Research Project Proposal: Integrative analysis of transcriptional, mutational and DNA structural profiles in ovarian cancer of chemotherapy sensitive vs. resistant patients

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

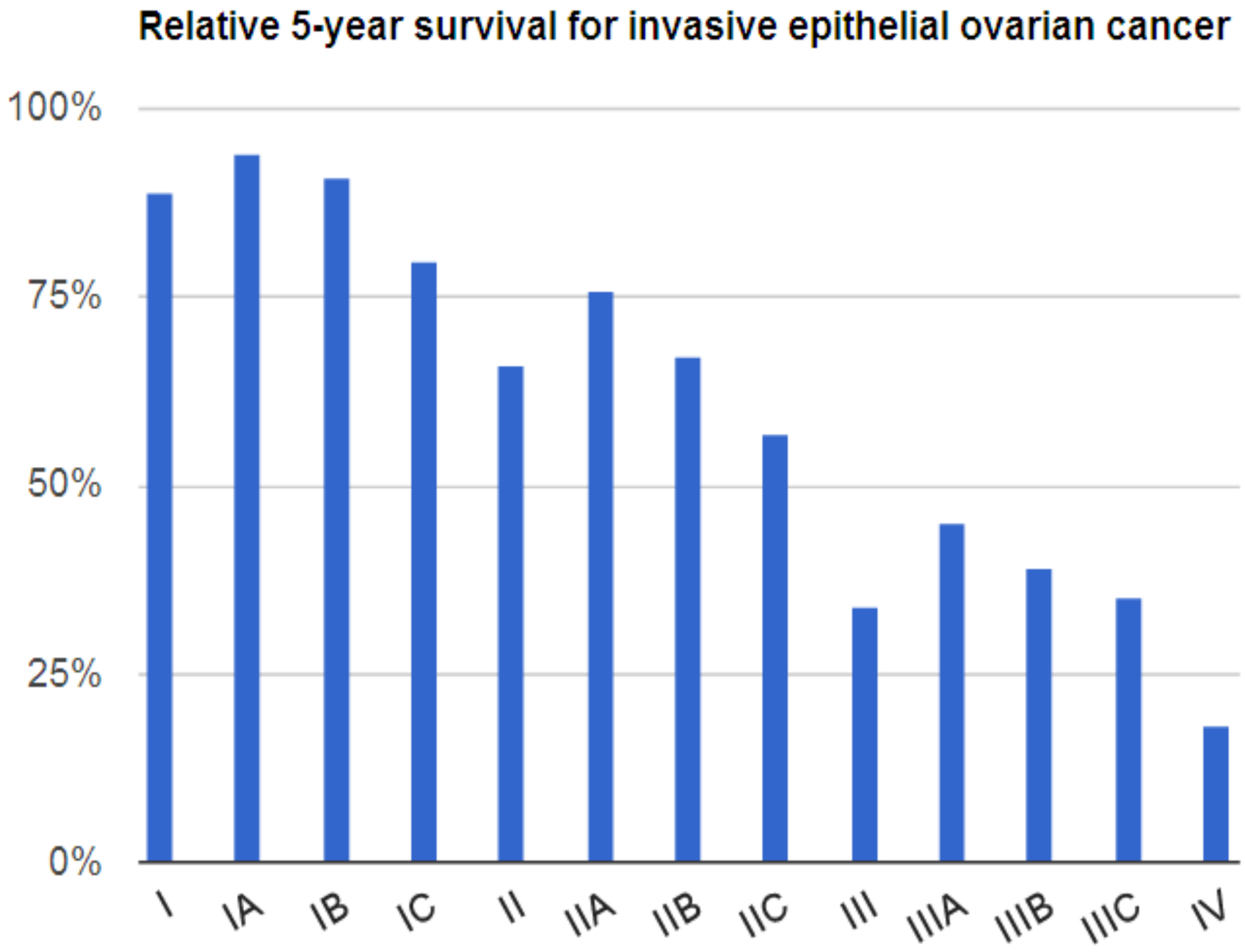
POLITECNICO MILANO 1863

HONOURS PROGRAMME HP-SR in Information Technology

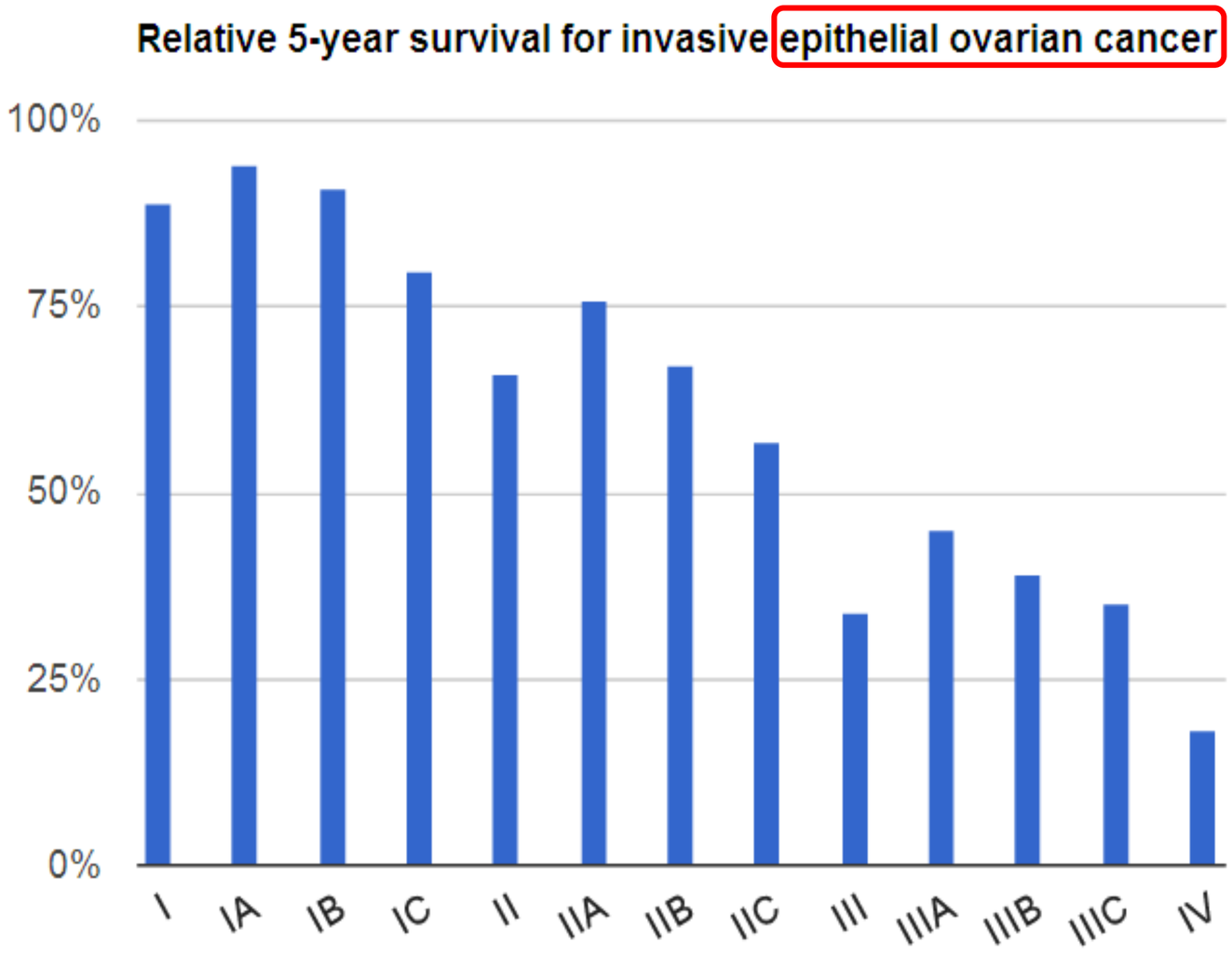
Genomic Computing

- Genomic computing is a new science focused on understanding the functioning of the genome.
- The aim is to make fundamental discoveries in biology and medicine.
- The challenge is to answer to relevant questions for biological and clinical research.

