POLITECNICO DI MILANO
School of Industrial and Information Engineering
Master of Science in Biomedical Engineering

Department of Electronics, Information and Bioengineering

Genome-based chemoresistance prediction in High-Grade Serous Ovarian Adenocarcinoma

Laboratory of Data Science and Bioinformatics

Supervisor: Prof. Stefano Ceri
Co-advisors: Prof. Francesca Ieva
Dr. Arif Çanakoğlu
Dr. Pietro Pinoli

Master Thesis of:
Giada Lalli, Student ID 884272

Academic Year 2018-2019

Il fato crudele
La daga affilata
Il rapido dardo
del tempo galante -
di questi oscuri Signori
non uno annebbia
le maestose colline
Instancabili
Violente
dei miei Desideri.

Contents

Abstract **xvii**

Sommario **xix**

Thanks and greetings **xxi**

1 Introduction **1**
1.1 General outline 1
1.2 Thesis structure 4

2 Ovarian Cancer **5**
2.1 Overview 5
2.2 Incidence and Prognosis 6
2.2.1 By Population/Age 6
2.2.2 Prognosis 6
2.3 Classifications and Histopathology 6
2.4 *Epithelial Ovarian Cancer* 7
2.4.1 Background 7
2.4.2 The Morphological and Molecular Heterogeneity of Epithelial Ovarian Cancer 8
2.4.3 The Cell of Origin of Most Epithelial Ovarian Cancer is not Ovarian 10
2.5 LGSC versus HGSC 17
2.5.1 Background 17
2.5.2 Behaviour 17
2.5.3 Pathogenesis 18
2.5.4 Clinical aspects, histologic features and selected diagnostic problems 22
2.6 Proposed case of study: High-Grade Serous Ovarian Adenocarcinoma 27
2.6.1 Resistance to platinum-based chemotherapy 28

3 Bioinformatics and tools **29**
3.1 Computational Biology 29
3.2 Genomics 30
3.2.1 Gene expression and its regulation 32
3.2.2 miRNA 34

3.2.3 Epigenetics and DNA methylation . . . 36
3.2.4 CNA regions . . . 38
3.3 GMQL . . . 38
3.4 Python Libraries . . . 40
3.5 R Libraries . . . 41
3.6 Cytoscape . . . 42

4 Datasets **43**
4.1 TCGA: The Cancer Genome Atlas . . . 43
4.1.1 Introduction . . . 43
4.1.2 The idea behind the project . . . 43
4.1.3 The Cancer Genome Atlas Research Network . . . 44
4.1.4 Platforms and data models . . . 45
4.1.5 Visualisation and examination of the genomic data . . . 47
4.2 Project Datasets . . . 48
4.2.1 Building datasets . . . 48

5 Computational Methods Foundation **55**
5.1 Multiple statistical test correction . . . 55
5.2 Features selection . . . 58
5.3 Classification . . . 61
5.3.1 Supervised Learning . . . 61
5.3.2 Evaluation of the classifier . . . 61
5.4 Standard methods of classification . . . 63
5.4.1 Random Forest . . . 63
5.4.2 Logistic Regression . . . 64
5.4.3 k-Nearest Neighbor . . . 66
5.4.4 Adaboost . . . 66
5.4.5 Suppor Vector Machine . . . 68
5.5 Survival Analysis . . . 69
5.5.1 Censoring . . . 69
5.5.2 Terminology and notation . . . 70
5.5.3 Survival data . . . 71
5.5.4 Kaplan-Meier survival estimate . . . 71
5.5.5 Performance measure: Concordance Index . . . 72
5.5.6 CoxPH Model . . . 73
5.6 Windowing . . . 75

6 Computational Methods **77**
6.1 Features selection . . . 78
6.2 Survival Analysis . . . 79
6.3 Classification . . . 80

7 Results **83**
7.1 Computational results . . . 83
7.1.1 Best performances of survival analysis . . . 83
7.1.2 Best performances of classifier . . . 86
7.2 Biological Results . . . 99
7.2.1 Enrichment Analysis . . . 99

8 Conclusions **113**

List of Figures

2.1 Key aspects of the types of epithelial ovarian cancers 7

2.2 Serous tubal intra-epithelial carcinoma (STIC). A. High magnification. Hematoxylin and eosin stain. B. Immunohistochemical stain for p53. An asterisk defines the boundary of the lesion. 13

2.3 Comparison of the immunohistochemical staining pattern for ovarian surface epithelium (mesothelium), normal fallopian tube epithelium, and high-grade serous carcinoma. PAX8 is a marker of Müllerian-type epithelium such as fallopian tube epithelium and calretinin is a marker of mesothelium. 13

2.4 Transfer of normal tubal epithelium to the ovary. A. Anatomical relationship of fallopian tube to the ovary at the time of ovulation. The fimbria envelops the ovary. B. Ovulation. The ovarian surface ruptures with expulsion and transfer of the oocyte to the fimbria. The fimbria is in intimate contact with the ovary at the site of rupture. C. Tubal epithelial cells from the fimbria are dislodged and implant on the denuded surface of the ovary resulting in the formation of an inclusion cyst. 15

2.5 Proposed development of low-grade (LG) and high-grade (HG) serous carcinoma. A. One mechanism involves normal tubal epithelium that is shed from the fimbria, which implants on the ovary to form an inclusion cyst. Depending on whether there is a mutation of KRAS/BRAF/ERRB2 or TP53 a low-grade or high-grade serous carcinoma develops respectively. Low-grade serous carcinoma often develops from a serous borderline tumor (SBT), which in turn arises from a serous cystadenoma. Another mechanism involves exfoliation of malignant cells from a serous tubal intraepithelial carcinoma (STIC) that implants on the ovarian surface resulting in the development of a high-grade serous carcinoma. B. A schematic representation of direct dissemination or shedding of STIC cells onto the ovarian surface where the carcinoma cells ultimately establish a tumor mass that is presumably arising from the ovary. Of note there may be stages of tumor progression that precede the formation of a STIC. 16
2.6 Small APST arisin in a serous cystadenoma. 18
2.7 APST. 18
2.8 MPSC. 19
2.9 Tumor progression in MPSC. 19
2.10 Microinvasion in MPSC. 19
2.11 Invasive implants. 19
2.12 Tubal intraepithelial carcinoma (TIC). The epithelium of the fallopian tube mucosa with TIC is thicker compared with normal mucosa (upper center). 21
2.13 Glandular pattern. 23
2.14 Diffuse, solid pattern. 23
2.15 Typical papillary pattern showing irregular slit-like spaces. 23
2.16 Micropapillary pattern. 23
2.17 The nuclei of high-grade serous carcinoma are larger with greater pleomorphism and larger nucleoli in respect to LGSC. 24
2.18 Endometrioid carcinoma-like pattern. 25
2.19 Serous carcinoma with clear cytoplasm. 25
2.20 Transitional cell carcinoma-like pattern. 25
2.21 Signet ring change simulating signet ring cells of metastatic adenocarcinoma involving the ovary. 27

3.1 DNA double helix structure. 30
3.2 The central dogma of molecular biology: it explains the flow of genetic information, from DNA to RNA, to make a functional product, a protein. 32
3.3 miRNA building process. 35
3.4 miRNA expression motifs. 36

3.5 Graphical explanation of the role of methylation in gene expression. 37
3.6 Graphical representation of CNA mutation. 38
3.7 GMQL web interface. 40

4.1 The Cancer Genome Atlas (TCGA) Research Network Centres flowchart. 45
4.2 Abbreviated example of the performed query on GMQL web interface to obtain files for Resistant class, gene expression data. 49
4.3 Example of the first patient for each class and the relative informations about gene expression. As can be seen also in Fig.4.4 and 4.5, the barcode of each patient has been replaced with the first letter associated to the class of belonging, followed by the rank. 50
4.4 Example of the first patient for each class and the relative informations about miRNA expression. 50
4.5 Example of the first patient for each class and the relative informations about DNA methylation. 50
4.6 Example of the first patient for each class and the relative informations about CNA regions. 51

5.1 Visual representation of hypergeometric test functioning. . . 58
5.2 CNA amplification profiles with a resolution of 10K for Resistant (red), Sensitive short (green) and Sensitive long (blue). 60
5.3 Graphical explanation of *k-fold cross-validation*. 62
5.4 Example of how Random Forest works: in the diagram each decision tree has voted or predicted a specific class. The final output or class selected by the Random Forest will be the Class N, as it has majority votes or is the predicted output by two out of the four decision trees. 64
5.5 Example of how Logistic Regression works with graphical representation of components. 65
5.6 Sigmoid fuction. 65
5.7 KNN working steps. 66
5.8 Illustration of AdaBoost algorithm for creating a strong classifier based on multiple weak linear classifiers. 67
5.9 Graphical example of margin maximization operated from SVM: the margin is defined as the distance between the separating hyperplane and the training sample that are closest to it. 68
5.10 How Nonlinear SVM overcomes the problem of data that cannot be separated from Linear SVM. From [33]. 69
5.11 Survival analysis: survival function. 70
5.12 Survival analysis: Kaplan-Meier curve. 72

5.13 Correlation Matrix of features extracted with the "Mild B-H Correction", without collinearities, miRNA expression data. 74
5.14 MiRNA features expressed on chromosome 8 for Resistant class: in *red*, features considered relevant in distinguishing Sensitive Long and Sensitive Short classes are shown; in *blue*, features considered relevant in distinguishing Sensitive Long and Resistant classes are shown; in *green*, features considered relevant in distinguishing Resistant and Sensitive Short classes are shown. 75

6.1 Progression Free Survival Kaplan Meier Estimate: time to relapse of each class is represented by different color described in label. 80

7.1 Prediction survival function of each class for gene expression data. 84
7.2 Prediction survival function of each class for miRNA data. . 85
7.3 Prediction survival function of each class for methylation data. 85
7.4 ROC curves of gene expression classifier, obtained by each binary comparison. 95
7.5 ROC curves of miRNA expression classifier, obtained by each binary comparison. 96
7.6 ROC curves of DNA methylation classifier, obtained by each binary comparison. 97
7.7 ROC curves for each comparison by merging the best features obtained from different data types (gene expression data, miRNA expression data, methylation data, CNA data). 98
7.8 Notch Signaling Pathway from DAVID Functional Annotation Tool. 101
7.9 Schematic of Notch signaling. 102
7.10 Connection network between relevant genes for progression of HGS-OC and miRNAs, obtained by the mean of Cytoscape tool. 104
7.11 Connection network between relevant genes for drug-response in HGS-OC and miRNAs, obtained by the mean of Cytoscape tool. 105
7.12 Gene expression comparison of 8 genes considered to be involved in drug-resistance mediated by Notch signaling pathwhay in Resistant and Sensitive classes. 106
7.13 DLL1 gene expression across all the classes. 107
7.14 CNA region alteration associated to DLL1 across all the classes. 107
7.15 CTBP2 gene expression across all the classes. 107
7.16 CNA region alteration associated to CTBP2 across all the classes. 107

7.17 CA9 gene expression across all the classes. 108
7.18 CNA region alteration associated to CA9 across all the classes. 108
7.19 ELAVL1 gene expression across all the classes. 109
7.20 CNA region alteration associated to ELAVL1 across all the classes. 109
7.21 HtrA1 gene expression across all the classes. 110
7.22 CNA region alteration associated to HtrA1 across all the classes. 110
7.23 RNASET2 gene expression across all the classes. 110
7.24 CNA region alteration associated to RNASET2 across all the classes. 110
7.25 BID gene expression across all the classes. 111
7.26 CNA region alteration associated to BID across all the classes. 111
7.27 URI1 gene expression across all the classes. 111
7.28 CNA region alteration associated to URI1 across all the classes. 111

List of Tables

4.1 Samples obtained from GMQL queries. 49
4.2 Samples obtained from GMQL queries after considering only common barcodes between data types. 51

7.1 Concordance index computed for each data type. 83
7.2 Performance related to gene expression data, using the most significant feature sets. 87
7.3 Performance related to miRNA expression data, using the most significant feature sets. 89
7.4 Performance related to methylation data, using the most significant feature sets. 90
7.5 Performance related to CNA data, using the most significant feature sets. 92
7.6 Performance obtained by merging different data types, using the most significant feature sets. 93
7.7 Brief description of the 33 genes related to ovarian cancer in the CNA amplification regions suitable for distinguishing Resistant and Sensitive classes. 100
7.8 p_values obtained comparing the distribution of the CNA values of classes for the different genes using the K-S test. . 112
7.9 p_values obtained comparing the distribution in the classes of the expression of the different genes using the K-S test. . 112

Abstract

Ovarian cancer is the most lethal gynecologic cancer, causing annually a large number of deaths throughout the world. In particular, high-grade serous ovarian adenocarcinoma (HGS-OC) is the most common type of ovarian epithelial carcinomas and has the worst prognosis; it is a rapidly growing carcinoma, it is believed to have a tubal origin with a high chromosomal instability.
Poor prognosis in HGS-OC is largely related to chemoresistance: although patients usually respond to initial therapy and cytoreductive surgery followed by adjuvant chemotherapy with platinum and paclitaxel, $\sim$70% of those with advanced-stage ovarian cancer experience recurrence; in many cases, the disease becomes incurable mainly due to the development of drug resistance.
This thesis is born as cooperation with *Istitituto di Ricerche Farmacologiche Mario Negri*, particularly driven from biologist Sergio Marchini, who has the intuition that using CNA data would provide interesting results in terms of early diagnosis of the disease. The analysis present in this work is done in collaboration with a parallel thesis, which is presented to a different committee. The focus of this work is on the biological aspects and results of the research, while the emphasis of the other one by Sara Sansone is on the computational methods implemented to achieve the final goal.
One of HGS-OC peculiarity its the relapse timing of patients, that may be used as a predictor or as a label. By using it in the first way, we performed survival analysis to discriminate the drug-responsiveness of patients through their time to relapse. The results obtained with this model were quite poor, therefore we decided to use the relapse timing as a label to classify patients in classes, and specifically, we focused on therapy-resistant and therapy-sensitive patients, where the former ones are identified by relapse within a short interval of just six months since diagnosis.
To do so, many different data types have been integrated - as gene and miRNA expression data, DNA methylation data and CNA data - because of their involvement in ovarian cancer spreading and development; these data-types were downloaded from The Cancer Genome Atlas (TCGA) repository and the information needed to implement the models have been

extracted through the use of GenoMetric Query Language (GMQL), developed at Politecnico di Milano. Building an effective classifier through an integrative approach was the most difficult part of the thesis; based on this approach we were able to find significant results. These results are promising, and the integration of multiple data-types can be considered an innovation for this case of study. The most significant biological contribution was then integration of the genome portions that characterize the classifier, useful for explaining from a biological point of view the main distinctive features of resistant patients.
An important outcome was the identification of the Notch Signaling Pathway, considered to be one of the most important signalling pathways in drug-resistance tumor cells.

Sommario

Il carcinoma ovarico é il tumore ginecologico più letale, e ogni anno causa un numero elevatissimo di decessi in tutto il mondo. In particolare, l'adenocarcinoma ovarico sieroso di alto grado (HGS-OC) é il tipo più comune di carcinoma epiteliale ovarico e presenta la prognosi peggiore; é un carcinoma caratterizzato da una rapida crescita, che si ritiene abbia origine tubarica con un'acuta instabilità cromosomica.
La prognosi sfavorevole dell'HGS-OC é soprattutto dovuta alla chemioresistenza: sebbene le pazienti di solito rispondano alla terapia iniziale e alla chirurgia citoriduttiva seguita da chemioterapia adiuvante con platino e paclitaxel, circa il 70% di quelle affette da carcinoma ovarico in stadio avanzato presenta recidiva; in molti casi, la malattia diventa incurabile principalmente a causa dello sviluppo della resistenza ai farmaci.
Questa tesi nasce come collaborazione con l' *Istitituto di Ricerche Farmacologiche Mario Negri*, in particolare dall'intuizione di Sergio Marchini che l'uso dei dati di *CNA* fornirebbe risultati interessanti in termini di diagnosi precoce della malattia. L'analisi presente in questo lavoro viene effettuata in collaborazione con una tesi parallela, che viene presentata a un comitato diverso. Il focus di questo lavoro riguarda gli aspetti e i risultati biologici della ricerca condotta, mentre l'enfasi di quello presentato da Sara Sansone é posta sui metodi computazionali implementati per raggiungere l'obiettivo finale. Una delle peculiarità di questa malattia é la tempistica di ricaduta delle pazienti, che può essere utilizzata come predittore o come etichetta.
Usandola nel primo modo, ha reso possibile l'esecuzione di un'analisi di sopravvivenza per discriminare la reattività farmacologica delle pazienti; tuttavia, i risultati ottenuti con questo modello sono stati piuttosto scarsi, quindi si é deciso di utilizzare i tempi di ricaduta come etichetta per classificare i pazienti in classi e, in particolare, ci siamo concentrati su pazienti resistenti alla terapia e sensibili ad essa, in cui i primi sono identificati da recidiva entro un breve intervallo di soli sei mesi dalla diagnosi.
Per fare questo, sono stati integrati molti diversi tipi di dati - come dati di espressione genica e miRNA, dati di metilazione del DNA e dati CNA - per via del loro coinvolgimento nella diffusione e nello sviluppo del cancro ovarico; questi tipi di dati sono stati scaricati da *The Cancer Genome At-*

las (TCGA) e le informazioni necessarie per implementare i modelli sono state estratte attraverso l'uso di *GenoMetric Query Language* (GMQL), sviluppato presso il Politecnico di Milano. Costruire un classificatore efficace attraverso un approccio integrativo é stata la parte più difficile della tesi; sulla base di questo approccio siamo riusciti a trovare risultati significativi. Questi risultati sono promettenti e l'integrazione di più tipi di dato può essere considerata un'innovazione per questo caso di studio. Il contributo biologico più significativo é stato quindi l'integrazione delle porzioni del genoma che caratterizzano il classificatore, utile per spiegare da un punto di vista biologico le principali caratteristiche distintive dei pazienti resistenti.
Un risultato importante é stata l'identificazione del *Notch Signaling Pathway*, considerato uno dei *pathway* più rilevanti per quanto riguarda lo studio delle cellule tumorali resistenti ai farmaci:

Acknowledgments and greetings

Mantenere significa, letteralmente, tenere per mano; quindi rassicurare, proteggere e, a volte, anche stritolare.

Al professor Stefano Ceri, relatore di questa tesi, che ha scelto di affrontare la difficoltà di avermi come tesista, che non ha mollato la stretta nonostante a volte si fosse più che allentata, che ha continuato a spronarmi e a chiedermi di esserci, di fare, di insistere, anche quando non credevo più che insistere fosse una buona idea.
Alla professoressa Francesca Ieva, senza la quale molte brillanti idee non sarebbero nate, alla sua infinita pazienza e alla gentilezza che mi ha sempre dimostrato.
Ai correlatori Arif Çanakoğlu e Pietro Pinoli, per aver continuamente alzato lo stardard, fino al punto da vederlo troppo lontano e irraggiungibile, ed avermi così costretta a non mollare la presa.
Ai ricercatori Luca Beltrame e Sergio Marchini, senza i quali questa tesi non sarebbe mai esistita, per aver avuto la brillante idea su cui si basa tutta questa ricerca: il vostro supporto e le vostre competenze mi hanno reso più chiaro che tipo di ricercatore voglio diventare.

E poi ai dottoranti Gaia Ceddia, Michele Leone e Luca Nanni, che più di chiunque altro sono stati pronti a sostenermi in questo percorso, affrontando le mie difficoltà e non lasciandosi spaventare da esse, consigliandomi, correggendomi, insegnandomi: avete fatto ben più di quello che vi spettava, e di questo vi sono grata.

A Sara Sansone, senza la quale probabilmente questa tesi sarebbe rimasta per sempre incompiuta, dalla quale ho imparato più che da chiunque altro con cui sia stata a contatto in questi mesi, e che mi ha sempre fornito la stretta più gradevole di tutte, quella di chi cammina con te, e se si accorge di star andando troppo veloce, rallenta per non perderti nelle retrovie.

Alla professoressa Elisa Fasoli, che ha seguito in parte gli sviluppi di questa tesi, e come un faro mi ha indicato il percorso da seguire, aiutandomi a colmare le mie lacune - il mal di mare sperimentato spesso durante il viaggio.

Alle professoresse Alessandra Pedrocchi e Carmen Giordano: la prima per avermi indirizzata nella scelta della tesi, e la seconda per avermi convinta a crederci fino alla fine.
Grazie poi a tutti i miei professori e alle mie professoresse, ognuno dei quali ha contribuito alla mia crescita, sia dal lato umano che da quello didattico, in particolar modo la professoressa Sara Mantero e il professor Gabriele Dubini.

Ai miei strepitosi coinquilini Donato De Patre e Carmelo Abate: siete stati costretti a sopportare e supportare tutto quello che questa tesi che portare nella mia vita, ma lo avete fatto con eleganza.

Infine, alle due persone più brillanti, coraggiose e sicure che io conosca, Florin Carp e Matteo Lalli: il vostro impegno e la vostra dedizione mi sono stati di stimolo e d'esempio per portare a termine questo progetto.

Queste mani mi hanno stretta, stritolata e sostenuta, e dalle loro carezze e dai loro schiaffi ho imparato moltissimo. Vi ringrazio dunque, per il contribuito che avete dato a questa tesi: mi avete aiutata a mantenere una promessa che avevo fatto a me stessa.

Chapter 1

Introduction

1.1 General outline

Ovarian cancer is the deadliest gynecologic malignancy, with a 5-year survival rate of approximately 47%, a number that has remained constant over the past two decades. Early diagnosis improves survival, but unfortunately, only 15% of ovarian cancers are diagnosed at an early or localized stage. Most ovarian cancers are originally epithelial and treatment prioritizes surgery and cytoreduction followed by cytotoxic platinum and taxane chemotherapy. While most tumours will initially respond to this treatment, recurrence is likely to occur within 16 months for patients with advanced-stage disease.
In this thesis, a particular type of ovarian cancer has been taken into consideration: high-grade serous ovarian adenocarcinoma (HGS-OC), which is a tumor-type arising from the serous epithelial layer in the abdominopelvic cavity and it is mainly found in the ovary; these carcinomas make up the majority of ovarian cancer cases and they have the lowest survival rates.
Patients diagnosed with high-grade serous ovarian adenocarcinoma who received initial debulking surgery followed by platinum-based chemotherapy can experience highly variable clinical responses: a small percentage of women experience exceptional long-term survival, while others develop primary resistance to therapy and succumb to the disease in less than 32 months.
Despite the promising results achieved with cytoreductive surgery and platinum-based chemotherapy, eventually between 70% and 80% of advanced-stage ovarian cancer patients develop a resistance to the treatment, which is the only available therapy at the moment. Predicting the drug responsiveness at the time of diagnosis is essential for an improved outcome.
What makes a reliable cure particularly difficult to find, are the distinctive traits of this pathology: it is a *rapidly growing carcinoma* believed to have *tubal origin* with a *high chromosomal instability*: in many cancer types,

chromosome instability (CIN; or abnormal numbers of chromosomes) is associated with *aggressive tumours*, the acquisition of *multi-drug resistance* and *poor patient outcome*.
Another one of its peculiarities stands in the relapse timing of the patients affected by it. Indeed, patients can be recognized and differentiated into three classes, according to the time elapsed from the end of the first-line therapy to relapse:

- relapse within 6 months since the end of treatment: *resistant*;
- relapse after 12 months since the end of treatment: *sensitive short term*;
- relapse after 32 months since the end of treatment: *sensitive long term*.

Since the aim of this study is to build an efficient method for chemoresistance prediction, it is fundamental to characterize genomic patterns in intrinsic or acquired drug resistance, identifying a molecular signature that could be used to predict response to therapy at the time of diagnosis: this would lead to an improvement in the quality of life of patients resistant to therapy, who could not be subjected to it as could consequently not be affected by its side effects, in the search for alternative therapies.
The types of data that have been analyzed in this work are *gene expression* data, *miRNA expression* data and *DNA methylation* data: these have been chosen due to their involvement in ovarian cancer spreading and development;

- regarding gene expression data, a particular category of genes belonging to this category was selected to carry out the study: these are protein-coding genes, equal to about 2% of the DNA of the human genome that encodes for protein. The choice of using this gene-type is derived from their involvement in pathways designed to perform various functions, which if affected can lead to the appearance of cancer.
- miRNA expression data were chosen because of the causal role they have in tumorigenesis; in particular, in comparison to the normal ovary, miRNAs are aberrantly expressed in human ovarian cancer;
- moreover, *methylation imbalance* is characteristic of cancer and it is known that changes in DNA methylation can be used diagnostically and that they may predict resistance to treatment.

To further improve these results, *copy number alteration* (CNA) data, concerning regions of the genome presenting deletions and amplifications, have been merged with those already mentioned: the merging of these data

types has led to an overall improvement in the results. The processing of this data-type has been performed in a parallel study, [1].
It has been decided to use also CNA data because these regions are particularly interesting for their possible prognostic and diagnostic involvement: these alterations could be considered "early events", so they are potential predictors of chemoresistance; furthermore, being probably indicative and characterizing therapy-resistant patients, they can also be investigated to evaluate the development of resistance to therapy for patients initially sensitive to it; this intuition came from *Istituto di Ricerche Farmacologiche Mario Negri* researchers. A key part of the pipeline to build a model suitable for the intended purpose was the feature selection process: various sets of features were selected according to the significance of their expression, with the significance being evaluated in both statistical and biological terms.
Once the feature sets were identified, we proceeded in two distinct ways: first, a *survival analysis* was performed to understand if this method could predict the patients' relapse time; once this parameter was known, it would have been possible to identify those patients as belonging to a specific class, and thus it would have been possible to prevent treatment for therapy-resistant patients, who would not benefit from receiving the treatment.
Unfortunately, the results of this analysis were quite poor. For this reason, it was decided to leave this method aside and move on to the classification, from which good results have been achieved.
The data types used in this study were downloaded from The Cancer Genome Atlas (TCGA) repository, and the information needed to implement the model has been extracted using appropriate queries on GenoMetric Query Language (GMQL): TCGA is a landmark cancer genomics program that sequenced and molecularly characterized a huge amount of cases of primary cancer, and GMQL is a next-generation query language for querying next-generation sequencing data.
The results obtained through the use of features sets selected as significant for the distinction of therapy-sensitive patients compared to those resistant to it turned out to be good, especially as regards those obtained through the integration of all data types: this integration is an innovation with respect to the studies already in the literature.
A very important achieved result was the identification of Notch Signaling Pathway, which is known to be involved in drug-resistance; its regulation can induce drug sensitivity, leading to increased inhibition of cancer cell growth, invasion and metastasis.
With the purpose of improving the declared results in the future, a modus operandi could be to better investigate the involvement of specific miRNAs in the development of ovarian cancer, their functional bonds with methylated genes and CNA deletion and amplification regions of the genome.
Being aware of the fact that resistance to the only available therapy

gives patients no other treatment options, it would be interesting, having treated methylation data, to investigate how DNA methylation may be useful for innovative cancer treatment and investigate the potential of DNA methylation-based markers for diagnosis, prognosis, screening and prediction of drug resistance for ovarian cancer patients.

1.2 Thesis structure

The structure of the thesis is the following.

Chapter 2: Ovarian Cancer This chapter is entirely devoted to the discussion and in-depth analysis of as many aspects as possible concerning ovarian cancer, necessary to provide a detailed description of the disease and its development, up to the understanding of the particular case of study analyzed, that of the high-grade serous ovarian adenocarcinoma.

Chapter 3: Bionformatics and Tools In this chapter, we provide the information necessary to understand more in detail the reasoning behind the choice of the declared data-types; the tools and interfaces used for their treatment will be described.

Chapter 4: Datasets In this chapter, we describe the structure of the datasets used during the discussion, their composition and the origin of the data used to construct them.

Chapter 5: Computational Methods Foundation This chapter is dedicated to the detailed explanation of the methods used to proceed with the analysis.

Chapter 6: Computational Methods This chapter explains the final model thanks to which the results have been extrapolated.

Chapter 7: Results The best results obtained through the performed analyzes are extensively described in this chapter.

Chapter 8: Conclusions In this final chapter, we critically evaluate results of this thesis and compare them with those in the literature, mentioning related works and future perspectives.

Chapter 2

Ovarian Cancer

2.1 Overview

Ovarian cancer is the preeminent cause of gynecologic cancer death worldwide while constituting only 3% of all female cancers, [2]. As of 2018, ovarian cancer was the seventh most common cancer worldwide in women, with around 240,000 new cases. Ovarian cancer is the second most common malignancy after breast cancer in women over the age of 40, particularly in developed countries. When looking at all types of cancers, ovarian cancer is the eleventh most common type in women, the fifth leading cause of cancer-related death in women, and, as mentioned before, the most fatal gynecologic cancer, [3].
Due to the lack of specific symptoms in the early stage, 70% of cases are not diagnosed until cancer has reached an advanced stage, FIGO Stages IIB to IV (spread of tumour within the pelvis or elsewhere in the abdomen). Early detection of ovarian cancer reportedly increases the 5-year survival rate by up to 92%; however, the actual overall 5-year survival rate is only 15%-45%, [2].
Ovarian cancer is characterized by a late-stage presentation and poor prognosis. Women often present with silent symptoms as abdominal bloating and pain, causing delayed referral for workup of a malignancy. The risk factors of nulliparity, early-onset menarche, and late-onset menopause and the protective factors of increased parity, extended time lactating, and use of oral contraceptive pills imply that ovarian cancer risk is proportional to the number of ovulations in a life-time. Besides ovulation number, family history is a strong risk factor. A hereditary predisposition is responsible for 1424% of ovarian cancers, with the majority attributable to inherited mutations in the BRCA1 or BRCA2 genes, [4].
The majority of the deaths (70%) are of patients presenting with advanced-stage, high-grade serous ovarian cancer (HGS-OvCa). The standard treatment is aggressive surgery followed by platinumtaxane chemotherapy. After therapy, platinum-resistant cancer recurs in approximately 25% of pa-

tients within six months, and the overall five-year survival probability is 31%. Approximately 13% of HGS-OvCa is attributable to germline mutations in BRCA1/2 and a smaller percentage can be accounted for by other germline mutations. However, most ovarian cancer can be attributed to a growing number of somatic aberrations, [5].
Despite advancements in cancer research and treatment, survival statistics have remained largely unchanged for many years. A better understanding of the molecular pathogenesis of ovarian cancer is needed in order to develop new drug therapies or diagnostic biomarkers and elucidate the role of environmental exposures to the individual predisposition to the disease, [2].

2.2 Incidence and Prognosis

2.2.1 By Population/Age

The Centers for Disease Control and Prevention reports that white women have the highest prevalence, with 11.3 out of every 100,000 women being affected. The highest incidence per ethnicity after whites are Hispanics, Asian/Pacic Islander, African Americans, and American Indian/Alaska natives, whose incident rates are 9.8, 9.0, 8.5, and 7.9 per 100,000, respectively. Ovarian cancer is rare in young women, particularly under the age of 30; risk increases with age, with the occurrence spiking drastically after the age of 50, and average diagnosis between the ages of 50 and 70 years, [3].

2.2.2 Prognosis

Prognosis for those women that develop ovarian cancer is directly related to the stage of disease at the time of diagnosis. Those diagnosed at stage I, have a 5-year survival rate of 90%. In those with regional disease (meaning the disease has spread to adjacent tissues), 5-year survival rates drop to around 80%, and 25% in those with metastatic disease. Over the last 30 years, mortality rates from ovarian cancer have narrowly dropped, [3]

2.3 Classifications and Histopathology

Ovarian cancer has three main types: epithelial (most common), germ cell, and sex-cord-stromal, with the latter two comprising only about 5% of all ovarian cancers. There are four primary histologic subtypes of epithelial ovarian cancer: serous, endometrioid, mucinous, and clear cell. Serous tumours are categorized into two classications: high-grade serous carcinomas (HGSC) or low-grade serious carcinomas (LGSC). HGSCs account for 70% to 80% of all subtypes of epithelial ovarian cancer, while LGSCs account for less than 5%.

Endometrioid, mucinous, and clear cell subtypes account for 10%, 3%, and 10%, respectively. Figure 2.1 from [3] highlights key aspects of the types of epithelial ovarian cancers.

Types of epithelial ovarian cancer.

- Type I *v* type II tumor types
  - Type I
    - Less lethal than type II
    - Causes are continued ovulation cycles, inflammation, and endometriosis
    - Typically present as low-stage disease
    - Ovary origin
  - Type II
    - Associated with fatal outcomes
    - Diagnosed later
    - Linked to genetic mutations
    - Fallopian tube origin
- High-grade *v* low-grade serous tumors
  - High-grade serous tumors
    - 90% of all tumor types
    - More fatal prognosis
    - 10-year mortality rate of 70%
  - Low-grade serous tumors
    - 10% of all tumor types
    - Diagnosed at younger age
    - Better prognosis than high-grade serous tumors
- Endometrioid carcinomas
  - Originate from endometriosis
  - Good prognosis
- Clear cell carcinomas
  - 10% of epithelial ovarian cancers
  - Often diagnosed in early stages
  - If diagnosed late has poor prognosis
- Mucinous carcinoma
  - Least common type of epithelial ovarian cancer
  - Associated with metastasis from gastrointestinal tract
- Germ cell ovarian cancer
  - Rare
  - Make up only 3% of all ovarian cancers
  - Frequently diagnosed in younger women
  - Histologic type similar to that of germ cell tumors in the testes of men
- Sex cord-stromal ovarian cancer
  - Least common ovarian cancer
  - Less than 2% of all primary ovarian cancers
  - Rarely malignant
  - Usually diagnosed early
  - Smoking can decrease risk

Figure 2.1: Key aspects of the types of epithelial ovarian cancers

The ovarian cancer we are interested in is the epithelial one, so we are gonna discuss about it and its features.

