UNIVERSITAôDEGLI STUDI DI PADOVA
Dipartimento di Scienze Chirurgiche, Oncologiche e Gastroenterologiche

SCUOLA DI DOTTORATO DI RICERCA IN ONCOLOGIA E ONCOLOGIA CHIRURGICA
XXVIII CICLO

An integrated proteomic and genomic approach to study FAP patients without *APC* and *MutHY* mutations

Direttore della Scuola : Ch.mo Prof. Paola Zanovello
Supervisore :Ch.mo Dott.. Marco Agostini

Dottorando : Lisa Agatea

SUMMARY

ABSTRACT

Familial Adenomatous Polyposis (FAP) is one of the most important clinical forms of inherited susceptibility to colorectal cancer, that is characterized by the development of hundreds to thousands of adenomas in the colon and rectum during the second decade of life. FAP is due to a germline mutation in the *APC* gene or to biallelic variations of *MutYH* gene. Almost all patients will develop cancer if the disease is not identified and surgically treated at an early stage.

The aim of this study was to characterize, by peptidomic and genetic approaches, 4 patients that, although at the colonoscopy showed many polyps, they did not present any mutations of *APC* and *MutYH* genes (defined here unresolved FAP).

Regarding the peptidomic study, MALDI-TOF analysis was performed on mutated and unresolved FAP patients. These data were compared with the one from adenoma patients, CRC patients and healthy control subjects. The peptide fingerprint of mutated FAP patients was obtained after performing statistical analysis. A subset of 45 ionic species was found differently expressed in the four groups considered, 12 of them peculiar of FAP patients. Four ionic species were found significantly different in the switch between adenoma and malignant carcinoma. In this study, the potentially prognostic peptides identified derive mainly from circulating proteins and some of them are involved in the inflammatory response. In particular, proteins such as Complement C3 and C4 are known to be cleaved by exoproteases that seem pathology-related.

In the case of unresolved FAP patients, in order to better define a specific pattern, the data from MALDI-TOF were combined with whole exome sequencing. The peptidomics data clearly mark a substantial difference between mutated and unresolved FAP patients. Indeed, unresolved FAP patients have characteristics similar to the control subjects, adenoma patients, CRC patients but not to mutated FAP patients. To understand the possible molecular pathway involved in the unresolved FAP cases, the whole exome sequencing (WES) was performed. From WES data analysis, 285 genes present in all the four unresolved FAP patients were filtered and selected. Among them, the O-linked glycans pathway of the mucins was the most represented.

In conclusion, in this study it was defined for the first time a specific panel of peptides for mutated FAP patients, that could be useful to monitor and predict the pathological evolution of adenocarcinoma malignancy. Furthermore, it was possible to characterize

a preliminary genetic variations pattern for unresolved FAP patients, in which mucin genes might represent the key of the molecular pathway involved.

However, further study are necessary to relate the identified mucin gene variations to their possible causative role in the polyposis. Future analysis of this pattern will be helpful, indeed, to better understand the interatome (the biological network that includes the whole set of direct and indirect molecular interactions in a cell) of these unresolved FAP patients.

RIASSUNTO

La poliposi adenomatosa familiare (FAP) è una delle più importanti forme cliniche di cancro colo-rettale ereditario ed è caratterizzata dallo sviluppo di centinaia/migliaia di polipi adenomatosi nel colon e nel retto durante la seconda decade di vita. La FAP è causata da una mutazione germinale del gene *APC* o da varianti bialleliche del gene *MutYH*. Quasi tutti i pazienti FAP sviluppano il cancro se la patologia non viene precocemente identificata e trattata chirurgicamente.

Lo scopo di questo lavoro è stato caratterizzare 4 pazienti in cui, nonostante lôesame colonscopico presentasse una poliposi conclamata, non risultavano mutazioni nei gene *APC* e *MutYH* (in questa tesi definiti pazienti FAP irrisolti) utilizzando un approccio integrato di peptidomica e genomica.

Riguardo la peptidomica, il MALDI-TOF è stato utilizzato per studiare il profilo peptidico plasmatico di pazienti FAP mutati ed irrisolti comparando i dati ottenuti con quelli derivanti dallo studio di pazienti con adenoma, cancro colo-rettale e soggetti sani di controllo. Dopo analisi statistica è stato ottenuto il *fingerprint* peptidico dei pazienti FAP mutati. Sono state ottenute 45 specie ioniche differentemente espresse nei quattro gruppi considerati, 12 delle quali peculiari per i pazienti FAP. Lôintensità di segnale di quattro di queste specie ioniche è stata trovata statisticamente alterata nello *switch* tra adenoma e carcinoma maligno. I peptidi potenzialmente prognostici identificati in questo studio derivano principalmente da proteine circolanti, alcune delle quali implicate nella risposta infiammatoria. In particolare è noto dalla letteratura che proteine del sistema del complemento come C3 e C4 vengono tagliate da esoproteasi che sembrano essere patologia correlate.

Riguardo ai pazienti FAP irrisolti, per definirne un *pattern* specifico, i dati derivanti dallôanalisi con il MALDI-TOF sono stati combinati con quelli ottenuti dal sequenziamento dellôesoma. I dati di peptidomica hanno chiaramente evidenziato le differenze tra pazienti FAP mutati e FAP irrisolti. Infatti i pazienti FAP irrisolti presentano caratteristiche simili a quelle dei soggetti di controllo, dei pazienti con adenoma e cancro colo rettale ma non a quelle dei pazienti FAP mutati. Allo scopo di capire la via di trasduzione del segnale implicata, è stato quindi eseguito il sequenziamento dellôesoma dei pazienti FAP irrisolti. Da questa analisi sono stati selezionati 285 geni variati in tut-

ti i pazienti e tra questi la via di trasduzione del segnale della O-glicosilazione delle mucine è risultata la più rappresentata.

In conclusione, in questo studio è stato definito per la prima volta un *set* peptidico specifico per i pazienti FAP mutati che potrebbe essere utilizzato per monitorare e predire lôevoluzione patologica della malattia. Inoltre è stato possibile caratterizzare un *pattern* preliminare per i pazienti FAP irrisolti in cui i geni delle mucine potrebbero rappresentare la chiave della via di trasduzione del segnale implicata. Ulteriori studi saranno necessari per correlare i geni delle mucine con la poliposi e costruire lôinteratoma (*network* biologico definito come lôinsieme di tutte le interazioni molecolari dirette e indirette che ci sono allôinterno di una cellula e di un organismo) di questi pazienti FAP irrisolti.

1. INTRODUCTION

1.1. LARGE INTESTINE ANATOMY AND HISTOLOGY

The large intestine is the terminal portion of the gastrointestinal tract (GI). The functions of the large intestine are: the completion of absorption, the production of vitamins, the formation of feces, and the expulsion of feces from the body.

The large intestine is about 1.5 m long and 6.5 cm in diameter and it extends from the ileum to the anus. Structurally, the four major regions of the large intestine are: the cecum, colon, rectum, and anal canal (Figure 1, panel a). The opening from the ileum into the large intestine is guarded by a fold of mucous membrane called the ileocecal sphincter (valve), which allows materials from the small intestine to pass into the large intestine. Hanging inferior to the ileocecal valve is the cecum, a small pouch about 6 cm long. Attached to the cecum is a twisted, coiled tube, measuring about 8 cm in length, called the appendix. The open end of the cecum merges with a long tube called the colon, which is divided into ascending, transverse, descending, and sigmoid portions. Both the ascending and descending colon are retroperitoneal; whereas the transverse and sigmoid colon are not. The ascending colon ascends on the right side of the abdomen, reaches the inferior surface of the liver, and turns abruptly to the left to form the right colic (hepatic) flexure. The colon continues across the abdomen to the left side as the transverse colon. It curves beneath the inferior end of the spleen on the left side as the left colic (splenic) flexure and passes inferiorly to the level of the iliac crest as the descending colon. The sigmoid colon begins near the left iliac crest, projects medially to the midline, and terminates as the rectum at about the level of the third sacral vertebra.

The rectum, the last 20 cm of the GI tract, lies anterior to the sacrum and coccyx. The terminal 2ï 3 cm of the rectum is called the anal canal (Figure 1 panel b). The mucous membrane of the anal canal is arranged in longitudinal folds called anal columns that contain a network of arteries and veins. The opening of the anal canal to the exterior, called the anus, is guarded by an internal anal sphincter of smooth muscle (involuntary) and an external anal sphincter of skeletal muscle (voluntary). Normally these sphincters keep the anus closed except during the elimination of feces.

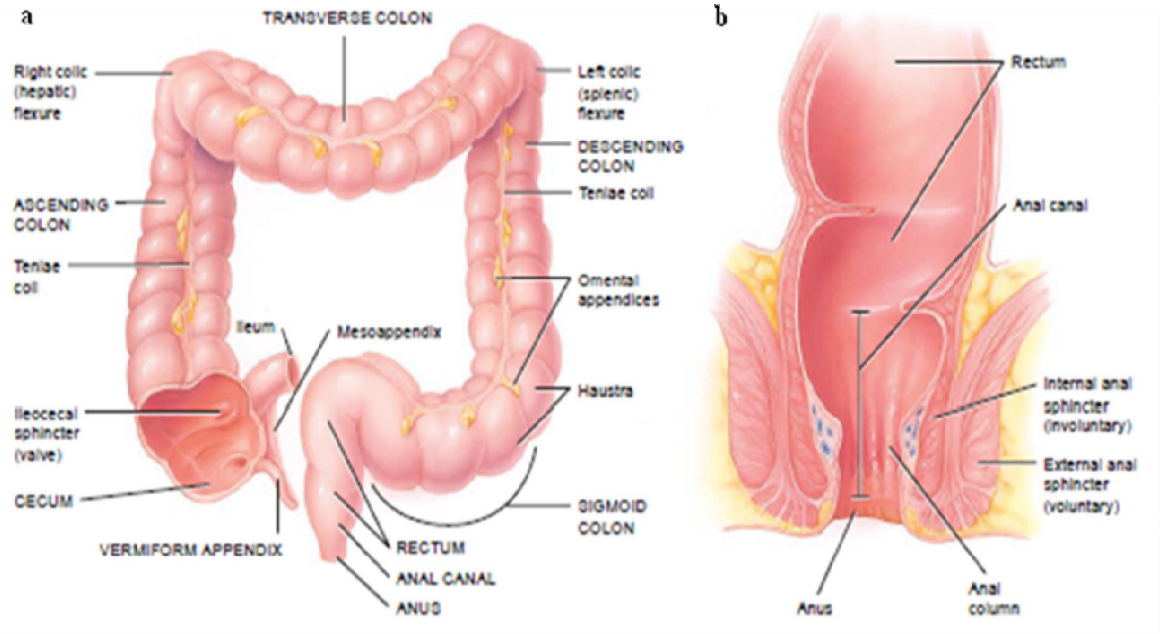


Figure 1: Anatomy of the large intestine. a: anterior view of the large intestine showing major regions; b: frontal section of anal canal. (Tortora and Derrickson, 2009)

The wall of the large intestine contains the typical four layers found in the rest of the GI tract: mucosa, submucosa, muscularis, and serosa. The mucosa consists of simple columnar epithelium, lamina propria, and muscularis mucosae (smooth muscle) (Figure 2 panel a). The epithelium contains mostly absorptive and goblet cells (Figure 2 panel b and c). The absorptive cells function primarily in water absorption; the goblet cells secrete mucus that lubricates the passage of the colonic contents. The major component of the mucus are the mucins. Mucins are filamentous high-molecular weight O-glycosylated glycoproteins found mainly in mucous secretions, but present also in the cell surface acting as transmembrane glycoproteins with the glycan exposed to the external environment. The mucins in the mucous secretions can be large and polymeric (gel-forming mucins) or smaller and monomeric (soluble mucins). Mucins' key characteristic is their ability to form gels; therefore they represent a key component in most of the gel-like secretions, having different functions such as the lubrication, the cell signaling and the formation of chemical barriers. Due to this function, mucins are present in many epithelial surfaces of the body, including the gastrointestinal, genitourinary, and respiratory tracts, where they shield the epithelial surfaces from physical and chemical damage and protect against pathogens infection.

Mature mucins present two distinct regions: the amino- and carboxy-terminal regions and the large central region rich in serine, threonine and proline residues. The first is

lightly glycosylated, but presents a high cysteines content, that causes the disulfide bonds between mucin monomers. The second one, the central region, is also called the variable number of tandem repeat region and can be highly O-glycosylated; about hundreds of O-GalNAc glycans are need to reach the saturation of this region. The protective role of mucins is achieved by the presence of oxygen bond that attach the long sugar side-chains to the tandem repeat regions, avoiding the mucin degradation by the inhibition of protease activity and preserving the viscosity and density of the mucus (Molaei et al., 2010). The O-linked carbohydrates account for up to 80% of the molecular weight of the mucins (Sheng et al., 2012).

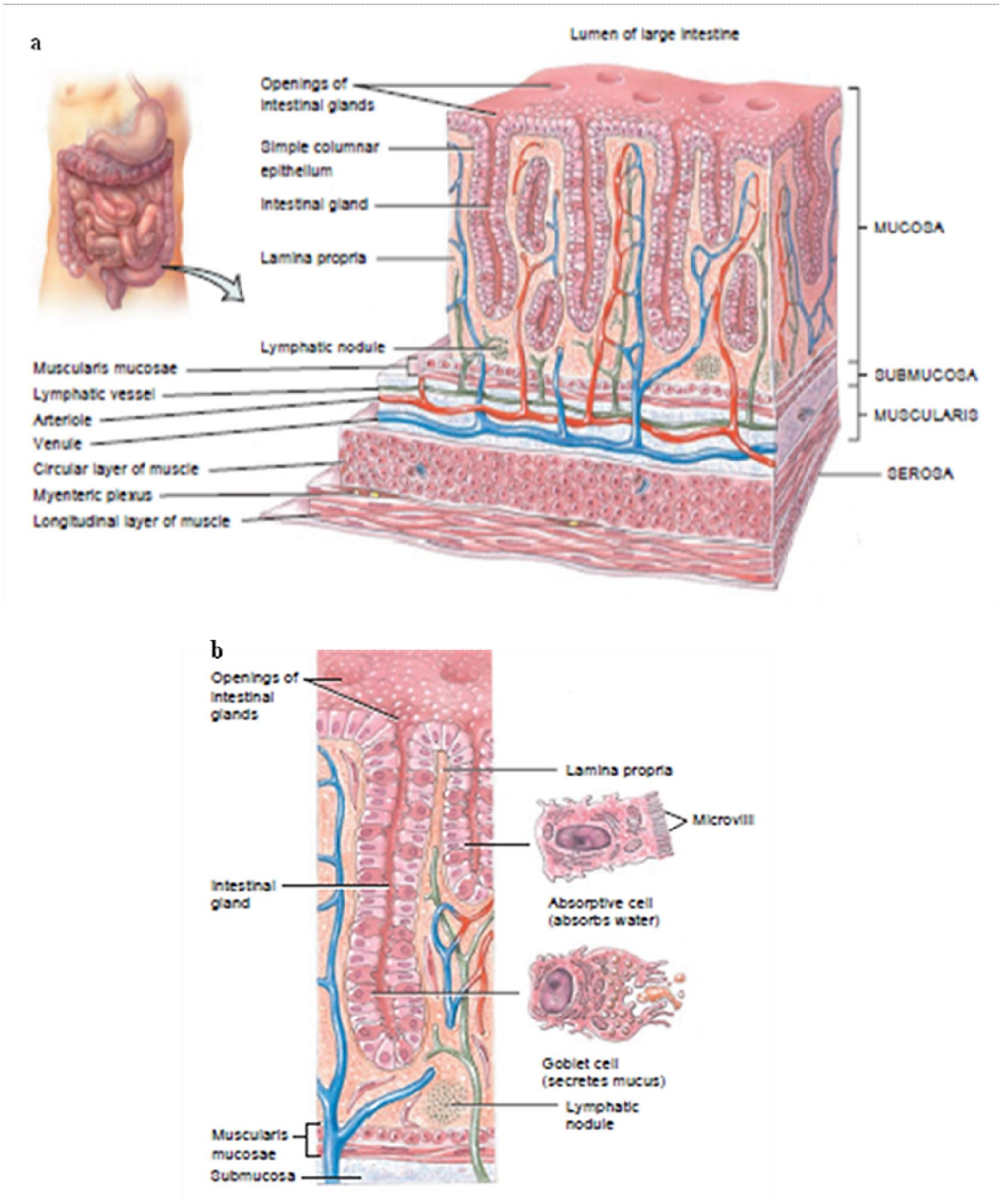
Up to now, 21 mucin genes have been described in literature. Among them, 15 are expressed in different regions of the gastrointestinal tract (Sheng et al., 2012) and are distinguished into secreted and cell surface mucins. In particular, among the 6 secreted mucins: five are oligomerizing secreted mucins (i.e. MUC2, MUC5AC, MUC5B, MUC6, MUC19) and one is a non-oligomerizing secreted mucin (i.e. MUC7). Except MUC19, the secreted mucins share a common evolutionary ancestor and are situated as a cluster on the chromosome 11p15.5. MUC2 is the predominant mucin produced by the intestinal goblet cells, whereas MUC5B is expressed in lower quantities by a subset of the goblet cells residing at the bottom of the colonic cryptic. There are nine genes encoding the major cell surface mucins in the intestine: *MUC1, MUC3A, MUC3B, MUC4, MUC12, MUC13, MUC15, MUC16* and *MUC17*. Intestinal cell surface mucins are highly expressed in enterocytes.

The mucin genes, their transcripts, the resulting mucin proteins and the attached O-GalNAc glycans show all an extreme variability. Different mucins show changes in the number and composition of the peptide repeats in their VNTR regions, whereas in the same mucin the repeats usually vary in their amino acid sequences.

Both absorptive and goblet cells are located in long, straight, tubular intestinal glands (crypts of Lieberkühn). Solitary lymphatic nodules are also found in the lamina propria of the mucosa and may extend through the muscularis mucosae into the submucosa. Compared to the small intestine, the mucosa of the large intestine does not have many structural adaptations that increase surface area.

The submucosa of the large intestine consists of areolar connective tissue. The muscularis consists of an external layer of longitudinal smooth muscle and an internal layer of circular smooth muscle. Unlike other parts of the GI tract, portions of the longitudinal muscles are thickened, forming three conspicuous bands called the teniae coli

that cover most of the length of the large intestine (Figure 2 panel a). The teniae coli are separated by portions of the wall with less or no longitudinal muscle. Tonic contractions of the bands gather the colon into a series of pouches called haustra which give the colon a puckered appearance. A single layer of circular smooth muscle lies between teniae coli. The serosa of the large intestine is part of the visceral peritoneum. Small pouches of visceral peritoneum filled with fat are attached to teniae coli and are called omental (fatty) appendices (Tortora and Derrickson, 2009).



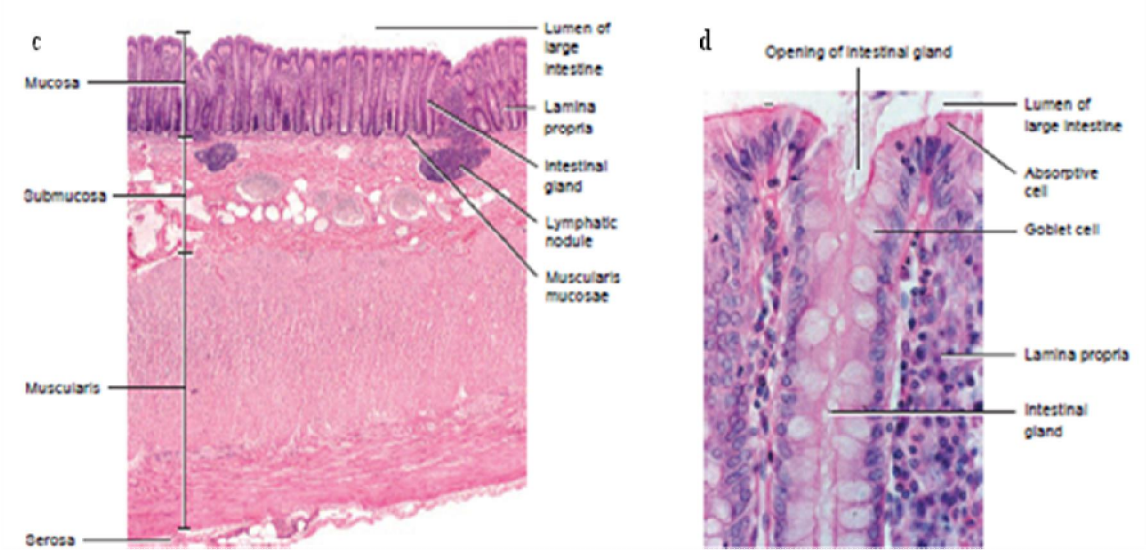


Figure 2: Histology of the large intestine. a: three-dimensional view of layer of the large intestine; b: sectional view of intestinal glands and cell types; c: portion of the wall of the large intestine; d: details of mucosa of large intestine (Tortora and Derrickson, 2009).