Research topic

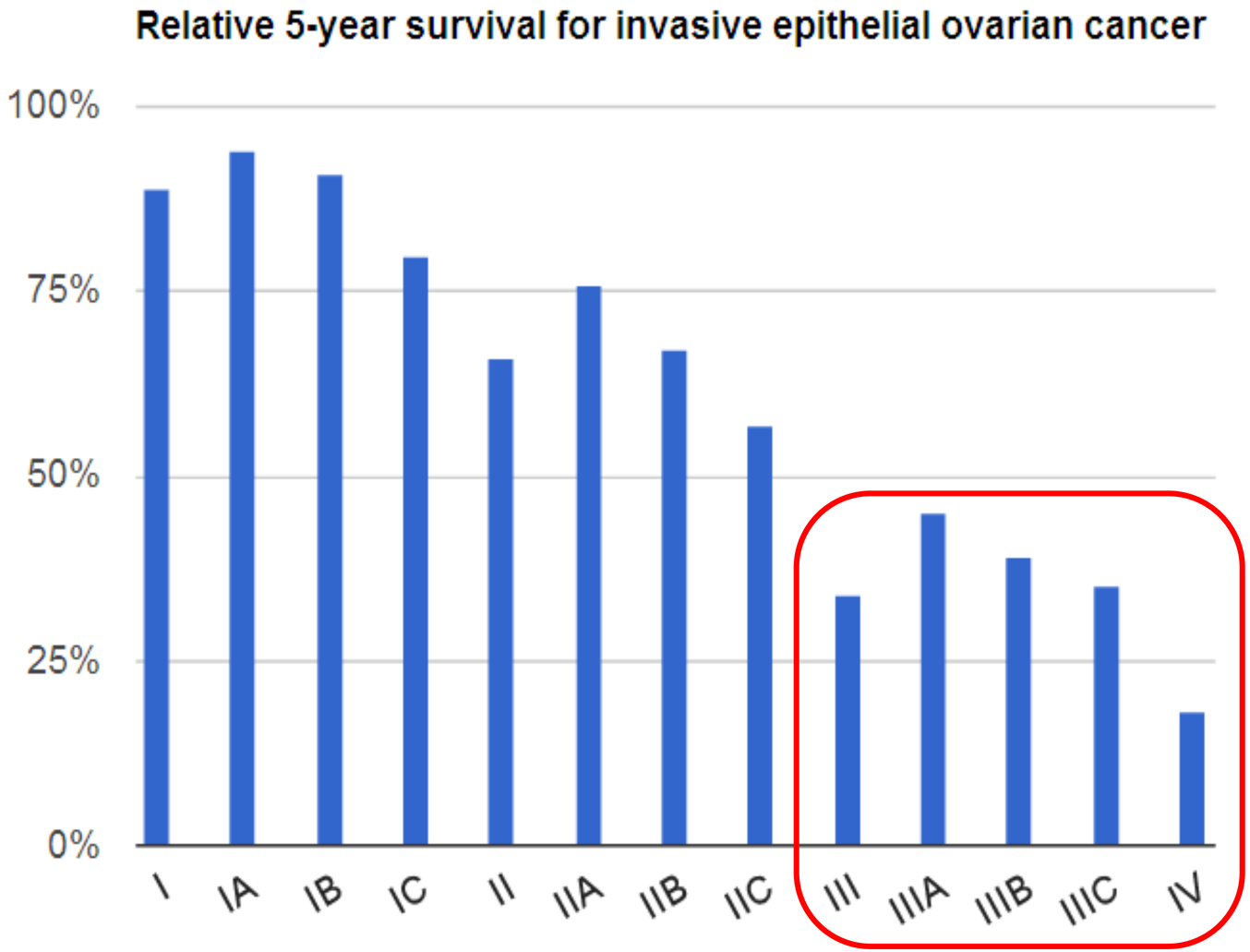


Research topic



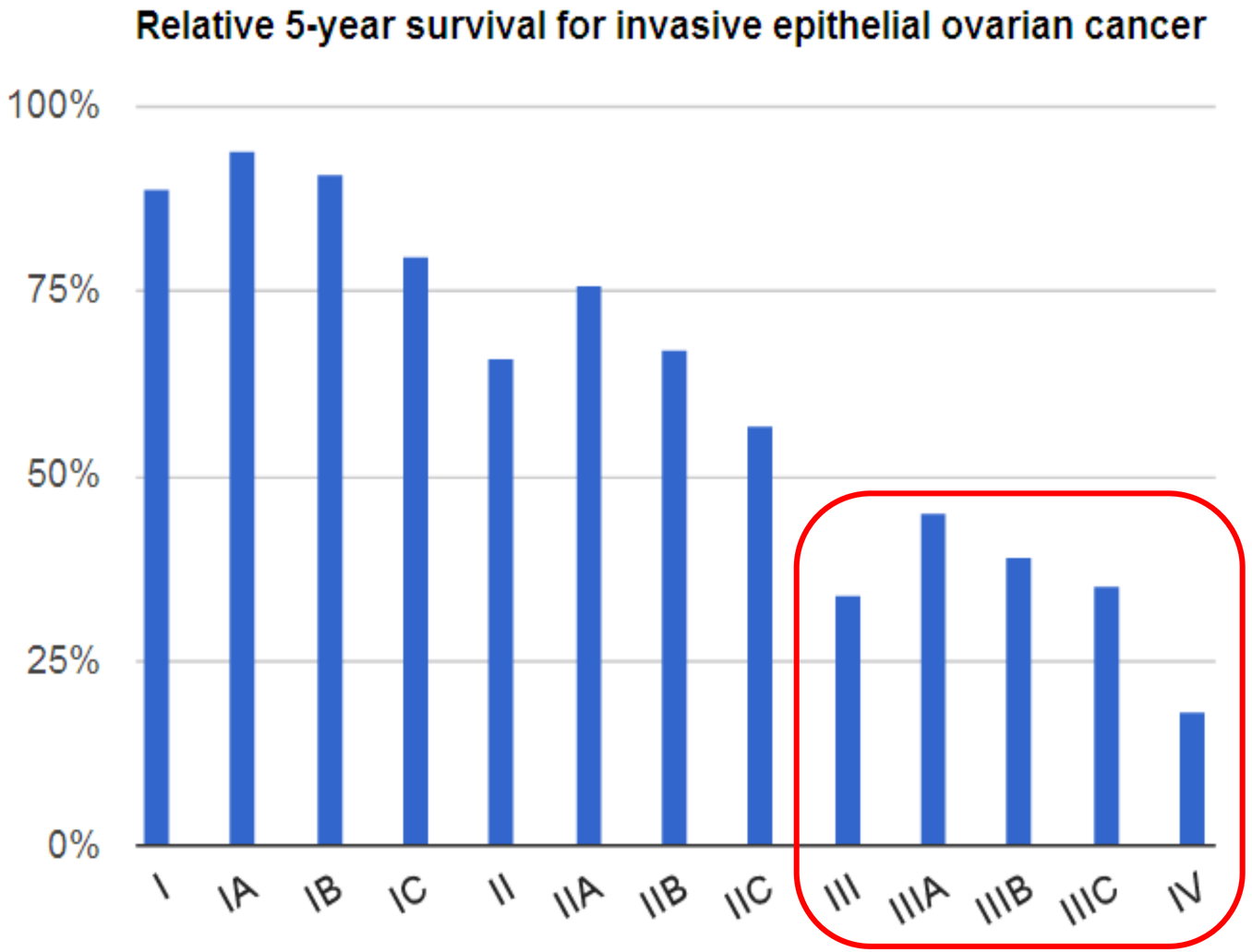
- Ovarian cancer