2.4 *Epithelial Ovarian Cancer*

2.4.1 Background

Epithelial malignancies will typically have three point-of-origin sites: ovarian, tubal, or other epithelial sites in the pelvis. Epithelial ovarian ones (which account for the majority of ovarian cancers) are subdivided into

two categories: type I and type II tumours. Type I tumours, which are not as lethal as type II tumours, are considered to be caused by continual ovulation cycles, inflammation, and endometriosis. The appearance of endometriosis seems to enhance a woman risk of ovarian cancer and it is associated with 5% to 15% of all epithelial ovarian cancers. Many of these cancers manifest as low-stage diseases and usually have a more favourable outcome than types that are not associated with endometriosis. Unfortunately, type II tumours are commonly associated with fatal outcomes. These cancers are usually diagnosed later and are often connected to the genetic mutations of the BRCA genes and p53 mutations, another tumour-suppressing gene. One theory is that these tumours have moved from the fallopian tubes, the point of origin for these cancers, [3].
Epithelial ovarian cancer (EOC) represents the largest subgroup (90%) of ovarian cancers. EOCs are distinguished by histology, of which papillary serous is the most frequent (75%). Serous carcinomas are further partitioned into high-grade and low-grade tumour types. High-grade and low-grade serous carcinomas behave differently in terms of disease progression and response to platinum-based chemotherapy: low-grade serous carcinomas (LGSC) are often associated with borderline serous tumours, indicating that they may arise from precursor lesions. LGSCs tend to follow a more indolent course and are relatively platinum-resistant, compared to high-grade serous tumours which are often aggressive and can respond to platinum therapies.
High-grade serous carcinomas (HGSC) are the most common serious tumours. Over 90% of high-grade serous ovarian cancers harbour somatic P53 mutations. The majority of P53 mutations found in ovarian cancer are missense mutations, most of which occur in the DNA-binding domain of the protein. This is also the area through which P53 exerts its major function as a tumour suppressor, by trans-activating target genes regulating cell cycle progression, proliferation, and apoptosis. P53 mutations not only deplete wild-type P53 tumour-suppressive functions but can also act in a dominant-negative fashion on tetramerization of wild-type P53 with its target DNA sequence. In addition, the mutant P53 protein frequently acquires an oncogenic gain-of-function in these tumours leading to uncontrolled proliferation, increased metastatic potential, and higher risk of acquiring resistance to specific treatments, all through transcriptional regulation of genes important for tumorigenesis, cancer progression, and metastasis, [4].

2.4.2 The Morphological and Molecular Heterogeneity of Epithelial Ovarian Cancer

Ovarian Cancer is a heterogeneous disease composed of different types of tumours with widely differing clinicopathologic features and behaviour.

Based on a series of morphological and molecular genetic comparisons, a dualistic model that classifies various types of ovarian cancer into two groups designated type I and type II is proposed, [6].
Type I tumours are clinically inactive and usually present at a low stage. They exhibit a shared lineage between benign cystic neoplasms and the corresponding carcinomas often through an intermediate (borderline tumour) step, supporting the morphological continuum of tumour progression in these neoplasms. This stepwise sequence of events parallels the adenoma-carcinoma sequence that occurs in colorectal carcinoma. Type I tumours include low-grade serous, low-grade endometrioid, clear cell and mucinous carcinomas. In contrast to the clear-cut and distinctive morphological differences among type I tumours, the morphological variations among the type II tumours are more complex and as a result, there is considerable overlap in the diagnosis of these tumours by different pathologists. Type II tumours exhibit papillary, glandular and solid patterns and they are diagnosed as high-grade serous, high-grade endometrioid and undifferentiated carcinomas depending on the dominant pattern. Generally, most pathologists classify them as high-grade serous carcinomas even though they bear little resemblance to the tubal-type epithelium (the basis for typing a tumour as serous); many of those lacking distinctive serous or endometrioid features could be classified as high-grade adenocarcinoma. In addition to these neoplasms, malignant mixed mesodermal tumours (carcinosarcomas) are included in the type II category because they have epithelial components identical to the pure type II carcinomas.
Type II tumours are highly aggressive and almost always present in an advanced stage. Since they account for approximately 75% of all epithelial ovarian carcinomas and have relatively similar morphological features and a uniformly poor outcome, ovarian cancer has been erroneously regarded as a single disease. The morphological differences between type I and type II tumours are mirrored by marked differences in their molecular genetic features. As a group, type I tumours are genetically more stable than type II tumours and display specific mutations in the different histologic cell types. Thus, KRAS, BRAF, and ERBB2 mutations occur in approximately two-thirds of low-grade serous carcinomas whereas TP53 mutations are rare in these tumours.
High-grade serous carcinoma, the prototypic type II tumour, is characterized by very frequent TP53 mutations (80% of cases) and CCNE1 (encoding cyclin E1) amplification but rarely mutations that characterize most type I tumours such as KRAS, BRAF, ERBB2, PTEN, CTNNB1 and PIK3CA7. Although only a small number of malignant mixed mesodermal tumours have been analyzed molecularly, these few have been displaying a similar molecular genetic profile.
In summary, type I tumours, as a group, are genetically more stable than type II tumours and display a distinctive pattern of mutations that occur

in specific cell types (low-grade serous, low-grade endometrioid, clear cell and mucinous).
In contrast, the type II tumours (high-grade serous, high-grade endometrioid, malignant mixed mesodermal tumours and undifferentiated carcinomas) show greater morphological and molecular homogeneity, they are genetically unstable with a very high frequency of TP53 mutations. These findings imply that different types of ovarian carcinomas develop along different molecular pathways.y of TP53 mutations. These findings suggest that different types of ovarian carcinomas develop along different molecular pathways.

2.4.3 The Cell of Origin of Most Epithelial Ovarian Cancer is not Ovarian

The origin of ovarian cancer and the mechanisms by which cancer develops have been long discussed. The traditional view of ovarian carcinogenesis has been that the various tumours are all originated from the ovarian surface epithelium (mesothelium) and that following metaplastic changes lead to the development of the different cell types (serous, endometrioid, clear cell, mucinous and transitional cell [Brenner[1]]) which morphologically resemble the epithelia of the fallopian tube, endometrium, gastrointestinal tract or endocervix and urinary bladder, respectively. The healthy ovary, however, has no constituents that resemble these tumours. Furthermore, the cervix, endometrium and fallopian tubes are derived from the Müllerian ducts whereas the ovaries develop from mesodermal epithelium on the urogenital ridge separate from the Müllerian ducts. Therefore, an alternate theory proposes that tumours with a Müllerian phenotype (serous, endometrioid and clear cell) are derived from Müllerian-type tissue, not mesothelium. This Müllerian-type tissue (columnar epithelium, often ciliated) lines cysts located in para-tubal and para-ovarian locations that have been referred to collectively as the secondary Müllerian system. According to this theory, ovarian tumours develop from these cysts. As the tumour enlarges, it compresses and eventually obliterates ovarian tissue resulting in an adnexal tumour that appears to have arisen in the ovary. More recently another theory has been advanced which argues that *the majority of ovarian carcinomas, which are high-grade serous carcinomas, arise from high-grade intra-epithelial serous carcinomas in the fallopian tube which then spread to the ovary.*

Evaluation of these hypotheses is problematical because it is hard to construct experimental systems, to test their validity. Consequently, this

[1]Extremely rare tumor of testis and paratesticular regions composed of transitional type epithelium; usually occurs in ovary; also called transitional cell tumor. Aetiology unknown; may originate from Walthard rests within tunica vaginalis or transitional epithelial nests located in testicular / paratesticular locations, [7].

evaluation is based on a critical analysis of these studies in light of observations made in the course of pathologic examination of ovarian tumours, [6].

The theory of origin from ovarian surface epithelium (mesothelium) has a number of limitations Histologically, the single layer of commonly attenuated mesothelium overlying the ovaries bears no correspondence to serous, endometrioid, mucinous, clear cell or transitional (Brenner) carcinomas. To account for this apparent contradiction, it was proposed that the mesothelium overlying the ovary invaginates into the underlying stroma to form so-called cortical inclusion cysts. These cysts under the influence of local factors, possibly steroid hormones, experience a metaplastic change, which results in the mesothelium being converted to Müllerian-type epithelium. These inclusion cysts, with their newly acquired Müllerian phenotype, can then undergo malignant transformation resulting in carcinomas corresponding to the different cell types (serous, endometrioid and clear cell carcinomas). Although cortical inclusion cysts lined by ciliated (Müllerian-type epithelium) are frequently observed in the ovarian cortex, well documented examples of what can be interpreted as a transition from these cysts to carcinoma have not been reported.
The limitations of the secondary Müllerian system theory are that precursor lesions resembling serous carcinomas have rarely, if ever, been reported in paratubal and paraovarian cysts.

The most compelling evidence suggests that the vast majority of what appear to be, primary ovarian cancers, namely serous, endometrioid and clear cell carcinomas, *are derived from the fallopian tube and endometrium, not directly from the ovary.* Sporadic reports of tubal carcinoma and dysplasia had been reported in the past but in 2001 a group of Dutch investigators described these lesions, which closely resemble high-grade ovarian serous carcinoma, in women with a genetic predisposition to ovarian cancer. This was a surprising finding, since numerous studies that carefully examined the ovaries of women with a genetic predisposition to ovarian cancer never reported similar lesions. In addition, other studies of normal appearing ovaries contralateral to sporadic (non-hereditary) unilateral ovarian carcinomas had never identified a convincing precursor lesion. These latter studies reported a number of morphological changes in grossly normal appearing ovaries, such as an increased number of inclusion cysts, surface papillae, cortical inclusions, including some displaying minor degrees of atypia. The data have been conflicting, some studies reporting a significant difference of these changes in cases versus controls and other studies reporting no difference. In any event, none of these changes, even remotely, resembles high-grade serous carcinoma. It was precisely because of a lack of convincing precursor lesions that the *de novo* hypothesis was proposed.

In hindsight, because it was assumed that precursors of ovarian carcinoma would logically be in the ovaries, the fallopian tubes were not carefully examined. Subsequent studies in which fallopian tubes were more carefully examined confirmed that in situ and small, early invasive tubal carcinomas occurred in women with a genetic predisposition for the development of ovarian cancer. This led to fallopian tube carcinoma being included as part of the cancer spectrum associated with inherited BRCA mutations. It was subsequently proposed that a proportion of ovarian carcinomas might develop as a result of implantation of malignant cells from the tubal carcinoma to the ovary. The next important step linking what had been termed tubal intra-epithelial carcinoma (TIC) and subsequently serous tubal intra-epithelial carcinoma (STIC) with ovarian carcinoma was the observation that over 70% of sporadic (non-hereditary) ovarian and peritoneal high-grade serous carcinomas demonstrated mucosal tubal involvement including STICs. This observation gave support to the proposal that STICs[2] may be the source of ovarian high-grade serous carcinoma in both women with hereditary mutations in BRCA as well as women who did not have a known genetic predisposition for ovarian cancer. Although it can be argued that mucosal tubal involvement could represent secondary spread from an ovarian carcinoma present in the same specimen, the presence of focal non-contiguous intra-epithelial lesions (STICs) would be an unusual manifestation of metastasis. Furthermore, the identification of STICs in prophylactic specimens from women, with a hereditary predisposition to ovarian cancer in which complete microscopic evaluation of the fallopian tubes and ovaries failed to identify invasive carcinoma in these organs, lends additional support to the concept that the serous neoplastic process may well begin in the fallopian tube rather than the ovary. Further support for this argument is the finding that nearly all STICs overexpress p53 similar to high-grade serous carcinoma, Figure 2.2.

Laser capture microdissection studies of these lesions have demonstrated that they harbor mutated TP53. In addition, STICs associated with a concomitant ovarian carcinoma share not only morphological features but also identical TP53 mutations indicating a clonal relationship between them. Adnexal malignant mixed mesodermal tumours (another type II tumor) have also been associated with STICs supporting the existence of a common precursor lesion for type II tumours. Further evidence, implicating the fallopian tube rather than ovarian surface epithelium as the site of origin of serous neoplasms, comes from a gene profiling study showing that the gene expression profile of high-grade serous carcinoma is more closely related to the fallopian tube than to ovarian surface epithelium. In addition high-grade serous carcinomas express PAX8, a Müllerian marker, but not calretinin, a mesothelial marker(Figure 2.3).

[2]STICs are almost always detected in the fimbria.

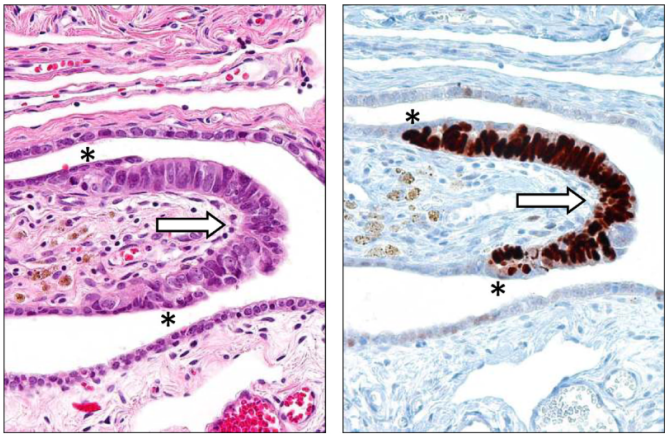


Figure 2.2: Serous tubal intra-epithelial carcinoma (STIC). A. High magnification. Hematoxylin and eosin stain. B. Immunohistochemical stain for p53. An asterisk defines the boundary of the lesion.

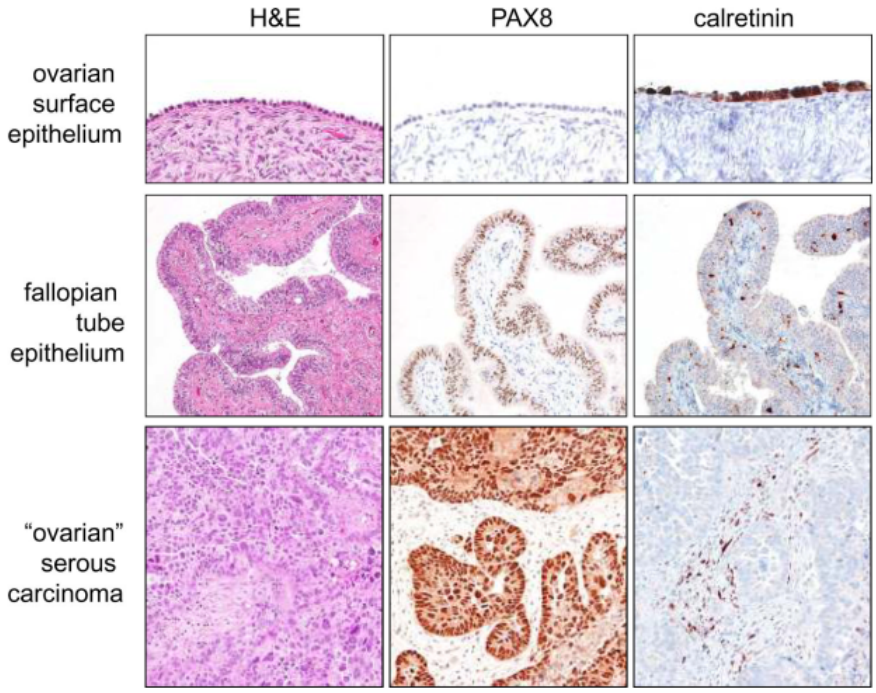


Figure 2.3: Comparison of the immunohistochemical staining pattern for ovarian surface epithelium (mesothelium), normal fallopian tube epithelium, and high-grade serous carcinoma. PAX8 is a marker of Müllerian-type epithelium such as fallopian tube epithelium and calretinin is a marker of mesothelium.

Generally, before a carcinoma acquires the ability to metastasize it must first invade and gain access to blood vessels or lymphatics.

Tubal intra-epithelial carcinomas are similar morphologically and immunohistochemically to endometrial intra-epithelial carcinomas, which are regarded as precursors or early forms of uterine serous carcinoma. These lesions have also been termed uterine surface serous carcinomas. They have been shown to disseminate throughout the peritoneal cavity presumably by passage of malignant cells through the fallopian tube without requisite myometrial invasion. The cells that comprise both endometrial and tubal intra-epithelial carcinomas are highly anaplastic and morphologically identical to high-grade serous carcinoma. The lesions form papillary tufts and the constituent cells are loosely cohesive.

In studies of ovarian and primary peritoneal high-grade serous carcinomas, in which the entire fallopian tubes were carefully sectioned, mucosal involvement of the tube, including STICs, were identified in approximately 70% of cases. The question arises as to the source of the remaining ovarian carcinomas that lack evidence of tubal involvement. There are a number of possible explanations; first, despite thorough sectioning, a small STIC could have been missed (unpublished data); second, on occasion *high-grade serous carcinomas are intimately associated with serous borderline tumours and low-grade serous carcinomas. In these cases the high-grade tumours have had KRAS mutations identical to those in the serous borderline tumours and lacked TP53 mutations. This finding suggests that some high-grade serous carcinomas arise from low-grade serous tumours and not by the usual (type II) pathway that begins with a TP53 mutation.* Third, clear-cut mucosal tubal involvement could have been obscured by overgrowth of the pelvic carcinoma. Fourth, the fimbria of the fallopian tube is normally in intimate contact with the ovarian surface at the time of ovulation. It is conceivable that when the ovarian surface epithelium is disrupted at the time of ovulation, normal tubal epithelial cells from the fimbria may be dislodged and implanted in the ovary to form an inclusion cyst from which a high-grade serous carcinoma could develop, (Figure 2.4). Evidence to support this notion is the observation that fallopian tube epithelial cells are easily obtained for culture by flushing the fallopian tube. This mechanism could also explain the development of endosalpingiosis, a lesion composed of glands and papillary structures lined by tubal-type epithelium that is found on peritoneal surfaces in the pelvis, omentum and beneath the capsule of pelvic and para-aortic lymph nodes. Endosalpingiosis is frequently found in association with low-grade serous tumours and has been viewed as a possible precursor of these tumours. Finally, the possibility that some high-grade serous carcinomas arise in cortical inclusion cysts, as a metaplastic process from the ovarian surface epithelium rather than from implantation of normal fallopian tube epithelium, cannot be entirely dismissed.

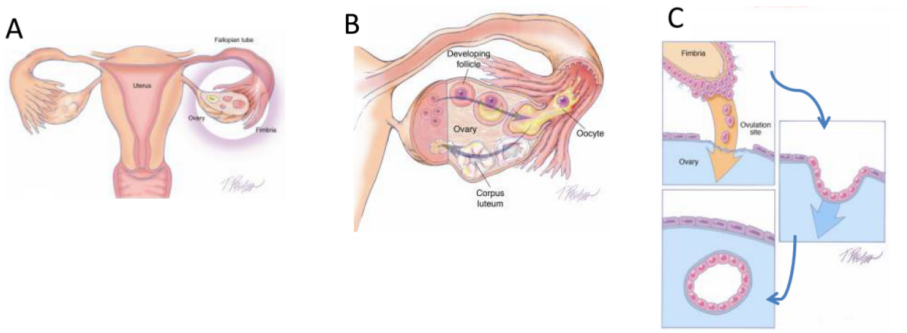


Figure 2.4: Transfer of normal tubal epithelium to the ovary. A. Anatomical relationship of fallopian tube to the ovary at the time of ovulation. The fimbria envelops the ovary. B. Ovulation. The ovarian surface ruptures with expulsion and transfer of the oocyte to the fimbria. The fimbria is in intimate contact with the ovary at the site of rupture. C. Tubal epithelial cells from the fimbria are dislodged and implant on the denuded surface of the ovary resulting in the formation of an inclusion cyst.

Direct implantation of tubal epithelium into the ovary to form an inclusion cyst, which in turn is the site of origin of ovarian serous carcinoma, is an attractive alternative theory to that of metaplasia from the surface epithelium (mesothelium). Implantation of fallopian tube epithelium from the fimbria at the time of ovulation, when the surface epithelium is disrupted, can explain the derivation of low- and high-grade serous carcinomas. In the case of a low-grade serous carcinoma the process slowly develops from a serous cystadenoma and then a serous borderline tumor after a KRAS or BRAF mutation whereas in the case of a high-grade serous carcinoma the process evolves rapidly, presumably from a cortical inclusion cyst after a TP53 mutation with the development of an intraepithelial carcinoma as an intermediate step. According to this view both low- and high-grade serous carcinomas are ultimately of tubal (Müllerian) origin and in a sense the ovary is involved secondarily, Figure 2.5.

In summary, none of the existing theories adequately reconcile all aspects of ovarian carcinogenesis. All of them have something to offer in explaining the development of ovarian carcinomas but none are all inclusive. It does appear that the vast majority of what have been thought to be primary epithelial ovarian and primary peritoneal carcinomas are, in fact, secondary. Thus, the most persuasive data support the view that serous tumours develop from the fimbriated portion of the fallopian tube. The concept, that the majority of epithelial ovarian carcinomas originates outside the ovary and involves it secondarily, has emerged only recently because in the past the default diagnosis of carcinomas involving the pelvis and abdomen was that they were ovarian. A carcinoma was classified as tubal in origin only when the bulk of tumor involved the fallopian tube rather than the ovary and there was evidence of an intra-epithelial (in situ)

tubal carcinoma. A diagnosis of primary peritoneal carcinoma is even more restrictive. Even with extensive tumor involving the peritoneum, omentum and other abdominal organs, a carcinoma is classified as primary ovarian if there is as little as 5 mm of tumor involving the ovaries. Thus, there has been an inherent bias in classifying pelvic tumours as being ovarian in origin.
Although the data, suggesting that epithelial ovarian carcinoma arises in extra-ovarian sites and involves the ovaries secondarily, are compelling, serous neoplasms (low- and high-grade) involve the ovaries and other pelvic and abdominal organs, such as the omentum and mesentery, much more extensively than the fallopian tubes.

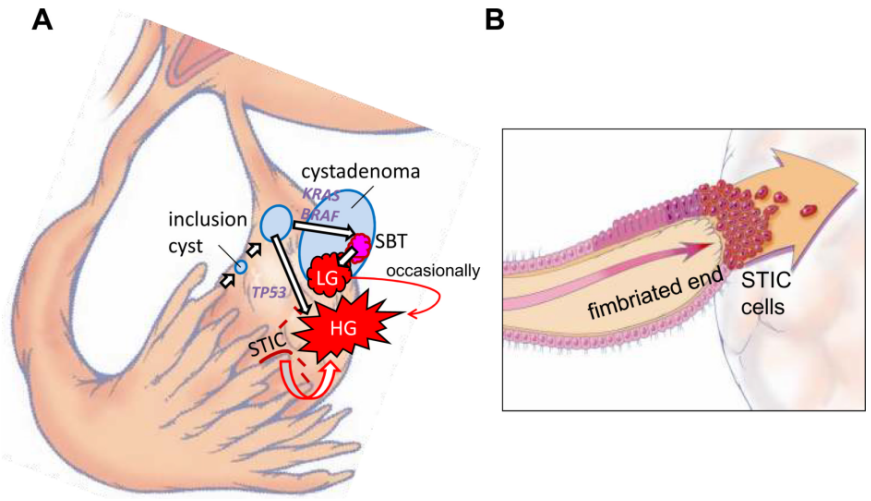


Figure 2.5: Proposed development of low-grade (LG) and high-grade (HG) serous carcinoma. A. One mechanism involves normal tubal epithelium that is shed from the fimbria, which implants on the ovary to form an inclusion cyst. Depending on whether there is a mutation of KRAS/BRAF/ERRB2 or TP53 a low-grade or high-grade serous carcinoma develops respectively. Low-grade serous carcinoma often develops from a serous borderline tumor (SBT), which in turn arises from a serous cystadenoma. Another mechanism involves exfoliation of malignant cells from a serous tubal intra-epithelial carcinoma (STIC) that implants on the ovarian surface resulting in the development of a high-grade serous carcinoma. B. A schematic representation of direct dissemination or shedding of STIC cells onto the ovarian surface where the carcinoma cells ultimately establish a tumor mass that is presumably arising from the ovary. Of note there may be stages of tumor progression that precede the formation of a STIC.

2.5 LGSC versus HGSC

2.5.1 Background

Ovarian serous carcinoma has traditionally been graded as well-, moderately, and poorly differentiated, suggesting that it is a homogeneous disease from the standpoint of pathogenesis. Multiple different grading systems have been used with variable results, including the FIGO[3] system based on percentage of solid architecture the WHO[4] system based on an impression of architecture and cytologic features, the Gynecologic Oncology Group (GOG) system based on histologic type, a system based on a combination of mitotic index and volume percentage of epithelium, a system based on presence/amount of hyperchromatic giant nuclei and solid or cribriform architecture, and a grading index based on a mean of the individual scores for architectural pattern, nuclear pleomorphism, nucleoli, nuclear-to-cytoplasmic ratio, mitotic index, pattern of invasion, capsule penetration, and vascular invasion. A 3-tier grading scheme that has gained much attention over the past several years is the universal grading system, which is also referred to as the Silverberg grade. In this system, points are assigned for each of 3 components: architecture (glandular, papillary, or solid), degree of nuclear atypia, and mitotic index. The points for each component are added, resulting in a total score which determines the grade, analogous to that used for breast carcinoma. More recently, a 2-tier grading system specifically for serous carcinoma, in which tumours are subdivided into low-grade and high-grade, has been proposed. A -tier grading system is easy to apply, reproducible, and based on underlying molecular biologic differences between low-grade and high-grade tumours, [2].

2.5.2 Behaviour

The few studies that have compared outcome between both types of serous carcinomas using the 2-tier system have shown that patients with low-grade tumours have better survival. In the study by Malpica *et al*[8], the 2-tier grading system was found to be of independent prognostic significance upon multivariate analysis, and the survival of patients with low-grade tumours was significantly higher than with high-grade tumours. In that study, death due to disease was more rapid with high-grade carcinoma. The median survival was 1.7 years for patients with high-grade

[3]The International Federation of Gynecology and Obstetrics is an international organization that links about 125 international professional societies of Obstetricians and Gynecologists. In 2011 FIGO recognized two systems designed to aid research, education, and clinical care of women with abnormal uterine bleeding (AUB) in the reproductive years.

[4]WHO's primary role is to direct international health within the United Nations' system and to lead partners in global health responses.

tumours compared to 4.2 years for women with low-grade tumours. Furthermore, in a large clinical study of only low-grade serous carcinoma, the median overall survival with stage III or IV disease was 6.8 years. Persistent disease after primary chemotherapy was the only variable associated with shorter survival time. With high-grade serous carcinoma, survival beyond 5 years is unusual, but survival over 10 years can be seen in a subset of low-grade serous carcinomas.
Few studies have compared survival using the 2-tier vs. 3-tier grading systems, where serous carcinomas were graded using the 2-tier low-grade/high-grade, 3-tier universal, and 3-tier FIGO systems; all 3 grading systems showed statistically significant prediction of survival, [8].
In view of its simplicity in application and excellent reproducibility, it is recommended to use the 2-tier system in routine practice, [9].

2.5.3 Pathogenesis

LGSC Low-grade serous carcinoma (invasive micropapillary serous carcinoma [MPSC]), has been hypothesized to arise from a serous cystadenoma (Figure 2.6) or adenofibroma which progresses to an atypical proliferative serous tumor (APST, atypical serous borderline tumor, in Figure 2.7, to non-invasive MPSC (micropapillary serous borderline tumor, in Figure 2.8), and then to invasive MPSC in a slow step-wise fashion. This has been described as the Type I pathway and is supported by several morphologic observations. First, invasive low grade serous carcinoma is associated with non-invasive MPSCs in over three fourths of cases, Figure 2.9. Second, in occasional tumours, the level of differentiation of the non-invasive tumor is intermediate between APST and non-invasive MPSC, suggesting a morphologically intermediate step. Third, true early invasion in an APST or non-invasive MPSC resembles low-grade serous carcinoma (Figure 2.10). Fourth, in several studies, non-invasive MPSCs have a higher frequency of invasive implants (Figure 2.11) compared with APST, and these implants are histologically identical to low-grade serous carcinoma.

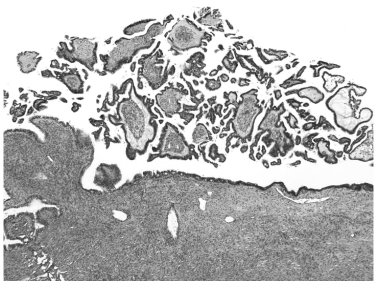


Figure 2.6: Small APST arisin in a serous cystadenoma.

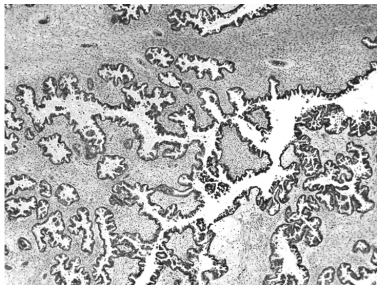


Figure 2.7: APST.

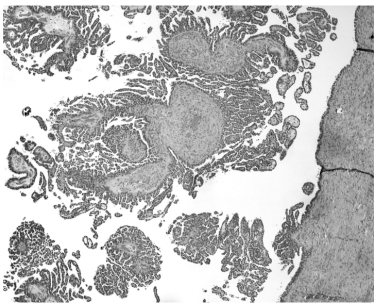


Figure 2.8: MPSC.

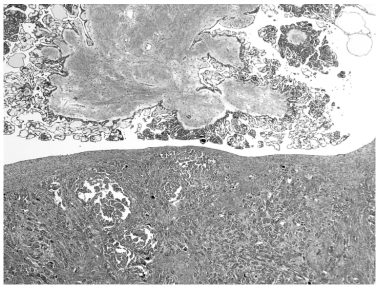


Figure 2.9: Tumor progression in MPSC.

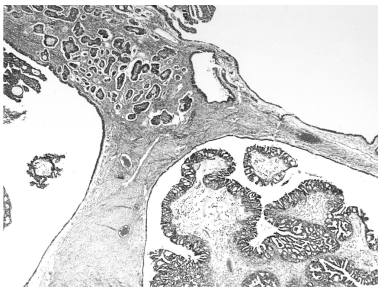


Figure 2.10: Microinvasion in MPSC.

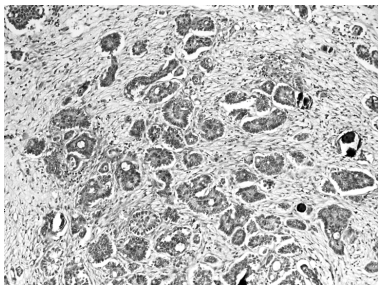


Figure 2.11: Invasive implants.

HGSC Much less is known about the pathogenesis of ovarian high-grade serous carcinoma compared with low-grade serous carcinoma; mutations of KRAS, BRAF, or ERBB2 occur very infrequently in high-grade carcinoma. In contrast, TP53 mutation occurs in 80% of high-grade tumours, and up-regulation and down-regulation of numerous other genes and various DNA copy number changes have been described.

Genome-wide analysis of DNA copy number alterations has demonstrated significant numbers of amplifications and deletions, including homozygous deletions. Among homozygous deletions, loci containing Rb1, CDKN2A/B, CSMD1, and DOCK4 were most common and except for the CDKN2A/B region, these homozygous deletions were not present in either serous borderline tumours or low-grade serous carcinomas.

The identification of the precursor lesion of high-grade serous carcinoma has been a topic of interest for decades. Since high-grade serous carcinoma nearly always presents with high-stage disease, the development of this tumor is thought to be rapid, and its origin has traditionally been presumed to be from surface epithelium or epithelial inclusions in the ovary. In an effort to detect putative precursors, researchers have focused on ovaries of women with a family history of ovarian cancer and

women with BRCA mutations. Increased p53 immuno-positivity has been noted in the epithelium of ovaries from these women compared with controls, but these findings have not been confirmed in other studies. Mutations and/or loss of heterozygosity of TP53 have been identified in early carcinomas and epithelial inclusions of the ovary, including identical mutations in the epithelium and adjacent carcinoma in the same cases. These molecular findings support to the role of TP53 mutation as an early event in the pathogenesis of high-grade serous carcinoma and that the origin for some tumours is the surface epithelium or epithelial inclusions of the ovary. Parenthetically, 10% of ovarian carcinomas are hereditary. Of the hereditary carcinomas, most are related to BRCA mutations, which appear to play a role in the pathogenesis of ovarian carcinoma in this subset of tumours. The vast majority of BRCA-related hereditary ovarian tumours are high-grade serous carcinoma.

Some studies report that:

- an incidental ovarian carcinoma in situ from a woman with a germline mutation of BRCA1 exhibited loss of heterozygosity of this gene;

- loss of heterozygosity of BRCA has been demonstrated in epithelial inclusions/surface epithelium in ovaries from prophylactic oophorectomy specimens;

- loss of heterozygosity has also been reported in invasive carcinoma and adjacent epithelium in stage I ovarian carcinomas from women with BRCA germline mutations;

these evidences suggest that loss of heterozygosity of BRCA is an early event in high-grade serous carcinoma for tumours with germline mutations. Similar to TP53, BRCA appears to function as a tumor suppressor gene. Thus, patients who inherit a germline mutation of BRCA and with somatic loss of the wild-type allele, develop carcinoma. The exact interaction between mutations of BRCA and TP53 in ovarian carcinoma is unclear. In addition to germline mutations, other molecular alterations leading to inactivation of BRCA include somatic mutation, promoter hypermethylation, and isolated loss of hetrerozygosity. These putative precursor lesions are detected in inclusions in the ovary or ovarian surface epithelium and they are characterized by tubal-type epithelium with varying degrees of cytologic atypia that have been termed dysplasia/carcinoma in situ. These findings, although they suggest that a morphologically identifiable precursor of high-grade serous carcinoma may exist in the ovary, they are very rarely detected, and, therefore, it has been suggested that these tumours arise *de novo*.

