UNIVERSITA' DEGLI STUDI DI PAVIA

FACOLTA' DI INGEGNERIA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

DOTTORATO DI RICERCA IN BIOINGEGNERIA E BIOINFORMATICA
XXIX CICLO - 2016

A NOVEL MOLECULAR-BASED APPROACH FOR THE PROGNOSIS IN ENDOMETRIAL CANCER

PhD Thesis by
FEDERICA TORRICELLI

Advisor:
Prof. Riccardo Bellazzi

PhD Program Chair:
Prof. Riccardo Bellazzi

A Mattia ,
il risultato più bello
ottenuto durante questo
Dottorato…

Abstract (Italiano)

Il cancro dell'endometrio è il tumore ginecologico più comune nei paesi sviluppati ed è al mondo il sesto tumore più comune nelle donne.

In quasi l'80% dei casi, il tumore dell'endometrio è di tipo I (istotipo endometrioide), per lo più di basso grado, mentre il restante 20% dei tumori si definisce di tipo II (non endometrioide) e normalmente ha prognosi più aggressiva rispetto ai tumori di tipo I.

Attualmente, la definizione del grado del tumore di tipo I avviene tramite analisi anatomo-patologica della morfologia tumorale, ma spesso la valutazione del tipo e del grado del tumore non è sufficiente a predire la prognosi. L'identificazione di variabili genetiche e/o molecolari associate al fenotipo e ai profili di aggressività di questi tumori è necessaria al fine di sviluppare nuovi e più efficaci strumenti prognostici.

Lo scopo di questo studio è lo sviluppo di un metodo di predizione basato sull'analisi mutazionale, che sia utile per stratificare i tumori dell'endometrio nelle categorie "buona prognosi" e "cattiva prognosi".

L'analisi è stata effettuata su 89 campioni di tumore dell'endometrio a diversi gradi. Ha previsto il sequenziamento high-throughput, mediante sequenziatore MiSeq Illumina, utilizzando il kit "Trusight Tumor 26" il quale consente l'analisi delle mutazioni somatiche descritte su 174 ampliconi all'interno di 26 geni selezionati tra quelli maggiormente associati a tumori solidi in letteratura.

Un'analisi di clustering non supervisionato è stata effettuata per suddividere i campioni in due gruppi, sulla base del profilo mutazionale. L'analisi di sopravvivenza effettuata ha mostrato un comportamento tendenzialmente differente dei due gruppi (Morti 18% nel cluster 1 e 11% nel cluster 2, Recidive 23 % nel cluster 1 e 11% nel cluster 2).

L'analisi ha inoltre mostrato l'associazione della clusterizzazione su base molecolare con il grado del tumore, in particolare distinguendo perfettamente i campioni di grado G1 e collocandoli tutti nel cluster 1 a buona prognosi.

L'analisi di frequenza dei geni mutati e del numero di mutazioni di ciascuno nei campioni dei due clusters ha evidenziato come i geni APC,

CTNNB1, KRAS, PIK3CA, PTEN siano differentemente mutati nei due gruppi

Le mutazioni sui geni CTNNB1, KRAS, PIK3CA, SMAD4 e TP53 sono risultate inoltre associate al diverso grado dei tumori.

Lo studio propone quindi un nuovo approccio su base molecolare, che prevedendo il sequenziamento di un piccolo pannello di geni, potrebbe affiancare l'analisi istologica del tumore dell'endometrio nella predizione dell'outcome tumorale, soprattutto dei casi a prognosi più incerta.

Abstract (English)

Endometrial cancer is the most common gynecologic cancer in developed countries and the world's sixth most common cancer in women. In almost 80% of cases, endometrial cancer is of type I (endometrioid histological type), mostly of low grade, while the remaining 20% of the tumors is called type II (non-endometrioid) and normally has poor prognosis.

Currently, grading classification of type I endometrial cancer takes place through pathologic analysis of tumor morphology, but often the assessment of the type and tumor grading is not sufficient to predict prognosis. In order to develop new and more effective prognostic tools the identification of genetic and/or molecular variables associated to the phenotype and profiles of aggressiveness of these tumors is necessary.

The purpose of this study is to develop a prediction method based on genes mutational status, useful to stratify the endometrial tumors in the categories "good prognosis" and "poor prognosis".

The analysis was performed on 89 samples of endometrial cancer of different degrees. High-throughput sequencing was performed on Illumina MiSeq sequencer using the "Trusight Tumor 26 kit", which allows the analysis of somatic mutations described to 174 amplicons within 26 genes selected among those most associated with solid tumors in literature. An unsupervised clustering analysis was carried out to divide samples into two groups, based on the mutational profile. Survival analysis showed a basically different behavior of the two groups (18% deaths in cluster 1 and 11% in cluster 2, 23% recurrence in cluster 1 and 11% in cluster 2).

The analysis also showed the association of the clustering with the grading of the tumors, in particular distinguishing perfectly grade G1 samples and placing them all in the “good prognosis” cluster 1.

The analysis of frequency of the mutated genes and their number of mutations in clusterized samples showed that APC, CTNNB1, KRAS, PIK3CA, PTEN mutation status was different in the two groups. Mutations on CTNNB1, KRAS, PIK3CA, SMAD4 and TP53 were also found associated with the different grading of endometrial tumors. The study therefore proposes a new approach based on molecular profiling, that trough the sequencing of a small panel of genes, could

support the histological analysis in endometrial cancer outcome prediction, especially in doubtful cases.

Contents

Abstract (Italiano) ..... **II**

Abstract (English) ..... **IV**

Contents ..... **VI**

1 Introduction to endometrial cancer **8**
1.1. Epidemiology 8
1.2. Risk factors 9
1.3. Diagnosis of endometrial cancer 10
1.4. Histological and molecular classification of endometrial cancer 11
1.4.1. The original classification: two histotypes of endometrial cancer 11
1.4.1.1. Histological grading of endometrial cancer 13
1.4.1.2. Surgical staging of endometrial cancer 13
1.4.1.3. Lax Kurman binary grading system 14
1.4.2. TCGA genomic classification of endometrial carcinoma 15
1.5. Treatment of endometrial cancer 17

2 Aim of the study **20**

3 Next generation approach of target resequencing **21**
3.1. Trusight Tumor 26 Kit Illumina 21
3.2. Miseq reporter Amplicon DS primary analysis 23
3.3. Variant Studio secondary analysis and Basespace variants annotation 25
3.3.1. Selection of genetic variants with effect on protein function 25
3.4. Results validation by Sanger sequencing 28

4 Study population **30**
4.1. Clinical and pathological description 30
4.2. Molecular description 30

5 Unsupervised identification of two population subgroups with different prognosis **34**
5.1. Unsupervised hierarchical clustering analysis 34
5.2. Overall survival and disease free survival analysis 37

6 Characterization of the generated prognostic clusters **41**
6.1. Statistical analysis of frequency of the clinical-pathological characteristics between the two clusters 41

6.2. Representation of the gene variants frequencies between the two clusters....43
6.3. Statistical analysis of frequency of genes mutations between the two clusters....45
6.4. Lasso and Elastic-Net Regularized Generalized Linear Model to investigate genes effect on patients survival....48
6.5. Data mining approaches used to define classification rules driving clusterization....50
6.5.1. Classification tree....50
6.5.2. CN2 analysis....52
6.5.3. Logistic regression and nomogram generation....54
6.6. Investigation of genes variants damaging effects in clusters....56
6.6.1. PaPI score calculation....56

7 Discussion....58
7.1. Proposal of an alternative molecular-based approach of prognosis prediction in Endometrial cancer....58
7.2. Related works and molecular models comparison....62
7.3. Further validations of the model....63

8 Conclusions....64

9 Methods and supplementary informations....65
9.1. Next generation sequencing....65
9.1.1. DNA extraction and quality evaluation....65
9.1.2. Trusight Tumor kit Illumina....65
9.1.3. MiSeq run....67
9.2. NGS data analysis....67
9.3. Sanger Sequencing....67
9.4. Hierarchical clustering analysis and statistical analysis....68
9.5. Data mining methods....68
9.5.1. Decision tree....68
9.5.2. CN2 analysis....69
9.5.3. Nomogram....69
9.6. PaPI Score analysis....69

References....70

Chapter 1

Introduction to endometrial cancer

1.1. Epidemiology

Endometrial cancer (EC) is the most common gynecological cancer in industrialized countries.

Each year, EC develops in about 142000 women worldwide, and an estimated 42000 women die from this cancer [1].

In Italy, EC is the fourth most common cancer among women, accounting for 5% of all malignant neoplasms with about 8200 estimated new cases each year. [2].

The typical age-incidence curve for EC shows that most cases are diagnosed after the menopause, with the highest incidence around the seventh decade of life. [1]

The early appearance of symptoms explains why about 70% of these patients have early-stage disease at presentation resulting in a favorable prognosis, with 5-year overall survival rate of 77%. However, for those women with more advanced or recurrent disease, response rates to conventional chemotherapy are low and clinical outcomes are extremely poor [3].

Commonly, EC is classified into two types based upon clinical-pathologic features [4]: Type 1 EC are endometrioid cancer, associated with hyperestrogeneism and typically preceded by endometrial hyperplasia. They are often diagnosed at an early stage, and have a good prognosis. Type 2 EC includes non-endometrioid cancers such as serous, clear cell, mixed cell, undifferentiated and carcinosarcoma. These neoplasms are not estrogen correlate, often occur in the presence of an atrophic endometrium and have a poor prognosis.

1.2. Risk factors

Several risk factors are reported in association to type I EC patients.

Age is demonstrated to be a risk factor and a predictor of poor prognosis in EC and over 90% of the cases are diagnosed after the age of 50 years [5].

The main risk factor for type I EC is a prolonged exposure to estrogen during women reproductive life [6], reason why early age at menarche and late age at menopause were associated with increased risk of endometrioid EC.

Additional risk factor for type I EC are indeed Chronic Anovulation and Polycystic Ovary Syndrome (PCOS) because characterized by elevated serum estrogen levels [7, 8].

Lower parity and/or null parity were found to increase the risk of developing endometrioid EC up to four fold, while any additional birth among parous women (after the birth of the second child) was demonstrated to decrease the risk of developing the disease by 10 % for every new child [9]. This is because pregnancy is characterized by a protective progesterone increase and a estrogen decrease, which results in suppressed endometrial mitotic activity.

Type I EC risk was also positively correlated with high-fat diet or high energy intake A high fat diet is considered a risk factor for type I EC both directly and indirectly because it promotes estrogen metabolism and leads to development of obesity, respectively [10].

Obesity is a well-identified risk factor for type I EC both in pre-menopausal and post-menopausal women [11]. The risk of type I EC is higher when obesity is associated with infertility or amenorrhea, conditions in which estrogen levels are already high, and obesity further increases estrogen exposure and insulin resistance. On contrary, type II EC is not correlated with obesity.

Diabetic women have 2 to 3 fold increased risk of developing type I EC, compared to not diabetic women and the risk is 6 fold higher when diabetes is associated with obesity [12].

Hypertension is an additional risk factor in EC, hypertensive women have 3-fold increased risk of developing type I EC compared to healthy women[7].

In patients with estrogen receptor positive breast cancer treated with Tamoxifen, a selective estrogen receptor modulator, EC risk increases with the duration of the therapy. Indeed, Tamoxifen stimulates

endometrial proliferation increasing the thickness of the endometrium [13].

Unexpectedly, smoking is considered a protective factor against EC. Smokers have lower endogenous estrogen levels compared to non-smokers and smoking also reduced the effect of estrogen by reducing the age of menopause and consequently the total number of menstrual cycles [14].

Type II carcinomas onset was demonstrated to be ordinarily independent from hormonal levels, age and obesity while it was demonstrated to be strongly associated with a first-degree family history of cancer [15].

About 5% of EC cases have a family history of the disease. In women less than 50 years old, about 9% of EC is due to mutations in mismatch repair genes (MSH1, MSH2, MSH6), that result in Lynch II syndrome, a multiple diseases condition in which Hereditary Non-Polyposis Colorectal Cancer (HNPCC) is associated with other cancers of the gastrointestinal tract or reproductive system [16].

Cowden Syndrome is also associated with an increased life time risk (5-10%) of EC due to the autosomal dominant germinal PTEN mutation [17].

1.3. Diagnosis of endometrial cancer

Uterine bleeding in a postmenopausal woman is the main presenting sign of endometrial carcinoma. Pre or perimenopausal women with acyclical bleeding should also undergo through diagnostic evaluation, particularly if they have risk factors for EC.

Targeted screening examinations for early detection, with endovaginal sonography followed by endometrial biopsy, may be reasonable for women at high risk (e.g., those with Lynch syndrome)[18].

Women with abnormal bleeding of the types described should undergo the following studies:

- Gynecological examination to localize the source of bleeding and determine its physical extent; transvaginal ultrasonography for evaluation of the endometrium and adnexa. In postmenopausal patients with uterine bleeding, an endometrial thickness exceeding 5 mm is considered suspect. In contrast, no reliable cut off has been

reported in pre- or perimenopausal women, as well as in postmenopausal women taking hormone replacement therapy or tamoxifen.

- Hysteroscopy and fractionated uterine curettage.

1.4. Histological and molecular classification of endometrial cancer

1.4.1. The original classification: two histotypes of endometrial cancer

In 1983 Bokhman proposed a dualic classification of endometrial tumorigenesis based on both etiology and clinical behaviour. To date EC are still broadly classified as type 1 and type 2 though this model is not entirely accurate and pathologic assignment of some uterine cancers remain controversial [4].

Type 1 tumors represent the 70-80% of sporadic cases of EC. Risk factors for this EC histotype include obesity, anovulation, nulliparity, and exogenous estrogen exposure. These lesions present endometrioid phenotype, arise in a background of hyperplasia and commonly express estrogen and progesterone receptors. Clinically, type 1 cancers are more often low-grade tumors with a favorable prognosis.

Type 2 ECs are less common, accounting for about 20% of total ECs. They are often of non endometrioid, high-grade tumors, usually papillary serous or clear cell. Clinically, type 2 cancers are characterized by an aggressive clinical course, and they have a propensity for early spread and poor prognosis [19].

The 5-year overall survival rate (OS) of Endometrioid cancer patients ranges from 75% to 86%, in contrast to 50% to 60% of non endometrioid EC patients.

Aside from their morphological and clinical features, type 1 and type 2 ECs are further distinguished by genetic alterations (Table 1) [20].