Research topic



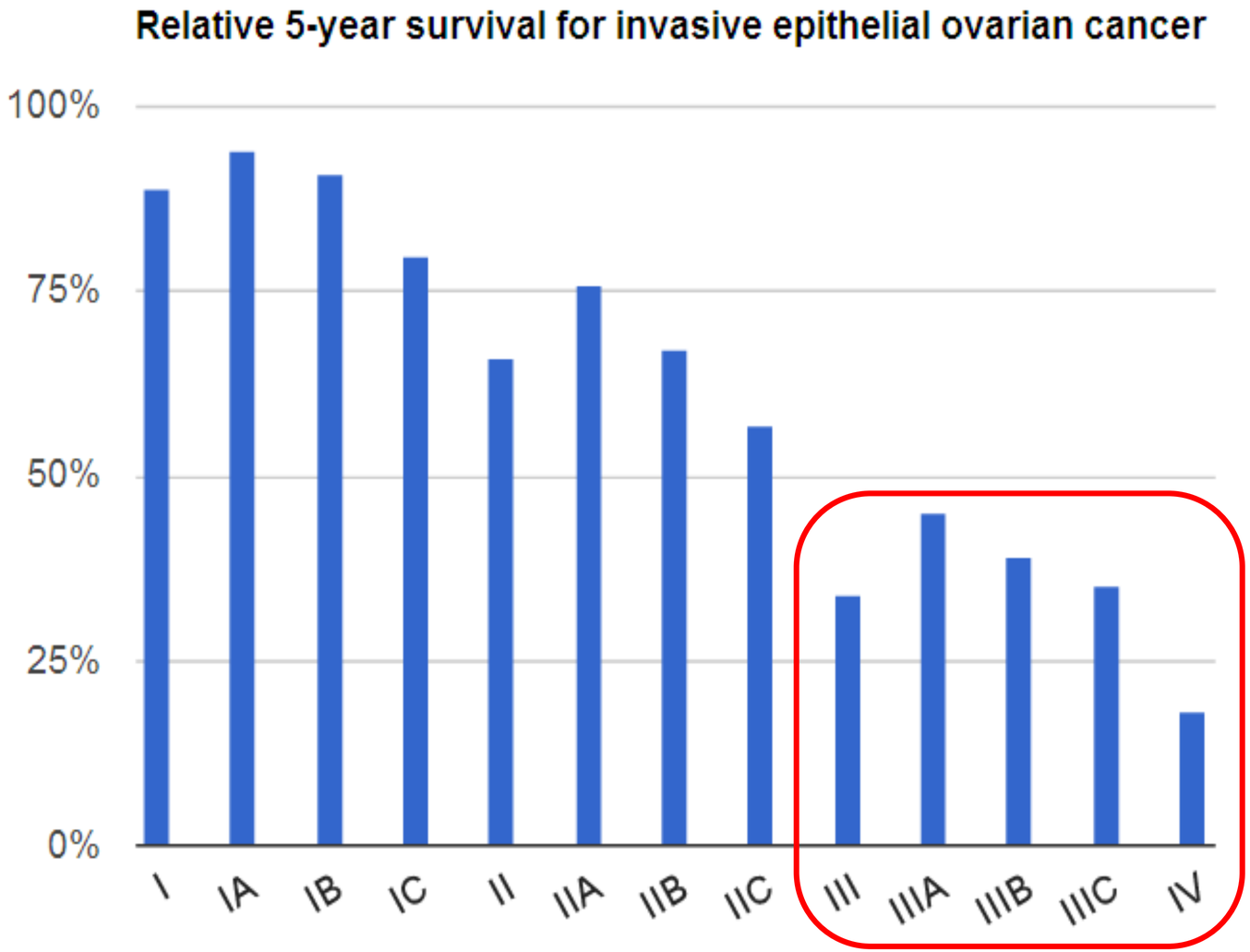
- Ovarian cancer
- HGS-OC: high-grade serous ovarian adenocarcinoma