1.2. COLORECTAL CANCER (CRC)

1.2.1. Epidemiology of CRC

Incidence and mortality rates of CRC vary markedly around the world. It has been estimated that 93,090 cases of colon cancer and 39,610 cases of rectal cancer will be diagnosed in 2015 (American Cancer Society: Cancer Facts and Figures 2015. Atlanta, Ga: American Cancer Society, 2015). CRC is the third most common cancer in both men and women. However, the incidence rates have been decreased during the past two decades due to changes in the risk factors and to the preventive screening among adults 50 years and older. CRC screening tests allow, indeed, the detection and the removal of colorectal polyps before they progress to cancer. From 2007 to 2011, the incidence rates declined by 4.3% per year among adults 50 years of age and older, but increased by 1.8% per year among adults younger than age 50. An estimated 49,700 deaths from colorectal cancer are expected to occur in 2015. CRC is the third leading cause of cancer death in both men and women and the second leading cause of cancer death when men and women are combined. From 2007 to 2011, the overall death rate declined by 2.5% per year. This trend reflects the declining incidence rates and the improvements in the early detection and treatment.

1.2.2. Risk and protective factors

Early CRC stages typically does not have symptoms, that is why preventive screening is usually necessary to detect this cancer early. Symptoms may include rectal bleeding, blood in the stools, a change in bowel habits or stools shape (e.g., narrower than usual), the feeling that the bowel is not completely empty, cramping pain in the lower abdomen, decreased appetite, or weight loss. In some cases, blood loss from the cancer leads to anemia, causing symptoms such as weakness and excessive fatigue. Timely evaluation of symptoms consistent with CRC is essential, even for adults younger than age 50, among whom CRC is rare, but increasing.

The risk of CRC increases with age; in 2011, 90% of cases were diagnosed in individuals 50 years of age and older. Modifiable factors associated with increased risk include obesity; physical inactivity; moderate to heavy alcohol consumption; long-term smoking; high consumption of red or processed meat; low calcium intake; and very low intake of whole-grain fiber, fruit, and vegetables.

Hereditary and medical factors that increase the risk include a personal or family history of CRC and/or polyps, a personal history of chronic inflammatory bowel disease (e.g., ulcerative colitis or Crohn disease), certain inherited genetic conditions (e.g., Lynch syndrome, also known as Hereditary Nonpolyposis Colorectal Cancer (HNPCC), type 2 diabete and Familial Adenomatous Polyposis (FAP)).

A large number of factors have been reported to be associated with a decreased risk of CRC. These include regular physical activity, a variety of dietary factors, the regular use of aspirin or nonsteroidal anti-inflammatory drugs (NSAIDs), and hormone replacement therapy in postmenopausal women. Regular use of aspirin and other NSAIDs are associated with a 20 to 40 percent reduction in the risk of colonic adenomas and CRC in individuals at average risk. How the aspirin produce this protective factor is not well understood yet. Proposed explanations include the increase and impairment of tumour cell growth by inhibition of cyclooxygenase-2 (Wang et al., 2012).

1.2.3. Colorectal cancer progression

In 1990, Fearon and Vogelstein (Fearon and Vogelstein, 1990) presented evidences for a multistep genetic model of the CRC formation. This model is based on the understanding that CRC is the result of mutations in some key genes, such as the inactivation

of tumour suppressor genes (TSG) and the activation of oncogenes. Furthermore, the accumulation of mutations occurs in a sequential manner, with mutations of some genes preceding the one of the others. This genetic paradigm is shown in Figure 3. As seen, there are two different pathways that lead to the formation of CRC. One way is through inactivation of the TSG, *APC* gene (Adenomatous Polyposis Coli) that accounts for approximately 85% of all CRC and it is germline mutated in patients with Familial Adenomatous Polyposis (FAP). The other CRC development pathway is through the mutational inactivation of a protein family involved in DNA mismatch repair (MMR), including *MLH1*, *MSH2*, and *PMS2* genes. MMR genes mutation is found in approximately 15% of all sporadic CRC and is responsible for the HNPCC syndrome if present as germline mutation.

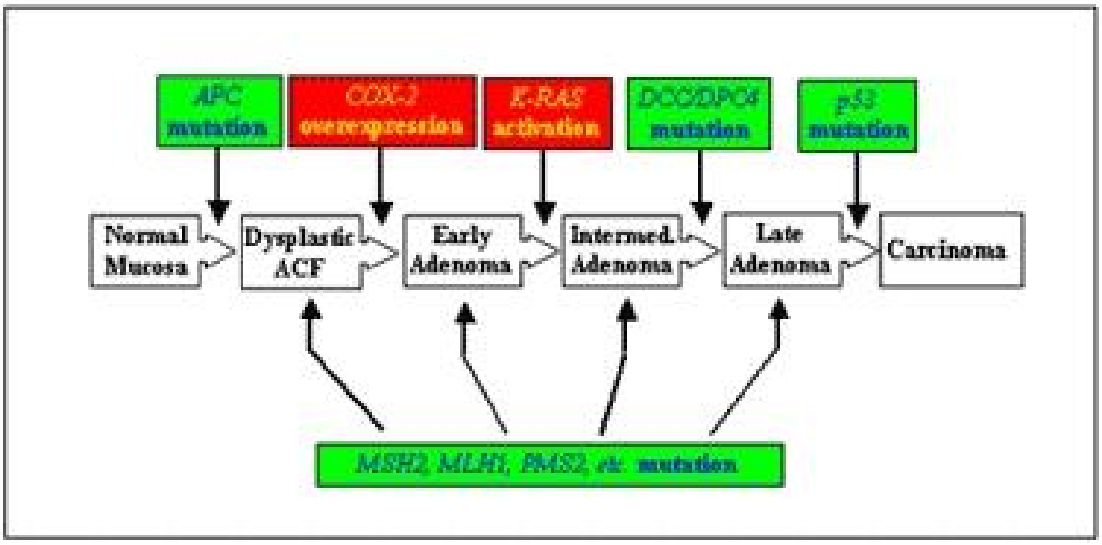


Figure 3: The genetic paradigm of colorectal cancer. The formation of CRC requires the sequential mutation of several genes. This includes the inactivation of TSGs (in green) and activation of oncogenes (in red) (Hisamuddin and Yang, 2004).

1.2. HEREDITARY FORMS OF CRC

Among the various causes of CRC, approximately 75% can be attributed to sporadic disease, where there is no apparent predisposing aetiology. The remaining cases of CRC are accounted for familial incidences (15%) and inflammatory bowel disease (IBD, 1%) (Figure 4). Although some familial cases consist of well-described hereditary CRC syndromes, including FAP (1%), HNPCC (5%), and the hereditary hamar-

tomatous polyposis syndromes (<1%), the majority of familial disease has no clearly identifiable genetic aetiology (Hisamuddin and Yang, 2004).

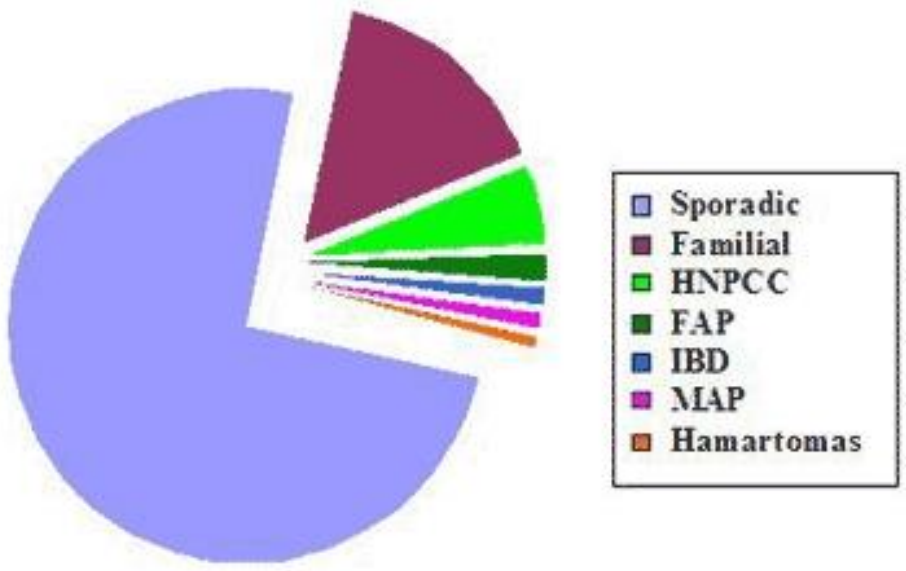


Figure 4: The distribution of CRC conditions in the population.

1.3.1. Familial Adenomatous Polyposis (FAP)

FAP is an autosomal dominant disease that is classically characterized by the development of hundreds to thousands of adenomas in the colon and rectum during the second decade of life (Half et al., 2009). Almost all patients will develop cancer if they are not identified and treated at an early stage. Clinically FAP arises equally in both sexes (Half et al., 2009). About half of the FAP patients develops adenomas at 15 years of age and 95% at 35 years of age (Petersen et al., 1991). Generally cancer starts to develop a decade after the appearance of the polyps. If the colon is intact the majority of the FAP patients develops CRC at the ages of 40-50 years. However, although uncommon, CRC can develop in children or in older people.

Individuals with FAP can also develop a variety of extra-colonic gastrointestinal manifestations: fundic gland polyps of stomach; adenomatous polyps in duodenum and periampullary region and small bowel adenomas.

Extra-intestinal manifestations are rarely malignant and include: cutaneous lesions, osteomas and dental abnormalities. The phenotypic variant, characterized by this extra-intestinal manifestation, is called the Gardner Syndrome. Other extra-colonic malig-

nancy associated with FAP disease are: congenital hypertrophy of the retinal pigment epithelium, desmoids, pancreatic mucinous adenocarcinomas, liver, brain and thyroid tumors.

A less aggressive variant of FAP is the Attenuated Familial Adenomatous Polyposis (A-FAP); this variant is characterized by fewer colorectal adenomatous polyps (usually 10 to 100), that in the later age evolves in adenoma (the average age of polyps diagnosis is 44 years old) and cancer (at the age of 56 years old) (Half et al., 2009).

Most of the FAP patients have a familiar history of colorectal polyps and cancer, however 25-30% of them develop *de novo* without clinical or genetic evidence of FAP in family members (Bisgaard et al., 1994).

FAP is a genetic disorder resulting from a mutation in the *APC* gene. The first evidence of the FAP gene location was supported by Herrera and colleagues, that demonstrated the presence of a constitutional deletion of chromosome 5q21 in a patient with Gardner Syndrome (Herrera et al., 1986). The following linkage analysis showed that 5q21 chromosome markers were related to the development of FAP (Bodmer et al., 1987). The gene responsible for FAP was identified in 1991 and was called *APC* gene (Kinzler et al., 1991). In adenomas and carcinomas developed in FAP patients, there is an evidence of the inactivation of the second copy of the *APC* gene (Levy et al., 1994).

1.3.2. APC gene and protein function

The *APC* gene is a tumor suppressor gene located on the long arm of the chromosome 5 in the band q21. It contains 17 exons (15 coding exons) and it has a transcript length of 10,619 base pairs (Ensembl). *APC* gene encodes a large polypeptide of 2843 amino acids with a molecular weight of 311646 Da (Uniprot).

The APC protein has multiple domains that mediate oligomerization as well as binding to a variety of intracellular protein and it is present in a variety of epithelial tissues, often in post-mitotic cells (Midgley et al., 1997). APC is a crucial member of the Wnt/β-catenin signaling pathway, that is an important determinant of cell proliferation, differentiation and apoptosis. Studies indicate that APC is involved in the regulation of a myriad of cellular functions including proliferation, differentiation, apoptosis, adhesion, migration and chromosomal segregation (van Es et al., 2001). Furthermore, APC

regulates cytoskeletal proteins including F-actin and microtubules, controlling cell adhesion, migration and mitosis.

The Wnt/ß-catenin signaling pathway is also essential in controlling intestinal epithelial cell proliferation (Reya and Clevers, 2005). A model for Wnt signaling is depicted in Figure 5. Wnt, a secreted glycoprotein, interacts with two cell surface receptors, Frizzled (Fz) and the low-density lipoprotein receptor-related protein (LRP) (Bhanot et al., 1996). In the absence of Wnt or in the presence of wild-type APC, ß-catenin is sequestered in the cytoplasm in a complex that includes APC, axin, and glycogen synthase kinase-3ß, where it is subjected to ubiquitin-mediated degradation (Reya and Clevers, 2005). In the presence of Wnt, or absence of APC (as occurs in many colon cancer), ß-catenin is released from the complex. Free ß-catenin is shuttled into the nucleus, where it binds the T-cell factor (TCF) and releases its repressors, CtBP and Groucho, activating the gene transcription (Behrens et al., 1996). Among the target genes stimulated by ß-catenin/TCF complex, there are c-Myc and cyclin D, both essential for the progression of the cell cycle during proliferation (He et al., 1998). The importance of Wnt/ß-catenin pathway in the pathogenesis of CRC is further demonstrated by the observation that in the absence of *APC* mutation, CRC sometimes contains inactivating mutation of axin (Jin et al., 2003) (Webster et al., 2000) or activating mutation of ß-catenin (Sparks et al., 1998).

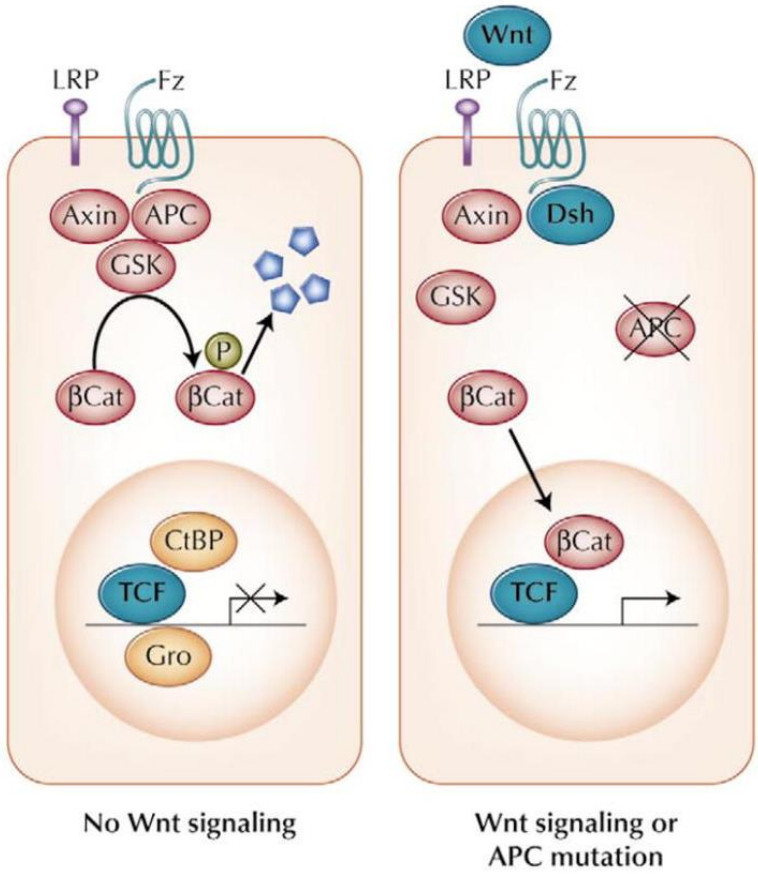


Figure 5: The Wnt/ß-catenin signaling pathway. ßCat ß-catenin; CtBPð C-terminal binding protein; Dshð Disheveled; Fzð Frizzled; Groð Groucho; GSKð Glycogen synthase kinase 3ß; LRPð Low-density lipoprotein receptor-related protein; TCFð T-cell factor (Hisamuddin and Yang, 2004).

Some studies indicate that the carboxyl terminal portion of APC binds cytoskeletal proteins. As a microtubule-associated protein, APC contributes to mitotic spindle formation and function. Cells lacking APC are prone to defects during chromosomal segregation (Fodde et al., 2001).

1.3.3. *APC* mutations

Mutations in the *APC* gene are present in up to 85% of FAP patients and in over 80% of sporadic CRC cases. It has been shown that loss of heterozygosity (LOH) is the main mechanism by which *APC* becomes inactivated. More than 300 mutations are recognized as cause of FAP. Most of these mutations result in the truncated protein. More than 60% of *APC* mutations are found in the central region of the gene, that is known as the mutation cluster region (MCR) and results in the expression of carboxyl-terminally truncated proteins. The MCR region coincides with the APC region that is important for the down-regulation of ß-catenin. *APC* mutations in the first or last third

of the gene are associated with an attenuated polyposis with a late onset and a small number of polyps. Non neoplastic cells of FAP patients are expected to retain normal APC function due to the presence of one wild-type allele, regardless of the position of the mutation in the affected allele.

1.3.4. *MutYH*-associated Polyposis (MAP)

Although the majority of FAP syndromes is attributed to germline mutations in the *APC* gene, a small number of familial cases of polyposis do not have any *APC* mutations but shows germline mutations in the *MutYH* gene. *MutYH* mutation causes the attenuated polyposis condition known as *MutYH*-associated Polyposis (MAP). It is recessively inherited and patients have either a homozygous or a compound heterozygous germline mutations of the *MutYH* gene.

The *MutYH* gene contains 16 exons (ensemble database) and it maps to chromosome 1p32-34. It encodes a protein of 546 amino acid that has a molecular weight of 60069 Da (uniprot database). This protein is a DNA glycosylase enzyme involved in the oxidative DNA damage repair. The enzyme is involved in the base excision repair (BER) process and excises adenine bases from the DNA backbone at the sites where the adenine was inappropriately paired with guanine, cytosine, or 8-oxo-7, 8-dihydroguanine (8-oxoG) during the DNA replication (Takao et al., 1999).

1.3.5. Genetic tests

Nowadays, a number of genetic tests are available to test for *APC* germline mutations. Currently, the most commonly used is the direct sequencing of the *APC* gene. The mutation detection rate, when the full gene sequencing is performed, is 70%. However, the large insertions and deletions, having a frequency of about 5%, need other assays (e.g. Multiplex Ligation-dependent Probe Amplification or MLPA). *MUTYH* mutations are responsible of the remaining cases. When the family's specific *APC* mutation is identified, genetic testing of all the first degree relatives should be performed. Parents of children at-risk should be informed that the genetic testing is recommended just before puberty or preferably in mid-adolescence. Familiar members that result negative

for a known mutation do not need further investigation or follow-up other than standard average-risk screening of the population.
Genetic testing for *MUTYH* mutation has been recommended for all patients who have tens to hundreds of colorectal adenomas with no identified germline mutation in the *APC* gene and with a family history compatible with an autosomal recessive mode of inheritance.

Nowadays, several approaches are available to study the cancer disease in the clinical field as the proteomic analysis and Whole Exome Sequencing (WES). The following two paragraphs will elucidate the use of these approaches in the clinical field.

1.4. PROTEOME STUDY

The proteome is the entire set of proteins produced by an organism at certain time or under defined conditions. The clinical proteomics focuses on the characterization of unique patterns of protein expression, or biomarkers, associated with specific diseases.
Matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) is one of the most widely used proteomic technology and it is usually coupled with the time of flight (TOF) ion analyzer. MALDI is an ionization technique widely employed for proteins and peptides analysis. The sample is mixed with a saturated solution of a matrix and deposited onto a plate to allow it to dry. When a laser bean hits the sample there is a production of ions. The ions are sent to the analyzer that separate the ions according to their mass/charge (m/z) ratio on the basis of the time that the ion takes to go from the source to the detector (Figure 6). MALDI is a ñsoftò ionization technique that allows a minimal fragmentation of peptides. Furthermore, it almost exclusively creates single charged ions, which mean that the m/z value represents a realistic measurement of the actual mass of the particles.