Recently, attention has been drawn to a lesion in the fallopian tube that has the cytologic appearance of high-grade serous carcinoma of the ovary and tubal intraepithelial carcinoma (TIC) has been designated, Figure 2.12, [9].

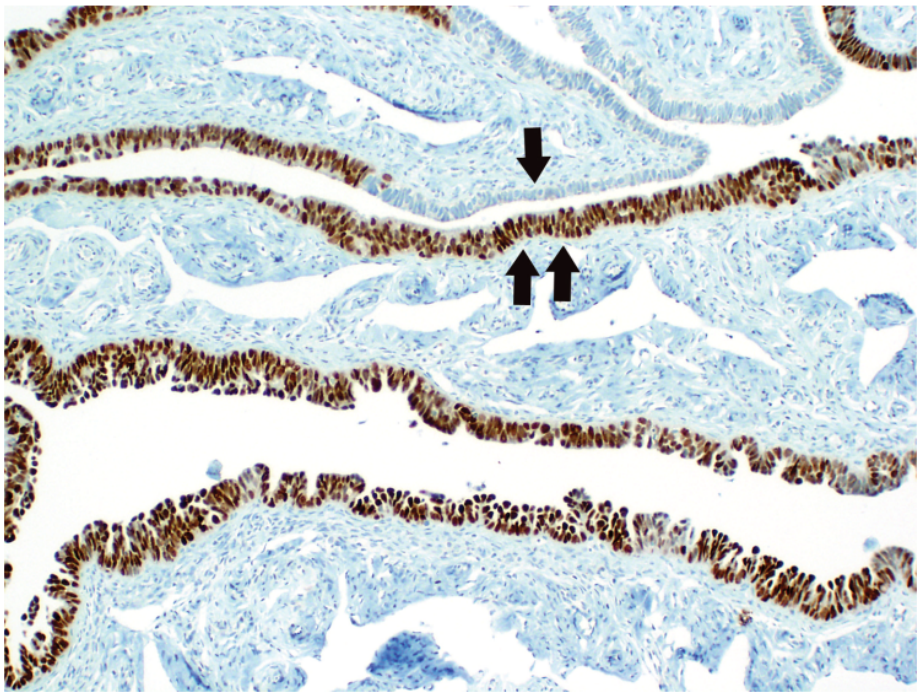


Figure 2.12: Tubal intraepithelial carcinoma (TIC). The epithelium of the fallopian tube mucosa with TIC is thicker compared with normal mucosa (upper center).

These lesions are almost always detected in the fimbriated end of the fallopian tube. The fimbriated end is in close proximity to the ovarian surface and it has been suggested that the tube is the origin of a subset of ovarian high-grade serous carcinomas. This is supported by the following:

- early serous carcinomas in prophylactic bilateral salpingo-oophorectomy specimens from women with BRCA mutations can be detected in the tube, especially the fimbriated end, in the absence of an ovarian tumor;

- identical TP53 mutations have been reported in TIC and synchronous ovarian high-grade serous carcinomas;

- identical TP53 mutations have been reported in TICs and in small foci of histologically normal tubal epithelium that diffusely expresses p53, which has been termed p53 signature. It has been suggested

that p53 signatures are precursors of TICs which in turn precede the development of high grade serous carcinoma.

Moreover, it has been proposed that when there is a synchronous TIC and ovarian high-grade serous carcinoma, the fallopian tube is the primary site of origin for the ovarian tumor.

The morphologic and molecular observations suggest that possibly half of ovarian high-grade serous carcinomas may be of tubal origin. In the other half of tumours, primary origin may have been ovarian or peritoneal. It should be noted that the criteria for distinction of primary ovarian vs. peritoneal origin are quite arbitrary. *Bona fide* well-defined precursor lesions in the ovary are rare and have not been identified in the peritoneum.

In summary, the pathogenesis of high-grade serous carcinoma (Type II pathway) is characterized by:

1. rapid development from what are now believed to be intraepithelial carcinomas very likely of tubal origin;

2. TP53 mutations;

3. a high level of chromosomal instability;

4. in hereditary tumours, BRCA germline mutations;

5. absence of mutations of KRAS, BRAF, or ERBB2.

2.5.4 Clinical aspects, histologic features and selected diagnostic problems

We will now leave the LGSC aside so that we can better concentrate on HGSC.

HGSC High-grade serous carcinomas may exhibit mixtures of papillary, glandular (Figure 2.13), nested, and diffuse/solid growth patterns (Figure 2.14) although any component may predominate in a given tumor. The papillae tend to be large and complex. The epithelium lining the papillae is usually stratified with an irregular slit-like configuration (Figure 2.15). Although a micropapillary growth pattern is typical of low-grade serous carcinoma, it should be emphasized that occasional high-grade carcinomas can also exhibit this architecture (Figure 2.16); however, they have high-grade nuclei and typically have an admixed solid growth pattern. The latter would be unusual for low-grade tumours.

Figure 2.13: Glandular pattern.

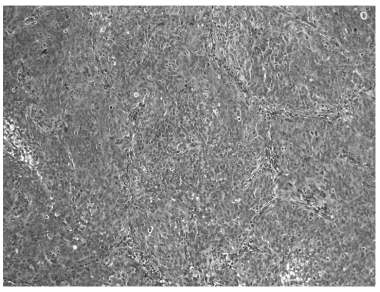


Figure 2.14: Diffuse, solid pattern.

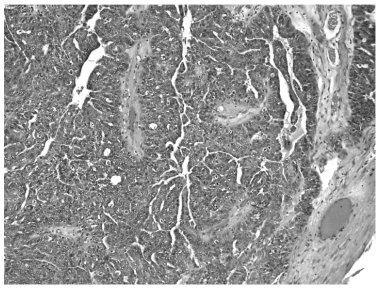


Figure 2.15: Typical papillary pattern showing irregular slit-like spaces.

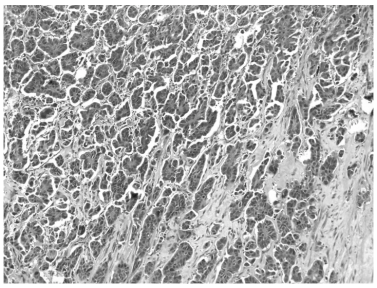


Figure 2.16: Micropapillary pattern.

The glands in high-grade serous carcinoma may be round and simple or complex with irregular slit-like spaces (Figure 2.13). Some tumours may have such extensive solid architecture with diffuse sheets of neoplastic epithelium that a careful search for a glandular or papillary component may be necessary for distinction from undifferentiated carcinoma. Obvious destructive stromal invasion is generally present, but some neoplasms may be predominantly intracystic and, therefore, misdiagnosed as APST/non-invasive MPSC. The presence of high-grade nuclei excludes that possibility. Necrosis is common in high-grade serous carcinoma. Psammoma bodies can be seen but they are typically less frequent compared with low-grade serous carcinoma.

The neoplastic epithelial cells are heterogeneous, they may be a mixture of low-cuboidal, columnar and hobnail shapes. Typically, there is marked variation in size and shape. The nuclear-to-cytoplasmic ratios are generally high, but at times, abundant eosinophilic cytoplasm is present. Most tumours have variable combinations of enlarged round or oval nuclei, irregular nuclear membranes, irregular chromatin distribution, hyperchromasia, large nucleoli and abundant mitotic figures, including atypical forms, Figure 2.17.

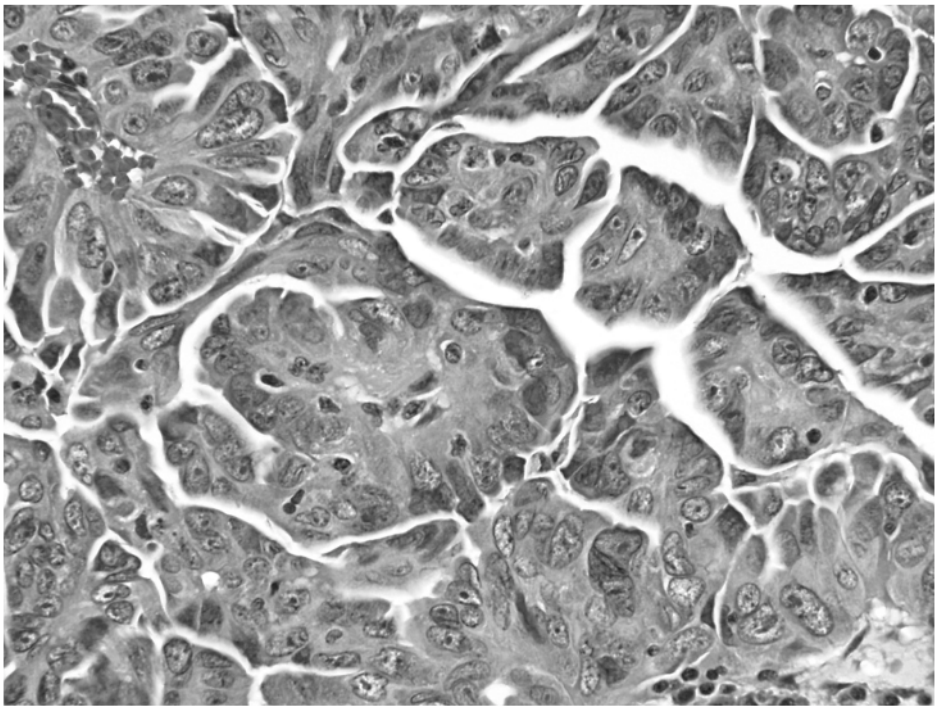


Figure 2.17: The nuclei of high-grade serous carcinoma are larger with greater pleomorphism and larger nucleoli in respect to LGSC.

At times high-grade carcinomas can display an appearance, mimicking endometrioid carcinoma (Figure 2.18). When the glands have irregular serrated luminal contours, large complex papillae lined by stratified epithelium with irregular slit-like patterns, hobnail cells, bizarre tumor giant cells and psammoma bodies, a serous carcinoma is favored. In contrast, high-grade endometrioid carcinoma is favoured by tumours with peripheral palisading of solid islands and nests, squamous metaplasia, or a background of atypical proliferative (borderline) endometrioid tumor or endometriosis. Immunohistochemical staining for WT-1 has been advocated as useful for this differential diagnosis, but in our experience it is not reliable. At times, distinction of high-grade serous carcinoma from high-grade endometrioid carcinoma (FIGO grade 2 or 3) is not possible and classification as high-grade adenocarcinoma, not otherwise specified with a descriptive comment is necessary.

Some tumours may contain cells with clear cytoplasm. If the tubulocystic and papillary patterns characteristic of clear cell carcinoma are not present, these tumours should not be interpreted as clear cell carcinoma, Figure 2.19. When the epithelium lining the surface of large rounded papillae is smooth, a transitional cell carcinoma-like appearance can be produced and may be mistaken for ovarian transitional cell carcinoma (TCC), Figure 2.20. Opinions vary among gynecologic pathologists as to

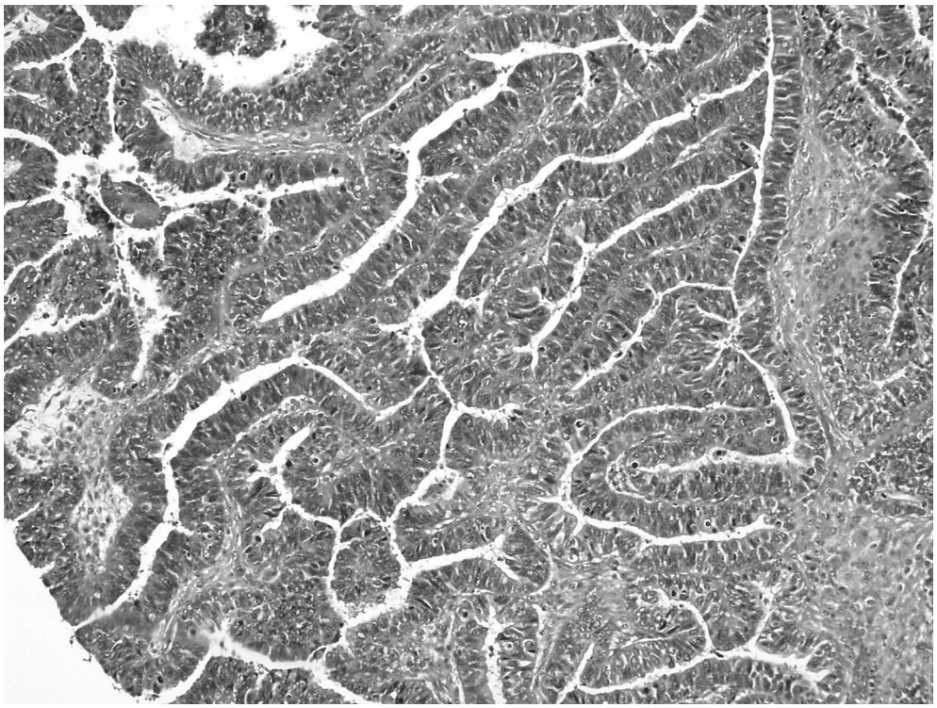


Figure 2.18: Endometrioid carcinoma-like pattern.

whether pure TCC of the ovary is a distinctive entity or simply a TCC-like pattern of high-grade serous carcinoma. At present, the jury is still out. Glandular differentiation has been described in transitional cell carcinoma; however, a diagnosis of serous carcinoma is favored when the glands merge with complex, branching papillae exhibiting epithelial tufting and solid nests surrounded by a space and irregular slit-like spaces are present. Also, psammoma bodies are more typical of serous carcinoma. Immunohistochemistry is not helpful as WT-1 expression has been described in both serous and transitional cell carcinomas.

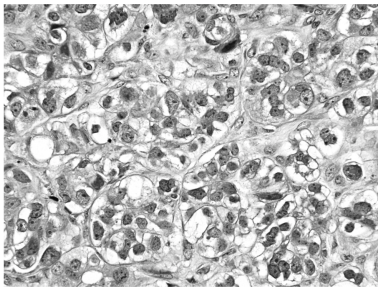


Figure 2.19: Serous carcinoma with clear cytoplasm.

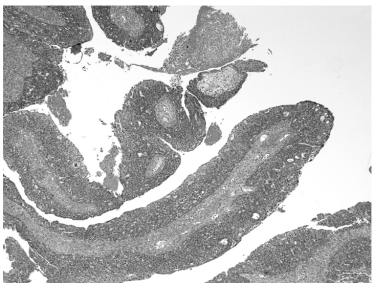


Figure 2.20: Transitional cell carcinoma-like pattern.

Rarely, microcystic or signet ring cell-like change can be seen in high-grade and also in low- grade serous carcinoma(Figure 2.21). Microcystic change in such tumours is produced by the presence of back-to-back cells with signet ring change and can simulate the reticular pattern of yolk sac tumor. The combination of older age, bilaterality, large papillae lined by complex and stratified epithelium, glands with irregular slit-like spaces and psammoma bodies favor serous carcinoma. On the other hand, the combination of younger age, unilaterality, microcystic patterns which blend with other classic patterns of yolk sac tumor, such as Schiller-Duval bodies, polyvesicular-vitelline, intestinal and myxoid patterns, and hyaline globules favor yolk sac tumor. An elevated serum AFP level is characteristic of yolk sac tumor. Immunohistochemistry may be helpful in that expression of WT-1, ER, PR, CK7, and EMA are more frequent in serous carcinoma while expression of AFP and absence of other markers are more typical of yolk sac tumor.
The distinction of low-grade from high-grade serous carinoma is based on nuclear features.
In most tumours, the nuclei of low-grade and high-grade serous carcinomas are typically grade 1 and grade 3, respectively, in a 3-tier system; thus, the diagnosis in the vast majority of tumours is straightforward. Some tumours (approximately 4% of serous carcinomas), however, exhibit nuclear features that are intermediate between low-grade and high-grade. These grade 2 nuclei are larger and have coarser chromatin, more mitotic activity, and larger nucleoli than grade 1 nuclei. They are also relatively uniform, smaller and less pleomorphic with less coarse chromatin than grade 3 nuclei. Thus, classification of these tumours with intermediate grade nuclei as low-grade versus high-grade serous carcinoma will be difficult.

High-grade serous carcinomas are architecturally heterogeneous: they correspond to moderately differentiated and poorly differentiated grades in 3-tier grading systems because some are predominantly papillary or glandular while others are mostly solid. However, they do not appear to be different from a molecular and in vitro drug resistance standpoint. In a study of high-grade serous carcinomas, in which moderately differentiated and poorly differentiated were compared, there were no significant differences in the frequency of TP53 mutation or extreme drug resistance for each of chemotherapeutic agents. In addition, the survival for patients with grades 2 and 3 serous carcinomas, using the universal grading system, is closer to each other compared with survival for patients with grade 1 and 2 tumours. These biologic and clinical findings suggest that moderately and poorly differentiated tumours can be combined into a single category, justifying the use of a 2-tier rather than a 3-tier grading system.

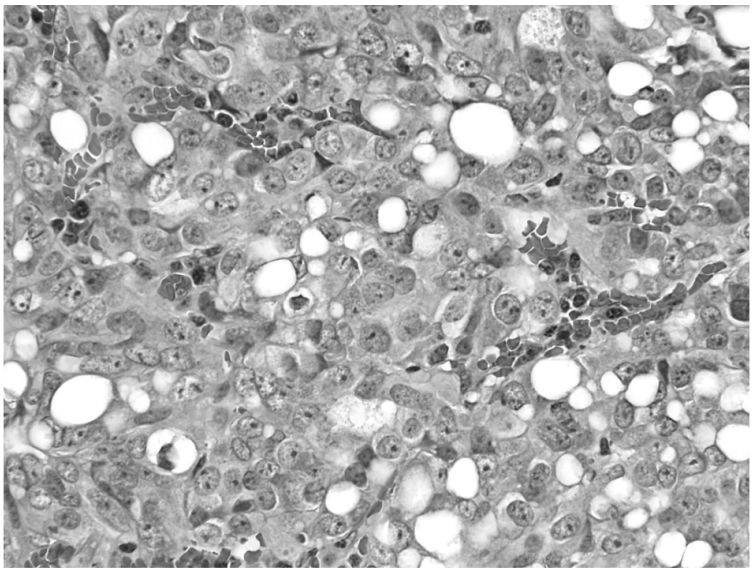


Figure 2.21: Signet ring change simulating signet ring cells of metastatic adenocarcinoma involving the ovary.

2.6 Proposed case of study: High-Grade Serous Ovarian Adenocarcinoma

As treated above, high-grade serous ovarian adenocarcinoma[5] is a rapidly growing carcinoma believed to have tubal origin with a high chromosomal instability; its peculiarity stands in the relapse timing of the patients affected by it. Indeed, patients can be recognized and differentiated into three classes (from sensitive to resistant) according to the time elapsed from the end of the first line therapy to relapse:

- relapse after 32 months since the end of treatment: *sensitive long term*;
- relapse between 6 and 32 months since the end of treatment: *sensitive*;
- relapse within 6 months since the end of treatment: *resistant*.

HGS-OC generally responds to platinum-based chemotherapy, but 80% of the patients relapse within 18 months from the diagnosis and progressively becomes resistant to treatment, up to becoming incurable: less than 20% of the patients survive after five years from the initial diagnosis.
This study involves the study of resistance in ovarian cancer patients basing on their transcriptional, mutational, and DNA structural profiles, in

[5]Adenocarcinoma: malignant epithelial tumor that originates specifically from cells of the glandular epithelium

particular of the molecular differences between patients that are sensitive to therapy compared to patients that are resistant, using data sets from the TCGA.
In particular, the ultimate aim is the identification of a molecular signature that could be used to predict the response to therapy (sensitive/resistant) at the time of diagnosis.

2.6.1 Resistance to platinum-based chemotherapy

Platinum-based combination chemotherapy with either cisplatin or carboplatin and paclitaxel is the standard treatment for ovarian cancer. However, resistance to chemotherapy remains a complex problem. Although 50% of the patients are already resistant to chemotherapy, a substantial number of those, who were originally responsive, develop resistance to platinum-based chemotherapy during the course of their treatment. In cell culture experiments, there is evidence that the efficacy of various chemotherapeutic agents, including cisplatin, requires a functional p53 protein for efficient induction of apoptosis and that loss of p53 function[6] enhances resistance to cytotoxic agents used in cancer therapy. The importance of identifying factors and molecular patterns that trigger resistance to this therapy is essential to distinguish which patients are suitable to receive the treatment and which patients would get worse after receiving it, given that prognosis relates to stage at diagnosis and sensitivity to platinum-based chemotherapy, [10].

[6]The p53 tumour suppressor gene plays a central role in cell cycle regulation and induction of apoptosis; p53 alterations influence the response to chemotherapy and clinical outcome in ovarian cancer patients.

Chapter 3

Bioinformatics and tools

3.1 Computational Biology

Computational biology and bioinformatics is an interdisciplinary field that develops and applies computational methods to analyse large collections of biological data, such as genetic sequences, cell populations or protein samples, to make new predictions or discover new biology. The computational methods used include analytical methods, mathematical modelling and simulation.
Initially, computational biology focused on the study of the sequence and structure of biological molecules, often in an evolutionary context. Beginning in the 1990s, however, it extended increasingly to the analysis of function. Functional prediction involves assessing the sequence and structural similarity between an unknown and a known protein and analyzing the proteins interactions with other molecules. Such analyses may be extensive, and thus computational biology has become closely aligned with systems biology, which attempts to analyze the workings of large interacting networks of biological components, especially biological pathways.
Biochemical, regulatory, and genetic pathways are highly branched and interleaved, as well as dynamic, calling for sophisticated computational tools for their modeling and analysis. Moreover, modern technology platforms for the rapid, automated (high-throughput) generation of biological data have allowed for an extension from traditional hypothesis-driven experimentation to data-driven analysis, by which computational experiments can be performed on genome-wide databases of unprecedented scale. As a result, many aspects of the study of biology have become unthinkable without the power of computers and the methodologies of computer science.

3.2 Genomics

Genomics is a branch of molecular biology involved in the study of the genome, including interactions of those genes with each other and with the person's environment, [11].
The genome is an organism's complete set of DNA, (Figure3.1). Virtually, every single cell in the body contains a complete copy of the approximately 3 billion DNA base pairs that make up the human genome, [12].

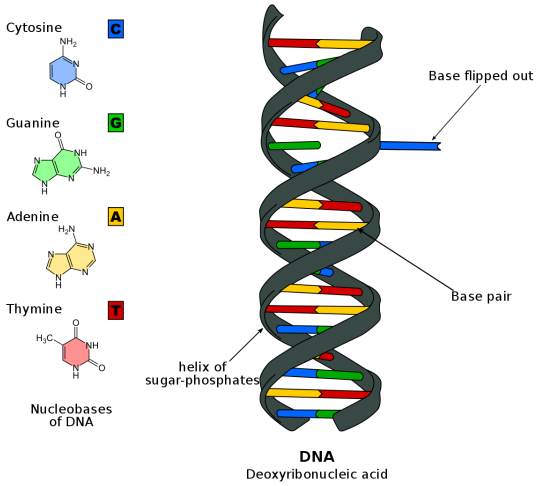


Figure 3.1: DNA double helix structure.

With its nucleotide-based language, DNA contains the information needed to build the entire human body. A gene traditionally refers to the unit of DNA that carries the instructions for making a specific protein or set of proteins. Each of the estimated 20,000 to 25,000 genes in the human genome codes for an average of three proteins.
Located on 23 pairs of chromosomes packed into the nucleus of a human cell, genes direct the production of proteins with the assistance of enzymes and messenger molecules. Specifically, an enzyme copies the information in a gene's DNA into a molecule called messenger ribonucleic acid (mRNA). The mRNA travels out of the nucleus and into the cell's cytoplasm, where the mRNA is read by a tiny molecular machine called a ribosome, and the information is used to link together small molecules called amino acids in the right order to form a specific protein.
Proteins make up body structures like organs and tissue, as well as control chemical reactions and carry signals between cells. If a cell's DNA is mutated, an abnormal protein may be produced, which can disrupt the body's usual processes and lead to a disease such as cancer.
Deoxyribonucleic acid (DNA) is the chemical compound that contains the instructions needed to develop and direct the activities of nearly all living

organisms. DNA molecules are made of two twisting, paired strands, often referred to as a double helix.
Each DNA strand is made of four chemical units, called nucleotide bases, which comprise the genetic "alphabet". The bases are adenine (A), thymine (T), guanine (G), and cytosine (C). Bases on opposite strands pair specifically: an A always pairs with a T; a C always pairs with a G. The order of the As, Ts, Cs and Gs determines the meaning of the information encoded in that part of the DNA molecule just as the order of letters determines the meaning of a word.
Sequencing simply means determining the exact order of the bases in a strand of DNA. Because bases exist as pairs, and the identity of one of the bases in the pair determines the other member of the pair, researchers do not have to report both bases of the pair.
In the most common type of sequencing used today, called sequencing by synthesis, DNA polymerase (the enzyme in cells that synthesizes DNA) is used to generate a new strand of DNA from a strand of interest. In the sequencing reaction, the enzyme incorporates into the new DNA strand individual nucleotides that have been chemically tagged with a fluorescent label. As this happens, the nucleotide is excited by a light source, and a fluorescent signal is emitted and detected. The signal is different depending on which of the four nucleotides was incorporated. This method can generate 'reads' of 125 nucleotides in a row and billions of reads at a time.
To assemble the sequence of all the bases in a large piece of DNA such as a gene, researchers need to read the sequence of overlapping segments. This allows the longer sequence to be assembled from shorter pieces, somewhat like putting together a linear jigsaw puzzle. In this process, each base has to be read not just once, but at least several times in the overlapping segments to ensure accuracy.
Researchers can use DNA sequencing to search for genetic variations and/or mutations that may play a role in the development or progression of a disease. The disease-causing change may be as small as the substitution, deletion, or addition of a single base pair or as large as a deletion of thousands of bases.
The study of tumours is one of the main branches of genomics, which is nowadays mainly taking advantages of the new possibilities provided by new advanced digital technologies regarding Big Data, Artificial Intelligence and machine learning algorithms, and Next Generation Sequencing (NGS), [13].
Contributing to the development of these technologies it is the H u m a n G e n o m e P r o j e c t , an international scientific research project that was designed to generate a resource that could be used for a broad range of biomedical studies. One such use is to look for the genetic variations that increase risk of specific diseases, such as cancer, or to look for the type of genetic mutations frequently seen in cancerous cells.

3.2.1 Gene expression and its regulation

Gene expression is the process by which the genetic code - the nucleotide sequence - of a gene is used to direct protein synthesis and produce the structures of the cell, (Figure 3.2).

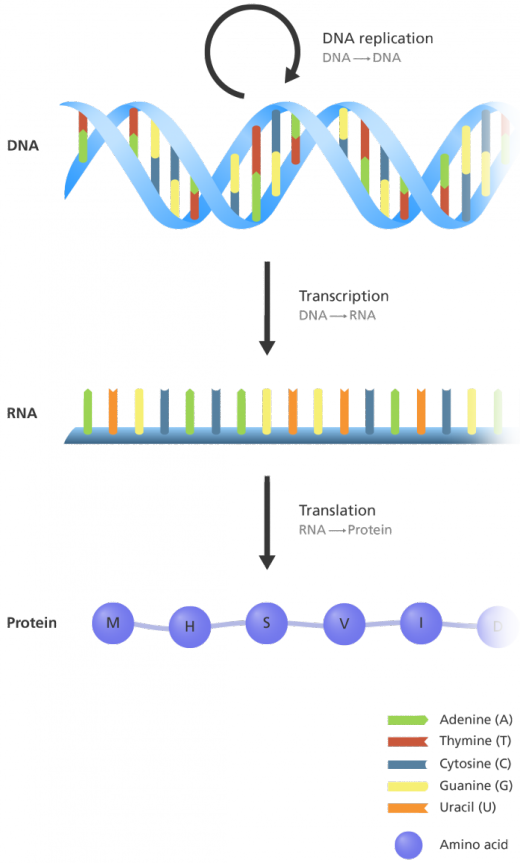


Figure 3.2: The central dogma of molecular biology: it explains the flow of genetic information, from DNA to RNA, to make a functional product, a protein.

Genes that code for amino acid sequences are known as "structural genes".
The process of gene expression involves two main stages:

- **Transcription**: is the process of RNA synthesis, controlled by the interaction of promoters and enhancers. Several different types of RNA are produced, including messenger RNA (mRNA), which specifies the sequence of amino acids in the protein product, plus transfer RNA (tRNA) and ribosomal RNA (rRNA), which play a role in the translation process; it involves four steps:

  1. initiation;

2. elongation;
3. termination;
4. processing.

- **Translation**: the mature mRNA molecule is used as a template to assemble a series of amino acids to produce a polypeptide with a specific amino acid sequence. The complex in the cytoplasm at which this occurs is called a ribosome. Ribosomes are a mixture of ribosomal proteins and ribosomal RNA (rRNA), and consist of a large subunit and a small subunit. It involves four steps:

  1. initiation;
  2. elongation;
  3. termination;
  4. post-translation processing of the protein.

Gene expression measurement is usually achieved by quantifying levels of the gene product, which is often a protein (i.e., measuring the expression level of cancer-causing genes (oncogenes) can help to determine a persons susceptibility to cancer).

Gene expression measure: RNA-seq (*RNA-sequencing*) is a technique that can examine the quantity and sequences of RNA in a sample using next generation sequencing (NGS). It analyzes the transcriptome of gene expression patterns encoded within our RNA. It's rapidly replacing gene expression microarrays because of the many advantages it has.
With RNA-seq more than just differential gene expression can be investigated. Although there are microarrays available for exon-level and microRNA analysis, most users are still interested in basic, probably 3 biased, differential gene expression. With RNA-seq, the attention can be driven at coding and non-coding RNA, at splicing and allele-specific expression, and possibly soon at full-length cDNA sequences, eliminating the need to infer or assemble isoforms.
Microarrays are also biased, as it has to be decided what content to place on the array. Since RNA-seq does not use probes or primers, the data suffer from much lower biases (although I do not mean to say RNA-seq has none).
RNA-seq provides digital data in the form of aligned read-counts, resulting in a very wide dynamic range, improving the sensitivity of detection for rare transcripts.
Expression is quantified by counting the number of reads that mapped to each locus in the transcriptome assembly step. Expression can be quantified for exons or genes using contigs or reference transcript annotations.

The common unit of measurement for the amount of the gene expression as a result of the RNA-seq process is the *Fragments Per Kilobase per Million reads* (FPKM), calculated from the number of reads that mapped to each particular gene sequence taking into account the gene length (one expects more reads to be produced from longer genes) and the sequencing depth (one expects more reads to be produced from the sample that has been sequenced to a greater depth):

$$FPKM = \frac{RC_g \cdot 10^9}{RC_{pc} \cdot L} \tag{3.1}$$

where:

- $\mathrm{RC_g}$ = number of reads mapped to the gene;
- $\mathrm{RC_{pc}}$ = total number of reads mapped to all protein-coding genes;
- L = gene length (in base pairs), calculated as the sum of the length of all the exons in the gene (i.e.,, the actual regions in the gene encoding information).

3.2.2 miRNA

MicroRNAs (*miRNAs*) are a class of short, endogenously-initiated non-coding RNAs that post-transcriptionally control gene expression via either translational repression or mRNA degradation. It is becoming evident that miRNAs are playing significant roles in regulatory mechanisms operating in various organisms, including developmental timing and host-pathogen interactions as well as cell differentiation, proliferation, apoptosis and tumorigenesis (i.e., miRNA genes frequently coincide with fragile sites and hot spots for chromosomal abnormalities or locate near cancer susceptibility loci that correlate with tumorigenesis), [14].
miRNAs are synthesized from primary miRNAs (pri-miRNAs) in two stages by the action of two RNase III-type proteins: Drosha in the nucleus and Dicer in the cytoplasm, Figure 3.3. The mature miRNAs are then bound by Argonaute (Ago) subfamily proteins. These miRNAs target mRNAs and thereby function as post-transcriptional regulators, [15].

It is possible to separate the genes subjected to the action of miRNAs from those that are not and categorize them into two groups:

- **target genes**, subjected to miRNA regulation: in general, the target genes have a longer 3'UTR (untraslated region);
- genes not subjected to miRNA regulation.

miRNAs can have both expression *tuning* and *buffering* motifs, all explained in Figure3.4:

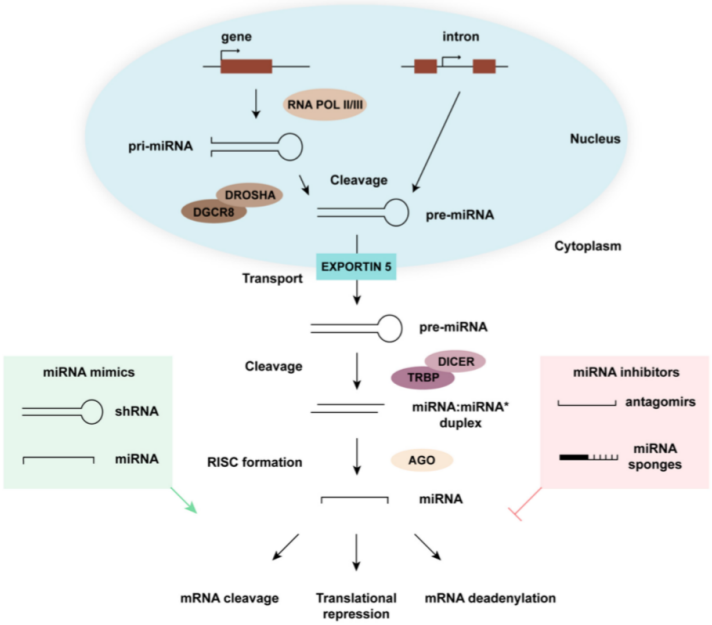


Figure 3.3: miRNA building process.