	Type I	Type II
Clinical, endocrinological, and morphological components (Bokhman classification)		
Distribution	60–70%	30–40%
Background endometrium	Hyperplasia	Atrophy
Oestrogen associated	Yes	No
Associated obesity, hyperlipidaemia, and diabetes mellitus	Yes	No
Myometrial invasion	Superficial	Deep
Potential for lymphogenic metastatic spread	Low	High
Prognosis	Favourable	Unfavourable
Sensitivity to progestagens	High	Low
Outcome (5-year survival)	86%	59%
Clinicopathological and molecular correlates		
Prototypical histological type	Endometrioid	Serous
Oestrogen-receptor or progesterone-receptor expression	High	Low
Stage at diagnosis	Early (FIGO stage I–II)	Advanced (FIGO stage III–IV)
Common genetic alterations		
PTEN mutation	52–78%	1–11%
PIK3CA mutation	36–52%	24–42%
PIK3R1 mutation	21–43%	0–12%
KRAS mutation	15–43%	2–8%
ARID1A mutation	25–48%	6–11%
CTNNB1 mutation	23–24%	0–3%
TP53 mutation	9–12%	60–91%
PPP2R1A mutation	5–7%	15–43%
HER2 amplification	0%	27–44%
Microsatellite instability	28–40%	0–2%

FIGO=International Federation of Gynaecology and Obstetrics

Table 1: Dualistic classification of epithelial EC, including clinical, pathological and common molecular genetic correlates [20]

Type 1 endometrioid ECs present high percentage of mutations in PTEN, KRAS, ARID1A and CTNNB1, as well as defects in DNA mismatch repair.

Type 2 non-endometrioid ECs frequently show aneuploidy, p53 mutations and HER2 amplification. PIK3CA mutations are frequent in both EC histotypes.

1.4.1.1. Histological grading of endometrial cancer

The EC cell differentiation is a key marker to predict outcome and to evaluate the most appropriate therapy.

Endometrioid type 1 ECs can be classified in histological grade G1, G2 and G3. G1 tumors have the best prognosis, cancer cells are still well differentiated and similar to normal cells and the solid component is lower than 5%. In G2 tumors the solid component varies from 6% to 50%, and the cancerous cells have major differences from their normal counterpart. The G3 tumors have the worst prognosis, with more than 50% of solid component.

In type 2 endometrial cancer grading classification is not necessary, since all tumors are considered grade tumors, with high percentage of undifferentiated cells [21].

1.4.1.2. Surgical staging of endometrial cancer

Surgical staging of endometrial cancer was first proposed in 1988, and the staging system was updated in 2009 [22]. In revised FIGO (International Federation of Gynaecology and Obstetrics) staging system, tumors classification is based on tissues invasion (Table 2) Tumors confined to the endometrium as well as those invading the inner half of the myometrium are designated as stage IA tumors and tumors invading the outer half of the myometrium are designated as stage IB tumors.

In 2009 FIGO staging system, tumors with endocervical glandular invasion are considered stage I tumors, while tumors with cervical stromal invasion are defined as stage II tumors.

Stage III comprised three groups: IIIA, IIIB, and IIIC. Stage IIIA tumors invade the serosa or adnexa, stage IIIB tumors invade the vagina or parametrium and stage IIIC is divided into stage IIIC1, which is

characterized by pelvic lymph node involvement, and stage IIIC2, which is characterized by paraaortic lymph node involvement.

Stage IV tumors present the worst prognosis: Stage IVA extend into adjacent bladder or bowel, and stage IVB tumors have distant metastases (e.g, to the liver or lungs) [23].

Stage	Description
IA	Tumor confined to uterus, <50% myometrial invasion
IB	Tumor confined to uterus, ≥50% myometrial invasion
II	Cervical stromal invasion
IIIA	Tumor invasion into serosa or adnexa
IIIB	Vaginal or parametrial involvement
IIIC1	Pelvic node involvement
IIIC2	Paraaortic node involvement
IVA	Tumor invasion into bladder or bowel mucosa
IVB	Distant metastases (including abdominal metastases) or inguinal lymph node involvemen

Table 2: 2009 FIGO Staing System for Endometrial Cancer

1.4.1.3. Lax Kurman binary grading system

In 2000 Lax and Kurman described a novel, binary architectural grading system that uses low-magnification assessment of amount of solid growth, pattern of invasion and presence of necrosis to divide endometrioid type I carcinomas into low and high grade tumors [24].

Based on these criteria a tumor can be classified as high grade if at least two of the following three criteria are present:

- More than 50% solid growth, without distinction between squamous versus nonsquamous differentiation
- A diffusely infiltrative growth pattern characterized by irregularly distributed glands, masses, cords or nests of tumor cells infiltrating the myometrium
- Tumor cell necrosis

1.4.2. TCGA genomic classification of endometrial carcinoma

The Cancer Genome Atlas Research Network (TCGA) has reported in 2013 a comprehensive genomic and transcriptomic analysis of endometrial cancers based on next-generation sequencing technologies, analysis of DNA methylation, reverse-phase protein array, and microsatellite instability[25].

The study focused on common histological types that, were further categorized into four genomic classes (Table 3):

- Ultra-mutated tumors (POLE) characterized by very high mutation rates and hotspot mutations in the exonuclease domain of POLE (a subunit of DNA polymerase), few copy number aberrations, mutations in PTEN, PIK3R1, PIK3CA, FBXW7, and KRAS, and favorable outcome;
- A microsatellite-instable group of tumors (MSI hyper-mutated), characterized by MLH1 promoter methylation, high mutation rates, few copy-number aberrations, recurrent RPL22 frameshift deletions, and KRAS and PTEN mutations;
- Low Copy Number tumors (endometrioid), comprising microsatellite-stable grade 1 and 2 endometrioid cancers with low mutation rates, characterized by frequent CTNNB1 mutations;
- High Copy Number tumors (serous-like), characterized by extensive copy number aberrations and low mutation rates, recurrent TP53, FBXW7, and PPP2R1A mutations, infrequent PTEN and KRAS mutations, and poor outcome.

	***POLE* (ultramutated)**	**MSI (hypermutated)**	**Copy-number low (endometrioid)**	**Copy-number high (serous-like)**
Copy-number aberrations	Low	Low	Low	High
MSI/*MLH1* methylation	Mixed MSI high, low, stable	MSI high	MSI stable	MSI stable
Mutation rate	Very high (232×10^{-6} mutations/Mb)	High (18×10^{-6} mutations/Mb)	Low ($2{\cdot}9 \times 10^{-6}$ mutations/Mb)	Low ($2{\cdot}3 \times 10^{-6}$ mutations/Mb)
Genes commonly mutated (prevalence)	*POLE* (100%) *PTEN* (94%) *PIK3CA* (71%) *PIK3R1* (65%) *FBXW7* (82%) *ARID1A* (76%) *KRAS* (53%) *ARID5B* (47%)	*PTEN* (88%) *RPL22* (37%) *KRAS* (35%) *PIK3CA* (54%) *PIK3R1* (40%) *ARID1A* (37%)	*PTEN* (77%) *CTNNB1* (52%) *PIK3CA* (53%) *PIK3R1* (33%) *ARID1A* (42%)	*TP53* (92%) *PPP2R1A* (22%) *PIK3CA* (47%)
Histological type	Endometrioid	Endometrioid	Endometrioid	Serous, endometrioid, and mixed serous and endometrioid
Tumour grade	Mixed (grades 1–3)	Mixed (grades 1–3)	Grades 1 and 2	Grade 3
Progression-free survival	Good	Intermediate	Intermediate	Poor

Table 3: Characteristics of four genomic classes of endometrioid and serous carcinomas [25]

The TCGA study revealed that also a subset of tumors diagnosed as high-grade endometrioid carcinomas harbored copy number and mutational profiles more similar to those of serous carcinomas and in general no mutations (excluding POLE) were identified as unique to any

of the four genomic classes. In view of the substantial genetic and morphological heterogeneity in endometrial carcinomas, these data suggested that the current approach of histopathology-based classification requires a revision, which could take into account also the complicated molecular profiles of these tumors [20].

1.5. Treatment of endometrial cancer

The International Federation of Gynecology and Obstetrics recommends systematic surgical staging for most patients, consisting of hysterectomy with bilateral adnexal removal and systematic pelvic and para-aortic lymphadenectomy. It should be performed by laparoscopic approach.

In many cases, laparoscopy seems to be as safe and effective as an open abdominal procedure and superior with respect to postoperative morbidity and recovery.

The findings obtained through this basic initial treatment serve as the definitive guide to the potential use of further adjuvant therapy, depending on the stage of disease.

Patients with tumor stage IA and grade 1 or 2 are unlikely to have lymph node involvement, and their prognosis is usually very good. Thus, systematic lymphadenectomy is not indicated for such patients as it can't guarantee any survival advantage.

On the other hand, patients with advanced disease and negative outcome can benefit from surgical intervention in addition to various palliative therapies.

With regards to the surgical removal of endometrial carcinoma, some controversy surrounds the question whether additional pelvic and para-aortic lymphadenectomy could grant more diagnostic or therapeutic benefits.[18]. Taken together, histotype classification, tumor grading, tumor size and myometrial infiltration, may help in the decision to perform or not lymphadenectomy.

Adjuvant treatment could be delivered only when the final stage and grade are known.

Well known clinical-pathological prognostic factors for EC include: age, FIGO stage, depth of myometrial invasion, tumor differentiation grade, tumor type (endometrioid versus serous and clear cell) and lymphovascular space invasion (LVSI). According to these factors,

patients were stratified in risk groups (Table 4), to guide adjuvant therapy use [26].

Risk group	Description
Low	Stage I endometrioid, grade 1–2, <50% myometrial invasion, LVSI negative
Intermediate	Stage I endometrioid, grade 1–2, ≥50% myometrial invasion, LVSI negative
High-intermediate	Stage I endometrioid, grade 3, <50% myometrial invasion, regardless of LVSI status
	Stage I endometrioid, grade 1–2,LVSI unequivocally positive, regardless of depth of invasion
High	Stage I endometrioid, grade 3, ≥50% myometrial invasion, regardless of LVSI status
	Stage II
	Stage III endometrioid, no residual disease
	Non-endometrioid (serous or clear-cell or undifferentiated carcinoma, or carcinosarcoma)
Advanced	Stage III residual disease and stage IVA
Metastatic	Stage IVB

Table 4: New risk groups to guide adjuvant therapy use in EC [26].

Based on risk group summarized below, in Italy AIOM (Italian Association of Medical Oncology) guidelines suggest:

- IA G3, IB G1-G2 EC patients (intermediate risk) require brachytherapy particularly in case of age > 60 years.

- IB G3 EC patients (high risk) and type II require external beam radiotherapy with brachytherapy and chemotherapy should be considered.
- Stage II-III EC patients require external beam radiotherapy with brachytherapy and chemotherapy.
- Stage IV EC patients require chemotherapy and palliative radiotherapy should be considered.

Chapter 2

Aim of the study

Although more than one classification of EC have been proposed based on histological and molecular characterization, numerous EC cases, in particular those with intermediate phenotype and grading (e.g. endometrioid tumor G2) still have uncertain prognosis.

While low-grade endometrioid and serous carcinomas integrate well into Bokhman's model (being, respectively, prototypical type I and II tumors), many in the range of endometrioid EC fall outside a simple dichotomous classification. Between 10% and 19% of endometrioid carcinomas are high grade and have clinical, histopathological and molecular features that are either intermediate between those of types I and II. By contrast, not all serous carcinomas behave as prototypical type II cancers. For example, 2% of serous carcinomas arise in association with endometrial hyperplasia, and at least 20% lack deep myometrial invasion [20].

TGCA molecular classification, on the other hand, offers a complete molecular characterization of different EC groups but the use of a so large and complex mutational screening in clinical routine, for rapid prognostic prediction and treatment choice, results still too expensive in terms of time and costs and far from being realistic. Furthermore TGCA classification was only partially associated with prognosis, giving results that seem to be in contrast with literature data and would need further investigation.

This study intends to propose a novel molecular-based approach to predict prognosis in EC. The model, based on DNA sequencing of few genes, subdivides endometrial tumors in "good prognosis" and "bad prognosis", to be applied for the investigation of doubtful cases, and support Bokhman's model and histological grading when the canonical approach is not sufficient to predict tumor outcome.

Chapter 3

Next generation approach of target resequencing

3.1. Trusight Tumor 26 Kit Illumina

With the aim to identify a reduced panel of genes able to stratify cases with good or bad prognosis we use Trusight Tumor 26 kit Illumina to analyze our population of EC Formalin Fixed Paraffin Embedded (FFPE) samples.

TruSight Tumor offers an approach of amplicon-based sequencing of 26 oncogenes and tumor suppressor genes selected for their involvement in common solid tumors (Table 5). Through 174 amplicons Trusight Tumor 26 kit provides coverage of hot-spot coding regions which variation were described and cataloged in the COSMIC database in oncogenes, and coverage of all exons in tumor suppressor genes.

Using a paired-end sequencing approach Trusight tumor 26 achieves limits of detection below 5% variant allele frequency, with a minimum of 1000X coverage.

AKT1	EGFR	GNAS	NRAS	STK11
ALK	ERBB2	KIT	PDGFRA	TP53
APC	FBXW7	KRAS	PIK3CA	
BRAF	FGFR2	MAP2K1	PTEN	
CDH1	FOXL2	MET	SMAD4	
CTNNB1	GNAQ	MSH6	SRC	

Table 5: Trusight Tumor 26 gene panel

Due to the high level of resolution reached by the system the sequencing can detect DNA damage, specifically DNA deamination, caused by formalin fixation. Deamination events effectively result in a C/T single nucleotide change on a single strand of DNA, which appear as a G/A variant when sequenced. TruSight Tumor assay was designed ad hoc for analysis of DNA from FFPE tissues: both strands of the DNA template were treated with two different pools of primers named FPA and FPB and so targeted with highly specific amplicon designs. In this way the method permit to compare the two strands of the same template, cytosine deamination results in a nucleotide change in one strand of a DNA molecule, but does not alter the complementary nucleotide on the opposite strand. Sequencing each strand independently will yield base calls that differ between the 2 strands and will be excluded because considered false positive. A true DNA mutation results in a nucleotide change in both strands of a DNA molecule. Sequencing each strand independently will yield the same variant call for both strands, the mutation will pass filters and will be considered a true positive (Figure.1)

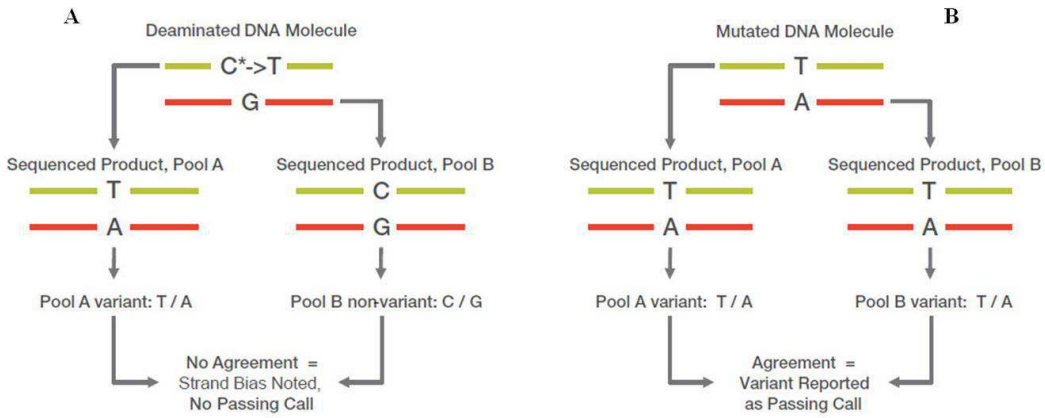

Figure 1: Detection and differentiation of DNA damage from mutation using Trusight Tumor chemistry

3.2. Miseq reporter Amplicon DS primary analysis

The Amplicon DS workflow is uniquely suited for detection of somatic mutations in formalin-fixed paraffin-embedded (FFPE) samples.