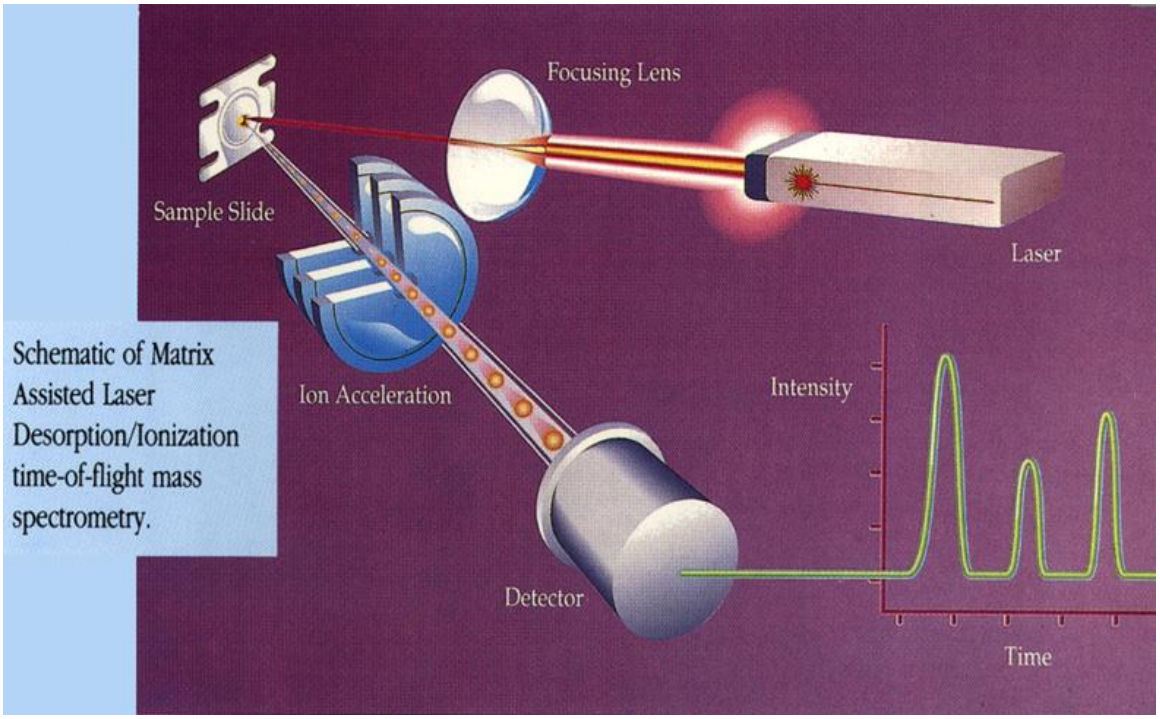


Figure 6: MALDI-TOF schematic representation. (Montana State University)

In accordance with the technologies progress, an increased interest has been arisen for the dynamic nature of sub-proteome presented in the blood. In particular, circulating low molecular weight (LMW) peptidome seems to be a source of disease-specific biomarkers. Indeed, peptides show better permeability between tissues and cell membranes than the corresponding full-length proteins. Furthermore, tumor specific proteolytic cascades within the tissue can generate fragments that diffuse into the circulatory system. Proteases are proteins that cleave amino acid sequences at precise positions and they participate both in the physiological and tumor processes. Tumor specific peptides are derived not only from a population of cancer cells but also from the microenvironment tissue surrounding the tumor cells. It also was proposed that some LMW peptide might be generated in the blood vessel by degradative proteases that are already present in the blood. (Figure 7) (Petricoin et al., 2006).

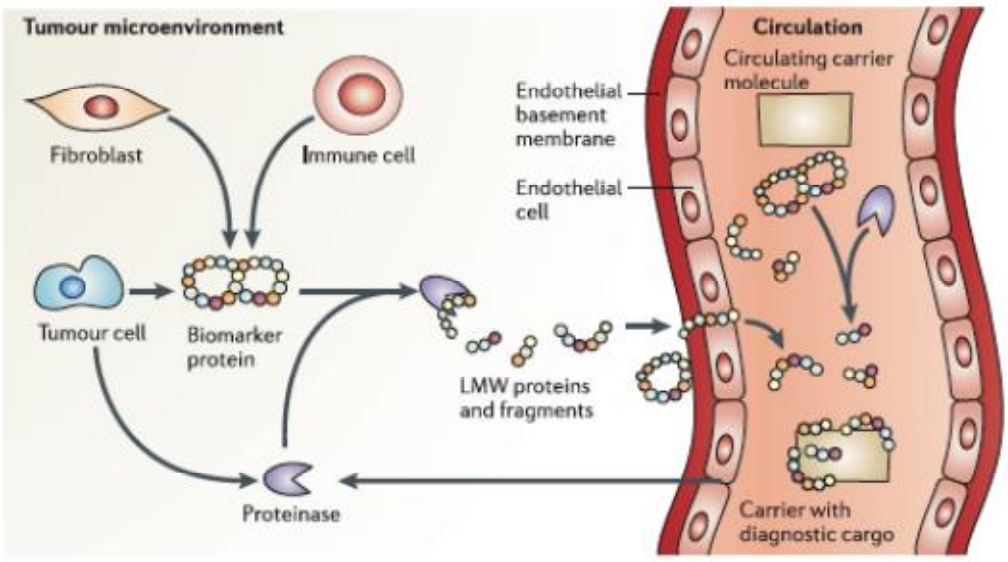


Figure 7: Different cascades of peptide generation and release into the circulatory system. Circulating peptides could be shed and generate from all the cell types present in the tissue microenvironment. (Petricoin et al., 2006).

1.5. WHOLE EXOME SEQUENCING (WES)

Human genome comprises more or less 3×10^9 bases having coding and non-coding sequences. About 3×10^7 base pairs (1%, 30 Mb) of the genome are the coding sequences. The genome assembly from the Genome Reference Consortium (GRCh37.p10, Feb 2009) in Ensembl Genome Browser includes about 20000 protein-coding genes, pseudogenes and non-coding genes (Rabbani et al., 2014). It is estimated that 85% of the disease-causing mutations are located in the coding and functional regions of the genome. For this reason, the complete sequencing of the coding regions (exome) has the potential to find the causes of a large number of rare genetic disorders (mostly monogenic) and as well as the predisposing variants present in common diseases and cancers.

Monogenic disease (also called Mendelian disease) are caused by a single gene variant conforms to the Mendelian inheritance laws. Over 6000 Mendelian diseases have been identified, however, the pathogenic genes of about half of these diseases remain unknown. Whole-exome sequencing (WES) can be used to perform NGS of exon-enriched samples and to identify protein-coding mutations, including missense, nonsense, splice site, and small deletion/insertion mutations. The application of WES definitely represents a revolutionary progress in Mendelian disease research (Zhang, 2014). Indeed, one of the major advantage in the use of WES in Mendelian disorders is

that it requires a low amount of clinical cases, indeed only a small number of individuals for each family are necessary for the analyses.

However, there is also an increased interest in leveraging the power of exome sequencing in diseases that do not exhibit a Mendelian mode of transmission (disease caused by non inherited or *de novo* mutations) and disease with a complex genetic component. WES technique could definitely help, at the clinical point of view, to understand the genetic mechanisms of the cancer formations and to find successful treatments. As mentioned in the literature, the first cancer exomes were sequenced soon after the completion of the Human Genome Project in 2001. Ley *et al.* used NGS to study the exomes of human acute myeloid leukemia (AML) cells in 2003 (Ley et al., 2003), whereas the first investigated solid tumor exomes derived from breast and colorectal cancer tissue samples (Sjoblom et al., 2006).

Although a lot of progress has been made, the currently available NGS technologies, such as HiSeqÊ (Illumina), SOLiD4Ê (Life TechnologiesÊ), and 454 GS FLX (Roche), generate hundreds of millions of short sequence reads with an average length in the range of 50 bp to 125 bp. This technique, as consequence, presents a sequencing error rate higher than the Sanger sequencing; thus, further validation using Sanger sequencing is essential (Zhang, 2014).

The three main steps of exome sequencing are:

1) Exome enrichment;
2) DNA high-throughput sequencing;
3) Biological interpretation.

Exome enrichment is the base of the exome sequencing. The genomic DNA is randomly sheared and used to form a shotgun library. The exome regions are enriched by hybridization capture (using oligonucleotide probes to hybridize the fragments of interest). The noncoding DNA sequences are then removed. After the sample preparation and sequencing, 20000 and 50000 variants are identified per sequenced exome. The disease-causing variants should be represented by the non-synonymous variants, splice acceptor-site or donor-site mutations, and insertions/deletions. Candidate variants are absent from the most databases of the known variant, such as dbSNP, and 1000 Genomes Project. Further specific candidate variants can be predicted by ANNOVAR or SIFT software tools. The predicted damaging variants can be sequenced by the traditional Sanger sequencing to confirm the real causative variant.

As in all the approaches, also WES analysis presents some limitations: it is known that approximately 85% of variants are in the coding exons, meaning that 15% are outside the exons, in the noncoding, conserved, or regulatory regions. Furthermore, some studies have revealed that 80% to 90% of the targeted regions are covered by more than 10X, but 4-8 Mb (or 1000 to 2000 genes) have no sufficient coverage for variant detection.

2. MATHERIALS AND METHODS

2.1. MUTATED and UNRESOLVED FAP PEPTIDOME

2.1.1. Patients selection and plasma preparation

All the sample were obtained from the Tissue Biobank at the First Clinical Surgery Section of the Department of Surgical, Oncological and Gastroenterological Science. Briefly, the plasma samples were collected from 13 mutated FAP patients (APC or MutYH mutated, see Table 4), 4 unresolved FAP patients (Table 7), 26 adenoma patients, 58 sporadic CRC patients and 38 control subjects resulted negative at the colonoscopy. A complete clinical history and written informed consent was obtained from each patient (Ethical Committee approval number 448).

Blood samples were collected in DB Vacutainer® Blood Collection Tubes (Becton Dickinson and Company, Franklin Lake, NJ, USA) containing K_3EDTA. Plasma was obtained after centrifugation for 10 mins at 3000 rpm and the aliquots were stored at -80°C until further analysis, following the Biobankôs protocol.

2.1.2. Sample preparation

Plasma samples (200 microliters) were diluted 1:4 in deionized water and 500 µl of the solution were centrifuged for 20 mins at 3000 g in Amicon Ultra-4 Centrifugal Filter Devices with 30 KDa MWCO (molecular weight cut-off) (Millipore, Merck KGaA, Darmstadt, Germany). Before MALDI-TOF analysis, the concentrate was discarded from the filter unit sample to eliminate high molecular weight proteins and the eluate was desalted and purified by ZipTip C18 pipette tips (Millipore, Merck KGaA, Darmstadt, Germany) following the procedure described in the userôs guide.

2.1.3. MALDI-TOF analysis

Peptide profile analysis was performed using a Voyager-DE PRO instrument (Applied Biosystems, Foster City, CA, USA), operating in reflectron positive ion mode. Ions, formed by a pulsed UV laser (λ = 337 nm) beam were accelerated to 20 keV. Instrumental set up was: mirror ratio 1.12; grid voltage 72%; extraction delay time 150 nsec;

grid wire 0.05%. External mass calibration was performed by the Calibration Mixture 2 of Sequazyme^{TM} Peptide Mass Standards Kit. Sample deposition was performed using alpha-cyano-4-hydroxycinnamic acid (CHCA, Sigma-Aldrich, St. Louis, MO, USA) in a 50% ACN/0.05% TFA (v/v) as matrix solution.

MALDI-TOF data files were exported as .txt files and processed with mMass open source software (Strohalm et al., 2008). Spectra have been denoised (precision 15, relative offset 25, arbitrary units) and smoothed (Savitzkyï Golay filtering, m/z 0.15) before manual peak picking (threshold set S/N=3).

The peptides were fragmented using an Ultraflex Extreme TOF-TOF instrument (Bruker Daltonics, Bremen, Germany) equipped with an Nd:YAG laser (ə=355nm). Fragmentation spectra were processed using flexAnalysis v3.3 software and a peaks list was generated after baseline subtraction using Snap as peaks detection algorithm. Peptides identification was obtained using Mascot v2.3 (Matrix Science, London, UK) and MS-Tag (UCSF, CA, USA) search engines. Search parameters were set as follow: database, NCBInr.2013.6.17; enzyme, no enzyme; taxonomy, *Homo sapiens*; precursor tolerance, 0.3 Da, fragment tolerance, 0.6 Da; variable modifications: acetyl (N-term); deamidated (N/Q); Gln->pyro-Glu (N-term Q/E) and oxidation (M/W/P).

2.1.4. ELISA assay

ELISA assays specific for C4b and C3b (MyBiosource kits) was used to analyzed in total 10 FAP, 8 Adenoma, 36 CRC patients and 22 health subjects.

The C4b assay kit (MBS700316) is based on the quantitative sandwich enzyme immunoassay technique. Briefly, the microplate was pre-coated with an antibody specific for C4b and the standard and sample were added in each well, so any C4b present was bound to the immobilized antibody. A biotin-coniugated antibody specific for C4b was added into the well, followed by an avidin-coniugated Horseradish Peroxidase (HRP). Before each step a wash was done in order to remove any unbound substance. Then, the staining reaction was blocked and the optical density (OD) was measured spectrophotometrically at ə=450 nm in a microplate reader.

The C3b assay kit (MBS745371) applies the competitive enzyme immunoassay technique using a monoclonal anti C3b antibody and C3b HRP -conjugated. The sample and standards in this case, were incubated together with C3b HRP conjugated in

precoated plate. The optical density was inversely proportional to the C3b concentration since C3b from the sample and C3b HRP-coniugated compete for the anti-C3b antibody binding site.

The standard curve of both assay was created using the ñCurve Expert 1.4ò software and it was plotted relating the intensity of the color (OD) to the concentration of standards. The C4b and C3b concentrations in each sample were interpolated from this standard curve.

2.1.5. Statistical analysis

The statistical analysis for MALDI dataset was performed using MetaboAnalyst 2.0 software. The peak list was normalized by sample median intensity and log transformed. Univariate analysis (one-way ANOVA, *t*-test, fold change, and Volcano plot) and multivariate analysis (Cluster Analysis (CA), Partial Least Square-Discriminant Analysis (PLS-DA) and Random Forest (RF)) were used. The Box plot was performed by GraphPad Prism 5 software.

2.2. WHOLE EXOME SEQUENCING

2.2.1. Genomic DNA extraction

Genomic DNA of 4 unresolved FAP patients was isolated from peripheral blood lymphocytes using QIAamp DNA Mini Kit (Qiagen) following the manufacturesô user guide.

2.2.2. Next Generation Sequencing analysis

The genomic DNA samples of the four unresolved FAP patients was sent to the Laboratory of Prof. Paolo Fortina (Kimmel Cancer Center, Philadelphia) for the analysis.

The whole exome sequencing considers an enrichment step in which are selected only the exome regions. The enrichment step was performed with SureSelect Human All Exon Target Enrichment System (Agilent Technologies). The genomic DNA (3µg) of four unresolved FAP patients was sheared into short fragments of about 300 bp. These

DNA fragments were hybridized into solution with RNA oligonucleotides, to specific target approximately 38 Mb of the human genome (1.22% of the genome) covering ~18,000 genes. Non selected regions were washed off and the selected fragments were eluted to create a library enriched for the protein coding sequences. Adapter sequences were ligated to DNA, before or after enrichment step, to lead the sequencing process (Figure 8).

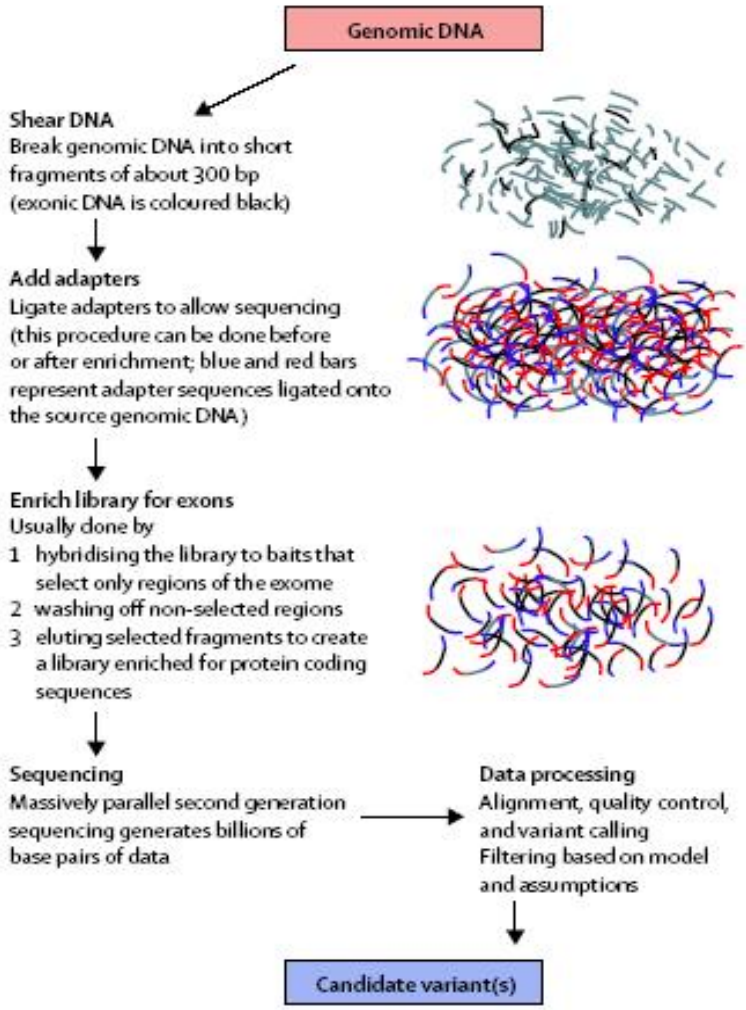


Figure 8: Exons library production.

WES analysis was performed by Sequencing by Oligonucleotide Ligation and Detection (SOLiD 4) instrument (Life Technologies Foster City, CA) with 80X coverage. The SOLiD system is a massively parallel DNA sequencer based on sequencing-by-ligation. The method consist of an emulsion PCR on glass beads. The beads are added in excess so that each bead links only one fragment and only the beads with the complementary adapter oligos are used. Briefly, after hybridization step 500 picomoles of the enriched exome library were used for the emulsion PCR in which the mineral oil

allows the formation of micelle, each of them acts like a single compartment. The single DNA molecules produced on the glass beads are then deposited onto a glass slide (Figure 9).

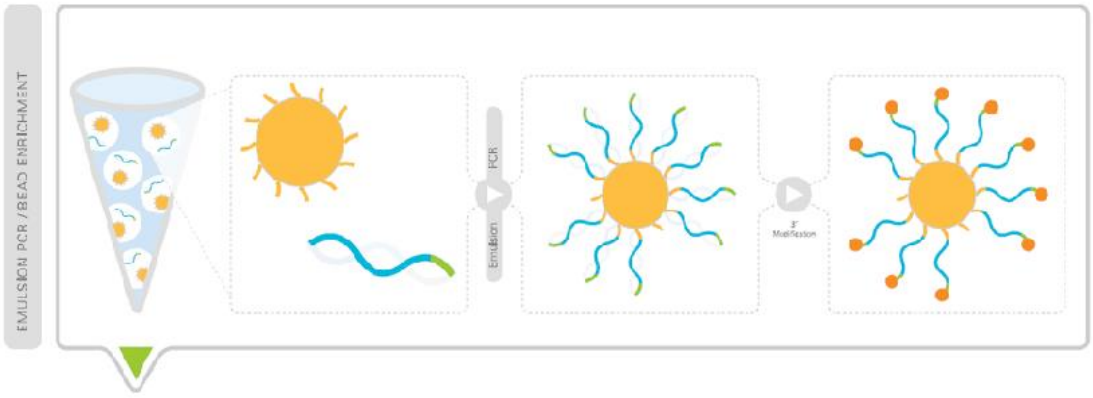


Figure 9: Emulsion PCR done on glass beads.

A universal primer is annealed and the selective hybridization and ligation of complementary probes occurs to establish the sequence. The universal primer recognizes the adapter sequence; DNA octamers (having a fluorescent dimer) bind DNA and are ligated. Afterwards, the last three bases are cut and a new step of hybridization occurs (Figure 10).

