POLITECNICO
MILANO 1863

HONOURS PROGRAMME HP-SR
in Information Technology

Prediction of Resistance to Chemotherapy in High Grade Serous Ovarian Adenocarcinoma

Sara Sansone
sara.sansone@mail.polimi.it
Track CSE - Data, Web and Society

Introduction to the Research Project:
A joint collaboration

POLITECNICO
MILANO 1863

Sara Sansone
Computer Science and Engineering

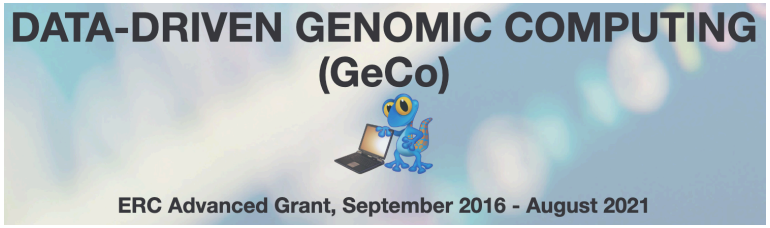
Giada Lalli
Biomedical Engineering

Introduction to the Research Project:
A joint collaboration

POLITECNICO
MILANO 1863

Sara Sansone
Computer Science and Engineering

Giada Lalli
Biomedical Engineering



Prof. Stefano Ceri
Supervisor

Dr. Arif Canakoglu, Dr. Pietro Pinoli
Co-supervisors

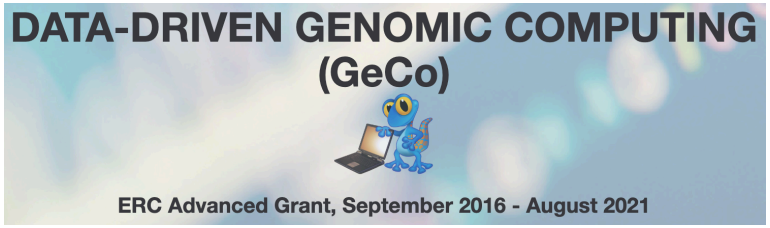
Prof. Francesca Ieva (MOX)
Co-supervisor

Introduction to the Research Project:
A joint collaboration

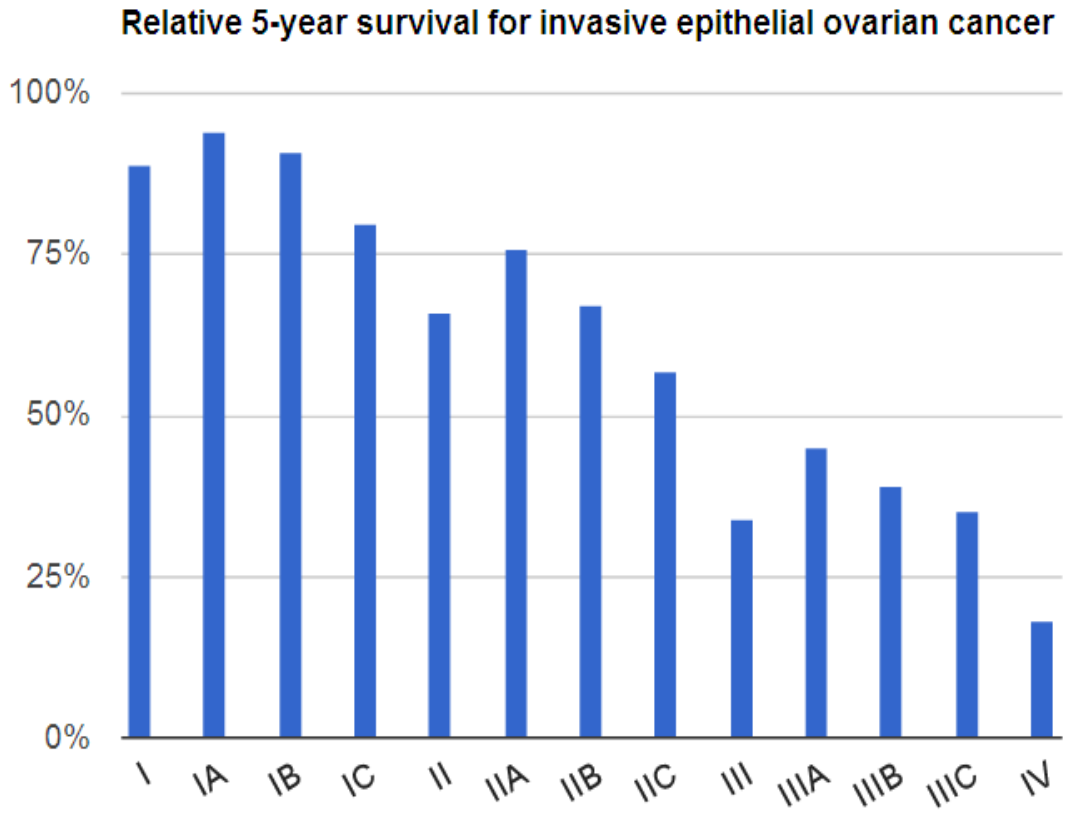
POLITECNICO MILANO 1863

ISTITUTO DI RICERCHE FARMACOLOGICHE MARIO NEGRI IRCCS

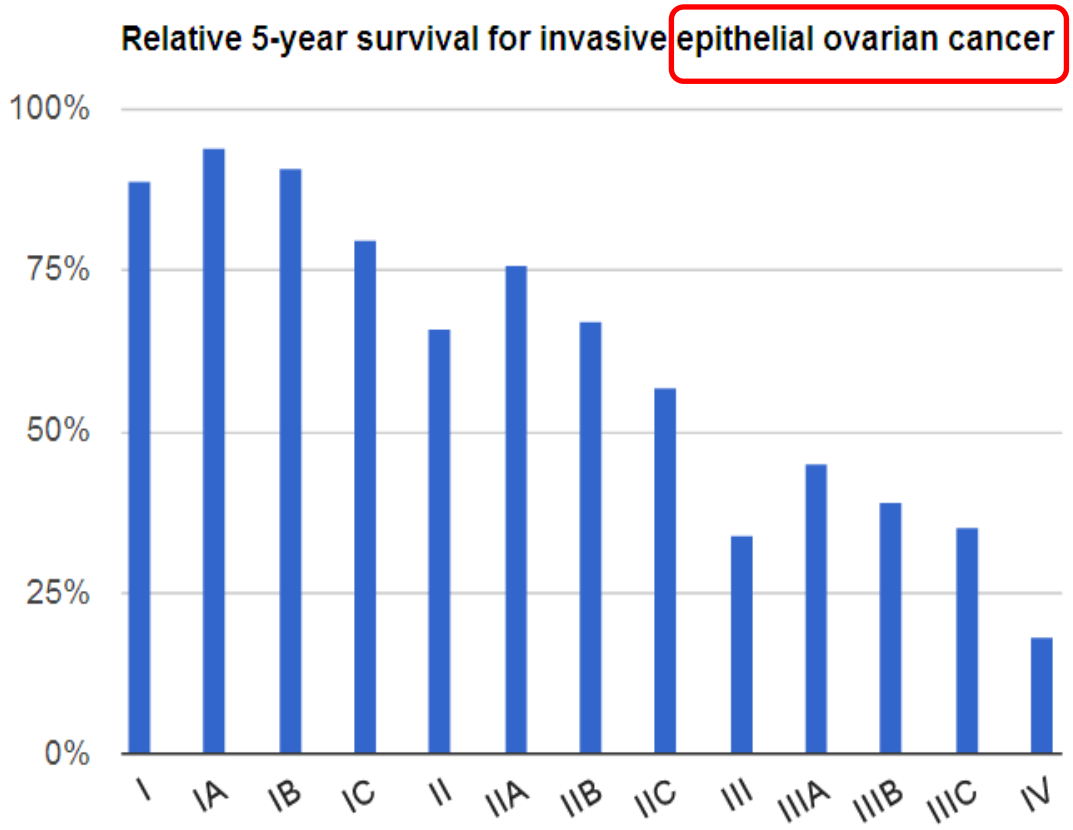
Sara Sansone	Giada Lalli	Sergio Marchini	Luca Beltrame
Computer Science and Engineering	Biomedical Engineering	Biologist	Bioinformatician



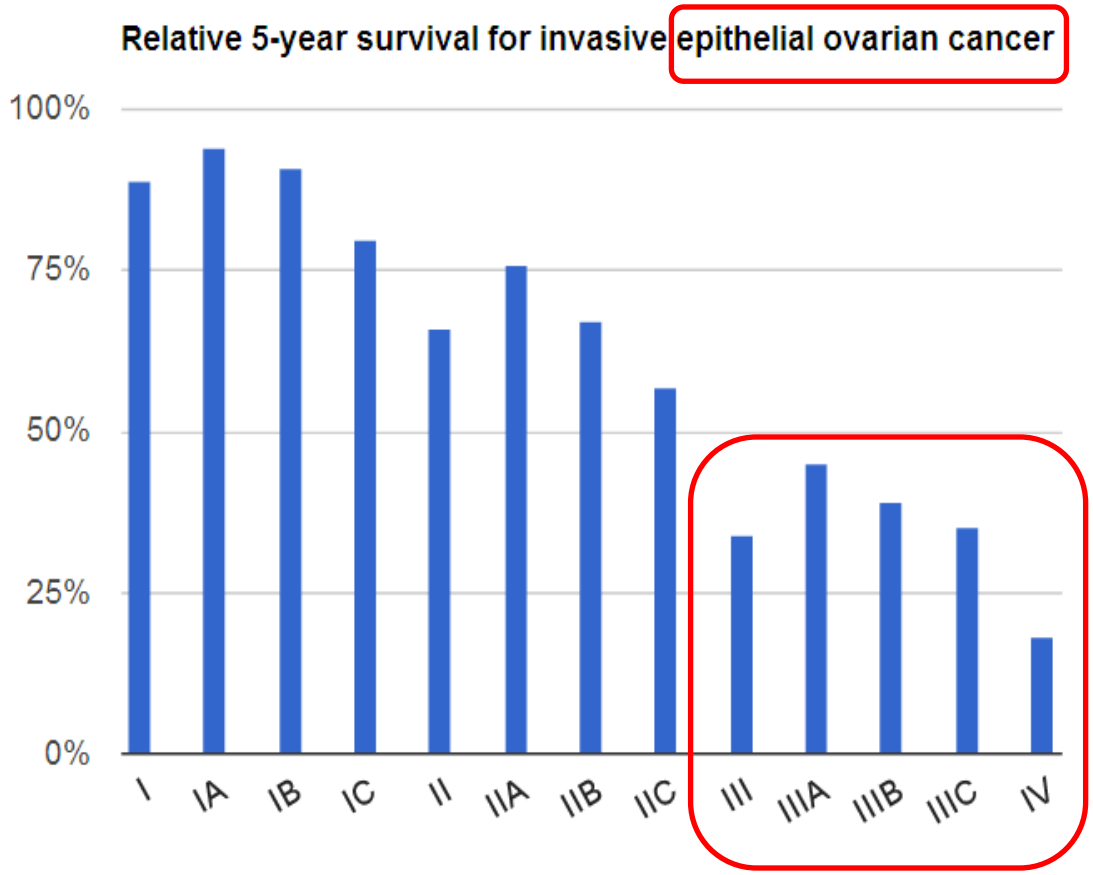
Prof. Stefano Ceri	Dr. Arif Canakoglu, Dr. Pietro Pinoli	Prof. Francesca Ieva (MOX)
Supervisor	Co-supervisors	Co-supervisor