Research topic



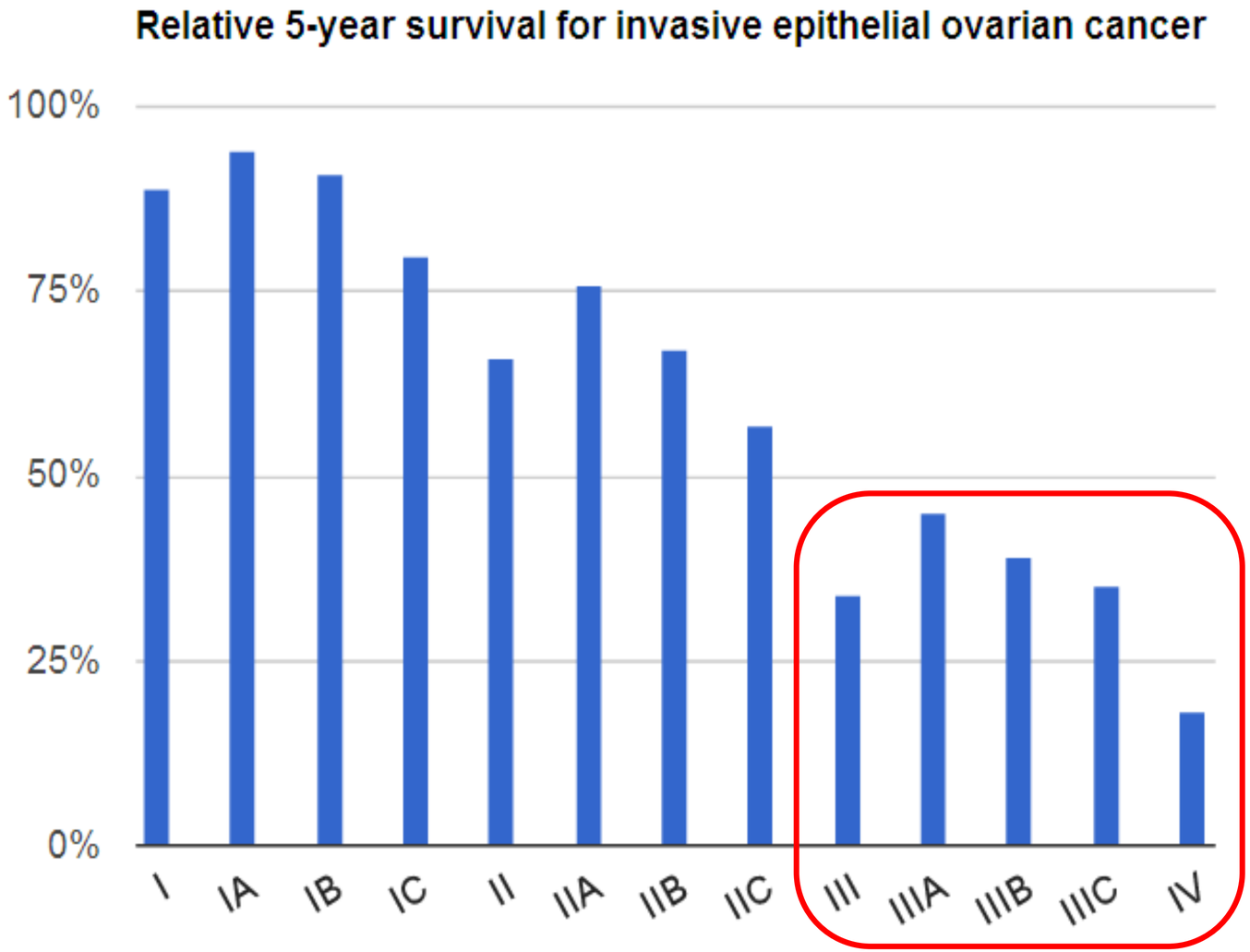
- Ovarian cancer
- HGS-OC: high-grade serous ovarian adenocarcinoma
- Treatment: surgery and cytoreduction followed by chemotherapy

Research topic



Problem with the treatment?

Research topic



Problem with the treatment?

- Relapse is likely to occur within a median of 16 months

Resistance to chemotherapy

The peculiarity of HGS-OC stands in the relapse timing of the patients affected by it:

Resistance to chemotherapy

The peculiarity of HGS-OC stands in the relapse timing of the patients affected by it:

- relapse within 6 months since the end of treatment: *resistant*;

Resistance to chemotherapy

The peculiarity of HGS-OC stands in the relapse timing of the patients affected by it:

- relapse within 6 months since the end of treatment: *resistant*;

- relapse after 12 months since the end of treatment: *sensitive*;

Resistance to chemotherapy

The peculiarity of HGS-OC stands in the relapse timing of the patients affected by it:

- relapse within 6 months since the end of treatment: *resistant*;
- relapse after 12 months since the end of treatment: *sensitive*;
- relapse after 36 months since the end of treatment: *sensitive long term*.

Relevance of the research project

- It is crucial to find a mechanism that allows to identify and discriminate resistant and sensitive patients, at the time of diagnosis.

Relevance of the research project

- It is crucial to find a mechanism that allows to identify and discriminate resistant and sensitive patients, at the time of diagnosis.

- New treatment options, which consider achievements in understanding of the pathophysiology of ovarian cancer, will then be needed to improve outcomes.

Relevance of the research project

- It is crucial to find a mechanism that allows to identify and discriminate resistant and sensitive patients, at the time of diagnosis.

- New treatment options, which consider achievements in understanding of the pathophysiology of ovarian cancer, will then be needed to improve outcomes.

- This study involves the analysis of resistance to chemotherapy in ovarian cancer patients, based on their transcriptional, mutational, and DNA structural profiles.

Aim of the research

- We will study the possibility of building a classifier able to predict the chemotherapy resistance of a patient affected by high grade serous ovarian cancer.

Aim of the research

- We will study the possibility of building a classifier able to predict the chemotherapy resistance of a patient affected by high grade serous ovarian cancer.

- The ultimate aim is the identification of a molecular signature (most likely the expression of a restricted list of genes) that could be used to predict the response to therapy (sensitive / resistant) at the time of diagnosis, starting from the Copy Number Alteration (CNA) profiles of the patients.

Aim of the research

- We will study the possibility of building a classifier able to predict the chemotherapy resistance of a patient affected by high grade serous ovarian cancer.

- The ultimate aim is the identification of a molecular signature (most likely the expression of a restricted list of genes) that could be used to predict the response to therapy (sensitive / resistant) at the time of diagnosis, starting from the Copy Number Alteration (CNA) profiles of the patients.

- The hope is that this classifier will achieve an accuracy of at least 80%.

Used technologies

- TCGA (The Cancer Genome Atlas): it is a public funded project that aims to catalogue and discover major cancer-causing genomic alterations to create a comprehensive “atlas” of cancer genomic profiles.

Used technologies

- TCGA (The Cancer Genome Atlas): it is a public funded project that aims to catalogue and discover major cancer-causing genomic alterations to create a comprehensive “atlas” of cancer genomic profiles.

- GMQL (GenoMetric Query Language): it provides parallel computation in the cloud, thereby supporting queries over thousands of samples, such as the ones provided by ENCODE and TCGA consortia.

Used technologies

- TCGA (The Cancer Genome Atlas): it is a public funded project that aims to catalogue and discover major cancer-causing genomic alterations to create a comprehensive “atlas” of cancer genomic profiles.