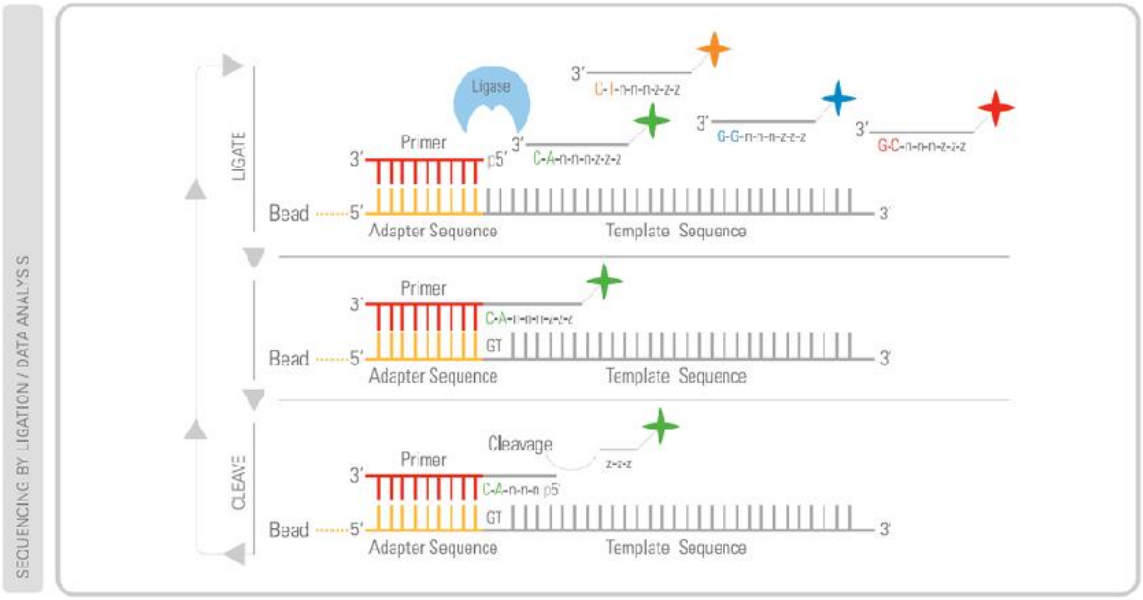


Figure 10: Sequencing-by-ligation.

Paired-end sequencing was performed. In this method, two ends of the same DNA fragment are sequenced in opposite directions which spans an inserted sequence of ~180 bp. The two sequenced fragments are 50 bp (F3 tag) and 35 bp (F5 tag). When

mapped back to the genome, the two-paired sequences should map to the same region and separated by a distance of the inserted fragment.

2.2.3. WES data analysis

The SHRiMP software was used for the data analysis, following three distinct steps. First, the color space reads were mapped to the hg19/GRCh37 Reference Genome (http://genome.ucsc.edu/) using an iterative mapping approach. The sequence coverage was determined as the proportion of targeted regions that was covered by at least one unique aligned read. Bases that aligned to the genome, but not to the targeted regions were not considered for further analysis. Additionally, only regions that had a coverage higher than 8X were considered for further analysis. Over 90% of the targeted exome regions had at least 20X depth of coverage in each sample. The second step of the bioinformatic screening was to identify genomic variants including SNPs and small insertion deletion variants (indels). The third step of the analysis included the interpretation of the data obtained.

2.2.4. First elaboration of the data

Variants were identified using the Genome Analysis Toolkit HaplotypeCaller.
Resulting variants were annotated using ANNOVAR and filtered considering the following criteria (Filter a):

1. Non synonymous SNVs
2. Not included in dbSNPs
3. At least 5% variant allele frequency
4. At least 20X coverage.

2.2.5. Second elaboration of the data

Variants were identified using a custom script.
ANNOVAR software was used to provide genomic and functional annotation of the variants found.

We filtered the variants considering the criteria discussed before (Filter a) and the following more stringent criteria (Filter b):

1. Non synonymous SNVs
2. Not included in dbSNPs
3. At least 5% variant allele frequency
4. At least 50X coverage
5. Statistical significance p-value threshold of 10^{-9}.

2.2.6. PCR and Sanger sequencing

The primers were drawn with Primer 3 plus software and validated with UCSC Genome Bioinformatics database (Table 1 and 2).

For *NBPF16* gene it was not possible to found primers specific for the region of interest showing only one amplification product, therefore they were manually drew. One common reverse (rw) primer and 2 forward (fw) primers were used: one specific for wild-type (wt) allele and one specific for mutated (mut) allele. The 2 forward primers differ for the last base in 3ô end of the sequence (Table 2, in bold). Using this strategy, if a subject is wild-type has the band only with the couple wt-fw + rw. If a subject presents the mutation in homozygous state has the band only with the couple mut-fw + rw, whereas if a subject has the mutation in heterozygous state has the band with both couples of primers. The PCR reactions were performed with GoTaq DNa Polymerase (Promega).

Gene	Protein Variant	Primer Forward (fw)	Primer Reverse (rw)
KMT2C	**Y816_I817delinsX**	TCCCCACATGAGGAAAGTAT	AAAAAGTTGCCATCCACACC
PABPC1	**E156fs**	GAAATAGGAACTGTGCAGTAATGG	CTTAAAATGAATGGATTTGGAATTA
PRSS3	**K152Q**	TCTCTTCCTGATCCTCACAGC	AGCCTCCCTCATTTCCCAAT
SARM1	**S182fs; 183_184del**	CTGTCCCCTTCCACTTTCAC	CTGCGCGCTTCTCTACCAT
CDC27	**V107delinsIV**	GAGGGATGAAGGAAGGAAGG	TGGAAATGCTTTTCTGACAGTT
MUC16	**N13088D**	CCGTCCTTCTCTCAGCAATC	CAGGCAGGAAGTGTGACCTC
MED12L	**2066_2095del**	CCCTCAAAAGGACGAATGAG	CATGAACAATGCAACCTGCT
FADS6	**P6delinsPMEPTEP**	CAGAGGAGCCAGGGTGTCT	CAAGGCGAAGAGGCTGAG
VSIG10L	**A859fs**	CAAGGATAGCCAGGCATTTT	GGGTCCAACCCTCAGTCATA
GRIA3	**G127fs**	AAGTTGCAGCAAAGGACCAT	CAGAGCAGGTGCCTCTTCTTA
ANAPC1	**Q465X**	CCCATACAGGTTGTTGCATGT	TCCCAATCTGAGCAAATTAACA
	L254X; L212X	TGAAAACTACAAAGAATGATCTGAA	CCCCTGTTTCCACAGTGATT
CDC27	**K169X**	AAGGCCTATTTCCTGTTTCCA	GAAAATTTCTACCAAGTCATTCACAA
	L155X; L144X	CAGGTTTTGGGGTTTGTAGC	TGATTCTTGAAACATACCGAAGA
KMT2C	**R904X**	AGATTTTTAAGTTGTGGAGTATACGTT	GCCTCACCCCAGGTAATACA
VEGFC	**X420L**	TTCCAGTCTGAAATTTAAAAACACA	TTTGTTAGCATGGACCCACA

Table 1: The primers used to check variants selected by WES analysis data.

Gene	Protein Variant	Primer Forward wild-type (fw-wt)	Primer Reverse	Primer Forward mutated (fw-mut)
NBPF16	**V646L**	TATCAGCTTCGCCCTTTAC**G**	CCTATGTCTGGGCTTCCAAA	TATCAGCTTCGCCCTTTAC**T**

Table 2: The primers for the variant in *NBPF*16 gene. The forward primers differ for the last base at 3ô end of the sequence (in bold) and are specific for wild-type allele or mutated allele.

The PCR product was purified using Illustra ExoProStar kit (GE Healthcare) that contains a mix of Illustra Alkaline Phosphatase and Exonuclease A, that remove unincorporated primers and nucleotides derived from the amplification reactions step.

The sequence reaction was performed with the BigDye Terminator v1.1 Cycle Sequencing Kit (Applied Biosystem). Afterwards, the sequences were purified with Ethanol/EDTA precipitation protocol and subjected to electrophoresis using the AB3130 XL Genetic Analyzers instrument (Applied Biosystem).

3. RESULTS

3.1. MUTATED FAP PEPTIDOME

3.1.1. MALDI-TOF analysis of plasma samples

Plasma samples of FAP patients (n=13), healthy subjects (n=38), adenoma patients (n=26) and CRC patients (n=58) were analyzed (Table 3). For CRC patients, the TNM (Tumour, Lymph Nodes, Metastasis) cancer classification system allowed the sub-grouping in early CRC (TNM staging I and II, n=29) and late CRC (TNM staging III and IV, n=29). Taking into account that FAP is a hereditary form of cancer characterized by an early onset of the pathology, the average age of FAP patients considered, resulted lower than the one of adenoma and CRC patients.

	Control	FAP	Adenoma	Early CRC	Late CRC
N	38	13	26	29	29
Gender M:F	20:18	07:06	14:12	18:11	18:11
Average age (years)	55	31	66	71	67
Range (years)	23-78	18-52	41-80	49-85	46-85

Table 3 Gender and age summary of controls subjects, FAP, adenoma, and early and late CRC patients

The clinical information collected from each FAP patient is reported in details in Table 4. Among the patients recruited, 7 were probands and 6 were patients, first or second degree relatives of these probands. Three FAP patients showed Gardner variant characterized by extra intestinal lesions such as cutaneous lesions, osteomas and dental abnormalities (Half et al., 2009).

Patient	Family	Proband relative	Age (y)	Surgery	Extraintestinal manifestation	Status of APC or MutYH gene	Sex
FAP1*	1	Proband	50	Proctocolectomy	Pulmonary nodels	APC: p.Y935X	M
FAP2	1	Daughter	19	Rectum resection	none	APC: p.Y935X	F
FAP42	1	Son	24	Rectum resection	none	APC: p.Y935X	M
FAP13*	9	Proband	37	Proctocolectomy	Osteomas and desmoids	APC: p.I1516EfsX9	F
FAP17*	10	Proband	22	No available	none	APC: Dup 4-5	F
FAP19	10	Second cousin	35	Colectomy	none	APC: Dup 4-5	M
FAP20	10	Second cousin	26	Proctocolectomy	Mandibular ostemas and renal cancer	APC: Dup 4-5	M
FAP35*	14	Proband	18	Total proctocoletomy	none	APC: p.S1436IfsX37	M
FAP26*	15	Proband	52	Total colectomy	Abdominal desmoids	APC: p.P1993PfsX50	F
FAP27	15	Daughter		No available	none	APC: p.P1993PfsX50	F
FAP24*	16	Proband	30	Proctocolectomy	none	APC: p.S905KfsX7	M
FAP25	16	Father	32	Proctocolectomy	none	APC: p.S905KfsX7	M
FAP37*	18	Proband		No available	none	MYH: c.1105delC	F

Table 4: The summary of the clinical information belonging to each mutated FAP patient.
***: proband; m: male; f: female; in bold: Gardner phenotype.**

For each patient, two spots of sample were analyzed by MALDI-TOF in order to reduce the instrumental variability. The spectra were acquired in the mass range 900-3000 m/z and the spectra signal was averaged to give a single spectrum for each sample and processed as previously described in the manual peaks picking procedure (Paragraph 2.1.3). A peaks list of 91 ionic species (m/z) and their relative signal intensities was obtained and subjected to statistical analysis.

For each specific group, the spectra obtained by MALDI-TOF are reported in Figure 11, where it is possible to observe the different intensity of the ionic species present in the groups considered. For instance, the spectra in Figure 11 (panel a and b) show a clear discrepancy in the peptides abundance between control subjects (panel a) and FAP patients (panel b). The ionic species at m/z 1896.03 and 2021.10 showed a higher intensity in FAP patients than in the control subjects. On the contrary, in the controls the most intense signal was detected at m/z 1060.56. Quite different patterns were observed in adenoma and in the two groups of CRC patientsô spectra (Figure 11, panel c, d, e), where the most intense signals were represented by the ionic species at m/z 1211.65, 1449.81, and 1855.86. These data suggest a possible correlation between the observed peptide profiles and the pathological state. However, the spectra from CRC patients stage I-II (Figure 11, panel d) and CRC patients stage III-IV (Figure 11, panel

e) resulted very similar between them. These results are in accordance with the data already published in literature (Zhu et al., 2013) and a possible explanation of this analogy could be that the peptide profile is not subjected to further alteration during CRC progression.

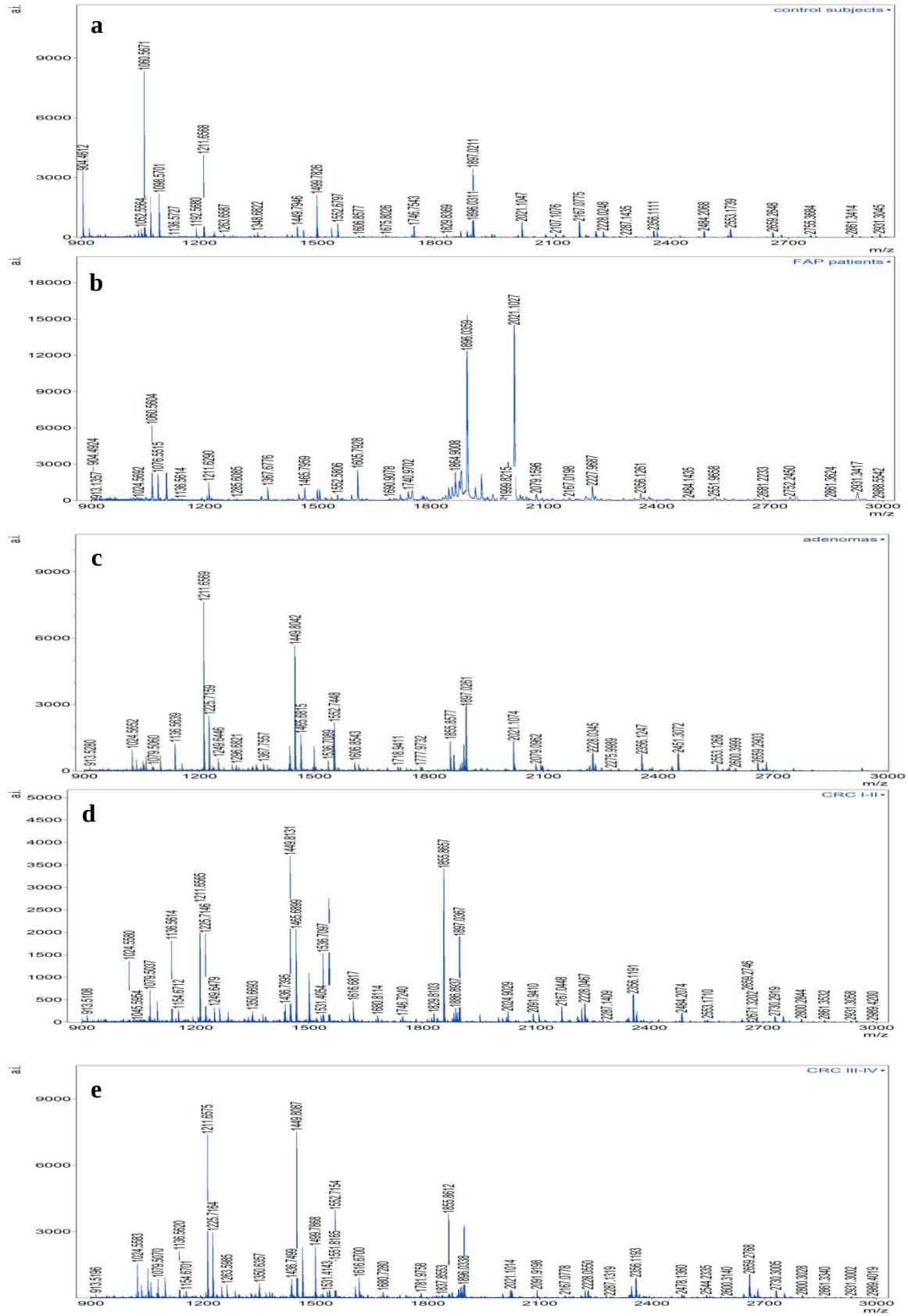


Figure 11: The MALDI-TOF spectra of control subjects (panel a), FAP patients (panel b), adenoma patients (panel c), stage I-II CRC patients (panel d) and stage III-IV CRC patients (panel e). a.i.: arbitrary units.

3.1.2. Statistical analysis

The peaks list obtained by MALDI data was subjected to statistical analysis. Both univariate and multivariate analysis techniques have been considered in order to choose ionic species with higher variance among groups. On the total of 91 ionic species detected, 50 resulted statistically significant at the ANOVA test (p value <0.05). First, a preliminary hierarchical clustering analysis among all the groups of study (controls subjects, FAP, adenoma, early CRC and late CRC patients) was performed. As represented in Figure 12, a higher dispersion among adenoma, early CRC and late CRC patients was found. In particular, adenoma patients were localized, some near CRC patients and other near the controls subjects. However, it was not possible to validate this clustering based on any known features (e.g. adenoma histotype, adenoma grading, CRC *RAS* mutational state). For this reason the early and late CRC were merged in a unique group called CRC.

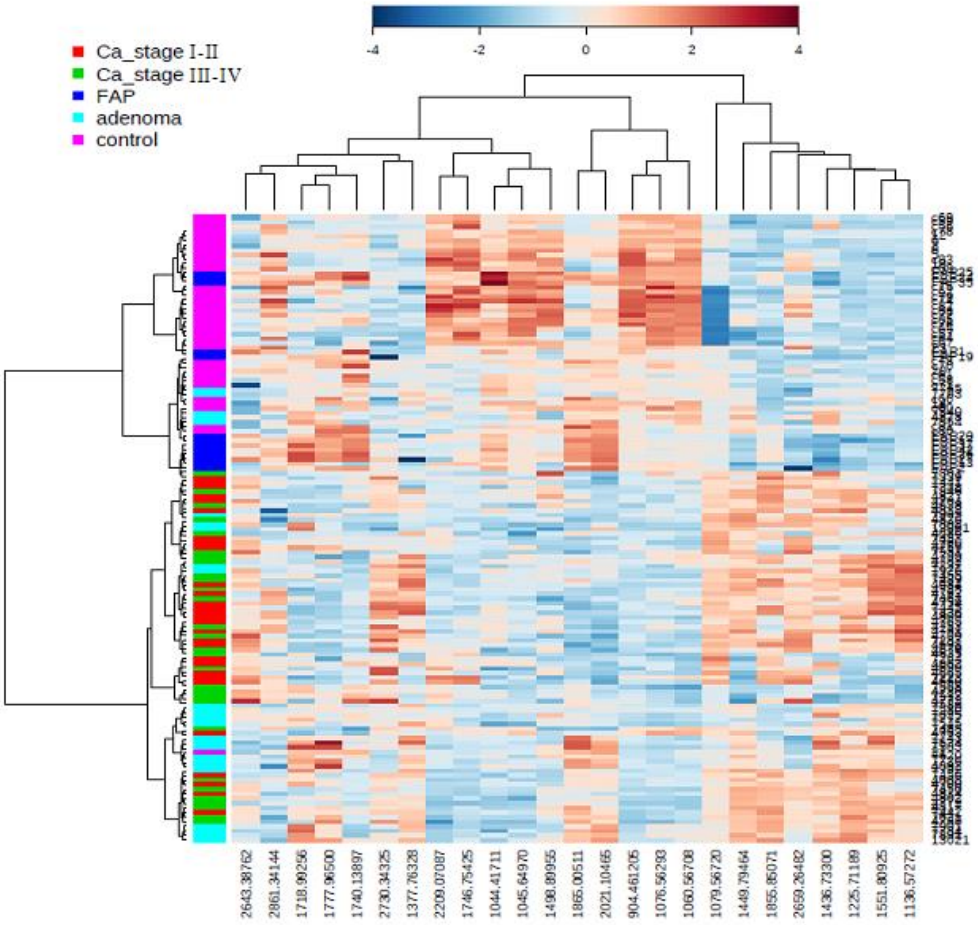


Figure 12 The heatmap of the clustering analysis (Euclidean distance and Ward algorithm) of control subjects, FAP, adenoma, early CRC (stage I-II) and late CRC (stage III-IV) patients.

Therefore, a different approach was considered and another hierarchical cluster analysis between FAP patients versus control subjects, adenoma patients and the two groups of CRC patients was performed. Using this different statistical strategy, it was possible to obtain a clear distinction among FAP patients and the other groups considered (Figure 13, 14 and 15). Indeed, the FAP patients showed a good classification compared to the controls and the adenoma patients. However, some CRC patients resulted to be erroneously clustered with FAP patients, due to the similar intensity profile of few ionic species.