- **expression tuning motifs**:
  1. *simple repression*;
  2. *coherent feed forward loop*;
  3. *double negative feedback loop*;
- **expression buffering motifs**:
  1. *incoherent feed forward loop*;
  2. *negative feedback loop*;
  3. *double incoherent feed forward loop.*

From bioinformatics point of view, two resources dedicated to miRNAs and the action they perform towards their targets are *miRBase*[7] and *TargetScan*[8].
The common unit of transcript expression is the *Reads Per Million Mapped reads* (RPM), that is:

- a normalized expression unit that explicitly ignores transcript length;

[7]Reachable at http://www.mirbase.org/
[8]Reachable at http://www.targetscan.org/vert_72/

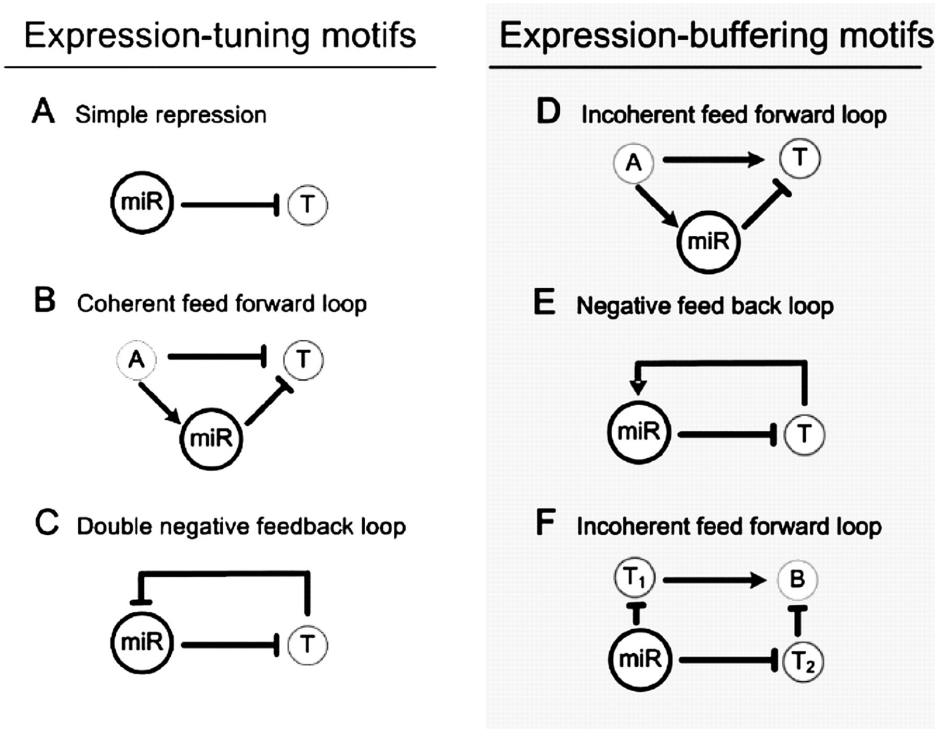


Figure 3.4: miRNA expression motifs.

- suitable for sequencing protocols where reads are generated irrespective of gene length;

it can be described by the formula:

$$RPM = \frac{N_{mg} \cdot 10^6}{Tot_{mr}} \tag{3.2}$$

where:

- N_{mg} = number of reads mapped to a gene;
- Tot_{mr} = total number of mapped reads from a given library.

3.2.3 Epigenetics and DNA methylation

Epigenetics has been defined has "the branch of biology which studies the causal interactions between genes and their products which bring the phenotype into being".
In the original sense of this definition, epigenetics referred to all molecular pathways modulating the expression of a genotype into a particular phenotype. Now, the definition of epigenetics has changed into the study of changes in gene function that are mitotically and/or meiotically heritable

and that do not entail a change in DNA sequence. The epigenetic modifications describe histone variants, posttranslational modifications of amino acids on the amino-terminal tail of histones, and covalent modifications of DNA bases [16].
Epigenetic mutations are able to change the chromatin structure, either generating euchromatin and promoting gene transcription, or producing heterochromatin and repressing gene transcription. One of the main epigenetic mechanisms regulating gene expression is the DNA methylation, explained in Figure 3.5.

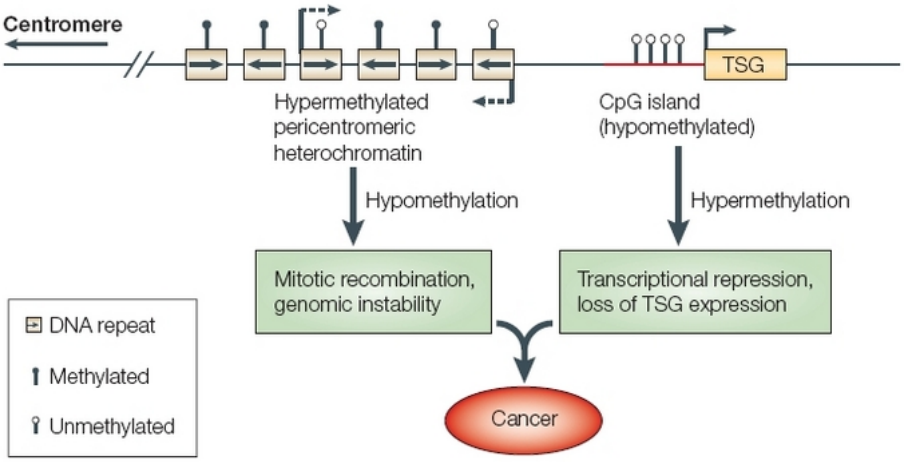


Figure 3.5: Graphical explanation of the role of methylation in gene expression.

DNA methylation is an epigenetic mechanism that occurs by the addition of a methyl (CH_3) group to DNA, thereby often modifying the function of the genes and affecting gene expression. The most widely characterized DNA methylation process is the covalent addition of the methyl group at the 5-carbon of the cytosine ring resulting in 5-methylcytosine (5-mC), also informally known as the fifth base of DNA. These methyl groups project into the major groove of DNA and inhibit transcription.
A common way to measure DNA methylation is the calculation of **beta values** (β) and it represents the probability that a coding gene is not transcribed due to methylation at the level of DNA. Beta values are the estimate of methylation level using the ratio of intensities between methylated and unmethylated alleles; these values are between 0 and 1, with 0 being unmethylated and 1 fully methylated.
Beta values are commonly described by the formula:

$$\beta = \frac{\phi^m}{\phi^m + \phi^{nm}} \tag{3.3}$$

where:

- ϕ^m = methylation intensity measurement;
- ϕ^{nm} = non-methylation intensity measurement.

3.2.4 CNA regions

A copy number alteration (CNA) [17] is when the number of copies of a particular gene varies from one individual to the next, as in Figure 3.6. Copy number variation is a type of structural variation where there is a stretch of DNA, which is duplicated in some people, and sometimes even triplicated or quadruplicated. And so when looking at that chromosomal region, it is shown a variation in the number of copies in normal people. Sometimes those copy number variants include genes, maybe several genes, which may mean that this person has four copies of that gene instead of the usual two, and somebody else has three, and somebody else has five. CNAs occur via a variety of mechanisms in cancer. Entire chromosomes may be gained/ lost during cell division, generating 3N or 1N copy number status for all genes on the chromosome. This occurs due to failed cell-division checkpoints resulting in chromosome missegregation. In contrast to such gains at the total chromosome level, tiny focal CNAs may alter a single gene (or even part of a gene), [18].
Aside from very infrequent gene losses paired with mutations, there are also a few CNAs which drive cancer through amplification of oncogenes. In the case of SOC, new cures are unlikely to be found unless somatic copy number alterations (SCNAs) are considered, so this study includes the use of this type of data.

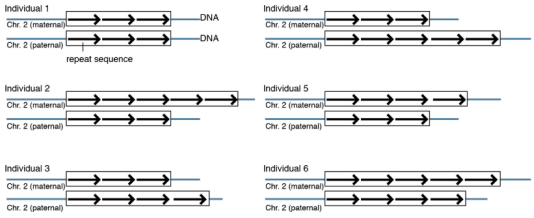


Figure 3.6: Graphical representation of CNA mutation.

3.3 GMQL

The GenoMetric Query Language (GMQL) is a high-level, declarative language used to perform queries on big genomic data, structured according to the Genomic Data Model (GDM), [19].

GMQL uses the GDM[9] based on the notion of genomic region: each region can be compared with millions of other regions, typically using metric properties; in addition, GDM also covers metadata of arbitrary structure. This model is mainly based on the notions of *datasets*, defined as collections of samples, and *samples*, representing different genomic data.
Each sample consists of two main parts:

- **Region data**, describing the physical coordinates of the genome areas and their features, encoded as specific fields having different values for each region;
- **Metadata**: descriptive attributes of a sample, describing its biological, clinical and experimental properties.

Regions are data format independent and provide an interoperability framework for comparing data on mutations, expression or regulation; while metadata are system independent and implement an interoperability framework for comparing samples based upon their biological aspects.
Formally, in the GDM a sample s is defined as a triple:

$$s =< id, \{r_i, v_i\}, \{m_j\} > \tag{3.4}$$

where id is the sample identifier; each pair ¡ r_i,v_i ¿ represents a region, with coordinates r_i and values v_i; m_j are attribute-value pairs, with values of type string.
Each sample s has specific attributes describing its region properties and an associated set of attribute-value pairs referred to as the metadata of s. The region data schema of s is the set of all attribute names used for region coordinates and values, and the region data type of s is the record of all the elementary types of the corresponding attributes.
The use of a type system to express region data allows for controlled arbitrary operations upon type-compatible values.
A GMQL query is expressed as a sequence of GMQL operations with the following structure:

$$< variable >= operator(< parameters >) < variables >$$

where each variable stands for a GDM dataset. Operators apply to one or more operand variables and construct one result variable; parameters are specific for each operator. Parameters of several operators include predicates, used to select and join samples; predicates are built by arbitrary Boolean expressions of simple predicates, as it is customary in relational

[9]The GDM is a formal framework used for representing in a uniform way genomic data with different formats.

algebra.
A typical GMQL query starts with a **SELECT** operation, which creates a dataset with only the data samples that it filters out from an input dataset by using a predicate upon their metadata attributes.
Then, the query proceeds by processing the selected samples in batch with operations applied on their region data and/or metadata.
Finally, it ends with a **MATERIALIZE** operation, which stores a dataset by saving the region data of each of its samples in an individual text file in standard GTF format and the related metadata in an associated tab delimited text file.
In this project, GMQL web interface (Figure 3.7) has been used in order to extract information of interest from TCGA data.

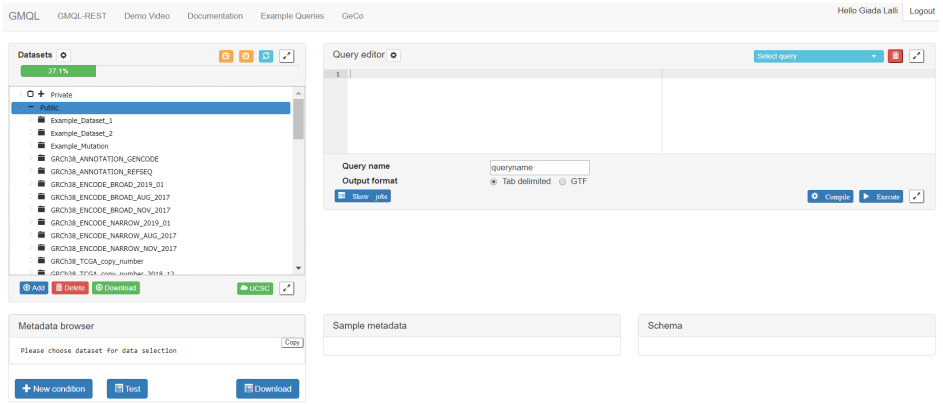


Figure 3.7: GMQL web interface.

3.4 Python Libraries

Most computations are performed using the Python programming language.
Python is an interpreted, high-level, general-purpose programming language, in which is not needed to compile the scripts into machine language instructions, and it allows to easily execute a lot of complex tasks, thanks to the availability of standard libraries and of a large number of resources.
Python has been choosen as it offers a wide set of functions for statistical modeling and machine learning analysis and it is the only programming language integrated with GMQL. Here, Python is used through *Anaconda*, an open source distribution for large-scale data processing and scientific computing and the most convenient framework for Python data science and machine learning, including at the installation more than 250 popular data science packages.
The main libraries used for the computations are detailed in the following paragraphs.

Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with relational or labeled data both easy and intuitive, [20]. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis / manipulation tool available in any language.

Scikit - learn is a Python library providing various classification, regression and clustering algorithms for performing machine learning operations and it is used in this project for normalizing input data of the linear regression process, using the *StandardScaler* class in the *preprocessing* module, consistently helping to compare results across models, [21].

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms, [22]. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits. By using this library, it is possible to generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc.

Lifelines is an implementation of survival analysis in Python, [23]; the benefits offered by it over other survival analysis implementations are:

- built on top of Pandas;
- internal plotting methods;
- simple and intuitive API;
- only focus is survival analysis;
- handles right, left and interval censored data

3.5 R Libraries

KaryoploteR is based on base R graphics and mimicks its interface. You first create a plot with `plotKaryotype` and then sequentially call a number of functions (`kpLines`, `kpPoints`, `kpBars`) to add data to the plot.
`karyoploteR` is only a plotting tool: that means that it is not able to download or retrieve any data. The downside of this is that the user is responsible of getting the data into R. The upside is that it is not tied to any data provider and thus can be used to plot genomic data coming from anywhere. The only exception to this are the ideograms cytobands, that by default are plotted using pre-downloaded data from UCSC.

3.6 Cytoscape

Cytoscape is an open source software platform for visualizing molecular interaction networks and biological pathways and integrating these networks with annotations, gene expression profiles and other state data. Although Cytoscape was originally designed for biological research, now it is a general platform for complex network analysis and visualization. Cytoscape core distribution provides a basic set of features for data integration, analysis, and visualization.
In this thesis, Cytoscape has been very useful for plotting the visual representation of the interaction between genes belonging to the same set.

Chapter 4

Datasets

4.1 TCGA: The Cancer Genome Atlas

4.1.1 Introduction

The Cancer Genome Atlas (TCGA) 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. So far, TCGA researchers have analysed large cohorts of over 30 human tumours through large-scale genome sequencing and integrated multi-dimensional analyses. Studies of individual cancer types, as well as comprehensive pan-cancer analyses have extended current knowledge of tumorigenesis. A major goal of the project was to provide publicly available datasets to help improve diagnostic methods, treatment standards, and finally to prevent cancer. This chapter is devoted to a review of the current status of TCGA Research Network structure, purpose, and achievements, as discussed in [24].

4.1.2 The idea behind the project

Cancer is deemed the most challenging illness to counteract.
More than 200 forms of cancer have been analyzed and each type can be distinguished by different molecular profiles requiring unique therapeutic strategies. Cancer involves dynamic changes in the genome: the architecture of occurring genetic aberrations - such as somatic mutations, copy number variations, changed gene expression profiles, and different epigenetic alterations - are individual for each type of cancer. The need for better diagnosis, treatment, and prevention of cancer has appeared and strongly correlates with a better understanding of genetic changes in the tumour. The latest progress in the technological development of genome-wide sequencing and bioinformatics has shed new light on the cancer genome, [25].
In 2005, The Cancer Genome Atlas (TCGA) and in 2008 the International Cancer Genome Consortium (ICGC) were launched as the two main

projects quickening the comprehensive understanding of the genetics of cancer using innovative genome analysis technologies, helping to generate new cancer therapies, diagnostic methods, and preventive strategies.
The National Institute of Health (NIH) launched TCGA Pilot Project to create a comprehensive atlas of cancer genomic profiles. The TCGA is a public funded project that intends to catalogue and discover major cancer-causing genome alterations in large cohorts of over 30 human tumours through large-scale genome sequencing and integrated multi-dimensional analyses. Providing publicly available cancer genomic datasets will allow the improvement of diagnostic methods, treatment standards, and finally cancer prevention. Phase I of the project (a 3-year pilot study) aimed to develop and test the research infrastructure based on the characterisation of chosen tumours having a poor prognosis: brain, lung, and ovarian cancers. Since 2009 (phase II) analyses have expanded to additional types reaching 30 different tumour types analysed by 2014. The TCGA project engaged scientists and managers from NIHs National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI) funded by the US government, as well as cooperating with institutions across the USA and Europe. To run the project, the NCI as well as the NHGRI each invested $50 million for the 3-year pilot study. Additional funding was also provided from different sources, such as the American Recovery and Reinvestment Act (ARRA), to help stimulate the US economy in the context of biomedicine.

4.1.3 The Cancer Genome Atlas Research Network

The structure of TCGA is well established and involves several cooperating centres engaged for collection and sample processing, followed by high-throughput sequencing and sophisticated bioinformatics data analyses. First, different Tissue Source Sites (TSSs) collect the required biospecimens (blood, tissue) from eligible cancer patients and deliver them to the Biospecimen Core Resource (BCR). Next, the BCR catalogue, process, and verify the quality and quantity of samples, and then submit clinical data and metadata to the Data Coordinating Center (DCC) and provide molecular analytes for the Genome Characterization Centers (GCCs) and Genome Sequencing Centers (GSCs) for further genomic characterisation and high-throughput sequencing. Then, sequence-related data are deposited in the DCC. The Genome Characterisation Centers also submit trace files, sequences, and alignment mappings to NCIs Cancer Genomics Hub (CGHub) secure repository. The generated genomic data is made available to the research community and Genome Data Analysis Centers (GDACs). The GDACs provide new information-processing, analysis, and visualisation tools to the entire research community to facilitate broader use of TCGA data. Furthermore, the information generated by the TCGA

Research Network is centrally managed at the DCC and entered into public free-access databases (TCGA Portal, NCBIs Trace Archive, CGHub), allowing scientists to constantly access the cancer datasets and to speed advancements in cancer biology and linked technologies (Figure 4.1, [24]).

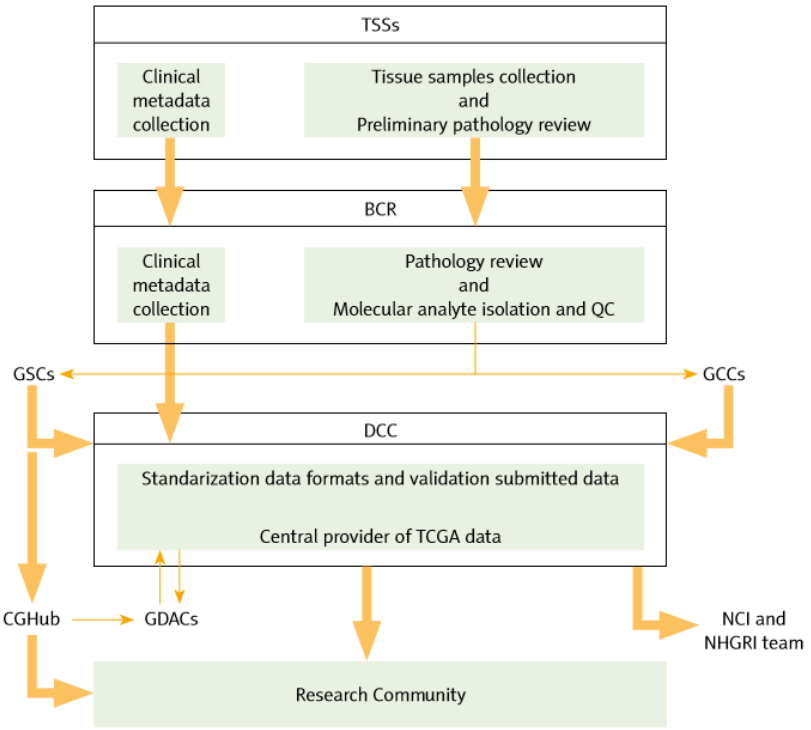


Figure 4.1: The Cancer Genome Atlas (TCGA) Research Network Centres flowchart.

4.1.4 Platforms and data models

To provide a comprehensive analysis of cancer genome profiles, TCGA applied high-throughput technologies based on microarrays (to test nucleic acids and proteins) and next-generation sequencing methods (for global analysis of nucleic acids). The research network structure comprises many centres employing different platforms to provide global information on cancer genomics.
Some of the applied methods are MicroRNA sequencing (miRNAseq), DNA sequencing (DNAseq), RNA sequencing (RNAseq), SNP-based platforms, Array-based DNA methylation sequencing and Reverse-phase protein array (RPPA).
The method by which the datasets used in this study were processed will be discussed below.

RNA sequencing (RNAseq)

RNAseq is a high-throughput technology for transcriptome (total RNA) profiling, deriving strand information with very high precision. RNAseq is able to quickly distinguish and quantify rare and common transcripts, isoforms, novel transcripts, gene fusions, and non-coding RNAs, among a wide range of samples, including low-quality samples. For transcriptome analysis, TCGA uses a platform based on the Illumina system. The TCGA deposited data includes information about both nucleotide sequence and gene expression. RNA sequence alignment provides different levels of information such as RNA sequence coverage, sequence variants, expression of genes, exon, or junction. The NCBI dbGaP database is the official repository for the actual sequence data, [24].

MicroRNA sequencing (miRNAseq)

MiRNAseq is a type of RNASeq, utilising material enriched in small RNAs, allowing the detection of specific sets of short, noncoding RNAs (mi RNAs) that have the capacity to regulate hundreds of genes within and across diverse signalling pathways. Moreover, miRNA-sequencing defines tissue-specific miRNA expression profiles, their isoforms, connection with diseases, and the discovery of unreported miRNAs, [24].

Array-based DNA methylation sequencing

Array-based DNA methylation sequencing is a high- throughput, genome-wide analysis of DNA methylation profile providing information of epigenetic changes in the genome. Abnormal profile of DNA methylation of CpG sites is among the earliest and most frequent alterations in cancer. The TCGA utilises DNA methylation assay mainly based on the Illumina platform, assuring single-base-pair resolution, high accuracy, easy workflows, and low input DNA requirements. Methylation profiling technologies are based on highly multiplexed genotyping of bisulphite-converted genomic DNA. The TCGA DNA methylation data files contain information of signal intensities (raw and normalised), detection confidence, and calculated beta values for methylated (M) and unmethylated (U) probes, [24].

DNA sequencing (DNAseq)

DNA sequencing (DNAseq) is a high-throughput method for determining the nucleotides within a DNA molecule, providing information about DNA alterations, such as insertions, deletions, polymorphism as well as **copy number variation** or mutation frequencies. To catalogue the genomic diversity across cancer types, TCGA Genome Sequencing Centers utilise DNA sequencing systems based on Sanger Sequencing.

Also other platforms can potentially produce this kinds of data, such as *reverse-phase protein array* (RPPA), which is a highly sensitive (detecting nanograms of proteins), reproducible, high-throughput, functional and quantitative proteomic method for large-scale protein expression profiling, biomarker discovery, and cancer diagnostics, or *SNP-based platforms*, that are used to analyse genome-wide structural variation across multiple cancer genomes. Array-based detection of single nucleotide polymorphisms (SNPs) included platforms able to define SNP and CNV across multiple samples, [24].

4.1.5 Visualisation and examination of the genomic data

Nowadays, next-generation sequencing (NGS) and array-based profiling methods produce massive numbers of diverse types of genomic data allowing researchers to analyse the cancer genome at an advanced level. Integrated multi-dimensional data visualisation is an indispensable component of cancer genomic data analysis. Consequently, the demand for advanced comprehensive visualisation tools has emerged allowing the emergence of numerous useful imaging tools and databases, as The cBioPortal for Cancer Genomics that has been used in this study.

The cBioPortal for Cancer Genomics

The cBioPortal for Cancer Genomics (http://cbioportal. org) is an open-access resource developed at the Memorial Sloan-Kettering Cancer Centre (MSKCC) for visualisation, analysis, and download of large-scale cancer genomics data sets. Additionally, the portal also grants for interactive exploration of custom datasets by access to OncoPrinter or Muttation-Mapper web tools. Currently, the portal collects data from 69 cancer genomics studies (datasets from literature and TCGA portal) including DNA copy-number data, mRNA and miRNA expression data, mutations, RPPA data, DNA methylation data, and limited clinical data related to survival. Visualisation type involves networks, matrices, and heatmaps. The cBio portal complements existing tools, such as the TCGA and ICGC data portals, the IGV, the UCSC Cancer Genomics Browser, and IntOGen.

Ovarian Cancer

Ovarian serous adenocarcinoma is a major type of ovarian cancer. The high mortality of ovarian cancer patients (only 31% of patients are expected to live for five years or more) is connected to a lack of methods for early detection and treatment. Recently, TCGA researchers performed a wide-range analysis of the genomic and epigenetic changes that occur in high-grade serous ovarian carcinoma (HGS-OvCa) and demonstrated several potential therapeutic targets. In their work published in 2011 in

Nature, TCGA scientists analysed 489 tumour samples and determined the presence of TP53 mutation in almost all specimens (96%) and a low but significant frequency of somatic mutations in nine further genes, including BRCA1 and BRCA2 (mutated in 22% of tumours). Integrated multidimensional analyses led to the identification of four ovarian cancer transcriptional subtypes, three miRNA subtypes, four promoter methylation subtypes, and a transcriptional signature that is associated with survival outcome. However, the main goal of TCGA research is to identify new therapeutic approaches. Therefore, TCGA scientists imply opportunities for therapeutic intervention in commonly dysregulated pathways: RB, RAS/PI3K, FOXM1, and NOTCH. Moreover, the research group from Johns Hopkins Medical Institution identified an amplified region in chromosome 19, containing a NACC1 gene known to contribute to chemoresistance. Analysing TCGA data, they demonstrated the correlation of amplified NACC1 with early tumour reoccurrence in ovarian cancer patients. Furthermore, TCGA data have helped to shed light on the effect of BRCA1/2 mutations on ovarian cancer patients survival. Recent findings from analyses of the ovarian cancer dataset have the potential to improve the therapeutic management of this deadly disease.

4.2 Project Datasets

4.2.1 Building datasets

To proceed with the development of a classifier for the prediction of the class belonging, it was first necessary to generate the input datasets.
The first step performed for this project was to examine, among the data available from TCGA, those concerning patients affected by Ovarian Cancer; subsequently, a file containing the clinical data[10] of these patients was downloaded: each patient was associated with an identification barcode, and to each barcode was assigned a series of data relating to the description of the status of the patient in question.

- In the first instance, to make a first distinction between patients due to their relapse timing, reference was made to the "Platinum Status" datum, which identifies patients as "Sensitive" or "Resistant".

- Consequently, given the temporal variety of patients belonging to the "Sensitive" category, a further division was made based on the Progression-Free Survival data (i.e., survival time before relapse); thus, from the split of the "Sensitive" class, it was possible to generate two distinct classes: patients with relapse between 6 and 32 months from therapy will constitute the "Sensitive Short Term"

[10]The file containing this data has been downloaded from https://gdc.cancer.gov/about-data/publications/ov_2011/

class, while patients who have not manifested relapse after 32 months will constitute the "Sensitive Long Term" class.

- Once the patients belonging to the three distinct classes were identified, their barcodes were selected; to obtain the metadata belonging to each patient, the GMQL web interface was used, in which for each class of patients, a query (Fig. 4.2) was inserted, from which two files for each patient have been obtained: a metadata file and a region file[11].

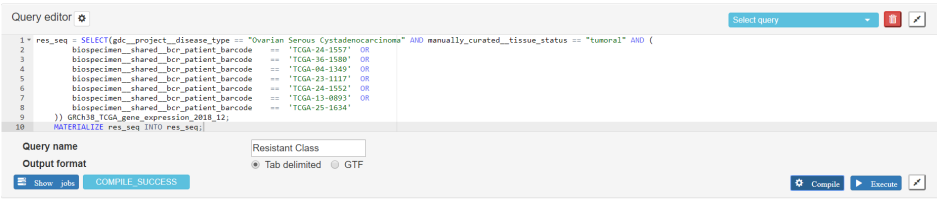


Figure 4.2: Abbreviated example of the performed query on GMQL web interface to obtain files for Resistant class, gene expression data.

- Then, using the region files containing information of interest, the descriptive information of the genetic profile of each sample is extracted by using Python (version 3.0): location of the feature, feature_id, feature_symbol, value of the expression.
- Once only the information relevant to our analysis was taken into account, the data relating to each patient were combined to form a single dataset; in Table 4.1 are shown the original number of samples available for each class and for each data type.

Table 4.1: Samples obtained from GMQL queries.

Class	**Gene expression**	**miRNA**	**Methylation**
Resistant	62	85	91
Sensitive Short	109	147	148
Sensitive long	34	40	40

This procedure was carried out for each class and for each type of data: where the type of data was of gene expression, the indicated expression value will be represented by the "fpkm", while in relation to a given miRNA, the same value will be identified by the "rpm"; a shortcut of the datasets is shown is Fig. 4.3, 4.4 and 4.5.

[11]This procedure has been iterated three times, in order to obtain from GMQL gene expression data, miRNA expression data and methylation of the patients (referring to assembly GRCh38).

patient	chrom	start	stop	ensemble_id	entrez_id	gene_symbol	fpkm_uq	fpkm
R_00000	chr1	69090	70008	ENSG00000186092.4	79501.0	OR4F5	0.000000	0.000000
patient	**chrom**	**start**	**stop**	**ensemble_id**	**entrez_id**	**gene_symbol**	**fpkm_uq**	**fpkm**
SL_00000	chr1	69090	70008	ENSG00000186092.4	79501.0	OR4F5	0.000000	0.000000
patient	**chrom**	**start**	**stop**	**ensemble_id**	**entrez_id**	**gene_symbol**	**fpkm_uq**	**fpkm**
SS_00000	chr1	69090	70008	ENSG00000186092.4	79501.0	OR4F5	0.000000	0.000000

Figure 4.3: Example of the first patient for each class and the relative informations about gene expression. As can be seen also in Fig.4.4 and 4.5, the barcode of each patient has been replaced with the first letter associated to the class of belonging, followed by the rank.

patient	chrom	start	stop	mirna_id	rpm	entrez_id	gene_symbol
R_00000	chr1	17368	17436	hsa-mir-6859-1	0.0	102466751.0	mir6859-1
patient	**chrom**	**start**	**stop**	**mirna_id**	**rpm**	**entrez_id**	**gene_symbol**
SL_00000	chr1	17368	17436	hsa-mir-6859-1	0.0	102466751.0	mir6859-1
patient	**chrom**	**start**	**stop**	**mirna_id**	**rpm**	**entrez_id**	**gene_symbol**
SS_00000	chr1	17368	17436	hsa-mir-6859-1	0.0	102466751.0	mir6859-1

Figure 4.4: Example of the first patient for each class and the relative informations about miRNA expression.

patient	chrom	start	stop	beta_value	gene_symbol	gene_type	ensemble_transcript_id	feature_type
R_00000	chr1	924804	924806	0.006785	SAMD11	protein_coding	ENST00000342066.6|ENST00000420190.4|ENST000004...	Island
patient	**chrom**	**start**	**stop**	**beta_value**	**gene_symbol**	**gene_type**	**ensemble_transcript_id**	**feature_type**
SL_00000	chr1	924804	924806	0.009356	SAMD11	protein_coding	ENST00000342066.6|ENST00000420190.4|ENST000004...	Island
patient	**chrom**	**start**	**stop**	**beta_value**	**gene_symbol**	**gene_type**	**ensemble_transcript_id**	**feature_type**
SS_00000	chr1	924804	924806	0.007873	SAMD11	protein_coding	ENST00000342066.6|ENST00000420190.4|ENST000004...	Island

Figure 4.5: Example of the first patient for each class and the relative informations about DNA methylation.

In order to then be able to perform a transversal analysis between all data types, only the barcodes common to all data types were selected: after extracting the samples, a match between the patients' id in the different data sets has been carried out, in order to uniquely identify each sample[12].
In addition to the data types described so far, CNA data[13] have been considered in order to carry out a cross-sectional analysis on the mutation profile of the various classes of patients; for this data type, the same pro-

[12]This means that only patients for whom gene expression, miRNA, and methylation data were known were considered in order to build the datasets.

[13]The original number of samples associated to each CNA class was: for *Resistant*: 86; for *Sensitive*: 148; for *Sensitive long*: 32.

cedure for constructing the datasets has been implemented, Fig. 4.6.

	patient	chrom	start	stop	n_probes	seg_mean
0	R_00000	1	3301765	8297161	3189	-0.0026
	patient	**chrom**	**start**	**stop**	**n_probes**	**seg_mean**
0	SL_00000	1	3301765	21771173	10406	0.0350
	patient	**chrom**	**start**	**stop**	**n_probes**	**seg_mean**
0	SS_00000	1	3301765	17266821	7519	0.1766

Figure 4.6: Example of the first patient for each class and the relative informations about CNA regions.

The number of final samples for each class and for each type of data will be shown in Table 4.2.

Table 4.2: Samples obtained from GMQL queries after considering only common barcodes between data types.