This workflow independently processes variants from the forward and reverse strands of the sample material, and then algorithmically reconciles the calls.

The Amplicon DS workflow consists in:

- **Demultiplexing:** data from pooled samples are separated based on short index sequences used to tag different libraries.
- **FASTQ file generation:** MiSeq Reporter generates intermediate analysis files in the FASTQ format. FASTQ files contain reads for each sample and their quality scores, excluding reads identified as inline controls and clusters that did not pass filter
- **Alignment:** Smith-Waterman algorithm aligns clusters from each sample against amplicon sequences specified in the manifest file. Each paired-end read is evaluated in terms of its alignment to the relevant probe sequences for that read. If the start of a read

matches a probe sequence with no more than 1 mismatch, the full length of the read is aligned against the amplicon target for that sequence. Alignments that include more than 3 indels are filtered from alignment results. Filtered alignments are written in alignment files as unaligned and are not used in variant calling.

Paired-End Evaluation: for paired-end runs, the top-scoring alignment for each read is considered.

Reads are flagged as an unresolved pair if either read did not align, the paired reads aligned to different chromosomes or if two alignments come from different amplicons or different targets of the manifests.

Finally, reads are sorted by sample and chromosome, and then by chromosome position. Results are written to one BAM file per sample.

- **Variant Calling:** SNPs and short indels are identified using the somatic variant caller tool developed by Illumina. The somatic variant caller identifies variants present at low frequency in the DNA sample and minimizes false positives. The somatic variant caller identifies SNPs in 3 steps:
  - Considers each position in the reference genome separately
  - Counts bases at the given position for aligned reads that overlap the position
  - Computes a variant score that measures the quality of the call.

    Variant scores are computed using a Poisson model that excludes variants with a quality score below Q20.

The model only calls variants for bases that are covered at 300x or greater for a single amplicon.

To exclude false positive due to DNA damage, first the variants for each pool (FPA and FPB) are called separately, then are compared and combined into a single output file.

If a variant meets the following criteria, the variant is marked as PASS in the variant file:

- Must be present in both pools
- Cumulatively have a depth of 1000 or an average depth of 500x per pool

- Have variant frequency of 3% or greater

In our study only PASS variants were considered for further analysis.

3.3. Variant Studio secondary analysis and Basespace variants annotation

MiSeq reporter Amplicon DS workflow finally produces a VCF v4.1 file that can be imported in Illumina VariantStudio desktop application for secondary analysis.

The Illumina VariantStudio desktop application provides commands to annotate variants, filter results using various filtering options, classify variant according to their biological impact, and export results to a report.

Variant studio aggregates information from multiple sources, capturing annotations at variant, gene and transcript level.

Through Variant Effect Predictor (VEP) Variant Studio consults databases such as NCBI Reference Sequence Database (RefSeq) and algorithms such as Polymorphism Phenotyping (PolyPhen)3 and SIFT. Information about known disease association can be obtained from the Catalogue of Somatic Mutations in Cancer (COSMIC), ClinVar, and Online Mendelian Inheritance in Man (OMIM), via the ClinVar database.

Resources such as dbSNP, the Ensembl 1,000 Genomes Project, and Exome Variant Server provide information about the occurrence and frequencies of variants within a population.

3.3.1. Selection of genetic variants with effect on protein function

In our study we used a sequencing kit designed to investigate mutations in a set of oncogenes and oncosoppressors genes.

In order to outline a molecular profile with prognostic potential we decided to consider only somatic genetic variants supposed to have an effect on protein coding. With this approach we wanted to focus only on genetic alterations occurred during neoplastic transformation and in particular on those that, modifying proteins sequence could be drivers in the acquisition of tumor cells aggressive phenotype.

The exclusion of germinal variants, polymorphisms and non coding variants preserved the model from bias do to patients genetic predisposition and susceptibility, permitting to define a “tumor specific” classification model more easily applicable to further different populations.

Consistently, after Variant Studio workflow, variants annotated as synonymous variants, 3’ prime UTR variants, inframe deletions, intron variants and non coding exon variants have been excluded from further analysis. Also missense variants considered polymorphisms because known to be highly frequent in normal population were excluded.

Table 6 summarizes all 1178 genetic alterations identified by sequencing: following criterions described above 608 variants were excluded and 285 were included in further analysis.

1178 Genetic variants identified												
285 Genetic variants included in the analysis							**608** Genetic variants excluded from the analysis					
	Missense variants	Frameshift variants	Splice acceptor variants	Splice donor variants	Stop gain variants	Tot	Synonymous variants	3 prime UTR variants	Polymorphic Missense variants	Intron variants	Non coding exon variants	Tot
AKT1	1	0	0	0	0	1	0	0	0	0	0	0
ALK	1	0	0	0	0	1	0	0	0	0	0	0
APC	2	2	0	0	5	9	78	0	0	0	0	78
BRAF	5	0	0	0	0	5	0	0	0	0	0	0
CDH1	1	0	0	0	0	1	8	0	0	0	0	8
CTNNB1	16	0	0	0	0	16	0	0	0	0	0	0
EGFR	6	0	0	0	0	6	2	0	0	1	66	69
ERBB2	0	0	0	0	0	0	2	0	0	0	0	2
FBXW7	19	0	0	0	5	24	1	0	0	0	0	1
FGFR2	10	0	0	0	0	10	1	0	0	0	0	1
FOXL2	0	0	0	0	0	0	0	0	0	0	0	0
GNAQ	3	0	0	0	0	3	0	0	0	46	0	46
GNAS	1	0	0	0	0	1	0	0	0	0	0	0
KIT	2	0	0	0	0	2	27	0	0	0	0	27
KRAS	18	0	0	0	0	18	0	35	0	0	0	35
MAP2K1	0	0	0	0	0	0	0	0	0	0	0	0
MET	14	0	1	0	0	15	128	0	3	0	0	131
MSH6	3	0	0	0	0	3	1	0	0	0	0	1
NRAS	6	0	0	0	0	6	0	0	0	0	0	0
PDGFRA	2	0	0	0	0	2	113	0	0	0	0	113
PIK3CA	53	0	0	0	0	53	4	0	0	0	0	4
PTEN	40	17	3	2	16	78	4	0	0	2	0	6
SMAD4	5	0	0	0	0	5	0	0	0	0	0	0
SRC	0	0	0	0	0	0	1	0	0	0	0	1
STK11	1	0	0	0	0	1	0	0	0	0	0	0
TP53	19	2	0	0	4	25	9	0	76	0	0	85

Table 6: Numeric report of all 1178 genetic variants identified by NGS. Only 285 somatic variants supposed to have effect on proteins functions were included in further analysis.

3.4. Results validation by Sanger sequencing

To confirm the NGS results reliability we decided to validate some mutations randomly using gold standard Sanger sequencing. In figure 2 electropherograms show results obtained on PIK3CA exon 20 and KRAS exon 2. A subgroup of WT and mutated samples were analyzed and all NGS results were confirmed.

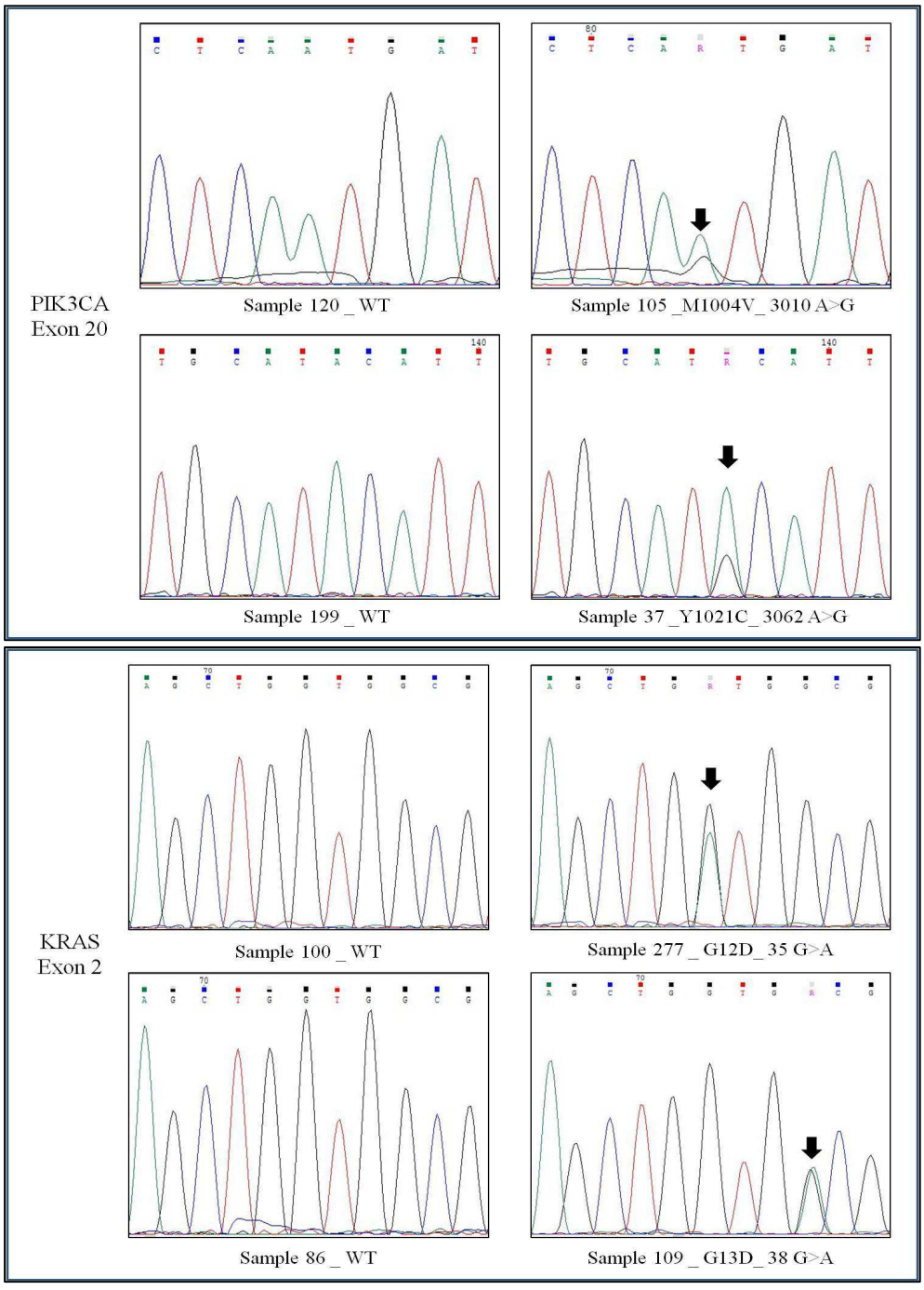

Figure 2: Electropherograms obtained by Sanger sequencing validation of KRAS and PIK3CA genes mutations.

Chapter 4

Study population

4.1. Clinical and pathological description

In this study we analyzed DNA from Formalin Fixed Paraffin Embedded tumoral tissues of 89 patients with EC.

Our study population had a mean age of 65 years (range 42-85 years) and a mean BMI of 30 (range 19-59). Fifty patients had hypertension and 18 were affected by diabetes.

The population was composed by 82 patients with endometrioid, type 1 EC and 7 patients with type 2 EC. Among the 82 type 1 patients 33 had a G1 well differentiated tumor, 16 had a G2 tumor and 33 patients had a G3 undifferentiated tumor.

Thirteen tumors were classified as FIGO stage IA, 36 as IB, 17 as IC; 8 patients had a tumor with FIGO stage II, 14 stage III (5 IIIA, 1 IIIB and 8 IIIC) and 1 stage IV.

49 cases were classified as Lax Kurman low grade, 36 as Lax Kurman high grade and 4 cases remained unclassified.

Mean follow up was 79 months (range 1-192 months), 14 patients had recurrence during follow up (14/89, 15.7%) and 11 patients died because o the tumor (11/89, 12.3%).

4.2. Molecular description

For this study only variants passing quality filter with a probable effect on proteins function were considered.

Seventy-six of 89 cases presented at least one mutation in one of the 26 genes analyzed by Trusight tumor while 13 didn’t present any somatic mutations in genes considered.

Four genes (ERBB2, FOXL2, MAP2K1, SRC) resulted non-mutated in all cases and were excluded.

Further analyses were conducted on the 22 mutated genes. Only somatic mutation with effect on protein coding were taken into account. All variants considered passed Variant Studio quality filter, with a minimum frequency of 3%, a sequency depth of at least 1000 and a minimum average depth of 500x.

PTEN and PIK3CA resulted the most mutated genes: 66/89 (74.1%) patients presented at least one mutation in PTEN, 37/89 (41.6%) in PIK3CA. 20 patients had more than one somatic mutation in PTEN and 12 had more than one in PIK3CA. Even 4 mutations for gene in the same patient were identified for PIK3CA and PTEN.

Twelve genes (APC, BRAF, CTNNB1, EGFR, FGFR2, KRAS, MET, NRAS, PIK3CA, PTEN, SMAD4, TP53) presented at least 5 mutations.

Five genes (AKT1, ALK, CDH1, GNAS, PDGFRA) were mutated in only one patient.

APC, CTNNB1, EGFR, APC, CTNNB1, EGFR, KRAS, MET, NRAS, PIK3CA, PTEN, TP53 genes presented in some patients the coexistence of more than one variant in different nucleotide positions. (Table 7).