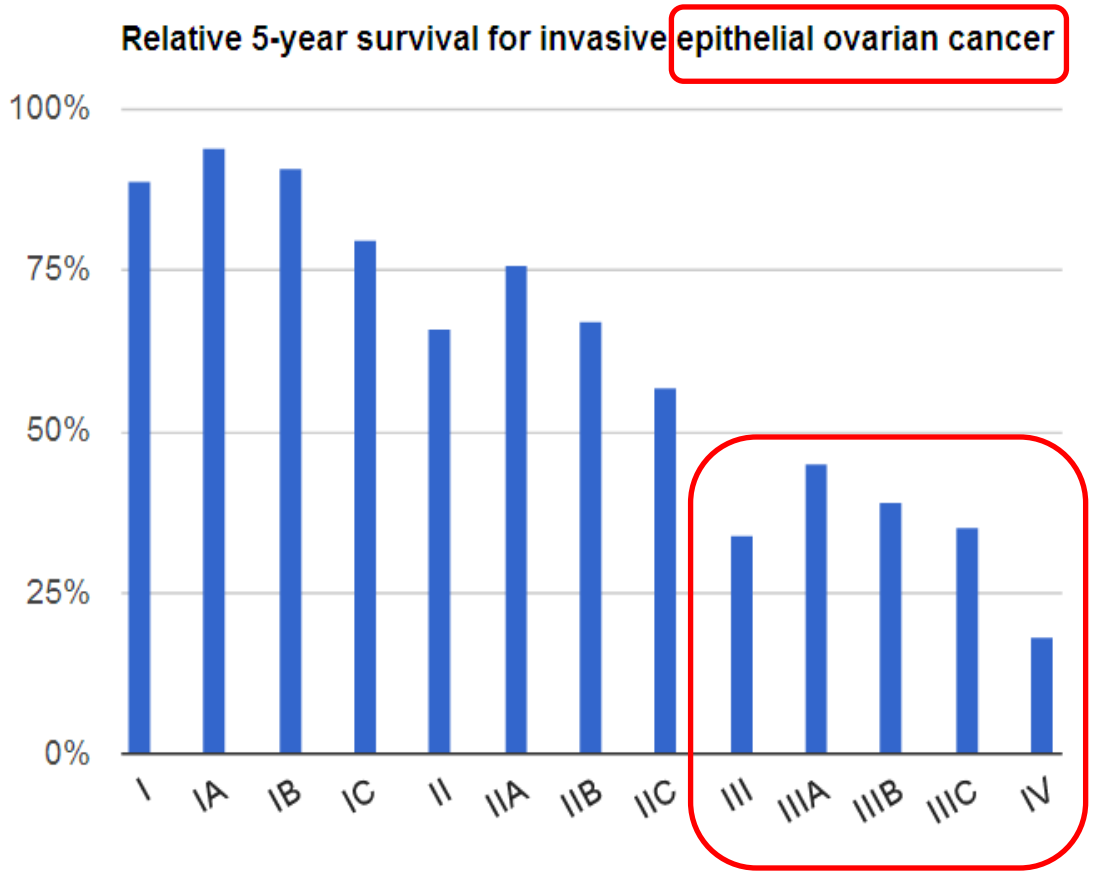
- Ovarian cancer



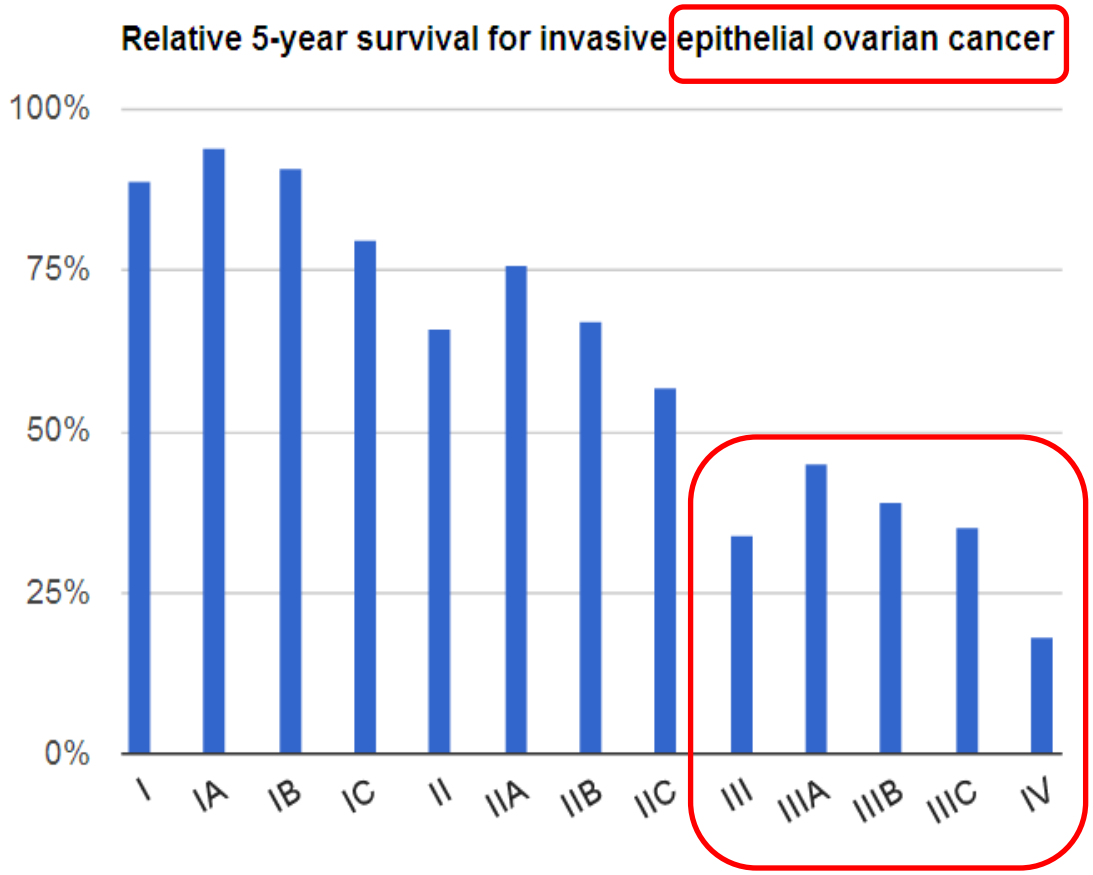
- Ovarian cancer



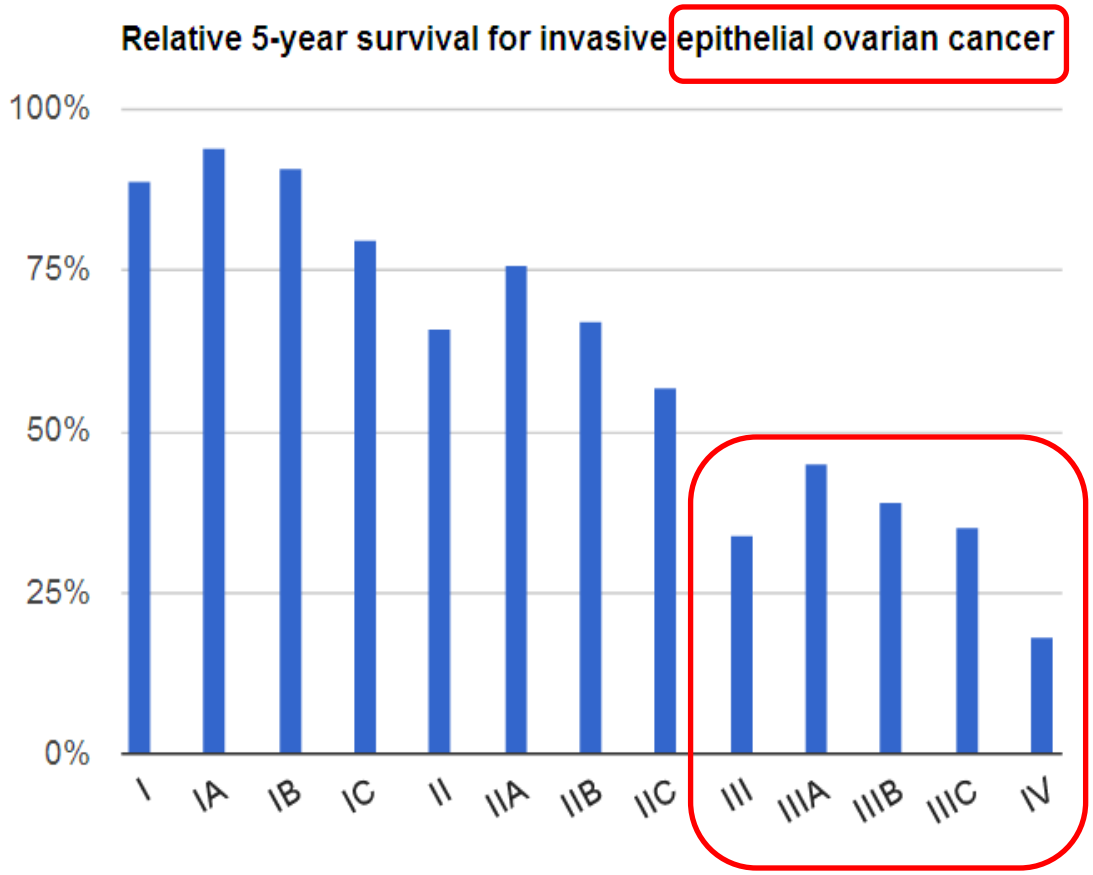
- Ovarian cancer
- High-Grade Serous Ovarian Adenocarcinoma (HGS-OC):



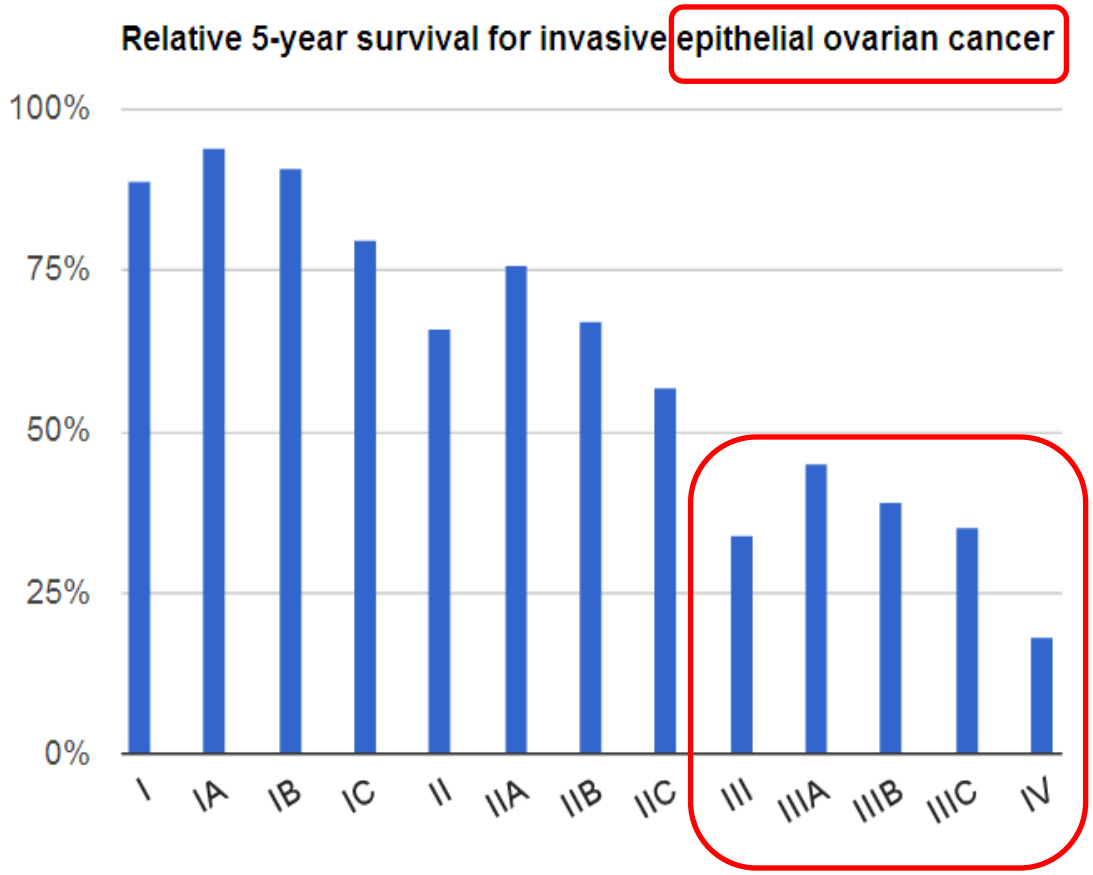
- Ovarian cancer
- High-Grade Serous Ovarian Adenocarcinoma (HGS-OC):
  - Rapidly growing carcinoma



- Ovarian cancer
- High-Grade Serous Ovarian Adenocarcinoma (HGS-OC):
  - Rapidly growing carcinoma
  - High chromosomal instability



- Ovarian cancer
- High-Grade Serous Ovarian Adenocarcinoma (HGS-OC):
  - Rapidly growing carcinoma
  - High chromosomal instability
  - All the patients have a relapse



- Ovarian cancer
- High-Grade Serous Ovarian Adenocarcinoma (HGS-OC):
  - Rapidly growing carcinoma
  - High chromosomal instability
  - All the patients have a relapse
  - They become progressively resistant to the treatment

Treatment:

Surgery and cytoreduction followed by platinum-based chemotherapy

Treatment:

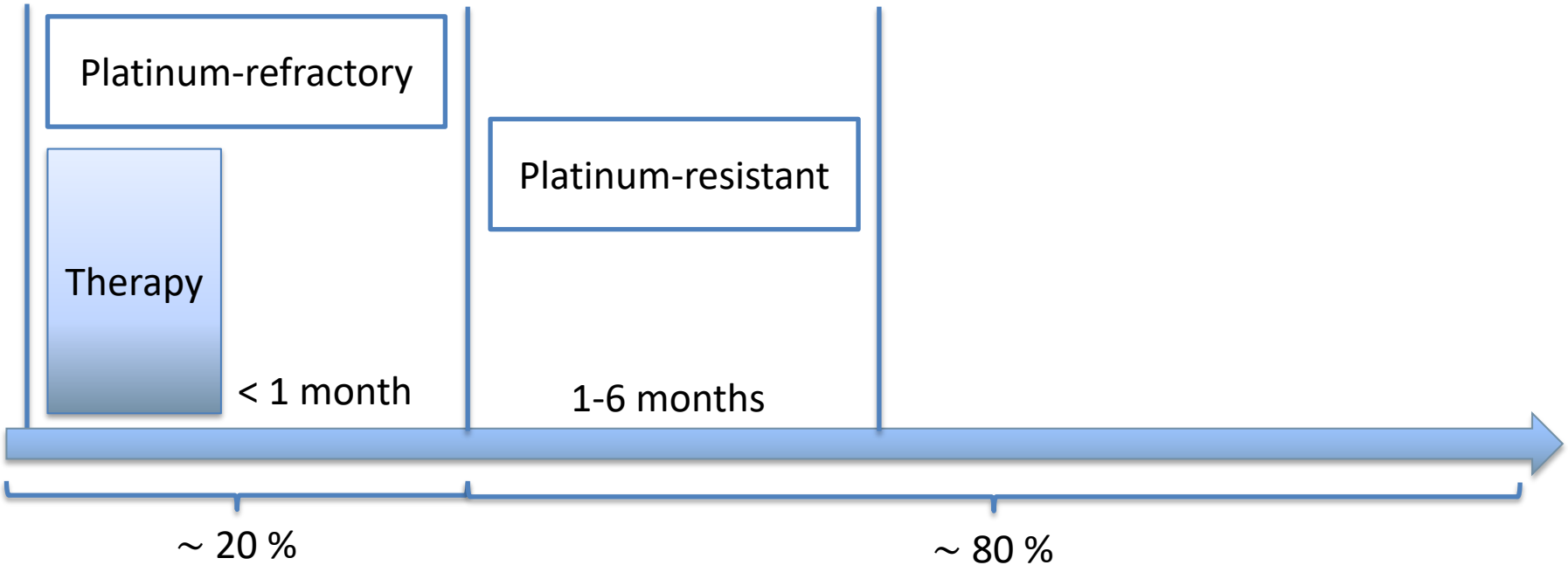
Surgery and cytoreduction followed by platinum-based chemotherapy

Patient's relapse timing:

Treatment:

Surgery and cytoreduction followed by platinum-based chemotherapy

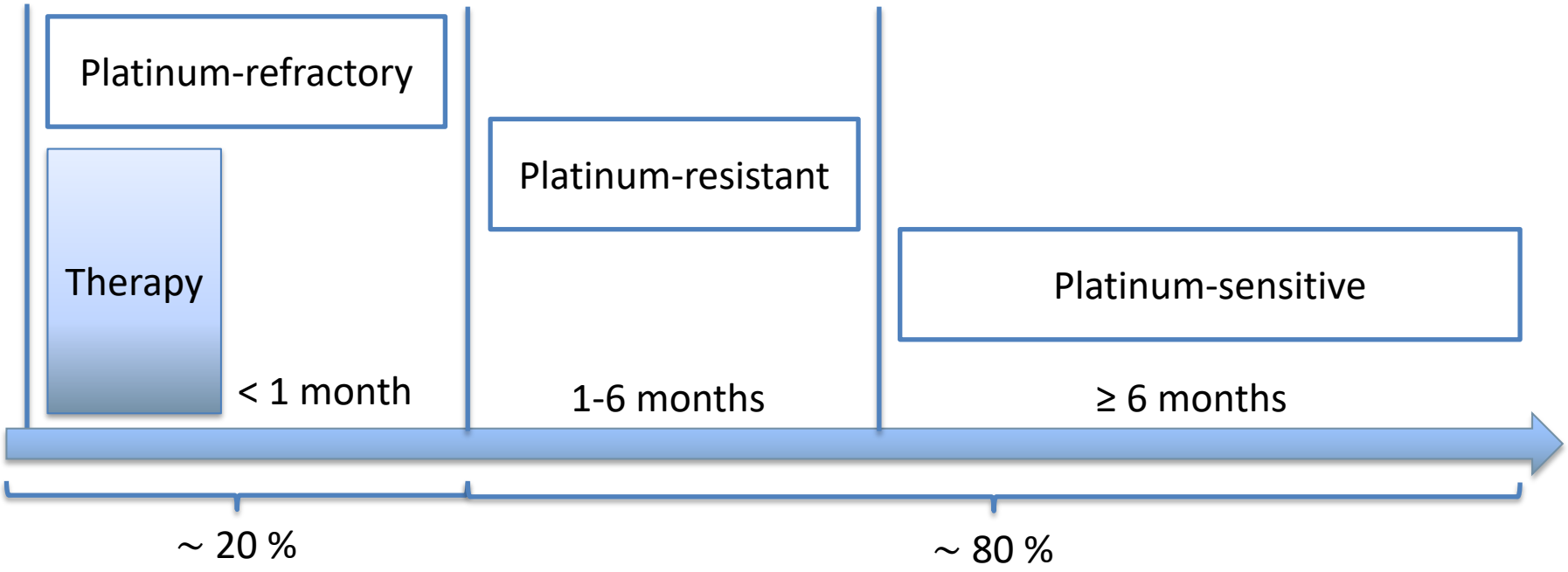
Patient's relapse timing:



Treatment:

Surgery and cytoreduction followed by platinum-based chemotherapy

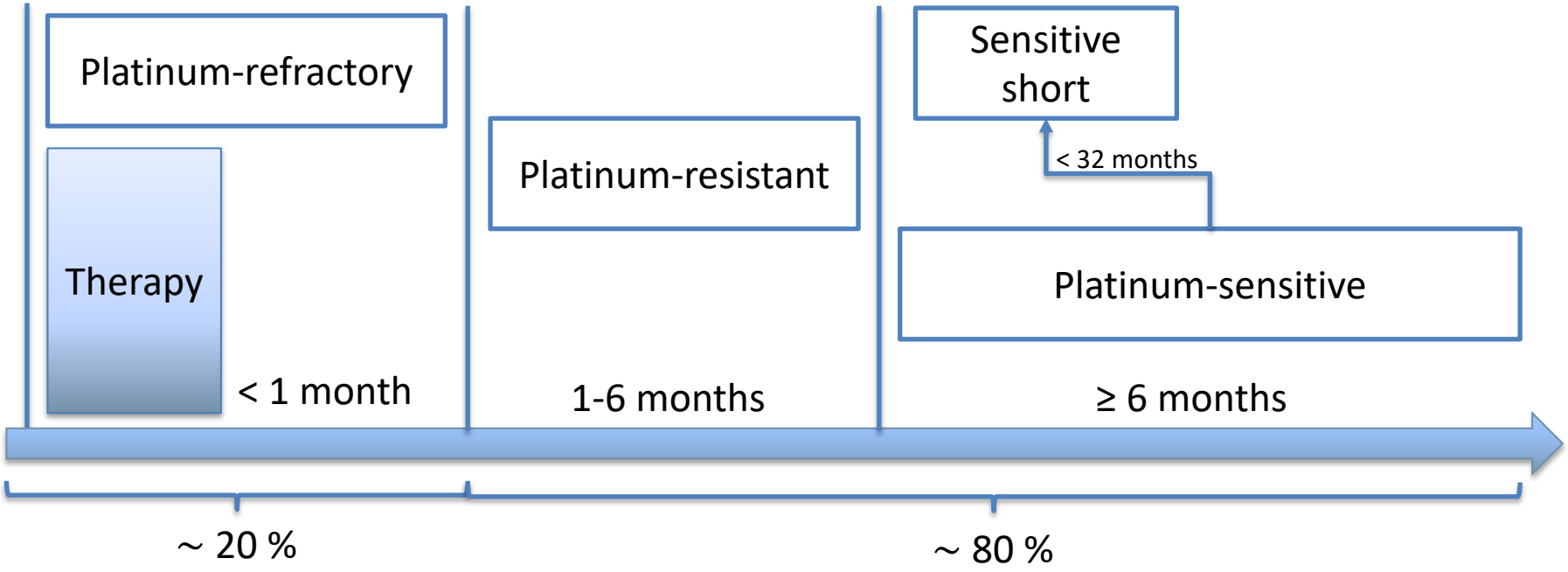
Patient's relapse timing:



Treatment:

Surgery and cytoreduction followed by platinum-based chemotherapy

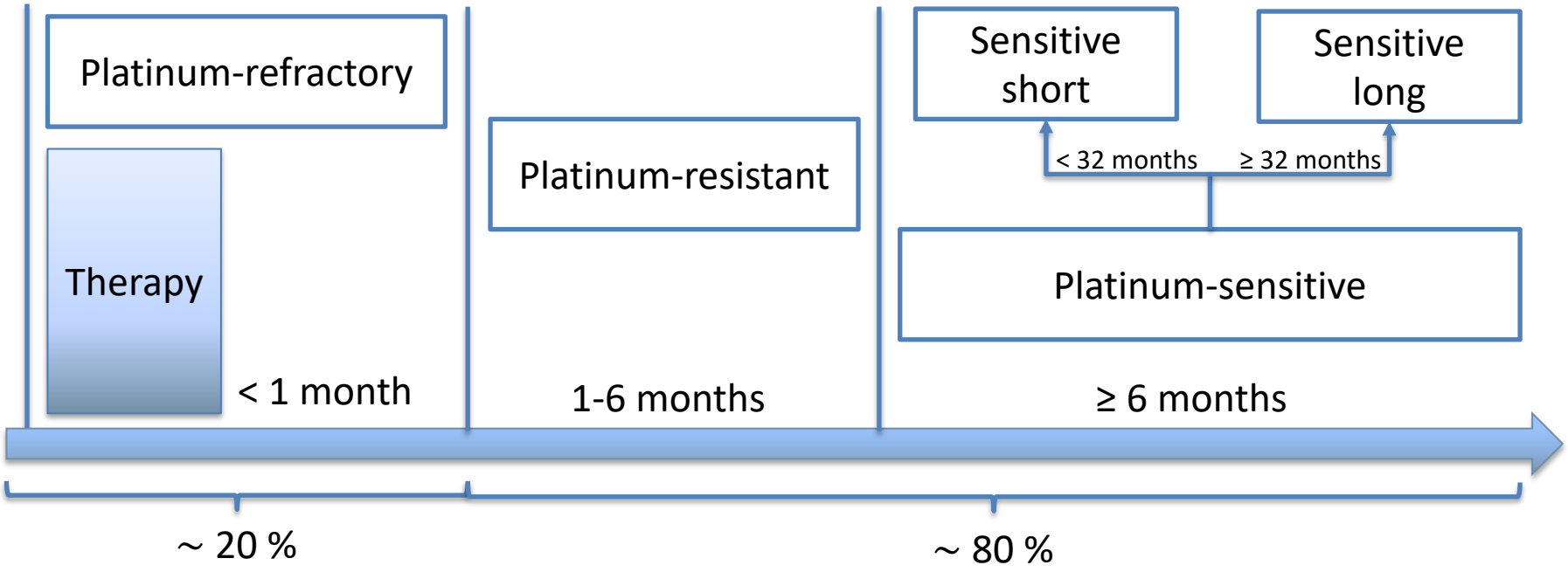
Patient's relapse timing:



Treatment:

Surgery and cytoreduction followed by platinum-based chemotherapy

Patient's relapse timing:



Exploit computational methods to identify a **molecular signature** that allows to:

Exploit computational methods to identify a **molecular signature** that allows to:

- Predict the response to therapy (resistant / sensitive)

Exploit computational methods to identify a **molecular signature** that allows to:

- Predict the response to therapy (resistant / sensitive)
- Understand the cause of chemoresistance