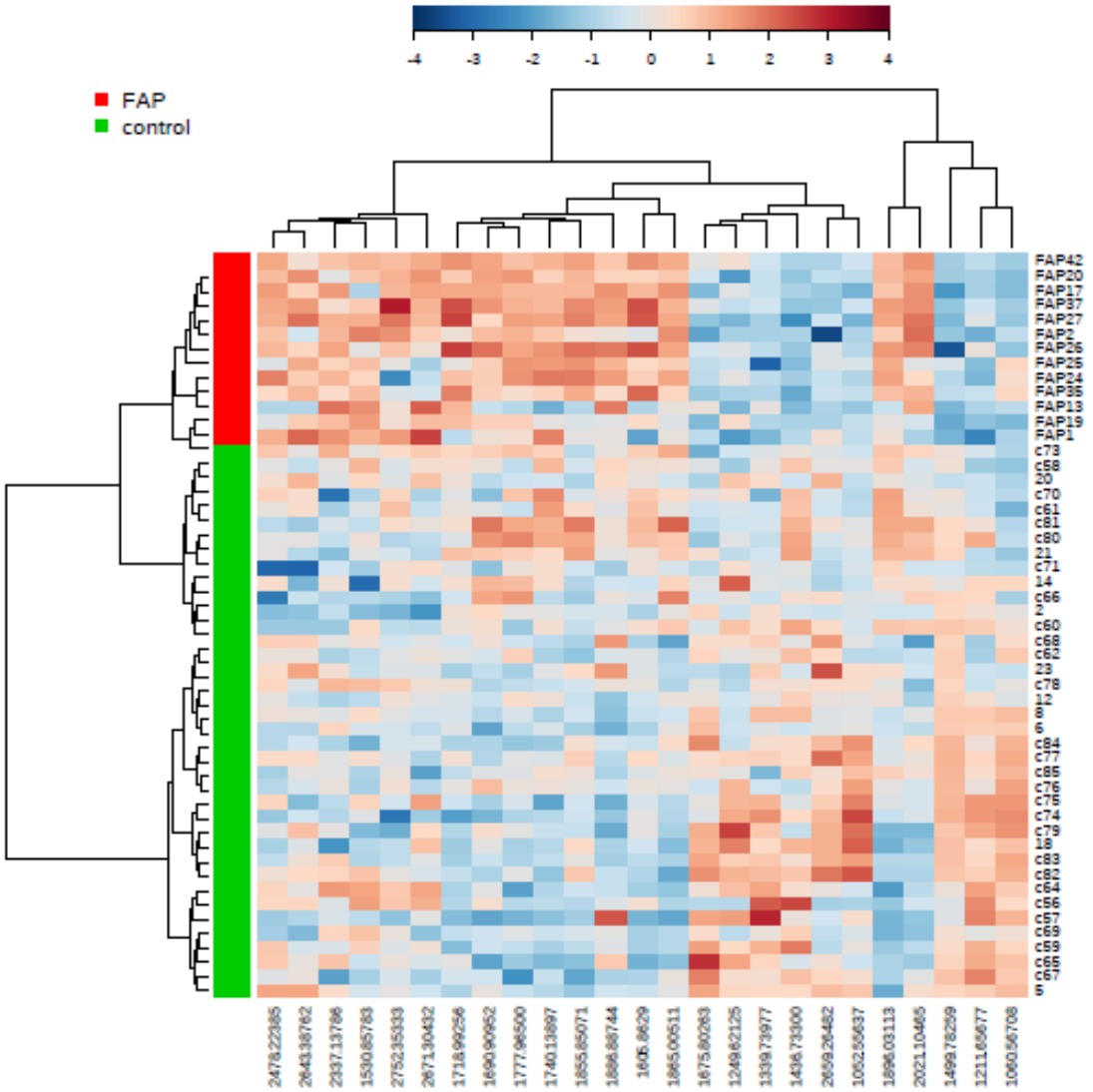


Figure 13: The heatmap of clustering analysis (Pearson distance and Ward algorithm) of FAP patients versus control subjects.

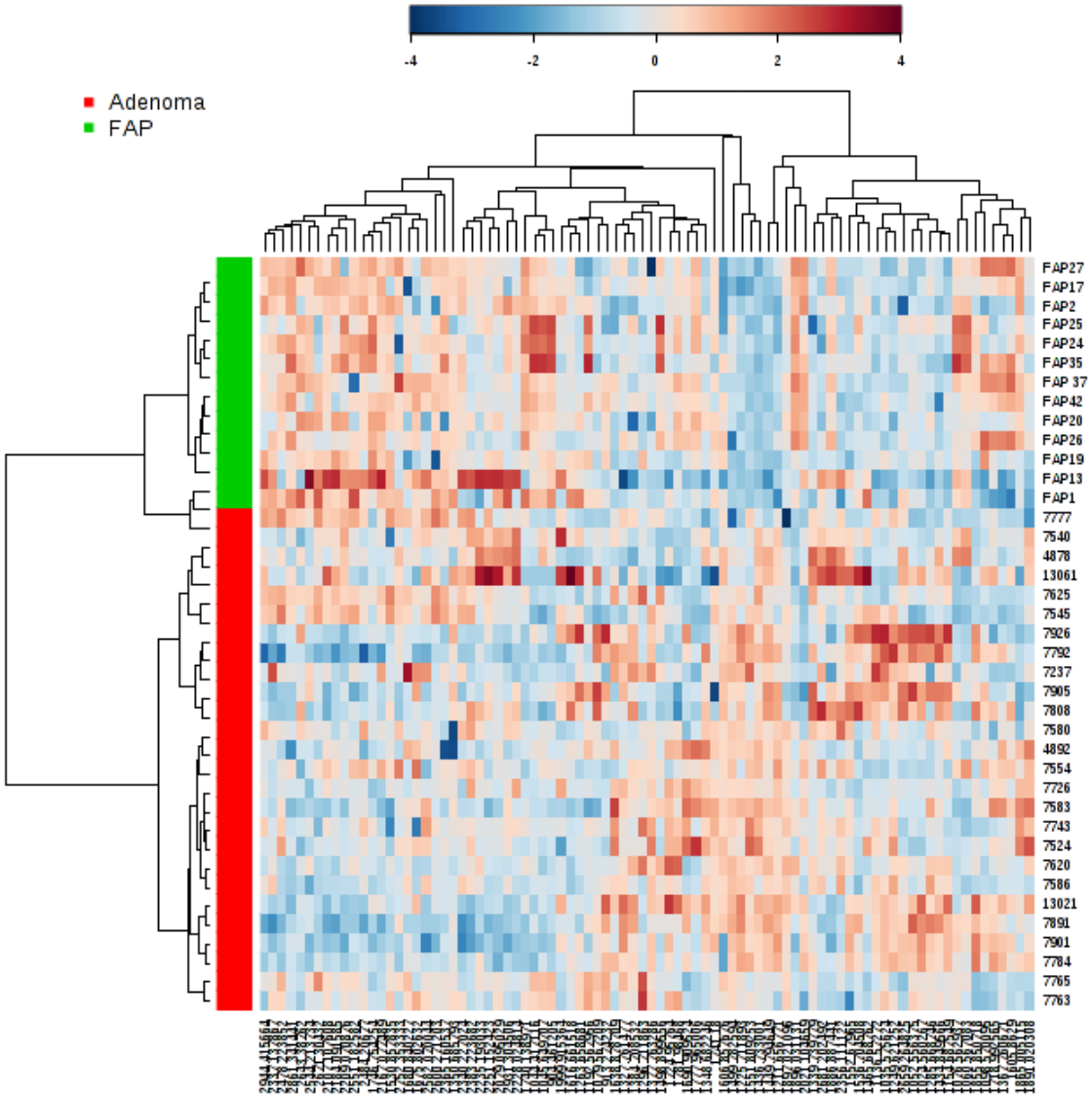


Figure 144: The heatmap of clustering analysis (Pearson distance and Ward algorithm) of FAP patients versus adenoma patients.

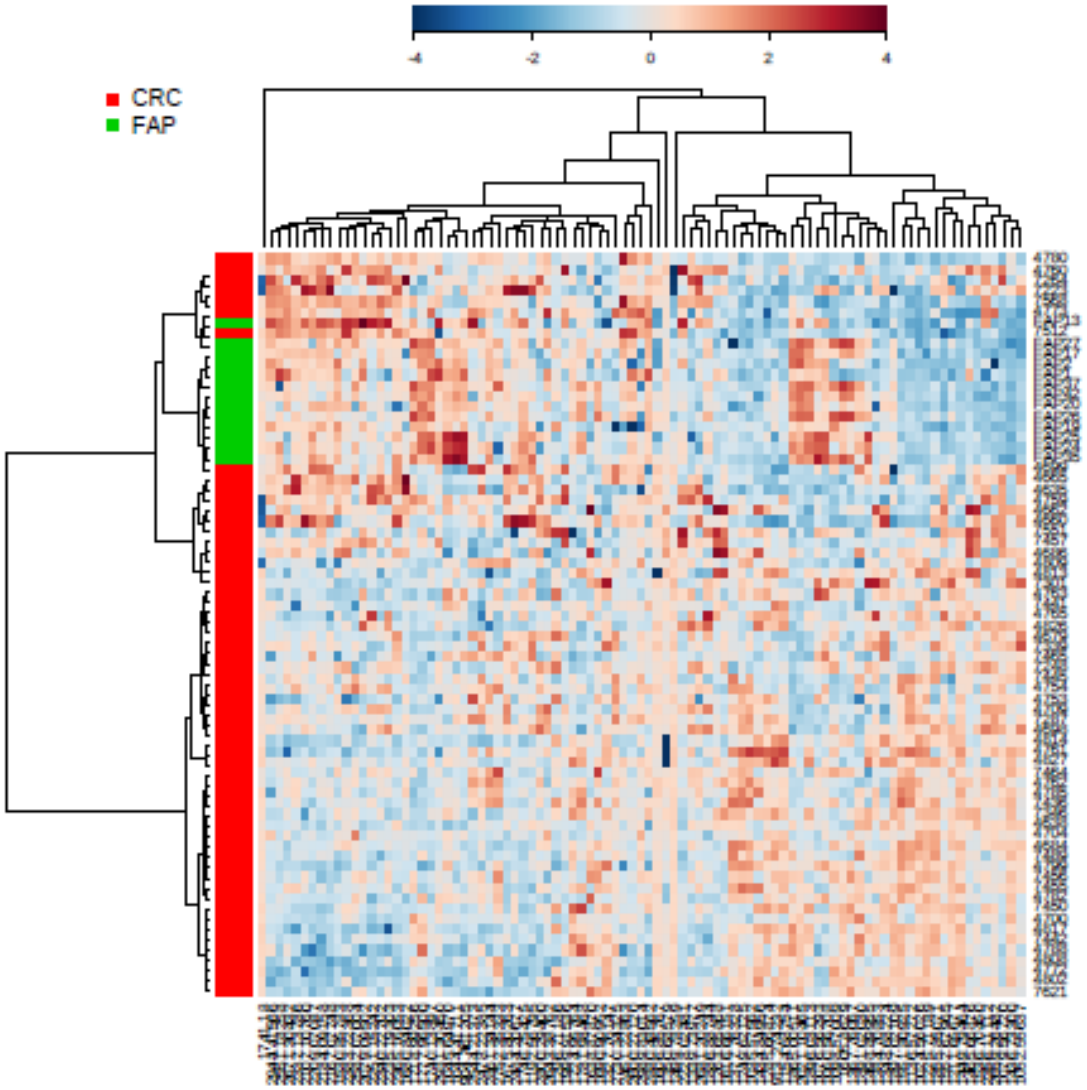


Figure 15: The heatmap of clustering analysis (Pearson distance and Ward algorithm) of FAP patients versus CRC patients.

On the contrary, considering more robust statistical methods, such as Volcano plot (univariate analysis), RF and PLS-DA (multivariate analysis), this overlap between FAP and CRC patients did not occur. Using these tests to compare FAP patients versus control subjects, adenoma patients and CRC patients respectively, three statistically significant subsets of 18, 40 and 17 m/z were obtained. In total, 45 ionic positive species were found, 29 of them showing a fold-change >2 in all the three statistical tests (Table 5).

m/z	Controls /FAP	Adenoma/ FAP	CRC/ FAP	ID peptide	Sequence
904.46	3.44	0.16	0.09	Des-Arg bradykinin	RPPGFSPF
1060.57	6.92	0.21	0.09	Bradykinin	RPPGFSPFR
1076.56	2.73	0.18	0.14	Oxidized bradykinin	RPPGFSPFR
2209.07	2.04	0.16	0.14	Kininogen fragment	KHNLGHGHKHERDQGHGHQ
2365.16	1.08	0.16	0.20	Kininogen HMW fragment	KHNLGHGHKHERDQGHGHQR
2021.10	**0.08**	**0.10**	**0.01**	**C3f**	**SSKITHRIHWESASLLR**
1035.58	0.42	1.60	1.99	C3f fragment	THRIHWESA
1136.57	0.46	2.42	4.61	C3f fragment	THRIHWESA
1211.66	**2.57**	**9.29**	**2.98**	**C3f fragment**	**IHWESASLLR**
1249.62	1.13	3.75	3.20	C3f fragment	ITHRIHWESA
1367.76	**0.15**	**0.61**	**0.18**	**C3f fragment**	**RIHWESASLLR**
1605.86	**0.10**	**0.37**	**0.10**	**C3f fragment**	**THRIHWESASLLR**
1718.99	**0.13**	**0.61**	**0.08**	**C3f fragment**	**ITHRIHWESASLLR**
1777.97	**0.14**	**0.36**	**0.11**	**C3f fragment**	**SKITHRIHWESASLL**
1865.01	**0.10**	**0.26**	**0.04**	**C3f fragment**	**SSKITHRIHWESASLL**
1855.85	0.08	0.51	3.92	C3c	SEETKENEGFTVTAEGK
1896.03	**0.04**	**0.04**	**0.02**	**C4A/B precursor fragments**	**NGFKSHALQLNNRQIR**
1154.69	0.77	3.14	2.29	C4A/B fragment	LQLNNRQIR
1225.71	0.97	13.76	17.79	C4A/B fragment	ALQLNNRQIR
1327.76	0.57	1.33	1.06	C4A/B fragment	FKSHALQLNNR
1436.73	**2.41**	**9.13**	**6.06**	**C4A/B fragment**	**GLEEELQFSLGSK**
1449.79	1.48	11.37	11.64	C4A/B deamidated fragment	SHALQLNNRQIR
1498.90	1.94	0.51	0.48	C4A/B fragment	NGFKSHALQLNNR
1499.78	**147.56**	**68.78**	**79.36**	**C4A/B fragment**	**SHALQLNNRQIR**
2861.34	0.58	0.16	0.28	Fibrinogen. alpha chain	QGVNDNEEGFFSARGH
1606.86	**11.78**	**61.36**	**81.10**	**Fibrinogen. alpha chain fragment**	**SSHHPGIAEFPSRGK**
1552.68	**2.30**	**6.04**	**14.06**	**Fibrinopeptide B**	**QGVNDNEEGFFSAR**
1746.75	1.93	0.31	0.27	Fibrinogen. beta chain fragment	SSHHPGIAEFPSRGK
2167.08	0.83	0.21	0.18	Pyro-glu ITIH4 fragment	QLGLPGPPDVPDHAAYHPFR

Table 5: The median intensity ratio (controls vs FAP; adenoma vs FAP and CRC vs FAP) of statistically significant m/z and their relative identification and sequence. In bold the distinctive peptides of FAP patients are reported.

3.1.3. Peptide identification

Fifty peptides out of the 91 present in the peaks list were identified by MS/MS experiments (Table 6).

Peptide	m/z	Sequence
Protein: P01024, Complement C3		
C3c	**1855.9**	SEETKENEGFTVTAEGK
C3f	**2021.1**	SSKITHRIHWESASLLR
C3f fragments	**1035.6**	THRIHWESA
	1098.6	HWESASLLR
	1136.6	THRIHWESA
	1211.7	IHWESASLLR
	1249.6	ITHRIHWESA
	1367.8	RIHWESASLLR
	1377.8	KITHRIHWESA
	1465.1	SKITHRIHWESA
	1551.8	SKITHRIHWESA
	1605.9	THRIHWESASLLR
	1719.0	ITHRIHWESASLLR
	1778.0	SKITHRIHWESASLL
	1865.0	SSKITHRIHWESASLL
Protein: P01042, Kininogen HMW		
Bradikinin	**1060.6**	RPPGFSPFR
	1076.6	RPPGFSPFR
Des-Arg Bradikinin	**904.5**	RPPGFSPF
	1348.7	RHDWGHEKQR
Kininogen-1 light chain	**2365.2**	KHNLGHGHKHERDQGHGHQR
	2209.1	KHNLGHGHKHERDQGHGHQ
Protein: P02671, Fibrinogen alpha chain		
Alpha chain	**2861.3**	MADEAGSEADHEGTHSTKRGHAKSRPV
	2659.3	DEAGSEADHEGTHSTKRGHAKSRPV
	2681.2	SSSYSKQFTSSTSYNRGDSTFESK
	1606.9	SSHHPGIAEFPSRGK
Fibrinopeptide A	**1536.7**	ADSGEGDFLAEGGGVR
	1616.7	ADSGEGDFLAEGGGVR
Protein: P02675, Fibrinogen beta chain		
Beta chain	**2235.2**	KREEAPSLRPAPPPISGGGYR
	2484.2	QGVNDNEEGFFSARGHRPLDKK
	2356.1	QGVNDNEEGFFSARGHRPLDK
	2228.0	QGVNDNEEGFFSARGHRPLD
	1999.9	QGVNDNEEGFFSARGHRP
	1746.8	QGVNDNEEGFFSARGH
Fibrinopeptide B	**1552.7**	QGVNDNEEGFFSAR
Protein: P08697, Alpha-2 antiplasmin		
Propeptide	**1285.7**	MEPLGRQLTSGP
Protein: P0C0L4/5, Complement C4A/B		
C4 b	**1896.0**	NGFKSHALQLNNRQIR
	1897.0	NGFKSHALQLNNRQIR
	1891.0	GLEEELQFSLGSKINVK
C4 b fragments	**1499.8**	NGFKSHALQLNNR
	1052.6	SHALQLNNR
	1154.7	LQLNNRQIR
	1225.7	ALQLNNRQIR
	1327.8	FKSHALQLNNR
	1384.7	GFKSHALQLNNR
	1436.7	GLEEELQFSLGSK
	1449.8	SHALQLNNRQIR
	1498.9	NGFKSHALQLNNR
	1782.0	GFKSHALQLNNRQIR
Protein: P10909, Clusterin		
Beta chain peptide	**1530.9**	RPHFFFPKSRIV
Protein: Q14624, ITIH4		
Proline-rich (PRR) peptide	**1675.8**	GPPDVPDHAAYHPFR

Table 6: The peptides identified among the total ones present in the peaksô list.

The Table 5 reports the identification of 29 peptides identified through the previously described statistical analysis. The peptides are mainly fragments of circulating proteins, such as: Kininogen, Fibrinogen, Complement 3, Complement 4A/B and Inter-Alpha-Trypsin Inhibitor Heavy Chain. Most of the observed peptides belong to few precursor peptides subjected to extensive ladder-like truncations. For instance, several fragments derive from the peptides identified at m/z 2021.10 and 1896.03. Twelve of the identified m/z had a distinctive signature in FAP patients (Table 5, in bold) showing the same trend in these patients but not in the other groups. These peptides were: C3f (at m/z 2021.10) and its fragments (at m/z 1211.66; 1367.76; 1605.86; 1718.99; 1777.96; 1865.01); the C4A/B precursor fragments (at m/z 1896.03) and its fragments (at m/z 1436.73; 1499.78); the Fibrinopeptide B (at m/z 1552.67) and Fibrinogen alpha chain peptide (at m/z 1606.86).

The C3f is a 17 amino acids peptide having a mass-average of 2020 Da and derives from the cleavage of C3b into iC3b done by the complement Factors I and H (Ishida et al., 2008). The m/z 2021.10 intensity (C3f) was higher in FAP patients and was found to be significantly decreased in the controls, adenomas and CRC patients (Figure 16, panel a). Additionally, C3f intensity was statistically significant different between FAP probands and FAP relatives, highlighting the effect of genetic mutation in the phenotype. The same tendency was observed for the intensity of its fragment at m/z 1367.76, 1605.86, 1718.99, 1777.89 and1865.01 (Figure 17, panel a). On the contrary, only one C3f fragment intensity (at m/z 1211.66) was decreased in FAP patients and not in the other groups. C3f fragments at m/z 1035.57 and 1136.57 were higher in all patient groups, whereas m/z 1249.62 was high only in adenoma and CRC patients.