Class	**Gene expression**	**miRNA**	**Methylation**	**CNA**
Resistant	57	57	57	57
Sensitive Short	104	104	104	104
Sensitive long	25	25	25	25

The information contained in the datasets will be explained more in detail below:

- for **gene expression data**, Figure 4.3:

  1. *patient*: the barcode of each patient has been replaced with the first letter associated to the class of belonging, followed by the rank;
  2. *chrom, start, stop*: location of each feature on the genome: the name of the chromosome, the starting and the ending position of the feature in the chromosome;
  3. *ensemble_id*: the Ensembl ID of the gene, including its version with . notation; an Ensembl stable ID consists of five parts: ENS(species)(object type)(identifier).(version);
  4. *entrez_id*: the Entrez gene ID, identifier for a gene per the NCBI Entrez database, of the gene related to the reported variant

5. *gene_symbol*: the symbol of the gene related to the reported variant;
6. *fpkm_uq*: the upper quartile FPKM (FPKM-UQ) is a modified FPKM calculation in which the total protein-coding read count is replaced by the 75th percentile read count value for the sample;
7. *fpkm*: Fragments Per Kilobase of transcript per Million mapped reads;

- for **miRNA expression data**, Figure 4.4:

1. *patient*: the barcode of each patient has been replaced with the first letter associated to the class of belonging, followed by the rank;
2. *chrom, start, stop*: location of each feature on the genome: the name of the chromosome, the starting and the ending position of the feature in the chromosome;
3. *mirna_id*: a valid miRBase ID (http://www.mirbase.org/);
4. *rpm*: millions of reads that mapped to a miRNA;
5. *entrez_id*: the Entrez gene ID, identifier for a gene per the NCBI Entrez database, of the gene related to the reported variant
6. *gene_symbol*: the symbol of the gene related to the reported variant;

- for **DNA methylation data**, Figure 4.5:

1. *patient*: the barcode of each patient has been replaced with the first letter associated to the class of belonging, followed by the rank;
2. *chrom, start, stop*: location of each feature on the genome: the name of the chromosome, the starting and the ending position of the feature in the chromosome;
3. *beta_value*: the ratio between the methylated array intensity and total array intensity, falling between 0 (lower levels of methylation) and 1 (higher levels of methylation);
4. *gene_symbol*: the symbol of the gene related to the reported variant;
5. *gene_type*: a general classification for each associated gene (e.g., protein coding, miRNA, pseudogene);
6. *ensemble_transcript_id*: Ensembl IDs of the transcripts related to the gene provided in the column *gene_symbol*, retrieved from the *gene_type* of the GDC DNA methylation file; an ensembl

gene includes any spliced transcripts (ENST) with overlapping coding sequence; transcript clusters with no overlapping coding sequence are annotated as separate genes;

7. *feature_type*: the position of the CpG site in reference to the island: Island, or N_Shore, or S_Shore, or N_Shelf, or S_Shelf;

- for **CNA data**, Figure 4.6:

1. *patient*: the barcode of each patient has been replaced with the first letter associated to the class of belonging, followed by the rank;
2. *chrom, start, stop*: location of each feature on the genome: the name or number of the chromosome where the CNV is located, the starting and the ending position of the CNV feature in the chromosome;
3. *n_probes*: the number of consecutive probes that comprise the genome segment with the CNV;
4. *seg_mean*: the estimated Copy Number (CN) ratio for the segment, that is the log2 ratio of the tumor intensity of CN to the normal intensity of CN.

Chapter 5

Computational Methods Foundation

To achieve the goal of distinguishing the response to therapy of HGS-OC patients, two methods have been tried: the first one is based on survival regression, in particular on CoxPH Model, used to predict the time to event - which in our case is the time to relapse experienced by patients. The other one, instead, uses classification models to directly discriminate the classes. Various levels of features selection have been performed, using statistical tests. To avoid false positives, and hence to select only features relevant for the analysis, multiple statistical test corrections have been applied. After performing various features selection levels, through which various subsets of features were created that are considered significant, the next step was to use the classifier for class distinction, verified by a 10-fold cross-validation.

5.1 Multiple statistical test correction

Bonferroni Correction Bonferroni correction is a conservative test that, although protects from Type I Error (the higher the chance for a false positive; rejecting the null hypothesis when you should not), is vulnerable to Type II errors (failing to reject the null hypothesis when you should in fact reject the null hypothesis).
Its procedure is to alter the p_value to a more stringent value, thus making it less likely to commit Type I Error.
To get the Bonferroni corrected/adjusted p_value, we have to divide the original -value by the number of analyses on the dependent variable. Then a new alpha for the set of dependent variables (or analyses) that does not exceed some critical value is assigned:

$$\alpha_{critical} = 1 - (1 - \alpha_{altered})^k \tag{5.1}$$

where k = the number of comparisons on the same dependent variable.

After correcting the value of the p_value associated with each feature of the original dataset, a threshold of 0.05 (corresponding therefore to an error rate of 5%) was imposed to select which features were significantly expressed after Bonferroni correction.

Given the low number of features extracted through the properly applied Bonferroni correction, we have chosen to modify one of the parameters constituting the equation, which can thus be rewritten:

$$p\_value_{corrected} = p\_value_{nominal} \cdot n\_tests \tag{5.2}$$

where n_tests = the total number of tests and p_value $_{\text{nominal}}$ = original p_value before the correction.
The parameter that has been modified is precisely n_tests, which is now equal to the sum of the number of patients belonging to the classes for which the differentially expressed features are sought. We will call this correction "mild Bonferroni correction".

False Discovery Rate, in particular the Benjamini-Hochberg Procedure The false discovery rate (FDR) is a method of conceptualizing the rate of type I errors in null hypothesis testing when conducting multiple comparisons. FDR-controlling procedures are designed to control the expected proportion of "discoveries" (rejected null hypotheses) that are false (incorrect rejections).
The Benjamini-Hochberg procedure (BH step-up procedure) controls the FDR at level α:

- For a given α, find the largest k such that:

$$P(k) \leq \frac{k}{m}\alpha$$

- Reject the null hypothesis for all $H_{(i)}$ for i = 1, ..., k.

The BH procedure is valid when the m tests are independent, and also in various scenarios of dependence, but is not universally valid. It can be rewritten as:

$$p\_value_{corrected} = p\_value_{nominal} \cdot \frac{n\_tests}{ranking} \tag{5.3}$$

where ranking = position of each p_value $_{\text{nominal}}$ in the ordered list[14].

After correcting the value of the p_value associated with each feature of the original dataset, a threshold of 0.05 (corresponding therefore to an error rate of 5%) was imposed to select which features were significantly

[14]P_values are ordered from the most significant (with a minor value) to the least significant (with a major value).

expressed after FDR correction.
Given the low number of features extracted through the properly applied FDR correction, we have chosen to modify this correction in the same way as we did for Bonferroni, so as to apply a "mild FDR correction".

Selection from p_value The p-value is the level of marginal significance within a statistical hypothesis test representing the probability of the occurrence of a given event; a smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.
Due to this, it was decided to proceed with a further selection without correcting the value of the original p_value, but choosing three thresholds: a threshold = 0.05 (statistically significant), a threshold = 0.005 (very significant) and a threshold = 0.0005 (extremely significant). This procedure was implemented in order to evaluate which features were naturally expressed with a significant p_value.

Hypergeometric test: Enrichment Analysis Enrichment analysis is a means to characterize biological attributes in a given gene set.
From a list of genes obtained from previous analyzes, a grouping is carried out in functional classes for the interpretation of the results. Rather than studying genes individually, the links in groups are analyzed. Depending on the different expressions of genes in a group, for different biological conditions, one can understand the cellular processes in which the genes of interest are involved [26].
The association of groups of genes with particular functions derives from experimental evidence or from computational inference. Establishing the list of genes is not enough, we must estimate the statistical significance, that is, how far this list differs from the genes obtained by sampling at random on the genome.
This type of test is called hypergeometric testing. The set of genes of interest C is compared with a universe set U (background) and a set of elements with a given property P. The population of C genes can be divided into P $\cap$ C genes associated with a certain function, and the rest in C not associated with it, as in Figure 5.1).
If $\mid P \cap C \mid = X$, $\mid U \mid = M$, $\mid P \mid = K$ and $\mid C \mid = N$, then the probability Pr of observing X or more elements with the property P is:

$$Pr = 1 - \sum_{i=0}^{X-1} \frac{\binom{K}{i}\binom{M-K}{N-1}}{\binom{M}{N}} = \sum_{i=X}^{N} \frac{\binom{K}{i}\binom{M-K}{N-1}}{\binom{M}{N}} \tag{5.4}$$

The zero hypothesis is that the presence in the list of a gene annotated by specific category is random, while in the alternative hypothesis the list is particularly enriched in the considered category.

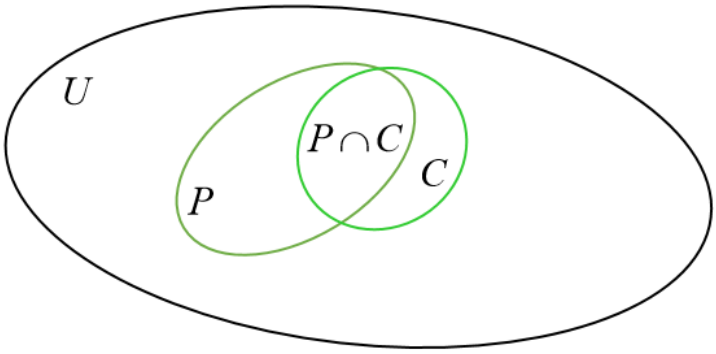


Figure 5.1: Visual representation of hypergeometric test functioning.

Various online tools are available for the calculation of functional enrichment, such as David [27]-[28], Gorilla and Fidea.
In addition to entering the list of genes of interest, you can specify the organism and choose a list of its own. Outgoing to the enrichment test, the result is an association of groups of genes with specific functions and to each category corresponds a value of p. The association is as significant as the pvalue is small.
The base of the functional analysis lies on two large databases: *KEGG* (Kyoto Encyclopedia of Genes and Genomes) and *Gene Ontology*, which attribute specific functions to genes. In particular, KEGG contains information on the metabolic pathway of the cell and on the molecules that are part of it.
In this project, functional analysis is used through David[15], since it is able to represent the results in a simple and understandable way.

5.2 Features selection

In order to create the most significant possible set of features, various levels of features selection have been carried out, which will be described below.

The number of initial features associated with each patient was:

- for *gene expression data*: **60483** genes;
- for *miRNA expression data*: **1881** miRNA;
- for *methylation data*: **14887** methylated genes.

Regarding the gene expression data, a first selection on the features that may be considered as significant to find a gene pattern that identified

[15]DAVID Functional Annotation Bioinformatics Microarray Analysis, reachable at https://david.ncifcrf.gov/home.jsp.

a profile for each class of patients was selecting, from the totality of genes, only the *protein-coding* ones[16].
In doing so, it has gone from a total of 60483 genes to **19814** *protein-coding* genes.

First, the median of all the features for each patient in each class was calculated, in order to have a single value of the expression for each of them.
In the case of *methylation* data, since more isoforms of the same gene are present, a median of the single features in the individual patient was calculated first, and then the median of the medians on the class of patients was calculated.
Then, using the Mann-Whitney test, the p-value associated with the binary comparison between classes was calculated for each feature.

The *Mann-Whitney test*, also known as the 2-sample rank test, can be completed in four steps:

- combine the data from the two samples into one;
- rank all the values, with the smallest observation given rank 1, the second smallest rank 2, etc.;
- calculate and assign the average rank for the observations that are tied (the ones with the same value);
- calculate the sum of the ranks of the first sample (the W-value).

Based on the W-value, the Mann-Whitney test now determines the p-value of the test using a normal approximation, which is calculated as follows:

$$\frac{Z_W = |W - \frac{(n(m+n+1))}{2}| - 0.5}{\sqrt{\frac{mn(m+n+1)}{12}}} \tag{5.5}$$

where W is Mann-Whitney test statistics, n is the size of sample 1 and m is the size of sample 2.
The resulting Z_W value translates for a both-sided test (+/- Z_W) and a normal approximation into a p-value.

[16]A protein-coding gene consists of a promoter followed by the coding sequence for the protein and then a terminator. The promoter is a base-pair sequence that specifies where transcription begins. The coding sequence is a base-pair sequence that includes coding information for the polypeptide chain specified by the gene. The terminator is a sequence that specifies the end of the mRNA transcript.

If there are ties in the data, the p-value is adjusted by replacing the denominator of the above Z statistics by:

$$\sqrt{\frac{nm}{12}\left[(m+n+1)-\frac{\sum_i^l(t_i^3-t_i)}{(m+n)(m+n-1)}\right]} \tag{5.6}$$

where $i = 1, 2, , \mathrm{l}$, l is the number of sets of ties and t_i is the number of tied values in the i-th set of ties.

CNA features selection Also for CNA data various levels of features selection have been performed.
CNA regions can be used to accomplish a genome wide analysis; to do so, in order to compare the alterations of different patients in specific portion of the genome, an homogeneous representation of the signal has been created. This representation will be addressed as *CNA profile*, in Figure 5.2, and it will allow to guess in which regions the signal is detected with a different intensity between the classes.
Given the length of the genome (which contains 3 billions of base pairs), three different resolutions have been tried (1K, 10K and 50K) to understand if doing averages over n portions might return misleading information: a resolution of 1K has been chosen.
This procedure has been done for both amplification and deletion regions. The regions significantly different between the classes has been selected

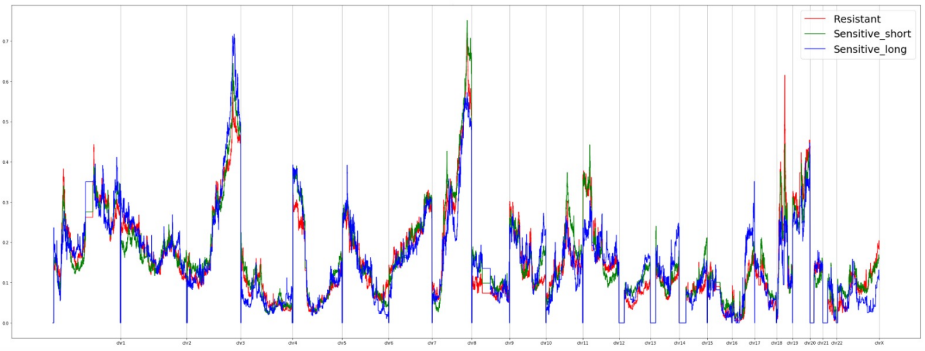


Figure 5.2: CNA amplification profiles with a resolution of 10K for Resistant (red), Sensitive short (green) and Sensitive long (blue).

by the mean of a permutation test, using a threshold for the p-values of 0.005.
This procedure is better explained in [1].

5.3 Classification

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).
Classification belongs to the category of supervised learning where the targets also provided with the input data.

5.3.1 Supervised Learning

Supervised (inductive) learning is the largest, most mature, most widely used sub-eld of machine learning.
Its purpose is to estimate the unknown model that maps known inputs to known outputs, given a training dataset including desired outputs:

$$D = \{< x, t >\} \Rightarrow t = f(x) \tag{5.7}$$

from some unknown function *f*.
It is used to solve problem of *classification*, *regression* and *probability estimation* by the mean of many techniques, as support vector machines and decision trees.
Supervised learning is particularly useful for finding a good approximation of f that generalizes well on test data.
The variables used are divided into:

- **input variables** x, also called features, predictors, attributes;
- **output variables** t, also called targets, responses, labels:
    1. if t is discrete: classification;
    2. if t is continuous: regression;
    3. if t is the probability of x: probability estimation.

The appropriate applications for its use are when:

- there is no human expert;
- humans can perform the task but cannot explain how;
- desired function changes frequently;
- each user needs a customized function *f*.

5.3.2 Evaluation of the classifier

After training the model the most important part is to evaluate the classifier to verify its applicability.

Cross-validation

Over-fitting is a common problem in machine learning which can occur in most models. *K-fold cross-validation* (Figure 5.3) can be conducted to verify that the model is not over-fitted. In this method, the data-set is randomly partitioned into k *mutually exclusive* subsets, each approximately equal size and one is kept for testing while others are used for training. This process is iterated throughout the whole k folds.

Figure 5.3: Graphical explanation of k-fold cross-validation.

Performance measures

Accuracy is a ratio of correctly predicted observation to the total observations:

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + F_N + T_N} \tag{5.8}$$

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations:

$$Precision = \frac{T_P}{T_P + F_P} \tag{5.9}$$

High precision relates to the low false positive rate.

Recall or Sensitivity is the ratio of correctly predicted positive observations to the all observations in actual class:

$$Recall = \frac{T_P}{T_P + F_N} \tag{5.10}$$

f1_score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

$$f1\_score = \frac{2 \cdot (Recall \cdot Precision)}{(Recall + Precision)} \tag{5.11}$$

ROC curve (Receiver Operating Characteristics)

ROC curve is used for visual comparison of classification models which shows the trade-off between the true positive rate and the false positive rate. The area under the ROC curve is a measure of the accuracy of the model. When a model is closer to the diagonal, it is less accurate and the model with perfect accuracy will have an area of 1.0

5.4 Standard methods of classification

5.4.1 Random Forest

Random Forests is an ensemble method, whose structure consists in building multiple trees and merging them together to get a more accurate and stable prediction.
Random Forest can be used either for classication or regression[17]:

- when used for classication, a random forest get a vote from each tree regarding a class, and then classies through using majority vote;
- when used for regression, the predictions from each tree for a target x are gathered and averaged.

When Random Forest is used for regression, it is not as good as in classification: it happens because prediction has not real continuous nature. Random Forest creates multiple decision trees on a subset of features and joins them for getting a prediction whose more stable, as in figure 5.4.

While growing the trees, instead of searching for the most relevant feature while dividing a node, it searches for the best feature among a random subset of features. So, only some features from a random subset will be taken into account when the split of a node will occur [inserisci cit]. Thanks to this randomness, the algorithm is resistant to overfit and able to handle datasets with high dimensionality. Another advantage of the random forest algorithm is its accuracy: for many data sets, it produces a highly accurate classifier; it also runs efficiently on large databases and can handle thousands of input variables without variable deletion. Finally, it gives estimates of what variables are important in the classification [29]. By the other hand, there is little control on what the model does, the typical hyper-parameters to be set in Random Forests regard:

[17] Random Forest can perform these different tasks by combining multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging

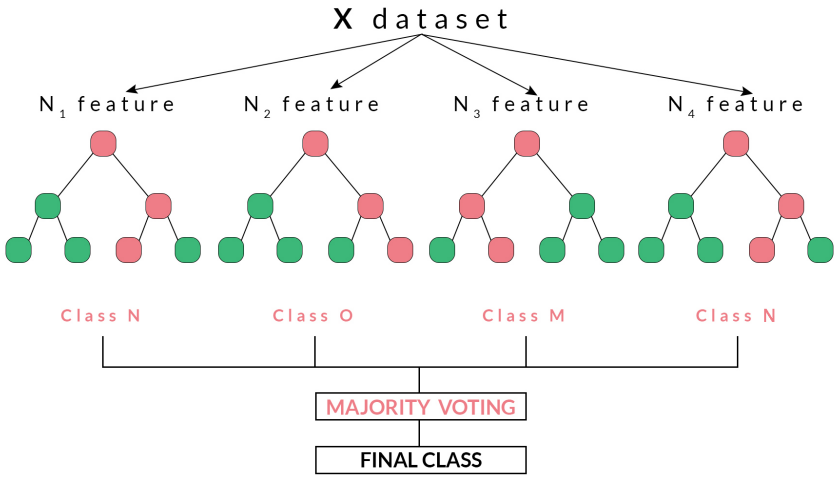


Figure 5.4: Example of how Random Forest works: in the diagram each decision tree has voted or predicted a specific class. The final output or class selected by the Random Forest will be the Class N, as it has majority votes or is the predicted output by two out of the four decision trees.

- the number of trees built, the most important to care;
- the maximum number of features allowed to be tried within a tree;
- the minimum number of leaves required to split an internal node.

To optimize those parameters, a compromise between accurate predictions and time performance is needed [30].

5.4.2 Logistic Regression

Logistic regression is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables, Figure 5.5, [31].
The class prediction is given by:

$$y(x_n) = \sigma(w_0 + x_{n1}w_1 + x_{n2}w_2) \tag{5.12}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{5.13}$$

where σ is called the logistic function or the sigmoid function [inserisci cit]. The loss measure used is the negative log likelihood, while the optimization

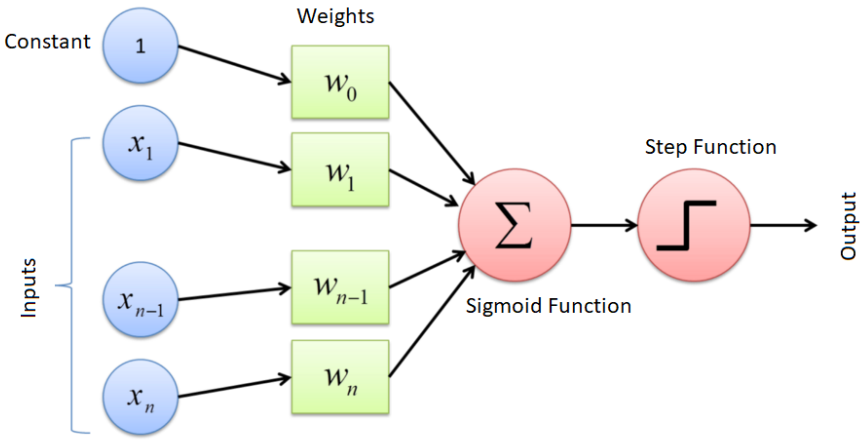


Figure 5.5: Example of how Logistic Regression works with graphical representation of components.

method is the Gradient Descend. The non-linearity of the sigmoid implies that there is no unique solution; what gradient descend returns is the maximum likelihood estimation. Given a threshold value (typically 0.5) it is decided to which class the sample belongs to (see Figure 5.6, [32]).

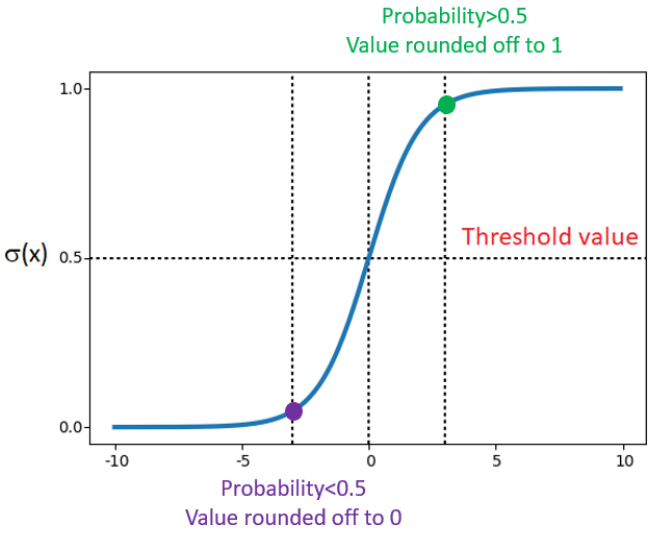


Figure 5.6: Sigmoid fuction.

5.4.3 k-Nearest Neighbor

K Nearest Neighbour is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. It is mostly used to classifies a data point based on how its neighbours are classified. Its initialization phase consists of:

- define a k, number of neighbor;
- define a *distance metric* (Manhattan distance, Hamming distance, etc);
- consider a set of labeled features.

Since it is a non-parametric method, no fitting is required; this algorithm memorizes the training dataset, [30]. This technique is suggested when data are huge and fitting would be time consuming. However, bottleneck arises when recomputing all distances. Classication works according to these easy consecutive steps, given a test sample.

1. Calculate the distance between the test sample and all other training samples.
2. Take the top k training samples based on the distance from the test sample.
3. Each neighbor votes for its label and so the most frequent class among the nearest k is assigned [Figure 5.7].

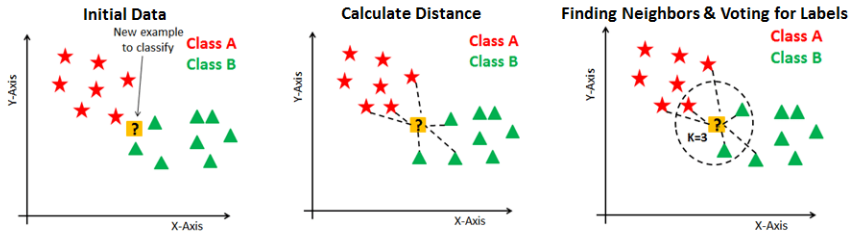


Figure 5.7: KNN working steps.

5.4.4 Adaboost

AdaBoost, short for Adaptive Boosting, is the first practical boosting algorithm [30].
It focuses on classification problems and aims to convert a set of weak

classifiers into a strong one. The final equation for classification can be represented as:

$$F(x) = sign(\sum_{m=1}^{M} \theta_m f_m(x)) \tag{5.14}$$

where f_m stands for the m_θ weak classifier and θ_m is the corresponding weight. It is exactly the weighted combination of M weak classifiers.
The procedure of the AdaBoost algorithm can be summarized as [Figure 5.8]:

- given a data set containing n points, initialize the weight for each data point;
- for iteration $m = 1, ..., M$:
  1. fit weak classifiers to the data set and select the one with the lowest weighted classification error;
  2. calculate the weight for the m_θ weak classifier;
  3. update the weight for each data point by using a normalization factor that ensures the sum of all instance weights is equal to 1.

After M iteration it's possible to get the final prediction by summing up the weighted prediction of each classifier.

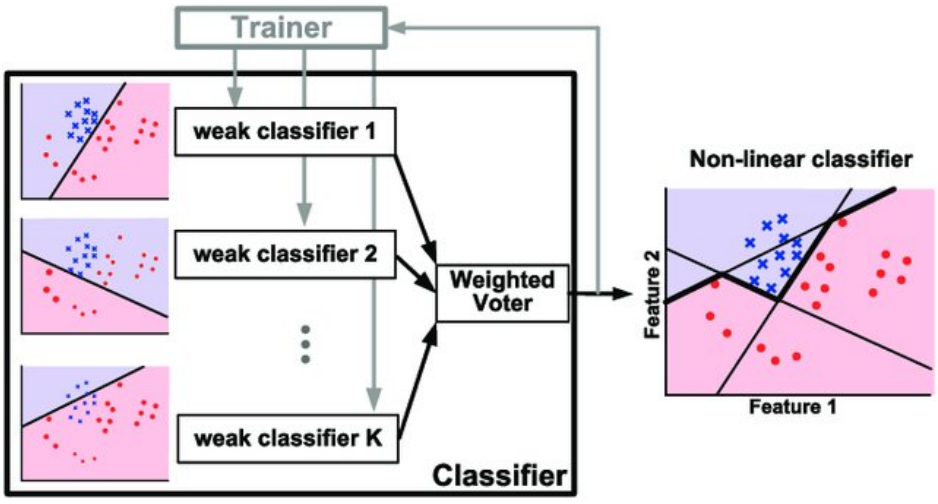


Figure 5.8: Illustration of AdaBoost algorithm for creating a strong classifier based on multiple weak linear classifiers.

5.4.5 Suppor Vector Machine

A Support Vector Machine is a supervised learning model considered one of the best classication methods due to its ability to recognize subtle patterns, [29].
A SVM is made of a subset of training examples x_m (support vectors), a vector of weights for them, and a similarity function k (the kernel). The hypothesis space is:

$$f(x_q) = sign(\sum_{m \in S} a_m t_m k(x_q, x_m) + b) \tag{5.15}$$

where S is the set of indices of the support vectors. SVM enables instancebased learning through the concept of margin, which is the smallest distance between the separating hyperplane and any of the samples.
There exist two kinds of SVM:

- *Linear SVM*, useful for linearly separable data. Receives the original feature space as input and finds the optimal hyperplane that separates the samples of the two classes, trying to maximize the margins (the distance of the decision boundaries). In case data are non-linearly separable, it is added a loss function to account also for misclassified samples. Graphical representation is at figure 5.9.

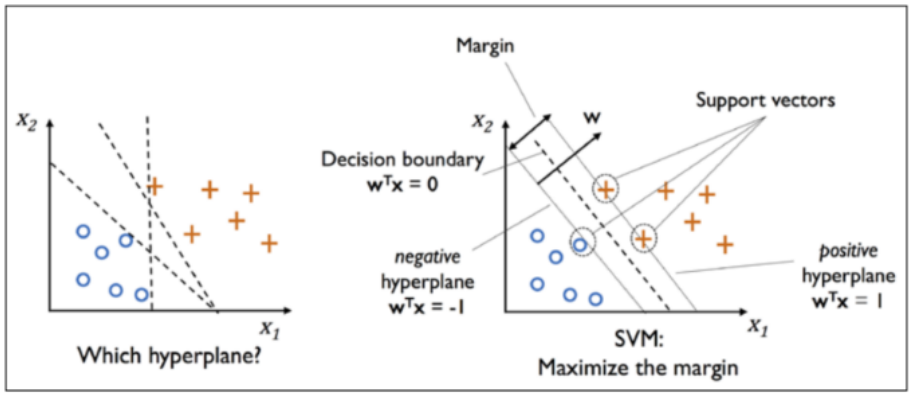


Figure 5.9: Graphical example of margin maximization operated from SVM: the margin is defined as the distance between the separating hyperplane and the training sample that are closest to it.

- *Nonlinear SVM*, that performs a non-linear mapping of the original feature space into a higher dimensional feature space able to better separate data (kernel trick). In this way it is possible to classify nonlinearly separable data, by using a non-linear margin as in Figure 5.10.

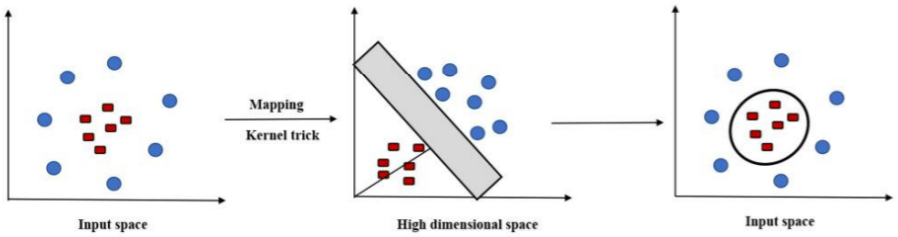


Figure 5.10: How Nonlinear SVM overcomes the problem of data that cannot be separated from Linear SVM. From [33].

5.5 Survival Analysis

Sometimes it's interesting to investigate how a risk factor or treatment affects time to disease or some other event, and if a study dropout is present therefore there will be subjects who we are not sure if they had disease or not.
In these cases, *survival analysis* is used to analyze data in which the time until the event is of interest. The response is often referred to as a failure time, survival time, or time to event, [34].
The survival time response:

- is always ≥ 0;
- is usually continuous;
- may be incompletely determined for some subjects, i.e., for some subjects is possible to know their survival time, which was at least equal to some time t. Whereas, for other subjects, their exact time of event is known;
- incompletely observed responses are called **censored**.

If there is no censoring, standard regression procedures could be used. However, these may be inadequate because:

- time to event is restricted to be positive and has a skewed distribution;
- the probability of surviving past a certain point in time may be of more interest than the expected time of event;
- the hazard function, used for regression in survival analysis, can lend more insight into the failure mechanism than linear regression.

5.5.1 Censoring

Censoring is present when some information about a subjects event time are available, but the exact event time is not known.

For the analysis methods that will be discussed to be valid, censoring mechanism must be independent of the survival mechanism.
There are generally three reasons why censoring might occur:

- a subject does not experience the event before the study ends;
- a subject is lost to follow-up during the study period;
- a subject withdraws from the study;

the reasons just explaines are all examples of right-censoring[18]. Regardless of the type of censoring, it must be assumed that it is non-informative about the event; that is, the censoring is caused by something other than the impending failure.

5.5.2 Terminology and notation

- T denotes the response variable, $T \geq 0$;
- the *survival function* (Figure 5.11) is:

$$S(t) = Pr(T > t) = 1 - F(t) \tag{5.16}$$

This function gives the probability that a subject will survive past

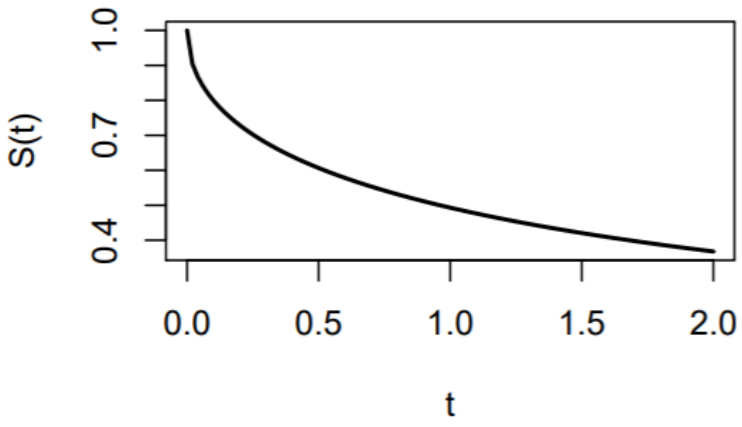


Figure 5.11: Survival analysis: survival function.

time t.
As t ranges from 0 to ∞ the survival function has the following properties:

[18]Right-censonring can be of different types, as *fixed type I censoring*, *random type I censoring*, and *type II censoring*.

1. it is *non-increasing*;
2. at time $t = 0$, S(t) = 1: the probability of surviving past time 0 is 1;
3. at time t = ∞, $S(t) = S(\infty) = 0$: as time goes to infinity, the survival curve goes to 0.

In theory, the survival function is smooth. In practice, events on a discrete time scale are observed.