Exploit computational methods to identify a **molecular signature** that

Genomic regions that differ between resistant and sensitive patients

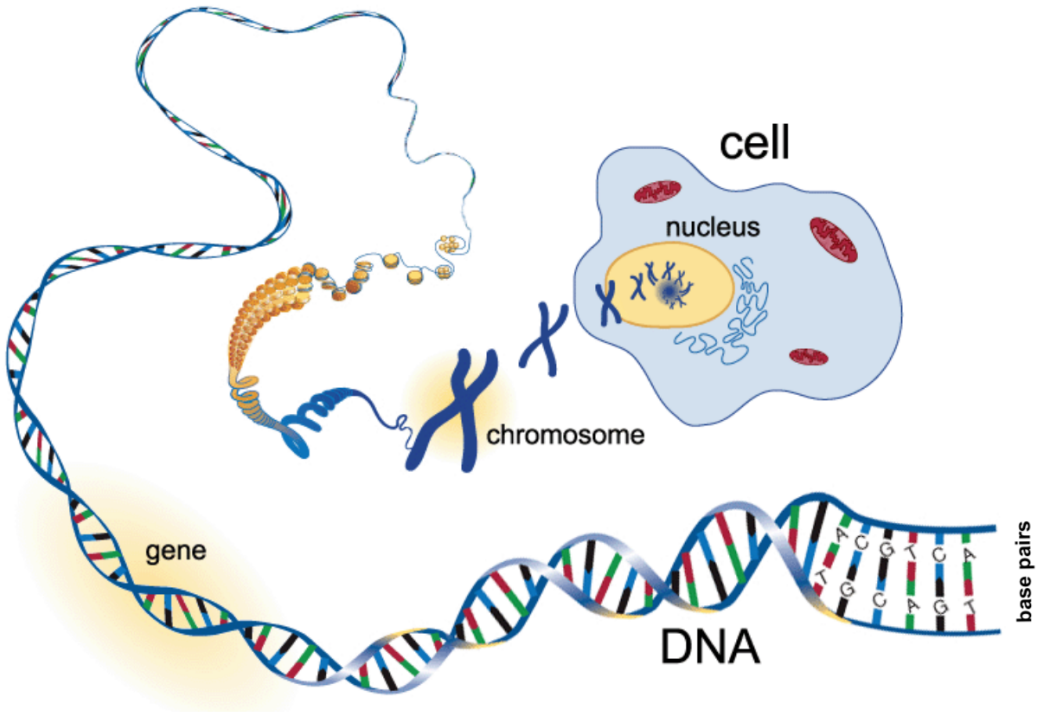
herapy (resistant / sensitive)

- Understand the cause of chemoresistance

Exploit computational methods to identify a **molecular signature** that

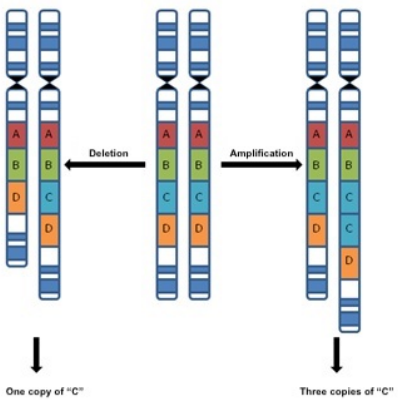
Genomic regions that differ between resistant and sensitive patients

- Understand the cause of

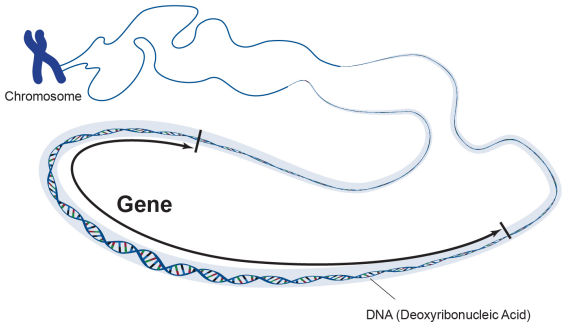


Data Description:
Genomic data used

Copy Number Alteration (CNA)

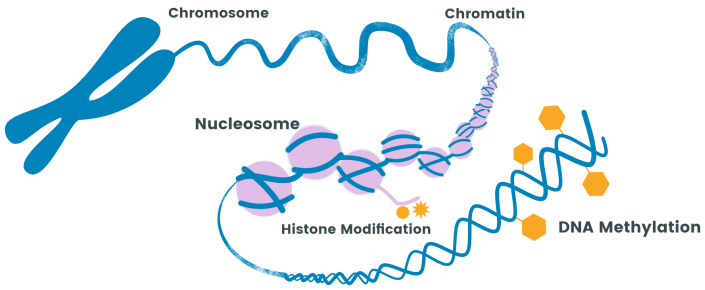


Gene expression



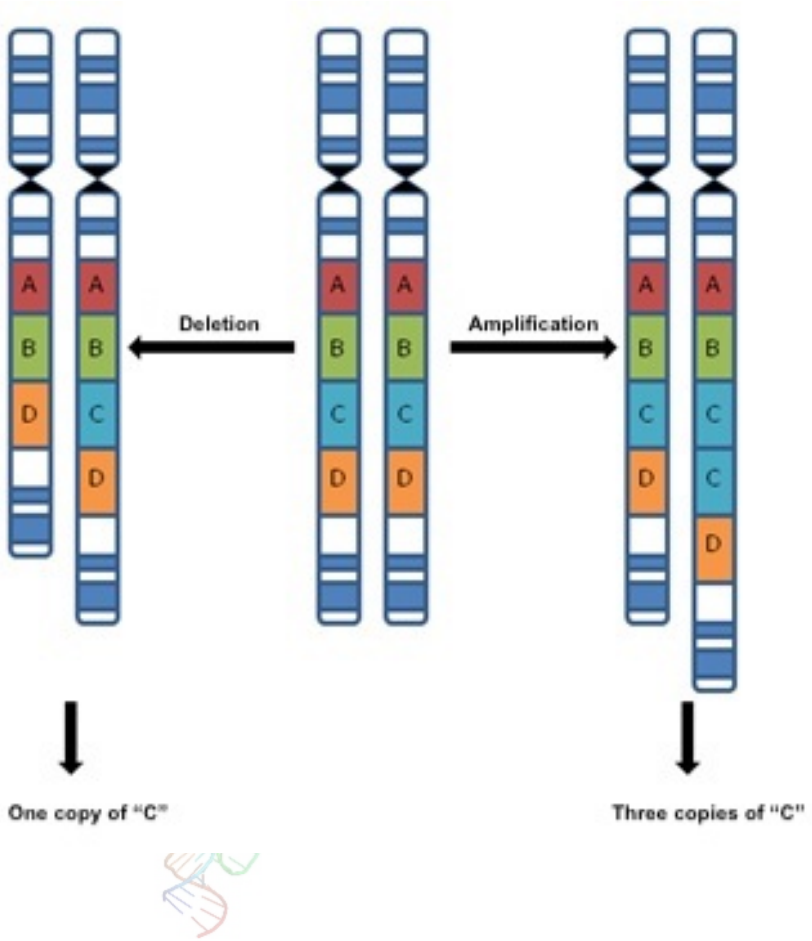
miRNA expression

DNA methylation



Data Description:
Genomic data used

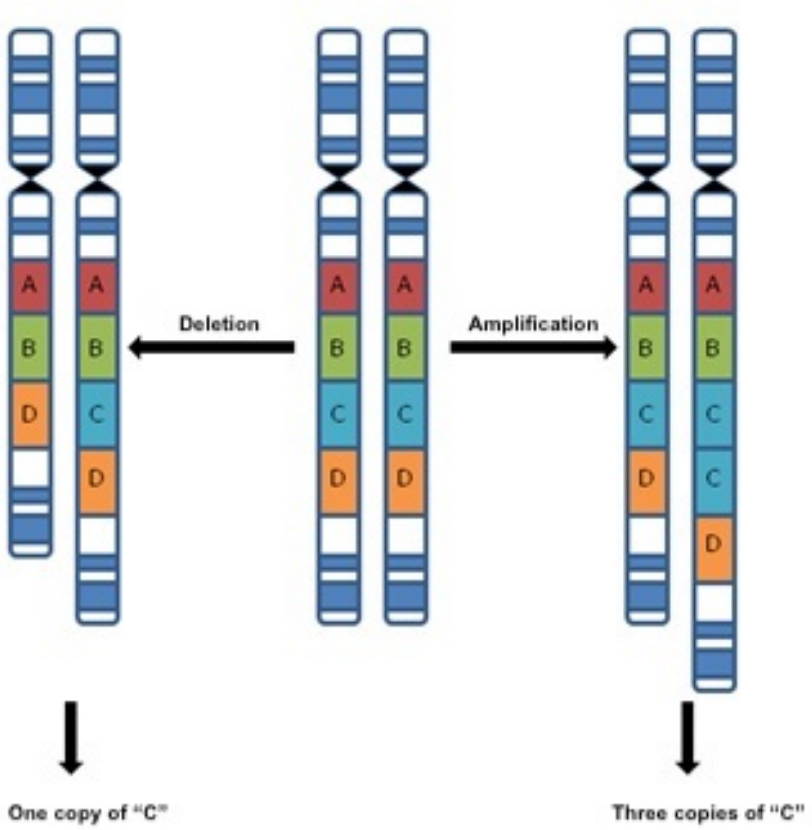
Copy Number Alteration (CNA)



Gene expression

DNA methylation

Copy Number Alteration (CNA)

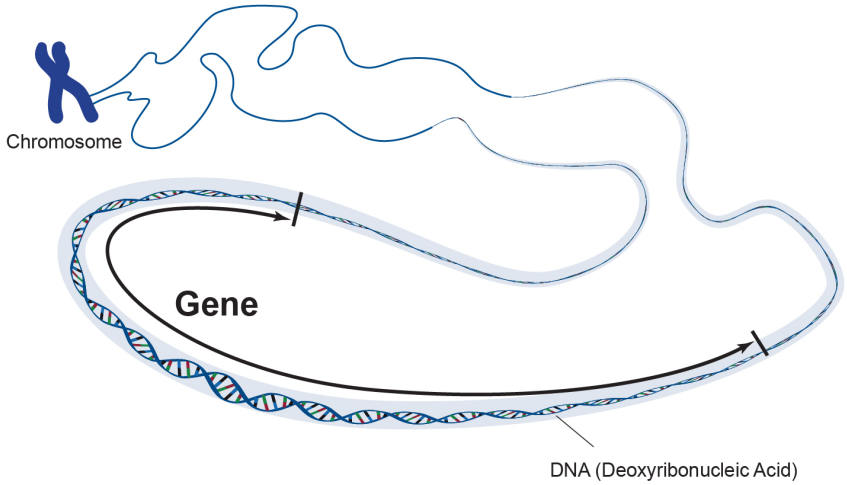


- A genomic region has normally two copies in the DNA, originating from the zygote formation
- CNAs alter this occurrence in two different ways: amplification and deletion
- The main focus is on CNA data:
  - Early events
  - May be a signal of the resistance to chemotherapy

Data Description:
Genomic data used

Copy Number Alteration (CNA)

Gene expression

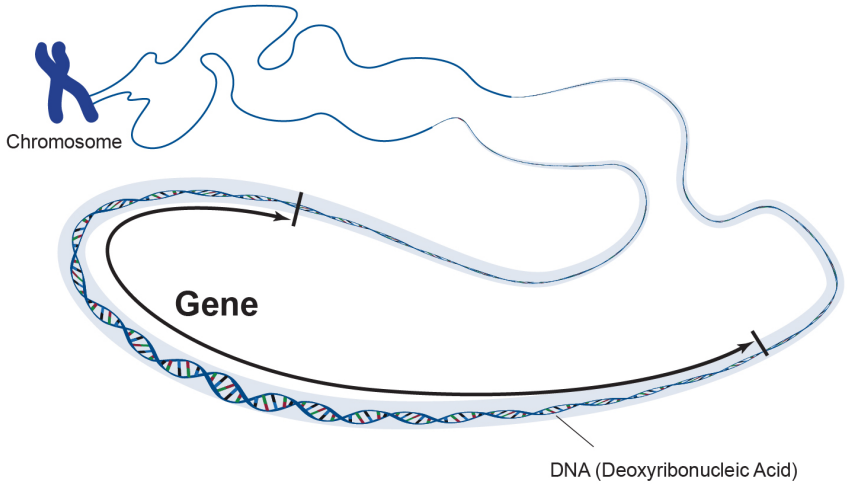


miRNA expression

DNA methylation

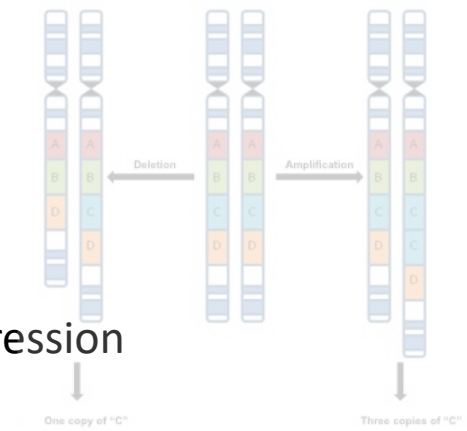
- A gene is the basic physical and functional unit of heredity
- The information encoded in the genes are used in the synthesis of functional products, such as proteins
- The process by which it is done is called gene expression
- We are mostly interested in *protein coding* genes:
  - They are related to many cellular functions and biological activities