- GMQL (GenoMetric Query Language): it provides parallel computation in the cloud, thereby supporting queries over thousands of samples, such as the ones provided by ENCODE and TCGA consortia.

- GISTIC 2.0: it identifies regions of the genome that are significantly amplified or deleted across a set of samples.

Research plan

- Data extraction

Research plan

- Data extraction
- Data analysis

Research plan

- Data extraction
- Data analysis
- Implementation

Research plan

- Data extraction
- Data analysis
- Implementation
- Validation

Research plan

- Data extraction
- Data analysis
- Implementation
- Validation

Samples were selected from TCGA repository as follows:

Research plan

- Data extraction
- Data analysis
- Implementation
- Validation

Samples were selected from TCGA repository as follows:

- Locate the barcode (TCGA-XXX-YYY) for samples marked as “Sensitive” or “Resistant” in the “Platinum status” column.

Research plan

- Data extraction
- Data analysis
- Implementation
- Validation

Samples were selected from TCGA repository as follows:

- Locate the barcode (TCGA-XXX-YYY) for samples marked as “Sensitive” or “Resistant” in the “Platinum status” column.
- Check the Progression-free Survival column to discriminate sensitive and sensitive long term (PFS months > 36).

Research plan

- Data extraction
- Data analysis
- Implementation
- Validation

Samples were selected from TCGA repository as follows:

- Locate the barcode (TCGA-XXX-YYY) for samples marked as “Sensitive” or “Resistant” in the “Platinum status” column.
- Check the Progression-free Survival column to discriminate sensitive and sensitive long term (PFS months > 36).
- Obtain three sets of data (one for each type of patients) after executing three different query on GMQL (GenoMetric Query Language).

Research plan

- Data extraction
- Data analysis
- Implementation
- Validation

A visual analysis of the data was carried out, in order to understand in which regions of the genome the three groups of patients differ the most.

Research plan

- Data extraction
- Data analysis
- Implementation
- Validation

A visual analysis of the data was carried out, in order to understand in which regions of the genome the three groups of patients differ the most.

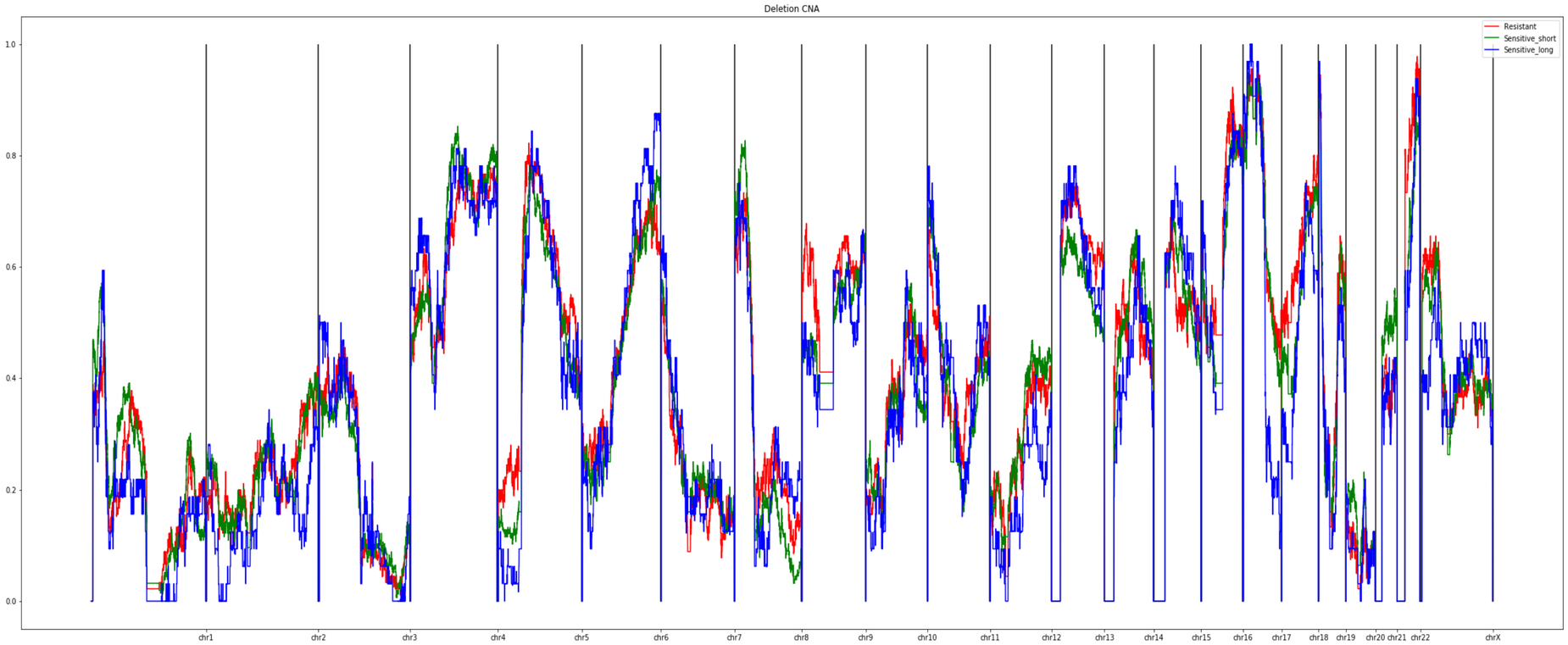
In particular, it was done on the CNA profiles of the three classes.

Research plan

- Data extraction
- Data analysis
- Implementation
- Validation

A visual analysis of the data was carried out, in order to understand in which regions of the genome the three groups of patients differ the most.

In particular, it was done on the CNA profiles of the three classes.



