# Genomic

MATTIA FRATELLO

ALESSIA LUPATTELLI

ILARIA MARINI

PAOLO MEROLA

STEFANIA SPEDALE

# Fonte dei dati: The Cancer Genome Atlas

Collaborazione tra il National Cancer Institute e il National Human Genome Research Institute

Mappe estese multidimensionali (espressioni geiche, mutazioni geniche, epigenetica) di 33 tipi di tumore, dei quali 10 rari

Primo dataset pubblico di questo genere, oltre 2 petabytes con 9976 geni associati a 33 tipi di tumore

- Tumori più osservati: carcinoma della mammella (brca), carcinoma della prostata (prad)
- Tumore meno osservato: colangiocarcinoma (chol)

# Panoramica sui dati

Dataset analizzato:

- 4476 **osservazioni**: campioni di materiale genetico di 30 diverse tipologie di cancro
- 9956 **attributi**:
  - *Id*: identificativo campione materiale genetico
  - 9954 *geni*: livello di espressione genica, inteso come numero di copie di RNA prodotte a partire dal gene
  - *Label* : tipologia di tumore (30 livelli)

*Osservazione*:

Rispetto al dataset originale del progetto *Atlas* sono stati eliminati:

- circa 30 geni
- 3 tipi di tumore, tra i più rari
- enorme numero di osservazioni

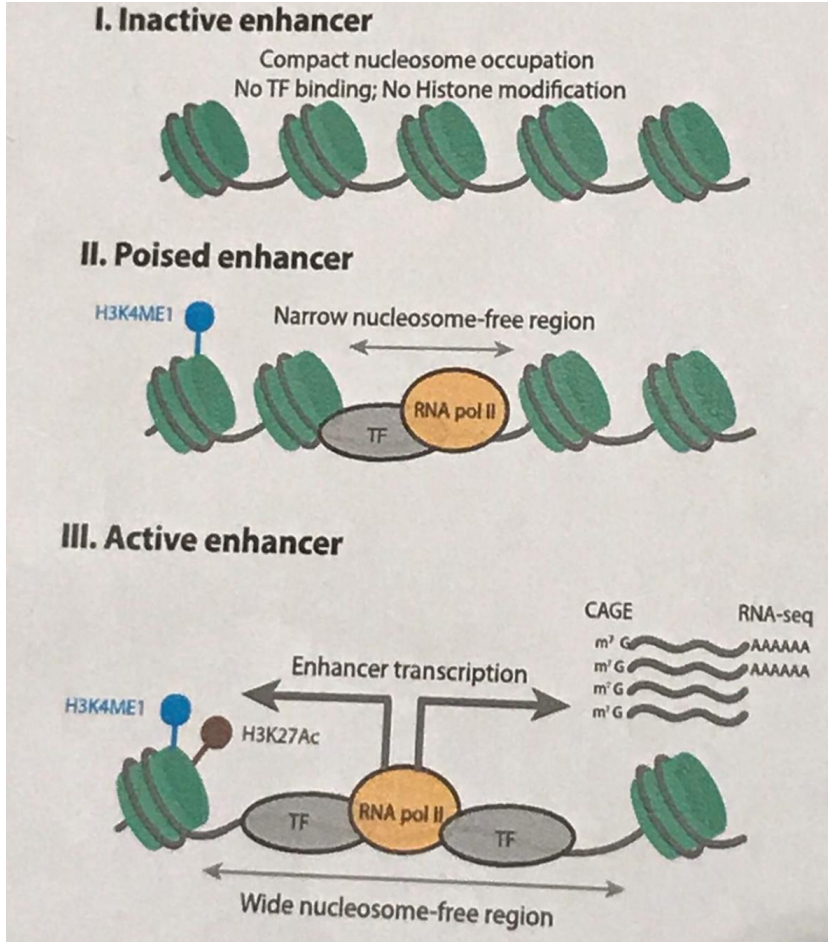
tuttavia la proporzione delle osservazioni rispetto alle label rispecchia abbastanza quella dei dati originali

# Contesto: alcuni concetti base

- **Genoma**: totalità dell'informazione ereditaria di un organismo (è codificata nel DNA)

- **Geni**: regione del DNA che viene tradotta in RNA. Alcuni tipi di RNA (mRNA) codificano per le componenti delle proteine; l'insieme delle proteine funzionali (proteoma) determina l'operatività della cellula (omeostasi).

  (Il genoma umano è costituiti da circa 20'000-25'000 geni)



# Obiettivo

Identificare espressioni geniche peculiari per i tipi di tumore con lo scopo di individuare quali sono le porzioni del patrimonio genetico sono peculiari per una specifica tipologia di cancro

*Perché* →

- ✓ **PREVENZIONE**: Individuare pazienti che hanno alto rischio di essere soggetti ad un tumore
- ✓ **TERAPIE**: possibilità effettuare cure mirate

↓

*In che modo?*

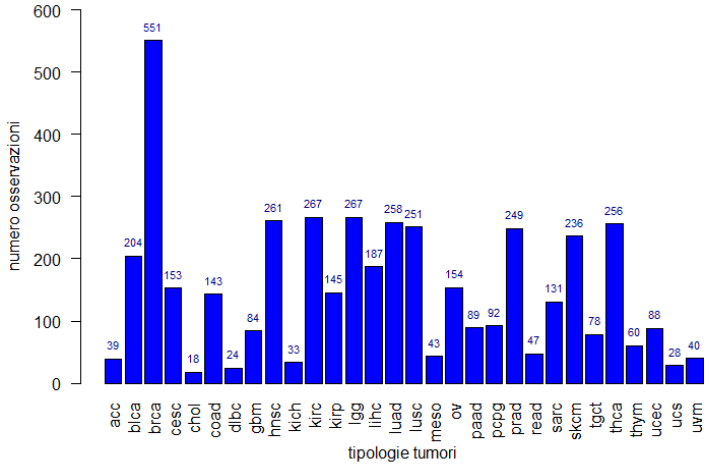
- **Modelli supervisionati**: predire la classe di tumore associata a un particolare tipo di campione di materiale genetico
- **Modelli non supervisionati**: individuare similarità nei livelli di espressione genica

# Preprocessing

- No missing values

Problemi:

- Sbilanciamento variabile target

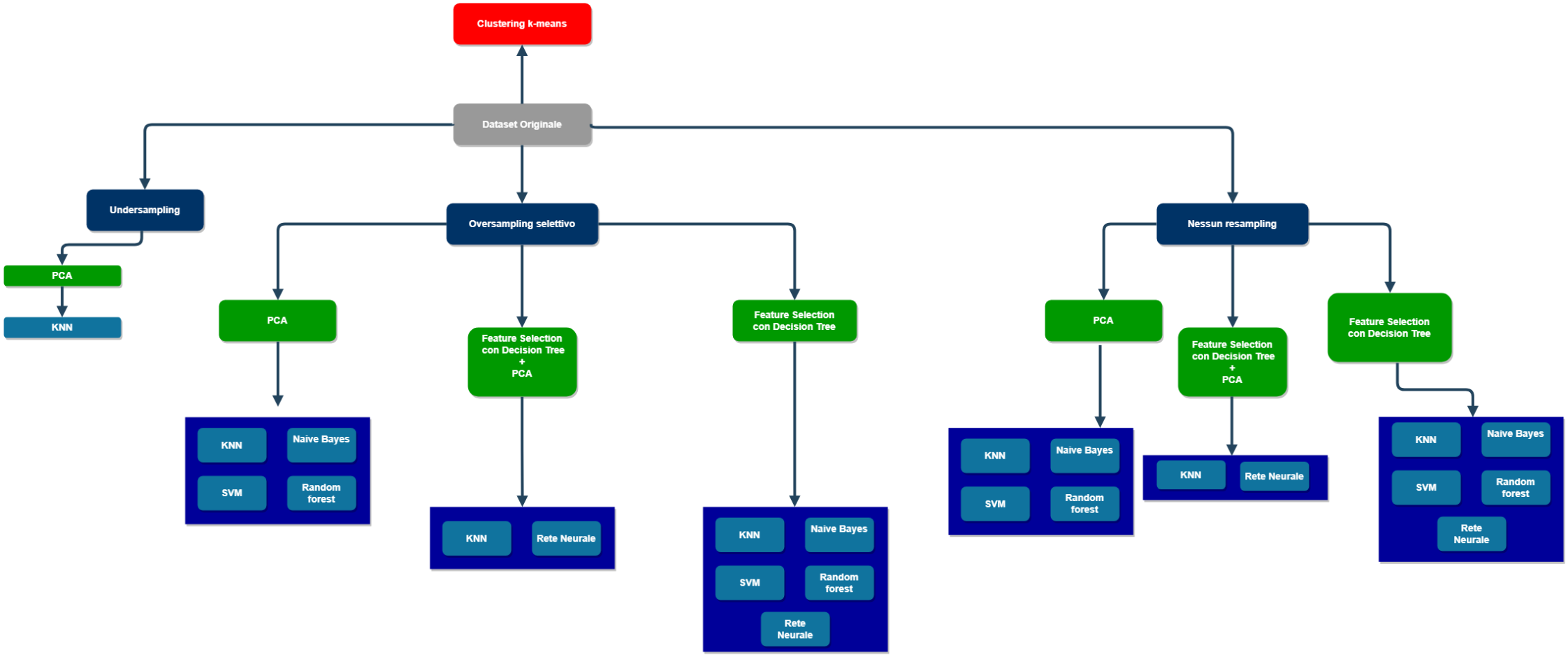



Over sampling bilanciato

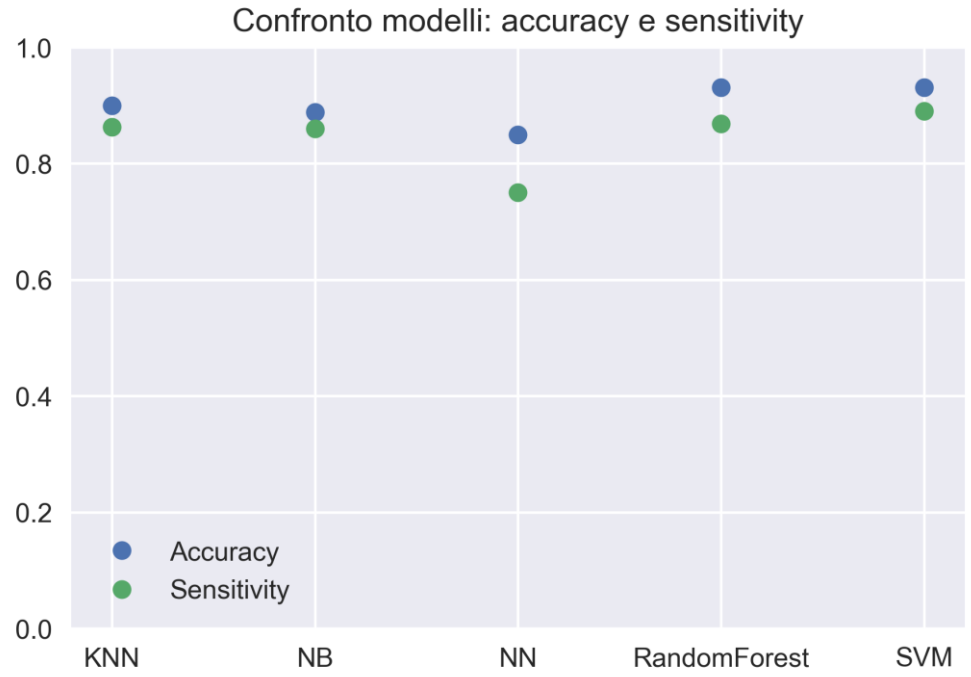
- Numero elevato di features:
  9954 espressioni geniche