Gene expression



Data Description:
Genomic data used

Copy Number Alteration (CNA)



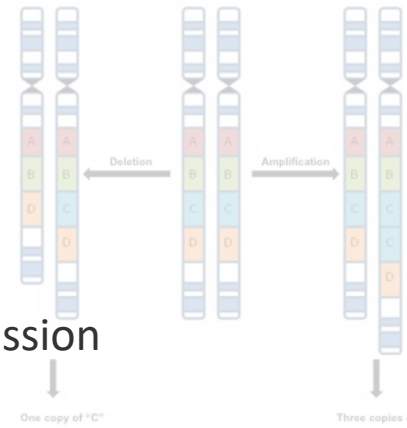
Gene expression

miRNA expression

miRNA

DNA methylation

Copy Number Alteration (CNA)



Gene expression

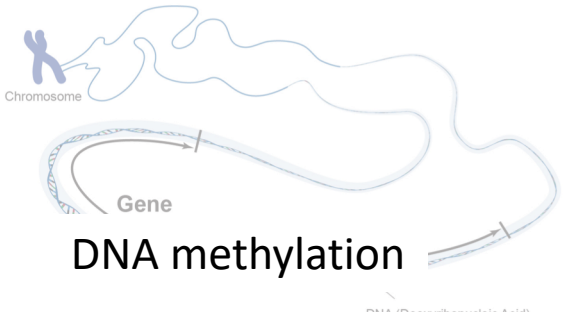
miRNA expression

- microRNAs (miRNAs) are small non-coding RNA molecules
- They target multiple genes and can either up-regulate or down-regulate their expression
- They have a causal role in tumorigenesis

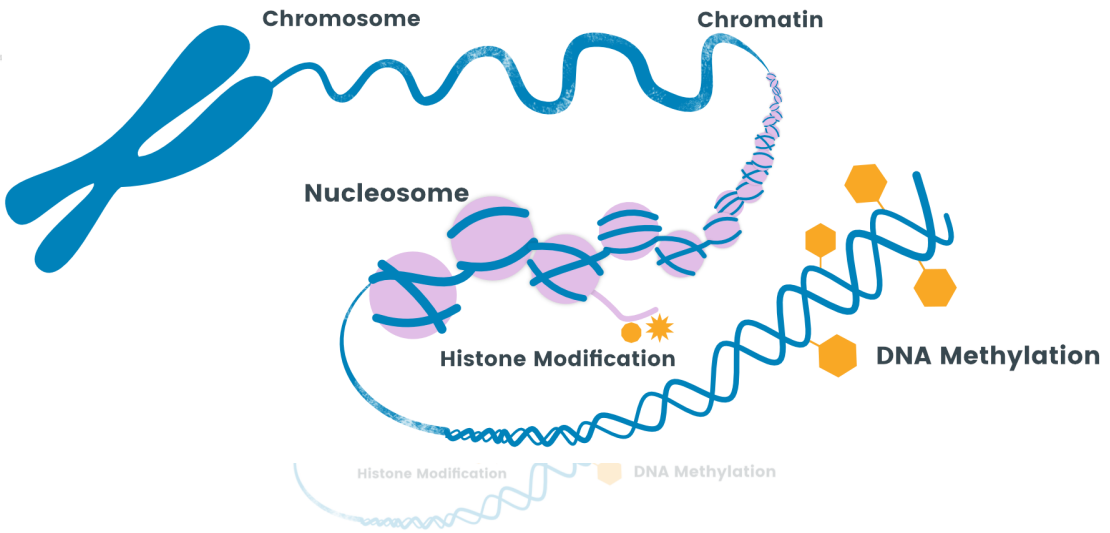
Data Description:
Genomic data used

Copy Number Alteration (CNA)

Gene expression



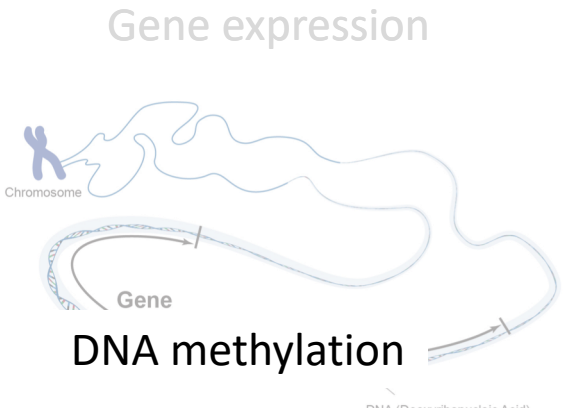
DNA methylation



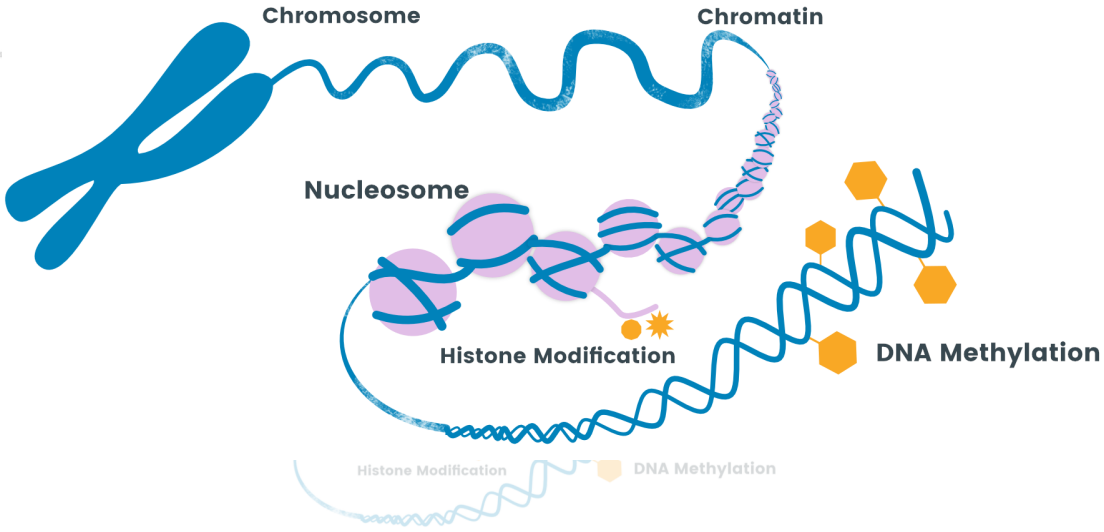
miRNA expression

Data Description:
Genomic data used

- DNA methylation is an epigenetic process by which methyl groups are added to the DNA molecule
- It can change the the function of each portion of the genome, by making it more or less accessible



DNA methylation



TCGA

THE CANCER GENOME ATLAS

National Cancer Institute
National Human Genome Research Institute

TCGA

THE CANCER GENOME ATLAS

National Cancer Institute
National Human Genome Research Institute

GMQL
GENOMETRIC QUERY LANGUAGE

TCGA

THE CANCER GENOME ATLAS

National Cancer Institute
National Human Genome Research Institute

GMQL
GENOMETRIC QUERY LANGUAGE

Obtain for each patient data about:

- CNA
- Gene expression
- miRNA
- DNA methylation

- **CNA**

patient	chrom	start	stop	num_mark	seg_mean
R0_TCGA-13-0720	chr1	3301764	16104539	7169	0.2480
R0_TCGA-13-0720	chr1	16108231	16162328	29	0.7084

Segmented mean: the log_2 ratio of observed intensity of alteration over reference intensity

- **CNA**

patient	chrom	start	stop	num_mark	seg_mean
R0_TCGA-13-0720	chr1	3301764	16104539	7169	0.2480
R0_TCGA-13-0720	chr1	16108231	16162328	29	0.7084

Segmented mean: the log_2 ratio of observed intensity of alteration over reference intensity

- **Gene expression**

patient	chrom	start	stop	gene_symbol	fpkm
R0_TCGA-13-0720	chr1	11868	14409	DDX11L1	0.000000
R0_TCGA-13-0720	chr1	14403	29570	WASH7P	23648.321087

FPKM (Fragments Per Kilobase Million): the value of expression, normalized for sequencing depth and gene length

- **miRNA expression**

patient	chrom	start	stop	mirna_id	rpm
R0_TCGA-13-0720	chr1	17368	17436	hsa-mir-6859-1	0.000000
R0_TCGA-13-0720	chr1	30365	30503	hsa-mir-1302-2	0.000000

RPM (Reads Per Million): the value of expression, normalized for sequencing depth

- **miRNA expression**

patient	chrom	start	stop	mirna_id	rpm
R0_TCGA-13-0720	chr1	17368	17436	hsa-mir-6859-1	0.000000
R0_TCGA-13-0720	chr1	30365	30503	hsa-mir-1302-2	0.000000

RPM (Reads Per Million): the value of expression, normalized for sequencing depth

- **DNA methylation**

patient	chrom	start	stop	gene_symbol	beta_value
R0_TCGA-13-0720	chr1	924804	924806	SAMD11	0.009892
R0_TCGA-13-0720	chr1	925936	925938	SAMD11	0.007828

Beta value: the ratio of intensities between methylated and unmethylated alleles

POLITECNICO
MILANO 1863

HONOURS PROGRAMME HP-SR
in Information Technology

First approach to solve the problem:

Use only CNA data

Steps performed

1. Data preprocessing

2. Feature selection

3. Methods: Classification vs Survival Regression

Problem

- A genome wide analysis is needed to identify regions with different CNA between the classes

Solution

Problem

- A genome wide analysis is needed to identify regions with different CNA between the classes

Solution

- We create, for each patient, two **CNA profiles** (for amplification alteration and for deletion alteration)

Problem

- A genome wide analysis is needed to identify regions with different CNA between the classes
- The genome contains 3 billions of base pairs

Solution

- We create, for each patient, two **CNA profiles** (for amplification alteration and for deletion alteration)

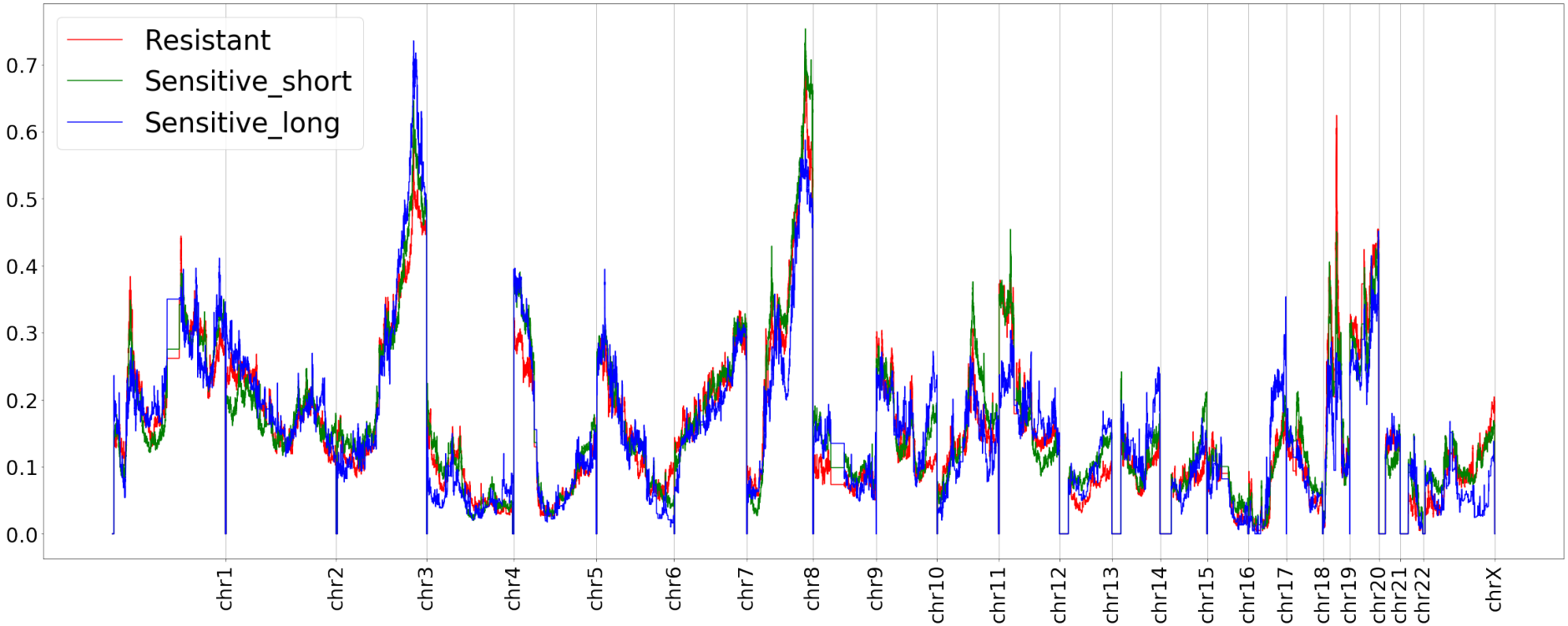
Problem

- A genome wide analysis is needed to identify regions with different CNA between the classes
- The genome contains 3 billions of base pairs

