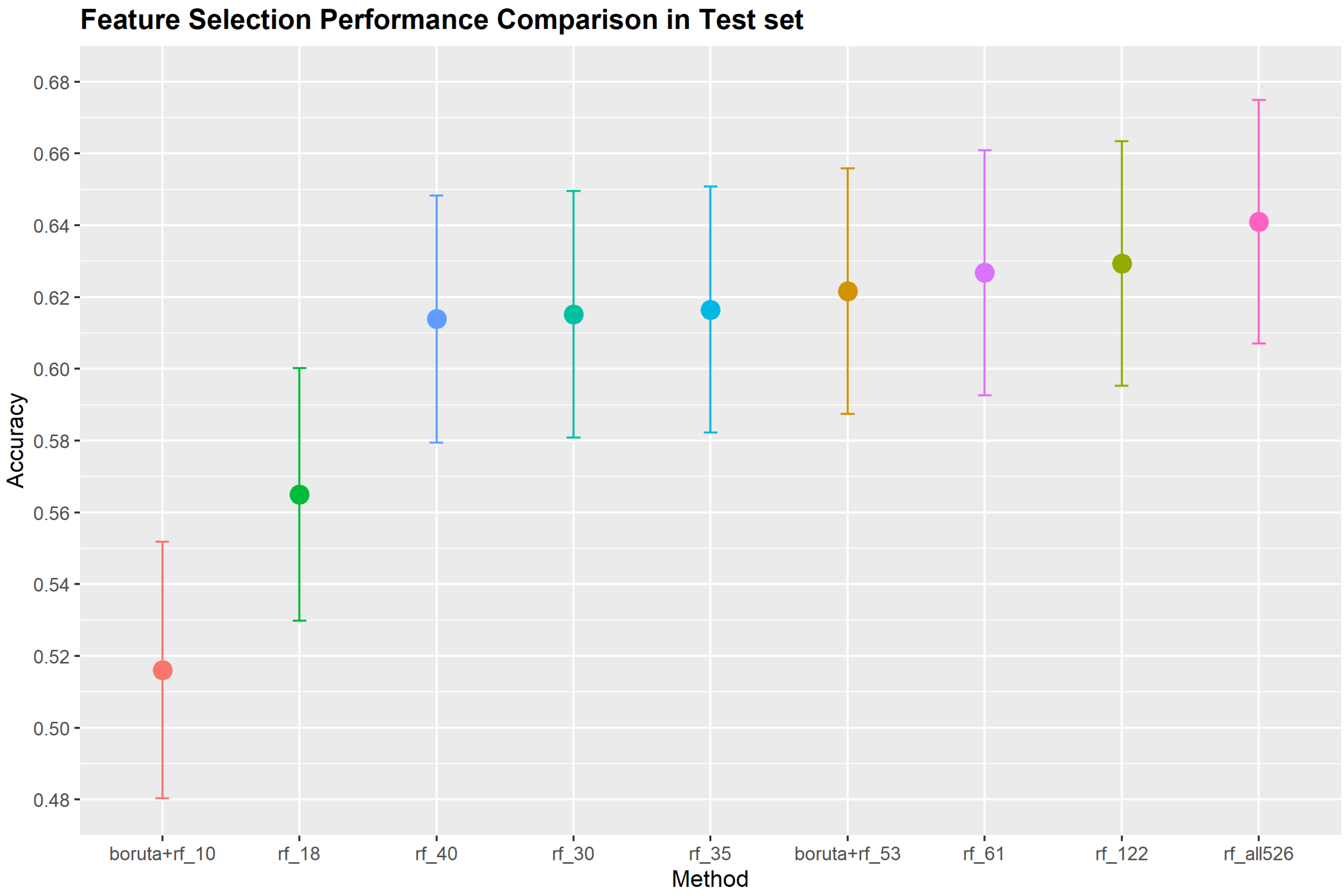
# Is there a small set of genes having a good predictive power?

- Used two approaches for gene selection:

  1. RandomForest, using different threshold values for Gini score (>2, >3, >4, >5, >6, >10)
  2. Boruta: a) selection of relevant genes in all CV runs;
     b) selection of genes defined as relevant in at least one CV run

- Training RandomForest with selected genes
- Model evaluation using the test set

# Is there a small set of genes having a good predictive power?



- Good accuracy for set of 30-40 genes and above
- Performance decreases significantly under 20 genes