- PCA: 136 componenti principali
- Decision Tree: 85 geni

# Procedimento



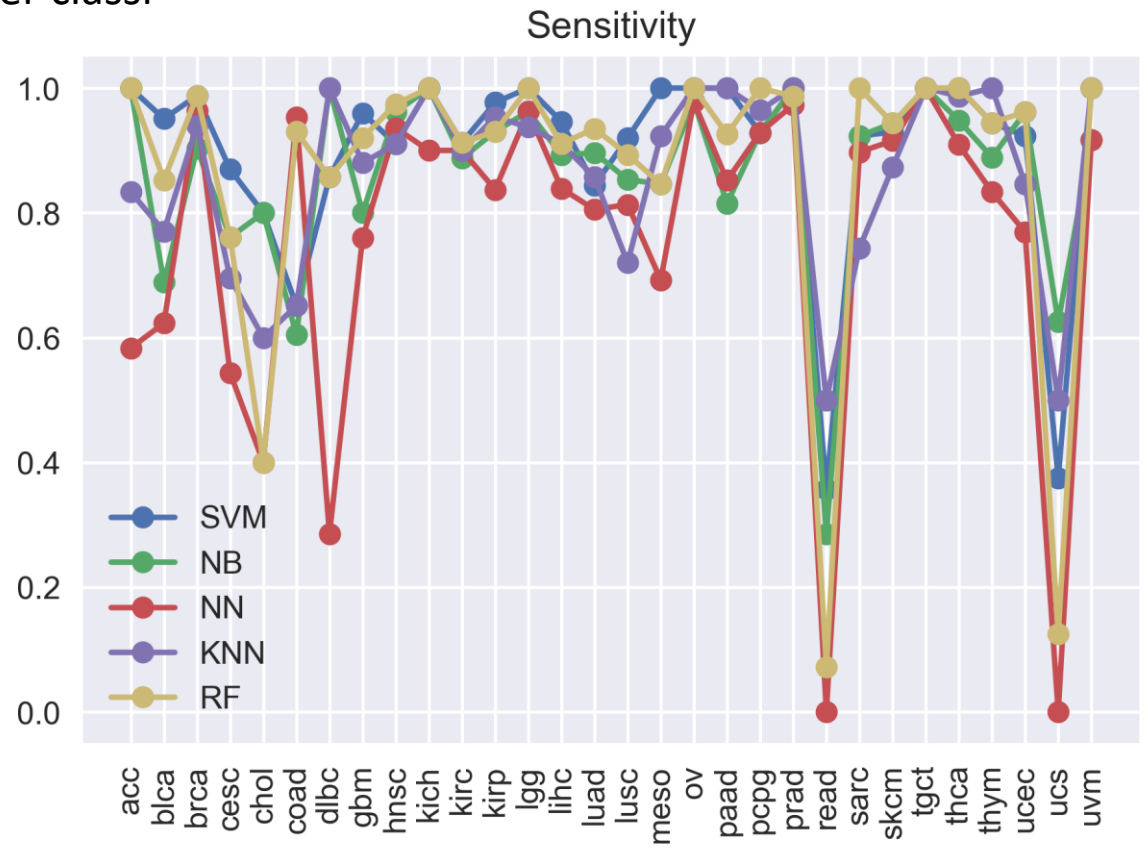
# Confronto modelli




| | Accuracy | Sensitivity |
|---|---|---|
| SVM | 0.932 | 0.898 |
| RF | 0.931 | 0.868 |
| KNN | 0.904 | 0.863 |
| NB | 0.888 | 0.867 |
| NN | 0.854 | 0.759 |

# Confronto modelli

Analizzando i risultati per classi



# Clustering: k-means

**Obiettivo:** individuare i tumori con espressioni geniche caratteristiche e trovare tumori con espressioni geniche simili

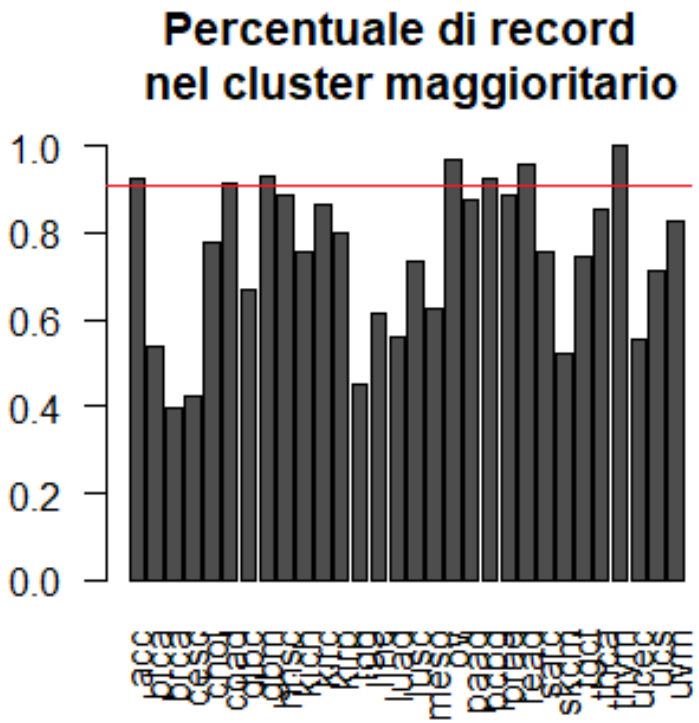
- sono stati generati clustering con k = 2,…,30

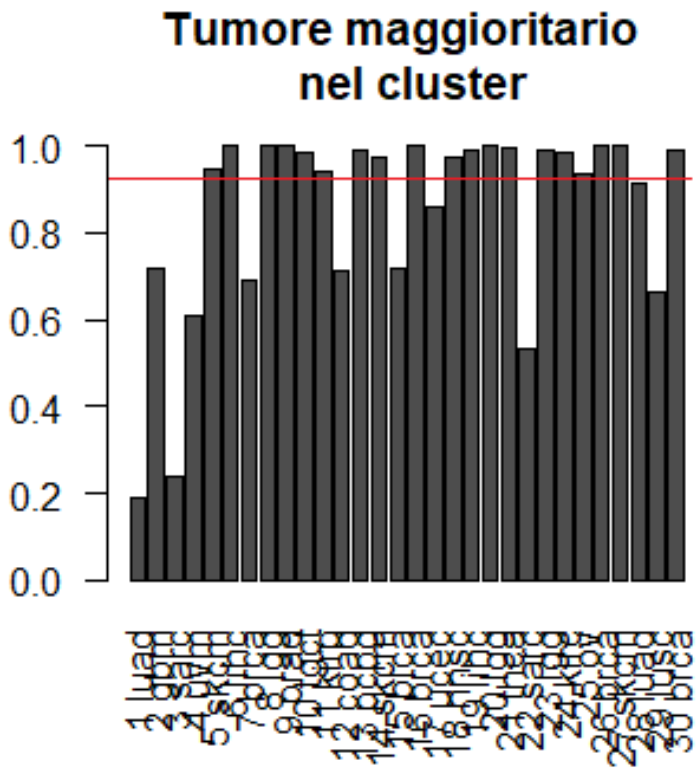
**Risultati:**

- distinzione tra cluster non ottima (max SSB/SST = 0,38, max Jaccard = 0,44)

- alcuni tumori sono stati ben identificati da un unico cluster

# Clustering: k-means




I tumori ov,  pcpg sono ben rappresentati in un unico cluster

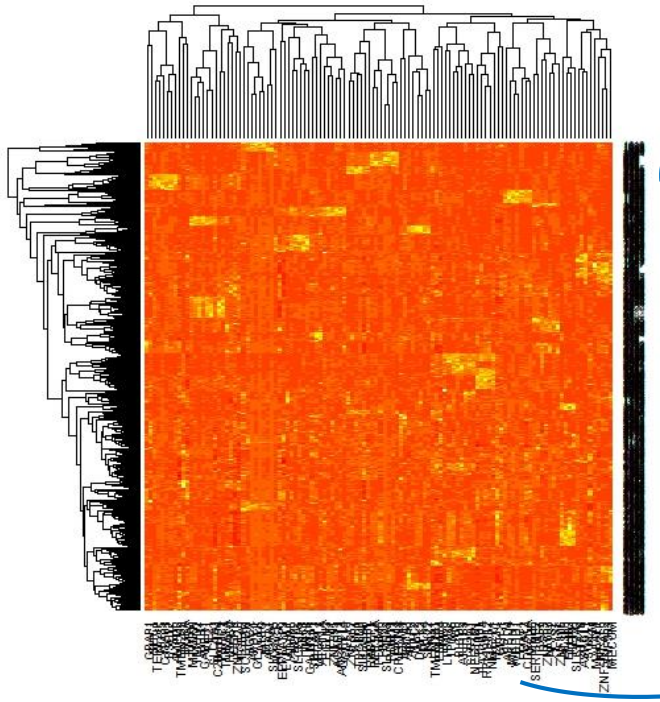
# Heat map e clustering gerarchico

Obiettivo:

- rappresentazione visiva espressione genica
- Identificare pattern comuni

Riduzione dimensionalità

- Undersampling
- Features selection con decision tree

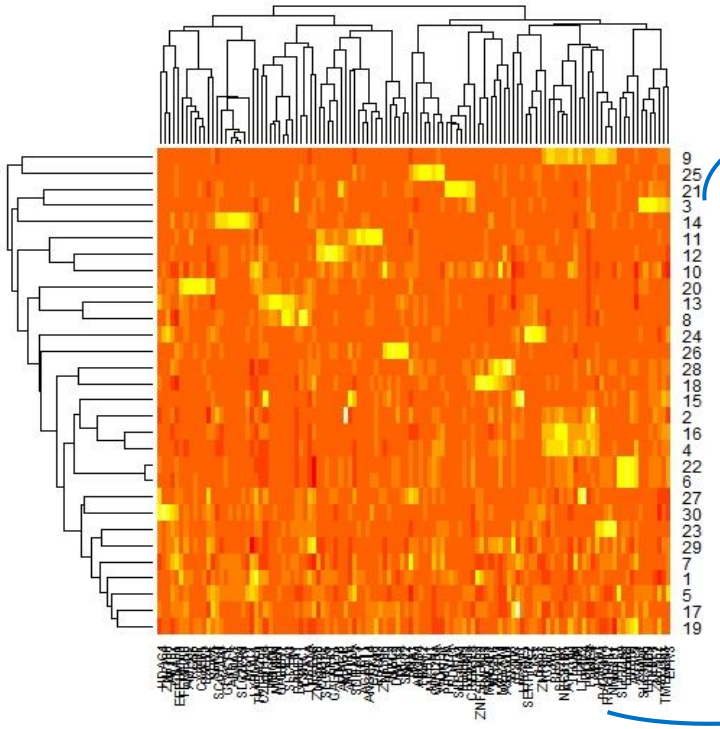


540 campioni di materiale genetico

116 geni (decision tree)

# Esempio

Identificare geni espressi nelle diverse tipologie di cancro



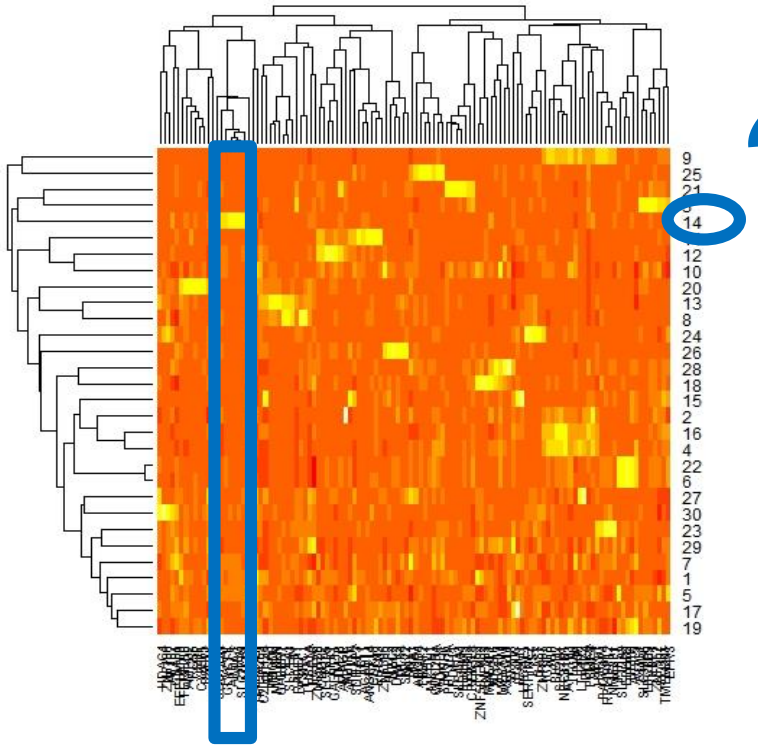

30 tipologie cancro

Si considera il livello medio di espressione genetica

116 geni (decision tree)

# Esempio

Identificare geni espressi nelle diverse tipologie di cancro



Esempio: cancro al fegato

Geni espressi:

- SLC45A3
- TMOD2
- A1BG
- CYB561
- AMPD2
- BSCL2
- EPHB2
- C20orf194
- TCEA2
- ASRGL1
- TMEM259
- ABCD4

# Grazie per l’attenzione

# Classificazione: RF

Random Forest:
- creazione di n di alberi
- ciascun albero, partendo da un subset casuale dei dati, fornisce una classificazione
- la classificazione che ha ricevuto più ‘‘voti’’ fornisce la classificazione finale

**PREPROCESSING**:
Due dataset: originale (senza modifiche) e con oversampling parziale
Due modalità di feature selection su ogni dataset: Decision Tree e PCA

# Classificazione: RF

**TUNING**

Stabilire il numero ottimale di alberi

**Metrica utilizzata**:

Accuracy

**Classificatore scelto**:

Random Forest con 300 alberi su dataset bilanciato e precedente feature selection tramite Decision Tree

# Classificazione: RF

**Risultati**

Accuracy = 93.13%

Sensitivity = 80.23%

**Risultati per classi**:

Accuracy & Sensitivity classe minoritaria (chol) = 69.96% & 40%

Accuracy & Sensitivity classe maggioritaria (brca) = 99.10% & 98.79%

# Classificazione: NN

L’algoritmo addestra una rete MLP (Multi-Layer Perceptron) e la utilizza per la classificazione.

**PREPROCESSING**:

Due dataset: originale (senza modifiche) e con oversampling parziale.