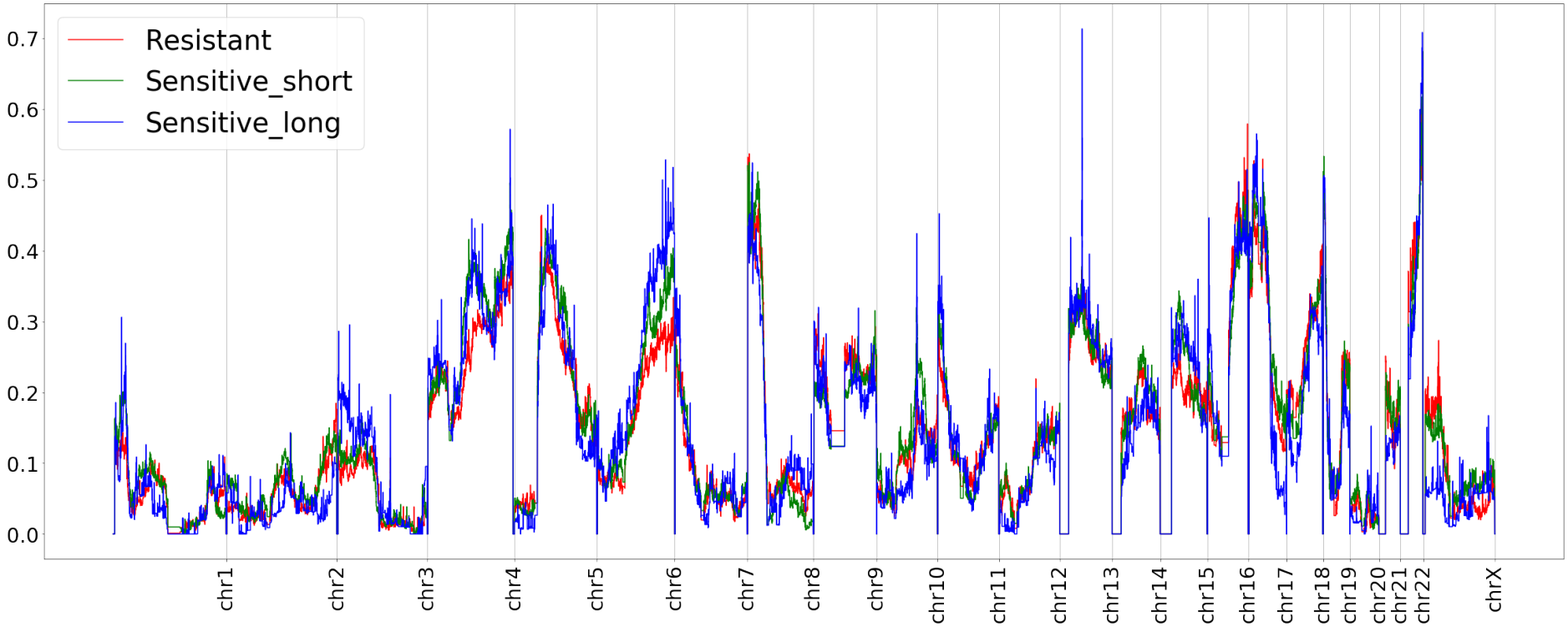
Solution

- We create, for each patient, two **CNA profiles** (for amplification alteration and for deletion alteration)
- We use bins of size n, i.e., we consider one position as the average of the values of n positions

Amplification profiles, resolution of 10Kb



Deletion profiles, resolution of 10Kb



Steps performed

1. Data preprocessing

2. Feature selection

3. Methods: Classification vs Survival Regression

We tried two different approaches to extract relevant CNA regions:

We tried two different approaches to extract relevant CNA regions:

1. Use **GISTIC2.0**, the state-of-the-art for CNA analysis

   GISTIC2.0 is a module able to find regions of the genome that are significantly amplified or deleted in a certain population

We tried two different approaches to extract relevant CNA regions:

1. Use **GISTIC2.0**, the state-of-the-art for CNA analysis

   GISTIC2.0 is a module able to find regions of the genome that are significantly amplified or deleted in a certain population

2. Compare **CNA profiles** of patients of different classes and compute the p-values for the regions using statistical tests:
   - Search for the more suitable test
   - Implementation of a permutation test
   - Use two different thresholds to select the p-values: 0.05, 0.005

Steps performed

1. Data preprocessing

2. Feature selection

3. Methods: Classification vs Survival Regression

1. Choose the most suitable classification algorithm

2. Choose the best set of features

3. Evaluate the model

- We tried different classification algorithms
- The ones giving the best performances were:
  - **KNN**, when using features from GISTIC2.0
  - **SVM**, in all the other cases

1. Choose the most suitable classification algorithm

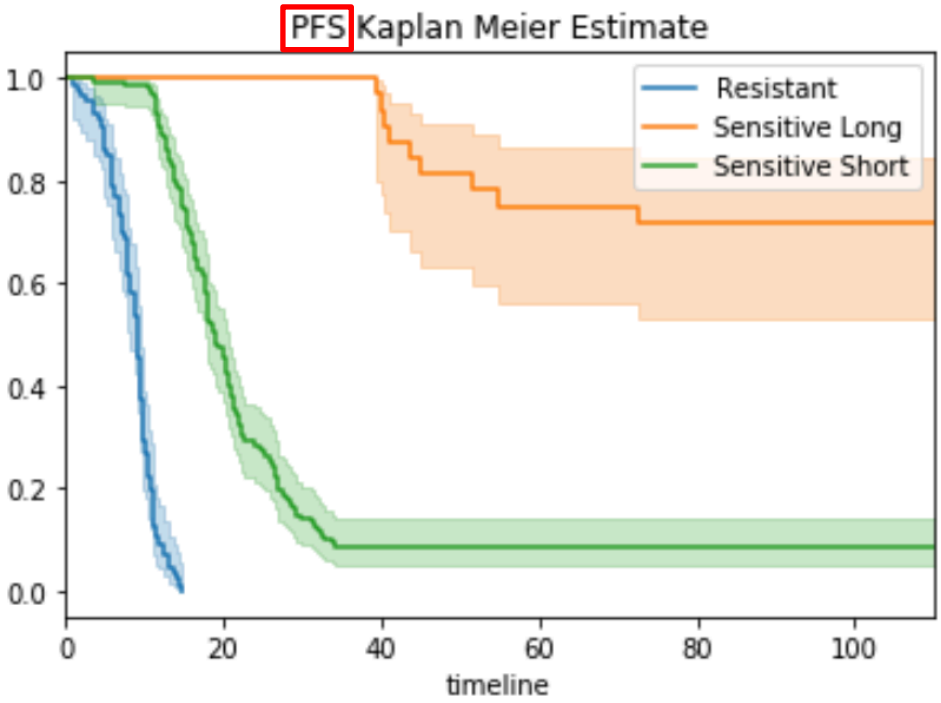
2. Choose the best set of features

3. Evaluate the model

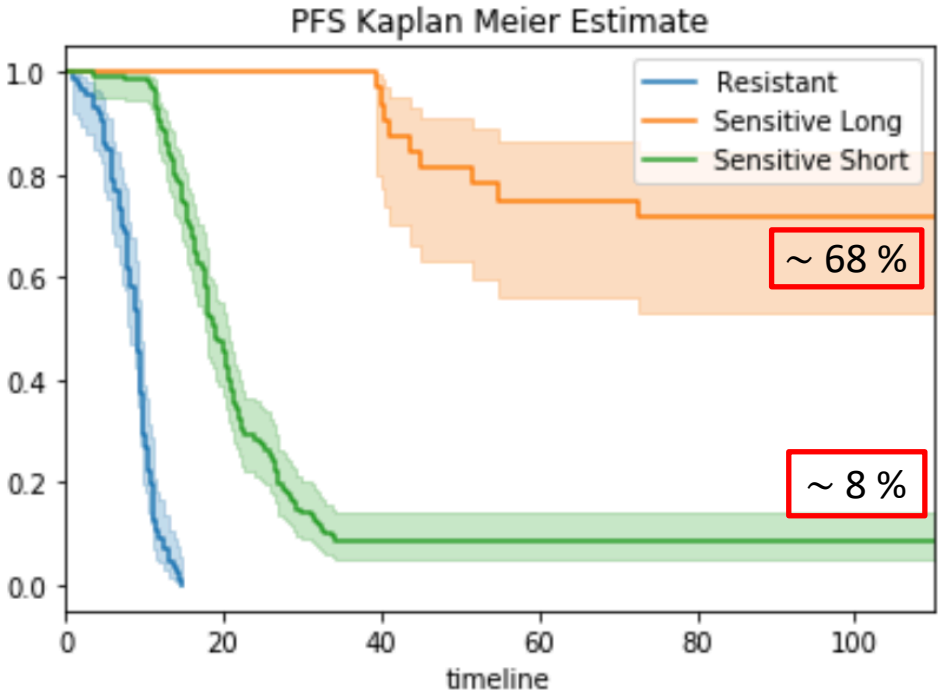
- Take the set of features obtained with the different features selection methods
- Compute for each of them precision, recall, accuracy and AUC through a 10-fold cross validation
- Select the features giving the best performances

1. Choose the most suitable classification algorithm
2. Choose the best set of features
3. Evaluate the model

- We did not achieve good results
- The best performances obtained for Resistant vs Sensitive were:
  - Average precision: 0.51 ± 0.10
  - Average recall: 0.61 ± 0.19
  - Average accuracy: 0.68 ± 0.07
  - Average AUC: 0.72 ± 0.11

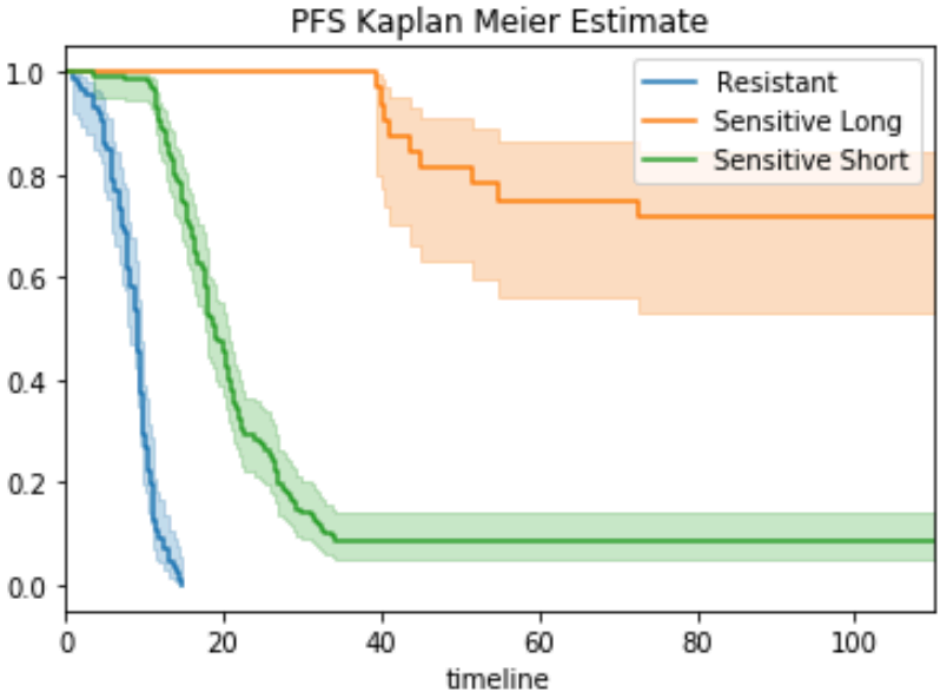


- *Progression Free Survival (PFS)*: the interval from the date of surgery to the date of progression, date of recurrence, or date of last known contact