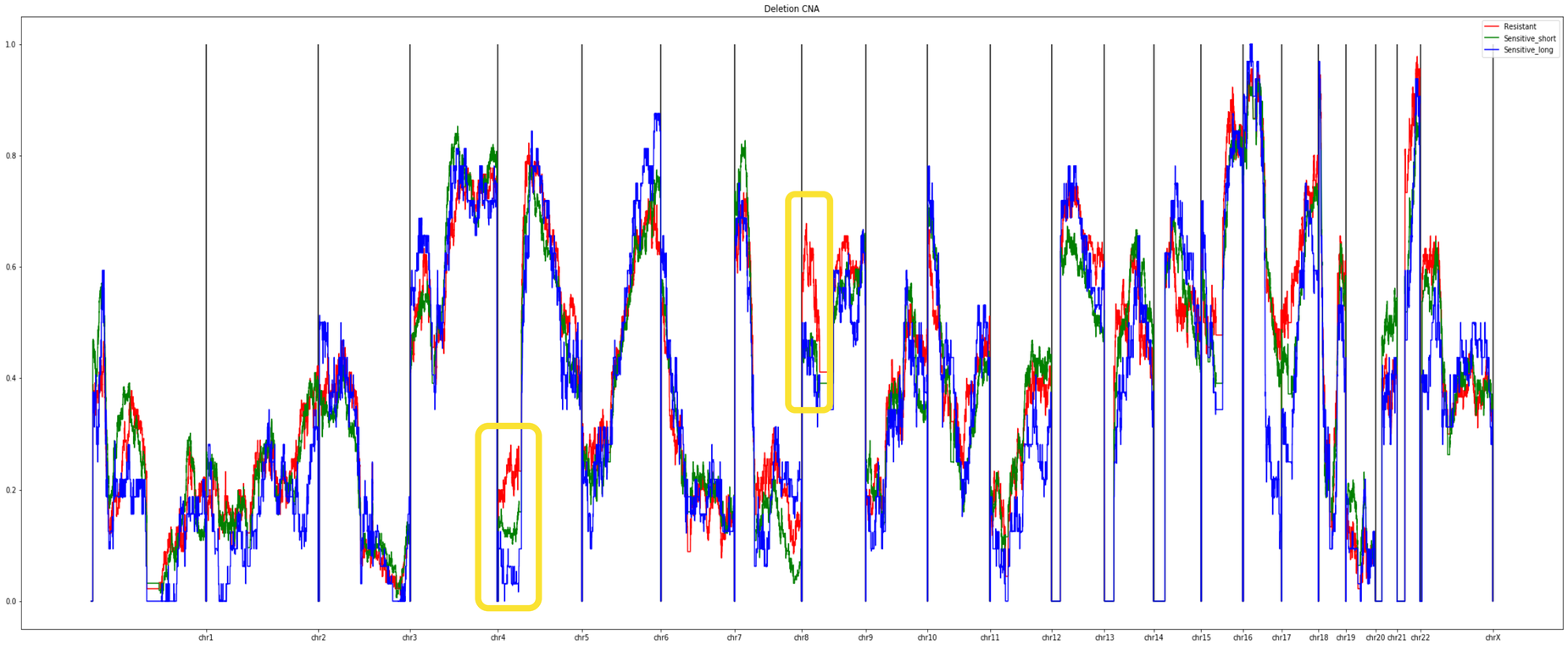
Resistant *Sensitive* *Sensitive long term*

Research plan

- Data extraction
- Data analysis
- Implementation
- Validation

A visual analysis of the data was carried out, in order to understand in which regions of the genome the three groups of patients differ the most.

In particular, it was done on the CNA profiles of the three classes.



Resistant *Sensitive* *Sensitive long term*

Research plan

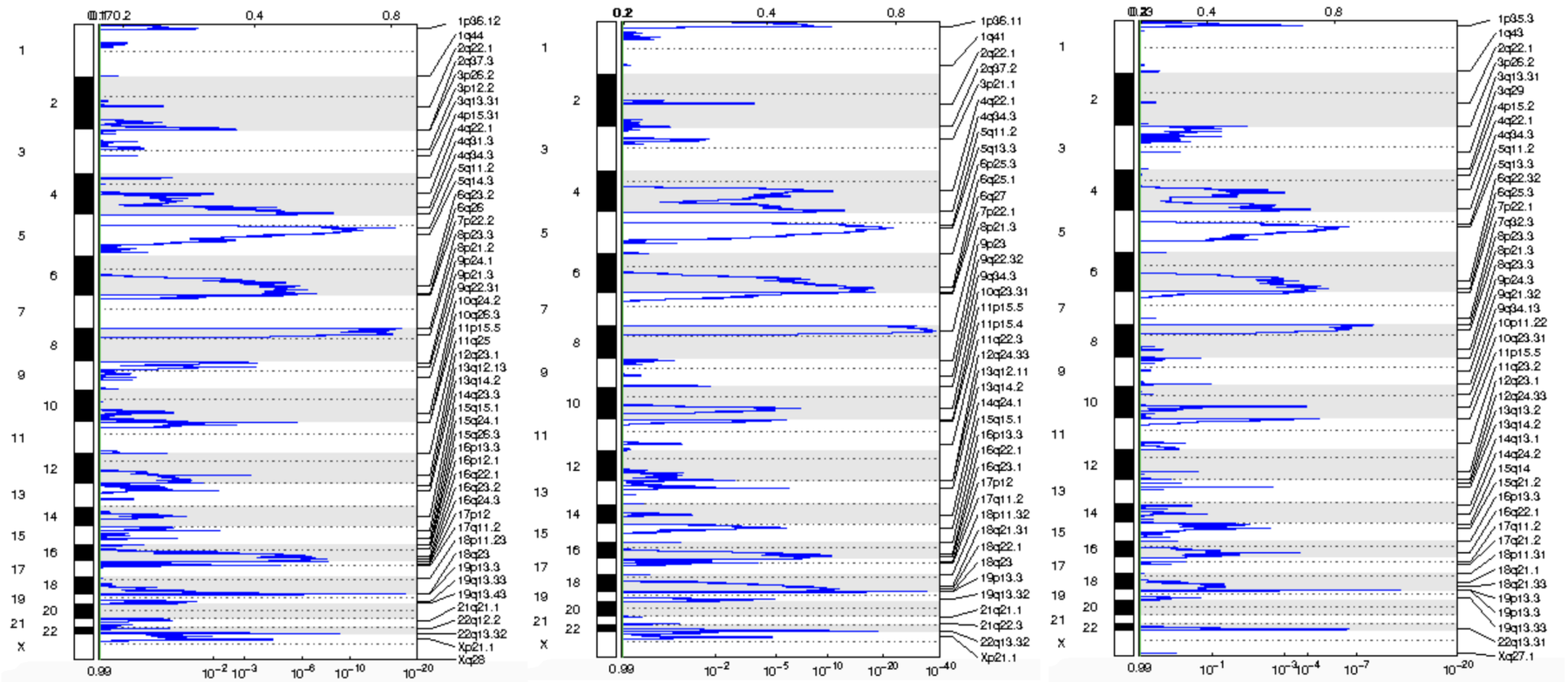
- Data extraction
- Data analysis
- Implementation
- Validation

We also used GISTIC 2.0 to visualize which were relevant regions of CNA identified by the tool and to have a further demonstration that different kind of patients do present difference in their CNA profiles.

Research plan

- Data extraction
- Data analysis
- Implementation
- Validation

We also used GISTIC 2.0 to visualize which were relevant regions of CNA identified by the tool and to have a further demonstration that different kind of patients do present difference in their CNA profiles.