- the hazard function, *h(t)*, is the instantaneous rate at which events occur, given no previous events:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{5.17}$$

5.5.3 Survival data

In the following will be described how survival data with censoring are recorded and represented:

- T_i denotes the response for the ith subject;
- C_i denotes the censoring time for the ith subject;
- δ_i denotes the event indicator:

$$\delta_i = \begin{cases} 1 : \text{if the event was observed}, (T_i \leq C_i)) \\ 0 : \text{if the response was censored}, (T_i > C_i)) \end{cases}$$

- the observed response is:

$$Y_i = min(T_i, C_i) \tag{5.18}$$

5.5.4 Kaplan-Meier survival estimate

The Kaplan-Meier (KM) [35] method is a non-parametric method used to estimate the survival probability from observed survival times. The survival probability at time t_i, *S(t_i)*, is calculated as follow:

$$S(t_i) = S(t_{i-1})(1 - \frac{d_i}{n_i}) \tag{5.19}$$

where *S(t_{i-1})* is the probability of being alive at t_{i-1}, n_i is the number of patients alive just before t_i, d_i is the number of events at t_i, t_0=*0* and *S(0)=1*.
The estimated probability (*S(t)*) is a step function that changes value only at the time of each event. Its also possible to compute confidence intervals for the survival probability.
The KM survival curve (Figure 5.12), a plot of the KM survival probability against time, provides a useful summary of the data that can be used to estimate measures such as median survival time, [36].

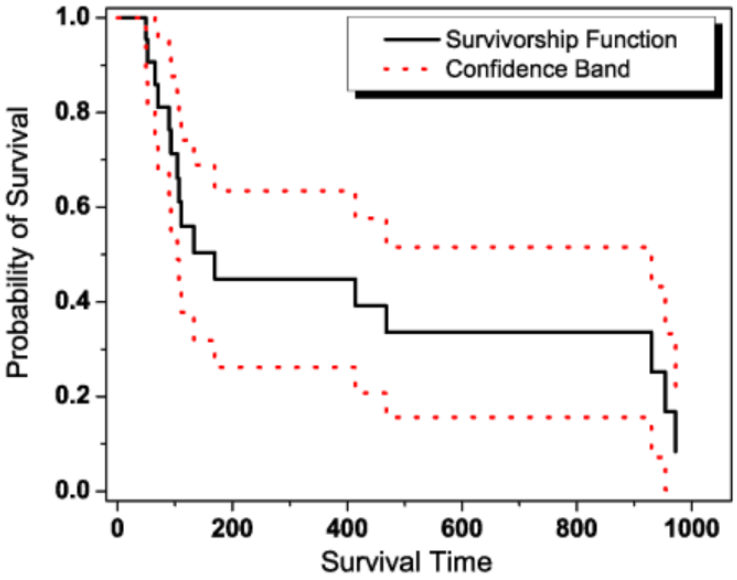


Figure 5.12: Survival analysis: Kaplan-Meier curve.

5.5.5 Performance measure: Concordance Index

The *C-statistic* (sometimes called the *concordance* statistic or *C-index*) is a measure of goodness of fit for binary outcomes in a logistic regression model. In clinical studies, the C-statistic gives the probability a randomly selected patient who experienced an event (e.g., a disease or condition) had a higher risk score than a patient who had not experienced the event. It is equal to the area under the Receiver Operating Characteristic (ROC) curve and ranges from 0.5 to 1.
Some consideration about the C-index:

- a value **below 0.5** indicates a **very poor model**;
- a value of 0.5 means that the model is no better than predicting an outcome than random chance;
- values over 0.7 indicate a good model;
- values over 0.8 indicate a strong model;
- a value of **1** means that the model **perfectly predicts** those group members who will experience a certain outcome and those who will not.

A *weighted c-index* is used when the cost of failing to predict a positive outcome (i.e., a test for cancer) is higher than benefit of correctly predicting

a negative outcome. Weighting penalizes models that result in small probability differences for positive and negative outcomes, but doesnt change the value of the C-statistic.
The C-statistic is sometimes paired with a confidence interval: in general, any result is not significant if it includes 0.5, even if it includes the relevant C-statistic.

5.5.6 CoxPH Model

The *Cox proportional-hazards model* [34] is a regression model commonly used statistical in medical research for investigating the association between the survival time of patients and one or more predictor variables. It works for both quantitative predictor variables and for categorical variables. Furthermore, the Cox regression model extends survival analysis methods to assess simultaneously the effect of several risk factors on survival time.
The purpose of the model is to evaluate simultaneously the effect of several factors on survival. In other words, it allows us to examine how specified factors influence the rate of a particular event happening (e.g., death) at a particular point in time. This rate is commonly referred as the *hazard rate*. Predictor variables (or factors) are usually termed *covariates* in the survival-analysis literature.
The Cox model is expressed by the hazard function, that can be interpreted as the risk of dying at time *t*. It can be estimated as (Chapter 5, page 70):

$$h(t) = h_0(t) \times exp(b_1x_1 + b_2x_2 + ... + b_px_p) \tag{5.20}$$

where t presents the survival time, *h(t)* is the hazard function determined by a set of p covariates (x_1,x_2,...,x_p), the coefficients (b_1,b_2,...,b_p) measure the impact of covariates; the term h_0 is called the *baseline hazard*: it corresponds to the value of the hazard if all the x_i are equal to zero (the quantity *exp(0)* equals 1). The t in *h(t)* reminds that the hazard may vary over time.
The Cox model can be written as a multiple linear regression of the logarithm of the hazard on the variables x_i, with the baseline hazard being an intercept term that varies with time.
The quantities *exp(b_i)* are called **hazard ratios** (HR). A value of b_i greater than zero, or equivalently a hazard ratio greater than one, indicates that as the value of the ith covariate increases, the event hazard increases and thus the length of survival decreases.
Put another way, a hazard ratio above 1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival.
In summary,

- HR = 1: No effect;

- HR < 1: Reduction in the hazard;
- HR > 1: Increase in Hazard.

So, the Cox model is a proportional-hazards model: the hazard of the event in any group is a constant multiple of the hazard in any other. This assumption implies that the hazard curves for the groups should be proportional and cannot cross.

Correlation Matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix (e.g., Figure 5.13) is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analysis.

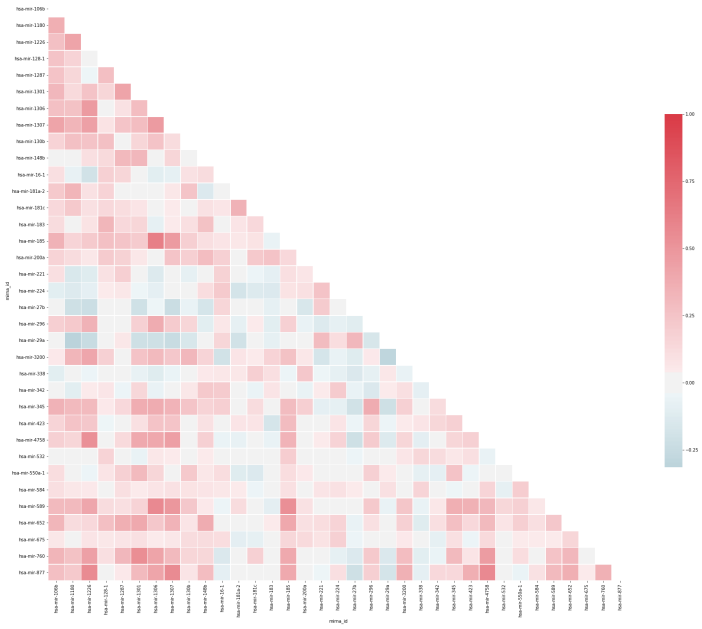


Figure 5.13: Correlation Matrix of features extracted with the "Mild B-H Correction", without collinearities, miRNA expression data.

5.6 Windowing

As explained before, many different data types with distinct biological meanings have been used to discriminate patients in classes.
In order to understand where the selected features were located, a new interesting method has been developed to visualize on each chromosome how many and which features were expressed in there: this method will be referred to as *windowing*.
The original idea is that to allocate a bar, which corresponds to the feature, on a graph that has chromosome length on x-axis and the feature's expression on y-axis. In order to made the graph printable, it has been performed a reduction of 1:10000 on the bins that make up the chromosome length, and regarding the bar heigth, the normalization is done by taking the maximum value associated with the features expressed on that gene and bringing it to 1, and comparing all the others to that value.
The first experiment to use this method was to plot, chromosome by chromosome, for each class, all the features coming from each binary comparison, in order to observe which were the chromosomes with greater density of significant features and which was the expression of those features at each comparison.

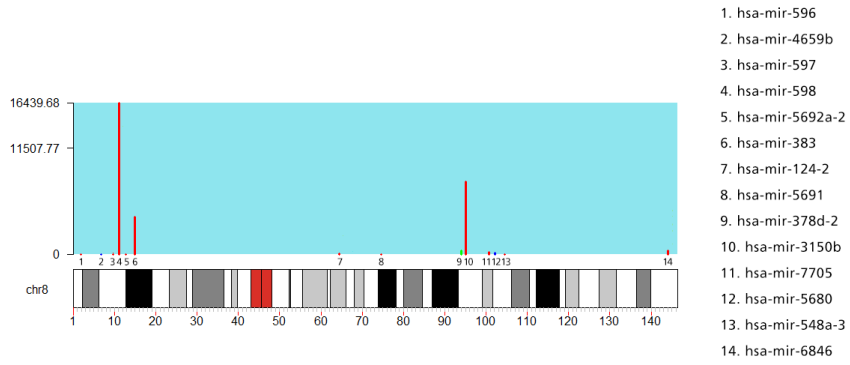


Figure 5.14: MiRNA features expressed on chromosome 8 for Resistant class: in red, features considered relevant in distinguishing Sensitive Long and Sensitive Short classes are shown; in blue, features considered relevant in distinguishing Sensitive Long and Resistant classes are shown; in green, features considered relevant in distinguishing Resistant and Sensitive Short classes are shown.

An example is shown in Figure 5.14, where for *Resistant* class, all the features extracted from three binary comparisons (*Resistant vs Sensitive Short*, *Resistant vs Sensitive Long* and *Sensitive Long vs Sensitive Short*) for performing survival analysis have been plotted; it has been chosen to

plot chromosome 8 because it seems the chromosome having much number of features in it.

This method can be used not only to assess which chromosome has greater density of relevant features, but also which are the regions of the genome considered significant through the use of various types of data.
Given the use in this analysis of regions of alteration and deletion due to copy number alterations, the windowing was also developed to understand if, which and how many features selected among the other types of data (protein coding genes, miRNAs and methylated genes) were contained in those regions.
The significance of understanding if multiple types of features overlap each other in the genome is associated to the possibility of identifying with confidence genomic regions of biological relevance that can be adequate predictors to distinguish patients from one another.
This could be considered a *method of intersection*: while the classifier works by using *all the features* selected by significance - this allowing to state it as a *method of union*, the windowing can grant the recognition of regions in which are included different genomic data, so to reduce the amount of processing to do on the genome, suggesting portions of moderate size to be analyzed.
Being able to identify genomic regions would lead to a speed-up in the analysis procedure, and a lower cost in acquiring and manipulating data for the diagnosis of this disease.

Chapter 6

Computational Methods

The response to therapy can be assessed both by trying to predict time to relapse and by attempting to distinguish between classes of patients.
However, both methodologies have in common a fundamental point, which concerns the features selection, performed through different steps for survival analysis and classification. The starting data concern gene expression, miRNA expression, DNA methylation and copy number alteration, related to the pre-treatment phase. The selection of features suitable for predicting chemoresistance is fundamental to identify a distinctive genomic pattern between therapy-resistant and therapy-sensory patients.
For this reason, it was decided to investigate the different expressions of data-types in order to verify which features show a greater variation in their expressions, relative to the different classes under consideration.
As for genes, a first selection has restricted the genomic material to be evaluated to only protein-coding genes, which result in many cellular functions and biological activities.
The DNA methylation data are those that present a greater variability, since for each methylated gene its isoforms are present, each having an expression value of its own.
Data related to miRNAs have been inserted due to their implication in downregulating their target genes: this change in gene expression can modify drug response.
All these types of data, however, suffer from a modification due to the administration of the therapy; for this reason, given that one of the main ambitions of the study of high-grade serous ovarian cancer is to identify patients who will develop resistance to therapy with an early diagnosis with good accuracy, among the analyzed data already mentioned was also introduced another data-type, related to the genomic regions of copy number alteration.
These regions, which cover a large part of the genome, show an alteration in the number of copies, which are considered early events, therefore eligible predictors of the response to therapy; the identification of precise

regions, in which it would in any case be possible to identify genes, would lead both to the possibility of using an integrative approach and to the possible development of new therapeutic options based on the targeting of those factors involved in the development of chemoresistance.
To better integrate the significant features from the point of view of the drug-response prediction, various levels of analysis have been carried out, starting from the features selection.

6.1 Features selection

To assess the significance of the features in terms of expression, the Mann-Whitney test was performed and then the relative p-value has been obtained for each feature for each binary comparison.
First of all, in order to understand which features could be considered significant based on their nominal p-value, a treshold equals to 0.05 was initially set, meaning statistically significant, which was then lowered twice (a second treshold = 0.005 and a third threshold = 0.0005, to evaluate the significance of the expression of that feature based on its p-value).
Later, the Bonferroni and Benjamini-Hochberg multiple test corrections were used: we chose to make both these corrections (therefore the most conservative and the least conservative) in order to verify the presence of biologically relevant features.
Unfortunately, almost no features outweighed the corrections: this supports the biological assumptions about the extreme variety of gene mutations within the development of this disease, which makes the evolution of this type of cancer very difficult to delineate precisely.
For this reason, the descriptive equations of the multiple test corrections have been slightly modified (Chapter 5, page 55), in order to still make a correction on the p-value, even if in a more bland way.
The reasoning behind this features selection procedure lies in the fact that it was preferred to have available various sets of features selected based on their p-value in order to evaluate the performance of each single set: not always a set including a large number of features will return the best reliability, as on too few features it will not be useful to develop an enrichment analysis since no pathway will be identified by them.
For performing the survival analysis with the most suitable features, some other steps have been added for each set:

1. first of all, the dataset is created, taking for each patient the expression value relative to the selected feature, and it is normalized. Then, a 5-fold cross validation is ran to evaluate the model;

2. at this point, in order to improve the performance obtained, the correlation matrix is computed and for the features that have a correlation greater than or equal to 0.9, only one candidate is selected.

Also, a second step of selection is performed, removing the features with a too low variance.
Again, a 5-fold cross validation is performed to evaluate the model;

3. finally, a further analysis is carried out to understand which of the selected features can be considered relevant for the prediction of the relapse time. This is done fitting the model on the entire dataset and checking for each features the exponential of the coefficient and its confidence interval. In particular, it is verified if the interval is either all below 1 or all above 1.
A 5-fold cross validation is ran one last time, using only the features that respect the condition required to be considered significant.

For CNA data only, the features selection is slightly different; starting from the *CNA profiles* (Chapter 5, section 5.2), various levels of selection have been carried out:

- we tried to select only the regions in which the average CNA value (for amplification and deletion, indipendently) was different between the classes;
- we moved to evaluate the real difference between regions by the mean of a *z-test*, a parametric test that verifies if the average values of two distributions are equal or not;
- a *permutation test* was implemented in order to find genomic regions in which the CNA profiles belonging to the classes have different alteration intensities.

A permutation test is a type of statistical significance test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points.
In order to combine all types of features (from gene and miRNA expression data, DNA methylation data and CNA data), the functional associations between features have been evaluated.
MiRNAs can play the role of downregulators in the gene regulation of their targets, so the causal relationships between the selected features were evaluated, through the use of interactions experimentally known[19] and high confidence predictions in silico[20].

6.2 Survival Analysis

One way to predict the response to therapy is to perform a survival regression to prognosticate the timing of patients' relapse.

[19]From http:// mirtarbase.mbc.nctu.edu.tw/php/index.php, [37].
[20]From https://targetscan.org, [38].

Considering that patients are divided into classes based on the time in which they show resistance to therapy, knowing the time at relapse also implies being able to distinguish the categories to which the patients belong.
First, in order to obtain a visual representation of the relapse timing of the patients' groups, the Kaplan-Meier curve has been plotted: in Figure 6.1, the y-axis represents the percentage of censored data and the x-axis is the time in months.

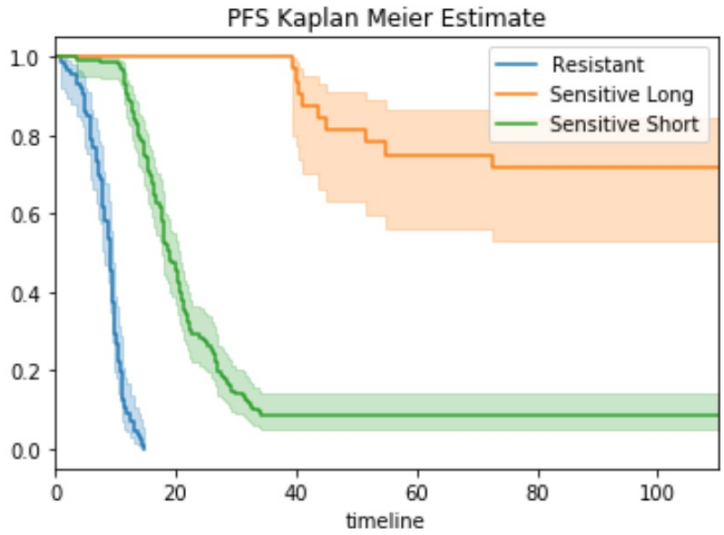


Figure 6.1: Progression Free Survival Kaplan Meier Estimate: time to relapse of each class is represented by different color described in label.

Then, the survival regression has been performed on CoxPH Model and concordance index has been computed, as performance measure.
This regression did not reach satisfactory values in terms of the predictive model, so we switched to the classification method.

6.3 Classification

To distinguish patients based on their belonging to a specific class, and thus predict the response to therapy, a classifier has been implemented, which uses the features extracted as explained above.
The classification was conducted separately for each data type and each set of features, to choose the set through which the distinction between classes was clearer.
Different types of classification algorithms have been used on each feature set to select the most suitable to treat our data and then the one with the best performance has been executed.

The results obtained through this method using the different data-types independently were quite good, as regards the distinction of the two classes with greater variety of expression, Resistant vs Sensitive Long; to improve them and try to clearly distinguish resistant patients from all sensitive ones, it was decided to implement a classifier that used all the best features extracted from each data-type (gene expression data, miRNA expression data, DNA methylation data and CNA data). First, a dataset was built in which there were only patients for whom all four types of data were available; this has reduced the number of patients associated with each class.
Subsequently, based on the results obtained through the different classification algorithms used previously on the individual data-types, the best set of features was chosen for each binary comparison. The dataset created for the classification is a matrix that has the patients in the rows and the features in the columns and each value corresponds to the relative expression data.
Once this is done, the support vector machine has been chosen as a classification algorithm because they work well in genomic analysis with many features and few samples, as in our case.
We executed a hyperparameter tuning to choose the best parameters for SVM and then we performed the classification: hyperparameter tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process, so many classification algorithms have been tried with relative hyperparameters, and then the one through which the best results are obtained has been choosen to perform the integrative classification.
From the merging of all types of features promising results have been obtained, discussed more in detail in Chapter 7.

Chapter 7

Results

7.1 Computational results

In this section, the best results obtained through the analysis methods implemented will be discussed.

7.1.1 Best performances of survival analysis

Concordance index

As already discussed in Chapter 5, subsection 5.5, page 72, the index of concordance is a "global" index for validating the predictive ability of a survival model and it is equivalent to a rank correlation.
Giving the fact that the index is not calculated for every observation/subject, c-index is not perfect because the predicted probabilities of outcome might be way off the observed probabilities of outcome, but as long as the ranking of these probabilities show that higher probabilities are associated with the event, discrimination/c-index might be relatively good.
Unfortunatly, the outcomes obtained for concordance index are quite poor, because none of them reach at least a value of C.I. = 0.7, which would mean a good model was used.
Actually, this model indicates no better prediction than random chance; for those reasons, survival analysis is not considered the best method to deal with this type of problem.

Table 7.1: Concordance index computed for each data type.

Data type	**Concordance Index**
Gene Expression data	0.61
MiRNA data	0.54
Methylation data	0.57

Survival curves

The fact of having a poor predictive model is also reflected on the graphic aspect of the survival functions: as it can be seen in Figure 7.1, 7.2 and 7.3, all the predicted survival functions show almost the same shape.
This result is disappointing to say the least, considering that the patients belonging to the three classes considered have extremely different relapse times; considering only that for the class of patients resistant to therapy, the range of relapse can reach a maximum of 6 months post-treatment, while for a patient belonging to the sensitive class, the range starts from 6 months and can reach up to over 3 years.
The predicted survival functions have confirmed the above, considering the results of the concordance index, which means that to be used to address problems of this type, the survival analysis must be implemented differently, on less discontinuous data; a chance to improve these results could be finding alternative ways to select the features.

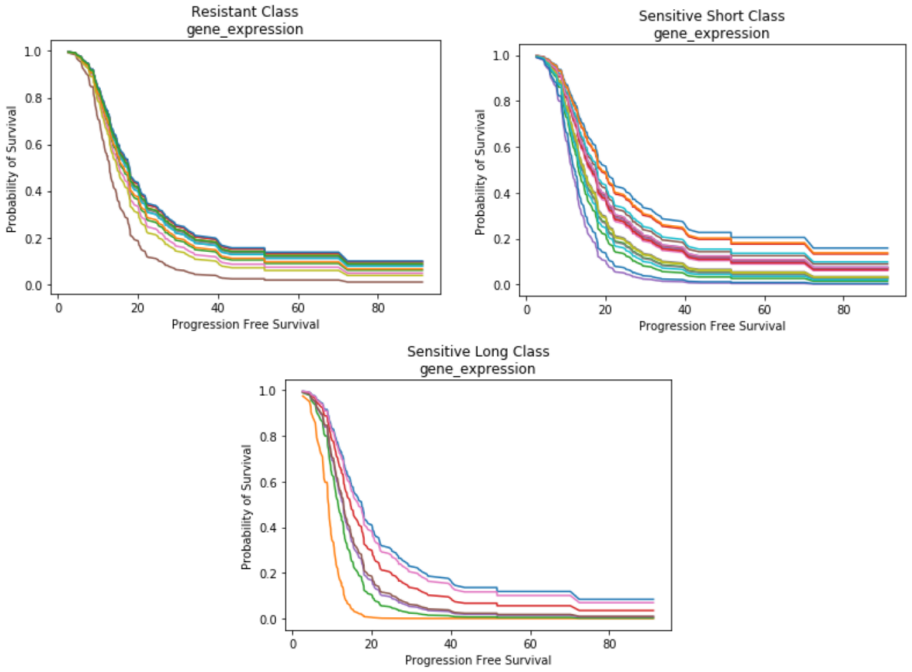


Figure 7.1: Prediction survival function of each class for gene expression data.

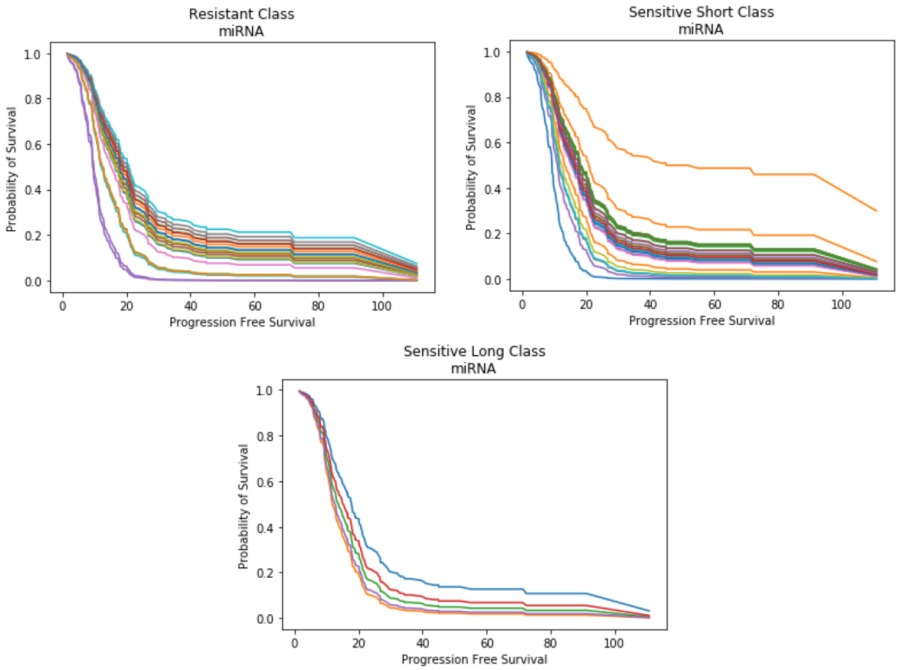


Figure 7.2: Prediction survival function of each class for miRNA data.

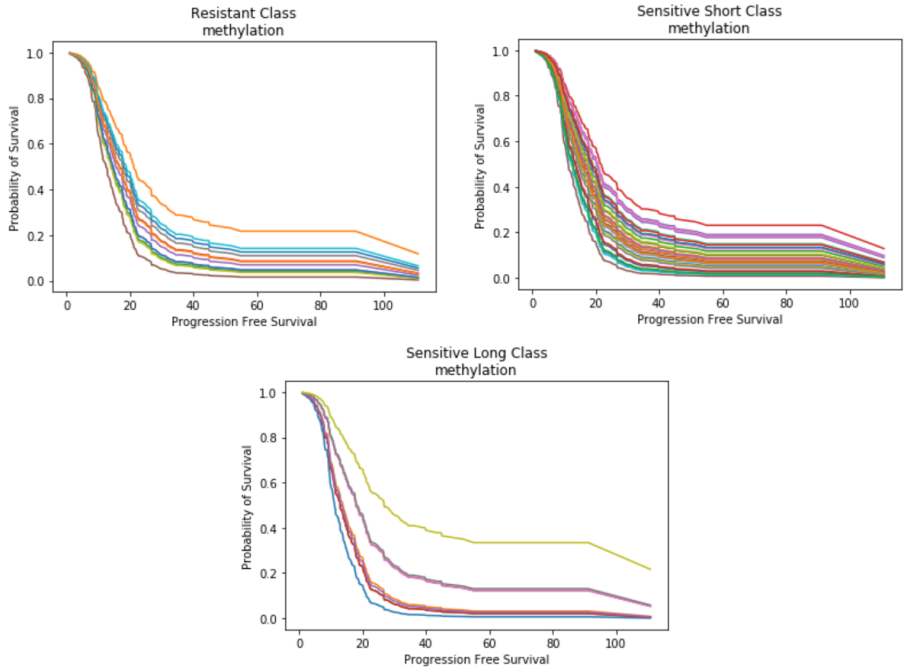


Figure 7.3: Prediction survival function of each class for methylation data.

7.1.2 Best performances of classifier

The feature sets with the best performances for each type of data analyzed for each comparison will be listed below.

Gene expression data

For each binary class comparison, the following will be described: the type of features selection that was used, the number of resulting features, the best performance obtained and the classification algorithm, with its related parameters, involved to obtain them.
In order to also have a visual result of the performances, ROC curves will be inserted.

Sensitive Long VS Sensitive Short The best classifier has been obtained using:

- features with **p_value** $<$ **0.0005**,
- **Support Vector Machine** as classification algorithm, with *C=10* and *kernel=rbf*.

The number of features resulting from this selection is **35**.
The visual results are shown in the ROC curve below (see Figure 7.4(c)).

Sensitive Long VS Resistant The best classifier has been obtained using:

- features with **p_value** $<$ **0.0005**,
- **Random Forest**, with *n_estimators=200* and *max_depth=15*.

The number of features resulting from this selection is **26**.
The visual results are shown in the ROC curve below (see Figure 7.4(a)).

Sensitive Short VS Resistant The best classifier has been obtained using:

- features with **p_value** $<$ **0.05** after applying the *Mild Bonferroni Correction*,
- **Random Forest**, with *n_estimators=200* and *max_depth=15*.

The number of features resulting from this selection is **40**.
The visual results are shown in the ROC curve below (see Figure 7.4(b)).

All Sensitive VS Resistant The best classifier has been obtained using:

- features with **p_value** $<$ **0.05** after applying the *Mild Bonferroni Correction*,
- **Random Forest**, with *n_estimators=200* and *max_depth=15*.

The number of features resulting from this selection is **20**.
The visual results are shown in the ROC curve below (see Figure 7.4(d)).

The performance of the classifier for each comparison are shown in Table 7.2.

Table 7.2: Performance related to gene expression data, using the most significant feature sets.

Gene Expression data Comparison	Precision	Recall	Accuracy	f_1 score
Sensitive Long vs Sensitive Short	0.606 ± 0.2	0.766 ± 0.2	0.801 ± 0.1	0.751 ± 0.1
Sensitive Long vs Resistant	0.862 ± 0.2	0.933 ± 0.1	0.845 ± 0.2	0.806 ± 0.3
Sensitive Short vs Resistant	0.718 ± 0.2	0.55 ± 0.2	0.766 ± 0.1	0.723 ± 0.2
Resistant vs All Sensitive	0.71 ± 0.2	0.37 ± 0.1	0.765 ± 0.05	0.474 ± 0.1

Considerations From the results shown in the Table 7.2, which refer to the best performances obtained with each set of features, it is possible to make some considerations:

- the binary comparison that gives the best performance is the one that discerns between the two most different classes, both in terms of relapse and gene profile: *Sensitive Long vs Resistant*;
- in this case, the binary comparisons between *Sensitive Long vs Sensitive Short* and between *Sensitive Short vs Resistant* give similar performance, although the second shows slightly worse results: this could have been predictable, because patients belonging to the Sensitive Short class can manifest relapse in very different times, therefore also very close (we speak of a difference of a few months) compared to those of therapy-resistant patient;
- The binary comparison between the Resistant class and the totality of the patients sensitive to the therapy returned good performances, despite the classes are strongly unbalanced between each other.

miRNA data

Sensitive Long VS Sensitive Short The best classifier has been obtained using:

- features with **p_value** $<$ **0.005**,
- **Random Forest** as classification algorithm, with *n_estimators=200* and *max_depth=15*.

The number of features resulting from this selection is **17**.
The visual results are shown in the ROC curve below (see Figure 7.5(c)).

Sensitive Long VS Resistant The best classifier has been obtained using:

- features with **p_value** $<$ **0.005**,
- **Random Forest** as classification algorithm, with *n_estimators=200* and *max_depth=15*.

The number of features resulting from this selection is **21**.
The visual results are shown in the ROC curve below (see Figure 7.5(a)).

Sensitive Short VS Resistant The best classifier has been obtained using:

- features with **p_value** $<$ **0.005**,
- **Random Forest** as classification algorithm, with *n_estimators=200* and *max_depth=15*.

The number of features resulting from this selection is **12**.
The visual results are shown in the ROC curve below (see Figure 7.5(b)).

All Sensitive VS Resistant The best classifier has been obtained using:

- features with **p_value** $<$ **0.005**,
- **Random Forest** as classification algorithm, with *n_estimators=200* and *max_depth=15*.

The number of features resulting from this selection is **11**.
The visual results are shown in the ROC curve below (see Figure 7.5(d)).

The performance of the classifier for each comparison are shown in Table 7.3.

Table 7.3: Performance related to miRNA expression data, using the most significant feature sets.

miRNA data Comparison	Precision	Recall	Accuracy	f_1 score
Sensitive Long vs Sensitive Short	0.666 ± 0.4	0.35 ± 0.3	0.837 ± 0.1	0.68 ± 0.3
Sensitive Long vs Resistant	0.77 ± 0.1	0.9 ± 0.1	0.75 ± 0.1	0.68 ± 0.3
Sensitive Short vs Resistant	0.7 ± 0.3	0.5 ± 0.3	0.736 ± 0.1	0.68 ± 0.2
Resistant vs All Sensitive	0.77 ± 0.3	0.37 ± 0.2	0.745 ± 0.1	0.45 ± 0.1

Considerations From the results shown in the Table 7.3, which refer to the best performances obtained with each set of features, it is possible to make some considerations:

- differently from what happened for the gene expression data, in this case the performances of all the binary comparisons are extremely similar;
- the binary comparison between *Sensitive Short vs Resistant* classes remains the one with the worst performances, as expected;
- the comparison between the Resistant class and that inclusive of all patients sensitive to therapy improves slightly compared to the results seen for the gene expression data; this indicates that the selected features of miRNA expression are able to better distinguish the two classes: this is positive, since the final aim of the project is to distinguish therapy-resistant patients from all others.

Methylation data

Sensitive Long VS Sensitive Short The best classifier has been obtained using:

- features with **p_value** $<$ **0.05** after applying the *Mild Bonferroni Correction*,
- **Random Forest** as classification algorithm, with *n_estimators=200* and *max_depth=15*.

The number of features resulting from this selection is **10**.
The visual results are shown in the ROC curve below (see Figure 7.6(c)).

Sensitive Long VS Resistant The best classifier has been obtained using:

- features with **p_value** $<$ **0.0005**,

- **Random Forest** as classification algorithm, with *n_estimators=200* and *max_depth=15*.

The number of features resulting from this selection is **18**.
The visual results are shown in the ROC curve below (see Figure 7.6(a)).

Sensitive Short VS Resistant The best classifier has been obtained using:

- features with **p_value < 0.005**,
- **Random Forest** as classification algorithm, with *n_estimators=200* and *max_depth=15*.

The number of features resulting from this selection is **31**.
The visual results are shown in the ROC curve below (see Figure 7.6(b)).