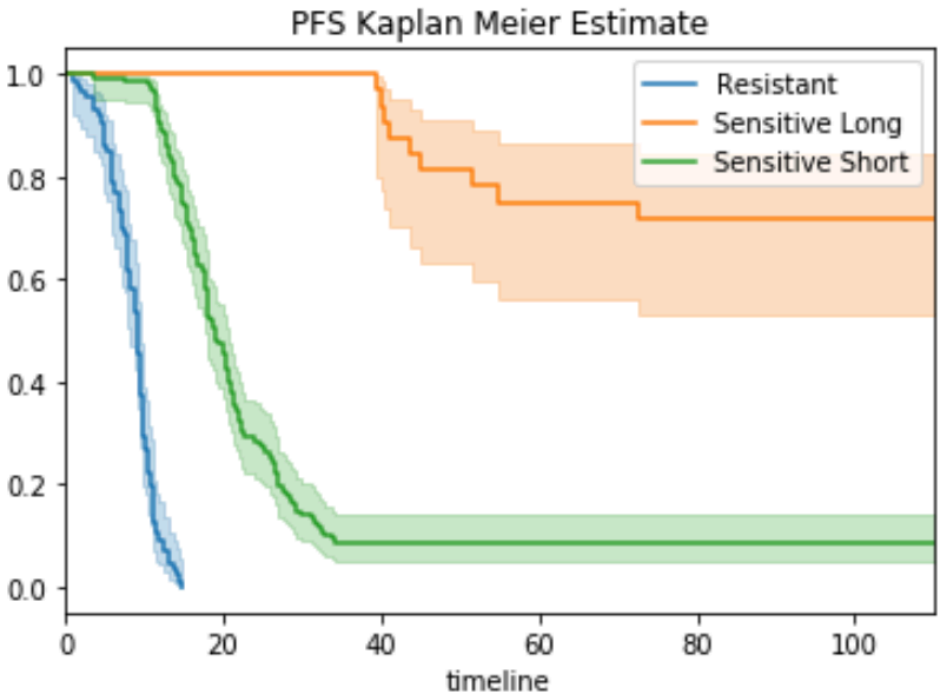
- *Progression Free Survival (PFS)*: the interval from the date of surgery to the date of progression, date of recurrence, or date of last known contact

- Censored data: patients who did not have the relapse up to the last contact



- *Progression Free Survival (PFS)*: the interval from the date of surgery to the date of progression, date of recurrence, or date of last known contact
- Censored data: patients who did not have the relapse up to the last contact
- How to predict *PFS*?
  - **Cox Regression Model**

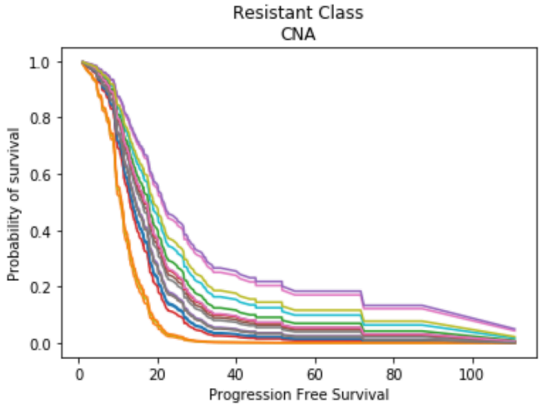
Methods:
Survival Regression

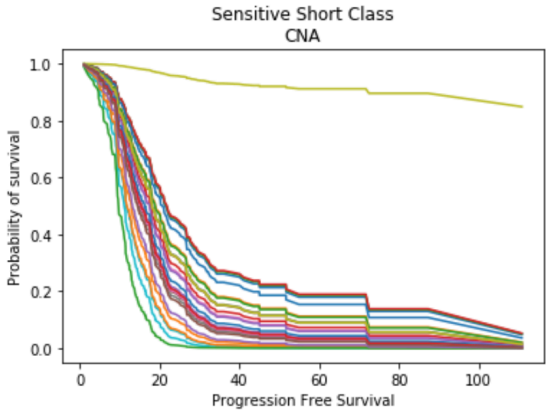


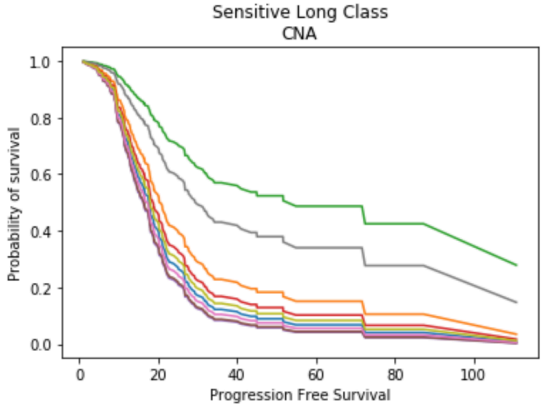
- *Progression Free Survival (PFS)*: the interval from the date of surgery to the date of progression, date of recurrence, or date of last known contact
- Censored data: patients who did not have the relapse up to the last contact
- How to predict *PFS*?
  - **Cox Regression Model**
- What features did we use?
  - The ones obtained through the **permutation test**

- We were not able to correctly predict the PFS times of the patients

- The best *concordance index* we got was equal to 0.58







POLITECNICO
MILANO 1863

HONOURS PROGRAMME HP-SR
in Information Technology

Second approach to solve the problem:

Use four types of genomic data

Steps performed

1. Feature selection for the other three types of data

2. Classification

- Compute the p-values, for the different genomic elements, using **Mann-Whitney** test (for each binary comparison)

- Compute the p-values, for the different genomic elements, using **Mann-Whitney** test (for each binary comparison)

- Try different thresholds for the p-values: 0.05, 0.005, 0.0005

- Compute the p-values, for the different genomic elements, using **Mann-Whitney** test (for each binary comparison)

- Try different thresholds for the p-values: 0.05, 0.005, 0.0005

- Try different correction for multiple testing:
  - *Bonferroni correction:*

$$p\_value_{corrected} = p_{value} \cdot n\_tests$$

  - *Benjamini-Hochberg correction*:

$$p\_value_{corrected} = p_{values} \cdot \frac{n\_tests}{ranking}$$

- Standard version: n_tests = total number of tests
- Mild version: n_tests = number of patients of the two classes

Steps performed

1. Feature selection for the other three types of data

2. Classification

Methods:
Classification with four types of genomic data

Patient_id	Amp:chr1:2000-2999	Del:chr4:37852-38402
R_00000		
R_00001		
R_00002		

Patient_id	ENSG00000223972.5	ENSG00000227232.5
R_00000		
R_00001		
R_00002		

Patient_id	hsa-mir-6859-1	hsa-mir-1302-2
R_00000		
R_00001		
R_00002		

Patient_id	SAMD11	GRID2
R_00000		
R_00001		
R_00002		

Select the best features for each type of genomic data

Methods:
Classification with four types of genomic data

Patient_id	Amp:chr1:2000-2999	Del:chr4:37852-38402
R_00000		
R_00001		
R_00002		

Patient_id	ENSG00000223972.5	ENSG00000227232.5
R_00000		
R_00001		
R_00002		

Patient_id	hsa-mir-6859-1	hsa-mir-1302-2
R_00000		
R_00001		
R_00002		

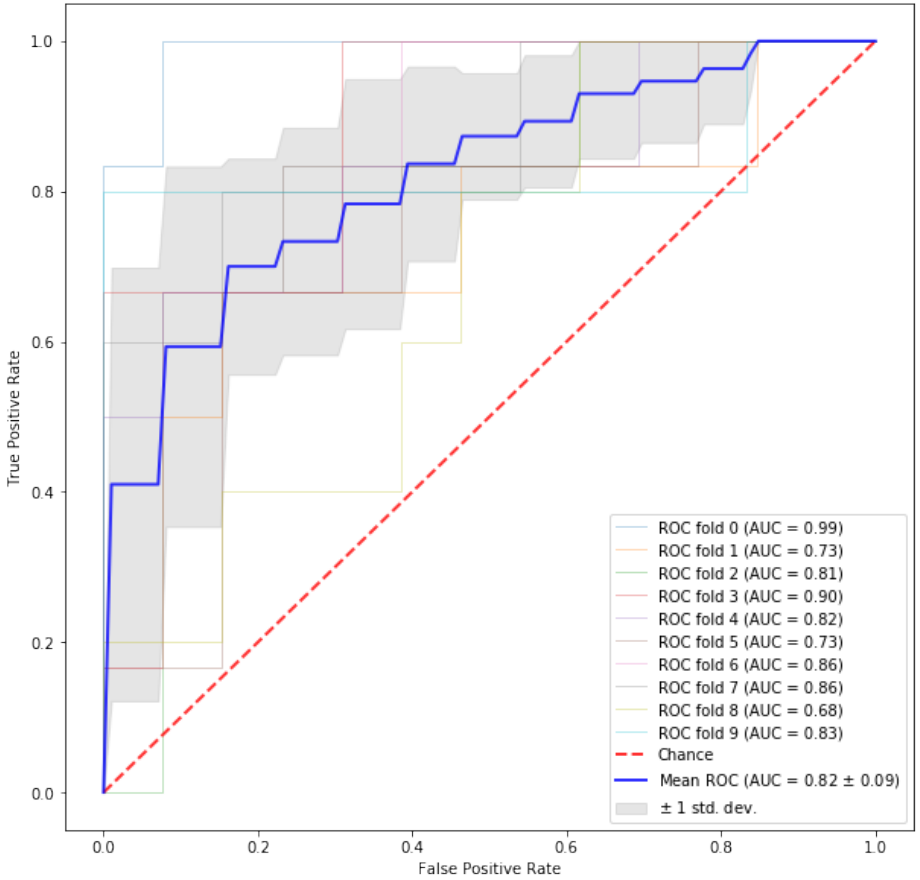
Patient_id	SAMD11	GRID2
R_00000		
R_00001		
R_00002		

Patient_id	Amp:chr1:2000-2999	Del:chr4:37852-38402	ENSG00000223972.5	ENSG00000227232.5	hsa-mir-6859-1	hsa-mir-1302-2	SAMD11	GRID2
R_00000								
R_00001								
R_00002								

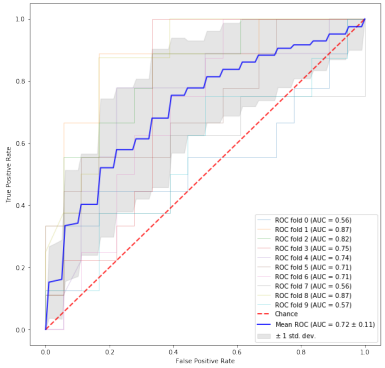
Select the best features for each type of genomic data → Merge the four datasets → Normalize → Classify using SVM