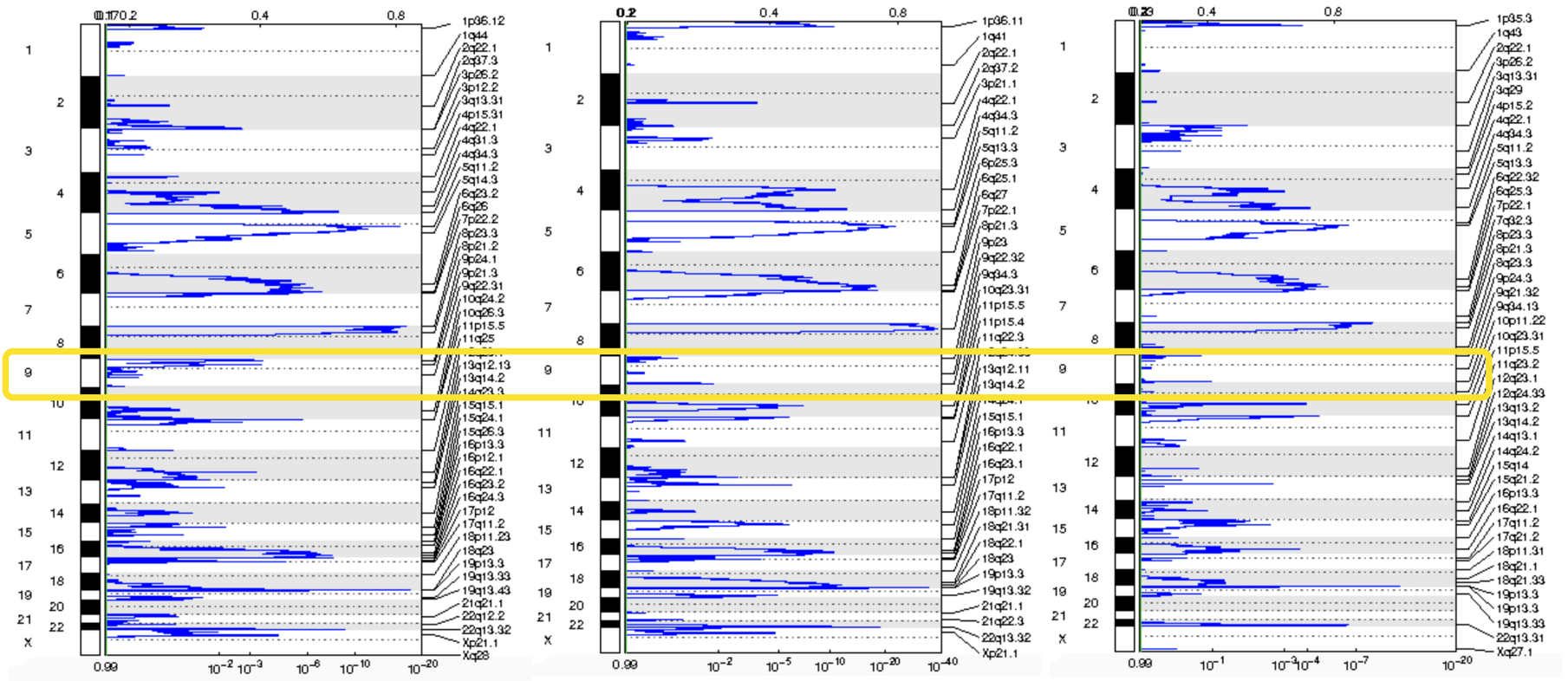


Resistant Sensitive Sensitive long term

Research plan

- Data extraction
- Data analysis
- Implementation
- Validation

We also used GISTIC 2.0 to visualize which were relevant regions of CNA identified by the tool and to have a further demonstration that different kind of patients do present difference in their CNA profiles.



Resistant Sensitive Sensitive long term

Research plan

- Data extraction
- Data analysis
- Implementation
- Validation

Different kind of classifier will be implemented and tested:

Research plan

- Data extraction
- Data analysis
- Implementation
- Validation

Different kind of classifier will be implemented and tested:

- A classifier that uses only CNA data.

Research plan

- Data extraction
- Data analysis
- Implementation
- Validation

Different kind of classifier will be implemented and tested:

- A classifier that uses only CNA data.
- A classifier that uses relevant CNA regions in order to identify a set of genes, whose expression will then be used to classify patients.

Research plan

- Data extraction
- Data analysis
- Implementation
- Validation

Different kind of classifier will be implemented and tested:

- A classifier that uses only CNA data.
- A classifier that uses relevant CNA regions in order to identify a set of genes, whose expression will then be used to classify patients.

The possibility to use the tool GISTIC 2.0 to identify those relevant regions will be considered.

Research plan

- Data extraction
- Data analysis
- Implementation
- Validation

Different kind of classifier will be implemented and tested:

- A classifier that uses only CNA data.
- A classifier that uses relevant CNA regions in order to identify a set of genes, whose expression will then be used to classify patients.

The possibility to use the tool GISTIC 2.0 to identify those relevant regions will be considered.

After creating the data set, we will use some known classifier, e.g. Random Forest, K-Nearest Neighbours or AdaBoost.

Research plan

- Data extraction
- Data analysis
- Implementation
- Validation

In order to identify the best model, a 10-fold cross validation will be executed for each proposed classifier.

Research plan

- Data extraction
- Data analysis
- Implementation
- Validation

In order to identify the best model, a 10-fold cross validation will be executed for each proposed classifier.

At the end, a test of the obtained model will be done using in-house data, which are never used during the training phase.

Implementation steps done so far

- Implementation with GISTIC 2.0

Implementation steps done so far

- Implementation with GISTIC 2.0
- Implementation without GISTIC 2.0

Implementation steps done so far

- Implementation with GISTIC 2.0

  We obtained from GISTIC 2.0 significantly amplified and deleted regions of the genome across the three set of samples (Resistant, Sensitive, Sensitive Long Term).

- Implementation without GISTIC 2.0

Implementation steps done so far

- Implementation with GISTIC 2.0
- Implementation without GISTIC 2.0

We obtained from GISTIC 2.0 significantly amplified and deleted regions of the genome across the three set of samples (Resistant, Sensitive, Sensitive Long Term).

We used those regions in two different ways:

Implementation steps done so far

- Implementation with GISTIC 2.0
- Implementation without GISTIC 2.0

We obtained from GISTIC 2.0 significantly amplified and deleted regions of the genome across the three set of samples (Resistant, Sensitive, Sensitive Long Term).

We used those regions in two different ways:

- As features for a classifier that uses only CNA data.

Implementation steps done so far

- Implementation with GISTIC 2.0
- Implementation without GISTIC 2.0

We obtained from GISTIC 2.0 significantly amplified and deleted regions of the genome across the three set of samples (Resistant, Sensitive, Sensitive Long Term).

We used those regions in two different ways:

- As features for a classifier that uses only CNA data.
- As a features selection tool for a classifier that uses gene expression data (using as features the genes that were mapped on those regions).

Implementation steps done so far

- Implementation with GISTIC 2.0
- Implementation without GISTIC 2.0

We obtained from GISTIC 2.0 significantly amplified and deleted regions of the genome across the three set of samples (Resistant, Sensitive, Sensitive Long Term).