Due modalità di preprocessing su ogni dataset: Standardizzazione o PCA

**FEATURE** SELECTION:

Due insiemi di variabili utilizzate:

Selezionate da Decision Tree (121 colonne)

Selezionate da Random Forest

# Classificazione: NN

**TUNING**

Stabilire il numero di layers e il valore di decay ottimali

**Metriche utilizzate**:

1. Accuracy
2. Sensitivity media

**Metodi di validazione**:

1. 10-fold cross validation

# Classificazione: NN

**RISULTATI**

**Migliore Accuracy:** 85%

- dataset originale

- feature selection con Random Forest

- Preprocessing: PCA

- 84-6-30, 1 hidden layer, decay = 0.1

**Migliore Sensitivity media 75**%

Stesso dataset

# Classificazione: NB

L'algoritmo classifica le osservazioni secondo il t. di Bayes, ovvero assegnando ad ogni nuova osservazione la classe che massimizza la probabilità a posteriori della variabile target, date le variabili esplicative, condizionalmente indipendenti tra loro

$$P(\omega / A) = \frac{P(A / \omega) P(\omega)}{P(A)}$$

Prob posteriori $\propto$ likelihood x Prob priori

- $w_j$ categoria del target con j = 1, … k
- A variabile esplicativa
- P(A) probabilità del campione
- P(w) probabilità a priori del class attribute
- P(w|A) probabilità a posteriori del class attribute rispetto alla variabile esplicativa
- P(A|w) verosimiglianza della variabile esplicativa dato il class attibute

# Classificazione: NB

**PREPROCESSING**

Dimensionalità del dataset ridotta con:

1. PCA
2. Feature selection con Decision Tree

**RISULTATI**

**Migliori Accuracy:** 88.8%

- dataset originale
- feature selection con Decision tree

**Sensitivity:**

- più elevata: 1 (su 5 variabili)
- più bassa: 0.28 (class attribute **read**)

**Risultati ottenuti dal classificatore non soddisfacenti..**

**…provo a migliorarne la performance agendo su:**

**multicollinearità** dei predittori
→ *eliminate una serie di var correlate dal set di variabili del Decision Tree*

**binarizzazione** della variabile target
→ *create 30 var target binarie, una per ciascuna classe di Label*

# Classificazione: NB

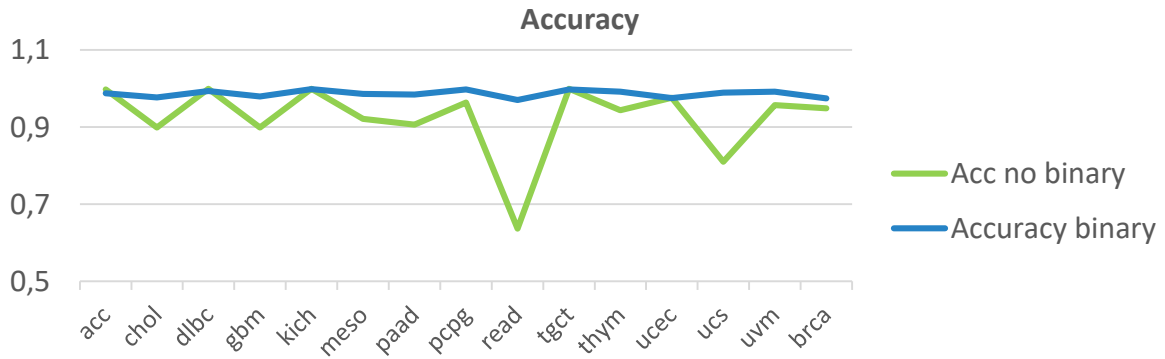
## Multicollinearità

- nessun aumento della performance, in termini di accuracy, dopo aver eliminato le variabili collineari (10 su 121)

## Binarizzazione della variabile target

- binarizzazione delle 14 classi meno rappresentate (<100 osservazioni) della variabile target **Label**
- costruzione del modello e predizione per ciascuna delle 14 variabili target considerate
- stesso processo eseguito anche per la variabile **brca**, la più rappresentata, per confronto

→ Forte impatto su tutte le metriche, performance del modello notevolmente migliorata!



# Classificazione: NB

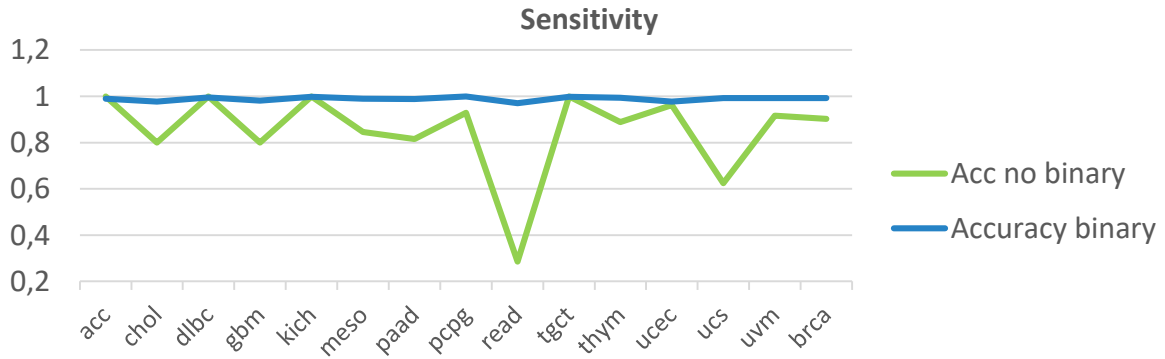
## Multicollinearità

- nessun aumento della performance, in termini di accuracy, dopo aver eliminato le variabili collineari (10 su 121)

## Binarizzazione della variabile target

- binarizzazione delle 14 classi meno rappresentate (<100 osservazioni) della variabile target **Label**
- costruzione del modello e predizione per ciascuna delle 14 variabili target considerate
- stesso processo eseguito anche per la variabile **brca**, la più rappresentata, per confronto

→ Forte impatto su tutte le metriche, performance del modello notevolmente migliorata!



# Classificazione: KNN

L'algoritmo classifica il dato in base alla maggioranza dei voti dei suoi *k* vicini, secondo la distanza euclidea.

**PREPROCESSING**

**Dimensionalità**

La dimensionalità del dataset è stata ridotta con:

1. PCA
2. Feature Selection con Decision Tree
3. Feature Selection con Decision Tree + PCA

**Varianza**

E' stata effettuata una standardizzazione degli attributi

# Classificazione: KNN

**TUNING**

Stabilire il valore di K ottimale

**Metriche utilizzate**:

1. Accuracy
2. Sensitivity media

**Metodi di validazione**:

1. 10-fold cross validation
2. Bootstrap

# Classificazione: KNN

**RISULTATI**

**Migliore Accuracy**: **90%**

- dataset originale

- feature selection con Decision Tree + PCA

- K = 5

**Migliore Sensitivity media tra le classi meno rappresentate: 86%**

**Migliore Sensitivity minima: 50%**

**-** dataset oversampling selettivo

**-** PCA

- K = 3

# Classificazione: SVM

L'algoritmo SVM classifica le osservazioni individuando l'iperpiano che massimizza la distanza tra gli elementi più vicini di due classi

Variabile **multiclasse**: - vengono comparate le classi a coppie
- assegnazione finale dell'osservazione alla classe maggiormente individuata nel corso del confronto a coppie

**PREPROCESSING**

La dimensionalità del dataset è stata ridotta con:

1. PCA
2. Feature Selection con Decision Tree

# Classificazione: SVM

**TUNING**

Stabilire la funzione di costo

**Metriche utilizzate**:

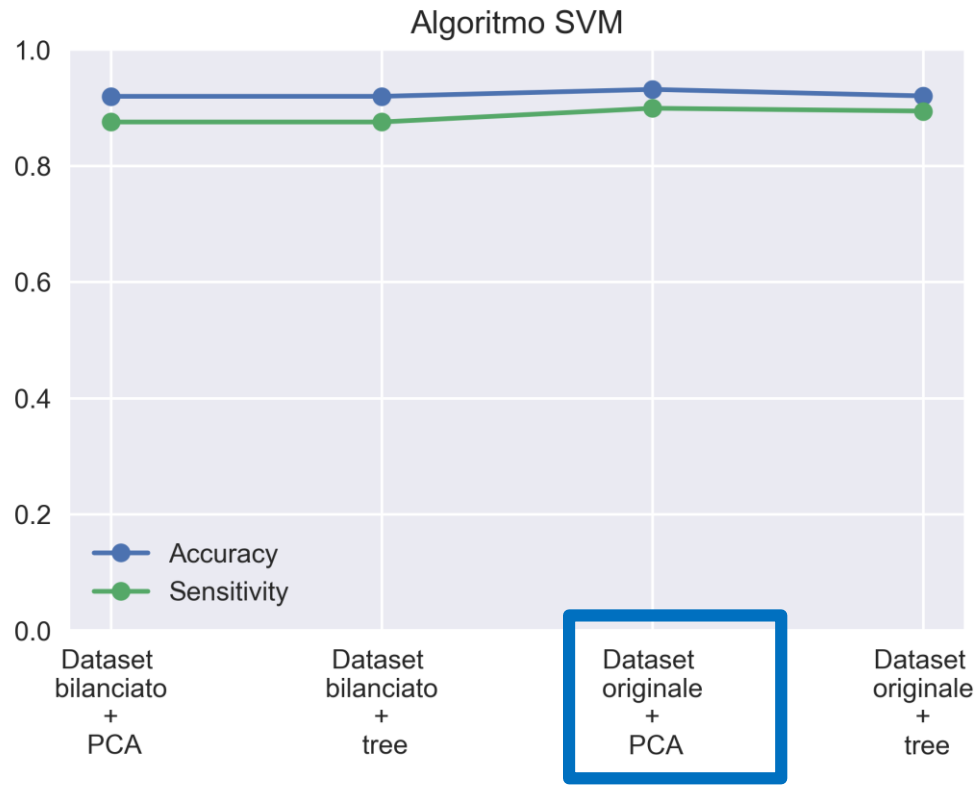
1. Accuracy
2. Sensitivity media

**Metodi di validazione**:

10-fold cross validation

# Classificazione: SVM

**RISULTATI**



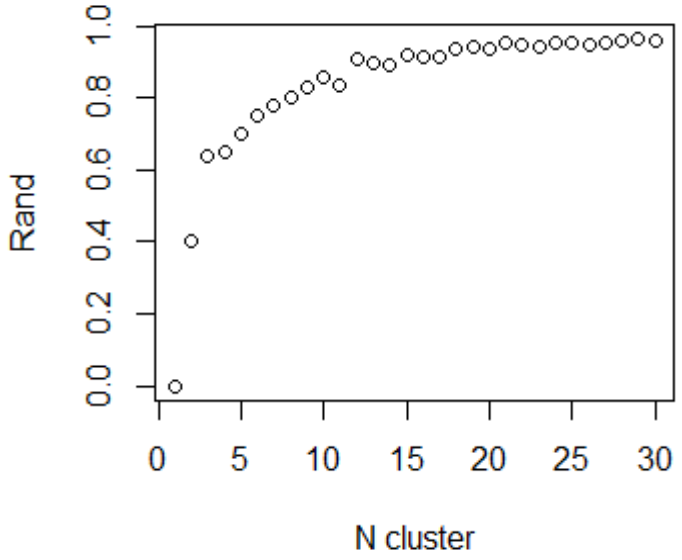

**Migliore Accuracy**: **93.23%**
**Miglior Sensitivity: 89.89%**
- dataset originale
- PCA
- costo = 10

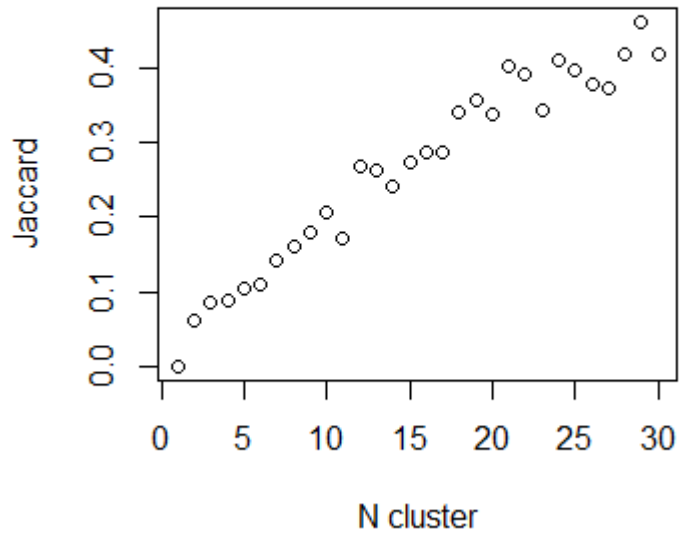
# Clustering: k-means

**INDICI SUPERVISED**

L'indice di Jaccard non indica una buona corrispondenza fra tipi di tumore e cluster:

- Clustering poco affidabile
- Espressioni geniche simili fra tumori diversi

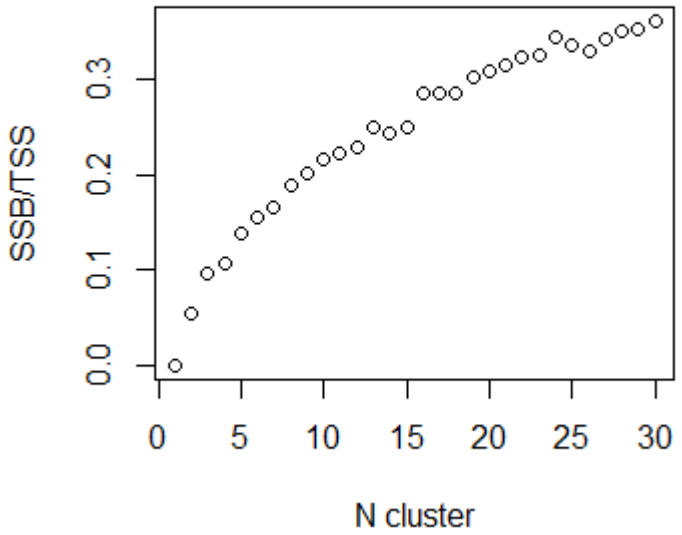


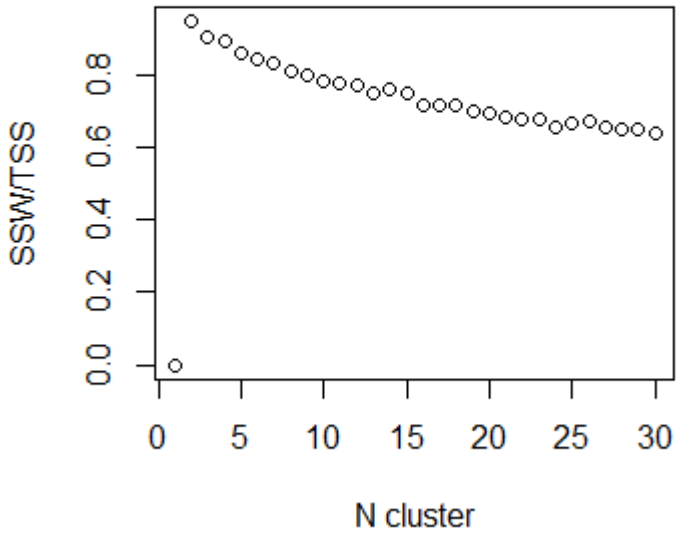



# Clustering: k-means

**INDICI UNSUPERVISED**

La qualità del clustering non è buona -> non si associano espressioni geniche simili a tumori diversi



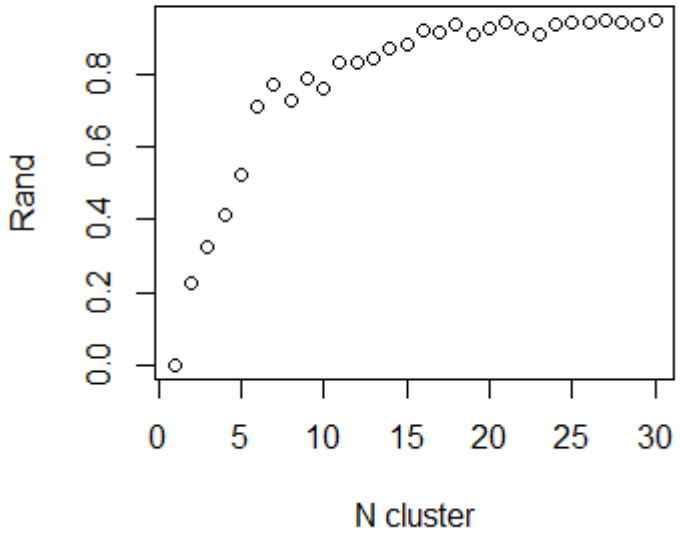
# Clustering: k-means

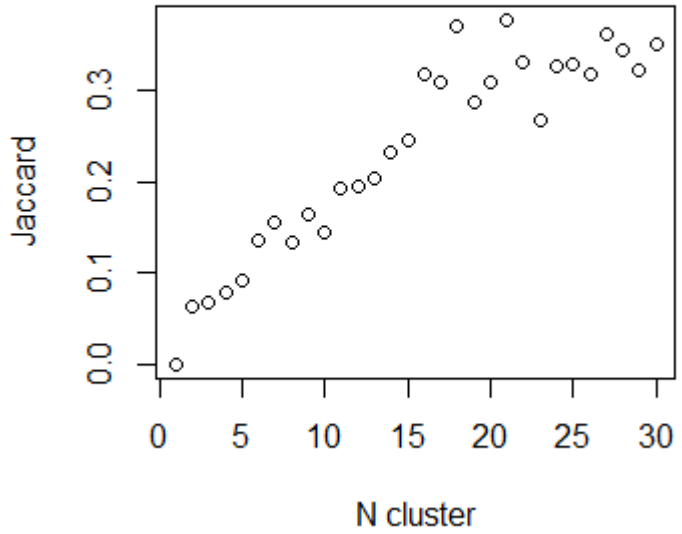
**FEATURE SELECTION**

**INDICI SUPERVISED**

L'indice di Jaccard non indica una buona corrispondenza fra tipi di tumore e cluster:

- Clustering poco affidabile
- Espressioni geniche simili fra tumori diversi

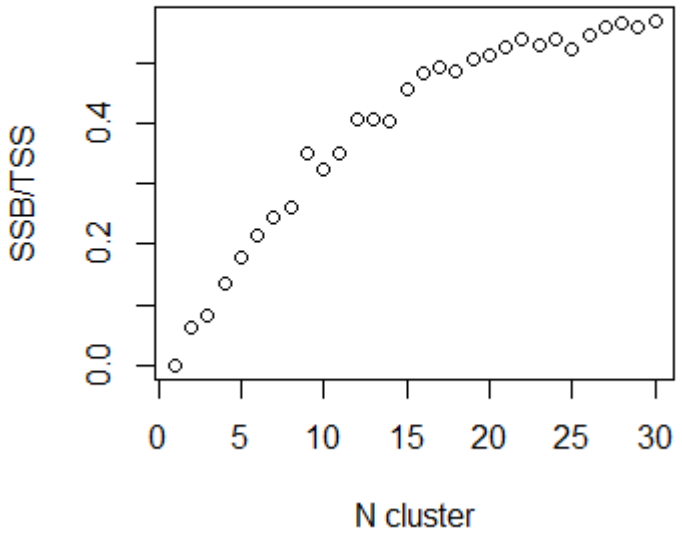


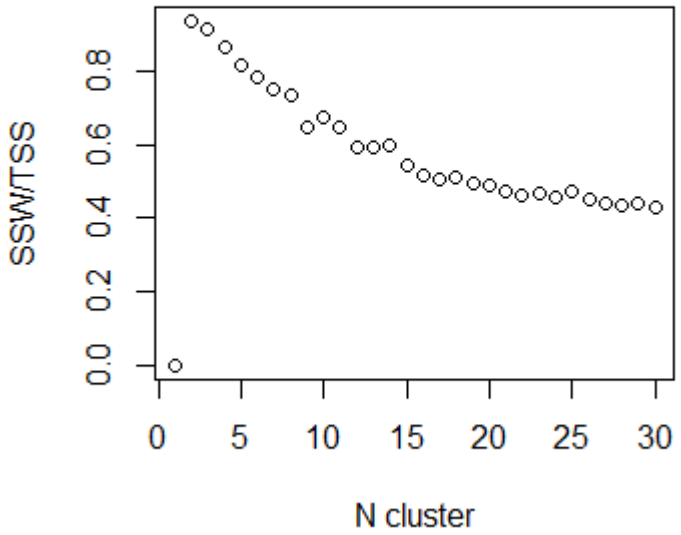



# Clustering: k-means

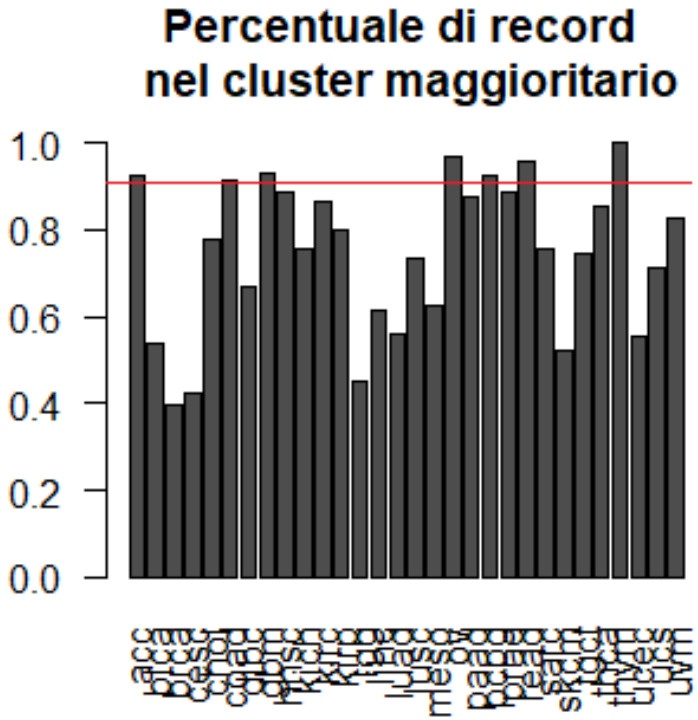
**INDICI UNSUPERVISED**

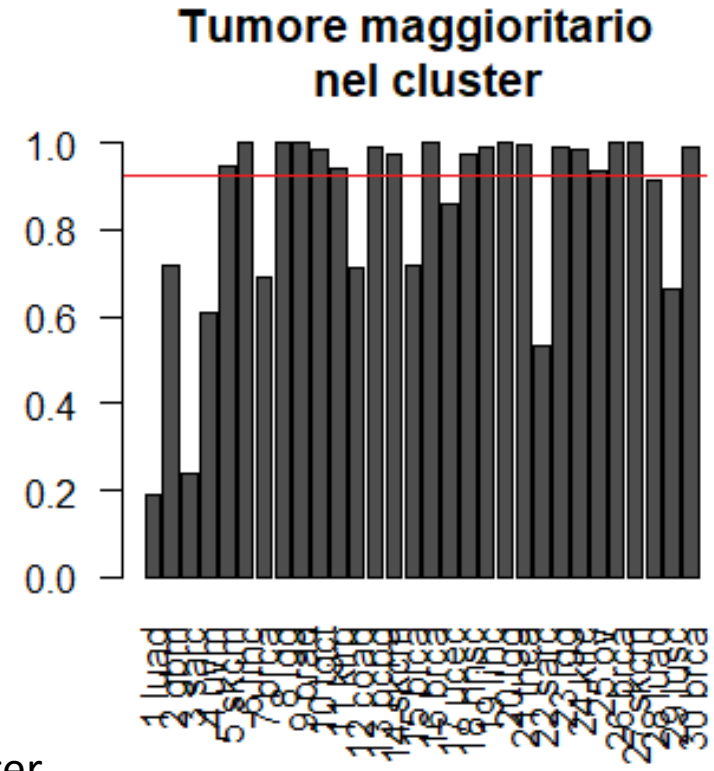
La qualità del clustering è migliorata, ma non è ottima -> non si associano espressioni geniche simili a tumori diversi



# Clustering: k-means

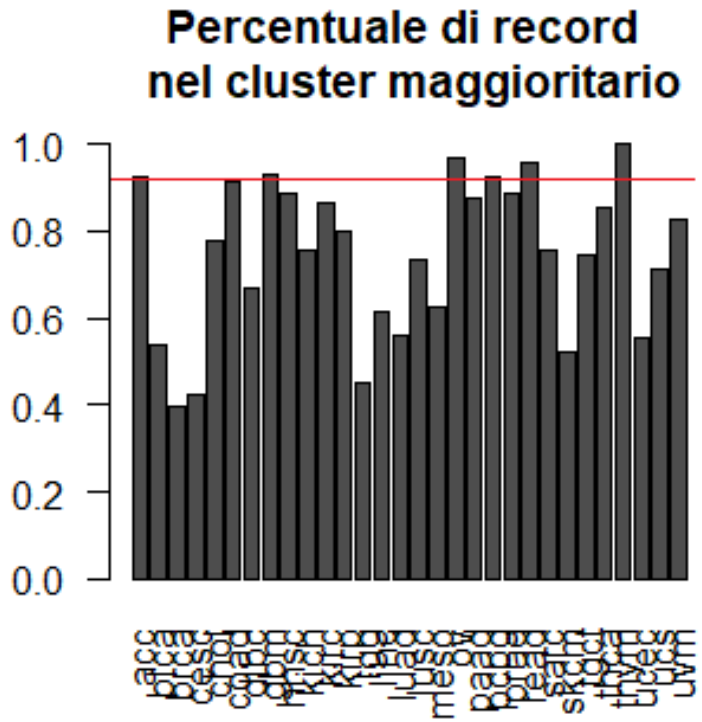


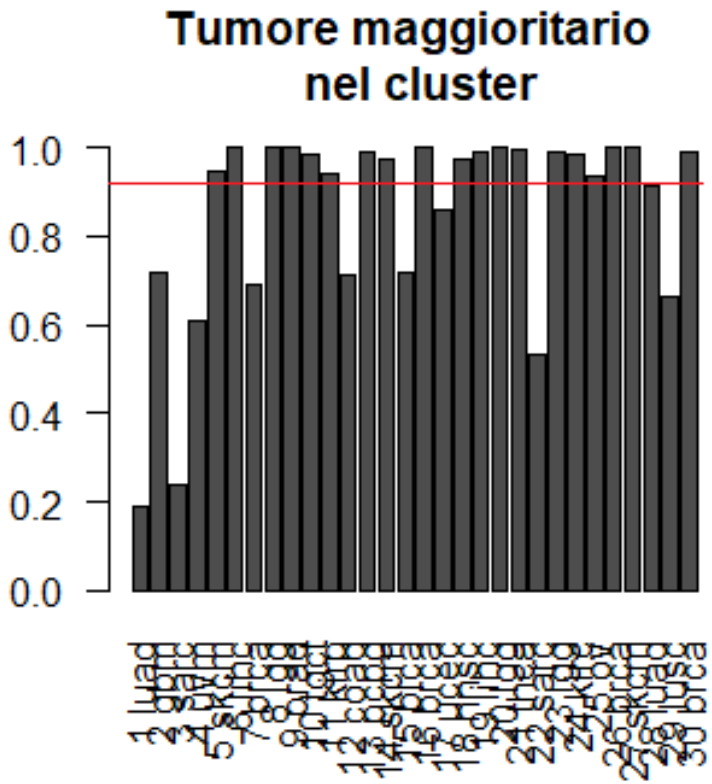




I tumori ov,  pcpg sono ben rappresentati in un unico cluster

# Clustering: k-means

CASO GENI SELEZIONATI




Il tumore ov è ben rappresentato in un unico cluster