Gene	**Total number of mutation**	**Number of mutated patients**	**HGVSc**
AKT1	1	1	c.142C>T
ALK	1	1	c.3521T>C
APC	9	5	c.3925G>T, c.4630G>T, c.4729G>T, c.2626C>T, c.4661delA, c.4738A>G, c.2677G>A, c.4385_4386delAG
BRAF	5	5	c.1328G>T, c.1805C>A
CDH1	1	1	c.1073C>T
CTNNB1	16	14	c.94G>C, c.122C>T, c.94G>A, c.101G>A, c.121A>G, c.101G>T, c.134C>T, c.98C>G, c.110C>T, c.100G>A, c.97T>G
EGFR	6	5	c.2505C>A, c.2505C>A, c.2491C>T, c.2258C>T, c.2591C>T

FBXW7	24	16	c.1660G>T, c.1719C>A, c.2009G>T, c.2065C>T, c.1660G>T, c.1513C>T, c.2066G>A, c.1345G>T, c.1436G>A, c.1634A>T, c.1393C>T, c.1694T>G, c.1552G>A, c.1268G>T, c.1394G>A
FGFR2	10	10	c.755C>G
GNAQ	3	3	c.524C>T, c.803C>T, c.562G>T
GNAS	1	1	c.2524C>T
KIT	2	2	c.1652C>A, c.1444G>A
KRAS	18	16	c.35G>T, c.35G>C, c.35G>A, c.312G>T, c.38G>A, c.35G>T, c.34G>T
MET	15	11	c.1586G>T, c.638C>T, c.1586G>T, c.3817C>A, c.4036C>A, c.3314-1G>T, c.901A>G, c.3029C>T, c.1688C>T, c.2962C>T, c.504G>T
MSH6	3	3	c.3232G>T, c.3319G>T, c.3388G>A
NRAS	6	5	c.191A>G, c.235C>A, c.405G>T, c.181C>A, c.35G>A, c.122G>A
PDGFRA	2	1	c.1780G>T, c.1921C>T
PIK3CA	53	37	c.263G>A, c.1337G>T, c.1634A>G, c.302_304delTAA, c.3140A>G, c.3143A>G, c.113G>A, c.112C>T, c.3139_3140delCAinsAT, c.3010A>G, c.333G>C, c.3104C>T, c.1258T>C, c.3169T>C, c.241G>A, c.1345C>T, c.419G>A, c.277C>T, c.329_331delAAA, c.1351G>A, c.3062A>G, c.3073A>G, c.278G>A, c.23G>A, c.1634A>C, c.1633G>A, c.1625A>C, c.317G>T, c.353G>A, c.1624G>A

PTEN	81	40	c.748delT, c.389G>A, c.462C>A, c.388C>T, c.361G>A, c.193T>G, c.395G>A, c.528delT, c.388C>G, c.64_69delGACTTA, c.517C>T, c.697C>T, c.224_228delATTAT, c.217_218insA, c.406T>C, c.403A>G, c.1031_1040delAGCTGTACTT, c.289C>T, c.511C>T, c.601G>T, c.794_795insA, c.380G>A, c.518G>A, c.94_96delATT, c.635-1G>A, c.740_741insA, c.253+1G>T, c.697_700delCGAC, c.274G>T, c.295G>T, c.19G>T, c.16A>G, c.100G>A, c.493G>A, c.389G>T, c.217G>T, c.634+1G>T, c.85_101delTATCCAAACATTATTGC, c.795delA, c.179A>C, c.431A>C, c.677C>T, c.237_238insA, c.464A>G, c.165-2A>G, c.631_632insG, c.665_678delTGAAGATATATTCC, c.37A>T, c.746T>G, c.635-1G>C, c.511C>A, c.610delC, c.389delG, c.385G>A, c.476G>T, c.323T>A
SMAD4	5	5	c.1544G>T, c.1612G>A, c.1609G>T, c.1487G>A
STK11	1	1	c.929G>A
TP53	25	22	c.659A>G, c.610G>T, c.1091C>A, c.645T>A, c.817C>T, c.659A>G, c.523C>T, c.359A>C, c.475G>A, c.637C>T, c.365_366delTG, c.524G>A, c.800G>A, c.380C>T, c.140delC, c.799C>T, c.742C>T, c.743G>A, c.452C>G, c.993G>T

HGVSc= Human Genome Variation Society coding sequences

Table 7: Summary of somatic mutations alleged to have effect on protein function and included in further analysis.

Chapter 5

Unsupervised identification of two population subgroups with different prognosis

5.1. Unsupervised hierarchical clustering analysis

We used an unsupervised hierarchical clustering analysis to automatically subdivide our EC patients in two groups with different molecular characteristics to principally understand if this small genetic profile obtained by Trusight Tumor could be sufficient to individuate any differences within our study population.

We decided to extract two clusters, hoping to obtain two numerically comparable groups of patients and to avoid the formation of subgroups too small to be statistically analyzed.

Unsupervised hierarchical clustering analysis was conducted considering the number of non-silent mutations occurred in each gene for each patient. Variables were considered as ordinal values with a range from 0 (no mutation) to 4. Euclidean distance was used to compute distance measures when clusters are generated in order to highlight the difference between the number of mutations. Ward agglomerative hierarchical clustering procedure was applied. Only data obtained from sequencing were used as attributes in the analysis, none clinical variable was included.

Clustering analysis derive a first cluster composed by 23 cases (cluster 1) and a second cluster composed by 66 cases (cluster 2). Interestingly, no G1 well differentiated EC are present in cluster 1 (Figure 3).

Unsupervised identification of two population subgroups with different prognosis

Unsupervised identification of two population subgroups with different prognosis

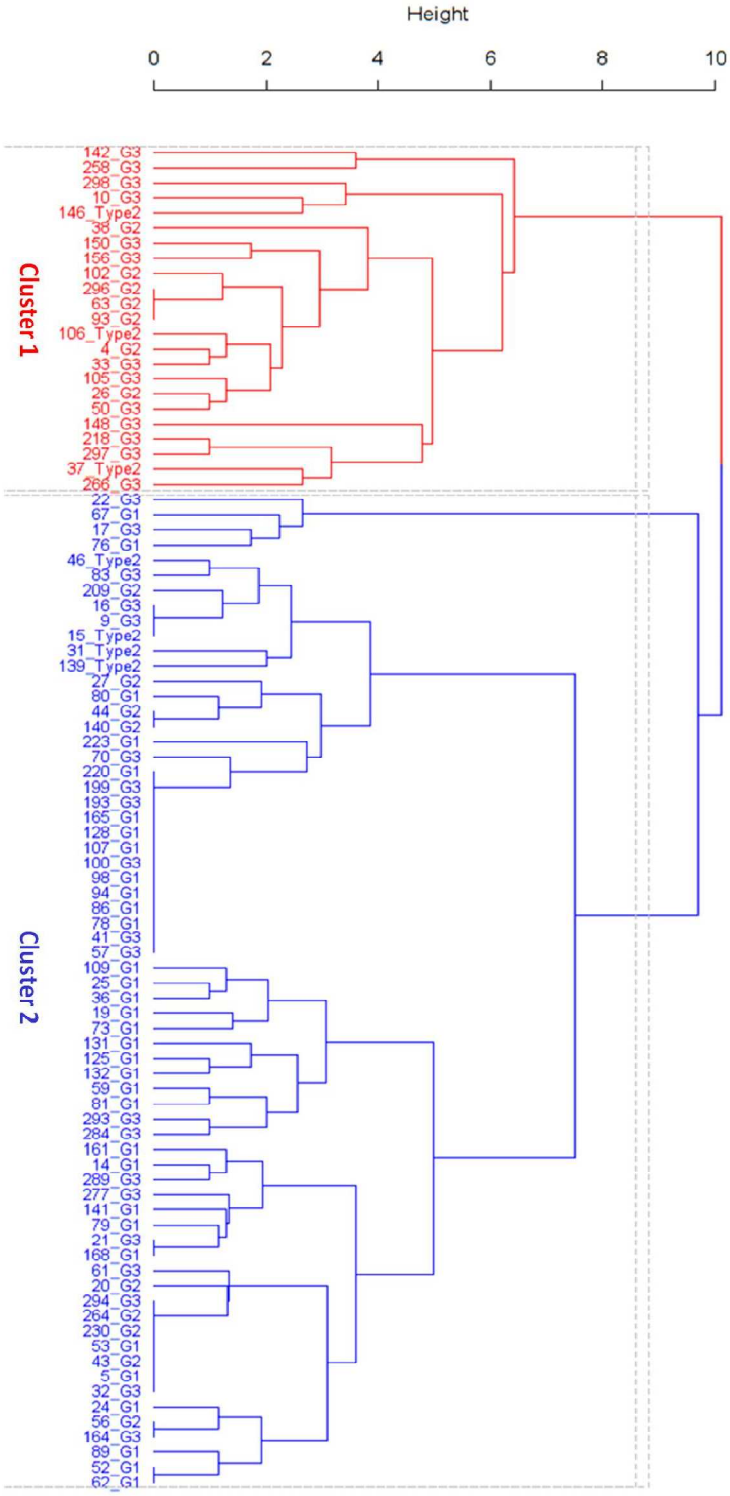

Figure 3: Binary unsupervised hierarchical clustering

5.2. Overall survival and disease free survival analysis

To determinate if molecular based clusterization could be useful to differentiate patients with different prognosis independently from histological characteristics, a Cox proportion hazard model was applied to compare overall survival and disease free survival in the two clusters.

Initially, the analysis was conducted considering total population (89 patients). Table 8 summarizes the number of events of death and recurrence registered in total population and reported the hazard ratio between the two clusters. Differences between the two groups didn't reached significance but, in particular for disease free survival, an interesting difference was observed in the percentage of events registered in the two groups. Considering EC, in which the events of death and recurrence are normally a small percentage, a difference from 22% for cluster 1 to 14% for cluster 2 could be clinically interesting, the lack of a significance in statistical analysis could probably be due to the small number of patients included in the study.

The Kaplan Meier curves (Figure 4) effectively show this different trend of DFS between the two clusters.

Total Population					
		Patients	**Events N,(%)**	**HR**	**Logrank P value**
Overall Survival	Cluster 1	23	4 (17%)	-	-
	Cluster 2	66	7 (11%)	0.46	0.205
Disease Free Survival	Cluster 1	23	5 (22%)	-	-
	Cluster 2	66	9 (14%)	0.42	0.119

Table 8: Cox proportional hazard model for overall survival and disease free survival comparison between the 2 clusters. Total population (89 patients) was considered.

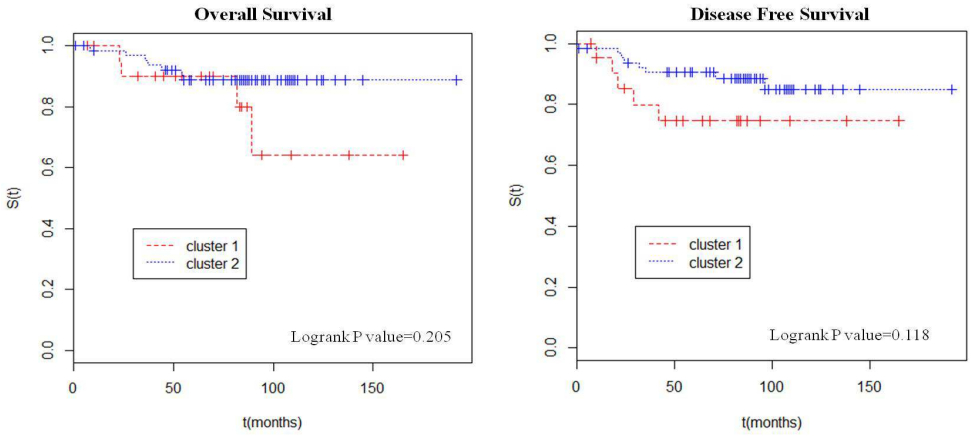

Figure 4: Kaplan Meier curves of overall survival and disease free survival in the 2 clusters. Total population (89 patients) was considered.

The same analysis was also carried on considering only patients with a histotype 1 EC (82 cases), to evaluate if molecular clusterization could be differently applied to subpopulations with different histological characteristics. In this case Cox proportion hazard model demonstrated a significant difference (Logrank P value= 0.033) in overall survival between the two clusters, in particular cluster 2 presented a 4 times lower risk of death because of the tumor (HR=0.26) (Table 9).

Disease free survival analysis didn't result in a significant difference between the two clusters but showed an interesting trend (Logrank P value= 0.108), observable in Kaplan Meier curves (Figure 5), as already described for total population. We think that the lack of significance of this data could be due the small number of events considered and our observation could be reinforced increasing the number of cases considered.

Histotype 1 Population					
		Patients	**Events N, (%)**	**HR**	**Logrank P value**
Overall Survival	Cluster 1	20	4 (20%)	-	-
	Cluster 2	62	5 (8%)	0.26	***0.033***
Disease Free Survival	Cluster 1	20	4 (20%)	-	-
	Cluster 2	62	7 (11%)	0.38	0.108

Table 9: Cox proportional hazard model for overall survival and disease free survival comparison between the 2 clusters. Type I EC population (82 patients) was considered.

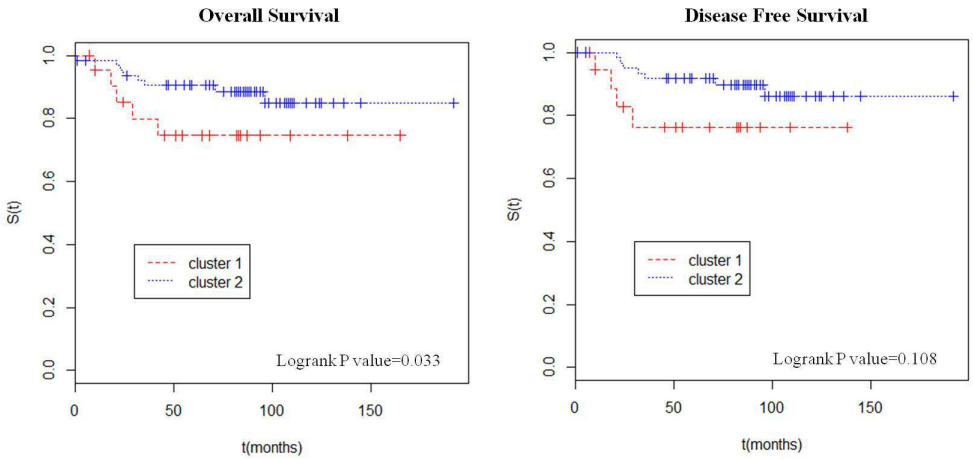

Figure 5: Kaplan Meier curves of overall survival and disease free survival in the 2 clusters. Type I EC population (82 patients) was considered.

Although it is clear the need to enlarge the study population to confirm survival differences between the two groups obtained by molecular clusterization, we judge the method as an useful tool for distinguish “good prognosis” EC patients from “poor prognosis” EC patients.

As clustering conformation suggested (figure 3), we decided to evaluate also samples clusterization and prognostic profiles considering to subdivide overall population in four clusters instead of two.
The new clustering analysis generated numerically heterogeneous groups of patients (Figure 6A) and Cox analysis demonstrated that not significantly differences in overall survival (Figure 6B) (Logrank P value =0.263) and disease free survival (Figure 6C) (Logrank P value =0.379) can be observed between the 4 population clusters.