We used those regions in two different ways:

- As features for a classifier that uses only CNA data.
- As a features selection tool for a classifier that uses gene expression data (using as features the genes that were mapped on those regions).

We discovered that the regions identified by GISTIC were not able to correctly discriminate the three classes.

Implementation steps done so far

- Implementation with GISTIC 2.0
- Implementation without GISTIC 2.0

We identified regions that have different CNA across the three set of samples (Resistant, Sensitive, Sensitive Long Term), starting from their values on the whole genome.

Implementation steps done so far

- Implementation with GISTIC 2.0
- Implementation without GISTIC 2.0

We identified regions that have different CNA across the three set of samples (Resistant, Sensitive, Sensitive Long Term), starting from their values on the whole genome.

We used again those regions in two different ways:

Implementation steps done so far

- Implementation with GISTIC 2.0
- Implementation without GISTIC 2.0

We identified regions that have different CNA across the three set of samples (Resistant, Sensitive, Sensitive Long Term), starting from their values on the whole genome.

We used again those regions in two different ways:

- As features for a classifier that uses only CNA data.

Implementation steps done so far

- Implementation with GISTIC 2.0
- Implementation without GISTIC 2.0

We identified regions that have different CNA across the three set of samples (Resistant, Sensitive, Sensitive Long Term), starting from their values on the whole genome.

We used again those regions in two different ways:

- As features for a classifier that uses only CNA data.
- As a features selection tool for a classifier that uses gene expression data (using as features the genes that were mapped on those regions).

Implementation steps done so far

- Implementation with GISTIC 2.0
- Implementation without GISTIC 2.0

We identified regions that have different CNA across the three set of samples (Resistant, Sensitive, Sensitive Long Term), starting from their values on the whole genome.

We used again those regions in two different ways:

- As features for a classifier that uses only CNA data.
- As a features selection tool for a classifier that uses gene expression data (using as features the genes that were mapped on those regions).

We tried first to discriminate the two classes that are more different, i.e. Resistant and Sensitive long term.

Preliminary relevant results

The classifier implemented without GISTIC 2.0 and using the identified regions to select 8875 relevant genes let us achieve promising results.

Preliminary relevant results

The classifier implemented without GISTIC 2.0 and using the identified regions to select 8875 relevant genes let us achieve promising results.

In particular, running a 10-fold cross validation and using AdaBoost as classification algorithm, we got the following performance in classifying Resistant against Sensitive long term:

Preliminary relevant results

The classifier implemented without GISTIC 2.0 and using the identified regions to select 8875 relevant genes let us achieve promising results.

In particular, running a 10-fold cross validation and using AdaBoost as classification algorithm, we got the following performance in classifying Resistant against Sensitive long term:

- Average precision: 0.75.

Preliminary relevant results

The classifier implemented without GISTIC 2.0 and using the identified regions to select 8875 relevant genes let us achieve promising results.

In particular, running a 10-fold cross validation and using AdaBoost as classification algorithm, we got the following performance in classifying Resistant against Sensitive long term:

- Average precision: 0.75.
- Average recall: 0.77.

Preliminary relevant results

The classifier implemented without GISTIC 2.0 and using the identified regions to select 8875 relevant genes let us achieve promising results.

In particular, running a 10-fold cross validation and using AdaBoost as classification algorithm, we got the following performance in classifying Resistant against Sensitive long term:

- Average precision: 0.75.
- Average recall: 0.77.
- Average accuracy: 0.68.

Preliminary relevant results

In order to improve the previous results, we normalized the values of expression of the selected 8875 genes in the dataset.

Preliminary relevant results

In order to improve the previous results, we normalized the values of expression of the selected 8875 genes in the dataset.

Then, we run again a 10-fold cross validation, using AdaBoost as classification algorithm, and we got the following performance:

Preliminary relevant results

In order to improve the previous results, we normalized the values of expression of the selected 8875 genes in the dataset.

Then, we run again a 10-fold cross validation, using AdaBoost as classification algorithm, and we got the following performance:

- Average precision: 0.84.

Preliminary relevant results

In order to improve the previous results, we normalized the values of expression of the selected 8875 genes in the dataset.

Then, we run again a 10-fold cross validation, using AdaBoost as classification algorithm, and we got the following performance:

- Average precision: 0.84.
- Average recall: 0.88.

Preliminary relevant results

In order to improve the previous results, we normalized the values of expression of the selected 8875 genes in the dataset.

Then, we run again a 10-fold cross validation, using AdaBoost as classification algorithm, and we got the following performance:

- Average precision: 0.84.
- Average recall: 0.88.
- Average accuracy: 0.79.

Future steps

- We will refine the way we select the relevant different regions.

Future steps

- We will refine the way we select the relevant different regions.
- In this way, we will obtain as features less genes, which are more meaningful for the discrimination of the two classes.

Future steps

- We will refine the way we select the relevant different regions.
- In this way, we will obtain as features less genes, which are more meaningful for the discrimination of the two classes.
- We will apply the same procedure in order to classify also Resistant against Sensitive and finally putting all the classes together.

Future steps

- We will refine the way we select the relevant different regions.
- In this way, we will obtain as features less genes, which are more meaningful for the discrimination of the two classes.
- We will apply the same procedure in order to classify also Resistant against Sensitive and finally putting all the classes together.
- This will hopefully lead us to a classifier with the desired performance.

Bibliography

- Ducie, J., Dao, F., Considine, M., Olvera, N., Shaw, P. A., Kurman, R. J., Shih, I.-M., Soslow, R. A., Cope, L., and Levine, D. A. Molecular analysis of high-grade serous ovarian carcinoma with and without associated serous tubal intra-epithelial carcinoma. *Nature Communications 8*, 1 (2017), 990.

- Gad Getz, Rameen Beroukhim, C. M. S. S., and Dobson, J. Genomic identification of significant targets in cancer. GISTIC Technical Report Release 2.0.23, Broad Institute, March 2017.

- Li, M., Balch, C., Montgomery, J. S., Jeong, M., Chung, J. H., Yan, P., Huang, T. H., Kim, S., and Nephew, K. P. Integrated analysis of dna methylation and gene expression reveals specific signaling pathways associated with platinum resistance in ovarian cancer. *BMC Medical Genomics 2*, 1 (Jun 2009), 34.

- Network, T. C. G. A. R., and Bell, e. A. Integrated genomic analyses of ovarian carcinoma. *Nature 474* (06 2011), 609 EP –.

- Tomczak, Katarzyna, Patrycja Czerwińska, and Maciej Wiznerowicz. "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge." *Contemporary oncology 19.1A* (2015): A68.

POLITECNICO
MILANO 1863

HONOURS PROGRAMME
HP-SR
in Information Technology