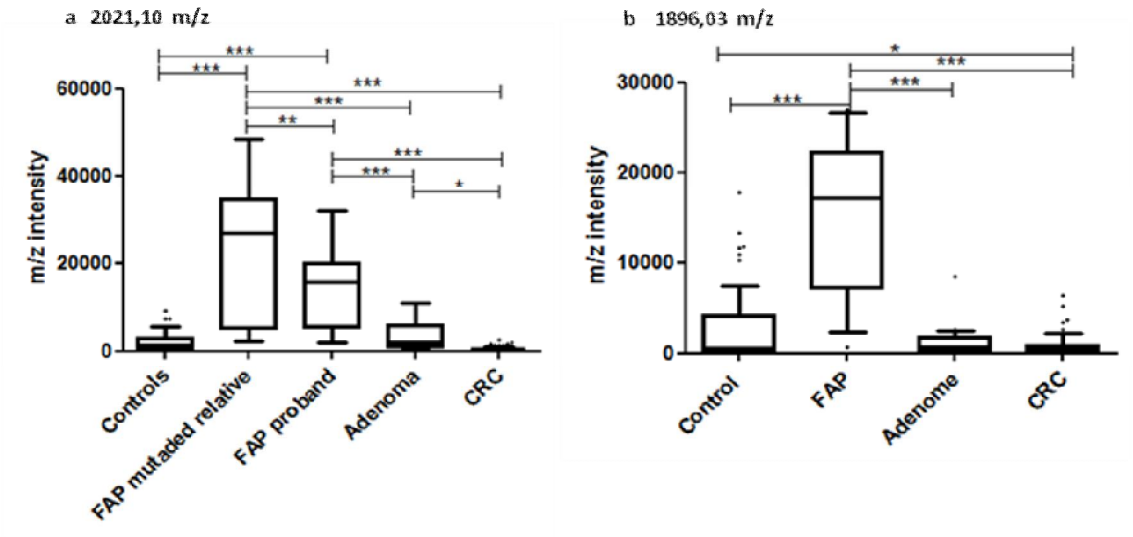
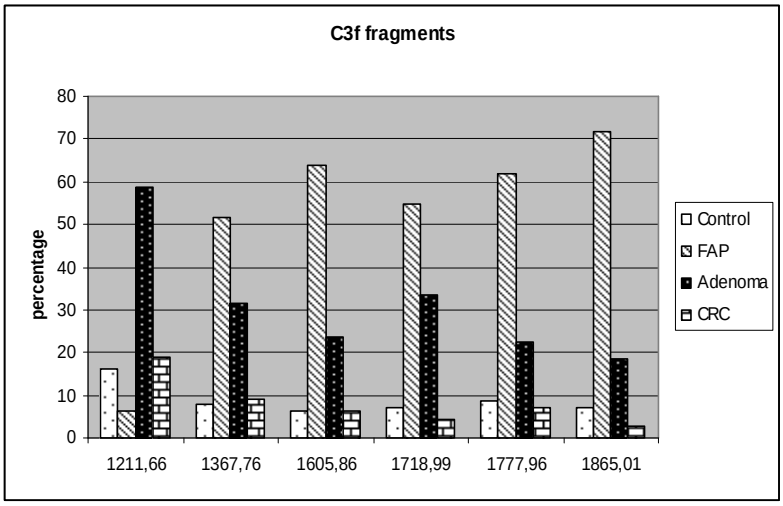


Figure 16: Box plot of ionic species significantly different among the samples of control subjects, FAP patients, adenoma and CRC patients. (Tukeyôs outliers are reported; *p<0.05; **p<0.01; *** p<0.001).

The m/z 1855.85 intensity (Complement C3c) was increased in FAP and in adenoma and it was found even higher in CRC patients.

The C4A/B fragment peptide at m/z 1896.03 was significantly increased in FAP patients (Figure 16, panel b). The intensity of the m/z 1436.73 and 1499.78 fragments were decreased in FAP patients whereas both of them were increased in adenoma and CRC patients (Figure 17, panel b). The C4A/B fragments derived from the ionic species at m/z 1896.03 (i.e. those at m/z 1154.68, 1225.71, 1327.76 and 1449.79) were higher in adenoma and CRC patients, instead the m/z 1498.89 was lower in these groups.

a



b

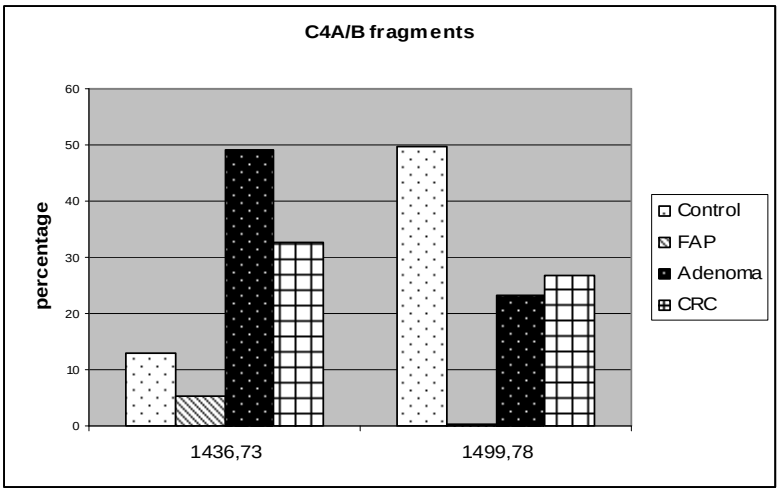


Figure 17: The percentage of C3f intensity (panel a) and C4A/B (panel b) fragments relative to their total amount in FAP, adenoma and CRC patients and controls groups. The percentage of each m/z species was calculated dividing the single intensity by the sum of their total intensity.

The intensity of ionic species at m/z 1552.67 (Fibrinopeptide B) and 1606.86 (the peptide of alpha chain fragment of Fibrinogen) were found to be higher in adenoma and CRC patients than in controls and FAP patients.

In order to understand which peptides could mark the switch between adenoma and malignant carcinoma, they were chosen only the ionic species showing an increase or decrease trend among FAP, adenoma and CRC patients and that were statistically different in all kind of comparisons. Using this criteria, it was possible to select four ionic species with a decreasing tendency: C3f peptide (at m/z 2021.10) and its fragments at m/z 1367.76, 1718.99 and 1865.01. These four peptides were differently present among the patientsô groups considered, although they showed a similar intensity both in the control subjects and in the CRC patients (Figure 16 and 18).

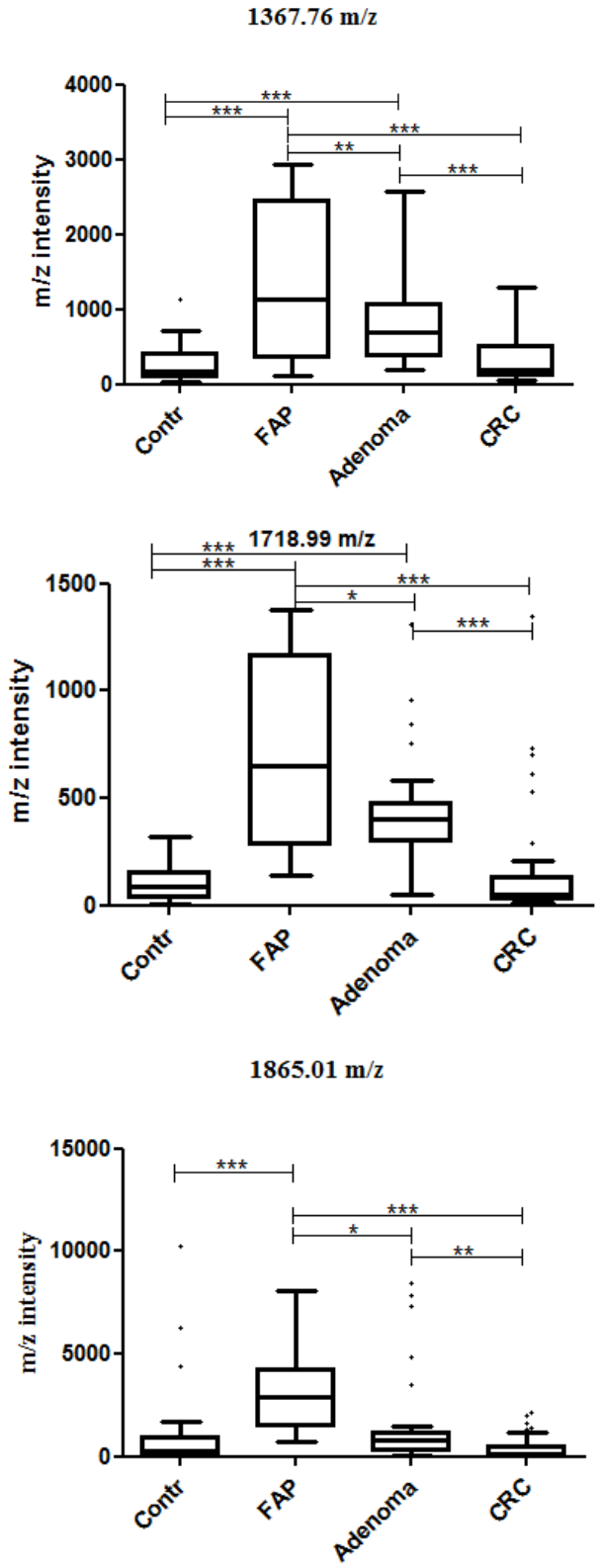


Figure 18: Box plot results about ANOVA significant C3f fragments with decrease trend among FAP, Adenoma and CRC patients. (Tukeyôs outliers are reported; *p<0.05; **p<0.01; *** p<0.001)

3.1.4. C3 and C4 quantification by ELISA

In order to understand if the observed changes in the peptides derived from complement proteins were related to changes in the precursor proteins C3b and C4b levels, the ELISA assay was performed. There were no differences in the quantity of C4b among the groups samples (Figure 19, panel b). Instead, the quantity of C3b between control group and CRC patients was different as well as between FAP and CRC patients; whereas between FAP patients and control group no significant expression changes were detected (Figure 19, panel a).

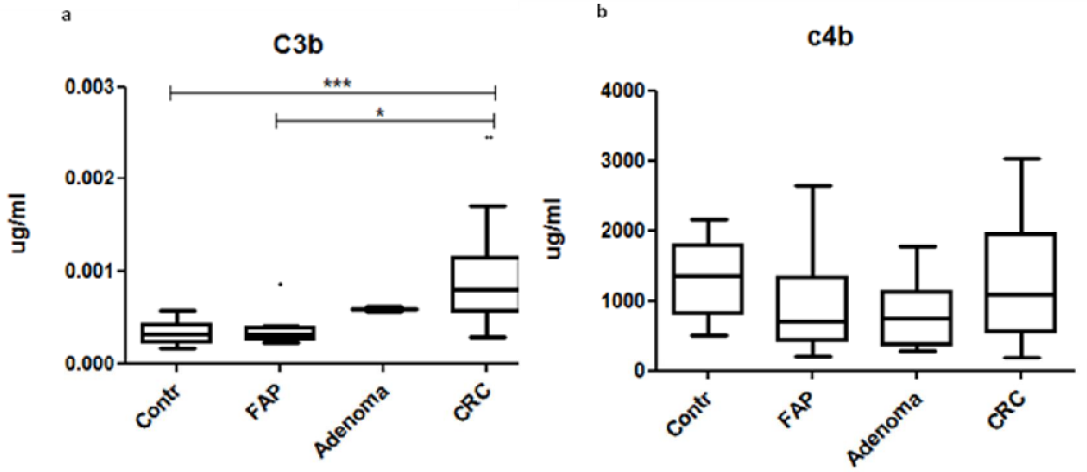


Figure 19: Box plot showing plasma amount of C3b (a) and C4b (b) in FAP, adenoma, CRC patients and control subjects. Highly significant C3b concentration changes were detectable only between CRC patients and healthy subjects and between FAP and CRC patients (a). (Tukeyôs outliers are reported; *p<0.05; **p<0.01; *** p<0.001).

3.2. UNRESOLVED FAP PEPTIDOME

The plasma sample of the four unresolved FAP patients was analysed and a peaks list of 92 ionic species, (m/z) representing plasmatic peptide and their relative signal intensities, was obtained. In this study, the attention was focused on the ionic species described in the previous paragraphs to be characteristic of mutated FAP patients. In particular, the C3f fragment at m/z 2021.10 and C4A/B precursor fragment at m/z 1896.03 in unresolved FAP patients have characteristic similar to the other groups (control subjects, adenoma patients and CRC patients) but not to the mutated FAP patients (Figure 20). These data further indicate and sustain the differences between unresolved and mutated FAP groups.

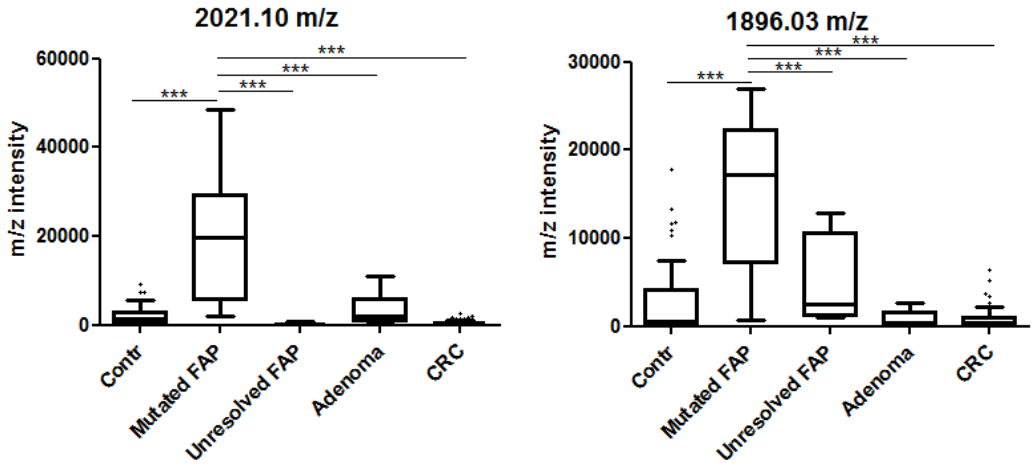


Figure 19: Box plot of ionic species significantly different among the samples of control subjects, mutated and unresolved FAP patients, adenoma and CRC patients. (Tukeyôs outliers are reported; *** p<0.001).

3.3. UNRESOLVED FAP WHOLE EXOME SEQUENCING

In order to be able to identify a possible specific genetic pattern for unresolved FAP, a WES analysis by SOLiD4 instrument was performed. The clinical information collected from each unresolved FAP patient are reported in details in the Table 7.

Patient	Age (y)	Surgery	Extra-intestinal Manifestations	Sex
FAP179*	68	Right hemicolectomy	Melanoma	m
FAP204*	54	Proctocolectomy	none	m
FAP207*	25	Subtotal colectomy	Osteomas, lipoma and supernumerary teeth	m
FAP211*	61	Subtotal colectomy	none	m

Table 7: Summary of the clinical information of each unresolved FAP patient.
***: proband; m: male; f: female; bold: Gardner phenotype.**

3.3.1. Filtered variants confirmation approach

The reads were mapped to the hg19/GRCh37 Reference Genome and the variants were identified as described previously in the paragraph first elaboration of the data in the materials and methods section (Paragraph 2.2.4) and filtered using filter a. Using these criteria and applying VCF tools program (vcftools is a suite of functions to use on genetic variation data in the form of VCF and BCF files; the tools provided is used mainly to summarize data, run calculations on data, filter out data, and convert data into other useful file formats) it was possible to select 16 variants present in all unresolved FAP patients (Table 8).

Gene	Name	Exonic Function	Reference sequence:coding sequence variant:protein variant	Sanger validation
LOC440563	Heterogeneous Nuclear Ribonucleoprotein C-Like2, pseudogene	nonsynonymous SNV	NM_001136561:c.A764G:p.D255G	n.p.
LOC440563	Heterogeneous Nuclear Ribonucleoprotein C-Like2	nonsynonymous SNV	NM_001136561:c.G763T:p.D255Y	n.p.
NBPF16	Neuroblastoma Breakpoint Family, Member 16	nonsynonymous SNV	NM_001102663:c.G1936T:p.V646L	wt
RBMXL1	RNA Binding Motif Protein, X-Linked-Like 1	frameshift deletion	NM_019610:c.1delA:p.M1fs	n.p.
MED12L	Mediator complex subunit 12-like	nonframeshift deletion	NM_053002:c.6198_6284del:p.2066_2095del	wt
KMT2C	Histone methyltransferase gene mixed-lineage leukemia 3 (Lysine specific Methyiltransferase 2C)	stopgain SNV	NM_170606:c.2447dupA:p.Y816_I817delinsX	wt
PABPC1	Poly(A) binding protein, cytoplasmic 1	frameshift insertion	NM_002568:c.467_468insG:p.E156fs	wt
PRSS3	Protease, serine, 3, mesotrypsinogen	nonsynonymous SNV	NM_001197098:c.A454C:p.K152Q	wt
SARM1	Sterile alpha and TIR motif containing 1	frameshift insertion	NM_015077:c.544_545insGC:p.S182fs	wt
			NM_015077:c.549_550del:p.183_184del	wt
CDC27 (APC3)	Cell division cycle 27 (anaphase promoting complex 3)	nonframeshift insertion	NM_001114091:c.318_319insATC:p.V107delinsIV	wt
FADS6	Fatty acid desaturase 6	nonframeshift insertion	NM_178128:c.17_18insGATGGAACCTACGGAGCC:p.P6delinsPMEPTEP	Sequencing error or polymorphism
KATNAL2	Katanin p60 subunit A-like 2	frameshift insertion	NM_031303:c.850_851insGC:p.E284fs	n.p.
VSIG10L	V-set and immunoglobulin domain containing 10 like	frameshift insertion	NM_001163922:c.2576dupC:p.A859fs	rs397723022 dbSNV, present also in healthy controls
MUC16 (CA125)	Mucin 16, cell surface associated	nonsynonymous SNV	NM_024690:c.A39262G:p.N13088D	wt
GRIA3	Glutamate receptor, ionotropic, AMPA 3	frameshift insertion	NM_001256743:c.382dupG:p.G127fs	wt

Table 8: List of 16 common variants obtained with first elaboration of data and filtered with filter a.
SNV: Single Nucleotide Variant; n.p.: not performed; wt: wild-type.

After a careful literature revision, out of the 16 common variants only 7 variants and 6 genes were described to be involved in cancer development. These genes were: *KMT2C*, *PABPC1*, *PRSS3*, *SARM1*, *CDC27* and *MUC16*. To confirm the selected variants, the Sanger sequencing was performed, but all the 6 genes resulted wild type.

To further sustain the robustness of this analytical approach, other variants present in the *NBPF16*, *MED12*, *FAD6*, *VSIG10L* and *GRIA3* genes were confirmed with other methods.

For the variants in *NBPF16* and *MED12* genes, the PCR primer design allowed a simple check of the PCR product on the agarose gel. The unresolved FAP patients presented wild type alleles for both gene variants.

Concerning *NBPF16* gene, a PCR was performed using a common reverse primer and two different forward primers, one specific for the wild type allele and the other specif-

ic for the mutated allele. The forward primers differ from each other for the last base at 3ô of the sequence. The unresolved FAP patients were wild type for this gene variant, indeed, as it is showed in figure 21, only the reverse/wild type forward primers combination shows the amplification band whereas the reverse/mutated forward primers combination did not amplify any band.

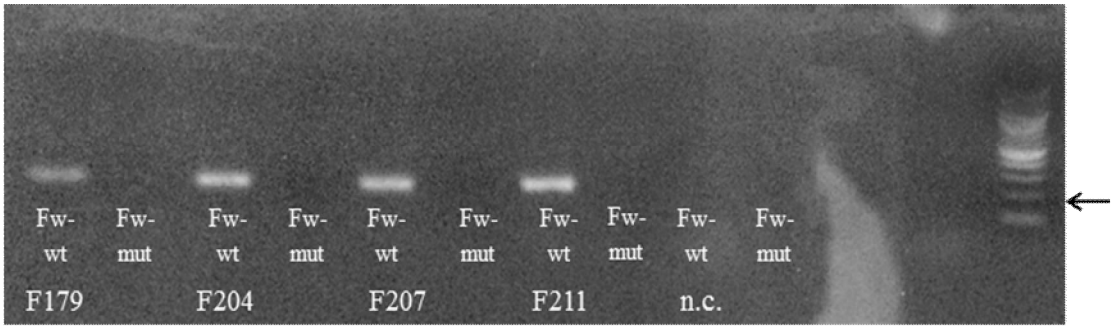


Figure 20: NBPF16 gene agarose gel picture. Unresolved FAP patients present only the amplification band resulted from the reverse/wild type forward primers combination and no band were detectable from the reverse + mutated forward primers combination. n:c: negative control; Fw wt: wild type forward primer ; Fw mut: mutated forward primer ; arrow: 200bp

Concerning *MED12* gene, the PCR product of the wild type has 86 bases more compared to the mutated subject, allowing a clear visualization of the amplified products using the agarose gel. As it is showed in figure 22 the unresolved FAP patients were all wild type for the MED12 deletion obtained from WES analysis.