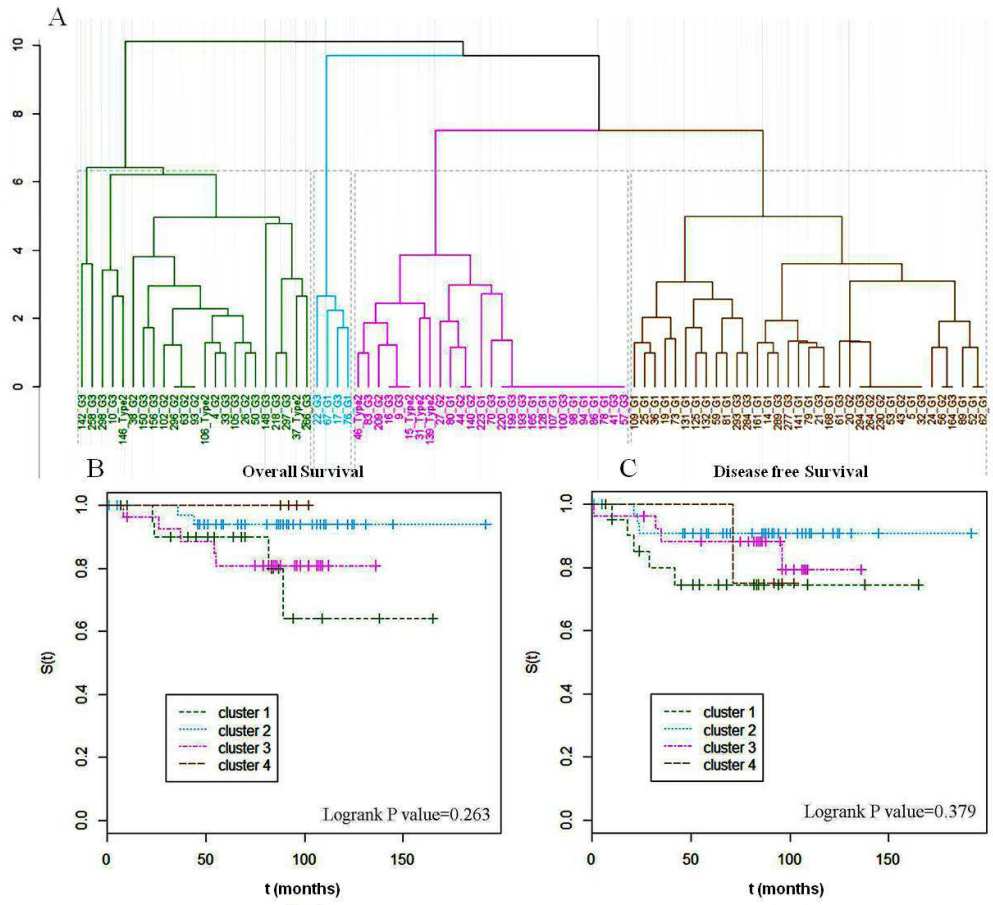

Figure 6: K=4 hierarchical clustering (A) and Kaplan Meier curves representing overall survival (B) and disease free survival (C) trend in EC population subdivided in 4 clusters.

Chapter 6

Characterization of the generated prognostic clusters

6.1. Statistical analysis of frequency of the clinical-pathological characteristics between the two clusters

After a first evaluation of the capability of the clusterization method to distinguish two groups with different prognosis we performed some statistical association analysis to investigate how the unsupervised approach subdivided samples.

Fisher test was used to analyze statistical association between clusterization and clinic-pathological characteristics of EC patients.

Table 10 summarizes the frequencies of clinical features within the two clusters. Very intriguingly data show a strong association between clusterization and tumor grading (P value <0.001), in particular the molecular model perfectly distinguish type 1 G1 tumors, localizing as expected all these cases in the “good prognosis” cluster 2. Also Lax Kurman histological classification resulted significantly associated with cluster subdivision, as expected about the 85% of the low grade tumors were classified in cluster 2.

No differences of age, BMI, FIGO stage, lymph nodes positivity between the two clusters was observed.

		CLUSTER		P value
		1 N, (%)	**2 N,(%)**	
Tot		23	66	
Age	64.8±10.1	66.6±11.7	64.1±9.6	0.328
BMI	30.7±8.4	31.8±10.0	30.4±7.9	0.523
Grade				***<0.001***
G1	33	0 (0.0)	33 (100.0)	
G2	16	7 (43.8)	9 (56.2)	
G3	33	13 (39.4)	20 (60.6)	
Histotype 2	7	3 (42.9)	4 (57.1)	
Lax Kurman				***0.037***
Low	49	7 (14.3)	42 (85.7)	
High	36	13 (36.1)	23 (63.9)	
NA	4	3	1	
FIGO Stage				0.522
I -II	74	18 (24.3)	56 (75.7)	
III-IV	15	5 (33.3)	10 (66.7)	
Lymph node positivity				1
0	80	21 (26.2)	59 (73.8)	
At least 1	9	2 (22.2)	7 (77.8)	

Table 10: Distribution of clinical features within the two clusters. P value were calculated performing Fisher test

These data suggest that the molecular clusterization based on Trusight tumor 26 profiling could be particularly useful to distinguish tumor at different histological grades and could be used to support cases where the histopathology classification results particularly difficult.
Subsequently, we tried to understand if the unsupervised clusterization was driven by specific mutated genes and if some genetic variants or a different mutational load were characteristic of the two clusters.

6.2. Representation of the gene variants frequencies between the two clusters

In order to investigate the molecular aspect of the two clusters and the drivers responsible for their formation, we generated an heat map that, preserving patients subdivision, represent the number of mutation occurred in each gene for each case (Figure 7). Y axis show clusters dendrogram, each row represent a patient and a color codify for the histological grade of the tumor while X axis reports the gene list. In each column the number of mutations of a gene in different samples are represented.
Some molecular aspects resulted particular evident in heat map representation:

- no mutations in APC gene were found in 66 patients in cluster 2, while 9 mutations in 5/23 patients were observed in cluster 1
- no mutations in KRAS were observed in cluster 1, while 14 patients in cluster 2 presented at least one KRAS mutation
- 23/23 patients in cluster 1 and 14/66 in cluster 2 presented PIK3CA mutations, but all tumors presenting more than one variant of the gene were localized in cluster 1
- In cluster 1 19/23 patients presented both PIK3CA and PTEN mutation. In cluster 2 the coexistence of these mutated genes was observed only in 9/66 cases.

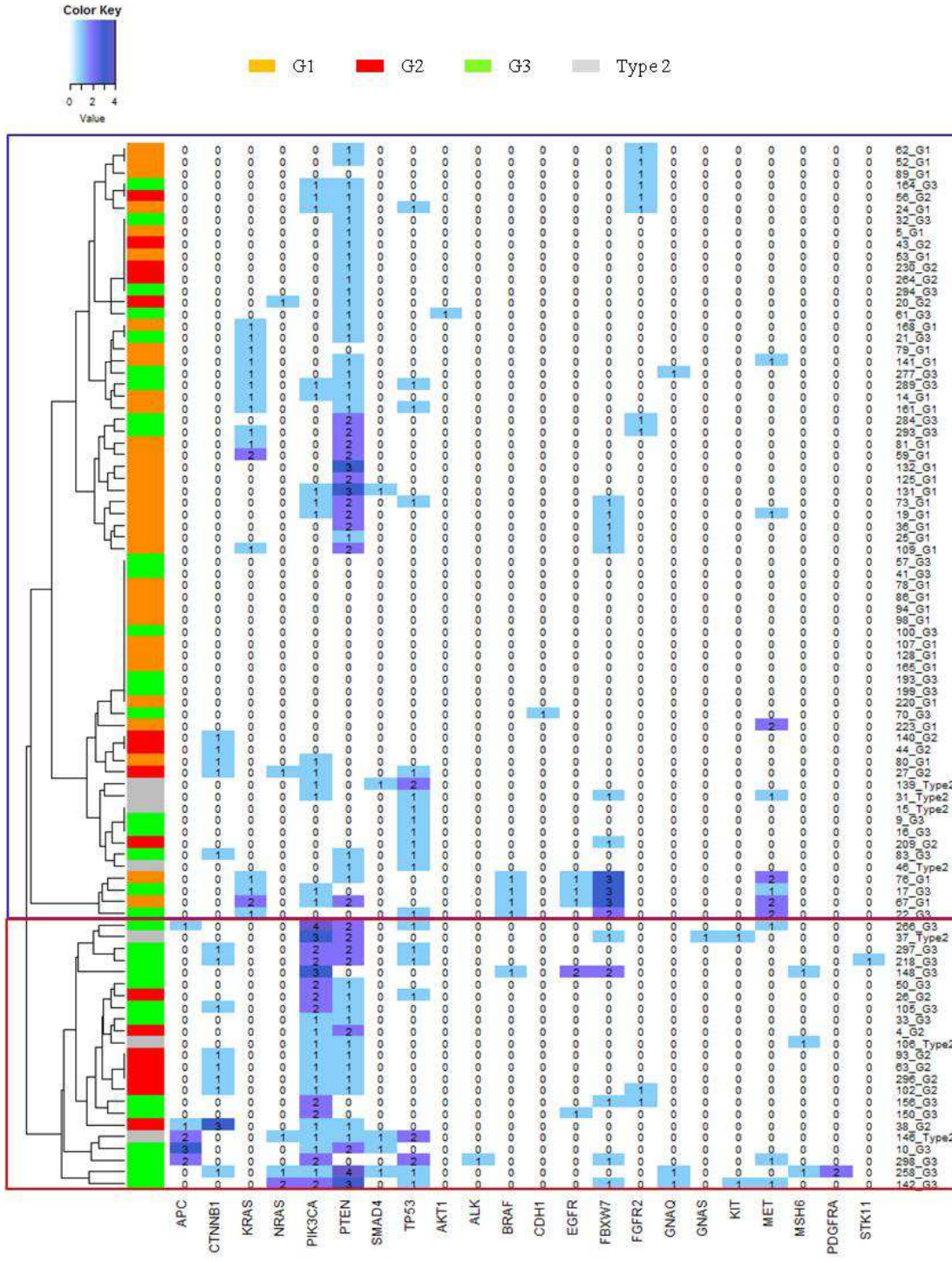

Figure 7: Heat map representation of gene mutations distribution in patients subdivided in 2 clusters

6.3. Statistical analysis of frequency of genes mutations between the two clusters

In order to statistically investigate the observations listed above and identify genes principally associated to the clustering, we analyzed the frequencies of mutations of each gene in the two clusters. Fisher test was used to recognize significant associations. Initially all genes were evaluated but only most interesting ones: with at least 5 mutations in our samples and with lower P values in gene-cluster association analysis, were showed in table 11. These genes (APC, CTNNB1, KRAS, PIK3CA, PTEN, SMAD4 and TP53) had been the only ones considered in further analysis.

A generalized linear model was conducted considering all 7 genes in a multivariate analysis, to discover most significant associations and to individuate principal drivers of hierarchical clustering.

Statistical univariate analysis confirmed the significantly different distribution of APC, CTNNB1, KRAS, PIK3CA, PTEN, as observed in heat map. In addition, total mutational load (calculated considering all 26 Trusight tumor genes) was found to be statistically different in the two clusters, too: while in “bad prognosis” cluster 1 a mean of 5.8 mutations for patient was observed, in cluster 2 the mean mutational load was only 2.3, suggesting as expected that the coexistence of a larger number of mutations could influence the development of a worse tumor phenotype.

Multivariate analysis confirmed in particular the significance of APC, CTNNB1 and PIK3CA mutations different distribution, suggesting a possible role of these gene mutational profiles as drivers of the cluster generation.

Based on the strong association described in table 11 and on the role of some genes mutations in clustering formation we also investigated if mutation of these genes resulted directly associated to different histological grades.

		Cluster			
		1	2	Univariate Analysis P	Multivariate analysis P
Total patients	89	23	66		
APC				***<0.001***	***0.002***
0	84 (94.4)	18 (78.3)	66 (100.0)		
1	2 (2.2)	2 (8.7)	0 (0.0)		
More than 1	3 (3.4)	3 (13.0)	0 (0.0)		
CTNNB1				***0.001***	***0.008***
0	65 (84.3)	14 (60.9)	61 (92.4)		
1	13 (14.6)	8 (34.8)	5 (7.6)		
More than 1	1 (1.1)	1 (4.3)	0 (0.0)		
KRAS				***0.021***	0.067
0	73 (82.0)	23 (100.0)	50 (75.8)		
1	14 (15.7)	0 (0.0)	14 (21.2)		
More than 1	2 (2.3)	0 (0.0)	2 (3.0)		
PIK3CA				***<0.001***	***<0.001***
0	52 (58.4)	0 (0.0)	52 (78.8)		
1	25 (28.1)	11 (47.8)	14 (21.2)		
More than 1	12 (13.5)	12 (52.2)	0 (0.0)		
PTEN				***0.049***	0.255
0	33 (37.1)	4 (17.4)	29 (43.9)		
1	36 (40.4)	11 (47.8)	25 (37.9)		
More than 1	20 (22.5)	8 (34.8)	12 (18.2)		
SMAD4				0.106	0.960
0	84 (94.4)	20 (87.0)	64 (97.0)		
1	5 (5.6)	3 (13.0)	2 (3.0)		
More than 1	0 (0.0)	0 (0.0)	0 (0.0)		
TP53				0.186	***0.042***
0	67 (75.3)	15 (65.2)	52 (78.8)		
1	19 (21.3)	6 (28.1)	13 (19.7)		
More than 1	3 (3.4)	2 (87.0)	1 (1.5)		
Total Mutational Load		5.8±3.1	2.3±2.3	***<0.001***	0.834

Table 11: Univariate and multivariate analysis of association between the mutational status of 7 genes (considered the most interesting) and clusters subdivision

Table 12 shows the significant association of CTNNB1, KRAS, PIK3CA, SMAD4 and TP53 with tumor grading.

	G1	P	G2	P	G3	P	Type 2	P
Total patients	33		16		33		7	
APC		-		0.596		0.114		0.078
0	33 (100.0)		15 (93.8)		30 (90.9)		6 (85.7)	
1	0 (0.0)		1 (6.2)		1 (3.0)		0 (0.0)	
More than 1	0 (0.0)		0 (0.0)		2 (6.1)		1 (14.3)	
CTNNB1		-		***<0.001***		0.239		0.861
0	32 (97.0)		8 (50.0)		28 (84.8)		7 (100.0)	
1	1 (3.0)		7 (43.8)		5 (15.2)		0 (0.0)	
More than 1	0 (0.0)		1 (6.2)		0 (0.0)		0 (0.0)	
KRAS		-		***0.008***		0.098		0.051
0	23 (69.7)		16 (100.0)		27 (81.8)		7 (100.0)	
1	8 (24.2)		0 (0.0)		6 (18.2)		0 (0.0)	
More than 1	2 (6.1)		0 (0.0)		0 (0.0)		0 (0.0)	
PIK3CA		-		***0.050***		***0.001***		***0.027***
0	26(78.8)		7(43.8)		17 (51.5)		2 (28.6)	
1	7 (21.2)		8 (50.0)		6 (18.2)		4 (57.1)	
More than 1	0 (0.0)		1 (6.2)		10 (30.3)		1 (14.3)	
PTEN		-		0.591		0.525		0.485
0	12 (36.4)		4 (25.0)		14 (42.4)		3 (42.9)	
1	11 (33.3)		11 (68.8)		11 (33.3)		3 (42.9)	
More than 1	10 (30.3)		1 (6.2)		8 (24.3)		1 (14.2)	
SMAD4		-		0.659		0.584		***0.008***
0	32 (97.0)		16 (100.0)		31 (100.0)		5 (71.4)	
1	1 (3.0)		0 (0.0)		2 (0.0)		2 (28.6)	
More than 1	0 (0.0)		0 (0.0)		0 (0.0)		0 (0.0)	
TP53		-		0.502		***0.021***		***<0.001***
0	30 (90.9)		13 (81.2)		22 (66.7)		2 (28.6)	
1	3 (9.1)		3 (18.8)		10 (30.3)		3 (42.8)	
More than 1	0 (0.0)		0 (0.0)		1 (3.0)		2 (28.6)	
Total Mutational Load	2.5±2.6	-	2.6±1.5	0.928	4.0±3.6	***0.044***	4.3±2.8	0.151

Table 12: Analysis of association between the mutational status of genes and tumor histological grading

Table shows that for some genes (CTNNB1, KRAS, TP53) the association between mutational status and histology of ECs was observed only for some specific grades of tumor differentiation. PIK3CA was the only gene which mutational status resulted associated to all tumor grades. These data sustain the hypothesis that the analysis of single genes could not be sufficient to distinguish tumor with different prognostic characteristics but the definition of a mutational profile based on the investigation of a small group of genes could be more effective in supporting tumor outcome prediction.