All Sensitive VS Resistant The best classifier has been obtained using:

- features with **p_value < 0.005**,
- **Random Forest** as classification algorithm, with *n_estimators=200* and *max_depth=15*.

The number of features resulting from this selection is **65**.
The visual results are shown in the ROC curve below (see Figure 7.6(d)).

The performance of the classifier for each comparison are shown in Table 7.4.

Table 7.4: Performance related to methylation data, using the most significant feature sets.

Methylation data Comparison	**Precision**	**Recall**	**Accuracy**	**f_1 score**
Sensitive Long vs Sensitive Short	0.575 ± 0.4	0.383 ± 0.3	0.821 ± 0.1	0.669 ± 0.3
Sensitive Long vs Resistant	0.79 ± 0.1	0.9 ± 0.1	0.77 ± 0.1	0.7 ± 0.2
Sensitive Short vs Resistant	0.65 ± 0.3	0.32 ± 0.2	0.69 ± 0.1	0.59 ± 0.2
Resistant vs All Sensitive	0.79 ± 0.3	0.35 ± 0.1	0.78 ± 0.1	0.47 ± 0.1

Considerations From the results shown in the Table 7.4, which refer to the best performances obtained with each set of features, it is possible to make some considerations:

- in this case, compared to the two precedents seen, the binary comparison between *Sensitive Short vs Resistant* is even worse, showing worse performance than all the other comparisons and also with respect to the same comparison carried out using different data types;

- the accuracy performances concerning the distinction between the Resistant class and the Sensitive one are the best views to date; however, accuracy is a great measure only when symmetric data sets are available, where values of false positive and false negatives are almost same. Therefore, you have to look at other metrics to evaluate the performance of your model, and unfortunately considering all the parameters constituting the performance of the classifier, the final result is that of a classically reliable classifier.

CNA data

As for the comparisons performed using CNA data as features, the best classifier was always obtained through:

- features selected from **permutation test**, with **p_value** $<$ **0.005** and **resolution = 1K**, from which we extracted:

  1. for **Resistant vs Sensitive Long** comparison: **153** features;
  2. for **Resistant vs Sensitive Short** comparison: **236** features;
  3. for **Sensitive Long vs Sensitive Short** comparison: **128** features;
  4. for **Resistant vs Sensitive** comparison: **225** features;

- **Support Vector Machine**, with:

  1. for **Resistant vs Sensitive Long** comparison: *C=1* and *kernel=rbf*;
  2. for **Resistant vs Sensitive Short** comparison: *C=1* and *kernel=rbf*;
  3. for **Sensitive Long vs Sensitive Short** comparison: *C=10* and *kernel=rbf*;
  4. for **Resistant vs Sensitive** comparison: *C=10* and *kernel=rbf*.

Considerations From the results shown in the Table 7.5, which refer to the best performances obtained with each set of features, it is possible to make some considerations:

Table 7.5: Performance related to CNA data, using the most significant feature sets.

CNA data Comparison	Precision	Recall	Accuracy	f_1 score
Sensitive Long vs Sensitive Short	0.30 ± 0.26	0.32 ± 0.21	0.74 ± 0.07	0.28 ± 0.1
Sensitive Long vs Resistant	0.89 ± 0.07	0.88 ± 0.10	0.83 ± 0.08	0.88 ± 0.2
Sensitive Short vs Resistant	0.53 ± 0.11	0.57 ± 0.11	0.64 ± 0.09	0.54 ± 0.09
Resistant vs All Sensitive	0.51 ± 0.10	0.61 ± 0.19	0.68 ± 0.07	0.54 ± 0.11

- the binary comparison that gives the best performance is the one that discerns between the two most different classes, both in terms of relapse and gene profile: *Sensitive Long vs Resistant*;

- the binary comparison between the Resistant class and the totality of the patients sensitive to the therapy returned better results than *Sensitive Short vs Resistant* comparison, despite the classes are strongly unbalanced between each other.

Merging data types

Finally, a last classifier was built by combining the best features of each data type (gene expression data, miRNA expression data, methylation data and CNA data).

Sensitive Long VS Sensitive Short The best classifier has been obtained using:

- **Support Vector Machine**, with *C=1* and **kernel=rbf**.

The number of features is equal to **190**, of which 128 of CNA, 35 of gene expression, 17 of miRNA and 10 of methylation.
The visual results are shown in the ROC curve below (see Figure 7.7(c)).

Sensitive Long VS Resistant The best classifier has been obtained using:

- **Support Vector Machine**, with *C=10* and **kernel=rbf**.

The number of features is equal to 218, of which 153 of CNA, 26 of gene expression, 21 of miRNA and 18 of methylation; some of them were removed due to a *correlation = 1* with other features, thus the final number is **213**.
The visual results are shown in the ROC curve below (see Figure 7.7(a)).

Sensitive Short VS Resistant The best classifier has been obtained using:

- **Support Vector Machine**, with *C=10* and **kernel=rbf**.

The number of features is equal to 319, of which 236 of CNA, 40 of gene expression, 12 of miRNA and 31 of methylation; some of them were removed due to *correlation = 1* with other features, thus the final number is **310**.
The visual results are shown in the ROC curve below (see Figure 7.7(b)).

All Sensitive VS Resistant The best classifier has been obtained using:

- **Support Vector Machine**, with *C=1* and **kernel=linear**.

The number of features is equal to 321, of which 225 of CNA, 20 of gene expression, 11 of miRNA and 65 of methylation; some of them were removed due to correlation=1 with other features, thus the final number is **311**.
The visual results are shown in the ROC curve below (see Figure 7.7(d)).

The performance of the classifier for each comparison are shown in Table 7.6.

Table 7.6: Performance obtained by merging different data types, using the most significant feature sets.

Merge data Comparison	**Precision**	**Recall**	**Accuracy**	**f_1 score**
Sensitive Long vs Sensitive Short	0.60 ± 0.39	0.48 ± 0.32	0.83 ± 0.11	0.83 ± 0.20
Sensitive Long vs Resistant	0.86 ± 0.09	0.95 ± 0.08	0.84 ± 0.08	0.91 ± 0.10
Sensitive Short vs Resistant	0.80 ± 0.16	0.65 ± 0.14	0.82 ± 0.09	0.83 ± 0.10
Resistant vs All Sensitive	0.68 ± 0.18	0.74 ± 0.11	0.80 ± 0.10	0.82 ± 0.09

Considerations From the results shown in the Table 7.4, which refer to the best performances obtained with each set of features, it is possible to make some considerations:

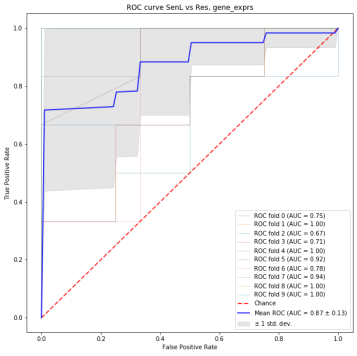
- the results obtained by combining the best features of different data types are generally good, considering the innovation of having integrated very different features (i.e., CNA regions and methylated genes) between them;

- exactly as for the performances analyzed before, the comparison with better performances turns out to be that between the Resistant and Sensitive Long classes: at this point it can be stated that all the subset of features considered are significant in the distinction of the two mentioned classes;

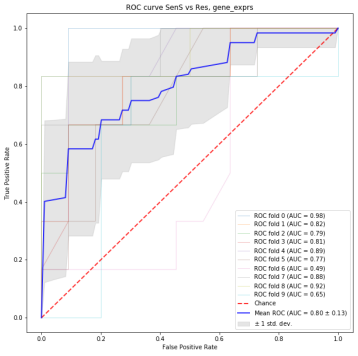
- the overall of the performances concerning the comparison between the Resistant and Sensitive classes is quite good: these results demonstrate that the features selection procedure used here, together with a careful choice of the classification algorithm and its parameters (hyperparameter tuning), is a winning strategy in order to discriminate among patients affected by a high-chromosomal instable cancer-type.

ROC curves

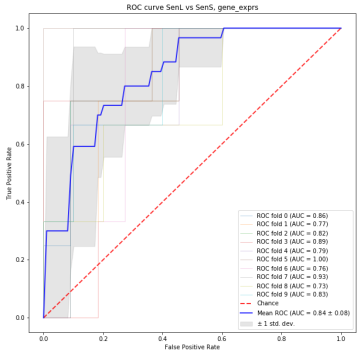
In a Receiver Operating Characteristic (ROC) curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. A test with perfect discrimination (no overlap in the two distributions) has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test.
Considering the above, the ROC curves describing the best overall accuracy of the various tests are in Figure 7.4(a), 7.5(a), 7.6(a) and 7.7(a), all referring to the binary comparison between the Resistant and Sensitive Long classes.
The curves that describe the worst performances are those related to the binary comparison between the two Resistant and Sensitive Short classes, while the comparison between the class of resistant patients and the class inclusive of all therapy-sensitive patients is generally good.



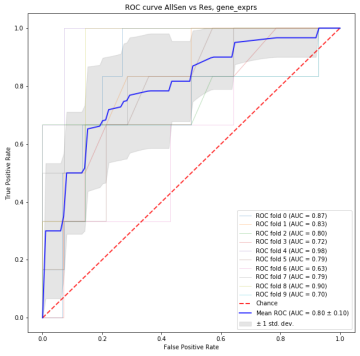
(a) Resistant vs Sensitive Long



(b) Resistant vs Sensitive Short

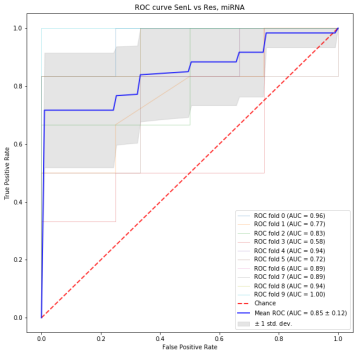


(c) Sensitive Long vs Sensitive Short

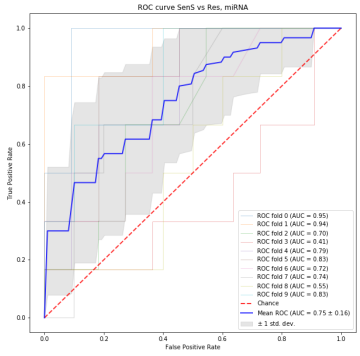


(d) Resistant vs All Sensitive

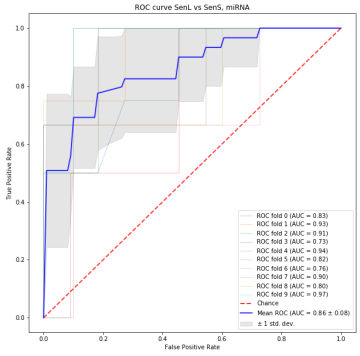
Figure 7.4: ROC curves of gene expression classifier, obtained by each binary comparison.



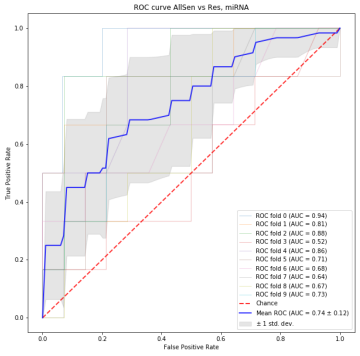
(a) Resistant vs Sensitive Long



(b) Resistant vs Sensitive Short

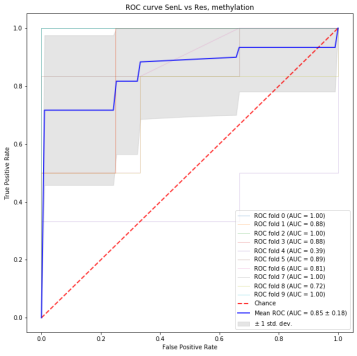


(c) Sensitive Long vs Sensitive Short

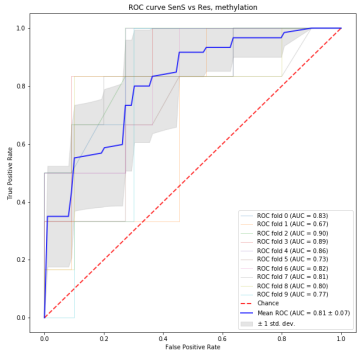


(d) Resistant vs All Sensitive

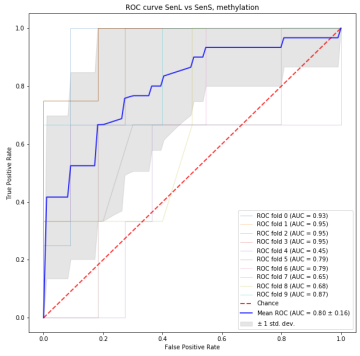
Figure 7.5: ROC curves of miRNA expression classifier, obtained by each binary comparison.



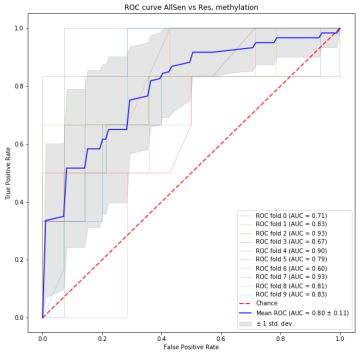
(a) Resistant vs Sensitive Long



(b) Resistant vs Sensitive Short

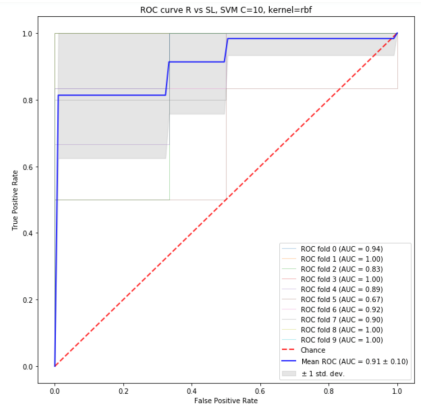


(c) Sensitive Long vs Sensitive Short

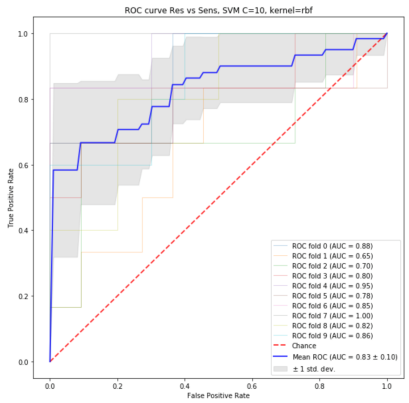


(d) Resistant vs All Sensitive

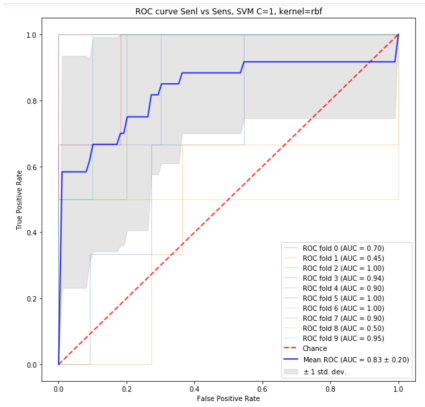
Figure 7.6: ROC curves of DNA methylation classifier, obtained by each binary comparison.



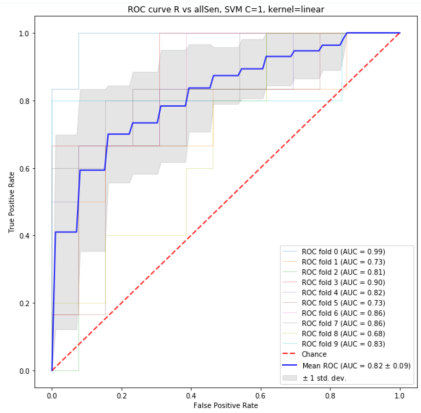
(a) Resistant vs Sensitive Long



(b) Resistant vs Sensitive Short



(c) Sensitive Long vs Sensitive Short



(d) Resistant vs Sensitive

Figure 7.7: ROC curves for each comparison by merging the best features obtained from different data types (gene expression data, miRNA expression data, methylation data, CNA data).

7.2 Biological Results

At this point, having various sets of considerable relevant features for each data-type, we have focused attention on the 137 CNA amplification regions relevant to the distinction of patients resistant to therapy compared to those who are not.
It was decided to analyze these regions in more detail due to the hypothesis that their intensity of alteration may be a good predictor of diagnosis, given their nature of early events: this means that it is not necessary to administer the therapy to understand the response to it because a particular expression of these regions could be indicative of a chemoresistance guessed *intrinsic*; in particular, the amplification regions have been chosen because, unlike those of deletion, these do not imply a sub-expression of genes such as to consider them silenced.
The identification of genomic regions distinctive of chemoresistance would represent great progress from the point of view of both the early diagnosis and the computational costs necessary for the acquisition and manipulation of data.
The use of an integrative approach based on the study of many different data-types showed good classification results among the classes, so it was decided to try to investigate whether, starting from the CNA features, it would be possible to bring the research back to a more targeted level.
Thus, a total of 183 protein-coding genes were extracted from the 137 CNA regions previously identified. Through functional annotations provided from DAVID, 33 genes considered relevant in the ovarian cancer study were selected. These genes and their brief descriptions are reported in Table 7.7.
Given the involvement of these 33 genes in the biology, aetiology and drug-resistance associated with ovarian cancer, a further selection was executed, including only the genes related to specific HGS-OC and its characteristics, resulting in 24 genes. A further selection has been done including only the genes that had relevance in terms of their contribution to the onset of chemoresistance.
The genes that proved to be relevant according to this selection are only *6* (in bold in Table 7.7).
At this point, to evaluate the existing interactions between genes and their involvement in the activation of any pathway relevant to the development and progression of HGS-OC and its relative chemoresistance, various levels of enrichment analysis have been performed, leading to the final result of *Notch signaling pathway* identification.

7.2.1 Enrichment Analysis

The first level of enrichment analysis was performed on the 24 genes involved in the development of HGS-OC and ovarian tumors, as epithelial

Table 7.7: Brief description of the 33 genes related to ovarian cancer in the CNA amplification regions suitable for distinguishing Resistant and Sensitive classes.

Official gene symbol	Gene name	Description
ADAM12	ADAM metallopeptidase domain 12	selectively expressed in ovarian tumor vasculature; its expression is associated with poor survival in patients with HGS-OC
BCCIP	BRCA2 and CDKN1A interacting protein	its germ line mutations are unlikely to be a major contributor to familiar ovarian cancer risk
BID	BH3 interacting domain death agonist	roles as predictive biomarker of chemotherapy resistance
BLM	Bloom syndrome RecQ like helicase	role in drug-resistance in leukemia
BMPR1B	bone morphogenetic protein receptor type 1B	role in EOC and influence on prognosis of OC patients
CA9	carbonic anhydrase 9	role in resistance to chemotherapy
CCSER1	coiled-coil serine rich protein 1	contribution to the chromosomal instability of cancer
CERS4	ceramide synthase 4	higher level in cancerour cell lines
CNTFR	ciliary neurotrophic factor receptor	may underlie treatment resistance, potential therapeutic target
CTBP2	C-terminal binding protein 2	ovarian cancer oncogene
CUZD1	CUB and zona pellucida like domains 1	promising biomarker for OC diagnosis
DLL1	delta like canonical Notch ligand 1	its overexpression may increase sensitivity of cells to chemotherapeutic agents
DPYSL4	dihydropyrimidinase like 4	associated with various systemic cancers
ELAVL1	ELAV like RNA binding protein 1	it mediates resistance to carboplatin in ovarian cancer cells
ERMP1	endoplasmic reticulum metallopeptidase	required for the organization of somatic cells and oocytes in the ovary
FANCG	Fanconi anemia complementation group G	interacts directly with BRCA2
FBN3	fibrillin 3	evidence for association with Polycystic ovary syndrome
FGFR2	fibroblast growth factor receptor 2	chemoresistance in neuroblastoma
FURIN	furin, paired basic amino acid cleaving enzyme	predictor of the disease outcome
GHSR	growth hormone secretagogue receptor	expressed in EOC in vivo and in vitro
HINT2	histidine triad nucleotide binding protein	downespression associated with low response to therapy in EOC
HtrA1	HtrA serine peptidase 1	its loss in OC may contribute to in vivo chemoresistance
IQGAP1	IQ motif containing GTPase activating protein 1	contribution to metastasis of OC
PARK2	parkin RBR E3 ubiquitin protein ligase	possible tumor suppressor gene defects in its expression may be involved in progression of OC
RECK	reversion inducing cysteine rich protein with kazal motifs	target-gene for anticancer effects of icariin
RNASET2	ribonuclease T2	its downregulation contributes to drug-resistance in OC
STOML2	stomatin like 2	overexpressed in EOC
TNFSF10	tumor necrosis factor superfamily member 10	good response to cisplatin
TPM2	tropomyosin 2 (beta)	overexpression in ovarian cancer compared to controls
UNC45A	unc-45 myosin chaperone A	expressed in OC proliferation and metastasis
UNC93A	unc-93 homolog A	located in a region of the genome frequently associated with OC
URI1	URI1, prefoldin like chaperone	role in development of chemotherapeutic resistance
VCP	valosin containing protein	highly sensitive serum tumor marker in several human cancers

OC, anticipating it. The result obtained through this first analysis was the identification of the Notch Signaling Pathway, considered relevant for its influence on drug-resistance.
This result highlighted the relevance of two genes present in our set that are activators of the pathway itself: the Delta-like 1 lingad DLL1 and the CTBP2 oncogene, as can be seen in Figure 7.8 from DAVID.

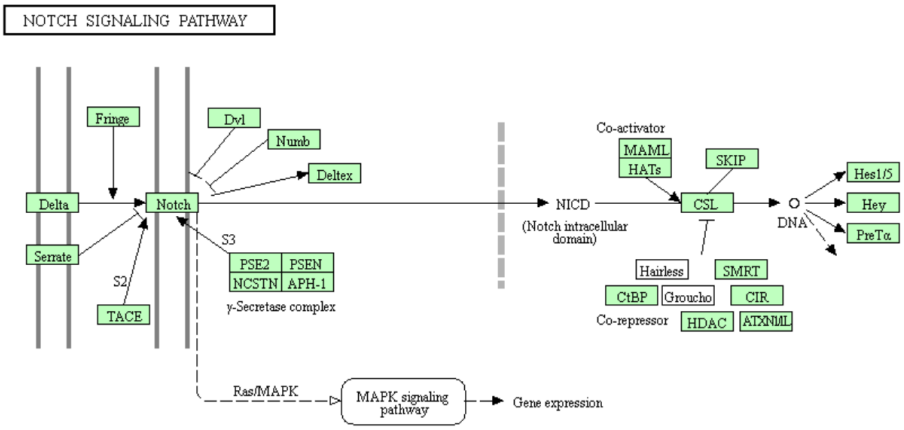


Figure 7.8: Notch Signaling Pathway from DAVID Functional Annotation Tool.

Notch Signaling Pathway

Notch Signaling pathway is one of the most important signaling pathway in drug-resistance tumor cells; its down-regulation could induce drug sensitivity, leading to increased inhibition of cancer cell growth, invasion and metastasis. It is a conserved ligand-receptor pathway which has critical roles in many aspects that affects the development and function of many organs, [39].
To date, four Notch receptors have been identified in mammals, including humans, such as Notch-1-4; the canonical ligands are designated as either Delta-like or Serrate-like ligands, known as Jagged-1 and Jagged-2.
Canonical Notch signals are transduced by a process called regulated intramembrane proteolysis. Notch receptors are normally maintained in a resting, proteolytically resistant conformation on the cell surface, but ligand binding initiates a proteolytic cascade that releases the intracellular portion of the receptor (ICN) from the membrane. The critical, regulated cleavage step is effected using ADAM metalloproteases and occurs at a site called S2 immediately external to the plasma membrane. This truncated receptor, dubbed NEXT (for Notch extracellular truncation), remains membrane tethered until it is processed at site S3 and additional sites by gamma secretase, a multiprotein enzyme complex.
After gamma secretase cleavage, ICN ultimately enters the nucleus, where it assembles a transcriptional activation complex that contains a DNA-binding transcription factor called CSL, and a transcriptional coactivator of the Mastermind family. This complex then engages additional coactivator proteins such as p300 to recruit the basal transcription machinery and turn on the expression of downstream target genes.
The whole Notch signaling is summarized in Figure 7.9. Notch signals influence a wide spectrum of cell fate decisions, both during development

Figure 7.9: Schematic of Notch signaling.

and in the adult organism. However, dysregulated signaling has also been implicated in a number of different human diseases ranging from neurodegeneration to cancer.
Recently, Notch pathway has been reported to be involved in drug resistance and many studies have demonstrated that Notch regulates the formation of cancer stem cells (CSCs) and contributes to the acquisition of the epithelialmesenchymal transition (EMT) phenotype, which are critically associated with drug resistance, [10].
Experimental evidence also revealed that Notch was involved in anti-cancer drug resistance, indicating that targeting Notch could be a novel therapeutic approach for the treatment of cancer by overcoming drug resistance of cancer cells, which may lead to the elimination of cancer stem cells or epithelialmesenchymal transition type cells which are typically drug-resistant, and are believed to be the root cause of tumor recurrence.
Moreover, this pathway seems to play a role in anti-taxol and platinum-based resistance, which are anti-cancer chemotherapy drugs used for the treatment of ovarian tumors, in particular of high-grade serous ovarian adenocarcinoma.
Also, evidences suggest that microRNAs (miRNAs) play important roles in the regulation of drug resistance. It is well known that the miRNAs elicit their regulatory effects in post-transcriptional regulation of genes by binding to the 3 untranslated region (3UTR) of target messenger RNA (mRNA). Some miRNAs are thought to have oncogenic activity while others have tumor suppressor activity:

- oncogenic miRNAs are up-regulated in cancer and contribute to its pathology through various mechanisms such as targeting tumor suppressor genes;
- in contrast, other miRNAs are considered to have tumor suppressor activity and are down-regulated in cancer.

One miRNA, namely miR-34, has been found to participate in Notch pathways regulation and has been reported to be involved drug resistance. The **miR-34** family is composed of three processed miRNAs: miR-34a is encoded by its own transcript, whereas miR-34b and miR-34c share a common primary transcript. The expression of miR-34a has been found to be lower or undetectable in many cancer types, suggesting that miR-34a could function as a tumor suppressor gene; in fact, many involvements of this miRNA in Notch regulation have been reported, such as:

- transfection of miR-34a to glioma cells down-regulated the protein expression of Notch-1, Notch-2, and CDK6;

- human gastric cancer cells with miR-34 restoration reduced the expression of target gene Notch;
- Notch-1 and Notch-2 are downstream genes of miR-34 in pancreatic cancer cells because restoration of miR-34 expression in the pancreatic cancer cells down-regulated the expression of Notch-1 and Notch-2;
- pancreatic cancer stem cells are enriched with tumor-initiating cells or cancer stem cells (CSCs) with high levels of Notch-1/2 and loss of miR-34, suggesting that miR-34 may be involved in pancreatic cancer stem cell self-renewal mediated by Notch signaling;
- miR-34a is down-regulated in drug-resistant prostate cancer cells, and ectopic overexpression of miR-34a resulted in growth inhibition and attenuated chemoresistance to the anti-cancer drug camptothecin;
- miR-34a was down-regulated in doxorubicin and verapamil resistance MCF-7 breast cancer cells.

Many reports clearly suggest the role of miR-34 in drug resistance, which is in part mediated through the regulation of Notch signaling.
Furthermore, the alteration of miR-200 family was also found in drug-resistant cells. The miR-200 family has five members: miR-200a, miR-200b, miR-200c, miR-141, and miR-429; many studies have shown that the miR-200 family regulates EMT which is associated with drug resistance:

- the expression of miR-200b was significantly down-regulated in docetaxel-resistant non-small cell lung cancer (NSCLC) cells;
- miR-200 expression regulates epithelialmesenchymal transition (EMT) in bladder cancer cells and reverses resistance to EGFR therapy;
- miR-200c restored microtubule-binding chemotherapeutic agents in breast and ovarian cancer cells;
- miR-200a, miR-200b, and miR-200c were down-regulated in gemcitabine-resistant pancreatic cancer cells, which show the acquisition of EMT phenotype;
- re-expression of miR-200 family resulted in the down-regulation of ZEB1, slug, E-cadherin, and vimentin and increased cell sensitivity to gemcitabine.

In addition, it has been found that Notch-1 could be one of miR-200b targets because overexpression of miR-200 family significantly inhibited

Notch-1 expression in gemcitabine-resistant pancreatic cancer cells and prostate cancer cells, suggesting that re-expression of miR-200 could increase drug sensitivity, which indeed could be mediated through the regulation of Notch signaling pathway.
It is important to underline the link of Notch with epithelial-mesenchymal transition, because it is associated with drug resistance and cancer cell metastasis; the processes of EMT has been shown to be important on conferring drug resistance characteristics to cancer cells against conventional therapeutics including taxol, vincristine, and oxaliplatin and Notch signaling pathway has been reported to be involved with the acquisition of EMT in drug-resistant cancer cells: the activation of Notch signaling is mechanistically linked with chemoresistance phenotype, which is consistent with the acquisition of EMT phenotype, and further suggesting that the inactivation of Notch signaling by novel strategies could be a potential targeted therapeutic approach for overcoming chemoresistance toward the prevention of tumor progression and/or treatment of human cancer for which current conventional therapeutic strategies are highly disappointing.

Then, by the mean of miRTarBase, which provides miRNA-target gene interactions experimentally known, and TargetScan, which provides high confidence prediction in silico, the relationships between the selected 24 genes with miRNAs have been investigated, leading to a network of connections shown in Figure 7.10.

As can be seen in the graph above, some of the selected genes have

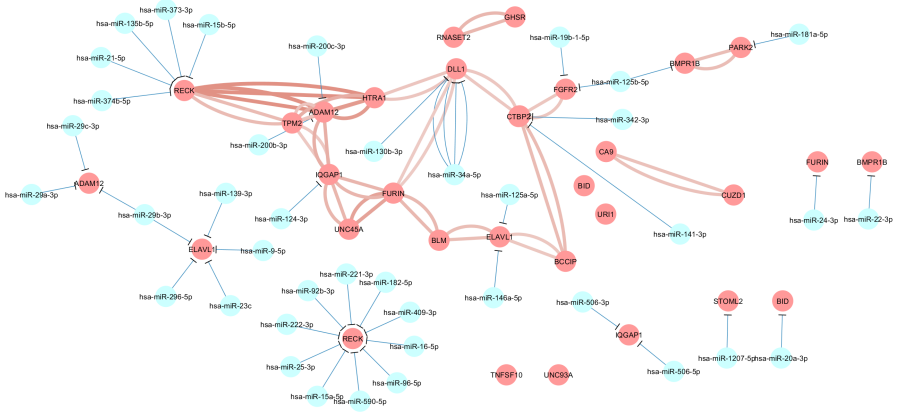


Figure 7.10: Connection network between relevant genes for progression of HGS-OC and miRNAs, obtained by the mean of Cytoscape tool.

shown to have experimental relationships with miRNAs involved in the Notch Signaling pathway, such as mir-34a, mir-200b and mir-200c.

To confirm the hypothesis that the genes involved in the drug-response,

selected within the CNA regions, were identifying a pattern that could lead back to the Notch Signaling Pathway, a second enrichment analysis was performed, using only the 6 protein-coding genes related to drug response plus the two activators DLL1 and CTBP2.

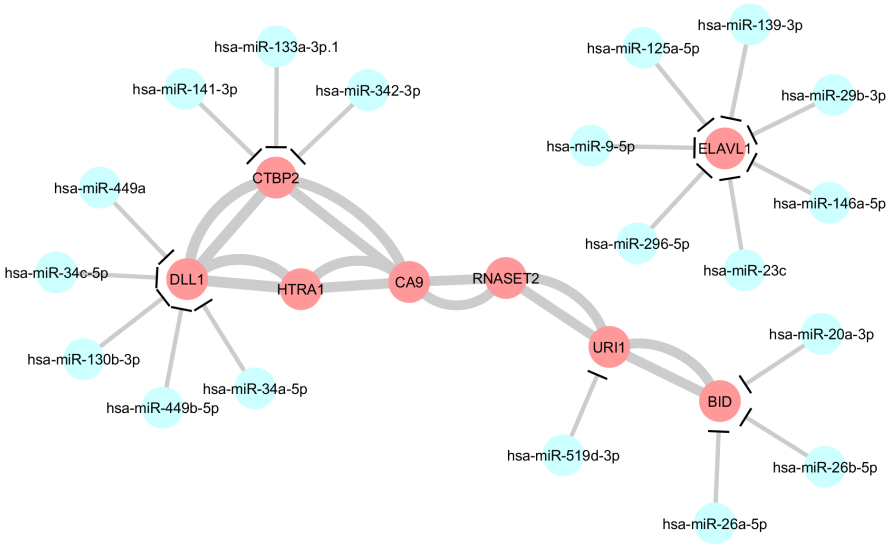


Figure 7.11: Connection network between relevant genes for drug-response in HGS-OC and miRNAs, obtained by the mean of Cytoscape tool.

This second analysis has again produced the identification of the Notch Signaling Pathway and the network of connection between those genes and miRNAs connected to them is shown in Figure 7.11. Also in this case the miRNA-target gene relationships were considered, which showed regulation by mir-34a and mir-34c, previously attributed to drug-resistance regulation mediated by the Notch signaling pathway.

Later, further analysis was performed to study in more detail the behavior of these 8 genes and their possible relevance within the found pathway.

To have a first visual impression of the difference in expression of the 8 genes considered to be involved in drug-resistance mediated by Notch signaling patwhay, the expression values of those genes were compared in Resistant and Sensitive classes, as shown in Figure 7.12. As can be seen from the graph in Figure 7.12, the expression of these genes is not significantly different in the classes that indicate a distinct drug-response to therapy.