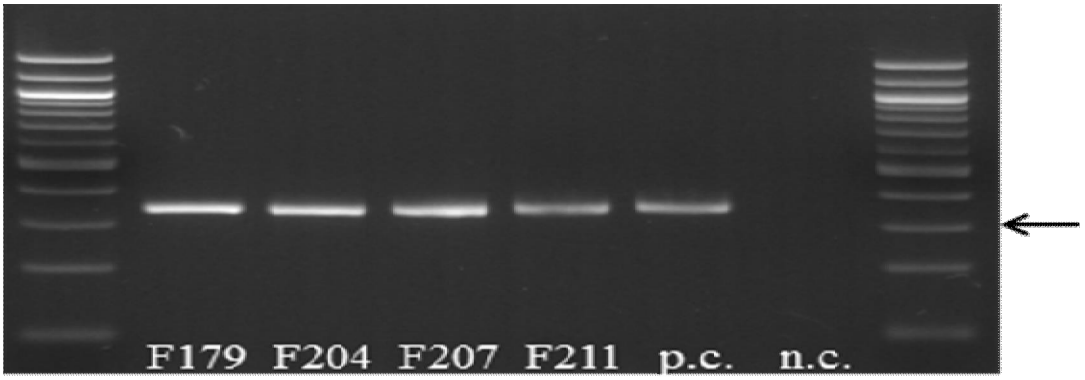


Figure 21: MED12 gene agarose gel picture. The wild-type allele has a 346 base pair long band, whereas the deleted allele has a 259 base pairs band. Unresolved FAP patients have only the 346 bp wild type band. p.c. positive control (wt), n:c: negative control; arrow: 300bp.

Sanger sequencing was instead performed for *FAD 6, VSIG10L* and *GRIA3* genes. In particular, *FAD6* and *VSIG10L* gene resulted mutated in FAP patients and in the healthy controls (n=5). In particular, for the *FAD6* gene, the UNIPROT database reported that: ñthe N-terminus of the sequence BAC04349.1 contains a duplication of the repeated MEPTEP sequence. It is unclear whether such duplication is due to a se-

quencing error or a polymorphismò. Concerning the *VSIG10L* gene, its variant is present in the dbSNPs database (i.e. rs397723022) and, using another software (i.e. Mutation Tasting Prediction software), is considered as a polymorphism. The *GRIA3* gene, instead, resulted to be wild-type using the Sanger sequencing.

At this point the variants were re-identified as described in the paragraph of the materials and methods section (the second elaboration of the data) and filtered using filter a. First it was taken in consideration the stop loss and the stop gain variants, identifying 15 common genes and 21 common variants (Table 9).

Gene	Name	Reference sequence:coding sequence variant:protein variant	Sanger sequencing
ANAPC1	Anaphase Promoting Complex Subunit 1	NM_022662:c.C1393T:p.Q465X	Present also in healthy controls
ANKRD36	Ankyrin Repeat Domain 36	NM_001164315:c.T4479G:p.Y1493X	n.p.
ANKRD36B	Ankyrin Repeat Domain 36B	NM_025190:c.T2715G:p.Y905X	n.p.
CCDC144NL	Coiled-Coil Domain Containing 144 Family, N-Terminal Like	NM_001004306:c.C533A:p.S178X	n.p.
CDC27	Cell division cycle 27 (anaphase promoting complex 3)	NM_001114091:c.T761G:p.L254X	wt
		NM_001114091:c.T635G:p.L212X	wt
		NM_001114091:c.A505T:p.K169X	wt
		NM_001114091:c.T464G:p.L155X	wt
		NM_001114091:c.T431G:p.L144X	wt
CTBP2	C-Terminal Binding Protein 2	NM_001083914:c.A22T: p.K8X	n.p.
HYDIN	Axonemal Central Pair Apparatus Protein	NM_001270974:c.C11713T:p.Q3905X	n.p.
		NM_001270974:c.T3052C:p.X1018Q	n.p.
KMT2C	Histone methyltransferase gene mixed-lineage leukemia 3 (Lysine specific Methyiltransferase 2C)	NM_170606:c.C2710T:p.R904X	wt
NCOR1	Nuclear Receptor Corepressor 1	NM_001190438:c.C241T:p.R81X	n.p.
PRAMEF1	PRAME Family Member 1	NM_023013:c.T314A:p.L105X	n.p.
SEC22B	SEC22 Vesicle Trafficking Protein Homolog B (S. Cerevisiae) (Gene/Pseudogene)	NM_004892:c.C394T:p.R132X	n.p.
SRGAP2B, SRGAP2C, SRGAP2D	SLIT-ROBO Rho GTPase Activating Protein 2B, 2C, 2D (pseudogene)	NM_001271872:c.T1096G:p.X366G	n.p.
TP53TG5	TP53 Target 5	NM_014477:c.G292T:p.E98X	n.p.
VEGFC	Vascular Endothelial Growth Factor C	NM_005429:c.A1259T:p.X420L	wt
ZNF717	Zinc Finger Protein 717	NM_001128223:c.A2518T:p.K840X	n.p.
		NM_001128223:c.G1957T:p.G653X	n.p.
		NM_001128223:c.G1858T:p.E620X	n.p.

Table 9: The list of 21 common stop gain and stop loss variants obtained with the first elaboration of data and filtered with filter b.
n.p.:not performed; wt: wild-type.

Among these, 4 genes (i.e. *CDC27, KMT2C, VEGFC, ANAPC1*) are known in literature to be involved in cancer development and in this thesis 8 variants of these genes were confirmed by Sanger sequencing. The variants in the *CDC27, KMT2C, VEGFC* genes resulted wild type, whereas the variant in *ANAPC1* gene resulted mutated in FAP patients and in the health control (n=5).

3.3.2. Pathways enrichment approach

In the case of the four unresolved FAP patients, a different strategy was considered in order to identify and select specific disease related genes and variants. The WES data were screened for the pathway and not anymore for the genes.

The variants were identified as described in the second elaboration of data paragraph in the materials and methods section and filtered using more stringent filter b. Using these criteria it was possible to identify a list of variants for each unresolved FAP patients (Table 10). The table shows that one of the patients (FAP179) resulted to have almost two times more variants than the others. This discrepancy could be explained in two different ways. The first could be that the patient, indeed, presents more variants in his genomes compared to the others. The second possibility is that, due to technical reasons, the quality of the DNA extract from this patient was not sufficient to obtain a clear and specific read out.

	Total variants	Filtered variants	Genes
FAP179	53464	2160	1462
FAP204	24754	1327	1002
FAP207	25413	939	718
FAP211	22779	490	379

Table 10: Number of variants and genes obtained with the second elaboration of the data and filtered with filter b for each unresolved FAP patients.

From these results, only the variants present in all the patients were selected and 372 common variants and 285 common genes were identified.

Considering all the common genes, the hypergeometric distribution (Figure 24) was used to perform the pathway enrichment analysis with the PathDIP software (http://ophid.utoronto.ca/pathDIP).

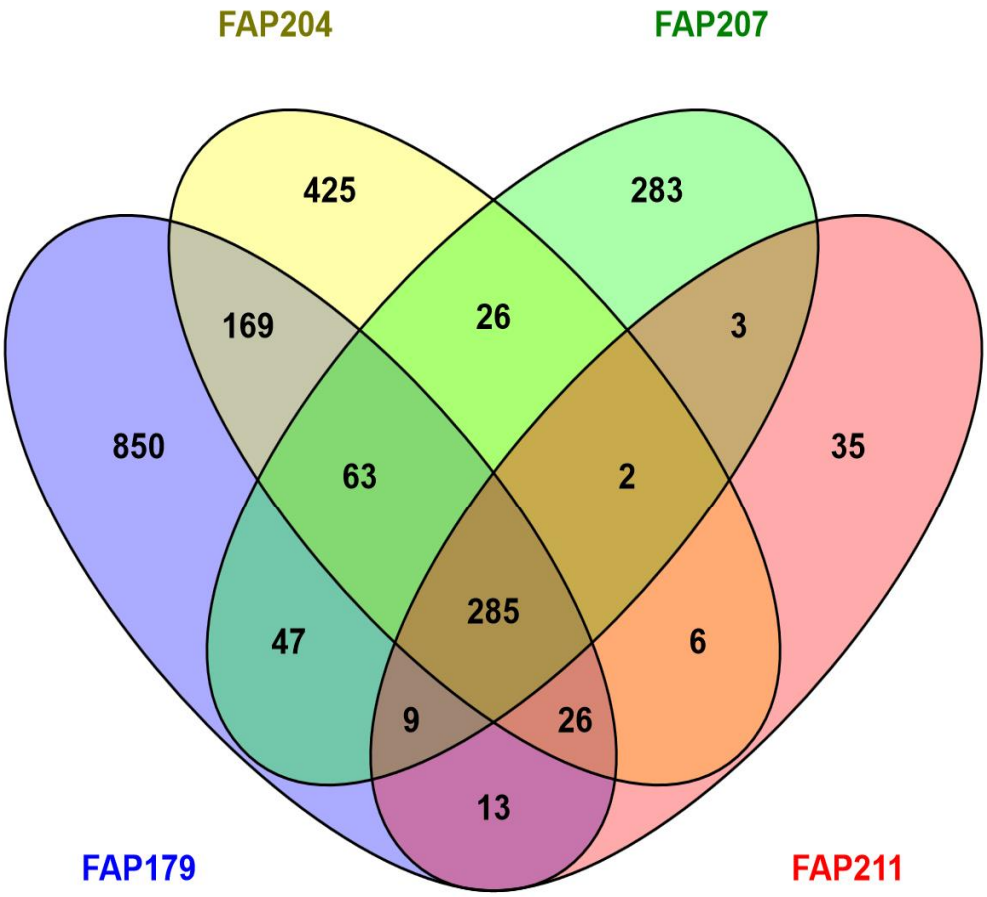


Figure 22: Hypergeometric distribution was used to gather the unresolved FAP patients common genes obtained from WES analysis.

Using this analysis, the O-linked glycans pathway of the mucins resulted to be highly enriched. In particular they were identified five common mucin genes related both to secreted mucins (e.g. *MUC2,* MUC5B, MUC6) and to cell surface mucins (e.g. MUC4, MUC16). The total number of variants belonging to these genes was 44, however only 11 of these resulted to be ñpossibly-probably damagingò using the Polyphen-2 prediction software (Table 11).

It will be important to further study the identified mucin gene variations in unresolved FAP patients, to confirm the possible causative role of these genes in polyposis.

gene	transcript_ID	cDNA_change	AA_change	Polyphen-v2	Score
MUC6	NM_005961	C5732T	T1911M	probably damaging	1
MUC5B	NM_002458	G7051A	V2351I	probably damaging	0,749
MUC5B	NM_002458	C13211G	A4404G	probably damaging	0,86
MUC16	NM_024690	C42148G	P14050A	probably damaging	0,986
MUC16	NM_024690	A40697G	Q13566R	possibly damaging	0,679
MUC16	NM_024690	C19331T	T6444I	possibly damaging	0,842
MUC16	NM_024690	A4200T	K1400N	possibly damaging	0,614
MUC4	NM_018406	G7693A	D2565N	possibly damaging	0,845
MUC4	NM_018406	C7685A	P2562H	probably damaging	0,994
MUC4	NM_018406	A6409G	T2137A	possibly damaging	0,528
MUC4	NM_018406	A6344T	D2115V	probably damaging	0,979

Table 11: List of possibly-probably pathogenetic variants of mucin genes selected with PolyPhen-v2 software.

4. DISCUSSION

Hereditary factors play a role in around 5ï 15% of all CRC cases. Colorectal cancer caused by highly penetrant mutations, such as those involved in tumor suppression genes or the DNA mismatch repair system, include Lynch syndrome, Familial Adenomatous Polyposis and MutHY-associated Polyposis. A prompt identification of individuals with a genetic predisposition to CRC is imperative, as it enables the use of preventative measures (i.e. the colonoscopy). Periodic examination by colonoscopy has been highly effective in reducing CRC-associated mortality in individuals at high risk of CRC. Polyposis syndromes are fairly easy to recognize, but some patients might have characteristics that overlap with other clinically defined syndromes. It has been shown in literature that a comprehensive analysis of the genes known to be associated with polyposis syndromes helps to establish the final diagnosis in these patients (Vasen et al., 2015). The FAP hallmark is the development of hundreds to thousands of colon and rectum adenomas, which confer a near 100% risk of CRC development by age 40, in the absence of any medical intervention. *APC* gene mutations are detected in more than 80% of patients with classical FAP and up to 30% of individual with classic polyposis but without a detectable *APC* mutation, may have bilallelic mutations in the *MutHY* gene (Kastrinos and Syngal, 2011).

In this study four patients, showing at the colonoscopy hundreds adenomatous polyps (commensurable to full-blown FAP patients) but not having mutations in *APC* or *MutYH* genes, were analyzed. In details, these patients were defined as ñunresolved FAP patientsò and their interactome was investigated, considering that they were characterized by a very severe phenotype leading to CRC development. The interactome is defined as a biological network including the whole set of direct and indirect molecular interactions in a cell. In the -*omics* era, it is possible to combine genomic and proteomic data to study the biological network of a subject. For example in the study performed by Agostini *et al.* (Agostini et al., 2015), the authors used the microarray data of the responder and not responder rectal cancer patients that had neoadjuvant radio chemotherapy before surgery. In order to remove redundant information, using statistical analysis together with a logistic model, they were able to select 4 genes. Only with this network analysis it was be possible to correlate 3 of these genes with another one, whereas with other methods they did not come up at all.

In this thesis the peptidome and genome data were combined to characterize unresolved FAP patient. The first step of this work was to perform a peptidomic analysis of the mutated FAP patients. Afterwards, the peptidomic and genomic analysis of unresolved FAP patients was done in order to define a specific pattern of these patients.

Peptidomic data of mutated FAP patients

In the peptidomic profile of mutated FAP patients, the low molecular weight fraction of the blood proteome, derived from the altered tumor protease activity, was taken in consideration. This represents, indeed, a promising source of novel human diseases biomarkers. The plasma peptide profile of FAP patients was compared with adenoma patients, CRC patients and healthy controls, to find a FAP disease specific peptide fingerprint. Twelve ionic species (Table 5, in bold), deriving from circulating peptides, showed a peculiar pattern in FAP patients. These peptides can be divided into three different groups: 1) C3f and its fragments; 2) a peptide from C4A/B and its fragments and 3) any blood clotting proteins. In order to give a prognostic biomarker for FAP patients, four peptides were selected as possible distinctive key regulators in the switch between adenoma and malignant carcinoma. Complement C3f and some of its fragments with a peculiar decrease trend from FAP to adenoma and CRC patients were identified. Among the identified peptides, there were mainly fragments belonging to the physiological circulating proteins and a group of these peptides are known to be involved in the inflammatory response. Inflammation orchestrates the microenvironment around tumors, contributing to the cell proliferation, survival and migration (Ungefroren et al., 2011). On the other hand, many cells of the immune system contribute to the cancer control, suppressing cancer growth (Gunn et al., 2012). FAP patients develop hundreds to thousands of adenomatous polyps that could explain the presence of a concomitant altered inflammatory response. The complement system is known to be involved in the innate immunity and plays an important role in the surveillance against tumors (Janssen et al., 2005). For decades, the complement has been recognized as an effector part of the immune system that contributes to the destruction of tumor cells. However, recent studies have challenged this paradigm by demonstrating a tumor-promoting role for the complement (Pio et al., 2013). Bonavita and collaborators (Bonavita et al., 2015) described for the first time that a regulatory component of the humoral part of the innate immunity, pentraxin *PTX3*, acts as an extrinsic oncosuppressor gene in mouse and human. There are three distinct pathways of the complement activation, all of them joined at the level of the complement component

C3 convertase that cleaves C3 to generate C3a and C3b. C3b stimulates the opsonization by phagocytic cells and enhances clearance of immune complexes. C3b can also take part of C5 convertase (Cook and Botto, 2006). The classical pathway of C3 convertase is composed by C2b and C4b, whereas the C5 convertase is composed by C3b and C3 convertase (Morgan et al., 2011). In this project, it was possible to describe a specific pattern, almost unique, for FAP disease characterized by lower levels of some of the C4b fragments and an increased level of C3f (Figure 17, panels a and b) and some of its ladder-like fragments (Table 5) derived from exoprotease activity. Altered levels of these peptides could be due to different plasma concentration of the C3b and C4b precursor proteins, or to different protease activity of Factor I and H, that have a central role in the regulation of complement system, or to tissue exoproteases related to cancer activity (Villanueva et al., 2006). From the ELISA assay it was possible to conclude that C4b and C3b precursors do not increase in FAP patients, suggesting that the altered levels of the C4A/B fragment at m/z 1896.03 seem more likely to derive from protease activity than from changes in the levels of circulating precursor. Furthermore, the observed increase of C3f was not related to changes in the precursor levels (Figure 19, panel a) and is due to an altered endoprotease activity. The same endoproteases determine the increase of the fragment at m/z 1211.66, whereas the increase of ladder-like end fragments could be derived from exoprotease activity.

Previous studies have reported increased levels of the C3f in the serum of patients affected by several neoplasm such as breast, bladder, thyroid, prostate cancer (Villanueva et al., 2006) but not in CRC cancer, where the presence of this peptide is decreased (Bedin et al., 2015) (Zhu et al., 2013). Although Ornellas et collaborators (Ornellas et al., 2012) found that the peptide C3f at m/z 2021.99 and C4A/B at m/z 1896.03 showed a low presence in squamous cell carcinoma of the penis patients, Profumo and collaborators (Profumo et al., 2013) found that the higher levels of C3f present in the serum of women affected by gross cystic disease (a breast benign condition predisposing to breast cancer) is correlated to the development of breast cancer even 20 years before. They explained the higher presence of C3f derived peptides as the result of a different concentration of the C3b precursor protein rather than a different peptidase activity.

Complement Factor I is responsible of cleaving both C4b and C3b. Factor I is a serine protease that cleaves and inactivates the complement components by preventing the assembly of the C3 and C5 convertase enzymes (GeneCards). Degradation of C4b into

the inactive fragments C4c and C4d by Factor I in the presence of the cofactors C4-binding protein and Factor H, blocks the generation of C3 convertase (UniProt database). The results in this project suggest that the peculiar FAP peptide profile can be a consequence of changes in Factor I activity or its cofactors. Okroj and collaborators (Okroj et al., 2008) reported that many non small cell lung cancer cell lines secrete the soluble complement inhibitors Factor I. Furthermore Riihila and coauthors described an overexpression of Factor I in the cutaneous squamous cell carcinoma *in vitro* (Riihila et al., 2015). Bonavita and collaborators (Bonavita et al., 2015) reported that the increase complement activation associated to *PTX3* deficiency due to the lack of recruitment of factor H has a major role in sustaining an exacerbated inflammatory response and enhanced carcinogenesis. PTX3 interacts and recruits Factor H by binding domain 19-20 and 7 through its N-terminal and glycosylated C-terminal domains respectively, without interfering with Factor H capacity to negatively regulate the complement cascade (Deban et al., 2008). *PTX3* gene methylation was detected in CRC stages I-IV as well as in the adenoma and this methylation progressively increased from normal colon epithelium, to adenoma and to CRC independently of the stage (Bonavita et al., 2015). It can be assumed that as the methylation of PTX3 increases, the recruitment of factor H and the inactivation of the complement system decrease.

Taken into account all these results, the combination between the change in the activity of Factor I and its cofactors in association with a deregulation of exoprotease activity (Villanueva et al., 2006) could explain the FAP peptide pattern that it was observed in this work.

Furthermore, in the ELISA assay it was observed an increase of C3b precursor protein in CRC patients compared to FAP patients and control subjects, whereas the peptidome showed an opposite trend for C3f and its fragments. However, for a correct analysis of this results it is necessary to keep in mind that cancer is a multifactorial disease. In this context, the increased levels of C3b in CRC patients probably indicates an activation of the complement system and it may be considered just one of these factors. In fact, activated complement proteins play a role in tumor defense directly through complement-dependent cytotoxicity and indirectly through antibody-dependent cell-mediated cytotoxicity. Neoplastic cells are known to express a wide variety of defenses against complement-mediated attack. The antagonist interaction between the complement and tumor cells, in which the tumor escape is relatively facilitated by the neutralization of the

complement attack, underscores the opposing roles of complement in carcinogenesis (Rutkowski et al., 2010).