6.4. Lasso and Elastic-Net Regularized Generalized Linear Model to investigate genes effect on patients survival

We applied a Lasso and Elastic-Net Regularized Generalized linear model (glmnet) in order to evaluate if the mutational status of a single gene could independently influence patients' overall survival and disease free survival.

As shown in figure 8A the glmnet applied to overall survival analysis presented an optimal lambda value of -3.2, associated with the most regularized model. At this lambda value the model maintained only two covariates coefficients activated: CTNNB1 and TP53 presented a coefficient of 0.91 and 0.26 respectively, suggesting a possible independent effect of these genes on patients' overall survival. (Figure 8B)

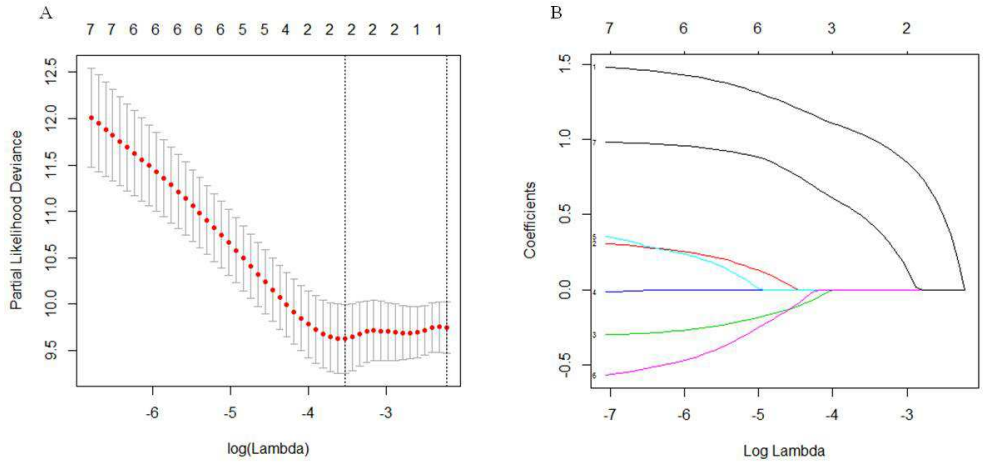

Figure 8: Glmnet used to investigate single gene influence on Overall survival. A. Cross validated error plot for the investigation of the optimal lambda associated with the most regularized model. B. Plot of single genes glmnet coefficient trend in relation to log (lambda).

The same analysis applied to disease free survival presented an optimal lambda at -2.8 but no covariant coefficients were kept activated, suggesting that none of the variables considered have an independent effect on patients disease free survival (Figure 9 A-B)

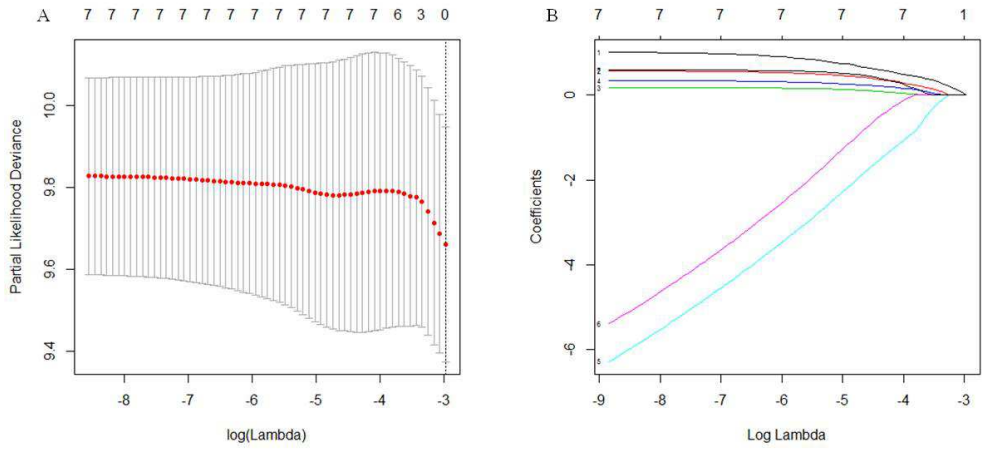

Figure 9: Glmnet used to investigate single gene influence on Disease free survival. A. Cross validated error plot for the investigation of the optimal lambda associated with the most regularized model. B. Plot of single genes glmnet coefficient trend in relation to log (lambda).

Considered together glmnet results supported the necessity to evaluate a more complex molecular profile that integrates the mutational status of the small set of genes considered to generate a prognostic model for endometrial cancer.

6.5. Data mining approaches used to define classification rules driving clusterization

After the evaluation of the statistical association existing between the presence of mutations in some genes, the formation of the two clusters and the patients' survival we wanted to generate a sequence of classification rules that could drive the definition of two groups of EC patients with different prognosis.

To do that we used Orange Canvas software to apply three different methods of data mining: classification tree, CN2 rules analysis and linear regression for nomogram generation. In all three approaches, we submitted the mutational status of genes APC, CTNNB1, KRAS, PIK3CA, PTEN, SMAD4, TP53 as attribute and considered cluster 1 and 2 as classes.

6.5.1. Classification tree

As first method of data mining for cluster generation we created a classification tree, using information gain ratio as attribute selection criterion (Figure 10).

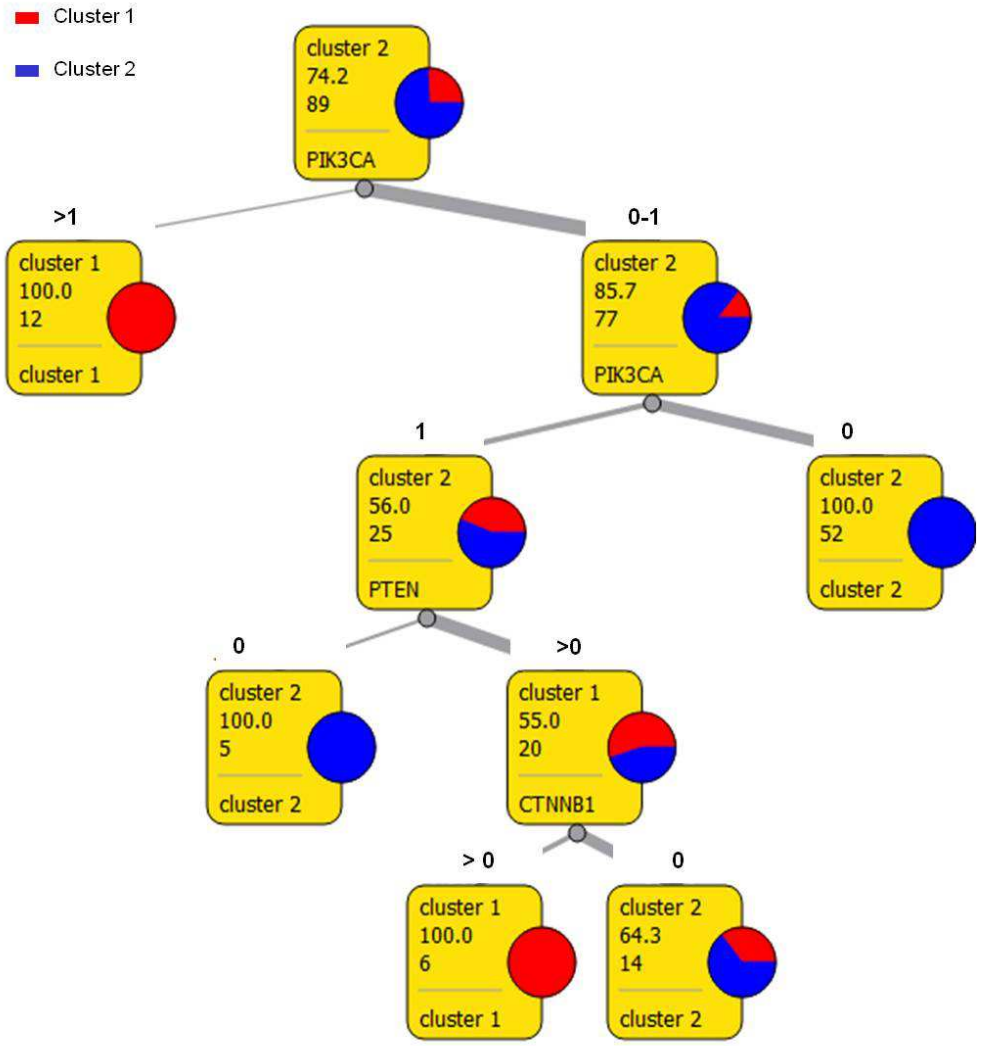

Figure 10: Decision tree for tumors classification in 2 prognostic clusters. The 7 genes mutational status considered were used as attributes. In squares: first row reports the majority class, second row expresses the frequency of the majority class, third row reports the number of instances considered in that leaf, fourth row shows the class of destination or the next attribute that should be evaluated.

We obtained a decision tree were PIK3CA, PTEN and CTNNB1 mutations resulted to be the principal drivers in cluster generation.
In particular, to generate two clusters of EC patients with different prognosis, the method proceeded as follow:

- First step: evaluation of PIK3CA mutational status
    - If tumor presented more than 1 PIK3CA mutation it was classified in cluster 1 (12 cases)
    - If tumor presented no mutation in PIK3CA gene it was classified in cluster 2 (52 cases)
    - If tumor presented only 1 PIK3CA mutation, PTEN had to be evaluated
- Second Step: evaluation of PTEN mutational status
    - If tumor presented 1 mutation in PIK3CA and no mutation in PTEN it was classified in cluster 2 (5 cases)
    - If tumor present 1 mutation in PIK3CA and at least one mutation in PTEN, CTNNB1 had to be evaluated
- Third step: evaluation of CTNNB1 mutational status
    - If tumor presented 1 mutation in PIK3CA, at least one mutation in PTEN and at least one mutation in CTNNB1, it was classified in cluster 1 (6 cases)
    - Otherwise tumors were classified in cluster 2 but for these cases the method was not so accurate

A 10-fold cross validation was used to evaluate this data mining method. Decision tree proposed above had 90% classification accuracy, 76% Matthew Correlation Coefficient, 74% sensitivity and 97% specificity.

6.5.2. CN2 analysis

As second method of data mining for the definition of rules driving our cluster generation we performed a CN2 analysis.

The CN2 algorithm is a classification technique designed for the efficient induction of simple, comprehensible rules of form “if *cond* then predict *class*”, even in domains where noise may be present.

Table 13 summarizes The CN2 rules generated for the distinction of the two clusters we created.

Rule quality	Coverage	IF	THEN
0.929	12	**PIK3CA>1**	**CLUSTER 1**
0.875	6	**CTNNB1>0** and **PTEN>0** and **PIK3CA>0**	**CLUSTER 1**
0.750	2	**APC>0**	**CLUSTER 1**
0.308	11	**PIK3CA>0** and **TP53=0**	**CLUSTER 1**
0.981	52	**PIK3CA=0**	**CLUSTER 2**
0.857	5	**PIK3CA<=1** and **PTEN=0**	**CLUSTER 2**
0.800	3	**KRAS>0**	**CLUSTER 2**

Table 13: CN2 rules summary and quality evaluation

In this case the method, evaluated by a 10-fold cross validation, presented 93% classification accuracy, 82% Matthew correlation coefficient, 74% sensitivity and 100% specificity.

Both data mining methods described confirmed that the majority of EC cases could be classified in 2 groups with different prognosis through the analysis at first of PIK3CA, PTEN and CTNNB1 mutational status. Small difference observed between rules generated with decision tree and CN2 were probably due to the analysis parameters fixed for each one.

6.5.3. Logistic regression and nomogram generation

Considering concordance and differences between the two data mining methods described above we decided to develop a third approach, a logistic regression and nomogram generation, to integrate together the effect of all considered genes mutational load.

The nomogram, based on a logistic regression analysis, was designed to calculate the probability for the patients to take part to cluster 1 (bad prognosis) on the basis of the number of variants sequenced on APC, CTNNB1, KRAS, PIK3CA, PTEN, SMAD4, TP53.

Nomogram represents an easy-to-use system, which can be easily applied in the clinic, to compute the probability for each patient to belong to the bad prognosis group starting from mutational profiling results.

To use this tool clinician have to regulate each ruler of the nomogram in reference to the number of mutation occurred in each gene, and automatically the method calculates the probability of cluster 1 membership (Figure 11).

In this case the method, evaluated by a 10-fold cross validation, presented 89% classification accuracy, 73% Matthew correlation coefficient, 78% sensitivity and 94% specificity.

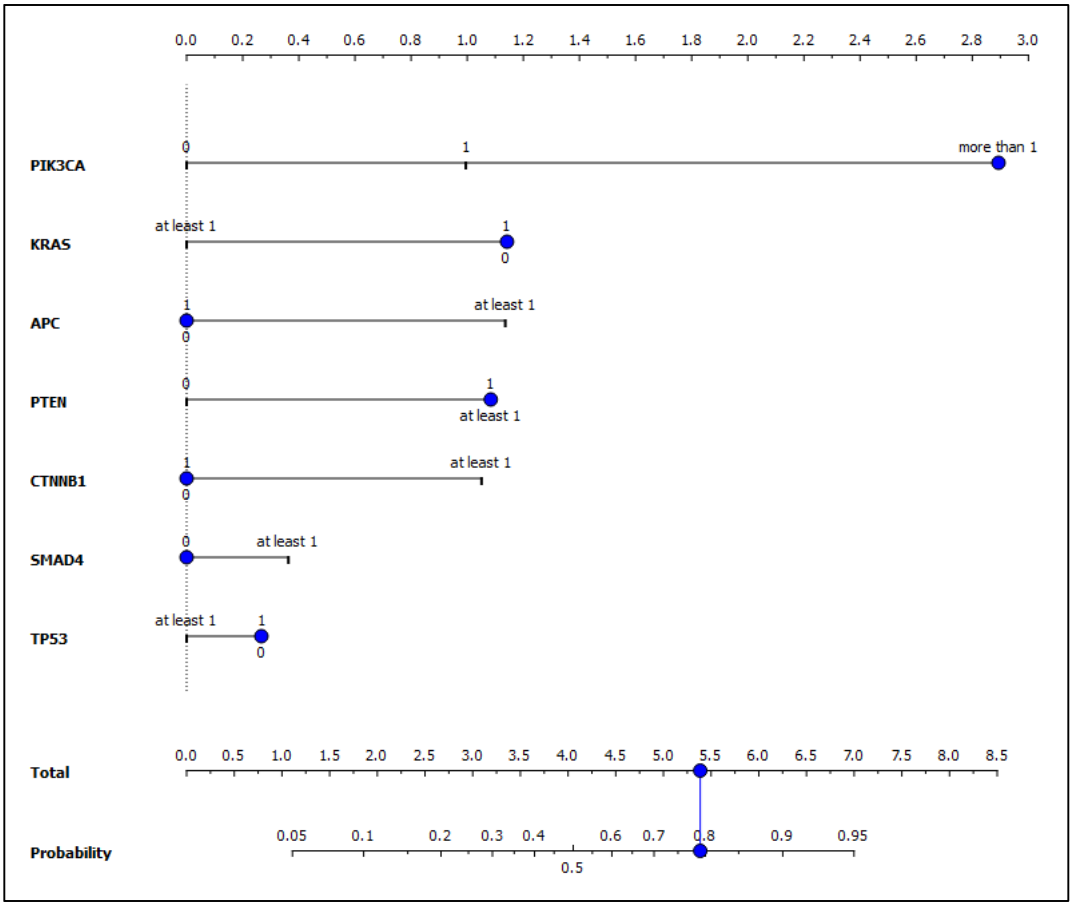

Figure 11: Nomogram tool to calculate patients' probability to belong to the cluster with bad prognosis. Probability scale is expressed as log odd ratio

6.6. Investigation of genes variants damaging effects in clusters

6.6.1. PaPI score calculation

The analysis described above suggested the role of some genes mutations to predict prognosis in endometrial tumors. Interestingly in some cases data demonstrated that not only the existence of mutated genes in tumor DNA, but also the number of mutation registered in each of these genes can influence patient's outcome.