This can be attributed to the fact that the data analyzed correspond only to the early stage, in which no administration of the therapy has taken place, which would induce a variation in the gene expression from which the development of chemoresistance would follow.

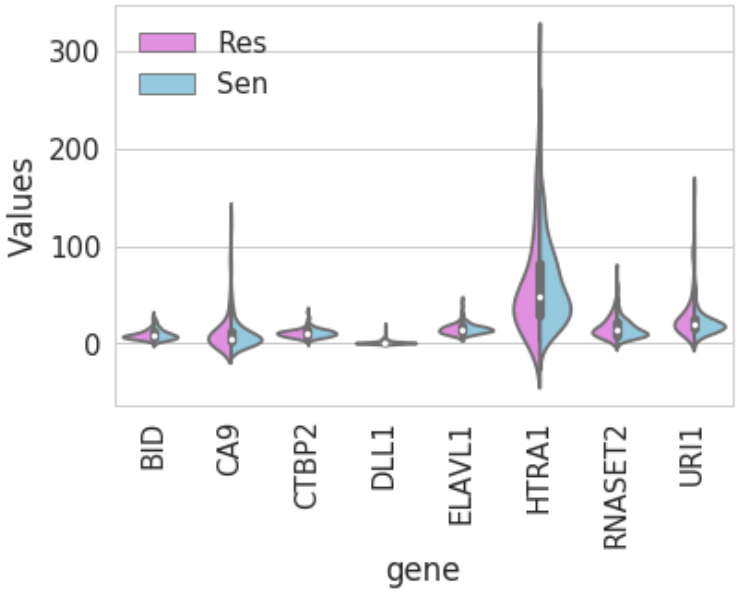


Figure 7.12: Gene expression comparison of 8 genes considered to be involved in drug-resistance mediated by Notch signaling pathwhay in Resistant and Sensitive classes.

For this reason, to verify the relevance of these genes in the induction of chemoresistance, the CNA regions associated with them were analyzed, to verify whether in those regions the intensity of alteration was already distinctive of the classes at the early stage.
Verifying a significant alteration in those CNA regions would validate the hypothesis that:

- the identified genes are drivers of chemoresistance at early stage;
- the CNA regions are predictors of drug-responsiveness at early stage, giving the possibility to do early diagnosis;

if these hypotheses were verified, there would be the possibility of targeting the identified genes and evaluating the development of other therapeutic options.

DLL1 Delta Like Canonical Notch Ligand 1 is a protein-coding gene, a human homolog of the Notch Delta ligand and it is a member of the delta/serrate/jagged family. The DLL1-induced Notch signaling is mediated through an intercellular communication that regulates cell lineage, cell specification, cell patterning and morphogenesis through effects on differentiation and proliferation. Also, its overexpression may increase the sensitivity of cells to chemotherapeutic agents.

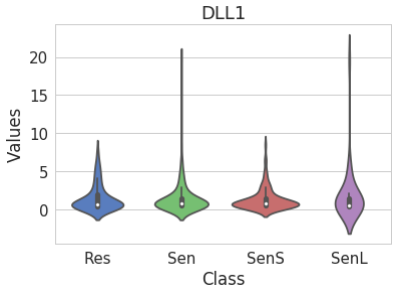


Figure 7.13: DLL1 gene expression across all the classes.

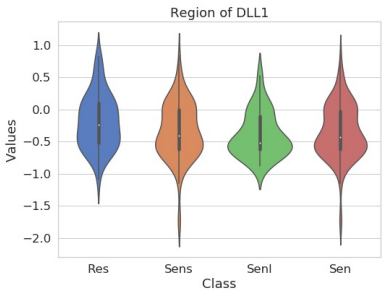


Figure 7.14: CNA region alteration associated to DLL1 across all the classes.

CTBP2 C-Terminal Binding Protein 2 is a protein-coding gene that produces alternative transcripts encoding two distinct proteins. One protein is a transcriptional repressor, while the other isoform is a major component of specialized synapses known as synaptic ribbons. Both proteins contain a NAD+ binding domain similar to NAD+-dependent 2-hydroxyacid dehydrogenases. A portion of the 3' untranslated region was used to map this gene to chromosome 21q21.3; however, it was noted that similar loci elsewhere in the genome are likely. Several transcript variants encoding two different isoforms have been found for this gene. The expression of this transcriptional co-repressor is elevated in human ovarian tumors. Downregulation of CtBP2 expression in ovarian cancer cell lines using short-hairpin RNA strategy suppressed the growth rate and migration of the resultant cancer cells. It has been proposed that CtBP2 is an ovarian cancer oncogene that regulates gene expression program by modulating histone deacetylase (HDAC) activity. CtBP2 expression may be a surrogate indicator of cellular sensitivity to HDAC inhibitors.

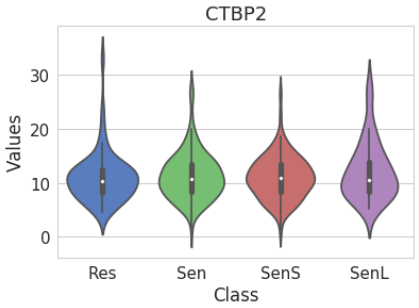


Figure 7.15: CTBP2 gene expression across all the classes.

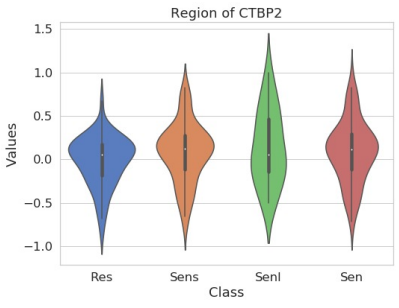


Figure 7.16: CNA region alteration associated to CTBP2 across all the classes.

Now, we will show the analyzes related to the genes considered most

important in the development of chemoresistance, in order of relevance.

CA9 Carbonic anhydrase IX is an enzyme that in humans is encoded by the CA9 gene, included in the large family of zinc metallo-enzymes that catalyze the reversible hydration of carbon dioxide. They participate in a variety of biological processes, including respiration, calcification, acid-base balance, bone resorption, and the formation of aqueous humor, cerebro-spinal fluid, saliva, and gastric acid. They show extensive diversity in tissue distribution and in their subcellular localization. CA9 is a transmembrane protein and is one of only two tumor-associated carbonic anhydrase isoenzymes known. Its expression in combination with that of vascular endothelial growth factor (VEGF) has been associated with decreased overall survival and response to therapy, because cancer growth, spread and chemotherapy resistance are promoted by hypoxic microenvironment which affects several genes through stabilization of hypoxia-inducibile factor 1-α, that triggers the promoters of CA9 and VEGF, [40]. The combined high expression CA9 and VEGF phenotype, described as high hypoxia profile group, showed significant positive correlation with resistance to chemotherapy and poor overall survival, suggesting that this phenotype may have a useful role in stratifying ovarian cancer for prognostic and therapeutic purposes.

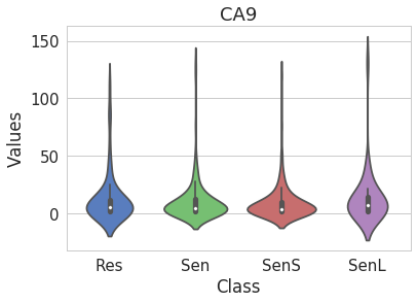


Figure 7.17: CA9 gene expression across all the classes.

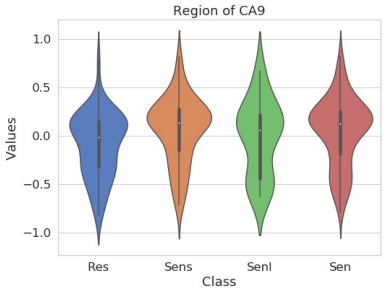


Figure 7.18: CNA region alteration associated to CA9 across all the classes.

ELAVL1 ELAV Like RNA Binding Protein 1, also known as Human antigen R (HuR), is a protein-coding gene. The protein encoded by this gene is a member of the ELAVL family of RNA-binding proteins that contain several RNA recognition motifs, and selectively bind AU-rich elements (AREs) found in the 3' untranslated regions of mRNAs. AREs signal degradation of mRNAs as a means to regulate gene expression, thus by binding AREs, the ELAVL family of proteins play a role in stabilizing ARE-containing mRNAs. This gene has been implicated in a variety of biological processes and has been linked to a number of diseases, including cancer. It is highly expressed in many cancers, and could be potentially

useful in cancer diagnosis, prognosis, and therapy.
HuR is regulated through NEDDylation post-translational modification: inhibition of this process should sensitise resistant tumour cells to carboplatin. It is showed in [41] that treatment of a tumour cell line with MLN4924, a NEDDylation inhibitor, overcame the resistance to carboplatin, this leading to the conclusion that inhibition of NEDDylation may be a useful strategy to resensitise tumour cells in patients that have acquired carboplatin resistance.

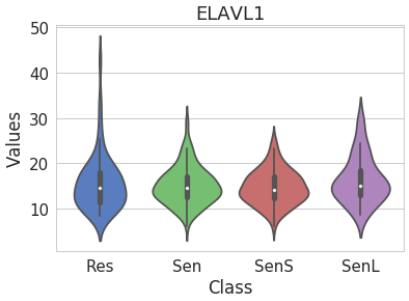


Figure 7.19: ELAVL1 gene expression across all the classes.

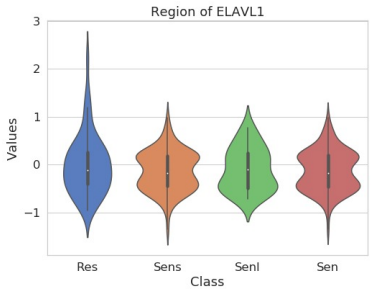


Figure 7.20: CNA region alteration associated to ELAVL1 across all the classes.

HtrA1 HtrA Serine Peptidase 1 is a protein-coding gene that encodes a member of the trypsin family of serine proteases. The expression of HtrA1, which is frequently downregulated in ovarian cancer, influences tumor response to chemotherapy by modulating chemotherapy-induced cytotoxicity. Downregulation of HtrA1 attenuated cisplatin- and paclitaxel-induced cytotoxicity, while forced expression of HtrA1 enhanced cisplatin- and paclitaxel-induced cytotoxicity. HtrA1 expression was upregulated by both cisplatin and paclitaxel treatment. This upregulation resulted in limited autoproteolysis and activation of HtrA1. Active HtrA1 induces cell death in a serine proteasedependent manner. The potential role of HtrA1 as a predictive factor of clinical response to chemotherapy was assessed in ovarian cancer patients receiving cisplatin-based regimens. Patients with ovarian tumors expressing higher levels of HtrA1 showed a higher response rate compared with those with lower levels of HtrA1 expression. These findings uncover what is believed to be a novel pathway by which serine protease HtrA1 mediates paclitaxel- and cisplatin-induced cytotoxicity and suggest that loss of HtrA1 in ovarian and gastric cancers may contribute to in vivo chemoresistance, [42].

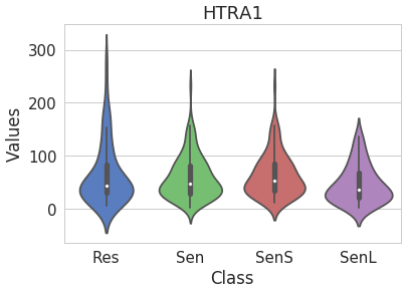


Figure 7.21: HtrA1 gene expression across all the classes.

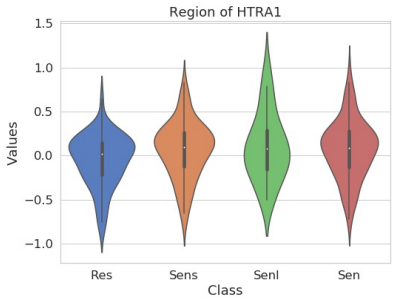


Figure 7.22: CNA region alteration associated to HtrA1 across all the classes.

RNASET2 Ribonuclease T2 is a protein-coding gene, member of the Rh/T2/S-glycoprotein class of extracellular ribonucleases. It is a tumor suppressor genes whose expression is significantly downregulated in drug-resistant ovarian cancer cells/tissues. The drug resistance-related functions of RNASET2 have been analyzed in [43], resulting that RNASET2 was co-expressed, co-localized, physically interacted and shared protein domains and pathways directly/indirectly with a number of proteins. Specifically, direct genetic interactions were established between RNASET2 and phosphatase and tensin homolog (PTEN): PTEN is a well-known TSG associated with cancer development through the ERK1/2 signaling and PI3K/Akt/mTOR pathways. It interacts with tumor suppressor genes (such as p53 and BRCA1) that contribute to the development of drug resistance in several types of cancer. In ovarian cancer, PTEN contributes to multidrug resistance through cell cycle regulation, apoptosis and the PI3K/Akt pathway.

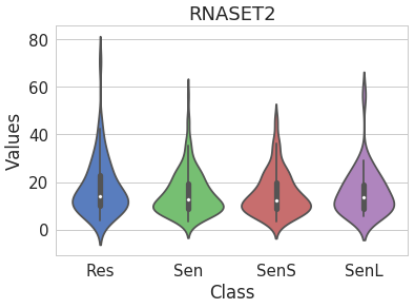


Figure 7.23: RNASET2 gene expression across all the classes.

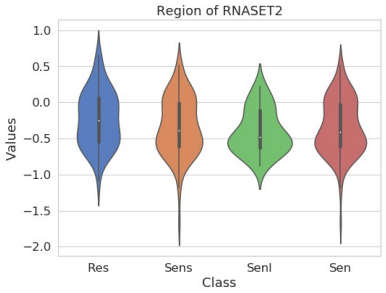


Figure 7.24: CNA region alteration associated to RNASET2 across all the classes.

BID BH3 Interacting Domain Death Agonist is a protein-coding gene that encodes a death agonist that heterodimerizes with either agonist BAX or antagonist BCL2. The encoded protein is a member of the BCL-2

family of cell death regulators. It is a mediator of mitochondrial damage induced by caspase-8 (CASP8); CASP8 cleaves this encoded protein, and the COOH-terminal part translocates to mitochondria where it triggers cytochrome c release. Multiple alternatively spliced transcript variants have been found, but the full-length nature of some variants has not been defined. BID preferentially activates BCL-2 antagonist or killer (BAK), affecting chemotherapy response, as stated in [44].

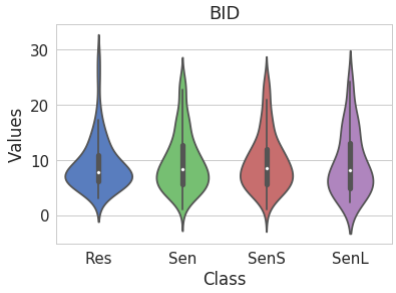


Figure 7.25: BID gene expression across all the classes.

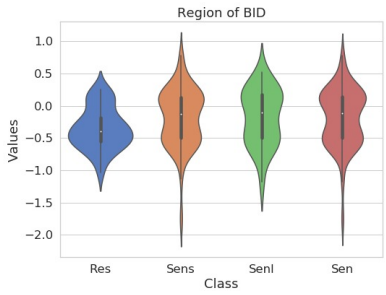


Figure 7.26: CNA region alteration associated to BID across all the classes.

URI1 URI1 Prefoldin Like Chaperone involved in gene transcription regulation. This gene may play a role in multiple malignancies including ovarian cancer and hepatocellular carcinoma. URI regulates tumorigenicity and chemotherapeutic resistance of multiple myeloma by modulating interleukin (IL)-6 transcription, [45].

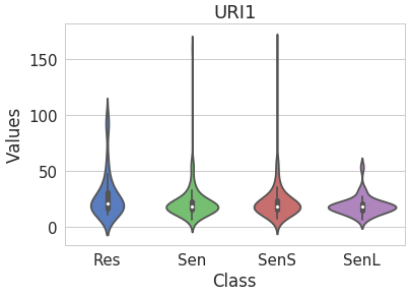


Figure 7.27: URI1 gene expression across all the classes.

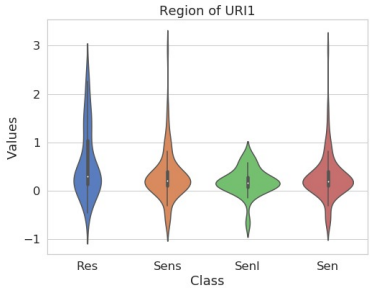


Figure 7.28: CNA region alteration associated to URI1 across all the classes.

Except for URI1, no significant differences in gene expression between the classes are shown; on the other hand, the alteration of the CNA regions seems to be more significant, in particular for BID, CA9, CTBP2, HtrA1 and DLL1, thus suggesting that the regions associated to those genes could actually be used as efficient predictors of drug-resistance at early stage.

To validate what has been said by commenting on the graphs that compare the gene expression with the intensity of alterations of the CNA region associated to the gene in analysis, a statistical verification was conducted by the mean of Kolmogorov-Smirnov[21] test, whose results are reported in Tables 7.8 and 7.9:

Table 7.8: p_values obtained comparing the distribution of the CNA values of classes for the different genes using the K-S test.

Comparison	BID	CA9	CTBP2	HtrA1	URI1	DLL1	RNASET2	ELAVL1
Resistant vs Sensitive	0.003	0.01	0.01	0.01	0.05	0.081	0.22	0.48

Table 7.9: p_values obtained comparing the distribution in the classes of the expression of the different genes using the K-S test.

Comparison	URI1	CTBP2	ELAVL1	RNASET2	BID	HtrA1	DLL1	CA9
Resistant vs Sensitive	0.03	0.18	0.22	0.38	0.49	0.49	0.69	0.88

The results shown in the tables above statistically validate the statement that for many of the genes, the CNA distributions are different between the classes while those of expression are practically the same.

In summary, from what has been said so far it can be stated that:

- the selected genes are actually involved in drug response, in particular their contribution modulates the chemoresistance;
- at early stage, these genes do not show a significantly different gene expression, but the CNA regions associated with them show a different intensity of alteration between classes;

this confirms that copy number alterations can be considered predictors of drug-response that can be used for early diagnosis and to assess whether the possible targeting of the selected genes would lead to the development of new therapeutic options.

[21] In statistics, the KolmogorovSmirnov test (KS test or KS test) is a nonparametric test of the equality of continuous (or discontinuous), one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample KS test), or to compare two samples (two-sample KS test).

Chapter 8

Conclusions

The aim of this study was the identification of a molecular signature that was distinctive of the onset of chemoresistance in patients with High-Grade Serous Ovarian Adenocarcinoma, the most lethal of ovarian tumors. Responsiveness to the platinum-based drug is a very relevant factor linked to the outcome of therapy, as most patients, who are initially sensitive, will develop a progressive resistance, becoming incurable due to the lack of therapeutic options. Moreover, in some cases, patients show an intrinsic drug-resistance, which makes them resistant from the beginning of the therapy: for this reason, it is necessary to identify those factors that discriminate therapy-resistant patients from therapy-sensitive ones, to be able to intervene promptly by subjecting them to a different type of treatment.
To do so, different genomic data has been downloaded from TCGA repository, like gene expression data, miRNA expression data and methylation data. Each of them has been analyzed to identify a set of relevant features which could be good predictors of the patient drug-responsiveness; in particular, for each data-type, different feature have been extracted to obtain the best set of predictors. that distinguish as the patients belonging to three classes, known as *resistant*, *sensitive long* and *sensitive short*.
Build said predictor was not easy. A first attempt to predict the response to therapy was done using each patient's time to relapse; this analysis was carried out using each data-type independently. The results obtained were quite poor since the maximum concordance index value reached was 0.61, lower a the threshold (c.i. = 0.7) that defines the model as a good predictor.
Given the unsatisfactory results of survival analysis, the focus shifted towards the modelling of a classifier capable of discriminating classes through binary comparisons, using as features the protein-coding genes, miRNAs and methylation of genes previously selected. In this case, the discrimination between classes has shown satisfactory results, both for the distinction of resistant patients to sensitive ones, and as regards the distinction be-

tween subclasses: the best results were those obtained in the separations of resistant versus sensitive long classes, which exhibit the greatest difference in relapse timing, reaching an accuracy of *0.84* for *gene expression data*, *0.75* for *miRNA expression data* and *0.77* for *DNA methylation data*.
To further improve these already promising results, all the different types of features have been merged, thus including also the CNA genomic regions, which also come from TCGA repository: this has led to an overall improvement in the performances, producing a good discrimination of resistant class compared to the sensitive one.
In order to develop a model of the rising of the resistance, we focused on copy number alterations, these being considered early events, and therefore possible predictors that would favor an early diagnosis. In particular, attention has focused on the 137 amplification regions indicative for the distinction between resistant and sensitive; within these genomic areas, corresponding to around 1% of the genome, 183 protein-coding genes were identified, subsequently selected based on two different criteria:

- the involvement of these genes in biology, aetiology and therapy-response of ovarian cancer, which allowed us to identify a set of *24* protein-coding genes;
- the influence of these genes with respect to drug responsiveness, which allowed us to identify a set of *6* protein-coding genes.

We conducted on both sets of genes, an enrichment analysis using DAVID tool, and that analysis allowed us to identify a relevant pathway implicated in drug-resistance: the *Notch Signaling Pathway*.
It was decided to perform a second enrichment analysis using only the genes involved in drug responsiveness plus the two pathway activators included in the first set of genes: DLL1 and CTBP2. Given that many evidences have suggested that miRNAs can play important roles in drug resistance regulation, the relationships between the selected genes with miRNAs have been investigated by the mean of miRTarBase and TargetScan.
Through this analysis, two miRNAs of the miR-34 family have been identified which, according to literature, are attributed to drug-resistance regulation mediated by the Notch signaling pathway: *miR-34a* - that is encoded by its own transcript and its lower expression in many cancer types suggests it could function as a tumor suppressor gene - and *miR-34c*. Further investigations have been performed to compare the different expressions of the selected genes with respect to the classes and compare them with the alterations present in the CNA regions corresponding to the location of those genes. Thanks to this procedure, it was possible to verify that:

- the genes selected according to their contribution to drug responsiveness are indicators of the development of chemoresistance;

- having used pre-treatment data, in which no alteration due to drug administration could be detected in the gene expression, it is possible to assume that:

    1. the CNA regions relative to the genes under examination, which show a difference in alteration between the classes also at early stage, can be considered predictors for early diagnosis;
    2. drug-resistance is triggered by drug intake, which induces expression changes in the selected genes.

This findings allow to formulating an interesting theory about the development of chemoresistance, linked to the activation of the pathway as a result of the regulation of the genes that we have identified, and that occurs for patients as a function of the number and form of the gene replications, well evidenced by the copy number alterations. Simplifying, our model implies that elevate alterations of copy number in the restricted areas of the genome that we have identified, already present at the diagnosis, lead to a greater probability that the Notch Signaling pathway is activated and that this leads to the rapid development of chemoresistance. This result validates on a biological level what has been hypothesized in the literature, and can lead to interesting therapeutic developments, aimed at downregulating the Notch signaling pathway, which would seem to induce drug-sensitivity.
Therefore, if our model will be confirmed by more detailed biological analysis, interesting advancements related to the knowledge of the disease are envisaged, and perhaps also a new direction for the development of targeted therapies, customized on the basis of a relatively simple examination to be performed (specific probes may be created to investigate the portions of the genome indicated by our model). On the other hand, the genes identified by us based on CNA regions are differentially activated when there is chemoresistance, as analyzes by various works; in particular:

- the Notch signaling pathway regulates the formation of cancer stem cells and contributes to the acquisition of the epithelial-mesenchymal transition phenotype, which are critically associated with drug-resistance; also, this pathway is found to be involved itself in anti-cancer drug-resistance, indicating that targeting and downregulating it could induce drug-sensitivy, [39];

- *DLL1* and *CTBP2*, two protein-coding genes identified by our procedure, are primary activators of the Notch signaling pathway, in particular DLL1 is a Delta-like 1 canonical ligand for Notch1 receptor and CTBP2 is an ovarian cancer oncogene that regulates gene expression;

- *RNASET2* is a strongly down-regulated protein-coding gene in drug-resistant cells and tissues; it is also associated with PTEN, a tumor suppressor gene that contributes to the development of drug resistance in many types of cancer; especially in ovarian cancer, PTEN contributes to a multi-drug resistance through its cell regulation cycle, apoptosis and the PI3K / Akt pathway, [43];

- *HtrA1* influences the cytotoxicity of paclitaxel and cisplatin drugs, specifically: its downregulation attenuates the cytotoxicity of the drug, while its upregulation promotes it; downregulation of this gene in ovarian cancer may represent an additional mechanism to chemoresistance, [42];

- *ELAVL1*, also known as Human antigen R (HuR), is implicated both in the sensitive response to the therapy and in the resistant one; this gene, regulated through NEDDylation post-translational modification, interacts with a series of genes that show an over-expression in clinical cases of ovarian cancer; this over-expression seems to be predictive of chemoresistance, [41];

- *CA9*, whise over-expression, combined with the over-expression of VEGF, results involved in chemoresistance and in the poor overall survival, [40];

- *URI1* is an oncogene whose upregulation contributes to drug resistance in ovarian cancer, [46], [45];

- BID, in interaction with BAK and BAX oncogenes modulates drug resistance, [47], [44], [48].

Our model can be verified with experimental data, that will be provided by *Instituto di Ricerche Farmacologiche Mario Negri*; however, we expect that very few individuals will match the model's applicaility. Then, further clinical studies could be performed, as well as to specify a low-cost method to collect the data necessary for the model to make a diagnosis prediction.Such other analyzes go beyond the specific context of this thesis, aimed at finding a molecular signature allowing the separation between sensitive and chemo-resistant patients; this task has been successfully completed.

Bibliography

[1] Sara Sansone. "Computational methods for chemoresistance prediction in High Grade Serous Ovarian Adenocarcinoma". MA thesis. Politecnico di Milano, 2019.

[2] Ourania Koukoura et al. "Epigenomics of Ovarian Cancer and Its Chemoprevention". In: *Epigenetics of Cancer Prevention*. Elsevier, 2019, pp. 333–358.

[3] Christine Stewart, Christine Ralyea, and Suzy Lockwood. "Ovarian Cancer: An Integrated Review". In: *Seminars in Oncology Nursing*. Elsevier. 2019.

[4] Sara Moufarrij et al. "Epigenetic therapy for ovarian cancer: promise and progress". In: *Clinical epigenetics* 11.1 (2019), p. 7.

[5] Cancer Genome Atlas Research Network et al. "Integrated genomic analyses of ovarian carcinoma". In: *Nature* 474.7353 (2011), p. 609.

[6] Robert J Kurman and Ie-Ming Shih. "The Origin and pathogenesis of epithelial ovarian cancer-a proposed unifying theory". In: *The American journal of surgical pathology* 34.3 (2010), p. 433.

[7] M.D. Jennifer Gordetsky. *Testis and epididymis, Paratesticular tumors: Brenner tumor*. 9 June 2017. URL: `http://www.pathologyoutlines.com/topic/testisbrenner.html`.

[8] Anais Malpica et al. "Grading ovarian serous carcinoma using a two-tier system". In: *The American journal of surgical pathology* 28.4 (2004), pp. 496–504.

[9] Russell Vang, Ie-Ming Shih, and Robert J Kurman. "Ovarian low-grade and high-grade serous carcinoma: pathogenesis, clinicopathologic and molecular biologic features, and diagnostic problems". In: *Advances in anatomic pathology* 16.5 (2009), p. 267.

[10] Sergio Marchini et al. "Resistance to platinum-based chemotherapy is associated with epithelial to mesenchymal transition in epithelial ovarian cancer". In: *European journal of cancer* 49.2 (2013), pp. 520–530.

[11] Terence A. Brown. *Genomes, 2nd Edition*. Wiley-Liss, June 15, 2002.

[12] National Human Genome Research Institute. *A Brief Guide to Genomics*. Available on line, `https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics`.

[13] Daniel C Koboldt et al. "The next-generation sequencing revolution and its impact on genomics". In: *Cell* 155.1 (2013), pp. 27–38.

[14] Yimei Cai et al. "A brief review on the mechanisms of miRNA regulation". In: *Genomics, proteomics & bioinformatics* 7.4 (2009), pp. 147–154.

[15] Fazli Wahid et al. "MicroRNAs: synthesis, mechanism, function, and recent clinical trials". In: *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* 1803.11 (2010), pp. 1231–1243.

[16] Cathérine Dupont, D Randall Armant, and Carol A Brenner. "Epigenetics: definition, mechanisms and clinical perspective". In: *Seminars in reproductive medicine*. Vol. 27. 05. © Thieme Medical Publishers. 2009, pp. 351–357.

[17] National Human Genome Research Institute. *Copy Number Variation*. Available on line, `https://www.genome.gov/genetics-glossary/Copy-Number-Variation`.

[18] Joe R Delaney and Dwayne G Stupack. "Genomic Copy Number Alterations in Serous Ovarian Cancer". In: *Ovarian Cancer: From Pathogenesis to Treatment* (2018), p. 111.

[19] Marco Masseroli et al. "GenoMetric Query Language: a novel approach to large-scale genomic data management". In: *Bioinformatics* 31.12 (2015), pp. 1881–1888.

[20] *Pandas Python library: Python Data Analysis Library*. Available on line, `https://pandas.pydata.org/`.

[21] *Scikit-learn Python library: Machine learning in Python*. Available on line, `https://scikit-learn.org/stable/index.html`.

[22] *Matplotlib Python library*. Available on line, `https://matplotlib.org/`.

[23] *Lifelines Python library*. Available on line, `https://lifelines.readthedocs.io/en/latest/`.

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

[25] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. "The cancer genome". In: *Nature* 458.7239 (2009), p. 719.

[26] Marco Pegoraro. "Rimozione della componente di variabilita'causata da fattori non biologici: un caso studio su dati di tumore all'ovaio". In: (2013).

[27] Brad T Sherman, Richard A Lempicki, et al. "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources". In: *Nature protocols* 4.1 (2009), pp. 44–57.

[28] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists". In: *Nucleic acids research* 37.1 (2008), pp. 1–13.

[29] Trevor Hastie et al. "The elements of statistical learning: data mining, inference and prediction". In: *The Mathematical Intelligencer* 27.2 (2005), pp. 83–85.

[30] Sebastian Raschka and Vahid Mirjalili. *Python machine learning*. Packt Publishing Ltd, 2017.

[31] David G Kleinbaum et al. *Logistic regression*. Springer, 2002.

[32] Christopher M Bishop. *Pattern recognition and machine learning*. Springer Science+ Business Media, 2006.

[33] Shujun Huang et al. "Applications of support vector machine (SVM) learning in cancer genomics". In: *Cancer Genomics-Proteomics* 15.1 (2018), pp. 41–51.

[34] TG Clark et al. "Survival analysis part I: basic concepts and first analyses". In: *British journal of cancer* 89.2 (2003), p. 232.

[35] Edward L Kaplan and Paul Meier. "Nonparametric estimation from incomplete observations". In: *Journal of the American statistical association* 53.282 (1958), pp. 457–481.

[36] Stuart J Pocock, Tim C Clayton, and Douglas G Altman. "Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls". In: *The Lancet* 359.9318 (2002), pp. 1686–1689.

[37] Sheng-Da Hsu et al. "miRTarBase: a database curates experimentally validated microRNA–target interactions". In: *Nucleic acids research* 39.suppl_1 (2010), pp. D163–D169.

[38] Li Li et al. "Computational approaches for microRNA studies: a review". In: *Mammalian Genome* 21.1-2 (2010), pp. 1–12.

[39] Zhiwei Wang et al. "Targeting Notch signaling pathway to overcome drug resistance for cancer therapy". In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1806.2 (2010), pp. 258–267.

[40] Emma Williams et al. "Co-expression of VEGF and CA9 in ovarian high-grade serous carcinoma and relationship to survival". In: *Virchows Archiv* 461.1 (2012), pp. 33–39.

[41] Graziele Fonseca de Sousa et al. “Chemogenomic study of carboplatin in Saccharomyces cerevisiae: inhibition of the NEDDylation process overcomes cellular resistance mediated by HuR and cullin proteins”. In: *PloS one* 10.12 (2015), e0145377.

[42] Jeremy Chien et al. “Serine protease HtrA1 modulates chemotherapy-induced cytotoxicity”. In: *The Journal of clinical investigation* 116.7 (2006), pp. 1994–2004.

[43] Fuqiang Yin et al. “Downregulation of tumor suppressor gene ribonuclease T2 and gametogenetin binding protein 2 is associated with drug resistance in ovarian cancer”. In: *Oncology reports* 32.1 (2014), pp. 362–372.

[44] Kristopher A Sarosiek et al. “BID preferentially activates BAK while BIM preferentially activates BAX, affecting chemotherapy response”. In: *Molecular cell* 51.6 (2013), pp. 751–765.

[45] JL Fan et al. “URI regulates tumorigenicity and chemotherapeutic resistance of multiple myeloma by modulating IL-6 transcription”. In: *Cell death & disease* 5.3 (2014), e1126.

[46] Xia Liu et al. “Oncogenes associated with drug resistance in ovarian cancer”. In: *Journal of cancer research and clinical oncology* 141.3 (2015), pp. 381–395.

[47] Janine T Erler et al. “Hypoxia-mediated down-regulation of Bid and Bax in tumors occurs via hypoxia-inducible factor 1-dependent and-independent mechanisms and contributes to drug resistance”. In: *Molecular and cellular biology* 24.7 (2004), pp. 2875–2889.

[48] N Goncharenko-Khaider et al. “The inhibition of Bid expression by Akt leads to resistance to TRAIL-induced apoptosis in ovarian cancer cells”. In: *Oncogene* 29.40 (2010), p. 5523.