# Outline

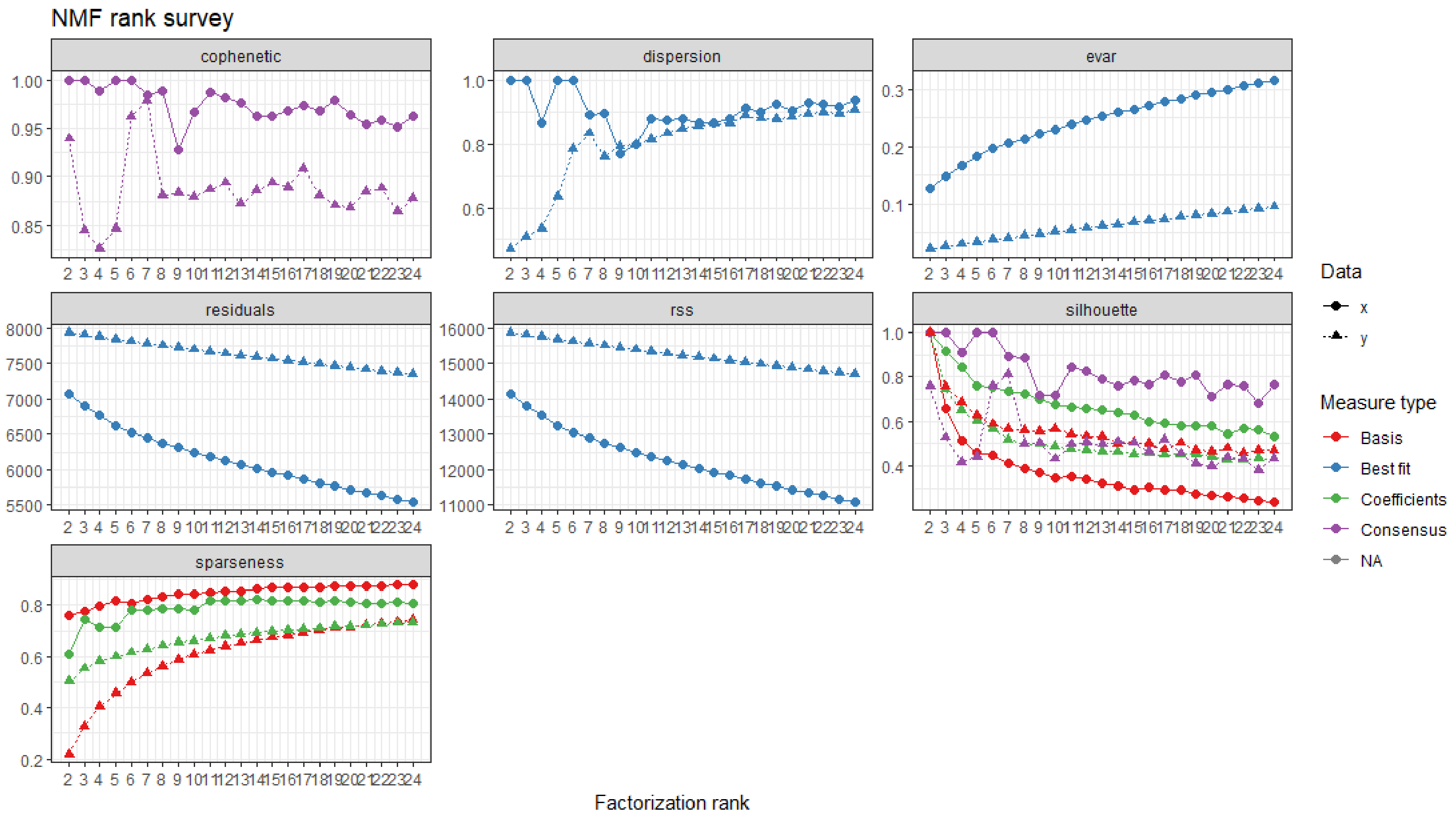
- Is it possible to predict cancer type based on genes with somatic mutation in a patient?

- Is there a «small» set of genes having a good predictive power, or at least as good as the entire set of genes?

- Does the patient grouping based on similarity of mutated genes reflect the grouping based on cancer type?

# Does the clustering based on mutated gene similarity reflect cancer type?

- NMF approach (Lee algorithm, 10 runs)



- Optimal K > 24, therefore higher than the number of cancer types

# Does the clustering based on mutated gene similarity reflect cancer type?

- Even when setting k = 11 (Lee algorithm, 30 runs), the clustering is different than real classes
- E.g., BRCA cancer is clusterized as follows:

| **k** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **cases** | **180** | 27 | 31 | **181** | 5 | 4 | 13 | **81** | 8 | **227** | 6 |

# Conclusions

- Is it possible to predict cancer type based on genes with somatic mutation in a patient?

  Is is possible to obtain a good predictive accuracy for specific cancer types whereas model performance is low for others

- Is there a «small» set of genes having a good predictive power, or at least as good as the entire set of genes?

  Yes, is is possible to select a set of 30-40 genes having a predictive power similar to the one obtained with all the genes

- Does the patient grouping based on similarity of mutated genes reflect the grouping based on cancer type?

  No, several cancer types seem to be composed by subsets of patients with different somatic mutation profiles

Thanks for your attention!