In conclusion, in this mutated FAP peptidomic study was described for the first time a specific plasma peptide pattern that could recognize the FAP patients and, moreover, that could predict the malignancy progression allowing the delay of the colonoscopy screening and surgery. In CRC patients it was observed an increased level of C3b precursor but not of the C3f inactivation product and its peptides; in FAP patients it was observed the opposite condition, whereas in adenoma patients there was an intermediate situation. Further studies could elucidate the role of complement system Factor I and its cofactor H in FAP pathology progression. In addition, an absolute quantification of C3f level in FAP patients should be done in order to define the maximum C3f level under which the surgery is necessary.

Peptidomic and WES data of unresolved FAP patients

In the study of **peptidomic profile** of unresolved FAP patients, the analysis was focused on the ionic species previously described in the results (unresolved FAP peptidome paraghraph of result section) and characteristic of mutated FAP patients. In particular, it was found that the C3f fragment at m/z 2021.10 and C4A/B precursor fragment at m/z 1896.03 in unresolved FAP patients have characteristic similar to the other groups (control subjects, adenoma patients, CRC patients) but not to mutated FAP patients. These data clearly indicate and sustain the differences between these two groups.

In order to be able to identify a specific pattern for unresolved FAP patients, **WES analysis** was performed. From these datasets, only the variants present in all the patients were selected, identifying 285 common genes. Considering these common genes, the pathway enrichment analysis was performed and the O-linked glycans pathway of the mucins was found to be the most represented. O-linked glycosylation is a common covalent modification of serine and threonine residues of mammalian glycoproteins that occurs in the Golgi apparatus in the eukaryotes. In particular a list of 44 variants present in 5 common mucin genes was identified (both secreted mucins *MUC2, MUC5B, MUC6* and cell surface mucins *MUC4, MUC16)*, 11 of these variants were classified as possibly-probably damaging using the Polyphen-2 prediction software.

Recent studies have demonstrated that the chronic inflammation leads to altered mucin expression and glycosylation (Sheng et al., 2012). Furthermore, the expression of mucin genes and the distribution of their product is dramatically altered in certain types of

colorectal polyps and neoplasms. These alterations take place through several mechanisms including the aberrant glycosylation of the mucin side chains, the immunoreactivity of the mucin core peptide, the deletion of the normally expressed antigens, the expression of the blood-group incompatible antigens and the *de novo* appearance of new antigens (Molaei et al., 2010).

Mucin depletion foci (MDF) are precancerous lesions of the colon, formed by dysplastic crypts lacking mucin production. They have been identified in carcinoma-treated rodents and high risk humans colorectal cancers. MDF carry molecular defects proper of colon tumor such as *APC* and *CTNNB1* gene mutations. Phenotypically, these lesion lead to a defective mucin production due to the lack of MUC2 expression. MUC2 is the main mucin produced by the intestinal goblet cells. It has been reported that genetically MUC2-deficent mice develop colitis and CRC spontaneously (Sakai E. et al., 2011). In humans several studies have showed diminished MUC2 mRNA expression in CRC patients (Sakai et al., 2012). Femia A.P. and collaborators have identified MDF both in the unsectioned colons of FAP patients and in patients with sporadic CRC, although in these at a lower density (Femia et al., 2012). Yoshimi and collaborators (Sakai et al., 2012) demonstrated that MDS in FAP patient show a moderate grade of dysplasia with a slight nuclear stratification, loss of polarity and Paneth cell metaplasia, whereas in sporadic CRC patients MDF change slightly in relation to a low-grade dysplasia. They also defined that the inflammatory cell infiltration is a specific histological feature of MDF. Furthermore, it has been shown that MUC1 and MUC4 can directly disrupt epithelial cell tight junctions through HER2 activation and promote and/or maintain loss of the epithelial cells polarity (Sheng et al., 2012). These results indicate that the alteration of MUC2 expression and the lack of the mucus protective layer might activate the local inflammation and contribute to MDF progression to a more advanced stage of colorectal carcinogenesis in humans.

Based on the importance of the mucins in the maintenance of a normal mucosa, their alterations in the various inflammatory states and cancer, and take in consideration that FAP patients present MDS (due to a deficiency in the mucus production), it will be important to further study the identified possibly-probably pathogenetic mucin gene variations, in unresolved FAP patients, to confirm the possible causative role of these genes in polyposis.

In conclusion, in this study, the combination of high-throughput techniques such as MALDI-TOF mass spectrometry together with exome sequencing gave the possibility

to define a preliminary pattern for unresolved FAP patients. Peptidomic analysis clearly define a substantial difference between mutated and unresolved FAP patients. Finally WES data suggest that mucin genes are most probably involved in the polyposis genesis of the unresolved FAP patients, although to confirm it future analysis of this pattern are necessary.

Clear results to sustain this hypothesis will be definitely fundamental and helpful not only to better understand the clinical cases that normally do not show the common FAP mutations, but also to be able to define the interactome of the unresolved FAP patients. Thus, merging peptidomic and WES data, could be useful to study the specific physical interactions among molecules (protein-protein interactions) and also describe sets of indirect interactions among genes (genetic interactions).

5. REFERENCES

- Agostini, M., Zangrando, A., Pastrello, C., D'Angelo, E., Romano, G., Giovannoni, R., Giordan, M., Maretto, I., Bedin, C., Zanon, C., *et al.* (2015). A functional biological network centered on XRCC3: a new possible marker of chemoradiotherapy resistance in rectal cancer patients. Cancer Biol Ther *16*, 1160-1171.
- Bedin, C., Crotti, S., Ragazzi, E., Pucciarelli, S., Agatea, L., Tasciotti, E., Ferrari, M., Traldi, P., Rizzolio, F., Giordano, A., *et al.* (2015). Alterations of the Plasma Peptidome Profiling in Colorectal Cancer Progression. J Cell Physiol.
- Behrens, J., von Kries, J. P., Kuhl, M., Bruhn, L., Wedlich, D., Grosschedl, R., and Birchmeier, W. (1996). Functional interaction of beta-catenin with the transcription factor LEF-1. Nature *382*, 638-642.
- Bhanot, P., Brink, M., Samos, C. H., Hsieh, J. C., Wang, Y., Macke, J. P., Andrew, D., Nathans, J., and Nusse, R. (1996). A new member of the frizzled family from Drosophila functions as a Wingless receptor. Nature *382*, 225-230.
- Bisgaard, M. L., Fenger, K., Bulow, S., Niebuhr, E., and Mohr, J. (1994). Familial adenomatous polyposis (FAP): frequency, penetrance, and mutation rate. Hum Mutat *3*, 121-125.
- Bodmer, W. F., Bailey, C. J., Bodmer, J., Bussey, H. J., Ellis, A., Gorman, P., Lucibello, F. C., Murday, V. A., Rider, S. H., Scambler, P., and et al. (1987). Localization of the gene for familial adenomatous polyposis on chromosome 5. Nature *328*, 614-616.
- Bonavita, E., Gentile, S., Rubino, M., Maina, V., Papait, R., Kunderfranco, P., Greco, C., Feruglio, F., Molgora, M., Laface, I., *et al.* (2015). PTX3 is an extrinsic oncosuppressor regulating complement-dependent inflammation in cancer. Cell *160*, 700-714.
- Cook, H. T., and Botto, M. (2006). Mechanisms of Disease: the complement system and the pathogenesis of systemic lupus erythematosus. Nat Clin Pract Rheumatol *2*, 330-337.
- Deban, L., Jarva, H., Lehtinen, M. J., Bottazzi, B., Bastone, A., Doni, A., Jokiranta, T. S., Mantovani, A., and Meri, S. (2008). Binding of the long

pentraxin PTX3 to factor H: interacting domains and function in the regulation of complement activation. J Immunol *181*, 8433-8440.

- Fearon, E. R., and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. Cell *61*, 759-767.
- Femia, A. P., Swidsinski, A., Dolara, P., Salvadori, M., Amedei, A., and Caderni, G. (2012). Mucin Depleted Foci, Colonic Preneoplastic Lesions Lacking Muc2, Show Up-Regulation of Tlr2 but Not Bacterial Infiltration. PLoS One *7*.
- Fodde, R., Kuipers, J., Rosenberg, C., Smits, R., Kielman, M., Gaspar, C., van Es, J. H., Breukel, C., Wiegant, J., Giles, R. H., and Clevers, H. (2001). Mutations in the APC tumour suppressor gene cause chromosomal instability. Nat Cell Biol *3*, 433-438.
- Gunn, L., Ding, C., Liu, M., Ma, Y., Qi, C., Cai, Y., Hu, X., Aggarwal, D., Zhang, H. G., and Yan, J. (2012). Opposing roles for complement component C5a in tumor progression and the tumor microenvironment. J Immunol *189*, 2985-2994.
- Half, E., Bercovich, D., and Rozen, P. (2009). Familial adenomatous polyposis. Orphanet J Rare Dis *4*, 22.
- He, T. C., Sparks, A. B., Rago, C., Hermeking, H., Zawel, L., da Costa, L. T., Morin, P. J., Vogelstein, B., and Kinzler, K. W. (1998). Identification of c-MYC as a target of the APC pathway. Science *281*, 1509-1512.
- Herrera, L., Kakati, S., Gibas, L., Pietrzak, E., and Sandberg, A. A. (1986). Gardner syndrome in a man with an interstitial deletion of 5q. Am J Med Genet *25*, 473-476.
- Hisamuddin, I. M., and Yang, V. W. (2004). Genetics of colorectal cancer. MedGenMed *6*, 13.
- Janssen, B. J., Huizinga, E. G., Raaijmakers, H. C., Roos, A., Daha, M. R., Nilsson-Ekdahl, K., Nilsson, B., and Gros, P. (2005). Structures of complement component C3 provide insights into the function and evolution of immunity. Nature *437*, 505-511.
- Jin, L. H., Shao, Q. J., Luo, W., Ye, Z. Y., Li, Q., and Lin, S. C. (2003). Detection of point mutations of the Axin1 gene in colorectal cancers. Int J Cancer *107*, 696-699.

- Kastrinos, F., and Syngal, S. (2011). Inherited colorectal cancer syndromes. Cancer J *17*, 405-415.
- Kinzler, K. W., Nilbert, M. C., Su, L. K., Vogelstein, B., Bryan, T. M., Levy, D. B., Smith, K. J., Preisinger, A. C., Hedge, P., McKechnie, D., and et al. (1991). Identification of FAP locus genes from chromosome 5q21. Science *253*, 661-665.
- Levy, D. B., Smith, K. J., Beazer-Barclay, Y., Hamilton, S. R., Vogelstein, B., and Kinzler, K. W. (1994). Inactivation of both APC alleles in human and mouse tumors. Cancer Res *54*, 5953-5958.
- Ley, T. J., Minx, P. J., Walter, M. J., Ries, R. E., Sun, H., McLellan, M., DiPersio, J. F., Link, D. C., Tomasson, M. H., Graubert, T. A., *et al.* (2003). A pilot study of high-throughput, sequence-based mutational profiling of primary human acute myeloid leukemia cell genomes. Proc Natl Acad Sci U S A *100*, 14275-14280.
- Midgley, C. A., White, S., Howitt, R., Save, V., Dunlop, M. G., Hall, P. A., Lane, D. P., Wyllie, A. H., and Bubb, V. J. (1997). APC expression in normal human tissues. J Pathol *181*, 426-433.
- Molaei, M., Mansoori, B. K., Mashayekhi, R., Vahedi, M., Pourhoseingholi, M. A., Fatemi, S. R., and Zali, M. R. (2010). Mucins in neoplastic spectrum of colorectal polyps: can they provide predictions? BMC Cancer *10*.
- Morgan, H. P., Jiang, J., Herbert, A. P., Kavanagh, D., Uhrin, D., Barlow, P. N., and Hannan, J. P. (2011). Crystallographic determination of the disease-associated T1184R variant of complement regulator factor H. Acta Crystallogr D Biol Crystallogr *67*, 593-600.
- Okroj, M., Hsu, Y. F., Ajona, D., Pio, R., and Blom, A. M. (2008). Non-small cell lung cancer cells produce a functional set of complement factor I and its soluble cofactors. Mol Immunol *45*, 169-179.
- Ornellas, P., Ornellas, A. A., Chinello, C., Gianazza, E., Mainini, V., Cazzaniga, M., Pereira, D. A., Sandim, V., Cypriano, A. S., Koifman, L., *et al.* (2012). Downregulation of C3 and C4A/B complement factor fragments in plasma from patients with squamous cell carcinoma of the penis. Int Braz J Urol *38*, 739-749.

- Petersen, G. M., Slack, J., and Nakamura, Y. (1991). Screening guidelines and premorbid diagnosis of familial adenomatous polyposis using linkage. Gastroenterology *100*, 1658-1664.
- Petricoin, E. F., Belluco, C., Araujo, R. P., and Liotta, L. A. (2006). The blood peptidome: a higher dimension of information content for cancer biomarker discovery. Nat Rev Cancer *6*, 961-967.
- Pio, R., Ajona, D., and Lambris, J. D. (2013). Complement inhibition in cancer therapy. Semin Immunol *25*, 54-64.
- Profumo, A., Mangerini, R., Rubagotti, A., Romano, P., Damonte, G., Guglielmini, P., Facchiano, A., Ferri, F., Ricci, F., Rocco, M., and Boccardo, F. (2013). Complement C3f serum levels may predict breast cancer risk in women with gross cystic disease of the breast. J Proteomics *85*, 44-52.
- Rabbani, B., Tekin, M., and Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. J Hum Genet *59*, 5-15.
- Reya, T., and Clevers, H. (2005). Wnt signalling in stem cells and cancer. Nature *434*, 843-850.
- Riihila, P., Nissinen, L., Farshchian, M., Kivisaari, A., Ala-aho, R., Kallajoki, M., Grenman, R., Meri, S., Peltonen, S., Peltonen, J., and Kahari, V. M. (2015). Complement factor I promotes progression of cutaneous squamous cell carcinoma. J Invest Dermatol *135*, 579-588.
- Rutkowski, M. J., Sughrue, M. E., Kane, A. J., Mills, S. A., and Parsa, A. T. (2010). Cancer and the complement cascade. Mol Cancer Res *8*, 1453-1465.
- Sakai, E., Morioka, T., Yamada, E., Ohkubo, H., Higurashi, T., Hosono, K., Endo, H., Takahashi, H., Takamatsu, R., Cui, C. X., *et al.* (2012). Identification of preneoplastic lesions as mucin-depleted foci in patients with sporadic colorectal cancer. Cancer Science *103*, 144-149.
- Sheng, Y. H., Hasnain, S. Z., Florin, T. H. J., and McGuckin, M. A. (2012). Mucins in inflammatory bowel diseases and colorectal cancer. J Gastroen Hepatol *27*, 28-38.
- Sjoblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N., *et al.* (2006). The consensus coding sequences of human breast and colorectal cancers. Science *314*, 268-274.

- Sparks, A. B., Morin, P. J., Vogelstein, B., and Kinzler, K. W. (1998). Mutational analysis of the APC/beta-catenin/Tcf pathway in colorectal cancer. Cancer Res *58*, 1130-1134.
- Strohalm, M., Hassman, M., Kosata, B., and Kodicek, M. (2008). mMass data miner: an open source alternative for mass spectrometric data analysis. Rapid Commun Mass Spectrom *22*, 905-908.
- Takao, M., Zhang, Q. M., Yonei, S., and Yasui, A. (1999). Differential subcellular localization of human MutY homolog (hMYH) and the functional activity of adenine:8-oxoguanine DNA glycosylase. Nucleic Acids Res *27*, 3638-3644.
- Tortora G.J., Derrickson B., Principles of anatomy and physiology (2009). John Wiley & Sons, Inc Edition.
- Ungefroren, H., Sebens, S., Seidl, D., Lehnert, H., and Hass, R. (2011). Interaction of tumor cells with the microenvironment. Cell Commun Signal *9*, 18.
- van Es, J. H., Giles, R. H., and Clevers, H. C. (2001). The many faces of the tumor suppressor gene APC. Exp Cell Res *264*, 126-134.
- Vasen, H. F., Tomlinson, I., and Castells, A. (2015). Clinical management of hereditary colorectal cancer syndromes. Nat Rev Gastroenterol Hepatol *12*, 88-97.
- Villanueva, J., Shaffer, D. R., Philip, J., Chaparro, C. A., Erdjument-Bromage, H., Olshen, A. B., Fleisher, M., Lilja, H., Brogi, E., Boyd, J., *et al.* (2006). Differential exoprotease activities confer tumor-specific serum peptidome patterns. J Clin Invest *116*, 271-284.
- Wang, J., Joshi, A. D., Corral, R., Siegmund, K. D., Marchand, L. L., Martinez, M. E., Haile, R. W., Ahnen, D. J., Sandler, R. S., Lance, P., and Stern, M. C. (2012). Carcinogen metabolism genes, red meat and poultry intake, and colorectal cancer risk. Int J Cancer *130*, 1898-1907.
- Webster, M. T., Rozycka, M., Sara, E., Davis, E., Smalley, M., Young, N., Dale, T. C., and Wooster, R. (2000). Sequence variants of the axin gene in breast, colon, and other cancers: an analysis of mutations that interfere with GSK3 binding. Genes Chromosomes Cancer *28*, 443-453.

- Zhang, X. (2014). Exome sequencing greatly expedites the progressive research of Mendelian diseases. Front Med *8*, 42-57.
- Zhu, D., Wang, J., Ren, L., Li, Y., Xu, B., Wei, Y., Zhong, Y., Yu, X., Zhai, S., Xu, J., and Qin, X. (2013). Serum proteomic profiling for the early diagnosis of colorectal cancer. J Cell Biochem *114*, 448-455.

PUBBLICATIONS

- **Agatea L**., Crotti S., Ragazzi E., Bedin C., Urso E., Mammi I., Traldi P., Pucciarelli S., Nitti D., Agostini M. Peptide patterns as discriminating biomarkers in plasma FAP patients. Clinical Colorectal Cancer: in press.
- Dyar K, Ciciliot S, Malagoli Tagliazucchi G , Pallafacchina G, Tothova J, Argentini C, **Agatea L**, Abraham R, Ahdesmäki M, Forcato M, Bicciato S, Schiaffino S, Blaauw B. The calcineurin-NFAT pathway controls activity-dependent circadian gene expression in slow skeletal muscle. Molecular Metabolism: novembre 2015; 4 (11): 823-833.
- Bedin C, Crotti S, Ragazzi E, Pucciarelli S, **Agatea L**, Tasciotti E, Ferrari M, Traldi P, Rizzolio F, Giordano A, Nitti D, Agostini M. Alterations of the Plasma Peptidome Profiling in Colorectal Cancer Progression. J Cell Physiol: settembre 2015.
- Blaauw B, Del Piccolo P, Rodriguez L, Hernandez Gonzalez VH, Agatea L, Solagna F, Mammano F, Pozzan T and Schiaffino S. No evidence for inositol 1,4,5-trisphosphateï dependent Ca2+ release in isolated fibers of adult mouse skeletal muscle. Journal Gen Physiology: luglio 2012; 140(2): 235-241.
- Blaauw B **, Agatea L**, Toniolo L, Canato M, Quarta M, Dyar KA, Danieli-Betto D, Betto R, Schiaffino S and Reggiani C. Eccentric contractions lead to myofibrillar dysfunction in muscular dystrophy. Journal of Applied Physiology; gennnaio 2010; 108(1): 105-11.
- Masiero E, **Agatea L**, Mammucari C, Blaauw B, Loro E, Komatsu M, Metzger D, Reggiani C, Schiaffino S and Sandri M. Autophagy is required to maintain muscle mass. Cell Metabolism; dicembre 2009; 10(6):507-15.
- Blaauw B, Canato M, **Agatea L**, Toniolo L, Mammucari C, Masiero E, Abraham R, Sandri M, Schiaffino S, Reggiani C. Inducible activation of Akt increases skeletal muscle mass and force without satellite cell activation. FASEB J.; novembre 2009; 23(11): 3896-905.
- Blaauw B, Mammucari C, Toniolo L, **Agatea L**, Abraham R, Sandri M, Reggiani C, Schiaffino S. Akt activation prevents the force drop induced by eccentric contractions in dystrophin-deficient skeletal muscle. Hum Mol Genet; dicembre 2008; 17(23): 3686-96.