Best computational results:
ROC curves for Resistant vs Sensitive

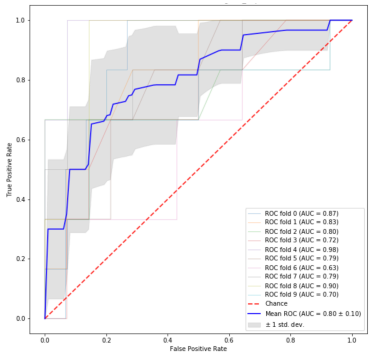
Merging of all genomic data



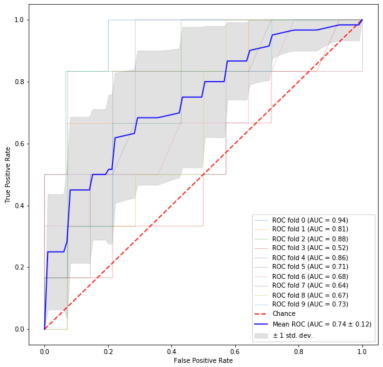
CNA



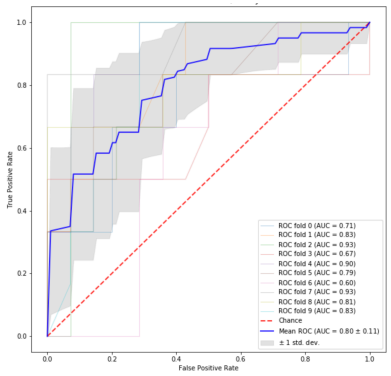
Gene expression



miRNA



Methylation



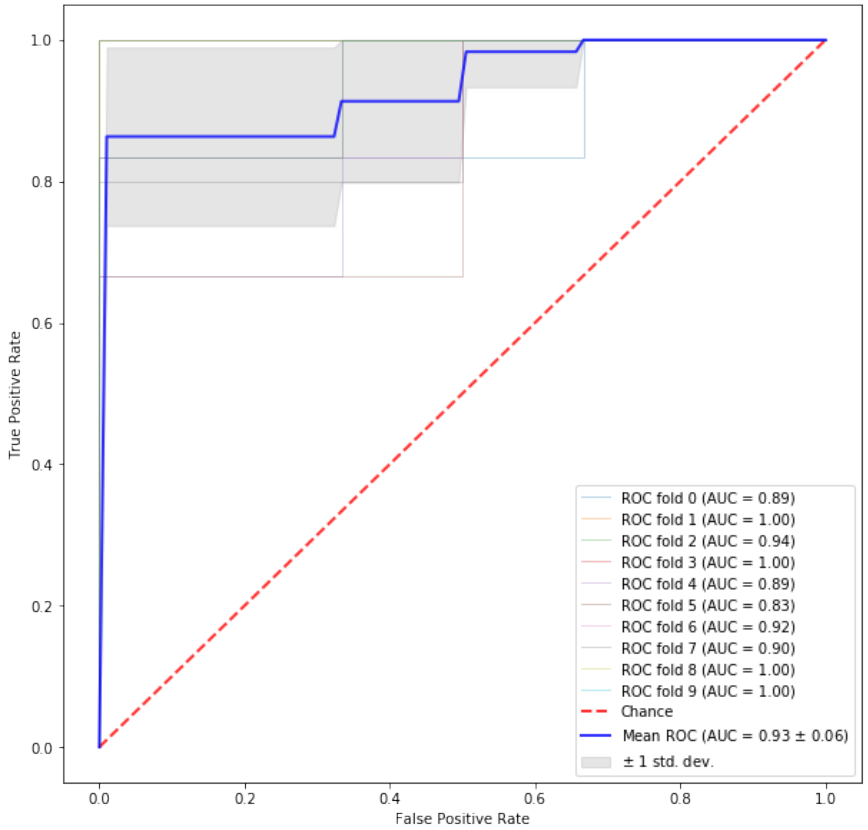
Best computational results:
For Resistant vs Sensitive

Type of data	N features	Precision	Recall	Accuracy	AUC
CNA	225	0.51 ± 0.10	0.61 ± 0.19	0.68 ± 0.07	0.72 ± 0.11
Gene expression	20	0.71 ± 0.20	0.37 ± 0.10	0.77 ± 0.10	0.79 ± 0.11
miRNA	11	0.77 ± 0.30	0.37 ± 0.20	0.75 ± 0.10	0.72 ± 0.15
Methylation	65	0.79 ± 0.30	0.35 ± 0.10	0.78 ± 0.10	0.78 ± 0.09
Merge	311	0.68 ± 0.18	0.74 ± 0.11	0.80 ± 0.10	0.82 ± 0.09

- A single genomic data is not enough to distinguish the two main classes: *resistant* and *sensitive*
- Four genomic signals together allow to achieve good performances ⇒ the recall is significantly better

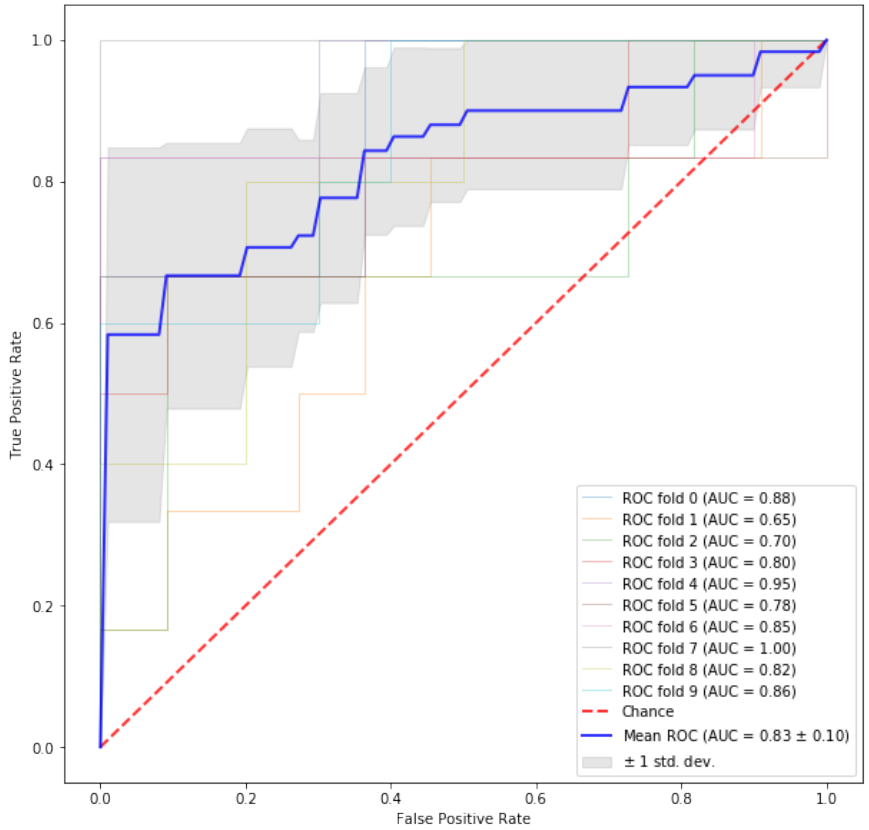
Best computational results:
ROC curves for the other binary comparisons

All genomic data



Resistant vs Sensitive Long

All genomic data



Resistant vs Sensitive Short

- The method is satisfying: it allows to achieve good results for all the performance measures, i.e., precision, recall, accuracy and AUC of the ROC curves

- **Innovation**: use four different genomic data-types and be able to classify the patients with good performances

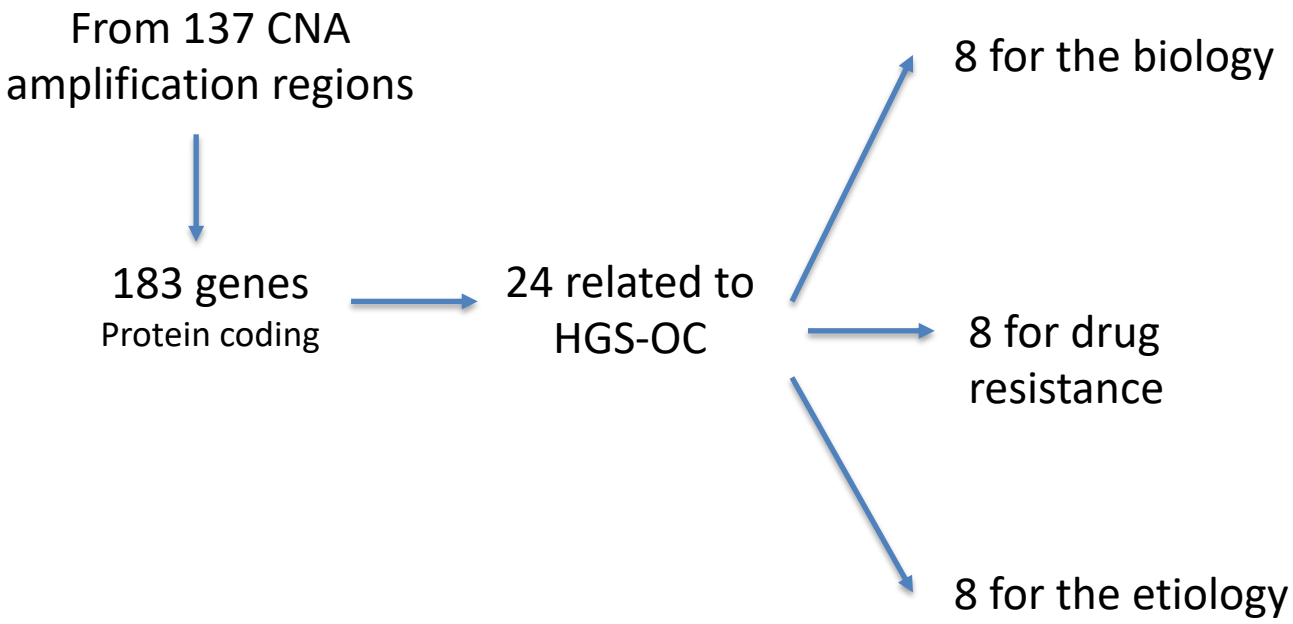
Biological Results:
Relevant features for Resistant vs Sensitive

From 137 CNA amplification regions

↓

183 genes
Protein coding → 24 related to HGS-OC

Relevant features for Resistant vs Sensitive



Biological Results:
Relevant features for Resistant vs Sensitive

From 137 CNA amplification regions

183 genes
Protein coding

24 related to HGS-OC

8 for the biology

8 for drug resistance

8 for the etiology

2 activators of the **Notch pathway** (DLL1 e CTBP2)

5 related to **resistance to platinum-based chemotherapy** (CA9, ELAV1, HTRA1, RNASET2, URI1)

1 related to the **response to chemotherapy** (BID)

- We further analyzed the 8 genes related to drug-resistance

- We further analyzed the 8 genes related to drug-resistance
- For 5 (DLL1, CTBP2, BID, CA9, HtrA1) of them, resistant and sensitive have:
  - **Different** CNA values distribution (at the time of diagnosis)
  - **Not different** Gene expression distribution (at the time of diagnosis)
  - **Different** Gene expression distribution (after therapy)

- We further analyzed the 8 genes related to drug-resistance
- For 5 (DLL1, CTBP2, BID, CA9, HtrA1) of them, resistant and sensitive have:
  - **Different** CNA values distribution (at the time of diagnosis)
  - **Not different** Gene expression distribution (at the time of diagnosis)
  - **Different** Gene expression distribution (after therapy)

N.B.: The last information is known from literature and need experimental confirmation

Exploiting computational methods we identified a **molecular signature** that allows to:

Exploiting computational methods we identified a **molecular signature** that allows to:

- Predict the response to therapy (resistant / sensitive)
- Understand the cause of chemoresistance

Exploiting computational methods we identified a **molecular signature** that allows to:

- Predict the response to therapy (resistant / sensitive)
- Understand the cause of chemoresistance

The goal of the project is accomplished

- We built a classifier with satisfying performances integrating four types of genomic data

- We built a classifier with satisfying performances integrating four types of genomic data

- With our model, we discovered 137 CNA regions of amplification (less then 1% of the genome) as discriminatory for the two main classes, *resistant* and *sensitive*

- We built a classifier with satisfying performances integrating four types of genomic data

- With our model, we discovered 137 CNA regions of amplification (less then 1% of the genome) as discriminatory for the two main classes, *resistant* and *sensitive*

- These regions contain 24 genes related to HGS-OC, **8** of which are directly **connected to chemoresistance**

- We built a classifier with satisfying performances integrating four types of genomic data
- With our model, we discovered 137 CNA regions of amplification (less then 1% of the genome) as discriminatory for the two main classes, *resistant* and *sensitive*
- These regions contain 24 genes related to HGS-OC, **8** of which are directly **connected to chemoresistance**
- Two of the 8 genes belongs to the **Notch Signaling Pathway**

- The results obtained lead to an interesting theory:

  Enhanced drug-resistance could be a direct consequence of the activation of the pathway, due to the alteration of the expression of the identified genes, which in turn occurs as a consequence of their greater replication at diagnosis within these genomic segments.

- The results obtained lead to an interesting theory:

  Enhanced drug-resistance could be a direct consequence of the activation of the pathway, due to the alteration of the expression of the identified genes, which in turn occurs as a consequence of their greater replication at diagnosis within these genomic segments.

- Interesting therapeutic options for resistant patients may be developed by targeting the Notch Signaling pathway

- The results obtained lead to an interesting theory:

  Enhanced drug-resistance could be a direct consequence of the activation of the pathway, due to the alteration of the expression of the identified genes, which in turn occurs as a consequence of their greater replication at diagnosis within these genomic segments.

- Interesting therapeutic options for resistant patients may be developed by targeting the Notch Signaling pathway

- An efficient test for copy number alterations at diagnosis could be performed using ad-hoc probes on a small set of genes

POLITECNICO
MILANO 1863

HONOURS PROGRAMME

HP-SR
in Information Technology

Thanks for your attention!