Based on this observation we used PaPI method to quantify the real damaging effect of each mutated genes taking into account both the type and the number of variations occurred on each gene, for each patient.

PaPI is a machine learning ensemble method [27], able to score the functional effect of coding single nucleotide variants, deletions, insertions and indels. The method is based on a pseudo amino acid composition model integrated with Polyphen2 and Sift algorithms of prediction. The PaPI score reflects the probability for the variant to be classified as damaging. The score varies from 0 to 1: values from 0.5 to 1 indicate that the variant is damaging, otherwise it is considered as benign.

The PaPI method takes into account also the existence of different transcripts for a single gene, giving in some case more than one more PaPI score for variant. We calculated the final score as follow:

- we selected the maximum PaPI score for each variant
- when more than one variant for gene coexisted in the same tumor we summed the maximum PaPI score of each one

In this way we obtained a single value of PaPI score for each gene, in each patient.

To evaluate if mutated genes had different damaging effect in the two generated clusters, we compared the PaPI score distribution for each gene in the two groups (Figure 12). Wilcoxon test was used to evaluate statistical differences between curves.

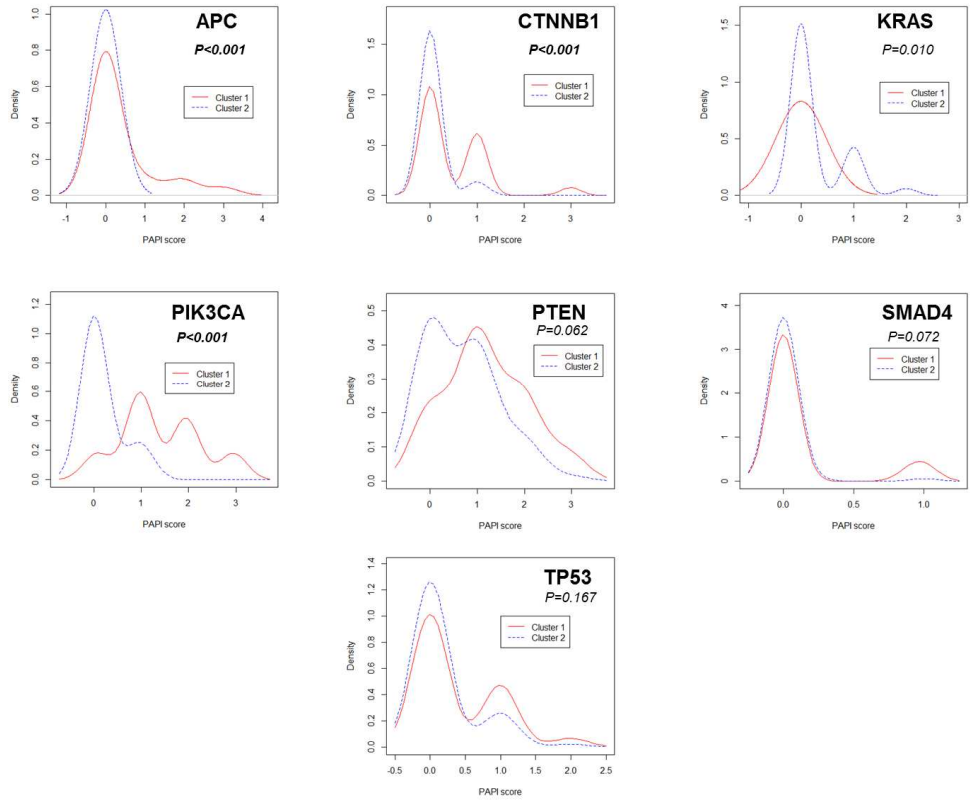

Figure 12: Curves of distribution of gene PaPI score in the 2 clusters. P values were calculated performing Wilcoxon test.

Noticeably, this approach further confirmed the previously described results. APC, CTNNB1, KRAS and PIK3CA displayed a PaPI score distribution significantly different between the two clusters. APC and CTNNB1 curves show that a PaPI score different to 0 is more frequent in cluster 1 (bad prognosis) while opposite situation was observed for KRAS. PIK3CA curves show a higher frequence of Papi Score ranging from 1 to 3 in cluster 1.

This data are in line with conclusions drawn from the analysis of the mutational load. PaPI evaluation of single variants indicated that most of those have a score near to 1 (damaging) so that gene PaPI score and mutational load often overlap.

Chapter 7

Discussion

7.1. Proposal of an alternative molecular-based approach of prognosis prediction in Endometrial cancer

To date histological characterization is the gold standard for EC prognosis. Different tumor histological criterion such as Bockman typing, FIGO stage, grading and Lax Kurman binary classification [4, 19-24] can be used to predict EC outcome.

Nevertheless, in some morphologically intermediate and doubtful cases, anatomo-pathological classification and risk based stratification turns out to be insufficient and inefficient.

In this study we explored the mutational profile of a selected cohort of EC with the aim of developing a genetic signature to improve the current risk based stratification of EC patients.

We used the Trusight tumor 26 kit Illumina to analyze the occurrence of mutations in a panel of 26 cancer related genes, in a population of 89 EC with different histological characteristics and different outcome. An unsupervised hierarchical clustering analysis demonstrated that the mutational profiles obtained from the Trusight tumor analysis effectively separate endometrial tumors in two groups characterized by a different prognosis (good and bad).

Subsequent analysis demonstrated that this clusterization independently identifies G1 well differentiated endometrioid EC, assigning all these tumors, characterized by a positive outcome, to the same “good prognosis” cluster.

Next, statistical analysis were performed to define which mutated genes investigated in the NGS panel could be considered drivers of the

prognostic clusterization. Three different data mining strategies (decision trees, CN2 analysis and logistic regression) were used to define a list of classification rules applicable to EC risk based stratification. APC, CTNNB1, PIK3CA, PTEN, SMAD4 and TP53 resulted the most interesting genes. Rules definition indicates that not only the presence or absence of somatic and damaging mutations on these genes, but also the number of variants occurred on the same gene in each sample can be determinant in predicting patient outcome.

Interestingly PIK3CA and PTEN mutations were the principal determinants of the patients prognostic clusterization. These mutations were already described as very frequent in EC and often coexistent in this kind of tumor. Accumulation of more than one PIK3CA mutations was sufficient to classify patient in the “bad prognosis” cluster 1.

By contrast, in case of presence of a single PIK3CA mutation, the existence of PTEN variants was considered the second classification criteria to predict negative outcome (cluster 1).

The phosphoinositide 3-kinase (PI3K) pathway regulates key aspects of cancer biology including metabolism, cellular growth, survival and resistance to apoptosis [28].

Upon ligand stimulation of tyrosine kinase receptors (RTK), PI3K phosphorylates the lipid phosphatidylinositol 4,5- biphosphate (PIP2), creating phosphatidylinositol 3,4,5- triphosphate (PIP3) [29]. PIP3 recruits pleckstrin homology domain-containing proteins, including the protein kinase AKT, to the membrane. Among its targets, AKT phosphorylates and inhibits tuberous sclerosis complex 2 (TSC2) within the multiprotein TSC complex, which indirectly inhibits mTOR complex 1 (mTORC1). Hence, PI3K-AKT signaling activates mTORC1, a key regulator of metabolism and biosynthetic processes[30]. PTEN hydrolyzes PIP3 back to PIP2, deactivating the pathway [31].

The PI3K/AKT/mTOR pathway is also involved in cross-talk with other signaling pathways, including the RAS/RAF/MEK [32] and estrogen receptor (ER) pathways [33] (Figure 13)[34].

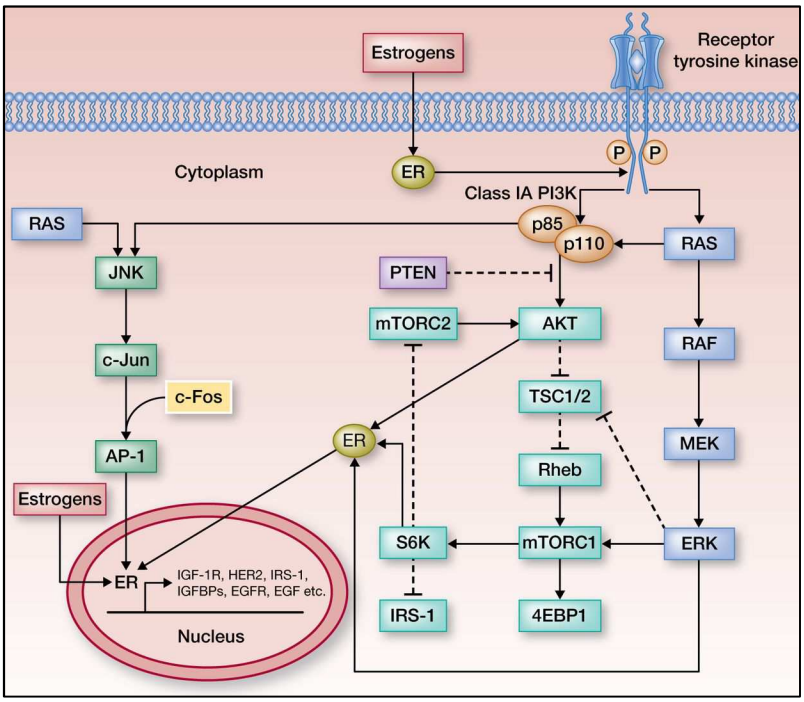

Figure 13: Overview of the PI3K/AKT/mTOR pathway, and cross-talk with other pathways relevant to endometrial cancer [34]

Noticeably, estrogenic signaling, that is considered the principal risk factor for type I EC, is known to heavily cooperate during tumor progression, confirming the existence of a functional correlation between tumor genetic profiles and phenotypic characteristics. Likely, a prolonged exposure to estrogen during women life induces activation of PI3K/AKT signaling, favoring oncogenic mechanisms. Later mutations on crucial genes could maintain PI3K/AKT pathway constitutively activated after the end of the estrogen exposure.

Based on our observations and on literature reports that constitutive activation of the PI3K/AKT pathway in endometrial cancer occurs most commonly through inactivating mutations of PTEN tumor suppressor or activating variants in PIK3CA [35].

Given the frequency of abnormalities in the PI3K/AKT pathway, this signaling pathway represents one of the most promising targets for endometrial cancer therapy. Thus, the identification of genetic mutations within key genes of this pathway, like the one we have identified in our analysis, could represent valuable markers for patient selection and

therapy response monitoring. Figure 14 summarizes the principal PI3K/AKT pathway inhibitors developed in preclinical studies. These molecules fall into 4 main categories: mTOR inhibitors, PI3K inhibitors, dual mTOR/PI3K inhibitors, and AKT inhibitors[34].

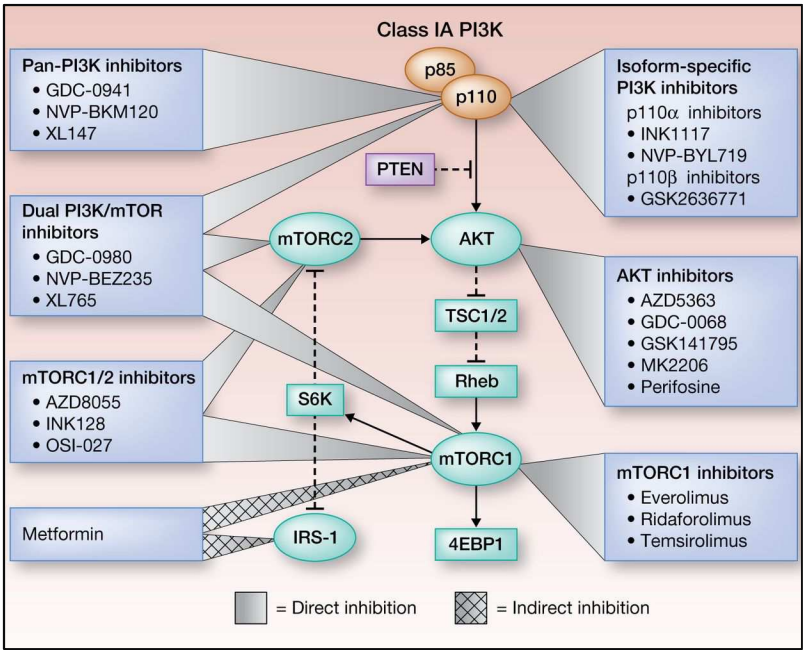

Figure 14: PI3K/AKT/mTOR pathway inhibitors [34]

These data suggest that the approach described by this work could became a double function tool:

- it represent an easy ad relatively economic molecular profiling of EC that could be associated to histological classification to make tumor prognosis in particular in doubtful, intermediate cases
- it could represent a rapid method to investigate mutational status of the genes that up to date are considered the most promising molecular target for EC therapy.

7.2. Related works and molecular models comparison

In 2013 TCGA provided new information on the genetic and molecular events that underlie EC progression and identified four classes of ECs with different molecular profiles that correlate with different patterns of survival.

As already discussed in paragraph 1.4.2 molecular profiles proposed by this work integrated different multiple -omics data and patients prognosis was based in particular on a complex molecular profiling based on genes mutational status, copy number alteration status and Microsatellite Instability [25].

Because the complexity of the approach this stratification results still far to be a applied in clinical routine nevertheless more than one subsequent works [3] [36] tried to combine only the most relevant assays suggested by TCGA to generate a lower labor-intensive and cost-prohibitive prognostic approach.

At the moment it is not possible to compare the method we described in this work with that proposed by TCGA and subsequently adapted for clinical use because not the same genetic features were taken in consideration.

In addition it is important to consider that also different NGS strategies were used and mutational results could not be totally comparable. TCGA applied a tumor/normal pairs exome sequencing, with a minimum read-depth of 20X to investigate mutations occurred during neoplastic transformation and present at high frequency in tumor specimens. We used an amplicon based target resequencing strategy, focused only on genes hot spot regions. Considering the heterogeneity of tumors and the consequent possible existence of different cancer cell clones in the same specimen we decided to apply a minimum read-depth filter of 500X to detect also mutations existing in a small percentage of tumor cells. These mutations, although present in subclones of tumor cells, may be responsible for relapse or therapy resistance, leading to a bad prognosis for the patient.

In future, it should be interesting to integrate our model with only the most effective molecular tests proposed by TCGA model such as POLE mutational status and MSI, to integrate our prognostic approach and try to correctly predict also the outcome of the little percentage of cases remained uncertain.

7.3. Further validations of the model

Even if promising the observations of this work are still preliminary and need to be corroborated in separates sets of EC. Based on the molecular profile, obtained by Trusight tumor next generation sequencing, the unsupervised hierarchical clustering generated 2 groups of patients with different prognosis. In particular significant differences in overall survival were observed between the two groups considering only 82 type I EC patients, while the analysis conducted on total population showed only a trend of difference between clusters' survival curves. Considering total population, DFS varied from 22% of cluster 1 to 14% of cluster 2; the data could be clinically interesting considering the low rate of death and recurrence in EC but P value resulted higher than significance probably because of the low number of patients. To confirm the efficacy of the method in stratifying EC with different prognosis we intend to collect a validation set of at least 100 tumors with various histological characteristics.

To perform molecular profiling in the new set of tumors we will generate a NGS custom panel comprehensive of the amplicons necessary to completely sequence only the seven genes resulted important for the prediction (APC, CTNNB1, PIK3CA, PTEN, SMAD4 and TP53). In this way we will create a smaller sequencing panel that could be economically advantageous both for the validation section but also, if data will be confirmed, for a subsequent introduction in clinical routine.

In parallel with the profiling of a new population by next generation sequencing we will continue to randomly analyze some DNA samples by gold standard Sanger sequencing to validate the reliability of the NGS results.

Chapter 8

Conclusions

This study presented a method for EC stratification in two groups with different prognosis.

The genetic profiling was able to recognize with 100% accuracy tumors with endometrioid G1 characteristics. This data indicate that this method could be particularly useful to improve risk based prediction in those cases with intermediate morphological profile (normally G2) in which histological characteristics are not sufficient to make prognosis.

The study proposes a user-friendly tool for the interpretation of molecular profiling results in order to support prognosis of EC doubtful cases.

If confirmed in an independent validation set this test could be easily introduced in clinical routine. Clinician could propose the NGS panel sequencing when the histology based prognostic interpretation of the EC cases result difficult. A nomogram or a decision tree tool could be equipped to easily interpret NGS panel results and to calculate patient probability to have a good or bad prognosis.

Chapter 9

Methods and supplementary informations

9.1. Next generation sequencing

9.1.1. DNA extraction and quality evaluation

Dna was extracted from Formalin fixed paraffin embedded (FFPE) EC tissues using Maxwell nucleic acid extractor (Promega).

Dna was quantified and quality evaluated using Kapa SYBR Fast qPCR kit as suggested by Illumina.

9.1.2. Trusight Tumor kit Illumina

Ten microliters of DNA, diluted as required after Kapa SYBR Fast qPCR kit, was used for each sample.

Trusight tumor library preparation was subdivided in the following phases.

- Oligo pool hybridization.

FPA and FPB oligo pool were mixed to DNA, placed in a pre-heated scigene block at 95°C for 1 minute and then incubated until the temperature reached 40°C (about 80 minutes)

- Removal of unbound Oligos

Two wash step using stringent wash buffer and filter capable of size selection were performed on samples.

- Extension-ligation of bound oligos.

Hybridized upstream and downstream oligos were connected together through the action of a DNA polymerase and a DNA ligase incubated at 37°C for 45 minutes.

- PCR amplification

The extension-ligation products are amplified using primers that add index sequencing for sample multiplexing as well as common adapters required for cluster generation.

PCR program:
95°C for 3 minutes
27 cycles of:
95°C for 30 seconds
62°C for 30 seconds
72°C for 60 seconds
72°C for 5 minutes
Hold at 10°C

- PCR clean-up

PCR products were purified using AMPure XP beads and 80% Ethanol washes and finally re-suspended in elution buffer

- Libraries quality control

Libraries quality control was performed using Agilent Technologies Bioanalyzer 2100_Agilent DNA 1000 kit.
Libraries size distribution between 300 and 330 bps were considered adequate.

- Libraries quantification and dilution

Libraries was quantified using Qubit fluorimeter. All library were diluted to a concentration of 4 nM

- Libraries denaturing and pooling.

Libraries prepared for a single Miseq Run were pooled together (5 ul for each).
Library pool was denaturated by incubation with 1N NaOH at room temperature for 5 minute and next diluted to a final concentration of 12.5 p M.

9.1.3. MiSeq run

For the sequencing of 89 EC DNA Twelve Miseq run using V2 cartridge 300 cycles (paired end sequencing 2x121) were performed.

MiSeq V2 cartridge 300 cycles produces about 5 Gb output and about 15 M reads. 8 patients (16 library FPA+FPB) were pooled in the same run to obtain a coverage of at least 1000x, as suggested by Illumina for low frequency somatic variants detection.

9.2. NGS data analysis

MiSeq Reporter was used to elaborate MiSeq row data and produce *.fastq and *.vcf files, as described in introduction paragraph.

Variant studio was used to visualize list of mutations occurred in each sample, annotate them and apply selection filters. Based on the high coverage obtained from these run, variant studio considered reliable mutation with a minimum frequency of 5%.

9.3. Sanger Sequencing

Sanger sequencing was performed by 3500Dx Genetic Analyzer (Applied Biosystem) using Big Dye Terminator V 3.1 Cycle sequencing kit.

KRAS exon 2 Sequencing primers
Forward: GTATTAACCTTATGTGTGACA
Reverse: GTCCTGCACCAGTAATATGC

PIK3CA exon 20 Sequencing primers
Forward: ATCATTTGCTCCAAACTGACCA
Reverse: TTGTGTGGAAGATCCAATCCAT

9.4. Hierarchical clustering analysis and statistical analysis

All analysis performed in this study were elaborated using R software.

To generate unsupervised hierarchical clustering we calculated euclidean distance between samples, and Ward clustering method was applied.

Survival analysis was conducted applying Cox model and Kaplan Meier curves were generated.

Analysis of association between clusters and clinical characteristics and between clusters and gene mutations were performed using Fisher test and generalized linear models. Association were considered statistically different if presented a P value lower than 0.05

9.5. Data mining methods

To generate decision tree, CN2 analysis and nomogram Orange Canvas software was used. In all three analysis mutational status of APC, CTNNB1, KRAS, PIK3CA, PTEN, SMAD4 AND TP53 were the only attributes considered and “cluster 1” and “cluster 2” were the two decision class. All attributes were defined as continuous.

Classification accuracy, sensitivity and specificity of these method were calculated after a 10-fold cross-validation.

9.5.1. Decision tree

To generate decision tree gain ratio was used as attribute selection criterion.

For pre-pruning a minimum of 5 instances for leave was fixed. For post-pruning the recursively merging of leaves with the same majority class was performed and m parameter was fixed to 1.

9.5.2. CN2 analysis

In CN2 analysis default parameters were fixed: rule quality estimation was conducted using Laplace method, pre-puning alpha value was fixed to 0.05 and the beam width was 5.

9.5.3. Nomogram

Nomogram was based on logistic regression.

Bad prognosis cluster 1 was considered the target class. The scale for class probability was expressed in log odd ratio.

9.6. PaPI Score analysis

PaPI score of each single variant was calculated submitting all variants to the informatics tool available on http://papi.unipv.it/.

A Perl script was used to select maximum PaPI score of each variant (in cases in which more than one transcript was considered) and then to summarize Papi score of the same gene in cases in which more than one variant for gene in the same patient was detected.

PaPI distribution curves representation was generated using R software and Wilcoxon test was performed to statistically compare the distribution of Papi score for each gene between the two groups of patients.

References

1. Amant, F., et al., Endometrial cancer. Lancet, 2005. 366(9484): p. 491-505.

2. Greggi, S., et al., Management of endometrial cancer in Italy: a national survey endorsed by the Italian Society of Gynecologic Oncology. Int J Surg, 2014. 12(10): p. 1038-44.

3. Talhouk, A. and J.N. McAlpine, New classification of endometrial cancers: the development and potential applications of genomic-based classification in research and clinical care. Gynecol Oncol Res Pract, 2016. 3: p. 14.

4. Bokhman, J.V., Two pathogenetic types of endometrial carcinoma. Gynecol Oncol, 1983. 15(1): p. 10-7.

5. Sorosky, J.I., Endometrial cancer. Obstet Gynecol, 2008. 111(2 Pt 1): p. 436-47.

6. Rice, L.W., Hormone prevention strategies for breast, endometrial and ovarian cancers. Gynecol Oncol, 2010. 118(2): p. 202-7.

7. Reis, N. and N.K. Beji, Risk factors for endometrial cancer in Turkish women: results from a hospital-based case-control study. Eur J Oncol Nurs, 2009. 13(2): p. 122-7.

8. Zucchetto, A., et al., Hormone-related factors and gynecological conditions in relation to endometrial cancer risk. Eur J Cancer Prev, 2009. 18(4): p. 316-21.

9. Pfeiffer, R.M., et al., Timing of births and endometrial cancer risk in Swedish women. Cancer Causes Control, 2009. 20(8): p. 1441-9.

10. Furberg, A.S. and I. Thune, Metabolic abnormalities (hypertension, hyperglycemia and overweight), lifestyle (high energy intake and physical inactivity) and endometrial cancer risk in a Norwegian cohort. Int J Cancer, 2003. 104(6): p. 669-76.

11. Lindemann, K., et al., Body mass, diabetes and smoking, and endometrial cancer risk: a follow-up study. Br J Cancer, 2008. 98(9): p. 1582-5.

12. Saltzman, B.S., et al., Diabetes and endometrial cancer: an evaluation of the modifying effects of other known risk factors. Am J Epidemiol, 2008. 167(5): p. 607-14.

13. Bergman, L., et al., Risk and prognosis of endometrial cancer after tamoxifen for breast cancer. Comprehensive Cancer Centres' ALERT Group. Assessment of Liver and Endometrial cancer Risk following Tamoxifen. Lancet, 2000. 356(9233): p. 881-7.

14. Zhou, B., et al., Cigarette smoking and the risk of endometrial cancer: a meta-analysis. Am J Med, 2008. 121(6): p. 501-508.e3.

15. Yang, H.P., et al., Endometrial cancer risk factors by 2 main histologic subtypes: the NIH-AARP Diet and Health Study. Am J Epidemiol, 2013. 177(2): p. 142-51.

16. Lu, K.H., et al., Prospective determination of prevalence of lynch syndrome in young women with endometrial cancer. J Clin Oncol, 2007. 25(33): p. 5158-64.

17. Committee opinion no. 634: Hereditary cancer syndromes and risk assessment. Obstet Gynecol, 2015. 125(6): p. 1538-43.

18. Denschlag, D., U. Ulrich, and G. Emons, The diagnosis and treatment of endometrial cancer: progress and controversies. Dtsch Arztebl Int, 2010. 108(34-35): p. 571-7.

19. Bansal, N., V. Yendluri, and R.M. Wenham, The molecular biology of endometrial cancers and the implications for pathogenesis, classification, and targeted therapies. Cancer Control, 2009. 16(1): p. 8-13.

20. Murali, R., R.A. Soslow, and B. Weigelt, Classification of endometrial carcinoma: more than two types. Lancet Oncol, 2014. 15(7): p. e268-78.

21. Creasman, W.T., et al., Carcinoma of the corpus uteri. FIGO 26th Annual Report on the Results of Treatment in Gynecological Cancer. Int J Gynaecol Obstet, 2006. 95 Suppl 1: p. S105-43.

22. Creasman, W., Revised FIGO staging for carcinoma of the endometrium. Int J Gynaecol Obstet, 2009. 105(2): p. 109.

23. Beddy, P., et al., FIGO staging system for endometrial cancer: added benefits of MR imaging. Radiographics, 2012. 32(1): p. 241-54.

24. Lax, S.F., et al., A binary architectural grading system for uterine endometrial endometrioid carcinoma has superior reproducibility compared with FIGO grading and identifies subsets of advance-stage tumors with favorable and unfavorable prognosis. Am J Surg Pathol, 2000. 24(9): p. 1201-8.

25. Kandoth, C., et al., Integrated genomic characterization of endometrial carcinoma. Nature, 2013. 497(7447): p. 67-73.

26. Creasman, W.T., et al., Surgical pathologic spread patterns of endometrial cancer. A Gynecologic Oncology Group Study. Cancer, 1987. 60(8 Suppl): p. 2035-41.

27. Limongelli, I., S. Marini, and R. Bellazzi, PaPI: pseudo amino acid composition to score human protein-coding variants. BMC Bioinformatics, 2015. 16: p. 123.

28. Myers, A.P., New strategies in endometrial cancer: targeting the PI3K/mTOR pathway--the devil is in the details. Clin Cancer Res, 2013. 19(19): p. 5264-74.

29. Whitman, M., et al., Type I phosphatidylinositol kinase makes a novel inositol phospholipid, phosphatidylinositol-3-phosphate. Nature, 1988. 332(6165): p. 644-6.

30. Ma, X.M. and J. Blenis, Molecular mechanisms of mTOR-mediated translational control. Nat Rev Mol Cell Biol, 2009. 10(5): p. 307-18.

31. Stambolic, V., et al., Negative regulation of PKB/Akt-dependent cell survival by the tumor suppressor PTEN. Cell, 1998. 95(1): p. 29-39.

32. Mendoza, M.C., E.E. Er, and J. Blenis, The Ras-ERK and PI3K-mTOR pathways: cross-talk and compensation. Trends Biochem Sci, 2011. 36(6): p. 320-8.

33. Miller, T.W., J.M. Balko, and C.L. Arteaga, Phosphatidylinositol 3-kinase and antiestrogen resistance in breast cancer. J Clin Oncol, 2011. 29(33): p. 4452-61.

34. Slomovitz, B.M. and R.L. Coleman, The PI3K/AKT/mTOR pathway as a therapeutic target in endometrial cancer. Clin Cancer Res, 2012. 18(21): p. 5856-64.

35. Westin, S.N. and R.R. Broaddus, Personalized therapy in endometrial cancer: challenges and opportunities. Cancer Biol Ther, 2012. 13(1): p. 1-13.

36. Talhouk, A., et al., A clinically applicable molecular-based classification for endometrial cancers. Br J Cancer, 2015. 113(2): p. 299-310.